Loan Prediction

Joe Pollastrini

June 18, 2020

Overview

This project's main goal is to see if a system can be built to help automate the loan approval process. An applicant will fill out an online form for variables such as income, number of dependents, loan amount, and more. The system will take that data, run it through a model, and predict whether the applicant's loan will be approved. Any applicant predicted to be approved can be targeted by representatives to expedite the loan process. This project focuses on building the predictive model.

Data

Exploration

The dataset was provided by a hackathon from Analytics Vidhaya (at the bottom of the page). Variables for the train and test data sets and their information can be found in Figure 1 and Figure 2, respectively. Note that the test set was missing the result variable (Loan_Status), and model accuracy was determined by submitting here.

The dependents variable was provided as a string, however, it is better to use it as a numerical variable. This change will help ensure the model does not "cherry pick" certain Dependents groups as having a higher probability of approval, and will instead treat them as an ordered value. Figure 3 illustrates the loan approval probability by Dependent group.

Education was one variable hypothesized to be an important indicator, however, while being a graduate was more favorable for loan approval, it was not as large of a difference as expected. Figure 4 illustrates the breakdown by Education.

360.0 terms was the most common value for Loan_Amount_Term, and most of the other value counts were too small to keep on their own, so grouping based on standard loan terms was implemented. The following are the groups, 15 and 30 year loan terms being the standards:

- Less than 15 year
- 15 year (180.0)
- 15 year to 30 year (non-inclusive)
- 30 year (360.0)
- Greater than 30 year

Figure 5 shows the value counts before and after grouping.

Credit History was another variable hypothesized to be an important indicator, however, the actual importance was not expected. After seeing the breakdown based on category, this will be the most explanatory variable. It also has the most missing values, so imputation here will be vital. Figure 6 illustrates the breakdown for Credit_History.

A higher family income was hypothesized as an indicator for a better chance of loan approval, however, breaking down the income into quartiles, and deciles did not confirm that thought. It's possible that income isn't an end all signal of financial health, and some better variables would need

to be created. Figure 7 illustrates loan approval by quartile for total income (ApplicantIncome + CoapplicantIncome).

The following variables and their formulas were created to help paint a better picture of the applicants financial standing:

- FamilyIncome - ApplicantIncome + CoapplicantIncome
- DualIncome_IO - 1 if CoapplicantIncome is greater than 0, otherwise 0
  - Designed to indicate multiple income streams or not
- Debt_Equity - (LoanAmount * 1000) / FamilyIncome
  - Designed to express how much an applicant is extending themself
- Debt_Equity_Annual - [((LoanAmount * 1000) / Loan_Amount_Term) * 12] / FamilyIncome
  - Designed to be a ratio of total income going to loan per year (Note there are no interest calculations)

Imputation

Figure 1 and Figure 2 show which variables needed to be imputed and how many values were missing. There were only 19 applications with 2 or more missing values.

Initially, many missing values were imputed with the mode, however, K-Nearest Neighbor models were utilized to have a more realistic real world method of imputation. Loan_Amount_Term, LoanAmount, and Self_Employed variables were imputed using a KNN model. Cross validation was utilized to select the best k value.

In order to utilize the float value of Loan_Amount_Term for Debt_Equity_Annual, any missing value was imputed using the following table based on the imputed LoanTermGroup.

LoanTermGroups	Loan_Amount_Term
< 15	90, (7.5 * 12)
15	180
(15, 30)	270, (22.5 * 12)
30	360
> 30	480, (40 * 12)

Credit_History required the most effort to find the best imputation because most models just selected having a history. Based on groupings, if an applicant was not educated and lived in an urban property area, a KNN model was used to predict Credit_History, otherwise, it was assumed the applicant had a history. Figure 8 shows a breakdown of Credit_History probability based on education and property area.

Missing Gender values were replaced as female if the applicant was not married and had 1 (one) dependent. Otherwise the applicant was assumed to be male. Figure 9 shows a breakdown of an applicant's expected gender based on Married and Dependents.

Missing Married values were replaced as not married if the applicant was a female. Otherwise it was assumed the applicant was a male. Figure 10 shows a breakdown of probability of marriage based on Gender.

Missing Dependents values were replaced with 1 if the applicant was married and male. Otherwise it was assumed there was no dependent. Figure 11 shows a breakdown of the average dependent size based on marriage and gender.

LoanAmount was log transformed in order to make the distribution closer to normal. The mean logged amount was calculated for each LoanTermGroups category. Any missing value in the logged LoanAmount was replaced with the average for its corresponding LoanTermGroups category.

Cleaning

Any variable that was Yes or No, or anything of that sort, was converted to an indicator, 1 in place of Yes (Male, Graduate), and 0 in place of No. Property Area was converted to a dummy variable. Dependents were converted to a numeric variable (3+ converted to 3).

Two indicators were created to show if an applicant's LoanAmount or FamilyIncome was an outlier (based on the training set only).

Model Build

Many models were built using multiple methods, such as logistic regression and random forests. A logistic regression model performed well, and can be seen later, however, ultimately, a random forest model was chosen as the best performing model. It was built using hyperparameter optimization and cross validation, as well as feature selection. Number of trees, maximum features per tree, maximum depth per tree, minimum samples for a split, and minimum samples for a leaf were all optimized. The final model used Credit_History, Debt_Equity, Debt_Equity_Annual, IncomePerMember, FamilyIncome, and LoanAmountLog in the model. Variables were kept if their importance was higher than the average importance for all variables.

Results

The benchmark rate, based on the percentage of loans approved from the training dataset, is 68.7%. Any model that does not improve on this rate is not worth implementing.

Figure 12 shows the final ANOVA table for the logistic regression output by statsmodels.api.Logit. Below are various scoring metrics and a confusion matrix for the logistic regression. These can also be found in Figure 13.

Accuracy	81.1%
Precision	0.792
Recall	0.983
F-Score	0.877
AUC-ROC	0.706
FPR	57.1%

Predicted

	Y	N
Y	351	6
N	92	69

Figure 14 shows the variable importance for the random forest model. Below are various scoring metrics and a confusion matrix for the random forest model. These can also be found in Figure 15.

Random Forest Model Score Statistics and Confusion Matrix

Accuracy	83.6%
Precision	0.822
Recall	0.974
F-Score	0.892
AUC-ROC	0.750
FPR	47.3%

Predicted

	Y	N
Y	370	10
N	80	89

Both models have surpassed the benchmark rate, and can be used to help better predict an applicant's loan approval. Furthermore, both models have a high false positive rate, but the random forest model was slightly better in nearly every scoring metric. Therefore, random forest was the model used for submission.

Conclusion

The random forest model predictions were submitted to the hackathon site for scoring. A final score of 0.7847 was awarded, putting me at 1125 of 59718 for the competition. The model built can predict, with fairly high accuracy, whether a loan will be approved based on applicants filling out an online form.

Next Steps

- Look into different models for different loan types. It's likely the loans with term lengths of 3, 4, 5 years are auto loans, and can be treated differently.
- Build out an automated system that can grab applicant answers from an online application, run it through the model, and send the loan approval prediction to an agent.

Appendix

Figure 1

Train Set Information

Variable	Description	Type	Total	% Missing
Loan_ID	Unique Loan ID	str	614	
Gender	Male/Female	str	601	2.1%
Married	Applicant married (Y/N)	str	611	0.5%
Dependents	Number of dependents	str	599	2.4%
Education	Applicant Education (Graduate/Not Graduate)	str	614	
Self_Employed	Self employed (Y/N)	str	582	5.2%
ApplicantIncome	Applicant income	int	614	
CoapplicantIncome	Coapplicant income	float	614	
LoanAmount	Loan amount **in thousands**	float	592	3.6%
Loan_Amount_Term	Term of loan **in months**	float	600	2.3%
Credit_History	Credit history meets guidelines	float	564	8.1%
Property_Area	Urban / Semiurban / Rural	str	614	
Loan_Status	(Target) Loan approved (Y/N)	str	614	

Figure 2

Test Set Information

Variable	Description	Type	Total	% Missing
Loan_ID	Unique Loan ID	str	367	
Gender	Male/Female	str	356	3.0%
Married	Applicant married (Y/N)	str	367	
Dependents	Number of dependents	str	357	2.7%
Education	Applicant Education (Graduate/Not Graduate)	str	367	
Self_Employed	Self employed (Y/N)	str	344	6.3%
ApplicantIncome	Applicant income	int	367	
CoapplicantIncome	Coapplicant income	float	367	
LoanAmount	Loan amount **in thousands**	float	362	1.4%
Loan_Amount_Term	Term of loan **in months**	float	361	1.6%
Credit_History	Credit history meets guidelines	float	338	7.9%
Property_Area	Urban / Semiurban / Rural	str	367	

Figure 3

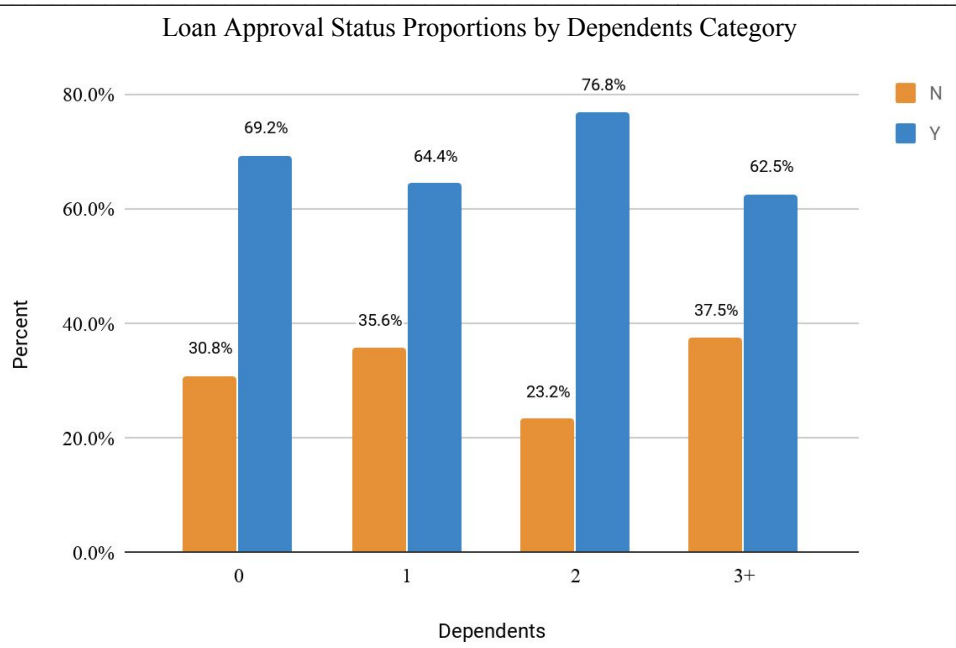

Figure 4

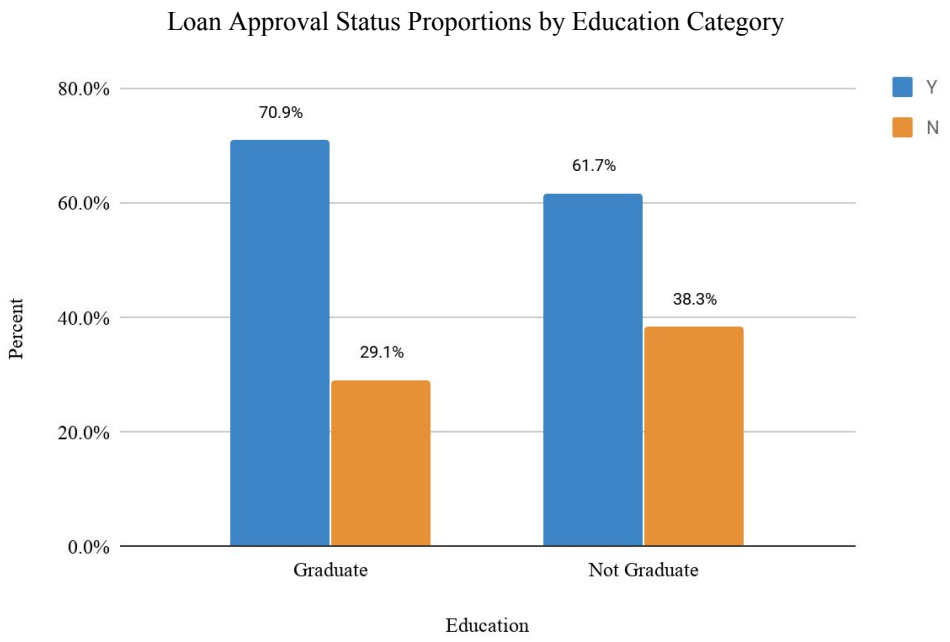

Figure 5

Loan_Term_Amount Counts & LoanTermGroup Counts

Before Grouping

Value	Count
360.0	512
180.0	44
480.0	15
300.0	13
84.0	4
240.0	4
120.0	3
36.0	2
60.0	2
12.0	1

After Grouping

Value	Count
30	512
15	44
(15, 30)	17
>30	15
<15	12

Figure 6

Loan Approval Status Proportions by Credit History Category

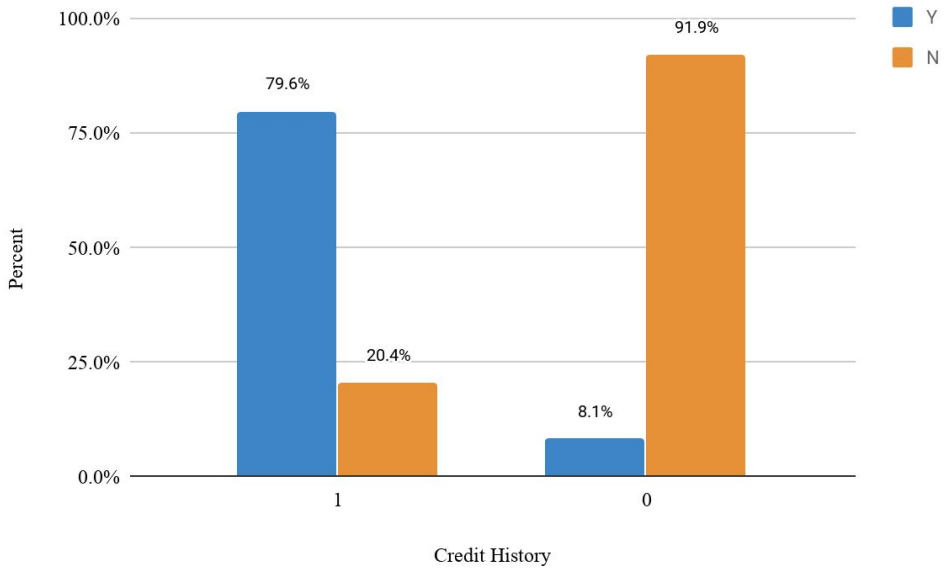

Figure 7

Loan Approval Status Proportions by Income (Quartiles)

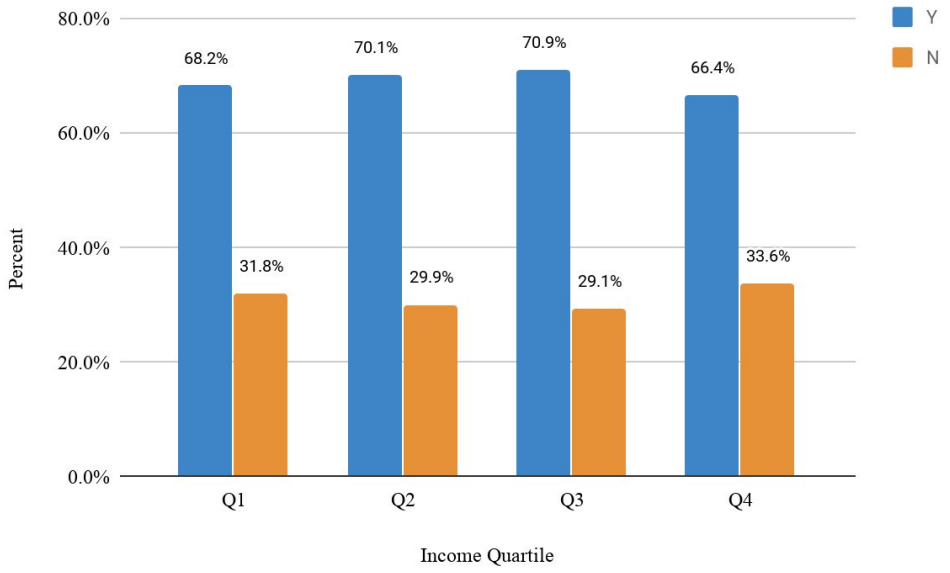

Figure 8

Proportion of Credit History by Education & Property Area Grouping

Education	Property Area	History	No History	P(No History)
No	**Rural**	36	8	18.2%
No	**Urban**	25	12	32.4%

No	**Semiurban**	34	6	15.0%
Yes	**Rural**	101	20	16.5%
Yes	**Urban**	126	19	13.1%
Yes	**Semiurban**	153	24	13.6%

Figure 9

Proportion of Male by Married & Dependents Grouping

Married	Dependents	Female	Male	P(Male)
No	**0**	60	109	64.5%
No	**1**	13	10	43.5%
No	**2**	2	6	75%
No	**3+**	3	3	50%
Yes	**0**	20	149	88.2%
Yes	**1**	6	72	92.3%
Yes	**2**	5	86	94.5%
Yes	**3+**	0	42	100%

Figure 10

Proportion of Married by Gender Grouping

Gender	No	Yes	P(Married)
Female	80	31	27.9%
Male	130	357	73.3%

Figure 11

Expected Dependents by Married & Gender Groupings

Married	Gender	Average Dependents
No	**Female**	0.33
No	**Male**	0.24
Yes	**Female**	0.52
Yes	**Male**	1.06

Figure 12

statsmodels.api.Logit ANOVA Output

```
Optimization terminated successfully.
         Current function value: 0.465285
         Iterations 6
                           Logit Regression Results
==============================================================================
Dep. Variable:            Loan_Status   No. Observations:                  518
Model:                          Logit   Df Residuals:                      512
Method:                           MLE   Df Model:                            5
Date:                Thu, 04 Jun 2020   Pseudo R-squ.:                 0.2492
Time:                        15:52:13   Log-Likelihood:                -241.02
converged:                       True   LL-Null:                       -321.03
Covariance Type:            nonrobust   LLR p-value:                 9.785e-33
==============================================================================
                     coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const             -1.7883      0.513     -3.483      0.000      -2.795      -0.782
DualIncome_IO      0.4763      0.237      2.007      0.045       0.011       0.941
CreditHistory_IO   3.7976      0.448      8.472      0.000       2.919       4.676
PA_Urban          -0.5958      0.284     -2.099      0.036      -1.152      -0.039
PA_Rural          -0.7902      0.282     -2.800      0.005      -1.343      -0.237
Debt_Equity       -0.0162      0.008     -2.134      0.033      -0.031      -0.001
==============================================================================
```

Figure 13

Logistic Regression Model Score Statistics and Confusion Matrix

Accuracy	81.1%
Precision	0.792
Recall	0.983

	Y	N
Y	351	6
N	92	69

F-Score	0.877
AUC-ROC	0.706
FPR	57.1%

Figure 14

Random Forest Model Variable Importance

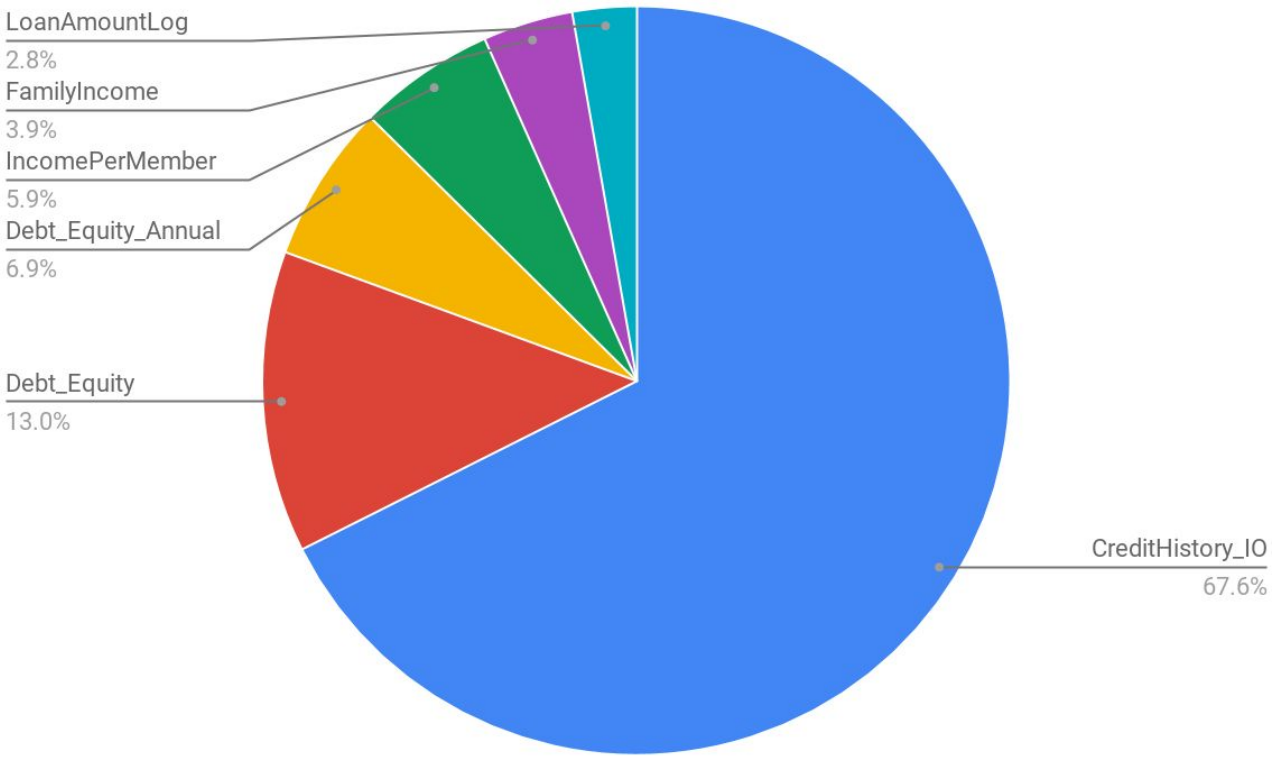

Figure 15

Random Forest Model Score Statistics and Confusion Matrix

Accuracy	83.6%

Predicted

Precision	0.822
Recall	0.974
F-Score	0.892
AUC-ROC	0.750
FPR	47.34%

	Y	N
Y	370	10
N	80	89
