Reunion Assessment

SHUBHAM NAIK

Presentation Overview

- Problem statement
- Data Information
- EDA & feature engineering
- Preparing Data for modeling
- Implementing Model
- Model summary
- Conclusion

Problem statement

Develop the ML model(s) to predict the credit risk(low or high) for a given applicant.

Business Constraint: Note that it is worse to state an applicant as a low credit risk when they are actually a high risk, than it is to state an applicant to be a high credit risk when they aren't.

Data Information

- **Attributes:**
  - Primary_applicant_age_in_years (numeric)
  - Gender (string)
  - Marital_status (string)
  - Number_of_dependents (numeric)
  - Housing (string)
  - Years_at_current_residence (numeric)
  - Employment_status (string)
  - Has_been_employed_for_at_least (string)
  - Has_been_employed_for_at_most (string)
  - Telephone (string)
  - Foreign_worker (numeric)
  - Savings_account_balance (string)
  - Balance_in_existing_bank_account_(lower_limit_of_bucket) (string)
  - Balance_in_existing_bank_account_(upper_limit_of_bucket) (string)

Data Information

- **Attributes:**

  - applicant_id (string)
  - Months_loan_taken_for (numeric)
  - Purpose (string)
  - Principal_loan_amount (numeric)
  - EMI_rate_in_percentage_of_disposable_income (numeric)
  - Property (string)
  - Has_coapplicant (numeric)
  - Has_guarantor (numeric)
  - Other_EMI_plans (string)
  - Number_of_existing_loans_at_this_bank (numeric)
  - Loan_history (string)

Data Processing :

#	Column	Non-Null Count	Dtype
0	applicant_id	1000 non-null	int64
1	Primary_applicant_age_in_years	1000 non-null	int64
2	Gender	1000 non-null	object
3	Marital_status	1000 non-null	object
4	Number_of_dependents	1000 non-null	int64
5	Housing	1000 non-null	object
6	Years_at_current_residence	1000 non-null	int64
7	Employment_status	1000 non-null	object
8	Has_been_employed_for_at_least	938 non-null	object
9	Has_been_employed_for_at_most	747 non-null	object
10	Telephone	404 non-null	object
11	Foreign_worker	1000 non-null	int64
12	Savings_account_balance	817 non-null	object
13	Balance_in_existing_bank_account_(lower_limit_of_bucket)	332 non-null	object
14	Balance_in_existing_bank_account_(upper_limit_of_bucket)	543 non-null	object
15	loan_application_id	1000 non-null	object
16	Months_loan_taken_for	1000 non-null	int64
17	Purpose	988 non-null	object
18	Principal_loan_amount	1000 non-null	int64
19	EMI_rate_in_percentage_of_disposable_income	1000 non-null	int64
20	Property	846 non-null	object
21	Has_coapplicant	1000 non-null	int64
22	Has_guarantor	1000 non-null	int64
23	Other_EMI_plans	186 non-null	object
24	Number_of_existing_loans_at_this_bank	1000 non-null	int64
25	Loan_history	1000 non-null	object
26	high_risk_applicant	1000 non-null	int64

- **1000 entries, 26 columns**

- Data processing is an important aspect of EDA. In the table you can see that the number of observation for each attributes are not same, which can affect our data while analyzing.

Data Processing :

- So there are total 26 columns and 1000 rows.
- 9 columns have missing values :

```
applicant_id                          0
Primary_applicant_age_in_years        0
Gender                                0
Marital_status                        0
Number_of_dependents                  0
Housing                               0
Years_at_current_residence            0
Employment_status                     0
empd_for_atleast                     62
Has_been_employed_for_at_most       253
Telephone                           596
Foreign_worker                        0
Savings_account_balance             183
A/c balance lower                   668
A/c balance upper                   457
loan_application_id                   0
Months_loan_taken_for                 0
Purpose                              12
Principal_loan_amount                 0
EMI_rate_prcnt                        0
Property                            154
Has_coapplicant                       0
Has_guarantor                         0
Other_EMI_plans                     814
Number_of_existing_loans_at_this_bank 0
Loan_history                          0
high_risk_applicant                   0
```

Treatment:

- Since Purpose have very few null values we can drop their entire row of total 12 observations. It wont affect the data significantly.
- In A/c balance lower, A/c balance upper I substituted the null values with the word ‘0’ for further analysis
- I have removed the column ‘Other EMI plans’ because it has more than 80% null values. Also removed the Telephone column, applicant_id and loan_applicant_id because by using ID column we cant gain any useful insights.
- In total_disbursement_amount I took the median and filled with it.

Data Processing :

Modified Dataset:

#	Column	Non-Null Count	Dtype
---	------	--------------	-----
0	applicant_id	988 non-null	int64
1	Primary_applicant_age_in_years	988 non-null	int64
2	Gender	988 non-null	object
3	Marital_status	988 non-null	object
4	Number_of_dependents	988 non-null	int64
5	Housing	988 non-null	object
6	Years_at_current_residence	988 non-null	int64
7	Employment_status	988 non-null	object
8	empd_for_atleast	988 non-null	object
9	Has_been_employed_for_at_most	988 non-null	object
10	Foreign_worker	988 non-null	int64
11	Savings_account_balance	988 non-null	object
12	A/c balance lower	988 non-null	object
13	A/c balance upper	988 non-null	object
14	loan_application_id	988 non-null	object
15	Months_loan_taken_for	988 non-null	int64
16	Purpose	988 non-null	object
17	Principal_loan_amount	988 non-null	int64
18	EMI_rate_prcnt	988 non-null	int64
19	Property	988 non-null	object
20	Has_coapplicant	988 non-null	int64
21	Has_guarantor	988 non-null	int64
22	Number_of_existing_loans_at_this_bank	988 non-null	int64
23	Loan_history	988 non-null	object
24	high_risk_applicant	988 non-null	int64

- Dataset is now optimized and we can proceed with analysis and visualization

Basic Exploration

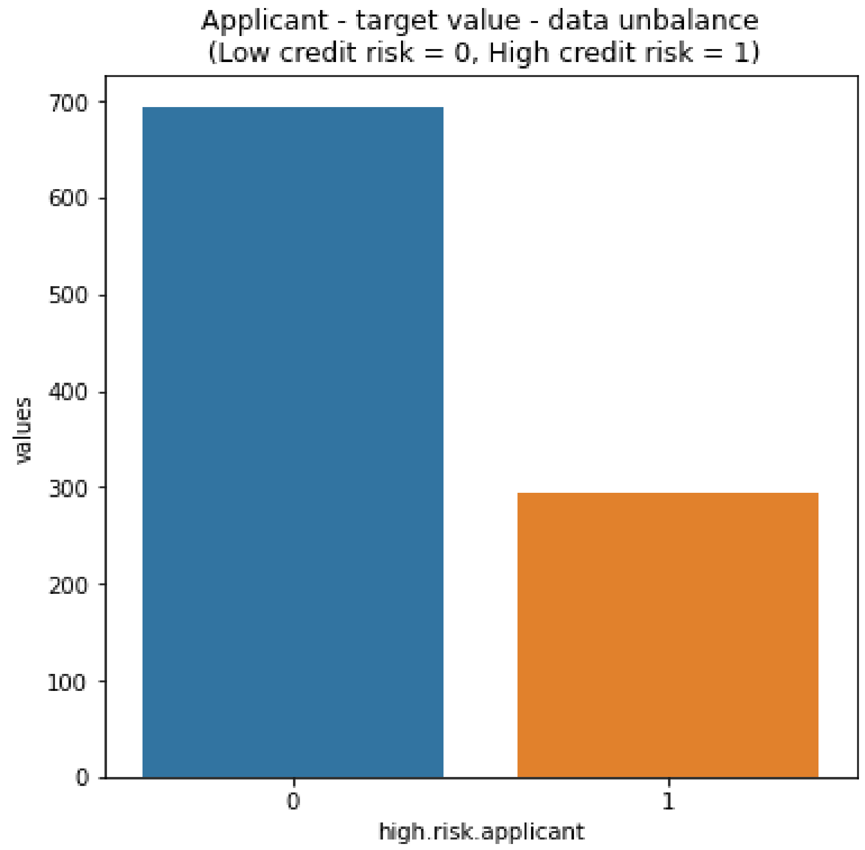
- With some quick analysis we found that there are more than 6 features that are highly correlated with each other.
- Over 69% of applicants are male whereas the remaining 31% are female.
- There is more applicant with marital status as single than married or divorced.
- Many applicants have their own housing property and their marital status is single.
- Over 400 applicants have been residing in the same residency for 4 years.
- Among all applicants 629 are skilled employee whereas 199 are unskilled.
- The main purpose for getting the loan is to buy electronic equipment or new vehicle.
- Over 620 applicants already have pending loan in the same bank and 693 of such applicants have been flagged as high-risk applicants.

