

Final Report

Northeastern University - ALY6980-20154

Instructor: Thomas Blanchard

Maryam Heidari

Mengmeng Du

Sony Kalita

Haoyuan Zhang

March 27th, 2020

**Introduction**

We are committed to optimizing the scoring method of the Enaible Productivity Score and trying to solve some potential problems that this application may cause, because the biggest feature of AI is to achieve self-learning by accumulating experience. AI can't consider some other existing problems. There are two typical problems here: measurement standards and privacy protection issues. For example, as people's work styles change, many people tend to work in their residences. Therefore, the AI ​​system needs to set some standards by determining the working hours by recording the working hours. Regarding privacy issues, because AI needs to obtain information for analysis, it may be able to gain access to certain information of employees, such as employees' web browsing records and emails. In order not to infringe the rights of employees, we must make certain settings for the system.

According to the characteristics of AI to realize self-learning by accumulating experience, we need to make some adjustments. For example, for different types of work, the system needs to set some standards for their productivity measurement. According to some data collected by the company before, we found that some types of work have significant differences in actual working hours and frequencies, such as sales supervisor. However, it is difficult for AI to distinguish. Thus, we need to make settings in the system after research.

Our reports are mainly divided into background and literature review, exploratory data analysis (EDA), correlation and forecasting. We will try to understand other applications of AI through literature review, market research and other means. We will summarize our findings and use the results in our optimization of The Enaible Productivity Score.

**Background**

The Enaible Productivity Score is the world's first comprehensive standardized productivity score. The Enaible team can use the existing system data of the client company to quantify company and team-level productivity through multi-dimensional calculations of capacity utilization, consistency and quality impact.

The Enaible team's AI Trigger-Task-Time ™ algorithm is the key to accurately measuring productivity. The leadership Trigger-Task-Time ™ algorithm is a breakthrough in the combination of leadership science and artificial intelligence. It takes into account complexity, order, internal and external factors, patterns, time of day and duration to provide the most accurate in the world Productivity score.

More and more companies want to know more about their employees. For company management, employee productivity is an important indicator. This will help management make better decisions about where employees will stay. Thus Enaible has a broad market.

**Literature review**

In modern society, more and more people believe that privacy protection has a great impact on employee loyalty and productivity. Some people very highlight the impact of privacy on work productivity across different generations. A group of researchers tried to understand the impact of factors such as privacy on employees. They found that Three variables emerged as impacts of office redesign perceived by respondents: friendship, collaboration and privacy(Haapakangas, M.Hallman, Mathiassen and Jahncke, 2018). Collaboration and privacy exert a positive influence on work productivity, while friendship does not. The relationship between privacy and work productivity is stronger for the Generation Y than for senior employees, namely, the Baby Boomers and Generation X. Actually, for example, one thing I want to say about monitoring is that companies are not a good idea to do so, but many companies are trying to use AI to monitor employees. We can definitely find other ways to protect the privacy of our employees. Simply collecting data or logging may be a good way.

In today's human resources (HR) management, there is still a big gap between people's hope and the reality of AI use(Tambe, Cappelli, Yakubovich, 2019). AI currently uses data science and technology in human resource tasks facing four challenges: the complexity of human resource phenomena, the constraints imposed by small data sets, issues of accountability related to impartiality, and other ethical and legal constraints, and employee adoption Data-based behaviors can have adverse reaction algorithms to management decisions.

We also need to further study the impact of management on employee productivity.

**Data Cleaning**

In this step of data cleaning, we will be checking all the 7 files available, if the data are in correct format, is there any missing values present and if yes, we will be imputing the missing values with the appropriate method such as by replacing with “0” or by average value. Further, for analysis we will be joining the dataset according to the requirement and split the variables if required.

Through the libraries such as “pandas” and “numpy” for importing and cleaning the datasets. We have imported 7 data files and view few of the records present in the data file.

**batch\_scores**: this is the daily time series of each employee’s score.

**wheels**: this is the autotimer tracker file indexed by day (yyyy-mm-dd) and the columns 0 to 1439 correspond to the minute of the day from 00:00 to 23:59.

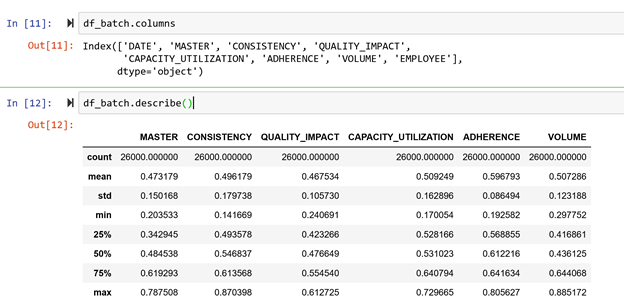
**quanta\_map**: this is the mapping from triggers (string) to trigger groups (int)

**allocator**: this is the results of the static allocator.

**Ideal\_times**: this is the ideal times computed for each trigger\_group.

**Employee**: this consists the information about all the employees

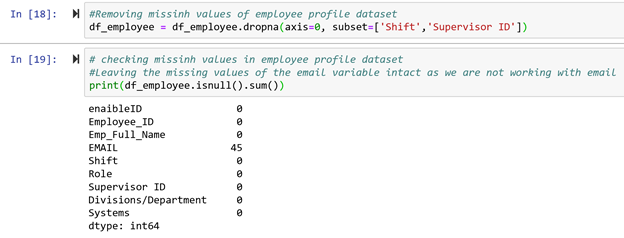
**Preprocessed\_data:** It consists of the employee name, Id, trigger and the trigger time allotted to each employee



The figure above shows the variables present in the batch\_score dataset and we can view the descriptive statistics of the dataset such as the mean, median, std, max etc.

