CSE590-Fundamentals of Data Science

Mini Project 2

Rahul Rishi Sharma: 110347475

NYC Jobs Open Data: Salary Prediction

Stony Brook University

Introduction

In the world of ever evolving technology , the number of jobs are always increasing and so is the data related to jobs(metadata). Some common metadata features include the Title, Civil Name, Qualification etc. This project aims at analyizing the NYC jobs' data and predicts the salary based on some of the features. This model will be of great help for the employers in making a decision while seeking a job for which salary is not publicaly available.

The data taken for the project is the NYC jobs' data.

Appraoch Taken

The approach was to analyse the data and see which features are contributing more towards salary fluctuations and then find out a way to classify the data based on a set of particular features.

The project was carried out in three phases:

- Cleaning the data
- Analysis of the data
- Prediction salary from data using naive bayes classifier

Cleaning the data

The data downloaded had lot of inaccurate records which needed to be cleaned like the work location place. For example New York as work location was written nyc,NY,Ny, so all these data points were needed to be converted to a single common value which would help in further analysis. Also the salary frequency of the data was varying from Hourly to Daily to Annually. Again this values were converted into Annually form based on the legal information about the average number of working hours in New York available online. Tools used for cleaning were Microsoft excel and ipython.

Analysis of the Data

Before getting to the task of prediction the data I planned to analyze the data and see what information I could retrive which would help me in predicting the salary.

The word map generated below is for the column 'Preferred Skills'. Figure 1 clearly tells us some of the skills the companies are looking out for in the market.Words like 'Java, Software and Code' have few occurrence in the overall map as compared to words like 'Manage, Data , Analytic'. This given some idea about jobs market avalilable in New York and they mostly inclinied towards analytics as compered to core software jobs.It can be also inferred that having some prior expirience is of great advantage.

Figure 1

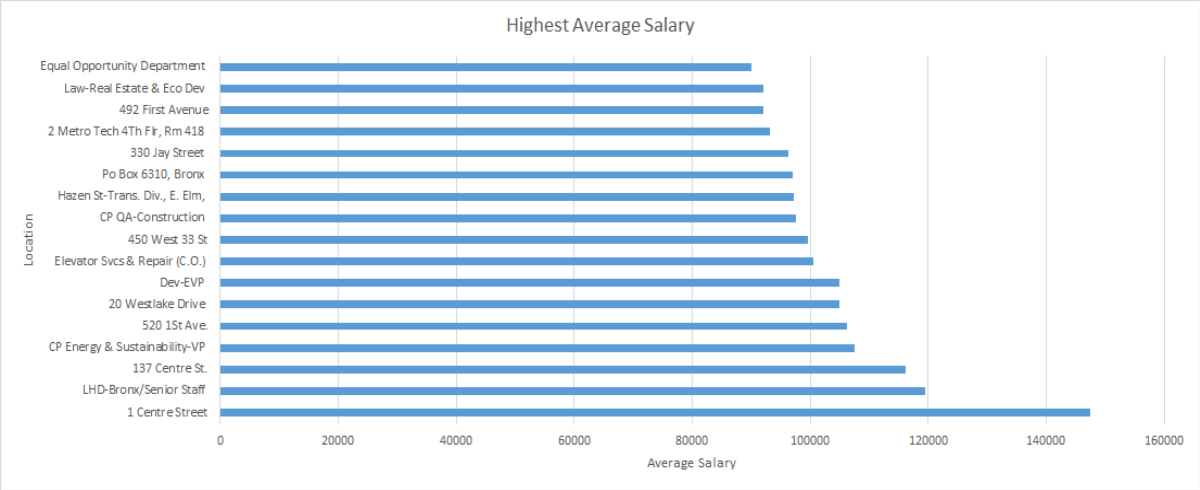


Figure 2

Figure 2 and Figure 3 help us analyze what a person can expect his average salary to be based on location.The salary range in the data set was from $28000 to $149000 anually. This images gives

