MASTER'S THESIS

For the attainment of the degrees
Master of Science in Finance at the Stockholm School of Economics
Master of Arts in Banking and Finance at the University of St. Gallen

A Causal Analysis of Cat Bond Markets

Part I:
Analyzing Secondary Market
Cat Bond Yields
with Random Forests

Part II:
Cat Bond Markets:
A Time Analysis of the
Causal Random Forest Approach

Tim Ludwig Leonard Matheis

MASTER'S THESIS

For the attainment of the degree
Master of Arts in Banking and Finance at the University of St. Gallen

Part I

Analyzing Secondary Market Cat Bond Yields with Random Forests

Tim Ludwig Leonard Matheis

Supervisor:	Prof. Dr. Alexander Braun Prof. Dr. Despoina Makariou
Submitted:	February 17, 2023

Abstract

This work is a contribution to the causal analysis of the catastrophe bond market, which has generated high excess returns over the last two decades. Since these excess returns remain partially unexplainable and the interest in catastrophe bonds is increasing, the causal study of the factors affecting their premiums is of high relevance. Most studies about the catastrophe bond market have used only linear models. However, more complex models such as random forests may be better suited for modeling this market. Considerable progress has been made in developing methods for inference in the specific setting of random forests. Causal random forests, especially when combined with double machine learning and Shapley values, represent a sophisticated empirical toolbox. They provide unbiased prediction intervals and allow the analysis of heterogeneous effects, while Shapley values help interpret individual predictions. I apply these methods to quantify uncertainties and heterogeneities of effects in the secondary catastrophe bond market. My results confirm that the effect of the analyzed factors on the premiums is often heterogeneous. In addition to mean effects, I present median effects, and their whole distribution. Expected loss is by far the most decisive factor. To date, the interactions among predictors have not been extensively studied, and there may be non-linearities in the explanatory predictors. My additional heterogeneity analysis provides answers as to which factors cause the variance in the impact of the expected loss on the premiums.

Contents

List of Figures

List of Tables

List of Abbreviations

ANOVA Analysis of variance
ATE Average treatment effect
CART Classification and regression trees (algorithm)
CAT Catastrophe
CATE Conditional average treatment effect
CEL Conditional expected loss
CRF Causal random forest
DML Double machine learning
EL Expected loss
EQ Earthquake
EU Europe
HTE Heterogeneous treatment effect
HU Hurricane
ILS Insurance linked security
JP Japan
LA Latin America
LIBOR London inter-bank offered rate
LIME Local interpretable model-agnostic explanations
NA North America
OLS Ordinary least squares
PCA Principal component analysis
PFL Probability of first loss
PLL Probability of exhaust
RCT Randomized controlled trial
RF Random forest
RFE Recursive feature elimination
RMSE Root-mean-square error
SD Standard deviation
SHAP Shapley additive explanations

1. Introduction

More than 20 years ago, the first cat bond was issued. Unlike most other assets and bonds, catastrophe bonds are usually issued with a comprehensive risk analysis, including the expected loss, from an independent third party. On the other hand, the only risk analysis that investors receive when deciding whether to invest in corporate bonds is an imprecise letter rating from a rating agency. One might therefore assume that predicting and explaining cat bond prices would be particularly straightforward compared to other asset classes. However, to date, pricing cat bonds has been a major challenge.

Random forests and other machine learning models often provide superior predictive power. As many fundamental problems are formulated as prediction problems, evaluating goodness of fit on a test set is sufficient in such a case (Mullainathan & Spiess, 2017). However, although better performance in terms of out-of-sample predictive power is valuable in practice, in some settings a valid confidence interval is more or equally important. Efron and Hastie (2021) criticize that prediction is an area where algorithmic developments have outstripped their inferential justification. One of the reasons for this is the model-free nature behind many well-performing prediction methods. A single decision tree is easy to interpret but has low predictive power. Conversely, random forests provide high predictive power at the cost of lower interpretability. It is unclear what it means to make a final prediction based on the predictions of multiple decision trees. Thus, there is a trade-off between interpretability and predictive power.

An average treatment effect of a parameter of interest may not be sufficient because it does not capture the degree of uncertainty. More traditional models, such as regressions, provide standard errors and confidence intervals. This quantification of uncertainty is critical when deciding whether to implement a treatment. Despite the absence of a "treatment" in the cat bond market, the analysis of the interaction of different parameters is very interesting and provides insights into the pricing of this asset. Since random forests can provide very good predictive results in the cat bond market (Makariou, Barrieu, & Chen, 2021), additional causal analysis of the results obtained with random forests may be important. A recent branch of economic literature focuses on adapting and tuning machine learning techniques to causal problems in which economists are interested (Athey & Imbens, 2019). This literature provides a toolbox for uncertainty quantification, unbiased treatment estimates, and interpretability of predictions.

I apply these methods to quantify uncertainty and heterogeneity in the impact of factors affecting premiums in the secondary cat bond market. To my knowledge, no similar causal analysis has been done before. My results are based on a large sample of cat bonds, most of which I collected myself. They confirm that the effect of factors on the premiums is often heterogeneous. In addition to mean effects, I present median effects, and their whole distribution. Expected loss is by far the most decisive factor. To date, the interactions among predictors have not been extensively studied, and there may be non-linearities in the explanatory predictors. My additional heterogeneity analysis provides answers as to which factors cause the variance in the impact of the expected loss on the premiums.

This work is a comprehensive contribution to the literature on asset pricing and empirical studies on machine learning, and in particular to the causal analysis of the catastrophe bond market. I not only apply machine learning methods to predict cat bond premiums in the secondary market, but also put special emphasis on the underlying causalities of the methods (causal machine learning). In the context of cat bonds, it is worth noting that this particular bond market is relatively small and new. Consequently, the application of machine learning methods is particularly challenging due to the limited availability of data (Khandani, Kim, and Lo (2010), Gu, Kelly, and Xiu (2020)). My analysis may help practitioners evaluate the potential of causal machine learning methods for cat bond pricing and asset pricing in general. In addition to previous studies that have already shown that machine learning can perform quite well on a relatively small data set (Götze, Gürtler, & Witowski, 2020), I also explain the underlying machine learning black box. This enhances the explainability of machine learning methods. In fact, the interpretability of machine learning models helps to improve them, build confidence in them, justify model predictions, and gain insights. The increasing use of machine learning makes its interpretability even more important.

In the following, I provide a brief overview of this work. The next chapter 2 provides an overview of the existing literature. The following chapter 3 explains the methodology for causal cat bond pricing methods. Hereby, the focus is on causal random forests. The main part of my work consists of chapter 4, which contains my quantitative analysis of the secondary cat bond market. First, I explain the sample selection and present the variables used in the analysis. The following empirical part comprises the analysis of conditional average treatment effects, focusing on the heterogeneity of the effect of the most crucial characteristic on the spread, expected loss. In chapter 5, I discuss the strengths and weaknesses of this work and suggest possible areas for future research.

2. Literature overview

In my literature review, I first discuss recent research on cat bond pricing. This research focuses mainly on achieving very accurate predictions. I then review recent advances in the causal machine learning literature. In this literature, some accuracy is abandoned in favor of more explainable, interpretable results with a quantification of the uncertainty in the results.

2.1. Recent cat bond pricing literature

In recent years, many empirical studies have analyzed the pricing of cat bonds using real-market data, mostly from the primary market. In these studies, the focus is usually on identifying the factors driving prices. That is, the explanatory variables that are statistically significant and theoretically relevant are extracted. Explanatory statistical models are usually used for this purpose. The simplest and most common model is a simple linear model.

For instance, Braun (2016) analyze the primary cat bond market and confirm expected loss as the most important factor in pricing cat bonds. Other critical parameters include the covered territory, the sponsor, the reinsurance cycle, and the spreads on comparably rated corporate bonds. Their econometric cat bond pricing model exhibits a robust fit across different calibration subsamples and achieves high in-sample accuracy. The out-of-sample accuracy is still decent and higher than in the previous specifications.

Braun, Herrmann, and Hibbeln (2022) build on the previous results to determine the actual bond's historically realized excess returns. This research shows whether bond-specific determinants of coupon and yield spreads lead to realized returns. Analyzing the determinants of realized cat bond returns, they find that of all known coupon and yield spread determinants, only event risk has a significant impact on the cross-section of realized returns. Their results support a three-factor asset pricing model based on the seasonality-adjusted probability of first loss (PFL), a corresponding seasonality amplitude factor, and a corporate bond market factor. However, these results still do not explain why the cat bond market has generated high excess returns over the past two decades. If natural disaster risk is diversifiable by capital market investors, and systematic risks from the broader financial markets are minimized to an almost negligible extent, this should not be the case.

There are some limitations in this literature. In general, data availability is a strong limitation in the literature on cat bond asset pricing (Braun, Ammar, & Eling, 2019). Extreme event risks require data from a long time horizon, because securitized events only occur very infrequently. Despite events such as Katrina, the Tohoku Earthquake, and the 2017 Atlantic hurricane season, historical analysis of cat bond performance remains difficult. The time series currently available only provide data for roughly the last two decades. However, there are also constraints that are easier to resolve. First, the results could be manipulated by a selection bias. Previously, the data samples often excluded bonds with certain characteristics, observations with missing entries, and unusual bond issuances (e.g., Braun (2016), Galeotti, Gürtler, and Winkelvos (2013)). Second, predictor interactions have not been examined, and there may be non-linearities in the explanatory predictors (Papachristou, 2011). Third, the vast majority

of studies only used linear models. However, more complex models may be better suited for modeling the cat bond market. Finally, the study objective in this literature was not purely predictive. Therefore, other, more complex models could potentially provide more accurate results.

To extract the relevant factors, most authors use a multivariate linear regression analysis based on empirical data from the primary ILS market. Occasionally, data from the secondary market is also used. This may be because it is more difficult to obtain. Two statistics are primarily used to compare different models: the (adjusted) R^2 and the standard error. The R^2 is an indicator for the explanation of the spread or premium from the explanatory variables. It ranges from 0 to 1. If it is high – close to 1 – most of the variation within the premium is explained by the explanatory variables. Indeed, many models exhibit a high R^2. But this does not necessarily mean that a model is also a good predictor of the future. When predicting spreads in the future for data not used to train the model, the performance is often rather disappointing. The standard error is a common measure for testing the quality of future predictions. When analyzing multiple data points, the $RMSE$ is often used. This measure is just a slight deviation from the standard error. Due to their unreliable predictive power, the question arises whether linear models are too restrictive for explaining cat bond spreads. Therefore, machine learning models that allow for more complex relationships may be more appropriate.

In fact, other, more complex models have also been studied. Makariou et al. (2021) and Götze et al. (2020) are among the first to use random forest approaches to predict cat bond premia. In these works, random forests are favored over other machine learning methods, as they provide highly accurate predictions by resolving the trade-off between over-fitting and prediction accuracy. Compared to other machine learning methods, random forests seem to outperform other statistical methods such as neural networks or linear regressions in the application domain of the cat bond market (Götze et al., 2020). Some of the previously discussed limitations can be overcome by random forests compared to linear regressions. First, the problem of potential non-linearities in the cat bond market can be addressed because random forests make no assumptions about the underlying data generative process. Second, the tree structure allows to capture potential interactions of factors without explicitly specifying them. Third, random forests are relatively robust to outliers because there are a variety of different regression trees. Even if the cat bonds are heterogeneous, this is not a major problem, and there is no loss of information since removal of outliers is not required. Fourth, there are measures of variable importance to filter out the most influential factors. Finally, data pre-processing and hyperparameter tuning is minimal since most steps are integrated into the method itself.

Makariou et al. (2021) focus on the analysis of random forest models, which are compared to linear regression models serving as a baseline, while no other advanced machine learning methods are considered. In contrast, Götze et al. (2020) analyze several machine learning models such as neural networks and "advanced" regression models such as Lasso. Nevertheless, the variables considered and the methodology are similar. First, the authors of both papers consider the time structure in the partitioning of the data into train and test data. When not taking into account the time structure of the data, much of the actual information in an out-of-sample forecasting model may be lost because the model cannot be tested for robustness to time shifts in the data set (Braun, 2016). Second, both include macroeconomic variables in their models because the literature shows their relevance for cat bond pricing (Braun, 2016). Third, in both papers hyperparameter tuning is performed, which is considered crucial for the

performance of machine learning models and to avoid overfitting.

Since there is not much causal theory for random forests, Makariou et al. (2021) assess the importance of factors by using methods such as permutation importance and minimal depth. In addition, they evaluate the sensitivity of the predictive accuracy of random forests versus their benchmark model by simultaneously removing predictors. This also sheds light on predictor interactions and the ability of existing variables to provide predictive power of missing factors.

2.2. Recent advances in causal machine learning literature

Machine learning models historically provide few explanations for their predictions, which is why such models are often referred to as black boxes. Supervised learning, as in the case of random forests, usually only provides a prediction function. Although the predictive models often achieve very high predictive accuracy, the coefficients and confidence intervals are often missing, and interpretability is usually much lower compared to traditional statistical models. However, causal machine learning has become popular and there are some tools for quantifying *uncertainty*, assessing *causality*, and obtaining *interpretable* results.

Uncertainty Regarding uncertainty quantification, conformal prediction aims at converting weak uncertainty scores into rigorous prediction intervals (Angelopoulos & Bates, 2021). Importantly, these intervals are valid in a distribution-free sense. More precisely, they possess explicit, non-asymptotic guarantees even without distributional assumptions or model assumptions. The underlying idea is to calibrate a heuristic uncertainty score derived from a separate calibration dataset.

Causality Causality is required to assess whether an intervention – e.g., a treatment or policy change – has an effect on the outcome. A recent branch of economic literature is concerned with adapting and tuning machine learning techniques to the problems that economists are interested in (Athey & Imbens, 2019). One very crucial adaptation is to exploit the structure of the problems. This includes for example the causal nature of many estimators, the endogeneity of variables, the configuration of complex data such as panel data, the nature of discrete choice among a set of substitutable products, and the presence of credible constraints motivated by economic theory. Another adaptation is to modify the optimization criteria of machine learning algorithms to prioritize considerations of causal inference. Examples include the need to control for confounders and the discovery of heterogeneity in treatment effects.

In the case of random forests, many recent papers have focused on capturing heterogeneity in a key parameter of interest, which is often the treatment variable (Athey, Tibshirani, and Wager (2019), Wager and Athey (2018), Athey and Imbens (2016)). The goal is to obtain unbiased estimators of the treatment effect for different sample groups. Strictly speaking, the treatment effect could depend on the other variables. Instead of looking at the average treatment effect, the treatment effect should be analyzed for different magnitudes of the covariates. If the treatment effect differs a lot between different sample groups, the average treatment effect is not helpful for causal inference. For instance, the treatment effect could be 0, although it is very positive for half of the sample and very negative for the other half because a covariate differs between the two sample groups. In this case, it would be helpful to partition the covariate space based on the heterogeneity of the treatment effects. This would be possible with a "normal" random forest if all counterfactuals were directly observable. The treatment could be the target

variable of the random forest. However, this is usually not possible, as counterfactuals are often missing. Other adaption techniques are sample splitting and orthogonalization. Sample splitting uses different data for model selection than for parameter estimation (Athey and Imbens (2016), Wager and Athey (2018)).

Orthogonalization as in Chernozhukov et al. (2018) can be applied to enhance the performance of machine learning estimators. A very common problem is the correlation of confounders with both the analyzed covariate, that could be the treatment, and the outcome variable. Typical approaches for correcting for unwanted correlations are the usage of instrumental variables and partial regressions. These concepts can also be applied to machine learning. When supervised machine learning is used to learn the functions, this tends to introduce overfitting and regularization biases. Double machine learning, also called orthogonal machine learning, tries to correct for these two biases (Chernozhukov et al., 2018). By training two models, causal inference is possible in supervised learning. One model is trained to predict the treatment from the confounders, and the other one is trained to predict the target from the confounders. This separation by two models fixes the problem of a potential regularization bias. In a final step, a linear regression is utilized to regress the previously predicted target on its predicted treatment. This takes care of the overfitting issue. In the case of random forests, this may lead to desirable properties such as asymptotic normality of the machine learning estimators (Athey et al., 2019).

Interpretability For the interpretation of results, some model-agnostic methods can be applied to any machine learning model (Molnar, 2022). Therefore, they offer a lot of flexibility. Hereby, global methods describe the average behavior of a model (e.g., permutation feature importance and partial dependence plots for feature effects), while local methods explain individual predictions (e.g., SHAP, LIME).

3. Methodology for random forests & causal inference

Machine learning offers automated procedures for predicting phenomena based on their past observations. In this way, underlying patterns in the data are revealed, which may provide new insights into (causal) relationships. However, the application of algorithms demands an understanding of the underlying mechanisms, assumptions, and limitations for the interpretations of their results.

This chapter provides an overview of the machine learning techniques applied in this work. Mainly, two classes of algorithms are discussed: decision trees (Breiman, Friedman, Olshen, & Stone, 1984) and random forests (Breiman, 2001). These algorithms have proven to be an accurate and robust tool for solving many machine learning tasks such as regression, classification, density estimation and semi-supervised learning (Criminisi & Shotton, 2013).

Since random forests are not optimal for causal inference, generalized random forest have been developed (Athey et al., 2019). Generalized random forests offer new methods for three statistical tasks: non-parametric quantile regression, conditional average partial effect estimation, and heterogeneous treatment effect estimation. In addition, I explain double machine learning and Shapley values, which are important tools in causal machine learning.

3.1. Machine learning

Machine learning can be described as the study of systems that learn from data without being explicitly programmed. The learning from data is reflected in an increasing performance measure as additional data is utilized in the learning process. Yet, machine learning should not be limited to producing algorithms that make accurate predictions. In addition, machine learning algorithms aim at providing insights into the predictive structure of the data (Breiman et al., 1984). More specifically, I am interested in extracting the variables and interactions between variables driving a phenomenon. Otherwise, it is difficult to accept or trust the results from a "black box" when the process leading to the results is incomprehensible.

A supervised learning task can be stated as learning a function $\varphi : \mathcal{X} \to \mathcal{Y}$ from a learning set $\mathcal{L} = (\mathbf{X}, \mathbf{y})$ (Louppe, 2014). The goal of the task is to find a model that yields predictions $\varphi(\mathbf{x})$, often denoted by the variable $\hat{Y}$, that closely approximate the true outcome variable Y. However, since one is usually interested in applying such a model to unseen data, the model should learn general relationships rather than over-fitting the data. Hence, instead of minimizing the error for the known learning set $\mathcal{L}$, one aims at minimizing the error for all possible values $(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}$. Precisely, the objective is to minimize the expected prediction error of the model $\varphi_{\mathcal{L}}$, as defined in Equation 3.1. In this formula, $\mathcal{L}$ is the learning set used to build the model $\varphi_{\mathcal{L}}$, and L is a loss function measuring the discrepancy between the two arguments, the predicted outcome and the actual outcome.

$$Err(\varphi_{\mathcal{L}}) = \mathbb{E}_{X,Y}\left\{L\left(Y, \varphi_{\mathcal{L}}(X)\right)\right\} \tag{3.1}$$

In this work, X represents a set of characteristics of a cat bond, such as the volume size and the sponsor. Therefore, $\mathbf{x}$ is specific combination of features of a bond. Since Y represents the expected cat bond returns in this work, it is a numerical variable, making the learning task a regression problem. For such a task, a simple loss function is the squared error loss, as displayed in Equation 3.2.

$$Err(\varphi_{\mathcal{L}}) = \mathbb{E}_{X,Y}\left\{(Y - \varphi_{\mathcal{L}}(X))^2\right\} \tag{3.2}$$

Whenever the outcome variable Y is instead a categorical variable, the learning task is a classification problem. To make the learning problem in this work a classification task, I could classify the expected cat bond returns according to their magnitude, e.g., high, moderate, low. This would transform Y from a numerical variable into a categorical variable. A possible loss function for a classification task would be the zero-one loss function, $L(Y, \varphi_{\mathcal{L}}(X)) = 1(Y \neq \varphi_{\mathcal{L}}(X))$, where all misclassifications are penalized equally. Here, the error would become the probability of misclassification.

In contrast, unsupervised machine learning algorithms learn from unlabeled data. This means that the learning set does not contain data couples of the features and the outcome, $(\mathbf{x}, y)$. Crucially, y is not given. A typical example is clustering. Unsupervised learning, however, is not relevant to this work.

3.2. Tree-based methods

The ambition behind the application of machine learning is usually twofold – making accurate predictions, while allowing the extraction of knowledge. Tree-based models are one of the most promising techniques, as they deliver reliable and understandable results. They are used for classification and regression tasks.

For simplicity, I explain the underlying logic for a classification problem. This is not a problem because the expected cat bond returns could be grouped by their values or simply rounded to make them a finite set of values. The expected cat bond returns Y can then be expressed as a partition over the universe $\Omega = \Omega_{c_1} \cup \Omega_{c_2} \cup ... \cup \Omega_{c_j}$. Thus, the output space consists of j categories. Likewise, a classifier φ can determine a partition of the universe Ω by approximating the "true" cat bond returns Y by $\hat{Y}$. In this case, the partition is defined over the input space X: $X = X^{\varphi}_{c_1} \cup X^{\varphi}_{c_2} \cup ... \cup X^{\varphi}_{c_j}$. Precisely, an input value – could be a vector of multiple explaining variables – is mapped to one of j output categories based on the magnitude and sign of the explaining variables. In the case of cat bonds, a cat bond could in theory be specified to have a high expected return if the bond is triggered by even mild wind storms at the West coast of the US. In this simple example, there could be two explaining variables – type of catastrophe and country – that allow for splitting the tree accordingly. Thus, the idea behind tree structured models is to approximate the partition of a model by recursively partitioning the input space X into subspaces. Consequently, predictive values $\hat{y}$, such as high expected return, can be assigned to all objects $\mathbf{x}$ within each terminal subspace.

I now briefly introduce some required concepts. Trees are a way of representing a model $\varphi : \mathcal{X} \to \mathcal{Y}$, where an outcome y is determined by its explaining variables X. A tree is a graph $G = (V, E)$, where V denotes its vertices and E its edges. In such a graph, any two vertices are connected by one path. This is a logical requirement for the concise mapping of features to an outcome. A tree usually has a root (rooted tree), which represents the whole input space X. Starting from the root, a path leads to the tree's leaves, the terminal vertices. When an

edge leads from one vertex to a vertex below, the upper vertex is called the parent vertex or node, while the bottom one is the child node. Each node t represents a subspace of the input space: $\mathcal{X}_t \subseteq \mathcal{X}$. The further down the tree, the smaller the subset becomes. On the way down, a splitting rule is applied at each edge to determine the child node. For example, this could be whether a catastrophe is insured in the US or somewhere else.

Decision trees, which underlie all the following tree-based methods, are particularly attractive because of their good properties. First, as they are non-parametric, they can model arbitrarily complex relationships between inputs and outputs, without any a priori assumption. Second, trees intrinsically implement feature selection, making them relatively robust to irrelevant or noisy variables. This also makes them relatively robust to outliers and labelling errors. Third, they can (simultaneously) handle heterogeneous data types, e.g., numerical, and categorical variables. Finally, since they can be represented graphically, they are relatively easy to interpret.

3.3. Decision trees

As foreshadowed before, the predicted output value $\varphi(\mathbf{x})$ is the label of the leaf reached by the instance $\mathbf{x}$, when propagated through the tree by following the splits s_t. The global error of the model, already defined in Equation 3.1, can be further resolved as in Equation 3.3. By minimizing the local error in the terminal nodes, the global error is now also minimized. Looking at the formula, the minimization is done over the set of terminal nodes $\tilde{\varphi}$. Additionally, the probabilities of the input variables are taken into account.

$$Err(\varphi) = \mathbb{E}_{X,Y}\left\{L\left(Y, \varphi(X)\right)\right\} = \sum_{t \in \tilde{\varphi}} P(X \in \mathcal{X}_t)\mathbb{E}_{X,Y|t}\left\{L\left(Y, \tilde{y}_t\right)\right\} \tag{3.3}$$

Figure 3.1 shows an exemplary decision tree for the classification of cat bonds into different categories according to the size of their (predicted) premium. When classifying an instance $\mathbf{x}$, in this example a particular cat bond, this instance is propagated from the top of the tree down to one of the leaves. This leaf is then the prediction. If the prediction differs from the true observation, the error is large. But if the splits are well chosen, the error (between true and predicted results) is hopefully small, and the prediction represents true relationships. These relationships are at least correlations and at best causal relationships.

To build a tree, there must be a measure that evaluates the goodness of a possible split at a node. Breiman et al. (1984) define such an impurity measure $i(t)$ that evaluates the goodness of a node t. When $i(t)$ is small, the node is regarded pure, leading to better predictions $\hat{y}_t(x)$ for all $x \in \mathcal{L}_t$, where $\mathcal{L}_t$ denotes the subset of learning samples falling into t, all $\mathbf{x} \in \mathcal{X}_t$. Equation 3.4 defines the impurity decrease of a binary split s – the most typical split in a decision tree – dividing a node t into a left node t_L and a right node t_R. In this formula, p_L is the proportion $\frac{N_{t_L}}{N_t}$ of learning samples from $\mathcal{L}_t$ going to the left node t_L, with N_t denoting the size of the subset $\mathcal{L}_t$. The proportion for the right node is determined symmetrically. Applying this measure, the entire learning set $\mathcal{L}$ can be iteratively divided to reach increasingly purer nodes.

$$\Delta i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R) \tag{3.4}$$

So far, the impurity measure is still very abstract. In fact, there are multiple impurity and

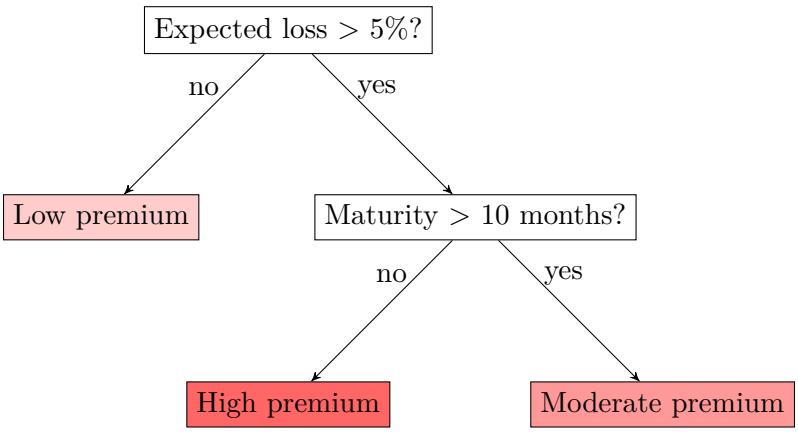


Figure 3.1.: This is an exemplary decision tree for determining the size of the premium. According to this example, a cat bond with an expected loss of less than 5% would be classified as a low premium bond. However, if a bond had an expected loss of more than 5% and a maturity of fewer than 10 months, the premium would be predicted to be high.

purity measures. The target of purity and impurity measures is to minimize the uncertainty of the outcome. For instance, a toss of a fair coin leads to an unpredictable outcome because the probability is the same for both sides. In this case, there is a lot of impurity. The degree of uncertainty is often described by the concept of entropy. Claude Shannon describes entropy as the average amount of information required to encode a randomly drawn value of a set X (Shannon, 1948). In the discrete case, entropy is defined as the expectation of all negative logarithmized probabilities of the possible outcomes: $H(X) = \mathbb{E}[-log\{P(X)\}] = -\sum_{i=1}^{N} P(X = i) log\{P(X = i)\}$. For example, a fair coin toss carries a high entropy, as the outcome of the toss is random. Hence, the outcome is uncertain and surprising. Knowing that the probability is 50% does not help to better predict the outcome of the coin toss. In contrast, when flipping an unfair coin, the entropy is much lower, as the outcome is not random anymore. Therefore, the best decision tree split leads to the largest improvement of purity, which is equivalent to a reduction in entropy. The information gain is a measure of how much information a feature provides about the outcome. When the information gain is maximized, the entropy is decreased the most. For instance, the CART decision tree in Breiman et al. (1984) uses Gini impurity to maximize the information gain from splitting the tree: $I_G = 1 - \sum_{l \in \{1,\dots,N\}} p_l^2$. The highest impurity is 1. The squared probability of all individual outcomes is subtracted from this value of 1. Thus, the impurity is close to 1 if there are many individual outcomes with a low probability, making the actual outcome unpredictable. The information gain in a decision tree is the difference between the entropy of the parent node and the average entropy of the child nodes. This means that a good decision tree will result in pure terminal nodes with low entropy.

A simple way of splitting the tree is to use a greedy algorithm that divides each node using a split that locally minimizes the impurity. Yet, a greedy strategy may be suboptimal. In fact, lookahead search could give better results, as the goodness of the split can also be assessed by evaluating deeper splits. However, this would come at the expense of higher computational power. Similarly, a deeper decision tree is not necessarily a better one. If the tree is shallow

(with few leaves), there is probably a high bias due to underfitting. This makes a very deep tree seem optimal at first. However, as with most machine learning algorithms, too much model complexity is likely to lead to overfitting. Theoretically, the tree could have as many leaves as data points, minimizing the error to 0. However, this tree would not be very useful for predicting new data, as the splitting rules would no longer be generally applicable. This means that there may be a point where improving training estimates – by reducing their error – does not further improve the test estimates. In fact, excessive complexity may worsen the test estimates. More specifically, the model should still be generalizable to some degree to deal with unseen data. To avoid overfitting, there are stopping criteria. The underlying idea is to find a good compromise between a tree that is neither too deep nor too shallow. For instance, a stopping criterion could be a maximal depth of a tree, a minimum size of a node, or a minimum required decrease in impurity for a split to be conducted.

3.4. Random forests

For reducing the generalization error, *ensemble methods* can be used. By introducing random perturbations into the learning process, several different models can be built based on a single learning set $\mathcal{L}$. In a second step, the combined predictions of the individual models form the prediction of the ensemble. A simple method of combining them is to take the average. Such an ensemble prediction has the advantage that the variance of the prediction is much lower compared to the prediction of a single tree if the individual trees differ, which in turn lowers the generalization error.

Random forests are a family of methods that make use of an ensemble of decision trees, that is also called *forest*. In the case of decision trees, this is very effective because they often have a high variance and low bias. Therefore, this high variance can be lowered by using a forest instead of a single tree. Random forest methods differ in the way they introduce random perturbations into the induction procedure – the construction of a decision tree.

To explain how they ensure that individual trees differ, I must first introduce two concepts: *bootstrapping* and *bagging*. Bootstrapping is a statistical sampling technique for estimating quantities such as descriptive statistics (e.g., mean, standard deviation) from a data sample. First, many random sub-samples are created from the initial sample with replacement and as many observations as the original sample. For instance, if the original sample consists of 100 cat bond observations, there could now be 1000 sub-samples, each containing 100 random observations from the original sample. Due to the replacement, the sub-samples usually contain many duplicates of the observations of the original dataset. Hence, the so called *out-of-bag* dataset for each of the bootstrap sub-samples consists of the observations of the original dataset that are not in the bootstrap sub-sample. Second, the needed descriptive statistic is calculated for each of the bootstrap sub-samples. For instance, if I am interested in the mean of a sample, the mean of all sub-samples would be computed. Finally, all individual statistics from the sub-samples are aggregated again. A simple way to do this is to take the mean of all sub-sample statistics. This bootstrapping procedure can greatly improve the estimation of a statistic.

