LOAN ELIGIBILITY PREDICTION

Lubna Rahman
Data Science Capstone Project

Springboard

Thanks to Springboard

PROBLEM STATEMENT

❑ Risk is always involved in approval of loans.

❑ Even after analyzing loan application data numerous times, approval decisions are not always correct.

❑ Dream Housing Finance company deals in home loans. They have presence across urban, semi- urban and rural areas.

❑ They want a smart loan approval process to reduce the risk and losses incurred by the company.

TARGET AUDIENCE

- ❑ Commercial Banks
- ❑ Savings and Loan Association
- ❑ Credit Union
- ❑ Brokerage Firm

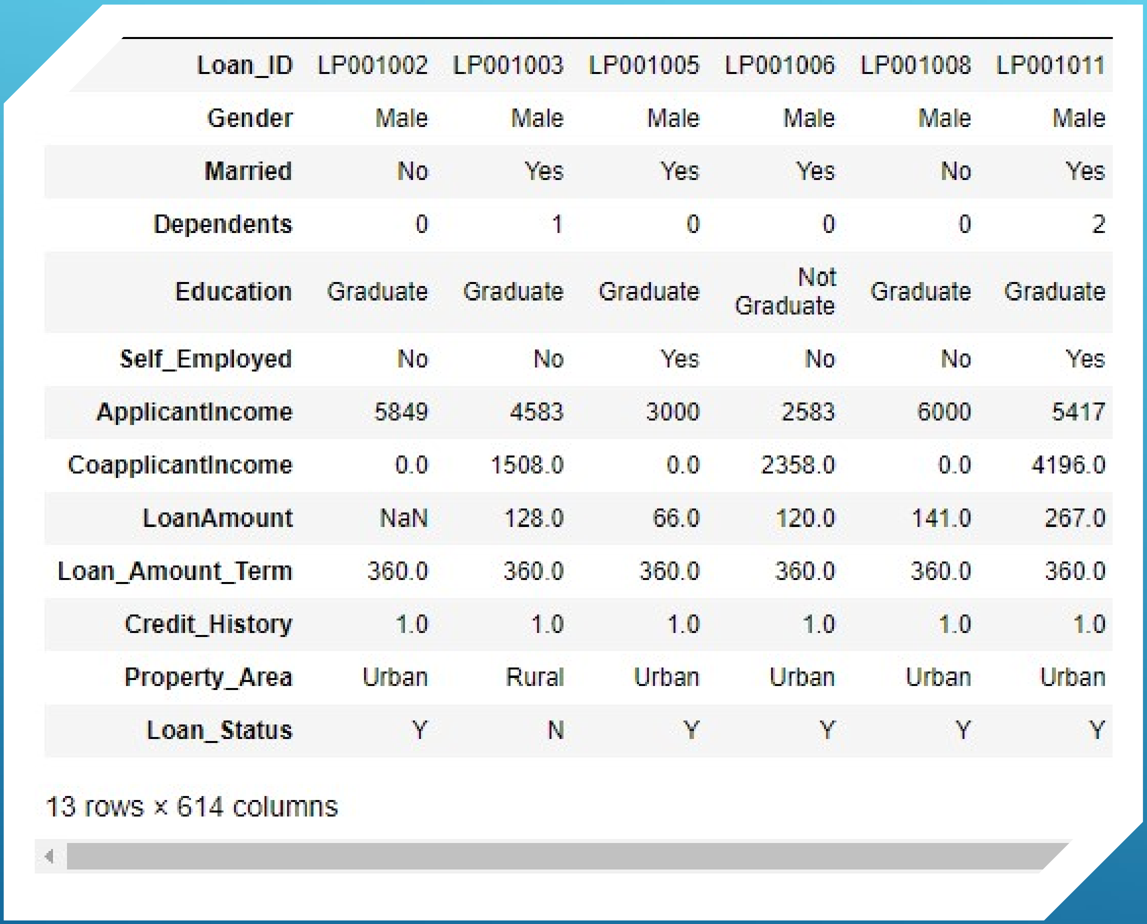
FACTORS WHICH AFFECT LOAN APPROVAL

- ❑ Credit Score
- ❑ Financial Profiling
- ❑ Education
- ❑ Demography
- ❑ Property Location

DATA ANALYSIS STEPS

- ❑ Data collection
- ❑ Data cleaning and Preprocessing: prepared the dataset for analysis.
- ❑ Exploratory Data Analysis(EDA): analyzed data sets to summarize their main characteristics using data visualization methods through univariate and bivariate analysis.
- ❑ Model building : used various Machine Learning algorithms
- ❑ Model performance and Evaluation: compared the various models and selected the best performing one.

