Fraud Detection for Transactions

Kausik Chattapadhyay

Bellevue University

DSC680: Applied Data Science

Prof. Amirfarrokh Iranitalab

04/02/2024

Fraud Detection for Transaction

Business Problem:

The Fraud Company aims to enhance its fraud detection service in the Brazilian/South America market. With a revenue-sharing model heavily dependent on correct fraud detection, the company faces significant financial risk if its models are not precise. This white paper explores the development of a machine learning model to predict whether a financial transaction is fraudulent or legitimate, aiming to mitigate financial risks associated with fraudulent transactions.

Background/History:

The Fraud Company is a company specialized in detecting fraud in financial transactions made through mobile devices. The company has a service called "Fraud" which guarantees the blocking of fraudulent transactions. The business model of the company is of the Service type with the monetization made by the performance of the service provided, in other words, the user pays a fixed fee on the success in detecting fraud in the customer's transactions.

However, the Fraud Company is expanding in Brazil/South America and to acquire customers more quickly, it has adopted an aggressive strategy. The strategy works as follows:

1. The company will receive 25% of the value of each transaction truly detected as fraud.
2. The company will receive 5% of the value of each transaction detected as fraud, but the transaction is truly legitimate.
3. The company will return 100% of the value to the customer, for each transaction detected as legitimate, however the transaction is truly a fraud.

With this aggressive strategy, the company assumes the risks of failing to detect fraud and is remunerated for assertive fraud detection. For the client, it is an excellent business to hire the Fraud Company. Although the fee charged is remarkably high on success, 25%, the company reduces its costs with fraudulent transactions detected correctly and even the damage caused by an error in the anti-fraud service will be covered by the Fraud Company itself.

For the company, in addition to getting many customers with this risky strategy to guarantee reimbursement in the event of a failure to detect customer fraud, it depends only on the precision and accuracy of the models built by its Data Scientists, in other words, how much the more accurate the "Blocker Fraud" model, the greater the company's revenue. However, if the model has low accuracy, the company could have a huge loss.

Business Assumption:

Fraud prevention is the implementation of a strategy to detect fraudulent transactions or banking actions and prevent these actions from causing financial damage and the reputation of the client and the financial institution. There are always financial frauds, and They can happen through virtual and physical ways. So, the investment in security has been increasing. The losses caused by fraud can reach R$ 1 billion - which corresponds to half the amount that institutions invest in technology systems aimed at information security every year, according to Febraban's 2020 Banking Technology Survey.

Business Questions:

- What is the model's Precision and Accuracy?
- How Reliable is the model in classifying transactions as legitimate or fraudulent?
- What is the Expected Billing by the Company if we classify 100% of transactions with the model?
- What is the Loss Expected by the Company in case of model failure?

- What is the Profit Expected by the Blocker Fraud Company when using the model?

Column Descriptions:

step: maps a unit of time in the real world. In this case 1 step is 1 hour of time. Total steps 744 (30 days simulation).

type: CASH-IN, CASH-OUT, DEBIT, PAYMENT and TRANSFER.

amount: amount of the transaction in local currency.

nameOrig: customer who started the transaction

oldbalanceOrg: initial balance before the transaction

newbalanceOrig: new balance after the transaction

nameDest: customer who is the recipient of the transaction

oldbalanceDest: initial balance recipient before the transaction. Note that there is not information for customers that start with M (Merchants).

newbalanceDest: new balance recipient after the transaction. Note that there is not information for customers that start with M (Merchants).

isFraud: This is the transactions made by the fraudulent agents inside the simulation. In this specific data set the fraudulent behavior of the agents aims to profit by taking control of our customers' accounts and trying to empty the funds by transferring them to another account and then cashing out of the system.

isFlaggedFraud: The business model aims to control massive transfers from one account to another and flags illegal attempts. An illegal attempt in this dataset is an attempt to transfer more than 200.000 in a single transaction.

Solution Strategy:

My solution to solve this problem will be the development of a data science project. This project will have a machine learning model which can predict whether a transaction is fraudulent or not.

Step 01. Data Description: In this first section the data will be collected and studied. The missing values will be threatened or removed. Finally, an initial data description will be carried out to know the data. Therefore, some calculations of descriptive statistics will be made, such as kurtosis, skewness, media, fashion, median and standard deviation.

Step 02. Feature Engineering: In this section, a mind map will be created to assist in the creation of the hypothesis and the creation of new features. These assumptions will help in exploratory data analysis and may improve the model scores.

