https://github.com/youming-qiu/BA888-Team3

BA888 Capstone Project Summary Report

Boston University
Questrom School of Business
MS in Business Analytics Program
Cohort B Team 3

Yue Gong
Jingcheng Huang
Youming Qiu
Yishuang Song
Minna Tang

Problem Statement & Setting

In today’s data-driven world, more and more people choose to join this aspiring industry and focus their career paths in the data science related field. As MSBA students, we are keen to explore the data science job market in the United States. The goal of this capstone project is to provide clear suggestions on what kind of technical skills and other requirements employers are looking for when hiring for data-related positions. Meanwhile, we collect data on future career choices, the range of salaries, and possible promotions that we can expect after graduation.

Skills & Other Requirements
Data Science Job Market
Salaries & Expectations

Figure 1: Project Goal

The steps are straightforward. Firstly, we need to acquire related information that includes specific skills and requirements for certain positions. Then, we will combine specific requirements related to various data-related jobs, and predict skills requirements along with salaries one can get in the current data science job market.

Why is it important?

We have listed the reasons below:

1. Data scientists, data engineers, and data analysts are among the most sought-after positions in the current American job market. Although the titles suggest that these positions require different skills, many have a hard time distinguishing the differences among them. As a result, job applicants find it difficult to correctly position themselves or stand out in a competitive setting.
2. Many existing and emerging workers yet don't have the full skillsets employers need. We try to explore employers’ needs to give the best skills combination suggestion for each position. Given job candidates’ limited resources, we would advise them regarding what kind of skills they should focus on.
3. Salary negotiation is an important skill in life. We want to create a salary prediction model that helps candidates address this issue with confidence. With certain skills, what amount of salary and bonus should one ask for? How does one know he/she is getting paid fairly? Our model is designed to predict data-related job salaries and let people have a general idea about wages, thus making better decisions on remaining or finding new jobs.

Data Sources

1. Data Scientist Job Market in the U.S.

https://www.kaggle.com/sl6149/data-scientist-job-market-in-the-us#alldata.csv

2. Indeed dataset - (Data Scientist/Analyst/Engineer)

https://www.kaggle.com/elroyggj/indeed-dataset-data-scientistanalystengineer

Dataset Summary

“Data Scientist Job Market in the U.S.” dataset is from the *Kaggle.com* website. In this dataset, we have in total of 6,964 rows and 5 columns. Columns include information such as position, company, description, reviews, and location.

“Indeed Dataset - (Data Scientist/Analyst/Engineer)” dataset is from the *Kaggle.com* website. This dataset is focusing specifically on the top three most popular positions in the data science job market which are Data Scientist, Analyst, and Engineer. In the original dataset, we have in total of 5,715 rows and 43 columns. Columns include information such as job title, salary, skill, description, location of the company and etc.

