Car Insurance Claims

Numann Malik

Project Description

- We analyzed behavioral data for the customers of a car insurance company.
- The goal was to identify key factors that are more likely to lead a customer to filing an insurance claim (e.g. for an accident).
- Our stakeholder would want to determine in particular, which customers are more accident prone and hence likely to file a claim.

Car Insurance Dataset

The dataset was sourced from the <u>Kaggle – Car Insurance</u> Dataset from the user 'Sagnik Roy'. This dataset utilizes 18 features for predicting auto insurance claims. The 'Outcome' target feature below indicates whether a customer filed a claim (1) or not (0).

| Feature Name	Description	Feature Name	Description	Feature Name	Description
Age	16-25, 26-39, 40-64, or 65+	Credit Score	Between 0 and 1 (scaled)	Annual Mileage	Rounded to nearest 1000 miles
Gender	Male or Female	Vehicle Ownership	0 (no) or 1 (yes)	Vehicle Type	Sedan or Sports Car
Race	Majority or Minority	Vehicle Year	Before 2015 or After 2015	Speeding Violations	Numerical Frequency
Driving Experience	0-9, 10-19, 20-29, or 30+ years	Married	0 (no) or 1 (yes)	DUIs	Numerical Frequency
Education	High School, University, or None	Children	0 (no) or 1 (yes)	Past Accidents	Numerical Frequency
Income	Poverty, Working Class, Middle Class, or Upper Class	Postal Code	10238, 21217, 32765, or 92101	Outcome	0 (no) or 1 (yes)

Stakeholder

- Our analysis will be utilized by a car insurance company
- This is to assist with deciding which potential customers to provide coverage to, and how much to charge for their premiums.
- Covering the cost of road accidents can be very expensive. Therefore, it is very important to investigate the history of current/former customers.

Relevant Observation

Relationship between owning a vehicle and making Insurance claims
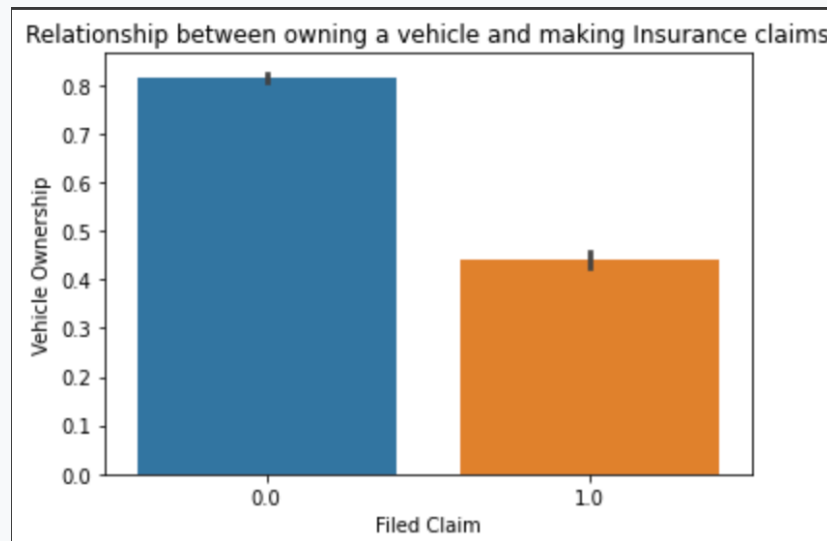
- Over 80% of customers who have never made an insurance claim owned their vehicle
- More than half of customers who don't own a vehicle made a claim
- Supporting evidence that those who lease their vehicles are more likely to file claims

