CREDIT CARD FRAUD DETECTION USING MACHINE LEARNING ALGORITHMS

Name: Gaddam Dileep
ID: 22084021
M.Sc. Data Science
Supervisor : Stephen kane

Credit Card Fraud Detection Using Machine Learning Algorithms

Research Question :

"How can resampling techniques and machine learning algorithms, such as Logistic Regression, Random Forest, and XG Boost, be optimized to accurately detect fraud transactions within highly imbalanced datasets?".

Short Summary of Project Topic & Background:

- Detecting fraud activity within credit card transactions using machine learning methods.

Background: Credit card fraud is a major concern for both banks and customers, causing considerable financial damage, The dataset used in this project comprises **284,807** transactions, of which only **492** are fraud highlighting an **extreme class imbalance** that complicates accurate detection. This project focuses on finding the best machine learning methods to accurately detect fraud in datasets where fraud cases are significantly outnumbered by legitimate ones, by exploring techniques specifically designed to handle imbalanced data.

Project Objectives

- **Data Exploration** : Examine the Credit card dataset to find any unusual transactions in the Class column

- **Address Class Imbalance:** Use resampling techniques, including the Synthetic Minority Over-sampling Technique (SMOTE) and under-sampling, to balance the data, making fraud cases more detectable within a majority of legitimate transactions

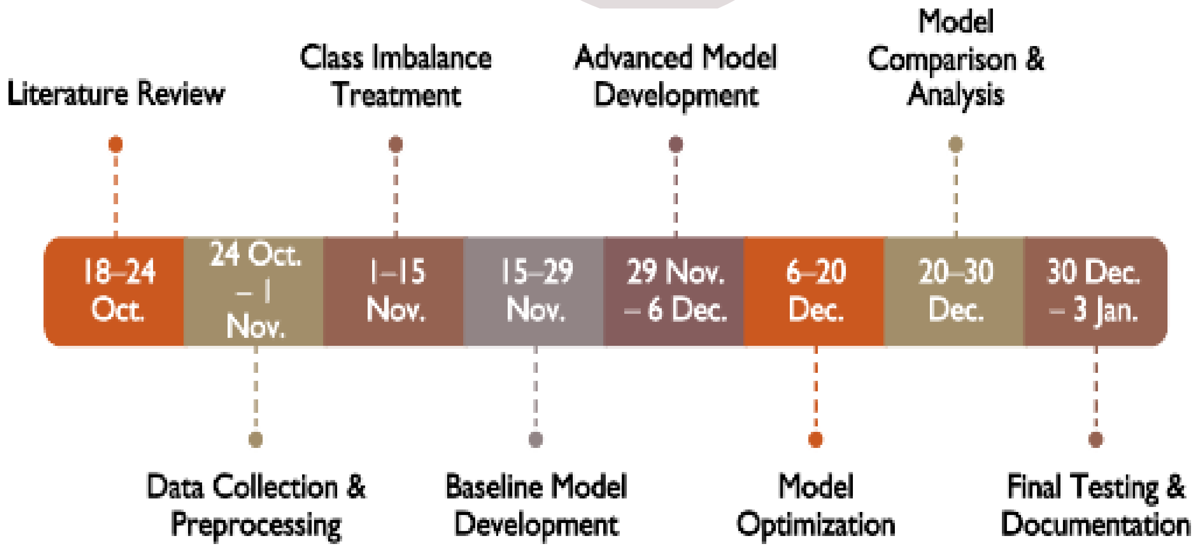
- **Model Development:** Develop and train machine learning models, specifically Logistic Regression, Random Forest, and XG Boost, to accurately identify potentially fraud transactions

- **Model Evaluation and Optimization:** Evaluate model performance using metrics tailored to imbalanced data, such as Precision, Recall, and F1-Score, with a focus on reducing false negatives to ensure actual fraud cases are detected

- **Hyperparameter Tuning:** Optimize model parameters, such as learning rate and depth, to enhance detection accuracy, ensuring the models are as effective as possible in real-world applications.

