PROJECT REPORT

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.

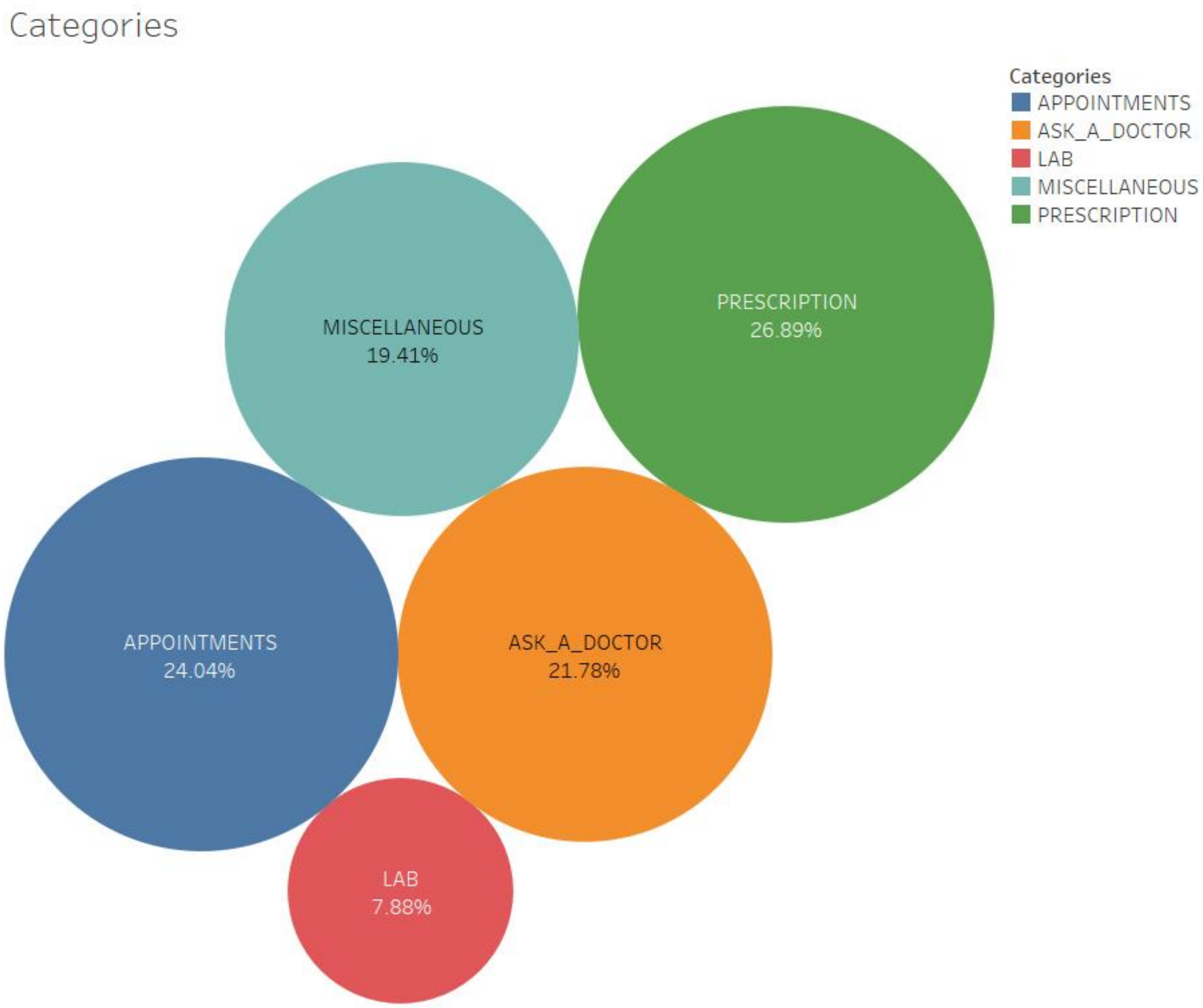
The problem is, based on the Text in the Summary and Description of the call, the ticket is to be classified to Appropriate Category (out of 5 Categories) and Subcategories (Out of 20 Sub Categories).

HYPOTHESEISI GENERATION

Generally phone calls are done by persons who are having ailments so the phonecall data contains appointment booking ,medication related , and followup of the visit of the data.one can find a relation between runtime in overall experience of the person starting from booking appointment to cure of his ailment.This wholecycle involves diagnosis of ailment, lab reports if necessary, mediction related, followup once dosage of medicine taken by the patient and results.this end to end cycle we can analyse from the patient id summary and description of call and according we can classify each and every patient experience and there by identifying pain points of the customers and providing better solutions for the patients there by enhanced services of the company in aholistic manner making it more competitive enough than before.For that to happen we need better model the text data of the information gathered from the description and summary of the phone data in better emanner leading to best classify in to best accuracy of classifying into respective categories and subcategories.

