ML1

Alvaro Cervan, Luca Renz, Rafaella Miranda-Sousa

2024-12-21

Contents

Introduction

One of the major challenges for insurance companies is to estimate the appropriate premiums to charge each customer while not risking any money loss. Therefore, this project aims to support an Ethiopian insurance company in understanding how their customers can benefit from having the most accurate and fair premium as they need and have to pay. Machine learning helps enormously in this case to understand what factors have a larger impact on the premium and how customers can be classified accordingly.

In this document, the reader may find different algorithms to solve various aspects of premium calculations and other related topics such as risk assessment or claim prediction.

Data Preprocessing

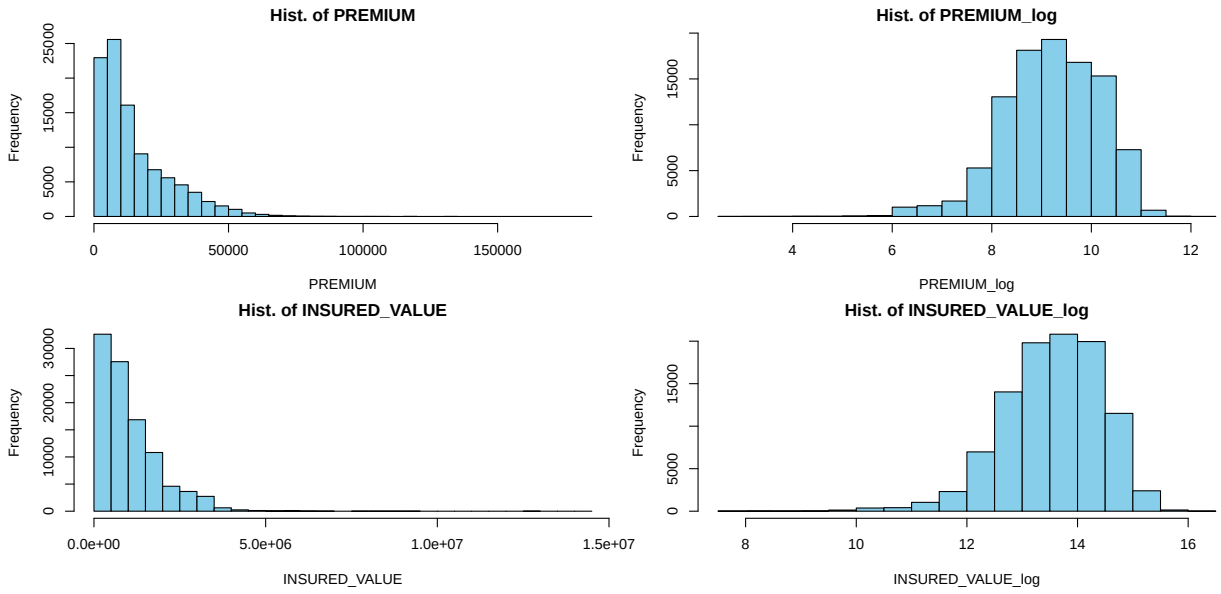
Lead: Rafaella Miranda-Sousa Wasser

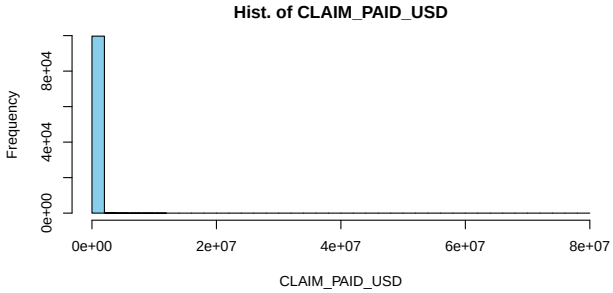
In order to apply such algorithms, the data has to be preprocessed. In a very brief summary, the script removes unnecessary columns and duplicates, and handles missing and zero values, particularly for columns like INSURED_VALUE and SEATS_NUM. It converts certain columns to more meaningful categories, such as transforming SEX into factors representing legal entities and genders. The script also filters data to exclude irrelevant vehicle types and usage, ensuring the final dataset contains only pertinent records. Finally, it summarizes and adjusts the dataset further by converting appropriate columns into factors and removing variables not required for analysis.

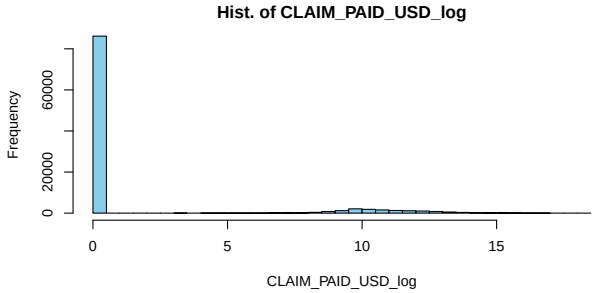
Exploratory graphical analysis

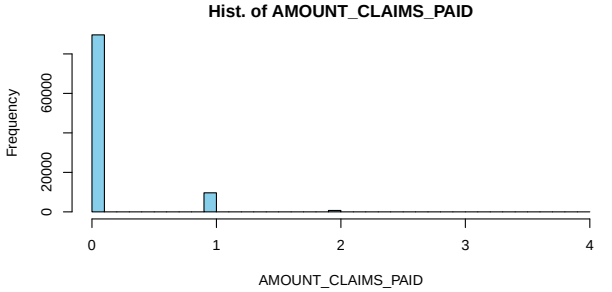
Lead: Rafaella Miranda-Sousa Wasser

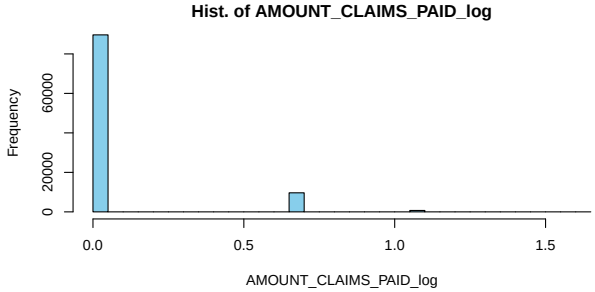
First, the distribution of the individual numerical variables was analysed to determine whether any transformations were necessary.

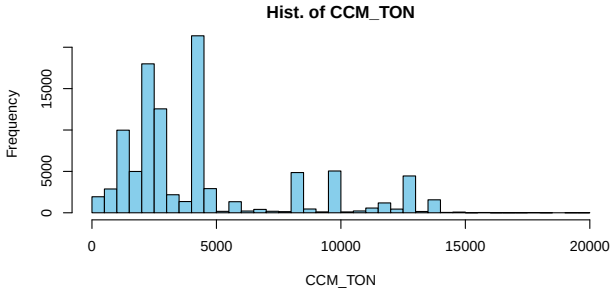


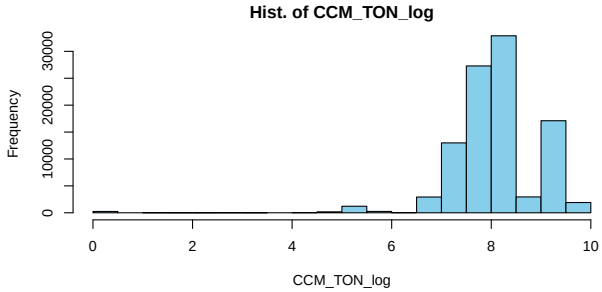

Hist. of CLAIM_PAID_USD
Frequency
0e+00
4e+04
8e+04
0e+00
2e+07
4e+07
6e+07
8e+07
CLAIM_PAID_USD

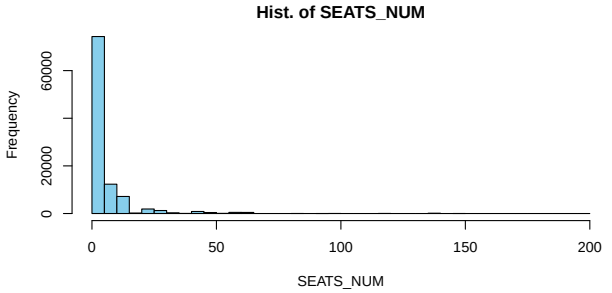

Hist. of CLAIM_PAID_USD_log
Frequency
0
20000
60000
0
5
10
15
CLAIM_PAID_USD_log

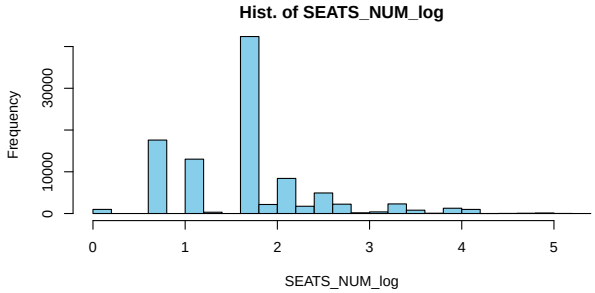

Hist. of AMOUNT_CLAIMS_PAID
Frequency
0
20000
60000
0
1
2
3
4
AMOUNT_CLAIMS_PAID

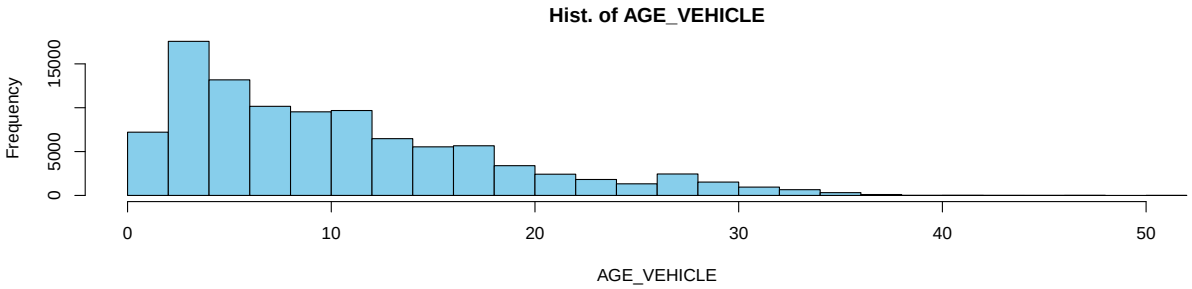

Hist. of AMOUNT_CLAIMS_PAID_log
Frequency
0
20000
60000
0.0
0.5
1.0
1.5
AMOUNT_CLAIMS_PAID_log

