04/23/2024

HOME CREDIT SCORE CARD MODEL

created by Group 14

Table of Contents

1. Problem Research

Project Background

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)

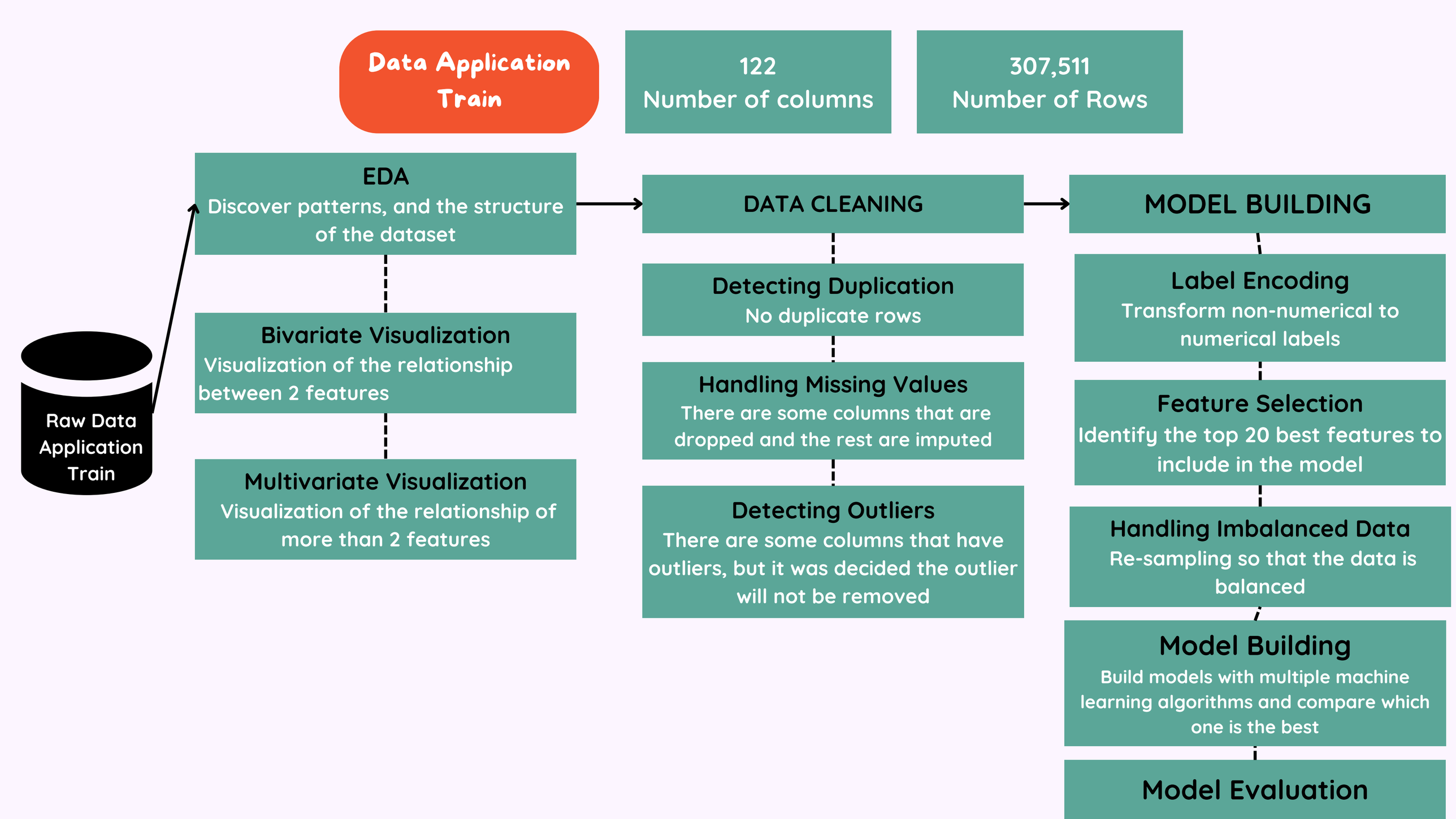
Objective

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


Data Application Train
122
Number of columns
307,511
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
Detecting Outliers
There are some columns that have outliers, but it was decided the outlier will not be removed
MODEL BUILDING
Label Encoding
Transform non-numerical to numerical labels
Feature Selection
Identify the top 20 best features to include in the model
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

