Assignment 2: ML Pipeline

CAPP 30254: Machine Learning for Public Policy, Spring 2016

Christine Chung

1. Introduction

Historical data was provided on 250,000 borrowers in order to generate a delinquency score. That is, the goal was to try and predict the probability that a person would experience 90 days past due delinquency or worse on a credit payment.

2. Process

2.1. Data Exploration

The variable of interest is *serious delinquencies*, noted by *serious_d1qin2yrs*. It is a binary variable equal to 1 when a person experienced 90 days past due delinquency or worse and 0 otherwise.

Figure 1 shows the histogram of serious delinquencies. Roughly 7 percent of the training sample experienced a serious delinquency.

The dataset also includes 10 features or independent variables describing each borrower. Figures 2 and 3 show the names of all of the variables and their summary statistics.

For additional histograms for each of the variables, please see Appendix 1.

2.2. Data Processing

Two columns were missing observations: *Monthly Income* and *Number of Dependents*. Of the two, monthly income was by far the most troublesome with 29,731 missing values.

In order to best fit the model, we first needed to fill in or impute the missing values with best guesses.

For monthly income, we imputed the missing values using the mean monthly income of $6670. We also tried using median of $5400, but it provided negligible gains. In this case, both measures of centrality had a similar impact on the distribution.

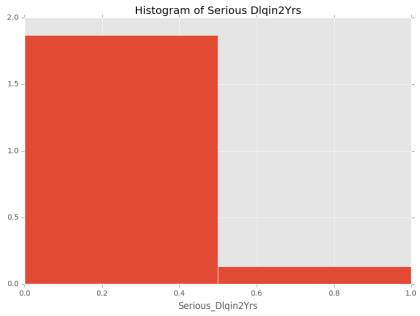

Figure 1. Histogram of Dependent Variable: Serious Delinquencies

Variable Number and Name	count	std
1 serious_dlqin2yrs	150000	0.25
2 revolving_utilization_of_unsecured_lines	150000	249.76
3 age	150000	14.77
4 number_of_time30-59_days_past_due_not_worse	150000	4.19
5 debt_ratio	150000	2037.82
6 monthly_income	120269	14384.67
7 number_of_open_credit_lines_and_loans	150000	5.15
8 number_of_times90_days_late	150000	4.17
9 number_real_estate_loans_or_lines	150000	1.13
10 number_of_time60-89_days_past_due_not_worse	150000	4.16
11 number_of_dependents	146076	1.12

Figure 2. Summary Statistics

Var. No.	min	25%	50%	75%	max	Mode	Mean
1	0	0.00	0.00	0.00	1	0	0.07
2	0	0.03	0.15	0.56	50708	0	6.05
3	0	41.00	52.00	63.00	109	49	52.30
4	0	0.00	0.00	0.00	98	0	0.42
5	0	0.18	0.37	0.87	329664	0	353.01
6	0	3400.00	5400.00	8249.00	3008750	5000	6670.22
7	0	5.00	8.00	11.00	58	6	8.45
8	0	0.00	0.00	0.00	98	0	0.27
9	0	0.00	1.00	2.00	54	0	1.02
10	0	0.00	0.00	0.00	98	0	0.24
11	0	0.00	0.00	1.00	20	0	0.76

Figure 3. Summary Statistics Continued

Variable	Number of Null Values
monthly_income 5000	29731
number_of_dependents 0	3924

Figure 4. Null Values

For the number of dependents, we used the mode value of 0. By far, most borrowers reported 0 dependents, so this seemed like the most appropriate value for imputation.

After cleaning and imputing the data, we split this dataset with 85 percent in training set and 15 percent in the testing set.

To review the data post imputation, please refer to the included document *imputed_data.csv*.

2.3. Feature Building

We first included every variable available to us:

1. Revolving Utilization of unsecured Lines
2. Age
3. Number of Times Past Due 30-59 days
4. Number of Times Past Due 60 - 89 days
5. Debt Ratio
6. Monthly Income
7. Number of Open Credit Lines and Loans
8. Number of Times 90 Days Late
9. Number of Real Estate Loans or Lines
10. Number of Dependents

In addition, we included three additional variables:

- **Age^2:** It is often the case that age has a non-linear effect on the dependent variable. In this case, one's risk of serious delinquencies could be positively correlated with age up until a certain point, at which point the effect of age could begin decreasing. Adding this term allows us to more accurately model the effect of age in the case of a non-linear relationship.
- **Log Income:** From Figure 6, we can see that monthly income is positively skewed, with most of the mass toward the left and long right tail. Logging the variable accounts for this asymmetric distribution.
- **Number of Dependents - Binary Variables:** We created binary variables for each number of dependents we saw in our training sample.

