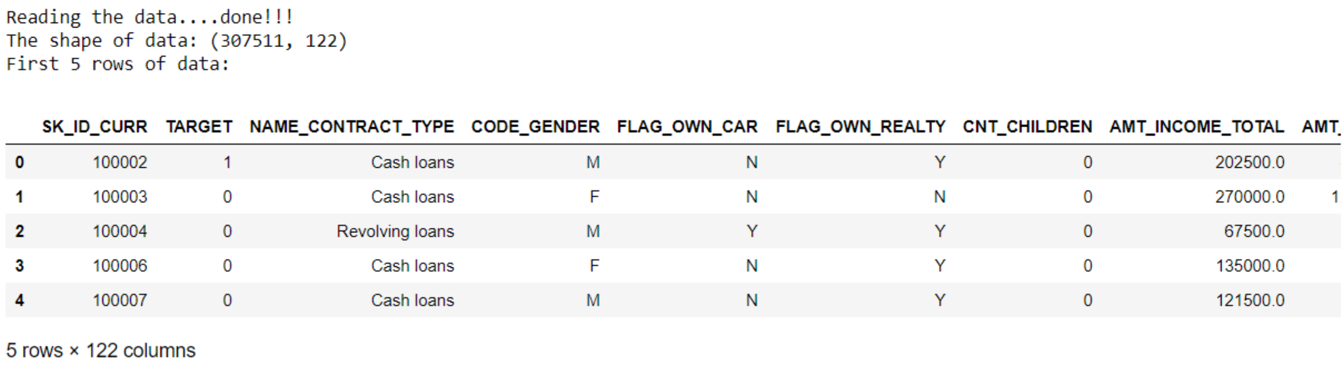
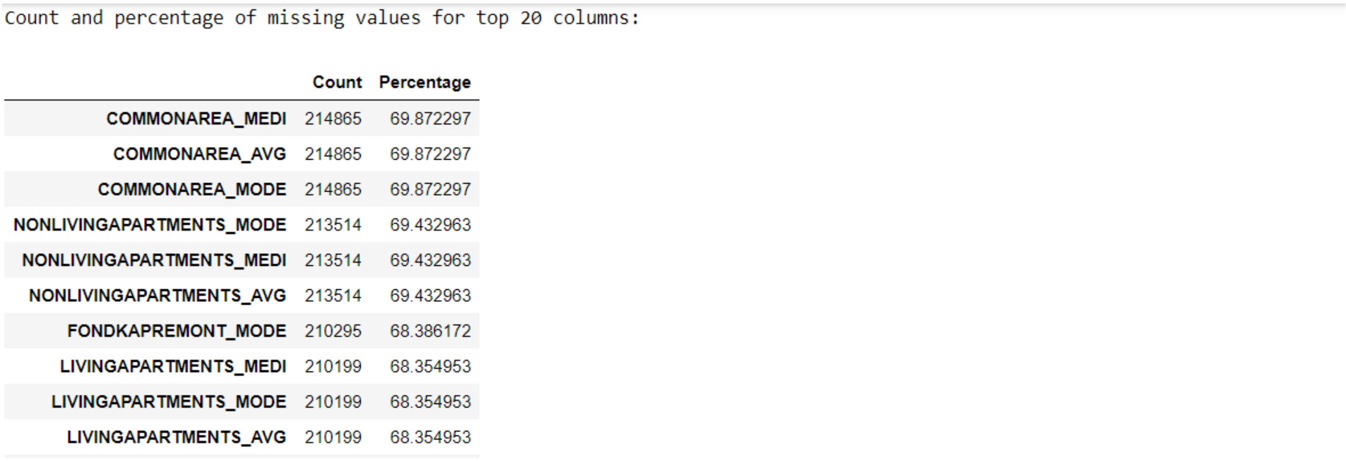
Project 1 : House Loan Data
Screen Shots



```
Reading the data....done!!!
The shape of data: (307511, 122)
First 5 rows of data:
```

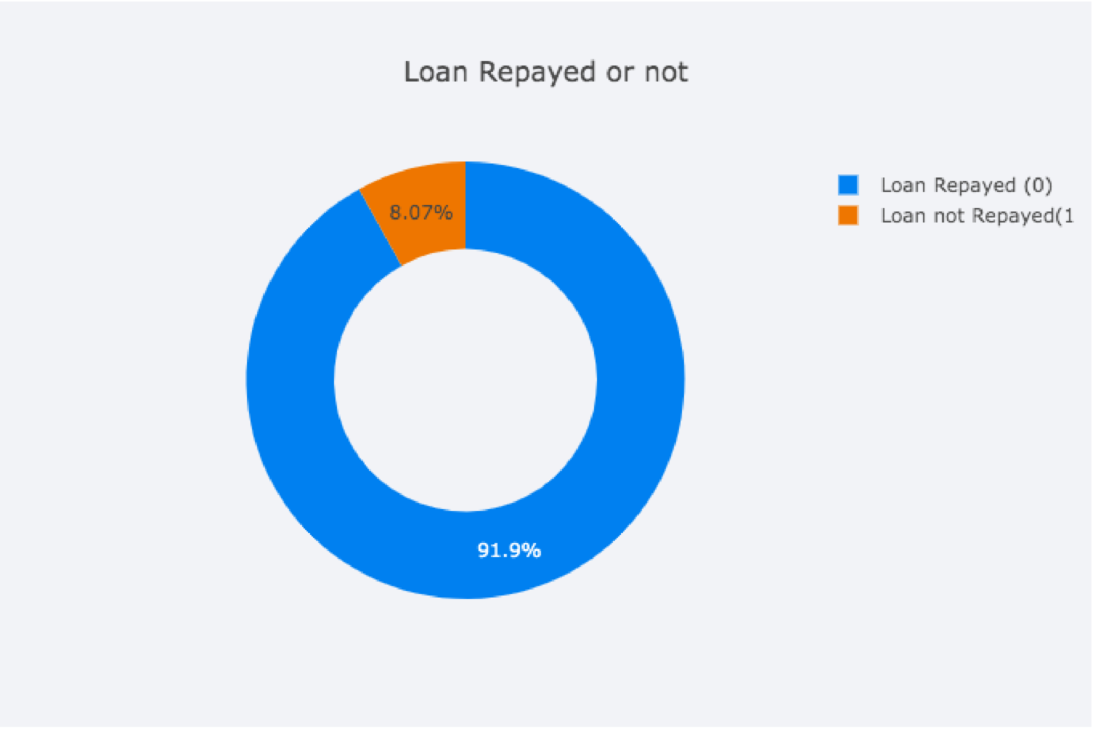
	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_
0	100002	1	Cash loans	M	N	Y	0	202500.0	
1	100003	0	Cash loans	F	N	N	0	270000.0	1
2	100004	0	Revolving loans	M	Y	Y	0	67500.0	
3	100006	0	Cash loans	F	N	Y	0	135000.0	
4	100007	0	Cash loans	M	N	Y	0	121500.0	

