

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

Introduction



- ☐ 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)
- \square 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

Implementation



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

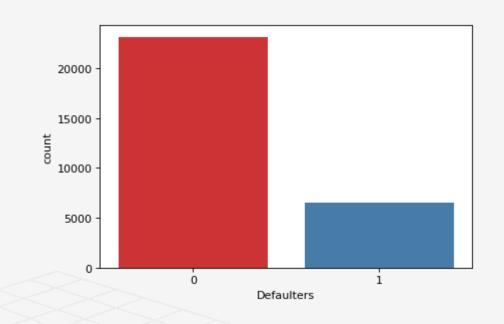


Martine 14 15 15 15 15 15 15 15																																				
PM MATERIAL PROPERTY 10 10 10 10 10 10 10 10 10 10 10 10 10	LIMIT_BAL -	1	0.14	-0.29			-0.28	-0.26																						-0.089			0.042			
Marine M	AGE -		1	-0.041	-0.052	-0.055	-0.052	-0.056	-0.051																	0.0071			0.082	-0.47	-0.74	-0.017	0.51			
Marine M	PAY_SEPT -			1	0.67	0.57																													0.011	
HILMATION 648 648 658 658 658 658 658 658 658 658 658 65	PAY_AUG -			0.67	1	0.76	0.66	0.62	0.57										-0.048																-0.0014	
Part	PAY_JUL -			0.57	0.76	1	0.78	0.68	0.63							0.0025																			-0.0093	
**************************************	PAY_JUN -				0.66	0.78	1	0.82	0.72							-0.0082			-0.044																	
SIL,MIT,SET 072 083 018 018 024 021 023 023 025 018 024 024 023 023 025 028 023 025 028 023 024 024 023 023 025 028 028 025 025 023 023 025 028 023 025 029 023 025 023 023 025 029 023 025 029 023 025 023 023 025 023 023 025 023 023 025 023 023 023 025 023 023 023 023 023 023 023 023 023 023	PAY_MAY -				0.62	0.68	0.82	1	0.82							-0.0068																				
BIL,AMT,AMS	PAY_APR -				0.57	0.63	0.72	0.82	1							-0.0036				-0.048		-0.043													-0.0043	
BILLAMT IN - 22	BILL_AMT_SEPT -									1	0.94	0.89	0.84	0.8	0.77	0.12																		-0.0048		-0.034
BIL_AMT_INA 02 02 027 018 023 029 028 028 028 028 028 028 028 028 028 028	BILL_AMT_AUG -									0.94	1	0.93	0.87	0.83	0.8	0.28																				-0.028
BIL_ANT_MR. 23 02.0 02.0 01.0 02.0 02.0 01.0 02.0 02.0	BILL_AMT_JUL -									0.89	0.93	1	0.92	0.87	0.83	0.24																	0.0089			
BILLAMT,APR 02 03 06 06 07 07 08 07 08 07 08 07 08 08 07 08 08 07 08 08 07 08 08 08 07 08 08 08 08 08 08 08 08 08 08 08 08 08	BILL_AMT_JUN -									0.84	0.87	0.92	1	0.93	0.89	0.22																				
PAY_ANT_SIGN - 019 019 020 030 030 030 030 030 030 030 030 030	BILL_AMT_MAY									0.8	0.83	0.87	0.93	1	0.94	0.2																				
PAY_ANT_JUL 10 0.01	BILL_AMT_APR									0.77	0.8	0.83	0.89	0.94	1	0.18						-0.0068	0.0068													
Fig.	PAY_AMT_SEPT -			-0.086	-0.086	0.0025	-0.0082	-0.0068	-0.0036						0.18	1	0.16									0.0046										-0.079
PRY_ANT_INA* - 01 017 000 010 000 000 000 000 000 000	PAY_AMT_AUG -															0.16	1	0.15																		
PRY ANT MAY 02 019 0.065 0.04 0.05 0.05 0.04 0.05 0.05 0.05 0.0	PAY_AMT_JUL -																0.15	1												-7.7e-05						
PAY ANT APR - 02 011 0.067 0.024 0.014 0.054 0.024 0.034 0.034 0.03 0.031 0.14 0.03 0.034 0.03 0.034 0.03 0.034 0.	PAY_AMT_JUN -				-0.048	-0.049													1																	
SEX_Female	PAY_AMT_MAY								-0.048										0.13	1	0.13														0.0043	
SKY Male - 0.03 0.09 0.05 0.071 0.06 0.06 0.05 0.071 0.06 0.06 0.05 0.072 0.07	PAY_AMT_APR -				-0.042	-0.044															1	-0.00075														
