Predicting Interest in Career Change

STAT 835: Categorical Reponse Variable Project

Brendan Stevens

Department of Biostatistics and Data Science
University of Kansas, USA
May 1, 2021

Contents

List of Tables

List of Figures

Abstract

A company that is involved with Big Data and Data Science was interested in predicting whether or not individuals would seek a career change based upon a number of factors, the most important of which was the total number of hours of training the individual surveyed had completed with the business. The business was also interested in what effect attending training had had on an individual's choice overall.

The analysis was completed using measures of association between the many predictor variables. Collinearity was assessed and an effort was made for its removal. A Logisitic Regression model was created for predicting responses.

Having had no expectations, the model was found to be moderately successful in predicting response outcomes.

Introduction

The imporantance of this study stems from financial efficiency. If, after this study, the business finds that no relationship exists between the amount of hours of training an individual attends with the business and whether or not that individual seeks to change careers, then the business will likely need to change or otherwise cease operations with their current training model. This study will begin with an analysis of association and correlation between the predictor variables, explain considerations and methods of variable removal from model consideration if needed, and then begin building a model for the prediction of the question at hand. Finally, a display of model effectiveness will be made and a conclusion will be drawn and discussed.

Materials and Methods

Data Sources

The data was obtained from Kaggle.com on April 17th, 2021. The data was originally uploaded on December 6th, 2020 by the user "Möbius". This data was recorded into one comma separated value file which includes 19158 observations, of which 8955 are used, due to missing information in the data. Variables in the data set include the **City Development Index** of the employee (scored from 0 to 1 with higher scores indicating better standards of living), their **Gender**, their **Relevant job experience**, or focused skillset specific to data science and big data, **Education level**, **College Major**, **Total years of work experience**, the **Size of the company** for which they worked, the **Type of company** for which they worked, **the amount of time that ellapsed between their current job and previous one**, the amount of **hours trained with the business**, and their response to the question as to whether they planned to **switch career targets or not**.

Statistical Analysis

This analysis was done using the statistical software program R, version 3.6.3 (2020-02-29) in RStudio, version 1.2.5033. The primary form of analysis used multiple logistic regression. The original data 19158 observations, of which some had missing values. This study removed observations with missing values, which left 8000 observations to consider.

Before any models were considered, a tremendous effort was made to identify and remove any collinearity. This was done through measures of association tests such as Chi-Squared Test of Independence, Fisher's Exact Test, and Cramer's V for nominal-nominal relationships. Spearman's correlation Coefficient was used for continuous-ordinal relationships. Both the Rank- and Point-Biserial Correlation Coefficients were used for continuous-nominal relationships when nominally dichotomous. The same test was applied for continuous-nominal that had multiple categories, as each of the multi-category nominal variables were subdivided into separate dataframes for pairwise correlation comparisons. Finally, Pearson's Correlation Coefficient was used for continuous-continuous relationships. There were no ordinal-ordinal relationships, as three of the four ordinal variables had greater than five categories which were deemed during analysis to be equidistant and thus treated as continuous for the association and correlation analysis.

Finally, a display of model effectiveness was included through an analysis of the Sensitivity and Specificity of the predicted and actual responses to the question at hand. This analysis was completed through use of a contingency table, a ROC curve, and a final Pearson's Correlation Coefficient between the predicted and actual responses.

Model Assumptions

All inferences are conducted using $\alpha = 0.05$ unless stated otherwise. The study is under the assumption that respondents were asked independently, and thus their conclusive answers to the question being posed was also independent.

The model assumes linearity in the association between the log of the odds and the predictor variables observed.

Exploratory Analysis Using Contingency Tables

Before screening the data, a check of the sparseness of the data should be considered to determine methods of further analysis. A selected table is shown here to represent this analysis. All other variables have at least 50 values in every category. However, the Degree that the observed each have is an issue, because three responses were given, "Masters", "PhD", and "Graduate". There is no information as to what the distinction is between Graduate and the other two responses. Since there might be individuals who responded "Graduate" that have a Masters or a PhD, this variable is useless. It will be removed from further consideration.

Table 1: College Majors of Individuals Observed

Major	Frequency
Arts	253
Business Degree	327
Humanities	669
No Major	223
Other	381
STEM	14492

Screening for Association and Correlation Between Predictors

Now that the data has been screened for sparseness, a study of Association and Correlation between pairwise comparisons needs to be made. The in-depth R code and analysis for this is available at the end of the document, but shown here for consideration are the correlations between the continuous variables.

	CDI	Years Exp	Size of Co.	Years Bet.	Hours Trained
City Development Index	1.00	0.33	0.07	0.16	-0.01
Years of Work Experience	0.33	1.00	0.09	0.38	0.00
Size of Company Worked For	0.07	0.09	1.00	0.10	-0.02
Years Betw. Curr. & Prev. Job	0.16	0.38	0.10	1.00	-0.01
Hours of Training Completed	-0.01	0.00	-0.02	-0.01	1.00

As can be seen, no correlation above the 0.70 correlation threshold for collinearity exists between the continuous predictors.

Now the nominal-nominal categorical variables are to be considered. There are four of them. Gender, Related Experience, Major, and Company type. Gender has three categories, Male, Female, and Other. Related Experience has two, essentially yes or no. Major has six categories, and company type has six categories. Below is the a selected contingency table for Gender and Major, the rest can be recreated using the provided R code at the end of this document.

	Arts	Business Degree	Humanities	No Major	Other	STEM
Female	24	16	82	5	21	656
Male	102	152	292	105	154	7268
Other	3	2	4	2	2	65

For any table with greater than twenty percent of cell values having values less than five for an entry, Fisher's Exact Test must be used over the Chi-Squared Test for Independence. Whenever possible, however, the Chi-Squared Test of Independence will be used. Cramer's V can provide an additional perspective score between 0 and +1 interpretted similarly to

the Pearson Correlation Coefficient and will be applied to all of these.

The results for the previous table are as follows. The p-value of the Fisher's Exact Test is 0.0005 and the Cramer's V is 0.07676.

For each of these tables, a p-value lower than 0.05 for either the Fisher's Exact Test or Chi-Squared Test for Independence means that the null hypothesis for independence is rejected, and association is assumed. So, in the case of Major and Gender, it is safe to assume there is association between one's gender and one's major of choice. A Cramer's V test of greater than 0.10 is considered the lower threshold for correlation. Therefore, we can assume that while there is association between Major and Gender in the dataset, there doesn't appear to be correlation.

Additional tests for Association and Correlation including the Spearman and two Biscerial tests between continuous-ordinal, continuous-nominal, and other relevant combinations can be retrieved from the code attached. It is not being included in here, as a Pearson Correlation Coefficient is being created for every continuous variable and pair of nominal levels. For example, a score calculated between the two levels of Relevant Experience versus City Index is easy to interpret and see, as it is only one score, but Major vs City Index has fifteen different correlation coeffients to consider, one for each pair of majors.

Correlation is present in two combinations in this dataset. It exists between the Relevant Experience a respondent has to the field of Data Science and Big Data and the Type of Company for which that individual works. It also exists between Company Type and Company Size.

Since Company Type is involved in both of these, dropping it will go a long way in eliminating collinearity, so moving forward it will be dropped from consideration in the model.

Primary Objective Analysis: Fitting the Model

The method for creating the model will be Purposeful Selection. The first step in this procedure is to construct an initial main effects model, which includes any beta coefficient with a p-value < 0.20, which can be thought of loosely as all the single predictor models that are better than merely using the mean instead to model the log of the odds. Included below are the first two of the nine single variable regression coefficients.

