Capstone Project One

Loan Approval Prediction

Sean Grant (Sept 2018 cohort)

Table of Contents

Introduction

Data Set Overview and Hypothesis

Data Wrangling

Exploratory Data Analysis

Machine Learning

Introduction

The American dream is to own a structure and call it home. There are many steps to the home buy process and each can be very time consuming. Most Americans don’t own a home because of the expense that accompanies and therefore another person’s money must be used the secure it. This is where securing a loan adds the to the process. 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. The two most critical questions in the lending industry are: 1) What is the risk level for the borrower? 2) After ascertaining the borrower’s risk, should we offer him/her a loan? Dream Home Financing wants to automate the loan eligibility process (real time) based on customer details provided while in an online application form. 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 Set Overview

The data was obtained from <https://www.analyticsvidhya.com/>. Data can be found here, h[ttps://goo.gl/6EsyqC](https://goo.gl/6EsyqC).

The original data contains 614 applications and 13 features (12 independent and 1 target)

**Features 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)

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.

Data Wrangling

The data was initially cleaned and wrangled using the following steps:

* For the ‘Dependents’ variable, the “3+” entry was simplified to “3” and then coerced to an integer format.
* Loan amount term is in months.
* Loan amount was assumed to be in thousands, this was concluded since there’s no logics in approval for a $98 loan with a term of 360 months and income of $3127, see LoanID LP002502.
* Income is entered as monthly (91% Applicant Income and 99% Co-applicant Incomes are below $10,000)
* Rename features for consistence by removing underscores
* The missing values for the ‘LoanAmountTerm’ variable was replaced by 360, 83% of all applications request a term of 360 and these loans range from $9,000 to $600,000.
* The missing values for the ‘SelfEmployed’ feature was replaced by ‘No’ for individuals with a monthly income less than $10,000 except for a rural applicant whose income is $7,333 (he is likely self-employed). Self-employed individuals normally earn a greater salary hence applicant incomes above $10,000 will be assigned ‘Yes.
* Dropped rows with missing loan amount and dependents, 94% of the data remained. A summary of the data before and after reveals that there isn’t much differences between the means for each numerical variable.
* Create features;
  + ‘TotalIncome’ (Applicant + Co-applicant),
  + ‘MonthlyPaymentNoInterest’ (‘LoanAmount’\*1000/’LoanTerm’) and
  + ‘LoantoIncomeRatio’ (‘MonthlyPaymentNoInterest’/’TotalIncome’)
* The LoanID LP002317 and LP001448 had a similar application with incomes of 81,000 and 23,803. LP002317 was denied a loan therefore 81,000 could possibly be an outlier that can safely be removed. It could be a yearly income, if this is the case the monthly income would be $6,750.

This process of wrangling and cleaning resulted in 578 observations remain. However, after the machine learning stage a few changes were made to how the missing values were treated.

* Loan Amount was replaced by the median.
* Dependents were replaced with 0, most applications had no dependents.
* Approved applications were given 1.0 for CreditHistory, however 0.0 were given for the applications that were denied.
* Columns were created for the log of TotalIncome and MonthlyPaymentNoInterest.

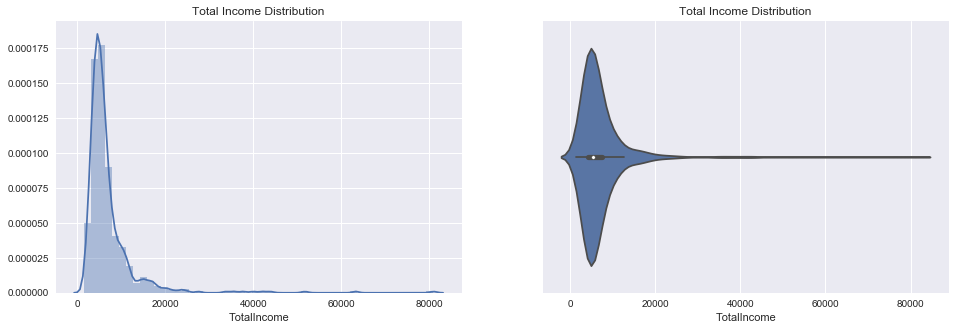
This resulted in 613 remaining applications.

