1/29/2019

*Rajas Khokle, Ph.D*

General Assembly

Project Artemis

The Job Hunt By a Data Scientist

**Contents**

[**Summary** 2](#_Toc536541512)

[**General Requirements** 3](#_Toc536541513)

[**Delivery Requirements** 3](#_Toc536541514)

[**Internal Deliverables** 3](#_Toc536541515)

[WebScraping 4](#_Toc536541516)

[ETL 4](#_Toc536541517)

[EDA 4](#_Toc536541518)

[Problem Formulation 4](#_Toc536541519)

[Regression Model 4](#_Toc536541520)

[Classification Models 4](#_Toc536541521)

# **Summary**

**The project focuses on finding a job for Data scientist professional by scraping data from Job boards. The data is collected by web scraping, followed by extraction of features from Job description using NLP. Then, the models are built that answer various questions related to the skills, experience, and salary expectations for a data scientist like roles.**

# **General Requirements**

|  |  |
| --- | --- |
| Id | Requirement |
| G1 | **Determine the Factors that impact salary** |
| G2 | **Determine the Factors that distinguish Job Category** |
| G3 | **Reduce the False Positives for higher salary data** |

# **Delivery Requirements**

|  |  |
| --- | --- |
| Id | Requirement |
| RD1 | **Scrape the data from Website.** |
| RD2 | **Create and compare at least two models for each G1 and G2** |
| RD3 | **Prepare a polished Jupyter Notebook with your analysis for a peer audience of data scientists.** |
| RD4 | **A brief writeup in an executive summary, written for a non-technical audience** |
| RD5 | **Use the model to explain the tradeoffs between detecting high vs low salary positions.** |
| RD6 | **Convert your executive summary into a public blog post of at least 500 words.** |
| RD7 | **components of a job posting that can distinguish data scientists from other data jobs.** |
| RD8 | **Features are important for distinguishing junior vs. senior positions.** |
| RD9 | **Requirements for different titles with industry.** |

# **Internal Deliverables**

|  |  |
| --- | --- |
| Id | Requirement |
| ID1 | **Formulate Correct Strategy for scraping the website – Crawler, API or ad-hoc.** |
| ID2 | **Clean and Extract the Features using NLP.** |
| ID3 | **Frame Correct Regression and Classification Problem for G1 and G2.** |
| ID4 | **Use Feature Engineering and PCA for dimensionality reduction.** |
| ID5 | **Develop Multiple Models and evaluate their performance.** |
| ID6 | **Answer of requirements RD7, RD8, RD9** |
| ID7 | **A Blog Post.** |

# WebScraping

# The Data Dictionary

The data consists of following columns.

1. Job Id – From the Website
2. Internal id – Unique Id for internal purpose
3. Job Search title – The Search Keyword
4. Job Title – The returned title of the job
5. Company – The company that posted the job – Nulls allowed
6. Job Description – The job description text
7. Job Posted Date – The date on which job is posted
8. Work Type – Permanent, Contract, Casual – Nulls are allowed
9. Location – City
10. Salary – Salary offered – Nulls allowed
11. Job Classification (Industry) – The industry to which the job belongs – Nulls allowed
12. Website – URL from which the job is scraped – Nulls not allowed.
13. Super cat – Include and exclude (Categorical variable) – Nulls are allowed.
14. Super int – Percentage of pay

# ETL

The algorithm

1: Outer loop of search terms

2: Inner loop of City

3: Get the first page

4: Get the list of pages

5: Inner loop of individual pages to grab the URLS and other info

6: Saver them in a data frame and csv

7: Load the CSV and look the website column and scrape the summary from the individual websites

Base Strategy

1. Look for duplicate job posts based on summary column.
2. Find the unique value and Clean the salary column – remove $ sign. Convert the range into means, convert daily rate into annual value (val x240), convert hourly rate into annual value (val x 40x50).
3. Fill the nulls in salary based on the median of salaries for a given job title.
4. Clean the job posted data formats.
5. In Job Posted date - Add scraping date if null.

# EDA

1. Average and median salary for a given job title
2. Average and median salary for a given location
3. Histogram of salaries across job titles
4. Average and median salary based on work type
5. Average and median salary based on industry

The web scraper pulled out 1665 job ads across 4 job titles – Data Scientist, Data Analyst, Data Engineer and Business Analyst out of which only 335 had salary information. Table 1 shows the number of records for which salary is available for each job title. The jobs for data scientist and data engineer are slightly underrepresented while Data Analyst and Business Analyst position are overrepresented. It should be noted the duplicates were removed based on the summary in the job description.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Job Title | Median | Average | No. of Records | % of Record |
| Data Scientist | 100 K | 99K | 37 | 11% |
| Data Analyst | 102 K | 103 K | 114 | 34 % |
| Data Engineer | 106 K | 110 K | 67 | 20% |
| Business Analyst | 97 K | 110 K | 117 | 35% |

