

**DS598: Project Report**

***MAPFRE DIGITAL WORKPLACE ANALYTICS***

**Submitted By**

|  |  |  |  |
| --- | --- | --- | --- |
| **Weinan Zhi**  M.Sc.  Data Science  WPI | **Chuchen Dai**  M.Sc.  Data Science  WPI | **Tahrifa Tarannum Lisa**  Ph.D.  Computer Science  WPI | **Omkar Kulkarni**  M.Sc.  Data Science  WPI |
|  | | | |

**GQP Sponsors**

**Beth Poplawski**

Knowledge Management Consultant,

Human Resources Department,

MAPFRE Insurance

**Hugh Allen**

Innovation Office Manager,

Human Resources Department,

MAPFRE Insurance

**Hurley Morgan**,

Manager Workforce Analytics,

Human Resources Department,

MAPFRE Insurance

**2. Description of the Problem**

This project will help set the stage and provide information that will be significant in the

Digital Workplace effort in 2020. This project will expand digital adoption and collaboration,

meeting improvements, and digital communication.

The project will be to provide advanced analytics and insights on Office 365 data on

an overall basis as well as by personas similar to what is provided in Microsoft MyAnalytics

on an individual basis. Analyzing data from meetings, emails, emails with

attachments, document collaborations, Skype and Team chats, Microsoft Teams usage etc.

The main point of this project is to boost productivity in the public workplace using

predictive modelling and app usage analytics insights. The questions we are trying to answer

is how the tools available at disposal of employees at MAPFRE are being used now and how

can that be improved for better productivity. What are the factors affecting that, what are the

key performance indices that reflect on the persona creation and finally how to make

employees proficient using those personas.

**3. Goals/Objectives**

1. Complete Analysis of O365 Data using 2019 data given to build a baseline model and

further analyze the 2020 data against that baseline.

(ii) Optimize use of O365 data with the data compiled, reviewed and analyzed to identify

Key Performance Indices (KPI’s)

(iii) Recommend additional KPIs based on the analysis if needed.

(iv) Design a model that can be used to measure employee’s adaption of new ways of

working by persona creation and company division.

**4. Attributes/Indicators**

The data can be clearly divided into two parts based on different tasks. The data for Time

Series Analysis comes from four applications: ‘Onedrive’, ‘Skype’, ‘SharePoint’ and

‘Teams’. In each application there exists over twenty distinct fields. Our solution to this is to

design several indicators for indicating the performance of each application so that we can

perform time series analysis on the indicators, which are way less than the initial fields,

and give predictions on those designed indicators for better understanding.

For the Persona task, data is provided as AuditLog of MAPFRE USA division within

seven days period with the maximum size of 25,000. For the time being, our idea is

to first process this data into individual-based data instead of activity-based data.

After that, we will try to join it with ‘Census’, which is a file containing

detailed information for each employee within MAPFRE USA division. For the

following procedure, clustering methods might be considered for this persona task.

Activity data consists of data from O365 applications like Skype, Onedrive,

Sharepoint, Teams, Email. It consists of Application activity and User activity data.

User activity is indexed by email ID’s of employees. While Application activity is

indexed by datetime. Dates range from Sep 2019 to Jan 2020 (180 days).

**5. Data Preprocessing**

**5.1 Data Description**

Office 365 is a huge ecosystem which is predominately used in lots of workforces.

It consists of multiple interconnected services (Exchange, Teams, Azure, OneDrive

and Skype) We are using four main types of datasets.

1. **Audit Log Data**

Auditlog tracks Office 365 activity for the purposes of auditing and includes

activity made on the services. Some attributes of Auditlog include:



**1. Census** **Data**

Census data is demographic data about employees at MAPFRE. It includes attributes Gender, Age-Range, Location, Tenure, Salary Grade, Organizational level.

**2.Survey Data**

Survey Data is generated from survey questions sent by team members and

corresponding responses.

**3.Activity Data**

**5.2 Time series indicators**

For the Time Series Analysis task, we first transform and join the data within each

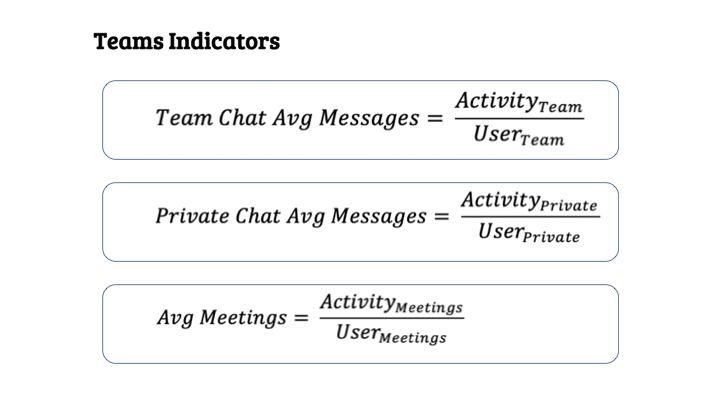
application as one file indexing on the ‘Report Date’. Besides these generated files, we

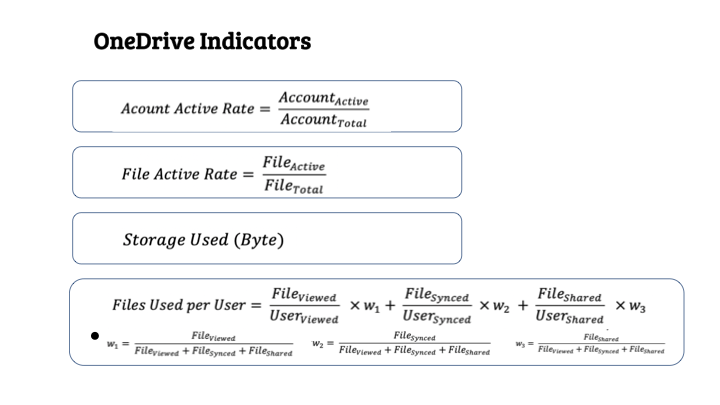
have got some data left indexing on the ‘Email’, which represents for each individual, will

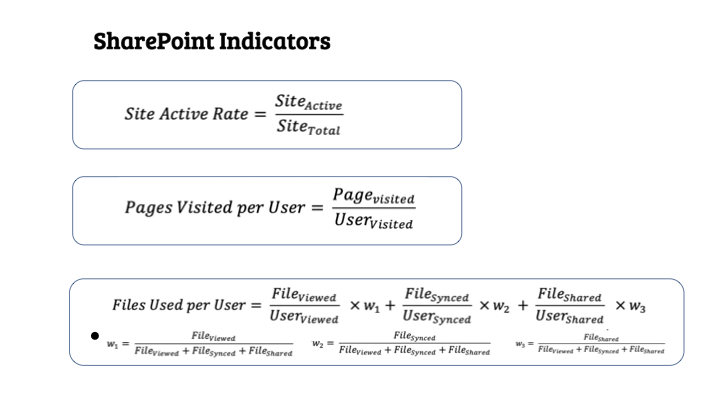
potentially be utilized for our Persona task. In addition to that, we’ve also designed

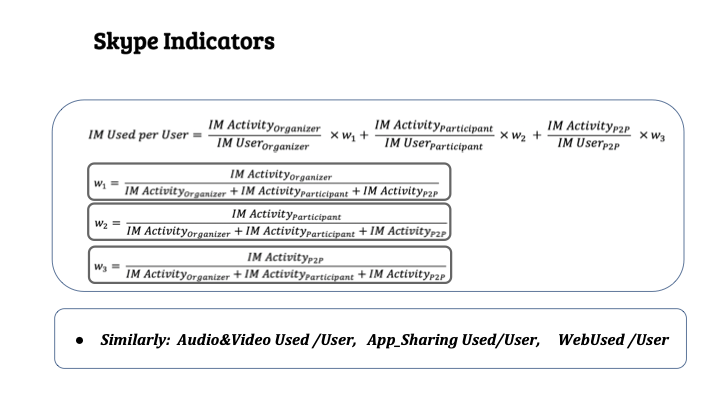
indicators mentioned above for generalizing the target agenda. The indicators are designed

as follows:









For the Persona task, we used 3 kinds of data from different sources to make the

final result reasonable. These data are: office 365 activities that show employees’ activity patterns and trends, employee census data that include some basic information about employees, survey results that we designed with the MAPFRE managers to know employees’ subjective ideas better.

Here we clean all the data separately. For the office 365 dataset, we chose some

user-based data that can show each user’s activity in each application. For the

AuditLog data, we transferred the json file into csv format, and made feature

selecting at the same time. Some features were changed into more meaningful

format, e.g, the “time” attribute was divided into “month”, “day”, “week”, “hour”, “clock time” etc. For census data, we will select some attributes that are relevant to the topic. For the survey, employees’ answers will be turned into their features in this part.

After the basic cleaning, we combined all the three parts of features together.

**6. Time Series Analysis**

       On the applications data provided, ARIMA models are the baseline models and

LSTM, PROPHET models were done and the most accurate results were chosen for analysis. For this analysis, the given data was for a period of 180 days. We divided the data into train and test by 18 days margin. On comparing the results it can be said that the outputs found from the LSTM model are more accurate than the ones

obtained from ARIMA and PROPHET models.

In order to label employees with different personas, clustering algorithms will definitely be adopted for better accuracy. The existing data will be treated as our unlabeled training set for training and validation purposes and the upcoming data will be used as a test set. For the dataset we will be relying on employee census using a survey questionnaire.

As for the algorithm choice, simple clustering methods as model-based

clustering, density-based clustering, hierarchical clustering, decision trees will be performed for general comparison. Afterwards, the bagging strategy will be

considered in order to pursue better clusters. The final proposed model will be expected to produce promising clustering results.

After we get the clusters and employees assigned to appropriate clusters, we will look for KPIs that highlight employee characteristics that boost productivity and do further persona creation.

**7. Time Series Analysis Results**

Results will be organized by each application since each application has its

distinct indicators designed as mentioned above. Each application result will consist

of two main parts. The first part will be the test result and the second part will be

the future 18 days predictions.

