HOME CREDIT DEFAULT RISK MODEL

created by Group 14

Table of Contents

1. Problem

Problem statement

Many people struggle to get loans due to insufficient or non-existent credit histories. Home Credit strives to broaden financial inclusion for the unbanked population by providing a positive and safe borrowing experience. In order to make sure this underserved population has a positive loan experience. Home Credit makes use of a variety of alternative data to predict their clients' repayment abilities. Doing so will ensure that clients capable of repayment are not rejected and that loans are given with a principal, maturity, and repayment calendar that will empower their clients to be successful.

Data Source

The data used are application train and application test. There are our main table, broken into two files for train(with Target) and test(without Target)

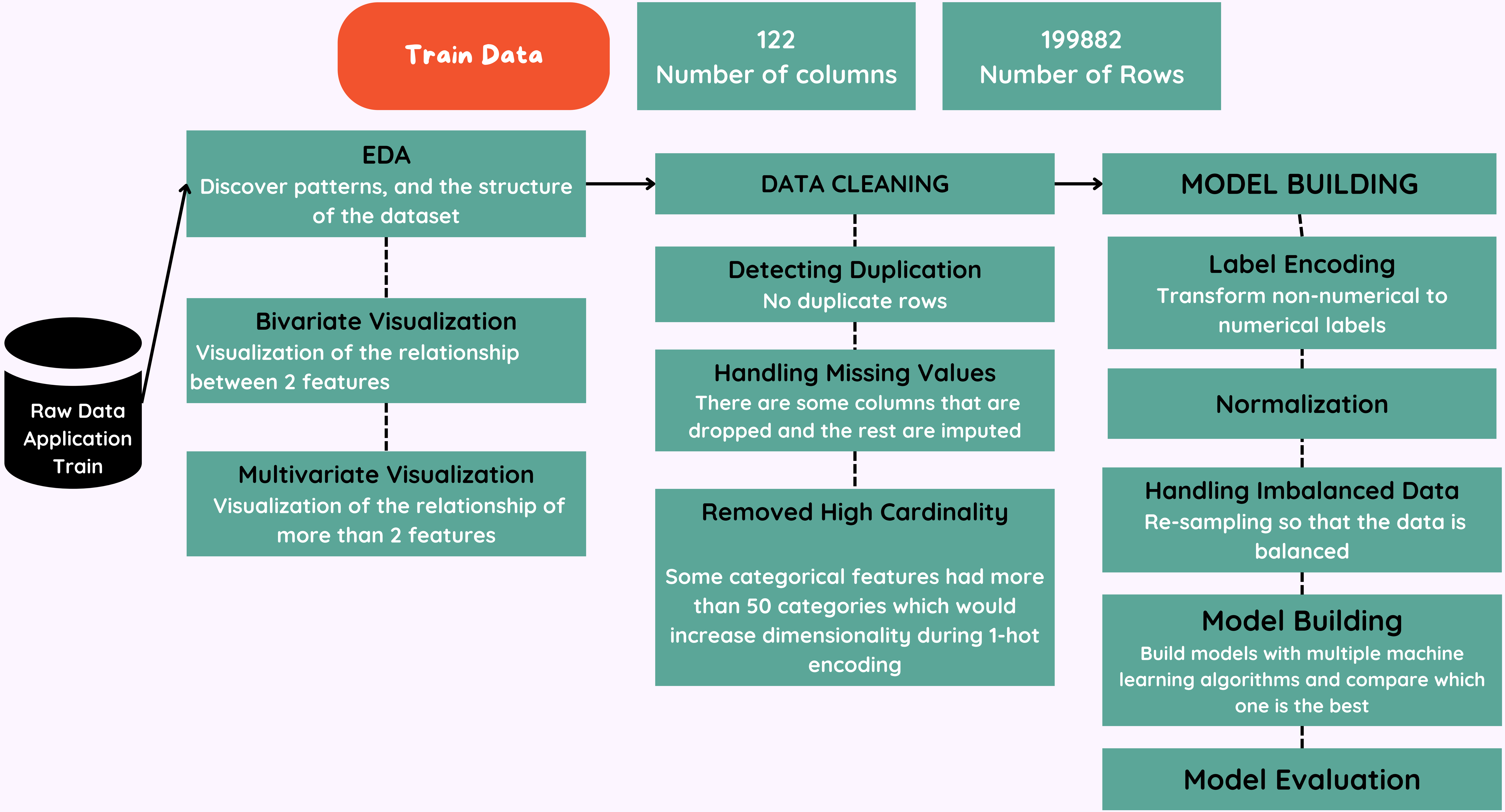
Objective

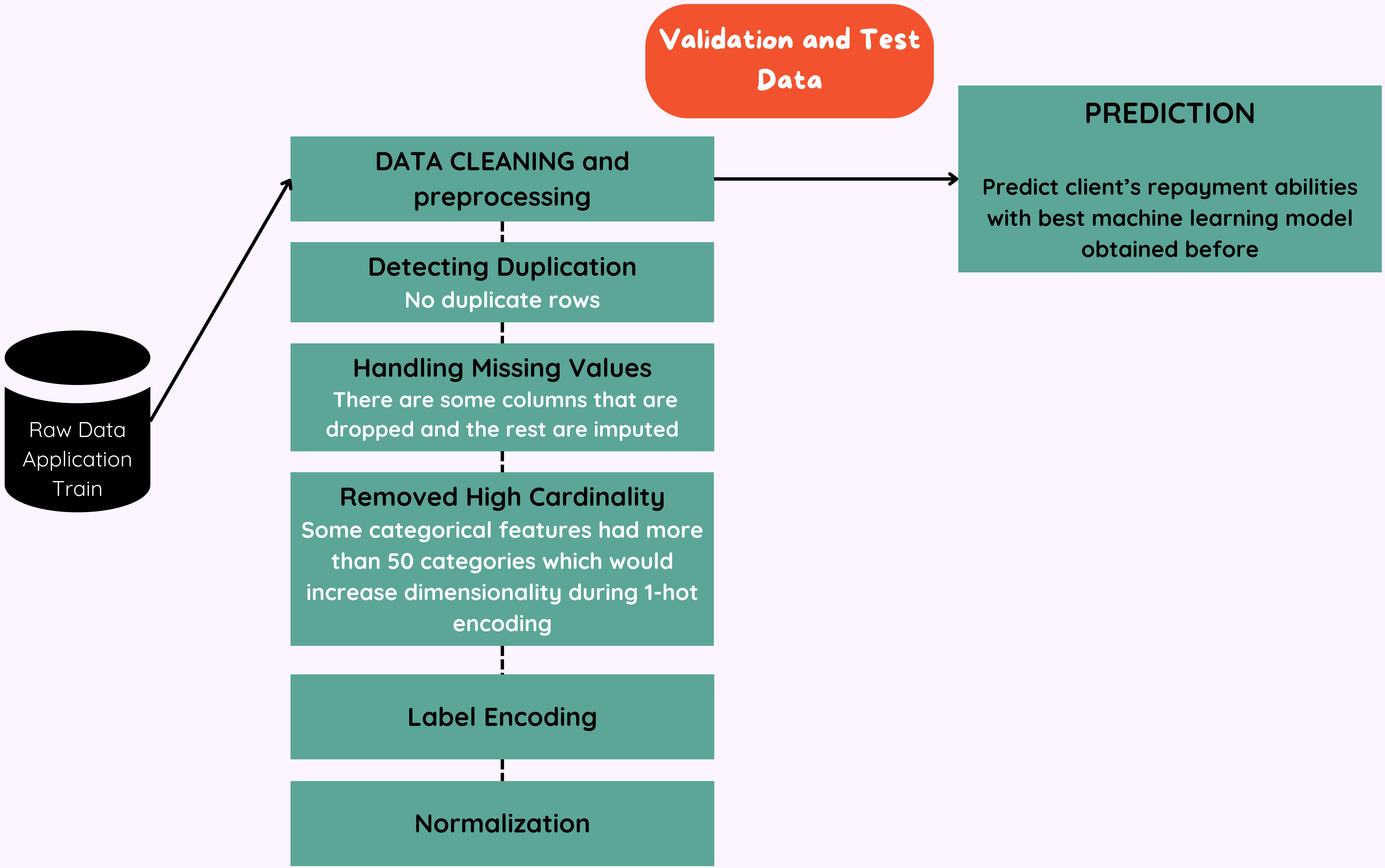
1. **Identify** characteristics of potential people who woll have difficulty repaying loans and who will not.
2. **Predict** client's repayments abilities.

