BANK PERSONNEL FRAUD DETECTION

EKREM DUMAN

Department of Industrial Engineering, Ozyegin University, Istanbul, Turkey
E-mail: ekrem.duman@ozyegin.edu.tr

Abstract - For building successful predictive models one should have enough number of examples for the class to be predicted (the positive class). When the number of examples of the positive class is very small, building strong predictive models becomes a very challenging task. In this study we pick up one such problem: predicting the bank personnel which might commit fraud (stealing money from customer accounts). For this problem, in order to have a strong enough predictive model, we decided to combine the powers of descriptive and predictive modeling techniques where we developed several descriptive models and used them as an input of a predictive model at the last stage. The results show that our solution approach perform quite well.

Keywords - Personnel Fraud, Predictive Modeling, Banking

I. INTRODUCTION

Altough it is very difficult to commit fraud and there will be big punishments when discovered, occasionally some bank personnel steal money from bank customer accounts [1]. These fraudsters know the banking software very well, are usually very successful (or, they seem so) at their job and commit their fraud little by little throughout time. Such cases are observed once upon a couple of years and when a case is resolved it can be realized that, the fraud case have been taking place for some years, many fraudulent transactions took place and the total loss reached up to a few million dollars. In the whole bunch of bank transactions statistically only a few of them are fraudulent but the economic losses can be very big. The problem is similar to finding a needle in a haystack. This problem is one of a classification problems where predictive machine learning algorithms [2] could be used (here the classes are fraudulent and legitimate). However, for the predictive algorithms to have a good enough learning, one should have enough number of examples from both classes. Unfortunately, in our case the number of examples from the fraudulent class is much less than the other which makes them negligible statistically. In other words, we have a highly imbalanced data. There are some techniques to handle the class imbalance problem like under-sampling the majority class or oversampling the minority class (using the same example multiple times) however, in our case even these techniques might not help. There is another set of techniques in machine learning which is named as descriptive machine learning (some name this unsupervised learning as opposed to the name of supervised learning of the above). This set includes techniques like clustering, anomaly detection, and association analysis. Anomaly detection is actually based on clustering where similar examples are grouped as clusters and the distance of each example from the centroid of its cluster is used in the calculation of the anomaly index [3-5]. In our approach to solve the personnel fraud detection problem, we thought we can use both types of machine learning and combine their powers. We calculated the anomaly index of the customers involved in a money transfer, and the corresponding personnel and then used these indexes in the predictive model together with the other information related to the transactions. We also made use of some hyper-variables like RFM (recency-frequency-monetary) to increase the power of the predictive model. As far as we know, this is a novel approach in literature and the results of the model have been found to be very useful by the bank inspectors. The structure of the paper is as follows. In the next section we provide a more detailed description of the problem. In Section 3, we explain the solution approach together with some background about the techniques used in this study. This is followed by some discussion on the model obtained and numerical performance analysis. The paper is concluded by summary and results in section 5.

II. PROBLEM DEFINITION

This study is inspired by a real life necessity in a big size Turkish bank. Similar to other banks in the country and perhaps in the majority of the world banks, not all bank employees are honest and they can steal money whenever they find an opportunity. Lokanan [6] describes the so called fraud triangle to describe how personnel frauds happen (Figure 1).



Figure 1. Fraud Triangle

At the first corner of this triangle we have pressure/motivation. The pressure may either be related with the high business targets that are difficult to reach or the financial difficulties that the employee might have. Here in this study we are concerned with the second reason more; the first reason could result in improper banking operations and these are tried to be determined by in the internal control departments of banks. The second corner of the triangle states opportunity which means the employee is looking for an opportunity, or a weak point, in the banking system. Typically the fraudsters are experienced users of the banking software, know the system very well and have been searching for some week points of the system. And again typically, they are successful employees in general. Then, the third corner of the triangle describes rationalization. After, having intentions to commit fraud, the employee tries to justify himself/herself why he or she can commit fraud. He/she starts thinking like this: "Well, I am very good personnel here, I cause my bank make a lot of money. However, the salary they pay me is very small and actually I deserve more money". After all three corners are visited, the employee is ready to commit fraud.

Now lets continue with what can an employee fraud look like and how fraudster employees typically behave. First, we should note that, due to the credibility and trustability concerns banks do not share the details of the fraud cases they were supposed to. Even, we do not know the full details of the 15 fraud cases in the bank we made this project (we were just given the list of fraudulent transactions). However, we can share a case as it is not possible to live such a case anymore: Once in a branch there was CR (customer representative) who was very successful in collecting money (time deposit accounts) and he has a very loyal list of customers. His customers preferred to talk to only him even when there were other CRs in the branch and he was not there. The reason of loyalty to him was that he had been giving considerably higher interest rates to deposits of his customers and he was saying to his customers to come to him again and get higher interest again at the time of renewal. He continued as such for two-three years and suddenly disappeared. Only after that his magic was discovered: he was entering the high interest rate at the time deposit opening screen, take a printout of the screen, sign and stamp the sheet and hand it over to customer, but then exit the screen using escape button without actually opening the account. To handle the renewals, he kind of created a parallel manual accounting system and the customers always heard the correct amount of money is lying there in their account until that day when the employee is disappeared with several millions of customer money. Preventing this case to take place again was easy, they have technically made the print screen function impossible from the terminals at the branches.

