Credit Card Approval Prediction

Ayush Bajracharya
Big Data
09/12/2024

Table of Contents

Introduction

Predict whether a customer will get their credit card approved

Data set,

Dataset Overview: Credit card approval prediction dataset, including features such as age, income, family size, etc.

Methodology: Feature engineering, model selection, training, evaluation, and optimization.

Dataset

application_record.csv

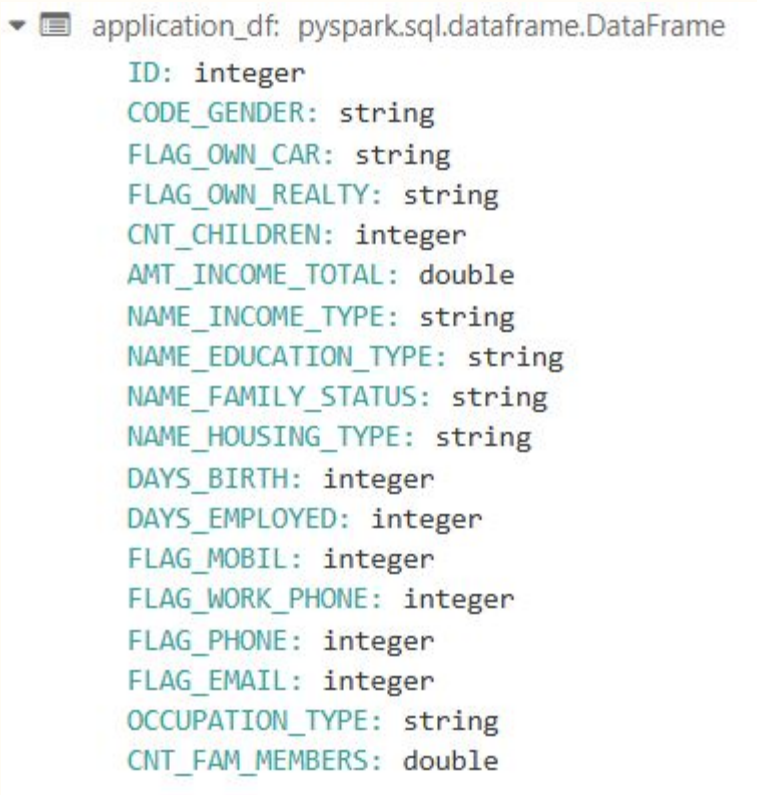
Size: 54 MB

8 categorical variables

10 numerical variables

DAYS_BIRTH = negative int (-1 means yesterday)

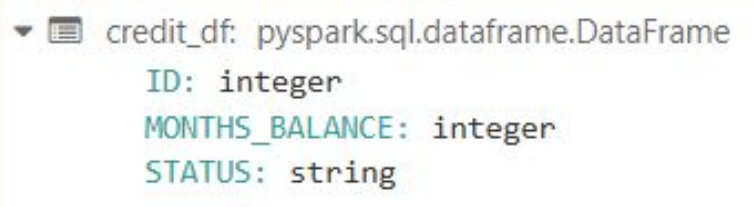
DAYS_EMPLOYED = negative int



Dataset

credit_record.csv

Size: 15 MB



MONTHS_BALANCE = negative int (-1 means last month)

