HOME CREDIT

HOME CREDIT

SCORECARD MODEL

BY: NOOR KHARISMAWAN AKBAR

PROJECT BACKGROUND

PROBLEM STATEMENT

Home Credit is currently using various statistical methods and Machine Learning to make credit score predictions in order to ensure customers who are able to make repayments are not rejected when applying for a loan.

GOAL & OBJECTIVE

Minimize the number of clients who are approved but actually defaulters and create predictive model to determine potential client and default client.

DATASET

- application_train.csv (with TARGET)
- application_test.csv (without TARGET)

MODEL EVALUATION

Model evaluated using area under ROC curve.

WORKING FLOW

APPLICATION_TRAIN.CSV

1 EDA

-Univariate visualization
-Bivariate visualization
-Multivariate visualization

2 DATA CLEANING

-Detecting duplication
-Handling missing values
-Detecting outliers

3 MODEL BUILDING

-Label encoding
-Feature selection
-Handling imbalanced data
-Model building
-Model evaluation

APPLICATION_TEST.CSV

1 DATA CLEANING

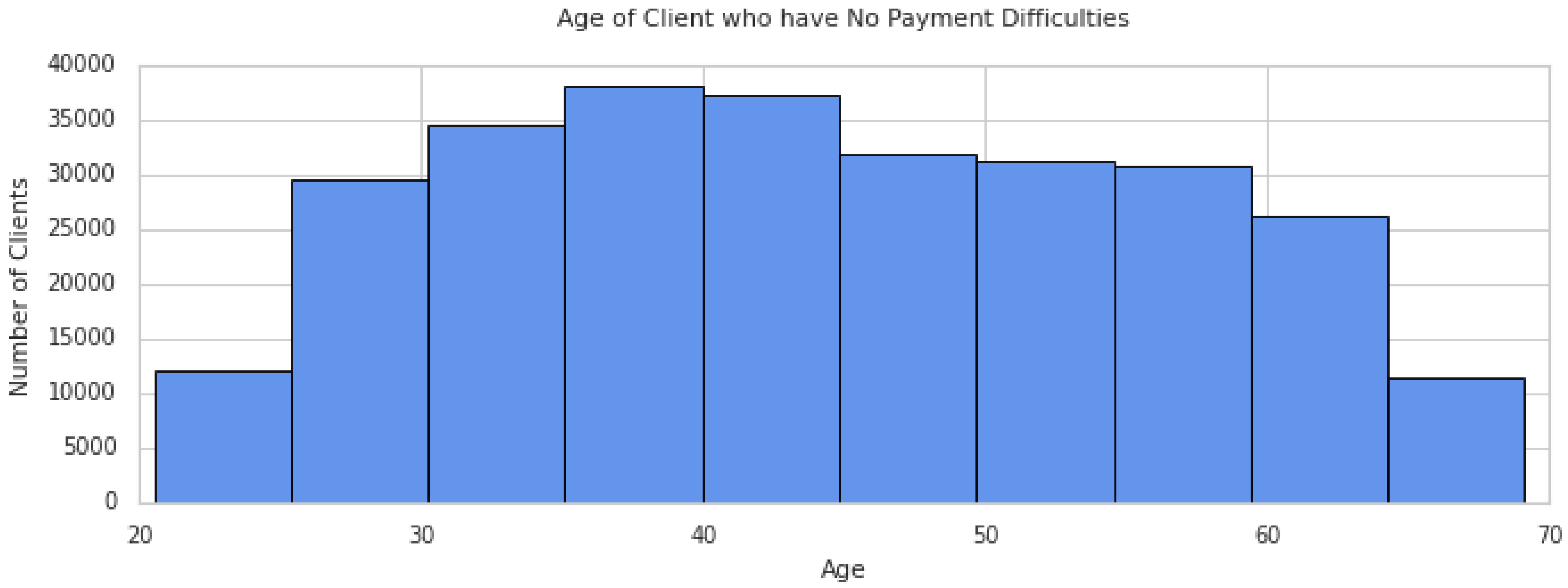
-Detecting duplication
-Handling missing values
-Detecting outliers

2 PREDICTION

Output is TARGET that classified by
-0 (Client with no payment difficulties)
-1 (Client with payment difficulties)

