Binghamton University - Fall 2021

Statistical Analysis as a Method to Gauge Interest in Prospective Applicants for a Data Analytics Jobs

Sofia Fasullo, Ranyerlyn Garcia, Julian Lieber, Datian Zhang

MATH 457 - Intro to Statistical Learning

Professor Vladislav Kargin

13 December 2021

Introduction

The U.S. Bureau of Labor Statistics sees strong growth for data science jobs skills and predicted that the data science field will grow about 28% through 2026, and Forbes described employment for data science as "exploding" (Schroeder, 2021). While a large portion of new data scientists are recent college graduates or just entering the workforce, there is no doubt that a significant portion of data scientists switched into the field from a previous job. That means that currently, there are several potential data scientists in other fields that need motivation to make the switch. This has drawn the attention of several employment researchers, most of all those who work for Data Analytics companies.

How can a company know which people are future employees and which are not? Unlike college graduates in new Data Science bachelor programs, older members of the workforce may not have a direct educational background in Data Science or other obvious indicators of their interest in Data Science. Furthermore, how will a company incentivize people to make the leap to apply for a new job, and one in their company?

The answer to this question for many organizations is offering Data Science bootcamps. An example of a Data Science bootcamp would be a 3-day, 15-hour immersion course in learning the basics of the R programming language. These free workshops allow for interaction between company associates and other people who may find interest in the company beyond the end of the workshop. They are also valuable opportunities to train potential employees, giving them an awareness of the type of work to expect in the company. Applicants would then be more knowledgeable about the reality of the Data Science job and would likely be more successful in the position if hired. Overall, there are a multitude of benefits for companies in providing free training workshops to non-employees.

However, these workshop trainings come at a cost to companies. They involve planning, outreach, and volunteer or paid hours on behalf of the instructors. In addition, the process of selecting which outside companies to provide training for or advertise to is a time-consuming process. If the cost of resources from a Data Science company for workshops exceeds the profit of new employee applications, then such workshops would be a poor investment. This is one of the underlying issues HR Analytics hope to answer. According to a journal article, "a number of scholars and practitioners expanded the benchmarking of HR metrics to include investments in training and developing employees, as well as in a broad array of other HR policies and practices" (Bassi, 2012). As a result, companies need accurate HR predictor models of interest in Data Analytics jobs to tailor their workshop outreach.

Dataset and Features

Our group has selected a dataset from Kaggle labeled "HR Analytics: Job Change of Data Scientists." A Data Science company wants to hire data scientists among people who complete some Data Science bootcamps, called "training" in this study. Most of these people already have jobs but may be considering leaving their jobs to apply to this company, although that is unknown. The company wants to know which of these candidates really would apply to work there after the training. This helps reduce the cost and time of research, increases quality of training, and categorizes candidates. This dataset is also designed to understand which factors might lead a person to leave their current job for a Data Analytics job.

Factors (predictor variables) were accumulated through candidates' signup and enrollment. There are over 10,000 observations in this dataset which represents each candidate. The factors represent candidates' demographics, education, experience, gender, previous company history, and training hours completed. Our response is a binary classification variable, **target**, which classes are Y=0 if a candidate wants to remain in their current job and Y=1 if a candidate is looking for a job change in the HR researcher role as well.

The goal of this study is to save the Data Science's company's resources by filtering in candidates with the highest potential and willingness for Data Science jobs. Therefore the research questions are to identify the influential factors that outputs candidates who will change jobs to work as data scientists in HR research and estimate the probability that a candidate will make the switch.

Methodology: Data Tidying and Manipulation

Firstly, the Kaggle dataset was separated into two datasets, for training and testing purposes. The test dataset, however, did not contain the 'target' response variable. This dataset would be useful for research on unsupervised learning, however for the purposes of this study we chose to focus only on the training data which contained a response. This dataset contained 19,158 observations and 14 predictor variables. When conducting regression analysis we split the training data into our own train and testing datasets. All further references on our part to training and testing data pertain to this split and not the true test dataset.

The dataset had a great deal of missing observations. This can be due to the fact people leave out information in training applications. These were represented by blank spaces which were not read by the na.omit function. We set our program to read the .csv file and consider "", or empty inputs, as missing values. Then we omitted these values. This reduced our dataset from 19,158 observations to 8,955. In other words, less than half of the people included in this dataset were being used to evaluate the research questions above. The source of the missing values could

be due to the design of the pre-bootcamp questionnaires used to obtain the data in this study; perhaps there were optional questions participants chose to ignore, resulting in missing data. Removing missing values allowed our later models to be fitted to the data without error messages. However, it is worth noting that only 1,483 of these 8,955 participants considered had 1 as their target value, or Y=1. This means that only 16.6% of observations had the desired response variable. One can see that currently, the data science bootcamps offered by this company were over 80% charity work, with no return job switch from the participants. If our models later predicted every response to be zero, the error rate would only be 16.6%. This is something we worked to minimize throughout our analysis. By minimizing this, we would ensure that participants selected through our model would have a higher return job switch rate than 16.6%.