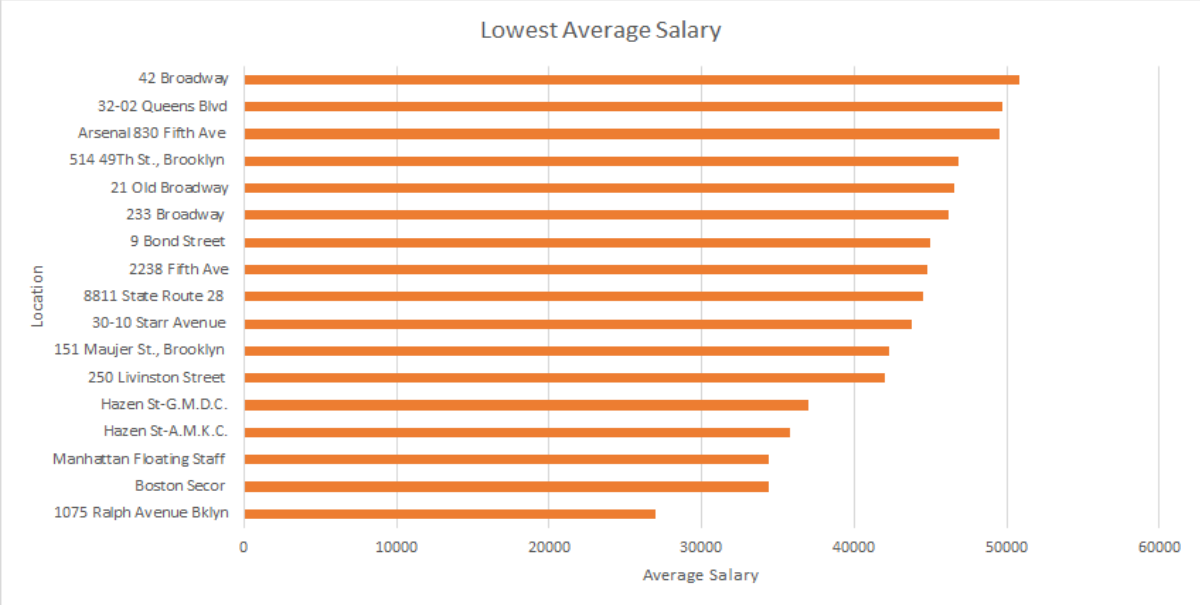


Figure 3

Figure 4 depicts the number of open positions a person can apply to based on departments.The graph containst the top 15 departments in new york government which have maximum openinigs.This gives a person an idea about what chances he has to get a job in a particular department.