Univariate Analysis

- (Lets have a look at the distribution of various variables in the Data set)

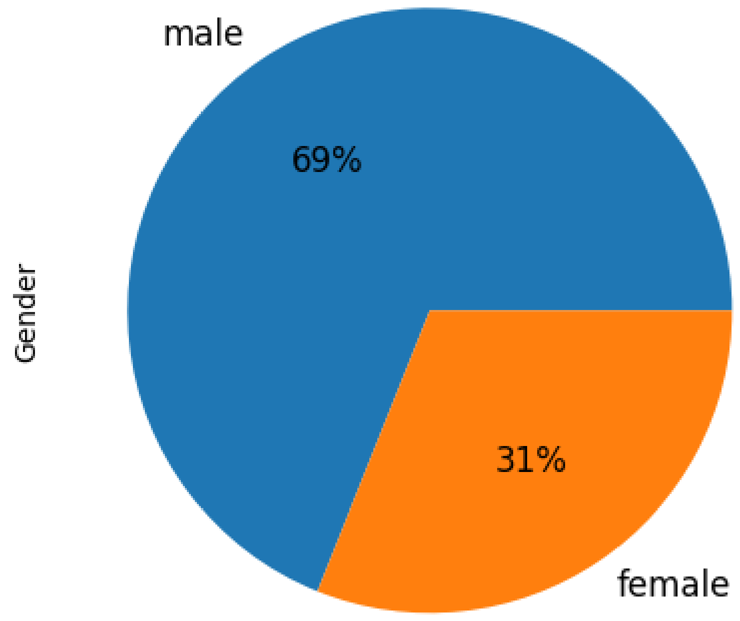
EDA

Distribution of target classes is imbalanced, non-risk applicant far outnumber risk applicants. This is common in these datasets since most people pay credit cards on time



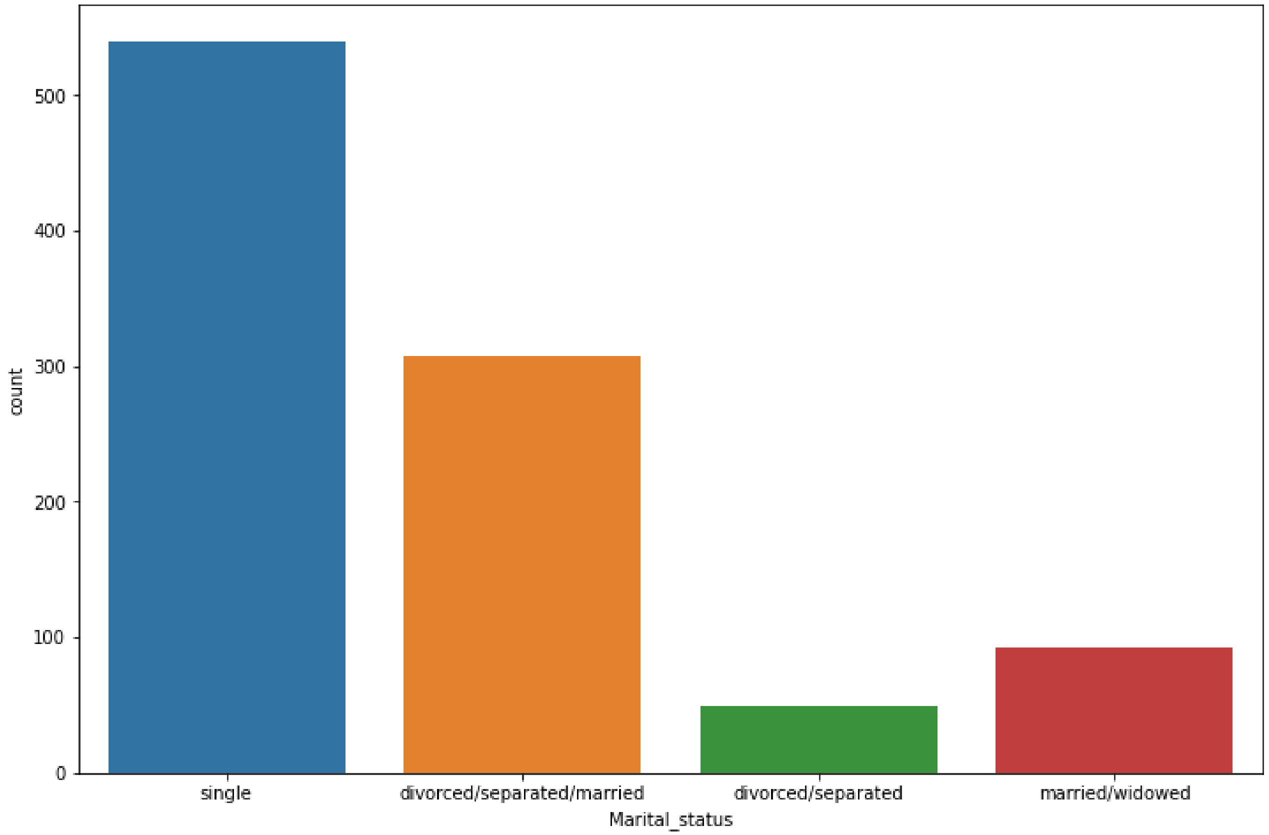
EDA

- By analyzing the Gender; we can conclude that the majority are of Male with occupancy of 69%, where as female occupies 31% among all applicants.



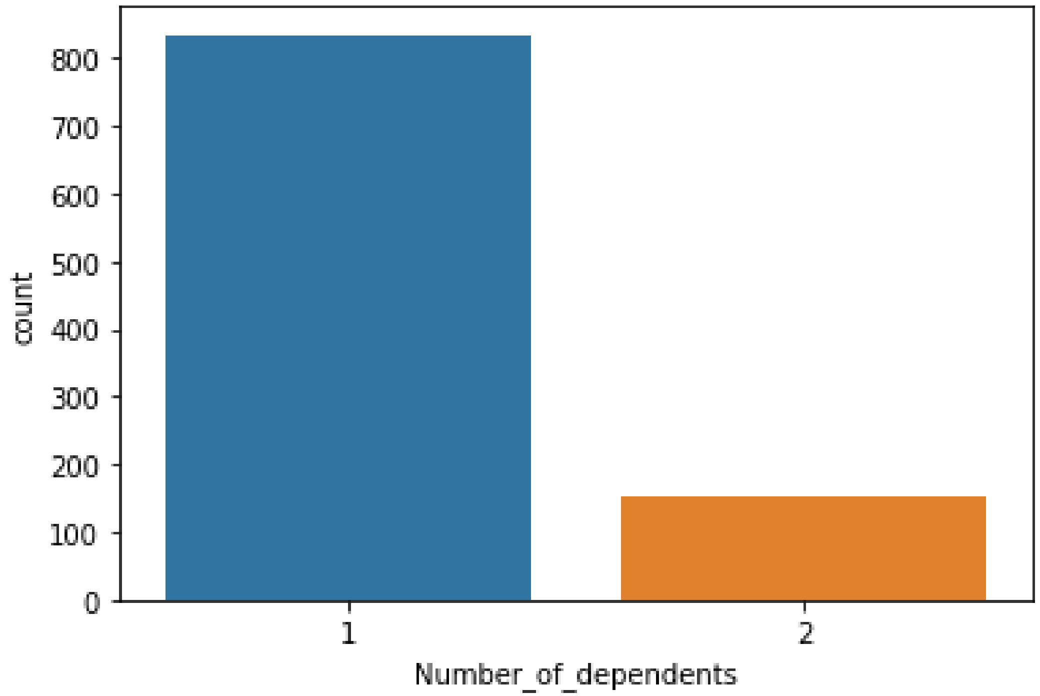
EDA

- From Marital status we can conclude that more than 500 of the applicants marital status is single followed by divorced/separated/married are more than 300 applicants



