**Text Classification Problem**

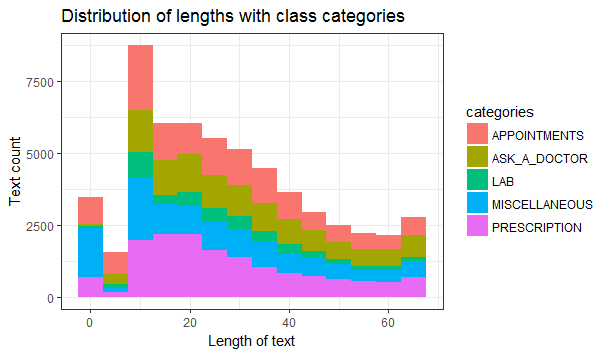
**Problem statement and Domain Knowledge:**

XYZ Health Services is a top ranked Health care provider in USA with stellar credentials and provides high quality-care with focus on end-to-end Health care services. The Heath Care Services range from basic medical diagnostics to critical emergency services. The provider follows a ticketing system for all the telephonic calls received across all the departments. Calls to the provider can be for New Appointment, Cancellation, Lab Queries, Medical Refills, Insurance Related, General Doctor Advise etc. The Tickets have the details of Summary of the call and description of the calls written by various staff members with no standard text guidelines.

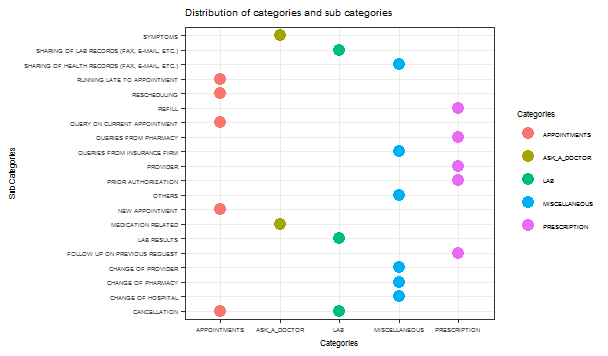
**Process**

**To classify the text into its categories and sub-categories, the module is divided into following process:**

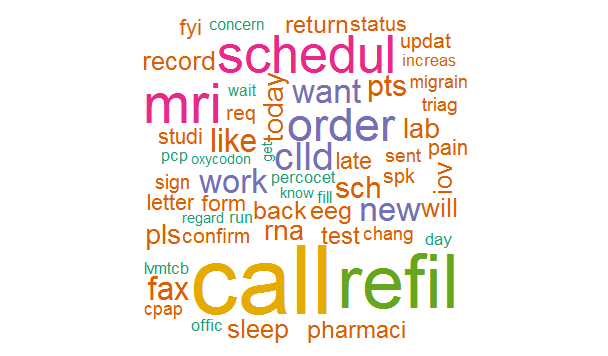
1. **Build and Save model**
   1. **Initialization**
      1. Clearing the directory
      2. Setting the working directory
      3. Loading library
      4. Read the csv file
   2. **Data cleaning**
      1. Unwanted columns field\_id and data is removed. Data with columns- "summary","categories","sub\_categories","previous\_appointment","id" is stored.
      2. Junk is an unwanted category and sub-category, not mentioned in the requirement document. So, its removed
      3. The categories and sub-categories are changed to lower case, for removing duplication.
      4. Data is analyzed to find the missing values. If any missing value is found, corresponding rows are removed. In this data set no missing values.
   3. **Data understanding**
      1. The proportion of categories and sub-categories in the dataset is found. After the data is split into train and test, data is checked for same proportion
      2. Understanding frequent words- To find the frequent words, length of the text is calculated.
      3. Histogram graph is plotted against the number of words in each category.

****

* + 1. Also, to understand the category – sub category distribution, multivariate scatter plot is plotted against category and sub-category



* **Term document Matrix**
  + **Text** is converted into corpus
  + **Corpus** is preprocessed by removing numbers, punctuations, whitespace, stop words, transforming to lower case and stemming the document
  + **Feature Engineering:** Techniques like bag of words, TF-IDF weights and N-gram tokenizer is assigned to calculate the document term matrix.
  + **Feature Extraction :** Dimensionality of the matrix is large. So, sparse terms(terms with maximum zero value) are removed and highly related columns are stored. **PCA**, is used. But due to low memory couldn’t process further
  + **Word cloud** is built to understand the distribution of frequently occurring terms.

****

* **Model building and saving**
  + Split the data into train and test
  + **Multinomial Logistic Regression** is built using multinom function.
  + **Multiclass naïve bayes** is built using naiveBayes function.
  + Model is saved using saveRDS function
* **Model validation** 
  + Model is loaded using readRDS function
  + Using predict command model is tested for the test data and confusion matrix is built
  + **Accuracy** is calculated by summing the diagonal values (true values) of matrix divided by total number of observations.
  + From the graphs plotted against actual value and predicted value , we can find that Multinomial Logistic regression is performing better compared to Naïve Bayes model.
  + **Result graph** is as follows –

