# CIS417 – Introduction to Business Analytics

**Fall 2017**

**Homework 6**

**Churn Part II**

Use the data frame churn saved in DSBchurn.Rda, that you created in the last homework, for this homework as well . If you cannot locate your copy of DSBchurn.Rda then redo steps 1-5 of the last homework to re-create it. Make sure that you did not discretize the numeric features (income, overage, leftover, house, phone_price, long_calls and av_duration). Rpart easily deals with measures and dimensions.

Business Scenario:

- There are 1,000,000 customers, of which 490,000 are likely to stay another year and 510,000 are likely to leave. The carrier does not know which customer is likely to stay or leave.
- A customer that stays will provide a NPV of $1,000 to the carrier. The NPV from a customer who leaves is $0.
- The firm has three strategies:
    a. Do nothing. Let the customers who are likely to leave, do so.
    b. Offer every customer a discount of $400 on a new phone. A customer who was going to stay will continue to stay and enjoy the discount. A customer who was likely to leave will now stay with a 0.50 probability. The discount is wasted on customers who end up leaving any way.
    c. Construct a prediction model to predict whether a customer will leave or stay. Use the prediction model to offer the discount only to the customers who are predicted to leave. The behavior of the customers who are predicted to leave and receive a discount remains as described above.
- Consider LEAVE to be the Negative Class and STAY to be the Positive class. Denote the FPR by α and FNR by β. Find the expected value of each of the three strategies. Write the expected value of strategy (c) in the form of:

$$v - a \,.\, \alpha - b \,.\, \beta$$

where $a$ is the False Positive Cost and $b$ is the False Negative Cost.

Use the DSBchurn.Rda data set to make a prediction model:

1. Split the data into training and test data sets with 66.7% and 33.3% of the observations, randomly chosen, respectively (use set.seed(3478) for reproduceability reasons).
2. Use rpart to build a large complex tree with a low value of cp and minsplit. Treating LEAVE as Negative and STAY as positive, determine the "Big Tree" error rates (FPR and FNR) using the test data set.
3. Find the best cp value to post-prune the tree. Use the test data set to find the "Pruned Tree" error rates. Save a PDF of a nicely formatted plot of the pruned tree.
4. Use ROCR to find the best threshold. Using this recommended threshold, determine the "Best Threshold Pruned Tree" error rates for the test data set.

5. For each of the error rates determined in steps 2, 3 and 4 above, find the expected values of strategy (c), for the firm. Display the result in a table with rows for Big Tree, Pruned Tree, and Best Threshold Pruned tree; and columns for FPR, FNR, Accuracy, Expected Value. The Expected Value column refers to the expected value of strategy (c).

Turn in 4 files:

(i) The R code
(ii) A PDF of a nice plot of the pruned tree.
(iii) A PDF of a Word document describing the calculations of the steps described above.
(iv) A PDF of a one page Executive Summary (fewer than 300 words), aimed at senior management describing the best strategy.