Post-Graduation Program in Artificial Intelligence & Machine Learning

Batch: Jan 20-B

Capstone Project Group: NLP Group 6

Great Lakes and TEXAS McCombs

Automatic Ticket Assignment

Submitted By:
Kunal Rathod
Dibakar Paul
Neha Vyas
Arpita Jain
Rohit Mundlapati

Mentor:
Saurabh Bansal

Submission Date: January 8th, 2021
Submitted in Partial Fulfilment of the requirements for PGP in AIML

Table of Contents

1. Introduction

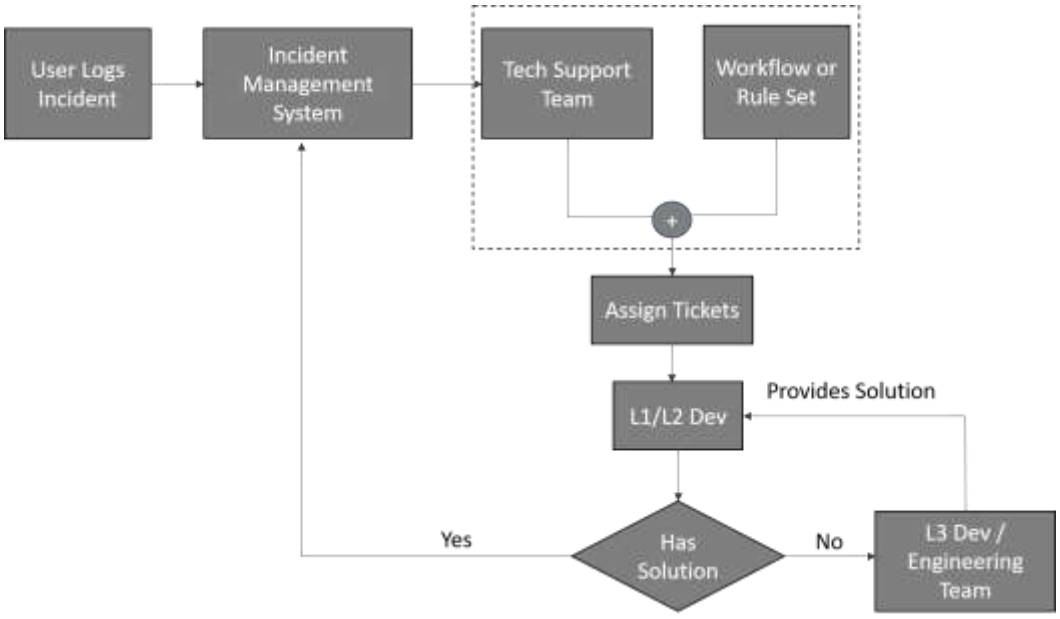
1.1 Background

The secret to business success is customer service. Good customer interactions create loyalty to the brand, drive more sales, and produce positive word-of-mouth. However, because of poor customer service, 66% of consumers switch products or services, and the key reasons for customer dissatisfaction are lack of efficiency and lack of speed. It's no wonder that consumers appreciate personalised, timely, and efficient customer service experiences in this age of technological innovation. However, companies are falling short in this since they don't have appropriate workflows to manage all their customer questions in location.

1.2 Business Problem Statement

Consumers meet firms, from social media networks and review pages to email and live chats, any time of the day, wherever they are, by using different channels. Moreover, the growing number of users of smartphones makes it easier for customers to get in touch than ever before and you can begin to imagine how tickets for customer service are starting to pile up.

When a support ticket drops into the help desk, first it needs to be processed and assigned a group or category so that it’s routed to the correct team member. This involves reading the ticket, so that agents know which category to choose.



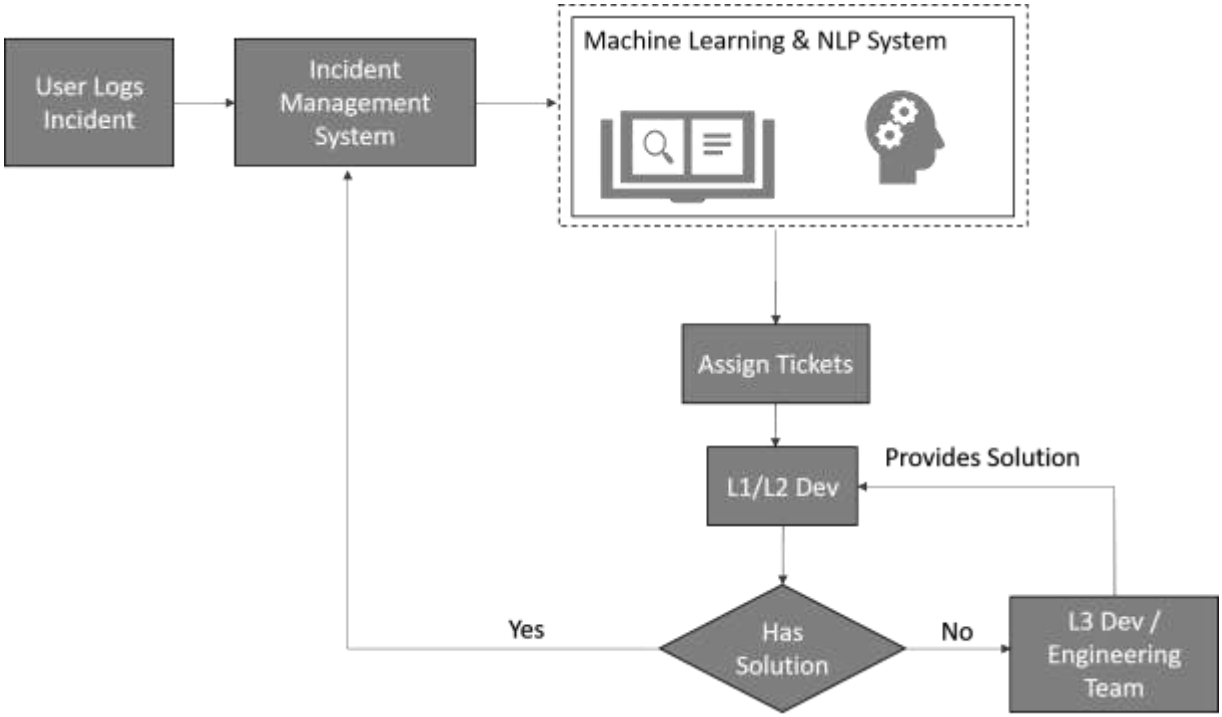
For any customer support team, the process of going through incoming tickets and allocating them to the agents best qualified to manage them is important, which suffers from below pain points in traditional incident management systems.

- Manually triaging high numbers of tickets is time consuming and extremely costly.
- It requires human efforts which may lead to inaccurate allocation of customer service agents due to human errors.

- Misaddressing of tickets leads to ineffective resource consumption.
- Around 25% of incidents are assigned to wrong functional groups.
- Additional effort needed for functional teams to reassign to right functional teams, during this process some of the incidents are in queue and are not addressed timely.
- Manual assignment increases the response and resolution times which result in user satisfaction deterioration / poor customer service.
- Cost involved in maintaining a team which works 24x7, including training cost for forming the team.

1.3 Proposed Solution

When businesses receive more customer questions through multiple channels, keeping up and maintaining lengthy queue is more difficult for support agents. Thus automatic incident assignment comes to rescue leveraging machine learning capabilities. Some of the major benefits of automatic classification of incidents are highlighted below:



1.4 Benefits of the Proposed Solution

Scalability: Categorize millions of incidents at a fraction of the cost of manual methods, save time so that agents can focus on more fulfilling tasks, and avoid inundating teams with heavy and repetitive workloads.

Availability: Available round the clock so incidents can get assigned immediately and can send a response in real-time, 24/7.

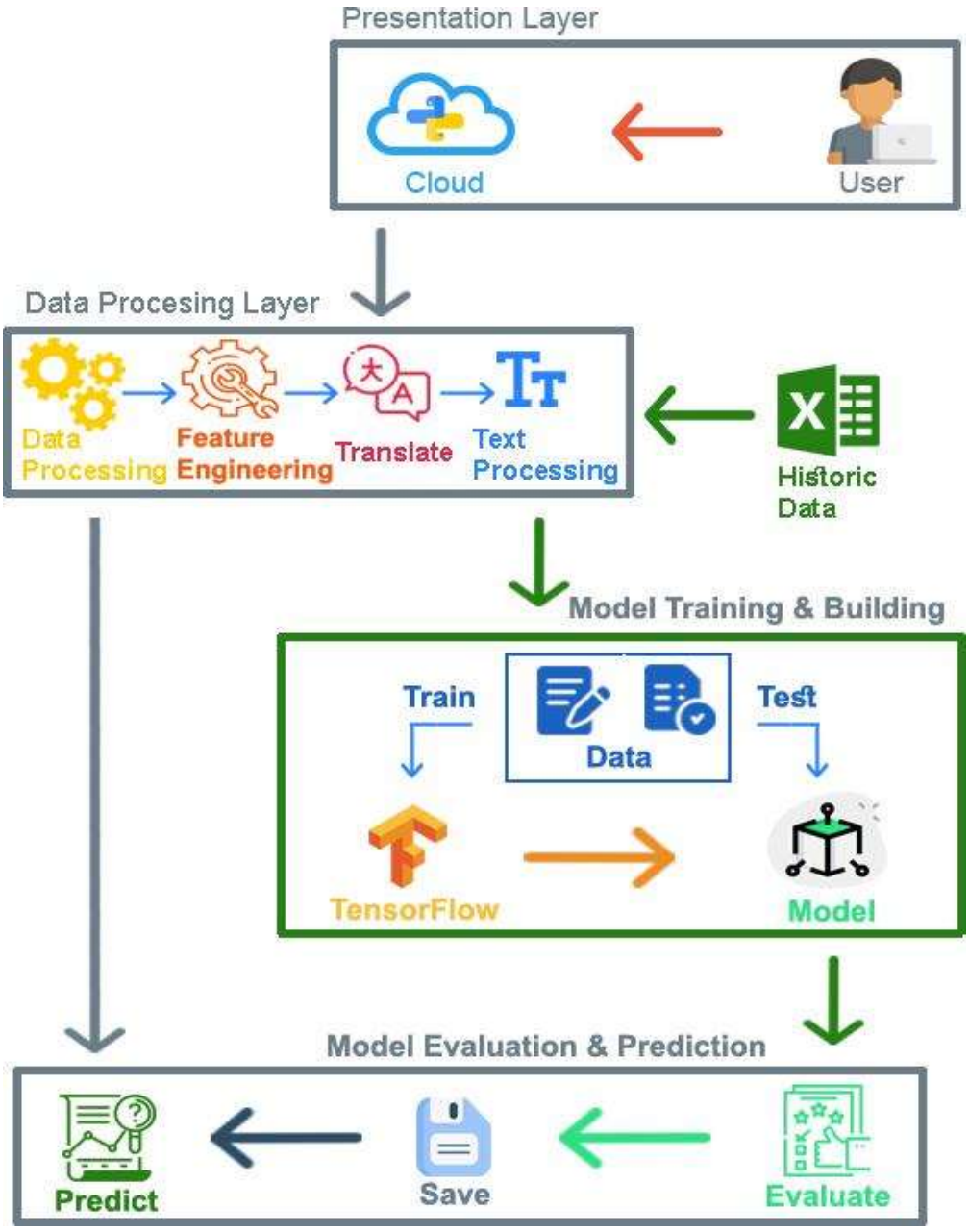
Real-time Analysis: Incident categorization data can provide valuable insights, to streamline processes of routing tickets to correct team members, and prioritizing tickets that are more urgent.

