Data Analyst Job Market:

A Machine Learning Insight Engine

By: Khushi Malik

Linkedin | github

2. Abstract

This project investigates over 2,000 job listings for Data Analyst positions to extract meaningful insights into hiring patterns, salary expectations, and significant job attributes. Utilizing data preprocessing techniques, exploratory data analysis, and a Random Forest Regressor, we developed a predictive model for average salaries based on variables such as company rating, location, and experience. The study demonstrates the practical application of data science in understanding job market dynamics and supporting informed career decisions.

3. Problem Statement

As the demand for data analysts continues to rise across industries, job seekers often struggle to make sense of a highly fragmented job market. The lack of consolidated insights into salary expectations, skill requirements, and company preferences creates uncertainty in career planning. There is a need for a data-driven solution that can decode patterns in job postings and forecast salary trends based on quantifiable features. This project addresses this gap by analyzing real-world job listing data to identify key drivers of compensation and develop a predictive model to assist job seekers, educators, and HR professionals in making informed decisions.

4. Objectives

- Analyze roles, companies, and regions leading hiring trends

- Identify features that significantly influence salary variations
- Predict average salaries using machine learning algorithms

5. Dataset Description

- Source: Web-scraped job listing dataset
- Number of Rows: 2000+
- Features: Company Rating, Job Location, Experience, Job Title, Industry, etc.
- Target Variable: Average Salary (in thousands)
- Preprocessing: Missing values handled using imputation, Label Encoding applied for categorical variables, data cleaned for consistency and completeness

6. Methodology

- Data Cleaning & Preprocessing: Addressed missing values, encoded categorical variables
- Label Encoding / Null Handling: Used LabelEncoder and filled NA values with appropriate strategies
- Exploratory Data Analysis (EDA): Performed using matplotlib and seaborn
- Feature Engineering: Identified key features influencing salary
- Model Building: Implemented Random Forest Regressor
- Hyperparameter Tuning: Performed using RandomizedSearchCV
- Evaluation Metrics: Model evaluated using R^2 score and RMSE

7. Tools and Technologies Used

- Python (pandas, NumPy, matplotlib, seaborn)
- scikit-learn (RandomForestRegressor, LabelEncoder, RandomizedSearchCV)
- Jupyter Notebook
- Excel (initial data filtering and exploration)

