INTERIM REPORT

Automatic Ticket CLassification

NLP

CAPSTONE PROJECT - GREAT LEARNING PGAIML

GROUP 9

SUPERVISED BY

Vaibhav Kulkarni

1. Summary of the Problem Statement, Data and Findings

1.1. Abstract

Excellent & Effective Customer Support is Quintessential to the running of any business organization, no matter its size—84% of organizations working to improve customer service report an increase in revenue. In the current scenario, various incidents faced by the business are all assigned to two L1/L2 teams. Only 54% of these incidents can be resolved at this level. For all the rest, the incidents are escalated to L3 teams to be resolved. Additionally, the manual reassignment to various functional groups was found to have an error rate of around 25%. This added overhead cost of time and resources of re-assigning the incidents is detrimental to the customer support efficiency causing delays and bad customer experiences. A better allocation and practical usage of the functional groups' resources will result in substantial cost, time savings and better customer support overall.

Hence, we aim to build a classifier using state-of-the-art NLP techniques to classify the tickets to various functional groups by just analyzing the text of the various issues, thereby driving direct business value in IT customer support.

1.2. Dataset

- Our dataset consists of 8500 data points each consisting of a short description of the issue, a longer description, the caller name (appears to be encrypted and anonymized in the given dataset to protect privacy) and the target class group to which the issue has to be assigned to.

```
dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8500 entries, 0 to 8499
Data columns (total 4 columns):
 #   Column             Non-Null Count  Dtype 
---  ------             --------------  ----- 
 0   Short description  8492 non-null   object
 1   Description        8499 non-null   object
 2   Caller             8500 non-null   object
 3   Assignment group   8500 non-null   object
dtypes: object(4)
memory usage: 265.8+ KB
```

- There seem to be missing values in the Short description and Description columns, which needs to be looked into and handled. There are 8 nulls/missing values present in the Short description and 1 null/missing value present in the description column.

```
dataset.isna().sum()  # Few missing values
```

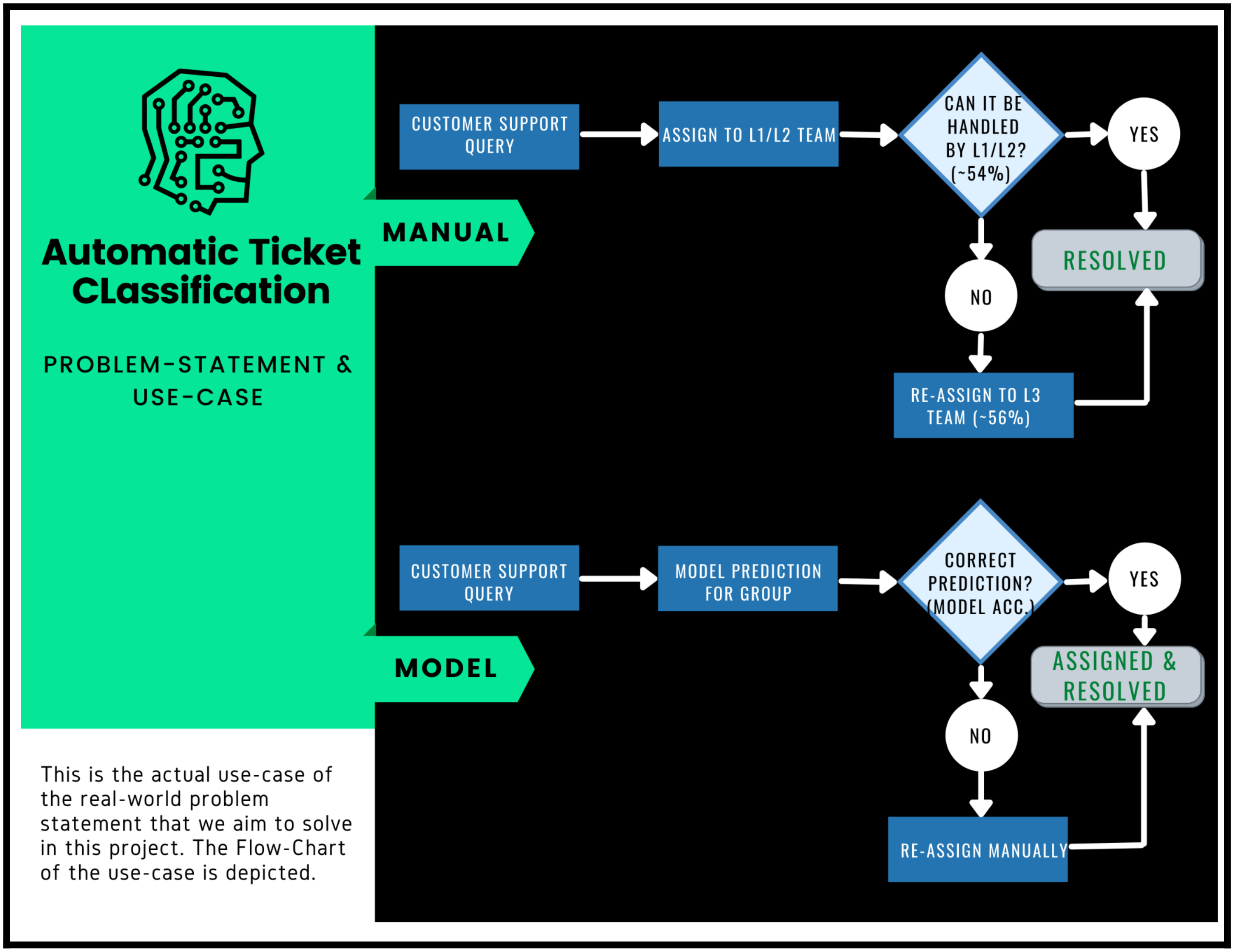
```
short_description    8
description          1
caller               0
group                0
dtype: int64
```

```
dataset[dataset.isna().any(axis=1)]  # check rows with missing values
```

	short_description	description	caller	group
2604	NaN	\r\n\r\nreceived from: ohdrnswl.rezuibdt@gmail...	ohdrnswl rezuibdt	GRP_34
3383	NaN	\r\n-connected to the user system using teamvi...	qftpazns fxpnytmk	GRP_0
3906	NaN	-user unable tologin to vpn.\r\n-connected to...	awpcmsey ctdiuqwe	GRP_0
3910	NaN	-user unable tologin to vpn.\r\n-connected to...	rhwsmefo tvphyura	GRP_0
3915	NaN	-user unable tologin to vpn.\r\n-connected to...	hxripljo efzounig	GRP_0
3921	NaN	-user unable tologin to vpn.\r\n-connected to...	cziadygo veiosxby	GRP_0
3924	NaN	name:wvqgbdhm fwchqjor\nlanguage:\nbrowser:mic...	wvqgbdhm fwchqjor	GRP_0
4341	NaN	\r\n\r\nreceived from: eqmuniov.ehxkcbgj@gmail...	eqmuniov ehxkcbgj	GRP_0
4395	i am locked out of skype	NaN	viyglzfo ajtfzpkb	GRP_0

