Credit Card Fraud Detection

Jashwanth Neeli - 2114682

DePaul University

Abstract:

The rise of digital banking has increased the occurrence of credit card transactions by multiple magnitudes. Coincidentally, there has been an acute escalation in the number of fraudulent activities, a major concern for the relevant parties. Despite the improved sophistication of security applications, skilled fraudsters continuously breach these boundaries with immaculate precision. This project seeks to eliminate the risks associated by incorporating models designed through machine earning models that can detect and predict occurrence of fraudulent transactions in real-time to determine the likelihood of such occurrences. The review of class-imbalance techniques and hyperparameter tuning is based on insights determined through customer-oriented analyses on data from Worldline and Machine Learning Group partnership to increase predictive models' confidence. It is essential for one to strive for a balance between improve detection capability and operational efficiency to deter the threat amplifying rich banking systems and destroying customer trust. This way, the financial industry will be able to suppress the threat animated by the growing number of transactional frauds.

Introduction:

The financial sector has undergone a complete overhaul due to the inclusion of digital technology, significantly bettering the convenience and efficiency of transactions. However, this has also led to an additional downside level of exposure to fraudulent activities, especially on multiple fronts of credit card transactions. Just in the fiscal year 2019, the number of reported credit/debit card fraud cases was at over 50,000, painting a grim picture of the urgent necessity for a powerful detection system. Therefore, this project tackles the grand challenge of credit card fraud by using advanced machine learning methods. Given the high cost so far, the goal is to minimize the overall risk of fraud not just to the monetary value but the score of lost confidence and brand affiliation. We deploy a wide variety of machine learning models to test the models' prediction results, including but not limited to the precision, recall, and the time to compute. We, however, realize that due to the poor observation of the fraudulent instances, we are sought to balance the class, called class imbalance. Therefore, we use the Random Over Sampling and Synthetic Minority Over-sampling Technique to fulfill the above-mentioned objective. Eventually, we understand the effort to fine-tune the model to improve performance.

Dataset:

The dataset used in this project was borrowed from Kaggle, a well-known platform that provides a wide range of datasets and challenges that support data science innovations. The dataset was composed of a large number of credit card transactions, specifically 284,807 transactions that were completed in September 2013 made by several European cardholders. It is important to point out that the dataset was heavily unbalanced. Only 492 transactions or 0.172% of the dataset contained fraudulent activity. It should be noted that this proportion is extremely rare, and fraud is very unlikely to happen. One of the key characteristics of the dataset is the anonymization of features by the users of the PCA. For reasons of confidentiality, the dataset contains many transformed numerical values named V1, V2, V3…V28. The features introduced as 'Time' define the number of seconds that have gone by between each transaction and the first transaction recorded in the dataset. The second feature introduced as 'Amount' is the price of the transaction. This feature can be used by prediction models as some of the models consider transaction prices for their predictions. The target variable in this dataset is the 'Class' variable, which can be found in two formats: 0 presents valid transactions, and 1 presents fraudulent transactions. As I have discussed earlier, there are 0.172% fraudulent transactions in this dataset, which will affect the response rate of the models. Therefore, in order to evaluate the performance of the model, the Area Under Precision-Recall Curve is used in this project.

Addressing the Core Challenges In Fraud Detection Imbalanced Data: Unsurprisingly, as is the case for many classification tasks, unbalanced data was a critical challenge. Our project addressed this by using techniques such as random oversampling and SMOTE to balance the number of instances of fraudulent transactions in the dataset. This approach sensitized our model for any existing patterns in other features, which in turn dampened the bias towards the majority class for all our models otherwise applied, i.e., Decision Trees, Random Forests, and XGBoost. Anonymized Features: The low-effort provided features by PCA transformation constrained the number of suitable models that could be considered to analyze the data. Indeed, predictive models capable of deriving meaningful patterns out of thin air, as opposed to other models that demand prior experience and expertise in the domain, were explored. The Decision Tree could have been used, albeit more deliberate, in addition to Random Forests and XGBoost for their feature importance analysis capabilities, crucial in determining the most actionable anonymized features. Accuracy and Fast Detection: In fraud detection, a trade-off between accuracy and time is present without negating one completely. Considering such an argument, we used AUC-ROC curves to compare the effectiveness of all our models for a range of ROC thresholds. A Decision Tree with depths of 4 and 8 and Gini and Entropy as the criteria was used. Random Forest and XGBoost used one and two hyperparameters, the learning rate adjusted in the latter. Cost-Sensitivity of Errors: The asymmetric cost distribution between false positives and negatives necessitates a rigorous consideration. Because of such complications, we used cost-based evaluated precision, recall, and F1 score comparison of our model's AUC-ROC deductions. Such a consideration is also applied in the determination of the best parameters in hyperparameter tuning, especially for XGBoost's learning rate and decision tree's criteria. Scalability and Real Time Prediction: Finally, a holistic foundational consideration for any machine learning in data was the lack of real-time prediction consideration. Decisions were made on our models' fitting speed without any predictive speed comparison between models: Decision Tree, despite tactful change for speed, and Random Forest and XGBoost in its speed and complexity adjustment measures. The overall decision on the applicability of our solution to achieve success concerning the number of fraudulent transactions was subjective.