From this and other similar cases we can start to think on what parameters (variables) can be useful in estimating whether a transaction is a part of a fraud case. The business experts state that the fraudsters are typically successful at their job as we can also see from the above example case. As they are doing something illegal, they do not want to go on holidays especially for long times, because when they are away the risk of being caught is higher. In our bank, later they brought the rule that each employee should go for a minimum of two weeks holiday every year and during the holiday he or she will not have access to emails and personnel accounts.

Another common feature of fraudster employees is that they do considerable amount of overtime (even if not paid for that). This can be expected since they might need more time to make alternative plans or calculations for the illegal things they did during the day.

After long discussions with the business experts we made a list of all variables that can be useful in identifying personnel frauds and give it to the IT department for their coding. The list contained 69 variables about the transaction details, 310 variables about the customer related with the transaction, 420 variables about the personnel performing the transaction and, 45 variables about the financial and non-financial (relative, neighbour etc.) relationship between the sender and receiver customers of the transaction.

III. SOLUTION APPROACH

A general overview of our solution approach is given in Figure 2. At the center of the figure (and the solution approach) we have the predictive model (shown by cone) trained by the past fraudulent and legitimate transactions (this model is then used to score all transactions in daily operations). The inputs of this predictive model are some detailed information about the transactions and the outputs of four intermediate models.

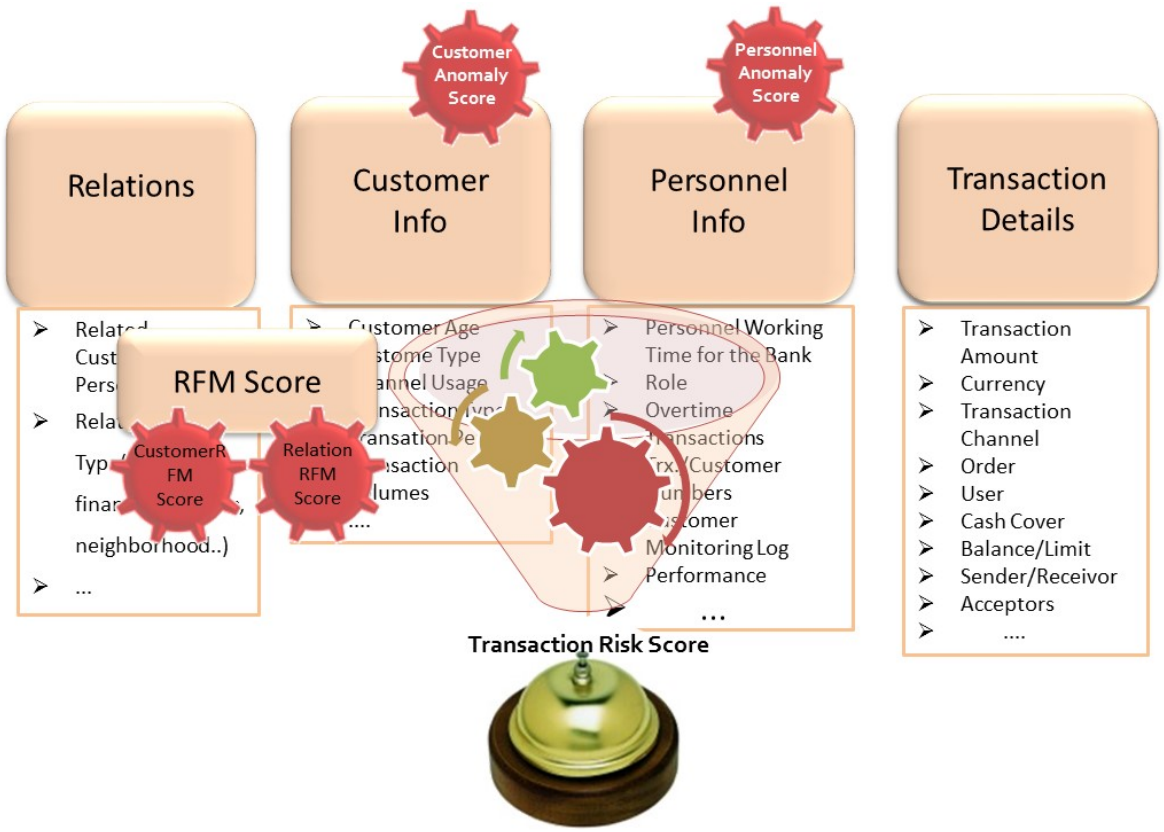


Figure 2. Overview of Our Solution Approach

Two of the intermediate scores are customer anomaly score and personnel anomaly score which are obtained by an anomaly detection algorithm. Anomaly detection algorithms first cluster the set of records which are also called peer groups and then measure the distance of each record from the centroid of its cluster. This distance, after normalization is called the anomaly index or anomaly score (refs). Naturally, some customers or some personnel can behave quite different than the others because of the strength of the relation with the bank or because of the set of responsibilities of the personnel and this is why similar customers or personnel are grouped together first. More information on the version of anomaly detection algorithm we implemented can be found in https://www.ibm.com/developerworks/community/.../AlgorithmsGuide.pdf.

