Project Title: CREDIT CARD FRAUD DETECTION

Team Members (Data Minds):
Hilda Ogamba
Joyce Malicha
Lynn Obadha

SECTION 1: OVERVIEW OF THE PROJECT

Project Overview: This project aims to detect fraudulent transactions in credit card data using various machine learning models. The dataset includes credit card transactions made by European cardholders over two days in September 2013, with 492 out of 284,807 transactions labeled as fraud. The dataset is highly imbalanced, with fraud constituting only 0.172% of all transactions.

Domain of Application: This project is in the financial sector and specifically targets fraud detection in credit card transactions.

Particular Issue: The main issue is detecting rare fraudulent transactions in a highly imbalanced dataset, challenging conventional machine learning models favoring the majority class (non-fraudulent transactions).

Approach Insight: We will implement and compare various machine learning models—logistic regression, random forests, and neural networks. To address class imbalance, techniques such as under sampling, oversampling, and SMOTE will be explored (Pozzolo et al., 2015). Additionally, given the imbalanced nature of the dataset, performance will be measured using metrics like the Area Under the Precision-Recall Curve (AUPRC) (Carcillo et al., 2021)rather than traditional accuracy.

Motivations: Credit card fraud is a serious issue for both financial institutions and consumers, and detecting fraud early can save significant financial resources. The ability to identify fraudulent transactions helps protect customers from financial losses and ensures trust in banking systems.

Goals: The goal is to develop a highly accurate fraud detection model that minimizes false positives (incorrectly labeling legitimate transactions as fraud) while ensuring fraud is detected as early as possible.

SECTION 2: RELATED WORK

Prior studies on fraud detection have explored both supervised and unsupervised machine learning techniques. Several projects have utilized logistic regression, decision trees, and random forests to detect fraudulent activities, while more recent work incorporates deep learning models for better performance (Dal Pozzolo et al., 2018).

Differentiation: This project will implement a supervised learning pipeline that compares multiple models, with a focus on handling class imbalance and optimizing performance using the AUPRC metric. We will also incorporate ensemble methods to enhance model performance and leverage PCA-transformed features effectively.

SECTION 3: DATA PLAN

Data Description:
The dataset contains 284,807 transactions, with 30 principal component features (V1-V28), the transaction "Time," the transaction "Amount," and the target class label ("Class"), which indicates fraud (1) or non-fraud (0). The "Time" feature measures the time elapsed between each transaction and the first transaction, while "Amount" represents the transaction value.

Data Source: The dataset is sourced from a research collaboration between Worldline and the Machine Learning Group at ULB (Université Libre de Bruxelles).

Preprocessing Steps:

- **Scaling:** Apply scaling techniques to the "Amount" feature as its range differs from the PCA-transformed features.
- **Time Feature:** The "Time" feature may be transformed or excluded depending on its correlation with fraud detection.
- **Class Imbalance:** Due to the extremely low fraud rate (0.172%), we will use techniques such as SMOTE or under sampling of the majority class to balance the dataset.
- **Train-Test Split:** The dataset will be split into training and test sets using stratified sampling to ensure both sets contain a proportional number of fraudulent transactions.

SECTION 4: IMPLEMENTATION PLAN

Data Mining Pipeline:

1. **Data Cleaning and Preprocessing:**
    - Handle missing values (if any).
    - Scale "Amount" and analyze the distribution of "Time."
    - Implement resampling techniques like SMOTE to balance the dataset.
2. **Model Development:**
    - Train multiple models, including Logistic Regression, Random Forests, and Neural Networks, to classify transactions.
    - Use cross-validation and GridSearchCV for hyperparameter tuning.
3. **Model Evaluation:**
    - Evaluate using AUPRC and ROC-AUC, focusing on precision, recall, and F1-score due to the imbalanced nature of the dataset.
4. **Ensemble Learning:**
    - Explore ensemble methods like XGBoost or Random Forests to boost accuracy and robustness in fraud detection.

Software and Libraries:

- **Pandas** and **NumPy** for data preprocessing.
- **Scikit-learn** for model building and evaluation.
- **Imbalanced-learn** for implementing SMOTE.
- **Matplotlib** and **Seaborn** for visualization.
- **TensorFlow/Keras** for neural network implementation (if needed).

SECTION 5: EVALUATION PLAN

Evaluation Methodology:

- The model performance will be evaluated using AUPRC, as it is more appropriate for imbalanced datasets where positive class detection (fraud) is crucial.
- Confusion matrices, precision, recall, and F1-score will be used to assess the trade-off between false positives and true positives.
- A comparative analysis of different models (Logistic Regression, Random Forest, Neural Networks) will be performed based on the AUPRC score.

Success Criteria: The model will be deemed successful if it achieves a high AUPRC score and effectively identifies fraudulent transactions with a low false positive rate while maintaining high recall.

SECTION 6: GROUP COLLABORATION

Collaboration Approach: Team members will collaborate using GitHub for version and Google Colab for development. Regular meetings will be held via Zoom and physically.

Code and Data Management: GitHub will be used to manage all code and datasets. Each member will be assigned specific tasks, such as data preprocessing, model implementation, and evaluation.

SECTION 7: TIMELINE

Week	Task
Week 9	Data exploration and preprocessing
Week 10	Implement baseline models (Logistic Regression)
Week 11	Hyperparameter tuning and resampling techniques
Week 12	Model evaluation and selection
Week 13	Prepare final report and presentation

SECTION 8: REFERENCES

References

Carcillo, F., Le Borgne, Y.-A., Caelen, O., Kessaci, Y., Oblé, F., & Bontempi, G. (2021). Combining unsupervised and supervised learning in credit card fraud detection. *Information Sciences*, *557*, 317–331. https://doi.org/10.1016/j.ins.2019.05.042

Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G. (2018). Credit Card Fraud Detection: A Realistic Modeling and a Novel Learning Strategy. *IEEE Transactions on Neural Networks and Learning Systems*, *29*(8), 3784–3797. https://doi.org/10.1109/TNNLS.2017.2736643

Pozzolo, A. D., Caelen, O., Johnson, R. A., & Bontempi, G. (2015). Calibrating Probability with Undersampling for Unbalanced Classification. *2015 IEEE Symposium Series on Computational Intelligence*, 159–166. https://doi.org/10.1109/SSCI.2015.33