TEAM MEMBERS:

VINEELA KOLAGANI

SAI SREE GAJJALA

LAKSHMI PRASANNA PEDDI

PRAVALLIKA PEDDI

OPTIMIZED HR RECRUITMENT FOR BIG DATA ROLES: A DATA DRIVEN APPROACH

Project Report

Table of Contents

1. Introduction: 2
2. Background: 2
3. Problem Statement 2
4. Dataset: 3
5. Dataset Cleaning, Text, and Manipulation 4
6. Data Analysis: 4
7. Model Selection and Model Evaluation: 11
8. Results: 11
9. Conclusion and Future Recommendations: 12
10. References: 13

Figures

Figure 1:Dataset 3
Figure 2:Job Titles Distribution 5
Figure 3:Most Repeated Companies' Ads 6
Figure 4:Job Area Distribution 7
Figure 5:Distribution of Job Title based on location ads. 8
Figure 6:Distribution of Ratings (Job Developers) 9
Figure 7:Most Repeated Words 1 10
Figure 8:Most Repeated Words 2 10
Figure 9:Models Results 11

1. Introduction:

The use of the power of big data analytics is crucial in the dynamic world of recruitment, where businesses are constantly looking for highly talented individuals to promote their success. Our project, "Optimized Recruitment," sets out to innovate conventional approaches to recruiting. By delving into the vast sea of data extracted from Indeed job listings, we aim to revolutionize recruitment strategies. Our main goal is to improve business hiring processes by using advanced data analytics. The goal of our research is to help HR departments better understand and anticipate hiring trends by analyzing thousands of actual job postings.

2. Background:

In recent years, the need for experts in data-related job sectors has increased considerably. Today's data-driven society relies heavily on specialists in fields like data science, machine learning engineering, and data engineering. The traditional recruitment procedure has changed highly to adapt to the transforming nature of work. Businesses are actively looking for new methods of recruiting talent. Job boards on the web like Indeed have become rich resources for people looking for work. Through postings of available positions, they make a lot of data available, such as job titles, business information, job descriptions, as well as more. These social media platforms have rapidly become significant for obtaining job seeking intelligence.

3. Problem Statement

According to the current dynamic job marketplace, companies have a challenging task of hiring for an extensive range of roles, each with its own specific skill demands and qualifications. The one-size-fits-all approach to recruiting is no longer effective. HR departments are facing great pressure to identify the right talent quickly and effectively. When trying to find the top employees out of a large pool, traditional hiring approaches often fall short. Companies are always trying to discover where they can get the best pools of talent and what positions are crucial to their development. Insights based on data are crucial for making educated choices regarding who and where to hire.

4. Dataset:

- **Source:** Indeed, which is among the most popular job portals on the internet, is where we obtained our data. There are a huge number of job postings offered.
- **Data Collection:** This data was meticulously web scraped via the Indeed web portal, using the Indeed API. This approach guarantees that our data set is accurate and up to date when compared to the job marketplace.
- **Key Features:** Essential features like Job Titles, Locations, as well as Company Ratings are all included in the dataset. These factors are critical to our analysis.
- **Purpose:** The primary objective of this data set is to analyze trends in jobs ads. We hope that by analyzing this data, we can help HR departments make better decisions based on information and improve the quality of their hiring processes.

```
Indeed_Jobs_Data_DB_ = pd.read_csv('/Users/vineelakolagani//Documents/PDS_PROJECT_FINAL/DATASET/jobs_indeed.csv')
Indeed_Jobs_Data_DB_
```

	Unnamed: 0	Title	Company	Location	Rating	Date	Salary	Description	Links
0	0	Data Scientist	Driven Brands	Benicia, CA	2.4	PostedPosted 26 days ago	NaN	You'll be working alongside a team of eight an...	https://www.indeed.com/rc/clk?jk=74d176d595225...
1	1	Business Analyst	Sabot Consulting	Remote	NaN	PostedPosted 4 days ago	80–120 an hour	Preferred candidates will have prior experienc...	https://www.indeed.com/rc/clk?jk=f662b2efb509b...
2	2	IT Business Intelligence Developer (FT) Remote...	Ballad Health	Remote in Blountville, TN	3.0	PostedPosted 30+ days ago	NaN	Job Details Apply Save Print this job Email a…	https://www.indeed.com/rc/clk?jk=58612836c63b8...
3	3	Data Engineer	Longevity Holdings Inc.	Remote in Minneapolis-Saint Paul, MN	NaN	PostedPosted 3 days ago	90,000–110,000 a year	Incorporate core data management competencies ...	https://www.indeed.com/company/TwentyFirst/job...
4	4	Network Administrator/dba developer	WKI Kenworth	Wichita, KS 67219	NaN	EmployerActive 2 days ago	50,000–70,000 a year	The Network Administrator provides 2nd level e...	https://www.indeed.com/pagead/clk?mo=r&ad=-6NY...
...	...	...	...	...	...	...	...	...	...
1195	1195	Senior Machine Learning Engineer	HyperScience	Remote	NaN	PostedPosted 24 days ago	191,000–235,000 a year	We turn ML lab experiments into enterprise-rea...	https://www.indeed.com/rc/clk?jk=e4a29b9e718fa...
1196	1196	Data Scientist	N9 it solutions	Remote	NaN	PostedPosted 4 days ago	40–60 an hour	Experience using a variety of data mining/data	https://www.indeed.com/company/N9-IT-SOLUTIONS...