The other two intermediate scores are the RFM (recency, frequency, monetary) scores. The first RFM score is about the customers: for each customer we identify how many days ago the last transaction took place (recency), on the average how many transactions are being made per month during the last months (frequency) and the average amount per transaction (monetary) are calculated and these three are combined with particular weights. This score is expected to be useful to identify those customers whose characteristics change. For example, customers with dormant accounts making a sudden transaction can be caught by this RFM score. The other RFM score measures the strength of the financial relation between the sender and receiver of the transaction. Non-financial relationship if there is any (relative, neighbour, children at the same school, etc.) is followed up by an independent variable and it is one of the input variables of the central predictive model.

IV. RESULTS

Since there can be great differences between the behaviours of retail and commercial customers and between active and passive customers, we made four groups of customers and applied anomaly detection algorithms on these separately. Similarly we grouped the personnel into three (retail and commercial customer representatives and the tellers) and developed different anomaly detection models for them.

In our training set obtained from five recent years there were 15 different fraud cases (stories). 46 different customers and 22 personnel were involved in these 15 cases. Although anomaly detection algorithms are descriptive methods and they are normally not used for prediction, we wanted to see if high anomaly index customers are more suspicious (Figure 3).

	Low Scores	Top Risky
Normal	265.899	2.670
Fraud	30	16

Figure 3. Customer Anomaly Confusion Matrix

Figure 3 displays the resulting confusion matrix when customers with top 1% highest anomaly score are marked as fraud victims. As we can see, if one can inspect top risky customers he would be able to identify 16 out 46, or 35%, of fraud victims. This information is useful itself. On the other hand, the personnel anomaly score did not show a similar performance and thus we did not tabulate its confusion matrix here.

Risky	Target	Hits	Hit %	Capture %	Lift
1%	839	219	26	62	62
5%	4,197	340	8	97	19
10%	8,395	351	4	100	10
TOTAL	83,950	351			

Figure 4. Predictive Model Performance Evaluation

The performance of the predictive model is shown in Figure 4. To obtain this figure, we have inserted the full list of 351 fraudulent transactions into the non-batch transactions of a randomly selected day and compared the scores of frauds and legitimates (we are told that there will be no frauds in batch processes such as automated bill payments). As we can see, top 1% risky transactions involved 65% of all frauds and when we go down to top 10% no fraud is missed. We can say that this model performs quite well.

SUMMARY AND CONCLUSIONS

In this study we have undertaken the problem of indetifying the bank personnel who are stealing money from customer accounts. This is a very difficult classification problem since the number of positive examples (frauds) is very very small as compared to normal transactions. For the solution of this problem we have developed a novel approach where we have combined the powers of descriptive and predictive data mining techniques. The performance of the model turned out to be very good so that if inspectors focus in only top 1% risky transactions they can identify most of the fraduluent transactions.

ACKNOWLEDGEMENTS

The authors would like to thank everyone, just everyone!

REFERENCES

[1] Y. Sahin, E. Duman, “An overview of business domains where fraud can take place, and a survey of various fraud detection techniques,” Proceedings of the 1st International Symposium on Computing in Science and Engineering, Aydin, Turkey, June, 2010.

[2] E. Duman, Y. Ekinci, A. Tanriverdi, "Comparing alternative classifiers for database marketing: the case of imbalanced datasets," Expert Systems with Applications, vol. 39, pp. 48-53, 2012.

[3] M. Ahmed, A. N. Mahmood, Md. R. Islam, “A survey of anomaly detection algorithms in financial domain”, Future Generation Computer Systems, vol. 55, pp. 278-288, 2016.

[4] J. Bi, T. Feng, H. Yuan, “Real-time and short term anomaly detection in GWAC light curves”, Computers in Industry, vol. 97, pp. 76-84, 2018.

[5] A. Ramchandran, A. K. Sangaiah, “Chapter-11 Unsupervised anomaly detection for high dimensional data – an exploratory analysis”, in Computational Intelligence for Multimedia Big Data on the Cloud with Engineering Applications. Academic Press, 2018, pp. 233-251, 2018.

[6] M. E. Lokanan, “Challenges to fraud triangle: questions on its usefulness”, Accouting Forum, vol. 39, pp. 201-224, 2015.

[7] IBM SPSS Modeler 17 Algorithms Guide, 2015, pp. 3-8, https://www.ibm.com/developerworks/community/.../AlgorithmsGuide.pdf

★★★