Project 4 Executive Summary

The purpose of the project was the following was to scrape data through job post website.

1. Determine the industry factors that are most important in predicting the salary amounts for these data.
2. Determine the factors that distinguish job categories and titles from each other. For example, can required skills accurately predict job title?

I scraped the data from mycareersfuture.sg and split the task into the following:

1. Webscraping
2. Cleaning and Dummy Variables
3. NLP
4. Modelling

I set both tasks as binary classification. For the first task, I set a target variable to determine whether the salary was under or over $6500 a month. I used Logistic Regression, Random Forest Classififier, SVM, and Multinomial Naïve Bayes and found that Random Forest had the most accurate results. With results ranging from 0.75 to 0.78. The best features were all related to the position and year’s experience.

Comparing the baseline of approximately 0.5, 0.75-0.78 is a decent improvement.

For the second task, I set the target variable to determine whether the job title included the terms either data science or machine learning. I used Logistic Regression, Random Forest Classififier, SVM, and Multinomial Naïve Bayes and found that Random Forest had the most accurate results again.

The best features were the following:

'data scientist.1', 'machine learning', 'machine learning techniques',

'data science', 'learning techniques', 'advanced analytics',

'actionable insights', 'business problems’

It can be seen that by using NLP, specifically count vectorizer and TDIDF, we found the words used often to determine whether the job title also had data scientist or machine learning inside.

Results as follows:

confusion matrix

[[202 0]

[ 16 4]]

precision score 1.0

recall score 0.2

f1 score 0.33333333333333337

auc score 0.6

With a baseline score approximately 0.9 based on the majority class, we are able to slightly improve the prediction to 0.93 to 0.95.