Sabirah Shuaybi
STAT-242
Final Report

Predicting Interview Attendance: Using Logistic Regression to Predict the Likelihood that a Candidate WIll Show Up for Their Interview

Abstract:

Lack of interview attendance is a growing concern in today's world. When candidates fail to show up for interviews, it negatively impacts the company by wasting time, resources and effort of the employees. The dataset "The Interview Attendance Problem" offers a variety of factors specific to candidates such as gender, marital status, resident location, interview location, industry, interview type, etc. By using a logistic regression model coupled with a best subsets procedure, we highlight which of these factors best predict the likelihood that a candidate will show up for their interview. A comprehensive understanding of these significant predictors can help companies adjust and refine their recruitment process around these identified factors to make it more cost-effective and streamlined.

Introduction:

The purpose of this analysis was to identify which set of variables are best suited for predicting interview attendance within this dataset and to then use these predictors to build a preliminary logistic regression model that will determine the probability that a candidate will show up for their interview. This rudimentary model was then refined by a best subsets process to obtain a model that can best predict this outcome. Logistic Regression was the chosen model and method of analysis for two main reasons. Firstly, a multiple linear regression model was found to severely violate many conditions such as linearity, constant variance and normality of errors. Because logistic regression does not require these assumptions, we chose this form of analysis. Secondly, logistic regression, due to its nature, was best designed to answer the research question at hand (The *Probability* of Interview Attendance). We were interested in a categorical response variable (named 'Showed' which contains 1 if candidate showed up for their interview and 0 if they did not). Following the steps outlined above, a five-predictor logistic regression model was obtained that had the highest adjusted R-squared and thus was comparatively better at explaining the variability in interview attendance.

Data:

The Dataset, named "The Interview Attendance Problem" was obtained from Kaggle. It contains information about interview attendance gathered primarily though survey techniques from the recruitment industry in India. The period of data collection ranged from September 2014 to January 2017. The data consists of details about the individual being interviewed, the nature of the interview itself, as well as answers from the interviewees to a set of questions asked by the recruiter. Examples of the original variables included in this data set are as follows: Industry, Position, Nature of Skill Set, Current Location of Candidate, Location of Interview, Location of Job, Gender, Marital Status, If they have taken out a printout of their resume, if they are clear on the details of the interview, etc. Since most variables in this data set have string values, the main challenge pre-analysis was cleaning the data, addressing NA values and re-creating variables for the sake of consistency.

Pre-existing variables were also combined into new variables that were then incorporated into the model. For example, a variable IntClose was created that contained 0 if the candidate's current location of residence was different than the interview venue. If these two locations were the same, the variable contained 1. Similarly, we created IntJob that checked it the candidates current location was the same as the location of the job. Moreover, other variables were transformed into more coherent levels. For example, one factor in the dataset was whether the candidate had a printed out version of their resume ready before the interview. The original character values of *"Yes"*,*"No"*, *"Not yet"*, and *"Will do so soon"* were encoded as numerical values 0 (for "No"), 1 ("for "Not yet"/"Will do so soon"), and 2 (for "Yes")*.

**Refer to code in the annotated appendix section for more details as to how the variables were filtered and encoded.*

Model:

In the preliminary regression model, the following predictors were included:

- *Gender*
- *Industry*
- *Marital Status*
- *Interview Type*
- *IntClose*
- *JobClose*
- *Clear* (If candidate was clear on where and when interview was taking place)
- *Printout* (Resume printed out before hand)
- *Call* (If the candidate gave permission for the interviewer to call 3 hours prior to the interview to confirm)

The reason behind this specific selection was based on logical inference and intuition that these factors might have an impact on interview attendance (as opposed to other variables that did not seem to be overtly relevant, such as date of interview, candidate native location, etc.) A summary of this model revealed that many of these predictors were not significant in explaining the variability in 'Observed Attendance'. Thus, to obtain a more concise and refined model, that could better predict the probability that a candidate will attend their interview, a best subsets procedure was conducted. The final model chosen was one that had the highest adjusted R-squared value. Based on the subsets outcome, this was a 5-predictor logistic regression model that used JobClose, IntClose, Clear, Printout and Call to predict the probability of attendance. This model had the highest R^2 value (0.0643) when compared to other models with 1, 2, 3, 4, 6, 7, 8 and 9 predictors. This model also had the highest adjusted R^2 value (0.0555). The summary of this new model revealed the following p-values:

- ➔JobClose: p-value = 0.022605 < 0.05
- ➔IntClose: p-value = 0.046235 < 0.05
- ➔Clear: p-value = 0.000606 < 0.05
- ➔Printout: p-value = 0.000466 < 0.05
- ➔Call: p-value = 0.221605 > 0.05

Thus, it can be seen that JobClose, IntClose, Clear, and Printout are all significant predictors of interview attendance. Call is not a significant predictor as it's p-value is higher than the significance level of 0.05. However, it is more significant than other variables and thus was kept in the model.