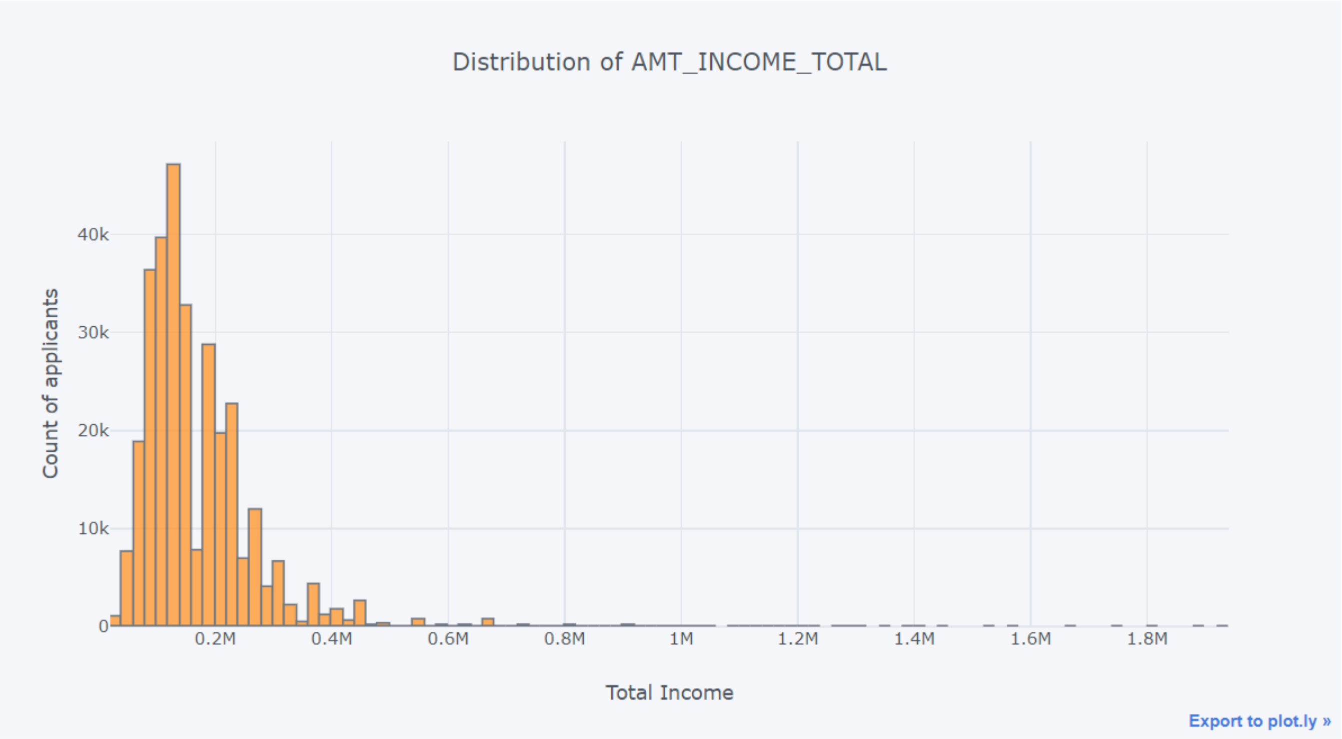
5 rows × 122 columns

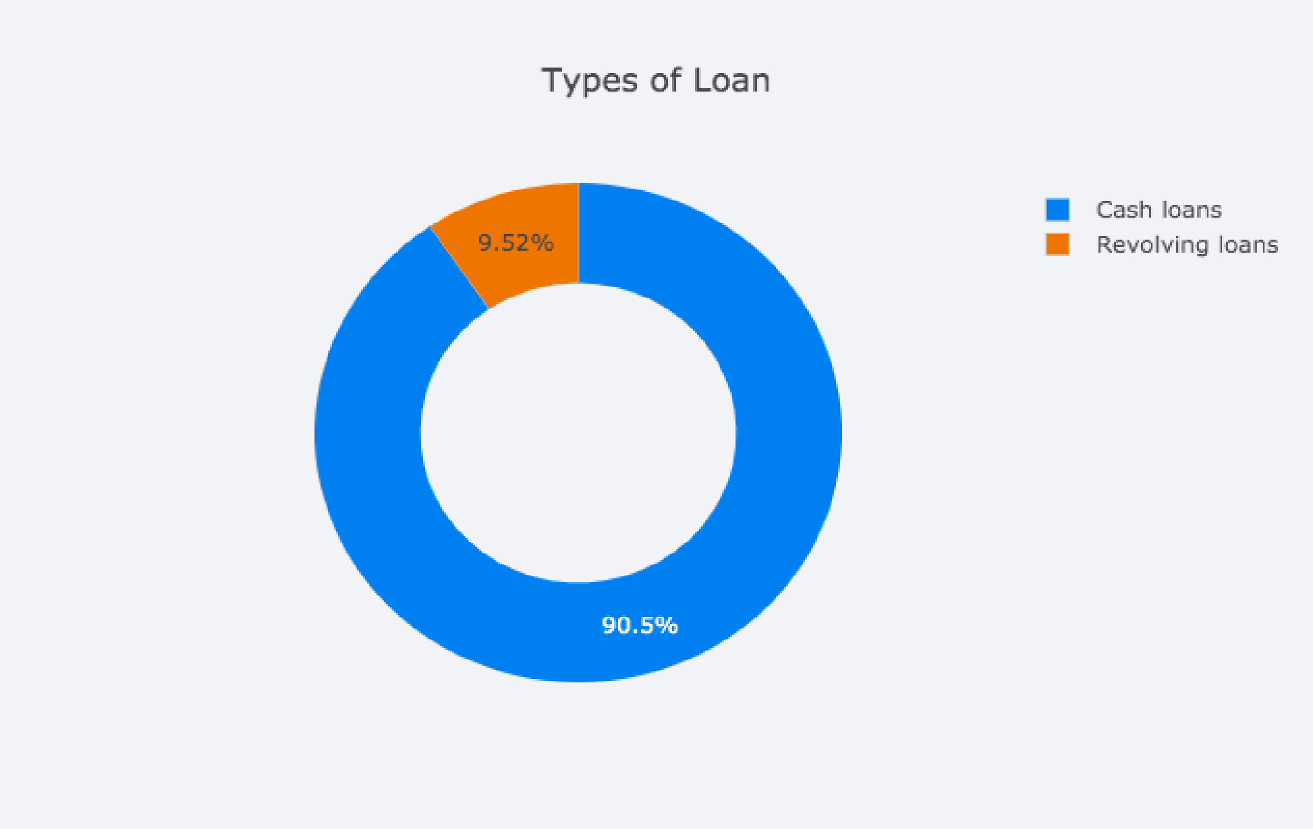


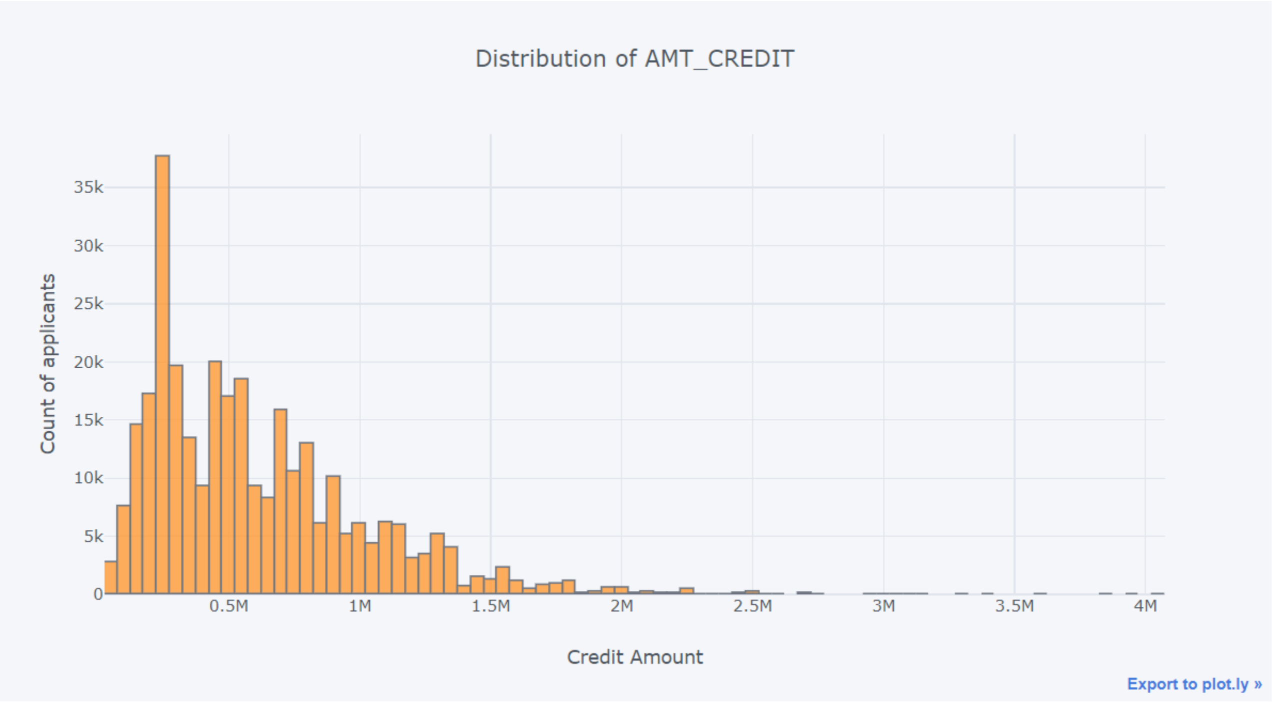
```
Count and percentage of missing values for top 20 columns:
```

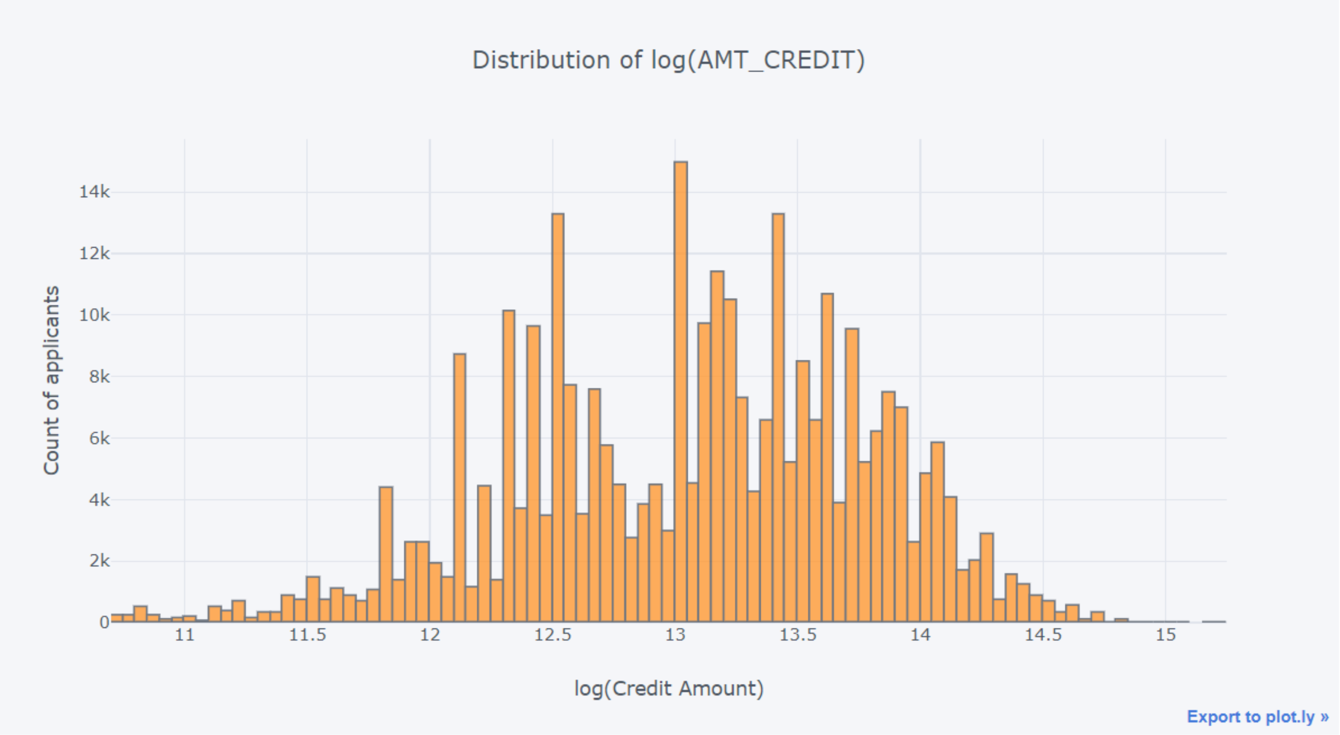
	Count	Percentage
COMMONAREA_MEDI	214865	69.872297
COMMONAREA_AVG	214865	69.872297
COMMONAREA_MODE	214865	69.872297
NONLIVINGAPARTMENTS_MODE	213514	69.432963
NONLIVINGAPARTMENTS_MEDI	213514	69.432963
NONLIVINGAPARTMENTS_AVG	213514	69.432963
FONDKAPREMONT_MODE	210295	68.386172
LIVINGAPARTMENTS_MEDI	210199	68.354953
LIVINGAPARTMENTS_MODE	210199	68.354953
LIVINGAPARTMENTS_AVG	210199	68.354953