Table 4: Gender Coefficients

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.6365	0.09550	-17.13738	0.00000
genderMale	0.0222	0.10007	0.22181	0.82446
genderOther	-0.0682	0.32802	-0.20792	0.83529

As can be seen in the two regression model summaries, City Development Index (CDI) is significant at the < 0.20 level, but Gender is not. Therefore CDI will be used for the initial

effects model, and Gender will not.

Similarly, Relevant Experience, Type of Enrollment, Major, Years of General Work Experience, and Years Between Jobs are significant. Gender, Company Size, and Hours of Training with the company were not significant at this level.

The next step is to compare the main effects model to each of the single predictor models. This will be done using the Likelihood Ratio Test to compare the deviances between models. Nine models are compared to the main effects model, shown below is the comparison between the model that uses CDI as the sole predictor versus the model that uses all nine predictors. If the p-value is below 0.05, then the nested model can be said to be less effective than the full model.

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
8953	6667.7	NA	NA	NA
8944	6601.4	9	66.319	0

The p-value here is small, and can be shown to be small when comparing the main effects model to the single predictor models in every case.

The next step is to give the three eliminated predictors and opportunity to prove their worth against the main effects model.

One variable, Company size, makes it back into the model.

While this is a predictive model, a certain degree of presentation-mided clarity has been requested insofar as model interpretation is concerned. For that reason, during this next step, which allows for the addition of interaction terms, only second degree interactions will be considered for the model.

Seven interaction terms appear to have significance at the 0.05 alpha threshold. For each of these terms, The model with company size added back in is being tested against the model with company size with one of the interaction terms added as well. Every model with a single interaction term appears to be better than the single main effects model that included company size. The same Likelihood Ratio Test and comparison of deviances is applied as before to reach this conclusion.

Finally, for good measure, the full model with all of the interaction terms included is pitted against the models with only one interaction term. Again, the full model with all seven interaction terms beats out the single interaction term models.

The final predictive model is as shown:
$logit(Y_i) = -1.10 - 2.09CDI - 1.54RelExp_{None} - 0.48Enrolled + 11.37Major_{Business}$
$+7.60Major_{Humanities} + 13.20Major_{NoMajor} + 9.20Major_{Other} + 8.12Major_{STEM}$
$-0.17Experience - 0.19LastNewJob + 2.22CDI * RelExp_{None} + 0.74CDI * Enrolled$
$-12.73CDI * Major_{Business} - 8.05CDI * Major_{Humanities} - 15.19CDI * Major_{NoMajor}$
$-10.82CDI * Major_{STEM} + 0.19CDI * Experience + 0.29CDI * LastNewJob$ $ -
0.01Experience*LastNewJob + 0.02LastNewJob*CompanySize$

Results

Summarizing Predictive Power of the Model

In order to analyze the effectiveness of the model, a comparison between the predicted answer to the target question, "Do you plan to change careers", needs to be made with the actual response of each of the observed individuals in the study. An at a glance representation of this relationship is shown below.

Table 6: Actual Responses (Left column) vs Predicted (Top Row)

	0	1
0	6044	1428
1	547	936

Here it can be seen that of the 8955 observations considered, 6081 did not seek work and were correctly predicted, 558 did seek work and were incorrectly predicted not to do so, 1391 did not seek work and were incorrectly predicted to do so, and 925 did seek work and were correctly predicted.

Sensitivity is the probability that the model will predict a person is seeking employment given that they actually are, or $925/(925+558) = 0.62$. Specificity is the probability that the model will predict a person is not seeking employment given that they are not, or $6081/(6081+1391) = 0.81$.

A ROC curve can also be constructed to evaluate the model. The model is shown below.

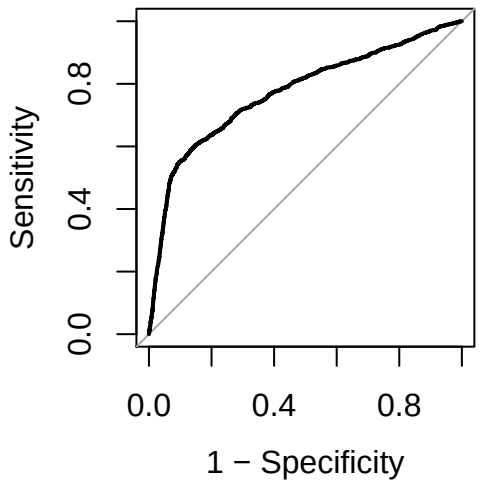

Figure 1: ROC Plot of the Predictive model vs the Actual Observed Values

The area under the curve of a ROC curve is a measure of interest, and here it is equal to

0.7672.

Finally, a Pearson's Correlation Coefficient between the actual and predicted responses can be calculated. That value is r = 0.4547879.

Discussion and Conclusion

A Sensitivity score of 62% is not great. But, its better than 50%, which was what the company was able to do without this analysis. A Specificity of 81% is much better. The model will predict that a person is not seeking employment, given that they are not 81% of the time.

The ROC Plot area under the curve is 0.7672, which is fair. A score of 0.50 means the model is no better at predicting than flipping a coin, and a score of 1.00 means that it perfectly predicts. For this reason, the score of 0.7672 is useful, but not impressive.

The Pearson Correlation Coefficient of 0.45 is a moderate correlation. With this large of a sample size, it can be stated that a 0.45 correlation is not chance, and is therefore rather accurate. While this model does not predict the outcomes correctly every time, it does do a better job of it by a substantial margin than random guessing on the end of the company.

Finally, the major question at hand, whether the number of hours spent in training with the company is worth all of the effort. This study would say "No", as there is no statistically significant linear relationship between the number of hours spent in training and the log-odds of a trained individual seeking a new career.

