CSC-40094: Final Assessment- Machine/ Deep Learning for a Real World Problem Report

Author:
Ozgun Gizlenci

Contents

Abstract

Nowadays it is very common to face any fraud problem and the fraud problems increased parallel with the development of technology. Today it is possible to protect and make provisions for users by using machine learning algorithms and deep learning techniques. In this report, a deep learning model is created with a real-world dataset named as credit card fraud detection dataset. This project aims to observe the machine learning model and see the process of credit card fraud detection by using the machine learning model. The project is completed in Python language and performance evaluation is based on the accuracy and confusion matrix score.

Introduction

This is a document about the application of the classification techniques for credit card fraud detection by using the necessary classification algorithms such as Random Forest Classifier, Support Vector Machine Classifier, and Decision Tree Classifier. The report includes textual and non-textual diagrams, tables, and figures.

Literature Review on Credit Card Fraud Detection Techniques

Abstract: Associated with the increase in credit card transactions, a vulnerability has come out as credit card fraud. There are many classification models developed but performance and accuracy have always been the main qualification. This study observes and evaluates the performance of the classification models on credit card fraud detection problems.

Introduction

Credit cards are extensively used by people with the advent of the age of technology. Online payment gateway systems make payment much easier for the user to save time such as Paypal, Google Pay and Alipay. However, transaction fraud causes a vast amount of money lost every year. Credit card fraud is a deliberate problem for users. There is approximately $24.2 billion lost globally due to fraud and 46% of the world faces credit card fraud. In addition, fraud is a certain crime by reason of acquiring users' financial information. Credit card fraud can be separated into inner and external card fraud. This study more focuses on the inner fraud problem.

Fraud detection is basically the process of displaying the transaction fraud action from a quantitative dataset. Since credit card usage increased in both online and offline transactions, the fraud rate also increased.

There are many machine learning methods to detect credit card fraud such as artificial neural networks, decision trees, genetic algorithms, and fuzzy logic. In this study, 3 machine learning classifier model is tested for usability in fraud detection, i.e. support vector machine, decision tree and random forest classifier to observe performance and accuracy.

Techniques

Random forest, support vector machine and decision tree are widely used classification algorithms for machine learning projects. In this experiment, we use those 3 classifier algorithms to observe accuracy, confusion matrix and performance. Random forest algorithm capacity depends on the correlation between trees. There are 2 types of random forest, random-tree based random forest and cart-based random forest. In the study, we use a random-tree based random forest. It is a base classifier of the random forest that simple appliance of a decision tree called the random tree. The training set of trees is randomly selected bootstraps from the main dataset.

$$leftCenter[k] = \frac{1}{n}\sum_{i=1}^{n} x_{ik} I(y = 0) \qquad (1)$$

$$rightCenter[k] = \frac{1}{n}\sum_{i=1}^{n} x_{ik} I(y = 1) \qquad (2)$$

Figure 1 – Working principle mathematical expression for random forest algorithm

0 is represented as “leftCenter” and 1 is represented as “rightCenter” for each centre. Decisions are determined according to their values.

According to (Poongodi and Kumar, 2021), the process of discretization, min-max normalization and attribute selection are easily handled by the support vector machine and reduce the attribute intervals. It is proposed to decompose the attribute values into small pieces. These pieces will be selected with the gain based algorithm. To determine credit card fraud, low values of information gain are used. Frequent itemsets are instanced by using the Apriori algorithm and pruning accomplished to cut itemset size down. SVM uses these frequent itemsets with information gain to detect credit card fraud.

(Shen, Tong and Deng, 2007) proposed decision tree has many benefits. It is a non-parameter and very flexible method for data dissemination. Nevertheless, it is in widespread use and accountable. The decision tree is actually like a tree as its name and contains nodes and the leaf nodes are signed by following more nodes.

(Shen, Tong and Deng, 2007) also proposed neural networks and Logistic regressions for the credit card fraud classification. As a statistical method application for credit card fraud detection, the neural network method is principal. Further, there are disadvantages of the neural network. The structure is not stable, efficiency depends and training time is uncertain. On the other hand, logistic regression is another statistical method to use in credit card fraud detection. Logistic regression is helpful in prediction situations. It is similar to the linear regression model except for dependent variables. Linear probability is commonly used for business failure predictions.

