Gabrielle Salamanca

Math 448, Tao He

April 22, 2022

Math 448: Project Progress Report II

I. Introduction

Names are an integral part of one's identity, and they will affect one's life in multiple ways. This is especially in the job aspect. This dataset, "Are Emily and Greg More Employable Than Lakisha and Jamal?", is from Vincent Arel-Bundock's Github projects under Rdatasets, which is sourced from *Introduction to Econometrics* by Stock and Watson. This book has sourced their data from the research paper *Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination* by Bertrand and Mullainathan.

This dataset has 4,870 rows and 27 columns containing information about the resumes of the applicants, about the job and its requirements, and the industry. The project's main objective is to see what variables are affecting the result of being called back, the variable call. We can explore if the applicants met the requirements,

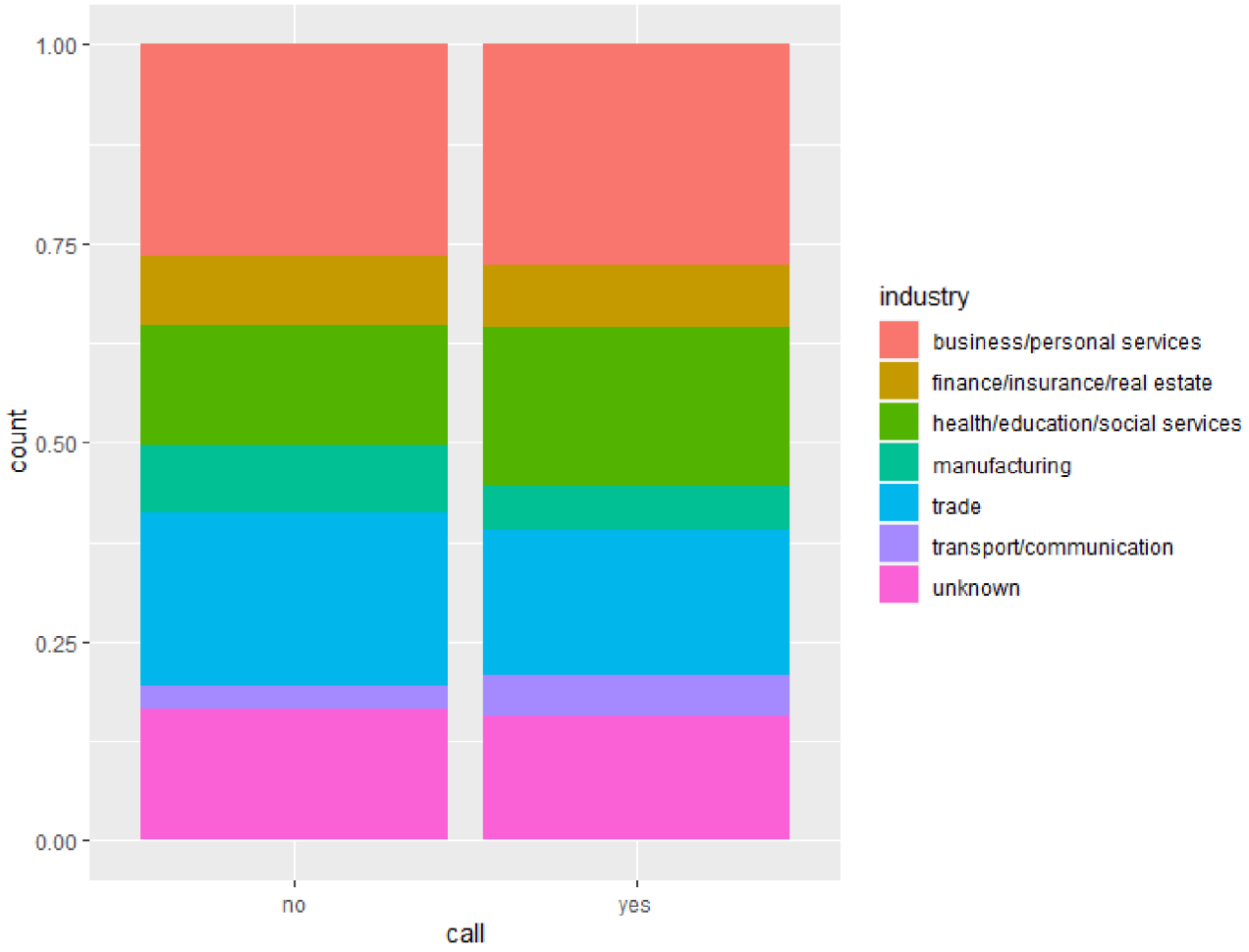
II. Data Cleaning

The first action I did was to downsample the dataset. There was an overwhelming amount of no's in the call variable, 4478 out of 4870. So, the downSample() function was used to shrink the no category down to 392 to match the yes category. I have also removed some columns: X, name, city, wanted, industry, and Class. The X variable was just the number of data entries, and the variables name, city, wanted, and industry were describing either the applicants or the type of job the application was sent to. Class was an added variable after down sampling the data, classifying if the applicant was called back or not, which was redundant.

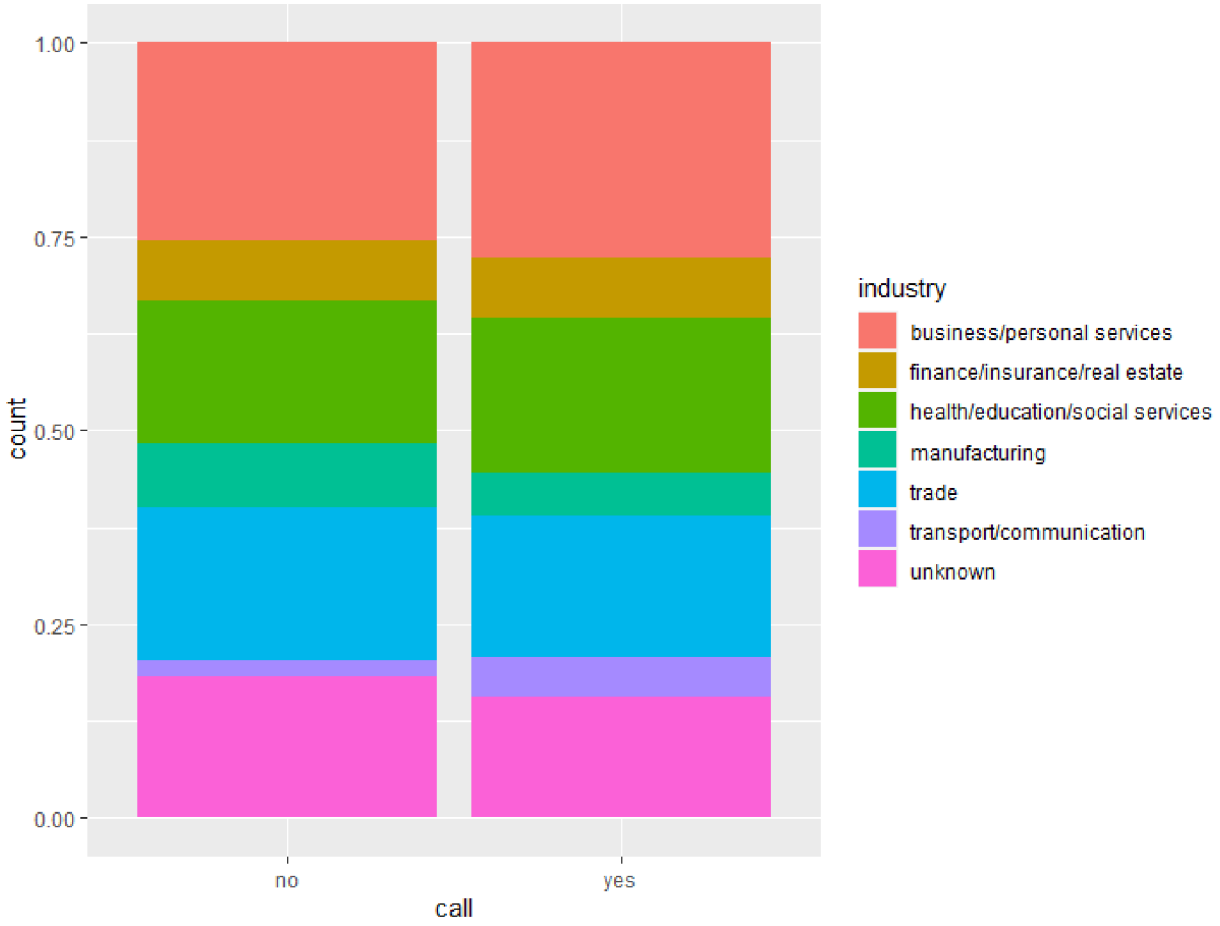
I then double-checked if there were any missing values using sum(is.na(NewNames)), and there were none thankfully. I also checked by using the View() function, and in the search bar, I typed in unknown. Thankfully, there was also none.

