2020 MIE PROPOSAL

Abstract

The document contains the proposed changes for MIE1624: Introduction to Data Science & Analytics and a proposal for a 16 month MSc in Data Science & Artificial Intelligence.

Anshul Verma
Hiba Doudar
Nada El Moghany
Obaid Muhammad
Zoha Sherkat-Masoumi

Table of Contents

INTRODUCTION

This report will examine a study on re-evaluating the current course MIE1624 Introduction to Data Science provided at the Department of Mechanical and Industrial Engineering at UofT, and propose a 16 month Professional Master degree in Data Science and program. The motivation for this project is rapidly increasing demand for data-scientists. The main dataset of this study is based on web-scraping the job postings on common job search engines such as Indeed, Monster, and Workpolis, to examine skill sets demanded by the industries in Data Scientist. We have also used the Kaggle survey 2019 dataset of. We have combined the results obtained from both the datasets to obtain inferences and results.

The survey revealed that annual salaries of Data Scientists in the United States and Canada has a median of 120,000 USD. We also discovered that most professionals with a Master's Degree or above are earning higher salaries. Furthermore, the survey also revealed that among the participants Coursera as a learning tool is more common than the courses offered at universities. This was true in particular for professionals with a Master's degree. This finding demonstrates a market need for developing and upgrading skills in Data Scientist at a Master specialization degree.

MIE1624 COURSE CURRICULUM

Course Overview

The objective of the course is to learn analytical models and overview quantitative algorithms for solving engineering and business problems. Considerable attention in the course is devoted to various data science and analytics techniques such as basic statistics, visualizations, regressions, machine learning and optimization modeling, text analytics, simulation modeling, artificial intelligence using Python. Practical aspects of the topics listed above are also emphasized. The course also devotes a lot of time demonstrating application of computational and modeling algorithms in finance, speech recognition, marketing, health care, predictive maintenance, web and social media analytics, etc. Various python libraries, IBM Watson Analytics, and AWS visualization software are also used and explained in this course.

Redesign of MIE 1624

In order to investigate the usefulness of topics covered in MIE1624, technical (and business) skills were extracted from 30,000+ job postings from Indeed and Workopolis. It was found that most of the topics discussed in the course align with the current trends of the job market. The topics that were not covered are highlighted in pink in the flowchart (Figure 1); while those highlighted in blue are topics

valued by the job market but are already implemented in the current course.

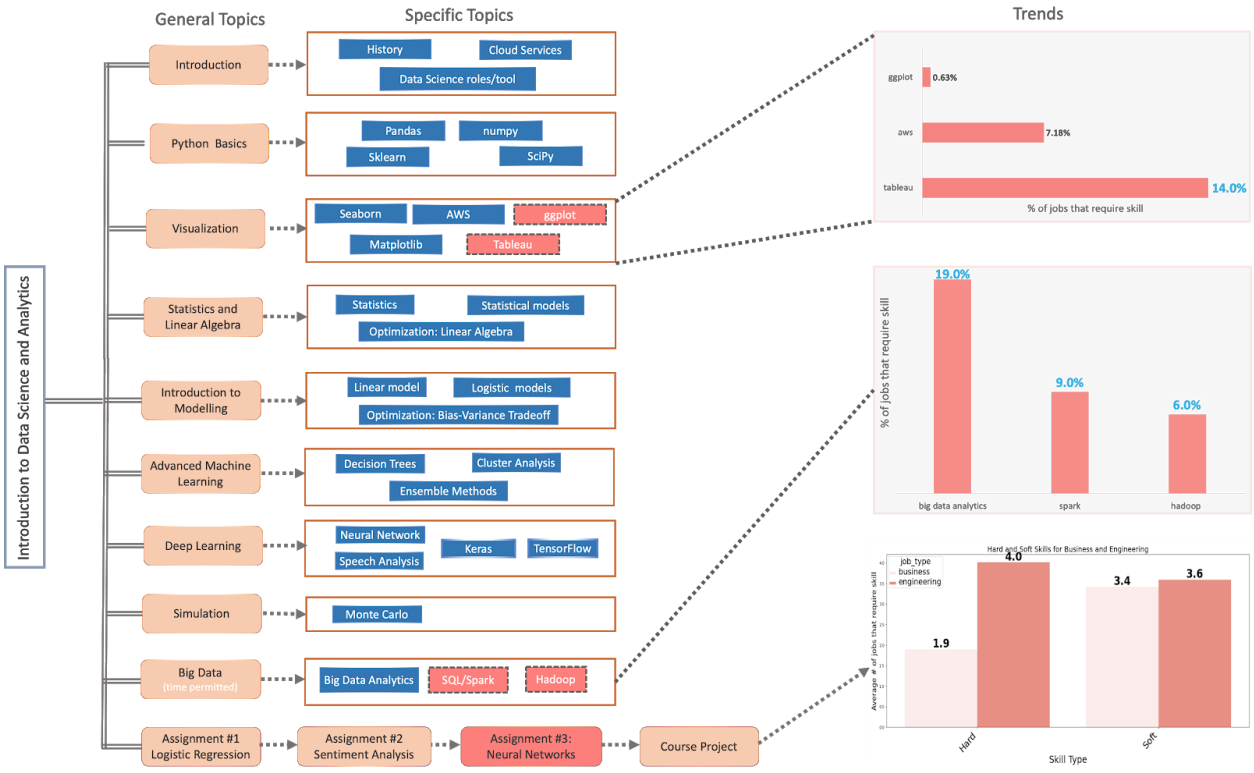


Figure 1: Course Redesign of MIE1624: Introduction to Data Science and Analytics

In terms of visualization techniques, tableau is a highly sought after skill where approximately 14% of employers are looking for individuals with knowledge of the software. Though not heavily covered in this course AWS is also quite popular. Big data is another desirable skill that was quantified by inquiring for topics such as sql, spark, and hadoop. Around 36% of jobs wanted working knowledge of SQL (Appendix fig: A5), while about 19% wanted individuals that can apply big data analytics topics such as data manipulation or data wrangling; and 9% are looking for individuals with spark knowledge. The suggestion is to cover big data only if time permits as python is the primary programming language in the course; teaching a separate language might overload future students.

Lastly, the course outline was slightly adjusted, instead of covering visualization libraries after the modelling techniques lecture, it should be taught after python basics as this will help the students with their assignment (which requires exploratory analysis). In addition, the deep learning lecture is recommended to be covered directly after machine learning (rather than after simulation) as these topics are correlated. Further in terms of course evaluations, the exam was replaced with a third assignment because the data showed that soft skills such as communication and reporting are equally important for engineering and business jobs. Completing projects is an effective way of improving these skills. The suggested topic for Assignment #3 should be neural networks as they were relatively as important as NLP (Appendix fig: A6) Whereas, Assignment #1 and #2 should be left unchanged since linear/logistic and NLP models are still valued in the current job market (Appendix fig: A6 and A7).

BUSINESS ANALYTICS GRADUATE PROGRAMS

There are numerous websites, blogs and posts on the internet ranking the top universities and graduate programs in the world. Some of the most popular university ranking agencies include Shanghai Ranking Consultancy ("Shanghai"), Times Higher Education ("THE") & Quacquarelli Symonds ("TopUniversities"). TopUniversities from these three ranking agencies are the ones which rank in the top Business Analytics Graduate Programs. Shanghai & THE doesn't provide ranking for the specific Graduate Programs, instead it's ranking is limited to the Field (i.e. Mathematics, Computer Science, Life Science, etc.).

Features to analyze different programs to skim out top Business Analytics Graduate Program, are the following rankings & their evaluation criteria:

- TopUniversities's 2019 Business Analytics Graduate Program Ranking
- THE's 2019 University's Computer Science Field Ranking
- Shanghai's 2019 University's Computer Science & Engineering Field Ranking

Current Programs

The following Business Analytics Graduate Programs are considered the Top 10 based on the overall ranking from TopUniversities with supplemental University's Computer Science / Engineering ranking from THE & Shanghai. Looking at an average ranking of these ranking agencies, MIT's Master of Business Analytics is considered the highest ranked program. Let's look deeper into what drives these rankings.

