Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 186 (2021) 721–726

Procedia Computer Science

www.elsevier.com/locate/procedia

14th International Symposium “Intelligent Systems”

Comparison of Poisson process and machine learning algorithms approach for credit card fraud detection

Anastasiia Izotova[a], Adel Valiullin[b,*]

[a]Tikhonov Moscow Institute of Electronics and Mathematics, National Research University Higher School of Economics,
34 Tallinskaya Street, Moscow, 123458, Russia
[b]Faculty of Computer Science, National Research University Higher School of Economics, 11 Pokrovsky blvd, Moscow, 109028, Russia

Abstract

This article describes the financial fraud detection in imbalanced data. We compare various approaches for credit card fraud detection problem. On the one hand, we use homogeneous and heterogeneous Poisson process to determine the probability of predicting fraud with the various intensity parametric functions. On the other hand, we solve classification problem using machine learning algorithms and different family of ensemble methods like boostings. The results of both methods are compared. The “false positive” problem is also discussed in the article.

Keywords: fraud detection; Poisson process; intensity model; machine learning; boosting ensembles; imbalanced data

1. Introduction

Currently, most banks provide their users with secure online services. One of the elements of such protection is a system for detecting and preventing fraudulent transactions.

In Russia, the share of non-cash expenses was 50.4% in the last quarter of 2019, and, according to analysts, the threshold of 50% will be overcome in the future. These figures indicate an increase in non-cash transactions, as well as an increase in fraudulent transactions. According to McAfee, cybercrime accounted for 0.8% of the world’s GDP in 2018, with the World Bank estimating annual losses to the global economy from cyberattacks at $445 billion. It is expected that there will be a significant increase in non-cash transactions due to the COVID-2019 pandemic and self-isolation measures taken by the end of the second quarter of 2020.

In this study, the fraud detection issue is solved by two approaches such as the Poisson process intensity model and supervised machine learning algorithms. Various unsupervised approaches is also used for that problem, Restricted Boltzmann Machine (RBM) and Generative Adversarial Networks (GAN) are highlighted [1].

[*] Corresponding author.
E-mail address: valiullin@hse.ru

1877-0509 © 2021 The Authors. Published by Elsevier B.V.

10.1016/j.procs.2021.04.214

Working with highly unbalanced sample significantly complicates the recognition of fraudulent transactions, so few samples of the minority learning class tend to be falsely detected by the classifiers. Several methods are created to solve this problem, for details see articles [2, 3].

2. Problem Statement

In this paper, we consider the problem of binary classification, exactly, the detection of fraudulent transactions with credit cards.

Let **X** be the m by n matrix of objects features, and **Y** be the m - dimensional vector of responses. The sample is presented as $\mathbf{D} = \{(x_1, y_1), ..., (x_m, y_m)\}$, where $x_i = (x_i^1, ..., x_i^n)$ is the feature vector of the i-th line and $0 \leq i \leq m$. After building the model, we get the algorithm $a(x) : \mathbf{X} \rightarrow \mathbf{Y}$. For the Poisson process, $a(x)$ is represented as an intensity using to predict responses $a(x_j)$. The percentage of correct responses is calculated as:

$$Q(a, X_l) = \frac{1}{l} \sum_{j=1}^{l} [a(x_j) = y_j],$$

where $l < m$. The task is to minimize this quality functionality for machine learning methods.

3. Poisson process

3.1. Mathematical concepts

Let the fraud time be $\tau \in [0, +\infty]$. Such a random variable is defined in a filtered probability space (Ω, F, Φ, P). Ω denotes the possible states of the world, F is the σ-algebra, $\Phi = (F_t)_{t\geq 0}$ is the filtration with F_t contains all information up to time t, and P is the probability measure describing the likelihood of certain events. Intuitively, one can determine whether or not the fraud time occurs before a certain deterministic time by observing the past up to time t, which is encoded in the filtration (F_t).

Now consider the sequence of time when fraudulent events occur as (τ_n) for $n \geq 0$. Denoting $X = \{X(\tau), \tau \geq 0\}$ for the counting process, we get:

$$X(t) = \sum_{n\geq 0} 1_{\{\tau_n \leq t\}}$$

In other words, $X(t)$ counts the number of fraud events between 0 and t. The trajectories of such process are piecewise constant, non-decreasing functions with jumps equal to one, see Fig. 1.