Actions

1. Perform **data cleaning**, and **visualization** for business insights.
2. **Build a models** with machine learning algorithms.
3. Provide **recommendations** for company to increase their clients succeed in applying for loans.

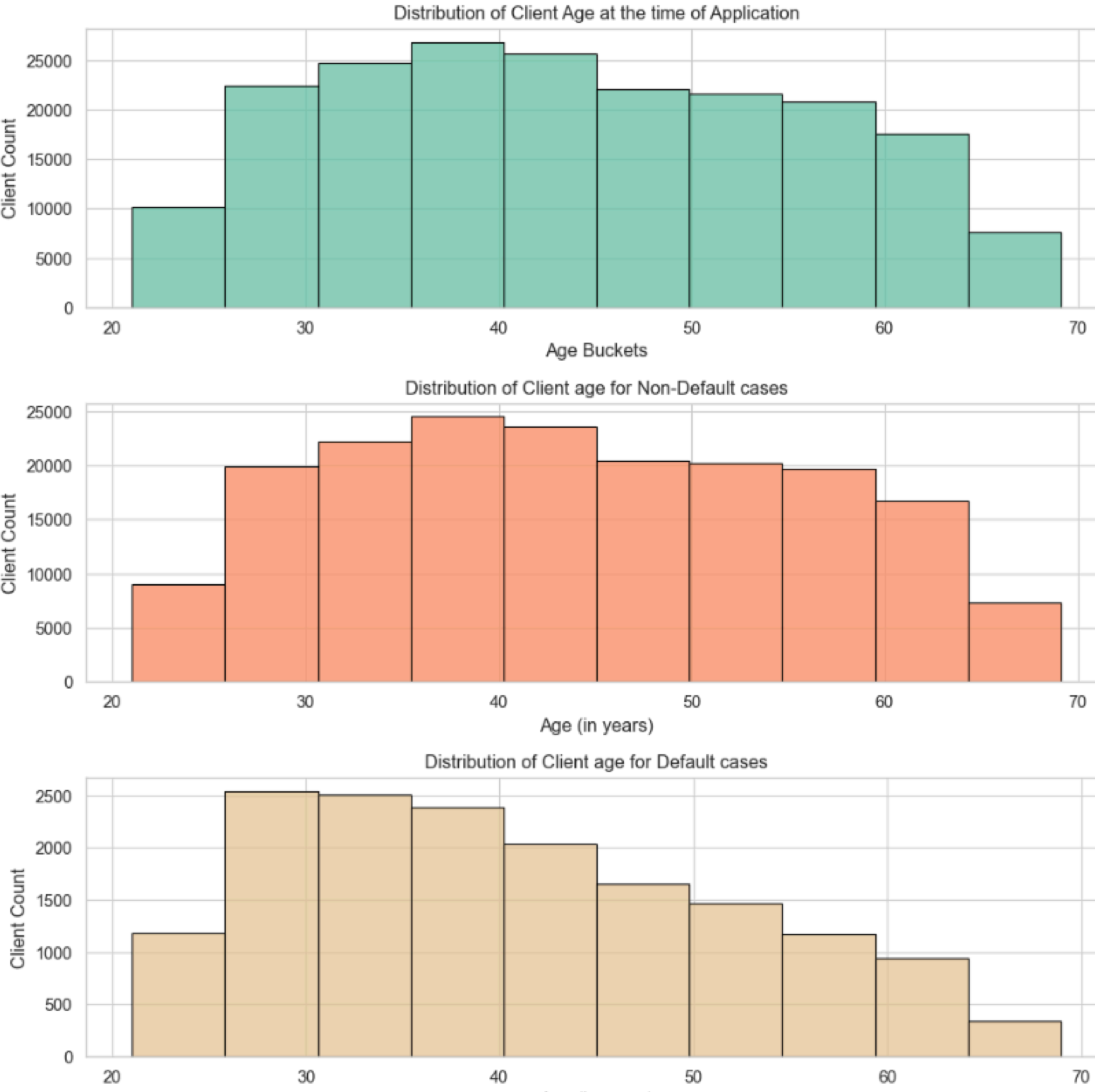
2 Data Preprocessing


Train Data
122
Number of columns
199882
Number of Rows
Raw Data
Application
Train
EDA
Discover patterns, and the structure of the dataset
Bivariate Visualization
Visualization of the relationship between 2 features
Multivariate Visualization
Visualization of the relationship of more than 2 features
DATA CLEANING
Detecting Duplication
No duplicate rows
Handling Missing Values
There are some columns that are dropped and the rest are imputed
Removed High Cardinality
Some categorical features had more than 50 categories which would increase dimensionality during 1-hot encoding
MODEL BUILDING
Label Encoding
Transform non-numerical to numerical labels
Normalization
Handling Imbalanced Data
Re-sampling so that the data is balanced
Model Building
Build models with multiple machine learning algorithms and compare which one is the best
Model Evaluation