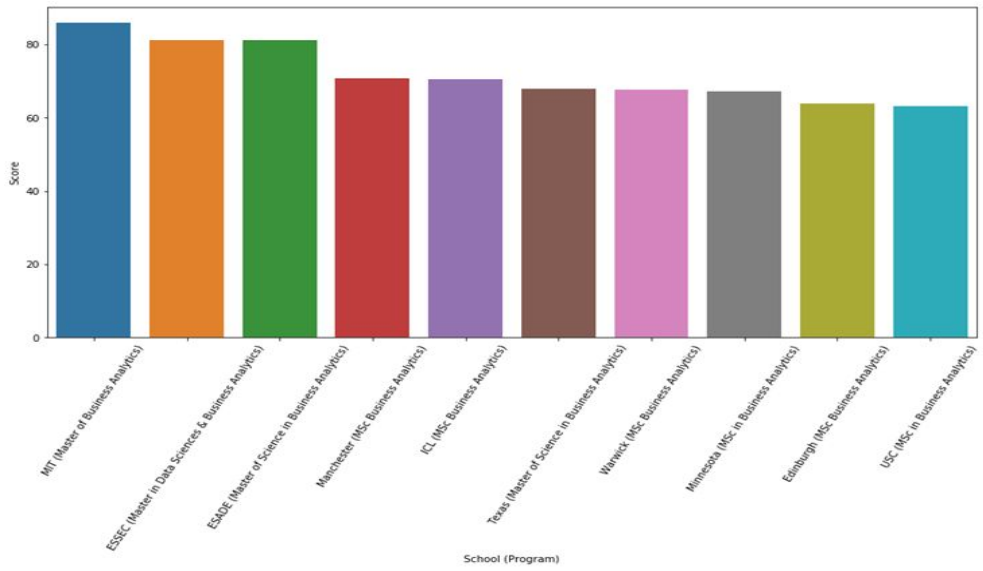


Figure2: Top Graduate Programs in Business Analytics

Evaluation Criteria

To propose our Graduate Program, let's consider the criteria for each of the popular ranking websites. For TopUniversities, the Business Analytics Graduate Programs are evaluated on: Value for Money, Diversity, Thought Leadership (Teaching), Alumni Outcomes & Employability. For TimesHigherEducation, University's Computer Science Field Ranking is evaluated on: Teaching, Research, Citations, Industry Income (Employability) & International Outlook (Diversity). Lastly, for Shanghai, University's Computer Science & Engineering Field Ranking is evaluated on: Diversity, Citations, Research.

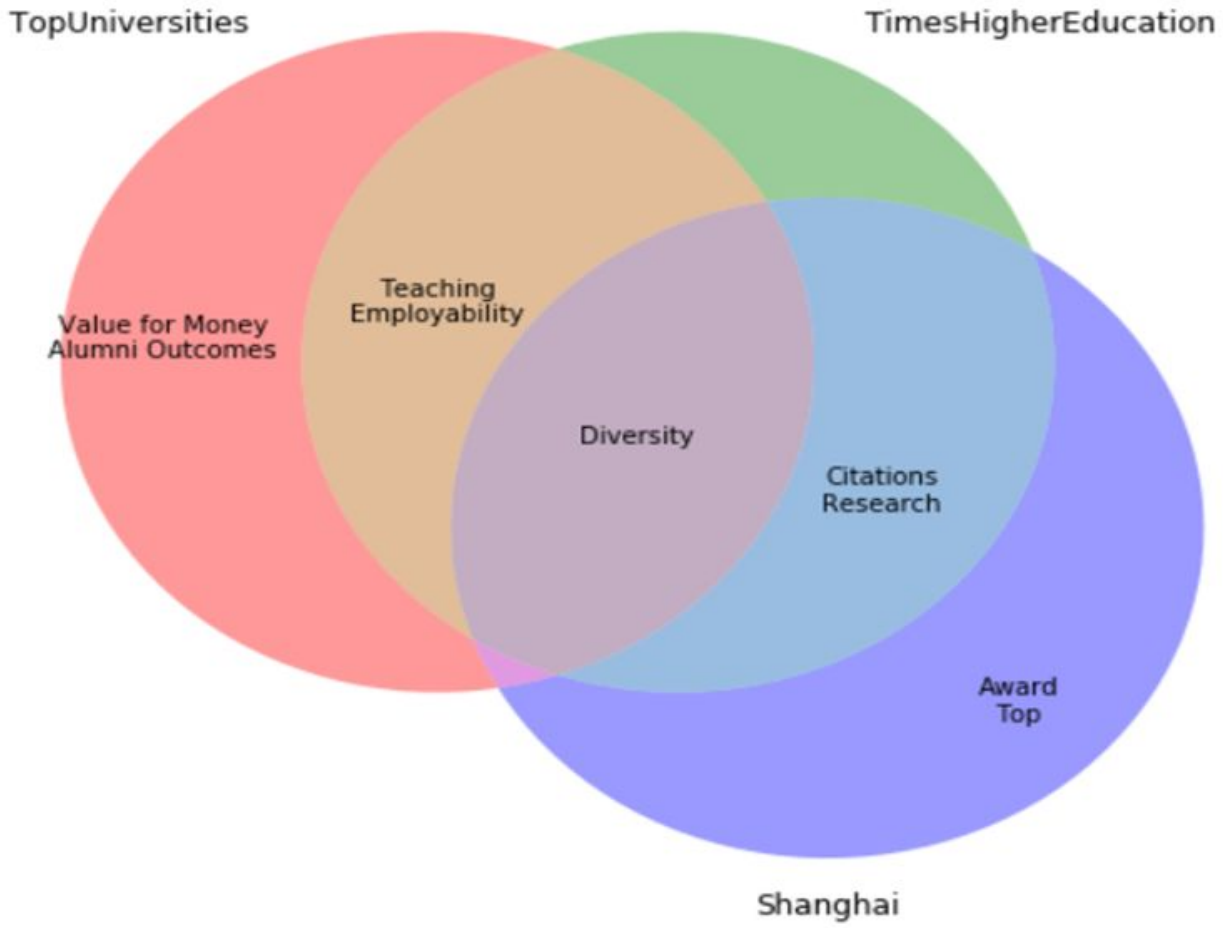


Figure3: Ranking criterion

The Venn Diagram above provides an overlap of the evaluation criteria used by each of the three popular ranking agencies. Using the overlap we design a framework to structure our graduate program suggestion:

- Diversity: To introduce diversity into the proposed graduate program, having an equal gender ratio and the amount of students from a diverse country of origin is considered important.
- Research: The research aspect of the program is driven by UofT's first ranking in the field of research in Canada from different rating agencies.

- Teaching: The teaching aspect is driven by the ratio of students to faculty. As such, the goal is to limit the class size to 20.
- Employability: To increase the employability, it's important to provide internship, alumni involvement working in the industry early on for coaching and courses designed based on the current market need (current job postings).