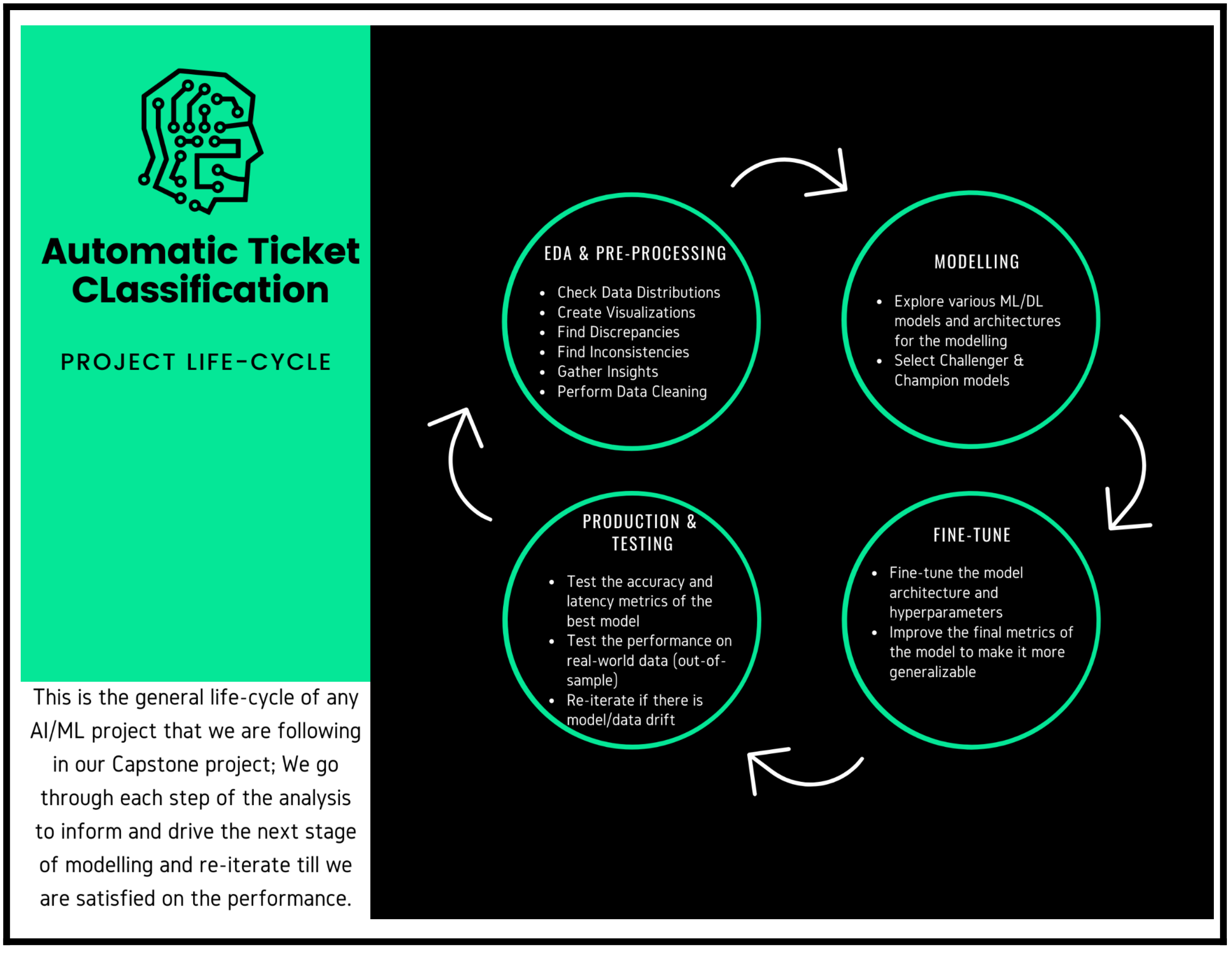
-
- The independent features are short descriptions and descriptions and the target/dependent feature is the group.

1.3. Use-Case

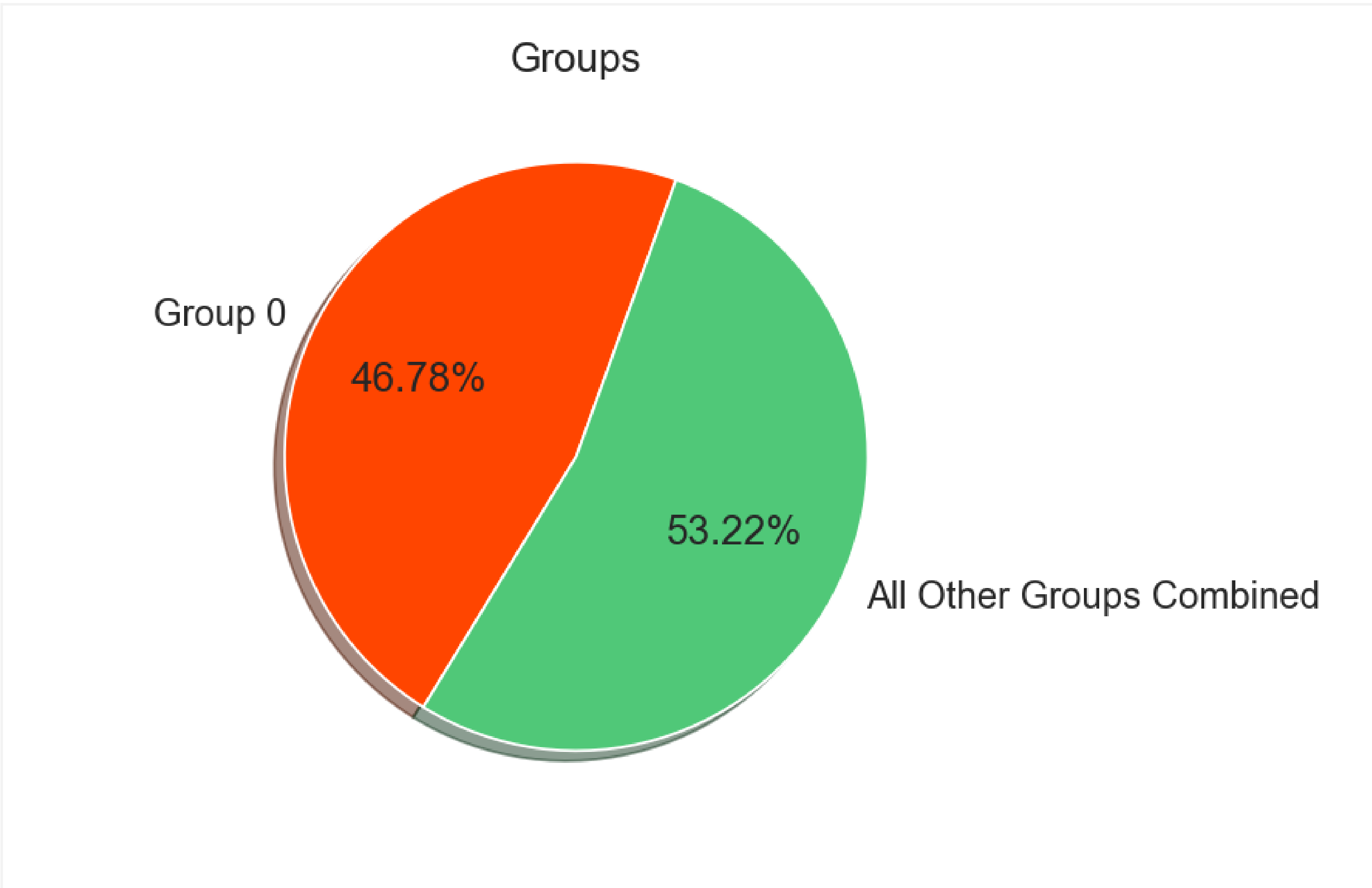

Automatic Ticket CLassification
PROBLEM-STATEMENT & USE-CASE
This is the actual use-case of the real-world problem statement that we aim to solve in this project. The Flow-Chart of the use-case is depicted.
MANUAL
CUSTOMER SUPPORT QUERY
ASSIGN TO L1/L2 TEAM
CAN IT BE HANDLED BY L1/L2? (~54%)
YES
RESOLVED
NO
RE-ASSIGN TO L3 TEAM (~56%)
MODEL
CUSTOMER SUPPORT QUERY
MODEL PREDICTION FOR GROUP
CORRECT PREDICTION? (MODEL ACC.)
YES
ASSIGNED & RESOLVED
NO
RE-ASSIGN MANUALLY

