Project 2

Machine Learning

Project 2

Machine Learning

Team Members

Miguel Sierra

Kashyap Suratia

Dustin Feinberg

Ryan weier

Purpose of the Project

Developing machine Learning application to efficiently and effectively detect fraudulent Credit card transactions .

Dataset

Dataset selected was very large with lots of variables. Including individual’s id, name,personal informations like location, with address,zip code,longitude and latitude, age etc.

It also included credit card numbers, transaction type, merchant type, day of the transaction, location of the transaction. etc

Dataset was highly imbalanced.

Analysis and Visualization

Extensive data analysis and Visualization was done to find correlation between different variables.

Most of the people in the data were in middle age gap(35 to 45).

* Fraudulent Transactions were very few compared to non-fraudulent transactions.
* More Female customers were transacted than Males customers.
* Merchant Longitude and Long variables were most correlated.
* Merchant Longitude and zip code were least correlated as from the correlation heatmap.

Methodology

The categorical data was then converted to numerical by Label encoding and a sample of the data taken to reduce time and memory consumption during training.

Since the target variable had an imbalance, random oversampling was done to create a balance between the binary variables

The independent variables were scaled in the range of 0 to 1 to reduce time taken during training.

Methodology

The data was then split into train and test data

The train data was fitted into machine learning models of scikit learn and tensor flow.

ML models were therefore tested on the test data.

Models Used in the project

* Logistic Regression
* Decision Tree Classifier
* Random Forest Classifier
* Naive Bayes
* TensorFlow
* Amazon Lex

Evaluation

The following methods of evaluation were implemented to evaluate the models' performance.

* Accuracy Score

* confusion Matrix

* ROC Curve

* Classification Report

Conclusion

The best performing model was Decision Tree and Random Forest Classifier

Logistic Regression performed satisfactory well but was the least performed when evaluated.