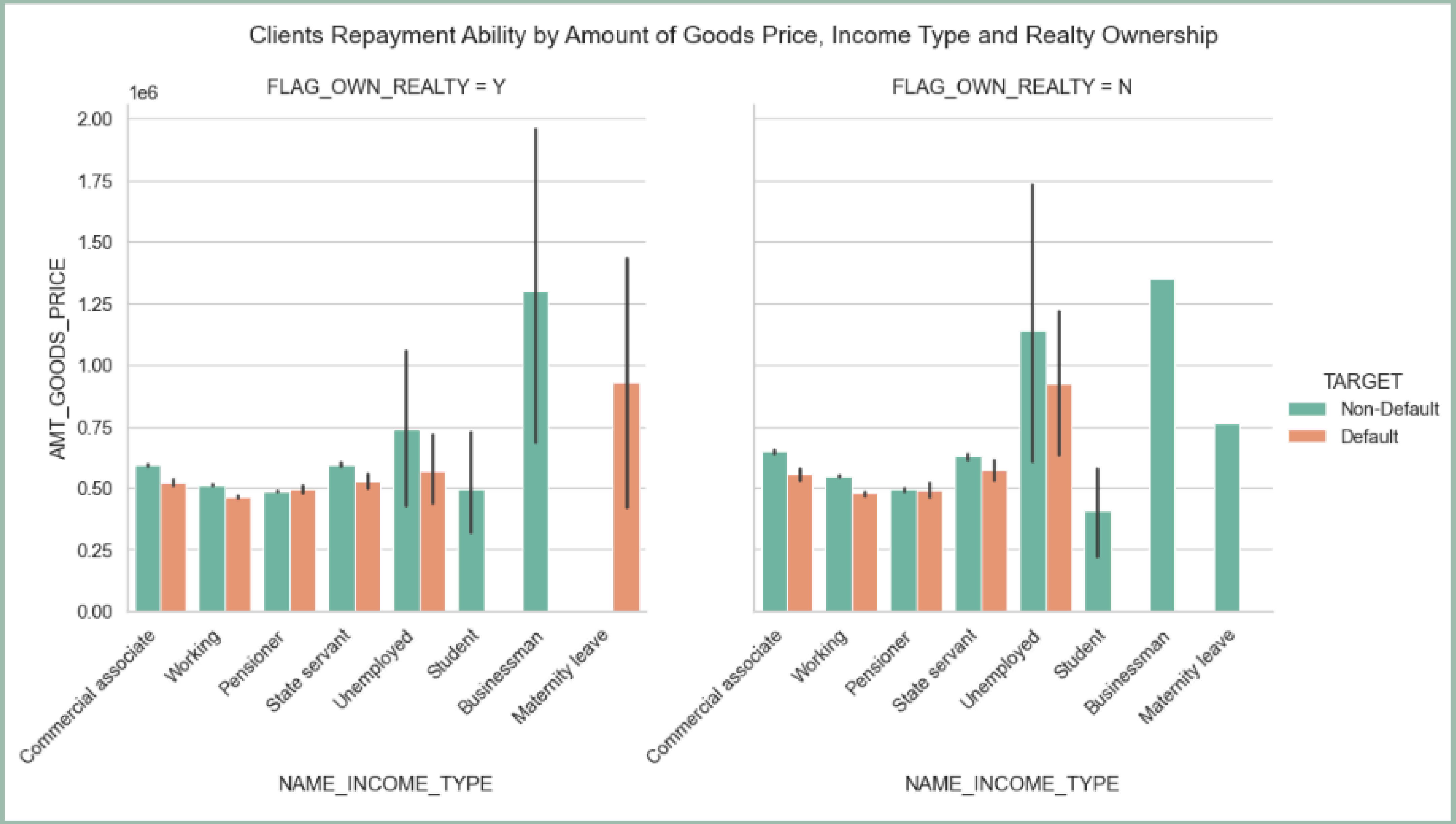
Validation and Test Data
PREDICTION
Predict client's repayment abilities with best machine learning model obtained before
DATA CLEANING and preprocessing
Detecting Duplication
No duplicate rows
Handling Missing Values
There are some columns that are dropped and the rest are imputed
Removed High Cardinality
Some categorical features had more than 50 categories which would increase dimensionality during 1-hot encoding
Label Encoding
Normalization
Raw Data Application Train

3 Data Insights



- Most number of clients who apply for loans are in the range of 35-40 years.
- Meanwhile, the number of applicants for clients aged <25 or age >65 is very low.

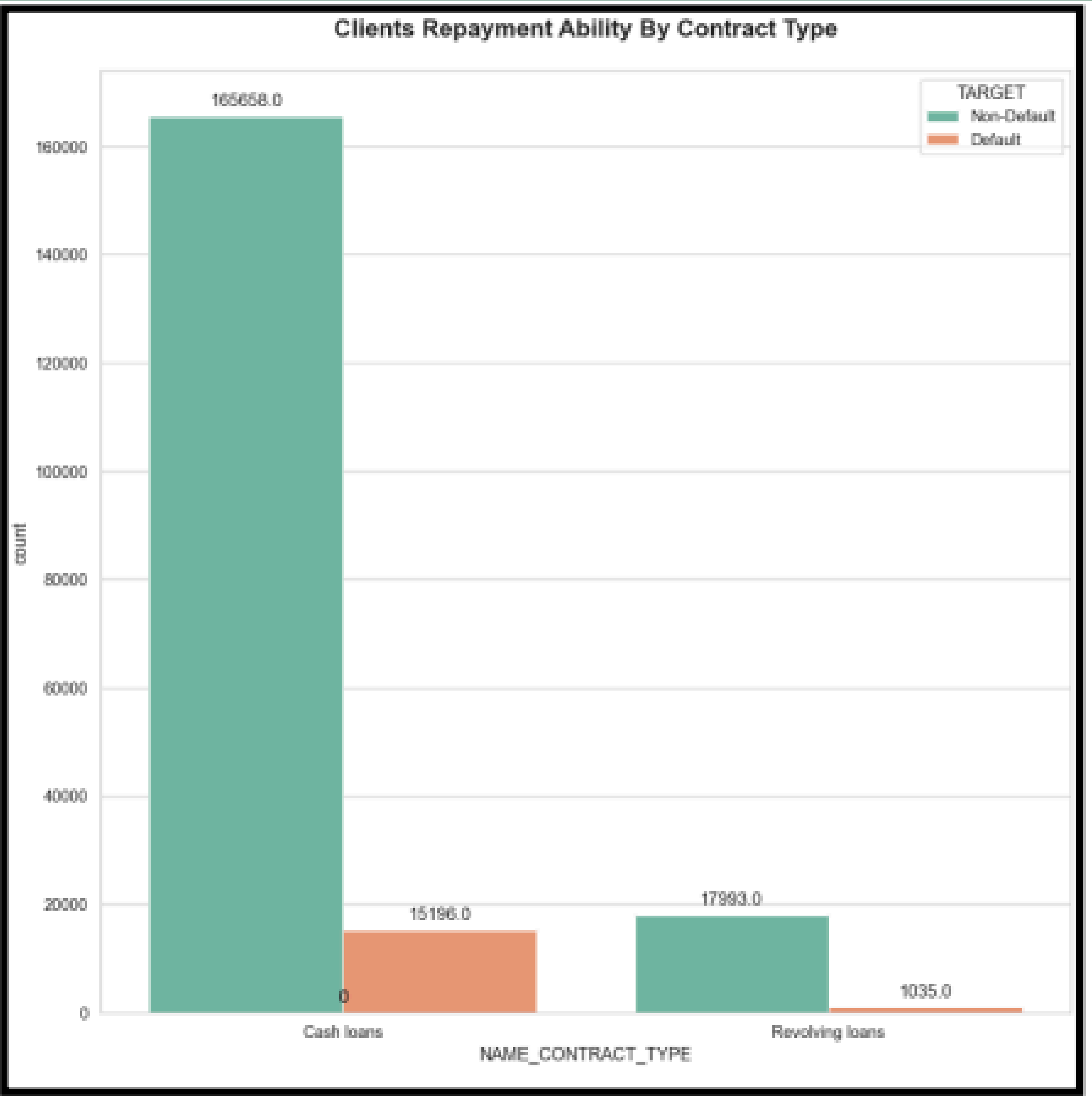
- Clients who have no payment difficulties are clients in the range of 35-45 years. You can target these clients as your priority.
- While clients who have payment difficulties are client the range of 25-35 years.