Figure 1:Dataset

5. Dataset Cleaning, Text, and Manipulation

- **Handling HTML Tags:** Job postings and other textual data typically include tags from HTML, that are used to format text for web display, when scraped from internet sources. The actual text can only be extracted if these tags are stripped away. The HTML can be parsed, and the clean text obtained with the use of parser tools or libraries such Beautiful Soup in Python.
- **Character Normalization:** Data stored as text may also include non-standard characters such as diacritics or symbols. Text that has been normalized to a consistent encoding (like ASCII) is more uniform and simpler to process.
- **Cleaning URLs:** Some URLs included in job descriptions may not be relevant to the purpose of the analysis and can be eliminated. These URLs are located and removed from the text data using regular expressions or manipulation of strings techniques.
- **Spaces and Special Characters:** Problems may arise in the analysis if the text has inconsistent spacing or inappropriate special characters. Spaces, unnecessary characters, and special symbols that don't contribute to the analysis can be cleaned up with the aid of regular expressions and other text processing techniques.

6. Data Analysis:

It is the goal of data analysis to discover and comprehend the data included within. To make intelligent decisions in the context of job postings on sites like Indeed, it is essential to derive insights about various qualities.

1. **Job Titles Distribution:** According to our findings, "Data Science" is the most popular job role in the dataset, accounting for 29.42% of postings. This highlights the significant need for data science experts in the job marketplace. In addition to Data Scientists, the industry has an insatiable demand for data-centric jobs such as Data Analysts and Data Engineers. The dataset supports the idea that businesses are increasingly looking to fill data-centric positions. Data scientists and analysts have grown increasingly important to the success of companies.

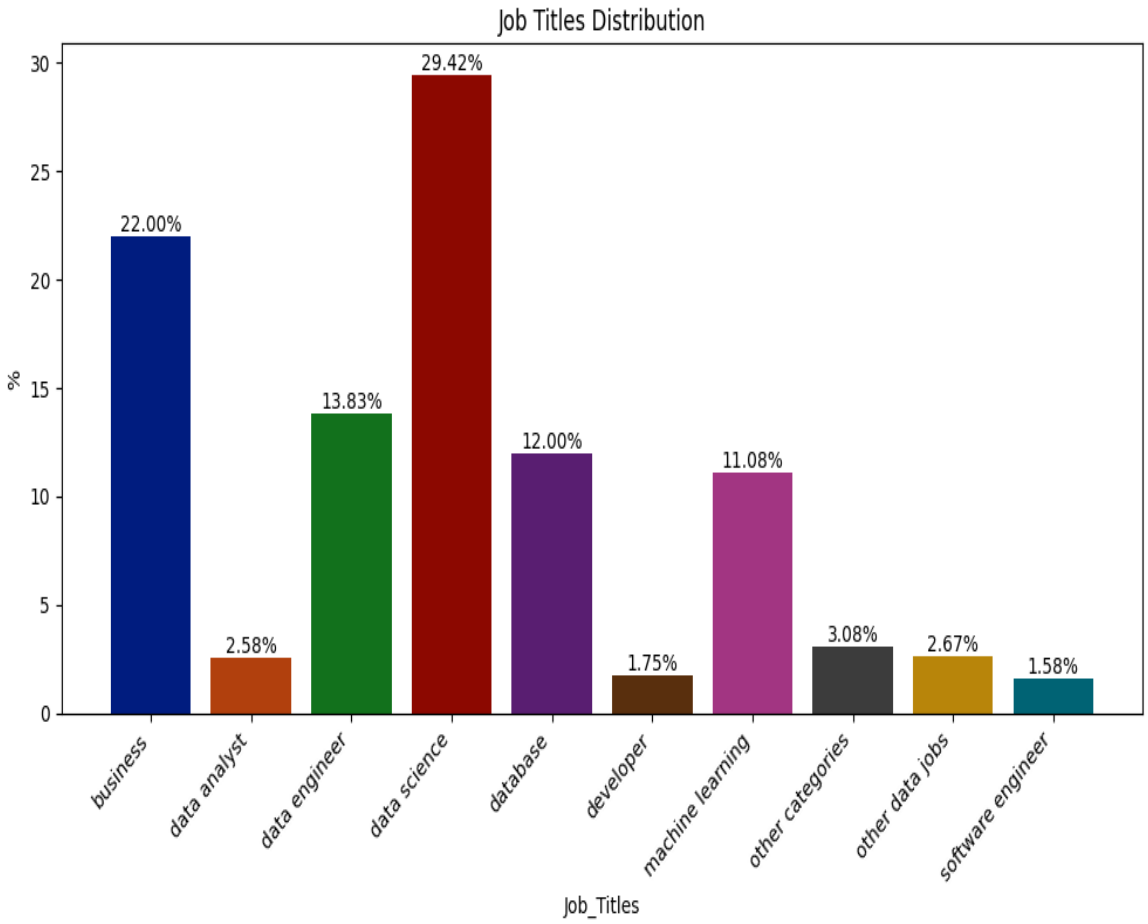

Figure 2:Job Titles Distribution

2. **Most Repeated Companies' Ads:**

Our research shows that positions in the fields of Data Science, Data Engineering, and Software Engineering are always in high demand. These jobs represent the most sought-after specializations by companies. We see that businesses are expanding their job boundaries to include a wider range of data-related roles. They have realized the value of data in many areas of their business, and their strategy reflects it. The dataset reveals a significant change in how businesses deploy their employees in their HR departments. They are hiring for positions involving both the management and utilization of data, such as Data Scientists and Software Developers. This demonstrates an all-encompassing strategy for utilizing data in working environments.

