

Capstone Project –3 Credit Card Default Prediction

By

Himanshu Tiwari

Data Science Trainee, AlmaBetter

Contents

- ❑ Introduction
- ❑ Problem Statement
- ❑ Data Summary
- ❑ Approach Overview
- ❑ Exploratory Data Analysis
- ❑ Modelling Overview
- ❑ Feature Importances
- ❑ Challenges
- ❑ Conclusion

- ❑ In today's world credit cards have become a lifeline to a lot of people so banks provide us with credit cards.
- ❑ Now we know the most common issue there is in providing these kind of deals are people not being able to pay the bills.
- ❑ These people are what we call “defaulters”.
- ❑ Our aim would be to predict these users.

Problem Statement

- ☐ The problem statement is to predict whether a customer can default the loan.
- ☐ We have been provided with the various features in the dataset and we shall be using them to predict it.
- ☐ In the banking domain we need this type of models to understand the person's data and then provide the loan.

Data Summary



- ☐ X1 - Amount of credit(includes individual as well as family credit)
- ☐ X2 - Gender (1 = male; 2 = female).
- ☐ X3 - Education (1 = graduate school; 2 = university; 3 = high school; 4 = others)
- ☐ X4 - Marital Status (1 = married; 2 = single; 3 = others)
- ☐ X5 - Age(year).
- ☐ X6 to X11 - History of past payments from April to September
- ☐ X12 to X17 - Amount of bill statement from April to September
- ☐ X18 to X23 - Amount of previous payment from April to September
- ☐ Y - Default payment next month

Understanding and cleaning.

- ☐ Null Values analysis
- ☐ Outlier treatment

Graphical

- ☐ Univariate analysis with visualization
- ☐ Bivariate analysis with visualization.

Machine Learning

- ☐ Logistic regression
- ☐ SVM
- ☐ Decision tree
- ☐ Random Forest
- ☐ XGBoost
- ☐ CatBoost
- ☐ Lightgbm

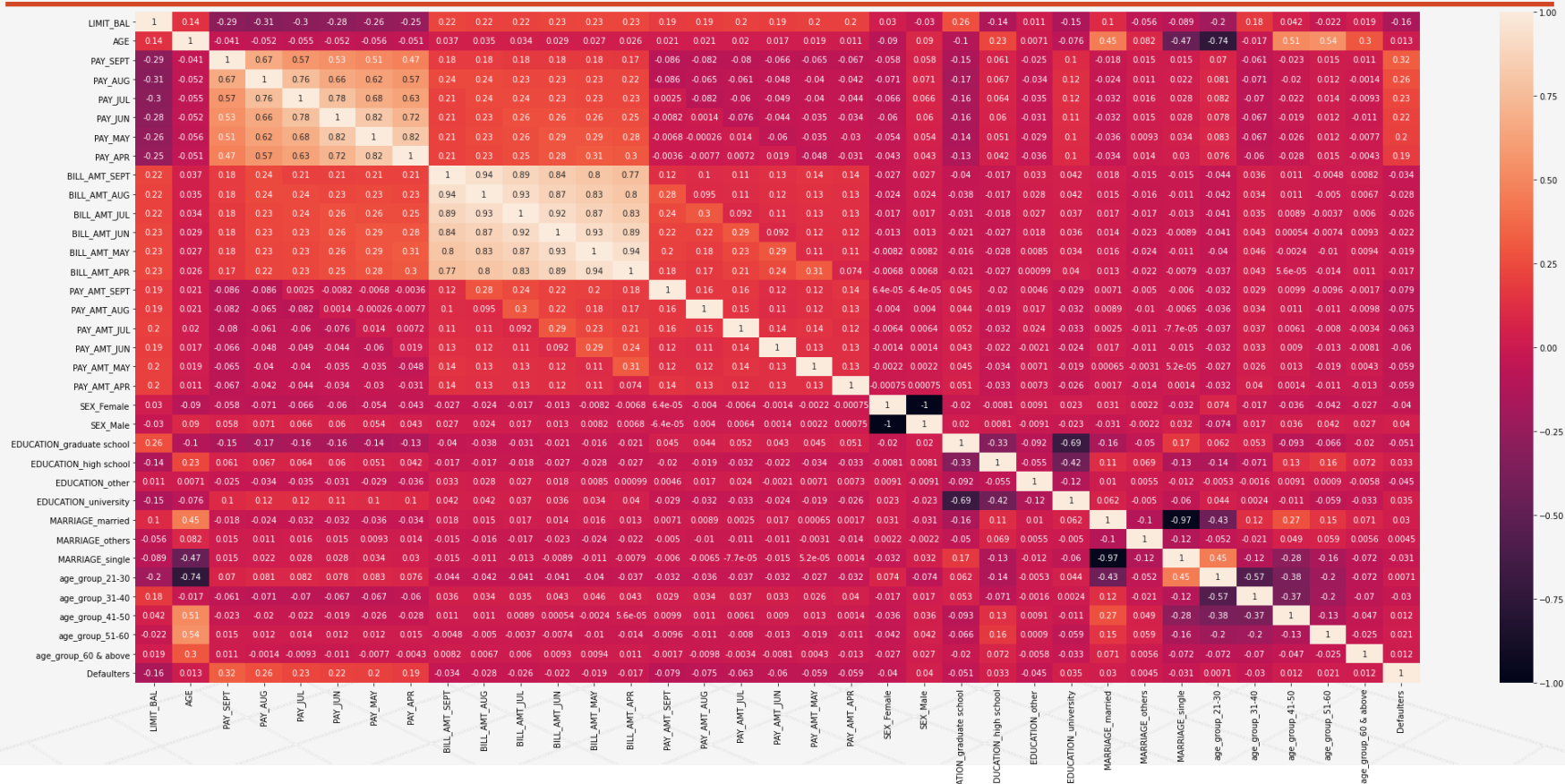
Basic Exploration



- ☐ Data of 30000 customers.
- ☐ 6 Months payment and bill data.
- ☐ No null data.
- ☐ 9 Categorical variables present.
- ☐ Various undocumented/wrong labels were present

