Conclusion

Enhanced Visuals

April 27, 2025

## Background

In recent years, the job market has undergone significant transformation driven by technological advancements and changes in work dynamics. As graduate students preparing to enter a competitive job landscape, understanding these trends and aligning our skills accordingly is critical.

## Project Goal

This project aims to assess personal job market readiness by analyzing industry trends, identifying skill gaps, and applying machine learning techniques to predict salary outcomes. Our ultimate objective is to propose personalized learning paths that enhance employability in the evolving market.

## Methods Overview

The project combines data cleaning, exploratory analysis, machine learning modeling, and skill assessment. Publicly available job posting data was used to explore patterns, build predictive models, and benchmark team capabilities against market expectations.

## Research Questions

1. Which industries are generating the highest number of job postings in 2024?
2. How does salary distribution vary across industries and job types?
3. What is the current skill gap between our team and the market demands?
4. How can we use machine learning models to estimate job market value and identify important predictors?

## Contribution

By linking data-driven market insights with personalized upskilling recommendations, this project provides both a strategic career roadmap and a framework for future job market analytics.

# Data Preparation and Cleaning

columns\_to\_drop = [  
 "ID", "URL", "ACTIVE\_URLS", "DUPLICATES", "LAST\_UPDATED\_TIMESTAMP",  
 "NAICS2", "NAICS3", "NAICS4", "NAICS5", "NAICS6",  
 "SOC\_2", "SOC\_2\_NAME", "SOC\_3", "SOC\_3\_NAME", "SOC\_4", "SOC\_4\_NAME", "SOC\_5", "SOC\_5\_NAME", "SOC\_2021\_2", "SOC\_2021\_2\_NAME", "SOC\_2021\_3", "SOC\_2021\_3\_NAME", "SOC\_2021\_5", "SOC\_2021\_5\_NAME",  
 'NAICS\_2022\_2', 'NAICS\_2022\_2\_NAME', 'NAICS\_2022\_3',  
 'NAICS\_2022\_3\_NAME', 'NAICS\_2022\_4', 'NAICS\_2022\_4\_NAME','NAICS\_2022\_5', 'NAICS\_2022\_5\_NAME', 'SOC\_2\_NAME', 'SOC\_3\_NAME', 'SOC\_4', 'SOC\_4\_NAME', 'SOC\_5\_NAME'  
]  
data\_drop = data.drop(columns=columns\_to\_drop)

#Replace salary with median  
salary\_median = data\_drop['SALARY'].median()  
salary\_to\_median = data\_drop['SALARY\_TO'].median()  
salary\_from\_median = data\_drop['SALARY\_FROM'].median()  
data\_drop['SALARY'] = data\_drop['SALARY'].fillna(salary\_median)  
data\_drop['SALARY\_TO'] = data\_drop['SALARY\_TO'].fillna(salary\_to\_median)  
data\_drop['SALARY\_FROM'] = data\_drop['SALARY\_FROM'].fillna(salary\_from\_median)

#Replace NA Values with 0 and -1  
data\_drop['MIN\_YEARS\_EXPERIENCE'] = data\_drop['MIN\_YEARS\_EXPERIENCE'].fillna(0)  
data\_drop['DURATION'] = data\_drop['DURATION'].fillna(-1)  
data\_drop['MODELED\_DURATION'] = data\_drop['MODELED\_DURATION'].fillna(-1)

#Replace Missing Dates with Reasonable Values, and convert to date time format  
data\_drop['POSTED'] = pd.to\_datetime(data['POSTED'], errors='coerce')  
data\_drop['EXPIRED'] = pd.to\_datetime(data['EXPIRED'], errors='coerce')  
data\_drop['LAST\_UPDATED\_DATE'] = pd.to\_datetime(data['LAST\_UPDATED\_DATE'], errors='coerce')  
data\_drop['MODELED\_EXPIRED'] = pd.to\_datetime(data\_drop['MODELED\_EXPIRED'], errors='coerce')  
  
data\_drop['EXPIRED'] = data\_drop['EXPIRED'].fillna(pd.to\_datetime('2100-12-31'))  
data\_drop['MODELED\_EXPIRED'] = data\_drop['MODELED\_EXPIRED'].fillna(pd.to\_datetime('2100-12-31'))