EDUCATION graduate school 0.00 0.01 0.01 0.00 0.00 0.00 0.00 0.0	SEX_Female 1								-0.043						-0.0068	6.4e-05					-0.00075	1	-1													
EDUCATION_high school - 0.14	SEX_Male -															-6.4e-05						-1	1													
EDUCATION, other - 0011 0071 0072 0073 0074 0075 0074 0075 0074 0075 0075 0074 0075 0075	EDUCATION_graduate school																						0.02	1	-0.33		-0.69	-0.16								
EDUCATION university: 0.15 0.076 0.1 0.12 0.12 0.11 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.	EDUCATION_high school																							-0.33	1	-0.055	-0.42									
MARRIAGE married - 01 045 0018 002 015 011 010 0015 0029 0039 014 0015 0019 0015 0010	EDUCATION_other													0.0085	0.00099	0.0046										1										-0.045
MARRIAGE chers - 0.05 0.082 0.01	EDUCATION_university																							-0.69	-0.42	-0.12	1									
MARRIAGE single - 0.089	MARRIAGE_married -		0.45																									1	-0.1	-0.97	-0.43					
age_group_13-40 - 0.12	MARRIAGE_others		0.082																									-0.1	1	-0.12			0.049			0.0045
age_group_51-60 - 0.022	MARRIAGE_single		-0.47																									-0.97	-0.12	1	0.45	-0.12	-0.28	-0.16		
age group 51-60 - 042 051 0.023 0.02 0.02 0.019 0.02 0.012 0.014 0.012 0.015 0.012 0.014 0.015 0.012 0.014 0.012 0.015 0.012 0.014 0.015 0.012 0.015 0.014 0.015 0.012 0.015 0.014 0.015 0.012 0.015 0.014 0.015 0.015 0.014 0.015 0	age_group_21-30 -		-0.74							-0.044	-0.042																			0.45	1	-0.57	-0.38			
## age group 51-60 - 0022 054 0015 0012 0014 0012 0014 0012 0015 0008 0005 0005 0005 0005 0005 0005	age_group_31-40		-0.017																												-0.57	1	-0.37	-0.2		
age_group_60 & above - 0.019 03 0011 0.0014 0.0093 0.011 0.0014 0.0093 0.011 0.0017 0.0043 0.0082 0.0067 0.006 0.0093 0.0014 0.0017 0.0098 0.0034 0.0081 0.0043 0.013 0.027 0.027 0.027 0.027 0.027 0.0058 0.033 0.011 0.0056 0.072 0.072 0.072 0.073 0.047 0.025 1 0.012 0.012 0.013 0.014 0.013 0.014 0.015 0.013 0.014 0.015 0.01	age_group_41-50		0.51																											-0.28	-0.38	-0.37	1	-0.13		
Defaulters - 0.16 0013 032 026 023 022 0.2 0.19 0.034 0.028 0.026 0.022 0.019 0.017 0.079 0.075 0.063 0.06 0.059 0.059 0.059 0.051 0.033 0.045 0.035 0.03 0.045 0.031 0.0071 0.03 0.012 0.021 0.012 1	age_group_51-60		0.54							-0.0048																							-0.13	1	-0.025	
AUG - APR -	age_group_60 & above -			0.011	-0.0014	-0.0093			-0.0043								-0.0098																	-0.025	1	0.012
	Defaulters -	-0.16																-0.063								-0.045			0.0045						0.012	1
		A.	- GE	F	- SU	H.	N	AY -	8.	Н	- DC	Ä	N	AY -	E.	Ы	- PG	a d	Š	AY -	PR-	ale	ale -	100	- 100	je	ity	- pa	- E	ale -	30 -	40 -	- 05	- 09	ove-	- SI
PAY AN PA		E .	₹	Y SE	AY AI	AY.	J/AY	AY M	AY A	T_SE	H_A	F	Ĕ	M.	A_A	T_SE	II A	IM	Į.	M.	A_M	Femi	×	schi	sch	to N	ivers	narri	oth	sing	21-	31.	41	- 51	× abc	aulte
BILL BILL BILL BILL BILL BILL BILL BILL		E		PA	7		- CONTRACTOR	ď	Œ	AM	LAN	IL.A	IL.A	LAN	LA	AM	Y AM	AY A	AY_AI	Y AN	NY A	SE.	Ж	duate	high	ATION	5	GE_	IAGE	IAGE	group	group	group	group	9 09	Def
										BILL	릚	<u>=</u>	=	H B	Ħ	Æ	PA	TO COLUMN TO STATE OF THE STATE	ď.	A.	4	OD BOOK OF THE PARTY OF THE PAR		grad	NOL	DOC	4TIO!	RRIA	IARR	MARR	900	age 6	ge g	age 5	roup	