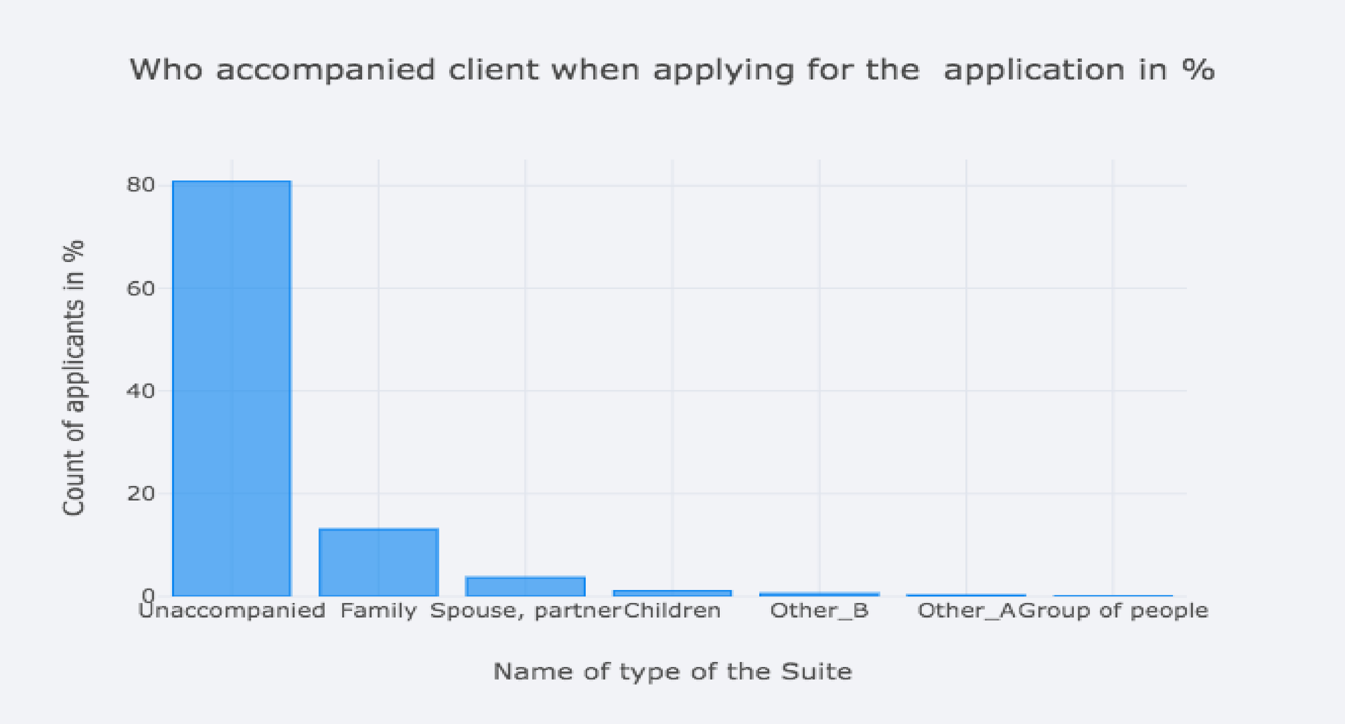


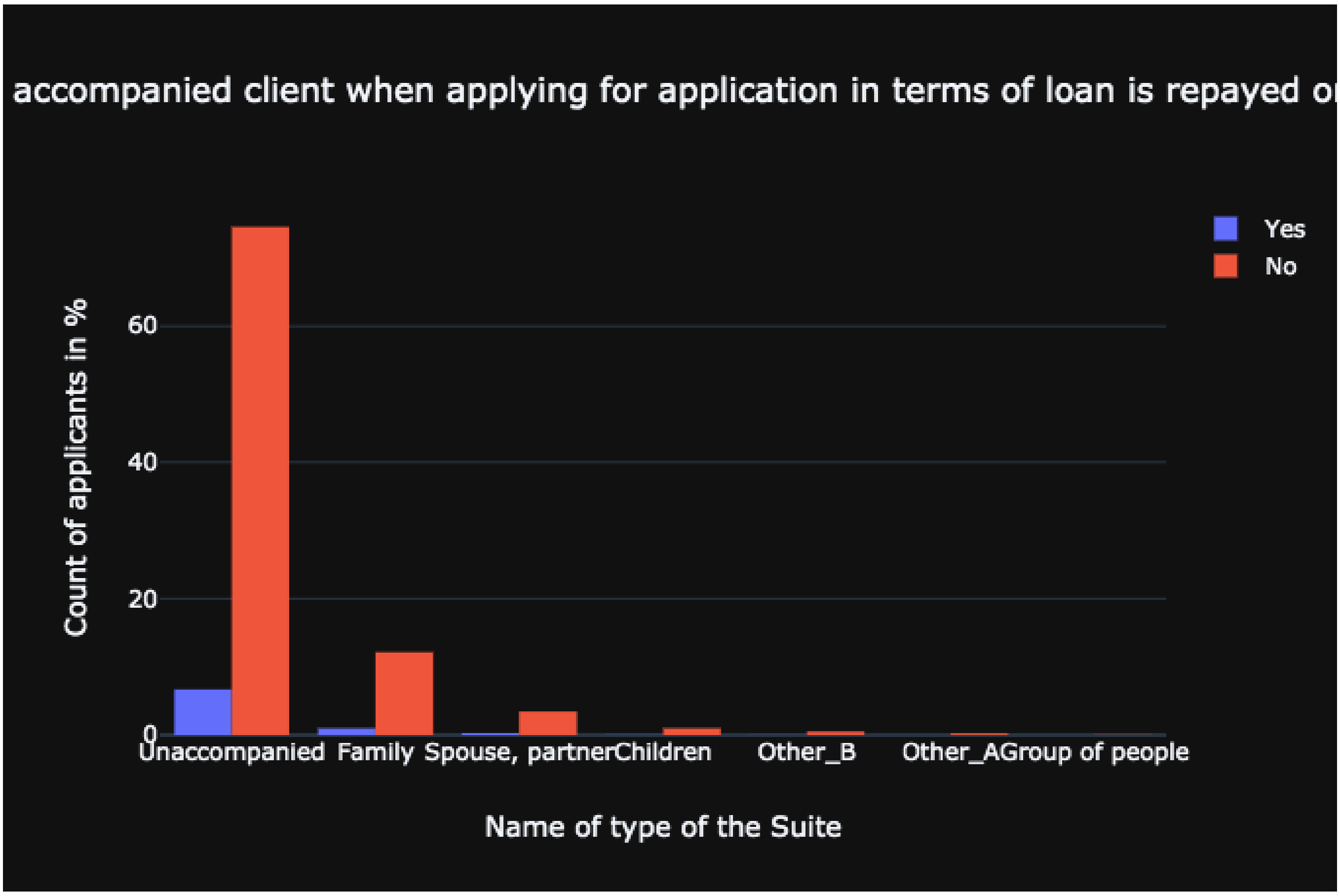

Distribution of AMT_INCOME_TOTAL
Count of applicants
0
10k
20k
30k
40k
0.2M
0.4M
0.6M
0.8M
1M
1.2M
1.4M
1.6M
1.8M
Total Income
Export to plot.ly »

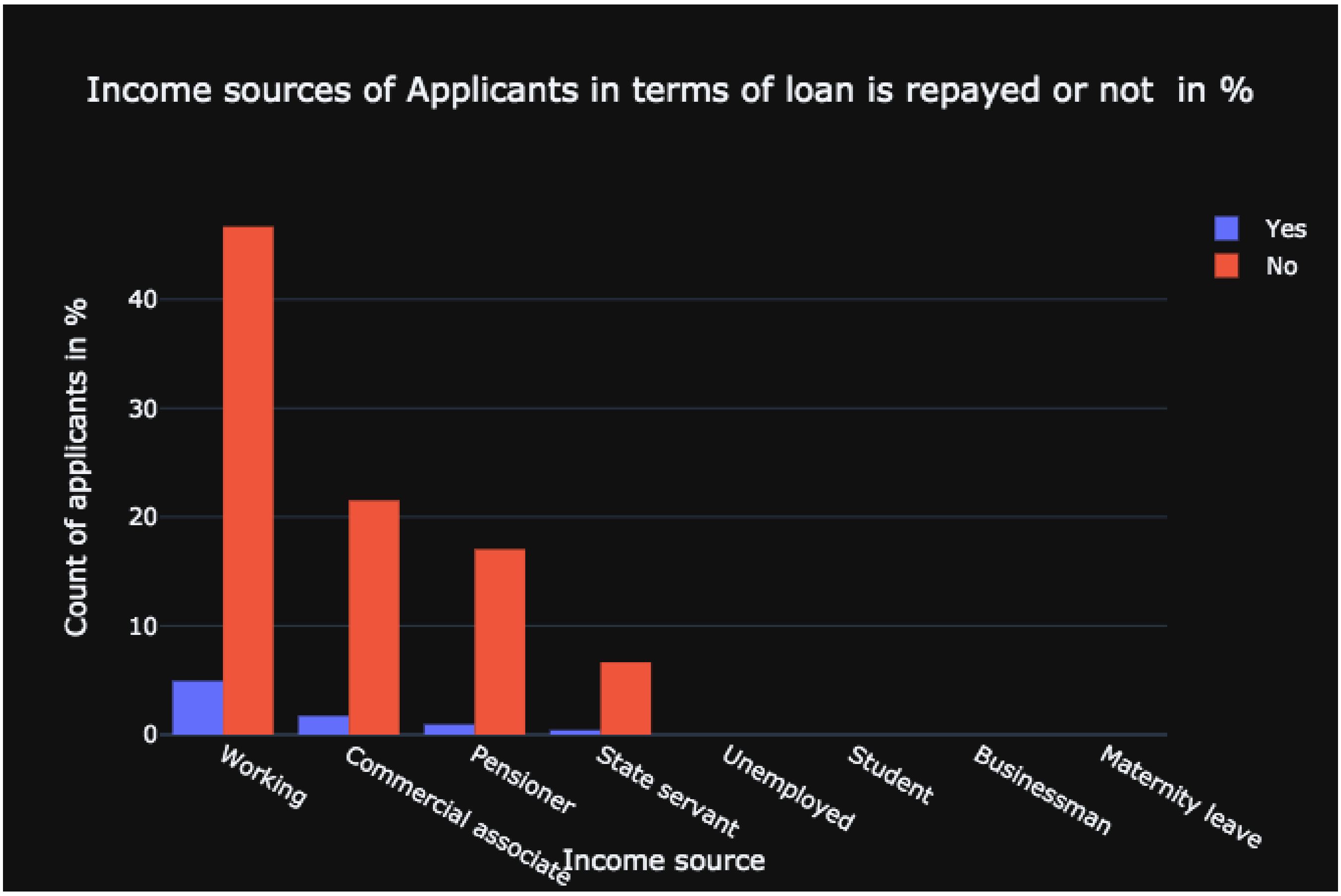

Types of Loan
9.52%
90.5%
Cash loans
Revolving loans

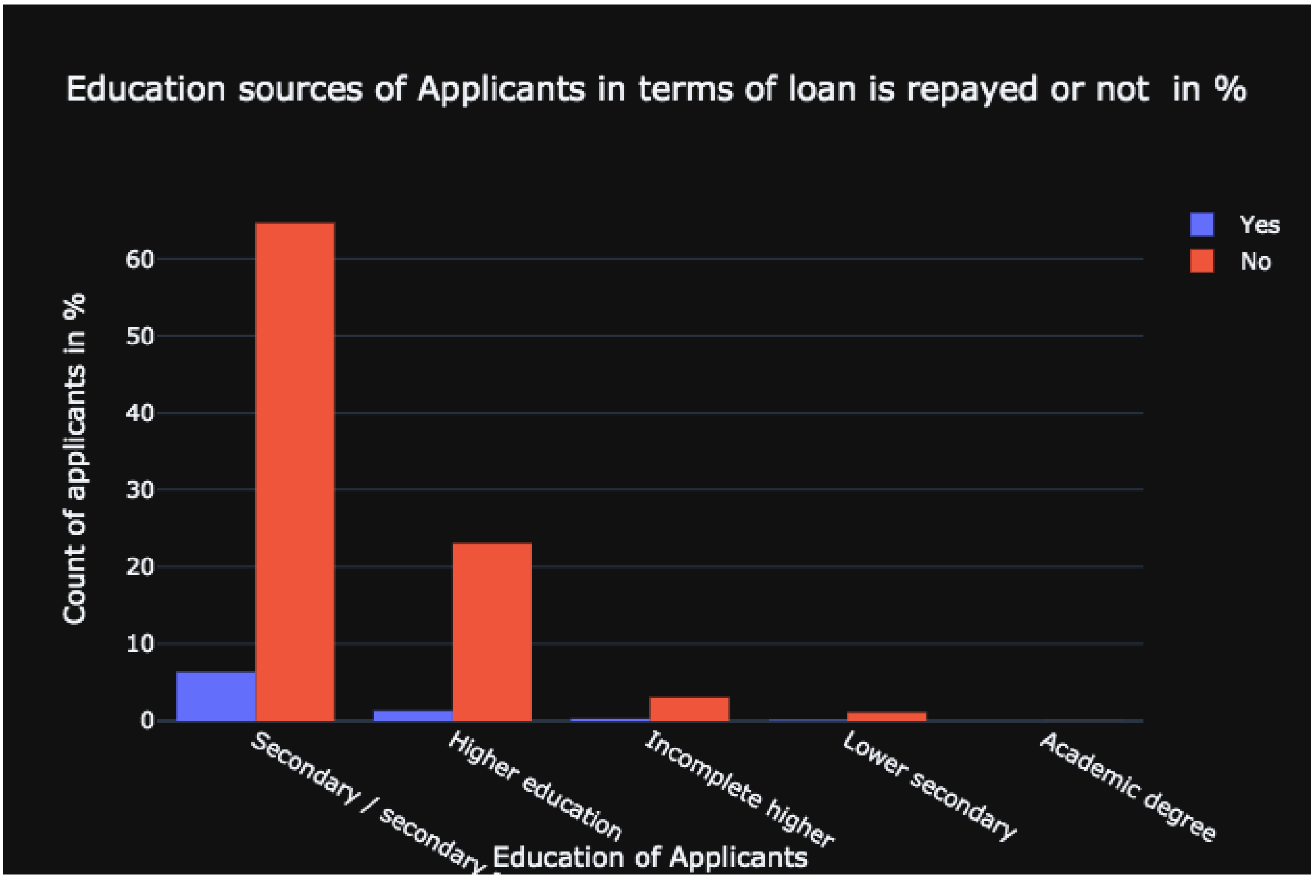

Distribution of AMT_CREDIT
Count of applicants
35k
30k
25k
20k
15k
10k
5k
0
0.5M
1M
1.5M
2M
2.5M
3M
3.5M
4M
Credit Amount
Export to plot.ly »

