**Week of March 31st- April 18th**

**Blog 7: April 18th**

In this week, I keep working on the modeling part. It is really interesting that we use AUC-ROC curve as the new measurement matrix and find out some variables contains future information. For example, in AmAPAC once a deal is closed, if they are won, their probability will be changed to 100 and if they are lost, their probability will be changed to 0. That’s the reason why we cannot use “opportunity” as one of the input variables. Also, there are some variables, such as companylevelaid, contains very few information that does not have any predictive power. So I just deleted all of them. Finally, the input variables include salesperson, product information, which are never changed during the whole process of transaction. And what surprised me is that even just using these variables, model has a really good performance (89%).

But there’s also a problem. For all these five variables, I use get\_dummy function to transform all variables to be dummy variables so that I can use them in the model. But since the variables I use are product and salesperson information, it is very likely that they will be changed in the future. For example, salespersons leave the company or new salesperson come into the company will change the whole data input structure, which means the training model cannot be reused. I think about how to solve that problem. There’s a solution: if I have a cleaner historical dataset, I could use other variables without any concern about including future information. For example, I could use the opportunity in the first stage. And using it as an input will not be influenced by the change of products or salespersons.

Also for the visualization part, our client said we should put more emphasis on it. Our client also wanted us to document our entire progress from the start of the semester to now, for every step: data cleaning, modeling, PowerBI, challenges and next steps onto their Wiki. We had to update each of those sections comprehensively with screenshots and it had to be as detail oriented as possible so that it is easy for someone to understand all the steps when they refer to it at a later stage.

I updated my modeling part and worked on the data merge because professor Jason gave us a good advice that maybe we should use the historical dataset to build the model by merging AmAPAC and Atlas datsets.