Project 4 - Job salary and title prediction

By Jason Wu

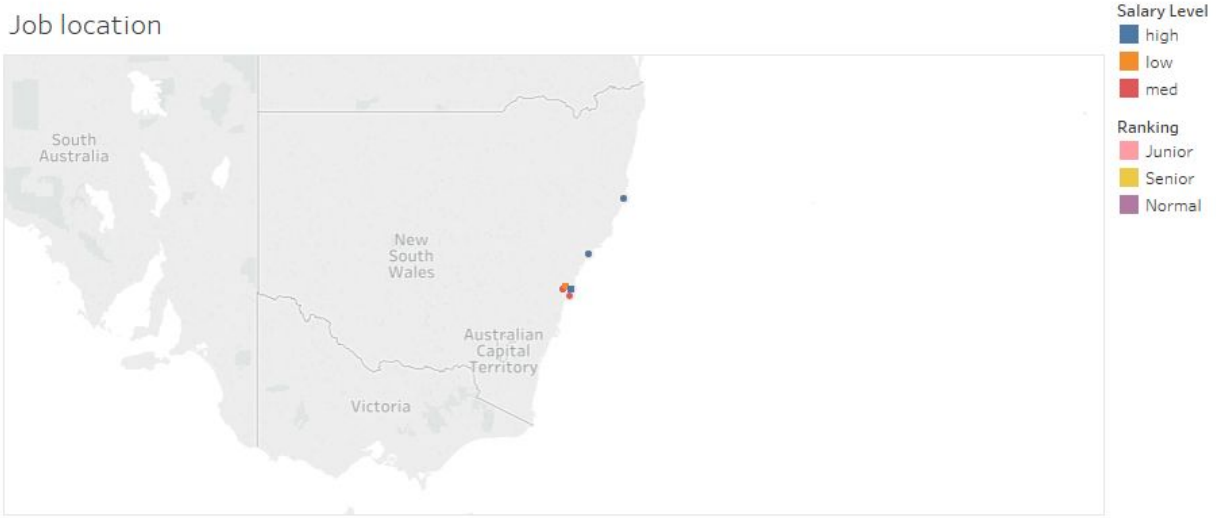
Obtaining the data

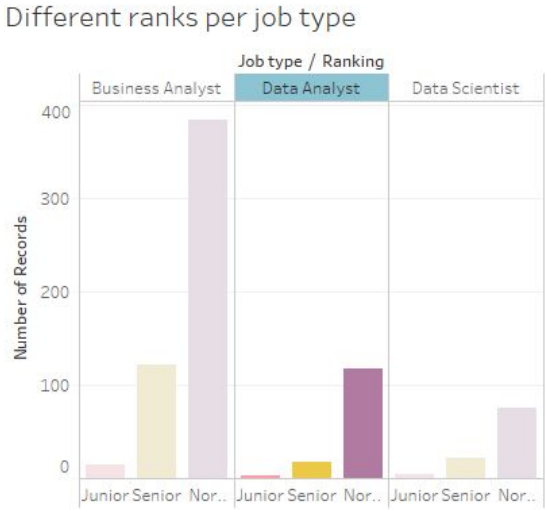
- Jobs scrapped from Indeed
  - Python - programming language used
  - BeautifulSoup - used to parse the html I scrapped
  - Selenium - the headless browser I used to load the website due to the use of widespread use of javascript
- Focused on business analyst, data analyst and data scientist across Australia
- Looked for job title, job description, salary and location of the job

