ISYS 620 COURSE PROJECT
TOPIC:
CREDIT CARD FRUAD DETECTION

AUTHORS
CHIKE JOSEPH OKWUDIAFOR
ERIC OYEMAM ATOBROWN
OBEHI IREKPONOR

SPRING 2020

Abstract

Credit card which is the mode through which many financial institutions uses to transact business between clients has become particularly important for easy access to funds anywhere at any time. But in recent years many financial institutions have encountered many issues of credit card fraud which has a higher effect on both the client of the bank and the institution. On this basis, this project is geared towards detecting credit card fraud and how we can use various machine Learning algorithms to detect the fraudulent transactions.

INTRODUCTION

Credit card fraud happens when a non- owner of a card uses a third parties' card to transact business without the knowledge of the person. In many cases, a fraudster can obtain your financial information and use it to make fraudulent purchases.

In view of some of the ways through which credit card fraud occurs, this project is aimed at using the available machine learning algorithms and technics to aid financial institutions and clients detect fraudulent credit card transactions. We hope that, we will be able to establish insights to help both parties to identify fraudulent transactions thereby saving them from loss of money.

The machine learning algorithms we used are Logistic Regression, Artificial Neural Network, Random Forest and K-Means Clustering in predicting genuine and fraudulent credit card transactions. In all, we seek to understand how the various chosen algorithms will be meaningful in establishing weather there is a higher magnitude of credit card fraud or no fraud. We will use the selected Machine learning Algorithms to analyze and predict large data looking for patterns and anomaly.

DESCRIBING THE DATA SET

A dataset of 284,807 sample size of credit card transaction in September 2013 by European cardholders was used for the analysis. This transaction was carried out in two days and it contains 492 fraud and its variable include, numerical input variables which are the result of a PCA transformation. Unfortunately, due to confidentiality issues, we cannot provide the original features and more background information about the data. The fraudulent transaction makes up only 0.1727% of all the transactions making the dataset highly unbalanced. This dataset and information was gotten from: (https://www.kaggle.com/mlg-ulb/creditcardfraud)

Features

V1, V2, … V28: are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'.
Time: contains the seconds elapsed between each transaction and the first transaction in the dataset.
Amount: is the transaction amount
Class: is the response variable and it takes value 1 in case of fraud and 0 otherwise.

Principal Component Analysis PCA transformation has previously been carried out on Features V1, V2, … V28 in other to manage confidentiality. Through this, we cannot provide the original features and more background information about the data.

Standard Scaling/Normalization: Variables that are measured at different scales do not contribute equally to the analysis and might end up creating a bias, a normalization was carried out in other to fit all the features in a particular range as it helps in speeding up the calculations in the algorithms.

Data Cleaning: - Fortunately, the dataset had no null values making it unnecessary for data manipulation.

Correlation: Correlation was carried out to check close similarities within the features and to re-confirm the PCA transformation.

Training and Testing: For the supervised learning algorithms carried out, the data was spilt into test and training data where the testing data was 30% of the dataset. we decided to split the testing data to 30% in other to involve or randomly capture as many fraudulent transactions as possible since it represents 0.1727% of the entire dataset.

DESCRIBING THE LEARNING ALGORITHMS

The analysis being carried out in this project is towards the prediction of binary outcomes (i.e. Yes or No denoted as 0 or 1). This binary outcome is a dummy variable to indicate the existence of an event which in this case is fraudulent credit card transactions. To this end, we decided to carry out the below listed classification algorithms; three supervised and one unsupervised algorithm. The response variable (Class) takes value 1 where the transaction is fraudulent and 0 otherwise.

Logistic Regression:

This is a supervised learning algorithm and a type of regression analysis used in predicting the outcome of binary variables as its logistic functions bounds the limits of the outcome variable to 0 and 1 thus overcoming the limitations of OLS.

Artificial Neural Network:

This is a machine learning algorithm which its inspiration is from the way the biological nervous systems such as the way the brain processes information. Artificial neural networks were built in this simple element called neurons, which take in a real value, multiply it by a weight, and run it through a non-linear activation function. This algorithm takes Xi inputs and attaches corresponding weights Wi which determines how much of the input is passed into the neuron to adjust the neuron's ability to learn which is adjusted during training time. The inputs are multiplied by its corresponding weights and assigned to a hidden layer before being estimated to the output layer using an Activation Function.

