Predictive Analysis

Introduction

We used NYC jobs dataset that we got from NYC Open Data. The dataset has more than 3000 data and 28 variables. However, the dataset does not have enough numerical variables for analysis. It also has missing values and inappropriate data. Therefore, preprocessed the data by converting it into IT / Non-IT job openings according to Work Location. The current dataset has 217 data. The variables are Work Location, salary range, salary frequency, total job openings, job openings based on IT / Non-IT, Part-time, Full time.

In an exploratory analysis, we analyzed data based on 5 NYC boroughs. However, here we took the 6th category as NYC because there is some data which has work location like Office for Exec Proj Manager, Office of Public Information, Real Estate, etc. All of these work locations show up all over NYC or in some NYC boroughs. We took the 6th category for that. We analyzed data by showing the relationship between two variables or finding min-max of salary or job opening. We also explored whether job openings are part-time or full-time based on work location.

For the prediction model, we are taking specific work location instead of NYC boroughs.

Model Building

Here we are using knn classification for the job category. In this model, we are predicting whether Job openings are related to IT or Non-IT using variables average salary, job openings, and work location.

- Describe your dependent and independent variable

Ans. We took the total number of jobs as an independent variable and average salary as a dependent variable. Here, the total openings number of job openings are according to IT/Non-IT field based on work location, and salary is the average salary of that number of job openings. Work location is independent because the salary also depends on that.

- Justify your model based on your dependent variable

Ans. Here, in this mode, the dependent variable is salary. If there are any number of job openings per location, then it has a salary. Without job openings, the salary cannot exist. The model predicts the job opening is related to IT/Non-IT with salary, job openings, and location.

- Identify any preprocessing you had to go through.

Ans. A variable number of openings has 1 or 2 digit values whereas the variable salary has 5 or 6 digit values. Knn classification calculates the distance between data points, so it is necessary to set a similar distance. Hence, to normalize these variables, we scaled it between 0 and 1 range.

Also, as knn works on numbers, so convert work location into numbers by creating a dummy variable using model.matrix function.

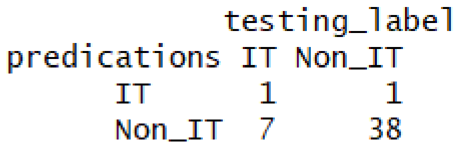
Model Results

Report on your model results

- What is the accuracy of your result?

Ans. Accuracy of the result of the model is 82%.

We can also study accuracy using the confusion matrix. For this model this is the confusion matrix:



```
              testing_label
predications IT Non_IT
      IT      1      1
      Non_IT  7     38
```

In the above matrix, columns are actual values and rows are predicted value. Diagonal values are predicted to be true and others are false predictions. Thus, from this matrix, we can know how accurately model predicts data.

- Identify any improvement processes you conducted

Ans. We changed k's value to improve accuracy. We also used a specific work location instead of NYC boroughs which also helped to improve the accuracy of the model.

- How do you interpret these results?

Ans. If we have job openings with its salary on specific work location, then we can predict whether the job opening is related to IT or Non-IT.

From the 82% model, accuracy says model predict data accurately by 82%.

In the confusion matrix, columns show actual value and rows show predicted values. Diagonal values are correctly predicted value and others are false.

Thus, in the confusion matrix, there are total job openings 39 in Non-IT which is actual value and prediction values say there are 38 Non-IT and 1 IT which model predicted for Non-IT data. Whereas actual job openings in IT is 8 and predication model shows 1 IT and 7 Non-IT job openings. So, the model predicting truly 39 out of the 47 data.

Conclusion

For predicting analysis, we predicted how Job opening is related to IT or Non-IT using three variables, which is the total number of job openings, salary and work location. If we have the total number of job openings with and its salary value on a particular location, then we can predict whether Job opening is for IT or Non-IT. From looking at our output, Job opening seems to be related to IT or Non-IT. There is 82% accuracy of the model based on the results. We also conducted the improvement process by changing k's value to improve the accuracy of the model.

To improve the model, we need more data in our dataset, and the presence of more data results in better and more accurate models. Also, we would add a more specific kind of profession in our job opening category to better improve the accuracy of the categorization of IT and Non-IT.