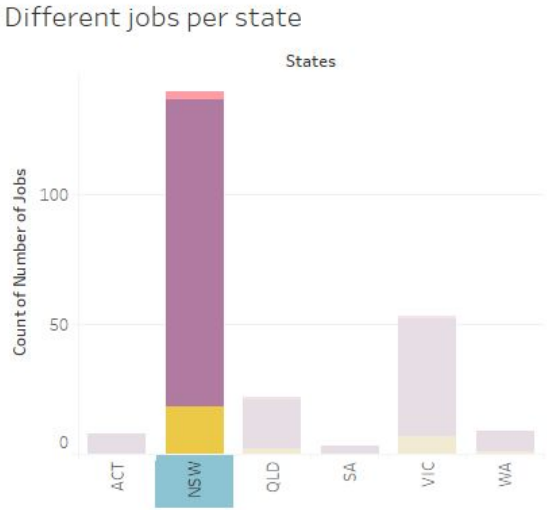
Cleaning

Started with a grand total of 8000 job listings and 4 columns, after removal of duplicates, irrelevant jobs and exploration, I was left with 1200 jobs and 9 columns.

	description	location	salary	title
89	The Company\r\n\r\nAre you currently in-betwee...	- Sydney NSW	BI AdministratorRobert Half Australia6,884 rev...	BI Administrator
389	3 Month Contract\r\nDesktop Support\r\n$36.50 ...	- Sydney NSW	Desktop Support OfficerFinite IT - Sydney NSW$...	Desktop Support Officer
451	Seeking Financial Accountant to assist the Fin...	- Sydney NSW	Financial AccountantMorgan McKinley19 reviews ...	Financial Accountant
480	Heavy vehicle mechanic needed for Afternoon or...	- Arncliffe NSW	Heavy Vehicle Mechanic (Afternoon or Bi weekl...	Heavy Vehicle Mechanic (Afternoon or Bi weekl...
490	ASAP – 15/02/2018\r\nGladesville\r\n$62 per Ho...	- Sydney NSW	Marketing & Communications OfficerFinite IT - ...	Marketing & Communications Officer







Visualisations

- Demo of tableau dashboard
- Data science roles across Australia seemed concentrated along the east coast
- There are 5 junior data science roles within NSW!
- Much higher demand for data analysts and business analysts within NSW

