Project Report: Credit Card Approval Predictor

1. Project Overview

This project focuses on building a predictive model to determine whether a credit card application will be approved or denied. Using machine learning techniques, the model is trained on a dataset of credit card applications, with the target variable indicating approval status.

2. Objectives

1. Preprocess and clean the dataset to handle missing and inconsistent data.
2. Encode categorical data for model compatibility.
3. Build and evaluate a logistic regression model to predict credit card approvals.
4. Optimize the model using hyperparameter tuning with GridSearchCV.

3. Dataset Description

The dataset contains credit card application information. The key features include demographic data, financial attributes, and application-related metrics. The target variable indicates the application approval status.

- **Columns:** 14 features (mix of numeric and categorical).
- **Missing Values:** Represented as `'?'`, later handled appropriately.
- **Data Type Composition:** Numeric and categorical values.

4. Methodology

Data Preprocessing

1. **Loading the Dataset**: Loaded using `pandas`.
2. **Missing Values**:
    - Replaced `'?'` with `NaN`.
    - Imputed numeric features with their mean values.
    - Replaced missing values in categorical columns with the placeholder `'Unknown'`.
3. **Encoding Categorical Data**:
    - Applied `LabelEncoder` to transform non-numeric data into numerical values.
4. **Feature Selection**: Dropped features 11 and 13 due to redundancy or irrelevance.
5. **Data Normalization**:
    - Used `MinMaxScaler` to scale feature values into the range [0,1].

6. **Train-Test Split**:
    - Split the dataset into 67% training and 33% testing data.

Model Development

1. **Model Choice**:
    - Implemented a Logistic Regression model for binary classification.
2. **Training**:
    - Fitted the model on the training data using scaled features.
3. **Evaluation**:
    - Predicted outcomes on the test set.
    - Calculated accuracy and confusion matrix.

Hyperparameter Tuning

- Applied GridSearchCV to optimize the model by varying `tol` and `max_iter`.
- Determined the best parameters for logistic regression.

5. Results

1. **Model Performance**:
    - **Accuracy**: Reported an accuracy score based on test predictions.
    - **Confusion Matrix**:
        - Provided insight into true positives, true negatives, false positives, and false negatives.
2. **Optimal Parameters**:
    - **Best Parameters**: Derived using GridSearchCV.
    - **Best Cross-Validated Score**: Displayed the highest accuracy achieved during cross-validation.

6. Tools and Libraries Used

1. **Data Processing**:
    - `pandas`
    - `numpy`
2. **Model Building**:
    - `sklearn` (for Logistic Regression, Label Encoding, Train-Test Split, and GridSearchCV)
3. **Evaluation**:
    - Confusion matrix and accuracy score.

7. Conclusion

The Logistic Regression model successfully predicts credit card approval with significant accuracy. Through effective data preprocessing and hyperparameter tuning, the model's performance was optimized. Future improvements may include exploring additional algorithms like Random Forest or XGBoost for enhanced predictive accuracy.

8. Future Work

1. Explore feature engineering techniques for deeper insights.
2. Experiment with other classification models and ensemble techniques.
3. Address potential class imbalances if present.

This report provides a comprehensive summary of the credit card approval predictor project, including the steps followed, results achieved, and potential areas for further enhancement.