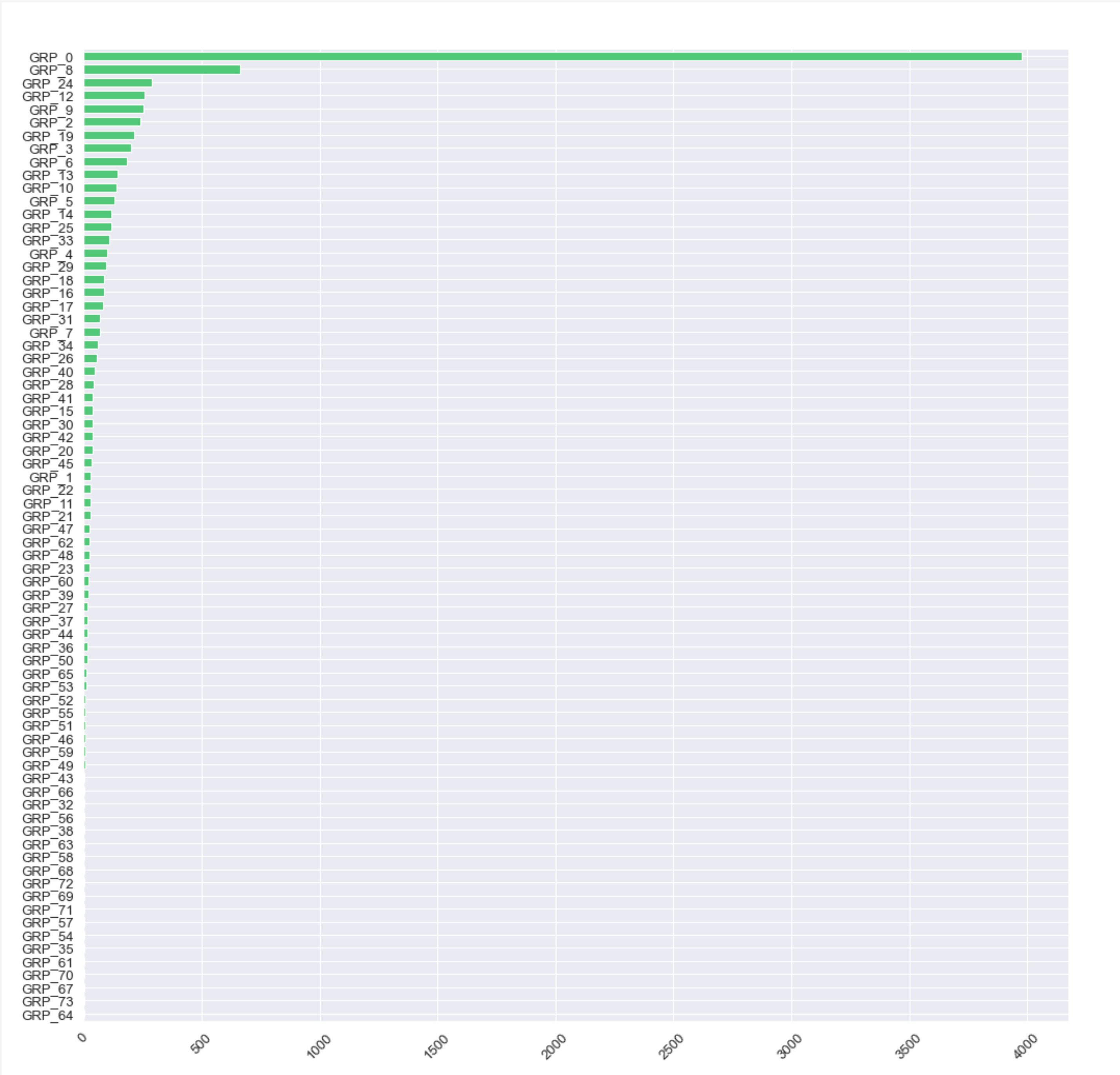
2. Summary of the approach to EDA and Pre-Processing

2.1. Project Life-Cycle



2.2.1. Target Distribution





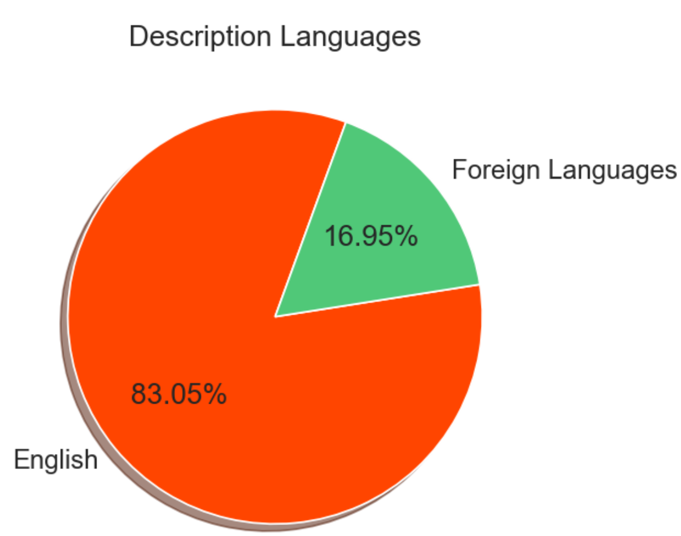
- The Target class distribution is extremely skewed and heavily imbalanced as the majority of incidents are from Group 0 followed by Group 8, 24, 12, 9, 2 and we find an imbalanced dataset for the rest of the groups.
- A large no. of entries for “Group 0” which account for ~47% of the data and remaining are grouped together as “Other” as there is not much information with the groups individually.

2.2.2. Choosing a Metric

- This is a multi-class classification problem, where the machine learning model will try to predict if each row is one of the 74 possibilities.
- The majority class is GRP_0, which occurs in 46.78% of the observations.
- The most common metrics for a multi-class classification problem are AUC, F1-score and accuracy.
- Accuracy is not suitable for an imbalanced classification problem. (Note that a model that always predicts GRP_0, will get an accuracy of 46.78%)
- We would choose F1-Score if the majority class is more important than the smaller classes.
- We would choose AUC if we also care about the smaller classes.

As we want to be able to classify the tickets into all functional groups and functional groups are given equal importance, we choose ***AUC*** *as the final metric to score model performance.*

2.2.3. Language Detection



- We have seen around 16.95% of the texts in the description. These foreign language texts include German, French, Chinese, Italian and other European languages in relatively small numbers.
- The models in Milestone-1 are working with only English texts with foreign languages dropped.
- We aim to deal with this variation in languages by using a language translation API or model to translate them first to English before using them in the model in Milestone-2.

2.2.4. Keyword Extraction

- YAKE is a lightweight unsupervised automatic keyword extraction using an unsupervised approach that rests on text statistical features extracted from single documents to select the most important keywords of a text and is independent of corpus, domain and language.
- The ten state-of-the-art unsupervised approaches followed here are TF.IDF, KP-Miner, RAKE, TextRank, SingleRank, ExpandRank, TopicRank, TopicalPageRank, PositionRank and MultipartiteRank.
- These keywords can then be used as a separate feature for the models and also for unsupervised clustering of groups based on the keywords later on.

```
print(test)
```

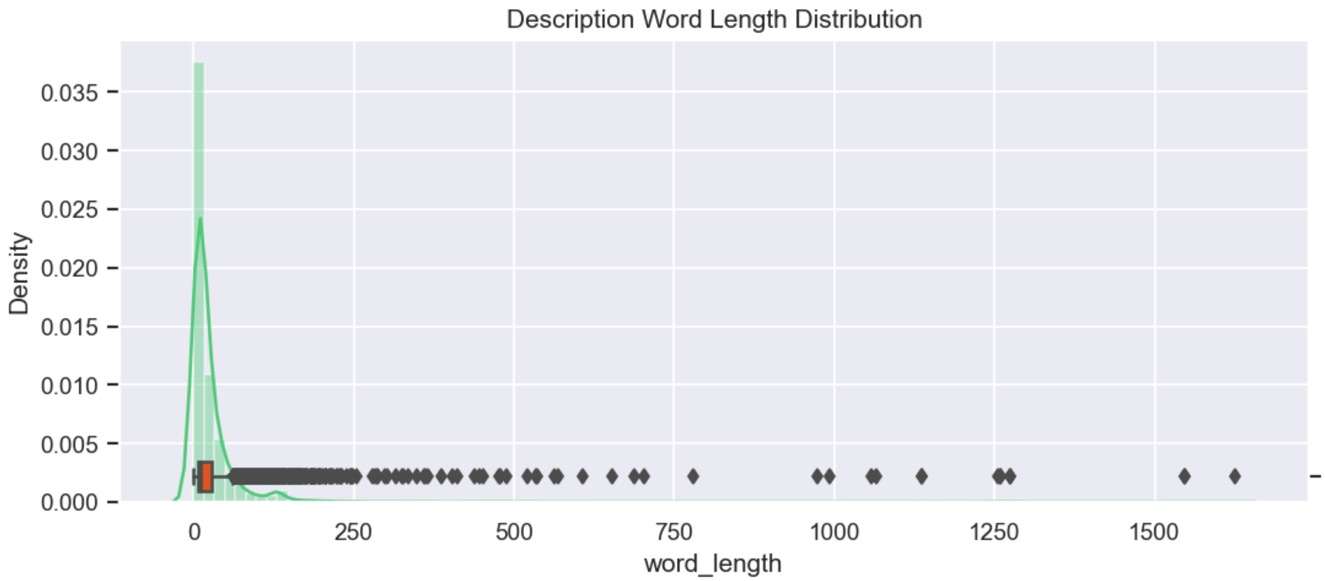
```
circuit outage: vogelfontein, south africa mpls circuit is down at 8:14 am et on 08/08
```

