Paul Mack
4/12/2016
HW2 Write Up

Predicting Delinquent Loans:

Introduction:

Having a loan more than 90 days past due is a relatively rare event. We were given the task to predict these serious loan delinquencies. Examining a historical record of 150,000 loans we only observed 6.684% of the loans were more than 90 days past due in the time period observed.

From the historical record of delinquent loans we have made predictions for an additional 101,503 of loan records whose current delinquent status is unknown. These predictions are saved in the file "predictions.csv"

Summary of Data:

The historical record contained 10 additional pieces of information on each loan besides whether it had passed the 90 day delinquency mark. We used these additional 10 pieces of information to predict the status. The variables included information on whether they had missed loan payments in the past 2 years, personal information like age and income, and information about their number of loans.

Many of the variables appeared to have different distributions for those who were delinquent and those who were not. A few examples are included in the Appendix, but in general it seems that those who were delinquent on loans tended to be younger and have less income. Interestingly, delinquent loans also tend to have a lower debt_ratio meaning that a person with a delinquent loan is, on average, paying less debt relative to their income than a person who is not delinquent.

Data Cleaning:

Monthly Income and Number of Dependents each were missing values for many records. Monthly income was missing nearly 20% of its records whereas Number of Dependents was missing for a little more than 2%. The missing values were filled in with overall mean in the population so that the observation could be included.

Constructing A Model:

We used a logistic regression to predict whether a person has a delinquent loan. The logistic model considers all of the 10 pieces of information provided and makes a prediction as to whether the loan is delinquent.

To construct our model we randomly split our historical record into a training set consisting of 80% of the data and a testing set on which we validated our model consisting of 20% of the data.

We also converted Revolving Utilization Of Unsecured Lines and Monthly Income into quartiles before including them into our model. Quartiles were chosen because much of the data is skewed quartiles evenly distributes them. The two variables above were chosen because when each was individually tested as a categorical variable it improved the accuracy of the predictions for the 20% of the data set aside.

Future Research:

In the future, we should closer examine how we are filling in the missing values for income and number of dependents since missing values are likely independent from whether a person will be delinquent on a loan.

We should also reexamine how we converted our continuous variables into quartiles, as quartiles are likely not the optimal bins for our categories.

Our cross validation process could be improved for how we selected which variables to include as categorical data – as we get different results when we take different samples into the training and testing sets. Furthermore, when each category is tested independently it does not appear that they necessarily improve the models accuracy from our baseline jointly.

Appendix:

Age Delinquent
max,101.0
mean,45.93
median,45.0
min,21.0
null_count,0.0
standard_deviation,12.92

Age Non-Delinquent:
max,109.0
mean,52.75
median,52.0
min,0.0
null_count,0.0
standard_deviation,14.79

Debt Ratio for Delinquent:
max,38793.0
mean,295.12
median,0.43
min,0.0
null_count,0.0
standard_deviation,1238.36

Debt Ratio for Non-Delinquent
max,329664.0
mean,357.15
median,0.36
min,0.0
null_count,0.0
standard_deviation,2083.28

Monthly Income for Delinquent:
max,250000.0
mean,5630.83
median,4500.0
min,0.0
null_count,1669.0
standard_deviation,6171.72

Monthly Income for Non-Delinquent
max,3008750.0
mean,6747.84
median,5466.0
min,0.0
null_count,28062.0
standard_deviation,14813.5