Elements of a Data Scientist’s Salary - Executive Summary

Andy Gonzalez

Introduction and Short Summary:

The project goal was to find out what characteristics or variables are most directly related to a high paying data scientist position. To collect data we ran a web scraper on indeed.com where we pulled data from job postings in over 40 U.S. cities. The data was then processed in order to create logistical elements that would best allow us to determine when a job offers pay above the median for the profession nationwide. The project concluded that aside from the more direct “data scientist” term, other features important to the decision are the terms “quantitative”, “engineer”, and “machine”. No specific skill or programming language made it into the top of important terms, however, the state of Arizona and Oregon both registered as negative influential characteristics.

Method:

The data was obtained using a web scraper that iterated over 350 pages for the term “Data Scientist”, and repeated this process for 42 U.S. cities. For each job posting the scraper found, it recorded the job title, location of the position, company name, and salary offerings. The vast majority of postings did not include salary data, and since they are the target of this project, I was only able to use entries that included said data.

The salary column was then converted into a float, and averaged if a range was provided. Salary data that came on a monthly/weekly/daily basis was converted to match the unit: dollars per year. When the salary float data was ready, I took the median of the data and converted any entry over and below that value to binary 1 and 0, respectively. This process was repeated when we were creating variables to use in the model. The rationale behind this is because we are trying to solve this problem using logistic classification, which works better with binary.

After making variables for a few key words in the job title and locations (by state) using the aforementioned process, I utilized a process that counts the words in the all of entries for job title, and ranks them according to how related they are to a successful outcome (i.e. correct prediction). The model chosen for this project is Random Forest, a classification aggregator that will use both the binary data for location and “word rank”.

Results:

The model yielded an 89% accuracy score when the predictions were compared to the test data set we generated before beginning the experiment. Among notable features considered important for the prediction, the model determined that outside of the obvious words like “data, scientist, research, analyst”, and combinations thereof, in the job title, there were a few others considered important: quantitative, machine, senior, and technician, in that order. The only state that registered among the top 15 characteristics, according to the model, is Arizona, at number 10.

When I ran the same variables over a logistic regression model using Scikit Learn, I ended up with a similar accuracy score of 84% using the same criterion as above. Although the measure of importance is not equivalent to a ranking of coefficient values, I was able to see that the characteristics my Random Forest model identified were also among the more influential variables in the logistic regression model: Quantitative, engineer, senior, Arizona, were all among characteristics that were very positively or negatively related to a successful prediction.

Below I have included 3 tables listing the words-as-features with their respective importance from the random forest model, and the words with their respective coefficients from the logistic regression model:





