NEW YORK UNIVERSITY STERN Leonard N. Stern School of Business

DATA MINING FOR BUSINESS ANALYTICS INTRODUCTION TO DATA SCIENCE

Spring 2015

Homework #3 **Due: by 11:55pm Fri 3/13/15**

Sahil Goel

(put your name above)

Total grade: _____ out of _____ points

Please answer all questions/follow all directions. Put your name above, and include your last name in the filename of your homework submission.

1) **Once again we will look at our churn problem from Homework 2. This time you should try to find the best possible model in terms of accuracy (with default cutoff at 0.5) and explore some of the evaluation concepts we have talked about in class. The first set of questions should be answered using the churn_train.arff ONLY!**

Include answers to the following:

a) What is the baserate of the churn task (there are a number of different ways to provide the correct answer – just be specific)? How would you characterize it (balanced vs. very skewed)?

Total number of entries in churn_train :- 14000
Class Leave :- 7339
Class zero : 6661

Base rate = 7339/14000 or 6661/14000
= 52.42% or 47.57%

b) What is the dimensionality of the features? In relative terms - do you think it is high or low?
24 Dimensions: On the lower side

c) What is the majority class?
Leave

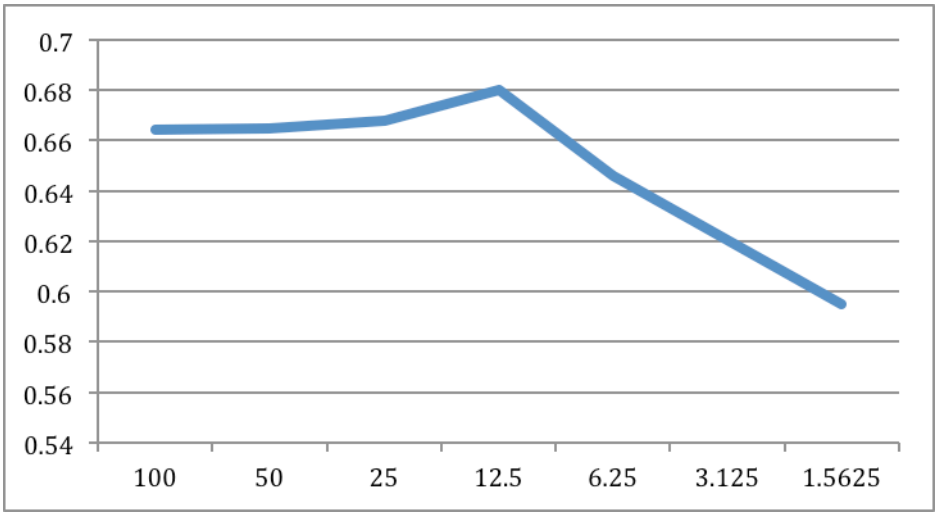
d) What accuracy would you expect to get if you always predict the majority class?
52.42%

2) Create two learning curves (using WEKA) of the out of sample AUC on the test set (churn_test.arfff) using both logistic regression and the decision tree J48 (just go with the default settings). In particular, starting from the full training set, after each iteration, reduce the training set to half until you reach less than 100 examples. Provide a plot with both curves (use some other tool such as EXCEL)

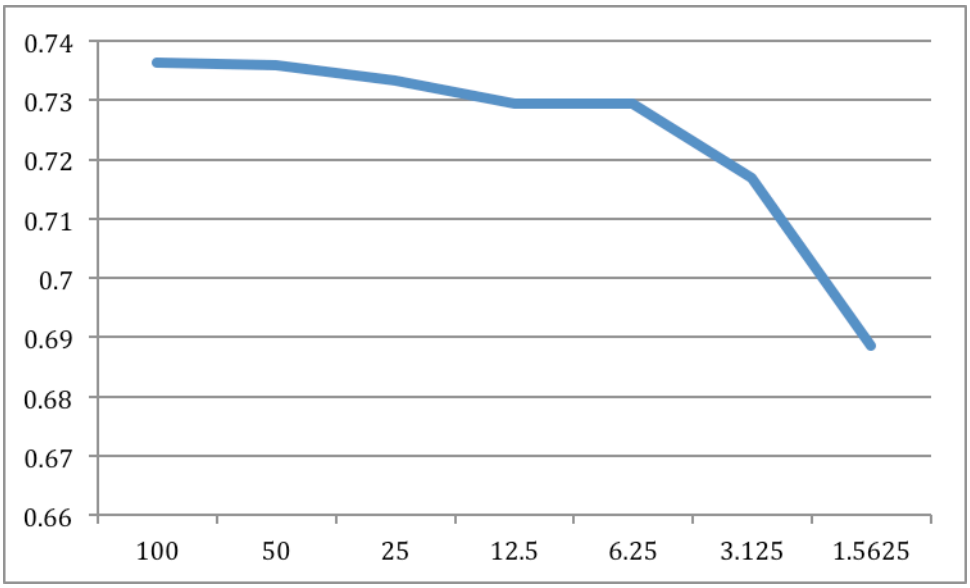
- *You can cut the dataset in half easily in Weka. In the Preprocess tab, in the box marked Filter, click on Choose. Under weka->filters->unsupervised->instance you will see RemovePercentage. Normally, it is a good idea first to run the filter Randomize, to make sure that you are removing the data randomly; real data often will be sorted based on some attribute, which can result in throwing away many data items with similar values. Don't Randomize for this assignment; the data for this assignment already will be randomized.*
- *The Undo button on the preprocess tab will undo the preprocessing (like Randomizing, RemovePercentage, etc.). Keep an eye on the data statistics (like the number of instances) in the preprocess tab to verify.*