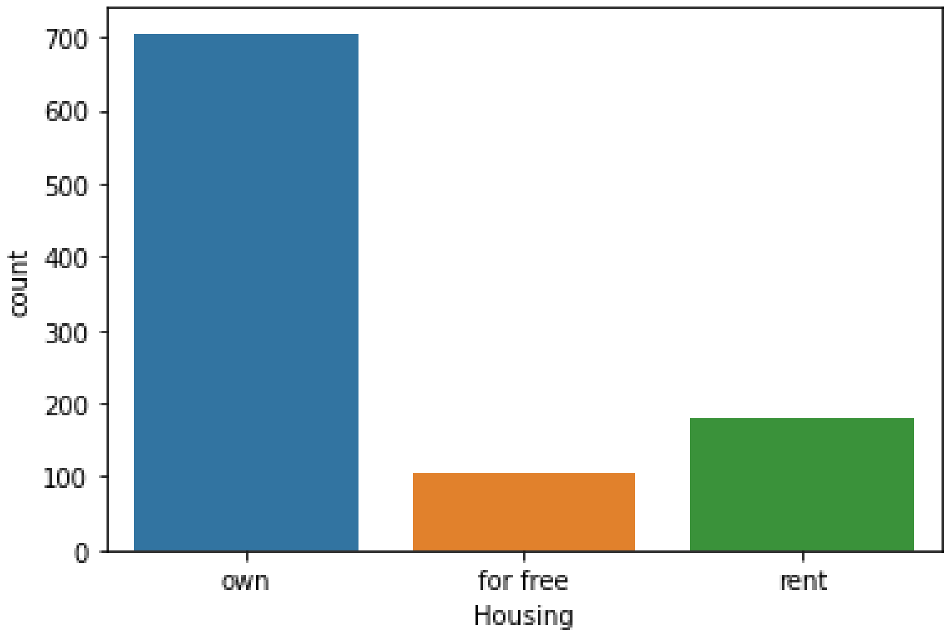
EDA

- From this plot we can conclude that more than 800 applicants have 1 dependent



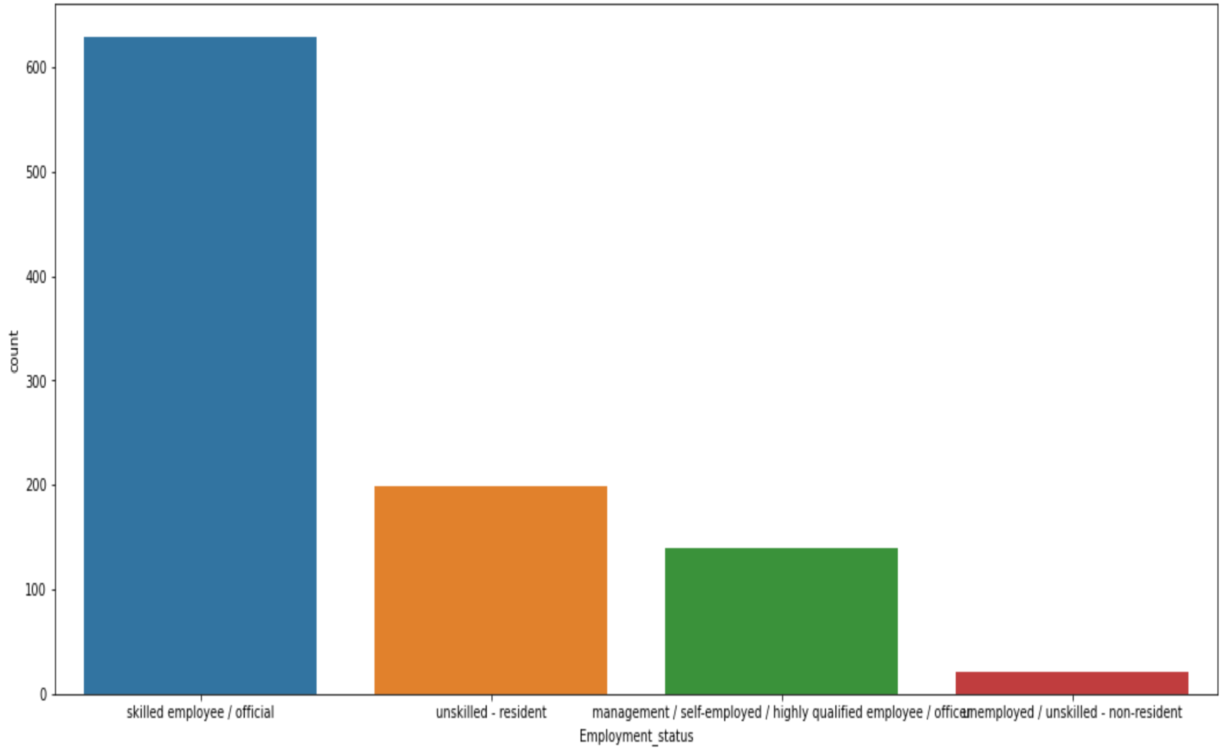
EDA

- From this plot we can conclude that most of applicants has their own housing which is almost 700 and almost 200 applicants are on rent.



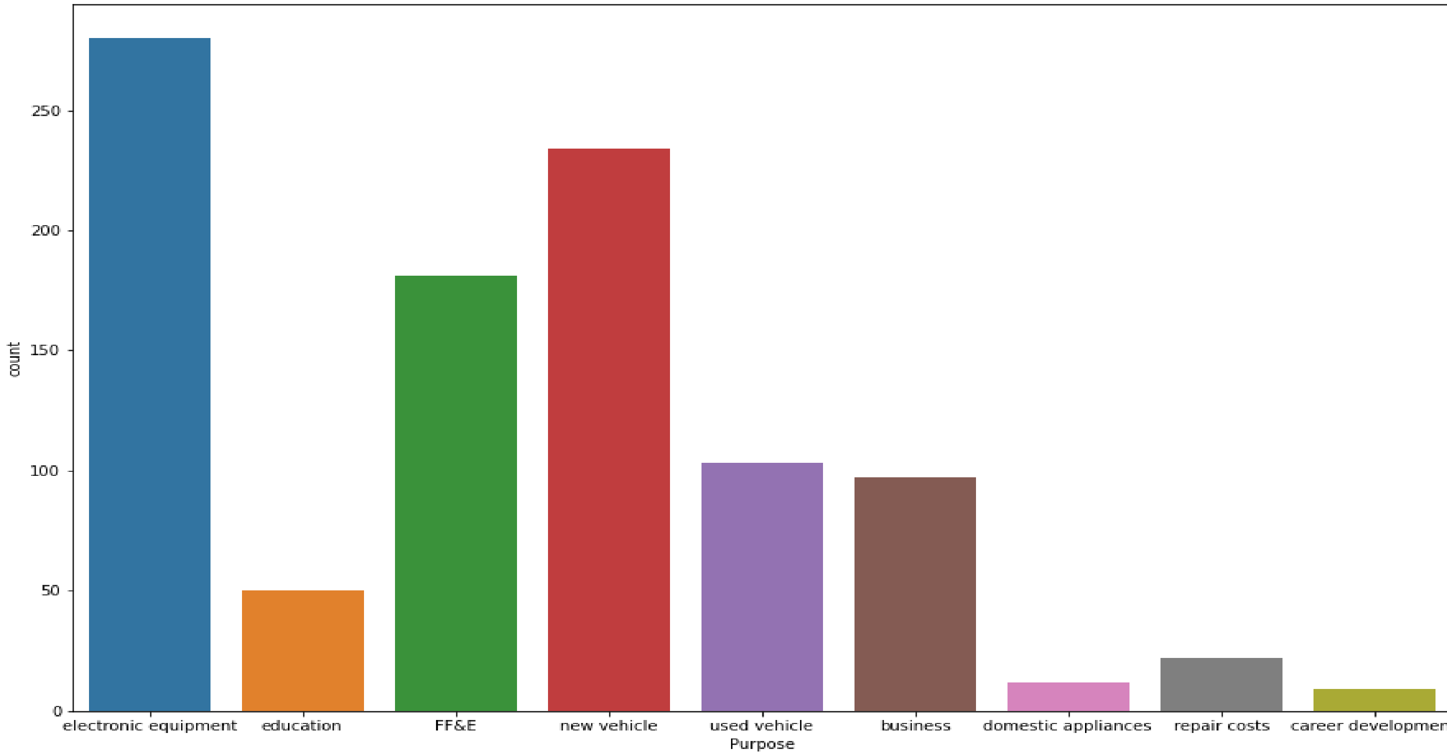
EDA

- From this plot we can conclude that more than 600 applicants are skilled employee or official and only 12 applicants are unemployed or unskilled non resident.



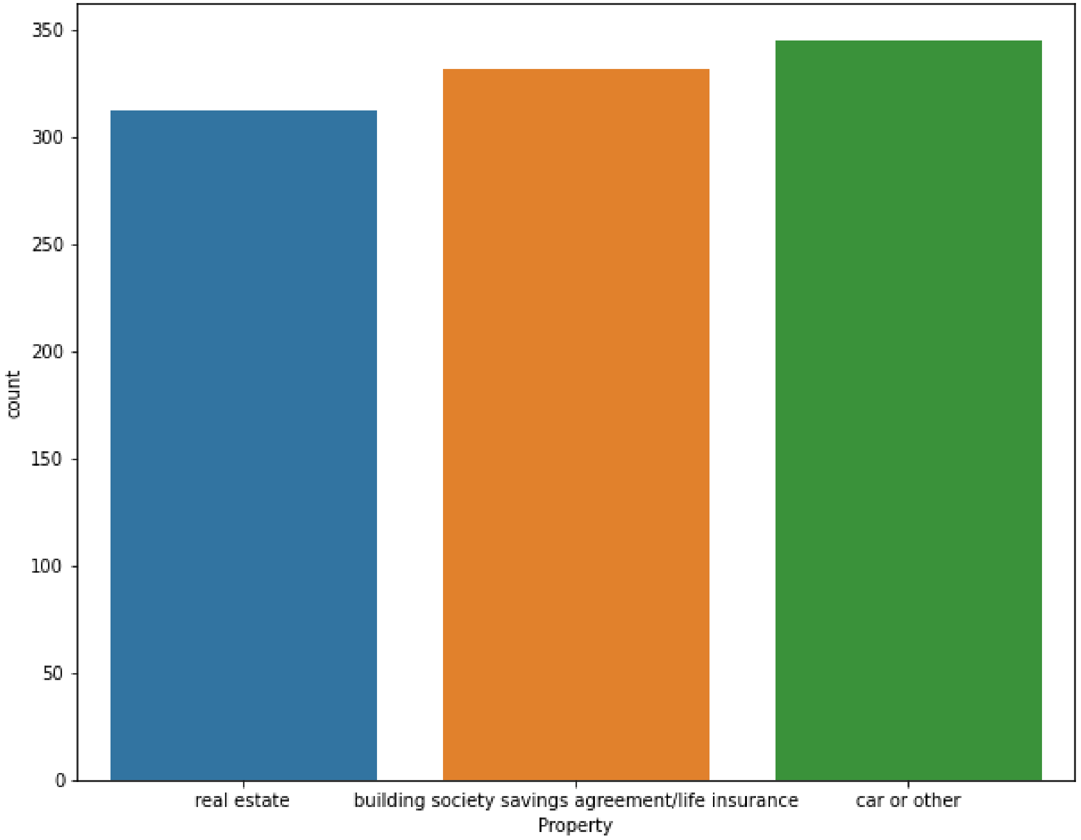
EDA

- From this plot we can conclude that most applicants loan purpose are electronic equipment which is 280 followed by new vehicle are 234 and least purpose was career development 9