After getting all kinds of data, we checked whether the batch\_score dataset contains any missing values, here the dataset does not have any missing values.

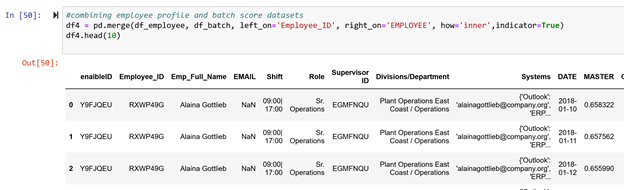
We performed similar steps for other data files. We have checked the type of data held by each file and whether any values are missing. For the employee dataset, we are missing values in the email, shift, and supervisor ID columns.



In the figure above, we have dropped the single row from shift and supervisor id variable to make the data consistent and we have deleted the email variable from the dataset later when we are joining the tables employee and batch\_score. As we are not using email for our analysis and we cannot impute the email id missing values, so we have decided to drop the variable.

After that, we have splitted the variable trigger/department of the allocator dataset into two separate variables for making analysis easy and understandable. Then we have replaced the missing values of the wheel’s dataset with “0”. As per instruction the wheels empty value NAN can be treated as “0”.

Finally, from the figure below, we have merged the employee and batch\_score dataset based on Employee id for further analysis.

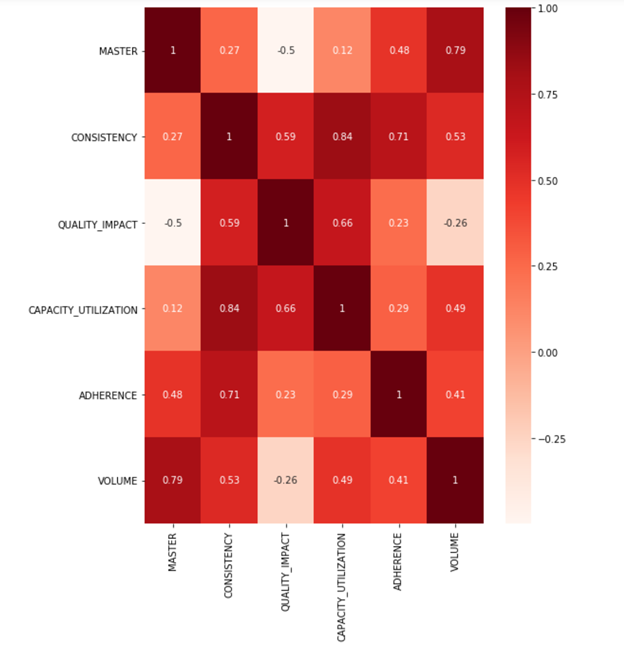


In the figure above, we can see that columns 'EMAIL, Shift, Systems, EMPLOYEE, \_merge' have been dropped after merging both the dataset as we want to keep the attributes that we are going to use for our analysis. Further, we have splitted the column “division/department” into two separate columns for analysis.

After merging the two datasets, the "EMAIL, Shift, Systems, EMPLOYEE, \_merge" columns have been removed because we want to preserve the attributes that will be used for analysis. In addition, we divided the "Department / Department" column into two separate columns for analysis.

**Correlation**

Correlation is used to test relationships between quantitative variables or categorical variables. In other words, it’s a measure of how things are related. The study of how variables are correlated is called correlation analysis. Correlations are useful because if you can find out what relationship variables have, you can make predictions about future behavior.