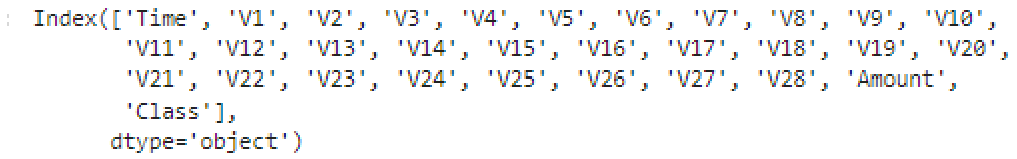
According to (Awoyemi, Adetunmbi and Oluwadare, 2017), using meta-learning and distribution of frauds in credit card fraud detection give better performance results. In the detection of fraud, decision trees surpass SVM in solving the problems. However, the decision tree’s accuracy matches the SVM classifier model but with the fraud detection count, it falls.

Conclusion

To sum up, credit card classification is performed with support machine vectors, decision trees and random forest algorithms. According to references, random forest is a better option for classification and its performance is better. On the other hand, this study includes other classification methods to evaluate and observe performance metrics such as accuracy, confusion matrix, and precision and recall.

Methodology

The dataset for the credit card fraud detection models is mainly a real-time transaction dataset that includes transaction times, class and a principal component. In general, credit card fraud detection analysis needs quantitative data to analyze properly and to get a meaningful result. On the other hand, numerical data is more efficient for predicting the values for fraud detection. The dataset is labelled with a class column, 0 means non-fraud and 1 means fraud transactions. All fraud and non-fraud transactions are used for training the model. Columns are shown in figure 2 below.



```
: Index(['Time', 'V1', 'V2', 'V3', 'V4', 'V5', 'V6', 'V7', 'V8', 'V9', 'V10',
         'V11', 'V12', 'V13', 'V14', 'V15', 'V16', 'V17', 'V18', 'V19', 'V20',
         'V21', 'V22', 'V23', 'V24', 'V25', 'V26', 'V27', 'V28', 'Amount',
         'Class'],
        dtype='object')
```

Figure 2 – Columns of the Credit Card Fraud Dataset.

Data was not a primary dataset. It is provided by Keele University with a link in the module. The dataset used in this study contains 284,807 rows and 31 columns. Besides, the dataset contains columns named as a class, time, amount and V1-V28 principal components.

Dataset is preprocessed by checking the missing values with the pandas library. Missing values are extracted from the dataset and the final dataset was ready for training.

In the learning system, the decision Tree contains the ID3 method and the decision tree works with a flowchart-like structure. In decision analysis, decision trees are such as support tool to calculate risks and aims. The decision tree model composes the high precision and small scale deep down. Nodes can be either ramification nodes or other nodes and one of them is signed as a leaf node by classification. The decision tree method has many benefits.

Firstly, it is very flexible and a non-parameter method. Besides, it forms well and is accountable.

Support vector machine classifier is a type of deep learning algorithm to help to find hyperplanes in N-dimensional space. It is very effective in high dimensional spaces either in situations such as the number of samples being greater than the number of dimensions. Another advantage of SVM is memory efficient and uses a subset of training points. The lower proportion of training data causes SVM to perform better. In this study, smaller sizes are selected with SVM selection algorithm.

In this experiment, we use the random forest algorithm as the main classifier. Decision tree models are famous for simplification in algorithms and flexibility in data mining. Although it is easy to overfit the training dataset, random forest is way more accurate than single classifiers. It is easy to combine many tree predictors. The correlation between different trees defines the capacity of the random forest.

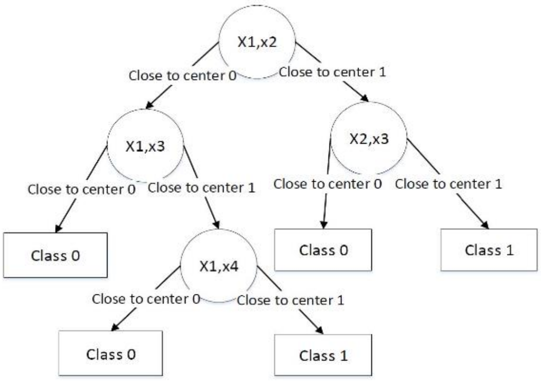


Figure 3 – Illustration of Random Tree.

Finally, about why this method is chosen, each method has its own weakness and strength. However, in credit card fraud detection the most preferred one is the random forest algorithm. The accuracy of the random forest is equal to the decision tree but model results can give differences. Random tree-based algorithms give more accurate results in classification

Results

The dataset of credit card fraud contains 284807 instances and 31 attributes. The dataset contains 28 variables as principle values. However, the dataset contains a class variable to detect the fraud transactions (0 = non-fraud, 1 = fraud).

