PURDUE UNIVERSITY | Krannert School of Management

DESCRIPTIVE & PREDICTIVE ANALYSIS OF DATA SCIENCE JOBS

Adarsh Reddy, Clay Marshall, Alex May, Julien Pham, Makram Assaf, Matthew A. Lanham
Purdue University, Krannert School of Management
reddy33@purdue.edu; marsh116@purdue.edu; may93@purdue.edu; assafm@purdue.edu; pham40@purdue.edu; lanhamm@purdue.edu

ABSTRACT

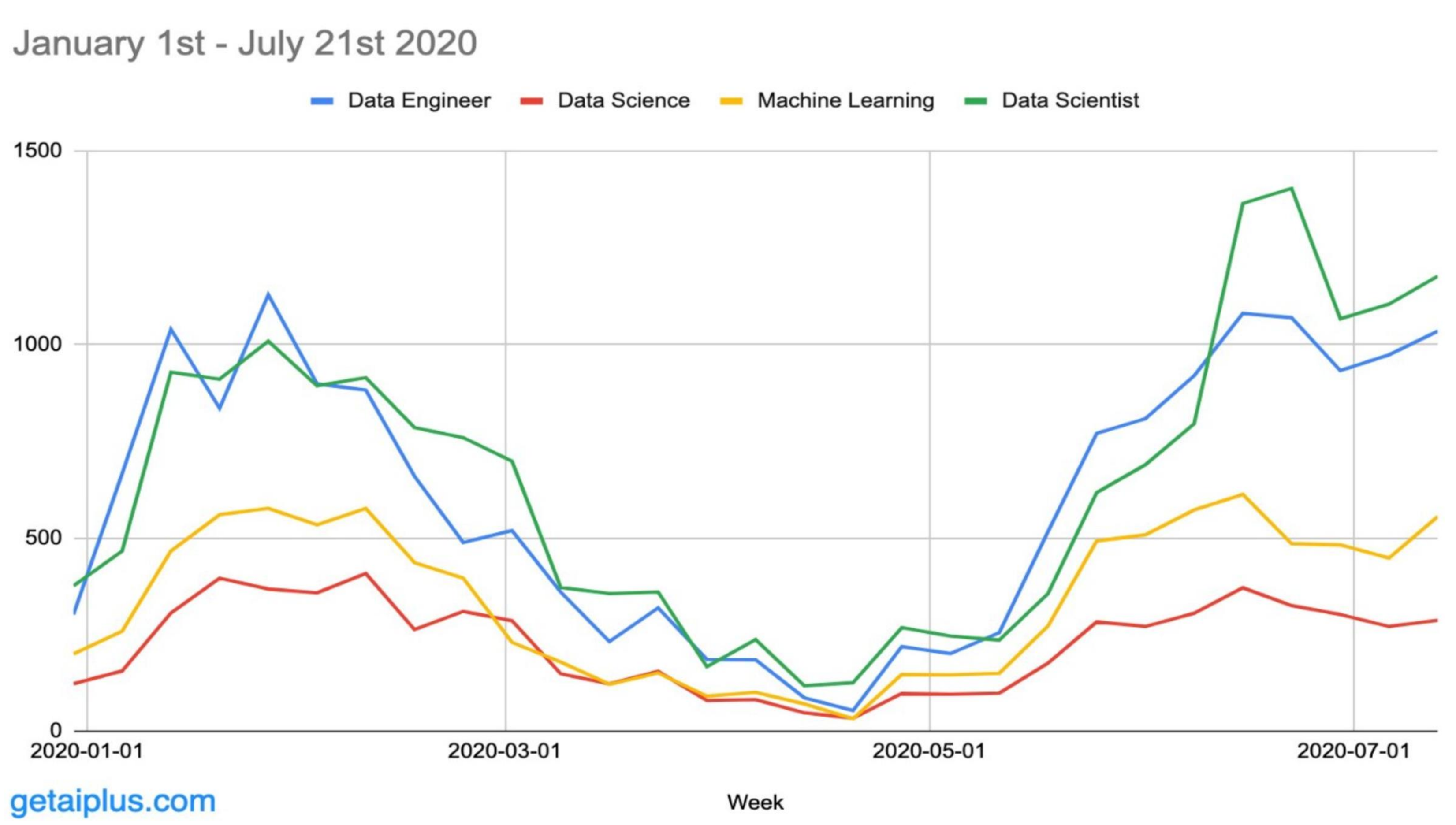
This project seeks to aid job-seekers to streamline their search for open positions in the post COVID-19 data science industry. This will be accomplished via statistical models that utilize text analysis to group roles into one of six categories and identify any key skills that were mentioned in online job postings.