Default Distribution



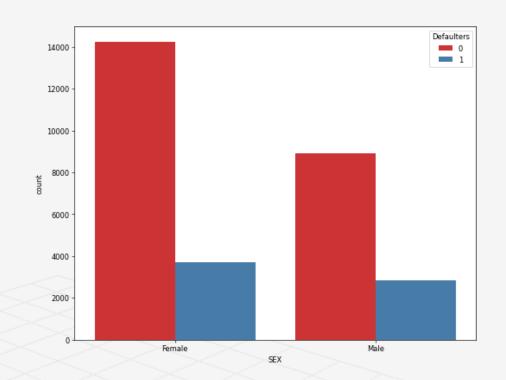
index	Is_Defaulter
0	77.88
1	22.12



Gender Distribution



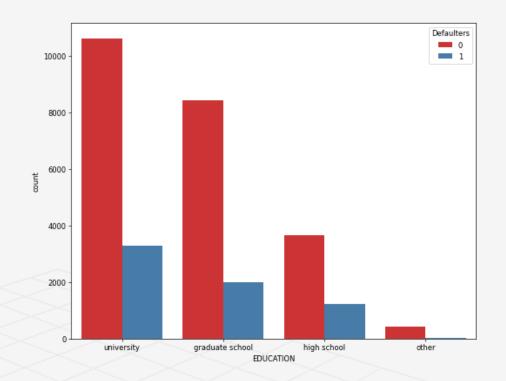
	SEX	Is_Defaulter
0	Female	20.776281
1	Male	24.167227



Education Distribution



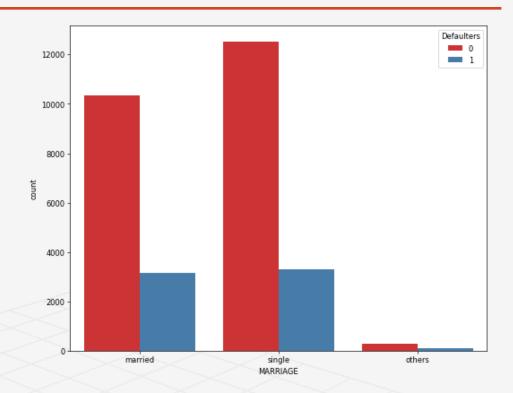
EDUCATION	Is_Defaulter
high school	25.157616
university	23.734854
graduate school	19.234766
others	7.051282



Marital Distributions



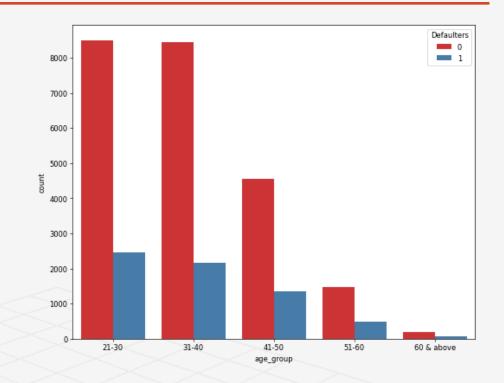
MARRIAGE	Is_Defaulter
Others	23.607427
Married	23.471704
Single	20.928339



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

Logistic Regression Model



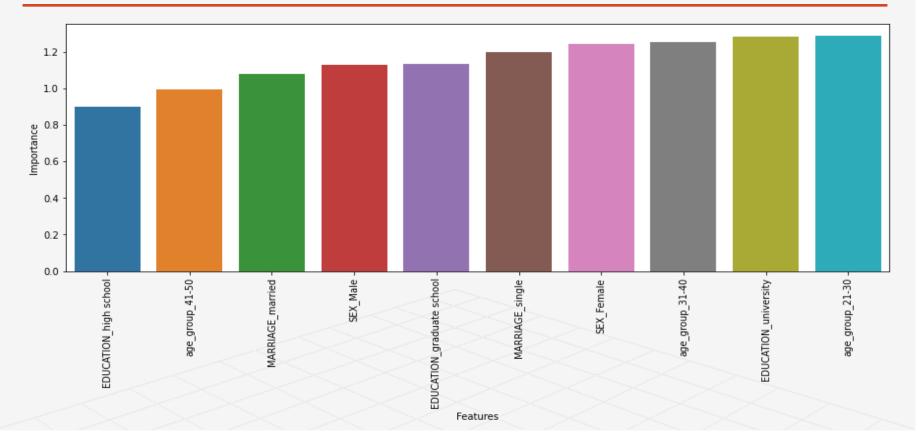
Parameters:

- \Box C = 0.01
- ☐ Penalty = L2

Classification Report								
pr	recision	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





Decision Tree



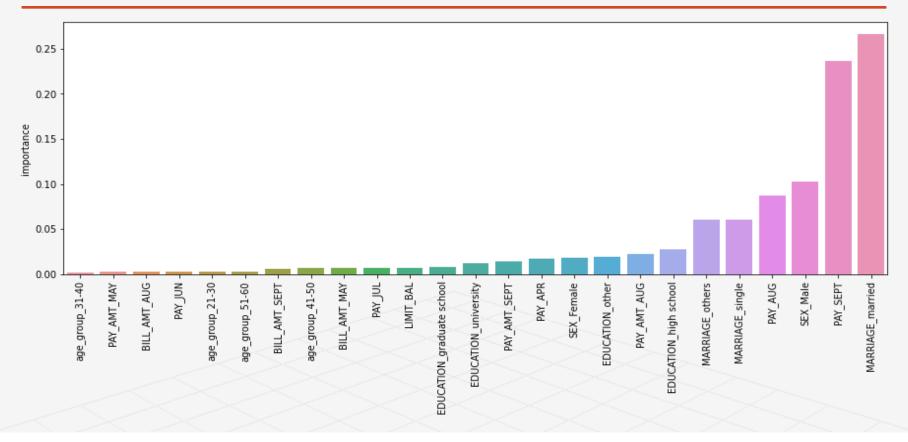
Parameters:

- ☐ 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





SVM Classifier



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						

Random Forest Metrics



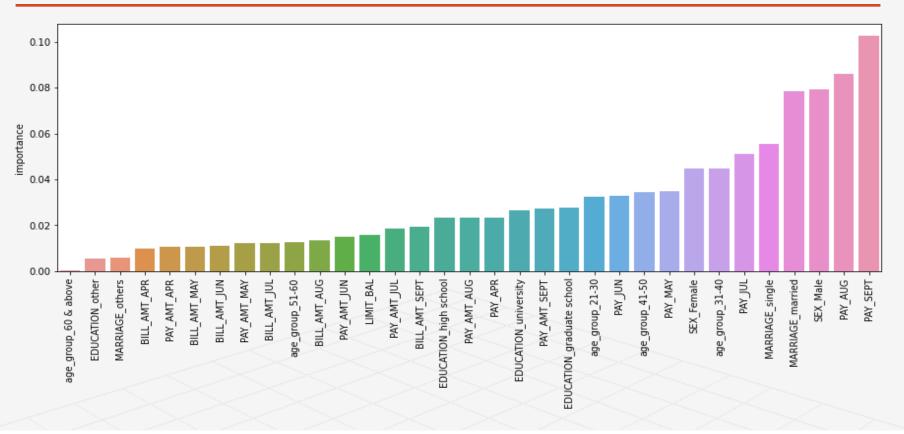
Parameters:

- ☐ 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





XGBoost Modelling



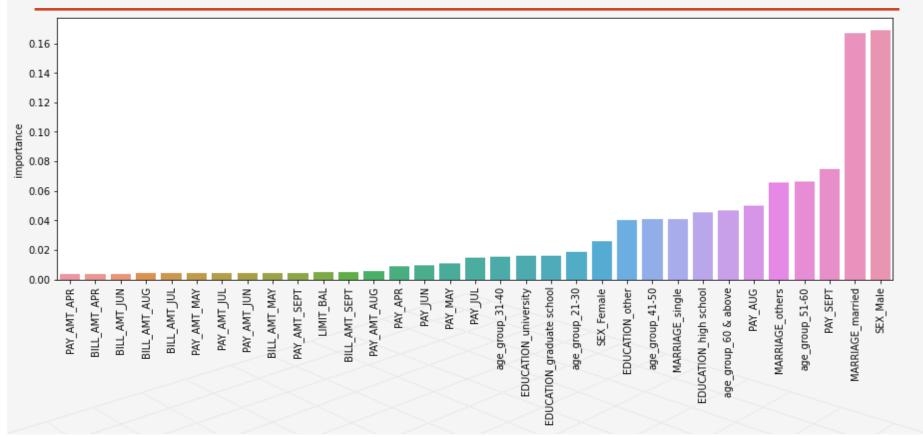
Parameters:

- ☐ max_depth= 9
- □ n_estimators=150

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

XGB feature importances







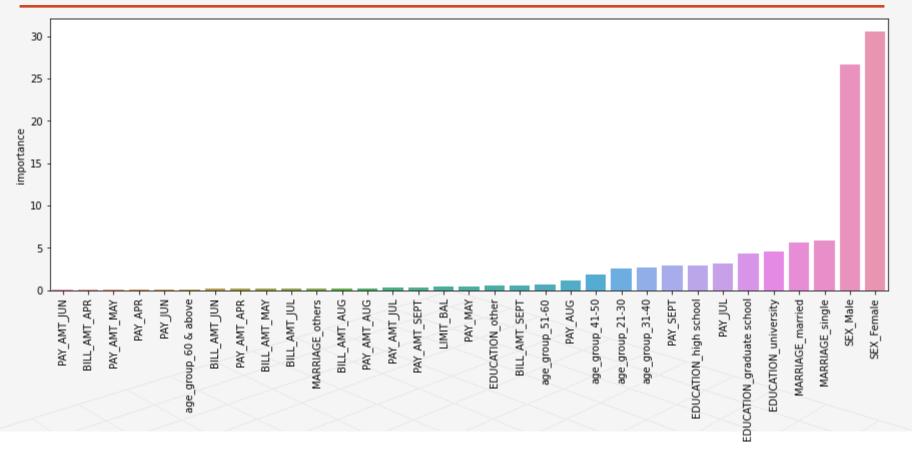
Parameters:

- □ 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







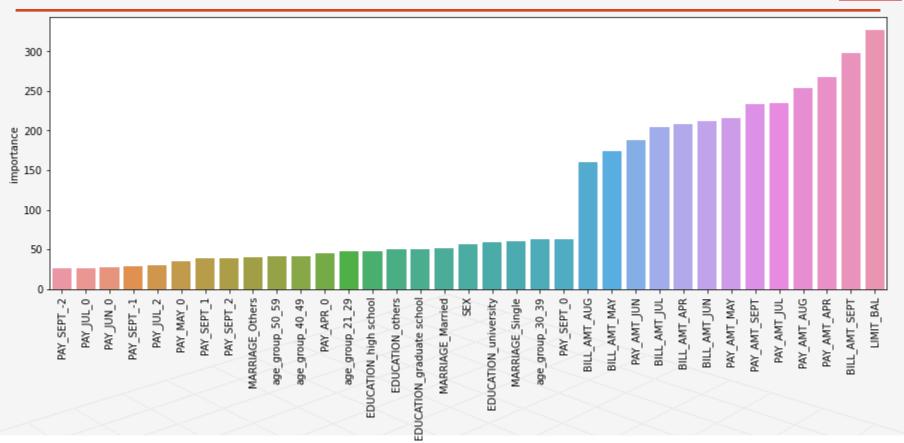
Parameters:

- max_depth=7n_estimators:
 - 150

Classification Report											
рі	recision	recall	f1-score								
0	0.84	0.94	0.88								
1	0.93	0.82	0.87								
accuracy			0.88								

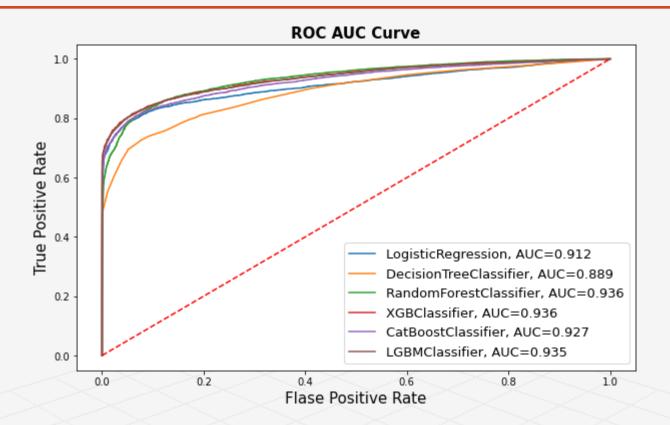
Random Forest Classifier





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 ...

