Lambton College – Mississauga

Leveraging Machine Learning for Credit Card Fraud Detection

David Higuera
Cecille Jatulan
Maria Melencio
Michael Montanez
Diana Reyes

2024W-T3 BDM 3035 - Big Data Capstone Project 01
Meysam Effati
April 16, 2024

Table of Contents

Abstract

The prevalence of credit card fraud in online transactions underscores the need for effective detection mechanisms. This project uses machine learning techniques to develop a robust credit card fraud detection system. We aim to create a model that accurately identifies fraudulent transactions while minimizing false positives to maintain a seamless user experience.

The project employs a dataset containing legitimate and fraudulent transactions. Machine learning algorithms, including logistic regression, random forests, and gradient boosting, are used for model training and evaluation. Feature engineering techniques are applied to extract relevant information from transaction data, enhancing model predictive power.

Methodologically, exploratory data analysis (EDA) is conducted to understand data variability and distribution. Model selection involves evaluating algorithms such as Random Forest Classifier and XGBoost Classifier and resampling techniques like RandomUnderSampler and SMOTE to address class imbalance. Model training, validation, and evaluation are performed, considering precision, recall, F1-score, and AUC-ROC metrics.

Results show that while the Random Forest Classifier and XGBoost Classifier achieve high accuracy, recall for fraudulent transactions varies. Balanced Random Forest Classifier and SMOTE-enhanced models exhibit improved recall but lower precision. A trade-off between recall and precision is observed, necessitating careful consideration of model performance.

In conclusion, techniques like RandomUnderSampler and SMOTE enhance recall, particularly for fraudulent transactions, albeit with a trade-off in precision. Future research could explore advanced ensemble methods or deep learning architectures to better balance recall and precision in fraud detection systems.

Introduction

In today's interconnected and cashless society, credit card transactions have become the backbone of modern commerce, facilitating convenient and secure payments across various platforms. However, with the proliferation of online transactions, the incidence of credit card fraud has also escalated, posing a significant threat to consumers and financial institutions. Detecting and preventing fraudulent activities in real time has thus become a paramount concern in economic security.

Background Information on the Problem Domain:

Credit card fraud encompasses a range of deceptive practices, including unauthorized transactions, stolen card information, and identity theft, among others. Fraudsters exploit vulnerabilities in payment systems to conduct illicit transactions, resulting in substantial financial losses and reputational damage. Traditional fraud detection methods, primarily rule-based systems, are often inadequate in detecting sophisticated fraudulent activities, necessitating the adoption of advanced analytical techniques to combat fraud effectively.

Statement of the Problem:

The overarching problem addressed in this capstone is the need for a robust and efficient credit card fraud detection system capable of accurately identifying fraudulent transactions while minimizing false positives. The challenge lies in developing a real-time model that can effectively distinguish between legitimate and fraudulent transactions, enabling timely intervention and preventing fraudulent activities.

Objectives of the Project:

The primary objective of this thesis is to develop and evaluate a comprehensive credit card fraud detection system using machine learning algorithms. Specifically, the project aims to achieve the following objectives:

To analyze and preprocess a large-scale dataset of credit card transactions, encompassing both legitimate and fraudulent instances.

To explore and implement various machine learning algorithms, including logistic regression, decision trees, random forests, and neural networks, for training and evaluating fraud detection models.

Optimize model performance through feature engineering, hyperparameter tuning, and ensemble techniques to enhance predictive accuracy and generalization capability.

The purpose of this study is to assess the effectiveness of the developed models in accurately detecting fraudulent transactions while minimizing false positives, using evaluation metrics such as accuracy, precision, recall, and F1-score.

The goal is to provide insights into the underlying patterns and features indicative of fraudulent activity through feature importance analysis and visualization techniques.

Overview of the Methodology Used:

The methodology employed in this thesis encompasses a systematic approach to credit card fraud detection, leveraging machine learning algorithms and advanced analytical techniques. The process involves data collection, preprocessing, feature engineering, model selection, training, evaluation, and performance optimization. Various machine learning algorithms are explored and evaluated, focusing on ensemble methods to improve model robustness and reliability. Additionally, feature importance analysis is conducted to gain insights into the factors contributing to fraud detection, thereby enhancing the interpretability and transparency of the developed models.