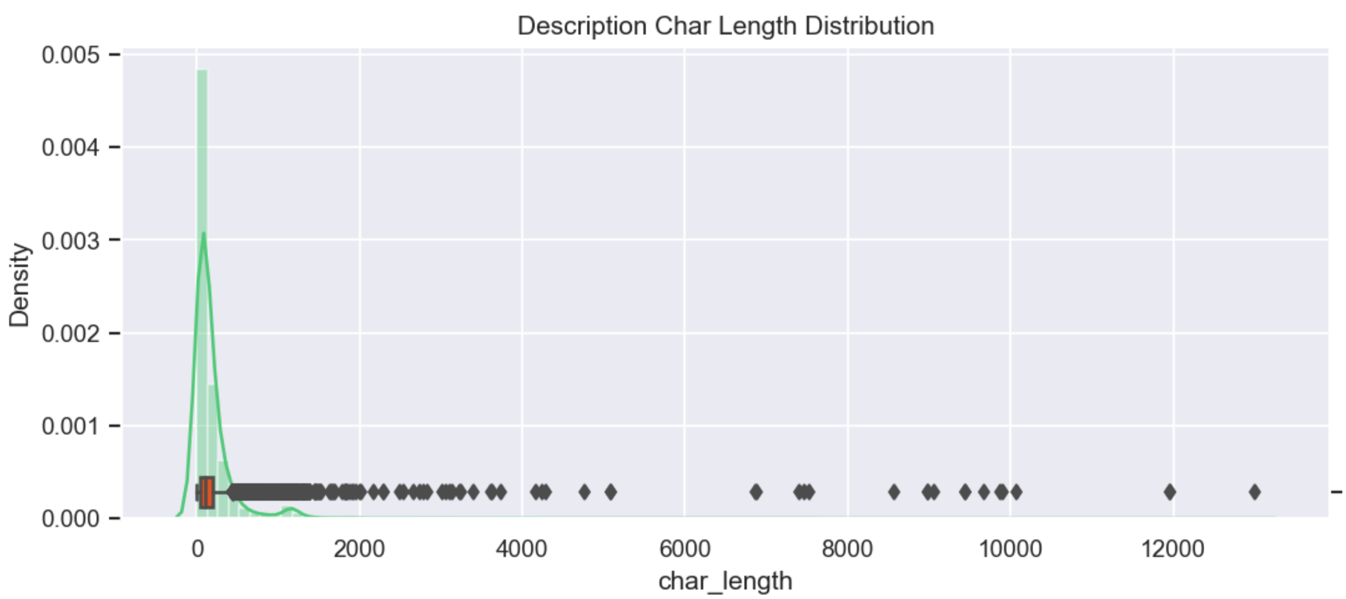
```
# !pip -q install yake
import yake

language = "en"
max_ngram_size = 5
duplication_threshold = 0.9
numOfKeywords = 1

custom_kw_extractor = yake.KeywordExtractor(lan=language,
                                            n=max_ngram_size,
                                            dedupLim=duplication_threshold,
                                            top=numOfKeywords,
                                            features=None)
```

```
k = custom_kw_extractor.extract_keywords(test)
k[0][0]
```

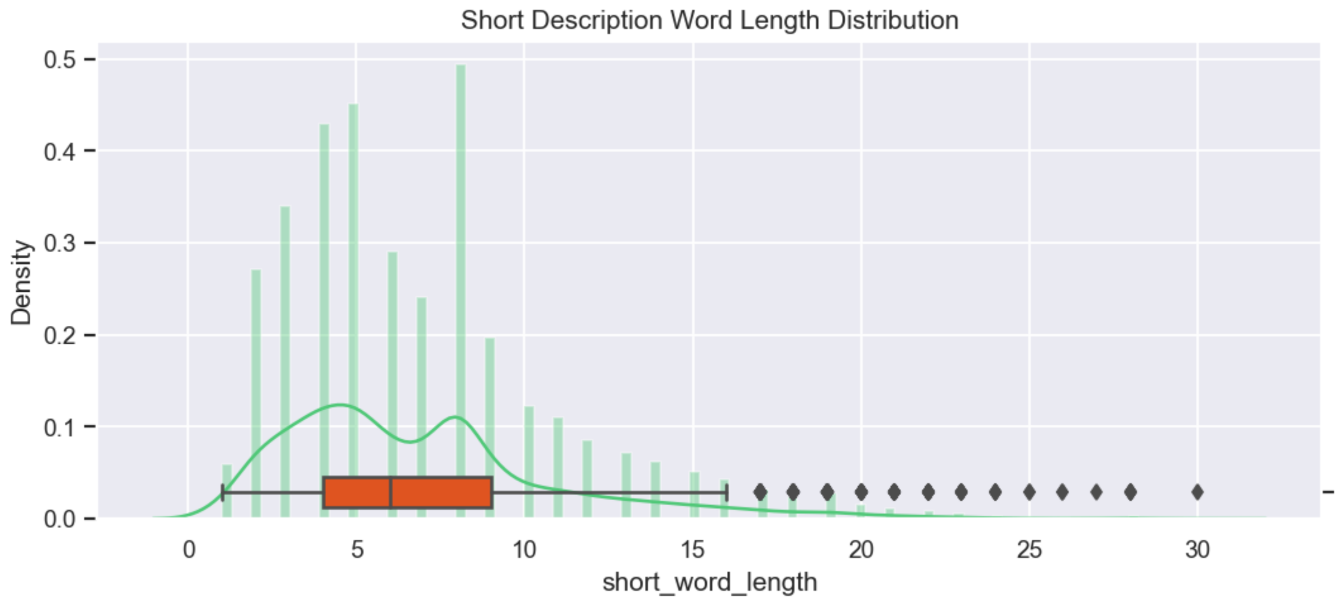
```
'south africa mpls circuit'
```

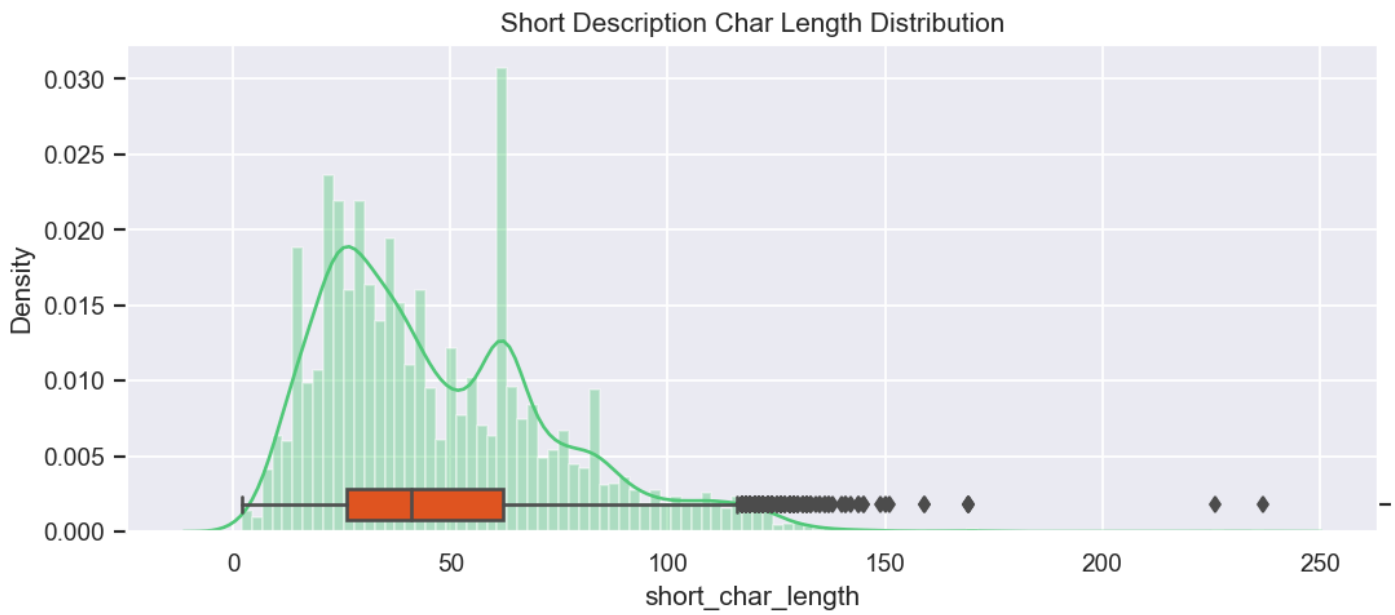
2.2.5. Description Word and Character Counts distribution





2.2.6. Short Description Word and Character Counts Distribution





- Most descriptions have between 6 and 28 words long with the median at 41 (106 characters) and the mean at 27.2 with relatively few outliers ranging to 1625 words.
- Most Short descriptions have between 4 and 9 words long with the median at 6 (41 characters) and the mean at 6.92 with relatively few outliers ranging to 28 words.

