Credit Card Approval Prediction

Nikunjkumar Mahida

Introduction

Project Purpose:
- Utilizing machine learning techniques for credit card approval prediction.

Importance of Credit Risk Assessment:
- Critical for financial institutions to mitigate potential losses.

Dataset Source and Structure:
- Utilized datasets: train_data.csv and test_data.csv.
- Structure includes features like demographics, financial indicators, and a target variable indicating high-risk applicants.

Relevance of the Project:
- Improves decision-making processes within financial institutions.
- Leads to better risk management strategies.
- Aims to enhance customer satisfaction.

Societal Impact:
- Facilitates access to financial services for deserving individuals.
- Safeguards against fraudulent activities.

Data Overview

Dataset Summary:

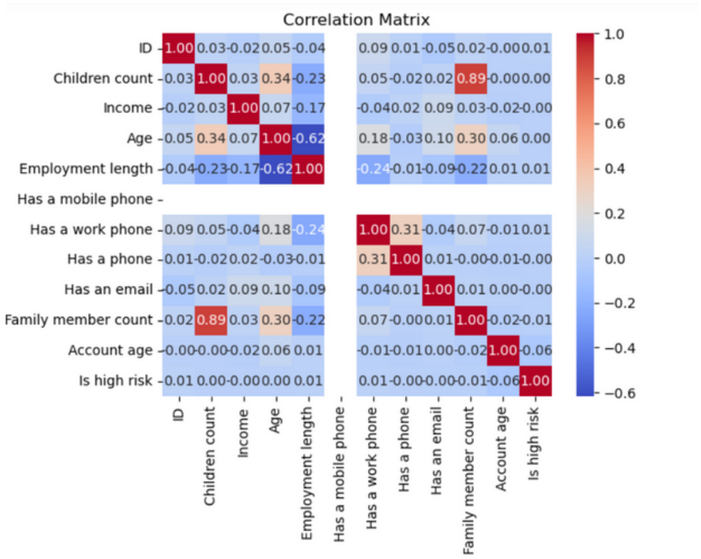
- The dataset comprises two main files: train_data.csv and test_data.csv, which contain information about credit card applicants.
- Each file consists of multiple entries, with each entry representing a single applicant.
- The dataset includes various features such as demographic information (e.g., age, gender, marital status), financial indicators (e.g., income, employment status), and a target variable indicating whether an applicant is classified as high risk or not.

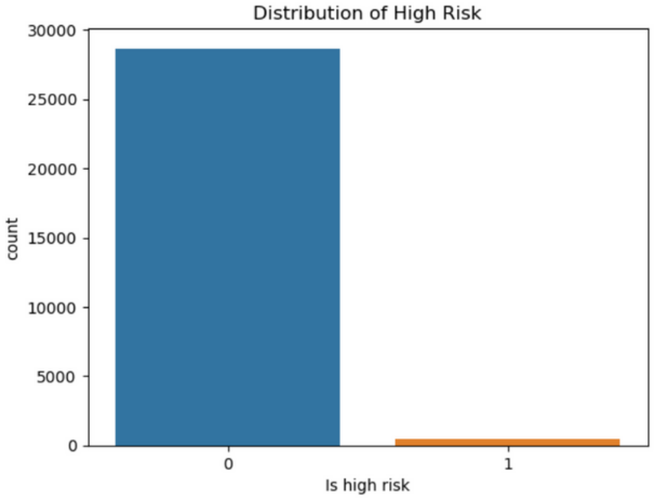
	ID	Gender	Has a car	Has a property	Children count	Income	Employment status	Education level	Marital status	Dwelling	Age	Employment length	Has a mobile phone	Has a work phone	Has a phone	Has an email	Job title	Family member count	Account age	Is high risk
0	5037048	M	Y	Y	0	135000.0	Working	Secondary / secondary special	Married	With parents	-16271	-3111	1	0	0	0	Core staff	2.0	-17.0	0
1	5044630	F	Y	N	1	135000.0	Commercial associate	Higher education	Single / not married	House / apartment	-10130	-1651	1	0	0	0	Accountants	2.0	-1.0	0
2	5079079	F	N	Y	2	180000.0	Commercial associate	Secondary / secondary special	Married	House / apartment	-12821	-5657	1	0	0	0	Laborers	4.0	-38.0	0
3	5112872	F	Y	Y	0	360000.0	Commercial associate	Higher education	Single / not married	House / apartment	-20929	-2046	1	0	0	1	Managers	1.0	-11.0	0
4	5105858	F	N	N	0	270000.0	Working	Secondary / secondary special	Separated	House / apartment	-16207	-515	1	0	1	0	NaN	1.0	-41.0	0