Correlation Heatmap

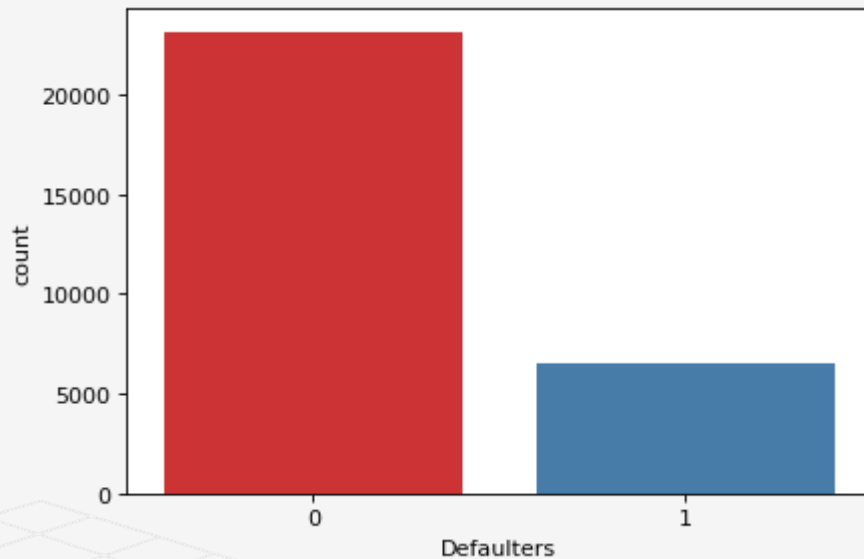
AI



Default Distribution



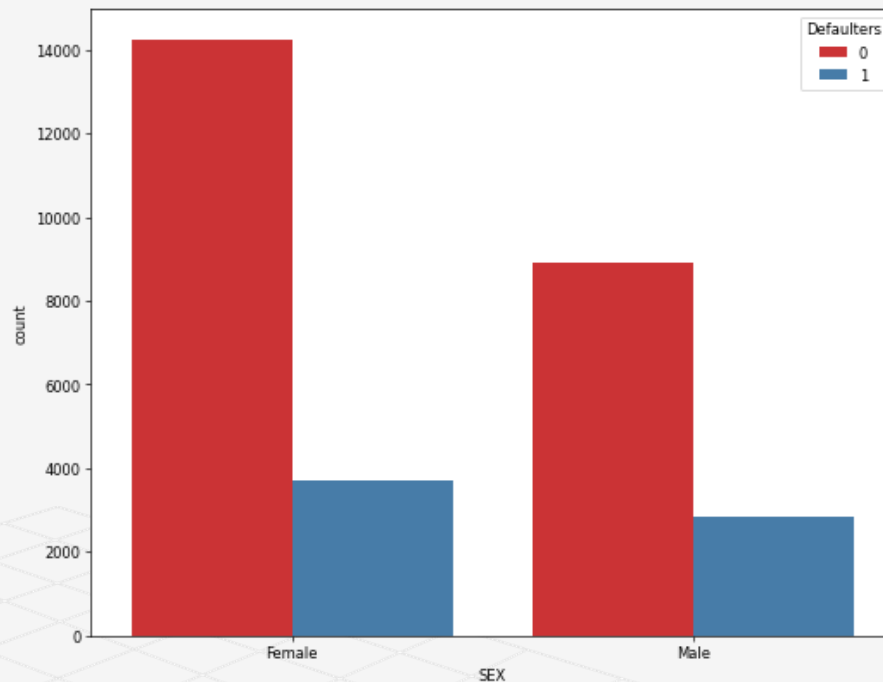
index	Is_Defaulter
0	77.88
1	22.12



Gender Distribution



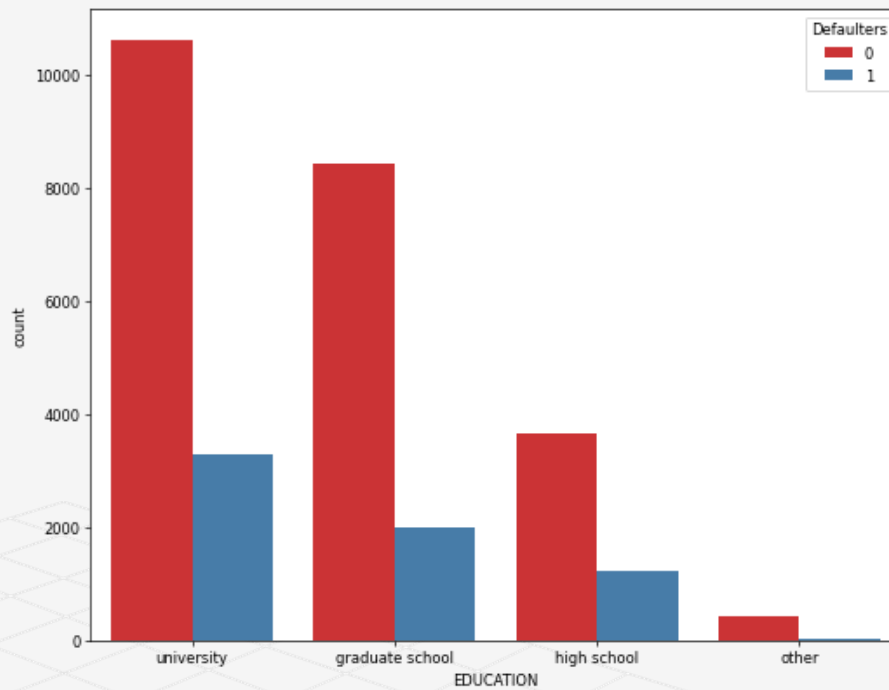
	SEX	Is_Defaulter
0	Female	20.776281
1	Male	24.167227



