DSA/ISE-5103

Project Report (Draft)

Daniel Casares and Felipe Olivera

Due December 4, 2017

Detecting fraud on credit card transactions

The University of Oklahoma

Executive Summary (1 page)

– Concise problem statement

– List of major concerns/assumptions (if any)

– Summary of findings

– Recommendations

Problem description and background (including related literature for problem, solution techniques, etc.)

The use of plastic cards (i.e. credit and debit cards) as a payment method has grown significantly over past years, unfortunately so has fraud (Bahnson 134). Plastic card fraud is defined as an unauthorized account activity committed by means of the debit and credit facilities of a legitimate account. Some successful fraud tactics observed in the industry are lost and stolen card fraud, counterfeit card fraud, card not present fraud, mail non-receipt card fraud, account takeover fraud and application fraud (Krivko 6070). Based on the latest figures gathered in 2015, card fraud accumulated $21.84 Billion worldwide in losses (The Nilson Report 6). When banks lose money due to credit card fraud, the losses partially are passed to customers through higher interest rates, higher membership fees and reduced benefits. Hence, it is both the banks' and cardholders' interest to reduce illegitimate use of credit cards (Maes 2).

In this work, we consider the problem of identifying whether a credit or debit card account has been subject to fraudulent activity, using real-life transaction data from a Latin American credit card processing company. The goal is to construct a supervised learning model that can detect fraud on new (previously unseen) plastic card transactions. Fraud detection is, given a set of credit card transactions, the process of identifying those transactions that are fraudulent. Thus, the transactions are classified as genuine or as fraudulent transactions (Maes 2). Different detection systems that are based on machine learning techniques have been successfully used for this problem, in particular: neural networks, bayesian learning, artificial immune systems, association rules, hybrid models, support vector machines, peer group analysis, decision tree techniques such as ID3, C4.5, and random forest, discriminant analysis, social network analysis and logistic regression (Bahnson 135, Mahmoudi 2510).

Exploratory data analysis (the highlights; not the kitchen sink)

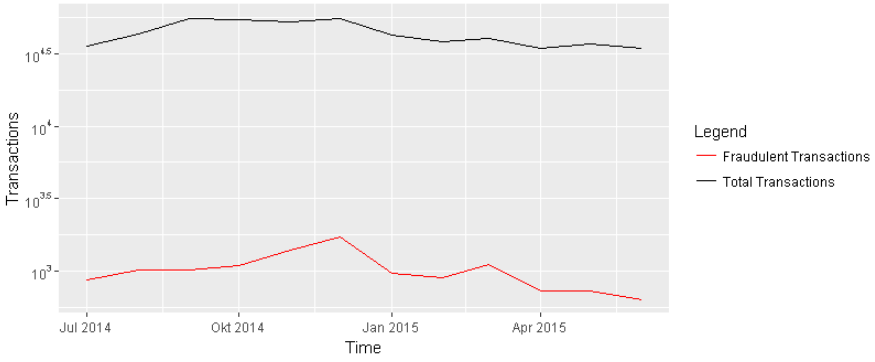
For this project we used a dataset provided by a Latin American card processing company. The dataset consists of fraudulent and legitimate transactions made with debit and credit cards between July 2014 and June 2015. The total dataset contains 41,091,288 individual transactions, each one with 13 attributes (as shown in the table below), including a fraud label indicating whenever a transaction is identified as fraud. This label was created internally in the card processing company, and can be regarded as highly accurate. In the dataset only 12,632 transactions were labeled as fraud, leading to a fraud ratio of 0.031%.

Attribute name	**Description**
Datetime	Date and time of the transaction
Tokenized_PAN	Identification of the credit card
Is_Upscale	Indicates if the card holder is an upscale customer
Pos_entry_mode	i.e. Chip and PIN, magnetic strip, etc.
Amount	Amount of the transaction in USD
Id_mcc	Identification of the merchant type
Mcc_group	Merchant group indentification
Country_code	Country of the transaction
Id_issuer	Issuer of the card
Is_fraud	

Kommentar [DC1]: Aca faltan algunos atributos y podes dar un ejemplo para issuer oft he card?

