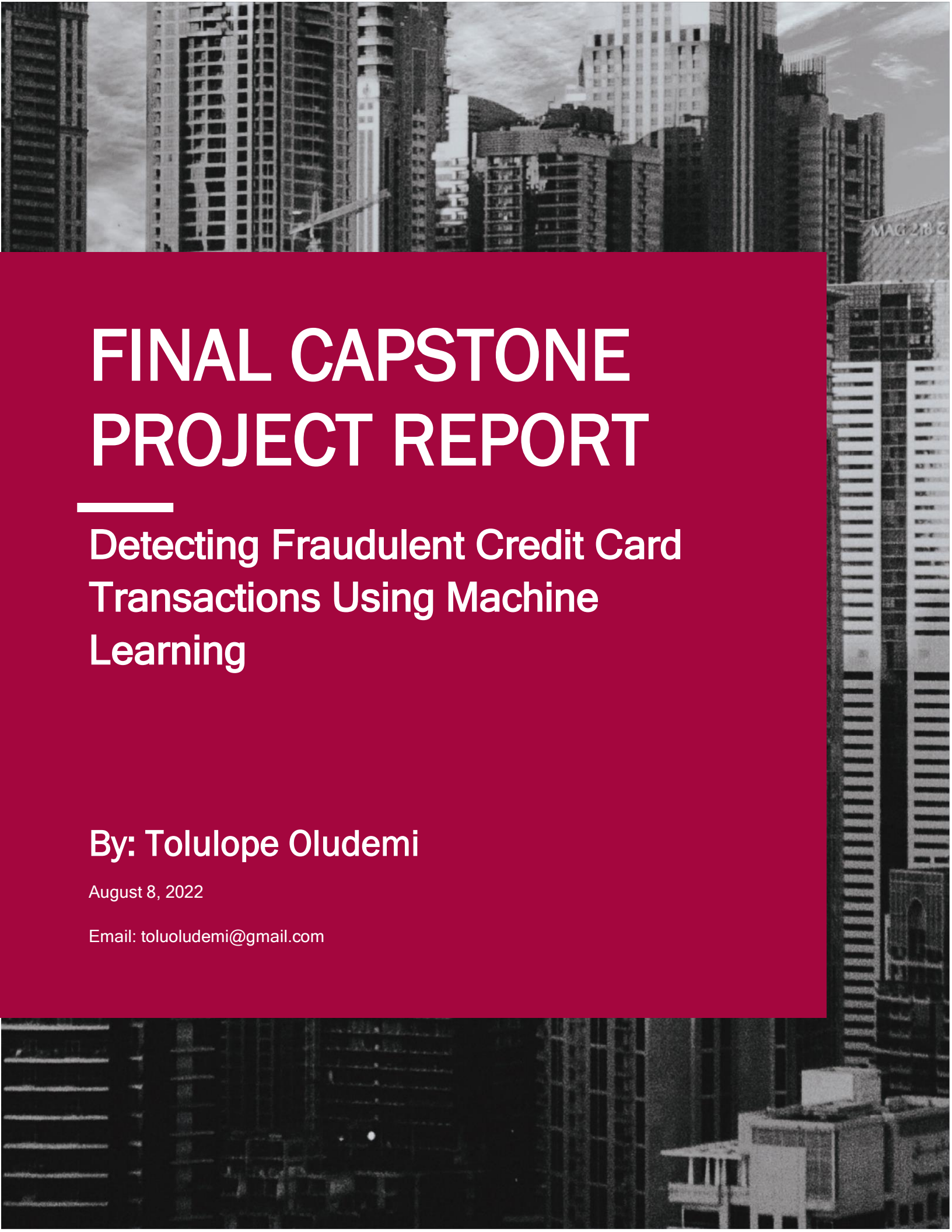


FINAL CAPSTONE PROJECT REPORT

Detecting Fraudulent Credit Card Transactions Using Machine Learning

By: Tolulope Oludemi

August 8, 2022

Email: toluoludemi@gmail.com

Introduction

This report summarizes the process of using machine learning to detect fraudulent credit card transactions, as well as key insights, practical applications, and results.

Background On the Subject Matter

Fraud is a major problem in the financial industry in Canada and United States. According to research, $3.3B was lost to fraud in the United States in 2021, and $300M was lost to fraud in Canada in 2021. There were $2.2M cases of fraud reported in the United States in the same year, and about 39% of Canadians feared using their credit cards due to the fear of their account being compromised. The source of this information is provided in the Reference section.