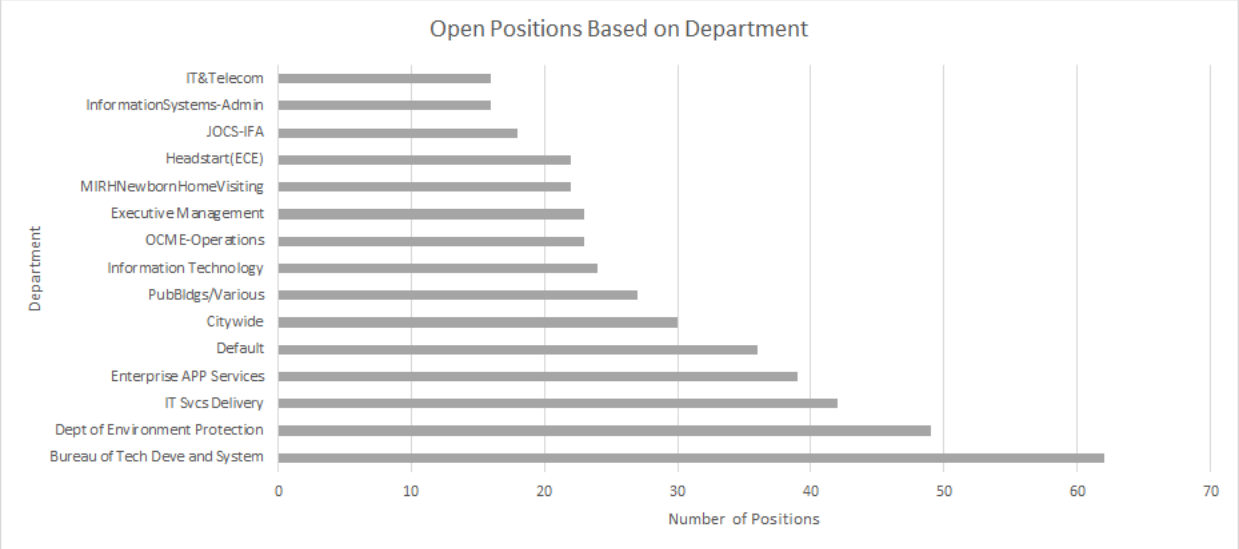


Figure 4

The reason to choose the three figures in bar chart was due to the axis label on the y – axis, these being long set of characters needed to be represented in a readble and clear way.

Salary Prediction

The prediction model developed was based on classification. I first calcluated the statistics on the salary model to come up with a classification criteria on the salary feature.

Mean	72898.94
Median	70460
Mode	77093

From the above stats it is clear the 70000 can be taken as a classification point. So for I manupilated the data by adding a new column which has a value 0 if salary is less than 7K and 1 other wise.

The approach taken for prediciton was using Naive Bayes Classifier

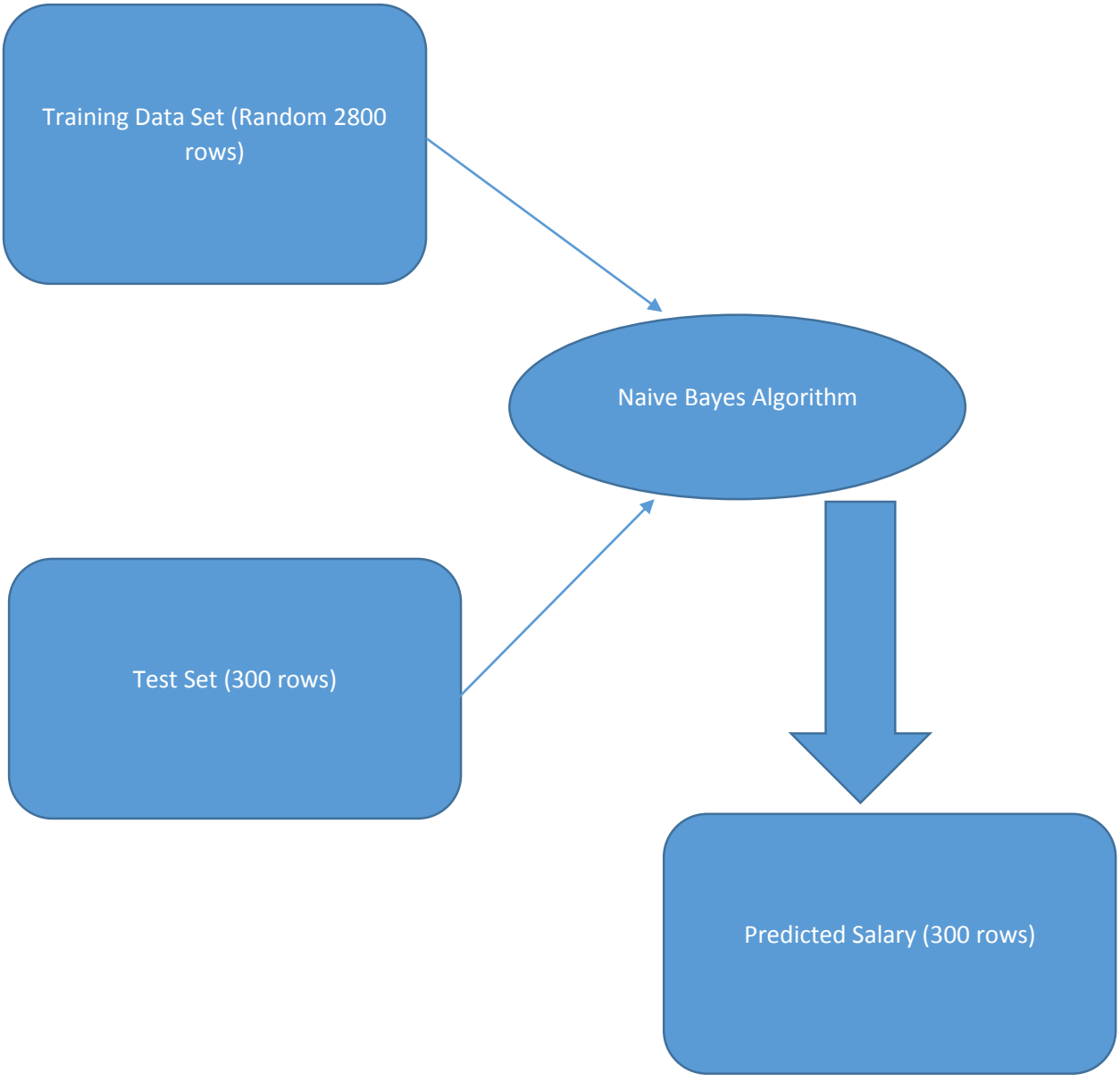
$$P(y \mid x_1, \ldots, x_n) \propto P(y) \prod_{i=1}^{n} P(x_i \mid y)$$