2)

(a) Decision trees AUC(y-axis) vs Percentage of training Data used(x-axis)



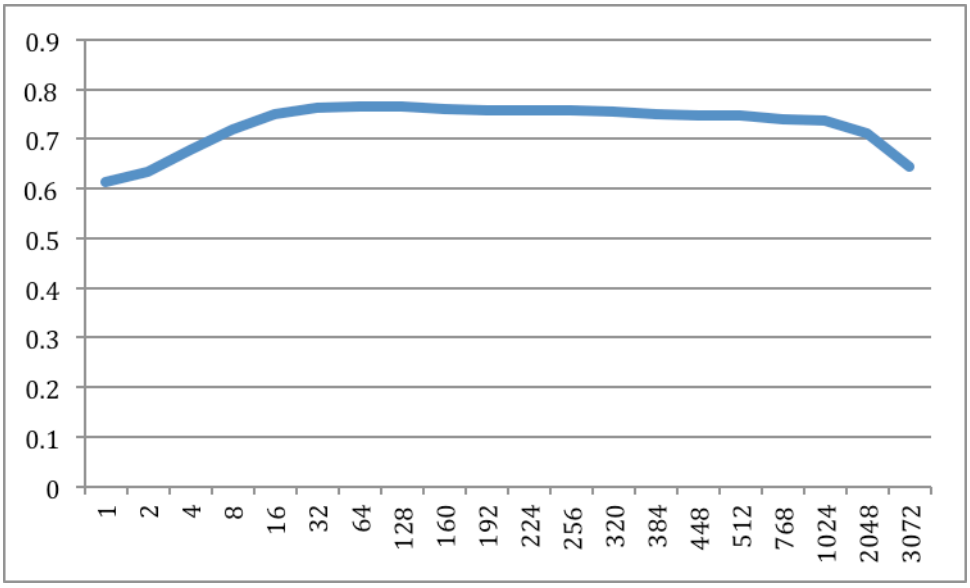
(b) Logistic Regression AUC(y-axis) vs Percentage of training Data used(x-axis)



3) Create a fitting curve of the generalization AUC for decision trees as a function of the MinNumObj parameter. First change the option 'unpruned' to 'true'. Provide a plot of the parameter and the resulting out of sample performance using either cross validation or a training/test split. What does the parameter do? What is the optimal selection for the parameter?

3).

Decision Trees: AUC(y-axis) using cross-validation v/s MinnumObj(x-axis)



Minnumobj :- Minimum number of instances a leaf can have in a decision tree.

Optimal Minnumobj :- In range of 128-160
At 128 AUC: .7663
At 160 AUC: .7619

4) Explore additional models and compare the performance using a training/test split of 33%. Include a decision tree, Logistic regression, Neuronal Networks (called MultiLayer Perceptron) and SVM (called SMO). The relevant complexity parameters for the neural network is the number of neurons in the hidden Layers for NN (default is the option 'a' – change to something else and read the description) and for the support vector machine the penalty term called c. Provide evidence (table) of your selection of what you think is the best overall method/parameter combination in terms of generalization AUC using only the churn_training.arff dataset and once you made your choice report the accuracy on the churn_test.arff data. Important here is that you can only look at the test set once at the end. All the selection and picking the best algorithm has to be done on the training set using either the recommended split or cross validation.

4).
Best AUC performance with tuned parameters(At percentage split of 66/33):
Decision Trees(128 Minnumobj): .7702

AUC	Minnumobj
0.7702	128

Logistic Regression: .7523

Neuron Networks: .7576

AUC	numofhidden layers
.7476	1
.7539	2
.7554	3
.7576	4
.7557	5
.7539	6

SVM: .6806

AUC	c
.6802	1.0
.6806	1.1
.6804	1.2
.6799	2.0

Best option is to go with decision trees and min num of objects around 128.
Accuracy on test set :- 70.08%

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.768	0.37	0.688	0.768	0.726	0.757	LEAVE
	0.63	0.232	0.718	0.63	0.671	0.757	STAY
wghtd Avg.	0.701	0.303	0.703	0.701	0.699	0.757	