Project Report: Text Classification Problem

1. Problem Statement:-

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.

The challenge is, based on the Text in the Summary and Description of the call, the ticket is to be classified to Appropriate Category (out of 5 Categories) and Subcategories (Out of 20 Sub Categories).

2. Tool's and Library's used:-

- R 3.4.1
- Rstudio-1.0.153
- caTools
- caret
- e1071
- tm
- wordcloud
- ggplot2
- scales
- xgboost

3. Exploratory Data analysis:-

The Data Input File was loaded into the environment with read.csv() function

1. Structure of all file was analyzed
2. Top rows of data were analyzed to get idea about variables

The Data contained

- 53913 observations
- 7 variables

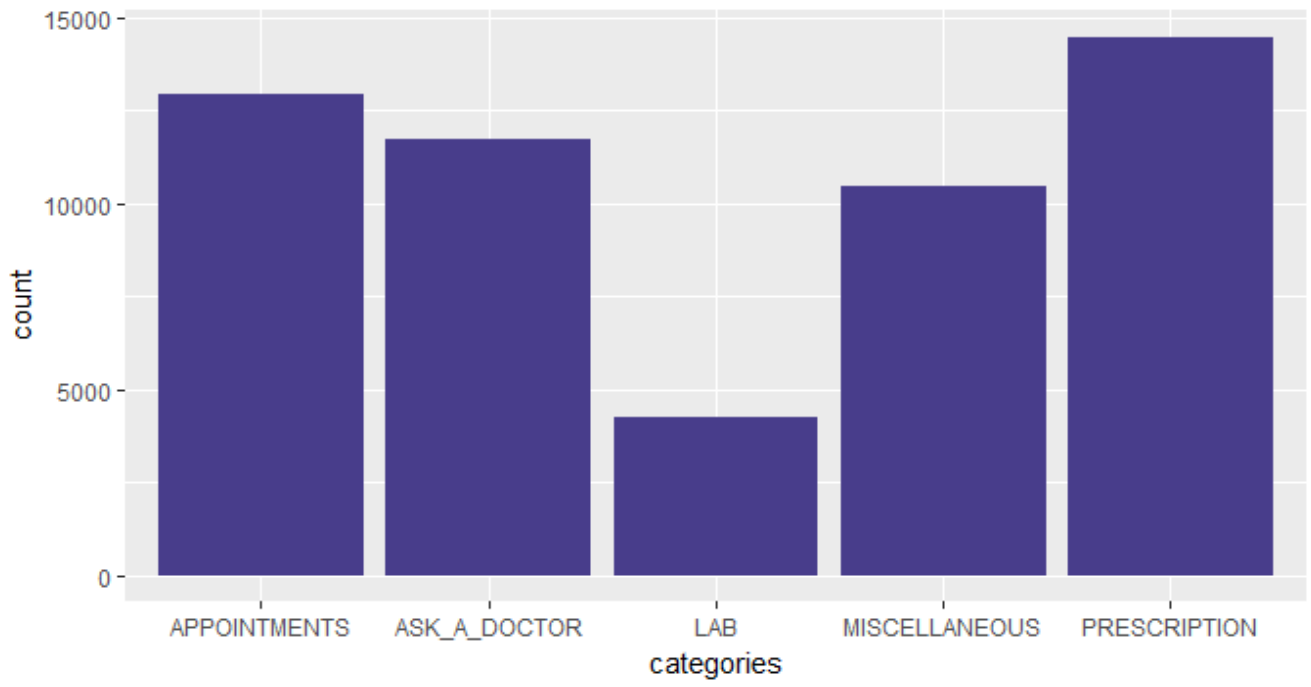
	Data type
◦ File ID – All observations contained same value of file id Numeric	
◦ SUMMARY- summary of call recording	Factor
◦ DATA- Original call recording data including Directory Paths	Factor
◦ Categories- First target variable containing categories of ticket classification	Factor
◦ Subcategories- Second target variable containing subcategories of ticket classification	Factor
◦ Previous_appointment – Contains Yes and No for patient	Factor

