**Predictive Analytics**

**HW5: Churn assignment**

Use the Churn data from a telecom company to understand what factors are good predictors of churn. Use a logistic regression model to build the best model (that is **best** in terms of model fit criteria).

Churn is the dependent variable that takes the value 1 if a customer has churned and 0 otherwise.

Make sure that there are no two explanatory variables that are highly correlated. Use correlation analysis to determine the correlation between the variables.

1. Include a table of coefficients, t-values, and odds ratio. Interpret the logistic output explaining AIC/BIC, meaning of coefficients, significance, prediction accuracy (percent concordance), odds-ratios etc.
2. Which are the top three factors that affect churn in your model.
3. What other variables (that if collected) would help to improve the fit of the model.
4. Compute the hit ratio for your model. Hit ratio is defined as the percentage of correct predictions using the logit model. Use the model to predict 1 or 0 using the same data.

Hint:

When you have a large number of variables how do you short list a few important variables that could affect churn?

1. Compute means of all x variables for churners and non-churners separately and compute the percentage difference for each variable.
2. Sort the variables from high to low based on percentage difference in means. The top 10 variables based on this criteria are good candidates for inclusion in the logit model as long as they are not highly correlated.
3. Make sure that your data is not reduced by a large percentage (80-90%) because you have included explanatory variables in your model, that have many missing values I expect you will use most of the data..