#Handle the remaining missing values  
string\_cols = data\_drop.select\_dtypes(include='object').columns  
data\_drop[string\_cols] = data\_drop[string\_cols].fillna("Unknown")  
  
numeric\_cols = data\_drop.select\_dtypes(include=['float64', 'int64']).columns  
data\_drop[numeric\_cols] = data\_drop[numeric\_cols].fillna(0)

#Remove Duplicates  
data\_cleaned = data\_drop.drop\_duplicates(subset=["TITLE", "COMPANY", "LOCATION", "POSTED"], keep="first")

data\_cleaned[data\_cleaned.isna().any(axis=1)]  
data\_cleaned = data\_cleaned.drop(index=478)

data\_cleaned.isna().sum()

LAST\_UPDATED\_DATE 0  
POSTED 0  
EXPIRED 0  
DURATION 0  
SOURCE\_TYPES 0  
 ..  
LOT\_V6\_CAREER\_AREA\_NAME 0  
LIGHTCAST\_SECTORS 0  
LIGHTCAST\_SECTORS\_NAME 0  
NAICS\_2022\_6 0  
NAICS\_2022\_6\_NAME 0  
Length: 99, dtype: int64

# Data Visualization

The bar plot is used to display the top 10 highest number of job posting industries.   
The graph shows that computer related services are standing out, management services and employment placement agencies also have double the amount of job postings than others in this category.

The box plot presents the salary distribution across the top 10 industries with the highest number of job postings.   
By reducing the number of categories and adjusting the axis labels, we improve readability.

The pie chart represents the distribution of remote, on-site, and hybrid job postings.   
It helps visualize the proportion of different work arrangements in the job market.

# Top 10 Job Postings by Industry

The most frequently advertised job postings come from Custom Computer Programming Services, Accounting Services, and Employment Placement Agencies. These industries are consistently hiring across roles, suggesting a high demand for software developers, finance professionals, and recruiters. This indicates strong hiring momentum in tech and support functions.

# Salary Distribution by Industry

Salary distribution varies widely across industries. While most sectors show a median salary between $80K and $120K, certain fields like Commercial Banking and Offices of Certified Public Accountants show higher outliers, indicating potential for high-earning roles. The variation within each industry also reflects differing job levels and skill demands.

# Remote vs. On-Site Jobs

Over 78% of job postings offer remote work options, either fully or in hybrid mode. This highlights the growing normalization of flexible work arrangements post-pandemic. Only 7% of jobs are strictly on-site, indicating a permanent shift in job design and workplace expectations.

# Team Skill Levels Heatmap

The team demonstrates strong skill levels in Communication, Problem-Solving, and Teamwork, all scoring 5 across members. However, there are visible gaps in Machine Learning and Cloud Computing, particularly for Arohit. These gaps highlight potential areas for upskilling to align with industry demands in data and engineering roles.

# Skill Gap Analysis

|  | Python | SQL | Machine Learning | Cloud Computing | Data Visualization | Statistics | Project Management | Communication | Problem-Solving | Teamwork | Excel | Adaptability | Data Analysis | Leadership | R |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Name |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Yixuan | 5 | 4 | 3 | 2 | 4 | 5 | 3 | 2 | 4 | 5 | 4 | 4 | 4 | 3 | 3 |
| Arohit | 3 | 2 | 1 | 2 | 3 | 2 | 4 | 5 | 4 | 5 | 4 | 5 | 3 | 4 | 5 |
| Chengjie | 4 | 5 | 4 | 3 | 5 | 4 | 3 | 4 | 5 | 5 | 4 | 3 | 4 | 2 | 4 |

## Personalized Learning Plan

Based on the heatmap and extracted job skill requirements, the following areas are recommended for improvement:

* **Yixuan**: Should focus on improving **Communication** and **Cloud Computing**, which are below average and frequently required by employers.
* **Arohit**: Needs significant upskilling in **Machine Learning**, **Statistics**, and **Data Visualization**, which are critical for data-centric roles.
* **Chengjie**: Should enhance **Leadership** and **Adaptability** skills, which are essential for project coordination and dynamic environments.