MSc DATA SCIENCE & ARTIFICIAL INTELLIGENCE

Program Overview

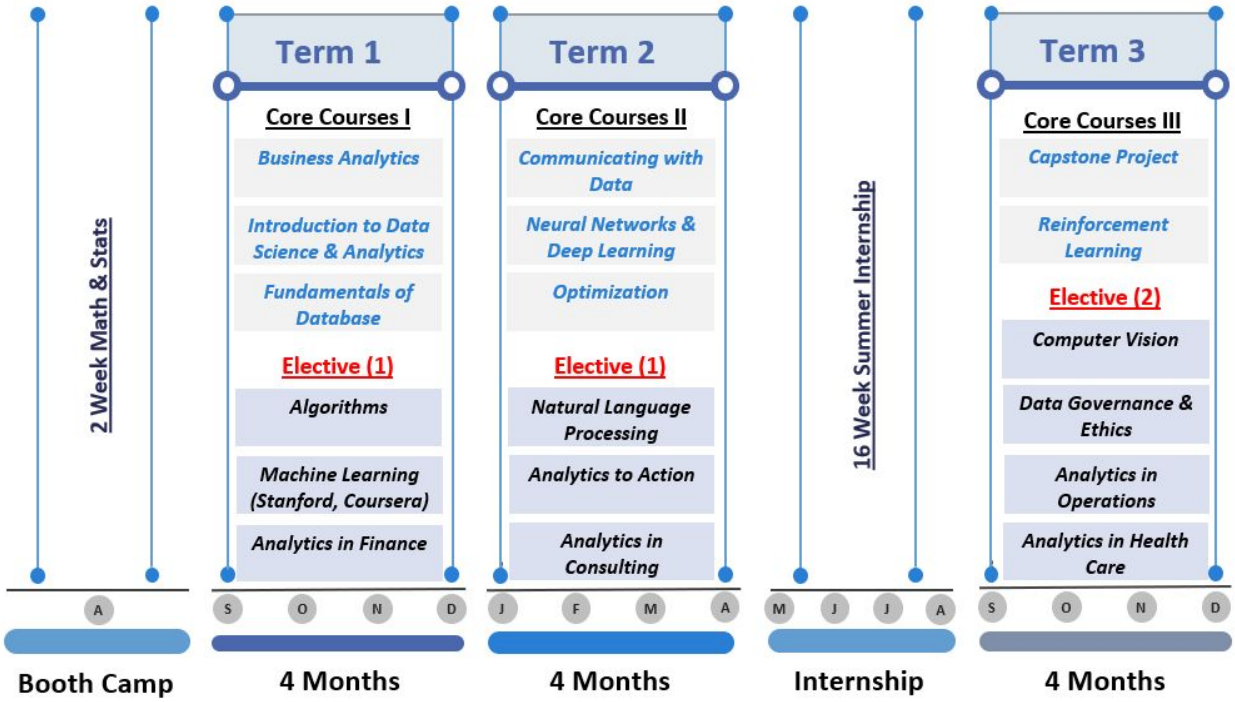


Figure 4: Program Overview

The MSc DS&AI should be a 16 month program which starts in Mid August. The Program should start off with a 2 week intensive boot camp that should be a quick review of math & statistics since these are core topics in DS&AI. The program consists of 12 courses including 8 core courses and 4 electives. These courses are designed based on the fundamental knowledge required to succeed in a data science job, technical and soft skills that the employers are looking for. There is also a 16 week summer internship which is very crucial since it provides students with much needed professional experience (Appendix fig: A8). Along with these the program should also include hackathon, and extra credits for active participation in kaggle competition since our analysis shows that coding experience is correlated to the salary(Appendix fig: A7).

Courses

Math Camp (1 Course)

The two week intensive preparatory course in Mathematics and Statistics which is every weekday for 6 hours during the month of August. The course is subdivided into Mathematics & Statistics. The Mathematics topics include Vectors, Matrices, Linear Algebra, Functions, & Calculus (Differentiation & Integration). The Statistics topics include Sample & Event Spaces, Probability (Axioms & Rules), Random Variables, Pdfs/Pmfs, Functions of Random Variables, MSE, CEF, BLP, Key Expectation Theorems & Probability Models.

Core Courses I (3 Courses)

Business Analytics

The course examines some of the most popular methodologies being used in the industry to draw inferences from data to transform businesses and industries. The course provides examples and case studies of analytics in finance, technology, energy, healthcare & industrial sectors.

Introduction to Data Science & AI

This course will cover every stage of processing data, working with raw datasets to evaluating and deploying ML and AI models. The course will teach using Python libraries, such as Pandas, Numpy, SciPy to extract and to clean dataset. Then using Python plotting libraries as Matplotlib, Seaborn, Plotly to visualize data. Then the course will train students to use Scikit-learn to build, evaluate and deploy Machine learning (ML) - on supervised and unsupervised learning and Artificial Intelligence (AI) models. Finally, case study will examine these skills in solving a public concern using data sources, as tweets, survey, Kaggle dataset.

Fundamentals of Databases

The course examines database fundamental concepts to apply foundation knowledge of SQL (or structured Query Language) to communicate and extract data from databases. Case study will provide hands-on experience to train students on working with real databases, creating databases in cloud, practice building and running SQL queries, and access databases from Jupyter notebooks using SQL and Python.

Core Courses II (3 Courses)

Communicating with Data

The course examines fundamental quantitative techniques of using data to make management decisions and enhance decision making skills. Exercises and case studies will be presented from marketing, finance, strategy and other operation management.

Neural Networks & Deep Learning

The course examines core concepts in neural networks and deep learning and application in image recognition, speech recognition and natural language processing. Case studies will illustrate the core concepts behind neural networks and deep learning.

Optimization

Optimization is a key tool in statistical, machine learning, and business operational models. This course will examine fundamental optimization tools available to data scientists and business analysts. Including on-hand experience in optimization projects using software packages such as CPLEX and modeling on AMPL.

Core Courses III (2 Courses)

Capstone Project

This should be an open ended research project where students should be able to demonstrate their learning from all the previous courses in any of the topics of their interest. The progress of a project must be analysed and graded weekly, to ensure proper guidance to struggling students and for them to know if what they are doing is in accordance to what is expected of the project by the advisor.

Reinforcement Learning

Reinforcement Learning is a vast and math intensive computer science topic. But in data-science the knowledge of reinforcement learning needs to be more application based. Therefore this course should be a more application based course focusing more on application of topics such as Convolutional Neural Networks, Recurrent Neural Network, End-to-End model, Generative Adversarial Networks, LSTM cells, GRU cells etc.

Electives (4 Courses)

Students should have an option to take 4 or more elective courses with 1 in first two terms and 2 elective courses in the third term. Due to the popularity of Coursera (Appendix fig: A3) students can choose to take one or two Machine Learning related courses offered by Coursera like the one offered by Andrew Ng of Stanford. The course is very popular with a rating of 4.9 based on ~130k ratings. Below is the list of other elective courses that should be offered:

Algorithms and Data Structures, Machine Learning (Coursera), Natural Language Processing, Analytics to Action, Computer Vision, Data Governance & Ethics, Analytics in Finance, Analytics in Consulting, Analytics in Operations, Analytics in Healthcare.

Application Process

Since, U of T is at the top of university ranking in Canada a master program like the one suggested will be highly in demand. The application process must vigorously classify candidates based on their education background, experience and leadership skills.

Fig 5: Application Process to the Professional Master Degree

The Figure 5 illustrates the application process. The department should be posting information brochures and contact details of the graduate office and alumni, so prospective students could connect with them early on. There should also be admission events where one can get online or in person to have their questions answered about the program admission details. In Step 3, where a candidate applies for the program after preparing their resume, letter of intent, and background experience, we must notify them that we want to hear their story.

Step 4 is a case interview assigning the communication and problem solving skills in selecting candidates. It will be a general case, not concentrating on the data analysis. It is aimed to explore candidate plans and future careers to understand better on how the prospect candidate will fit in the program. Atleast, one faculty member or an alumni working in the field should be present during the Case Presentation. In final review the candidate can be asked some technical and behavioural questions to give the test taker more knowledge on how the candidate would perform in a competitive master's program. Final decision must be based on the faculty members and selected aluminis judgement.

STARTUP IDEA

Our startup idea is based on resume analysis, job recommendation, and skills to learn recommendation. We have named our startup idea as Data Analytics Counselor because it aims to guide the students to have a better future in the field of data analytics. We have not only designed the mechanism on what we want from our startup but we have also made a prototype web-application to showcase its working.

The demonstrated working prototype is totally data based but we would like to add opinions and suggestions of some pioneers in data-science in near future to make the overall suggestions more dependable . It should be an important step which would convince and attract more people to our app in near future because for some people data is still not the substitute of experience.

Data Analytics Counselor Start-up Overview

Data Analytics Counselor is a start-up aiming to accommodate students to find jobs in data science. It will help students save time proving them job postings matching their resume and it will provide them much needed guidance to learn and add new technical topics into their resume to improve their chances of getting a job in data science. Data Analytics Counselor is a web-based application aiming to automate job searching and counselling for the data-science field. It aggregates job postings available in websites such as Indeed and Monster and matches students to the most suitable jobs available in the market. It also utilizes the information scraped from job postings and Kaggle surveys to discover important skills in the field. Using the set of important skills, it ranks a resume, compares skills in resume with that required for a particular job posting and if the overall match is over 90% then it suggests that particular job to the resume holder.

Resume Analysis

The analysis is all based on a user's resume. We accept a .pdf form resume of users and extract skills and topics mentioned in the resume. Based on the weights of skills in the resume and skills in demand in the market the resume is scored out of 1. Then using the scores of various skills in the resume and other in demand skills in the job market. We find the most weighted missing skills in the user's resume and then we divide the big skill topics into subtopics and suggest users to learn more in demand skills missing from their resume.

Job Recommendation

Based on the skills extracted from the resume and the skills required for different jobs that we scraped for part1 . We look for the match of profile of the job posting and the user, we find this match attribute by comparing the weights of skills in the resume with the max possible weights of required skills in the job and if there is an over 90%(score, 0.9) match then that matching job is suggested to the user.