Consistent Criteria: Incident classification with machine learning enables grouping incidents accurately because it applies the same criteria to measure each set of data, plus a machine will never be subjective, lack alertness, and rush through tickets without understanding them properly.

Cost Saving: Saves cost to employ multiple employees or losing business due to customer dissatisfaction.

1.5 Architecture of the Proposed Solution

The proposed solution can further be integrated with any of the ITSM service based tools for end to end automation. A Flask microservice would be developed to deploy the model based classification as a Web Service which can further be exposed using a Restful API to communicate with any ITSM client service tool.



1.6 Data Source

The details of the data to build this classification model that can classify the tickets is available at the below link:

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

The Data source consists of a single spreadsheet with different attributes of the generated tickets. The shape of this historical data is **(8500, 4)** i.e. **8500 rows & 4 columns**.

A sample data is shown below:

Original Data Shape: (8500, 4)

	Short description	Description	Caller	Assignment group
0	login issue	-verified user details.(employee# & manager na...	spxjnwir pjlcoqds	GRP_0
1	outlook	\r\n\r\nreceived from: hmjdrvpb.komuaywn@gmail...	hmjdrvpb komuaywn	GRP_0
2	cant log in to vpn	\r\n\r\nreceived from: eylqgodm.ybqkwiam@gmail...	eylqgodm ybqkwiam	GRP_0
3	unable to access hr_tool page	unable to access hr_tool page	xbkucsvz gcpydteq	GRP_0
4	skype error	skype error	owlgqjme qhcozdfx	GRP_0
5	unable to log in to engineering tool and skype	unable to log in to engineering tool and skype	eflahbxn ltdgrvkz	GRP_0
6	event: critical:HostName_221.company.com the v...	event: critical:HostName_221.company.com the v...	jyoqwxhz clhxsoqy	GRP_1
7	ticket_no1550391- employment status - new non-...	ticket_no1550391- employment status - new non-...	eqzibjhw ymebpoih	GRP_0
8	unable to disable add ins on outlook	unable to disable add ins on outlook	mdbegvct dbvichlg	GRP_0
9	ticket update on inplant_874773	ticket update on inplant_874773	fumkcsji sarmtlhy	GRP_0
10	engineering tool says not connected and unable...	engineering tool says not connected and unable...	badgknqs xwelumfz	GRP_0
11	hr_tool site not loading page correctly	hr_tool site not loading page correctly	dcqsolkx kmsijcuz	GRP_0
12	unable to login to hr_tool to sgxqsuojr xwbeso...	unable to login to hr_tool to sgxqsuojr xwbeso...	oblekmrw qltgvspb	GRP_0
13	user wants to reset the password	user wants to reset the password	iftldbmu fujslwby	GRP_0
14	unable to open payslips	unable to open payslips	epwyvjsz najukwho	GRP_0
15	ticket update on inplant_874743	ticket update on inplant_874743	fumkcsji sarmtlhy	GRP_0
16	unable to login to company vpn	\n\nreceived from: xyz@company.com\n\nhi,\n\ni...	chobktqj qdamxfuc	GRP_0
17	when undocking pc , screen will not come back	when undocking pc , screen will not come back	sigfdwcj reofwzlm	GRP_3
18	erp SID_34 account locked	erp SID_34 account locked	nqdyowsm yqerwtna	GRP_0
19	unable to sign into vpn	unable to sign into vpn	ftsqkvre bqzrupic	GRP_0
20	unable to check payslips	unable to check payslips	mrzgjdal whnldmef	GRP_0

2. Exploratory Data Analysis

2.1 Data Pre-Processing

2.1.1 Explore Data

On exploring the data source in greater detail we were able to derive below insights that will further drive the solution of this problem statement.

Inference

Short Description & Description:

These columns contain the issue description about the ticket raised. It also has various special characters, HTML tags, email ids, text from multiple languages etc. which needs to be handled as a part of the Data Pre-processing steps

Caller:

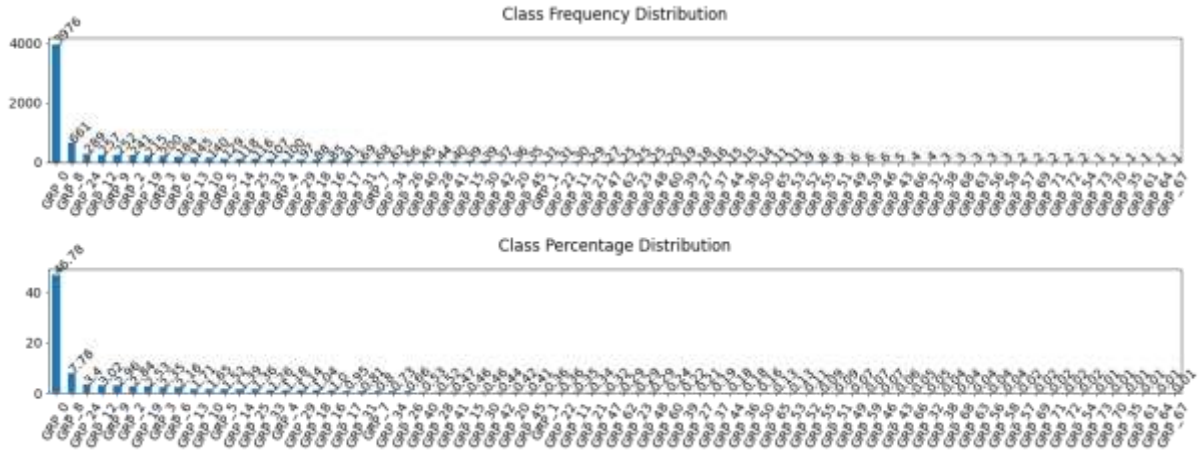
It has values that don't signify any particular feature. There is no specific pattern observed with remaining columns. For e.g. text "afkstcev utbnkyop" is getting related with multiple assignment groups like GRP_0, GRP_12, GRP_16, GRP_19, GRP_2, GRP_3, GRP_30, GRP_31, GRP_33, GRP_39, GRP_47, GRP_50 and GRP_69. Thus it doesn't help in inferencing anything significant hence we can drop it from the dataset.

Assignment group:

It has the list of groups to which the tickets are actually getting allotted. It will be our Target column since we need to assign the incoming tickets to these groups.

2.1.2 Explore Target Column

To solve any problem statement we have to understand what is actual outcome that is expected based on which the proposed solution is designed. In our case, the Target is Assignment Group which we have to predict from the available data. We will understand it further with the help of a graph



Inference

- After analyzing the graph & value counts it is evident that the Target column distribution is extremely skewed. There are 74 distinct assignment groups available in the dataset
- GRP_0 is the most dominant assignment group with accounts for 46.78% of data

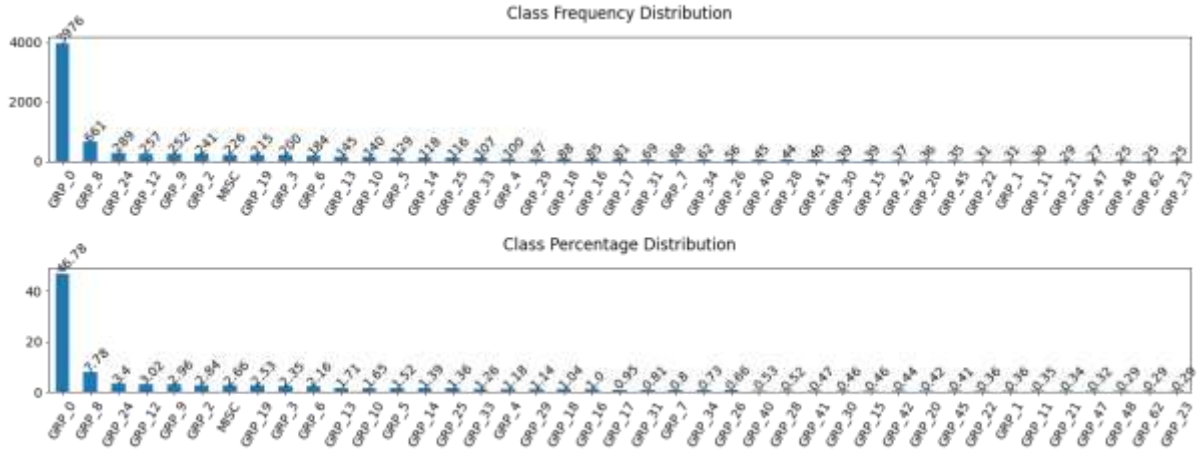
- There are 26 groups that have less than 10 assigned to them. Out of which 11 groups like GRP_58, GRP_57, GRP_69, GRP_70, GRP_67, GRP_71, GRP_72, GRP_64, GRP_61 and GRP_35 etc. have only 1 or 2 entries
- Thus, it implies that we have a very imbalanced data in hand

Challenge

Reduce the imbalance in data

Solution

We merge the groups with small entries which are contributing less than 0.25 % (or < 25 entries) to the data into a miscellaneous group. This will reduce the skewness to an extent. This has a business implication wherein the miscellaneous group has to be manually managed by someone and assign the tickets to the individual groups that are merged. Post merging, we have the following outcome.



We were able to bring down the distinct number of groups from 74 to 41.

2.1.3 Duplicate Data

Challenge

Data consist of good amount of duplicate rows as shown below

	Short description	Description	Assignment group
51	call for ecwtrjnq jpecxuty	call for ecwtrjnq jpecxuty	GRP_0
81	erp SID_34 account locked	erp SID_34 account locked	GRP_0
123	unable to display expense report	unable to display expense report	GRP_0
157	ess password reset	ess password reset	GRP_0
229	call for ecwtrjnq jpecxuty	call for ecwtrjnq jpecxuty	GRP_0
235	erp SID_34 account unlock and password reset	erp SID_34 account unlock and password reset	GRP_0
242	windows password reset	windows password reset	GRP_0
274	windows account locked	windows account locked	GRP_0
301	windows password reset	windows password reset	GRP_0
312	erp SID_34 account unlock	erp SID_34 account unlock	GRP_0
333	windows password reset	windows password reset	GRP_0
380	unable to login to erp SID_34	unable to login to erp SID_34	GRP_0
391	password reset request	password reset request	GRP_0
393	password reset	password reset	GRP_0
422	password reset	password reset	GRP_0

GRP_0	547
GRP_8	16
GRP_17	13
MISC	4
GRP_24	4
GRP_21	1
GRP_6	1
GRP_4	1
GRP_15	1
GRP_5	1
GRP_12	1
GRP_19	1

Solution

We exclude these rows from all further processing. There are 591 such duplicate rows in the dataset which are excluded. This reduces the shape of original data to (7909, 3)

2.1.4 Missing Data

Challenge

The data has very few missing values which can create issues later while processing. There are 6 missing values in the dataset as shown below

	Short description	Description	Assignment group
2604	NaN	\r\n\r\nreceived from: ohdrnswl.rezuibdt@gmail...	GRP_34
3383	NaN	\r\n-connected to the user system using teamvi...	GRP_0
3906	NaN	-user unable tologin to vpn.\r\n-connected to...	GRP_0
3924	NaN	name:wvqgbdhm fwchqjor\nlanguage:\nbrowser:mic...	GRP_0
4341	NaN	\r\n\r\nreceived from: eqmuniov.ehxkcbgj@gmail...	GRP_0
4395	i am locked out of skype	NaN	GRP_0

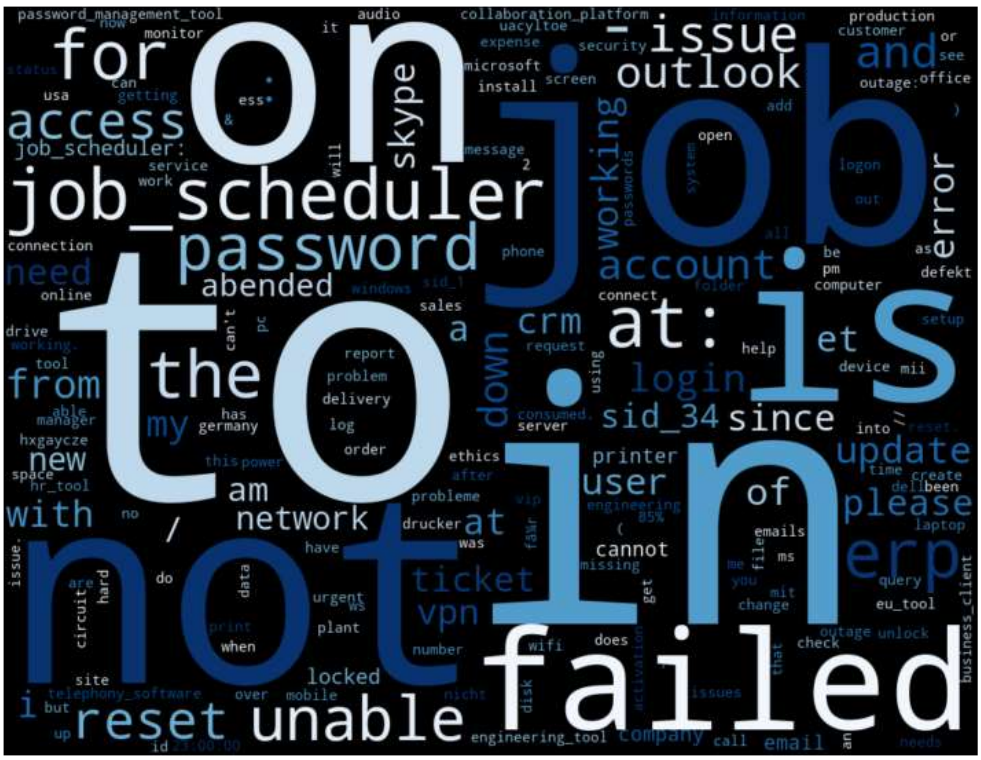
Solution

We replace the missing values with empty string. This enables us to handle the missing values as well as retain the information available within these rows.

2.2 Text Features

2.2.1 Short Description

It consists of brief description about the ticket which could also be considered as the title of the ticket. We can visualize its features using a word cloud.



The graph indicates that the top words in Short Description are stop words i.e. to, in, at, on, is etc. The other prominent words are job, job_scheduler, failed, unable, reset, erp etc.

2.2.2 Description

It consists of detailed description of the issue raised via ticket. In some cases, the short description & description have the same text as well. We can visualize its features using a word cloud.

The graph indicates that the top words in Description also include stop words i.e. to, in, the, from, is etc. The other prominent words include received, company, password, outlook, gmail etc.

Challenge

It is evident from the word clouds that both the columns i.e. Short Description & Description are focusing on different set of words (excluding the stop words). Indicating that both are required to predict the correct assignment group.

Solution

Looking at the data we could imply that both columns gives some information about the ticket. In some cases both are same & in some cases short description has proper text & description doesn't have any meaningful text. Hence, we merge Short Description & Description as one column.

This merging leads to emergence of duplicate words which we will eliminate first to make the Full Description look more appropriate and clean. The word cloud for this merged column is given below.

The word cloud of Full Description clearly show that it is an amalgamation of both Short Description & Description. Along with stop words, we can also observe unicode characters, numbers and special characters in the Word Cloud.

2.3 Text Pre-Processing

Pre-processing is mandatory to get the data in a consistent format. We will perform the following list of pre-processing on the merged text column Full Description (refer it as Description)

- Handle Unicode Characters
- Make text lowercase
- Remove hyperlinks
- Remove punctuation
- Remove numbers
- Remove special characters
- Remove HTML tags, unwanted spacing
- Specific words etc.

2.3.1 Unicode Characters

Challenge

The text column in the dataset contains Unicode characters as shown in the example below

é'å²›å…´å^æœºç"µshipment notificationé,®ç®±è®¾ç½® from: sent: friday, october 28, 2016 7:20 am to: nwfodmhc exurcwkm subject: re: dearï¼Œ pls help to update customer 4563729890 shipment notification email address : abcdegy@gmail.com b.

Solution

We have used a python library i.e. **ftfy** to convert these unicode characters into their respective non-English language format.

青岛兴合机电shipment notification邮箱设置 from: sent: friday, october 28, 2016 7:20 am to: nwfodmhc exurcwkm subject: re: dear, pls help to update customer 4563729890 shipment notification email address : abcdegy@gmail.com b.

2.3.2 Translation

Challenge

The text column also has data in other non-English languages along the converted text from unicode characters mentioned earlier.

青岛兴合机电shipment notification邮箱设置 from: sent: friday, october 28, 2016 7:20 am to: nwfodmhc exurcwkm subject: re: dear, pls help to update customer 4563729890 shipment notification email address : abcdegy@gmail.com b.
an mehreren pc`s lassen sich verschiedene prgramdntyme nicht öffnen. bereich cnc.
无法登陆hr_tool考勤系统 显示java插件无法加载,所需版本1.8.0.-45或更高版本。

Solution

We have used another python library i.e. **google_translate_new** for language translation from non-English to English words. This also improves the overall features since we have more information about the ticket after translation.

Qingdao Xinghe Electromechanical shipment notification email setting from: sent: friday, october 28, 2016 7:20 am to: nwfodmhc exurcwkm subject: re: dear, pls help to update customer 4563729890 shipment notification email address: abcdegy@gmail.com b.
Different programs cannot be opened on several pc's. area cnc.
Unable to log in to the hr_tool attendance system. It shows that the java plug-in cannot be loaded. The required version is 1.8.0.-45 or higher.

2.3.3 Text Cleansing (Noise removal)

Challenge

As mentioned earlier, the text column are having numbers, email ids, special characters, hyperlinks, punctuations, HTML tags, unwanted space and keywords like to:, from:, received from: and subject etc. These are not useful in predicting in fact it can lead to extra word count which might affect the model performance.

Qingdao Xinghe Electromechanical shipment notification email setting from: sent: friday, october 28, 2016 7:20 am to: nwfodmhc exurcwkm subject: re: dear, pls help to update customer 4563729890 shipment notification email address: abcdegy@gmail.com b.
Different programs cannot be opened on several pc's. area cnc.
Unable to log in to the hr_tool attendance system. It shows that the java plug-in cannot be loaded. The required version is 1.8.0.-45 or higher.

