Alden Golab
4/12/2016
CAPP 30254
PA #2

Credit worthiness is a significant prediction problem for banks whose results can be improved through the use of Machine Learning techniques. Here, we analyze historical data for 150,000 customers provided by the client to train and develop a supervised logistic classifier. I go over the nuances below and achieve an accuracy of 93%. I then apply this to the test dataset of approximately 101,000 individuals.

A basic look through the dataset revealed substantial missing data: approximately 19.8% of entries were missing incomes and 2.6% of entries were missing dependents data. As these seem quite relevant to the prediction at hand, these entries were accounted for using mean and conditional mean methods so that we were able to use the full dataset, improving the predictive power of the other variables without harming our overall accuracy. By far the majority of individuals in the dataset make less that $20,000 per month; the median is around $5,400 for an annualized gross income of $64,800. Additionally, after imputing data for child dependents, we see that a substantial number of individuals within the dataset have no children.

Indeed, looking through the data, we find that this seems to be a remarkably well-off group of individuals: no children, high median income, and an average of 8.4 credit lines or loans per customer. It is thus no surprise, then, that very few have been seriously delinquent within the past two years: only 6%.

Figure 1: Histogram of Monthly Incomes below $20,000, after imputation

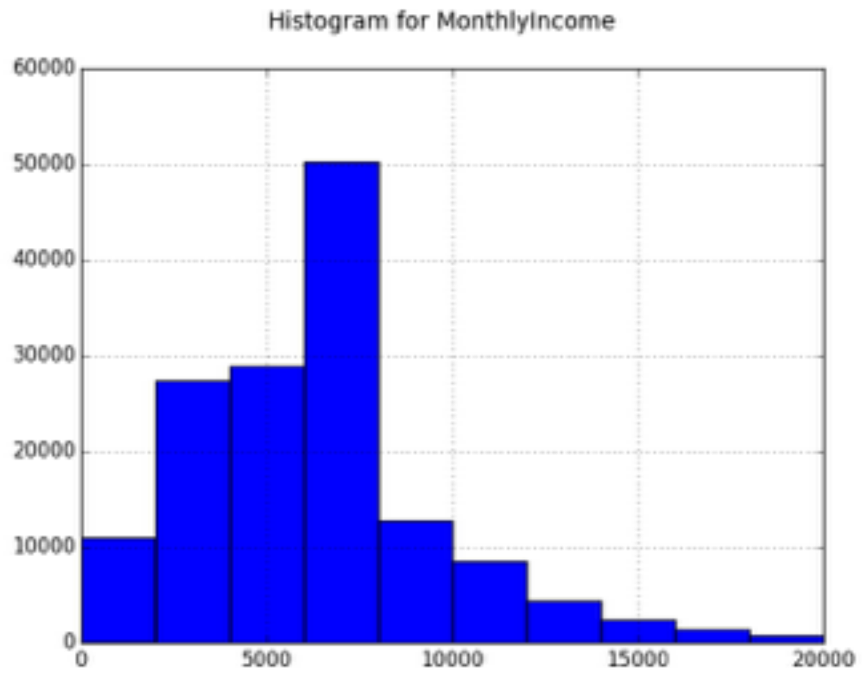
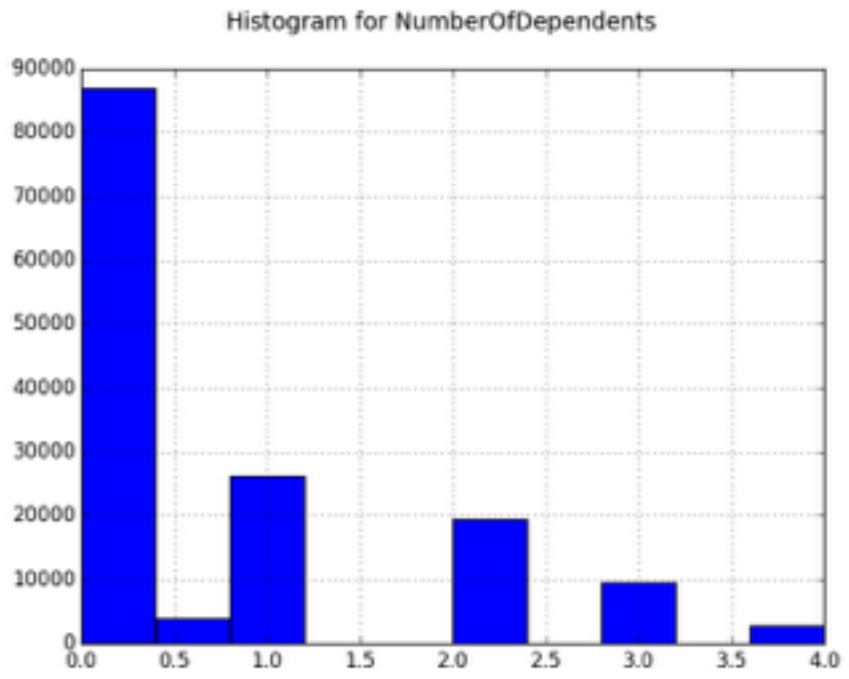


