

**Course:**

Intro to Data Science – DS-GA-1001 /

Data Mining for Business Analytics - INFO-GB.3336.11

Fall 2014

**Instructor:**

Brian Dalessandro

**Homework 3 – Due 10/22/2014 at 5 pm**

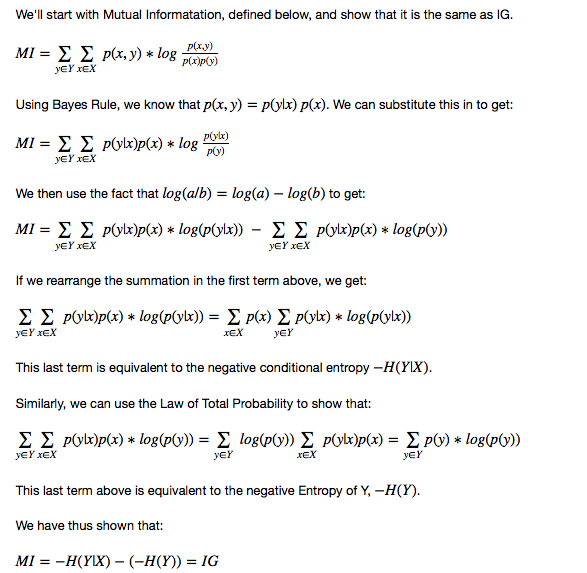
**(Submissions must be a WORD or PDF document)**

Name:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Grade Total: \_\_\_\_\_\_\_\_\_\_ out of 10**

**Part 1 (2 Points)**

**Show mathematically that Mutual Information = Information Gain,**



**Part 2 (4 Points)**

**This is a hands-on task where we build a predictive model using Decision Trees discussed in class.**

1. Download the Cell2Cell\_data.csv from NYU Classes > Resources > Data. Also read the Cell2Cell\_info.pdf to understand the contents of the data file.

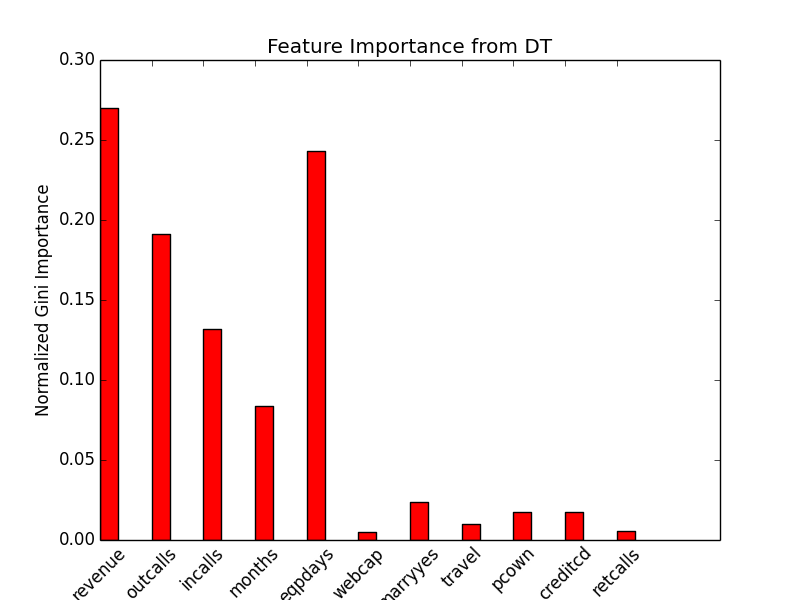
No write up expected

1. Load the data and prepare it for modeling:
   1. Define a training set (80% of data)
   2. Define a test set (20% of data)

No write up expected

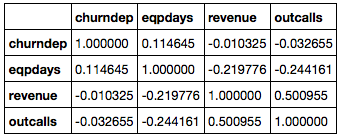
1. Now build a decision tree classifier on the training data to predict the churndep target variable. Make sure to use criterion='entropy' when instantiating an instance of the DecisionTreeClassifier. Also, use all of the default options.

**Question 2c: Using the results of above show a bar plot of features ranked by feature importance (hint: check the attributes of the DecisionTreeClassifier object).**

****

**Is the relationship between the top 3 most important features (as measured here) negative or positive?**

The following correlation table shows that eqpdays is positively correlated with churn while outcalls and revenue are negatively correlated.

