Capstone Project 1: Milestone Report

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 Description

The original data contains 614 instances/observations and 13 features/variables (LoanID, Gender, Marital Status, Self Employed, Education, Number of Dependents, Applicant Income, Coapplicant Income, Loan Amount, Loan Amount Term, Property Area, Credit History, and Loan Status)

The data was obtained from https://www.analyticsvidhya.com/. Data can be found here, ttps://goo.gl/6EsyqC.

The data was cleaned and wrangled using the following steps:

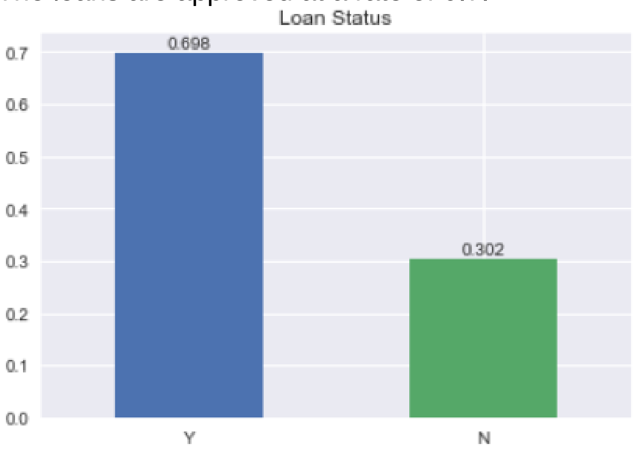
- For the 'Dependents' variable, the "3+" entry was simplified to "3" and then coerced to an integer format.
- Loan amount term is in months.
- Loan amount was assumed to be in thousands, this was concluded since there's no logics in approval for a $98 loan with a term of 360 months and income of $3127, 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 will be 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.
- Drop 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.
- The LoanID LP002588 monthly income is lower than the expected no interest monthly payment and therefore it will be removed.

After wrangling and cleaning we were left with 578 observations and 16 features.

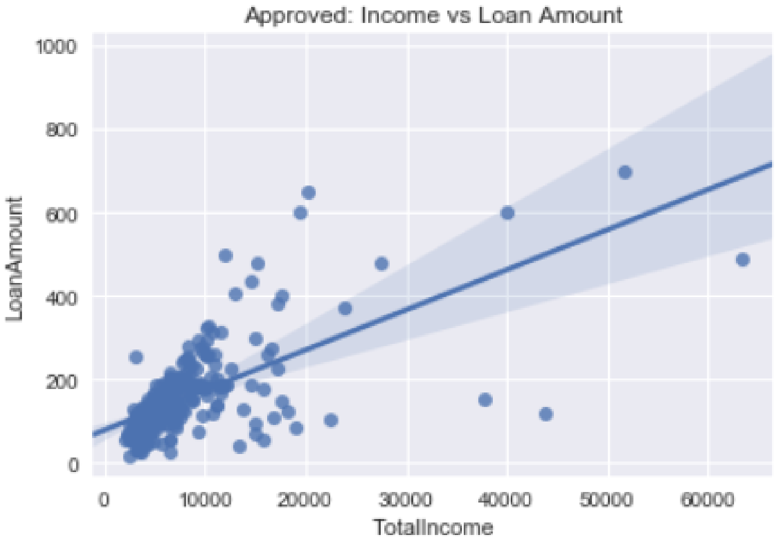
Summary, Visuals, and Statistics of findings.

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

What is the rate of loan approval?
The loans are approved at a rate of 0.7.

