# https://ga-dash.s3.amazonaws.com/production/assets/logo-9f88ae6c9c3871690e33280fcf557f33.pngProject 4: Web Scraping Job Postings

This project is a test of three major skills: collecting data by scraping a website, using natural language processing, and building a binary classifier.

### I collected salary information on data science jobs from [https://www.mycareersfuture.sg](https://www.mycareersfuture.sg/). Then using various features, such as job title, industries, job types and job description, I attempted to predict the salary of the job. Data is collected using web scraping tools such as selenium, which mimics human behavior on web browsers, and BeautiFulSoup, which allows us to parse html source code easily. This is an extremely useful way to obtain large amount of information from job posting sites. Furthermore, with salary information hidden for many jobs found online, the ability to predict them will be very useful as well.

**Webscraping**

Data was collected from individual job pages. The features collected are: job\_title, company, address, job\_type, seniority, categories, salary\_from, salary\_to, salary\_type, description, url.

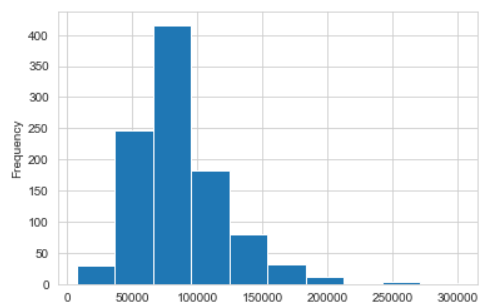
|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |

Above is a sample of the DataFrame of the data collected.

**EDA**

Data was cleaned and feature engineering done before applying models. Some Exploratory Data Analysis (EDA) was performed to get a feel of the data. As salary was given in a min to max range, average salary, taken annually, was calculated in order to provide even ground for comparison across all the different job postings. Findings of the EDA are as follows:

Distribution of the annual mean salary:

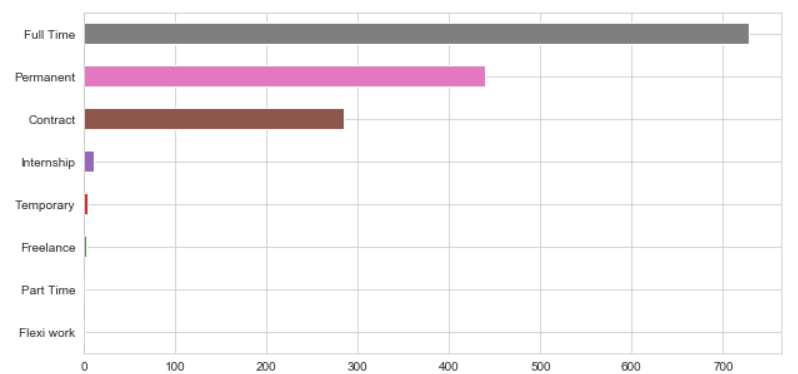


Mean Salary: $86,290.33

Median Salary: $81,000.00

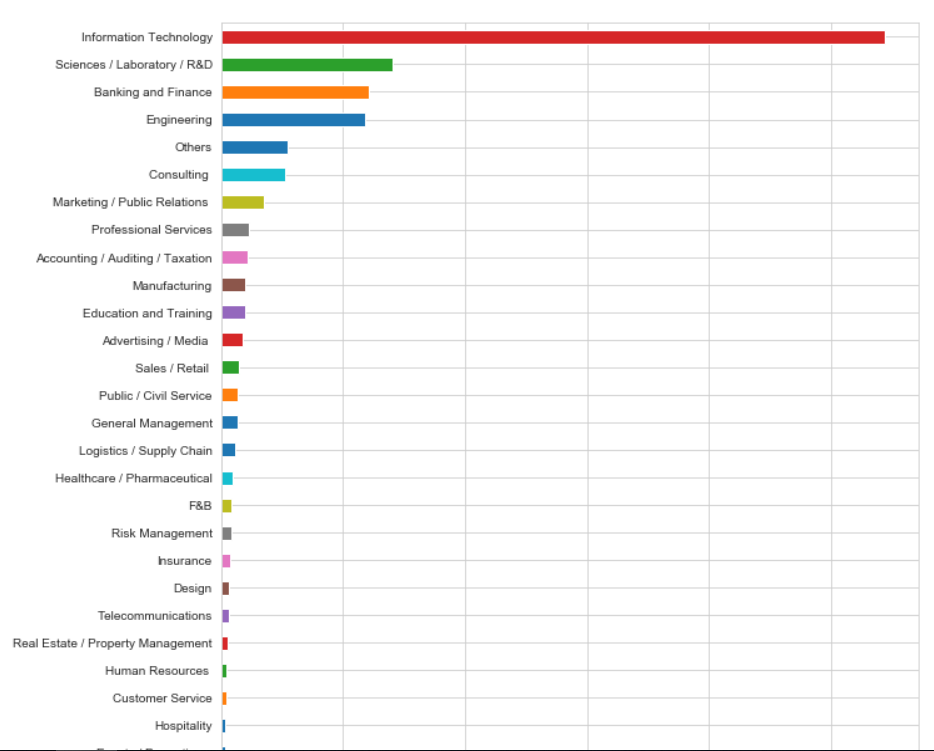
Annual mean salary was skewed to the right, meaning there were a number of outliers that were much higher than the rest of the dataset.

Number of jobs per job\_type:



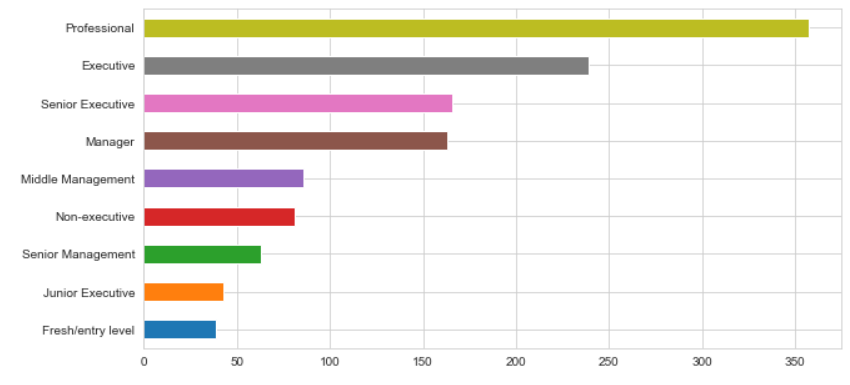
The most common type of jobs were Full Time and Permanent. Together with Contract jobs, these take up the majority of the job postings on the website.

Most common industries:



Most of the job postings were overwhelmingly in the IT industry.

Most common job seniority levels:



Most jobs were Professional in nature, with Executive level jobs the next common.

These 3 type of features were identified as huge factors in predicting job salaries, so they were chosen to be kept as dummy variables (binary values indicating which type each job belongs to) for further analysis.

**Question 1: Predicting Salary**

This problem was treated as a classification problem, with salaries above the median of $18,000.00 considered high, and below median considered low.

Count vectorizers (transformation tools that help translate words into features with the frequency they appear in the document) were used on the job titles and job descriptions.

Random Forest Classifiers (RFC) and Support Vector Machines(SVM) models were applied in this part. RFC was used it is more robust in fitting unbalanced data with outliers and reduces overfitting. SVM was used as it is also avoids overfitting the data. However, RFC has shown better scores during the fitting of this particular dataset.

As it turns out, the most important factors that help predict salaries were words in the job description, followed by the job seniority (e.g. a job that’s senior management will pay higher.)