Problem Statement

The problem statement is **using machine learning on credit card transactions, can I create a model that predicts fraudulent transactions to allow financial institutions to provide protection to their customers' finances, improving their service, and overall business growth**. For example, a model that could detect 60% of fraudulent transactions could save banks like RBC or TD, millions of dollars.

Data Collection

Banking information is confidential, so the data used is a simulated credit card transaction dataset. It was synthetically made using a "Sparkov Data Generation tool", and it was generated to create a more realistic representation of simulated transactions within the United States. The generator was collected from a GitHub repository and stored in Kaggle. The dataset contains both legitimate and fraudulent transactions from 2019 to 2020, and it represents transactions between customers and merchants. The data has about 1.2M rows and 22 columns, containing numeric and categorical columns with detailed information on each customer, including the transaction and merchant information.

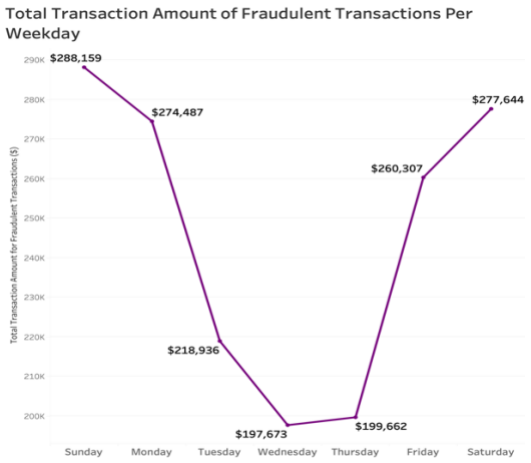
Pre-Modeling Data Transformation

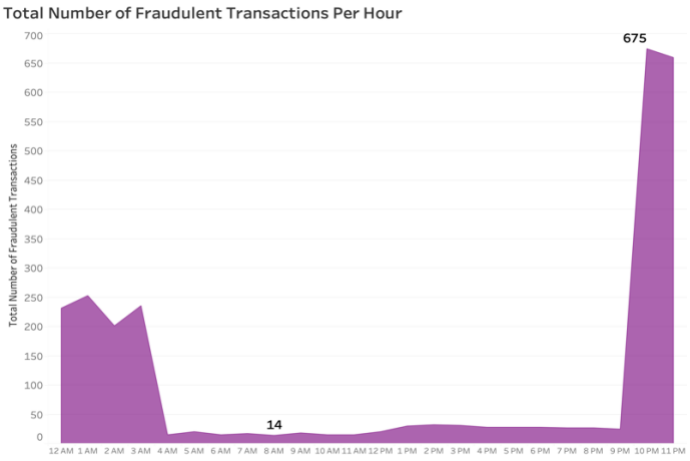
The data cleaning and transformation process included changing datatypes appropriately, for example dates initially saved as an 'object' type was changed to a 'datetime' datatype. During the exploratory data analysis (EDA) process, an extreme class imbalance was observed with over 99% of the data being non-fraudulent, and under 1% being fraudulent.

During preprocessing and feature engineering, bi-variate analysis was conducted with the purpose of increasing the percentage of fraudulent transactions by decreasing the percentage of non-fraudulent transactions. This entailed removing categories that had low to no occurrences of fraud, this allowed the model to be more sensitive to the patterns of fraudulent transactions, leading to better model performance at detecting fraudulent transactions. The final class distribution was 97.93% of the data being non-fraudulent, and 2.07% of the data being fraudulent, which is an improvement.

Some insights gained from EDA was the understanding of fraudulent patterns. For example, by dollar cost, the highest amount of fraudulent transactions occurred on the weekends - Sunday and Saturday.

More fraudulent transactions occur in the night, mainly from 10PM to about 3AM. The visuals showing these insights are presented below.