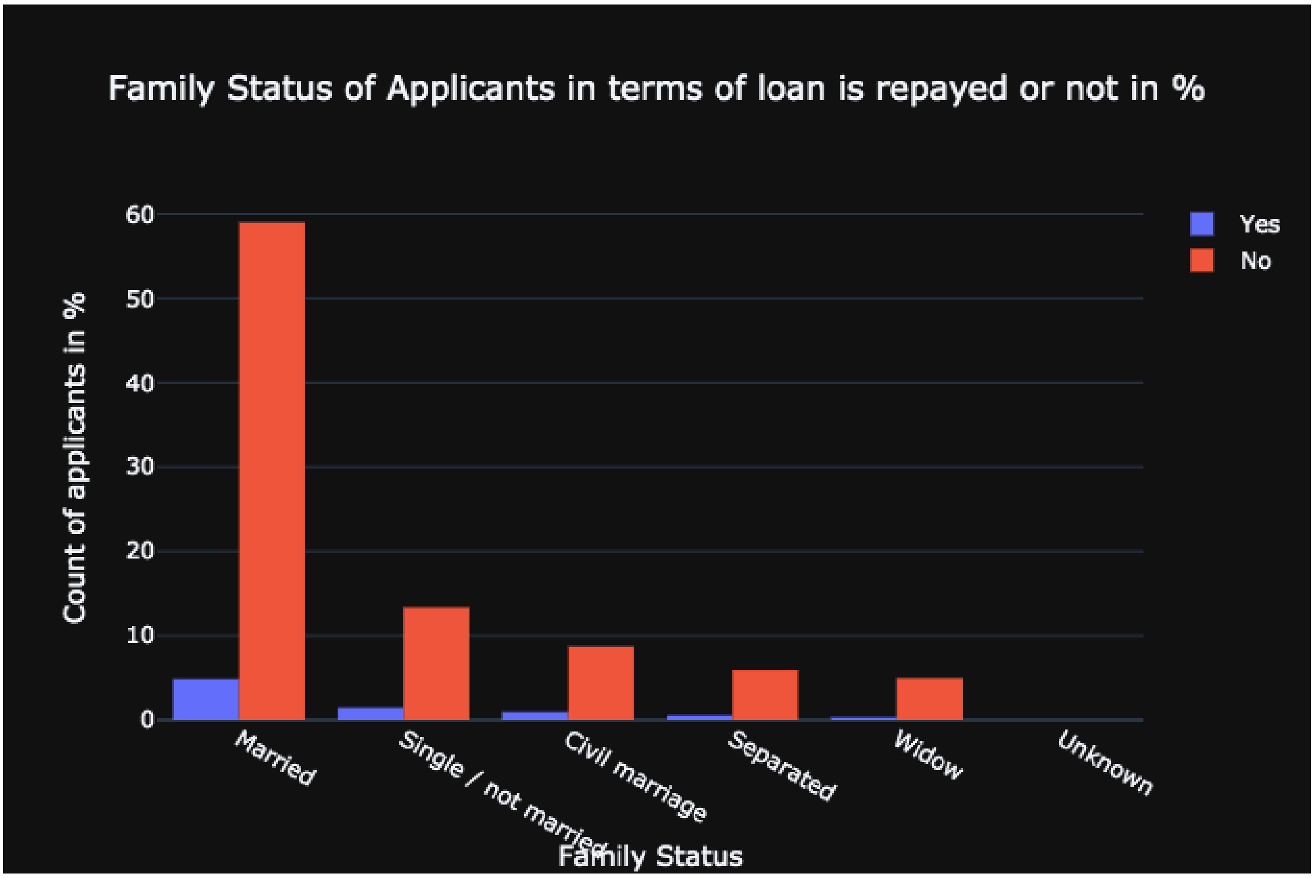

Distribution of log(AMT_CREDIT)
Count of applicants
14k
12k
10k
8k
6k
4k
2k
0
11
11.5
12
12.5
13
13.5
14
14.5
15
log(Credit Amount)
Export to plot.ly »

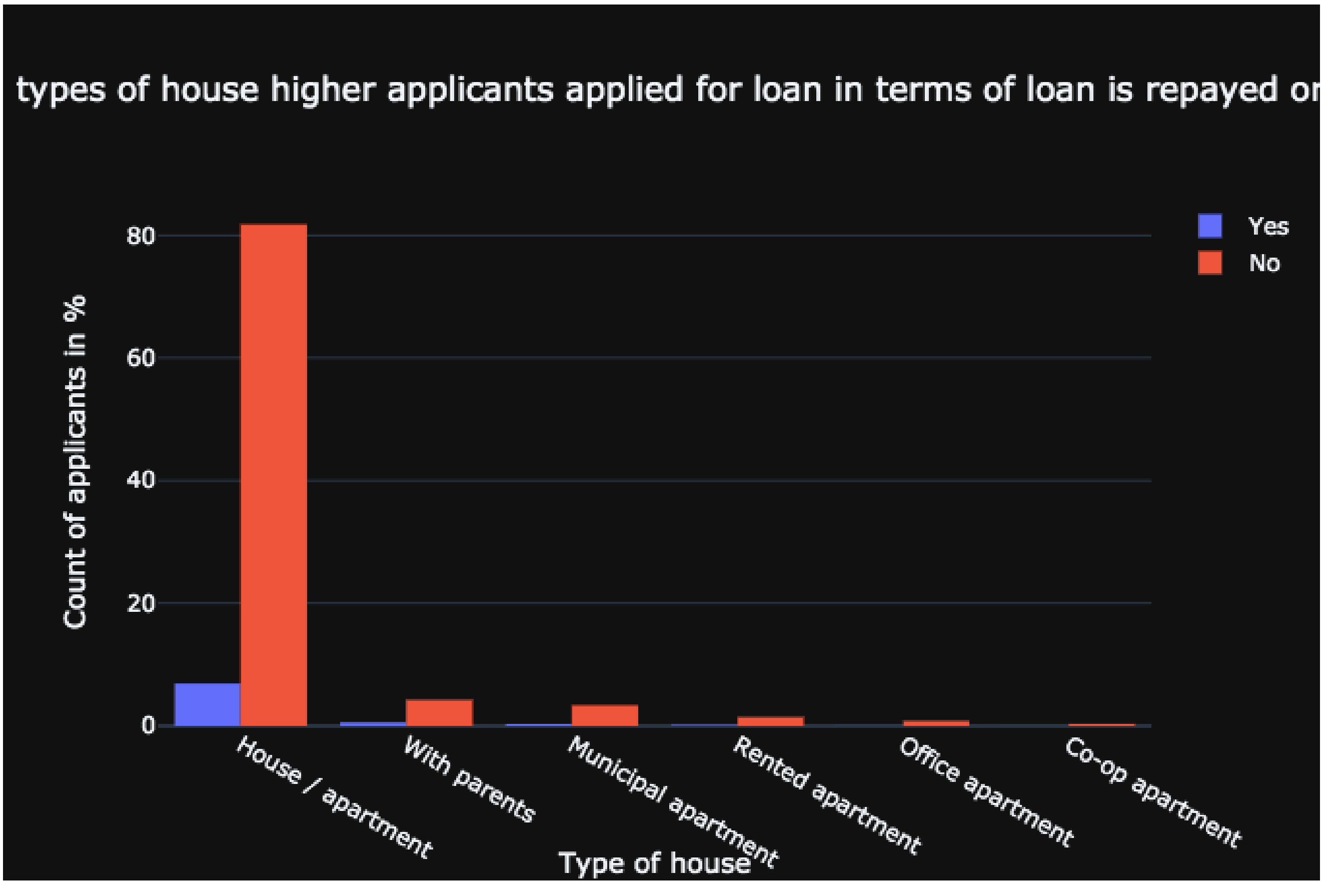

Who accompanied client when applying for the application in %
80
60
40
20
0
Count of applicants in %
Unaccompanied
Family
Spouse, partner
Children
Other_B
Other_A
Group of people
Name of type of the Suite

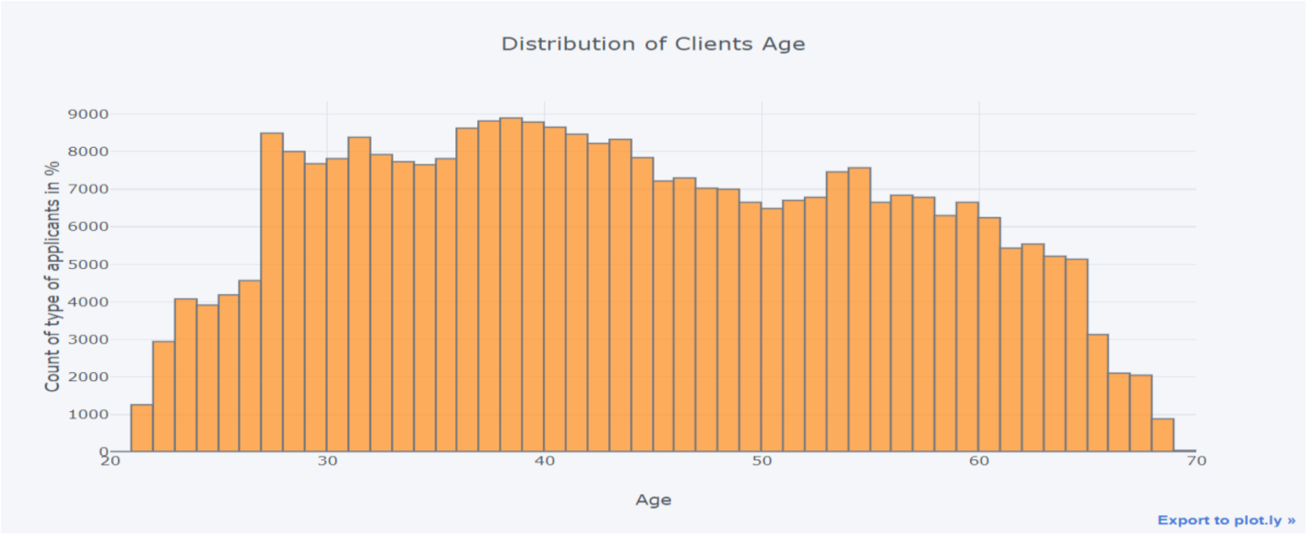

accompanied client when applying for application in terms of loan is repayed o
Yes
No
60
40
20
0
Count of applicants in %
Unaccompanied
Family
Spouse, partner
Children
Other_B
Other_A
Group of people
Name of type of the Suite

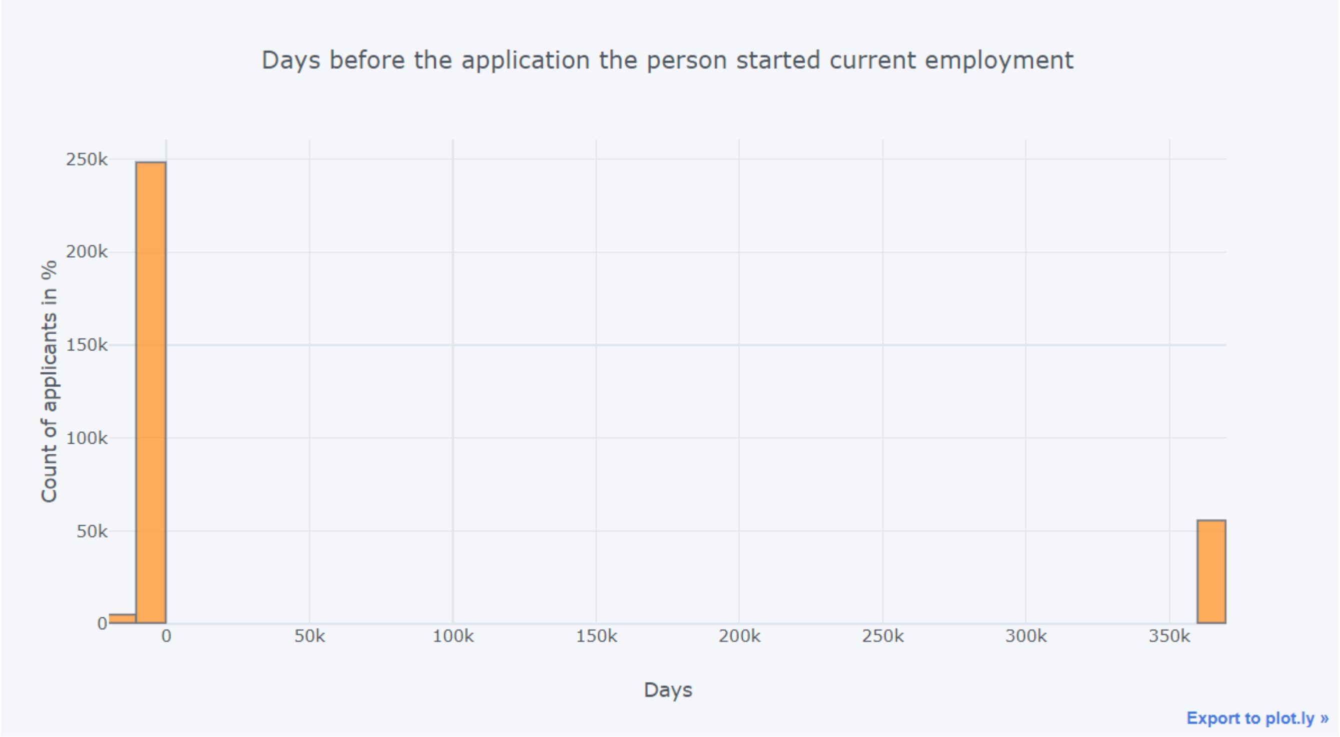

Income sources of Applicants in terms of loan is repayed or not in %
Yes
No
Count of applicants in %
0
10
20
30
40
Working
Commercial associate
Pensioner
State servant
Unemployed
Student
Businessman
Maternity leave
Income source