Literature Review:

The document focuses on the challenges and techniques applied in solving credit card fraud detection. It includes problems such as data imbalance, non-stationarity, and more assessment strategies. The researchers based the research on the models using machine learning, testing them with a variety of original datasets, and stressing the necessity of model adaptation for changes in data aggregates using special balancing methods. Focused strategies – such as the combination of information flows in real and preliminary data and the balancing theory, viz. SMOTE and EasyEnsemble – have proven fade performance in fraud detection. The paper firmly verifies the sophistication maturing the dynamic aspects of the fraud detection and the importance of incremental training and batching.

The document has critically presented an in-depth explanation and recommended measures for improving credit card fraud detection systems. These involve the need to tackle class imbalance, concept drift, and verification latency. It further identifies machine learning algorithms as a perfect fit that efficiently adjusts to the evolving patterns in transaction data. The research further introduces a new learning algorithm that uses feedbacks and the delayed supervised samples to boost the accuracy of the detections. The empirical validation of their method using actual datasets guarantees the applicability of their approach in non-stationary settings and the sampling bias instigated by interaction with the alert-feedback.

The "Credit Card Fraud Detection" dataset hosted by Kaggle, offered by the Machine Learning Group of ULB, contains transactions by European cardholders in September 2013. This dataset enables work on an extremely imbalanced data that represents the transactions that occurred over two days, where the positive class consists of the fraudulent transactions (492 cases). It encompasses 284,807 transactions, which occur in two days, and include 492 that are fraud cases. The dataset consists of anonymized features obtained from a PCA, in addition to the 'Time' and 'Amount' features, and the 'Class' feature as the label.

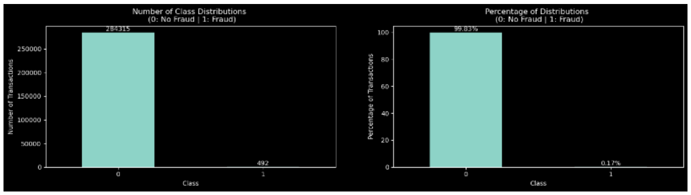
This article is about the need for accurate detection of fraudulent transactions by systems to prevent massive financial loss. It also addresses the difficulty caused by the diverse characteristics of fraud behavior by different types of transactions and domains. The paper introduces two domain adaptation methods in a deep neural network (DNN) framework, targeting the transfer of models from learning e-commerce transactions to face-to-face (F2F) transactions. These two methods are tested on an extensive dataset of over 80 million transactions. The results suggest how transfer learning can improve fraud detection within any payment context.

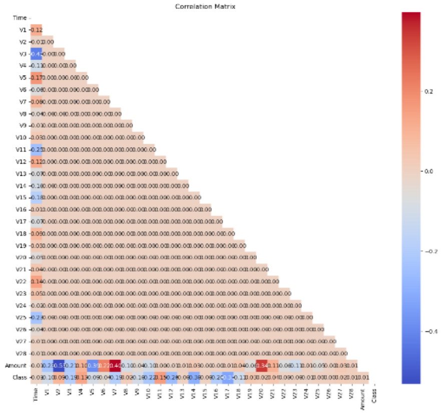
The document contains a full review of credit card fraud detection research, outlining the key challenges of data unbalance and non-stationarity and the choice of performance indicators. It covers a wide range of machine learning methods and techniques applied to the relevant datasets, underlining the significance of re-learning provided matches are truly random and data balance thanks to undersampling, SMOTE, EasyEnsemble, and other methods. The paper also supports the relevance of adaptive learning that takes parameters based on each new example and precision of the decision with the inability of fraud departments to verify as many fraud alerts as they produce.

