Internship Project Details: Credit Card Fraud Detection with MLflow

Project Title:

"Credit Card Fraud Detection using Machine Learning and MLflow"

Project Objective:

To develop a machine learning pipeline for detecting fraudulent credit card transactions, leveraging MLflow to manage the lifecycle of machine learning experiments.

Project Description:

Context:
Credit card fraud detection is a critical task for financial institutions. Your goal is to create a robust solution to identify fraudulent transactions, focusing on handling class imbalance and ensuring reproducibility using MLflow.

Dataset Details:

- **Transactions:** Credit card transactions made by European cardholders in September 2013.
- **Time Period:** 2 days.
- **Total Transactions:** 284,807.
- **Fraud Cases:** 492 (0.172%).
- **Features:**
  - 28 principal components obtained via PCA (V1, V2, ..., V28).
  - `Time`: Seconds elapsed between a transaction and the first transaction in the dataset.
  - `Amount`: Transaction amount.
  - `Class`: Response variable (1 for fraud, 0 otherwise).
- **Challenge:** The dataset is highly imbalanced.

Project Tasks:

1. **Data Preprocessing and Exploration:**
   - Analyze and visualize the data distribution.
   - Handle missing or outlier values if required.
   - Scale the `Amount` and `Time` features.
2. **Modeling and Experimentation:**
   - Split the dataset into training and test sets.

    - Experiment with different resampling techniques (e.g., SMOTE, undersampling) to address class imbalance.
    - Train models such as Logistic Regression, Random Forest, or Gradient Boosting.
    - Evaluate models using metrics suitable for imbalanced datasets, such as:
        - Area Under Precision-Recall Curve (AUPRC).
        - F1-score.
3. **MLflow Integration:**
    - Use MLflow to track experiments, including:
        - Model parameters.
        - Evaluation metrics.
        - Model artifacts.
    - Log the best-performing model.
    - Create a comparison dashboard using MLflow UI.
4. **Model Deployment:**
    - Save the trained model as a reusable artifact.
    - Demonstrate a simple deployment strategy (optional).
5. **Documentation and Presentation:**
    - Document the experiment workflow, findings, and results.
    - Present a final report with key insights and recommendations.

Technologies to Use:

- **Programming Language:** Python
- **Tools and Libraries:**
    - MLflow for experiment tracking and model management.
    - scikit-learn, pandas, numpy, matplotlib, seaborn.
    - Imbalanced-learn for handling class imbalance.
    - Jupyter Notebook for coding and visualization.

Expected Deliverables:

1. **Codebase:** A well-structured repository with scripts and notebooks.
2. **MLflow Logs:** A comprehensive MLflow tracking dashboard.
3. **Report:**
    - Data insights and visualizations.
    - Model evaluation metrics and comparison.
    - Challenges faced and solutions implemented.
4. **Model Artifact:** The best-performing trained model saved using MLflow.

Learning Outcomes:

- Understand the end-to-end workflow of a machine learning project.
- Gain experience in working with imbalanced datasets.
- Learn to use MLflow for managing machine learning experiments.
- Enhance problem-solving and critical thinking skills.