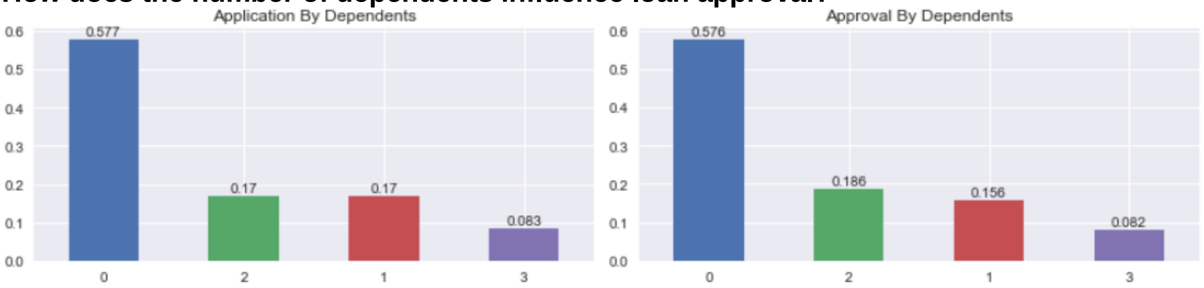


What's the correlation between loan amount and total income?
H_o: There is no linear relationship between total income and loan amount.
H_A: There is a linear relationship between total income and loan amount.



The correlation coefficient of our observed data is r = 0.657 and the p-value is 4.7e-51.
With a p-value that is infinitesimally below 0.05 there is very strong evidence to reject H_O for H_A.

How does the number of dependents influence loan approval?

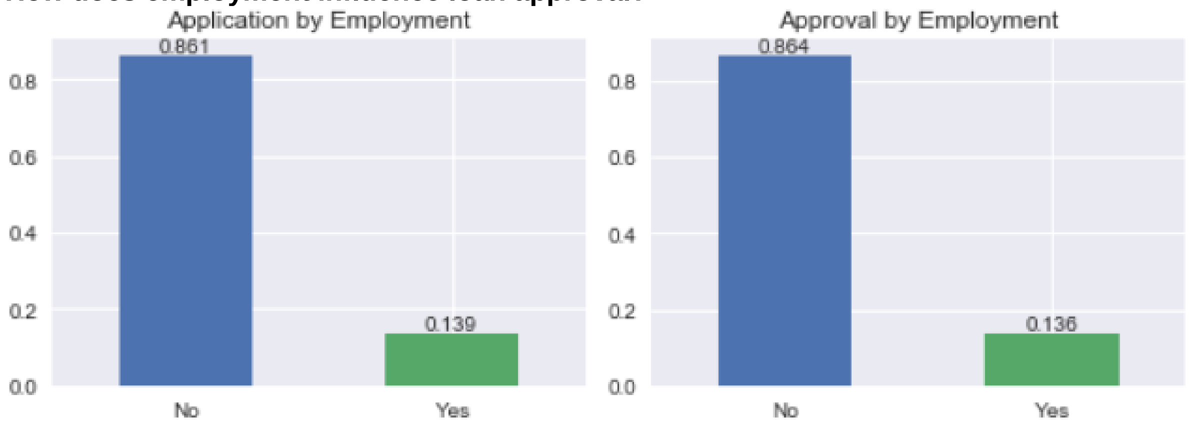


The rates of application and approval are indeed similar. Therefore, the number of dependents may not be taken into consideration when approving or denying a loan application.

H_o: Loan approval is not related to the number of dependents.
H_A: Loan approval is related to the number of dependents.

The chi-square statistics is 3.55 and the p-value is 0.314. With a p-value (0.314) that is above 0.05 there isn't enough evidence to reject H_O for H_A.

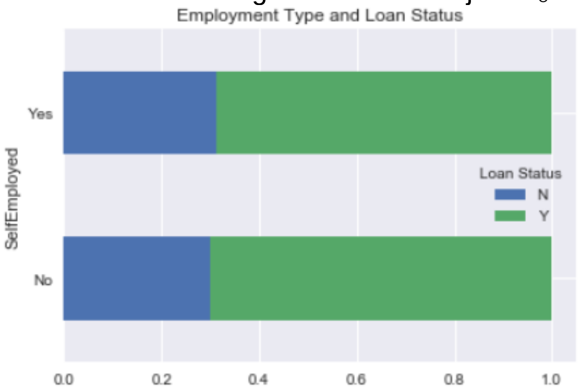
How does employment influence loan approval?



The rates of application and approval are indeed similar. Therefore, self-employment status may not be taken into consideration when approving or denying a loan application.