Exploratory Data Analysis - EDA

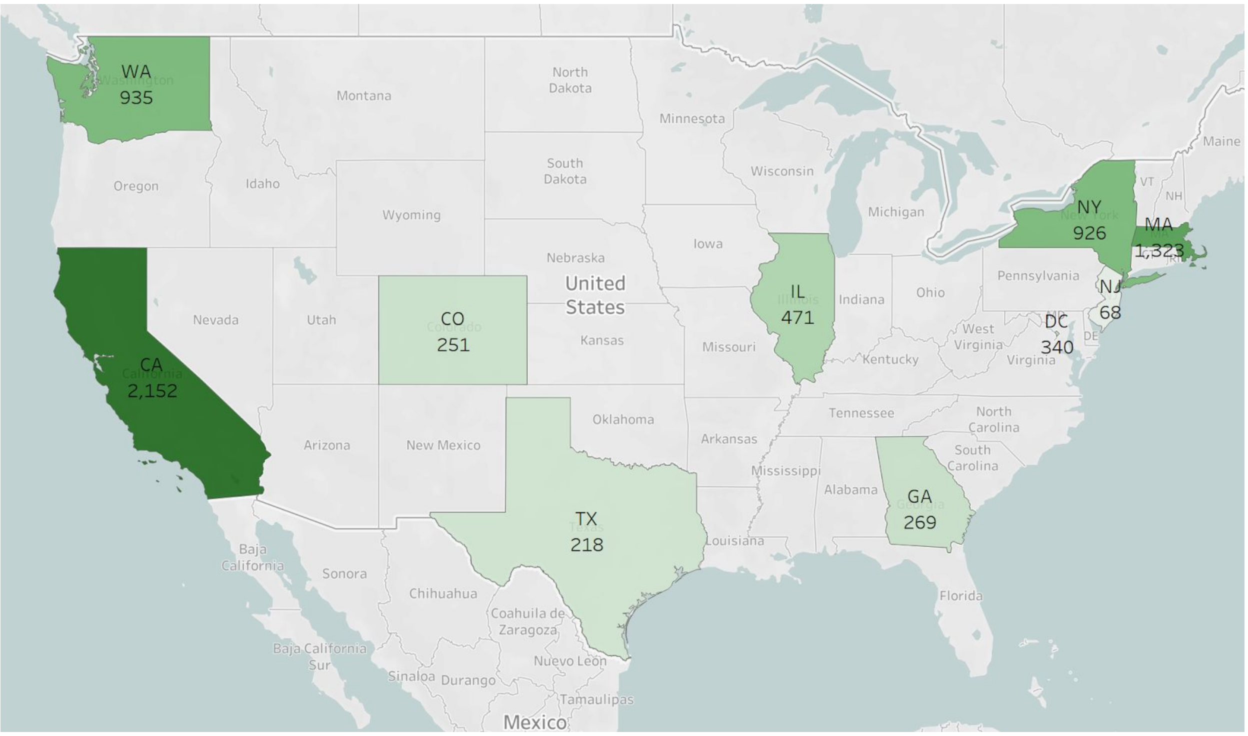


Figure 2: Top 10 States With the Largest Number of Data Scientists Jobs

First, we found the top 10 states with the largest number of data scientist jobs. As we can see from Figure 2, California is the top one among the states. As the leading state of technologies and creativity, California has a lot of technology companies with opening data related positions, followed by Massachusetts, Washington, New York, and Illinois. If one is looking for a data scientist job, these states are better locations since there are more job opportunities. In general, the west coast has more data scientist jobs, especially in California and Washington, but if one prefers to find future careers on the east coast, then Massachusetts, New York, and D.C. can be good choices. There are also opportunities in the middle of the country like Illinois and Colorado.

