ERASMUS SCHOOL OF ECONOMICS

MASTER THESIS
DATA SCIENCE AND MARKETING ANALYTICS

IMPROVING PREDICTIVE PERFORMANCE WITHIN DRIVER´S INSURANCE CROSS-BORDER CLAIMS THROUGH A COLLABORATION BETWEEN ML AND HUMAN EXPERTS

Author: Gabriel Andrade Calderon (611216)

Supervisor: Prof. Dr. Michiel van Crombrugge

Co-reader: Prof. Dr. Aurore Archimbaud

Date: July 31, 2022

Acknowledgements

This achievement is not only mine and I would like to share it with the people that helped along the way. I would like to thank Erasmus University Rotterdam, my professors, and thesis supervisor for all of their knowledge shared. I would like to thank Van Ameyde International and all of the people working there for their trust in my thesis project, their advice, as well as the data they provided. I would like to thank all of the people behind the Holland Scholarship 2021-22. Without it, none of this would have been possible. I would like to thank my family and friends for their support in this meaningful stage of my life which has made me grow as a professional and as a person.

Abstract

In a globalized world there is a large number of car accidents involving people visiting a country instead of being resident of the country. Driver insurance multilateral agreements such as the Green Card System in Europe functions to cover these situations, so that third party victims are compensated for loss caused both in material damage and in bodily injury. These cross-border arrangements encompass an entire new array of challenges when compared to local arrangements, as they must deal with different laws and regulations, protocols, prices, currencies, languages, and more. Because of its complexity, it is important for insurance companies in this context to make accurate prediction of costs ahead of payment dates. Consequently, Artificial Intelligence and supervised Machine Learning become very useful tools. However, there is a vast amount of data generated before, during, and after the car accident. Some of the data relates to the ´who´, ´what´, ´when´, ´where´, and ´why´ of the car accident. This study dives into big data in order to investigate, through Global Interpretation Methods applied to black-box models in the form of Random Forest and CatBoost algorithms, which type of predictors are more important and how when predicting a driver´s insurance claim cost. It finds that some of the most relevant predictors are not directly related to the car or object damaged, but to the context such as time, location, parties involved, and cause of the accident. Additionally, this study configures a setting in which AI is used in collaboration with human expert claim handlers to achieve better predictive performance than either AI or human experts by themselves. The increase in predictive performance when including human expert input into the trained models is remarkable. This sheds new light into the currently discussed topic of AI versus Human Expert and suggests a successful implementation of a synergy between the two into insurance claim handling. Finally, it investigates how the progressive inflow of information about the car accident during the lifecycle of a claim changes the relative importance of AI versus human expert input for the model to make an output of the cost.

Keywords: Artificial Intelligence, Machine Learning, Random Forest, CatBoost, Driver´s Insurance, Green Card System, Claim Handling, Insurance Claim Predicting, AI + Human Expert Collaboration

CONTENTS

I. INTRODUCTION 6
II. LITERATURE REVIEW 7
The current state of AI and ML applications 7
AI and ML in insurance 9
Studies and ML in Driver’s Insurance 10
AI + Human collaboration 12
III. CONCEPTUAL FRAMEWORK 13
Overview 13
The AI side of the collaboration 14
LASSO Regression 16
Random Forest 17
CatBoost 19
The Human Expert side of the collaboration 21
Research Questions 24
IV. DATA 25
Subcategory selection 28
V. RESULTS 28
Solely AI models 29
AI + Human Expert Collaboration Models 30
In-depth analysis of the models at month 1 35
LASSO 35
Random Forest 37
CatBoost 38
Interpretation of the variables in the best performing model 39
VI. CONCLUSION, LIMITATIONS, AND FUTURE RESEARCH 47
VII. REFERENCES 49
APPENDIX A: Exploratory Plots – Before Modeling 54

Figure 1. Claim cost estimation progression with increasing degree of human knowledge ... 14
Figure 2. Data Science process workflow ... 16
Figure 3. Gradient Descend ... 20
Figure 4. Word Cloud: Most common words or phrases in description about how Claim Handlers estimate reserves ... 23
Figure 5. RMSE of the reserve and models as time progresses ... 31
Figure 6. MAE of the reserve and models as time progresses ... 32
Figure 7. random forest: relative importance of the variable reserve versus all the other variables across time ... 34
Figure 8. CatBoost: relative importance of the variable reserve versus all the other variables across time ... 35
Figure 9. lasso coefficient shrinkage ... 36
Figure 10. random forest: out-of-bag error across number of trees ... 37
Figure 11. Random Forest: variable importance plot ... 40
Figure 12. PDP: EventDateYear (left) & PDP: ClientName_Freq (right) ... 41
Figure 13. PDP: IsJuridicalExpert (left) & PDP: IsMedicalExpert (right) ... 42
Figure 14. PDP: EventCountry ... 43
Figure 15. PDP: EventDateMonth ... 44
Figure 16. PDP: VehicleMake ... 45
Figure 17. PDP: AccidentCause ... 46
Figure 18. Exploratory boxplot between EventCountry & Cost ... 54
Figure 19. Exploratory boxplot between ClaimedObjectSubType & Cost ... 55
Figure 20. Exploratory boxplot between VehicleMake & Cost ... 56
Figure 21. Exploratory boxplot between accidentcause & Cost ... 57

Table 1. Independent variables ... 26
Table 2. CatBoost: two-step consecutive grid search results for hyperparameters . 38
Table 3. ranking of the algorithms used based on RMSE, MAE, and Training/Tunning time ... 39

I. INTRODUCTION

The growth of the insurance industry seems to have regained pace after slowing down during the Covid-19 pandemic in 2020 and 2021, as analyzed by The Global Insurance Report from McKinsey & Company (Bernard, P.I. et al., 2022). Even more so, the global growth was larger from the period of 2020 to 2021 than from the period of 2018 to 2019, pre-pandemic times. This rebound has been more pronounced in the American continent, followed by Europe, Middle East, and Africa, with Asia and Asia Pacific at the end. Similarly, according to the Insurance Industry Outlook from Deloitte (Shaw, 2021), the insurance industry will have an accelerated growth from 2022 onwards. This suggests that it will be a key player in the global markets in the following years. From the same Deloitte study, it is estimated that the largest increase in spending within the specific industry, when it comes to emerging technologies, will be destined to Artificial Intelligence (AI), followed by Cloud Computing and Storage, Data Privacy, Data Acquisition and Processing, and Cybersecurity. The larger increase in spending is expected to be in AI and thus in Machine Learning, which constitutes a key aspect of AI. The Covid-19 pandemic has been stimulating the digital transformation of the economy (Levantesi & Piscopo, 2021). Companies have been forced to invest in technology to satisfy consumer needs in a remote manner. This has brought a hastened modernization of the industry. Within this modernization, AI has the potential to improve processes and results. This might happen for example in the form of automation, such as AI chatbots, as well as in the form of predictions, such as claim cost predictions. The former is one of the key motivations behind this study. Lots of data is generated within insurance. For instance, a car accident and a subsequent insurance claim involves data about the insured, such as demographic information and behavioral patterns; data on the other parties involved, such as victims, lawyers and law makers, the police, Governmental Authorities, and more; data on the location of the accident, such as country, city, the type and condition of the road; data on the time of the accident, such as month and season, time of the day, weather; data on the accident itself, such as the cause of the accident, the vehicle brand, the objects involved; data on the insurance company, such as how many claims it handles every year, the type of coverage; and a lot more.

These are just a handful of examples. It is well known that there is a lot of data, but it is not clear which data is more important when trying to predict a driver's insurance claim cost.

That being said, not all types of insurance products are the same. For example, there is a considerable difference between driver's insurance and life insurance. For the same reason, the study zooms in the driver's insurance Green Card System, which is a multilateral agreement between 47 countries which stipulates that a person driving to a foreign participating country must be insured so that if the person causes or is involved in a car accident, any third party victims are not negatively affected by the fact that the driver was from a foreign country and not from the country in which the accident took place. This facilitates the crossing of borders, and also the loss settlement for both material damages and bodily injuries for the victims. Every year, more than 400.000 car accident in Europe fall under the Green Card category (Council of Bureaux, 2022).

Furthermore, it is a hot topic in data science the discussion about how AI works either against or in synergy with humans. Against in the sense that it might displace workers by taking over their tasks. However, this study experiments with Machine Learning settings and showcases how a configuration in which AI is used in collaboration with human expertise delivers better results when predicting driver's insurance Green Card claims. AI should not be focused on competing against years of know-how but on complementing and improving it. These is another key motivation behind the study.

In brief, the scope of study is important because the insurance industry is revitalizing itself; it is current because it analyzes how to implement AI and ML in the insurance context which is precisely what a lot of insurance companies are venturing on at this moment in time; and it is innovative by experimenting and suggesting a new take on how to establish a beneficial collaboration between AI and Human Experts with improved results compared to how some insurance companies are estimating costs at present.

II. LITERATURE REVIEW

THE CURRENT STATE OF AI AND ML APPLICATIONS

From the past decades, Artificial Intelligence (AI) and Machine Learning (ML) have been in constant development as fields of study, with constant new techniques and algorithms being engineered, but also as applications to real life problems and business settings with an increasing number of industries and companies implementing data-driven solutions beyond calculations done by humans and shifting towards less linear calculation done by computers. According to a study done by McKinsey (Chui et al., 2021) about the "State of AI", 56% of all respondent organizations affirm that they have implemented AI in at least one of their activities. The adoption of AI seems to be more predominant in the areas of service operations, product and/or service development, and marketing and sales, in that order. AI is no longer a futuristic vision, but something that is happening at this moment in time, and is making an impact on the world today. AI is increasing in importance in a variety of sectors because of the possibilities it brings in regards to integrating information that is in constant genesis and in continuous update, analyzing big and unstructured data, and utilizing its outputs to improve decision-making (West & Allen, 2018).

As gathered and summarized by Sarker (2021), some of the ML applications that have gained a lot of attention in recent times are the following: *predictive analytics and intelligent decision-making*, which is commonly applied in determining an unknown outcome such as identifying criminal suspects, detecting credit card fraud, avoiding out-of-stock situations, and anticipating sales (Adewumi & Akinyelu, 2017); *cybersecurity and threat intelligence*, which involves protecting networks, systems, hardware, and data from digital attacks (Dasgupta et al., 2022); *internet of things and smart cities*, which focuses on turning everyday objects into smart objects through data transmission and task automation, an example being smart industrial robots in warehouses (Souza et al., 2019); *traffic prediction and transportation*, which consists of managing traffic through future traffic prediction and optimization of elements such as traffic lights and parking spots (Boukerche & Wang, 2020); *healthcare and the Covid-19 pandemic*, which serves as a tool to assist in practices such as diagnosis based on symptomatology and patient management (Shailaja et al., 2018), and in the Covid-19 context, in practices such as patient risk classification and forecasting where the virus is likely to spread and potential peaks in cases, hospitalizations, and mortality (Islam et al., 2020); *e-commerce and product recommendation*, which is one of the most predominant aspects of any e-commerce website nowadays and uses consumer's previous behavior and previous purchases to predict future behavior and future purchases as well as to personalize preferences and experiences (Micol Policarpo et al., 2021); *natural language* *processing (NLP) and sentiment analysis*, which covers a broad spectrum of language analysis and language processing by computers through text and speech, in order to read, understand, interpret, adjust, and respond, some examples being chatbots, machine language translations, and the analysis of emotions in product reviews (Mehta & Pandya, 2020); *image, speech and pattern recognition*, which uses less structured data in the form of images for applications such as facial recognition, landmark identification, and social media tagging recommendation (Pak & Kim, 2017); *sustainable agriculture*, which seeks to optimize the resources obtained from agricultural practices while reducing the negative impact to the environment (Sharma et al., 2020); and *user behavior analytics and context-aware smartphone applications*, which bases on gathering information of the surrounding, often in real-time, to provide a context-sensitive output such as geolocated advertisement (Al-Saedi et al., 2022). With all of these applications, it is no wonder that this technological surge (with ML at its core) is commonly being referred to as the fourth industrial revolution (4IR, Industry 4.0).

AI AND ML IN INSURANCE

The insurance industry is no exception to the mentioned trend, as ML algorithms are being applied to fraud prevention, risk management, claims processing, and others (Nevins, 2021). An example of a relevant application is Severino & Peng (2021) study which compared nine different statistical methods and Machine Learning algorithms to detect fraudulent claims in a Brazilian insurance company. In their study, the most accurate models turned out to be a Random Forest and a Gradient Boosting algorithm, even more so than a Neural Network. However, they only included 8 variables in their analysis. Similarly, Levantesi et al. (2020) applied ML to longevity calculations in life insurance and compared the accuracy of the ML predictions with the predictions made with more traditional methods used in the life insurance industry since many decades ago, such as the Lee-Carter and the Renshaw-Haberman models. The ML methods proved to be more accurate in all of their tests.

In regard to the claim handling application, the studies are mostly focused on automation. Oza et al. (2020) studied the implementation of Robotic Process Automation (RPA) so that an AI performs business tasks in a similar manner to human users, with the purpose of reducing costs, handling times, and errors. Other studies have been focused on predictive analysis. Blier-Wong et al., (2021) recompiled 77 publications that have used some type of ML in property & causality insurance. They mention how linear models have predominated in the field during the past decades but at the same time how they may be too simple to reflect reality, as relationships between the variables that compromise insurance activities are seldom just linear relationships. Thus, the overdue implementation of ML algorithms in insurance has been an important milestone. According to the recompilation, the number of ML publications in the mentioned field in 2018 alone is larger than in the entire interval of 2000-2014. Moreover, the largest proportion of supervised ML has been focused on the prediction of pricing (the price of the prime for the insured), and a smaller portion has been focused on the prediction of reserves[1] or insurance claim costs. This is so, first, because reserve data is usually unstructured since the number of payments and the time until a clam is settled are unknown at the time of registering a claim for the first time; second, because data is often aggregated either by group such as portfolio level or by time such as quarterly; and third, because some variables are dynamic and may change over time during the claim handle process (Blier-Wong et al., 2021). Overall, AI and ML is blossoming in the insurance industry, and the majority of the studies carried out seem to suggest that the implementation of these methods tend to lead to more precise predictions and more efficient operations.

STUDIES AND ML IN DRIVER'S INSURANCE

As seen, plenty of research has been done in the insurance industry, which is very wide. However, the findings for one type of insurance product are not necessarily applicable to other types of insurance products because, even though they have in common the purpose of mitigating a risk, they involve very different activities and processes. Some examples can be seen; life insurance is projected for the long term, while travel insurance is projected for the short term; health is a human right; natural disasters are a rare occurrence: most insurance products are optional; the Green Card System is compulsory in order to drive in the participating states. Therefore, it is important to look at previous studies in the relevant subtype of insurance product.

[1] A claim reserve is defined by the International Risk Management Institute as "an amount of money set aside to meet future payments associated with claims incurred but not yet settled at the time of a given date" (2022)

Motor car or driver’s insurance has been studied since many decades ago. For instance, already in 1978, Zehnwirth proposed a hierarchical model for the estimation of claim rates in a motor car insurance portfolio. Further studies have also been focused on examining factors that increase the frequency of road accidents and the consequential number of claims. Braver & Trempel (2004) studied the relationship between the occurrence of claims and the age of the driver and found out that older drivers tend to be involved in more road accidents than younger drivers, and that they are more often than average liable for the collision. Nonetheless, Twisk & Stacey (2007) discovered that that young driver’s relative risk to road accidents has been consistently increasing in Europe, especially in men when compared to women, due to attitudes such as driving at night, at high speed, under the influence of alcohol/drugs, and negligence with the use of the seatbelt. Furthermore, they also investigated this increase in risk across countries in the EU. More recent studies have been conducted on driver’s behavior and on proposed Pay-how-you-drive (PHYD) insurance approaches instead of Pay-as-you-drive (PAYD) (Tselentis et al., 2017).

