Journal of Data Science,17(1). P. 81 - 106,2019
DOI:10.6339/JDS.201901_17(1).0004

A ENSEMBLE MACHINE LEARNING BASED SYSTEM FOR MERCHANT CREDIT RISK DETECTION IN MERCHANT MCC MISUSE

Chih-Hsiung Su[1], Fengjun Tu[2], Xinyu Zhang[3], Ben-Chang Shia[4], Tian-Shyug Lee[5*]

[1]*Department of Accounting Information, Chihlee University of Technology, New Taipei City, Taiwan*

[2]*School of Business Administration, Guizhou University of Finance and Economics,Guiyang, China*

[3]*Guiyang No. 1 High School, Guiyang, China*

[4]*College of Management, Taipei Medical University, Taipei, Taiwan*

[5]*Graduate Institute of Business Administration, College of Management, Fu Jen Catholic University, New Taipei City, Taiwan*

ABSTRACT

Although credit score models have been widely applied, one of the important variables-Merchant Category Code (MCC)-is sometimes misused. MCC misuse may cause errors in credit scoring systems. The present study aimed to develop and deploy an MCC misuse detection system with ensemble models, gives insights into the development process and compares different machine learning methods. XGBoost exhibited the best performance, with overall error, sensitivity, specificity, F_1 score, AUC and PRAUC of 0.1095, 0.7777, 0.9672, 0.8518, 0.9095 and 0.9090, respectively. MCC misuse by merchants can be predicted with satisfactory accuracy by using our ensemble-based detection system. The paper can thus not only suggest the MCC misuse cannot be overlooked but also help researchers and practitioners to apply new ensemble machine learning based detection system or similar problems.

Keywords: MCC misuse ,credit risk ,ensemble machine learning

* Corresponding author .
Email: 036665@mail.fju.edu.tw

1 Introduction

With the continuous development of Big Data, the processing and analysis of massive volumes of data has become a major challenge. Extracting valuable information from huge amounts of data has become a goal that is continually pursued. With credit cards being an important payment tool, the promotion of peoples credit consciousness, and the need to improve credit risk assessment, many machine learning methods applicable to credit risk evaluation have been developed. However, with the increasing use of Internet transactions, e-commerce and Bitcoin, credit card risk problems have become increasingly widespread. It should be noted that credit risk not only results from cardholders credit but also from merchants credit, which is one of the most important channels of credit card transactions. For various reasons, merchants trading information is sometimes changed in order to reduce costs or avoid supervision, which is illegal behaviour. The existing quantitative research on merchant credit risk, however, is insufficient.

The credit score model has been widely applied to assist commercial banks and international credit card organizations to identify and manage cardholders risk. An accurate credit score model can adequately protect these organizations from bad debt. In addition, credit risk models on cardholders not only include application scoring, behaviour scoring and collection scoring, but also other scoring systems on interest-bearing assets, cross marketing (such as insurance telemarketing), customer churn and cash-out.

The issue of personal credit scoring systems on cardholders has long been investigated. Erdem (2008) used 474 Turkish credit card customers data to construct a structural equation model, and found that education level, marital status, number of children, spouses employment status and other factors have certain influence on default risk. Schreiner (2004) used data from Bolivia, and found that womens credit default rates were lower than mens, and that gender had different effects on credit risk. Carow and Staten (1999) found that credit card users tend to be younger, more educated and have more credit cards, and these peoples credit risk is often higher. Traditional Logistic Regression, Logistic Regression with penalized variable selection, Support Vector Machine, Neural Network and Random Forest are the most widely used machine learning techniques in this field.

In the present study, we used the ensemble-based method to conduct feature selection and build a MCC misuse detection system, and examined the performance of the classification models using several indices. The marginal effect of important variables was also discussed with reference to partial dependence plots.

2 Merchant case study

Although many credit score models have evaluated most aspects of cardholders profiles, most of these models depend on cardholders expenditure behaviour. However, we always overlook whether the merchant has been accurately categorised. Şahin and Duman (2011), Gadi et al. (2008) and Bhattacharyya et al. (2011) have built several credit card fraud models using Merchant Category Code (MCC) as their independent variable . Hand (2007) also mentioned that the MCC is intrinsically more likely to be associated with fraud. A Merchant Category Code (MCC) is a four-digit number assigned to a business by a bank or card organization (such as American Express, MasterCard, Visa, UnionPay) in order to accept one of these cards as a form of payment. The code reflects the category in which the merchant does business and may be used by credit card companies to offer cash back rewards or reward points for spending in specific categories. The MCC includes hotel, catering, entertainment, jewellery, real estate, wholesale, air ticketing, refuelling, supermarket, hospital, school classes and general merchant categories with hundreds of codes. A merchants POS terminal can only have one code which represents what kind of goods the merchant sells. If the code does not match the merchants actual situation, we would treat them as an MCC misuse merchant.

For authorized merchants in the settlement of credit card spending, the issuing bank will charge a certain ratio of transaction fees based on their industry categories; this ratio is called the merchant discount rate. For example, for UnionPay merchants, the discount rate is 0.45% for general category, 0.351% for the livelihood category and zero for the public welfare category. If a merchant have 100,000 per day trading flow, the rate difference of 0.45%, thus it will lost 165,000 in a year. For Visa, MasterCard and AMEX merchants, the discount rate is in 1% to 4%. Also, each category has its own transaction regulation model to predict some illegal transactions. MCC is the key which connects the card issuing bank, c ardholder, card-acquiring bank, merchant, card organization and related firms in the transaction network. In the first half of 2014 in China, up to 460,000 merchants

misused their MCCclose to 6% of all merchant activity. Alliston (2002) also built a system to detect incorrect merchant codes.

In recent years, some so-called big data techniques to predict credit risk and cardholder behaviour models have treated MCC as an important variable. However, few studies have focused on MCC misuse. MCC misuse may result in fake merchants, merchants malicious closures, merchants financial deterioration, merchants illegal cash out, merchants theft of cardholder information, merchant fraud and other risks. Therefore, a MCC misuse detection system can effectively reduce the risk of financial institutions. Zhang (2015) also mentioned this problem and use Group Bridge-Logistic model to identify the MCC misuse merchants. However, the overall prediction accuracy is just 62.12%. The objective in this study is to develop a system which will result in a higher automation level, while the accuracy is as higher as possible.

3 Machine learning techniques for credit risk

Machine learning methods have frequently been used in the analysis of the credit scoring system because they require fewer assumptions and deliver higher analytical accuracy. The ensemble model is one of the commonly-used algorithms in current machine learning techniques. The tree-based ensemble model using bagging and boosting is especially popular. Given that the hierarchical tree structure can model non-linear associations, this method is typically used for regression and classification, and is likely to perform well for complex, independent variables. Random Forest (RF) is an ensemble machine learning method that uses multiple trees as classifiers using bagging. After taking the majority vote over all classifiers, the RF method combines information across all trees to reveal variable importance. Boosting methods such as Adaptive Boosting (AdaBoost) and Gradient Boosting Machine (GBM) are another kind of ensemble method with strong similarity to RF. Ensemble models have been applied to many financial studies, such as studies concerning credit risk, customer profit, stock prices and automated trading.

