G-SUITE METRICS

Anusha Vanama, Lavanya Goluguri, Advisor: Somya D. Mohanty

***Abstract*—G Suite is a set of cloud computing, productivity and collaboration tools, software and products developed by Google. G Suite comprises G-mail, Drive, Docs, Calendar, Hangouts Chat, Hangouts Meet, Mobile, Voice, Slides, Sites, Sheets, Keep and Tasks, Groups, Google+, Forms, Contacts, Cloud Search, Chrome Browser as some some of there metrics. Currently,the G-suite data that we are working on is from UNCG ITS department, the data for Google (G Suite) report metrics is pulled automatically into Splunk from Google on daily basis.Our main goal for this project is analyze the data, find anomalies in the data and predict the data for the major services in the G-suite metrics and develop a dash board to have better visualization of the G-suite data."Organizing 'Information Technology Services' data to make it useful, accessible, actionable and aligned with data-informed decision making ideals."**

I. Introduction

In order to achieve the goal of data prediction, finding anomalies with in the data and statistical analysis. We have first started with basic data analysis of understanding the data. In understanding the data we have looked at different attributes of the data set and the metrics that attribute belongs to. We have only considered the prominent attributes in the corresponding metrics. We have reconfigured the data set for each metric individually, so that it gets easier for querying and to have a better understanding of that metric.

We have performed data cleaning, where few observations with most missing values are dropped entirely, few of the observation with 1 or 2 missing values and significant information in the remaining features are replaced with mean values. There are also few dates on which the data is not being pulled into Splunk. These missing dates are not replaced in few cases while in other case they are replaced with the mean values.

Before going into further details of our project, the main metrics that we have focused on doing data analysis, statistics and predictions are Google drive, Gmail, Google Meet, Google calender and Google Accounts.

II. Metrics

GOOGLE DRIVE: Google drive reflect aggregated user behavior on Google Drive items in the domain. The items include details about the forms, sheets, presentations, drawings etc. Few of the metrics in drive are as shown below

Num_owned_items_with_visibility_*_added: The Docs or Drive items with at least one addition type event performed on them such as create, uploads, untrash events, or ownership transfer. This activity gets reported regardless of the final state of the item. Multiple addition events to the same file do not increment the count.

Num_owned_items_with_visibility_*_removed: The Docs or Drive items with at least one removal type event performed on them, such as trash, permanent delete, or ownership transfer. This activity gets reported regardless of the final state of the item.

Num_owned_items_with_visibility_*_delta: The net change in number of Google Docs and Drive items owned for the entire domain, segmented by user. For example, if one item gets added and another gets removed, the added and removed metrics each report 1, but the delta metric reports 0.

The metrics that are mainly focused on in the drive data are:

1day_active_users
30day_active_users
Google_forms_created
Google_forms_edited
Google_documents_created
Google_documents_edited
Google_presentations_created
Google_presentations_edited
Google_Spreadsheets_created
Google_spreadsheets_edited
Google_drawings_created
Google_drawings_edited
Google_forms_created_delta.
Google_documents_created_delta.

GMAIL: The metric Gmail gives about the information related to the number of email received, number of email emails sent, information about the spam email numbers. Few metrics of the Gmail are:

Is gmail enabled: the user's Gmail service is enabled
Num email sent : Number of emails sent by the user
Num spam emails received: Number of emails received by the user's marked as spam mail
Num emails received: Number of emails received by the user
Num emails exchanged : The total number of emails exchanged. This is the total of num emails sent and num emails received.The metrics we maiinly focused on gmail data are:

num_emails_received.
num_emails_sent
num_inbound_rejected_emails
num_inbound_spam_emails
num_inbound_non_spam_emails

GOOGLE MEET:The google meet data the hangout meet data, few of the metrics in google meet are:

Total call minutes: Gives the number of call minutes on that day.

Num 30day active users: Gives the number of active users for hangout meet for 30 days.
Num meetings: Number of meeting on that particular day.
Num calls: Number of hangout calls on that day
Average meeting minutes: Average number of meeting minutes on a given day.
Total meeting minutes: Total number of meeting minutes on a given day. The metrics mainly focused on the meet data are:

google.meet:total_call_minutes
google.meet:num_30day_active_users
google.meet:num_meetings
google.meet:num_calls
google.meet:average_meeting_minutes
google.meet:total_meeting_minutes

GOOGLE CALENDER:This metric gives information related to the calender events that are scheduled on a particular day
GOOGLE ACCOUNTS:This metric gives information related to the number accounts that are created, the size in mb used by each account. etc. The metrics mainly focused on the accounts are:
num_users
google.accounts:num_7day_logins
google.accounts:num_disabled_accounts
google.accounts:num_30day_logins
google.accounts:num_1day_logins
google.accounts:num_users_2sv_not_enforced
google.accounts:num_suspended_users
google.accounts:num_authorized_apps
google.accounts:drive_used_quota_in_mb
google.accounts:gmail_used_quota_in_mb
google.accounts:used_quota_in_mb
google.accounts:gplus_photos_used_quota_in_mb
google.accounts:total_quota_in_mb