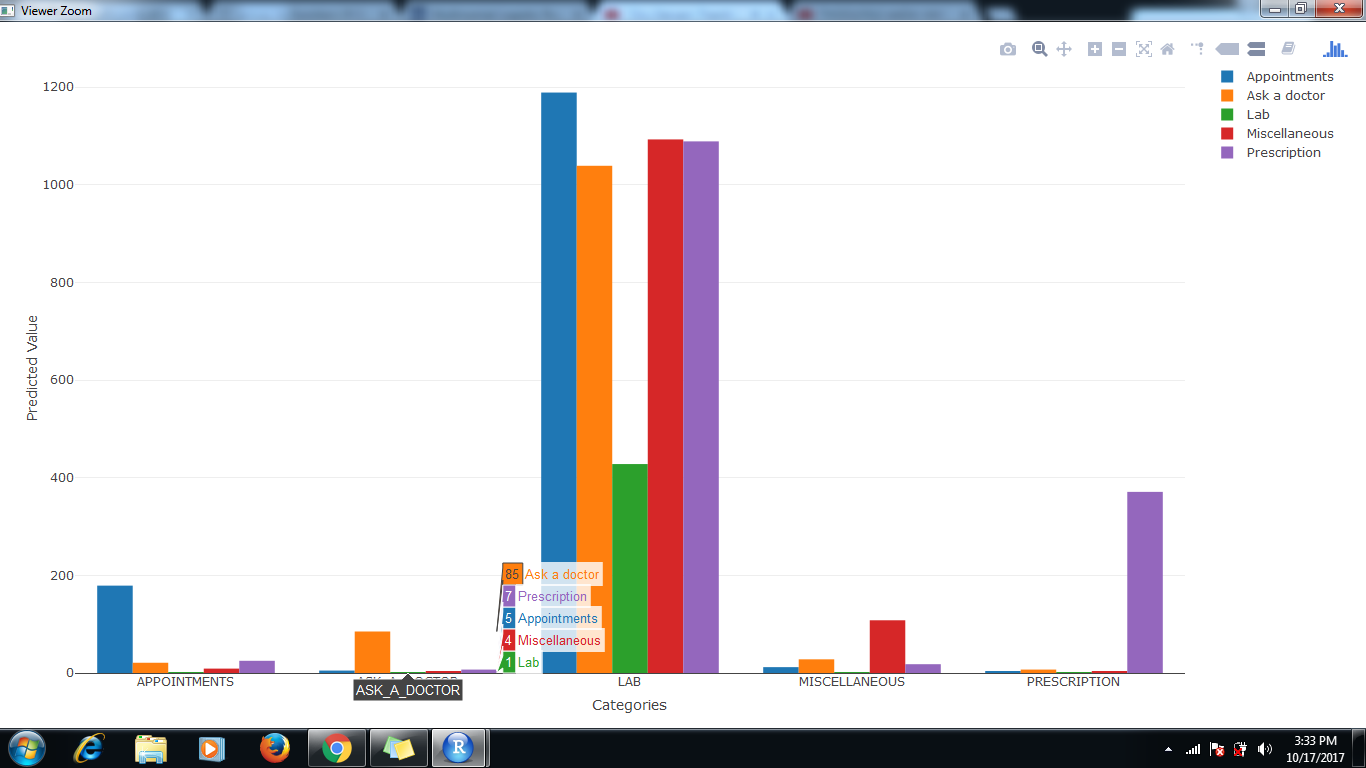
****

Figure 1Naive Bayes op



Figure 2Multinomial Logistic

**Recommendations and Insights**

* Actual and predicted value of category and sub-catgeory is stored as a data frame
* Multinomial Logistic Regression gives more accuracy compared to Naïve Bayes model
* From the result, we can **recommend following observations**
  + XYZ health service is working better for Prescription related activities. And Lab facilities need to be improved.

**About the model**

**Multinomial Logistic** **Regression**

Multinomial Logistic Regression (MLR) is a form of linear regression analysis conducted when the dependent variable is nominal with more than two levels. It is used to describe data and to explain the relationship between one dependent nominal variable and one or more continuous-level (interval or ratio scale) independent variables.

**Working**

The multinomial logistic regression estimates a separate binary logistic regression model for each dummy variables.  The result is M-1 binary logistic regression models.  Each model conveys the effect of predictors on the probability of success in that category, in comparison to the reference category.

Each model has its own intercept and regression coefficients—the predictors can affect each category differently.

**Limitation**

The category to which an outcome belongs to, does not assume any order in it. For example, if we have N categories, all have an equal probability. In reality, we come across problems where categories have a natural order.

**Naive Bayes Classifier:**

It is a classification technique based on Bayes’ theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Due to this, the result can be (potentially) very bad - hence, a “naive” classifier.

Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c).

***Pros:***

* It is easy and fast to predict class of test data set. It also perform well in multi class prediction
* When assumption of independence holds, a Naive Bayes classifier performs better compare to other models like logistic regression and you need less training data.
* It perform well in case of categorical input variables compared to numerical variable(s). For numerical variable, normal distribution is assumed (bell curve, which is a strong assumption).

***Cons:***

* If categorical variable has a category (in test data set), which was not observed in training data set, then model will assign a 0 (zero) probability and will be unable to make a prediction. This is often known as “Zero Frequency”. To solve this, we can use the smoothing technique. One of the simplest smoothing techniques is called Laplace estimation.
* On the other side naive Bayes is also known as a bad estimator, so the probability outputs from predict probability are not to be taken too seriously.
* Another limitation of Naive Bayes is the assumption of independent predictors. In real life, it is almost impossible that we get a set of predictors which are completely independent.