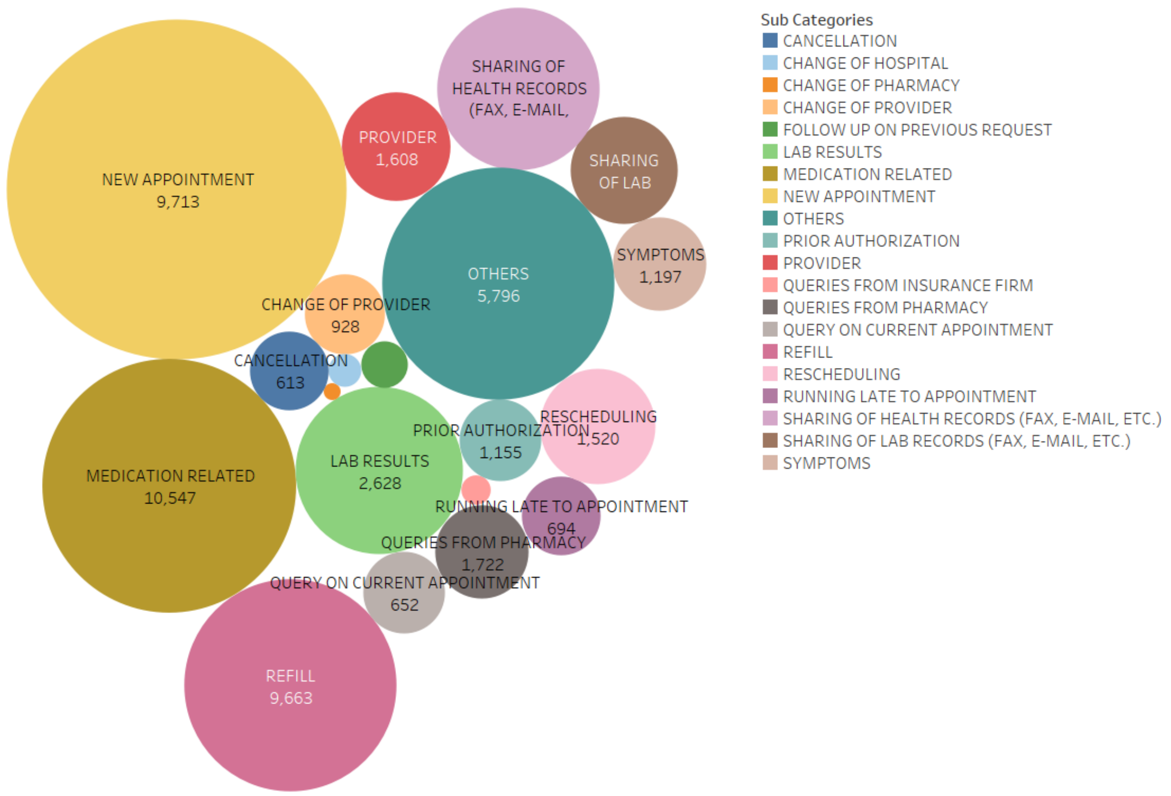
Exploratory Data Analysis

All are categorical variables mostly. since there is missing data of variable of data which accounts for 5% of data, so excluding that overall that data around 57,910 observations



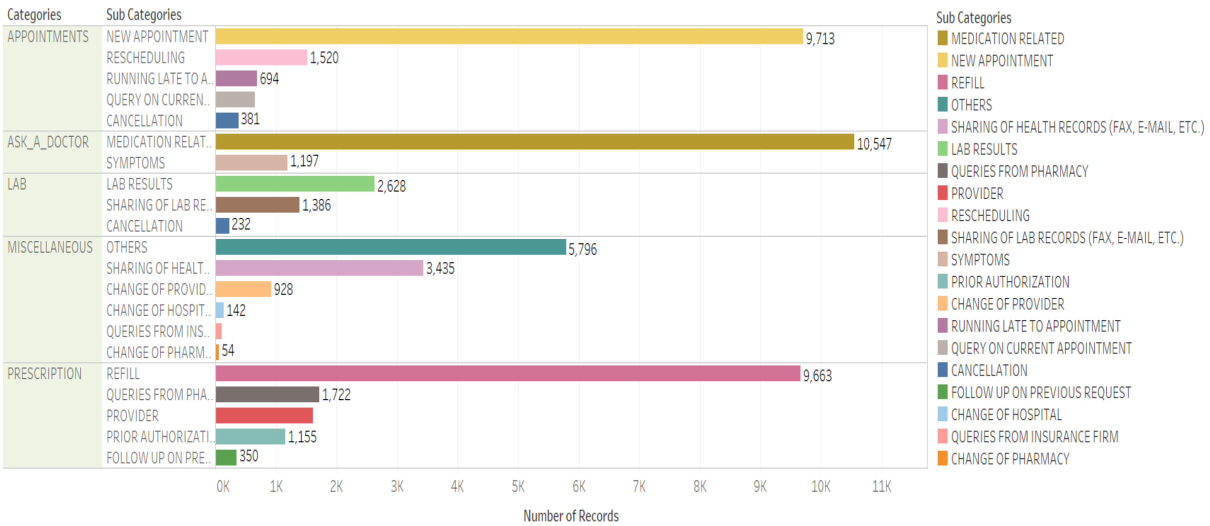
Categories and % of Total Number of Records. Color shows details about Categories. Size shows % of Total Number of Records. The marks are labeled by Categories and % of Total Number of Records. Details are shown for Categories. Percents are based on each cell of each pane of the table.

