Part 1: Overview of the Data

The Dream House finance company wants to automate the process of deciding whether a customer is eligible for a home loan, based on data from their previous loan applicants.

We will be looking at data corresponding to 480 loan applicants and relevant information about the applicants. This includes their gender (male or female), marital status (yes or no), number of dependents, education (graduate or not graduate), self-employment status (yes or no), applicant income, co-applicant income, loan amount, loan amount term, whether the credit history matches the company's guidelines, type of property area (urban, semi-urban, rural), and whether or not they were approved for the loan. This data was found in Kaggle.

We will start by looking at an overview of the data, and the relationships between variables, especially looking at differences in the data between those who were determined to be eligible for a loan and those who weren't. All results were found using SAS procedures. A significance level of 0.05 is used for all tests.

Numerical Overview: Applicant Income, Co-applicant Income, Loan Amount

We will start by looking at the numerical variables in the data. We can use the corr procedure in SAS to see that none of these variables are highly correlated with each other. Table 1 shows us that loan amount and applicant income have the highest magnitude correlation of 0.49531. Multicollinearity will not be an issue.

Pearson Correlation Coefficients, N = 480 Prob > |r| under H0: Rho=0			
	ApplicantIncome	**CoapplicantIncome**	**LoanAmount**
ApplicantIncome	1.00000	-0.11259 0.0136	0.49531 <.0001
CoapplicantIncome	-0.11259 0.0136	1.00000	0.19074 <.0001
LoanAmount	0.49531 <.0001	0.19074 <.0001	1.00000

Table 1

Categorical Overview: Gender, Married, Dependents, Education, Self-Employed, Loan Amount Term, Credit History, Property Area, Loan Status

For each categorical variable, we are going to look at the counts at each level. If there is a level with a very small number of observations, we will consider combining levels. Loan amount term is being treated as a categorical despite appearing numerical, because it consists of a small number of unique numbers with several observations.

Loan_Amount_Term	Frequency	Percent
36	2	0.42
60	2	0.42
84	3	0.63
120	3	0.63
180	36	7.50
240	2	0.42
300	9	1.88
360	411	85.63
480	12	2.50

Table 2

Loan_Term_Group	Frequency	Percent
Long	423	88.13
Medium	47	9.79
Short	10	2.08

Table 3

Looking at the frequency tables in appendix tables A1-A8, we can see that there are an adequate number of observations for all levels of all categorical variables except for Loan Amount Term in Table 2. We can see some levels with 2 or 3 observations. To solve this, we will group Loan Amount Term into 3 levels: Long, Medium, and Short. If the length is less than 180, it will be considered short, if it is 180 and above and less than 360, it will be medium, if it is 360 or above it will be long. We can see in table 3 that there are now an adequate number of observations for each level.

Distribution of Numerical Variables

To determine if each continuous variable has a normal distribution or not, we can look at histograms, box plots, and the various tests for normality that SAS gives us.

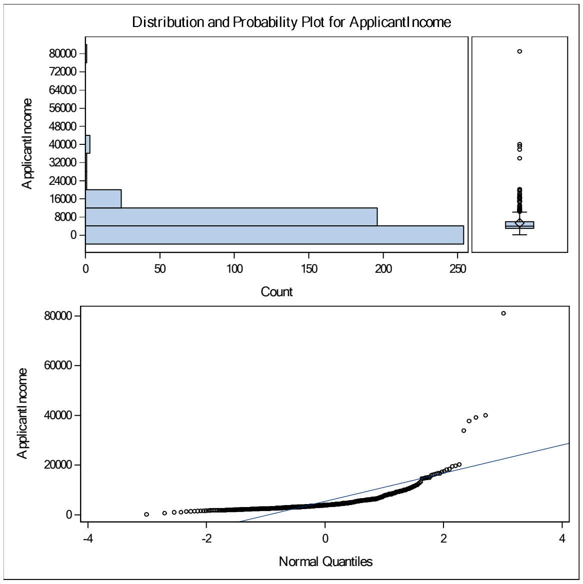


Figure 1

Table 3: Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.493112	**Pr < W**	<0.0001
Kolmogorov-Smirnov	**D**	0.237375	**Pr > D**	<0.0100
Cramer-von Mises	**W-Sq**	10.72196	**Pr > W-Sq**	<0.0050
Anderson-Darling	**A-Sq**	56.39212	**Pr > A-Sq**	<0.0050

For Applicant Income, we can look at the histogram in figure 1 to see that the distribution appears heavily right-skewed, which indicates that Applicant Income does not follow a normal distribution. To confirm this conclusion, we can see that every test for normality in Table 3 results in a p-value smaller than 0.05, so we reject the null hypothesis that the distribution is normal. The Q-Q plot in Figure 1 shows us that the points do not follow the straight line, so once again we are led to believe that the distribution is not normal. The visualizations also show us some potential outliers, and table A9 in the appendix tells us what those are. We will keep those in mind before we model.