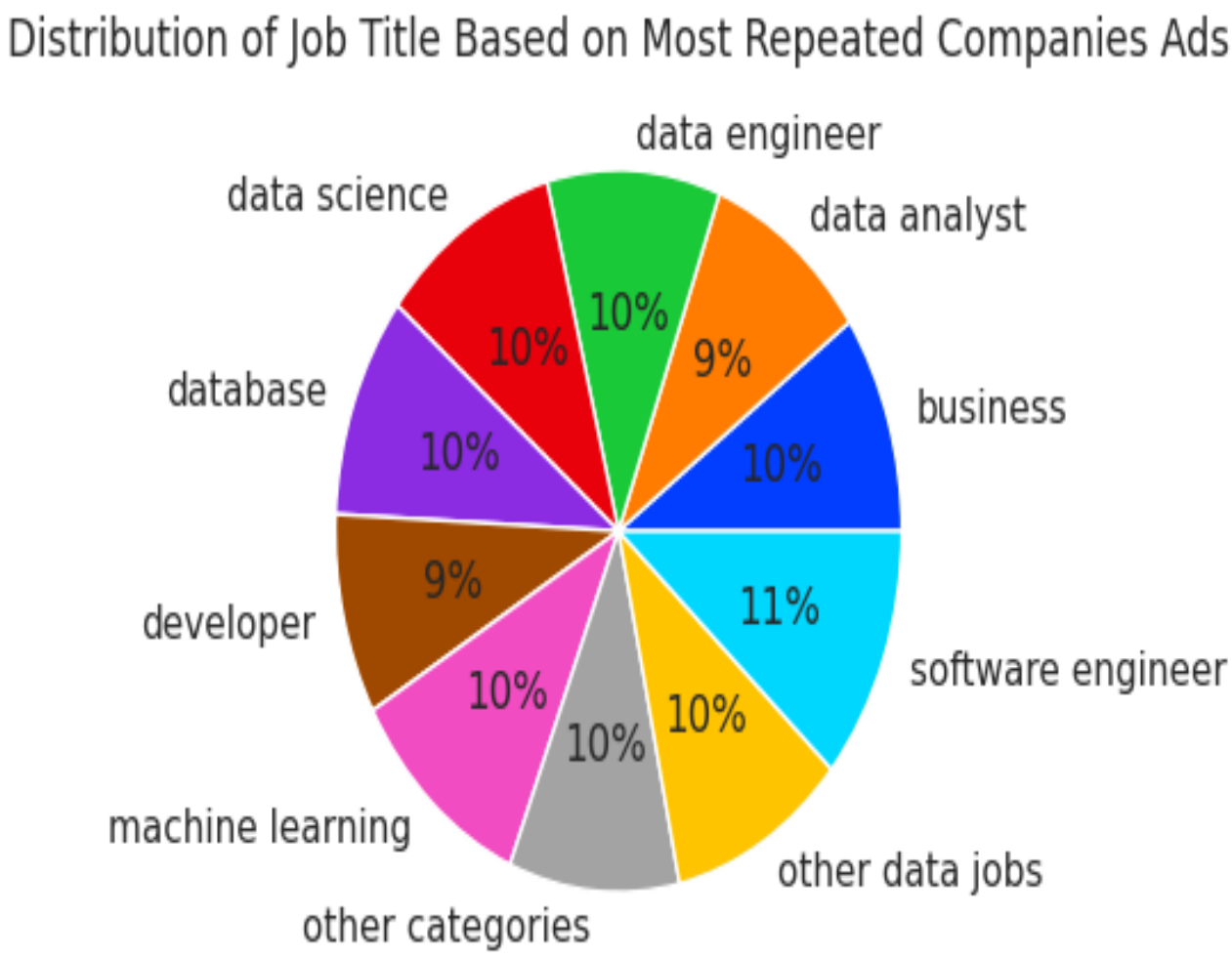

Figure 3:Most Repeated Companies' Ads

3. **Job Area Distribution:**

Our analysis finds that the most popular jobs location in this group is "Remote," showing an increasing trend to possibilities for remote employment. This connects with the worldwide shift to flexible work arrangements. Particularly, powerful tech businesses may be found in both California and New York, both of which are major employment hubs for data-related jobs. Careers in this industry can be found in a variety of states, with Texas, Illinois, among Massachusetts emerging as key locations.

