

# **A simulation study on the properties of KCV-SMOTE and KFS in credit card fraud prediction**

**STATS 771 - Project Proposal**

Pao Zhu Vivian Hsu (Student Number: 400547994)

2023-12-14

## Introduction

Credit cards are an electronic form of payment that are popular for the convenience and protection they offer on everyday purchases. Due to its popularity, detecting credit card fraud is crucial for financial companies to maintain customer satisfaction and minimize losses. Machine learning techniques have become the predominant method to predict credit card fraud over the years and companies are constantly looking for ways to improve their predictions. In this paper, we propose a simulation study that explores how unbalanced and noisy data impacts Kang and Zhang’s method for online credit anti-fraud, a method that combines the K-fold cross-validation and synthetic minority oversampling technique (KCV-SMOTE) with key feature scanning (KFS) on a classifier (2022).

## Literature Review

To this date, a variety of machine learning techniques are being used to predict credit card fraud. Some techniques mentioned in recent literature include random forest, neural networks, SVM, k-nearest neighbours, logistic regression, and Naïve Bayes Itoo et al. (2021).

As described by Sulaiman et al. (2022), researchers often have challenges with the imbalance of fraudulent and genuine transactions when it comes to studying credit card fraud . Kang and Zhang introduced the KCV-SMOTE and KFS method to improve poor model performance caused by this concern (2022). The method was applied to a logistic regression approach and the final model had an AUC of 98.62% (Kang & Zhang, 2022).

No additional work has been done to explore the properties of the KCV-SMOTE and KFS method since their paper was published. Therefore, our proposed study will investigate the method’s predictive performance under different proportions of unbalanced data and ratios of signal-to-noise.

## Methods

### Simulation Study

The proposed study will use a public data set on credit card fraud as the baseline for simulation. This data set is collected by Worldline and the Machine Learning Group of ULB, the Université

Libre de Bruxelles (Worldline & Machine Learning Group - ULB, 2017), and is frequently used in credit card fraud research including the work done by Kang and Zhang Itoo et al. (2021). The ULB has created a transaction data simulator using this data (Le Borgne et al., 2022) which we will use to simulate data sets for our study.

To study how unbalanced data impacts prediction performance, we will create 5 sets of simulated data where the proportions of fraud transactions are 0.001%, 0.01%, 0.5%, 1%, and 5% of the full data set. These proportions were selected to mimic the proportion of fraud in real data sets, which tends to be less than 1% (Le Borgne et al., 2022). To study how noisy data impacts prediction performance, we will create 5 sets of simulated data where the ratios of signal predictors to noise predictors are 1:9, 3:7, 5:5, 7:3, and 9:1.

## **Application**

Once the data sets are simulated, we will then build a logistic regression model that utilizes Kang and Zhang’s KCV-SMOTE and KFS methodology (2022) for each data set. Here are details on how we will build each model based on Kang and Zhang’s work (2022):

- 1) Split the data into a training set and a testing set.
- 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 logistic regression classifier for each sub-training set separately and get a set of AUROC values. Sort the AUROC models and select the sub-training sets with the highest scores.
- 5) Train the intersection of the sub-training sets from 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) as Kang and Zhang have done (Kang & Zhang, 2022). We will then compare these metrics between models to draw a conclusion on KCV-SMOTE and KFS performance. All analyses will be performed using Python.

## **Properties of the KCV SMOTE and KFS**

This study will provide greater insight into the effectiveness of Kang and Zhang's method (2022) under different proportions of unbalanced data and ratios of signal-to-noise. We expect the method to perform best on less imbalanced data and smaller signal-to-noise ratios. Results of our study will inform further research into the method and guide the industry in building more accurate credit card fraud models.

## **Timelines**

There are two main components to this project, a paper and a presentation. By January 10, 2023, we will complete the data simulation and data splitting. By January 31, 2024, we will complete the application. By February 28, 2024, we will compute performance metrics and compare the results between models. By March 31, 2024, we will complete a draft of the paper. By March 20, 2024, we will finish the presentation slides and finalize the paper. Finally, during the last week of April, the presentation will occur.

## 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>
- Le Borgne, Y.-A., Siblini, W., Lebichot, B., & Bontempi, G. (2022). *Reproducible machine learning for credit card fraud detection - practical handbook*. Université Libre de Bruxelles. <https://github.com/Fraud-Detection-Handbook/fraud-detection-handbook>
- Sulaiman, R. B., Schetinin, V., & Sant, P. (2022). Review of machine learning approach on credit card fraud detection. *Human Centric Intelligent Systems*, 2, 55–68. <https://doi.org/10.1007/s44230-022-00004-0>
- Worldline, & Machine Learning Group - ULB. (2017). *Credit card fraud detection*. <https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud>