A stochastic ordering method with applications in fraud detection in Anti-Money Laundry system

Kai-Shiang Fan
Advisers: Prof. Ming-Hsuan Kang and Prof. Huei-Wen Teng
National Yang Ming Chiao Tung University

Abstract

Anti-Money Laundering (AML) refers to a set of policies, procedures, and technologies that prevents fraudulent transaction, and plays a critical role in financial systems. It is implemented within government systems and large financial institutions to monitor potentially fraudulent activity. When analyzing the fraudulent transactions, we observe that a majority of them turn out to have same characteristics. Therefore we use these observations to construct a stochastic ordering method, and due to the strict law to protect person privacy, we use synthetic data to implement our analysis. In our research, we provide a stochastic ordering method with high AUC and F1-score compared with Logistic Regression, Naive Bayes, Decision Tree, LightGBM and XGBoost. Furthermore, the time consuming of our method is also much better than Logistic Regression, Decision Tree, LightGBM and XGBoost.

Keywords: Fintech, anti-money laundering, machine learning, data mining, synthetic data

Contents

1. Introduction

With the rapid development of technology nowadays, a lot of problems have come to pass, while fraud is one that can not be neglected. Financial infrastructure does not remain indifferent to this development. Online transactions increased dramatically and brought convenience, but on the other hand, created opportunities for cyber thieves. In order to prevent this from happening, financial institutions employ anti-money laundering (AML) solutions. AML policies and procedures exist to aid financial institutions combat money laundering by stopping criminals from engaging in transactions to disguise the origins of funds connected to illegal activity [12]. Although AML policies cover a limited range of transactions and criminal behavior, their implications are far-reaching. There are research that apply data mining methods to deal with this issue [5, 9].

When analyzing fraudulent data, some researchers also monitor customer's activity to identify unusual behavior and detect potential money-laundering situations [10]. Moreover, some scientists introduce machine learning methods, aiming to solve these problems. One of these research aimed to provide a comprehensive survey of machine learning algorithms and some methods were applied to detect suspicious transactions [3]. Watkins et al. [17] provided an overview of the money laundering problem in the USA and overseas, furthermore exploring the use of various innovative data mining and artificial-intelligence-based solutions to assist financial investigators and enhanced law enforcement. [7] formulated an AML conceptual model by following [14] decision-making process model and developed a novel and open multi-agent AML system prototype based on the above-mentioned model.

Throughout the time, the importance of personal privacy increased, making real-transaction datasets even harder to obtain from banks. Therefore, researchers decided to use or create synthetic financial datasets in order to experiment on. Gaber et al. [6] proposes a synthetic data generator, while Ahmed et al. [1] discusses the lack of real data in the world and how synthetic data has been used to validate recent fraud detection techniques. There are pros

and cons using synthetic data: obtaining a dataset can be a lot easier, but whether synthetic data is accurate or not may be a question [2]. In other words, synthetic data facilitates the analysis of fraud detection in AML for the time-consuming procedure of applying an authorization to use a real data can be waived. Meanwhile, whether a synthetic data can actually represent a real data is still questionable.

While we analyze a synthetic dataset in the field of AML, but more precisely, our analysis is more about the part of binary classification in AML which is similar to default detection in credit scoring. Credit scoring is of great importance to minimize credit loss for financial institutions, while credit scoring models play an important role in this field. They predict financial risk to customer lending [11], aiding decision makings of the amount and whether or not to lend a specific customer. In order to predict the delinquency of a credit card holder, information of customers' are used. Repayment performances on other loans to demographic characteristics such as gender, age, and income all serve as variables for predictions [4]. Credit scoring models not only reduce the cost of credit analysis, but also enable faster credit decisions while allowing closer monitoring of existing accounts and collections simultaneously. Thomas [15] did surveys of statistical and operational research based techniques adopted in this field, while pointing out the need to incorporate economic conditions into credit scoring systems. Machine learning techniques have also been employed in the research of credit scoring, while customers' behavior can be classified into binary outcomes by their predicted default probability [8], which is similar to the detection of fraudulent behavior.

In the rest of this paper, we organize the article as follows. Section 2 reviews Naive Bayes and the stochastic ordering method. Section 3 describes the dataset. Section 4 provides our study plan and results. The last section concludes the whole research and gives our future work.

2. Our methodology

Let y denote the response variable (also called the dependent variable), which is a binary label indicating whether a fraud occurs, with $y = 1$ for a fraud and $y = 0$ otherwise. As a result, we focus on the binary classification. A problem instance to be classified, denoted by $x = (x_1, \ldots, x_p)$, representing p features (also called explanatory variables or independent variables). For simplicity, we assume that each x_i takes n_i possible outcome or classes $c_{i,j}$ for $j = 1, \ldots, n_i$.

2.1. Review on Naive Bayes

Naive Bayes assigns the posterior probability given an instance x:

$$p(y = k|x)$$

for $k = 1, 2$. Using Bayes' theorem, the conditional probability can be written as

$$p(y|x) = \frac{p(y)p(x|y)}{p(x)}.$$

The numerator is equivalent to the joint probability, $p(y, x_1, \ldots, x_p)$ which can be written using the chain rule for repeated applications of the definition of the conditional probability:

$$\begin{aligned} p(y, x_1, \ldots, x_p) &= p(x_1, \ldots, x_p, y) \\ &= p(x_1|x_2, \ldots, x_p, y)p(x_2, \ldots, x_p, y) \\ &= p(x_1|x_2, \ldots, x_p, y)p(x_2|x_3, \ldots, x_p, y)p(x_3, \ldots, x_p, y) \\ &= \cdots \\ &= p(x_1|x_2, \ldots, x_p, y)p(x_2|x_3, \ldots, x_p, y) \cdots p(x_{p-1}|x_p, y)p(x_p|y)p(y) \end{aligned}$$

Assuming that conditional on the category y, all features in x are mutually independent. we obtain

$$p(x_i|x_{i+1}, \ldots, x_p, y) = p(x_i|y).$$

Therefore, the joint model can be expressed as

$$\begin{aligned} p(y|x_1,\ldots,x_p) &= \frac{p(y,x_1,\ldots,x_p)}{p(x_1)p(x_2)\cdots p(x_p)} \\ &= \frac{p(y)p(x_1|y)p(x_2|y)\cdots}{p(x_1)p(x_2)\cdots p(x_p)} \\ &= \frac{p(y)\prod_{i=1}^{p} p(x_i|y)}{\prod_{i=1}^{p} p(x_i)}. \end{aligned}$$

With this formulation, the naive Bayes classifier assigns the value of y, denoted by $\hat{y}$, given an instance x, through the following equation:

$$\hat{y} = \arg\max_{k\in\{0,1\}} \frac{p(y=k)\prod_{i=1}^{p} p(x_i|y)}{\prod_{i=1}^{p} p(x_i)}.$$

2.2. The stochastic ordering method

We define the stochastic ordering score as follows:

$$\text{score}(y=1|x) = \prod_{i=1}^{p} p(y=1|x_i). \tag{1}$$

In practice, we estimate $p(y=1|x_i)$ by sample probability.

The above procedure assigns a score given an instance x. When predicting the y label, we first set up a threshold h. If the score is larger then h, we predict y to be 1. Otherwise, we predict y to be 0. The threshold is determined with a pilot study, by optimizing a preferred measure, such as the accuracy or F-1 score.

2.3. An Example

Table 1 gives a simple example to contrast the Naive Bayes and our stochastic ordering method. The probability given in Table 1 means x_1 and x_2 happen when the y-label is 0 and 1. Moreover, in Table 1, we assume that the probability of $x_1 = 0$ and $x_2 = 0$ is a, the probability of $x_1 = 0$ and $x_2 = 1$ is b, and so on.

Suppose we want to predict $x = (x_1, x_2) = (0, 0)$ with Naive Bayes. Let $p_0 = p(y = 0) = a + b + c + d$ and $p_1 = p(y = 1) = A + B + C + D$, then we

calculate $p(y=0|x)$ as follows:

$$\begin{aligned} p(y=0|x) &= p(y=0)p(x_1=0|y=0)p(x_2=0|y=0) \\ &= p_0 \times \frac{a+c}{p_0} \times \frac{a+b}{p_0} \times \frac{1}{p_0(p_0(a+c)+p_1(A+C))(p_0(a+b)+p_1(A+B))} \end{aligned}$$

Similarly, $p(y=1|x)$ equals

$$\begin{aligned} p(y=1|x) &= p(y=1)p(x_1=0|y=1)p(x_2=0|y=1) \\ &= p_1 \times \frac{A+C}{p_1} \times \frac{A+C}{p_1} \times \frac{1}{p_1(p_0(a+c)+p_1(A+C))(p_0(a+b)+p_1(A+B))} \end{aligned}$$

After that, we compare the two probability calculated from above, and choose the bigger one to be our prediction.

With the proposed stochastic ordering method and the sample $x=(x_1,x_2)=(0,0)$, we have the stochastic ordering score as

$$\begin{aligned} \text{score}(y=1|x) &= \prod_{i=1}^{2} p(y=1|x_i=c_{i,j}) = p(y=1|x_1=0)p(y=1|x_2=0) \\ &= \frac{A+C}{p_0(a+c)+p_1(A+C)} \times \frac{A+B}{p_0(a+b)+p_1(A+B)} \end{aligned}$$

Therefore, following the above rule, we estimate the conditional probability $p=P(y=1|x=x_i)$, for $i=1,\ldots,p$. Then, to decide whether y = 1 or y = 0 for a given x = x_0, we set a threshold q. If the final probability is higher than q, we predict y as 1. Otherwise, predicting y as 0.

2.4. Connections between the Naive Bayes and the Stochastic ordering method

The Naive Bayes and the stochastic ordering method has high similarity with each other. From the above example, we can observe that when the sample is $x=(x_1,x_2)=(0,0)$, the probability when $y=1$ given by Naive Bayes is

$$\frac{(A+C)(A+B)}{p_1(p_0(a+c)+p_1(A+C))(p_0(a+b)+p_1(A+B)))}$$

On the other hand, the probability when $y = 1$ given by stochastic ordering method is

$$\frac{(A+C)(A+B)}{(p_0(a+c)+p_1(A+C))(p_0(a+b)+p_1(A+B))}$$

Comparing these two probabilities, we can observe that the difference between them are just $\frac{1}{p_1}$. In general, the difference between them will be $(\frac{1}{p_1})^{p-1}$, where the p in the power $p-1$ indicates the number of features. Moreover, Naive Bayes sets its threshold to be 0.5, but we decide our threshold to maximize the F1-score which is a moving threshold method.

3. Data

The dataset we used in this research is an open dataset -*Synthetic Financial Datasets For Fraud Detection* [16], which can be found on Kaggle. This synthetic dataset is generated by a simulator called PaySim. PaySim uses aggregated data from some private dataset to generate a synthetic dataset which simulates real transactions including some malevolent behavior. In other words, PaySim simulates mobile money transactions based on real transactions which are extracted from one-month financial records of a mobile money service implemented in an African country. The original records were provided by a multinational company, which is the provider of the mobile financial service and is currently running in more than fourteen countries around the world. This synthetic dataset is scaled down one-fourth of the original dataset, containing $6,362,620$ observations with $8,231$ observations labeled as fraudulent. This making it an extremely unbalanced dataset as the fraud rate is only 0.13%. Furthermore, a binary variable is set up as the y-label, and the following 10 features in Table 2 are the explanatory variables. Table 3 below shows the information of the ten explanatory variables, including average, standard deviation, minimum, median, and maximum, and Table 4 shows the information of one categorical variable and the y-label. Note that the notation given in Table 2 will be used to replace the acronym of these ten features.

3.1. Data analysis

First of all, since the two features, $nameOrig$ and $nameDest$, both represent the customers' ID, we ignore both explanatory variables in our research. Afterwards, all features are further analyzed in order to separate them into numerical and categorical features. Since the input of our method has to be categorical, we convert each of the numerical features into 20 categories by their percentage. For example, data that falls between the range of 0 to 5 percent will be transformed into 0 as a categorical representation. After all features are transformed into categorical representations, further examinations of the dataset are made.

To start with, we observe the correlation of these 8 features by plotting a heatmap which is shown in Figure 1. From both graphs, we can see that two pairs of features, x_4, x_5 and x_6, x_7, have high correlation with each other for they represent an account's balance before and after a transaction. Furthermore, we can also see that x_1 and x_3 have a relatively high relationship with the y-label, indicating that the time when the transaction happens and the amount of money during a transaction effects the outcome of fraudulent behavior the most.

We also take a look at the fraud rate plots of each variable which are shown in figure 2. By doing so, we can see whether there is a characteristic in each variable when fraudulence happens.

From figure 2, it is easy to realize that the fraud rate in this data rises tremendously when some feature provides a certain value. Let's take feature x_2 for example. According to its scatter plot, we can know that when the type of a data is equal to 3, which means the transaction type is transfer, the probability of the data to be fraudulent is high compared with the other values in feature x_2.

3.2. Mutual information score

Information score is a measurement which is based on the amount of information a feature contains. To put it another way, it can also be regarded as a

criteria evaluating the strength of correlation between two clusters. Within our research, we calculate the information score between a feature and the y-label.

Mutual information is the quantity which is firstly defined by Claude Shannon in his remarkable paper, *A Mathematical Theory of Communication*[13]. However, he didn't call the value as mutual information at that time. The term, "mutual information", was created by Robert Fano later.

The mutual information of two jointly discrete random variables X and Y is calculated as a double sum:

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p_{(X,Y)}(x,y) log(\frac{p_{(X,Y)}(x,y)}{p_X(x) p_Y(y)}) \tag{2}$$

In order to compute the information score of all explanatory variables, all of them will have to be categorized into two different groups depending the type of value of the data. If the explanatory variable is categorical, we use Eq. (2) directly. However, if the explanatory variable is numerical, we use its sample mean or median to divide the value into groups. Those bigger than the mean or median of the feature will fall into a group, while those smaller than the mean, on the other hand, will fall into another one. When the preprocess is done, we can compute mutual information score of each data by using the above-mentioned formula, with X as one of a feature and Y as the feature "y". By doing so, the mutual information score represents the mount of a feature X's information sharing with the y-label.

Table 5 shows the mutual information score for each explanatory variables. We can see that *step* and *type* provide higher information score than others, indicating that the time when transaction happens and the type of a transaction have a large impact on the occurrence of fraudulence.

4. Empirical analysis

4.1. Study plan

Owing to the data analysis aforementioned, we construct our experiment as shown in Figure 3. First of all, we separate 10 features into two categories, numerical and categorical. We then convert all the numerical features into categorical representation using the method mentioned above. Then, we split the whole data into 80% for training and 20% for testing. As we want to use the characteristic of fraud ratio in each feature, we calculate the fraud ratio of each value under all features. After finishing this process, we obtain the probabilities that fraud happens for each of the 7 features a data has according to the value the data obtains in a specific feature. With the 7 probabilities just gained, we multiply them together to get the final probability which will be later used to predict the y-label. Yet setting a threshold is necessary to get the outcome, so a precision-recall-curve is plotted in order to find out which threshold will maximize F1_score.

4.2. Results

We consider 80%/20% training-testing data and record the computing time (in seconds), accuracy, precision, recall, F1-score, and AUC in Table 6.The fitting time means the time used to calculate the model and fit the model to data, the predictint time means the time used to calculate the estimated or predicted probability for both training data and testing data, the threshold finding time is the time used to find the threshold, and the calculating time is used to calculate the other measures, including accuracy, precision, recall, F1-score, and AUC. By examining Table 6, we can observe that our method, namely stochastic ordering, performs much better than Logistic Regression, Naive Bayes, Decision Tree and LightGBM, in terms of AUC. Futhermore, although LightGBM provides the highest F1_score, our method's F1-score is just a bit lower than that of LightGBM, which still outstands all the other models. Additionally, even though both our method's AUC and F1-score are just barely higher than XGBoost's,

it is obvious that XGBoost is 10 times more time-consuming than our method. Therefore, we find out that our method not only outperforms in terms of AUC and F1_score but also increases the efficiency significantly.

5. Conclusion

5.1. Summary

In our research, we provide a stochastic ordering method based on the characteristic of each data's fraud rate. From the analysis of the dataset, we can realize that the fraud rate of features under certain values will be higher than that of other values obviously. Taking advantage of that, we decide to multiply all fraud rates of each feature accordingly within a data to obtain a final probability, getting the predicted y-label by a certain threshold. In conclusion, our method provides the best result in terms of AUC, and above all, significantly increases the efficiency of the model.

5.2. Future work

In this research, we only execute the training experiment once. In the future, we aim to achieve a more precise result in training by considering a 5-fold cross-validation. Moreover, since we obtain our result by using our method just once, we look forward to improving it by imitating the concept of XGBoost.

References

[1] Ahmed, M., A. N. Mahmood, and M. R. Islam (2016). A survey of anomaly detection techniques in financial domain. *Future Generation Computer Systems 55*, 278–288.

[2] Cahill, M. H., D. Lambert, J. C. Pinheiro, and D. X. Sun (2002). Detecting fraud in the real world. In *Handbook of Massive Data Sets*, pp. 911–929. Springer.

[3] Chen, Z., E. N. Teoh, A. Nazir, E. K. Karuppiah, K. S. Lam, et al. (2018). Machine learning techniques for anti-money laundering (AML) solutions in suspicious transaction detection: a review. *Knowledge and Information Systems 57*(2), 245–285.

[4] Crook, J. N., D. B. Edelman, and L. C. Thomas (2007). Recent developments in consumer credit risk assessment. *European Journal of Operational Research 183*(3), 1447–1465.

[5] Fawcett, T. and F. Provost (1997). Adaptive fraud detection. *Data Mining and Knowledge Discovery 1*(3), 291–316.

[6] Gaber, C., B. Hemery, M. Achemlal, M. Pasquet, and P. Urien (2013). Synthetic logs generator for fraud detection in mobile transfer services. In *2013 International Conference on Collaboration Technologies and Systems (CTS)*, pp. 174–179. IEEE.

[7] Gao, S. and D. Xu (2009). Conceptual modeling and development of an intelligent agent-assisted decision support system for anti-money laundering. *Expert Systems with Applications 36*(2), 1493–1504.

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

[9] Khac, N. A. L. and M.-T. Kechadi (2010). Application of data mining for anti-money laundering detection: A case study. In *2010 IEEE International Conference on Data Mining Workshops*, pp. 577–584. IEEE.

[10] Kingdon, J. (2004). AI fights money laundering. *IEEE Intelligent Systems 19*(3), 87–89.

[11] Li, X.-L. and Y. Zhong (2012). An overview of personal credit scoring: techniques and future work. *Scientific Research Publishing*.

[12] Masciandaro, D. (1999). Money laundering: the economics of regulation. *European Journal of Law and Economics 7*(3), 225–240.

[13] Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal 27*(3), 379–423.

[14] Simon, H. A. (1960). The new science of management decision.

[15] Thomas, L. C. (2000). A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International Journal of Forecasting 16*(2), 149–172.

[16] Wang, S. and X. Yao (2012). Multiclass imbalance problems: Analysis and potential solutions. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 42*(4), 1119–1130.

[17] Watkins, R. C., K. M. Reynolds, R. Demara, M. Georgiopoulos, A. Gonzalez, and R. Eaglin (2003). Tracking dirty proceeds: exploring data mining technologies as tools to investigate money laundering. *Police Practice and Research 4*(2), 163–178.

Table 1: The joint probability of (x_1, x_2, y)

$y = 0$	$x_1 = 0$	$x_1 = 1$
$x_2 = 0$	a	b
$x_2 = 1$	c	d
$y = 1$	$x_1 = 0$	$x_1 = 1$
$x_2 = 0$	A	B
$x_2 = 1$	C	D

Table 2: Data Descriptions

Notation	Acronym	Description	Type
x_1	step	Maps a unit of time in the real world. Each step represents an hour	Numerical
x_2	type	Transaction type	Categorical
x_3	amount	Amount of the transaction in local currency	Numerical
x_4	oldbalanceOrg	Initial balance before the transaction	Numerical
x_5	newbalanceOrig	Customer's balance after the transaction	Numerical
x_6	oldbalanceDest	Initial recipient balance before the transaction	Numerical
x_7	newbalanceDest	Recipient's balance after the transaction	Numerical
y	isFraud	Identifies a fraudulent transaction	Categorical

