Ravelin

Detecting fraudulent activity from customer data
Description of analytical approaches (semi-technical)

Viktoria Csink

I. Problem definition

The objective of the current exercise was to explore whether it is possible to predict fraudulent financial activity from customer data. The dataset contained information whether a transaction was fraudulent or not ("fraudulent": True/False), therefore I approached this exercise as a supervised machine learning problem.

Firstly, I tried to understand the data by exploring how the variables were related to each other (Part 1.)

I then continued with data visualisation to see which features are likely to be important in determining whether a transaction is fraudulent or not. (Part 2.)

Thirdly, based on the exploratory data analysis above, I carried out feature engineering to discard variables that are unlikely to be informative, to combine the information from several features into one feature, and finally to encode the categorical variables into the appropriate format for machine learning. (Part 3.)

Finally, I built a supervised machine learning model to determine which transactions are likely to be fraudulent. This model showed a high accuracy in identifying fraudulent transactions. Most importantly, the metric for "false negatives" (i.e. missed fraudulent activity) also indicated very good model performance. (Part 4.)

II. The data

A number of variables in the dataset contained personal data linked to *customers* (e.g. email address, billing address, IP address, etc), and other variables were linked to specific *orders* and *transactions* (e.g. transaction amount, whether the transaction failed or not, the shipping address of the order, etc).

The first thing that caught my eye about the data was that the predicted variable (fraudulent/not fraudulent) was linked to *customers*, rather than specific transactions. In other words, all customers who carried out a fraudulent transaction always did so, and the rest of the customers never engaged in fraud. I thought that this was an interesting aspect of the data, as I would have imagined that - outside of the scope of this exercise - the same customer would carry out both fraudulent and non-fraudulent activities. However, it is possible that all fraudulent activity is carried out with fake details, therefore identifying fraudulent customers is enough to detect fraudulent activity. I would be very interested to follow up on this question from a domain knowledge perspective.

III. Analytical approaches

3.1. Exploring variables

I decided that the first task was to understand how the different features are linked to one another and what the units of observation are.

I noticed that some variables were linked to customers (e.g. email address, billing address, etc) , and others to specific transactions (transaction id, paymentMethodId, etc). I wanted to find a way to link all these variables to one another. I noticed that 'transactions' were linked to 'orders' through the dictionary key 'orderId' and 'transactions' were linked to 'paymentMethods' through the key 'paymentMethodId'.

Based on this information, I decided that the units of observation will be 'transactions', and I created a pandas data frame where each row is a transaction, and all the information corresponding to that transaction appears in the same row.

During the exploration of the variables, I also noticed that - with the exception of 'transactionAmount' - all the other variables were ordinal or nominal variables. This means that nominal variables with plenty of unique values (such as a unique orderId for each order) will not be useful for machine learning.

3.2. Visual exploration

Subsequently, I conducted a visual exploration of the data to understand how the features were linked. I explored whether the *payment method*, the *payment provider* and the *transaction amount* were indicative of whether a transaction is fraudulent or not.

Regarding payment methods, I found that PayPal was a safer method than the rest, i.e. this payment method was associated with the largest difference between fraudulent and non-fraudulent activity, with far more transactions being non-fraudulent. In contrast, paying with bitcoin carried a higher risk of fraud compared to the other payment methods (see /visualisation/payment_methods.pdf).

Secondly, I found that certain payment providers were safer than others; for instance, Visa 13 digit and Voyager were associated with more risk than the rest of the providers, whereas MasterCard seemed to be the safest payment provider (see /visualisation/payment_providers.pdf)

Thirdly, I found that more expensive transactions were also more likely to be fraudulent (see visualisation/transaction_amounts.pdf). Interestingly, judging from the error bars on the graph, the mean transaction amounts of fraudulent and non-fraudulent activity were not statistically significant. However, statistical significance in traditional inferential statistics is highly dependent on the variance and on sample sizes, therefore these insights should be treated with caution. (Also fraudulent and non-fraudulent sample sizes were not equal in this dataset, therefore a parametric t-test was no appropriate to perform in this case).