Data Collection and Preprocessing

Data Collection:

The dataset used for this project is the Credit Card Fraud Detection Dataset sourced from Kaggle. It is available at the following link: Credit Card Fraud Detection Dataset. The dataset contains transactions made by credit cards in September 2013 by European cardholders. It consists of numerical input variables which are the result of a PCA transformation and anonymized due to confidentiality issues. The dataset also includes the 'Time' and 'Amount' features, representing the seconds elapsed between each transaction and the first transaction in the dataset and the transaction amount, respectively. The target variable, 'Class', indicates whether a transaction is fraudulent (1) or not (0).

Data Preprocessing:

1. **Reading the Data:** The CSV file containing the dataset was successfully read into a pandas DataFrame, allowing for further analysis and manipulation.
2. **Identifying Data Types:** The data types of the features were inspected using df.dtypes to ensure proper handling of each variable during preprocessing and modeling.
3. **Checking for Missing Values:** Utilizing df.isnull().sum(), it was determined that there are no missing values present in the dataset, simplifying the preprocessing pipeline.
4. **Handling Missing Values:** Since no missing values were detected, no imputation or removal of records was required. However, strategies like imputation or removal could have been employed if missing values were present.
5. **Handling Class Imbalance:** The 'Class' feature revealed a significant class imbalance, where fraudulent transactions are substantially less represented compared to non-fraudulent ones. Techniques such as undersampling, or using algorithms robust to class imbalance were considered to address this issue.
6. **Feature Scaling/Normalization:** Scaling features to a similar range was considered to potentially enhance the performance of certain machine learning algorithms.
7. **Feature Engineering:** The possibility of creating new features based on domain knowledge or feature interactions was acknowledged to potentially improve the model's predictive power.
   - *Outlier Detection/Removal:* Due to the imbalanced nature of the dataset, traditional outlier removal techniques may not be suitable as they could remove valuable information related to fraudulent transactions. Instead, robust models and techniques capable of handling imbalanced data and outliers, such as Random Forest or Gradient Boosting, were selected. Anomaly detection algorithms like Isolation Forest or One-Class SVMs were also considered effective for outlier identification without requiring a balanced dataset.

Challenges Encountered:

One of the main challenges encountered during data preprocessing was the handling of outliers. Given the imbalanced nature of the dataset and the goal of detecting fraudulent transactions, traditional outlier removal techniques were deemed inadequate as they might remove crucial information related to fraud. To address this, robust models and techniques capable of handling imbalanced data and outliers were prioritized, along with anomaly detection algorithms. Additionally, the evaluation stage will consider metrics suitable for imbalanced datasets, such as precision, recall, F1-score, and AUC-ROC, to comprehensively assess the model's performance in detecting fraudulent transactions while accounting for class imbalance.

Methodology

Exploratory Data Analysis (EDA):

- Obtain summary statistics for the DataFrame.
- Identify the distribution of each column in the dataset to understand the data's variability and potential skewness.
- Create histograms to explore the distribution of transaction amounts, particularly focusing on the frequency of transactions with zero amounts.
- Explore the relationship between transaction amounts and fraudulent transactions through a histogram, providing insights into potential differences in transaction amounts between fraudulent and non-fraudulent transactions.
- Utilize boxplots to visualize the distribution and identify potential outliers for all features in the dataset, aiding in the detection of anomalous data points.
- Create a pie chart to identify the proportion of fraudulent transactions within the dataset, providing an overview of the class distribution and highlighting the imbalance between fraudulent and non-fraudulent transactions.
- Explore the quantity of transactions over time by creating a histogram, allowing for the identification of temporal patterns or anomalies in transaction volume.
- Use a boxplot to identify outliers in the 'Amount' feature, facilitating the detection of extreme values that may require further investigation or treatment during preprocessing.

Model Selection:

- *Random Forest Classifier:* Selected for its robustness, ability to handle complex data, and effectiveness in handling imbalanced datasets through ensemble learning.
- *XGBoost Classifier:* Chosen for its efficiency and effectiveness in sequential boosting, which is beneficial for improving model performance on imbalanced datasets.
- *Resampling Techniques:* Utilize RandomUnderSampler, Balanced Random Forest Classifier, and SMOTE to address class imbalance by either undersampling the majority class or oversampling the minority class.

Model Training, Validation, and Evaluation:

Training: Train each model on the training set using default hyperparameters or after hyperparameter tuning.

Validation: Validate model performance on the testing set using the chosen evaluation metric, primarily focusing on recall to minimize false negatives.

