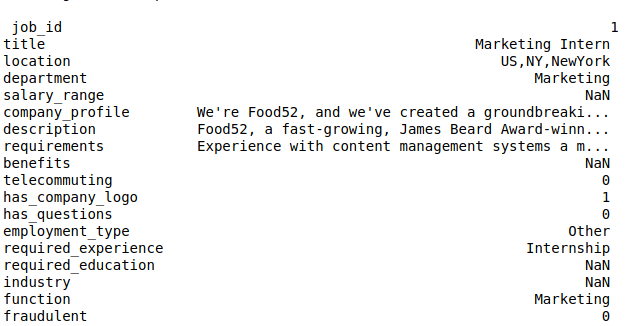
COMS3007: Machine Learning Assignment 2020

Group Members

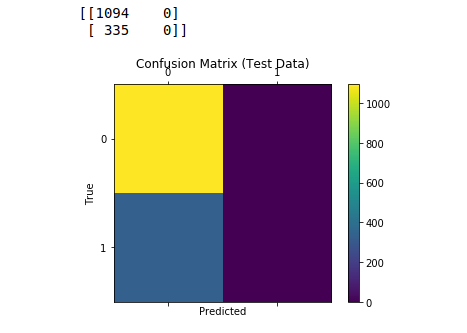
|  |  |
| --- | --- |
| Bakang Majova | 1733164 |
| Maccauley Marota | 1637447 |
| Thabo Rachidi | 1632496 |

1. Our dataset has just under 18 000 job descriptions and around 800 of them are fraudulent. The data set contains 13 string/text columns, 4 boolean(0/1) columns and an ID column. We are using this dataset to create classification models that can learn whether a job description is real or fraudulent.

|  |  |  |
| --- | --- | --- |
| Attribute | Description | Type |
| Title | The title of the job posting | String/Text |
| Location | Geographical location of the job posting. | String/Text |
| Department | Corporate department (e.g. Sales). | String/Text |
| Salary range | Indicative salary range. | String/Text |
| Company profile | A brief company description. | String/Text |
| Description | The detailed description of the job posting. | String/Text |
| Requirements | Enlisted requirements for the job opening | String/Text |
| Benefits | Enlisted offered benefits by the employer | String/Text |
| Telecommuting | True for telecommuting positions | Boolean |
| Company logo | True if company logo is present | Boolean |
| Questions | True if screening questions are present | Boolean |
| Employment type | The type of employment for the job posting (Full-type, Part-time, Contract, etc.) | String/Text |
| Required Experience | The experience required for the job posting (Executive, Entry level, Intern, etc.) | String/Text |
| Required Education | The education required for the job posting (Doctorate, Master’s Degree, Bachelor, etc.) | String/Text |
| Industry | The industry that the job posting belongs to (Automotive, IT, Health care, Real estate, etc.) | String/Text |
| Function | The function the job posting belongs to (Consulting, Engineering, Research, Sales etc.) | String/Text |
| Fraudulent | Classification attribute | Boolean |

  
Sample Data point

1. The data set we used was in a excel spreadsheet. We used the pandas library and the string functions in python to remove punctuation marks and to turn all strings to lower case. We also replaced all blank entries and entries marked as null to missing as this was easier to work with. Then we created a new excel spread sheet with the normalised and prepocessed data which we could use to train our models.  
   The data was split 60% training , 20% validation and 20% testing data.
2. Classification Algorithms we used:
   * Naive Bayes Classifier:
     + We essentially split the dataset into training, validation and testing data. We then used the training dataset to compute the prior probabilites, create a class conditional model by computing the counts of each unique feature and the probabilites of each unique feature. We could then use this to make predictions using our test data.
     + We decided not include certain features in our dataset for concerns of computation time, but had we included them our model may have performed better
     + Errors of the model in the form of a confusion matrix:



# Dataset link:

https://www.kaggle.com/shivamb/real-or-fake-fake-jobposting-prediction