2.2.7. Inconsistencies & Discrepancies found during EDA

Clean up the unwanted information from initial observations. Imputing the dataset which has no data, one and two-word length by their corresponding short description.

- No word length ⇒ Imputed the description with the corresponding short description

	short_description	description	caller	group	char_length	word_length	short_char_length	short_word_length
6371	authorization add/delete members	\r\n\r\n	hpmwliog kqtnfvrl	GRP_0	5	0	33	3
7397	browser issue :	\r\n	fgejnhux fnkymoht	GRP_0	2	0	16	3

- One word length ⇒ Drop the row as the descriptions have no discernible information

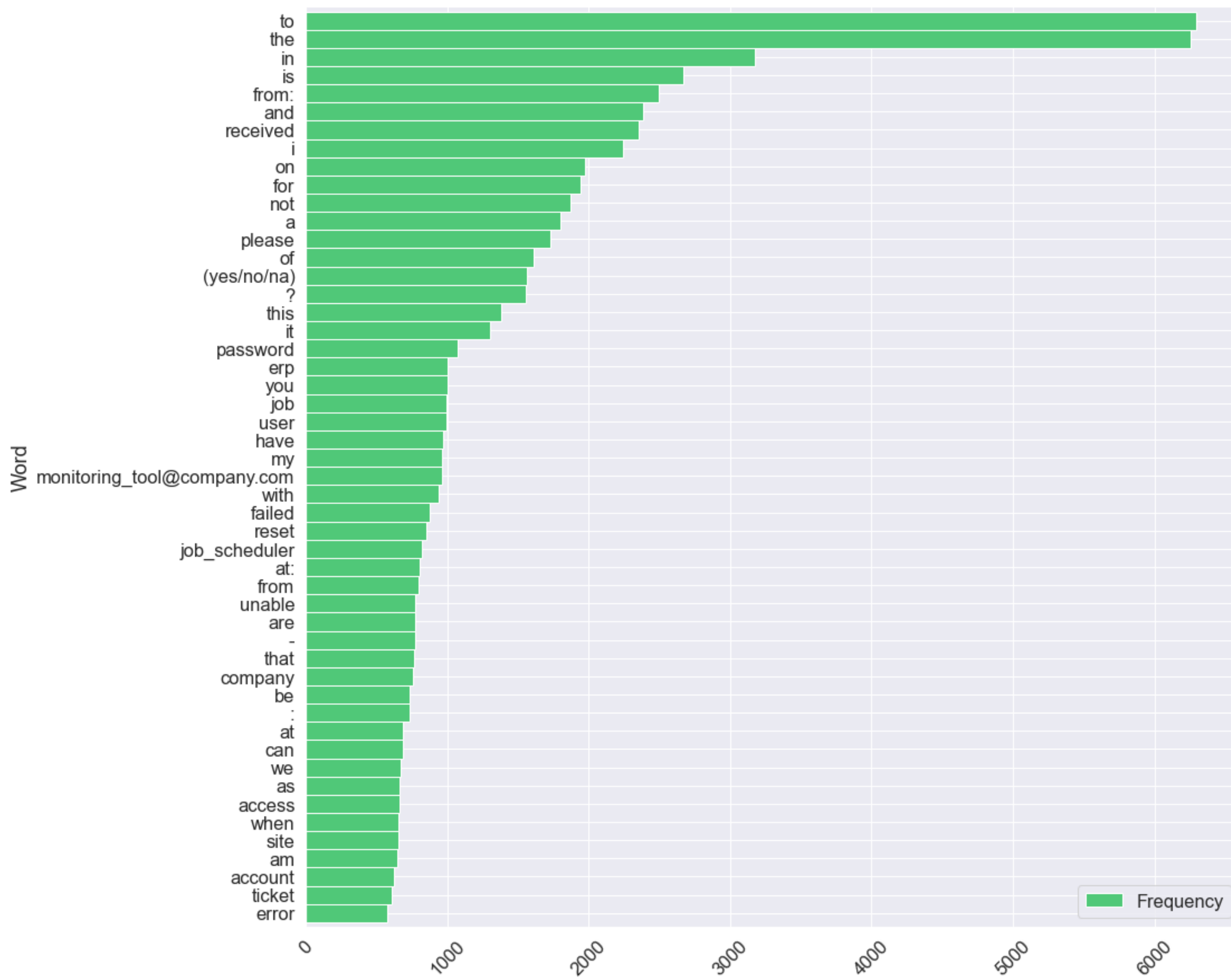
	short_description	description	caller	group	char_length	word_length	short_char_length	short_word_length
1860	s	s	gzjtweph mnslwfqv	GRP_0	1	1	1	1

- Fix the encoding

	short_description	description	caller	group	char_length	word_length	short_char_length	short_word_length
6106	ç”µè„‘ä¸ èƒ½å¼€æœº	æ—©ä¸Šä¸Šç ç”µè„‘æ‰“ä¸ å¼€ã€,	mzerdtop xnlytczj	GRP_30	30	1	18	1
276	outlookæ”¶åˆ°ç®±ä¸folderå ˜ä¸ºæ¯ å¤©ä¸€ä¸ªfol...	outlookæ”¶åˆ°ç®±ä¸folderå ˜ä¸ºæ¯ å¤©ä¸€ä¸ªfol...	bxfdkiol mdqlszvc	GRP_30	73	1	73	1

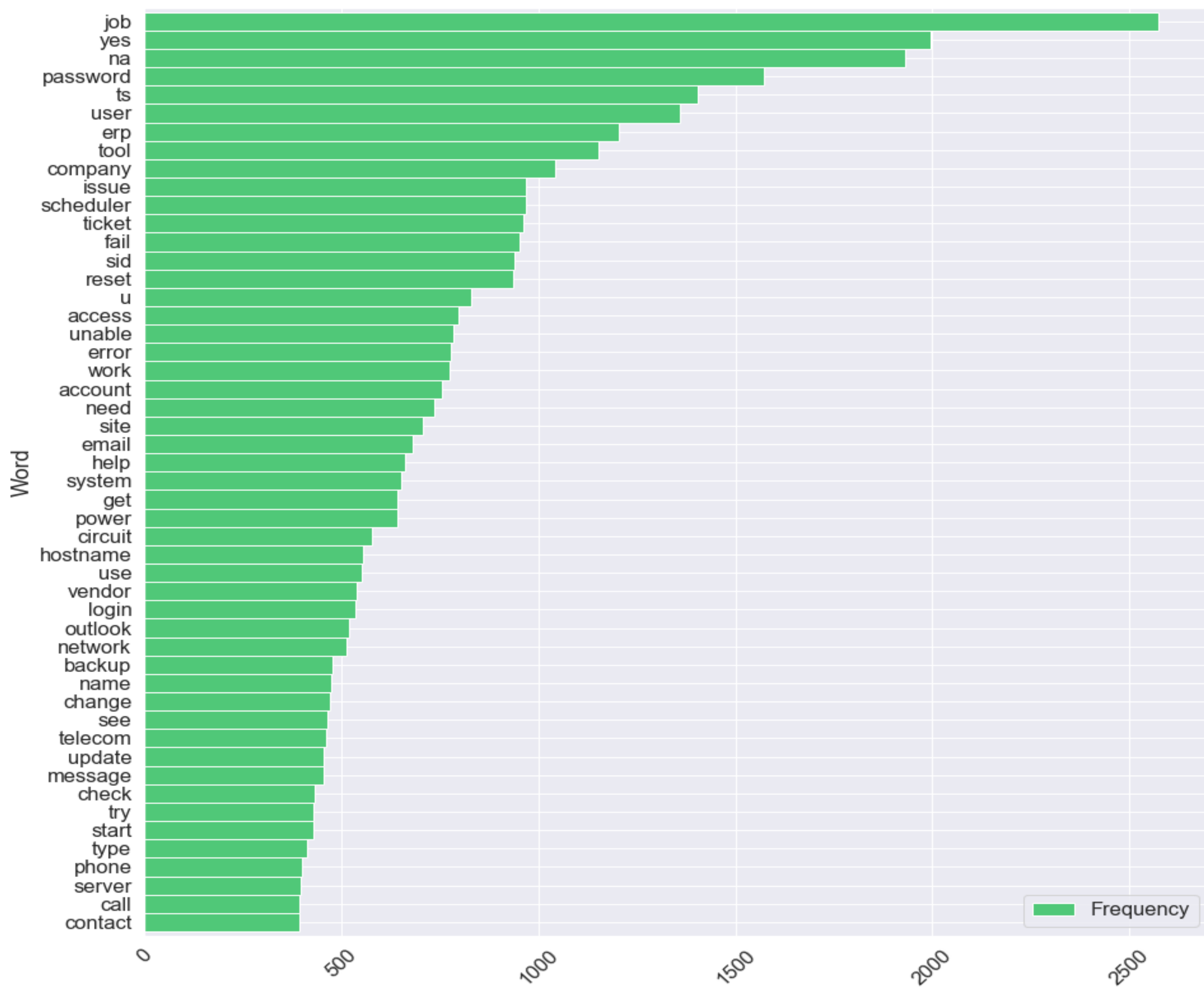
Description column and cleaned up data is generated for further analysis.