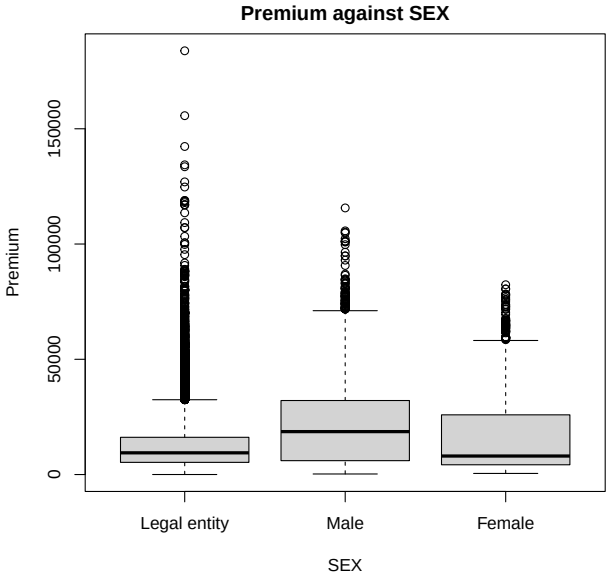

Hist. of CCM_TON
Frequency
0
5000
15000
0
5000
10000
15000
20000
CCM_TON

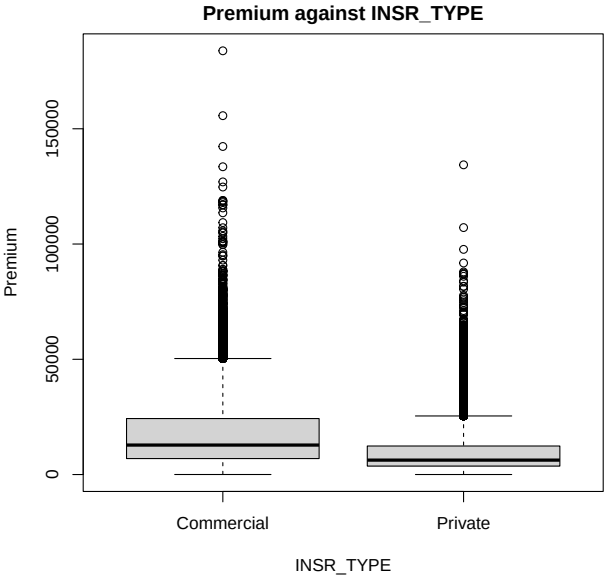

Hist. of CCM_TON_log
Frequency
0
10000
20000
30000
0
2
4
6
8
10
CCM_TON_log

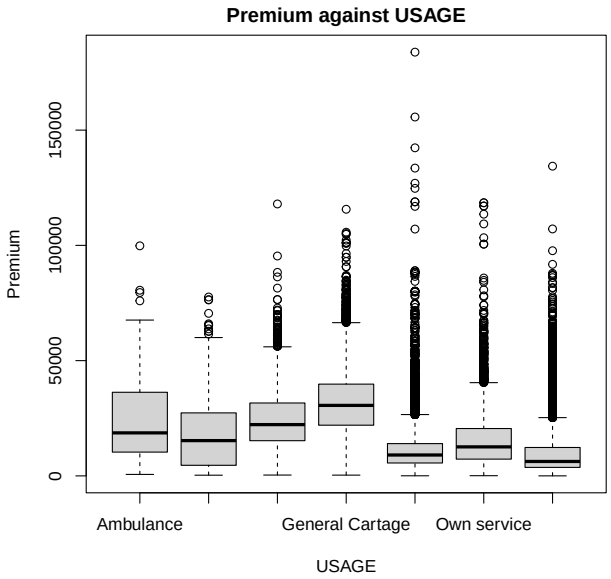

Hist. of SEATS_NUM
Frequency
0
20000
60000
0
50
100
150
200
SEATS_NUM

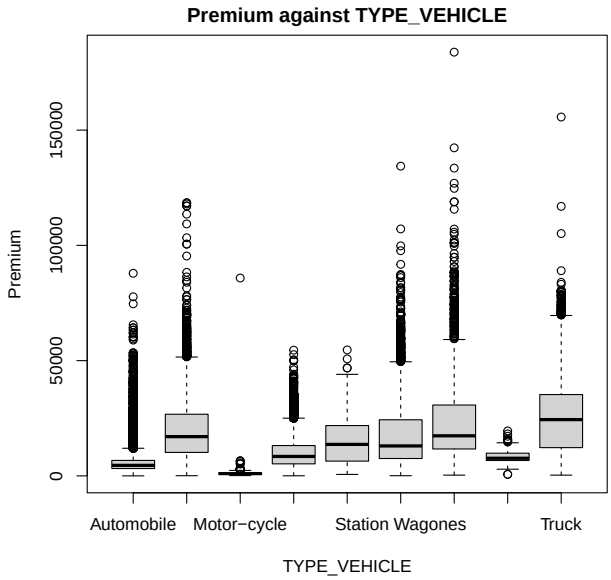

Hist. of SEATS_NUM_log
Frequency
0
10000
30000
0
1
2
3
4
5
SEATS_NUM_log

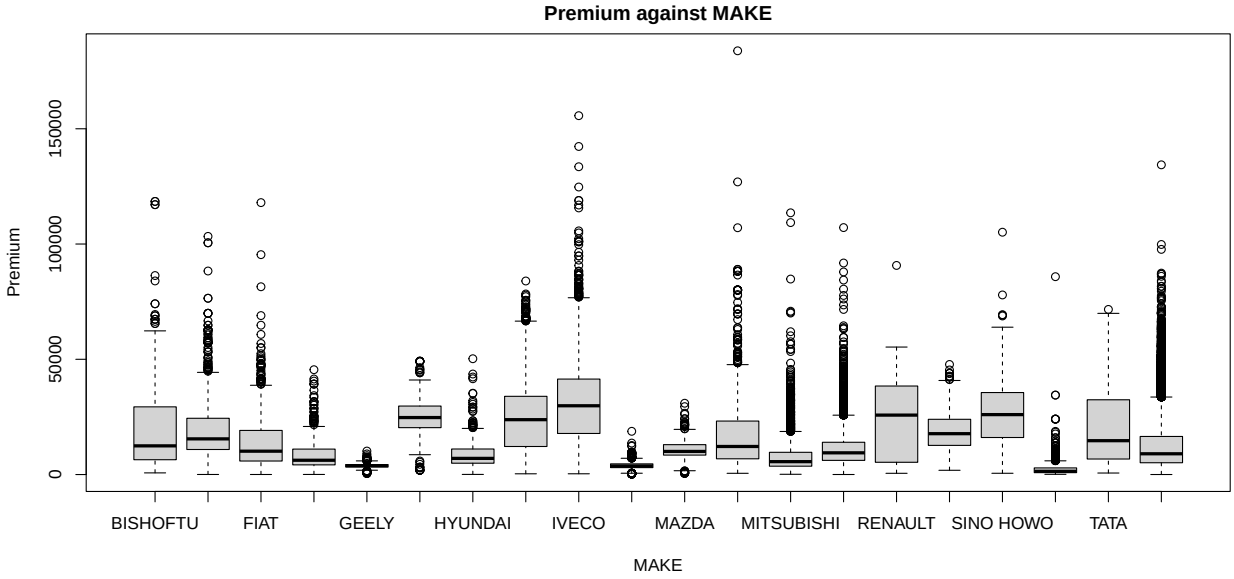

Hist. of AGE_VEHICLE
Frequency
0
5000
15000
0
10
20
30
40
50
AGE_VEHICLE

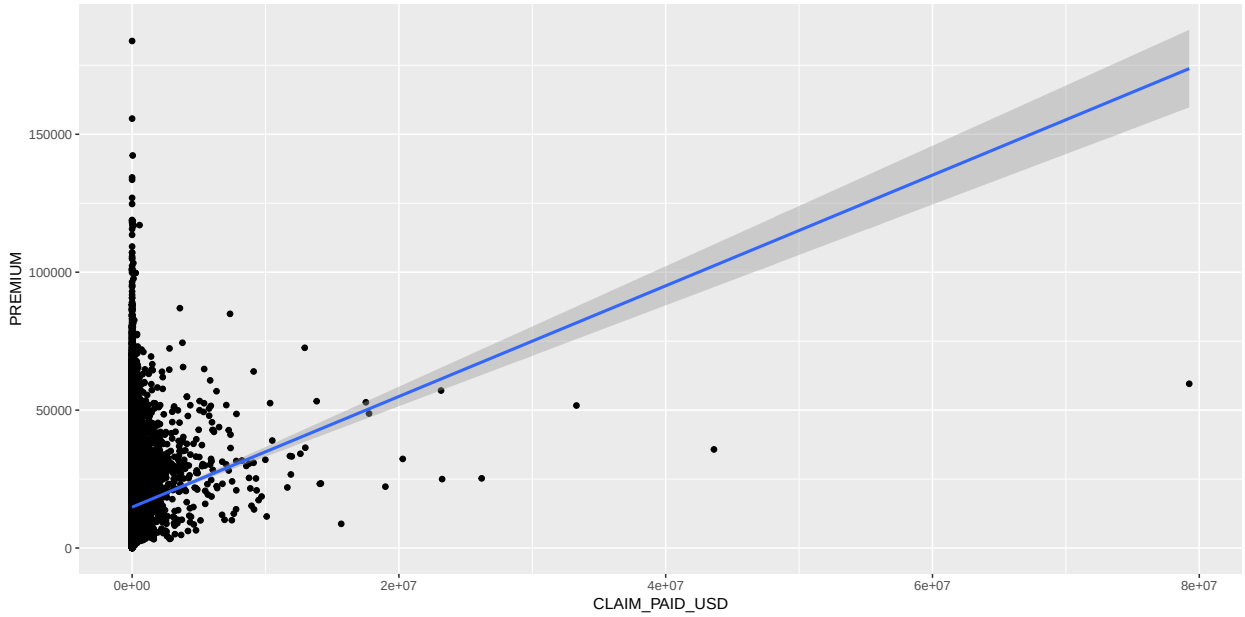

Premium against SEX
Premium
0
50000
100000
150000
Legal entity
Male
Female
SEX

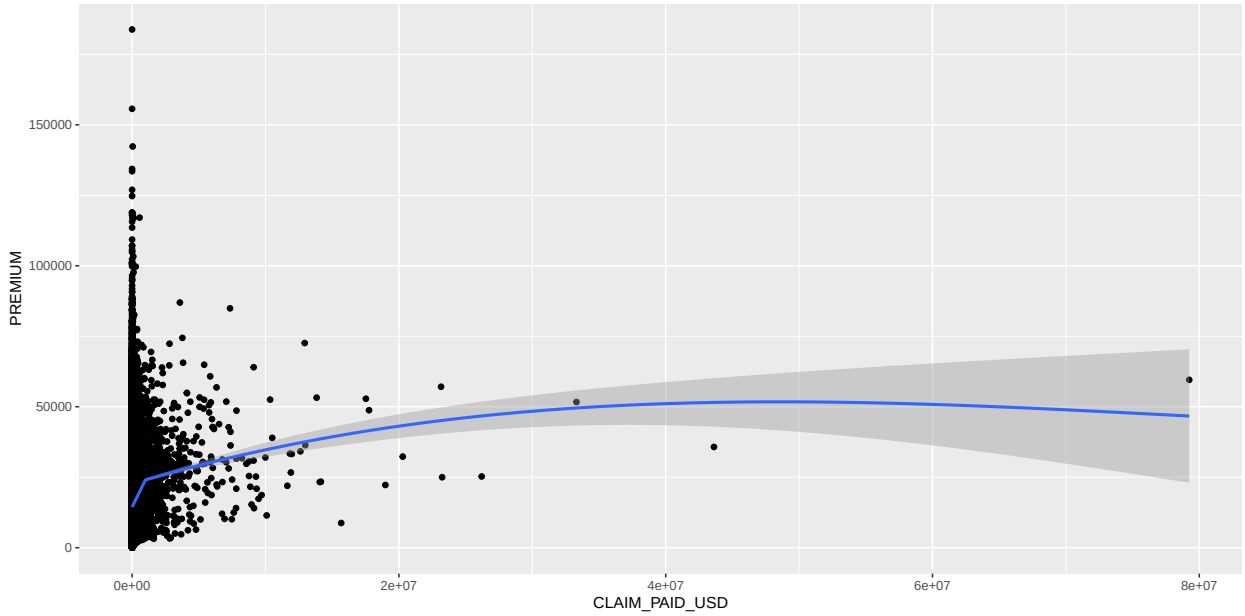

Premium against INSR_TYPE
Premium
0
50000
100000
150000
Commercial
Private
INSR_TYPE


Premium against USAGE
Premium
0
50000
100000
150000
Ambulance
General Cartage
Own service
USAGE


Premium against TYPE_VEHICLE
Premium
0
50000
100000
150000
Automobile
Motor−cycle
Station Wagones
Truck
TYPE_VEHICLE


Premium against MAKE
Premium
0
50000
100000
150000
BISHOFTU
FIAT
GEELY
HYUNDAI
IVECO
MAZDA
MITSUBISHI
RENAULT
SINO HOWO
TATA
MAKE





The histograms show that the variables PREMIUM, INSURED_VALUE, CLAIM_PAID_USD, AMOUNT_CLAIMS_PAID, CCM_TON and SEATS_NUM are right-skewed and require a log transformation. The transformed variables will be inserted in the later regression models instead of the original variables.