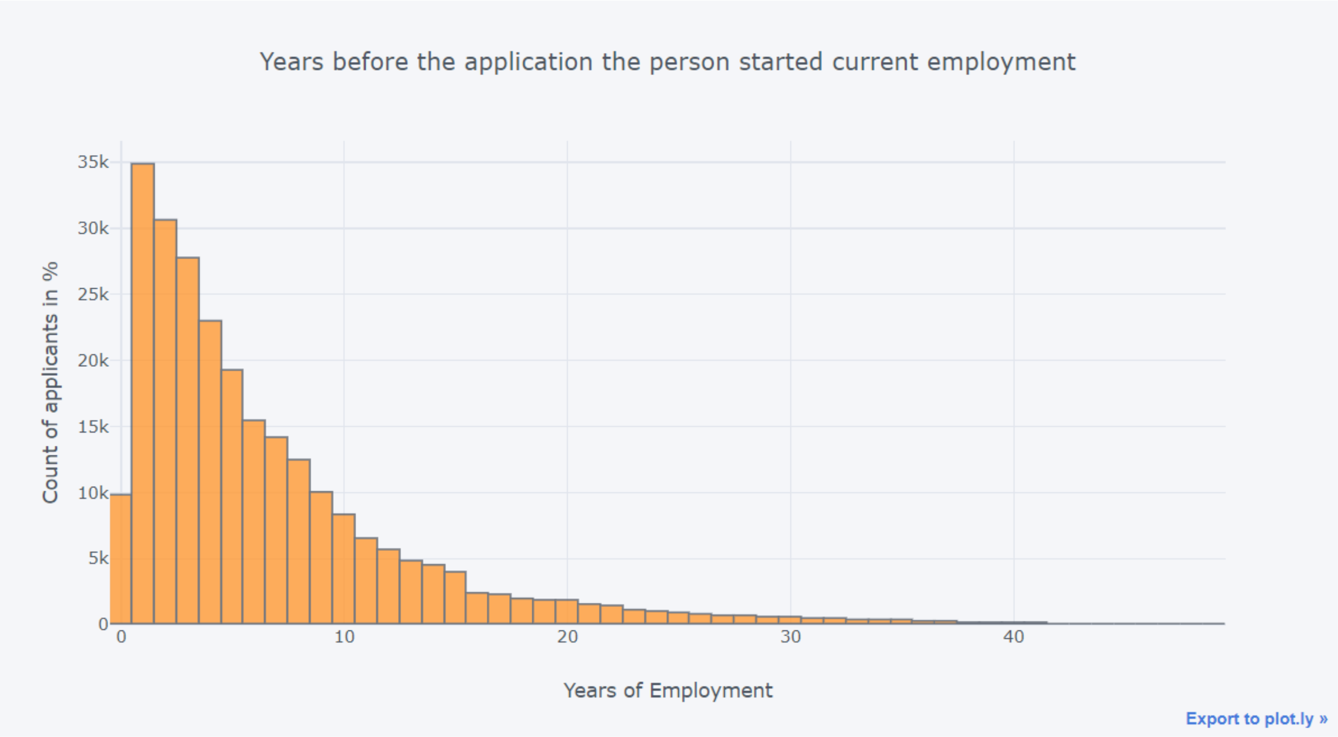

Education sources of Applicants in terms of loan is repayed or not in %
Yes
No
Count of applicants in %
0
10
20
30
40
50
60
Secondary / secondary
Higher education
Incomplete higher
Lower secondary
Academic degree
Education of Applicants

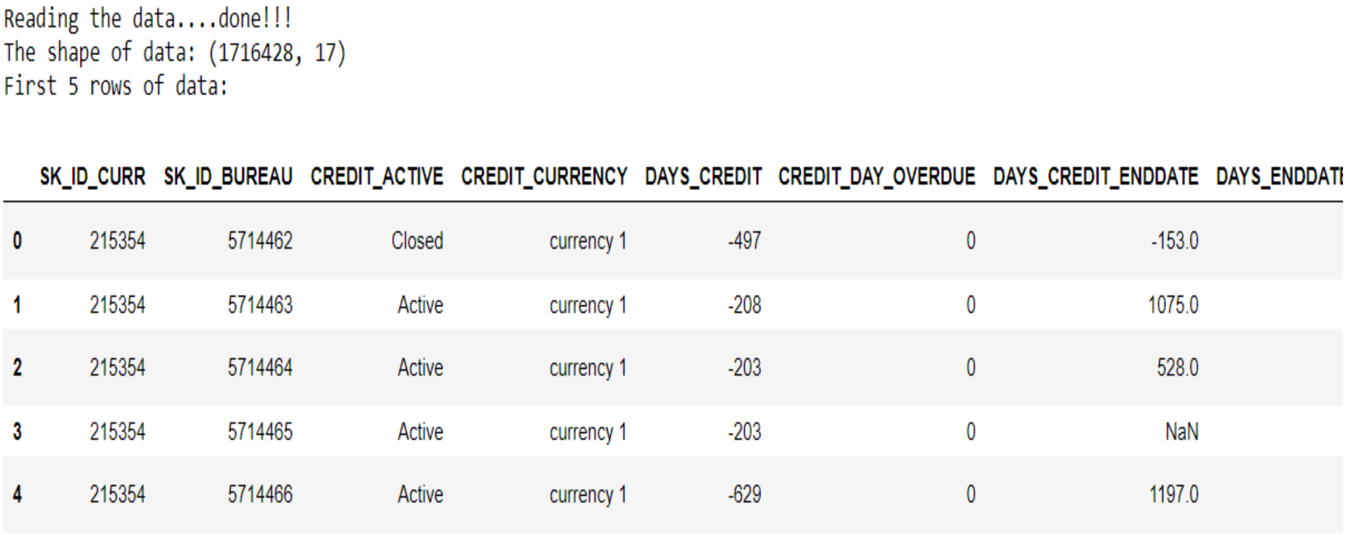

Family Status of Applicants in terms of loan is repayed or not in %
Yes
No
Count of applicants in %
60
50
40
30
20
10
0
Married
Single / not married
Civil marriage
Separated
Widow
Unknown
Family Status


types of house higher applicants applied for loan in terms of loan is repayed or
Yes
No
Count of applicants in %
80
60
40
20
0
House / apartment
With parents
Municipal apartment
Rented apartment
Office apartment
Co-op apartment
Type of house


Distribution of Clients Age
Count of type of applicants in %
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
20
30
40
50
60
70
Age
Export to plot.ly »

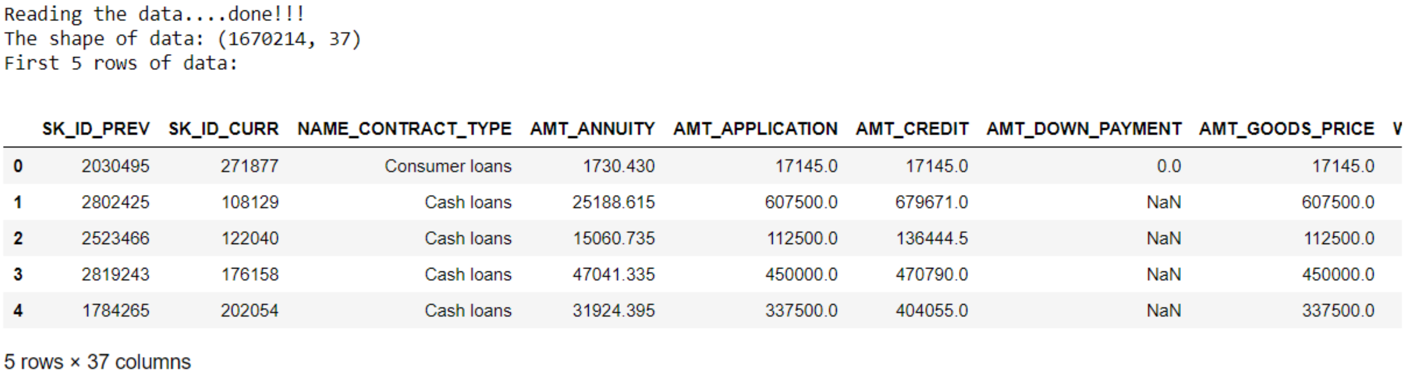

Days before the application the person started current employment
Count of applicants in %
0
50k
100k
150k
200k
250k
0
50k
100k
150k
200k
250k
300k
350k
Days
Export to plot.ly »





```
Reading the data....done!!!
The shape of data: (1716428, 17)
First 5 rows of data:
```

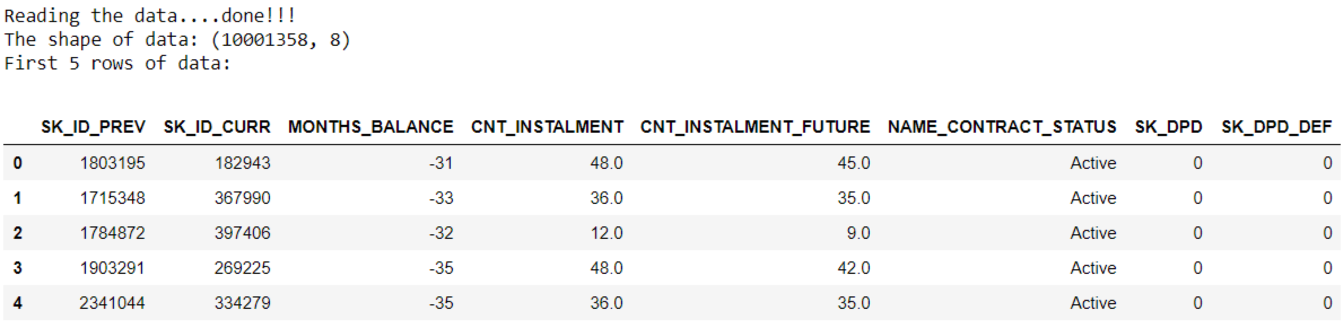
	SK_ID_CURR	SK_ID_BUREAU	CREDIT_ACTIVE	CREDIT_CURRENCY	DAYS_CREDIT	CREDIT_DAY_OVERDUE	DAYS_CREDIT_ENDDATE	DAYS_ENDDATI
0	215354	5714462	Closed	currency 1	-497	0	-153.0	
1	215354	5714463	Active	currency 1	-208	0	1075.0	
2	215354	5714464	Active	currency 1	-203	0	528.0	
3	215354	5714465	Active	currency 1	-203	0	NaN	
4	215354	5714466	Active	currency 1	-629	0	1197.0	



```
Reading the data....done!!!
The shape of data: (1670214, 37)
First 5 rows of data:
```