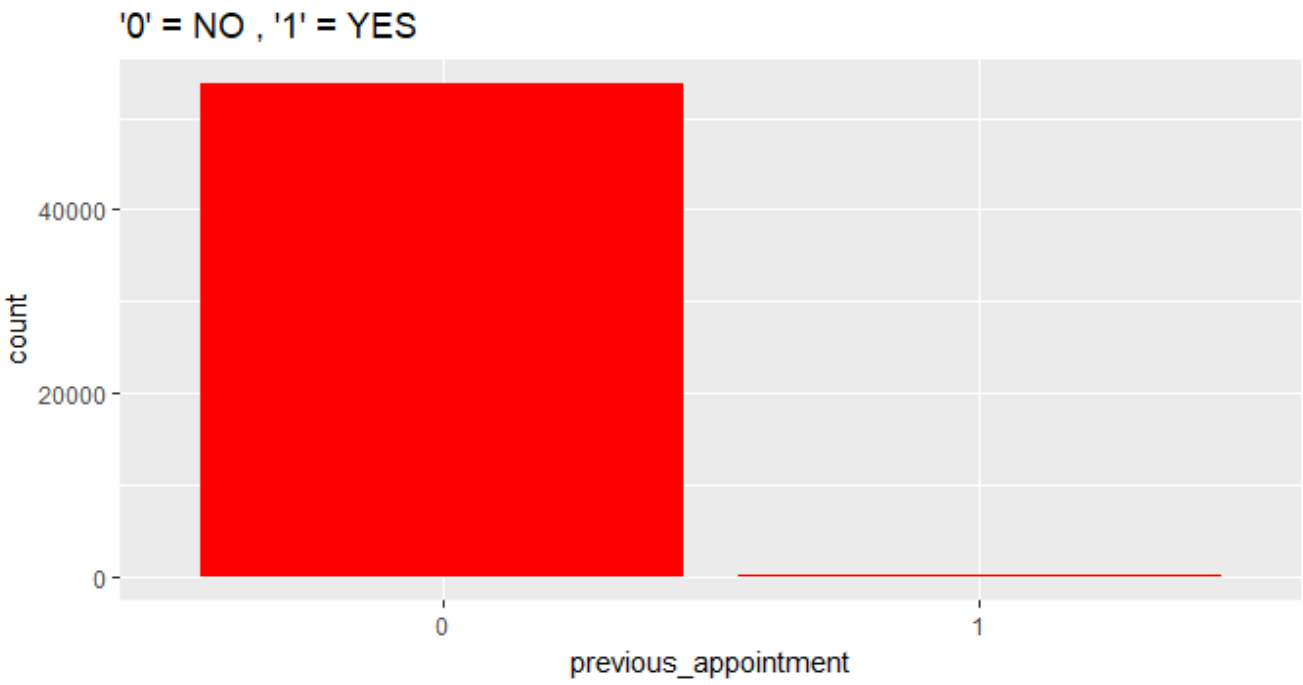
previous appointment
- ID- Date of the call received Factor

