Miranda Gibbons

Project 3 Written Report

In determining what features are most important in predicting salary for data science jobs, I first decided to separate salaries into two distinct categories: those below the national median, and those above the national median. That median would be determined by the data that I collected, and not by pre-existing demographic metrics. I collected job postings from the website Indeed for data scientists, and looked at those jobs that explicitly listed salary. It is important to note that of the 7000-plus salaries that I collected, only 300 contained salary information – it was approximately 4 percent of the job postings. Therefore, any conclusions we may make about what predicts salary well is from a small sample of a small sample, and whether those jobs that explicitly list salary are somehow not representative of all jobs, and whether those jobs that post regularly to Indeed are not representative of all data science jobs in the United States are all important considerations.

From my data collection, I determined median national salary for data science jobs to be $117,500. This seems high, considering external reports of data science salaries, and such a discrepancy is an important source of error to consider. I therefore was looking at what about a job posting may predict whether a salary would be above or below that number. Using two separate modeling techniques, and attempting to use multiple different features such as location, key words in the job title, and key words in the job description, I determined that key words in the description of the job was the best predictor of whether a job was a high salary or low salary job, and using a logistic regression model gave the best performance. Key words such as “algorithm”, “machine,” “data” (unsurprisingly), “analysis”, and “processing” were associated with higher salaries.

How does this information help us as a firm? We can now infer what our competitors are offering job seekers by analyzing the job descriptions, and offer a more competitive salary to win the best minds in the data scientist scene here in DC.