Out of all the matching jobs for a user we select and display five job postings randomly so that every time the user logs in he/she gets to see and apply to a new or different set of jobs.

Data Analytics Counselor Manual

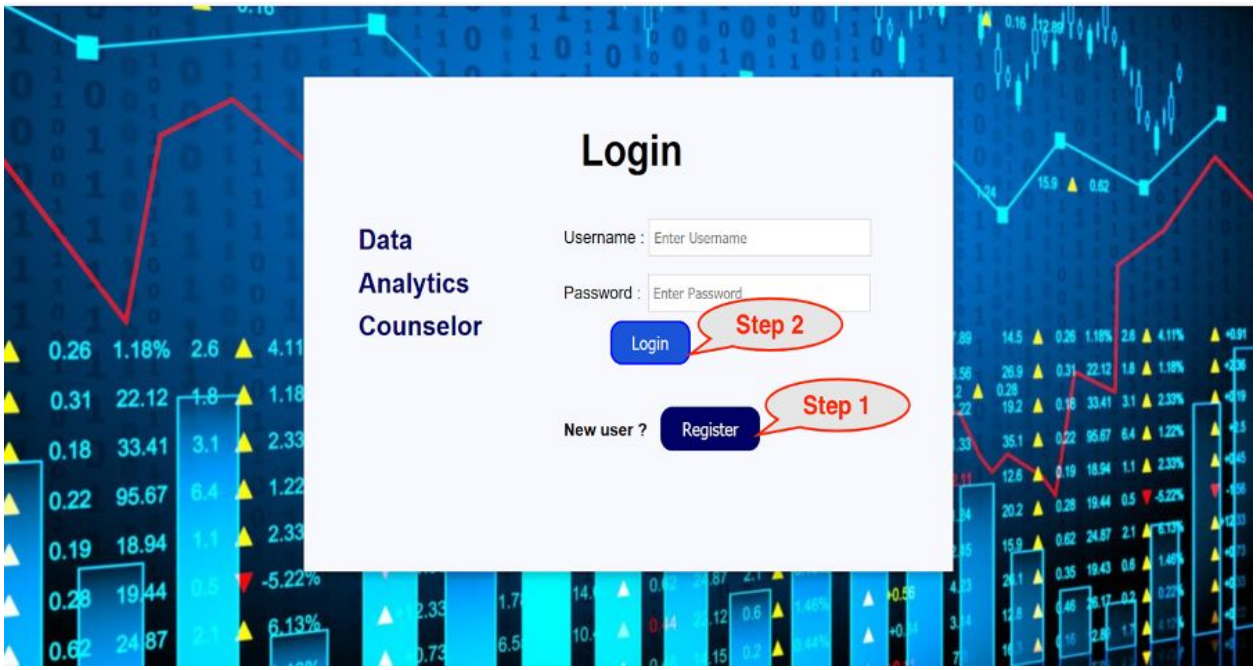
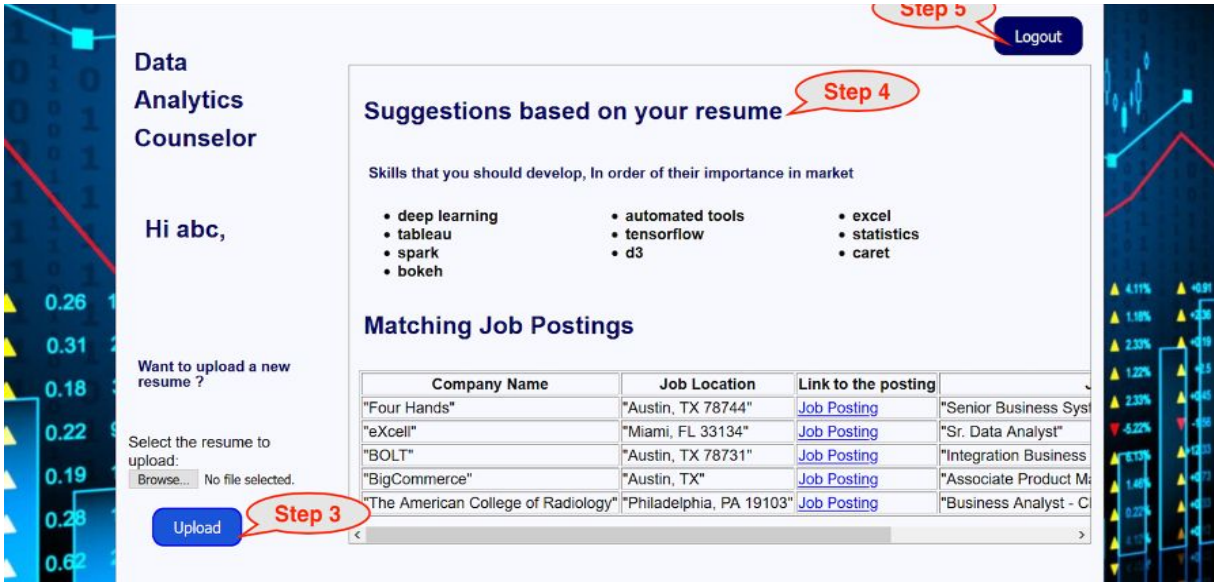


Fig 6: Prototype Main Page

The following two steps are needed to login to Data Analytics Counselor website:

1. Register: New users are required to register before accessing the website. In order to register, students need to provide a unique username and password. Passwords are encrypted and saved in the database to ensure security to the registered user.

2. Login: Users enter their username and password, then press login for the main page to open.



, Fig 7: Prototype user=abc/Home

User requires to take next three steps for the resume analysis to be done successfully:

3. Upload: Users Need To browse a .pdf type resume in their PCand press upload.
4. Suggestion: Skill improvement with the highest increase in chance of a user to land a data analytics job are suggested along with job postings matching the resume.
5. Logout: To exit from home page logout is pressed.

Suggestions include two categories which are explained in detail below:

1. Skills you should develop: Top important technical skills are compared against skills present in the resume uploaded, the missing skills with high enough importance are suggested to users. The application only stores one resume per user and at the time a second upload is made it automatically deletes the old resume and provides the user with the analysis result for the new updated resume. Also the resume uploaded by one user can in no way be accessed by another user because we have used a relational database and the resume information table has a unique foriegn key which is different for each user. This was done to protect the app against any data breach.

2. Matching Job Postings: Resume is matched with available jobs based on the important skills in the resume which is compared to job postings available and the matched weight is given to each job posting. The jobs with the highest weights are suggested. The company name, link, company location and salary for each suggested job is displayed to speed up the job searching procedure.

Restrictions

1. The app is just a prototype and has some restrictions, one of the most important restrictions is that it only suggests technical skills.
2. While analysing the resume it looks for keywords in the resume which currently only has a fixed set of words based on our analysis for part 1. These keywords are clubbed together into a topic

to assign weight to each skill topic present in the resume. So, if the resume has uncommon words which are not there in the list then it's not counted towards the resume score. To overcome this, we need to improve the resume analysis part of the application by using more complex NLP models.

3. The set of job postings include 30,000 jobs which we web scraped for part 1 and it's a constant dataset, but in the real world the job postings are constantly changing and we need to come up with a way such that the job postings are automatically updated in the app. To overcome this, we need to constantly update the job posting dataset in the app.
4. The prototype only runs on one AWS EC2 instance currently and will crash if the load is high. To overcome this, we will need to implement a load-balancer with an auto-scaling worker pool so that the app can manage high traffic.

Data Analytics Counselor in near future

In future this website will be extended to include other engineering and business fields to accommodate a wider range of students. We also plan to include course recommendations for students to fill their skill gaps. This will include online and in person courses for their specific skill requirements to increase their chances of getting a good data-science job.