Table 3: Summary statistics for numerical features

Acronym	Average	SD	min	median	max
step	243.40	142.33	1	239	743
amount	179861.9	603858.2	0	74871.94	92445516.64
oldbalanceOrg	833883.1	2888243	0	14208	59585040.37
newbalanceOrig	855113.7	2924049	0	0	49585040.37
oldbalanceDest	1100702	3399180	0	132705.7	356015889.35
newbalanceDest	1224996	3674129	0	214661.4	356179278.92

Table 4: Summary statistics for categorical features

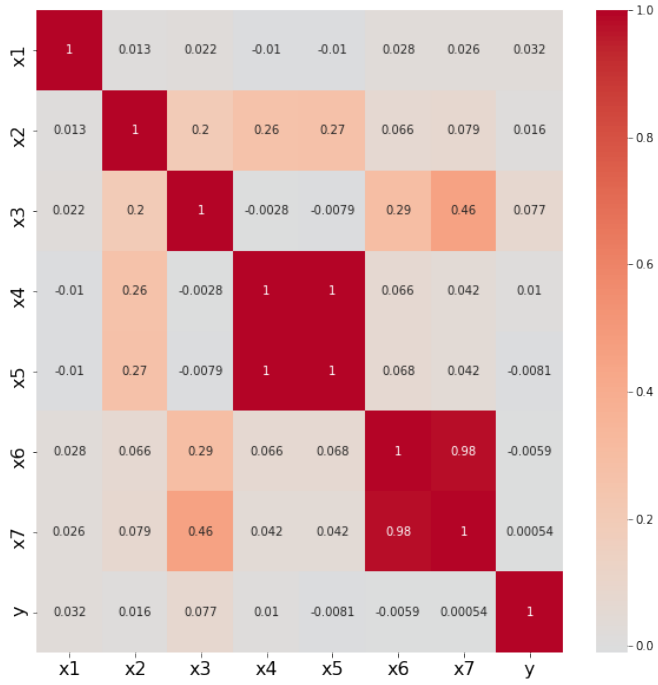
Acronym	Value Descriptions
type	CASH-OUT = 0, PAYMENT = 1, CASH-IN = 2, TRANSFER = 3 and DEBIT = 4
isFraud	0 = normal data, 1 = fraud data

Table 5: Information score

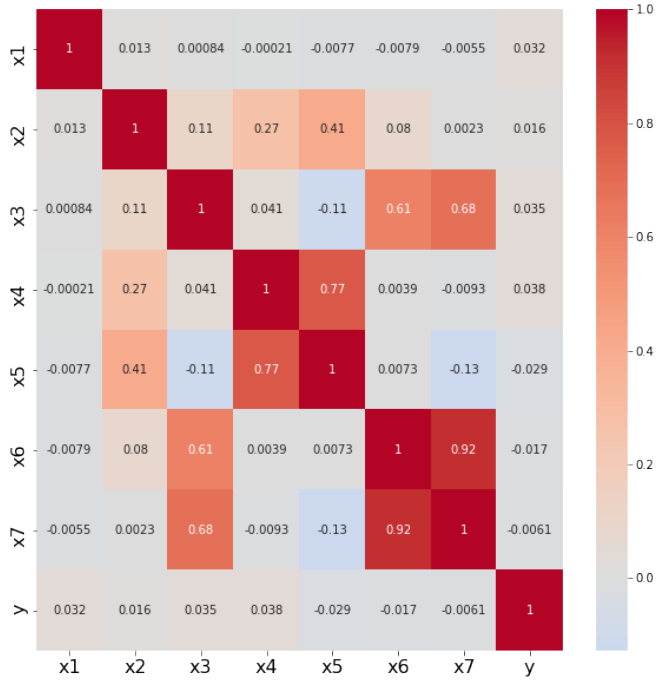
Acronym	Original	Mean	Median	20-category
step	0.004689			
type	0.001380			
amount		0.000412	0.000304	0.001068
oldbalanceOrg		0.000260	0.000650	0.001095
newbalanceOrig		0.000099	0.000615	0.000641
oldbalanceDest		0.000069	0.000161	0.000170
newbalanceDest		0.000000	0.000017	0.000053