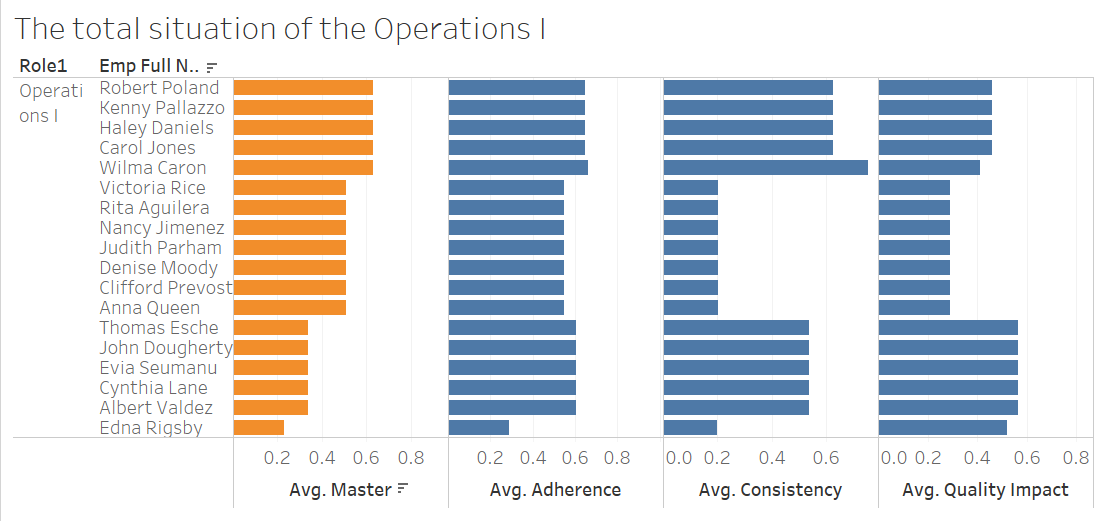
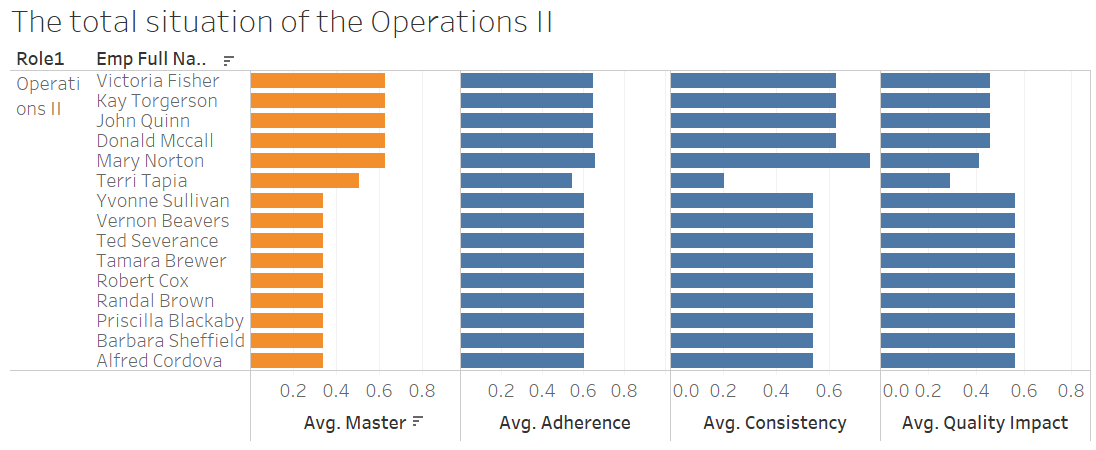


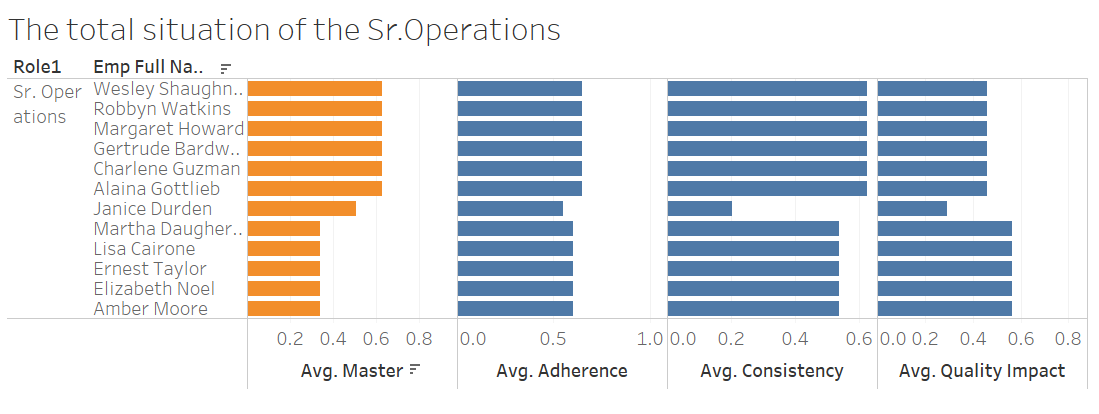
From the figure above, we can find that the “Master” variable has positive correlation with volume, adherence, capacity\_utilization, consistency and negative correlation with Quality\_impact attribute.

**Exploratory Data Analysis**

The data we studied came from data provided by Enaible. These data introduce us to some of the current situation of customers of Enaible and the measurement of productivity of Enaible for some company employees. With EDA, we try to make an overall assessment of the situation of these customers and analyze the factors that may affect employee productivity.