SubCategories



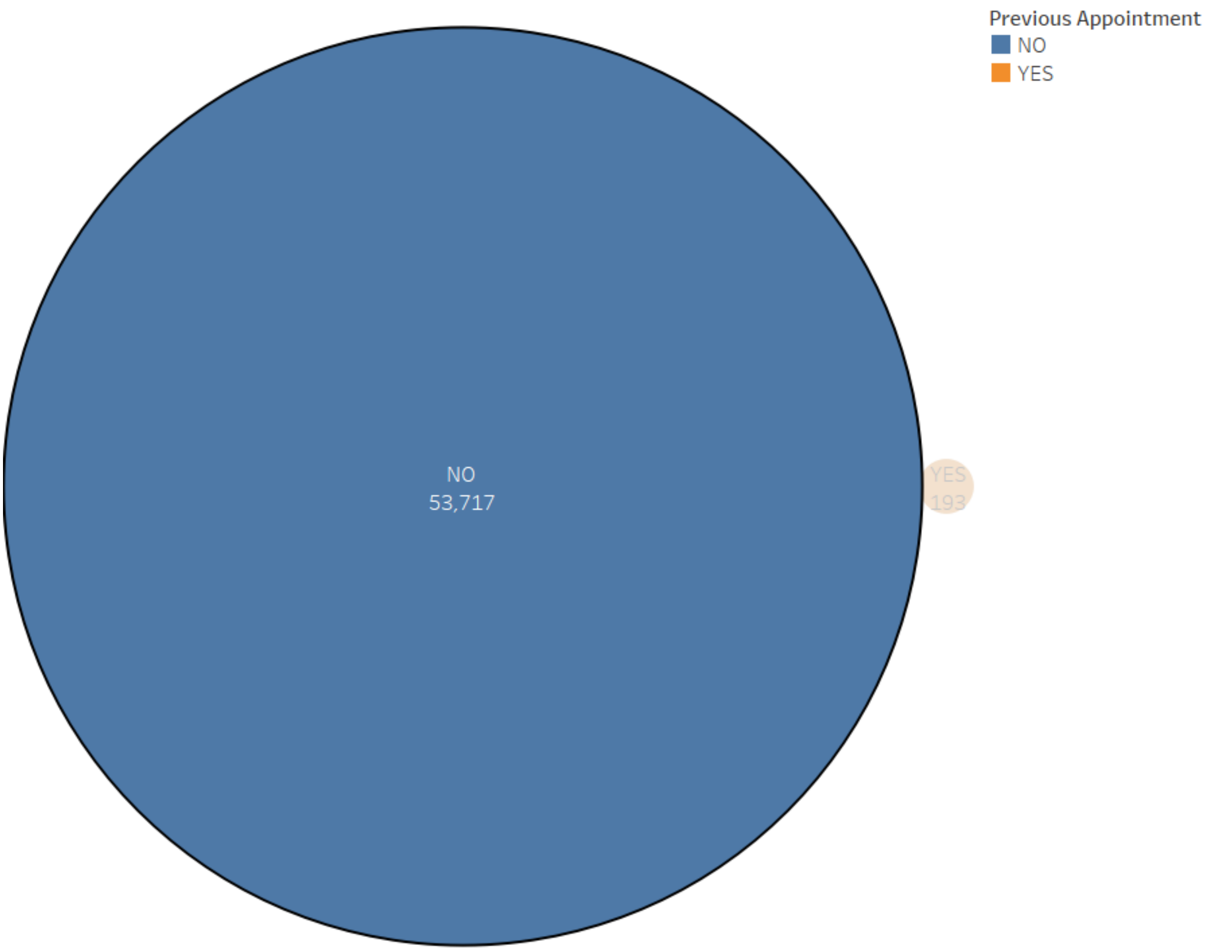
Sub Categories and sum of Number of Records. Color shows details about Sub Categories. Size shows sum of Fileid. The marks are labeled by Sub Categories and sum of Number of Records.

Categories & SubCategories correlation



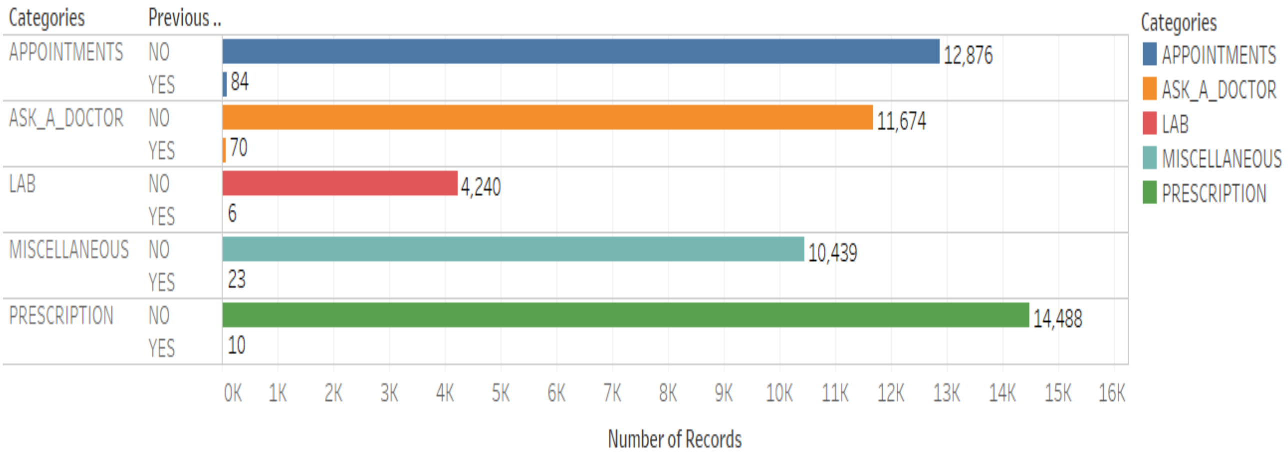
Sum of Number of Records for each Sub Categories broken down by Categories. Color shows details about Sub Categories. The marks are labeled by sum of Number of Records.