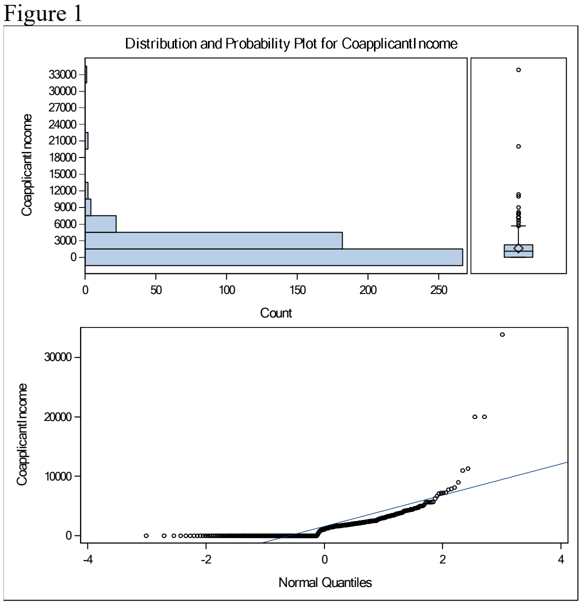


Figure 2

Table 4: Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.555886	**Pr < W**	<0.0001
Kolmogorov-Smirnov	**D**	0.272921	**Pr > D**	<0.0100
Cramer-von Mises	**W-Sq**	7.224572	**Pr > W-Sq**	<0.0050
Anderson-Darling	**A-Sq**	41.8139	**Pr > A-Sq**	<0.0050

For Coapplicant Income, we see similar trends. The distribution is skewed right, as seen in Figure 2. Each of the tests for normality in Table 4 is rejected, so it is safe to conclude that the distribution of Coapplicant Income is not normal.

Figure 3

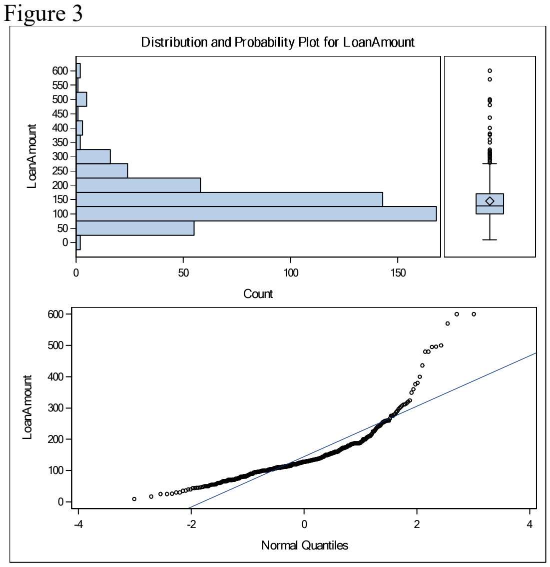


Table 5: Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.807411	**Pr < W**	<0.0001
Kolmogorov-Smirnov	**D**	0.147727	**Pr > D**	<0.0100
Cramer-von Mises	**W-Sq**	3.653832	**Pr > W-Sq**	<0.0050
Anderson-Darling	**A-Sq**	20.70463	**Pr > A-Sq**	<0.0050

The histogram for Loan Amount appears in Figure 3 slightly less skewed than the previous variables, but the Q-Q plot and tests for normality in Table 5 still tell us that the distribution is not normal.

Part 2: Association Between Predictors and Target

In this part we will spend some time assessing the relationship between each of the predictor variables and the target variable, loan status. We will be looking for differences in the variables when they are determined to be eligible for a loan versus when they aren't.

Categorical Variables Grouped by Loan Status

We can start by looking at hypothesis tests for group independence. This way we can see if there are differences in the groups, and we can determine if there is a relationship. In other words, do different levels of a given predictor change the frequency of being approved for a loan? We will also consider counts of cross-tabulation of each variable with loan status. These can be found in the appendix, Tables A12-A19

Table 6	DF	Value	Prob
Chi-Square	1	1.9972	0.1576
Likelihood Ratio Chi-Square	1	1.9456	0.1631
Continuity Adj. Chi-Square	1	1.6496	0.1990
Mantel-Haenszel Chi-Square	1	1.9930	0.1580
Phi Coefficient		0.0645	
Contingency Coefficient		0.0644	
Cramer's V		0.0645	

Gender:
First, we will look at gender, and we find that the observed counts do not stray too far from what we would expect if gender and loan status were independent. We are led to believe that there is no relationship, and our intuition is confirmed when we look at the chi-square p-value in Table 6, which is at 0.1576, which is above 0.05. We do not need to look at Fisher's exact test because our counts for each cell are sufficiently large.

Table 7	DF	Value	Prob
Chi-Square	1	6.0557	0.0139
Likelihood Ratio Chi-Square	1	5.9665	0.0146
Continuity Adj. Chi-Square	1	5.5571	0.0184
Mantel-Haenszel Chi-Square	1	6.0431	0.0140
Phi Coefficient		0.1123	
Contingency Coefficient		0.1116	
Cramer's V		0.1123	

Married:
Next, we can look at marriage status. Here we find that there is more deviance from the expected counts in our observed counts. The chi-square test p-value in Table 7 is 0.0139, which is below 0.05, so we can conclude that marriage status does have a relationship with loan status. Which is intuitive, if you are married you are likely to have two sources of household income, and more financial stability in general, which means you are more likely to be approved for a home loan.