Evaluation Metric: Use evaluation metrics such as precision, recall, F1-score, and AUC-ROC to assess model performance comprehensively.

Model Comparison: Compare the performance of different models to identify the most effective approach for fraud detection.

Hyperparameter Tuning and Optimization:

Grid Search: Implement Grid Search on models like Balanced Random Forest Classifier to exhaustively search through hyperparameter grids and find optimal settings.

Randomized Search Cross-Validation: Apply Randomized Search Cross-Validation on models like Balanced Random Forest Classifier after SMOTE to efficiently sample hyperparameters and reduce computation time while still achieving good performance.

Fine-Tuning: Iterate through multiple rounds of parameter tuning to optimize model performance based on evaluation metrics, ensuring the best trade-off between precision and recall.

Feature Engineering:

Time Feature Modification: Modify the 'Time' feature to represent the time of day for each transaction, potentially capturing time-dependent patterns in fraudulent activity.

Feature Scaling: Scale features to avoid imbalanced weights among independent variables, evaluating model performance both with and without scaling to assess its impact.

Results

The credit card fraud detection project involved the evaluation of several machine learning models. The performance of each model was assessed based on precision, recall, F1-score, and accuracy metrics. The models considered include Random Forest Classifier, XGB Classifier, Balanced Random Forest Classifier, and variations of Random Forest Classifier with RandomUnderSampler and SMOTE techniques.

Performance Metrics

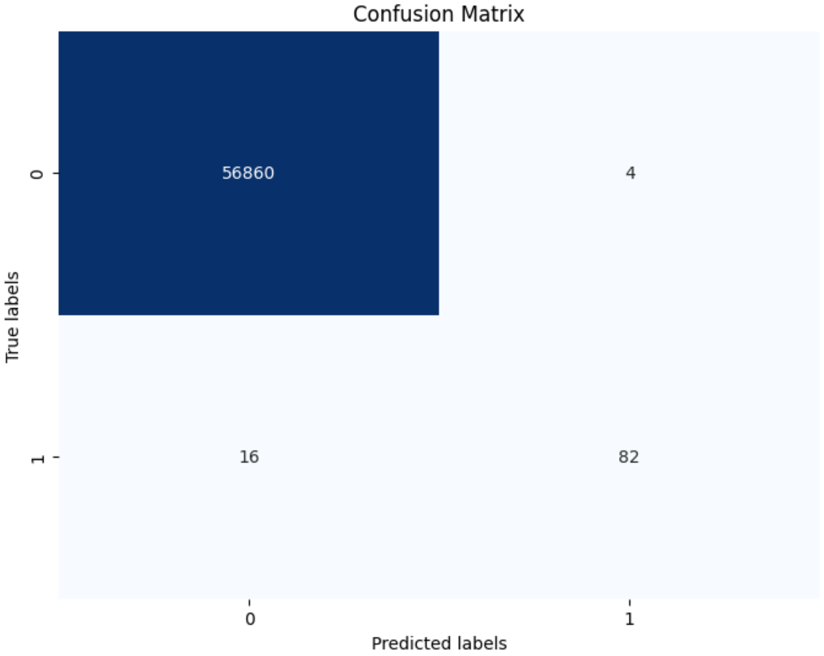
The following performance metrics were used for evaluation:

- Precision: Ability of the model to correctly identify fraudulent transactions out of all predicted frauds.
- Recall: Ability of the model to correctly identify fraudulent transactions out of all actual frauds.
- F1-score: Harmonic mean of precision and recall, providing a balanced measure between the two metrics.
- Accuracy: Proportion of correctly classified transactions out of the total transactions.

Comparison of Different Models and Techniques

Random Forest Classifier (Model 1):

