

Final Project Report

“Fraud detection”

Hevorhyan Sofiya

Zubyk Olha

January 2020

Lviv, Ukraine

### Problem Description and Importance

Nota Bene: The data and all information about the client are hidden in terms of NDA. Due to that purpose, all names, some features, relations, and industry are changed or transformed.

The ‘fraud detection’ problem is usually well-known in different business areas and is crucially important for large enterprises as well as for small firms. Typically, it deals with dishonest behavior or swindler users which are usually called frauds.

In our case, we have a large chain of petrol stations and stores which have a well-known bonus structure for their clients. Typically, when clients come to the station to buy petrol or products they can use their user card to earn bonus points for their purchase or to get a discount using earned ones. All those transactions are recorded and bounded to a specific user with a specific bonus card. That gives the client an incentive to stay with the store in the future, be loyal to it and also opens the door to customer analytics.

However, when a person without the store’s card shows up some cashiers act dishonestly and disloyal. They use their own cards to gain points for the client’s purchases and after dozens of such transactions can use earned points to get discounts. They are called ‘frauds’ in our case and represent a potential loss for the company and weakening of its reputation.

For that purpose, a dataset with approximately *9 million* transactions was given for analyzing the behavior of different users and detecting those cashiers that do not act properly. Also, there is some recorded information about already caught frauds and the history of transactions that were made with their user cards.

The goal is to detect users in the first dataset that have a suspiciously abnormal set of transactions.

### Data Investigation and Preprocessing

As was mentioned above, there are several datasets with transactions and they require lots of preprocessing steps before actual modeling.

First, there are two files *basket.csv* and *transaction.csv* that represent information about products bought and other details. The unique id that will identify each user is the combination of columns *TRN\_WALLET\_WAL* and *TRN\_WAL\_PERSON\_ID*.

Other files such as *fraud\_data.csv* and *fraud\_11\_07.xlsx* contain transactions of already detected frauds. The behavior of those transactions is reviewed and relabeled in order to use only those ones that have typically ‘fraud’ behavior and won’t confuse our model with the normal transactions in the future.

Each transaction contains information about time and place, products that were bought and the amount paid for those products. It also has unique ids of card and user, and the number of points that were earned and used for a discount.

We use such typical for preprocessing stage steps as changing format and type for different columns. But besides that, to analyze and model different behaviors we also need a unique set of features for each of our user which we extract from the whole set of transactions.

Feature engineering is the first and one of the most important steps before modeling. It helps to extract relevant information from our data and will distinguish frauds and normal users. To come up with the relevant features we, first of all, should know a lot about the business part and industry analyzed. The business task indicates that the types of consuming among different users is important - and for that purpose, we should know the main actions that are typical for frauds.

The main part of the logic behind feature engineering are insights that were given by people that have good experience in that industry and similar type of tasks. The understanding of these features is quite intuitive as long as they good describe the typical incentives of fraud cashiers.

First, they typically have a bigger amount of transactions because they `steal` other clients’ purchases and these purchases are usually made at one place because each cashier is bounded to the specific cashier machine.

The next step is industry analysis and how it can influence consumer behavior. As long as we deal with data gathered on petrol stations one of the best parameters to analyze is petrol. Because it is the typical product bought with the high price and most users do not need more than n liters per day/week, there are some features that will show a great difference between frauds and normal clients. For example, it makes sense to add a feature to each of our users that will show the amount of petrol bought, the number of types of petrol, the frequency of purchases and the information about places where these transactions occur.

The next product that efficiently illustrates the individual type of consumption is cigarettes. Their purchases can be strange and inconsistent in frauds as long as people usually stick to one or two marks while frauds don’t usually control the number of marks bought on their bonus card.

These and many other features constructed for each user will be the training set for our model. The main purpose of that stage is to come up with such features that will show fraud behavior the greater the difference the better.

The problem is also complicated by a big gap between the number of frauds and normal users so in modeling we should take into account huge class imbalance.

### Related Works

The existing works study a variety of cases where fraud behavior is present (such as credit card fraud, insurance fraud, tax form fraud and so on). Given the industry specifics, ways to detect fraud have been different(especially rule-based), but with artificial intelligence advances, the universal approaches became applicable.

A lot of researchers use supervised methods for fraud detection Among the most popular are Naive Bayes, SVM, and decision tree-based algorithms. To improve performance, bagging classifiers are built [1]- different algorithms are run on the resampled dataset, and then “voting” is performed by each algorithm to decide on the final label.

An intuitive approach would be to treat frauds as outliers, as there is a fewer number of them and they are behaving abnormally. However, it’s not sufficient to identify the single outliers from the dataset. The reason is that one fraud may have similar behavior to the other ones, and they will form a distinct cluster in the data, which the algorithm will perceive as a part of dataset distribution. In this case, unsupervised and semi-supervised learning comes in handy, such as clustering[2].

Gaussian Mixture Model is an especially common unsupervised algorithm for anomaly detection. It fits several normal(Gaussian) distributions to the dataset, with random initial means and covariances. Unlike k-means, GMM-clusters can be of different form. Covariance factor between the clusters is also taken into account. Overall, this is one of the best algorithms in unsupervised learning that can distinguish fraud clusters. Some researchers use it as a baseline model when comparing performance of deep-learning-based algorithm[3]

A more novel approach\* includes deep learning as a contrary to extensive feature generation. The new articles combine classifying, clustering and deep learning, creating brand-new algorithms that outperform the earlier approaches. Lang & Mettenmeier[4] use a 3-layer RNN and propose the input generating approach (sessions). Deng and Ruan[5] combine GAN and autoencoder to combine classifying and clustering and detect new fraud patterns on-the-run. Marchal and Sczyller[6] propose the recursive agglomerative clustering approach. An interesting thing that is common to 3 papers above is that they use the e-commerce data, the amount of which has increased in the last years with the increasing popularity of online shopping.