Education Distribution



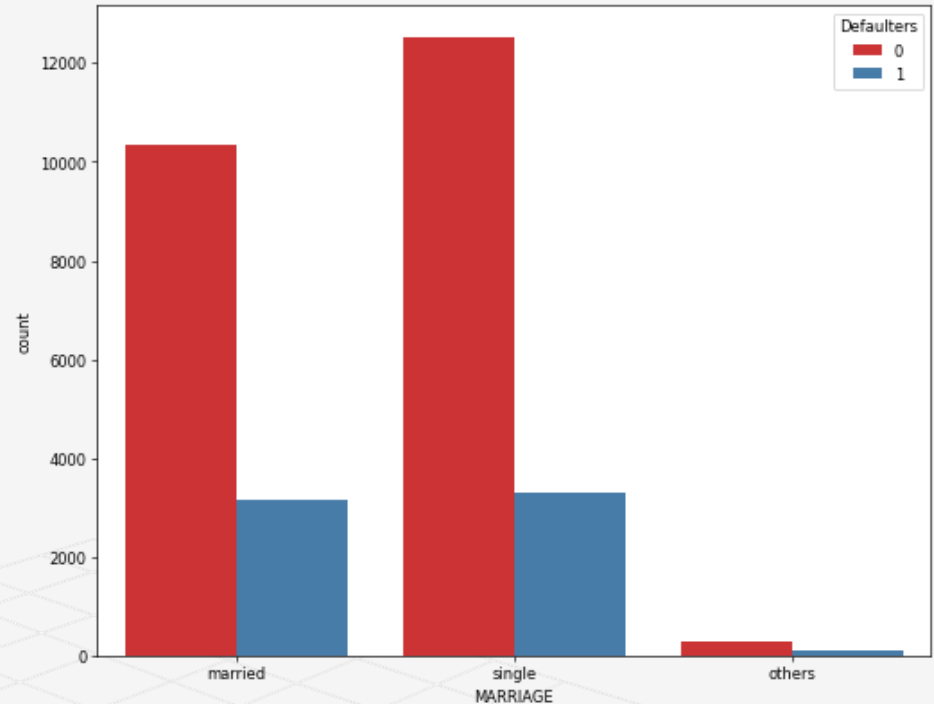
EDUCATION	Is_Defaulter
high school	25.157616
university	23.734854
graduate school	19.234766
others	7.051282



Marital Distributions

AI

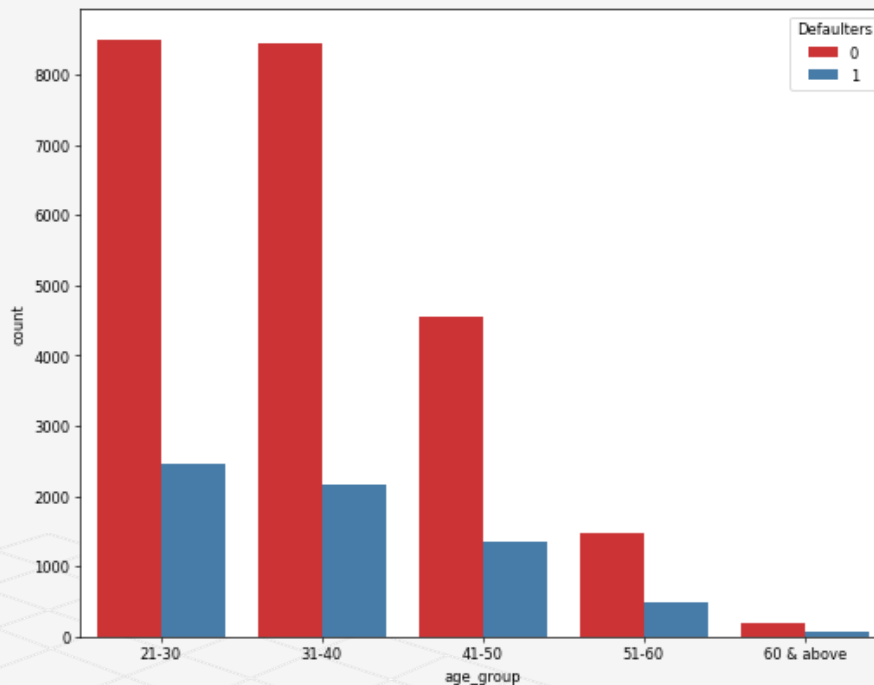
MARRIAGE	Is_Defaulter
Others	23.607427
Married	23.471704
Single	20.928339



Age Distribution



age_group	Is_Defaulter
60_&_above	28.318584
50_59	24.861170
40_49	22.973391
21_29	22.842587
30_39	20.252714



Modelling Overview

- ❑ Supervised learning
 - ❑ Binary Classification
- ❑ Imbalance data with 22% defaulters

Models Used:

- ❑ Logistic Regression
- ❑ Decision Trees
- ❑ Random Forest
- ❑ SVM

Modelling Steps



- ❑ Feature selection
- ❑ Feature engineering
- ❑ Train test data split(75%-25%)
- ❑ SMOTE oversampling
- ❑ Start with default model parameters
- ❑ Hyperparameter tuning
- ❑ Measure scores on training & test data
- ❑ Model testing
- ❑ Compare models

Parameters :