APPENDIX

Figures

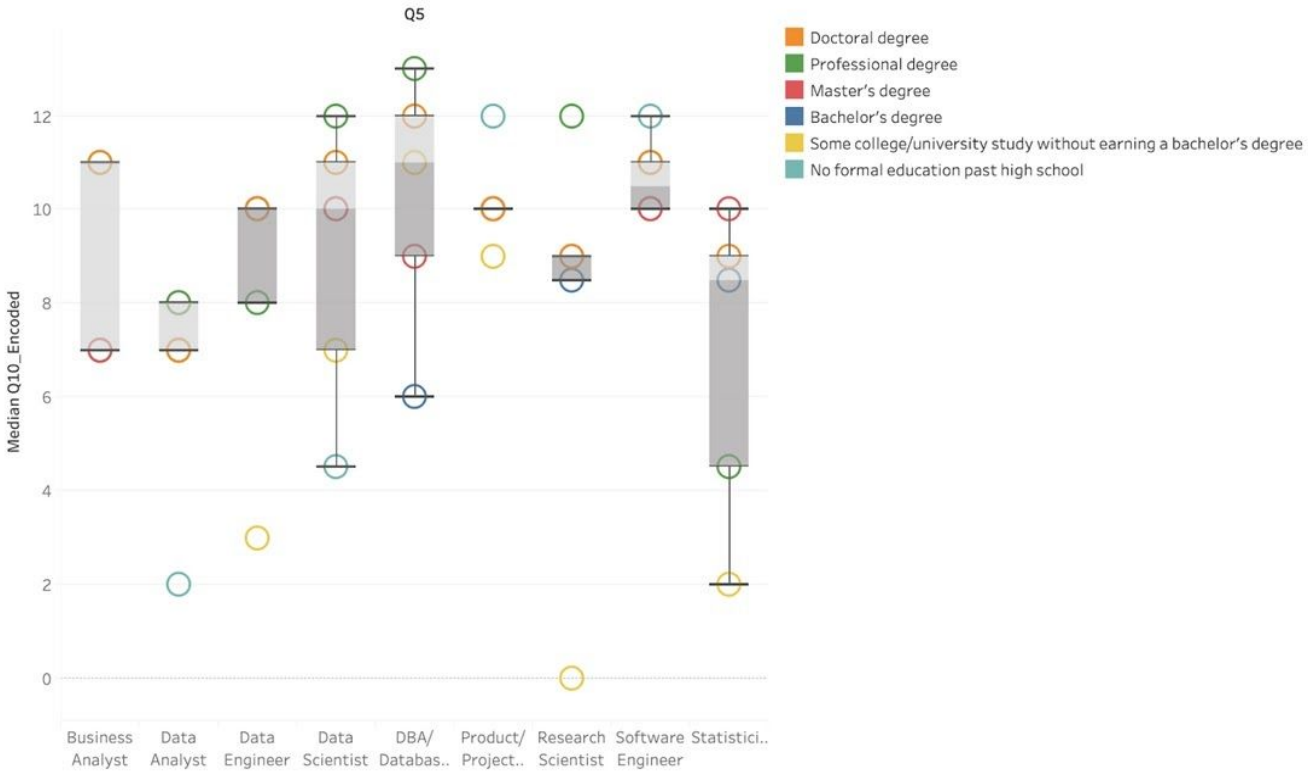


Figure A1: Salary Breakdown by Degree

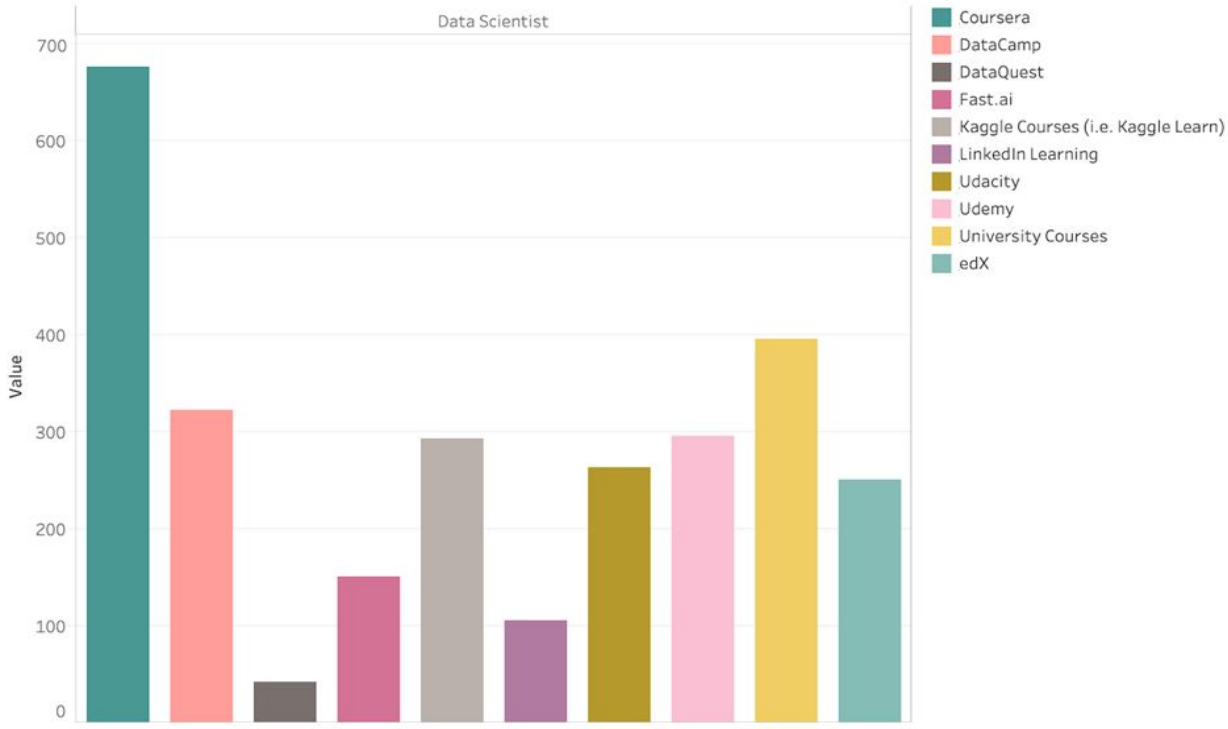


Figure A2: Coursera Popularity amongst Data Scientists.

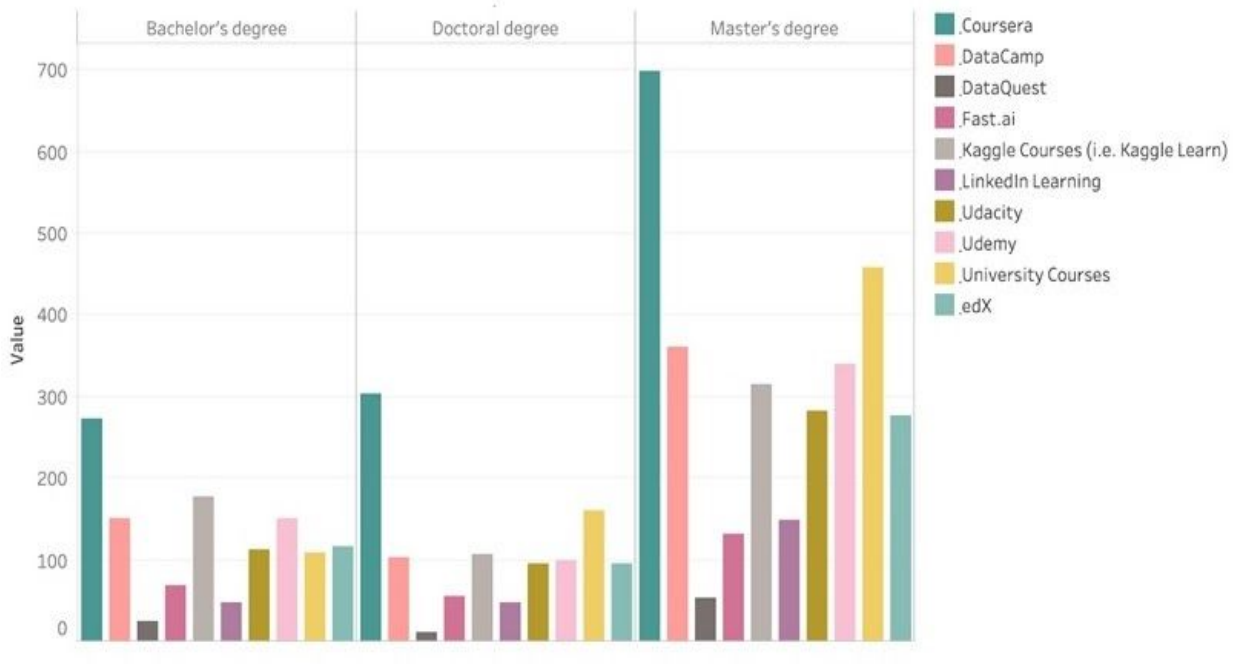


Figure A3: Coursera Popularity regardless of Education Levels.