Appendix: R-code

```
# Stevens Final Project

library(readr)
library(rcompanion) # For use of Cramer's V (Correlation between cat. vars)

work <- read_csv("aug_train.csv")

str(work)
str(work$experience)
unique(work[c("experience")])
unique(work[c("gender")])


# Not experienced with imputation, so deleting all of the NA values
cleaned <-subset(work, (!is.na(work[, 1])) & (!is.na(work[, 2]))
                & (!is.na(work[, 3])) & (!is.na(work[, 4]))
                & (!is.na(work[, 5])) & (!is.na(work[, 6]))
                & (!is.na(work[, 7])) & (!is.na(work[, 8]))
                & (!is.na(work[, 9])) & (!is.na(work[, 10]))
```

```
              & (!is.na(work[, 11])) & (!is.na(work[, 12]))
              & (!is.na(work[, 13])) & (!is.na(work[, 14])))

# Went from 19158 observations to 8955, so n is still rather large.

# Checking to see if the categorical response values are sparse at any level
table(unlist(cleaned$gender)) # 804 Female
                           # 8073 Male
                           # 78 Other

table(unlist(cleaned$relevent_experience)) # 7851 Has experience
                                      # 1104 No experience

table(unlist(cleaned$enrolled_university)) # Full time 832
                                      # No Enrollment 7594
                                      # Part time 529
# Categories need reordered, as they could be made ordinal

table(unlist(cleaned$education_level)) # Graduate 6252
                                  # Masters 2449
                                  # PHD 254
# There doesn't appear to be any documentation on why Graduate is an available
# response option while Masters and PHD were also choices, so this analysis
# will ignore this categorical variable
cleaned <- cleaned[, -7] # degree is variable 7 in dataframe

table(unlist(cleaned$major_discipline)) # Arts 129
                                   # Business 170
                                   # Humanities 378
                                   # No Major 112
                                   # Other 177
                                   # STEM 7989

table(unlist(cleaned$experience)) # Varied responses
# Categories need reordered, as they could be made ordinal
# All of the response categories were present from 1-20 year with at least
# 50 values in each category, making for a great spread of data
# There is also a category for less than a year and greater than 20
# Likely going to replace with 0 and 21 (might increase the 21 to 24ish
# to simulate the unknown average of these ages, which is likely higher than
# just 21. Could also just outright remove these observations)

table(unlist(cleaned$company_size)) # Good spread of responses
# Categories need reordered, as they could be made ordinal

table(unlist(cleaned$company_type)) # Early State Startup 385
                                # Funded Startup 784
```

```
                                # Non Government Organization 356
                                # Other 72
                                # Public Sector 564
                                # Private Limited Company 6794

table(unlist(cleaned$last_new_job)) # Never (no time between) 373
                                # 1 Year 3838
                                # 2 Years 1570
                                # 3 Years 610
                                # 4 Years 599
                                # More than 4 Years 1965
# Categories need reordered, as they could be made ordinal
# Might value "the "More than 4 Years" to 7ish to simulate the unknown
# average of this category, which, based solely on intuition, is likely
# right skewed due to the how the other data behaves (right skewed)

table(unlist(cleaned$target)) # Did not seek employment 7472
                           # Did seek employment 1483


# Reordering the levels of the ordinal variables and changing them to numbers

# No Enrollment = 1, Part time = 2, Full time = 3
cleaned$enrolled_university <- as.factor(cleaned$enrolled_university)
levels(cleaned$enrolled_university) # original order
cleaned$enrolled_university <- factor(cleaned$enrolled_university,
                                levels = c("no_enrollment",
                                           "Part time course",
                                           "Full time course" ))
levels(cleaned$enrolled_university) # from lowest to highest
cleaned$enrolled_university <- as.numeric(cleaned$enrolled_university) - 1

# Assigning numeric values to each category
cleaned$experience <- as.factor(cleaned$experience)
levels(cleaned$experience) # original order
cleaned$experience <- factor(cleaned$experience,
                                levels = c("<1", "1", "2", "3",
                                           "4", "5", "6", "7",
                                           "8", "9", "10", "11",
                                           "12", "13", "14", "15",
                                           "16", "17", "18", "19",
                                           "20", ">20"))
cleaned$experience <- as.numeric(cleaned$experience) - 1
# Switching 21+ years of experience with 24
cleaned$experience[cleaned$experience == 21] <- 24

# Changing order of company size to correct order and assigning values
```

```
cleaned$company_size <- as.factor(cleaned$company_size)
levels(cleaned$company_size)
cleaned$company_size <- factor(cleaned$company_size,
                               levels = c("<10", "10/49", "50-99",
                                          "100-500", "500-999", "1000-4999",
                                          "5000-9999", "10000+"))

cleaned$company_size <- as.numeric(cleaned$company_size)

# Changing order of last new job response
cleaned$last_new_job <- as.factor(cleaned$last_new_job)
levels(cleaned$last_new_job)
cleaned$last_new_job <- factor(cleaned$last_new_job,
                               levels = c("never", "1", "2", "3", "4", ">4"))
cleaned$last_new_job <- as.numeric(cleaned$last_new_job)
cleaned$last_new_job[cleaned$last_new_job == 5] <- 7




############################################################
# Step 0: TESTING FOR CORRELATION OR ASSOCIATION SECTION
############################################################

# Checking variables for correlation, test choices found in article below
# https://journals.sagepub.com/doi/pdf/10.1177/8756479308317006

# Starting with the continuous vs continuous comparison between the variables
# (Pearson's Correlation Coefficient)
# treating ordinal values with higher than 5 categories as continuous for this
# https://www.statisticssolutions.com/can-an-ordinal-likert-scale-be-a-continuous-variable/
continuous <- cleaned[, c(3, 8, 9, 11, 12)]
cor(continuous) # no multicollinearity present within continuous vs cont

# Nominal vs Nominal association and correlation (Cramer's V)
# comparing nominal variables using chi squared tests for independence
# gender vs relevant experience
gender_relevant <- xtabs(~gender + relevent_experience, data=cleaned)
gender_relevant
# no sparse relations, use chi-squared test
chisq.test(gender_relevant) # dependency exists
stdres_gender_relevant <- chisq.test(gender_relevant)$stdres # standardized residuals

# https://www.statisticshowto.com/what-is-a-standardized-residuals/
stdres_gender_relevant
# the "Rule of Thumb" for residuals is that...
```

```
# residuals are above/below 2, so the cells are greater and lower than
# expected, with men having a higher chance to have relevant experience

# using Cramer's V test to check for correlation
# http://www.acastat.com/statbook/chisqassoc.htm#:~:text=between%20the%20variables.-,It%20is%20
# "[Cramer's V] is interpreted as a measure of the relative (strength)
# of an association between two variables. The coefficient ranges from
# 0 to 1 (perfect association). In practice, you may find that a Cramer's V
# of .10 provides a good minimum threshold for suggesting there is a
# substantive relationship between two variables."
cramerV(gender_relevant) # 0.08, so while there is association, no correlation
# and therefore no collinearity

# gender vs major
gender_major <- xtabs(~gender + major_discipline, data=cleaned)
gender_major <- as.matrix(gender_major) # have some small values in the other category
# using Fisher's Exact test to deal with small sample values
fisher.test(gender_major, simulate.p.value = TRUE)
# reject null of independence, there is association
cramerV(gender_major) # 0.08, association exists, but not correlation

# gender vs company_type
gender_company <- xtabs(~gender + company_type, data=cleaned)
gender_company
fisher.test(gender_company, simulate.p.value = TRUE)
# reject null of independence, there is association
cramerV(gender_company) # 0.04, association but no correlation

# relevant experience vs major
relevent_major <- xtabs(~relevent_experience + major_discipline, data=cleaned)
relevent_major
chisq.test(relevent_major)
# reject null of independence, there is association
cramerV(relevent_major) # 0.09, association no correlation

