Auto Insurance Claim Prediction

IEOR 242 Group Project
Team Members:
Yike Chen
Yiting Gan
Keke Lin
Anna Xu

1. Motivation

The motivation for developing models to predict auto insurance claims is to assist clients and companies in gaining a comprehensive understanding of the variables that impact auto insurance premiums and how to effectively reduce associated costs. The primary objective of this model is to facilitate consumers' identification of the key determinants that govern their auto insurance claim amounts, including age, gender, location, and number of policies, etc. By gaining mastery over this information, clients can make more informed decisions when selecting insurance types and providers, ultimately resulting in cost savings during insurance acquisition. Moreover, insurance companies can gain insights into trends and changes in the auto insurance industry. This can help them develop more competitive insurance products, better meet the ever-changing needs of customers, and improve overall business performance. Overall, we aim to use our model to improve the efficiency and effectiveness of the entire auto insurance industry as much as possible and drive positive change within the industry.

2. Data

2.1 Data Source

Our project used a published dataset online about claims of one auto insurance company. As we observed the dataset, we noticed that it contains attributes representing state, and we decided to add external factors that may provide background information of different states to assist us when we analyze the data. Those extra factors are Consumer price index (CPI), Gross domestic product (GDP), and Accident Fatality rate. The sources of them are credential, because they are released from the U.S. government.

After merging these tables by states, the dataset expanded from 26 to 32 attributes with 9134 records. There are numerical and categorical factors. In order to further analyze the distribution of data before building the model, we conducted Exploratory Data Analysis (EDA).

2.2 Exploratory Data Analysis

Through EDA, we aimed to discover the relationships between the claim amount and 8 categorical variables. We used Python to visualize the results using boxplots. The following four graphs have the same properties: similar median and distribution of views. These properties represent that the results of these categorical attributes do not make much impact on claim amount. In other words, Claim amount is unlikely to be affected by which state you live in, or the response from the insurance company, or the gender of customers, or the marital status of customers. Because state does not influence claim amount, the three additional factors added early based on state could be removed.

