CS3244-12

# Comparative Study of Supervised Learning Methods for Credit Card Fraud Detection

Au Liang Jun, Dexter Wah Yixiang, Joycelyn Ng, Kenny Ng Jian Liang, Liew Jia Hong, Yan Hong Yao Alvin

## Abstract

Despite the implementation of fraud analytics and Europay, MasterCard and Visa (EMV) technology, credit card fraud rates have risen over the years due in part to increasing prevalence of cashless payments. In this study, we compare the effectiveness of traditional supervised learning methods and deep learning methods in detecting fraudulent credit card transactions via an appropriate performance metric. The findings from this comparative study could help credit card companies improve their fraud detection technology, and could be extended to detect other types of fraud.

## Metric

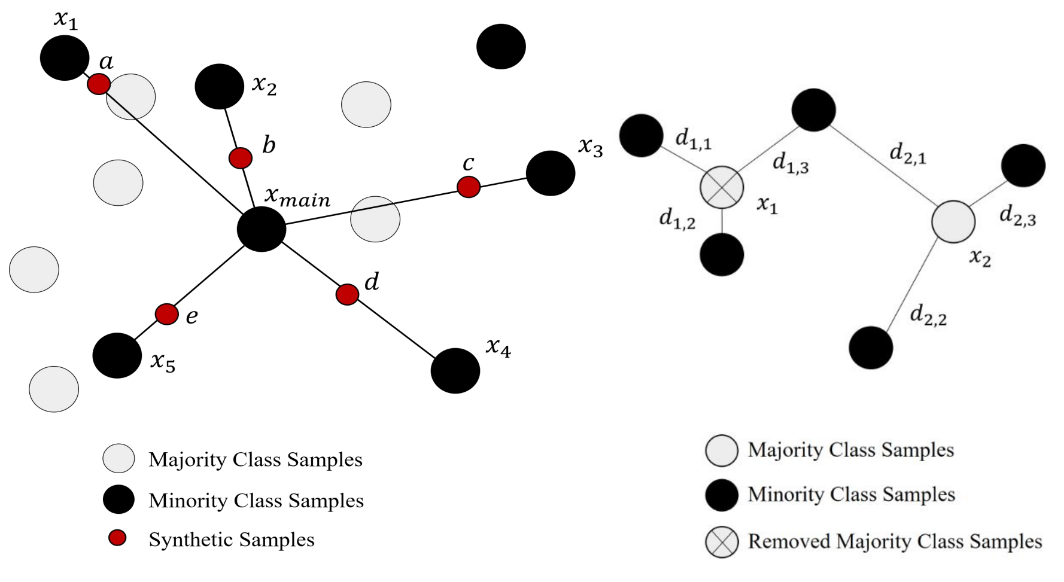
Fraudulent transactions greatly outnumber the non-fraudulent transactions in our dataset. To illustrate, a model can achieve a high classification accuracy of 99.8% by predicting all the transactions as non-fraudulent. Thus, high classification accuracy does not necessarily indicate good predictive power. We will assess our models with the Fβ score, which calculates a weighted harmonic mean of precision and sensitivity based on a pre-defined β value.

$$F_\beta = \frac{(1+\beta^2) \times Precision \times Sensitivity}{(\beta^2 \times Precision) + Sensitivity}$$

We define β as 2 in our study to give more weight to sensitivity, because we believe it is more important to identify every fraudulent transaction, even though we may inadvertently have some false positives.

## Methods

The class imbalance causes the models to be biased towards the majority class, resulting in poor classification of minority class instances. To balance the classes, we implement two types of resampling methods.