Exploratory Data Analysis

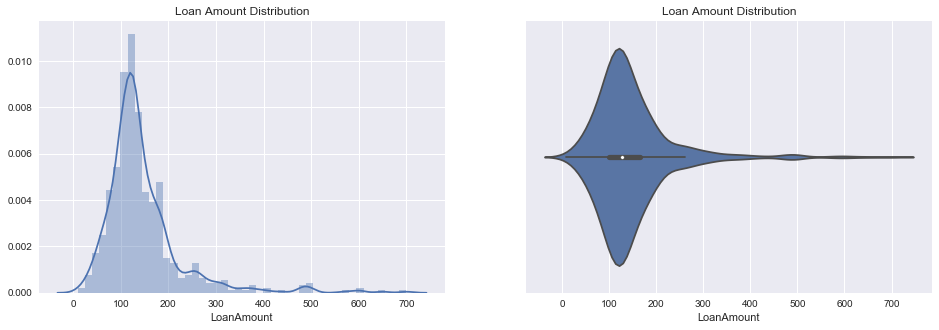
The variables in the data includes 2 quantitative and 7 categorical. For the quantitative data a correlation was evaluated using the linear regression model. While for the categorical data, a chi-square test was applied and for those variables where visual inspection may lead to an incorrect assumption a proportion z-test will be completed.



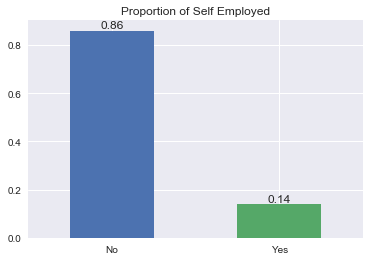
In the plot below, it can be seen that the rate of loan approval is approximately 70%.

****

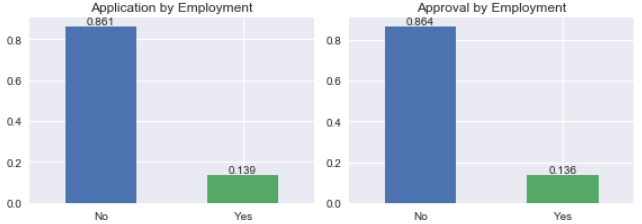
The income of the applicants is concentrated below 10,000 and extends beyond 80,000.

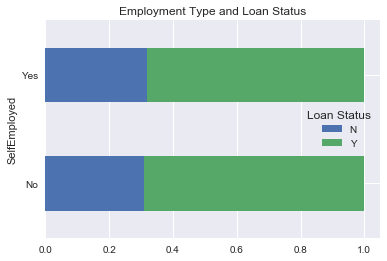
****

Loan amounts is concentrated between 50k and 200k and extends to a much as 750k

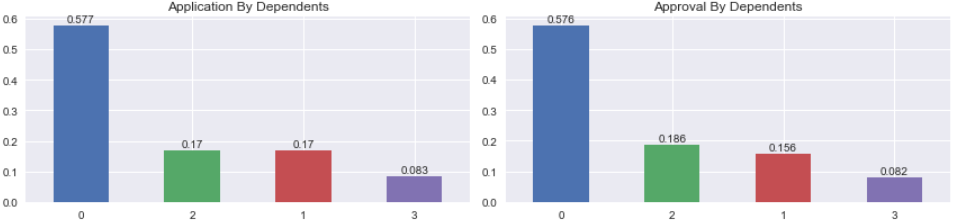
****

The portion of non-self-employed was expected, being greater than the amount of self-employed, more workers than employers.