Data Application Test

121
Number of columns

48,744
Number of Rows

DATA CLEANING

Detecting Duplication
No duplicate rows

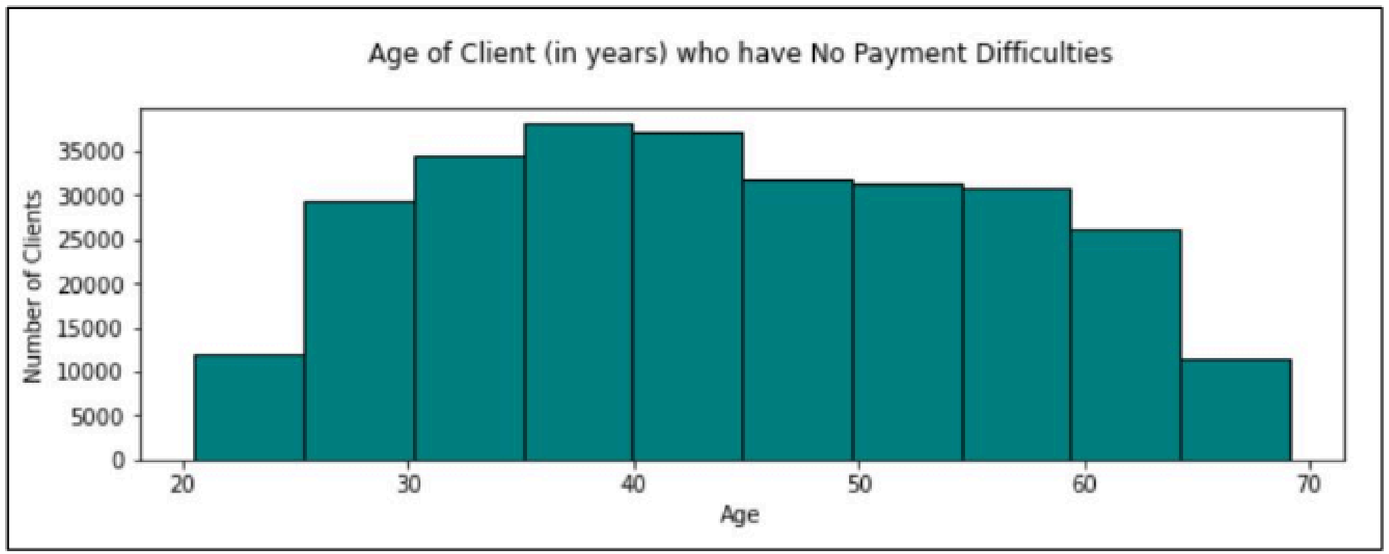
Handling Missing Values
There are some columns that are dropped and the rest are imputed