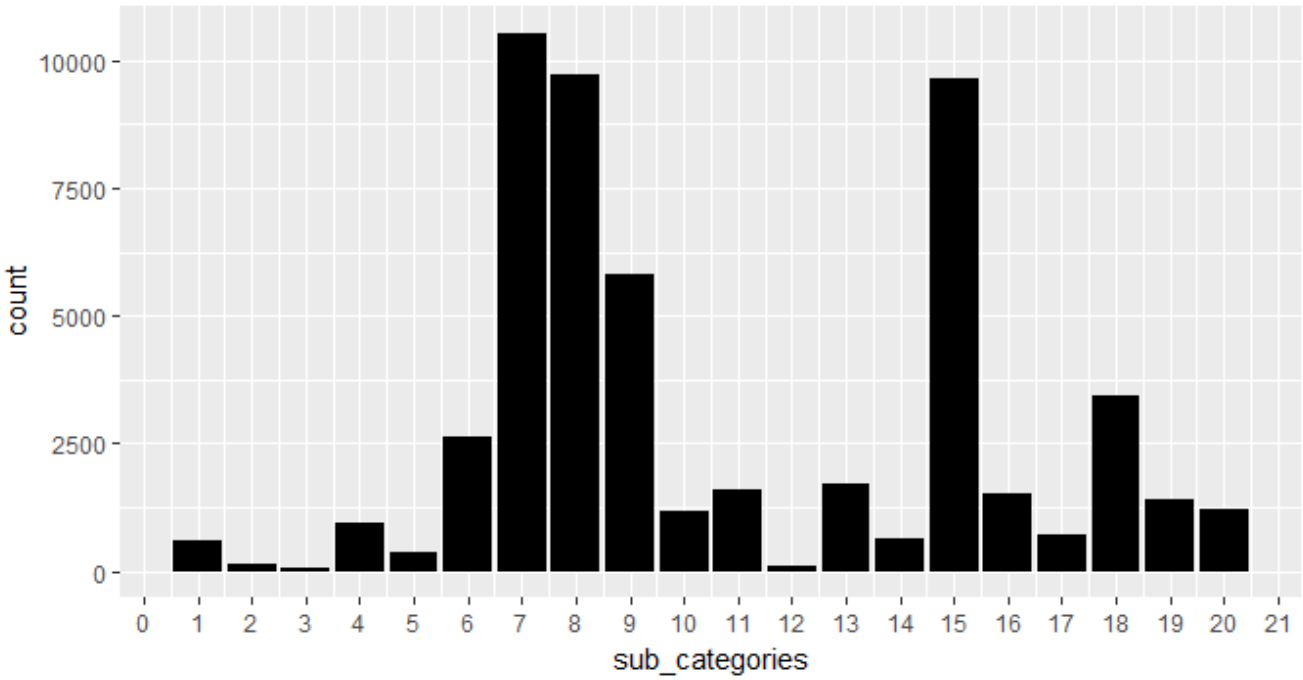
4. Feature Engineering

- First the two variables ID and FileID were dropped as they didn't contained any Importance to the target variable
- The Categories variable ,subcategories variable and Previous_appointment variable were capitalized and converted into factor form
- The “JUNK” category was found out to have least frequency in both categories and subcategories.
- Then the observations Containing “JUNK” were removed
- Levels of categories and subcategories variable were reduced so to get 5 levels in categories and 20 level in subcategories variable.
- Then in the previous_appointment variable the levels “YES” and “NO” were converted into Binary 1, 0 respectfully.
- The Summary and Data variables were converted into character data types
- The Summary variable contained missing values so those observations were dropped.

5. Plots




'0' = NO , '1' = YES
40000
20000
0
count
0
1
previous_appointment


10000
7500
5000
2500
0
count
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21
sub_categories

6.Text Mining:-

- The text corpus for Data and Summary were made
- Pre-processing steps performed:
  - 1. Removing Punctuation Marks
  - 2. Removing Numbers
  - 3. Case folding
  - 4. Removing Stop words
  - 5. Removing White spaces
  - 6. Stemming
- Wordcloud was made over corpus to observe the importance of words

- Document term matrix of the text corpus was made on TF-IDF as weighting criteria
- The generated Document term matrix of both Data and Summary variable were cbinded with the Categories, subcategories and previous_appointment variable from the original data set.

7. Model building and validation:-

- Two seprate model were build and tested
- As there were two different target variables so the model for each target variable was built seperately
- Error metrics used Confusion Matrix

- **Naive Bayes model for classification of Categories variable**
  Overall Statistics

  Accuracy : 0.6693
  95% CI : (0.662, 0.6765)
  No Information Rate : 0.256
  P-Value [Acc > NIR] : < 2.2e-16

  Kappa : 0.5839
  Mcnemar's Test P-Value : < 2.2e-16
  Time difference of 3.883619 secs

- **Naive Bayes model for classification of Sub-Categories variable**
  Overall Statistics

  Accuracy : 0.0963
  95% CI : (0.0918, 0.101)
  No Information Rate : 0.4661
  P-Value [Acc > NIR] : 1

  Kappa : 0.0797
  Mcnemar's Test P-Value : NA
  Time difference of 4.984946 secs

- **Xgboost model for classification of Categories variable**
  Overall Statistics

  Accuracy : 0.9963
  95% CI : (0.9952, 0.9972)
  No Information Rate : 0.2699
  P-Value [Acc > NIR] : < 2.2e-16

  Kappa : 0.9952
  Mcnemar's Test P-Value : NA
  Time difference of 1.709981 mins

- **Xgboost model for classification of Sub_Categories variable**
  Overall Statistics

  Accuracy : 0.887
  95% CI : (0.882, 0.8918)
  No Information Rate : 0.2105
  P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.8697
Mcnemar's Test P-Value : NA
Time difference of 31.94608 mins