Education:
Next up is education. Like gender, our observed counts aren't too far from what we would expect, and our p-value in Table 8 is 0.1338, which is above our significance level of 0.05. We fail to conclude that there is a relationship between education level and loan status.

Table 8	DF	Value	Prob
Chi-Square	1	2.2482	0.1338
Likelihood Ratio Chi-Square	1	2.1936	0.1386
Continuity Adj. Chi-Square	1	1.8942	0.1687
Mantel-Haenszel Chi-Square	1	2.2435	0.1342
Phi Coefficient		-0.0684	
Contingency Coefficient		0.0683	
Cramer's V		-0.0684	

Self Employed:
Observed counts are even closer to our expected counts for self-employment status. We have a very high p-value in Table 9 of 0.4469, so we cannot say there is a relationship between self-employment and loan status.

Table 9	DF	Value	Prob
Chi-Square	1	0.5785	0.4469
Likelihood Ratio Chi-Square	1	0.5678	0.4512
Continuity Adj. Chi-Square	1	0.3808	0.5372
Mantel-Haenszel Chi-Square	1	0.5773	0.4474
Phi Coefficient		-0.0347	
Contingency Coefficient		0.0347	
Cramer's V		-0.0347	

Credit History:
The results for credit history are unsurprising. We can see there is a clear relationship between credit history and loan status. It should be expected that people who have a credit history that meets the company's guidelines are approved for home loans more often than those with a worse credit history. The p-values in Table 10 are below 0.0001, which are the lowest we've seen yet.

Table 10	DF	Value	Prob
Chi-Square	1	134.5217	<.0001
Likelihood Ratio Chi-Square	1	129.0268	<.0001
Continuity Adj. Chi-Square	1	131.2933	<.0001
Mantel-Haenszel Chi-Square	1	134.2414	<.0001
Phi Coefficient		0.5294	
Contingency Coefficient		0.4679	
Cramer's V		0.5294	

Property Area:
Property area also has a relationship with loan status. Looking at the expectations, it appears that trying to get a home loan for properties in purely rural or urban areas makes you less likely to get approved for the loan than we would expect. In contrast, people trying to get loans for semiurban properties appear to be more likely to get approved for the loans than we would expect assuming independence. The chi-square p-value in Table 11 is well below 0.05 at 0.0022. The Mantel-Haenszel chi-square p-value is high, this is because it is testing for a linear relationship specifically. We can see that this is not linear if we know that you are less likely to get a loan in rural, then more likely in semiurban, then less likely again in urban. This does not follow a linear trend.

Table 11	DF	Value	Prob
Chi-Square	2	12.2259	0.0022
Likelihood Ratio Chi-Square	2	12.4916	0.0019
Mantel-Haenszel Chi-Square	1	0.4711	0.4925
Phi Coefficient		0.1596	
Contingency Coefficient		0.1576	
Cramer's V		0.1596	

Table 12	DF	Value	Prob
Chi-Square	3	2.9006	0.4072
Likelihood Ratio Chi-Square	3	2.9953	0.3923
Mantel-Haenszel Chi-Square	1	0.6012	0.4381
Phi Coefficient		0.0777	
Contingency Coefficient		0.0775	
Cramer's V		0.0777	

Dependents:
Our observed counts based on the number of dependents the loan applicant has are not far from what we would expect under independence. We also have a large p-value in Table 12 of 0.4072, so we cannot conclude that there is a relationship between number of dependents and loan status.

Table 13	DF	Value	Prob
Chi-Square	2	1.3613	0.5063
Likelihood Ratio Chi-Square	2	1.3150	0.5181
Mantel-Haenszel Chi-Square	1	0.6729	0.4120
Phi Coefficient		0.0533	
Contingency Coefficient		0.0532	
Cramer's V		0.0533	

Loan Term Group:
Our final categorical variable is the one we made from Loan_Term. We find that our observed values aren't far from our expectations, and we have a p-value in Table 13 of 0.5063. We cannot conclude there is a relationship between the length of the loan and loan status.

Numerical Variables Grouped by Loan Status

We have previously determined that the normality assumption is not met with our numerical variables. So, we will test for differences in distributions when grouped by loan status by using the results from the npar1way procedure in SAS. This procedure performs nonparametric tests that will allow us to interpret the results.

We are looking at Wilcoxon scores for each of our numerical variables. That is, applicant income, co-applicant income, and loan amount. Their test statistics and box plots for Wilcoxon scores are displayed Figures 4-6. We can see that none of the numerical variables have a significant result, so none of these numerical variables appear to be significant predictors of loan status.