In practice, the ensemble model can combine many weak, simple models to obtain a stronger ensemble prediction. In real credit risk applications, many results shows that traditional, single-prediction models have lower prediction accuracy and are less robust than ensemble models, especially in high-dimensional or large sample data sets. In this research, ensemble models were the majority concerned.

In addition to accuracy and robustness, computational efficiency is also an important factor. To demonstrate the performance of ensemble models, we also compared them against traditional algorithms such as Decision Tree (CART) and Support Vector Machine (SVM).

This study adopted a 10-fold cross-validation method on feature selection: each original dataset has been randomly divided into ten stratified parts of equal (or approximately equal) size. For each fold that was employed as testing data, the other nine folds were employed as training data. In the final prediction model, the data was divided randomly 100 times with 90% used for training and the remaining 10% for testing.

The model performance evaluation examined by several indices. First, prediction accuracy and Cohen's Kappa coefficient were used to assessed feature selection of different independent variable set models. The Receiver Operating Characteristic (ROC) curves with Area Under the Curve (AUC) were plotted in the final prediction model, and the sensitivity, specificity and F_1 score for each prediction model was calculated.

3.1 Random Forest

Bagging is one of the ensemble algorithms in machine learning used to improve the stability and accuracy of machine learning algorithms. Random forest (RF), proposed by Breiman, is one such algorithm. RF can also help identify the truly relevant predictor variables so that feature selection can be conducted by the model. Furthermore, some results also illustrate the importance of the choice of the number of variables in each tree and it is found to be optimal with respect to prediction accuracy in empirical studies.

3.2 AdaBoost.M1

Boosting is another kind of ensemble algorithm for improving the accuracy of any given learning algorithm, and it means that a weak learning algorithm better than random guessing in a Probability Approximately Correct (PAC) model can be boosted into a strong learning algorithm. The Adaptive Boosting (AdaBoost) algorithm solved many of the practical difficulties with the earlier boosting algorithms. AdaBoost.M1 is used to extend AdaBoost to multi-class cases in generalization.

3.3 eXtreme Gradient Boosting

eXtreme Gradient Boosting (XGBoost), derived by Chen and He (2015), is one kind of GBM model. Both XGBoost and GBM follow the principle of gradient boosting, but there are differences in modelling details. Specifically, XGBoost uses a more regularized model formalization to control over-fitting, which grants better performance. XGBoost has used second derivative information, and ordinary GBM only uses first-order derivatives. XGBoost models greatly optimize the traditional gradient boosting model, and is one of the fastest learning algorithm of gradient boosting algorithm.

3.4 Light Gradient Boosting Machine

Light Gradient Boosting Machine (LightGBM) a gradient boosting framework that uses tree-based learning algorithms. It is highly efficient and scalable, and can support many different GBM algorithms. This method was developed by Microsoft Research Asia. LightGBM has been shown to be several times faster than existing implementations of gradient boosting trees, due to its fully greedy tree-growth method, histogram-based memory and computation optimization. LightGBM adds a maximum depth limit on the leaf-wise algorithm to ensure high efficiency and prevent overfitting.

4 Data understanding and preparation

For our analysis we focus on one international card organization due to the different card organization have some different coding on MCC. The raw data contains 56,129 merchant records from 2014 belonging to general merchants of China UnionPay. Specifically, to each record there corresponds a merchant, and each record contains the annual summary information on the merchant. We have collected 74 independent variables involving transactions in different times of the day (morning, lunch, afternoon, evening and night), different card types (credit card and debit card), different days of the week (weekday and weekend) and other merchant information.

The dependent variable of each merchant record in our study has to be labelled as a binary variable regarding whether MCC misuse or not. Two criteria are taken into consideration when defining the merchant of the dataset and we created a variable which is called IS_TY with a value of 0 or 1, such that

1. a merchant was classified as MCC misuse by bank or card organization was labelled as MCC misuse (IS_TY=1); and
2. all other merchants were labelled as no MCC misuse (IS_TY=0).

It should be note that not all merchants can be labelled, due to the long time it takes for the bank or card organization to realise that a merchant abnormal transactions has occurred. We restricted our analysis to merchant records which satisfy the following criteria:

3. no much missing value. Among the dataset of missing data, each record has fewer than 9 missing variables. Data imputation methods cannot fit the data well if lots of variables are missing; and
4. no abnormal data. The record does not include abnormal values (e.g., negative amounts or frequencies; whole year transaction amount being less than the sum of morning, lunch, afternoon, evening and night).

These criteria were chosen to make sure the merchants were really functioning and the records had no errors or missing information. Then, 38,365 samples were used in the following analysis. All merchants were grouped into 2 categories by MCC misuse, where 15,674 were no MCC misuse and 22,691 were MCC misuse. The meaning description of the independent variable is shown in Table1. Most of the independent variables are continuous

variables, and only one, ACQ_3RD_COMPANY, is a categorical variable.

Tabel 1: Independent Variables Description of the Dataset

Name	Description	Name	Description
ACQ_3RD_COMPANY	Is Third-Party Acquirer	LUNCH_TRANS_AMT	Lunch Transaction
ACTIVE_MONTH_H1	1st Half Year Active Months	LUNCH_TRANS_AMT_PCT	
ACTIVE_MONTH_H2	2nd Half Year Active Months	LUNCH_TRANS_AMT_PF	
AFTERNOON_AMT	After Transaction	LUNCH_TRANS_FREQ	
AFTERNOON_AMT_PCT		LUNCH_TRANS_FREQ_PCT	
AFTERNOON_AMT_PF		MAX_AMT	Maximum Transaction Amount

AFTERNOON_FREQ		MIN_AMT	Minimum Transaction Amount
AFTERNOON_FREQ_PCT		MORNING_AMT	
CREDIT_N_CUSTOMER	Count of Credit Card	MORNING_AMT_PCT	
CREDIT_TRANS_AMT		MORNING_AMT_PF	Morning Transaction
CREDIT_TRANS_AMT_PCT		MORNING_FREQ	
CREDIT_TRANS_AMT_PF		MORNING_FREQ_PCT	
CREDIT_TRANS_FREQ	Credit Card Transaction	N_CARD_YEAR	Year Number of Card
CREDIT_TRANS_FREQ_PCT		N_CUSTOMER_YEAR	Year Number of Customer
CV	Coefficient of Variation	N_ISS_BANK_NM	Number of Interbank Transactions
DEBIT_N_CUSTOMER	Count of Debit Card	NIGHT_AMT	
DEBIT_TRANS_AMT		NIGHT_AMT_PCT	Night Transaction
DEBIT_TRANS_AMT_PCT		NIGHT_AMT_PF	
DEBIT_TRANS_AMT_PF	Debit Card Transaction	NIGHT_FREQ	
DEBIT_TRANS_FREQ		NIGHT_FREQ_PCT	
DEBIT_TRANS_FREQ_PCT		RANGE_AMT	Range of Monthly Transaction
EVENING_AMT		SETTLE_DT	Days Used
EVENING_AMT_PCT		TRANS_AMT	Transaction Amount
EVENING_AMT_PF	Evening Transaction	TRANS_AMT_PER_CARD	Transaction Amount per Card
EVENING_FREQ		TRANS_AMT_PF	Transaction