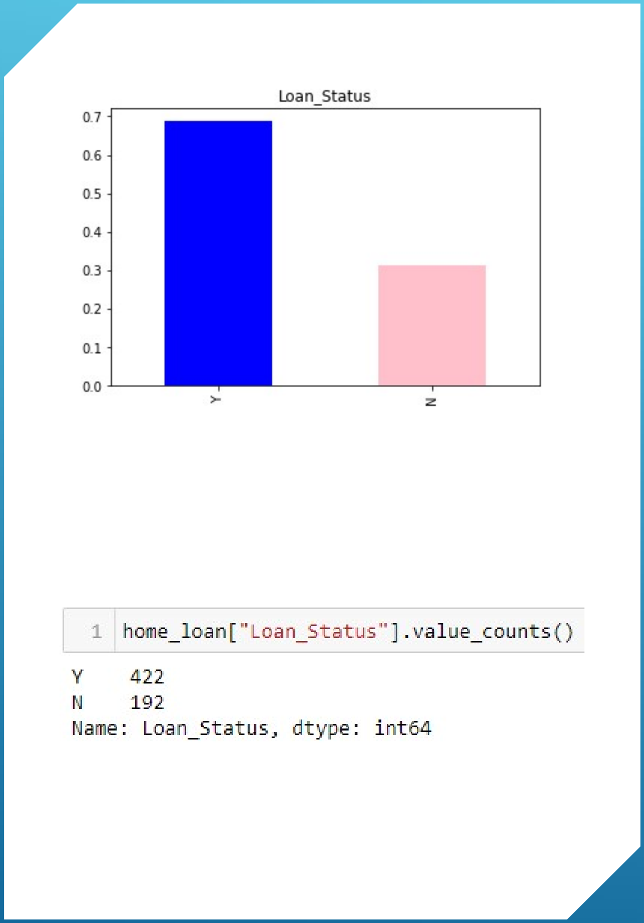
DATA COLLECTION



Loan_ID	LP001002	LP001003	LP001005	LP001006	LP001008	LP001011
Gender	Male	Male	Male	Male	Male	Male
Married	No	Yes	Yes	Yes	No	Yes
Dependents	0	1	0	0	0	2
Education	Graduate	Graduate	Graduate	Not Graduate	Graduate	Graduate
Self_Employed	No	No	Yes	No	No	Yes
ApplicantIncome	5849	4583	3000	2583	6000	5417
CoapplicantIncome	0.0	1508.0	0.0	2358.0	0.0	4196.0
LoanAmount	NaN	128.0	66.0	120.0	141.0	267.0
Loan_Amount_Term	360.0	360.0	360.0	360.0	360.0	360.0
Credit_History	1.0	1.0	1.0	1.0	1.0	1.0
Property_Area	Urban	Rural	Urban	Urban	Urban	Urban
Loan_Status	Y	N	Y	Y	Y	Y

13 rows × 614 columns

- ❑ Source: https://www.kaggle.com/sazid28/home-loan-prediction/data
- ❑ Number of records: 614
- ❑ Number of fields: 13
- ❑ Imbalanced Data