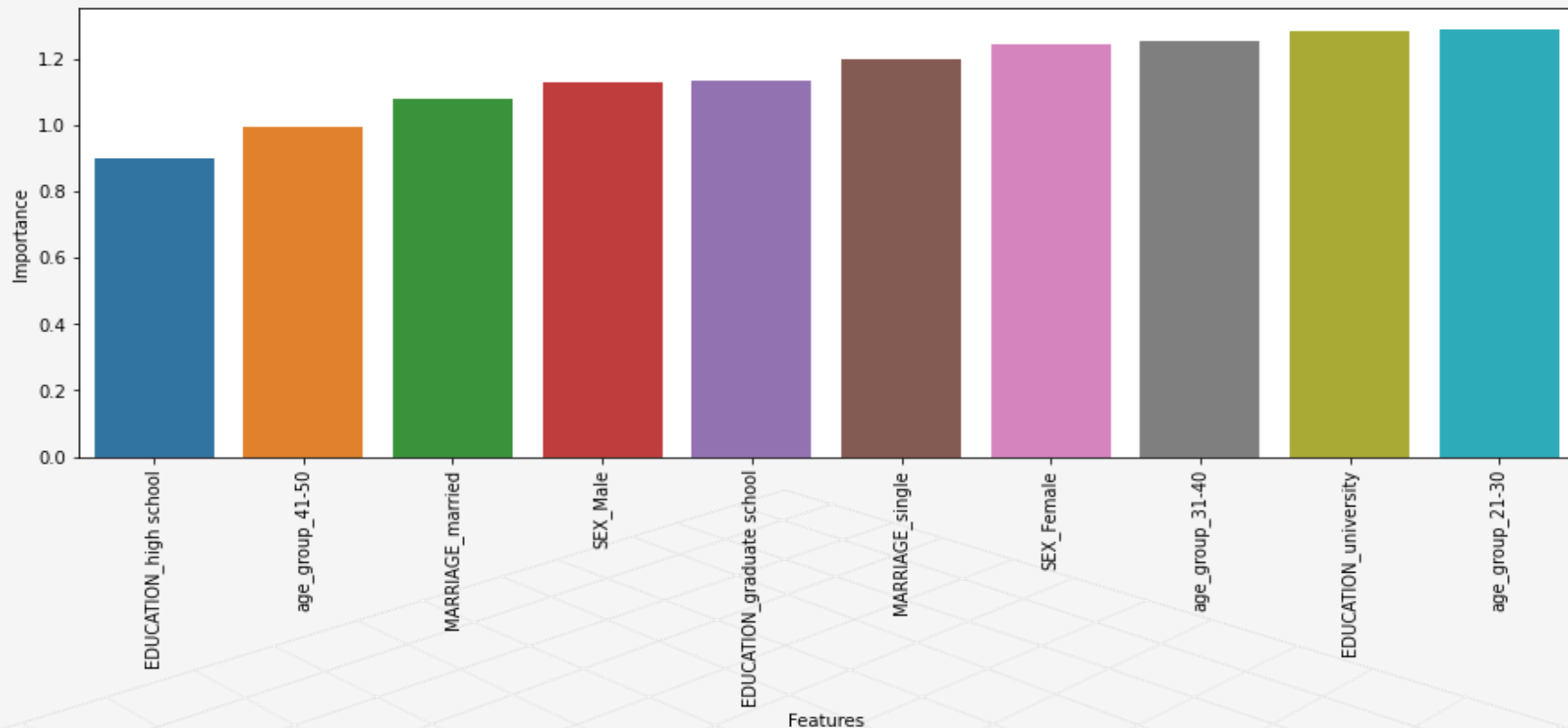
❑ **C = 0.01**

❑ **Penalty = L2**

Classification Report

	precision	recall	f1-score
0	0.81	0.97	0.88
1	0.96	0.77	0.85
accuracy			0.87

Feature Importances for Logistic Regression



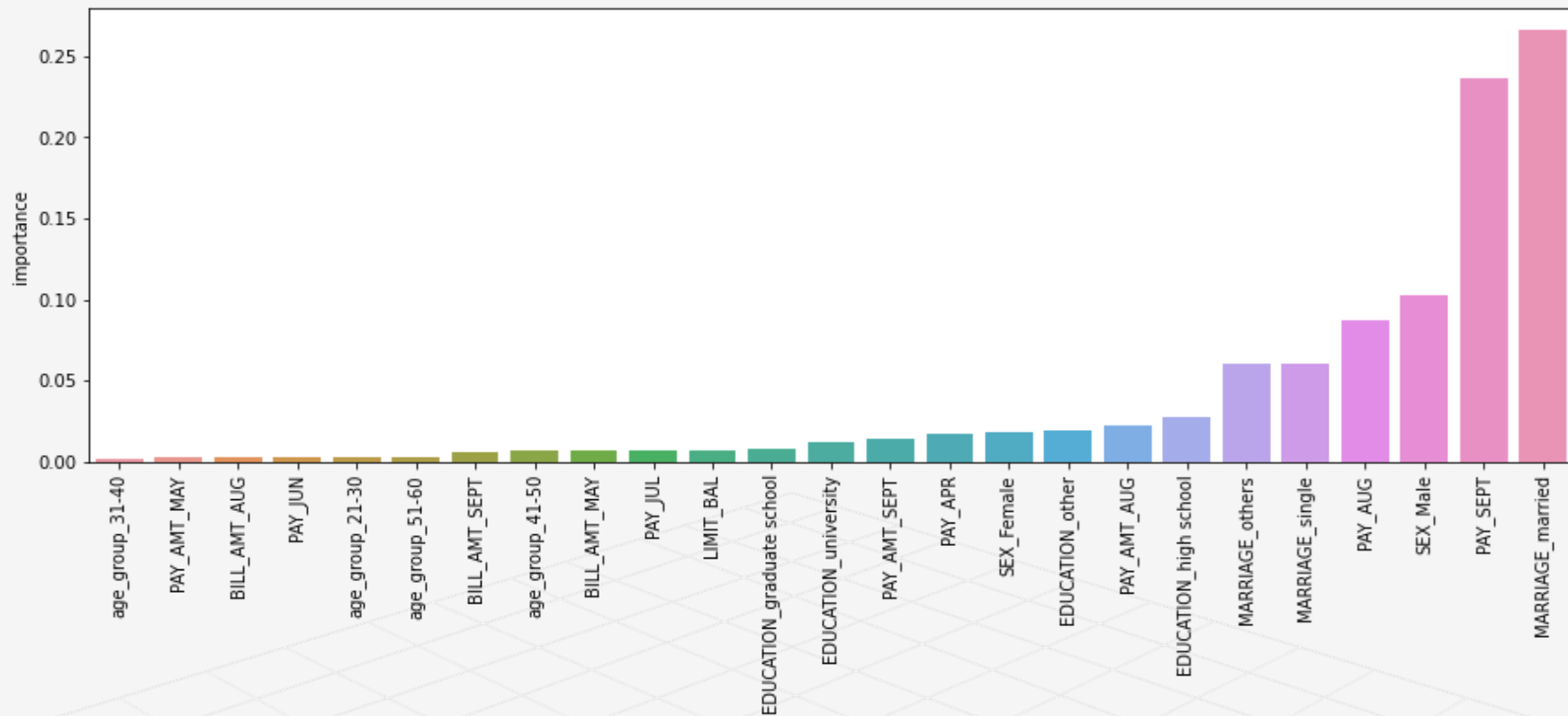
Parameters :

