7PAM2002-0901-2024 - Data Science Project

**Project and Data Management (PDM) Plan for Credit Card Fraud Detection**

Student name: Kavibharathi Somasundaram (ks23aad@herts.ac.uk)

# Project Overview

**Project Title**:

**Detecting Credit Card Fraud Using Machine Learning**

**Project Summary**:

Credit card fraud detection is a critical issue for financial institutions, as the volume of fraudulent transactions continues to rise due to increased online transactions. The project's objective is to develop a machine learning model to identify fraudulent credit card transactions, using an imbalanced dataset of realworld credit card transactions from European cardholders in 2013. The dataset consists of approximately 284,807 transactions, where only a small fraction (492 transactions) is fraudulent, making class imbalance the primary challenge.

**Research Question**:

How can machine learning techniques be used to accurately detect credit card fraud, and what methods are effective in addressing the imbalance between fraudulent and legitimate transactions?

**Project Objectives**:

1. **Data Pre-processing**: Clean and pre-process the credit card fraud dataset, handling the class imbalance using resampling techniques (SMOTE, under sampling).
2. **Model Building**: Build and compare several machine learning models (Logistic Regression, Random Forest, XGBoost) for fraud detection.
3. **Performance Evaluation**: Use precision, recall, F1-score, and AUC-ROC as key performance metrics to evaluate and select the best model.
4. **Interpretability**: Analyse feature importance using SHAP (Shapley Additive explanations) to understand key predictors of fraud.

**References**: 1. Dal Pozzolo, A., Caelen, O., Le Borgne, Y. A., Waterschoot, S., & Bontempi, G. (2014).

**"Calibrating Probability with Undersampling for Unbalanced Classification"**. In IEEE Symposium on Computational Intelligence and Data Mining (CIDM). Link

1. Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). **"Isolation Forest"**. In 2008 Eighth IEEE International Conference on Data Mining. Link
2. Breiman, L. (2001). **"Random Forests"**. Machine Learning, 45(1), 5-32. Link

**Project Plan: Task List and Project Timeline Task:**

**Week 1: (Sept 25 - Oct 1)**

* Project and data understanding.
* Set up GitHub for version control and tracking.
* Download and explore the dataset from Kaggle.
* **Deadline**: Initial dataset exploration completed by **Oct 1**.

**Weeks 2-3: (Oct 2 - Oct 15)**

* Data preprocessing: Handle missing data, scale Time and Amount, and address class imbalance with SMOTE or undersampling.
* Commit preprocessed data to GitHub.
* **Deadline**: Data preprocessing completed by **Oct 15**.

**Weeks 3-4: (Oct 16 - Oct 29)**

* Develop baseline models (Logistic Regression, Decision Trees) and evaluate on key metrics (precision, recall, F1-score).
* Commit baseline model code to GitHub.
* **Deadline**: Baseline model development completed by **Oct 29**.

**Weeks 5-6: (Oct 30 - Nov 12)**

* Implement advanced models (Random Forest, XGBoost, Neural Networks).
* Hyperparameter tuning using GridSearchCV or RandomSearchCV.
* Commit models and tuning code to GitHub.
* **Deadline**: Advanced model development completed by **Nov 12**.

**Week 7: (Nov 13 - Nov 19)**

* Model evaluation using AUC-ROC, precision, recall, and F1-score.
* Analyze feature importance using SHAP.
* **Deadline**: Model evaluation and feature analysis completed by **Nov 19**.

**Week 8: (Nov 20 - Dec 3)**

* Final model selection and deployment plan.
* Prepare presentation slides and final report.
* Submit project report and present PDM plan.
* **Deadline**: Presentation and submission of PDM plan by **Dec 3**.

# Data Management Plan

**Overview of the Dataset**: The dataset used in this project contains real-world credit card transactions collected in September 2013 from European cardholders. It has 284,807 transactions, with 30 features (v1 to V28, derived from PCA), along with Time, Amount, and Class (the target label, where 1 = fraud, 0 = nonfraud).

* **Source**:

The dataset is publicly available on Kaggle. The original purpose of the data collection was for research in fraud detection. [Kaggle Dataset](https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud)

**Data Collection**:

Data will be sourced from Kaggle. No personal data is used in this project, and sensitive features are anonymized using PCA, so the dataset does not violate privacy regulations.

**Metadata**:

* **Format**: CSV file.
* **Size**: Approximately 144 MB.
* **Records**: 284,807 rows and 31 columns.
* **Expected Code Size**: Around 10-20 MB for scripts and model files.

**Document Control (GitHub)**:

* **GitHub URL**:<https://github.com/ivakbharathi/Data-Science-Project>
* **Commit Frequency**: Code and data will be committed weekly, following incremental development milestones (preprocessing, modelling, evaluation).
* **Version Control**: Git will be used for version control. File versions will be tracked by dates, e.g., model\_v1.0.py, preprocessing\_v1.0.py.

**ReadMe File**: The ReadMe file will provide:

* A summary of the project.
* Instructions on how to run the code.
* Dependencies (libraries, Python version).
* Expected outcomes (model performance metrics).
* Guide for future developers to extend the codebase.

**Security and Storage**:

* **Backup Frequency**: Weekly backups will be made on GitHub and OneDrive.
* **Data Sharing**: Data will be shared via GitHub for code access and the dataset will be provided as a link to Kaggle to ensure it complies with dataset license and sharing permissions.

**Ethical Requirements**:

1. **GDPR Compliance**: The dataset does not contain any personal identifiers, as it has been anonymized through PCA. Therefore, GDPR requirements are met.
2. **UH Ethical Policies**: The dataset is publicly available and ethically shared under Kaggle’s open data license. The project does not involve any human participants directly.
3. **Permission to Use Data**: The dataset is publicly available for academic and research purposes, and I have obtained permission to use it via Kaggle’s terms.
4. **Ethical Data Collection**: The original dataset was collected ethically for research in fraud detection by the ULB Machine Learning Group and shared on Kaggle for academic purposes.