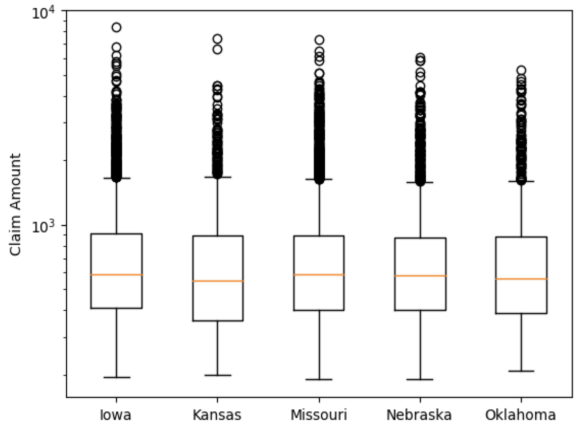

Figure 1: Claim Amount v.s. State

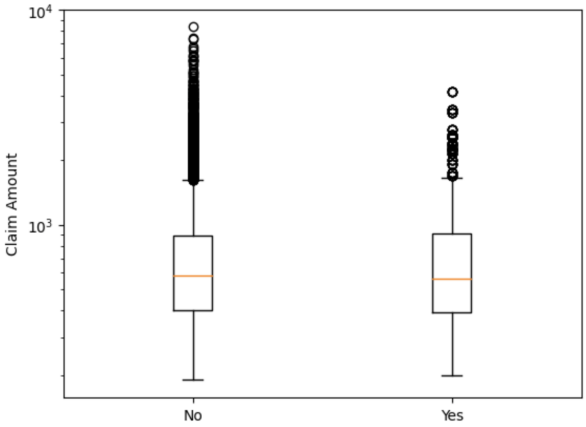

Figure 2: Claim Amount v.s. Response

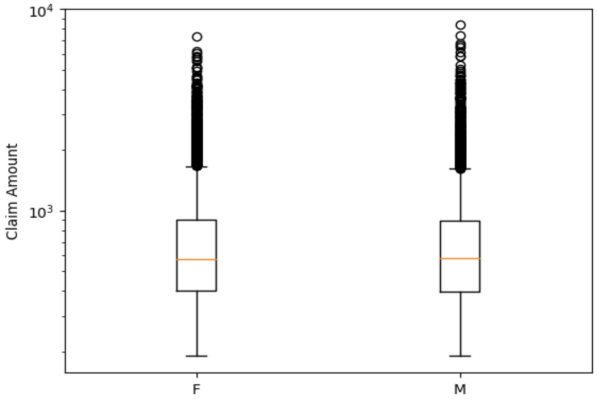

Figure 3: Claim Amount v.s. Gender

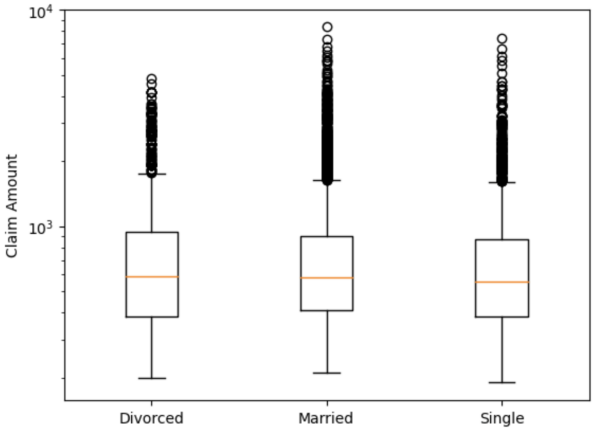

Figure 4: Claim Amount v.s. Marital Status

Other factors affecting claim amount include policy types, claim reasons, coverage, and vehicle class. The following are the corresponding graphs:

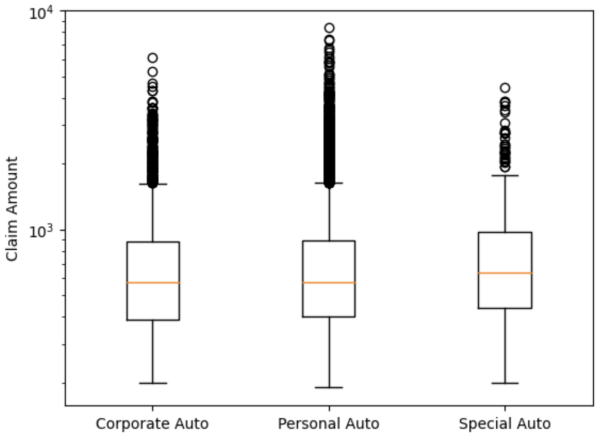

Figure 5: Claim Amount v.s. Policy Type

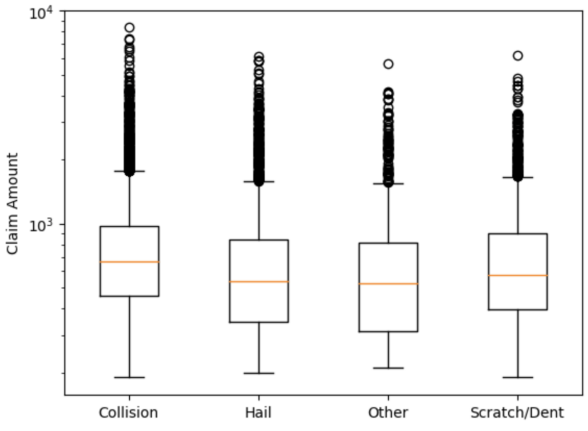

Figure 6: Claim Amount v.s. Claim Reason

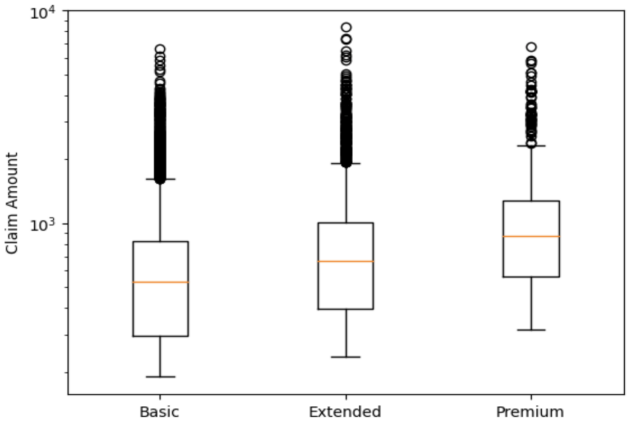

Figure 7: Claim Amount v.s. Coverage

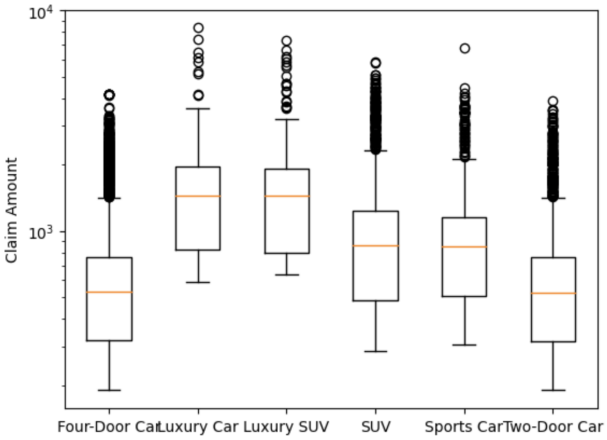

Figure 8: Claim Amount v.s. Vehicle Size

From our observations of the four diagrams, while they share similar shapes, the differences in median and data distribution should be further explored during the analysis process. In terms of policy type, the special auto type generally has a higher claim amount compared to the other two types, which have similar medians and distributions. This suggests that the special auto type is much more expensive than the other two. When it comes to claim reasons, collision and scratch/dent claims are relatively short in comparison to others, indicating that most accidents of this nature have similar claim amounts. Furthermore, due to their relatively high medians, these two claim reasons will result in higher claim amounts. In regards to coverage, the box plots display similar shapes with increasing median claim amounts for different coverage plans. For vehicle size, we noticed that some box plots have unevenly sized sections. For instance, the luxury SUV box plot indicates that for this car size, it has similar claim amounts at certain parts of the scale, but in other parts of the scale, the claim amount is more variable in terms of the car size. The long upper whisker in this example signifies that the claim amount is varied amongst the most positive quartile group, while it is very similar for the least positive quartile group.

3. Analysis Model

In this part, we are going to introduce the model we choose to analyze and predict the claim amount, which includes a baseline model, linear regression model and CART model with cross-validation and Random Forest model with cross-validation.

3.1 Baseline Model

The baseline model we choose is simply taking the mean of claim amount in the training dataset.

RMSE	MAE	OSR2
680.581	444.041	-0.000

Table 1: Test Result of Baseline Model

And the result for the baseline model is shown in the chart above, we can see that the OSR2 for the baseline model in the test set is 0.

3.2 Linear Regression

For the linear regression model, we take all the variables in the training data except for the "Claim Amount". And then we select the top 10 variables with the highest coefficient and then sort them in the graph below.