- ☐ **max_depth=10**
- ☐ **max_leaf_nodes=45**
- ☐ **criterion= Entropy**

Classification Report

	precision	recall	f1-score
0	0.77	0.93	0.84
1	0.91	0.72	0.81
accuracy			0.83

Decision Tree feature importances



Parameters

- ❑ **C = 10**
- ❑ **Kernel = 'rbf'**

Classification Report

	precision	recall	f1-score
0	0.82	0.96	0.88
1	0.95	0.79	0.86
accuracy			0.87

Parameters :

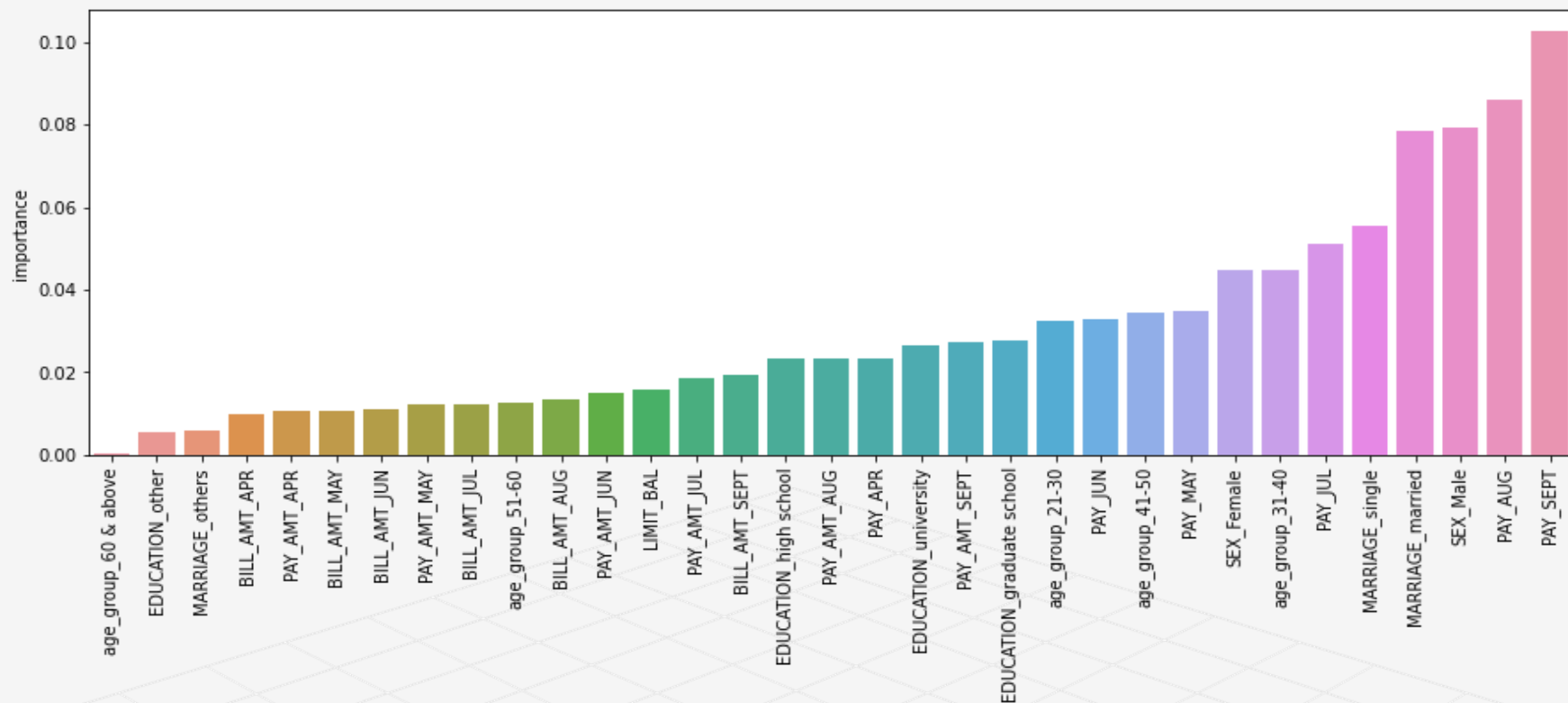
- ❑ **max_depth=9**
- ❑ **n_estimators=200**
- ❑ **criterion: entropy**