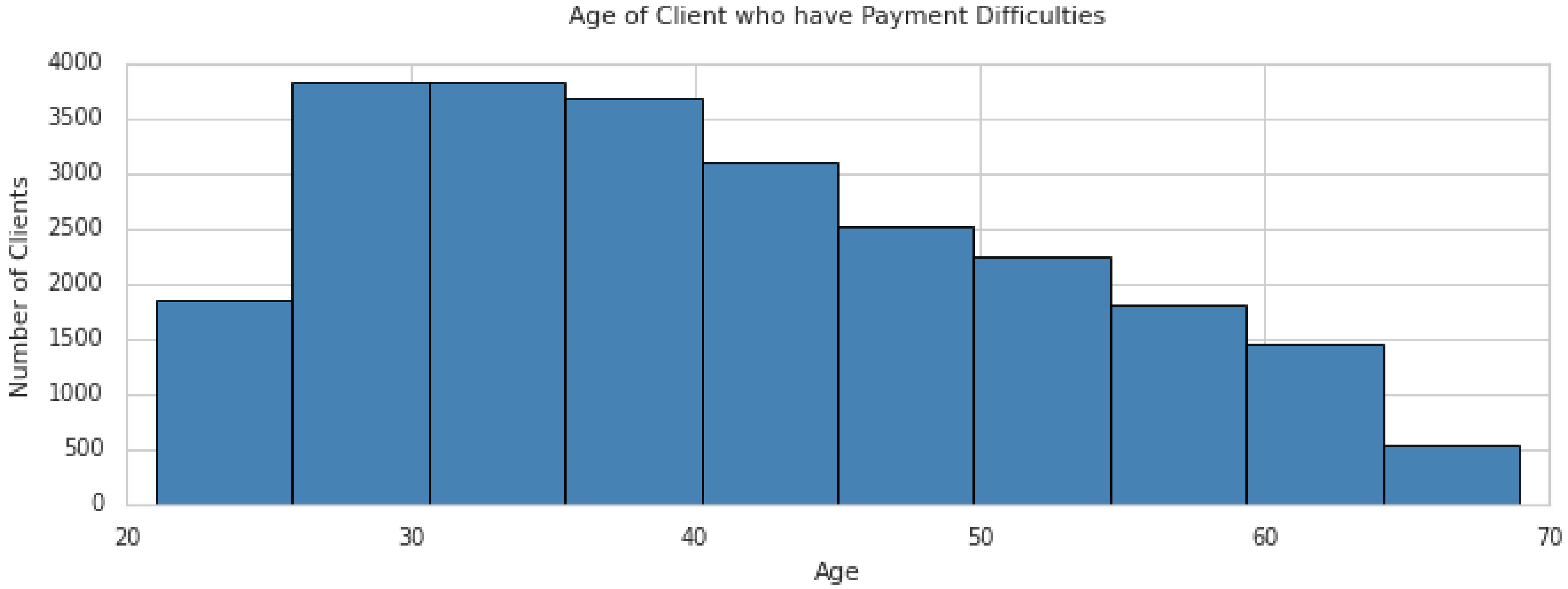
BUSINESS INSIGHT

Part 1



✓

Clients who **have no payment difficulties** are client the range of **35-45 years old.**