STATUS = {0, 1, 2, 3, 4, 5, X, C}

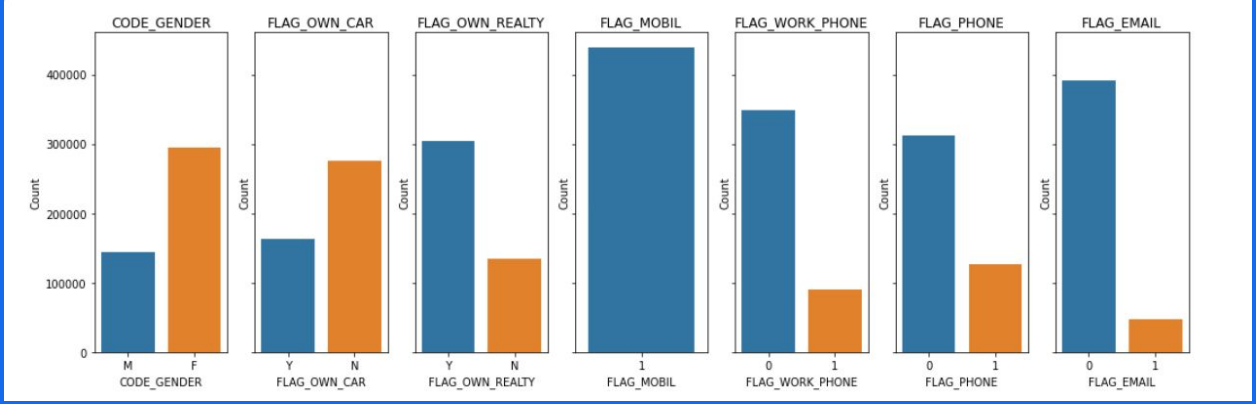
0-5: scale of late repayments

X: no loans

C: not late for repayments

Exploratory Data Analysis

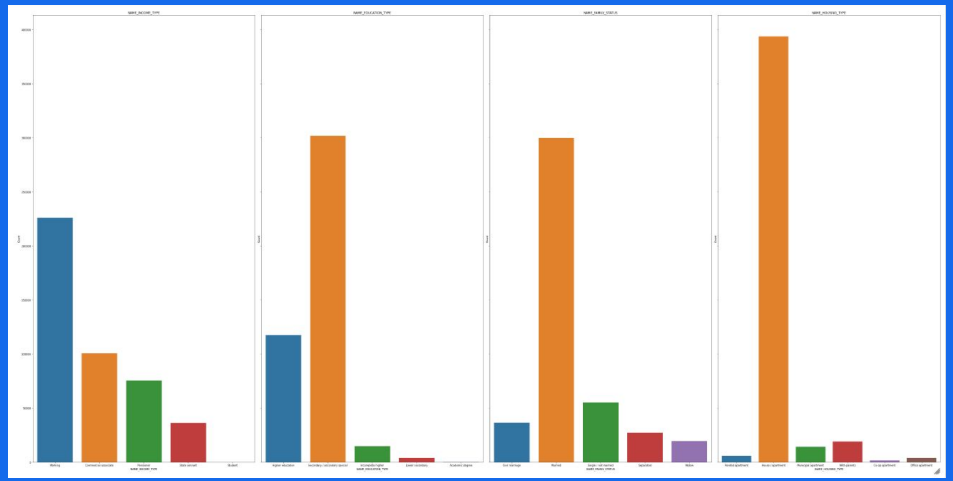
Binary classes



Exploratory Data Analysis

Socio-Economic Status

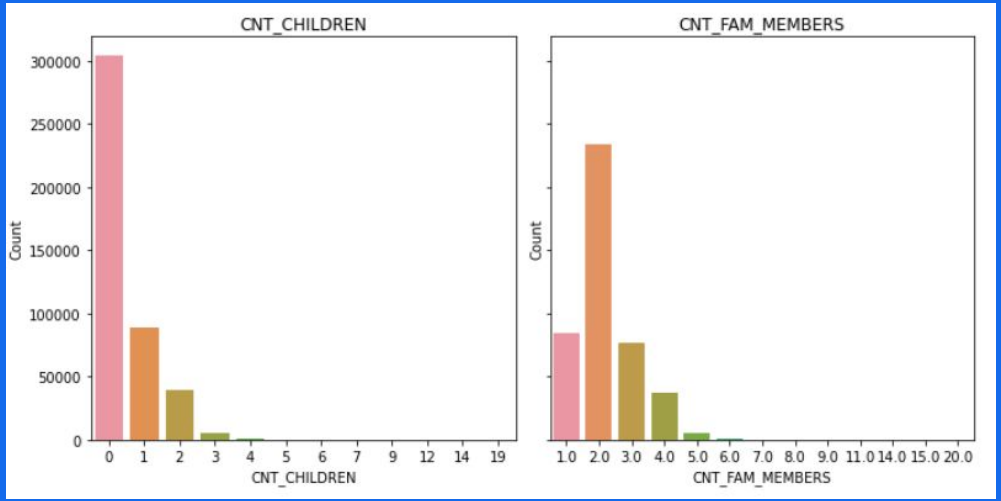
- Income type
- Education level
- Family status
- Housing type