To review the data used for training, please refer to the included file *training.csv*.

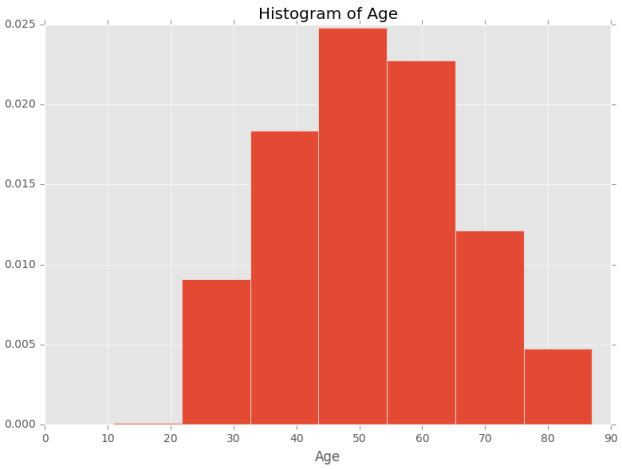

Figure 5. Histogram of Age

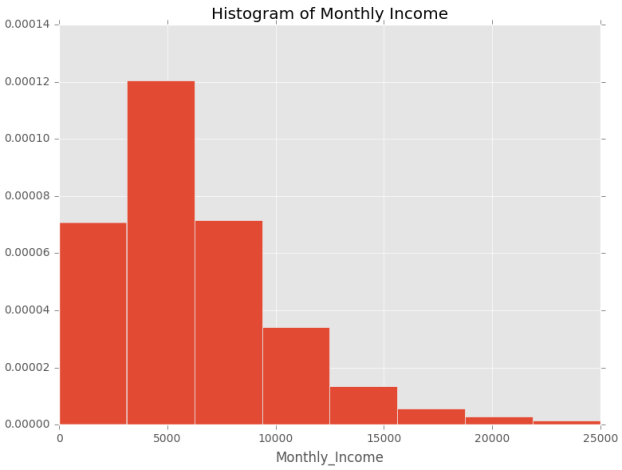

Figure 6. Histogram of Monthly Income

Model	Accuracy Score
Logistic Regression	0.933
K-Nearest Neighbors	0.932

Figure 7. Accuracy Scores

2.4. Model Building

We tried two different models: Logistic and K-Nearest Neighbors (KNN).

We used an accuracy measure to score the models (see Figure 7). This score measures the number of observations in the split test set that were correctly predicted by our model.

Both models scored similarly. Over several 5 trials, logistic regression consistently scored better than KNN, though the difference in accuracy was negligible.

In the end, we decided to go with logistic regression because of the slight edge that it provided and the ability to analyze coefficients.

2.5. Results

Of the 101,503 observations in our testing set, we predicted that 512 or .5% of people were at risk of delinquency. That is, 512 observations received a greater than 50% chance that they would experience a 90 day past due delinquency or worse.

This percentage is significantly less than the percentage of 7% that we saw in our training set. In addition, our accuracy score of 93% is the same as the percentage of people who are were not seriously delinquent in our training set. This means that we could have predicted that every borrower would not be delinquent and still obtain the same accuracy. This is likely the case since the model predicts very few serious delinquencies.

Going forward, we will tune the model by refining our list of features and trying additional classifiers.

For more discussion of the coefficient and interpretation of the model, please see Appendix 2.

The results and predictions can be found in the file *logistic_predictions.csv*. It includes two additional columns: *predictions* is a binary variable with 1 if the borrower has a greater than 50% chance of becoming seriously delinquent and 0 otherwise. *Probability* is a number between 0 and 1 that reflects the probability a borrow will be come seriously delinquent.

3. Appendices

3.1. Appendix 1: Histograms of Features

See next page.