INTRODUCTION

There are many applicants for the data science industry with limited openings due to the impact of COVID-19. Therefore, it is crucial to understand the distribution of roles in order to assist applicants in finding appropriate positions for their skillset. The goal of this study is to utilize real-world data and incorporate predictive analysis to answer the following **research questions:**

1. How can we optimally categorize online data science job postings?
2. What are the key skills employers seek and how do they differ between different roles within industry?

Data Science Industry Job Posting Trends (2020 Q1 thru Q2)



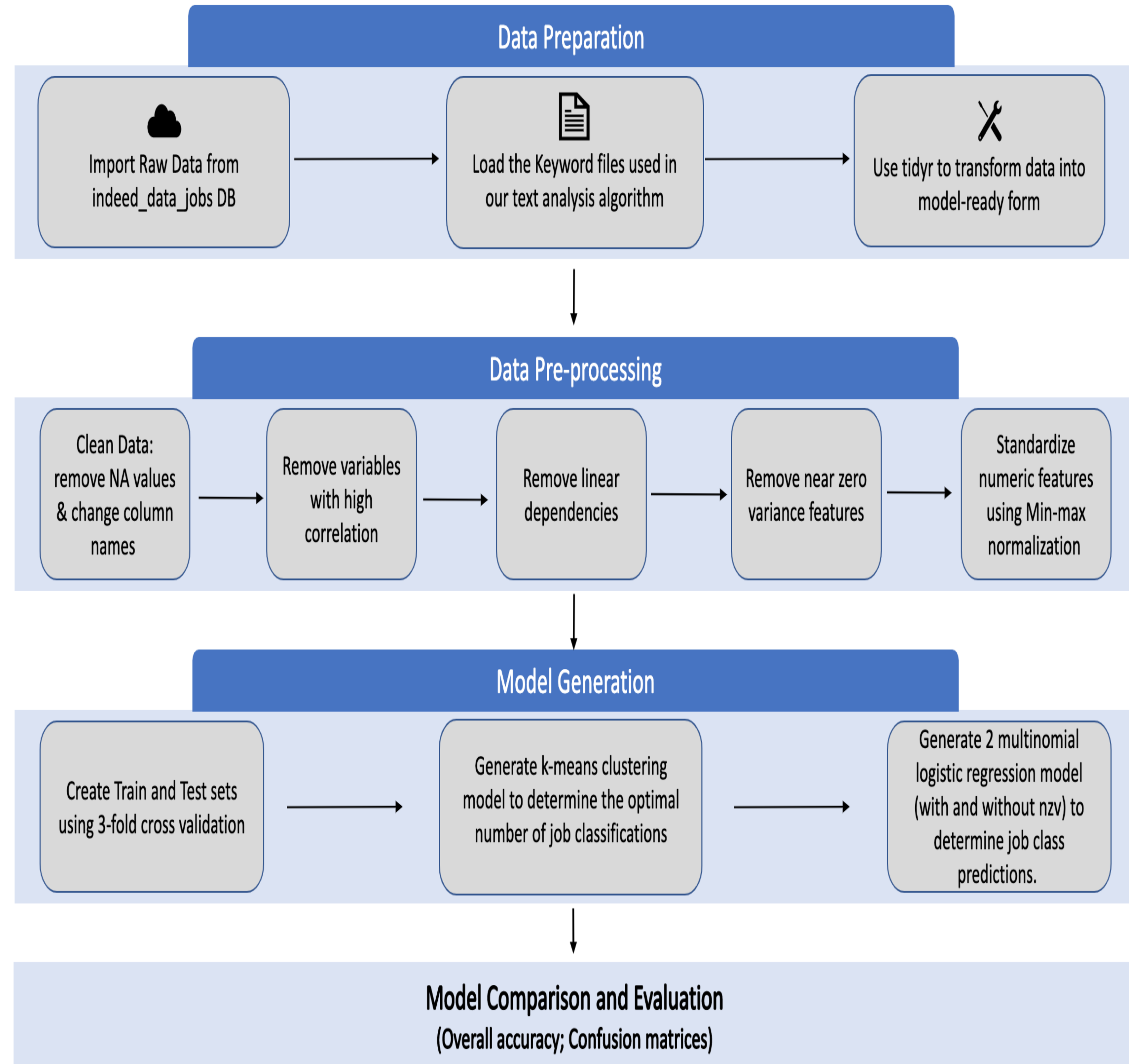
Source: https://medium.com/@ODSC/looking-for-data-science-jobs-in-the-pandemic-good-news-and-not-so-goodnews-1add9367c861