Figure 2: Histogram of Number of Dependents after imputation



Indeed, there is very little variation in most of the variables: Number of times past due on any measure of time, number of real estate loans or lines, and revolving utilization of unsecured lines. Even age was skewed older, with the mean age of 52—far above average.

For this reason, predicting new cases of default is expected to be difficult. I trained a model using the historical training set and achieved a 93.4% accuracy rate, which looks great. However, since 6.68% of individuals within the training set have been seriously delinquent, this means that guessing 'No serious delinquency' for all individuals within the dataset would result in a 93.3% accuracy rate. That is to say: the trained model only achieves a .1 percentage point increase in accuracy over always predicting no delinquency.

Looking at the coefficients, we also find the that the data provided were not terribly helpful with our prediction. The two most impactful predictors were the number of times an individual was 30-59 days past due and number of times an individual was 90 days late. As mentioned before, neither of these variables has substantial variation, despite their predictive power. You'll notice in Table 2 that, while Standard Deviation is high, the mean in below one for both values; indeed, looking at the histograms in Figure 3, we see that these values are highly skewed, presenting very little actual variation to use for model building.

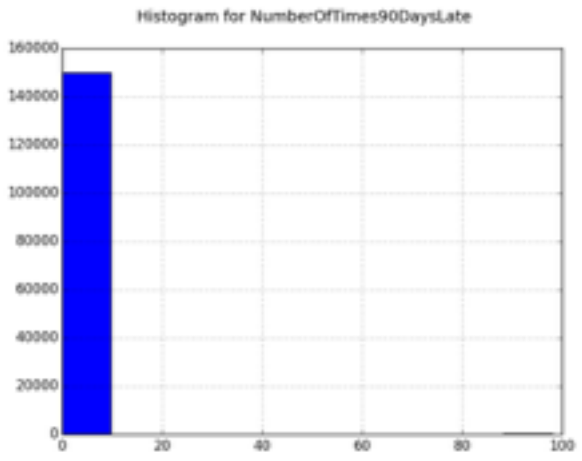
Table 1: Coefficients from logistic model

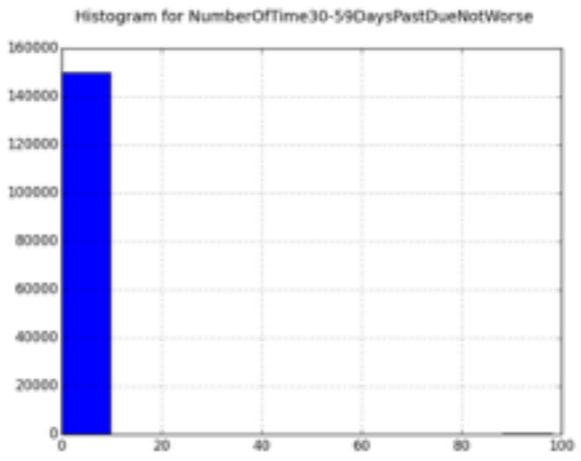
Variable	Coefficient
RevolvingUtilizationOfUnsecuredLines	-0.0000392
age	-0.030
NumberOfTime30-59DaysPastDueNotWorse	0.485
DebtRatio	-0.0000182
MonthlyIncome	-0.0000368
NumberOfOpenCreditLinesAndLoans	-0.008
NumberOfTimes90DaysLate	0.418
NumberRealEstateLoansOrLines	0.057
NumberOfTime60-89DaysPastDueNotWorse	-0.871
NumberOfDependents	0.086

Table 2: Descriptive statistics for selected variables

Variable	Standard Deviation	Mean
Number of times 30-59 days past due	4.19	0.42
Number of times 90 days past due	4.17	0.27

Figure 3: Histograms for selected variables





As such, we cannot expect reasonable accuracy in the model. Indeed, running on the test set, we see a predicted .004% rate of delinquency, which is too small to be taken seriously. Next steps are to try another Machine Learning classification technique and see if improvements can be made. Additionally, provision of additional covariates around customers by the Client may be helpful; nevertheless, this may not be possible.