**Final Project Report**

By Mike Stegmuller

**Problem Statement**

Insurance Fraud is a serious crime and should not be taken lightly. It is a problem for both the insurance companies and honest consumers. Identifying factors that lead to insurance fraud is important, so companies don’t have to pay our unnecessary compensation, but also keeps premiums down. Identifying what factors lead to fraud will not only help companies detect fraud better, but will also prevent people from committing fraud and going to jail.

**Data**

The dataset I chose had car accident data from the states of Illinois, Indiana, and Ohio from 2015. It shows various details such as incident severity, police reports, occupation, sex, and collision type about the auto accident and the insured customer details.

**Data Wrangling**

When examining the dataset, I noticed that one column was all Null values so I dropped the entire column from the data frame. I then checked the data types and unique values. The collision type, police report, and property damage columns had some cells filled with unknown information and question marks. I decided to replace the unknowns and question marks with the mode. I found out that the fraud percentage was 24.7%, which was close to a quarter of the data.

A graph of two people

Description automatically generated

**Modeling**

The machine learning models I chose were the decision tree, ada boost classifier and gradient boost classifier. After I ran the notebook unbalanced, I then balanced the notebook using SMOTE to see if there was a difference in results.

**Results**

When the data was unbalanced, the Gradient boosting was the most accurate model for predicting fraud. It was slightly better than the ada boosting classifier at 82%. In the decision tree model the major damage incident severity was the most important feature to predict fraud. In the ADA boosting model, the total claim amount and months as a customer were the most important features. Age was also a significant feature. Incident severity with major damage was also the most important feature in Gradient boosting.

**Results w/ SMOTE**

The results when the data was balanced with SMOTE were slightly higher for all the machine learning models. Gradient boosting at 84%, ada boost at 82% and the decision tree at 78%. The incident severity in the decision tree was still the most important feature, however trivial damage was second behind major damage. In the ada boost, non family relationship was the most important feature, followed by incident severity. I read this as when a non family member causes an accident with the insureds vehicle the insured is more likely to commit fraud. In Gradient boosting, major damage and total loss combined for 35% of feature importance for committing fraud.

**Analysis**

The Gradient boosting classifier produced the best results especially when balanced with SMOTE. Incident severity was the most important factor in both balanced and unbalanced datasets. Major damage and total loss at a combined 35% feature importance makes sense as the most important feature because of the amount of money involved in these types of claims. The hobbies seem random to me. Chess and CrossFit are the two hobbies that had the most fraud reported. Witnesses make sense because you can corroborate with a witness or there may not be a witness so an insured could potentially make up details about an accident. Policy deductibles also make sense as a key feature for fraud because if an insured is in a situation where their deductible is too high, they may try to lie to get out of paying it. A person being underinsured can lead to insurance fraud.

A screenshot of a data

Description automatically generated

Gradient Boosting results w/ SMOTE

**Recommendations**

1. I recommend that insurance companies use this information to see what factors lead customers to commit fraud in the first place. Use questionnaires, check to see there is police involvement in their accident, are they getting sued, what family issues are they having if any, does their salary play a role?
2. Use this information to show the insureds that companies know what factors lead to fraud. Hopefully insureds knowing this information will help dissuade people from committing fraud. This will keep people out of jail with potential felonies and lower premiums as companies won’t have to pay out on false claims or spend money on litigation.
3. Use this information to assess the accident details. If an insured has a claim with major damage or total loss, claim adjusters need to be on high alert that fraud may be the case since 35% of all fraud cases involved these types of losses.

**Ideas for Future work**

1. I would get more information from other states and see if the information is constant throughout the country. Seeing if the factors are consistent with the rest of the country is very beneficial to everyone involved.
2. I would also like to see where fraud is happening. I would like to create a US state map with counties, towns, and zip codes showing where fraud happens the most. Companies will be able to see that and use it to create rates and premiums.