Sean R Grant

Home Loan Prediction

Dream Housing Finance

Dream Home Financing is a company that is dedicated to helping consumers to find the right loan program and lender to suit their needs. They have presence across all urban, semi urban and rural areas. The application process requires that the customer's eligibility for the loan is validated first. Dream Home Financing wants to automate the loan eligibility process (real time) based on customer details provided while in an online application form.

Problem Statement

How can automating the loan eligibility process identify highly qualified customers?

How can automation help with marketing by targeting potential customers?

How can eligibility automation reduce risk and increase profitability?

Data Source and Description

The data for this project was acquired from,

https://datahack.analyticsvidhya.com/contest/practice-problem-loan-prediction-iii/

The original data contains 614 applications and 13 features (LoanID, Gender, Married, Self Employed, Education, Dependents, Applicant Income, Coapplicant Income, Loan Amount, Loan Amount Term, Property Area, Credit History, and Loan Status)

Feature Descriptions

LoanID – Unique code that identifies the loan application

Gender – Male/Female

Marital Status – Married (Yes/No)

Self Employed – Self Employed (Yes/No)

Education – Graduate/Undergraduate

Dependents - Number of Dependents

Applicant Income – Main Applicant Income

Coapplicant Income – Co-applicant Income

Loan Amount – Borrowing amount in thousands

Loan Amount Term – Length of loan in months

Property Area – Location of property under consideration

Credit History – Having good credit history

Loan Status – Status of the application (Y/N)

Challenge: Treating Missing Data

Feature	Number of Missing Data	Handling	Reason
Gender	13	Male	81% of applicants are males
Married	3	Yes	65% of applicants are married
Dependents	15	0	53%, 0; 17%, 1; 17%, 2; 9%, 3+
Self Employed	32	Not Self Employed	86% Not Self Employed
Loan Amount	22	Median	Avoid the effects outliers
Loan Amount Term	14	360 months	83% borrowed under a 360 month term and loan amounts were $77k+
Credit History	50	Approved loans were filled as 1.0 Denied loans were filled as 0.0	81% of approved loans had credit history