Solution

We have cleaned the data using python library **clean-text** as it better manages text containing emails, file location, hostnames etc. Using **regular expressions** to identify noisy patterns and keywords not

helping in prediction. Finally using another python library named **nostril** that enables us to remove some of the random words but not all.

qingdao xinghe electromechanical shipment notification email setting from sent friday october am to re dear pls help to update customer shipment notification email address b
different programs cannot be opened on several pc area cnc
unable to log in to the hr tool attendance system it shows that the java plug in cannot be loaded the required version is or higher

2.3.4 Lemmatization

Challenge

In the description text we can see different transformation of the same word. E.g. shows, showed, showing etc. We as human can understand that all these words are related to the act of show but for machine these are separate words and are treated differently.

prpf instead of prir for usa location and order operation mii showed the in status looking at erp was partially confirmed followed by automatic pause which adds change log shows all entries were done miiadmin it weird since erp is showing expected status provided but itself does not make sense reporting this issue so team can do deep dive come up with solution

Solution

We can resolve these particular scenario by performing Lemmatization of data. It is the process of grouping together the inflected forms of a word so they can be analyzed as a single item. We used a python library called **spacy** for this task. Thus, facilitates in reducing the number of unique word counts and eventually increase the overall performance of the model.

prpf instead of prir for usa location and order operation mii show the in status look at erp be partially confirm follow by automatic pause which add change log show all entry be do miiadmin it weird since erp be show expect status provide but itself do not make sense report this issue so team can do deep dive come up with solution

2.3.5 Stop words

Challenge

As observed in the word cloud and examples shared earlier, the text contains words like to, from, in, the, is, on, and, for, at etc. These are the most frequent words across the whole corpus of data available. In case of text classification, these words are of very little help in classifying the correct class. In fact it increases the computation time of the model.

prpf instead of prir for usa location and order operation mii show the in status look at erp be partially confirm follow by automatic pause which add change log show all entry be do miiadmin it weird since erp be show expect status provide but itself do not make sense report this issue so team can do deep dive come up with solution

Solution

In text processing, these words are known as Stop words. We have excluded these words using the same python library **spacy**. It tokenizes each sentence into words and identifies that each token/word is a stop word or not.

We have excluded stop words like not from the list and added custom stop words based on the data.

prpf instead prir location order operation mii status look erp partially confirm follow automatic pause add change log entry miiadmin weird erp expect status provide not sense report issue team deep dive come solution

2.3.6 Word Translation

Challenge

The Language translation performed earlier translates the whole text together hence in some case where majority of the words are in English and very few word in foreign language then it will not translate those words. The google translate works this way. Hence, our data still has words in foreign language

machine nao esta funcionando unable access utility finish drawer adjustment setting network
qingdao xinghe shipment notification email set send friday october subject dear help update customer shipment notification email address

Solution

Implemented a separate logic to translate the individual words into English language. Before that we created an exclusion list of words that are specific to business and doesn't need any translation. Words like skype, erp, hcm, mii, vpn, crm, Verizon, inplant, iphone, payslip, hostname, hrtool, infopath, workflow etc. are excluded from any further text processing

machine brain this working unable access utility finish drawer adjustment setting network
Please come galaxy shipment notification email set send friday october subject dear help update customer shipment notification email address

2.3.7 Exclude Non English Words

Challenge

Even after word translation, in the description text of some of the rows we could see few non-English words then did not make any sense. This increased the overall word count of the corpus.

access business client drawing xvgftyr tryfuh language browser microsoft internet explorer email customer number summary not netweaver application installs computer doesn page colleague
reset password wseacnvi erp qa block login jdhdw

Solution

Used **nltk** package and downloaded multiple corpus like **words, brown & webtext** to create a list of English words that should be allowed. These words along with words identified in exclusion list were allowed and others were excluded from the data corpus.

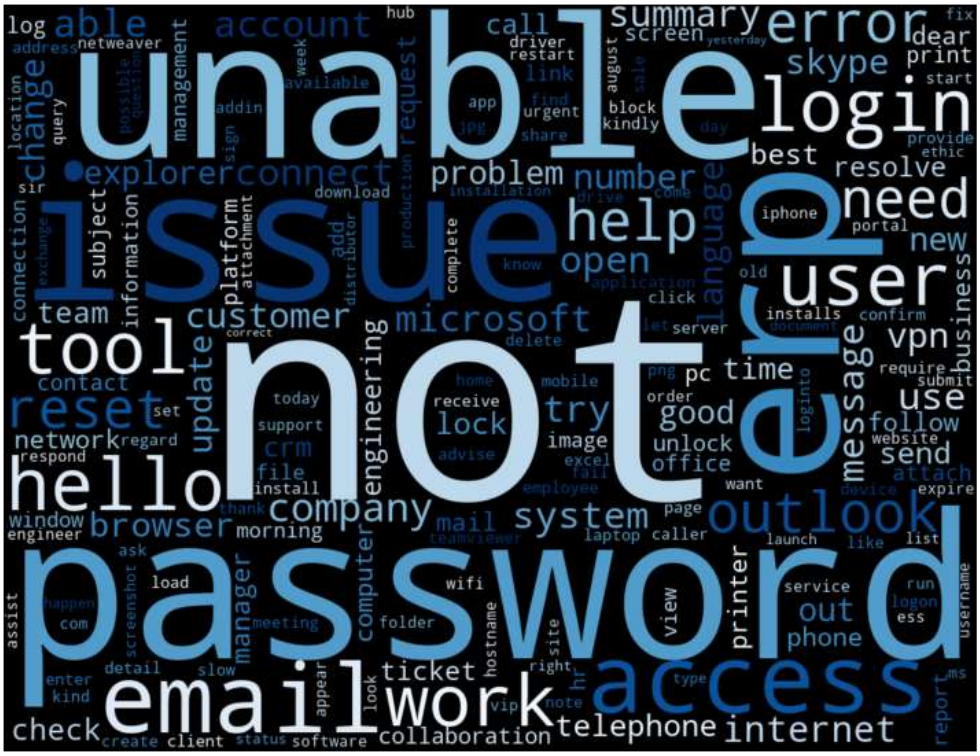
access business client drawing language browser microsoft internet explorer email customer number summary not netweaver application installs computer doesn page colleague
reset password erp qa block login

2.3.8 Data Pattern

Challenge

Now the dataset highlighted a pattern in which the rows has different texts but the eventual meaning was same for all rows especially when the assignment group was GRP_0. For e.g. ad

account lock, ad lock password, account lock ad, unlock ad account, ad lock, active directory lock etc. All these text means that the active directory account is locked and need password reset. The word cloud of GRP_0 data is shown below.



Solution

Implemented custom logic to identify such patterns and replace all the text with a single one and remove the redundancy in the data. Pattern matching was carried out in two ways, one matching whole text & another checking whether keywords are contained in the text. In this case, we replaced the text with “ad active directory account lock need unlock password reset” ensuring the main keywords are still present which will help model to learn. This step helped us in cleaning up the data to an extent. The shape of the GRP_0 dataset changed from **(3428, 2)** to **(2699, 2)**

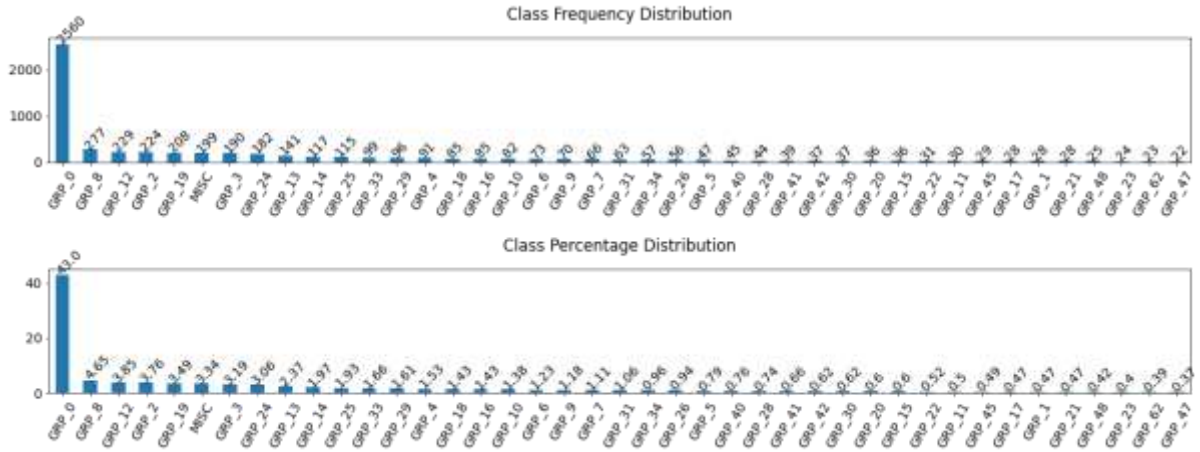
2.3.9 Final Output

We performed another round of stop word removal & removed all the duplicate data. The final output after performing various text preprocessing is represented in the word cloud below

Now, it looks much cleaner than the original one. The important words like issue, not, work, access, tool, error, password, email, user, erp, ticket, network reset, vpn etc. are correctly visible in the figure. To conclude, the text preprocessing task is completed successfully and the actual objective is achieved.

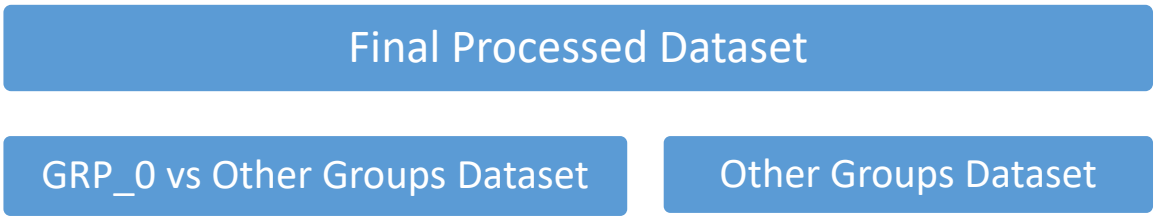
3. Proposed Solution Approach

After completing the EDA, data pre-processing and text pre-processing task we could still see that the data is imbalanced and GRP_0 is still the major contributor in the data. It still contributed 43% of the total data.



