**CREDIT CARD FRAUD DETECTION**

Abstract

As credit card transactions are a commonly used payment method, we could see the rise in credit card fraud. This is a result of technological advancements and a growth in online transactions, which lead to frauds and significant financial losses. Effective strategies to lessen the loss are therefore required. Fraudsters also find ways to obtain the user’s credit card information via phishing attack, fake calls and SMS, and so on. This project aims to predict the occurrence of fraud through Machine Learning algorithm such as Logistic Regression, Decision Tree, and Neural Network. In addition, differentiation of the supervised machine learning and deep learning techniques was done to determined if the difference between fraud and non-fraud transactions.

1. Introduction

In recent years, most people now use credit cards to pay for their necessities due to technological advancements, with this the number of credit card fraud cases has been steadily increasing. Nowadays, credit cards are accepted as payment methods by practically all business, regardless of the size of business. Credit card fraud is happening across all industries. Numerous techniques, such as data mining and machine learning, are used in conjunction with algorithmic approaches to detect credit card fraud, however it did not get considerable result. Hence, there is a increasing need of efficient and effective algorithms to be developed that works significantly. By using machine learning and Neural Network we will try to avoid fraudster using our credit. Card before the transaction get approved.

1.1. Problem statement

The majority of them now use credit cards to purchase items that they desperately need but are currently unable to afford. Credit cards are used to satisfy needs, but there is also an increase in credit card fraud, thus it is necessary to create a model that fits well and makes predictions with more accuracy (Oluwole, 2023).

1.2. Objectives

* Finding fraudulent credit card transactions is the primary goal of the research.
* Comparison between deep learning versus supervised learning, the deep learning algorithm performed better in terms of accuracy.

1.3. Existing system

The existing systems that are currently in use are machine learning algorithms which include Logistic Regression, Decision Tree, and others (Afriyie et al., 2023). Some of these algorithms also make use of random data sets.

1.4. Proposed system

The proposed system for this project is Neural Network to identify fraud in credit card transaction fraud. While Neural Networks have not been the traditional go-to method for detecting credit card fraud, recent advancements in deep learning have led to their increased application in this field. Studies have shown that Neural Network can identify complex patterns in transaction data to detect fraudulent activity (Lei et al., 2023).

1.5. Dataset

The dataset is from the Kaggle website. The data set includes credit card transactions made by European cardholders over a period of two days in September 2013. Out of a total of 2,84,807 transactions, 492 were fraudulent. This data set is highly unbalanced, with the positive class (frauds) accounting for 0.172% of the total transactions. The data set has also been modified with Principal Component Analysis (PCA) to maintain confidentiality. Apart from ‘time’ and ‘amount’, all the other features (V1, V2, V3, up to V28) are the principal components obtained using PCA. The feature 'time' contains the seconds elapsed between the first transaction in the data set and the subsequent transactions. The feature 'amount' is the transaction amount. The feature 'class' represents class labelling, and it takes the value 1 in cases of fraud and 0 in others.

2. Methods

This section describes the implementation that includes the algorithm used for implementation of proposed system.

In this project the implementation starts from loading the dataset. The data will go through pre-processing which includes data cleansing and normalising the data. The dataset than will be spliced into two dataset as train data and test data and will be model to be trained and tested. Lastly, the model will predict whether transactions is fraud or non-fraud

2.1. Packages and libraries used

* Numpy
* Pandas
* Logistic Regression
* Decision Tree Classifier
* Keras

2.2. Classification techniques

Why SVM was not tried for model building and Random Forest was not tried for few cases?

In the dataset we have 284807 datapoints and in the case of Oversampling we would have even more number of datapoints. SVM is not very efficient with large number of datapoints beacuse it takes lot of computational power and resources to make the transformation. When we perform the cross validation with K-Fold for hyperparameter tuning, it takes lot of computational resources and it is very time consuming. Hence, because of the unavailablity of the required resources and time SVM was not tried.

For the same reason Random forest was also not tried for model building in few of the hyperparameter tuning for oversampling technique.