The output value can be continuous, binary or categorical. The output of one perceptron model can be the input of the next perception and so on. When a neuron network contains 2 or more hidden layers, it is then referred to as Deep Neural Network. The main goal is to minimize the cost function through back propagation in order to get fewer errors which makes our model more reliable.

Random Forest:

Random forest is a supervised learning algorithm that considers multiple decision trees. A decision tree is a classification/regression model that works in a tree-like structure. The tree has nodes, and each node shows a feature from the input, each branch is a decision, and each leaf is a corresponding output value. Decisions trees are however quite sensitive to the data and small changes to the training set can result in significantly different tree structures. Even though each of those trees might not be ideal, overall, on average they can perform very well. Random forest takes advantage of the sensitivity of data using the bagging process. In this model, it provided the strengths of the decision tree algorithm, and is highly effective at preventing overfitting and thus much more accurate.

K-Means Clustering:

This is an unsupervised learning algorithm for clustering. Clustering refers to the collection of data points aggregated together because of certain similarities. The objective of k-means is to group similar data points and discover certain underlying patterns. K-means looks for a fixed number(k) of clusters in the dataset. This is a stochastic algorithm so it would not give the same answer twice. In order to avoid the clusters getting stuck in a local minimum, k-means++ is initialized.

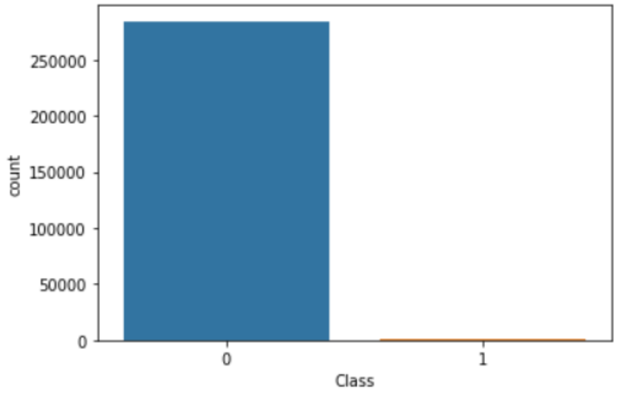
DISCUSSIONS AND CONCLUSION

Class:

0 – No Fraud
1 – Fraud

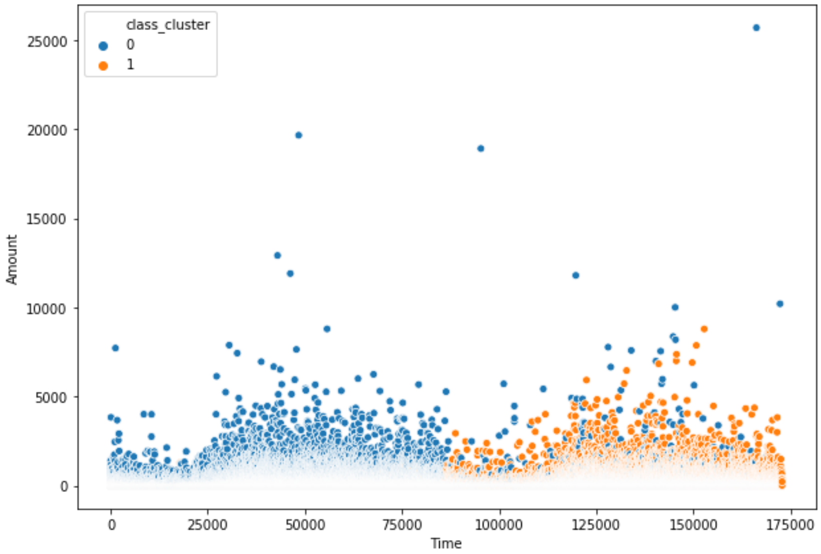
Count Plot of the Class Feature

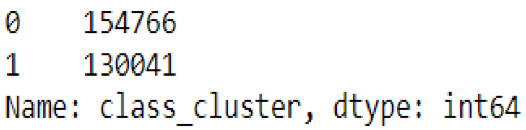
The count plot below shows the distribution of the Class variable where we have
0 - 284,315
1 - 492
This shows an unbalanced dataset and as a result Precision Recall Curves would be plotted to capture a better picture of the accuracy of the models.



