Scotia Data Science Day
Winter 2023

Background

The widespread use of credit cards in Canada, coupled with the increase in online transactions during the Covid-19 pandemic, has made credit card fraud a significant problem. With over 80[1] million credit cards in circulation in 2021 and more than 6 billion[2] purchase transactions made each year in Canada, fraudsters have been using credit cards to obtain goods and services without paying. This not only harms the profitability of financial institutions and the security of consumers, but it also poses a threat to the stability of the financial system as a whole. As the intensity of transaction fraud continues to increase, it is essential for financial institutions and consumers to take steps to prevent and mitigate the impact of this crime.

Scenario

As a young and talented data scientist, you joined a credit card company recently and met a group of analysts who share the same interests as you. Your team is well-suited to tackle the challenge of detecting and preventing fraud.

When you begin analyzing the customer transaction data, you quickly notice that fraud losses have been steadily increasing in the latest two years. This alarming trend is causing concern for the profitability of the credit card line of business. After conducting some research, you discover that the company's fraud detection models have not been updated in years, making them vulnerable to attacks from increasingly sophisticated fraudsters.

In order to address this problem, your team is tasked with building a new fraud detection model that can accurately identify fraudulent transactions. Additionally, you are asked to provide insights and recommendations on how to prevent fraud in the future. This will require you to analyze the data, identify patterns and trends, and develop strategies for detecting and preventing fraud.

Ask

Your team needs to develop a transaction fraud detection model based on the given historical data and any publicly available external data. Furthermore, you need to look for any insights from the data that can help you conduct statistical/business analysis to answer the following questions:

How to prevent fraud more effectively without creating operation overhead?
Which transactions should we decline to reduce fraud?
What are some key attributes that help to make the decline decision?
How to minimize the negative impact to customer experience while preventing fraud?
Can you make any long-term suggestions?
How did you arrive at the conclusions?

---

[1] "Number of Credit Cards in Canada 2000-2021." Statista, 2 June 2022, https://www.statista.com/statistics/594299/number-of-credit-cards-canada/
[2] "Canada: Credit card use per capita" Statista, 5 September 2022, https://www.statista.com/statistics/1309039/total-number-of-credit-card-payments-in-canada/

Tools and Methods

You can use any common statistical or programming languages such as Python, R, etc. You can try different machine learning methods like logistic regression, XGBoost, anomaly detection, etc.

Teams

Each team should have up to five students. Students who have registered without a team will be grouped randomly into teams.

Data

- 89,230 records in the training dataset; 22,307 records in the test dataset
- Each record is a credit card transaction, 1 ID column (TRANSACTION_ID) to uniquely identify a transaction
- 1 target column (FRAUD_FLAG): whether a transaction is labeled as fraud after fraud review
- 173 features, description of features is available in data dictionary

Submission

Submission should be made to the private MS Teams channel before 12pm on Sunday, January 22, 2023. Participants are expected to submit all four items below and follow the naming conventions. Details of what each item should include are provided in the Judging Criteria document.

1) prediction for the testing dataset in csv format
   - The file should have three columns: TRANSACTION_ID, PROBABILITY (probability of fraud), and PREDICTION (value of 0 indicating not fraud, or 1 indicating fraud)
   - [team_name]_prediction.csv
2) code in any format (Jupyter Notebook, R scripts, etc.)
   - [team_name]_code
3) slideshow presentation (Max 10 pages)
   - [team_name]_slides.pptx/pdf
4) brief write-up (Max 500 words)
   - [team_name]_brief.pdf

Assessment

There are two rounds of assessment evaluated by Scotiabank Director & VP and University of Waterloo faculty members. Detailed judging criteria is provided in the Judging Criteria document.

Criteria for Round 1:

- Performance of the proposed fraud prevention model, evaluated on the testing dataset using one of the key measurements: F1 score, defined as the harmonic mean of precision and recall for the fraud class. Other metrics, e.g., area under the ROC curve might also be considered.
- Insights, recommendation, and visualization in slides
- Evidence of analytical rigor and creativity

- Use of external data to improve model performance and derive business insight

The six best teams (determined by overall score calculated based on Judging Criteria) will be selected to present their work in a 6-minute presentation to the judges.

Criteria for Round 2:

- Clarity and organization of thought
- Overall presentation
- Analytical insight

Prize Structure

- First Place: $2,000
- Second Place: $1,500 x 2
- Third Place: $1,000 x 3
- Invite First Place to meet Scotiabank D&A Teams / Execs for dinner on Sunday

Important Notes

- Use MS Teams public channel to drop your questions or attend the mentorship session on Saturday
- Formal Kick-off Friday (Jan 20th), please join to check-in and ask questions!
- Please submit your team result with required format on time & prepare presentation ahead of time
- Final presentations on Sunday (Jan 22nd), team must attend in person to qualify for Top 6 winning teams

Thank you for your interest, and good luck!