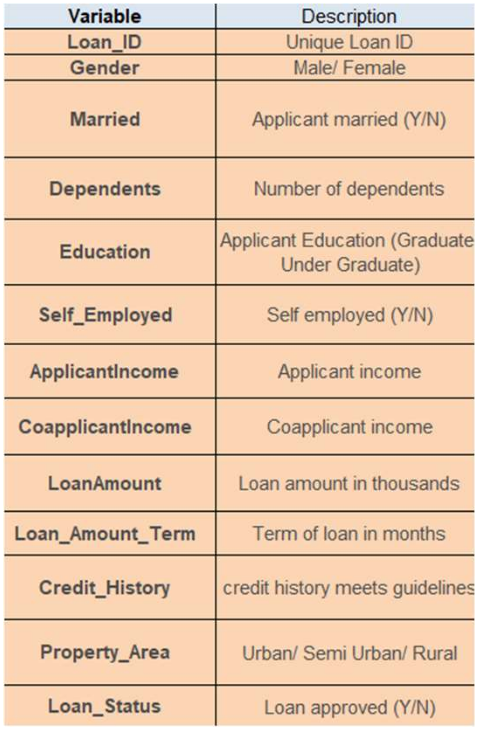


```
home_loan["Loan_Status"].value_counts()
```

```
Y    422
N    192
Name: Loan_Status, dtype: int64
```

IMBALANCED DATA ANALYSIS

- ❑ Based on the value counts of Approval(Yes-Y) and Rejection(No-N), we infer that our data is imbalanced.
- ❑ Out of total 614 applications, 422 loans were approved and 192 were rejected.



Variable	Description
Loan_ID	Unique Loan ID
Gender	Male/ Female
Married	Applicant married (Y/N)
Dependents	Number of dependents
Education	Applicant Education (Graduate Under Graduate)
Self_Employed	Self employed (Y/N)
ApplicantIncome	Applicant income
CoapplicantIncome	Coapplicant income
LoanAmount	Loan amount in thousands
Loan_Amount_Term	Term of loan in months
Credit_History	credit history meets guidelines
Property_Area	Urban/ Semi Urban/ Rural
Loan_Status	Loan approved (Y/N)

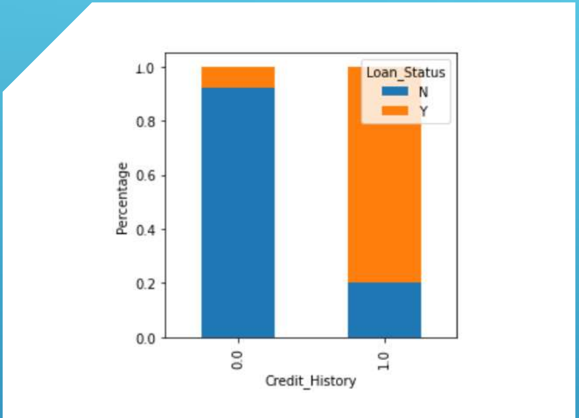
DATA INFORMATION

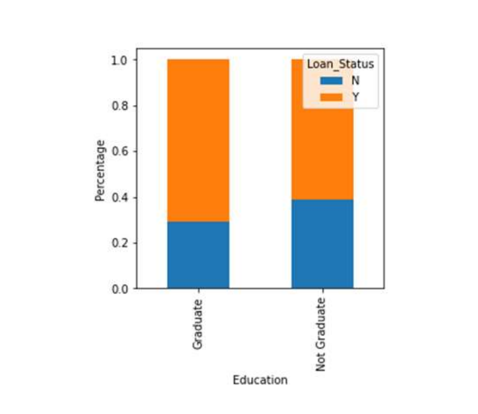
The table presents a brief description of all the 13 features (predictor variables and the target variables)in our dataset. Loan Status is the target variable, and the other 12 are the predictor variables. Majority of the loans are for 360 Months (30 years).

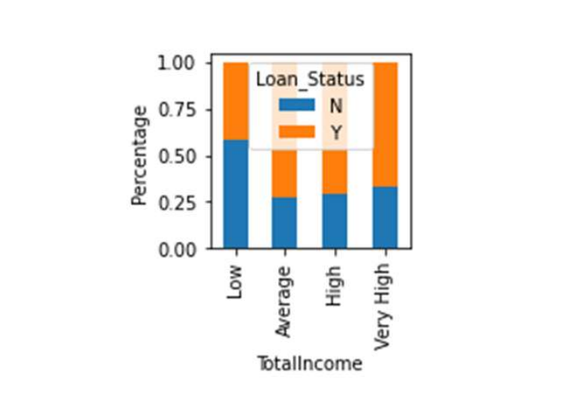
Through feature engineering, we added 4 new features (predictor variables) :