Performance metrics for this experiment are accuracy, precision and confusion matrix. Accuracy calculates how uniquely classified. The ratio of correctly predicted measurement over total measurement outcome is the accuracy value. Accuracy can be measured by the formula represented below.

$$Accuracy = \frac{\mathrm{TP+TN}}{\mathrm{P+N}}$$

Figure 4 – Accuracy calculation formula.

TP, TN, P and N indicates numbers of true positive, true negative, positive and negative.

Another metric is precision. Precision detects correct positive predictions made in the model and for the minority, class precision measures the accuracy. The ratio of the accurately predicted positive measurement to total positive predictions. Precision can be measured by the formula represented below. FP indicates the number of false positives.

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP+FP}}$$

Figure 5 – Precision calculation formula.

Based on classification techniques, the performance is evaluated for each algorithm such as Random Forest, Decision Tree, SVM and Logistic Regression and compared to each other. Furthermore, for the attribute reduction process, the following algorithm is used with SVM.

Algorithm: Attribute Reduction

Input: Full set of attributes.

Output: Favorable attributes.

Start -> Read attributes -> Measure information entropy -> Measure information gain -> Sort the attributes -> Attribute results -> End

After preprocessing and clearing the missing values from the dataset, the class column is visualized to see the fraud/non-fraud transactions. There are 284315 normal transactions and 492 fraud transactions to use in detection models.

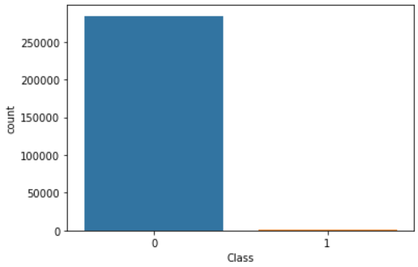


Figure 6 – Distribution of fraud and non-fraud transactions over the class column.

Moreover, following the distribution graph, another data frame is created to observe the correlation between the class variable and other variables. This correlation graph assisted to observe class value relation with the other variables.

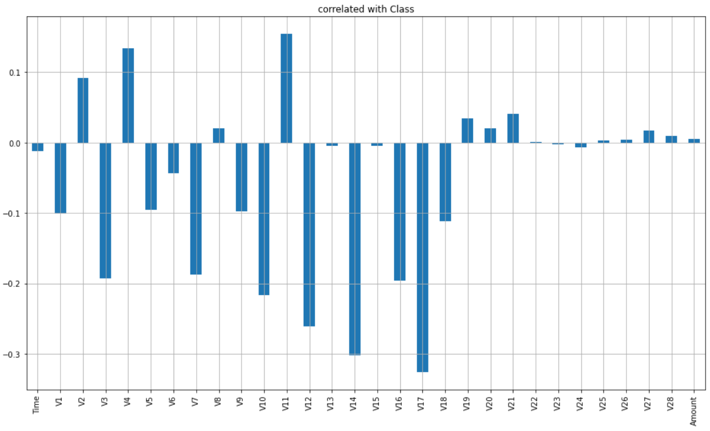


Figure 7 – Correlation of class variable with other variables.

In decision algorithms, heat maps show the data in a visual form to simplify the complex information with the help of colours and it is easy to assimilate the data. Lastly, a heatmap is created to observe the colour representation of the dataset.

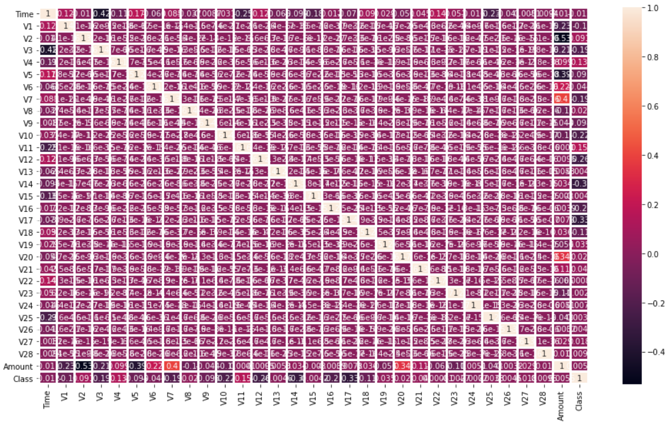


Figure 8 – Heat map representation of the credit card fraud dataset.

Independent and dependent variables of the dataset are identified and the dataset is split into train and test sets with the help of scikit-learn library's train and test split method. Subsequently, the train and test set shape is observed. 227845 attributes are separated for training in the company with 56962 attributes as a test set which is approximately 30% of the main dataset. Value distribution is resized by using the standard scaler and then train and test sets are fitted.