When it comes to ML, a large proportion of the studies seem to be focused on binary classification, such as detecting if a claim is fraudulent or not. For instance, Wang & Xu (2018) used text analytics and deep learning to detect automobile insurance fraud with 91.4% out-of-sample accuracy. Yet, the study does not go into explaining which characteristics of the text tend to lead more often than not to a fraudulent claim. Moreover, Explainable Machine Learning focuses on the accuracy of the predictions but also on the interpretability of the effects of the independent variables on the dependent variable. Maillart (2021) used Global Interpretability Methods on a Random Forest to explain the predicted occurrence of driver’s insurance claims in Belgium which incorporated telematics that capture the driving behavior of the insured with information such as the distance driven during the ensured period, the number of trips, the road type, and the different regions within Belgium. Maillart’s study already finds differences in claims across different regions in Belgium. However, this study was aimed at predicting claim occurrence and not claim cost. Guelman's (2012) study uses a set of predictors classified as either characteristics of the driver, characteristics of the policy, characteristics of the vehicle, or accident/conviction history to predict insurance loss cost, with loss cost being the multiplication of claim frequency and claim severity. Using Gradient Boosting (GB), it finds that the most important variables for predicting the severity of a claim are characteristics of the vehicle, such as vehicle age and vehicle price; characteristics of the driver, such as years licensed; and only one characteristic of the policy, this being collision deductible. This study provides an interesting classification of features which is expanded upon. Yet, the algorithm used (GB) is currently considered somewhat outdated and usually outperformed after the emergence of other more developed forms of Gradient Boosting such as Extreme Gradient Boosting (XGBoost) and Categorical Boosting (CatBoost), as tested in claim predictions within the driver's insurance industry (Fauzan & Murfi, 2018). Additionally, because of the difficulty of obtaining consistent (comparable) and sufficient (large enough sample size) data about car accidents and subsequent insurance claims, some studies have used simulated data to test theoretical hypotheses that are expected to hold true in real life. An example is Baudry & Robert's (2019) study on the prediction of reserve costs to showcase the use of individual and not aggregate data for claim analysis.

AI + HUMAN COLLABORATION

As AI develops into a more complex tool, and as it becomes more entwined in business activities, the questions about how and to what extent will it displace human tasks and human labor becomes more relevant. Evidence in the service sector seems to suggest that AI has progressively developed key characteristics needed in order to potentially replace human intelligence in service tasks, these being mechanical, analytical, intuitive, and empathic intelligence characteristics (Huang & Rust, 2018). According to the authors, intuitive AI (or strong AI) is being developed by ambitious projects from large companies, and it involves the machine's ability to adjust to unexpected situations, to learn from previous experiences, to perform with imperfect information, and maybe even to be self-conscious, aware of its own limitations, and more importantly aware of what it needs in order to overcome its own limitations. It can be said that AI is evolving; however, the discussion about whether AI can develop empathic intelligence, at least in a practical (being able to read emotions from others and adjust as in 'what is the socially correct response to this display of emotion') and not organic (actually being able to feel emotions) manner is an ongoing discussion. For this very same reason, some scholars argue that AI can never replace humans when empathy is required for the task in question. Montemayor et al. (2021) divides empathy into emotional empathy, cognitive empathy, and motivational empathy. Emotional empathy is defined by viscerally or biologically experiencing emotions and is followed by motivational empathy (drive to help others) in the form of a genuine concern due to the similar emotions that others might be experiencing. On the other hand, cognitive empathy is based not on experiencing but on recognizing emotions or cues of emotions which produces a far more dissociated type of empathy. As such, the motivational empathy that follows after a cognitive empathy might be based on different reasons, including non-altruistic ones such as self-interest and manipulation. A chatbot AI trained to read and respond to the emotions of others ultimately is also trained with an outcome in mind, such as to improve customer satisfaction or to get better service quality reviews. Therefore, the question remains: is a self-interest driven empathic response really empathy? But on the other hand: is empathy really ever totally and completely altruistic? Even for humans? There are no objective answer for these questions. Thus, the most composed response to the AI and human relationship dilemma is not to replace humans with AI or to neglect AI entirely, it is to establish a collaboration between AI and Human Experts to harness the strengths of each side and adjust depending on the specifics of the task in question.

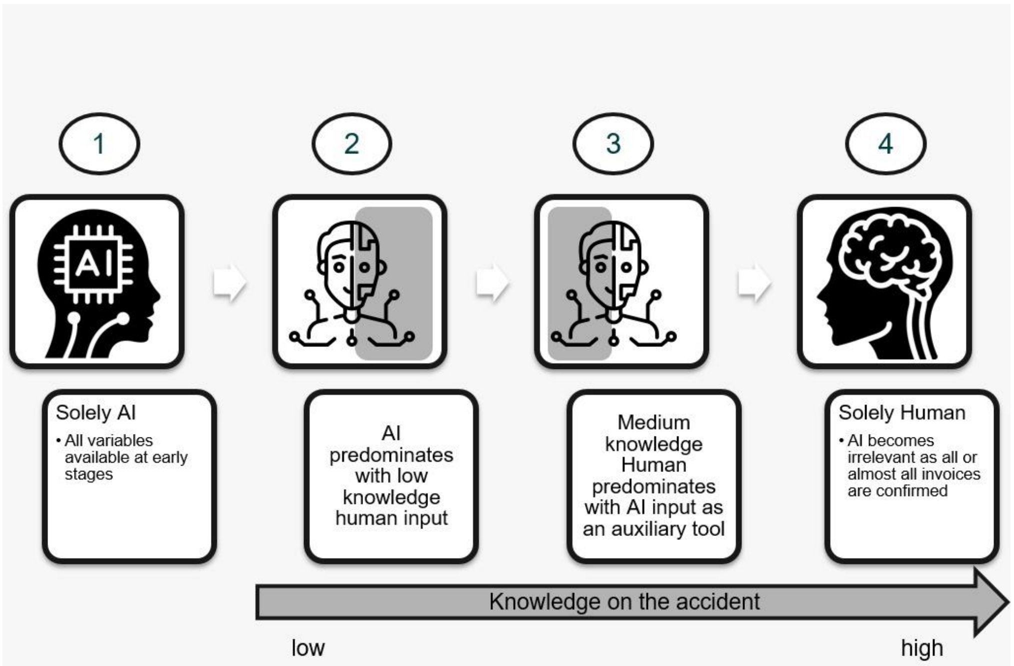
III. CONCEPTUAL FRAMEWORK

OVERVIEW

Leveraging the data that was made available, this study seeks to build on previous research and overcome some of the common limitations by applying predictive and interpretable Machine Learning to real life data that is sufficient and consistent across countries in Europe.

Moreover, the objective is to predict Green Card System driver’s insurance claim costs ahead of payment dates and compare the predictive performance of models that rely solely on AI versus models that establish a collaboration between AI and Human Experts with different degrees of knowledge about the car accident and the subsequent insurance claim. This sheds new light into how AI can be implemented successfully into the driver’s insurance industry. It also puts into perspective how AI and Human Experts can work in synergy. Finally, it establishes a pathway in time about how claim cost calculations in insurance progress from a solely AI estimation to a solely Human estimation as knowledge about the claim (from the expert claim handler) increases from very limited at Stage 1 (only the most basic information about the accident is known to the claim handler), to low at Stage 2 (some more early details about the accident arrived and a rough reserve estimate is modified by the claim handler accordingly), to high at Stage 4 (costs about the damages or injuries are confirmed to the claim handler, reserve estimate is modified accordingly again, and payments are ready to be wired). The setting is a case of imperfect information versus perfect information, and all the partial information instances in between.

FIGURE 1. CLAIM COST ESTIMATION PROGRESSION WITH INCREASING DEGREE OF HUMAN KNOWLEDGE



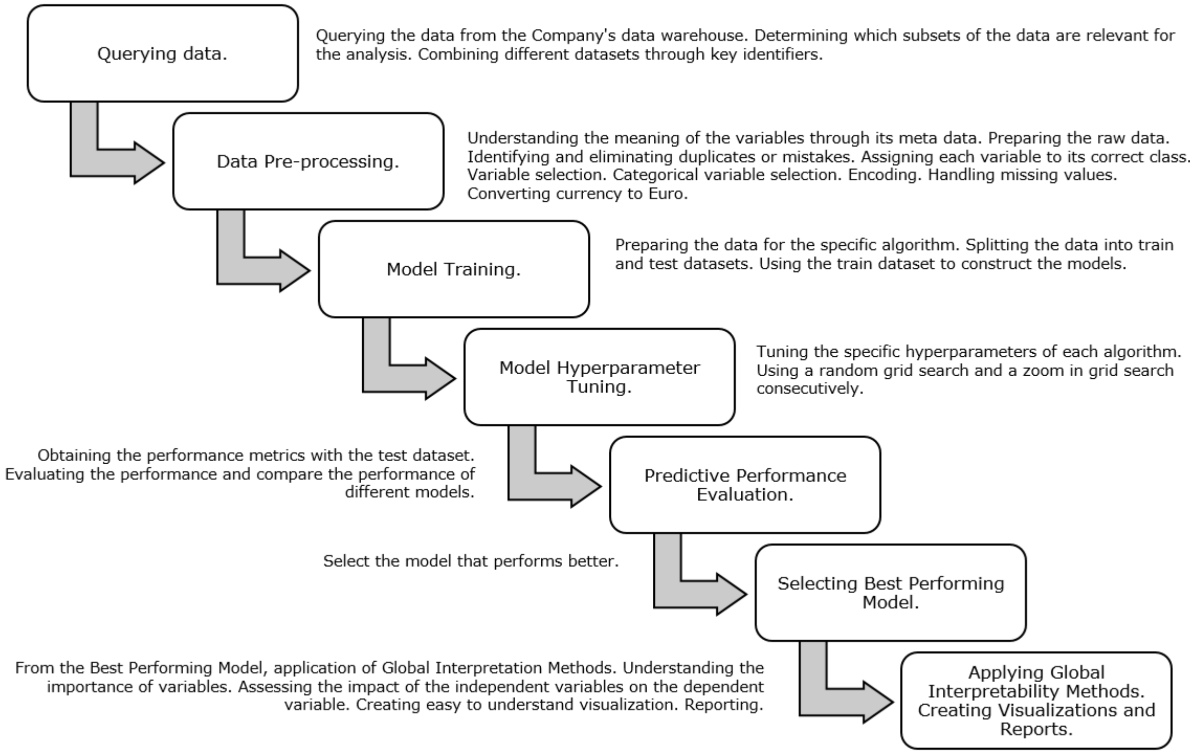
THE AI SIDE OF THE COLLABORATION

To predict driver's insurance claim costs ahead of the payment date, AI input was used in the form of Machine Learning algorithms. These algorithms were trained on the data available. The data will be discussed more thoroughly in Section IV. The task of the study falls under the category of Supervised Machine Learning, which constructs algorithms that are capable of finding and reproducing patterns by using externally supplied instances to predict future instances (Singh et al., 2016). Moreover, the data was analyzed in a predictive and interpretative manner using algorithms that were fit for the type of data, these being LASSO Regression, Random Forest, and

Boosting though Categorical Boosting (CatBoost)[2]. Furthermore, an out-of-sample test was carried out, in which 60% of the data was used to train the models and 40% of the data was used to test the performance of the models. The models were used to solve a regression problem, more specifically to predict the numeric cost of a driver's insurance claim, and as such the main performance metrics considered were the Root Mean Square Error (RMSE) and the Mean Absolute Error (MAE), which are the most commonly used performance metrics for ML regressions. The predictive performance of the various models was compared and a best performing model (or the one with the smallest error in this train-test setting) was selected. Finally, the interpretative aspect of the analysis was carried out using Global Interpretation Methods, specifically Permutation Feature Importance and Partial Dependence Plots (PDP). The former measures the increase in the prediction error of the model when the features are iteratively permuted, as in to break any relationship between the values of the feature and the outcome. The latter depicts the marginal effect that one or two variables have on the predictions of a trained model; this relationship between the variables and the predictions can be linear, monotonic, or more complex (Molnar, 2020). In brief, the first provides interpretation on the impact of each variable on the predictions, taking into consideration that the impression is about the model (which has an error term) and not directly about reality (Is the fact that the car involved in the accident was a Ferrari important when predicting the claim cost?). And the second provides interpretation on the direction that a variable has on the predictions (Do Ferraris on average lead to lower cost claim predictions or higher cost claim predictions?). The summary data analysis process workflow from querying the data to elaborating visualizations and results can be observed in Figure 2.

[2] Regular linear regression models were not used due to there being too many variables – "The curse of dimensionality". Neural networks were not used due to the fact that dozens of models were trained and the significant increase in complexity and training/tuning times for neural networks was not justified by a possible but not certain increase in performance in tabular data.

FIGURE 2. DATA SCIENCE PROCESS WORKFLOW



LASSO REGRESSION

Least Absolute Shrinkage and Selection Operator or LASSO is a type of linear regression model in which a penalization term is added in order to prevent overfitting when many independent variables are included in the analysis. Overfitting can lead to overestimating the predictive performance of the model as it performs particularly well on its training dataset but poor on the testing dataset and beyond. LASSO achieves variable selection by identifying the regression coefficients associated to features that minimize the prediction error and shrinks the coefficients of less relevant variables to a value very close to 0 or even 0 (Ranstam & Cook, 2018). Variables whose coefficient are shrunken to 0 are excluded from the model. As defined by Tibshirani (1996) and seen in equation (1), LASSO solves a minimization problem between the real outcomes and the predicted outcomes in an OLS setting constrained by a fixed value (λ), which limits the sum of the absolute values of all the coefficients. Thus, coefficients are shrunken in order to reach an optimal solution.

$$\underset{\beta}{\text{minimize}}\left\{\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j}\beta_j x_{ij}\right)^2\right\} \textit{ subject to } \sum_{j=1}^{p}\left|\beta_j\right|\le\lambda \qquad (1)$$

From the equation (1), y is the real value of the dependent variable, β_0 is the coefficient of the intercept, β_j is the array of coefficients for all independent variables x_j, where $x = x_1, x_2, \dots, x_j$, and λ is a tuning parameter that regulates the shrinkage of the coefficients. When $\lambda = 0$, the results are the same as in a linear regression. When λ increases, coefficients are shrunken in a greater degree, and more coefficients are shrunken to 0.

LASSO served as a good starting point in this analysis due to the large number of independent variables available. More so, many of the variables were categorical variables with numerous subcategories, and in a linear regression context each subcategory (excluding the base subcategory) is given a coefficient. A regular Multiple Regression or a Ridge Regression would have resulted in too many variables. The LASSO Regression model also served the purpose of a benchmark model to compare to the more advanced Machine Learning algorithms discussed further.

RANDOM FOREST

The Random Forest algorithm (Breiman, 2001) builds on Bootstrap Aggregation (Bagging) and Classification and Regression Trees (CART). Depending on the purpose of the task, it can be used to solve either a regression or a classification problem. It follows the logic of an ensemble method, a technique that uses multiple models instead of a single model to obtain better results. In the case of the Random Forest, multiple decision trees are trained creating a forest of decision trees. Each tree is trained and evaluated on different subsets of the data by implementing Bootstrapping. For each tree in the forest, a portion (usually two thirds) of the dataset is used to train the tree and the remaining portion of the dataset (the remaining one third) is used to evaluate the performance of the tree. This leads to an internal metric of error called the out-of-bag (OOB) error. Bagging uses a random selection of observations on the tree growing process, the Random Forest algorithm expands on that by additionally using a random selection of features to decide on the splits of the nodes. This helps in building trees that are less correlated between each other. Other benefits of a random selection of features and consequently of the Random Forest algorithm are, according to Breiman (2001), an accuracy as good or even better than Adaboost; robustness to outliers and noise; computational speeds faster than bagging or boosting; useful internal estimates of error, strength, correlation and variable importance; and easy parallelization. After the decision trees in the Random Forest are trained, the definitive output of the model is given by either the majority of votes (tree one votes for outcome A, tree two votes for outcome A, and tree three votes for outcome B, thus through a majority of votes, the output is determined to be A) or the mean of the outcomes (tree one estimates 4, tree two estimates 7, and tree three estimates 3, thus by calculating the mean, the outcome is 4.66). This of course depends on whether it was a classification or a regression problem.

