STATS 771 - Project Proposal Predicting credit card fraud using KCV-SMOTE, KFS, and SVM

Pao Zhu Vivian Hsu (Student Number: 400547994)

Invalid Date

Introduction

Credit cards are a form of electronic payment that are popular for its convenience, purchase protection, and rewards. Due to its popularity, credit card companies heavily rely on fraud detection to minimize losses and maintain satisfaction among their customers. Companies have started using machine learning techniques to predict fraudulent activity and are in constant search for stronger methods to improve fraud detection.

In this paper, we propose a study to further investigate Kang and Zhang's K-fold cross-validation and synthetic minority oversampling technique (KCV-SMOTE) and key feature scanning (KFS) suggestion (2022) with a Support Vector Machine (SVM) model. The aim is to determine if such a model would produce a stronger prediction on credit card fraud.

Literature Review

To this date, a variety of machine learning techniques are being used to predict credit card fraud. This section provides a summary of some recent models found in literature.

Itoo et al. (Itoo et al., 2021) built a logistic regression, K-nearest neighbour and Naïve Bayes model. Their logistic regression model was the strongest with an accuracy of 95%.

Kang and Zhang (2022) introduced the KCV-SMOTE method to improve poor model performance caused by the imbalance of fraudulent and non-fraudulent data. They combined this method with a linear regression approach using publicly available data from online sources. Their final model

No additional work has been done to explore the KCV-SMOTE method on other classifiers since their paper was published.

Data and Methods

The proposed study will use a public data set on credit card fraud collected by Worldline and the Machine Learning Group of ULB, the Université Libre de Bruxelles (Worldline & Machine Learning Group - ULB, 2017). This is the same data used in Kang and Zhang's study (2022) and many other studies on credit card fraud (Itoo et al., 2021).

We will follow Kang and Zhang's procedure (2022) but replace logistic regression with SVM. We chose to use SVM because it is a basic classifier that is strong for classification problems, such as credit card fraud. Here are the steps involved:

1) Divide the data into a training set and a test set according to a certain ratio.
2) Use KCV-SMOTE on the training set to obtain a synthetic training set.
3) Perform key feature scanning on the synthetic training set to obtain a sub-training set.
4) Train the SVM classifier for each sub-training set separately and getting a set of AUROC values. Sort the AUROC models and select the sub-training sets with the highest scores.
5) Train the intersection of several sub-training sets obtained in the previous step to obtain the best model, and input the test set into the best model to obtain the final prediction result.

Once the models are built, we will compute their accuracy, precision, recall, F1-score, and area under the precision-recall curve (AUROC). These metrics will be used to compare models within the study. Since we are using the same data source and metrics as Kang and Zhang (2022), we will also compare the SVM models with Kang and Zhang's logistic regression models.

All analyses will be performed using Python.

Expected Results

We expect the results to perform better than linear regression model and the normal classifier.

Timelines

There are two main components to this project, a written report and a presentation. We aim to finish data collection, preliminary visualization, and data splitting by November 30, 2023. By December 31, 2023, we will finish the modelling, including a model for the default classifier method and another model using KCV-SMOTE and KFS version. By January 31, 2024, we will assess each model's performance and compare the results between models. The results will also be compared with Kang and Zhang's study (2022). By February 28, 2024, we will finish a draft of the report. By March 31, 2024, we will prepare the presentation slides. By April 20, 2024, the report and presentation will be finalized. The presentation will occur during the last week of April.

References

Itoo, F., Meenakshi, & Singh, S. (2021). Comparison and analysis of logistic regression, naïve bayes and KNN machine learning algorithms for credit card fraud detection. *International Journal of Information Technology*, *13*, 1503–1511. https://doi.org/10.1007/s41870-020-00430-y

Kang, H., & Zhang, H. (2022). A new improved method for online credit anti-fraud. *Automatic Control and Computer Sciences*, *56*, 347–355. https://doi.org/10.3103/S0146411622040046

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Worldline, & Machine Learning Group - ULB. (2017). *Credit card fraud detection*. https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data/data