Analyzing Data Science Salaries

By: Neil Liberman

Statement of Goal

- Understand what keywords most directly impact data science salaries.
- Build a model that predicts whether a job will fall above or below the median salary.
- Limit the number of clients that are incorrectly told to expect a high salary.

Methodology: Using Estimated Salaries

- Not concerned with perfectly accurate salaries because of the large spread of salary and the goal to predict above or below the median.
- More concerned with having a large dataset to build a larger corpus for Natural Language Processing.

What Features and Models Were Used

Features:

Count vectorizer (CVEC), term frequency inverse document frequency (TFIDF), job location (Bos, Chi, DC, Hou, NY, SF), the title of the position, and the company hiring.

Models:

Logistic Regression, Random Forest, Gradient Boosting

What are these models?

Logistic Regression - a classification model that estimates a probability that an observation falls into a given class.

Random Forest Classifier - an ensemble classification model that utilizes random selection of data and random selection of variables.

Gradient Boosting - converts a sequence of weak models into a more complex model .

What features are most important

- City: Houston had the lowest median salary at $95,500, with Chicago second lowest at $100,800. DC and Boston were in the middle with salaries of $104,400 and $105,300 respectively while the top two income locations were San Francisco at $111,800 and New York at $113,200.
- Job Title
- Company

Which Keywords Are Most Predictive According to CVEC?

- ‘Data’
- ’Scientist’,
- ‘Lead’,
- ‘Machine’,
- ‘Senior’,
- ‘Interpret’
- ‘Analytics’
- ‘Research’,
- ‘Big’,
- ‘Models’
- ‘Risk’.

Which Keywords Are Most Predictive According to TFIDF?

- ’Scientist’,
- ‘Data’,
- ‘Machine’,
- ‘Databases’,
- ‘Research’,
- ‘Analyst’,
- ‘Lead’,
- ‘Reports’
- ‘Engineer’
- ‘Senior’
- ‘Big’,
- ‘Risk’

Evaluating the Models

Models implementing CVEC performed slightly better than models using TFIDF

The model’s which were most accurate in predicting classes correctly were Logistic Regression and Gradient Boosting which predicted correctly about 70% of the time.

However, performing random forest on a specific list of words (based on SME) returned the lowest rate of false positives. As previously stated, we would like to avoid telling a client to expect a high salary job and end up with a low salary job. Therefore, limiting false positives is of high importance.

Hand Chosen Words Based on SME

‘Junior',

'Senior',

'Manager',

'Masters',

'PHD',

'Entry',

'Scientist',

'Machine',

'Research'

Conclusion

There’s a difference in the results based on which models and features are used.

Important to understand what it is you are most concerned in measuring.

Being that we are most concerned with false positives, using the hand selected words will be our best option