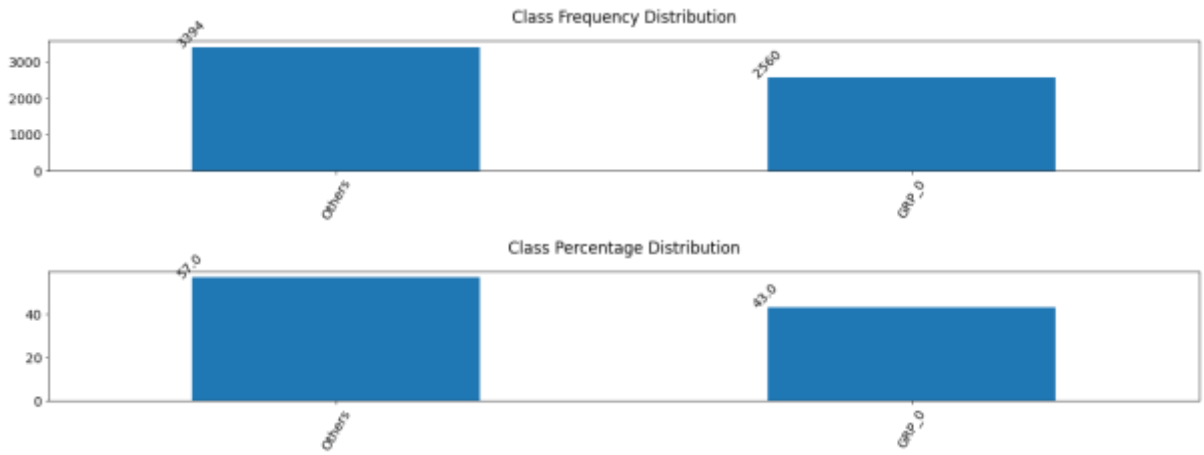
3.1 Logical Breakdown / Split

We performed a logical breakdown or split of the final processed dataset into two datasets as shown in the diagram below.



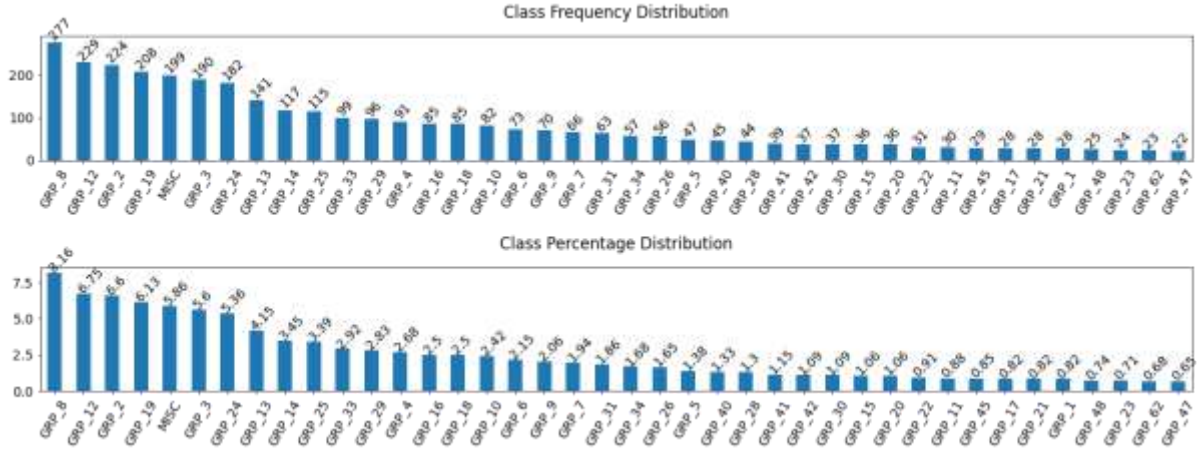
GRP_0 vs Other Groups

We introduced a new column named Is_GRP_0 to identify whether a particular row belongs to GRP_0 or not. Thus the shape of this dataset was updated to (5954, 3). Now, this dataset will fall under Binary Classification.



Other Groups

We filtered out the Final processed dataset on the basis of assignment group & fetched only those rows in which the group was not GRP_0. The shape of this new dataset is (3394, 2). This dataset will still fall under Multi-Class Classification.



Both these datasets will be trained separately by the models so we will have 2 models that will be finalized at the end. The final prediction of the assignment group will happen by the amalgamation of these 2 models wherein based on the output of the first model, the second model will be initiated. The detailed explanation is provided in Model Prediction section later.

3.2 Data Resampling

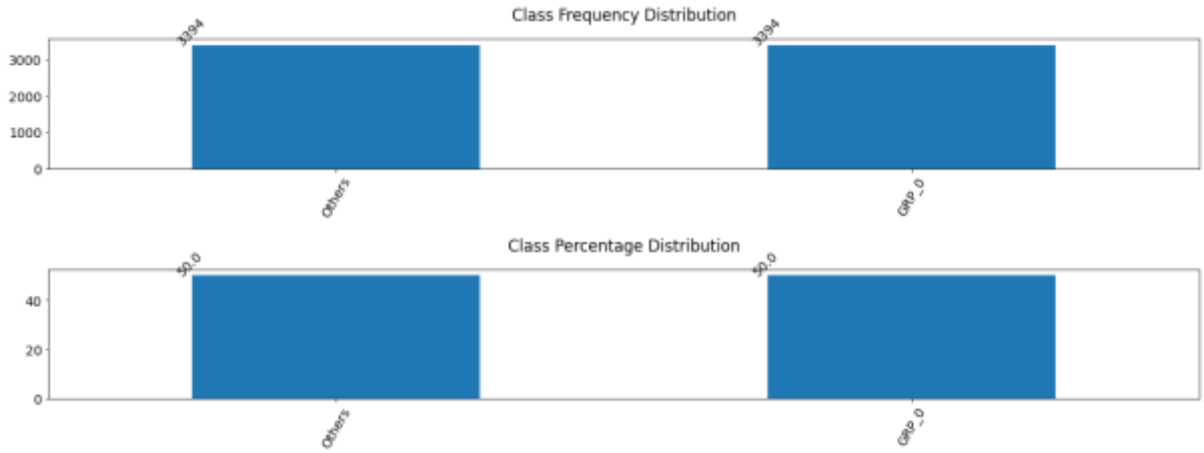
Challenge

The Logical Split has facilitated us with segregation of data but still the data is imbalanced. The second dataset i.e. other groups one is much skewed and we need to be rectified before passing to the models.

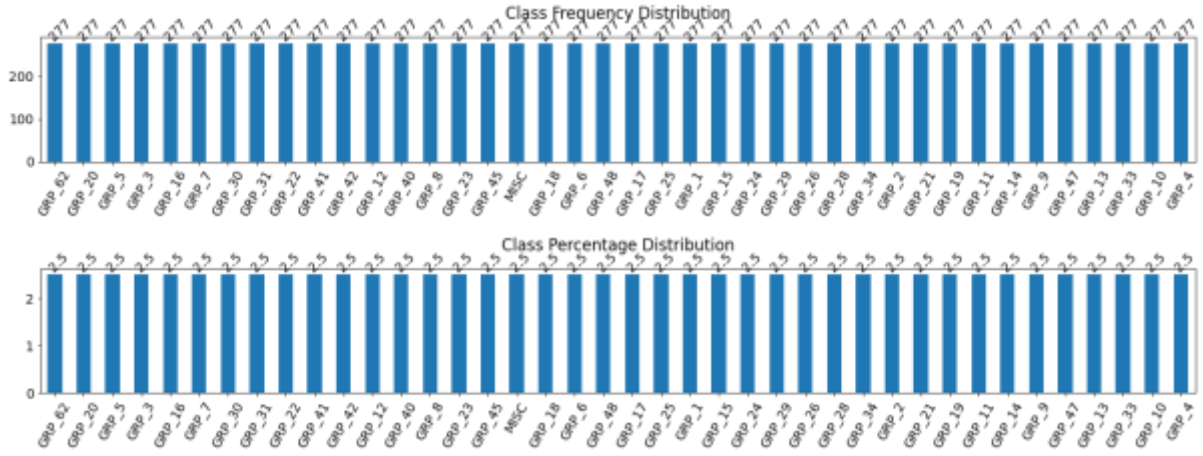
Solution

We have used the **resample** utility of **sklearn** python library to carry out resampling of the data. It resamples the data in a consistent way and implements the bootstrapping procedure. Bootstrap is a resampling technique that creates the bootstrap sample by sampling data points from the original dataset with replacement.

GRP_0 vs others dataset is also over sampled to ensure better learning & predictability of the models



Other group’s dataset is over sampled based on the maximum value count of the group within the dataset such that the groups with lower ticket count could be learned by the model and the overall accuracy of the model is improved.



4. Model Building

4.1 Nature of Problem

With reference to the Business Problem statement for this exercise, we are to analyze the text input from an end user describing an issue and thereafter predict the appropriate support group to assign the ticket automatically. So it is prudent that the nature of the problem is Multiclass Text Classification utilizing NLP capabilities.

Text Classification (a.k.a. text categorization or text tagging) is the process of classifying documents into predefined categories based on their content. It is the automated assignment of Natural Language texts to predefined categories.

Text Classifiers can be used to organize structure and categorize pretty much any kind of text. It takes text as an input, analyze its content and then automatically assign relevant tags.

Machine Learning, Natural Language Processing (NLP) and other AI-guided techniques are being used to automatically classify text in a faster, more cost-effective and more accurate manner.

4.2 Models

Multiclass Text Classification can be achieved using a number of different ML or DL based models for achieving optimal classification output.

4.2.1 Machine Learning Models

4.2.1.1 Logistic Regression

It is a supervised learning classification algorithm used to predict the probability of a target variable. It is one of the simplest ML algorithms that can be used for various classification problems such as spam detection, text classification etc. Logistic Regression is the most preferred ML algorithm when the dependent variables are Categorical.

We will use Multinomial type of Logistic Regression in which the dependent variable can have 3 or more possible unordered types or the types having no quantitative significance. It is easy to implement and does not require too many computational resources.