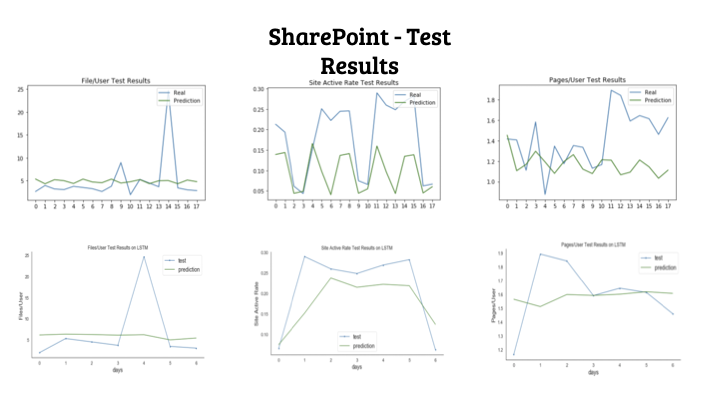
**SharePoint:**

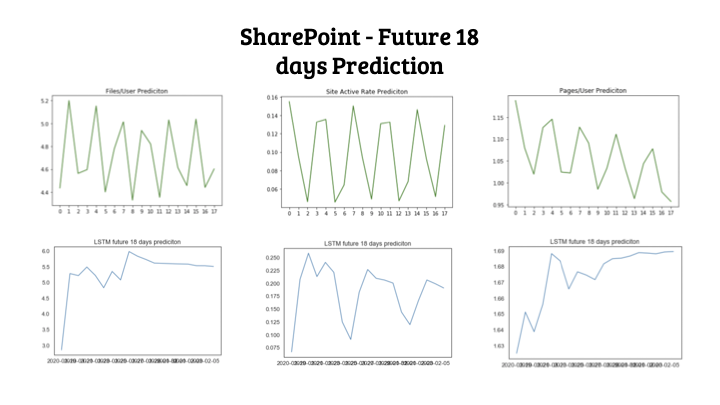
The plots on the first row are the baseline models, which is the ARIMA model and

the plots on the second row are LSTM models in this application. The plots on the first

column is for ‘File/User’ indicator, the plots on the second column are for ‘Site Active

Rate’ indicator and the plots on the third column are for ‘Pages/User’ indicator.

****

****

From the results shown, it is obvious that the LSTM models outperformed ARIMA models. Though the test errors do not show too many differences

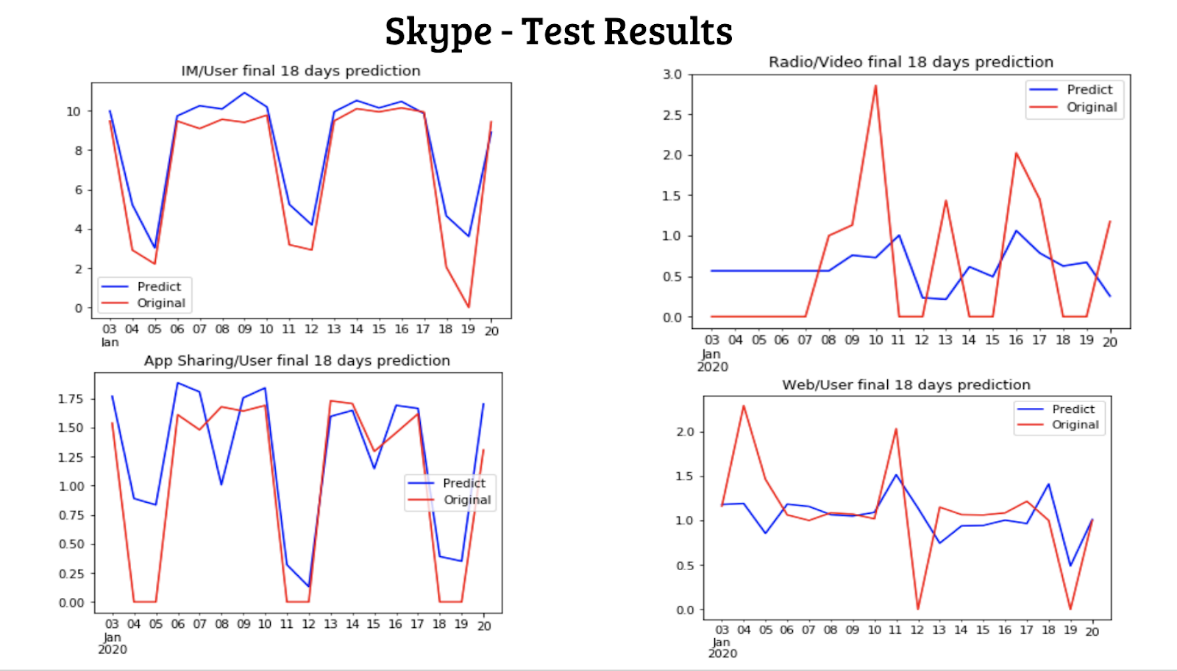
between ARIMA models and LSTM models, the predictions of three indicators are performing drastically differently. In the ARIMA models, ‘File/User’ and ‘Site

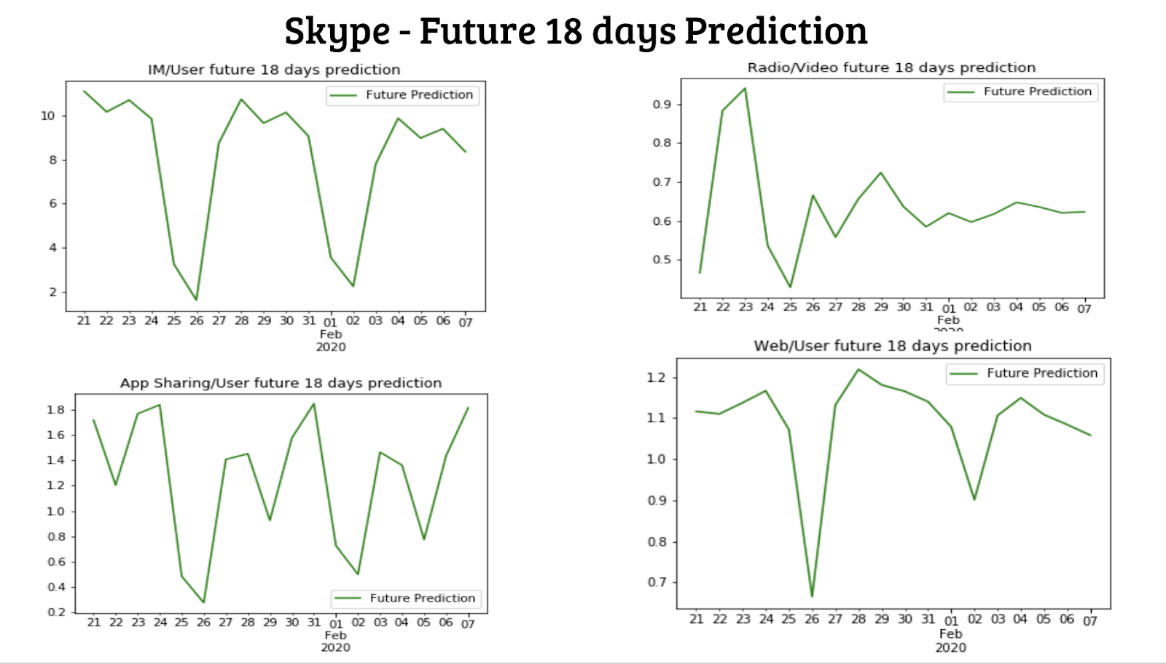
Active Rate’ are showing repeating patterns with no trend, ‘Pages/User’ is

showing repeating patterns with a steadily decreasing trend.

On the other hand, in LSTM models, obvious trends indicate much clearer performance on SharePoint Usage. Both ‘File/User’ and ‘Pages/User’ are showing a drastically increasing trend, which indicates SharePoint will become more frequently used in the future.  ‘Site Active Rate’ is showing a slowly increasing trend which indicates among sites created within the MS SharePoint, active sites will take more and more roles in the future.

**Skype**:





For the Skype results, IM/User and App Sharing were predicted better because

of the obvious trend of training data. Low count of holidays may be an important reason. For the Radio/Video result, the prediction looks not very accurate due to the sudden floating of testing data, but it’s good that the stable trend of the first several days of January can be predicted. About Web/User results, the going down trend of

the whole January was predicted well, especially for the mid-January part. Although Skype prediction shows some fault in detail, overall the results are worth to be used

to help the company for future prediction and preparation.

**OneDrive:**

For Onedrive KPI’s, following are the results for ARIMA and PROPHET models.

PROPHET:

Model consists of

\* A piecewise linear or logistic growth curve trend.

\* A yearly seasonal component modeled using Fourier series.

\* A weekly seasonal component using dummy variables.

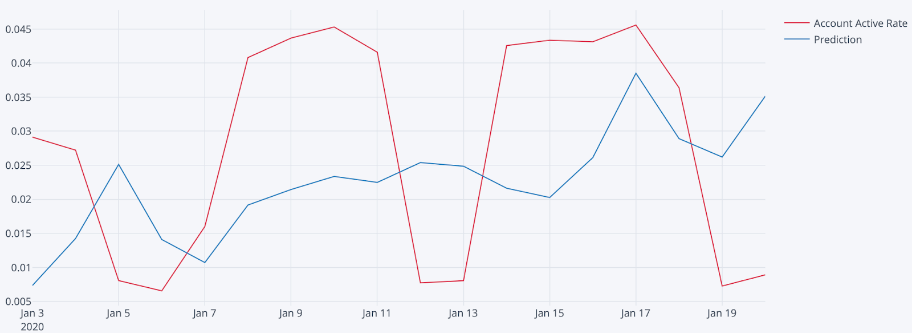
\* User provided list of Holidays

It also provides smoothing parameters for seasonality that allow you to adjust how closely to fit historical data cycles, as well as smoothing parameters for trends that allow you to adjust how aggressively to follow historical trend changes.

For Files Per User we get unusual prediction results because of extreme outlier which is a sudden spike in the amount of files transferred on that particular day. Apart from that PROPHET model works better for half KPI’s while ARIMA shows best fit for rest.

