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 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.

Missing values were replaced with the mode for Self_Employed and Loan_Amount_Term. There were no groupings that helped to indicate any other relationship.

Credit_History required the most effort to find the best imputation. No groupings were helpful in finding a breakdown. A model was created to help pick apart relationships, however the model just predicted the mode (Credit history was satisfactory). Missing values were replaced with the mode for Credit_History as well.

Missing Gender values were replaced as female if the applicant was not married and had 1 dependent. Otherwise the applicant was assumed to be male. Figure 8 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 9 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 10 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).

Model Build

Many model types and methods were built, in order to find the best method. Consistently, logistic regression and random forest with limited nodes were the top performing models. Removing outliers was also the best performing method. Therefore, FamilyIncome outliers and LoanAmountLog outliers were removed from the data. A logistic regression model and a random forest model with a maximum depth of 5 nodes were built. The model builds started with all variables, even if EDA suggested they wouldn't be predictive, and removed the least important variable until all variables met a threshold level. In order to be kept for the logistic regression model, the variable's P-value had to be less than 0.05. For the random forest model, in order to be kept, the variable importance had to be greater than 0.01.

Results

Figure 11 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 12.

Logistic Regression Model Score Statistics and Confusion Matrix

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 13 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 14.

Random Forest Model Score Statistics and Confusion Matrix

Accuracy	81.9%
Precision	0.802
Recall	0.978
F-Score	0.881
AUC-ROC	0.722
FPR	53.4%

	Predicted	
	Y	N
Y	349	8
N	86	75

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

- Lower the false positive rate of the models to improve accuracy. Try starting with a better imputation technique of Credit_History.
- 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

Loan Approval Status Proportions by Dependents Category

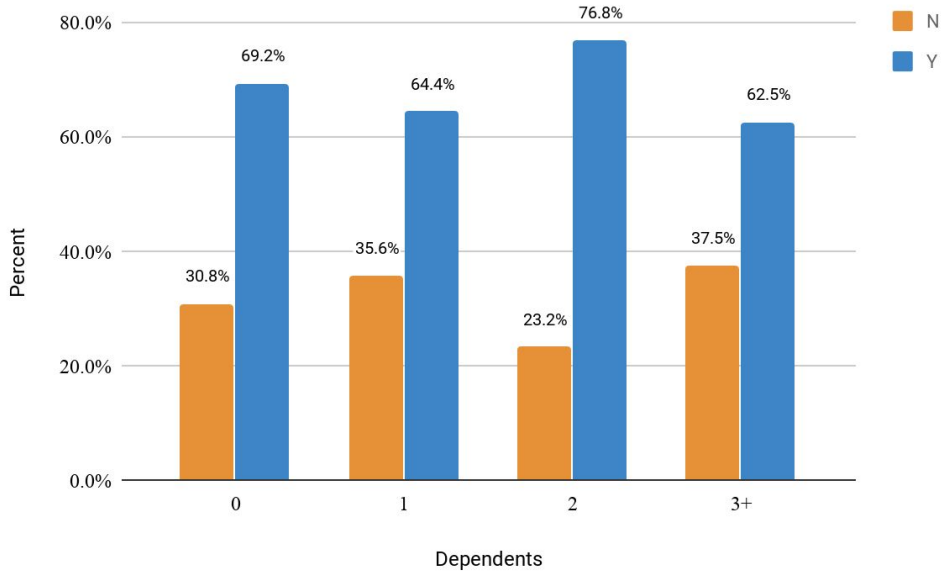

Figure 4

Loan Approval Status Proportions by Education Category

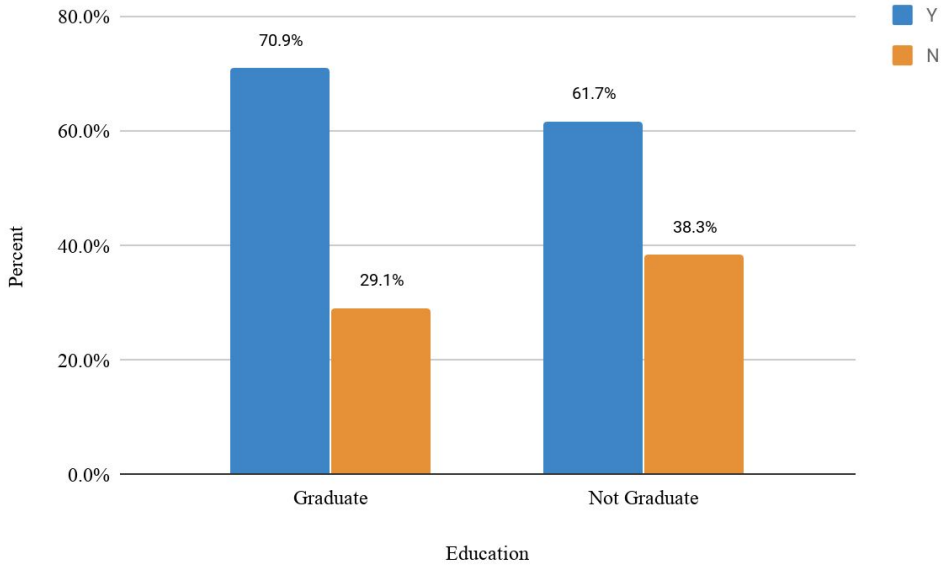

Figure 5

Loan_Term_Amount Counts & LoanTermGroup Counts

Before Grouping

Value	Count
360.0	512

After Grouping

Value	Count
30	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

15	44
(15, 30)	17
>30	15
<15	12

Figure 6

Loan Approval Status Proportions by Credit History Category

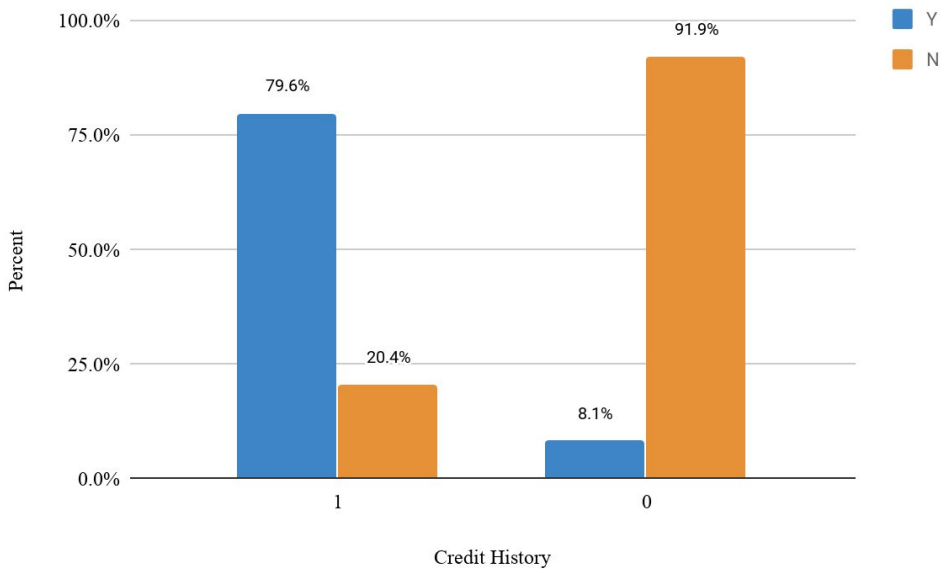

Figure 7

Loan Approval Status Proportions by Income (Quartiles)

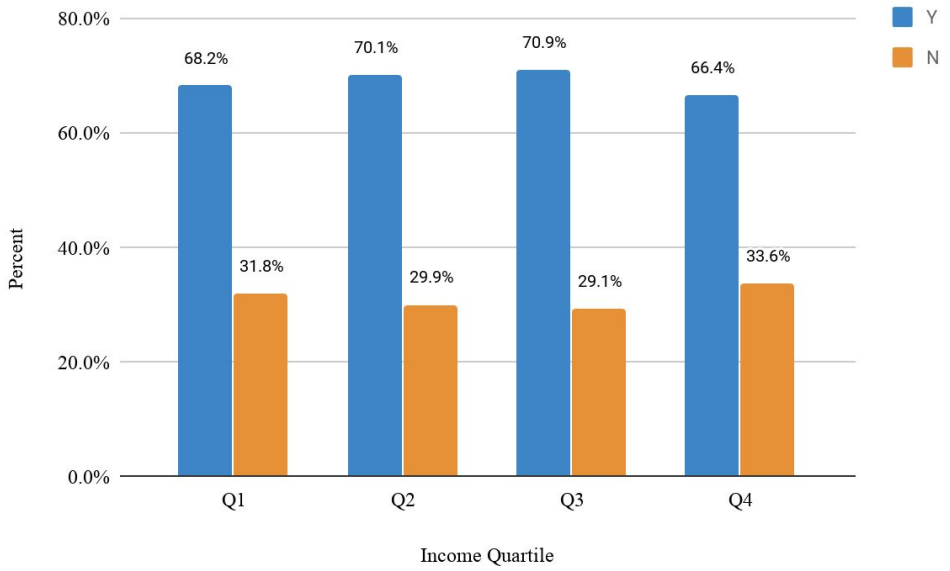

Figure 8

Proportion of 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 9

Proportion of Married by Gender Grouping

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

Figure 10

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 11

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 12

Logistic Regression Model Score Statistics and Confusion Matrix

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

	Y	N
Y	351	6
N	92	69

Figure 13

Random Forest Model Variable Importance

CreditHistory_IO	0.551
Debt_Equity	0.179
Debt_Equity_Annual	0.107
LoanAmountLog	0.058
Education_IO	0.03
PA_Semiurban	0.029
PA_Rural	0.027
Dependents	0.019

Figure 14

Random Forest Model Score Statistics and Confusion Matrix

Predicted

Accuracy	81.9%
Precision	0.802
Recall	0.978
F-Score	0.881
AUC-ROC	0.722
FPR	53.4%

	Y	N
Y	349	8
N	86	75