III. Data description

Initially the data set that was given to us as below. The details of Data Set which we are using for this project:
• Size: 611,914 rows and 649 metric names.
• Date Range: March 23rd, 2015 to August 17th, 2019
• Fields: time, metric name, metric value
Time - The day for which the data is collected. (format: yy-mm-dd hr:min:sec:msec)
Metric Name - it is name of the google service used with their parameter.
Metric Value - It is a numeric value related to metric parameter, it can be no. of users the metric has on that day, no.of devices that the used the metric parameter, no. of. documents or it can be a size in mb. The data set is similar to the one shown below.
Later the data set has been divided based on the name of the metric, As the data set is reshaped we were able to created different data sets for different metrics, The data set size for Google drive data is 831 rows X 144 columns. The data set for

time	metric_name	metric_value
2015-03-25T00:00:00.000-0400	google.docs:num_7day_active_users	7468
2015-03-25T00:00:00.000-0400	google.docs:num_docs	1899985
2015-03-25T00:00:00.000-0400	google.docs:num_docs_edited	7178
2015-03-25T00:00:00.000-0400	google.docs:num_docs_externally_visible	84246
2015-03-25T00:00:00.000-0400	google.docs:num_docs_internally_visible	1815783
2015-03-25T00:00:00.000-0400	google.docs:num_docs_not_edited_for_12months	1031157
2015-03-25T00:00:00.000-0400	google.docs:num_docs_not_edited_for_3months	1613176
2015-03-25T00:00:00.000-0400	google.docs:num_docs_not_edited_for_6months	1388920
2015-03-25T00:00:00.000-0400	google.docs:num_docs_not_viewed_for_12months	810481
2015-03-25T00:00:00.000-0400	google.docs:num_docs_not_viewed_for_3months	1484887
2015-03-25T00:00:00.000-0400	google.docs:num_docs_not_viewed_for_6months	1203530
2015-03-25T00:00:00.000-0400	google.docs:num_docs_shared_outside_domain	44
2015-03-25T00:00:00.000-0400	google.docs:num_docs_viewed	14113
2015-03-25T00:00:00.000-0400	google.docs:num_docs_with_visibility_anyone_with_link	78631
2015-03-25T00:00:00.000-0400	google.docs:num_docs_with_visibility_people_at_domain	2878
2015-03-25T00:00:00.000-0400	google.docs:num_docs_with_visibility_people_at_domain_with_link	26493
2015-03-25T00:00:00.000-0400	google.docs:num_docs_with_visibility_private	1786412
2015-03-25T00:00:00.000-0400	google.docs:num_docs_with_visibility_public	5571
2015-03-25T00:00:00.000-0400	google.docs:num_drawings	2084

Fig. 1. G-suite original data set

Gmail data is 28 columns X 1545 row. The data for different metrics is collected form different time frames, For example drive data is collected from March 2017 to August 2019. Where as gmail data is present from March 2015 to August 2019. there are also few metrics that are no longer continued.

	time	1day_active_users	1day_google_documents_active_users	1day_google_drawings_active_users	1day_google_forms_active_users	1day_google_p
0	2017-03-12 05:00:00	1543.0	2499.954068	7.240157	54.104987	
1	2017-03-14 04:00:00	2495.0	2499.954068	7.240157	54.104987	
2	2017-03-16 04:00:00	2403.0	2499.954068	7.240157	54.104987	
3	2017-03-19 04:00:00	2848.0	2499.954068	7.240157	54.104987	
4	2017-03-20 04:00:00	5302.0	2499.954068	7.240157	54.104987	

Fig. 2. G-suite drive data set after reshaping the data frame

IV. Basic Data Exploration

A. Statistical evaluation of data

1) Google Meet: There are 55 unique metric names, 29434 observations under Google Meet recorded from January 14th, 2018 to August 17th, 2019. There are 17731 observations in 2018 and 11703 in 2019. Some of the Metric Values of Google Meet includes Number of users, Number of Minutes, Number of Meetings, Number of Calls, Usage of Chromebox and Usage of Chromebase.
Mean value is greater than median for total call minutes, num of calls, num of meetings, total meeting minutes. Maximum Outlier values are present for all these metrics as shown in (Fig. 3). Mean value is lower than median for number of 30day active users, average meeting minutes. Minimum outlier values are present for these metrics as shown in (Fig. 4). Variance and Standard Deviation are not the right measures of spread as our data is having outlier values Median absolute deviation is the right measure of spread for all the above metrics.