Bagging – **b**ootstrap **agg**regat**ing** – is an ensemble algorithm used to reduce variance and avoid overfitting. In the case of decision trees, sub-samples with replacement are again generated. Based on each sub-sample, a decision tree is generated. Finally, the results of the individual decision trees are aggregated. Again, a simple way of doing so would be to take the average of the individual results. One additional source of variance stems from the similarity

of the individual decision trees although their samples already differ. This is directly related to feature importance. Even if the samples differ, the same or similar features are likely to decide the splits of the trees. Algorithms are usually greedy and try to minimize errors (without being forward looking). By limiting the known features for each tree and selecting different known features for each tree, more different sub-trees are generated due to different splitting criteria. The trees are increasingly split based on "weak learners", i.e., predictors that are not very strong, instead of "strong learners" that may not be known. This mechanism decorrelates the trees, reducing the variance of the bagged estimator. Averaging many rather uncorrelated quantities leads to a larger reduction in variance than averaging many strongly correlated quantities. This is demonstrated in Equation 3.5. The aggregated variance is decreased when the covariance term (almost) vanishes. This is only the case when the trees contain many weak learners, so that the trees tend to be uncorrelated. If the trees contain mostly strong learners, they are very similar, which increases the covariance term.

$$Var\{\hat{f}_{Bag}(x)\} = Var\{\frac{1}{B}\sum_{b=1}^{B}\hat{f}_b(x)\} = \frac{1}{B^2}\left[\sum_{b=1}^{B} Var\{\hat{f}_b(x)\}\sum_{c\neq d} Cov\{\hat{f}_c(x), \hat{f}_d(x)\}\right] \tag{3.5}$$

Random forests are particularly effective in settings with many features that are unrelated to the outcome, i.e., settings with sparsity. Since the splits generally ignore unrelated covariates, the performance remains strong in such settings.

3.5. Causal random forests

In the machine learning literature, the focus has been on out-of-sample performance as the criterion of interest (Athey & Imbens, 2019). This has been at the expense of the ability to perform inference, which has been a major focus in the statistics and econometrics literature. A typical application of inference is the construction of confidence intervals that are valid. Recently, the development of methods for inference has made substantial progress for low-dimensional functions in specific settings. Random forests have especially profited from a range of novel literature (e.g., Wager and Athey (2018), Athey and Imbens (2016), Athey et al. (2019)). Generalized random forests offer new methods for three statistical tasks: nonparametric quantile regression, conditional average partial effect estimation, and heterogeneous treatment effect estimation via instrumental variables. In generalizing random forests, many core elements of Breiman's forest are preserved (Breiman, 2001), including recursive partitioning, subsampling, and random split selection, but the final estimate is not obtained by averaging estimates from each member of an ensemble of trees. Instead, the forests are treated as a type of adaptive nearest neighbor estimator. Doing so opens many beneficial statistical extensions.

Motivation Before explaining general random forests, I need to explain why "normal" random forests are sometimes insufficient. When using a multiple linear regression, interaction terms between variables of interest can be leveraged to gain information about heterogeneous treatment effects and the interconnectedness of different variables. For instance, if I am interested in the treatment indicator w and how it is connected to another variable x_1, I could use the regression $Y = \beta_0 + \beta_1 w + \beta_2 x_1 + \beta_3 (w * x_1)$ to find out. In this example, the treatment

effect would be $\beta_1 + \beta_3 \times x_1$. Depending on the value of x_1, the treatment effect could be heterogeneous between different observations. However, when there are many variables, the number of possible interactions quickly increases to a number at which the linear model suffers from low statistical power. Furthermore, the linear model only allows for linear relationships, unless additional polynomials are added, which again reduces statistical power. Hence, a regression is not an optimal option for finding complex relationships. This is the reason why random forests are increasingly used for estimating heterogeneous effects, computing quantile regressions, etc. However, a "traditional" random forest is optimized for minimizing the mean squared error of the outcome variable. In the case of causal random forests that enable to draw causal conclusions, this is not the best way to optimize. Instead, Wager and Athey (2018) add two additional features to a "traditional" random forest to adapt it to causal purposes: a different **splitting criterion** and **honesty**.

Random forests choose splits for the variable and value at each tree node such that the greatest reduction in the mean squared error with respect to the outcomes Y is achieved. In contrast, causal random forests adjust the splitting criterion by searching for a partitioning where the treatment effects differ the most. In addition, it accounts and corrects for how the splits affect the variance of the parameter estimates.

An honest tree has a high accuracy, namely a bias that asymptotically disappears, low standard errors, and low confidence intervals. This leads to consistent estimates and valid confidence intervals. To make a tree honest, the training data is split into two subsamples. One is a splitting subsample used to perform the splits. The other one is an estimating subsample used to make the predictions. More precisely, first a tree is grown using the splitting subsample. Then, all observations from the estimating subsample are dropped down the previously grown tree until they fall into terminal nodes. Ultimately, the prediction of the treatment effects is determined by the delta of the average outcomes between the treated and the untreated observations of the estimating subsample in the terminal nodes.

Methodology To explain general random forests, some notation is needed. The notation is very general, making it applicable for many different statistical methods. Random forests is only one of them. This also leads to an abstract notation. But I will later explain the derivation of the estimator for the special case of regression trees. I assume to have data $(X_i, O_i) \in \mathcal{X} \times \mathcal{O}$, for $i = 1, \ldots, n$. Here, X_i denotes the covariates used to predict the quantity of interest $\theta(x)$, and O_i denotes the observable quantities encoding relevant information for predicting $\theta(x)$. The quantity of interest $\theta(x)$ is defined by a so-called "local estimating equation" presented in Equation 3.6. In this equation, $\psi(\cdot)$ is a scoring function and $\nu(x)$ is an optional nuisance parameter. The equation can be used in most cases to determine the maximum likelihood parameters $(\theta(x), \nu(x))$. Additionally, conditional means, quantiles, and average partial effects can be identified, adding valuable information to predictions derived from random forests.

$$\mathbb{E}\left[\psi_{\theta(x),\nu(x)}(O_i) \mid X_i = x\right] = 0 \quad \text{for all } x \in \mathcal{X} \tag{3.6}$$

Breiman (2001) make a prediction for a point x by pushing x down each of a certain number of trees until it ends up in a terminal node. The prediction from each tree b is then $\hat{\mu}_b$. The random forest's final prediction for x is the mean prediction of all trees: $\frac{1}{B}\sum_{b=1}^{B}\hat{\mu}_b$. In contrast, for generalized random forests, one observes which training examples fall into the same terminal node as x for each tree (Athey et al., 2019). $L_b(x)$ denotes the set of training examples falling into the same terminal node as x for each tree T_b. The following two equations

describe how the similarity of a training example X_i and the prediction x is determined.

$$\alpha_{bi}(x) = \frac{1\{X_i \in L_b(x)\}}{|L_b(x)|}, \qquad \alpha_i(x) = \frac{1}{B}\sum_{b=1}^{B}\alpha_{bi}(x).$$

For each training example $i = 1, \ldots, n$ and each tree b, the function $\alpha_{bi}(x)$ is computed. This function analysis whether the training point X_i falls into the same terminal node as x. If X_i and x are in the same terminal node, $\alpha_{bi}(x)$ becomes $\frac{1}{|L_b(x)|}$. If not, it becomes 0. Since one is interested in a global measure for the similarity of a training sample X_i and the prediction x, $\alpha_i(x)$ is a weighting function that aggregates over all trees B. The average $\alpha_{bi}(x)$ over all trees is computed. This measure $\alpha_i(x)$ increases towards 1 with the similarity of X_i and x and decreases towards 0 when they (almost) never end up in the same terminal node. If they are similar, the training sample X_i should receive a bigger weight when predicting at x. Hence, $\alpha_i(x)$ can be regarded a weighting function. Figure 3.2 is an illustration of the weighting function. The final weighting of each training example i, $\alpha_i(x)$, represented by the size of the circle (per observation) in the bottom row, is the average of the weights of all the individual decision trees $\alpha_{bi}(x)$, as shown in the top row. The idea is that similar observations should often end up in the same terminal node as the quantity of interest x.

After determining the $\alpha_i(x)$, the estimator $\hat{\theta}$ is determined by solving the minimization problem Equation 3.7.

$$\left(\hat{\theta}(x), \hat{\nu}(x)\right) \in \text{argmin}_{\theta,\nu} \quad \left\| \sum_{i=1}^{n} \alpha_i(x)\psi_{\theta,\nu}(O_i) \right\|_2 \tag{3.7}$$

Generalized random forests for random forests "Normal" random forests (Breiman, 2001) are a special case of generalized random forests (Athey et al., 2019). In the following, I show that the estimator derived from the generalized random forest equals the one from Breiman (2001). Athey et al. (2019) just provide a more general solution of the special case in Breiman (2001). The notation simplifies in this special case because the observable quantity equals the response of interest: $O_i = Y_i \in \mathbb{R}$. I am interested in estimating the conditional mean function of x defined by the following.

$$\theta(x) = \mu(x) = \text{argmin}_{\mu} \quad \mathbb{E}\left[(Y_i - \mu)^2 \mid X_i = x\right]$$

After differentiating with respect to μ and setting this equal to 0 to arrive at the optimum of the minimization problem, the problem becomes the following.

$$\begin{aligned} 0 &= \mathbb{E}[Y_i - \mu(x) \mid X_i = x] \\ &= \mathbb{E}[\psi_{\mu(x)}(Y_i) \mid X_i = x] \end{aligned}$$

In this equation, $\psi_{\mu(x)}(Y_i) = Y_i - \mu(x)$ is the same form as Equation 3.6. This means that it is now possible to plug it into Equation 3.7. Doing this, taking the derivative and then solving for the solution of the minimization problem, while considering that the sum of the weights

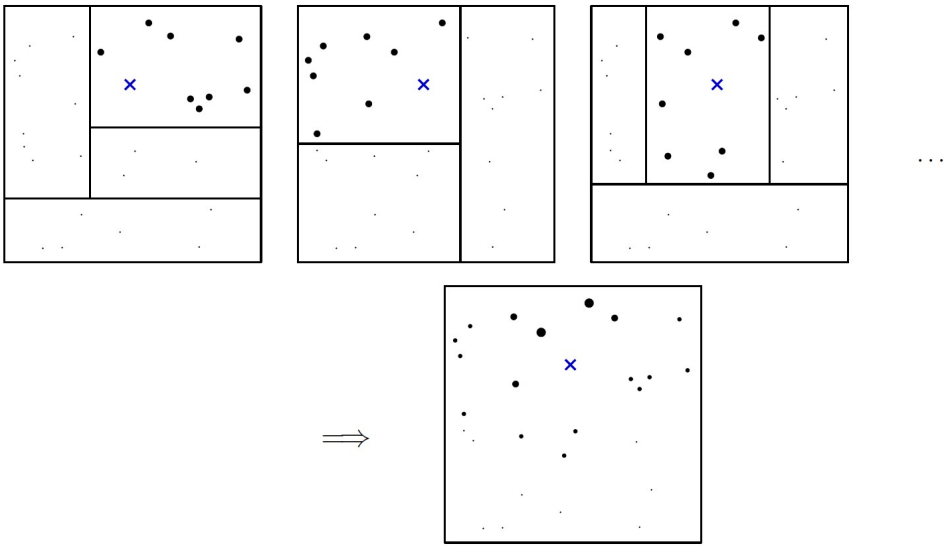


Figure 3.2.: This graphic illustrates the random forest weighting function. Each square in the top row corresponds to a decision tree and the small rectangles inside them correspond to terminal nodes of the tree. To find a good prediction for the test point of interest x, shown as a blue cross, the other observations (or circles here) are weighted. For each tree (or square here), the observations that fall into the same terminal node as x are weighted by 1 (and are large in the graphic), the others by 0 (and are small in the graphic). Then, the final weighting, as depicted in the bottom row, is the average of all tree-based weightings of the top row. Thus, if an observation often falls into the same terminal node as x, it is weighted more heavily than an observation that falls less frequently into the same node. The final weighting is emphasized by the size of the circles. This illustration is taken from Athey et al. (2019).

of the training examples is 1 ($\sum_{i=1}^{n} \alpha_i(x) = 1$), I arrive at $\hat{\mu}(x) = \hat{\mu}_b(x)$. Importantly, this is exactly the same estimator as derived from "normal" random forests (Breiman, 2001).

$$\begin{aligned} \hat{\mu}(x) &= \text{argmin}_{\mu} \quad \left\| \sum_{i=1}^{n} \alpha_i(x) \psi_{\mu}(Y_i) \right\|_2 \\ &= \text{argmin}_{\mu} \quad \left(\sum_{i=1}^{n} \alpha_i(x)(Y_i - \mu) \right)^2 \end{aligned}$$

$$\Rightarrow \quad \sum_{i=1}^{n} \alpha_i(x)(Y_i - \hat{\mu}(x)) = 0$$

$$\begin{aligned}
\hat{\mu}(x) &= \sum_{i=1}^{n} \alpha_i(x) Y_i \\
&= \sum_{i=1}^{n} \sum_{b=1}^{B} \frac{1}{B} \frac{1\{X_i \in L_b(x)\}}{|L_b(x)|} Y_i \\
&= \frac{1}{B} \sum_{b=1}^{B} \frac{Y_i 1\{X_i \in L_b(x)\}}{|L_b(x)|} \\
&= \frac{1}{B} \sum_{b=1}^{B} \hat{\mu}_b(x)
\end{aligned}$$

3.6. Double machine learning

The previous section on generalized random forests sounds very compelling. However, it relies on data from a randomized control trial, as the assignment of the treatment must be random. For most practical purposes, however, this is not the case. In fact, it is often impossible to perform experiments and historical observational data is usually also non-experimental. If the assignment of the treatment is not random, a common problem is the correlation of confounders with both the analyzed covariate (the treatment) and the outcome variable. This is the reason why double machine learning is additionally needed when making use of general random forests.

Typical approaches for correcting undesired correlations are the use of instrumental variables and partial regressions. These concepts can also be applied to machine learning. When supervised machine learning is used to learn the functions, it tends to lead to overfitting and regularization biases. Double machine learning, also called orthogonal machine learning, attempts to correct both biases (Chernozhukov et al., 2018). The "double" comes from the use of primary and auxiliary predictive models. By training two models, causal inference is possible in supervised learning.

I now briefly summarize the overall process of double machine learning. First, the data are divided into two equal sets through sample splitting. This should eliminate the problems of overfitting later. One model is trained to predict the treatment from confounders. The treatment $\tilde{T}$ residuals are calculated based on this model. The other model is trained to predict the target from the confounders. The outcome $\tilde{Y}$ residuals are computed based on this model. This separation by having two models fixes the problem of a potential regularization bias. As a final step, the previously predicted target $\tilde{Y}$ (in this context the premium spread) is regressed on its predicted treatment $\tilde{T}$. This allows to obtain the overall treatment effect, as depicted in Equation 3.8. In the case of random forests, this process should lead to desirable properties such as asymptotic normality of the machine learning estimators (Athey et al., 2019).

$$\tilde{Y} = \theta(X) \times \tilde{T} + \epsilon \tag{3.8}$$

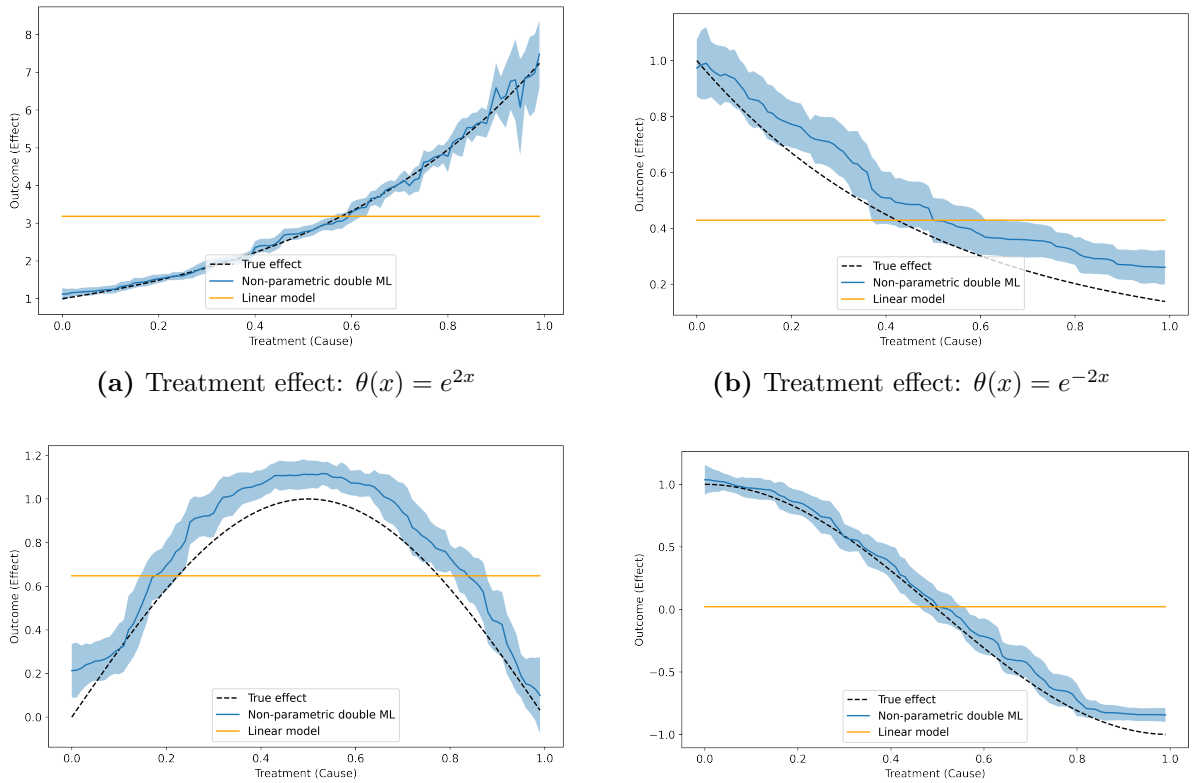
To illustrate how causal effects can be detected and disentangled, I analyze synthetic datasets. My synthetic dataset consists of 20 features with 20 thousand rows of data. The synthetic

data follow the partially linear regression model. In this model, the outcome is defined by $Y = T\theta + g(X) + \epsilon$. Here, θ is the causal parameter that I am interested in determining. However, the feature of interest T depends on X: $T = m(X) + \epsilon$. In these equations, ϵ are the irreducible error contributions. And g and m are the nuisance functions. The first equation is the main equation. By also having the second equation, one can correct for an omitted-variable bias.

In the synthetic dataset:

- Y is the target
- T is the feature of interest, also called treatment
- X are the control vectors $\sim \mathcal{U}(0, 1)$, that one needs to control for
- W is a matrix of the confounders $\sim \mathcal{N}(0, 1)$

In Figure 3.3, I test four different treatment effects. Double machine learning can reveal them all.



(a) Treatment effect: $\theta(x) = e^{2x}$

(b) Treatment effect: $\theta(x) = e^{-2x}$

(c) Treatment effect: $\theta(x) = sin(\pi x)$

(d) Treatment effect: $\theta(x) = cos(\pi x)$

Figure 3.3.: These simulations with synthetic data demonstrate that heterogeneous treatment effects can be identified by using double machine learning. A linear model leads to an incorrectly estimated treatment effect. For example, in the case of (d), the estimated effect is around 0. However, in reality, the treatment effect ranges from 1 to -1, depending on the value of the control variable X.

3.7. Interpretability tests: Shapley values

The Shapley value is an idea from the field of cooperative game theory (Shapley, 1953). Players cooperate in a coalition and receive payoffs depending on their contribution to the total payoff. This idea is transferable to machine learning predictions. Now, the "game" becomes the prediction task for a single instance of the dataset, the "payoff" becomes the difference between the actual prediction for a specific instance and the average prediction for all instances, and the "players" become the cooperating feature values of the instance (Molnar, 2022).

The interpretation of Shapley values sounds a bit abstract at first. Given a current set of feature values, the estimated Shapley value is the contribution of a feature value to the difference between the actual prediction and the mean prediction. To obtain a Shapley value for a specific feature, the marginal effect of this feature for all possible coalitions of features must be observed. The Shapley value is then the average marginal effect. For instance, I may want to predict cat bonds spreads with three characteristics: expected loss, trigger type, and sponsor. In this example, I already know that the spread is 5% for an expected loss of 10%, an indemnity trigger, and Swiss Re. I also know that the spread is 6% for an expected loss of 10%, a different trigger, and Swiss Re. Both bonds have the same features except for the trigger type (here a dummy). Now I know that the additional 1% of spread for the second bond must be due to the different trigger type. But this 1% is not the average effect of a bond with no indemnity trigger. Therefore, I would have to repeat this computation for all possible coalitions to finally arrive at the correct estimation for a bond with no indemnity trigger. In this example, there are only three features, but there are already plenty of possible coalitions because the expected loss and sponsor can have many different values.

Importantly, although Shapley values are great for interpretation, they do not serve prediction purposes. They do, however, allow for contrastive explanations. For instance, different subsets of data can be compared, or a single data point can be compared to the whole dataset or a subset. In this way, the effect of certain features becomes more understandable. When analyzing cat bonds, this becomes a powerful tool. Figure 3.4 illustrates why the predicted spread of one particular bond is 13.28%. For the entire training set, the average spread is around 7.5%, which is described as the base value. The red features drive the predicted spread to a value above the average, while the blue features reduce the predicted spread. In addition, the feature arrows indicate how much each feature value affects the predicted spread. In the example, the expected loss of around 5.6%, which is well above the average expected loss of around 2.5%, is the strongest feature causing most of the upward shift in the predicted premium.

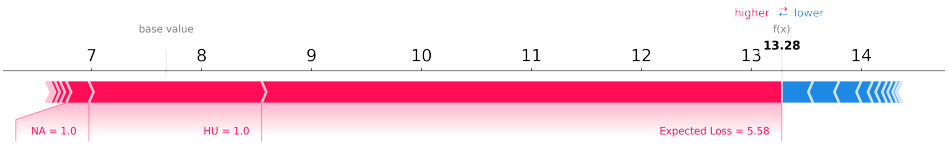


Figure 3.4.: Shapley value example for one cat bond compared to the whole dataset. This bond's predicted spread is higher than for the "average bond" due to its high expected loss.

4. A causal random forest approach for the secondary cat bond market

This chapter contains my analysis of the premium in the secondary cat bond market. I make use of the previously introduced methodology of chapter 3, that includes causal random forest approaches, Shapley values, and double machine learning. In contrast to recent papers such as Makariou et al. (2021), my approach focuses on causality rather than spread prediction. My aim is to shed light into the machine learning algorithms to make their predictions explainable. For pure predictions, the models can remain "black boxes" since the main goal is accurate prediction. In contrast, I am not trying to achieve high accuracy, but to quantify the effects and their uncertainty.

4.1. Data

I use a data collection on 736 cat bonds issued between March 2002 and March 2021 for which premium at issue, secondary market prices and all explanatory variables are available. Observations with missing or implausible data are excluded from this dataset. The first half of the data is from Gürtler, Hibbeln, and Winkelvos (2016) and contains 332 bonds after cleaning the data. This dataset contains cat bonds issued between March 2002 and March 2012. For the remaining time period between April 2012 and March 2021, I collected data for additional 404 cat bonds after cleaning the dataset. The premiums – the yield spreads over the LIBOR – form the dependent variable in my analysis. I extracted these premiums from the annual reports of Lane Financial LLC[1].

The explanatory variables are bond-specific or describe the macroeconomic state at the time when the bond was issued or when the secondary market prices were recorded. The set of variables included in this empirical study is based on recent papers on cat bond pricing (Götze and Gürtler (2020), Braun (2016), Gürtler et al. (2016)). I use the Artemis Deal Directory[2] and Lane Financial LLC for the collection of bond specific data. Data on the trigger mechanism, bond issuance volume, insured peril types, and peril locations are obtained from the Artemis Deal Directory. The issuance volume is a potential proxy for a bond's liquidity. The trigger type is critical because it exposes the investor to an additional source of risk. For instance, indemnity-trigger cat bonds potentially evoke ex-ante or ex-post moral hazard. Settling claims is a costly activity for the insurer. As its benefit is partly borne by the investors when using an indemnity trigger, this might create a conflict of interest between the bond's sponsoring insurer and its investors. Namely, the sponsor's loss adjustment policy may become laxer, which may affect either the probability of loss before or after an insured event (Götze & Gürtler, 2020). Regarding the trigger type, I only distinguish between indemnity and non-indemnity (like

[1] The annual reports are available on the website of Lane Financial LLC: http://www.lanefinancialllc.com/content/blogcategory/41/67/ (retrieved on 10/10/2022).

[2] The deal directory can be found on the following website: https://www.artemis.bm/deal-directory/ (retrieved on 10/10/2022).

Götze and Gürtler (2020)) to reduce the complexity and increase the sample size per group. A high complexity through multi-peril and multi-location bonds may also be reflected by a higher spread as compensation. Similarly, locations that are considered peak territories such as the US as well as peak perils such as hurricane cat bonds could be assumed to require an additional compensation.

The other bond-specific data is obtained from Lane Financial LLC, including the bonds' sponsors, the bonds' S&P ratings, the maturity and exposure term, the expected loss (EL), the probability of first loss (PFL), the probability of exhaust (PLL), and the conditional expected loss (CEL). Since the EL is defined as the first moment of the principal loss distribution of cat bonds, a higher EL should convert into a higher spread. The EL is widely assumed to be the most important price-determining factor. The PFL directly reflects event risk, while the CEL captures downside risk. Previous research has shown that some sponsors such as Swiss Re may achieve better conditions due to their market leading position (Braun, 2016). The information about the sponsor could be easily included through a dummy variable for each sponsor. Since there are many different sponsors, I decide to include only the type of sponsor: 'reinsurer', 'insurer', and 'other'. This results in a larger sample size per category.

The macroeconomic variables are included to capture market developments. All necessary data for these variables are extracted from Refinitiv and Guy Carpenter. First, the development of equity markets could influence cat bond prices, or the underlying price-driving factors could be similar. Hence, I include the monthly return of the S&P500 as an indicator for the development of equity markets. For doing so, I obtain the monthly S&P500 closing prices from Refinitiv[3]. Second, cat bond prices could co-move with prices of more traditional reinsurance products because cat bonds are a potential substitute (Braun (2016), Gürtler et al. (2016)). Therefore, I include a proxy for the reinsurance price cycle, namely the annual relative change in the Guy Carpenter Global Property Catastrophe Rate-on-Line Reinsurance Price Index[4] (Carpenter, 2012). As the general development of prices in the cat bond market should also be accounted for, I include the "Swiss Re Global Cat Bond - Total Return Index". This catastrophe bond market index calculated by Swiss Re Capital Markets should reflect the returns of the catastrophe bond market. I calculate the relative change in that index on a monthly basis. For the date of interest, this monthly change is considered. The monthly data used is collected from Refinitiv Datastream. Additionally, I include a monthly corporate credit spread. This spread captures the difference between the yields on corporate bonds and government bonds. The data is obtained from FRED[5]. Previous papers such as Götze and Gürtler (2020) observed the spread for each rating class. As most of the cat bonds in my dataset do not have an S&P rating, this does not seem to be a practical solution for my dataset.

Table 4.1 presents the summary statistics for all cardinal cat bond specific and macroeconomic variables. Comparing my statistics with those in Götze and Gürtler (2020), only the corporate spreads differ significantly. This is due to the different definition. In my analysis, I do not examine the spread for each risk category. All other statistics are similar. As I have more, different, and more recent observations, the statistics should naturally differ slightly. After analyzing the dataset including the previously mentioned PFL, PLL and CEL, I decided to remove them. Including them lets the dataset shrink, as there are some missing values.

[3] I retrieved the data on 12/10/2022.

[4] I obtain the data on the Guy Carpenter Global Property Catastrophe Rate-On-Line Index from https://www.artemis.bm/global-property-cat-rate-on-line-index/ (retrieved on 11/10/2022).

[5] ICE BofA US Corporate Index Option-Adjusted Spread from https://fred.stlouisfed.org (retrieved on 11/10/2022).

But more importantly, they do not seem to provide much additional information. When including them in a linear regression on the spread in addition to the other features, none of these risk measures was significant at a 10% level. Additionally, these variables contain similar information as the EL. A high correlation between the EL and those variables may worsen the interpretation of the impact of the EL on the spread later.

Table 4.1.: Summary statistics: cardinal cat bond specific and macroeconomic variables (reported at the issue level)

	Obs.	Mean	SD	Min.	Max.
Cat bond specific variables					
Premium (at issue, in %)	736	7.66	5.12	0.65	49.20
Expected loss (annual, in %)	736	2.56	2.45	0.00	15.75
Volume (in USD million)	736	139.63	118.28	1.75	1500.00
Maturity (in month)	736	37.31	11.60	5.00	69.00
No. perils	736	1.72	0.76	1.00	4.00
No. locations	736	1.25	0.64	1.00	5.00
Macroeconomic variables					
Reins. index (annual change at issue, in %)	736	1.68	12.69	-11.20	36.59
S&P500 (monthly change at issue, in %)	736	0.48	4.06	-12.51	12.68
Corp. spread (monthly at issue, in %)	736	1.52	0.69	0.83	5.92
Cat bond index (monthly change at issue, in %)	736	0.50	0.59	-3.74	2.64

Table 4.2 presents the summary statistics for all nominal and ordinal cat bond specific variables. Except for the rating statistics, my statistics resemble the ones from Götze and Gürtler (2020). Most of the bonds in my dataset are unrated, as most recent bonds do not have a rating at issuance anymore. However, bonds issued a few years ago often had a rating at issuance. Since my dataset contains more recent observations, this leads to a higher proportion of unrated bonds. The sample size for some ratings is very low, especially the AA, A, and CC rating. Hence, coefficients for these variables estimated by any model should be interpreted with great caution. Importantly, a single bond can have multiple peril types and locations. Consequently, neither the peril types nor the peril locations sum up to the amount of cat bonds in the dataset. However, each bond is assigned to only one rating category, one trigger type, and one sponsor type.

Since the secondary market premiums are reported quarterly and the bonds have an average maturity of around 37 months, the number of observations increases to 7,624 after removing outliers. Before dealing with the outliers, I observed the premium value for the 1st percentile and the 99th percentile. Since these values roughly match the minimum and maximum value of the premia at issuance, I use these values as cutoffs. Precisely, this removes the zero premia for defaulted bonds and very high premia greater than 50%. These values are not representative for the whole dataset. Extreme values could be misreported values, or they could stem from wildly fluctuating bond prices during (potential) defaults. Importantly, premia during or before defaults should be removed because the most important feature, the expected loss, changes during or before a default. As I only know the expected loss at issue level, I cannot use these premia. Premia during defaults may be crucial for making correct predictions, but they are not necessary for making causal conclusions about the interaction of variables. I could have applied winsorizing if I had believed these outliers to belong to the distribution. However, as the percentiles resemble the range of issuance premia and the premia during defaults are

Table 4.2.: Summary statistics: nominal and ordinal cat bond specific variables (reported at the issue level)

	Obs.	Percentage
Trigger		
Indemnity	309	41.98
Non-indemnity	427	58.02
Peril type		
EQ	467	63.45
HU	219	29.76
Wind	329	44.70
Other	228	30.98
Peril location		
EU	152	20.65
JP	95	12.91
NA	600	81.52
Latin America	34	4.62
Asia/Australia	32	4.35
Sponsor		
Reinsurer	331	44.97
Insurer	364	49.46
Other	41	5.57
Rating		
S&P Rating AA	1	0.14
S&P Rating A	4	0.54
S&P Rating BBB	17	2.31
S&P Rating BB	210	28.53
S&P Rating B	106	14.40
S&P Rating CC	1	0.14
No rating	397	53.94

difficult to interpret, I think trimming is better than winsorizing in this case.