3.3. Feature engineering

Before training a machine learning model, I decided to restrict the feature space to variables that are likely to be informative in detecting fraud.

Firstly, I eliminated nominal variables with many unique values which are unlikely to be useful, such as 'transactionId', 'orderId' or 'paymentMethodId'.

Secondly, I explored incongruities in the customers' personal data: i.e. a customer registering with a single email address and several different phone numbers, billing addresses or IP addresses. I found that a single customer registered with one email address but 6 different numbers/billing addresses and carried out 28 fraudulent transactions. Therefore, I added a column ('incongruities') to denote such discrepancies and dropped the rest of the personal data from the analysis.

I also considered it an incongruity if the billing address was different from the shipping address - although this may not necessary be an indicator of crime (see the Future directions section).

Lastly, I encoded the categorical and the boolean variables using OneHotEncoding into a numerical format that is interpretable for the machine learning algorithm.

3.4. Modelling

Once I had had an initial understanding about the data, the dataset was cleaned and the features were organised for machine learning, I built a supervised machine learning model to predict whether a transaction is likely to be fraudulent or not.

I decided to use Random Forest, because in my experience this model performs well with imbalanced classes - i.e. that there will always be less fraudulent than non-fraudulent transactions both in the training data as well as in the real world.

This model yielded strong results: the overall accuracy was 85% (correct predictions / all predictions made). More importantly, recall (i.e. transactions correctly identified as fraudulent / (transactions correctly identified as fraudulent + missed fraudulent transactions) was 87%. In other words, 87% of fraudulent transactions were actually identified as fraudulent by the model.

In this model, I had lowered the decision boundary to 0.4, which results in a model that is geared towards catching fraudulent activity, even at the expense of possibly incorrectly identifying a few non-fraudulent transactions. I made this decision because I understood the main objective as the identification of fraudulent activity, and therefore I decided to optimise the model for recall.

I then conducted permutation analysis on the features to understand which feature is most important in predicting fraud and to rank and quantify the importances of the features (/results/feature_importances.csv).

IV. Summary of findings

Indeed, the dataset was appropriate for building a machine learning model that flags fraudulent transactions with a high accuracy.

The most important feature in determining the presence or absence of fraud was 'transactionAmount' - as it was also indicated by the plot.

Secondly - as it was also indicated by the visual exploration of the data - certain payment providers carried a higher risk of fraud than others. As also apparent on the graph, the ranking of feature importances confirmed that Voyager was associated with more fraudulent activity than the rest of the providers.

Importantly, the feature "incongruity", which encompasses various types of customer information, such as address, phone number and IP address, emerged as a highly important feature. This indicates that financial crime is strongly linked to certain

customers, therefore identifying customers correctly is extremely important in predicting whether their activity is fraudulent or not.

V. Limitations and future directions

I thoroughly enjoyed carrying out this analysis and I would be very interested to explore some of these questions further.

Firstly, I wonder whether the most appropriate units of observation for this type of analysis are “transactions” or “customers”. In this dataset, all transactions of a customer were either fraudulent or not. If this is the case, then identifying the customer and understanding the personal data of customers is the highest priority, rather than exploring the variables that relate to specific transactions.

Secondly, I found it very interesting that the incongruities in the customers’ personal data emerged as an important feature in predicting financial crime. While it seems intuitively true that if multiple transactions are linked to the same IP address, and yet they are linked to different names and postal addresses, then these transactions are rightfully suspicious. However, customers might change their phone numbers, move to a new area or use a different device, and this does not necessarily mean that their transactions are fraudulent. This question has made me recognise the importance of specific domain knowledge when building such predictive models.

Lastly, from a machine learning point of view, I am intrigued by the problem of imbalanced classes when it comes to identifying financial crime. In this dataset, I identified 257 fraudulent and 366 non-fraudulent transactions, which indicates a slight class imbalance. However, in the real world I anticipate this ratio to be much more tilted towards non-fraudulent activity, with very few cases of financial crime compared to all transactions. This in turn poses the very important and very interesting challenge of predicting rare events in the midst of ordinary events.