Capstone Two - Project Proposal
Chris Chung

Credit Card Fraud Detection

Problem Statement
How accurate can banks detect fraudulent credit card transactions by learning a consumer's spending behavior?

Context
Instead of cash transactions, more and more credit cards are being used for day-to-day purchases. While this allows for more opportunities to spend money, it also opens up the risk of credit card fraud, which affects both the merchant and the consumer.

Alarmingly, 46% of the world's credit card fraud takes place in the United States. In 2018 alone, there were more than 200,000 breached credit card accounts. As for affecting businesses, in 2018 merchants lost $2.94 in revenue for every $1 in fraud.

Our goal is to create a data-driven model that measures the likelihood that a transaction is fraudulent.

(*Source*: https://paymentdepot.com/blog/credit-card-fraud-detection/)

Data Source
We will be using a simulated credit card transaction dataset from kaggle:

https://www.kaggle.com/kartik2112/fraud-detection

The dataset was simulated by using Sparkov Data Generation tool created by Brandon Harris. It contains legitimate and fraudulent transactions from the duration 01 Jan 2019 - 31 Dec 2020. It covers credit cards of 1000 customers doing transactions with a pool of 800 merchants.

Relevant variables include:

- trans_date_trans_time
- cc_num
- merchant
- category
- amt
- first
- last
- gender
- street
- city
- state

- job
- dob
- is_fraud
- ...

Criteria for success
The goal is to use the train set to create a model that can predict the likelihood that a transaction is fraudulent as accurately as possible. We will then use the test set to confirm that our trained model works as expected. Ideally, we'd like our model to be able to identify at least 95% of fraudulent transactions correctly.

Process
This project will be approached as a classification problem (assigning a given transaction as valid or fraudulent). Steps taken will be:

1. Clean the train dataset using Python3. Relevant variables for the prediction model will be manually selected, based on preliminary knowledge of variables.
2. Train dataset will be explored with graphs and statistical summaries.
3. Try different models (KNN, Decision Tree, etc). Use GridSearch and cross validation to train multiple models and look for the best parameters.
4. Identify top important features that are relevant in predicting whether a transaction will be fraudulent or not.
5. Test the model using the test set.

Deliverables

A GitHub repository will be created for this project, containing the original dataset, cleaned dataset, and jupyter notebook files for each step of the process. The finalized project will be presented with a project report document and a slide deck.