Detecting Credit Card Fraud
Milestone 1
Lincoln Brown
DSC680-T301
Professor Iranitalab

Topic

Detecting Credit Card Fraud Using Machine Learning Models. I work in the banking industry, where credit card fraud is an ongoing issue for both financial institutions and our customers. I want to use this project as an opportunity to investigate patterns in synthetic credit card transaction data to develop predictive models for identifying credit card fraud.

Business Problem

Identifying credit card fraud is critical for financial institutions. Credit card fraud has become a widespread problem, and I want to focus on several questions:

- Which geographic areas are associated with fraud the most?
- Does more fraud occur online or in person?
- Does time of day or transaction amount correlate with fraudulent activity?
- What machine learning models perform the best when predicting fraud?
- What other features are important for identifying fraud?

Dataset

I am using a synthetic dataset I found on Kaggle. The author is Erik Altman and contains 20 million transactions that were generated from a multi-agent virtual world simulation performed by IBM. There are several different csv files included in the downloaded dataset. I will primarily be focusing the transactions dataset but may incorporate the users dataset if I need to improve my model's performance by including more features (or different features).

There are 24,386,900 transactions in the dataset and it includes 15 features. The column names are UserID, CardID, Year, Month, Day, Time, Amount, Use Chip, Merchant Name, Merchant City, Merchant State, Zip, MCC, Errors?, and Is Fraud?. It will likely not be necessary to use all of these features, but several of them will be useful to identify both the timeframe of when fraud occurs and if there are areas that it occurs in the most (or online purchases). Furthermore, the synthetic dataset was created to emulate United States customers who also travel abroad, which will help provide diverse geographic information.

Methods

The Kaggle dataset page describes the use of TabFormer (Bidirectional Encoder Representations from Transformers (BERT)-based model), as well as a Long Short Term

Memory Model, and Gated Recurrent Unit (GRU) model. Another Kaggle user describes the use of a Graph Neural Network (GNN). Considering this, I will be investigating the use of Neural Networks on the dataset. I plan to implement a LSTM model and possibly GNN model as well. However, given my unfamiliarity with neural networks, I will also be implementing simpler baseline models such as Logistic Regression and Random Forest for comparison.

Ethical Considerations

I am confident that I do not have a lot of ethical considerations to worry about when working with this data. It does not need to be masked or otherwise protected because it is a synthetic model created specifically for the purpose of learning. However, due to the synthetic nature of this data, there may be bias present. With this in mind, it will be important to treat any solutions as educational experiences that would need to be vetted and retrained before being implemented in any real-world scenarios.

Challenges/Issues

I have not worked extensively with neural networks before, so I am a little leery of how smoothly the implementation will go. I will be doing research to learn more about these neural networks and how to implement them. Additionally, I could run into issues with computational resources given the size of the dataset. If I encounter computational restrictions, I will sample the data and work with a smaller subset of the data. I am also considering working only with the online transactions if necessary.

References

The original paper written by the dataset creator:
Altman, E. R. (2019, October 4). *Synthesizing credit card transactions*. arXiv.org. https://arxiv.org/abs/1910.03033

A Medium article on LSTM models:
J, R. T. J. (2024, March 6). LSTMs Explained: A Complete, Technically Accurate, Conceptual Guide with Keras. Medium. https://medium.com/analytics-vidhya/lstms-explained-a-complete-technically-accurate-conceptual-guide-with-keras-2a650327e8f2

A paper on Graph Neural Networks:
Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. S. (2020). A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(1), 4–24. https://doi.org/10.1109/tnnls.2020.2978386

I will also be using neural network walk-throughs and demonstrations as I progress through the project.