Project Timeline

- Literature Review: Oct 18 – Oct 24
- Data Collection & Preprocessing: Oct 24 – Nov 1
- Class Imbalance Treatment: Nov 1 – Nov 15
- Baseline Model Development: Nov 15 – Nov 29
- Advanced Model Development: Nov 29 – Dec 6
- Model Optimization: Dec 6 – Dec 20
- Model Comparison & Analysis: Dec 20 – Dec 30
- Final Testing & Documentation: Dec 30 – Jan 3


Literature Review
18–24 Oct.
Data Collection & Preprocessing
24 Oct. – 1 Nov.
Class Imbalance Treatment
1–15 Nov.
Baseline Model Development
15–29 Nov.
Advanced Model Development
29 Nov. – 6 Dec.
Model Optimization
6–20 Dec.
Model Comparison & Analysis
20–30 Dec.
Final Testing & Documentation
30 Dec. – 3 Jan.

Data Management Plan

Overview of the Dataset

Dataset Name: Credit Card Fraud Detection Dataset

- **Source:** The dataset is sourced from OpenML, created by researchers at the ULB Machine Learning Group.

- **Purpose:** This dataset includes anonymized credit card transaction data, which will be used to develop and analyze fraud detection models. It supports the creation of machine learning algorithms that can predict fraudulent transactions based on past transaction patterns.

Details about the Dataset:

- Full Dataset: Contains **284,807** transactions made by European cardholders over two days, with **492** labeled as fraud (**0.172%**).

Data Fields:

- Amount: The transaction amount, which can be used for certain fraud detection features.

- Class: The target variable, with "1" indicating fraud and "0" indicating non-fraud.

- Anonymized Features: The dataset includes 28 anonymized features labeled V1 through V28, derived from a PCA transformation for confidentiality

Document Control

Version Control:

- Git Hub Repository : All code and project files will be managed using GitHub to maintain version control. Weekly commits will be made to track changes and updates.

- Repository Link: Gaddamdileep · GitHub

Security and Storage :

- Backup Frequency:

Weekly backups will be made to GitHub and a cloud storage solution (e.g., OneDrive or Google Drive) to ensure data safety and prevent loss.

- Data Sharing:

Project data and code will be shared with supervisors and collaborators via GitHub for controlled access.

OneDrive will serve as an additional backup and provide easy access

Reference List

Key References:

● Ghosh, S. and 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. https://ieeexplore.ieee.org/abstract/document/323314/

● Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C. and Bontempi, G., 2017. Credit card fraud detection: a realistic modeling and a novel learning strategy. IEEE transactions on neural networks and learning systems, 29(8), pp.3784-3797.https://ieeexplore.ieee.org/abstract/document/803800 8/

● Srivastava, A., Kundu, A., Sural, S. and Majumdar, A., 2008. Credit card fraud detection using hidden Markov model. IEEE Transactions on dependable and secure computing, 5(1), pp.37-48.https://ieeexplore.ieee.org/abstract/document/4358713/

● Jemai, J., Zarrad, A. and Daud, A., 2024. Identifying Fraudulent Credit Card Transactions using Ensemble Learning. IEEE Access. https://ieeexplore.ieee.org/abstract/document/10477993/

Ethical Requirements

- GDPR Compliance:The dataset is fully anonymized, containing no personal or identifiable information, which ensures compliance with GDPR regulations.
- University Ethical Policies:The project aligns with the University of Hertfordshire's ethical policies, ensuring responsible data use.
- Permission to Use Data:The dataset is publicly available on OpenML for academic research purposes, with full permission for use.
- Ethical Data Collection:The data was ethically collected and made available by a reputable research group, ensuring adherence to ethical standards in data acquisition

Conclusion

This project effectively tackles the issue of credit card fraud detection by systematically developing and optimizing machine learning models.Through a well-structured timeline, each phase—from data collection to model evaluation—ensures a focused approach to addressing class imbalance and improving fraud detection accuracy.The use of algorithms like Logistic Regression, Random Forest, and XGBoost, along with techniques such as SMOTE, enhances the reliability of detecting fraudulent transactions in highly imbalanced datasets.This research not only improves fraud detection methods but also offers practical solutions for real-world financial security challenges.