BAD CUSTOMER DETECTION

Group 5:
Roman Burekhin
Thien An Trinh
Shoaib ahmed khan
Joy Vahini Varatharaajah
Athira Devan
Pratheep kumar venkatrangam
Ahmad AlHammad
Esther Yu

CONTENTS

1. Business Value
2. Exploratory Data Analysis
3. Data preprocessing
4. Default models
5. Tuning model
6. Feature selection
7. Retune model with new refined data
8. Threshold Analysis

1. BUSINESS VALUE

Purpose: predicting default on a loan among individuals (bad client detection)

Model relevance:

- credit risk reduction
- reduction of banks' reserves, and, accordingly, profit growth, due to an advanced approach to assessing credit risks.

Who is interested in this model:

- Any financial institutions
- Marketplaces

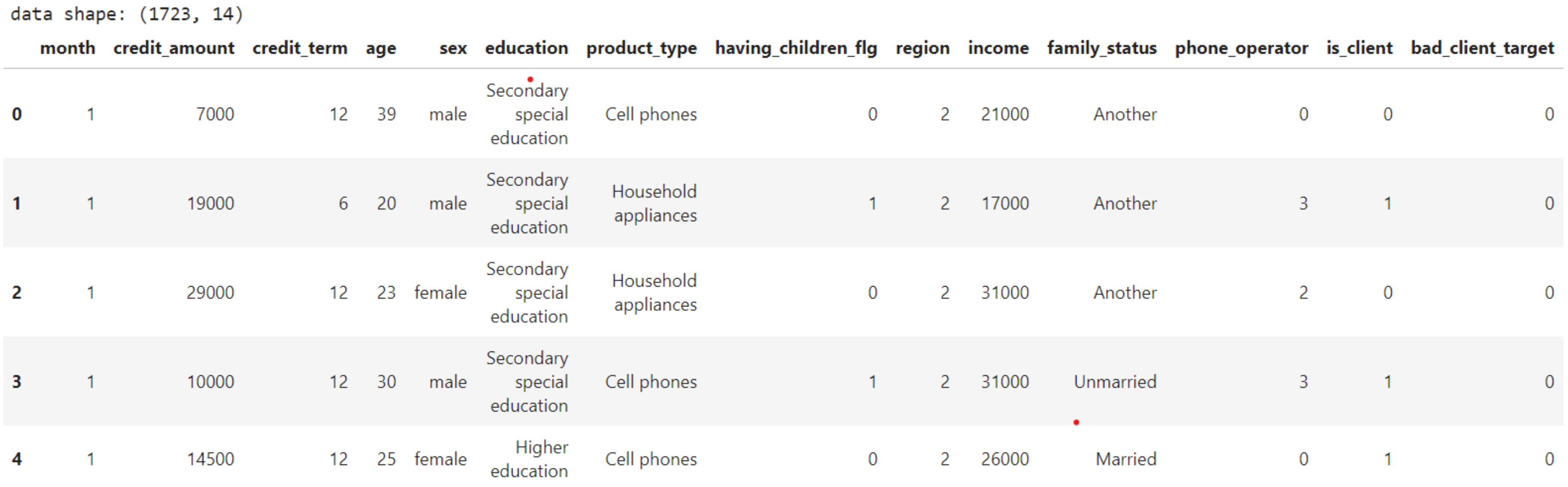
Example:

- Raiffeisen bank case: decrease in reserves by 27 billion rub. (550 million CAD) in Russia in 2020 *

Raiffeisen BANK

* Source: https://www.raiffeisen.ru/about/investors/annualreport/

DATA DESCRIPTION



```
data shape: (1723, 14)
```

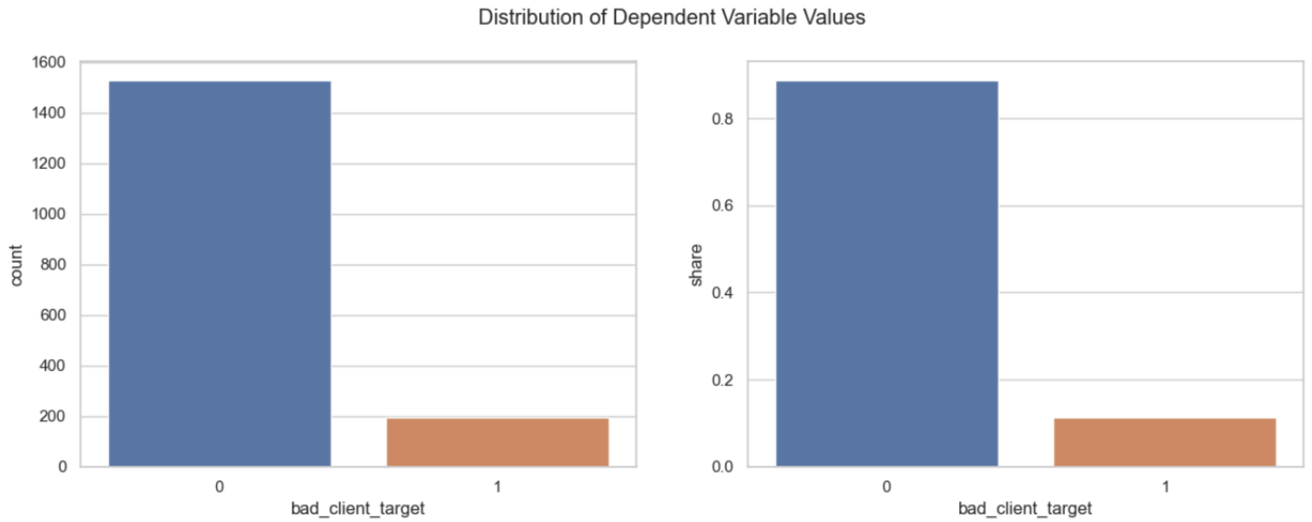
	month	credit_amount	credit_term	age	sex	education	product_type	having_children_flg	region	income	family_status	phone_operator	is_client	bad_client_target
0	1	7000	12	39	male	Secondary special education	Cell phones	0	2	21000	Another	0	0	0
1	1	19000	6	20	male	Secondary special education	Household appliances	1	2	17000	Another	3	1	0
2	1	29000	12	23	female	Secondary special education	Household appliances	0	2	31000	Another	2	0	0
3	1	10000	12	30	male	Secondary special education	Cell phones	1	2	31000	Unmarried	3	1	0
4	1	14500	12	25	female	Higher education	Cell phones	0	2	26000	Married	0	1	0

The dataset contains 1723 rows and 14 columns.

