**Setup:** Job board data represents a great opportunity for businesses that want to gain a competitive market edge for recruitment activity and also for job bards that want to increase accuracy of user searches. This paper explores seven different modelling techniques using a few different packages within the R distribution to determine if a model can be built using a range of independent variables like job title, job description, company and location to predict the dependant variable job salary.

The dependant variable average salary AUD has a mean of 80,708 and kurtosis of 5.5, for some high volume roles like account manager this is as high as 24. Anything higher than 3 is leptokurtic therefore I have set the null hypothesis for each model with an absolute mean error greater than $30,000 based my experience with role design and skill composition for particular roles. Additional analysis performed on average min/max by role has a variance of $270k as identified by the high kurtosis therefore business acceptance of anything less than $30,000 is acceptable for the alternative hypothesis.

The data consists of ~370,000 records collected over a month period from the 13 of April to the 17th of May. ~100,000 records have a job salary; a train subset of 70% and validation of 30% has been created to test model performance on these records with an identified salary.

The effectiveness of each model will be measured against absolute mean error. Using metrics like recall and precision do not make much sense in this case as I would like to measure how much variance between train and validation was achieved through each model vs. comparison of specific outcome labels through a confusion matrix.

Each of the independent variables have a range of different levels and will need to be tested for significance within <0.1% for the linear regression models. In addition the model will need to be tested for skewness, kurtosis and heteroscedasticity.

**Approach and Results:** Baseline model build ran into memory issues due to the number of levels contained within job title, company and location at around ~78,000 levels. Additional features were created that reduced this number to around 70 levels based on experience with workforce analytics and segmentation. This was achieved by creating a job role dictionary and using SQL to search for key words within the job title. Companies were filtered down to ‘blue-chip’ identified as listed companies on the ASX200, S&P500 and FTSE100 and company narrowed to a few major cities. These features were further tuned based on overall counts to ensure that each level had at least 200 observations to test.

There are 30 potential independent variables within the model with 9 new features. By applying common sense this is reduced to 10 and a baseline model is created by using the following variables; 1) job\_role, 2) major\_city, 3) title\_junior, 4) title\_manager, 5) title\_senior 6) desc\_junior, 7) desc\_senior, 8) desc\_manager, 9) blue\_chip and 10) source with average salary aud as the dependant variable.

Average Salary in AUD does not pass normal distribution tests; it has a high positive skewness of 1.81 and a kurtosis of 5.5. This variable was log transformed resulting in a skewness of 0 and a kurtosis 3.16 which is a little high but acceptable. Additional features were created through text analysis on job title and job description that searched for key terms; Manager, Junior and Senior, this resulted in an increase from .24 adjusted R2 to .29, a significant improvement. Baseline model assumptions pass both skewness and kurtosis. A test on Heteroscedasticity is acceptable with a value of 1.443 with a p-value of 0.23.

Model 2: Stepwise backward linear regression is performed on the baseline model. The resulted in desc\_junior not passing a level of 0.01 significance. Model 3: Random Forest of baseline formula with 500 trees, Model 4: Support Vector Machine of baseline formula and Model 5: Generalised Boosted Modelling of baseline formula with 1,000 trees.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Linear Regression Baseline | Stepwise Linear Regression | Random Forest | Support Vector Machine | Generalised Boosted Modelling |
| RMSE | 71759.8 | 71758.0 | 38427.0 | 38979.7 | 42801.9 |
| MAE | 55255.3 | 55254.5 | 24718.7 | 24864.6 | 28746.0 |
| Hypothesis | *H0* | *H0* | *H1* | *H1* | *H1* |

There is a significant difference for the random forest, support vector machine and generalised boosted modelling with a mean absolute error (MAE) less than 30,000 which pass the alternative hypothesis. The linear regression models did not pass the alternative hypothesis therefore null. Computational power was an overall limitation as my machine had 16gb of ram, I would revisit model design to include job title and company if I had more ram.

Overall the random forest model had achieved the best result with a MAE of 24,719 significantly under the 30,000 threshold. In terms of production the Random forest took 30min to process for 100,000 records where the generalised boosted modelling model took less than 1min to process. When dealing with limitations to computer memory for more than one month, the X model may be preferred even with a higher (MAE).

For the models that passed the alternative hypothesis my recommendation would be to place these into production as they solve the problem in identifying missing salaries providing that I looked at the following actions post production;

* This model should be reviewed in the next few months to account for market seasonality as a snapshot of one month has been taken. This will need to be tested in addition to individual job role changes to market supply e.g. Mining shift and lower demand for geologists and specialised engineers. With supply/demand this may decrease average salaries increasing overall (MAE).
* Ideally ongoing model review should be completed quarterly to test overall assumptions and baseline improvement to (MAE). Key will be to ensure role mapping to position title is updated to pick up new positions not contained within the dictionary as market definitions evolve.
* Feature text analysis has been tested lightly i.e. key word searches on job title and descriptions are only based on three key words to help determine role complexity. Deeper text analysis could be completed to better understand role complexity and job design. Secondly, company linkage to high performing listed companies is very narrow, this needs to be broadened to cover key industry’s not present in the data collection. This data could be supplemented with searches from seek.com where industry and job subcategory is present.

Overall I have learnt much about regression techniques and I feel that I have only scratched the surface. I would like to test additional models such as ridge and lasso regression and neural net to how this compares to existing models. In addition I would like to spend more time to implement these models in Python as packages like ‘scikit-learn’ should be more efficient and this will better round out my data modelling skillset to provide an improved contingency dependant on problem and business requirement.