Machine Learning Engineer Nanodegree

Capstone Report

Sai Sumanth

Project Overview

- Credit Card transactions has been a larger share of US payment system. In today's increasingly electronic society and with the rapid advances of electronic commerce on the Internet, the use of credit cards for purchases has become convenient and necessary.
- Credit card transactions have become the de facto standard for Internet and Web based e-commerce. The US government estimates that credit cards accounted for approximately US $13 billion in Internet sales during 1998. This figure is expected to grow rapidly each year.
- However, the growing number of credit card transactions provides more opportunity for thieves to steal credit card numbers and subsequently commit fraud. When banks lose money because of credit card fraud, cardholders pay for all of that loss through higher interest rates, higher fees, and reduced benefits. Hence, it is in both the banks' and the cardholders' interest to reduce illegitimate use of credit cards by early fraud detection.
- Credit Card companies are approaching data scientists in order find a better solution to this problem. To solve this problem we need to build a model which flags the fraud transactions and gives an alert to the companies and the cardholders. This model can be designed using both supervised learning and unsupervised learning methods. For supervised learning methods, we need to have labelled data to train our algorithm. Whereas for unsupervised learning methods, we can tag the outlier transactions as fraud.
- A research collaboration of Worldline and the Machine Learning Group of ULB on big data mining and fraud detection have worked on the datasets provided by Kaggle in order to find a better model to tag the fraud credit card transactions.

Problem Statement

- The goal of this project is to find out whether a cardholder transaction is fraud or not.
- In supervised learning, the model is trained on data which consists of label feature, it gives an information about a transaction, 1 if the transaction is fraud, and 0 if the transaction is good.
- In unsupervised learning, the model is directly trained on the transaction data without the label feature. The transactions which lie farther away (i.e., outliers) from the normal transactions are considered as fraud transactions.
- The model performance is evaluated using recall score rather than accuracy score. Since, we care about tagging the fraud transactions, we need to have less number of false negatives.

Metrics

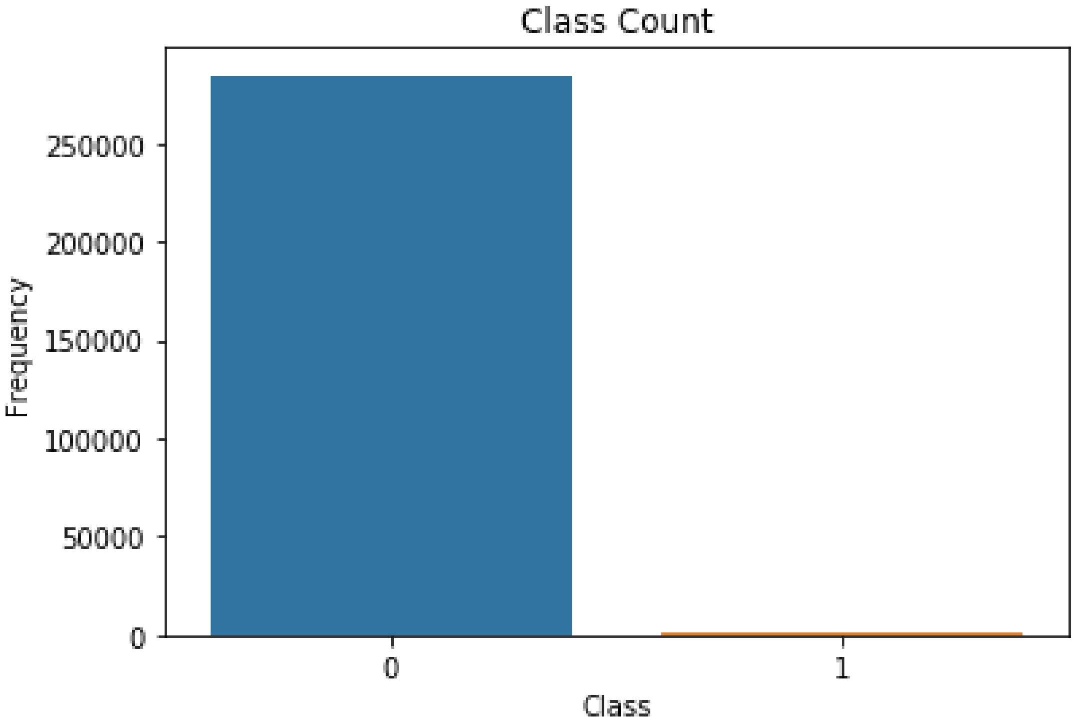
- For supervised model, I'm going to evaluate my model using Area under the Precision-Recall Curve and Recall Score to evaluate its performance.
- For unsupervised model, the outliers are tagged as 1 and the remaining instances are tagged as 0. The predicted results are then compared with the actual results to calculate Area under the Precision-Recall Curve and Recall Score to evaluate its performance.
- The precision-recall curve shows the trade-off between precision and recall for different threshold. A high area under the curve represents both high recall and high precision, where high precision relates to a low false positive rate, and high recall relates to a low false negative rate. For our problem statement, the model should have area under precision-recall curve nearly equal to 1.
- The reason to consider area under precision-recall as a performance metric for both the supervised and unsupervised models is our business problem is to capture the fraudulent transactions, which are false negatives in the prediction results. Therefore, the less the number of false negatives, the better the model captured the fraudulent transactions. The area under precision-recall is equal to 1, if the number of false negatives are zero.
- Recall Score is the ratio of True Positives and Sum of (True Positives and False Negatives). For our problem statement, the model should have a recall score nearly equal to 1. Similarly, the less the number of false negatives, the better the model captured the fraudulent transactions. If the recall score is equal to 1, the number of false negatives are zero.