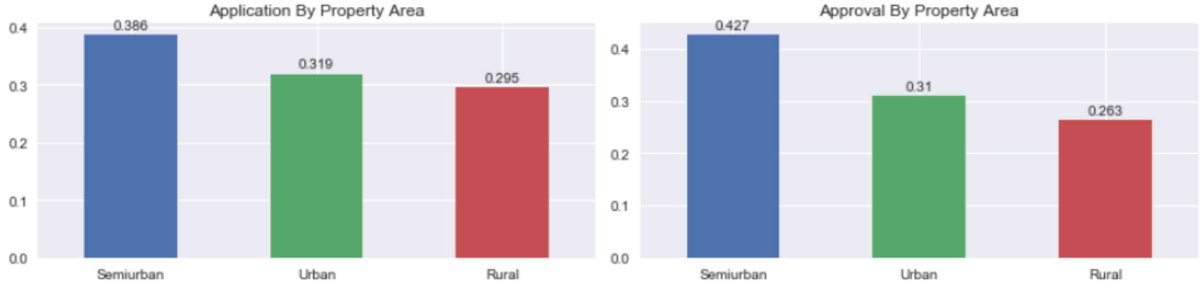
H_o: There is no relationship between loan approval and employment type.
H_A: There is a relationship between loan approval and employment type.

The chi-square statistics is 0.01 and the p-value is 0.922. With a p-value (0.922) that is above 0.05 there isn't enough evidence to reject H_O for H_A.



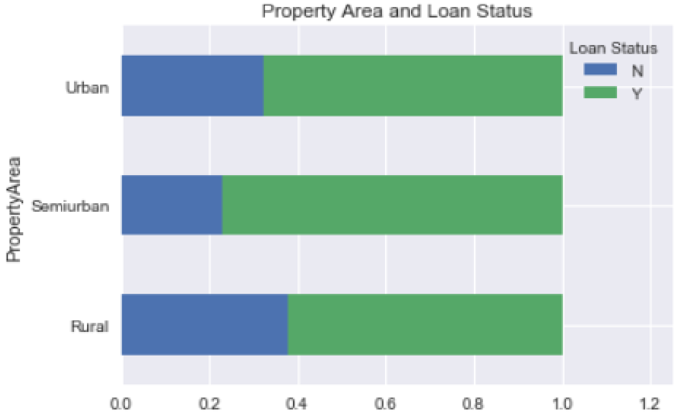
The probability of receiving a loan is similar for both types of employment.

Does location affect loan approval rate?



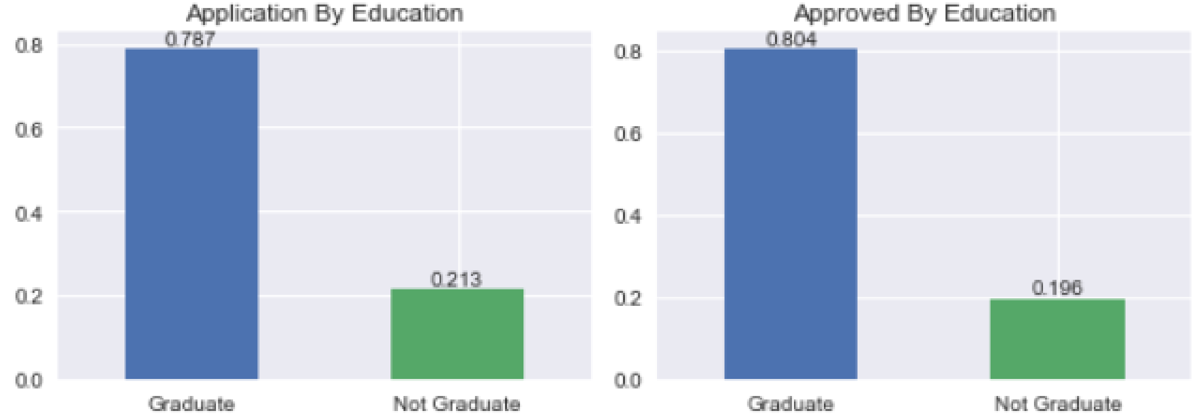
The ratio of approved (0.427) semi-urban loans is greater than the ratio of semi-urban applicants (0.386). This can lead to the conclusion that your location can influence your application.
H_o: The variables loan approval and property location are independent.
H_A: The variables loan approval and property location are not independent.

