**Main progress**

* Data column clarification
* Data Classification
* Data Exploration – Probability accuracy test

**1. Data column clarification**

In the phone meeting this week, Mariem kindly explained several variables that I used before are not in the meaning that I thought before.

Coverage: It means the sales person in Credit Suisse who deal with this case. Thus this information is related to the sales people in Credit Suisse instead of customers

Business\_Stages: Business stage can’t tell me if the deal is closed or not. There is 4 stages and 1 status in the dataset.

Status: Only Status can tell if the deal is closed or not. Thus the status should be the target variable. To be more clear, rows containing “win” in status column mean the deal is won by Credit Suisse and rows containing “lose” in status column mean the customer refused to make this deal with Credit Suisse.

**2. Data Classification**

Since now I have better understanding of the dataset, I figured it out that we have mainly three kinds of variables that influence the success of winning a deal

1. Customer Variables: Variables that contain the information of customers, such as amount

2. Product Variables: Variables that contain the information of the product, such as product name and product level

3. Coverage Variables: Variables that contain the information of the sales person, such as the region and id. Sometimes there is just 1 coverage and sometimes there are 2 coverages. Id can be used as categorical variables here because the unique id number is not big.

I divide them into 3 big categories because I think they influence the probability of wining a deal interactively. Thus in order to determine the effect of different factor(customer, product or the sales person)

**3. Data Exploration – Probability Accuracy Test**

Now I’m able to do some really interesting analytics. The first thing I did is to change the target variable into a binary form. “1” means winning the deal an “0” means didn’t get it. I’m interested in evaluate the accuracy of the probabilities. The reason I want to focus on it is that “Gauge the probability” is on the task description of the project and a stuff in the Credit Suisse has mentioned that the probabilities they set now is relatively subjective. They would like to have it to be more accurate and scientific. And from my knowledge I also know it’s a big deal to get more accurate probabilities since better probabilities provide better forecast and Credit Suisse will know how much net asset they can get, thus making investment strategies in advance.

I find the probabilities only have 6 outcomes. And after group win and lose number by the probability, the result is very interesting. Nearly only the record with 100% probability will be won. The other probabilities only have several win case and most of them failed.

Personally, I think the reason behind this is the record only have the final status and doesn’t have any history data here. Which means they will change the probability to 100% if they win the deal. Otherwise, they either keep the original probability or change it to below 100%.