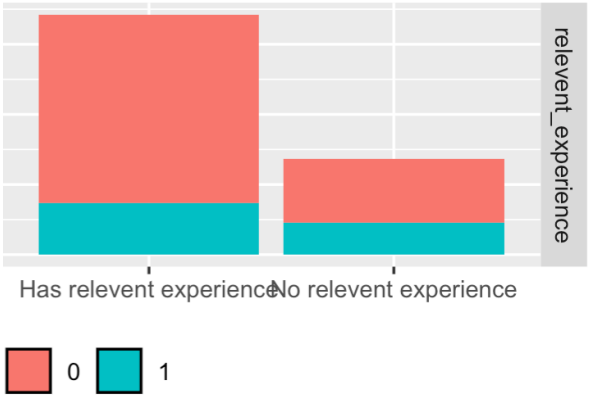
Before fitting models, we modified the data for ease of use. Since 7,989 (89%) of the 'major_discipline' variables were "STEM" and the rest were split between 5 different categories, we created using the mutate() function in the dplyr package a binary variable called 'STEM' to reduce the number of dummy variables created by 'major_discipline'. However, neither variable proved to be useful in the final models created. The variable 'company_size' had inputs of "<10", "10/49" and "50-99" among many others. These could not be read as numeric values by R and therefore the variable would be split into 8 dummy variables in a model which had no ranked order. We fixed this by removing the symbols and averaging the ranges in each category so the inputs were numbers and could be read as such. We continued this for the variables 'city_development_index', 'experience', 'last_new_job', and 'training_hours'. For 'training_hours', we also had inputs of ">20" which was converted to 20. Lastly, some variables were categorical but had a natural rank, such as 'education_level' which had categories "Graduate", "Masters" and "PhD" we converted to 1, 2, 3. We repeated this with the variables 'relevent_experience' and 'enrolled_university'.

After our data was cleaned and missing values omitted we proceeded to fit a full OLS model to the data. This created an error rate of 32% and a meager R-squared value of 0.25 R-squared values measure the proportion of variation of the response attributed to the predictors in the model, so considering this our model performed poorly. We removed the variable 'city' because it had 99 categories (thus 99 dummy variables were produced) and had no natural rank, while the variable 'city_development_index' ranked the cities in numeric value, making 'city' redundant. After refitting the OLS model, our result was a 34% error rate. As discussed before, both are worse than predicting 0 for every response.

Methodology and Results: Models

We ran a logistic regression with just training hours and experience as predictors to directly study the relationship between hours trained and experience. These variables worked in a logistic regression because the more hours trained and experience an individual has, usually the better they are at a specific job. However, after a few years of experience, the marginal value of a year of experience decreases. Also, we used the experience variable as a proxy for how likely a person is to switch jobs. Our hypothesis was that if a person was more competent, they would be willing to switch more so that they get better pay or company experience. The p value for 'training_hours' was around .01 while 'experience' was negligible. The estimate for the coefficients of training_hours was -.001 while coefficients for experience was -.079. This means that people with more years of experience and more training hours leads to people leaving their job less.