# relevant experience and company type
relevent_company <- xtabs(~relevent_experience + company_type, data=cleaned)
relevent_company
chisq.test(relevent_company)
# reject, there is association
cramerV(relevent_company) # 0.21 assocation AND correlation here

# major and company type
major_company <- xtabs(~major_discipline + company_type, data=cleaned)
major_company
fisher.test(major_company, simulate.p.value = TRUE)
# reject, there is association
```

```
cramerV(major_company) # 0.05, association no correlation

# Continuous vs Ordinal Assoctiation (Spearman's test)
# city index and enrolled in university
cor(cleaned$city_development_index,
    cleaned$enrolled_university,
    method = "spearman")
# r = -0.1096387, which means extremely low correlation, so no collinearity

# Enrolled in Uni vs experience
cor(cleaned$enrolled_university,
    cleaned$experience,
    method = "spearman")
# r = -0.255773, which means low correlation, no collinearity

# Enrolled in Uni vs company size
cor(cleaned$enrolled_university,
    cleaned$company_size,
    method = "spearman")
# r = -0.03709883, no correlation or collinearity

# Enrolled in Uni vs Years passed between previous and current job
cor(cleaned$enrolled_university,
    cleaned$last_new_job,
    method = "spearman")
# r = -0.1237778, no correlation or collinearity

# Enrolled in Uni vs Hours of Company training completed
cor(cleaned$enrolled_university,
    cleaned$training_hours,
    method = "spearman")
# r = 0.002190174, no correlation or collinearity

# Correlation between Continuous and Nominal
# City_index vs Gender
cor(cleaned$city_development_index,
    as.numeric(as.factor(cleaned$gender)) - 1)
# since the nominal value is binary, can just use pearson instead
# of the point-biserial, since they're equivalent here
# r = -0.01179183, so no correlation

# Gender vs experience
cor(as.numeric(as.factor(cleaned$gender)) - 1,
    cleaned$experience)
# r = 0.07766857, no correlation

# Gender vs company size
```

```
cor(as.numeric(as.factor(cleaned$gender)) - 1,
    cleaned$company_size)
# r = -0.003053524, no correlation

# Gender vs Years passed between previous and current job
cor(as.numeric(as.factor(cleaned$gender)) - 1,
    cleaned$last_new_job)
# r = 0.03042108, no correlation

# Gender vs Training Hours
cor(as.numeric(as.factor(cleaned$gender)) - 1,
    cleaned$training_hours)
# r = -0.01327347, no correlation

# Relevant Experience vs city index
cor(as.numeric(as.factor(cleaned$relevent_experience)) - 1,
    cleaned$city_development_index)
# r = 0.005042932, no cor

# Relevant Experience vs experience
cor(as.numeric(as.factor(cleaned$relevent_experience)) - 1,
    cleaned$experience)
# r = -0.1190207, no cor

# Relevant Experience vs Company Size
cor(as.numeric(as.factor(cleaned$relevent_experience)) - 1,
    cleaned$company_size)
# r = 0.05430268, no cor

# Relevant Experience vs Years passed between previous and current job
cor(as.numeric(as.factor(cleaned$relevent_experience)) - 1,
    cleaned$last_new_job)
# r = -0.02864176, no cor

# Relevant Experience vs Training Hours
cor(as.numeric(as.factor(cleaned$relevent_experience)) - 1,
    cleaned$training_hours)
# r = -0.01758188, no cor

# Doing sub biserial correlations between each category of the remaining
# nominal variables which have more than 2 levels
# first is major, which has Arts, Business Degree, Humanities, No Major,
# Other, and STEM
# levels(as.factor(cleaned$major_discipline))
major_partition1 <- subset(cleaned,
                        major_discipline == "Arts" |
                        major_discipline == "Business Degree")
```

```
major_partition2 <- subset(cleaned,
                        major_discipline == "Arts" |
                        major_discipline == "Humanities")
major_partition3 <- subset(cleaned,
                        major_discipline == "Arts" |
                        major_discipline == "No Major")
major_partition4 <- subset(cleaned,
                        major_discipline == "Arts" |
                        major_discipline == "Other")
major_partition5 <- subset(cleaned,
                        major_discipline == "Arts" |
                        major_discipline == "STEM")
major_partition6 <- subset(cleaned,
                        major_discipline == "Business Degree" |
                        major_discipline == "Humanities")
major_partition7 <- subset(cleaned,
                        major_discipline == "Business Degree" |
                        major_discipline == "No Major")
major_partition8 <- subset(cleaned,
                        major_discipline == "Business Degree" |
                        major_discipline == "Other")
major_partition9 <- subset(cleaned,
                        major_discipline == "Business Degree" |
                        major_discipline == "STEM")
major_partition10 <- subset(cleaned,
                        major_discipline == "Humanities" |
                        major_discipline == "No Major")
major_partition11 <- subset(cleaned,
                        major_discipline == "Humanities" |
                        major_discipline == "Other")
major_partition12 <- subset(cleaned,
                        major_discipline == "Humanities" |
                        major_discipline == "STEM")
major_partition13 <- subset(cleaned,
                        major_discipline == "No Major" |
                        major_discipline == "Other")
major_partition14 <- subset(cleaned,
                        major_discipline == "No Major" |
                        major_discipline == "STEM")
major_partition15 <- subset(cleaned,
                        major_discipline == "Other" |
                        major_discipline == "STEM")

cor(as.numeric(as.factor(major_partition1$major_discipline)) - 1,
    major_partition1$city_development_index)
cor(as.numeric(as.factor(major_partition2$major_discipline)) - 1,
    major_partition2$city_development_index)
```

```
cor(as.numeric(as.factor(major_partition3$major_discipline)) - 1,
    major_partition3$city_development_index)
cor(as.numeric(as.factor(major_partition4$major_discipline)) - 1,
    major_partition4$city_development_index)
cor(as.numeric(as.factor(major_partition5$major_discipline)) - 1,
    major_partition5$city_development_index)
cor(as.numeric(as.factor(major_partition6$major_discipline)) - 1,
    major_partition6$city_development_index)
cor(as.numeric(as.factor(major_partition7$major_discipline)) - 1,
    major_partition7$city_development_index)
cor(as.numeric(as.factor(major_partition8$major_discipline)) - 1,
    major_partition8$city_development_index)
cor(as.numeric(as.factor(major_partition9$major_discipline)) - 1,
    major_partition9$city_development_index)
cor(as.numeric(as.factor(major_partition10$major_discipline)) - 1,
    major_partition10$city_development_index)
cor(as.numeric(as.factor(major_partition11$major_discipline)) - 1,
    major_partition11$city_development_index)
cor(as.numeric(as.factor(major_partition12$major_discipline)) - 1,
    major_partition12$city_development_index)
cor(as.numeric(as.factor(major_partition13$major_discipline)) - 1,
    major_partition13$city_development_index)
cor(as.numeric(as.factor(major_partition14$major_discipline)) - 1,
    major_partition14$city_development_index)
cor(as.numeric(as.factor(major_partition15$major_discipline)) - 1,
    major_partition15$city_development_index)
# No correlations above 0.7 exist, so no correlation between major
# and city index

# Now for major and experience
cor(as.numeric(as.factor(major_partition1$major_discipline)) - 1,
    major_partition1$experience)
cor(as.numeric(as.factor(major_partition2$major_discipline)) - 1,
    major_partition2$experience)
cor(as.numeric(as.factor(major_partition3$major_discipline)) - 1,
    major_partition3$experience)
cor(as.numeric(as.factor(major_partition4$major_discipline)) - 1,
    major_partition4$experience)
cor(as.numeric(as.factor(major_partition5$major_discipline)) - 1,
    major_partition5$experience)
cor(as.numeric(as.factor(major_partition6$major_discipline)) - 1,
    major_partition6$experience)
cor(as.numeric(as.factor(major_partition7$major_discipline)) - 1,
    major_partition7$experience)
cor(as.numeric(as.factor(major_partition8$major_discipline)) - 1,
    major_partition8$experience)
cor(as.numeric(as.factor(major_partition9$major_discipline)) - 1,
```

```
    major_partition9$experience)
cor(as.numeric(as.factor(major_partition10$major_discipline)) - 1,
    major_partition10$experience)
cor(as.numeric(as.factor(major_partition11$major_discipline)) - 1,
    major_partition11$experience)
cor(as.numeric(as.factor(major_partition12$major_discipline)) - 1,
    major_partition12$experience)
cor(as.numeric(as.factor(major_partition13$major_discipline)) - 1,
    major_partition13$experience)
cor(as.numeric(as.factor(major_partition14$major_discipline)) - 1,
    major_partition14$experience)