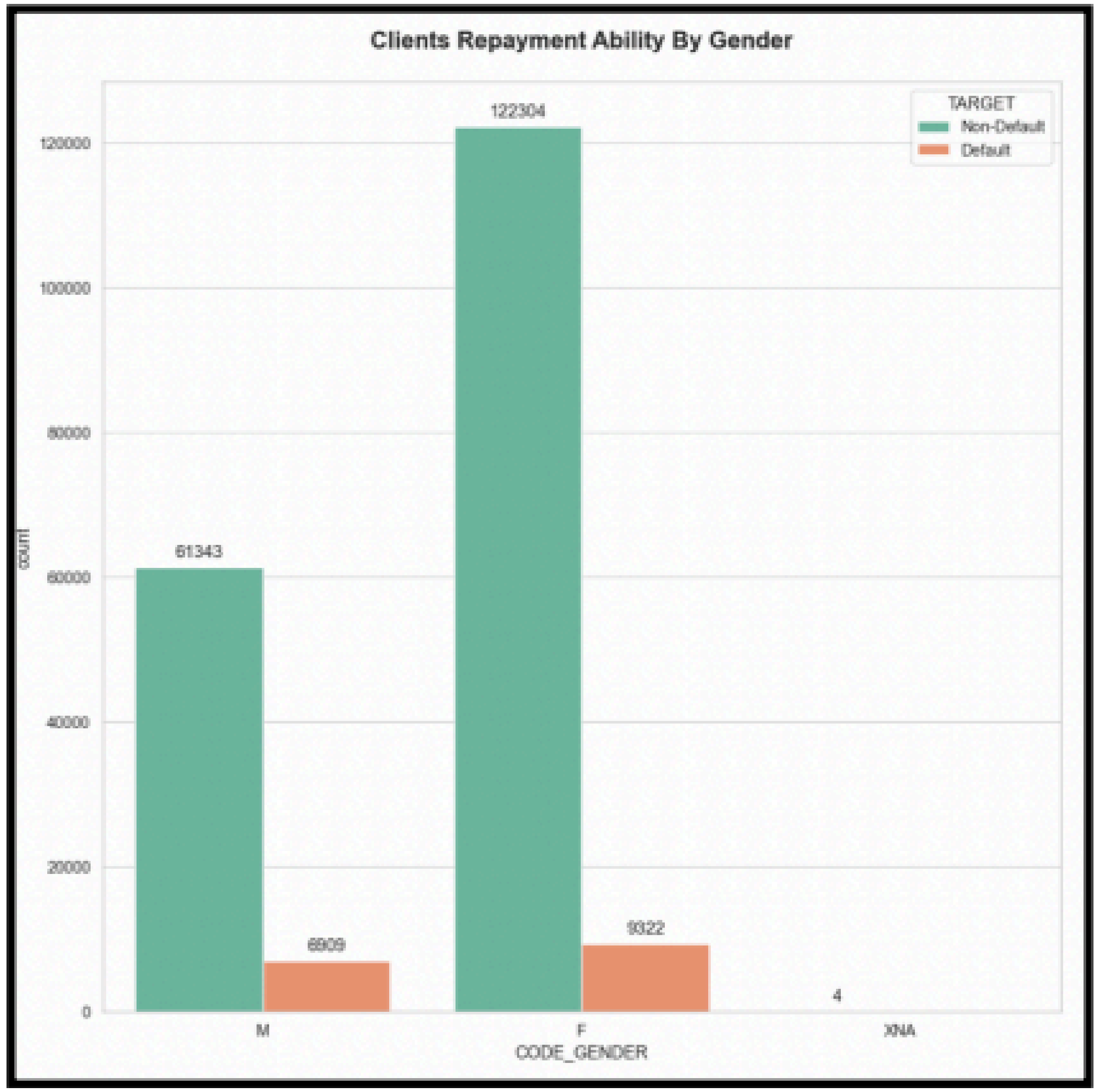


All student clients have no difficulty repaying the loans whether with cash loan or revolving loan for a low to medium credit amount of the loan.

For the income type of maternity leave with cash loans, all the clients have problems repaying the loans for a medium credit amount of the loan. While all clients with maternity leaves and revolving loans have no difficulty repaying the loans.

For unemployed clients with cash loans, more than 50% of clients have problems repaying loans with medium credit amounts of the loan. While all unemployed clients with revolving loans have no difficulty repaying the loan.