III. Exploratory Data Analysis

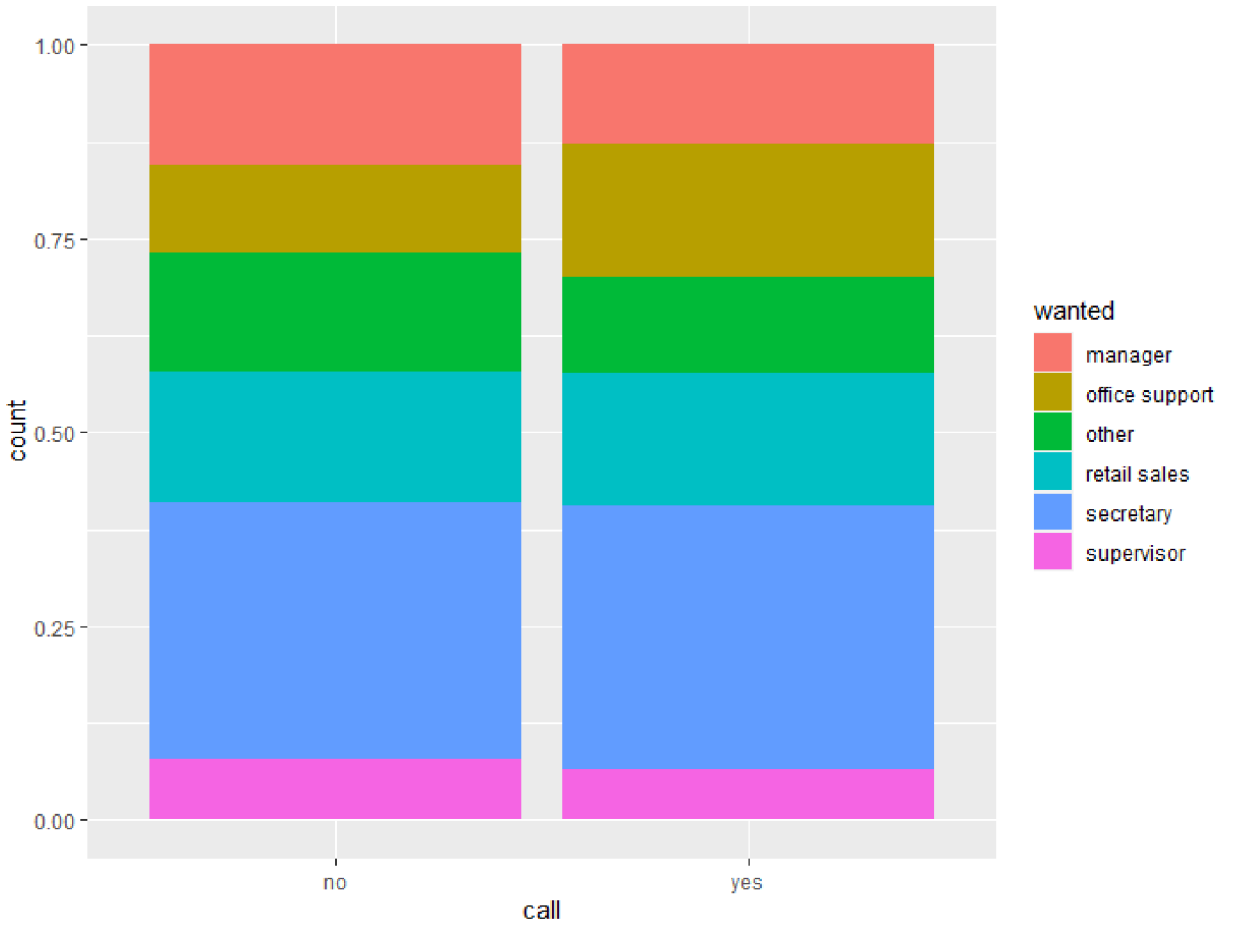
Once I tweaked my dataset, I decided to check on graphs again with the original data and the down sampled data, but keeping the industry and wanted. In a separate R script, I decided to make four plots with three variables: call, industry, and wanted. Each one had a chance to be the x variable, and the fill was either industry or wanted.