2. EXPLORATORY DATA ANALYSIS

Dependent Variable

- Bad_client_target

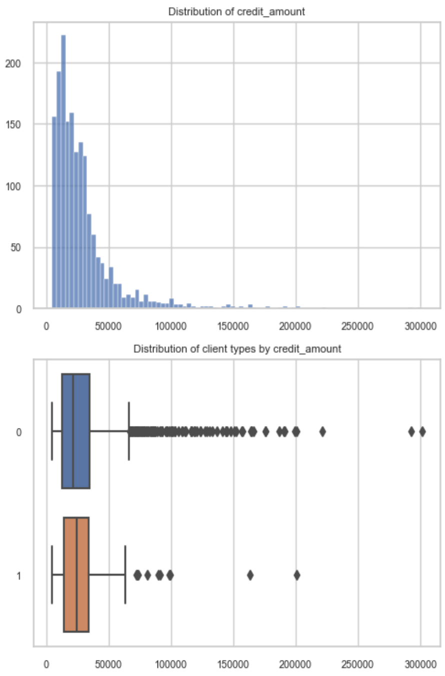


Independent Variable

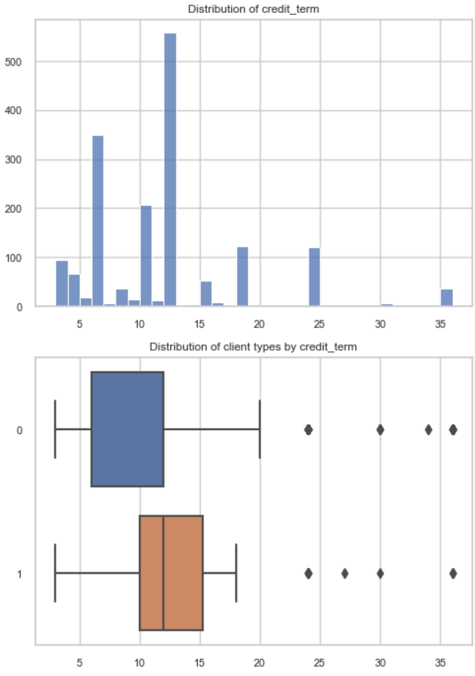
- **Numerical Variable**
  - Credit_amount
  - Credit_term
  - Age
  - Income

- **Categorical Variable**
  - Month
  - Sex
  - Education
  - Product_type
  - Having_children_flg
  - Region
  - Family_status
  - Phone_operator
  - Is_client

