Predicting Car Insurance Fraud

Problem Statement

Car insurance fraud can be difficult to detect due to the wide range of fraudulent activities that can occur and because they are far less common than legitimate claims. Fraud can be as simple as lying about a residential address or as extreme as faking an accident. Fraudulent claims are unethical and cause significant revenue loss for insurance companies each year. A predictive model for determining if claims are fraudulent will give insurance companies more confidence in pursuing a challenge to the claim and hopefully preventing more loss from these individuals in the future.

Data Wrangling

The data was obtained from kaggle via a scientific paper "Minority report in fraud detection: classification of skewed data", which provided all necessary data for this project. In this step the data was collected, organized, defined, and cleaned. The source code for this project is available for reference of these steps. The raw data contained 11,565 rows and 34 columns, and very little missing data. The dataset provides a column that represents fraud and reflects a 1 if it has been detected and a 0 if no fraud was found. There was no column representing what kind of fraud occurred. Therefore exploratory data analysis was conducted on available features to assess correlations.

Exploratory Data Analysis

Each numerical column was first examined through histograms. This resulted in further exploration into the 'Age' and 'ClaimSize' columns due to their distributions. Non-numerical columns were visualized through tables identifying fraud rate of each feature. Outliers were also identified. Of the 11,565 rows, only 685 of them represented claims that were found as fraudulent. 5.9% of the claims were fraudulent, while 94.1% were non-fraudulent. The largest fraud rate found with the non-numerical data was the policy type 'Utility - All Perils' at 0.131, so all correlations were quite insignificant.