			Amount per Frequency
EVENING_FREQ_PCT		TRANS_FREQ	Transaction Frequency
EVENING_FREQ_PCT		TRANS_FREQ	Transaction Frequency
HOLIDAY_ND_AMT	National Day Transaction	TRANS_FREQ_PER_CARD	Transaction Frequency per Card
HOLIDAY_ND_AMT_PCT		WEEKDAY_AMT	Weekday Transaction
HOLIDAY_ND_AMT_PF		WEEKDAY_AMT_PCT	
HOLIDAY_ND_FREQ		WEEKDAY_AMT_PF	
HOLIDAY_ND_FREQ		WEEKDAY_AMT_PF	
HOLIDAY_ND_FREQ_PCT		WEEKDAY_FREQ	
HOLIDAY_NY_AMT	New Year Transaction	WEEKDAY_FREQ_PCT	
HOLIDAY_NY_AMT_PCT		WEEKEND_AMT	Weekend Transaction
HOLIDAY_NY_AMT_PF		WEEKEND_AMT_PCT	
HOLIDAY_NY_FREQ		WEEKEND_AMT_PF	
HOLIDAY_NY_FREQ_PCT		WEEKEND_FREQ	
LUNCH_N_CUSTOMER	Lunch Customers	WEEKEND_FREQ_PCT	

In order to get the best performance of the machine learning algorithms, the data must be clean and complete. We chose random forest to imputing values in the missing variables. The characteristics of the merchants on MCC misuse is shown in Table5.

All analyses in this study were implemented using R Software Version 3.3.2 (www.r-project.org). We used the adabag package for AdaBoost.M1, the randomForest package for RF, the xgboost package for XGBoost, the lightgbm package for LightGBM, the rpart package for CART, and the kernlab package for SVM.

5 Modelling and evaluation

5.1 Feature selection

As 74 independent variables were recorded, feature selection was necessary before classifier construction. Ensemble models Adaptive Boosting M1 (AdaBoost.M1), Random Forest (RF), eXtreme Gradient Boosting (XGBoost) and Light Gradient Boosting Machine (LightGBM) can rank the importance of variables in the model so that we can also conduct feature selection by these models. A 10-fold cross-validation model was developed with 7 kinds of top variable sets, which are sorted by each models importance ranking. The number of top variables was chosen on 19, 37, 42, 47, 52, 56 and 74 (all). Using prediction accuracy and kappa to evaluate the model performance, the results are shown in Table2. Almost every model suggested that the top 42 variables can perform better in terms of prediction accuracy and consistency (compared with other numbers of top variables). Finally, combined with the four models relative importance, leaving the top 42 variables for the classifier construction (Fig.1). First Half Year Active Months (ACTIVE_MONTH_H2) and Days Used (SETTLE_DT) were the 2 most important variables in all models. The importance of active months in the first six months of a year is much greater than the number of days for which the POS terminal was settled for all models except the random forest model. New Year Amount per Transaction (HOLIDAY_NY_FREQ_PCT), Maximum Transaction Amount (MAX_AMT) and Second Half Year Active Months (ACTIVE_MONTH_H1) are ranked the top 5 important variables for the boosting models (AdaBoost.M1, XGBoost and LightGBM). However, in the bagging model (RF), the third to fifth-most important variables were Maximum Transaction Amount (MAX_AMT), Transaction Frequency per Card (TRANS_FREQ_PER_CARD) and Transaction Amount per Card (TRANS_AMT_PER_CARD).

Tabel 2: Model Performance of Different Variables

Number of Variables	Index	AdaBoost.M1	Random Forest	XGBoost	LightGBM
19	Accuracy	0.8611	0.8668	0.8799	0.8524
	Kappa	0.7048	0.7172	0.7696	0.6903
37	Accuracy	0.8603	0.8820	0.8828	0.8540
	Kappa	0.7024	0.7491	0.7510	0.6935
42	Accuracy	0.8608	0.8836	0.8841	0.8540
	Kappa	0.7035	0.7525	0.7536	0.6936
47	Accuracy	0.8611	0.8817	0.8825	0.8540
	Kappa	0.7042	0.7484	0.7502	0.6934
52	Accuracy	0.8610	0.8827	0.8826	0.8557
	Kappa	0.7037	0.7505	0.7505	0.6971
56	Accuracy	0.8589	0.8803	0.8835	0.8531
	Kappa	0.6999	0.7455	0.7524	0.6915
74(ALL)	Accuracy	0.8586	0.8803	0.8827	0.8577
	Kappa	0.6987	0.7456	0.7508	0.7010

