CREDIT CARD PREDICTION

Presented By : Tanaya Tipre

Mentorness | MIP-ML-09

INTRODUCTION

Credit card prediction plays a pivotal role in determining creditworthiness, managing risk, and ensuring fair access to financial services.

Our primary objective is to develop a robust machine learning model that predicts an individual's eligibility for a credit card based on various demographic and financial attributes.

PROBLEM STATEMENT

- The primary objective of this project is to predict the approval or rejection of credit card applications.

- This prediction is crucial for minimizing the risk of default and fraud for financial institutions while ensuring fair and accessible credit opportunities for consumers.

DATA COLLECTION

Train Dataset:

https://drive.google.com/file/d/1ONmjxVLbAvMoas5Rqai9pZ0TbWCrsDOC/view?usp=drive_link

Test Dataset:

https://drive.google.com/file/d/1WbUpvudclwmaBzJwPwtE4kDw6uw5t7aq/view?usp=drive_link

EXPLORATORY DATA ANALYSIS

Categorical columns

1. Gender
2. Has a car
3. Has a property
4. Employment status
5. Education level
6. Marital status
7. Dwelling
8. Employment length
9. Has a mobile phone
10. Has a work phone
11. Has a phone
12. Has an email
13. Job title
14. Is high risk

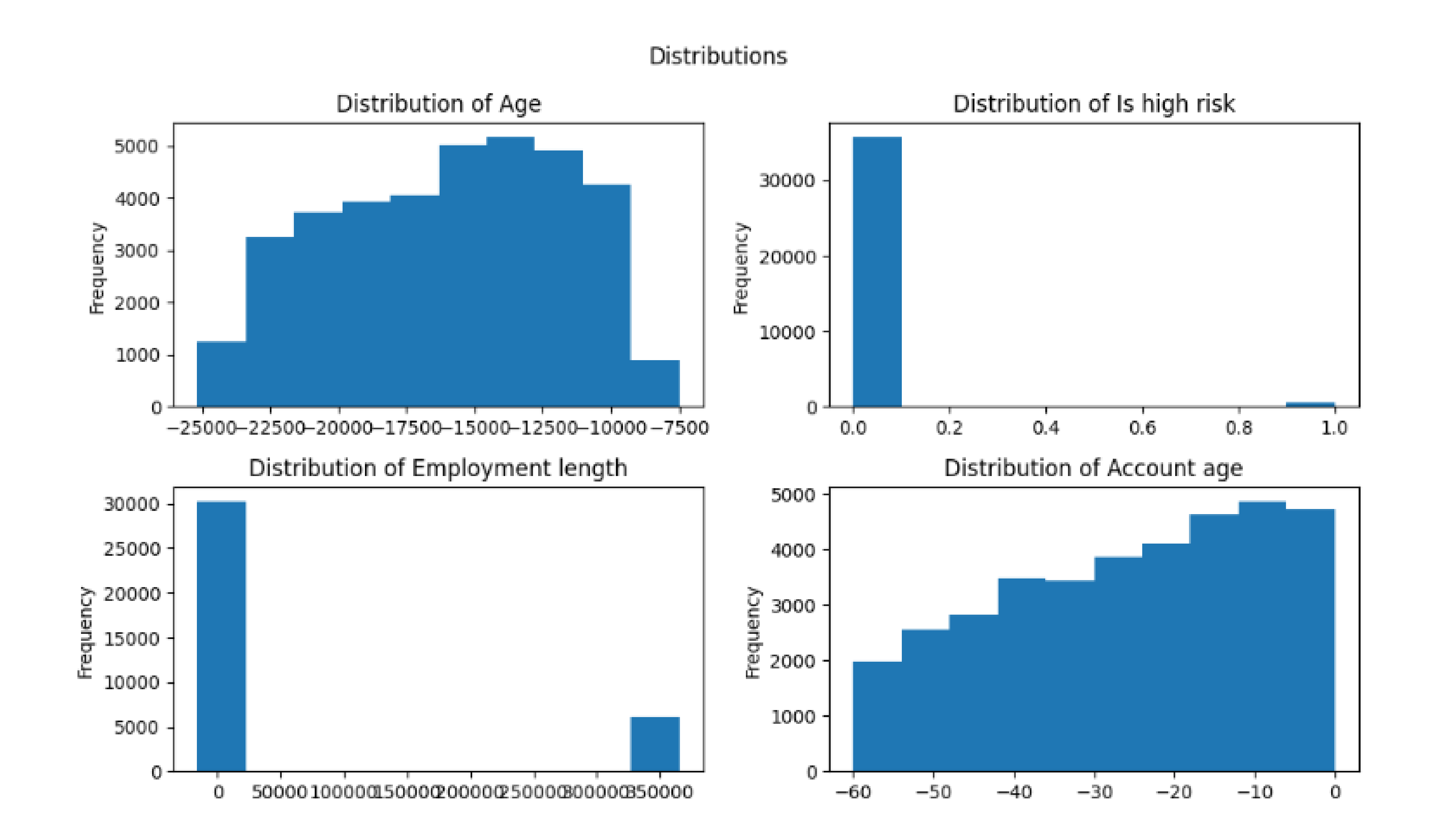
- Number of categorical columns: 14
- Number of continuous columns: 6