Step 03. Data Filtering: Data filtering is used to remove columns or rows that are not part of the business. For example, columns with customer ID, hash code or rows with age that does not consist of human age.

Step 04. Exploratory Data Analysis: The exploratory data analysis section consists of univariate analysis, bivariate analysis and multivariate analysis to assist in understanding of the database. The hypothesis created in step 02 will be tested in the bivariate analysis.

Step 05. Data Preparation: In this fifth section, the data will be prepared for machine learning modeling. Therefore, they will be transformed to improve the learning of the machine learning model, thus they can be encoded, oversampled, subsampled, or rescaled.

Step 06. Feature Selection: After the data preparation in this section algorithms, like Boruta, will select the best columns to be used for the training of the machine learning model. This reduces the dimensionality of the database and decreases the chances of overfitting.

Step 07. Machine Learning Modeling: Step 07 aims to train machine learning algorithms and how they can predict the data. For validation the model is trained, validated and applied to cross validation to know the learning capacity of the model.

Step 08. Hyperparameter Fine Tuning: Selected the best model to be applied in the project, it's important to make a fine tuning of the parameters to improve its scores. The same model performance methods applied in step 07 are used.

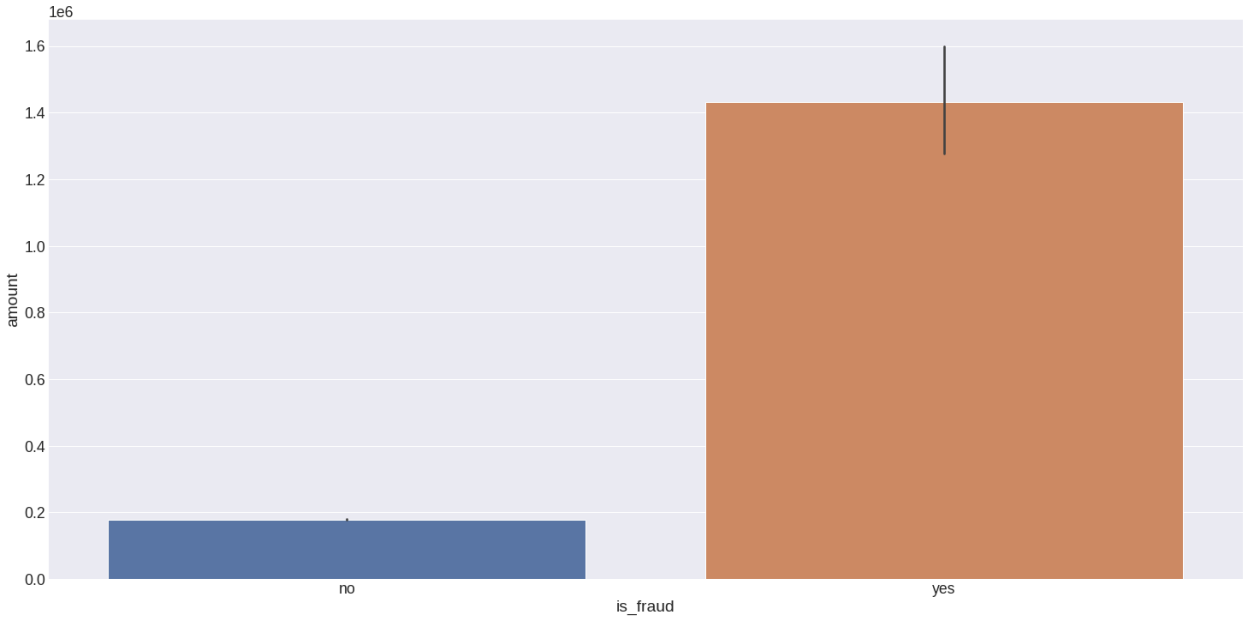
Step 09. Conclusions: This is a conclusion stage in which the generation capacity model is tested using unseen data. In addition, some business questions are answered to show the applicability of the model in the business context.

Step 10. Model Deploy: This is the final step of the data science project. So, in this step the flask Api is created, and the model and the functions are saved to be implemented in the Api.