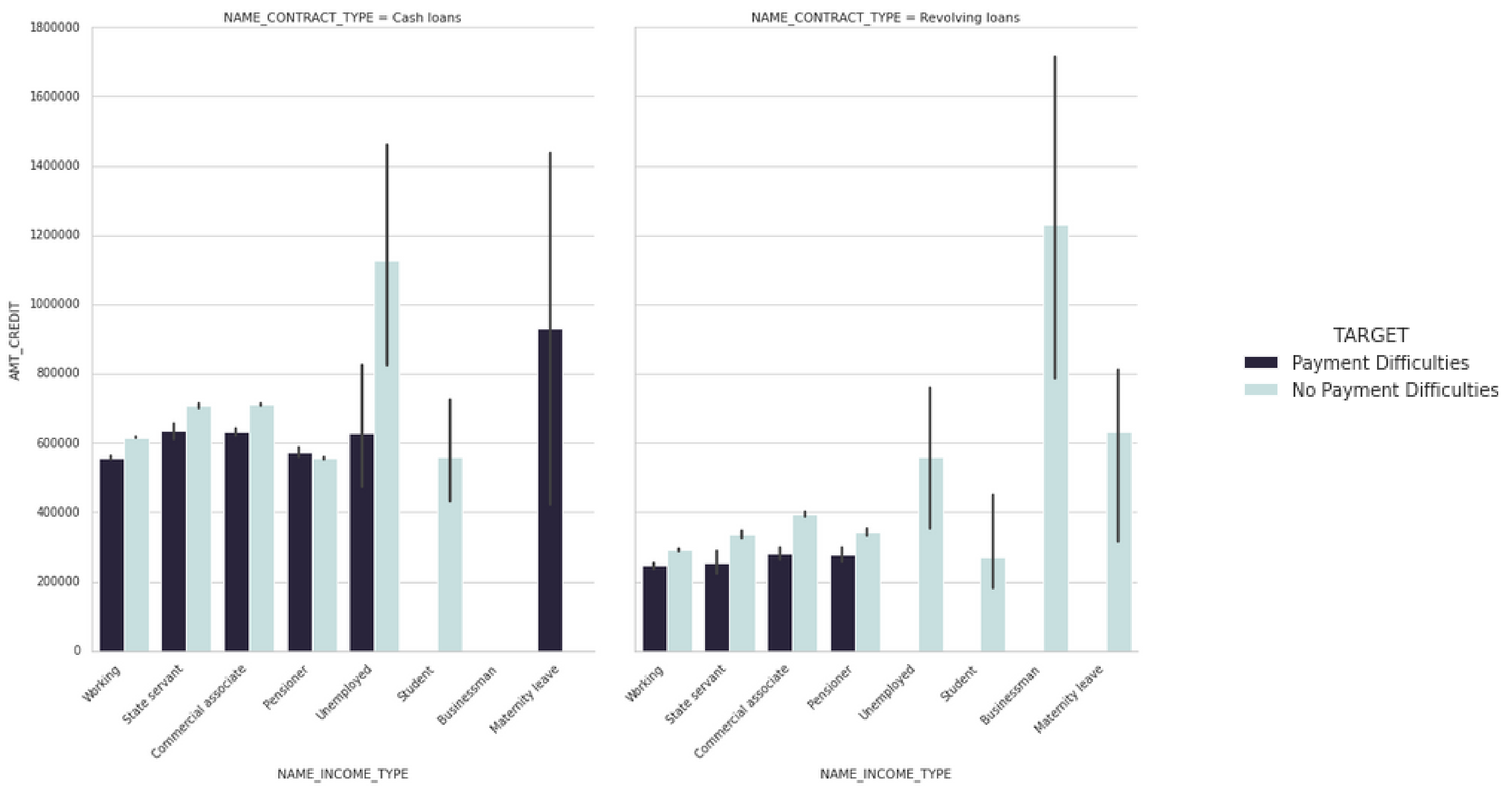


⚠

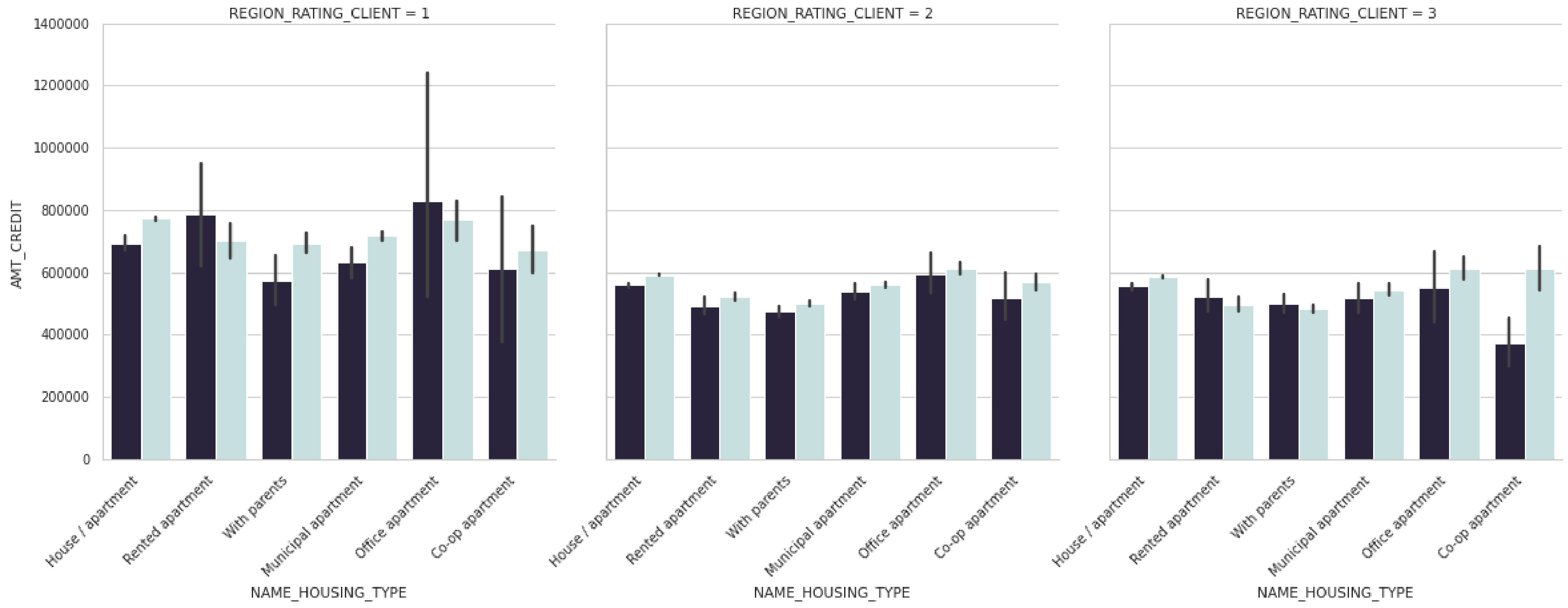
While clients who **have payment difficulties** are client the range of **25-35 years old**.

BUSINESS INSIGHT

Part 2



- >50% of clients who are **unemployed** have difficulty repaying cash loans, but have no difficulty paying back revolving loans.
- All **student** clients have no difficulty repaying loans either with cash loans or revolving loans for low to medium-credit loan amounts.
- All clients with an income type of **maternity leave** had difficulty repaying cash loans, but had absolutely no difficulty paying revolving loans.



- In **region 1**, Clients who lived in **rented** & **office apartment** have difficulties on repaying loans.
- In **region 2**, have **no difficulties** on repaying loans in any housing type.
- In **region 3**, clients who lived in **rented apartment** and **with parents** have difficulties on repaying loans.

MACHINE LEARNING MODELLING

Models	Training Accuracy Score	Testing Accuracy Score	Error	ROC Score	
Random Forest	**100.00%**	**99.64%**	**0.36%**	**0.9964**	✓
Decision Tree	100.00%	88.36%	-11.64%	0.8836	
K-Nearest Neighbor	91.56%	88.07%	-3.49%	0.8806	
Neural Network	69.59%	69.05%	0.54%	0.6906	
Logistic Regression	67.16%	67.29%	0.13%	0.6729	
Gaussian Naive Bayes	60.24%	60.39%	0.15%	0.604	

From the results of the model evaluation, it was found that the **highest accuracy and lowest error** were found in the **Random Forest classifier model**. Based on the ROC score, it was also found that Random Forest had a much higher score than the others. Which means that the **model has the minimum under-fitting and over-fitting**.

