**Telcom Customer Churn Analysis using Machine Learning Techniques**

#### Group 7 - Purvi Thakor, Akshay Kamath

**\* Motivation:**

For all companies that bill customers on a regular basis, one of the main variables is churning. Having worked on some telecom projects in a professional capacity, this area wasn’t new for us.

However, applying ML techniques to predicting customer behavior isn’t something that either one of us has done before. We were keen to see if we could identify the key reasons behind customer churn & if possible, create a model that would help retain the same customers. We came across a telecom company’s customer churn [dataset](https://www.ibm.com/communities/analytics/watson-analytics-blog/guide-to-sample-datasets/) on the IBM Watson Analytics website.

…………………………………………………………………………………………………………………………..

**\* Methodology:**

The dataset consists of 21 variables in all. A few are continuous, rest are categorical. The control variable was customer churn with 2 levels Y/N (i.e. customer has left or not). Initially, we started out with some basic EDA & some visual plots that would help us understand the data better. Since many machine learning algorithms cannot operate on categorical variables, we had to convert this data into numerical variables. We decided to use the Label Encoder method to convert the variable type. After the initial analysis, we proceeded to use techniques like Feature Importance & Feature Selection to see if we had any variables that were redundant & could be discarded in the process of building the models. We decided to use models like Logistic Regression, Random Forest, Decision Trees, kNN, & Naïve-Bayes Classifier for this analysis. We split the data into train & test & built a model using each of these classifiers. We used K-fold Cross-validation to evaluate the quality of the predictive models by partitioning the original data into a training set to train the model, and a test set to evaluate it. We also plotted the ROC curves to check the performance of the binary classifiers.

…………………………………………………………………………………………………………………………..

**\* Observations**

**EDA:**

We started out with some basic EDA. We have around 7000 observations. We got rid of 11 incomplete observations from our dataset. We have 20 features & 1 target variable. Only 4 features out of 19 were numeric. All the rest were categorical variables. Our customer churn v/s stay data split is in the ratio of 1:3. Ratio of males to females is around 50:50. When we plotted the Churn variable against the customer’s tenure, one interesting observation was that most of the customers who leave the telco provider, usually do it within the first year. Beyond the first year, they tend to stick around. Another interesting bit is that their biggest customer base are their oldest customers followed by the newest. Obviously, we would not like to have multicollinearity in our dataset. So, we plotted a heatmap/correlation plot for all the variables. The variables with the highest positive correlation are TotalCharges, MonthlyCharges & Tenure. Looking at the data, we figured that TotalCharges is nothing but Tenure times MonthlyCharges. Hence, we discarded the TotalCharges variable. Now, customerID is another such redundant feature, so we got rid of that variable, too.

**Data Preprocessing, Feature Importance & Recursive Feature Elimination:**

As the first step in our data preprocessing, we split the variables into Feature variables & Target variables. Then, after this we split our data into training, & testing. To bring the variables on the same scale, we standardized the data. MonthlyCharges & Tenure are the features with the highest Feature Importance (~22%). No other feature exceeds more than 10% in Feature Importance. In order to check if we can get rid of features that are not contributing to the model much, we tried to use the Recursive Feature Elimination method. The output of the RFE is surprising since it mentions that the optimal number of features are **18**. We decided to include all the features in our analysis & move ahead with our Model building & comparison.