The chi-square statistics is 10.47 and the p-value is 0.0053. The p-value (0.0053) is below 0.05 there is enough evidence to reject H_O for H_A.



There’s a greater probability of receiving a loan if the property is in the semi-urban area.

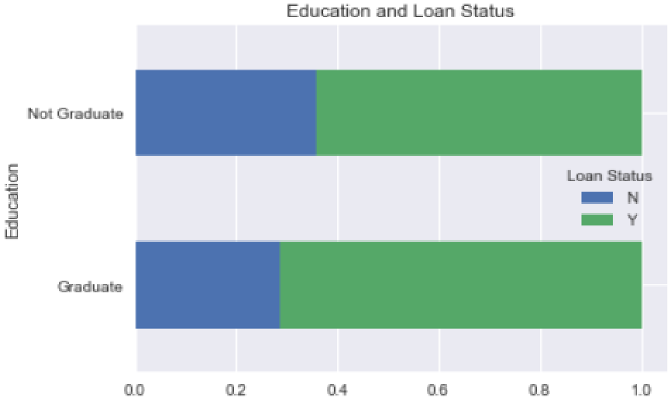
Is the education of the applicant taken into consideration when approving a loan?



A greater portion of more educated persons loans were approved (0.804) compared to the proportion that applied (0.787). While a great proportion of less educated individuals were denied (0.253) than the proportion that applied (0.213). There's a higher likelihood that those who are less educated earn less and therefore this can affect their chances of receiving a loan.

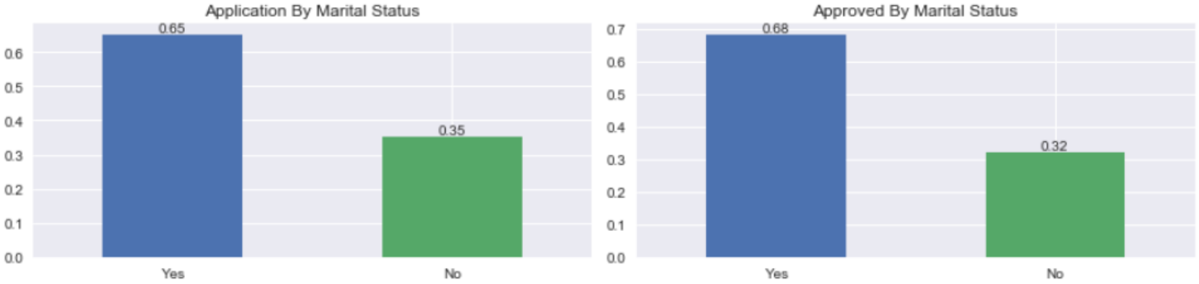
H_o: There is no statistically significant relationship between loan approval and educational status.
H_A: There is a statistically significant relationship between loan approval and educational status.

The chi-square statistics is 2.01 and the p-value is 0.156. The z-test statistics is -0.65 and the p-value is 0.514. Both a p-values are above 0.05, therefore there isn't enough evidence to reject H_O for H_A.



There’s a greater probability of getting a loan approved with a higher education.

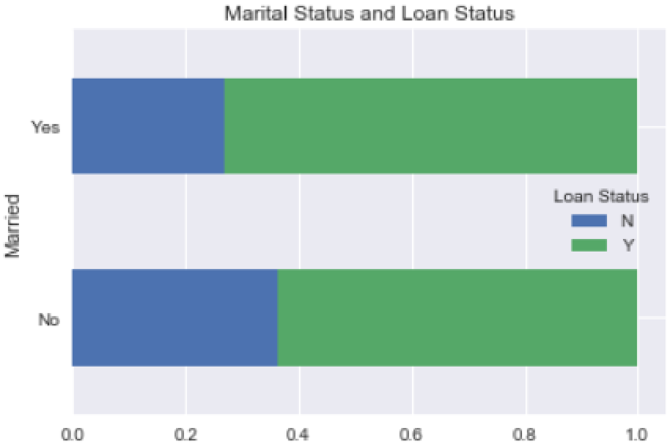
How does marital status affect loan approval?



