**Week of March 18th- March 30th**

**Blog 6: March 30th, 2019**

In these weeks, we took the meeting with both professor Jason and Mariem. Both of them gave us great ideas. We found that the largest dataset is really hard to use because there’s only one “data\_change” column contains all change information. But if we would like to use it, we need first to match the index with them. However, this dataset is so dirty that we cannot correctly match all of them. So our decision is to just use the rest of them, which are “Atlas.csv” and “AmAPAC.csv”. Atlas dataset contains the historical information before the system changes. For the CRM system of Credit Suisse, it has been moved from old system to new one on June 2017 and Atlas only keeps the change information before that date. And AmAPAC dataset contains the latest information of all transactions. We can use old opportunity id to match them. We decide to descriptive analysis first.

Our descriptive analysis focused on some key areas like: usage of salespeople, region and coverage distribution and product level. My mainly focus is try to build a binary-classification model to predict whether a certain deal will won or lost. It is a very simple question. But one problem is that I have to work on the AmAPAC datset because Atlas dataset is dirty. But since AmAPAC is the latest information of all transactions, I feel worried about including some future information.

Firstly, I worked with Doris and picked some variables that we think will have the predictive power. Then we put all of them into the logistic regression and random forest model. We found that the model performance is extremely good and accuracy reaches 98%, which makes me feel even more worried about including some future information. And Doris applied two model in R and calculate the feature importance of all these 50 variables. Unfortunately, we found almost all of them are significant.

Also I found I used a wrong measure matrix. I should not use accuracy because our dataset is really unbalanced that the lost deals contains almost 76% of the whole dataset. In other word, if we use a naïve model and set the all prediction value to be 0, its accuracy will still reach 76%. I decided to use AUC-ROC curve. AUC - ROC curve is a performance measurement for classification problem at various thresholds settings. ROC is a probability curve and AUC represents degree or measure of separability. It tells how much model is capable of distinguishing between classes. After using the AUC-ROC curve, I can get a better measurement.