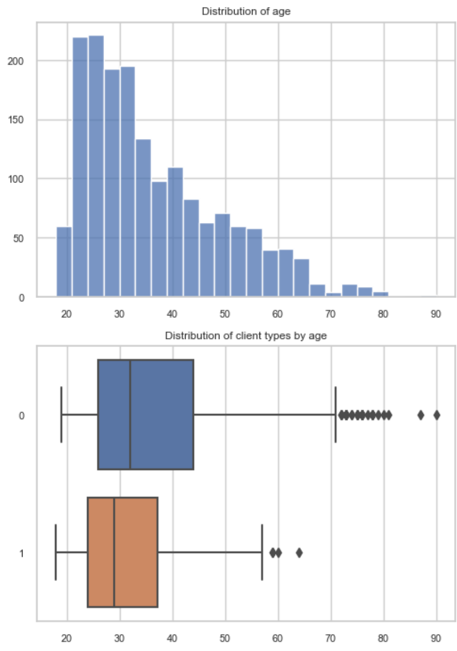
Credit_Amount



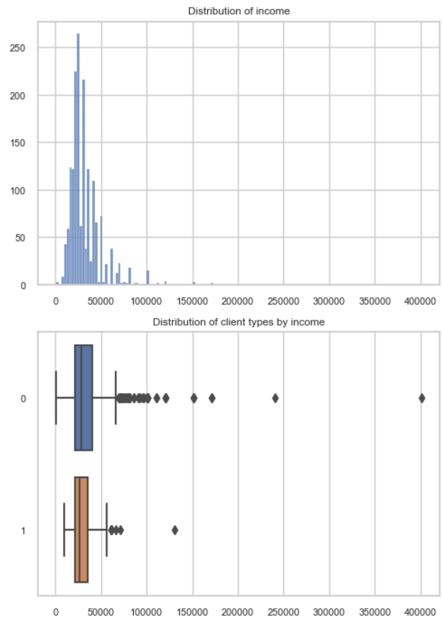
Credit_Term



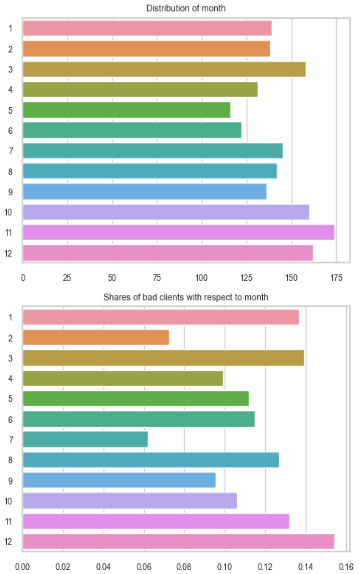
Age



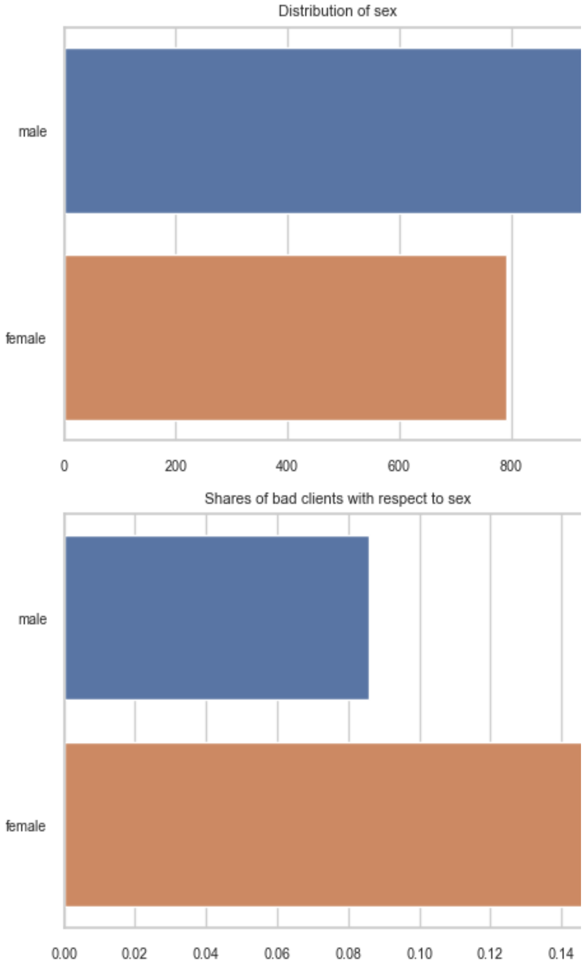
Income



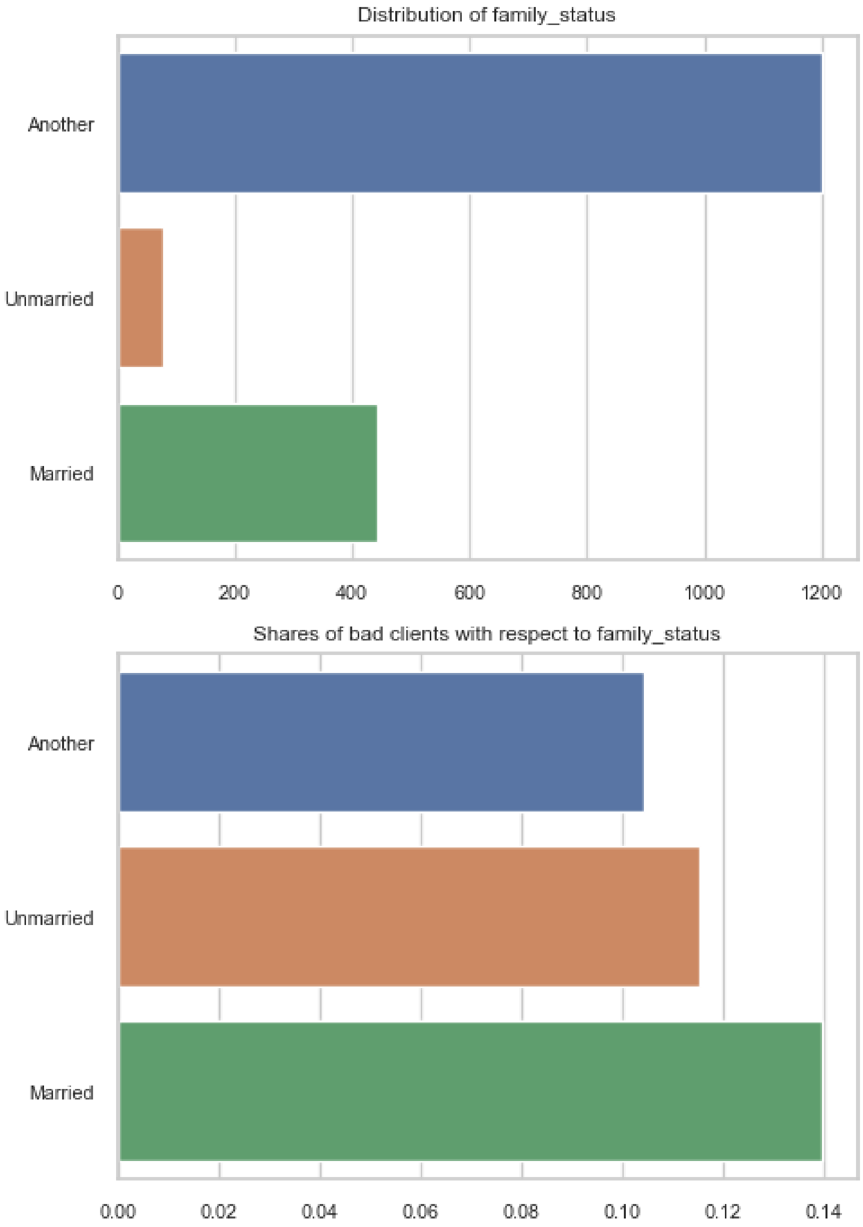
Month



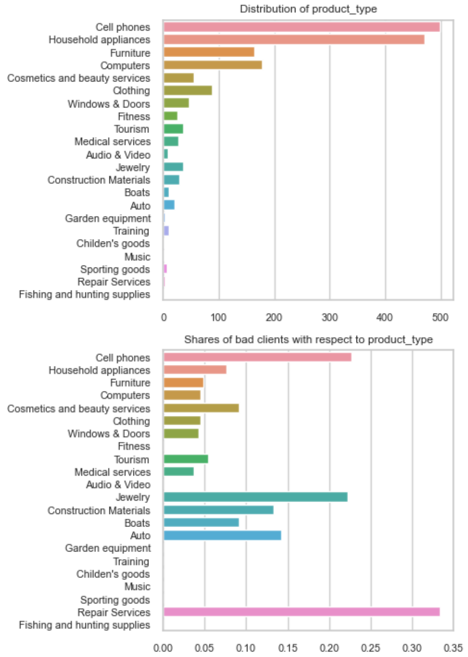
Sex



Family_status



Product Type



1. Most bad client are in December, minority are in July
2. There are outliers in credit amount.
3. Bad clients having credit time
4. Female are bad client more frequently
5. People with only secondary education are bad clients more frequently
6. It seems that loan for cell phone is the most risky
7. People with children are less risky clients
8. People from region 3 are riskier
9. There are outliers in income
10. Marriage status, phone operator don’t influence the probability of client default
11. Clients of bank are more risky