K-Means Clustering

The K-Means Clustering which is an unsupervised learning algorithm in using the Time and Amount features of the dataset, clustered the credit card transactions as follows.





```
0    154766
1    130041
Name: class_cluster, dtype: int64
```

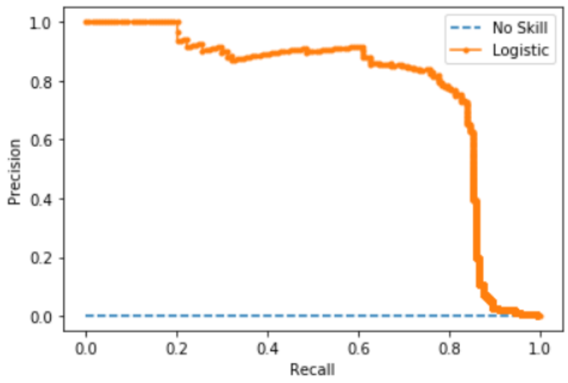
This shows that of the 284,807 transactions in the dataset, these are clusters that show similar patterns in the transactions with relation to the time and Amount variables. These underlying patterns give a deeper insight to what might be going on with the transactions. In retrospect, through deeper analysis, it is possible that most of cluster 1 transactions are found between 0 and 75,000 seconds, and the cluster 2 is found between 100,000 to 175,000 seconds. Through this, it can be deduced that the clusters differ mostly through the Time feature. This gives an extra variable to consider so in order, to increasing the efficiency of the other algorithms, we decided to add these class.

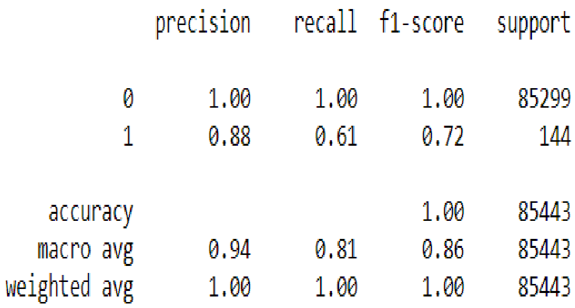
The scatter plot above shows the clusters.

Logistic Regression:

The Precision Recall Curve and Classification Report of the Logistic Regression depicts the following predictions.

Precision Recall Curve and Classification Report

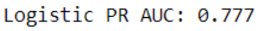




```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     85299
           1       0.88      0.61      0.72       144

    accuracy                           1.00     85443
   macro avg       0.94      0.81      0.86     85443
weighted avg       1.00      1.00      1.00     85443
```



```
Logistic PR AUC: 0.777
```

The performance of the Logistics Regression is good as the Precision Recall Curve shows that the Logistic Regression model achieves a PR AUC of about 0.777 in predicting the fraudulent transactions. Though the area under the curve is not perfectly positioned on the upper right corner however this still shows that the Logistic Regression Classifier was able to capture about 0.777.

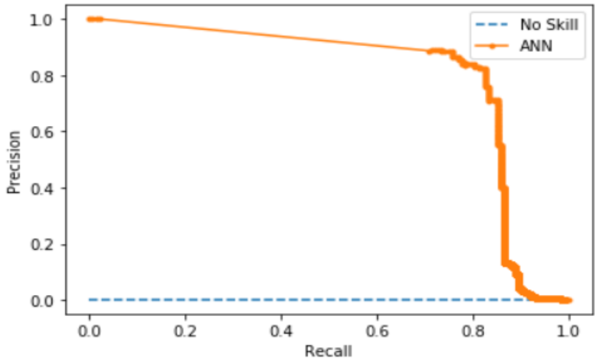
The Classification Report also shows that it predicts the no fraud transactions perfectly with a f1-score of 1 while it predicted the fraudulent transaction with f1-score of 0.72. This model depicts an ideal system with high precision

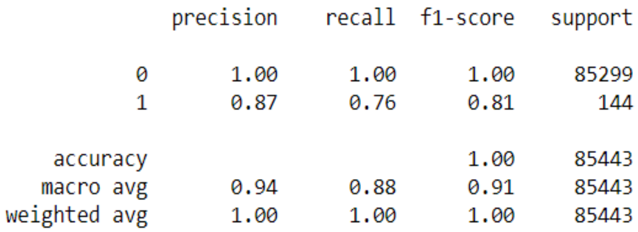
With this report, the model is cable of identifying when a credit card transaction is not fraudulent with a manageable accuracy thereby allowing the systems of the financial institution to authorize the genuine transactions where necessary in an ideal situation.

