MSBANA DATA SCIENCE JOB RECOMMENDER

By –

Vaishali Pawar

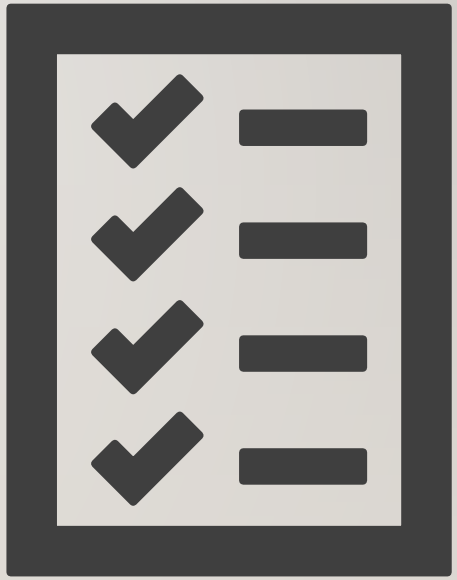
Rohan Sharma

Manoj Tomar

Aditya Nawal

AGENDA

- Problem Definition – What are we trying to solve
- Data description and Feature Engineering
- Solution Approach
  - Personalized job Finder
  - Skills Recommender
- Walkthrough of the Power BI Dashboard
- Overall Insights and Recommendations



Problem Definition | MSBANA Data Science Job Recommender Would help students secure Data Scientist Jobs

Current State

- MS Business Analytics students find it difficult to secure Data Scientist jobs upon completion of the program
- The OBAIS Department is of the opinion that - It would be helpful for the students to know beforehand about the tools needed, availability of opportunities and job titles, and the salaries associated with the roles
- Understanding what would be best for students to pursue and what are the best suited job profiles, is important in their job search

Gap

Identify the "best" Data Science jobs and have an understanding about the skills required for these jobs

Key Question

What are the jobs in Data Science domain and what is the average skillset required for these jobs

Desired Future State

- The Power BI dashboard helps the MSBANA students to find Data Science jobs based on their current skillset, and recommends additional courses that can be helpful to land desired jobs
- Creating a similarity matrix on existing and auxiliary features gives best suited jobs for a candidate and suggests similar jobs he/she can apply to
- Feature engineering and adding auxiliary variables to the existing dataset helps us generate insights on the popular tools and technologies required to acquire Data Science jobs

Solution Approach | Why not Clustering or Regression?

- The Data Scientist Salary data had ~765 records, out of which we found that there were duplicates in the data. The Final row count after data cleaning was 469
- For any statistical or machine learning algorithm to give a good accuracy, we need a good amount of training data. With a smaller dataset, the recommendations that the model would suggest would be biased
- Moreover, the problem statement background is to help MSBANA students secure Data Scientist jobs. Our solution approach is more focused on what the students should or should not do, to land Data Scientist jobs
- While developing this tool and analyzing the problem statement, we found very interesting insights that would we helpful for us in our job hunt
- This tool is scalable and with more historical data, we would be able to give personalized recommendations for all the MSBANA students to help them streamline their job search process

Data Description | Engineering Features and adding auxiliary variables helped in finding meaningful insights

Data Scientist Salary

Historic Dataset

Job Description	Salary	Python	…..
Data Scientist	80000	1	……
Sr. Data Scientist	95000	1	…….

+

Feature Engineering

Job Description

Experience Required - YOE	Type (Service /Product Based)	Job Description – Stop Words	Job in Headquarters
3	Service	Data, experience	FALSE
5	Product	……	TRUE

+

Auxiliary Variables

Relevant Data Scrape

Average Expenditure in the city	Net Saving	H1 B Approvals	Average Salary of State
	1169	23	43526
	1674	34	----

Final Dataset

Job description	Salary	Python	Years of Experience	Type (Service/Product based)	Job Description – Stop Words	Median Rental Income	Average Salary of state	Standardized Salary	FICO Score	Job in Headquarter
Data Scientist	80000	1	3	Service	Data, experience	1169	93156		717	FALSE
Sr. Data Scientist	95000	1	5	Product	……	1674	89139		734	TRUE

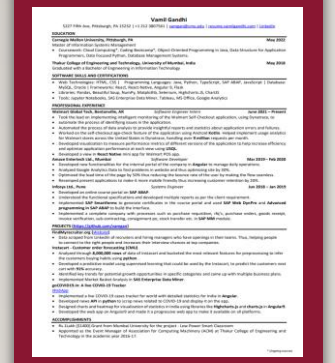
Metadata →

Data Description | Definitions of Engineered and Auxiliary Variables

Auxiliary/ Engineered Features	Definition
Experience Required -YOE	Years of Experience required for the job
Type (Service /Product Based)	Is the Company Service based or Product based
Job Description – Stop Words	Relevant information after removing Connecting words
Job in Headquarters	If Job Location is the Headquarter of the Company
Net Saving	Average salary of job – Average cost of Living for the city
H1 Approvals	Number of H1 approvals for the company – Important field to consider for the International Students
Average Salary of State	Average salary of the state

Solution Approach | Our Approach helps students land best jobs and also improve their skillset

Sample Resume



What are the best 10 jobs you should apply, based on current skillset?