Data Exploration

- For this project, I am going to use the datasets which contains transactions made by credit cards in September 2013 by European cardholders which are available on Kaggle.
- The dataset contains 492 frauds out of 284,807 transactions that occurred in 2 days. Due to confidentiality issues, the dataset contains only numerical input variables which are the result of a PCA transformation. It consists 28 Principal Components, along with Time and Amount of the Transaction.
- Since, we have only 492 frauds, the dataset is considered as class imbalanced. In order to overcome this, we have to perform undersampling or upsampling on the dataset to make it balanced.
- Feature Time contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature Amount is the transaction Amount, this feature can be used for dependant cost-sensitive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise. The remaining 28 Principal components gives us the transformed personal information about the cardholder.
- For the feature distribution, we are going to consider only Amount feature, because most of the other features are principal components and wouldn't give much information about the real features.
- Average Amount from all the transactions is 88.349619 and has a Standard Deviation of 250.120109
- The Amount value at 75 percentile is 77.165000, which gives us an information that the transactions with high amounts are less.
- More insights on Amount features is discussed in Exploratory Visualization.

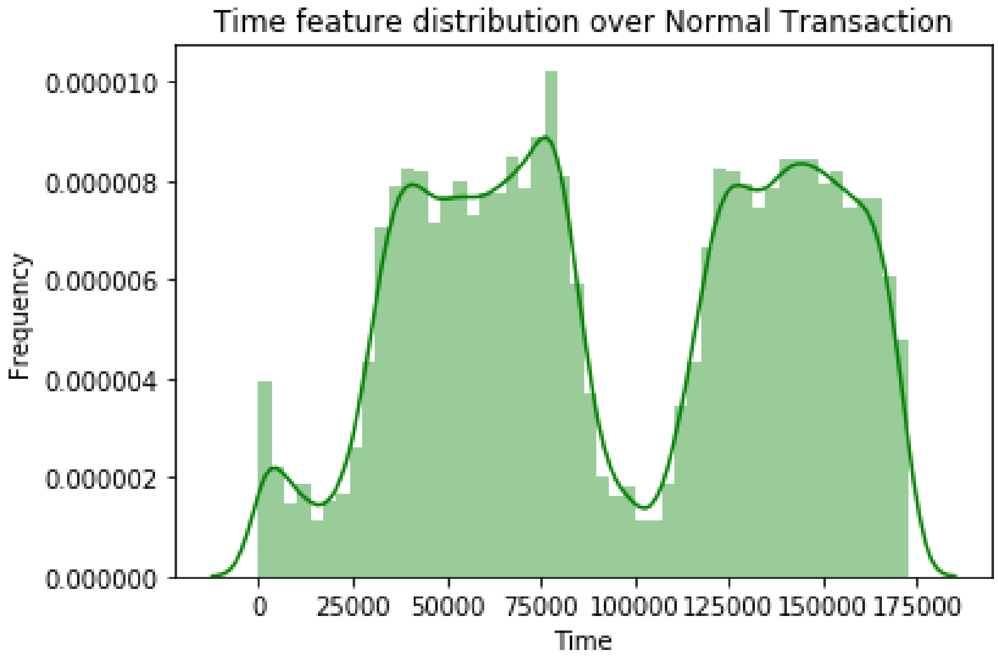
Exploratory Visualization

Class Count

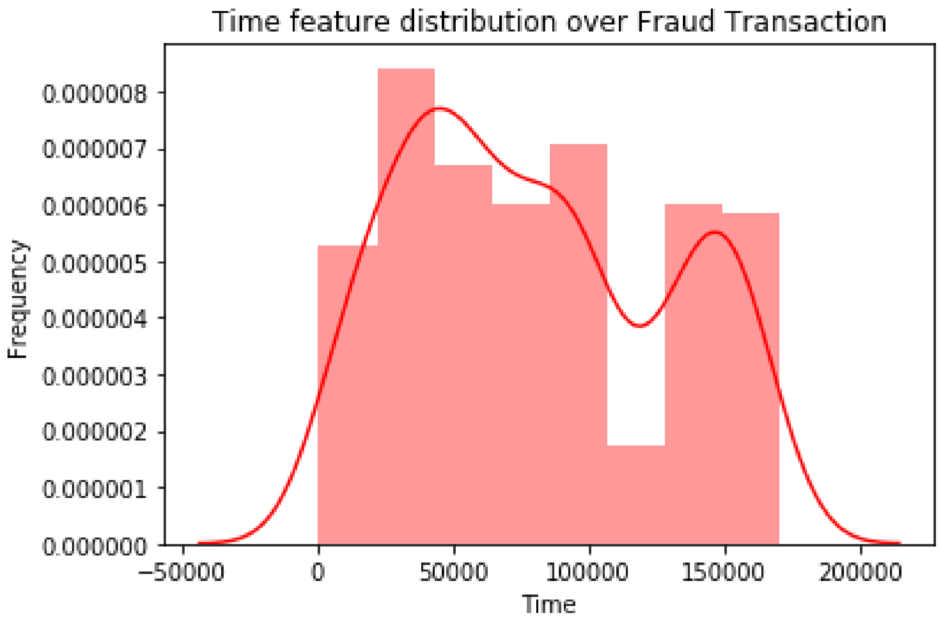


- We can see that it's a class imbalance problem, because the normal transactions are more than 2500000 but the fraud transactions are very less

Time Distribution

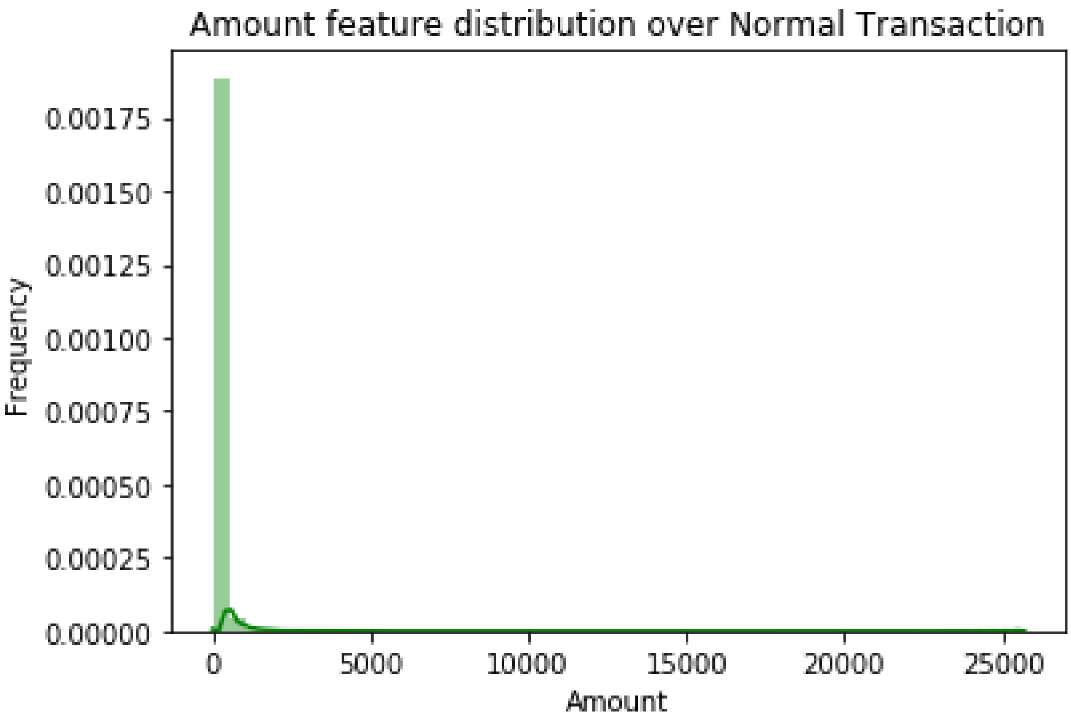


- We can see the distribution of Time over Normal Transactions, we can interpret that they are 2 peaks in the distribution and nothing unusual.