### Solution

There are different ways to work on this problem once we perform feature engineering. From now on, we have a relevant set of features for each of our users that describes his/her consumer behavior. Typically, these indicators are connected to the amount of petrol bought, its types, statistics, and frequency. In such cases, when coming up with a new feature we can see some clear differences between numbers in fraud dataset and data with normal users.

After that, several supervised approaches were used, such as logistic regression, decision tree, and random forest. They showed different results on the same test dataset.

* Logistic regression
  + automatically adjusted weights
  + GridSearch class weights

The first logistic model with automatically assigned weights shows probably the best results as long as it identifies almost all negative examples and gives a large amount (nearly 200) false positives. This is the set of users that can potentially be fraud cashiers and now their behavior can be investigated much closer - instead of checking all of 4 000 other users that were labeled as normal.

On the other hand, the model with class weights found by GridSearch performs differently. It gives only 5 false positives and does not identify all already caught frauds as ‘1’. Together with the previous model, it can give us some insights about users. First of all, it can be an indicator of which users should be checked in the first place. If we have a lack of resources, e.g. time, out of 208 false positives caught in the first model we can start checking those ones that are also caught by the second one - that indicated that they have very noticeable fraud behavior. Second, we can look at those users in the false-negative category - on their features and behavior. It will give us insights about some brand new features to come up with that will try to catch the behavior of those fraud group - which is probably omitted in the present case.

Altogether, two models working on the same task will give us different results based on which we can extend and improve our decisions in several ways.

* Decision Tree and Random Forest

Decision tree and random forest do not show such great results as logistic regression - they typically catch less false positives and often confuse already caught frauds. However, they are also a powerful instrument for analysis. Looking closely at their feature importance we not only can see some features at the top of both ‘lists’ but also analyze the information that they bring.

We can see that the number of petrol types bought at the main station is the relevant indication and that gives the incentive to add more information about the main station to our training features. The importance of the main station seems pretty logical as long as cashiers are bounded to the specific cash machine and place of work and lots of fraud transactions occur at the same place while normal users usually mix stations and shops. By analyzing information about the time difference between transactions we can dive deep into the main station parameters and add a new important feature.

* Unsupervised

Even though it is unlikely that the cluster with only fraud users can be formed, we would consider this method a success if more than 70% of frauds will be in one cluster - this narrows our dataset to only 1 cluster, and we can continue applying other algorithms on it. As all the users in one cluster will be similar in some way, it means that those kinds of frauds will also share some similar features. This makes easier to detect them, and even supervised learning could be done on a cluster.

However, we cannot fully rely on such a pipeline, as our model will be prone to the frauds with unseen before the behavior. So, occasional clustering should do the checking if any new clusters have emerged.

### Results and evaluation

As were mentioned above, we applied several approaches to our model. First, supervised methods show a variety of different results. With the best ‘accuracy’ performed logistic regression with automatical weights - it predicts negatives with the best accuracy and shows a significant amount of potential fraud. After that, LDA showed pretty good results when looking at the confusion matrix. Logistic regression with GridSearch class weights performed worse than two others but combined with another logistic regression it can be a good tool to analyze potential frauds and improve our models. Decision Tree and Random Forest showed the worst results.

* + DBSCAN

As a baseline unsupervised approach, we decided to use the DBSCAN clustering algorithm. We chose the value of eps(radius of expanding clusters) using GridSearch, based on the measure of unsimillarity between clusters(silhouette score). Resulting 3 clusters are spread unevenly - the biggest cluster has more than 90% of the user, while two others don’t exceed 2%. Of course, most of the frauds are located in the biggest cluster, and that doesn’t make our life any easier.

* + GMM

Number of clusters in GMM has to be given manually - we have chosen 4 based on BIC score. In this case, there were 2 clusters that have 90% of frauds, and the actual size of these clusters was just 50% of the data! So, only 2 clusters can be enough for the further development of the model.

### Code

Link to Github repository with:

* Jupyter notebook
* Final report
* saved models and files

Sources:

[1] Zareapoor, M., & Shamsolmoali, P. (2015). Application of Credit Card Fraud Detection: Based on Bagging Ensemble Classifier. Procedia Computer Science, 48, 679–685. DOI: 10.1016/j.procs.2015.04.201

[2] Vaishali, V. (2014). Fraud Detection in Credit Card by Clustering Approach. International Journal of Computer Applications, 98(3), 29–32. DOI: 10.5120/17164-7225

[3] M. Taniguchi, M. Haft, J. Hollmen and V. Tresp, "Fraud detection in communication networks using neural and probabilistic methods," *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181)*, Seattle, WA, USA, 1998, pp. 1241-1244 vol.2.

doi: 10.1109/ICASSP.1998.675496

[4] Lang, T., & Rettenmeier, M. (n.d.). Understanding Consumer Behavior with Recurrent Neural Networks.

[5] Deng, R., & Ruan, N. (n.d.). FraudJudger: Real-World Data-Oriented Fraud Detection on Digital Payment Platforms. Retrieved from <https://arxiv.org/pdf/1909.02398.pdf>

[6] Marchal, S., & Szyller, S. (2019). Detecting organized eCommerce fraud using scalable categorical clustering. Proceedings of the 35th Annual Computer Security Applications Conference on - ACSAC 19. DOI: 10.1145/3359789.3359810

\* Due to the lack of time, we have not implemented some of the approaches mentioned. After consulting with our mentor, we have decided to enhance the current algorithms that we had. We are going to try deep learning approaches in the next semester of our internship. Stay tuned!