Probability Form of the Final Model:

$$P = \frac{e^{(-5.4+(2.7*\mathbf{Clear})+(1.3*\mathbf{Printout})+(-1.3*\mathbf{JobClose})+(1.2*\mathbf{IntClose})+(0.96*\mathbf{Call})}}{1 + e^{(-5.4+(2.7*\mathbf{Clear})+(1.3*\mathbf{Printout})+(-1.3*\mathbf{JobClose})+(1.2*\mathbf{IntClose})+(0.96*\mathbf{Call})}}$$

(Where ***P*** represents the probability that a given candidate will show up for their interview).

Results:

From the final model obtained, it can be observed from the corresponding slope coefficients that the variables, Clear, Printout, IntClose and Call are positively associated with the log of 'Showed'. In other words, this means that candidates who are clear on the details of the interview, have a printout of their resume, live in the same city as the interview location and have received a call confirming the interview are more likely to show up for their interview. Interestingly, it can also be noted that JobClose has a negative coefficient. This suggests that candidates who live in a *different* city than the job location are more likely to attend their interview (due to how this variable was encoded). This probability model was then tested by inputting various values for each of the five factors and computing the probability of attendance *(The code for the predict function and test values are also available in the annotated appendix section)*.

Clear=1, Printout=2, JobClose=0, IntClose=1, Call=1

Probability = 0.903

When entering these values (above) for each predictor, there is a 90.3% probability that this candidate will show up, based on this model.

Clear=0, Printout=0, JobClose=1, IntClose=0, Call=0
Probability = 0.001

On the other hand, when entering these values for each predictor, the model predicts a 0.1% likelihood of this candidate showing up for their interview.

**Note: these two examples were deliberately extreme on each end, to showcase the impact of these predictors on probability of attendance*
These results answer the original research inquiry of what factors influence interview show-rate by demonstrating the importance of location (interview as well as job location), preparedness (resume printed) and clarification (clear on details) on interview attendance.

Conclusion:

The application of this analysis is to provide companies with a holistic understanding of the dynamics of the interview process and inform them as to which factors are most significant in predicting interview attendance. Through this analysis, we have obtained a 5-predictor model that is adept at explaining a portion of the variation in interview attendance. Companies can then use this model to improve their resource allocation. For example, a company could input data for these 4 significant factors (IntClose, JobClose, Printout, and Clear) into the model for any given candidate to calculate an approximate probability of their show rate. If the probability came out to be quite low, employees would have a back-up plan ready in the case that the candidate does in fact, fail to show up. On the other hand, if the probability was presented to be fairly high, employees could be more mentally prepared and assured of a given candidates attendance. Furthermore, the identification of these significant factors is useful in enhancing the interview process and attendance rate by implementing new practices such as clarification calls to candidates and requests for a resume upfront (when scheduling the interview) or even having multiple interview venues, if possible, for improved accessibility and thus, higher attendance.

This analysis is undermined by several limitations of the data and statistical methods used. Firstly, the data was primarily obtained through a series of questions/surveys of candidates which could lead to cases of response bias thus rendering some answers as unreliable. Secondly, because the range of data collection for this dataset is quite narrow (data from a specific time frame in a specific country only), these results may not be applicable to other countries, which undoubtedly have different economic conditions. Furthermore, because the model only predicts a binary categorical response (either the candidate shows up, or they do not show up), it loses some of the real-life complexity involving intermediate outcomes (such as arriving late to the interview, or cancelling due to a sickness or emergency). Lastly, while this logistic regression model identified a couple of significant and tangible predictors for Observed Attendance, it was not particularly useful for explaining the majority of the variability in the response. In fact, the final model only accounted for 6.4% of the total variability.

While there are obviously numerous limitations of an integrated application of this analysis into the recruitment industry, stemming from the dataset itself (extrapolation issues), as well as the real-life multiplicities of the interview process (such as the vagueness of intermediate probabilities → what can an employee do with a 50/50 chance of attendance?), this analysis is nonetheless valuable in unravelling the complexity of the interview attendance problem and highlighting which factors can be addressed in the future to ensure higher attendance rates as well as more sophisticated interview resource allocation.