Classification Report

	precision	recall	f1-score
0	0.85	0.89	0.87
1	0.89	0.85	0.87
accuracy			0.87

Random Forest feature importances

AI



Parameters :

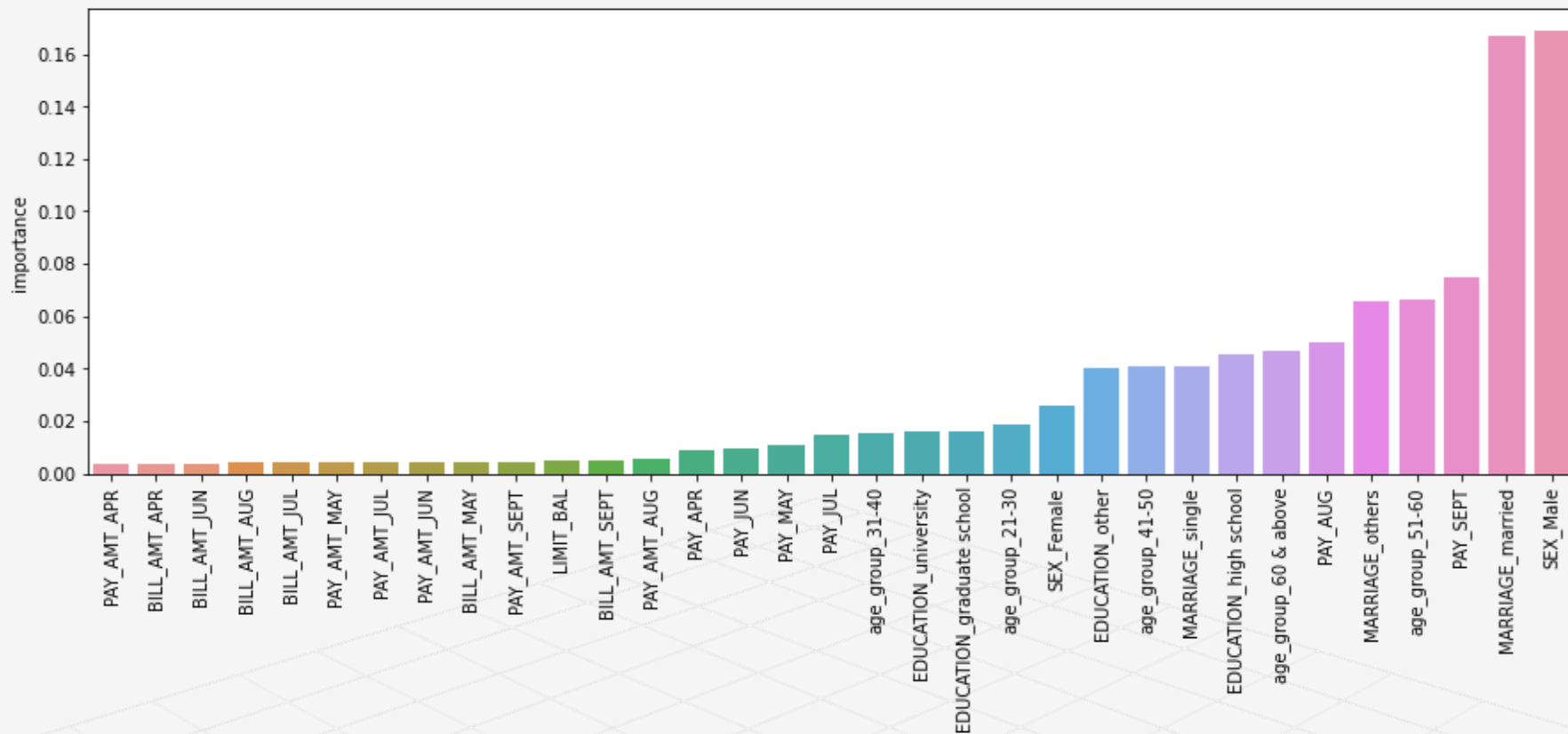
- ❑ **max_depth= 9**
- ❑ **n_estimators=150**