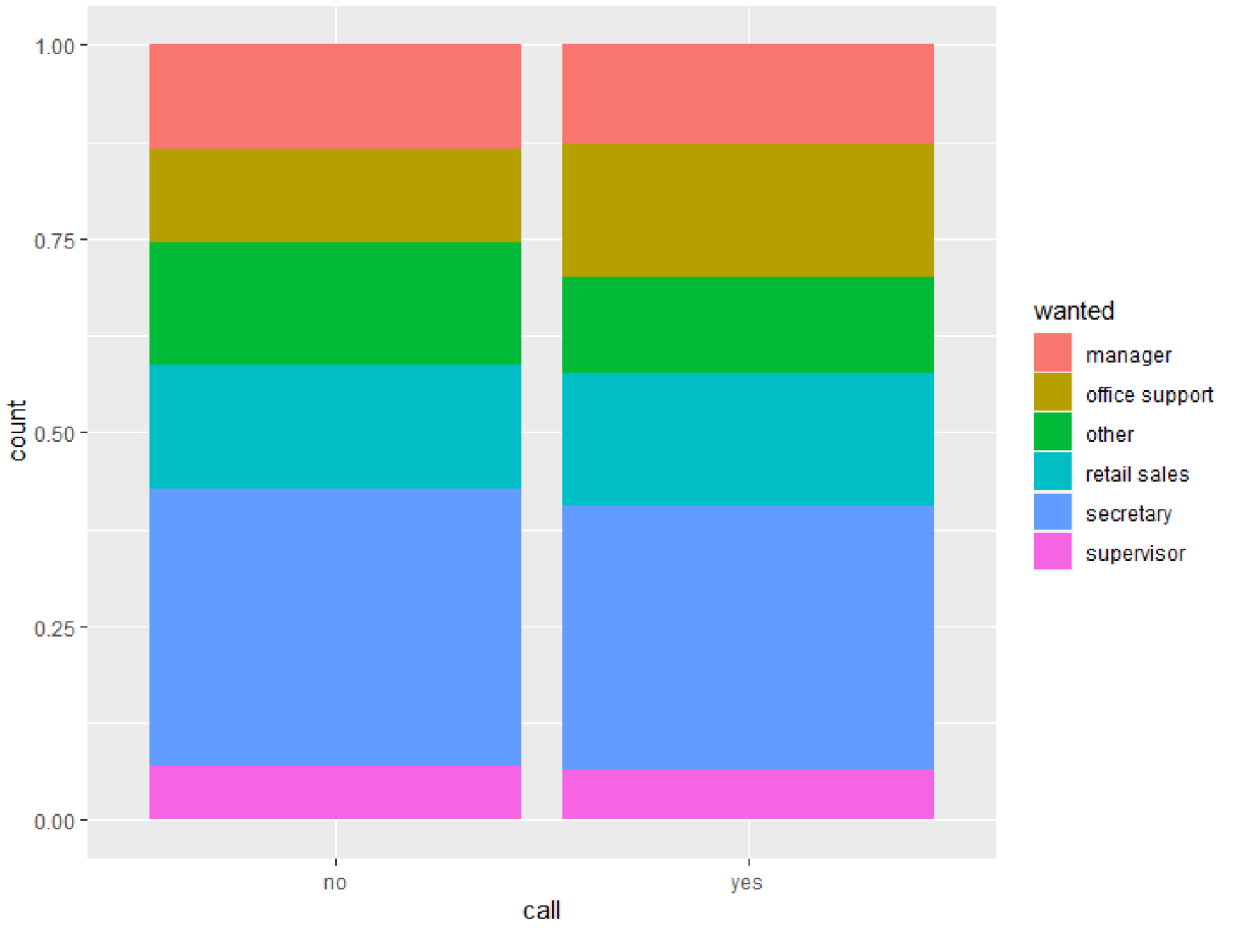
From the original dataset, called Rnames, the callbacks were relatively similar for each industry. The noticeable differences were from finance/insurance/real estate, health/educational/social services, manufacturing, trade, and transport/communication industries.



And comparing it with the down sampled dataset, called Names, there were a few that either increased or shrunk in the no column. The ones that shrunk were business/personal services, health/educational/social services, and transport/communication. The ones that increased were manufacturing, trade, and unknown. Finance/insurance/real estate and trade seemed to stay relatively the same



From Rnames, the wanted variable was mostly similar in the call variable, noting that manager, office support., and other were minorly different.