Another Relevant Observation

Scatter Plot between Past Accidents and Speeding Violations
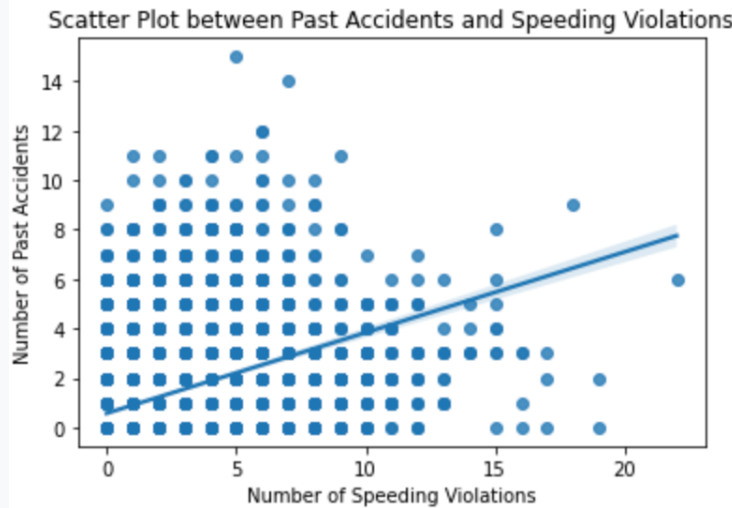
- There is a positive correlation between the number of past accidents and speeding violations.
- We should determine both for each potential customer before deciding to grant any coverage.

Model Development

Several Models were evaluated on the dataset:
- Decision Tree Classification
- Random Forest Classification
- K-Nearest Neighbors Algorithm
- Logistic Regression
- Gradient Boosting
- LightGBM
- XGBoost

Multiple metrics were used for evaluation:
- Accuracy
- Precision
- Recall
- F1-score

Model Evaluation

```
              precision    recall  f1-score   support

        0.0       0.90      0.89      0.90      1716
        1.0       0.77      0.79      0.78       781

   accuracy                           0.86      2497
  macro avg       0.84      0.84      0.84      2497
weighted avg      0.86      0.86      0.86      2497
```

- Our stakeholder, the car insurance company, will want the model that minimizes the proportion of false negatives.
- The reason is that it will cost more money to pay out for claims that were not accurately predicted.
- Therefore, we mostly we care about the recall score.
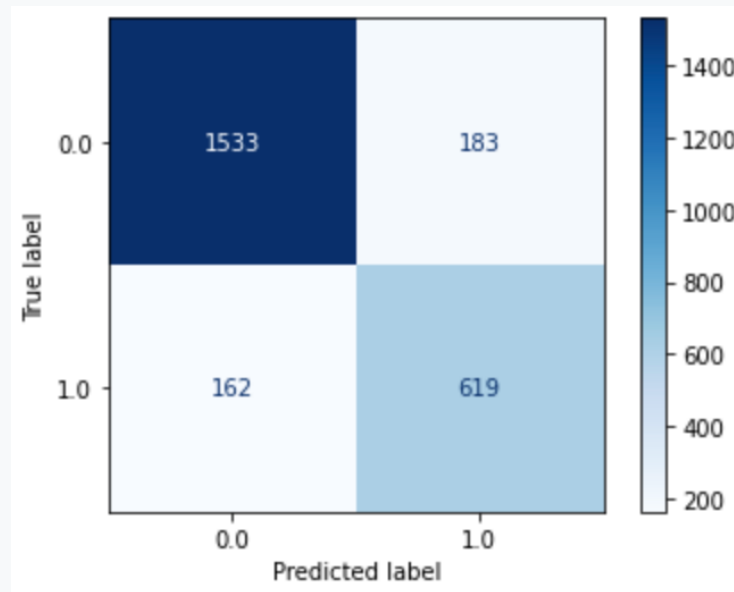
Recommendations

- The model with the best recall score is our tuned **LightGBM** with 79%, i.e. 79% of customers who file claims will have been predicted to correctly, with test accuracy of 86.2%.
- We also find relatively high precision (which minimizes false positives). This is interpreted as accurately predicting when customers do not file a claim. This reassures us that the insurance company can reliably pick customers who will continue to pay premiums, while not costing the insurance company more money with claims.
- Finally, the insurance company should scrutinize potential customers more who lease (or who otherwise don't own) their vehicles, as well as those who have a history of speeding violations and accidents.