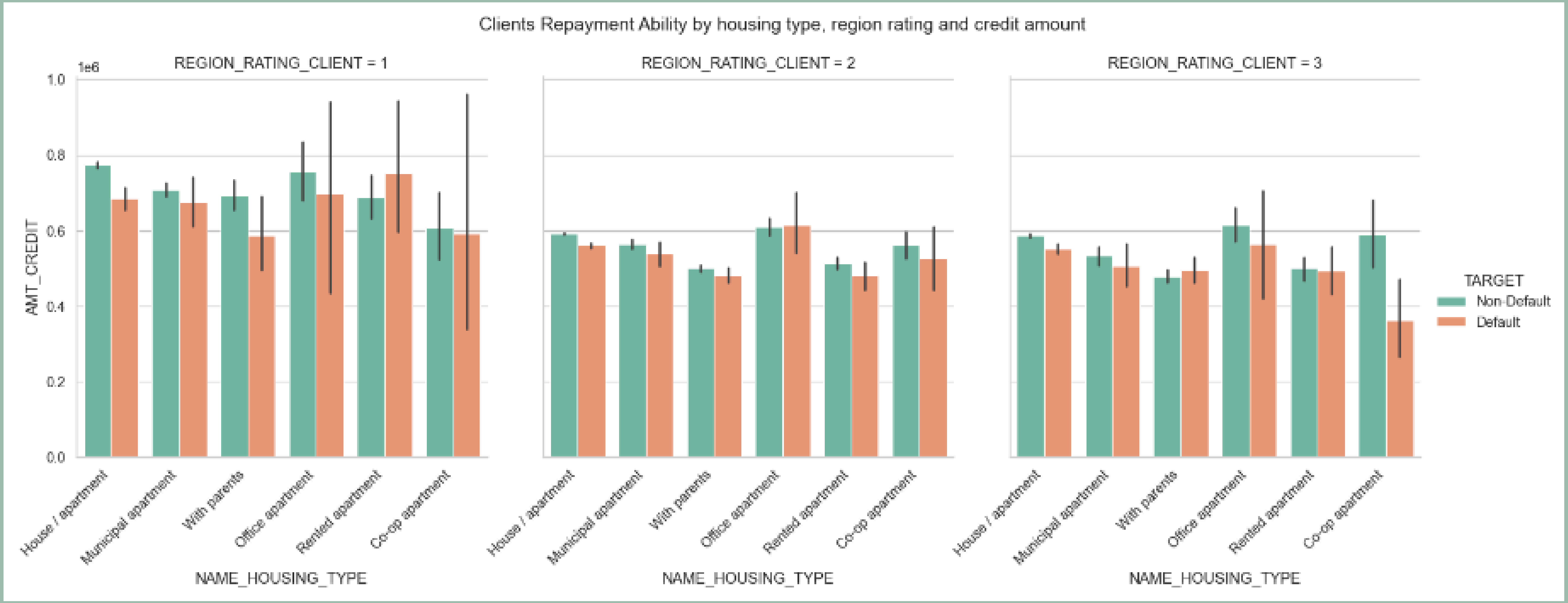


TARGET	Default	Non-Default
NAME_CONTRACT_TYPE		
Cash loans	8.40	91.60
Revolving loans	5.44	94.56

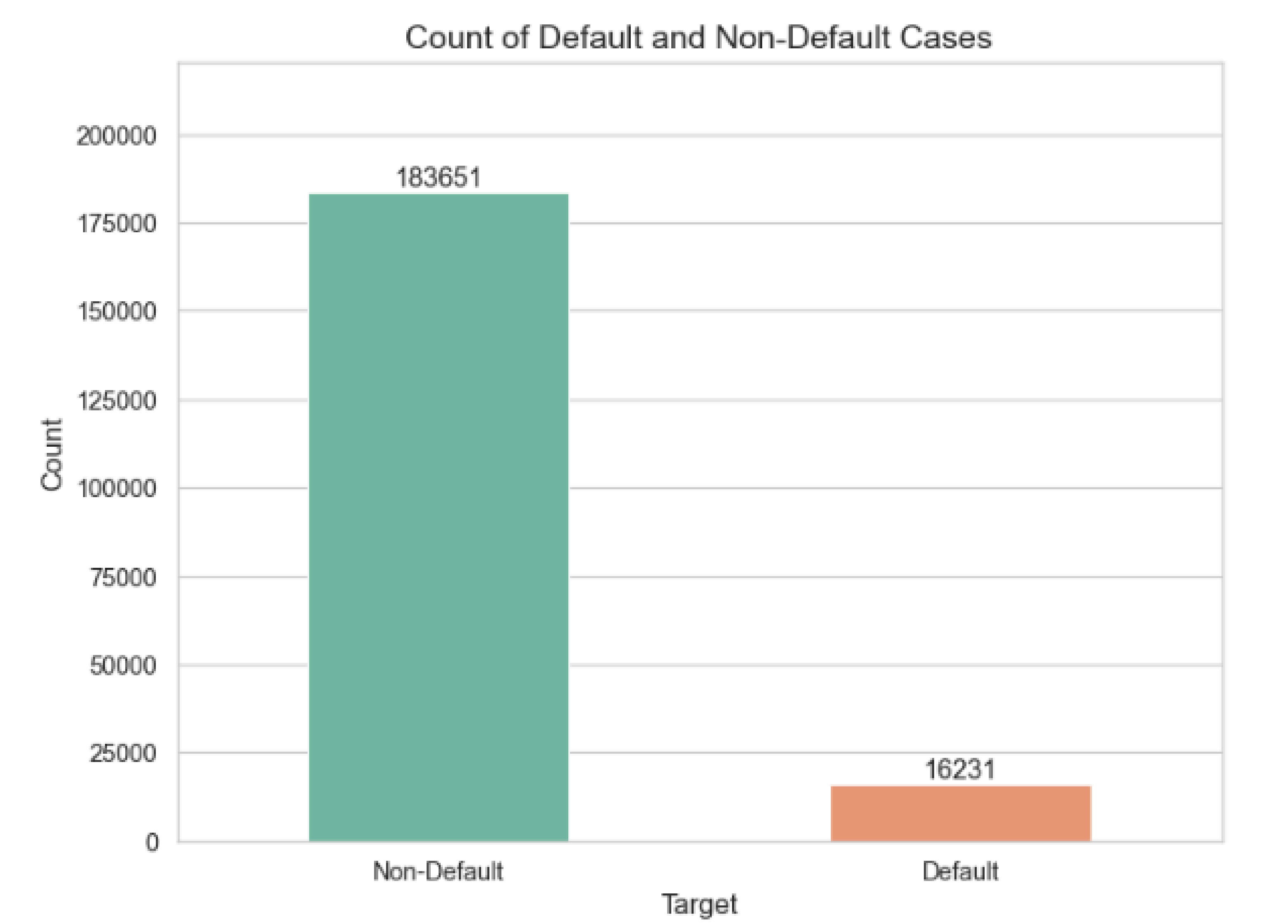
TARGET	Default	Non-Default
CODE_GENDER		
Female	7.08	92.92
Male	10.12	89.88

- We observe that there is a much larger number of clients taking out cash loans compared to revolving loans. Cash loans have higher defaults (8.4%) compared to revolving loans (5.4%

- Women applied for most loans with 202,448 applications versus 105,059 by men.
- Men have a higher default rate at approximately 10%, compared to women's 7%.