cor(as.numeric(as.factor(major_partition15$major_discipline)) - 1,
    major_partition15$experience)
# No correlation between major and experience

# Now for major vs company size worked for
cor(as.numeric(as.factor(major_partition1$major_discipline)) - 1,
    major_partition1$company_size)
cor(as.numeric(as.factor(major_partition2$major_discipline)) - 1,
    major_partition2$company_size)
cor(as.numeric(as.factor(major_partition3$major_discipline)) - 1,
    major_partition3$company_size)
cor(as.numeric(as.factor(major_partition4$major_discipline)) - 1,
    major_partition4$company_size)
cor(as.numeric(as.factor(major_partition5$major_discipline)) - 1,
    major_partition5$company_size)
cor(as.numeric(as.factor(major_partition6$major_discipline)) - 1,
    major_partition6$company_size)
cor(as.numeric(as.factor(major_partition7$major_discipline)) - 1,
    major_partition7$company_size)
cor(as.numeric(as.factor(major_partition8$major_discipline)) - 1,
    major_partition8$company_size)
cor(as.numeric(as.factor(major_partition9$major_discipline)) - 1,
    major_partition9$company_size)
cor(as.numeric(as.factor(major_partition10$major_discipline)) - 1,
    major_partition10$company_size)
cor(as.numeric(as.factor(major_partition11$major_discipline)) - 1,
    major_partition11$company_size)
cor(as.numeric(as.factor(major_partition12$major_discipline)) - 1,
    major_partition12$company_size)
cor(as.numeric(as.factor(major_partition13$major_discipline)) - 1,
    major_partition13$company_size)
cor(as.numeric(as.factor(major_partition14$major_discipline)) - 1,
    major_partition14$company_size)
cor(as.numeric(as.factor(major_partition15$major_discipline)) - 1,
    major_partition15$company_size)
# No correlation between major and company size worked for
```

```
# Now for major and Years not working between previous and current job
cor(as.numeric(as.factor(major_partition1$major_discipline)) - 1,
    major_partition1$last_new_job)
cor(as.numeric(as.factor(major_partition2$major_discipline)) - 1,
    major_partition2$last_new_job)
cor(as.numeric(as.factor(major_partition3$major_discipline)) - 1,
    major_partition3$last_new_job)
cor(as.numeric(as.factor(major_partition4$major_discipline)) - 1,
    major_partition4$last_new_job)
cor(as.numeric(as.factor(major_partition5$major_discipline)) - 1,
    major_partition5$last_new_job)
cor(as.numeric(as.factor(major_partition6$major_discipline)) - 1,
    major_partition6$last_new_job)
cor(as.numeric(as.factor(major_partition7$major_discipline)) - 1,
    major_partition7$last_new_job)
cor(as.numeric(as.factor(major_partition8$major_discipline)) - 1,
    major_partition8$last_new_job)
cor(as.numeric(as.factor(major_partition9$major_discipline)) - 1,
    major_partition9$last_new_job)
cor(as.numeric(as.factor(major_partition10$major_discipline)) - 1,
    major_partition10$last_new_job)
cor(as.numeric(as.factor(major_partition11$major_discipline)) - 1,
    major_partition11$last_new_job)
cor(as.numeric(as.factor(major_partition12$major_discipline)) - 1,
    major_partition12$last_new_job)
cor(as.numeric(as.factor(major_partition13$major_discipline)) - 1,
    major_partition13$last_new_job)
cor(as.numeric(as.factor(major_partition14$major_discipline)) - 1,
    major_partition14$last_new_job)
cor(as.numeric(as.factor(major_partition15$major_discipline)) - 1,
    major_partition15$last_new_job)
# No correlation between major and years in between work

# Now for major and hours of training taken with company
cor(as.numeric(as.factor(major_partition1$major_discipline)) - 1,
    major_partition1$training_hours)
cor(as.numeric(as.factor(major_partition2$major_discipline)) - 1,
    major_partition2$training_hours)
cor(as.numeric(as.factor(major_partition3$major_discipline)) - 1,
    major_partition3$training_hours)
cor(as.numeric(as.factor(major_partition4$major_discipline)) - 1,
    major_partition4$training_hours)
cor(as.numeric(as.factor(major_partition5$major_discipline)) - 1,
    major_partition5$training_hours)
cor(as.numeric(as.factor(major_partition6$major_discipline)) - 1,
    major_partition6$training_hours)
```

```
cor(as.numeric(as.factor(major_partition7$major_discipline)) - 1,
    major_partition7$training_hours)
cor(as.numeric(as.factor(major_partition8$major_discipline)) - 1,
    major_partition8$training_hours)
cor(as.numeric(as.factor(major_partition9$major_discipline)) - 1,
    major_partition9$training_hours)
cor(as.numeric(as.factor(major_partition10$major_discipline)) - 1,
    major_partition10$training_hours)
cor(as.numeric(as.factor(major_partition11$major_discipline)) - 1,
    major_partition11$training_hours)
cor(as.numeric(as.factor(major_partition12$major_discipline)) - 1,
    major_partition12$training_hours)
cor(as.numeric(as.factor(major_partition13$major_discipline)) - 1,
    major_partition13$training_hours)
cor(as.numeric(as.factor(major_partition14$major_discipline)) - 1,
    major_partition14$training_hours)
cor(as.numeric(as.factor(major_partition15$major_discipline)) - 1,
    major_partition15$training_hours)
# No correlation between major and hours of training completed


# Doing sub biserial correlations between the levels of company type
# nominal variables which have more than 2 levels
# levels are Early Stage Startup, Funded Startup, NGO, Other, Pub Sect, Pvt Ltd
# levels(as.factor(cleaned$major_discipline))
company_partition1 <- subset(cleaned,
                         company_type == "Early Stage Startup" |
                         company_type == "Funded Startup")
company_partition2 <- subset(cleaned,
                         company_type == "Early Stage Startup" |
                         company_type == "NGO")
company_partition3 <- subset(cleaned,
                         company_type == "Early Stage Startup" |
                         company_type == "Other")
company_partition4 <- subset(cleaned,
                         company_type == "Early Stage Startup" |
                         company_type == "Public Sector")
company_partition5 <- subset(cleaned,
                         company_type == "Early Stage Startup" |
                         company_type == "Pvt Ltd")
company_partition6 <- subset(cleaned,
                         company_type == "Funded Startup" |
                         company_type == "NGO")
company_partition7 <- subset(cleaned,
                         company_type == "Funded Startup" |
                         company_type == "Other")
company_partition8 <- subset(cleaned,
```

```
                        company_type == "Funded Startup" |
                        company_type == "Public Sector")
company_partition9 <- subset(cleaned,
                        company_type == "Funded Startup" |
                        company_type == "Pvt Ltd")
company_partition10 <- subset(cleaned,
                        company_type == "NGO" |
                        company_type == "Other")
company_partition11 <- subset(cleaned,
                        company_type == "NGO" |
                        company_type == "Public Sector")
company_partition12 <- subset(cleaned,
                        company_type == "NGO" |
                        company_type == "Pvt Ltd")
company_partition13 <- subset(cleaned,
                        company_type == "Other" |
                        company_type == "Public Sector")
company_partition14 <- subset(cleaned,
                        company_type == "Other" |
                        company_type == "Pvt Ltd")
company_partition15 <- subset(cleaned,
                        company_type == "Public Sector" |
                        company_type == "Pvt Ltd")

# Checking correlation between individual levels of company type and city
# development index
cor(as.numeric(as.factor(company_partition1$company_type)) - 1,
    company_partition1$city_development_index)
cor(as.numeric(as.factor(company_partition2$company_type)) - 1,
    company_partition2$city_development_index)
cor(as.numeric(as.factor(company_partition3$company_type)) - 1,
    company_partition3$city_development_index)
cor(as.numeric(as.factor(company_partition4$company_type)) - 1,
    company_partition4$city_development_index)
cor(as.numeric(as.factor(company_partition5$company_type)) - 1,
    company_partition5$city_development_index)
cor(as.numeric(as.factor(company_partition6$company_type)) - 1,
    company_partition6$city_development_index)
cor(as.numeric(as.factor(company_partition7$company_type)) - 1,
    company_partition7$city_development_index)
cor(as.numeric(as.factor(company_partition8$company_type)) - 1,
    company_partition8$city_development_index)
cor(as.numeric(as.factor(company_partition9$company_type)) - 1,
    company_partition9$city_development_index)
cor(as.numeric(as.factor(company_partition10$company_type)) - 1,
    company_partition10$city_development_index)
cor(as.numeric(as.factor(company_partition11$company_type)) - 1,
```

```
    company_partition11$city_development_index)
cor(as.numeric(as.factor(company_partition12$company_type)) - 1,