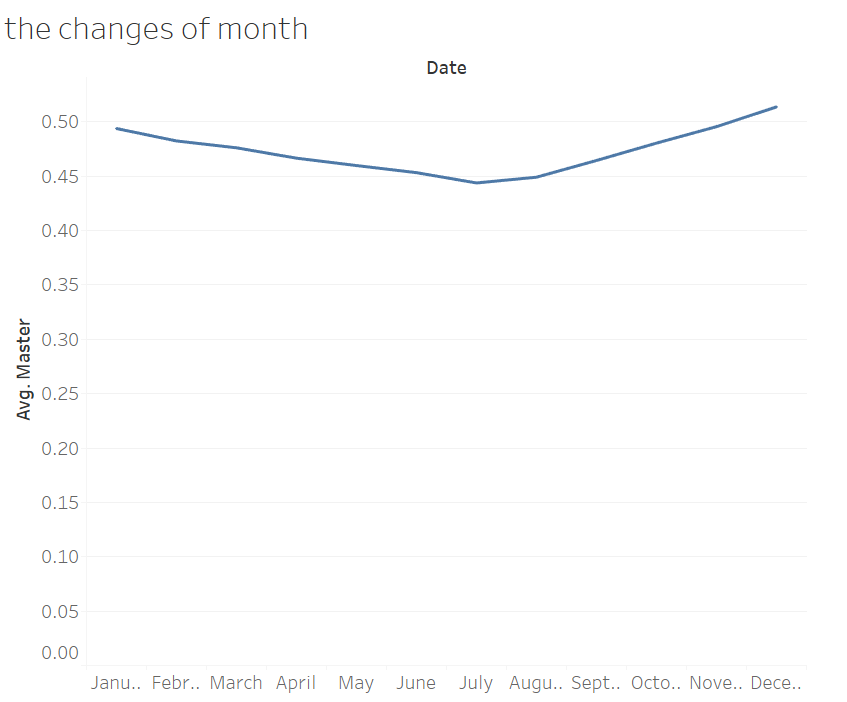
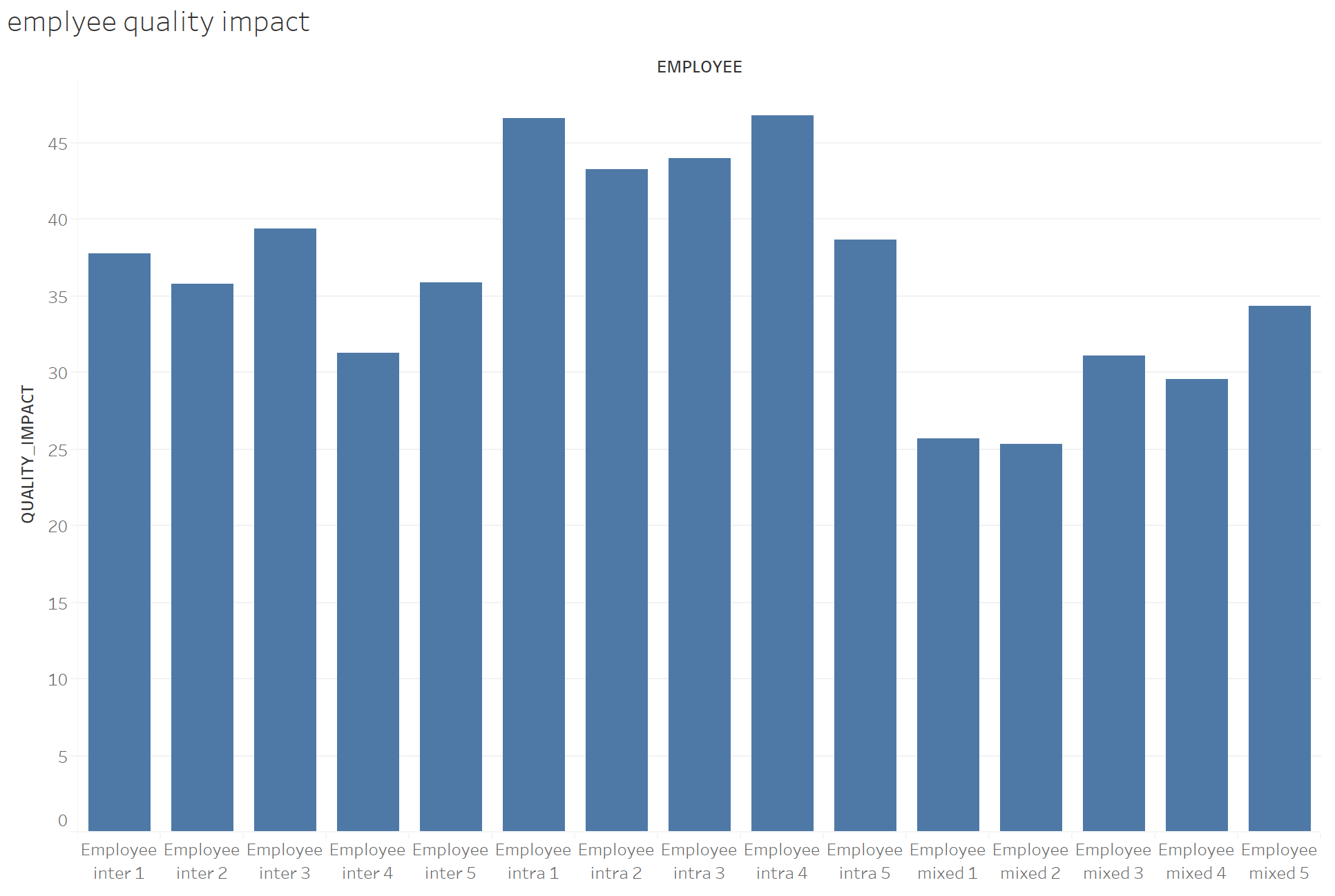
We made an overall analysis of the operations that appear in the data. Their average adherence is all in the 0.6-0.8. This is a point worthy of attention. Because the surveyed operations are in different industries and their working frequencies are very different. We can do further analysis. Overall, the company should pay attention to the employees whose scores are not very good but others are good.





We also tried to study employee productivity through time series. Unfortunately, we don't get many regular results. However, we found some interesting points. In June and July, the productivity of the surveyed employees reaches its lowest level in a year. When the time comes to December, employees' productivity is still increasing. In general, as the Christmas holidays approach, employees may have more absences than usual (because of holiday arrangements, etc.), resulting in reduced employee productivity. We should focus on these points.