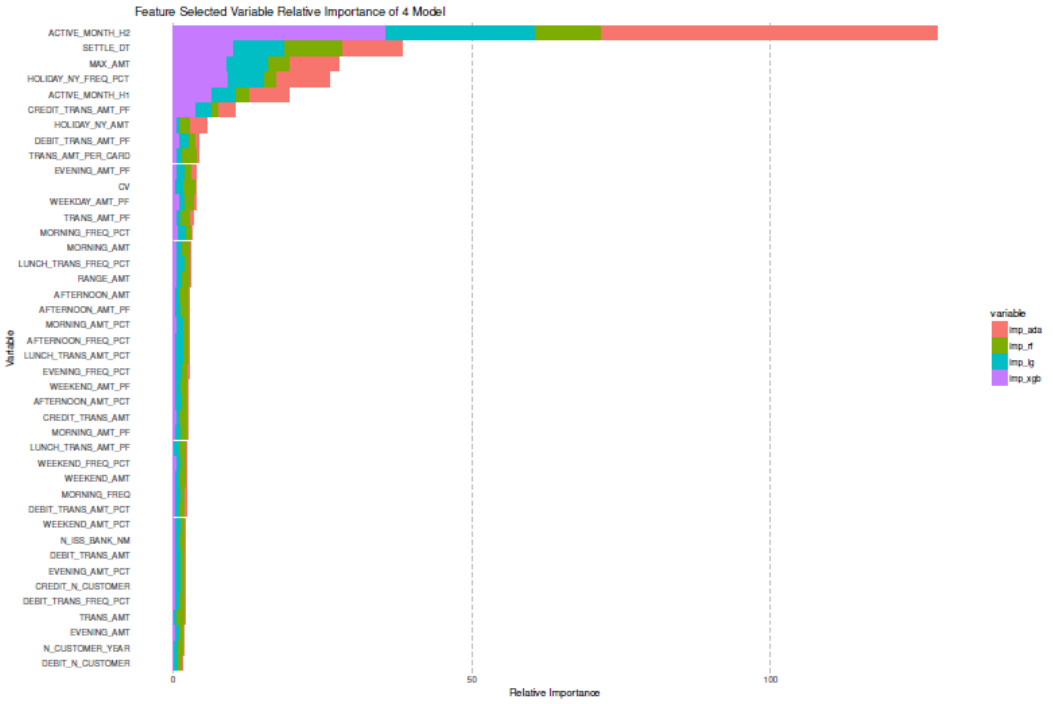


Figure 1: Combined Relative Variable Importance

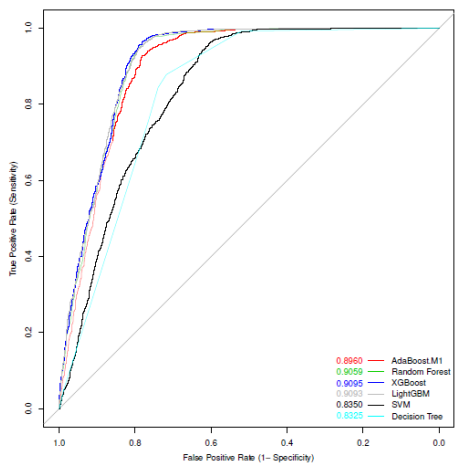
5.2 Classification performance

The performance of ensemble models was compared with Classification And Regression Tree (CART) and Support Vector Machine (SVM). All results were averaged over 100 times on the data sets, with 90% for training and the remaining 10% for testing.

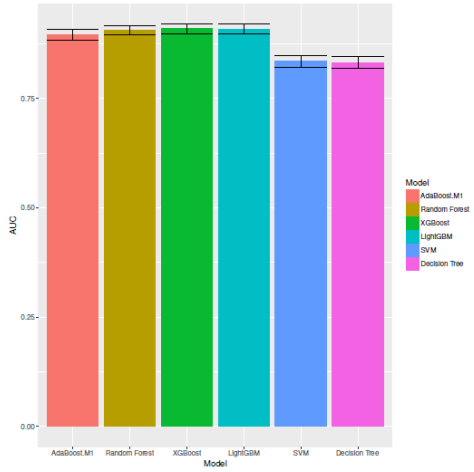
The average prediction errors are shown in Table3, indicating that XGBoost performed better than other models on all types of errors. The ensemble models performed significantly better than traditional models. All models performed better on MCC misuse than on No MCC misuse. However, the bagging model RF here had higher accuracy on MCC misuse than LightGBM but lower accuracy on no MCC misuse. That is because the bagging model was using bootstrap random select samples and variables and combining their vote to the final model. This type of voting will improve misclassification equally with growing number of trees. XGBoost and LightGBM are based on gradient boosting models where each new model is created so that the residuals of the previous model are reduced in the direction of the gradient. In this case, the misclassification of No MCC misuse was much larger than MCC misuse, so it had greater weight in the iteration. The Receiver Operating Characteristic (ROC) curve and average Area Under the Curve (AUC) also suggested that the ensemble models performed much better than traditional models (Fig.2a and Fig.2b). The AUC of ensemble models was almost over 0.9000. While the models had a consistent hierarchy of accuracy for ensemble models, with the XGBoost model being the most accurate and the AdaBoost.M1 model being the least accurate, the XGBoost model had the best AUC at 0.9095 (95% CI: 0.8989-0.9202).

Tabel 3: Model Prediction Errors

Prediction Errors	AdaBoost.M1	Random Forest	XGBoost	LightGBM	CART	SVM
MCC Misuse	0.0499	0.0350	0.0328	0.0403	0.1217	0.0482
No MCC Misuse	0.2532	0.2307	0.2223	0.2184	0.2822	0.3879
Overall	0.1322	0.1142	0.1095	0.1124	0.1867	0.1856



(a) ROC Curve and AUC on Testing Data



(b) Comparison of AUC with 95% CIs

Figure 2: ROC and AUC of testing data on different models

Table4 provides several indices to examine the performance of the classification models. XGBoost, LightGBM and RF were more robust than the other models. The corresponding sensitivity and specificity for XGBoost were 0.7777 and 0.9672, respectively. Also, XGBoost had the highest F_1 score of all the models.

Table 4: The Performance of the Classification Models

Index	AdaBoost.M1	Random Forest	XGBoost	LightGBM	CART	SVM
Sensitivity	0.7468	0.7693	0.7777	0.7816	0.7176	0.6121
Specificity	0.9501	0.9650	0.9672	0.9597	0.8783	0.9518
F_1	0.8205	0.8450	0.8518	0.8491	0.7568	0.7274

In addition, due to the XGBoost model performed the best here. We focus on XGBoost for further information regarding predictions. XGBoost model can find out not only the different importance between variables, but also the different relationship of inner variables. One way to investigate these relations is with partial dependence plots. These plots are graphical visualizations of the marginal effect of a given variable (or multiple variables) on an outcome. Fig.3 shows partial dependence plots of the marginal effect of the top 10 important variables against the independent variable. Note that ACTIVE_MONTH_H1, ACTIVE_MONTH_H2 and SETTLE_DT are all related to the merchants activity status. However, the partial dependence of these three variables do not demonstrate the same trend. SETTLE_DT had a positive correlation on the probability of MCC

misuse; when the POS terminal was settled over one year, the probability of MCC misuse was almost equal to one. ACTIVE_MONTH_H2 concerns the months from July to December; if a merchant did not have any transactions, they would have a very low probability of MCC misuse. Conversely, if a merchant has transactions in all of the six months, they are more likely to be an MCC misuse merchant. ACTIVE_MONTH_H1 had the opposite impact compared to the first six months; if the merchant is more active in these six months, the MCC misuse probability decreases. It also suggests that if a merchant always has a large number of transactions, no matter whether credit cards or debit cards are used, they are probably not an MCC misuse merchant (refer to plots for CREDIT_TRANS_AMT_PF and DEBIT_TRANS_AMT_PF). In contrast, if the average transaction amount is small, but the largest transaction amount is very high, there will be a higher risk of MCC misuse (refer to plots for MAX_AMT, CREDIT_TRANS_AMT_PF and DEBIT_TRANS_AMT_PF). New year is a special holiday; it is the first day of a year. HOLIDAY_NY_FREQ_PCT shows that if these transactions account for too high a proportion of annual transactions (just over 5% of annual transactions), the risk of MCC misuse grows rapidly. CV is the coefficient of variation, and is defined as the ratio of the standard deviation σ to the mean μ. CV can reflect the volatility of the merchant; if the CV value is higher the risk of MCC misuse is also higher. Here there maybe have an outlier around 0.2. If the weekend transaction frequency accounts for a larger proportion of all transactions than weekdays, the risk will reduce (WEEKEND_FREQ_PCT). If the proportion of transaction frequency at lunch is over 0.5, the curve is flat at a low risk. Otherwise, the risk of misuse will decrease with the increase of proportion (LUNCH_TRANS_FREQ_PCT).