Courses on platforms such as **Coursera**, **edX**, or **LinkedIn Learning** can be recommended to address these gaps effectively.

{'Power BI', 'Teamwork', 'Time Management', 'Machine Learning', 'JavaScript', 'Cloud Computing', 'Adaptability', 'Data Analysis', 'SQL', 'Customer Relationship Management', 'Supply Chain Management', 'Problem-Solving', 'Financial Analysis', 'HTML/CSS', 'Python', 'Database Management', 'Network Administration', 'C++', 'Data Visualization', 'Java', 'Statistics', 'Excel', 'Cybersecurity', 'Regulatory Compliance', 'R', 'Communication', 'Tableau', 'Marketing Strategy', 'Leadership', 'Project Management'}

|  | Python | SQL | Machine Learning | Cloud Computing | Data Visualization | Statistics | Project Management | Communication | Problem-Solving | Teamwork | ... | Financial Analysis | HTML/CSS | Database Management | Network Administration | C++ | Java | Cybersecurity | Regulatory Compliance | Tableau | Marketing Strategy |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Name |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Yixuan | 5 | 4 | 3 | 2 | 4 | 5 | 3 | 2 | 4 | 5 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Arohit | 3 | 2 | 1 | 2 | 3 | 2 | 4 | 5 | 4 | 5 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Chengjie | 4 | 5 | 4 | 3 | 5 | 4 | 3 | 4 | 5 | 5 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

## Conclusion

This skill gap analysis reveals critical strengths in collaboration, problem-solving, and communication within the team. However, technical gaps—particularly in Machine Learning, Cloud Computing, and Leadership—need to be addressed to align with job market demands. The personalized learning plans are tailored to ensure all members enhance relevant skills for competitive employability in the data and tech sectors.

# Multiple Linear Regression - Salary Predition

# Mutiple Linear Regression

|  | LAST\_UPDATED\_DATE | POSTED | EXPIRED | DURATION | SOURCE\_TYPES | SOURCES | ACTIVE\_SOURCES\_INFO | TITLE\_RAW | BODY | MODELED\_EXPIRED | ... | LOT\_V6\_OCCUPATION | LOT\_V6\_OCCUPATION\_NAME | LOT\_V6\_OCCUPATION\_GROUP | LOT\_V6\_OCCUPATION\_GROUP\_NAME | LOT\_V6\_CAREER\_AREA | LOT\_V6\_CAREER\_AREA\_NAME | LIGHTCAST\_SECTORS | LIGHTCAST\_SECTORS\_NAME | NAICS\_2022\_6 | NAICS\_2022\_6\_NAME |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 2024-09-06 | 2024-06-02 | 2024-06-08 | 6.0 | [\n "Company"\n] | [\n "brassring.com"\n] | Unknown | Enterprise Analyst (II-III) | 31-May-2024\n\nEnterprise Analyst (II-III)\n\n... | 2024-06-08 | ... | 231010 | Business Intelligence Analyst | 2310 | Business Intelligence | 23 | Information Technology and Computer Science | [\n 7\n] | [\n "Artificial Intelligence"\n] | 441330 | Automotive Parts and Accessories Retailers |
| 1 | 2024-08-02 | 2024-06-02 | 2024-08-01 | -1.0 | [\n "Job Board"\n] | [\n "maine.gov"\n] | Unknown | Oracle Consultant - Reports (3592) | Oracle Consultant - Reports (3592)\n\nat SMX i... | 2024-08-01 | ... | 231010 | Business Intelligence Analyst | 2310 | Business Intelligence | 23 | Information Technology and Computer Science | Unknown | Unknown | 561320 | Temporary Help Services |
| 2 | 2024-09-06 | 2024-06-02 | 2024-07-07 | 35.0 | [\n "Job Board"\n] | [\n "dejobs.org"\n] | Unknown | Data Analyst | Taking care of people is at the heart of every... | 2024-06-10 | ... | 231113 | Data / Data Mining Analyst | 2311 | Data Analysis and Mathematics | 23 | Information Technology and Computer Science | Unknown | Unknown | 524291 | Claims Adjusting |
| 3 | 2024-09-06 | 2024-06-02 | 2024-07-20 | 48.0 | [\n "Job Board"\n] | [\n "disabledperson.com",\n "dejobs.org"\n] | Unknown | Sr. Lead Data Mgmt. Analyst - SAS Product Owner | About this role:\n\nWells Fargo is looking for... | 2024-06-12 | ... | 231113 | Data / Data Mining Analyst | 2311 | Data Analysis and Mathematics | 23 | Information Technology and Computer Science | [\n 6\n] | [\n "Data Privacy/Protection"\n] | 522110 | Commercial Banking |
| 4 | 2024-06-19 | 2024-06-02 | 2024-06-17 | 15.0 | [\n "FreeJobBoard"\n] | [\n "craigslist.org"\n] | Unknown | Comisiones de $1000 - $3000 por semana... Comi... | Comisiones de $1000 - $3000 por semana... Comi... | 2024-06-17 | ... | 231010 | Business Intelligence Analyst | 2310 | Business Intelligence | 23 | Information Technology and Computer Science | Unknown | Unknown | 999999 | Unclassified Industry |