Table 6: Result

	Stochastic Ordering	Logistic	Naive Bayes	Decision Tree	LGBM	XGBoost
Train						
Fitting time	14.63	43.80	0.97	30.45	42.861	330.84
Predicting time	0.14	0.28	0.98	0.30	1.18	1.16
Threshold time	0.71	1.36	0	0	0.71	0.67
calculating time	5.49	5.91	5.29	5.13	5.42	6.47
Accuracy	0.9993	0.9991	0.9927	0.9991	0.9995	0.9995
Precision	0.8404	0.8358	0.0337	0.7695	0.8958	0.9592
Recall	0.5768	0.3671	0.1677	0.4520	0.7026	0.6729
F1-score	0.6840	0.5101	0.0561	0.5695	0.7875	0.7910
AUC	0.9899	0.8875	0.5807	0.7259	0.8512	0.9826
Test						
Predicting time	0.04	0.06	0.22	0.07	0.24	0.46
Calculating time	1.24	1.47	1.44	1.39	1.40	1.71
Accuracy	0.9993	0.9991	0.9928	0.9991	0.9995	0.9995
Precision	0.8346	0.8293	0.0361	0.7608	0.8943	0.9538
Recall	0.5537	0.3793	0.1793	0.4634	0.7116	0.6665
F1-score	0.6657	0.5202	0.0600	0.5760	0.7925	0.7846
AUC	0.9893	0.8901	0.5003	0.7316	0.8557	0.9826