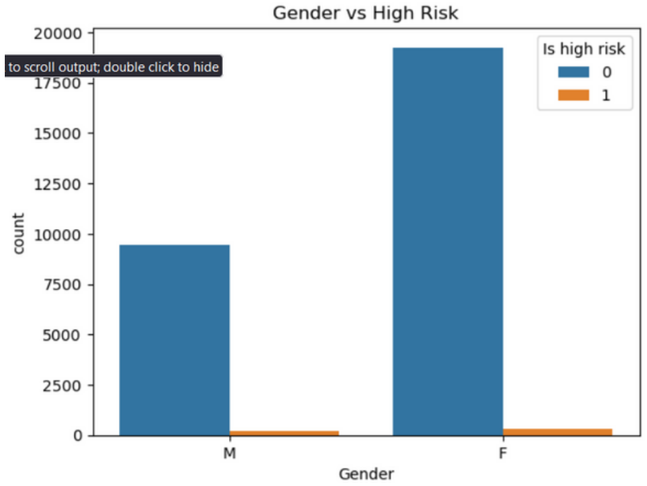
Data Visualization

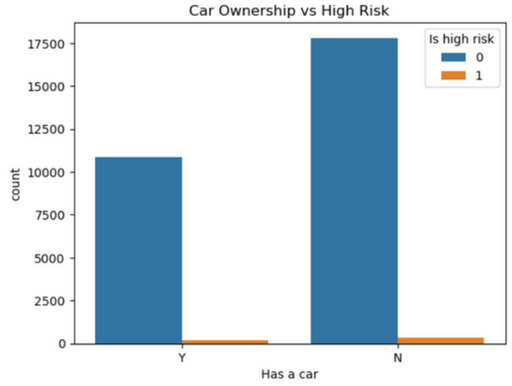
Visualizations:

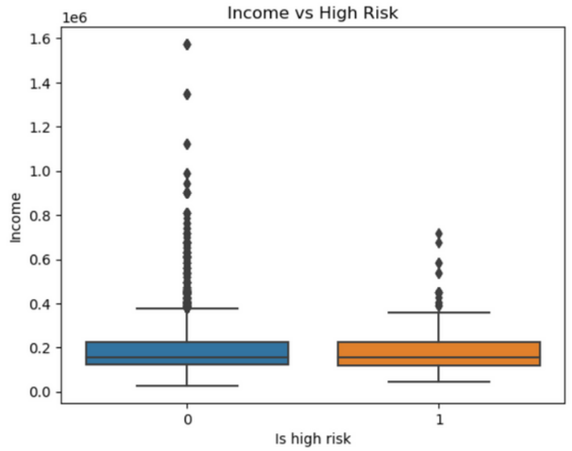
- To gain insights into the dataset's characteristics and distribution, various visualizations were employed.
- A countplot of the target variable distribution was used to visualize the proportion of high-risk applicants in the dataset.
- Additionally, a heatmap of the correlation matrix was generated to examine the relationships between different features and the target variable.
- Individual countplots were created for selected categorical variables, while boxplots were used for numerical variables to visualize their distributions and potential relationships with the target variable.