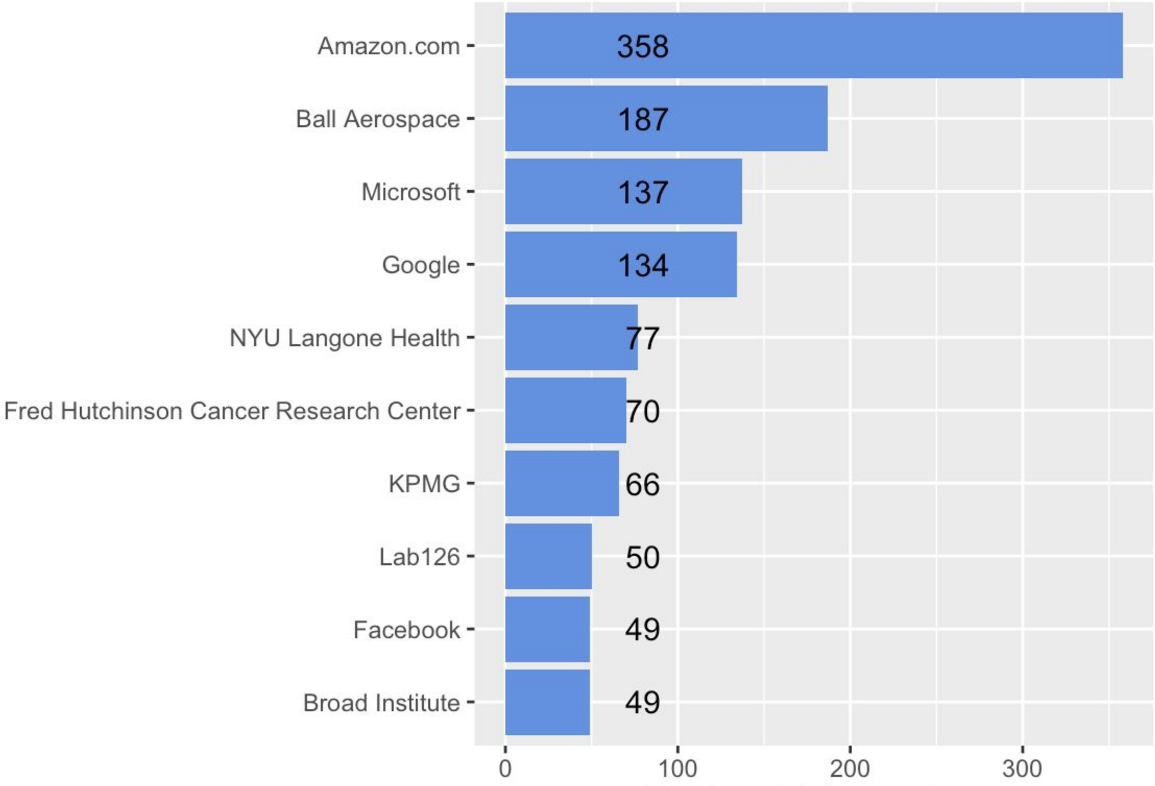


Figure 3: Companies with most data science job openings

Then, we identified the companies with most data science job openings. It’s not surprising that Amazon is the top one ranking company with 358 opening positions in our dataset. As a world-famous technology company based in Seattle, Amazon is providing millions of jobs for data-science people across the globe. According to Figure 3, Ball Aerospace, Microsoft, and Google also have more than 100 opening positions for data science related jobs. It is noticeable that when looking for data science related jobs, people should not only focus on traditional tech companies like Amazon, Microsoft, Google, and Facebook, but can also look for jobs in other types of companies like KPMG, which is one of the “Big Four” accounting companies, and also medical companies like the Fred Hutchinson Cancer Research Center. These companies also have a demand for data science related jobs and their job hiring competition may be less than traditional technology companies.