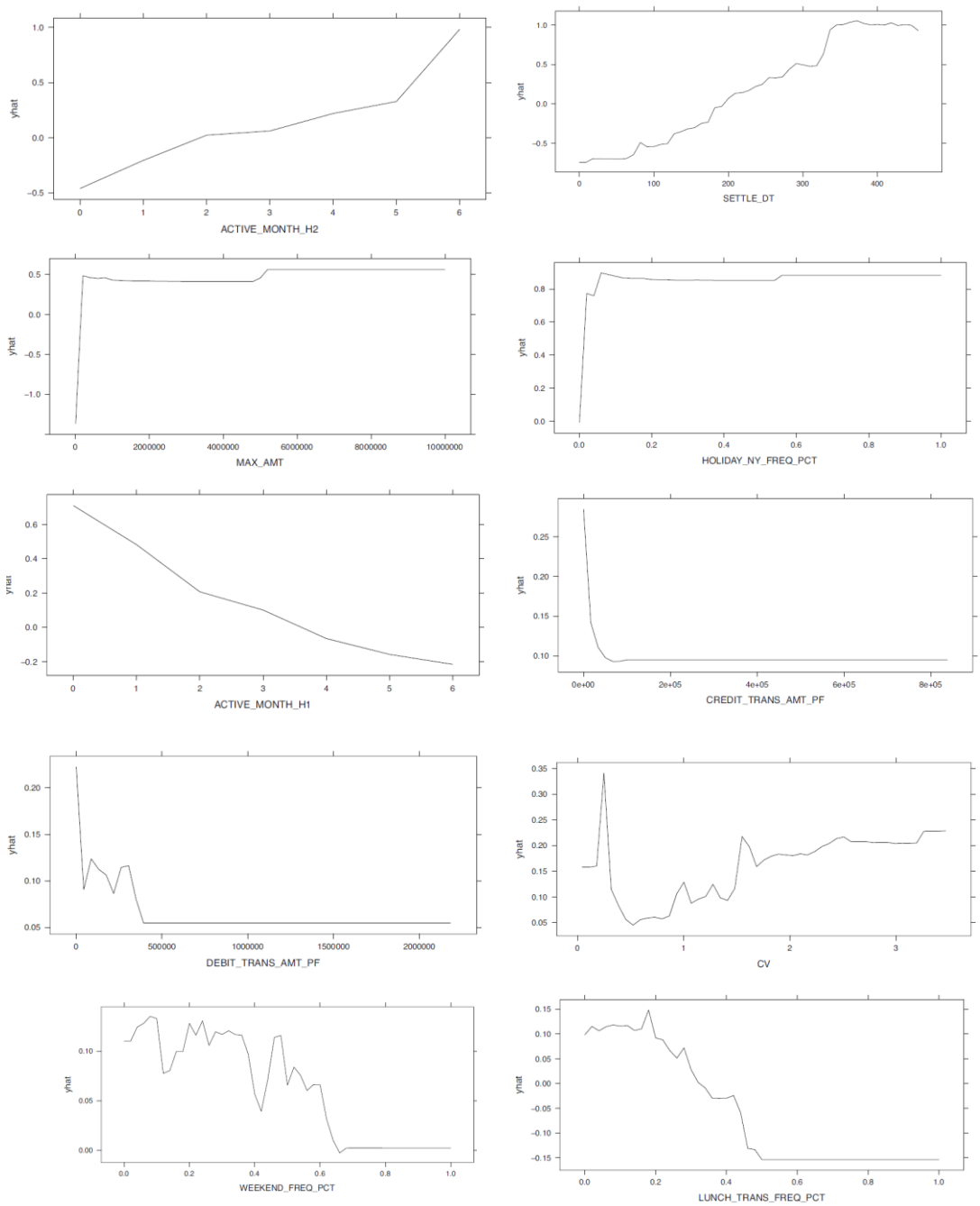
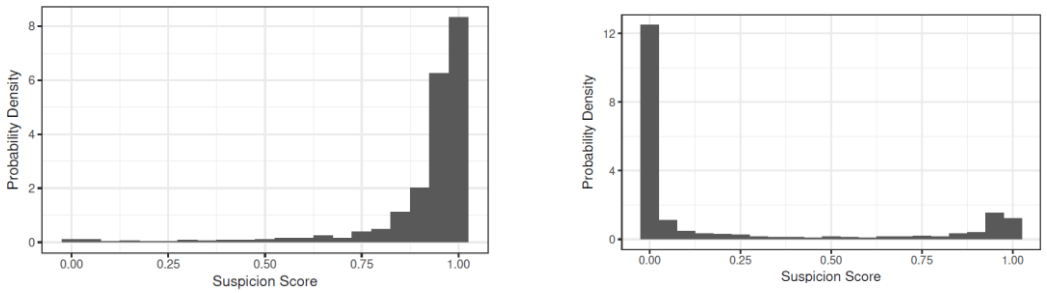


Figure 3: Top 10 Important Variables Partial Dependence Plots on XGBoost



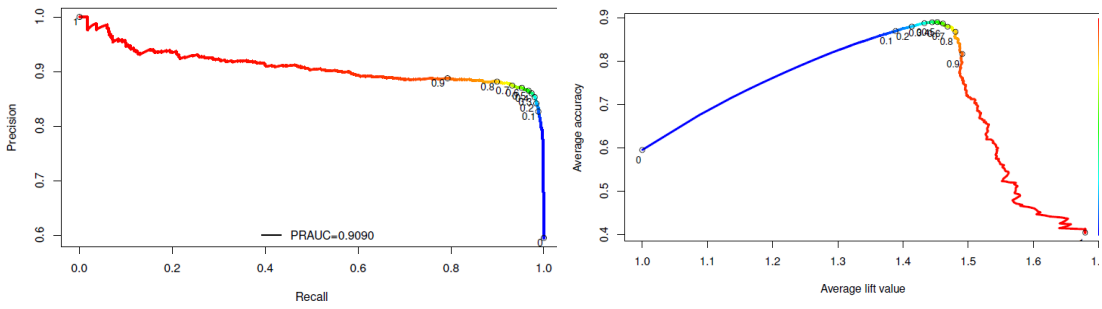
(a) Scores of MCC misuse (b) Scores of no MCC misuse

Figure 4: Average testing scores density distribution histogram of XGBoost.