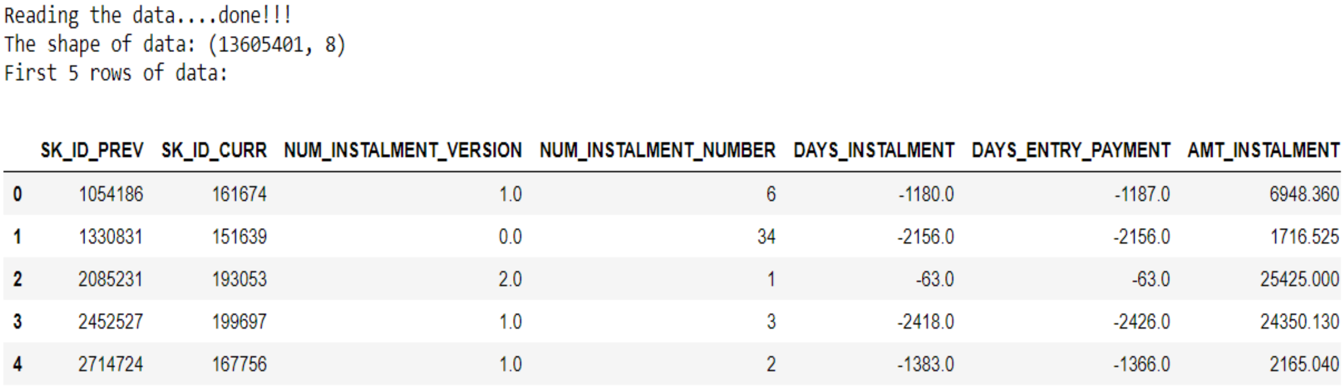
	SK_ID_PREV	SK_ID_CURR	NAME_CONTRACT_TYPE	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_DOWN_PAYMENT	AMT_GOODS_PRICE	V
0	2030495	271877	Consumer loans	1730.430	17145.0	17145.0	0.0	17145.0	
1	2802425	108129	Cash loans	25188.615	607500.0	679671.0	NaN	607500.0	
2	2523466	122040	Cash loans	15060.735	112500.0	136444.5	NaN	112500.0	
3	2819243	176158	Cash loans	47041.335	450000.0	470790.0	NaN	450000.0	
4	1784265	202054	Cash loans	31924.395	337500.0	404055.0	NaN	337500.0	

5 rows × 37 columns



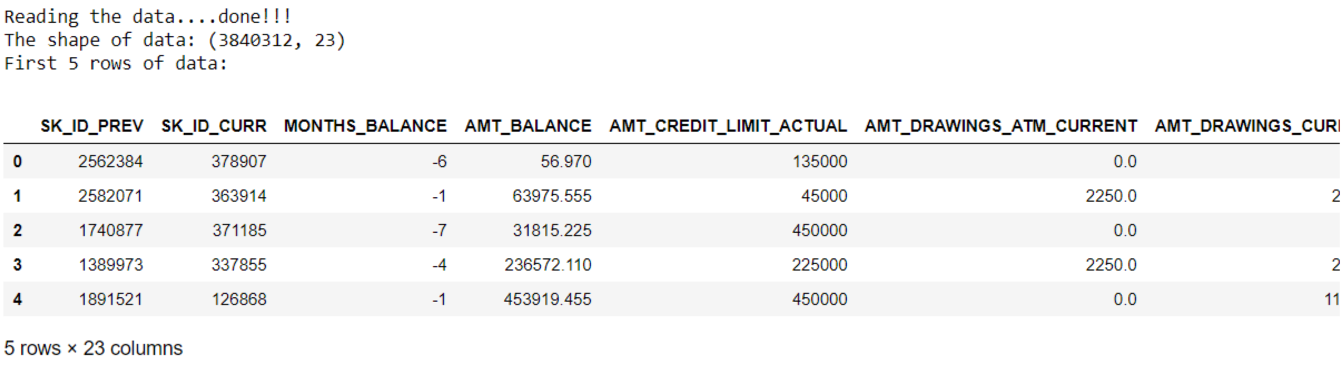
```
Reading the data....done!!!
The shape of data: (10001358, 8)
First 5 rows of data:
```

	SK_ID_PREV	SK_ID_CURR	MONTHS_BALANCE	CNT_INSTALMENT	CNT_INSTALMENT_FUTURE	NAME_CONTRACT_STATUS	SK_DPD	SK_DPD_DEF
0	1803195	182943	-31	48.0	45.0	Active	0	0
1	1715348	367990	-33	36.0	35.0	Active	0	0
2	1784872	397406	-32	12.0	9.0	Active	0	0
3	1903291	269225	-35	48.0	42.0	Active	0	0
4	2341044	334279	-35	36.0	35.0	Active	0	0



```
Reading the data....done!!!
The shape of data: (13605401, 8)
First 5 rows of data:
```

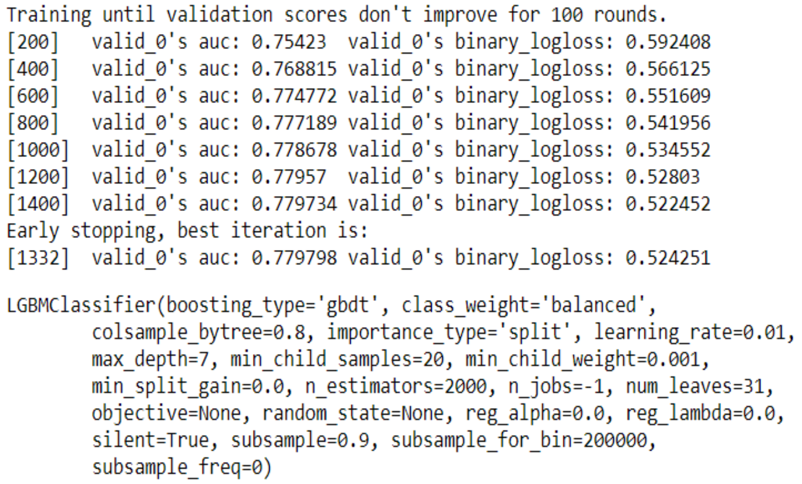
	SK_ID_PREV	SK_ID_CURR	NUM_INSTALMENT_VERSION	NUM_INSTALMENT_NUMBER	DAYS_INSTALMENT	DAYS_ENTRY_PAYMENT	AMT_INSTALMENT
0	1054186	161674	1.0	6	-1180.0	-1187.0	6948.360
1	1330831	151639	0.0	34	-2156.0	-2156.0	1716.525
2	2085231	193053	2.0	1	-63.0	-63.0	25425.000
3	2452527	199697	1.0	3	-2418.0	-2426.0	24350.130
4	2714724	167756	1.0	2	-1383.0	-1366.0	2165.040



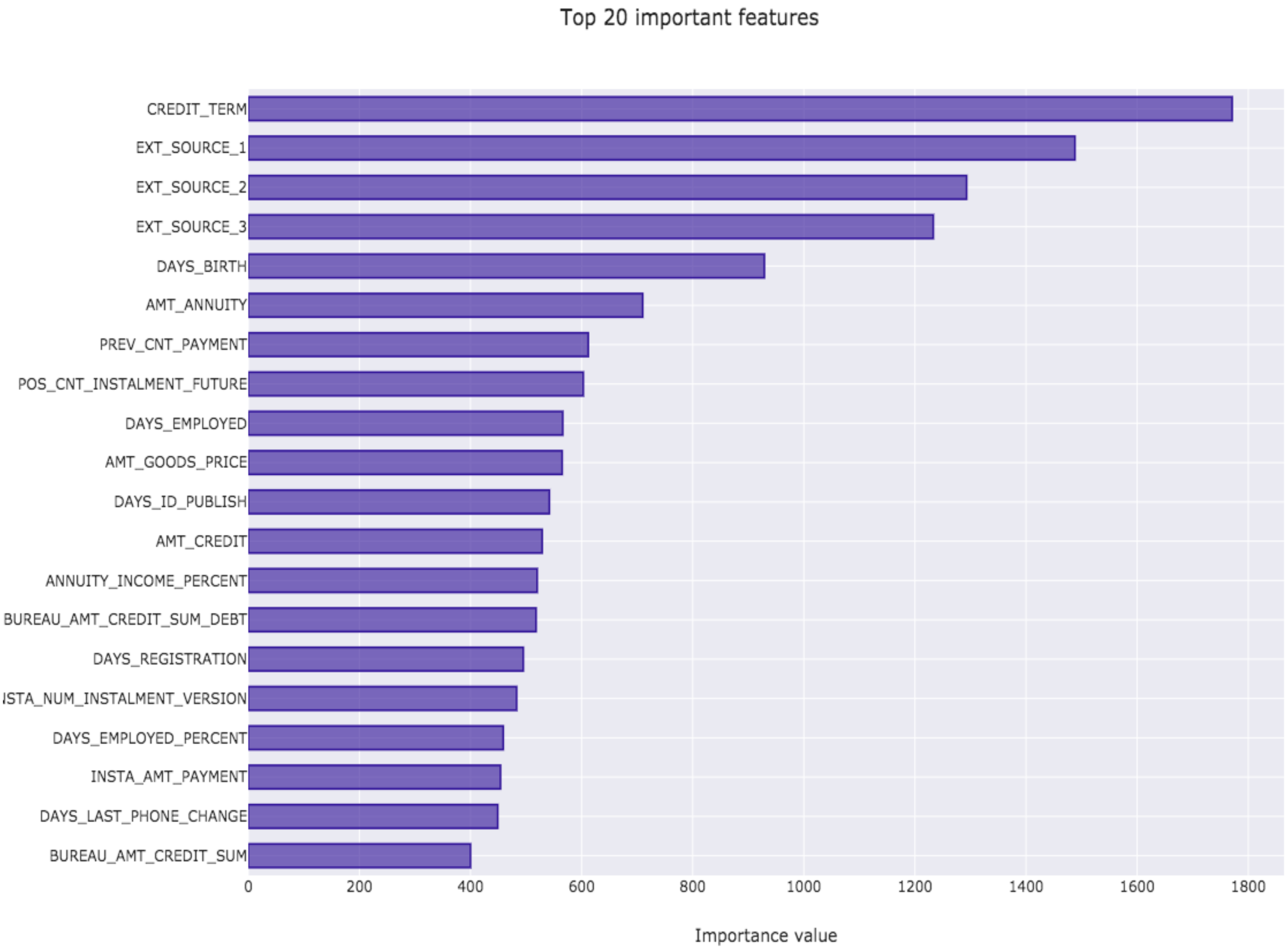
```
Reading the data....done!!!
The shape of data: (3840312, 23)
First 5 rows of data:
```

	SK_ID_PREV	SK_ID_CURR	MONTHS_BALANCE	AMT_BALANCE	AMT_CREDIT_LIMIT_ACTUAL	AMT_DRAWINGS_ATM_CURRENT	AMT_DRAWINGS_CURI
0	2562384	378907	-6	56.970	135000	0.0	
1	2582071	363914	-1	63975.555	45000	2250.0	2
2	1740877	371185	-7	31815.225	450000	0.0	
3	1389973	337855	-4	236572.110	225000	2250.0	2
4	1891521	126868	-1	453919.455	450000	0.0	11

5 rows × 23 columns



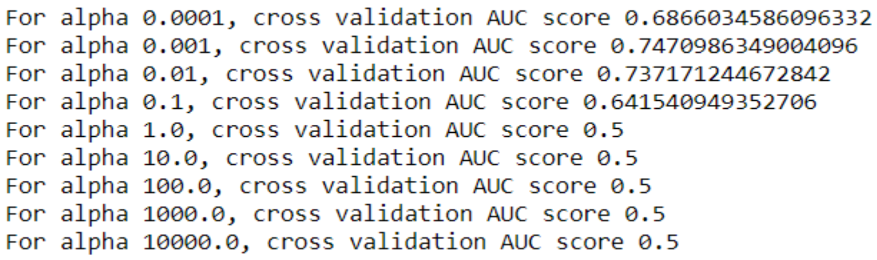
```
Training until validation scores don't improve for 100 rounds.
[200]	valid_0's auc: 0.75423	valid_0's binary_logloss: 0.592408
[400]	valid_0's auc: 0.768815	valid_0's binary_logloss: 0.566125
[600]	valid_0's auc: 0.774772	valid_0's binary_logloss: 0.551609
[800]	valid_0's auc: 0.777189	valid_0's binary_logloss: 0.541956
[1000]	valid_0's auc: 0.778678	valid_0's binary_logloss: 0.534552
[1200]	valid_0's auc: 0.77957	valid_0's binary_logloss: 0.52803
[1400]	valid_0's auc: 0.779734	valid_0's binary_logloss: 0.522452
Early stopping, best iteration is:
[1332]	valid_0's auc: 0.779798	valid_0's binary_logloss: 0.524251

LGBMClassifier(boosting_type='gbdt', class_weight='balanced',
        colsample_bytree=0.8, importance_type='split', learning_rate=0.01,
        max_depth=7, min_child_samples=20, min_child_weight=0.001,
        min_split_gain=0.0, n_estimators=2000, n_jobs=-1, num_leaves=31,
        objective=None, random_state=None, reg_alpha=0.0, reg_lambda=0.0,
        silent=True, subsample=0.9, subsample_for_bin=200000,
        subsample_freq=0)
```