****

The correlation chart here is not necessary to answer the question, but it is a quick way to see the direction of the relationship between the top 3 features and the outcome.

**If your marketing director asked you to explain the top 3 drivers of churn, how would you interpret the relationship between these 3 features and the churn outcome?**

Revenue and Outcalls are correlated and they generally describe the amount of usage the person has on their phone. We should expect that if the person uses their service a lot then they are both satisfied and getting their money’s worth. Thus, we should expect they are less likely to churn.

Eqpdays is less obvious, but we might assume that customers that have had their equipment longer are more likely to want a new phone. When confronted with the choice of a new phone, the customer might also consider shopping around with other providers to get the best deal.

**What “real-life” connection can you draw between the variable and churn.**

1. Using the classifier built in 2c:

**Question 2d: What is the accuracy on the test data?**

I got 53% using the default options and no optimization. This is not very good, which is what might happen when you rely on default settings. The exact accuracy will vary because of randomness in the train/test sets showing, but they shouldn’t be too much more or less than this.

**Part 3 (4 Points)**

**The default options for your decision tree may not be optimal. We need to do analysis on whether tuning the parameters can improve the accuracy of the classifier. For the following options [min\_samples\_split, min\_samples\_leaf]:**

* 1. Generate a range of 10 values of each that make sense to test
  2. For each combination of 3a (there should be 100), build a new classifier and check the classifier’s accuracy on the test data.

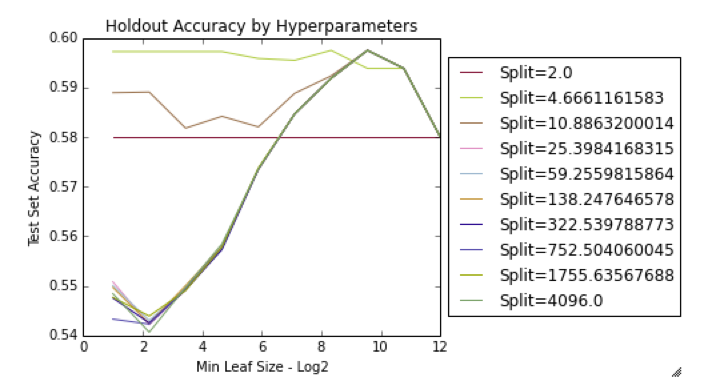
For my analysis I chose:

splits = np.logspace(1,12,num=10,base=2)

leaves = np.logspace(1,12,num=10,base=2)

For both I chose 10 numbers evenly spaced by powers of 2 to cover an even range on a logarithmic scale.

**Question 3b: Plot the test set accuracy against the options above. Use the values of min\_samples\_split as the x-axis and generate a new series (line) for each of min\_samples\_leaf. Copy and paste the plot below.**

****

**Which configuration returns the best accuracy?**

There are effectively a bunch of ties here. If the split is >10 then the best min\_leaf\_size is around 750. If using this analysis for model selection, I would choose a min\_split between 5 and 10, as the evidence above shows that these are the most robust. The lower one (around 5), has very little sensitivity to min\_leaf\_size. We can see that with split~5 that any leaf size is around the maximum accuracy

**What is this accuracy?**

With some tuning we can reach an accuracy of a little over 59%.

**If you were working for a marketing department, how would you use your churn production model in a real business environment? Explain why churn prediction might be good for the business and how one might improve churn by using this model.**

Customer churn is a negative outcome for a business, as it both reduces revenues and it increases the chance that customers might give negative reviews of the service. The negative reviews can then influence other customers to churn or prevent other customers from signing up. It is important for the business to predict churn to design marketing strategies that focus on those who are most at risk of churning. There are both short and long term strategies to be employed here.

In the short term, the company might want to find the most likely churners and offer some special incentive to convince them to stay with the company. An example might be a discount on a new phone or more minutes for the same price.

For the long-term strategy, the company can use the churn model to better understand the drivers of churn. With this understanding the company can think of ways to better improve the customer experience so that churn is prevented even before there is a direct intervention.