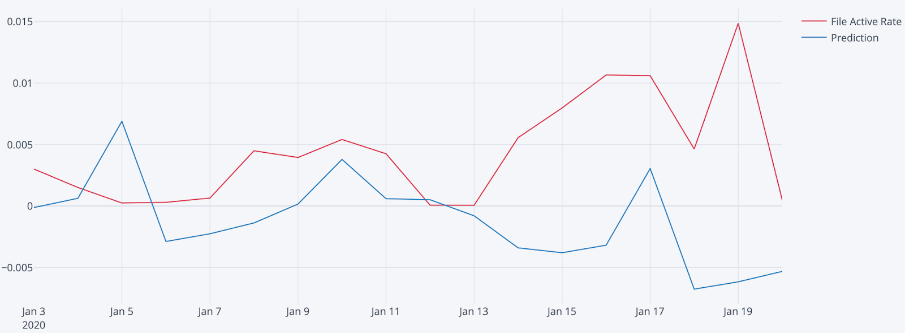
**OneDrive - ARIMA Test Prediction**

Account Active Rate



LS

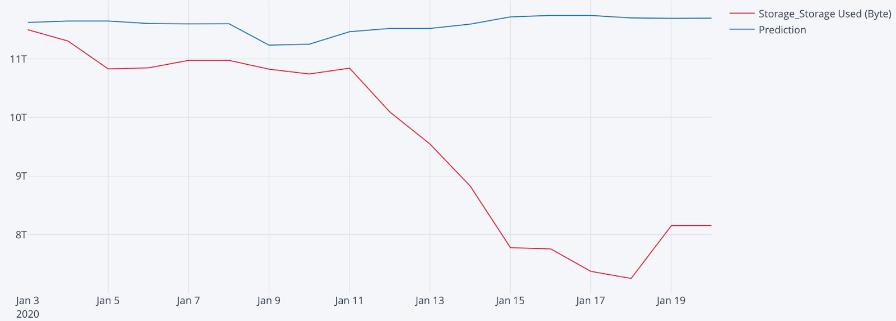
File Active Rate



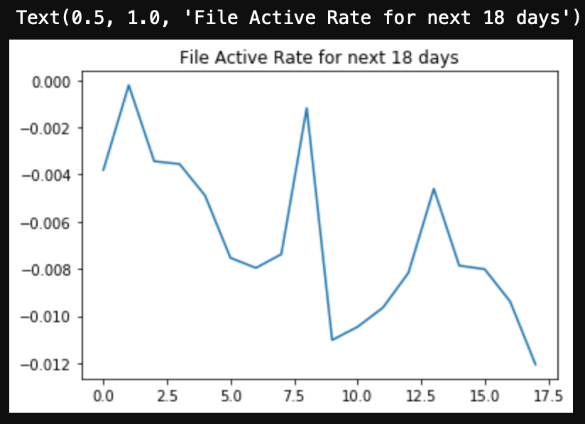
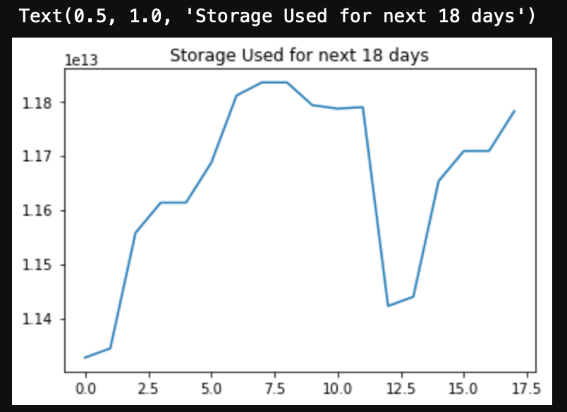
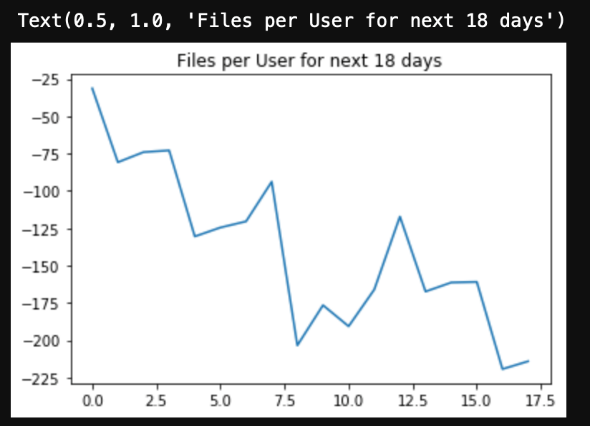
File/User



Storage

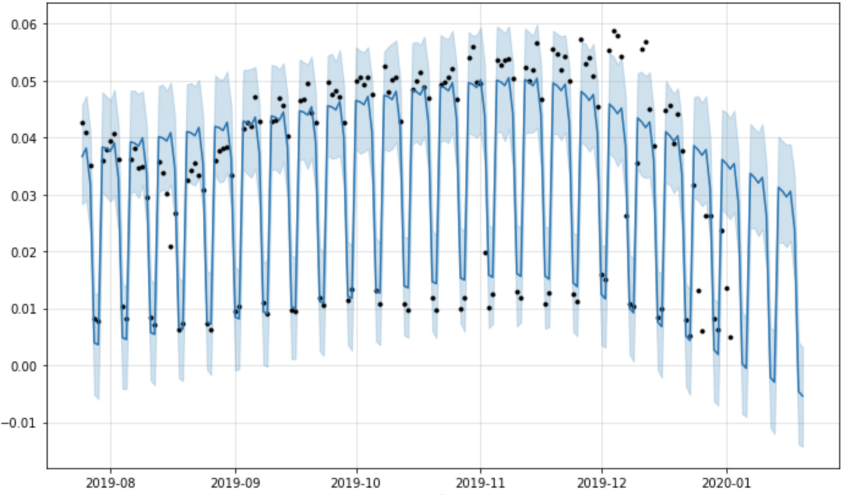


**OneDrive - ARIMA Future 18 days Prediction**



**OneDrive - PROPHET Test Prediction**

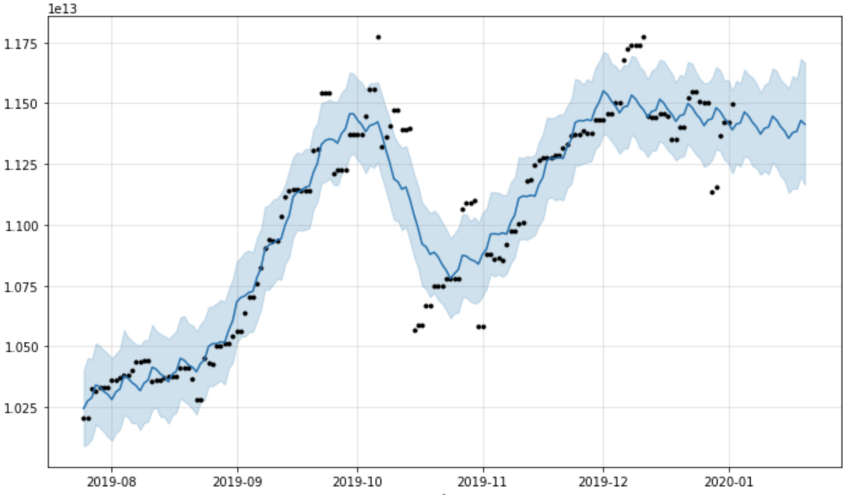
Files Per User



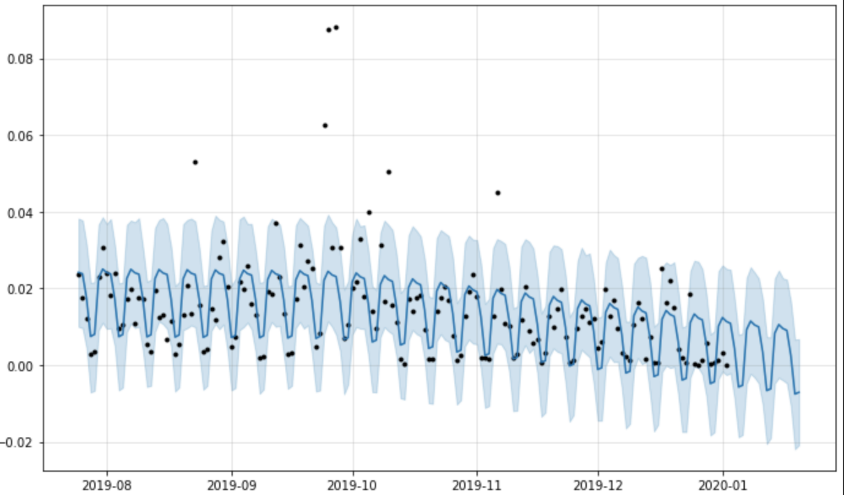
Account Active Rate



      Files Active Rate



Storage Used (Bytes)



For Storage used (Bytes) indicator, while predicting ARIMA does model linear

curve. Although the prophet doesn’t model a steep curve like we have in the data prediction it shows a downward trend.

**Teams**

Following are the results for ARIMA and LSTM models. For ARIMA model, Teams

chat average message and average meetings show repeating pattern with no trend.