The list of important words, along with the coefficients indicating their relative importance is as follows.

business 0.006254

bachelor degree 0.005105

development 0.004863

team 0.004637

responsible 0.004499

design 0.004405

Senior Management 0.003848

time 0.003475

guidance 0.003436

algorithms 0.003171

amended 0.002991

leadership 0.002976

prevailing recruitment 0.002966

business problems 0.002948

based singapore 0.002913

value 0.002849

sets 0.002741

senior 0.002625

spark 0.002580

role 0.002556

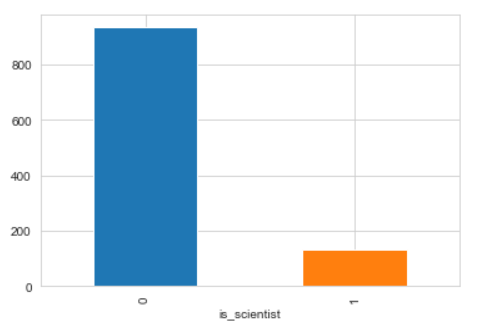
Overall, qualifications such as bachelor degree and business related roles rank at the top in terms of importance in getting high salaries.

Scores and predicted values can be found in the accompanying Jupyter notebook.

**Question 2: Predicting job roles**

Next part of this project is about predicting job roles. This is further split into investigating whether a role is a data scientist or not, and whether it’s a junior role or senior role. More feature engineering was done to create a binary feature which indicates whether a role is a scientist or not, as well as whether a role is senior or not.

For the first part, Logistic Regression was used. This is because the resulting dataset was very unbalanced.



In the above figure, 1 means those jobs are scientist roles. 0 means they are not. As shown, there were a lot more non-scientist roles than scientist roles.

Logistic Regression has a built in weight class argument that helps balance the weighting of the minority class, which is why it was applied here.

Logistic Regression scored very well, at 94% accuracy, although that probably had to do with the fact that ‘scientist’ jobs are defined as having the word ‘scientist’ in the job title. While that particular word was taken out as a feature during analysis, there remain other related words like root words ‘Science’ and ‘Laboratory’, which led to very easy classification by the model.

Distinguishing between junior and senior roles was more difficult. Multiple features were lumped together to create new features as target variables. In particular, junior roles were defined as roles classified under 'Fresh/entry level', 'Junior Executive', and 'Executive', while senior roles were defined as 'Senior Executive', 'Manager', and 'Senior Management', where senior roles were given a value of 1 and junior roles a 0. All other classifications were dropped for this analysis.

As Naïve Bayes model was better at Natural Language Processing, this model was applied here. The dataset itself was more balanced than the previous problem. The model scored just below 67% accuracy, and the confusion matrix was given as follows:

|  | **Predicted**  **0** | **Predicted**  **1** |
| --- | --- | --- |
| **True**  **0** | 42 | 26 |
| **True**  **1** | 27 | 65 |

The key factor in determining whether roles were senior or not depended on the job type. Full Time and Permanent job tended to be more for senior roles. The top 10 list of key words are listed below:

business -0.297391

Full Time -0.421904

Permanent -0.576976

development -0.609766

Information Technology -0.636794

big data -1.066524

computer science -1.166054

manager -1.329941

senior -1.371906

complex -1.493266

As for whether are there any differences affecting roles across different industries, 3 industries were picked:  'Engineering', 'Sciences/Laboratory/R&D' and 'Banking and Finance'. These 3 industries were picked as they have roughly the same number of occurences in the dataset.

A pipeline was set up to model the data for these 3 industries separately. As it turned out, 2 of them had the same unbalanced data problem as the previous one. It was decided to use Logistic Regression again for all 3 of them.

Logistic Regression scored very well, with over 90% accuracy for all 3 industries. There were a few key differences for the key factors influencing role seniority in each sector though. For Science / Laboratory / R&D Industries, the key factors affecting job roles tend to be fluency in big data or real time data. Job types also play a significant role. For Banking and Finance however, data scientist and computer science play a bigger role in setting salary. For Engineering, the most important feature is the same as Banking and Finance, data scientist. However, besides that, other important skills and knowledge seem to be in the field of Business and Insurance. While the ranking might vary, all 3 jobs seem to share the same top few features. From Engineering and Banking and Finance sectors, it seems that pay is related to knowledge outside of the domain.

Overall, salary seemed to be tied to skills, particularly in big data, as well as experience, in particular those over 10 years. A candidate with a wide range of skills, both in information technology as well as business sense, will do well in this job market, based on the analysis of current data science related jobs.