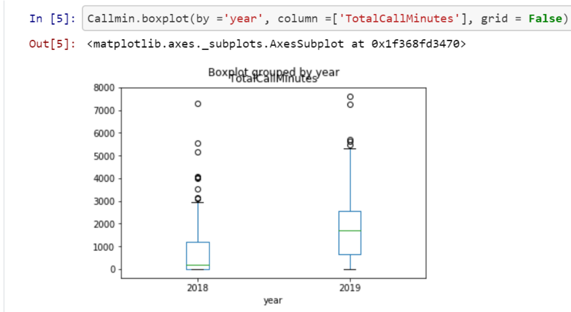

Fig. 3. Maximum Outlier Values for Total Call Minutes metric

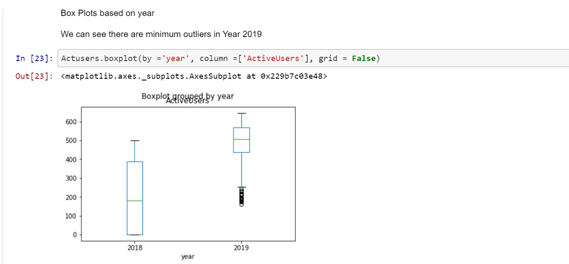

Fig. 4. Minimum Outlier Values for Total Call Minutes metric

Below are few more interesting observations under Google Meet

1) Maximum number of total call minutes is 7623 minutes recorded on 2019-07-01.
2) Maximum number of calls are 159 calls recorded on 2019-03-21.
3) Maximum number of meetings are 39 meetings recorded on 2019-03-21.
4) Observed that there are more number of calls and meetings are scheduled during fall advising period which was started on March 18th 2019.
5) Maximum number of total meeting minutes is 1358 minutes recorded on 2018-12-11.
6) Maximum number of 30 day active users are 646 recorded in the month of April (2019-04-16).
7) Maximum number of Average meeting minutes is 174 minutes recorded on 2018-12-09.

2) Google Calendar: There are 3 unique metric names, 4683 observations under Google Calendar. Metric values represents the number of 1 day active users, number of 7 day active users, number of 30 day active users.
Highest number of daily active users are recorded on May 4 2018 with 10979 daily active users as shown in (Fig. 5). It happened to be the day where UNCG celebrates largest graduating class at may commencement.

B. Point Estimates

1) Google Meet: Used Point estimates to estimate the mean of total call minutes for all days. Point estimate based on a sample of 400 observations overestimated the true call minutes mean by 22.560137 minutes. The distribution has high

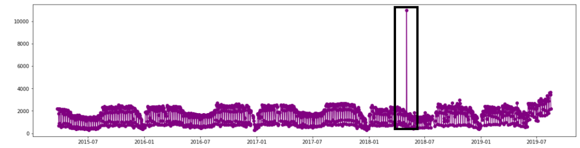

Fig. 5. Highest number of daily active users in Google Calendar

skewness, and the plot reveals the data is clearly not normal as shown in (Fig. 6). The sample drawn from this data is also having same shape as shown in (Fig. 7). This suggests that we can't apply techniques that assume a normal distribution to this data set, since it is not normal.
Created a sampling distribution by taking 300 samples from our call minutes data and then making 300 point estimates of the mean. The sampling distribution appeared to be normal after applying Central Limit Theorem as shown in (Fig. 8) and the mean of the sampling distribution approached the true mean, there is only a difference of 3 minutes.
A point estimate can give you a rough idea of a distribution parameter like the mean, but estimates are prone to error and taking multiple samples to get improved estimates may not be feasible. A confidence interval is a range of values above and below a point estimate that captures the true distribution parameter at some predetermined confidence level.
The confidence interval we calculated is very close the true call minutes mean of 1148.051565. 24 out of 25 Samples overlapped the red line marking the true mean in a 95 percent confidence interval as shown in (Fig. 9).

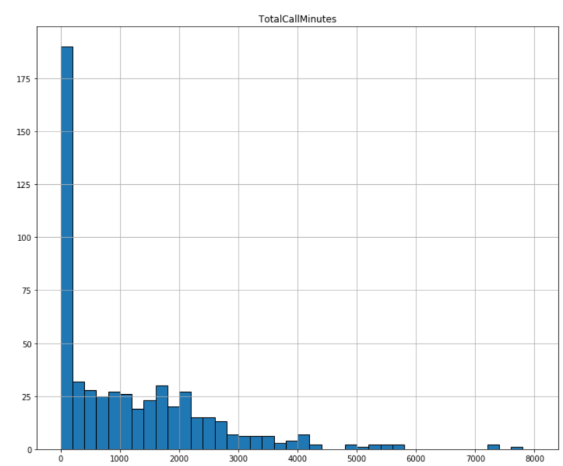

Fig. 6. Distribution of data in Total Call Minutes metric

C. Distribution Modeling

1) Google Meet: Observed that google meet application usage is very less in UNCG from Jan 2018 to June 2018. Metric values are 0 for most of the days. Eg: Number of meetings held using google meet is 0 for 159 days out of 543 days as shown in (Fig. 10).
Plotted Total Call Minutes data using Histograms to observe what sort of distribution fits the data. Total call minutes comes