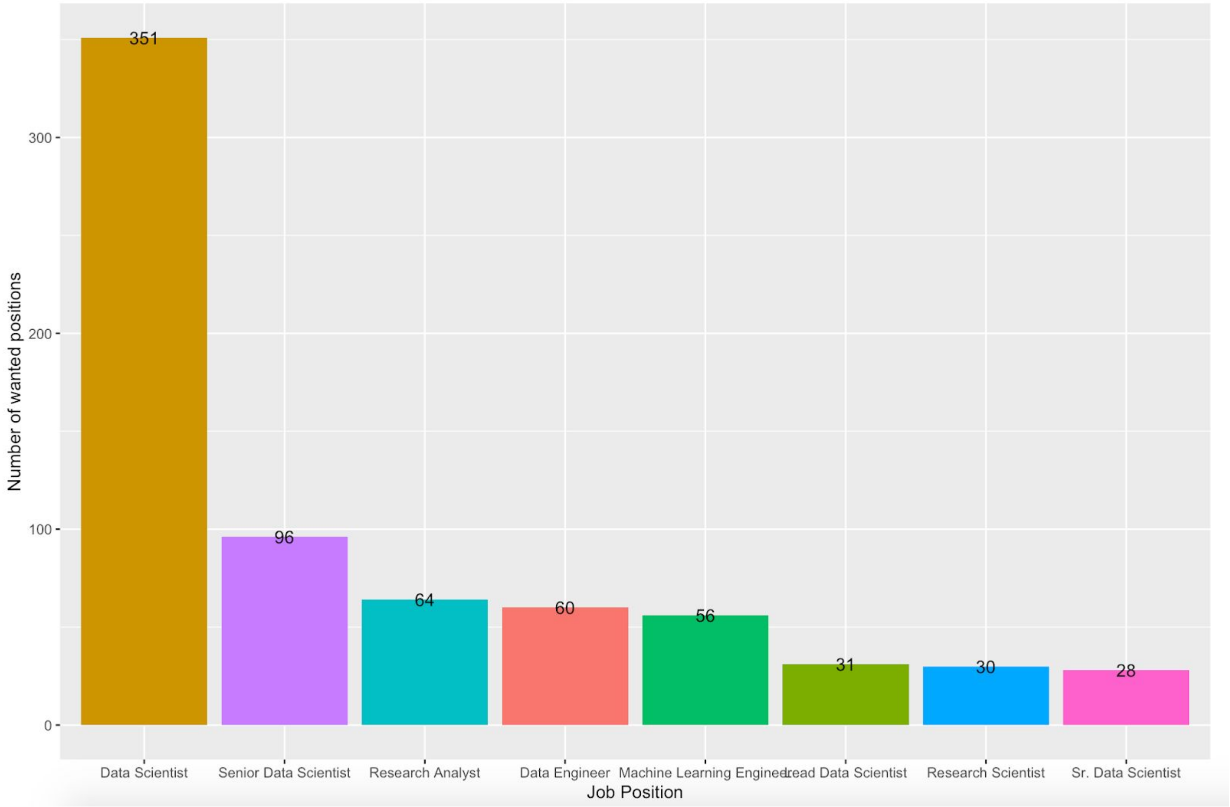


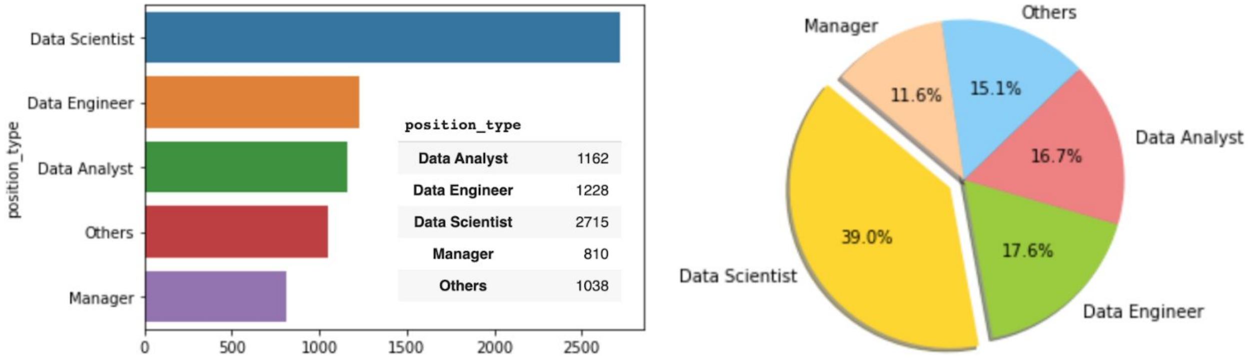
Figure 4: Top 8 Most Wanted Positions

In today's data-driven world, there are various kinds of data-related positions. Thus, our team is curious about what kind of data-related position would be popular when we are looking for jobs after graduation. As we can see from Figure 4, we listed the top 8 most wanted positions within the data science industry. The top three categories are data scientists, senior data scientists, and research analysts. Data engineers and machine learning engineers are also among the top 5. Without a doubt, the data scientist position is the most popular job type among all positions within the data science field.

Data Cleaning Process

In the “Data Scientist Job Market in the U.S” dataset, we have a column called “location” which includes both city, state information, and as well as zip codes. We decided to remove zip codes and split the “location” into two separate columns which are “city” and “state” which makes our following analysis easier. Since there are many similar positions in the data science job market, we want to categorize them into 5 main types which are Data Scientist, Data Analyst, Data Engineer, Manager, and Others. We came up with a list of keywords for each position and searched these keywords within the job titles and job descriptions.

Then, we counted the frequency of each position in the dataset. In Figure 5, the top three most popular positions are Data Scientist, Data Analyst, and Data Engineer. Thus, our team decided to focus on these positions in our following analysis.



position_type	
Data Analyst	1162
Data Engineer	1228
Data Scientist	2715
Manager	810
Others	1038

Figure 5: Top 3 Most Popular Positions

<u>The initial version of the plan. How to tackle the problem?</u>

As stated earlier, we want to help aspiring students who are interested in starting a career in the field of data science to navigate their job searching process. There are two main directions we want to take:

Direction I: Key qualifications the data science job market wants

Since the datasets include job descriptions for data science positions, we hope to conduct text analysis on the job descriptions. The result would help identify key skills, degree requirements, the preferred number of years of experience, and other possible factors that companies take into consideration during the recruitment process. It is important to recognize that every job description is subjective to the hiring firm specifically, so the uniqueness of each posting would potentially create difficulty in analyzing for a broader audience. Thus, we decide to try two analytical approaches:

- We use various packages to run an initial text cleaning on the job descriptions. The process will be challenging because every job description is one single text in the file. By tokenizing every word, certain meaningless tokens might appear more frequently than the actual important factors, such as python and R. The number of words within the ngram we use also matters. If we use bigram, factors such as the “Bachelor's degree” bigram will be obtained, but we would miss out on important information about the number of “years of experience” required.
- Since we have learned certain criteria that employers are looking for, we would come up with a list of qualifications on our own. For example, programming languages, communication skills, degree types, and the number of years of work experience would be on the list. After the list is done, we filter all job descriptions

to look for the overall frequency of each qualification. The goal is to find the most popular skills, degree requirements, and other possible important factors that companies are searching for among their job applicants. However, one drawback is that the list will be relatively subjective, so there might be factors that we might have missed when initiating the list.