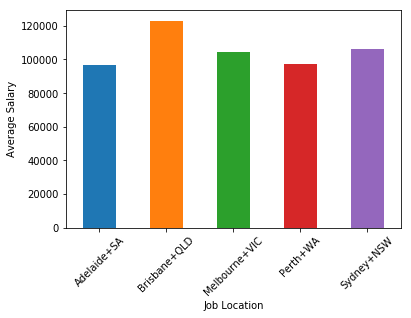
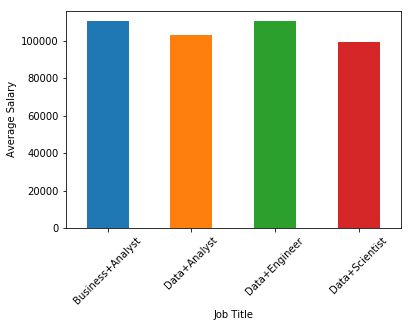
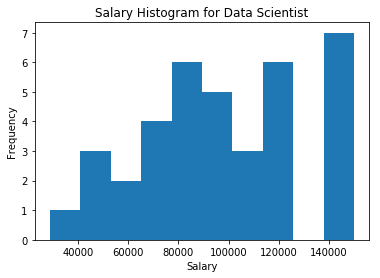
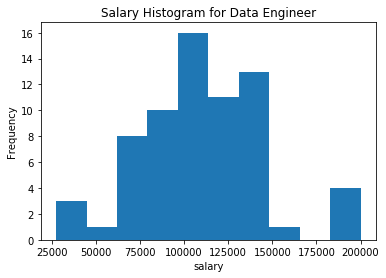
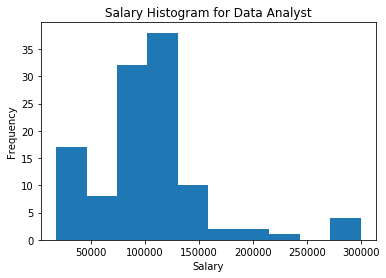
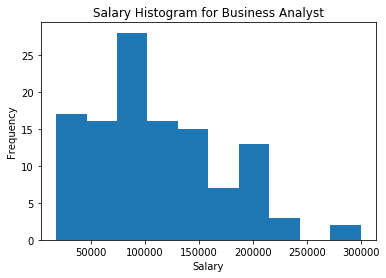


Fig. 1 (a) Average Salary for different job titles. (b) Average Salary based on different locations in Australia.

From Fig. 1b, it is seen that Adelaide and Brisbane have lowest and highest average salary respectively while average salary in Sydney and Melbourne is almost same.

1. (b)

(c) (d)

Fig.2 Histograms of salaries across different job searches. (a) Data Scientist (b) Data Engineer

(c) Data Analyst (d) Business Analyst

The Data scientist and data engineers have two distinct groups in salary. In one group, the salaries are normally distributed while there is group that has salaries around 140 K and more.

The Job summary is heavily text based so NLP is performed to extract the set of features that can be used for predicting the salary and job titles.

# NLP Steps

1. Convert text to lowercase.
2. Remove stopwords (This is iterative process. Create stop word list, lemmatize text, perform TFIDF and see the top ten features. If they contain non meaningful words, add them to the stop list and rerun)
3. Lemmatize the text
4. Perform count vectorization and TFIDF
5. Tokenize the text
6. Stem the text
7. Remove the stopwords
8. Get rid of numbers in the job description using regex
9. Find the words for different job titles
10. Vectorize (count vector) and TFIDF
11. Build Classifier for each case.

# Regression Model – Predict Salary

Basic model – Based on Title, location, industry, work type – GLM and Random Forest. None of them work well in predicting the salary as a continuous variable. Their accuracy ranges from 0.05 to 0.2 and have standard deviation of 0.2.

# Classification Models – Predict Salary Class

Target – Create low high salary as the target.

Predicting High Salary: The top characteristics of the Highly paid Data Scientist are as follows

1. Problem Solving
2. Business analyst
3. Years of Experience
4. Contract
5. Machine Learning
6. Business Process
7. Communication Skills

For solving this as a classification problem, following methods are tried and their accuracies and standard deviations of scores is provided in the table 1.

Table 1. Comparison of different model for classification of the job salary.

|  |  |  |
| --- | --- | --- |
| **Method** | **Accuracy** | **Standard Deviation** |
| Logistic Regression (1500 Features) | 0.81 | 0.09 |
| Random Forest (10 Features) | 0.85 | 0.08 |
| Bernoulli Naïve Bayes (4000) | 0.83 | 0.09 |
| KNN (neighbors = 100) | 0.65 | 0.03 |
| SVM (Degree =100) | 0.89 | 0.01 |

Since random forest is very open and interpretable, it give feature importance for the required categories which is shown in Table 2.

Table 2. Importance of Skills for highly paid salary data.

* Problem Solving Skills 17 %
* Verbal Communication Skills 14 %
* Experience 11 %
* Business Process 10 %
* Machine Learning 9 %
* Contract Based Jobs 8 %
* Stakeholder Management 7 %

Predicting the Job Title

Random Forest performs best.Table 3 shows the confusion matrix while predicting the Job titles from the job description.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Business** | **Business**  **Analyst** | **Data Analyst** | **Data Engineer** | **Data Scientist** |
| Business Analyst | 105 | 10 | 1 | 1 |
| Data Analyst | 9 | 99 | 1 | 5 |
| Data Engineer | 6 | 8 | 52 | 1 |
| Data Scientist | 1 | 1 | 1 | 34 |

Baseline: 0.35

Naïve Bayes : 0.61 (Naïve Bayes)

Accuracy Score: 0.86 (Random Forest)