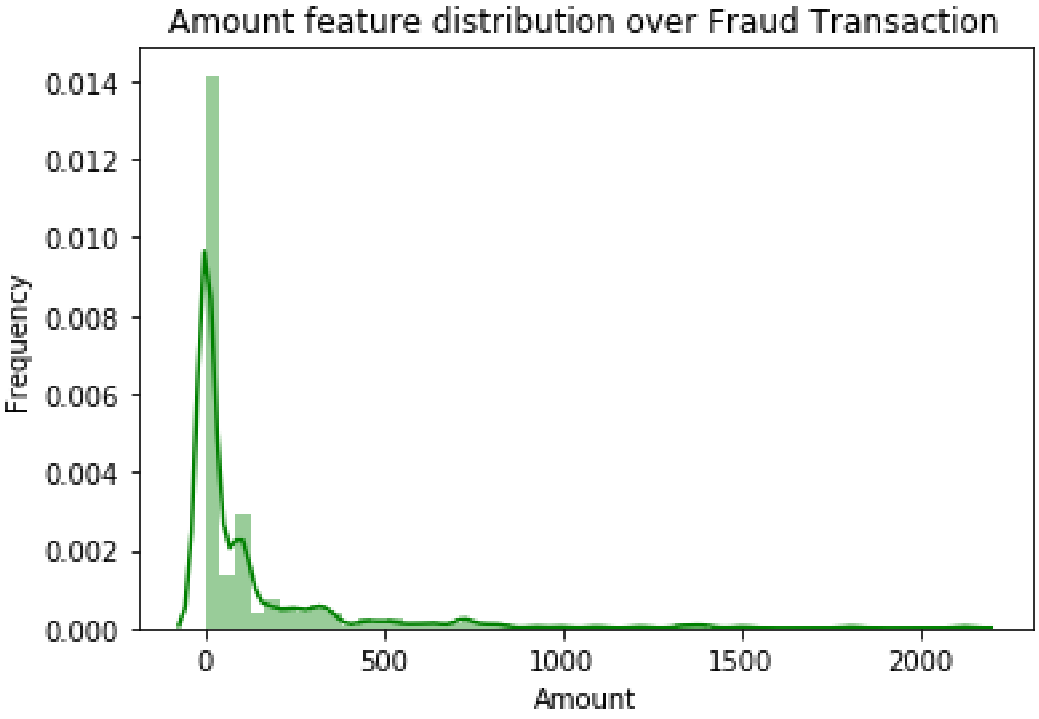


- We can see the distribution of Time over Fraud Transactions, we can interpret that it is a normal distribution and nothing unusual.

Amount Distribution

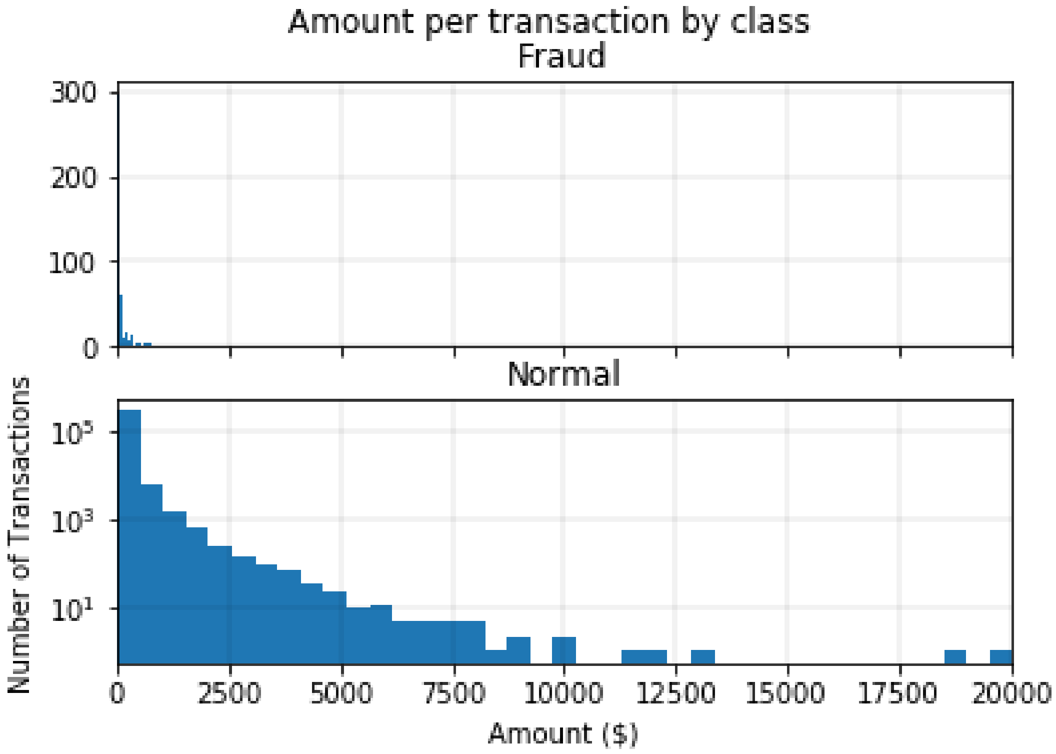


- We can see the distribution of Amount over Normal Transactions, we can interpret there is peak at the beginning but it becomes flat after the peak
- Therefore, the Amount features has a right skew distribution in Normal Transactions



- We can see the distribution of Amount over Fraud Transactions, we can interpret there is huge peak at the beginning but it becomes flat after 900.
- Therefore, the Amount features has a right skew distribution in Fraud Transactions

Number of Transactions over Amount



- We can see that Normal has more number of transactions over Amount when compared to Fraud Transactions.

