



Problem Statement & Context

- Which skills should I focus on learning first, where should I focus my job search, and in which industry should I focus my job search in to increase my odds of getting a job as a Data Scientist or Data Analyst in 2024?
- Data Scientist and Data Analyst jobs require a specialized set of skills. I want to identify the most in demand skills in the Glassdoor job descriptions for a Data Scientist and Data Analyst, so I know which ones to focus on developing first to be a more highly rated candidate during my job search. Understanding where these jobs are located and in which industries will allow me to more efficiently focus my search in the appropriate state and industry to increase my odds of getting a job as a Data Scientist or Data Analyst.

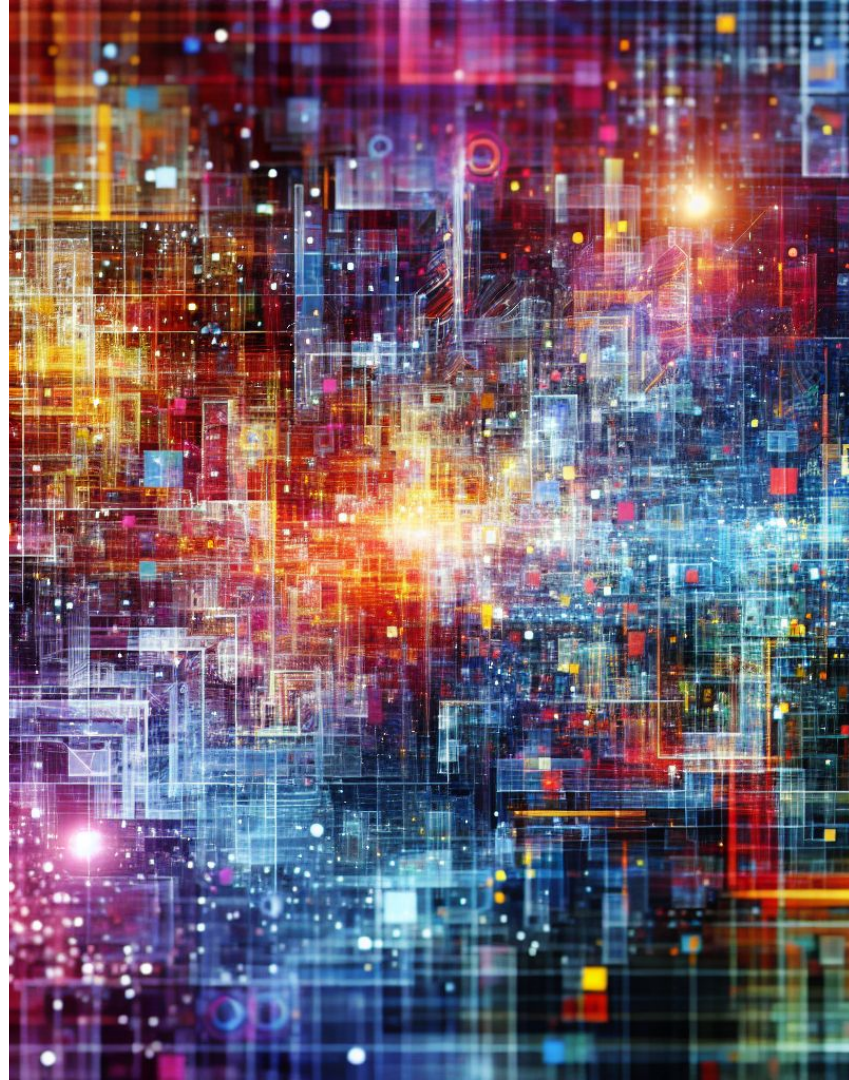
Criteria for Success

Identify the most common skills, location, and industry for a Data Scientist and Data Analyst job in the United States.



Data Collection

Data was gathered from Glassdoor's Data Science job listings using systematic web scraping to provide insights into essential job requirements and trends.





Methodology

Three classification models were employed:

- Logistic Regression
- Random Forest
- Stacked Model combining LightGBM and XGBoost



Model Performance

- Random Forest was selected for its scalability, predictive power, and efficiency.
- Precision and recall scores indicated variability in how well different skills were predicted.

	Skill	Accuracy	Precision	Recall	F1 Score
0	Target_Spark	0.990291	0.666667	0.666667	0.666667
1	Target_Python	0.888350	0.526316	0.416667	0.465116
2	Target_SQL	0.927184	0.384615	0.416667	0.400000
3	Target_R	0.927184	0.733333	0.500000	0.594595
4	Target_NLP	0.985437	1.000000	0.500000	0.666667
5	Target_Statistics	0.941748	0.800000	0.571429	0.666667
6	Target_Excel	0.980583	0.000000	0.000000	0.000000
7	Target_Power BI	0.985437	0.000000	0.000000	0.000000
8	Target_scikit-learn	0.985437	0.000000	0.000000	0.000000
9	Target_Azure	0.985437	0.000000	0.000000	0.000000
10	Target_Cloud	0.936893	0.300000	0.333333	0.315789
11	Target_Machine Learning	0.815534	0.693878	0.596491	0.641509
12	Target_Deep Learning	0.980583	0.750000	0.500000	0.600000
13	Target_Keras	0.995146	0.000000	0.000000	0.000000
14	Target_Data Analysis	0.927184	0.363636	0.333333	0.347826
15	Target_TensorFlow	0.980583	0.333333	0.333333	0.333333
16	Target_AWS	0.980583	0.000000	0.000000	0.000000
17	Target_Hadoop	1.000000	1.000000	1.000000	1.000000
18	Target_Tableau	0.970874	0.666667	0.500000	0.571429
19	Target_Data Mining	0.975728	1.000000	0.500000	0.666667

Conclusion and Future Work

- The analysis indicates variability in the model's ability to predict different skills accurately. Some skills are well predicted, while others are not detected at all.
- Improvements can be made by focusing on better feature representation, addressing class imbalance, optimizing model parameters, and utilizing more advanced NLP models. Emphasis should also be given to enriching the dataset to ensure that all skills are sufficiently represented.



Recommendations

1. Use the model to identify the most in demand skills in the job market for Data Scientist and Data Analysts by determining which skills are the most important.
2. This model can be used to help curate a training curriculum based on the most important skills to upskill individuals looking to work as a Data Scientist or Data Analyst.
3. Help understand which industries are hiring the most Data Scientists and Data Analysts and how the companies are rated (according to GlassDoor) to aid job seekers in focusing their job search.