LITERATURE REVIEW

Most studies we found analyzed predicted key skills or job titles through K-means Clustering and Multinomial Logistic Regression. A key differentiating factor in our analysis is the focus put on the impact of Covid-19 on the job market.

Study	Algorithm(s) Used
Mihet et al. (2019)	Times-series analysis, Logistic Regression
Fayyad et al. (2020)	Logistic Regression, Naive Bayes
Paul et al. (2020)	K-Means Clustering
Nguyen et al. (2019)	K-Means Clustering
Radovilsky et al. (2018)	SVD Plot, Term-based Frequency tables

METHODOLOGY



STATISTICAL RESULTS

Based on the nature of our research problem, we decided that overall accuracy was the best measure on which to judge the efficacy of our models. This is because the goal of the models generated was to accurately predict a particular job class rather than see the effects of the various input variables on our target variable.

Statistical Output Table	Train Dataset	Test Dataset	Difference
MODEL 1 Pre-processed dataset ***including*** near zero variance (nzv) variables: 308 input variables	Overall Accuracy: 0.8891	Overall Accuracy: 0.7701	0.119
MODEL 2 Pre-processed dataset ***excluding*** near zero variance (nzv) variables: 62 input variables	Overall Accuracy: 0.7021	Overall Accuracy: 0.6398	0.0623

Key take-aways:

- Model 1 is more accurate overall but slightly overfit. Considering the trade-offs, we will use Model 1 to classify job descriptions.
- Accuracy was our statistical performance measure used as the # of job postings in each group were fairly balanced. Confusion matrices (not pictured) are another way to compare these models as they allow for a more detailed interpretation of classification accuracy within individual job types.

JOB INSIGHTS

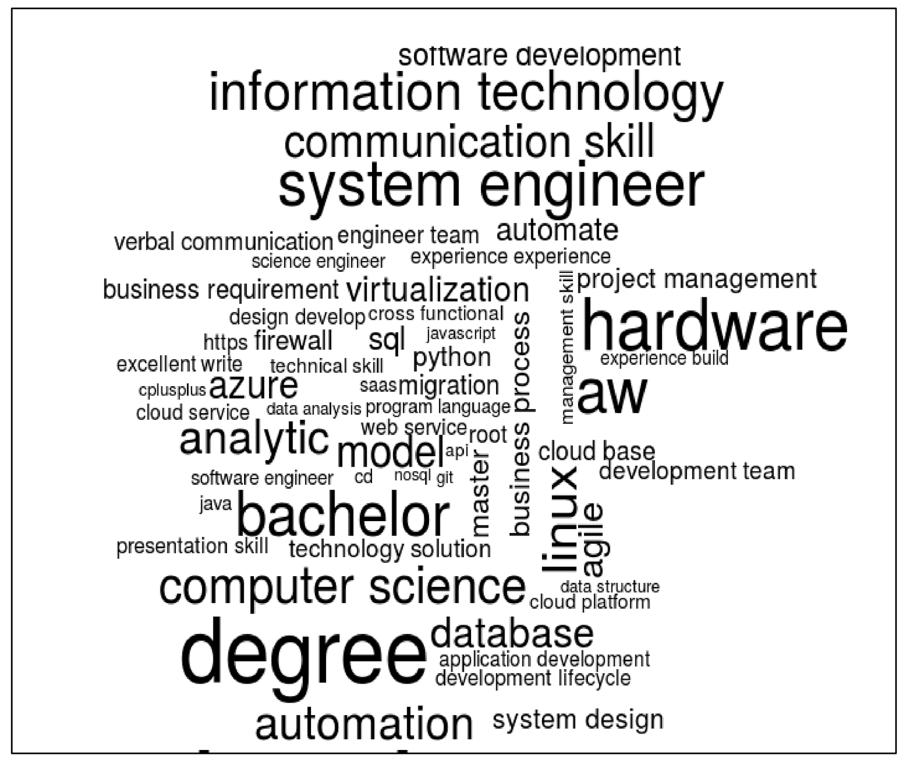
Our study will provide key insights that will aid how data science candidates search for open positions. These insights will be realized via a R-Shiny app that has been built with our models as the underlying framework. Specifically, based on a user-provided job description, the output will display a job type and word cloud detailing skills relevant to that role. The figures below show outputs for two different user input scenarios (user input not shown).