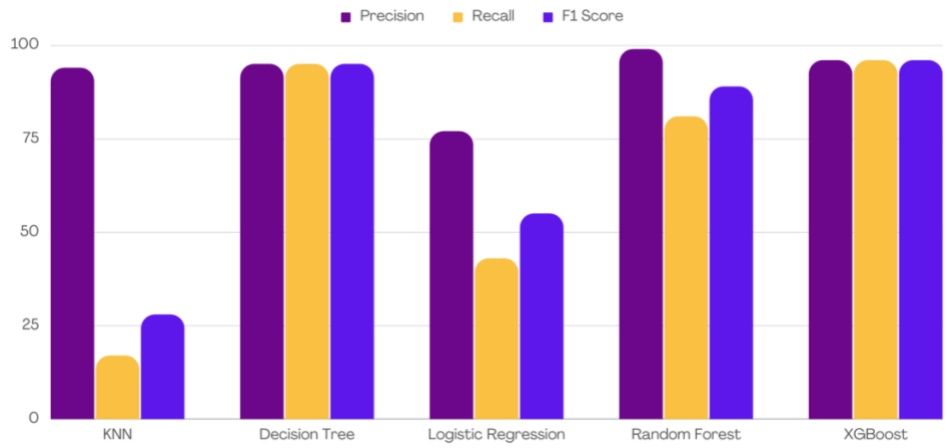




Summary Of Modeling

The target is a column identifying if the transaction is legitimate or fraudulent, with Class 0 being non-fraudulent, and Class 1 being fraudulent. Therefore, binary classification models (supervised learning) were used in the process. Prior to modeling, I set an evaluation metric to determine how the best performing model would be chosen. The goal was to detect as many fraudulent transactions as possible, without misclassifying too many legitimate transactions as fraudulent. So, if the model had a high recall score (ability to detect many fraudulent transactions) and a high f1 score (the harmonic mean, indicating the balance between the recall and precision score) then it was a good model.

The machine learning models used were K-Nearest Neighbor, Decision Tree, Logistic Regression, Random Forest, Neural Network, and eXtreme Gradient Boosting (XGBoost). The initial process was to create pipelines and use grid search to find the best parameters for each model, including up-sampling and down-sampling, but this proved to be computationally intensive. The final process was using default parameters on each model, then based on the model performance, the best model was chosen and then the parameters of the best performing model was optimized to see if better results can be achieved.

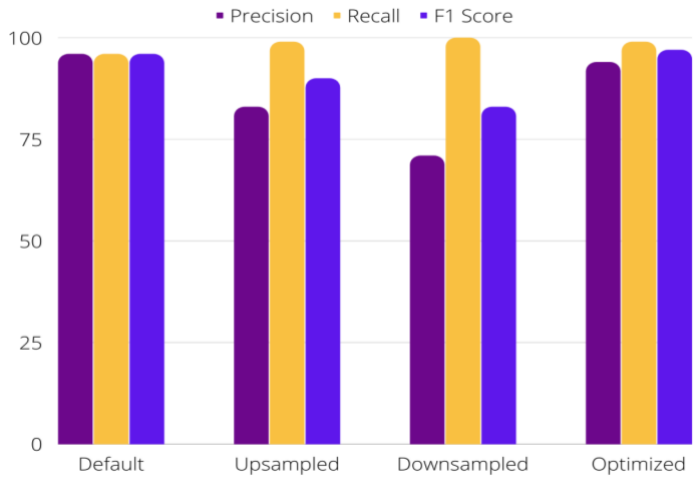


In this case, the best performing model with default parameters was XGBoost. Using this model, up-sampling and down-sampling was conducted to correct the class imbalance, and it was concluded that up-sampling performed even better than down-sampling and the default parameters of XGBoost on

the original class distribution. Afterwards, the parameters of the up-sampled XGBoost model were optimized using a grid search which resulted in an even better model, with 99% Recall, 97% F1 Score, and 94% Precision.

Findings and Practical Applications

The best performing model at detecting fraudulent transactions was the XGBoost model, accurately detecting 99% of the fraudulent transactions, with a good balance of recall and precision having 97% F1 Score.



Practically speaking, the model can be applied in financial institutions, FinTech companies, and credit card companies. In the dataset, about $1.8M was lost to fraudulent transactions. With the XGBoost model being able to detect 99% of fraudulent transactions, the model was able to detect $1.78M fraudulent transactions out of $1.8M, which could potentially be a money saver to companies that have credit cards in their business.

Next Steps

Some next steps and future directions are to be able to predict customers with increased likelihood of being defrauded based on their financial state (i.e., how much debt they have, their credit limit, etc.), and to further develop the model to be able to detect transactions outside of just credit cards or even further in spam detection using other machine learning methods.

Reference

Sources for Research:

- https://mint.intuit.com/blog/planning/credit-card-fraud-statistics/
- https://www.simplerate.ca/credit-card-fraud-statistics-canada/

Sources for Data and Generation Tool:

- https://www.kaggle.com/datasets/kartik2112/fraud-detection
- https://github.com/namebrandon/Sparkov_Data_Generation