From Data Analyst to Data Scientist?

Evidence from real job postings

using text mining and NLP techniques

Author: Eric (Chengcheng) Jiang
LinkedIn: https://www.linkedin.com/in/cc-eric-jiang/

Background

Fact:

- Different companies/industries have different definitions about data-related jobs.
- A variety of information can be found around this topic, but many are misleading.

Dilemma:

- Many job-seekers generally assume data scientist and data analyst the same, and blindly apply to both positions.
- Many job-seekers do not know how to match their skills with a specific data-related position.

Failure from the employer side:

- Increase in recruiting cost due to employees' low job satisfaction rate and low retention rate.
- Potential profit decrease due to rising cost.

Failure from the employee side:

- Fail to secure a position that suits them best.
- Fail to transitition from data analyst to data scientist successfully.

Techniques Used

- Web scraping
- Text mining (wordcloud visualization, top frequency words extraction, etc.)
- Clustering (TF-IDF, PCA, k-means Clustering)
- Classification (Logistic Regression, Support Vector Machine, Naïve Bayes Classifier)
- More advanced NLP algorithms (Neural Networks) – still in progress

Implementation

- Data Source: Job postings from Indeed

✓ positions: Data Scientist, Data Analyst (fuzzy matching)

✓ locations: Toronto, Montreal, Vancouver, Calgary (within 100km of each city)

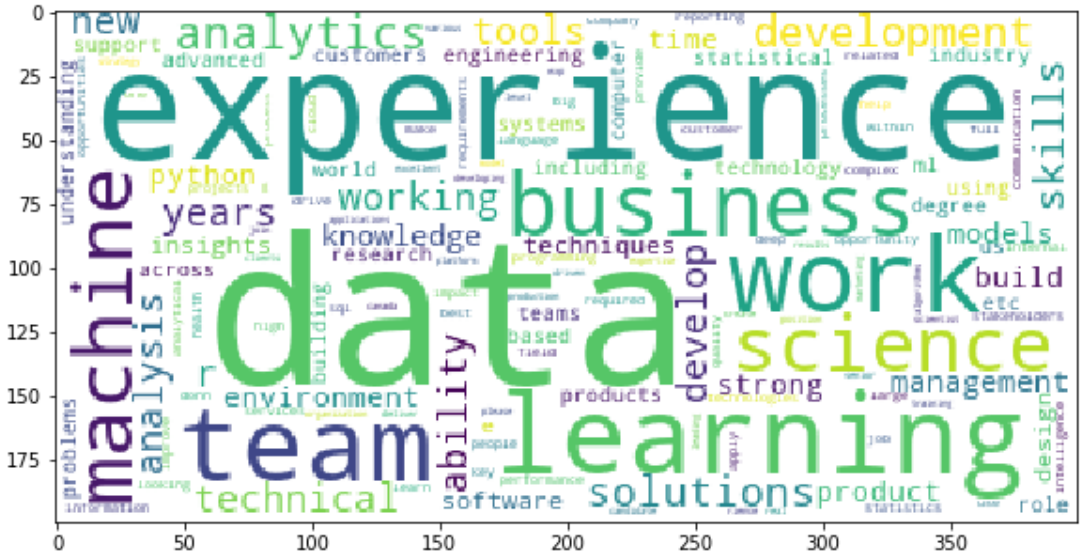
✓ additional: full-time, sort by date

✓ dataset: 784 observations (Data Scientist:222; Data Analyst:562)

- Sample link:

✓ https://ca.indeed.com/jobs?q=data+scientist&l=Toronto,+ON&radius=100&jt=fulltime&sort=date&start=10

From visualization, data scientist deals more with machine learning using Python, and data analyst lies more on the business and management side.

