IBM 322: Analytics in Managerial Decision-Making
Group Project

Problem Statement

Predicting best suited Job profiles for a resume:

Job seekers encounter difficulty in finding suitable roles that align with their skills. An intelligent system is needed to analyze resumes and offer personalized job recommendations, streamlining the job search process and empowering individuals to make informed career choices.

Team Members

Anupriya Dey	21114014
Gujar Neha Pankaj	21114039
Komal Gupta	21114052
Naqiyah Kagzi	21114064
Rohan Kalra	21114083

MOTIVATION

The job market is slow these days, so to land the job of your dreams you need to be fast, efficient and competitive. It is crucial for a person to waste as little time as possible because of the tight deadlines and the sheer volume of applicants might make you miss your opportunity in a matter of seconds.

Most companies only allow a limited number of applicants, and this number is reached within a few hours if not minutes from the time of posting. Also, to stand out among the applicants your resume needs to match the industry requirements. It needs to align with the skills that the job description lists as most companies filter out the candidates by using some ML models on your resume to determine if your experiences and qualifications match.
Our approach helps you find the type of listings that your profile matches and what are the most common types of job listings. Also, how you can tailor your resume to include only the details that are relevant to the most common jobs in your dream industry. A concise and precise resume might be what sets you apart.
We look for specific keywords that your resume has, to find the job that suits you the most. We also study a sample of job listings on LinkedIn that were posted in 2023 to identify the major industry types and broadly how many industries can these listings be classified into. Then we identify the most common keywords that these industries look for.

METHODOLOGY

Resume text analysis and modelling for category specification

Dataset used: Resume dataset from livecareer.com (link)

Resume text analysis

We received this dataset from Kaggle containing both HTML and string representations of resume documents. Our focus during the data preprocessing stage was on the removal of extraneous entries to ensure retention of clean dataset.

We categorized the entire set of resume profiles into 24 distinct categories, encompassing roles such as Human Resources, Aviation, and Sales Management etc. which were provided in the dataset.

Utilizing the HTML format of resumes, we extracted key components, including 'Summary,' 'Experience,' 'Education,' ‘years of experience’ and 'Skills,' using “BeautifulSoup” library. Note that each resume had different html format, so extracting these fields were challenging and although for maximum resumes, fields were extracted, we defined null for rest of them to avoid interference with model formulation. We then applied a rigorous text cleaning process to refine the extracted data by tokenization, removal of stopwords and lemmatization. This comprehensive preprocessing approach establishes a solid foundation for subsequent analysis and modelling within our project.

Modelling for Category Specification

In the pursuit of identifying the most fitting model for our project, we conducted training and testing across four distinct models:

- ### LOGISTIC REGRESSION

  The dataset was split into 80-20 training and testing sets. A TF-IDF transformation was applied to the training and testing data, and a logistic regressor was employed for modelling.

- ### Support Vector Machine (SVM)

  The dataset was divided into an 80-20 ratio for training and testing. The TF-IDF matrix of the training set data was fitted onto a Support Vector Machine.

- ### Multi - Layer Perceptron (MLP)

  The dataset underwent an 80-20 split for training and testing. The TF-IDF matrix of the training set data was fitted onto an MLP with 3 layers of 10, 5, and 10 neurons, respectively along with activation function ‘relu’. Due to overfitting, this model could not achieve equivalent accuracy on test dataset as with train dataset.

- ### Cosine Similarity

  The dataset was divided into an 80-20 ratio for training and testing. For the training dataset, 24 documents were created for each of the 24 categories and concatenated together. The DTM of the resulting document was calculated. For the testing dataset, cosine similarity was calculated with each of the 14 documents, and the resume was assigned the category with the highest cosine similarity.

We observe that, SVM provides the best accuracy for category prediction. Following the category prediction, the cleaned data is subsequently processed for job profile searches, marking a crucial phase in our analytical pipeline.

Job Profile modelling for category specification and best match recommendation

Dataset used: LinkedIn dataset (link)

Data Pre-processing

Our methodology commenced with the pre-processing of the LinkedIn dataset, which was loaded into a Pandas DataFrame for convenient manipulation. During this pre-processing stage, we handled missing data by substituting empty or null values with "Not Specified" in the pertinent columns. This ensures uniformity in the dataset and provides a clear indication of where specific information is unavailable.

Job Type Prediction

In predicting job types from LinkedIn data, we transformed job descriptions and skills into a consolidated feature. Using TF-IDF vectorization and cosine similarity, we matched each entry with predefined job types. The job type with the highest similarity was assigned to each entry in the dataframe. This streamlined methodology enhances the dataset for subsequent analysis and categorization of diverse job roles.