The above formula states that the probablity of y given the probablity of conditional independence probablity of x_i is directly propotional to the prodcut of the each feature probablity x_i given the occurrence of y.In our case the features x_i will be the coloumn based on which we will predict y(salary is above 7K or not. In order to choose conditional indepence features I took the following coloumns from the data .

Agency	Business Title	Civil Service Title	Work Location	*Salary*
DEPARTMENT OF BUSINESS SERV.	Account Manager	CONTRACT REVIEWER	110 William St. N Y	56000
DEPT OF ENVIRONMENT PROTECTION	Project Specialist	ENVIRONMENTAL ENGINEERING INTERN	253 Broadway New York Ny	52400
ADMIN FOR CHILDREN'S SVCS	Claims Analyst	ADMINISTRATIVE STAFF ANALYST	150 William Street, New York N	77632

Based on the values of Agency, Buisness Title , Civil Serive Tile and Work Location the idea is to predict whether the expected salary is above 7k or below 7k.

The data set had about 3400 rows, So I divided the training set into 3000 rows out of which 2800 will be chosed randomly over a 10 times iteration and the test set into 400 rows to avoid either accurate or vague predictions.



The text in the feature columns was tokenixed using sci-kit pakages and each column was converted into a matrix of the form [2800 X n]. n is calculated based on the how each word in a sentence gets tokenized.

vector = CountVectorizer()
agency_training_data = vector.fit_transform(agency_training_data).toarray()

The above matrix is calculated for all the features and combined together into a single matrix of the form 2800 x(n+m+..). Also I have the salary of the training data set which is a single column matrix (2800 x 1). These two sets are gived to the naive bayes algorthim with along with the test set as below.

gnb = GaussianNB()
salary_predict = gnb.fit(training_data,training_salary).predict(test_data).

The output is a matrix of the form (300 x 1) , which is the predicted salary based on the traning data set. The output is in the form of 0 and 1, where 1 is for salary greater than 70k and 0 for salary less than 70k.

On comparing the output of the the guassian classifier with the test data salary i was able to get an accurcy of approximately 73% for the 10 interations.
The reasons for doing 10 were to be able to get a good precition value, as initially the columns chosen to predict the salary were only Agency and Job tiltle but with that the accury was about 58%.

Conclusion

The project solves a major problem of salary prediction both from employee and employer prospective. In the above model there might be a small possibility of some features being dependent, The prediction could have been better if the data had more conditional independence features like college grade, work experience, previous salary etc.

References

http://scikit-learn.org/stable/modules/naive_bayes.html

https://tagul.com/