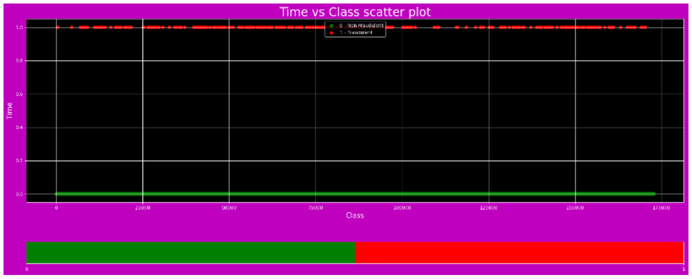
Methodology:

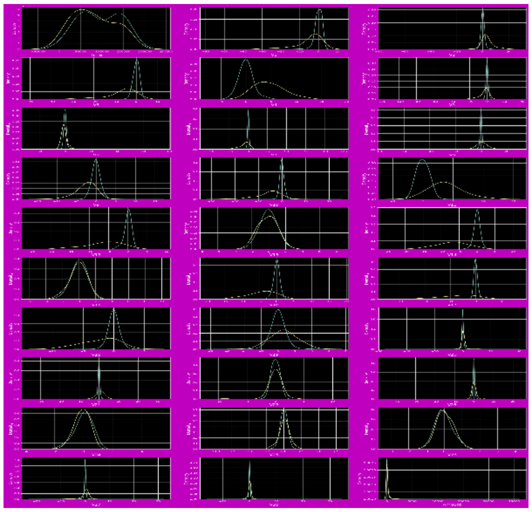
I checked the column names in a dataframe to understand the structure and attributes of the dataset used for credit card fraud detection. I printed the data types of columns, the shape of the dataframe, and displayed detailed information about the dataframe, including the count of columns, count of non-null values, and datatype of each column. I also checked for the presence of any null entries in each column and calculated the total number of missing values and their percentage relative to the entire dataset. No missing values are present in the dataframe. Since our entire dataset was transformed with PCA, I am assuming that the outliers have already been addressed. Therefore, I will not be performing any outlier treatment on the dataframe, even though outliers are still observable.A collection of density plots is generated to show the distribution of values for each feature in the dataset. Each plot represents one feature and includes two distributions: one for non-fraudulent and the other for fraudulent transactions. Some features exhibit distinct distributions, potentially aiding in the differentiation between fraudulent and non-fraudulent transactions. Conversely, other features display overlapping distributions, indicating they might not offer precise discriminative power in models. Notably, the density plot for the "Amount" feature reveals a significantly skewed distribution, particularly for non-fraudulent transactions. This analysis suggests that most features have overlapping distributions for both fraud and non-fraud transactions, pointing to the complexity of distinguishing between the two based solely on individual feature distributions.

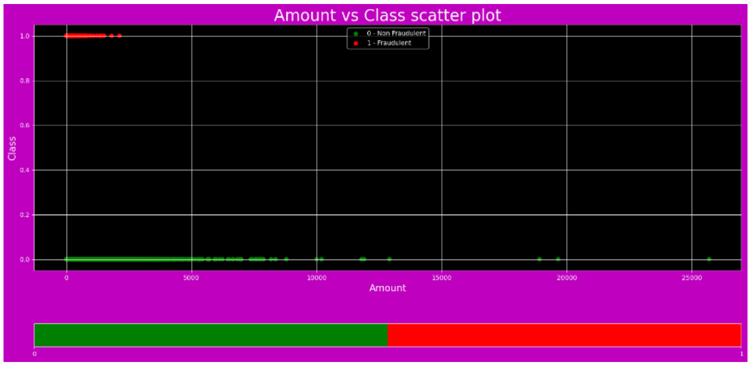
Most features peak at certain values, with frequencies diminishing as values deviate from these peaks. The histogram for "Amount" showcases a right-skewed distribution, indicating smaller transaction amounts are more common. These patterns help uncover underlying data trends and could indicate normal or aberrant transaction behaviors. Many features exhibit high skewness, prompting the use of skewness checks and potentially power transformations for normalization, especially for features like V1, V3, V5, V6, V8, V13, V15, V20, V21, V22, V23, and "Amount." Such skewed features can introduce bias in machine learning models, which perform better with normally distributed data. Employing power transformations may improve data distribution, enhancing model accuracy.

