3. WebEx Meeting Data Analysis

**1) What is the function of WebEx meeting KPP?**

KPP = Key Performance Parameters

KPP = Key Performance Parameters

The most important KPP we care about is JMT(means join meeting time) which means how long did a customer join a WebEx meeting when he clicks the join meeting button. The JMT consists of ClientJMT, CbServerJMT, PageJMT, PagePluginJMT, gpyDoneJMT and GpcJMT. Different teams are responsible for different JMT.

My work is helping them to analyze JMT on our data analysis platform(Platfora). For example, the change trend of JMT on different WebEx version, operation system, browser, region and so on.

The visualization vizborad can help them find the potential problem for their JMT which may help them debug or optimize the software.

**2) Data analysis platform**

In order to output the analysis dashboard, there are four steps on Platfora.

1) Define Data Sources to Connect to Raw Data in Hadoop.

Usually, I need to write SQL on Hue to generate the required raw data.

2) Create Datasets to describe the structured of the data

3) Build a Lens to pull data from Hadoop into Platfora

It submits a series of MapReduce jobs to Hadoop, collects the results, and brings the results back into Platfora. From the perspective of an ETL workflow, the lens build is the load part of the process.

Lens build is time-consuming and one lens may depend on multiple Datasets. Therefore, I should carefully choose how to build a lens for future analyze.

4) Create vizboards to analyze and visualize the data

Visualizations can take various forms such as charts, graphs, maps or cross-table. Then I can share the dashboard with relevant teams to help them reduce workloads.

**3) About the "Data Extraction and Migration", data scale, difficulties you met and how to solve it.**

Since we have test and production environment for our data platform. A Data analysist usually create dashboard on test environment and then move to the production environment. A data migration tool can help them quickly do this job.

At that time, Platfora was acquired by another company named Workday. So, Ploatfora will not be updated and customer responses are slow. In the latest version of the document, it support REST API but did not have detailed document for this. So I have to do this step by step.

1. Called the API successfully to download the JSON file for a simple dataset, lens,

Vizboard.

1. Understand JSON file meaning and then change then content of it. Finally, put the

JSON file to production environment and created a simple dataset.

1. Successfully download a dataset/lens/vizboard JSON file from test environment, change

the content and then put it to production environment.

1. Finally, successfully do the same job for a complicated dataset/lens/vizboard.

The whole process like read, guess, test. Also I have to do many tedious work, like naming rules unification on test and production environment.

**5) About the "Fault Prediction System": what does data look like? and what is the targeted "fault"?**

Fault WebEx meeting means the JMT is larger than the experience value, I labeled this part of data as Fault WebEx meeting and the other as normal WebEx meeting. When I first tried to do this, I was going to use the decision tree to predict fault WebEx meeting since it is simple to understand and interpret.

Firstly, I have to extract and parse the log files from different data sources since different team store log fires in different ways. Then, I transformed the data to CSV files and each row contained all the features for a meeting.

In the beginning, the number of the feature is too large with many problems like the same repeat features categorical features, unique value features, low variance features. At that time, it is better to use Random Forest. Finally, I can set the hyperparameters for the random forest to improve the training data results. The hyperparameters includes ‘max\_features’, ‘max\_depth’, ‘min\_samples\_split’, ‘min\_samples\_leaf’, ‘n\_estimatiors’, ‘bootstrap’.

Afterward, I realized that feature engineering is the most difficult part when using Random Forest for fault prediction. Some part of work can be done by myself like drop the repeat features, unique value features, and low variance features. When I training the model, I found that category encoding is the key point. One-hot encoding is not a good solution compared to encode by experience engineer. So my experience during this project is that iterative feature engineer and test to get better results.