Fraud Detection in Insurance Claims

Zahra Khalafi
UIN: 655096362
Zkhala3@uic.edu

Github link for this project:
https://github.com/zahrakhalafi/Project_CS_418.git

Problem statement:

Problem: Developing a machine learning model to detect fraudulent insurance claims. Insurance fraud can result in significant financial losses for insurance companies and policyholders, as well as higher insurance premiums for everyone. Detecting fraudulent claims can help reduce the financial impact of fraud and improve the accuracy and fairness of the insurance industry.

Question to Answer: Can we accurately predict which insurance claims are fraudulent using the available data on insurance claims? The machine learning model developed can be used to automatically flag potentially fraudulent claims for further investigation or denial.

Why did I chose this topic: It is a significant and persistent issue in the insurance industry. Insurance fraud can take many forms, from exaggerated claims to deliberate accidents, and can occur in any type of insurance policy. I can also use several machine learning techniques in predicting the fraud claims.

Hypothesis 1: Insurance claims on vehicle with higher values are more likely to be fraudulent than claims on vehicle with lower values.

- **Reasoning**: Fraudsters may be more motivated to commit fraud if the potential payout is high, as it represents a greater financial gain for them. Additionally, high claim amounts may be more difficult to verify and investigate thoroughly, making them more vulnerable to fraudulent activity.

Hypothesis 2: Insurance claims on vehicle with lower deductible are more likely to be fraudulent than claims on vehicle with higher deductible amount.

- **Reasoning**: Insurance policies with lower deductibles have higher premiums, which means that the policyholders may be more likely to file claims in order to recoup their costs. This could potentially create an incentive for individuals to file fraudulent claims, as they may see an opportunity to receive a payout that exceeds the amount they paid in premiums.

Data

The data that I will be using is the Vehicle Insurance Claim Fraud Detection.

Shape of the initial data is 15420 rows and 35 columns. After a lot of data cleaning it became 15420 rows by 15 columns.

I was able to access the data through this link:

https://github.com/mahmoudifard/fraud_detection

The dataset contains information related to insurance claims, with features including the month, week of the month, and day of the week the claim was made. It also includes information about the person making the claim, such as their sex, marital status, age, and past number of claims. There is information about the vehicle involved in the claim, including the make, category, price, and age. Other features include the type of policy, deductible, driver rating, and the number of days between the policy and the accident or claim. Additionally, there are features related to the claim itself, such as whether a police report was filed or a witness was present. The target variable, 'FraudFound_P', indicates whether or not fraud was found in the claim.

DATA Exploration

Step1: Initial Exploratory data analysis

- Running a loop for each column to find the type, number of unique values, number of missing values, and a sample of the data

```
Column Name: Make
Data Type: object
Number of Unique Values: 19
Number of Missing Values: 0
Sample Values: ['Mazda', 'Toyota', 'Toyota', 'Toyota', 'Honda']
```

Step2: Looking for null values

- Running another loop to find the null values in each column

```
Column 'Month' has 0 null values.
Column 'WeekOfMonth' has 0 null values.
Column 'DayOfWeek' has 0 null values.
Column 'Make' has 0 null values.
Column 'AccidentArea' has 0 null values.
Column 'DayOfWeekClaimed' has 0 null values.
```

Preprocessing the data

- **Finding the categorical columns**
  - Loop through the columns to find numerical vs categorical features
- **Mapping the vehicle price into ordinal numbers**
  - As one of the main features, a code specifically run to convert the vehicle price into ordinal numbers.

```
mappings = {'more than 69000': 6,
            '60000 to 69000': 5,
            '40000 to 59000': 4,
            '30000 to 39000': 3,
            '20000 to 29000': 2,
            'less than 20000': 1}
```

- **Removing the constant columns**
  - Since the constant values will not add any values in terms of modeling, they should be removed.
- **Encoding the categorical columns**
  - As the next step before starting the modeling we need to encode the categorical variables and then scale them. The StandardScaler method is used to scale the numerical columns in the dataset to have zero mean and 1 as variance.