Previous Appointments



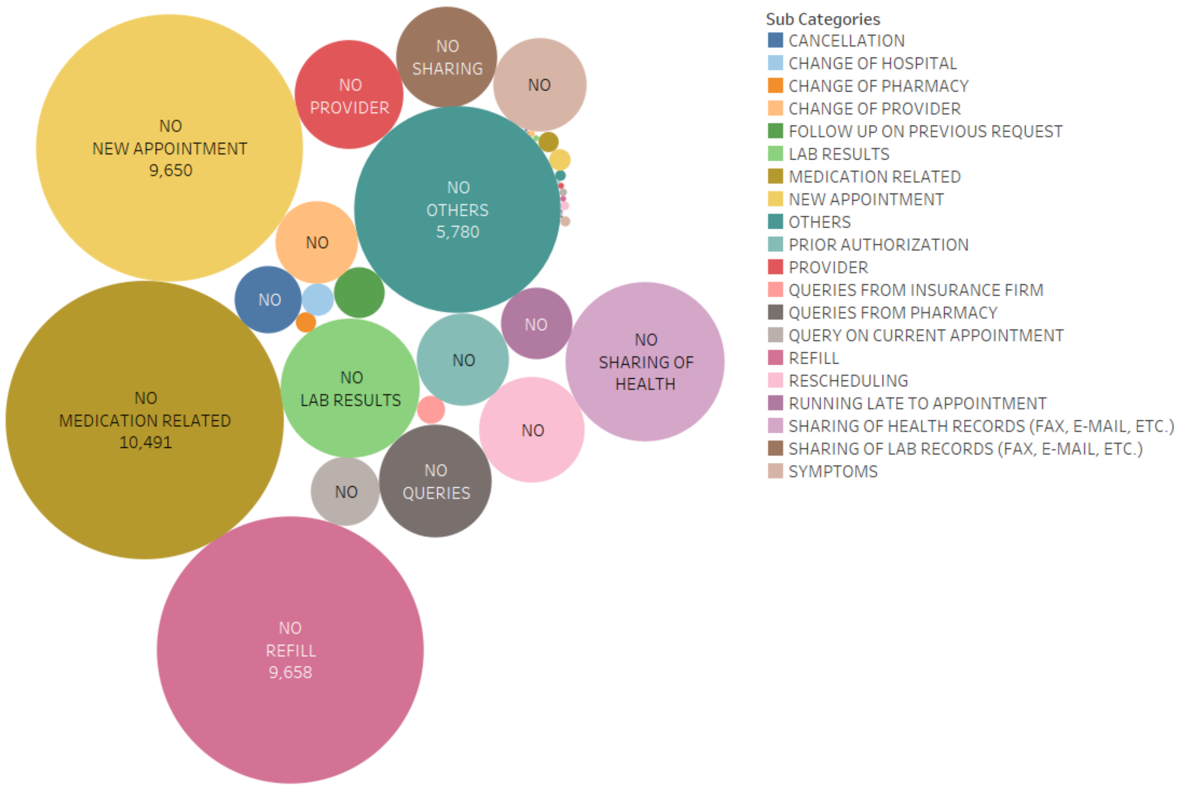
Previous Appointment and sum of Number of Records. Color shows details about Previous Appointment. Size shows sum of Number of Records. The marks are labeled by Previous Appointment and sum of Number of Records.

Categories Vs Previous Appointments



Sum of Number of Records for each Previous Appointment broken down by Categories. Color shows details about Categories. The marks are labeled by sum of Number of Records.

SubCategories Vs PreviousAppointment



Previous Appointment, Sub Categories and sum of Number of Records. Color shows details about Sub Categories. Size shows sum of Number of Records. The marks are labeled by Previous Appointment, Sub Categories and sum of Number of Records.


method
question
usedtab
start
patientparf
work
check
print
xxxxxxxxf
test
return
receiv
end
patientparfsbparfsb
subject
atparb
pharmac
cell
addit
now
mari
note
just
pain
use
tablet
talk
tri
still
sinc
chang
call
caller
phone
bid
sleep
backfsbparb
send
rxidtab
done
due
need
dose
offic
let
medfus
momparparfsb
fill
list
verbal
appt
patientparb
say
per
med
action
queu
parfsbparb


patientparb
fax
per
left
mari
dose
contact
pmparfsbpar
say
rna
action
convert
thank
method
sign
use
week
mri
adviseparb
advi
clinic
hcl
patientpar
phone
come
increa
place
ask
name
refillparb
phonexxxxxxxxpar
due
changesparfspar
now
bxxxxxxxxb
still
see
medic
fill
end
med
tablet
detail
tall
parfsbparb
wrotepar
usedtab
form
send
flag
work
just
test
enter
author
tri
check
updatedparb
bxxxxxxxx
lpn
signedpar
prescript
follow
note


prescript
med
adviseparb
sinc
par
pick
start
notifi
offic
bid
talk
xxx
bytab
still
parfsbparb
due
test
parb
print
pmparb
signedpar
provid
let
take
question
dose
note
left
clinic
xxxxxxxx
verbal
now
check
reason
complet
use
author
send
just
per
tri
increa
ask
receiv
hcl
fsbpar
chang
end
method
queu
place
followup
return
prn
bxxxxxxxb
done
authorizedxxxxxxx