Overall, we will work on both methods to get a better text analysis result. We plan to consult with professors, such as Prof.Brock Tibert, along the process to make sure that we are working effectively and efficiently.

Direction II: Salary prediction for candidates

The *Indeed dataset* provides a clean data frame with specific skills and qualifications listed. Since our text analysis is relatively limited compared to the analysis *Indeed* has done for the dataset, we plan to use the *Indeed* dataset for our machine learning model. We will employ various supervised machine learning methods, such as backward/ forward linear regression, lasso/ridge model, random forest, boosted trees to create predictions for how much a person with his/her unique background will make.

Analysis and Results

Unsupervised Machine Learning - Text Analysis

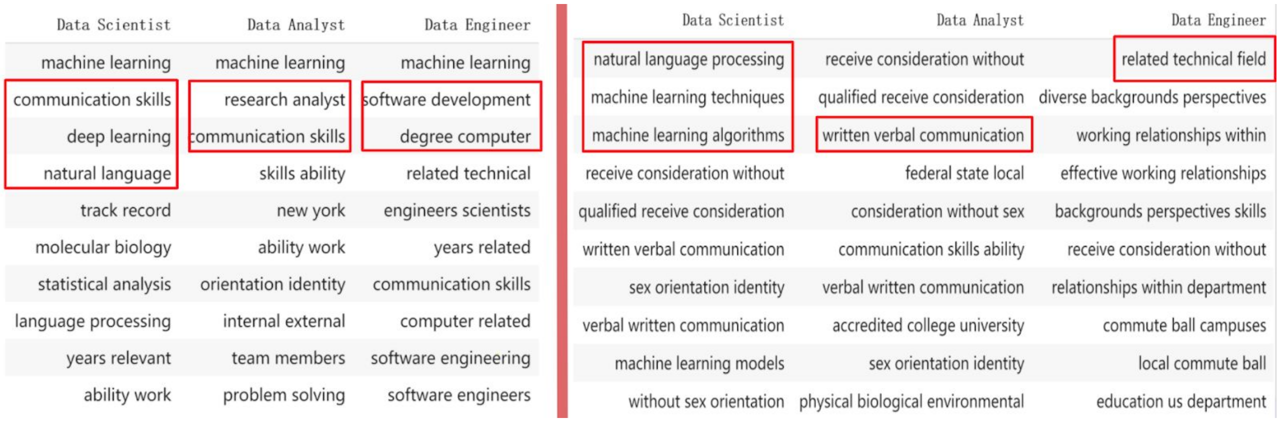
1. Word cloud

Firstly, we generated word cloud plots for each of these three positions - Data Scientist, Data Analyst, and Data Engineer. By plotting the word clouds, we gained better insight into the job responsibility of each position. Here are what we found: top three keywords for data analysts are “team”, “data”, and “client”; top three keywords for data scientists are “machine learning”, “communication skills”, and “data science”; top three keywords for data engineers are “machine learning”, “computer science”, and “big data”. It is clear to conclude that data analysts work with their team to analyze the data and report to their clients most of the time, whereas data engineers require a higher level of technical skill and knowledge to write and run algorithms at the backend. Data scientists are people whose job responsibility is between data analysts and data engineers. Data scientists are required to conduct the machine learning analysis and to have good communication skills in order to communicate their findings to their clients or managers.

2. N-gram

We conducted an N-grams analysis to tokenize the job description of each position into consecutive sequences of words. We set n to 2 and 3 to examine pairs of two and three consecutive words, often called “bigrams” and “trigrams”. As one might expect, a lot of the most common bigrams are pairs of common words, such as “equal opportunity” and “race color” which are employer-related. To make the result more skill-oriented, we added recruitment-related keywords, such as “race”, “gender”, “color”, and “equal” to our customized “stop-words” list. After implementing the stopwords, we noticed that for data analysts, words like “research”, “communication skill” and “problem solving” are being mentioned a lot; for data scientists, words like “machine learning” and “natural language processing” are emphasized more; for data engineers, words like “software development”, “software engineering” and “related technical field” were shown and have not been seen in the other two positions’ job descriptions. By comparing the results, we learned that “machine learning” is important for all three positions, while different positions have different priorities.

