# Reflective Essay on Combination of Machine Learning Techniques and Oversampling Ratios for Credit Card Fraud Detection: A Comparative Analysis

## Introduction

The journey of researching and delivering my MSc Project has been both enlightening and challenging. Looking into complexity of fraud detection using machine learning techniques provided not just an academic challenge but also a opportunity to appreciate the real world implication of these technologies. The objective this research for the better choice of algorithm for credit card fraud detection is crucial, as it's impact is directly related to consumer and financial institutions globally. However, as it is with any complex studies, the pathway to understanding and mastering it is never easy and comes with numerous hurdles, new learning opportunities and various moments of self observation.

This essay aims to explore my personal and academic journey through this project, from the initial sparks of curiosity to the final implementation of machine learning models and their evaluation. Through this essay, I aim to portray my experiences, challenges faced, skills acquired and learning outcome of this research.

## Motivation and Objective

Primarily my motivation arose during my second semester course on "Data Analytics", where I was introduced to the domain of data science and machine learning. The concept of an algorithm's capability to process data, train itself to recognize patterns in the data, learn from it and subsequently make predictions based on it, captivated me. Realizing the potential or machine learning compelled me to embark on a project that focuses on its principles and applications. My goal was to gain a deeper insight and broaden my understanding of this field by the end of this research.

Choosing an appropriate theme proved challenging for this project initially. So I had looked through online suggestions for interesting works possible with machine learning. To overcome this, I explored various resources online for potential application of machine learning. During this research, I came across the concept of credit card fraud detection leveraging machine learning techniques. While this idea initially appeared straightforward, primarily involving the development of a predictive model utilizing algorithms, I came across a study by Y. Jain and N. Tiwari and others (2019)[1]. This study is a comparative analysis of multiple machine learning techniques and assessments of their efficacy in fraud detection. However, I inserted an additional element by testing each algorithm at different oversampling ratios using the Synthetic Minority Oversampling Technique (SMOTE). It is noteworthy that for my research, I opted for a synthetic dataset, as crafted by Kartik Shenoy, utilizing a tool developed by Brandon Harris.

Beyond technical aspects, personally I hoped to challenge myself to dig deeper into the realm of machine learning and grow as a researcher. Exploring this topic allowed me to utilize my personal interest and academic experience. I am eager to see where this journey takes me.

*Acknowledgements*

*I would like to extend special thanks to both Brandon Harris and Kartik Shenoy for their invaluable contributions to the materials utilized in this project.*

## Ethical Considerations

For this research I remained careful of ethical issues. Since the moment I chose the topic of credit card fraud detection it introduces ethical concerns. Original datasets would contain real details of real people, getting permission to utilize such dataset can be tricky, time consuming and in certain cases impossible. Given this issue I opted for a synthetic dataset I found on Kaggle.com[2]. This makes sure this research does no unintentionally result in the misuse of any sensitive information.

All algorithms used in this research were derived from Scikit learn's online documentations and some other online sources which are linked and highlighted within the notebook. Scikit Learn is a python library used for implementing machine learning models and statistical modelling. The goal with adopting widely accepted and well documented methods was to maintain the integrity and openness of my research process.

If not used carefully, machine learning models, especially in credit card fraud detection, can unintentionally reinforce existing biases. Even though the dataset I used was synthetically generated. The bias we see in this dataset leans toward the genuine transaction. This is because fraudulent transaction usually tend to be a very small fraction of the majority. Integration of SMOTE was primarily to handle this issue.

## Challenges and Limitations:

During the course of my research, I encountered a significant setback due to a severe facial injury. Due to an unfortunate event, I was compelled to undergo facial surgery, specifically around my eye area. The aftermath of the surgery left me in significant swelling, causing a misalignment in my vision. This condition rendered me incapable of reading or writing any material for a duration of slightly over two months, as attempts to do so induced severe headaches, watery eyes, and doubled vision even after the swelling had gone down externally. According to my doctors' advice, this was due to the prolonged healing process of the internal swellings. As a consequence, the timeline for my project was severely shortened, requiring a more accelerated pace of work upon my recovery.