Algorithms and Techniques

- Initially, I will visualize the distribution of the data and the skewed features are transformed using logarithmic transformation or square root transformation.
- The features are then standardized in order to achieve zero mean and equal variance.
- I'm going to build models using both supervised and unsupervised learning methods, and compare their performance results.
- For supervised learning, the dataset with Label feature is fitted to a supervised classifier and the Area under the Precision-Recall Curve and Recall score are calculated to evaluate its performance.
- For unsupervised learning, the dataset without the Label feature is fitted to unsupervised algorithm, and the outliers are tagged as fraud transactions, and they are compared with the actual Label results to evaluate the performance of the model.
- For supervised learning, I'm going to build Random forest classifier. It's an Ensemble method over Decision trees, which is trained using the bagging method. The bagging method is a combination of learning models, where the learning models are decision tress in case of random forest. Area under the Precision-Recall Curve and Recall Score are calculated, and the model is tuned accordingly to achieve high scores.
- For unsupervised learning, I'm going to use a k-means clustering algorithm. The instances which are higher than 95% percentile distance from the cluster centres are considered as fraud transactions. The comparison of tagged outlier instances with the real label feature gives us the performance of the model.
- Random Forest Classifier and K-means algorithm are well known for Anomaly Detection.

Benchmark

- I plan to compare my models with Basic Logistic Regression Classifier as my bench mark model, since it is considered as a go to method for binary classification problems.
- I want to fit Logistic Classifier to the labelled dataset and calculate the Area under the Precision-Recall Curve and Recall scores.
- I would like to choose these performance metrics as a threshold to evaluate my supervised and unsupervised models.

Data Pre-processing

- The amount values are higher than principal component features, the dataset is scaled using MinMaxScaler.
- Since the business problem is class imbalanced, I have used upsampling and down sampling methods to have a balanced class dataset.
- For the benchmark model, I have used SMOTE up sampling method.
- SMOTE up sampling method increases the fraudulent transactions in dataset, so that the normal transactions and fraud transactions are of equal quantity.
- For the Supervised model, I have used SMOTETomek, which is a combination of upsampling and down sampling methods.
- SMOTETomek method increases the fraudulent transactions and decreases the normal transactions until both of them are equal in quantity in dataset.
- For the Unsupervised model, I have used TomekLinks down sampling method.
- TomekLinks down sampling method decreases the normal transactions in dataset, so that the fraud transactions and normal transactions are of equal quantity.
- After sampling the data, I have divided the balanced datasets into training and testing sets by stratified shuffle splitting.

Implementation

- Initially, the dataset is class imbalanced because the normal transactions are more compared to the fraudulent transactions.
- Therefore, we have to perform sampling methods to make it a balanced dataset.
- I have used various up sampling and down sampling methods like SMOTE, SMOTETomek and Tomeklinks
- I have performed visualization on the features to derive insights from the data.
- But the visualization did not give away much information about the data.
- Scaling of data is key factor to be considered because the amount feature is higher than the other features
- Therefore, in order to have good data for model building, the data needs to be scaled.
- The dataset must be split into training and testing datasets.
- I have used stratified shuffle splitting which splits the datasets randomly.
- The training dataset is used for model building and tuning.
- The testing dataset is used to evaluate the model performance.
- Initially, I built the Logistic Regression on SMOTE up sampled training dataset as a benchmark model to evaluate Random Forest Classifier and K-means clustering models.
- I calculated the performance metrics for the Logistic Regression Model for future comparison.
- Then I built Random Forest Classifier on SMOTETomek sampled training dataset.
- I calculated the performance metrics for the Random Forest Model to compare the results with Logistic Regression.
- Then I built K-Means Classifier on Tomeklinks sampled dataset.
- I have calculated the distance between the instances and the cluster centroids
- Then I calculated the distance mean by taking the average of the distances

- I tagged the instances greater than 95 percentile distance mean as fraudulent transactions.
- I calculated the performance metrics by comparing the predicted fraudulent transactions with the real fraudulent transactions to compare the results with Logistic Regression
- Normally, accuracy of the model is calculated to evaluate its performance.
- Since, it’s a credit card fraud detection business problem, we would have to consider Recall and AUC score, which gives us a higher score when there are less number of false negatives.
- Because if we have less number of false negatives, then we have captured almost all the fraud transactions.
- I compared the performance metrics from all the three models
- I visualized that Random Forest has high accuracy , recall score and auc score over Logistic Regression and K-Means Clustering
- Logistic Regression has high accuracy, recall score and auc score over K-Means Clustering.
- Finally, I deduced that Ensemble Supervised Learning Methods like Random Forest Classifier gives us a better chance to capture almost all the fraudulent transactions.
- While model building I have faced difficulty while evaluating the Unsupervised Model, but I came up with a solution by using the cluster centres and centroids to interpret the fraudulent transactions.