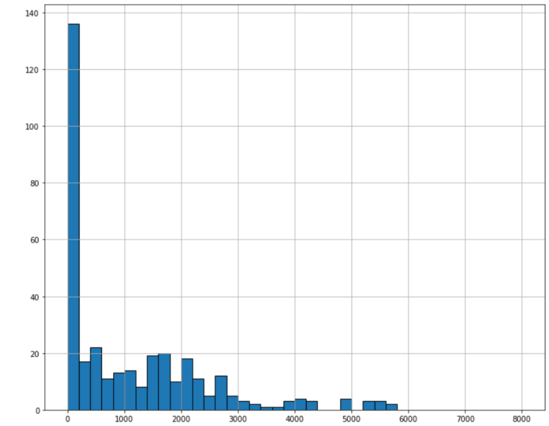

Fig. 7. Distribution of Sampled data in Total Call Minutes metric

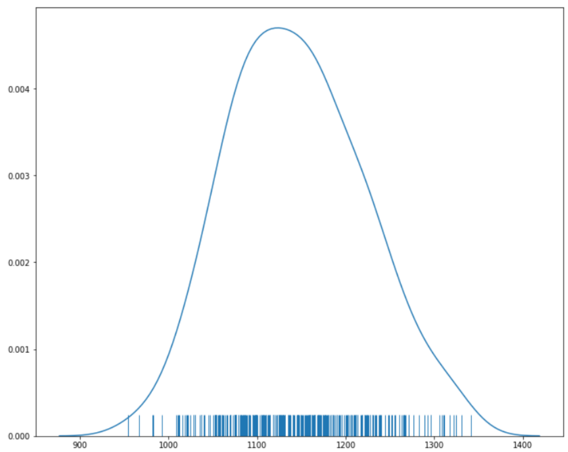

Fig. 8. Distribution after applying central limit theorem

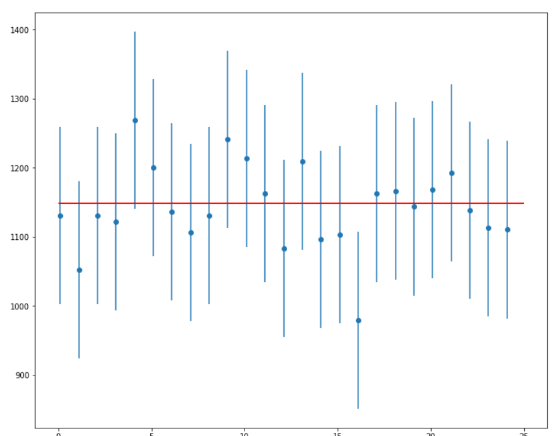

Fig. 9. Confidence Intervals of Total Call Minutes metric

under continuous data and it followed exponential distribution. The exponential started high and it has a long tail that trails off to the right as shown in (Fig. 11) . Kernel Density Estimator is used to find probability density function (PDF) for Total Call Minutes and Total meeting Minutes data.

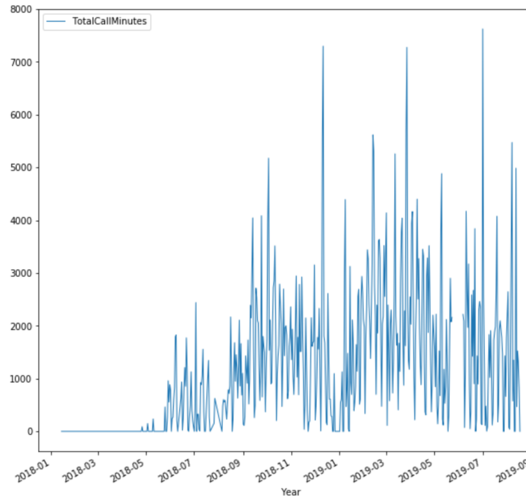

Fig. 10. Metric Values for Total Call Minutes metric

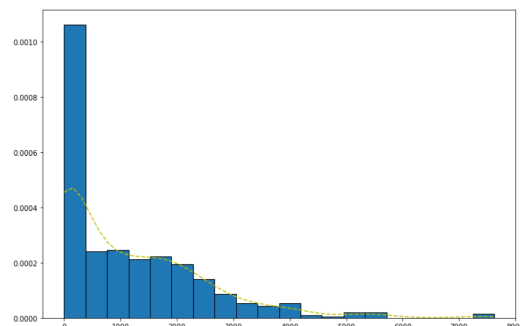

Fig. 11. Distribution Modeling for Total Call Minutes metric

2) Google Drive: Both the 1day active users and 30day active users data is discrete and both of them followed Poisson distribution as shown in (Fig. 12).

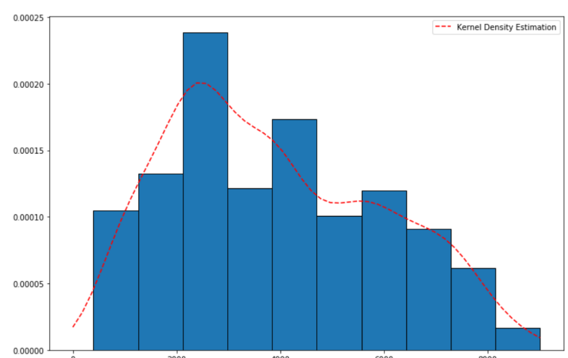

Fig. 12. Distribution Modeling for 1day active users metric in Google drive

3) Gmail: All the attributes in Gmail are discrete and the distribution that will work with the data is the Poisson distribution as shown in (Fig. 13), because it shows how many times an event likely to occur over specific time. Here analysis is done on how users are using a service in G-Suites, more specifically Gmail, over time. Performed ks-tests to find which estimator has the better fit. It has a good fit for majority of the attributes.

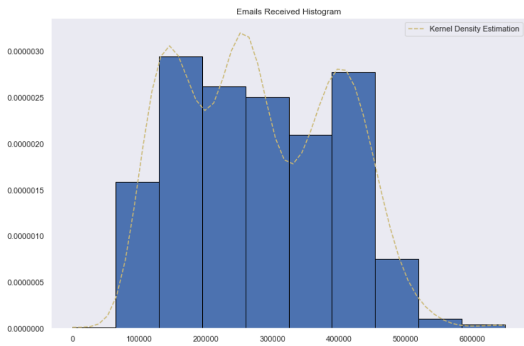

Fig. 13. Distribution Modeling for Emails Received metric in Gmail

4) Google Calendar: Here, daily active users followed bi modal distribution as shown in (Fig. 14). Largest peak is for spring/fall classes which has a higher attendance of students, and the smaller peak is for summer or winter semesters where there is a lower rate of students attending classes.