The first model is created with a linear logistic regression model. The Logistic regression classifier variable is created and the train set is fitted over the model subsequently, the test set is used to predict values. Lastly, the accuracy and precision and confusion matrix are calculated the observe the results. The accuracy of the logistic regression model is very high with a 99.91% value and the average precision score is 0.55 and the confusion matrix result was as it is shown below:

```
[[56852    9]
 [   37   64]]
```

After logistic regression, the high demand algorithm random forest is used in the experiment to observe accuracy and precision scores. The Random forest algorithm imported from the sci-kit-learn library and train set is fitted over the random forest classifier variable. The accuracy is 99.94% as it is expected and the precision score is 0.71 with a confusion matrix result shown below:

```
[[56854    7]
 [   22   79]]
```

Second to last, a linear support vector machine classifier model is created to test the credit card fraud detection and the SVM accuracy result is 99.93% with a 0.61 precision score with a confusion matrix result shown below:

```
[[227417    37]
 [   119   272]]
```

Lastly, a decision tree model is created to observe the accuracy and precision score. The model is fitted into 3 folds with a total of 27 fits and the maximum leaf nodes range between 2-5. The minimum sample is split into [2,3,4]. The grid search algorithm is used for hyperparameters. The accuracy result is 99.92% with a 0.60 precision score with a confusion matrix result shown below:

```
[[56844   17]
 [   26   75]]
```

Classification Model	Accuracy	Precision
Logistic Regression	91.91%	0.55
Random Forest	99.94%	0.71
SVM	99.93%	0.61
Decision Tree	99.92%	0.60

Table 1 – AccuracuResults of the classification models used in the experiment.

Classification Model	Confusion Matrix
Logistic Regression	[[56852 9] [37 64]]
Random Forest	[[56854 7] [22 79]]
SVM	[[227417 37] [119 272]]
Decision Tree	[[56844 17] [26 75]]

Table 2 – Confusion matrix results of the classification models.

Finally, we got a comparison Table 1 represented as above to observe performance evaluation for credit card detection with each classifier. The random forest-based classification model is slightly more accurate than the support vector machine and decision tree but it is well ahead of the logistic regression classifier accuracy result. Foremost, the precision score is higher than the other classification model used in this experiment. As it is expressed in the literature review section, random forest is ahead of the other classification for credit card fraud detection experiments.

Conclusion

The project examined and demonstrated the performance of different classifier models for credit card detection. Successfully evaluated the performance and compared it to each model. The study shows although the used classification models results are so close, random tree-based classifiers are more efficient and give more accurate results than the other classifiers for the fraud detection system.

List of Appendices

Jupyter Notebook File- Appendix 1: 21007505- Ozgun Gizlenci.ipynb

Class Countplot- Appendix 2: Classcountplot.png

Correlation Graph- Appendix 3: Correlation.png

Heat Map of Credit Card Dataset- Appendix 4: Heatmap.png

Credit Card Dataset- Appendix 5: Creditcard.csv

References

S. Xuan, G. Liu, Z. Li, L. Zheng, S. Wang and C. Jiang, "Random forest for credit card fraud detection," 2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC), 2018, pp. 1-6, DOI: 10.1109/ICNSC.2018.8361343.

De Sá, Alex & Pereira, Adriano & Pappa, Gisele. (2018). A Customized Classification Algorithm for Credit-Card Fraud Detection. Engineering Applications of Artificial Intelligence. 72. 10.1016/j.engappai.2018.03.011.

A. Shen, R. Tong and Y. Deng, "Application of Classification Models on Credit Card Fraud Detection," 2007 International Conference on Service Systems and Service Management, 2007, pp. 1-4, DOI: 10.1109/ICSSSM.2007.4280163.

Lin, T. and Jiang, J., 2021. Credit Card Fraud Detection with Autoencoder and Probabilistic Random Forest. *Mathematics*, 9(21), p.2683.

Awoyemi, J. O., Adetunmbi, A. O. and Oluwadare, S. A. (2017) ‘Credit card fraud detection using machine learning techniques: A comparative analysis’, *Proceedings of the IEEE International Conference on Computing, Networking and Informatics, ICCNI 2017*, 2017-Janua, pp. 1–9. doi: 10.1109/ICCNI.2017.8123782.

Poongodi, K. and Kumar, D. (2021) ‘Support vector machine with information gain based classification for credit card fraud detection system’, *International Arab Journal of Information Technology*, 18(2), pp. 199–207. doi: 10.34028/IAJIT/18/2/8.