The selection of the specific dataset introduces some constrains. Primarily, I am using the test dataset for both training and testing the models within the scope of this research.. This decision stems from a few reasons. With the current dataset the some algorithms, notably Gradient boosting and Random Forest already take a significant amount of time to run. Combine that with 5 fold cross validation at every single oversampling ratios used, introduces algorithm and cross validation run-time combination of over 30 minutes at a single over sampling ratio. Given these time constraints, the choice to utilize a single dataset for both training and testing was deemed more time-efficient, albeit with some limitations. I acknowledge this methodology can potentially influence the performance of the models. but due to the limitation of time I chose to go with what I thought would be a more time efficient approach. Furthermore, I understand that using synthetic data, while alleviating privacy concerns, also brings its set of limitations, notably regarding the generalizability of results which I clarified in the research paper to the best of my ability.

## Outcomes and Insights

As I navigated through this research, I found several outcomes and personal insights that reshaped my understanding of both the subject and my capabilities as a researcher. One of the foremost realizations was the role of preprocessing in machine learning. While raw data can be misleading and unmanageable, preprocessing can transform such datasets into insightful resources. And visualization of outliers in the project highlights the need for such measures. Primarily I selected models primarily based on accuracy metrics. However as I got deeper into the research I started seeing the complexities that come into play, such as overfitting and understanding the mechanics of the algorithms. It was enlightening to see how different SMOTE oversampling strategies could impact the results of the same algorithm.

During this project I came across numerous instances where I found myself confronted with challenges, particularly when adjusting oversampling strategies. Deciding on the ideal strategy required multiple iterations, and each decision had to be weighed against its potential impact on the results. Which is why I chose over sampling ratios at an increment of 0.2 apart from the ration 1 which skips an iteration. This decision was also due to the time limitations mentioned previously as originally my plan was to re-iterate the algorithms in each oversampling ratios on an increment of +0.1.

Importance of cross validation was another was another significant takeaway. For this research I used 5-fold cross validation technique. While a higher fold cross validation would provide more insight, given the computational intensity and run-time duration with some of the algorithms previously discussed, I determined that a 5-fold approach was both practical and appropriate for the academic context of this research.

## Future Perspectives

The iterative nature of machine learning, along with nature of the research often required me to go back and forth and re-asses my decisions. On a personal level, this experience improved my problem-solving skills and expanded my knowledge in machine learning further. Although, the scope of my research extends to comparing results of different algorithms at different sampling ratios. I identified several techniques to boost the performance of this research even further. These techniques include using the separate training and test datasets to run the algorithms, refine implementation of SMOTE at all ratios 0.1 – 1.0 at an increment of +0.1 as well as test other types of oversampling or resampling techniques like “RandomOverSample” and use a higher fold cross validation and finally include more machine learning algorithms.

Reflecting on this project, I recognize the vastness and complexities of this field. Although I learned various new techniques when working on this project, I realize I still have a long way to go If I am ever to master this field. For which I am eager to take forward what I learnt from this research and continue to work on it and expand the research with the techniques mentioned previously.

## Conclusion

In conclusion, this reflective process itself has been insightful as it has allowed me to go over the span of my effort to both appreciate and criticize my efforts and recognize the valuable insights I have acquired on the way. Moving forward, I can carry with me knowledge of the machine learning techniques their implementation on fraud detection, effects and necessity of over sampling in relation to fraud detection that I gained from this research, which hopefully will inspire my future academic and professional pursuit.

## References


1. Jain, Y. & Tiwari, N. & Dubey, S. & Jain, Sarika. (2019). A comparative analysis of various credit card fraud detection techniques. International Journal of Recent Technology and Engineering. 7. 402-407.
2. Dataset Created by Kartik Shenoy using Brandon Harris's tool. Link to dataset: https://www.kaggle.com/datasets/kartik2112/fraud-detection