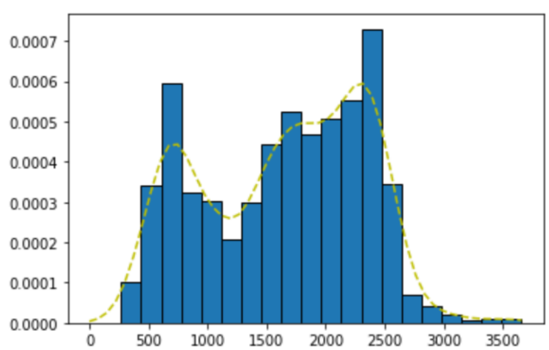

Fig. 14. Distribution Modeling for daily active users metric in google calendar

D. Hypothesis testing

One Sample t-test - A one-sample t-test checks whether a sample mean differs from the population mean.
Two Sample t-test - A two-sample t-test investigates whether the means of two independent data samples differ from one another.

1) Google Meet: Tested whether the mean of number of meetings for Google Meet in the month of March 2019 differs from the mean of Number of meetings for Google meet in Year 2019 using One-Sample t-test. Here, p-value is 0.17177339793916663 which is greater than 0.05. So, we fail to reject the null hypothesis.
Tested whether the mean of total call minutes for Google Meet in the year 2019 differs from year 2018 using Two Sample t-test Null Hypothesis is that both the means are similar and Alternate hypothesis is that they are different. Here, p value is 0.000000000000000000113 which is smaller than 0.05. So, we reject the null hypothesis.

2) Google Drive: Below are the Hypothesis testing results for various comparisons in Google Drive

- Active Users - Tested whether 1 day active users is similar to 30 day active users. Yes they are similar as the P value is 0.17.
- Forms - 2018 spring semester data is similar to 2019 spring semester data. 2018 spring semester data is similar to overall data as the P value is 0.4, but it is different from the 2019 spring semester data as the P value is 0.02.
- Documents - Tested whether documents created is similar to documents edited. Yes they are similar as the P value is 0.2.
- Spread Sheets - Tested whether April Month spreadsheets and May month spreadsheets data is similar. April data is similar to the overall data as the P value is 0.11. May data is different from the overall data and the April data as the p value is 0.001.
- Presentations - Tested whether 2017 presentations data is similar to 2018 presentations data. 2017 data is similar to the overall data , but 2018 data is different form the 2017 and overall data as well.

3) Gmail: Hypothesis is that the number of the 1 day active users in December is smaller when compared to active users for the overall year. One sample t test is performed with 1 day active users in December(sample) and the 1 day users mean (population). The p value is less than 0.05 which means that we reject the null hypothesis. Since the mean for December is 14764 and the mean for overall year is 17342, and the p value is much less than 0.025 it is concluded that month of December has less active users.

4) Google Calendar: Hypothesis - The third week of the month has a significant difference to the rest of the weeks.
T-test - Mean of third week compared to mean of rest of the weeks
Result - Here, p-value of 0.27 is greater than 0.05. So, we accept the null hypothesis and the conclusion is that the third week of month is not significantly different.

E. Correlation and Co-variance within the data

Correlation and Co-variance describe how two random variables are related. Co-variance measures the extent to which the relationship between two variables is linear.

1) Google Meet: Our mentor is interested in the comparison of the types of devices (iOS/Android etc). Tried finding Correlation between Total call minutes made using iOS, Total call minutes made using Android, Total call minutes made using all devices in UNCG.
Below are the results

- Correlation of Android and iOS: 0.08
- Correlation of Android and all devices: 0.23
- Correlation of iOS and all devices: 0.51

The closer to 0 the correlation coefficient is, the weaker the relationship between the variables. Here, there is weaker relationship between Android and iOS usage of google meet as shown in (Fig. 15). There is better relationship between Total usage and iOS usage of google meet as shown in (Fig. 17) when compared to the relationship between Total Usage and Android Usage as shown in (Fig. 16).

