Predictive Analysis

Introduction

We used NYC jobs dataset which we got from NYC Open Data. 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. Now, dataset has 217 data. Its variables are Work Location, salary range, salary frequency, total job openings, job openings based on IT / Non-IT, Part time, Full time.

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

For 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 that whether Job openings are related to IT or Non-IT using variables Average salary, Job openings and work location.

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

- Justify your model based on your dependent variable
  Ans. Here, in this mode, dependent variable is salary. If there is number of job openings per location, then it has salary. Without job opening, salary cannot be existing. 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. Variable number of openings has 1 or 2 digit values whereas variable salary has 5 or 6 digit values. Knn classification calculate distance between data points, so it is necessary to set similar distance. Hence, normalize these variables, we scaled it between 0 and 1 range. In

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

Model Results

Report on your model results

- What is the accuracy of your result?
  Ans. The Accuracy of the result of the model is 82%.

  In addition, we can also study accuracy using confusion matrix. For this model this is the confusion matrix:

	testing_label	
predications	IT	Non_IT
IT	1	1
Non_IT	7	38

  In above matrix, columns are actual values and rows are predicted value. Diagonal values are predicted truly and others are false prediction. Thus, from this matrix we can know how much accurately model predict data.

- Identify any improvement processes you conducted
  Ans. We changed k's value to improve accuracy. Also, we used specific work location instead of NYC boroughs which also helped to improve 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 that the job opening is related to IT or Non-IT.

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

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

  Thus, in confusion matrix, there are total job openings 39 in Non-IT which is actual value and prediction values says 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 total number of job opening, salary and work location. If we have total number

of job openings with and its salary value on 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 are 82% accuracy of the model based on the result. And 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, the presence of more data results in better and more accurate models. In addition, we would add more kinds of profession in our Job opening category to better improve the accuracy on the categorization of IT and Non-IT.