**Week of February 25st - February 31th**

**Blog 5: February 29th , 2019**

During this week, we separate our group to be two teams because we have more than one tasks need to be solved. We could see it as a prediction problem and also try to find the relationship between variables and get some business insights.

In our group yuming worked on the second part, find the relationship between variables. It is interesting that he separated the dataset according to some variables and after he has different groups, he will try to analyze and compare the difference between groups.

And other teammates will focus other problems. We have already submitted a report to Mariem and she gave us some advice that red line means that question is not important, yellow line means that question is what they want but we have not mentioned and others are questions that they want and we also mentioned. I try to separate the question to each team member. In the past, our group has little progress because most of us are trying to figure out the data, the variable definition, business process, problem definition. After that, I think we could begin to solve the problem. But the tasks we met are small tasks, if all of us work on a certain task, it will be inefficient. Instead, if each of us work on one task, I believe we will have a better progress.

In this week, we have worked on the variable selection part. Since we don’t have the variable dictionary, we have to pick variables one by one. We picked the independent variables one by one from APAC csv file. There are also some confusing variables, such as: pricename, which is a binary variable that contains value ‘Yes’ or ‘No’; also ‘competitorstrengthweakness’, which is also a binary variable containing value ‘Yes’ or ‘No’. Without clarification, we cannot input those values in Logistics Regression.

During the meeting with Mariem, we asked the question about these variables and we understand about the relationship between three csv files. I just have the guess that we have a wrong understanding about the largest csv file. That file is just the merge result out of APAC and AuditHistroy without opportunity and do not have the same structure as the Atlas CSV file.

Doris and I have also worked on analyzing which model should we use. We decide to use Logistic Regression and Decision Tree model, compare the results and find the relationship between different variables. We decide to use Random Forests to overcome the over-fitting problem. As for SVM, it works by projecting your feature space into kernel space and making the classes linearly separable. The best thing about support vector machines is that they rely on boundary cases to build the needed separating curve. But it can be tricky to find appropriate kernel sometimes.

In the next week, we will keep working on the model construction and training part, try to find the relationship between different variables. After that, we will apply the model structure to more historical data and see the results.