Exploratory Analysis of Job Skill Demand in New York City via Machine Learning Techniques

Darrell L. Nelson II, Taylor Moorman, and Wes Stanis

September 2019

Introduction

Unemployment is widely recognized as a key indicator of labor market performance and is used as a barometer to gauge overall health of an economy. When workers are unemployed, the government loses their economic contribution in terms of potential goods and services. Unemployment also decreases the population's purchasing power leading to less revenue in taxes and creates unemployment for other workers. This negative feedback loop can seriously cripple an economy.

Unemployment has also been linked with depression, increased crime, and a reduction in community service. Furthermore, in a study published by the American Journal of Public Health unemployed men made significantly more visits to their physicians, took more medications, and spent more days in bed sick than did employed individuals despite the number of diagnoses in the two groups being similar.[1] Although there are many underlying causes of unemployment, this paper will focus on qualified employment scarcity.

Qualified employment scarcity is the continued dropping rate of qualified workers in the job market. The Bureau of Labor Statistics (BLS) reported that at the end of April 2018, there were 6.7 million job openings. Many of which are mid-level or higher (experience required) or entry level with an accredited degree. By the end of May in the same year there were over 6 million people classified as unemployed, continuing a dangerous trend that could potentially see job openings eclipse the labor pool for the first time.[2] With all of the new openings, why are there so many unemployed? Does everyone just not want to work? Where is this mismatch coming from? According to the U.S. Census Bureau over a third of all adult Americans have a college degree, that equates to 70 million people. Of that 70 million, there was a 2.1% unemployment rate in April 2018 equating to 1.5 million college graduates missing from the workforce. The workforce with an associate's degree and (1-3 years) little work experience totals out to 9 million. That's 10 million certified workers with degrees that have been lost through the cracks! This project aims to match this disenfranchised workforce with the jobs that they qualify for.

[1] (LINN, SANDIFER and STEIN 1985)

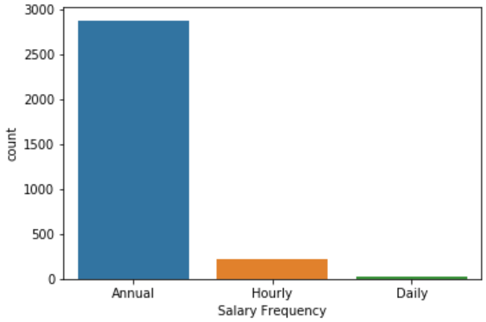
[2] (CNBC 2018)

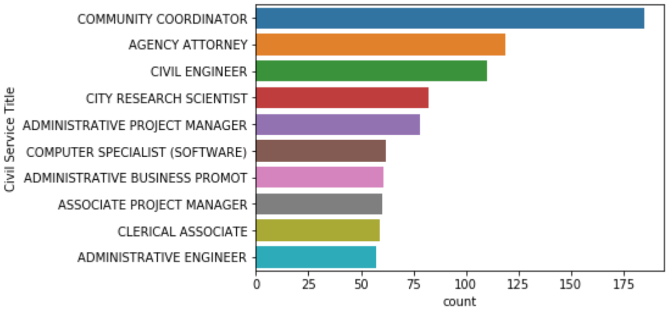
About the Data

Dataset is a comma-separated value file (CSV) comprised of 3,123 job postings available on the City of New York's official jobs site (http://www.nyc.gov/html/careers/html/search/search.shtml). Some of the key 28 attributes include: Agency, Posting Type (available to all applicants or closed off to New-Yorkers), Job Level, and Minimum Qual Requirements. The CSV was then read into a Python IDE (Google Colab) and stored as a pandas dataframe.

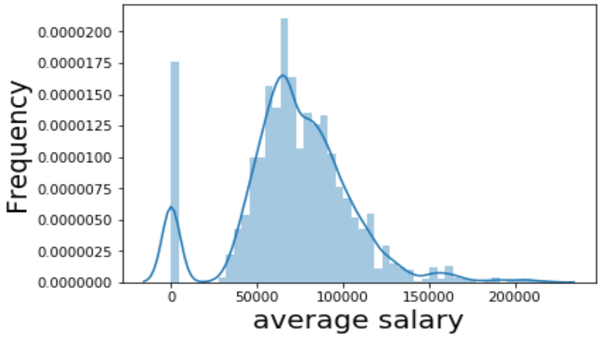
EDA and Cleaning the Data

All superfluous attributes were removed and salary frequency (how different jobs will pay these positions) and the top 10 needed positions were visualized.

