Solution Approach

The problem of credit card fraud detection is approached through a systematic machine learning pipeline to ensure that the models can effectively identify fraudulent transactions in a highly imbalanced dataset. The steps are as follows:

1. Data Understanding and Exploration

- Load the dataset and study its structure, dimensions, and distribution of variables.
- Identify the key features (Time, Amount, V1–V28) and the target variable (Class).
- Understand the imbalance in the dataset where fraudulent transactions account for only 0.172%.

2. Data Cleaning

- **Handling Missing Values**: Check for and impute/remove any missing data (if present).
- **Outlier Treatment**: Detect and handle extreme transaction values or anomalies that could skew results.

3. Exploratory Data Analysis (EDA)

- **Univariate Analysis**: Study individual feature distributions to check skewness and unusual patterns.
- **Bivariate Analysis**: Explore relationships between features and the target variable to identify significant predictors.

4. Data Preparation for Modeling

- **Skewness Handling**: Apply appropriate transformations to reduce skewness for fair modeling.
- **Data Imbalance Treatment**: Use techniques like under-sampling, over-sampling, or SMOTE (Synthetic Minority Over-sampling Technique) to balance the dataset since fraud cases are extremely rare.

5. Train/Test Split and Scaling

- Split the dataset into training and testing sets for model validation.
- Normalize or standardize features (especially Amount and Time) to bring them to a comparable scale.

6. Model Building

- Train multiple machine learning algorithms including Logistic Regression, Support Vector Machines (SVM), Decision Trees, Random Forest, and XGBoost.
- Perform **Hyperparameter Tuning** using Grid Search with Cross Validation to identify optimal parameters for each model.

7. Model Evaluation

- Since accuracy is not meaningful for imbalanced data, evaluate models based on:
    - **Precision**: Ability to correctly identify fraud cases among predicted frauds.
    - **Recall (Sensitivity/TPR)**: Ability to detect actual frauds.
    - **F1-score**: Balance between precision and recall.
    - **ROC-AUC**: Overall ability to distinguish between fraud and non-fraud, focusing on maximizing True Positive Rate (TPR) while minimizing False Positive Rate (FPR).

8. Final Outcome

The best-performing model is selected based on its ability to detect fraudulent transactions with high recall and precision, while maintaining a good ROC-AUC score. This ensures fewer false negatives (missed frauds) and an acceptable level of false positives (legitimate transactions flagged as fraud).