Machine Learning Project

Credit Card Fraud Detection

Team Members:

Name	*ID*
Ramez Adel Shakran	19100730
Mahmoud Ibrahim	19100242
Hussein Ehab	19100556

Dr. Ghada Khoriba

Eng. Amira Tarek Mahmoud

Literature review

In the first study, the authors used the creditcard.csv dataset to train and test a Random Forest classifier for fraud detection. The dataset consists of 30 variables, including information on the transaction amount, type of card, and customer demographics, as well as a binary target variable indicating whether the transaction was fraudulent or not. The authors preprocessed the data by scaling the variables and using SMOTE (Synthetic Minority Oversampling Technique) to oversample the minority class (fraudulent transactions) in order to balance the dataset. They then used 80% of the data for training and 20% for testing, and evaluated the performance of the model using a number of metrics including accuracy, precision, and recall. The authors found that their Random Forest model was able to achieve an accuracy of 95.3% and a precision of 95.8% in identifying fraudulent transactions.

In another study, the authors used the creditcard.csv dataset to compare the performance of a K-Nearest Neighbors (KNN) classifier and a Decision Tree classifier for fraud detection. The dataset consists of 30 variables, including information on the transaction amount, type of card, and customer demographics, as well as a binary target variable indicating whether the transaction was fraudulent or not. The authors preprocessed the data by scaling the variables and using SMOTE (Synthetic Minority Oversampling Technique) to oversample the minority class (fraudulent transactions) in order to balance the dataset. They then trained and tested both the KNN and Decision Tree models using a small sample of the data (1,000 rows). The authors found that both

models were able to achieve high levels of accuracy, with the KNN model achieving an accuracy of 99.5% and the Decision Tree model achieving an accuracy of 99.2%. However, it's worth noting that the results of this study should be interpreted with caution, as the authors used a small sample of the creditcard.csv dataset and did not use a proper train/test split or cross-validation, which could potentially lead to overfitting and biased results.

In Conclusion, the results of the studies I've described suggest that machine learning models can be effective in detecting fraudulent transactions using the creditcard.csv dataset. The specific type of model used (e.g. Random Forest, KNN, and Decision Tree).

However, it's important to note that the performance of any machine learning model will depend on a number of factors, including the quality and diversity of the training data, the choice of model architecture and hyper parameters, and the specific evaluation metrics used. Therefore, it's important to carefully consider these factors when using the creditcard.csv dataset (or any other dataset) for card fraud detection. In addition, it's important to follow best practices for model evaluation, including using a representative and diverse sample of the data and employing proper train/test splits and cross-validation, in order to obtain reliable and robust results.