Why KNN was not used for model building?

KNN is not memory efficient. It becomes very slow as the number of datapoints increases as the model needs to store all the data points. It is computationally heavy because for a single datapoint the algorithm has to calculate the distance of all the datapoints and find the nearest neighbors.

At the start of the project, the dataset goes through cleaning and normalizing. Then the dataset will undergo train/split test. Finally the dataset will go through modelling of each of the machine learning models below:

1. Logistic Regression

Logistic regression is a widely used statistical method for modeling the relationship between a dependent variable and one or more independent variables. In the context of credit card fraud detection, logistic regression can be applied to predict the likelihood of a transaction being fraudulent based on various features extracted from the transaction data. The model estimates the coefficients of the linear equation that best fits the relationship between the input features and the target variable. Despite its simplicity, logistic regression provides a baseline for comparison with more complex machine learning models.

Ahmed et al. (2019) investigated the application of logistic regression for credit card fraud detection, highlighting its interpretability and computational efficiency. However, the study also identified limitations, such as the assumption of linear relationships between features and the target variable, which may not hold true in real-world scenarios.

1. Decision Tree

Decision trees are non-parametric supervised learning algorithms that recursively partition the feature space into regions, based on the values of input features, to make predictions. In the context of credit card fraud detection, decision trees can effectively capture complex decision boundaries and interactions between features, making them well-suited for detecting fraudulent transactions.

Jain and Bhattacharyya (2019) evaluated the performance of decision trees for credit card fraud detection and observed promising results in terms of accuracy and interpretability. However, decision trees are prone to overfitting, especially when dealing with high-dimensional and imbalanced datasets. Regularization techniques and ensemble methods, such as random forests, can mitigate overfitting and improve the robustness of decision tree models.

1. Neural Network

Neural Networks are powerful machine learning models inspired by the structure and function of the human brain. They consist of interconnected layers of artificial neurons (nodes) that process input data and learn complex patterns through iterative optimization algorithms, such as gradient descent. In the context of credit card fraud detection, Neural Networks offer the flexibility to model non-linear relationships and capture intricate patterns in high-dimensional data.

Chen et al. (2020) proposed a deep learning Neural Network architecture for credit card fraud detection, incorporating techniques such as dropout regularization and batch normalization to improve model generalization and stability. The study demonstrated the superiority of Neural Networks over traditional machine learning models in terms of accuracy and scalability, highlighting their potential for real-time fraud detection systems.

A diagram of a network

Description automatically generated

Neural Network Overview (Chen et al, 2020)

2.3. Data Flow

Step 1: Importing the necessary packages.

Step 2: Reading dataset .read.csv (file name) # reads the data set file.

Step 3: Data learning and processing of data.

* If there is any null value, then eliminate.
* Determine correlation of variables.
* Over sampling of data is done.
* Dataset is split into two sets as train data and test data using split () on training data is used to split the data.
* Data is scaled and normalized.

2.3.1. Modelling

1. Logistic Regression

Step 1: Importing the necessary packages.

Step 2: Finding the optimal hyperparameter C.

Step 3: Fit model with the optimal hyperparameter C.

Step 4: Training the data using the LogisticRegression()

Step 5: Calculate the fraud transaction and valid transactions, the calculating the ROC, precision, and accuracy.

1. Decision Tree

Step 1: Importing the necessary packages.

Step 2: Finding the optimal hyperparameters.

Step 3: Fit model with the optimal hyperparameters.

Step 4: Training the data using the DecisionTreeClassifier()

Step 5: Calculate the fraud transaction and valid transactions, the calculating the ROC, precision, and accuracy.

1. Neural Network

Step 1: Importing the necessary packages.

Step 2: Training the model using the tf.keras.Sequential().

Step 3: Compile Model.

Step 4: Fit model and train the data.

Step 5: Calculate the fraud transaction and valid transactions, the calculating the ROC, precision, and accuracy.

Step 6: Check for overfitting.

Step 7: Regularization Training to adjust overfitting.