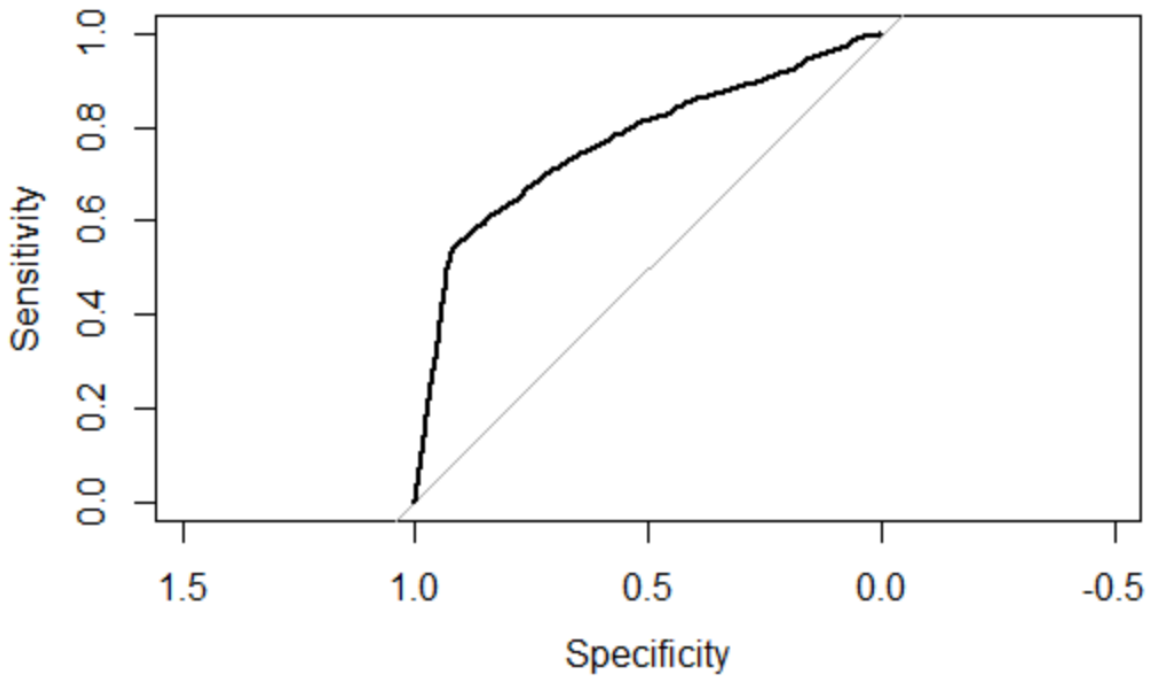
One explanation for the negative coefficients is that experience is not necessarily a good proxy for whether somebody will switch companies. People who just started their career may want to explore different companies to figure out their path, which results in a higher rate of job switching. Also, people who have more experience may be placed in better positions and are more satisfied with their job and career progression, resulting in fewer job changes. Because the p-value for 'training_hours' was higher and the coefficient closer to 0, we do believe this negative correlation is weak and conclusions about experience are more important.



A ggplot of 'city_development_index' and target shows that when the city development index is <.85 or lower, people who leave are more likely to leave their jobs (blue = leaves, red = stays), where people in cities that are more developed (higher index) are more likely to stay at their current jobs in developed cities. Later, we constructed regression trees to predict the target variable and found that the same variable was split at <0.6245 predicting 'target'=1, that people would leave their jobs.

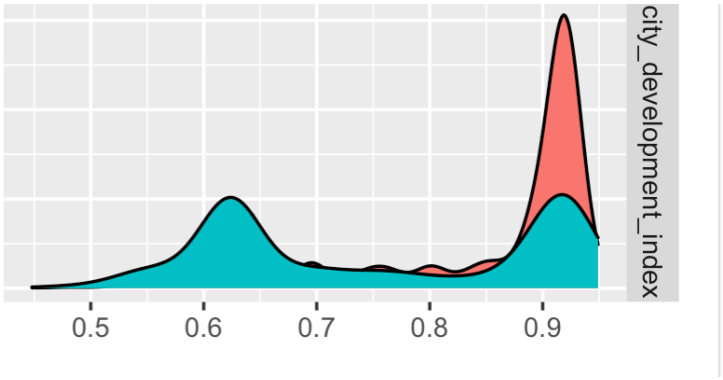
Generally, the population in developed cities is higher than in less developed cities, and it appears that the job satisfaction rate is stronger in developed cities as well. This graph may suggest that more people leave their jobs because the blue dominates for all of the <.85 indices, but as we stated before 16% of people left their job for data science so the magnitude of job retention is very large compared to job switching. One interpretation of city development and job

retention is that people who are in worse cities tend to leave and move to bigger cities, where they are happier and end up staying.



Logistic regression with these predictors alone resulted in an error rate of 12.5%. This was greatly improved from the OLS model and as we will demonstrate later, proved to be the best result from any other model generated in this study. When creating a logistic regression model for a binary variable, a threshold must be set for the classification of the predicted response. Moving this threshold changes the associated confusion matrix, and as a result some thresholds are better than others. In our case, we would prefer to have fewer false positive (Type I) errors with our analysis of professionals who want to switch their career, since there are already so many Type I errors (the rate is over 80% in the original dataset). This ROC graph below plots the true false positive rate (sensitivity) against the false positive rate (specificity). Since we want to accept fewer false positives, it appears that a threshold close to 1 would be ideal. This means that we need models that have a harsher criteria and group more people as "no"s for target, in other words, assuming that most people will not be interested in switching their jobs. The area under the curve (AUC) for the logistic regression model is 0.789, which is a good number considering that a higher AUC score is better. Overall, the logistic regression model performs very well.