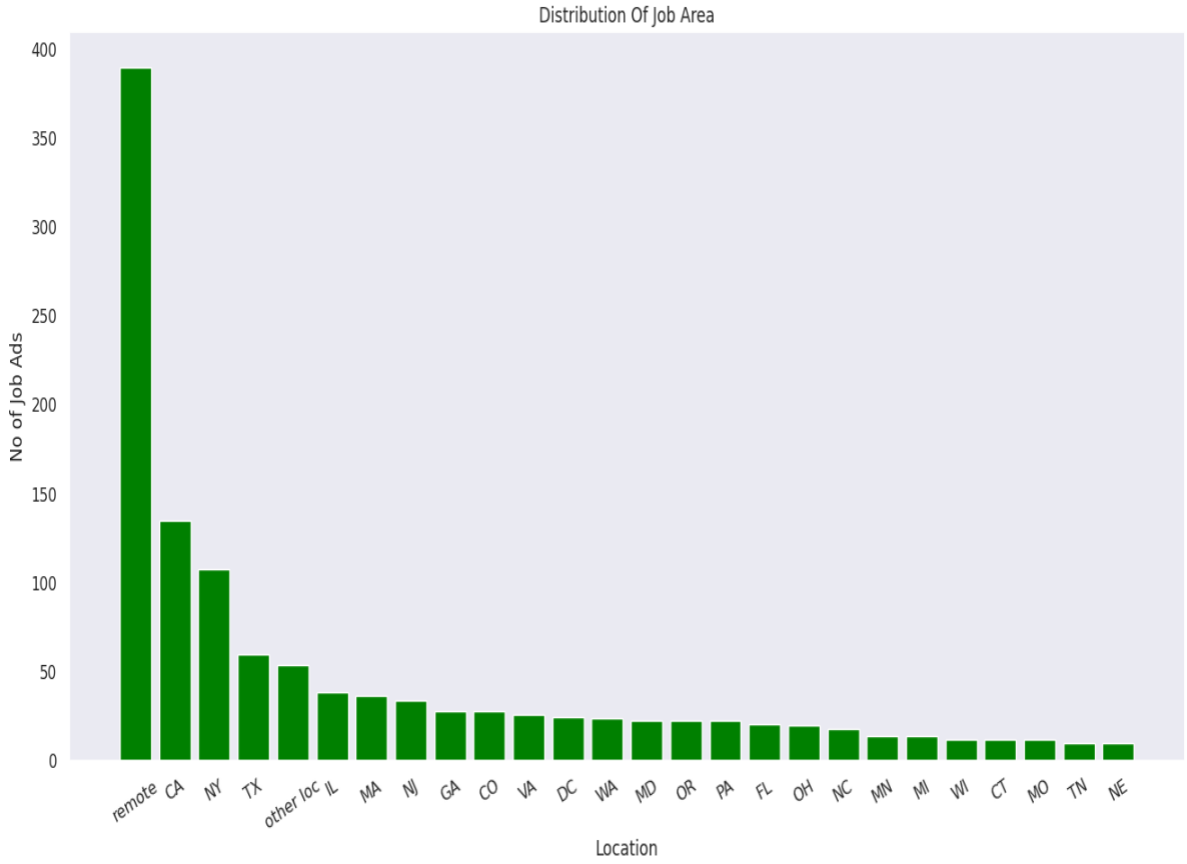

Figure 4:Job Area Distribution

4. **Job Title based on Location: T**he distribution of job titles by location is shown in the graphic, which highlights the importance of business and data-centric careers in the modern workforce. At 27%, the "Business" title dominates and denotes a flourishing corporate ecosystem in particular areas. Data experts are in high demand, as seen by the significant prevalence of "Data Engineer" and "Data Science" professions (21% and 16%, respectively), which highlights the changing nature of the labor market and its strong preference for data-driven skills and knowledge.

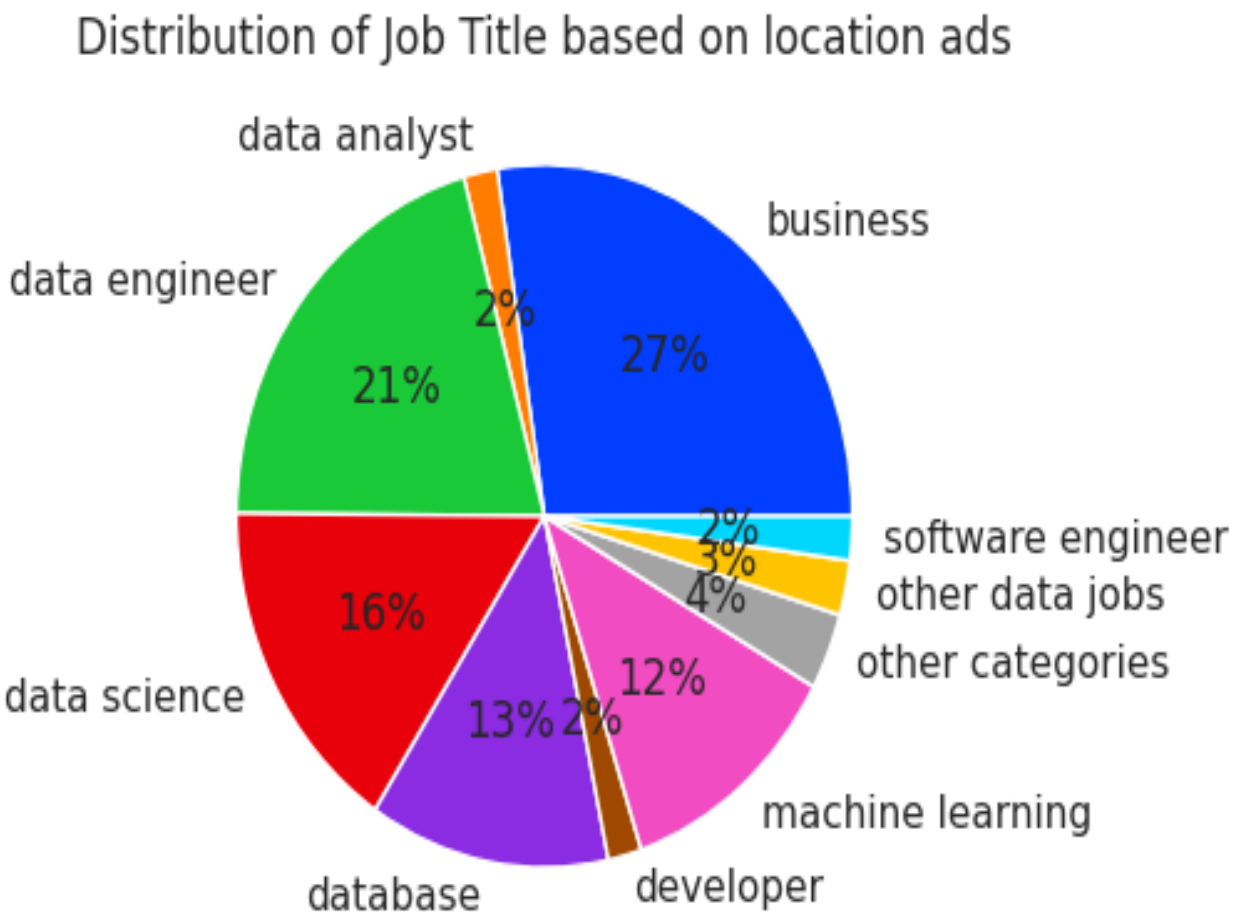

Figure 5:Distribution of Job Title based on location ads.