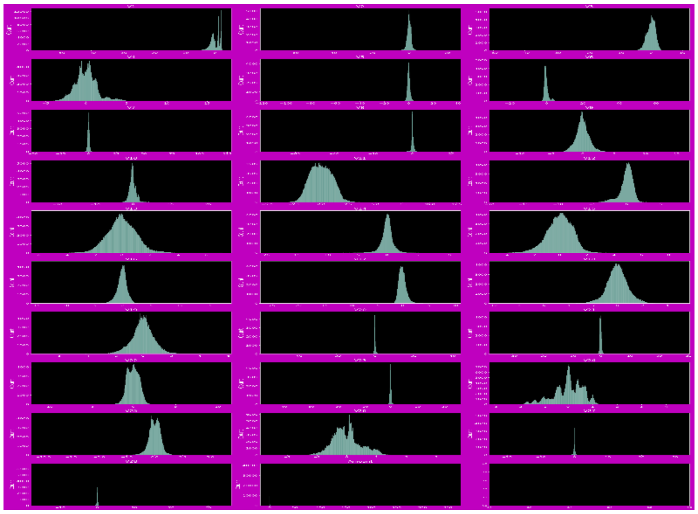

Number of Class Distributions
(0: No Fraud | 1: Fraud)
284315
492
Number of Transactions
250000
200000
150000
100000
50000
0
0
1
Class
Percentage of Distributions
(0: No Fraud | 1: Fraud)
99.83%
0.17%
Percentage of Transactions
100
80
60
40
20
0
0
1
Class


Correlation Matrix
Time
V1
V2
V3
V4
V5
V6
V7
V8
V9
V10
V11
V12
V13
V14
V15
V16
V17
V18
V19
V20
V21
V22
V23
V24
V25
V26
V27
V28
Amount
Class
0.2
0.0
−0.2
−0.4


Time vs Class scatter plot
0 - Non Fraudulent
1 - Fraudulent
Time
Class
0
25000
50000
75000
100000
125000
150000
175000




Amount vs Class scatter plot
0 - Non Fraudulent
1 - Fraudulent
1.0
0.8
0.6
0.4
0.2
0.0
Class
0
5000
10000
15000
20000
25000
Amount
0
1



Results

When trained on random oversampled and SMOTE balanced datasets, all models outperform the unbalanced dataset in terms of performance metrics (ROC-AUC, F1-Score, Precision, and recall).

Random oversampling and SMOTE increase performance, particularly in terms of recall and precision, showing a greater capacity to properly categorize minority class occurrences (fraudulent transactions).

XGBoost consistently beats Decision Tree and Random Forest on all datasets, proving its ability to handle class imbalance and generalize effectively to new data.

Random Forest and XGBoost trained on random oversampled and SMOTE balanced datasets had greater Recall values than the unbalanced dataset, indicating that they can detect more instances of fraudulent transactions. While Random Forest and XGBoost perform similarly on random oversampled and SMOTE balanced datasets, XGBoost has somewhat higher Precision and Recall scores, showing superiority in handling unbalanced datasets.

Model	**Dataset**	**ROC-AUC**	**F1-Score**	**Precision**	**Recall**
Decision Tree	Unbalanced Data	0.9314	0.8200	0.8039	0.8367
	Random Oversampling	0.9485	0.9194	0.9286	0.9379
	SMOTE	0.9485	0.9194	0.9286	0.9379
Random Forest	Unbalanced Data	0.9637	0.8200	0.8039	0.8367
	Random Oversampling	0.9798	0.7504	0.8573	0.9998
	SMOTE	0.9798	0.7504	0.8573	0.9998
XGBoost	Unbalanced Data	0.9698	0.8241	0.8119	0.8367
	Random Oversampling	0.9861	0.8702	0.9299	0.9985
	SMOTE	0.9861	0.8702	0.9299	0.9985

what worked and what didn't:

There are indications of overfitting in some cases, especially with Decision Tree models trained on unbalanced data:

1. Decision Tree – Unbalanced Data: The first model was Decision Tree, and the ROC-AUC score was noticeably lower on the testing score compared to the training score (0.9314 vs. 0.9874). Furthermore, the F1-Score, Precision, and Recall were also much lower, which could indicate overfitting. The way to overcome overfitting might include using regularization methods, such as restriction to the maximum depth of the decision tree or using pruning techniques.