But private chat average message shows repeating pattern with slightly decreasing

trend. On the other hand, the results obtained from the LSTM model is slightly better

than the ones from the ARIMA model. In LSTM, Teams chat average message shows

slightly increasing trend. Though Private chat average message and Average Calls

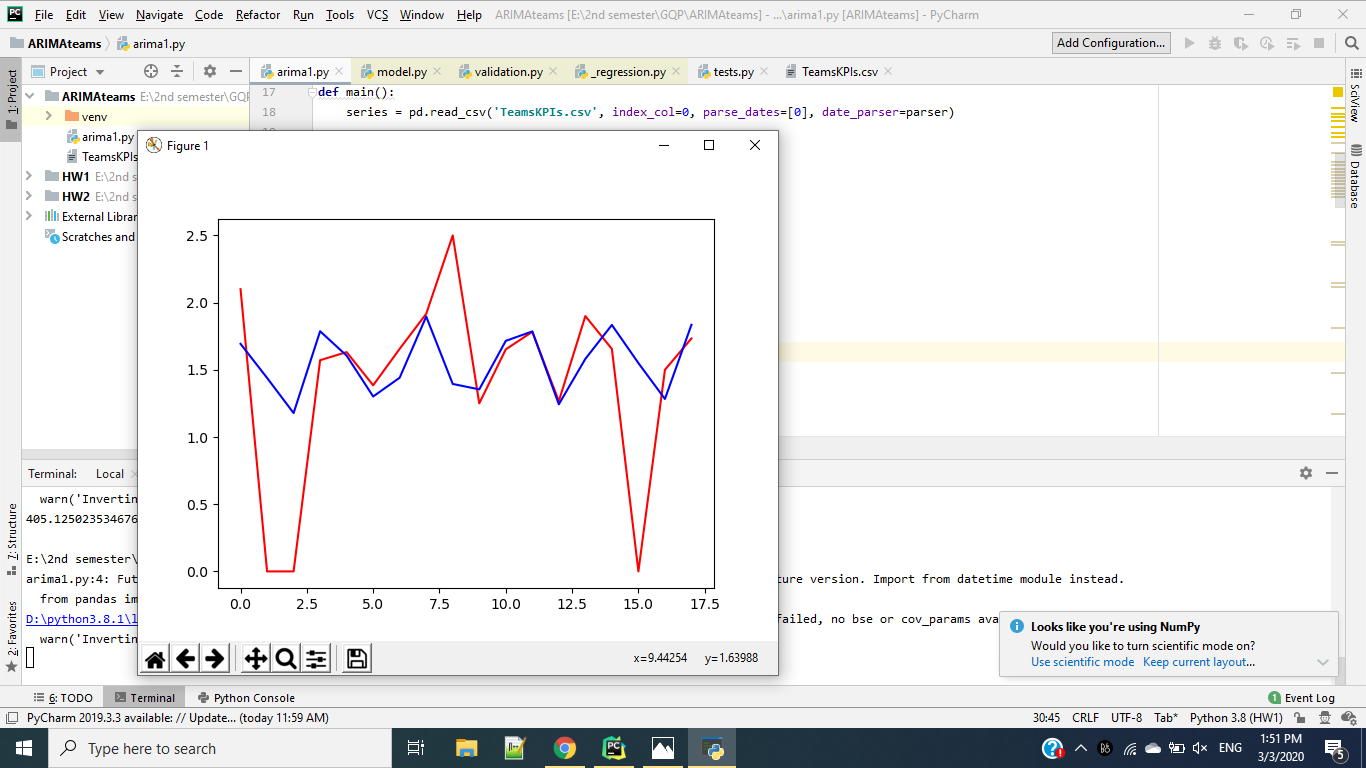
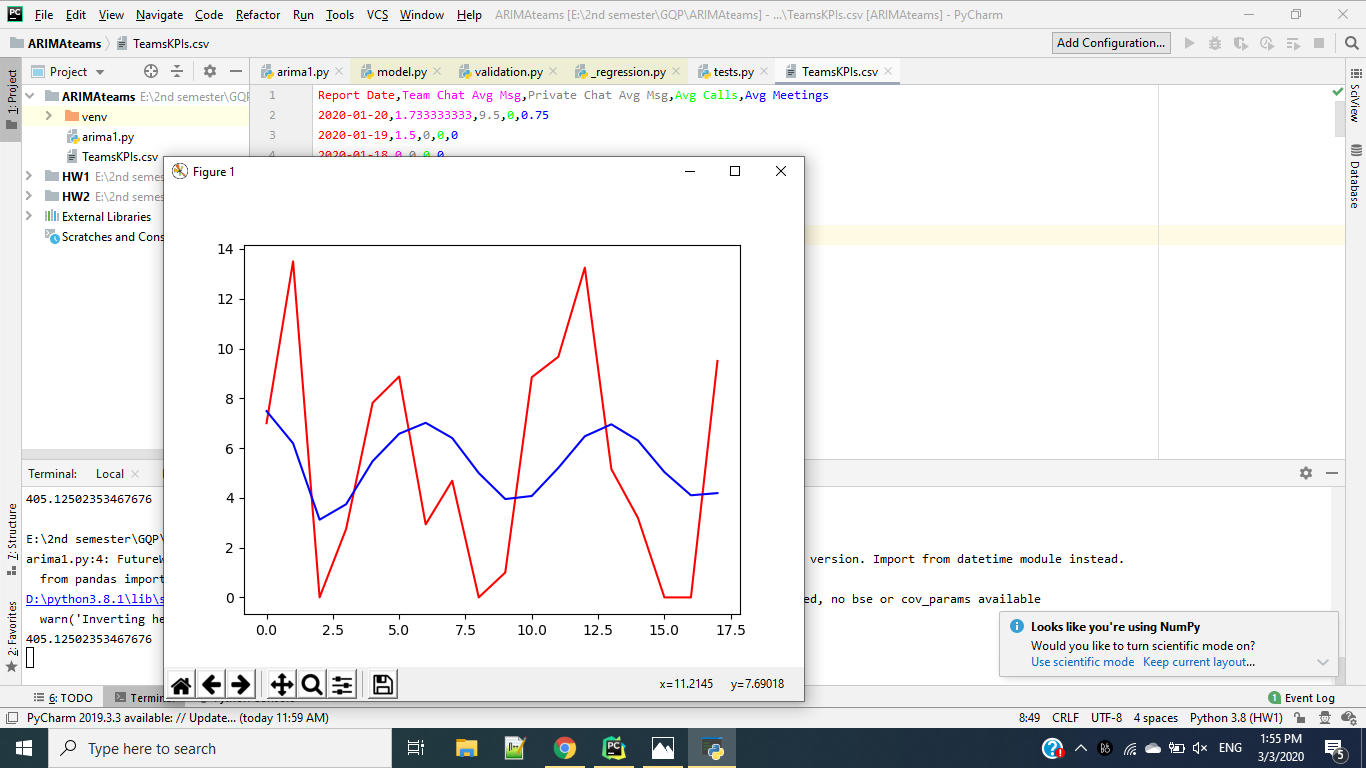
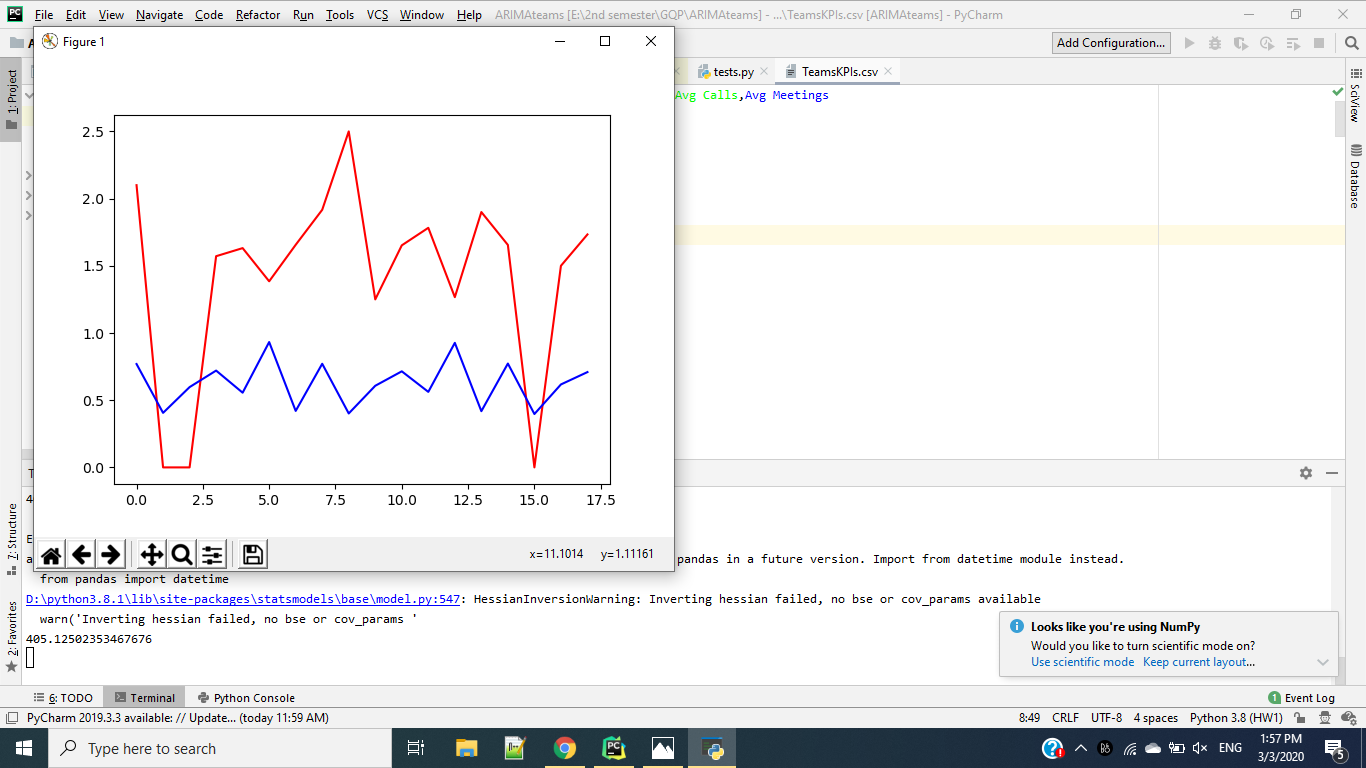
predictions display mostly decreasing trends, Average Meetings prediction shows