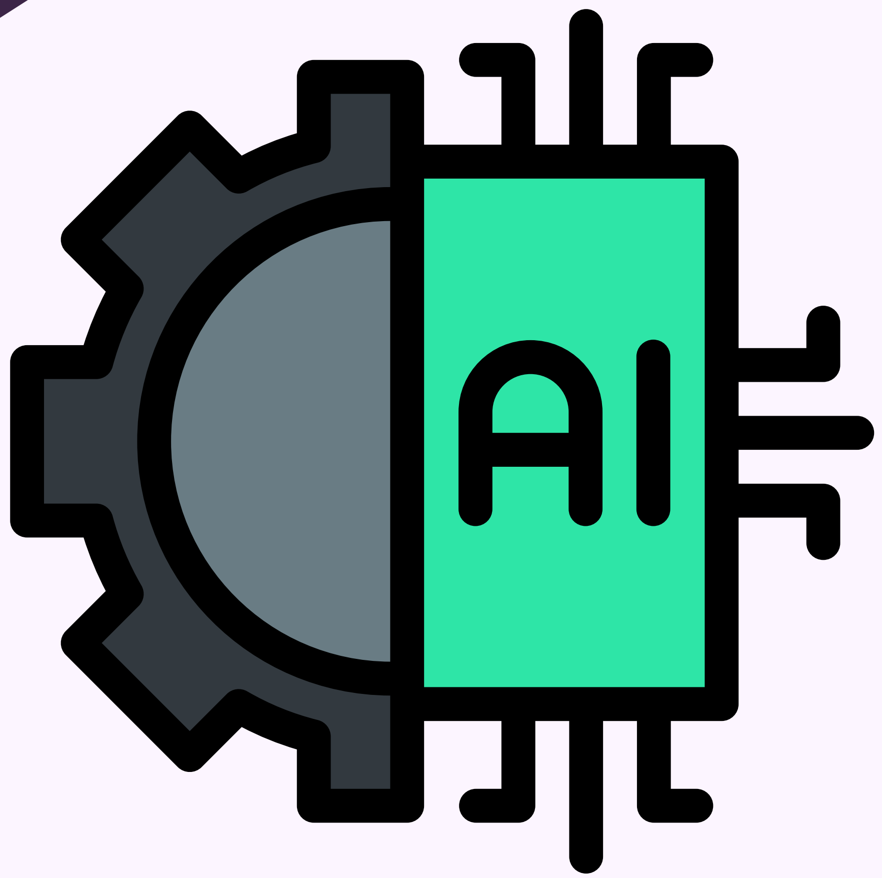


- Clients who lives in rented apartment and office apartment and their region have a rating of 1, have a problem repaying the loans compared to client in region with rating of 2 for a medium credit amount of the loan.



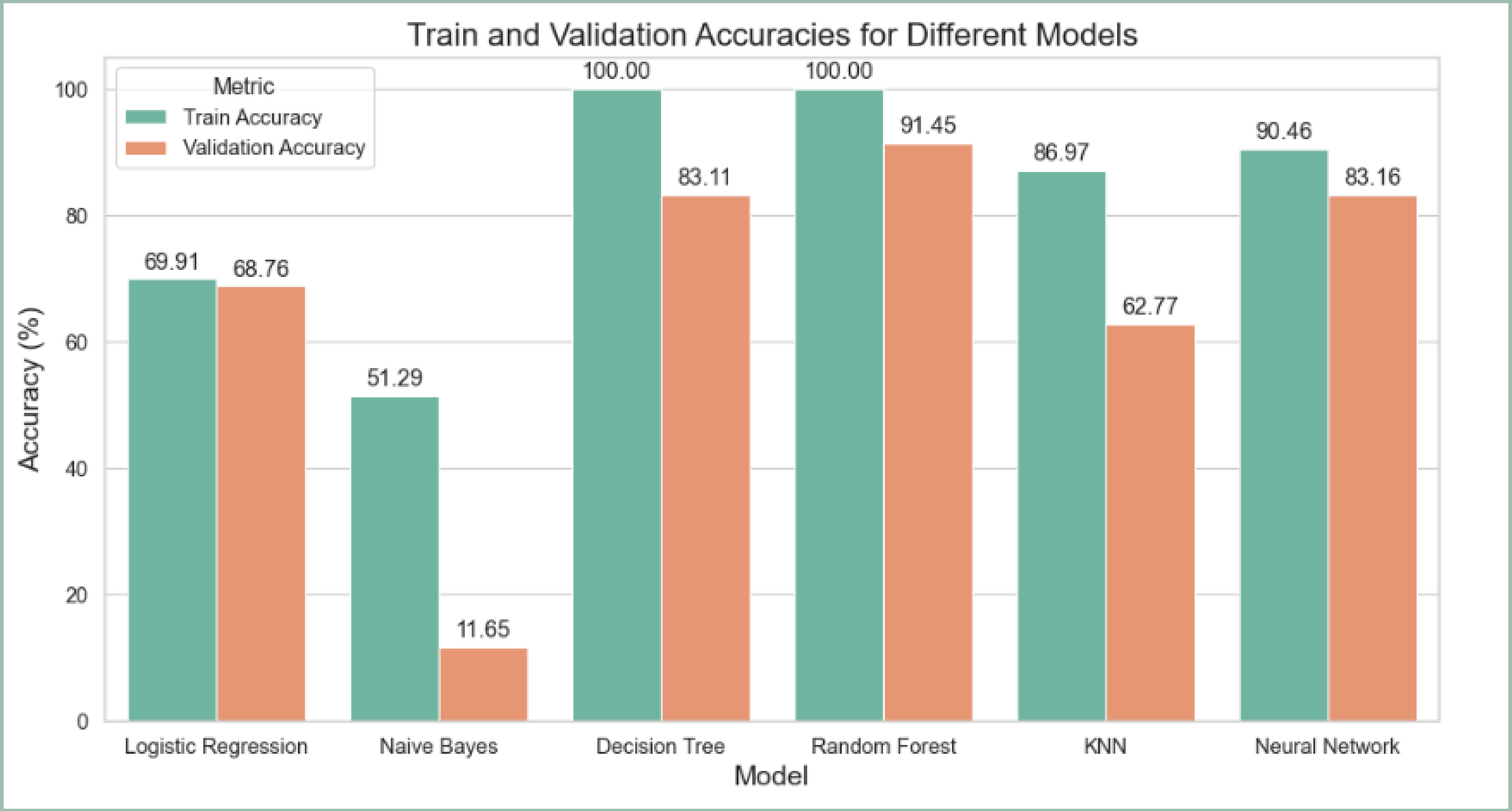
- There is a much higher number of Non-Default cases (Class 0) compared to default cases (Class 1): 183,651 vs 16,231.
- Thus, we see a large class imbalance. We will need to address this before training our model using oversampling the minority class using methods like SMOTE.

