Project Progress Report

Predictive Modeling of Auto Insurance Claims

Cyril Odiwuor
Patrick Odongo
Nixon Okiomeri
Jacinta Babu

Project Progress Overview

The tasks which have been completed so far are:

a) **Data Acquisition and Preprocessing.** In this stage, we were able to find the data, clean it, and apply preprocessing to the data. We split the data into training and testing data. The data was in the form Training set size (8000, 15) and Testing set size (2000,15). We were able to explore and plot graphs on missing values of the credit score column and annual mileage column as seen below in Fig. 1 and Fig. 2, respectively:

Figure 1: Credit Score Column Count

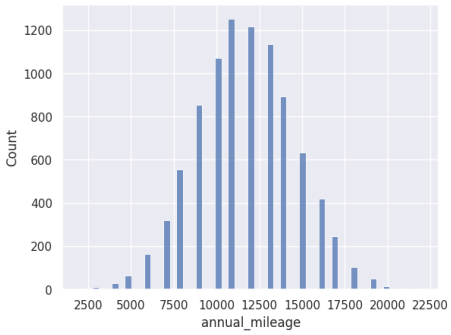

Figure 2: Annual Mileage Column Count

It would not be a good idea to use the median or mean to fill in missing values for both graphs above, as they will distort the distribution.

We explored outliers in the data and plotted them out as seen below in Fig. 3:

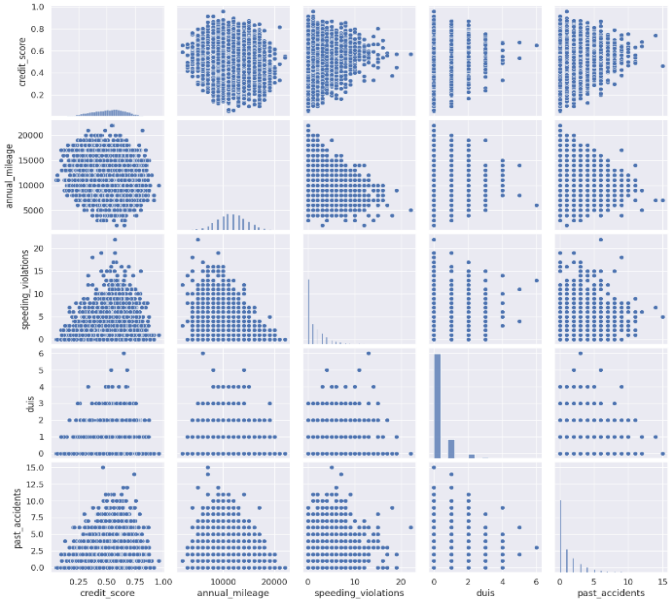

Figure 3: Outliers in the dataset

There are some outliers across our suspected columns, but there also does not seem to be a linear relationship between them such that we are unable to determine the trend line to best fit them. We will need to scale the data later.

We established that the target variable will be **outcome** from the dataset. We came up with the plot for the distribution of the outcome of taking a claim or not, as shown below in Fig. 4:

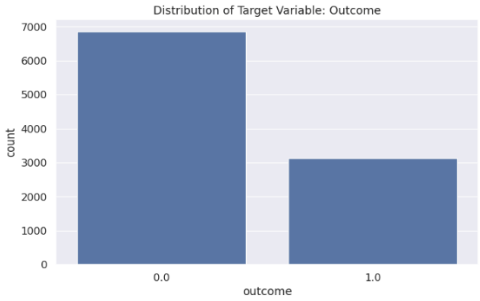

Figure 4: Distribution of Target Variable: Outcome

On the x-axis, 0.0 represented the outcome of not taking a claim, while 1.0 represented the outcome of taking a claim. The likelihood of taking a claim is lower than that of not taking one.

b) **Feature Engineering**. In this stage, we used polynomial features, which help capture **non-linear patterns** that linear models might miss.
c) **Model Development and Training**. In this stage, we tried to implement an array of multiple regression techniques by using Root Mean Square Error as a comparison metric of the models first. A lower root mean square error is better for a model, and we compared them using a bar graph as shown below in Fig. 5:

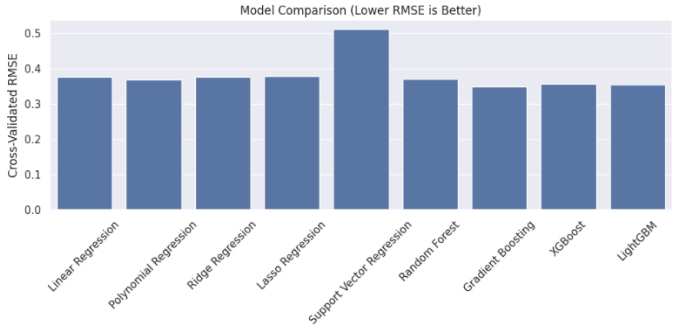

Figure 5: Distribution of Target Variable: Outcome

The results made it pretty clear which models stood out. Gradient Boosting took the top spot with the lowest RMSE, proving to be the most reliable at predicting insurance claims. XGBoost wasn't far behind, landing in second place, with Polynomial Regression coming in third. Some of the other models showed promise, but they couldn't quite match the accuracy of the front-runners.

Challenges

We faced challenges while looking for the right dataset to suit our needs. Moreover, we faced an uphill task of dealing with missing data and cleaning up the data. We consulted each other and came up with the target variable and the dependent variables which will be used in the project. Model training and applying each model to come up with the result is another challenge we are facing, and we are working towards achieving the phase of model evaluation and comparison. We have addressed these challenges by dropping the columns with the missing data, since it will hinder the modeling of the data.

Collaboration

We get together twice a week to go over any roadblocks anyone's running into with the dataset. Everyone's pitching in and sharing ideas to keep things moving forward.

Next Steps

The next steps are to do Model Evaluation and Comparison. This includes applying each model to the test data, calculating and comparing performance metrics, visualizing predicted results, analyzing any model residual, and assessing model explainability using SHAP values. Moreover, we are going to fine-tune the other steps to see if there are areas we can improve in the project.