    company_partition12$city_development_index)
cor(as.numeric(as.factor(company_partition13$company_type)) - 1,
    company_partition13$city_development_index)
cor(as.numeric(as.factor(company_partition14$company_type)) - 1,
    company_partition14$city_development_index)
cor(as.numeric(as.factor(company_partition15$company_type)) - 1,
    company_partition15$city_development_index)
# No correlation between company type and city index

# Now for company type and experience
cor(as.numeric(as.factor(company_partition1$company_type)) - 1,
    company_partition1$experience)
cor(as.numeric(as.factor(company_partition2$company_type)) - 1,
    company_partition2$experience)
cor(as.numeric(as.factor(company_partition3$company_type)) - 1,
    company_partition3$experience)
cor(as.numeric(as.factor(company_partition4$company_type)) - 1,
    company_partition4$experience)
cor(as.numeric(as.factor(company_partition5$company_type)) - 1,
    company_partition5$experience)
cor(as.numeric(as.factor(company_partition6$company_type)) - 1,
    company_partition6$experience)
cor(as.numeric(as.factor(company_partition7$company_type)) - 1,
    company_partition7$experience)
cor(as.numeric(as.factor(company_partition8$company_type)) - 1,
    company_partition8$experience)
cor(as.numeric(as.factor(company_partition9$company_type)) - 1,
    company_partition9$experience)
cor(as.numeric(as.factor(company_partition10$company_type)) - 1,
    company_partition10$experience)
cor(as.numeric(as.factor(company_partition11$company_type)) - 1,
    company_partition11$experience)
cor(as.numeric(as.factor(company_partition12$company_type)) - 1,
    company_partition12$experience)
cor(as.numeric(as.factor(company_partition13$company_type)) - 1,
    company_partition13$experience)
cor(as.numeric(as.factor(company_partition14$company_type)) - 1,
    company_partition14$experience)
cor(as.numeric(as.factor(company_partition15$company_type)) - 1,
    company_partition15$experience)
# No correlation between company type and experience

# Now for company type and company size
cor(as.numeric(as.factor(company_partition1$company_type)) - 1,
    company_partition1$company_size)
```

```
cor(as.numeric(as.factor(company_partition2$company_type)) - 1,
    company_partition2$company_size)
cor(as.numeric(as.factor(company_partition3$company_type)) - 1,
    company_partition3$company_size)
cor(as.numeric(as.factor(company_partition4$company_type)) - 1,
    company_partition4$company_size)
cor(as.numeric(as.factor(company_partition5$company_type)) - 1,
    company_partition5$company_size)
cor(as.numeric(as.factor(company_partition6$company_type)) - 1,
    company_partition6$company_size)
cor(as.numeric(as.factor(company_partition7$company_type)) - 1,
    company_partition7$company_size)
cor(as.numeric(as.factor(company_partition8$company_type)) - 1,
    company_partition8$company_size)
cor(as.numeric(as.factor(company_partition9$company_type)) - 1,
    company_partition9$company_size)
cor(as.numeric(as.factor(company_partition10$company_type)) - 1,
    company_partition10$company_size)
cor(as.numeric(as.factor(company_partition11$company_type)) - 1,
    company_partition11$company_size)
cor(as.numeric(as.factor(company_partition12$company_type)) - 1,
    company_partition12$company_size)
cor(as.numeric(as.factor(company_partition13$company_type)) - 1,
    company_partition13$company_size)
cor(as.numeric(as.factor(company_partition14$company_type)) - 1,
    company_partition14$company_size)
cor(as.numeric(as.factor(company_partition15$company_type)) - 1,
    company_partition15$company_size)
# There is one correlation of >0.70 in there, meaning that there is
# a relationship between company type and company size, which makes sense
# It may be a good idea to drop one of these, since I'm not familiar
# with linearly combining them to one category or Principal component analysis

# Now for company type and Years of not working between jobs
cor(as.numeric(as.factor(company_partition1$company_type)) - 1,
    company_partition1$last_new_job)
cor(as.numeric(as.factor(company_partition2$company_type)) - 1,
    company_partition2$last_new_job)
cor(as.numeric(as.factor(company_partition3$company_type)) - 1,
    company_partition3$last_new_job)
cor(as.numeric(as.factor(company_partition4$company_type)) - 1,
    company_partition4$last_new_job)
cor(as.numeric(as.factor(company_partition5$company_type)) - 1,
    company_partition5$last_new_job)
cor(as.numeric(as.factor(company_partition6$company_type)) - 1,
    company_partition6$last_new_job)
cor(as.numeric(as.factor(company_partition7$company_type)) - 1,
```

```
    company_partition7$last_new_job)
cor(as.numeric(as.factor(company_partition8$company_type)) - 1,
    company_partition8$last_new_job)
cor(as.numeric(as.factor(company_partition9$company_type)) - 1,
    company_partition9$last_new_job)
cor(as.numeric(as.factor(company_partition10$company_type)) - 1,
    company_partition10$last_new_job)
cor(as.numeric(as.factor(company_partition11$company_type)) - 1,
    company_partition11$last_new_job)
cor(as.numeric(as.factor(company_partition12$company_type)) - 1,
    company_partition12$last_new_job)
cor(as.numeric(as.factor(company_partition13$company_type)) - 1,
    company_partition13$last_new_job)
cor(as.numeric(as.factor(company_partition14$company_type)) - 1,
    company_partition14$last_new_job)
cor(as.numeric(as.factor(company_partition15$company_type)) - 1,
    company_partition15$last_new_job)
# No correlation between company type and years between


# Now for company type and training hours completed
cor(as.numeric(as.factor(company_partition1$company_type)) - 1,
    company_partition1$training_hours)
cor(as.numeric(as.factor(company_partition2$company_type)) - 1,
    company_partition2$training_hours)
cor(as.numeric(as.factor(company_partition3$company_type)) - 1,
    company_partition3$training_hours)
cor(as.numeric(as.factor(company_partition4$company_type)) - 1,
    company_partition4$training_hours)
cor(as.numeric(as.factor(company_partition5$company_type)) - 1,
    company_partition5$training_hours)
cor(as.numeric(as.factor(company_partition6$company_type)) - 1,
    company_partition6$training_hours)
cor(as.numeric(as.factor(company_partition7$company_type)) - 1,
    company_partition7$training_hours)
cor(as.numeric(as.factor(company_partition8$company_type)) - 1,
    company_partition8$training_hours)
cor(as.numeric(as.factor(company_partition9$company_type)) - 1,
    company_partition9$training_hours)
cor(as.numeric(as.factor(company_partition10$company_type)) - 1,
    company_partition10$training_hours)
cor(as.numeric(as.factor(company_partition11$company_type)) - 1,
    company_partition11$training_hours)
cor(as.numeric(as.factor(company_partition12$company_type)) - 1,
    company_partition12$training_hours)
cor(as.numeric(as.factor(company_partition13$company_type)) - 1,
    company_partition13$training_hours)
```

```
cor(as.numeric(as.factor(company_partition14$company_type)) - 1,
    company_partition14$training_hours)
cor(as.numeric(as.factor(company_partition15$company_type)) - 1,
    company_partition15$training_hours)
# No correlation between company type and training hours

# Checking Gender vs Enrolled in University
cor(as.numeric(as.factor(cleaned$gender)) - 1, cleaned$enrolled_university)
# no correlation

# Checking Related Experience vs Enrolled in University
cor(as.numeric(as.factor(cleaned$relevent_experience)) - 1, cleaned$enrolled_university)
# no correlation

# Checking Enrolled in University vs major
cor(as.numeric(as.factor(major_partition1$major_discipline)) - 1,
    major_partition1$enrolled_university)
cor(as.numeric(as.factor(major_partition2$major_discipline)) - 1,
    major_partition2$enrolled_university)
cor(as.numeric(as.factor(major_partition3$major_discipline)) - 1,
    major_partition3$enrolled_university)
cor(as.numeric(as.factor(major_partition4$major_discipline)) - 1,
    major_partition4$enrolled_university)
cor(as.numeric(as.factor(major_partition5$major_discipline)) - 1,
    major_partition5$enrolled_university)
cor(as.numeric(as.factor(major_partition6$major_discipline)) - 1,
    major_partition6$enrolled_university)
cor(as.numeric(as.factor(major_partition7$major_discipline)) - 1,
    major_partition7$enrolled_university)
cor(as.numeric(as.factor(major_partition8$major_discipline)) - 1,
    major_partition8$enrolled_university)
cor(as.numeric(as.factor(major_partition9$major_discipline)) - 1,
    major_partition9$enrolled_university)
cor(as.numeric(as.factor(major_partition10$major_discipline)) - 1,
    major_partition10$enrolled_university)
cor(as.numeric(as.factor(major_partition11$major_discipline)) - 1,
    major_partition11$enrolled_university)
cor(as.numeric(as.factor(major_partition12$major_discipline)) - 1,
    major_partition12$enrolled_university)
cor(as.numeric(as.factor(major_partition13$major_discipline)) - 1,
    major_partition13$enrolled_university)
cor(as.numeric(as.factor(major_partition14$major_discipline)) - 1,
    major_partition14$enrolled_university)
cor(as.numeric(as.factor(major_partition15$major_discipline)) - 1,
    major_partition15$enrolled_university)
# No correlation between major and enrollment in university
```

```
# Checking the company type worked for vs whether they attend Uni
cor(as.numeric(as.factor(company_partition1$company_type)) - 1,
    company_partition1$enrolled_university)
cor(as.numeric(as.factor(company_partition2$company_type)) - 1,
    company_partition2$enrolled_university)
cor(as.numeric(as.factor(company_partition3$company_type)) - 1,
    company_partition3$enrolled_university)
cor(as.numeric(as.factor(company_partition4$company_type)) - 1,
    company_partition4$enrolled_university)
cor(as.numeric(as.factor(company_partition5$company_type)) - 1,
    company_partition5$enrolled_university)
cor(as.numeric(as.factor(company_partition6$company_type)) - 1,
    company_partition6$enrolled_university)
cor(as.numeric(as.factor(company_partition7$company_type)) - 1,
    company_partition7$enrolled_university)
cor(as.numeric(as.factor(company_partition8$company_type)) - 1,
    company_partition8$enrolled_university)
cor(as.numeric(as.factor(company_partition9$company_type)) - 1,
    company_partition9$enrolled_university)
cor(as.numeric(as.factor(company_partition10$company_type)) - 1,
    company_partition10$enrolled_university)
cor(as.numeric(as.factor(company_partition11$company_type)) - 1,
    company_partition11$enrolled_university)
cor(as.numeric(as.factor(company_partition12$company_type)) - 1,
    company_partition12$enrolled_university)
cor(as.numeric(as.factor(company_partition13$company_type)) - 1,
    company_partition13$enrolled_university)
cor(as.numeric(as.factor(company_partition14$company_type)) - 1,
    company_partition14$enrolled_university)
cor(as.numeric(as.factor(company_partition15$company_type)) - 1,
    company_partition15$enrolled_university)
# No correlation


# Correlation exists between relevant experience vs company type
# as well as company type and company size
# to decrease multicollinearity, company type will be dropped from the model

cleaned <- cleaned[, -10]

############################################################
# Step 0: TESTING FOR CORRELATION OR ASSOCIATION SECTION
############################################################


# Now to create the model using Purposeful Selection
```

```
############################################################
# Step 1: INITIAL MAIN EFFECTS MODEL
############################################################

# Any solo betas with p<0.20 will be included in the main effects

model_city <- glm(target ~ city_development_index,
                  family = binomial(link="logit"),
                  data = cleaned)
summary(model_city) # p < 0.20, meaning it is significant, keep city_dev_index
                    # for main effects model
model_gender <- glm(target ~ gender,
                    family = binomial,
                    data = cleaned)
summary(model_gender) # p > 0.20, gender not significant, don't include

model_rel_exp <- glm(target ~ relevent_experience,
                     family=binomial,
                     data = cleaned)
summary(model_rel_exp) # significant, keep for model

model_enrolled <- glm(target ~ enrolled_university,
                      family=binomial,
                      data = cleaned)
summary(model_enrolled) # keep

model_major <- glm(target ~ major_discipline,
                      family=binomial,
                      data = cleaned)
summary(model_major) # some significance in STEM major

model_experience <- glm(target ~ experience,
                      family=binomial,
                      data = cleaned)
summary(model_experience) # significant, keep

model_company_size <- glm(target ~ company_size,
                      family=binomial,
                      data = cleaned)
summary(model_company_size) # not significant, drop

model_lnj <- glm(target ~ last_new_job,
                      family=binomial,
                      data = cleaned)
summary(model_lnj) # significant, keep
```

```
model_training <- glm(target ~ training_hours,
                      family=binomial,
                      data = cleaned)
summary(model_training) # training hours, not significant

# Initial main effects model includes all variables except gender and training
# hours attended for each individual

model_initial_effects <- glm(target ~ city_development_index
                             + relevent_experience
                             + enrolled_university
                             + major_discipline
                             + experience
                             + last_new_job,
                             family = binomial(link="logit"),
                             data = cleaned)

#############################################################
# Step 1: INITIAL MAIN EFFECTS MODEL
#############################################################

# Pitting the inital effects model against each of the
# individual effects as the sole predictor to see if the
# bigger model is significantly different

#############################################################
# Step 2: BACKWARDS ELIM, INITAL EFFECTS MODEL VS INDIVIDUAL
#############################################################

anova(model_city, model_initial_effects, test = "LRT")
# p-value that is less than 0.05 for the likelihood ratio test
# means that the full model is "significantly better" than the
# nested model
anova(model_gender, model_initial_effects, test = "LRT")
anova(model_rel_exp, model_initial_effects, test = "LRT")
anova(model_enrolled, model_initial_effects, test = "LRT")
anova(model_major, model_initial_effects, test = "LRT")
anova(model_experience, model_initial_effects, test = "LRT")
anova(model_lnj, model_initial_effects, test = "LRT")
# the initial effects model is better than all of these

#############################################################
# Step 2: BACKWARDS ELIM, INITAL EFFECTS MODEL VS INDIVIDUAL
#############################################################

# The next step is seeing if any of the variables that didn't
# make it in the main effects model are significant when
```

```
# added after the fact. This is also their last chance!

############################################################
# Step 3: A SECOND CHANCE FOR THE INSIGNIFICANT SOLO VARS
############################################################

model_in_ef_gender <- glm(target ~ city_development_index
                        + gender
                           + relevent_experience
                           + enrolled_university
                           + major_discipline
                           + experience
                           + last_new_job,
                           family = binomial(link="logit"),
                           data = cleaned)

model_in_ef_company_size <- glm(target ~ city_development_index
                        + company_size
                           + relevent_experience
                           + enrolled_university
                           + major_discipline
                           + experience
                           + last_new_job,
                           family = binomial(link="logit"),
                           data = cleaned)

model_in_ef_training <- glm(target ~ city_development_index
                        + training_hours
                           + relevent_experience
                           + enrolled_university
                           + major_discipline
                           + experience
                           + last_new_job,
                           family = binomial(link="logit"),
                           data = cleaned)

anova(model_initial_effects, model_in_ef_gender, test = "LRT")
# gender is not significant, and will therefore not make the model

anova(model_initial_effects, model_in_ef_company_size, test = "LRT")
# company size is significan, and will make it back into the model

anova(model_initial_effects, model_in_ef_training, test = "LRT")
# training hours is not significant, drop from model

############################################################
# Step 3: A SECOND CHANCE FOR THE INSIGNIFICANT SOLO VARS
```

```
############################################################

############################################################
# Step 4: CHECKING FOR PLAUSIBLE INTERACTION TERMS
############################################################

# For the sake of complexity, this model will only look at
# pairwise comparisons for interactions

model_all_pairwise <- glm(target ~ (city_development_index
                        + relevent_experience
                        + enrolled_university
                        + major_discipline
                        + experience
                        + company_size
