**INTERIM REPORT**

**AI Enabled IT Ticketing Service Tool**

*Submitted By*

*Disha Palan*

*Parita Desai*

*Gloria Preet*

*Vivek UImale*

Table of Contents

[Summary of problem statement, data and findings 3](#_Toc74499165)

[Problem Statement 3](#_Toc74499166)

[Objective 3](#_Toc74499167)

[Data 4](#_Toc74499168)

[Summary of the Approach to EDA and Pre-processing 5](#_Toc74499169)

[EDA 5](#_Toc74499170)

[Findings 7](#_Toc74499171)

[Other findings 9](#_Toc74499172)

[Further Data Analysis 9](#_Toc74499173)

[Word cloud for Short description 9](#_Toc74499174)

[Word cloud for Description 10](#_Toc74499175)

[Assignment group distribution of top 10 callers as compared to other callers 11](#_Toc74499176)

[Assignment group distribution amongst top 10 callers 12](#_Toc74499177)

[Assignment group distribution 12](#_Toc74499178)

[Feature Engineering 14](#_Toc74499179)

[Data Pre-processing 15](#_Toc74499180)

[Data Cleaning 15](#_Toc74499181)

[NER and POS Tagging 15](#_Toc74499182)

[Deciding Models and Model Building 16](#_Toc74499183)

[Models 17](#_Toc74499184)

[Traditional Models 17](#_Toc74499185)

[Sequential Models 18](#_Toc74499186)

[How to improve your model performance? 18](#_Toc74499187)

# Summary of problem statement, data and findings

## Problem Statement

In any of the IT industry, incident management plays an important role in delivering quality and timely support to its customers across the globe.

The incidents are generally created by various stakeholders like end users, vendors, IT users, etc. They might not have right information as to which team the ticket should go to. Hence, to improve and retain customer satisfaction, it is very important that the ticket is assigned to the right group of people for faster and appropriate resolution. In many Organizations this is still a manual process. There are few problems with the manual process:

1. Manual assignment of incidents is time consuming
2. It requires human efforts
3. There may be mistakes due to human errors and resource consumption is carried out ineffectively because of the misaddressing
4. Manual assignment increases the response and resolution times which result in user satisfaction deterioration / poor customer service

L1 / L2 needs to spend time to review Standard Operating Procedures (SOPs) before assigning to Functional teams (Minimum 25–30% of incidents needs to be reviewed for SOPs before ticket assignment).

15 mins are being spent for SOP review for each incident. Minimum of 1 FTE effort needed only for incident assignment to L3 teams.

During the process of incident assignments by L1 / L2 teams to functional groups, there were multiple instances of incidents getting assigned to wrong functional groups.

Around 25% of Incidents are wrongly assigned to functional teams. Additional effort needed for Functional teams to re-assign to right functional groups

During this process, some of the incidents are in queue and not addressed timely resulting in poor customer service and loss of business.

### Objective

We are building an AI solution which will enable organizations to classify incidents to the right functional group by implementing the best suited machine learning model and leading to customer satisfaction.

Guided by AI, organizations can reduce the resolution time and focus on more productive tasks. This will overcome and save time with below losses:

1. Time latency due to review of SOPs before assigning to right functional group
2. Incorrect assignments to functional groups

## Data

Reference: <https://drive.google.com/open?id=1OZNJm81JXucV3HmZroMq6qCT2m7ez7IJ>

The given dataset has below four columns:

1. Short description
2. Description
3. Caller
4. Assignment group

Out of above four columns we have 3 features namely, short description, description and caller and one target group namely assignment group

Top 10 records of our dataset :

Graphical user interface, text

Description automatically generated

# Summary of the Approach to EDA and Pre-processing

## EDA

We have used SweetViz and Panda profiling to Visualize and analyse our dataset. Below are attached reports for and observations from reports are in next sections. The reports can be viewed in GitHub link - https://github.com/dishapalan02/AI-Enabled-IT-Service-Ticketing-tool/tree/main

**SweetViz**:

Graphical user interface, application

Description automatically generated

**Pandas Profiling:**

A picture containing chart

Description automatically generated

Graphical user interface

Description automatically generated with medium confidence

Graphical user interface

Description automatically generated

Graphical user interface, application

Description automatically generated

Graphical user interface

Description automatically generated with low confidence

Chart, bar chart

Description automatically generated

## Findings

From Above two reports we have below observations:

1. Shape of the data - { Rows : 8500, Columns : 4 }
2. Total features - 3

2.1. Short Description - Text

2.2. Description - Text

2.3. Caller - Text

1. Target Column - 1

3.1 Assignment Group - Categorical

1. There are 84 duplicate records in total. Strategy to handle duplicates and the approach taken is defined in the pre-processing section below.
2. New features are required or not needs to be analysed further and also to check if below hidden patterns can be figured out:

   A. Common Issues -> user can be trained if possible   
   B. Common Caller -> May be user needs training or help with hardware or software  
   C. #ToDo To find if issue is controllable or not --> Check if possible .  
   D. To find if customer is happy with service or needs further imporvement and assistance

Now let's have a look at individual features:

1. **Short description**

A. Total values - 8492 ( > 99% )

B. Missing values - 8 ( < 1% )

C. Distinct values - 7481 (88%)

D. Mostly occurring value - password reset ( 0.4% )

E. We can also see the number of times each value is being repeated

F. Max length of statement - 159

G. It contains:

 Characters -> Lowercase Letter, Punctuation,   
 Uppercase Letter, Decimal Number,   
 Math Symbol, Math Symbol,   
 Modifier Symbol, Other Number,   
 Other Symbol, Currency Symbol  
   
 Scripts -> Common(ASCII) and Latin

H. Point G indicates that we have to translate the texts in the dataset based on the scripts as part of data pre-processing.

1. **Description**

A. Total values - 8499 ( > 99% )

B. Missing values - 1 ( < 1% )

C. Distinct values - 7817 ( 92% )

D. Mostly occurring value - it shows "the" ( 0.7% ) but will analyse further after the removal of stop words. But we consider the next which is windows password reset ( 0.3% )

E. We can also see the number of times each value is being repeated

F. Max length of statement - 13001

G. It contains:

 Characters -> Lowercase Letter, Punctuation,   
 Uppercase Letter, Decimal Number,   
 Math Symbol, Math Symbol,   
 Modifier Symbol, Other Number,   
 Other Symbol, Currency Symbol  
   
 Scripts -> Common(ASCII) and Latin

H. Point G indicates that we have to translate the texts in the dataset based on the scripts as part of data pre-processing.

1. **Caller**

A. Total values - 8500 ( 100% )

B. Missing values - no missing value

C. Distinct values - 2950 ( 35% )

D. Mostly occurring value - bpctwhsn kzqsbmtp (10%)

E. We can also see the number of times each value is being repeated

F. Max length of statement - 30

G. It contains:

 Characters -> Lowercase Letter, Space Separator,   
 Uppercase Letter, Connector Punctuation  
   
 Scripts -> Common(ASCII) and Latin

H. Point G indicates that we have to translate the texts in the dataset based on the scripts as part of data pre-processing.

1. **Assignment Group**

A. Total values - 8500 ( 100% )

B. Missing values - no missing value

C. Distinct values - 74

D. Mostly occurring value - GRP\_0 (47% ~ nearly half of the data --> Hence we can say that target class is highly imbalanced, so needs a strategy to be employed to reduce the bias here)

E. We can also see the number of times each value is being repeated

G. This indicates we can merge few assignment groups with smaller percentage to reduce overall number of categories.

### Other findings

1. There are duplicates which needs to be tackled
2. There are mojibake texts in the description and short description which needs to be processed
3. There are texts belonging to different languages which needs translations
4. There are email ids, blank spaces, dates, numbers which needs to be processed
5. There are missing values to be treated

## Further Data Analysis

### Word cloud for Short description

Text

Description automatically generated

### Word cloud for Description

Text, letter

Description automatically generated

### Assignment group distribution of top 10 callers as compared to other callers

Graphical user interface, text

Description automatically generated with medium confidence

### Assignment group distribution amongst top 10 callers

Graphical user interface, application

Description automatically generated with medium confidence

### Assignment group distribution

Graphical user interface

Description automatically generated

Chart, pie chart

Description automatically generatedA picture containing table

Description automatically generated  
From above we see that there is significant class imbalance with GRP\_0 being the majority class. Now let’s see the distribution of classes other than two majority classes GRP\_0 and GRP\_8 an d also for those classes which have ticket count less than or equal to 30 ( The number 30 is chosen using central limit theorem here)

Now we will see if we can ignore some classes or merge them into some other assignment group considering that other group has capability to resolve the ticket for this assignment group. For this we first divided the classes in some ticket bins and found the distribution as

Chart, pie chart

Description automatically generated

We see above that majority of assignment groups are one with tickets between 3 and 8 i.e. 17.6%. Also, from the above chart, we can see that Assignment group <= 2 tickets contributes to 14.87% i.e. (8.11 + 6.76%).