Customized Job Recommendations:

We receive resume processed before to be used for classification of job profiles. Employing TF-IDF vectorization and cosine similarity, the system evaluates skill compatibility and sorts the dataset based on salary information. To ensure clarity, duplicate entries are removed, and the unique sorted dataset consisting of top ten most suitable jobs is presented in a new CSV file ‘best_predicted_jobs’, offering users tailored job recommendations that align with their skills and preferences.

Top 10 Highest-Paying Industries:

An examination of industry-specific median salaries highlights key insights into the compensation landscape. Grouping data by industry names, we identified and visualized the top 10 industries with the highest median salaries. This analysis provides a succinct overview of salary distributions across sectors, offering valuable benchmarks for industry-specific compensation considerations.

Clustering for Job Role Exploration:

In pursuit of meaningful clusters within the dataset, a combined textual feature was created by merging job descriptions and skills. Leveraging the TF-IDF transformation and TruncatedSVD for dimensionality reduction, the textual data was condensed into a 2D space. Subsequently, KMeans and GMM clustering were applied, resulting in a plot that aids in exploring distinct job roles within the dataset.

- ### K-Means Clustering

  Applying K-Means clustering to the reduced 2D space derived from job descriptions and skills, distinct job role clusters were identified. Visualizing this in a 2D projection, the plot color-codes data points according to their cluster labels, providing a succinct representation of the identified job role groupings. This visualization serves as a valuable tool for comprehending the inherent structure and diversity present within the dataset, aiding in the interpretation of job role patterns and facilitating targeted analyses.

- ### Gaussian Mixture Model Clustering

  Utilizing GMM, various models were fitted to the 2D space derived from job descriptions and skills, assessing different cluster configurations. The optimal number of clusters, determined through the Bayesian Information Criterion (BIC), was found. A concise 2D projection visualization, color-coded by GMM cluster labels, provides a clear representation of job role groupings, aiding in the interpretation of inherent structures within the dataset.

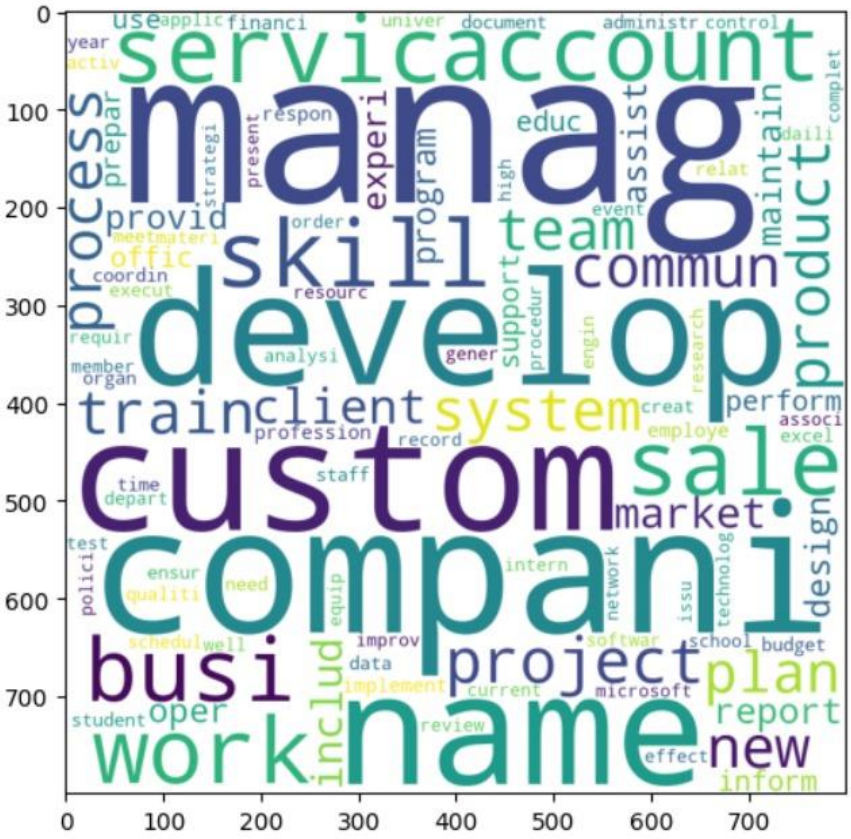
Model Evaluation Metrics:

To gauge the effectiveness of clustering models the Davies-Bouldin Index, measuring cluster compactness and separation, was calculated for both KMeans and GMM. Lower Davies-Bouldin Indices signify more robust clustering. These metrics offer concise insights into the performance and suitability of the clustering models for uncovering patterns within the job dataset.