The boxplots reveal distinct differences in premium distributions across categories, such as higher premiums for "Legal Entity" compared to "Male" and "Female," and significant variations among vehicle types like "Motorcycle" and "Truck." These patterns suggest that the predictor variables have a substantial influence on premium amounts, warranting further formal analysis.

The relationships between PREMIUM and all numerical predictors except INSURED_VALUE appear to be linear and are included in the models as linear effects. The effect of INSURED_VALUE is clearly non-linear. The smoothing indicates that we could model this relationship as a quadratic effect.

The analysis shows that there are no clear indications of significant interactions for the time being.

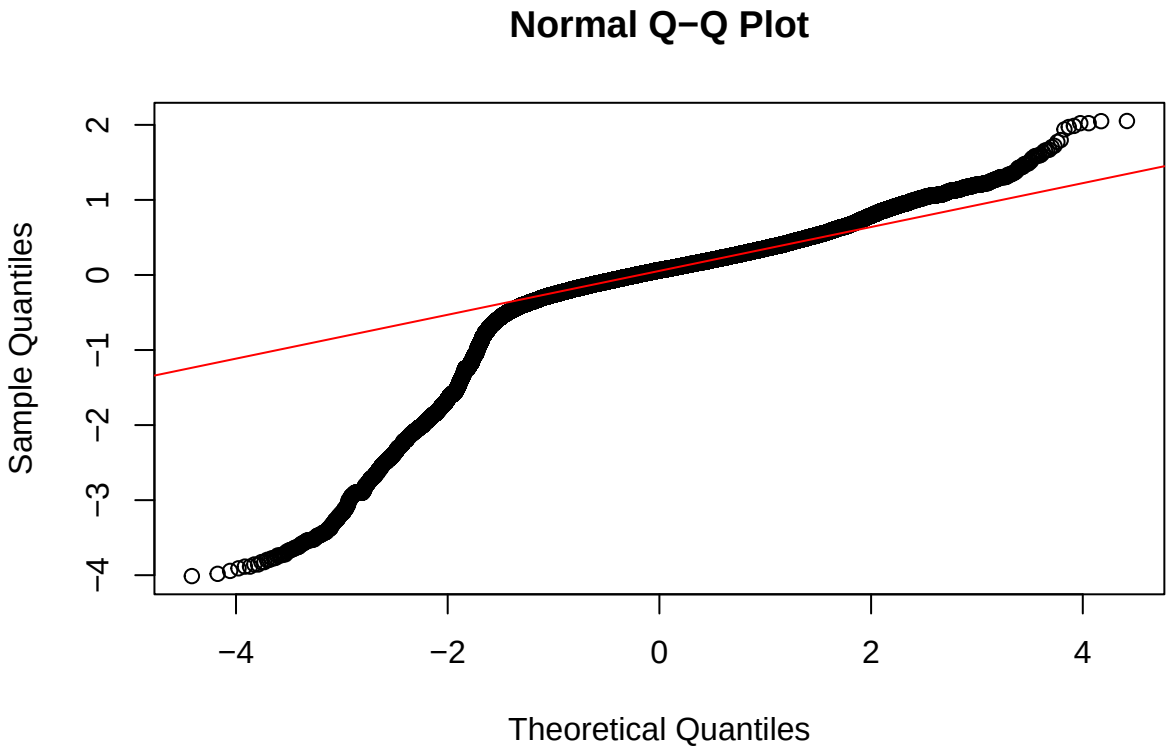
Models

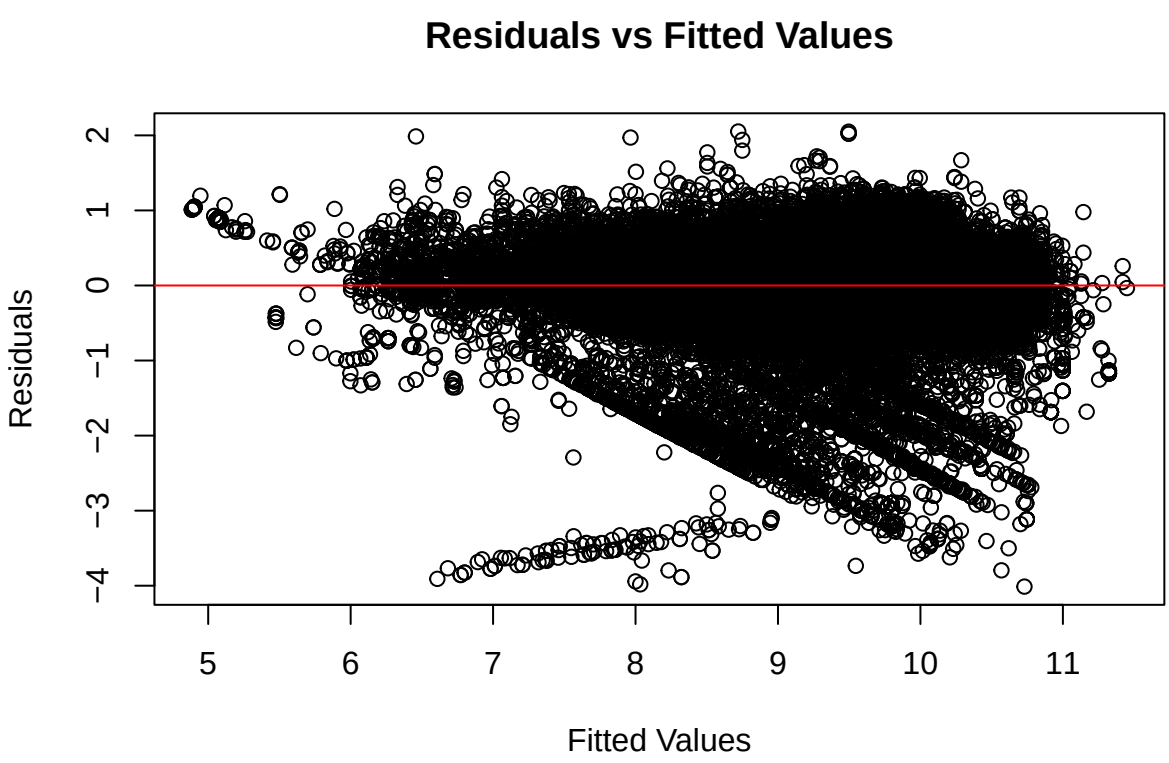
Once the pre-processing was completed and a good overview about the given data and domain knowledge what acquired, the team focused on defining models using several different methods shown in followed subsections.

Linear model

Lead: Rafaella Miranda-Sousa Wasser

The aim of the following analysis is to identify the key factors influencing premium amounts. Since the data originates from an Ethiopian insurer, the objective is to assess whether the premium setting is reasonable. The choice of PREMIUM_log as the target variable is based on the strongly right-skewed distribution of the original variable PREMIUM, which could violate the assumptions of linear regression. In constructing the linear model, CLAIM_PAID_USD_log is excluded, as premiums are determined at the start of the contract, making its inclusion technically inappropriate. Instead, a bonus-malus system is incorporated by adding AMOUNT_CLAIMS_PAID_log, reflecting the cumulative claim history and aligning with premium adjustments based on risk exposure.





All explanatory variables except CCM_TON_log appear to be significant. Therefore, CCM_TON_log is removed from the model

The model summary shows that the Multiple R-squared value is 0.7307, indicating that the model can explain approximately 73.07% of the variance in premiums. This suggests that the model provides a good fit to the data. The F-test for the overall model is significant ($p < 2.2e-16$), indicating that the predictors as a group have a substantial effect on the premium. All predictors have a significant impact on the target variable PREMIUM_log. The Modell explains 73.06% of the variance in PREMIUM_log (Adjusted R-squared: 0.7306). The Mean Squared Error (MSE) of 0.2536 indicates a good fit with relatively low average prediction errors.

Men pay, on average, 8.14% (-0.0814) and women 7.67% (-0.0767) lower logarithmic premiums compared to Legal Entity, with minimal differences between the genders. Vehicles used for commercial purposes (Fare Paying Passengers) have significantly higher premiums compared to the reference level "Ambulance," with an increase of 52.59%, while private or self-use vehicles such as Own Goods -50.52% and Private -47.28% have substantially lower premiums compared to "Ambulance." High-end brands like ISUZU, TOYOTA, and MERCEDES tend to have higher premiums, while GENLYON and RENAULT exhibit lower premiums, which may be attributed to differences in repair costs or target customer groups. Commercial vehicles such as Tanker and Truck have significantly higher premiums due to their higher risk, while smaller vehicles like Motor-cycle are associated with lower premiums.

INSURED_VALUE_log has the strongest impact with a coefficient of 0.762, meaning that a 1% increase in the insured value leads to an approximately 76.2% increase in the logarithmic premium. Similarly, AMOUNT_CLAIMS_PAID_log (0.208) indicates that a 1% increase in claims paid corresponds to an approximately 20.8% increase in the logarithmic premium, and SEATS_NUM_log (0.016) suggests a 1% increase in the number of seats results in a 1.6% increase in the logarithmic premium. Meanwhile, AGE_VEHICLE shows a smaller but still significant effect with a coefficient of 0.0026, where each additional year increases the logarithmic premium by approximately 0.26%.

VIF: An analysis of multicollinearity revealed that the Variance Inflation Factor (VIF) for the variable

INSR_TYPE is 5.88, which suggests possible multicollinearity. This could affect the model's stability and interpretability and should be considered in further model optimization.

Residuals Analysis Residuals vs. Fitted Plot: The Residuals vs. Fitted Plot displays a funnel-shaped pattern, indicating heteroskedasticity. The variance of the residuals increases with higher predicted values, meaning that the model is less accurate for larger premium values. This violates the assumption of constant variance, suggesting that homoskedasticity is not fully met.

Normal Q-Q Plot: The Normal Q-Q Plot shows that the residuals do not lie perfectly along the line, indicating significant deviations from the theoretical normal distribution, particularly at the tails. These "heavy tails" suggest a non-normal distribution of residuals, potentially due to outliers or unmodeled non-linear relationships.