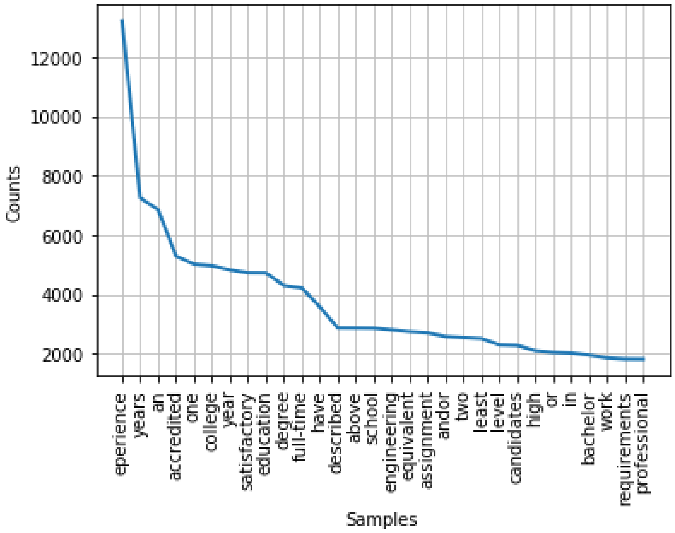




Salaries were averaged out by position from their distinct ranges and plotted.



Cleaning the textual data was very important in this dataset. There were many out of place textual characters. The minimum qualification requirements attribute was vectorized based on white space and a document matrix was created. The most prolific words in the dataset were:



Machine Learning Methods

In order to predict the salary given the qualification requirements, the average salary needed to be discretized. It was split into six categories:

1. Low $1 - $20,000
2. Low average $20,000 - $50,000
3. Average $50,000 - $70,000
4. Average high $70,000 - $100,000
5. High $100,000 - $150,000
6. Very high $150,000 - $220,000

They were labeled as numbers from 0 – 5, with 0 being low and 5 being very high.

The three models chosen to predict salary from the requirements were: 1) Multinomial Naïve-Bayes with a unigram count vectorizer, 2) Linear Kernel Support Vector Machine with a unigram count vectorizer, and 3) a SVM with a bigrams vectorizer.

Results

Naïve-Bayes

Initially a 60/40 split on the dataset received an accuracy score of 65%.

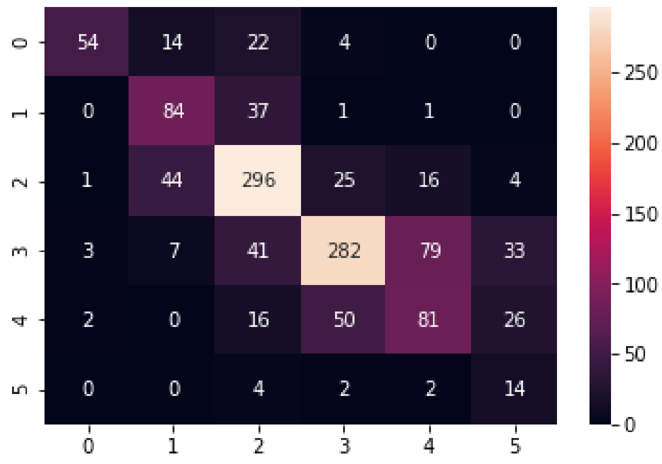
The top 10 words relating to the low salary level, and the log value that determined it are below:

```
[(-4.8329031896361085, 'science'), (-4.792081195115853, 'semester'), (-
4.776207845959563, 'must'), (-4.7683646684985375, 'equivalent'), (-
4.730045804196401, 'conditioned'), (-4.730045804196401, 'credits'), (-
4.715120153979725, 'continuance'), (-4.715120153979725, 'described'), (-
4.715120153979725, 'matriculation'), (-4.700414006590029, 'acceptable'),
```

Here are the top 10 related to very high salary:

```
[(-5.152568789984922, 'systems'), (-5.072526082311386, 'business'), (-
5.072526082311386, 'computer'), (-5.072526082311386, 'programming'), (-
5.034785754328539, 'candidates'), (-5.034785754328539, 'responsible'), (-
4.9984181101576635, 'including'), (-4.963326790346393, 'fulltime'), (-
4.896635415847721, '18'), (-4.896635415847721, 'data'),
```

Confusion matrix for model performance on testing data:



The 10-fold cross validation method was also used with an accuracy of 60%.

Support Vector Machines

The Support Vector Machine model with the bigram feature set performed better than the unigram feature set. This is likely because of the nature of this dataset with many different minimum qualifications spanning more than just a single word (e.g. “Two years” versus just “years”).

The Unigram model showed an accuracy of ~74%.

The top 10 words relating to the low salary level, and the log value that determined it:

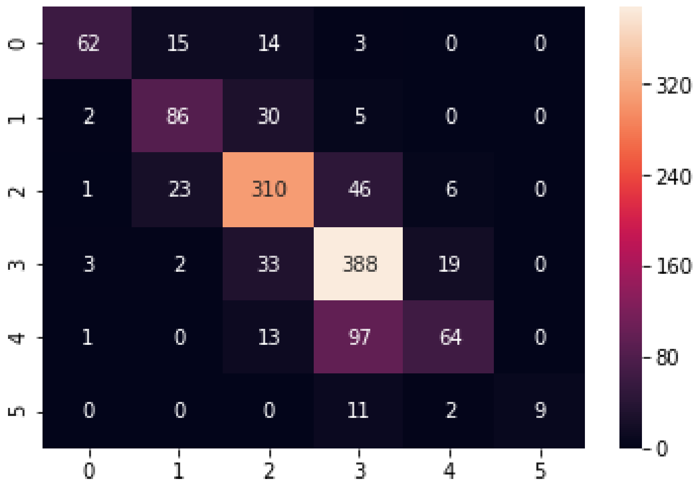
```
(0.22621662256147454, 'endorsement')
(0.2266068684981074, 'buildings')
(0.228268929466962, 'enrolled')
(0.22862705130337097, '10')
(0.22919268083370575, 'sufficient')
```

```
(0.23376226592893262, 'credited')
(0.2338161032615198, 'license')
(0.23672768660644433, 'ability')
(0.23690963247191416, 'law')
(0.24054771463484692, 'student')
```