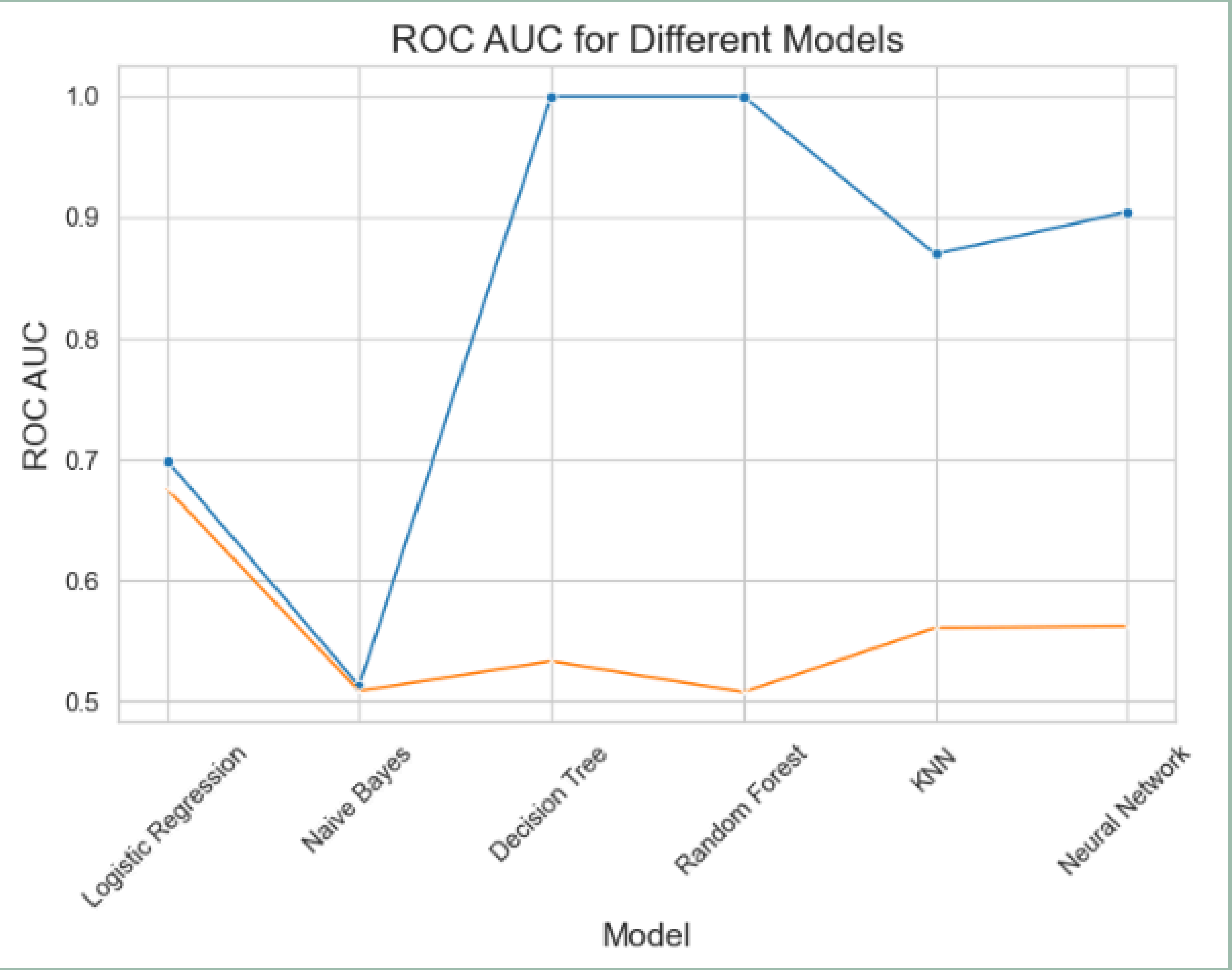
4 Machine Learning Model



Model Comparison

Algorithm	Training Accuracy Score	Validation Accuracy Score	Training ROC AUC	Validation ROC AUC
Logistic Regression	69.91%	68.76%	0.69	0.67
Gaussian Naive Bayes	51.29%	11.65%	0.5129	0.50
Decision Tree	100%	83.11%	1.0	0.53
Random Forest	100%	91.45%	1.0	0.50
K-Nearest Neighbor	86.97%	62.77%	0.86	0.56
Neural Network	90.46%	83.16%	0.90	0.56

