**STATS 415 Final Project**

Ruifan Liu, Yuan Shi, Ruochen Wang, Yichun Liu

1. **Introduction**

Interview lies in the center of human resource recruiting process, which requires extensive preparation from the company side. However, it is not guaranteed whether a candidate will attend the interview or not. Consequently, it is at recruiter’s best interest to know what factors might affect a candidate’s attendance to a scheduled interview. Can the actual attendance be determined by the questions asked by the interviewer? Or are there demographic factors of the interviewer that might weight in to the prediction? In this project, we aim at both predicting interview attendance as well as identifying the factors that affect it.

1. **Data Overview**

In this project, we analyzed data about the recruitment industry in India from 2014 to 2016. 23 factors were collected from 1233 candidates, resulting in a total of 28359 observations to analyze interview attendance. From these 23 factors we choose 17 as predictors along with 1 as response variable(Observed interview attendance). We eliminate other 5 variables because they look random and unrelated to the prediction of interview attendance, such as ID and interview date. Then we divided these 17 predictors into 2 groups according to the nature of the variables. The first group contains the basic information about the candidates and the interview, such as the candidate’s gender and location, and the type and location of the interview. The second group contains 7 questions that are asked by a recruiter while scheduling the candidate, for example, “Have you obtained the necessary permission to start at the required time”, and expected interview attendance predicted by recruiter.

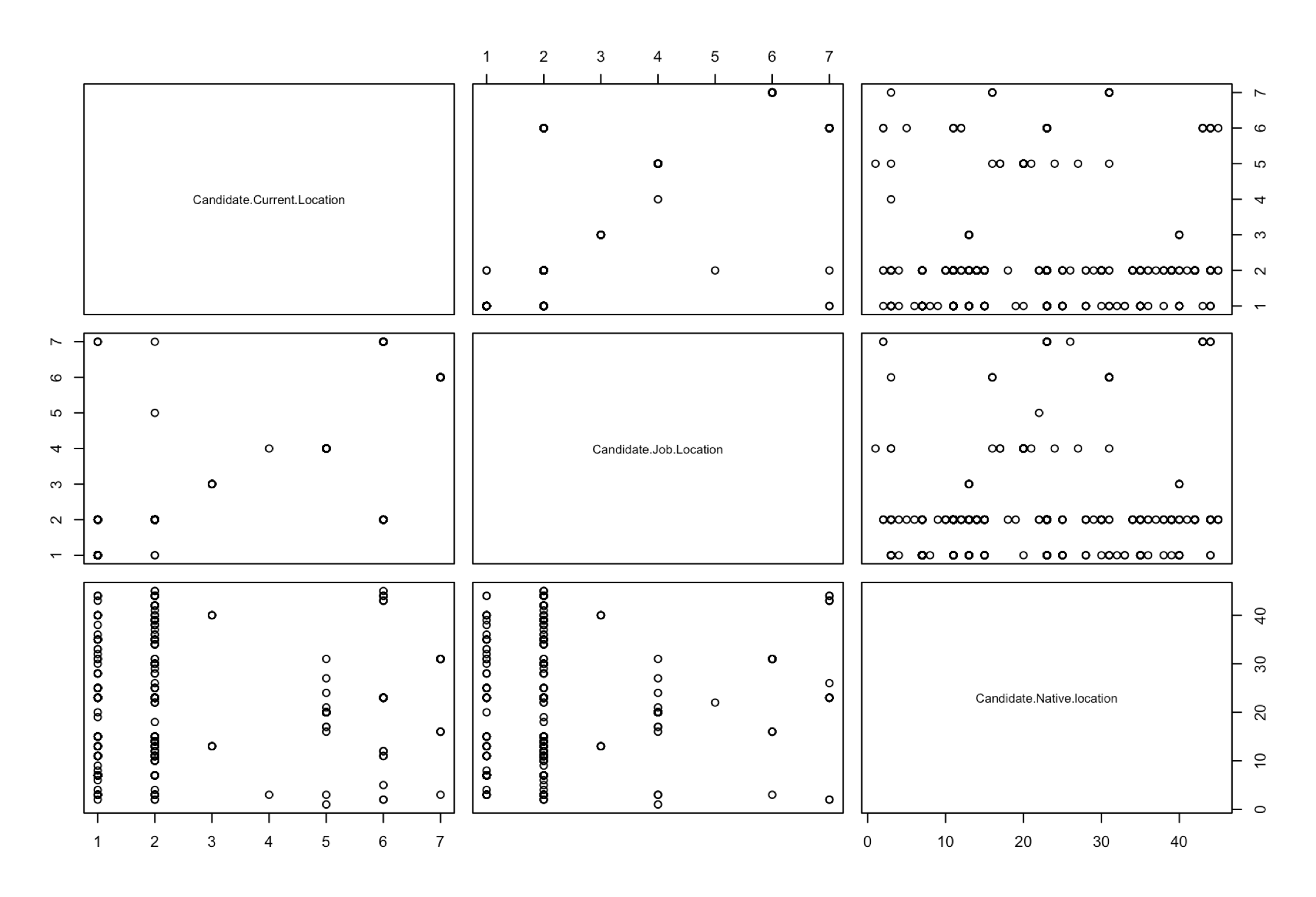
We conducted two steps in the data cleaning process. In our original data, we found that there are both capital letters and lower-case letters in the candidate’s response. For example, R identifies “Yes” with a capital Y and “yes” with a lower-case y as two different levels, which make no sense, so we had change all the responses into consistent lower-case letters. Also, when we looked at the data, we found that there are around 20% missing data in our predictors. Since our data are all about choices made by individuals, it is hard to predict the trend and fill out the data, so we eliminated these missing values.

