Credit Card Transactions: How to Detect Fraud with Apache Spark

# Matt B. Jackson DSC 680 - Summer 2020 https://mbjackson.github.io/

# References

Atharva. (2019, November 24). *How to build machine learning model at large scale with Apache Spark and LightGBM?* Retrieved from Towards Data Science: https://towardsdatascience.com/how-to-perform-credit-card-fraud-detection-at-large-scale-with-apache-spark-and-lightgbm-61a982de8672

Awoyemi, J. O., Adetunmbi, A. O., & Oluwadare, S. A. (2017). Credit card fraud detection using machine learning techniques: A comparative analysis. *2017 International Conference on Computing Networking and Informatics (ICCNI)*.

Brownlee, J. (2020, March 11). *Imbalanced Classification with the Fraudulent Credit Card Transactions Dataset*. Retrieved from Machine Learning Mastery: https://machinelearningmastery.com/imbalanced-classification-with-the-fraudulent-credit-card-transactions-dataset/

Gonzalez-Fierro, M. (2018, August). *Introduction to Fraud Detection Systems*. Retrieved from KDNuggets: https://www.kdnuggets.com/2018/08/introduction-fraud-detection-systems.html

Jacobs, K. (2017, June 15). *Machine Learning Fraud Detection: A Simple Machine Learning Approach*. Retrieved from Data Blogger: https://www.data-blogger.com/2017/06/15/fraud-detection-a-simple-machine-learning-approach/

Machine Learning Group - ULB. (2018, March 22). *Credit Card Fraud Detection*. Retrieved from Kaggle: https://www.kaggle.com/mlg-ulb/creditcardfraud

Madhavan, R. (2018, April 18). *Credit Card Fraud Detection using Python*. Retrieved from Kaggle: https://www.kaggle.com/renjithmadhavan/credit-card-fraud-detection-using-python

Nathan, L. S. (2018). *Train, Evaluate, Repeat: Building a Credit Card Fraud Detection System*. Retrieved from codetalks.tv: https://codetalks.tv/talk/train-evaluate-repeat-building-a-credit-card-fraud-detection-system-leela-senthil-nathan-v8sqikqq-pw

Patel, A. A. (n.d.). *Chapter 4. Anomaly Detection*. Retrieved from O'Reilly: https://www.oreilly.com/library/view/hands-on-unsupervised-learning/9781492035633/ch04.html

Sharma, A. (2019, October 26). *Credit Card Fraud Detection in Python using Scikit Learn*. Retrieved from Medium: https://medium.com/analytics-vidhya/credit-card-fraud-detection-in-python-using-scikit-learn-f9046a030f50

# Data Source

1. “Credit Card Fraud Detection” – Kaggle (Machine Learning Group - ULB, 2018)
   1. Dataset consists of 1 CSV files each with 31 variables and almost 250,000 rows. The entire dataset takes up 144 MB of data.
   2. https://www.kaggle.com/mlg-ulb/creditcardfraud

# Research Questions

1. How can a model be created in Apache Spark to detect fraudulent credit card charges?
2. What options are there to deploy this model into production? Which ones are optimized best for this application?

# Methods

In this project the dataset will be processed in Apache Spark. The reason for using Spark is that while this dataset isn’t huge, the application requires the ability to process large amount of data with ease. The classifier that will be used to determine if fraud exists is the LightGBMClassifier. This classifier is a gradient boosting one that utilizes a tree-based learning algorithm.

# Potential Issues

The biggest issue I think I would have with this project is Spark. I tried to use Spark for my last project and did not have the best results. Because of this I had to abandon the use of Spark and pivot to scikit-learn. Since the data here is numeric, it might pose less of a problem for me. Another potential problem is that the dataset is not balanced. Most of the data is legitimate transactions while fraudulent ones take up a small portion of the dataset.

# Concluding Remarks

As society continues to embrace “virtual” currency over physical money, they are lots of opportunities for criminals to do some bad acts. Credit card companies have been getting better at identifying potential instances of fraud by temporarily shutting off a credit card, alerting the customer about a transaction, or just preventing a transaction from happening without the user acknowledging it. The questions about all of these methods is how do the credit card companies know which transactions may be fraudulent? This project will look at how a machine learning model can be created in Apache Spark to identify any anomalous transactions.