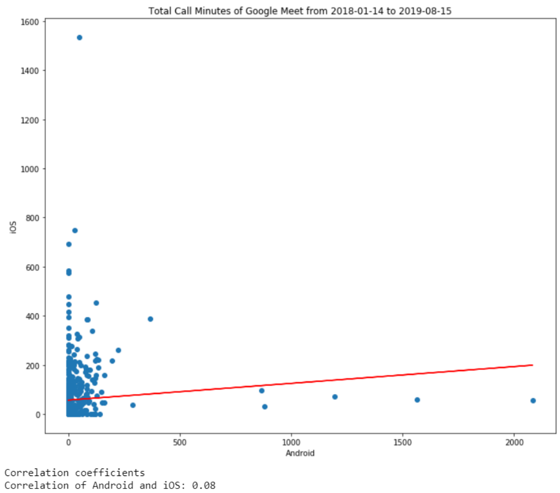

Fig. 15. Correlation of Android and iOS

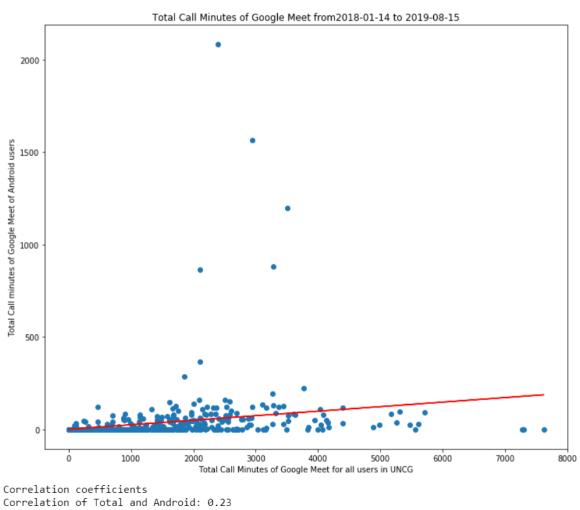

Fig. 16. Correlation of Android and all devices

2) Google Drive: The 1day active users and 30day active users have a correlation of 0.45 as shown in (Fig. 18). There is only a slight correlation between the two, It is not like every time there is a growth in 1day active users, there is growth in 30day active users.

3) Gmail: Below are the correlation results in Gmail

- Emails received and emails sent correlation: 0.95
- Emails receive and inbound non spam emails correlation: 0.88
- Emails sent and inbound non spam emails correlation: 0.83

These correlations are very strong since it very close to 1.

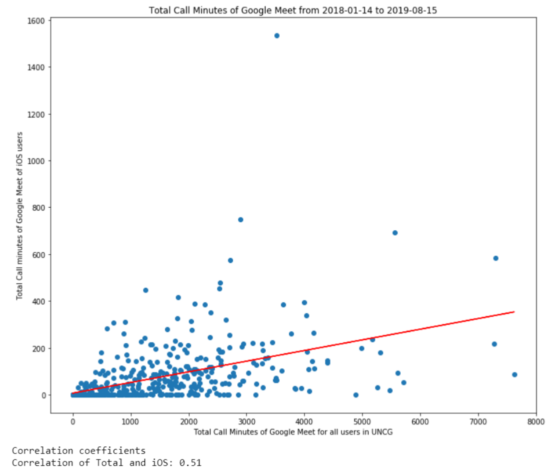

Fig. 17. Correlation of iOS and all devices

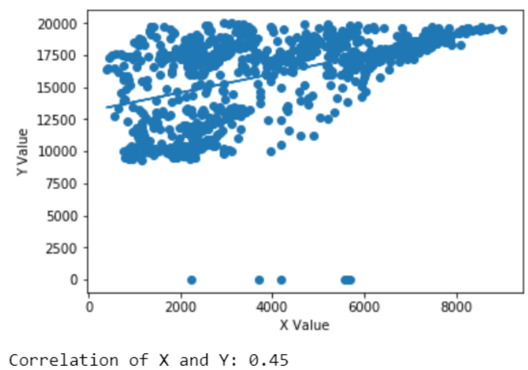

Fig. 18. Correlation of 1day active users and 30day active users

V. Machine Learning

A. Methods

Time series forecasting is a technique for the prediction of events through a sequence of time. The techniques predict future events by analyzing the trends of the past, on the assumption that future trends will hold similar to historical trends. Prophet is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. Prophet is open source software released by Facebook's Core Data Science team. Facebook prophet is used to perform time series analysis on google suite metrics. The input to Prophet is always a data frame with two columns: ds and y. The ds(date stamp) column and the y column represents the values of a metric we wish to forecast. We fit the model by instantiating a new Prophet object. Predictions are made for the future period of 365 days(1 Year). The predict method is used which will assign each row in future a predicted value. Plotted the forecast using Prophet.plot method. Prophet.plot components method is used to see the trends and weekly seasonality of the time series.

B. Outcomes and their Implications

1) Google Meet: Cross validations are done to assess prediction performance on a horizon of 70 days, starting with 210 days of training data in the first cutoff and then making predictions every 15 days. Performance metrics is used to compute some useful statistics of the prediction performance. The statistics computed are mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), mean absolute percent error (MAPE), and coverage of the yhat lower and yhat upper estimates.

The mean absolute percent error (MAPE) expresses accuracy as a percentage of the error. Because the MAPE is a percentage, it can be easier to understand than the other accuracy measure statistics.

During the Initial analysis, MAPE value is not displayed as most of the actual metric values in Google Meet data are zero. Because MAPE divides the absolute error by the actual data, values close to 0 can greatly inflate the MAPE.

Observed that the metric values are zero from Jan 2018 to May 2018 for most of the Google Meet metrics. The reason is likely that metric didn't exist then, or google discontinued temporarily. Performed time series analysis by ignoring them. The predicted trends for the next year is similar to the trends in the previous years. Google meet usage is very less at the beginning of the semester and the usage increases slowly and at the end it is high. Again, after the semester ends usage is decreasing. The same pattern is followed in 2019 and the predicted pattern is also the same for 2020 as shown in (Fig. 19). Even though after removing them, MAPE value is greater than 20 percent. Dots show the absolute percent error for each prediction. The blue line shows the MAPE, where the mean is taken over a rolling window of the dots as shown in (Fig. 26). Here, MAPE does not give good results as we have a low volume of data. In Google Meet metrics, all the actual values are very small.