Text analytics

Business Analyst

- Focus on communication skills

- Client and customer base

Business Analyst

Text analytics

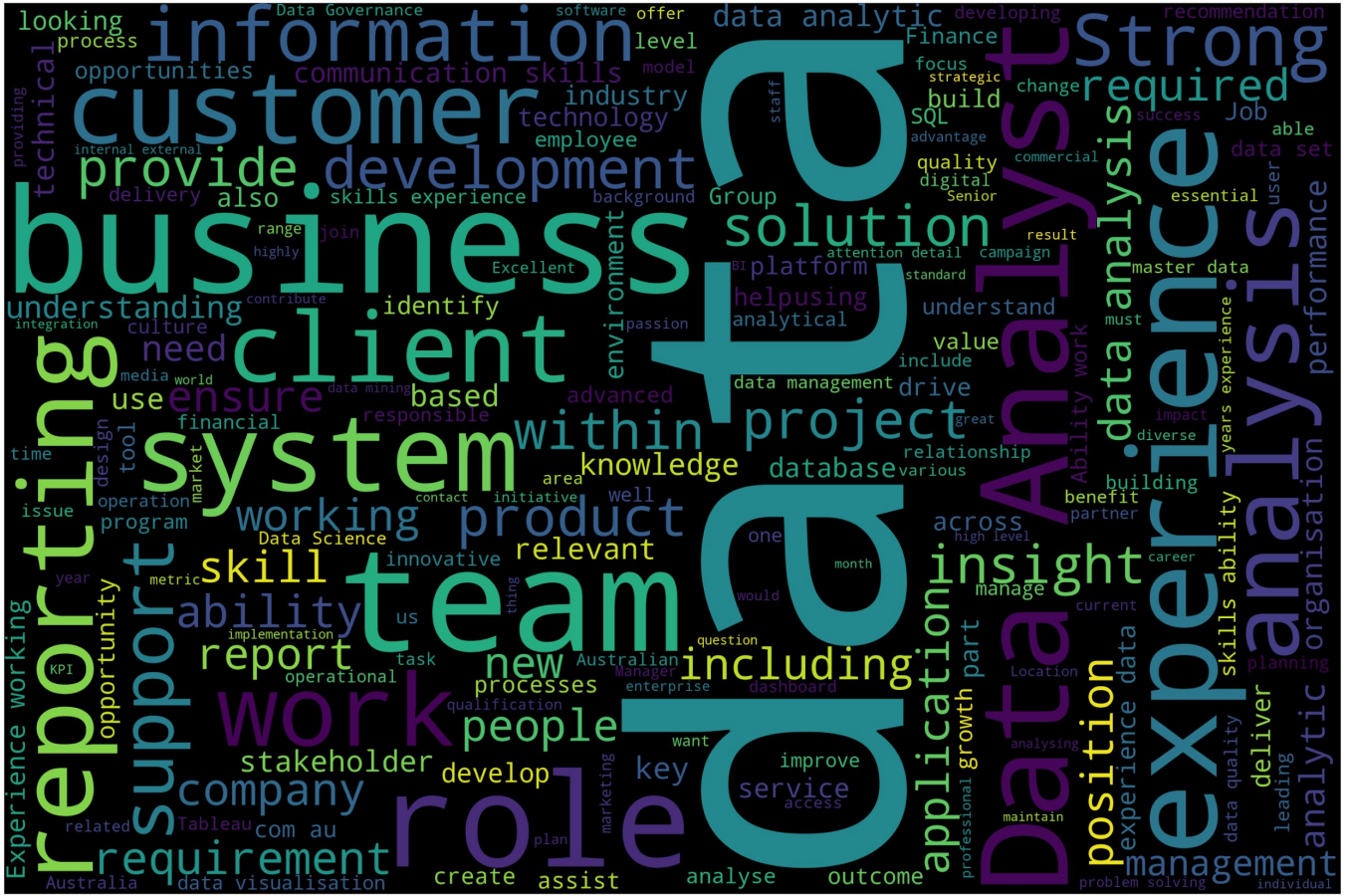
Data Analyst

- Focus on reporting

- Working in a team environment

- slight database and SQL knowledge

Data Analyst



Text analytics

Data Scientist

- Machine learning

- Big data, sql, python

- Advanced analytics and insights

Data Scientist

Question 1 - predicting salary

- A total of 4 models fit (linear regression, lasso linear regression, random forest regressor and XGboost)
- Best regression model was random forest

```
Lasso train score: 0.31332890180632267
Lasso test score: 0.15479666738896813
RMSE: 29164.320012346598
```

```
Random forest train score: 0.9032714218326503
Random forest test score: 0.23701563063879216
RMSE: 27709.52461573283
```

- Only 0.237 of the variance within the salary could be explained by the description and location
- RMSE means on average my predictions were off by $27,709!

Question 1

- Made a mistake by trying to take it as a regression problem
  - Due to how I imputed salaries, I effectively made it a classification problem without realising
- When taken as a classification the result was much better
  - Baseline was 0.64

	precision	recall	f1-score	support
high	0.66	0.61	0.63	61
low	0.75	0.68	0.71	66
med	0.85	0.88	0.86	268
avg / total	0.80	0.81	0.80	395

```
Variable: data analyst          Importance: 0.12
Variable: senior busi           Importance: 0.11
Variable: busi analyst          Importance: 0.06
Variable: senior busi analyst   Importance: 0.04
Variable: commun skill          Importance: 0.02
Variable: data analysi          Importance: 0.02
```

Question 2 - predicting job field

- Predicting either business analyst, data analyst or data scientist
  - Baseline 0.67
- I used 3 classifiers for this question, logistic regression, random forest classifier and XGboost classifier
- Utilized more stop words, bigrams and TFIDF