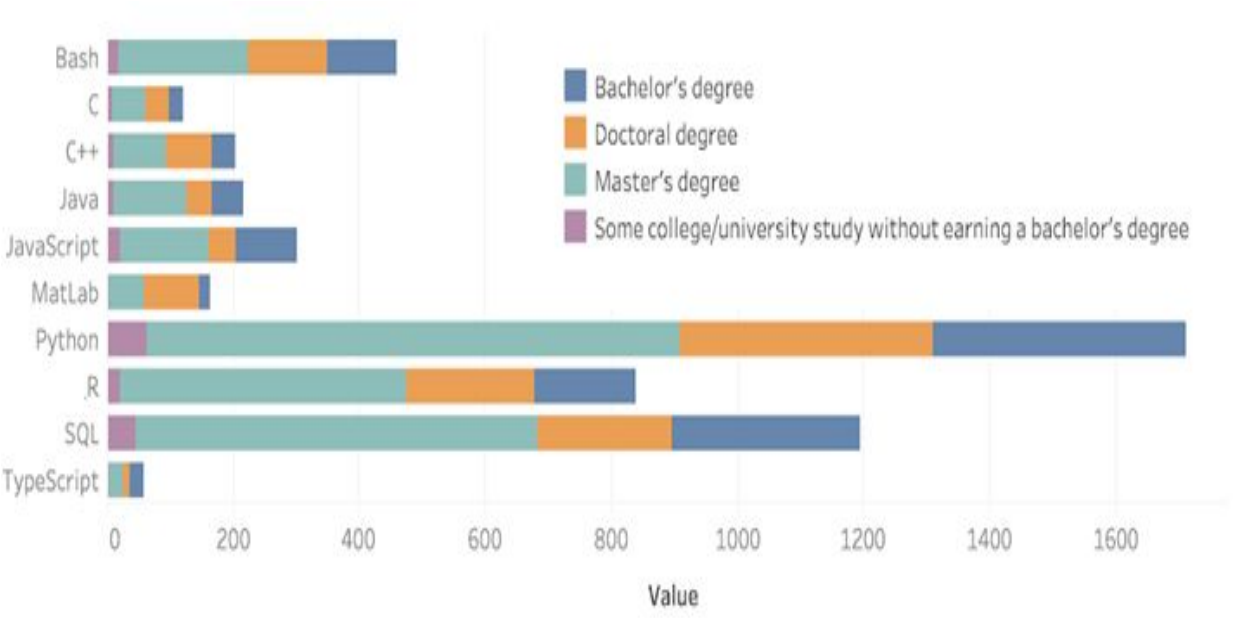


Figure A4: Python & SQL are most popular.

