**Variables**

[1] "Date.of.Interview" X

[2] "Client.name" X

[3] "Industry" E

[4] "Location" X

[5] "Position.to.be.closed"

[6] "Nature.of.Skillset" X

[7] "Interview.Type" E

[8] "Name.Cand.ID." X

[9] "Gender" I

[10] "Candidate.Current.Location" I

[11] "Candidate.Job.Location" E

[12] "Interview.Venue" E

[13] "Candidate.Native.location" I

[14] "Q1"

[15] "Q2"

[16] "Q3"

[17] "Q4"

[18] "Q5"

[19] "Q6"

[20] "Q7"

[21] "Expected.Attendance"

[22] "Observed.Attendance"

[23] "Marital.Status"

**Goals**

* Predict the interview attendance of a candidate from his/her answers to the questions - the external factors (attendance - questions asked).
* if there is any internal factor that will affect the candidates’ response (response - properties).
* 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 (external - internal).
* 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 (select significant externals which influence response).

**Methods**

* Boxplot to show collinearity between variables
* visualization, summary statistics, PCA(sophisticated dimensionality reduction method)
* Past methods:

1. Visualizations ---------------- General comparison of genres

2. Dimension Reduction ---------------- Components that predict a song’s rating

3. Classification ---------------- Using derived components for classification/prediction

4. Clustering ---------------- Users grouped via rating behavior

5. Limitations ---------------- Discussion of possible pitfalls in analysis

1. LDA, QDA (Ruochen)
2. Ridge (Ruifan)
   1. Should we omit na? -> 80% data remaining
   2. Only Ridge, does better in prediction -> no feature selection
   3. Is “finding significant variables” still of interests? -> feature selection
   4. train accuracy = 0.7662, test accuracy = 0.67
3. BACKWARD, FORWARD (Yuan) logistic-see hw5
   1. Best subset + logistic?

Different than Proposal

Other methods:

PCA/PCR unsupervised, try (select variables -> fit into logistic)

Clustering can try but eliminate the response...

Random Forest, ntree = 1000, select the overlap problems.

Adaboost can try...

Group lasso try

LDA/QDA X only numerical predictors

KNN not work very well... similarity same or different 0,1 may fail in large dimension