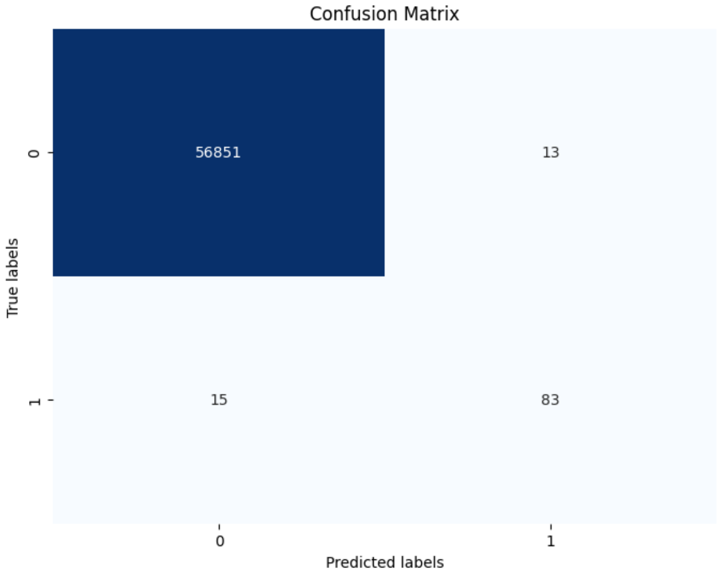
- Achieved an accuracy of almost 100% with a recall of 84% for fraudulent transactions.
- Outperformed by the XGB Classifier in terms of recall for fraudulent transactions.



XGB Classifier (Model 2):

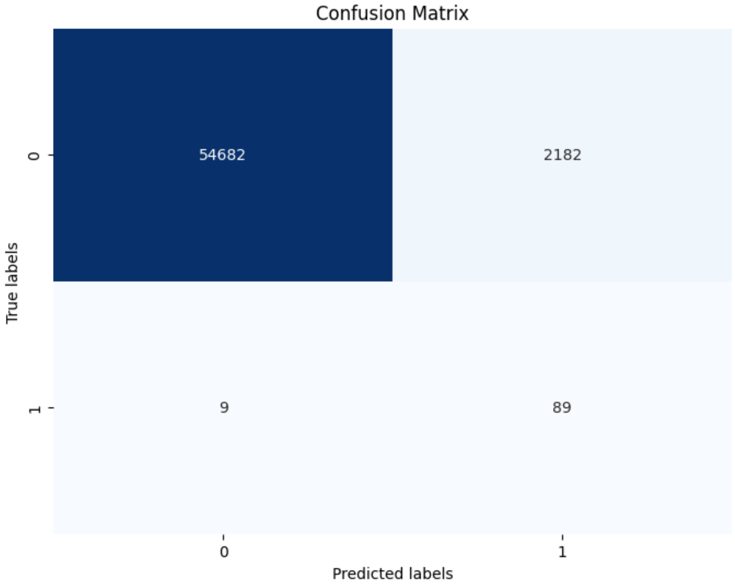
- Achieved an accuracy of almost 100% with a recall of 85% for fraudulent transactions.

- Slightly higher recall for fraudulent cases compared to the Random Forest Classifier.



Random Forest Classifier after RandomUnderSampler (Model 3):

- Achieved an accuracy of 96% with a recall of 91% for fraudulent transactions.
- Significant improvement in recall for fraudulent cases compared to the original Random Forest Classifier but at the cost of overall accuracy.

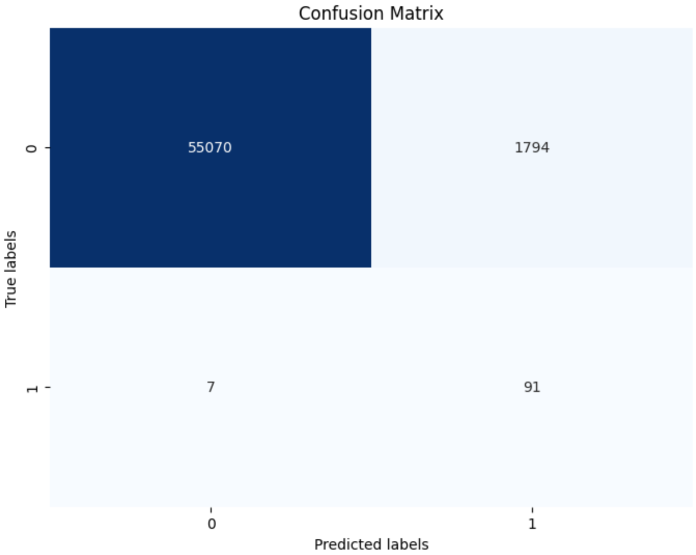


This model correctly identified 89 out of 98 fraudulent transactions (91% recall). However, it incorrectly classified 2182 non-fraudulent transactions as fraudulent, resulting in a relatively low precision for fraud detection.

Balanced Random Forest Classifier (Model 4):

- Achieved an accuracy of 97% with a recall of 93% for fraudulent transactions.