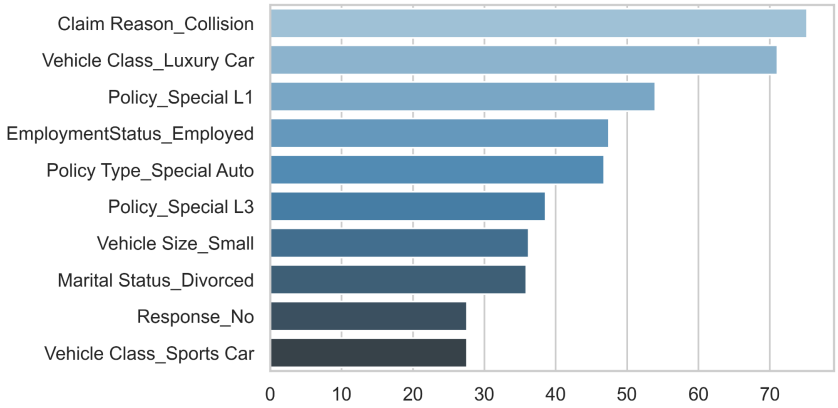

Figure 9: Coefficients of Linear Regression

We can see that "Claim Reson_Collision" and "Vehicle Class_Luxury Car" are valued the most in the linear regression model, which means that the collision type of claim reason and luxury cars will contribute to predicting the claim amount.

RMSE	MAE	OSR2
633.379	390.661	0.134

Table 2: Test Result of Linear Regression Model

From the result in the test set, we know that the OSR2 is slightly improved than the baseline model.

3.3 Regression Tree with Cross-Validation

We then apply the classification and regression tree model (CART) with the cross validation.

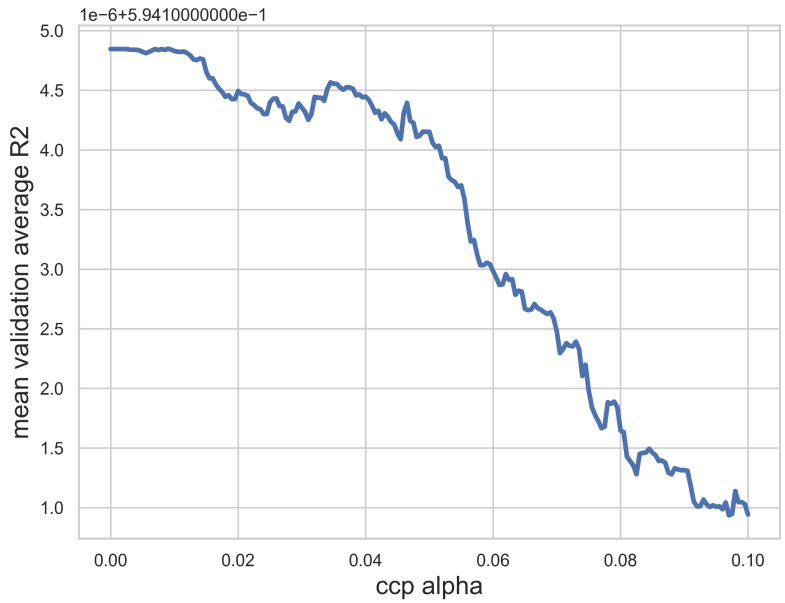

Figure 10: ccp alpha v.s. R^2

We explore the CCP value at a range of [0, 0.1] with 100 sample points. And we find that the CART model performs best with the highest R square when the ccp value is 0.009.

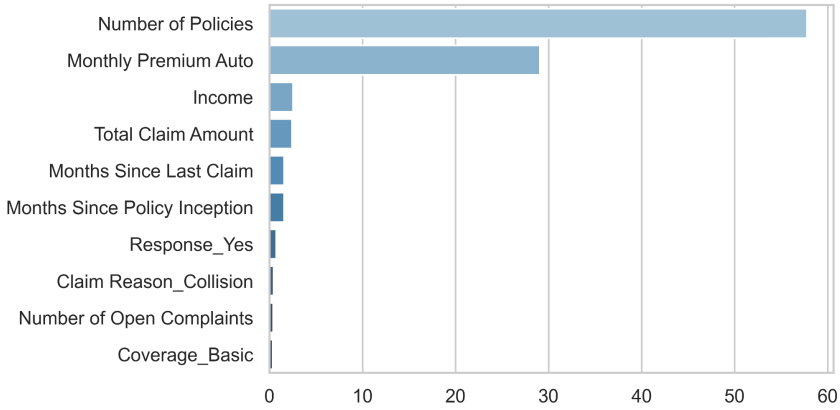

Figure 11: Feature Importance of CART

The feature importance is sorted and listed in the figure above. We find that the most important attributes that the CART model values are the “Number of Policies” and “Monthly Premium Auto”.