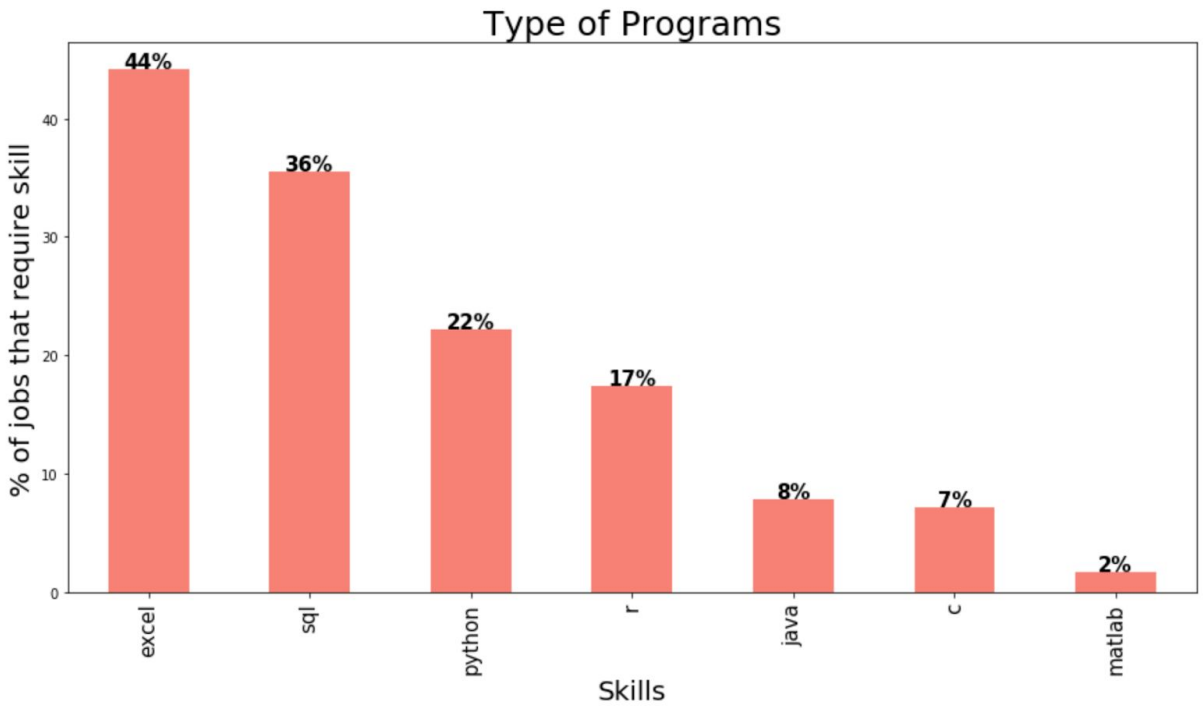


Figure A5: The most sought after programming knowledge by employers

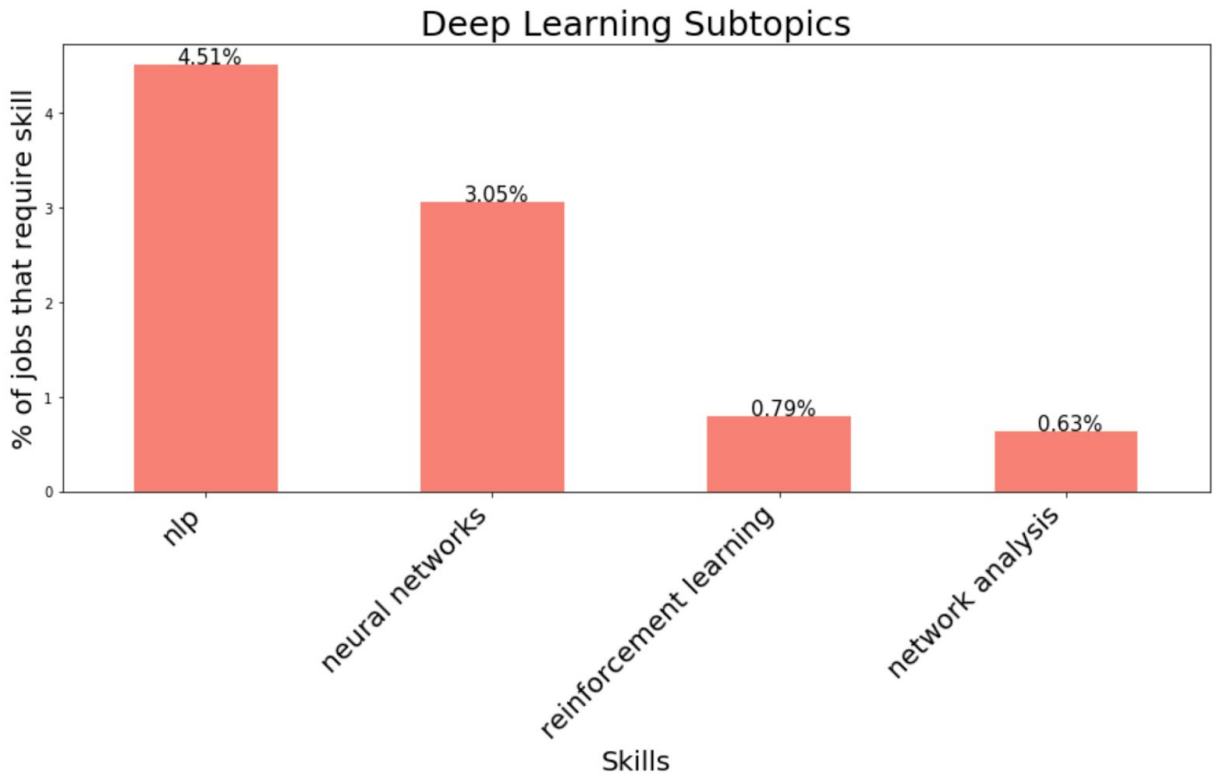


Figure A6: Most sought after deep learning techniques by employers

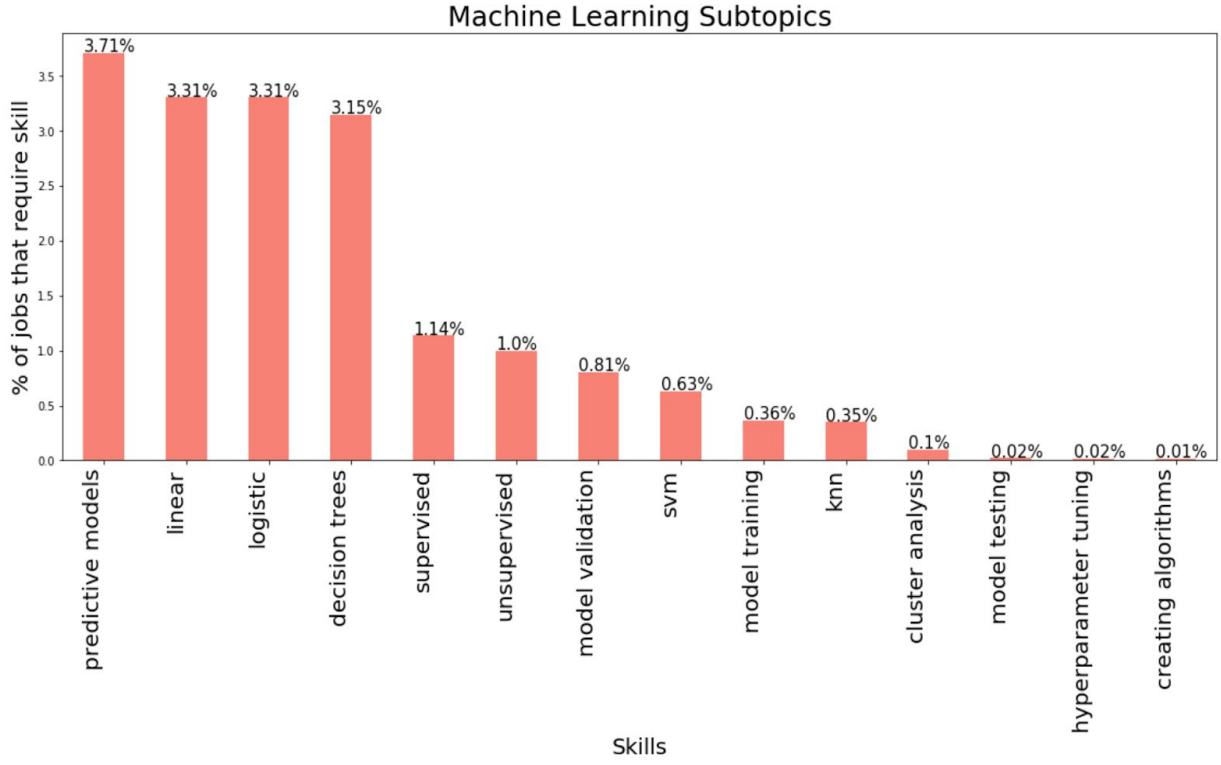


Figure A7: Most sought after machine learning techniques by employers .

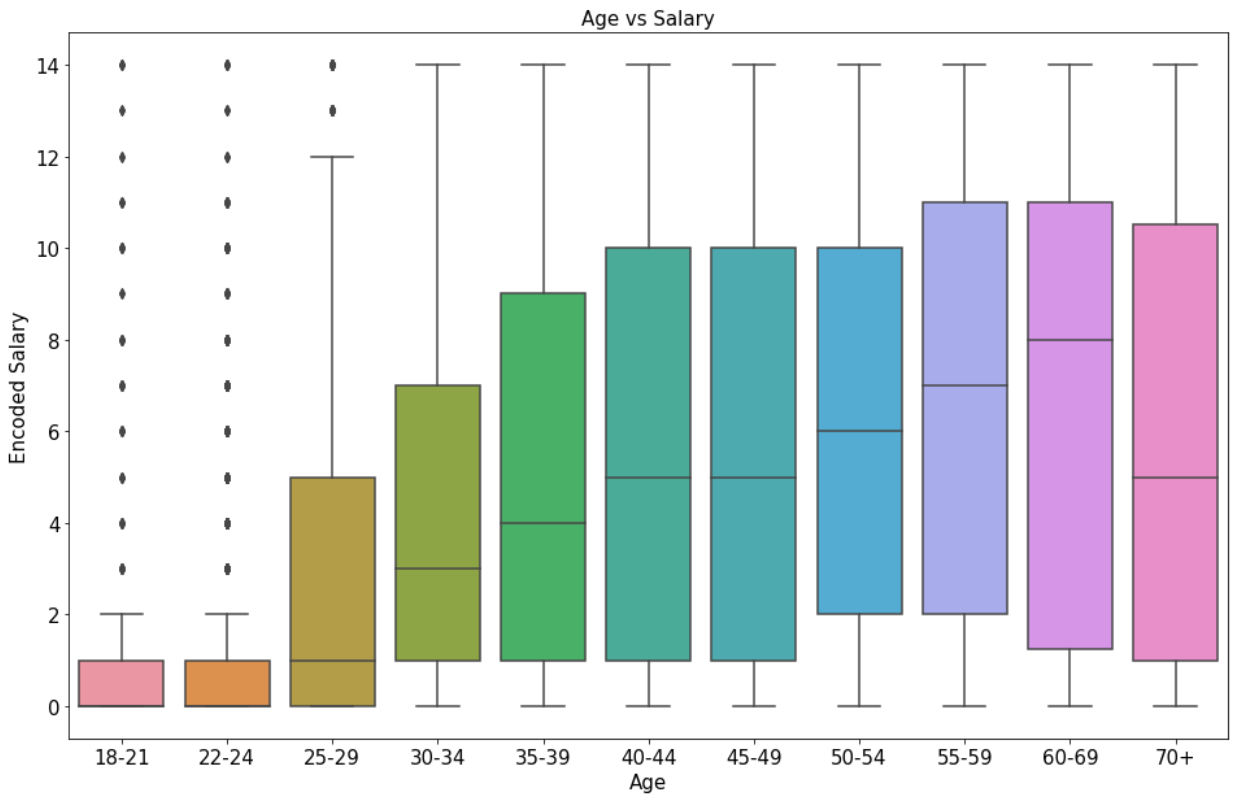


Figure A8: Age vs Salary.

Data Analytics Counselor

Details of Data Analytics Counselor Application:

1. Resume is inserted in pdf format
2. Resume is cleaned to enhance searching
3. Important skills and their weights are obtained
4. Skills you should develop:
    1. The skills are searched in resume, and if missing the missing skills are stored
    2. Missing skills are displayed to user in order of importance
5. Matching Job Postings
    1. The job posting and their required skills from scrapping are obtained
    2. Each job posting is compared with resume and matching weight is assigned to job posting according to skill weights
    3. Five jobs with highest weights are suggested
    4. Details of the job like, company name, link, salary and location are displayed