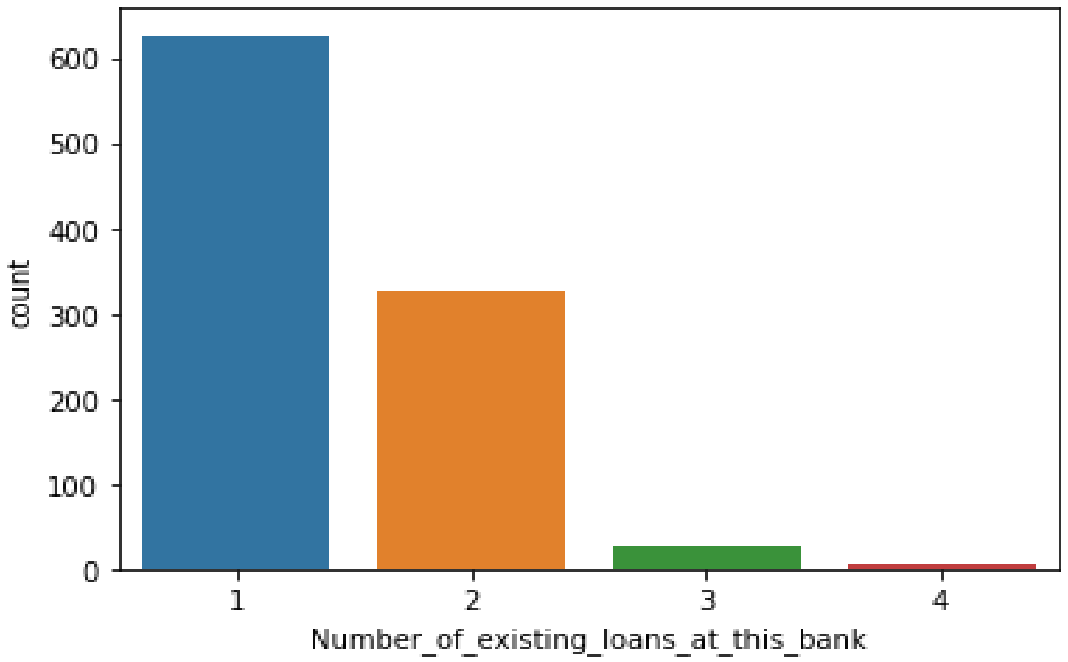
EDA

- From this plot we can see that among all applicants the property type count are almost same.



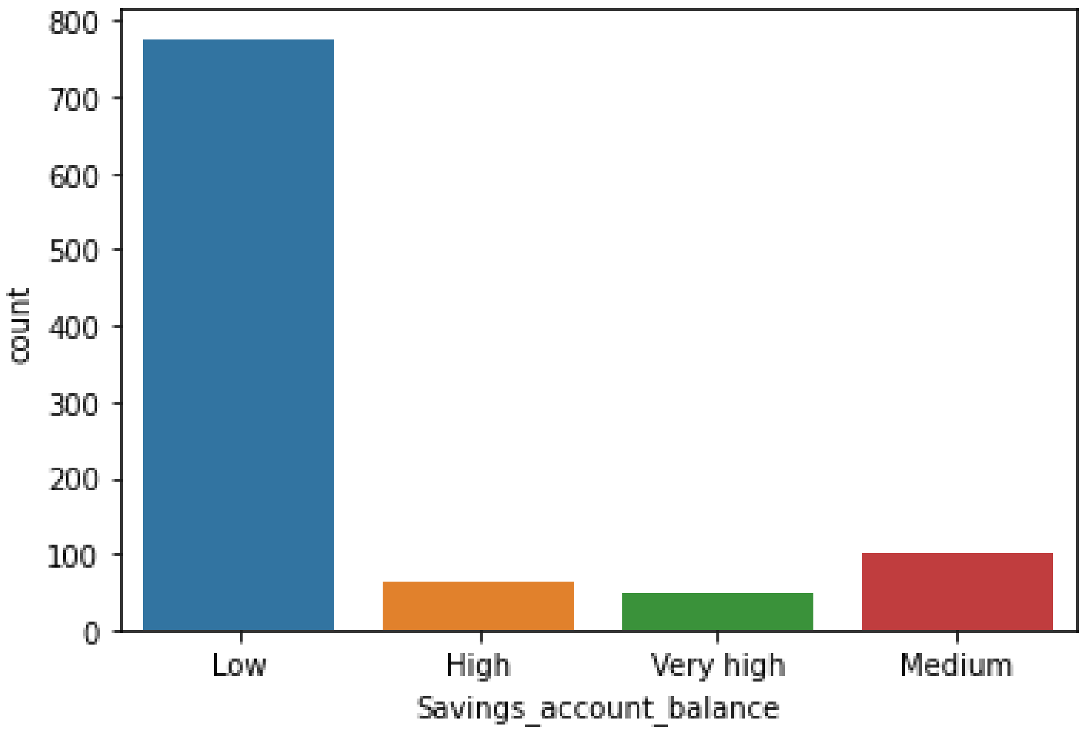
EDA

- From this plot we can conclude that most of applicants has 1 or 2 loan in existing bank



EDA

- From this plot we see that most applicants savings accounts balance are low

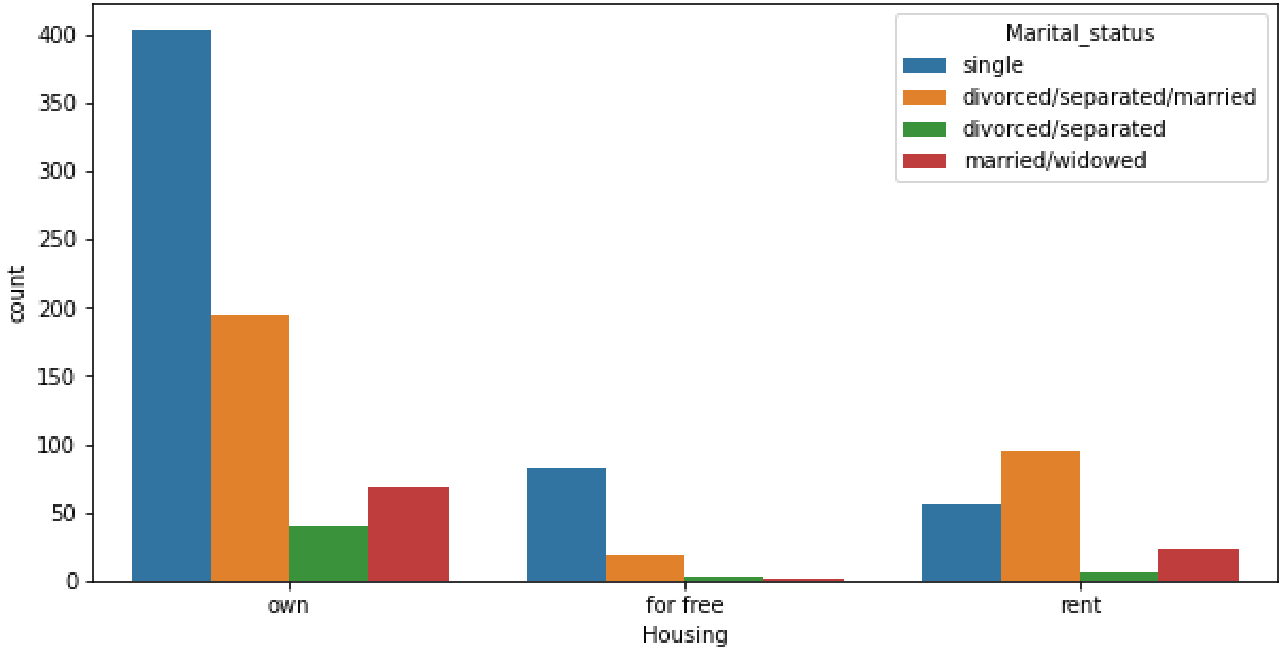


Bivariate Analysis

- (Bivariate Analysis involves finding relationships, patterns, and correlations between two variables.)