2.2.8(a) Word Frequency Distributions



- We have observed that words like "to", "the", "in".. etc., are occurring most frequently in the descriptions. These words will not add any predictive power to the models and will need to be removed as part of the stopword removal process during data pre-processing.
- Also, anchor words like "from:" and "received" and email addresses, punctuations, numbers are also occurring relatively frequently. We will remove these as well as part of the pre-processing.

2.2.8(b) Word Frequency Distribution after Data Cleaning



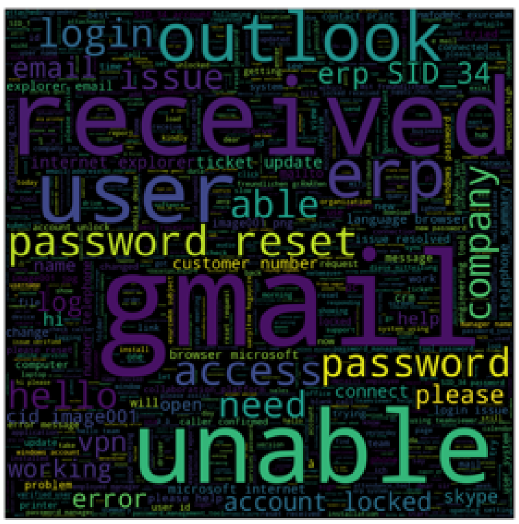
- The distribution of words is shown above after the data cleaning including the removal of stopwords, anchor words, numeric tokens, extra punctuation.
- Indicative words like "job", "password", "user" and "issue" are among the most frequent words.

2.2.9. Analysis using Word Clouds

- Descriptions WordCloud

- Short Descriptions WordCloud

- Group 0 WordCloud



- Other Groups WordCloud

- WordCloud is a data visualization technique used for representing text data in which the size of each word indicates its frequency or relative importance within the dataset.
- Significant textual data points can be highlighted using a word cloud. WordClouds have been generated with All available words & top 50 words.
- We have also inferred a few observations over the target class – Assignment groups with word clouds for the top 50 words from each group.

2.2.10. Cleaning Security Logs

```
# example string of a security log
test_string = '''
source ip: 10.16.90.249
source hostname: android-ba50a4497de455a
source port: 55198
source mac address: 50:2e:5c:f0:f6:98
system name :
user  name:
location :
sep , sms status :
field sales user ( yes / no) :
dsw event log:
-------------------------------------------------------------------------------------------------------------
=========================
event data
=========================
related events:
event id: 84682727
event summary: internal outbreak for 137/udp
occurrence count: 505
event count: 1

host and connection information
source ip: 10.16.90.249
source hostname: android-ba50a4497de455a
source port: 55198
source mac address: 50:2e:5c:f0:f6:98
destination hostname: [no entry]
destination port: 137
connection directionality: internal
protocol: udp

device information
device ip: 80.71.06.702
device name: company-european-asa.company.com-1
log time: 2016-09-26 at 08:23:55 utc
action: blocked
cvss score: -1

scwx event processing information
sherlock rule id (sle): 537074
inspector rule id: 186739
inspector event id: 639601949
ontology id: 200020003203009162
event type id: 200020003203009062
agent id: 103761
event detail:
sep 26 08:23:55 80.71.06.702 %asa-4-106023: deny udp src inside:10.16.90.249/55198 dst noris:100.74.211.4/137 by access-group "acl_inside" [0x30e3d92a, 0x0]

[correlation_data]
sep 26 04:23:54 60.43.89.120 dhcpd[23598]: dhcpack on 10.16.90.249 to 50:2e:5c:f0:f6:98 (android-ba50a4497de455a) via eth2 relay 10.16.88.2 lease-duration
691200 (renew)
sep 26 08:23:55 80.71.06.702 %asa-4-106023: deny udp src inside:10.16.90.249/55198 dst noris:100.74.211.1/137 by access-group "acl_inside" [0x30e3d92a, 0x0]

[correlation_data]
sep 26 04:23:54 60.43.89.120 dhcpd[23598]: dhcpack on 10.16.90.249 to 50:2e:5c:f0:f6:98 (android-ba50a4497de455a) via eth2 relay 10.16.88.2 lease-duration
691200 (renew)
sep 26 08:23:55 80.71.06.702 %asa-4-106023: deny udp src inside:10.16.90.249/55198 dst noris:100.74.211.12/137 by access-group "acl_inside" [0x30e3d92a, 0x0]

[correlation_data]
sep 26 04:23:54 60.43.89.120 dhcpd[23598]: dhcpack on 10.16.90.249 to 50:2e:5c:f0:f6:98 (android-ba50a4497de455a) via eth2 relay 10.16.88.2 lease-duration
691200 (renew)
sep 26 08:23:55 80.71.06.702 %asa-4-106023: deny udp src inside:10.16.90.249/55198 dst noris:100.74.211.14/137 by access-group "acl_inside" [0x30e3d92a, 0x0]

[correlation_data]
sep 26 04:23:54 60.43.89.120 dhcpd[23598]: dhcpack on 10.16.90.249 to 50:2e:5c:f0:f6:98 (android-ba50a4497de455a) via eth2 relay 10.16.88.2 lease-duration
691200 (renew)

ascii packet(s):
[no entry]
hex packet(s):
[no entry]'''
```

```
> clean_sec_logs(test_string)

>after initial cleanup:

cleaipsource  androidbae  portsource  addressffsystem name  user  name location  sep  sms status  field sales user      dsw event log   event data  related
event idevent summary internal outbreak forudp occurrence countevent count host and connection information  ipsource  androidbae  portsource
addressffdestination  no entry destination portconnection directionality internal protocol udp  device information device ipdevice name
companyeuropeanasa.company.comlog timeatutc action blocked cvss score   scwx event processing information sherlock rule id sleinspector rule idinspector event
idontology idevent type idagent id event detail sepasa deny udp src insidedst norisby accessgroup acl_inside   correlation_data sepdhcpd dhcpack
ontoffandroidbae via ethrelayleasedurationrenew sepasa deny udp src insidedst norisby accessgroup acl_inside   correlation_data sepdhcpd dhcpack
ontoffandroidbae via ethrelayleasedurationrenew sepasa deny udp src insidedst norisby accessgroup acl_inside   correlation_data sepdhcpd dhcpack
ontoffandroidbae via ethrelayleasedurationrenew sepasa deny udp src insidedst norisby accessgroup acl_inside   correlation_data sepdhcpd dhcpack
ontoffandroidbae via ethrelayleasedurationrenew sepasa deny udp src insidedst norisby accessgroup acl_inside   correlation_data sepdhcpd dhcpack
ontoffandroidbae via ethrelayleasedurationrenew sepasa deny udp src insidedst norisby accessgroup acl_inside   correlation_data sepdhcpd dhcpack
ontoffandroidbae via ethrelayleasedurationrenew sepasa deny udp src insidedst norisby accessgroup acl_inside   correlation_data sepdhcpd dhcpack
ontoffandroidbae via ethrelayleasedurationrenew sepasa deny udp src insidedst norisby accessgroup acl_inside   correlation_data sepdhcpd dhcpack
ontoffandroidbae via ethrelayleasedurationrenew sepasa deny udp src insidedst norisby accessgroup acl_inside   correlation_data sepdhcpd dhcpack
ontoffandroidbae via ethrelayleasedurationrenew sepasa deny udp src insidedst norisby accessgroup acl_inside   correlation_data sepdhcpd dhcpack
ontoffandroidbae via ethrelayleasedurationrenew sepasa deny udp src insidedst norisby accessgroup acl_inside   correlation_data sepdhcpd dhcpack
ontoffandroidbae via ethrelayleasedurationrenew sepasa deny udp src insidedst norisby accessgroup acl_inside   correlation_data sepdhcpd dhcpack
ontoffandroidbae via ethrelayleasedurationrenew sepasa deny udp src insidedst norisby accessgroup acl_inside   correlation_data sepdhcpd dhcpack
ontoffandroidbae via ethrelayleasedurationrenew sepasa deny udp src insidedst norisby accessgroup acl_inside   correlation_data sepdhcpd dhcpack
ontoffandroidbae via ethrelayleasedurationrenew sepasa deny udp src insidedst norisby accessgroup acl_inside   correlation_data sepdhcpd dhcpack
ontoffandroidbae via ethrelayleasedurationrenew sepasa deny udp src insidedst norisby accessgroup acl_inside   correlation_data sepdhcpd dhcpack
ontoffandroidbae via ethrelayleasedurationrenew  ascii packets no entry  hex packets no entry

>after duplicates removal (final cleaned up string):

ipsource androidbae portsource addressffsystem name user location sep sms status field sales dsw event log data related idevent summary internal outbreak forudp
occurrence countevent count host and connection information addressffdestination no entry destination portconnection directionality protocol udp device ipdevice
companyeuropeanasa.company.comlog timeatutc action blocked cvss score scwx processing sherlock rule id sleinspector idinspector idontology type idagent detail
sepasa deny src insidedst norisby accessgroup acl_inside correlation_data sepdhcpd dhcpack ontoffandroidbae via ethrelayleasedurationrenew ascii packets hex"""
```