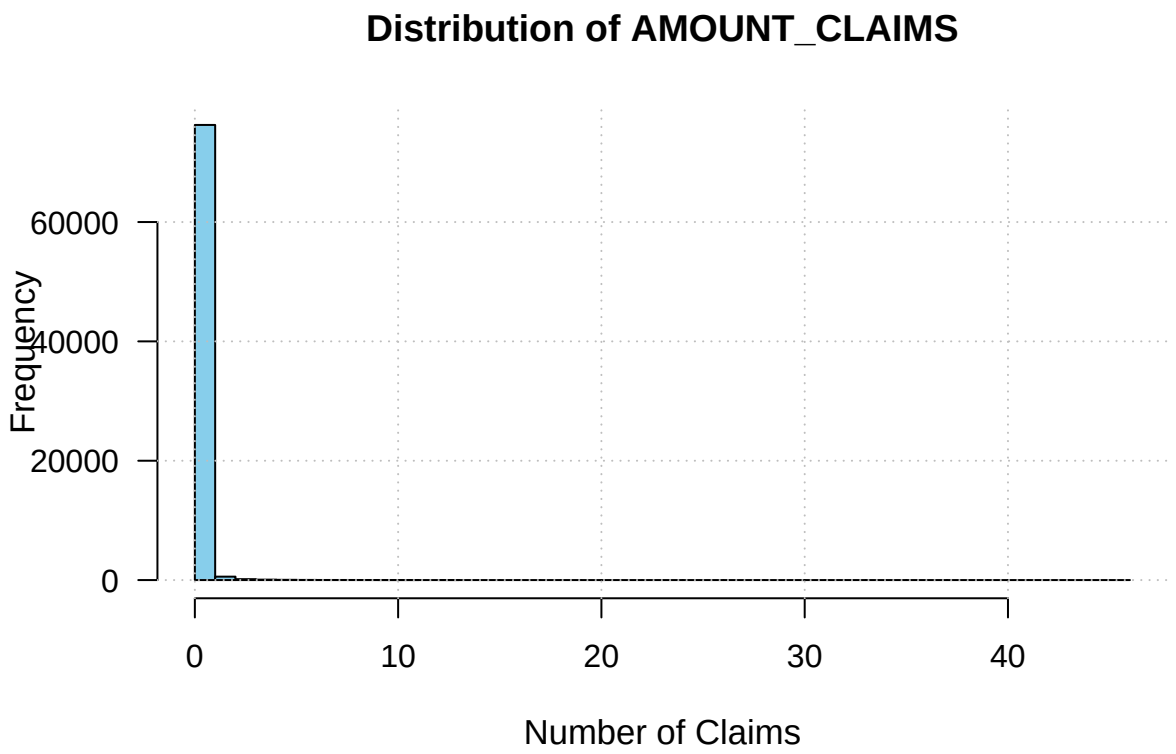
To improve the model, various measures could be considered. One approach would be to transform the target variable, for example, using a Box-Cox transformation, to reduce heteroskedasticity and achieve a more stable residual variance. Additionally, incorporating non-linear relationships by including polynomial terms or using a generalized linear model (GLM) could be beneficial. This would allow the model to better capture complex relationships between variables, thereby enhancing predictive accuracy.

Poisson model

Lead: Rafaella Miranda-Sousa Wasser

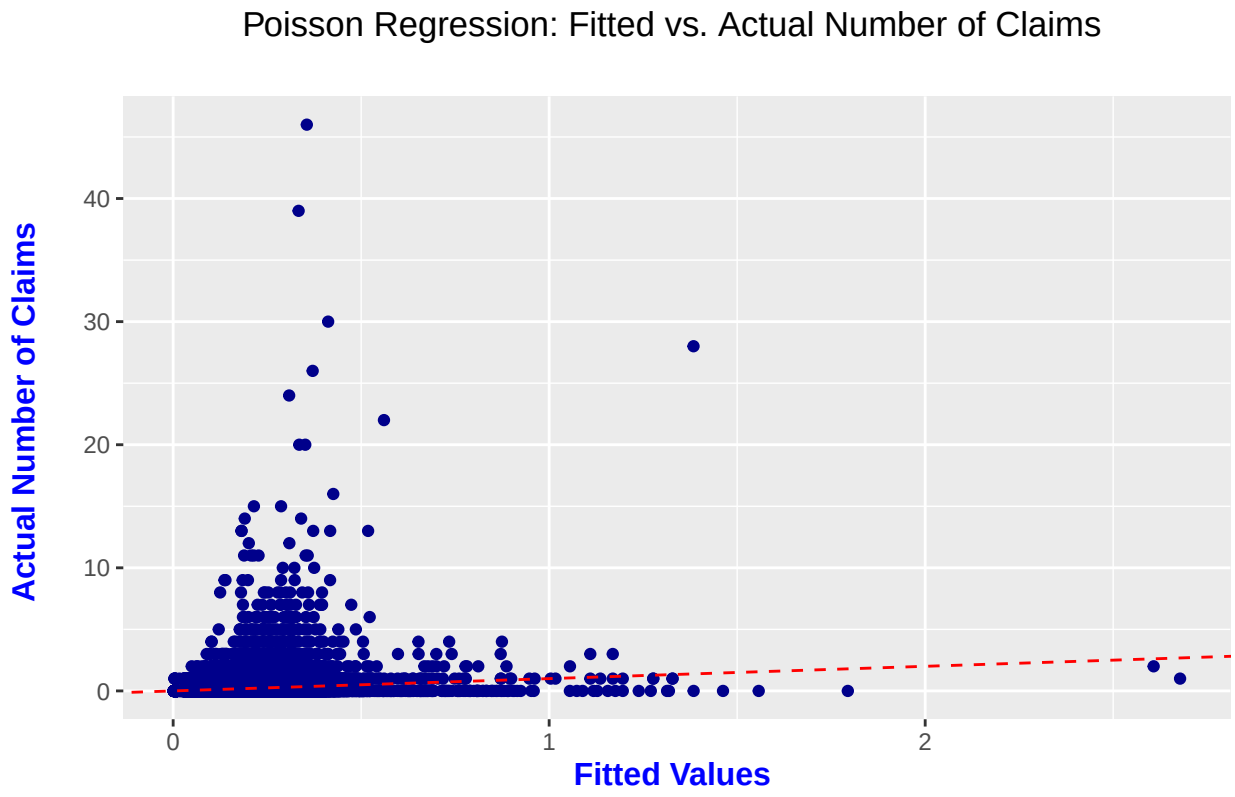
A Poisson model is fitted to predict the number of claims over a 5-year period based on the characteristics SEX, INSR_TYPE, USAGE, TYPE_VEHICLE, MAKE, AGE_VEHICLE, SEATS_NUM, CCM_TON, INSURED_VALUE, and PREMIUM.

First, the data is grouped accordingly, and the results are analyzed to gather insights.



The analysis of the distribution of the target variable AMOUNT_CLAIMS reveals that a large portion

of the values are zero. This concentration of zero values is confirmed by the median, as well as the 1st and 3rd quartiles, which are also at zero. Additionally, the distribution shows some high outliers with a maximum value of 46, indicating an uneven distribution with a few high values. The low mean of 0.1791 further supports this observation, suggesting a significant number of zero values. Given these distribution characteristics, the use of a Zero-Inflated Poisson (ZIP) model could be appropriate, as such a model can account for both random and structural zeros. Initially, however, a Poisson model will be fitted.



The Poisson model analysis shows no signs of overdispersion, as the overdispersion value of 0.688 is below 1. This indicates a possible underdispersion, but this is confirmed by the goodness-of-fit test (p-value = 1), which confirms a well-fitting model.

The Poisson regression model for predicting the number of claims reveals that several variables show statistically significant relationships with claim frequency. The model indicates statistically significant differences in claim frequency across categories (p-value < 0.001). The group of legal entities, which serves as the reference category, exhibits the highest claim rate. Compared to legal entities, males have a rate ratio of 0.666, reflecting a 33.4% lower claim rate, while females have the lowest claim frequency, with a rate ratio of 0.623, or 37.7% below that of legal entities.

For the insurance type (INSR_TYPE), it was found that INSR_TYPEPrivate has a rate ratio of 1.284, indicating that private insurers have a 28.4% higher claim probability compared to the reference category INSR_TYPECommercial. The variables TYPE_VEHICLE and MAKE also show significant differences in claim rates. Among vehicle types, Pick-up has the highest claim rate, with a rate ratio of 1.121, representing a 12.1% increase in claim probability compared to the reference category Automobile; however, this effect is not statistically significant (p-value = 0.120). Conversely, Motor-cycle has the lowest claim rate, with a rate ratio of 0.039, indicating an approximately 96% reduced claim probability and a highly significant result (p-value < 0.001).

Among vehicle brands, GEELY shows the highest claim rate with a rate ratio of 1.040, which, however, represents no meaningful change compared to the reference brand BISHOFTU and is statistically insignificant

(p-value = 0.709). Conversely, MERCEDES has the lowest claim rate, with a rate ratio of 0.403, indicating a 59.7% lower claim probability compared to BISHOFTU and is highly significant (p-value < 0.001).

These results suggest that Pick-up and GEELY exhibit the highest, though statistically insignificant, claim rates, while Motor-cycle and MERCEDES show the lowest and statistically significant claim rates relative to their respective reference categories.

Further analysis indicates that vehicle age (AGE_VEHICLE) has a rate ratio of 0.956, meaning that the claim rate decreases by approximately 4.4% with each additional year (p-value < 0.001). The number of seats (SEATS_NUM) shows a rate ratio of 1.009, indicating that each additional seat slightly increases the claim probability, though significantly. Engine capacity (CCM_TON) shows no practical change in claim rate with a rate ratio of 1.000031, though it is statistically significant (p-value < 0.001). Insured value (INSURED_VALUE) has a rate ratio of 0.99999986, effectively showing no influence on claim frequency, although the effect is statistically significant. Premium amount (PREMIUM) exhibits a rate ratio of 1.000019, suggesting a minimal increase in claim probability with rising premiums; again, the effect is significant but very small.

The VIF values of all predictors are below the critical limit of 5, which indicates that there are no multicollinearity problems. The variables INSR_TYPE (3.75) and TYPE_VEHICLE (1.42) have the highest values, which indicates a moderate correlation with other predictors, but does not cause any stability problems in the model. Overall, the low VIF values support the robustness of the model estimates.

