Credit Card Fraud Detection using Data Science

Chandramouli Yalamanchili
DSC680 - T302 Applied Data Science (2215-1)
https://chandu85.github.io/DSC680-Site/

Which Domain?

Considering my work experience in banking industry, credit card processing to be specific, I chose to work on fraud detection in credit card transactions as my first project.

As I am looking to build a machine learning model to predict the fraud detection, I am going to review the work that is done already in this area and use the best options available to achieve high accuracy models to predict fraudulent transaction. Below are several references I am planning to refer as part this project:

1. Chan, P. K., Fan, W., Prodromidis, A. L., & Stolfo, S. J. (1999). Distributed data mining in credit card fraud detection. IEEE Intelligent systems, (6), 67- 74.
    - https://cs.fit.edu/~pkc/papers/ieee-is99.pdf
    - I will review this reference and see how they handled the skewed data for their experiment as credit card transaction data is highly skewed (legitimate transactions are very high in number compared to fraudulent transactions).
2. Brause, R., Langsdorf, T., & Hepp, M. (1999). Neural data mining for credit card fraud detection. In Proceedings 11th International Conference on Tools with Artificial Intelligence(pp. 103-106). IEEE.
    - http://sphinx.rbi.informatik.uni-frankfurt.de/asa/papers/ICTAI99.pdf
    - I will review this reference to gather details on what type of advanced data mining techniques they have applied as well what type of neural network algorithms they have used.
3. Ghosh, S., & Reilly, D. L. (1994, January). Credit card fraud detection with a neural-network. In System Sciences, 1994. Proceedings of the Twenty- Seventh Hawaii International Conference on (Vol. 3, pp. 621-630). IEEE.
    - http://bit.csc.lsu.edu/~jianhua/quang.pdf
    - This particular reference has some good information about P-RCE neural network that was used to implement FDS system in a financial institute. I will review this paper further to see if I can use any of the techniques mentioned as part of the machine learning model or neural network I will be building as part of this project.
4. Chan, P. K., & Stolfo, S. J. (1998, August). Toward Scalable Learning with Non-Uniform Class and Cost Distributions: A Case Study in Credit Card Fraud Detection. In KDD (Vol. 98, pp. 164-168).
    - https://www.aaai.org/Papers/KDD/1998/KDD98-026.pdf
    - This reference, similar to first one talks about dealing with the skewed data, I will review this to see if I can use any of the data distribution techniques they have used in their experiement.
5. Maes, S., Tuyls, K., Vanschoenwinkel, B., & Manderick, B. (2002, January). Credit card fraud detection using Bayesian and neural networks. In Proceedings of the 1st international naiso congress on neuro fuzzy technologies (pp. 261-270).
    - https://arxiv.org/pdf/1908.11553.pdf

    - This reference also concentrates on how to improve the minority sample in the input data as well as to reduce the noise. I will see if I can apply of these techniques into my project to achieve model with high accuracy.