Due to the low proportion of the target class (i.e. frauds) in the given dataset, the class imbalance problem arises. Classification of imbalanced data is difficult because standard classifiers are driven by accuracy, thus the minority class may simply be ignored (Visa 67). Generally all classifiers present some performance loss when the data is unbalanced (Prati 253). Additionally, many imbalanced datasets experience problems related to its intrinsic characteristics, such as lack of density and information. To illustrate these issues, a dataset containing of 5 : 95 minority-majority examples and a dataset of 50 : 950 are compared. Though the imbalance factor is the same as in both datasets in the first case the minority class is poorly represented and suffers more from the lack of information factor than in the second case. In order to reduce these problems in our modeling, a smaller subset of transactions with a higher fraud ratio is selected from the original data. This *new* dataset contains 523,049 transactions and a fraud ratio of 2.33%. In this dataset, the total financial losses due to fraud are 1,876,697 USD. It was selected considering all the fraudulent transactions in the original dataset, in addition to all the legitimate transactions for the corresponding customers. Next, transactions for some customers that have never been victims of fraud were added.

From plotting the amount of fraudulent and total transactions over time we can see that the proportion of fraudulent transactions varies over time.



Analysis plan

Explanation of modeling choice – Why choose this technique? Strengths, weaknesses? Dani

We investigated the performance of three techniques in predicting fraud: Logistic Regression (LR), Support Vector Machines (SVM), and Random Forest (RF).

Which modelling approaches did you chose? Logistic regression and trees

Feature selection, engineering, missing value, outlier plan (felipe)

The raw data contained on each transaction detail some characteristics of it: amount, date and time, merchant type (e.g. gas station), entry mode, among others (as stated above). Just with those attributes, fraud may be identified at the transaction level, however, as seen in (Bolton & Hand, 2001), a single transaction information is not enough to detect a fraudulent transaction since it leaves behind important information such as the customer spending behavior. In order to fulfill this problem, a better strategy is proposed by (Whitrow et al., 2008): to perform transaction aggregation.

The derivation of the aggregation features consists in grouping the transactions made during the last given number of hours by card number, followed by calculating the number of transactions and the total amount spent on those transactions. We processed those new attributes for time windows of 1 day, 2 days, 1 week and 30 days respectively, resulting in 8 new features for the model.

Validation plan (including how do your findings compare with others?) Dani

Results and validation of analysis

Conclusion

References

Bahnsen, Alejandro Correa, et al. “Feature engineering strategies for credit card fraud detection.” *Expert Systems with Applications*, vol. 51, 2016, pp. 134–142.

Krivko ,M. “A hybrid model for plastic card fraud detection systems.” *Expert Systems with Applications*, vol. 37, 2010, pp. 6070–6076.

Maes, Sam et al. “Credit Card Fraud Detection Using Bayesian and Neural Networks”. *Proceedings of NF*, 2002.

Mahmoudi, Nader, et al. “Detecting credit card fraud by Modified Fisher Discriminant Analysis.” *Expert Systems with Applications,* vol. 42, 2015, pp. 2510–2516.

“The Nilson Report.” David Robertson, 17 Oct. 2016, www.nilsonreport.com/upload/content_promo/The_Nilson_Report_10-17-2016.pdf. Accessed 02 Dec. 2017.

Prati, R.C. et al. “Class imbalance revisited: a new experimental setup to assess the performance of treatments methods.” *Knowledge and Information Systems*, vol. 45, 2015, pp. 247–270.

Visa, Sofia. “Issues in mining imbalanced data sets – a review paper.” Proc. 16th Midwest Artificial Intelligence and Cognitive Science Conference, 2005, pp. 67–73.

Appendix

– Data visualizations, tables, etc. which support the work, but are not of primary importance

– List of data transformations, missing value imputations, outlier treatment, etc.

– List of any important assumptions not otherwise included

– Important code excerpts or algorithms used / developed if any.