Top 3 Data Insights:

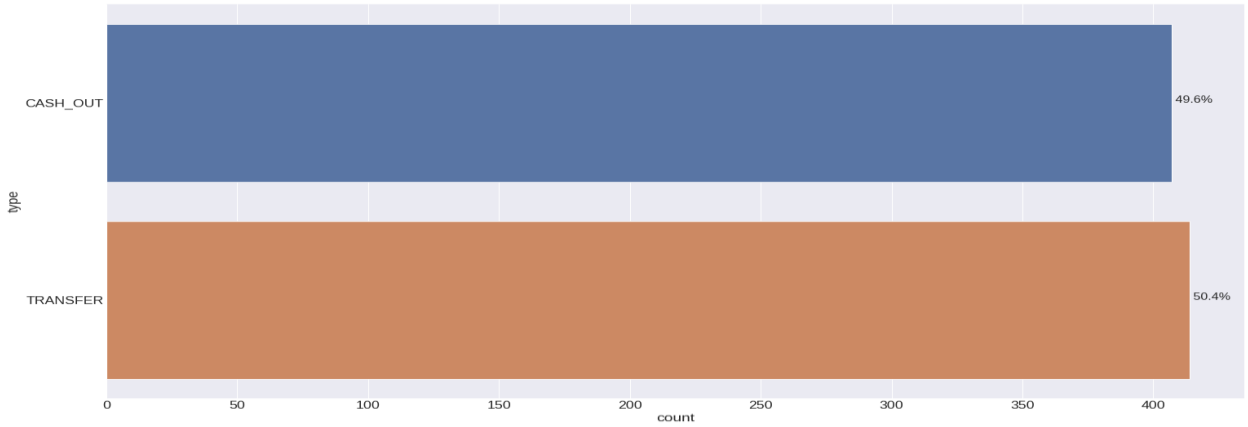
- **All the fraud amount is greater than 10.000.**

TRUE: The values are greater than 10.000. But it is important to note that the no-fraud values are greater than 100.000 also.



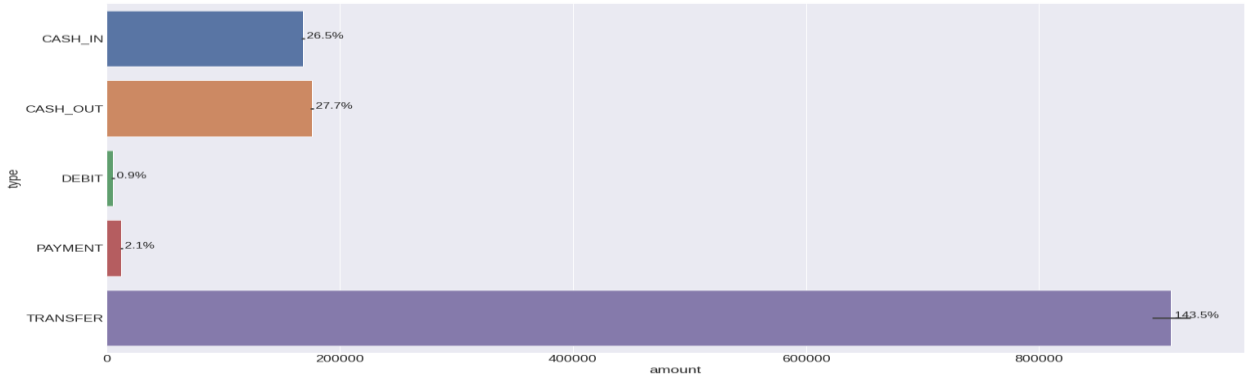
- **60% of Fraud transactions occur using the cash-out-type method.**

FALSE: The fraud transaction occurs in transfer and cash-out type. However, they are almost the same value.



- **Values greater than 100.000 occurs using the transfer-type method.**

FALSE: Most transactions occur in transfer-type, however transactions greater than 100.000 occur in cash-out and cash-in too.



Machine Learning Applied:

Here is all cross-validation results of the machine learning model with their default parameters. The cross-validation method is important to show the capacity of the model to learn.

