**Machine Learning Techniques for Credit Card Fraud Detection: A Comparative Analysis**

**Zami Nizam Uddin**

**Student ID: 221057490 Submission Date:**

**MSc Computing and Information system final project**

**Queen Mary university of London, 327 Mile End Rd, Bethnal Green, London E1 4NS**

**Abstract:**

This research paper enters upon a comparative analysis of distinct machine learning algorithms: Logistic regression, Random Forest, Gradient Boosting, Gaussian Naïve Bayes and K-Nearest Neighbors (k-NN) for credit card fraud detection. After preprocessing a dataset of, each algorithm was trained and evaluated for both training and test sets, to assess for potential overfitting. A major aspect of this study is the use of “Synthetic Minority Over-Sampling Technique” or SMOTE with varying oversampling ratios (1.0, 0.7, 0.5, 0.3, 0.1) to address the class imbalance between fraudulent and non-fraudulent samples, which is inherent in a fraud detection dataset. Each algorithm’s performance was thoroughly validated using 5-fold crass validation technique for each oversampling ratio. The research further enhances understanding of the models through visual integration of confusion matrices, Precision-recall curves and accuracy plots for each algorithm as well as make prediction on unseen data. This helps provide insights into real-world applicability. The goal of this research is to guide the choice of the most effective machine learning algorithm for credit card fraud detection, clarifying a theoretical understanding of primary application in digital financial security.

**Table of Contents**

**1. Introduction**

**1.1. Background:**

With the rapid growth of digital transaction worldwide, the integrity and security of financial data has become ever so crucial. Use of credit cards are on a rise as they introduce incredible convenience and benefits, however, comes with possible threats of fraud. Therefore, introduces an imperative need to explore advanced methodologies that are both adaptive and efficient.

Machine learning is a subset of artificial intelligence, it offers promising solutions to such problems. Leveraging different machine learning models to learn patterns in vast data, machine learning models cam potentially identify unusual and potentially fraudulent activities with higher accuracy than traditional systems like Blacklisting, Manual verification, velocity checks etc. which have been foundational in detecting fraud but with the evolving nature of fraud, the system needs to be a more adaptive and sophisticated solution like machine learning. However, every solution comes with its own challenges. One of the main and significant challenge in terms of credit card fraud detection is imbalance of fraudulent and non-fraudulent classes in a dataset as fraudulent cases usually tend to occur less frequently in comparison to genuine ones. Such imbalances can cause algorithms to perform well in terms of accuracy but perform poorly in recall causing it to likelier to miss most fraudulent transaction.

This research dives into this problem by comparing various machine learning techniques in their ability to detect fraudulent transactions. This can be improved by balancing the data for which the use of various Synthetic Minority Over-Sampling Technique (SMOTE) ratios proved to be crucial. Further, use of cross validation for each algorithm on each over sampling ration helps highlight the best combination of oversampling ratio and machine learning algorithm that offers the best accuracy for detecting fraudulent transaction.

In further sections of this research, I will be developing the complexities regarding machine learning models and their significance in their detection accuracy and the importance of handling imbalanced data. By the end I hope to establish a clear understanding of the most potent among the selected algorithms to tackle this problem.

**1.2. Objective:**

Primary objective of this research is to conduct a thorough, comprehensive and comparative analysis of multiple machine learning techniques naming: Logistic Regression, Random Forest, Gradient Boosting, Gaussian Naïve Bayes and K Nearest Neighbors (K-NN). For this research I am using a simulated dataset I found on “kaggle.com”, the aim is to probe into the effects of different SMOTE over sampling ratios on the performance of the algorithms. Through this research my goal is to determine the optimal machine learning technique and over sampling ratio that yields the highest predictive accuracy, precision and recall. Through this investigative research the aim is to offer a dependable and efficient model recommendation for real world application in terms of credit card fraud detection.