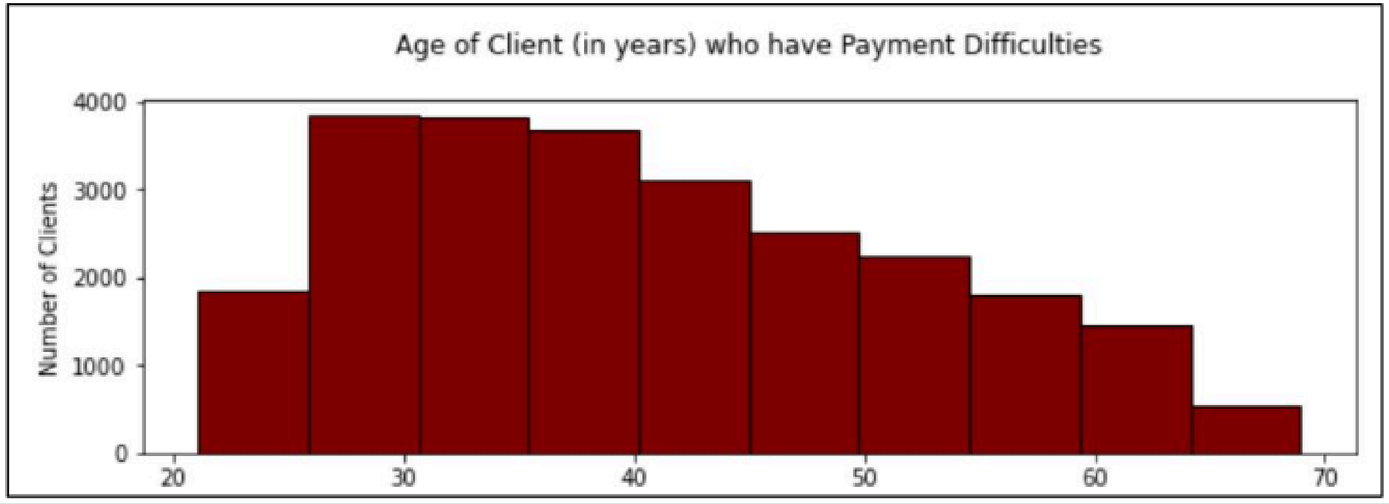
Detecting Outliers
There are some columns that have outliers, but it was decided the outlier will not be removed

PREDICTION

Predict client's repayment abilities with best machine learning model obtained before

