Autoencoder and Logistic Regression: Algorithms for Fraud Detection and Their Results

Erekle Shishniashvili, Lizi Mamisashvili, Anano Turkiashvili

e-mail: ersho.ge@gmail.com

computer sciences: technical informatics, computer science, Tbilisi State University, University street. N13

email: mamisashvili.lizi@gmail.com

computer sciences: technical informatics, computer science, Tbilisi State University, University street. N13

e-mail: turkiashvilianano@gmail.com

computer sciences: technical informatics, computer science, Tbilisi State University, University street. N13

"Card Not Present" (CNP) transactions became popular in the 21st century. To explain, it is possible to make an order without the card physically present to the merchant. Along with the advantages of such a convenient system, there is a big problem adversely affecting bank consumers. According to the statistics of the retail company "Shift Credit Card Processing", in 2018, $24.26 Billion was fraudulently transitioned from the credit card holders' accounts (Credit Card Fraud Statistics, 2019).

A novel approach for protecting consumer's account information is connected to the intelligent systems differentiating legitimate transactions from the fraudulent ones. While the use of machine learning algorithms seems promising, there are particular challenges concerning fraudulent actions. First, the data where one can find many true positive (fraud) samples is scarce. Second, due to confidentiality, it is impossible to retrieve the data about the details of user information that might be helpful for training the machine learning model. Third, fraudsters mimic legitimate transactions so that it becomes harder to detect patterns in the data (Jurgovskya et al, 2018). Finally, fraud detection is a const-sensitive problem meaning that presence of false positives and false negatives have different effects on performance evaluation (Gomez et al, 2018).

Among the different supervised and unsupervised algorithms, there is a special type of feedforward artificial neural network model – autoencoder – well-known for its use in fraud detection problem. However, there is not a thorough experimentation done on its performance with the different hyperparameters or compared to the classical machine learning algorithms. In our project, we use the imbalanced data retrieved from the website – Kaggle – describing the three days of credit card transactions by European cardholders. Fraudulent samples comprise only 0.172% of the data. In our project, we build an autoencoder reducing 30 features to 10 on the latent layer. Additionally, we trained the logistic regression model using the results of the autoencoder. In this case, as we wanted the final output to be more accurate, the threshold was changed again. As we wanted our ANN model to be compared to the performance of a classical machine learning algorithm, we used logistic regression independently on the original data. After training the model, we used PR (precision vs recall) curve to evaluate the performance. Finally, we compared the performance of autoencoder stacked with logistic regression and plain logistic regression models using confusion matrices and AUCPR score.

References

"Credit Card Fraud Statistics". 2020. Shiftprocessing.Com.

Gómez, Jon Ander, Juan Arévalo, Roberto Paredes, and Jordi Nin. 2018. "End-To-End Neural Network Architecture For Fraud Scoring In Card Payments". *Pattern Recognition Letters* 105: 175-181.

Jurgovsky, Johannes, Michael Granitzer, Konstantin Ziegler, Sylvie Calabretto, Pierre-Edouard Portier, Liyun He-Guelton, and Olivier Caelen. 2018. "Sequence Classification For Credit-Card Fraud Detection". *Expert Systems With Applications* 100: 234-245.