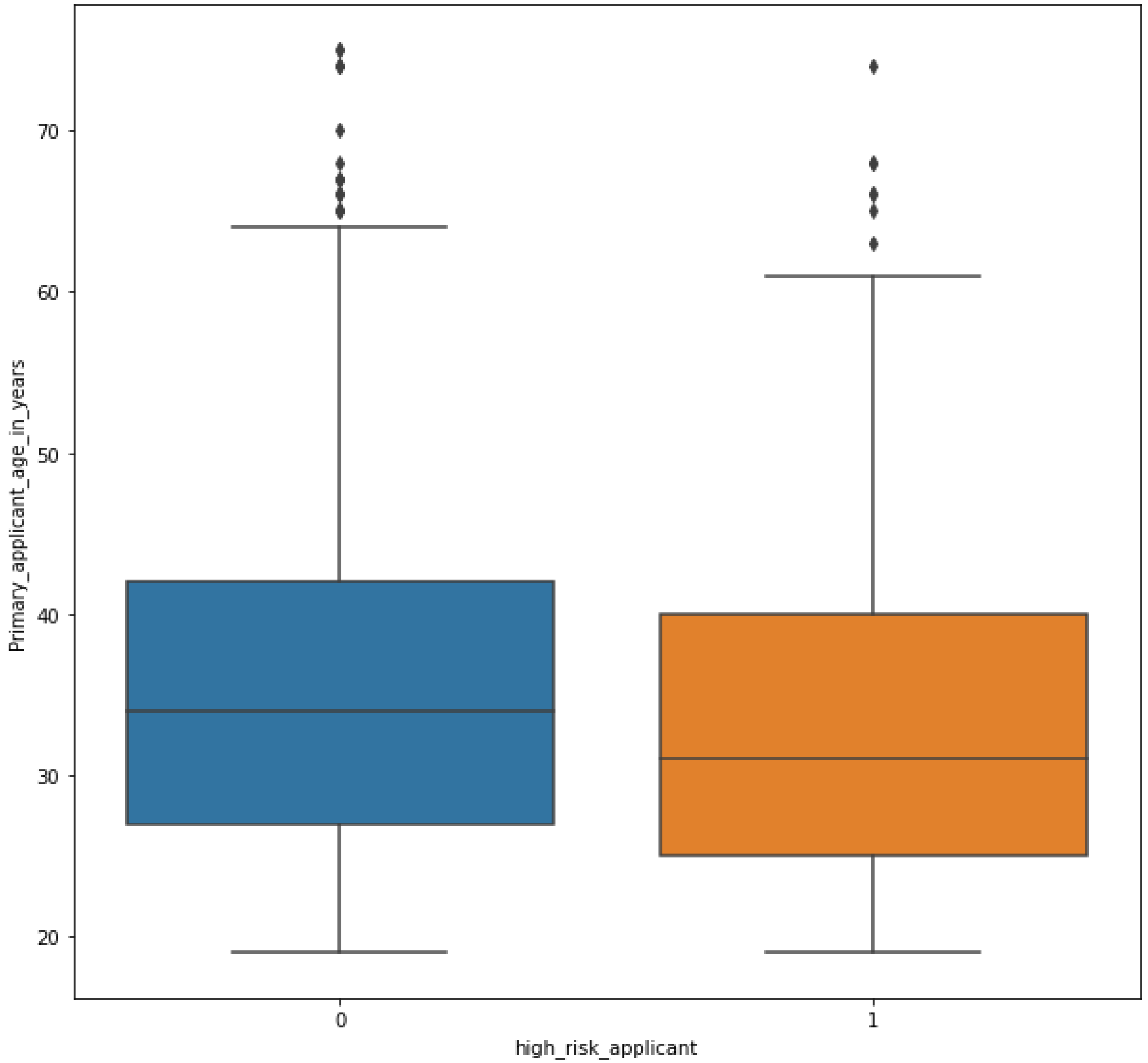
EDA

- From this plot we can see that the applicants marital status is single has maximum count of own housing.



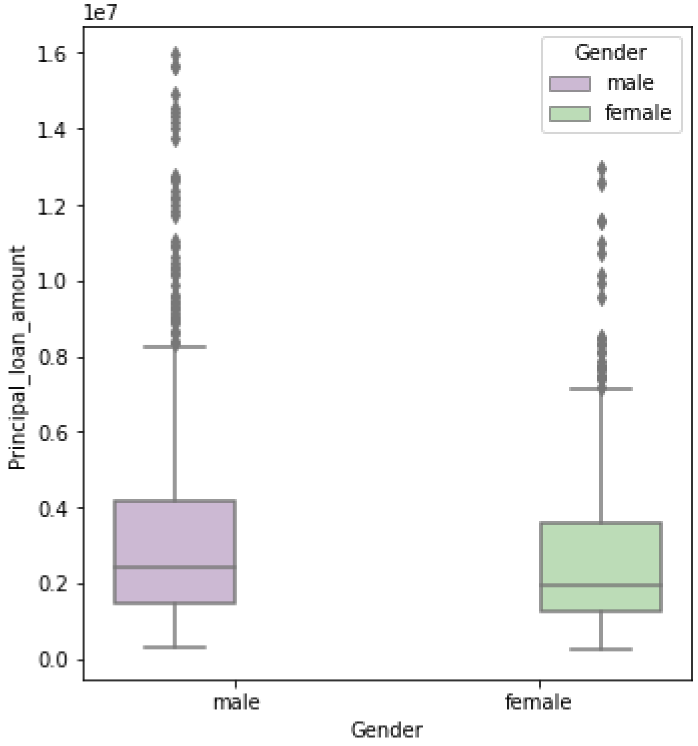
EDA

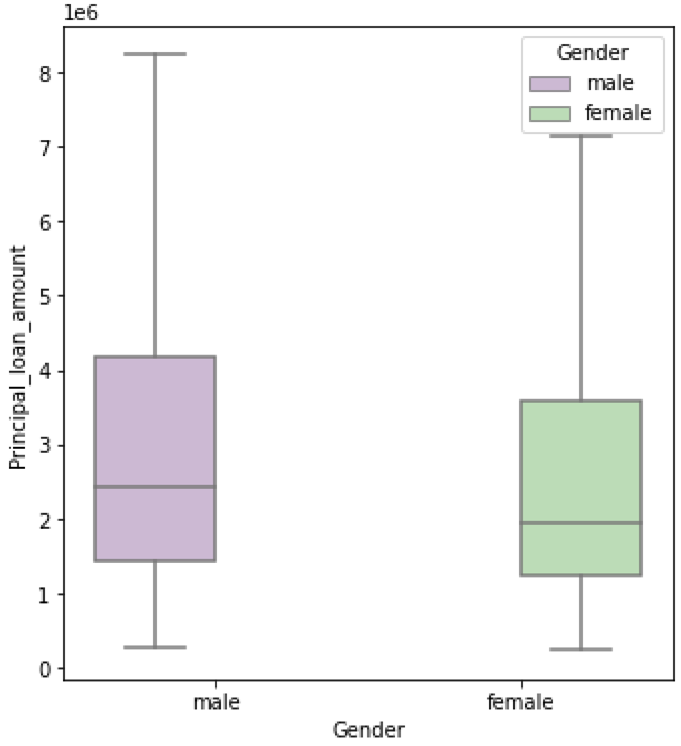
- From this plot we can see that applicants age varies from 19 to 75 for both high risk and low risk applicants