Findings & Recommendations

Findings & Recommendations

In Categories section prescription tops of all followed by appointments and ask a doctor categories,which indicates better performance of the company which are major revenue generators.so better investing more in qualitative customer support further boost up the company revenue a lot

In sub category section ,customers tops at new medication related ,new appointment and refill these are the top sub categories which for major revenue generation of the company but follow upon requests which is one of the least sub category need to focus on that area which improves better care for patients and improving recurring revenues for the company Also keep tracking of health records, lab records of each and every patients helps in devoloping better predictive measures in future for those lready diagnosed patients.

Need to decrease the appointment cancellation requests finding reasons for that and according decreas the cancellations of appointments tends to zero improve the revenue share of the company.

The previous appointment category "NO" is highest indicates either patients are feeling their experience on a whole is not upto the mark churning out to other hospitals are high, it is complex and need to be monitored in better way

As a whole new appointments, medication and doctor suggestions tops of all which is keeping the company at top position but if labrecords, medication reports are taken care well and data related to that of each and every patient record when kept maintained and analysed properly will give more opportunities, open new doors and strategic advantage to this healthcare company against other since it helps in predictive measures for the patients and new patients can be trated well providing a better holistic experience which is utmost need of the hour in present health care market,

Libraries used

Matrix'

Tidyr

Tm

Dplyr

Tibble

Mlr

RandomforestSRC

Rferns

MLMetr*ics*

Feature Engineering

In this multilabel classification we are doing one hot encoding to Category section and Sub Category variables into binary coding into 1's and 0's using sparse model matrix and then bringing back to data frame structure. And for text mining since we need to model two types of textual variables we are combining both these textual vectors and then applying tf-idf technique to bring best variables out of them.<u>Labelled the target variables from v1 t0 v25 and these are binary logical in nature.</u> Encoded as TRUE & FALSE taking "0 "as FALSE and "1" as TRUE.

```
 [v1] "categoriesAPPOINTMENTS"
 [v2] "categoriesASK_A_DOCTOR"
 [v3] "categoriesLAB"
 [v4] "categoriesMISCELLANEOUS"
 [v5] "categoriesPRESCRIPTION"
 [v6] "sub_categoriesCANCELLATION"
 [v7] "sub_categoriesCHANGE OF HOSPITAL"
 [v8] "sub_categoriesCHANGE OF PHARMACY"
 [v9] "sub_categoriesCHANGE OF PROVIDER"
[v10] "sub_categoriesFOLLOW UP ON PREVIOUS REQUEST"
[v11] "sub_categoriesLAB RESULTS"
[v12] "sub_categoriesMEDICATION RELATED"
[v13] "sub_categoriesNEW APPOINTMENT"
[v14] "sub_categoriesOTHERS"
[v15] "sub_categoriesPRIOR AUTHORIZATION"
[v16] "sub_categoriesPROVIDER"
[v17] "sub_categoriesQUERIES FROM INSURANCE FIRM"
[v18] "sub_categoriesQUERIES FROM PHARMACY"
[v19] "sub_categoriesQUERY ON CURRENT APPOINTMENT"
[v20] "sub_categoriesREFILL"
[v21] "sub_categoriesRESCHEDULING"
[v22] "sub_categoriesRUNNING LATE TO APPOINTMENT"
[v23] "sub_categoriesSHARING OF HEALTH RECORDS (FAX, E-MAIL, ETC.)"
[v24] "sub_categoriesSHARING OF LAB RECORDS (FAX, E-MAIL, ETC.)"
[v25] "sub_categoriesSYMPTOMS"
```

Model Building

Since in this problem we need to predict both category section and sub category section at a time, it is multilabel classification problem in text mining Here we need to predict 25 target variables variables at a time applying MLR library and implement two machine learning techniques for better accuracy and distinction.

They are1) RandomForestSurvivalRegressionClassification known as RFSRC

2)Random Ferns knowns as Rferns

We got only these two machine learning algorithms to predict these variables.

Model evaluation

Both the models are giving high accuracies above 90% .