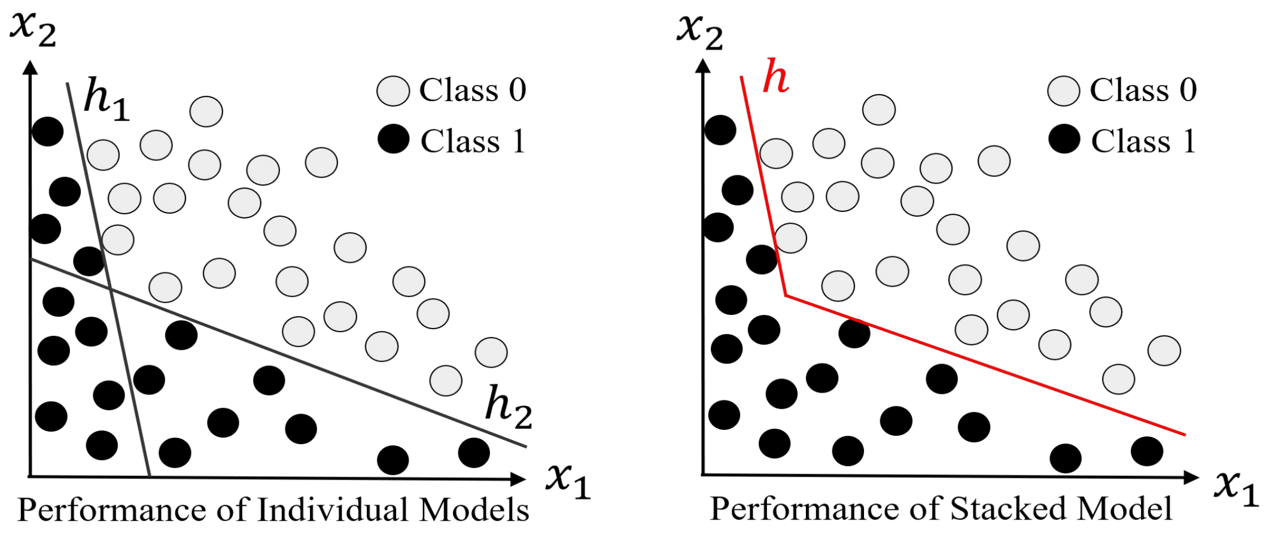


**SMOTE** **NearMiss**

The Synthetic Minority Oversampling Technique (SMOTE) is applied to oversample the minority class. New samples of the minority class are generated based on randomly selected minority class instances and their corresponding nearest neighbours. The NearMiss technique is applied to undersample the majority class. It removes majority class instances that are closest to minority class instances, based on average Euclidean distance.

Different types of supervised learning methods are implemented and compared, namely Logistic Regression, *k* Nearest Neighbours, Naive Bayes, Decision Tree, Random Forest, Single Layer Perceptron and Multi Layer Perceptron.

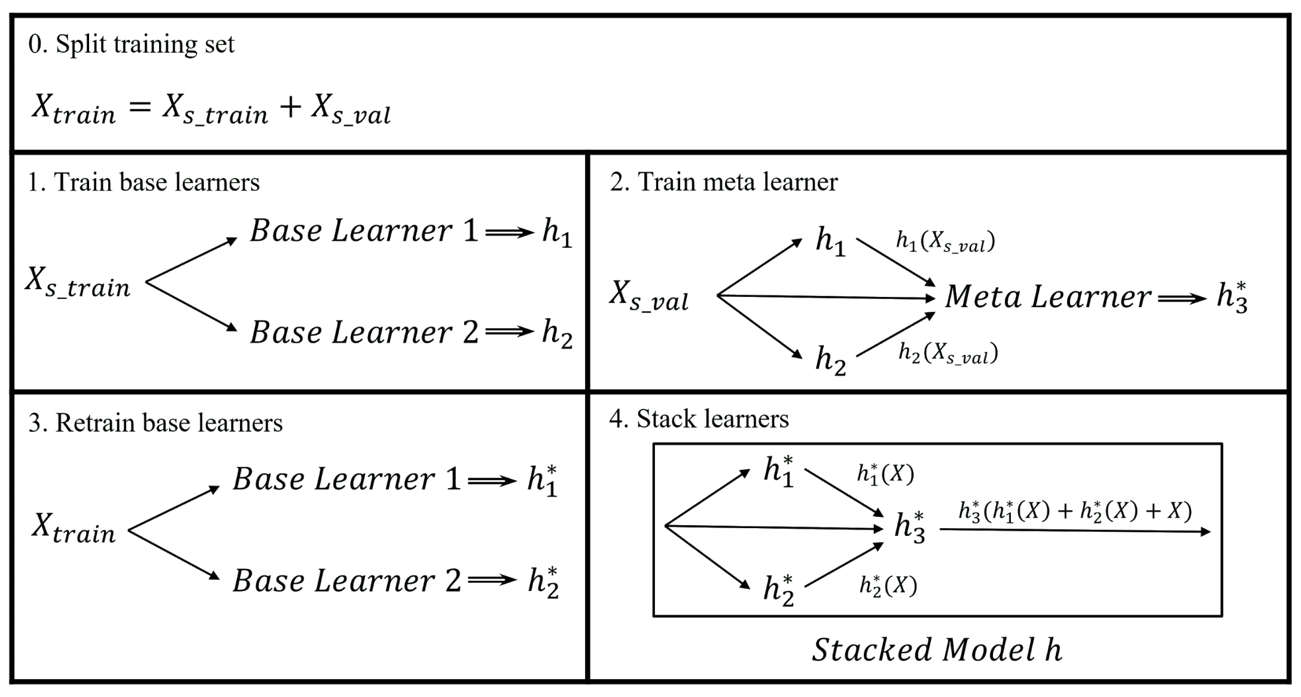
To find out if merits of individual models can be amalgamated, we explored model stacking, an ensemble learning method in which predictions from base learners are fed as inputs to a meta learner, to generate a final prediction.