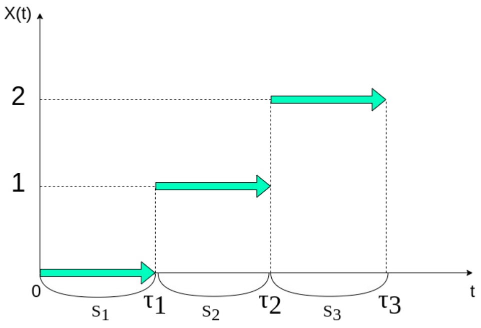

Fig.1 Poisson process trajectories

For more details see article[4]

3.2. Variation of the Poisson process

Definition. A counting process whose events occur with some intensity λ is called a Poisson process if:
1. $X(0) = 0$ with P(this event) = 1;
2. The process has independent increments;
3. For $0 \leq s < p, X(p) - X(s)$ is Poisson distributed with parameter $\int_s^p \lambda(x)dx$:

$$P(X(p) - X(s) = k) = \frac{e^{-\int_s^p \lambda(x)dx}(\int_s^p \lambda(x)dx)^k}{k!}$$

An equivalent definition is contained in article [5].

If the intensity is a deterministic function of time $\lambda = \lambda(t)$, such a process is called *non-homogeneous* Poisson process.

A Poisson process with an intensity that does not change over time and has both independent and stationary increments is called *homogeneous*.

For the homogeneous case, λ is constant; for the non-homogeneous case, $\lambda = \lambda(t)$. Note that the intensity always takes non-negative values. This should be taken into account when setting limits on the coefficients of the intensity function.

The waiting times between successive events, or arrivals follow an exponential distribution of parameter $\int_s^p \lambda(x)dx$. $S_1 = \tau_1$ and $S_i = \tau_i - \tau_{i-1}$ are the waiting times. Using MLE, the estimation of the homogeneous process intensity takes the form:

$$\hat{\lambda} = \frac{1}{\bar{S}},$$

To find an estimate of the intensity for non-homogeneous case, we need to determine the parameter that maximizes the logarithmic likelihood function. This function is derived from the joint density of fraudulent events:

$$l = logL(\lambda; \tau = \tau_1, \tau_2, \ldots, \tau_n) = \sum_{i=0}^{n} \log \lambda(\tau_i) - \int_0^T \lambda(x)dx$$

Let us know the information about the fraud events up to the moment T. The new transaction arrived at the time $T + \ delta$. The final formula for predicting the fraud of homogeneous and non-homogeneous Poisson processes:

$$P(a\, fraud\ occurs\ at\ t = T + \delta) = 1 - e^{-\int_T^{T+\delta} \lambda(x)dx}$$

The proof can be viewed [6].

4. Ensembles

In machine learning ensembles includes a set of several algorithms that are trained to solve the same problem. Thus, ensembles give a better forecasting efficiency than each algorithm individually contained in the ensemble. Several gradient boosting models are sequentially built using an anti-gradient. In the learning process, the models indicate the direction of future corrections in the predictions of the current ensemble model in order to reduce errors.

- LightGBM

This algorithm is an advanced implementation of gradient boosting. LGBM is used to increase the gradient with a tree-based training algorithm. The main difference from other algorithms is the growth of the tree in depth, or by leaves.

You should also pay attention to the name of this algorithm. ”Light” indicates a high speed of execution. LightGBM handles large amounts of data with the lowest memory requirements. The focus on prediction accuracy is another advantage of this algorithm. For a detailed description of hyperparameters see documentation[7].

- XGBoost

XGBoost is a machine learning algorithm based on the decision tree. It is an upgrade of the gradient boosting framework through system optimization and algorithm improvement.XGBoost has a wide range of applications: it can be used to solve regression, classification, ordering, and custom prediction problems. XGBoost provides a parallel tree boosting that solve many data science problems in a fast and accurate way [8].