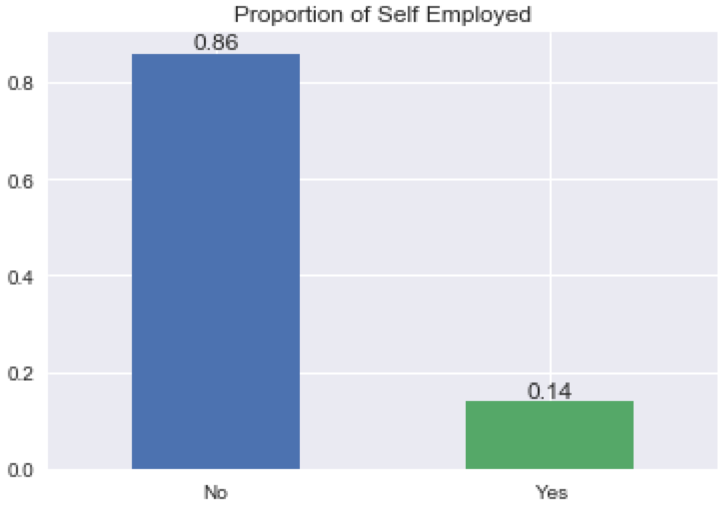
Loan Approval Rate

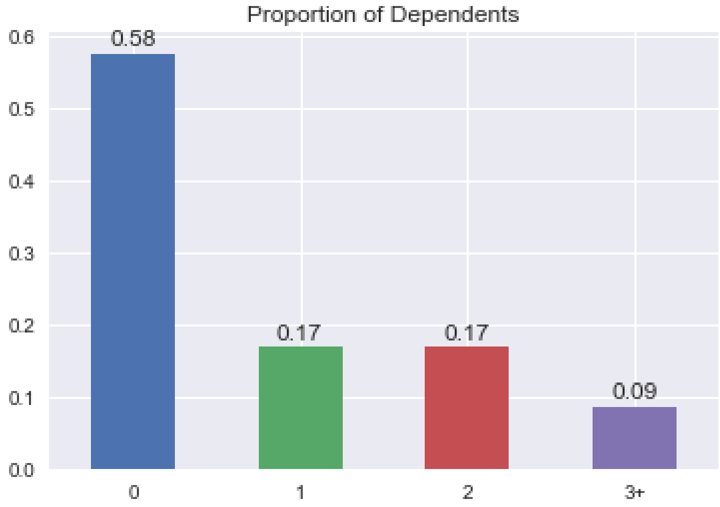
Hypothesis

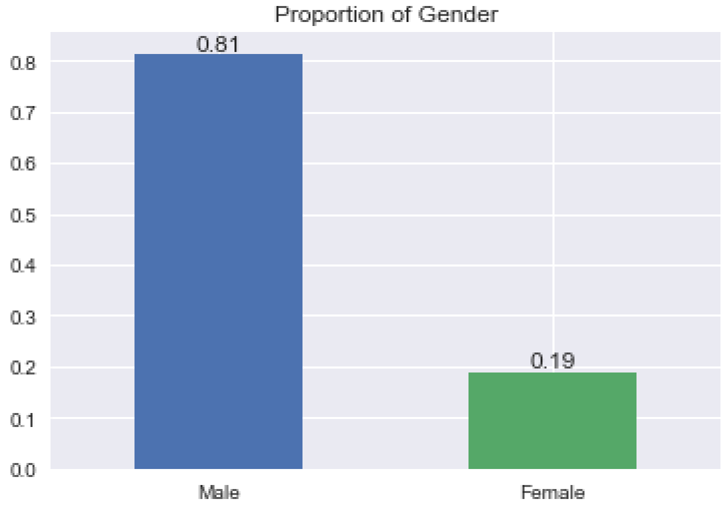
Factors that may affect the approval of a loan are the factors worth considering when determining the hypothesis. Here are a few,

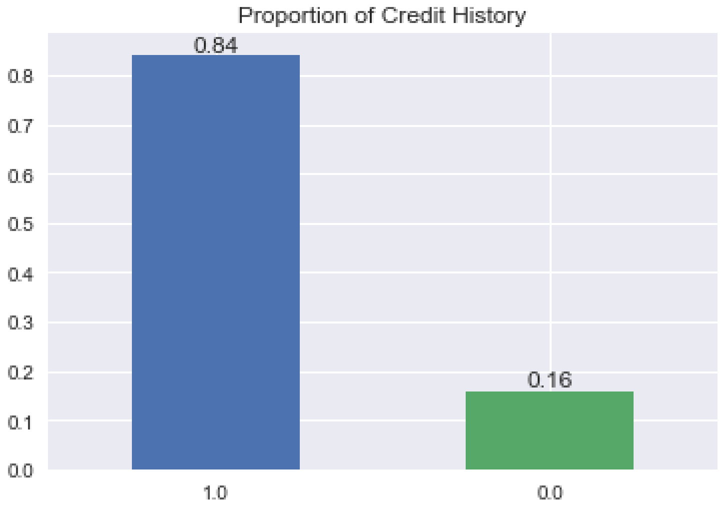
- Income: The total income (combined applicant and coapplicant) is a determining factor. The higher the total income the higher the chances of being approved.
- Credit History: Having a history of previous loans help to improve approval possibility because it indicates loan worthiness.
- Monthly payments: The lower to monthly payment to income the greater the chances of approval.
- Term: The shorter the term, the chances of approval increases however this is also dependent previously mentioned factors.
- Loan amount: The lower the loan amount the greater the chance of approval.