4.2.1.2 K-Nearest Neighbors

It is a supervised machine learning algorithm that can be used to solve both classification and regression problems. The algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other.

Very effective for text datasets and it naturally handles the multi-class classification problem but computationally this model is very expensive.

4.2.1.3 Support Vector Machine

SVMs are powerful yet flexible supervised machine learning algorithms which are used both for classification and regression. They are extremely popular because of their ability to handle multiple continuous and categorical variables. They are generally used in classification problems.

Robust against overfitting problems especially for text dataset due to its representation in high dimensional space. It also results in lack of transparency and memory complexity. Still usually offers good accuracy and uses less memory.

4.2.1.4 Naive Bayes

It is a classification technique based on applying Bayes' theorem with a strong assumption that all the predictors are independent to each other i.e. the presence of a feature in a class is independent to the presence of any other feature in the same class.

We will use Multinomial Naïve Bayes classifier for our problem. It works very well with text data. Very easy to implement and converges faster. It requires less training data and it highly scalable in nature.

4.2.1.5 Decision Trees

Decision tree analysis is a predictive modelling tool which can be constructed by an algorithmic approach that can split the dataset in different ways based on different conditions. The two main entities of a tree are **decision nodes**, where the data is split and **leaves**, where we got outcome.

Very fast algorithm for both learning and prediction. It can easily handle categorical features. But it is extremely sensitive to data and can over fit easily.

4.2.1.6 Random Forest

Random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result.

Very flexible, less variance then decision trees, works well with large range of data and possess very high accuracy. Complexity is high, constructing it is harder, more computational resources required and prediction process is time-consuming.

4.2.1.7 Bagging

It is an ensemble learning method that combines the weak learners. It often considers homogeneous weak learners, learns them independently from each other in parallel and combines them following some kind of deterministic averaging process or sits on top of the majority voting principle. It is also known as Bootstrap Aggregating.

The recursive nature of picking the samples at random with replacement can improve the accuracy of an unstable machine learning model. Additionally, it prevents overfitting and makes your model generalize better on unseen data. On the downside, it has large computational complexity and requires careful tuning.

4.2.1.8 Boosting

It is also an ensemble learning method that combines the weak learners. It often considers homogeneous weak learners, learns them sequentially in a very adaptive way (a base model depends on the previous ones) and combines them following a deterministic strategy.

The core concept of boosting focuses on those specific training samples that are hard to classify. When a weak-classifier misclassifies a training sample, the algorithm then uses these very samples to

improve the performance of the ensemble. It is known to decrease bias. On the downside, it also has large computational complexity.

We will use Gradient Boosting & XG Boosting Classifier for model building.

4.2.2 Deep Learning Models

4.2.2.1 LSTM

Long short-term memory (LSTM) is an artificial recurrent neural network (RNN) architecture. Unlike standard feedforward neural networks, LSTM has feedback connections. It can not only process single data points (such as images), but also entire sequences of data (such as speech or video).

A common LSTM unit is composed of a cell, an input gate, an output gate and a forget gate. The cell remembers values over arbitrary time intervals and the three gates regulate the flow of information into and out of the cell.

LSTM networks are well-suited to classifying, processing and making predictions. LSTMs were developed to deal with the vanishing gradient problem that can be encountered when training traditional RNNs.

4.2.2.2 Bi-Directional LSTM

Bidirectional LSTM are really just putting two independent RNNs together. This structure allows the networks to have both backward and forward information about the sequence at every time step.

Unlike LSTMs, Using bidirectional will run your inputs in two ways, one from past to future and one from future to past. In this way both the layers are co-trained simultaneously thus helping to maintain the context of input.

4.2.2.3 GRU

GRU supports gating and a hidden state to control the flow of information. Unlike LSTM, GRU does not have an output gate and combines the input and the forget gate into a single update gate.

GRU is less complex than LSTM and is significantly faster to compute and the results are almost similar to LSTMs.

4.3 Model Selection

Model selection is the process of choosing one among many candidate models for a predictive modeling problem. There may be many competing concerns when performing model selection beyond model performance, such as complexity, maintainability, and available resources. There are two main classes of techniques to approximate the ideal case of model selection

4.3.1.1 Probabilistic Measures

It involves analytically scoring a candidate model using both its performance on the training dataset and the complexity of the model. A hold-out test set is typically not required. For e.g. a highly biased model like the linear regression algorithm is less complex and on the other hand, a neural network is very high on complexity.

A fair bit of disadvantage however lies in the fact that probabilistic measures do not consider the uncertainty of the models and has a chance of selecting simpler models over complex models.

4.3.1.2 Resampling Methods

It estimate the performance of a model on out-of-sample data. These are simple techniques of rearranging data samples to inspect if the model performs well on data samples that it has not been trained on. In other words, resampling helps to understand if the model will generalize well.

4.3.1.2.1 Random Splits

Random Splits are used to randomly sample a percentage of data into training and testing sets. The advantage of this method is that there is a good chance that the original population is well represented in all the three sets. In more formal terms, random splitting will prevent a biased sampling of data.

We are going to use the train/test split method which is random splits technique under resampling methods for model selection.

4.3.1.2.2 Stratified K-Fold Cross-Validation

The cross-validation technique works by randomly shuffling the dataset and then splitting it into k groups. Thereafter, on iterating over each group, that group needs to be considered as a test set while all other groups are clubbed together into the training set. This helps in determining how the model generalizes and provides robust estimate of the performance of a model on unseen data.

The stratified k-fold ensures that each test fold gets an equal ratio of the target classes when compared to the training set. This makes the model evaluation more accurate and the model training less biased.

4.3.1.2.3 Bootstrap

Bootstrap is one of the most powerful ways to obtain a stabilized model. It is close to the random splitting technique since it follows the concept of random sampling. The first step is to select a sample size. Thereafter, a sample data point must be randomly selected from the original dataset and added to the bootstrap sample. After the addition, the sample needs to be put back into the original sample. This process needs to be repeated for N times, where N is the sample size.

Therefore, it is a resampling technique that creates the bootstrap sample by sampling data points from the original dataset with replacement. This means that the bootstrap sample can contain multiple instances of the same data point.

In this exercise, we are going to use all the resampling methods mentioned above for model selection. The random split technique is used to evaluate the model performance and accuracy. The stratified k-fold cross validation technique is used to check how the model generalize on unseen data. The bootstrap with replacement technique to resample the original dataset and create a more balanced dataset and evaluate model performance on the balanced dataset.

4.4 Evaluation Metrics

Models can be evaluated using multiple metrics. However, the right choice of an evaluation metric is crucial and often depends upon the problem that is being solved. A clear understanding of a wide range of metrics can help the evaluator to chance upon an appropriate match of the problem statement and a metric.

Evaluation metrics considered for our problem statement are mentioned below

Accuracy

It is the simplest metric and can be defined as the number of test cases correctly classified divided by the total number of test cases. It can be applied to most generic problems but is not very useful when it comes to unbalanced datasets.

Precision (Specificity)

Precision is the metric used to identify the correctness of classification. It is the ratio of correct positive classifications to the total number of predicted positive classifications. The greater the fraction, the higher is the precision, which means better is the ability of the model to correctly classify the positive class.

Recall (Sensitivity)

Recall tells us the number of positive cases correctly identified out of the total number of positive cases. Unlike precision that only comments on the correct positive predictions out of all positive predictions, recall provides an indication of missed positive predictions. It gives a measure of how accurately our model is able to identify the relevant data.

F1-Score

F1 score is the harmonic mean of Recall and Precision and therefore, balances out the strengths of each. It is useful in cases where both recall and precision can be valuable – like in the identification of plane parts that might require repairing. Here, precision will be required to save on the company's cost (because plane parts are extremely expensive) and recall will be required to ensure that the machinery is stable and not a threat to human lives.

Cross Validation Score

Cross validation is a technique for assessing how the statistical analysis generalizes to an independent data set. It is a gold standard for evaluating machine learning models by training several models on subsets of the available input data and evaluating them on the complementary subset of the data. Hence, the cross validation score is important as it provide the robust estimate of the performance of the model.

Execution Time

Execution time of the model is of equal importance. Along with being highly accurate we also need our model to have less training time and executes instantly. This will give an idea as to how much time the model will take if in case we need to retrain the model in future for any reason. Thus making it an important metric for model evaluation.

4.5 Model Performance

Using the above evaluation metrics we have evaluated all the models. The performance of each model is shown in detailed below based on the datasets on which the models are trained.

4.5.1 GRP_0 vs Other Groups Dataset

4.5.1.1 Machine Learning Models

Support Vector Machine and Logistic Regression model are neck to neck in terms of performance on each evaluation metric except for execution time where SVM is taking quite longer to get up and running. These models is closely followed by Multinomial Naïve Bayes model.

	Precision Score	Recall Score	Training Score	Testing Score	KFold CV	F1 Score	Accuracy Score	Execution Time
Logistic Regression	91.8524	91.6544	99.7685	91.6544	88.613	91.6447	91.6544	0.227701
K-Nearest Neighbors	89.4942	88.758	99.9158	88.758	85.9189	88.7058	88.758	1.07455
Support Vector Machine	91.9871	91.8508	99.9158	91.8508	88.8234	91.8443	91.8508	23.3751
Naive Bayes	90.1504	89.5925	96.927	89.5925	85.8768	89.5566	89.5925	0.013634
Decision Trees	85.6331	84.8797	99.9158	84.8797	85.8762	84.8	84.8797	1.54786
Random Forest	89.5551	88.8071	99.8948	88.8071	88.1498	88.7543	88.8071	53.4112
Bagging	88.9101	87.8252	99.9158	87.8252	87.8551	87.7404	87.8252	113.443
Gradient Boosting	88.9263	88.4634	99.6211	88.4634	88.1286	88.4294	88.4634	54.4109
XG Boosting	77.2779	77.2705	81.0356	77.2705	77.4572	77.2688	77.2705	5.65043

4.5.1.2 Deep Learning Models

Bi-directional LSTM and LSTM models are going hand in hand on all the parameters with almost same results. They are giving the same accuracy & f1 score. The only difference is execution time where LSTM is slower by 18-19 seconds. These models are followed by GRU model.