3. DATA PREPROCESSING

Used method: one hot encoding

Education
education_Higher education
education_Incomplete higher education
education_Incomplete secondary education
education_PhD degree
education_Secondary education
education_Secondary special education

Product_type	
product_type_Audio & Video	product_type_Repair Services
product_type_Auto	product_type_Sporting goods
product_type_Boats	product_type_Tourism
product_type_Cell phones	product_type_Training
product_type_Construction Materials	product_type_Windows & Doors
product_type_Fitness	product_type_Household appliances
product_type_Furniture	product_type_Cosmetics and beauty services
product_type_Garden equipment	product_type_Fishing and hunting supplies
product_type_Jewelry	product_type_Childen_good
product_type_Medical services	product_type_Clothing
product_type_Music	product_type_Computers

Family_status
family_status_Another
family_status_Married
family_status_Unmarried

sex
Sex_female
Sex_male

4. DEFAULT MODELS

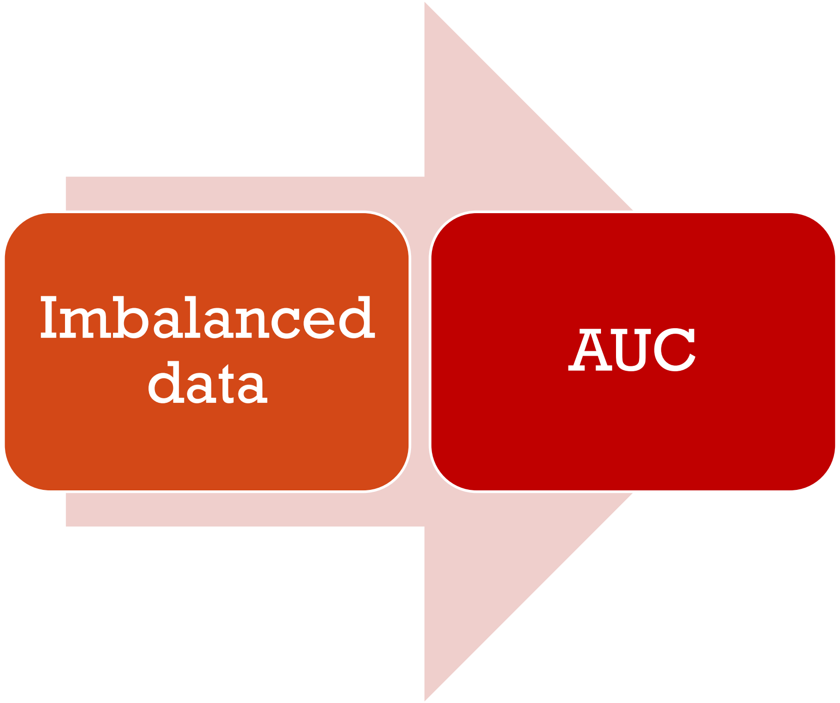
Train and test split

- Splitting ratio

- 517 instances : 1206 instances

Bad_client_target	Train
0	1073
1	133

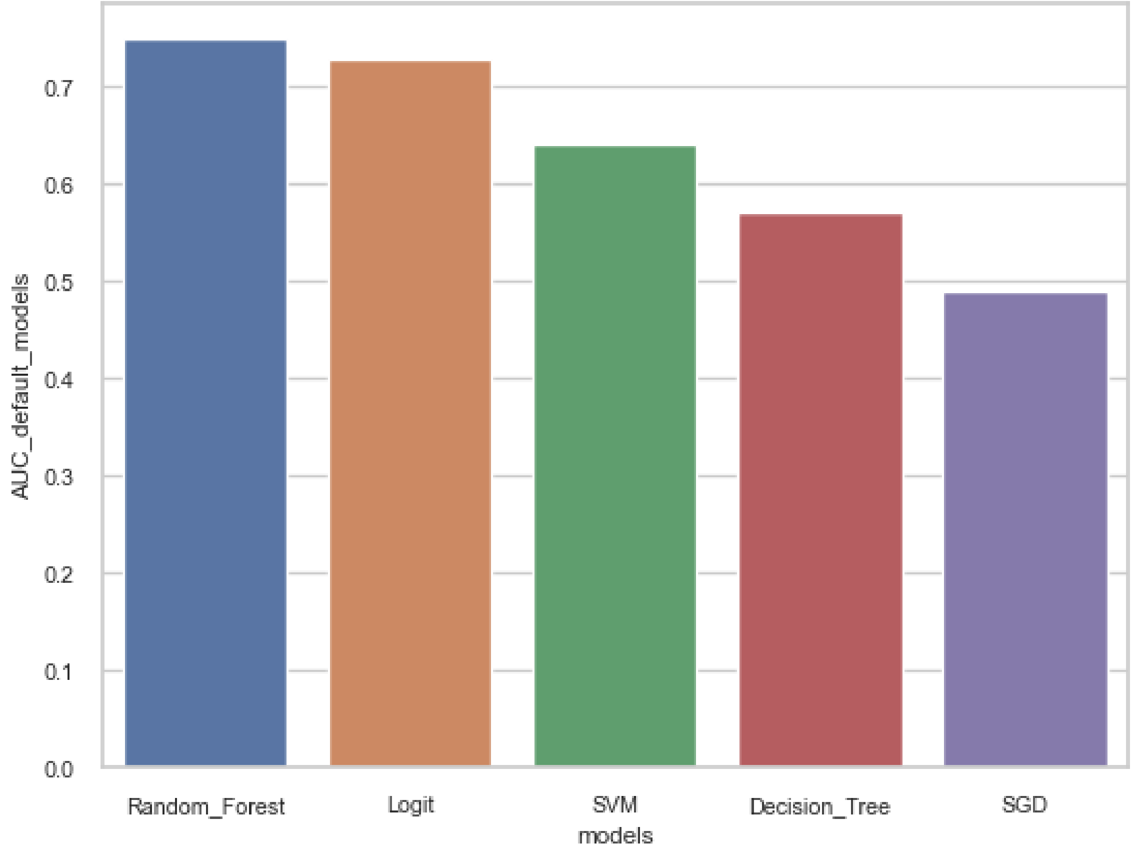
Bad_client_target	Test
0	454
1	63



Imbalanced data

AUC

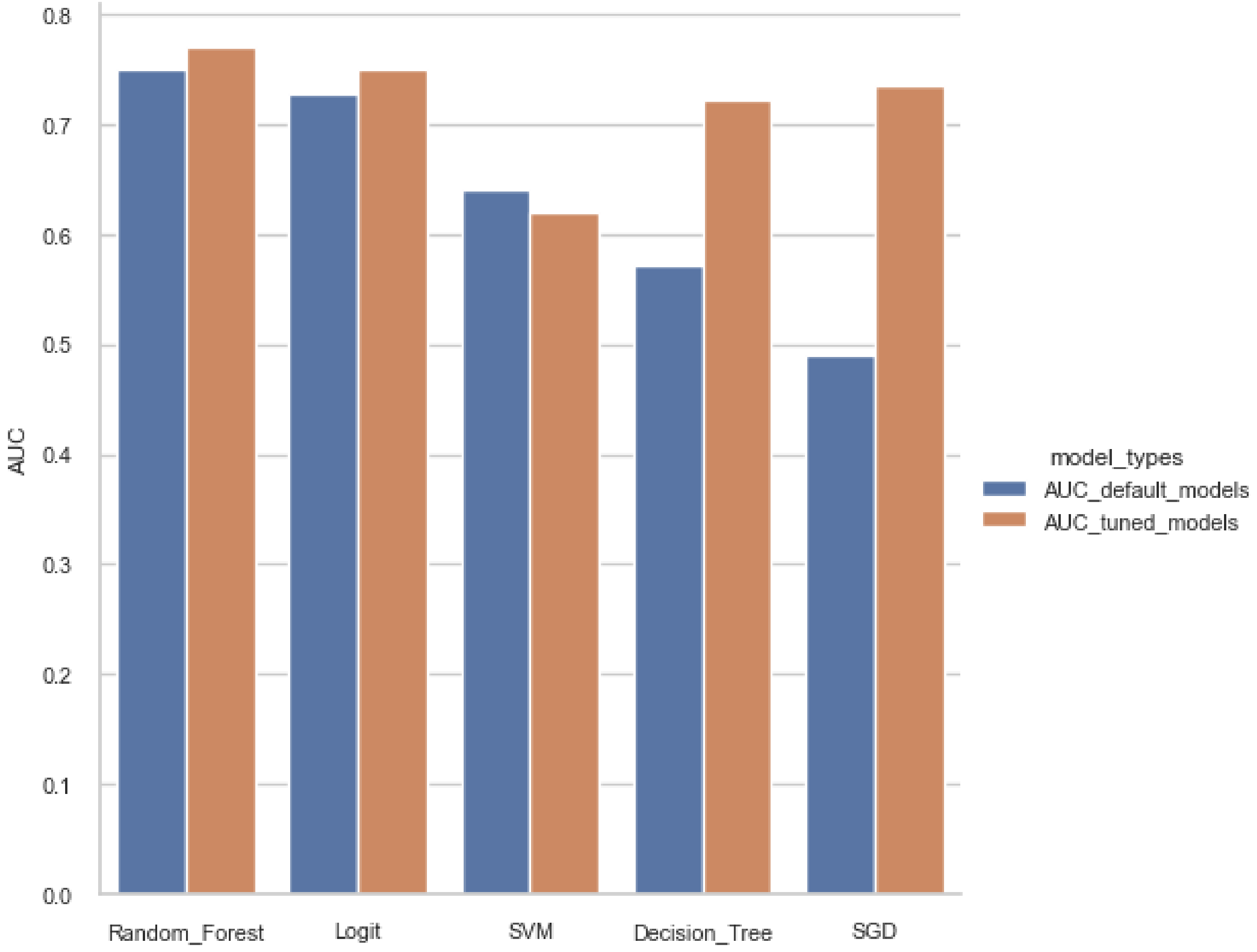
Models	AUC default models
Random Forest	0.74
Logit	0.72
SVM	0.64
Decision Tree	0.57
SGD	0.49



- The best default model is Random Forest

5. TUNING MODEL

Models	AUC default models	AUC tuned models
Random Forest	0.74	0.77
Logit	0.72	0.74
SVM	0.64	0.62
Decision Tree	0.57	0.72
SGD	0.49	0.73