A decision tree has nodes and splits. In order to select the variables that determine these nodes and splits, a Random Forest applied to a classification task frequently uses the Gini impurity metric. The logic behind is that some variables are more polarizing, in the sense that they lead to certain outcomes more than other variables, or in a more determinant or consistent manner. Intuitively, Gini impurity across variables is heterogeneous. The variables with the lowest impurity are selected for the splits. The formula for Gini impurity (Daniya et al., 2020) is shown in Equation (2), where i is the number of classification labels from i to n and p_i is the probability that the data point belongs to classification label i.

$$Gini\ Impurity = 1 - \sum_{i=1}^{n} p_i^2 \qquad (2)$$

For a regression task, a Random Forest determines its splits by selecting the variables that reduces the variance using the Mean Square Error or the Mean Absolute Error.

Furthermore, there are some hyperparameters in a Random Forest that can be tuned in order to improve its performance. The ones tuned in this study are the following:

1. The number of trees in the Forest (*ntree*).

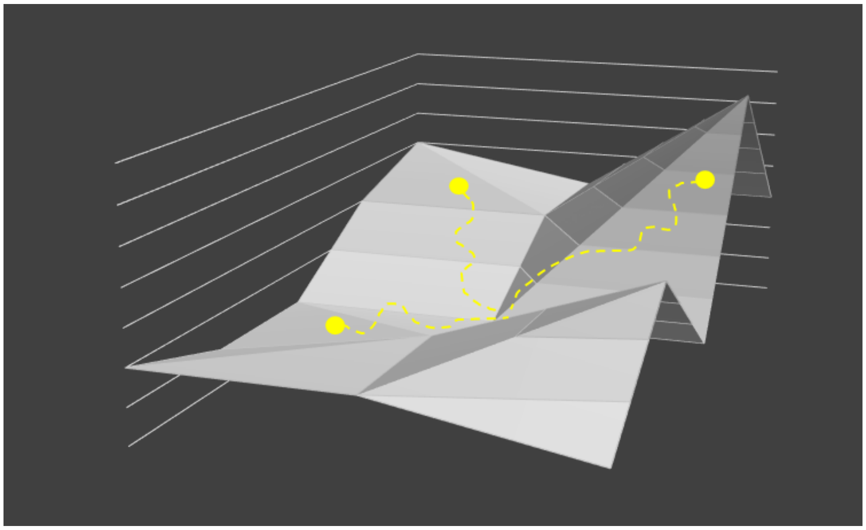
2. The number of features to be randomly sampled at each split (*mtry*).

The motivation to use a Random Forest algorithm in the study comes from its versatility when handling different types of variables, its ability to grow diverse trees using a random selection of features and observations, and its robustness against noise variables and overfitting (Hastie et al., 2009). Additionally, since a large number of models were trained and tested (at numerous points in time during the claim handling process), its superior computational efficiency was advantageous. On top of that, Random Forest algorithms tend to have high predictive performance. Hastie et al. (2009) compared the performance of Bayesian Neural Networks, boosted trees, boosted Neural Networks, Random Forests, and bagged Neural Networks using five datasets of different dimensions and from a variety of domains. The Random Forest algorithm was outperformed only by the Bayesian Neural Network. However, the average computation time ratio of the Bayesian Neural Network versus the Random Forest was of 202 to 1. Then, in the sense of performance per time, it can be said that the Random Forest was the evident winner.

CatBoost

CatBoost (Prokhorenkova et al., 2019) stands for Categorical Boosting and it is a type of gradient boosting algorithm based on decision trees. Like the Random Forest, it is an ensemble method that uses an n number of models together to obtain better results. The key difference with bagging is that the models are trained not independently from one another, but sequentially one after the other by also implementing a learning rate. The individual models are weak learners that are improved progressively at every stage. The model trained at $n + 1$ focuses on improving on the mistakes made by the model trained at n. The algorithm seeks to minimize the error term through gradient descend. Gradient descend, briefly put, iteratively finds global or local minimums by means of first-order optimization. As seen in Figure 3, there are random initial points, then a gradient is calculated and a step is taken into the opposite direction of the gradient (since the objective is to find a minimum), and these steps are repeated until a minimum is reached. It must be taken into account that the procedure can lead to local minimums instead of global minimums, that is the reason and importance behind the numerous iterative random initial points.

FIGURE 3. GRADIENT DESCEND



Other gradient boosting algorithms based on decision trees exist, such as XGBoost, which is quite popular. However, CatBoost has comparable performance and is optimized for categorical variables. It is especially useful for sparse data or data that has categorical variables with high cardinality. XGBoost and other algorithms handle categorical variables through One-hot Encoding which greatly increases the dimensionality of the data and its sparsity, especially when there are categorical variables with high cardinality. Alternatively, CatBoost uses an innovative manner to handle categorical variables through Ordered Target Statistics which relies on permutations and ends up converting the categorical variables into numerical variables. Hancock & Khoshgoftaar (2020) reviewed 403 records of articles using CatBoost, XGBoost and/or LightGBM in diverse fields such as biology, medicine, cyber-security, finance, marketing, astronomy, and others and concluded that CatBoost is indeed a very good candidate for ML choice of algorithm and in various applications outperformed other algorithms particularly in the mentioned circumstances, when there are many categorical variables each with many subcategories. However, the algorithm seems to be quite sensitive to hyperparameter tuning, highlighting the importance of being meticulous in this stage of the model building process.

For this reason, the hyperparameters that were tuned for this study where various:

1. The maximum tree depth for each individual tree (*depth*).

2. The learning rate which determines how large should the change to the model in $n + 1$ be as a response to the error of the previous model in n (*learning_rate*).
3. The number of random initial points (*iterations*).
4. L2 regularization that penalizes less important features similar to a Ridge regression *(l2_leaf_reg*).
5. An additional random parameter in the determination of the splits to help against overfitting. (*random_strenght*).

Due to the high predictive accuracy of gradient boosting algorithms, it made sense to train a boosting model together with the Random Forest to compare their performances and ultimately select a best performing model. The CatBoost algorithm was preferred over XGBoost or LightGBM due to the fact that the data used for the study has a major number of categorical variables of which several have high cardinality.

THE HUMAN EXPERT SIDE OF THE COLLABORATION

When an insurance claim is first registered following a car accident, a reserve monetary value is set by the insurance company or in this case by Van Ameyde across its offices in the countries in which it operates (Van Ameyde is the Company that provided the data for the analysis. More about this will be discussed in Section IV.). The idea is for the reserve to be as exact as possible to the cost of the claim once valuations of the damages and other expenses are made throughout the entire duration of a claim handling process. The lapse in time between a claim first being registered and it being settled varies greatly across types of insurance products. In the context of the Green Card System driver’s insurance, claims are open for longer than average intervals of time due to the fact that they encompass cross-border procedures, the potential involvement of several parties including Governmental Authorities, and different laws and regulations. During this interval of time, the reserve value must be set with the information available and adjusted as information is updated. For Van Ameyde, information is constantly arriving in the form of phone conversations with the client (the insurance company), the insured (insured with the client, involved in the accident, and many times liable of the accident), or the third party (no relationship with the client, but involved in the accident, and many times a non-liable victim); emails; photos of the accident or objects damaged; expert valuation of the damage; medical information in the case of bodily injuries; progress on the medical treatment; second medical opinions (SMO); progress on trials; court rulings; updates on the loss of ability to work for the parties involved; return to work for the parties involved, and so on. Every case is different. Taking all of this into consideration, reserves are essentially predictions of how much claims will cost ahead of payment dates. Yet, these predictions or calculations have historically not been done by AI or Machine Learning algorithms but by Human Experts in the form of Claim Handlers, Risk Analysts, or similar. Therefore, using AI to predict the cost of a claim will potentially help the Human Experts in setting more accurate reserves. Both (Human Experts estimating reserves and AI predicting the cost of a claim) are different approaches to the same goal that can be used together. In summary, the reserve estimation of claim i is a function of the information available x for that claim i in the stage k of the claim handling process and of other unknown aspects z, plus an error term. z is not necessarily specific to the claim in question, but it may be a set of attributes related to the Human Expert or the insurance company.

$$Reserve_i = f\left(x_i^k, z\right) + \varepsilon \qquad (3)$$

For a better depiction of what z compromises, interviews to claim handlers were conducted with the purpose of getting insights on how they set claims without the use of any AI or ML, as well as to understand the logic behind their thought processes, the skills or tools they use, and/or the protocols they follow. These interviews were done to six claim handlers from the Van Ameyde offices in three different countries. The average duration of an interview was 30 minutes. Some of the questions that were asked to the claim handlers were "*How would you define a reserve?*", "*What aspects in a claim do you consider are more important when estimating a reserve?*", "*How often in the claim handling process do you modify a reserve?*", "*How does the information inflow from an average claim look like?*", "*How often do you receive new information?*", "*What are the usual sources of new information?*", "*How difficult is it to estimate a reserve that ends up being close to the actual real cost of a claim?*", and "*can you tell me the five most important aspects (in single words or short phrases) that describe how claim handlers determine the reserve values (these can be skills or tools or anything that comes into your mind)*". The answers to these questions provide some information about z. The Word Cloud in Figure 4 shows the responses gathered from the last question. All respondents mention experience as a fundamental aspect that claim handlers rely on when estimating reserves. They also use similar claims as reference, vehicle market values, historical data, data mostly through the form of averages, their general idea about the industry, they differentiate claims by clients, by whether there was a bodily injury, sometimes provide rough estimates, and wait for the arrival of more information.

FIGURE 4. WORD CLOUD: MOST COMMON WORDS OR PHRASES IN DESCRIPTION ABOUT HOW CLAIM HANDLERS ESTIMATE RESERVES

These are not results of the research questions of the study. Instead, they provide evidence that reserve history, their values and modifications, can serve the purpose of a proxy variable for Human Expert input in the prediction of claim costs ahead of payment dates. When a claim is registered for the first time into the platform used by Van Ameyde (called ECHO), a default value of the reserve is set automatically by the system. This default value is configured by a very general calculation of the averages of previous claim costs. The default values may vary across clients as there are individual arrangements with some. However, as soon as new information on the claim arrives, the default values of the reserves are modified at the discretion of the claim handlers; this is exemplified further by a quote that stood out from one of the claim handler interviews: "*If an identical claim is handled by ten different claim handlers, there will be ten different reserve values*". Another aspect to consider is the duration time a claim takes to be handled from the moment it is registered to the moment it is settled. Within the driver's insurance Green Card

System, claims are usually open for around 9 months, being the minimum time of 1 day and the maximum time of 132 months (some very few claims can be open for longer but the cutoff point in the dataset used for the analysis was 2010). The time unit used in the study was months. AI + Human Expert models were trained using the reserve values at every month from when a claim is first registered for a timeframe of 3 years. Thus, there were models trained for month 1, 2, 3, 4, 5, 6 and all the way to month 36, which is the last month analyzed[3] (example: for a claim a reserve was initially set at month 1, then modified at months 2, 3, 6, and 7. The claim was ultimately settled and closed at month 7. At month 1, the claim handler had very few information about the claim but at month 7 it had all the confirmed invoices, therefore the modification of the reserve at month 7 made it so that the reserve became equal to the actual cost. A model trained at month 2 uses this claim as an observation, but the variable Reserve takes the value that was modified at month 2, so it does not use the initial reserve value of month 1 nor the perfect reserve value of month 7). In conclusion, the human side of the collaboration is captured by including a variable *Reserve* in the models mentioned previously (except on the model that uses solely AI input).

RESEARCH QUESTIONS

After defining AI and Human Expert inputs in the context of this study, the research questions to be explored are the following:

RQ1: How does the predictive performance of models that rely solely on AI compare to models that integrate a collaboration of AI + Human Expert input when predicting Green Card driver's insurance claim costs?

RQ2: How does the progression from low information to high information change the relative importance of AI versus Human Expert input when predicting claim costs. (What is the

[3] The starting point of time for each claim is the registration date of that claim. This is different for every claim. If claim X was registered in February 2018, then the reserve value at month 3 for that claim is the reserve that was currently in place in the system in April 2018 for that claim. In the same manner, if claim Y was registered in July 2014, the reserve value at month 3 is the one that was currently in place in the system in September 2014 for that claim. Since time is analyzed, it is important to understand that time is time passed for a claim since that claim was first registered.

moment in time in which AI > Human Expert becomes AI < Human Expert with regards to variable importance).

RQ3: Which variables have the largest importance when predicting a Green Card System driver's insurance claim cost?

IV. DATA

The data for the analysis was provided by Van Ameyde International, one of the leading Third Party Administrator (TPA) companies in Europe. A TPA is a company that provides an operational service such as claim handling to another company. This is often a common practice within the insurance industry. Van Ameyde is present in 30 countries and works with over 1000 business clients including large international insurance companies. They handle around 750.000 insurance claims every year across most type of insurance products except life and health insurance. The study was focused on their operations of the Green Card System driver's insurance product, for which they handle claims on behalf of over 500 insurance companies. Since for this specific insurance product they represent the insurance companies, it is Van Ameyde who has direct contact with the insured, receives the information, handles the payment, and records the data in their system. Thus, even if the data is from insurance policies of different insurance companies, it follows the same structure. This type of data is difficult to come by, since in other settings it is usually handled by different insurance companies, hence the data is gathered in dissimilar manners, the variables are defined differently, they are inconsistent, and problematic to compare. On the other hand, the data for this study is in the ideal setting because the Green Card System is a European agreement and thus is standardized.

The data was queried directly from Van Ameyde's data warehouse. Datasets about accidents, claims, costs, damages, objects involved, parties involved, and reserves history were used. Some additional steps on the data that are worth mentioning are the following:

- Only '*closed*' claims were used to train the model. This means claims that are fully settled and are not expected to be modified anymore.

- Claims from 2015 to 2022 were included. However, the dataset is very unbalanced. Naturally, there are a lot more low-cost claims or medium-cost claims than high-cost claims. This was addressed by manual upsampling of high-cost claims. This means that only for high cost, the sample included claims from 2010 to 2022. This solution was implemented because it adds not only more observations of high-cost claims, but it adds more diversity on the dataset.[4]
- Claim costs that were not in euros were converted to euros using the monthly average conversion rate of the month-year in which the claim was registered for the first time in the system.
- Missing values were imputed using the *missRanger* function of the *missRanger* package (Wright et al., 2022) in *R*. This means that missing values were replaced by predictions from a Random Forest using the feature with missing values as dependent variable and all of the other features as independent variables, iteratively until all features had no missing values.
- The outcome variable '*Cost*' is the aggregate of all the costs Van Ameyde incurred for the settlement of the claim both for material damages and body injuries except the handling fee, which is not exogenous but is established by the company and thus did not make sense to include.

After the mentioned steps that were taken together the rest of the data-preprocessing, the data that was used to train the models consisted of 341.669 observations across the list of variables referred to in Table 1.

TABLE 1. INDEPENDENT VARIABLES

Variable	**Description**
ClientName	The insurance company for which Van Ameyde provides the claim handling service. The insured has its policy with the client.
Product	The type of Green Card insurance product.
Branch	The subtype of the insurance product.