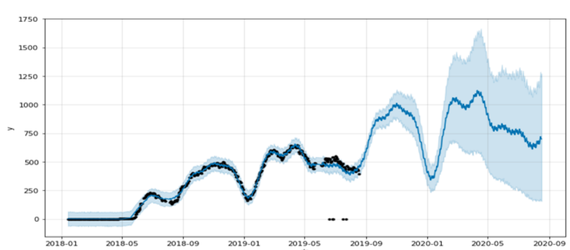

Fig. 19. Time Series Forecasting for Number of 30 day Active users in Google Meet

2) Google Drive: we find the prediction were following the similar trends as the input data in many cases, there are clear variations that can be identified in the beginning of the semester, semester break, end of the semesters for the drive data metrics. The predictions are more accurate in the initial days of the forecast like the first 50 days than compared to the end of the days of the prediction. If we consider the the attributes 30day_active_users, the predictions

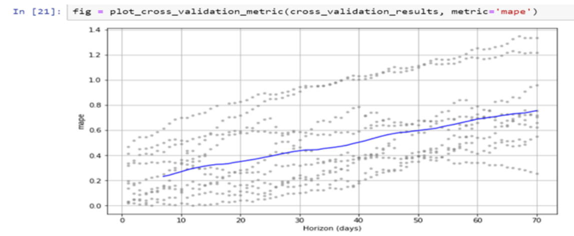

Fig. 20. MAPE for Number of 30 day Active users in Google Meet

are almost similar to the existing data and the mape(mean absolute percentage error) scores for the initial days is only 4 percent, and in the end days of the prediction it is around 20-25 percent. In the middle of semester, the error value of the mape in the semester breaks is more as, this is how mape works, as the y value decreases, mape is inflated, So, we see more error when the data is low. I was also able to see the trends in the data for some attributes like in which the metric is used more, the overall trend across the year, The monthly trend on which month the metric is used more, on which month it is less.The weekly usage on what day of the week the usage is more, which can help in understanding how the google drive is being used by members of UNCG. If we

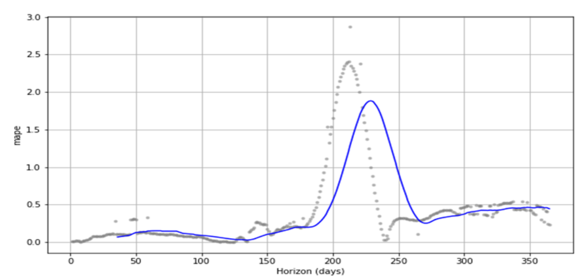

Fig. 21. Mape error for the metric 30_day_active_users

look at the above figure, mape error for 30_day_active_users for the drive data, the prediction in the initial stages were very much similar to the actual data, but there was a huge spike in the error value almost around the mid year. This is because that time is a summer time where there are very less number of students enrolled in the courses, even the staff and faculty of the university may not be working in the summer. So, there is huge dip in the data which caused the mape to inflate the error as we are calculating mape by considering the actual value in the denominator. The trends for the 30_day_active_users are seen in the figure 22, from that we can clearly see that there is a steady increase in the drive users with time as per our predicted and the actual data, so we can see more drive users in future. The trends in the more months are also seen, we see increase in usage in the months of March and November and decrease in usage in January and May. these make sense as we know the exams time and most the assignment submissions happen around March and November months. January and May are the semester breaks. If we look at the trends in the

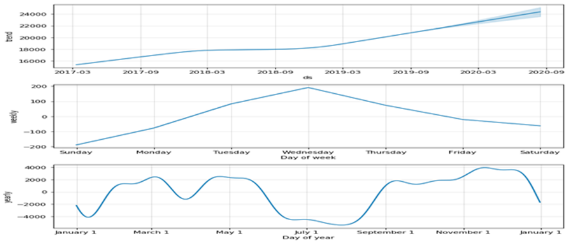

Fig. 22. Trends in the metric 30_day_active_users

usage at days of the week, there is an increased usage in the middle of the week and very less usage over the weekends.

3) GMAIL: : The predictions for emails_Received, emails_sent, 1day_active_users, email_exchange, inbound_non_spam_emails, inbound_delivered_emails will give the ITS department a firm understanding what is happening with Gmail throughout the year and allow the ITS department to prepare in advance to seasonal and overall trends in Gmail. The big question is why these services are not growing proportional to the growth of UNCG? These possible conclusions are the following: 1) Students don't send emails as much as they used to 5 years ago. 2) People are using chat more than email now. 3) Google is doing a better job of blocking spam from being received in the first place. 4) Active users however are expected to see similar growth in relation to the overall population. It's possible students/users aren't using their accounts as much as they used to. 5) Students who graduated or Faculty leaving UNCG might not be the same "quality" users as users coming into UNCG, as in a new users are not as in grained as older users and will not use it that much in the beginning. With a influx of new users over the years it could explain why the growth is not proportional. The anomalies found