Raw data



Transformed data

Figure 1: Heatmaps

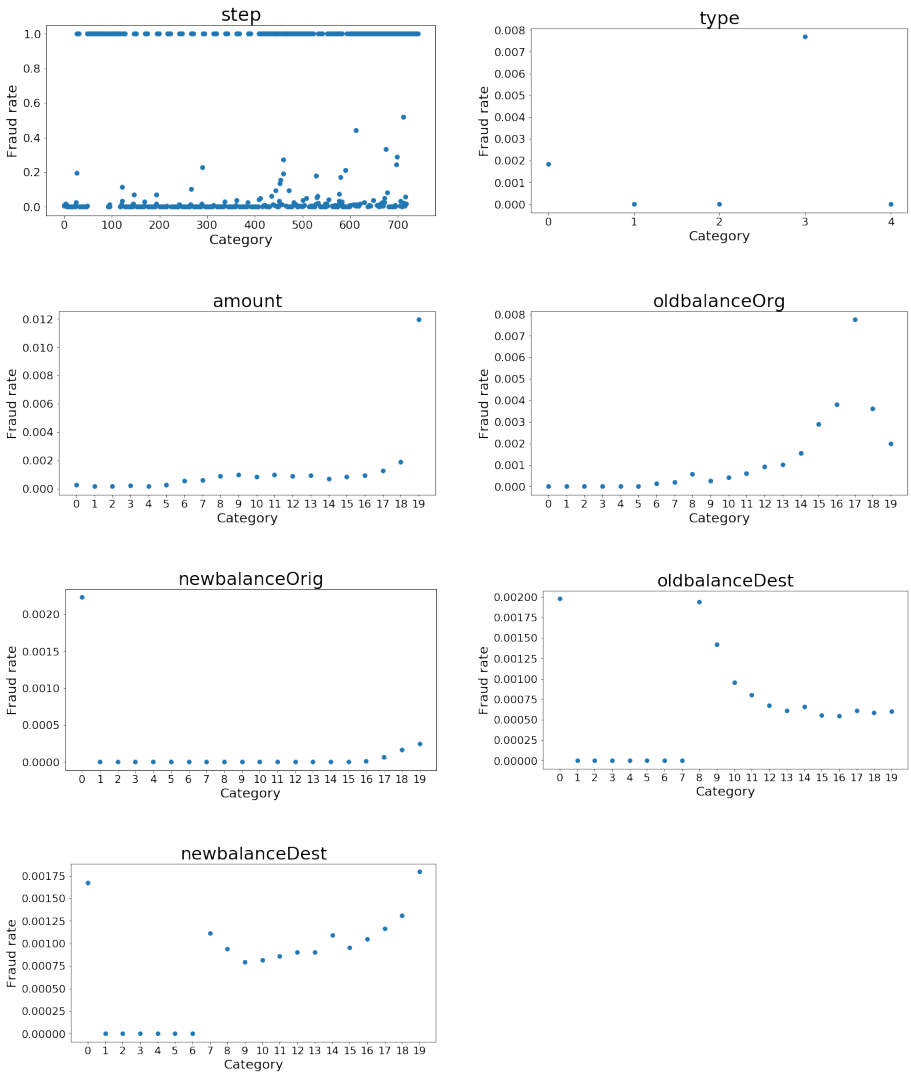


Figure 2: Fraud rate of each explanatory variable

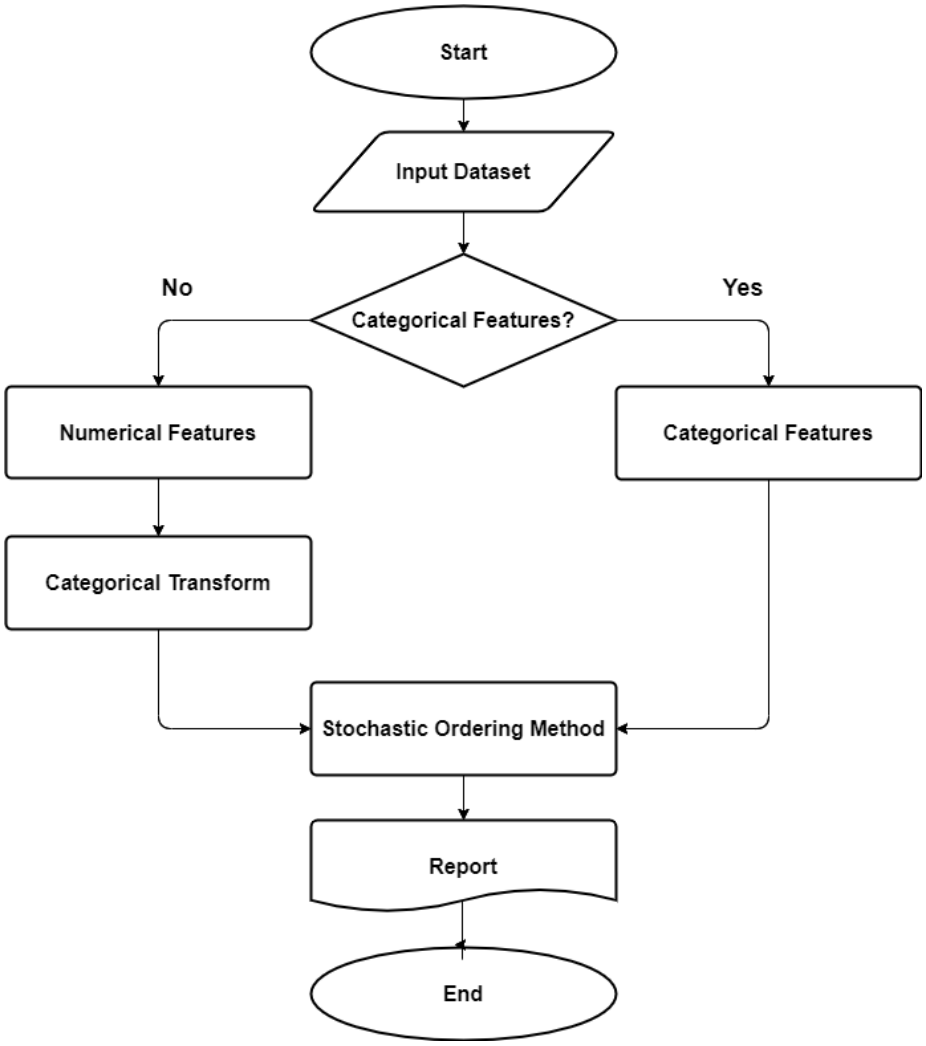


Figure 3: Experiment Process