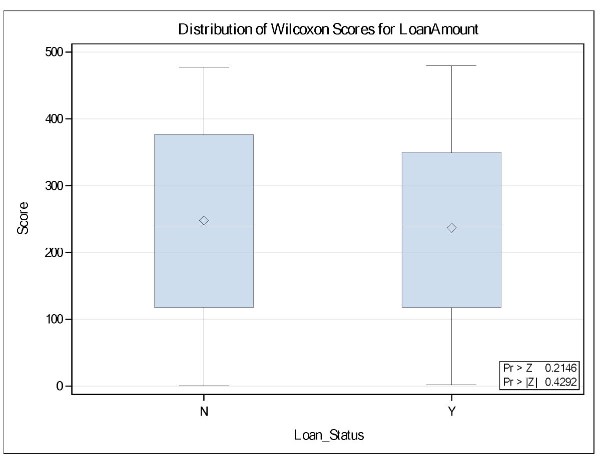


Figure 4

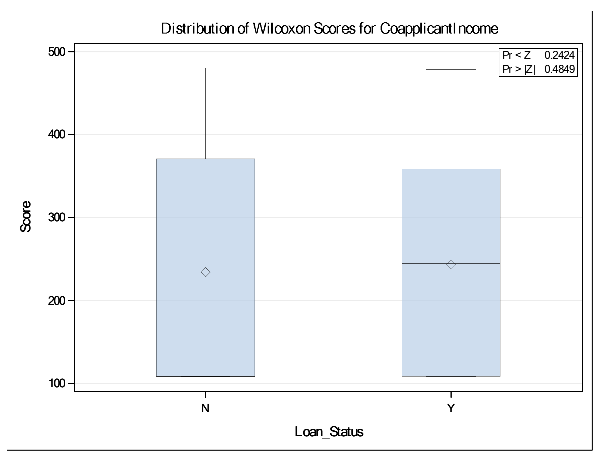


Figure 5

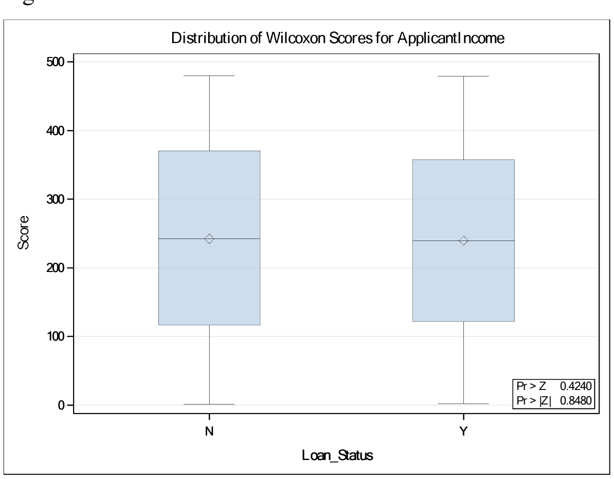


Figure 6

Part 3: Modeling

The goal of this analysis is to be able to determine whether new loan applicants should be approved for a loan or not based on the historical data. To do this, we will need to create a model that will allow us to predict with some degree of accuracy. When we are trying to classify observations into groups of a binary variable, the best way to do this is to use a logistic regression model. Logistic regression will force predictions into a realistic range of 0 to 1. We are predicting the probability of an applicant being approved, so we are in fact looking for values between 0 and 1.

Feature Selection

The data set given has several different variables with differing levels of interaction with loan status, as we found in our preliminary analysis. This means that the full model that includes all the possible variables as predictors is unlikely to be the best performing model. We will have to perform a feature selecting algorithm to determine which variables to include and which to leave out. We will specifically be using stepwise selection. This algorithm starts without any of the variables, and includes the most significant one first, given that it is significant at all to begin with. It will then repeat the act of checking to see if any of the included variables are insignificant and adding significant variables. The cutoff to include or remove a variable is a p-value of 0.05 for the chi-square test. This is essentially a combination of forward selection and backward elimination procedures.

Table 14: Summary of Stepwise Selection

Step	Effect		DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
	Entered	Removed					
1	Credit_History		1	1	134.5217		<.0001
2	Property_Area		2	2	12.3102		0.0021
3	Married		1	3	5.9189		0.0150

As we see in Table 14, the only variables that remain after performing stepwise selection are credit history, property area, and married. All the other variables were determined to not have enough predictive power to justify including in the model. This aligns with our intuition, and our findings in Part 2. We found that the only categorical variables that rejected the null hypothesis of independence with loan status were these variables. We also found that none of the continuous numerical variables had significant differences in distributions between loan statuses. It is unsurprising that these variables are left out of the model.

Model Evaluation

Now that we have our final model, we need to determine how effective the model is.

Table 15: Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	147.4722	4	<.0001
Score	147.6848	4	<.0001
Wald	81.3792	4	<.0001

Table 15 compares our model to the model without any of our predictors, just the intercept, to see if there is a significant improvement. The null hypothesis is that none of our predictors are significant in predicting loan status, which means that all the coefficients, or betas, would be zero. This null hypothesis is rejected by all 3 tests, so we can conclude with confidence that there is sufficient evidence to suggest that at least one of the predictors has a non-zero coefficient. So the model with these variables is significantly better than the one without.