Classification Report

	precision	recall	f1-score
0	0.84	0.94	0.89
1	0.93	0.82	0.87
accuracy			0.88

XGB feature importances

AI



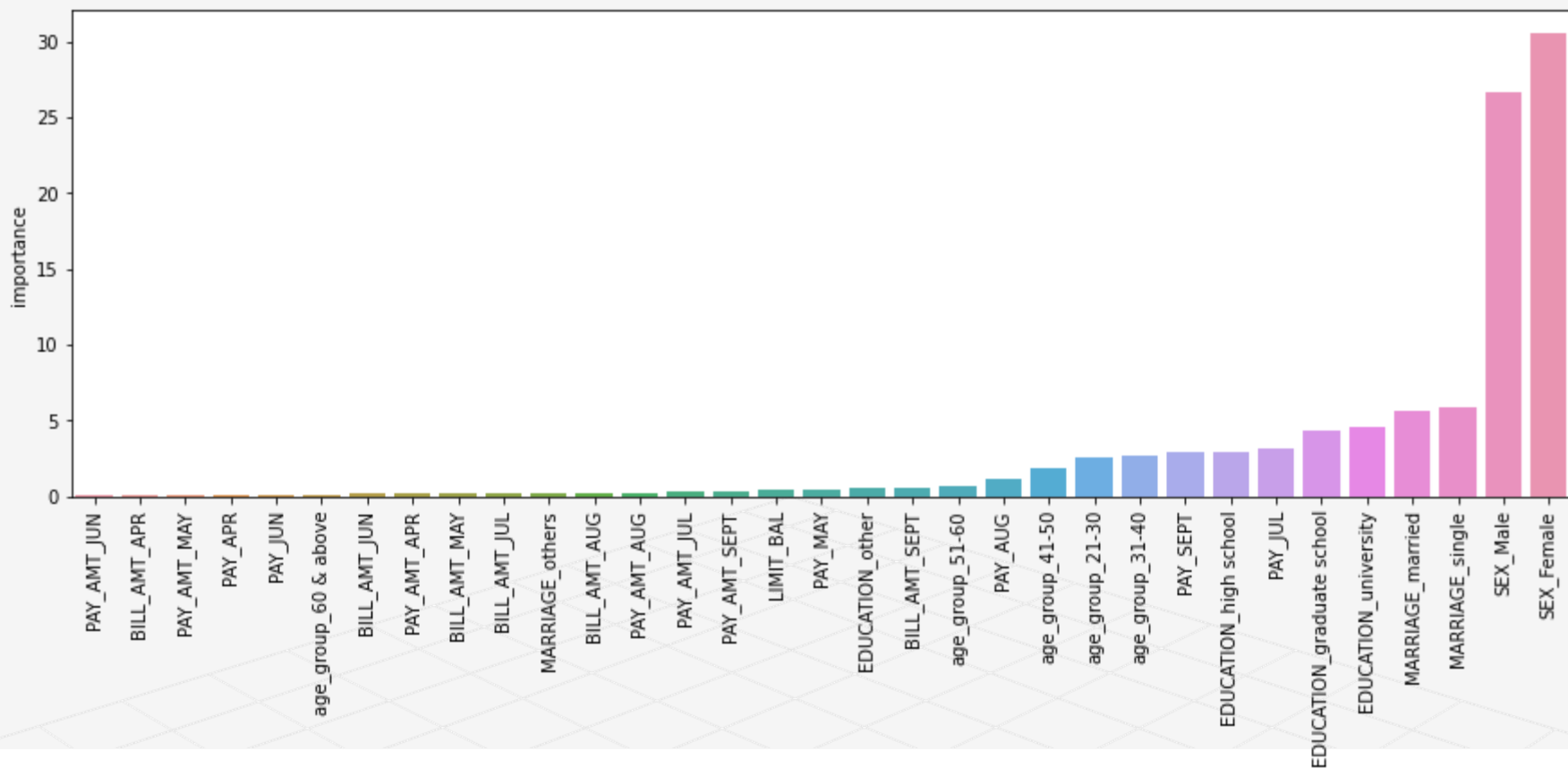
Parameters :

- ❑ **max_depth=3,**
- ❑ **n_estimators: 150**

Classification Report

	precision	recall	f1-score
0	0.83	0.95	0.88
1	0.94	0.80	0.86
accuracy			0.87

CatBoost feature importances



Parameters :

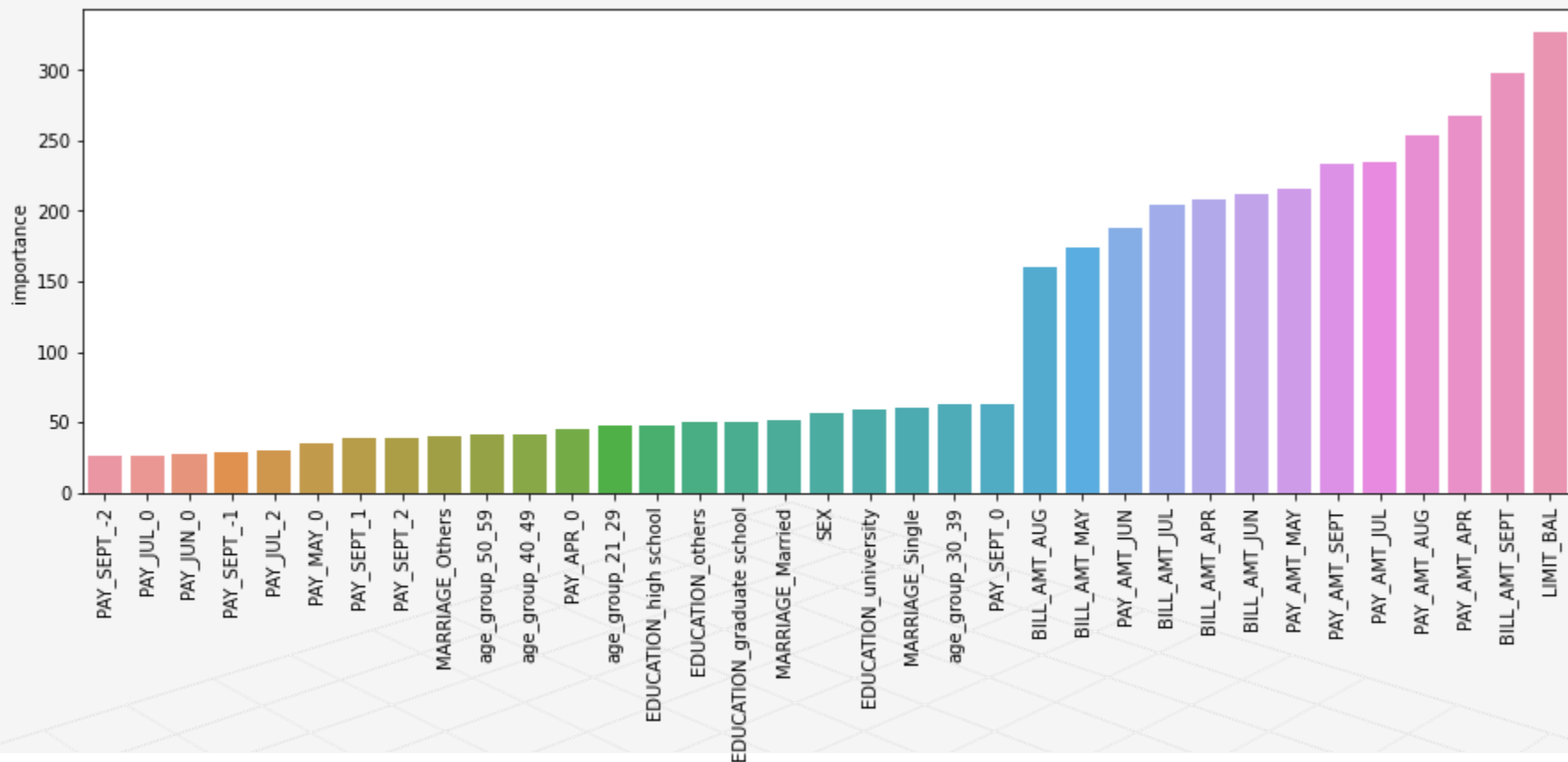
- ❑ **max_depth=7**
- ❑ **n_estimators:**
150