RMSE	MAE	OSR2
441.008	167.826	0.580

Table 3: Test Result of CART Model

The OSR2 for the CART model with cross validation is significantly better than the linear regression model, which is 0.580.

3.4 Random Forest with Cross Validation

For the random forest model, we tune the max features between 15 and 45, and then we get the results as follows:

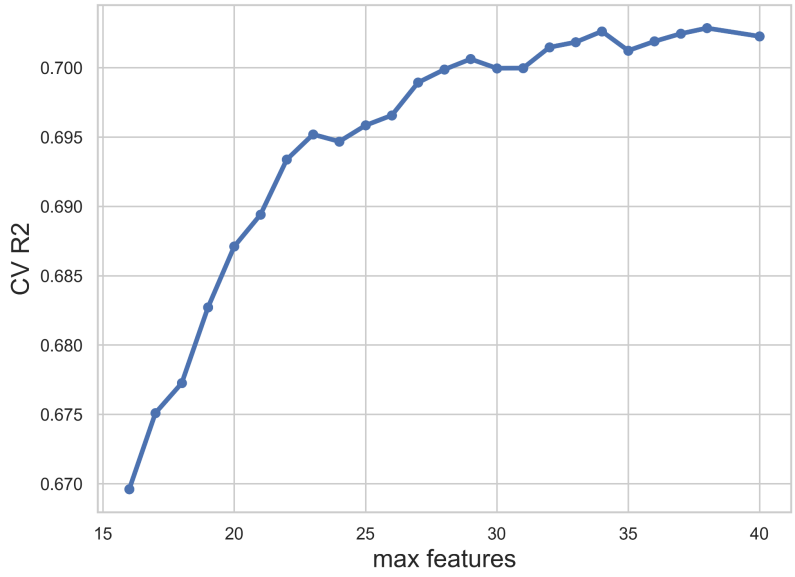

Figure 12: Max Features vs. CV R2

We find that the random forest model achieves its best performance when the max feature is 38 and the R square is maximum.

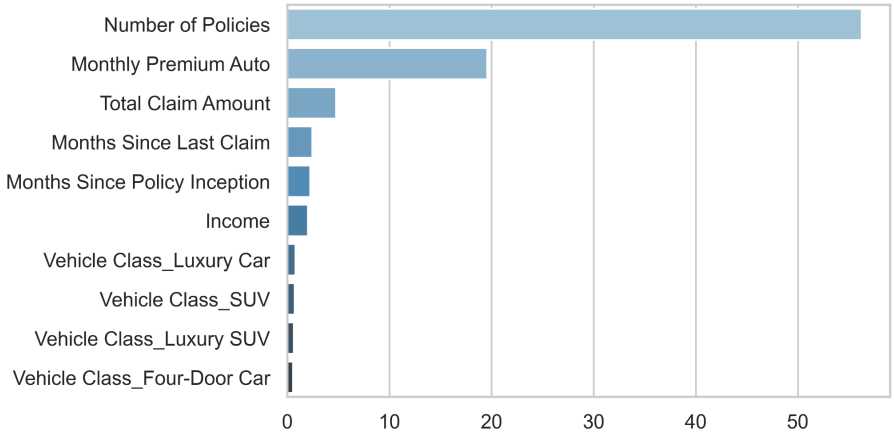

Figure 13: Feature Importance of Random Forest

The feature importance graph shows that the random forest model with cross validation puts more emphasis on “Number of Policies” and “Monthly Premium Auto” which is similar to the decision tree model.