OBSERVATIONS AND FINDINGS

WORD CLOUD OF CLEAN RESUME

Resume in form of string cleaned by tokenization and lemmatization is displayed as a word cloud to observe the most frequently occurring tokens.



RESULTS OF MODELING ON RESUMES

- Logistic Regression

  Accuracy on training set: 80.69%
  Accuracy on testing set: 66.87%

- Support Vector Machine (SVM)

  Accuracy on training set: 91.87%
  Accuracy on testing set: 62.40%

- Multiple Neuron Perceptron (MLP)

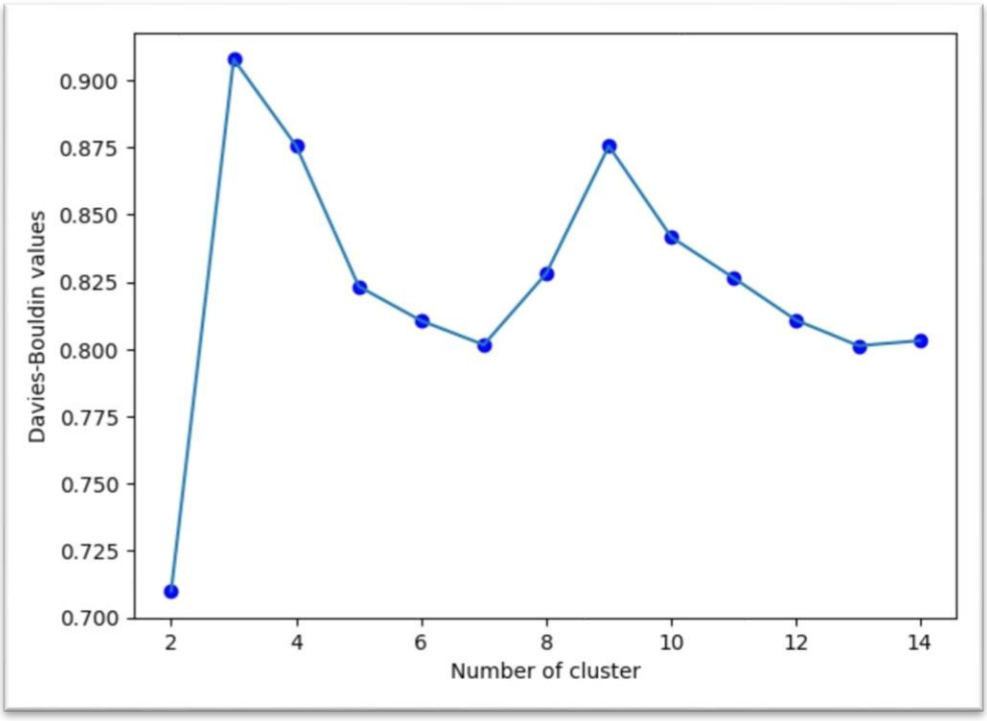
  Accuracy on training set: 98.53%
  Accuracy on testing set: 23.78%