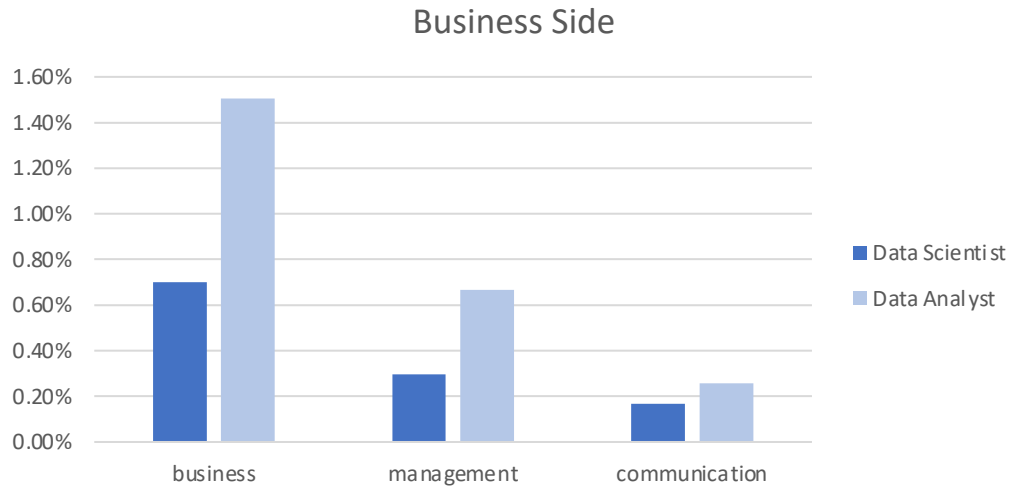


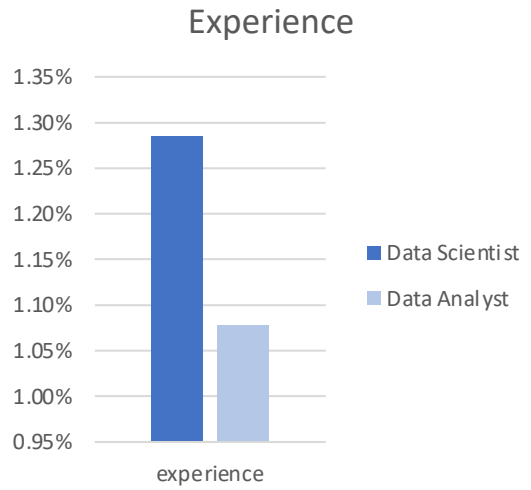
Word Cloud of ‘Data Scientist’

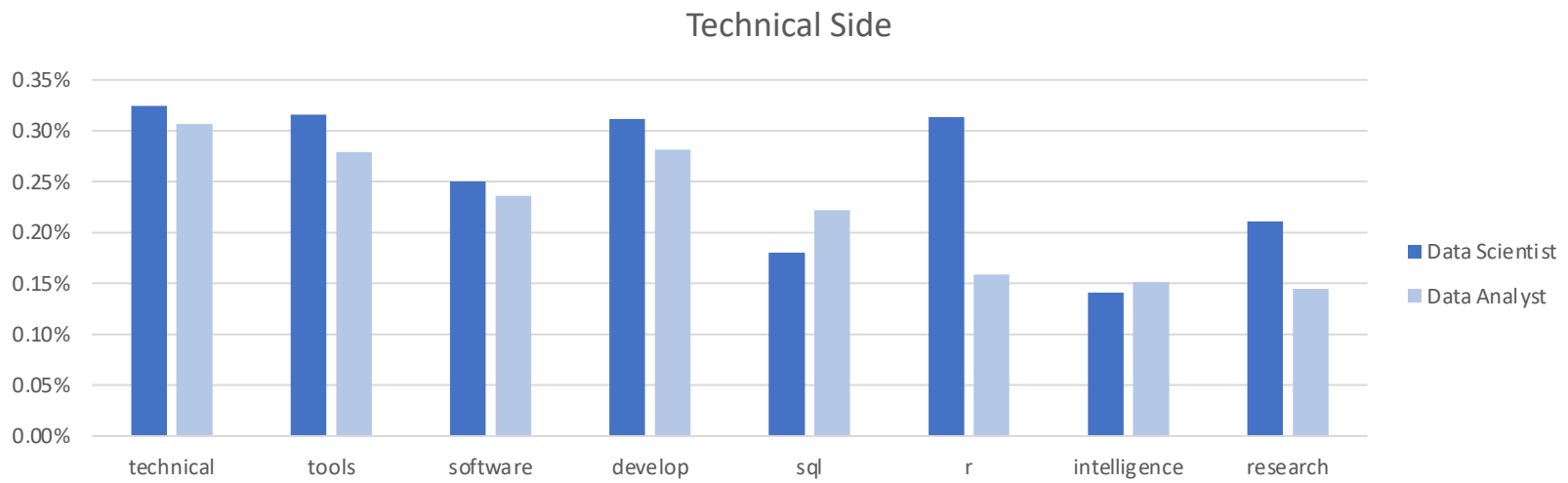
Word Cloud of ‘Data Analyst’