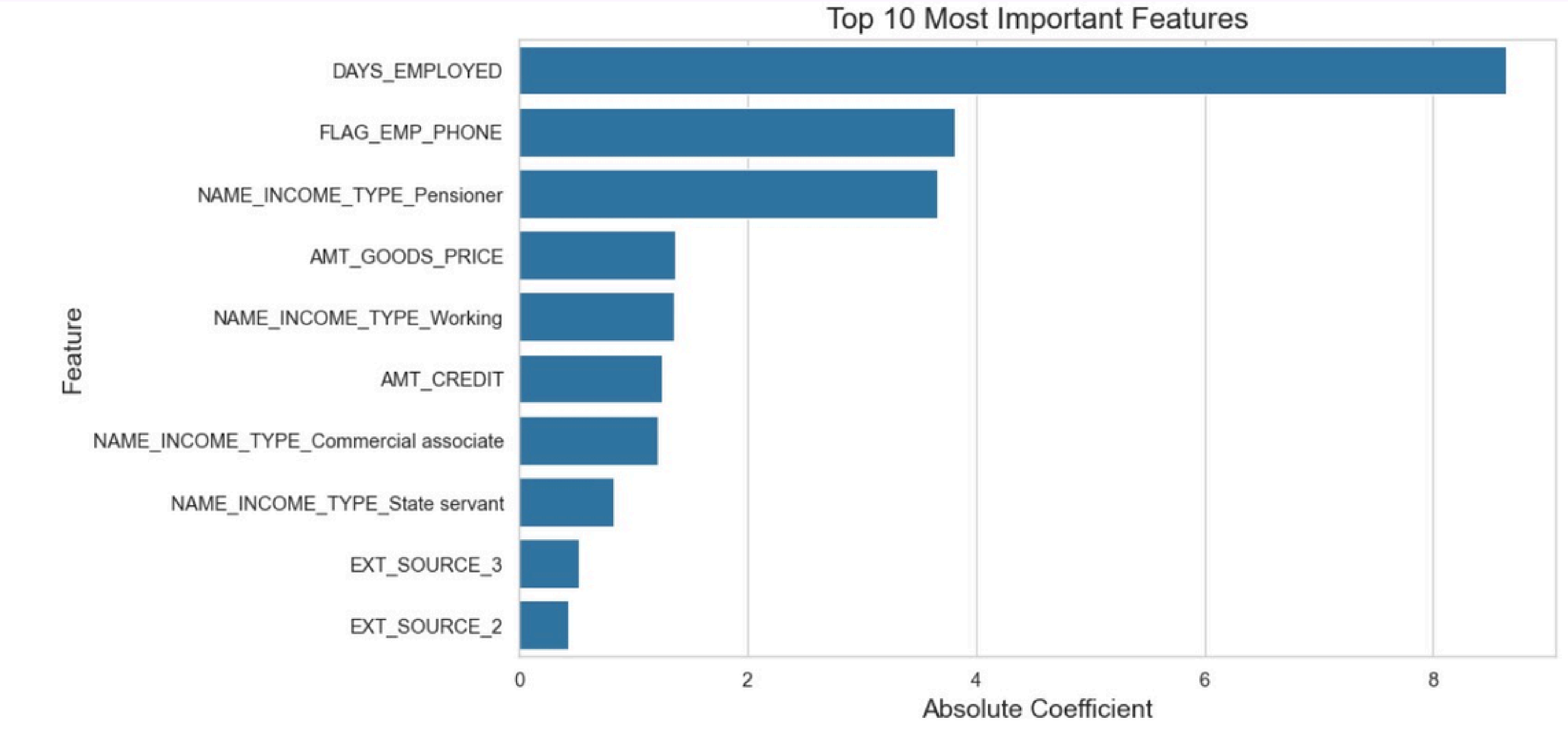

Train and Validation Accuracies for Different Models
Metric
Train Accuracy
Validation Accuracy
Accuracy (%)
0
20
40
60
80
100
69.91
68.76
51.29
11.65
100.00
83.11
100.00
91.45
86.97
62.77
90.46
83.16
Logistic Regression
Naive Bayes
Decision Tree
Random Forest
KNN
Neural Network
Model



ROC AUC for Different Models
ROC AUC
1.0
0.9
0.8
0.7
0.6
0.5
Logistic Regression
Naive Bayes
Decision Tree
Random Forest
KNN
Neural Network
Model

Logistic Regression had moderately good accuracies on training and validation sets. It had the least difference between validation and training accuracy compared to all other models. This means that logistic regression is the closest to being a good fit on the data compared to all other models. This is solidified by the fact that logistic regression has the best ROC AUC score on the validation set outperforming all other models.

Best Model



Algorithm

Logistic Regression

Performance

- Minimal train-validation accuracy gap
- Highest validation ROC AUC

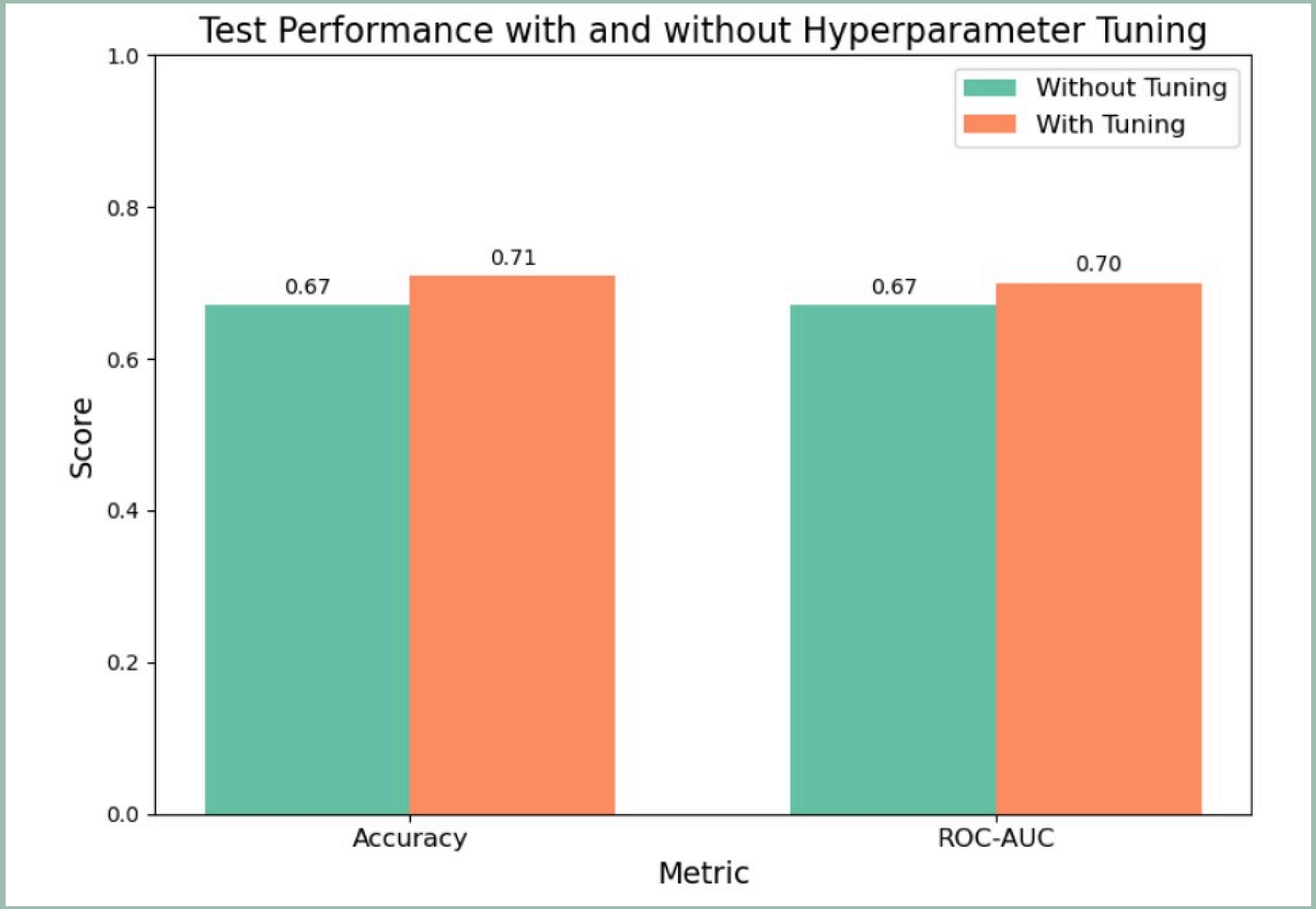
The 5 most important features

- DAYS_EMPLOYED
- FLAG_EMP_PHONE
- NAME_INCOME_TYPE_pensioner
- AMT_GOODS_PRICE
- NAME_INCOME_TYPE_WORKING

Hyperparameter Tuning

```
Parameter Grid:
C: [0.01, 0.1, 1, 10]
penalty: ['l1', 'l2']
solver: ['liblinear', 'saga']

Best Parameters:
C: 10
penalty: l1
solver: liblinear
```



Conclusion

- In conclusion, our data analysis provides insights for better lending practices:

1. Identified key demographic and socio-economic factors linked to repayment issues.
2. Examined loan characteristics like credit amount and type for their repayment influence.
3. Analyzed employment status and occupation for their effects on payment behavior.
4. Assessed living arrangements to understand their impact on repayment capacity.
5. Highlighted payment difficulty variances between different loan types and borrower profiles.
6. These findings will refine lending criteria and enhance risk management, improving client borrowing experiences.
7. Highlighted important features that potentially can influence loan approval process.
8. Built a predictive model to classify a loan application case as default/non default with good accuracy.

Thank You!

Shivie Saksenaa

Kushagra Trivedi

Anirudha Kuchibhotla