- Cosine Similarity

  Accuracy on training set: 50.40%
  Accuracy on testing set: 58.90%

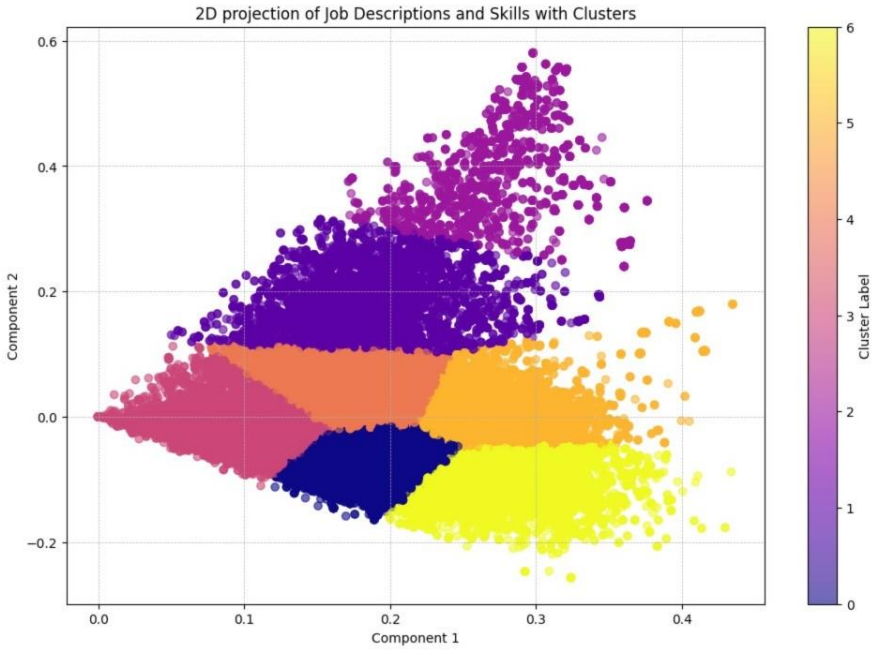
RESULTS OF CLUSTERING

The analysis of our dataset involved the application of both K-Means and Gaussian Mixture Model (GMM) clustering. Utilizing methodologies such as the elbow method and Davies-Bouldin (DB) index, we determined that K-Means exhibited optimal clustering when set to 7 clusters.

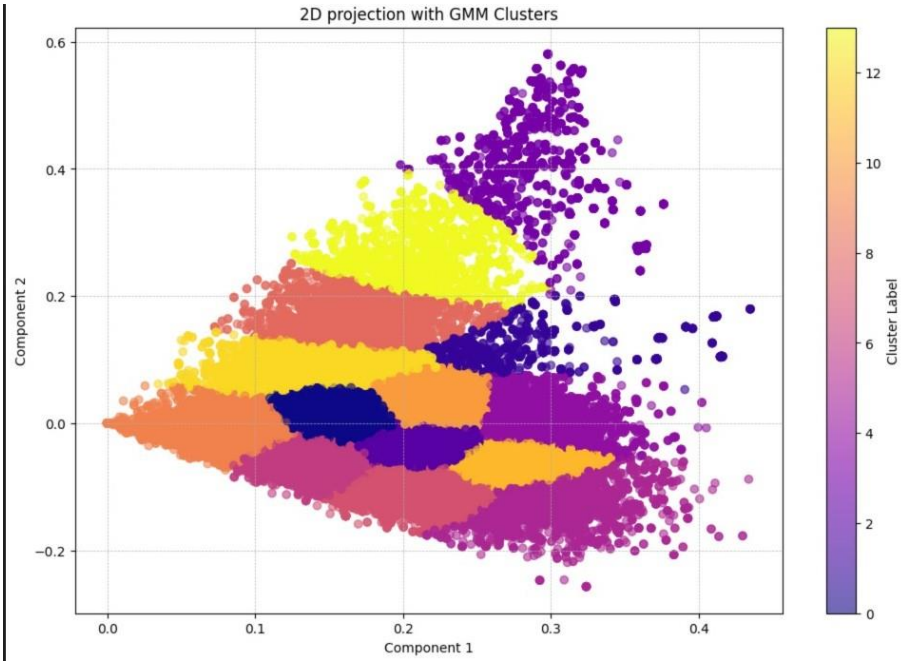


On the other hand, GMM resulted in 14 clusters using the Bayesian Information Criterion (BIC) method. It was observed that the Davies-Bouldin indices were slightly higher for the GMM algorithm. This implies that K-Means achieved a more favorable balance between compactness and separation of clusters compared to GMM in our dataset.

RESULTS WITH K-MEANS:

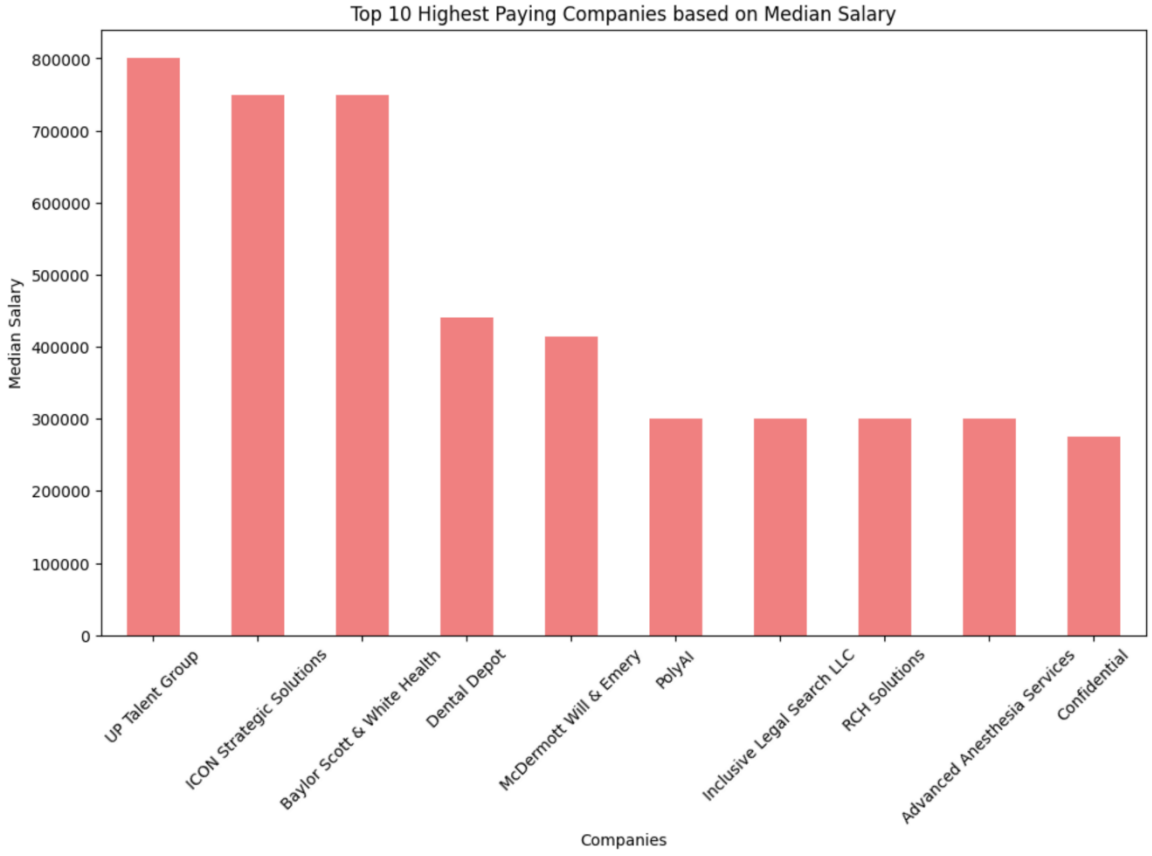


RESULTS WITH GMM:



SALARIES LAYOUT

We found out that the median salaries were maximum for the company UP Talent Group followed closely by ICON Strategic Solutions and Baylor Scott & White Health and then Dental Depot.



FINAL PREDICTIONS

After all the analysis is done and the model is set, we are finally ready to predict the best job profiles for a resume

PREDICTIONS:

A	B	C	D	E	F	G	H	I	J	K	L	M	N	
b_id	company_id	title	descriptic	max_	med_salar	min_s	pay_perio	formatted	location	applies	original_listed_ti	remote_al	views	job_
3693066662	212201	ServiceNow Platform Architect	Title:		250000		YEARLY	Full-time	New York,	4	1.69274E+12	Unknown	26	http
3699410802	11130470	HR Business Partner, Go To Market and S	About		240000		YEARLY	Full-time	San Franci	10	1.69275E+12	Unknown	120	http
3697385977	5353	Program Advisor	Summar		136908		YEARLY	Part-time	Washingtc	12	1.69274E+12	Unknown	86	http
3699420827	5353	Supervisory Operations Coordinator	Summar		132368		YEARLY	Part-time	Washingtc	0	1.69285E+12	Unknown	1	http
3693052455	117983	Human Resources Manager (Bilingual)	Jacuzzi		105000		YEARLY	Full-time	Roselle, IL	22	1.69274E+12	Unknown	91	http

INSTRUCTIONS TO US

- Download all necessary libraries
- Run `observation_on_job_postings.ipynb` to see analysis on current job trends extracted from LinkedIn dataset
- Replace corresponding values of Skills, Category, Years of Experience, Education, Summary in `predict_job_type()` function, otherwise you can use other entries from the given `Resume.csv` dataset for prediction.
- Run `resume_analysis.ipynb`
- Read the results from `best_predicted_jobs.csv`

CONCLUSION

As users input their resumes, the system extracts key information such as job category, skills, education, summary, experience and years of experience. Leveraging a comprehensive dataset of companies and job postings, our algorithm then recommends the best-suited job, streamlining the job search process and providing tailored career guidance for individuals. This solution aims to enhance the efficiency and effectiveness of job matching, empowering candidates to discover opportunities aligned with their professional strengths and aspirations.