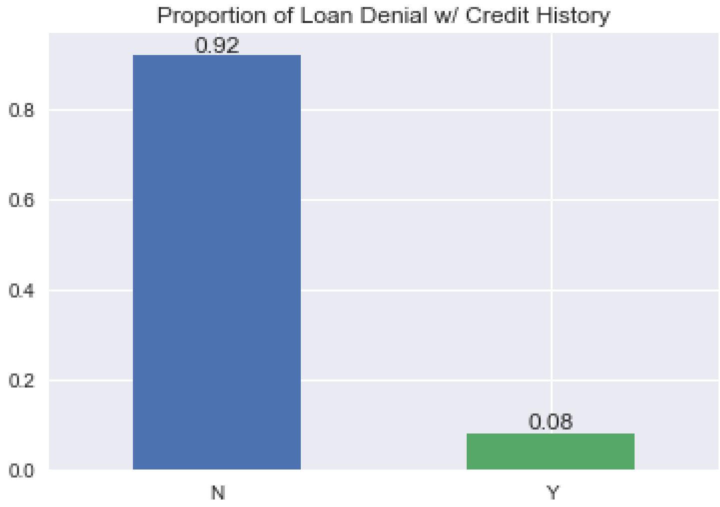
Feature Proportions

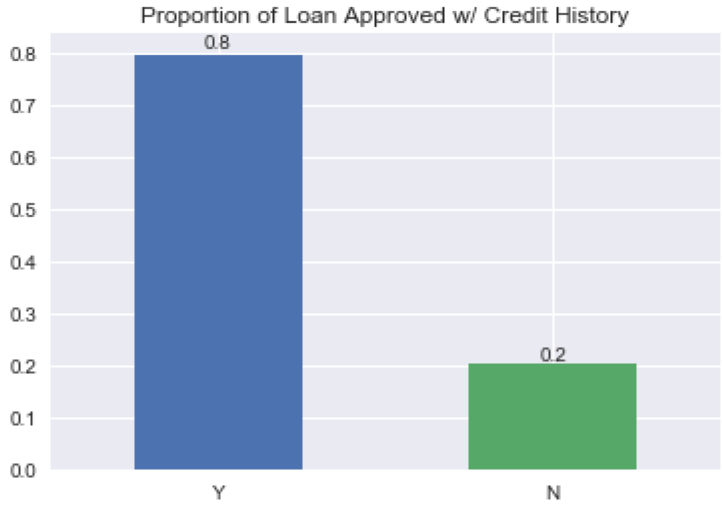




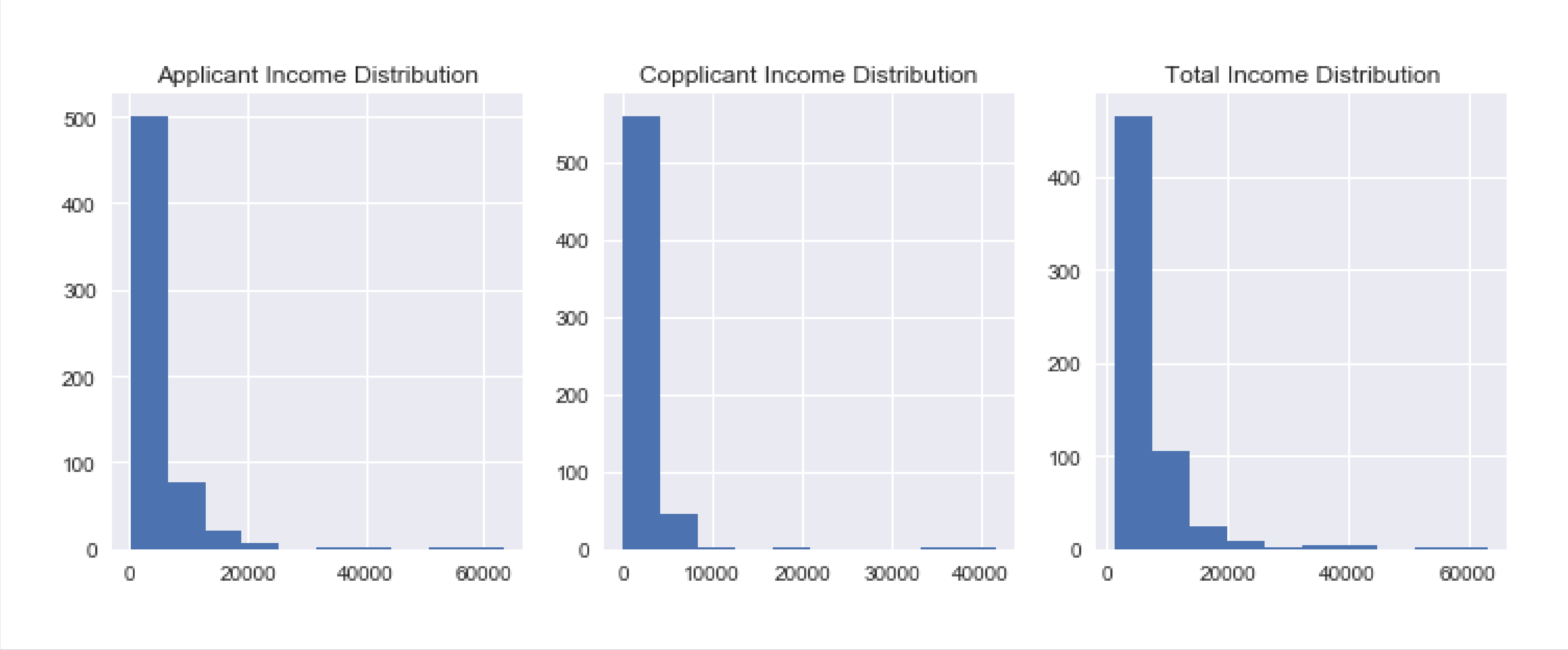