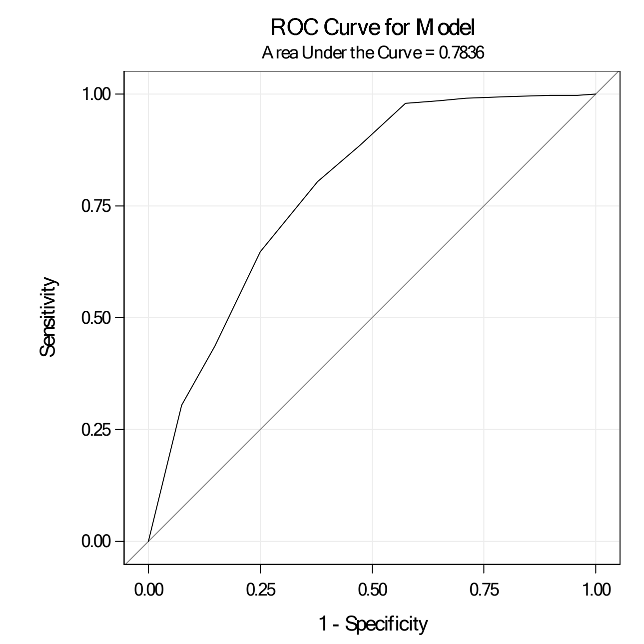


Figure 7

Figure 7 graphs the ROC curve for our model. This is a visual representation of the tradeoff between sensitivity and specificity. Sensitivity is proportion of correctly classified positives, which in our case is loan status being yes. Specificity is the proportion of correctly classified negatives, where loan status is no. On the x-axis is 1 – specificity, which is the false positive rate. When we create a logistic regression model, we choose a cutoff, usually 0.5 to determine whether to classify as a 1 or 0. When we change this cutoff, sensitivity and specificity will change. If you raise the cutoff, sensitivity goes down and specificity goes up. This means you miss more positives, but negatives are more accurately classified. Lowering the cutoff will have the opposite effect. Each point on this curve is a classification performance at different cutoffs.

The model has an area under the ROC curve of 0.7836, indicating it has good ability to differentiate between approved and denied loans. In 78.36% of randomly selected pairs of approved and denied applicants, the model assigns a higher probability to the approved applicant.

Table 17 below shows an example of sensitivity and specificity if we set our cutoff as 0.5. We can see that the sensitivity is 97.9%, 325 loan approvals were correctly classified as loan approvals, while 7 loan approvals were incorrectly classified as denials. Specificity is 42.6%, where 63 loan denials were correctly classified as denials, while 85 loan denials were incorrectly classified as approvals. This means that at a cutoff of 0.5, the model is classifying many people as loan approvals that shouldn't be approved. A different cutoff might be necessary depending on the company's goals.

Table 17: Classification Table

	Correct		Incorrect		Percentages				
Prob Level	**Event**	**Non-Event**	**Event**	**Non-Event**	**Correct**	**Sensi-tivity**	**Speci-ficity**	**Pos Pred**	**Neg Pred**
0.500	325	63	85	7	80.8	97.9	42.6	79.3	90.0

Table 17: Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	**DF**	**Pr > ChiSq**
1.4027	5	0.9240

We can test to see if the model fits the data well using a goodness-of-fit test. The null hypothesis is that the model fits the data well, which is to say that the predicted probabilities match the observed outcomes. We can look at Table 17 to see that our p-value is very high, which means that we fail to reject the null hypothesis and there is insufficient evidence to suggest that the fit is not good.