**1.3. Scope:**

The extent of this research covers:

* Dataset preprocessing: For this research I’m using a simulated dataset that represents typical credit card transactions. The dataset being simulated ensures confidentiality while replicating a real-world transactional pattern.
* Analyzed Machine learning Techniques: For this research I selected 5 machine learning models to examine. They are:
  + Logistic Regression
  + Random Forest
  + Gradient Boosting
  + Gaussian Naive Bayes
  + K-Nearest Neighbors (K-NN)
* handling data Imbalance: In this research I will probe into the challenges posed by severely unbalanced datasets, which tends to be a common issue in fraud detection. I am using a technique called “Synthetic Minority Oversampling Technique” in short SMOTE to address the issue. The analysis is done on synthetic oversampling ratios of 1.0, 0.7, 0.5, 0.3, 0.1.
* Evaluation Metrics: Each model is evaluated using metrics such as Precision-Recall, f1-Score, accuracy and confusion matrix. Further each model is evaluated on both training and test sets to detect potential overfitting.
* Visualizations: Outliers, imbalance and evaluation metrics are visualized to give a clearer picture of the models
* Model Validation: To ensure dependability of the algorithms, each algorithm is cross-validated (5-folds) on each oversampling ratio.
* Predictive Capability: Beyond the analysis the models are also used to predict on new unseen data in order to showcase potential real worlds application.

This research does not extend to other machine learning models apart from the ones mentioned previously, datasets other than simulated one used, oversampling technique beyond SMOTE and real-world deployment or integration of the model into a transaction system.

The aim here is to provide a thorough comparison of the selected algorithm’s performances in context of credit card fraud detection and on the efficacy of different SMOTE oversampling ratios. The scope ensures a focused exploration of the topic with the potential to lead to actionable insights and recommendations.

**2. Literature Review**

**2.1. Credit Card Fraud Detection:**

Credit card fraud detection has always been a pressing challenge in the modern banking and financial sector. Because of the fast digitization of the financial world, opportunities for malicious activities have also increased. Traditional credit card fraud detection system depended on rule-based systems which uses predefined rules based on known fraudulent patterns to flag potential or suspicious transactions. For example, a Cumulative transaction system; where a rule can dictate if a transaction for a particular account exceeds a certain amount within a certain period it would trigger an alert to catch potential fraudulent transaction. But in such a scenario a genuine transaction meeting the same points most likely will also trigger an alert which will in turn can cause problems for a genuine user of a transaction service. However, these systems, while adept at detecting known fraud types, often faltered when encountering novel fraudulent tactics (Sahin, Bulkin, & Duman, 2013).

Realizing the limitations of rule-based systems, researchers are studying more dynamic approaches like machine learning and data driven methods. These techniques can detect patterns within massive datasets, this enhancing the detection’s accuracy (Bhattacharyya et al., 2011).

The aspect of data imbalance is a prevailing issue for the literature as it is inherent in a dataset for fraud detection as the amount of fraudulent transaction tends to be much lower than genuine ones. This is a problem because it can introduce potential bias in a model’s predictive capability. To deal with such issues researchers have developed various oversampling techniques, one of which is SMOTE, which is represented during the fitting and training of the data (Chawla et al., 2002).

In summary the field of credit card fraud detection has transformed significantly and for better from static rule-based systems to more dynamic and sophisticated approaches like machine learning. However, the issue with handling data imbalance and real time detection remains areas of active research.

**2.2 Machine Learning Techniques in Fraud Detection:**

**References**

* Sahin, Y., Bulkan, S., & Duman, E. (2013). A cost-sensitive decision tree approach for fraud detection. Expert Systems with Applications, 40(15), 5916-5923. https://doi.org/10.1016/j.eswa.2013.05.021
* Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, J. C. (2011). Data mining for credit card fraud: A comparative study. Decision Support Systems, 50(3), 602-613. https://doi.org/10.1016/j.dss.2010.08.008
* Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357. https://doi.org/10.1016/j.eswa.2013.05.021