Instead of observing the macroeconomic variables at issuance level, I calculate them for the respective reporting period of the secondary market prices. Thus, the change in the reinsurance index, the change in the S&P500, the corporate spread, and the cat bond index are updated depending on the reporting period. In contrast, the cat bond specific variables mostly remain the same across all reporting periods: the expected loss, the volume, the number of peril types and locations remain constant. Of course, the secondary market premium for each bond changes over time. Similarly, the maturity is updated by subtracting the time difference between the issue date and reporting date from the initial maturity. The maturity becomes the remaining maturity. In less than 1% of the data sample, the remaining maturity becomes marginally negative due to (slightly) inaccurate dates. In these cases, I replaced the negative number with 0. The summary statistics, as reported in Table 4.3, resemble those at the issue level in Table 4.1. However, the premium and expected loss are slightly lower, which could be due to trimming. Also, the mean of the reinsurance index is negative instead of positive such as in the analysis at the issue level. I assume that there are more reported premia in periods with a negative change in my dataset. This does not mean that the equally weighted annual change was negative in the observed time period. All other variables, reported in Table 4.4, remain constant over time. The percentages in this table differ slightly from those in Table 4.2

due to the different maturities of the bonds. For reference, the statistics without trimming are reported in Table A.1 and Table A.2.

Table 4.3.: Summary statistics: cardinal cat bond specific and macroeconomic variables (reported at secondary market level)

	Obs.	Mean	SD	Min.	Max.
Cat bond specific variables					
Secondary market premium (annual, in %)	7624	6.49	5.44	1.00	49.00
Expected loss (annual, in %)	7624	2.32	2.17	0.00	15.75
Volume (in USD million)	7624	143.53	121.40	1.75	1500.00
Maturity left (in month)	7624	22.04	13.48	0.00	67.00
No. perils	7624	1.75	0.77	1.00	4.00
No. locations	7624	1.26	0.63	1.00	5.00
Macroeconomic variables					
Reins. index (annual change, in %)	7624	-0.96	10.25	-11.20	36.59
S&P500 (monthly change, in %)	7624	0.34	4.03	-12.51	8.76
Corp. spread (monthly, in %)	7624	1.64	0.91	0.83	6.39
Cat bond index (monthly change, in %)	7624	0.55	0.63	-1.13	2.41

Table 4.4.: Summary statistics: nominal and ordinal cat bond specific variables (reported at secondary market level)

	Obs.	Percentage
Trigger		
Indemnity	3100	39.71
Non-indemnity	4524	57.96
Peril type		
EQ	4857	62.22
HU	2579	33.04
Wind	3316	42.48
Other	2372	30.39
Peril location		
EU	1705	21.84
JP	1105	14.16
NA	6061	77.65
Latin America	322	4.13
Asia/Australia	366	4.69
Sponsor		
Reinsurer	3355	42.98
Insurer	3867	49.54
Other	402	5.15
Rating		
S&P Rating AA	11	0.14
S&P Rating A	28	0.36
S&P Rating BBB	218	2.79
S&P Rating BB	2577	33.01
S&P Rating B	1176	15.07
S&P Rating CC	13	0.17
No rating	3601	46.13

4.2. Base model

Previous works have compared the performance of random forest models to a base model. However, their objective was to assess and compare the predictive power of the models. In contrast, my goal is to make the predictions more understandable. A simple linear baseline model can still be helpful for explaining interpretive differences between linear models and random forest models. Additionally, they provide an initial indication of the most important features and how well the features explain the outcome. Although linear models may also be considered machine learning models, they are rather simple. In this section, I first regress on all features before using recursive feature elimination to only regress on the features that seem most important. In the literature, variable selection methods (i.e., forward, backward, and stepwise selection) are often used to arrive at a base model. Other models such as penalization methods (i.e., Lasso, Ridge, and elastic net regression methods) are also popular for deriving base models.

I then check whether any other potential features out of all remaining variables are strongly correlated. If two features are strongly correlated, it is not clear from the model's perspective which of them should be used as a predictor. Since the model is usually not able to (correctly) prefer certain variables, the importance of the two variables could be wrongly estimated. For example, if an impurity has already been removed by one of them, the other variable's impor-

tance could be (wrongly) reduced although that variable could have also removed the impurity. In such cases, the reported importance may be biased. In the case of random forests, this interpretability problem is somewhat reduced by the random selection of features at each node creation. Nevertheless, this remains a potential problem to be aware of. Therefore, it is reasonable to use feature selection to remove features that are redundant. Figure 4.1 shows that there are no very strong correlations among the remaining features.

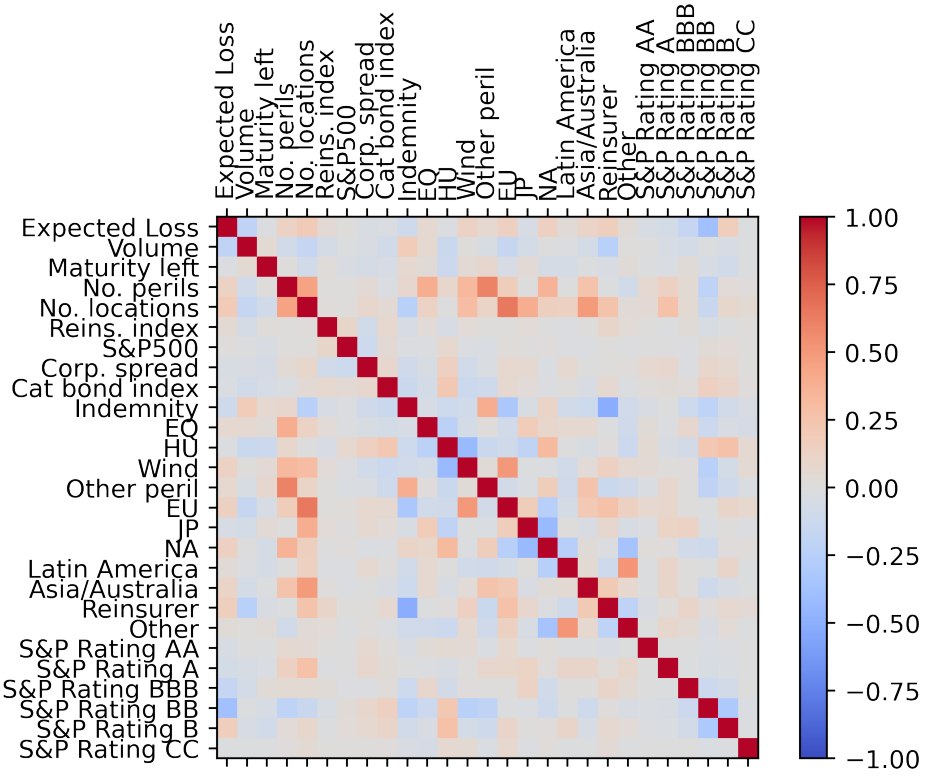


Figure 4.1.: Correlation matrix of all variables after removing strongly correlated features.

To arrive at an initial linear model, I first regress the outcome on all features in the dataset (after having removed the ones likely to cause multicollinearity issues). As shown in Table 4.5, the R^2 is not moderately high (0.525), indicating that some important characteristics may be missing in the dataset. It is unclear whether the data quality, the missingness of important features or the randomness in the cat bond spread causes the R^2 to not be higher. Although the adjusted R^2 is only slightly lower (0.524), not all of the features seem to be in fact decisive. Some p-values are very high, indicating that features such as the S&P500 (p-value of 0.84) may not explain much of the variance in the premium. The coefficients of the model are easy to interpret because it is a multivariate linear regression. For instance, increasing the expected loss by one additional percentage point (ceteris paribus) lifts the premium by around 1.56 percentage points according to the model. A bond that is exposed to hurricane perils, has on average a premium that is 0.65 percentage points higher than a bond without this risk.

To find the features that explain most of the variance in the premium between cat bonds, feature selection can be helpful. There are various methods for feature selection, such as uni-

variate selection (e.g., an ANOVA), recursive feature elimination (RFE), principal component analysis (PCA), and feature importance (e.g., using bagged decision trees). Here, I use RFE because it is easily implementable, and this work does not focus on feature selection. However, detecting the most important features (for a linear model) may be valuable later. The idea behind RFE is to recursively remove features and then build a new (linear) model on the remaining features. By analyzing the model accuracy, RFE identifies which features and combinations of features contribute most to predicting the outcome, the spread. Here, I remove features one by one until most of the remaining features are at least significant at the 10% level. I do not remove features one by one until all of them are significant because the features that remain unsignificant among the last features are ratings with a high coefficient. Due to their high standard errors, they are not significant, but they do influence the estimated premium significantly. Out of the 27 features of the "original" OLS analysis, only 9 remain in this slimmer model. Nevertheless, the R^2 only decreases slightly. Table 4.6 illustrates that next to the EL, the corporate spread, the hurricane peril, the Japan peril location, the North America peril, the reinsurer dummy and three ratings are also especially substantial for explaining the model outcome, the spread. Compared to the previous linear model, the coefficients also change. Thus, the interpretation is also slightly different. For example, the effect of a hurricane peril is now 0.97 percentage points instead of the previous 0.65 percentage points. At the end of this chapter, I will analyze the coefficients in more detail in comparison to the coefficients from the causal random forest. For comparison, I also briefly analyzed the primary cat bond market to show that the R^2 of the same explanatory variables is much higher in this market (Table A.3, Table A.4).

Table 4.5.: OLS regression results

Dep. Variable:	Secondary market premium (annual, in %)	**R-squared:**	0.525
Model:	OLS	**Adj. R-squared:**	0.524
Method:	Least Squares	**F-statistic:**	311.6
Date:	Sat, 03 Dec 2022	**Prob (F-statistic):**	0.00
Time:	22:47:03	**Log-Likelihood:**	-20886.
No. Observations:	7624	**AIC:**	4.183e+04
Df Residuals:	7596	**BIC:**	4.202e+04
Df Model:	27		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	-0.7941	0.253	-3.144	0.002	-1.289	-0.299
Expected Loss	1.5594	0.024	64.856	0.000	1.512	1.607
Volume	-0.0029	0.000	-7.606	0.000	-0.004	-0.002
Maturity left	0.0037	0.003	1.127	0.260	-0.003	0.010
No. perils	0.3390	0.177	1.917	0.055	-0.008	0.686
No. locations	0.1667	0.296	0.562	0.574	-0.414	0.748
Reins. index	0.0690	0.004	16.052	0.000	0.061	0.077
S&P500	0.0022	0.011	0.206	0.837	-0.019	0.024
Corp. spread	1.1104	0.050	22.324	0.000	1.013	1.208
Cat bond index	0.0509	0.072	0.709	0.478	-0.090	0.192
Indemnity	-0.0732	0.123	-0.596	0.552	-0.314	0.168
EQ	-0.3127	0.179	-1.744	0.081	-0.664	0.039
HU	0.6472	0.191	3.394	0.001	0.273	1.021
Wind	0.0051	0.197	0.026	0.979	-0.380	0.391
Other peril	-0.0934	0.217	-0.431	0.667	-0.518	0.331
EU	0.0149	0.320	0.047	0.963	-0.613	0.643
JP	1.1554	0.311	3.711	0.000	0.545	1.766
NA	1.7315	0.306	5.655	0.000	1.131	2.332
Latin America	-0.3202	0.380	-0.842	0.400	-1.066	0.425
Asia/Australia	-0.8989	0.359	-2.505	0.012	-1.603	-0.195
Reinsurer	-0.1783	0.113	-1.572	0.116	-0.401	0.044
Other	-0.4847	0.259	-1.873	0.061	-0.992	0.023
S&P Rating AA	-3.1763	1.152	-2.757	0.006	-5.434	-0.918
S&P Rating A	-2.1469	0.792	-2.710	0.007	-3.700	-0.594
S&P Rating BBB	-0.7255	0.286	-2.540	0.011	-1.285	-0.166
S&P Rating BB	0.0519	0.125	0.414	0.679	-0.194	0.298
S&P Rating B	0.5309	0.144	3.697	0.000	0.249	0.812
S&P Rating CC	-0.8197	1.055	-0.777	0.437	-2.889	1.249

Table 4.6.: OLS regression results after recursive feature elimination

Dep. Variable:	Secondary market premium (annual, in %)	**R-squared:**	0.501
Model:	OLS	**Adj. R-squared:**	0.500
Method:	Least Squares	**F-statistic:**	848.0
Date:	Sat, 03 Dec 2022	**Prob (F-statistic):**	0.00
Time:	23:44:07	**Log-Likelihood:**	-21081.
No. Observations:	7624	**AIC:**	4.218e+04
Df Residuals:	7614	**BIC:**	4.225e+04
Df Model:	9		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	-1.2480	0.148	-8.435	0.000	-1.538	-0.958
Expected Loss	1.6134	0.021	75.715	0.000	1.572	1.655
Corp. spread	1.0716	0.049	21.660	0.000	0.975	1.169
HU	0.9654	0.100	9.612	0.000	0.769	1.162
JP	1.5752	0.141	11.154	0.000	1.298	1.852
NA	2.0504	0.127	16.207	0.000	1.802	2.298
Reinsurer	0.1716	0.091	1.884	0.060	-0.007	0.350
S&P Rating AA	-2.5072	1.165	-2.153	0.031	-4.790	-0.224
S&P Rating A	-1.7606	0.741	-2.377	0.017	-3.213	-0.308
S&P Rating BBB	-0.5159	0.273	-1.890	0.059	-1.051	0.019

4.3. Random forests

Since causal random forest models build on top of random forests, I analyze the results of random forests in this section before arriving at causal random forests. In addition, I explain how well these models can be utilized to draw causal conclusions in the cat bond market. To some extent, I already use causal methods such as Shapley values.

Unlike regressions, random forests are effective in settings with many features and even features which are unrelated to the outcome. Thus, I do not restrict the set of features prior to my analysis. Any of the 27 features can be used for a splitting criterion. If a feature is not helpful or worse than another feature, it is simply not used for a splitting criterion. Moreover, decision trees can reveal more complex structures than a linear model, allowing previously insignificant variables to become very important. My random forest consists of 2,000 trees, with a maximum of 10 features per tree, and a maximum depth of 3. The maximum number of features implies that for each decision tree the "best" features (according to the splitting criterion) out of 10 random features from the complete set of features are chosen for the splits. The criterion that measures the quality of the split in my random forest is the Gini impurity. Not allowing all features for all trees, there is more diversity, as "weaker" features are also used for splits. This decorrelates the tree, which in turn leads to a lower variance in the estimator. The maximum depth of 3 means that for each of the 2,000 trees, there are a maximum of 3 splits from the top node to one of the end nodes. The sample, sample size, as well as the features differ between the trees. I use bootstrapping, meaning that only a sample of the entire dataset is used to build each tree. As an example, Figure 4.2 shows one decision tree made of 15 nodes. This tree therefore has the maximum number of splits, which is not the case for all 2,000 trees. The input space in this example is

$$X = EL \times HU \times No.locations \times \cdots \times Maturity = [0,1] \times \{0,1\} \times \{1,2,\ldots,5\} \times \cdots \times \{5,6,\ldots,69\}$$

for the regression problem of explaining the cat bond spread. Out of the 10 random features, only the most meaningful four are used in this specific tree. If the maximum depth were greater, potentially other features would have been used too. It is also possible that the same features are used multiple times. The node on the top is the root node and corresponds to the whole input space. The input space is then divided into two disjoint subsets depending on the value of the hurricane peril. The set of all input vectors with $HU \leq 0.5$ (which is in fact only $HU = 0$, as this is a dummy variable) ends up in the left node on the next bottom layer. This is also described by the "True" arrow. Similarly, the set on the right side of this second layer with the "False" arrow contains all other input vectors which do not satisfy $HU \leq 0.5$, which is the case for $HU > 0.5$ (which is in fact only $HU = 1$). According to one split criterion per node, the input vectors end up in one of the 8 terminal nodes on the bottom. Each split further reduces the MSE, since the goal is to divide the input space into more and more homogeneous subspaces. However, splitting too often bears the risk of overfitting and leaving too few samples per terminal node.

As plotted in Figure 4.3, the predictive accuracy of a random forest depends on the number of trees as well as the maximum number of features selected. For both plots, I split the entire sample into a train set containing 80% of the observations and a test set containing the remaining 20%. The random forest regressor for accuracy as a function of the number of trees is trained with 200 trees, a maximum number of features of 10, and a maximum depth of 5. The effect of adding trees leads to a large gain in prediction accuracy in the test set for

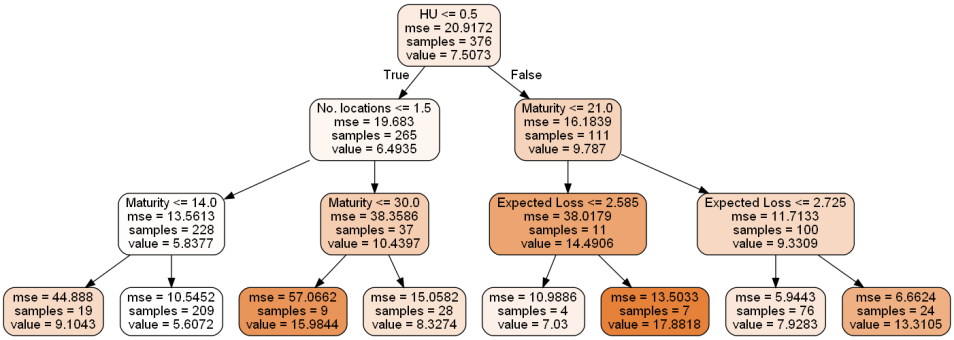
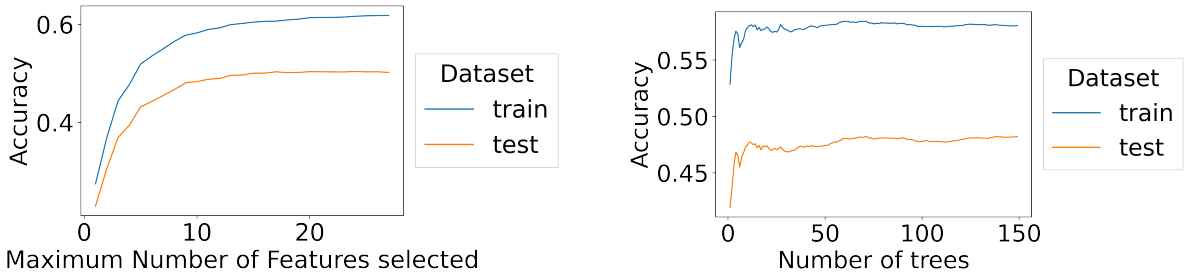


Figure 4.2.: One of many decision trees that form a random forest.

the first few trees. However, with many trees, the gain decreases. In comparison, the random forest regressor for the accuracy in dependence of the number of features is trained on 500 trees. The random forest is then fit on a maximum number of features that increases from 1 to 27, the number of available features. Naturally, more features per tree improve the accuracy in the train set, but this is not necessarily the case for the test set. In fact, the test accuracy continuously appears to slightly decrease from a maximum number of features of around 20. For comparison, Figure A.1 shows that the accuracy is much higher for the primary market premiums.



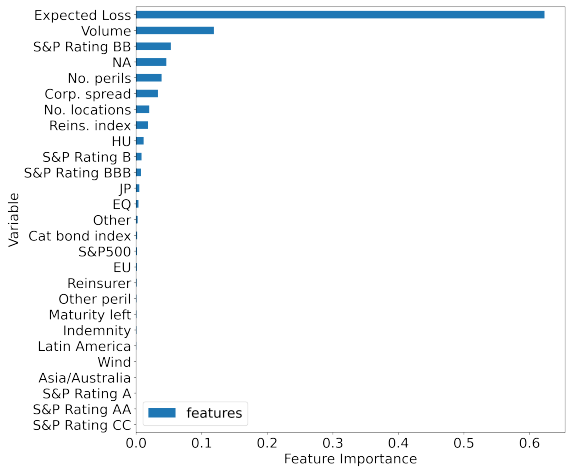
(a) Random forest accuracy in dependence of number of features.

(b) Random forest accuracy in dependence of number of trees.

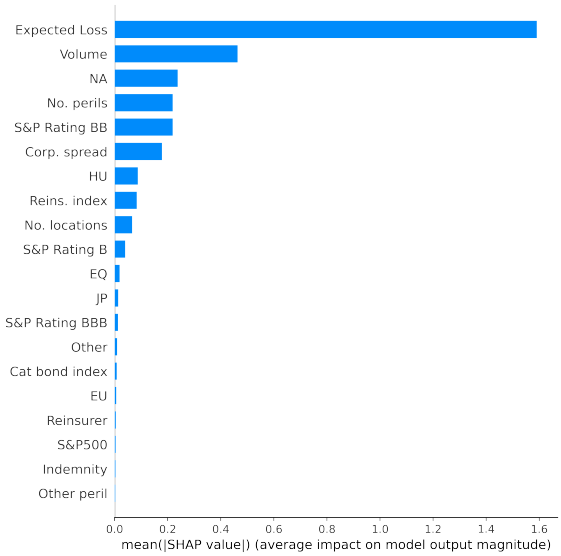
Figure 4.3.: Random forest accuracy in dependence of number of features and trees.

In determining which features are especially helpful at predicting the outcome, the feature importance is of great use. Figure 4.4 shows that the expected loss is by far the most important feature and the volume size of a bond follows in importance. Other important features include some rating dummies, the North America peril, the number of perils, the corporate spread, the number of locations, and the reinsurance index. The feature importance, shown on the left, is calculated as the mean and standard deviation of accumulation of impurity decrease within each tree. Hence, if splits based on a feature are especially good at lowering the impurity, their importance is high. In this case, it seems like making a split based on the EL helps the most. The SHAP feature importance, shown on the right, is measured as the mean absolute Shapley value. For each feature, the mean of the absolute Shapley values is calculated across all observations. Absolute values are calculated, as I do not want positive and negative values to offset each other. However, the results from the SHAP feature importance do not differ much

from the impurity feature importance, as the same features are determined to be relevant to a similar extent. Importantly, the importance scores do not tell me much about the interplay of different features. Additionally, they do not help at assessing why a feature may be better or worse than another feature, as the scores are based purely on correlations. A correlation could be random, or an underlying true factor may be missing in my dataset but correlated with another feature in the dataset.



Feature importance based on mean decrease in impurity.



SHAP feature importance measured as the mean absolute Shapley values.

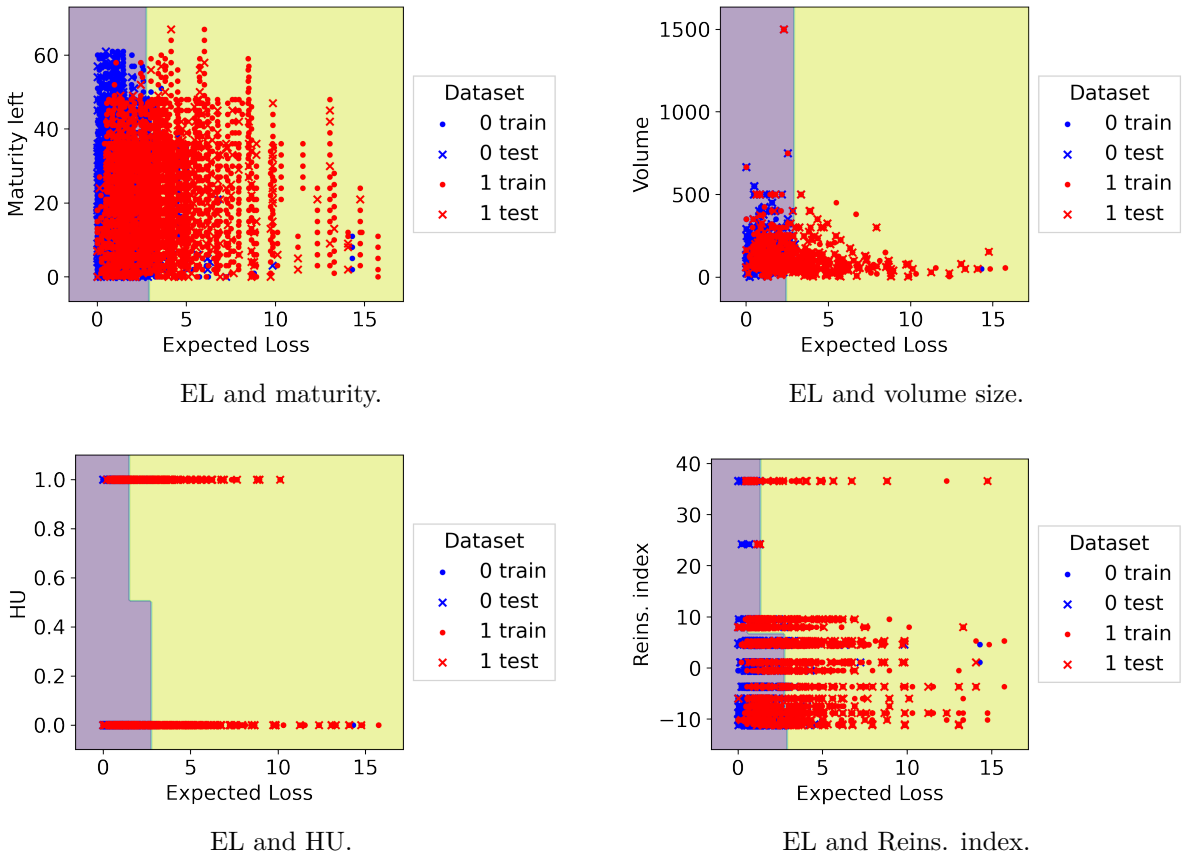
Figure 4.4.: Random forest feature importance: expected loss and volume seem especially decisive in determining the premium. Oppositely, features such as the EU peril location do not seem to influence the premium much.

Now that I know that expected loss is by far the most important feature, it becomes interesting to analyze how its effect is related to other features. This is done in Figure 4.5. To simplify the analysis, I first classify the cat bonds as having a high or low premium, depending on whether their premium is lower or higher than the median premium of all bonds in the dataset. In this way, I can predict whether a bond is a high or low premium bond, which is now a classification task rather than a regression task. Hence, I can now use a random forest classifier instead of a regressor. This has the advantage that the result is much easier to interpret. I then divide the entire dataset into a test set (20%) and a training set (80%). The random forest classifier is fitted using only the expected loss as well as one additional feature. Using contour lines and filled contours, all observations are plotted in dependence of the expected loss and the other feature. The blue dots represent observations with a lower than median premium, while the red ones have higher than median premium. Importantly, the test and train observations are plotted. The cross markers represent the test observations, whereas the dot markers indicate the train observations. If the classifier works well, the train and test observations should not differ much. More interestingly, the classifier's predictions for all combinations of the expected loss values with the values of the other feature are represented by the yellow and purple color. The yellow area indicates that the classifier would predict any

observation falling within this area as high premium, whereas the violet area contains all feature combinations with a low premium prediction. As HU is a dummy feature, the classifier's prediction should only depend on whether $HU = 1$ (which is the same as $HU \geq 1$ in this case) or not. This is also clearly the case for the random forest classifier. However, some relationships are more complex. For instance, a change in the reinsurance index by more than around 8% apparently changes the classifier's prediction dramatically. As such an interaction dummy is not normally included in a linear regression model, this is a very interesting result. These plots already help at interpreting random forest's predictions and allow some interpretations. For instance, a high EL leads to a high premium, but the EL is not the only decisive feature. At the same time, the interaction of different features is very crucial. Here, I only present the interaction of two features per plot. For instance, a bond with a hurricane peril seems to feel riskier for investors, as it only takes a lower EL to move it to the high premium side. The line between the violet and yellow area shifts from around 3% EL for a bond without hurricane peril to around 2% EL for a bond with hurricane peril. According to the graphs, the reinsurance index, the volume and the hurricane peril appear to influence the effect of the expected loss, while maturity does not seem to change its effect much. In reality, the interactions of more than two features may also be very important. For comparison, I did the same analysis for the primary cat bond market. As shown in Figure A.2, the effects of the reinsurance index and the hurricane peril appear to be very similar, while the maturity and volume have a much greater impact on the effect of the EL in this market. For example, in this market, a large volume bond seems to feel safer to investors, as it takes a higher EL to move it to the high premium side.

Figure 4.6 illustrates the certainty of the random forest classifier. The yellow area indicates that the probability is very high. Namely, the classifier is very certain that the prediction falls into the analyzed category. For example, in the top left graph, bonds with a large volume and a low expected loss have a high probability of being low premium bonds. In contrast, the violet area indicates that the probability of an observation falling into the analyzed category is very low. For instance, in the same plot at the top left, bonds with a low volume and a high expected loss have a low probability of being low premium bonds. The graphs help at interpreting how certain the classifier is. This is very important because a probability of 50% may not be enough to make a decision. In fact, a probability of 50% would mean that the classifier must guess the predicted outcome in this case. One can infer several results from the plots. At least when classifying bonds into low and high premium bonds, a random forest is very confident in classifying bonds with a low expected loss, a high maturity, and a large volume as low premium bonds. The interplay of the variables is also important. For example, some bonds with a relatively low expected loss still have a high premium due to their low volume size and/or low maturity. The plots on the right implicate that the classifier believes bonds with a high expected loss and low volume size and/or maturity to be high premium bonds. Furthermore, the classifier assigns bonds with a low expected loss and a high maturity and/or a high volume only a very low probability of falling into the high premium bond category. These results seem to be transferable to the primary market (Figure A.3).

One of the advantages of Shapley values is the analysis of heterogeneous feature effects on the model output and the analysis of single model outputs. In the summary plot of Figure 4.7, the values are grouped by the features on the y-axis and ordered by importance, their mean Shapley values. Each point represents a Shapley value for a feature and instance. While the x-axis represents the Shapley value, the feature value is represented by the color of the instance. For instance, the expected loss has the greatest feature importance, most instances have a high



EL and maturity.

EL and volume size.

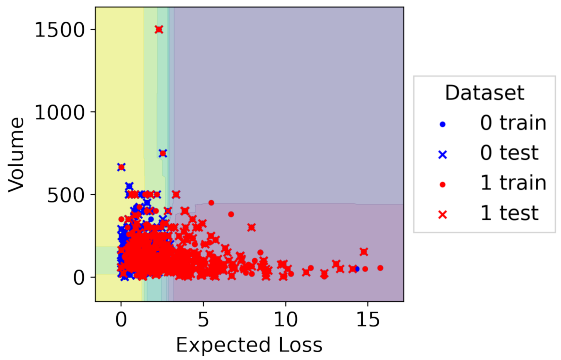
EL and HU.

EL and Reins. index.

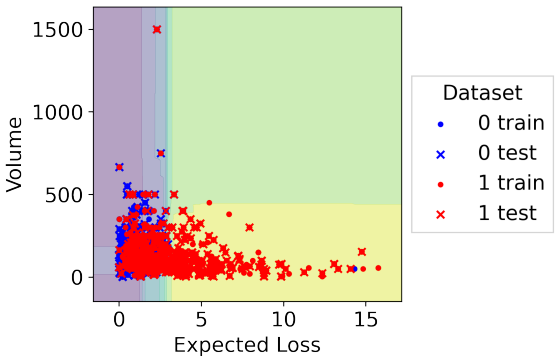
Figure 4.5.: The relationship of the most important feature, EL, with four other important features is analyzed. The blue markers ("0") represent observations with a lower than median premium, whereas the red ones ("1") have higher than median premiums. The background color depicts whether the random forest classifier predicts a high premium (yellow) or a low premium (violet).

EL Shapley value, and the EL Shapley value is especially high for large feature values. Other examples are that a high expected loss, low volumes, a North America peril, large corporate spreads, and a positive change in the corporate spread enlarge the estimated premium of a bond. The decision plot on the right side of Figure 4.7 shows how exactly the model arrives at specific output values for ten observations. With the help of this plot, a specific output becomes explainable. Nevertheless, all effects only describe the behavior of the model. Thus, these effects are not necessarily causal.