A larger portion of applicants are married and the proportion of approved (0.680) is greater than that which applied (0.651). Therefore, it is more likely that married applicants may have their loans approved. It is more likely that they are willing to apply jointly improving their chances due to a greater combined income.

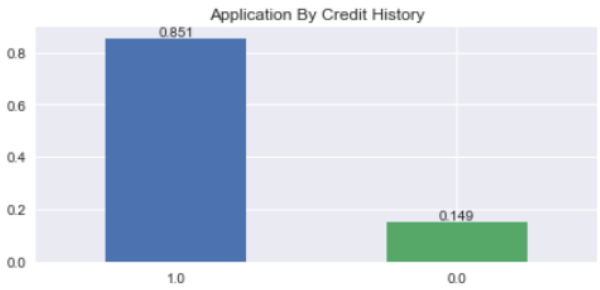
H_o: There is no relationship between loan approval and marital status.
H_A: There is a relationship between loan approval and marital status.

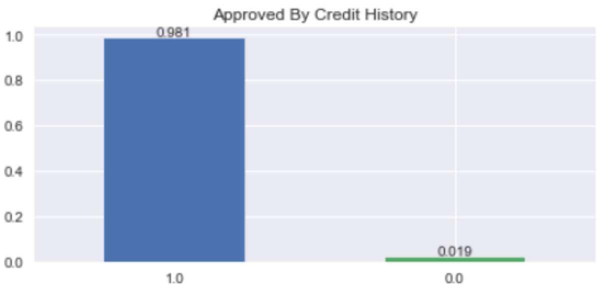
The chi-square statistics is 4.85 and the p-value is 0.028. The p-value (0.028) is below 0.05, there is enough evidence to reject H_O for H_A.



There’s a greater probability of getting a loan approved with a higher education.

Does having a credit history increase the chances of securing a loan?



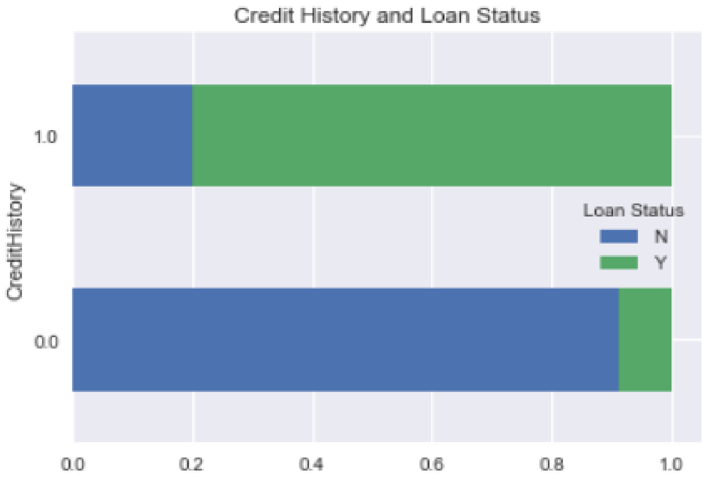


Having a previous credit history (1.0) does increase the likelihood of receiving an approval. Creditors are more comfortable with someone who has already proven their credit worthiness, of the approved loans only 0.019 had no credit history.

H_o: There isn’t a statistical difference between loan approval and credit history.
H_A: There is a statistical difference between loan approval and credit history.

The chi-square statistics is 156.75 and the p-value is 5.81e-36. With a p-value that is infinitesimally below 0.05 there is very strong evidence to reject H_O for H_A.

Based on the statistical evaluation of the data set it can be concluded that the factors with greatest consideration when determining the approval of a loan are Credit History and Property location.



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