In "batch scores. csv", there is an important value "quality impact". We can understand that it is a score for employee productivity. We found that for different types of employees, their quality impact will vary greatly. We can see this in the picture below. We found that compared to “inter employees” and “mixed employees”, intra employees had a higher quality impact. We need to further understand the working characteristics of intra employees. Their working characteristics will be valuable for helping other companies improve employee productivity.



**Time series and Forecasting**

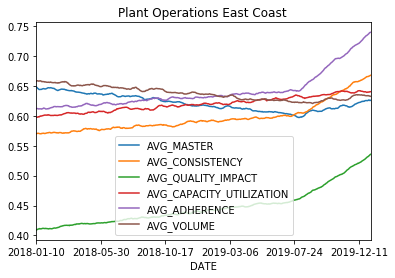
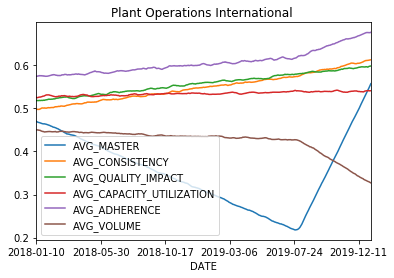
In most of the companies what matters is the productivity of the team.

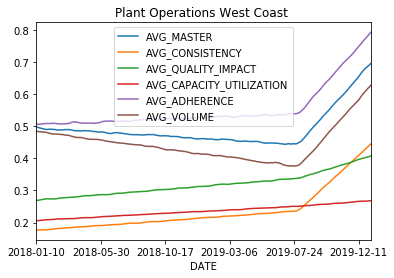
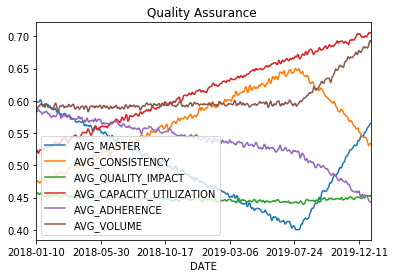
In our opinion we need to analyse the team productivity and then if we see any problem with a team, we need to dig in more and analyse the individual productivity.

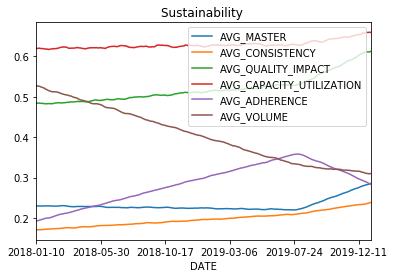
All of the data that we have belongs to the operational department. We assumed we could find each team based on the divisions and each division belongs to one team.

Then I plot the time series for each team once by average of the values in each column of the batch\_scores and then by standard deviation.

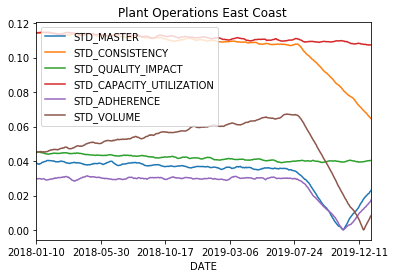
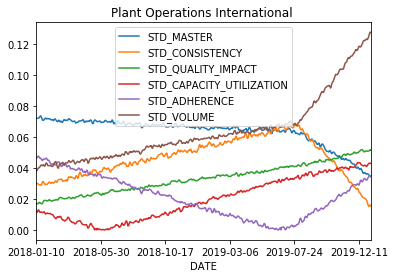
The result for average is:

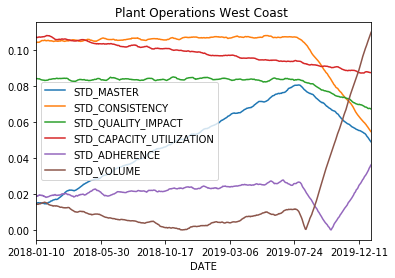
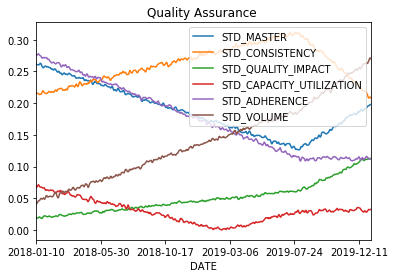
 

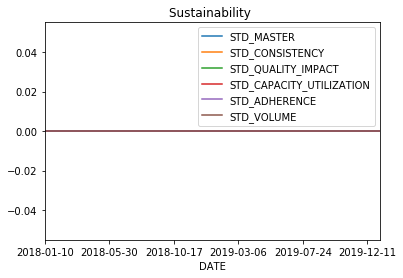
 



And the result of standard deviation are:

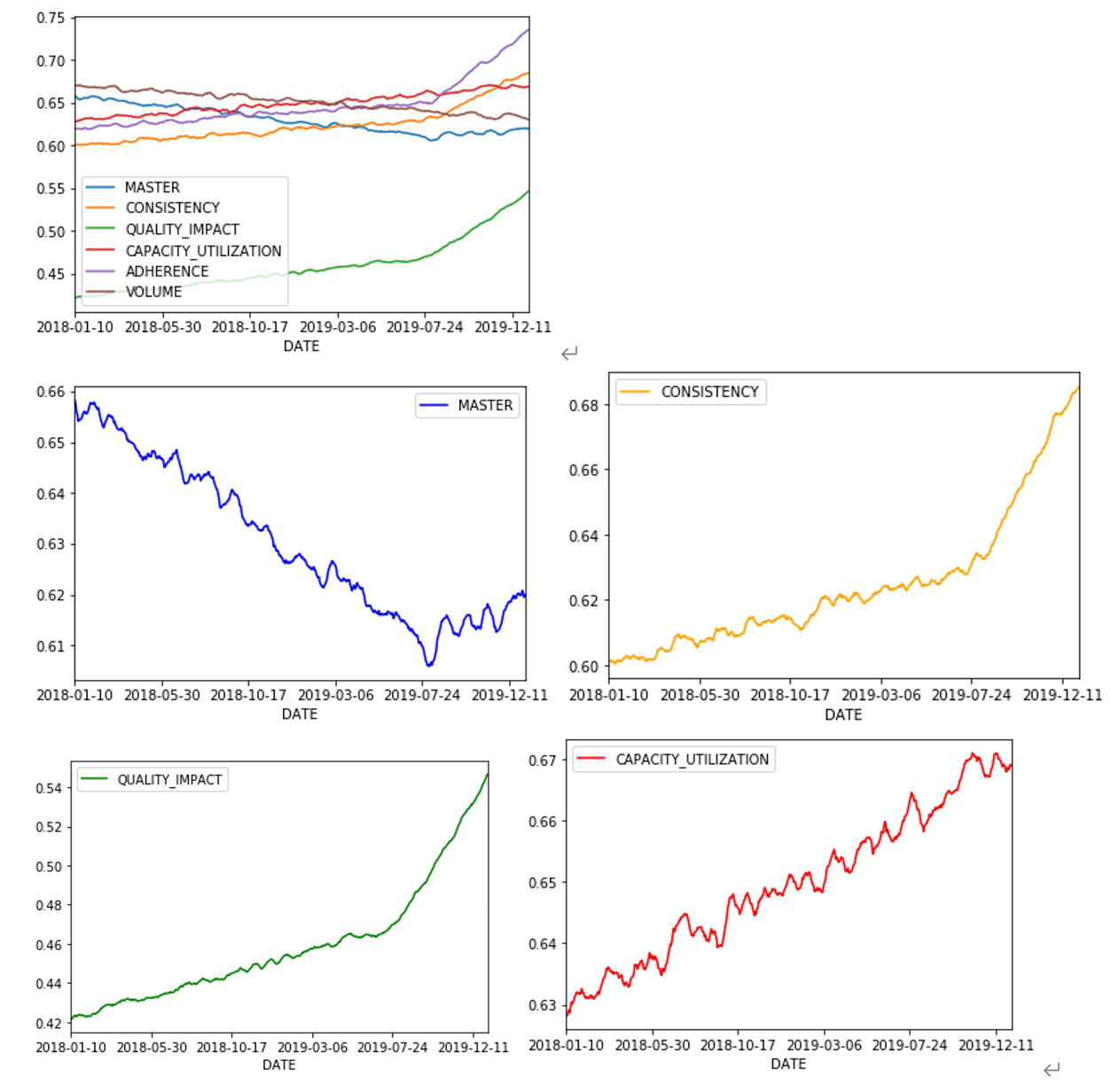
 

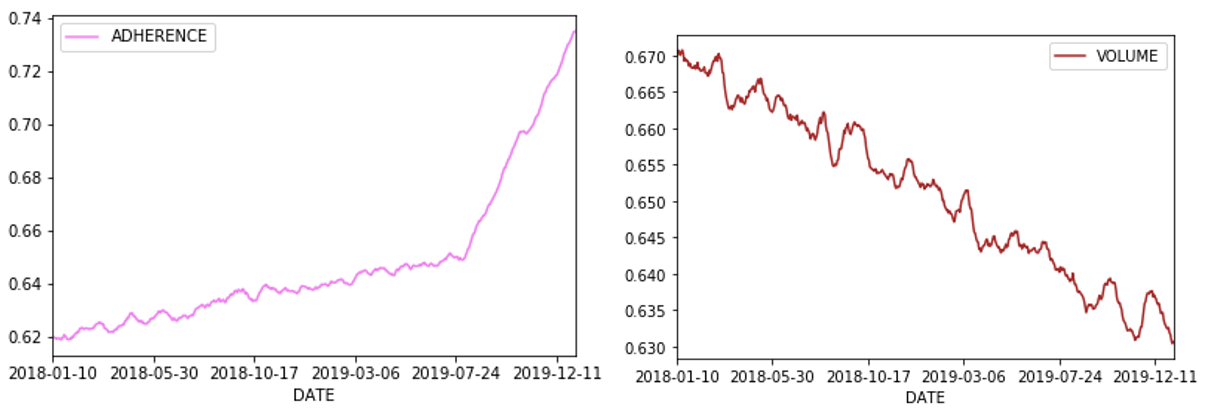


Based on these plots, the employer can decide which group has a problem and which team they want to dig in more by going through each employee ID productivity that belongs to that team.

For the individual time series, I decided to plot all the information in batch\_scores for each employee id to understand if there is any trend or seasonality. I do that for “RXWP49G” as an example, the company can make an interactive graph that changes by changing the employee id.

The result for RXWP49G ID is:





As you can see except the volume and CAPACITY\_UTILIZATION , the others don’t have obvious seasonality.