[4] In order to not overestimate the importance of year on high-cost claim predictions, since an additional sample of older claims was included in the analysis only for high-cost claims, the information for the year of these observations was deleted and treated as a missing value, which was later imputed like all other missing values.

DamageType	Either Material Damage ir Bodily Injury.
Coverage	The general type of coverage the insured ha with the client.
EventDateYear	The number of years since the accident took place.
EventDateMonth	The month of the year in which the accident took place.
AccidenCauseCategory	The general category of the cause of the accident.
AccidentCause	The more specific cause of the accident.
PolicyStatus	The policy status of the insured at the time of the accident.
LiabilityStatus	The status concerning the determination of liability for the accident.
LiabilityPercentage	The percentage of liability of the insured.
ClaimedObjectType	The general type of object corresponding to the claim.
ClaimedObjectSubType	The more specific type of object corresponding to the claim.
ClaimedObjectBelongsToClient	If the object corresponding to the claim belongs to the client or not.
EventCountry	The country where the accident took place.
FirstNoification	The person or institution that made the first notification about the claim.
BusinessDaysUntilNotification	How many business days passed between the accident and it being notified.
OwnDamage	If the object damaged belongs to the insured or not.
ThirdPartyDamage	If the object damaged belongs to a third party or not.
OwnInjury	If the injury was sustained by the insured or not.
ThirdPartyInjury	If the injury was sustained by a third party or not.
VehicleMake	The vehicle corresponding to the claim main's brand.
VehicleTotalLoss	If the vehicle corresponding to the claim was totally lost or not.
IsAssesor	If an assessor got involved in the claim or not.
IsBroker	If a broker got involved in the claim or not.
IsCourt	If the claim went to court or not.
IsHealthInsurance	If a health insurance covered (totally or partially) the cost of the injury or not.
IsIntermediary	If an intermediary got involved in the claim or not.
IsJuridicalExpert	If a juridical expert got involved in the claim or not.
IsMedicalExpert	If a medical expert got involved in the claim or not.
IsHospital	If a hospital got involved in the claim or not.
IsPolice	If the police got involved in the claim or not.
IsRepresentative	If a representative got involved in the claim or not.
IsSolicitor	If a solicitor got involved in the claim or not.
IsVA_Agent	If a veteran assistance agent got involved in the claim or not.

ClientName_Freq	The number of claims handled with a client.
Reserve	The reserve set ahead of time to cover the costs of a claim.

SUBCATEGORY SELECTION

15 variables are categorical, 18 are binary, and 5 are numeric, totaling 38 variables. More so, some of the categorical variables have high cardinality. For example, the variable *VehicleMake* had 135 subcategories, as there are lots of car makes that are involved in accidents. Another example is the variable Claime*ObjectSubType* which had 76 subcategories. Not all subcategories were included in the models as that would have made the training and the hyperparameter tuning very computationally heavy and the computation time would have been excessive. Also, some of the subcategories are only represented in a handful of observations which are not sufficient to assess the effect of that specific subcategory on the outcome variable. For these reasons, subcategory selection was carried out using a combination of criteria: the most frequent subcategories, the subcategories that have the highest and lowest claim cost mean, the subcategories that have the highest and lowest claim cost median, the subcategories that were the most significant according to the variable selection carried out by the LASSO model, and research-oriented subcategories, such as the most expensive car brands according to Global Cars Brands (Vroom, 2021). Therefore, it can be summarized that the LASSO was trained under all of the subcategories, as its penalization term allows for it, but the Random Forest and the CatBoost models were trained after subcategory selection was performed. Tables on exploratory analysis between the variables with high cardinality and the outcome variable can be seen in Appendix A.

V. RESULTS

As mentioned previously, claims can remain open for many months until costs are confirmed and settled. During this period of time, reserves are updated to adjust for new information about the claim that was recently attained but was not previously known by the claim handler.

From the utilized dataset, claims were open on average for 8.95 months. 79.9% of the claims were closed within 1 year, and 96.9% within 3 years. The timeframe analyzed was of 3 years.

This means that, for each ML algorithm (Radom Forest and CatBoost) the number of models trained was the following:

- 1 only AI model without the variable Reserve.
- 36 AI + Human Expert collaboration models with the variable Reserve being updated at every month for a timeframe of 3 years.

Therefore, in total 74 models were trained in this subsection of the results to analyze first, the performances of the models as time goes by in the claim handling process, and second, the relative importance of the variable *Reserve* (which represents the Human Expert input) to the other variables (which represents AI input) in this setting. Furthermore, the hyperparameter of these models were not tuned as the purpose of this subsection was not to achieve a higher predictive performance but to assess how the models change across time. In the case of tuning, difference in results could have been attributed to differences in tuning. Instead, 500 trees/iterations and default parameters of the corresponding packages were used. In the case of the Random Forest, the *ranger* package (Wright et al., 2022) in *R* was used. The default *mtry* is the square root of the number of independent variables. For the CatBoost, the *CatBoost* package (Prokhorenkova et al., 2019) was used, and its default parameters are a *depth* of 6, a *learning_rate* of 0.095, a *12_leaf_reg* of 3, and a *random_strenght* of 1.

For all of the models ahead, the dataset was split, randomly and without repetition, into a 60% subset corresponding to the train dataset and a 40% subset corresponding to the test dataset. The metrics used to evaluate the performance of the models (RMSE and MAE) refer to the test dataset.

SOLELY AI MODELS

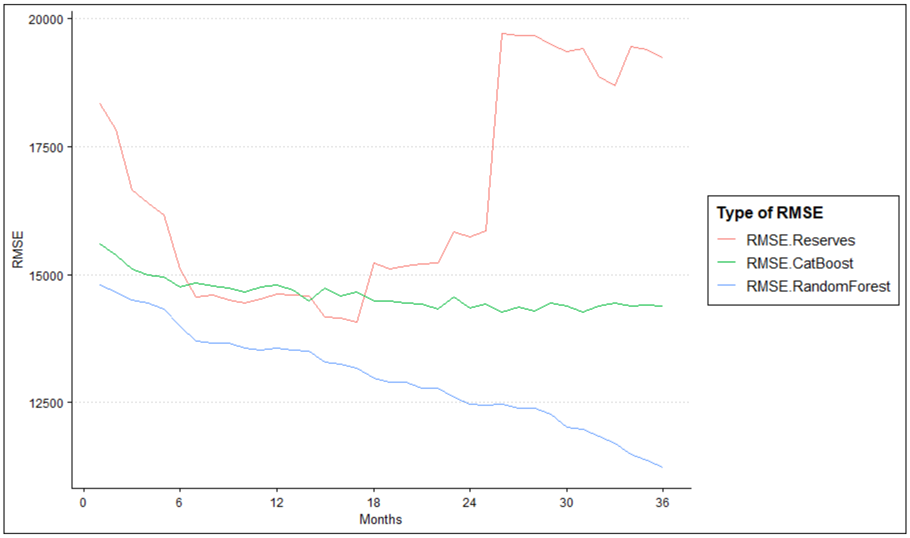
The first two models (one Random Forest algorithm and one CatBoost algorithm) were trained using 37 independent variables available at the early stages of the claim handling process, including information about the accident itself, such as the cause of the accident; information about the objects and parties involved, such as the vehicle brand and the possible involvement of juridical or medical experts; and information about the location and time, such as the country and month of the year. However, estimations on the reserves made by expert claim handlers were not included as predictors. More specifically, these models used the variables mentioned in the Data section of this study, with the exception of the variable *Reserve*.

When it comes to performance, the Random Forest yielded a Root Mean Square Error (RMSE) of 16,704 and a Mean Absolute Error (MAE) of 2,603. The CatBoost had a RMSE of 17,030 and a MAE of 2,963. These numbers do not indicate much by themselves and thus will be compared later to the error terms of the models that did include the variable *Reserve*. Yet, the RMSE and MAE generated between the actual costs and the mean of the actual costs was used as a preliminary benchmark. This was found to be 18,006 and 3634 for the RMSE and the MAE respectively. Thus, even without the variable *Reserve*, the ML predictions do provide a decrease in the error when compared to this benchmark.

AI + Human Expert Collaboration Models

Predictions were obtained for all the following models which did include the variable Reserve along with all the other variables (36 Random Forest algorithms and 36 CatBoost algorithms as this part of the analysis compromised 36 months). First, a comparison was gathered between the actual confirmed cost of the claims and the value of the reserves (at each given month) versus the actual cost of the claims and the predictions of the models (at each given month), to determine if using AI does improve predictions. To test this, the metrics used were the RMSE and the MAE. During the first month the RMSE of the actual costs and the reserves was of 18,334, while that of the actual costs and the predictions of the models trained for that month were of 14,791 for the Random Forest and 15,601 for the CatBoost. Similarly, for month 2 the RMSE values were respectively 17,832, 14,659 and 15,380, and for month 3 they were 16,666, 14,496 and 15,115. The values for all of the months within 3 years can be seen in Figure 5.

FIGURE 5. RMSE OF THE RESERVE AND MODELS AS TIME PROGRESSES

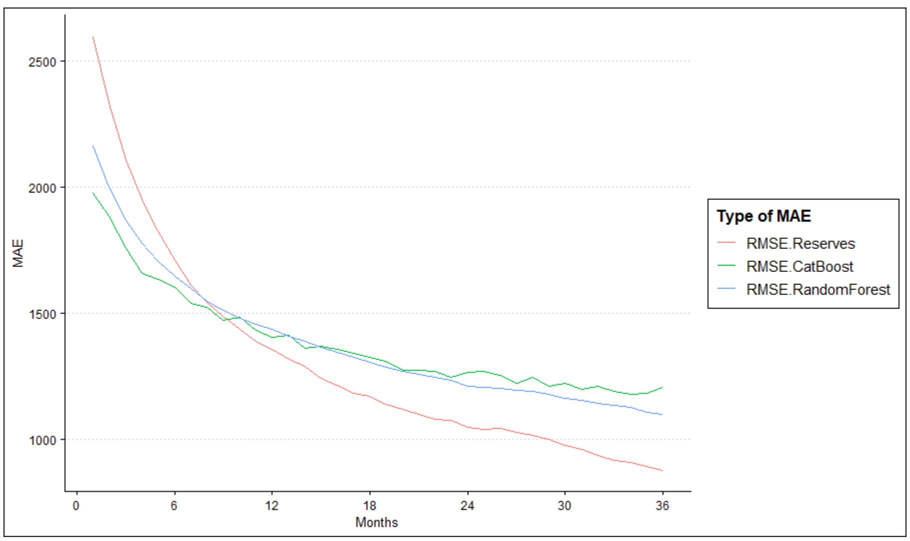


In all the months, the error term was lower when implementing a ML algorithm, with the exception of month 7, 8, 9, 10, 11, 12, 13, 15, 16, and 17 when it comes to the CatBoost. The RMSE of the Random Forest was the smallest during all of the months. This denotes the resourcefulness of having a complex tool that takes in data and finds relationships between the variables in forms that are not evident to people, even to expert claim handlers with years of experience. Within the 3-year timeframe, the RMSE of the reserves decreased until reaching a minimum of 14,074 at month 17. After that, it increased again reaching a maximum of 19,706 at month 26. When it comes to claims, usually the longer it remains open the higher its cost, therefore reserves that are modified at the later months of the timeframe comprise high-cost claims. Since the RMSE metric punishes large error values, and error values are larger in high-cost claims, this might be a reason behind the increase of RMSE of the reserves at the later months.

Regarding the MAE, it decreases both for the reserves and the predictions as months go by, new information about the claim arrives, and reserves are modified, as expected. The moment in time in which the error become smaller for human expert reserves versus model predictions occurred at month 8 for the Random Forest, point in which the MAE of the reserves was 1,544 and that of the predictions was slightly higher at 1,548; and at month 10 for the CatBoost, moment in which the MAE of the reserves was 1,438 and that of the predictions was 1,483. Hence, the CatBoost had a smaller MAE than the reserves for a longer duration within the timeframe than the Random Forest.

According to this error term, at month 9 it is better to use the reserves as an estimate of the cost than the predictions of the Random Forest, but it is still better to use the predictions of the CatBoost above the other two. Before these thresholds, the error was consistently smaller when implementing a ML algorithm. At month 1 the MAE of the actual cost and the reserves was 2,595, while that of the actual costs and the predictions of the model trained for that month was of 2,162 for the Random Forest and of 1,976 for the CatBoost. At month 2, these values were 2,335, 2,002, and 1,886 respectively, and at month 3 they were 2.106, 1.867, and 1,748. At the last stage, in month 36, the MAE of the reserves was 878 while that of the predictions were 1,102 and 1,208 for the Random Forest and CatBoost. The values for all the months within 3 years are found in Figure 6.

FIGURE 6. MAE OF THE RESERVE AND MODELS AS TIME PROGRESSES



Moreover, it can be seen that as time progresses, the distance between the ML algorithm lines and the reserves lines narrows down until they intercept at month 8 and 10, and then they widen out. The difference in the error between not having AI and implementing AI decreases as the reserves become more accurate representations of the final cost, and then the reserves actually become more accurate than the predictions, at least while using the MAE as the performance metric. Finally, it can also be noted that the error of the CatBoost was smaller than that of the Random Forest at the beginning, but then ended up the other way around. It was at month 10 that the performance of the

Random Forest caught that of the CatBoost and then it continued to perform with a smaller MAE until the end of the timeframe.

Theoretically, it can be expected that at month 132 (when the last claim is settled), the error (both RMSE and MAE) of the actual costs and the reserves would be 0 (or very close to 0), as the reserves would be identical (or very close to identical) to the actual costs. In a similar manner, it can be understood that when the reserves are identical to the actual costs, all other independent variables become irrelevant. Of course, at this stage, using AI becomes pointless. Using AI at a late stage in the claim handling process would also be not recommended. Under the assumption that predictions are done early or relatively early, since that is the purpose of predicting in the first place, the results are consistent with the literature mentioned on the topic (Blier-Wong et al., 2021), (Levantesi et al., 2020), (Severino & Peng, 2021), (Wang & Wu, 2018) which indicates that AI improves results in insurance and insurance claim handling. During all of the early months, the reserves estimated by humans solely have a larger error than when AI is also implemented.

Moreover, if Human Expert contribution to the predictions is represented by the variable *Reserve*, then AI contribution to the predictions is represented by all other independent variables. If predictions are done early, while human experts have low knowledge about the claim, then naturally AI predictors become comparatively more relevant. On the other hand, if predictions are done late, when human experts have significantly more knowledge about the claim including information about actual costs, then the opposite happens, and the predictor *Reserve* becomes comparatively more relevant. These values can be observed in Figures 7 and 8. At month 1 the variable *Reserve* accounts for 24% of the total variable importance for the Random Forest, and 22% for the CatBoost. For both algorithms, this relative importance continues to grow. Although the growth is more pronounced in the Random Forest. Moreover, at what month does Human Expert contribution become larger than AI contribution when predicting? The threshold occurs at month 25 for the Random Forest, in which the variable importance for the variable *Reserve* becomes larger than half of the total for the first time in the timeframe, with a value of 50.4%. This means that from month 25, the estimations done by the expert claim handlers contribute more to the prediction of the costs than all the other variables combined. Yet, this never happens for the CatBoost algorithm within the 3-year timeframe. Even though, *Reserve* is still the most important variable for the CatBoost, its importance, as percentage of the total importance, never reaches 50%. The highest value occurs at month 31 and 36 with 40.9%.