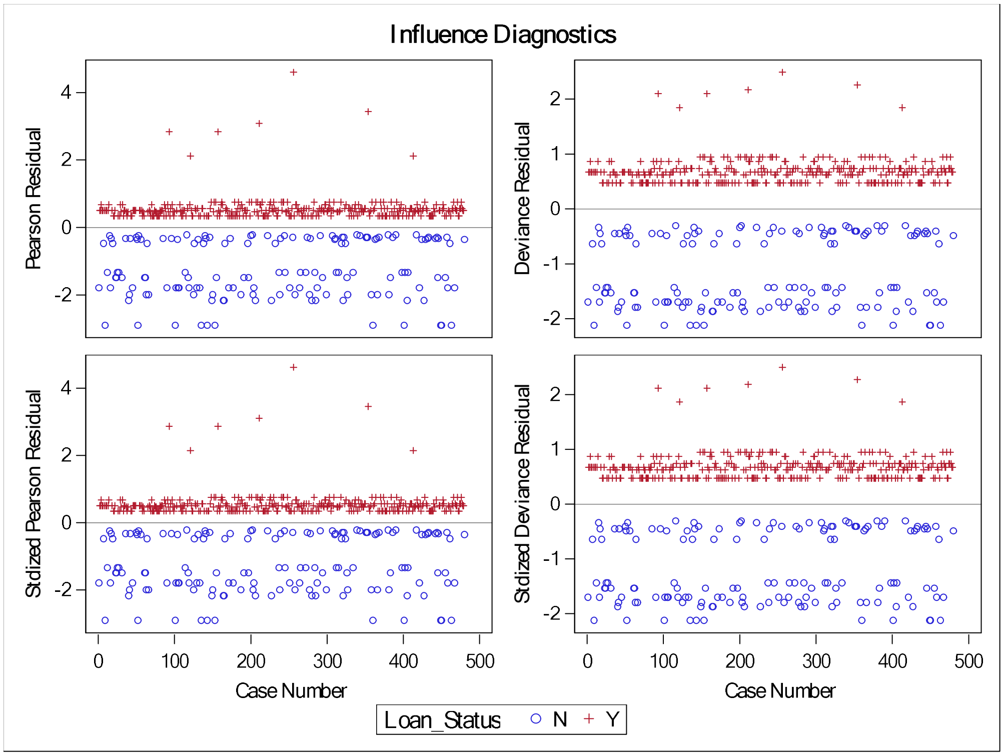


Figure 8

Sometimes models have certain observations that are highly influential and may be damaging the effectiveness of the model. We can look at several metrics above in Figure 8 to identify any highly influential observations, but it appears that nothing is influential enough to warrant further investigation.

Model Interpretation

We have determined that our model is moderately effective at predicting loan status. Now we will explore what each coefficient in the model means and what they tell us about each variable's ability to help predict loan status.

Figure 9

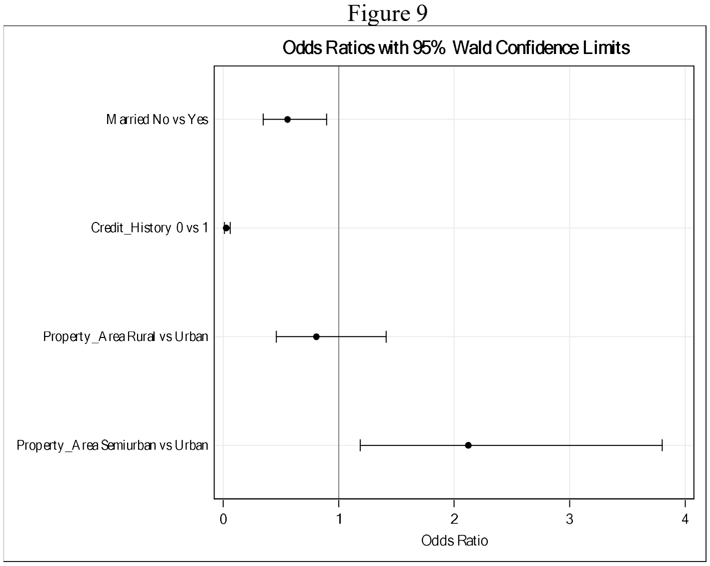


Figure 9 above visualizes the 95% confidence limits of each predictor's odds ratio. The odds ratio is found by exponentiating the coefficient, or beta. We do this because the coefficients are modeling the log odds of loan status being yes, since we are using logistic regression. We can compare each odds ratio confidence interval to 1. This is because an odds ratio of 1 means that this predictor multiplies the odds of loan status being yes by 1, which means it is not significant. If beta is negative, exp(beta) will be below 1, if beta is zero, exp(beta) will be 1, and if beta is positive, then exp(beta) will be above 1. We see that the entire confidence intervals for married and credit history are negative. This means that an unmarried applicant has lower odds of approval relative to married applicants, and an applicant with an undesired credit score has lower odds of approval relative to an applicant with a desired credit score. The entire confidence interval for semiurban vs urban property area is positive. This means that an application with a semiurban property area has higher odds of approval relative to an application with an urban property area. Rural vs urban has a point estimate below 1, but 1 is within the confidence interval, so the effect is not significant. Note that property area has two dummy variables because there are three groups, so urban is the reference.

Table 18: Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-0.5517	0.2139	6.6493	0.0099
Married	**No**	1	-0.2928	0.1212	5.8379	0.0157
Credit_History	**0**	1	-1.8145	0.2130	72.5564	<.0001
Property_Area	**Rural**	1	-0.3944	0.1668	5.5918	0.0180
Property_Area	**Semiurban**	1	0.5735	0.1734	10.9428	0.0009

Table 18 displays the exact estimates for the coefficients. These tell us how each dummy variable relates to the expected log odds of approving a loan. These interpretations are not very meaningful, so we can exponentiate to find the odds ratios that we analyzed before. Interpretations for odds ratios are as follows:

- Intercept: The odds of a married loan applicant with a desired credit history and a property in an urban area being approved for a loan are exp(-0.5517) = 0.576 to 1, or about 36.5%
- Unmarried: The odds of loan approval for an unmarried applicant are exp(-0.2928) = 0.557 times the odds of loan approval for a married applicant, all else held constant.

- Poor Credit History: The odds of loan approval for an applicant with poor credit history are exp(-1.8145) = 0.027 times the odds of loan approval for an applicant with good credit history, all else held constant.
- Rural vs Urban Area: The odds of loan approval for an application in a rural area are exp(-0.3944) = 0.806 times the odds of loan approval for an application in an urban area, all else held constant.
- Semiurban vs Urban Area: The odds of loan approval for an application in a semiurban area are exp(0.5735) = 0.1734 times the odds of loan approval for an application in an urban area, all else held constant.