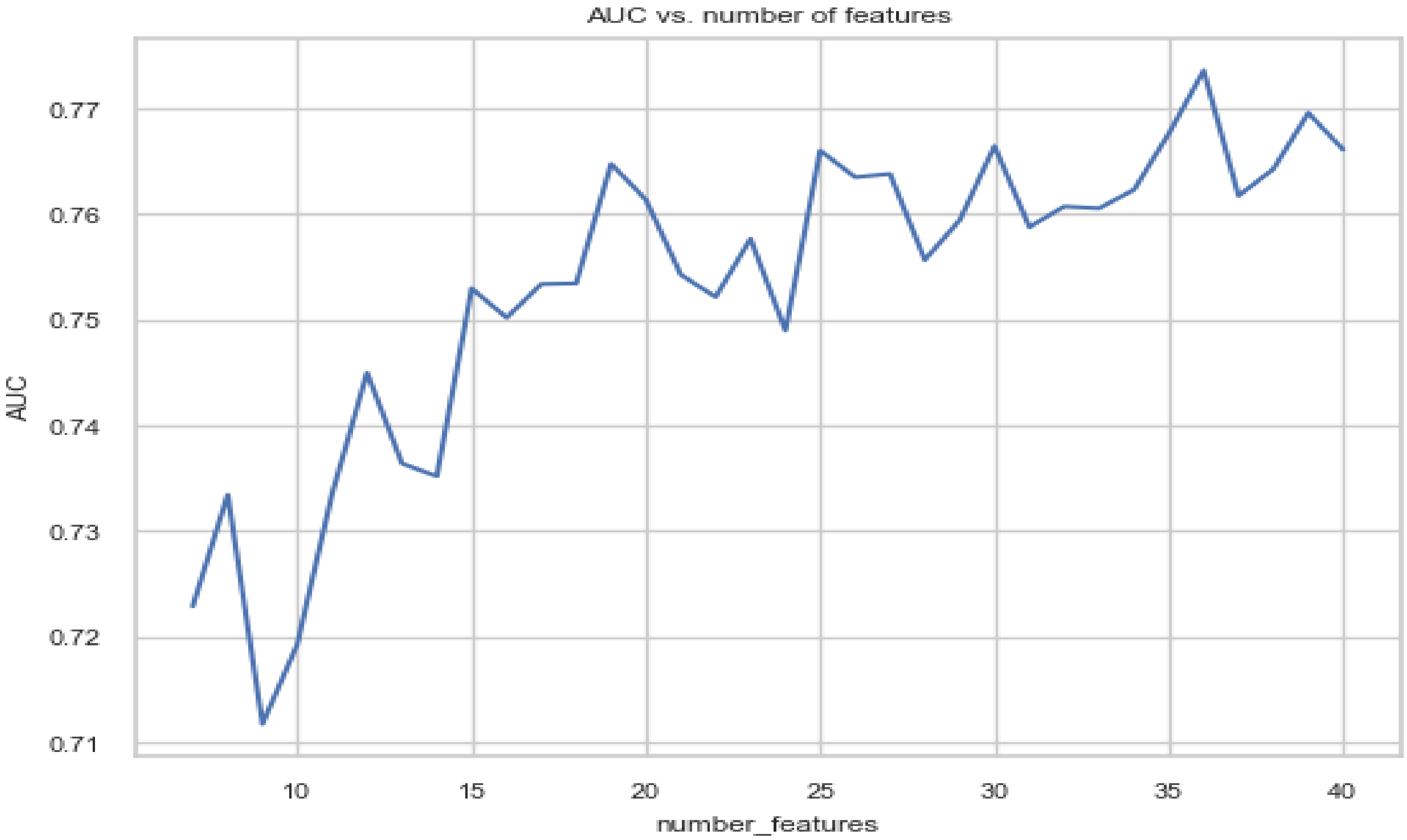


- The best model is **Random Forest**
- With:
  - Criterion = gini
  - Max_depth = 6
  - Max_features = 7
  - N_estimators = 100

All of the models attained better AUC results except SVM

6. FEATURE SELECTION

Recursive feature Elimination RFE: Fits a model and removes the weakest feature (or features) until the specified number of features is reached