Example 1

Outputted Job Class

This Job Description looks more like an opening for a..... Data.Analyst !

Outputted Word Cloud

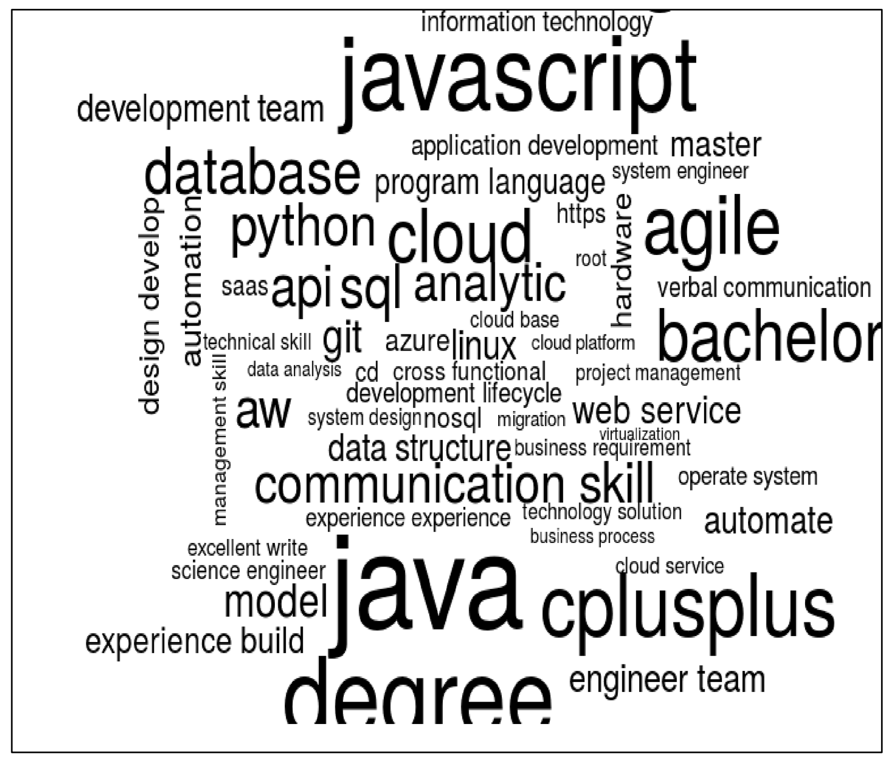


Example 2

Outputted Job Class

This Job Description looks more like an opening for a..... Software.Developer !

Outputted Word Cloud



CONCLUSIONS

I. The use of k-means clustering in our analysis allowed us to identify six categories of data science jobs that we feel most postings will fall into. These being, "Data" or "Business" analysts and "Cloud", "Network", "Software" or "System" engineers. Furthermore, our use of multinomial logistic regression techniques allowed us to measure how accurately (~77%) a role might fall into one of these classes.

II. Our analysis also provided us with the means to visualize the top skills associated with each role. Across all classes, we have identified that **programming** and **communication skills** are the most commonly sought-after skills while specialized expertise, such as **networking** and **visualization skills**, are specific to certain types of roles.

ACKNOWLEDGEMENTS

We would like to thank Professor Matthew Lanham, Xinyu Wang, Theo Ginting, and our graduate student mentors, Ali Faiz and Cavin Vivekh, for their guidance and support on this project.