Walsh Assignment 3 – DSCI 401

Brian Walsh

1. The response variable is Churn and the predictor variables are Gender, Age, Income, FamilySize, Education, Calls, and Visits.
2. I deleted the ‘CustID’ column, since that had no influence on churn. I also got dummy variables for the categorical columns, such as Gender and Income.
3. I decided on a Random Forest model, since it yielded the best results over my tests. My model development process consisted of building a variety of models (Logistic Regression, K Nearest Neighbors, Random Forest, Support Vector Machine, Decision Tree, and Gaussian Naïve Bayes). I iterated through different parameters for each model built, making sure I selected the best model. I also saved the F1 score for each model, and then used the average of three different cross validation methods (K-Fold, Leave One Out, and Shuffle Split) to determine which of my models performed the best overall.
4. I chose the F1 score since it considers both the precision and the recall of the model. My performance was:
   1. Logistic Regression F1 Score: .737
   2. K Nearest Neighbors F1 Score: .698
   3. Random Forest F1 Score: .884
   4. Support Vector Machine F1 Score: .714
   5. Decision Tree F1 Score: .714
   6. Gaussian Naïve Bayes F1 Score: .667
5. I used all the predictor variables since I believe that all of them correlate to the likelihood of churn.
6. The F1 score for the validation data varies due to the nature of the random forest model, but it seems to be about .65 to .7.
7. It is not necessarily a bad model, since it is better than just randomly predicting whether a customer will churn or not. It is true that the model is not as accurate as it could be, but using this model would still improve a business’ performance.