both increasing and decreasing trends overall. LSTM future 18 days prediction

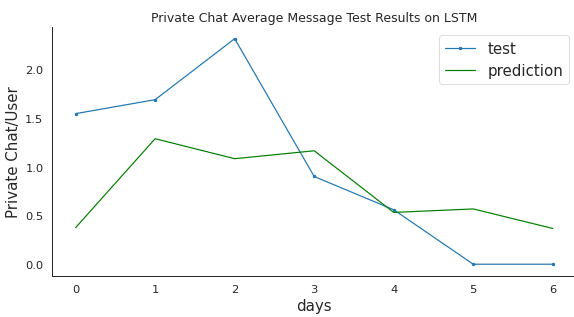
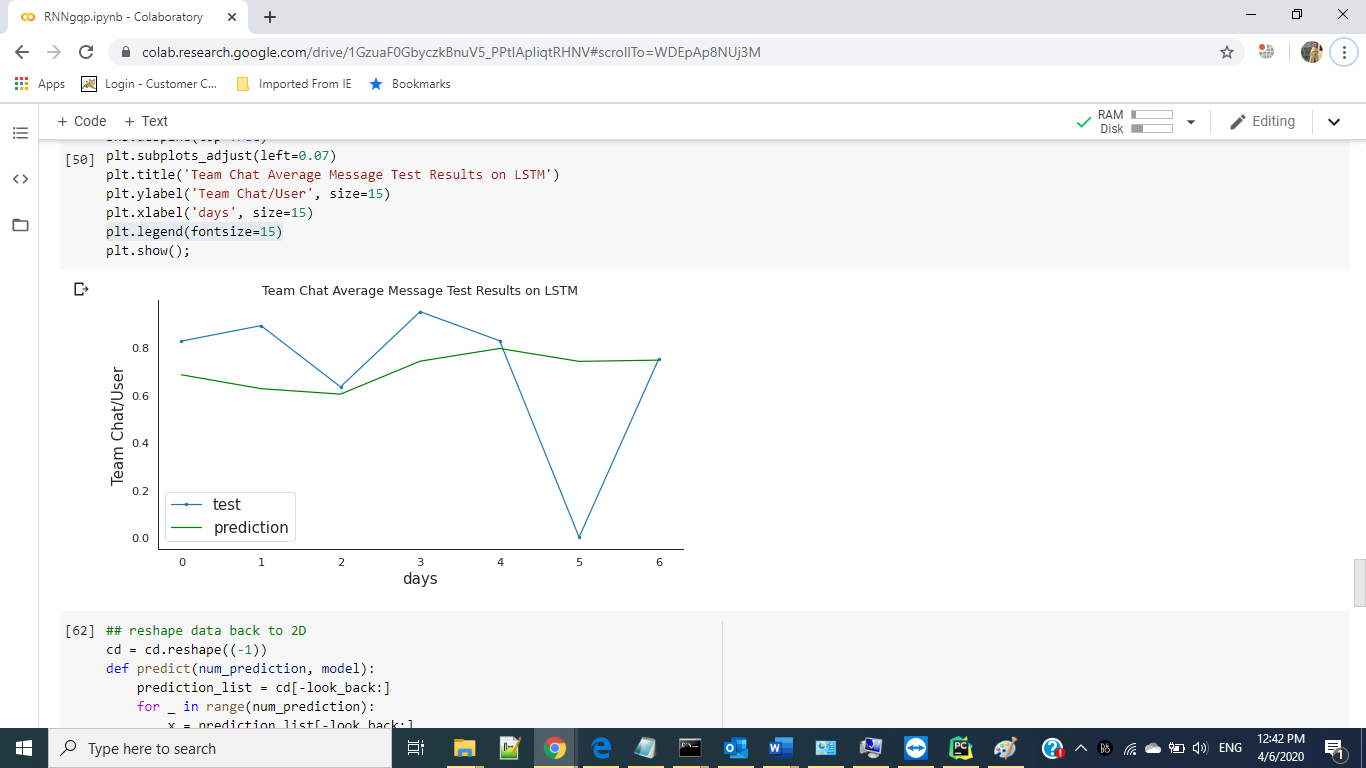
definitely shows better trends using the LSTM model.

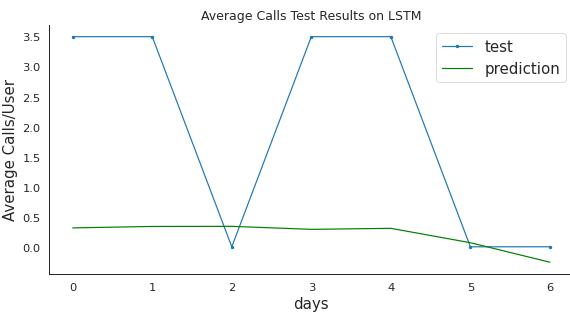
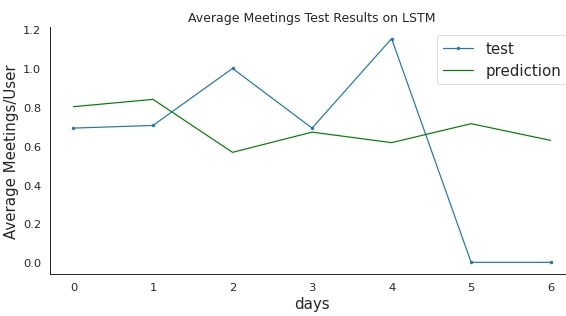
**Teams ARIMA test prediction**

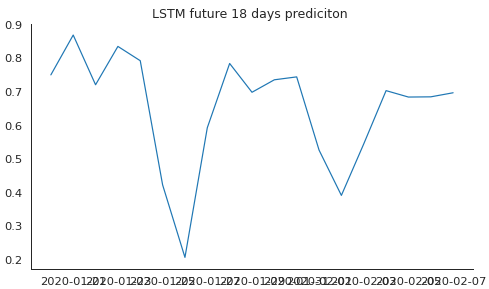
Teams Chat Average Message   Private Chat Average Message Average Meetings

**Teams LSTM test prediction**



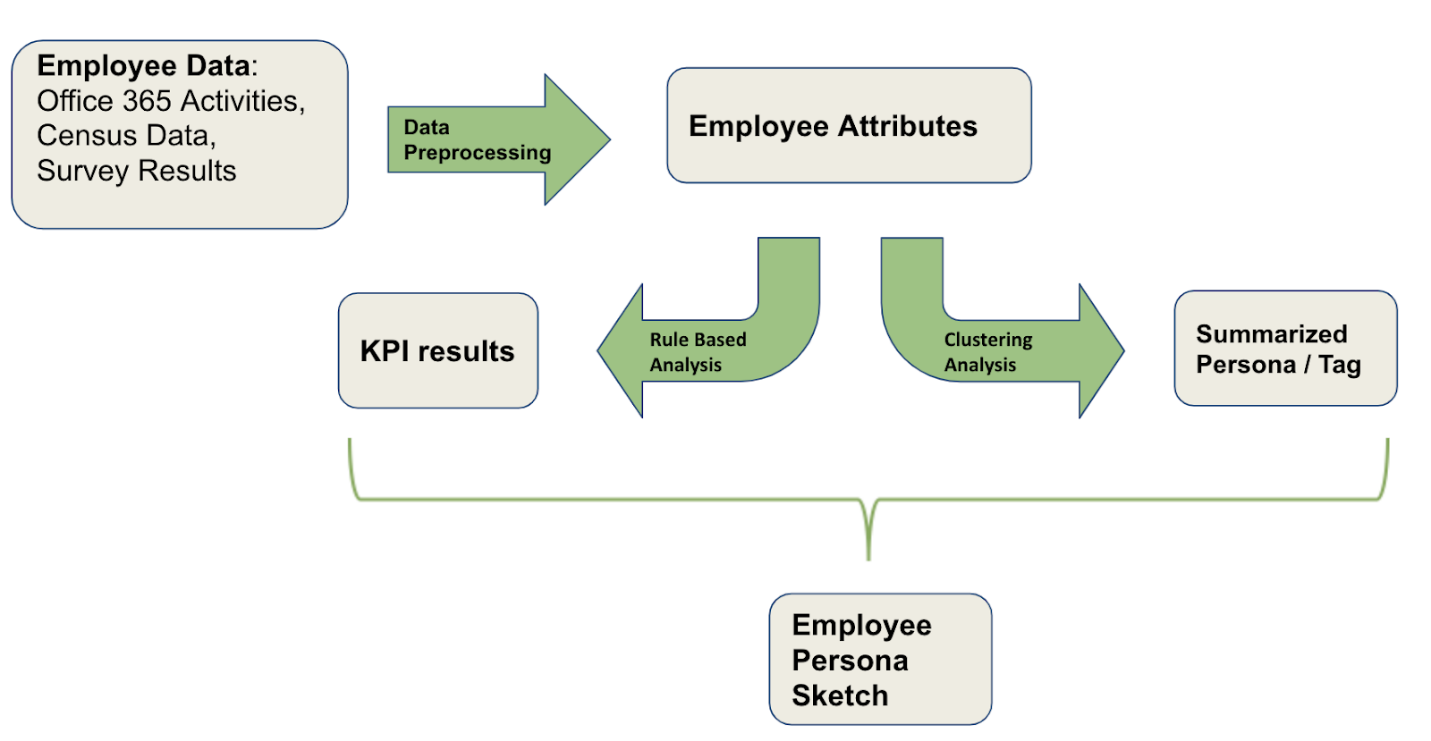


**8. Persona System**

In our design, the employee persona system will be composed of KPI results and

summarized persona tags that are generated by clustering analysis.

This is our workflow:



We preprocessed the data and did feature selection to get the most related

attributes. This was our input data in the clustering analysis part, and was used in

KPIs generation. An employee will have one persona tag and 3 KPIs’ scores.

The KPIs we designed are: Active Level / Adaptation; Communication Ability;

Working Responsibility.

**9. KPIs Generation**

When designing the KPIs, we considered what data we have and what the

managers may need to know, to get a reasonable result. The 3 KPIs are generated

from different data source, and we used hierarchical rules dividing people into small

groups layer by layer to get a final score for each one. The full mark is 100, and the

lowest score should be about 40.

- In **Active Level** (adaptation) generation:

This KPI showed how often the employees use office 365 applications, and the deep meaning behind this is how well the employees get used to new things.

It is mainly generated from: Office 365 Auditlog data, survey results, census data.

In Office 365 Auditlog data, we calculated the activity frequency of employees from the formula: Activity Frequency = Activity Times / used days.

In survey results, we collected two question answers as input: “How often do you use Teams?” and “How to share files?”

In census data, we picked the attribute “Age Range”.