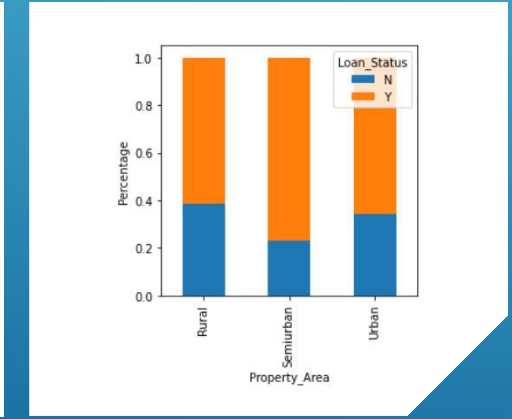
- ❑ Total Income
- ❑ EMI
- ❑ Balance Income
- ❑ Debt to Income (DTI)

EDA(BIVARIATE ANALYSIS)



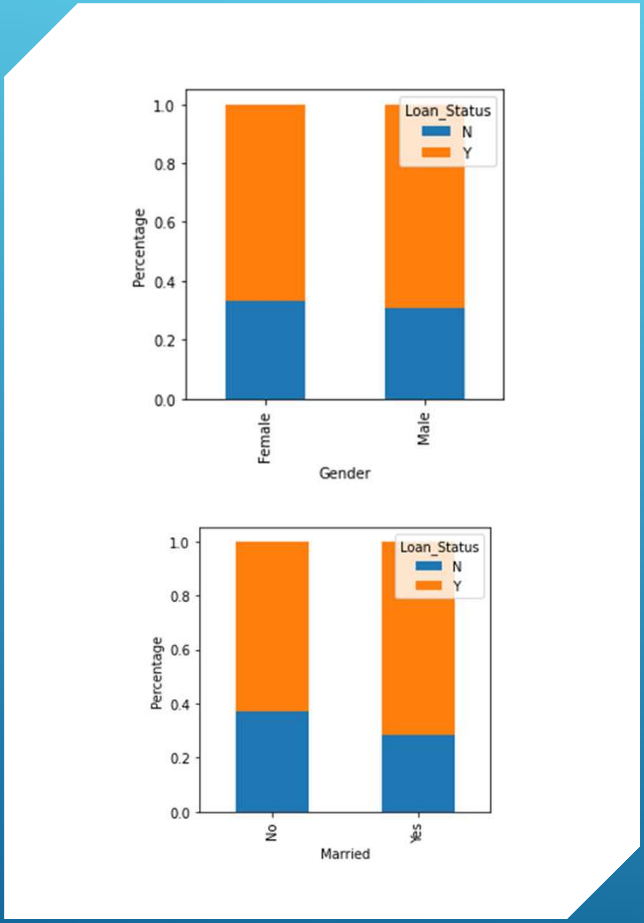






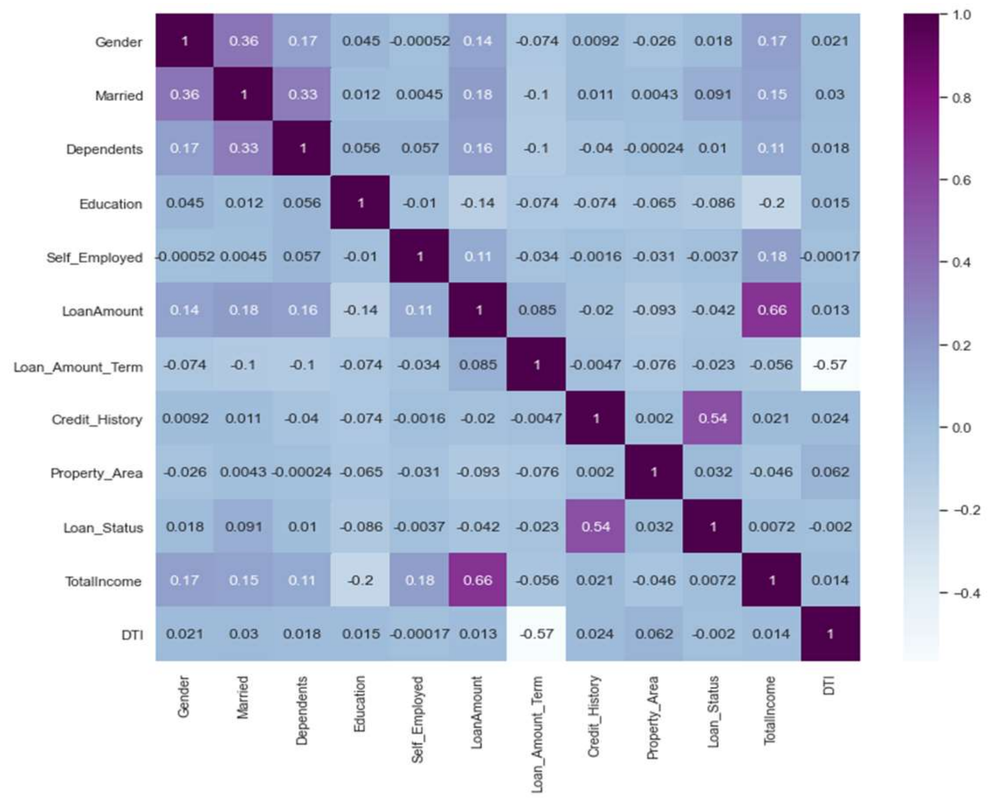
- Credit History: Applicant with good credit history are far more likely to be accepted.
- Education: About 5/6th of the population is a ‘Graduate’ and graduates have higher proportion of loan approval.
- Total Income: Applicants with higher total income are more likely to have loans approved.
- Property Area: More applicants are from Semi-urban and also more likely to be granted loans.

EDA(BIVARIATE ANALYSIS)



- ❑ Gender : There are more Male (81%) applicants than Female(19%). Males have an approval rate of around 69% whereas females have around 67%.
- ❑ Martial Status: 2/3rd of the population in the dataset is Married; Married applicants are more likely to be granted Home Loans.

CORRELATION HEATMAP



	Gender	Married	Dependents	Education	Self_Employed	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status	TotalIncome	DTI
Gender	1	0.36	0.17	0.045	-0.00052	0.14	-0.074	0.0092	-0.026	0.018	0.17	0.021
Married	0.36	1	0.33	0.012	0.0045	0.18	-0.1	0.011	0.0043	0.091	0.15	0.03
