COMP5310 Project Stage 1 – Sean Preusse, spre7600

Project: Classification Model to Predict Advertised Job Salary

# PROBLEM STATEMENT

There are many companies that use social media and online job boards to advertise vacancies. Applicants can use a wide range of mediums e.g. career sites, social media, print to apply for a vacancy. The preferred tool is career sites with up to 80% of applicants using this to apply to a vacancy *(careerbuildercommunications.com, 2016).*

There are many different career sites in the market e.g. seek.com.au, careerone, indeed.com. In this project, the focus will be on job data from indeed.com as this has been identified as number one job site in the world attracting 180 million unique visitors each month *(indeed.com, 2016).* In addition there is more information to scrape with greater accessibility to complete these functions across many thousands of pages for this site.

Two data science opportunities exist with this data, one specifically to improve on applicant search capabilities for indeed.com and another more generally on market research for any business that would like to gain a competitive edge in the market place. The issue is access to data and missing salaries (74% of vacancies on average do not have a posted salary.

* When searching for a job with salary in mind, many mixed roles appear – some junior, some very senior. This could be optimised to improve on salary search accuracy to improve customer experience and customer return / referral rates.
* In my experience many organisations struggle in finding the right salary when advertising for a role. In addition, the competitive landscape moves fast and there may be future retention issues where salary is not aligned to the employee’s skills for the role and where a competitor is willing to offer more. Having access to this information with a meaningful lens can be very powerful for market intelligence.

Project State 1 – Build of predictive model to predict missing salaries. Such a model should improve search accuracy and also build upon a company’s market intelligence. The focus will be on the Australian market, with the opportunity to incorporate US and UK market data to improve the performance of the model.

Out of Scope – Forecast model on job vacancy salary and expected market demand over time. Once a predictive model has been built to address missing salaries, a future capstone project could look at this type of forecasting with the addition of Industry and Market variables to enhance the forecast.

**Predictive Model Research Questions**

* Is there a relationship of the country, company, job title and location to salary advertised and what elements impact the model the most?
* There are over 65,000 discrete positions, does a grouping of 20 or 50 roles improve model performance
* There are over 30,000 companies listed in the dataset. Does a grouping by industry (up to 10) improve performance?
* Internationally, does the grouping of high performing listed companies under FTSE 250, ASX 200, S&P 500 improve model performance?
* Does the removal of outliers improve performance e.g. long tail has been identified in average salary AUD.

# DATA

Source - Indeed.com (Primary salary prediction data)

* How the data was sourced: I originally built a website scraping tool using a package in R called ‘rvest’. This worked well, however greater productivity could be achieved through import.io, a free web scraping tool. In order to use this tool, a URL generator in R was developed to enable scale by role and country.
* Data Cleaning: The web-scraped data needed to go through a range of different cleaning processes to make it usable.
  + Not all fields in the data extract have been used. Reviews and review URL have data integrity issues that will be difficult to fix in this time period. Employment category is populated for less than 10% of vacancies. Other URL’s have been removed as they do not add value to the model. Once key variables have been identified they are then given more meaningful names.

Page 1 of 2

* + Salary – The salary is a string with salary amounts varying from hourly, daily, fortnightly, monthly and annually. The text was first removed using ‘gsub in R’. The Salary then needed to be converted in an annual basis based on the pay frequency described.
  + Salary\_AUD – Cleaned salary is converted to AUD based on current currency rate at the time of extraction.
  + Published Date – Published Date is a string that is not representative of a date. It will list how long ago the job was posted. To create a date, I first converted seconds, hours and days into minutes from the string and then deducted this from the date extracted to get the published date.
  + State – State has been created from the location string by using a dictionary. Australia has been completed; UK and US will depend on variable importance outside of the Australian market.
  + Role Title – New field to give position title greater meaning. A dictionary of 50 roles has been created to group position titles. SQL was used with ‘like’ on position titles. The order of assignment is prioritised on position length to increase accuracy of assignment
* Fields of interest: **Country** (Australia, United Kingdom, United States), **URL** (Unique ID), Job Title, **Published Date** (Date the vacancy was posted), **Extracted Date** (Date data was extracted), **Company**, **Location**, **Salary** (Annual Salary)**, Salary\_AUD** (Annual salary in Australian currency), **Reviews** (number of company reviews), **State** (Australian States), **Role Title** (Group of many position title into key volume roles).
* Size: On two day periods, 123,741 unique job vacancies have been collected represented by AU (28%), UK (36%) and US (37%). New daily vacancies are around 20,000. For this exercise, 9 other periods will be collected bringing the total number of expected job vacancies to over 300,000 for this analysis.
* Completed but out of scope: Market and Industry variables have been completed from sources such as ABS, RBA, Yahoo and Google. This data will be used for a second project, likely capstone to build a forecast model on this predicted dataset.

**Exploratory analysis**

Detailed analysis contained in the attachment ‘Exploration Analysis’.

* 123,741 vacancies have been scraped from indeed.com (AU, UK and US). There is an even distribution of vacancies by country. 32,287 (26%) or one in four vacancies posted has a salary. This is not evenly distributed by country with the UK having the highest completeness 42% and the US having 16%. Australia is at a reasonable rate of 19%. This should be sufficient for analysis.
* 65,158 distinct job titles have been identified. To make this field usable, 63% of these have been mapped to 1 of 49 roles allowing for greater insight.
* Salary AUD has a positive skew of 1.78. The maximum salary is 450k, model performance may be impacted by outliers. Further investigate log transformation, min/max by role to see if this is an issue.

# PROPOSAL

* Proposal - Build a linear model to predict salary. Ongoing model will need to be refreshed based on seasonality as the job market fluctuates (Independent Google trend analysis 2004-2016).
* Model Build - Random sample of test and train (25% test, 75% train) with the following models: Linear Regression, Ridge Regression, K-Nearest Neighbours, Mean Leaner, Random Forest, SVM regression and Naïve Bayers classification.
* Model Validation Approach - For each of the models I will be running 10 k-fold cross validation and or repeat random sampling of test and train data. I will test model performance in a few different ways. Firstly analysis of confusion matrix (Accuracy, Specificity and Precision), analysis of ROC and Lift Curve and finally mean absolute error.
* Additional features to improve model performance.
  + Reduce number of discrete job titles and companies through a location hierarchy, introduction of job role group and industry group. This should reduce the number of discrete variables significantly.
  + Test transformation of Salary AUD into smaller groupings from dollars into 1k and 10k groups.

# Attachments:

* Script
* Exploratory Analysis

Page 2 of 2