Homework 2

Michael Fosco

4/12/2016

1 Results

We are examining the rate of delinquency in the past two years on people's credit. The rate of delinquency appears low at around .06684 and so when we run our classifier we will need an accuracy rate above 93% to be better than always assuming no delinquency. We have a wide range of ages, with the mean being the mid fifties, suggesting that we have a decent representative swath of the working populace. It appears that the number of open credit lines and loans is skewed to the right like many other variables, as should be expected since there is a lower bound of zero for the poor but no effective upper bound on the rich. The only other main information that can be gleamed from the correlation table and summary statistics is that the greater the number of days late one is, the higher the correlation with a serious delinquency in the past two years. This relationship is what should be expected since if you fall behind on your payments then you are more likely to fall into delinquency. I decided to fill in the the missing values for any given variable with that variable's mean.

I tested three different types of classifiers using the machine learning pipeline that I built. The three types of classifiers I tested included logistic regression, linear SVM, and K-nearest neighbors (KNN). For the logistic regression model I varied the tolerance level for its rate of learning from $1e-7$ to 1 as well as the C parameter from .1 to 25. The C parameter functions similarly to the parameter determining the bandwidth for SVM. For SVM, I varied the tolerance parameter from 1e-7 to 1 as well as letting the max number of iterations vary from 500 to 2,000. The C parameter for SVM was also varied from .1 to 25. For KNN I varied whether the weight of a neighbor was uniform versus dictated by the distance of the points. I also varied the number of neighbors included from 2 to 1,000 as well as the leaf size from 15 to 120.

I varied all of these parameters in an attempt to find the best classifier and parameter combination based on the calculated accuracy on the training dataset. The best classifier was a KNN model with weight given by the calculated distance and a leaf size of 120. The calculated accuracy was .9996 which is much better than an accuracy of .93316, which is what we would get if we only guessed no. However, this model is most likely overfitting since no validation set was created to test the model against. Nevertheless, predictions were created for the testing dataset and stored in the github repository within hw/hw2/predictions.csv. The pipeline created in python that easily enabled this analysis is stored in the github repository within hw/hw2/hw2.py.

2 Tables

2.1 Descriptive Statistics

	Unnamed: 0	SeriousDlqin2yrs	RevolvingUtilizationOfUnsecuredLines	age
count	150000	150000	150000	150000
mean	75000.5	0.06684	6.04844	52.2952
std	43301.4	0.249746	249.755	14.7719
min	1	0	0	0
25%	37500.8	0	0.0298674	41
50%	75000.5	0	0.154181	52
75%	112500	0	0.559046	63
max	150000	1	50708	109
missing values	0	0	0	0

	NumberOfTime30-59DaysPastDueNotWorse	DebtRatio	MonthlyIncome
count	150000	150000	120269
mean	0.421033	353.005	6670.22
std	4.19278	2037.82	14384.7
min	0	0	0
25%	0	0.175074	3400
50%	0	0.366508	5400
75%	0	0.868254	8249
max	98	329664	3.00875e+06
missing values	0	0	29731

	NumOfOpenCreditLinesAndLoans	NumOfTimes90DaysLate	NumRealEstLoansOrLines
count	150000	150000	150000
mean	8.45276	0.265973	1.01824
std	5.14595	4.1693	1.12977
min	0	0	0
25%	5	0	0
50%	8	0	1
75%	11	0	2
max	58	98	54
missing values	0	0	0

	NumberOfTime60-89DaysPastDueNotWorse	NumberOfDependents
count	150000	146076
mean	0.240387	0.757222
std	4.15518	1.11509
min	0	0
25%	0	0
50%	0	0
75%	0	1
max	98	20
missing values	0	3924

2.2 Correlation Table

	Unnamed: 0	SeriousDlqin2yrs	RevolvUtilOfUnsecLines
Unnamed: 0	1.000000	0.002801	0.002372
SeriousDlqin2yrs	0.002801	1.000000	-0.001802
RevolvUtilOfUnsecLines	0.002372	-0.001802	1.000000
age	0.004403	-0.115386	-0.005898
NumTime30-59DaysPastDueNotWorse	-0.000571	0.125587	-0.001314
DebtRatio	-0.002906	-0.007602	0.003961
MonthlyIncome	0.002356	-0.018002	0.006565
NumOpenCreditLinesAndLoans	0.004586	-0.029669	-0.011281
NumOfTimes90DaysLate	-0.001104	0.117175	-0.001061
NumRealEstateLoansOrLines	-0.000666	-0.007038	0.006235
NumTime60-89DaysPastDueNotWorse	-0.000777	0.102261	-0.001048
NumOfDependents	-0.000055	0.045621	0.001539

	age	NumTime30-59DaysPastDueNotWorse	DebtRatio
Unnamed: 0	0.004403	-0.000571	-0.002906
SeriousDlqin2yrs	-0.115386	0.125587	-0.007602
RevolvUtilOfUnsecLines	-0.005898	-0.001314	0.003961
age	1.000000	-0.062995	0.024188
NumTime30-59DaysPastDueNotWorse	-0.062995	1.000000	-0.006542
DebtRatio	0.024188	-0.006542	1.000000
MonthlyIncome	0.032984	-0.007636	-0.005355
NumOpenCreditLinesAndLoans	0.147705	-0.055312	0.049565
NumOfTimes90DaysLate	-0.061005	0.983603	-0.008320
NumRealEstateLoansOrLines	0.033150	-0.030565	0.120046
NumTime60-89DaysPastDueNotWorse	-0.057159	0.987005	-0.007533
NumOfDependents	-0.208102	-0.002525	-0.038287