Conclusions

From our modeling, we have an in-depth understanding of how effective each socio-economic feature is at predicting whether a loan should be approved or denied.

- We know specifically which features are insignificant in predicting loan status
- Applicants with the company's desired credit history are far more likely to be approved. This is the most significant predictor of loan status.
- Married applicants are more likely to be approved for loans.
- Loans for properties in urban areas are more likely to be approved than in rural areas, but less likely to be approved than in semiurban areas.
- With the given features, loan status can be predicted with moderate levels of accuracy. The model classifies somewhat well but likely needs some improvement before being implemented in the Dream House finance company for loan approval automation.

There are several ways to take this analysis to the next level to improve model performance. We can search for more relevant information about our applicants. The given variables will not tell the whole story, so more information could prove useful. We could find alternative ways to select the best model. Stepwise selection is not always the most robust choice, sometimes the best combination of predictors is missed by the algorithm. One way is LASSO regularization. It would also be valuable to further evaluate the categorical variables and how they interact with each other.

This model is a solid starting point but there is certainly some work to be done before the Dream House finance company is ready to automate loan approval.

Appendix

Categorical Counts

Education	Frequency	Percent
Graduate	383	79.79
Not Graduate	97	20.21

Table A1

Dependents	Frequency	Percent
0	274	57.08
1	80	16.67
2	85	17.71
3+	41	8.54

Table A2

Self_Employed	Frequency	Percent
No	414	86.25
Yes	66	13.75

Table A3

Married	Frequency	Percent
No	169	35.21
Yes	311	64.79

Table A4

Credit_History	Frequency	Percent
0	70	14.58
1	410	85.42

Table A5

Loan_Status	Frequency	Percent
N	148	30.83
Y	332	69.17

Table A6

Gender	Frequency	Percent
Female	86	17.92
Male	394	82.08

Table A7

Property_Area	Frequency	Percent
Rural	139	28.96
Semiurban	191	39.79
Urban	150	31.25

Table A8

Extreme Observations

Extreme Observations ApplicantIncome			
Lowest		**Highest**	
Value	**Obs**	**Value**	**Obs**
150	171	33846	143
645	391	37719	348
1000	63	39147	145
1025	328	39999	121
1299	13	81000	320

Table A9

Extreme Observations CoapplicantIncome			
Lowest		**Highest**	
Value	**Obs**	**Value**	**Obs**
0	480	10968	9
0	479	11300	138
0	477	20000	313
0	476	20000	327
0	472	33837	455

Table A10

Extreme Observations LoanAmount			
Lowest		**Highest**	
Value	**Obs**	**Value**	**Obs**
9	445	496	472
17	13	500	381
25	99	570	287
25	76	600	121
26	434	600	439

Table A11

Counts and Expected Counts

Table of Education by Loan_Status			
Education	**Loan_Status**		
Frequency Expected Percent	**N**	**Y**	**Total**
Graduate	112 118.09 23.33	271 264.91 56.46	383 79.79
Not Graduate	36 29.908 7.50	61 67.092 12.71	97 20.21
Total	148 30.83	332 69.17	480 100.00

Table A12

Table of Married by Loan_Status			
Married	**Loan_Status**		
Frequency Expected Percent	**N**	**Y**	**Total**
No	64 52.108 13.33	105 116.89 21.88	169 35.21
Yes	84 95.892 17.50	227 215.11 47.29	311 64.79
Total	148 30.83	332 69.17	480 100.00

Table A13

Table of Gender by Loan_Status			
Gender	**Loan_Status**		
Frequency Expected Percent	**N**	**Y**	**Total**
Female	32 26.517 6.67	54 59.483 11.25	86 17.92
Male	116 121.48 24.17	278 272.52 57.92	394 82.08
Total	148 30.83	332 69.17	480 100.00

Table A14

Table of Self_Employed by Loan_Status			
Self_Employed	**Loan_Status**		
Frequency Expected Percent	**N**	**Y**	**Total**
No	125 127.65 26.04	289 286.35 60.21	414 86.25
Yes	23 20.35 4.79	43 45.65 8.96	66 13.75
Total	148 30.83	332 69.17	480 100.00

Table A15

Table of Credit_History by Loan_Status			
Credit_History	**Loan_Status**		
Frequency Expected Percent	**N**	**Y**	**Total**
0	63 21.583 13.13	7 48.417 1.46	70 14.58
1	85 126.42 17.71	325 283.58 67.71	410 85.42
Total	148 30.83	332 69.17	480 100.00

Table A16

Table of Property_Area by Loan_Status			
Property_Area	**Loan_Status**		
Frequency Expected Percent	**N**	**Y**	**Total**
Rural	54 42.858 11.25	85 96.142 17.71	139 28.96
Semiurban	42 58.892 8.75	149 132.11 31.04	191 39.79
Urban	52 46.25 10.83	98 103.75 20.42	150 31.25
Total	148 30.83	332 69.17	480 100.00

