


Part 1: Job Role

Find a job advertisement for a data science position that covers the type of data science role you see yourself doing.

a) **Include a copy of the job description.**



Bluefin Resources
Data Analytics Recruitment Solutions

Data Scientist

- 12 month Fixed Term Contract with Strong Longer Term Potential
- Financial Services
- Machine Learning, AI, NLP, Predictive Modelling and Analytics

Data Scientist - Machine Learning

12 month Fixed Term Contract

The Company

My client is a collaborative, people focused financial services company; this role has been newly created due to investment and expansion of the data science team.

The Role

Reporting in to a passionate, innovative and highly capable Head of Analytics you will have fantastic opportunities to grow and develop. You will be working across a wide range of projects across the financial services platform to develop solutions to make informed decisions through the application of data mining, predictive analytics and machine learning techniques.

Key Responsibilities

- Apply Machine Learning techniques to create automated, personalised and next best action solutions
- AI, text mining and NLP
- Deliver models that can be put into production on a large scale
- Build a deep understanding of the relevant data & systems landscape & optimise the data extraction processes
- A clear ability to generate insights from data, engage and communicate those insights to non-technical business stakeholders

About You

- An innovative , out of the box thinker who can bring new concepts and ideas to life through data
- **A minimum of 5 years'** experience in statistical modelling in a big data environment (R/Python)
- Experience in developing predictive models in a commercial environment; **NLP experience would be highly advantageous**
- Knowledge of a variety of statistical and data mining techniques (regression, statistical tests and proper usage, random forest, clustering, decision tree learning, deep learning text mining etc.) and their real-world advantages/drawbacks
- Consultative and commercially focused; a proven understanding and ability to articulate and deliver solutions that can demonstrate business value and ROI
- Tertiary qualified in a quantitative discipline
- Strong R programming essential with SQL skills

This is a collaborative innovative company and team with huge investment in systems, technology and people. You will get to work within a highly consultative and innovative business with a team of exceptionally talented analysts and data scientists.

Please apply on-line or for more information please contact Marie Thow at Bluefin Resources on [02 9270 2640](tel:0292702640).

IMPORTANT: By submitting your email address and any other personal information when you APPLY to a job, you consent to such information being collected, held, used and disclosed in accordance with our COLLECTIONS NOTICE and PRIVACY POLICY.

<http://www.bluefinresources.com.au/privacy-policy>

www.bluefinresources.com.au

b) Which domain/field is the job position related to?

This job position of Data Scientist in Bluefin Resources falls under Employment and Recruitment domain.

c) Is the role looking for insights into the data or making predictions? Justify your answer.

The role is looking for the insights into the data using data mining techniques. The role requires building classification models bring into production and presenting the insights to non-technical business stakeholders. The role description includes having knowledge

of data mining techniques like clustering and classification, which makes it clear that the role is looking for the insights or trends/patterns in the data.

Part 2: Data Set

Find a public data set that could be construed as relevant in some way to the particular job ad you identified (for example, a data set that can be used for training and testing a binary classifier).

a) Include a URL and a brief description of the data set.

URL: <https://www.kaggle.com/HRAnalyticRepository/job-classification-dataset>

Description

This is a dataset containing some fictional job class specs information. Typically job class specs have information which characterize the job class- its features, and a label- in this case a pay grade - something to predict that the features are related to.

The data is a static snapshot. The contents are

ID column - a sequential number

Job Family ID

Job Family Description

Job Class ID

Job Class Description

PayGrade- numeric

Education Level

Experience

Organizational Impact

Problem Solving

Supervision

Contact Level

Financial Budget

PG- Alpha label for PayGrade

b) Very briefly describe why you chose it (e.g. how it relates to the data science position).

The Bluefin Resources Company is Recruitment Solution Company which matches the employers with the Hirers. So the company would have a database which would include all the information related to recruitment companies such as their Job family which means different job roles, company description, ratings and reviews, company experience, pay grade that is an amount an employer may receive, etc. The company database might also include details of different Employers/job seekers like their name, mobile no, address, education background, CV upload (link), certifications, previous company experience if applicable, etc. This dataset includes job class specs like ID, Job Family, Description, Pay Grade, Experience, Financial impact, etc which closely relates

to the database Bluefin Resources Company might have. However, this dataset is only related to recruitment companies and rather not about the Employers/job seekers. So, this is the reason working on this dataset would better as it related with job position of Bluefin Resources company.