In general, the development of many events is related to the development of time. Therefore, forecasting a time series is often of great commercial value. Our research and prediction are also based on this reason. Based on the data set, we can find that there is no seasonality in this time series. Therefore, in order to make predictions on this data set, we decided to use the ARIMA method. This is a prediction algorithm based on the idea that the information in the past values of the time series can be used alone to predict future values. Any `` non-seasonal '' time series exhibiting a pattern and not random white noise can be modeled using the ARIMA model.

An ARIMA model is characterized by 3 terms: p, d, q. The p is the order of the AR term, q is the order of the MA term, d is the number of differencing required to make the time series stationary.

In general, the development of many events is related to the development of time. Therefore, forecasting a time series is often of great commercial value. Our research and prediction are also based on this reason. Based on the data set, we can find that there is no seasonality in this time series. Therefore, in order to make predictions on this data set, we decided to use the ARIMA method. This is a prediction algorithm based on the idea that the information in the past values ​​of the time series can be used alone to predict future values. Any `` non-seasonal '' time series exhibiting a pattern and not random white noise can be modeled using the ARIMA model.

There are several main steps in modeling through the ARIMA model:

1. To see whether series are stationary or not.

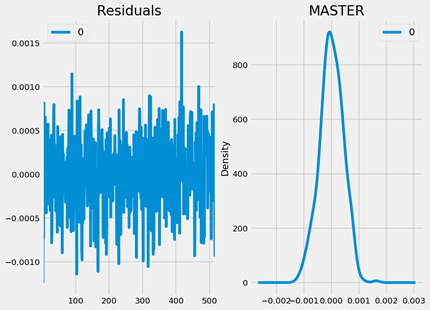
2. To identify whether the model needs any AR terms. We can find out the required number of AR terms by inspecting the Partial Autocorrelation (PACF) plot. Through this step, we temporarily fix p to 0.

3. To find the MA term. An MA term is technically, the error of the lagged forecast. We can look at the ACF plot for the number of MA terms.

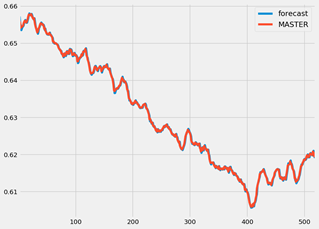
Finally, we made a model for ARIMA for d = 2, q = 1 and p = 1. The model summary reveals a lot of information. The table in the middle is the coefficients table where the values ​​under 'coef' are the weights of the respective terms.

here the coefficient of the MA term is close to zero and the P-Value in ‘P>|z|’ column is ideal because it should ideally be less than 0.05 for the respective X to be significant.

And below, you can see the plot of the residuals to ensure there are no patterns:



The residual errors seem fine with near zero mean and uniform variance. Let’s plot the actuals against the fitted values using plot\_predict().

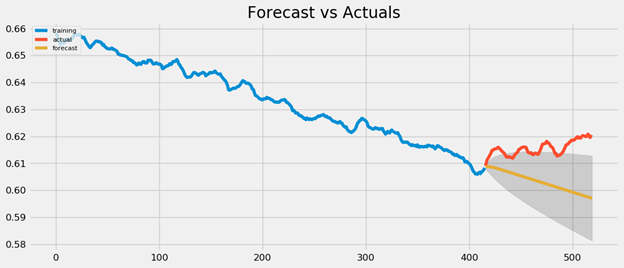


When you set dynamic=False the in-sample lagged values are used for prediction.

That is, the model gets trained up until the previous value to make the next prediction. This can make the fitted forecast and actuals look artificially good.

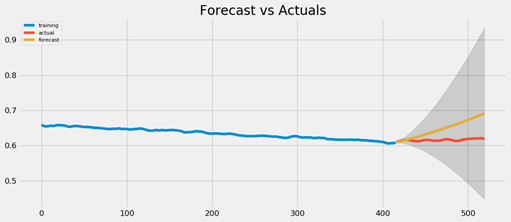
Now that we have the model, it is time to make the forecasting model.

To do so, we need to train and test the data set. I divided the data set to 80% train and 20% test. Now we can build the ARIMA model on the training dataset, forecast and plot it.

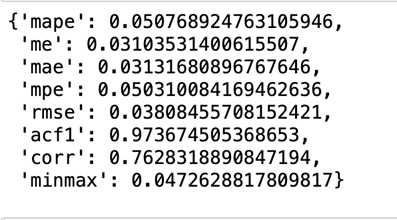


As you can see the forecast line is not even close to the actual line, so I change the ARIMA model to model = ARIMA(train, order=(3, 2, 1))

And the result of the forecast is:



And now, to measure the accuracy of the forecast:



Around 51% MAPE implies the model is about 94.9% accurate in predicting the next observations. Now you know how to build an ARIMA model manually.

**Conclusions and suggestions**

1. We recommend that, in order to increase productivity, company management should increase the adherence, capacity\_utilization, and consistency of employees, and then reduce the impact of quality impact on employees. Enaible's AI system also needs to adjust the scoring proportions and scoring rules for all aspects of employees. This will lead to a better measurement.

2. From the time series of the whole year, the productivity of employees during the summer (June to July) is at the lowest level of the year. Companies can think of activities in the summer to increase employee motivation. In combination with the findings of the literature review, we also recommend that companies create a better working environment for their employees.

3. Since most of the observed employees come from sales positions, we suggest that companies can add more benefits to sales positions to increase productivity. Enterprises should pay more attention to some employees who have poor overall scores but are very productive to get some methods.