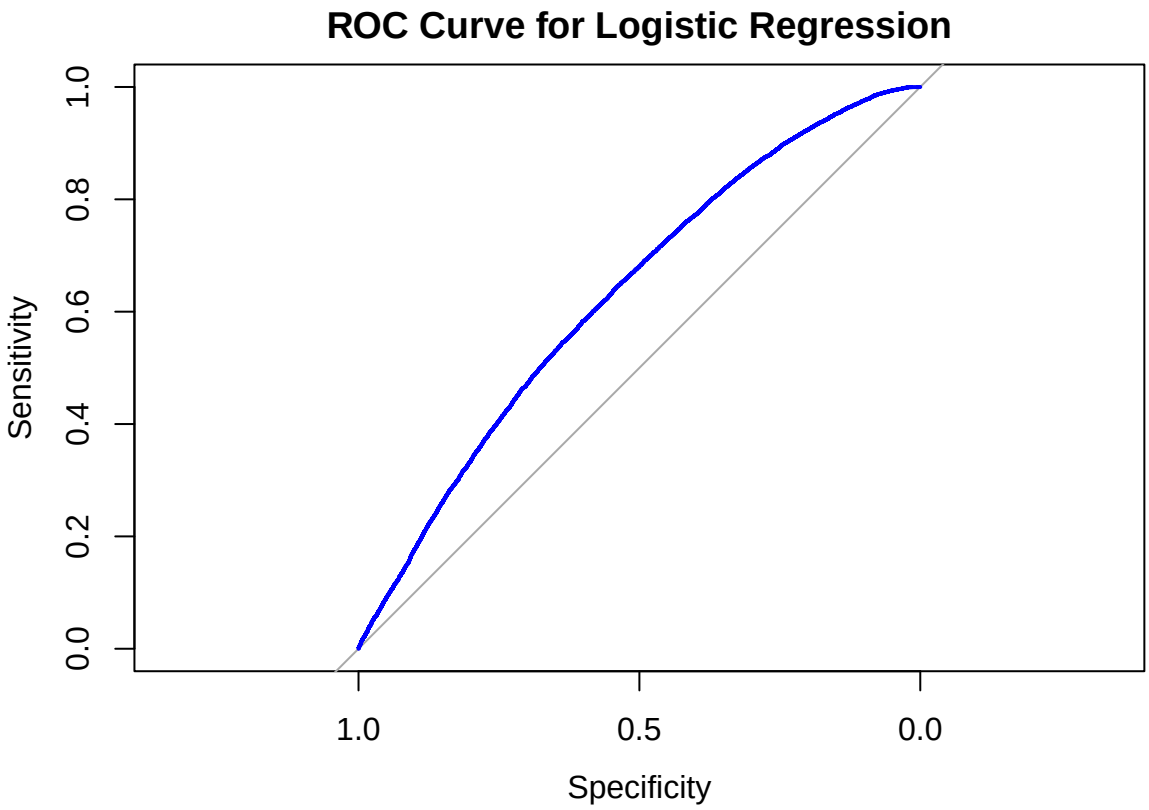
The plot illustrates the relationship between the predicted values (Fitted Values) and the actual number of claims (Actual Number of Claims). Most actual values are concentrated in the lower range (0 to 10), while the predicted values are almost entirely clustered near zero. The model struggles to predict higher claim counts (>10), as evident from the significant deviations for extreme values. The red dashed line, representing the ideal fit between predicted and actual values, shows that many points fall below the line, indicating a systematic underestimation of actual claims by the model. High variability or extreme values in the data can cause the model to perform poorly in capturing such cases.

The plots of estimated vs. actual values show that the Poisson model has difficulties in accurately modelling the distribution of claims, especially for higher claims values. Most of the predicted values are close to zero and systematically underestimate the actual loss frequencies as they increase. This systematic underestimation and the high number of zero claims indicate that the simple distribution of the Poisson model may not be sufficient to fully represent the structure of the data.

Given the high number of zero values in the data, a Zero-Inflated Poisson (ZIP) model could represent a useful alternative. Such a model can distinguish between structural zeros (cases where no claims occur) and random zeros (cases where claims could occur but did not), potentially improving predictive accuracy for higher claim counts without violating model assumptions about variance. As a further alternative, simplifying the model, for example by removing fewer significant variables, could be a sensible measure to improve the model.

Binomial model

Lead: Rafaella Miranda-Sousa Wasser



All explanatory variables except INSR_TYPE appear to be significant. Therefore, INSR_TYPE is removed from the model.

The analysis shows that some variables, such as SEX, TYPE_VEHICLE, MAKE, AGE_VEHICLE, INSURED_VALUE_log, PREMIUM_log, and AMOUNT_CLAIMS_PAID_log, are highly significant ($p < 0.001$). These variables make a substantial contribution to explaining the variance. On the other hand, certain categories, particularly within the variables USAGE and MAKE, show low significance, indicating a limited explanatory effect of these predictors. The global test statistic (TD) is 2713.2777, with a highly significant p-value < 2.2e-16, indicating that the model as a whole provides a statistically significant fit to the data. The overdispersion ratio of 0.7766 indicates that the model does not suffer from overdispersion, confirming the appropriateness of the binomial logistic regression assumption.

Men have a 20.2% lower likelihood of receiving a claim payout compared to the reference group (legal entities), while women have a 22.1% reduced likelihood. These effects are highly significant ($p < 0.001$). The TYPE_VEHICLE also shows significant differences. Motorcycles have approximately a 96.5% lower likelihood of claim payouts, making them the group with the lowest payout frequency. Similarly, station wagons and trailers exhibit significantly lower probabilities, with reductions of around 21.5% and 80.3%, respectively. In contrast, pick-ups do not show significant differences compared to the reference level (automobiles). MAKE further influences payout frequency. Vehicles from the brand DAEWOO have a 39.4% lower likelihood of receiving a claim payout compared to the reference brand, FIAT shows a reduction of 47%, and MERCEDES has a 51.8% lower likelihood. These effects are all statistically significant. Other MAKE's, such as GEELY or MAZDA, do not show significant differences compared to the reference level. PREMIUM_log proves to be one of the strongest predictors. An increase in the logarithmic premium amount leads to a significant 71.3% increase in the likelihood of a claim payout. Conversely, a higher INSURED_VALUE_log reduces the payout likelihood by 24.8%. The claim amount also positively correlates with payout frequency: higher AMOUNT_CLAIMS_PAID_log significantly increase the likelihood of payouts. The age of the vehicle has a negative effect on payout frequency: with each additional year, the likelihood of a payout decreases by approximately 2.8%.

The model has an AIC value of 77716.2. The ROC curve confirms the moderate discriminatory ability of the model. The pseudo-R^2 values (e.g., Nagelkerke: 0.0338) indicate limited explanatory power for the model. This is supported by the AUC (Area Under the Curve) of 0.6295, reflecting a low to moderate discriminatory ability. The model is only slightly better than random at distinguishing between positive and negative cases.

The model is statistically significant but shows limited predictive power and variance explanation. Further investigation into additional predictors or non-linear effects might improve model performance.

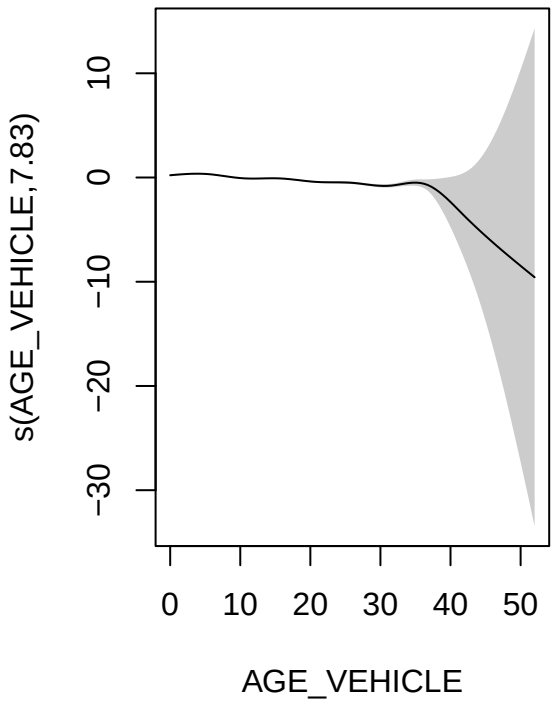
The imbalanced distribution of classes (CLAIMS_PAID: "YES"/"NO") in the dataset significantly impacts the model's discriminatory ability. The moderate AUC of 0.629 indicates that the model struggles to reliably recognize the minority class ("YES"). To improve discriminatory performance, balancing strategies such as oversampling, class weighting, or adjusting the decision threshold could be implemented.

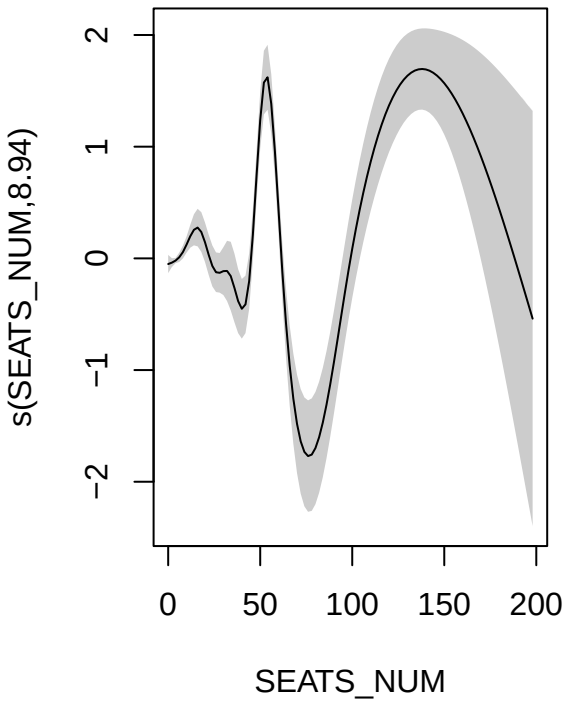
Generalised Additive Model (GAM)

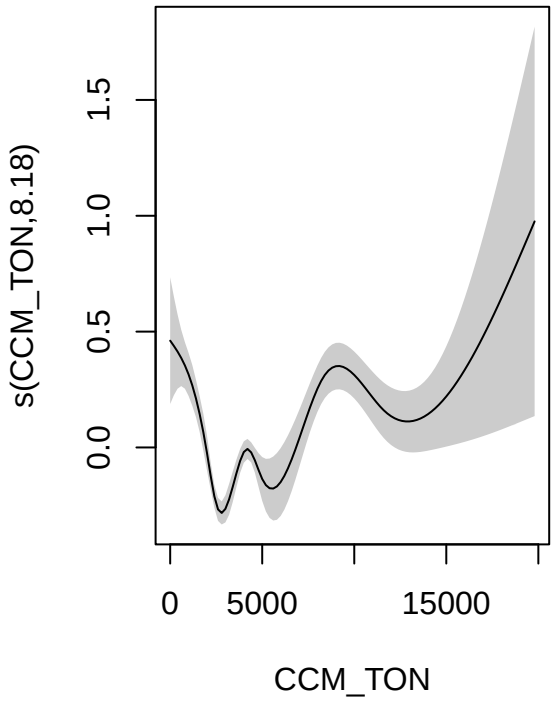
Lead: Alvaro Cervan

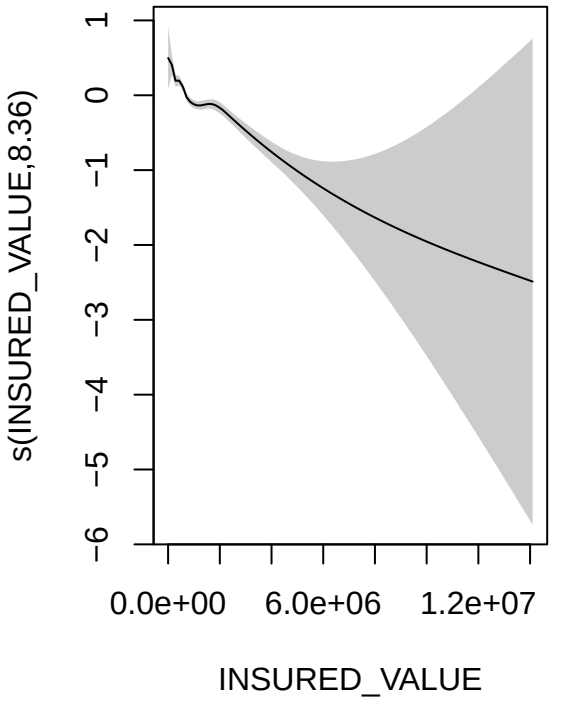
A General Additive Model (GAM) is fitted to predict the number of claims based on the characteristics SEX, INSR_TYPE, USAGE, TYPE_VEHICLE, MAKE, AGE_VEHICLE, SEATS_NUM, CCM_TON, INSURED_VALUE, and PREMIUM. Similar to the Poisson model, the GAM model aims to capture the relationship between the predictors and the number of claims, allowing for non-linear relationships and interactions between variables. AUC value of the GAM model: 0.6175357 RMSE of the GAM model: 0.5938353