We then moved on to ridge regression and the lasso. These two methods introduce a cost parameter lambda, which gives penalties to the coefficients in the model such that their values are reduced towards 0. Essentially, this makes our model consider only a few variables as important, which reduces the overall variance of the model predictions. Using the glmnet() and cv.glmnet() functions from the glmnet package, we were able to find the lambda value that produces the smallest error, and use it to fit our ridge regression model on our training data. We see a significant drop in test misclassification rate compared to our OLS model, with a new error rate of about 15.7%. This reduces the error rate below 16.5%, however it does not outperform logistic regression. With ridge regression, coefficients converge towards 0, but never quite make it. The Lasso, on the other hand, has the property of shrinking coefficients all the way to 0, thereby removing them from the model all together. In some cases, this can help reduce rate even further as compared to ridge regression, and this was found to be true in our case. Our test error rate from the Lasso model was shrunken to about 15.4%.



Although Ridge regression and the Lasso were successful in reducing our error rate, they still are far from ideal. Perhaps regression models are not the way to go for our dataset. In hopes to reduce our error rate even further, we used tree based methods to make our predictions. Using the Tree package, we first grew a tree using our training data with the tree() function. This tree only considered one variable: 'city_devleopment'. A split was made at the city_development value of .632, into two terminal nodes. This tree reduced our previous test misclassification rate significantly down to 13.4%. After taking a closer look at our predictions, it was found that we accurately predicted target values of 0 about 91% of the time, and correctly predicted target values of 1 about 60% of the time.

In hopes to reduce these rates more, we used the randomForest package to grow a random forest, with trees considering 5 variables at each split. Typically, the random forest will outperform a regular tree, because it randomly generates uncorrelated trees with a random subset of variables being considered when constructing each split of the tree. It then averages all of the predictions together to produce the most accurate result.

Unfortunately, for our dataset, this actually was not the case, our new test error rate rose to 14.7%. We tried to grow random Forests with a different amount of variables considered at each split, but in each case the error was actually raised even more. Looking closely at the confusion matrix produced, it was observed that we accurately predicted target values for 0 about

95% of the time, whereas we only predicted target values of 1 correctly about 36% of the time. It seems as though this model increased our accuracy for 0 values, but the opposite was true for 1 values. We then used the varImpPlot() function to see which variables were important, it showed that 'city_development_index' was by far the most valuable variable in our dataset, which could possibly explain why the tree that only considered city_development outperformed the random forest.

Discussion and Conclusion

We achieved a test error of 12.5% using logistic regression which best models datasets with a binary response such as ours. This has been our lowest error yet which reduced it by 4.1%. We also achieved 13.4% error using a simple classification tree which utilized the 'city_development_index' as a variable to indicate whether someone left their job to work as a Data Analyst. This improvement reduced the error rate by over 3% compared to the model which predicts that no individuals will apply to Data Analysis jobs. While these decreases in error may seem small, they are crucial to reducing the time and cost associated with investing in false positive errors, such as when a Data Analytics company devotes 150 hours to a Data Analysis training hosted at a company from which no employees apply to the Data Analytics job.

We discovered that the magnitude of people in developed cities are a lot higher than anywhere else, and high rates of job retention are coming primarily from developed cities. However, people who came from underdeveloped cities are choosing to leave their jobs because of the prospect that they will thrive more in a developed city with Data Science work opportunities. Another discovery we found is that younger professionals who just started out their careers would not mind testing their potential elsewhere so they also will be prone to leave their jobs. This will lead employers to filter in candidates with the least experience in the field but this discovery was nonsensical.

These findings significantly improved the allocation of resources by our client Data Science company in their outreach towards training for prospective employees. Human Resources researchers at Data Analytics companies should target their training to companies in cities with lower city development indices and to individuals with less professional experience. In turn, those people who have an unrealized passion for work in data science will be selected for a free Data Analytics bootcamp and get the opportunity to improve their skills and their motivation to take the leap of faith and apply for a Data Analyst position.

References:

Bassi, Laurie. “Raging Debates in HR Analytics.” *Human Resource Management International Digest*, vol. 34, no. 2, 2012, pp. 14–18., https://doi.org/10.1108/hrmid.2012.04420baa.010.

Schroeder, Bernhard. “The Data Analytics Profession and Employment Is Exploding-Three Trends That Matter.” *Forbes*, Forbes Magazine, 10 Dec. 2021, https://www.forbes.com/sites/bernhardschroeder/2021/06/11/the-data-analytics-profession-and-employment-is-exploding-three-trends-that-matter/?sh=2624b1cc3f81.