The purpose of this analysis is to predict the salaries from various data-relevant jobs posted on jobs aggregation tool, e.g. indeed. Another purpose is to distinguish 'data scientist' job from other data jobs such as 'data engineer' and 'data scientist'. The analysis was conducted through main steps:

1. **DATA SCRAPING:**

The dataset is constructed through different key words searching from indeed.con.au. Final data table includes jobs from different industries, companies and salary ranges. One main challenge of this analysis is that companies do not usually provide salary range in their job posts. However, the problem is solved by utilising salary estimation tool on indeed - which is claimed to be quite accurate due to their huge database of companies' profiles.

2. **DATA CLEANING:**

Data scraped from the web are usually messy and need lots of data cleaning. Some main problems to deal with are various forms of salary ranges, job titles and job descriptions. NLP technique such as Tfidf were used to extract information from data under text from.

3. **DATA MODELLING**:

- My first selected model is Elastic Net, which is a regularization regression technique used to deal with excessive amount of text features derived from various text fields in the data set. The model returns a small MSE, but also a really small R squared, just around 7% for the test set. Further investigations are needed to fix the model accuracy.

- The second model is a Random Forrest Classifier used to classify 'Data Scientists' jobs from other jobs. Although the model was overfitted and performed terrible on the test set, there are some insights that we can see when extracting features importance.

Recommendations

* Being more selective wih text features to remove noisy features and gather valuable predictors. Use other techniques such as customized stop words and lemmatizing.

* Do more Grid Search with wider range of hyper parameters to look for the optimised model.

* Remove any features that could be related to data scientist, data analyst and data engineers job posts to improve accuracy on classification model.