- CatBoost

An open software library developed by Yandex that implements a unique patented algorithm for building machine learning models using one of the original gradient enhancement schemes [9]. Almost any modern gradient-based method works with numeric attributes. If the data set contains not only numerical, but also categorical signs, then we need to convert categorical signs to numerical, which leads to a potential decrease in the accuracy of the model. CatBoost is a gradient enhancement library whose main advantage is that it works equally well with two types of attributes.

5. Data preprocessing

The dataset consists of 95 662 transactions between November 15, 2018 and February 13, 2019. Transaction data belongs to 3 633 clients. All tuples of this set contain information (a label) about belonging to class 0 or 1. An important feature of the data set is a strong data imbalance: the percentage of fraudulent transactions is slightly less than 0.2%.

To predict the probability of an object belonging to a particular class using the Poisson process basics, you need three attributes: the client ID, the time when the transaction arrived, and its label.

For the Poisson model, the training and test sets must have the same clients with at least 2 transactions. Thus, 812 transactions were deleted from the entire data set, which is the same as 812 clients. Also, for this model, the solution to the ”False Positive answers” problem is to set $\lambda = 0$ for ”clean” clients.

A few statistics are provided based on the results of data processing:

- Number of transactions: 94 850;
- Number of clients: 2 821;
- Number of fraudulent transactions: 183;
- Percentage of fraudulent transactions: 0.19%;
- Number of fraudulent clients: 42;
- Maximum amount of fraud per client: 31;
- Minimum amount of fraud per client: 0.

Distribution of the clients number relative to the number of fraudulent transactions: (Fig. 2).

Splitting the data set into training and test parts for the Poisson and gradient boosting process: the first 80% of a specific client’s transactions make up the training set, the rest of the transactions fall into the test and train/test/validation as 70/15/15 accordingly.

6. Computational process

As a result of building models on training data and their operation on a test set, the probabilities of transactions belonging to a particular class are obtained. The ROC-AUC metric is used to evaluate the quality of these models.

In this research 3 models of the Poisson process are identified as HomoModel, LinearModel and QuadraticModel, which are Poisson process model with intensity like constant, linear function and quadratic function respectively.

Gradient boosting models such as LGBM, XGBoost and CatBoost are also presented.

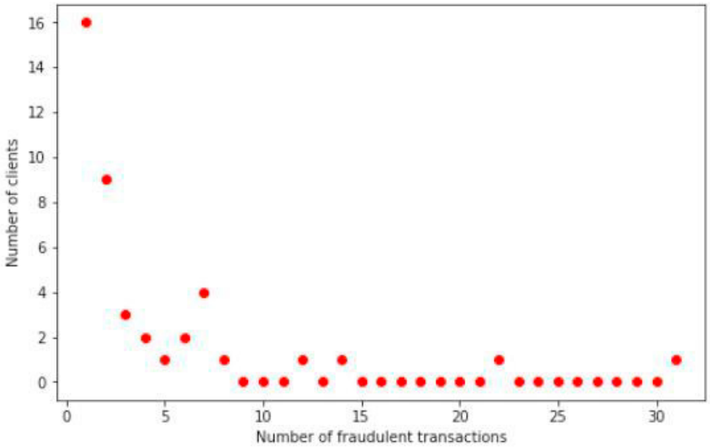

Fig.2 The distribution of fraudulent transactions

The values of the ROC-AUC quality metric or the area under the error curve are contained in Table 1. Note that gradient boosting models give a huge accuracy close to one. Models of the Poisson process showed accuracy lower but higher than random selection, which shows the predictive ability of the models.

Table 1. ROC-AUC metric.

Models	HomoModel	LinearModel	QuadraticModel	LGBM	XGBoost	CatBoost
AUC_test	0.748951	0.786327	0.769392	0.999074	0.999760	0.999693

Still, such models have an advantage: they require only three attributes, which clearly simplifies the computational process. And the time to build such models is much less than the time to learn gradient boosting.

For a visual comparison of the two approaches, two ROC-AUC curves are presented: the best result among the Poisson process models is highlighted in yellow, similar to gradient boosting models (pink). These curves are plotted in the coordinates True Positive Rate and False Positive Rate from point (0, 0) to (1, 1). From the example of the pink curve we can note that the curve is closer to point (0, 1), then the more accurate the model prediction is (Fig. 3).