Continuous columns

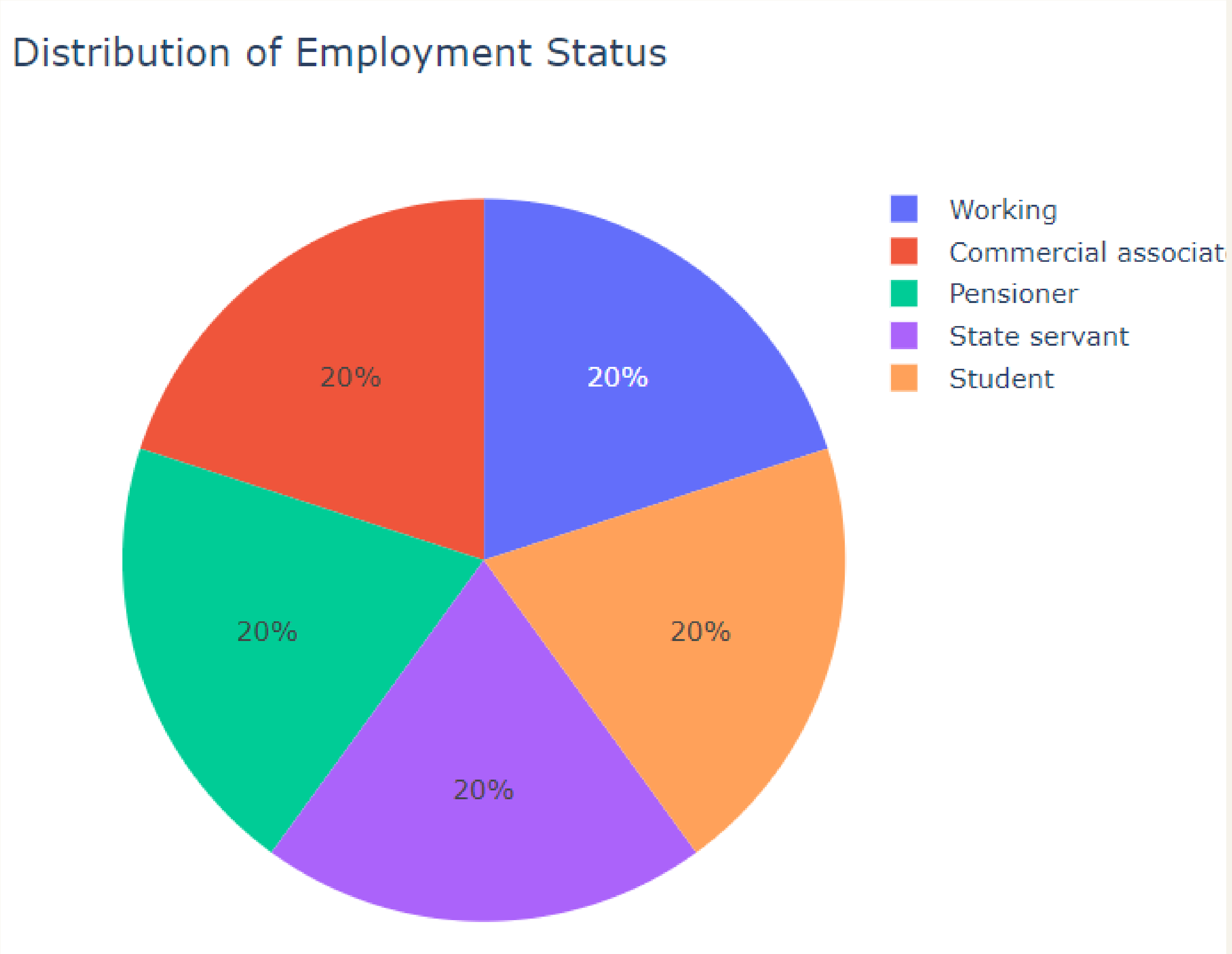
1. ID
2. Children count
3. Income
4. Age
5. Family member count
6. Account age

