Review of Research Papers

Fraud Detection

Two Research/Code pertaining to fraud detection will be examined in this review.

The code will be reproduced to compare results. The results will be analysed and recommendations will made which would have enhanced the code and hence the Machine Learning Model.

Review 1: Credit Card Fraud Detection

Credit Card companies should be able to recognize fraudulent credit card transactions so that customers are not charged for items that they did not purchase.

Summary

The dataset used in this research contains transactions made by credit cards in September 2013 by European cardholders.
This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.

It contains only numerical variables which are the result of a PCA transformations. Due to confidentiality issues, the original features and more background information about the data were not provided. 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'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' is the transaction amount. Feature 'Class' is the target variable and it takes value 1 in case of fraud and 0 otherwise.

Research Dataset: https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud

Research Notebook: https://www.kaggle.com/code/janaelshikh/credit-card-fraud-detection/notebook

Reproduced Notebook:
https://github.com/Baz177/capstone_project/blob/main/review_credit_card_fraud_detection.ipynb

Analysis and Recommendations

1. No hypertuning was done in this research notebook. Hyperparameters with values LogisticRegression(max_iter = 120, random_state = 0, n_jobs = 20, solver = ‘liblinear’) produces a slightly higher accuracy.
2. The notebook used an undersampling method to balance the data. Though this helps in increasing accuracy, oversampling could produce a higher accuracy. One way of oversampling is using SMOTE(sampling_strategy = ‘auto’, random_state = 42, k_neighbors = 4)
3. Outliers were not analysed in the notebook. Outliers should have been analysed and removed in the EDA process.
4. Though ‘accuracy_score’ is usually a quick and easy method of measuring accuracy, ROC-AUC_score is better for imbalance data.
5. Different ML models should have been tried to compare result accuracy.
6. Auto ML models could also have been implemented which would reduce the need for finding the model as well as hyperparameter tuning.

Review 2: Credit Card Fraud Detection II

Credit Card companies should be able to recognize fraudulent credit card transactions so that customers are not charged for items that they did not purchase.

Summary

The dataset used in this notebook was synthetically produced to simulate credit card transactions. The use case is primarily to create Machine Learning models which could be used for Fraud Detection. It contains several columns that represent both user and transaction information, including features for detecting fraudulent activities. The data includes a mix of categorical, numerical, and datetime values, which need to be processed for machine learning.

Research Dataset: https://www.kaggle.com/code/harsh221upadhyay/fraud-detection/input

Research Notebook: https://www.kaggle.com/code/harsh221upadhyay/fraud-detection

Reproduced Notebook: https://github.com/Baz177/capstone_project/blob/main/reveiw_credit_card_fraud_detection_2.ipynb

Analysis and Recommendations

1. The accuracy in this notebook was poor. A new model should definitely be attempted to improve accuracy and performance.
2. There was little hyperparameter tuning. A grid search would have helped to improve results
3. The pairplot in the Data exploration process did not give much information. A correlation heatmap would have helped to show relationships if there were any.
4. The notebook used ordinal Encoding instead of OneHot Encoding. OneHot Encoding may be a better option since there is no direct order for ‘profession’.

Conclusions

- Data Exploration is critical in finding and analysing relationships between features
- Hyperparameter tuning is a significant step in improving model performance
- It's important to try different models to enhance and compare performance
- In imbalance data where the minority class was grossly underrepresented oversampling may be a better option than undersampling as
- Outliers should be taken into consideration in the data exploration process
- For categorical data, it's important to choose the right encoding method. Deciphering whether there is order to the data is critical in choosing the right encoding method.