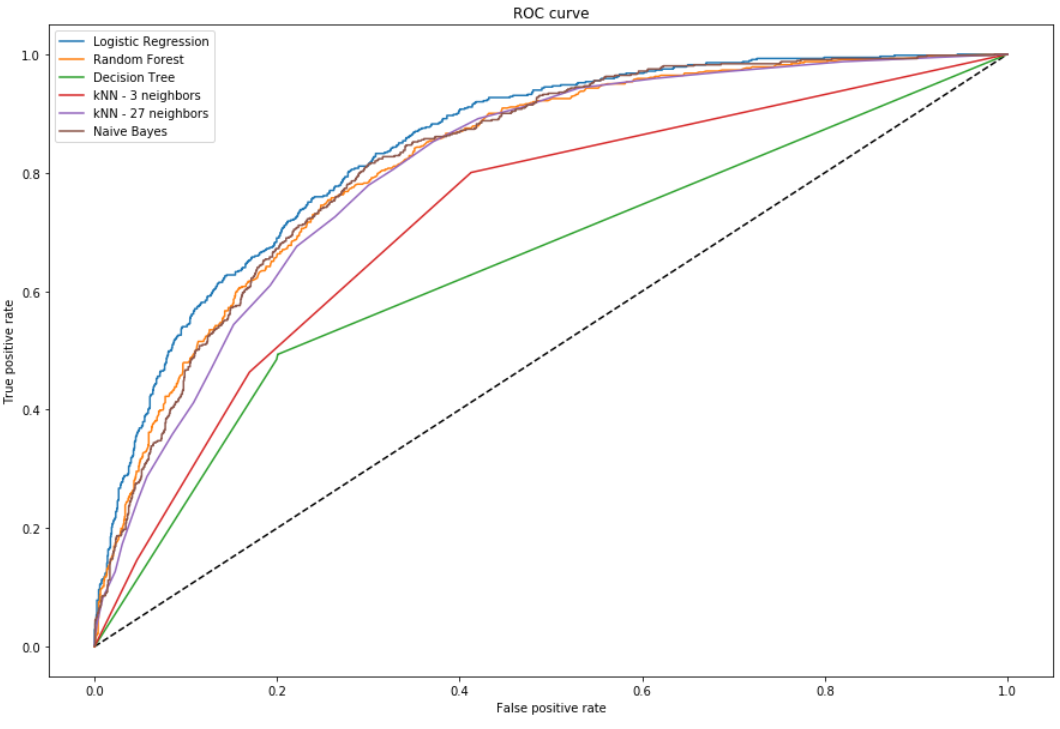
**Model Building & Comparison:**

We've used quite a few models to check which fits best on our data. Models used are –

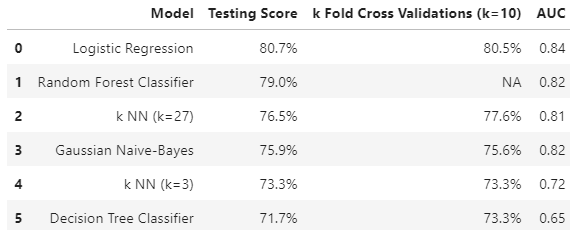
* Logistic Regression
* Random Forest Classifier
* Decision Tree Classifier
* k-NN Classifier
* Gaussian Naive-Bayes Classifier

Our target variable is a binary variable. The customer either stays or leaves. Hence, we decided to use the Logistic Regression as our first model to check how the model fits the data.

To check the model’s predictive performance, we’ve used the K-fold Cross Validation across almost each and every classifier barring Random Forest (since that is essentially what a Random Forest does). For the kNN classifier, we initially used k=3 to predict our target label. However, since we did not know what the optimal value should be for k we tried to fit the model against all odd k’s between 1 & 50. The model with k = 27 had the highest accuracy i.e. the lowest error. Hence, using k = 27, we built another kNN model. We also plotted the ROC curve for all the classifiers as a diagnostic test to evaluate the performance of the models. Another fancy way to check at model accuracy is to look at the model precision, recall & fscore or basically, a Confusion Matrix.



As we can see, the model with the best ROC curve is the Logistic Regression. Naïve Bayes & Random Forest are competing for the second place. Comparing the 2 kNN models, we can see that the one with k as 27 is a lot better than the one with 3 neighbors. The Decision Model looks to be our worst model. Our observations from the AUC concur with the test data scores across all the models, as shown below.



**\* Closing Remarks**

\* It would be interesting to have more features, some continuous ones preferably, in our model & do this exercise all over again. This would probably increase the efficiency of our models.

\* One of the more surprising outputs was that Recursive Feature Elimination (RFE) didn’t help eliminate any redundant features.

\* We came across a couple of articles that mentioned that Decision Trees handle categorical variables better than continuous variables. That did not seem to be the case in our models.

\* Based on the model scores, to predict customer churn Logistic Regression seems to be the best model for this dataset.

…………………………………………………………………………………………………………………………..