**Predicting Auto Claims**

**White Paper**

Donald Price

Bellevue University

DSC 680: Applied Data Science

Catherine Williams

**Business Problem & Background**

Auto insurers paid $198 billion for auto claims incurred in 2021 (Insurance Information Institute, 2023). Underwriting margins (premiums from policyholders less claim payments and company expenses) are tight. In 2021, the auto insurance industry paid 1.9% more in claims and expenses than they collected in premiums. Claims represent nearly 80% of an insurer’s expenses. Insurance carriers have a strong profit incentive to accurately predict claim costs per policyholder and price the insurance policy appropriately. Developing a model to identify risk rating factors would greatly improve a carrier’s ability to improve profit margins.

**Stakeholders and Goals**

Our model will attempt to predict a policyholder’s likelihood of getting into an auto accident. We will model our solution as a binomial classification (predicted loss vs. predicted loss-free). Confidence intervals or predicted probabilities will help our clients, the insurers, better understand which policyholders are more likely to generate a claim. Insurers will combine claim frequencies from our model with their own estimate of damages per claim to produce an expected loss estimate for each policy. The loss estimate plus a provision for expenses and profit will determine the proposed premium. Our model will meet an insurer’s need if they can produce a more accurate, estimated premium when using our model compared to what they use currently.

**Data Explanation**

The data set originates from an unnamed auto insurer who provided nearly 60,000 policyholder records. Each record identifies whether the policyholder incurred an accident during that policy period. The data set is believed to contain claims data for US policyholders as one of the data attributes provides a safety rating issued by the National Highway Traffic Safety Administration (NHTSA), which is a US federal agency.

The data set includes anonymized information about the policy holder (driver), location of the car, and details on the car covered by the policy. In total, there are 80+ potential fields for model building covering the key risk properties of an auto policy.

Standardization of features won’t be required during preprocessing as this step was performed before releasing to the public. Preprocessing work for modeling will include removal of ids and correlated features.

**Modeling/Analytical Approach and Methods**

The model will predict whether a policyholder gets into an accident (i.e. auto frequency). Random forest and XGBoost classifiers will be used, both tuned for optimal results. Exhibit 1 shows an imbalanced target class, which requires resampling of our training data to produce a reliable model. The initial approach will use SMOTE, a synthetic oversampling technique to balance classes.

**Exhibit 1: Distribution of Target Variable**

Chart, bar chart

Description automatically generated

Models will be evaluated using the F1 score. This metric focuses on how many claims we accurately predict while minimizing our false predictions. This measure better fits the insurers use cases as the F1 metric targets accuracy for predicting a claim.

**Analysis**

We hypothesize that driver characteristics are key variables for predicting an auto claim. The data set provides policyholder age. Exhibit 2 does suggest a slight relationship between age and claim occurrence.

**Exhibit 2: Policyholder’s Age and the Influence on Claim Activity**

Chart, box and whisker chart

Description automatically generated

Exhibit 3 shows three primary variables for predicting claims: age of car, policy tenure, and policyholder age. As noted above, we anticipated the policyholder’s age would provide predictive ability. We would expect younger, less experienced drivers to cause more accidents and potentially older drivers with less physical response times to also cause more accidents. Age of car is likely suggesting that newer cars carry additional safety features that are impacting the likelihood of claims. Lastly, policy tenure impacts likelihood of claims. This makes sense as a customer with relatively few claims will be encouraged by the insurance company to renew. Conversely, someone who gets into many accidents will likely get cancelled by the insurance company.

**Exhibit 3: Factors Influencing the Model’s Predictions**

Chart, bar chart

Description automatically generated

**Conclusion**

Our target metric was the F1 score. This metric focuses on the accuracy of predicting when a claim is likely to occur, which is the primary focus of an insurance company. The target is imbalanced with 6% of policyholders getting into an accident, so accuracy was avoided as a useful metric.

The final random forest model achieved an F1 score of 0.17. While low in absolute standards, the model produced a 50% better result than predicting that all policies will have a claim. Exhibit 4 demonstrates the improved identification of claims when the model predicts a claim will occur.

**Exhibit 4: Model’s F1 Score on Test Set**

Table

Description automatically generated

**Assumptions**

The model includes relatively few attributes of the driver. Any changes to the customer mix is likely to impact the claims frequency. Additionally, underlying environmental trends (such as increasing use of cell phones and distracted driving) would not be identified by the model.

**Implementation, Future Uses, & Recommendations**

Depending on existing models, the result seems strong enough to use for pricing a policy. There could be additional improvements made by adding additional information about the driver. According to the Insurance Information Institute, the most significant variables for rating a policy’s risk relate to the driver (Insurance Information Institute, 2023).

**Limitations & Challenges**

Missing variables lessen the accuracy of this model. The claims data set provides only 1 attribute related to the policyholder and driver. It’s highly likely that claims are influenced by driver experience and tendencies. Excluding these variables reduces the overall accuracy. Geographic variables may also be important. Congested areas, areas with significant weather conditions, and improper road maintenance or construction may affect driving risks.

Discussions with insurance leaders should explain the data limitations and encourage leaders to provide these variables for modeling.

**Ethical Assessment**

Few ethical issues are present with this project. The Kaggle data set has been pre-scrubbed to remove any sensitive or personal information. The collective data provided is insufficient for re-identifying the policyholder. Only the training data will be used, because the testing data set does not specify whether the policyholder had an accident.

Uses of this information are ethical. Insurers already make assessments based on the personal characteristics of the driver and car. In this data set, age is the only factor provided and that is an acceptable rating factor for carriers.

**Sources**

Most of the data used for this project will come from the Kaggle data set. Details on these sources and supplemental reading materials are listed in the ‘References’ section. Additional information from insurance companies, such as the factors outlined by Allstate on their consumer education site (Allstate 2023), will be used to assess the outputs against the data carriers already use to rate policies. An article from Forbes (Kilroy and Metz, 2022) states that credit ratings are still heavily used by auto insurers. This suggests auto insurers may not fully be leveraging the data they own for ratings, which leaves models like the one we intend to create during this project as highly valuable to these companies.

**Appendix**

**Appendix I:** Modeling Features – Categorical Variable Count Plots

Diagram

Description automatically generated

**Appendix II:** Modeling Features – Numerical Variable Count Plots

Chart, bar chart

Description automatically generated

**References**

Allstate (2023). What factors may affect your car insurance premium? *Allstate Website.*

h<ttps://www.allstate.com/resources/car-insurance/what-affects-premiums-and-r>ates#:~:text=Some%20factors%20that%20may%20affect,car%20and%20your%20driving%20record.

Insurance Information Institute (2023). Facts + Statistics: Auto Insurance. *The III Institute.*

<https://www.iii.org/fact-statistic/facts-statistics-auto-insurance>.

Insurance Information Institute (2023). What Determines the Price of an Auto Insurance Policy. *The III Institute.* https://www.iii.org/article/what-determines-price-my-auto-insurance-policy

Kilroy, Ashley and Metz, Jason, March 10, 2022. 9 Factors that Affect Your Auto Insurance

Rates. *Forbes.* https://www.forbes.com/advisor/car-insurance/factors-in-rates/.