Top 20 important features
CREDIT_TERM
EXT_SOURCE_1
EXT_SOURCE_2
EXT_SOURCE_3
DAYS_BIRTH
AMT_ANNUITY
PREV_CNT_PAYMENT
POS_CNT_INSTALMENT_FUTURE
DAYS_EMPLOYED
AMT_GOODS_PRICE
DAYS_ID_PUBLISH
AMT_CREDIT
ANNUITY_INCOME_PERCENT
BUREAU_AMT_CREDIT_SUM_DEBT
DAYS_REGISTRATION
ISTA_NUM_INSTALMENT_VERSION
DAYS_EMPLOYED_PERCENT
INSTA_AMT_PAYMENT
DAYS_LAST_PHONE_CHANGE
BUREAU_AMT_CREDIT_SUM
0
200
400
600
800
1000
1200
1400
1600
1800
Importance value

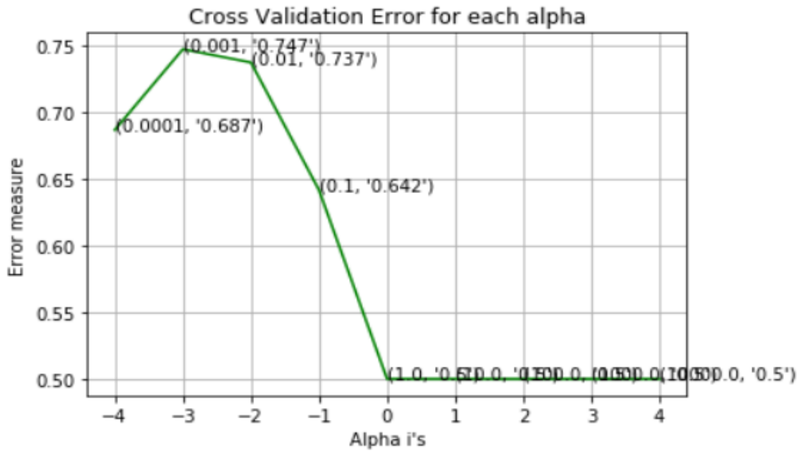
Machine Learning Models:

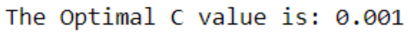
Logistic regression with selected features:

Logistic Regression finds a hyperplane which best seperates the given positive and negative data points.



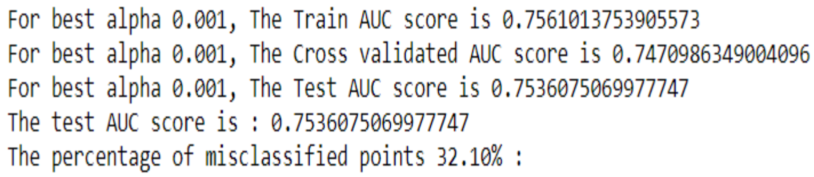
```
For alpha 0.0001, cross validation AUC score 0.6866034586096332
For alpha 0.001, cross validation AUC score 0.7470986349004096
For alpha 0.01, cross validation AUC score 0.737171244672842
For alpha 0.1, cross validation AUC score 0.641540949352706
For alpha 1.0, cross validation AUC score 0.5
For alpha 10.0, cross validation AUC score 0.5
For alpha 100.0, cross validation AUC score 0.5
For alpha 1000.0, cross validation AUC score 0.5
For alpha 10000.0, cross validation AUC score 0.5
```



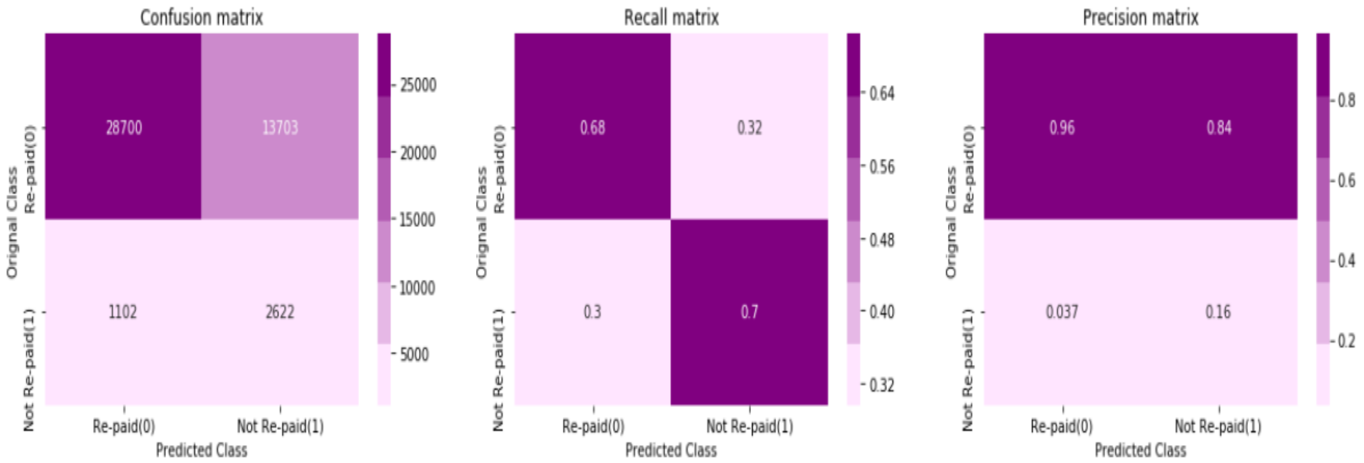


```
The Optimal C value is: 0.001
```

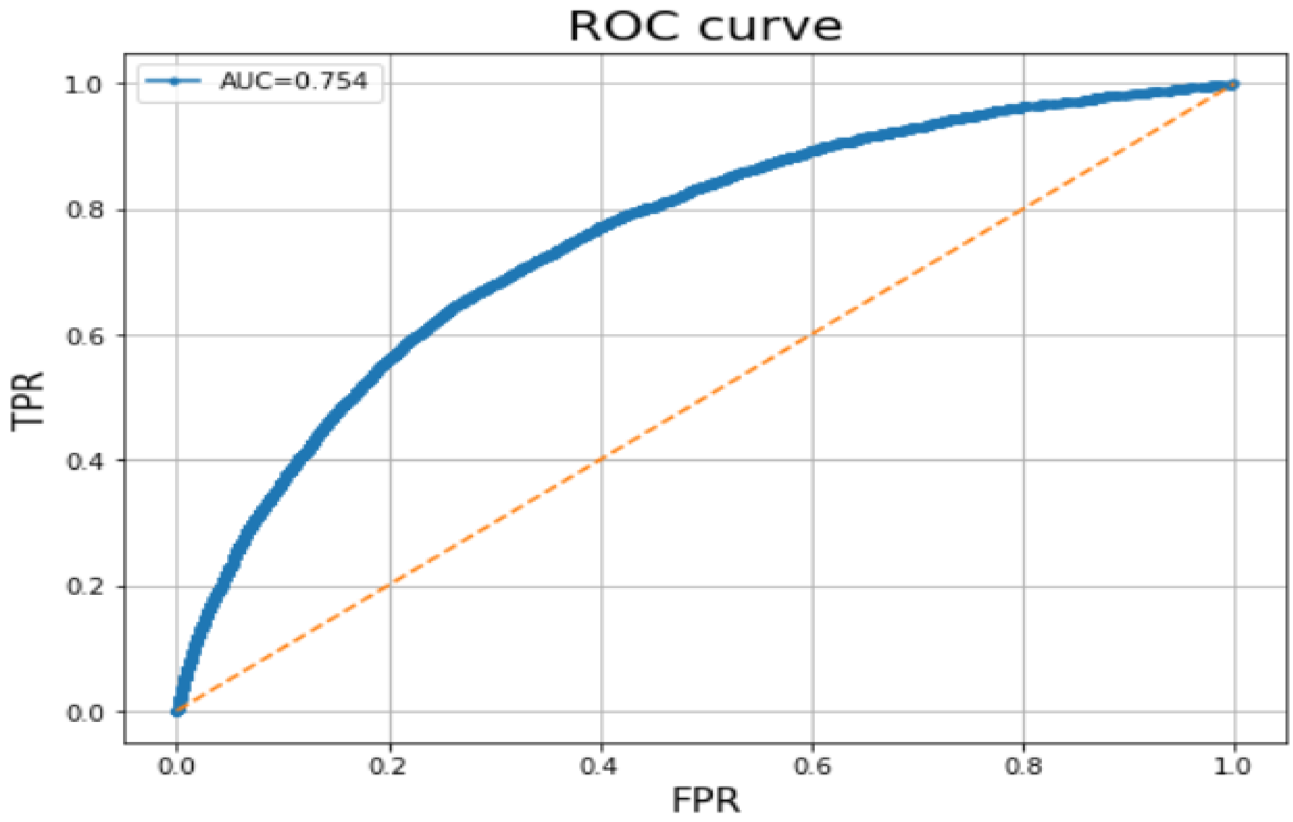
Cross validation results and plot for Logistic Regression model.



```
For best alpha 0.001, The Train AUC score is 0.7561013753905573
For best alpha 0.001, The Cross validated AUC score is 0.7470986349004096
For best alpha 0.001, The Test AUC score is 0.7536075069977747
The test AUC score is : 0.7536075069977747
The percentage of misclassified points 32.10% :
```