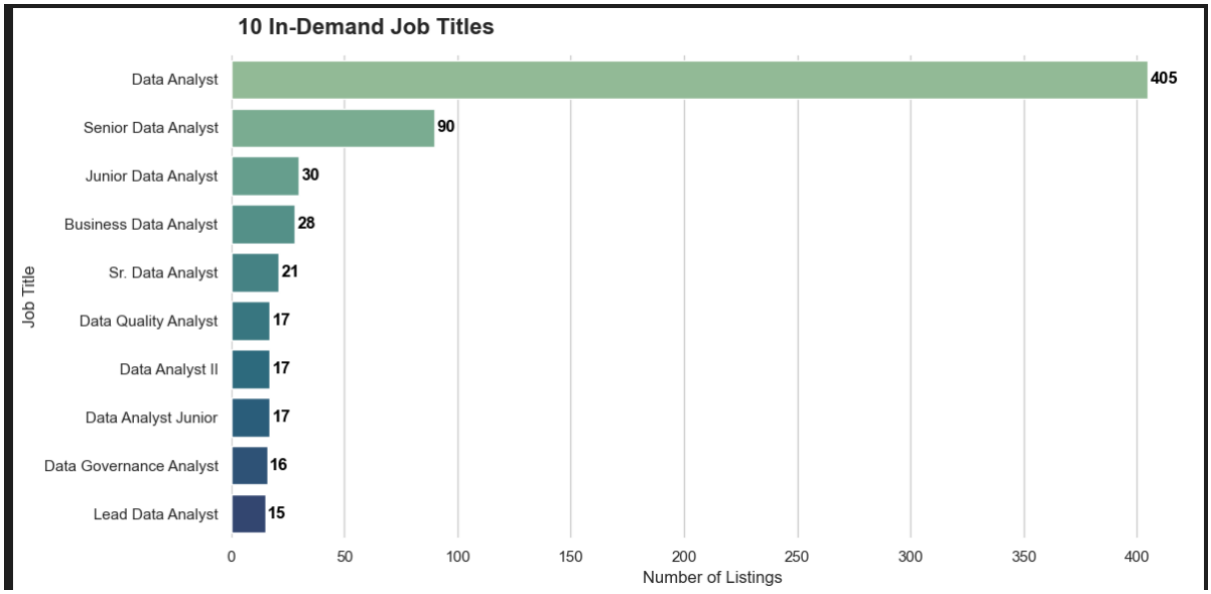
8. Key Findings / Results

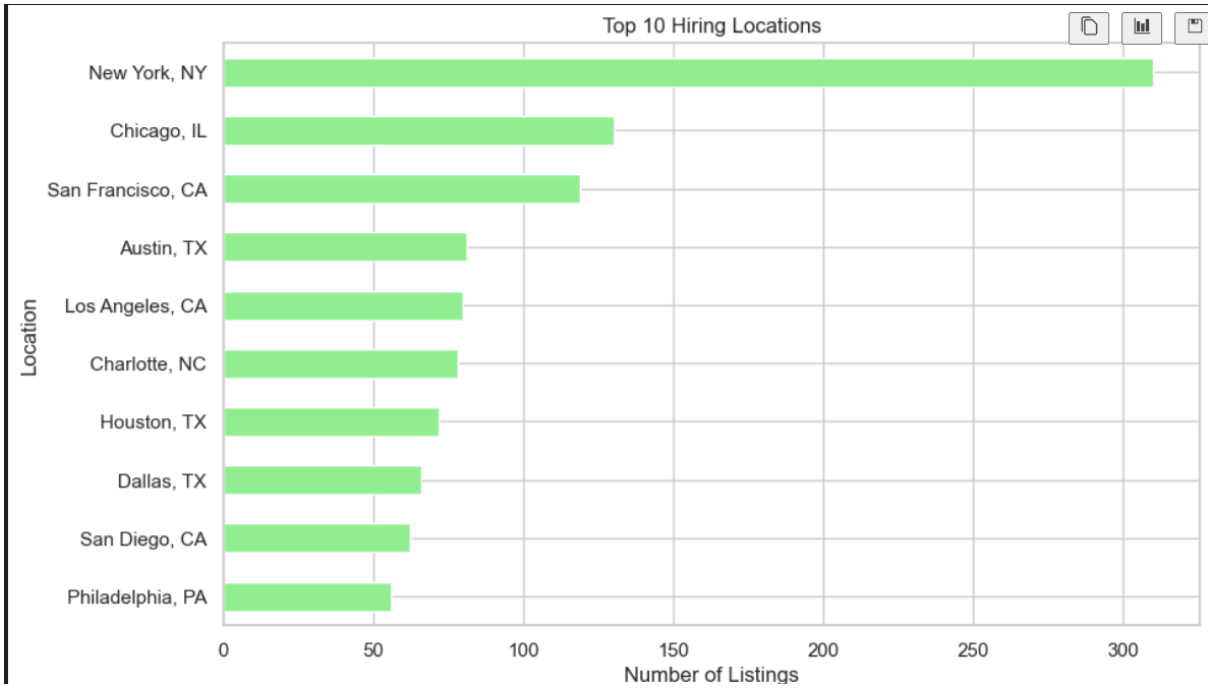
- Top predictors of salary include company rating, job location, and required experience
- R^2 score indicates strong model performance for salary forecasting