1. **Methodology**

As a recap of the introduction section, we have 2 goals in this project. The first is to select the most reasonable model with the highest prediction accuracy. The other one is that we apply the model to two groups respectively to see which group, questions+expected attendance or demographic information, has more impact on interview attendance, and to find the most significant variables in both cases.

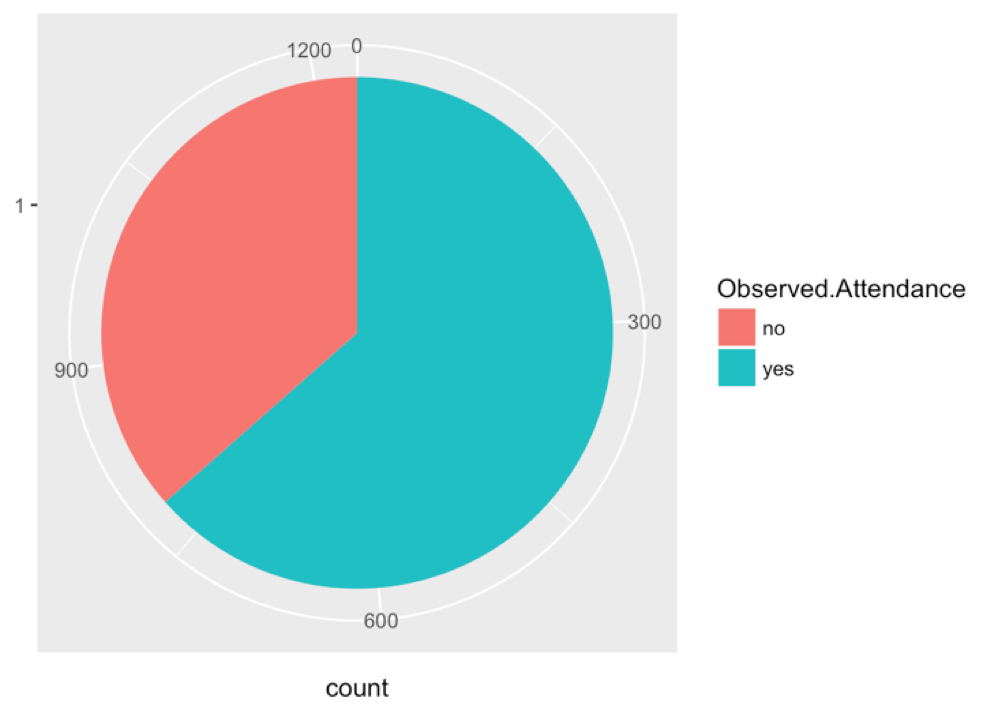
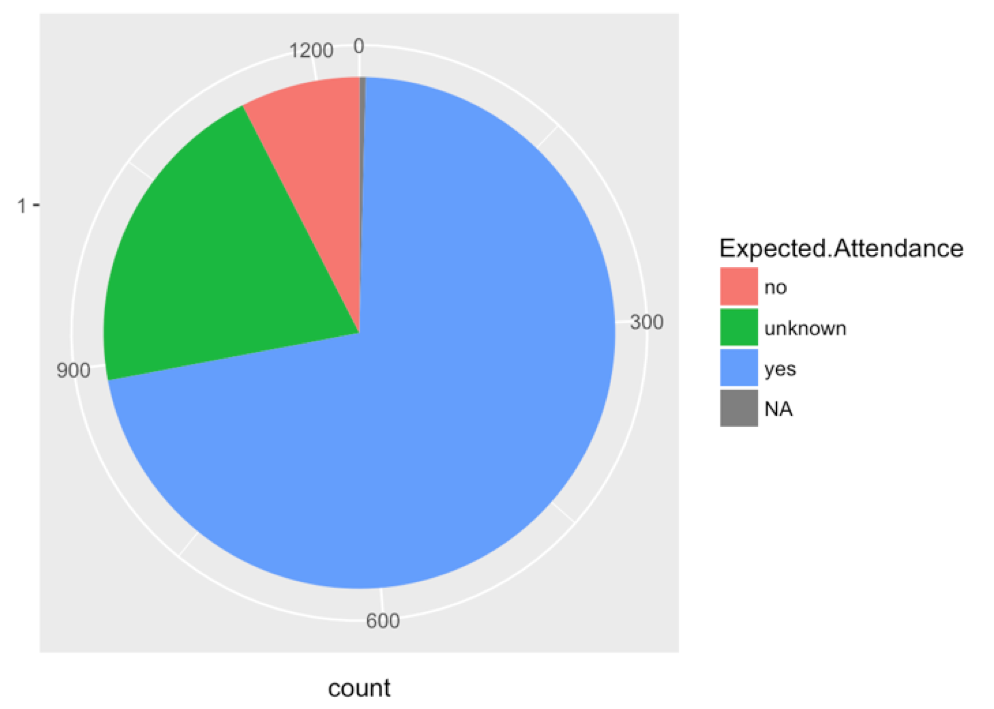
With classification techniques, we were able to predict whether a specific candidate would attend an interview, and compare the prediction accuracy with the baseline, i.e., always predict “yes”, to see the improvement of applying dimension reduction to the models. Furthermore, we applied the optimized model to two sets of variables, Questions and Non-Questions, to select the most significant variables in each set. This led us to the influential conclusion - what factors should be appreciated in encouraging interview attendance in the future.