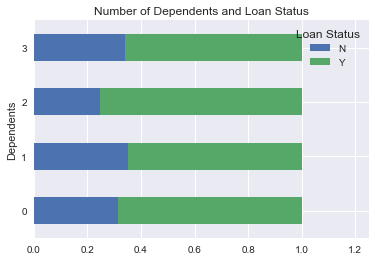


****

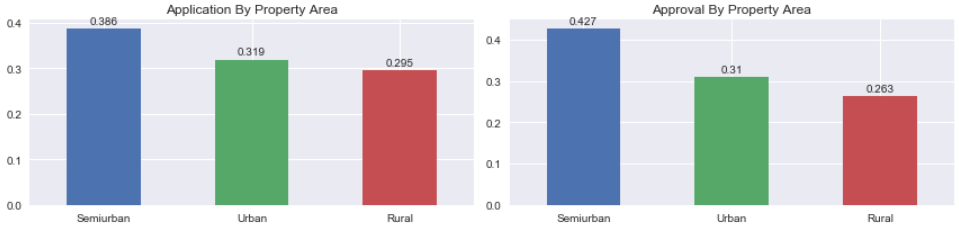
The comparison of loan outcomes for this categorical (SelfEmployed) variable reveals that the rate of approval is approximately the same. Most individuals try to acquire homes of value they can reasonable sustain/afford, if you work more you can afford more that would result in asking for a loan amount that fits your income.

****

The rates of application and approval are indeed similar. Therefore, the number of dependents may not be taken into consideration when approving or denying a loan application.



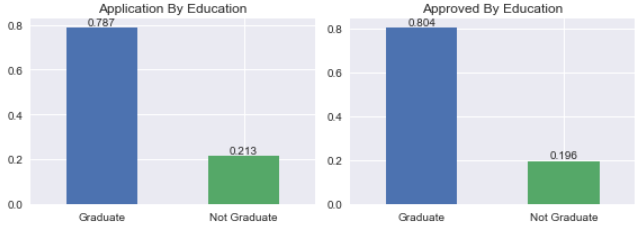
In the graph above we see that the rate of approval for applications with no dependents and 3 dependents are quite similar. This could lead to the conclusion that the number of dependents isn’t taken into consideration. However, I’m leaning on the side that having dependents is important, this data set does include applicant’s debt and having more dependents results in having less for loan repayment.



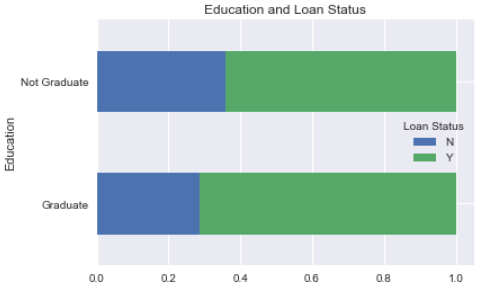
A greater portion of applicants are for properties in the Semi-urban (Suburbs). This can be a result of urban areas becoming more populated giving rise to more developments outside urban limits.



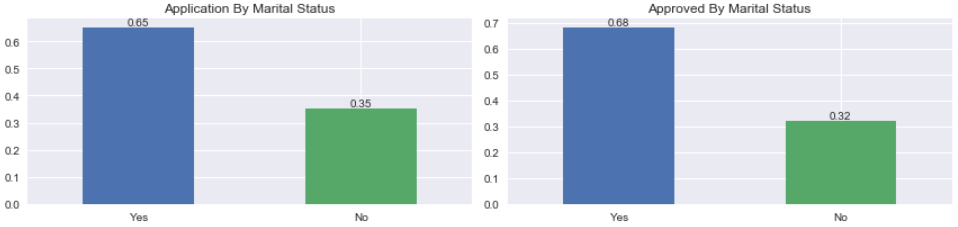
There’s a greater probability of receiving a loan if the property is in the semi-urban area. A hypothesis is that these areas have lower risk, owing to the fact that they are more likely to be detached family homes.

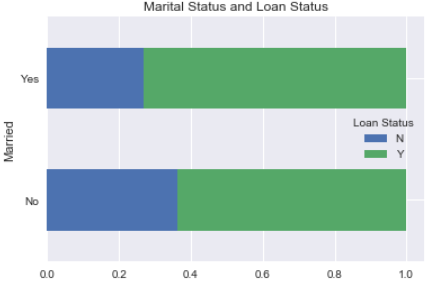


A greater portion of applicants has a graduate degree. There is relationship between the salary and education of an individual, higher degree higher salary. Because of this you will find that most graduates can afford a loan which will impact the comparative approval rate as see below. There’s a greater probability of getting a loan approved with a higher education.

