Credit Card Fraud Detection Case Study Summary
Candidate: Chang Liu, CFA

Overview

The dataset presents an extensive collection of 24 million credit card transactions, sourced from IBM's financial database. Capturing a wide spectrum of user interactions, the data provides a detailed snapshot of transaction behaviors, patterns, and potential vulnerabilities across a relatively long time horizon. Details include: User & card info, transaction time and dollar amount, transaction types, merchant details, fraud detection flag. The goal for this analysis is to examine the dataset for its quality and spending pattern, and then to train ML models to predict fraud activities.

Approach and Methodology

- **EDA:** Analyze time series trends, gain insights into total transaction spending, fraud transactions and system vulnerabilities.
- **ML Modeling:** Develop models based on past fraudulent transactions to predict and mitigate future fraud, aiming to enhance overall security and user trust.

Q1: Merchant Spending Analysis (30 mins)

- I used multiple definitions for top 5 merchants, as ranked by 1) total transaction counts, and 2) total transaction dollar amount. Each of the two can be further represented by either 1) the entire time frame, or 2) the top 5 ranks on each day, meaning the top 5 merchants might be different everyday.
- The overall pattern for total daily spending by transaction counts vs. by dollar amount is quite similar. We see significant acceleration in transaction amount and count starting the year 2000, likely due to the proliferation of the Internet and ecommerce transactions.

Q2: Spending Pattern Analysis (30-45 mins)

- **Total daily spending & Individual daily spending approach:** I first created a boxplot for all the ‘total daily spending’ and looked at the distribution and its median, the 1st and 3rd quartile, the max and min values, and tried to see what would be considered as “higher spending” using the 3rd quartile, and which days experienced those higher total spending, and investigated those days to see whether is any pattern.
- I used the IQR method to detect all the “outliers”, or relatively high spending days, where the upper bound threshold is the smaller of Q3 + IQR * 1.5, or $200,000 total daily spending. Then I found that Mon, Sat, and Sun are usually days with higher spending, which is not entirely surprising given it is the weekend.
- Then I used the distribution of individual daily spending and looked for outliers instead of aggregate total daily spending. Both the distribution and median spending showed consistent patterns across all seven days, with Mon, Fri, and Sat spending slightly higher.
- **Seasonal Decomposition:** The challenge is the clear upward trend over time, and it is challenging to spot significant seasonality with strong upward growth trends over the years. Therefore I also used seasonal decomposition to decompose the data into trends,

seasonality, and randomness. After running seasonal decomposition, by looking at only 1Q2023, we see the same seasonal pattern every week, and if we look at all years, it was the same pattern, with Sat, Sun, Mon spending significantly higher than the Tue, Wed, Thu. We see the same pattern for all weeks throughout the entire time frame, likely due to the fact that it is machine generated data, and the seasonality is algorithmically built in.

Q3: Bias Assessment in Datasets (1 hour)

- **Data Imbalance:** Fraud is only less than 1% of total transaction counts most of the time. I identified the imbalance by calculating the number of fraud vs. non-fraud transactions over time and identified its presence.
- **Geographical representation bias:** Here I used the 'Merchant State' column to look at how transactions are distributed geographically, and then merged it with USA population data to identify any pattern or misrepresentation. I found that the USA is disproportionately represented, and other countries may be under-indexed. However within the U.S. I found that the transaction distribution is very closely matched with the U.S. population by state after I joined the data with real US population data.
- **MCC bias:** I grouped all the transactions by MCC code and identified which MCC category is over-indexed vs. under-indexed. I also merged the aggregated data with MCC documentation data published by Visa and Mastercard to understand what each of the MCC codes represents.
    - Certain merchant categories carried a significantly higher percentage of transactions, such as grocery stores and convenience stores (combined 22.7% of total transactions), service stations like gas stations (10.8%), restaurants (7.4%), drug stores and pharmacies (5.8%), etc.
    - The top 10 categories accounted for 68.3% of the total number of transactions, and hence this might introduce bias towards these categories.
    - Several MCC categories accounted for the biggest share of fraud transactions, including Department Stores, Wholesale Clubs, Discount Stores, Money Orders, Wire Transfer, Drug Stores and Pharmacies, etc.

Q4: Fraud Prediction Model (1 hour)

- I used an under-sampling method to resample the data, using only 40,000 transaction records with 20% being fraud and 80% non-fraud. This will reduce the bias in the model caused by the data imbalance.
- I then experimented several models:
    - Logistic Regression
    - Decision Trees
    - Random Forest
    - Gradient Boosting Machines (GBM): LightGBM, XGBoost, CatBoost
- The initial cross validation results (accuracy, precision, f1-score, recall, auc-roc) showed that all models performed well and similarly, except for the logistic regression model. Of all models, XGBoost performed the best in all metrics except for recall score, leading me to pick it to do hyperparameter tuning.

- For tuning, I used grid search method to go through different parameter values across
    - 'n_estimators'
    - 'learning_rate'
    - 'max_depth'
    - 'subsample'
    - 'min_child_weight'
- Key constraint is time limit and memory and compute constraints. Due to the large data size, I had to resample the dataset into a smaller random sample, and cut down the number of models to experiment. I left out some models such as Bayesian based models, as well as reducing the number of parameters to fine tune.

Q5: Model Output Analysis (45 mins)

- For model result evaluation, I used the same metrics (precision, recall, f1-score) as well as confusion matrix to determine the performance of the XGBoost model.
- Overall, the model showed a high accuracy (95%) and performed well in classifying non-fraudulent transactions (class 0). Meanwhile, it showed a slightly lower recall (89%) for fraudulent transactions (class 1), indicating that it missed some fraudulent cases.
- **Online Transactions** is far more impactful than any other factors in predicting transaction frauds. This means that online transactions might be more vulnerable compared to physical transactions like chip or swipe.
- **Certain merchant states are also important features**, meaning that if transactions took place in states like OH, CA, TX, FL, the chances of encountering fraud are significantly higher.
- **MCC is also an important feature.** This indicated that certain merchant categories such as Department Stores, Wholesale Clubs, Discount Stores, Money Orders, Wire Transfer, Drug Stores and Pharmacies are prone to fraud.