This is the hierarchical structure of this KPI:



We divide employees into groups based on the age range first, because we think

mature employees with longer experience will be less inclined to use new platforms,

and those young employees should get familiar with new things faster. The scores of

mature groups should be less strict than youth. To be specified, if a mature employee

and a young employee have the same situation of other criteria, the score of the

mature employee will be higher than the young employee.

Then we divided them by activity frequency. Here we used binning technology

cutting all the input data into 3 pieces evenly to get the boxes “high”, “middle”, “low”.

Next layers are the two question results of the survey, we used the question options

to divide the groups. After these, each employee got a final score of “active level”.

- In **Communication ability** generation:

This KPI showed how often the employees communicate with others on office 365

applications. It is mainly generated from: Office 365 Activities data, Auditlog data,

census data.

In Office 365 Activities data, we used records counts of three applications: Skype,

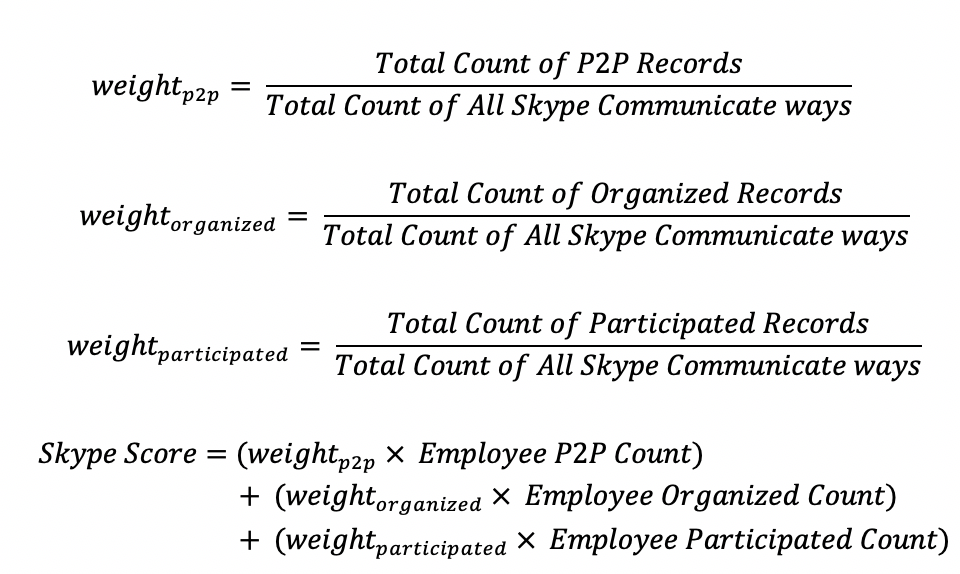
Teams and Email. In order to make the final score more reasonable and fair for

everyone, we customized some formulas here:

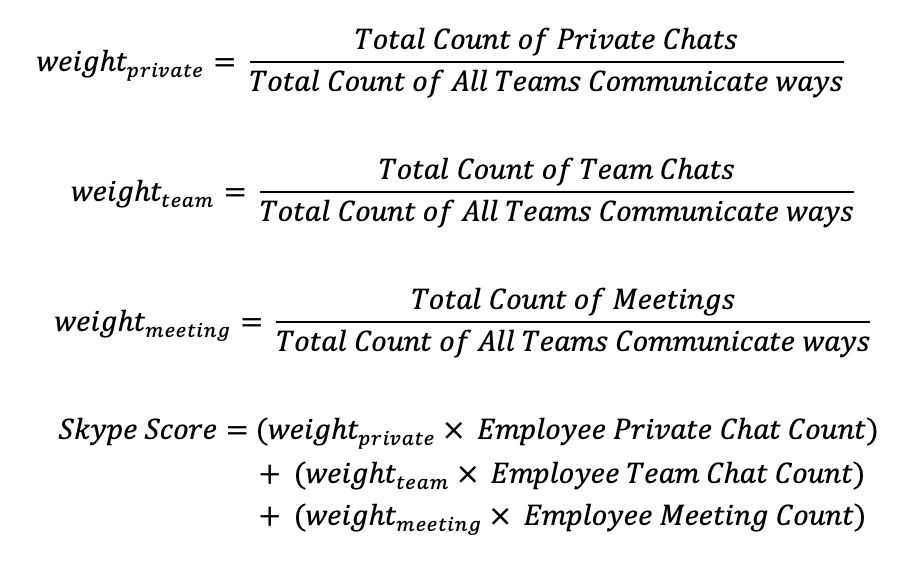
For Skype data, there are 3 approaches to show communication frequency. We

calculated a weight for each approach, then we put the weight times the approach to

get a final score of this application:

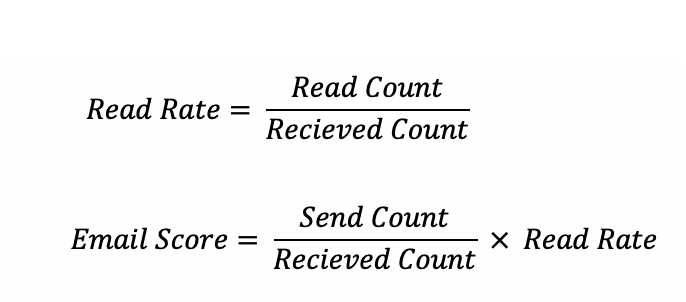


For Teams, we used similar methods:



For Email, we calculated the read frequency first, then used the read rate times

the reply rate to get a final score.

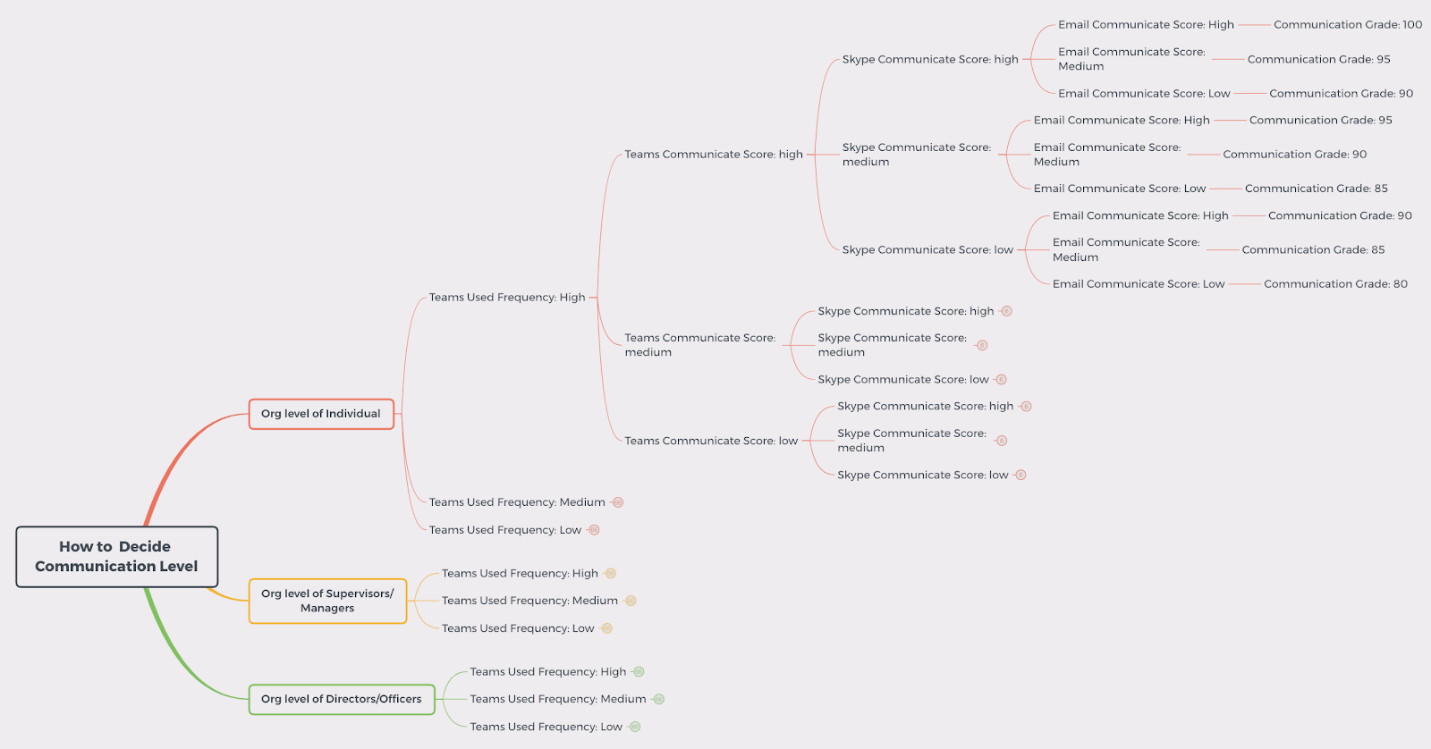


In Office 365 Auditlog data, we calculated the Teams activity frequency of

employees from the formula: Teams Activity Frequency = Teams Activity Times / used

days. In census data, we picked the attribute “Org Level”.

This is the hierarchical structure of this KPI:



In the structure, we divide employees by organization level first, because we

believe people with higher rank org level should manage more people, and they will

have more chances to communicate. For them, the grade should be stricter. Then we

grouped people by teams using frequency, since we think if people used Teams more

frequently than others, they should have more chances to communicate, and their

scores should be stricter. Next layers we used binning on the three formula results to

cut data into “high”, “middle”, “low” levels. Final results combine all the above

considerations.

- In **Working Responsibility** generation:

This KPI showed how conscientious the employees are from their habits in using

office 365 applications and skills they had. Here we suppose that people who spend

more of their free time on work are more responsible, and people with more skills the

department required are more responsible.

It is mainly generated from: Office 365 Auditlog data, survey results, census data.