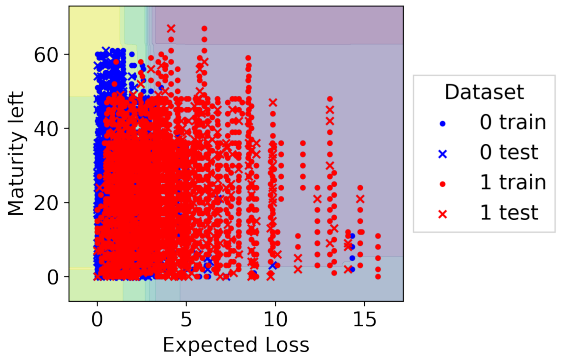
The summary plot provides a first hint at the relationship between the value of a feature and its impact on the prediction. To analyze the relationship in more detail, I compute the SHAP dependence plots for the most important feature, the expected loss, and the four features most likely to interact with it. The four features – volume, indemnity, wind, and the "other" sponsor – are derived by computing the approximate interactions first. As shown in Figure 4.8, large volumes appear to increase the effect of the expected loss for low expected losses. The plots of the other three features are more difficult to interpret. It seems like the "other" sponsor peril and the indemnity trigger lower the effect of the expected loss. The wind peril may reduce the



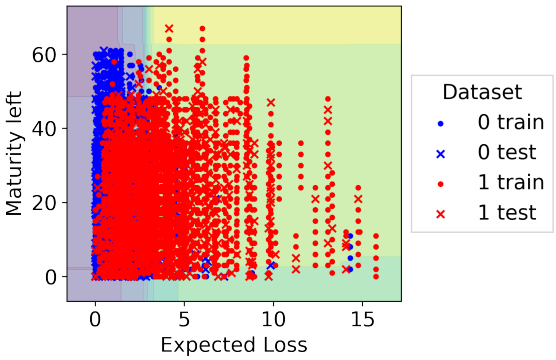
Probabilities for being a below median premium bond for different combinations of expected loss and volume size.



Probabilities for being an above median premium bond for different combinations of expected loss and volume size.

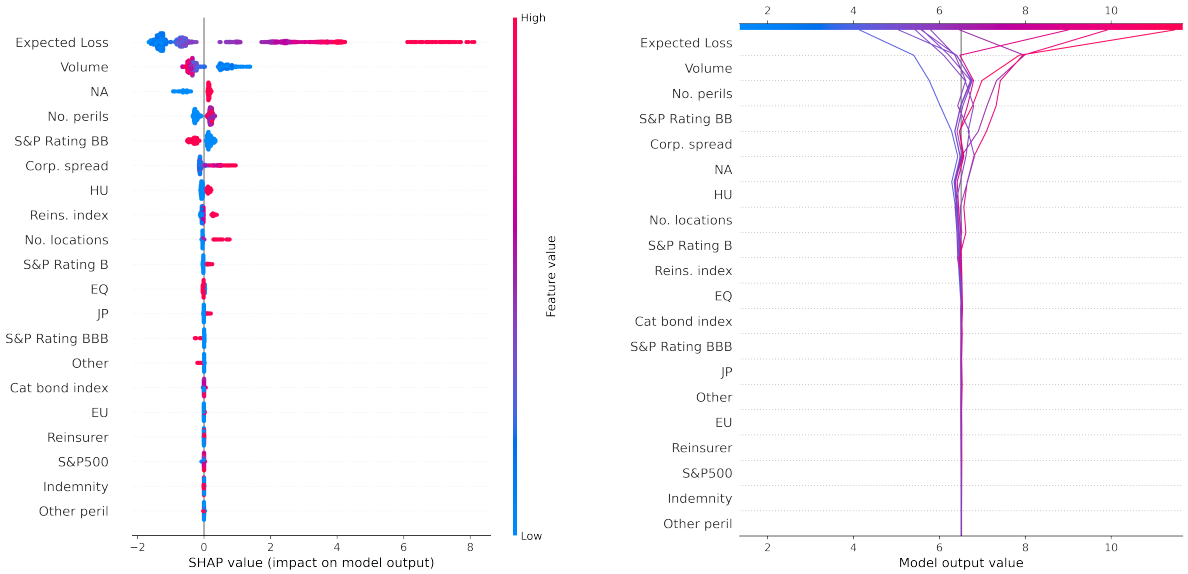


Probabilities for being a below median premium bond for different combinations of expected loss and maturity.



Probabilities for being an above median premium bond for different combinations of expected loss and maturity.

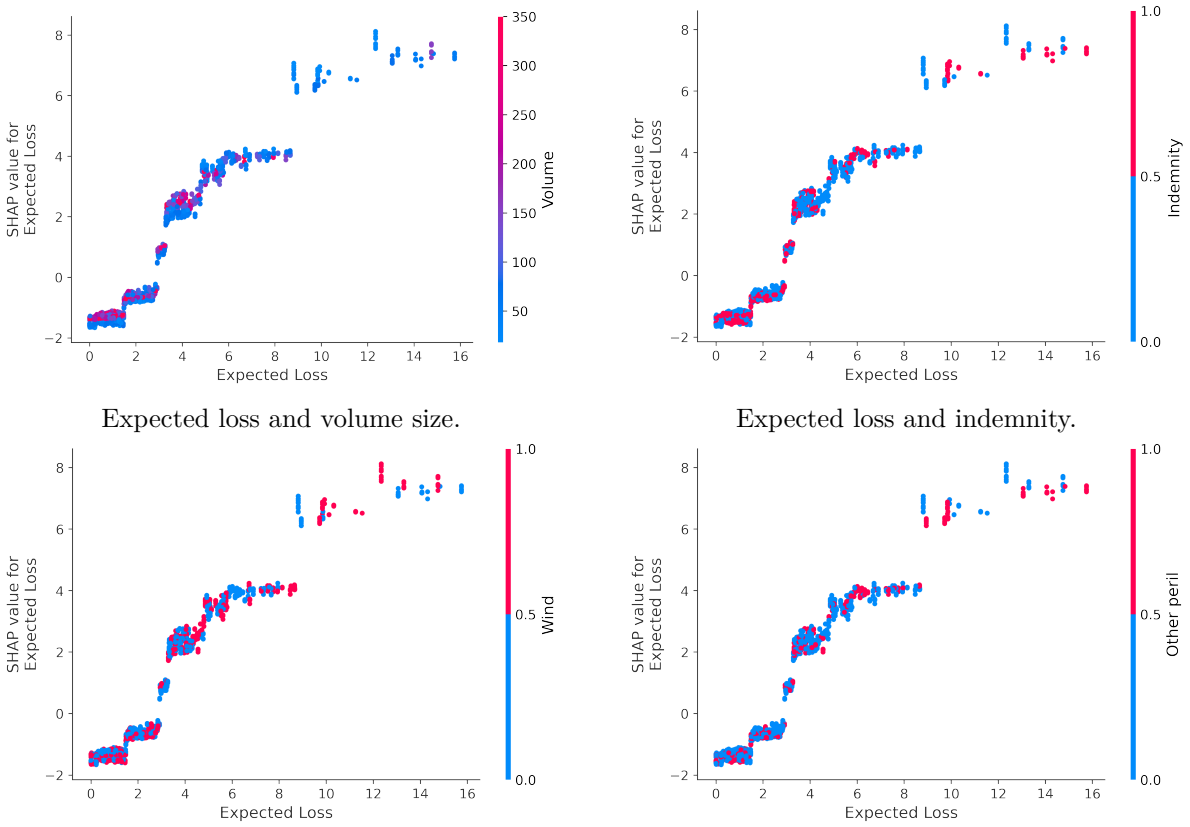
Figure 4.6.: The prediction probabilities are depicted by different background colors. Yellow stands for a high probability, and violet a low probability. The blue markers ("0") represent observations with a lower than median premium, whereas the red ones ("1") have higher than median premiums.



Summary plot. Decision plot.

Figure 4.7.: SHAP analysis: Shapley values for a random forest.

effect of the expected loss for low expected losses and increase it for high expected losses.



Expected loss and volume size.

Expected loss and indemnity.

Expected loss and wind peril.

Expected loss and other peril.

Figure 4.8.: SHAP analysis: Dependence plot for the four features most likely to interact with the expected loss.

4.4. Causal random forests

Random forests are well suited for predictive tasks. To make a good prediction, correlations are sufficient. In fact, these correlations could be purely random. For instance, a true underlying factor could be missing but correlated with another factor that is one of the features. By mistake, this included variable could now be considered the causal factor. In this example, an instrument is missing. This is the reason why A/B tests and randomized control trials are often referred to as the gold standard for causal inference. As randomized settings allow to exclude unwanted biases, unwanted effects can be avoided to optimally analyze causality.

Causal random forests are a great method for analyzing treatment effects (or other interesting effects). Because of their honest trees, they lead to asymptotically normally distributed estimators. Therefore, they allow to interpret coefficients and other statistical measures derived from a model. They can be used to determine not only the average treatment effect (ATE) but also the conditional average treatment effect (CATE). This is important because, for example, the effect of the EL on the cat bond spread might be heterogeneous. Indeed, it could be smaller or larger for bonds with a smaller or larger volume. This is very crucial because heterogeneity may lead to a misleading ATE that is likely to be misinterpreted. Additionally, a quantile

regression enables to observe variances around coefficients. I use the *EconML* python package for the causal analysis of one especially relevant factor, the "treatment".

However, causal random forests work best with data derived from an RCT. Unfortunately, in most practical situations, data from such optimal studies are not obtainable. In the case of cat bonds, this is simply not possible. Cat bonds would have to be issued at the same time and differ in only one factor to allow direct causal inference. Moreover, even a large sample size of such similar cat bonds would be needed to rule out randomness. Double machine learning may be a solution to this problem. Fortunately, *double machine learning* is implemented for many methods of the *EconML* package.

Importantly, interpretability relies strongly on the assumption that there is no unobserved confounding. Namely, all important features that significantly influence the outcome variable, the cat bond spread, should be included. If this assumption is not fulfilled and there are unobserved confounders, it is no longer possible to draw interpretive conclusions.

Figure 4.9 illustrates the interplay of all features and how they affect the cat bond spread Y. Both X and W are features used to predict both the outcome Y and the treatment T. However, only the features X are assumed to influence the strength of the relationship between Y and T. More specifically, the assumption is that the treatment effect θ is a function of X but not W. By having this setup, the feature of interest T (also called treatment) can be studied in detail when using causal forests.

In my cat bond dataset:

- Y is the outcome/target: here the cat bond spread
- T is the feature of interest: here the expected loss
- X are all features that may have a heterogeneous effect on T
- W are all remaining features

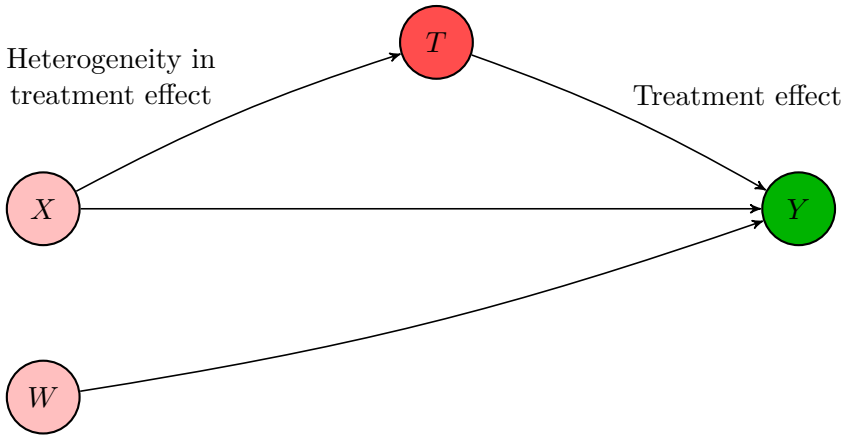


Figure 4.9.: Illustration how the "honest" effect of a feature T on the outcome Y, the cat bond spread, can be studied.

4.5. Causal random forests – Conditional average treatment effects

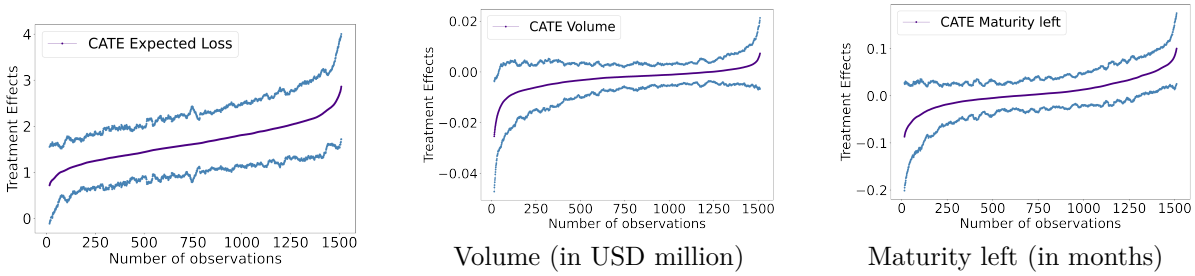
In this section, I analyze all features as the treatment factor, respectively. All other features are treated as X, features that may have a heterogeneous effect on T. By doing so, the clean,

"honest" effect on the outcome, the premium, may be observable. The causal forest is trained on 80% of the data. Since I want to study the effect of all features, one of the features is assigned to be T and all other features are X. Later, I reproduce this estimation for all possibilities of T. By using one random forest regressor each first for T and then also for Y, also called double machine learning, the "honest" effect of T should become analyzable. The splitting criterion is the mean squared error, 1,000 decision trees are used, and three-fold cross-validation. I then compute the causal effect as well as the confidence intervals for the training and test samples. For the confidence interval, I use $\alpha = 0.05$, meaning that the 95%, two-sided confidence interval is reported. Ideally, the effects should be similar for the test and train sample. This is indeed the fact, as I cannot find a big difference in the predictions between the train and test sample for any of the features.

Figure 4.10 shows the conditional treatment effect for all numerical features in the test sample. As I compute the heterogeneous treatment effect, the estimated effects differ between the sample. For interpretive purposes, I sort the treatment effects by their estimated magnitude and plot the rolling mean with a window of 30 observations. The individual treatment effects help assess whether the effect appears to be consistent or heterogeneous. In general, the confidence intervals become wider for especially low and high estimated effects. The most important feature, the expected loss, has an effect that appears to vary in its magnitude and exhibits a relatively large volatility, but the effect is always positive and ranges from around one to three percentage points. An additional percentage point in the expected loss leads to an increase in the premium of around 1.65 percentage points. The confidence interval is around ± 0.8 around this value. The volume effect is mostly negative. A larger volume size might lead to more trust from the investors, more liquidity, or the volume size is in fact correlated with a better known sponsor which in turn leads to more trust. To get a higher treatment effect, I could have also used 100 million as the unit. For instance, a treatment effect of -0.01 means that an increase in volume of 100 million would lead to a reduction in the premium of one percentage point. The results from the regression model with feature elimination does not show any effect of the volume on the premium. Here, the causal model gives a more ambivalent picture. In some cases, volume might matter a lot. But the effect is not clear. The "maturity left" feature is very heterogeneous, so that the mean and median values are close to zero. While the number of perils has a larger average coefficient of -3.38, the median coefficient is lower at -2.75. In comparison, the number of locations has a median effect of only 0.17. It is unclear whether the strong heterogeneity in this feature is random. Similarly, the median effect of the reinsurance index is only 0.04, but the mean of 0.26 is much larger, showing the strong heterogeneity. The effects of the S&P500 and the corporate spread appear to be mostly positive, while the cat bond index mostly has a negative effect. Based on this analysis, the expected loss, the number of perils, the cat bond index and the corporate spread seem to be the most decisive features, as their mean and magnitude coefficient values have a large absolute value. It is very crucial to consider the differences in units, as the predicted coefficients alone do not tell much. Moreover, most features exhibit a lot of heterogeneity. Unfortunately, this analysis does not provide an answer to the question of what this heterogeneity is due to.

Additionally, Figure 4.11 and Figure 4.12 describe the conditional treatment effect for all dummy features in the test sample. The effects of the earthquake peril type, Europe peril location, Japan peril location, the Latin America peril location, and the Asia/Australia peril location are very heterogeneous and can be both positive and negative, which complicates the interpretation of the coefficients. The effect of the hurricane peril type is mostly negative, while the effects of the wind peril type, the "other" peril type, the Japan peril location, and

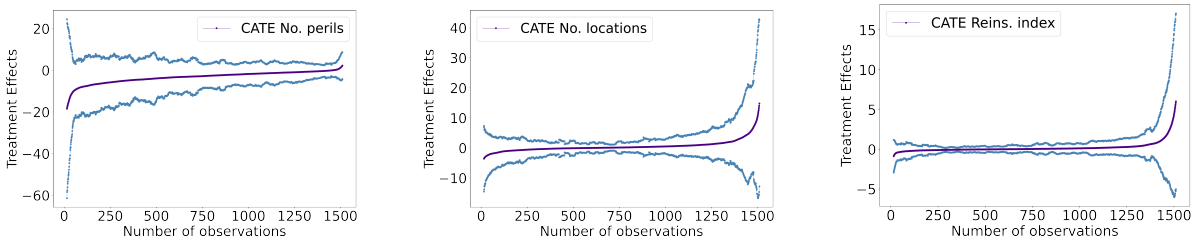
the North America peril location are mostly positive. The conditional average treatment effect of the indemnity trigger type is estimated to be both positive and negative, which complicates the interpretation, but signals that heterogeneity is crucial in the analysis and interpretation. If the indemnity trigger increases the premium, this would fit with investors preferring index triggers that are easier to understand and cause less disagreement between sponsor and investor. Interestingly, the effect of having a reinsurer as a sponsor instead of an insurer appears to have a negative effect. Here, a better bond rating leads to a higher premium, which is not intuitive. In fact, the AA and A rating have a positive mean effect. As the sample size is very low (at least when counting the number of bonds), great caution is needed when interpreting these figures. The coefficient of the BB rating is slightly positive, which seems realistic, and the sample size is sufficiently large. Moreover, mostly older bonds are rated. Therefore, the coefficients of the ratings could be biased because of time effects, which are not considered in my analysis.



Expected Loss (in %)
(mean: 1.65; median: 1.62)

Volume (in USD million)
(mean: -0.0029; median: -0.0019)

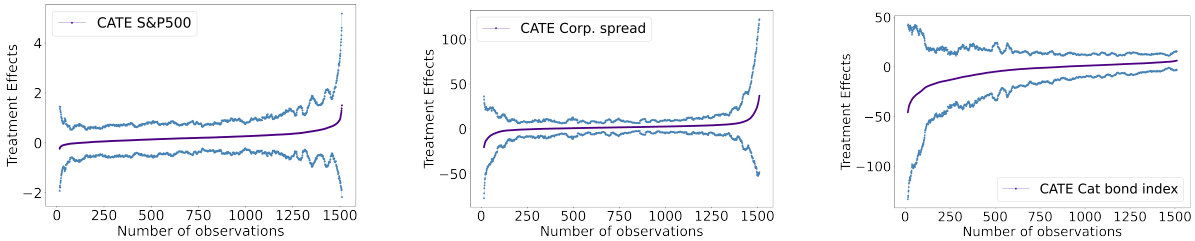
Maturity left (in months)
(mean: 0.0067; median: 0.0037)



No. perils
(mean: -3.38; median: -2.75)

No. locations
(mean: 0.70; median: 0.17)

Reins. index (in %)
(mean: 0.26; median: 0.04)



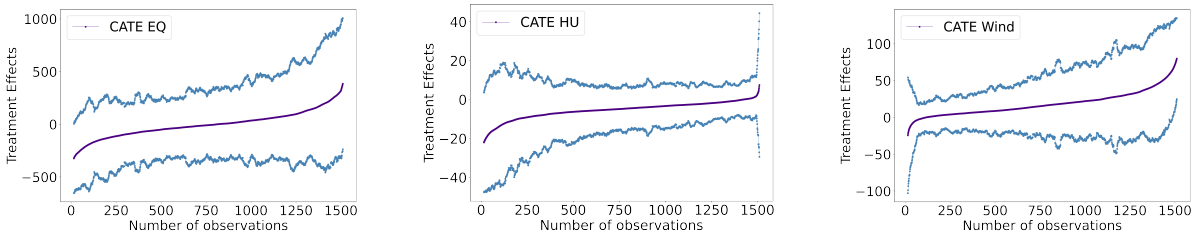
S&P500 (in %)
(mean: 0.25; median: 0.21)

Corp. spread (in %)
(mean: 2.34; median: 1.87)

Cat bond index (in %)
(mean: -4.49; median: -0.93)

Figure 4.10.: CATE for numerical features.

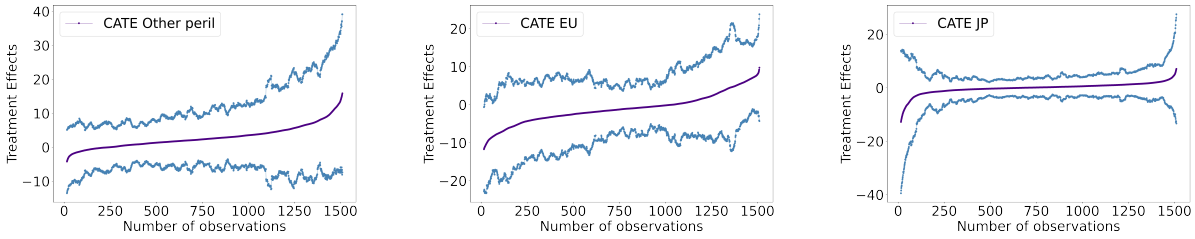
Table 4.7 provides a summary of the results of the linear estimators and conditional average treatment effects for the secondary market. When comparing the results from the linear models



Peril type: EQ
(mean: -0.28; median: -7.78)

Peril type: HU
(mean: -5.40; median: -4.80)

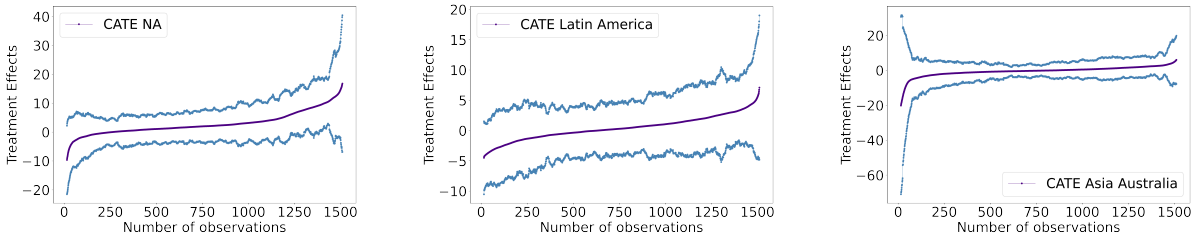
Peril type: Wind
(mean: 18.54; median: 15.56)



Peril type: Other
(mean: 3.06; median: 2.40)

Peril location: EU
(mean: -1.03; median: -1.17)

Peril location: JP
(mean: -0.10; median: 0.16)

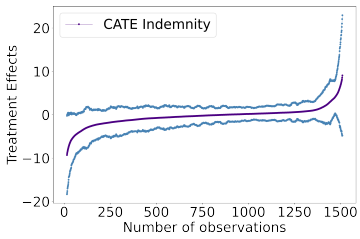


Peril location: NA
(mean: 2.92; median: 2.04)

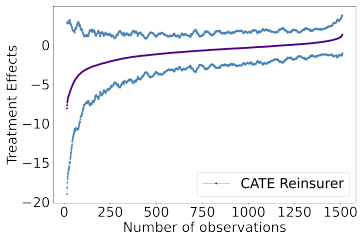
Peril location: Latin America
(mean: 0.47; median: 0.32)

Peril location: Asia/Australia
(mean: -0.59; median: -0.19)

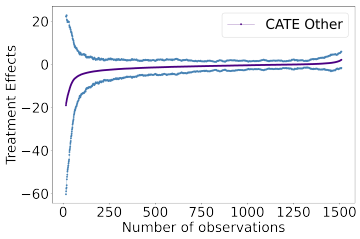
Figure 4.11.: CATE for dummy features: peril type, peril location.



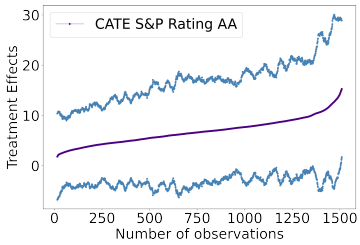
Trigger type: Indemnity
(mean: -0.36; median: -0.19)



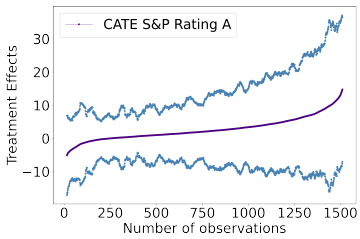
Sponsor: Reinsurer
(mean: -1.04; median: -0.68)



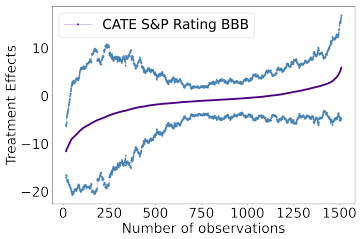
Sponsor: Other
(mean: -1.53; median: -0.79)



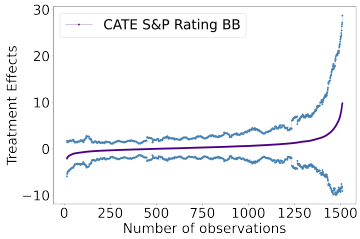
S&P Rating: AA
(mean: 6.88; median: 6.67)



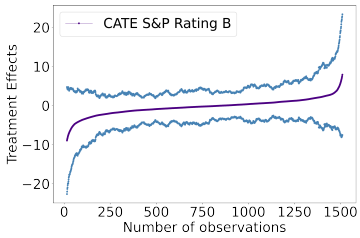
S&P Rating: A
(mean: 2.82; median: 2.06)



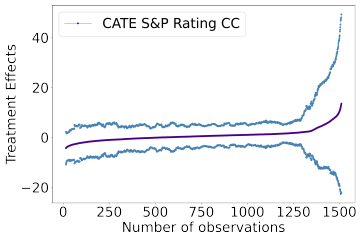
S&P Rating: BBB
(mean: -1.50; median: -1.01)



S&P Rating: BB
(mean: 0.68; median: 0.28)



S&P Rating: B
(mean: -0.32; median: -0.22)



S&P Rating: CC
(mean: 0.88; median: 0.53)

Figure 4.12.: CATE for dummy features: trigger type, sponsor type, S&P rating.

for the secondary market with the results for the primary market (Table A.3, Table A.4), most estimated coefficients are very similar. For instance, the expected loss in the secondary market is estimated to be 1.54 and 1.56 in the primary market. Out of the variables which are significant at the 1% level, the effect of the reinsurance index, the corporate spread, the Japan location peril, the North America location peril, and the rating B are also very similar. Interestingly, the volume is only significant in the secondary market, and the maturity only in the primary market. The effect of the hurricane peril differs immensely between both markets. The estimators in the simple regression could be biased because the model includes variables which are not significant. However, the regression with prior feature elimination leads to similar results. The expected loss has a coefficient of 1.61 in both markets. This means that one additional percentage point in the expected loss would lead to an increase in the premium of 1.61 percentage points. The other coefficients are also mostly similar. The estimated impact of the hurricane peril differs: 0.97 versus 2.78 percentage points. The reinsurer dummy feature also has a very divergent value of 0.93 in the primary market and 0.17 in the secondary market (only significant at the 10% level). Overall, this is a strong indicator that both markets are influenced by similar factors. It is also a strong indicator that my results are at least somewhat reliable.

For comparing the results for the causal random forest models, I added the median estimated coefficient, which provides information about the distribution of the estimated values and allows an assessment of whether the effect can be interpreted based on the visualizations of the CATE plots. The green numbers mark the ten variables which are most important according to the feature importance analysis. My analysis focuses on them. First, it is interesting that these variables are not always significant according to the linear models. A random forest can also reveal complex interactions, which is not possible in a linear model. For example, the number of peril locations is identified as important although this feature is not significant in the linear models. As the mean and median values differ greatly and the effect is highly volatile, the linear estimator is not well suited to discover such a heterogeneous effect. The expected loss has a clear positive effect. Moreover, its magnitude is fairly similar across the different models. The effect of the volume is negative, while the maturity effect's magnitude is vanishingly low. For example, a bond with a maturity of additional 10 months is estimated to have a premium which is only 0.07 percentage points higher, ceteris paribus. The number of perils is very important, as indicated by the high absolute coefficient of the mean and median. Some features such as the number of locations and the reinsurance index are difficult to interpret because the treatment effect is very volatile. The effects of the corporate spread and the North America location peril are estimated to be greater in comparison with the linear models. Interestingly, the hurricane peril type has a very negative effect, although it is rather low for the linear models.

This analysis provides an educated guess of the true mean effect, and makes the volatility in the features' effects tangible. Given the three different models for the secondary market, I am confident in quantifying some effects such as the expected loss, which is roughly the same across all models. In addition, the number of peril locations and types, and the corporate spread are much more important according to the causal forest models than for the linear estimators. The results for the hurricane peril are inconclusive and probably require additional analysis. The brief comparison with the primary market suggests that some features may have a very similar effect in different markets.

Table 4.7.: Comparison of results for the secondary cat bond market

	Secondary premium				
	Linear model		**Causal RF**		
	Simple	FE	Mean	Median	Clear effect
Expected loss	1.56***	1.61***	1.65	1.62	yes
Volume (100M)	-0.29***		-0.29	-0.19	yes
Maturity (10 mos.)	0.04		0.07	0.04	no
No. perils	0.34*		-3.38	-2.75	yes
No. locations	0.17		0.70	0.17	no
Reins. index	0.07***		0.26	0.04	no
S&P500	0.00		0.25	0.21	yes
Corp. spread	1.11***	1.07***	2.34	1.87	yes
Cat bond index	0.05		-4.49	-0.93	no
Indemnity	-0.07		-0.36	-0.19	no
EQ	-0.31*		-0.28	-7.78	no
HU	0.65***	0.97***	-5.40	-4.80	yes
Wind	0.01		18.54	15.56	yes
Other peril	-0.09		3.06	2.40	yes
EU	0.01		-1.03	-1.17	no
JP	1.16***	1.58***	-0.10	0.16	no
NA	1.73***	2.05***	2.92	2.04	yes
Latin America	-0.32		0.47	0.32	no
Asia/Australia	-0.90**		-0.59	-0.19	no
Reinsurer	-0.18	0.17*	-1.04	-0.68	yes
Sponsor: Other	-0.48*		-1.53	-0.79	no
S&P Rating AA	-3.18***	-2.51**	6.88	6.67	yes
S&P Rating A	-2.15***	-1.76**	2.82	2.06	yes
S&P Rating BBB	-0.73**	-0.52*	-1.50	-1.01	no
S&P Rating BB	0.05		0.68	0.28	no
S&P Rating B	0.53***		-0.32	-0.22	no
S&P Rating CC	-0.82		0.88	0.53	yes

The significance levels in the linear models are indicated by the stars next to the coefficients. If a p-value is less than 0.1, it is flagged with one star (*). If a p-value is less than 0.05, it is flagged with two stars (**). If a p-value is less than 0.01, it is flagged with three stars (***).
The coefficients colored in green highlight the features that are most important according to the causal forest model.

4.6. Causal random forests – Heterogeneous effects

When analyzing how factors influence the premium, it is useful to determine whether their effect is heterogeneous due to a particular factor. In this context, l define causality as the influence a single continuous feature such as the expected loss has on the outcome, the cat bond spread, while holding all but one other characteristic constant. This one other feature X, that is not kept constant, is assumed to affect the strength of the treatment effect. The feature under investigation is denoted T since causal forests are especially popular for analyzing treatment effects. However, this feature does not necessarily need to be a treatment in the classical economic context. The final outcome is denoted Y. Both X and W are used to predict both outcome Y and treatment T. But only the features X are assumed to influence the strength of the relationship between Y and T. Precisely, the assumption is that the treatment effect θ is a function of X but not of W. I could, of course, include all remaining variables in X instead of W, which I do in section 4.5, but there are also cases where the effect is expected to be heterogeneous only with respect to only one or a few variables. Hence, the effect of T is assumed to depend on the values of X. This heterogeneity of the treatment effect, or as here, just the effect of the feature under study, is examined in this section.