Refinement

- The Random Forest Classifier, did not require any refinement because I got an AUC score of 99.9%, which means my model captured all the fraudulent transactions.
- The K-Means Classifier, hasn’t performed very well in capturing the fraudulent transactions because it is mainly used of segmentation
- Therefore, for Credit Card Fraud Detection, I recommend people to use Ensemble Supervised Models like Random Forest Classifier over Unsupervised Models to achieve a better solution

Model Evaluation and Validation

1) Logistic Regression, Benchmark Model prediction has given the following performance metrics

- Accuracy: 94.4836%
- Recall: 91.5164%
- AUC: 94.4929%
- Since, it is class imbalanced problem, I am going to evaluate model performance using AUC Score.
- For the benchmark model, the AUC score is 94.5%
- The Logistic Regression is a go to model for a binary classification dataset, it clearly did a better by giving a performance over 94%

2) Random Forest Classifier, Supervised Model prediction has given the following performance metrics

- Accuracy: 99.9859%
- Recall: 100.0000%
- AUC: 99.9859%
- For the supervised model, the AUC score is 99.9%
- Since the Random Forests are built using ensemble methods, where it learns from the all the decision trees generated which finally gave a very promising score of more than 99.9%

3) K-Means Clustering, Unsupervised Model prediction has given the following performance metrics

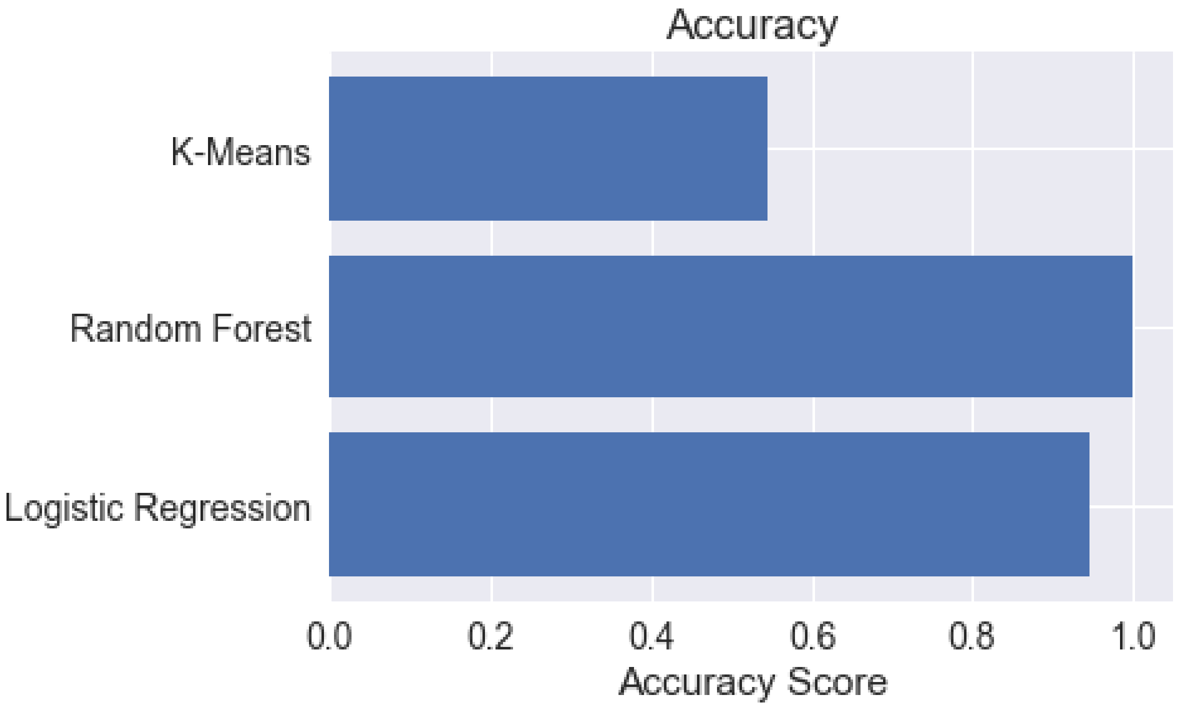
- Accuracy: 54.3814%
- Recall: 9.3815%
- AUC: 54.3814%
- For the unsupervised model, the AUC score is 54.3%
- Since the K-means Clustering are mainly used for segmentation, it didn't perform a better job in classifying the fraudulent transactions, and gave a score of less than 60%