From Names, the wanted variable mostly evened out. Office support was still larger in the yes column to the no column, but secretary did increase in the no.

As for the industry and wanted graphs, they also follow a similar pattern. To note within Rnames, business/personal services, finance/insurance/real estate, health/education/social services, and unknown were mainly looking for secretaries. While manufacturing, trade, transport/communication, and unknown were mainly looking for retail sales and retail sales.

I took note of these, because depending on what the applicants applied for in a certain industry, it could affect also affect the call variable. But when these plots were generated, it seemed there wasn't a large affect, especially when we down sampled the data. For now, I will be leaving out these two variables.

IV. Selection of Performance Measure

The main objective of this project was classification, specifically predicting the call variable. So, the performance measure for this type of problem would be the accuracy, which is the percentage of correctly classified examples out of all the predictions made. Because there are still 22 variables after down sampling the dataset, I will have to find the best glm fit to use in the the feature selection.

V. GLM Fitting

Before running any of the methods, I split the data into 80% training and 20% test set. Using this, I ran multiple glm fits to find the best subset of variables that will give us the highest accuracy. Of course, with each iteration, the numbers and significant variables will change. For this project, I used glm.fit4. The variables used were ethnicity, quality, jobs, honors, volunteer, holes, school, special, reqeduc, and reqcomp. Of the ten variables, the most significant variables were ethnicity and special, which is the variables that asks if the applicant's resume mentioned some special skills. It's accuracy, when ran, was 62.422038%.

VI. Feature Selection

The main objective of this project was classification, so I would be using these classification methods: Quadratic Discriminant Analysis, Linear Discriminant Analysis, and K-Nearest Neighbors. I have also decided to choose three other methods: Best Subset Selection, Ridge Regression, and Lasso. These are linear model selection methods, and they have been chosen, because some variables use in this dataset may not actual associate with Y. So these methods are going to be used to help choose the related predictors and perhaps improve the classification accuracy.

A. Quadratic Discriminant Analysis

The first method that was ran was QDA. It performs identically to LDA, assuming that each categorical predictor has a normal distribution, however they each have a different variance. This means each class estimation has its own mean and covariance.

B. Linear Discriminant Analysis

LDA is a classification method that does the same task as logistic regression, predicting Y but with categorical variables. It assumes that each categorical predictor in the dataset has a normal distribution, and all the predictors share a variance. This means each class has its own mean, but they share a common covariance.

C. K-Nearest Neighbors

KNN is a non-parametric classification method for estimating the conditional probability by considering the K nearest neighbor-observations and look at their classes. Let's say we want to classify a data point, we would look at the nearest annotated data point, also known as the nearest neighbor. For example, we choose a Y and look at its nearest neighbors. If a majority are the color purple, then we predict that this Y is purple. Also to note, the smaller the parameter K is, the more flexible the method will be.

However, KNN does not tell us which predictors are important, there is no table of coefficients.

D. Best Subset Selection

Best subset selection is a method that aims to find the subset of variables that best predict the outcome Y, which is call in this project, by considering all possible combinations of them. Once it found the best subset, it then fits the model with it. Choosing the best subset could be choosing the predictors that have the smallest RSS or the largest R-squared.

E. Ridge Regression

Ridge regression is a shrinkage method that involves shrinking the estimates' coefficients towards zero with a shrinkage penalty. This method decreases the variance,which in turn reducing the test MSE, but it does increase the bias. It is computationally better than best subset selection, but it includes all of the predictors in the final model instead of choosing the best subset.

F. LASSO

LASSO, which stands for Least Absolute Shrinkage and Selection Operator, is an alternative shrinkage method to ridge regression. It shrinks the coefficient estimates towards 0, but in a different way. It has a penalty called L1 that has the effect of forcing some of them to be exactly equal to 0 when the tuning parameter is large. It is a sparser model compared to ridge regression, which is easier to interpret

VII. Evaluation Metrics Comparison

Algorithm	Metrics	Results
QDA	**N/A**	61.78344% accuracy
LDA	**N/A**	63.05732% accuracy
K-Nearest Neighbors	**K = 1**	57.96178% accuracy
	K = 5	61.1465% accuracy
Best Subset Selection	**Adjusted RSq**	> 0.065 at 6 variables
	Cp	5 at 6 variables
	BIC	< -20 at 3 variables
Ridge Regression	**N/A**	67.51592% accuracy
Lasso	**N/A**	66.87898% accuracy