

LOAN ELIGIBILITY PREDICTION

Data Science Capstone Project



By,

Lubna Rahman

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 and a more approachable 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: preparing the dataset for analysis and evaluation.
- ❑ Exploratory Data Analysis(EDA): Analyzing data using visualization methods through univariate and bivariate analysis.
- ❑ Model building and Training: using various Machine Learning algorithms to train the model.
- ❑ Model performance and Evaluation: comparing the various models and selecting the best performing one.

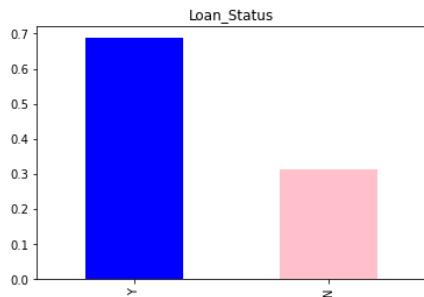
DATA COLLECTION

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

IMBALANCED DATA ANALYSIS



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

```
Y    422
```

```
N    192
```

```
Name: Loan_Status, dtype: int64
```

- ❑ 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.

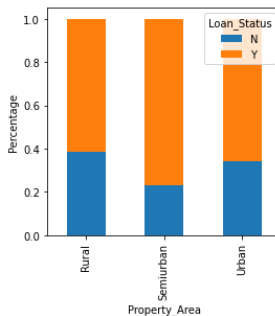
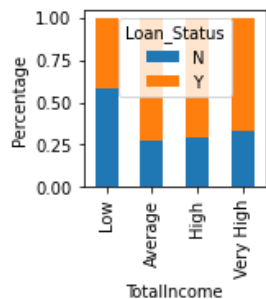
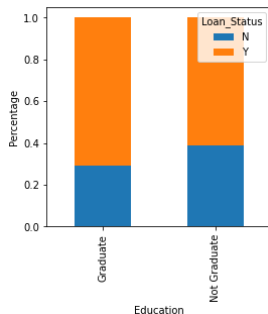
DATA INFORMATION

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)

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).

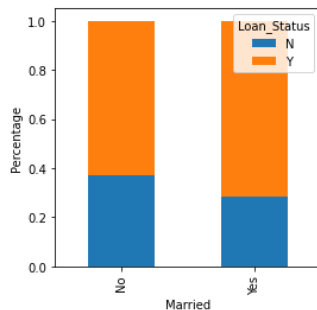
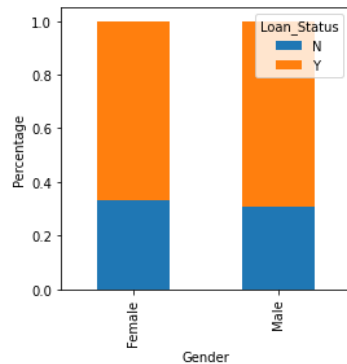
Through feature engineering, we added 4 new features (predictor variables) :

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



EDA(BIVARIATE ANALYSIS)

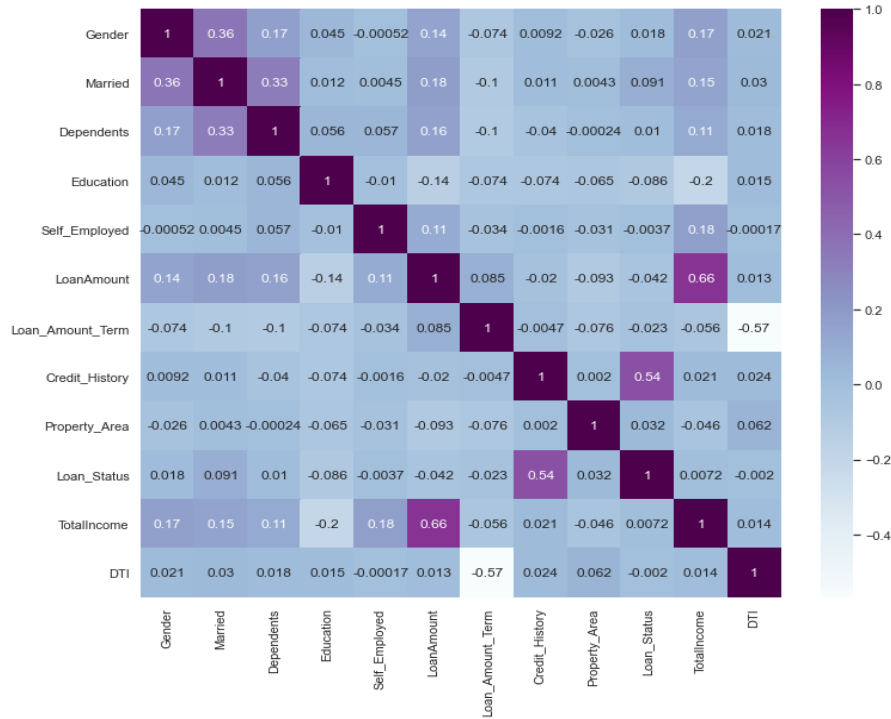
- ❑ Credit History: Applicant with good credit history are far more likely to be accepted.
- ❑ Education: About 5/6th of the population is a 'Graduate' and graduates have higher proportion of loan approval.
- ❑ Total Income: Applicants with higher total income are more likely to have loans approved.
- ❑ 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