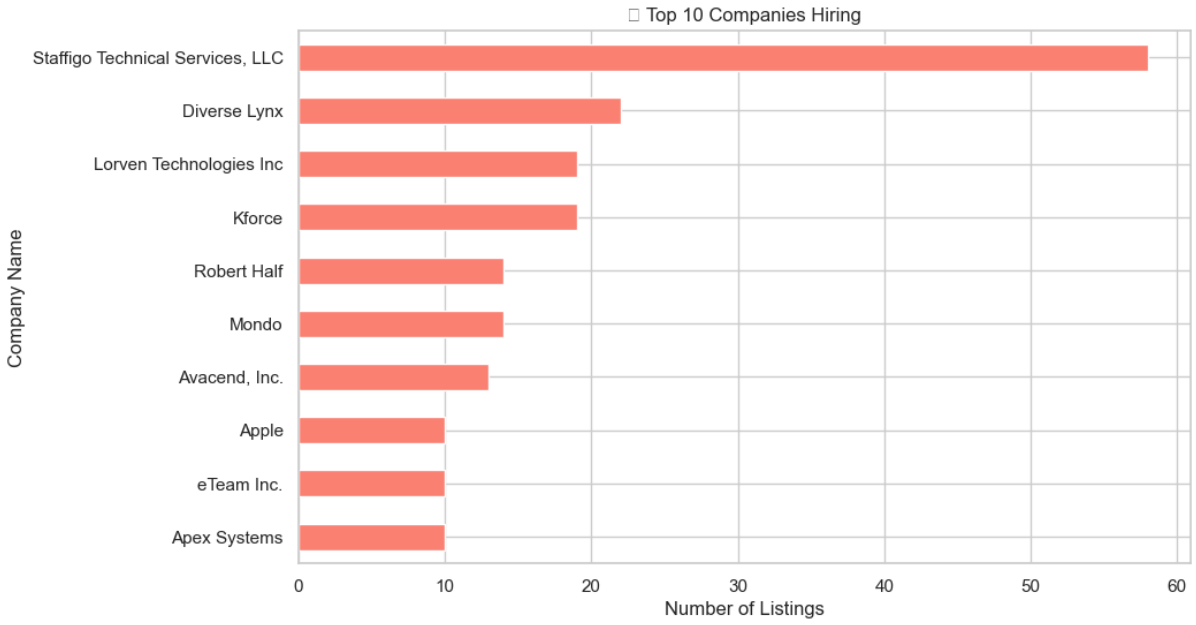
	Model	R2 Score	RMSE	MAE
0	RandomForestRegressor	0.9996	0.4776	0.0624
1	GradientBoostingRegressor	0.9188	7.0540	6.1089
2	SVR	0.0555	24.0533	17.4822