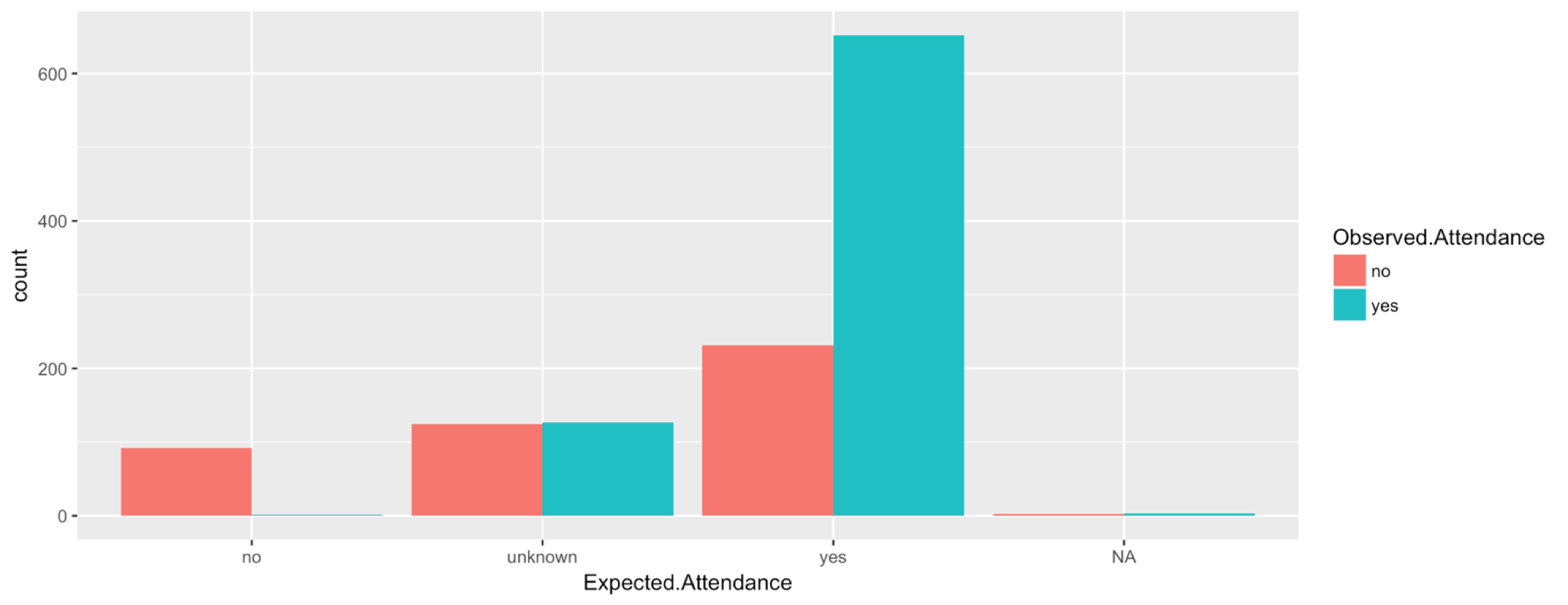
1. **Visualization**
   1. **Candidate Current Location, Candidate Job Location, Candidate Native Location**



A positive trend is observed with the candidate’s current location and the candidate’s job location, which indicates us to be aware of the collinearity of those two variables in the predictive models.

* 1. **Expected Attendance and Observed Attendance**





The expected attendance generally agrees with the observed attendance, with the exception that there are more people expected to attend the interview than people who actually showed up to the interview.

1. **Model Selection**

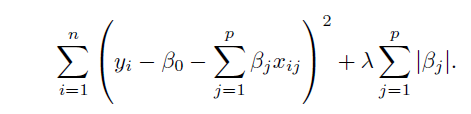
In the selection of models to address our classification question, we first listed out several suitable models: Logistic Regression, Linear Discriminant Analysis, Quadratic Discriminant Analysis, and more. After our research, most of the regression methods can also be applied to classification questions, with the specification of “family” parameters. This largely expanded our choices!

However, since our predictors and response are all categorical variables, we eliminated some choices: Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA), as they are only applicable to numerical predictors. We gave up with K-Nearest Neighbourhood (KNN) as well, because it may fail on high dimensions (In our case, there are 18 variables in total). We also abandoned the Adaboost method, because it is likely to overfit the data.