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 scikit-learn, pandas, NumPy, Matplotlib, Seaborn in Jupyter notebook.

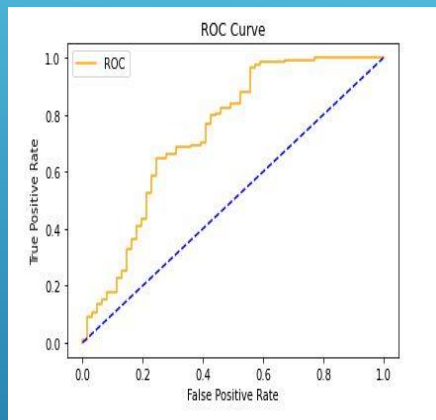
MODELING STEPS

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

CLASSIFICATION ALGORITHMS

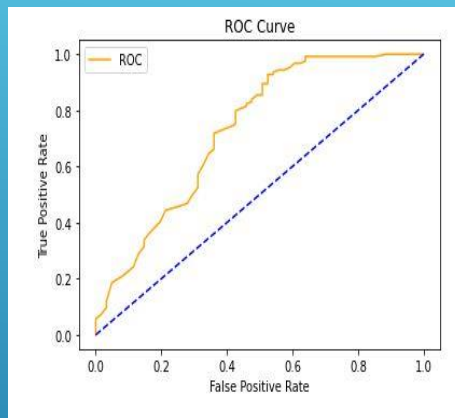
- ❑ Logistic Regression
- ❑ Random Forest
- ❑ Decision Tree

ROC-AUC CURVE COMPARISON



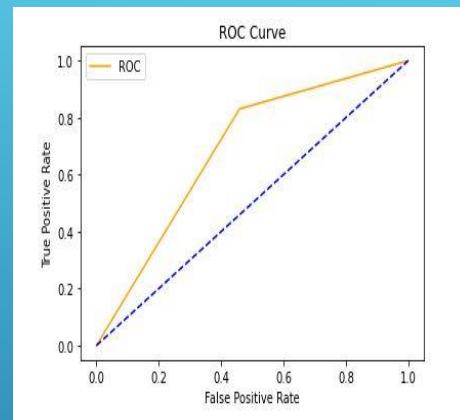
LOGISTIC REGRESSION

78.9%



RANDOM FOREST

76.7%



DECISION TREE

69.1%

MODEL COMPARISON

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%

- ❑ The ROC(Receiver Operating Characteristic) Curves-AUC(Area Under Curve) scores comparison for the dataset.
- ❑ ROC-AUC scores were taken due to the imbalanced nature of the data.

LOGISTIC REGRESSION: BEST FEATURES

The most important features in predicting the Loan Status were found to be:

- ❑ Credit History,
- ❑ Total Income,
- ❑ Debt-to-Income(DTI), and
- ❑ Education

CONCLUSION

- ❑ Logistic Regression provided the best results compared to Random Forest and Decision Tree.
- ❑ 4 new features 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 in future.



THANK YOU