- The student uploads his/her resume to the app
- The algorithm scans the resume and extracts important tools and technologies mentioned in the resume
- Then the extracted tools and technologies are matched with the records in the database
- Based on the matches found, the algorithm suggests 10 'best' jobs that the student can apply for

What courses you should take, to boost your skillset?

- The 2nd part of the algorithm, compares the extracted tools and technologies from the resume with the popular tools and technologies required for Data Science jobs
- The tools and technologies that are missing in the resume are highlighted
- The algorithm also suggests the BANA and Data Camp courses the student should take to boost their skills which increases their chances to land Data Science jobs

Solution Approach | Step 1 - How do we suggest 10 best jobs you should apply ?

The Student uploads his/her resume

We extract the Description and remove stop (connecting words)

The Extracted Key skills using **Natural Language Processing** library in python to find candidates possessed skills

Then we vectorize each document to convert them into numeric vectors using **TF-IDF**

We find cosine similarity of each document with the existing records

Then sort top 10 records based on Cosine Similarity value

Term Frequency (TF) is the relative frequency of a word in a document and is given as (term instances/total instances). Inverse Document Frequency (IDF) is the relative count of documents containing the term is given as log(number of documents/documents with term) The overall importance of each word to the documents in which they appear is equal to TF * IDF

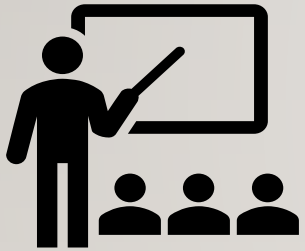
Solution Approach | Step 2 - What courses you should take, to boost your skillset?

The Student uploads his/her resume

We extract the Description and remove stop (connecting words)

The Extracted Key skills using **Natural Language Processing** library in python to find candidates possessed skills

Same as Step 1



Compare skills required in the best jobs recommended and skills from the Student's resume

Create a flag –

1. **Required** – if the skill is missing in Resume
2. **Not Required** – If the skill is in Resume but not required for the job
3. **Satisfied** – If the skill is required for the job and is present in the student's resume

Recommend courses from BANA curriculum

Insights | Streamlining Insights for groups would help students find best suited jobs

Product Companies that need 0 to 3 YOE

Words to use: Data, Learning, Team, Support, Scientist
Best Possible Net Saving PA: 156K
Best Job : Lead DE @Credit Sesame
Top3 Skills: Python, SQL, Excel

Product Companies that need 0 to 3 YOE & Sponsor H1B Visas

Words to use: Data, Learning, Complex, Skills, Team
Best Possible Net Saving PA: 124K
Best Jobs: Data Scientist @ Pfizer
Top3 Skills: Python, SQL, Excel

Product Companies that need >5 YOE

Words to use: Data, Learning, Team, Support, Scientist
Best Possible Net Saving PA: 220K
Best Job : Director Data Science @ Liberty Mutual
Top3 Skills: Python, SQL, Excel

Product Companies that need >5 YOE & Sponsor H1B Visas

Words to use: Data, Development, Solutions, Science
Best Possible Net Saving PA: 220K
Best Jobs: Director Data Science @ Liberty Mutual
Top3 Skills: Python, SQL, Excel

** All net savings' are annual and in USD($)*

Insights | Streamlining Insights for groups would help students find best suited jobs

All Companies that need 0 to 3 YOE

Words to use: Data, Learning, Team, Support, Models
Best Possible Net Saving PA: 156K
Best Job : Lead DE @Credit Sesame
Top3 Skills: Python, SQL, Excel

All Companies that need 0 to 3 YOE & Sponsor H1B Visas

Words to use: Data, Learning, Statistical, Support
Best Possible Net Saving PA: 147K
Best Jobs: Lead Data Scientist @ Visa
Top3 Skills: Python, SQL, Excel

All Companies that need >5 YOE

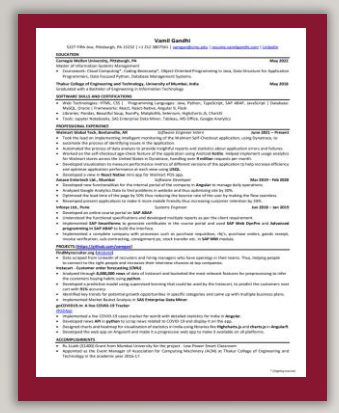
Words to use: Data, Learning, Team, Support, Scientist
Best Possible Net Saving PA: $ 220K
Best Job : Director Data Science @ Liberty Mutual
Top3 Skills: Python, SQL, Excel