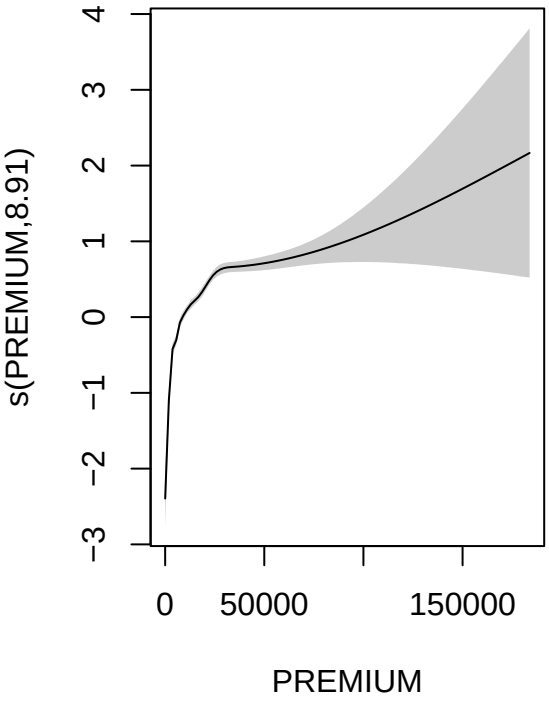
Different variations were used, with and without smoothing predictos and using B-splines with cubic regression. Cubic regression was chosen as it provided the best fit for the data with the smallest RMSE value, all values where ~0.02 of difference. The AUC value of the GAM model is 0.61, very similar to the results in the Poisson model. This indicates that the GAM model has a moderate ability to discriminate between the number of claims and the predictors. The RMSE value of the GAM model is 0.59, which is relatively low and indicates that the model's predictions are close to the actual values.

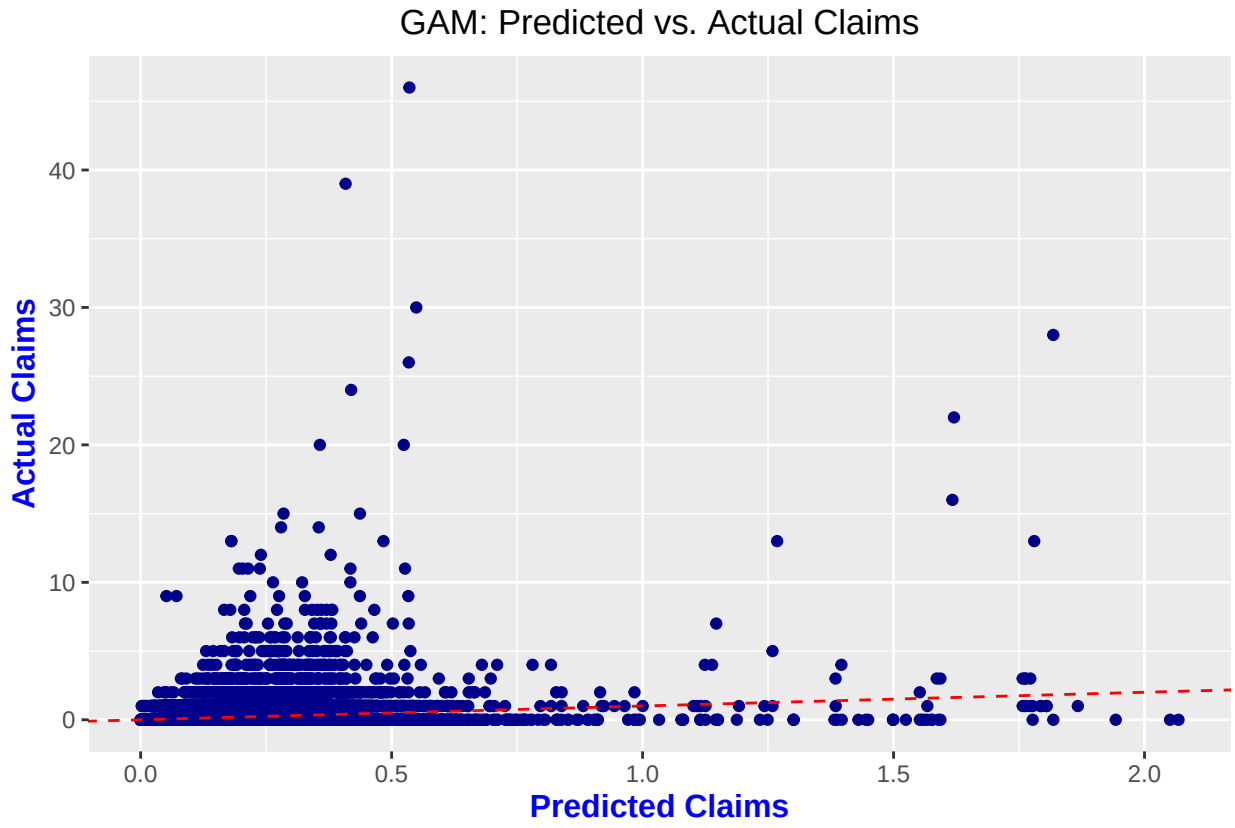

s(AGE_VEHICLE,7.83)
10
0
−10
−20
−30
0
10
20
30
40
50
AGE_VEHICLE

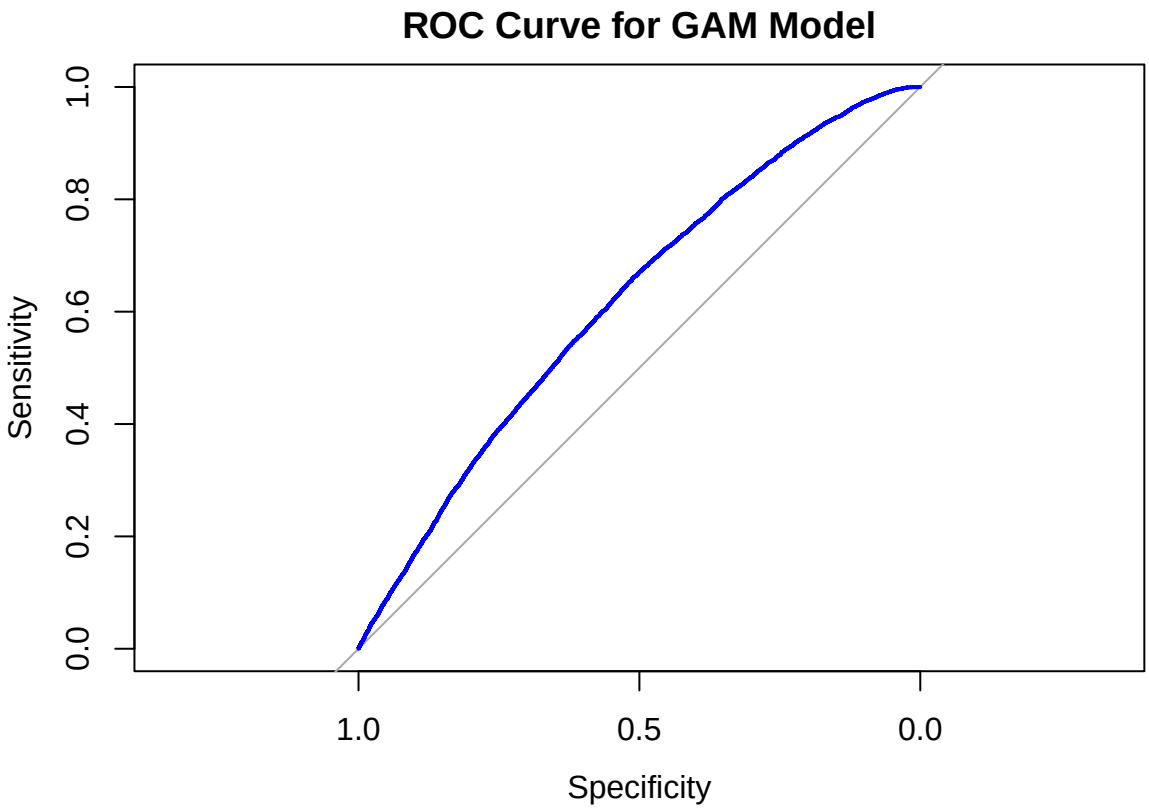

s(SEATS_NUM,8.94)
2
1
0
−1
−2
0
50
100
150
200
SEATS_NUM


s(CCM_TON,8.18)
1.5
1.0
0.5
0.0
0
5000
15000
CCM_TON


s(INSURED_VALUE,8.36)
1
0
−1
−2
−3
−4
−5
−6
0.0e+00
6.0e+06
1.2e+07
INSURED_VALUE


s(PREMIUM,8.91)
4
3
2
1
0
−1
−2
−3
0
50000
150000
PREMIUM


GAM: Predicted vs. Actual Claims
Actual Claims
40
30
20
10
0
0.0
0.5
1.0
1.5
2.0
Predicted Claims



Some interesting plots are shown, to illustrate how the predictors are related to the number of claims. The plots show the smooth functions of the predictors AGE_VEHICLE, SEATS_NUM, CCM_TON, INSURED_VALUE, and PREMIUM. The plots illustrate the non-linear relationships between these predictors and the number of claims, capturing the complex interactions and patterns in the data.

For AGE_VEHICLE, is steady until 40 years where it declines rapidly, indicating that older vehicles have fewer claims.

For SEATS_NUM, the number of claims increases with the number of seats related to "consumer vehicles", up to ~15 seats, where it starts to decline. There is a spike about 50 seats, probably related to commercial vehicles. The next valley and peak are related to very high number of seats, which are outliers and related to commercial or custom vehicles.

For CCM_TON, the number of claims decreases with the engine capacity untill 2500cc, which is where most of the vehicles are, from motorcicles and utility cars. After that, the number of claims increases, probably related to commercial and sport vehicles. Around 5000cc is a common engine size for commercial vehicles such as busses and trucks, after that the number of claims increases rapidly.

