Credit Card Transaction Analysis

Mining transaction data to detect fraudulent purchases

By: Kevin Corboy & Brendan McDonald

Credit Card Transactions Dataset

Overview:

The dataset includes ~1.8M rows and 24 columns of detailed credit card transaction data. Each transaction includes purchase information about time and value, customer and merchant details (age, location, market segment, demographic data), and a binary value indicating whether the transaction is fraudulent or not. The dataset spans ~18 months, which also allows for analysis of seasonality and year-over-year spending trends.

Questions:

- Do some categories of spending (value, demographic, market segment, etc) have a higher probability and prevalence of fraud?
- What patterns can we identify in fraudulent transactions compared to normal transactions?
- How accurately can we predict fraudulent transactions using machine learning?
- How can we apply this information to improve security in different products and lines of business?

Prior Work

The credit card industry is responsible for processing *billions* of dollars in transactions on a daily basis, which makes timely and reliable detection of fraud paramount. An entire industry supports the banking and credit card transaction market in fraud detection, using all variants of regression analysis, decision trees and neural networks, and all types of AI/ML analysis to describe and make decisions using data.

This dataset was found on Kaggle and there are a number of analysis notebook submissions from different developers available. Since the authors are unknown and the quality is variable, we are not referencing these analyses in our work.

Datasets/References:

- *Dataset home: https://www.kaggle.com/datasets/priyamchoksi/credit-card-transactions-dataset*
- *Other analyst notebooks: https://www.kaggle.com/datasets/priyamchoksi/credit-card-transactions-dataset/code*

Proposed Work and Tools

1. Load data into a suitable coding environment
2. Clean data to populate missing fields and remove observations that are not able to be completed
3. Perform initial mining to find high-level correlations
4. Perform data reduction on unnecessary fields to reduce analysis workload
5. Review results of correlation analysis and visualize data via plot to better detect patterns that aren't obvious
6. Test various correlations to implement ML model to detect fraudulent transactions
7. Evaluate model performance compared to test sample set
8. Tune model performance as needed

Tools:

- Analysis: Pandas, NumPy
- ML algorithms: Scikit Learn
- Visualization: Matplotlib, Seaborn, Tableau

Dev Env: Jupyter Notebook
Github: github.com/brendan-mcdonald-cspb/4502GROUP10

Evaluation

To evaluate our results, we will hold-back a subset of the data prior to performing our analysis. This subset will serve as our sample to conduct testing on once we have our model developed.

If our model is working, it should be able to detect fraudulent transactions using customer data provided at the point of purchase.

We can compare model estimates to the binary flag that tells us whether or not the transaction was fraudulent, and tune our model as needed.