- Higher recall for fraudulent cases compared to the Random Forest Classifier, but still relatively low.

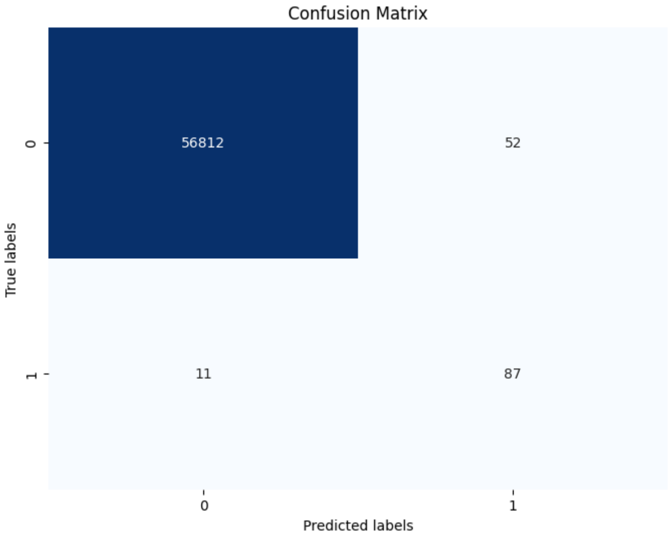


This model achieved a higher recall compared to the Random Forest Classifier. However, it still had a relatively low precision due to misclassifying non-fraudulent transactions.

Balanced Random Forest Classifier After SMOTE (Model 5):

Achieved an accuracy of almost 100% with a recall of 89% for fraudulent transactions.

- Improved recall for fraudulent transactions compared to the original Balanced Random Forest Classifier.



By using SMOTE to balance the dataset, this model achieved an improved recall for fraudulent transactions while maintaining high precision, resulting in a more balanced performance.

The following table summarizes the performance of the models, focusing on precision and recall, which are crucial for evaluating the detection of fraudulent transactions.

Metrics	Model 1	Model 2	Model 3	Model 4	Model 5
Precision	0.953	0.865	0.039	0.048	0.626
Recall	0.837	0.847	0.908	0.929	0.888

Discussion

Precision: Precision measures the accuracy of the optimistic predictions made by the model. It tells us what proportion of transactions are predicted as fraudulent.

Model 1 (Random Forest Classifier) achieved the highest precision at 0.953, indicating that when it predicts a transaction as fraudulent, it is correct 95.3% of the time. This suggests that Model 1 is very effective at minimizing false positives, which is crucial in fraud detection to avoid inconveniencing legitimate customers with unnecessary security measures.

Model 2 (XGB Classifier) has a slightly lower precision at 0.865 compared to Model 1. While still high, it suggests that Model 2 may be less accurate in identifying true positives than Model 1.

Models 3, 4, and 5 (Random Forest Classifier after RandomUnderSampler, Balanced Random Forest Classifier, and Balanced Random Forest Classifier after SMOTE, respectively) all have significantly lower precision scores, indicating a higher rate of false positives compared to Models 1 and 2. This implies that these models may be less effective in distinguishing fraudulent transactions from legitimate ones, leading to a higher risk of mistakenly flagging legitimate transactions as fraudulent.

Recall: Recall measures the model's ability to correctly identify all positive instances, including actual and false negatives. It tells us the proportion of actual fraudulent transactions correctly identified by the model.

Models 3, 4, and 5 achieved higher recall scores than Models 1 and 2. This indicates that these models better capture a more significant proportion of fraudulent transactions, minimizing false negatives. However, this comes at the expense of precision, as noted in the lower precision scores for Models 3, 4, and 5.

Model 4 (Balanced Random Forest Classifier) achieved the highest recall score at 0.929, suggesting it is the most effective at identifying fraudulent transactions among all models tested.

Models 1 and 2 also have respectable recall scores, indicating that they can still capture a significant portion of fraudulent transactions despite prioritizing precision. However, the precision gap with the other models is too big, so despite that recall, we consider models 1 and 2 as the best models.

Implications:

The choice of model depends on the specific objectives and constraints of the credit card fraud detection system. If minimizing false positives (precision) is a top priority to avoid inconveniencing legitimate customers, then Models 1 and 2 may be preferred.