Part 3: Experiment

Use **two machine learning algorithms** (e.g., including decision trees, kNN, SVM, or neural networks) to gain insight into the data. One of the machine learning models has to be different from the ones used in the Practical Data Science course.

a) What insights do the machine learning algorithms give you?

Decision Tree and Naïve Bayes classifier algorithms were used to gain the insights into the data. The dataset had multi-label target variable as Pay Grade Alpha with ordinal values from 0 to 9. The dataset is split into 75-25% so that distribution of training and testing on unseen data is balanced.

Decision Tree algorithm is widely used for interpretation and rather not prediction. The algorithm can be use to understand the behavior of recruitment companies and job seekers and to build a decision framework in engaging the existing companies or attracting new companies.

Naïve Bayes algorithm works considering all the features as independent i.e. assuming no correlation between features. The algorithm uses existing knowledge to calculate the probability of any event. It can be used to calculate the overall/average ratings by job applicants or existing employers for companies.

The Decision Tree had an accuracy score of 94% whereas the Naïve Bayes had an accuracy score of 88%. This clearly means that the Decision Tree has a better predict score as compare to the Naïve Bayes algorithm. This could also mean that some features are correlated with each other.

b) Compare the predictive power of the models produced by the two algorithms and visualise their effectiveness. Clearly justify the evaluation metrics you used to compare the effectiveness of your the models.

	precision	recall	f1-score	support
0	0.00	0.00	0.00	0
1	0.00	0.00	0.00	1
2	1.00	1.00	1.00	3
3	1.00	1.00	1.00	2
4	1.00	1.00	1.00	1
6	1.00	1.00	1.00	1
7	1.00	1.00	1.00	3
8	1.00	1.00	1.00	3
9	1.00	1.00	1.00	3
accuracy			0.94	17
macro avg	0.78	0.78	0.78	17
weighted avg	0.94	0.94	0.94	17

The fig shows the classification report for the Decision Tree classifier. Classification report is one of the evaluation metrics used to measure the quality of predictions from a classification algorithm. The metrics are calculated by using true and false positives, true and false negatives.

The first 2 labels weren't correctly identified whereas the 2-9 labels were predicted correctly. Support is the no of occurrences of given labels where the results show that the first label has 0 occurrences. The recall is the no of classes found in the whole number of classes. In this case it seems accurate for 2-9 labels and 0 for first 2 labels. The F1-score is the mean between precision and recall. The greater the F1 Score, the better is the performance of the model. The decision tree shows a perfectly balanced F1 score.

	precision	recall	f1-score	support
1	0.00	0.00	0.00	1
2	0.75	1.00	0.86	3
3	1.00	1.00	1.00	2
4	1.00	1.00	1.00	1
5	0.00	0.00	0.00	0
6	0.00	0.00	0.00	1
7	1.00	1.00	1.00	3
8	1.00	1.00	1.00	3
9	1.00	1.00	1.00	3
accuracy			0.88	17
macro avg	0.64	0.67	0.65	17
weighted avg	0.84	0.88	0.86	17

The fig shows the classification report for the Naïve Bayes classifier. The 1, 5 and 6 labels were identified incorrectly with a precision of 0. The score for the 5th label is 0 regarding the correct prediction or found occurrences. The recall score for 1, 5 and 6 was 0 i.e. these classes weren't found in the whole number of classes. The F1 score of Naïve Bayes classifier is good which means the model performs better.

Overall, it can be said that the Decision Tree performs better as compared to Naïve Bayes. However, it is possible that some classes may have been missed with Decision Tree classifier as this is a multi-label classification resulting in nearly perfect accuracy.

References

- [1]"Job Classification Dataset", *Kaggle.com*, 2020. [Online]. Available: <https://www.kaggle.com/HRAnalyticRepository/job-classification-dataset>. [Accessed: 03-Aug- 2020].
- [2]"Bluefin Resources, a specialist in recruitment", *Bluefinresources.com.au*, 2020. [Online]. Available: <https://www.bluefinresources.com.au/>. [Accessed: 03- Aug- 2020].