5.3 Finding a score threshold

Furthermore, in XGBoost model, the average testing scores density distribution histogram for MCC misuse or not are shown in Fig.4. The left one is the MCC misuse scores distribution where the right one is the scores distribution of no MCC misuse. Almost every MCC misuse merchants are close to value 1. However, for no MCC misuse merchants, there are also few merchants are scored to value 1. This is likely to be a consequence of the most of the variables statistical units are in years may lose some detail.

Precision-recall (PR) curves have been used as an alternative performance measure to ROC. Fig.5a. The XGBoost model performance very well on both ROC curve PR curve (with AUC at 0.9095 and PRAUC at 0.9090 respectively). Fig.5b provides the relationship of average lift value versus average accuracy. Both Fig.5a and Fig.5b are labelled and colorized the threshold of the XGBoost model. The threshold between 0.1 to 0.9 are located very close which means that this model is very robust and the median value 0.5 can be the best threshold of the model. Compared to other models, our XGBoost model does not require too much human intervention (such as variable standardization, discretization) but can achieve very good results. Also, the distribution on Fig.4 point out that the scored distribution are towards to both side with very few ambiguous results. This XGBoost classifier seems to yield very good results in increasing the automation level and having a high rate of recall ratio. These values meet all the goals of the banks and card organizations.



(a) PR curve (b) Lift-Accuracy curve

Figure 5: PR and Lift-Accuracy curve of XGBoost on testing data

5.4 Deployment

In order to ensure that the classifier does not degenerate, it is important to update the data for training. The system will update the new data for training monthly. The classification service runs monthly by services application programming interface (API). XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It can runs on major distributed environment (Hadoop, SGE, MPI) and can solve problems beyond billions of data. XGBoost not only provide native interfaces for C++, R, Python, Julia and Java but also on Hadoop, Spark and Fink with GPU accelerated. Fig.6 illustrates the main parts of detection system on Spark. Every month, after the transaction data has been updated, the system will do ETL steps to obtain the training samples for the past one year. The application seamlessly embeds XGBoost into the processing pipeline and exchange data with other Spark-based processing phase through Spark's distributed memory layer. Also, the threshold and parameters can be optimized by current situation if needed. The suspicion score dashboard can alarm which merchant is occurring misuse the MCC.

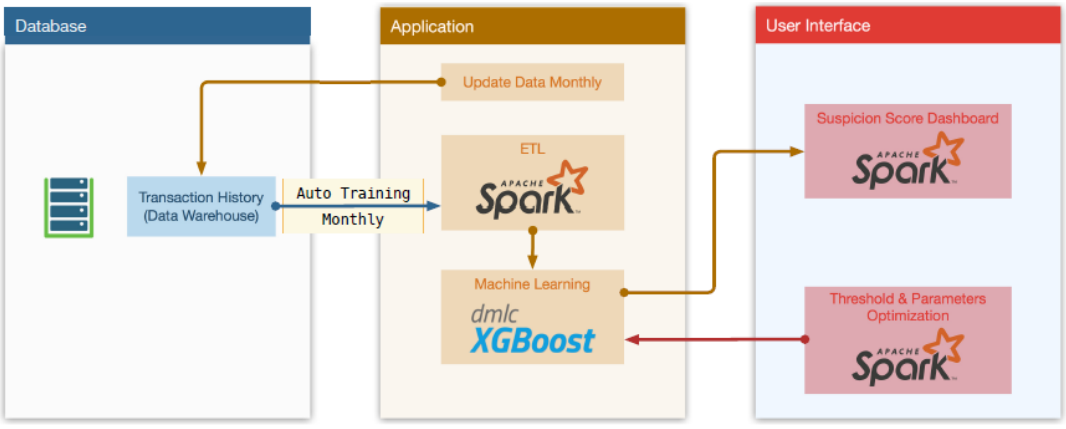


Figure 6: Diagram of the MCC misuse detection system.

6 Conclusions and future work

These ensemble models all suggest that time-based variables are very important in predicting MCC misuse among merchants. Such as the activities in months and days, all of them have a big impact on MCC misuse. The trend characteristics of transactions have significant differences in terms of MCC misuse. However, the average transaction amount does not have significant difference on credit card and debit card transactions. In this data, the merchants belong to general department store and wholesale categories. The general department store include travel, ticketing, department store, medical, alcoholic, tobacco, general service, professional service, hotel, restaurant, entertainment, estate etc., which always have lower discount rates than merchants like public welfare which sometimes will replace their POS terminal with that of a general merchant MCC. A real general merchant usually does not have large amounts per transaction, so CV is not very high. Also, this kind of merchant always operates at weekends (such as travel and retail stores). This is consistent with the results of Zhang (2015). Variables about number of customers do not have significant relationships to MCC misuse, which implies that merchants MCC misuse depends more on transaction amount and frequency rather than the number grouped by customer and card.

Our results indicated that after feature selection, the use of fewer variables can allow more accurate predictions than using all available variables. Banks and card organizations always have more dimensions of the merchant information, but effective feature selection is very important for prediction purposes. Many studies suggest that the ensemble model feature selection will always have better performance than other feature selection models such as Least Absolute Shrinkage and Selection Operator (LASSO), Principal Component Analysis (PCA), Information Value (IV), etc..

This study gives a new view of credit risk detection systems, in which merchants with incorrect MCC codes affect other scoring systems reliability and can incur losses for banks and card organizations. Ensemble models help us to distinguish the MCC misuse merchants. Although traditional models are single models, in this study XGBoost and LightGBM ran much faster than SVM. AdaBoost. M1 and RF were still very slow because of their almost-continuous variables; in each of the trees they need to separate into several intervals.

Compared with traditional credit scoring systems, the current study was successful in providing novel ensemble approaches that can predict merchant MCC misuse more accurately and conveniently. Generally speaking, the performance of the ensemble algorithms was very similar in predicting MCC misuse. The XGBoost performed best of the models. It should be note that the LightGBM is a very fresh algorithm of GBM which can be more faster and higher accuracy than traditional GBM model, but now the package in R do not have more parameters to tune. A more comprehensive evaluation of these algorithms is needed to come to a final conclusion. It should also be noted that, in the current study, most of the variables are counted in years, so certain details may be lost. Compared with penalized variable selection Logistic model in Zhang (2015), the accuracy is much higher than traditional models (0.6212 v.s. 0.8905). Although the Logistic model have a better interpretation and faster calculation efficiency than machine learning models. The XGBoost and LightGBM model can take into account both interpretation and speed, and to achieve higher prediction accuracy.

There are some limitations to this study. First, no other potentially-important factors were included, such as merchant type, merchant location, and other merchant information factors. Second, the prediction error on no MCC misuse was approximately 0.25. The reason for this is twofold: (i) most of the variables statistical units are in years and some detail may be lost; and (ii) in reality, the proportion of no MCC misuse is much higher than is represented in this data. This also results in a maximum of around 400 days for SETTLE_DT. The data may be omitting some of the merchants that have been settled for a long time without MCC misuse. Thus, the prediction error increases.

