**Week of February 18st - February 24th**

**Blog 4: February 22th , 2019**

During this week, our group put emphasis on the initial data processing part. Since we don’t have any variable description, we have some question about these variables.

* What’s the meaning of won-unfunded?  - win the deal but still wait to receive money
* Salesperson change the close date multiple times - which is interesting and worth digging into
* What should be our target variable?

After discussion with Mariem, we figured out that the target variable is “status\_code”, which has six categories: dual-dillgence, won(unfunded), won(funded), lost and so on. The key is that we can conclude all categories excepted won/lose to be outliers. We also use our dataset to test that. We found over 80% of the data’s status is either won or lose. So we make a reasonable speculation that for the APAC dataset, it is a slice of current time. In other words, it only contains the latest information of each transaction.

Another two dataset is the historical data, which includes all the historical changes. For example, the change of amount, probability, stage and so on. One of the key difficulties our team met is to merge these three datasets. Also we have questions about these two historical datasets as follows.

* How much money they want to invest- so the money value changes-
* Old value and new value- for eg changed from 5%-25%
* Is something always changing these stats?
* Now when something went to a higher probability- multiple rows

When we try to merge these three datasets, one of the keys is to find the key index of each file. Then we will meet the 1 to many merge problems. We found that one of the historical dataset’s index is opportunity id, which matches with “previous\_id” variable in the APAC csv file. So I check the time of the each file and found the data of one of historical file is from 2009 to 2012 and the APAC file is from 2007 to 2019. I guess that the company’s database system had update in 2012 and that’s the reason why APAC csv file has a column named “previous\_id”, which matches with the historical dataset’s key index. By using the key index, we could try to merge them.

But there exists one other problem that, how can we solve it as a time series problem. Time series problem means that we should reconstruct the dataset to be a clean and elegant dataset with each line of data has the same time interval.

We will work on it next week and I believe we will have some progress!