The top 10 words related to very high salary:

```
(0.11125855044319603, 'qualify')
(0.11445306789497403, 'paid')
(0.11962543035809009, 'similar')
(0.12056152660757999, 'level')
(0.13687264882562636, 'university')
(0.14249764495366457, 'mentioned')
(0.1485700437447347, '1a')
(0.15843484965358529, 'services')
(0.15917480623633493, 'executive')
(0.16871962425576664, 'staff')
```

Here is the confusion matrix performance on the testing data:



	0	1	2	3	4	5
0	62	15	14	3	0	0
1	2	86	30	5	0	0
2	1	23	310	46	6	0
3	3	2	33	388	19	0
4	1	0	13	97	64	0
5	0	0	0	11	2	9

The Bigram model resulted in an accuracy of 75% and a cross validation accuracy of 69%.

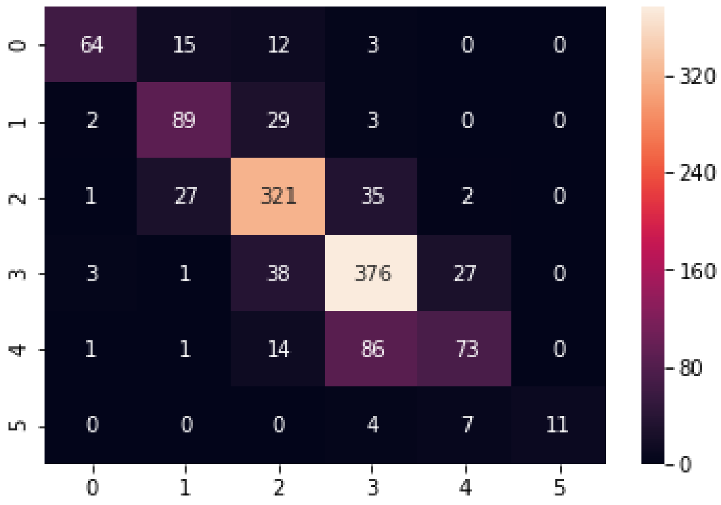
The top 10 words/bigrams relating to low salary level, and the log value that determined it:

```
(0.12753043068710007, 'university six'),
 (0.12858394333777867, 'make equivalent'),
 (0.128883530135879, 'employment'),
 (0.12894086379381656, 'program accredited'),
 (0.1308748823740406, 'license'),
 (0.1326105286771497, 'enrolled'),
 (0.13261052867714976, 'degree program'),
 (0.14103381100266157, 'school health'),
 (0.14215676022268378, 'student'),
 (0.14759400694031846, 'archival')
```

The top 10 words/bigrams related to very high salary:

```
(0.08667536494464705, 'college three'),
 (0.08848106106100671, 'organizational'),
 (0.08850575888259359, 'satisfactory fulltime'),
 (0.09198976434381526, 'practice'),
 (0.09216273110242894, 'capacity'),
 (0.09281638461905727, 'similar'),
 (0.09480112140414564, 'combination education'),
 (0.09707227164025206, '2a'),
 (0.10391908047499766, 'duties'),
 (0.10833636436768833, 'organization')
```

Here is the confusion matrix performance on the testing data:



As seen in the top features in both of the SVM models, the bigram model painted a much more accurate picture of the specific requirements needed for each job and their salary range. It simply means more to have the diction of a phrase rather than just one word to describe a requirement.

Conclusion

The SVM Model outperformed the NB model by ~9% using 10-fold cross validation. The most frequently desired skills in NYC are: experience, degree, and engineering. This should come to very little surprise as many jobs that require a specific skill-set either require experience in that field or a degree that confirms the applicant’s understanding of the fundamentals. Engineering is a trickier one to dissect due to its broad applications. Engineers can be: mechanical, civil, chemical, administrative, sanitation, and so on. Although this finding does not provide enough granularity into specific types of engineering, it does provide stiff evidence that STEM majors are highly sought after in NYC. Especially computer scientists and engineers.

Some of the top words associated with high paying jobs in NYC was computer and programming. On par with words such as: executive and business. So, not only is there a skills demand for tech-savy employees, companies are also willing to pay them a high starting salary.

For the city of New York it has been demonstrated that just by using the job requirements, the category level of pay can be estimated with a 70% accuracy. For someone looking to get a desired salary, this also provides insight into what requirements necessitate that pay. It also allows unemployed degree holders to see what types of jobs they qualify for.