In conclusion, we successfully applied ensemble machine learning approaches to identify MCC misuse merchants for the construction of a risk score model. Our results suggested that several ensemble models achieved AUC over 0.9000 in MCC misuse. These models can assist card organizations and banks to improve their credit scoring systems in the future.

References

[1] D. McCoy, H. Dharmdasani, C. Kreibich, G. M. Voelker, S. Savage, Price- less: The role of payments in abuse-advertised goods, in: Proceedings of the 2012 ACM conference on Computer and communications security, ACM, 2012, pp. 845–856.

[2] B. Boding, N. Wood, M. McGirr, Acquirer facing fraud management system and method, uS Patent App. 14/292,684 (Dec. 4 2014).

[3] A. Keramati, N. Yousefi, A proposed classification of data mining techniques in credit scoring, in: the Proceeding of 2011 International Conference of Industrial Engineering and Operations Management, Kuala Lumpur, Malaysia, 2011, pp. 22–4.

[4] M. Stone, R. Shaw, Database marketing for competitive advantage, Long Range Planning 20 (2) (1987) 12–20.

[5] D. J. Hand, W. E. Henley, Statistical classification methods in consumer credit scoring: a review, Journal of the Royal Statistical Society: Series A(Statistics in Society) 160 (3) (1997) 523–541.

[6] L. J. Mester, et al., What's the point of credit scoring?, Business review 3 (Sep/Oct) (1997) 3–16.

[7] C. Erdem, Factors affecting the probability of credit card default and the intention of card use in turkey, International Research Journal of Finance and Economics 18 (August) (2008) 159–171.

[8] M. Schreiner, Scoring arrears at a microlender in bolivia, ESR Review 6 (2) (2004) 65.

[9] K. A. Carow, M. E. Staten, Debit, credit, or cash: survey evidence on gasoline purchases, Journal of Economics and Business 51 (5) (1999) 409–421.

[10] E. K. Laitinen, Predicting a corporate credit analyst's risk estimate by logistic and linear models, International Review of Financial Analysis 8 (2) (1999) 97–121.

[11] H. Wang, Q. Xu, L. Zhou, Large unbalanced credit scoring using lasso logistic regression ensemble, PloS one 10 (2) (2015) e0117844.

[12] T. Bellotti, J. Crook, Support vector machines for credit scoring and discovery of significant features, Expert Systems with Applications 36 (2) (2009) 3302–3308.

[13] B. Baesens, R. Setiono, C. Mues, J. Vanthienen, Using neural network rule extraction and decision tables for credit-risk evaluation, Management science 49 (3) (2003) 312–329.

[14] J. Kruppa, A. Schwarz, G. Arminger, A. Ziegler, Consumer credit risk: Individual probability estimates using machine learning, Expert Systems with Applications 40 (13) (2013) 5125–5131.

[15] Y. G. Şahin, E. Duman, Detecting credit card fraud by decision trees and support vector machines.

[16] M. F. A. Gadi, X. Wang, A. P. do Lago, Credit card fraud detection with artificial immune system, in: International Conference on Artificial Immune Systems, Springer, 2008, pp. 119–131.

[17] S. Bhattacharyya, S. Jha, K. Tharakunnel, J. C. Westland, Data mining for credit card fraud: A comparative study, Decision Support Systems 50 (3) (2011) 602–613.

[18] D. J. Hand, Mining personal banking data to detect fraud, in: Selected Contributions in Data Analysis and Classification, Springer, 2007, pp. 377– 386.

[19] Wikipedia, Merchant category code — the free encyclopedia, [Online; accessed 26-October-2016] (2016).

[20] H. Zhang, Comparative analysis of variable selection and its application in the merchant paraphrase risk assessment, Master's thesis (2015).

[21] R. Alliston, Method and system for detecting incorrect merchant code used with payment card transaction, uS Patent App. 10/116,870 (Apr. 5 2002).

[22] A. Adjaoute, Reducing false positives with transaction behaviour forecasting, uS Patent App. 14/521,386 (Oct. 22 2014).

[23] Y. Kültür, M. U. Çağlayan, A novel cardholder behaviour model for detecting credit card fraud, in: Application of Information and Communication Technologies (AICT), 2015 9th International Conference on, IEEE, 2015, pp. 148–152.

[24] Z. Zojaji, R. E. Atani, A. H. Monadjemi, et al., A survey of credit card fraud detection techniques: Data and technique oriented perspective, arXiv preprint arXiv:1611.06439.

[25] G. Wang, J. Ma, Study of corporate credit risk prediction based on integrating boosting and random subspace, Expert Systems with Applications 38 (11) (2011) 13871–13878.

[26] K. Fang, Y. Jiang, M. Song, Customer profitability forecasting using big data analytics: A case study of the insurance industry, Computers & Industrial Engineering 101 (2016) 554–564.

[27] A. Lemmens, C. Croux, Bagging and boosting classification trees to predict churn, Journal of Marketing Research 43 (2) (2006) 276–286.

[28] M. Ballings, D. Van den Poel, N. Hespeels, R. Gryp, Evaluating multiple classifiers for stock price direction prediction, Expert Systems with Applications 42 (20) (2015) 7046–7056

[29] A. Booth, E. Gerding, F. Mcgroarty, Automated trading with performance weighted random forests and seasonality, Expert Systems with Applications 41 (8) (2014) 3651–3661.

[30] A. Natekin, A. Knoll, Gradient boosting machines, a tutorial, Frontiers in neurorobotics 7 (2013) 21.

[31] S. Bakiev, Data mining techniques for predicting the survival of passengers of the titanic.

[32] R. Song, S. Chen, B. Deng, L. Li, extreme gradient boosting for identifying individual users across different digital devices, in: International Conference on Web-Age Information Management, Springer, 2016, pp. 43–54.

[33] D. Petrov, Y. Dodonova, L. Zhukov, M. Belyaev, Boosting connectome classification via combination of geometric and topological normalizations, in: Pattern Recognition in Neuroimaging (PRNI), 2016 International Work- shop on, IEEE, 2016, pp. 1–4.

[34] L. Breiman, Random forests, Machine learning 45 (1) (2001) 5–32.

[35] C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, A. Zeileis, Conditional variable importance for random forests, BMC bioinformatics 9 (1) (2008) 307.

[36] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, B. P.Feuston, Random forest: a classification and regression tool for compound classification and qsar modelling, Journal of chemical information and computer sciences 43 (6) (2003) 1947–1958.

[37] M. J. Kearns, L. G. Valiant, Learning Boolean formulae or finite automata is as hard as factoring, Harvard University, Center for Research in Computing Technology, Aiken Computation Laboratory, 1988.

[38] M. Kearns, L. Valiant, Cryptographic limitations on learning boolean formulae and finite automata, Journal of the ACM (JACM) 41 (1) (1994) 67–95.

[39] Y. Freund, R. E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, in: European conference on computational learning theory, Springer, 1995, pp. 23–37.