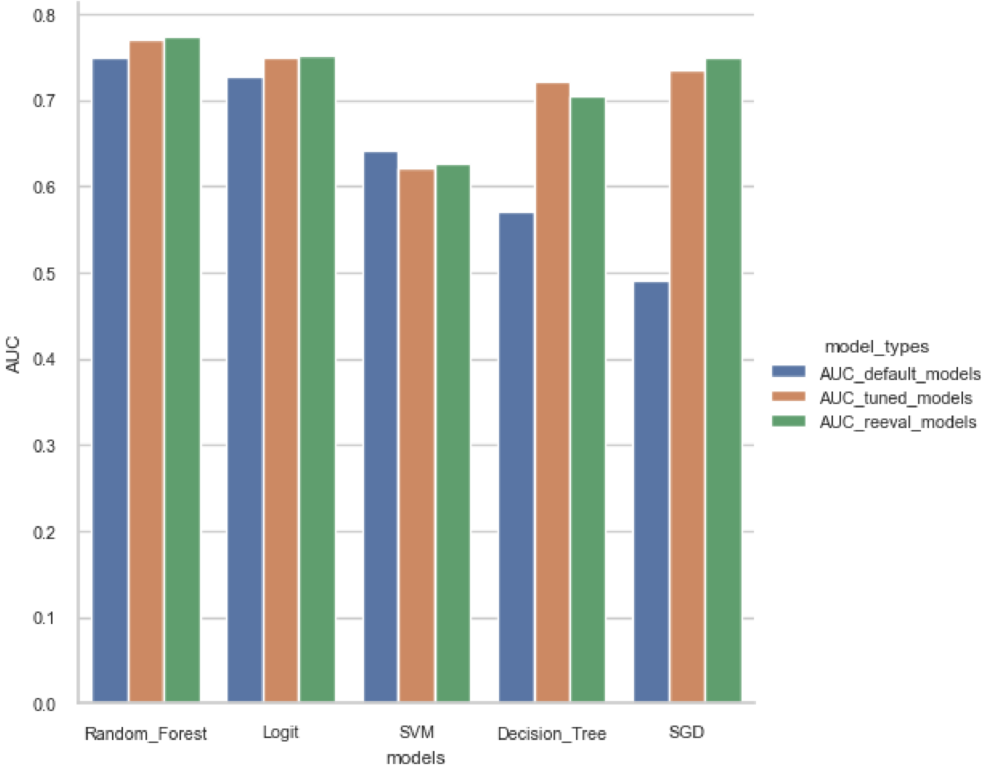


- Month
- credit_amount
- credit_term
- Age
- having_children_flg
- Region
- Income
- phone_operator
- is_client
- education_Higher education
- education_Incomplete higher education
- education_Incomplete secondary education
- education_Secondary education
- education_Secondary special education
- product_type_Audio & Video
- product_type_Auto
- product_type_Boats
- product_type_Cell phones
- product_type_Clothing
- product_type_Computers
- product_type_Construction Materials
- product_type_Cosmetics and beauty services
- product_type_Fitness
- product_type_Furniture
- product_type_Garden equipment
- product_type_Household appliances
- product_type_Jewelry
- product_type_Medical services
- product_type_Sporting goods
- product_type_Tourism
- product_type_Training
- product_type_Windows & Doors
- family_status_Another
- family_status_Married
- family_status_Unmarried
- sex_female

Selected Features= 36

Default vs. Tuned vs. Re-evaluated Models

Models	AUC_default_models	AUC_tuned_models	AUC_reeval_models
Random_Forest	0.74	0.77	0.77
Logit	0.72	0.74	0.75
SVM	0.64	0.62	0.62
Decision_Tree	0.57	0.72	0.70
SGD	0.49	0.73	0.74