7. Conclusion

To solve the problem of fraud detection, two approaches were applied: the Poisson process and machine learning. In the first case, we considered various intensity functions that are involved in predicting fraudulent events. Gradient boosters LightGBM, XGBoost, and CatBoost were used as machine learning algorithms.

Solutions to the problem of data imbalance, ”False positive answers” and the presence of ”clean” clients were also provided.

During the work, new data was processed and generated. To determine the estimate of the Poisson process intensity, it is sufficient to know the deterministic intensity function, the arrival time of the fraud transaction, and the label. Gradient boosting models used all the features of the updated data set.

In the future, it is planned to build Poisson process models using the sliding window method; in addition, more complex intensity functions for the non-homogeneous process, such as the Fourier series, will be considered. On the machine learning side, ensembles from the above ensembles will be added to the research. As a result of building such models with different approaches for detecting fraudulent transactions, it is assumed to combine the results, so the disadvantages of one algorithm will be compensated by the benefits of another[10]. Combined with machine learning, the approach of applying the Poisson process to financial datasets can lead to the most appropriate method for fraud detection.

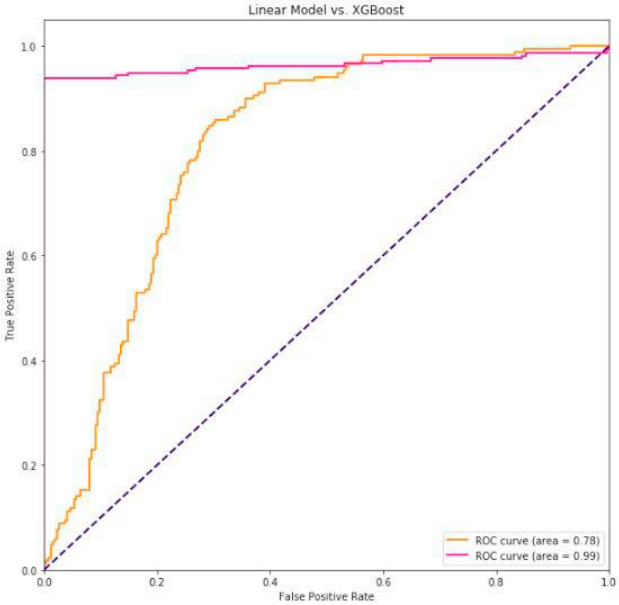

Fig.3 The best models' ROC-AUC curves

References

[1] Niu, X., Wang, L. and Yang X. (2019). *A comparison study of credit card fraud detection: Supervised versus Unsupervised*. [online] arXiv:1904.10604 [cs.LG]. Available at: https://arxiv.org/abs/1904.10604.

[2] Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5, pp.1–12.

[3] Wedge, R., Kanter, J.M. and Veeramachaneni, K. (2017). *Solving the "false positives" problem in fraud prediction*. [online] arXiv:1710.07709 [cs.AI]. Available at: https://arxiv.org/abs/1710.07709.

[4] Reiss, R.D. (1993). *A Course on Point Processes. 1st ed.*. Springer Series in Statistics.

[5] Drazek, L.C. (2013). Intensity estimation for poisson processes. Master Thesis, The University of Leeds.

[6] Ross, S.M. (2010). *Introduction to probability models. 10th ed.*. Boston: Academic press.

[7] LightGBM (2020). *Documentation LightGBM*. [online] LightGBM. Available at: https://lightgbm.readthedocs.io/en/latest/.

[8] XGBoost (2020). *Documentation XGBoost*. [online] XGBoost. Available at: https://xgboost.readthedocs.io/en/latest/.

[9] CatBoost (2020). *Documentation CatBoost*. [online] CatBoost. Available at: https://catboost.ai/.

[10] Zojaji, Z., Atani, R.E. and Monadjemi, A.H. (2016). *A survey of credit card fraud detection techniques: data and technique oriented perspective*. [online] arXiv:1611.06439 [cs.CR]. Available at: https://arxiv.org/abs/1611.06439.