FIGURE 7. RANDOM FOREST: RELATIVE IMPORTANCE OF THE VARIABLE RESERVE VERSUS ALL THE OTHER VARIABLES ACROSS TIME

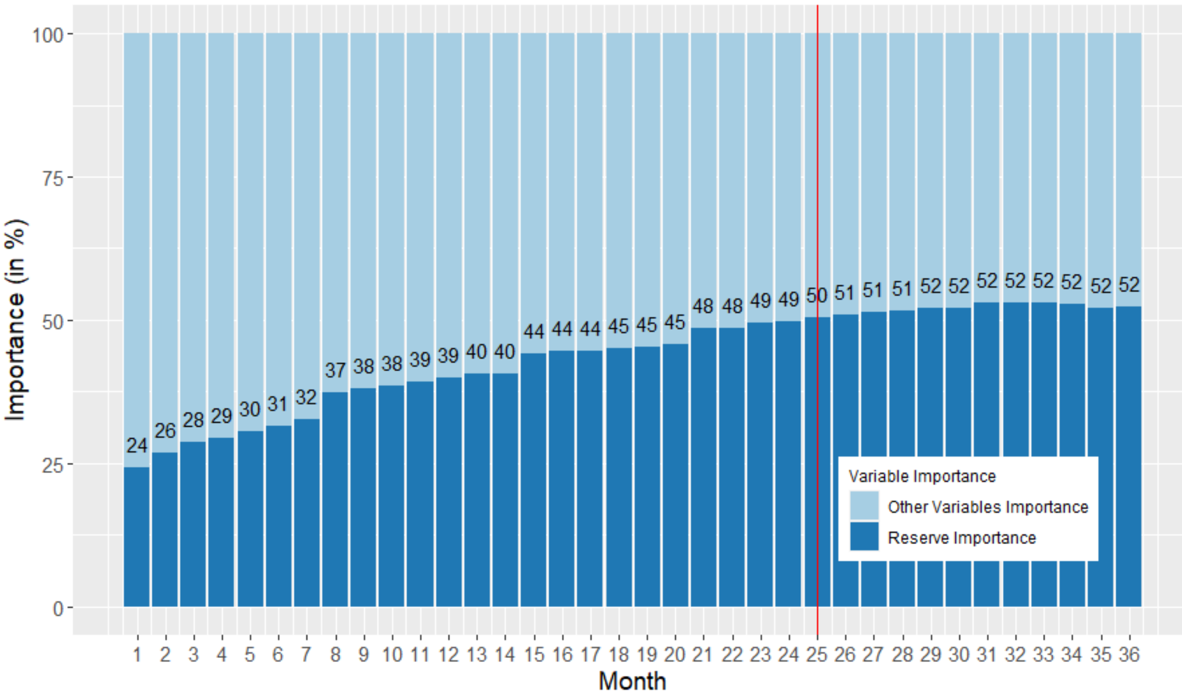
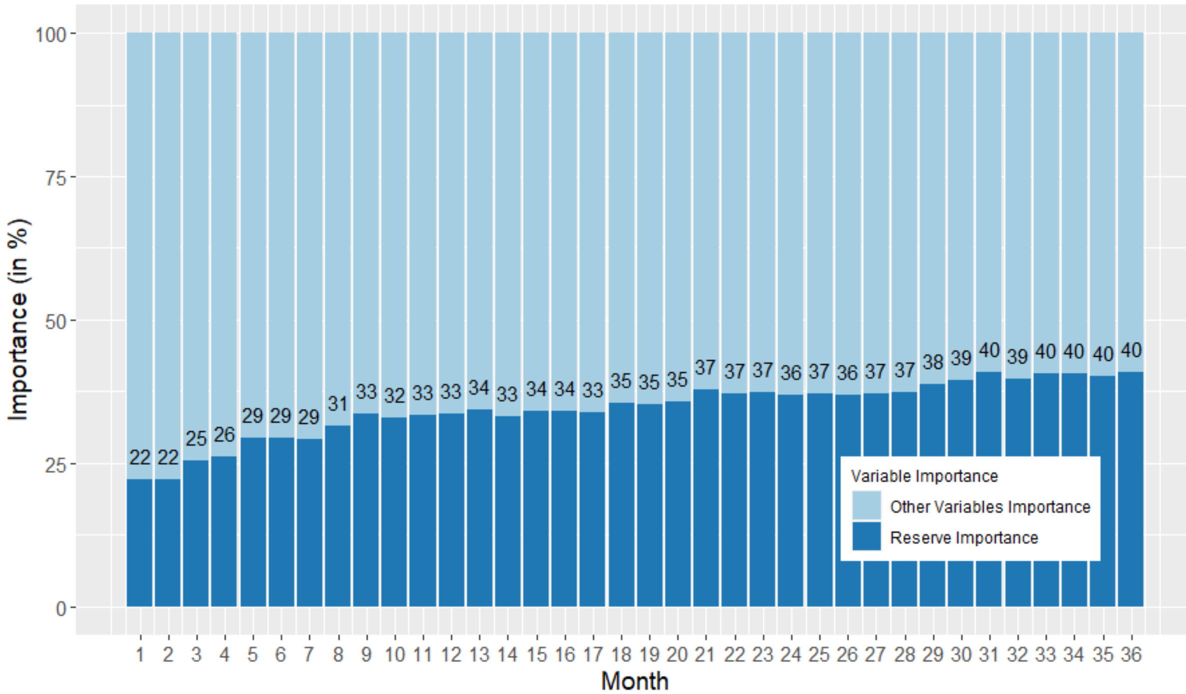


FIGURE 8. CATBOOST: RELATIVE IMPORTANCE OF THE VARIABLE RESERVE VERSUS ALL THE OTHER VARIABLES ACROSS TIME



IN-DEPTH ANALYSIS OF THE MODELS AT MONTH 1

Continuing with the analysis, the idea of a predictive analysis in this setting is to make predictions at an early stage of the claim handling process. Therefore, the analysis was expanded at month 1 to find first, the best performing model and second, to interpret the independent variables used and their relationship with the outcome variable cost[5].

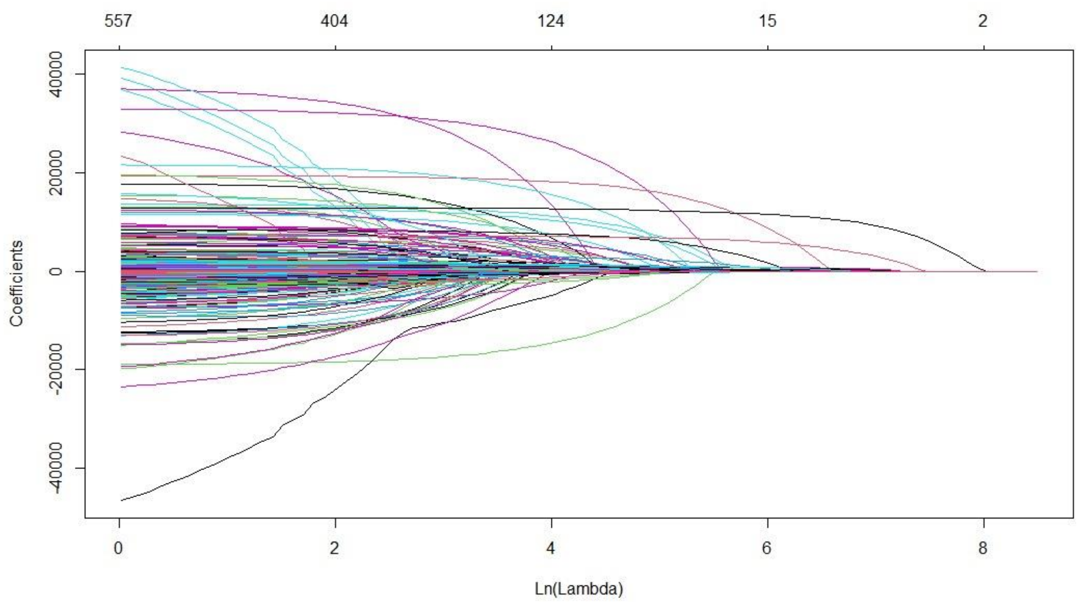
LASSO

A LASSO regression was used as a first model. An optimal lambda (λ) of 15 was determined through 10-fold cross validation. With this setting, the model conserved 308 variables with a non-zero β coefficient from 636 possible variables. This means that it shrank to zero the coefficient of

[5] The hyperparameters of the models in this subsection were carefully tuned using the grid search technique, as compared to the 72 models of the previous subsection which were not tune, as mentioned.

328 variables, from which a large proportion were binary corresponding to different insurance company clients, from the *ClientName* variable; different insurance coverages, from the *Coverage* variable; different accident causes, from the *AccidentCause* variable, different objects claimed in the accident, from the *ClaimedObjectSubType* variable, different vehicle brands from the *VehicleMake* variable, and different countries from the *EventCountry* variable. The shrinkage can be observed in Figure 9.

FIGURE 9. LASSO COEFFICIENT SHRINKAGE



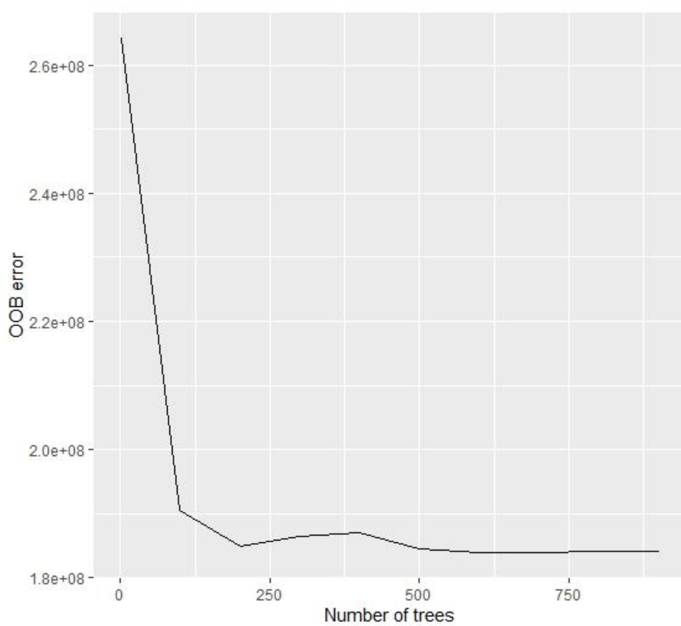
Interesting insights from the LASSO can be obtained from some of the variables with the largest coefficients. Among these are when there is a juridical expert involved (12,805), when the accident involves an Aston Martin (5,426) or a Donkervoort (9,131) car, when the accident takes place in Ireland (8,122), when the object claimed is a train (19,806), when the accident is caused by a staff member (15,053), when it is a robbery (-12,682), when the accident is caused by uncommon forces of nature (9,925), when the cause of the accident is an overturn (10,503), when the cause of the accident is the condition of the road (12,917), when the accident is a collision with an animal (-7828), when the insurance coverage is *Third Party Liability* (31,127), and when the insurance coverage is *Comprehensive Cover Extra* (30,712).

The performance of the LASSO was also assessed with an out-of-sample test using the RMSE and MAE as performance metrics. These were found to be 14,946 and 3,106 respectively.

RANDOM FOREST

Next, the analysis moved into black-box Machine Learning algorithms. The first black-box model used was a Random Forest. The Random Forest hyperparameters were determined. As seen in Figure 10, the Out-of-bag error decreases very steeply just before 125 trees; however, it also decreases slightly just before 500 trees. Thus, the number of trees used was 500. Additionally, the number of variables to be randomly selected for each split (*mtry*) was 6. The square root of the Out-of-bag error was 13,594.

FIGURE 10. RANDOM FOREST: OUT-OF-BAG ERROR ACROSS NUMBER OF TREES



The RMSE of the Random Forest was 13,735, which is very similar to the Out-of-bag error. This indicates a good fit to the data. The MAE was 1,885. It can be seen that the Random Forest outperforms the LASSO.

CatBoost

The second black-box Machine Learning algorithm used was a CatBoost algorithm. Boosting algorithms like this one allow for more hyperparameter tuning (but also depend more on it). Due to this process being computationally heavy and time consuming, the grid search carried out to find the optimal hyperparameters followed a two-step consecutive zoomed in grid search: first a random grid search with a broader range of hyperparameters to test iteratively; and second a zoomed in grid search with a narrower range of hyperparameters to test iteratively based on the results of the random grid search. The results can be seen in Table 2.

Table 2. CatBoost: two-step consecutive grid search results for hyperparameters

depth	11
12_leaf_reg	9
iterations	700
random_strength	0.5
learning_rate	0.12

After considerable hyperparameter tuning, the performance of the CatBoost improved notoriously with regards to the MAE, but not as much with regards to the RMSE. The resulting error terms were a MAE of 1,699 and a RMSE of 15,711.

Table 3 ranks the performance of the three models used based on three criteria: first, the lowest RMSE; second, the lowest MAE; and third, the lowest computational resources and time to train and tune. The RMSE of the LASSO is slightly lower than that of the CatBoost but its MAE is significantly higher compared to both the Random Forest and the CatBoost, yet it is the fastest to train and tune. The CatBoost is by far the one that takes the longest time to tune. The Random Forest ranks number 1 on the lowest RMSE, while the CatBoost ranks number 1 on the lowest MAE. However, the CatBoost ranks last on both the RMSE and the training time. Thus, it is the Random Forest that is selected as the best performing model in the analysis. It must also be noted though that all of these models perform better than the solely AI models discussed in the beginning of the Results section, except for the MAE of the LASSO. Undoubtedly, including the variable *Reserve* into the ML algorithms improves their performance.

TABLE 3. RANKING OF THE ALGORITHMS USED BASED ON RMSE, MAE, AND TRAINING/TUNNING TIME

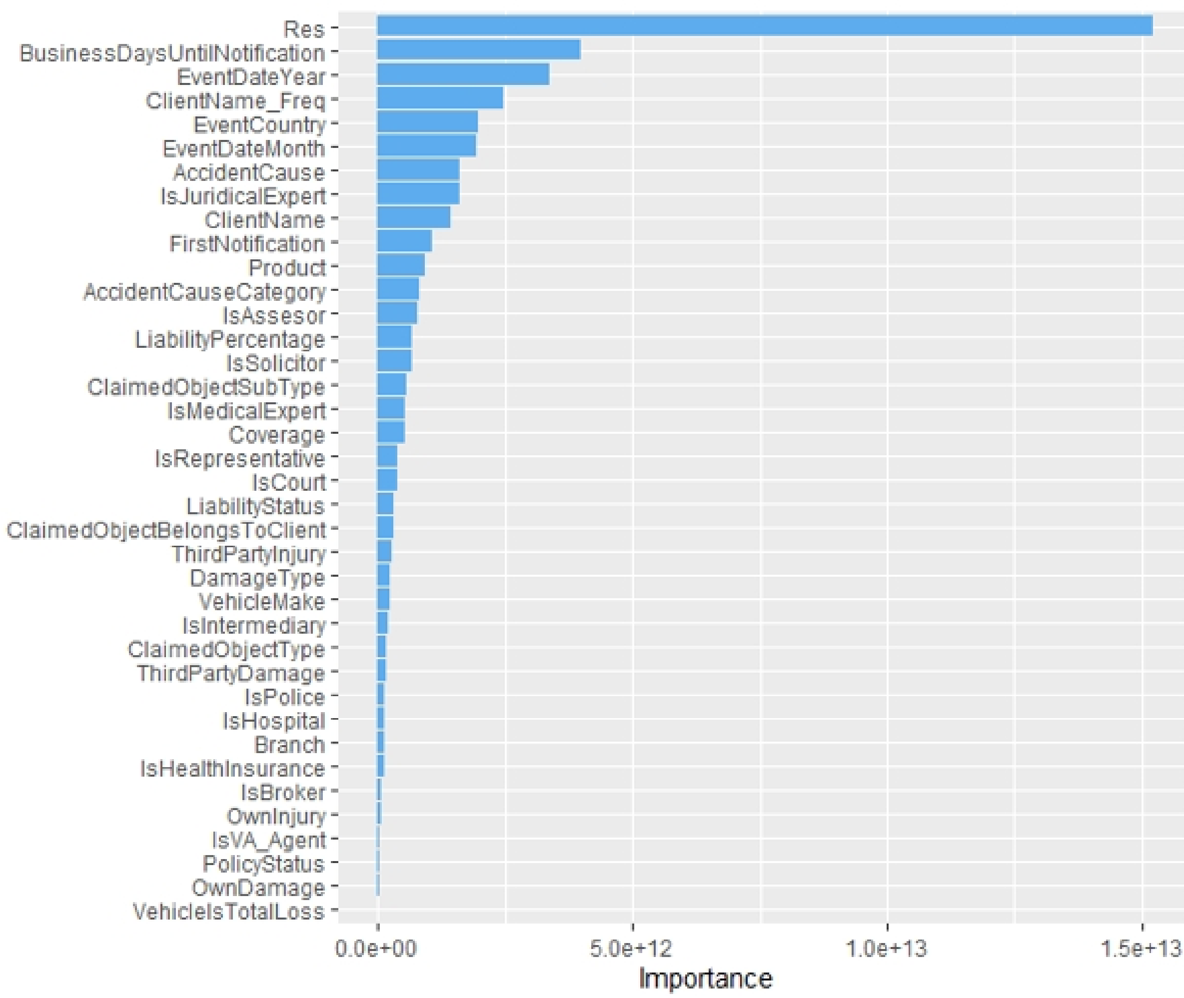
	RMSE	MAE	Training and tuning time
LASSO	14,946 (2)	3,106 (3)	1
Random Forest	13,735 (1)	1,885 (2)	2
CatBoost	15, 711 (3)	1,699 (1)	3

INTERPRETATION OF THE VARIABLES IN THE BEST PERFORMING MODEL

Once a best performing model was selected, in this case the Random Forest from the previous subsection, interpretation on the effects of the independent variables on the dependent variable cost was done.