**Annotated Appendix:

```
#Who shows up for interviews?
@Author: Sabirah Shuaybi
attach(Int)

#Data Cleaning/Filtering and Renaming Variables
Int$Call[Int$Can.I.Call.you.three.hours.before.the.interview.and.follow.up.on.your.attendance.for.the.interview=="No"|Int$Can.I.Call.you.three.hours.before.the.interview.and.follow.up.on.your.attendance.for.the.interview=="Na"|Int$Can.I.Call.you.three.hours.before.the.interview.and.follow.up.on.your.attendance.for.the.interview=="No Dont"]<-0
Int$Call[Int$Can.I.Call.you.three.hours.before.the.interview.and.follow.up.on.your.attendance.for.the.interview=="yes"|Int$Can.I.Call.you.three.hours.before.the.interview.and.follow.up.on.your.attendance.for.the.interview=="Yes"]<-1

Int$Showed[Int$Observed.Attendance=="no"|Int$Observed.Attendance=="No"|Int$Observed.Attendance=="NO"]<-0
Int$Showed[Int$Observed.Attendance=="yes"|Int$Observed.Attendance=="Yes"]<-1

Int$IntType[Int$Interview.Type=="Scheduled"]<-"Sch"
Int$IntType[Int$Interview.Type=="Scheduled Walkin"|Interview.Type=="Scheduled Walk In"]<-"SchW"
Int$IntType[Int$Interview.Type=="Walkin"]<-"W"

Int$Clear[Int$Are.you.clear.with.the.venue.details.and.the.landmark.=="No"|Int$Are.you.clear.with.the.venue.details.and.the.landmark. == "No- I need to check"]<-0
Int$Clear[Int$Are.you.clear.with.the.venue.details.and.the.landmark.=="Yes"]<-1
```

```r
Int$Printout[Int$Have.you.taken.a.printout.of.your.updated.resume..Have.you.read.the.JD.and.understood.the.same=="Yes"]<-2
Int$Printout[Int$Have.you.taken.a.printout.of.your.updated.resume..Have.you.read.the.JD.and.understood.the.same=="No- will take it soon"|Int$Have.you.taken.a.printout.of.your.updated.resume..Have.you.read.the.JD.and.understood.the.same=="No"|Int$Have.you.taken.a.printout.of.your.updated.resume..Have.you.read.the.JD.and.understood.the.same=="Not yet"]<-1
Int$Printout[Int$Have.you.taken.a.printout.of.your.updated.resume..Have.you.read.the.JD.and.understood.the.same=="No"]<-0

#Creating a new variable called IntClose (comparison of candidate loc vs. interview loc)
for(i in 1:1234) {
  if((as.character(Int$Candidate.Current.Location[i])) == (as.character(Int$Interview.Venue[i]))) {
    Int$IntClose[i]=1
  }
  else {
    Int$IntClose[i]=0
  }
}

#Creating a new variable called JobClose (comparison of candidate loc vs. job loc)
for(j in 1:1234) {
  if((as.character(Int$Candidate.Current.Location[j])) == (as.character(Int$Candidate.Job.Location[j]))) {
    Int$JobClose[j]=1
  }
  else {
    Int$JobClose[j]=0
  }
}

#Exploratory Analysis of Dataset
prop.table(table(Int$Observed.Attendance, Int$Gender))
prop.table(table(Int$, Int$Gender))
prop.table(table(Int$`Gender`, Int$`Showed`), 1)

#Multiple Linear Regression Model (Not used as part of Analysis due to Violation of Conditions)
mlrMod <-lm(Showed~Gender+Industry+Marital.Status+IntType+IntClose+JobClose+Clear+Printout+Call, data=Int)
summary(mlrMod)
plot(mlrMod)
```

```
#Logistic Regression
modLog<-glm(Showed~Gender+Industry+Marital.Status+IntType+IntClose+JobClose+Clear+Printout+Call, family=binomial, data=Int)
summary(modLog)

#Best Subsets
library(leaps)
mymodels=regsubsets(Showed~Gender+Industry+Marital.Status+IntType+JobClose+IntClose+Clear+Printout+Call, data=Int)
summary(mymodels)
summary(mymodels)$rsq
summary(mymodels)$adjr2

#Building a better model from the best subsets procedure (5-predictor model)
modbest<-glm(Showed~Clear+Printout+JobClose+IntClose+Call, family=binomial, data=Int)
summary(modbest)

#Obtaining Predicted Values
logpred<-predict(modbest)
mean(logpred[Int$Clear==1], na.rm=TRUE)

newdata = data.frame(Clear=1, Printout=2,  JobClose=0, IntClose=1, Call=1)
predict(modbest, newdata, type="response") #Probability = 0.903

newdata2 = data.frame(Clear=0, Printout=0, JobClose=1, IntClose=0, Call=0)
predict(modbest, newdata2, type="response") #Probability = 0.001
```