[40] Y. Sun, M. S. Kamel, Y. Wang, Boosting for learning multiple classes with imbalanced class distribution, in: Data Mining, 2006. ICDM'06. Sixth International Conference on, IEEE, 2006, pp. 592–602.

[41] T. Chen, T. He, Xgboost: extreme gradient boosting, R package version 0.4-2.

[42] M. R. Asia, Distributed machine learning toolkit, http://www.dmtk.io, accessed: 2017-02-01 (2016).

[43] Q. Meng, G. Ke, T. Wang, W. Chen, Q. Ye, Z.-M. Ma, T. Liu, A communication-

efficient parallel algorithm for decision tree, in: Advances in Neural Information Processing Systems, 2016, pp. 1271–1279.

[44] N. Carneiro, G. Figueira, M. Costa, A data mining based system for credit card fraud detection in e-tail, Decision Support Systems.

[45] E. U. Weber, S. Shafir, A.-R. Blais, Predicting risk sensitivity in humans and lower animals: risk as variance or coefficient of variation., Psychological review 111 (2) (2004) 430.

[46] N. Chen, B. Ribeiro, A. Chen, Financial credit risk assessment: a recent review, Artificial Intelligence Review 45 (1) (2016) 1–23.

[47] Y. Wang, Y. Li, W. Pu, K. Wen, Y. Y. Shugart, M. Xiong, L. Jin, Random bits forest: a strong classifier/regressor for big data, Scientific Reports 6.

[48] G. Valdes, J. M. Luna, E. Eaton, C. B. Simone, et al., Mediboost: a patient stratification tool for interpretable decision making in the era of precision medicine, Scientific Reports 6.

Table 5: Sample Characteristics of the Subgroups with Merchants

Variable Name	MCC_Misuse=TRUE	MCC_Misuse=FALSE	p-value
N_ISS_BANK_NM	17.29(14.01)	16.55(15.77)	<0.001
N_CUSTOMER_YEAR	919.63(8026.88)	673.09(11142.15)	0.0175
LUNCH_N_CUSTOMER	211.4(1958.66)	155.2(3017.84)	0.0402
DEBIT_N_CUSTOMER	606.36(5529.85)	533.36(14801.16)	0.5554
CREDIT_N_CUSTOMER	747.77(7329.36)	400.99(5932.51)	<0.001
N_CARD_YEAR	1354.14(12536.14)	934.35(18520.36)	0.0134
ACTIVE_MONTH_H1	4.29(2.31)	4.35(1.92)	0.0055
ACTIVE_MONTH_H2	5.43(1.13)	3.02(2.48)	<0.001
CV	1.05(0.6)	1.41(0.76)	<0.001
SETTLE_DT	367.77(105.16)	247.82(148.94)	<0.001
RANGE_AMT	793340.32(3997307.05)	659073.53(5061821.91)	0.0055
MAX_AMT	152316.91(414724.13)	107881.68(506026.06)	<0.001
MIN_AMT	104.6(523.79)	60.33(1239.93)	<0.001
TRANS_AMT	4385986.48(30864962.46)	3286175.11(40896404.14)	0.0043
TRANS_AMT_PER_CARD	14535.29(41781.76)	10904.33(60136.67)	<0.001
MORNING_AMT	1254847.06(11511389.26)	1085115.05(13552827.07)	0.2002
AFTERNOON_AMT	2540841.41(18193129.41)	1733038.5(19531807.68)	<0.001
EVENING_AMT	579147.09(3914843.58)	439634.4(10422182.97)	0.1097
NIGHT_AMT	11150.92(281983.76)	28387.16(1030376.57)	0.0412
WEEKDAY_AMT	3309979.68(23682659.16)	2534825.26(31158317.78)	0.0085
WEEKEND_AMT	1076006.81(7724543.26)	751349.85(10673097.52)	0.0011
LUNCH_TRANS_AMT	723241.17(5955350.36)	509794.76(6090720.35)	<0.001
DEBIT_TRANS_AMT	2625648.26(25113163.87)	2468372.46(36726779.04)	0.6411
CREDIT_TRANS_AMT	1760338.22(10044811.07)	817802.65(10942457.85)	<0.001
HOLIDAY_NY_AMT	16738.45(146220.6)	9568.92(188494.29)	<0.001
HOLIDAY_ND_AMT	55139.14(693695.7)	54324.24(780187.19)	0.9162
TRANS_FREQ	1864.84(18122.38)	1307.08(28356.78)	0.0297
TRANS_FREQ_PER_CARD	0.85(0.41)	0.67(0.97)	<0.001
MORNING_FREQ	390.18(3478.23)	304.24(7093.07)	0.1601

Table 5: Sample Characteristics of the Subgroups with Merchants

Variable Name	MCC_Misuse=TRUE	MCC_Misuse=FALSE	p-value
AFTERNOON_FREQ	850.45(8401.23)	582.97(12042.74)	0.0162
EVENING_FREQ	613.96(6755.33)	398.4(9080.49)	0.0115
NIGHT_FREQ	10.25(373.26)	21.47(649.83)	0.0511
WEEKDAY_FREQ	1225.5(11499.01)	900.69(20179.87)	0.0686
WEEKEND_FREQ	639.34(6714.7)	406.39(8318.86)	0.0036
LUNCH_TRANS_FREQ	244.04(2359.53)	190.79(3927.44)	0.1289
DEBIT_TRANS_FREQ	794.2(7636.15)	766.18(23001.14)	0.8831
CREDIT_TRANS_FREQ	1070.64(10956.87)	540.89(9028.48)	<0.001
HOLIDAY_NY_FREQ	9.96(110.46)	4.37(104.37)	<0.001
HOLIDAY_ND_FREQ	39.64(476.86)	28.19(506.69)	0.0259
TRANS_AMT_PF	12994.68(24037)	11716.45(45521.59)	0.0013
MORNING_AMT_PF	13362.82(31616.27)	11596.41(50229.78)	<0.001
AFTERNOON_AMT_PF	13043.73(24882.47)	11398.21(41831.64)	<0.001
EVENING_AMT_PF	11268.19(30480.3)	8628.97(47799.11)	<0.001
NIGHT_AMT_PF	1172.97(15838.91)	588.03(10482.21)	<0.001
WEEKDAY_AMT_PF	13393.34(25159.92)	11961.3(46308.15)	<0.001
WEEKEND_AMT_PF	11869.88(24106.17)	10228.77(46492.68)	<0.001
LUNCH_TRANS_AMT_PF	12685.29(30324.43)	10440.23(39896.47)	<0.001
DEBIT_TRANS_AMT_PF	11670.84(33149.38)	10396.81(50104.35)	0.0053
CREDIT_TRANS_AMT_PF	8499.53(13187.91)	5789.85(10889.23)	<0.001
HOLIDAY_NY_AMT_PF	4075.52(28434.16)	2541.85(60364.39)	0.0031