**1.Summary of problem statement, data and findings**

**\*\*\*need to delete(Every good abstract describes briefly what was intended at the outset, and summarizes findings and implications.)\*\*\***

ATA is a classification problem which comes under the Supervised Machine Learning category and plays a key role for successfully running any system. Especially in a very large system that provides numerous services and each service has multiple categories and subcategories. Manually tagging tasks to specific categories and subcategories required user training, manpower and also prone to human error that can impact overall service delivery. ATA uses machine learning techniques to assign tasks to appropriate groups automatically that can improve overall turnaround time of service delivery.

This saves First Response Time(FRT) and Average Response Time(ART) from shooting up. So as you can see, auto-assignment is a smart way to improve your support team's efficiency. With Automatic Ticket Assignment, you can make your support agents accountable for a ticket and reduce the overall wait time of the customer.

ATA also helps in avoiding SLA breach and can help companies save penalties and helps in improving TAT

While the manual assignment is time consuming and ineffective, it also incurs cost that arises from wrong assignment of tickets. Routing tickets manually makes it difficult to track the availability of the right technicians for responding to issues. Tracking down the end user is also problematic and the time consumed in this makes only less time available for actual problem resolution.

**2. Summary of the Approach to Pre-processing/Data cleansing and EDA**

**\*\*\*need to delete(Include any insightful visualization you have teased out of the data. If you’ve identified particularly meaningful features, interactions or summary data, share them and explain what you noticed. Visual displays are powerful when used well, so think carefully about what information the display conveys.)\*\*\*\***

Below analysis can be performed on dataset ->

1. Distribution of tickets among Group/assignment group/department in form of histogram.
2. Distribution of Tickets among Category/Sub categories.
3. Incident count per assignment group.
4. Repetitive ticket count per assignment group.
5. Most frequent keyword distribution in each assignment group.
6. Most frequent keywords distribution across assignment groups.
7. Any time frame in which most of the incident is coming.
8. Average resolution time/turn around of incidents.

Brief on data sets -

1. There are 8500 rows with 4 columns in the dataset.
2. We can remove the 'caller' column and retain both "description" and "Short description" independent variables.
3. There are 9 rows with null values that can be removed.
4. There are 1717 duplicate records. These duplicates are important because users may have faced the same issue multiple times, it will help us to understand the pain area of the system and also provide automation opportunities to fix repetitive issues.
5. Below are the most repetitive tickets and mostly these tickets belong to GRP\_0 group.

windows password reset 28

password reset 25

windows account locked 22

account locked in ad 22

erp SID\_34 account unlock 17

account unlock 15

account locked. 14

erp SID\_34 account locked 14

blank call 14

unable to connect to vpn 14

unable to launch outlook 13

erp SID\_34 account locked 13

erp SID\_34 password reset 13

erp SID\_34 account unlock and password reset 13

erp SID\_34 password reset. 12

1. After removing duplicates, performed 'Assignment group' distribution. There are 74 groups. and these are imbalanced. Most of the group has just 1 record.
2. We will have to perform some clustering techniques to group these 74 groups to category and sub category. It is quite impractical to have 74 categories for any application.
3. Analyzed tokens for Description column . Below are the observations.  
   \*\*\* top 10 most frequent token  
   com 2826  
   gmail 2139  
   yes 2013  
   please 1938  
   na 1892  
   company 1599  
   received 1578  
   user 1291  
   tool 1162  
   e 1139  
   \*\*\*Single character tokens need to be removed.  
   \*\*\* Top 10 tokens are general email terms that need to be removed.  
   \*\*\* Create custom list for token to removed ((com, job, received, gmail, yes, please, na,tool)
4. Analyze description column for individual group (GRP\_0), found similar token as above , so we are good to remove these custom token list.
5. Analyzing short description token for individual group (GRP\_0)- below are the most frequent tokens.  
   password 598  
   erp 467  
   unable 408  
   reset 315  
   issue 295  
   tool 293  
   outlook 290  
   account 220  
   access 205  
   login 197  
   sid 178  
   user 172

## **\*\*\* these token are very clean, don't need any token cleanup**

## **\*\*\* from above token we can see in group 0, most of the issues are related to login and access**

## **\*\*\* Short description is much clearer than the description column.**

## **\*\*\* We can also check ngram tokens for a short description.**

**3. Deciding Models and Model Building**

Based on the nature of the problem, decide what algorithms will be suitable and why?

Experiment with different algorithms and get the performance of each algorithm.

**4. How to improve your model performance?**

What are the approaches you can take to improve your model? Can you do some feature selection, data manipulation and model improvements.