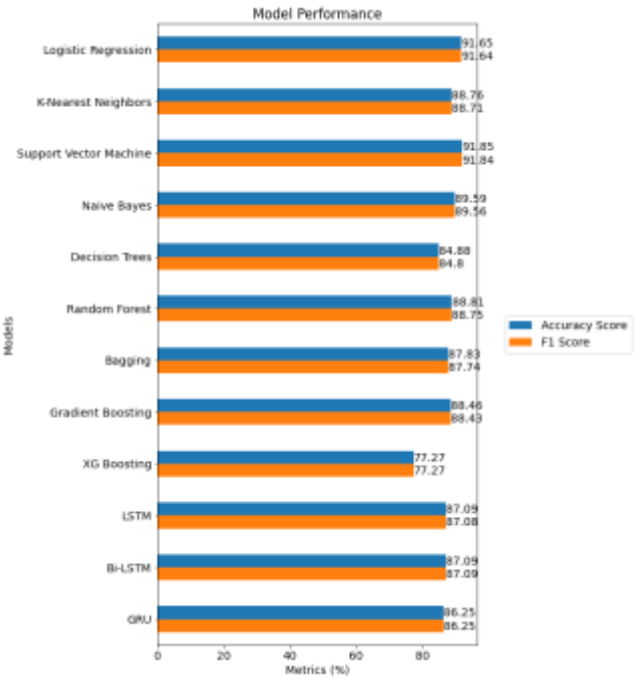
NOTE

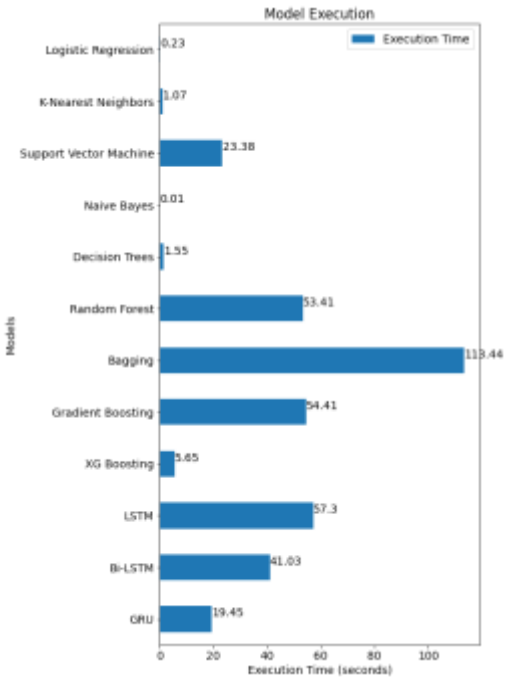
Training of Deep Learning models was performed using GPU hence we can see it was executed within a minute otherwise it can take up to 15-20 minutes to train them without GPU.

	Training Loss	Training Accuracy	Val. Loss	Val. Accuracy	Testing Loss	Testing Accuracy
LSTM	0.0236805	99.7895	0.707179	87.9075	0.737553	87.0889
Bi-LSTM	0.11099	98.5263	0.626171	88.1178	0.701387	87.0889
GRU	0.245902	95.5263	0.528897	88.4332	0.636744	86.2543

	Precision Score	Recall Score	F1 Score	Accuracy Score	Execution Time
LSTM	87.1611	87.0889	87.0824	87.0889	57.2959
Bi-LSTM	87.1045	87.0889	87.0874	87.0889	41.031
GRU	86.2579	86.2543	86.254	86.2543	19.4546

Overall the accuracy & f1 score of all the models along with their respective execution time on the **GRP_0 vs Other Groups dataset** can be illustrated in the figure below





4.5.2 Other Groups Dataset

4.5.2.1 Machine Learning Models

Again Logistic Regression and Support Vector Machine models are very close to each other in terms of performance of each evaluation metric. These models are followed by Multinomial Naïve Bayes and Random Forest which are very close among themselves in terms of performance but a bit far from first two models.

	Precision Score	Recall Score	Training Score	Testing Score	KFold CV	F1 Score	Accuracy Score	Execution Time
Logistic Regression	94.184	94.1937	98.8525	94.1937	92.5091	94.096	94.1937	38.6424
K-Nearest Neighbors	91.9913	91.6366	98.7107	91.6366	88.5897	91.3716	91.6366	5.42081
Support Vector Machine	94.4466	94.4043	98.8525	94.4043	91.8646	94.3468	94.4043	62.0406
Naive Bayes	93.5136	93.4116	98.4657	93.4116	90.7944	93.3625	93.4116	0.150626
Decision Trees	92.1077	92.0277	98.8525	92.0277	89.1956	91.8153	92.0277	3.0239
Random Forest	93.3947	93.231	98.8009	93.231	91.2328	93.1066	93.231	15.1945
Bagging	92.5426	92.509	98.8525	92.509	90.1625	92.336	92.509	260.584
Gradient Boosting	92.5875	92.4789	98.8525	92.4789	89.9304	92.367	92.4789	290.706
XG Boosting	86.8245	86.8532	92.8056	86.8532	83.935	86.341	86.8532	235.779

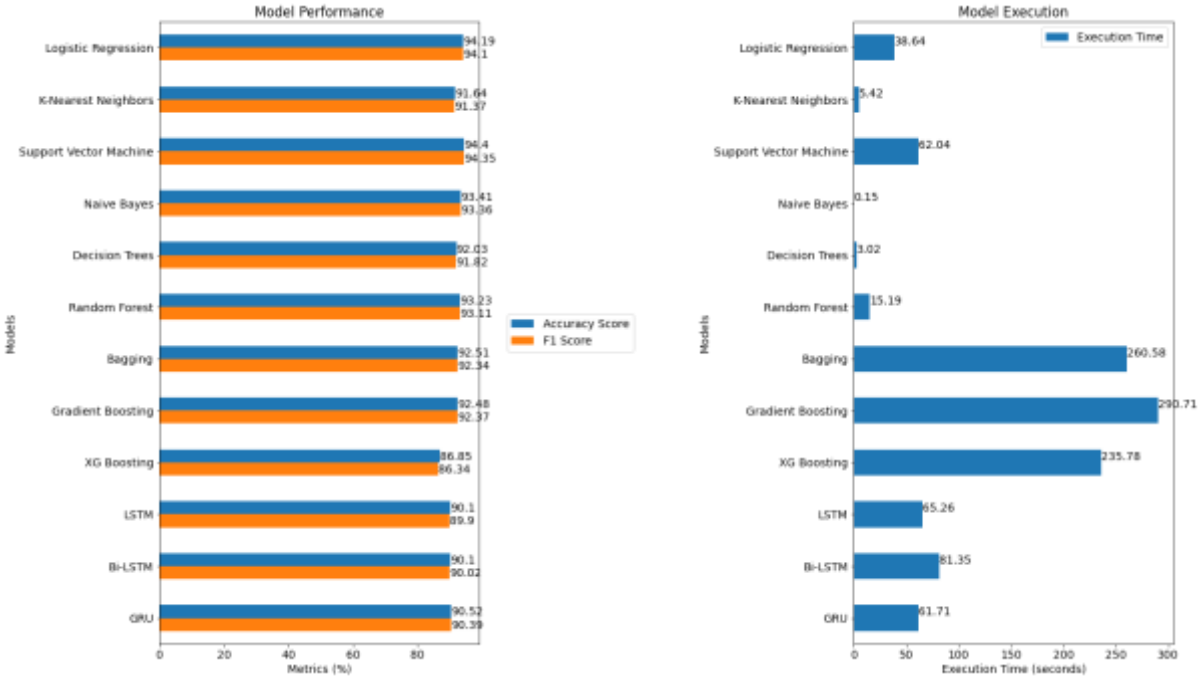
4.5.2.2 Deep Learning Models

GRU model is the better and faster one among the deep learning models for this dataset and very closely followed by the other two models i.e. Bi-directional LSTM and LSTM. Here too they are giving almost same results with difference in execution time.

	Training Loss	Training Accuracy	Val. Loss	Val. Accuracy	Testing Loss	Testing Accuracy
LSTM	0.134545	98.4526	0.581638	89.9485	0.551795	90.1023
Bi-LSTM	0.317699	98.1464	0.733411	90.6572	0.736556	90.1023
GRU	0.132399	98.6783	0.535683	89.6263	0.490796	90.5235

	Precision Score	Recall Score	F1 Score	Accuracy Score	Execution Time
LSTM	90.1792	90.1023	89.8961	90.1023	65.2628
Bi-LSTM	90.3396	90.1023	90.0171	90.1023	81.3502
GRU	90.5884	90.5235	90.3946	90.5235	61.7078

Overall the accuracy & f1 score of all the models along with their respective execution time on the **Other Groups dataset** can be illustrated in the figure below



4.6 Identify Best Model

The previous section clearly illuminates the best models for each dataset but we need to choose a single best model for each dataset to provide a robust and final solution to the business problem.