- In the security log cleanup, we have removed ip addresses, special characters, extra whitespace in the event data descriptions and the duplicate entries
- This function is specific to the security/event logs present in the dataset which start with a specific pattern

2.3 Key Insights and Takeaways

- 74 Assignment groups found - Target classes
- ***Group 0*** is the majority class which accounts for ~47% of the data and the remaining groups are relatively much less frequent resulting in ***highly imbalanced data***.
- Around 17% of the descriptions were found to be in ***Non-English languages***
- Several Emails were found in the description
- Some descriptions have entire security/event logs
- Symbols & other non-ascii characters were detected in the description
- Hyperlinks, URLs, Email Addresses, Telephone Numbers & other irrelevant information was found in the descriptions
- Blanks found either in the short description or description field
- Few descriptions same as the short description
- Few words were combined together

- Spelling mistakes and typos were found in the data
- Contraction words found in the merged Description and expanded for ease of word modelling

2.4 Final Pre-Processing Techniques applied

```
print(test)
```

```
received from: tbvpkjoh.wnxzhqoa@gmail.com

i need access to the following path.  please see pmgzjikq potmrkxy for approval.



tbvpkjoh wnxzhqoa
company usa plant controller
tbvpkjoh.wnxzhqoa@gmail.com<tbvpkjoh.wnxzhqoa@gmail.com>

ticket update on inplant_872683
unable to login to collaboration_platform // password reset
all my calls to my ip phone are going to warehouse_toolmail, it is not even ringing.
sales area selection on opportunities not filtering to those in which the account
```

```
%%time

cleaned = preprocess_text(test)
pprint(cleaned, compact=True)
```

```
('need access follow path see pmgzjikq potmrkxy approval company usa plant '
 'controller ticket update inplant 872683 unable login collaboration platform '
 'password reset call ip phone go warehouse toolmail even ring sale area '
 'selection opportunity filtering wch account')
Wall time: 49.5 ms
```

Below steps have been performed for initial pre-processing and clean-up of data in the preprocess_text function:

- Fix text encoding using ftfy.fix_text — A lot of text in the data was being misinterpreted as some gibberish text (æ‰“å¼€ outlook) when in reality they were Chinese characters (打开 outlook)
- Parse email messages to retain only subject and body — Parse the mails to strip out headers, salutations, attachments etc., to retain only the relevant message.
- Clean up emails, links, website links, telephone numbers — Strip out any of this unnecessary information using regex patterns.
- Clean up anchor words like: 'Received from:', 'name:', 'hello', 'hello team', 'cid'... etc., — Strip out any of these filler words which add no information to the model

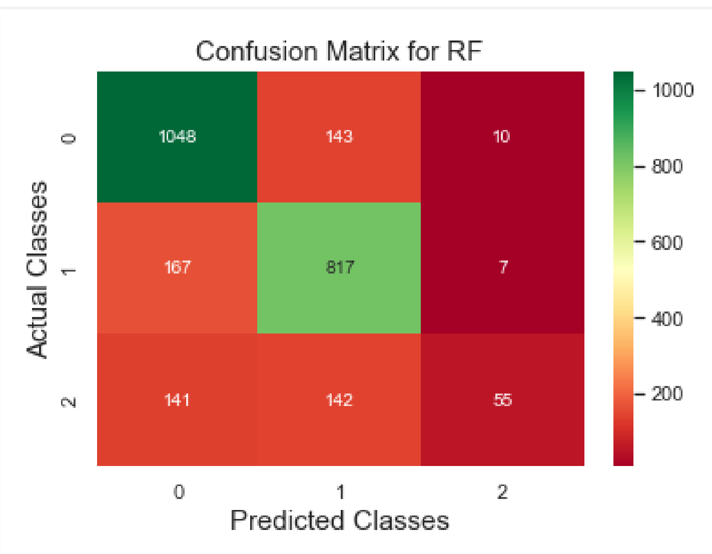
- Clean up security logs — Clean the logs in the data by removing unnecessary information using regex patterns
- Clean HTML tags wherever they exist in the data
- Clean Blank (/r /n) characters
- Strip caller names in descriptions — Caller names were found to be present in the descriptions as well, these values were tokenized and stripped out if found in the descriptions
- Translate/Normalize accented characters (á -> a)
- Convert Unicode characters to Ascii
- Expand contractions (they're -> they are)
- Clean stopwords & a few custom stopwords were found by analyzing the text
- Clean up extra whitespaces between words & Tokenize
- Remove gibberish — A lot of gibberish was still found to be in the text, this was stripped out using regex patterns
- Remove extra punctuation
- Changed the case sensitivity of words to lowercase
- Lemmatize the tokens in the final string
- Replaced Null values in Short description & description with space

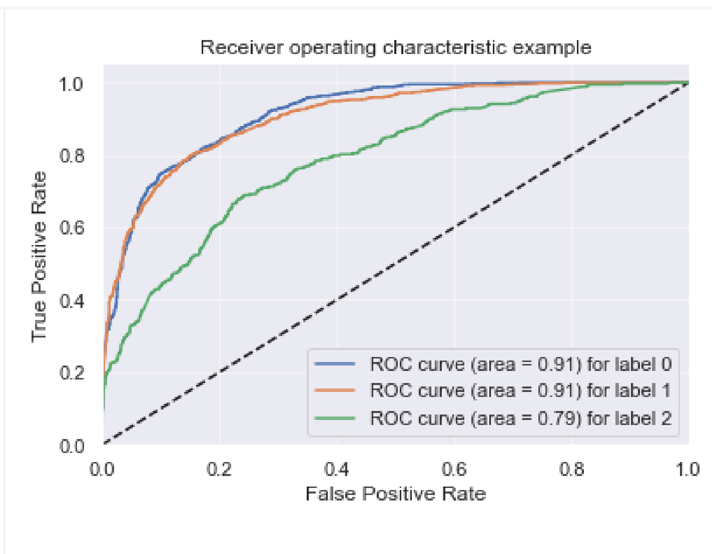
3. Model Building

3.1. Machine Learning Models

- The machine learning models were built based on random search cross-validation, two separate data sets were used for model building
- The training dataset for the ML models consisted of 3 target classes: ***L1, L2 and L3,*** these groups were collapsed based on the frequency of tickets received by each group.
- Metrics for the models on the first iteration:

Logistic Regression	83.07%
K-Neighbors Classifier	78.53%
Gaussian Naïve Biased	74.84%
Support Vector Machine	79.74%
Decision Tree	71.46%
Random Forest	**85.57%**
Gradient Boosting Classifier	83.96%
XGB Classifier	**85.17%**
Light GBM	**85.07%**

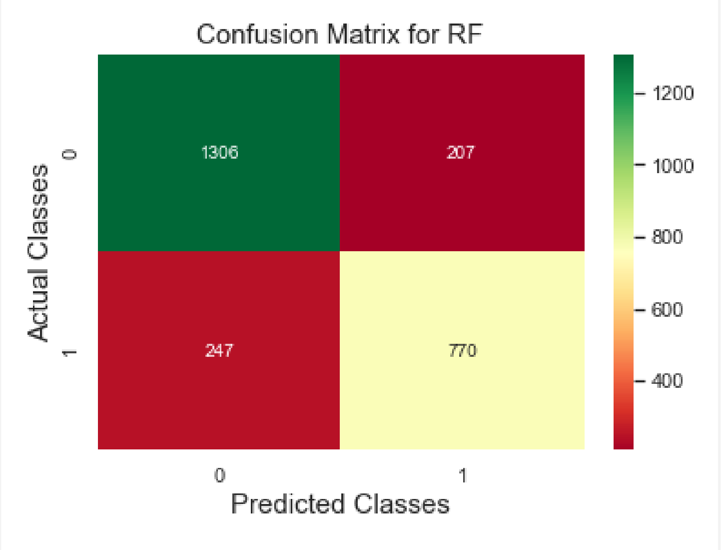


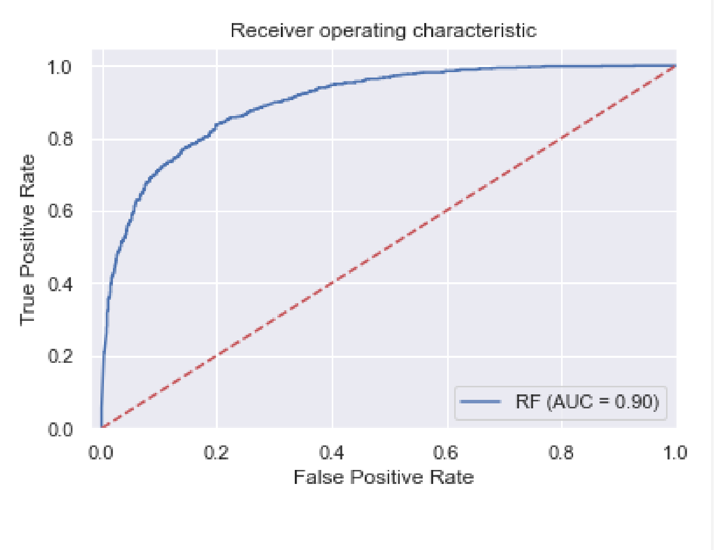


- The second iteration of ML models were built by further collapsing the 3 groups into 2 as there was a bit of an overlap with the groups L1 and L3 which later merged into ***L13 and L2.***
- Metrics for the models on the second iteration:

Logistic Regression	88.45%
K-Neighbors Classifier	84.70%
Gaussian Naïve Biased	84.78%
Support Vector Machine	86.35%
Decision Tree	76.12%
Random Forest	**90.49%**
Gradient Boosting Classifier	88.71%
XGB Classifier	89.99%
Light GBM	**90.20%**

During model building with 3 target classes and with 2 target classes the best performing model turned out to be Random forest based on the test accuracy as well as precision, recall, ROC, and the F1 score. However, the data needs to be further pre-processed to allocate tickets to the right groups based on keywords manually and run the models again.





3.2. Deep Learning Models

- The deep learning models were built based on two separate data sets with different grouping of the target classes.
- The training dataset for the first iteration models consisted of all target classes as they exist in the dataset without any further treatment.
- Metrics for the models on the first iteration:

Simple Feed-Forward Neural Net	60.40
Feed-Forward NN + Batch Norm	63.43
Feed-Forward NN + Dropout	64.73
Feed-Forward Nn + Pre-trained GloVe embeddings	61.53
LSTM	*49.91*
Bi-Directional LSTM	**65.87**
Convolution + MaxPool Blocks (Dimensionality Reduction) + LSTM	54.71
Convolution + MaxPool Blocks (Dimensionality Reduction) + Bi-LSTM	59.87
TfIdf Vectorization + Feature Selection + Feed-Forward Neural Net	**66.80**

- The training dataset for the second iteration models consisted of groups: **GRP0/Other**
- During the experimentation we have empirically found that this split works better than the **L13/L2** split for deep learning-based models.
- Metrics for the models on the first iteration:

Simple Feed-Forward Neural Net	**86.25%**
Feed-Forward NN + Batch Norm	83.76%
Feed-Forward NN + Dropout	85.83%
Feed-Forward Nn + Pre-trained GloVe embeddings	82.75%
LSTM	*65.44%*
Bi-Directional LSTM	**85.24%**
Convolution + MaxPool Blocks (Dimensionality Reduction) + LSTM	84.47%
Convolution + MaxPool Blocks (Dimensionality Reduction) + Bi-LSTM	85.00%
TfIdf Vectorization + Feature Selection + Feed-Forward Neural Net	85.77%
Stratifiedkfold Validation +TfIdf Vectorization + Feature Selection + Feed-Forward Neural Net	**86.40%**

- The custom model with TF-IDF vectorization and feature selection approach worked best for the current iteration of training. Here, we are processing more contextual information as compared to the generalized nature of word embeddings and through Feature Selection, we are able to capture most important features and avoid noise going into the model.
- We have observed that the Bidirectional LSTM model is performing much better relative to the LSTM model. One reason behind this could be that the bidirectional model takes into account the past and the future context of a sequence and is hence more robust in dealing with the noise that might be biasing the vanilla LSTM model and also understanding the context of the words in a description.

4. Key Learnings & Further Improvements

4.1. Learnings:

- The dataset is highly imbalanced which could be an inherent limitation that affects the performance of our classification model.
- The presence of other languages in the dataset is an inherent limitation within the dataset which will limit the language models to learn properly. This is further complicated by the fact that we may receive similar foreign language descriptions in out-of-sample data which have to be translated first adding to the latency of the inference pipeline.
- Even though we found that the caller column had significant correlation with the target column by performing a $\chi 2$ test for two distributions. We have decided to drop the caller column as an input feature to the models as considered in the tradeoff:
    - New queries from an old caller could give some prior probabilities/info to the model.
    - But it wouldn't help on new callers in out-of-sample data as the caller doesn't really indicate what their ticket or issue is So, Real-world performance will degrade.
- In view of this tradeoff, we have dropped the caller id column as a feature.

4.2. Planned Improvements for Milestone-II:

- Tune the hyper-parameters of various models using a coarse to fine search methodology by searching randomly followed by a grid search in a tighter range.
- Translate Foreign Languages using a custom model or Google Translate API
- Find and remove custom stopwords from the text
- Oversampling of minority classes with text augmentation to handle the imbalance in the data
- Stacked model on top of the binary classifier which further classifies among the other groups
- Use other pre-trained embeddings like word2vec, BERT, ELMO, Flair Word Embeddings to capture the context/meaning of the natural language words
- Clustering of groups using unsupervised clustering or with manual intervention to do effective sub-grouping of the target classes
- Transfer Learning using some pre-trained NLP models to utilize the language understanding of these models and fine-tune them for ticket classification
- Add Attention Layer to the model architecture to capture context and significant keywords in the descriptions better
- Train transformer based model architectures (BERT)
- Productionize the model by deploying locally or on cloud