**Model Stacking Intuition**

Stacking hinges on the intuition that certain models may perform better at certain regions in the multidimensional space. By learning these regions, it is possible to determine which model's prediction should be retained under particular circumstances.

The following model stacking algorithm is implemented in this study.



0. Split training set

$X_{train} = X_{s\_train} + X_{s\_val}$

1. Train base learners

$X_{s\_train}$: $Base\ Learner\ 1 \Longrightarrow h_1$, $Base\ Learner\ 2 \Longrightarrow h_2$

2. Train meta learner

$X_{s\_val}$: $h_1$, $h_2$; $h_1(X_{s\_val})$, $h_2(X_{s\_val})$ → $Meta\ Learner \Longrightarrow h_3^*$

3. Retrain base learners

$X_{train}$: $Base\ Learner\ 1 \Longrightarrow h_1^*$, $Base\ Learner\ 2 \Longrightarrow h_2^*$

4. Stack learners

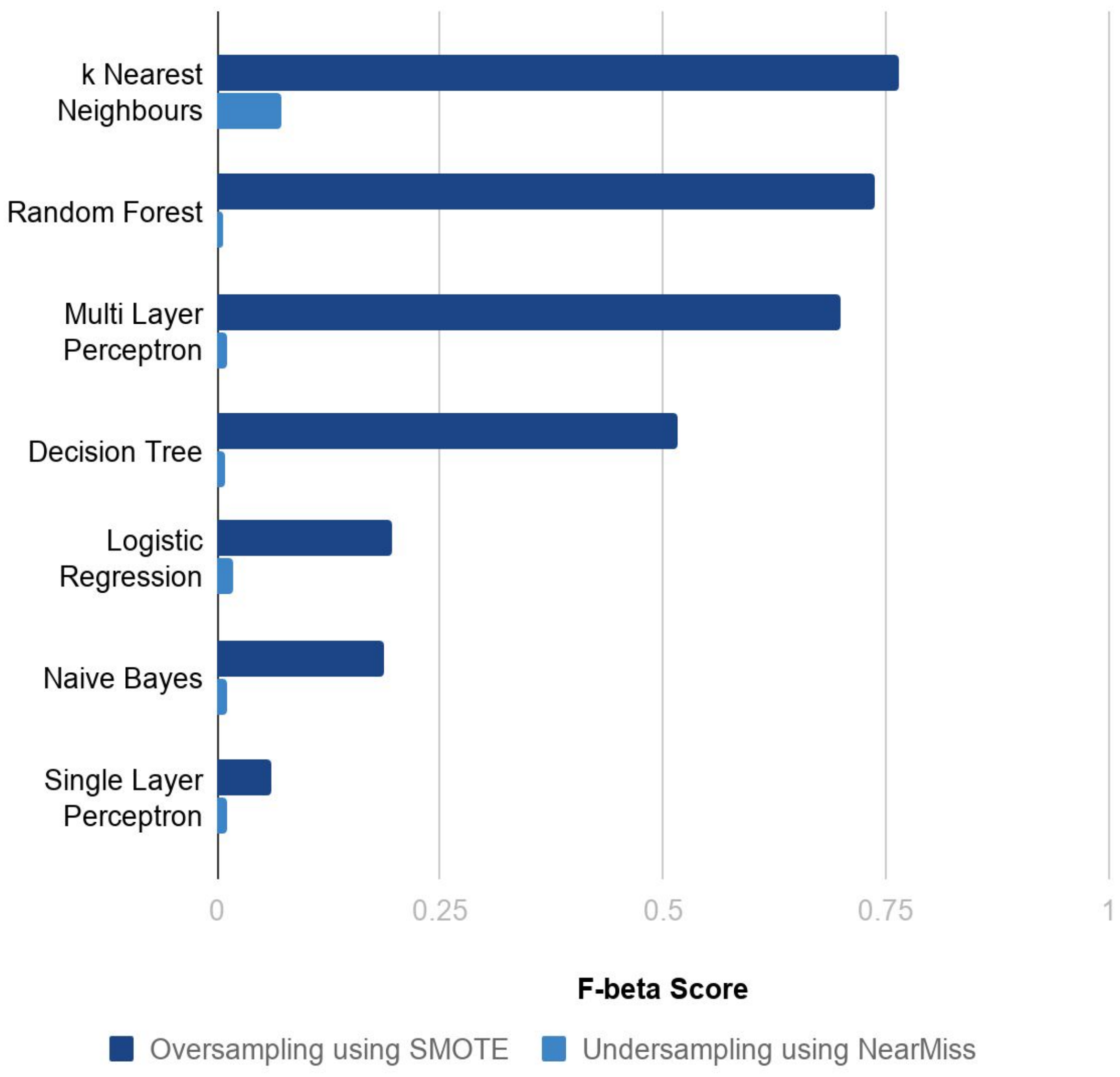
$h_1^*$, $h_2^*$; $h_1^*(X)$, $h_2^*(X)$ → $h_3^*$; $h_3^*(h_1^*(X) + h_2^*(X) + X)$

