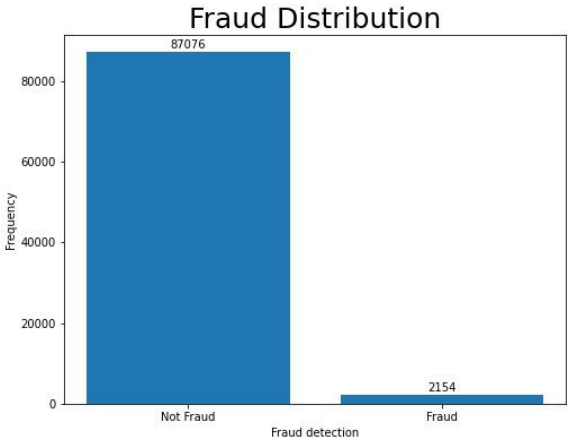
Winter 2023 Scotia DSD Problem
Fraud Detection
Brief Write-up

Team: B4U
Tao Shan, Yuxuan Liu, Fanghe Lin, Peiyi Zheng

Jan 21, 2023

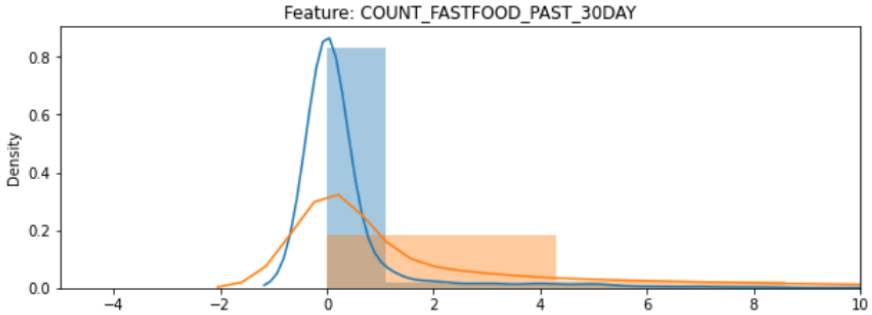
Fraud has always been a threat to financial institutions. In this report, we illustrate our approach to preventing fraud using XGBoost model and a fraud scoring system, which allows us to prevent fraud while minimizing the negative impact on customer experience.



Fraud is a complex issue that can be driven by a variety of factors. One of which is people's pressure, specifically financial stress caused by high levels of debt, job loss, and reduced income. The COVID-19 pandemic has increased financial pressure on individuals, and in Canada, the average debt per person has been increased continuously, which providing an opportunity for someone to commit fraud for financial gain.

Based on that, we choose to add two extra data set from outside, Canada's 2021 Covid-19 positive and death rate and Canada's 2021 bank of interest rates.
link:Covid-19,Bank-of-Canada

The data was preprocessed by removing duplicate and similar columns by features' distribution analysis, and cost-sensitive learning and correlation-based feature selection were used to solve the problem of unbalanced dataset. Downsampling plus SMOTE method was attempted but did not work as well as other methods.