In Figure 6, we drew a table below to demonstrate the importance of technical skills for each position. We calculated the percentage of different skills mentioned in each position’s description. We learned that Python, SQL and R are popular skills required for data scientist jobs. For data analysts, Excel is the most important skill. For data engineers, C++ and java - considered as two of the hardest programming languages - are required. Based on the comparison, we concluded that a data engineer is the most technical position while a data analyst is the least technical position.



Data Scientist	Data Analyst	Data Engineer
machine learning	machine learning	machine learning
communication skills	research analyst	software development
deep learning	communication skills	degree computer
natural language	skills ability	related technical
track record	new york	engineers scientists
molecular biology	ability work	years related
statistical analysis	orientation identity	communication skills
language processing	internal external	computer related
years relevant	team members	software engineering
ability work	problem solving	software engineers

Data Scientist	Data Analyst	Data Engineer
natural language processing	receive consideration without	related technical field
machine learning techniques	qualified receive consideration	diverse backgrounds perspectives
machine learning algorithms	written verbal communication	working relationships within
receive consideration without	federal state local	effective working relationships
qualified receive consideration	consideration without sex	backgrounds perspectives skills
written verbal communication	communication skills ability	receive consideration without
sex orientation identity	verbal written communication	relationships within department
verbal written communication	accredited college university	commute ball campuses
machine learning models	sex orientation identity	local commute ball
without sex orientation	physical biological environmental	education us department

Figure 6: Important Keywords for each position

3. Word Frequency

In Figure 7, we compared the sum of keywords frequency by each position. The graph demonstrates that data analysts tend to focus more on data analysis, data visualization, and research, and data scientists tend to focus more on machine learning, modeling, and statistical analysis.

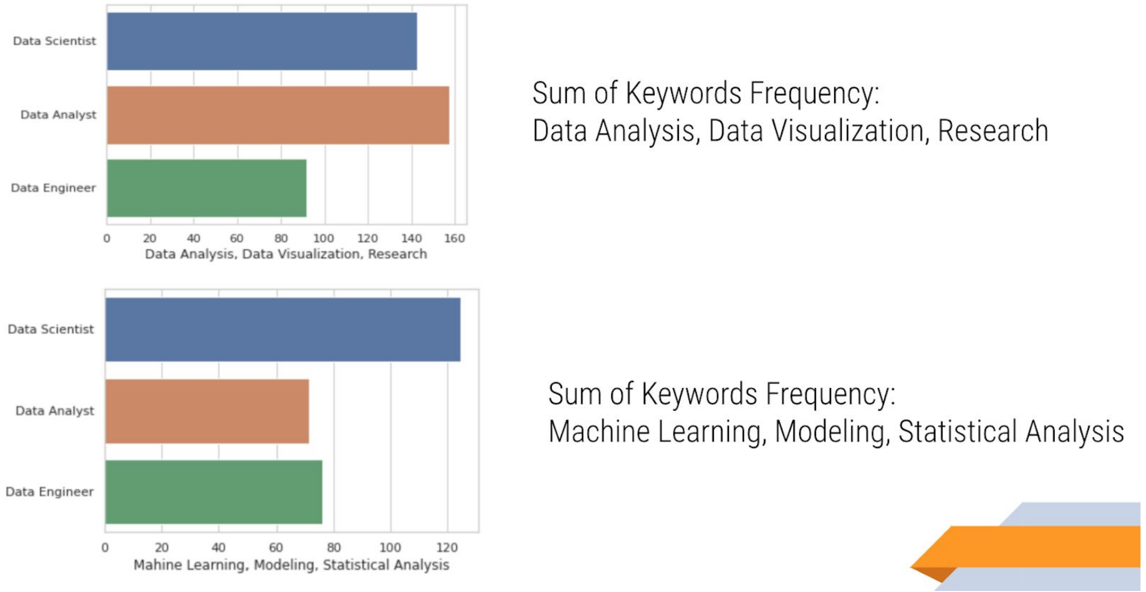


Figure 7: Sum of Keywords Frequency for each position

By comparing the degree requirement for each position in Figure 8, we learned that all three positions require at least a bachelor's degree while higher degrees are preferred. Specifically, the data scientist position tends to favor Ph.D. candidates the most.