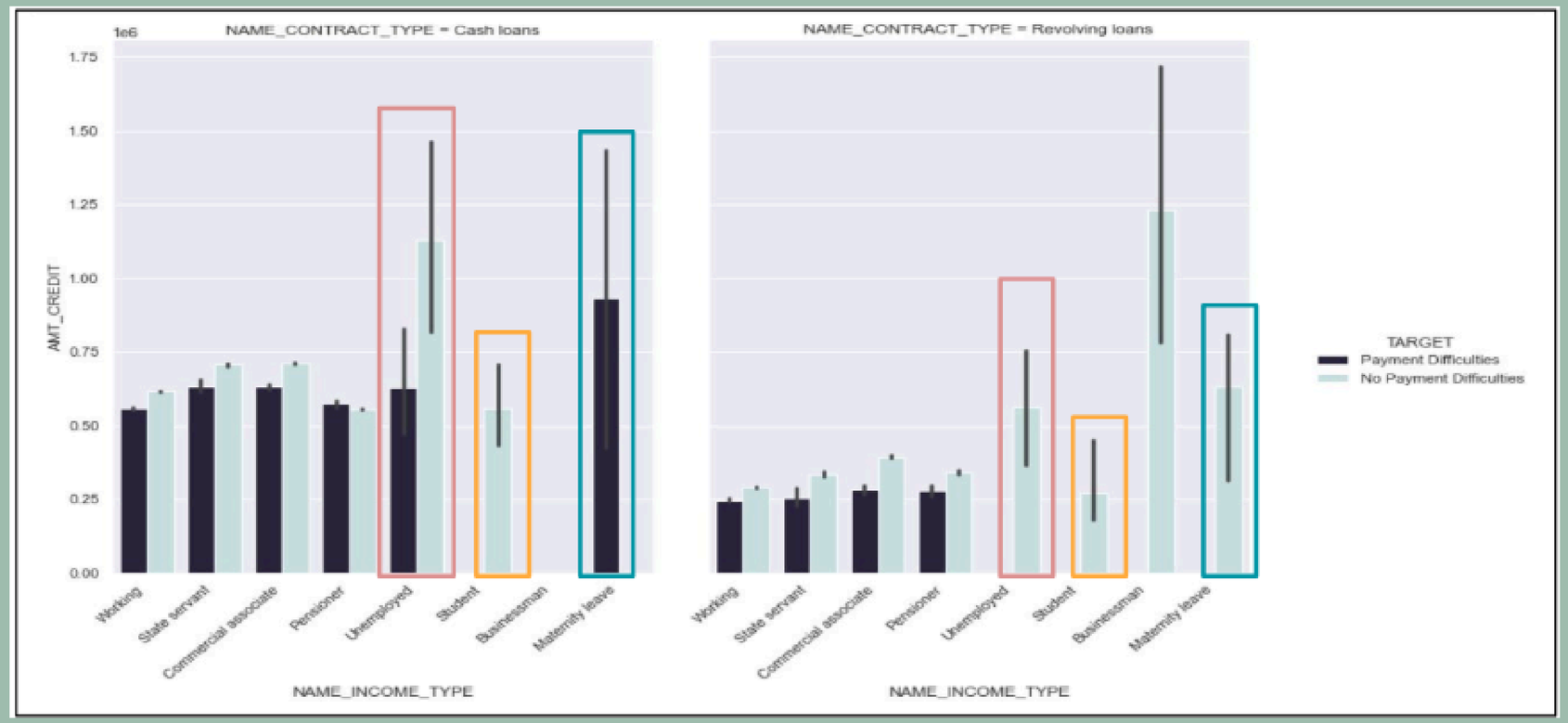
3 Data Insights





- 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.

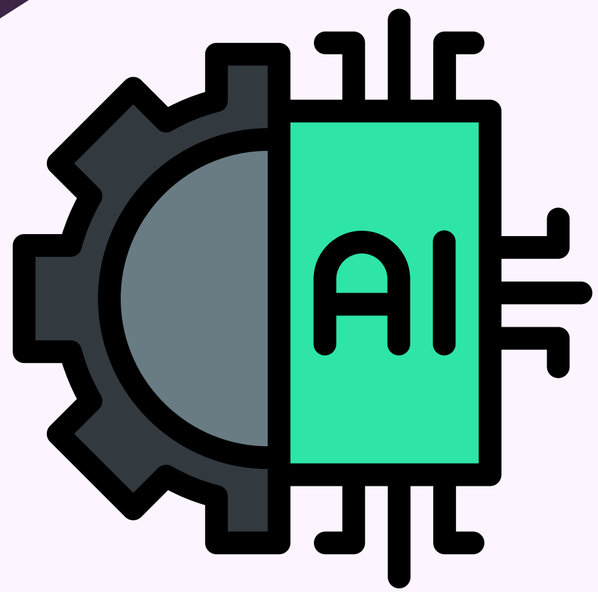


All student clients have no difficulty repaying the loans whether with cash loan or revolving loan for a low to medium credit amount of the loan.

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.