Artificial Neural Network:

The Precision Recall Curve and Classification Report of Artificial Neural Network depicts the following predictions

Precision Recall Curve and Classification Report





```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     85299
           1       0.87      0.76      0.81       144

    accuracy                           1.00     85443
   macro avg       0.94      0.88      0.91     85443
weighted avg       1.00      1.00      1.00     85443
```

```
ANN: f1=0.810 auc=0.802
```

The performance of the Artificial Neural Network (ANN) is particularly good as the Precision Recall Curve shows it achieves a PR AUC of about 0.8 in predicting the fraudulent transactions. The area under the curve is not perfectly positioned on the upper right corner however this still shows that the ANN model was able to capture about 0.802 with an f1-score of 0.81.

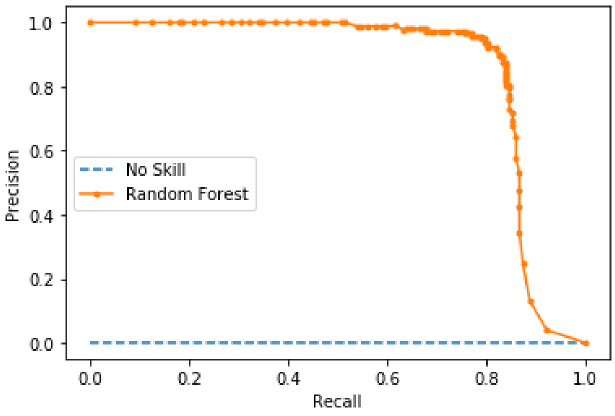
Similarly, the Classification Report shows that it predicts the no fraud transactions perfectly with a f1-score of 1 while it predicted the fraudulent transaction with f1-score of 0.81.

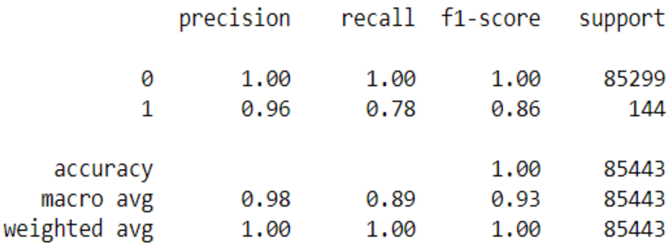
Overall, this model performed particularly good in identifying when a credit card transaction is not fraudulent with a good accuracy thereby allowing the systems of the financial institution to authorize the genuine transactions where necessary in the ideal situation.

Random Forest:

The Precision Recall Curve and Classification Report of the Random Forest Classifier depicts the following predictions.

Precision Recall Curve and Classification Report





```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     85299
           1       0.96      0.78      0.86       144

    accuracy                           1.00     85443
   macro avg       0.98      0.89      0.93     85443
weighted avg       1.00      1.00      1.00     85443
```

```
Random Forest PR AUC: 0.859
```

The performance of the Random Forest algorithm is quite exceptional as the Precision Recall Curve shows it achieves a PR AUC of about 0.859 in predicting the fraudulent transactions. The area under the curve is not perfectly in the upper right corner however this still shows that the classifier was able to capture about 0.859, which is quite a lot.

Similarly, the Classification Report shows that it predicts the no fraud transactions (0) perfectly with a f1-score of 1 while it predicted the fraudulent transaction (1) with f1-score of 0.86.

This model has the highest Precision- Recall Accuracy score making it the ideal Algorithm model that we would recommend that should be used.

Comparison of Artificial Neural Network and Random Forest Models

In terms of predicting whether a credit card transaction is fraudulent or not, we can see that the models selected above performed quite well in determining whether transactions flagged as fraud or no fraud. Three supervised learning algorithms were used in our analysis and the two best models in terms of performance are the Artificial Neural Network and Random Forest.