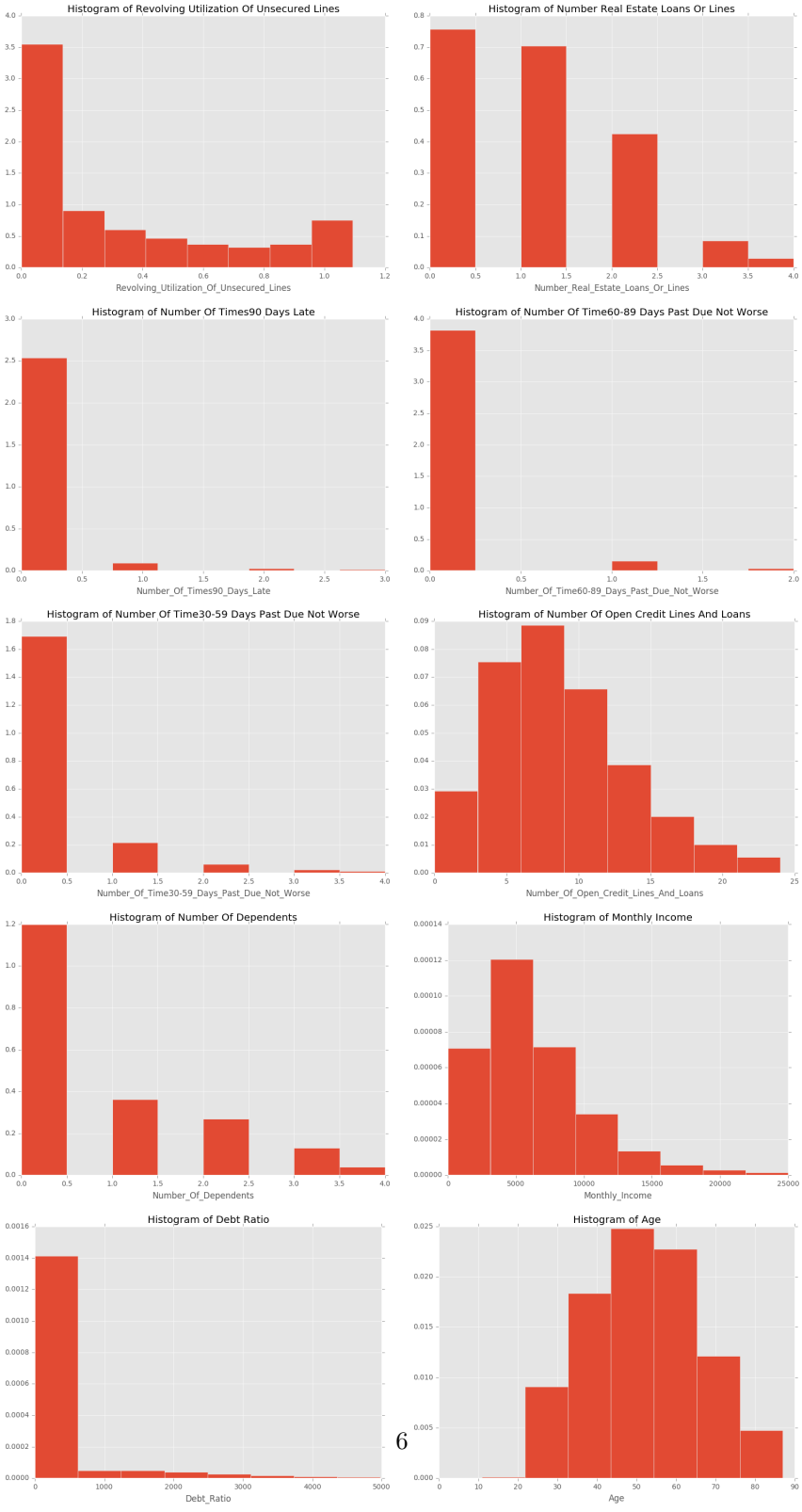

Figure 8. Histograms of Features

3.2. Appendix 2: Coefficients from Logistic Regression

Below are the coefficients for the logistic regression. We assume that a positive coefficient denotes that the variable has a positive relationship with serious delinquency (more likely to be delinquent) and that a negative coefficient represents a negative relationship (less likely to be delinquent).

Some interesting insights are that the number of times you are past due 30-59 days or 90 days late, the number of real estate loans you have, age, and number of dependents are all positive predictors of becoming seriously delinquent.

However, it is worth noting that many of these variables are not significant. The next model would would refine the features used and provide more useful, interpretable results.

The numbered columns represent binary variables for number of dependents.

Coefficient	**Variable**
-2.3674e-05	Revolving Utilization of Unsecured Lines
-5.601e-02	Age
5.212e-01	Number of Times Past Due 30-59 days
-1.998e-05	Debt Ratio
-2.384e-05	Monthly Income
-9.426e-03	Number of Open Credit Lines and Loans
4.444-01	Number of Times 90 Days Late
4.4783e-02	Number of Real Estate Loans or Lines
-9.340e-01	Number of Times Past Due 60 - 89 days
1.072e-02	Number of Dependents
-4.146e-02	Log Monthly Income
4.500e-02	Age^2
-1.420e-01	0
2.384e-04	1
-2.991e-02	2
-8.531e-03	3
1.917e-02	4
-6.992e-03	5
6.316e-03	6
-1.044e-04	7
-1.043e-03	8
-3.232e-04	9
-6.792e-04	10
-1.905e-04	13
-3.611e-04	20