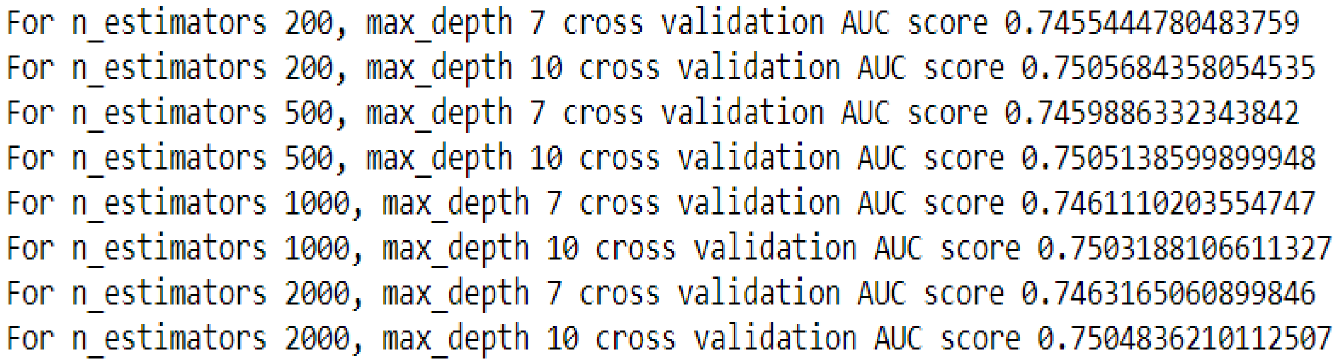


Logistic Regression model results



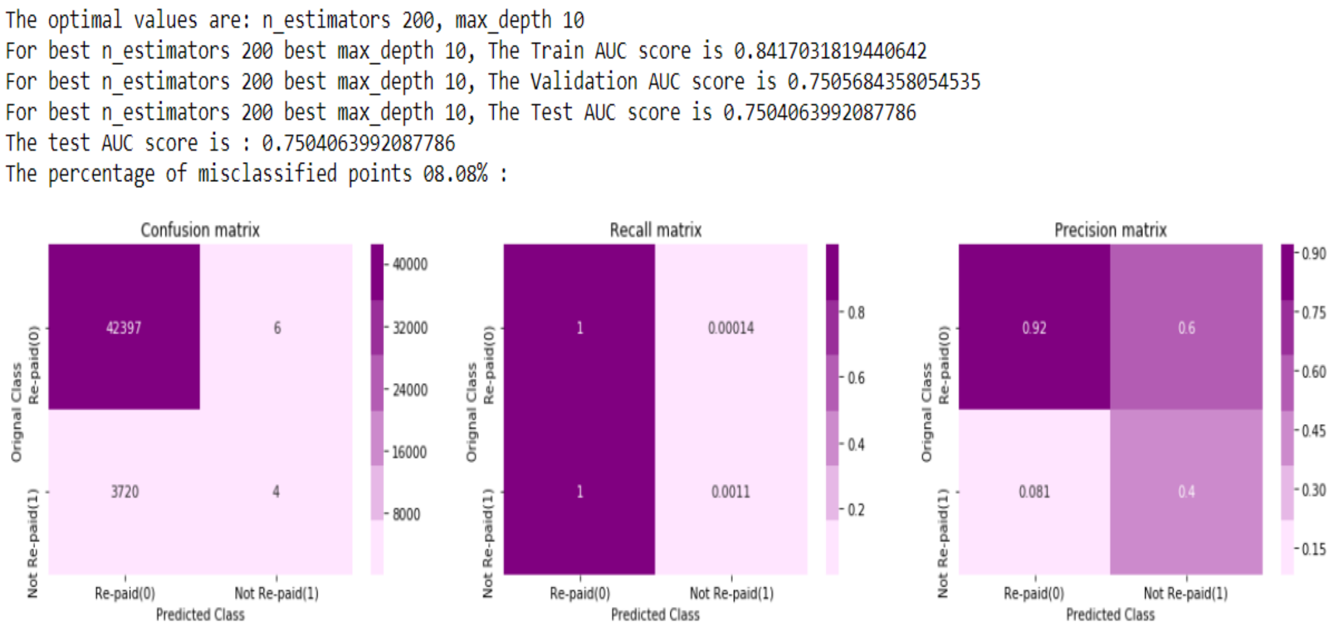
ROC curve for Logistic Regression model with AUC=0.754

Random Forest with selected features:



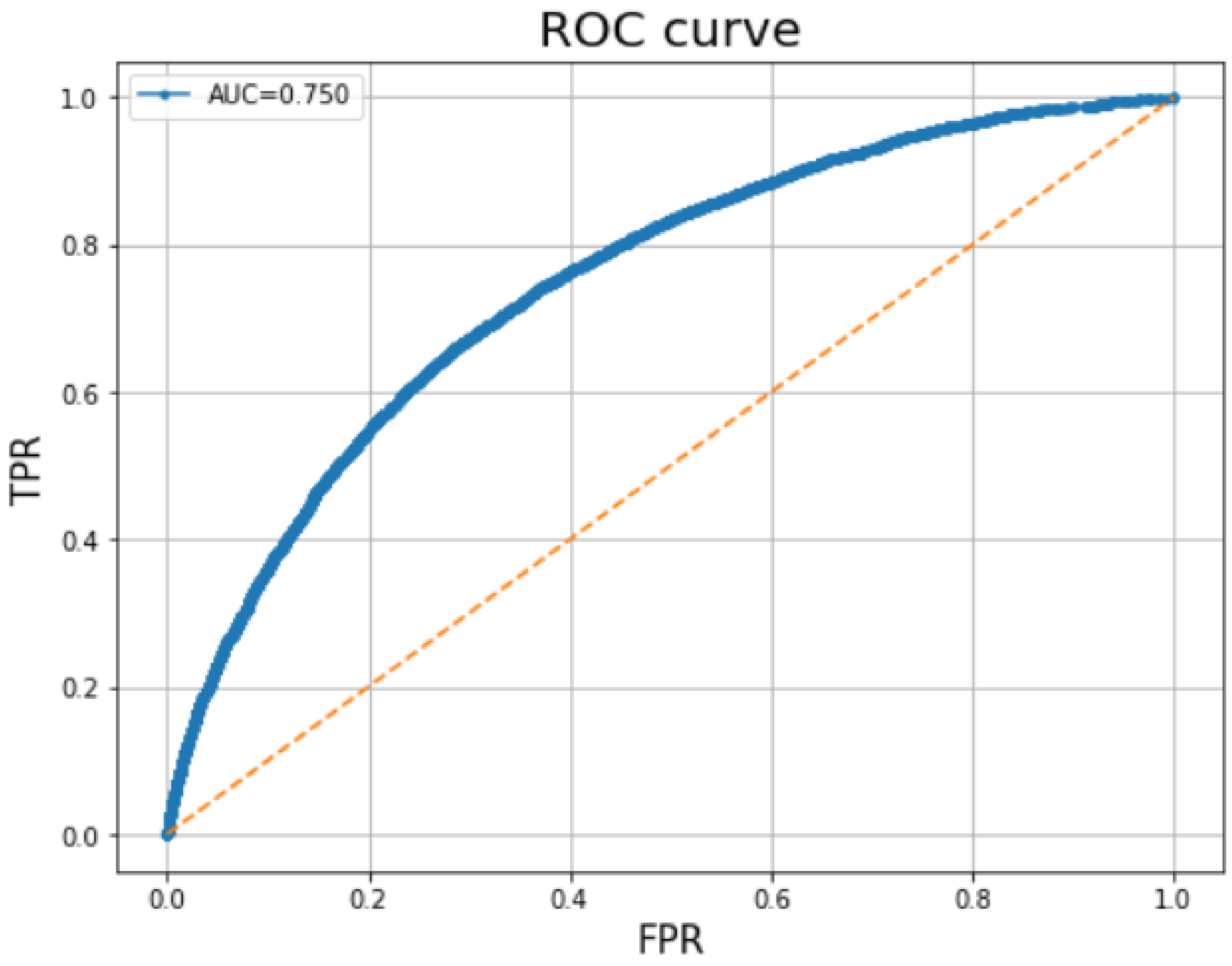
```
For n_estimators 200, max_depth 7 cross validation AUC score 0.7455444780483759
For n_estimators 200, max_depth 10 cross validation AUC score 0.7505684358054535
For n_estimators 500, max_depth 7 cross validation AUC score 0.7459886332343842
For n_estimators 500, max_depth 10 cross validation AUC score 0.7505138599899948
For n_estimators 1000, max_depth 7 cross validation AUC score 0.7461110203554747
For n_estimators 1000, max_depth 10 cross validation AUC score 0.7503188106611327
For n_estimators 2000, max_depth 7 cross validation AUC score 0.7463165060899846
For n_estimators 2000, max_depth 10 cross validation AUC score 0.7504836210112507
```

Cross validation results for Random Forest model.



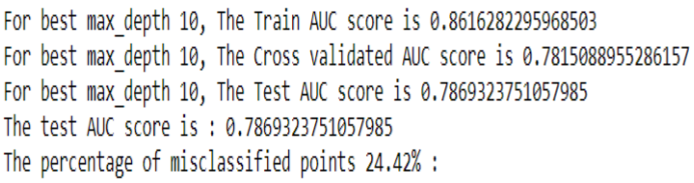
```
The optimal values are: n_estimators 200, max_depth 10
For best n_estimators 200 best max_depth 10, The Train AUC score is 0.8417031819440642
For best n_estimators 200 best max_depth 10, The Validation AUC score is 0.7505684358054535
For best n_estimators 200 best max_depth 10, The Test AUC score is 0.7504063992087786
The test AUC score is : 0.7504063992087786
The percentage of misclassified points 08.08% :
```

Random Forest model results.

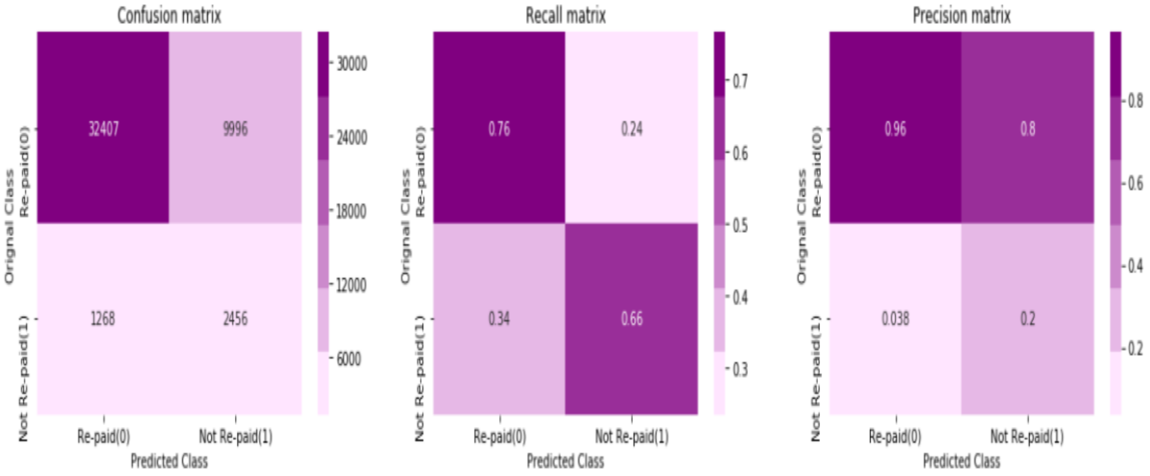


ROC curve for Random Forest model with AUC=0.75

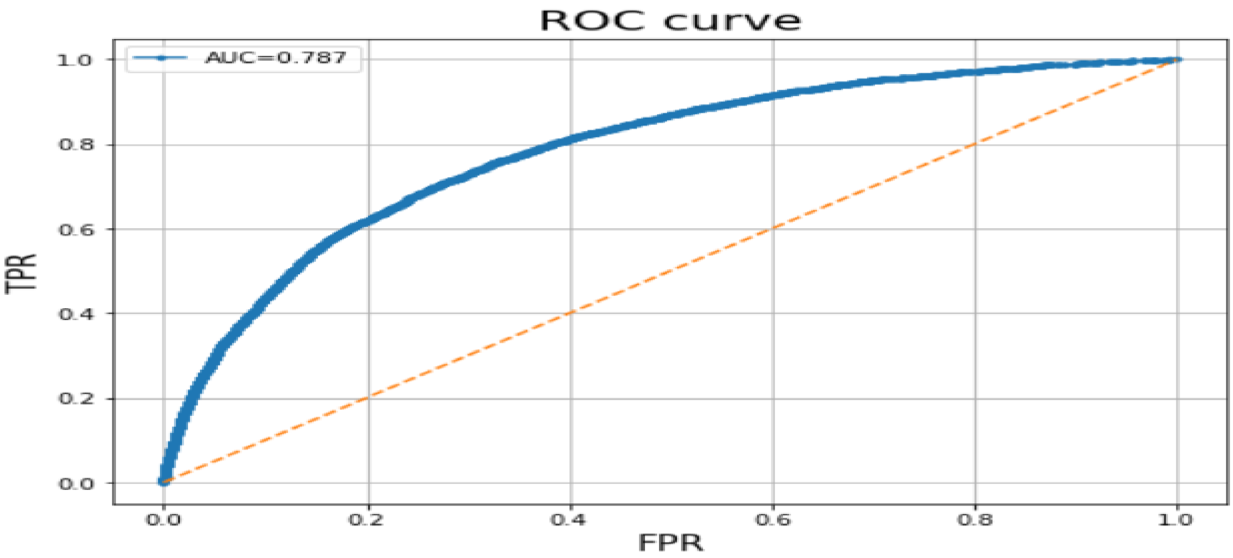
LightGBM with selected features:



```
For best max_depth 10, The Train AUC score is 0.8616282295968503
For best max_depth 10, The Cross validated AUC score is 0.7815088955286157
For best max_depth 10, The Test AUC score is 0.7869323751057985
The test AUC score is : 0.7869323751057985
The percentage of misclassified points 24.42% :
```



LightGBM model Results



ROC curve for LightGBM model with AUC=0.787

Overview of Results:

Model	Train AUC	Validation AUC	Test AUC
Logistic Regression with Selected features	0.756	0.747	0.753
Random Forest with Selected features	0.841	0.751	0.751
LightGBM with Selected features	**0.861**	**0.781**	**0.787**

LightGBM gives the best performance and it is also faster to train when compared to Xgboost.