I could analyze the treatment for different groups based on their heterogeneity. For instance, I could split my observations into two groups, one with high and the other with low changes in reinsurance index values. The (average) treatment effect of the expected loss could then be expected to differ between the groups. This would lead to additional insights into the heterogeneous effect of the expected loss on the premium. Figure 4.13 illustrates how this would work. If there is indeed a heterogeneity in the treatment effect, the conditional average treatment effect should not be the same for the two subgroups.

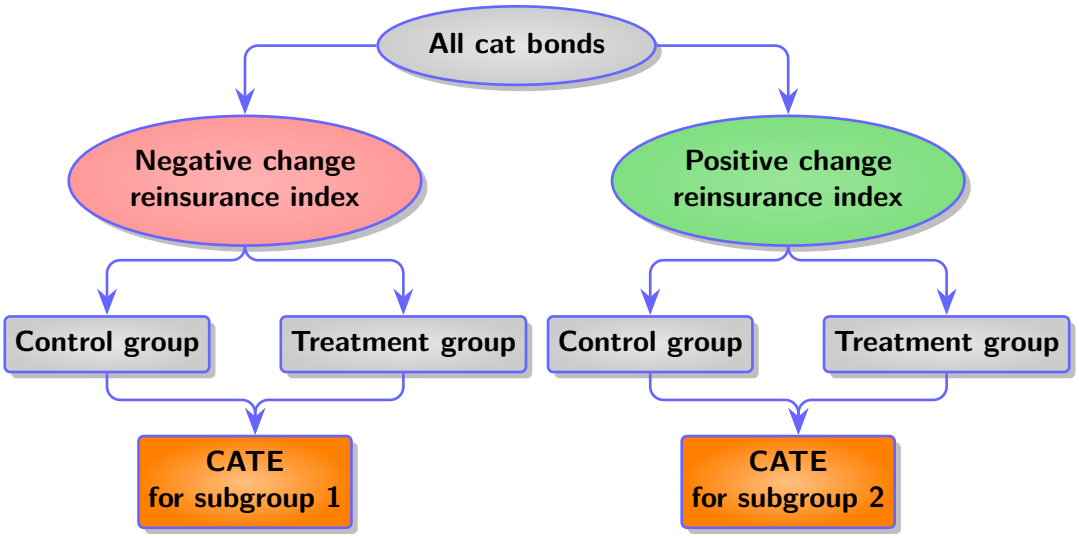


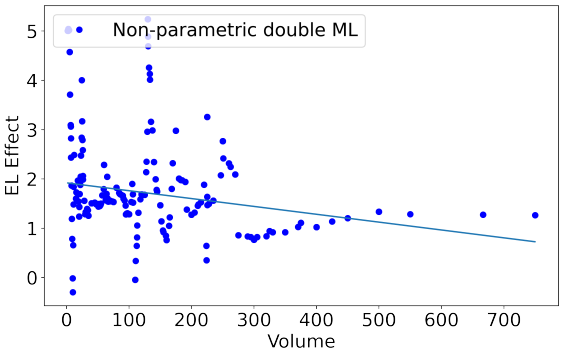
Figure 4.13.: Causal analysis: An illustration of the analysis of a potential heterogeneous effect of the expected loss on the premium based on the sign of the change in the reinsurance index. The conditional average treatment effect (CATE) could differ between the two subgroups.

To simplify my analysis, I do not split my observations into groups. Instead, I analyze

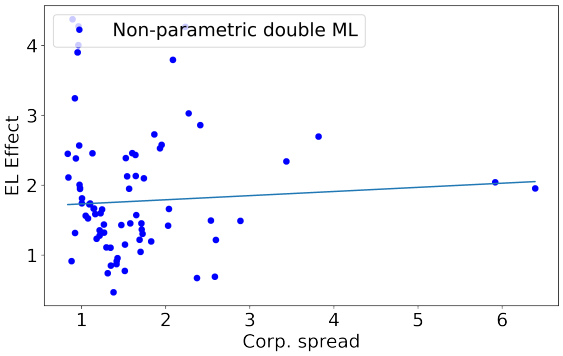
the treatment effect for the whole spectrum of possible values in the feature that may cause heterogeneity. Knowing that the expected loss is by far the most important feature, I examine whether any of the other characteristics strongly influence the effect of expected loss on the premium and whether this effect is heterogeneous. I do so by using a causal forest including double machine learning (the *CausalForestDML* method of the *EconML* package). Similar to before, I train on 80% of the data and test the results on the remaining 20%. The splitting criterion is not the mean squared error, but the heterogeneity in the split. This means that the quality of a split in a decision tree is measured by the heterogeneity score. Since I am mainly interested in a potential heterogeneous effect of the expected loss, this splitting criterion seems more appropriate for this analysis than the mean squared error. The causal forest is formed by 10,000 decision trees, and 10-fold cross-validation should stabilize the results. After training the causal forest model on the train sample, I compute the causal effect for the train and test sample. Ideally, the effects for the test and train sample should be similar. This is indeed the fact, as I cannot find much difference in the predictions between the train and test sample for any of the features.

The results for the test sample are plotted in Figure 4.14. I analyze only the most interesting features in detail. A higher volume may reduce the effect of the expected loss, and the effect seems to vary extensively for low volumes. This seems to be in line with Figure 4.5. There, I showed that the premium tends to be lower for high volume bonds. A major reason could be that the effect of the expected loss diminishes for these bonds. High volume bonds may mostly be issued by big, trustworthy sponsors, which may lower the required premium. In addition, higher volumes usually lead to more liquidity. A high volume size appears to stabilize the effect of the expected loss according to the plot. The effects of the corporate spread, the maturity left, the indemnity trigger, and the number of peril locations and types are not clear from the plots. The dots of the predicted effects on the expected loss do not follow a clear structure. It is important to note that scaling is important. Initially, it looks like the effect of the indemnity trigger has a large impact, but this is not the case, since there is only a difference of around 0.04 percentage points. A positive change in the reinsurance index instead of a negative one seems to amplify the effect of the expected loss. Similarly, a positive change in the S&P500 appears to increase the effect of the expected loss. This is not intuitive, as a positive change in the S&P500 may lead to a positive business environment, in which risks are not perceived as harmful. But this would lead to a lower effect of the EL in good economic times.

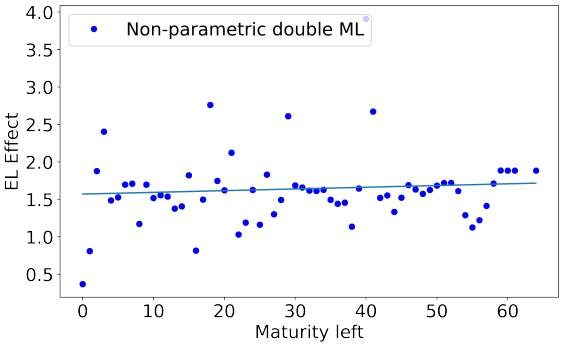
The Shapley values can again be used to observe heterogeneity in the treatment, as shown in Figure 4.15. This additional analysis hopefully verifies the previous results. According to the average impact analysis, the largest impacts on the expected loss stem from the corporate spread, the volume, the hurricane peril, the cat bond index, and the maturity left. The graph on the right helps at observing the most influential features and the strength of their impact on the expected loss. This graph analyzes the absolute effect using absolute Shapley values. The left graph helps further analyze the heterogeneity of this effect by also considering the sign of the effects. For example, the hurricane peril, the third feature of the graph, appears to have the third largest effect on the expected loss effect. This effect is heterogeneous. If the hurricane peril is positive, i.e., a bond is exposed to this peril, this leads to a high Shapley value. Thus, when there is a hurricane peril, the expected loss appears to have a larger impact on the premium. In contrast, when this not the case, as plotted in blue, the Shapley values are affected negatively. As the blue dots are still relatively close to the 0 impact line, this impact appears to be less significant than for positive feature values. The interpretation of the corporate spread is slightly more complex. A low corporate spread may cause a positive



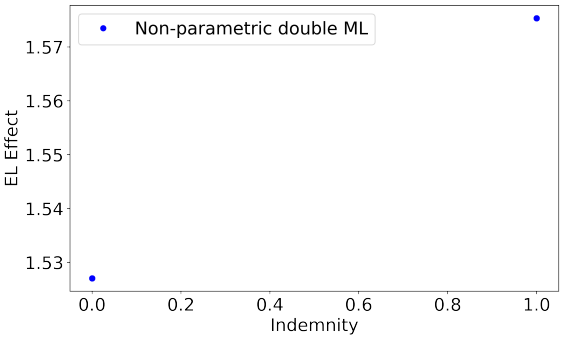
(a) Volume



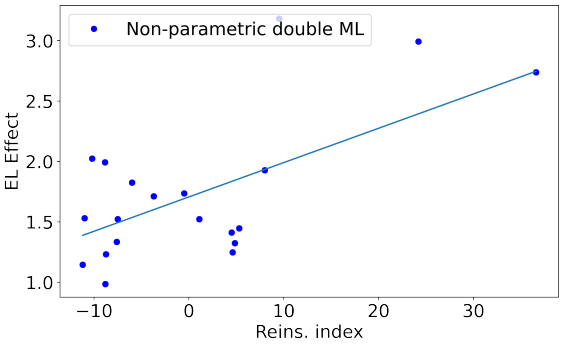
(b) Corporate spread



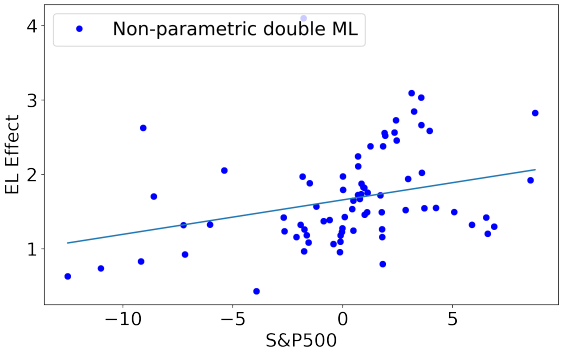
Maturity (in months)



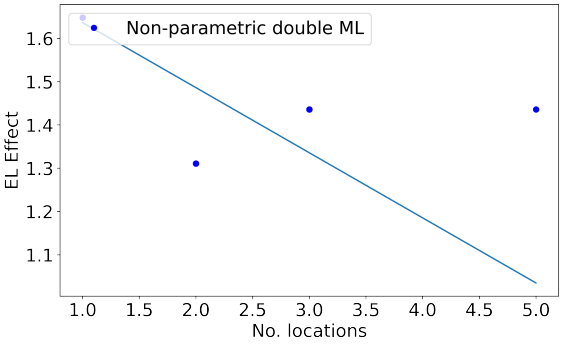
Indemnity



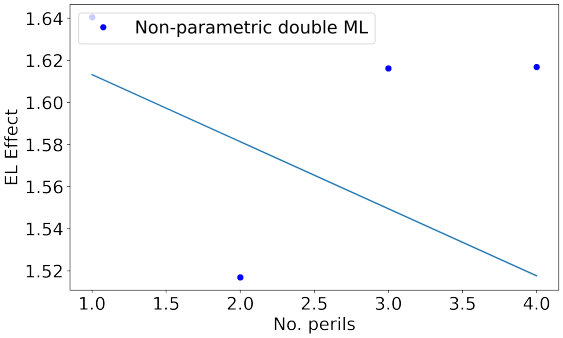
Change in reins. index



Change in S&P500

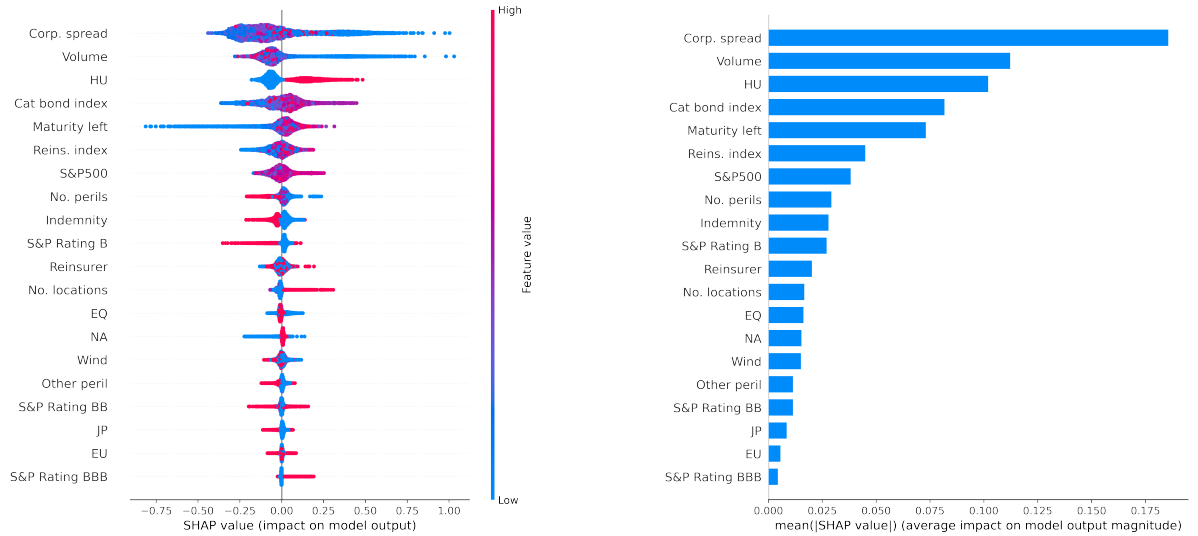


Number of peril locations



Number of peril types

Figure 4.14.: HTE of expected loss.



(a) Impact on model output for EL as treatment.

(b) Average impact on model output for EL as treatment.

Figure 4.15.: Causal SHAP analysis: Shapley values for causal forest with expected loss as treatment.

effect on the Shapley value. However, many blue dots are also plotted for low Shapley value. This means that a low corporate spread can have both a positive and a negative effect on the impact of the expected loss on the premium. This is consistent with the results in Figure 4.14, as the EL effect is highly volatile for varying corporate spreads.

5. Conclusion and future research

Recently, considerable progress has been made in developing methods for inference in the specific setting of random forests (Wager & Athey, 2018). I apply these methods in my work and quantify uncertainties and heterogeneities of effects in the cat bond market. For many other methods, however, the construction of (asymptotically) valid confidence intervals remains impossible (Athey & Imbens, 2019).

Before using causal methods, the question arises whether confidence intervals are as important as the traditional econometric literature suggests. The mere possibility of inference is not a reason to use it, as it often reduces predictive performance (Athey & Imbens, 2019). When causal inference is the aim of a machine learning algorithm, its optimization criteria must be modified. A model with low bias and low variance would be desirable, but this is usually not achievable. In fact, there is a tradeoff between bias and variance. Using causal random forests in this work, the confidence intervals of the conditional average treatment effects are very large. Nevertheless, they give a first indication of the uncertainty. Since I also use regressions, I can compare the estimated effects from different models.

Despite the bias-variance tradeoff, causal random forests provide a sophisticated empirical toolbox, which is applicable across a range of domains. So far, the interactions among predictors have not been extensively studied, and the explanatory predictors may be non-linear. My heterogeneity analysis shows which factors most influence the effect of the expected loss on the premiums. This partially explains the heterogeneity of the effect. Capturing heterogeneity in a key parameter of interest is crucial because an average treatment effect is not helpful if it fluctuates widely across sample groups. Shapley values add an additional layer of analysis, as they make individual predictions understandable.

Importantly, the interpretability of my results relies heavily on the assumption that there are no unobserved confounders. If an important feature that significantly affects the cat bond spread is missing, then there are unobserved confounders. This would make any interpretative conclusions impossible. Since my data are not from a randomized experiment, I must use double machine learning to account for the potential correlation of confounders with both the "treatment" variable and the cat bond premium.

There are a few other limitations to my findings that leave room for further research. First, data availability is generally a strong limitation in the literature on cat bond asset pricing (Braun et al., 2019). Extreme event risks require data from a long time horizon, as securitized events only occur very infrequently. Second, I do not consider time effects. It is possible that certain effects change over time. This could also explain some of the heterogeneity. Similarly, my results could be manipulated by a selection bias. For instance, I had to exclude bonds with missing entries. Although my results provide additional insights, they still do not explain why the cat bond market has generated high excess returns over the past two decades. If natural disaster risk is diversifiable by capital market investors, and systematic risks from the broader financial markets are minimized to an almost negligible extent, this should not be the case. Finally, causal methods are not only applicable to random forests. It would be interesting to apply them to neural networks and other machine learning models.

References

Angelopoulos, A. N., & Bates, S. (2021). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*.

Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, *113*(27), 7353–7360.

Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, *11*(1), 685-725. Retrieved from https://doi.org/10.1146/annurev-economics-080217-053433 doi: 10.1146/annurev-economics-080217-053433

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, *47*(2), 1148–1178.

Braun, A. (2016). Pricing in the primary market for cat bonds: new empirical evidence. *Journal of Risk and Insurance*, *83*(4), 811–847.

Braun, A., Ammar, S. B., & Eling, M. (2019). Asset pricing and extreme event risk: Common factors in ils fund returns. *Journal of Banking & Finance*, *102*, 59–78.

Braun, A., Herrmann, M., & Hibbeln, M. T. (2022). Common risk factors in the cross section of catastrophe bond returns. *Available at SSRN 3901695*.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Routledge.

Carpenter, G. (2012). Catastrophes, cold spots and capital. navigating for success in a transitioning market. *Guy Carpenter, New York*.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). *Double/debiased machine learning for treatment and structural parameters*. Oxford University Press Oxford, UK.

Criminisi, A., & Shotton, J. (2013). *Decision forests for computer vision and medical image analysis*. Springer Science & Business Media.

Efron, B., & Hastie, T. (2021). *Computer age statistical inference, student edition: Algorithms, evidence, and data science* (Vol. 6). Cambridge University Press.

Galeotti, M., Gürtler, M., & Winkelvos, C. (2013). Accuracy of premium calculation models for cat bonds—an empirical analysis. *Journal of Risk and Insurance*, *80*(2), 401–421.

Götze, T., Gürtler, M., & Witowski, E. (2020). Improving cat bond pricing models via machine learning. *Journal of Asset Management*, *21*(5), 428–446.

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, *33*(5), 2223–2273.

Gürtler, M., Hibbeln, M., & Winkelvos, C. (2016). The impact of the financial crisis and natural catastrophes on cat bonds. *Journal of Risk and Insurance*, *83*(3), 579–612.

Götze, T., & Gürtler, M. (2020). Risk transfer and moral hazard: An examination on the market for insurance-linked securities. *Journal of Economic Behavior & Organization*, *180*, 758-777. Retrieved from https://www.sciencedirect.com/science/article/pii/S0167268119302008 doi: https://doi.org/10.1016/j.jebo.2019.06.010

Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, *34*(11), 2767–2787.

Louppe, G. (2014). Understanding random forests: From theory to practice. *arXiv preprint arXiv:1407.7502*.

Makariou, D., Barrieu, P., & Chen, Y. (2021). A random forest based approach for pre-

dicting spreads in the primary catastrophe bond market. *Insurance: Mathematics and Economics*, *101*, 140–162.

Molnar, C. (2022). *Interpretable machine learning* (2nd ed.). Retrieved from https://christophm.github.io/interpretable-ml-book

Mullainathan, S., & Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87–106.

Papachristou, D. (2011). Statistiscal analysis of the spreads of catastrophe bonds at the time of issue. *ASTIN Bulletin: The Journal of the IAA*, *41*(1), 251–277.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379-423. doi: 10.1002/j.1538-7305.1948.tb01338.x

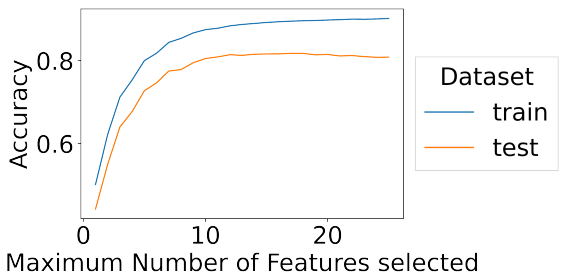
Shapley, L. S. (1953). A value for n-person games. In H. W. Kuhn & A. W. Tucker (Eds.), *Contributions to the theory of games (am-28), volume ii* (pp. 307–318). Princeton: Princeton University Press. Retrieved from https://doi.org/10.1515/9781400881970-018 doi: doi:10.1515/9781400881970-018

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, *113*(523), 1228–1242.

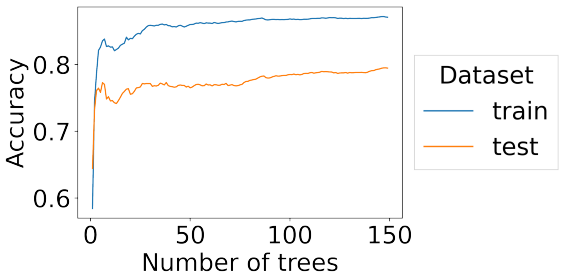
A. Additional figures and tables for the analysis of the cat bond market

Table A.1.: Summary statistics without trimming: cardinal cat bond specific and macroeconomic variables (reported at secondary market level)

	Obs.	Mean	SD	Min.	Max.
Cat bond specific variables					
Secondary market premium (annual, in %)	7806	23.86	598.51	0.00	35452.00
Expected loss (annual, in %)	7806	2.38	2.25	0.00	15.75
Volume (in USD million)	7806	142.41	120.58	1.75	1500.00
Maturity left (in month)	7806	21.85	13.50	0.00	67.00
No. perils	7806	1.76	0.77	1.00	4.00
No. locations	7806	1.26	0.63	1.00	5.00
Macroeconomic variables					
Reins. index (annual change, in %)	7806	-0.89	10.20	-11.20	36.59
S&P500 (monthly change, in %)	7806	0.33	4.04	-12.51	8.76
Corp. spread (monthly, in %)	7806	1.63	0.91	0.83	6.39
Cat bond index (monthly change, in %)	7806	0.54	0.63	-1.13	2.41



(a) Random forest accuracy in dependence of number of features.



(b) Random forest accuracy in dependence of number of trees.

Figure A.1.: Random forest accuracy in dependence of number of features and trees for the issue level.

Table A.2.: Summary statistics without trimming: nominal and ordinal cat bond specific variables (reported at secondary market level)

	Obs.	Percentage
Trigger		
Indemnity	3231	41.39
Non-indemnity	4575	58.61
Peril type		
EQ	4983	63.84
HU	2601	33.32
Wind	3424	43.86
Other	2464	31.57
Peril location		
EU	1730	22.16
JP	1123	14.39
NA	6232	79.84
Latin America	324	4.15
Asia/Australia	366	4.69
Sponsor		
Reinsurer	3398	43.53
Insurer	3999	51.23
Other	409	5.24
Rating		
S&P Rating AA	11	0.14
S&P Rating A	38	0.49
S&P Rating BBB	224	2.87
S&P Rating BB	2586	33.13
S&P Rating B	1181	15.13
S&P Rating CC	13	0.17
S&P Rating NR	3753	48.08

Table A.3.: OLS regression results – Issue level

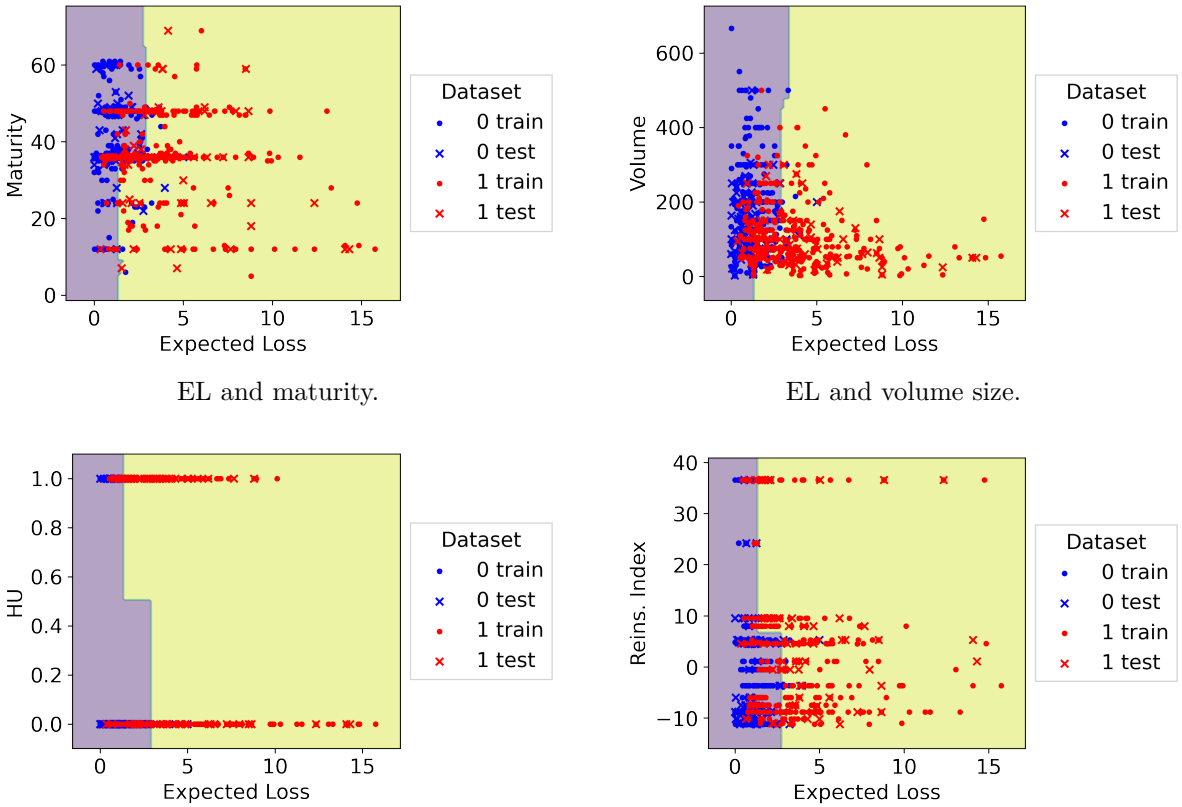
Dep. Variable:	Premium (at issue, in %)	**R-squared:**	0.821
Model:	OLS	**Adj. R-squared:**	0.814
Method:	Least Squares	**F-statistic:**	120.5
Date:	Mon, 05 Dec 2022	**Prob (F-statistic):**	8.10e-244
Time:	15:39:48	**Log-Likelihood:**	-1611.9
No. Observations:	736	**AIC:**	3280.
Df Residuals:	708	**BIC:**	3409.
Df Model:	27		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	0.6199	0.569	1.090	0.276	-0.497	1.736
Expected Loss	1.5449	0.040	38.936	0.000	1.467	1.623
Volume	-9.12e-05	0.001	-0.120	0.904	-0.002	0.001
Maturity	-0.0328	0.008	-4.072	0.000	-0.049	-0.017
No. perils	0.5407	0.358	1.510	0.131	-0.162	1.243
No. locations	0.4030	0.408	0.989	0.323	-0.397	1.203
Reins. Index	0.0769	0.007	10.591	0.000	0.063	0.091
S&P500	0.0179	0.021	0.849	0.396	-0.024	0.059
Corp. spread	0.9619	0.129	7.472	0.000	0.709	1.215
CAT bond Index	-0.1479	0.151	-0.980	0.328	-0.444	0.148
Indemnity	-0.0082	0.232	-0.035	0.972	-0.463	0.447
EQ	-0.6099	0.356	-1.711	0.088	-1.310	0.090
HU	2.1397	0.387	5.530	0.000	1.380	2.899
Wind	-0.2374	0.386	-0.615	0.539	-0.995	0.520
Other peril	-0.4480	0.433	-1.035	0.301	-1.298	0.402
EU	-0.2219	0.463	-0.479	0.632	-1.131	0.688
JP	1.3239	0.450	2.940	0.003	0.440	2.208
NA	1.4380	0.428	3.363	0.001	0.599	2.278
Latin America	-1.2482	0.595	-2.097	0.036	-2.417	-0.079
Asia/Australia	-1.1870	0.571	-2.078	0.038	-2.308	-0.066
Reinsurer	0.3854	0.219	1.759	0.079	-0.045	0.816
Other	0.2153	0.477	0.452	0.652	-0.720	1.151
S&P Rating AA	-6.3756	2.269	-2.809	0.005	-10.831	-1.920
S&P Rating A	-3.0780	1.265	-2.434	0.015	-5.561	-0.595
S&P Rating BBB	-1.6888	0.593	-2.848	0.005	-2.853	-0.525
S&P Rating BB	-0.2292	0.244	-0.940	0.347	-0.708	0.249
S&P Rating B	0.8270	0.276	2.999	0.003	0.286	1.368
S&P Rating CC	-0.5177	2.233	-0.232	0.817	-4.901	3.866

Table A.4.: OLS regression results after recursive feature elimination – Issue level

Dep. Variable:	Premium (at issue, in %)	**R-squared:**	0.775
Model:	OLS	**Adj. R-squared:**	0.772
Method:	Least Squares	**F-statistic:**	277.9
Date:	Mon, 05 Dec 2022	**Prob (F-statistic):**	2.19e-228
Time:	15:47:52	**Log-Likelihood:**	-1696.6
No. Observations:	736	**AIC:**	3413.
Df Residuals:	726	**BIC:**	3459.
Df Model:	9		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	-0.6458	0.329	-1.965	0.050	-1.291	-0.000
Expected Loss	1.6127	0.038	42.267	0.000	1.538	1.688
Corp. spread	0.7780	0.138	5.656	0.000	0.508	1.048
HU	2.7780	0.224	12.399	0.000	2.338	3.218
JP	1.5034	0.296	5.084	0.000	0.923	2.084
NA	1.7646	0.260	6.778	0.000	1.253	2.276
Reinsurer	0.9331	0.184	5.068	0.000	0.572	1.295
S&P Rating AA	-3.1197	2.455	-1.271	0.204	-7.940	1.701
S&P Rating A	-1.9896	1.244	-1.599	0.110	-4.432	0.453
S&P Rating B	0.9349	0.271	3.447	0.001	0.402	1.467



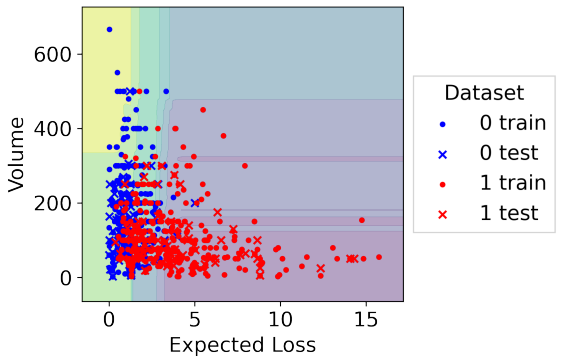
EL and maturity.

EL and volume size.

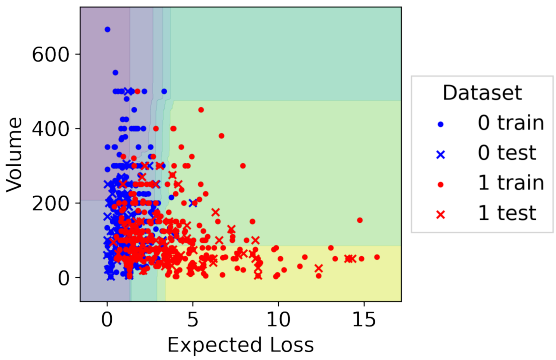
EL and HU.

EL and Reins. index.