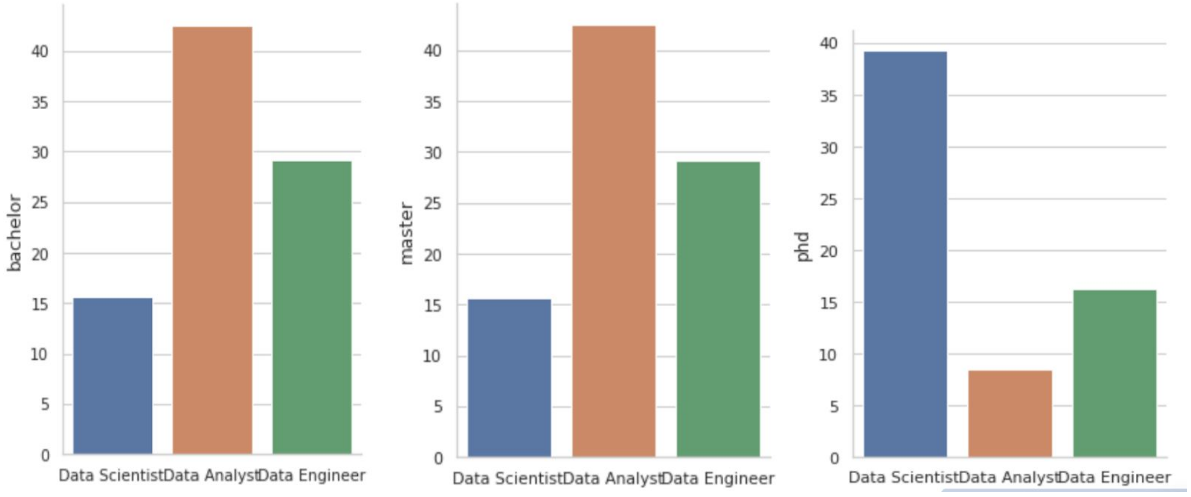


Figure 8: Degree Requirement for each position

Supervised Machine Learning - Predictive Model

We used the indeed dataset to train a supervised machine learning model to help predict a candidate's future salary in the data science field. The model was trained in the following methods:

- Linear Discriminant Analysis (LDA)
- Stepwise Regression
- K-Nearest Neighbors (KNN)
- Support Vector Machines (SVM) with a linear Kernel
- Random Forest (RF)
- Boosted Trees

The input variables are factors that indicate candidates' and company's backgrounds; the output variable is a numeric salary value or a salary range.

In the original dataset, the salary for each job posting was a range, such as $80,000 - $100,000. As a result, we looked at it as a classification problem and trained a supervised machine learning model that predicted a salary range. However, the accuracy score for the best model was only 49%. If the correct salary is $79,999, but the predicted salary range is $80,000 - $100,000, the model would consider this prediction as wrong. Considering that salaries are often negotiable and the difference between $79,999 and $80,001 is not drastic, we decided to recalculate the accuracy score by also counting the neighboring buckets as accurate. This change improved our accuracy score to 89%.

We also used the same input variables to train a second model where we look at the issue as a regression problem. We assigned the median value of the correct salary range to each job posting. For example, if a position lists their salary ranges between $80,000 and $100,000, the newly assigned salary value would be $90,000. Thus, we were able to predict a numeric salary value instead of a range. Regardless of the output formats, the inputs and the training process for both models are exactly the same.

The dataset was divided into 80% training set and 20% testing set. 10-fold cross-validation was applied to the training set to improve our model. Out of all the models, the random forest is the best with the lowest RMSE value and accuracy score.

We asked our networks to fill out a salary prediction survey. They were asked to indicate their technical skills, industry preference, work location preference, and etc. Out of the 60 survey respondents, 73.33% have **at least** 5 technical skills - Python, R, SQL, Tableau, and Machine Learning. The most popular locations are New York, Massachusetts, and California. The most popular industries are consulting and technology. The predicted salaries range from $94, 424 - $142,819 with a mean of $116, 252.

