Project Name	Property Insurance Claim Analytics
Objective	To develop a model to predict fraudulent claims based on historic data analysis of property insurance claims in 2017 and 2018 for all states in the US
Github Url	https://github.com/AnupamGH/insurance-claim-analytics.git
Author(s)	Anupam Shukla (M13469377) Ankit Kumar(M13436962) Mudit Verma(M13500805)
Date	October 5, 2019

Proposal:

Claims are a major expense for any insurance company, and it is important for these companies to identify the fraudulent ones. We have tried to answer this question using predictive analytics on property claims and customer datasets (source: Kaggle) covering the entire US region for the year 2017 and 2018(data till October):

a. **claims.csv (source: Kaggle)** – This dataset has details such as customer id, claim amount, claim date, claim incident cause, police_report, claim type, fraudulent(Y/N)
b. **cust_demographics.csv (source: Kaggle)** – This dataset has details like customer age, gender, date of birth, resident state and segment
c. **American_States.csv** – This dataset has mapping details from US states to different regions and subregions. We created this dataset to group claims by region and subregions in the US.

Analysis and Observations:

Our analysis is divided into 3 steps:

a. Exploratory data analysis
b. Deriving important statistics from the data
c. Developing a model to predict whether claims are fraudulent or not.

a. Exploratory Data Analysis:

After combining the three datasets, valuable insights from the data were uncovered –

- The difference in amounts claimed by female and male did not vary much except in East South Central Region where amounts claimed by males was almost 50% more than that by females.(To achieve this result, we created pivot table with subregion as index and gender as columns)
- The average amount claimed by the customers from the various segments(Gold, Platinum and Silver) are distributed uniformly.
- Most number of claims come from the South Atlantic region in the southern region of the United States
- The total number of claims made by males(51.2%) are a little higher than those by females(48.8%)
- Fraudulent claims in the adult category (age between 30 to 60) are almost 23% of total claims while for the youth category(age between 18 and 30) it is almost 21%

- Most claims are made in the months of April and July while very few claims are made in the months of August and September

b. Deriving important statistics from the data –

In this section, we have studied the various statistical measures across classification variables and their relationships with the claim amount.

Findings and Insights:

1. The average amount of fraudulent claims is higher for males as compared to females.
2. The average amount of fraudulent claims is higher for adults as compared to youth.
3. The overall mean claim amount is $12354 whereas the median is $2749 which indicates that the data is skewed and has very low and extremely high claim values.
4. While checking further, we found that either the claim 60% of claims have amount less than $4000 and 40% claims have amount more than $13000.
   We have classified them as low-value claims and high-value claims.
5. In low-value claims, the proportion of incident causes is similar with crime being the lowest but in high-value claims, the distribution is different with natural causes and crime being much lower than the others.
6. The mean and median claim amounts for males are higher overall but in low-value claims data, the same are lower than for females.
7. The hypothesis testing for mean claim amount for males and females suggests that they are not significantly different within a 95% confidence interval.

c. Developing a model to predict whether claims are fraudulent or not.

Findings and Insights:

1. Percentage of Fraudulent claims are 23% whereas percentage of non-fraudulent claims are 77%.
2. Incident cause does not seem to have an impact on Fraudulent behaviour. Fraud percentage is very similar for all incident causes.
3. Difference between the proportion of frauds between Auto and Home is very minimal.
4. Materials and Injury and Material Only types tend to have higher fraud percentage as compared to Injury only.
5. Mean claim amount for Fraudulent Yes is $12879.4, Fraudulent No is $12166.59 hence very minimal difference between the two.
6. Proportion of Frauds is very similar across both Male and Female.
7. Proportion of Frauds is very similar across the 3 segments – Gold, Platinum and Silver.

Modelling:

1. Created Dummy variables for categorical variables – gender, segment, incident cause, claim area, claim type, total policy claims.
2. Split the data into independent and dependent variables and prepared Unseen data. (on which the models were tested).

3. Created test and train sets and fit, scored and predicted using the following models –
    a. Logistic regression model score – 76.01%
    b. Gaussian NB model score – 76.38%
    c. Decision tree classifier score – 70.47%

Conclusion:

We can either go for logistic regression or Gaussian NB classifier as they have the maximum accuracy in predicting the fraudulent claims.