FICO Analytic Challenge
Week 2: Data Analysis

Summary

Our goal for this project is to train a model that can detect fraudulent transactions. This week, we will start by diving into the data and identifying the differences between fraud and non-fraud transactions that we can leverage for fraud detection in the coming weeks.

First, we will explore the dataset to understand its structure and contents. We will look at various fields and values to see what they represent. We will also cover best practices for cleaning and prepping the data for analysis. Next, we will focus on identifying what makes a transaction more likely to be fraudulent. We will analyze the dataset in detail, splitting it into subpopulations like Domestic vs. International transactions to compare their fraud rates.

Throughout the week, we will use different analytical tools and techniques. We will calculate important statistical measures and create visualizations to spot trends and patterns. These insights will help us identify features that can distinguish between fraudulent and non-fraudulent transactions.

By the end of the week, students will have a solid grasp of the dataset, know how to visualize and analyze the data, and be able to identify key features that will aid in detecting fraudulent transactions.

Key Takeaways

- Introduce students to common fraud detection terminology.
- Introduce the dataset for the project, exploring the various fields and their meanings.
- How to use descriptive statistics and data visualization to identify patterns distinguishing fraud from non-fraud transactions.
- Calculate fraud rates and use basis points for precise expression of small changes.
- Describe how to divide the dataset into subpopulations, such as CP/CNP and domestic/cross-border and use fraud rates to identify fraud-prone groups.

Overall Learning Goal

- How to load the dataset using data types and *to_datetime* function.

- Using functions like *head()*, *info()*, *shape()*, and *describe()* to analyze the data and monitor/confirm the changes we make to it along the way.
- Finding and handling missing values
- Using descriptive statistics to look for patterns that can help us distinguish between fraud and non-fraud transactions.
- The concept of fraud and non-fraud accounts:
    - Non-fraud account: an account (pan) that does not have any fraudulent transactions.
    - Fraud account: an account (pan) with at least one fraudulent transaction.
- Introducing the concept of fraud rate and basis points:
    - Fraud rate: Fraud rate is calculated as the ratio of fraudulent transactions to the total number of transactions.
    - Basis points: basis points can be used to express small changes or differences in fraud rates in a precise manner. One basis point is equal to one-hundredth of one percent, or 0.01%. Thus, 100 basis points are equal to 1%.
- Using histograms and pie plots to visualize distribution of variables to detect patterns in data. For example: time of day histograms for fraud and non-fraud
- Understanding Card Present (CP) and Card Not Present (CNP) transactions, based on the "category" field in the dataset:
    - A transaction is classified as CNP if the category field contains the word "net"
    - If the category field does not include "net" the transaction is classified as CP.
- Explaining the distinction between domestic and cross-border transactions and how to determine them from the dataset:
    - A transaction is considered domestic if the merchCountry and cardholderCountry are identical.
    - A transaction is classified as cross-border if the merchCountry is different from the cardholderCountry.

Resources

Matplotlib:

- Frequently used in this week's notebook for visualization.

    - basic matplotlib tutorial
    - Official matplotlib tutorial- subplots

Pandas:

- Essential for manipulating dataframes.
    - Manipulating DataFrames with Pandas
    - Pandas cheat sheet
    - Official tutorial

Additional Resources:

- Exploratory data analysis by IBM
- Exploratory data analysis in python

Exercises & Challenges

Exercises

1. Create histograms that show the counts by merchCountry for frauds and non-frauds. (hint: use value_counts() function)
- Calculate fraud rates for each merchCountry.
2. Analyze cross-border vs. domestic transactions.
- Use a pie plot to visualize the fraud/non-fraud counts within each sub-population (cross-border vs. domestic).
- Calculate transaction-level fraud rates in cross border and domestic subpopulations.
- Analyze and explain the differences in fraud rates.

Challenge: Explore Additional Predictive Features

- Examine the existing data fields and brainstorm other features or subpopulations that could be predictive of fraud.