MIDTERM PROJECT REPORT

Data Analytics for Information Systems – IS665 – Fall 2022

Team Members:

Vishnu Santhana Gopal – vs92
Krushna Mahadev Akhade – ka544
Aravindakshan Viswanathan Sarma – av67

Fraud Detection Using Machine Learning Techniques:

We import all the libraries we require, including numpy, pandas, seaborn, matplotlib, gc, and sklearn, as well as the data sets (transaction.csv and identity.csv). The headers of the two csv input files, as well as the shapes, are checked. The two training sets were then integrated based on the TransactionID column, allowing us to work on a single combined training set. The frequency of the IsFraud binary column is then shown, and we see that only a small percentage of transactions are fraudulent.

Exploratory Data Analysis:

We started with the analysis of each of the columns and its relationship with the IsFraud column.

*1. **TransactionDT -** This column was compared with the IsFraud column, and a histogram was plotted to check the distribution where IsFraud is 0 and 1*

*2. **TransactionAmt** – The TransactionAmt column was compared with the Fraudulent data (IsFraud = 1) and non- Fraudulent data (IsFraud = 0). The average TransactionAmt was calculated for the Fraudulent & Non-Fraudulent data.*

3. To reduce the imbalance among the dataset, we introduced a new column called LogTransactionAmt. The LogTransactionAmt was plotted for both Fraudulent & Non-Fraudulent data sets. From the graph, we could infer that the LogTransactionAmt for legit transaction ranges from 3-5 and fraudulent transaction was higher than 3 and less than 7.

4. We used scatter plot to analyze the TransactionAmt and TransactionDT for the Fraudulent transactions.

5. To see which goods have the most fraud, we plotted the data and grouped the columns by IsFraud and ProductCD. We see that product C has the highest likelihood of fraud, followed by W,H,R, and S.

6. We analyzed the density of Fraudulent transactions on card1 which was 0.00016 and for card2 it was 0.010 and for card3 it was 0.4.

7. Upon analyzing the card4, which has the information about the type of cards. We could infer that Discover is prone to have the most fraudulent transactions than the legit transactions which is then followed by Visa and Mastercard.

8. The Visa and Master card account for a total of 60% and 30% of the fraudulent transactions.

9. The Card 6 tends to have higher chances of fraud at 225 and legit transactions happen mostly at 160.

10. Upon analyzing card6 which holds the information if a transaction had happened through credit or debit card. We can conclude that the credit card is prone to more fraudulent transactions than the legit transactions which is then followed by debit card.

11. The charge card and debit/credit card value has been combined to represent one value since it doesn't have any adverse effect on the datasets.

12. After analyzing the address columns, we can conclude that most of the transactions had occurred in the country with code 87. The country with code 87 amounts to 88.14% of the total data.

13. Based on the email columns, we may deduce that Gmail, Yahoo, and Hotmail should include most of the transaction data. Apparently, these three domains were prone to most of the fraudulent transactions.

14. The C and D columns contain correlations, which will be verified later in the modeling process.

15. We identified the NaN values in the V features which has been taken care later in the preprocessing steps, and we could identify some correlations among the M features.

16. After analyzing the Device Info and Device Type column, we found that the mobile device is prone to more fraudulent transactions than the legit transactions.

Feature Engineering and Selection:

Categorical values are determined, and two columns, id 23 and id 27, with largely NaN values are eliminated. Other discovered NaN values are replaced with the least disruptive value, -999. The categorical values were substituted with -999 in the train identity data sets. We next employ down sampling to reduce values from the dataset in order to balance it, because utilizing the complete dataset will result in overfitting and bias in our model. Both data frames are joined since they share the TransactionID column, and the final train dataset is ready for modeling.

Modelling:

We have selected a model and fit the model using the trained data sets, in order to predict the IsFraud values and test the accuracy of the model using AUC (Area under the roc curve). An ROC (Receiver Operating Characteristics Curve) shows the performance of a classification models at all threshold levels.

Below are the steps we followed to create and predict a model:
1. We imported the train test split from the sklearn model selection and assigned X with all the rows under all columns except IsFraud. Also, we assigned y with all the rows under IsFraud column.

2. X_train, X_test, y_train and y_test were assigned with values using the train test split function with X_train having 80% of the records and X_test having 20% of the records. Likewise, y_train has 80% of the records and y_test has 20% of the IsFraud values.

3. We used Decision Tree Classifier model to predict the accuracy of the data. We then fit the X_train and y_train before assigning the predicted results to the y_pred. The accuracy of this test was around 95.28%

4. The Random Forest Classifier model was picked because it is ideal for binary targets, such as the IsFraud column; it is ideal for large datasets; and it makes fair predictions. It also eliminates overfitting.

5. We fit the X_train and y_train, and ran the same model under the X_test data. Finally we assigned the result to the y_pred_dt.

6. The AUC score for this model is seems to be around 85.24%

7. The Decision Tree Classifier will be our final model which predicts the fraudulent transactions around 95% of the time.