Justification

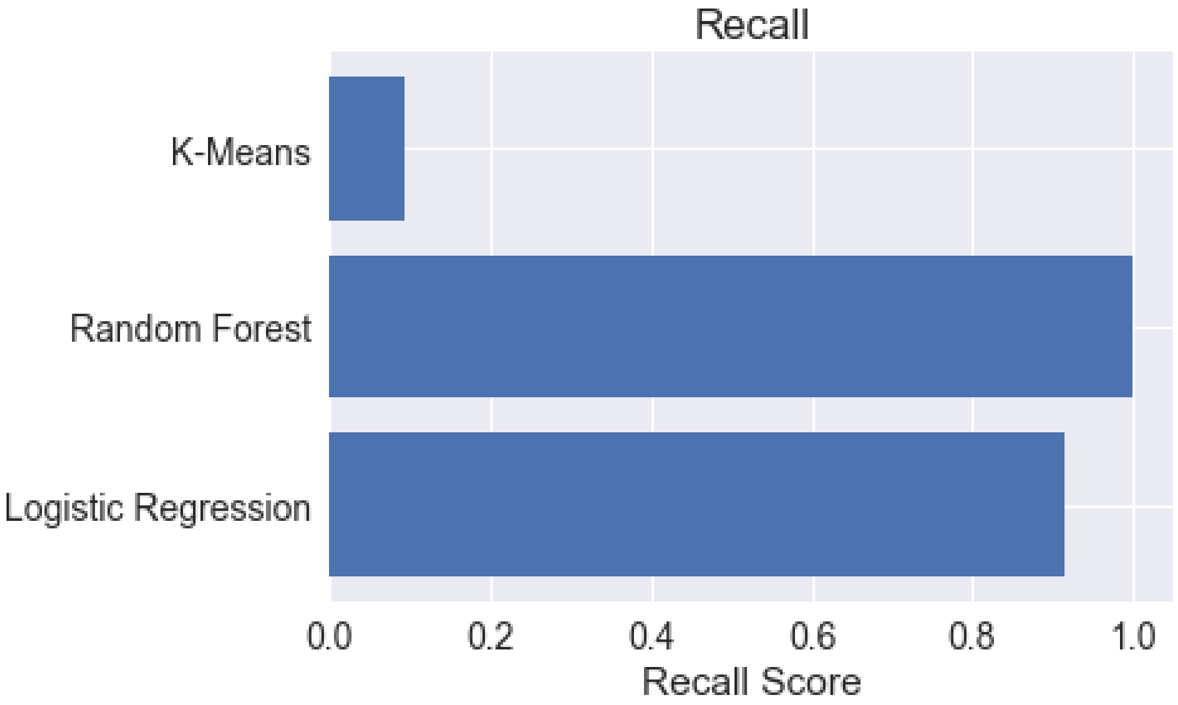
- Supervised Learning Methods clearly dominates Unsupervised Learning Methods for a class imbalance dataset.
- Random Forest Classifier has higher accuracy than Logistic Regression followed by K-Means Clustering.
- Similarly, Random Forest Classifier has higher recall score than Logistic Regression followed by K-Means Clustering.
- Similarly, Random Forest Classifier has higher auc score Logistic Regression followed by K-Means Clustering.
- My supervised model did a better job than my benchmark model but not the unsupervised model.
- Therefore, I would like to conclude that Supervised Learning Methods such as Random Forest Classifier gives a better prediction over Unsupervised Methods for Credit Fraud Detection.

Free-Form Visualization

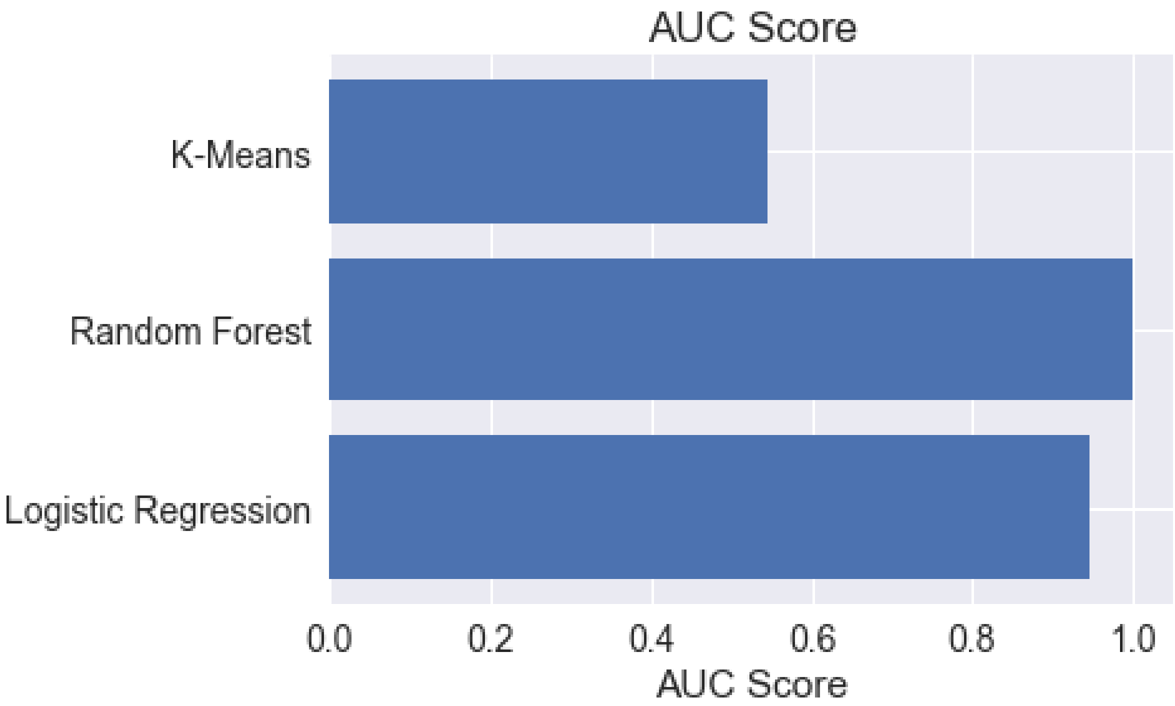
Visualization of Performance Metrics for K-means Clustering, Random Forest and Logistic Regression