EXPLORATORY DATA ANALYSIS

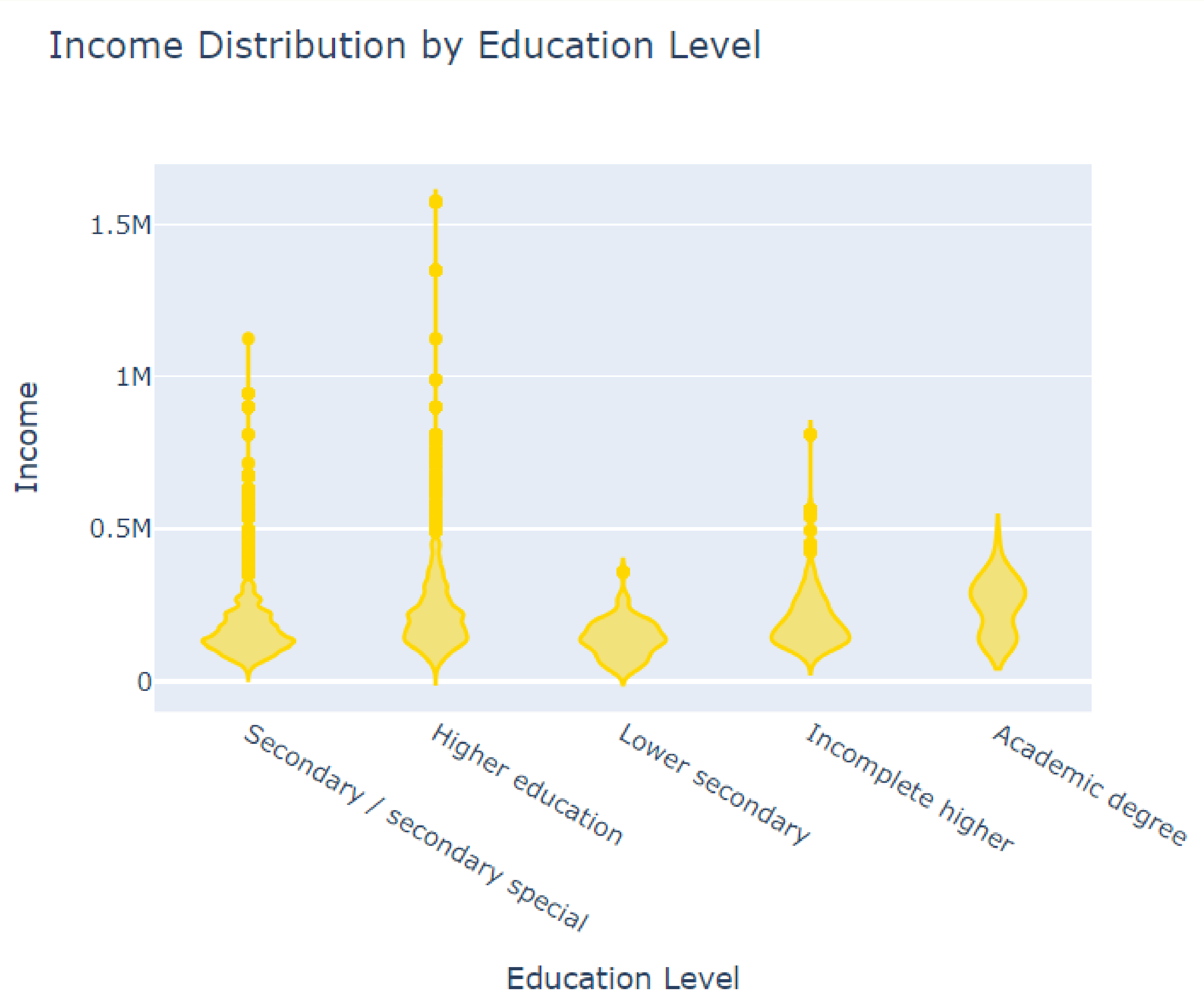
Distribution of different features of the dataset:



EDA



Distribution of Employment Status
Working
Commercial associat
Pensioner
State servant
Student
20%
20%
20%
20%
20%



Income Distribution by Education Level
Income
1.5M
1M
0.5M
0
Secondary / secondary special
Higher education
Lower secondary
Incomplete higher
Academic degree
Education Level

FEATURE ENGINEERING

- ## Label Encoding

Label encoding assigns a unique numerical value to each category in a categorical column. This encoding preserves the ordinal relationship between categories.