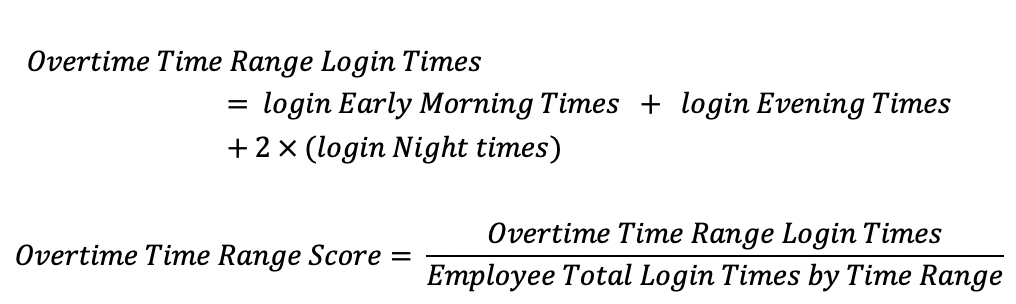
In Office 365 Auditlog data, we used employees’ login time counts by time range and

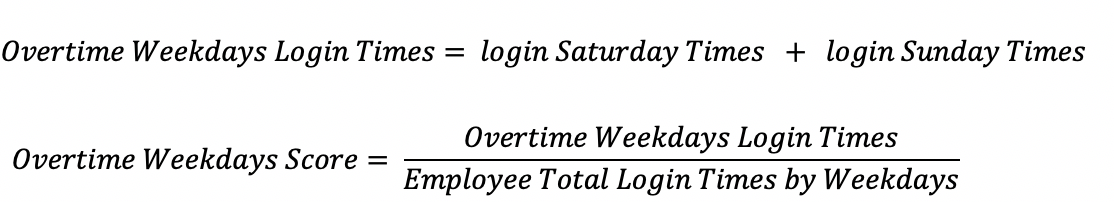
weekdays from different O365 applications. Also, we customized some formulas here:

For login time by time range, we got the count of how many times the employee

logged in on unusual time ranges, then we got a rate of their working overtime

frequency. Similarly work overtime by weekdays were obtained.





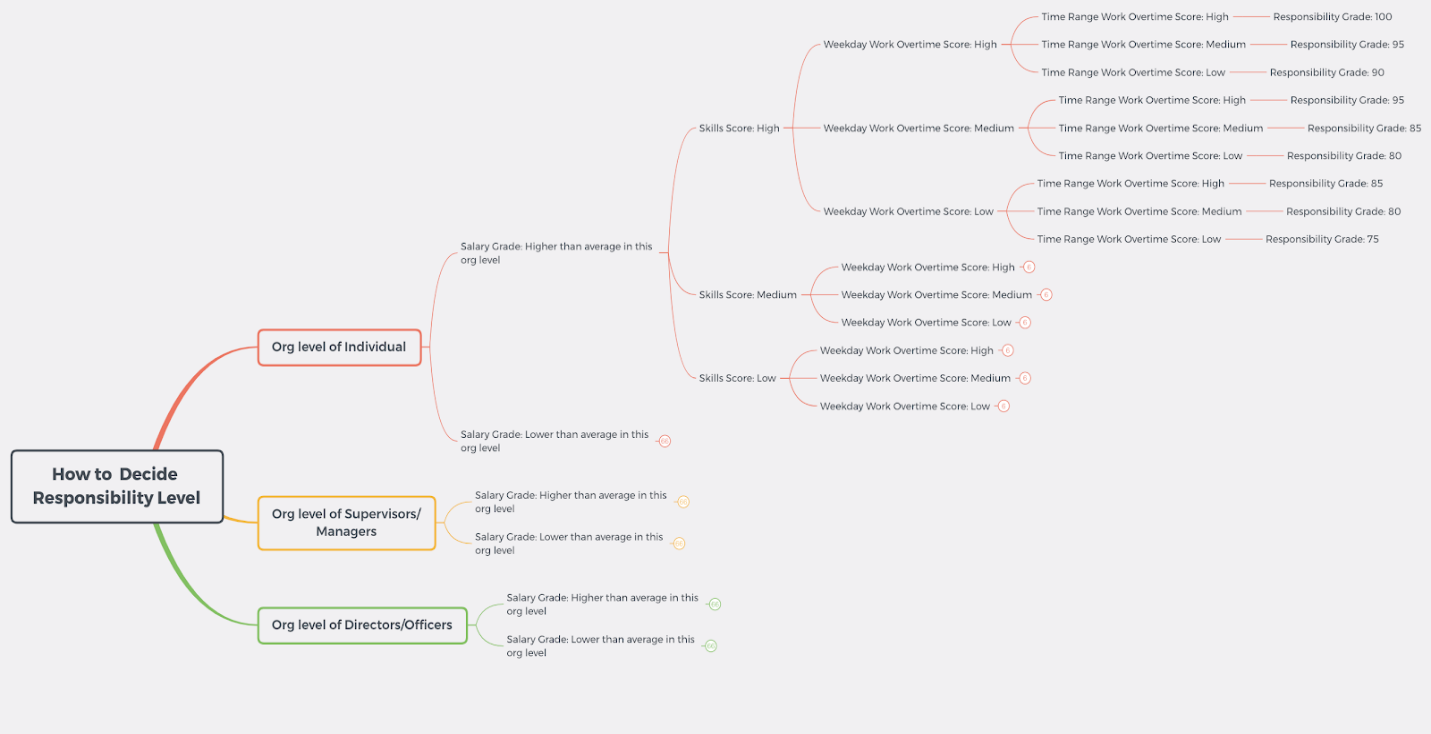
In survey results, we collected employees’ departments and how many skills they

mastered in the department required list. The formula is: Skills Score = Employee

Owned Skills Count / Department Required Skills Count.

In census data, we picked the attribute “Org Level” and “Salary Grade”.

This is the hierarchical structure of this KPI:



We firstly divide employees by org level, too. After that, we used salary grades to

divide employees, because we think that people who earned more should contribute

more to their work. Then the grades for the higher salary group would be stricter.

Similar as last one, next layers we used the formula results after combining.

**10.1 Clustering Analysis**

In order to decide the persona employee system, various types of clustering have been performed. Using K-means clustering as a baseline model, Hierarchical clustering, DBScan clustering and SOM model have been implemented.

**10.2 Clustering Analysis Results**

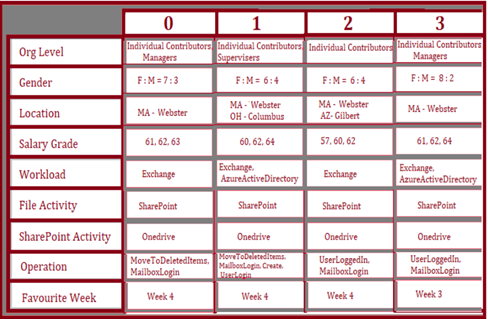
**K-means:**

K-Means clustering is an unsupervised learning algorithm that finds a fixed number

of clusters in a set of data. k-means clustering aims to partition n observations into k

clusters in which each observation belongs to the cluster with the nearest mean, which

serves as a prototype of the cluster.



As you can see, K-means results fail to generalize employees in clusters. Due to

outliers k-means doesn’t work well in this case. We tried to use K-means as a baseline

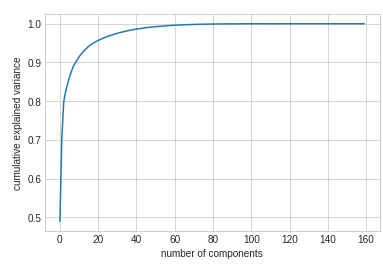
model.

**Hierarchical Clustering:**

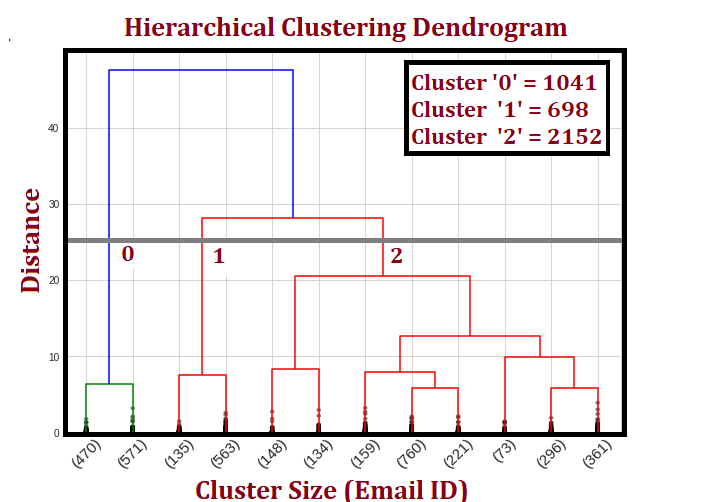
Hierarchical clustering is an algorithm that builds hierarchy of clusters. It is a

distance based clustering method. This method predicts data by finding distance between each data point and its nearest neighbors. Agglomerative clustering which

is a Bottom-up approach has been used. A variance minimizing algorithm Ward portrays the dissimilarities. Ward minimizes variance inside a cluster and maximizes variance among clusters. Number of components of clustering against cumulative explained variance are plotted below.



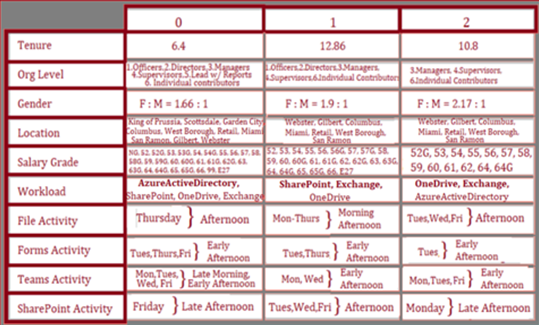
We have created the following dendrogram to show hierarchical clustering.



.

The number of clusters have been calculated by using principal component analysis (PCA). Principal Component Analysis is a statistical procedure that uses an orthogonal transformation which converts a set of correlated variables to a set of uncorrelated variables. According to the dendrogram Cluster 0, Cluster 1 and Cluster 2 have 1041, 698 and 2152 Email ID’s respectively

The statistics obtained from this clustering is shown below in a chart.



**DBScan Clustering:**



The clustering results generated by DBScan is given above. The cluster 0 has all

the officers and their application availabilities are ranged through the weekend

during afternoon and evening. Besides, employees from Scottsdale, AZ and Columbus,

OH are all clustered in this cluster with provided AuditLog data. The cluster 1 has over

90 percent of Teams users in our integrated data and employees in this cluster also

tend to have a full-week Teams availability especially at afternoon and evening time.