Classification Report

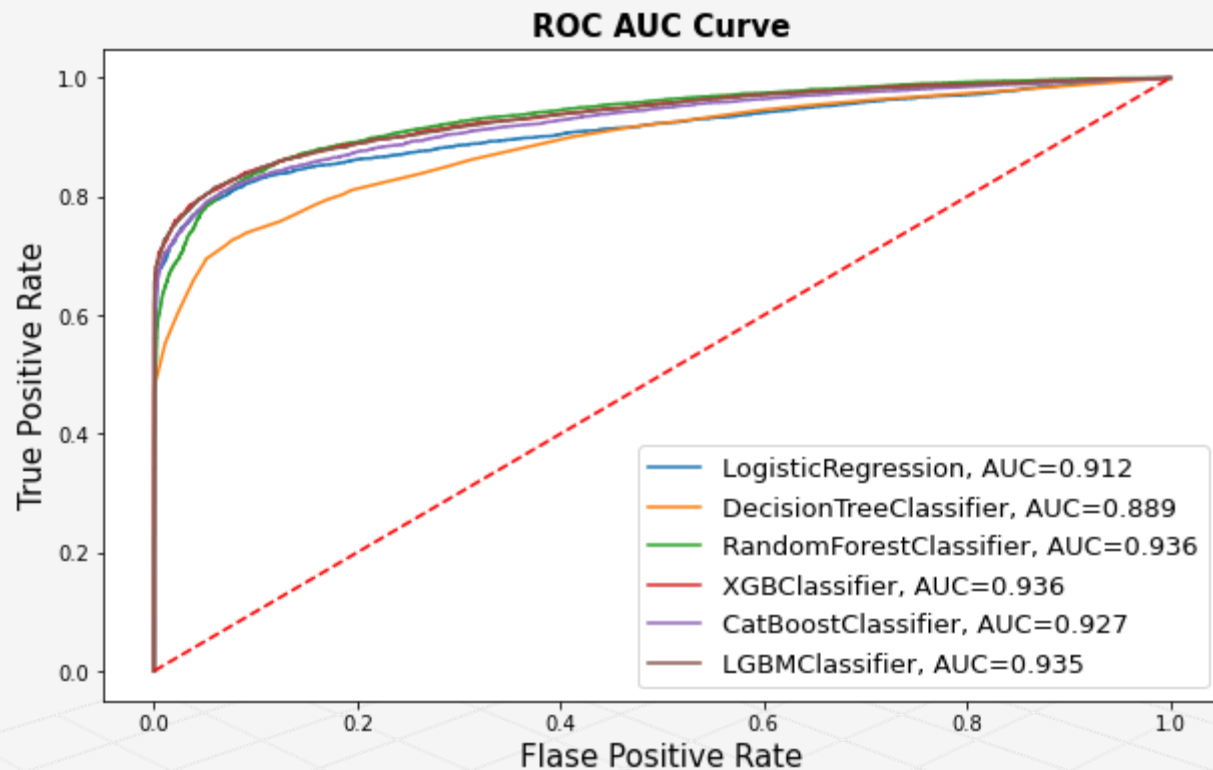
	precision	recall	f1-score
0	0.84	0.94	0.88
1	0.93	0.82	0.87
accuracy			0.88

Random Forest Classifier

AI



AUC-ROC curve comparison



Score Matrix



	Models	accuracy	precision	recall	f1	roc_auc
0	Logestic Regrestion	0.863511	0.953632	0.766049	0.849610	0.864155
1	grid_log_regg	0.864115	0.950053	0.770512	0.850915	0.864734
2	Desision Tree	0.823428	0.891674	0.738929	0.808147	0.823986
3	Random forest	0.874482	0.902152	0.841916	0.870994	0.874697
4	grid random forest	0.869730	0.901749	0.831789	0.865357	0.869981
5	SVM	0.869039	0.943051	0.787333	0.858185	0.869579
6	Grid SVM	0.868435	0.931076	0.797631	0.859203	0.868903
7	XGboost	0.867312	0.921249	0.805184	0.859315	0.867722
8	Grid Xgboost	0.875173	0.925092	0.818229	0.868385	0.875549
9	CATBoost	0.872927	0.926877	0.811535	0.865379	0.873332
10	Grid Catboost	0.869039	0.912519	0.818229	0.862805	0.869375
11	LightGBM	0.876469	0.932337	0.813594	0.868928	0.876884
12	Grid LightGBM	0.877332	0.933320	0.814452	0.869844	0.877748

Challenges Faced



- ❑ **Understanding the columns.**
- ❑ **Feature engineering.**
- ❑ **Getting a higher recall on the models.**

Conclusion



- ❑ The default rate is higher for males, increases as the education increases, also increases as the age of a person increases. i.e clients whose age over 60 was higher than mid-age and young people.
- ❑ In all of these models, our recall revolves in the range of 76 to 84%.with the best fit model as random forest

Thank you ...

