Final Project Report
Introduction to Data Analytics

Project Title:
Prediction/Analysis of insurance claim legitimacy

Prepared by:
Shubham Patel (N01624539)
Calist Dsouza (N01717873)

AIGC 5000 – Fall 2024
Humber College

1. Problem Statement

Prediction/Analysis of insurance claim legitimacy using demographic, policy, and claim data to identify fraudulent claims while minimizing false positives

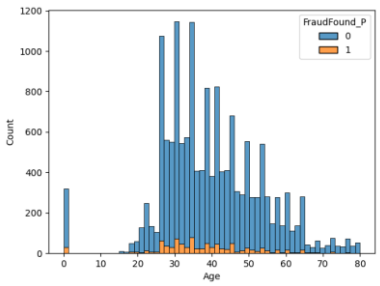
2. Dataset Description

The dataset contains information about insurance claims, including policy details, claim specifics, party information, vehicle characteristics, and repair data.

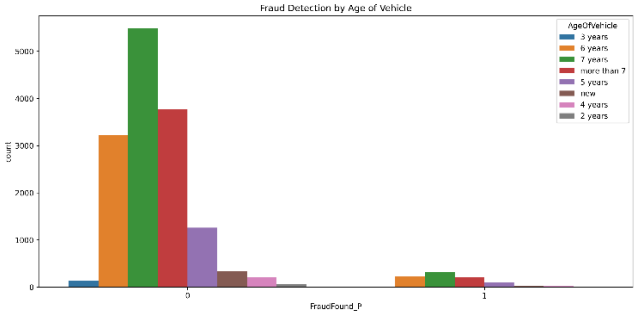
The dataset includes dependent variables for fraud detection, covering temporal data, claim specifics, policy details, etc. The target variable FraudFound_P is a binary indicator (0 = no fraud, 1 = fraud detected).

3. Dataset Analysis and Observations

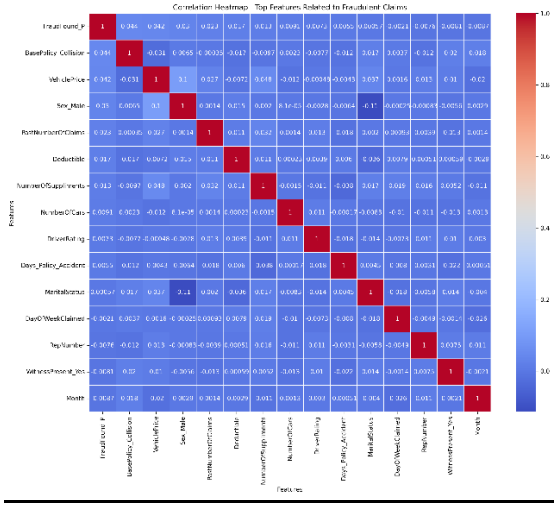
For our data analysis, we utilized histograms to explore univariate distributions, countplots to examine bivariate relationships, and correlation heatmaps to analyze multivariate interactions.



This plot will show how age is distributed among cases where fraud was found versus cases where it wasn't.



This plot aims to show the distribution of fraud cases across different vehicle age categories.



This is a correlation heatmap displaying relationships between various features in a dataset.

Most features show very low correlation (close to 0) with the target variable FraudFound_P, indicating weak linear relationships.

Some features are slightly correlated with each other. VehiclePrice and Deductible show some correlation (~0.1). Days_Policy_Accident and RepNumber have slight negative relationships.

4. Proposed Analytical/Prediction Model

We developed a predictive model to determine whether an insurance claim is fraudulent based on the dependent variables in the dataset. To accomplish this, we trained two machine learning models: Logistic Regression and Random Forest Classifier.

Before training, we preprocessed the data by handling missing values, encoding categorical variables, and scaling numerical features. We split the dataset into training and testing sets to evaluate the models' performance on unseen data.

5. Results and Discussions

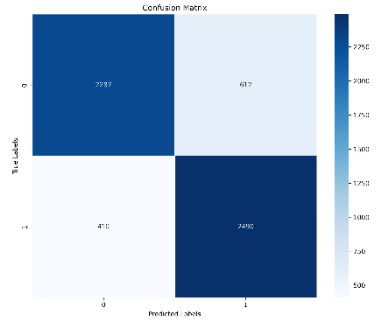
To evaluate and compare the performance of these models, we used several key metrics, including F1 Score, Recall, Precision, and Accuracy. These metrics provide a comprehensive view of the models' effectiveness in identifying fraudulent claims and ensuring robust predictions.

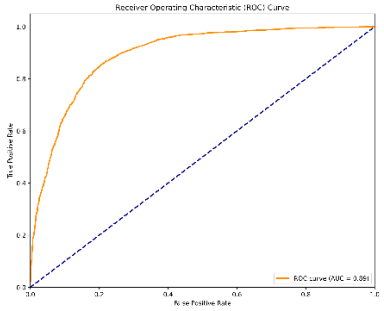
Logistic Regression Model

```
Accuracy: 0.82376271770995
Classification Report:
              precision    recall  f1-score   support

           0       0.85      0.79      0.82      2899
           1       0.80      0.86      0.83      2900

    accuracy                           0.82      5799
   macro avg       0.83      0.82      0.82      5799
weighted avg       0.83      0.82      0.82      5799
```



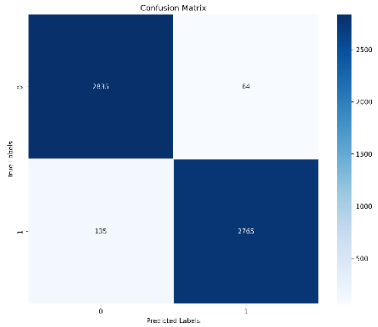


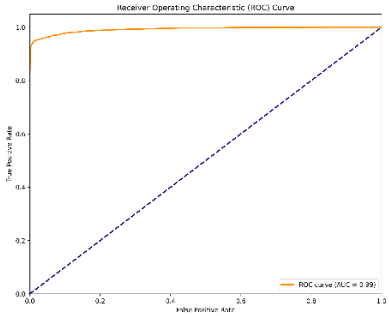
Random Forest Classifier

```
Accuracy: 0.9656837385756165
Classification Report:
              precision    recall  f1-score   support

           0       0.95      0.98      0.97      2899
           1       0.98      0.95      0.97      2900

    accuracy                           0.97      5799
   macro avg       0.97      0.97      0.97      5799
weighted avg       0.97      0.97      0.97      5799
```





Conclusion:

The Random Forest Classifier outperforms Logistic Regression in detecting fraudulent claims, with higher accuracy, precision, and recall. Its ability to capture complex patterns makes it the preferred model for reliable fraud detection.