**Blog 5: March 15th, 2019 (Week of Mar 1st - Mar 15th)**

During these weeks, our team moved from initial data preprocessing and explorative data analysis to modeling. Enzhe and I focused on one of the core problem: What are the factors influencing the success of a deal? Is there an identified cluster of timings or types of deals are being closed? What are the factors that are heavily correlated?

Since it is a binary problem, we planned to use classification methods. Considering that some statistically significant variables may not have business meaning, we decided to tore our problem into 2 parts. The first part is to pick up some variables from our business understanding and intuition, and construct logistics regression model to find the important factors. The second part is to input all the independent variables and use the most popular and accurate machine learning methods, like Decision Trees, Random forest, Support Vector Machines etc., to avoid missing important variables.

Since we don’t have data description, I 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.

In the meeting with Mariem today, I asked her about these questions and she will replied to us ASAP. Also she suggested us to ignore those binary variables since there are too many missing values in those variables. According to her suggestion, Enzhe and I created an initial logistics regression model. Initially, we input all samples we have into the logistics regression model rather than split samples into training and test dataset. The accuracy score for logistics regression is above 90%. We also found some significant variables. We will work on other machine learning methods by using cross-validation, and training and test dataset and see its performance before next Tuesday.

I also dug into the comparison of Decision Trees, Support Vector Machines and Logistics Regression. I found that logistic regression will work for classification problems where classes are approximately linearly separable. It is robust to small noise in the data and is not particularly affected by mild cases of multi-collinearity (with L2 regularization). However, it cannot handle too many categorical variables or too many features and relies on transformations for non-linear features. Decision Trees are suitable for classification problems where boundary is non-linear and can be approximated by cutting feature space into rectangles. It can handle non-linear features but it is highly biased to training set. So we may 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 much needed separating curve. But it can be tricky to find appropriate kernel sometimes.

So I will follow the suggestions:

1. Always start with logistic regression, if nothing then to use the performance as baseline.

2. See if decision trees (Random Forests) provide significant improvement.

3. Go for SVM for large number of features/ observations.