2. Random Forest – Unbalanced Data: The second model used was Random Forest. The ROC-AUC score was lower on the testing score compared to the training score (0.9637 vs. 0.9630).

Therefore, regularization methods could include minimizing the minimum samples a leaf can have or restricting the maximum number of features randomly selected.

3. XGBoost – Unbalanced Data: The third model was XGBoost, and the performance was relatively stable between the training and testing datasets. Nevertheless, the performance on the testing set was lower – this may warrant the need for similar regularization methods to prevent possible overfitting as the model becomes more complex.

If things failed, talk about why they perhaps failed and what could be tried in the future.

Imbalanced data distribution: The possible future would be exploring more advanced techniques for handling class imbalance, including more sophisticated sampling methods like SMOTE, ADASYN, and cost-sensitive learning.

Hyperparameter tuning: Suboptimal hyperparameters can have a significant effect on the performance of the models. If hyperparameters are not properly tuned, they may fail to capture the complex relationships of the data or even overfit a training set. Possible strategies for hyperparameter tuning, like exhaustive search grid search or randomized search, could be employed to find the optimal settings that maximize model performance.

In conclusion, the analysis of model performance on unbalanced data, random oversampled data, and SMOTE balanced data offers insightful information regarding the challenges and opportunities in credit card fraud detection. Addressing class imbalance is crucial, as it significantly impacts the performance of machine learning models, especially for tasks targeting the minority class, like credit card fraud detection. The presence of class imbalance notably undermines model performance. Techniques such as random oversampling and SMOTE have shown effectiveness in mitigating class imbalance, enhancing model accuracy.

Moreover, the selection of the appropriate model is paramount. Different machine learning models show varied levels of performance on imbalanced datasets. The choice of models, including decision trees, random forests, and gradient boosting methods like XGBoost, should be based on the dataset's specific characteristics and the task at hand to ensure the optimal balance between performance and computational efficiency.

Furthermore, relying solely on accuracy as an evaluation metric is insufficient in the context of imbalanced classes. Other metrics like ROC-AUC, Precision, Recall, and F1-Score offer a more nuanced understanding of a model's capacity to accurately classify fraudulent transactions while minimizing false positives. These insights underscore the importance of adopting a comprehensive approach in model evaluation to effectively combat credit card fraud.

Reference:

Andrea Dal Pozzolo, Olivier Caelen, Reid A. Johnson and Gianluca Bontempi. Calibrating Probability with Undersampling for Unbalanced Classification. In Symposium on Computational Intelligence and Data Mining (CIDM), IEEE, 2015

Dal Pozzolo, Andrea; Caelen, Olivier; Le Borgne, Yann-Ael; Waterschoot, Serge; Bontempi, Gianluca. Learned lessons in credit card fraud detection from a practitioner perspective, Expert systems with applications,41,10,4915-4928,2014, Pergamon

Dal Pozzolo, Andrea; Boracchi, Giacomo; Caelen, Olivier; Alippi, Cesare; Bontempi, Gianluca. Credit card fraud detection: a realistic modeling and a novel learning strategy, IEEE transactions on neural networks and learning systems,29,8,3784-3797,2018,IEEE

Dal Pozzolo, Andrea Adaptive Machine learning for credit card fraud detection ULB MLG PhD thesis (supervised by G. Bontempi)

Carcillo, Fabrizio; Dal Pozzolo, Andrea; Le Borgne, Yann-Aël; Caelen, Olivier; Mazzer, Yannis; Bontempi, Gianluca. Scarff: a scalable framework for streaming credit card fraud detection with Spark, Information fusion,41, 182-194,2018,Elsevier

Carcillo, Fabrizio; Le Borgne, Yann-Aël; Caelen, Olivier; Bontempi, Gianluca. Streaming active learning strategies for real-life credit card fraud detection: assessment and visualization, International Journal of Data Science and Analytics, 5,4,285-300,2018,Springer International Publishing

Bertrand Lebichot, Yann-Aël Le Borgne, Liyun He, Frederic Oblé, Gianluca Bontempi Deep-Learning Domain Adaptation Techniques for Credit Cards Fraud Detection, INNSBDDL 2019: Recent Advances in Big Data and Deep Learning, pp 78-88, 2019