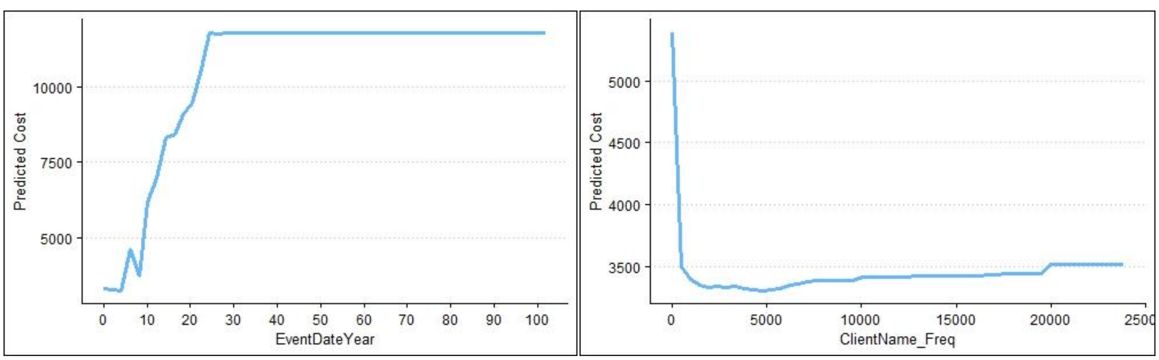
From the Variable Importance Plot (Figure 11) below, it can be observed that besides from the variable Reserve, which was already discussed about previously, other variables are the top ten most relevant for the model to make predictions on the cost. These are the business days until an accident is notified from the moment it occurred, the number of years passed since the accident took place, the number of claims of the insurance companies that are clients of Van Ameyde, the country, the month of the year, the cause of the accident, the possible involvement of a juridical expert in the accident claim, the insurance company (highlighting that the costs are not uniform across insurance companies), the source from which the first notification arrives, and the type of insurance product. On the other hand, variables that are not relevant for the model include whether the vehicle was a total loss (this specific variable had the highest number of missing values which explain the lowest position in the variable importance plot when actually it could have been a rather important variable since it gives information on the severity of the accident), if the claimed object that was damaged was owned by the insured or by the third party, the policy status, the possible involvement of agents specialized in assisting veterans, and if the injury was sustained by the insured or by the third party. It is important to mention as well that the effect of some variables may be captured by other similar variables, which is something a Random Forest is capable of handling when making predictions, but still must be considered when making interpretations. Some examples may be the variables *IsMedicalExpert* and *DamageType*, it is expected that medical experts only get involved when the damage type is an injury. Another example is the *LiabilityStatus* and *LiabilityPercentage*. *LiabilityStatus* indicates the stage in the process of determining liability, whether liability has been acknowledged or not, or whether it is in dispute, etc. *LiabilityPercentage*, thus, depends on *LiabilityStatus*. Nonetheless, calculating variable importance does provide useful markers to understand better how the model makes predictions.

FIGURE 11. RANDOM FOREST: VARIABLE IMPORTANCE PLOT



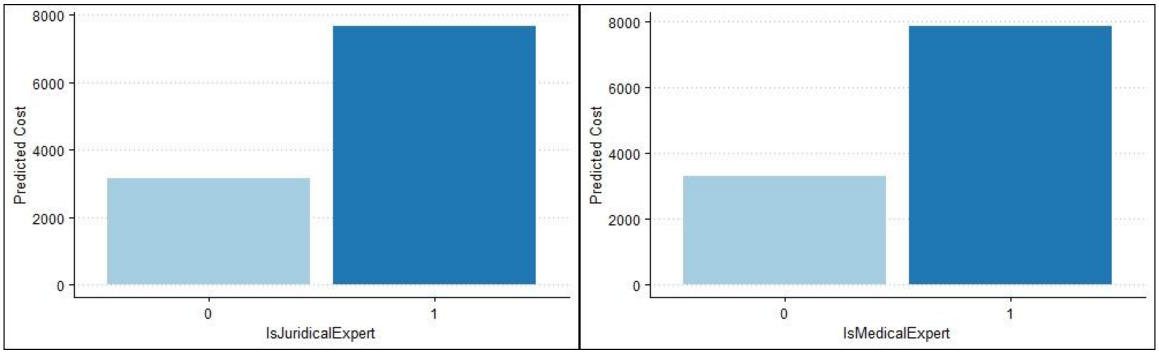
Compared to Guelman´s (2012) study in which the variables related to the vehicle were among the most important, the results of this study point into a different direction as the variables related to the vehicle or the object involved in the accident such as *VehicleMake* or *ClaimedObjectSubType* are shadowed by other more important variables. These other variables were not included in Guelman´s study. However, other variables such as the vehicle age were not included in this study. While a one-on-one comparison cannot be drawn, findings suggests that some of the most relevant information is not necessarily related to the vehicle itself but to settings about the accident or surrounding the accident, such as the timing, location, and parties involved. Variables related to the policy or coverage have a medium importance in both studies. Maillart´s (2021) study found that location is an important variable in driver´s insurance claims. This study provides similar results as the variable *EventCountry* is among the top five most important. That being said, the variable importance plot does not indicate the direction of the effect of the variables on the cost. This can be assessed with a Partial Dependence Plot (PDP). PDPs of some important or interesting variables were drawn ahead.

FIGURE 12. PDP: EVENTDATEYEAR (LEFT) & PDP: CLIENTNAME_FREQ (RIGHT)



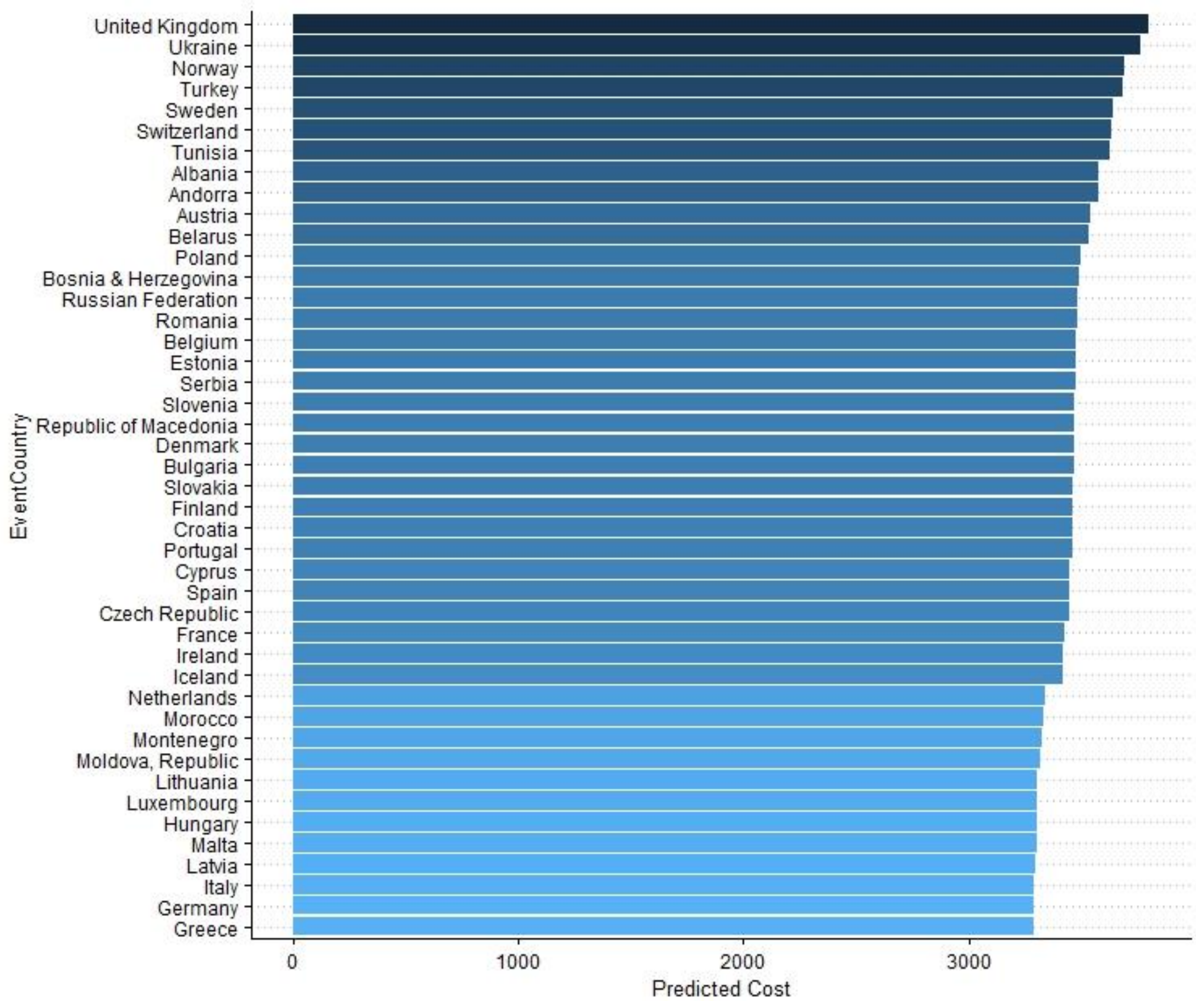
The PDP plot about EventDateYear (left) indicates how the prediction of the costs vary as the number of years since the accident occurred increase. There is not a large difference on the cost of claims when the accident took place between 1 and 4 years ago. However, there is a sharp rise at around 9 years which stabilizes at around 25 years. If a claim is open for so long then most likely there is some type of disability and/or loss of the ability to work, which first is an ongoing cost and second it significatively increases the cost of the claim. The PDP on the right suggests that the cost from claims of infrequent clients (insurance companies that have only handled a handful of claims with Van Ameyde) are notoriously higher than those of frequent clients (that have handled thousands of claims).

FIGURE 13. PDP: ISJURIDICALEXPERT (LEFT) & PDP: ISMEDICALEXPERT (RIGHT)



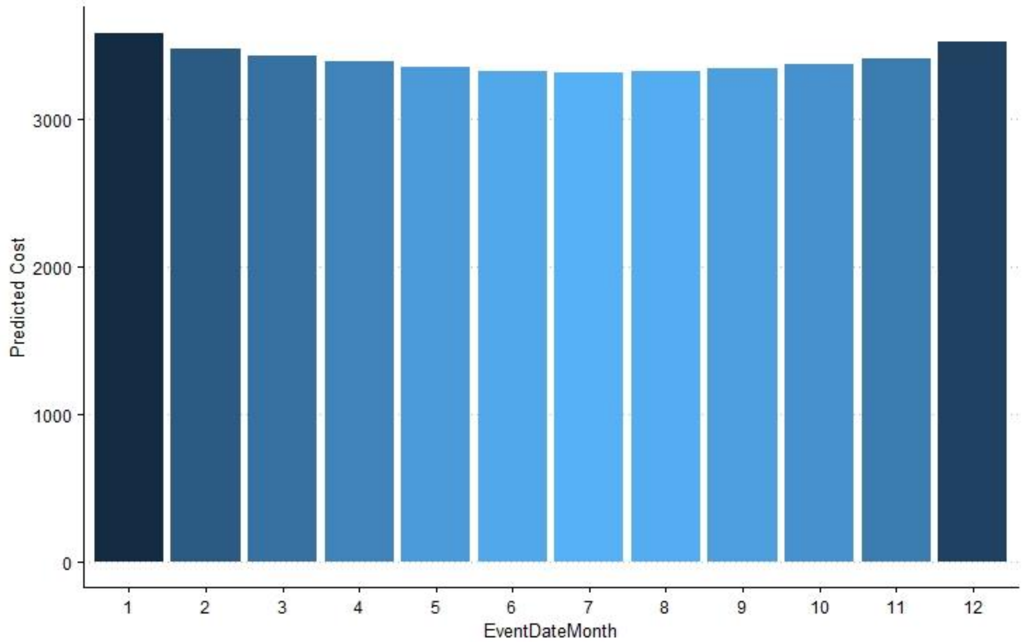
The two PDPs above (Figure 13) are very similar to each other. Both, when there is a juridical expert or a medical expert involved in the claim, the average predicted cost increases in a similar proportion.