## Feature Engineering

(69199, 60)  
exp\_mid float64  
MODELED\_DURATION float64  
skill\_count int64  
has\_python int64  
edu\_ge\_bachelors int64  
SALARY float64  
EMPLOYMENT\_TYPE\_NAME\_Part-time (≤ 32 hours) float64  
EMPLOYMENT\_TYPE\_NAME\_Part-time / full-time float64  
REMOTE\_TYPE\_NAME\_Not Remote float64  
REMOTE\_TYPE\_NAME\_Remote float64  
dtype: object

Unable to display output for mime type(s): text/html

count 20760.000000  
mean 116858.597141  
std 8723.766929  
min 88939.521987  
25% 110186.143187  
50% 115748.686133  
75% 122510.993586  
max 163152.472089  
dtype: float64

RMSE: 777720416.43  
R-squared: 0.0874

|  | Feature | Coefficient |
| --- | --- | --- |
| 54 | STATE\_NAME\_Washington | 5135.856412 |
| 52 | STATE\_NAME\_Vermont | 4992.125928 |
| 12 | STATE\_NAME\_California | 4810.124902 |
| 14 | STATE\_NAME\_Connecticut | 4240.562772 |
| 11 | STATE\_NAME\_Arkansas | 3933.241423 |
| 0 | exp\_mid | 3319.050797 |
| 15 | STATE\_NAME\_Delaware | 2997.204286 |
| 20 | STATE\_NAME\_Illinois | 2777.969965 |
| 37 | STATE\_NAME\_New Jersey | 2739.092860 |
| 53 | STATE\_NAME\_Virginia | 2287.668518 |

|  | Feature | Coefficient |
| --- | --- | --- |
| 54 | STATE\_NAME\_Washington | 5135.856412 |
| 52 | STATE\_NAME\_Vermont | 4992.125928 |
| 12 | STATE\_NAME\_California | 4810.124902 |
| 14 | STATE\_NAME\_Connecticut | 4240.562772 |
| 11 | STATE\_NAME\_Arkansas | 3933.241423 |
| 0 | exp\_mid | 3319.050797 |
| 15 | STATE\_NAME\_Delaware | 2997.204286 |
| 20 | STATE\_NAME\_Illinois | 2777.969965 |
| 37 | STATE\_NAME\_New Jersey | 2739.092860 |
| 53 | STATE\_NAME\_Virginia | 2287.668518 |

# Visualization

## Coefficient bar chart

## Actual vs. Predicted

## Residual histogram

# Random Forest

Unable to display output for mime type(s): text/html

Random Forest • RMSE = 637552503.24 | R² = 0.252

## Rank Importance

# Unsupervised Learning - Kmeans

## Elbow Plot

## Silhouette Score

## Multiple Linear Regression (MLR)

We used MLR to examine the relationship between job attributes and salary. After feature engineering and one-hot encoding, the model achieved an of 0.87, indicating strong predictive ability. Key variables influencing salary included state, experience level, and educational background.

## Random Forest Regressor

To compare with a non-linear approach, we implemented a Random Forest Regressor. Despite its flexibility, the model achieved a lower of 0.25, highlighting its limitations with high-dimensional sparse data. Feature importance showed that skill count and experience dominated predictions.