- We can see that Random Forest out performs both K-Means and Logistic Regression by having an accuracy nearly equal to 1



- We can see that Random Forest out performs both K-Means and Logistic Regression by having an recall nearly equal to 1



- We can see that Random Forest out performs both K-Means and Logistic Regression by having an auc score nearly equal to 1

Reflection

- Initially, the dataset is class imbalanced because the normal transactions are more compared to the fraudulent transactions.
- Therefore, we have to perform sampling methods to make it a balanced dataset.
- I have used various up sampling and down sampling methods like SMOTE, SMOTETomek and Tomeklinks
- I have performed visualization on the features to derive insights from the data.
- But the visualization did not give away much information about the data.
- Scaling of data is key factor to be considered because the amount feature is higher than the other features
- Therefore, in order to have good data for model building, the data needs to be scaled.
- The dataset must be split into training and testing datasets.
- I have used stratified shuffle splitting which splits the datasets randomly.
- The training dataset is used for model building and tuning.
- The testing dataset is used to evaluate the model performance.
- Initially, I built the Logistic Regression on SMOTE up sampled training dataset as a benchmark model to evaluate Random Forest Classifier and K-means clustering models.
- I calculated the performance metrics for the Logistic Regression Model for future comparison.
- Then I built Random Forest Classifier on SMOTETomek sampled training dataset.

- I calculated the performance metrics for the Random Forest Model to compare the results with Logistic Regression.
- Then I built K-Means Classifier on Tomeklinks sampled dataset.
- I have calculated the distance between the instances and the cluster centroids
- Then I calculated the distance mean by taking the average of the distances
- I tagged the instances greater than 95 percentile distance mean as fraudulent transactions.
- I calculated the performance metrics by comparing the predicted fraudulent transactions with the real fraudulent transactions to compare the results with Logistic Regression
- Normally, accuracy of the model is calculated to evaluate its performance.
- Since, it's a credit card fraud detection business problem, we would have to consider Recall and AUC score, which gives us a higher score when there are less number of false negatives.
- Because if we have less number of false negatives, then we have captured almost all the fraud transactions.
- I compared the performance metrics from all the three models
- I visualized that Random Forest has high accuracy , recall score and auc score over Logistic Regression and K-Means Clustering
- Logistic Regression has high accuracy, recall score and auc score over K-Means Clustering.
- Finally, I deduce that Ensemble Supervised Learning Methods like Random Forest Classifier gives us a better chance to capture almost all the fraudulent transactions.
- The interesting aspect while building the models was that ensemble methods outperforms any other algorithms and gives us a better solution.
- The biggest challenge I faced was evaluating the Unsupervised Model, but I came up with a solution by calculating Cluster Centroid and Instances distance, and tagging the Instances which are greater 95 percentile distance as fraudulent transactions.
- I have learned that we have to adapt Ensemble Supervised Learning Methods for Anomaly Detection.

Improvement

- Ensemble Methods needs higher computing power to give the results in fraction of seconds.
- I would like to make improvements to Unsupervised Learning Methods like K-Means Clustering, in order to predict fraudulent transactions.
- Since my data consisted of label feature, I was able to evaluate my model in a sensible way
- But if the Label feature is missing , then we would have to adapt Unsupervised Methods only
- Therefore, we would have to come up with a better solution to increase the performance of the Unsupervised Models to capture a fraudulent transaction.

References

1) Kaggle:

https://www.kaggle.com/mlg-ulb/creditcardfraud

2) Research Paper on Credit Card Fraud Detection:

https://pdfs.semanticscholar.org/7456/adc49f8b3d2bfba97064284aa81364ad7dbc.pdf

3) Random Forest:

http://ai.ms.mff.cuni.cz/~sui/nezvalovapopelinsky.pdf

4) K-Means:

http://pmg.it.usyd.edu.au/outliers.pdf