6. Masoumeh Zareapoor, & Pourya Shamsolmoali (2015). Application of Credit Card Fraud Detection: Based on Bagging Ensemble Classifier. Procedia Computer Science (Vol. 48, pp. 679-685). https://doi.org/10.1016/j.procs.2015.04.201.
    - https://cyberleninka.org/article/n/324468.pdf
    - This paper has applied different mining techniques to compare the performance of different methods, I want to review this again as part of modeling process to ensure I use the right model for my project that would yield high accuracy.
7. Stolfo, S., Fan, D. W., Lee, W., Prodromidis, A., & Chan, P. (1997, July). Credit card fraud detection using meta-learning: Issues and initial results. In AAAI-97 Workshop on Fraud Detection and Risk Management.
    - https://www.aaai.org/Papers/Workshops/1997/WS-97-07/WS97-07-015.pdf
    - This research paper, similar to the previous one tested several machine learning algorithms as well as more importantly using the meta-learning strategies. I want to review this paper to get more understanding about the meta-learning strategies and see if I can use any of these strategies in my project as well as I have same data related issues as this project refers to.
8. Luke Sun (July 2020). Credit Card Fraud Detection.
    - https://towardsdatascience.com/credit-card-fraud-detection-9bc8db79b956
    - This article speaks about the data sampling issues as well as building different models to compare the performance of different modeling techniques. I will use this reference to build the sample data as well as to build the machine learning models.
9. Dorronsoro, J. R., Ginel, F., Sgnchez, C., & Cruz, C. S. (1997). Neural fraud detection in credit card operations. IEEE transactions on neural networks, 8(4), 827-834.
    - https://repositorio.uam.es/bitstream/handle/10486/663701/neural_dorronsoro_ITNN_1997_ps.pdf;jsessionid=28C549CC8D6DFFC1AB4F2A16D511F89F?sequence=1
    - This paper talks about the Minerva fraud detection system, a real time fraud detection system using neural networks. I want to review this paper to understand how they are using neural network for fraud detection. Also one more interesting factor for this paper is their mainframe implementation, I wanted to understand how they are integrating the model to the IBM mainframe components so that I can see if it's feasible for our application at work as well.
10. Sánchez, D., Vila, M. A., Cerda, L., & Serrano, J. M. (2009). Association rules applied to credit card fraud detection. Expert systems with applications, 36(2), 3630-3640.
    - http://didawiki.cli.di.unipi.it/lib/exe/fetch.php/dm/ar-creditcard-fraudedetection.pdf
    - This paper reviews the use of association rules to determine fraudulent transactions. This paper probably will not be of much help for this particular project as most of the data I have is transformed. But I still wanted to review to see if I can gather any of the insights from this work for my project, also to see if this is something I can implement in my work environment.

Which Data?

I am planning to use the below dataset from Kaggle for this project.

Dataset Link - https://www.kaggle.com/mlg-ulb/creditcardfraud

- This dataset has only few columns in clear, rest of the columns have been PCA transformed due to the confidentiality of the data.
- Time, Amount and Fraud indicator are the columns that are in clear, this dataset has remaining 29 columns or features that had gone through PCA transformation.

Research Questions? Benefits? Why analyze these data?

- We currently have multiple fraud detection tools and most of them highly dependent on manual intervention to build and maintain for more efficiency in catching fraud.
- Through this project, I want to research and see how data science and machine learning can help in this area to catch the fraud automatically with very less to no need of human supervision.
- Credit card fraud has always been one of the major concerns for financial institutions as it could them financially in term of penalties as well as impact their reputation.
- Being able to detect fraud efficiently in real time with less human support would be a great help for the financial institutions.

Method?

- I am planning to use Python for this project.
- I will initially do some data visualization to understand any trends I can derive out of the data.
- Evaluate the data and apply data engineering as needed to be used for modelling.
- Build the deep learning model to be able to predict the fraudulent transaction based on features provided.

Potential Issues?

- Having the transformed and normalized column data for most of the features would restrict me to a very few data visualizations I can derive from the data. I am planning to build my visualization using only the three columns in clear as the visualizations with other features would not add any value when we don't know the attribute behind that feature.
- We only have 284,807 transactions as part of this dataset, that might not be enough to train the model to achieve maximum performance. So I might have to find additional datasets or apply different data engineering and modeling techniques to make the most out of the data available.

Concluding Remarks

We have witnessed an enormous evolution in credit card processing over last few years, issuing chip-based credit cards, starting mobile device-based wallets like Apple Pay is a significant change done to secure credit card transactions.

Despite financial institutions (banks) working hard to eliminate fraud in credit card transactions, credit card fraud has been continuously rising over the last few years. Fraudsters are getting smarter and using latest technologies to steal cardholder's information, either through hacking or through social engineering.

Increasing fraud in the industry makes fraud prediction very critical to be able to identify and stop fraud in real time, and data science plays a significant role in analyzing and being able to predict fraud based on transactional and cardholder information. The scope of this project is to research and identify different types of predictive analysis algorithms available that can be applied to determine and stop fraudulent transactions.