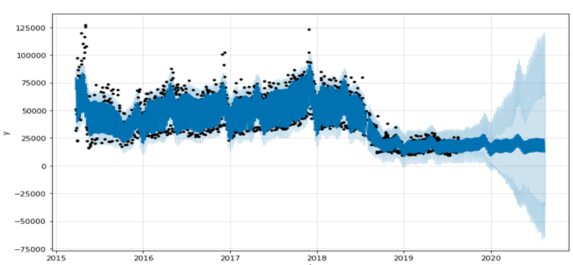

Fig. 23. The prediction for Spam emails

in inbound_rejected_emails and inbound_spam_emails can be explained by multiple of reasons. Over the years the ITS department took the "best defense is the best offense" approach by being more proactive. Additionally it is likely Google is doing a better job of handling rejections than they were before. IE : better at spam and phishing prevention. These reasons explains why there is a huge drop off in these attributes. The implication here is that there is progress being made to limit spam or hazardous emails and that is a good thing.

VI. Dashboard development

A dash board has been developed as a part of the project goals. The software that we have used to develop the dash board is the google data studio. This dashboard allows us to visualize the metrics that are present in the data set after the data cleaning. The basic dash board visualization of a metric shows the average, mean, median ,max and aggregate values of that metric. The visualization is as shown. The line graphs are used to represents the data trends.

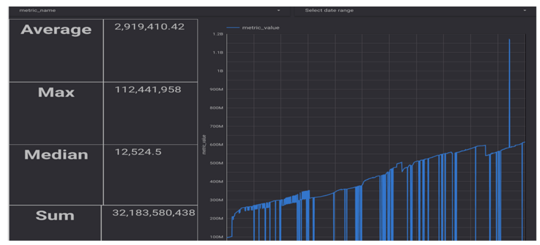

Fig. 24. Dash board visualization of a metric

The dash board also allows us to apply the filter on the metric we have selected. It allows us to apply the filter as per our time requirement. If we want to visualize the data of a metric for only 6 months, then we can select the time and apply it to the metric. From the time series analysis of the current

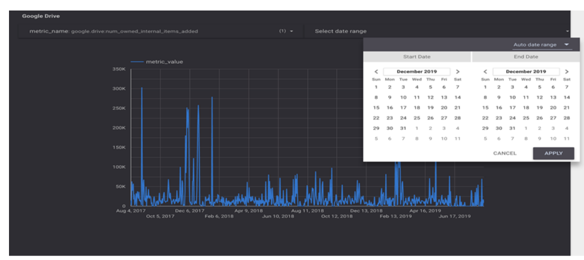

Fig. 25. Dash board visualization of a metric

data we were able to predict the trends in the future data using fb prophet. New predicted data set have been created using the time series analysis results. The results for the prediction data set and the actual data set. The trends that it follows, like is there any increase in the usage of that metric, decrease in the usage of the metric are also visualized in the dash board. It is similar to the figure shown below. In the figure if we see in detail the first pie chart represents the data used by google accounts for the gmail, drive and google plus. If we see the predicted data we can see there is decrease in the gmail quota in mb and there is a clear increase in the drive quota mb usage. from which we can interpret that in future there is going to be increase number of users in the drive, which results in the increased drive data quota and there is decrease in the gmail allocated mb, which depicts lesser number of users in future. There is also a decrease in the google plus quota.

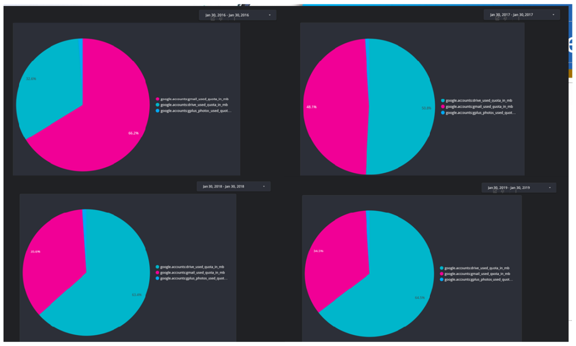

Fig. 26. Dash board visualization of a Predicted metric data and actual data

VII. Conclusion

In conclusion, team has worked in analysing the raw data that has been provided to us by UNCG ITS. We have cleaned the data, performed statistical analysis, did the time series prediction for the year 2020, based on which we can easily identify that which metrics are going to used more, which are used less. We also predicted the trends like on what day the usage is more, on which months the usage is more and so on. These results are helpful for UNCG ITS to make decisions regarding the G-suite metric services, like planning down time, providing the resources based on usage. We have also developed the dash board using google data studio which can help in better visualization of the G-suite data. The machine learning model, the dash board can be used even for the future data of the G-suite metrics to do the analysis, prediction and visualization.

References

[1] *Forecasting Using Facebook Prophet tutorials*. URL: https://towardsdatascience.com/time-series-analysis-in-python-an-introduction-70d5a5b1d52a.

[2] Somya Mohanty. *Data Science methods and approaches*. URL: https://github.com/UNCG-CSE/CSC-405-605_Fall_2019/tree/master/Class_Resources.

[3] *Predictive Analysis*. URL: http://pythondata.com/forecasting-time-series-data-with-prophet-part-1/.

[4] *Time Series Forecasting Using Facebook Prophet*. URL: https://facebook.github.io/prophet/docs/quick_start.html.