Step 8: Calculate the fraud transaction and valid transactions, the calculating the ROC, precision, and accuracy.

3. Results and Discussion

3.1. Evaluation measure

As the data is heavily biased, where only 0.17% transactions are fraudulent we should not just look at Accuracy for evaluating the model. We will also measure ROC-AUC score for a fair evaluation of the model.

The result is evaluated based on the confusion matrics and precision, accuracy and ROC is calculated. It contains two classes: actual class and predicted class. The confusion metrics depends on these features:

**True Positive**: in which both the values positive that is 1.

**True Negative**: it is case where both values are negative that is 0.

**False Positive**: this is the case where true class is 0 and non-true class is 1.

**False Negative**: It is the case when actual class is 1 and non-true class is 0.

Precision defined as follows:

Precision = true positive / Actual result

Precision = true positive/(true positive + false positive)

Accuracy defined as:

Accuracy = (true positive + true negative)/ total

ROC is defined as a graph that shows the performance of a classification model at all classification thresholds.

3.2. Model Comparison

Table 1 shows the result of the three algorithms on the performance metrics such as accuracy, precision and ROC.

| **Algorithms** | **Accuracy** | **Precision** | **ROC** |
| --- | --- | --- | --- |
| Logistic Regression | 0.94 | 0.97 | 0.99 |
| Decision Tree | 0.95 | 0.96 | 0.97 |
| Neural Network | 0.94 | 0.98 | 1.00 |

Table 1. Represent accuracy, precision, and ROC.

Based on the data we collected, almost all the models performed more or less good. But we should be interested in the best model. To choose the best model, we be looking at Accuracy, Precision and ROC score. Precision is important for banks as a high false positive will impact the reputation of the financial institution.

Among the three models using oversampling to balance the data, we can say that Neural Network is the best model as its precision score is 0.98 and ROC score of 1.00. Even though Decision Tree has a higher accuracy, the difference is too little to impact the outcome as shown in figure 1.

Hence, we can conclude that the Neural Network model is the best model for its precision and ROC.

A graph showing a comparison of a model

Description automatically generated with medium confidence

Figure 1

4. Implementation

In Malaysia, as in many other parts of the world, the common tools for machine learning deployment include a variety of MLOps platforms that support the end-to-end machine learning lifecycle. These tools often provide capabilities for experiment tracking, model metadata management, workflow orchestration, data and pipeline versioning, model deployment and serving, and model monitoring in production (Bellamkonda, n.d.). These are the MLOps platforms that support the end-to-end machine lifecycle:

* 1. Azure Machine Learning Studieo
  2. Seldon.io
  3. BentoMLAmazon
  4. SageMaker
  5. Google Cloud AI Platform
  6. IBM Watson Studio
  7. Heroku
  8. Algorithmia

4.1. Deployment

Here are the steps to deploy using Azure Machine Learning Studio:

1. **Register the Model**: Save the trained model in MLOps tool so it can be accessed in other projects or deployed.
2. **Prepare Entry Script and Dependencies**:
   * **Entry Script (score.py)**: This script is responsible for loading the trained model, processing input data, performing predictions, and returning the results.
   * **Dependencies File (conda\_env.yaml)**: Specifies the necessary pip and conda packages required by your service.
3. **Configure the Deployment**:
   * Define the compute target where the model will be deployed.
   * Configure the entry script and dependencies file.
   * Set up the deployment configuration with the required CPU, memory, and other settings.
4. **Deploy the Model**:
   * Deploy the model to the chosen compute target.
5. **Test the Deployment**:
   * Once deployed, test the endpoint to ensure it’s working correctly by sending sample requests and verifying the responses.
6. **Monitor and Manage the Endpoint**:
   * Use monitoring tools to keep track of the model’s performance and health.
   * Manage the endpoint through scaling, updating, or redeployment as needed.

4.2. Cost benefit analysis

We have compared the three machine learning models as of now. We have notice that most of the models have performed more or less well in terms of Accuracy, Precision, and ROC score.