RMSE	MAE	OSR2
376.199	154.594	0.694

Table 4: Test Result of Random Forest Model

The OSR2 for random forest is 0.694, which is the highest for all the models we have trained so far.

3.5 Blending

We utilize ordinary least squares to blend the model we have trained, and get a blending model as below.

Blending = -0.0005 * Linear Regression + 0.1210 * CART + 0.8861 * Random Forest

RMSE	MAE	OSR2
377.891	155.942	0.692

Table 5: Test Result of Blending Model

The test result for the blending model shows that the OSR2 is 0.692 which is slightly lower than the Random Forest model. So in the following section, we will extend our analysis based on the random forest model.

4. Bootstrap Analysis

To prove the reliability of our random forest model in the test result, we run a bootstrap analysis on 1000 samples for the three metrics we choose, MAE, RMSE and OSR2.

4.1 MAE

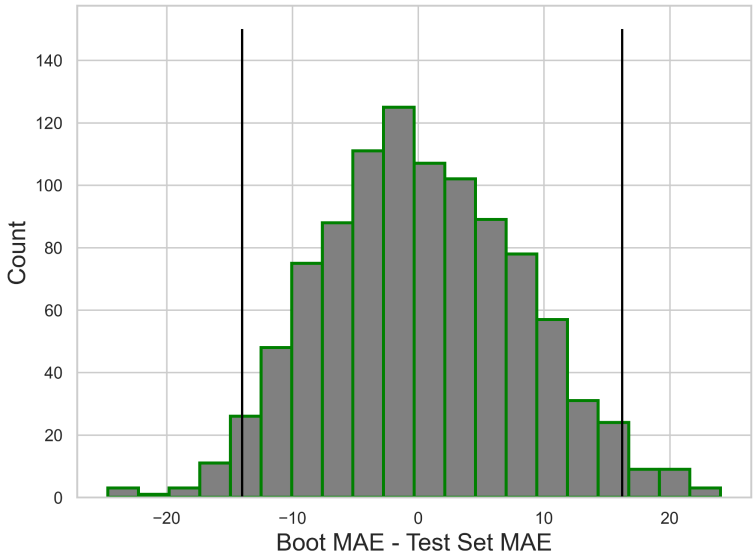

Figure 14: 95% Confidence Interval for Boot MAE v.s. Test MAE

The graph shows that the 95 % confidence interval of Bootstrap MAE - Test SET MAE contains 0, which means that the MAE of the test set in the previous section is trustable.

4.2 RMSE

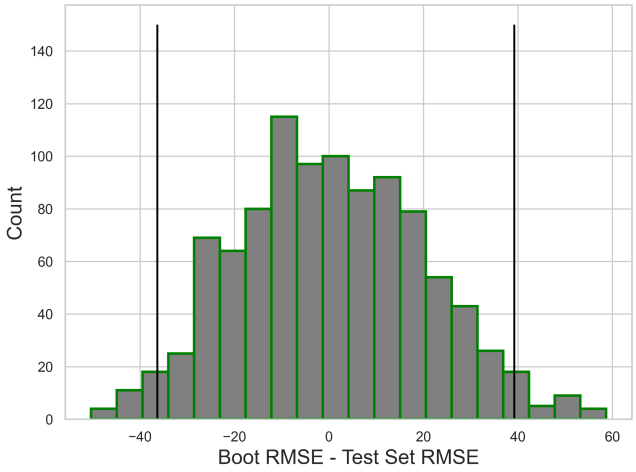

Figure 15: 95% Confidence Interval for Boot RMSE v.s. Test RMSE

We perform the same analysis on the RMSE and find that the 95 % confidence interval of Bootstrap RMSE - Test SET RMSE contains 0, which means that the RMSE of the test set in the previous section is trustable.