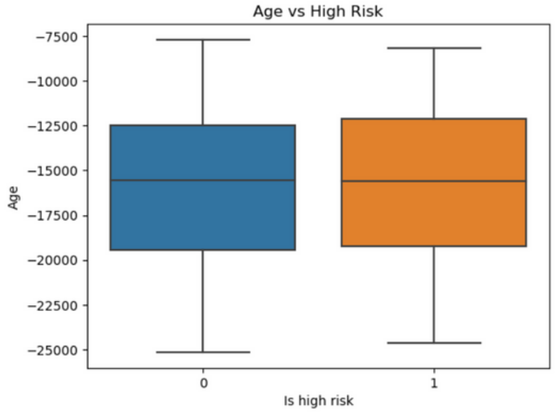

Correlation Matrix
ID
Children count
Income
Age
Employment length
Has a mobile phone
Has a work phone
Has a phone
Has an email
Family member count
Account age
Is high risk


Distribution of High Risk
count
Is high risk


Gender vs High Risk
to scroll output; double click to hide
Is high risk
count
Gender


Car Ownership vs High Risk
Is high risk
count
Has a car


Income vs High Risk
1e6
Income
Is high risk


Age vs High Risk
Age
Is high risk

Data Preprocessing

1. Preprocessing Steps:
    - Data preprocessing is a crucial step in preparing the dataset for modeling and analysis.
    - Several steps were undertaken to ensure the dataset's cleanliness, consistency, and suitability for machine learning algorithms.
2. Column Removal:
    - Unnecessary columns that do not contribute to the prediction task were removed from the dataset.
    - For example, the "ID" column was dropped as it does not provide meaningful information for credit card approval prediction.
3. Handling Missing Values:
    - Missing values in the dataset were addressed using the median imputation strategy.
    - Features with missing values, such as "Family member count," "Account age," and "Employment length," were imputed with the median value of each respective feature.
    - This approach ensures that missing values are replaced with plausible estimates, maintaining the dataset's integrity for subsequent analysis.

Data Preprocessing

4. Categorical Variable Encoding:
    - Categorical variables in the dataset were encoded into numerical format using LabelEncoder.
    - This transformation is necessary as most machine learning algorithms require numerical input.
    - Each categorical variable was encoded with a unique integer, enabling the algorithm to process and learn from these features effectively.
5. Quality Assurance:
    - Data preprocessing ensures that the dataset is clean, consistent, and free from inconsistencies that could potentially impact model performance.
    - By standardizing the format and structure of the dataset, we facilitate the modeling process and improve the accuracy and reliability of the predictive models.
6. Data Integrity Preservation:
    - Throughout the preprocessing steps, care was taken to preserve the integrity and representativeness of the original data.
    - By addressing missing values and encoding categorical variables appropriately, we maintain the dataset's fidelity while preparing it for modeling.

Model Selection and data splitting

1. Model Selection:
   - Choosing an appropriate model is crucial for building an effective credit card approval prediction system.
   - In this project, a Random Forest classifier was selected for its robustness, scalability, and ability to handle complex datasets with high dimensionality.
   - Random Forest is an ensemble learning technique that combines multiple decision trees to improve predictive performance and reduce overfitting.
2. Data Splitting:
   - Before training the model, the dataset was split into features (X) and the target variable (y).
   - Further, the data was divided into training and validation sets using the train_test_split function from scikit-learn.
   - This step ensures that the model's performance can be evaluated on unseen data, helping to assess its generalization capability.

Model Training

1. Feature Scaling:
    - Feature scaling is essential for ensuring that all features contribute equally to the model's learning process.
    - StandardScaler from scikit-learn was used to scale the features by removing the mean and scaling to unit variance.
    - Scaling ensures that features are on the same scale, preventing certain features from dominating others during model training.
2. Model Training:
    - Once the data was preprocessed and split, a Random Forest classifier was trained on the training set.
    - The RandomForestClassifier from scikit-learn was configured with 100 estimators (trees) to build a robust ensemble model.
    - Training the model involves fitting the classifier to the training data, allowing it to learn patterns and relationships between features and the target variable.

Evaluation of Model:

Model Robustness and Scalability:

- Random Forest classifiers are known for their robustness to noise and outliers and their ability to handle large datasets with high dimensionality.
- By training a Random Forest classifier, we aim to build a model that can effectively predict credit card approvals while minimizing overfitting and maintaining scalability.