Dummy Model

Balanced Accuracy	Precision	Recall	F1	Kappa
0.499 +/- 0.0	0.0 +/- 0.0	0.0 +/- 0.0	0.0 +/- 0.0	-0.001 +/- 0.0

Logistic Regression

Balanced Accuracy	Precision	Recall	F1	Kappa
0.565 +/- 0.009	1.0 +/- 0.0	0.129 +/- 0.017	0.229 +/- 0.027	0.228 +/- 0.027

K Nearest Neighbors

Balanced Accuracy	Precision	Recall	F1	Kappa
0.705 +/- 0.037	0.942 +/- 0.022	0.409 +/- 0.074	0.568 +/- 0.073	0.567 +/- 0.073

Support Vector Machine

Balanced Accuracy	Precision	Recall	F1	Kappa
0.595 +/- 0.013	1.0 +/- 0.0	0.19 +/- 0.026	0.319 +/- 0.0373	0.319 +/- 0.037

Random Forest

Balanced Accuracy	Precision	Recall	F1	Kappa
0.865 +/- 0.017	0.972 +/- 0.014	0.731 +/- 0.033	0.834 +/- 0.022	0.833 +/- 0.022

XGBoost

Balanced Accuracy	Precision	Recall	F1	Kappa
0.88 +/- 0.016	0.963 +/- 0.008	0.761 +/- 0.033	0.85 +/- 0.023	0.85 +/- 0.023

LightGBM

Balanced Accuracy	Precision	Recall	F1	Kappa
0.701 +/- 0.089	0.18 +/- 0.1	0.407 +/- 0.175	0.241 +/- 0.128	0.239 +/- 0.129

Machine Learning Performance:

The chosen model was **XGBoost** and it was tuned to improve their parameters and scores. Below there's a table with the capacity of the model to learn.

Balanced Accuracy	Precision	Recall	F1	Kappa
0.881 +/- 0.017	0.963 +/- 0.007	0.763 +/- 0.035	0.851 +/- 0.023	0.851 +/- 0.023

It's possible to determinize the capacity of the model to generalize using unseen data. In other words, capcity of the model to classify new data as shown.

Balanced Accuracy	Precision	Recall	F1	Kappa
0.915	0.944	0.829	0.883	0.883

Business Results:

The company receives 25% of each transaction's value truly detected as fraud.

The company can receive R$ 60,613,782.88 detecting fraud transactions.

The company receives 5% of each transaction value detected as fraud, however the transaction is legitimate.

For wrong decisions, the company can receive R$ 183,866.98.

The company gives back 100% of the value for the customer in each transaction detected as legitimate, however the transaction is actually a fraud.

The company must return the amount of R$ 3,546,075.42.

What is the model's Precision and Accuracy?

For unseen data, the value of balanced accuracy is equal 91.5% and precision is equal 94.4%.

How reliable is the model in classifying transactions as legitimate or fraudulent?

The model can detect 76.3% +/- 3.5% of the fraud. However, it detected 0.829 of the frauds from a unseen data.

What is the revenue expected by the company classify 100% of transactions with the model?

Using the model the company can get revenue of R$ 60,797,649.86. Using the current method to detect fraud the revenue is 0.00.

What is the loss expected by the Company if it classifies 100% of the transactions with the model?

For wrong classifications the company must return the amount of R$ 3,546,075.42. In contrast, for wrong classifications using the current method, the company must return the amount of R$ 246,001,206.94.

What is the profit expected by the fraud company when using the model?

The company can expect profits of R$ 57,251,574.44. The profit value of the current method is R$ -246,001,206.94.

My Code:

https://github.com/chatkausik/chatkausik.github.io/blob/master/Fraud%20Detection%20For%20Transactions/notebooks/transaction-fraud-detection-cycle1.ipynb

Conclusions:

The data is extremely unbalanced; however, it was possible to make all the data analysis and create with good scores. The company may expect a revenue of R$ 57,251,574.44. This result may show the capacity of a project of data science and help the company.

Ethical Considerations:

There are several ethical considerations to address in this project, including data privacy, bias in model predictions, and potential impacts on individuals falsely identified as fraudulent. It is essential to ensure that the model's predictions are fair and unbiased, and measures should be taken to protect sensitive customer information.

Challenges/Issues:

One significant challenge is dealing with imbalanced data, where fraudulent transactions are rare compared to legitimate ones. This imbalance can lead to biased models favoring the majority class. Additionally, ensuring the model's interpretability and explainability is crucial, especially in a financial context where decisions have significant consequences.

Deliver Fraud Company a production model in which my access will be done via API, that is, customers will send their transactions via API so that my model classifies them as fraudulent or legitimate.

Lessons Learned:

- Even when the classes are unbalanced, it is possible to create a model with good scores.
- It is possible to create a model that can classify classes with less than 1% of samples.

Next Steps:

- Test at most more than 10 hypotheses.
- Implement oversampling or subsampling techniques to improve the model scores.
- Implement the Api on the Heroku platform.

References

https://www.kaggle.com/datasets/ealaxi/paysim1