Dependents	0.17	0.33	1	0.056	0.057	0.16	-0.1	-0.04	-0.00024	0.01	0.11	0.018
Education	0.045	0.012	0.056	1	-0.01	-0.14	-0.074	-0.074	-0.065	-0.086	-0.2	0.015
Self_Employed	-0.00052	0.0045	0.057	-0.01	1	0.11	-0.034	-0.0016	-0.031	-0.0037	0.18	-0.00017
LoanAmount	0.14	0.18	0.16	-0.14	0.11	1	0.085	-0.02	-0.093	-0.042	0.66	0.013
Loan_Amount_Term	-0.074	-0.1	-0.1	-0.074	-0.034	0.085	1	-0.0047	-0.076	-0.023	-0.056	-0.57
Credit_History	0.0092	0.011	-0.04	-0.074	-0.0016	-0.02	-0.0047	1	0.002	0.54	0.021	0.024
Property_Area	-0.026	0.0043	-0.00024	-0.065	-0.031	-0.093	-0.076	0.002	1	0.032	-0.046	0.062
Loan_Status	0.018	0.091	0.01	-0.086	-0.0037	-0.042	-0.023	0.54	0.032	1	0.0072	-0.002
TotalIncome	0.17	0.15	0.11	-0.2	0.18	0.66	-0.056	0.021	-0.046	0.0072	1	0.014
DTI	0.021	0.03	0.018	0.015	-0.00017	0.013	-0.57	0.024	0.062	-0.002	0.014	1

From the heatmap, we can infer that:

- The target variable, Loan Status shows a positive correlation to applicants Credit history, marital status, total income and property area.
- Loan Status shows a negative correlation to DTI (Debt To Income) and Loan Amount

MODELING

- ❑ Supervised Machine Learning
- ❑ Binary classification
- ❑ Imbalanced Data
- ❑ Tools used: Python’s sklearn, pandas, numpy, matplotlib, seaborn in Jupyter notebook