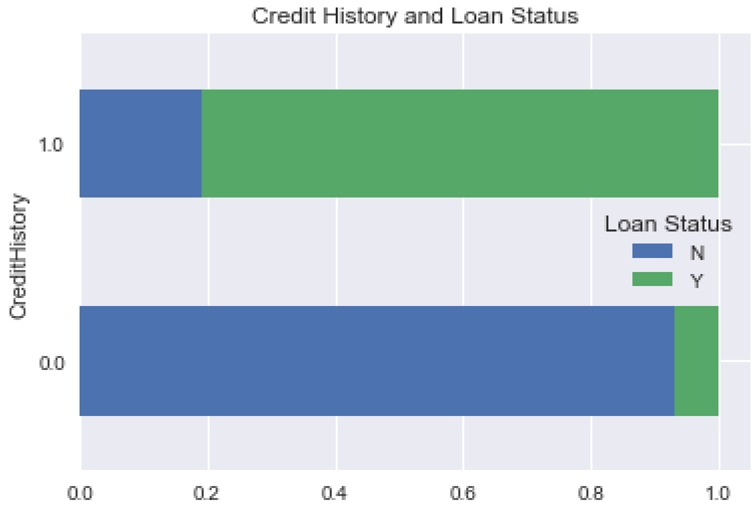
Income Distribution



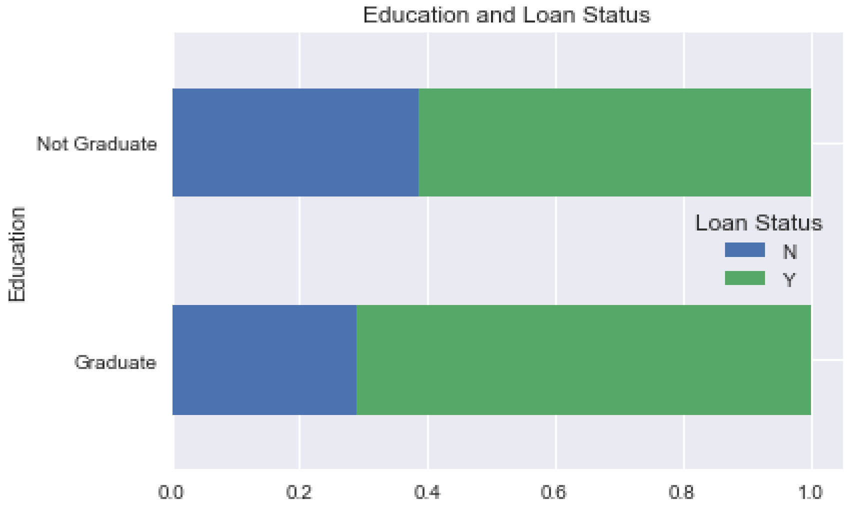
Loan Amount and Total Income Correlation