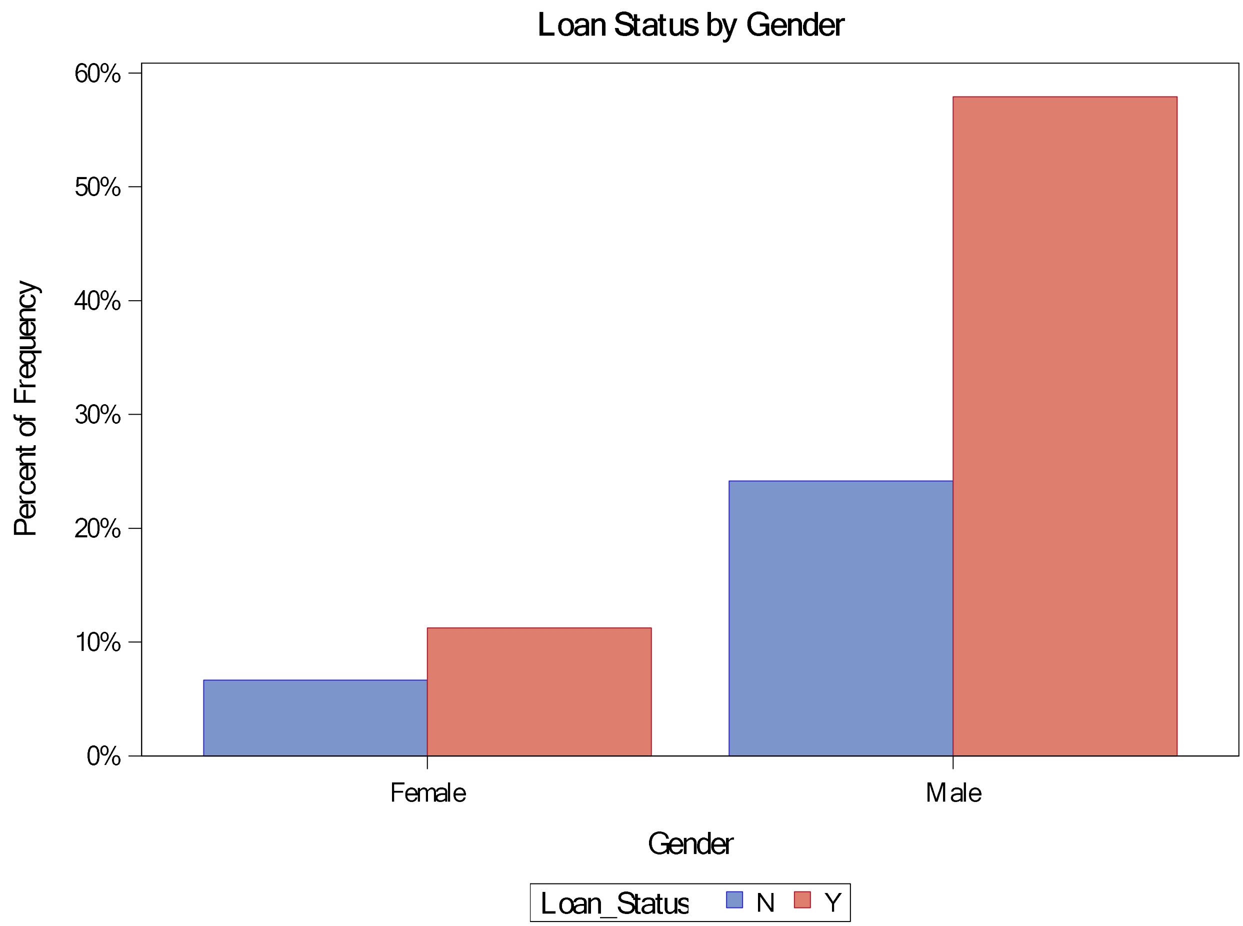
Table A17

Table of Dependents by Loan_Status			
Dependents	**Loan_Status**		
Frequency Expected Percent	**N**	**Y**	**Total**
0	87 84.483 18.13	187 189.52 38.96	274 57.08
1	28 24.667 5.83	52 55.333 10.83	80 16.67
2	20 26.208 4.17	65 58.792 13.54	85 17.71
3+	13 12.642 2.71	28 28.358 5.83	41 8.54
Total	148 30.83	332 69.17	480 100.00

Table A18

Table of Loan_Term_Group by Loan_Status			
Loan_Term_Group	**Loan_Status**		
Frequency Expected Percent	**N**	**Y**	**Total**
Long	127 130.43 26.46	296 292.58 61.67	423 88.13
Medium	18 14.492 3.75	29 32.508 6.04	47 9.79
Short	3 3.0833 0.63	7 6.9167 1.46	10 2.08
Total	148 30.83	332 69.17	480 100.00

Table A19


Loan Status by Gender
60%
50%
40%
30%
20%
10%
0%
Percent of Frequency
Female
Male
Gender
Loan_Status
N
Y