FEATURES IMPORTANCE PLOT

1 MACHINE LEARNING MODEL

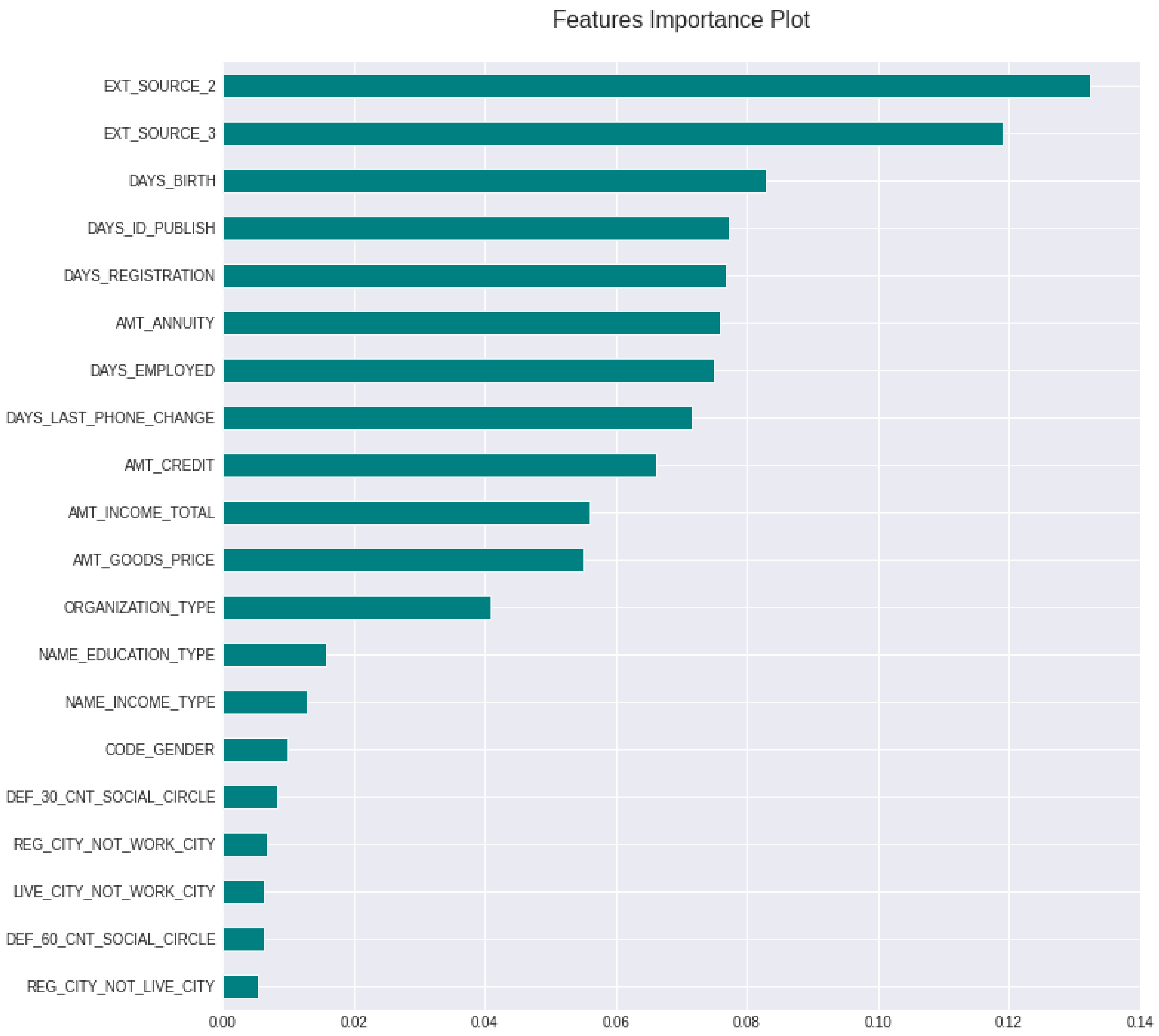
RANDOM FOREST CLASSIFIER

2 PERFORMANCE ACCURACY

- Train data: 100%
- Test data: 99.64%
- Error margin: 0.36%

3 TOP 5 MOST IMPORTANT FEATURES

1. **EXT_SOURCE_2**: Normalized score from external data source 2
2. **EXT_SOURCE_3**: Normalized score from external data source 3
3. **DAYS_BIRTH**: Client's age in days at the time of application
4. **DAYS_ID_PUBLISH**: Days before the application did client change the identity document with which he applied for the loan
5. **DAYS_REGISTRATIO**N: Days before the application did client change his registration



MODEL PREDICTION

USING RANDOM FOREST CLASSIFIER

Based on the important columns, prediction model build to determine the target

SK_ID_CURR	TARGET
100001	0
100005	0
100013	0
100028	0
100038	0

The results of the 5 samples above is **all clients have no difficulties** on repaying loans.

RECOMMENDATION

CREATE A CAMPAIGN

For **clients aged 35-45 years who work as students, accountants, high-skill tech staff, managers in region 2**. According to the analysis, they have no difficulty repaying loans.

FURTHER CONSIDERATION NEEDED

For clients with **unemployed and maternity leave** type income. Where they are more likely to be able to repay revolving loans.

DEEPER RESEARCH NEEDED

Focussing on top 5 in features important plot (EXT_SOURCE_2, EXT_SOURCE_3, DAYS_BIRTH, DAYS_ID_PUBLISH, DAYS_REGISTRATION)