

# ***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. Below are several references I will be using for 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.
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.
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.
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).
5. Dhok SS, Bamnote GR. Credit Card Fraud Detection Using Hidden Markov Model. *International Journal of Advanced Research in Computer Science*. 2012;3(3):816.
6. 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).
7. 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>.
8. 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*.
9. 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.
10. 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.

## **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.