Evaluation Strategy:

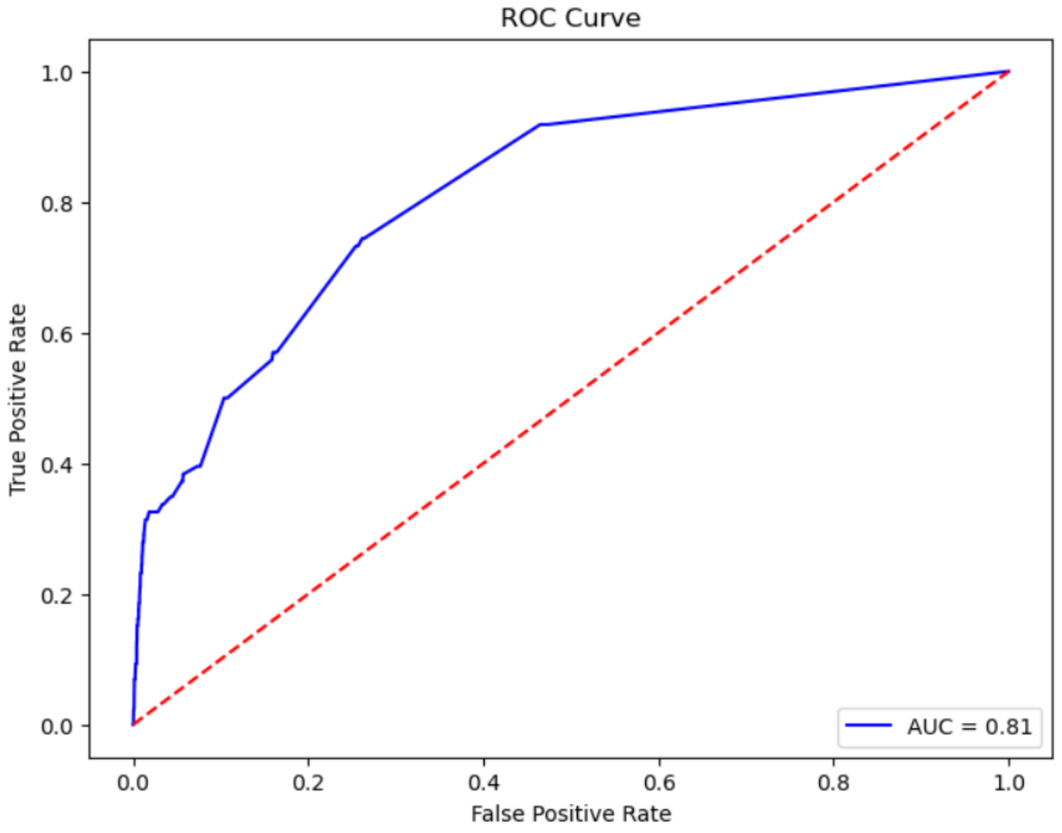
- In addition to traditional evaluation metrics such as accuracy, precision, recall, and F1-score, the model's performance will also be assessed using the AUC-ROC curve.
- The AUC-ROC curve provides a comprehensive evaluation of the classifier's performance across various threshold values, depicting the trade-off between true positive rate (sensitivity) and false positive rate (1-specificity).
- A higher AUC-ROC score indicates better discrimination between positive and negative instances, reflecting the model's ability to correctly classify high-risk and low-risk credit card applicants.