While picking the best performing model is important, there are also other factors that need to be considered when choosing the best machine learning models. These factors are whether we have the required resources, infrastructure or computational power to run the model or not (Ileberi et al., 2022).

For example, models such as Decision Tree, SVM, XGBoost, and Neural Network require heavy computational resources thus, the cost will increase when building the infrastructure for deploying the model. On the other hand, the simpler model such as Logistic regression require less computational resources, so the cost of building the model is less.

5. Conclusion and Recommendations

Almost all the models performed more or less good. Based on what we gathered so far, we can conclude that Neural Network is the best model to detect credit card fraud with an ROC score of 1.00 and the highest precision of 0.98. Each models’ accuracy have too little difference for comparison.

Even though Neural Network is the best model comparatively financial institutions need to consider the cost of deploying Neural Network as a model. This is because, for computing complex models the deployment cost will be high. On the flip side, simple models like Logistic Regression have a lower cost due to requiring less computational resources.

In conclusion, Neural Network is the best Model but if resources are limited, logistic regression is a good option as the simplistic nature of Logistic Regression makes the resources is much lesser thus, making the cost lower.

For reference of the modeling here is the link to the handling of data and modeling: <https://github.com/Nabz3/Credit-Card-Fraud-Detection.git>

6. Reference

Afriyie, J. K., Tawiah, K., Pels, W. A., Addai-Henne, S., Dwamena, H. A., Owiredu, E. O., Ayeh, S. A., & Eshun, J. (2023). A supervised machine learning algorithm for detecting and predicting fraud in credit card transactions. Decision Analytics Journal, 6, 100163. <https://doi.org/10.1016/j.dajour.2023.100163>

Ahmed, I., Khan, S., Ali, H., & Khan, M. A. (2019). Application of Linear Regression for Credit Card Fraud Detection. International Journal of Advanced Computer Science and Applications, 10(8), 133-138.

Bellamkonda, S. (n.d.). Top 10 MLOPs Tools to Optimize & Manage Machine Learning Lifecycle - KDNuggets. KDnuggets. <https://www.kdnuggets.com/2022/10/top-10-mlops-tools-optimize-manage-machine-learning-lifecycle.html>.

Chen, J., Zhang, T., & Tang, X. (2020). Deep Neural Network for Credit Card Fraud Detection. Journal of Big Data, 7(1), 1-12.

Cherif, A., Badhib, A., Ammar, H., Alshehri, S., Kalkatawi, M., & Imine, A. (2022). Credit card fraud detection in the era of disruptive technologies: A systematic review. *Journal of King Saud University - Computer and Information Sciences*, *35*(1). <https://doi.org/10.1016/j.jksuci.2022.11.008>

Chauhan, M. K. S. (2023, April 19). The ultimate guide to linear regression: Mastering R-squared, Adjusted R-Squared, AIC, BIC & Mallow’s CP for Data-Driven Insights. Medium. <https://medium.com/@mukul.mschauhan/the-ultimate-guide-to-linear-regression-mastering-r-squared-adjusted-r-squared-aic-bic-69bba3b5dd6c>.

Ileberi, E., Sun, Y., & Wang, Z. (2022). A machine learning based credit card fraud detection using the GA algorithm for feature selection. Journal of Big Data, 9(1). https://doi.org/10.1186/s40537-022-00573-8

Jain, K., & Bhattacharyya, S. (2019). Evaluation of Decision Trees for Credit Card Fraud Detection. International Journal of Computer Applications, 182(45), 29-33.

Lei, Y., Ma, C., Ren, Y., Chen, X., Narayan, S., & Huynh, N. Q. A. (2023). A distributed deep neural network model for credit card fraud detection. Finance Research Letters, 58, 104547. <https://doi.org/10.1016/j.frl.2023.104547>

Oluwole, O. (2023). PREDICTIVE ANALYTICS ON CREDIT CARD FRAUD DETECTION USING CLASSIFICATION MODELS. ResearchGate. <https://doi.org/10.13140/RG.2.2.26369.02404>