<u>RESULTS</u>

Results of each and every variable since every variable comes under binary classification computed both AUC and Accuracy metrics for each and every variable with their respective accuracies and AUC's are compared for each and every target variable and compared both model efficiencies

```
> Accuracy(y_pred=pred_rfsrc_N$data$response.v1,y_true=test_data$v1)

[1] 0.9269152

> #Accuracy of  rfsrccmodel

> Accuracy(y_pred=pred_rfsrc_N$data$response.v2,y_true=test_data$v2)

[1] 0.8791504

> #Accuracy of  rfsrccmodel

> Accuracy(y_pred=pred_rfsrc_N$data$response.v3,y_true=test_data$v3)

[1] 0.9633649

> #Accuracy of  rfsrccmodel

> Accuracy(y_pred=pred_rfsrc_N$data$response.v4,y_true=test_data$v4)

[1] 0.8642181

> #Accuracy of  rfsrccmodel

> Accuracy(y_pred=pred_rfsrc_N$data$response.v5,y_true=test_data$v5)

[1] 0.9027082

> #Accuracy of  rfsrccmodel

> Accuracy(y_pred=pred_rfsrc_N$data$response.v6,y_true=test_data$v6)
```

```
[1] 0.9893341
> #Accuracy of  rfsrccmodel
> Accuracy(y_pred=pred_rfsrc_N$data$response.v7,y_true=test_data$v7)
[1] 0.9967538
> #Accuracy of  rfsrccmodel
> Accuracy(y_pred=pred_rfsrc_N$data$response.v9,y_true=test_data$v9)
[1] 0.9819143
> #Accuracy of  rfsrccmodel
> Accuracy(y_pred=pred_rfsrc_N$data$response.v10,y_true=test_data$v10)
[1] 0.9936004
> #Accuracy of  rfsrccmodel
> Accuracy(y_pred=pred_rfsrc_N$data$response.v11,y_true=test_data$v11)
[1] 0.9784827
> #Accuracy of  rfsrccmodel
> Accuracy(y_pred=pred_rfsrc_N$data$response.v12,y_true=test_data$v12)
[1] 0.8737711
> #Accuracy of  rfsrccmodel
> Accuracy(y_pred=pred_rfsrc_N$data$response.v13,y_true=test_data$v13)
[1] 0.9194027
> #Accuracy of  rfsrccmodel
> Accuracy(y_pred=pred_rfsrc_N$data$response.v14,y_true=test_data$v14)
[1] 0.9182897
> #Accuracy of  rfsrccmodel
> Accuracy(y_pred=pred_rfsrc_N$data$response.v15,y_true=test_data$v15)
[1] 0.9824708
> #Accuracy of  rfsrccmodel
> Accuracy(y_pred=pred_rfsrc_N$data$response.v16,y_true=test_data$v16)
[1] 0.9728251
> #Accuracy of  rfsrccmodel
```

```
> Accuracy(y_pred=pred_rfsrc_N$data$response.v17,y_true=test_data$v17)
[1] 0.9986088
> #Accuracy of  rfsrccmodel
> Accuracy(y_pred=pred_rfsrc_N$data$response.v18,y_true=test_data$v18)
[1] 0.9841402
> #Accuracy of  rfsrccmodel
> Accuracy(y_pred=pred_rfsrc_N$data$response.v19,y_true=test_data$v19)
[1] 0.9883139
> #Accuracy of  rfsrccmodel
> Accuracy(y_pred=pred_rfsrc_N$data$response.v20,y_true=test_data$v20)
[1] 0.9513077
> #Accuracy of  rfsrccmodel
> Accuracy(y_pred=pred_rfsrc_N$data$response.v21,y_true=test_data$v21)
[1] 0.9741235
> #Accuracy of  rfsrccmodel
> Accuracy(y_pred=pred_rfsrc_N$data$response.v22,y_true=test_data$v22)
[1] 0.9952699
> #Accuracy of  rfsrccmodel
> Accuracy(y_pred=pred_rfsrc_N$data$response.v23,y_true=test_data$v23)
[1] 0.9437952
> #Accuracy of  rfsrccmodel
> Accuracy(y_pred=pred_rfsrc_N$data$response.v24,y_true=test_data$v24)
[1] 0.976164
> #Accuracy of  rfsrccmodel
> Accuracy(y_pred=pred_rfsrc_N$data$response.v25,y_true=test_data$v25)
[1] 0.975422
```