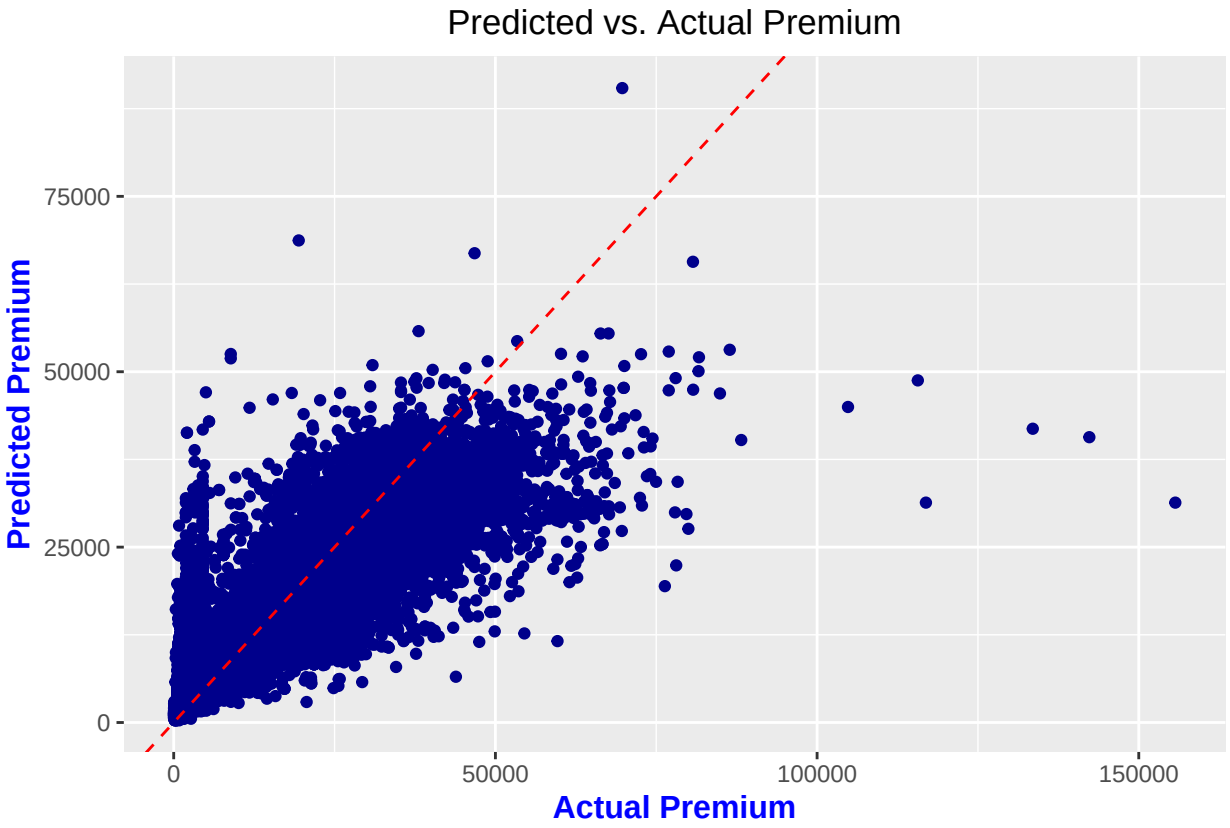
For INSURED_VALUE, the number of claims decreases with the insured value, indicating that more expensive vehicles have fewer claims. This is expected as more expensive vehicles are usually driven more carefully and less often.

For PREMIUM, the number of claims increases with the premium amount, starting in negative for the cheapest premium values, probably related to the insurance type and the insured value. After that, the number of claims increases with the premium amount, indicating that higher premiums are associated with more claims. This could be due to higher premiums for higher-risk drivers or vehicles or simply due to the increased value of the insured vehicles and the related need to keep it in good condition, making smaller defects a claim, which would not be claimed in cheaper vehicles.

Neural Network

Lead: Alvaro Cervan

A neural network model is fitted to predict the premium amount based on the characteristics of the insured vehicles and drivers. The model is trained using the cleaned and transformed data, and the results are analyzed to evaluate the model's performance.



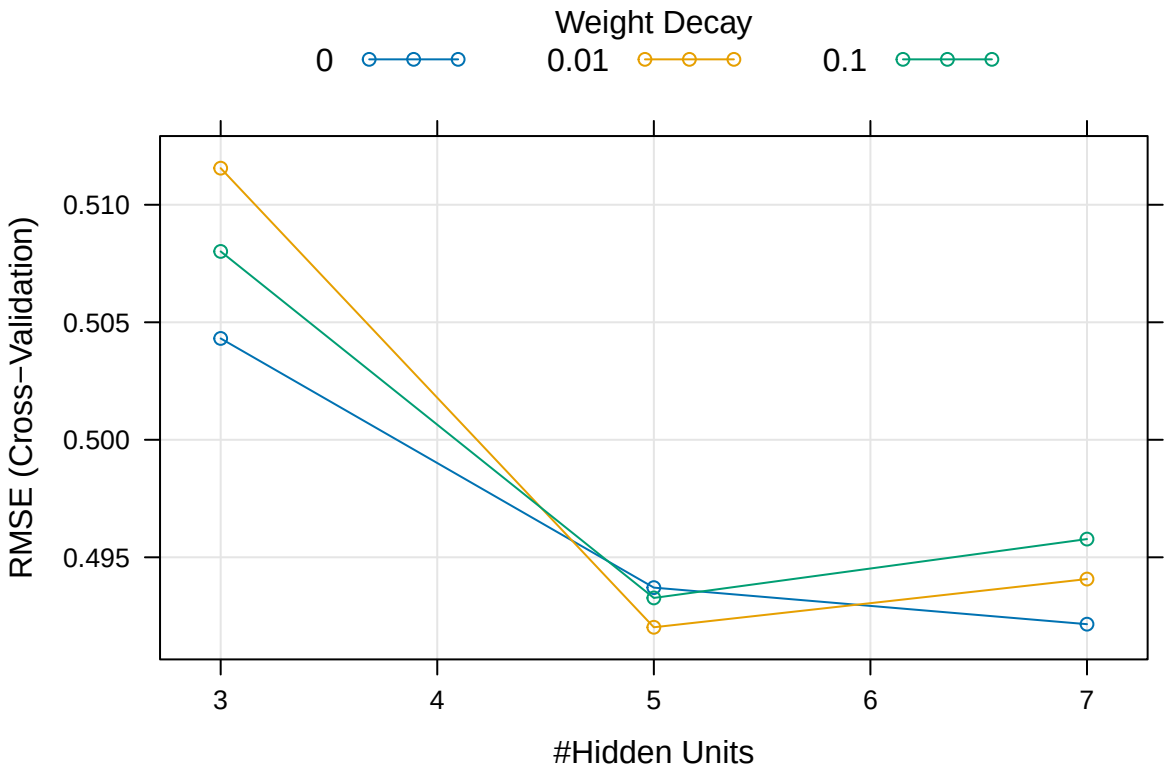
The neural network model was trained to predict the log-transformed premium amount based on the characteristics of the insured vehicles. The model was fitted using the nnet package, with the training data split into 80% training and 20% testing sets. The neural network model was trained with a hidden layer size of 5 neurons, linear output for regression tasks, and a maximum of 100 iterations for training.

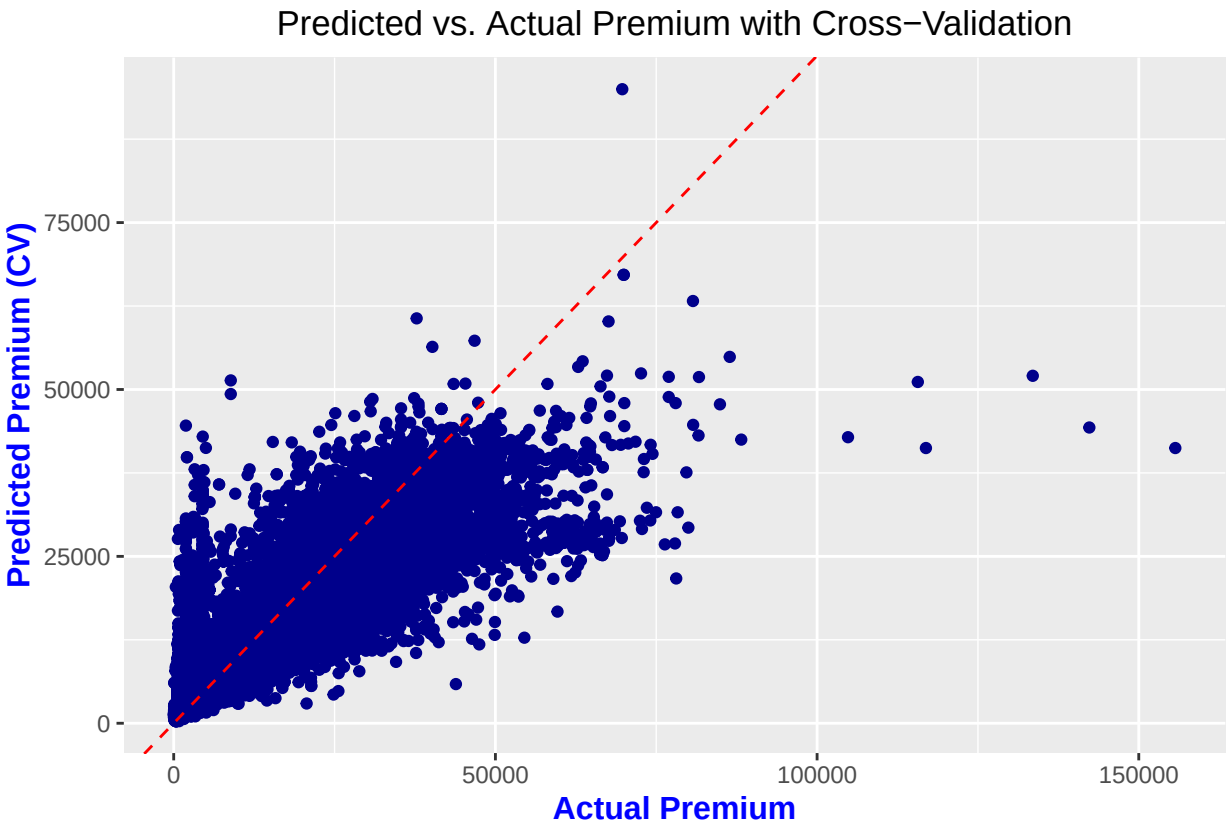
The plots of predicted vs. actual premium amounts show that the neural network model generally performs well in predicting the premium amounts. The points are clustered around the diagonal line, indicating a good alignment between the actual and predicted values. The model captures the general trend of the premiums, with some deviations for higher premium values. The model's performance can be further evaluated by considering additional metrics such as the R-squared value, RMSE, MAE, and MAPE, which provide insights into the model's accuracy and predictive power.

Neural Network Cross Validation

Nevertheless, we cannot be sure that those values for the model above are truly correct or it was luck that the model performs well at a first instance. To solve this question, the NN will be run again using **k-fold Cross Validation** with hyperparameter tuning. This approach will help ensure that the model's performance is robust and not due to overfitting or random chance. The k-fold Cross Validation will produce a more reliable estimate of the model's performance by splitting the data into k = 10 subsets, training the model on k-1

subsets, and validating it on the remaining subset. This process is repeated k times, and the results are averaged to provide a comprehensive evaluation of the model.