FIGURE 14. PDP: EVENTCOUNTRY



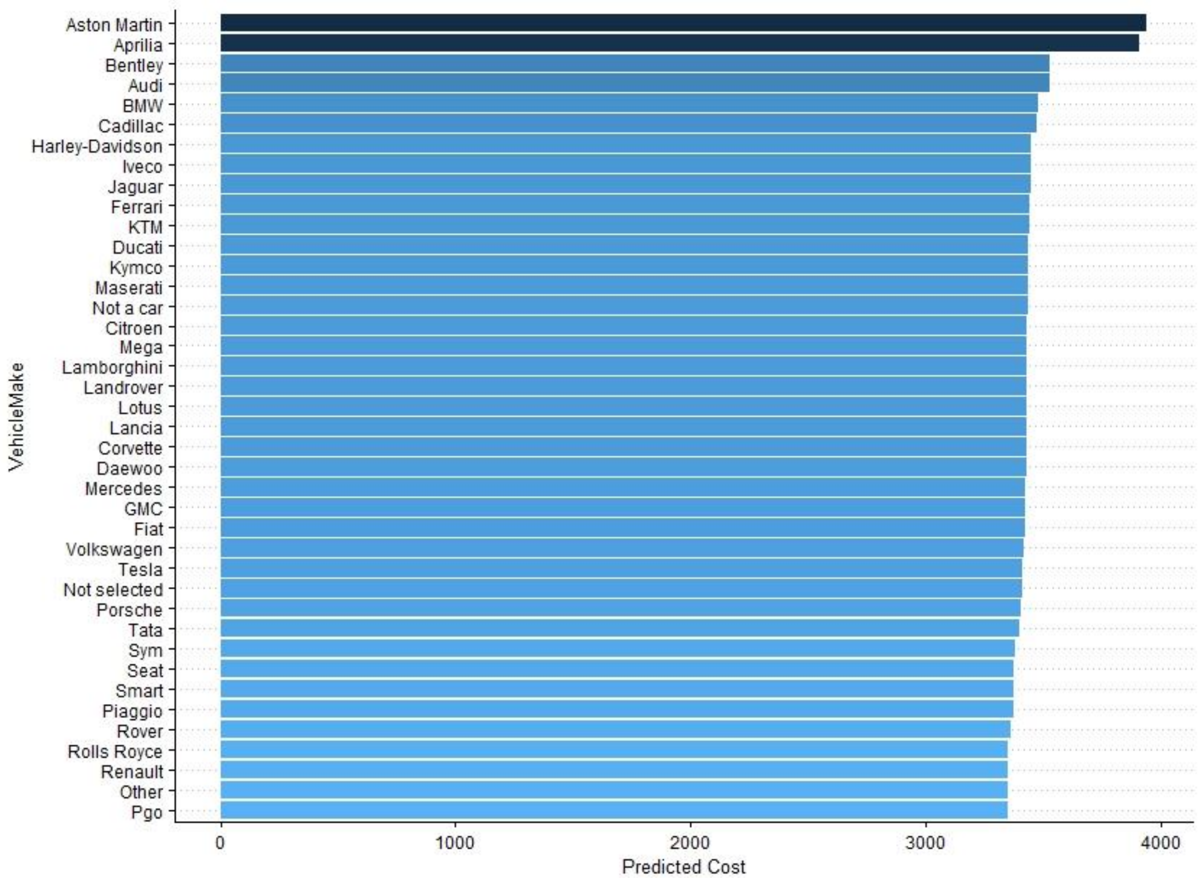
The countries that on average lead to higher cost predictions are the United Kingdom, Ukraine, Norway, Turkey, Sweden, and Switzerland. On the other hand, Greece, Germany, Italy, Latvia, and Malta tend to generate lower cost predictions. The Netherlands is also positioned in the bottom half of the graph.

FIGURE 15. PDP: EVENTDATEMONTH



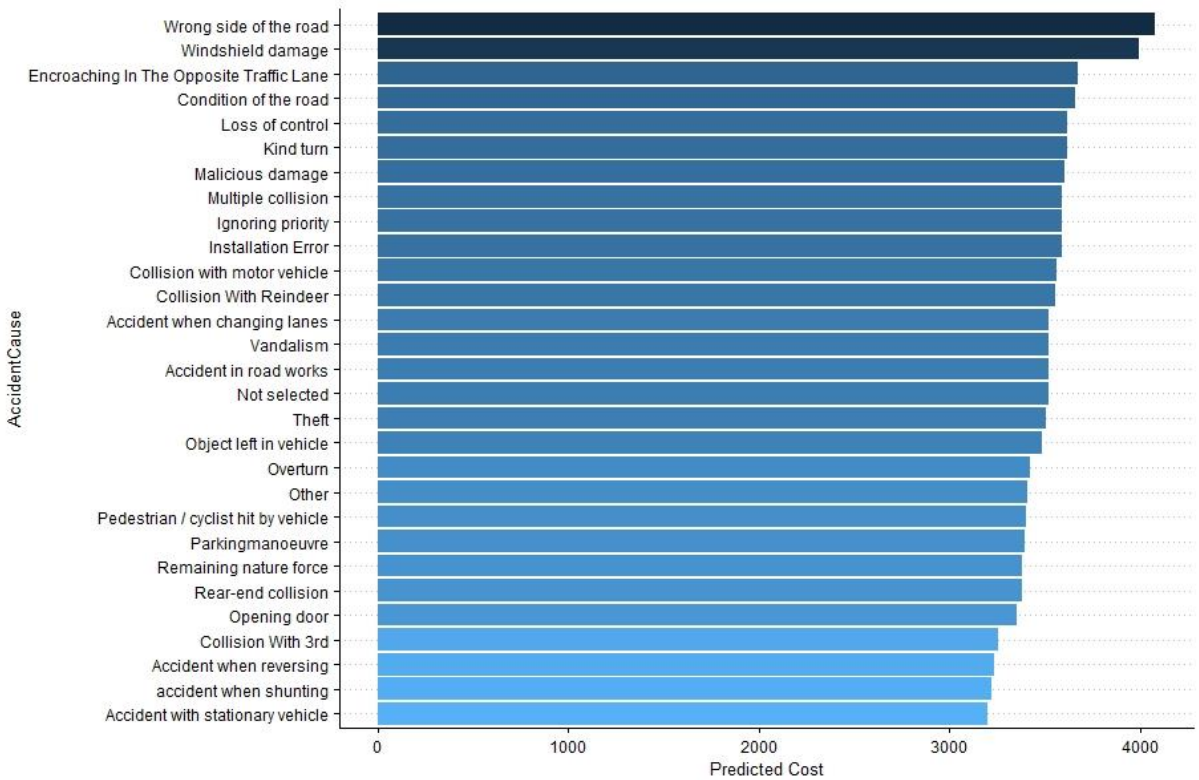
Accidents that occur towards the beginning/end of a year, between December and January seem to lead to higher cost claims. Moreover, the costs seem to progressively decrease reaching a minimum in summer around July and the progressively increase again.

FIGURE 16. PDP: VEHICLEMAKE



When it comes to vehicle make, there are two brands that are clearly above the rest. These are Aston Martin and Aprilia. These results are interesting since, while an Aston Martin is among the most expensive cars in the world and hence its repair is expected to be expensive, Aprilia is mostly a scooter and motorcycle brand. The reason why an Aprilia might be so high in the charts is because of potential bodily injuries. These can be expected to be more severe when there is a scooter or motorcycle involved instead of a car.

FIGURE 17. PDP: ACCIDENTCAUSE



The accident leads to a higher cost claim when the cause is being on the wrong side of the road or when encroaching the opposite traffic lane, which possibly leads to a frontal impact between vehicles. A windshield damage possibly causes limited visibility and also generates a high-cost claim on average. Lower cost claim accident causes include stationary vehicles and similar, such as an accident when reversing or opening a door.

To summarize, these are some of the results gathered about the performance of the different algorithms used when predicting Green Card System driver's insurance claims as well as how aspects such as variable importance vary across time when knowledge of the claim increases. It also goes into interpreting how other variables related to the *what (DamageType, ClaimedObjectSubType)*, *where (EventCountry)*, *why (AccidentCause)*, *when (EventDateYear, EventDateMonth)*, and *who (ClientName, IsJuridicalExpert)* of the accident have an effect on the final cost.

VI. CONCLUSION, LIMITATIONS, AND FUTURE RESEARCH

This study sought to make use of very interesting and difficult to come by data to analyze aspects relevant within insurance, more specifically Green Card System driver's insurance. It also has a broader relevance since it analyzed potential manners in which AI can collaborate with Human Experts and how this collaboration leads to better results. Additionally, it investigates how the interaction between the two parts (AI and Human Experts) is dynamic and might vary across time as knowledge or information is updated.

When Human Experts are given the task to predict an outcome based on their experience, or previous knowledge, or instinct, or calculations they make, etc. It is a good idea to use this as an additional predictor in a Machine Learning algorithm along with other variables to potentially obtain better predictions.

While this study initially had the idea of 'seeing how AI predictions are better than claim handler estimation of reserves' it was soon after the start of the study that the focus shifted and became 'how can AI be used along claim handler estimation of reserves to better predict ahead of time the cost of a claim.' This shift from AI versus Human Expert to AI plus Human Expert was a breakthrough in the study and ultimately allowed for very interesting insight. This can certainly be replicated in lots of setting, even in different industries and fields.

Moreover, regarding the predictions, the best performing model turned out to be a Random Forest algorithm above a LASSO regression or a Boosting algorithm in the form of Categorical Boosting (CatBoost). Global Interpretation Methods allowed to determine pointers within the variables that usually lead to higher or lower cost claims. As an illustrative example, one of the most expensively possible claims for Van Ameyde (according to the model) would be from an accident involving an Aston Martin (even more so than a Lamborghini or Rolls Royce) cause by encroaching the opposite traffic lane, in the United Kingdom, in July, in which a lawyer was involved, there was bodily injury and a medical expert was involved, in which there was a delay between the accident taking place and the claim being notified, and it came from an infrequent client. Of course, other variables also come into scene, and there are also interaction effects between the variables, but having some elements of interpretability do help especially in a business context.

The study also faced some important limitations. Even though the models provide improved predictions, they did lack information within the training dataset. For instance, no variable captured information about the severity of an injury. There was a variable that determined if there was an injury or not, but it was not known if the injury referred to a broken finger or a pierced lung. Other variables that could have been very useful such as *VehicleTotalLoss* were not consistently recorded in the system by the claim handlers. This highlights the importance of establishing a stringent data recording process. If some box fields in the system are optional when registering a claim, most claim handlers will leave that box field empty. It is recommended to make all fields compulsory to register in the system.

Other limitations included the vast number of subcategories within the categorical variables. Any form of subcategory selection implicates some degree of loss of information. With more computational resources, maybe more subcategories could have been included to find more specific insights. Further research can test different criteria of subcategory selection or variable encoding. Further research could also expand the timeframe of 3 years.

Other studies can expand on the Human Expert information and include variables about the Human Expert such as years of experience of the claim handler, approach taken (optimistic approach, conservative approach, best estimate). Maybe the Human Expert proxy variable improves the model under certain characteristics, or in a greater/lesser degree.

To conclude, predicting insurance claim costs is a complex topic and thus any collaboration between AI and Human Experts is welcome in a business context (for a company such as Van Ameyde), but also in an academic context, since the world is still discovering how AI will work with humans and how does the outlook of a fourth industrial revolution or a technological revolution look and what implications it will bring.

VII. REFERENCES

Adewumi, A. O., & Akinyelu, A. A. (2017). A survey of machine-learning and nature-inspired based credit card fraud detection techniques. *International Journal of System Assurance Engineering and Management*, *8*(S2), 937–953. https://doi.org/10.1007/s13198-016-0551-y

Al-Saedi, A. A., Boeva, V., Casalicchio, E., & Exner, P. (2022). Context-Aware Edge-Based AI Models for Wireless Sensor Networks—An Overview. *Sensors*, *22*(15), 5544. https://doi.org/10.3390/s22155544

Baudry, M., & Robert, C. Y. (2019). A machine learning approach for individual claims reserving in insurance. *Applied Stochastic Models in Business and Industry*, *35*(5), 1127–1155. https://doi.org/10.1002/asmb.2455

Bernard, P.I., Binder, S., D´Amico, A., de Combles de Nayves, H., Ellingrud, K., Kotanko, B., Klais, P., & Strovink, K. (2022). *Creating value, finding focus: Global Insurance Report 2022 | McKinsey*. https://www.mckinsey.com/industries/financial-services/our-insights/creating-value-finding-focus-global-insurance-report-2022

Blier-Wong, C., Cossette, H., Lamontagne, L., & Marceau, E. (2021). Machine Learning in P&C Insurance: A Review for Pricing and Reserving. *Risks*, *9*(1), 4. https://doi.org/10.3390/risks9010004

Boukerche, A., & Wang, J. (2020). Machine Learning-based traffic prediction models for Intelligent Transportation Systems. *Computer Networks*, *181*, 107530. https://doi.org/10.1016/j.comnet.2020.107530

Braver, E. R., & Trempel, R. E. (2004). Are older drivers actually at higher risk of involvement in collisions resulting in deaths or non-fatal injuries among their passengers and other road users? *Injury Prevention*, *10*(1), 27–33.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chui, M., Hall, B., Singla, A., & Sukharevsky, A. (2021). *Global survey: The state of AI in 2021 | McKinsey*. https://www.mckinsey.com/business-functions/quantumblack/our-insights/global-survey-the-state-of-ai-in-2021

Council of Bureaux. (2022). *The Green Card system | Council of Bureaux*. https://www.cobx.org/article/3/green-card-system

Daniya, T., Geetha, M., & Suresh Kumar, K. (2020). CLASSIFICATION AND REGRESSION TREES WITH GINI INDEX. *Advances in Mathematics: Scientific Journal*, *9*(10), 8237–8247. https://doi.org/10.37418/amsj.9.10.53

Dasgupta, D., Akhtar, Z., & Sen, S. (2022). Machine learning in cybersecurity: A comprehensive survey. *The Journal of Defense Modeling and Simulation*, *19*(1), 57–106. https://doi.org/10.1177/1548512920951275

Fauzan, M. A., & Murfi, H. (2018). The accuracy of XGBoost for insurance claim prediction. *Int. J. Adv. Soft Comput. Appl*, 10(2), 159-171.

Guelman, L. (2012). Gradient boosting trees for auto insurance loss cost modeling and prediction. *Expert Systems with Applications*, *39*(3), 3659–3667. https://doi.org/10.1016/j.eswa.2011.09.058

Hancock, J. T., & Khoshgoftaar, T. M. (2020). CatBoost for big data: An interdisciplinary review. *Journal of Big Data*, *7*(1), 94. https://doi.org/10.1186/s40537-020-00369-8

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Science & Business Media.

Huang, M.-H., & Rust, R. T. (2018). Artificial Intelligence in Service. *Journal of Service Research*, *21*(2), 155–172. https://doi.org/10.1177/1094670517752459

International Risk Management Institute. (2022). *Claims Reserve | Insurance Glossary Definition |*. https://www.irmi.com/term/insurance-definitions/claims-reserve

Islam, M. N., Inan, T. T., Rafi, S., Akter, S. S., Sarker, I. H., & Islam, A. K. M. N. (2020). A Systematic Review on the Use of AI and ML for Fighting the COVID-19 Pandemic.

IEEE Transactions on Artificial Intelligence, *1*(3), 258–270. https://doi.org/10.1109/TAI.2021.3062771

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning: With Applications in R*. Springer Nature.

Levantesi, S., Nigri, A., & Piscopo, G. (2020). Longevity risk management through Machine Learning: State of the art. *Insurance Markets and Companies*, *11*(1), 11–20. https://doi.org/10.21511/ins.11(1).2020.02

Levantesi, S., & Piscopo, G. (2021). COVID-19 Crisis and Resilience: Challenges for the Insurance Sector. *Advances in Management and Applied Economics*, 1–12. https://doi.org/10.47260/amae/1131

Maillart, A. (2021). Toward an explainable machine learning model for claim frequency: A use case in car insurance pricing with telematics data. *European Actuarial Journal*, *11*(2), 579–617. https://doi.org/10.1007/s13385-021-00270-5

Mehta, P., & Pandya, D. S. (2020). *A Review On Sentiment Analysis Methodologies, Practices And Applications*. *9*(02), 10.

Micol Policarpo, L., da Silveira, D. E., da Rosa Righi, R., Antunes Stoffel, R., da Costa, C. A., Victória Barbosa, J. L., Scorsatto, R., & Arcot, T. (2021). Machine learning through the lens of e-commerce initiatives: An up-to-date systematic literature review. *Computer Science Review*, *41*, 100414. https://doi.org/10.1016/j.cosrev.2021.100414

Molnar, C. (2020). Interpretable Machine Learning - A Guide for Making Black Box Models. Lulu.com.