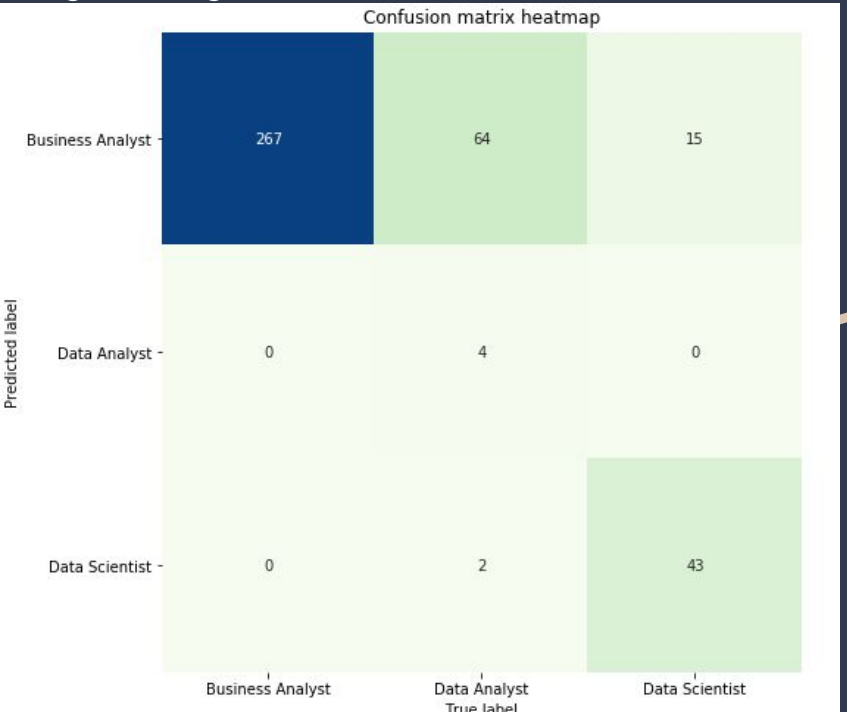
Question 2 - predicting job field

- I had quite a small spread in terms of scores for all three models.
  - Logistic regression F1=0.73
  - Random forest F1=0.85
  - XGboost F1=0.82

- Although logistic regression had the lowest F1 score, as a model it is much easier to interpret than the others

Question 2 - predicting job field

Logistic regression classifier



- Most misclassifications were for the data analysts being mislabeled as business analysts
- No data scientists or business analysts mislabeled as data analysts

Question 2 – predicting job field

```
Class: Business Analyst

Top 5 positive coefs
user stories                6.861131
process mapping             4.529036
functional requirements     2.676896
stakeholder management      1.735461
project management          1.376746
Name: Business Analyst, dtype: float64

Top 5 negative coefs
power bi                   -2.545008
customer insights          -2.904655
science team               -3.148081
predictive models          -3.700114
machine learning          -16.043441
Name: Business Analyst, dtype: float64
```

```
Class: Data Analyst

Top 5 positive coefs
customer insights    3.798015
power bi             3.741001
strong sql           3.586418
experience sql       2.869020
cloud based          2.601069
Name: Data Analyst, dtype: float64

Top 5 negative coefs
functional requirements   -1.010817
stakeholder management    -1.410913
process mapping           -3.150783
machine learning          -3.521603
user stories              -4.661324
Name: Data Analyst, dtype: float64
```

```
Class: Data Scientist

Top 5 positive coefs
machine learning          23.750666
predictive models          7.421287
science team               5.278090
mathematics statistics     4.471347
analytics science          3.671269
Name: Data Scientist, dtype: float64

Top 5 negative coefs
functional specifications    0.0
functional requirements      0.0
experience sql               0.0
experience similar           0.0
high level                   0.0
Name: Data Scientist, dtype: float64
```

Summary

Question 1

I was able to create a model to predict the salary at a reasonable rate when the salaries are simply classed into high, low or medium. The model was consistent for all the classes with consistent precision and recall scores.

Question 2

The skills focus that identify data related jobs from each other include machine learning(data scientists),sql(data analysts) and communication skills(business analysts)

Thanks for listening!

Any questions?