- ## Min-Max Scaler

Min-max scaling normalizes the continuous variables within a specified range, typically between 0 and 1. By scaling features to a common range, the model can converge faster and make more accurate predictions.

DATA PREPROCESSING

- **Removed**
  - 'ID': Unnecessary unique identifier.
  - 'Has a mobile phone': Limited predictive power.
  - 'Children count': Low relevance to eligibility.
  - 'Age': Redundant with other features.
  - 'Family member count': Limited impact on prediction.

MODEL SELECTION

- Logistic Regression: Linear model for binary classification.
- Random Forest: Ensemble method combining multiple decision trees.
- XGBoost (Extreme Gradient Boosting): Boosted tree ensemble algorithm.
- K-Nearest Neighbors (KNN): Instance-based learning for classification.

METHODOLOGY-LR

Logistic Regression

It is well-suited for binary classification problems, such as predicting credit card eligibility (eligible or not eligible), which aligns with the nature of the problem statement.

```
Confusion Matrix :
 [[2551 1132]
 [1059 2550]]
Classification Report:
              precision    recall  f1-score   support

           0       0.71      0.69      0.70      3683
           1       0.69      0.71      0.70      3609

    accuracy                           0.70      7292
   macro avg       0.70      0.70      0.70      7292
weighted avg       0.70      0.70      0.70      7292

The test accuracy of Logistic Regression is :  69.95337356006583 %
```

METHODOLOGY-RFC

Random Forest Classifier

Random Forest can capture complex, non-linear relationships between features and the target variable, making it suitable for a wide range of classification tasks.

```
Accuracy Score is: 0.775644541963796
              precision    recall  f1-score   support

           0       0.79      0.76      0.77      3683
           1       0.77      0.79      0.78      3609

    accuracy                           0.78      7292
   macro avg       0.78      0.78      0.78      7292
weighted avg       0.78      0.78      0.78      7292
```

METHODOLOGY-XGB

```
Accuracy Score is: 0.8289906747120132
confusion matrix
[[3327  356]
 [ 891 2718]]


0.8289906747120132
              precision    recall  f1-score   support

           0       0.79      0.90      0.84      3683
           1       0.88      0.75      0.81      3609

    accuracy                           0.83      7292
   macro avg       0.84      0.83      0.83      7292
weighted avg       0.84      0.83      0.83      7292
```

XGBoost Classifier

XGBoost is an implementation of gradient boosting, a powerful ensemble learning technique that builds a sequence of trees, where each tree corrects the errors of the previous one, leading to high predictive accuracy.

METHODOLOGY-KNN

```
confusion matrix
[[3214  469]
 [ 177 3432]]


0.9114097641250686
              precision    recall  f1-score   support

           0       0.95      0.87      0.91      3683
           1       0.88      0.95      0.91      3609

    accuracy                           0.91      7292
   macro avg       0.91      0.91      0.91      7292
weighted avg       0.91      0.91      0.91      7292
```

KNN Classifier

KNN takes into account the local structure of the data, making it effective in capturing complex decision boundaries and handling nonlinear relationships between features and the target variable.

CONCLUSION

LR

It provides a decent performance with an accuracy of 70.0%. However, its precision, recall, and F1 scores are relatively lower compared to other models.

XGB

XGBoost exhibits superior performance compared to Logistic Regression and Random Forest with an accuracy of 82.9%. It achieves high precision and recall for class 0 but slightly lower recall for class 1.