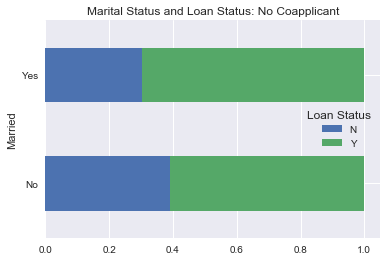


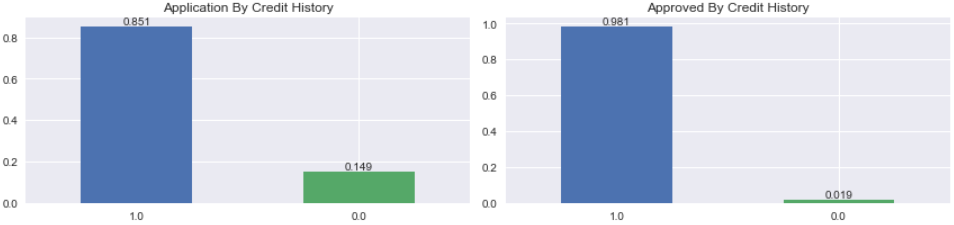
At the time of buying a home most people are at the point of starting a family and therefore it may affect the ratio of married and not married (single or divorce). It is twice a likely that an applicant is married as revealed below.



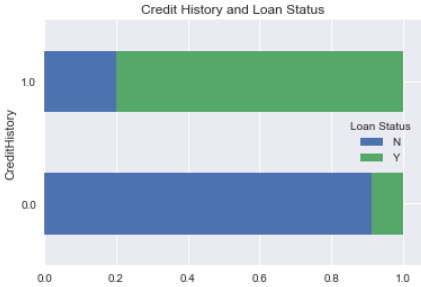


It is more likely that married applicants may have their loans approved, since each spouse is willing to apply jointly improving their chances due to a greater combined income. However, after further evaluation of applications without a co-applicant suggest that it could be because of marital status.





Having a previous credit history (1.0) does increase the likelihood of receiving an approval. Creditors are more comfortable with someone who has already proven their credit worthiness.



There’s an extremely probability of getting a loan approved with a credit history.

**Machine Learning**

This problem seeks to identify whether a loan an be approved or not, therefore it will require the application of a binary classification algorithm. Key features were tested using supervised learning; Linear Regression, SVM, Random Forest, Naïve Bayes.

I took the data as was first wrangled and clean and applied the models using the following features: 'Married', 'Dependents', 'Education', 'SelfEmployed', 'LoanAmount', 'LoanAmountTerm', 'CreditHistory', 'PropertyArea', 'TotalIncome', 'MonthlyPaymentNoInterest'.

This resulted in the results as shown in the table below.

|  |  |  |  |
| --- | --- | --- | --- |
| Logistic Regression | Random Forest | Naïve Bayes | SVM |
| Precision: 0.791  Recall : 1.0  Accuracy : 0.834 | Precision: 0.833  Recall : 0.979  Accuracy : 0.855 | Precision: 0.784  Recall : 1.0  Accuracy : 0.828 | Precision: 0.784  Recall : 1.0  Accuracy : 0.828 |

Instead of removing instances the missing data was replaced by appropriate values, result is summaries in the following table.

|  |  |  |  |
| --- | --- | --- | --- |
| Logistic Regression | Random Forest | Naïve Bayes | SVM |
| Precision: 0.845  Recall : 0.973  Accuracy : 0.851 | Precision: 0.862  Recall : 0.946  Accuracy : 0.851 | Precision: 0.852  Recall : 0.973  Accuracy : 0.857 | Precision: 0.84  Recall : 0.982  Accuracy : 0.851 |

Two features were added, TotalIncome\_log and MonthlyPaymentNoInterest\_log, to treat outliers. Additionally, C optimization and feature selection (based on coef) of Linear Regression model and GridSearchCV for RandomForest model was done. These techniques resulted in accuracies as shown below.

|  |  |  |  |
| --- | --- | --- | --- |
| Logistic Regression | Random Forest | Naïve Bayes | SVM |
| Precision: 0.846  Recall : 0.982  Accuracy : 0.857 | Precision: 0.846  Recall : 0.982  Accuracy : 0.857 | Precision: 0.846  Recall : 0.982  Accuracy : 0.857 | Precision: 0.846  Recall : 0.982  Accuracy : 0.857 |

Naïve Bayes seems like the best model to the generalization purpose.