Also after comparing the descriptions and understanding that other groups have capability to resolve same tickets. We found that we can either merge or ignore below assignment groups :

A picture containing text, bird, plant

Description automatically generated

## Feature Engineering

***Null value treatment***

***Removal of Duplicates***

***N-gram analysis***

***Word cloud for each assignment group***

## Data Pre-processing

***Translation***

***Detecting different languages***

***Fix Encoding***

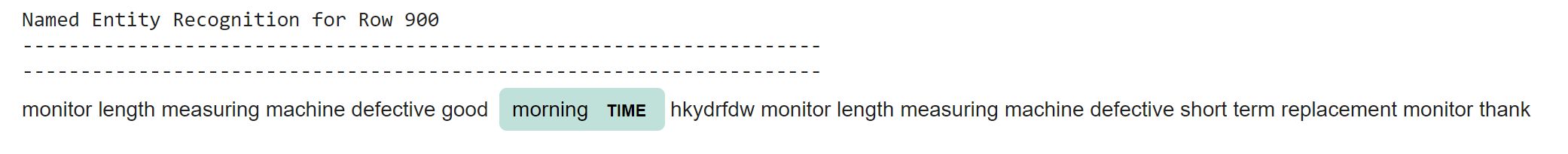
## Data Cleaning

***Removal of stop words***

***Lemmatization***

***Tokenization***

## NER and POS Tagging



Diagram

Description automatically generated

# Deciding Models and Model Building

Before building the model below steps were performed

Label Encoding

Splitting into train and test (80:20)

Treating class imbalance

The Assignment group before and after class balancing is shown as below:

Chart, pie chart

Description automatically generatedChart, pie chart

Description automatically generated

## Models

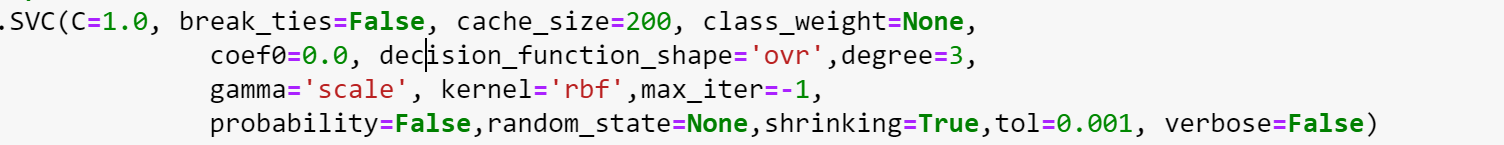
As this is a classification problem where we need to classify the assignment groups using the ticket description and short descriptions. We have selected below traditional models and sequential models

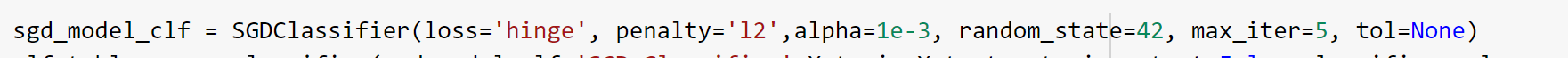
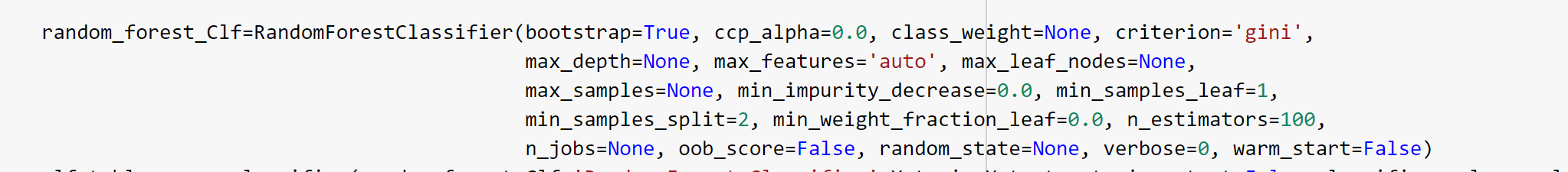
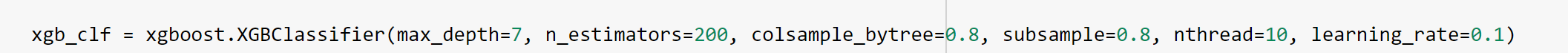
### Traditional Models

1. Multinomial NB Classifier



1. SVC Classifier



1. KNN Classifier
2. SGD Classifier
3. Random Forest Classifier
4. XGBOOST

We got below training and test accuracies and F1 score

|  |  |  |  |
| --- | --- | --- | --- |
| **Classifier** | **Train Accuracy** | **Test Accuracy** | **F1 Score** |
| Multinomial NB Classifier | 69.57% | 59.69% | 71.14% |
| SVC Classifier | 90.01% | 65.04% | 69.74% |
| KNN Classifier | 76.33% | 62.48% | 66.70% |
| SGD Classifier | 72.95% | 64.15% | 71.91% |
| Random Forest Classifier | 95.36% | 64.03% | 71.56% |
| XGBOOST | 92.70% | 66.59% | 71.07% |
|  |  |  |  |

### Sequential Models

1. CNN
2. LSTM

|  |  |  |  |
| --- | --- | --- | --- |
| **Classifier** | **Train Accuracy** | **Test Accuracy** | **F1 Score** |
| CNN |  |  |  |
| LSTM |  |  |  |

# How to improve your model performance?

* Still we are working with balancing the classes ( We are also trying SMOTE for this)
* We want to use PCA as there seems to be too much difference between training and test accuracies
* Plotting the accuracy graphs is in progress