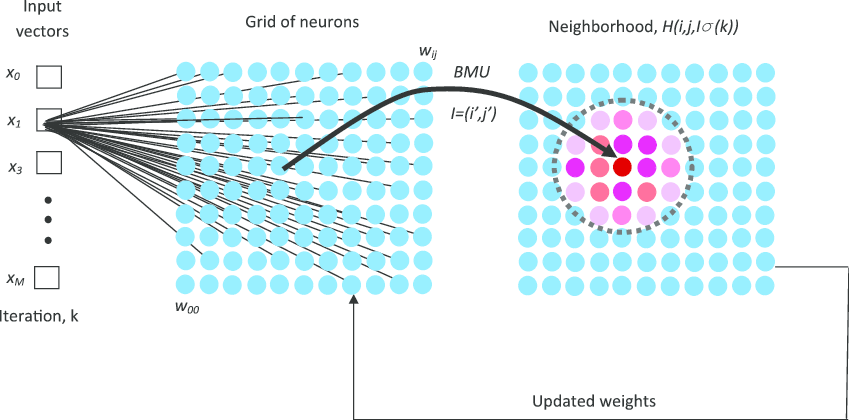
The cluster 2 and 3 are mainly female employees with similar “Org Level”, “Location”,

“Workload” and so on. The slightly distinguishable characteristics are the Teams

availability.

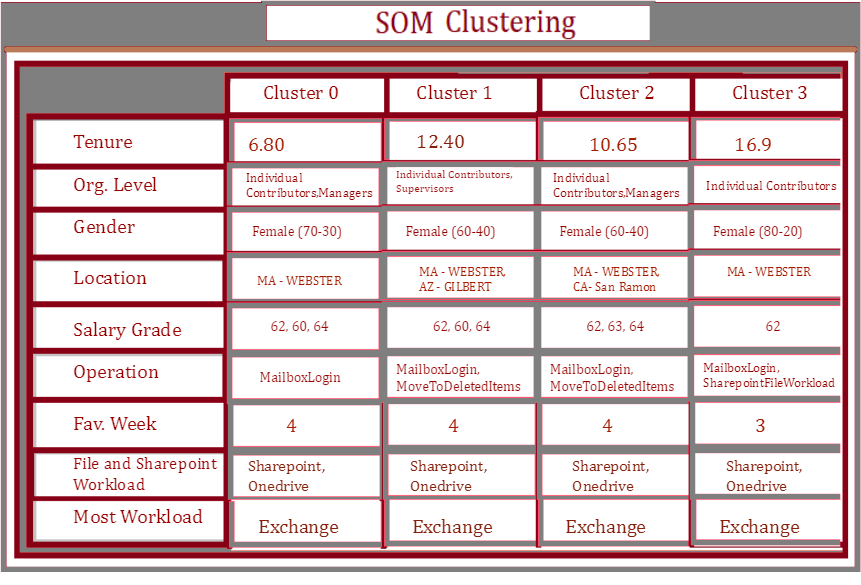
**Neural Network Clustering:**

For Neural Network based clustering we have decided to use Self-organizing maps. Below is the figure which explains SOMs in an intuitive way. Self-organizing maps are a class of supervised learning neural networks used for feature detection. As opposed to traditional neural networks using error correlated learning, SOM uses competitive learning. Refer the below figure:



On the left side we have input vectors for our input data. In the middle, the 2-dimensional grid of neurons is characterized by local coordinates (I, j) and weights w ij. The SOM grid is a 2-d manifold space onto which each observation in the m-d space is mapped via its similarity with the vectors for each cell in the SOM grid. During the k’th iteration, an input vector x m is presented to the grid, and the neuron with the smallest distance to the input vector is considered the best matching unit (BMU) vector. At the end of the learning process, several samples from our data can be comfortably represented by one unit of the SOM, and our data points are now “clustered” around it. The Gaussian neighborhood function h(I, j, I, r(k)) determines the strength of association among neurons.

Refer to below figure for results for SOM model,



Although these results are better than K-means clustering, due to lack of hyper-parameter tuning and optimized training these are not sufficient. These results

cannot cluster employees by preserving similarity within cluster while keeping maximum dissimilarity between clusters.

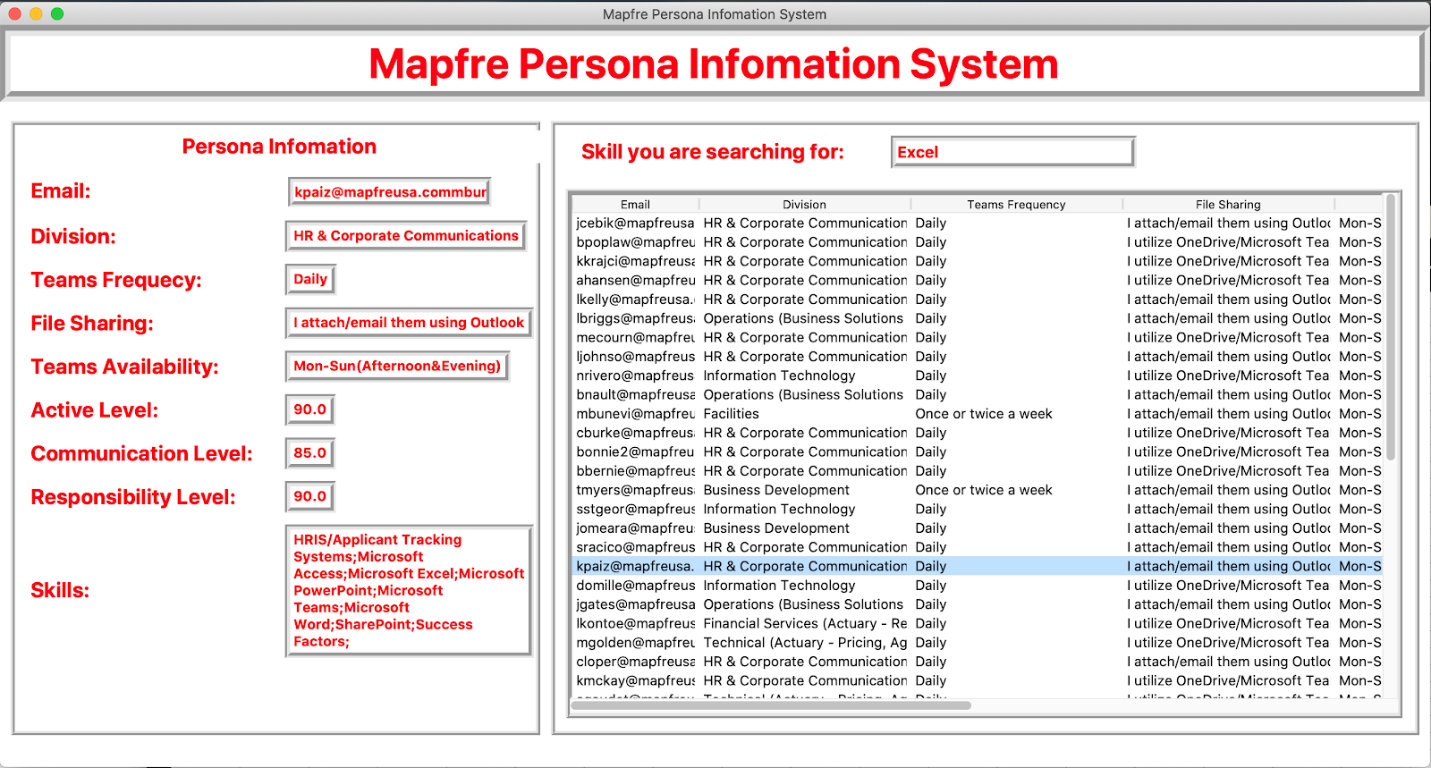
**11. Clustering Decision**

Overall, we proposed four Clustering results. As a baseline model, K-means

results show the least informative Clustering results as expected. Our expectation on the Clustering results is the Neural Network Clustering will outperform all the other Clustering methods with enough depth and complexity. Due to the time limit of our project, we do not have enough time to discuss and design a sound architecture of Clustering Neural Network and tune the hyperparameters accordingly. We are left

with a choice between Hierarchical Clustering results and DBScan results. Generally, these two results are showing similar level of dissimilarity between clusters. We suggest the DBScan as by far the best results since it shows more characteristics on Teams activity which is the focus of our project and can be further utilized into GUI visualization.

**12. Persona GUI**



This GUI is built based on the survey results and our clustering results (DBScan results used as in the Demo). It mainly has two functionalities. First, on the left side manage frame, user can search “Division”, “Teams Frequency”, “File Sharing”, “Teams Availability”, “Active Level”, “Communication Level”, “Responsibility Level” and “Skills” based on “Email”. This search function is activated by pressing the <ENTER> button on your keyboard. After that, there exists a skill search detail frame. Customizing for employees’ talents acquirement purpose, this functionality will enable user to get all the employees within MAPFRE equipped with required skill. Once user type in the skill

in the entry box above and hit <ENTER> button, the treeview box below will show the details of all the employees who have the skill. In addition to that, a click event is created within this frame. Once user clicked one of the returned results in the treeview below, the manage frame on the left side will automatically update with the clicked employee’s Persona information. An instruction will be delivered as a separate file just to illustrate two ways of executing this GUI.

**References**

[1] MAPFRE Kickoff Meeting PPT

[2]        MAPFRE GQP PDF

1[KERAS API](http://www.keras.io)

2*Key Performance Indicators for dummies* by Bernard Marr

3 “A General Coefficient of Similarity and Some of Its Properties” by J. C. Gower

4 [ARIMA Models for Time Series Forecasting](https://people.duke.edu/~rnau/411arim.htm)

5[Facebook PROPHET Python API](https://facebook.github.io/prophet/docs/quick_start.html#python-api)

6 [Microsoft O365 Activity Report Value Mappings](https://docs.microsoft.com/en-us/microsoft-365/admin/activity-reports/activity-reports?view=o365-worldwide)