Project Proposals

**Team members:**

Yichun Liu (GSI: April Cho), Ruifan Liu (GSI: Nick Seewald), Ruochen Wang (GSI: Weijing Tang), Yuan Shi (GSI: April Cho)

**Presenter:**

Yichun Liu

**Dataset description:**

The data pertains to the recruitment industry in India for the years 2014-2016 and deals with candidate interview attendance for various clients (Kaggle). It contains a set of questions that are asked by a recruiter while scheduling the candidate (we will call these external factors in the context of this research), information about the candidates such as location, industry, position, gender, etc. (which will be referred as internal factors), whether a candidate is expected to show up on the interview, and whether a candidate actually attended the interview.

**Questions:**

We would like to predict the interview attendance of a candidate from his/her answers to the questions - the external factors. Beyond this, we would also like to know if there is any internal factor that will affect the candidates’ response. Furthermore, whether some of the candidate’s answers to the recruiter’s questions could be explained by those internal factors, therefore, implicitly affect the response. In this setting, we plan to analyze how much the external factors will affect the candidate’s decision, holding other factors fixed. After fully understanding the correlations between different predictors and their relationships with the response, we will be able to reasonably conduct prediction on candidate’s response, with respect to the selected recruiting questions.

**Tools:**

To answer our questions, we will build predictive models to predict if a candidate will attend the interview based on the external factors as well as internal factors included in the dataset. Firstly, we will perform various EDAs to gain a better understanding of the dataset. Then, we will use Naive Bayes Model, Logistic Regression, LDA, QDA and KNN as classifiers. We will also use dimensionality reduction methods, including forward/backward variable selection, to select valuable/deterministic recruiting questions, so as to improve our model and infer the usefulness of the variables.