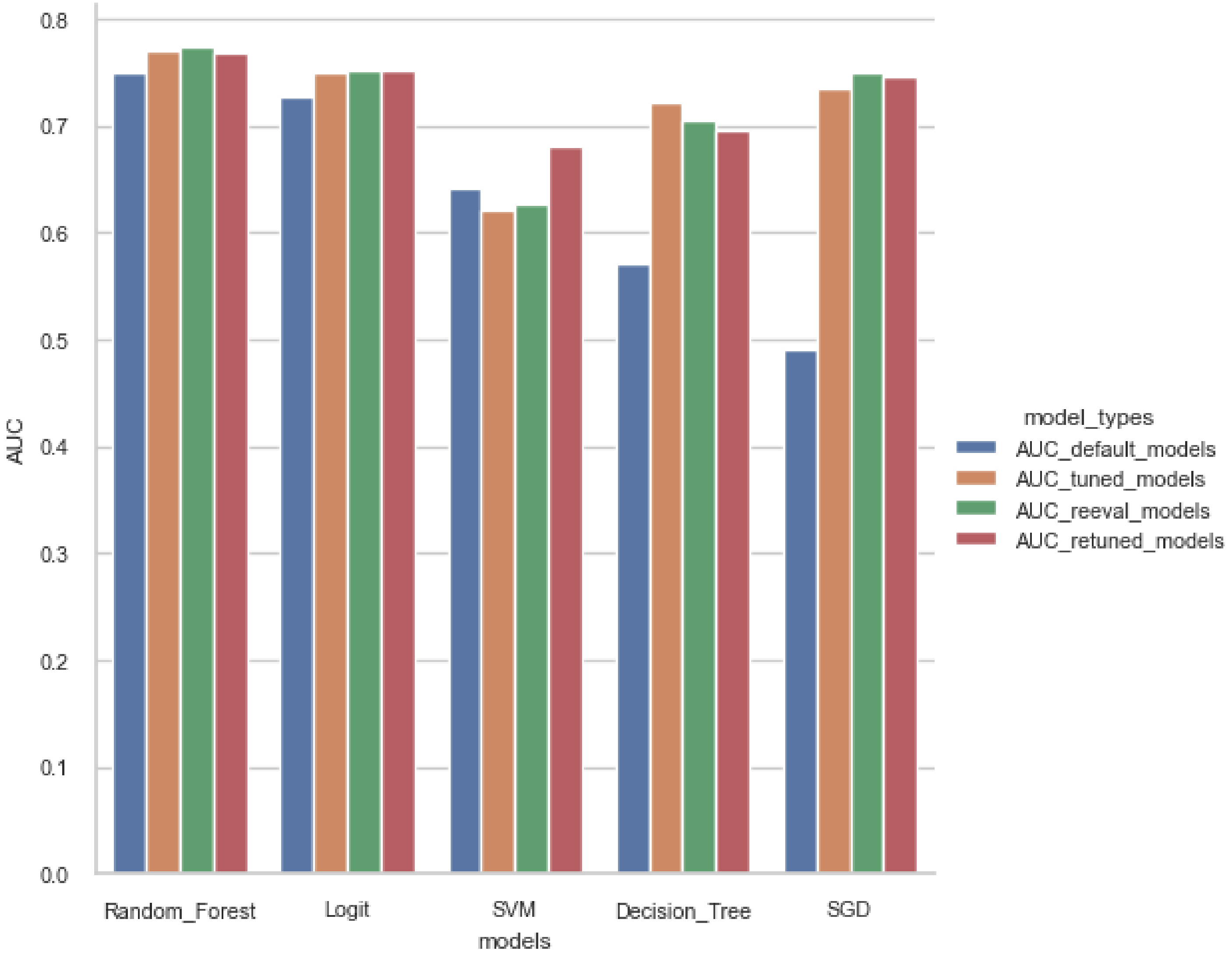


The best model after feature selection is Random Forest

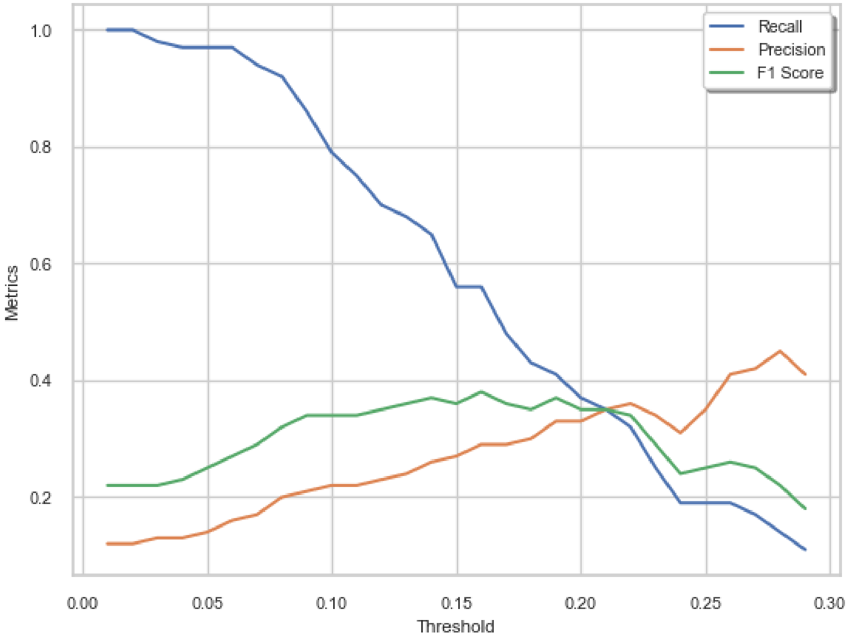
7. RETUNE MODEL WITH NEW REFINED DATA

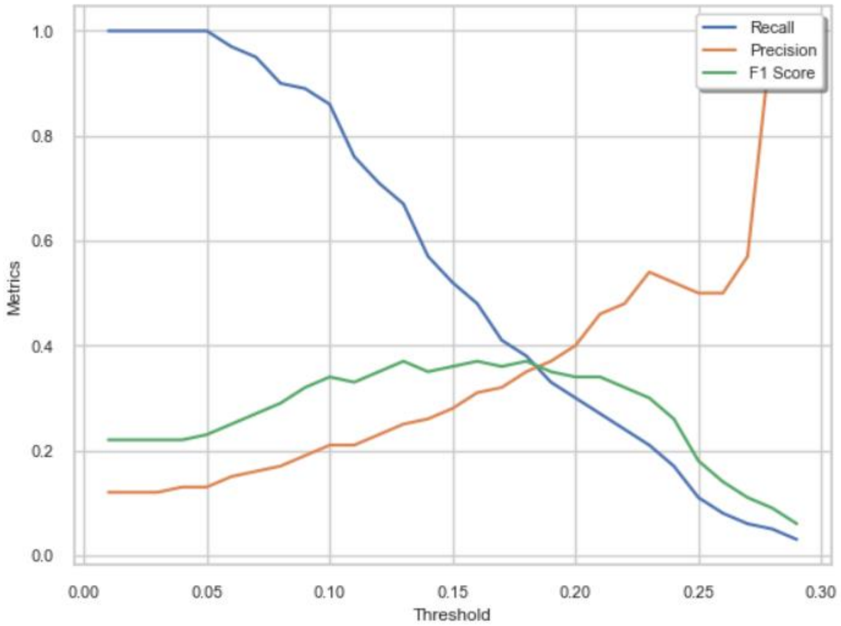
The best model after feature selection and fine tuning is Random Forest

Models	AUC Default models	AUC Tuned models	AUC Re-eval models	AUC Retuned models
Random_ Forest	0.74	0.77	0.77	0.76
Logit	0.72	0.74	0.75	0.75
SVM	0.64	0.62	0.62	0.68
Decision Tree	0.57	0.72	0.70	0.69
SGD	0.49	0.73	0.74	0.74



8. THRESHOLD ANALYSIS

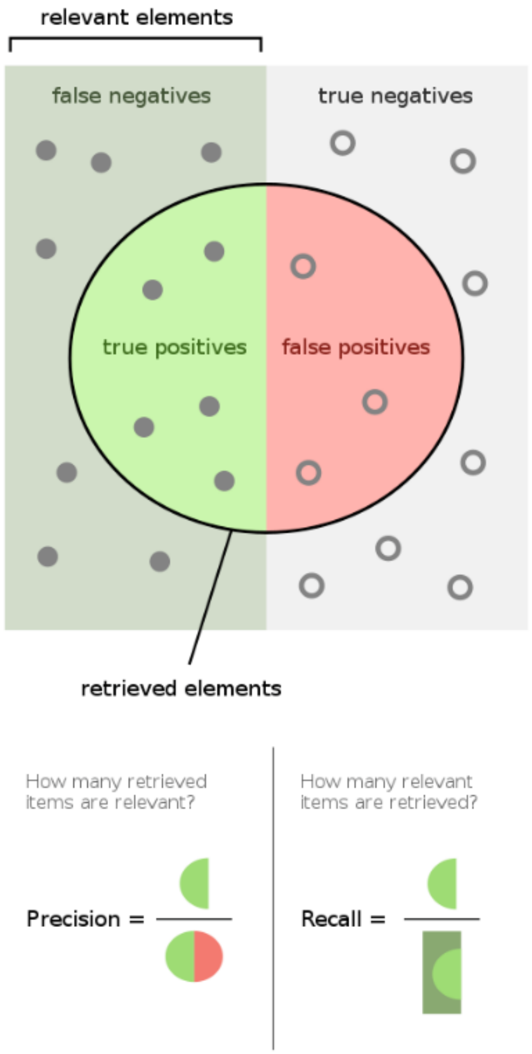




$$F = \frac{2}{\frac{1}{Recall} + \frac{1}{Precision}}$$

$$F = 2\,\frac{Precision\ x\ Recall}{Precision + Recall}$$

- Selecting the threshold is a trade-off between Recall and Precision



Source: Wikipedia

CONCLUSION

In these experiments, Random Forest algorithm with the refined dataset gave the highest ROC AUC.

After finding the best method of detecting bad clients, threshold analysis should be done, and the threshold value is set based on specific business purposes.