4.3 OSR2

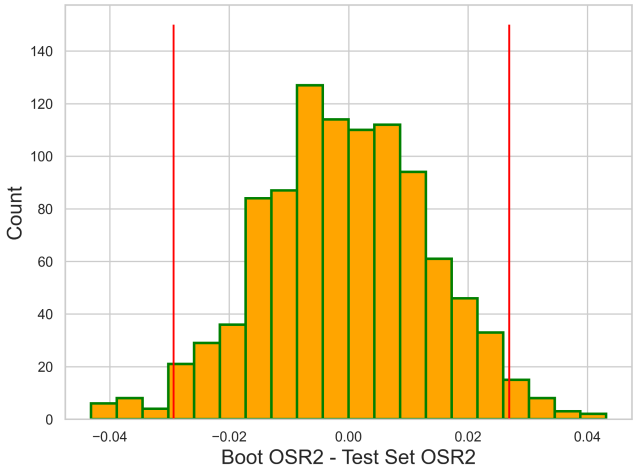

Figure 16: 95% Confidence Interval for Boot OSR2 v.s. Test OSR2

Lastly, we notice that the 95 % confidence interval of Bootstrap OSR2 - Test SET OSR2 contains 0, which means that the OSR2 of the test set in the previous section is also trustable.

5. Result Analysis

From four models we implemented which are Linear Regression model, Regression Tree with Cross Validation model, Random Forest with Cross Validation model and a blending model. The Random Forest with Cross Validation model outperformed with the highest OSR^2 of 0.694. In the RFCV model, the number of policies and monthly premium auto are considered the most important factors that contribute to the claim amount. It is reasonable to account for the number of policies to be one of the important factors since it could reveal insured risk profiles such as risk-averse, risk-taking, etc. Monthly premium auto is another important factor since it directly

reflects clients' assets value. It is closely related to claim amount since clients with more valuable assets may be prone to file claims since the cost would be comparatively high. Insurance companies may utilize these characteristics when personalized auto insurance for clients. Based on the results from all models, a severe claim reason such as collison filed by a special policy holder who has a luxury car is considered as important factors. All three models' feature importance analysis have similar results on that.

6. Impacts, Limitations and Future Potentials

The prediction models generated can benefit both clients and insurance companies. On the one hand, insurance companies can use the model to better estimate the risk of insuring different drivers and vehicles, which can lead to more accurate pricing of premiums. This can help insurance companies stay profitable and avoid underwriting losses. On the other hand, policyholders can benefit from more personalized pricing, with premiums more closely aligned with their risk profile.Additionally, with a more accurate estimation of the claim amount, insurance companies can better allocate resources to handle claims, reduce claim processing times, and provide more timely and accurate payouts. The impact of a model to predict auto insurance claim amounts can vary across different subpopulations of interest. This is because different subpopulations may have different risk factors and driving behaviors that can affect their likelihood of filing a claim, as well as the amount of the claim.For example, younger drivers may have a higher risk of getting into accidents and filing claims compared to older drivers, due to less driving experience and higher rates of risky behaviors. Similarly, drivers in certain geographic regions may face different driving conditions that affect their risk of filing a claim, such as areas with high rates of accidents or theft.

Based on the analysis using the data sources, there may be some limitations since there may be missing data or comparatively not reliable data due to potential self-reporting. As we stated for our dataset, the dataset is limited to only 5 states. In order to expand the scope of our analysis, in the future, we may collect the entire US dataset on auto insurance claims. And furthermore, we need to assess our resources and gather more comprehensive data. As we expand the research, there may appear some potential negative consequences such as privacy information leak and bias and discrimination as certain groups of people are unfairly penalized by the model.

Appendix:
The demo code and original data file are included in our github repository:
https://github.com/sh0829kk/IEOR242/blob/main/analysis.ipynb