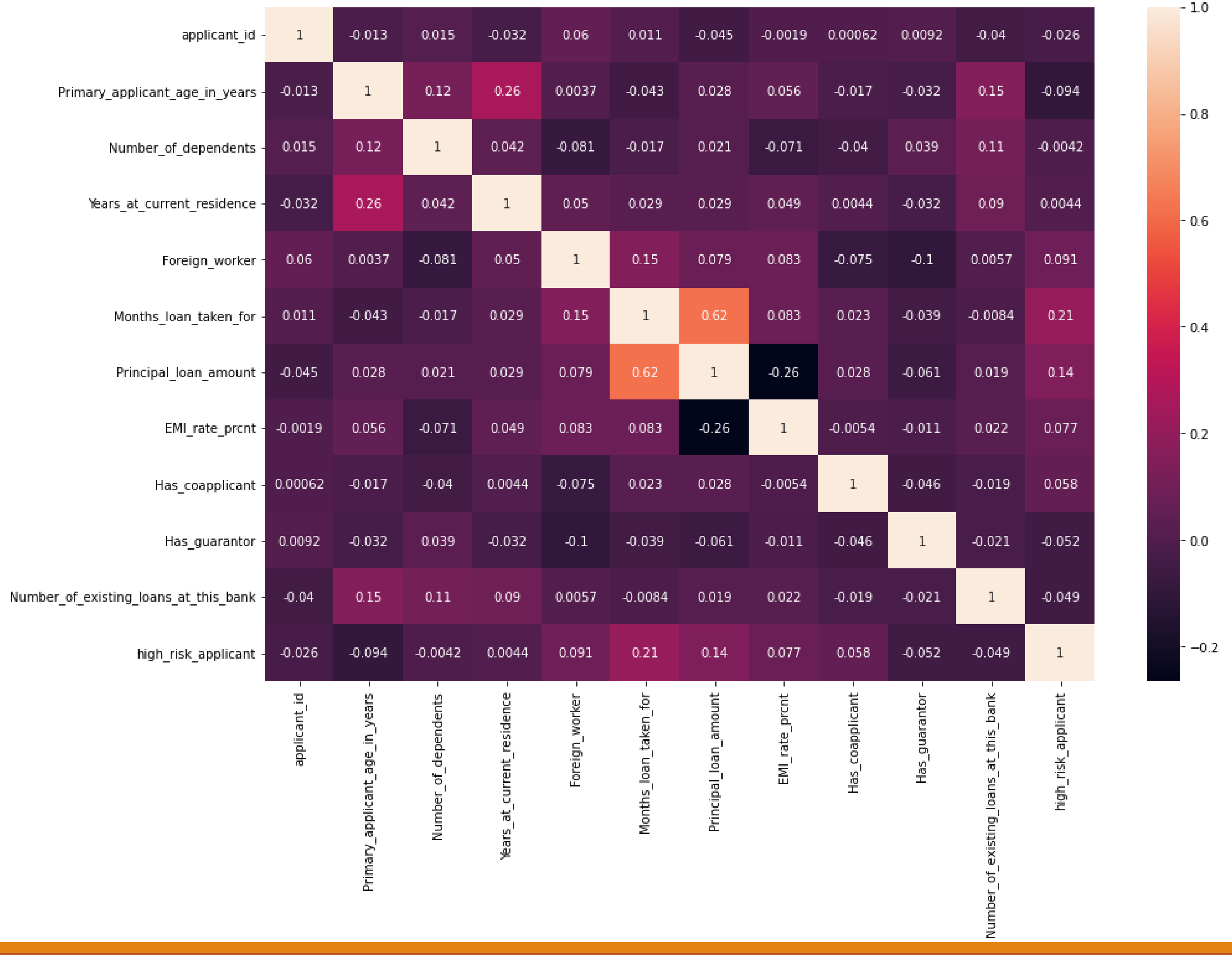
EDA

- From this box plot we can see that male gender raise more loan amount than female





EDA


applicant_id 1 -0.013 0.015 -0.032 0.06 0.011 -0.045 -0.0019 0.00062 0.0092 -0.04 -0.026
Primary_applicant_age_in_years -0.013 1 0.12 0.26 0.0037 -0.043 0.028 0.056 -0.017 -0.032 0.15 -0.094
Number_of_dependents 0.015 0.12 1 0.042 -0.081 -0.017 0.021 -0.071 -0.04 0.039 0.11 -0.0042
Years_at_current_residence -0.032 0.26 0.042 1 0.05 0.029 0.029 0.049 0.0044 -0.032 0.09 0.0044
Foreign_worker 0.06 0.0037 -0.081 0.05 1 0.15 0.079 0.083 -0.075 -0.1 0.0057 0.091
Months_loan_taken_for 0.011 -0.043 -0.017 0.029 0.15 1 0.62 0.083 0.023 -0.039 -0.0084 0.21
Principal_loan_amount -0.045 0.028 0.021 0.029 0.079 0.62 1 -0.26 0.028 -0.061 0.019 0.14
EMI_rate_prcnt -0.0019 0.056 -0.071 0.049 0.083 0.083 -0.26 1 -0.0054 -0.011 0.022 0.077
Has_coapplicant 0.00062 -0.017 -0.04 0.0044 -0.075 0.023 0.028 -0.0054 1 -0.046 -0.019 0.058
Has_guarantor 0.0092 -0.032 0.039 -0.032 -0.1 -0.039 -0.061 -0.011 -0.046 1 -0.021 -0.052
Number_of_existing_loans_at_this_bank -0.04 0.15 0.11 0.09 0.0057 -0.0084 0.019 0.022 -0.019 -0.021 1 -0.049
high_risk_applicant -0.026 -0.094 -0.0042 0.0044 0.091 0.21 0.14 0.077 0.058 -0.052 -0.049 1
applicant_id
Primary_applicant_age_in_years
Number_of_dependents
Years_at_current_residence
Foreign_worker
Months_loan_taken_for
Principal_loan_amount
EMI_rate_prcnt
Has_coapplicant
Has_guarantor
Number_of_existing_loans_at_this_bank
high_risk_applicant
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

Modeling Steps

Data Preprocessing

- Feature selection
- Train test data split (70%-30%)

Data Fitting & Tuning

- Start with default model parameters
- Hyperparameter tuning
- Measure RUC-AOC on training data

Model Evaluation

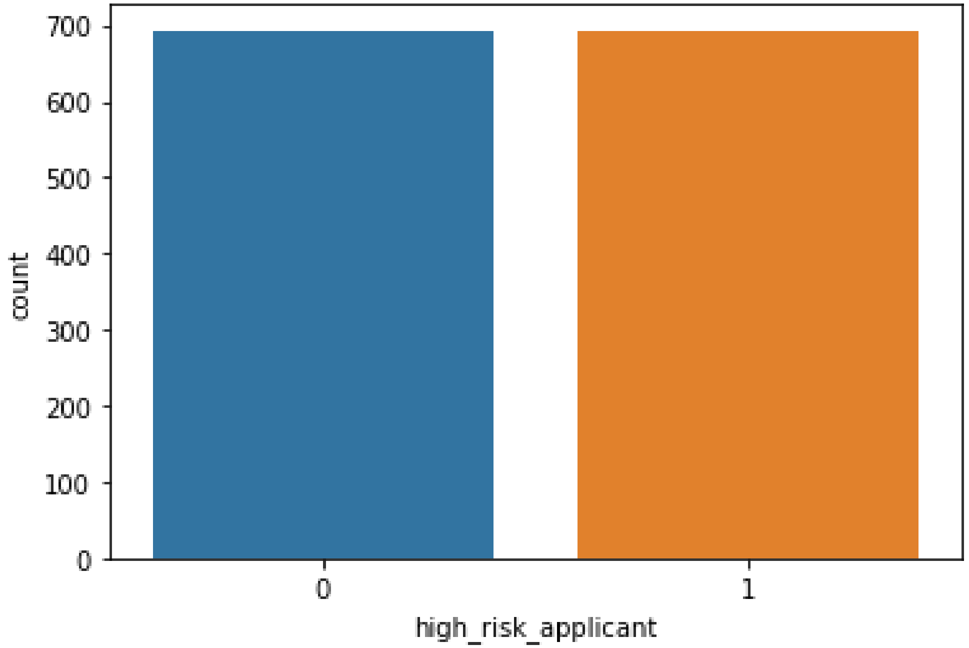
- Model testing
- Precision_Recall Score
- Comapare with the other models

Preparing Data for modeling

As we have seen earlier that we have imbalanced dataset. So to remediate Imbalance we are using SMOTE(Synthetic Minority Oversampling Technique)

Original dataset shape 988

Resampled dataset shape 1386



Applying Model

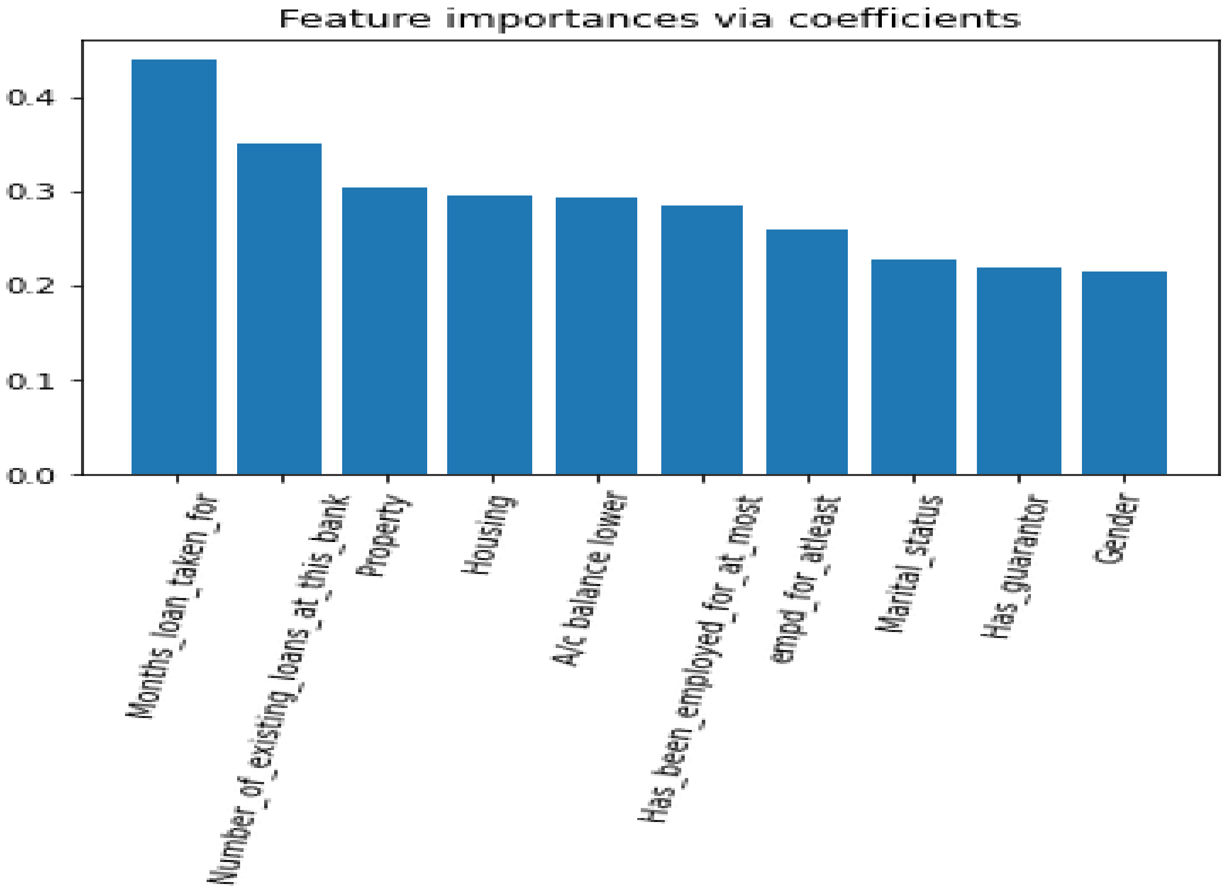
- **Logistic Modelling**

  - **Parameters:**

```
The accuracy on test data is  0.6899038461538461
The precision on test data is  0.6923076923076923
The recall on test data is  0.6889952153110048
The f1 on test data is  0.6906474820143885
The roc_score on test data is  0.6899082356748261
```