5. **Distribution of Ratings (Job Developers):**

   The pie chart displaying the distribution of evaluations for job developers across varied titles indicates a good equilibrium. Each section marking a spread of around 9-11% expresses an overarching satisfaction or uniform performance review across the board, independent of the specific domain. This fair rating shows a constant quality of work and professionalism by job developers in diverse professions.

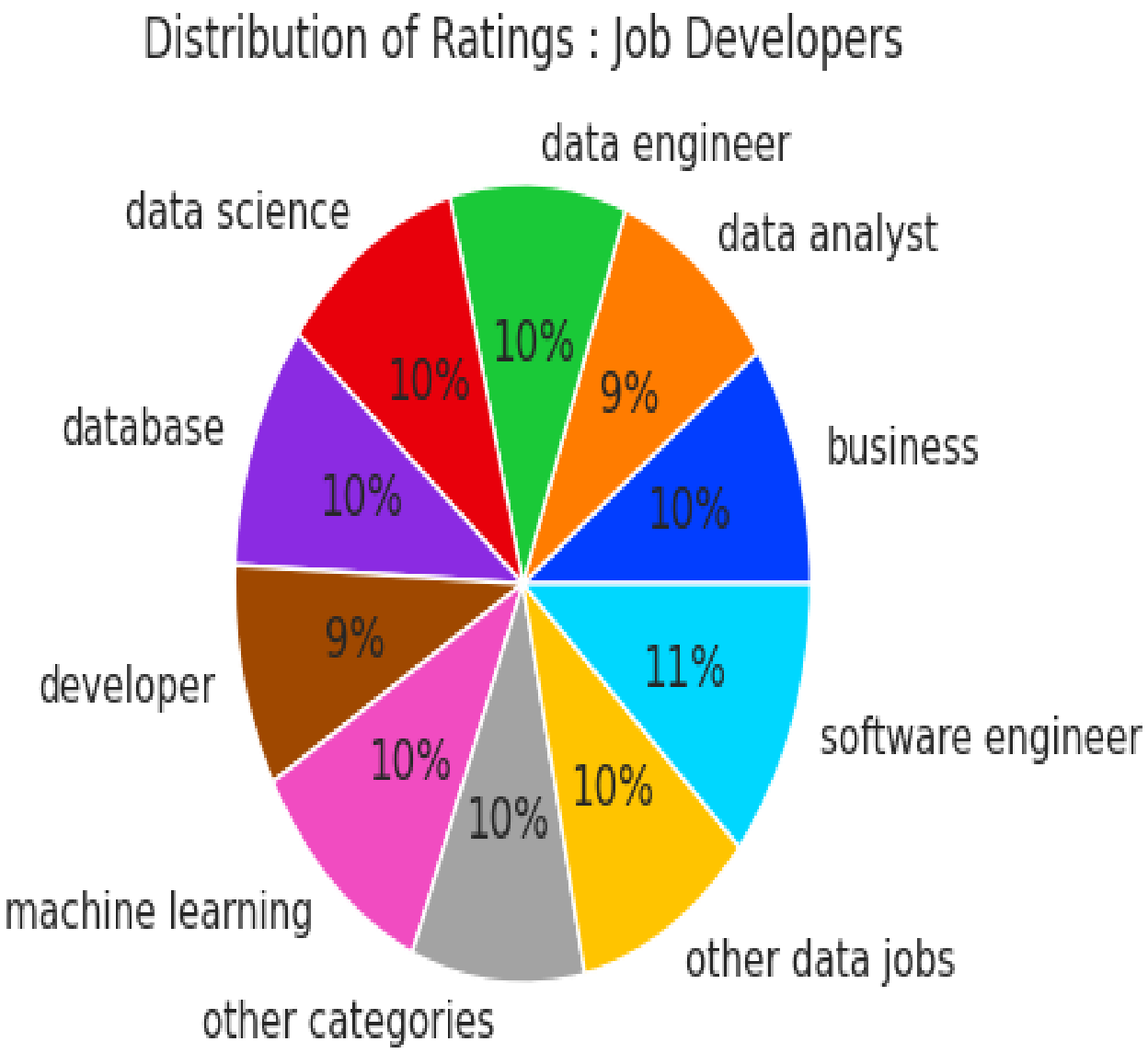

Figure 6:Distribution of Ratings (Job Developers)

6. **Most Repeated Words:**

Regarding data analysis, the word cloud highlights the present industrial zeitgeist since it is filled with frequently occurring terms from job descriptions. The prevalence of words such as "data", "experience", "business", "analysis", "learning", and "database" suggests a clear desire for experts in these fields, explaining the current and possibly growing need for such knowledge in the labor market.