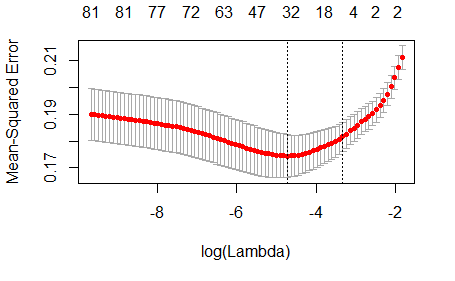
Next, in order to test whether using dimension reduction would improve the overall prediction accuracy, we decided to apply greedy approach to select the threshold for the three models: Lasso, Logistic, and Ridge. Details are as follows.

* 1. **Group Lasso**

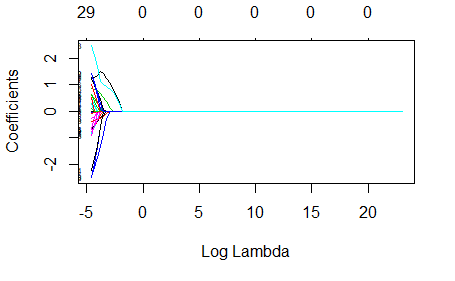
In our first trial of Naive Lasso, we encountered a problem: after dummy coded each level of categorical variables, we had nearly a hundred variables. But Lasso only shrinked several levels of a variable, which did not help with the variable selection. So, we decided to explore a new method - Group Lasso. This method views all levels of a variable as a whole, and filters out the entire variables that are of less significance.



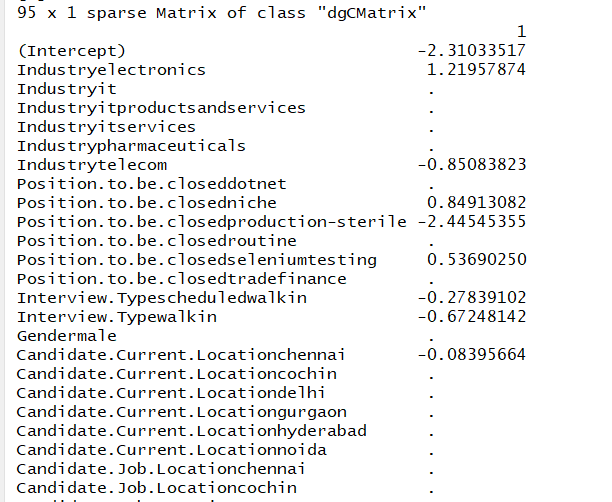
After decided on the method, we used Cross Validation to choose the optimal penalty factor,lambda, based on training errors.



Group Lasso shrinked out 6 insignificant variables: *Gender, Question 2 to 5, and Marital.Status*, and left us 12 significant variables. The test prediction accuracy is above 68.52%. In other words, it suggests the interviewer’s Question 1, 6, 7 are effectively affecting the candidate’s response. Besides, most of the demographic factors are decision-defining.



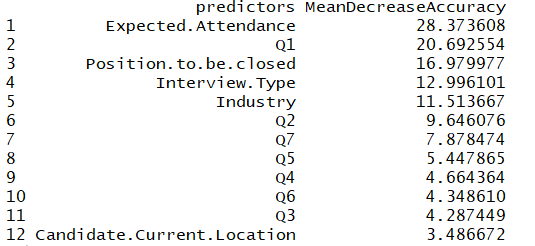
Here is a snapshot of the shrinking result from Group Lasso, we only eliminated the variable whose all levels were shrinked out.



* 1. **Random forest + logistic**

Logistic regression is a classification method without the assumption of data distribution, which makes it a good candidate for our task. Concretely, we first perform dimensionality reduction with threshold chosen by 10 fold cross validation, then apply logistic regression on the reduced set of predictors.

We use decrease-mean-accuracy, obtained from random forest with 1000 instances, as a criteria for dimensionality reduction.