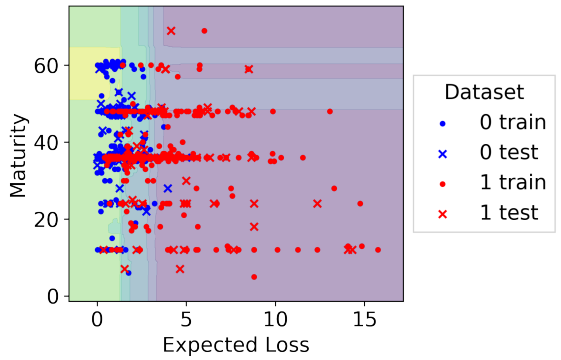
Figure A.2.: The relationship of the most important feature, EL, with four other important features is analyzed for the primary market. The blue markers ("0") represent observations with a lower than median premium, whereas the red ones ("1") have higher than median premiums. The background color depicts whether the random forest classifier predicts a high premium (yellow) or a low premium (violet).



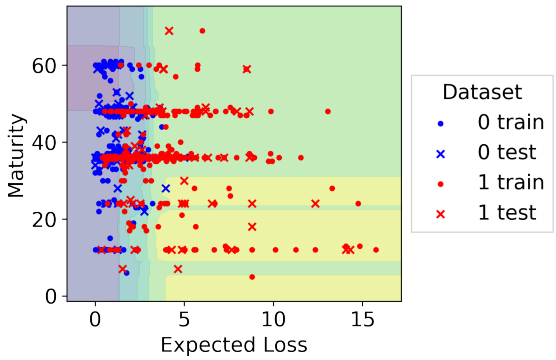
Probabilities for being a below median premium bond for different combinations of expected loss and volume size.



Probabilities for being an above median premium bond for different combinations of expected loss and volume size.



Probabilities for being a below median premium bond for different combinations of expected loss and maturity.



Probabilities for being an above median premium bond for different combinations of expected loss and maturity.

Figure A.3.: The prediction probabilities are depicted by different background colors. Yellow stands for a high probability, and violet a low probability. The blue markers ("0") represent observations with a lower than median premium, whereas the red ones ("1") have higher than median premiums.

Declaration of Authorship

I hereby declare

- that I have written this thesis without any help from others and without the use of documents and aids other than those stated above;
- that I have mentioned all the sources used and that I have cited them correctly according to established academic citation rules;
- that I have acquired any immaterial rights to materials I may have used such as images or graphs, or that I have produced such materials myself;
- that the topic or parts of it are not already the object of any work or examination of another course unless this has been explicitly agreed on with the faculty member in advance and is referred to in the thesis;
- that I will not pass on copies of this work to third parties or publish them without the University's written consent if a direct connection can be established with the University of St.Gallen or its faculty members;
- that I am aware that my work can be electronically checked for plagiarism and that I hereby grant the University of St.Gallen copyright in accordance with the Examination Regulations in so far as this is required for administrative action;
- that I am aware that the University will prosecute any infringement of this declaration of authorship and, in particular, the employment of a ghostwriter, and that any such infringement may result in disciplinary and criminal consequences which may result in my expulsion from the University or my being stripped of my degree.

St. Gallen, February 17, 2023

Tim Mathes

(Signature of the candidate)

By submitting this academic term paper, I confirm through my conclusive action that I am submitting the Declaration of Authorship, that I have read and understood it, and that it is true.

MASTER'S THESIS

For the attainment of the degree
Master of Science in Finance at the Stockholm School of Economics

Part II

Cat Bond Markets: A Time Analysis of the Causal Random Forest Approach

Tim Ludwig Leonard Matheis

Supervisor: Tobias Sichert, PhD

Submitted: February 17, 2023

Abstract

This work is a contribution to the causal analysis of the catastrophe bond market, which has generated high excess returns over the past two decades. Since these excess returns remain partly unexplained and the interest in catastrophe bonds is growing, the causal analysis of the factors affecting their premiums is of high relevance. Most studies about the catastrophe bond market have used only linear models. However, more complex models such as random forests may be better suited for modelling this market. In addition, the interactions among predictors have not been extensively studied, and there may be non-linearities in the explanatory predictors. Considerable progress has been made in developing methods for inference in the specific setting of random forests. Causal random forests – especially when combined with double machine learning and Shapley values – represent a sophisticated empirical toolbox. They provide unbiased prediction intervals and allow the analysis of heterogeneous effects, while Shapley values help interpret individual predictions. In Part I of my thesis, I have already carried out a detailed analysis of the interactions between other predictors. In this Part II of my thesis, I apply these methods to quantify uncertainties and heterogeneities of effects in the primary catastrophe bond market with respect to the issue date of the bonds. In particular, I analyze whether the expected loss has a time-varying effect on the catastrophe bond premiums. My results confirm that the effect of the expected loss seems to have decreased in the primary catastrophe bond market in recent years.

Contents

List of Figures

List of Tables

List of Abbreviations

ANOVA Analysis of variance
ATE Average treatment effect
CART Classification and regression trees (algorithm)
CAT Catastrophe
CATE Conditional average treatment effect
CEL Conditional expected loss
CRF Causal random forest
DML Double machine learning
EL Expected loss
EQ Earthquake
EU Europe
HTE Heterogeneous treatment effect
HU Hurricane
ILS Insurance linked security
JP Japan
LA Latin America
LIBOR London inter-bank offered rate
LIME Local interpretable model-agnostic explanations
NA North America
OLS Ordinary least squares
PCA Principal component analysis
PFL Probability of first loss
PLL Probability of exhaust
RCT Randomized controlled trial
RF Random forest
RFE Recursive feature elimination
RMSE Root-mean-square error
SD Standard deviation
SHAP Shapley additive explanations

1 Introduction

Random forests and other machine learning models often provide superior predictive power. As many essential problems are formulated as prediction problems, evaluating goodness of fit on a test set is sufficient in such a case (Mullainathan & Spiess, 2017). However, although better performance in terms of out-of-sample predictive power is valuable, a valid confidence interval is often more or equally important. Efron and Hastie (2021) criticize that prediction is an area where algorithmic developments have run far ahead of their inferential justification. One of the reasons for this is the model-free nature behind many well-performing prediction methods. A single decision tree is easy to interpret but has low predictive power. Conversely, random forests provide high predictive power at the cost of lower interpretability, as it is unclear what it means to make a final prediction based on the predictions of multiple decision trees. Thus, there is a trade-off between interpretability and predictive power.

An average treatment effect of a parameter of interest may not be sufficient because it does not capture the degree of uncertainty. But the quantification of uncertainty is critical when deciding about whether to implement a treatment or not. Despite the absence of a "treatment" in the cat bond market, the analysis of the interaction of various parameters is very interesting and provides insights into the pricing of this asset. Since random forests can provide very good predictive results in the cat bond market (Makariou, Barrieu, & Chen, 2021), additional causal analysis of the results obtained with random forests is important. A recent branch of economic literature focuses on adapting and tuning machine learning techniques to causal problems (Athey & Imbens, 2019).

In this part II of my thesis, I apply these methods to quantify uncertainties and heterogeneities in the effects of the expected loss on the premiums in the primary cat bond market in dependence of the issuance date of the bonds. Hence, I analyze whether the issuance date or the events during that time have a strong effect on the premiums. In particular, I focus on whether the effect of the expected loss on the premiums has changed over time. My results confirm that the effect of the expected loss on the premiums seems to have decreased in recent years.

This work is a contribution to the literature on asset pricing and in particular the causal analysis of the catastrophe bond market. I not only apply machine learning methods for predicting cat bond premiums, but also put special emphasis on the underlying causalities of the methods. My analysis may help practitioners evaluate the potential of causal machine learning methods for asset pricing. The interpretability of machine learning models helps improve them, build confidence in them, justify model predictions, and gain insights.

The next chapter 2 provides an overview of the literature on the time changes in the cat bond market in addition to the extensive literature review in part I of my thesis. The following chapter 3 explains the methodology for causal cat bond pricing methods. Although it is almost the same as in part I of my thesis, I still included it as it is needed for my analysis. The main part of my work, chapter 4, contains my quantitative analysis of the primary cat bond market with specific focus on the expected loss over time. In chapter 5, I discuss the strengths and weaknesses of this work and suggest possible areas for future research.

2 Literature on the time changes in the cat bond market

In this literature review, I only review recent research on the impact of the time factor on cat bond prices and spreads. I already provide an overview of the general cat bond pricing literature and the causal machine learning literature in part I of my thesis. The reader is encouraged to read those sections in addition.

Early research on cat bonds shows that natural catastrophes, such as Hurricane Katrina, or financial crises, such as the financial crises 2008, have an impact on cat bond premiums (Gürtler, Hibbeln, & Winkelvos, 2016). According to this research, the increased premium caused by catastrophes mostly stems from an increased expected loss coefficient. This early research suggests that the time – or the events occurring within a month or year – is relevant for the analysis of cat bond premiums. Some of the heterogeneity found in part I of my thesis may be due to time effects.

Carayannopoulos, Kanj, and Perez (2020) find an overall decreasing trend in the price of expected loss risk for the period 1999-2016. Over the same period, cat bond prices increased by an average of 34% due to large catastrophes. In contrast to prior research, supporting that investors' perceptions about catastrophe risk changed, Carayannopoulos et al. (2020) argue that the pricing differences may be caused by a change in investors' effective risk aversion. According to their theory, catastrophic events that trigger cat bonds make investors converge to their habit consumption levels. In addition, they find that contagion effects from financial markets have only a minor effect. These effects seem to be relevant only during major financial crises.

Cat bond spreads are strongly affected by seasonality as the probability of cat bonds being triggered depends on the season. For example, Braun, Herrmann, and Hibbeln (2022) construct a seasonality amplitude factor that describes how much the PFLC (constant probability of first loss per month) is scaled up (e.g., in the peak season when, for example, hurricanes are likely to occur) or down (e.g., in the off-season) for each bond and month. At times when the underlying catastrophe risk is high, the spreads are also higher. For example, the spread of a hurricane bond is highest at the start of the hurricane season and lowest at the end of it (if it was not triggered). Herrmann and Hibbeln (2021) develop a framework to model this seasonality. Their results suggest that seasonality accounts for a large fraction of market variation in spreads. For example, up to 47% of market fluctuations in the spreads of single-hurricane bonds are caused by seasonality according to their model. Seasonality is probably not a factor that needs to be considered in the primary cat bond market because the bonds tend to be issued during the same low-risk seasons. However, the secondary market prices are clearly affected by seasonality. Although seasonality affects cat bond prices, the relationship between the seasonality of catastrophic events and cat bond spreads remains unexplored in the empirical literature (Herrmann & Hibbeln, 2021). One reason may be that the empirical literature on cat bonds has not extensively examined the secondary market so far.

3 Methodology for random forests & causal inference

Machine learning offers automated procedures for predicting phenomena based on their past observations. In this way, underlying patterns in the data are revealed, which may provide new insights into (causal) relationships. However, the application of algorithms demands an understanding of the underlying mechanisms, assumptions, and limitations for the interpretations of their results.

This chapter provides an overview of the machine learning techniques applied in this work. Mainly, two classes of algorithms are discussed: decision trees (Breiman, Friedman, Olshen, & Stone, 1984) and random forests (Breiman, 2001). These algorithms have proven to be an accurate and robust tool for solving many machine learning tasks such as regression, classification, density estimation and semi-supervised learning (Criminisi & Shotton, 2013).

Since random forests are not optimal for causal inference, generalized random forest have been developed (Athey, Tibshirani, & Wager, 2019). Generalized random forests offer new methods for three statistical tasks: non-parametric quantile regression, conditional average partial effect estimation, and heterogeneous treatment effect estimation. In addition, I explain double machine learning and Shapley values, which are important tools in causal machine learning.

3.1 Machine learning

Machine learning can be described as the study of systems that learn from data without being explicitly programmed. The learning from data is reflected in an increasing performance measure as additional data is utilized in the learning process. Yet, machine learning should not be limited to producing algorithms that make accurate predictions. In addition, machine learning algorithms aim at providing insights into the predictive structure of the data (Breiman et al., 1984). More specifically, I am interested in extracting the variables and interactions between variables driving a phenomenon. Otherwise, it is difficult to accept or trust the results from a "black box" when the process leading to the results is incomprehensible.

A supervised learning task can be stated as learning a function $\varphi : \mathcal{X} \to \mathcal{Y}$ from a learning set $\mathcal{L} = (\mathbf{X}, \mathbf{y})$ (Louppe, 2014). The goal of the task is to find a model that yields predictions $\varphi(\mathbf{x})$, often denoted by the variable $\hat{Y}$, that closely approximate the true outcome variable Y. However, since one is usually interested in applying such a model to unseen data, the model should learn general relationships rather than over-fitting the data. Hence, instead of minimizing the error for the known learning set $\mathcal{L}$, one aims at minimizing the error for all possible values $(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}$. Precisely, the objective is to minimize the expected prediction error of the model $\varphi_{\mathcal{L}}$, as defined in Equation 3.1. In this formula, $\mathcal{L}$ is the learning set used to build the model $\varphi_{\mathcal{L}}$, and L is a loss function measuring the discrepancy between the two arguments, the predicted outcome and the actual outcome.

$$Err(\varphi_{\mathcal{L}}) = \mathbb{E}_{X,Y}\left\{L\left(Y, \varphi_{\mathcal{L}}(X)\right)\right\} \tag{3.1}$$

In this work, X represents a set of characteristics of a cat bond, such as the volume size and the sponsor. Therefore, $\mathbf{x}$ is specific combination of features of a bond. Since Y represents the expected cat bond returns in this work, it is a numerical variable, making the learning task a regression problem. For such a task, a simple loss function is the squared error loss, as displayed in Equation 3.2.

$$Err(\varphi_{\mathcal{L}}) = \mathbb{E}_{X,Y}\left\{\left(Y - \varphi_{\mathcal{L}}(X)\right)^2\right\} \tag{3.2}$$

Whenever the outcome variable Y is instead a categorical variable, the learning task is a classification problem. To make the learning problem in this work a classification task, I could classify the expected cat bond returns according to their magnitude, e.g., high, moderate, low. This would transform Y from a numerical variable into a categorical variable. A possible loss function for a classification task would be the zero-one loss function, $L(Y, \varphi_{\mathcal{L}}(X)) = 1(Y \neq \varphi_{\mathcal{L}}(X))$, where all misclassifications are penalized equally. Here, the error would become the probability of misclassification.

In contrast, unsupervised machine learning algorithms learn from unlabeled data. This means that the learning set does not contain data couples of the features and the outcome, $(\mathbf{x}, y)$. Crucially, y is not given. A typical example is clustering. Unsupervised learning, however, is not relevant to this work.

3.2 Tree-based methods

The ambition behind the application of machine learning is usually twofold – making accurate predictions, while allowing the extraction of knowledge. Tree-based models are one of the most promising techniques, as they deliver reliable and understandable results. They are used for classification and regression tasks.

For simplicity, I explain the underlying logic for a classification problem. This is not a problem because the expected cat bond returns could be grouped by their values or simply rounded to make them a finite set of values. The expected cat bond returns Y can then be expressed as a partition over the universe $\Omega = \Omega_{c_1} \cup \Omega_{c_2} \cup ... \cup \Omega_{c_j}$. Thus, the output space consists of j categories. Likewise, a classifier φ can determine a partition of the universe Ω by approximating the "true" cat bond returns Y by $\hat{Y}$. In this case, the partition is defined over the input space X: $X = X^{\varphi}_{c_1} \cup X^{\varphi}_{c_2} \cup ... \cup X^{\varphi}_{c_j}$. Precisely, an input value – could be a vector of multiple explaining variables – is mapped to one of j output categories based on the magnitude and sign of the explaining variables. In the case of cat bonds, a cat bond could in theory be specified to have a high expected return if the bond is triggered by even mild wind storms at the West coast of the US. In this simple example, there could be two explaining variables – type of catastrophe and country – that allow for splitting the tree accordingly. Thus, the idea behind tree structured models is to approximate the partition of a model by recursively partitioning the input space X into subspaces. Consequently, predictive values $\hat{y}$, such as high expected return, can be assigned to all objects $\mathbf{x}$ within each terminal subspace.

I now briefly introduce some required concepts. Trees are a way of representing a model $\varphi : \mathcal{X} \to \mathcal{Y}$, where an outcome y is determined by its explaining variables X. A tree is a graph $G = (V, E)$, where V denotes its vertices and E its edges. In such a graph, any two vertices

are connected by one path. This is a logical requirement for the concise mapping of features to an outcome. A tree usually has a root (rooted tree), which represents the whole input space X. Starting from the root, a path leads to the tree's leaves, the terminal vertices. When an edge leads from one vertex to a vertex below, the upper vertex is called the parent vertex or node, while the bottom one is the child node. Each node t represents a subspace of the input space: $\mathcal{X}_t \subseteq \mathcal{X}$. The further down the tree, the smaller the subset becomes. On the way down, a splitting rule is applied at each edge to determine the child node. For example, this could be whether a catastrophe is insured in the US or somewhere else.

Decision trees, which underlie all the following tree-based methods, are particularly attractive because of their good properties. First, as they are non-parametric, they can model arbitrarily complex relationships between inputs and outputs, without any a priori assumption. Second, trees intrinsically implement feature selection, making them relatively robust to irrelevant or noisy variables. This also makes them relatively robust to outliers and labelling errors. Third, they can (simultaneously) handle heterogeneous data types, e.g., numerical, and categorical variables. Finally, since they can be represented graphically, they are relatively easy to interpret.

3.3 Decision trees

As foreshadowed before, the predicted output value $\varphi(\mathbf{x})$ is the label of the leaf reached by the instance $\mathbf{x}$, when propagated through the tree by following the splits s_t. The global error of the model, already defined in Equation 3.1, can be further resolved as in Equation 3.3. By minimizing the local error in the terminal nodes, the global error is now also minimized. Looking at the formula, the minimization is done over the set of terminal nodes $\tilde{\varphi}$. Additionally, the probabilities of the input variables are taken into account.

$$Err(\varphi) = \mathbb{E}_{X,Y}\left\{L\left(Y, \varphi(X)\right)\right\} = \sum_{t \in \tilde{\varphi}} P(X \in \mathcal{X}_t)\mathbb{E}_{X,Y|t}\left\{L\left(Y, \tilde{y}_t\right)\right\} \tag{3.3}$$

Figure 3.1 shows an exemplary decision tree for the classification of cat bonds into different categories according to the size of their (predicted) premium. When classifying an instance $\mathbf{x}$, in this example a particular cat bond, this instance is propagated from the top of the tree down to one of the leaves. This leaf is then the prediction. If the prediction differs from the true observation, the error is large. But if the splits are well chosen, the error (between true and predicted results) is hopefully small, and the prediction represents true relationships. These relationships are at least correlations and at best causal relationships.

To build a tree, there must be a measure that evaluates the goodness of a possible split at a node. Breiman et al. (1984) define such an impurity measure $i(t)$ that evaluates the goodness of a node t. When $i(t)$ is small, the node is regarded pure, leading to better predictions $\hat{y}_t(x)$ for all $x \in \mathcal{L}_t$, where $\mathcal{L}_t$ denotes the subset of learning samples falling into t, all $\mathbf{x} \in \mathcal{X}_t$. Equation 3.4 defines the impurity decrease of a binary split s – the most typical split in a decision tree – dividing a node t into a left node t_L and a right node t_R. In this formula, p_L is the proportion $\frac{N_{t_L}}{N_t}$ of learning samples from $\mathcal{L}_t$ going to the left node t_L, with N_t denoting the size of the subset $\mathcal{L}_t$. The proportion for the right node is determined symmetrically. Applying this measure, the entire learning set $\mathcal{L}$ can be iteratively divided to reach increasingly purer nodes.

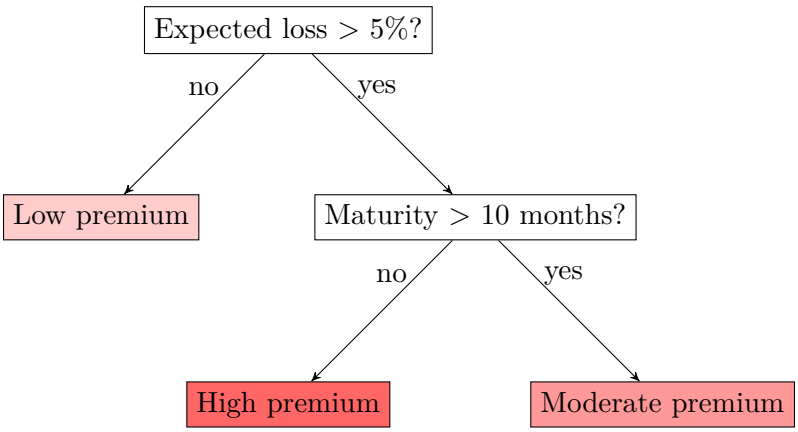


Figure 3.1: This is an exemplary decision tree for determining the size of the premium. According to this example, a cat bond with an expected loss of less than 5% would be classified as a low premium bond. However, if a bond had an expected loss of more than 5% and a maturity of fewer than 10 months, the premium would be predicted to be high.

$$\Delta i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R) \tag{3.4}$$

So far, the impurity measure is still very abstract. In fact, there are multiple impurity and purity measures. The target of purity and impurity measures is to minimize the uncertainty of the outcome. For instance, a toss of a fair coin leads to an unpredictable outcome because the probability is the same for both sides. In this case, there is a lot of impurity. The degree of uncertainty is often described by the concept of entropy. Claude Shannon describes entropy as the average amount of information required to encode a randomly drawn value of a set X (Shannon, 1948). In the discrete case, entropy is defined as the expectation of all negative logarithmized probabilities of the possible outcomes: $H(X) = \mathbb{E}[-log\{P(X)\}] = -\sum_{i=1}^{N} P(X=i) log\{P(X=i)\}$. For example, a fair coin toss carries a high entropy, as the outcome of the toss is random. Hence, the outcome is uncertain and surprising. Knowing that the probability is 50% does not help to better predict the outcome of the coin toss. In contrast, when flipping an unfair coin, the entropy is much lower, as the outcome is not random anymore. Therefore, the best decision tree split leads to the largest improvement of purity, which is equivalent to a reduction in entropy. The information gain is a measure of how much information a feature provides about the outcome. When the information gain is maximized, the entropy is decreased the most. For instance, the CART decision tree in Breiman et al. (1984) uses Gini impurity to maximize the information gain from splitting the tree: $I_G = 1 - \sum_{l \in \{1,\dots,N\}} p_l^2$. The highest impurity is 1. The squared probability of all individual outcomes is subtracted from this value of 1. Thus, the impurity is close to 1 if there are many individual outcomes with a low probability, making the actual outcome unpredictable. The information gain in a decision tree is the difference between the entropy of the parent node and the average entropy of the child nodes. This means that a good decision tree will result in pure terminal nodes with low entropy.

A simple way of splitting the tree is to use a greedy algorithm that divides each node using

a split that locally minimizes the impurity. Yet, a greedy strategy may be suboptimal. In fact, lookahead search could give better results, as the goodness of the split can also be assessed by evaluating deeper splits. However, this would come at the expense of higher computational power. Similarly, a deeper decision tree is not necessarily a better one. If the tree is shallow (with few leaves), there is probably a high bias due to underfitting. This makes a very deep tree seem optimal at first. However, as with most machine learning algorithms, too much model complexity is likely to lead to overfitting. Theoretically, the tree could have as many leaves as data points, minimizing the error to 0. However, this tree would not be very useful for predicting new data, as the splitting rules would no longer be generally applicable. This means that there may be a point where improving training estimates – by reducing their error – does not further improve the test estimates. In fact, excessive complexity may worsen the test estimates. More specifically, the model should still be generalizable to some degree to deal with unseen data. To avoid overfitting, there are stopping criteria. The underlying idea is to find a good compromise between a tree that is neither too deep nor too shallow. For instance, a stopping criterion could be a maximal depth of a tree, a minimum size of a node, or a minimum required decrease in impurity for a split to be conducted.

3.4 Random forests

For reducing the generalization error, *ensemble methods* can be used. By introducing random perturbations into the learning process, several different models can be built based on a single learning set $\mathcal{L}$. In a second step, the combined predictions of the individual models form the prediction of the ensemble. A simple method of combining them is to take the average. Such an ensemble prediction has the advantage that the variance of the prediction is much lower compared to the prediction of a single tree if the individual trees differ, which in turn lowers the generalization error.

Random forests are a family of methods that make use of an ensemble of decision trees, that is also called *forest*. In the case of decision trees, this is very effective because they often have a high variance and low bias. Therefore, this high variance can be lowered by using a forest instead of a single tree. Random forest methods differ in the way they introduce random perturbations into the induction procedure – the construction of a decision tree.

To explain how they ensure that individual trees differ, I must first introduce two concepts: *bootstrapping* and *bagging*. Bootstrapping is a statistical sampling technique for estimating quantities such as descriptive statistics (e.g., mean, standard deviation) from a data sample. First, many random sub-samples are created from the initial sample with replacement and as many observations as the original sample. For instance, if the original sample consists of 100 cat bond observations, there could now be 1000 sub-samples, each containing 100 random observations from the original sample. Due to the replacement, the sub-samples usually contain many duplicates of the observations of the original dataset. Hence, the so called *out-of-bag* dataset for each of the bootstrap sub-samples consists of the observations of the original dataset that are not in the bootstrap sub-sample. Second, the needed descriptive statistic is calculated for each of the bootstrap sub-samples. For instance, if I am interested in the mean of a sample, the mean of all sub-samples would be computed. Finally, all individual statistics from the sub-samples are aggregated again. A simple way to do this is to take the mean of all sub-sample statistics. This bootstrapping procedure can greatly improve the estimation of a statistic.

Bagging – **b**ootstrap **agg**regat**ing** – is an ensemble algorithm used to reduce variance and

avoid overfitting. In the case of decision trees, sub-samples with replacement are again generated. Based on each sub-sample, a decision tree is generated. Finally, the results of the individual decision trees are aggregated. Again, a simple way of doing so would be to take the average of the individual results. One additional source of variance stems from the similarity of the individual decision trees although their samples already differ. This is directly related to feature importance. Even if the samples differ, the same or similar features are likely to decide the splits of the trees. Algorithms are usually greedy and try to minimize errors (without being forward looking). By limiting the known features for each tree and selecting different known features for each tree, more different sub-trees are generated due to different splitting criteria. The trees are increasingly split based on "weak learners", i.e., predictors that are not very strong, instead of "strong learners" that may not be known. This mechanism decorrelates the trees, reducing the variance of the bagged estimator. Averaging many rather uncorrelated quantities leads to a larger reduction in variance than averaging many strongly correlated quantities. This is demonstrated in Equation 3.5. The aggregated variance is decreased when the covariance term (almost) vanishes. This is only the case when the trees contain many weak learners, so that the trees tend to be uncorrelated. If the trees contain mostly strong learners, they are very similar, which increases the covariance term.

$$Var\{\hat{f}_{Bag}(x)\} = Var\{\frac{1}{B}\sum_{b=1}^{B}\hat{f}_b(x)\} = \frac{1}{B^2}\left[\sum_{b=1}^{B} Var\{\hat{f}_b(x)\}\sum_{c\neq d} Cov\{\hat{f}_c(x), \hat{f}_d(x)\}\right] \tag{3.5}$$

Random forests are particularly effective in settings with many features that are unrelated to the outcome, i.e., settings with sparsity. Since the splits generally ignore unrelated covariates, the performance remains strong in such settings.

3.5 Causal random forests

In the machine learning literature, the focus has been on out-of-sample performance as the criterion of interest (Athey & Imbens, 2019). This has been at the expense of the ability to perform inference, which has been a major focus in the statistics and econometrics literature. A typical application of inference is the construction of confidence intervals that are valid. Recently, the development of methods for inference has made substantial progress for low-dimensional functions in specific settings. Random forests have especially profited from a range of novel literature (e.g., Wager and Athey (2018), Athey and Imbens (2016), Athey et al. (2019)). Generalized random forests offer new methods for three statistical tasks: nonparametric quantile regression, conditional average partial effect estimation, and heterogeneous treatment effect estimation via instrumental variables. In generalizing random forests, many core elements of Breiman's forest are preserved (Breiman, 2001), including recursive partitioning, subsampling, and random split selection, but the final estimate is not obtained by averaging estimates from each member of an ensemble of trees. Instead, the forests are treated as a type of adaptive nearest neighbor estimator. Doing so opens many beneficial statistical extensions.

Motivation Before explaining general random forests, I need to explain why "normal" random forests are sometimes insufficient. When using a multiple linear regression, interaction

terms between variables of interest can be leveraged to gain information about heterogeneous treatment effects and the interconnectedness of different variables. For instance, if I am interested in the treatment indicator w and how it is connected to another variable x_1, I could use the regression $Y = \beta_0 + \beta_1 w + \beta_2 x_1 + \beta_3 (w * x_1)$ to find out. In this example, the treatment effect would be $\beta_1 + \beta_3 \times x_1$. Depending on the value of x_1, the treatment effect could be heterogeneous between different observations. However, when there are many variables, the number of possible interactions quickly increases to a number at which the linear model suffers from low statistical power. Furthermore, the linear model only allows for linear relationships, unless additional polynomials are added, which again reduces statistical power. Hence, a regression is not an optimal option for finding complex relationships. This is the reason why random forests are increasingly used for estimating heterogeneous effects, computing quantile regressions, etc. However, a "traditional" random forest is optimized for minimizing the mean squared error of the outcome variable. In the case of causal random forests that enable to draw causal conclusions, this is not the best way to optimize. Instead, Wager and Athey (2018) add two additional features to a "traditional" random forest to adapt it to causal purposes: a different **splitting criterion** and **honesty**.

Random forests choose splits for the variable and value at each tree node such that the greatest reduction in the mean squared error with respect to the outcomes Y is achieved. In contrast, causal random forests adjust the splitting criterion by searching for a partitioning where the treatment effects differ the most. In addition, it accounts and corrects for how the splits affect the variance of the parameter estimates.

An honest tree has a high accuracy, namely a bias that asymptotically disappears, low standard errors, and low confidence intervals. This leads to consistent estimates and valid confidence intervals. To make a tree honest, the training data is split into two subsamples. One is a splitting subsample used to perform the splits. The other one is an estimating subsample used to make the predictions. More precisely, first a tree is grown using the splitting subsample. Then, all observations from the estimating subsample are dropped down the previously grown tree until they fall into terminal nodes. Ultimately, the prediction of the treatment effects is determined by the delta of the average outcomes between the treated and the untreated observations of the estimating subsample in the terminal nodes.

Methodology To explain general random forests, some notation is needed. The notation is very general, making it applicable for many different statistical methods. Random forests is only one of them. This also leads to an abstract notation. But I will later explain the derivation of the estimator for the special case of regression trees. I assume to have data $(X_i, O_i) \in \mathcal{X} \times \mathcal{O}$, for $i = 1, \ldots, n$. Here, X_i denotes the covariates used to predict the quantity of interest $\theta(x)$, and O_i denotes the observable quantities encoding relevant information for predicting $\theta(x)$. The quantity of interest $\theta(x)$ is defined by a so-called "local estimating equation" presented in Equation 3.6. In this equation, $\psi(\cdot)$ is a scoring function and $\nu(x)$ is an optional nuisance parameter. The equation can be used in most cases to determine the maximum likelihood parameters $(\theta(x), \nu(x))$. Additionally, conditional means, quantiles, and average partial effects can be identified, adding valuable information to predictions derived from random forests.

$$\mathbb{E}\left[\psi_{\theta(x),\nu(x)}(O_i) \mid X_i = x\right] = 0 \quad \text{for all } x \in \mathcal{X} \tag{3.6}$$

Breiman (2001) make a prediction for a point x by pushing x down each of a certain number of trees until it ends up in a terminal node. The prediction from each tree b is then $\hat{\mu}_b$.