>Now observing AUC of each and every target variable in randomforestSRC Model

```
> AUC(y_pred=pred_rfsrc_N$data$response.v1,y_true=test_data$v1)
[1] 0.884844
> #Accuracy of  rfsrccmodel
> AUC(y_pred=pred_rfsrc_N$data$response.v2,y_true=test_data$v2)
[1] 0.7710847
> #Accuracy of  rfsrccmodel
> AUC(y_pred=pred_rfsrc_N$data$response.v3,y_true=test_data$v3)
[1] 0.8049867
> #Accuracy of  rfsrccmodel
> AUC(y_pred=pred_rfsrc_N$data$response.v4,y_true=test_data$v4)
[1] 0.7062303
> #Accuracy of  rfsrccmodel
> AUC(y_pred=pred_rfsrc_N$data$response.v5,y_true=test_data$v5)
[1] 0.8477085
> #Accuracy of  rfsrccmodel
> AUC(y_pred=pred_rfsrc_N$data$response.v6,y_true=test_data$v6)
[1] 0.5128373
> #Accuracy of  rfsrccmodel
> AUC(y_pred=pred_rfsrc_N$data$response.v7,y_true=test_data$v7)
[1] 0.5
> #Accuracy of  rfsrccmodel
> AUC(y_pred=pred_rfsrc_N$data$response.v8,y_true=test_data$v8)
[1] 0.5
> #Accuracy of  rfsrccmodel
> AUC(y_pred=pred_rfsrc_N$data$response.v9,y_true=test_data$v9)
[1] 0.5
> #Accuracy of  rfsrccmodel
> AUC(y_pred=pred_rfsrc_N$data$response.v10,y_true=test_data$v10)
[1] 0.5071997
```

```
> #Accuracy of  rfsrccmodel
> AUC(y_pred=pred_rfsrc_N$data$response.v11,y_true=test_data$v11)
[1] 0.8407655
> #Accuracy of  rfsrccmodel
> AUC(y_pred=pred_rfsrc_N$data$response.v12,y_true=test_data$v12)
[1] 0.7298236
> #Accuracy of  rfsrccmodel
> AUC(y_pred=pred_rfsrc_N$data$response.v13,y_true=test_data$v13)
[1] 0.8171566
> #Accuracy of  rfsrccmodel
> AUC(y_pred=pred_rfsrc_N$data$response.v14,y_true=test_data$v14)
[1] 0.6429286
> #Accuracy of  rfsrccmodel
> AUC(y_pred=pred_rfsrc_N$data$response.v15,y_true=test_data$v15)
[1] 0.614958
> #Accuracy of  rfsrccmodel
> AUC(y_pred=pred_rfsrc_N$data$response.v16,y_true=test_data$v16)
[1] 0.5715041
> #Accuracy of  rfsrccmodel
> AUC(y_pred=pred_rfsrc_N$data$response.v17,y_true=test_data$v17)
[1] 0.5
> #Accuracy of  rfsrccmodel
> AUC(y_pred=pred_rfsrc_N$data$response.v118y_true=test_data$v18)
Error: unexpected '=' in "AUC(y_pred=pred_rfsrc_N$data$response.v118y_true="
> #Accuracy of  rfsrccmodel
> AUC(y_pred=pred_rfsrc_N$data$response.v18,y_true=test_data$v18)
[1] 0.8035025
> #Accuracy of  rfsrccmodel
> AUC(y_pred=pred_rfsrc_N$data$response.v19,y_true=test_data$v19)
```

[1] 0.5

> #Accuracy of rfsrccmodel

> AUC(y_pred=pred_rfsrc_N$data$response.v20,y_true=test_data$v20)

[1] 0.8852304

> #Accuracy of rfsrccmodel

> AUC(y_pred=pred_rfsrc_N$data$response.v21,y_true=test_data$v21)

[1] 0.5258427

> #Accuracy of rfsrccmodel

> AUC(y_pred=pred_rfsrc_N$data$response.v22,y_true=test_data$v22)

[1] 0.8111111

> #Accuracy of rfsrccmodel

> AUC(y_pred=pred_rfsrc_N$data$response.v23,y_true=test_data$v23)

[1] 0.616871

> #Accuracy of rfsrccmodel

> AUC(y_pred=pred_rfsrc_N$data$response.v24,y_true=test_data$v24)

[1] 0.5725639

> #Accuracy of rfsrccmodel

> AUC(y_pred=pred_rfsrc_N$data$response.v25,y_true=test_data$v25)