However, in comparing the results of the Artificial Neural Network and Random Forest models. After looking at the Performance, Accuracy, and complexity, we can invariably recommend the implementation or adoption of the Random Forest model as a mechanism for identifying what transaction should be flagged as fraudulent or not. For banks or financial institutions, this model is important to know how to accurately capture the transactions that are fraudulent or not fraudulent.

Though the two models predicted the genuine transactions perfectly which can be attributed to the fact the number of no fraud transactions in the dataset used are far larger than the fraud transactions, and also the fact that these two models have complex natures of finding the best predictions through their networks/trees.

This made it possible for the models to perfectly train the test data to identify no fraud transactions. The Random Forest model works very well in handling both numerical and non-numerical data and in this case, it achieved higher prediction values.

<u>Statement of individual contribution</u>

Chike Joseph Okwudiafor

- Contributed with analyzing and interpreting the results of the models in the report.
- Compiled and researched machine learning methodologies used in the Project.
- Worked with the team to share the responsibility of running each Machine learning Algorithm
- In collaboration with the team, worked to compile the running of the codes in the Jupyter Notebook.
- Worked to contribute to coining the business problem and solutions for the project.

Eric Oyemam-Ato Brown

- Contributed with analyzing and interpreting the results of the models in the report. ‘
- Worked with the team to share the responsibility of running each Machine learning Algorithm
- In charge of editing and compiling the Report in one document.
- Worked to contribute to coining the business problem and solutions for the project.

Obehi Irekponor

- Contributed with analyzing and interpreting the results of the models in the report.
- Dataset Information search/selection
- Worked with the team to share the responsibility of running each Machine learning Algorithm
- In charge of delegating tasks and duties for each member.
- Worked to contribute to coining the business problem and solutions for the project.

References

Andrea Dal Pozzolo, Olivier Caelen, Reid A. Johnson and Gianluca Bontempi. Calibrating Probability with Undersampling for Unbalanced Classification. In Symposium on Computational Intelligence and Data Mining (CIDM), IEEE, 2015

Bertrand Lebichot, Yann-Aël Le Borgne, Liyun He, Frederic Oblé, Gianluca Bontempi Deep-Learning Domain Adaptation Techniques for Credit Cards Fraud Detection, INNSBDDL 2019: Recent Advances in Big Data and Deep Learning, pp 78-88, 2019

Brownlee, Jason. "ROC Curves and Precision-Recall Curves for Imbalanced Classification." *Machine Learning Mastery*, 14 Jan. 2020, www.machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-imbalanced-classification/

Carcillo, Fabrizio; Dal Pozzolo, Andrea; Le Borgne, Yann-Aël; Caelen, Olivier; Mazzer, Yannis; Bontempi, Gianluca. Scarff: a scalable framework for streaming credit card fraud detection with Spark, Information fusion,41, 182-194,2018,Elsevier

Carcillo, Fabrizio; Le Borgne, Yann-Aël; Caelen, Olivier; Bontempi, Gianluca. Streaming active learning strategies for real-life credit card fraud detection: assessment and visualization, International Journal of Data Science and Analytics, 5,4,285-300,2018,Springer International Publishing

Dal Pozzolo, Andrea Adaptive Machine learning for credit card fraud detection ULB MLG PhD thesis (supervised by G. Bontempi)

Dal Pozzolo, Andrea; Boracchi, Giacomo; Caelen, Olivier; Alippi, Cesare; Bontempi, Gianluca. Credit card fraud detection: a realistic modeling and a novel learning strategy, IEEE transactions on neural networks and learning systems,29,8,3784-3797,2018,IEEE

Dal Pozzolo, Andrea; Caelen, Olivier; Le Borgne, Yann-Ael; Waterschoot, Serge; Bontempi, Gianluca. Learned lessons in credit card fraud detection from a practitioner perspective, Expert systems with applications,41,10,4915-4928,2014, Pergamon

Fabrizio Carcillo, Yann-Aël Le Borgne, Olivier Caelen, Frederic Oblé, Gianluca Bontempi Combining Unsupervised and Supervised Learning in Credit Card Fraud Detection Information Sciences, 2019

Machine Learning Group — ULB, Credit Card Fraud Detection (2018), Kaggle https://data-flair.training/blogs/data-science-machine-learning-project-credit-card-fraud-detection/

Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011