The random forest's final prediction for x is the mean prediction of all trees: $\frac{1}{B}\sum_{b=1}^{B}\hat{\mu}_b$. In contrast, for generalized random forests, one observes which training examples fall into the same terminal node as x for each tree (Athey et al., 2019). $L_b(x)$ denotes the set of training examples falling into the same terminal node as x for each tree T_b. The following two equations describe how the similarity of a training example X_i and the prediction x is determined.

$$\alpha_{bi}(x) = \frac{1\{X_i \in L_b(x)\}}{|L_b(x)|}, \qquad \alpha_i(x) = \frac{1}{B}\sum_{b=1}^{B}\alpha_{bi}(x).$$

For each training example $i = 1, \ldots, n$ and each tree b, the function $\alpha_{bi}(x)$ is computed. This function analysis whether the training point X_i falls into the same terminal node as x. If X_i and x are in the same terminal node, $\alpha_{bi}(x)$ becomes $\frac{1}{|L_b(x)|}$. If not, it becomes 0. Since one is interested in a global measure for the similarity of a training sample X_i and the prediction x, $\alpha_i(x)$ is a weighting function that aggregates over all trees B. The average $\alpha_{bi}(x)$ over all trees is computed. This measure $\alpha_i(x)$ increases towards 1 with the similarity of X_i and x and decreases towards 0 when they (almost) never end up in the same terminal node. If they are similar, the training sample X_i should receive a bigger weight when predicting at x. Hence, $\alpha_i(x)$ can be regarded a weighting function. Figure 3.2 is an illustration of the weighting function. The final weighting of each training example i, $\alpha_i(x)$, represented by the size of the circle (per observation) in the bottom row, is the average of the weights of all the individual decision trees $\alpha_{bi}(x)$, as shown in the top row. The idea is that similar observations should often end up in the same terminal node as the quantity of interest x.

After determining the $\alpha_i(x)$, the estimator $\hat{\theta}$ is determined by solving the minimization problem Equation 3.7.

$$\left(\hat{\theta}(x), \hat{\nu}(x)\right) \in \text{argmin}_{\theta,\nu} \quad \left\| \sum_{i=1}^{n} \alpha_i(x)\psi_{\theta,\nu}(O_i) \right\|_2 \tag{3.7}$$

Generalized random forests for random forests "Normal" random forests (Breiman, 2001) are a special case of generalized random forests (Athey et al., 2019). In the following, I show that the estimator derived from the generalized random forest equals the one from Breiman (2001). Athey et al. (2019) just provide a more general solution of the special case in Breiman (2001). The notation simplifies in this special case because the observable quantity equals the response of interest: $O_i = Y_i \in \mathbb{R}$. I am interested in estimating the conditional mean function of x defined by the following.

$$\theta(x) = \mu(x) = \text{argmin}_{\mu} \quad \mathbb{E}\left[(Y_i - \mu)^2 \mid X_i = x\right]$$

After differentiating with respect to μ and setting this equal to 0 to arrive at the optimum of the minimization problem, the problem becomes the following.

$$\begin{aligned} 0 &= \mathbb{E}[Y_i - \mu(x) \mid X_i = x] \\ &= \mathbb{E}[\psi_{\mu(x)}(Y_i) \mid X_i = x] \end{aligned}$$

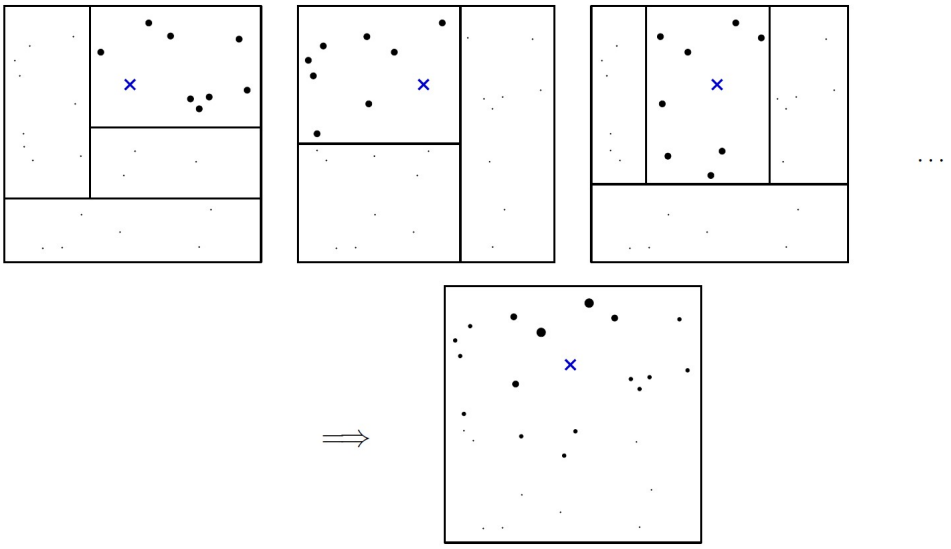


Figure 3.2: This graphic illustrates the random forest weighting function. Each square in the top row corresponds to a decision tree and the small rectangles inside them correspond to terminal nodes of the tree. To find a good prediction for the test point of interest x, shown as a blue cross, the other observations (or circles here) are weighted. For each tree (or square here), the observations that fall into the same terminal node as x are weighted by 1 (and are large in the graphic), the others by 0 (and are small in the graphic). Then, the final weighting, as depicted in the bottom row, is the average of all tree-based weightings of the top row. Thus, if an observation often falls into the same terminal node as x, it is weighted more heavily than an observation that falls less frequently into the same node. The final weighting is emphasized by the size of the circles. This illustration is taken from Athey et al. (2019).

In this equation, $\psi_{\mu(x)}(Y_i) = Y_i - \mu(x)$ is the same form as Equation 3.6. This means that it is now possible to plug it into Equation 3.7. Doing this, taking the derivative and then solving for the solution of the minimization problem, while considering that the sum of the weights of the training examples is 1 ($\sum_{i=1}^{n} \alpha_i(x) = 1$), I arrive at $\hat{\mu}(x) = \hat{\mu}_b(x)$. Importantly, this is exactly the same estimator as derived from "normal" random forests (Breiman, 2001).

$$\begin{aligned} \hat{\mu}(x) &= \text{argmin}_\mu \quad \left\| \sum_{i=1}^{n} \alpha_i(x) \psi_\mu(Y_i) \right\|_2 \\ &= \text{argmin}_\mu \quad \left(\sum_{i=1}^{n} \alpha_i(x)(Y_i - \mu) \right)^2 \end{aligned}$$

$$\Rightarrow \quad \sum_{i=1}^{n} \alpha_i(x)(Y_i - \hat{\mu}(x)) = 0$$

$$\begin{aligned}
\hat{\mu}(x) &= \sum_{i=1}^{n} \alpha_i(x) Y_i \\
&= \sum_{i=1}^{n} \sum_{b=1}^{B} \frac{1}{B} \frac{1\{X_i \in L_b(x)\}}{|L_b(x)|} Y_i \\
&= \frac{1}{B} \sum_{b=1}^{B} \frac{Y_i 1\{X_i \in L_b(x)\}}{|L_b(x)|} \\
&= \frac{1}{B} \sum_{b=1}^{B} \hat{\mu}_b(x)
\end{aligned}$$

3.6 Double machine learning

The previous section on generalized random forests sounds very compelling. However, it relies on data from a randomized control trial, as the assignment of the treatment must be random. For most practical purposes, however, this is not the case. In fact, it is often impossible to perform experiments and historical observational data is usually also non-experimental. If the assignment of the treatment is not random, a common problem is the correlation of confounders with both the analyzed covariate (the treatment) and the outcome variable. This is the reason why double machine learning is additionally needed when making use of general random forests.

Typical approaches for correcting undesired correlations are the use of instrumental variables and partial regressions. These concepts can also be applied to machine learning. When supervised machine learning is used to learn the functions, it tends to lead to overfitting and regularization biases. Double machine learning, also called orthogonal machine learning, attempts to correct both biases (Chernozhukov et al., 2018). The "double" comes from the use of primary and auxiliary predictive models. By training two models, causal inference is possible in supervised learning.

I now briefly summarize the overall process of double machine learning. First, the data are divided into two equal sets through sample splitting. This should eliminate the problems of overfitting later. One model is trained to predict the treatment from confounders. The treatment $\tilde{T}$ residuals are calculated based on this model. The other model is trained to predict the target from the confounders. The outcome $\tilde{Y}$ residuals are computed based on this model. This separation by having two models fixes the problem of a potential regularization bias. As a final step, the previously predicted target $\tilde{Y}$ (in this context the premium spread) is regressed on its predicted treatment $\tilde{T}$. This allows to obtain the overall treatment effect, as depicted in Equation 3.8. In the case of random forests, this process should lead to desirable properties such as asymptotic normality of the machine learning estimators (Athey et al., 2019).

$$\tilde{Y} = \theta(X) \times \tilde{T} + \epsilon \tag{3.8}$$

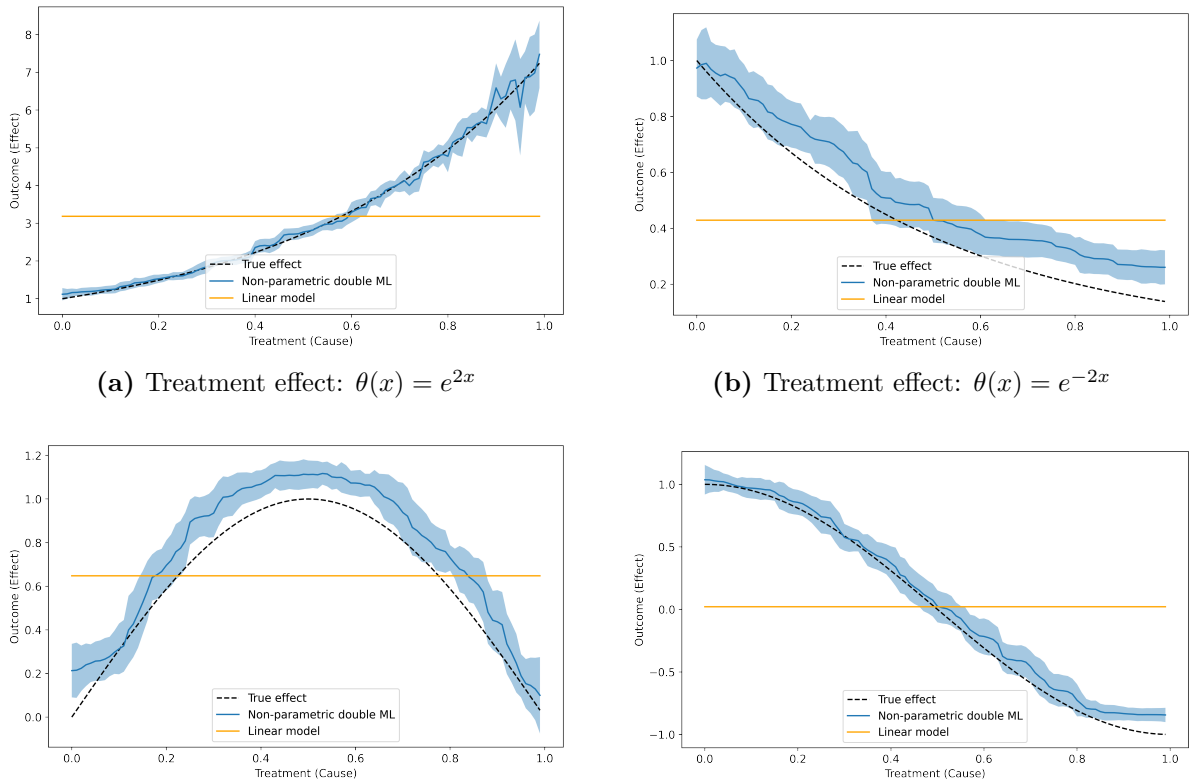
To illustrate how causal effects can be detected and disentangled, I analyze synthetic datasets. My synthetic dataset consists of 20 features with 20 thousand rows of data. The synthetic

data follow the partially linear regression model. In this model, the outcome is defined by $Y = T\theta + g(X) + \epsilon$. Here, θ is the causal parameter that I am interested in determining. However, the feature of interest T depends on X: $T = m(X) + \epsilon$. In these equations, ϵ are the irreducible error contributions. And g and m are the nuisance functions. The first equation is the main equation. By also having the second equation, one can correct for an omitted-variable bias.

In the synthetic dataset:

- Y is the target
- T is the feature of interest, also called treatment
- X are the control vectors $\sim \mathcal{U}(0, 1)$, that one needs to control for
- W is a matrix of the confounders $\sim \mathcal{N}(0, 1)$

In Figure 3.3, I test four different treatment effects. Double machine learning can reveal them all.



(a) Treatment effect: $\theta(x) = e^{2x}$

(b) Treatment effect: $\theta(x) = e^{-2x}$

(c) Treatment effect: $\theta(x) = sin(\pi x)$

(d) Treatment effect: $\theta(x) = cos(\pi x)$

Figure 3.3: These simulations with synthetic data demonstrate that heterogeneous treatment effects can be identified by using double machine learning. A linear model leads to an incorrectly estimated treatment effect. For example, in the case of (d), the estimated effect is around 0. However, in reality, the treatment effect ranges from 1 to -1, depending on the value of the control variable X.

3.7 Interpretability tests: Shapley values

The Shapley value is an idea from the field of cooperative game theory (Shapley, 1953). Players cooperate in a coalition and receive payoffs depending on their contribution to the total payoff. This idea is transferable to machine learning predictions. Now, the "game" becomes the prediction task for a single instance of the dataset, the "payoff" becomes the difference between the actual prediction for a specific instance and the average prediction for all instances, and the "players" become the cooperating feature values of the instance (Molnar, 2022).

The interpretation of Shapley values sounds a bit abstract at first. Given a current set of feature values, the estimated Shapley value is the contribution of a feature value to the difference between the actual prediction and the mean prediction. To obtain a Shapley value for a specific feature, the marginal effect of this feature for all possible coalitions of features must be observed. The Shapley value is then the average marginal effect. For instance, I may want to predict cat bonds spreads with three characteristics: expected loss, trigger type, and sponsor. In this example, I already know that the spread is 5% for an expected loss of 10%, an indemnity trigger, and Swiss Re. I also know that the spread is 6% for an expected loss of 10%, a different trigger, and Swiss Re. Both bonds have the same features except for the trigger type (here a dummy). Now I know that the additional 1% of spread for the second bond must be due to the different trigger type. But this 1% is not the average effect of a bond with no indemnity trigger. Therefore, I would have to repeat this computation for all possible coalitions to finally arrive at the correct estimation for a bond with no indemnity trigger. In this example, there are only three features, but there are already plenty of possible coalitions because the expected loss and sponsor can have many different values.

Importantly, although Shapley values are great for interpretation, they do not serve prediction purposes. They do, however, allow for contrastive explanations. For instance, different subsets of data can be compared, or a single data point can be compared to the whole dataset or a subset. In this way, the effect of certain features becomes more understandable. When analyzing cat bonds, this becomes a powerful tool. Figure 3.4 illustrates why the predicted spread of one particular bond is 13.28%. For the entire training set, the average spread is around 7.5%, which is described as the base value. The red features drive the predicted spread to a value above the average, while the blue features reduce the predicted spread. In addition, the feature arrows indicate how much each feature value affects the predicted spread. In the example, the expected loss of around 5.6%, which is well above the average expected loss of around 2.5%, is the strongest feature causing most of the upward shift in the predicted premium.

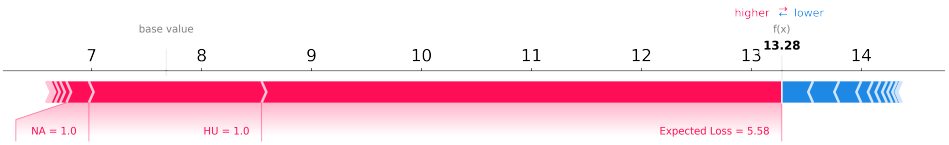


Figure 3.4: Shapley value example for one cat bond compared to the whole dataset. This bond's predicted spread is higher than for the "average bond" due to its high expected loss.

4 A causal random forest approach to analyzing the impact of expected loss over time

This chapter contains my analysis of the premium in the primary cat bond market. In particular, I analyze whether cat bond specific and macroeconomic variables have a changing effect on the premium over time. To do so, I make use of the previously introduced methodology of chapter 3, that includes causal random forest approaches, Shapley values, and double machine learning. In contrast to recent papers such as Makariou et al. (2021), my approach focuses on causality rather than spread prediction. My aim is to shed light into the machine learning algorithms to make their predictions explainable. For pure predictions, the models can remain "black boxes" since the main goal is accurate prediction. In contrast, I am not trying to achieve high accuracy, but to quantify the effects and their uncertainty.

4.1 Data

I use a data collection on 736 cat bonds issued between March 2002 and March 2021 for which premium at issue, secondary market prices and all explanatory variables are available. Observations with missing or implausible data are excluded from this dataset. The first half of the data is from Gürtler et al. (2016) and contains 332 bonds after cleaning the data. This dataset contains cat bonds issued between March 2002 and March 2012. For the remaining time period between April 2012 and March 2021, I collected data for additional 404 cat bonds after cleaning the dataset. The premiums – the yield spreads over the LIBOR – form the dependent variable in my analysis. I extracted these premiums from the annual reports of Lane Financial LLC[1].

The explanatory variables are bond-specific or describe the macroeconomic state at the time when the bond was issued or when the secondary market prices were recorded. The set of variables included in this empirical study is based on recent papers on cat bond pricing (Götze and Gürtler (2020), Braun (2016), Gürtler et al. (2016)). I use the Artemis Deal Directory[2] and Lane Financial LLC for the collection of bond specific data. Data on the trigger mechanism, bond issuance volume, insured peril types, and peril locations are obtained from the Artemis Deal Directory. The issuance volume is a potential proxy for a bond's liquidity. The trigger type is critical because it exposes the investor to an additional source of risk. For instance, indemnity-trigger cat bonds potentially evoke ex-ante or ex-post moral hazard. Settling claims is a costly activity for the insurer. As its benefit is partly borne by the investors when using an indemnity trigger, this might create a conflict of interest between the bond's sponsoring insurer

[1] The annual reports are available on the website of Lane Financial LLC: http://www.lanefinancialllc.com/content/blogcategory/41/67/ (retrieved on 10/10/2022).

[2] The deal directory can be found on the following website: https://www.artemis.bm/deal-directory/ (retrieved on 10/10/2022).

and its investors. Namely, the sponsor's loss adjustment policy may become laxer, which may affect either the probability of loss before or after an insured event (Götze & Gürtler, 2020). Regarding the trigger type, I only distinguish between indemnity and non-indemnity (like Götze and Gürtler (2020)) to reduce the complexity and increase the sample size per group. A high complexity through multi-peril and multi-location bonds may also be reflected by a higher spread as compensation. Similarly, locations that are considered peak territories such as the US as well as peak perils such as hurricane cat bonds could be assumed to require an additional compensation.

The other bond-specific data is obtained from Lane Financial LLC, including the bonds' sponsors, the bonds' S&P ratings, the maturity and exposure term, the expected loss (EL), the probability of first loss (PFL), the probability of exhaust (PLL), and the conditional expected loss (CEL). Since the EL is defined as the first moment of the principal loss distribution of cat bonds, a higher EL should convert into a higher spread. The EL is widely assumed to be the most important price-determining factor. The PFL directly reflects event risk, while the CEL captures downside risk. Previous research has shown that some sponsors such as Swiss Re may achieve better conditions due to their market leading position (Braun, 2016). The information about the sponsor could be easily included through a dummy variable for each sponsor. Since there are many different sponsors, I decide to include only the type of sponsor: "reinsurer", "insurer", and "other". This results in a larger sample size per category.

The macroeconomic variables are included to capture market developments. All necessary data for these variables are extracted from Refinitiv and Guy Carpenter. First, the development of equity markets could influence cat bond prices, or the underlying price-driving factors could be similar. Hence, I include the monthly return of the S&P500 as an indicator for the development of equity markets. For doing so, I obtain the monthly S&P500 closing prices from Refinitiv[3]. Second, cat bond prices could co-move with prices of more traditional reinsurance products because cat bonds are a potential substitute (Braun (2016), Gürtler et al. (2016)). Therefore, I include a proxy for the reinsurance price cycle, namely the annual relative change in the Guy Carpenter Global Property Catastrophe Rate-On-Line Reinsurance Price Index[4] (Carpenter, 2012). As the general development of prices in the cat bond market should also be accounted for, I include the "Swiss Re Global Cat Bond - Total Return Index". This catastrophe bond market index calculated by Swiss Re Capital Markets should reflect the returns of the catastrophe bond market. I calculate the relative change in that index on a monthly basis. For the date of interest, this monthly change is considered. The monthly data used is collected from Refinitiv Datastream. Additionally, I include a monthly corporate credit spread. This spread captures the difference between the yields on corporate bonds and government bonds. The data is obtained from FRED[5]. Previous papers such as Götze and Gürtler (2020) observed the spread for each rating class. As most of the cat bonds in my dataset do not have an S&P rating, this does not seem to be a practical solution for my dataset.

Table 4.1 presents the summary statistics for all cardinal cat bond specific and macroeconomic variables. Since I analyze whether factors have a changing effect on the premium over time, I split my datasample into three time periods, which each contain five years of data. Their sample size is fairly similar, ranging from 151 to 188 bonds. When exploring whether

[3] I retrieved the data on 12/10/2022.

[4] I obtain the data on the Guy Carpenter Global Property Catastrophe Rate-On-Line Index from https://www.artemis.bm/global-property-cat-rate-on-line-index/ (retrieved on 11/10/2022).

[5] ICE BofA US Corporate Index Option-Adjusted Spread from https://fred.stlouisfed.org (retrieved on 11/10/2022).

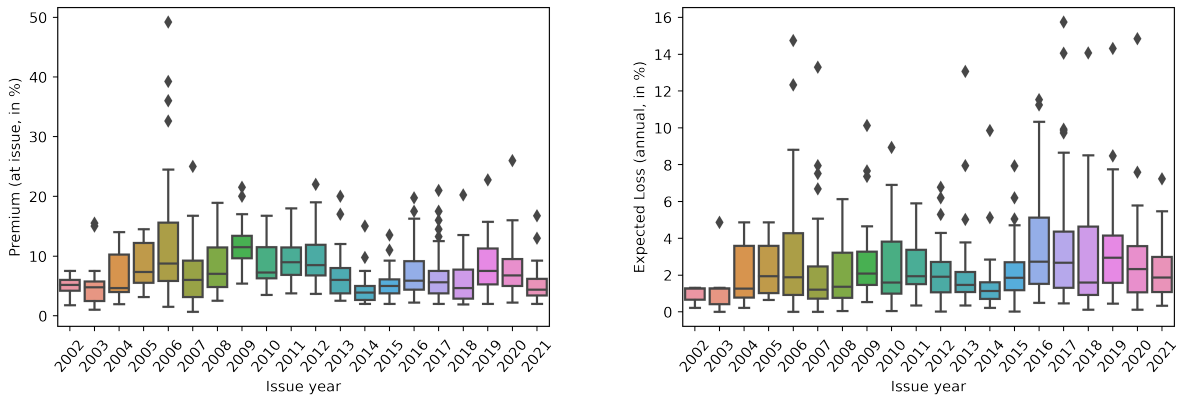
certain factors had a changing influence on the premium, I either investigate effects in these periods or on a yearly basis. As already described in part I of my thesis, my overall statistics for the whole time period from 2002 to 2021 are very similar to those in Götze and Gürtler (2020). I reference to part I of my thesis for additional information. After analyzing the dataset including the previously mentioned PFL, PLL and CEL, I decided to remove them. They do not seem to provide much additional information. When including them in a linear regression on the spread in addition to the other features, none of these risk measures was significant at a 10% level. Additionally, these variables contain similar information as the EL. A high correlation between the EL and those variables may worsen the interpretation of the impact of the EL on the spread later. When looking at the three time periods in particular, the mean and standard deviation fluctuate over time. Interestingly, the premium decreases from 9.6% in the first period (2006 to 2010) to 7.5% in the second period (2011-2015) and then again to 6.9% in the most recent time period (2016-2020). Although the premium was the lowest in the most recent time period, the expected loss was the highest at 3.4%. In contrast to the period from 2006-2010, this is almost an increase of one percentage point. If the expected loss is indeed the most decisive factor influencing the premium, the question arises whether the risk is not rewarded to the same extent anymore.

Table 4.1: Summary statistics: cardinal cat bond specific and macroeconomic variables (reported at the issue level)

	2006-2010			**2011-2015**			**2016-2020**		
	Obs.	Mean	SD	Obs.	Mean	SD	Obs.	Mean	SD
Cat bond specific variables									
Premium (%)	171	9.64	7.27	151	7.51	4.24	188	6.93	4.21
Expected Loss (ann., %)	171	2.56	2.70	151	2.05	1.78	188	3.40	2.98
Volume (USD million)	171	103.45	91.01	151	167.03	153.47	188	165.79	119.71
Maturity (months)	171	30.02	12.10	151	40.36	8.65	188	40.91	10.87
No. perils	171	1.61	0.86	151	1.74	0.68	188	1.95	0.73
No. locations	171	1.41	0.91	151	1.20	0.42	188	1.21	0.57
Macroeconomic variables									
Reins. Index (ann. change, %)	171	8.90	20.38	151	-1.57	7.98	188	-1.23	5.12
S&P500 (mthly. change, %)	171	0.08	3.60	151	0.98	2.77	188	0.75	2.88
Corp. spread (%)	171	1.78	1.18	151	1.67	0.43	188	1.24	0.22
CAT bond Index (mthly. change, %)	171	0.81	0.43	151	0.46	0.76	188	0.36	0.58

When examining a boxplot of the premiums and expected loss per year, a clear connection of these two variables becomes visible (Figure 4.1). Whenever the expected loss increases over time, the premiums are also very likely to do so. Similar to other financial products, there must be a risk compensation in form of the premium. For a more detailed analysis of the correlations of the analyzed variables, have a look at part I of my thesis.

Table 4.2 presents the summary statistics for all nominal and ordinal cat bond specific variables. As before, I analyze the variables for the three time periods. As mentioned in part I of my thesis, my statistics for all years resemble the ones from Götze and Gürtler (2020). Since the sample size for some ratings is very low, especially the AA, A, and CC rating, coefficients for these variables estimated by any model should be interpreted with great caution. Importantly, a single bond can have multiple peril types and locations. Consequently, neither the peril types nor the peril locations sum up to the amount of cat bonds in the dataset. However, each bond is assigned to only one rating category, one trigger type, and one sponsor type.



(a) Premiums **(b)** Expected loss

Figure 4.1: Distribution of cat bond characteristics aggregated by year: Premiums and expected loss of cat bonds co-move because they are strongly correlated.

Table 4.2: Summary statistics: nominal and ordinal cat bond specific variables (reported at the issue level)

	2006-2010 (%)	2011-2015 (%)	2016-2020 (%)
Trigger			
Indemnity	15.20	50.33	61.17
Non-indemnity	84.80	49.67	38.83
Peril type			
EQ	56.14	57.62	70.74
HU	47.37	55.63	0.53
Wind	28.07	34.44	68.62
Other	15.79	25.83	55.32
Peril location			
EU	29.24	20.53	15.43
JP	21.05	7.28	5.85
NA	73.68	84.11	86.17
Latin America	7.60	5.96	4.26
Asia/Australia	3.51	1.99	9.04
Sponsor			
Reinsurer	64.91	33.11	34.57
Insurer	29.82	63.58	57.45
Other	5.26	3.31	7.98
Rating			
S&P Rating AA	0.58	0.00	0.00
S&P Rating A	1.75	0.00	0.00
S&P Rating BBB	2.34	0.00	0.00
S&P Rating BB	44.44	39.74	1.06
S&P Rating B	22.22	24.50	1.60
S&P Rating CC	0.00	0.66	0.00
S&P No rating	28.65	35.10	97.34

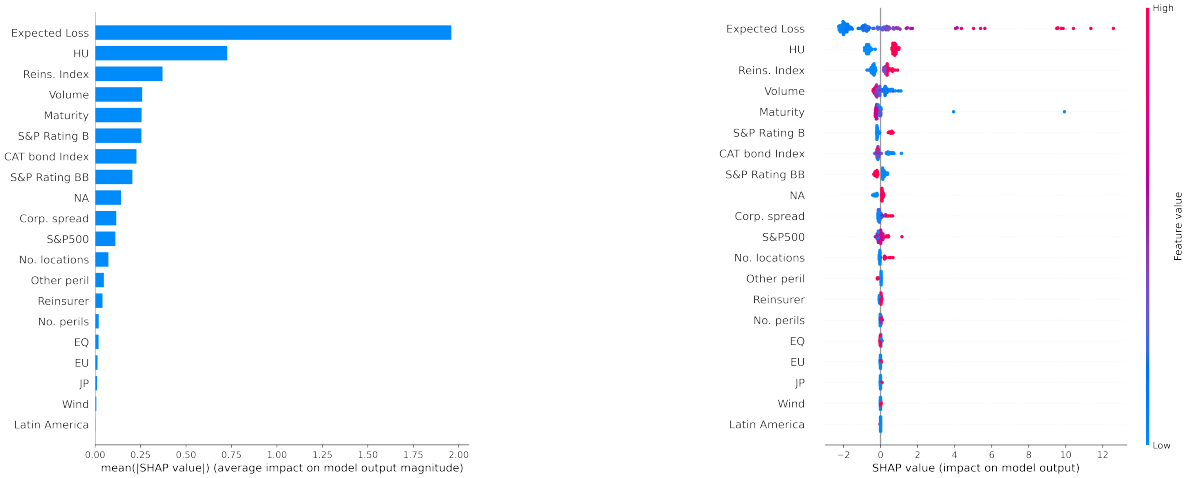
4.2 Random forests – Shapley values

In determining which features are especially helpful at predicting the outcome, the feature importance is of great use. To get a first overview of the effects of the predictors on the

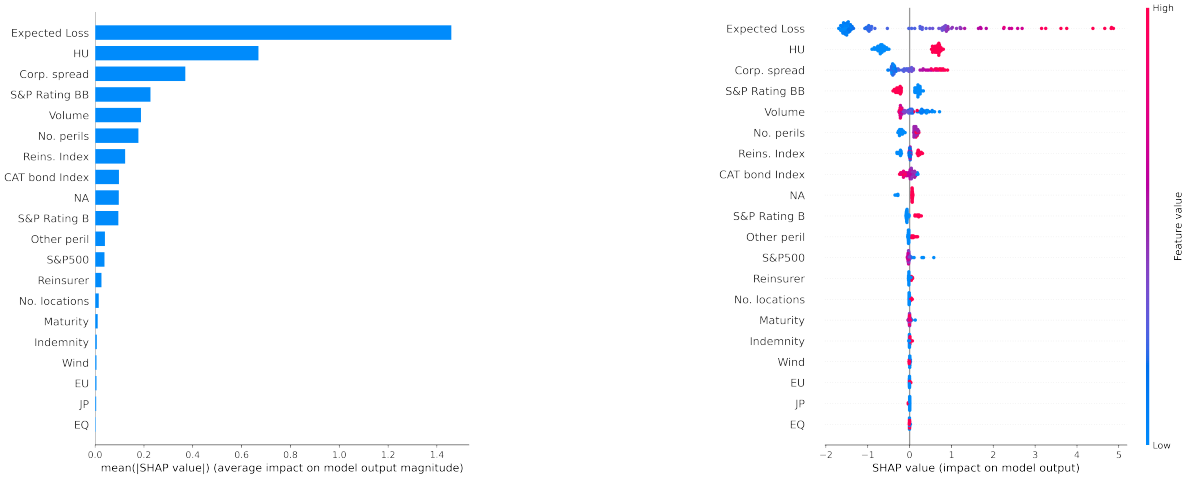
premiums, I determine the Shapley feature importance by training a random forest on 80% of the sample data for all three time periods. Each random forest is formed by 2,000 decision trees, the maximal depth is three and the maximum number of features is 10. The SHAP feature importance is measured as the mean absolute Shapley value. For each feature, the mean of the absolute Shapley values is calculated across all observations. Absolute values are calculated, as I do not want positive and negative values to offset each other. The left graphics of Figure 4.2 show that the expected loss is by far the most important feature in all three time periods. However, its average impact on the premium is around 2 both for the years 2006-2010 and 2016-2020, and only around 1.4 for the years 2011-2015. One of the advantages of Shapley values is the analysis of heterogeneous feature effects on the model output and the analysis of single model outputs. In the beeswarm plots on the right side of Figure 4.2, the values are grouped by the features on the y-axis and ordered by importance, their mean Shapley values. Each point represents a Shapley value for a feature and instance. While the x-axis represents the Shapley value, the feature value is represented by the color of the instance. Most instances of the expected loss have a high EL Shapley value, and the EL Shapley value is especially high for high feature values. Except for the partially very large SHAP values in the period from 2006 to 2010, the distribution of observations looks very similar across the three time periods for the expected loss.