Balancing precision and recall based on the organization's risk tolerance and priorities is crucial. Additionally, ongoing monitoring and fine-tuning of the models may be necessary to adapt to changing fraud patterns and minimize false positives while maximizing fraud detection.

Strengths and Weaknesses of the Models:

Random Forest Classifier (Model 1):

Strengths:

She achieved the highest precision, indicating high accuracy in identifying fraudulent transactions.

A decent recall score suggests that it captures a reasonable proportion of fraudulent transactions.

Weaknesses:

It may struggle with capturing all instances of fraud due to its default settings prioritizing precision over recall.

Prone to overfitting on imbalanced data, leading to suboptimal performance.

XGB Classifier (Model 2):

Strengths:

High precision, though slightly lower than Model 1.

Good balance between precision and recall.

Weaknesses:

Similar to Model 1, it may not capture all instances of fraud due to its default settings prioritizing precision.

It may require more computational resources and longer training times than simpler models like Random Forest.

Random Forest Classifier after RandomUnderSampler (Model 3):

Strengths:

We achieved higher recall than Models 1 and 2, indicating better detection of fraudulent transactions.

Weaknesses:

Significantly lower precision, suggesting a higher rate of false positives.

We may need to conserve valuable information by undersampling the majority class, potentially leading to a loss of predictive power.

Balanced Random Forest Classifier (Model 4):

Strengths:

She achieved the highest recall among all models, indicating superior detection of fraudulent transactions.

Weaknesses:

Lower precision compared to Models 1 and 2, suggesting a higher rate of false positives.

It may require more hyperparameter tuning and computational resources to achieve optimal performance.

Balanced Random Forest Classifier after SMOTE (Model 5):

Strengths:

Balanced precision and recall, distinguishing between capturing fraudulent transactions and minimizing false positives.

Weaknesses:

It may introduce synthetic samples that need to accurately represent the underlying distribution of the data, potentially leading to overfitting.

Unexpected Outcomes or Observations:

Model 3, which employed RandomUnderSampler, demonstrated a surprisingly low precision despite achieving a high recall. This suggests that undersampling the majority class may have resulted in losing important information, leading to an increased rate of false positives.

Models 4 and 5, which utilized techniques like Balanced Random Forest and SMOTE, achieved higher recall than Models 1 and 2 but at the expense of precision. This trade-off highlights the challenge of balancing the competing objectives of maximizing fraud detection while minimizing false positives.

Comparison with Prior Work and Contribution to Existing Knowledge:

Prior credit card fraud detection work has often focused on traditional machine learning algorithms like Random Forest and XGBoost. This project builds upon existing knowledge by systematically evaluating the performance of different models, including those specifically designed for handling imbalanced data.

The use of techniques like RandomUnderSampler, Balanced Random Forest, and SMOTE represents advancements in addressing class imbalance, which has been a common challenge in fraud detection in recent years.

By comparing the performance of various models in terms of precision and recall, this project provides insights into the trade-offs involved in designing an effective fraud detection system. This contributes to the existing knowledge by offering practical guidance for selecting the most suitable model based on the organization's priorities and constraints. Also, it allows us to continue searching for more techniques and try to dive deeper into the current algorithms to improve performance.

Conclusions

- The Random Forest Classifier and XGB Classifier achieved high accuracy but varied in recall for fraudulent transactions. The Balanced Random Forest Classifier and SMOTE-enhanced models showed improvements in recall but at the expense of precision.
- There is a trade-off between recall and precision. Models with higher recall tend to misclassify more non-fraudulent transactions as fraudulent, impacting precision.
- RandomUnderSampler and SMOTE techniques helped improve the recall of the models, especially for the minority class of fraudulent transactions. However, these techniques need to be balanced with the impact on precision and overall model performance.
- Future work could explore more advanced ensemble methods or deep learning architectures to improve the balance between recall and precision.

References

- Mrozek, P., Panneerselvam, J., & Bagdasar, O. (2020). Efficient resampling for fraud detection during anonymised credit card transactions with unbalanced datasets. 2020 IEEE/ACM 13th International Conference on Utility and Cloud Computing (UCC). https://doi.org/10.1109/ucc48980.2020.00067
- Windari, L. (2024, April 16). Credit Card Fraud Detection. Medium. https://medium.com/analytics-vidhya/credit-card-fraud-detection-how-to-handle-imbalanced-dataset-1f18b6f881
- OpenAI. (2023). ChatGPT [Computer software]. https://www.openai.com/chatgpt