[1] 0.5

<u>Pros   of rfsrc</u>

- **Robust model for the best accuracy performs well on almost any data**
- **Prediction is very good**
- **Accuracy results in more than 90%**
- **<u>Cons of RFSRC</u>**
- **Run time is very high more than 15 mins**

```
> #Accuracy of rferns model
> Accuracy(y_pred=pred_rferns$data$response.v1,y_true=test_data$v1)
[1] 0.9269152
> #Accuracy of rferns model
> Accuracy(y_pred=pred_rferns$data$response.v2,y_true=test_data$v2)
[1] 0.8787794
> #Accuracy of rferns model
> Accuracy(y_pred=pred_rferns$data$response.v3,y_true=test_data$v3)
[1] 0.9633649
> #Accuracy of rferns model
> Accuracy(y_pred=pred_rferns$data$response.v4,y_true=test_data$v4)
[1] 0.8642181
> #Accuracy of rferns model
> Accuracy(y_pred=pred_rferns$data$response.v5,y_true=test_data$v5)
[1] 0.9027082
> #Accuracy of rferns model
> Accuracy(y_pred=pred_rferns$data$response.v7,y_true=test_data$v7)
[1] 0.9967538
> #Accuracy of rferns model
> Accuracy(y_pred=pred_rferns$data$response.v8,y_true=test_data$v8)
[1] 0.998516
> #Accuracy of rferns model
> Accuracy(y_pred=pred_rferns$data$response.v9,y_true=test_data$v9)
[1] 0.9819143
> #Accuracy of rferns model
> Accuracy(y_pred=pred_rferns$data$response.v10,y_true=test_data$v10)
[1] 0.9936004
> #Accuracy of rferns model
```

```
> Accuracy(y_pred=pred_rferns$data$response.v11,y_true=test_data$v11)
[1] 0.9784827
> #Accuracy of rferns model
> Accuracy(y_pred=pred_rferns$data$response.v12,y_true=test_data$v12)
[1] 0.8740493
> #Accuracy of rferns model
> Accuracy(y_pred=pred_rferns$data$response.v13,y_true=test_data$v13)
[1] 0.9194027
> #Accuracy of rferns model
> Accuracy(y_pred=pred_rferns$data$response.v14,y_true=test_data$v14)
[1] 0.9182897
> #Accuracy of rferns model
> Accuracy(y_pred=pred_rferns$data$response.v15,y_true=test_data$v15)
[1] 0.9824708
> #Accuracy of rferns model
> Accuracy(y_pred=pred_rferns$data$response.v16,y_true=test_data$v16)
[1] 0.9728251
> #Accuracy of rferns model
> Accuracy(y_pred=pred_rferns$data$response.v17,y_true=test_data$v17)
[1] 0.9986088
> #Accuracy of rferns model
> Accuracy(y_pred=pred_rferns$data$response.v18,y_true=test_data$v18)
[1] 0.9841402
> #Accuracy of rferns model
> Accuracy(y_pred=pred_rferns$data$response.v19,y_true=test_data$v19)
[1] 0.9883139
> #Accuracy of rferns model
> Accuracy(y_pred=pred_rferns$data$response.v20,y_true=test_data$v20)
[1] 0.9513077
```

```
> #Accuracy of rferns model
> Accuracy(y_pred=pred_rferns$data$response.v21,y_true=test_data$v21)
[1] 0.9741235
> #Accuracy of rferns model
> Accuracy(y_pred=pred_rferns$data$response.v22,y_true=test_data$v22)
[1] 0.9952699
> #Accuracy of rferns model
> Accuracy(y_pred=pred_rferns$data$response.v23,y_true=test_data$v23)
[1] 0.9437952
> #Accuracy of rferns model
> Accuracy(y_pred=pred_rferns$data$response.v24,y_true=test_data$v24)
[1] 0.976164
> #Accuracy of rferns model
> Accuracy(y_pred=pred_rferns$data$response.v25,y_true=test_data$v25)
[1] 0.975422


>Now evaluating with  AUC metric
> #Accuracy of rferns model
> AUC(y_pred=pred_rferns$data$response.v1,y_true=test_data$v1)
[1] 0.884844
> #Accuracy of rferns model
> AUC(y_pred=pred_rferns$data$response.v2,y_true=test_data$v2)
[1] 0.7703884
> #Accuracy of rferns model
> AUC(y_pred=pred_rferns$data$response.v3,y_true=test_data$v3)
[1] 0.8049867
> #Accuracy of rferns model
> AUC(y_pred=pred_rferns$data$response.v4,y_true=test_data$v4)
```

```
[1] 0.7062303
> #Accuracy of rferns model
> AUC(y_pred=pred_rferns$data$response.v5,y_true=test_data$v5)
[1] 0.8477085
> #Accuracy of rferns model
> AUC(y_pred=pred_rferns$data$response.v6,y_true=test_data$v6)
[1] 0.5128373
> #Accuracy of rferns model
> AUC(y_pred=pred_rferns$data$response.v7,y_true=test_data$v7)
[1] 0.5
> #Accuracy of rferns model
> AUC(y_pred=pred_rferns$data$response.v8,y_true=test_data$v8)
[1] 0.5
> #Accuracy of rferns model
> AUC(y_pred=pred_rferns$data$response.v9,y_true=test_data$v9)
[1] 0.5
> #Accuracy of rferns model
> AUC(y_pred=pred_rferns$data$response.v10,y_true=test_data$v10)
[1] 0.5071997
> #Accuracy of rferns model
> AUC(y_pred=pred_rferns$data$response.v11,y_true=test_data$v11)
[1] 0.8407655
> #Accuracy of rferns model
> AUC(y_pred=pred_rferns$data$response.v12,y_true=test_data$v12)
[1] 0.7303595
> #Accuracy of rferns model
> AUC(y_pred=pred_rferns$data$response.v13,y_true=test_data$v13)
[1] 0.8171566
> #Accuracy of rferns model
```

```
> AUC(y_pred=pred_rferns$data$response.v14,y_true=test_data$v14)
[1] 0.6429286
> #Accuracy of rferns model
> AUC(y_pred=pred_rferns$data$response.v15,y_true=test_data$v15)
[1] 0.614958
> #Accuracy of rferns model
> AUC(y_pred=pred_rferns$data$response.v16,y_true=test_data$v16)
[1] 0.5715041
> #Accuracy of rferns model
> AUC(y_pred=pred_rferns$data$response.v17,y_true=test_data$v17)
[1] 0.5
> #Accuracy of rferns model
> AUC(y_pred=pred_rferns$data$response.v18,y_true=test_data$v18)
[1] 0.8035025
> #Accuracy of rferns model
> AUC(y_pred=pred_rferns$data$response.v19,y_true=test_data$v19)
[1] 0.5
> #Accuracy of rferns model
> AUC(y_pred=pred_rferns$data$response.v20,y_true=test_data$v20)
[1] 0.8852304
> #Accuracy of rferns model
> AUC(y_pred=pred_rferns$data$response.v21,y_true=test_data$v21)
[1] 0.5258427
> #Accuracy of rferns model
> AUC(y_pred=pred_rferns$data$response.v22,y_true=test_data$v22)
[1] 0.8111111
> #Accuracy of rferns model
> AUC(y_pred=pred_rferns$data$response.v23,y_true=test_data$v23)
[1] 0.616871
```

> #Accuracy of rferns model

> AUC(y_pred=pred_rferns$data$response.v24,y_true=test_data$v24)

[1] 0.5725639

> #Accuracy of rferns model

> AUC(y_pred=pred_rferns$data$response.v25,y_true=test_data$v25)

[1] 0.5

- **pros of RFERNS**
- **Least run time just about a minute**
- **Accuracy is above 90%**

The results contains prediction of of test results containing prediction of target variables.