Note: Pictures are generated by Python scripts using ’WordCloud’ package.

Both positions require certain degree of experience, business and technical knowledge, while different focuses exist.



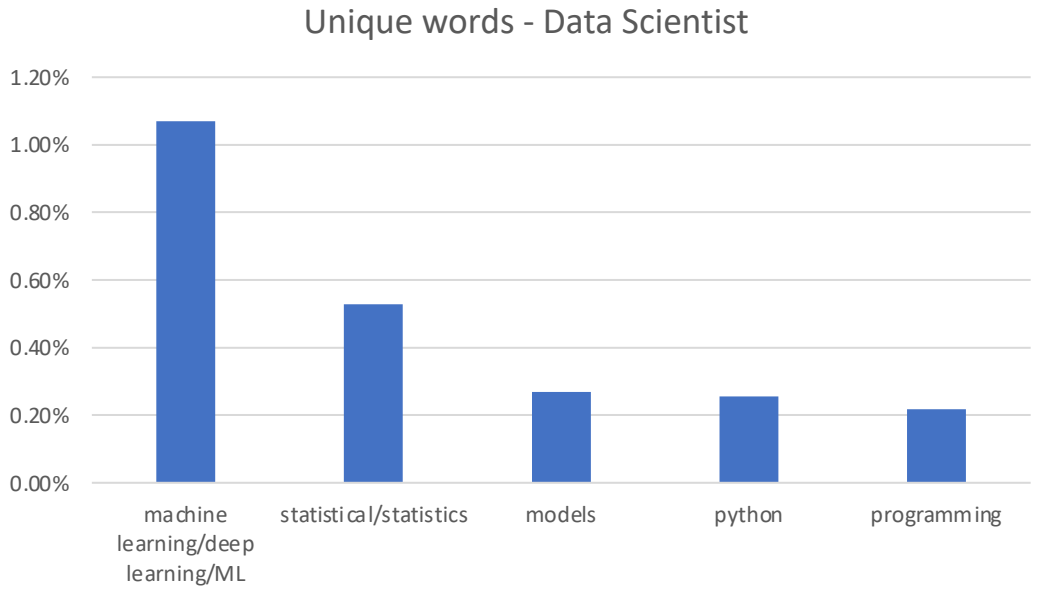


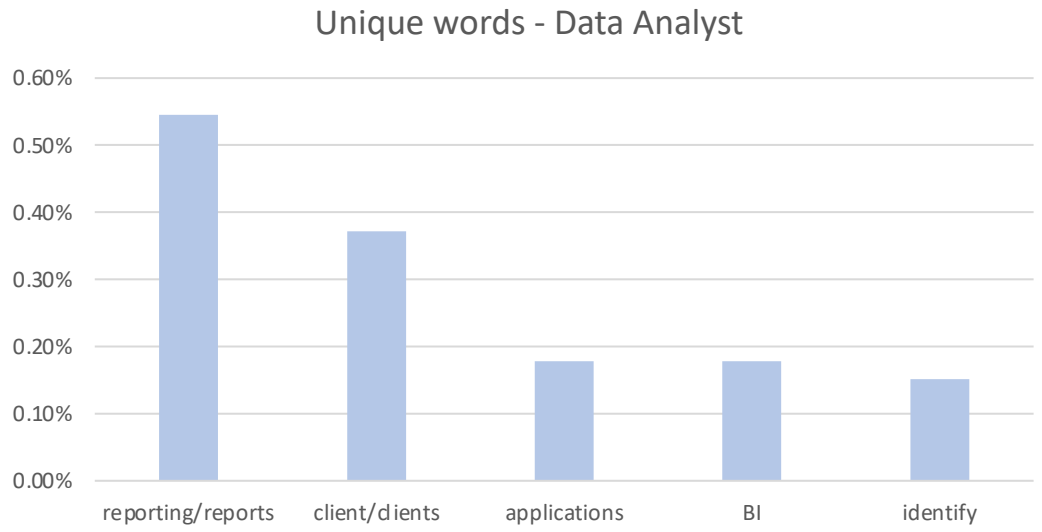


- Data Analyst is more business related.
- Both roles require some technical knowledge – R and research are more relevant with Data Scientist while SQL and intelligence with Data Analyst.
- Overall, Data Scientist usually requires more related experience as it involves more technical things.

Note: The value on y-axis is the ratio of frenquency of a specific word to total number of words in the corpus (after stopword removal).

Building ML models and conducting statistical analysis are critical parts in Data Scientist role; Data reporting and facing clients are major things for Data Analyst.





- Data Scientist roles focus on machine learning/deep learning algorithms, as well as statistical analysis.
- Mostly use programming languages (eg. Python) to deal with data modeling.

- Data Analyst roles are more client-based, in which data reporting is of great significance.
- Mostly use data analytics techniques (eg. Business Intelligence tools) to identify problems and improve applications.

Note: The value on y-axis is the ratio of frenquency of a specific word to total number of words in the corpus (after stopword removal).

Analytics & Insights, Software Development, Product Development are some typical workstreams for Data Scientist.

Cluster representative words	Rank	Summary
time,analysis,world,new,solutions,build,skills,analytics,learning,science,team,work,business,experience,data	1	General key words of Data Scientist roles
work,including,management,learning,science,statistical,skills,support,technical,insights,analytics,analysis,experience,business,data	2	Analytics & Insights direction – analytics, insights, business, management, statistical analysis, etc.
engineering,systems,computer,development,develop,models,years,deep,work,team,software,data,experience,machine,learning	3	Software/Systems development direction – software/systems, development, ML/deep learning, years experience, etc.
development,deep,best,software,techniques,research,products,models,work,team,experience,machine,data,learning,ml	4	Product development (R&D) direction – products, research, ML/deep learning, development, etc.
quality,analysis,computer,science,team,statistics,job,world,work,data,health,research,learning,deep,statistical	5	Roles in healthcare-related industries – health, research, deep learning, statistics, etc.