$Stacked\ Model\ h$

**Model Stacking Algorithm**

## Results

The Fβ scores for the various supervised learning algorithms are plotted for comparison.



The Fβ scores obtained from stacking the top three performing models, Multi Layer Perceptron, Random Forest and *k* Nearest Neighbours are tabulated below.

| Base Learners | Meta Learner | Fβ Score |
|---|---|---|
| *k* Nearest Neighbours and Random Forest | Multi Layer Perceptron | 0.7661 |
| *k* Nearest Neighbours and Multi Layer Perceptron | Random Forest | 0.7735 |
| Random Forest and Multi Layer Perceptron | *k* Nearest Neighbours | 0.7723 |

## Analysis

### Sampling Methods

Higher Fβ scores are observed on oversampled data as compared to undersampled data.

- SMOTE's method of generating synthetic instances between existing instances of the minority class allow models to better recognise the range of traits that fraudulent transactions typically possess.
- Undersampling with NearMiss reduces the amount of training data dramatically. With fewer majority class instances, it is likely that the models learn specific features of this smaller set of majority class instances, thus overfitting the training data for the majority class.

### Supervised Learning Methods

*k* Nearest Neighbours yields the highest Fβ score.

- Fraudsters are likely to follow a modus operandi, thus fraudulent transactions are likely to form clusters in multidimensional space. Making predictions based on the nearest neighbours principle is thus highly effective at identifying points belonging in these clusters.

Logistic Regression, Naive Bayes and Single Layer Perceptron yield low Fβ scores.

- The performance of these methods hinge on assumptions of linear separability and conditional independence. The poor performance could indicate that these assumptions are false.

Multi Layer Perceptron and Random Forest achieve relatively high Fβ scores.

- Multi Layer Perceptron and Random Forest form complex and expressive models, and hence are able to better approximate the target function.

The stacked models achieve slightly higher Fβ scores as compared to individual learners.

- The improved scores suggests that the meta learner is able to leverage on the strengths of the base learners to perform better than the individual learners. However, the marginal improvement may indicate that our implementation is lacking diversity, which is crucial in ensembling methods.

## Conclusion

Our findings show that in a fraud dataset, oversampling of the minority class yields better results compared to undersampling of the majority class. *k* Nearest Neighbours, Random Forest and Multi Layer Perceptron, as well as techniques like model stacking, work well in detecting fraudulent credit card transactions.

The workflow used in study can also be applied to other problems with imbalanced datasets, such as the detection of rare diseases where there are very few diagnosed cases.

## References

Dataset Source: Kaggle.com. (2018). Credit Card Fraud Detection. [Online] Available at: https://www.kaggle.com/mlg-ulb/creditcardfraud