	MonthlyIncome	NumOpenCredLines+Loans	NumTimes90DaysLate
Unnamed: 0	0.002356	0.004586	-0.001104
SeriousDlqin2yrs	-0.018002	-0.029669	0.117175
RevolvUtilOfUnsecLines	0.006565	-0.011281	-0.001061
age	0.032984	0.147705	-0.061005
NumTime30-59DaysPastDueNotWorse	-0.007636	-0.055312	0.983603
DebtRatio	-0.005355	0.049565	-0.008320
MonthlyIncome	1.000000	0.082319	-0.009484
NumOpenCreditLinesAndLoans	0.082319	1.000000	-0.079984
NumOfTimes90DaysLate	-0.009484	-0.079984	1.000000
NumRealEstateLoansOrLines	0.113823	0.433959	-0.045205
NumTime60-89DaysPastDueNotWorse	-0.008259	-0.071077	0.992796
NumOfDependents	0.058542	0.064507	-0.009579

	NumRealEstLoans,Lines	NumTime60-89DayPastDue	NumDepend.s
Unnamed: 0	-0.000666	-0.000777	-0.000055
SeriousDlqin2yrs	-0.007038	0.102261	0.045621
RevolvUtilOfUnsecLines	0.006235	-0.001048	0.001539
age	0.033150	-0.057159	-0.208102
NumTime30-59DaysPastDueNotWorse	-0.030565	0.987005	-0.002525
DebtRatio	0.120046	-0.007533	-0.038287
MonthlyIncome	0.113823	-0.008259	0.058542
NumOpenCreditLinesAndLoans	0.433959	-0.071077	0.064507
NumOfTimes90DaysLate	-0.045205	0.992796	-0.009579
NumRealEstateLoansOrLines	1.000000	-0.039722	0.123370
NumTime60-89DaysPastDueNotWorse	-0.039722	1.000000	-0.010277
NumOfDependents	0.123370	-0.010277	1.000000

2.3 Graphs

Figure 1:

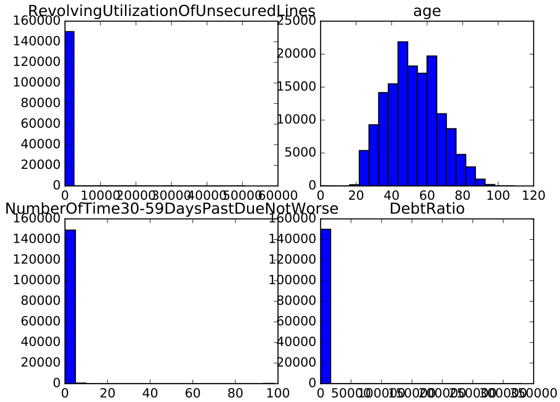


Figure 2:

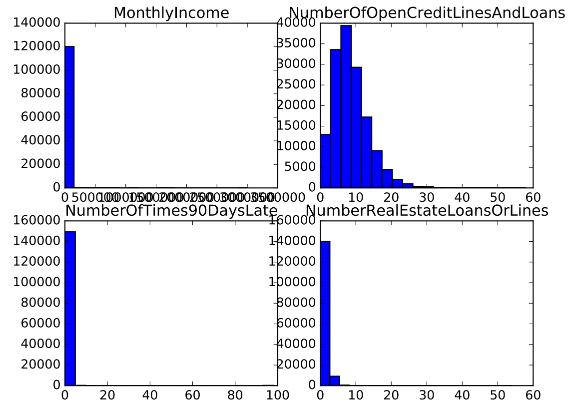


Figure 3:

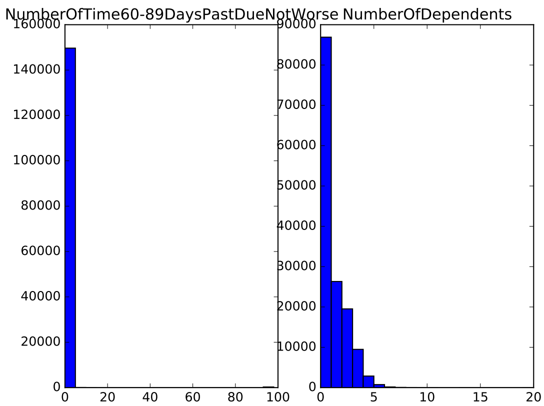
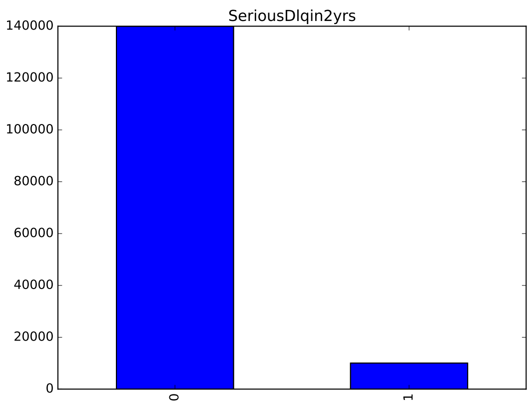


Figure 4:


SeriousDlqin2yrs
140000
120000
100000
80000
60000
40000
20000
0
0
1