**Reference**

Prasanna Tambe, Peter Cappelli, Valery Yakubovich (2019). Artificial Intelligence in Human Resources Management: Challenges and a Path Forward. Berkeley Haas, Volume: 61 issue: 4, page (s): 15-42.

Annu Haapakangas, David M.Hallman, Svend Erik Mathiassen and HelenaJahncke(2018). Self-rated productivity and employee well-being in activity-based offices: The role of environmental perceptions and workspace use. Building and Environment, Volume 145, November 2018, Pages 115-124.

**Appendix**

**Tableau**

Figure 1

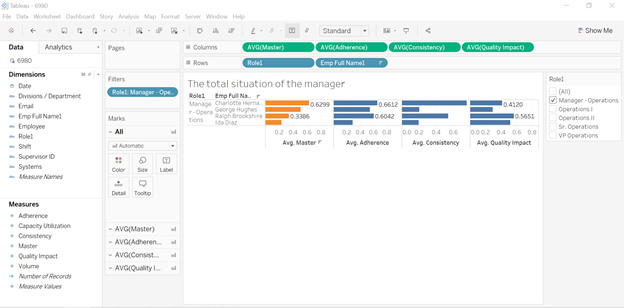
****

Figure 2

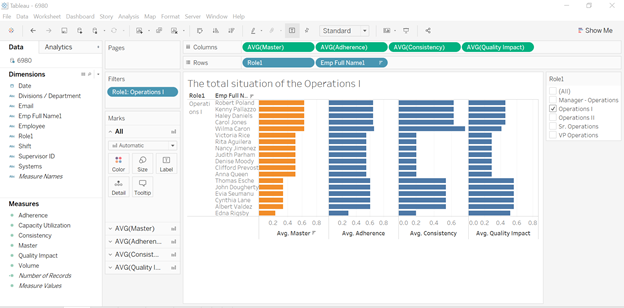
****

Figure 4

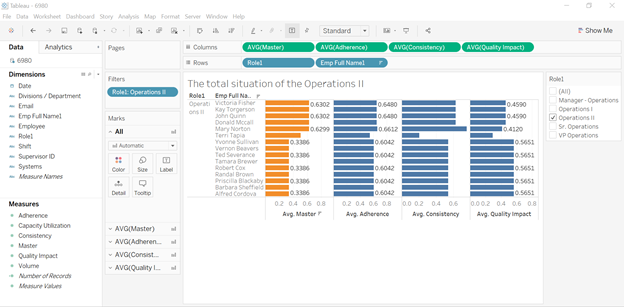
****

Figure 6

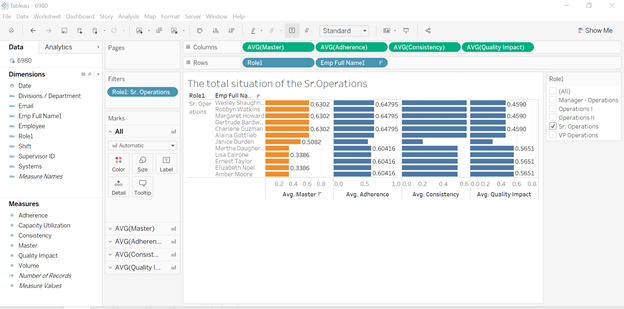
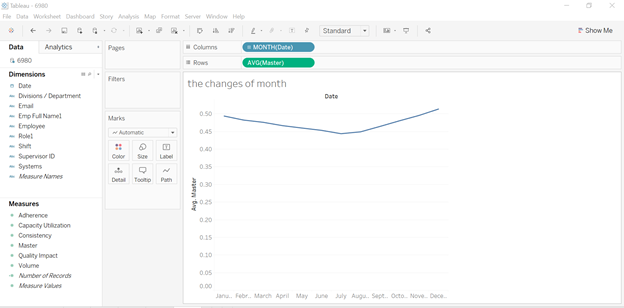
****

Figure 7

****

**Figure 8**

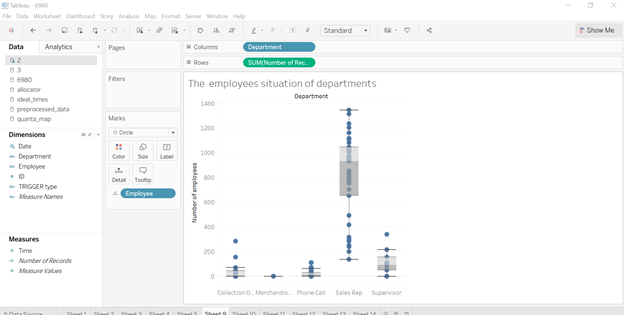
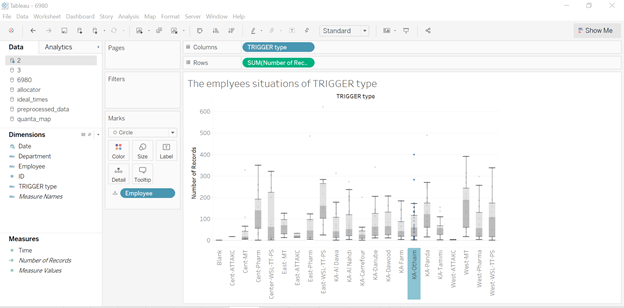
****

Figure 9

****