Montemayor, C., Halpern, J., & Fairweather, A. (2021). In principle obstacles for empathic AI: Why we can’t replace human empathy in healthcare. *AI & SOCIETY*. https://doi.org/10.1007/s00146-021-01230-z

Nevins, B. (2021). Machine learning: Practical applications for the insurance industry. *Property & Casualty 360*. https://www.proquest.com/docview/2592962628/abstract/3391B6F62DDB4F93PQ/1

Oza, D., Padhiyar, D., Doshi, V., & Patil, S. (2020). *Insurance Claim Processing Using RPA Along With Chatbot* (SSRN Scholarly Paper No. 3561871). https://doi.org/10.2139/ssrn.3561871

Pak, M., & Kim, S. (2017). A review of deep learning in image recognition. *2017 4th International Conference on Computer Applications and Information Processing Technology (CAIPT)*, 1–3. https://doi.org/10.1109/CAIPT.2017.8320684

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2019). *CatBoost: Unbiased boosting with categorical features* (arXiv:1706.09516). arXiv. https://doi.org/10.48550/arXiv.1706.09516

Ranstam, J., & Cook, J. A. (2018). LASSO regression. *British Journal of Surgery*, *105*(10), 1348. https://doi.org/10.1002/bjs.10895

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, *2*(3), 160. https://doi.org/10.1007/s42979-021-00592-x

Severino, M. K., & Peng, Y. (2021). Machine learning algorithms for fraud prediction in property insurance: Empirical evidence using real-world microdata. *Machine Learning with Applications*, *5*, 100074. https://doi.org/10.1016/j.mlwa.2021.100074

Shailaja, K., Seetharamulu, B., & Jabbar, M. A. (2018). Machine Learning in Healthcare: A Review. *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, 910–914. https://doi.org/10.1109/ICECA.2018.8474918

Sharma, R., Kamble, S. S., Gunasekaran, A., Kumar, V., & Kumar, A. (2020). A systematic literature review on machine learning applications for sustainable agriculture supply chain performance. *Computers & Operations Research*, *119*, 104926. https://doi.org/10.1016/j.cor.2020.104926

Shaw, G. (2021). *2022 Insurance Industry Outlook*. Deloitte Insights. https://www2.deloitte.com/us/en/insights/industry/financial-services/financial-services-industry-outlooks/insurance-industry-outlook.html

Singh, A., Thakur, N., & Sharma, A. (2016). A review of supervised machine learning algorithms. *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, 1310–1315.

Souza, J. T. de, Francisco, A. C. de, Piekarski, C. M., & Prado, G. F. do. (2019). Data Mining and Machine Learning to Promote Smart Cities: A Systematic Review from 2000 to 2018. *Sustainability*, *11*(4), 1077. https://doi.org/10.3390/su11041077

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Tselentis, D. I., Yannis, G., & Vlahogianni, E. I. (2017). Innovative motor insurance schemes: A review of current practices and emerging challenges. *Accident Analysis & Prevention*, *98*, 139–148. https://doi.org/10.1016/j.aap.2016.10.006

Twisk, D. A. M., & Stacey, C. (2007). Trends in young driver risk and countermeasures in European countries. *Journal of Safety Research*, *38*(2), 245–257. https://doi.org/10.1016/j.jsr.2007.03.006

Vroom, A. (2021, December 8). These Are the 20 Most Expensive Car Brands. *Global Cars Brands*. https://www.globalcarsbrands.com/most-expensive-car-brands/

Wang, Y., & Xu, W. (2018). Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud. *Decision Support Systems*, *105*, 87–95. https://doi.org/10.1016/j.dss.2017.11.001

West, D. M., & Allen, J. R. (2018). *How artificial intelligence is transforming the world*. http://www.cemas.org.uk/index.php/17-content/publications/2009-how-artificial-intelligence-is-transforming-the-world

Wright, M. N., Wager, S., & Probst, P. (2022). *ranger: A Fast Implementation of Random Forests* (0.14.1). https://CRAN.R-project.org/package=ranger

Zehnwirth, B. (1979). A hierarchical model for the estimation of claim rates in a motor car insurance portfolio. *Scandinavian Actuarial Journal*, *1979*(2–3), 75–82. https://doi.org/10.1080/03461238.1979.10413713

APPENDIX A: EXPLORATORY PLOTS – BEFORE MODELING

FIGURE 18. EXPLORATORY BOXPLOT BETWEEN EVENTCOUNTRY & COST

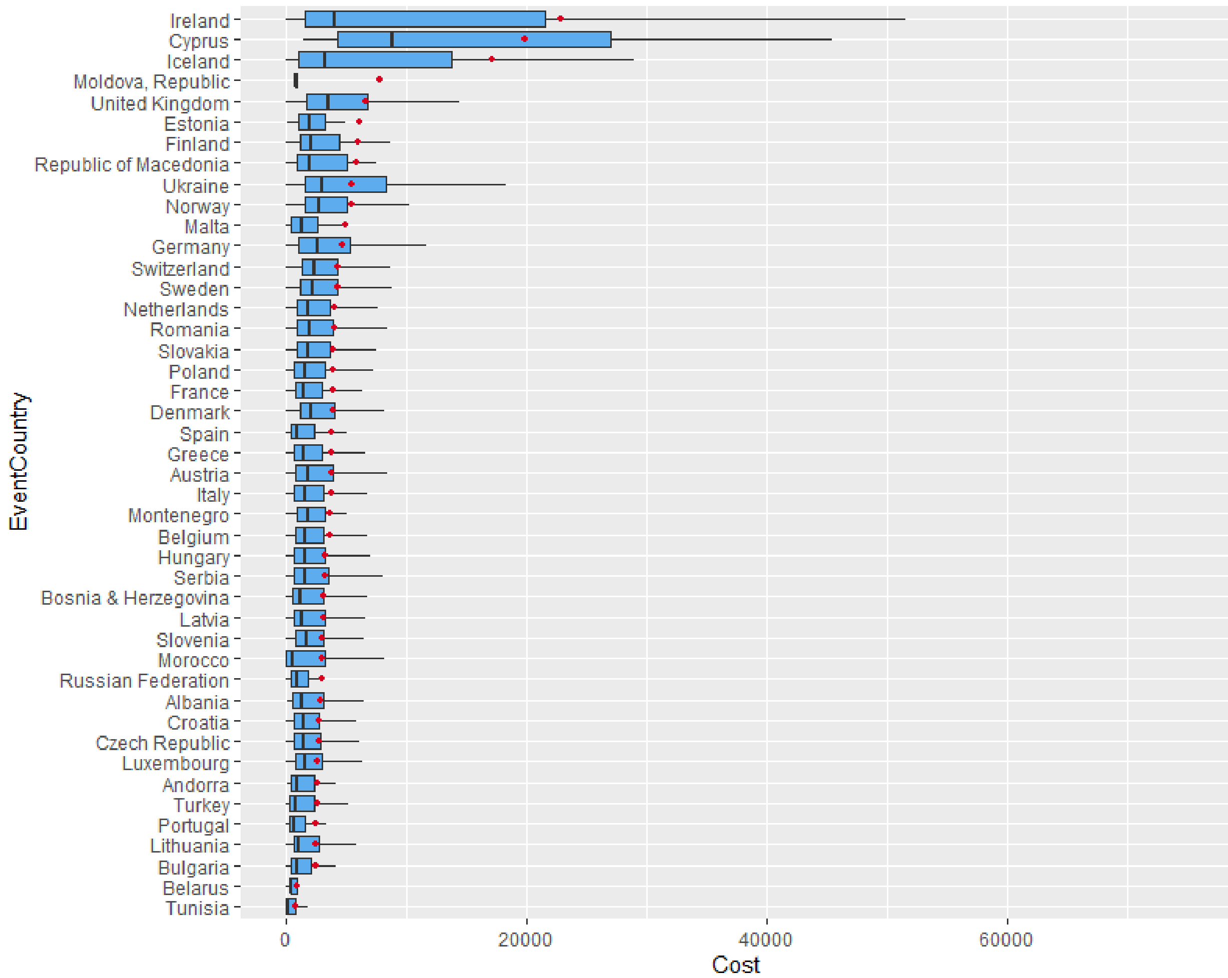


FIGURE 19. EXPLORATORY BOXPLOT BETWEEN CLAIMEDOBJECTSUBTYPE & COST

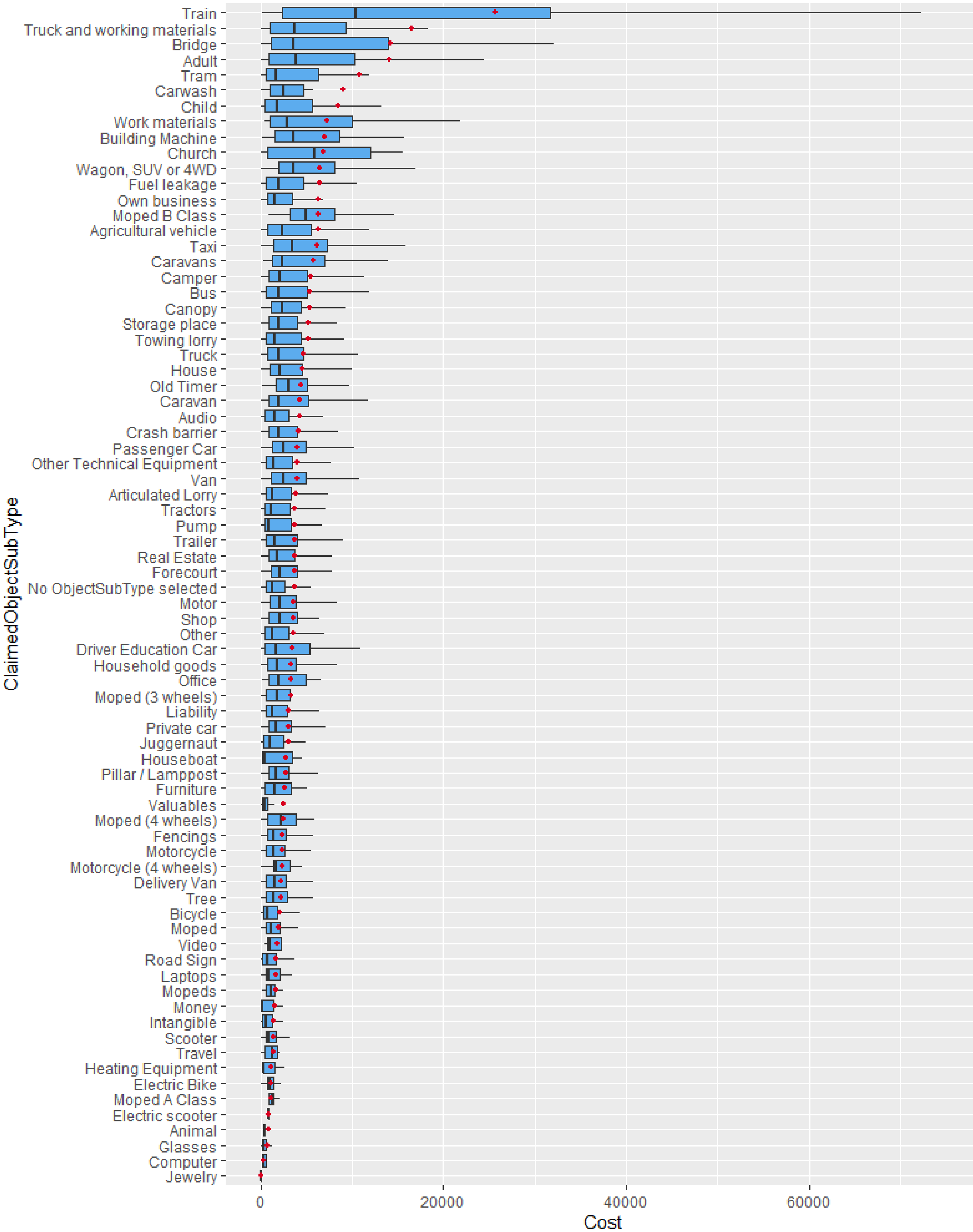


FIGURE 20. EXPLORATORY BOXPLOT BETWEEN VEHICLEMAKE & COST

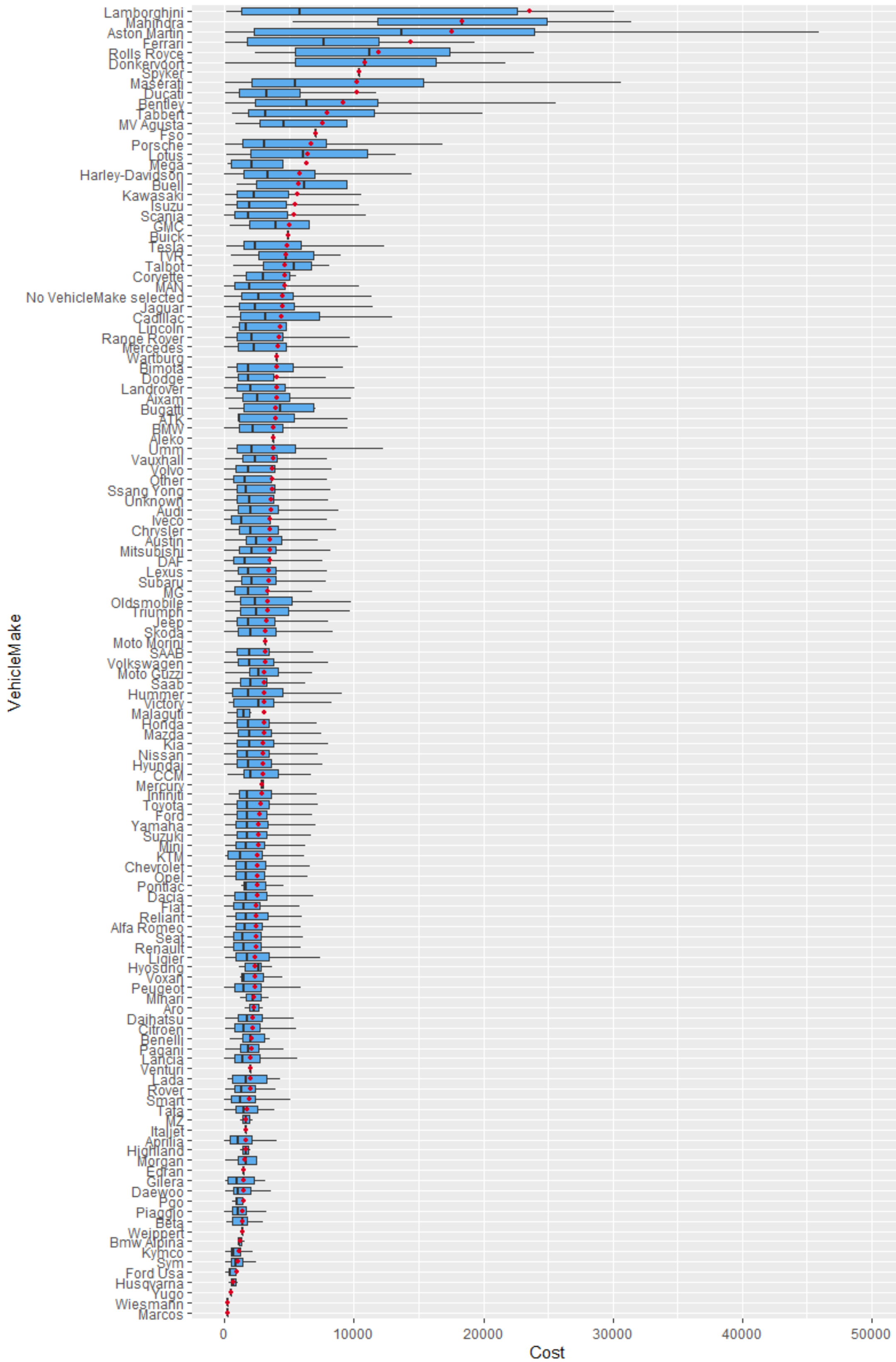


FIGURE 21. EXPLORATORY BOXPLOT BETWEEN ACCIDENTCAUSE & COST