Feature Selection

- **Correlation matrix**
  - As another part of preprocessing a correlation matrix was designed to remove the values that are highly correlated. We used a 80% threshold to remove those variables
  - Highly correlated features: {'VehicleCategory', 'AgeOfPolicyHolder', 'Year'}

- **Feature selection analysis using Recursive Feature Elimination with cross-validation**
  - To avoid overfitting a feature selection analysis was designed. The method that was used is called Recursive Feature Elimination with cross-validation.

```
[('Sex', 0.03),
 ('VehiclePrice', 0.03),
 ('Days_Policy_Accident', 0.02),
 ('Deductible', 0.01),
 ('DayOfWeek', -0.01),
 ('Make', -0.01),
 ('PoliceReportFiled', -0.01),
 ('AccidentArea', -0.03),
 ('PolicyNumber', -0.03),
 ('AgeOfVehicle', -0.04),
 ('Age', -0.05),
 ('AddressChange_Claim', -0.07),
 ('Fault', -0.32),
 ('BasePolicy', -0.36)]
```

Machine learning (logistic Regression)

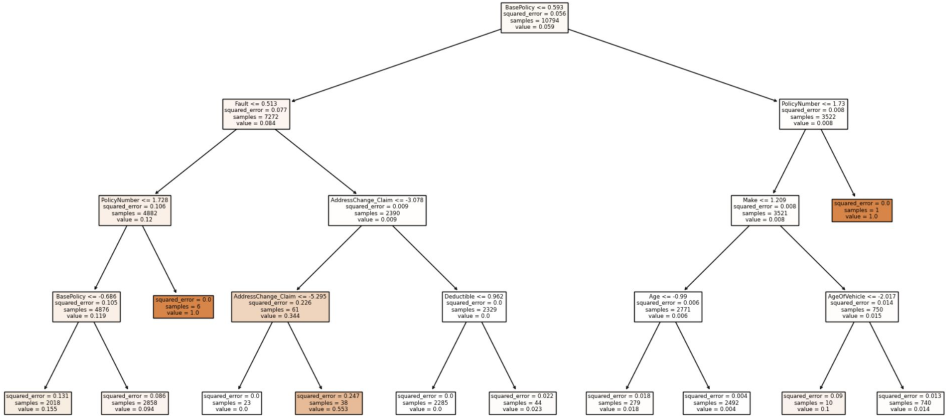
The evaluation of the logistic regression model involves the use of the confusion matrix and classification report. The former presents the true positives, false positives, true negatives, and false negatives in the model's predictions, showing that it correctly predicted most cases where the event did not occur but missed nearly all the cases where the event did occur. On the other hand, the classification report summarizes the precision, recall, and F1-score for each class, indicating that class 0 had high precision and recall, while class 1 had perfect precision but very low recall. The model has an accuracy of 0.94, but its performance is considered inadequate due to the low recall for class 1.

```
Logistic Regression:
Confusion Matrix:
 [[4341    0]
 [ 285    0]]
Classification Report:
              precision    recall  f1-score   support

           0       0.94      1.00      0.97      4341
           1       1.00      0.00      0.00       285

    accuracy                           0.94      4626
   macro avg       0.97      0.50      0.48      4626
weighted avg       0.94      0.94      0.91      4626
```

Machine learning (Decision Tree)

The decision tree model has a mean squared error (MSE) of 0.11 and an R2 score of -0.84. The negative R2 score indicates that the model does not fit the data well. The p-values for the features show that many of them are not statistically significant, with only a few having p-values less than 0.1. The F-statistic and p-value show that the overall model is not statistically significant.
Overall, these results suggest that the decision tree model is not a good fit for this dataset and that there may be other models that could perform better. Additionally, the significance of the features suggests that some of them may not be necessary for predicting fraud, and further analysis could be done to determine which features are most important for the task.