Fig. 1 Age Distribution

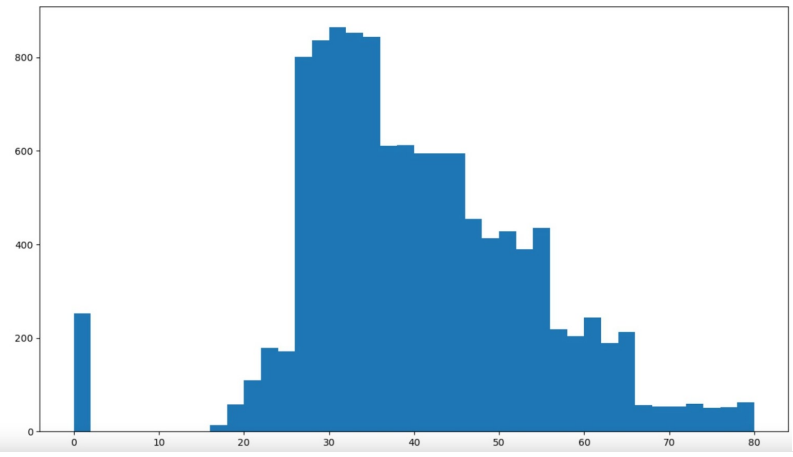


Fig. 2 Age Distribution for fraud and no fraud claims

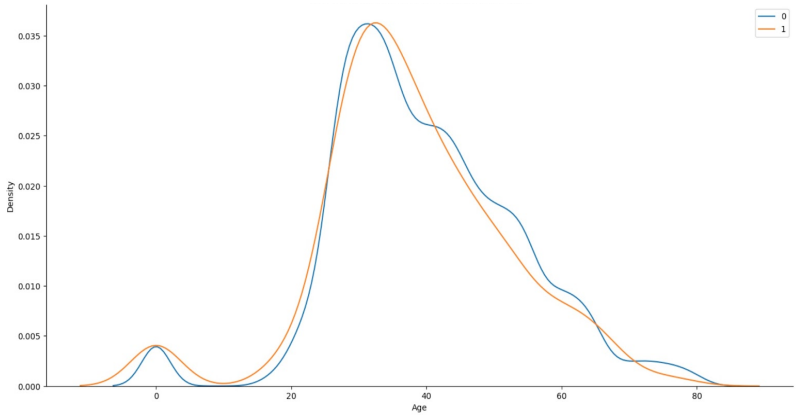


Fig. 3 ClaimSize Distribution

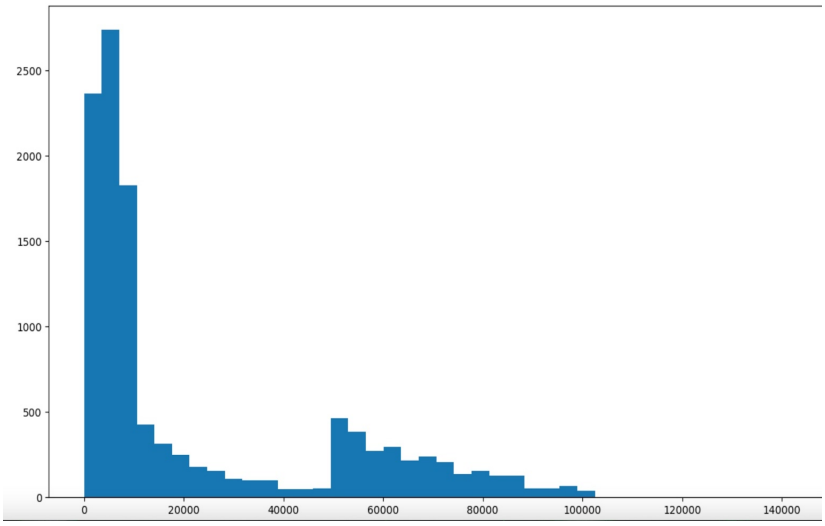


Fig. 4 ClaimSize Distribution for fraud and no fraud claims

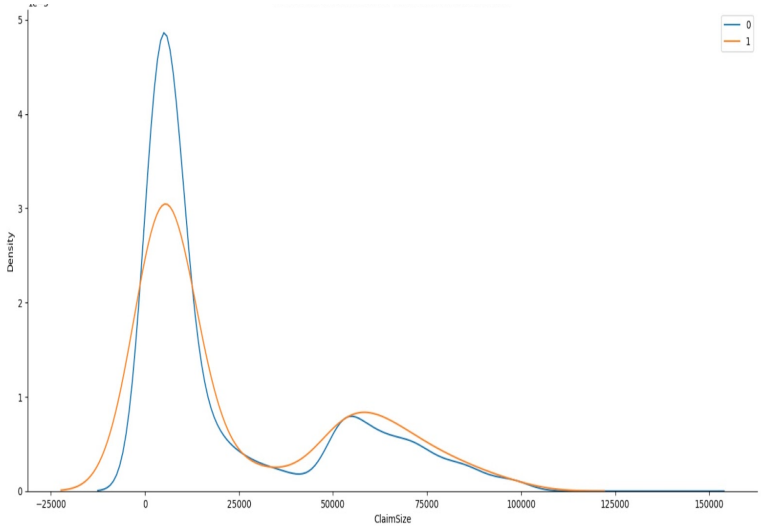
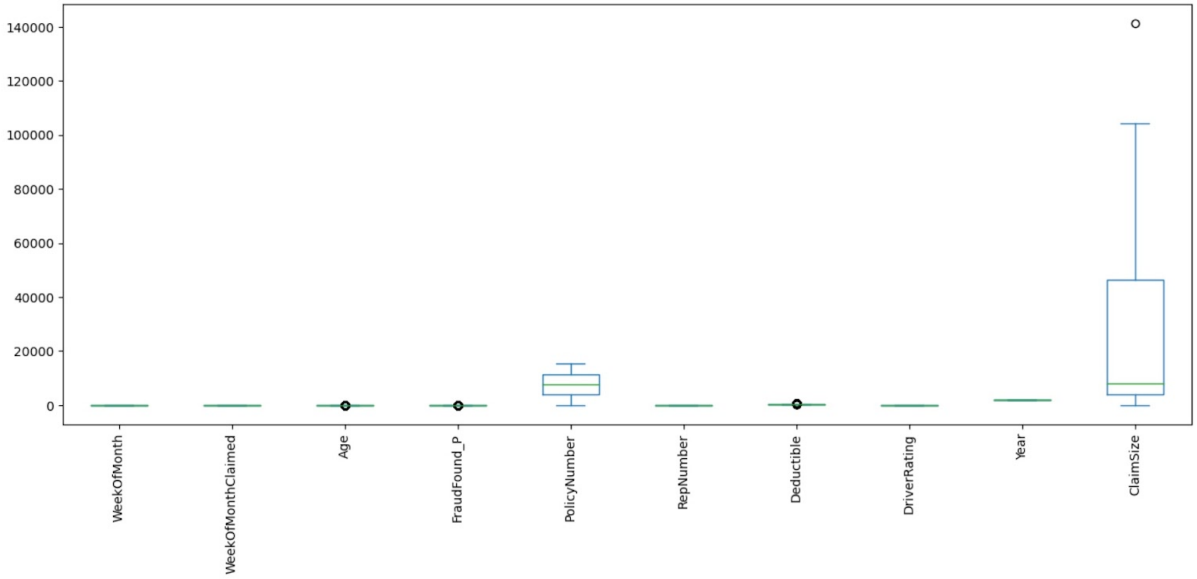


Fig. 5 Outlier Analysis



Modeling

The data was pre-processed for modeling and dummy features were created. Non-numerical features such as ‘MonthClaimed’, ‘DayOfWeekClaimed’, ‘VehiclePrice’, ‘AgeOfVehicle’, and ‘AgeOfPolicyHolder’ were converted to numerical values. The data was scaled before determining the appropriate model. Logistic regression, decision tree, and random forest were tested with the data. After running each model, random forest was determined to be the best fit for the data with an accuracy score of 0.94. Future work could include a larger data set, which would open the data to more modeling with potentially more accuracy.