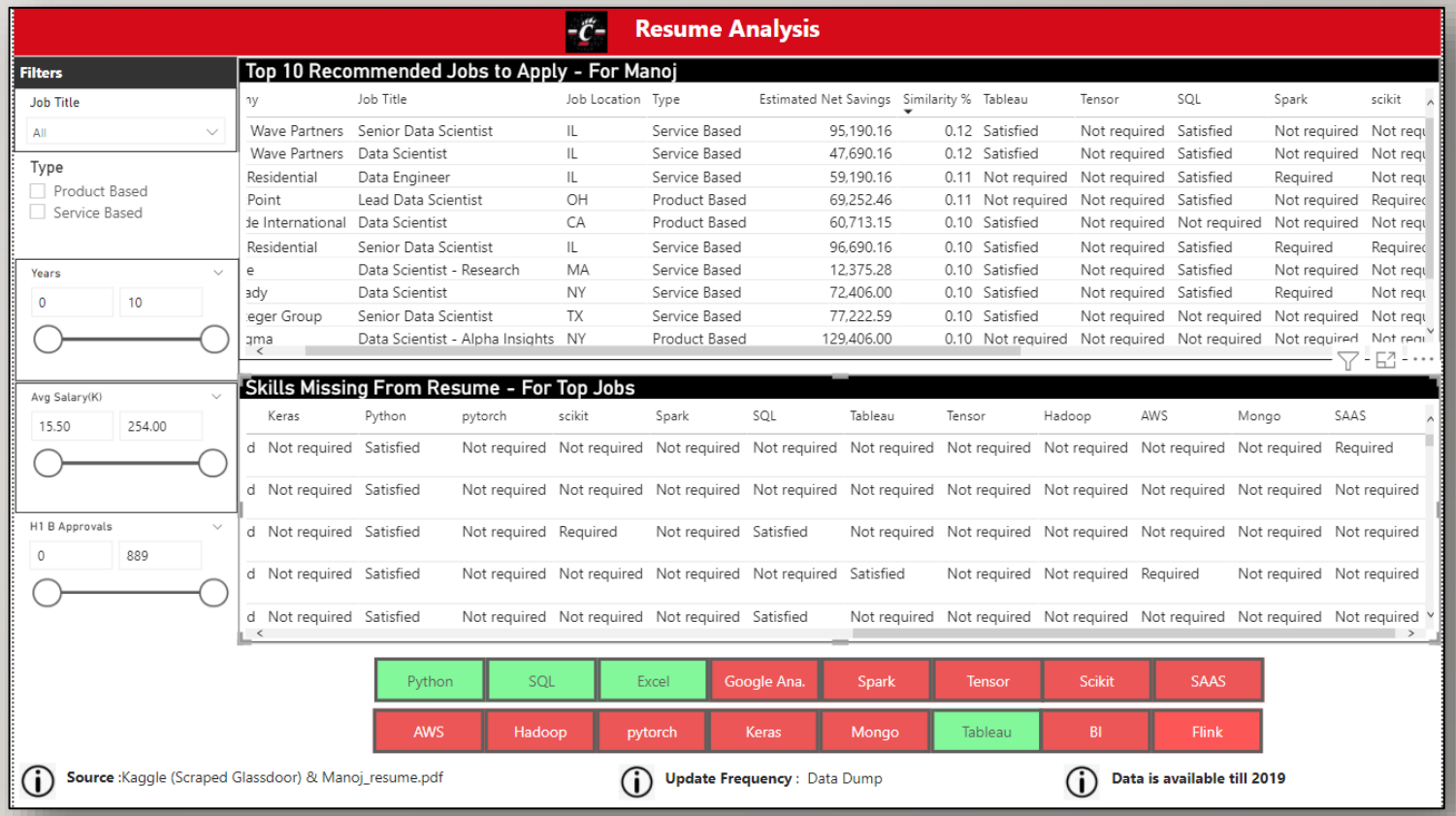
All Companies that need >5 YOE & Sponsor H1B Visas

Words to use: Data, Learning, Complex, Skills, Team
Best Possible Net Saving PA: 220K
Director Data Science @ Liberty Mutual
Top3 Skills: Python, SQL, Excel

** All net savings' are annual and in USD($)*

Insights | Manoj needs to work on Scikit, AWS and SAAS to support his candidature for top 5 jobs





❖ Hovering over the buttons shows the target BANA course / Data Camp course to gain particular skill, users can click the buttons in the dashboard to be redirected to the course links.

References for Auxiliary Variables

- H1B data : US CIS Gov. website: https://www.uscis.gov/
- https://wise.com/us/blog/cost-of-living-in-the-usa
- https://advisorsmith.com/data/coli/
- https://meric.mo.gov/data/cost-living-data-series