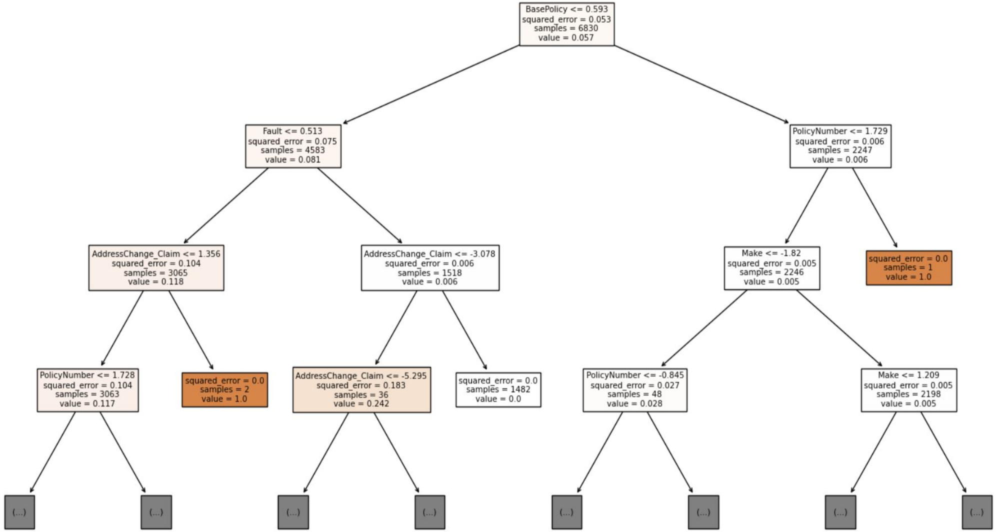
```
Decision Tree:
MSE: 0.11
R2 score: -0.84
P-values: const                        0.00
VehiclePrice               0.10
Sex                        0.03
Days_Policy_Accident       0.46
Deductible                 0.26
Make                       0.20
DayOfWeek                  0.18
PoliceReportFiled          0.33
PolicyNumber               0.07
AccidentArea               0.03
AgeOfVehicle               0.03
Age                        0.00
AddressChange_Claim        0.00
Fault                      0.00
BasePolicy                 0.00
dtype: float64
F-statistic: 0.00
P-value: 1.00
```



Machine learning (Random Forest)

The mean squared error (MSE) is 0.06, indicating that the average squared difference between the predicted and actual values is low. The R2 score is 0.05, indicating that the model explains only 5% of the variance in the data. The permutation importance scores show the relative importance of each feature in predicting the target variable. The p-values and F-statistic are shown for each feature in the model, but many features have p-values less than 0.1, indicating they are not statistically significant.

```
Random Forest:
MSE: 0.06
R2 score: 0.05
Permutation importance: [ 0.01  0.    0.    0.01  0.   -0.01  0.    0.01  0.01  0.01
  0.16  0.14]
P-values: const                 0.00
VehiclePrice            0.10
Sex                     0.03
Days_Policy_Accident    0.46
Deductible              0.26
Make                    0.20
DayOfWeek               0.18
PoliceReportFiled       0.33
PolicyNumber            0.07
AccidentArea            0.03
AgeOfVehicle            0.03
Age                     0.00
AddressChange_Claim     0.00
Fault                   0.00
BasePolicy              0.00
dtype: float64
F-statistic: 0.00
P-value: 1.00
```



Machine learning (Neural Network)

In this case, the neural network model has an MSE of 0.06 and an R2 score of 0.04, indicating that the model's performance may be suboptimal for detecting fraudulent claims. The R2 score suggests that the model explains only a small proportion of the variability in the data, while the MSE indicates that there may be considerable errors in the model's predictions. Further analysis and refinement of the model may be necessary to improve its accuracy and robustness.

```
Neural Network:
MSE: 0.06
R2 score: 0.04
```

Hypothesis testing

Hypothesis 1: Insurance claims on vehicle with higher values are more likely to be fraudulent than claims on vehicle with lower values.

It seems that Hypothesis 1 is supported by the data:

Proportion of fraudulent claims in vehicles with high price group: 0.02

Proportion of fraudulent claims in vehicles with low price group: -0.01

Difference in proportions: 0.03

Z-statistic: 24.15
P-value: 0.0000

Hypothesis testing

Hypothesis 2: Insurance claims on vehicle with lower deductible are more likely to be fraudulent than claims on vehicle with higher deductible amount.

Based on the statistical analysis, the variable Deductible seems to be not statistically significant, according to the statistical analysis of the variables in the decision trees and random forest tree. it seems that the pvalue for deductible is 0.26 which indicates it is less than 90% confidence intervals. In another word we can not reject nor accept the null hypothesis based on the data.

References:

https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8

https://scikit-learn.org/stable/modules/tree.html#classification

https://www.geeksforgeeks.org/ml-label-encoding-of-datasets-in-python/#

https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

https://towardsdatascience.com/an-implementation-and-explanation-of-the-random-forest-in-python-77bf308a9b76

https://www.tensorflow.org/tutorials/keras/classification