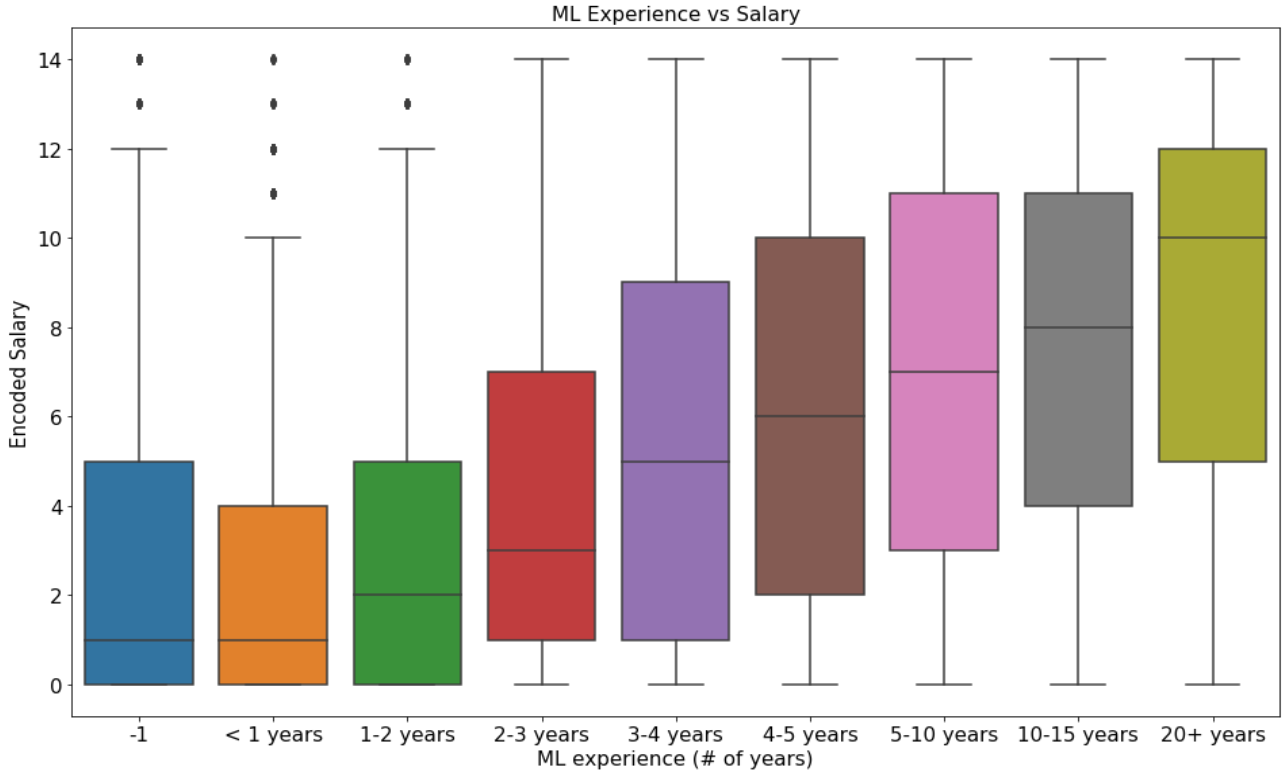


Figure A8: ML experience vs Salary.

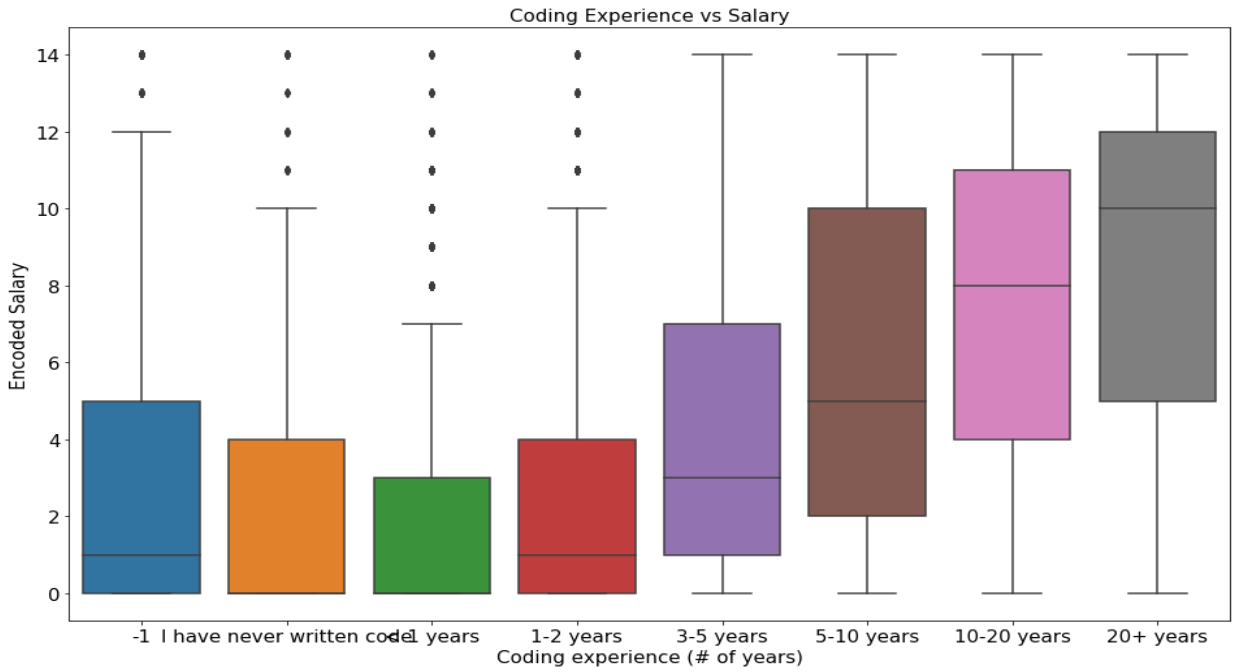


Figure A9: Coding Experience vs Salary.

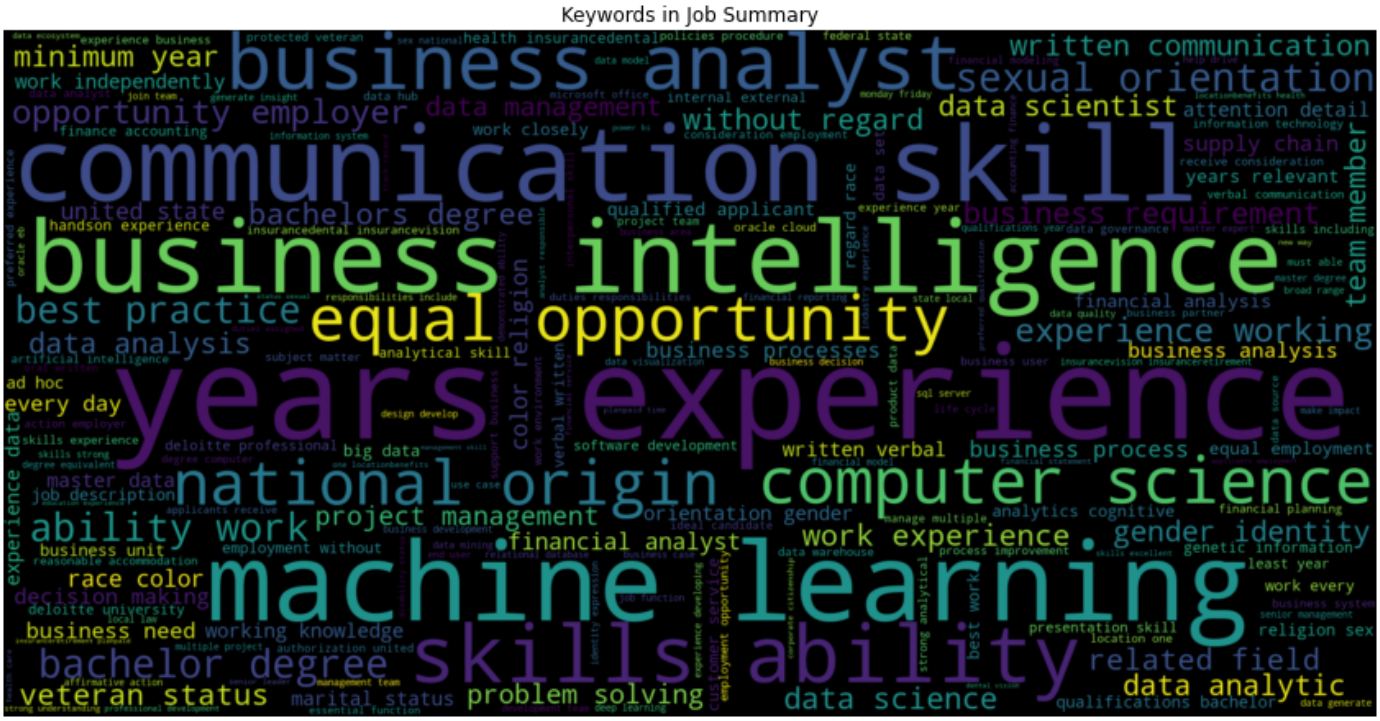


Fig A10: Keywords in cleaned Job summary.

Fig A11: Technical Keywords in Job summary.