4 Machine Learning Model

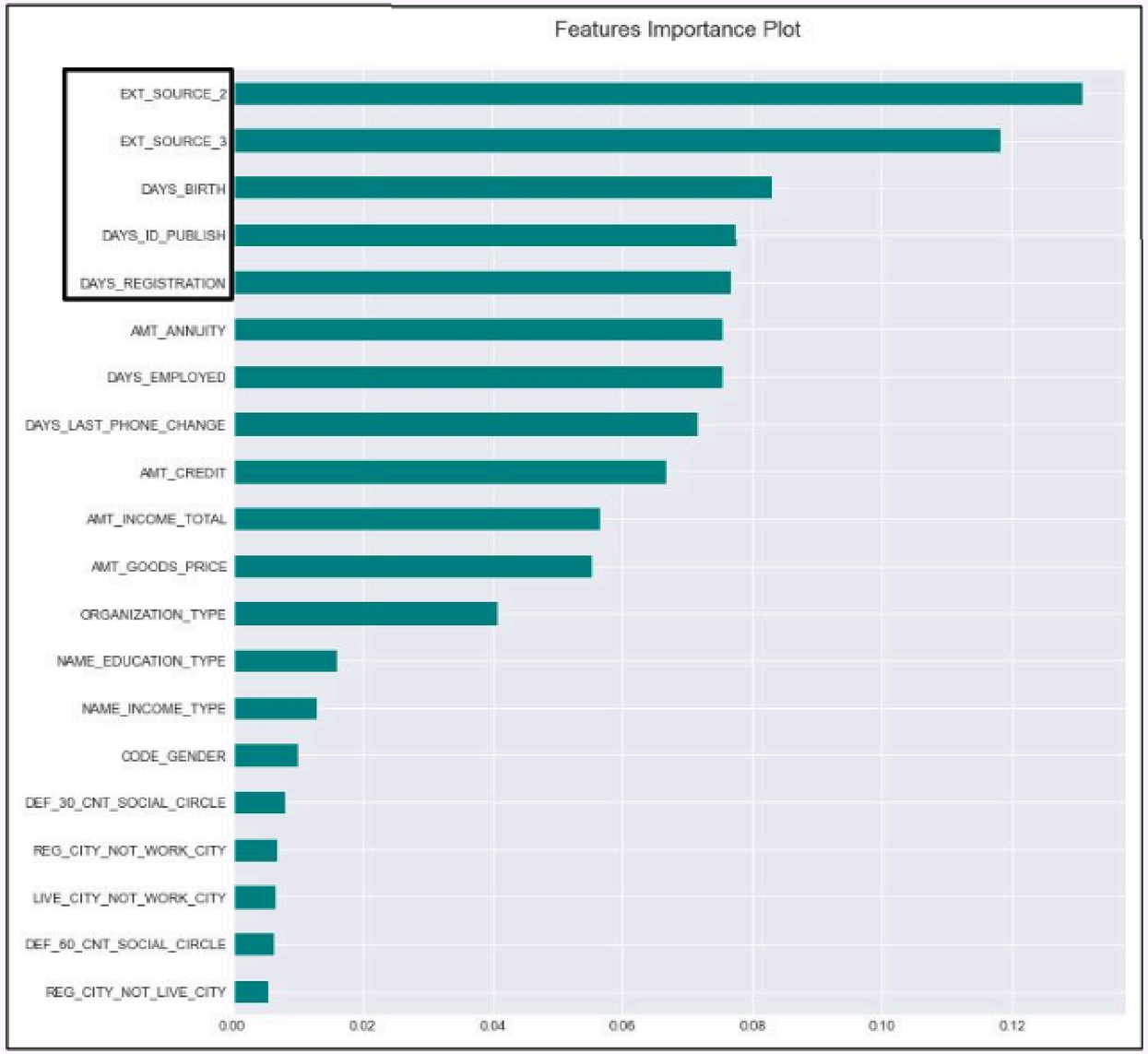


Model Comparison

Algorithm	Training Accuracy Score	Testing Accuracy Score	Error Margin	ROC Score
Logistic Regression	67.16%	67.29%	0.13%	0.6728
Gaussian Naive Bayes	60.24%	60.39%	0.15%	0.604
Decision Tree	100%	83.9%	11.74%	0.8826
Random Forest	100%	99.65%	0.35%	0.9965
K-Nearest Neighbor	91.56%	88.07%	3.79%	0.8806
Neural Network	70.01%	69.48%	0.58%	0.6948

The prediction accuracy of the train and test data in Random Forest model has a value that is not much different, it can be said that the model is very good, which is there is no underfitting or overfitting. So the Random Forest model was chosen as the best model to predict client’s repayment abilities.

Best Model



- **Algorithm** — Random Forest
- **Performance**
  - Random forest model gives 100% correct results
  - There is 0.35% error margin
- **The 5 most important features**
  - Score from external data source 2
  - Score from external data source 3
  - Client's age in days
  - Days ID publish
  - Days registration

Conclusion

Thank You!