Results

Model	Mean Squared Error (MSE)	R-squared	Root Mean Squared Error (RMSE)	Mean Absolute Error (MAE)	Mean Absolute Percentage Error (MAPE)
Neural Network	0.2345529	0.7510002	0.4843067	0.3009332	3.488952 %
Neural Network Cross Validation	0.2343864	0.751177	0.4841347	0.2981682	3.462526 %

The weights decay plot shows how the RMSE (Root Mean Square Error) from cross-validation varies with the number of hidden units in a neural network for different weight decay values (0, 0.01, and 0.1). As the number of hidden units increases from 3 to 5, RMSE decreases across all weight decay values, suggesting improved model accuracy with additional capacity. However, beyond 5 hidden units, RMSE levels off or slightly increases, especially when weight decay is low or absent, indicating potential overfitting. Weight decay, a regularization technique to prevent overfitting, has a noticeable effect as the number of hidden units increases; while it slightly raises RMSE at lower hidden units, it helps to control error at higher hidden units. The optimal configuration, with the lowest RMSE, occurs at 5 hidden units regardless of weight decay, though weight decay of 0.1 becomes more beneficial as the model complexity increases, particularly at 6 and 7 hidden units.

The results from both the neural network and the neural network with cross-validation are very similar, with only minor differences in the evaluation metrics. This consistency suggests that the model is robust and performs well regardless of the validation method used. The cross-validation approach confirms the

reliability of the neural network model, indicating that it is not overfitting and generalizes well to unseen data.

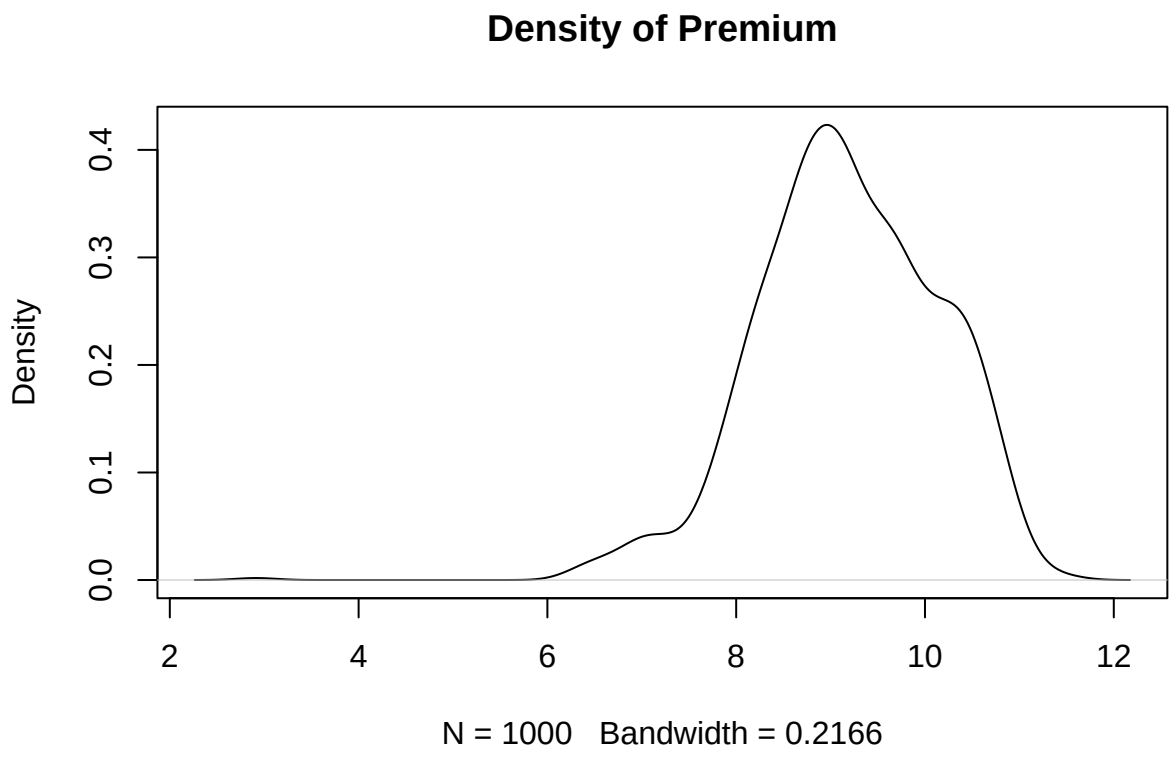
The evaluation of the neural network model revealed a Mean Squared Error (MSE) of 0.23, indicating the average squared difference between the actual and predicted log-transformed premium amounts. The R-squared value of 0.75 suggests that the model can explain approximately 75.10% of the variance in the log-transformed premiums, indicating a good fit to the data. The Root Mean Squared Error (RMSE) of 0.48 represents the square root of the MSE, providing a measure of the model's prediction accuracy. The Mean Absolute Error (MAE) of 0.30 indicates the average absolute difference between the actual and predicted log-transformed premiums. The Mean Absolute Percentage Error (MAPE) of 3.46% represents the average percentage difference between the actual and predicted premiums, providing a measure of the model's relative accuracy.

Overall, the neural network model demonstrates good performance in predicting the premium amounts based on the characteristics of the insured vehicles and drivers. The model captures the underlying patterns in the data and provides accurate predictions of the premium amounts. The evaluation metrics indicate that the model has a high level of accuracy and predictive power, which could make it a valuable tool for premium prediction in the insurance industry.

Support Vector Machine (SVM)

Lead: Luca Renz

For the scenario of SVM models, it has been decided to do multiple-classifications for the premiums and divide it into 4 levels from low to very high.



This analysis explores the performance of a multiclass classification task using an SVM model with a radial kernel. The goal was to classify PREMIUM_log into categories: "low," "medium," "high," and "very_high,"

employing hyperparameter tuning and parallel computation for efficiency. Data preparation included sampling 1000 entries from the cleaned dataset and visualizing the distribution of PREMIUM_log, followed by a 70/30 split for training and testing. It is important to note that the dataset is imbalanced, which may pose challenges in model training and evaluation, potentially impacting the reliability of certain performance metrics.

Model training was carried out with 10-fold cross-validation repeated three times, leveraging parallel processing to speed up the evaluation. A grid search was performed to find optimal values for the hyperparameters c (cost) and sigma, ensuring the model's robustness and generalizability.

The evaluation showed that the "very_high" category had the highest sensitivity at 0.73, while the "low" category excelled in specificity at 0.94 and positive predictive value at 0.81. Nevertheless, as accuracy, sensitivity, specificity and other measurements might be useful for balanced data, a different approach is used for unbalanced data as such. Therefore, the MCC was used and has revealed for each class strong performance overall as it can be seen below with the confusion matrix and corresponding MCC values.

Prediction	Low	Medium	High	Very High
Low	54	11	0	2
Medium	17	48	12	1
High	2	14	49	17
Very High	2	2	14	55

Category	MCC
Low	~0.69
Medium	~0.50
High	~0.49
Very High	~0.58
Overall	~0.57

The code process incorporated parallelized cross-validation and grid search, facilitating comprehensive hyperparameter tuning. The findings highlighted an overall accuracy of 0.69 and a Kappa statistic of 0.58, with McNemar's test yielding a significant P-value below 0.05 and mcc-value of roughly 0.57, suggesting a noteworthy difference from random classification.

Measurement	Value
Accuracy	~0.69
Kappa	~0.58
p-value	<0.05

To improve the model, checking the training set's class distribution and considering resampling techniques like random oversampling or SMOTE could be performed to further improve the model.

In conclusion, while the model showed strong results for the "low" and "very_high" categories, further optimization is needed for "medium" and "high" to enhance overall performance.

Nevertheless, the model demonstrates reliable performance in classifying insurance premiums into the four categories with MCC of about 0.73 indicating a moderate to strong correlation between prediction and actual category. Therefore, the robust model can be used to classify premiums. Further refinement of tailored features may improve overall performance, especially for medium and high categories.

Conclusion

This project successfully evaluated multiple machine learning models to improve insurance premium calculations for an Ethiopian insurance company. The analysis provided critical insights into predictive accuracy and practical applications, enabling data-driven recommendations.

Among the models evaluated, the neural network emerged as the most robust for predicting premium amounts, achieving an R^2 of 75.1% with minimal overfitting, as confirmed by cross-validation. Its ability to handle complex interactions and deliver high accuracy makes it the most reliable choice for implementation. The linear regression model, while simpler, also performed well, explaining 73.08% of the variance in premiums. However, challenges with heteroskedasticity and residual normality limit its utility, especially for larger premium values.

For claims prediction, the Poisson model provided valuable insights into claims frequency and highlighted the influence of factors such as vehicle age and insured value. However, systematic underestimation for higher claims suggests that alternative approaches, such as Zero-Inflated Poisson models, could address these shortcomings. Similarly, the binomial logistic regression model was effective in identifying predictors of claim payouts, but its limited explanatory power ($R^2 \approx 4.8\%$) indicates a need for further refinement.

The support vector machine (SVM) demonstrated strong potential for classifying premiums into discrete categories, achieving an accuracy of 77.7%. Its strength lies in distinguishing "low" and "very_high" premium categories effectively. However, its performance for "medium" and "high" premiums could be improved through better feature engineering and data balancing techniques. The generalized additive model (GAM) captured non-linear relationships in the data, but it did not outperform simpler models like linear regression or neural networks in terms of accuracy, limiting its practical application.

Based on these findings, we recommend adopting the neural network model for premium prediction due to its high accuracy and flexibility. For claims prediction, incorporating a Zero-Inflated Poisson model would provide a more nuanced approach to handling data with a high frequency of zero claims and outliers. Additionally, the SVM model is well-suited for premium categorization and can aid in customer segmentation, especially if combined with strategies to enhance accuracy for the "medium" and "high" categories.

In conclusion, integrating insights from these models into pricing policies will enable the client to implement fair and accurate premiums, optimize risk management, and improve overall profitability. Factors such as insured value, vehicle brand, and age should be strategically incorporated into pricing and marketing strategies. With these data-driven enhancements, the company is well-positioned to leverage machine learning for sustained competitive advantage in the insurance market.

Usage of Generative AI

In the group project, generative AI was employed to facilitate coding tasks, generate text, and clarify complex concepts. This technology proved beneficial for automating repetitive tasks and assisting in the assembly of report sections, especially in presenting complex ideas in a clear manner.

However, challenges were encountered in the precise formulation of prompts; imprecise prompts occasionally led to AI-generated solutions that did not meet specific project needs. Consequently, all AI-generated outputs required thorough verification to ensure their relevance and accuracy. In some instances, modifications were necessary to align the AI-produced code with project specifications or to optimize performance. The text generated by the AI also needed careful examination to confirm its alignment with project objectives and adherence to academic standards.

Generative AI struggled with tasks requiring deep contextual understanding or specialized knowledge unique to the project. While it significantly enhanced productivity and facilitated the drafting process, active human oversight was crucial to not apply irrelevant or incorrect changes.

Verification of AI suggestions against trusted sources and empirical data was consistently performed, particularly in the context of complex statistical analyses and interpretations.Using AI offered considerable

advantages but required a focused and hands-on approach to fully leverage its capabilities in the academic context. Summing up, it has definitely supported the team in terms of explaining difficult concepts while also increasing efficiency.