Statistically and Practically Significant

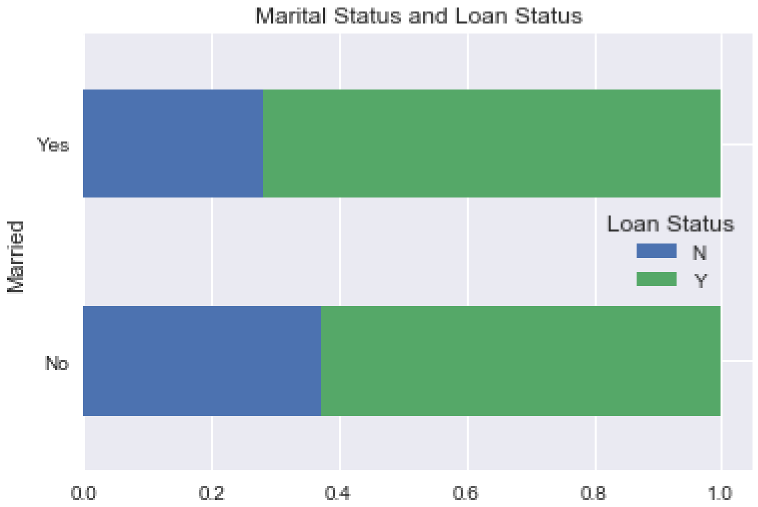
P-value = 0.0025



P-value = 4.61e-48



P-value = 0.04



P-value = 0.026

Model Selection

The Approach

Test key features using supervised learning algorithm; Linear Regression, SVM, Random Forest, Naïve Bayes.

Using the Linear model to evaluate the effects of each key feature using coefficient values.

Identify the best model after understanding the model results.

	Precision	Recall	Accuracy
Logistic Regression	0.791	1.0	0.834
Random Forest	0.833	0.979	0.855
Naives Bayes	0.784	1.0	0.828
SVM	0.784	1.0	0.828

Data Cleaning Done:

Removing rows with missing data for Loan Amount, Dependents, and Credit History.

Self-Employed if income is above $7,000 monthly.

Loan term 360.

Challenges:

This may not result in a generalize model.

ML Model Results

Missing Data Handling

Feature	Number of Missing Data	Handling	Reason
Gender	13	Male	81% of applicants are males
Married	3	Yes	65% of applicants are married
Dependents	15	0	53%, 0; 17%, 1; 17%, 2; 9%, 3+
Self Employed	32	Not Self Employed Self Employed for incomes over $7k	86% Not Self Employed Higher salaries for Self-Employed
Loan Amount	22	Median	Avoid the effects outliers
Loan Amount Term	14	360 months	83% borrowed under a 360 month term
Credit History	50	Approved loans were filled as 1.0 Denied loans were filled as 0.0	81% of approved loans had credit history

	Precision	Recall	Accuracy
Logistic Regression	0.845	0.973	0.851
Random Forest	0.862	0.946	0.851
Naives Bayes	0.852	0.973	0.857
SVM	0.84	0.982	0.851

Advantage: A model that is more generalize and therefore can handle new data well.

ML Model Results

	Precision	Recall	Accuracy
Logistic Regression	0.845	0.973	0.851
Random Forest	0.862	0.946	0.851
Naives Bayes	0.852	0.973	0.857
SVM	0.846	0.982	0.857

Feature engineering: Created logs for TotalIncome and MonthPaymentNoInterest

ML Model Results

ML Model Results

	Precision	Recall	Accuracy
Logistic Regression	0.846	0.982	0.857
Random Forest	0.846	0.982	0.857
Naives Bayes	0.846	0.982	0.857
SVM	0.846	0.982	0.857

C optimization and Feature selection based on coef of Linear model

GridSearchCV for RandomForest model

Conclusion

Naïve Bayes seem to the model that is best fitted for generalization. Therefore it is the model I would choose for application to the problem.

Recommendations

Gather additional data based on current debt and savings.

- Helps with identifying highly qualified customers.
- This will better estimate the likelihood of repayment, preventing defaults.