RFC

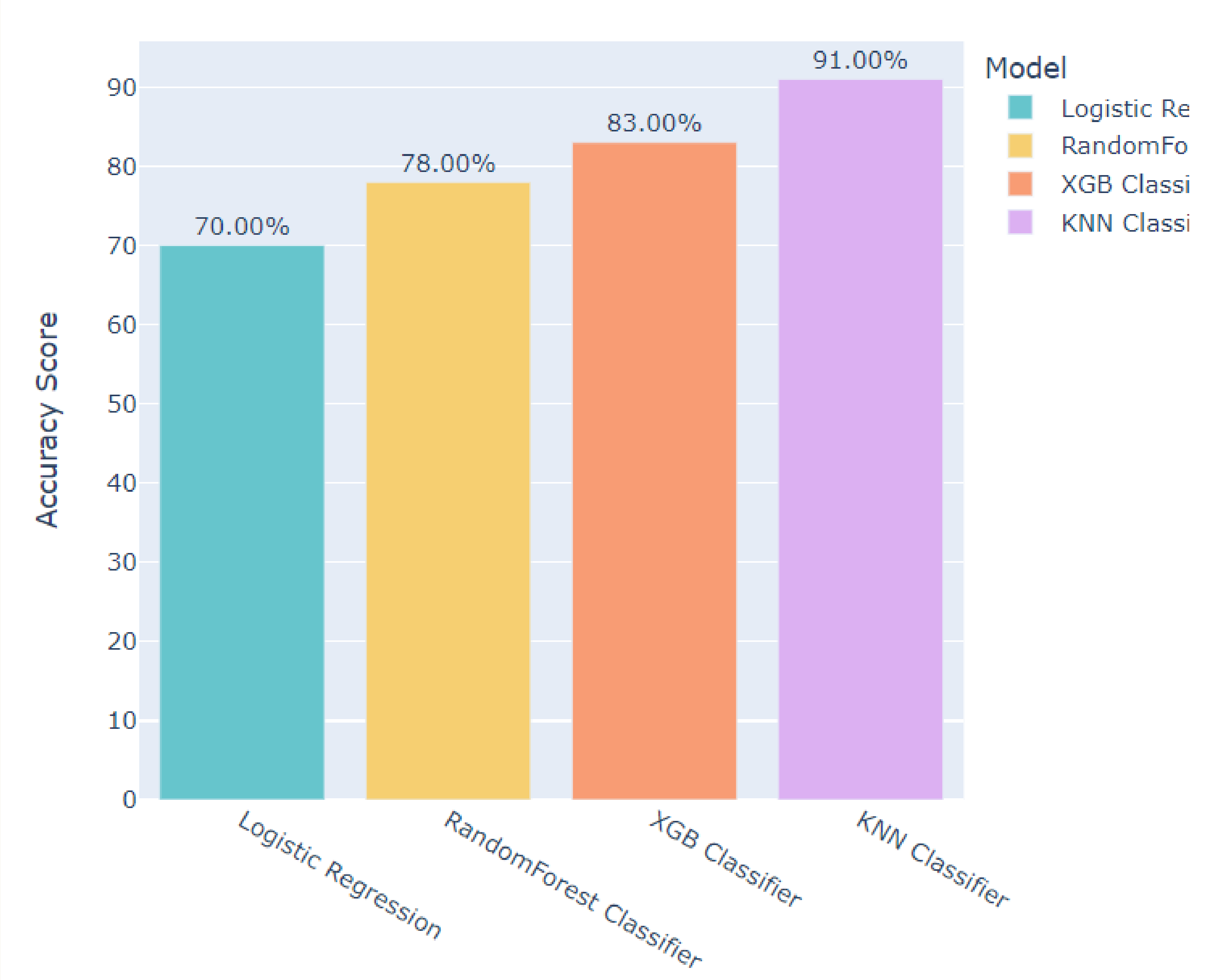
Random Forest performs better than Logistic Regression with an accuracy of 77.6%. It achieves balanced precision, recall, and F1-scores for both classes.

KNN

KNN outperforms all other models with the highest accuracy of 91.1%. It demonstrates excellent precision, recall, and F1 scores for both classes, indicating robust performance across the board.

RESULT

Based on the evaluation metrics and overall performance, K-Nearest Neighbors (KNN) emerges as the best performing model for the credit card prediction task, with an accuracy of 91.1% and balanced performance across all metrics.



RECOMMENDATION

- **Recommendation I**
    - **Based on the results obtained, the recommendation is to deploy K-Nearest Neighbors (KNN) as the primary model for credit card prediction due to its high accuracy and balanced performance across metrics.**

I.

- **Recommendation 2**
    - **Explore ensemble techniques, such as stacking or boosting, to further improve model performance and achieve even higher accuracy in credit card prediction.**

THANK YOU