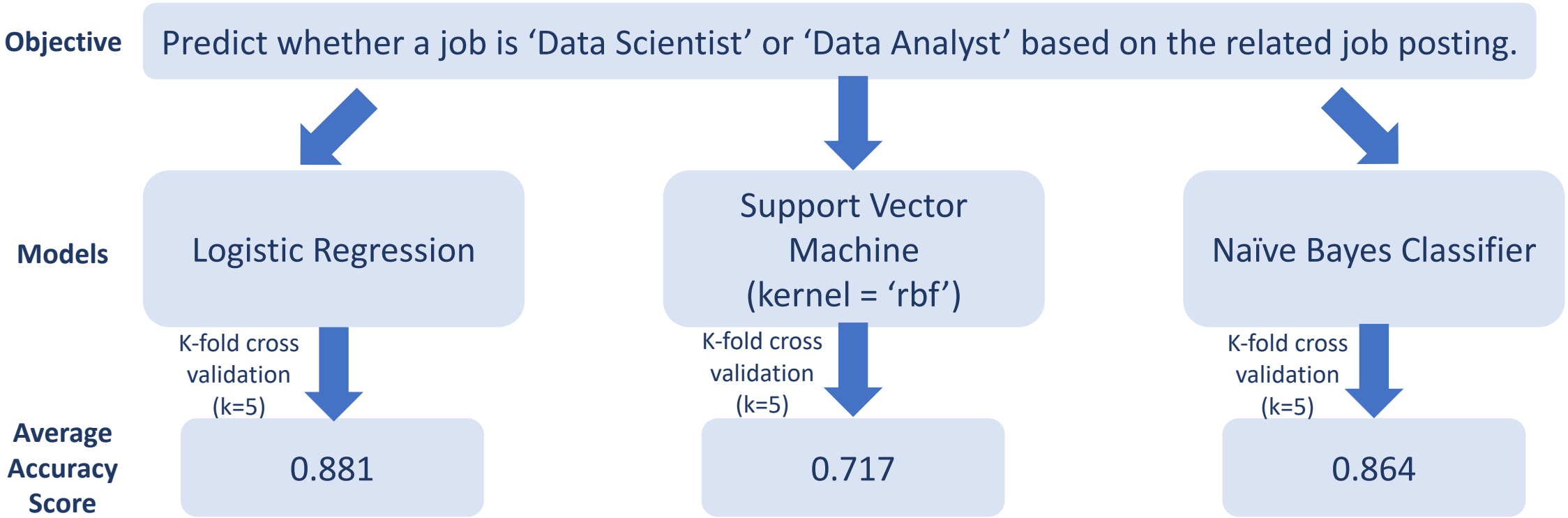
Note: Top 5 clusters (each accounts for over 10%) in Data Scientist roles using k-means clustering.

Customer/User Analytics, Project Management, Process Management are some typical workstreams for Data Analyst.

Cluster representative words	Rank	Summary
process,support,reporting,customer,strong,team,financial,analysis,work,ability,experience,data,skills,management,business	1	Customer analytics in financial industry direction – customer, financial, analysis, reporting, etc.
ability,user,analysis,support,skills,solutions,processes,work,system,management,requirements,experience,process,project,business	2	User analysis & Project/Process Management direction – user, analysis, project, process, solutions, etc.
knowledge,support,ability,sql,tools,analytics,team,management,skills,quality,work,analysis,business,experience,data	3	General key words of Data Analyst roles
analysis,design,work,management,development,support,data,team,project,system,technical,experience,requirements,systems,business	4	Project/System management direction – design, project, system, technical, development, etc.
time,team,applications,experience,clients,business,skills,development,work,job,services,data,financial,health,research	5	Roles in healthcare/financial services direction – health, research, financial, services, clients, applications, etc.

Note: Top 5 clusters (each accounts for over 10%) in Data Analyst roles using k-means clustering.

Logistic Regression, Naïve Bayes Classifier are two suitable models for this text classification problem, achieving an average accuracy of over 85%.



Notes:

i) Accuracy is the main performance metric in this case, as the main focus of the problem is to classify a job posting to 'Data Analyst' or 'Data Scientist' category correctly.

ii) Due to the scale of the dataset (relatively small), k-fold cross validation is used to enhance the robustness of the model performance.

iii) Three representative classification algorithms are selected here. Deep learning based algorithms (Neural Networks, etc.) will be discussed in the future.

The text mining and NLP process in this case can be applied to many industries, and can help individuals and businesses in different ways.

Job search in data field	Help job seekers get an overview of the data-related jobs, and evaluate what they are suitable for based on their skillset.
E-learning	The whole process can be used in online education platforms, helping learners understand what is the similarity and difference between two related concepts.
Customer Analytics	Match customers' comments with their NPS (Net Promoter Score), and study what are the successful aspects of products/serivces and what are unsuccessful.
Professional Services	Help consultants have a better understanding of various documents (financial, law, etc.) from a data-driven approach, in order to better serve their clients.

Stay tuned for more … (Neural Networks, etc.)

- Thanks!