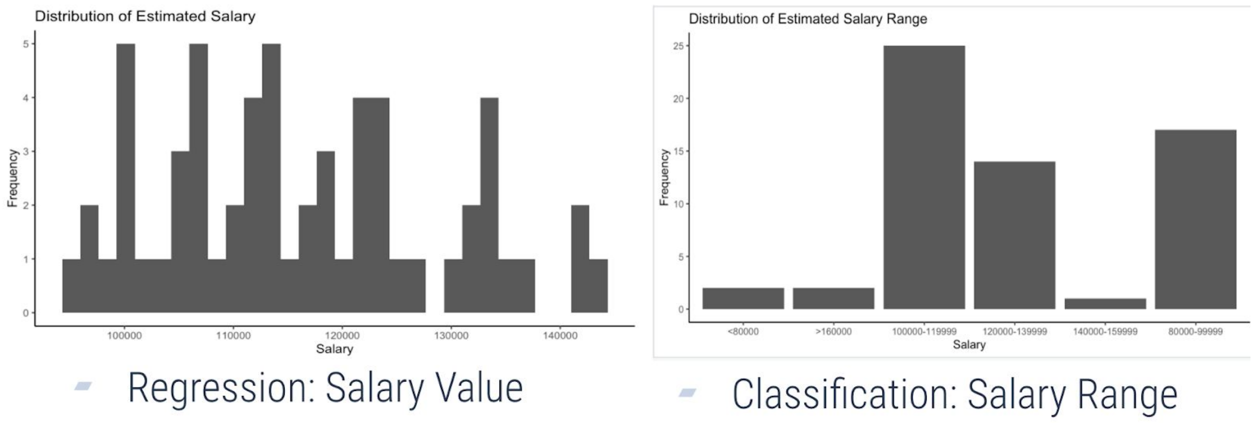


Figure 9: Results of Prediction Model Based on Our Survey

Discussion around Results

By creating an average profile for MSBA students, we were able to conduct salary prediction for students in the BU MSBA program. The profile information is below:

- Skills: Python, R, SQL, Tableau, ML
- Location: NY
- Industry: Consulting
- Position: Data Analyst
- Company: NA

As a result, we were able to predict that on average, the salary range would be $100,000 - 119,999 and the numeric value would be $114,208.33.

This model is relatively inflated for various reasons. In our dataset, we did not have information regarding education, years of experiences required, and some other key factors that determine a candidate's salary. More importantly, our survey responses are largely biased towards big companies, such as Google and Amazon, at which they tend to pay their employees above the industry average.

Limitations, Challenges and Future Work

We faced some challenges and limitations throughout the process. Firstly, data cleaning is very hard at the beginning because our job description is a very complex text and includes too much information. Besides, we have unverified categorization when putting different job positions into one of the chosen three main types. Next, due to the sample size of our survey, the results might not very representative. As we stated earlier, another problem is that our model has an inflated response. More than half of people want to work at big companies such as Amazon or Google. For further analysis, if we have more time, we will try our best to gather more survey responses and get more accurate predictions. Lastly, the most difficult part of our project is that our team was unable to communicate face to face because of this unexpected epidemic.

Conclusion

In conclusion, If one wants to become a data scientist, he or she needs to have proficiency in technical skills such as Python, R, and SQL. If one wants to become a data analyst, Excel, SQL, and R are the most important skills compared to other programming languages. If one wants to become a data engineer, the more technical skill one needs to have is Java, and even C/C++ or Python. Even though different positions have different job skills emphasis, both three categories have the same essential skills which are machine learning, data analysis, and data visualization. Additionally, when applying for jobs, resumes are the core part of the first impression. Based on our analysis, we want to remind our MSBA classmates to read job descriptions carefully and make sure to include your technical skills on your resume.

Acknowledgments

Our team would like to take this opportunity to express our sincere gratitude to those who have given us guidance, inspiration, and encouragement throughout the process. Firstly, this project would not be possible without the support of Professor Mohammad Soltanieh-ha. His dedication to students was remarkable because we would have not completed this project without bi-weekly project meetings with him.

All professors at the MS in Business Analytics program have provided their expertise to help us navigate through the project. The staff member at Questrom School of Business assisted us to make this project possible logistically.

Lastly, we feel very grateful to many people, especially our classmates and team members ourselves, who have made valuable contributions and put lots of effort into this project. We thank all the people for their help directly and indirectly to complete our capstone project.