Other important features are the bond maturity, volume size of a bond, whether the bond is exposed to a hurricane peril, the change in the reinsurance index at issuance date, the corporate spread at issuance as well as a few other features. The magnitude and the ranking of the features differs slightly between the three periods. Some changes are interesting. For instance, the hurricane peril does not seem to have an impact for the years 2016-2020 although it has the second largest impact in the other two periods. As I focus on the effect of the expected loss in this part II of my thesis, I do not analyze the other features in more detail.

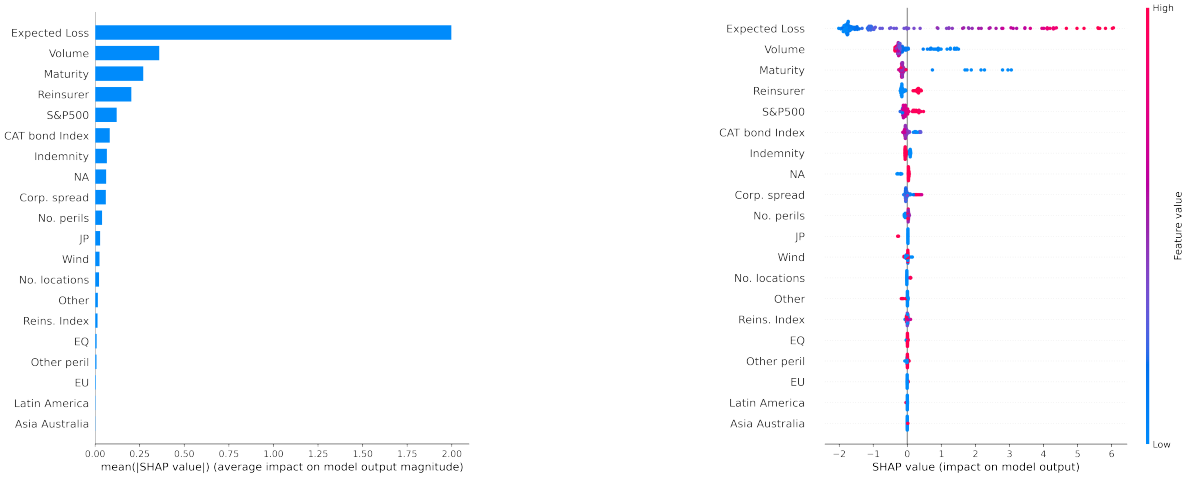
Importantly, the importance scores do not tell me much about the interplay of different features. Additionally, they do not help at assessing why a feature may be better or worse than another feature, as the scores are based purely on correlations. A correlation could be random, or an underlying true factor may be missing from my dataset but correlated with another feature in the dataset. With the help of these plots, a specific output becomes explainable. Nevertheless, all effects only describe the behavior of the model. Thus, these effects are not necessarily causal.



(a) Years 2006-2010



(b) Years 2011-2015



(c) Years 2016-2020

Figure 4.2: SHAP analysis: Mean absolute Shapley values (left side) and beeswarm plots (right side) using data samples from different time periods.

4.3 Causal random forests – An introduction

Random forests are well suited for predictive tasks. To make a good prediction, correlations are sufficient. In fact, these correlations could be purely random. For instance, a true underlying factor could be missing but correlated with another factor that is one of the features. By mistake, this included variable could now be considered the causal factor. In this example, an instrument is missing. This is the reason why A/B tests and randomized control trials are often referred to as the gold standard for causal inference. In such settings, unwanted effects can be avoided to optimally analyze causality. A randomized setting allows to exclude unwanted biases.

Causal random forests are a great method for analyzing treatment effects (or other interesting effects). Because of their honest trees, they lead to asymptotically normally distributed estimators. Therefore, they allow to interpret coefficients and other statistical measures derived from a model. They can be used to determine not only the average treatment effect (ATE) but also the conditional average treatment effect (CATE). This is important because, for example, the effect of the EL on the cat bond spread might be heterogeneous. Indeed, it could be smaller or larger for bonds with a smaller or larger volume. Similarly, the effect of the EL could be time dependent, which is analyzed in this part II of my thesis. The analysis whether an effect is heterogeneous is very crucial because heterogeneity may lead to a misleading ATE that is likely to be misinterpreted. In the context of this part II of my thesis, it could be the case that the effect of the EL differs substantially between the years. For instance, if the effect is in some years positive and in others negative, the individual yearly effects could even cancel each other out. Although this is unlikely, the example still illustrates that my analysis is insightful. Additionally, a quantile regression enables to observe variances around coefficients. I use the *EconML* python package for the causal analysis of one especially relevant factor, the "treatment".

However, causal random forests work best with data derived from an RCT. Unfortunately, in most practical situations, data from such optimal studies are not obtainable. In the case of cat bonds, this is simply not possible. Cat bonds would have to be issued at the same time and differ in only one factor to allow direct causal inference. Moreover, even a large sample size of such similar cat bonds would be needed to rule out randomness. Double machine learning may be a solution to this problem. Fortunately, *double machine learning* is implemented for many methods of the *EconML* package.

Importantly, interpretability relies strongly on the assumption that there is no unobserved confounding. Namely, all important features that significantly influence the outcome variable, the cat bond spread, should be included. If this assumption is not fulfilled and there are unobserved confounders, it is no longer possible to draw interpretive conclusions.

Figure 4.3 illustrates the interplay of all features and how they affect the cat bond spread Y. Both X and W are features used to predict both the outcome Y and the treatment T. However, only the features X are assumed to influence the strength of the relationship between Y and T. More specifically, the assumption is that the treatment effect θ is a function of X but not W. By having this setup, the feature of interest T (also called treatment) can be studied in detail when using causal forests.

In my cat bond dataset:

- Y is the outcome/target: here the cat bond spread
- T is the feature of interest: here the expected loss

- X are all features that may have a heterogeneous effect on T: here the issue date
- W are all remaining features

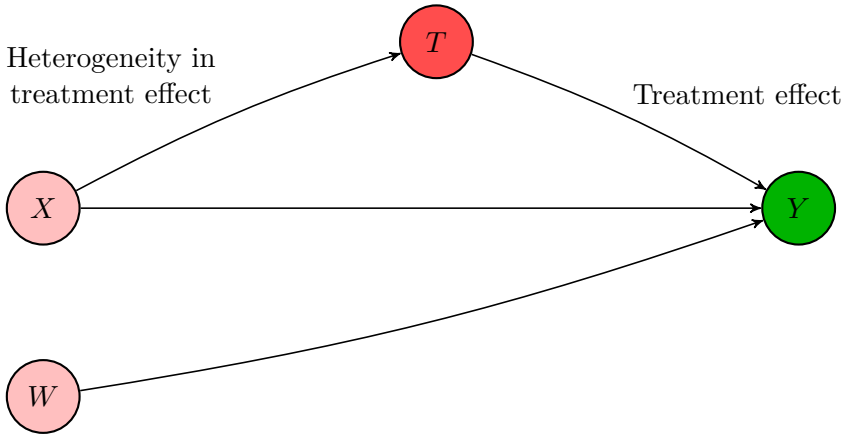


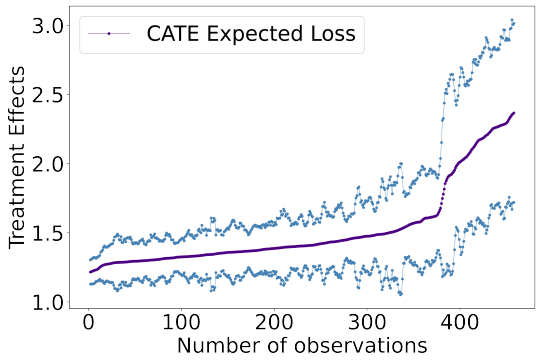
Figure 4.3: Illustration how the "honest" effect of a feature T on the outcome Y, the cat bond spread, can be studied.

4.4 Causal random forests – Conditional average treatment effects

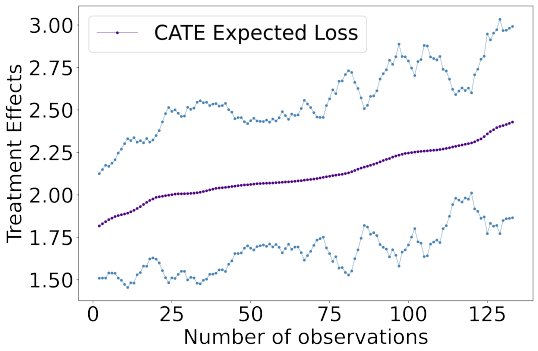
In this section, I analyze the expected loss as the treatment factor. All other features are treated as X, features that may have a heterogeneous effect on T. By doing so, the clean, "honest" effect on the outcome, the premium, may be observable. In order to find time effects, I analyze the effect of the expected loss for a time period. My three subsets of the whole dataset each comprise bonds issued within 5 years. For each subset, I train a causal forest to estimate the effect. Each causal forest is trained on 80% of the data. By using one random forest regressor each first for T (the expected loss) and then also for Y (the premium), also called double machine learning, the "honest" effect of T should become analyzable. The splitting criterion is the mean squared error, 1,000 decision trees are used, and three-fold cross-validation. I then compute the causal effect as well as the confidence intervals for the training and test samples. For the confidence interval, I use $\alpha = 0.05$, meaning that the 95%, two-sided confidence interval is reported. Ideally, the effects should be similar for the test and train sample. This is indeed the fact, as I cannot find a big difference in the predictions between the train and test samples.

Figure 4.4 shows the conditional treatment effects for the three time periods and the whole period from 2006 to 2020, respectively. As I compute the heterogeneous treatment effect, the estimated effects differ within a sample. For interpretive purposes, I sort the treatment effects by their estimated magnitude. The individual treatment effects help assess whether the effect appears to be consistent or heterogeneous. For the whole period from 2006 to 2020, the treatment effect of the EL ranges from around 1.3 to around 2.2. Thus, an additional percentage point in the expected loss would lead to an increase in the premium of around 1.3 to 2.2 percentage points. The confidence interval is around ± 0.2 around this value for low estimates but increases dramatically for high estimates. In contrast, when looking at the three time periods, the estimated effects per time period have different ranges. For an easier assessment of the effects, I computed boxplots of them (Figure 4.5). The estimated effect of the EL on the premium ranges from around 1.8 to 2.4 for the period from 2006 to 2010, while it

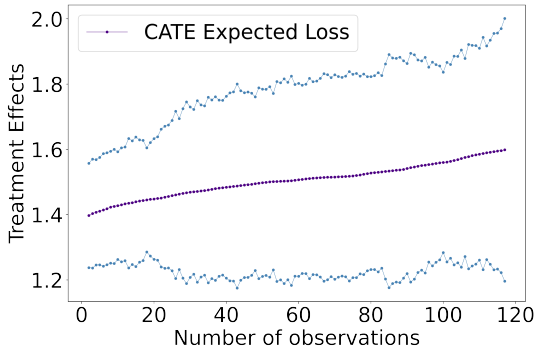
is only around 1.5 for 2011 to 2015, and 1.3 for 2016 to 2020. Overall, the effect seems to have decreased over time. Maybe even more interestingly, the volatility of the effect seems to have decreased as well. The boxes, indicating the interval between the 25% and 75% percentile, shrink enormously over time.



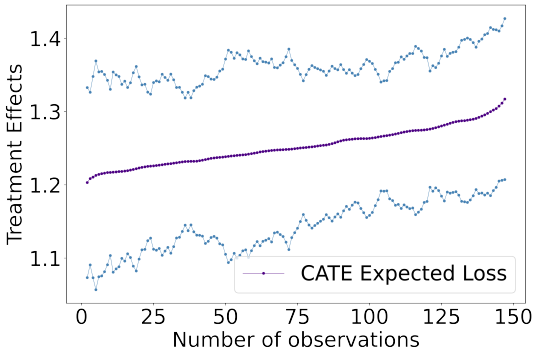
Years 2006-2020



Years 2006-2010



Years 2011-2015



Years 2016-2020

Figure 4.4: CATE of expected loss (in %) for three time periods of 5 years and the whole period from 2006 until 2020.

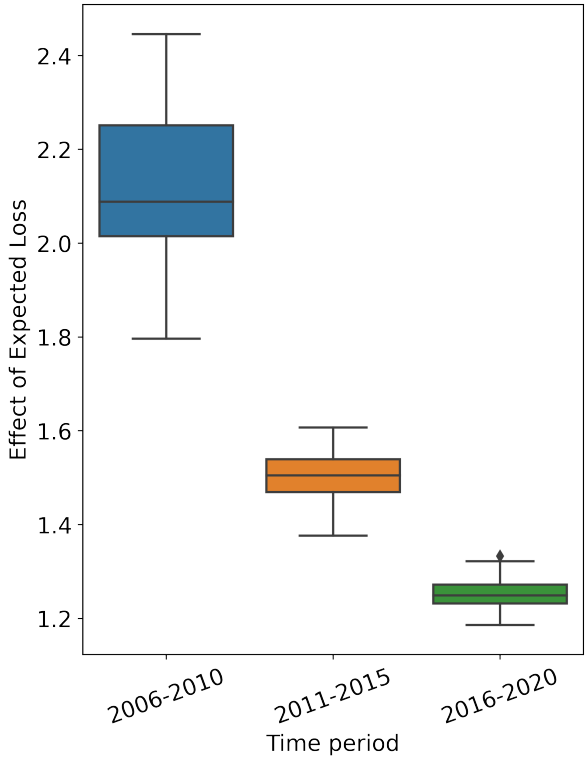


Figure 4.5: Boxplots of CATE of the expected loss on the premium for three time periods of five years.

4.5 Causal random forests – Heterogeneous effects

When analyzing how factors influence the premium, it is useful to determine whether their effect is heterogeneous due to a particular factor. In this context, l define causality as the influence a single continuous feature, the expected loss, has on the outcome, the cat bond spread, while holding all but one other characteristic constant. This one other feature X, that is not kept constant, is assumed to affect the strength of the treatment effect. In this part II of my thesis, X is the year in which a bond is issued. The feature under investigation is denoted T since causal forests are especially popular for analyzing treatment effects. However, this feature does not necessarily need to be a treatment in the classical economic context. The final outcome is denoted Y. Both X and W are used to predict both outcome Y and treatment T. But only the features X are assumed to influence the strength of the relationship between Y and T. Precisely, the assumption is that the treatment effect θ is a function of X but not of W. I could, of course, include all remaining variables in X instead of W, which I do in section 4.4, but there are also cases where the effect is expected to be heterogeneous only with respect to one or a few variables. Hence, the effect of T is assumed to depend on the values of X. This heterogeneity of the treatment effect, or as here, just the effect of the feature under study, is examined in this section.

To make this more understandable, I use the following example. I can analyze the treatment, the effect of the expected loss, for different groups based on X, the year of the issue of the bonds. For instance, I could split my observations into two groups, one with cat bonds issued

between 2002 to 2011 and another with cat bonds issued between 2012 and 2021. The (average) treatment effect of the expected loss could then be expected to differ between the groups. This would lead to additional insights into the heterogeneous effect of the expected loss on the premium. Figure 4.6 illustrates how this would work. If there is indeed a heterogeneity in the treatment effect, the conditional average treatment effect should not be the same for the two subgroups. In this example, I could discover whether the effect of the expected loss on the premium has changed over time.

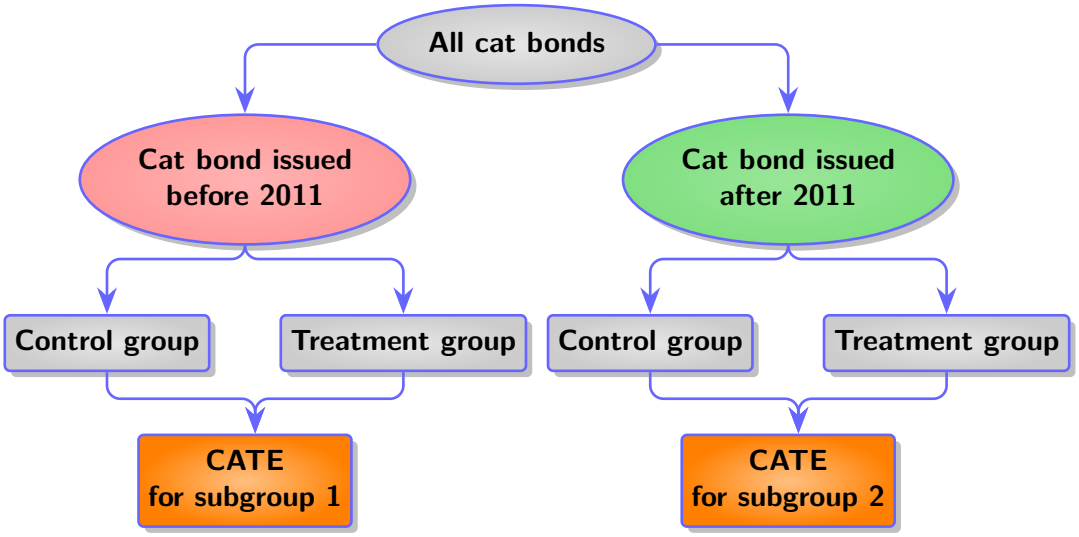


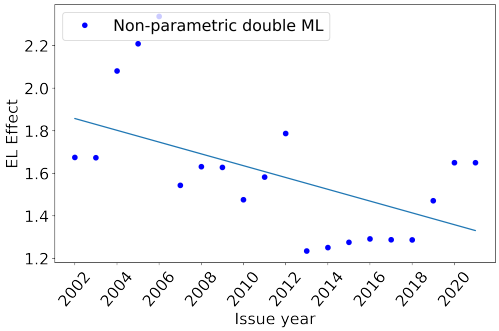
Figure 4.6: An illustration of the analysis of a potential heterogeneous effect of the expected loss on the premium based on the issue date of the bond. The conditional average treatment effect (CATE) could differ between the two subgroups.

I examine whether the issue date of a bond strongly influence the effect of the expected loss on the premium and whether this effect is heterogeneous. I do so by using a causal forest including double machine learning (the *CausalForestDML* method of the *EconML* package). Like before, I train on 80% of the data and test the results on the remaining 20%. The splitting criterion is not the mean squared error, but the heterogeneity in the split. This means that the quality of a split in a decision tree is measured by the heterogeneity score. Since I am mainly interested in a potential heterogeneous effect of the expected loss, this splitting criterion seems more appropriate for this analysis than the mean squared error. The causal forest is formed by 10,000 decision trees, and 10-fold cross-validation should stabilize the results. After training the causal forest model on the train sample, I compute the causal effect for the train and test sample. Ideally, the effects for the test and train sample should be similar. This is indeed the fact, as I cannot find much difference in the predictions between the train and test sample for any of the features.

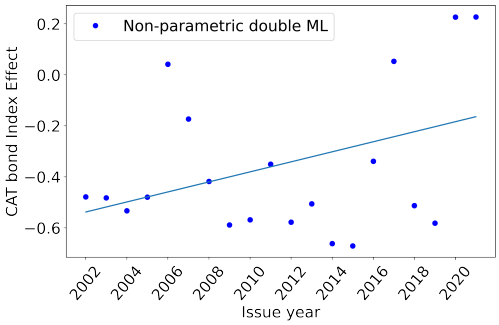
The results for the test sample are plotted in Figure 4.7. In addition to the expected loss, I also analyze the effect of the most interesting other features on the premium in dependence of the issue year in detail. As plotted in graph (a), the effect of the expected loss on the premium seems to have decreased over time. An effect of 1.5 means that one additional percentage point in the expected loss would increase the premium by 1.5 percentage points. In the years 2012 to 2018, the estimated effect is especially low at around 1.3. Nevertheless, the most recent

estimated effects for the years 2019 and 2020 do not fit into the downward trend as they are relatively high at around 1.7.

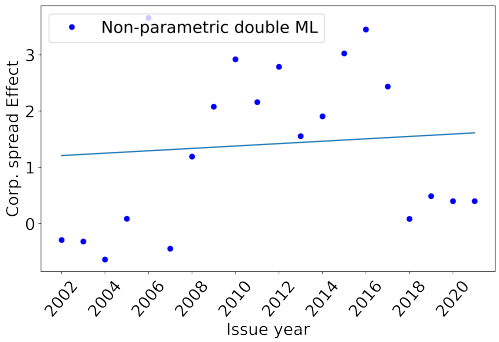
When analyzing the effect of the other features on the premium in dependence of the year, there is no clear trend for most of them. I fit a linear trend to each of the estimated features. The cat bond index seems to have lost its impact over time, as the effect seems to converge to 0. However, there is strong volatility around this "trend". The evolution of the estimated effect of the corporate spread does not seem to follow any particular trend. However, the evolution of the effect looks like an inverted U-shape. From around 2009 to 2016, the effect was at around 2.5, while for the other years it is around 0. The maturity's negative effect appears to have become stronger over time. However, the effects are relatively small. A maturity effect of -0.005 means that, ceteris paribus, a bond with an additional 10 months to maturity would have a premium that is -0.05 percentage points lower. The effect of the change in the reinsurance index seems to have reversed from around 0.3 to around -0.2 in recent years. This is interesting because an estimated effect for the whole time period cannot capture such a reversed effect at the risk of estimating the effect as irrelevant. The change in the S&P500 seems to have lost importance over time as the estimated effect converges to around 0. Similarly, the number of peril locations and types fluctuate at around 0. Although there is a slight downward trend, these effects do not seem to help explain the premium.



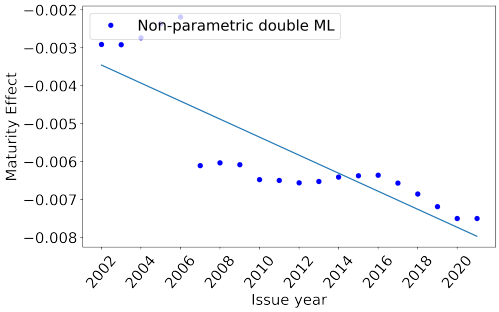
(a) Expected loss



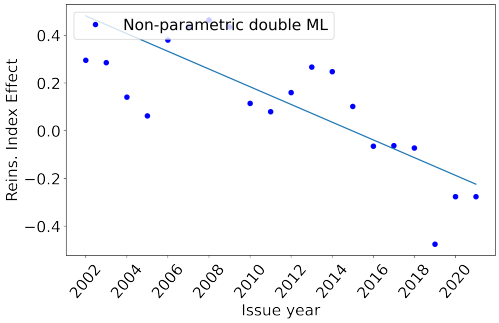
(b) Cat bond index



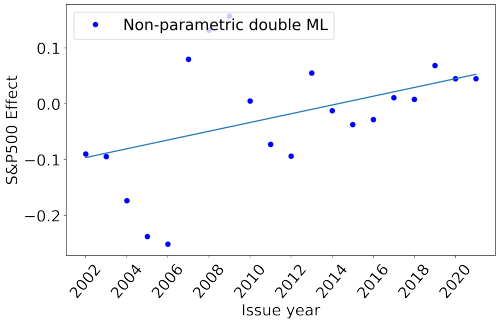
(c) Corporate spread



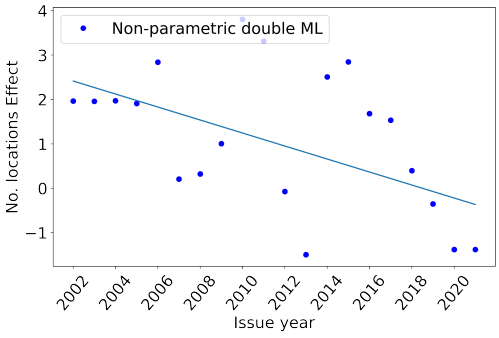
(d) Maturity



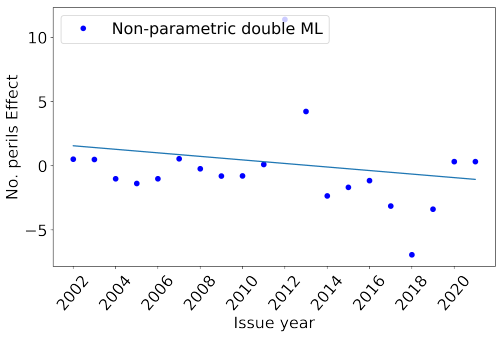
(e) Change in reins. index



(f) Change in S&P500



(g) Number of peril locations



(h) Number of peril types

Figure 4.7: HTE of the expected loss and other important factors on the premium over time.

5 Conclusion and future research

Recently, considerable progress has been made in developing methods for inference in the specific setting of random forests (Wager & Athey, 2018). Despite the bias-variance trade-off, causal random forests provide a sophisticated empirical toolbox, which is applicable across a range of different fields. I apply these methods in my work and quantify uncertainties and heterogeneities of effects in the cat bond market. In this part II of my thesis, I analyze if the expected loss has a time-varying effect in the primary cat bond market.

Through my analysis, the heterogeneity of the effect of the expected loss on the premiums becomes partially explainable. My heterogeneity analysis shows that the effect of the expected loss on the premiums has decreased in recent years. Capturing heterogeneity in a key parameter of interest is crucial because an average treatment effect is not helpful if it varies widely across sample groups. Although the confidence intervals of the conditional average treatment effects are very large, they give a first indication of uncertainty. Interestingly, this uncertainty seems to have decreased over time as the confidence intervals have shrunk.

Importantly, the interpretability of my results relies heavily on the assumption that there are no unobserved confounders. If an important feature that significantly affects the cat bond spread is missing, there are unobserved confounders. This would make any interpretative conclusions impossible. Since my data is not from a randomized experiment, I must use double machine learning to account for the potential correlation of confounders with both the "treatment" variable and the cat bond premium.

There are a few other limitations to my findings that leave room for further research. First, data availability is generally a strong limitation in the literature on cat bond asset pricing (Braun, Ammar, & Eling, 2019). Extreme event risks require data from a long time horizon, because securitized events only occur very infrequently. Second, I only conduct my analysis for the primary cat bond market. An additional analysis for the secondary market would help clarify if my results apply to both markets. Similarly, my results could be manipulated by a selection bias. For instance, I had to exclude bonds with missing entries. Although my results provide additional insights, they still do not explain why the cat bond market has generated high excess returns over the past two decades. If natural disaster risk is diversifiable by capital market investors, and systematic risks from the broader financial markets are minimized to an almost negligible extent, this should not be the case. Finally, causal methods are not only applicable to random forests. It would be interesting to apply them to neural networks and other machine learning models.

References

Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, *113*(27), 7353–7360.

Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, *11*(1), 685-725. Retrieved from https://doi.org/10.1146/annurev-economics-080217-053433 doi: 10.1146/annurev-economics-080217-053433

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, *47*(2), 1148–1178.

Braun, A. (2016). Pricing in the primary market for cat bonds: new empirical evidence. *Journal of Risk and Insurance*, *83*(4), 811–847.

Braun, A., Ammar, S. B., & Eling, M. (2019). Asset pricing and extreme event risk: Common factors in ils fund returns. *Journal of Banking & Finance*, *102*, 59–78.

Braun, A., Herrmann, M., & Hibbeln, M. T. (2022). Common risk factors in the cross section of catastrophe bond returns. *Available at SSRN 3901695*.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Routledge.

Carayannopoulos, P., Kanj, O., & Perez, M. F. (2020). Pricing dynamics in the market for catastrophe bonds. *The Geneva Papers on Risk and Insurance-Issues and Practice*, 1–31.

Carpenter, G. (2012). Catastrophes, cold spots and capital. navigating for success in a transitioning market. *Guy Carpenter, New York*.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). *Double/debiased machine learning for treatment and structural parameters*. Oxford University Press Oxford, UK.

Criminisi, A., & Shotton, J. (2013). *Decision forests for computer vision and medical image analysis*. Springer Science & Business Media.

Efron, B., & Hastie, T. (2021). *Computer age statistical inference, student edition: Algorithms, evidence, and data science* (Vol. 6). Cambridge University Press.

Götze, T., Gürtler, M., & Witowski, E. (2020). Improving cat bond pricing models via machine learning. *Journal of Asset Management*, *21*(5), 428–446.

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, *33*(5), 2223–2273.

Gürtler, M., Hibbeln, M., & Winkelvos, C. (2016). The impact of the financial crisis and natural catastrophes on cat bonds. *Journal of Risk and Insurance*, *83*(3), 579–612.

Götze, T., & Gürtler, M. (2020). Risk transfer and moral hazard: An examination on the market for insurance-linked securities. *Journal of Economic Behavior & Organization*, *180*, 758-777. Retrieved from https://www.sciencedirect.com/science/article/pii/S0167268119302008 doi: https://doi.org/10.1016/j.jebo.2019.06.010

Herrmann, M., & Hibbeln, M. (2021). Seasonality in catastrophe bonds and market-implied catastrophe arrival frequencies. *Journal of Risk and Insurance*, *88*(3), 785-818. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/jori.12335 doi: https://doi.org/10.1111/jori.12335

Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, *34*(11), 2767–2787.

Louppe, G. (2014). Understanding random forests: From theory to practice. *arXiv preprint*

arXiv:1407.7502.

Makariou, D., Barrieu, P., & Chen, Y. (2021). A random forest based approach for predicting spreads in the primary catastrophe bond market. *Insurance: Mathematics and Economics*, *101*, 140–162.

Molnar, C. (2022). *Interpretable machine learning* (2nd ed.). Retrieved from https://christophm.github.io/interpretable-ml-book

Mullainathan, S., & Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87–106.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379-423. doi: 10.1002/j.1538-7305.1948.tb01338.x

Shapley, L. S. (1953). A value for n-person games. In H. W. Kuhn & A. W. Tucker (Eds.), *Contributions to the theory of games (am-28), volume ii* (pp. 307–318). Princeton: Princeton University Press. Retrieved from https://doi.org/10.1515/9781400881970-018 doi: doi:10.1515/9781400881970-018

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, *113*(523), 1228–1242.

Declaration of Authorship

I hereby declare

- that I have written this thesis without any help from others and without the use of documents and aids other than those stated above;
- that I have mentioned all the sources used and that I have cited them correctly according to established academic citation rules;
- that I have acquired any immaterial rights to materials I may have used such as images or graphs, or that I have produced such materials myself;
- that the topic or parts of it are not already the object of any work or examination of another course unless this has been explicitly agreed on with the faculty member in advance and is referred to in the thesis;
- that I will not pass on copies of this work to third parties or publish them without the University's written consent if a direct connection can be established with the Stockholm School of Economics or its faculty members;
- that I am aware that my work can be electronically checked for plagiarism and that I hereby grant the Stockholm School of Economics copyright in accordance with the Examination Regulations in so far as this is required for administrative action;
- that I am aware that the University will prosecute any infringement of this declaration of authorship and, in particular, the employment of a ghostwriter, and that any such infringement may result in disciplinary and criminal consequences which may result in my expulsion from the University or my being stripped of my degree.

St. Gallen, February 17, 2023

Tim Mathers

(Signature of the candidate)

By submitting this academic term paper, I confirm through my conclusive action that I am submitting the Declaration of Authorship, that I have read and understood it, and that it is true.