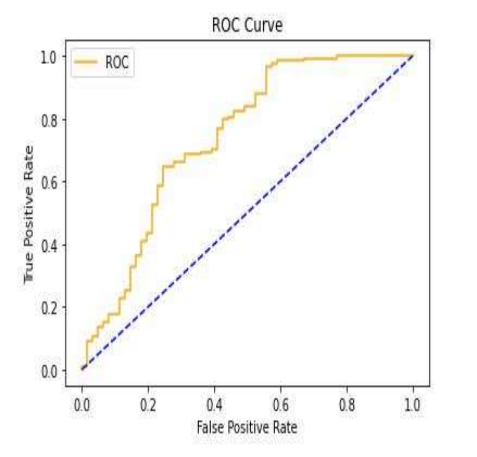
MODELING STEPS

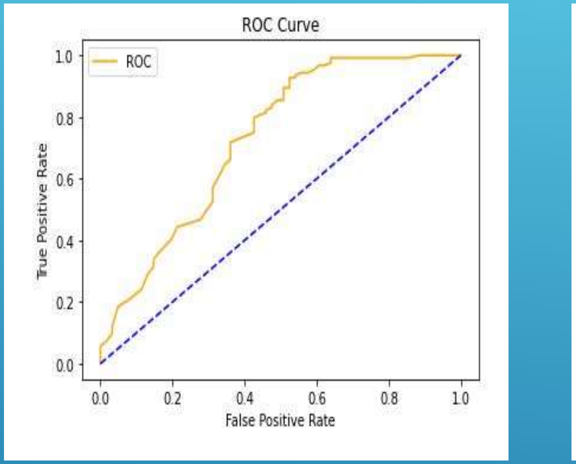
- ❑ Data Pre-processing
- ❑ Feature Engineering
- ❑ Train-Test Split (70/30) and Hyperparameter Tuning(5-fold Cross Validation)
- ❑ Classifier training using optimal parameters
- ❑ Fit the data
- ❑ Model Evaluation
- ❑ AUC/ROC score

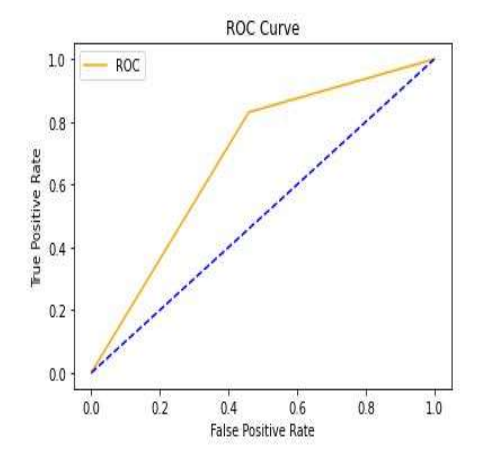
CLASSIFICATION ALGORITHMS USED

- ❑ Logistic Regression
- ❑ Random Forest
- ❑ Decision Tree

ROC-AUC CURVE COMPARISION







LOGISTIC REGRESSION

78.9%

RANDOM FOREST

76.7%

DECISION TREE

69.1%

MODELS	SCORE	ROC-AUC	ACCURACY
LOGISTIC REGRESSION	78.9%	72%	79.5%
RANDOM FOREST	76.7%	68.8%	78.9%
DECISION TREE	69.1%	67.7%	69.2%

MODEL COMPARISION

ROC(Reciever Operating Characteristic) Curves-AUC(Area Under Curve) scores for the dataset:

ROC-AUC scores were taken because of the imbalanced nature of the data. These score summarizes the curves and used to compare classifiers.

LOGISTIC REGRESSION: BEST FEATURES

- ❑ From the best performing model on our dataset, Logistic Regression, we found that 'Credit History', 'Total Income', 'DTI' , 'Education’, features are most important in predicting the target variable (Loan Status).

CONCLUSION

❑ Out of the three supervised classification models, Logistic Regression provided the best results compared to Random Forest and Decision Tree.

❑ 4 more features(predictor variables) were added to the dataset(through feature engineering) for detailed data analysis.

❑ Due to the imbalanced and limited nature of the data, the accuracy may not be of the correct measure.

❑ With more data and ideas, the model can be improved.

THANK YOU