  - C = 0.01
  - Penalty = L2

Logistic feature importance



Applying Model

- Support Vector Classifier

  - Parameters:

```
The accuracy on test data is  0.7235576923076923
The precision on test data is  0.7067307692307693
The recall on test data is  0.7313432835820896
The f1 on test data is  0.7188264058679706
The roc_score on test data is  0.7238111766747658
```

  - C = 10
  - Kernel = ‘rbf’

Applying Model

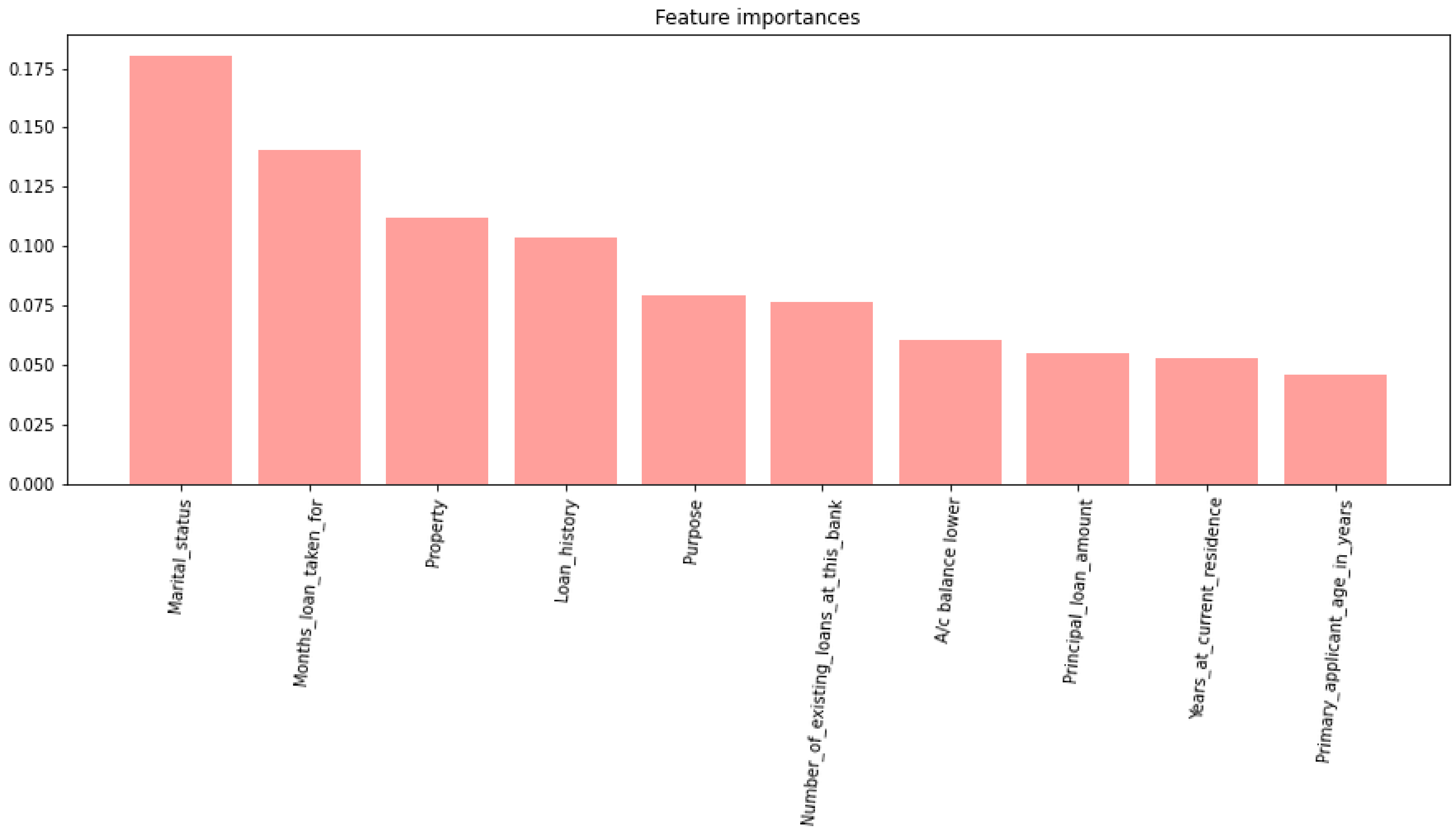
- **Decision Tree Classifier**

  - **Parameters:**

```
The accuracy on test data is  0.6442307692307693
The precision on test data is  0.6634615384615384
The recall on test data is  0.6388888888888888
The f1 on test data is  0.6509433962264151
The roc_score on test data is  0.6444444444444444
```

  - max_depth=50
  - min_samples_split=0.1

Random Forest feature importance



Applying Model

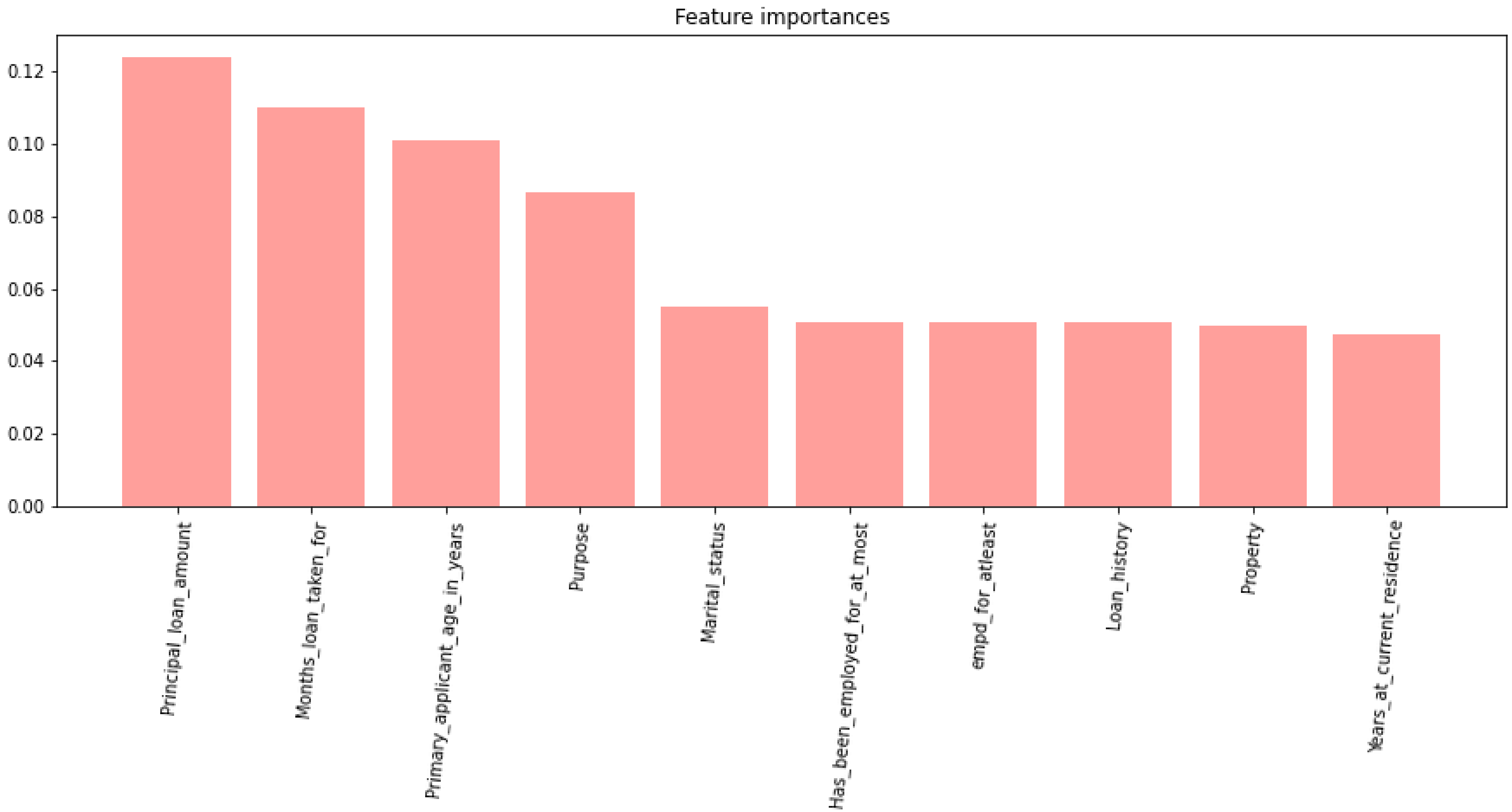
- **Random Forest**

  - **Parameters:**

```
The accuracy on test data is  0.7548076923076923
The precision on test data is  0.7355769230769231
The recall on test data is  0.765
The f1 on test data is  0.7500000000000001
The roc_score on test data is  0.7551851851851852
```