Exploratory Data Analysis

Family Structure

- Number of children
- Number of family members



Feature Engineering

- **Labeling Credit Status**
- Account Opened Month
- Computing Age
- Credit History Length
- Recent Activity
- Years Employed
- Group Employment type
- Drop Unnecessary Columns
- Drop Low Frequency Categories
- Log Transformation - **Total Income**
- Age Binning
- Derived Features - Income per year, Income per family member etc.
- Feature Grouping - Total Income Activity Ratio

Preprocessing Pipeline

String Indexer

We use the StringIndexer to convert categorical variables into numerical indices. This transformation helps in representing each category as a unique numeric value.

One Hot Encoder

The OneHotEncoder is used to convert the indexed values into one-hot encoded vectors, which are binary vectors representing each category.

Vector Assembler

The VectorAssembler is used to combine both the one-hot encoded categorical features and numerical features into one unified vector column

Standard Scaler

The StandardScaler standardizes the feature vector by removing the mean and scaling the features to unit variance. It helps to normalize the range of the features, ensuring that the machine learning models run optimally

Machine Learning Model

ML Models Evaluation

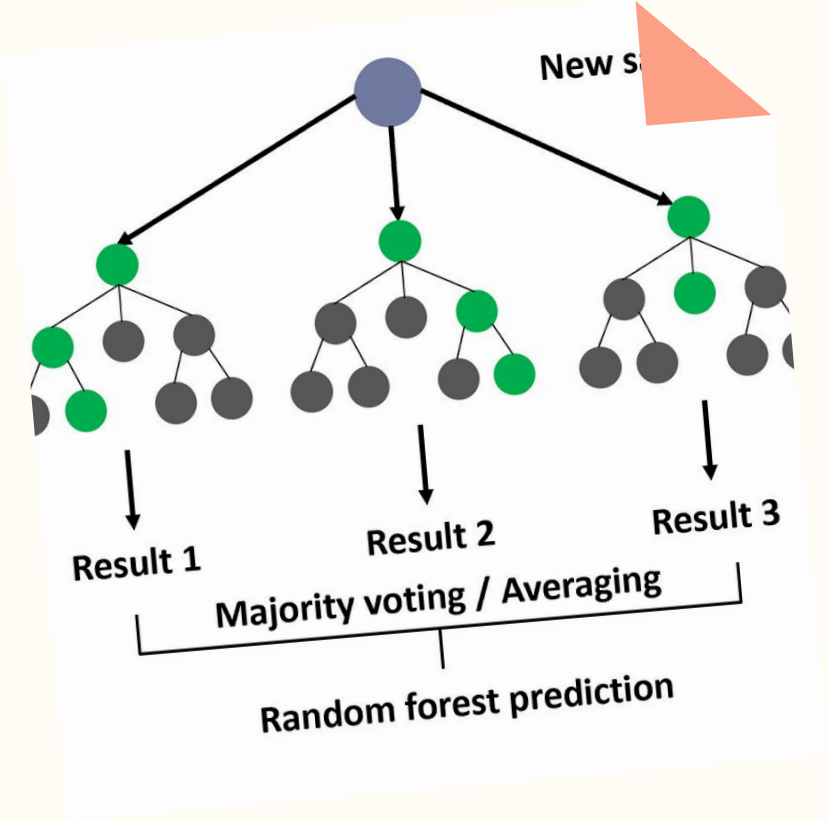
	ROC AUC Score
Logistic Regression	81.9%
Decision Tree	74.7%
Random Forest	85.4%
Gradient Tree Boosting	82.4%
Multi Layer Perceptron	81.7%

Optimizing through Hyperparameter Tuning

Cross-validation: Used for hyperparameter tuning and model selection.

Grid Search: Optimized hyperparameters for Logistic Regression, Random Forest, and Gradient Boosting models.

Improved Metrics: Fine-tuning improved ROC AUC and F1 scores



Conclusion

Best Model: Random Forest Classifier achieved the highest performance with a ROC AUC score of 0.854 and accuracy of 0.806.

Questions?