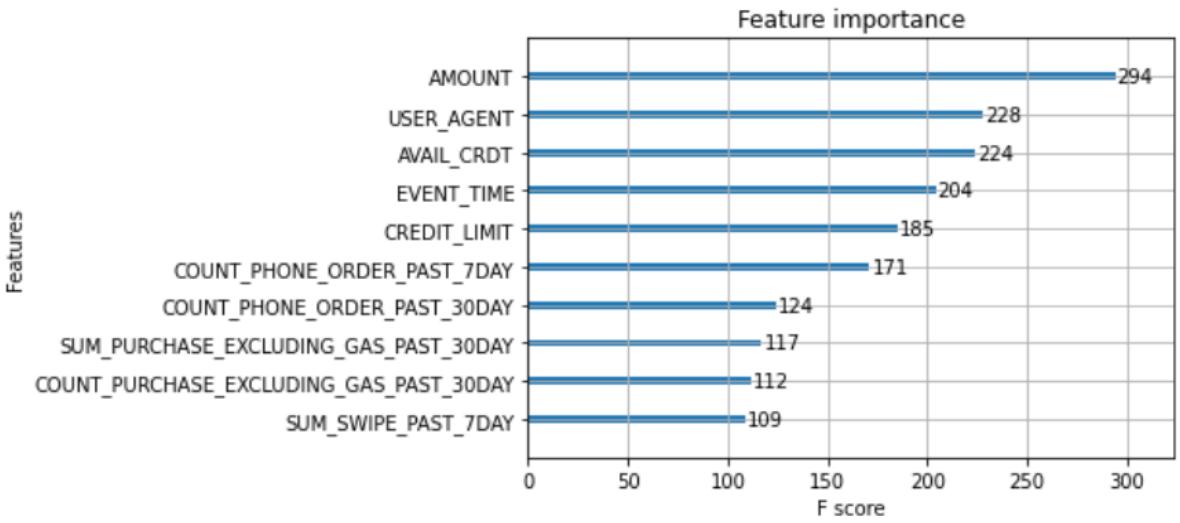


XGBoost is used to model and predict fraud transactions, which is more effective in preventing fraud without generating operational overheads. The model is trained on previous transaction data as well as external data, and can automatically detect and flag potential fraud cases, eliminating the need for manual processes and saving money by reducing investigation costs.

	Logistic Regression	Rigid Regression	XGBoost	XGBoost (Fine Tuning)
ROC-AUC	0.922	0.914	0.956	0.962
Recall	0.165	0.026	0.270	0.383
Precision	0.706	0.772	0.764	0.777
Accuracy	0.978	0.976	0.980	0.982

The model was evaluated using AUC (Area Under the ROC Curve) as the evaluation metric and obtained a score of 0.962, indicating a high accuracy in identifying fraud and non-fraud cases. Additionally, an F1-score of 0.609 indicates that the model has a fair balance of precision and recall.

To decrease fraud, we use feature importance to identify key attributes and sort them using weighted F-score to decide which transactions to decline.



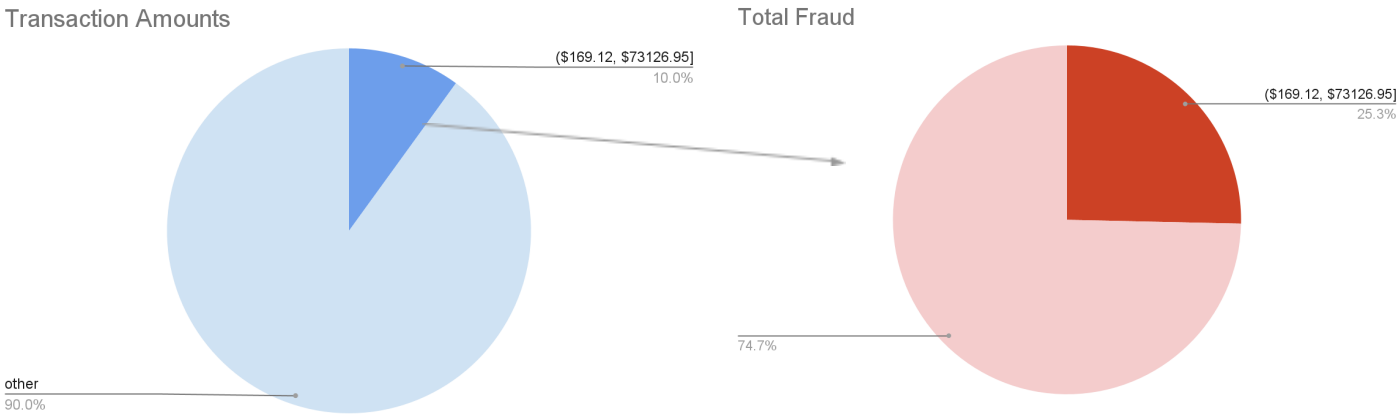
We also introduced a fraud scoring system based on the predictions and probabilities, which is similar to the credit score. We take the predicted probability multiplied by 1000

and classify the fraud score in three risk groups: low, medium and high. Customized fraud prevention strategies will be developed for those groups while collabrating with other departments to enhance the customer experience.

Groups	Fraud Score	Probability of Fraud
low risk	0-200	0.021%
medium risk	200-700	55.862%
high risk	700-1000	99.680%

A few suggestions for future fraud-prevention work include: continuously monitoring and gathering data, updating the model, exploring new technologies like AI/NLP, and collaborating with other organizations to share information and intelligence.

Interesting Facts we found from the dataset:



- Transaction Amounts in the range (169.12, 73126.95] account for 10% of data but 25.3% of fraud flag
- 12:00 am to 8:45 am each day, fraud rate is more than twice
- Less credit limit leads to more fraud
- Number of phone orders/number of purchase for excluding gas < 3 times for the past 30 days takes up to 45.6% and 47.12% of fraud flag respectively