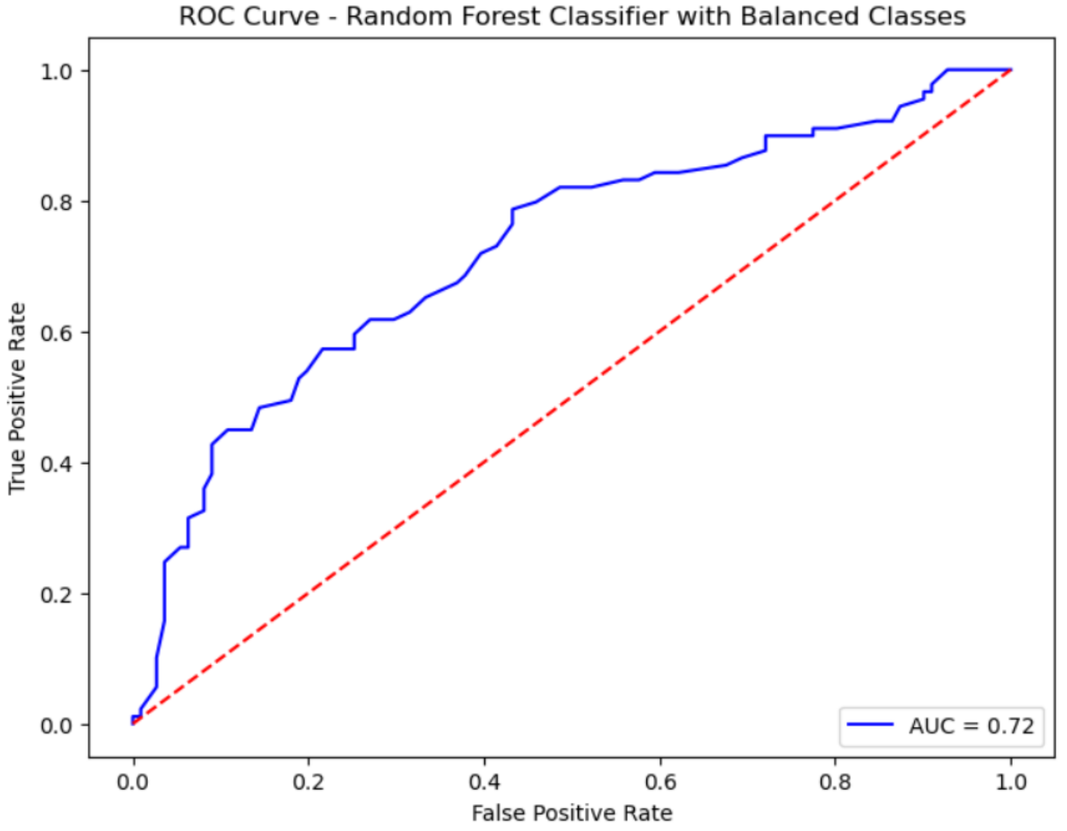
Problems

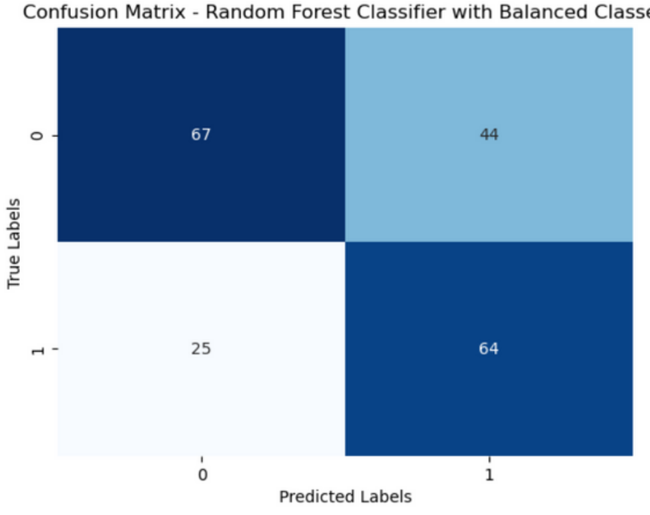
The Dataset had around 28000 values of 0 class in the target variable while it had just 500 values of 1 class. This made the dataset very unbalanced and gave a very low precision and recall. Even after using the techniques like regularization of parameters ,the model failed to meet the requirements.

So to solve the problem i trained the model with same number of 0 class values and 1 class values in the target variable. The precision and recall values showed a great improvement but the model slightly showed a lesser accuracy then before. The highly unbalanced nature of this dataset is the main reason behind the poor performance of the model used.

Even after training it with many different classification models Random Forest Regressor gave the best Accuracy.


ROC Curve
True Positive Rate
False Positive Rate
AUC = 0.81
0.0
0.2
0.4
0.6
0.8
1.0


ROC Curve - Random Forest Classifier with Balanced Classes
True Positive Rate
False Positive Rate
AUC = 0.72
0.0
0.2
0.4
0.6
0.8
1.0


Confusion Matrix - Random Forest Classifier with Balanced Classes
True Labels
Predicted Labels
67
44
25
64
0
1

Thank you

Nikunjkumar Mahida

9510575076

nikunjjmahida@gmail.com