                        + last_new_job)^2,
                        family = binomial(link="logit"),
                        data = cleaned)

summary(model_all_pairwise)

# Looks like the following terms have below a 0.05 level of
# significance as pairwise interaction terms
       # city_development_index:relevent_experience
       # city_development_index:enrolled_university
       # city_development_index:major_discipline
       # city_development_index:experience
       # city_development_index:last_new_job
       # experience:last_new_job
       # company_size:last_new_job

# Creating the interaction term models and testing them against the current
# full model without interaction
model_interaction_1 <- glm(target ~ city_development_index
                         + relevent_experience
                         + enrolled_university
                         + major_discipline
                         + experience
                         + company_size
                         + last_new_job
                         + city_development_index:relevent_experience,
                        family = binomial(link="logit"),
                        data = cleaned)

anova(model_initial_effects, model_interaction_1, test = "LRT")
# Significant, keep for model
```

```
model_interaction_2 <- glm(target ~ city_development_index
                         + relevent_experience
                         + enrolled_university
                         + major_discipline
                         + experience
                         + company_size
                         + last_new_job
                         + city_development_index:enrolled_university,
                         family = binomial(link="logit"),
                         data = cleaned)

anova(model_initial_effects, model_interaction_2, test = "LRT")
# Significant, keep

model_interaction_3 <- glm(target ~ city_development_index
                         + relevent_experience
                         + enrolled_university
                         + major_discipline
                         + experience
                         + company_size
                         + last_new_job
                         + city_development_index:major_discipline,
                         family = binomial(link="logit"),
                         data = cleaned)

anova(model_initial_effects, model_interaction_3, test = "LRT")
# Significant, keep

model_interaction_4 <- glm(target ~ city_development_index
                         + relevent_experience
                         + enrolled_university
                         + major_discipline
                         + experience
                         + company_size
                         + last_new_job
                         + city_development_index:experience,
                         family = binomial(link="logit"),
                         data = cleaned)


anova(model_initial_effects, model_interaction_4, test = "LRT")
# Significant, keep

model_interaction_5 <- glm(target ~ city_development_index
                         + relevent_experience
                         + enrolled_university
                         + major_discipline
```

```
                          + experience
                          + company_size
                          + last_new_job
                          + city_development_index:last_new_job,
                          family = binomial(link="logit"),
                          data = cleaned)
anova(model_initial_effects, model_interaction_5, test = "LRT")
# Significant, keep

model_interaction_6 <- glm(target ~ city_development_index
                          + relevent_experience
                          + enrolled_university
                          + major_discipline
                          + experience
                          + company_size
                          + last_new_job
                          + experience:last_new_job,
                          family = binomial(link="logit"),
                          data = cleaned)

anova(model_initial_effects, model_interaction_6, test = "LRT")
# Significant, keep

model_interaction_7 <- glm(target ~ city_development_index
                          + relevent_experience
                          + enrolled_university
                          + major_discipline
                          + experience
                          + company_size
                          + last_new_job
                          + experience:last_new_job,
                          family = binomial(link="logit"),
                          data = cleaned)

anova(model_initial_effects, model_interaction_7, test = "LRT")
# Significant, keep

# Comparing the full model with interaction terms to the models with
# single interaction terms

model_full_interaction <- glm(target ~ city_development_index
                            + relevent_experience
                            + enrolled_university
                            + major_discipline
                            + experience
                            + last_new_job
                            + city_development_index:relevent_experience
```

```
                            + city_development_index:enrolled_university
                            + city_development_index:major_discipline
                            + city_development_index:experience
                            + city_development_index:last_new_job
                            + experience:last_new_job
                            + company_size:last_new_job,
                            family = binomial(link="logit"),
                            data = cleaned)

anova(model_interaction_1, model_full_interaction, test = "LRT")
anova(model_interaction_2, model_full_interaction, test = "LRT")
anova(model_interaction_3, model_full_interaction, test = "LRT")
anova(model_interaction_4, model_full_interaction, test = "LRT")
anova(model_interaction_5, model_full_interaction, test = "LRT")
anova(model_interaction_6, model_full_interaction, test = "LRT")
anova(model_interaction_7, model_full_interaction, test = "LRT")
# Model with all 7 interaction terms is better than each model with
# single interaction term

############################################################
# Step 4: CHECKING FOR PLAUSIBLE INTERACTION TERMS
############################################################


############################################################
# Summarizing Predictive Power of Model
############################################################

prop <- sum(cleaned$target)/nrow(cleaned) # sample proportion of 1's for target
prop

predicted <- as.numeric(fitted(model_full_interaction) > prop)
# predict y=1 when est.> 0.6416

xtabs(~ cleaned$target + predicted)
# of the 8955 observations considered
    # 6081 did not seek work and were correctly predicted
    # 558 did seek work and were incorrectly predicted not to do so
    # 1391 did not seek work and were incorrectly predicted to do so
    # 925 did seek work and were correctly predicted

# Sensitivity is the probability that the model will predict a person is
# seeking employment given that they actually are, or 925/(925+558) = 0.62
# Specificity is the probability that the model will predict a person is
# not seeking employment given that they are not, or 6081/(6081+1391) = 0.81
```

```
# ROC Curve
library(pROC)
rocplot <- roc(target ~ fitted(model_full_interaction), data=cleaned)
plot.roc(rocplot, legacy.axes=TRUE) # Specficity on x axis if legacy.axes=F
auc(rocplot) # auc = area under ROC curve = concordance index

# Area under the ROC curve is 0.7672

# https://www.sciencedirect.com/science/article/pii/S1889186116300063
# I read on the internet that a good score rating system is...
# "The Area Under an ROC Curve
#  .90-1 = excellent (A)
#  .80-.90 = good (B)
#  .70-.80 = fair (C)
#  .60-.70 = poor (D)
#  .50-.60 = fail (F)
# So this model is fair

# Correlation between real and predicted outcomes
cor(cleaned$target, fitted(model_full_interaction))
# r = 0.4547879, so not strong, but useful

############################################################
# Summarizing Predictive Power of Model
############################################################
```

References

1. "Measures of Association: How to Choose". Author: Harry Khamis, PhD. (Published May, 2008) https://journals.sagepub.com/doi/pdf/10.1177/8756479308317006
2. Johnson, D.R., & Creech, J.C. (1983). Ordinal measures in multiple indicator models: A simulation study of categorization error. American Sociological Review, 48, 398-407. https://www.statisticssolutions.com/can-an-ordinal-likert-scale-be-a-continuous-variable/
3. "Predicting risk of violence through a self-appraisal questionnaire", Volume 8, Issue 2. Authors: José Manuel Andreu-Rodríguez, María Elena Peña-Fernández, Wagdy Loza, The European Journal of Psychology Applied to Legal Context. (Published 2016) https://www.sciencedirect.com/science/article/pii/S1889186116300063