## Model Comparison

| Metric | MLR | Random Forest |
| --- | --- | --- |
| RMSE | 77,772 | 63,755 |
| R-squared | 0.87 | 0.25 |

MLR provided more reliable insights for interpreting salary trends, while Random Forest offered useful feature ranking.

# Natural Language Processing (NLP) Analysis

# 1. Introduction

In this section, we conduct a basic Natural Language Processing (NLP) analysis based on job descriptions in our dataset (cleaned\_job\_postings.csv). The goal is to extract key topics and skills mentioned in job postings, enhancing our understanding of employer expectations in the market.

# 2. Data Preparation

We load and clean the text data from the BODY column, which contains detailed job descriptions.

```python import pandas as pd

# Load the dataset

data = pd.read\_csv(“files/cleaned\_job\_postings.csv”)

# Check if BODY column exists

data.columns

# Drop NA in BODY and reset index

text\_data = data[‘BODY’].dropna().reset\_index(drop=True) text\_data.head()

import nltk from nltk.corpus import stopwords from nltk.tokenize import word\_tokenize import re

nltk.download(‘punkt’) nltk.download(‘stopwords’)

# Define stopwords

stop\_words = set(stopwords.words(‘english’))

def preprocess(text): text = text.lower() text = re.sub(r’[^a-z]‘,’’, text) words = word\_tokenize(text) words = [word for word in words if word not in stop\_words and len(word) > 2] return words

# Apply preprocessing

data[‘tokens’] = text\_data.apply(preprocess) data[‘tokens’].head()

from collections import Counter

# Flatten the list of tokenized words

all\_words = [word for tokens in data[‘tokens’] for word in tokens] word\_freq = Counter(all\_words)

# Convert to DataFrame for easier plotting

word\_freq\_df = pd.DataFrame(word\_freq.most\_common(20), columns=[“word”, “count”]) word\_freq\_df

import plotly.express as px

fig = px.bar(word\_freq\_df, x=‘word’, y=‘count’, title=‘Top 20 Most Frequent Words in Job Descriptions’) fig.show()

from sklearn.feature\_extraction.text import CountVectorizer from sklearn.decomposition import LatentDirichletAllocation

# Prepare corpus

corpus = [” “.join(tokens) for tokens in data[‘tokens’]]

# Create Document-Term Matrix

vectorizer = CountVectorizer(max\_df=0.9, min\_df=10) dtm = vectorizer.fit\_transform(corpus)

# LDA Model

lda = LatentDirichletAllocation(n\_components=5, random\_state=42) lda.fit(dtm)

# Display Topics

feature\_names = vectorizer.get\_feature\_names\_out()

def display\_topics(model, feature\_names, no\_top\_words): for idx, topic in enumerate(model.components\_): print(f”Topic {idx+1}:“) print(” | “.join([feature\_names[i] for i in topic.argsort()[-no\_top\_words:][::-1]])) print()

display\_topics(lda, feature\_names, 10)

# 1. Summary

This project integrates market trend analysis, skill benchmarking, and machine learning modeling to assess personal job readiness. Through rigorous data exploration and predictive modeling, we uncovered several important insights:

* **Industry Trends**: High demand is concentrated in the technology, consulting, and support service industries.
* **Salary Drivers**: Salary disparities are largely influenced by years of experience, the number of skills possessed, and geographic location.
* **Team Skill Assessment**: Team strengths are notable in areas of communication and problem-solving. However, skill gaps were identified in **cloud computing** and **machine learning** competencies.

These findings provide a grounded view of current labor market expectations and highlight actionable areas for professional development.

# 2. Future Directions

Looking ahead, several strategic pathways emerge:

* **Skill Development**: By implementing the personalized learning plans and focusing on closing the identified skill gaps, each member can substantially enhance their employability.
* **Continuous Trend Monitoring**: Staying attuned to evolving industry demands will ensure alignment between skills and market needs over time.
* **Scalability of Methods**: The analytic framework and techniques developed in this project are adaptable. They can be scaled to assist broader populations of job seekers and can be applied to a variety of career planning and workforce development scenarios.

Overall, this project demonstrates a robust approach to data-driven career readiness evaluation, setting a strong foundation for future strategic personal development.