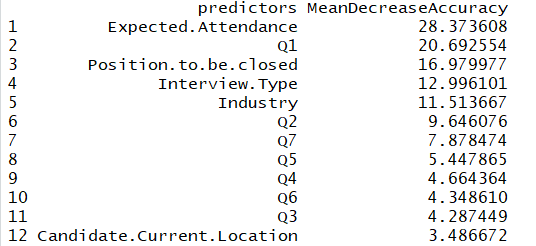


We found 6 variables - *Expected.Attendance, Question 1, 2, Position.to.be.closed, Interview.Type and Industry* - to be effective for our logistic model.

The model can be written as :

* 1. **Random forest + ridge**

We also investigate the performance of ridge regression in our analysis. Similar to the above model, we first build random forest with 1000 trees; then select variables by decrease-mean-accuracy with threshold chosen via 10 fold cross validation.



The lambda of the penalty term is also selected by a separate cross validation after each model is built. We found 12 variables - *Expected.Attendance, all the Questions, Position.to.be.closed, Interview.Type, Industry, Candidate.Current.Location* to be effective for ridge regression. The lambda for our finalized ridge model is 0.1349724.

1. **Model Application**

From the prediction accuracy applied to test data, we conclude that group lasso performs best:

|  |  |  |
| --- | --- | --- |
| **Methods** | **Train accuracy** | **Test accuracy** |
| Group Lasso | 0.7623888 | 0.6852792 |
| Random Forest + Logistic | 0.7242694 | 0.6649746 |
| Random Forest + ridge | 0.7623888 | 0.6598985 |

In order to figure out whether Question set or non-Question set is more important to the candidate’s decision, we applied group lasso to two groups respectively, prediction accuracy are as follows:

|  |  |  |
| --- | --- | --- |
| **Group Lasso** | **Train accuracy** | **Test accuracy** |
| Questions+expected attendance | 0.7534943 | 0.6954315 |
| non-questions | 0.7217281 | 0.6497462 |

We noticed that the test accuracy for group 1 is even better than the prediction using full model. However, the performance of non-question predictors are at baseline level. Our selected significant questions from group lasso are: Question 1, 6, 7, and Expected Attendance. And the significant non-question variables are: Gender, Candidate.Job.Location, and Marital.Status. This verified our hypothesis within tolerance.

1. **Conclusion**

From our analysis, we conclude that a candidate’s attendance can be predicted at a reasonable accuracy. Moreover, while the interview attendance of a candidate can be mainly explained by his/her answers to the questions asked by the interviewer, it seems to be irrelevant to the non-question factors, such as the position marital status and gender. Perhaps in order for the HRs to get a sense of whether a candidate will attend the interview and thus make preparation accordingly, it seems to be all coming down to what questions they should ask and how the candidate answers them.

Although we have achieved reasonable results over the baseline, our conclusion is still prone to several influential factors. First, our analysis is based on a dataset collected in India, which might not be applicable to other regions. Second, some of the non-question variables contain too many levels, and some of them only have one data in that, and thus we need more data to better fit the models and obtain more robust analysis. Lastly, our conclusion on the effect of question + expected attendance versus non-questions on the prediction separately are drawn from the test accuracies along. More complicated multiple testing should be applied to further confirm the results.

**8. Individual contribution:**

We worked closely to choose our questions, goals and models for analysis, while each of us has our own responsibility. Ruochen cleaned data and wrote code for random forest. He coordinates final code we used in our report as well. In lab report, he did introduction, models 5.2, 5.3 and conclusion part. Ruifan wrote codes for group lasso and ridge. She comes up with creative ideas all the time and also designs format for both our presentation and final report. In report, she is responsible for model 5.1, graphs and model application. Yuan wrote codes for logistic model, and explored other conductible models, such as Grouped KNN, Adaboost, etc. She is responsible for creating PowerPoint for group presentation and communicating with professors about model selection. In report, she did data overview, methodology and adding tables for model application. Yichun is our presenter and did visualization part for both presentation and our final report.