GRP_0 vs Other Groups Dataset

Support Vector Machine model has edge over Logistic Regression model, Random Forest & deep learning models to classify between GRP_0 & Other Groups dataset. It has good accuracy, f1 score & cross validation score even though it takes more time then Logistic Regression model. It has **accuracy of 91.85%, f1 score of 91.84%, cross validation score of 88.82% and execution time of 23.37 seconds.**

Other Groups Dataset

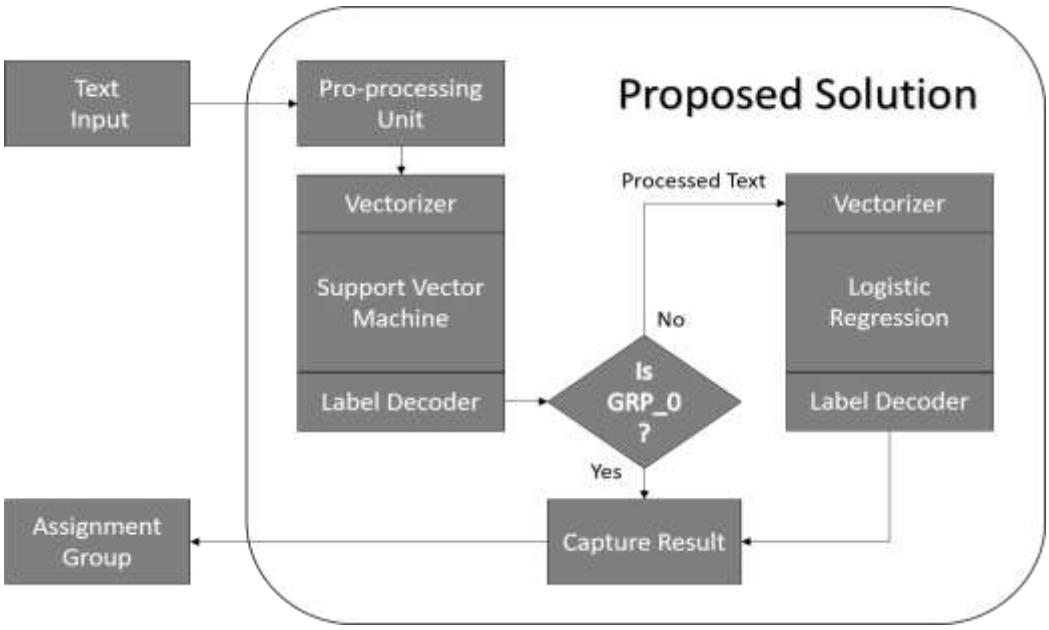
Support Vector Machine and Logistic Regression models were neck-to-neck (almost similar) to classify assignment group from the remaining non GRP_0 assignment groups. Hence it was a challenge to shortlist one model.

But, we have nominated **Logistic Regression model** as it has a better cross validation score i.e. better generalization and it faster than SVM but still marginally behind in accuracy and f1. It has **accuracy of 94.19%, f1 score of 94.10%, cross validation score of 92.51% and execution time of 38.66 seconds.**

5. Model Prediction

5.1 Chaining Models

The final stage of the proposed solution is the Model Prediction. In this stage we will implemented a strategic approach wherein we will chain both the models together to work in association with each other as illustrated below



Solution Flow

Input

The text entered by the user in the ticketing platform

Output

Target Assignment group based on the text supplied in the input.

Processing Steps

1. Input text goes to pre-processing unit for text cleansing and noise removal.
2. Processed text is provided to Vectorizer to generate model interpretable vectors
3. Support Vector Machine model predicts the output based on input vectors
4. Label Decoder will decode the model output to actual assignment group
5. Value of Assignment group
    a. If GRP_0 -> Processing ends. Jump to step 8.
    b. If Others -> Processing continues.
6. Pass the processed text to Vectorizer of Logistic Regression model.
7. Model will predict the output and Label Decoder will decode the model output to actual assignment group similar to earlier model.
8. Capture the assignment group & Return it as the final output of proposed solution.

5.2 Prediction Accuracy

In order to calculate the prediction accuracy we are using the processed test data which is not seen by the models while training. This data is already pre-processed so no additional steps are required.

We have structured an automated flow that will generate 10 different random test data sets each of around 100 samples from the test data and is passed to the chained models for prediction. Post that we calculated the prediction accuracy of each test set and derived the mean accuracy.

```
Prediction Accuracy of Set 1 : 90.0990099009901
Prediction Accuracy of Set 2 : 92.15686274509804
Prediction Accuracy of Set 3 : 91.17647058823529
Prediction Accuracy of Set 4 : 87.25490196078431
Prediction Accuracy of Set 5 : 95.09803921568627
Prediction Accuracy of Set 6 : 95.09803921568627
Prediction Accuracy of Set 7 : 93.13725490196079
Prediction Accuracy of Set 8 : 94.11764705882352
Prediction Accuracy of Set 9 : 94.11764705882352
Prediction Accuracy of Set 10 : 97.02970297029702
Mean Accuracy of the Test Data Prediction is 92.92855756163851
```

The strategic approach of chaining two models resulted in overall prediction accuracy of **92.93%**. It is a jump of **17.93%** in accuracy then the existing system which has **75%** accuracy.

6. Benchmark Comparison

Accuracy of the existing system is around 75% which also says that there is still misclassification of around 25%. This was set as a benchmark by us to begin with and we targeted to provide better accuracy than 75% and reduce the misclassification as much as possible.

The final solution that we are offering is able to achieve **92.93% accuracy** which is a significant improvement than the benchmark results. The primary reason of this improvement is the following list of machine learning techniques that we have incorporated in our solution.

- Data Pre-processing
    - Treating missing values and duplicate records
- Feature Engineering
    - Feature Selection
    - Feature Extraction
- Text Pre-processing
    - Decode unicode characters
    - Language & Word Translation
    - Noise Removal
    - Lemmatization
    - Treat Stop words & exclude Non English words
    - Treat Data Pattern
- Data Resampling
- Applying multiple Classification algorithms
- Tuning algorithm to produce best results
- Various Evaluation metrics

7. Implications

Abridged Manual Intervention

The proposed solution would automatically assign the tickets to their respective assignment groups. This will reduce the involvement of L1/L2 team to classify the tickets.

Superior Accuracy

Misclassification in the current setup is around 25%. This will be significantly reduced with the introduction of this automated solution.

Consistency

Proposed AI based solution will provide more consistency in accurate assignment of tickets.

Availability

System could be made available 24/7 with automatic ticket assignments

Cost Saving

Automation will facilitate in reducing the overall cost of maintaining the ticketing system.

Business Prospects

Substantial amount of human resources can be diverted to focus on more fulfilling tasks and growing the business rather than investing their time in such a repetitive task.

Enhance Ticketing Platform

More equipped ticketing platform with better categorization of tickets can help in streamlining and improving the structure of data. E.g. instead of user entering password reset there could be type a ticket called as Password Reset. Under that there could be a list of applications from which the user can select and raise ticket. This will prevent abnormalities caused due to manual text entry within the ticket thus ensuring better quality of data for training and predictions.

More Automated Solution

Self Service based automated solution for repetitive issues like Password Reset and Account unlock for various applications. This will completely eliminate SLAs and will make the user independent as their issues will be resolved almost instantly.

Revamp Classification Strategy

The data consists of overlapping tickets across multiple groups. A revamped classification strategy could assist in better segregation of tickets and also reduce the number of classifications itself. This will make the platform more manageable and efficient.

Endorse Collaboration

It would be always beneficial to engage the business users & solution architects to work collectively towards achieving the goal of providing a robust solution to the business problem. In this case, reviving the ticketing system

NOTE

Some recommendations made here will have its share of cost implications involved so it totally depends on the business to make a judicious decision that is in their best interest and could serve them better in future.

8. Limitations and Future Work

Unstructured Data

The better the structure of source data, the lesser is the need for pre-processing steps and eventually facilitates in arriving at the result in quicker time. Operations involved in Data pre-processing, Feature Engineering and Transformation, Text pre-processing are very time intensive. In this case, the data was unstructured like it consists of unicode characters, unwanted characters, and multiple language texts etc. thus processing it was time consuming job. Time taken was proportional to the total size of the data.

Manual Intervention

The proposed solution doesn't consider Assignment Groups with lower ticket count i.e. less than 25. These groups would be classified as "MISC" group by the model. Thus, manual intervention is required to classify them further.

Error-Prone

The solution could be error prone since we were not able to train the models with clean & accurate data for the problem we are trying to solve.

Data Acquisition

It is another pain point as we need more data to be collected for the lower Assignment Groups to perform better prediction but this can increase the cost.

Offline Training

It is required to understand new patterns within the system on which model has be further re-trained to carry out predictions.

Scope to identify Data patterns

The proposed solution has taken care of a number data pattern but still there is scope to identify more pattern and correct the data. This will help reducing the data imbalance.

Scope to attempt different techniques

Look to incorporate Data Augmentation techniques to reduce the imbalance within data. Give a shot at other sampling techniques and find out whether it can improve the performance further.

Usability

The designed solution doesn't have the capability to perform Real Time Classification of the Assignment Groups which could be achieved by integrating the solution with any of the ITSM or Ticket Submission platform.

Scalability

The proposed solution is a standalone solution being executed locally on a single machine. As a future enhancement for deployment, Cloud Services can be opted for robust utilization for real time prediction.

9. Closing Reflections

Data is the King

Learned the importance of spending time with actual data. Understanding the data in detail makes the road ahead easier.

No Thumb Rule

There is no thumb rule in designing an ML solution. Nothing is right or wrong while working on it, you have try multiple things simultaneously and find out the best that works for the given problem.

Working with Unstructured Data

As discussed in earlier sections, the dataset was unstructured. It posted us with a lot of challenge to remove the abnormalities and streamlined it for predictive modelling.

Working with Imbalanced Data

Along with being unstructured, the data was highly skewed. For some groups we had single digit records. This forced us to incorporate devise a strategic solution to the problem.

Looking for Patterns & Data Cleansing

Identifying pattern during EDA is an important step. It helped us to perform data cleansing and also understand granular details which filled us with confidence to provide logical recommendations to the business.

Model Diversity

Working with multiple models made us realize that every model has their own specific attributes & features. We need to deep dive into it to get the best out of those models.

Model Tuning

It is an exhaustive and time consuming task but still an important one to enhance the performance of models. It takes a lot of research to identify which parameter will provide the best results.

Model Generalization

Learned techniques that could help in generalizing models on unseen data. We implemented cross validation for ML models and Dropout layers for DL models.

Explore Different Techniques

Always look forward to explore different machine learning & deep learning techniques that could help in enhancing the overall performance of the solution.

Horses for Courses Approach

Each problem statement is unique in its own way. Each dataset has its own features and attributes. Each model cater differently to different dataset. Hence, the approach for designing each solution will be different based on the problem at hand.

Hands on Experience

We got a hands on experience about how an AIML solution is approached within the industry and what to expect while working on such solutions in future.