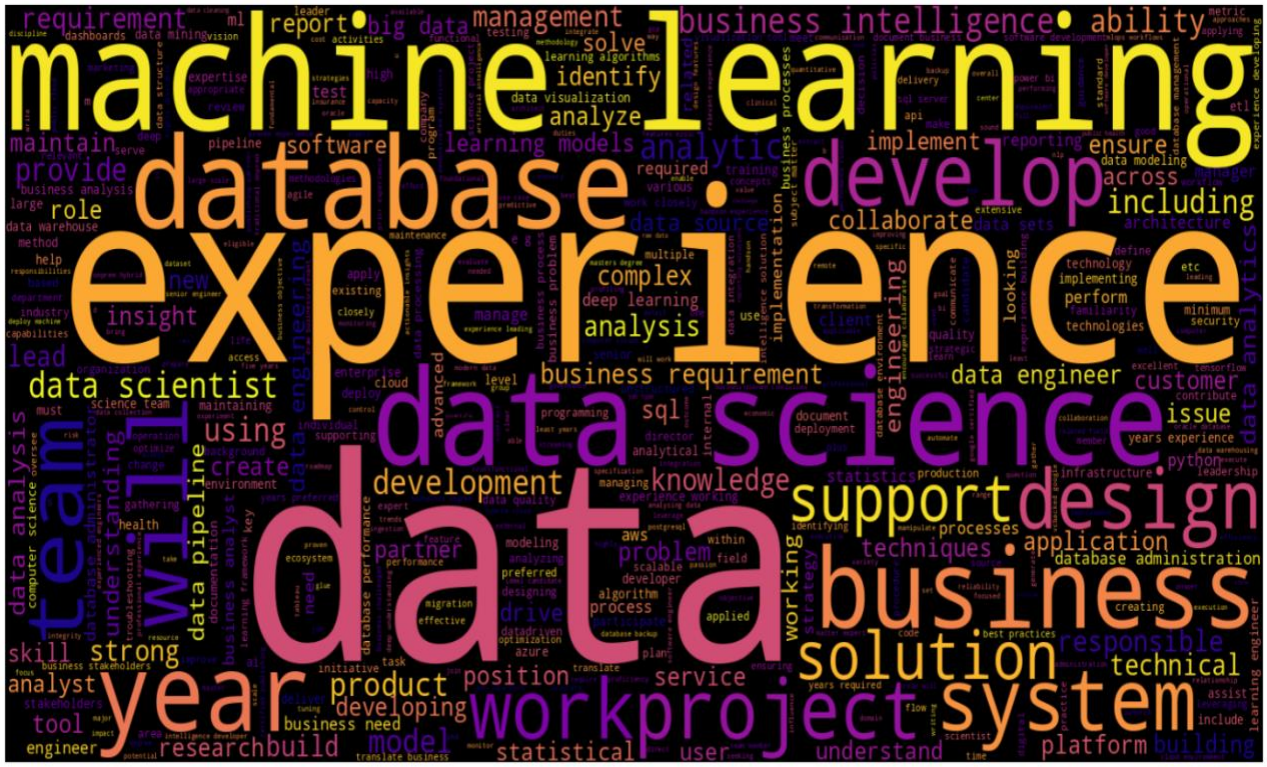

Figure 7:Most Repeated Words 1

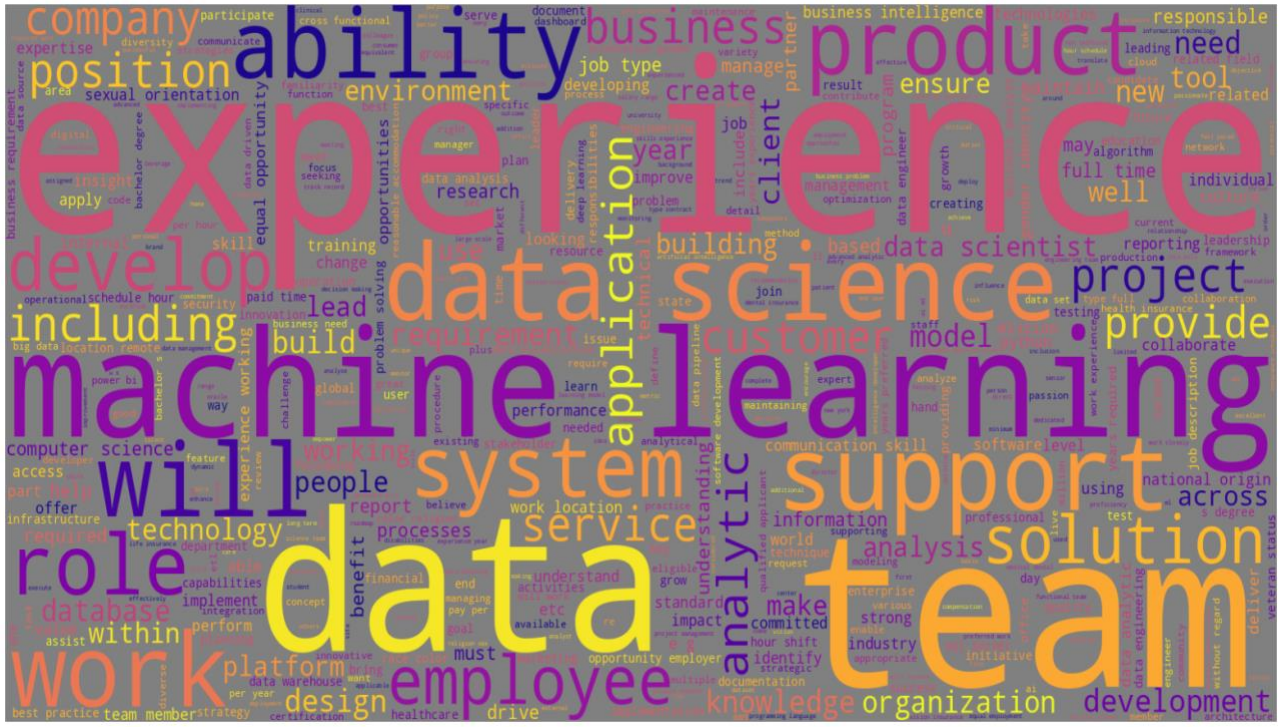

Figure 8:Most Repeated Words 2

7. Model Selection and Model Evaluation:

Regression models such as Random Forest, Gradient Boosting, Linear Regression, & Support Vector Machines (SVM) are used for model evaluation and selection. We use the following criteria for training and evaluating each model:

- **Training Time:** The amount of time needed to train the model using data from Job Indeed.
- **R2 Score (Coefficient of Determination):** Assesses how much of the observed variation has been resolved by the model.
- **RMSE (Root Mean Squared Error):** This value represents the typical inconsistency between predicted and observed values.
- **MAE (Mean Absolute Error):** A measure of the typical discrepancy between predicted and observed values.

8. Results:

Different machine learning models can be utilized for regression tasks, and the outcomes of these models reflect their abilities.

Default Model	Training_time	train_r2_score	val_r2_score	train_RMSE	val_RMSE	train_MAE	val_MAE
GradientBoosting	0.071116	0.543504	0.054878	0.460573	0.656006	0.280076	0.358419
LinearRegression	0.001209	0.017381	0.039071	0.675730	0.661469	0.393230	0.379923
RandomForest	0.257448	0.882628	0.116807	0.233541	0.634150	0.127473	0.345283
SVM	0.027151	0.100703	0.042852	0.646446	0.660167	0.319026	0.357780

Figure 9:Models Results

As shown above by these results, all models perform outstandingly, with R2 values close to 1.0 for both the training and validation sets. Small error values (very close to 0) are possible with Linear Regression, suggesting a very close fit. While still quite effective, Random Forest and Gradient Boosting have a bit more error than Linear Regression does. Although SVM performs well, it has more significant RMSE and MAE errors compared to the other models. Also different is the time required to train each model; Linear Regression, for example, takes the least time to learn. Overall, the predictive capabilities and error rates of all models are rather good, demonstrating their efficacy in producing correct predictions.

9. Conclusion and Future Recommendations:

Based on the data, it appears that the Random Forest model has the greatest balanced performance metrics. With its high accuracy and low error, it is suitable for making forecasts. Overfitting must be considered, though, particularly with models like linear regression that produced flawless ratings. One of the recommendations for the future is to further optimize the Random Forest model by performing extensive hyperparameter tuning. To find out if the prediction accuracy can be improved, it would also be advantageous to investigate neural network topologies and ensemble approaches. Models may be kept accurate and relevant over time by periodically updating their data and retraining them.

Further optimization of the Random Forest model by detailed hyperparameter tuning is a key move for future improvements. In addition, investigating cutting-edge methods like neural network topologies and ensemble approaches may help improve the reliability of predictions.

To keep models up-to-date and correct, regular reevaluation and retraining are recommended. The models can get more accurate over time because of an ongoing supply of updated data about the job market. Taking these preventive actions is critical to ensuring their future usefulness as well as efficiency in real-world recruiting applications.