  - max_depth=30
  - n_estimators=100

Random Forest feature importance



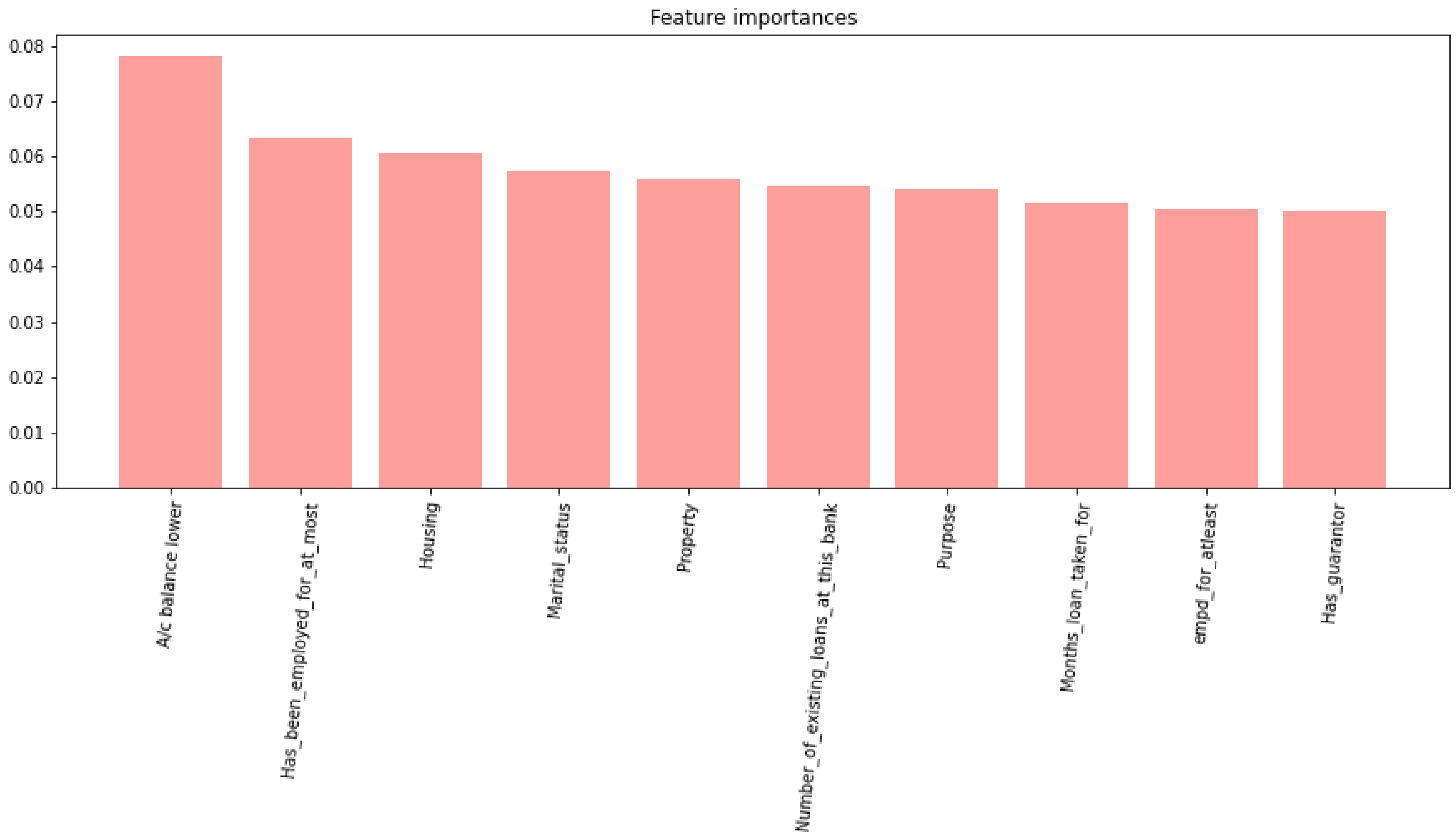
Applying Model

- XGBoost
  - Parameters:

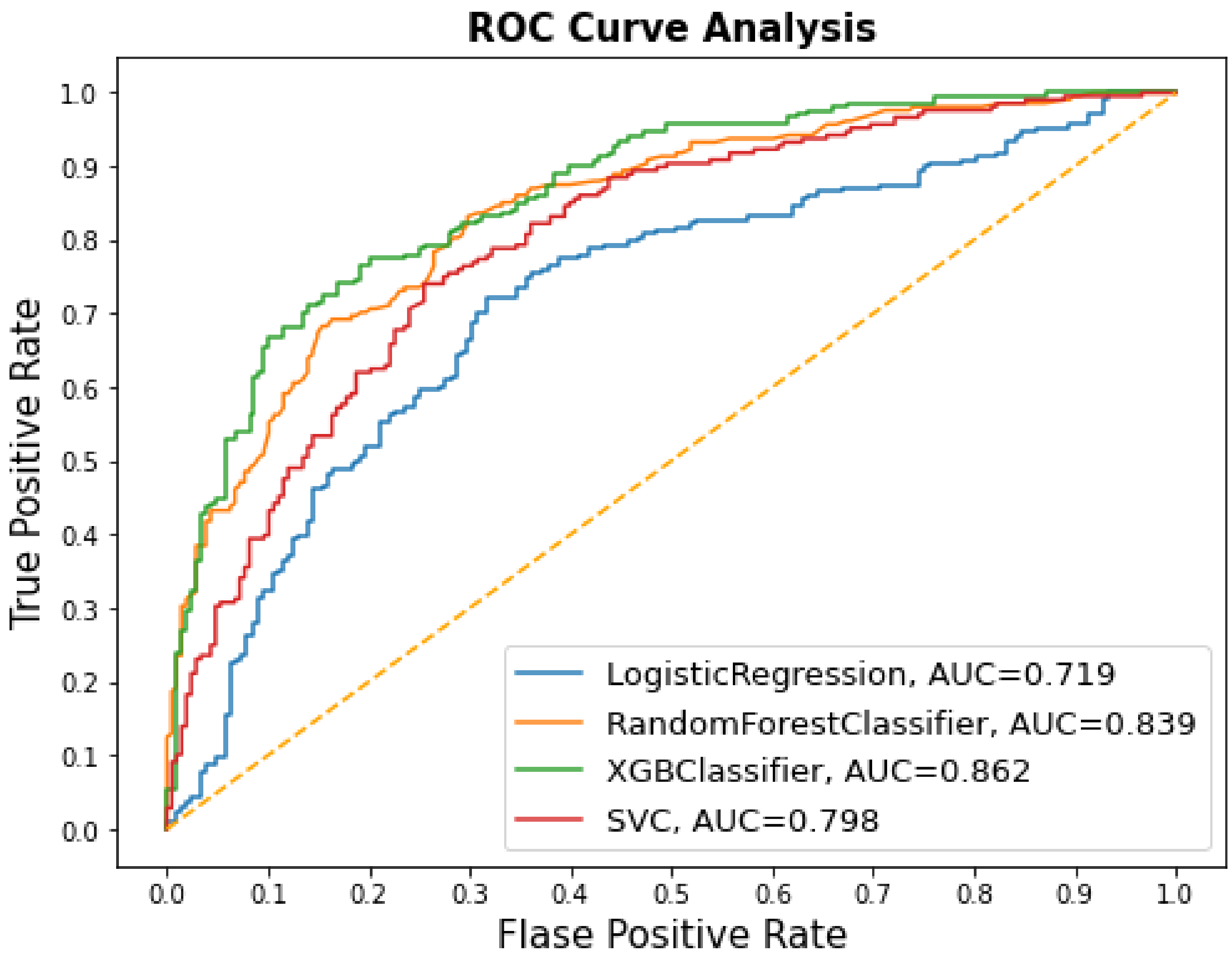
```
The accuracy on test data is  0.7331730769230769
The precision on test data is  0.7355769230769231
The recall on test data is  0.7320574162679426
The f1 on test data is  0.7338129496402878
The roc_score on train data is  0.7331784665880775
```

  - max_depth=9
  - min_child_weight=1

XGBoost feature importance



AUC-ROC curve comparison



Conclusion

- XGBoost provided us the best results giving us a recall of 76 percent(meaning out of 100 risk applicants 76 will be having high chances of paying loan back)
- Random Forest also had good score as well.
- Logistic regression being the least accurate with a recall of 69.

	Classifier	Train Accuracy	Test Accuracy	Precision Score	Recall Score	F1 score
0	Logistic Regression	0.721649	0.694712	0.697115	0.693780	0.695444
1	SVC	0.967010	0.737981	0.740385	0.736842	0.738609
2	Random Forest CLf	1.000000	0.742788	0.740385	0.743961	0.742169
3	Xgboost Clf	1.000000	0.771635	0.778846	0.767773	0.773270

Thank You