- EDA revealed salary concentration around certain regions and company types

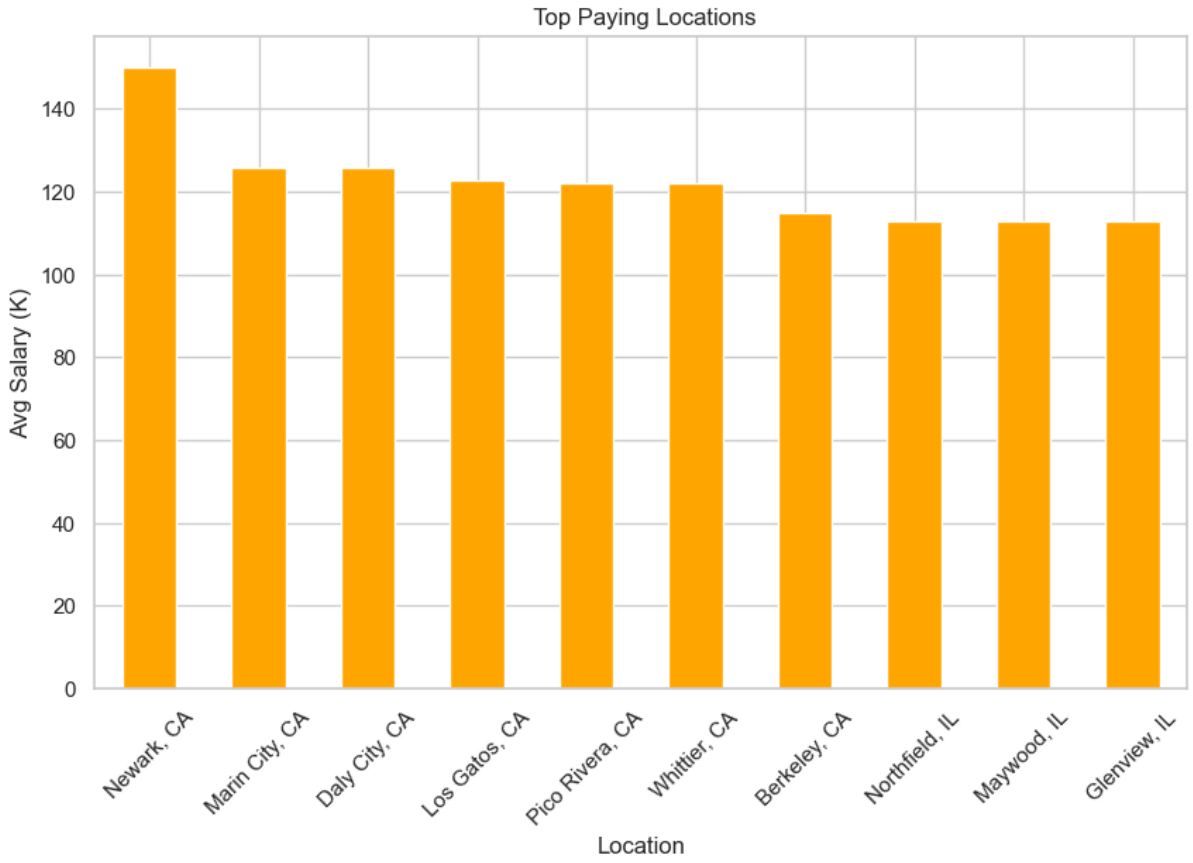
9. Visualizations

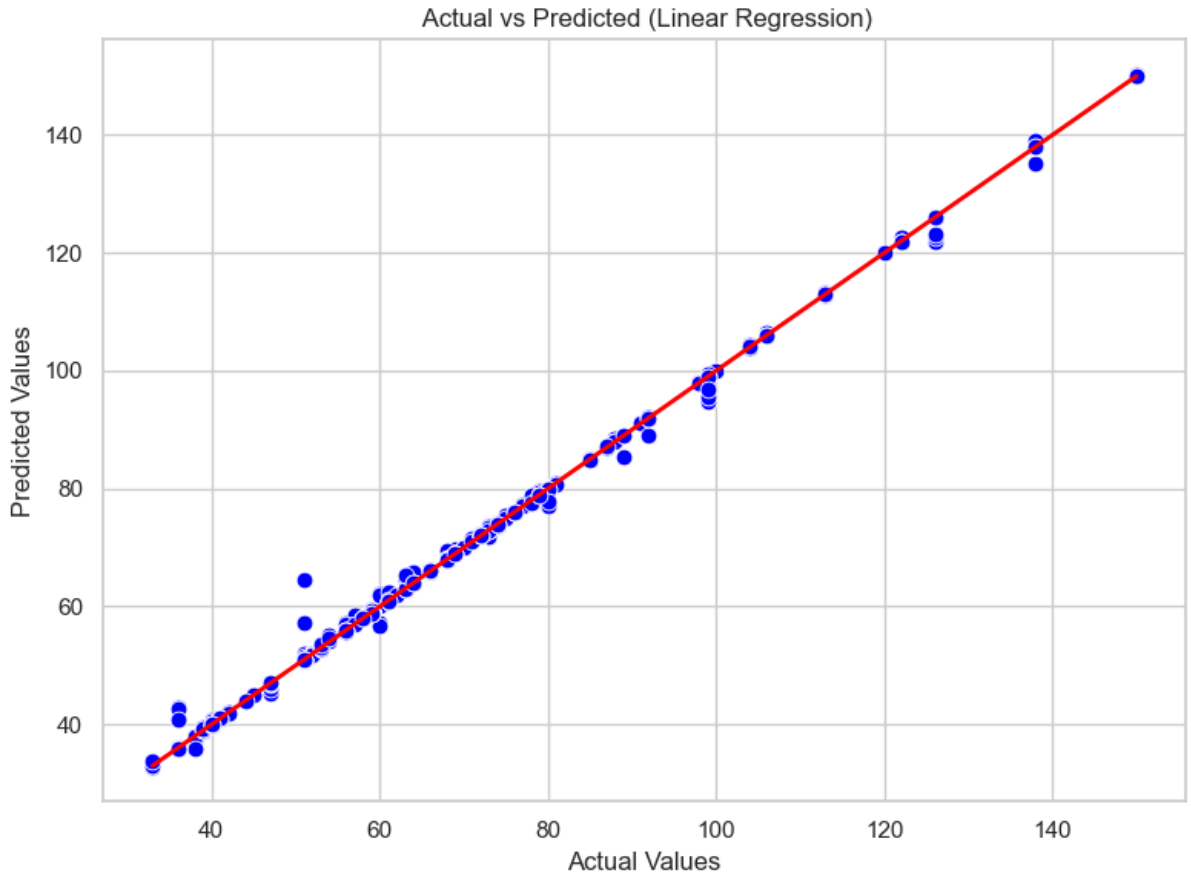


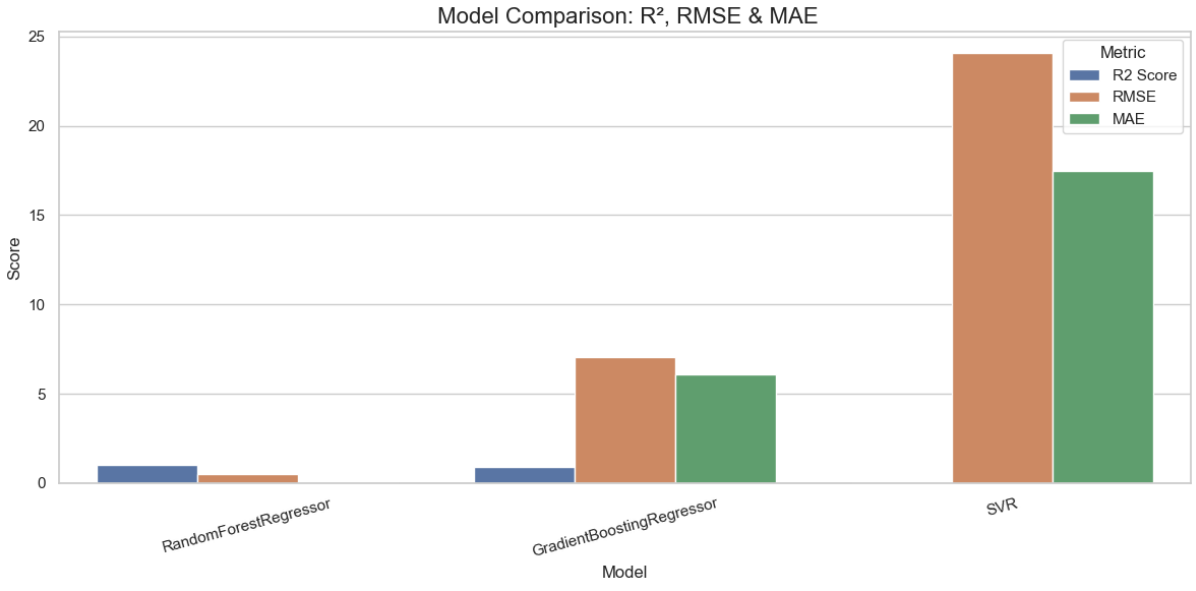

Top 10 Hiring Locations
New York, NY
Chicago, IL
San Francisco, CA
Austin, TX
Los Angeles, CA
Charlotte, NC
Houston, TX
Dallas, TX
San Diego, CA
Philadelphia, PA
Location
0
50
100
150
200
250
300
Number of Listings

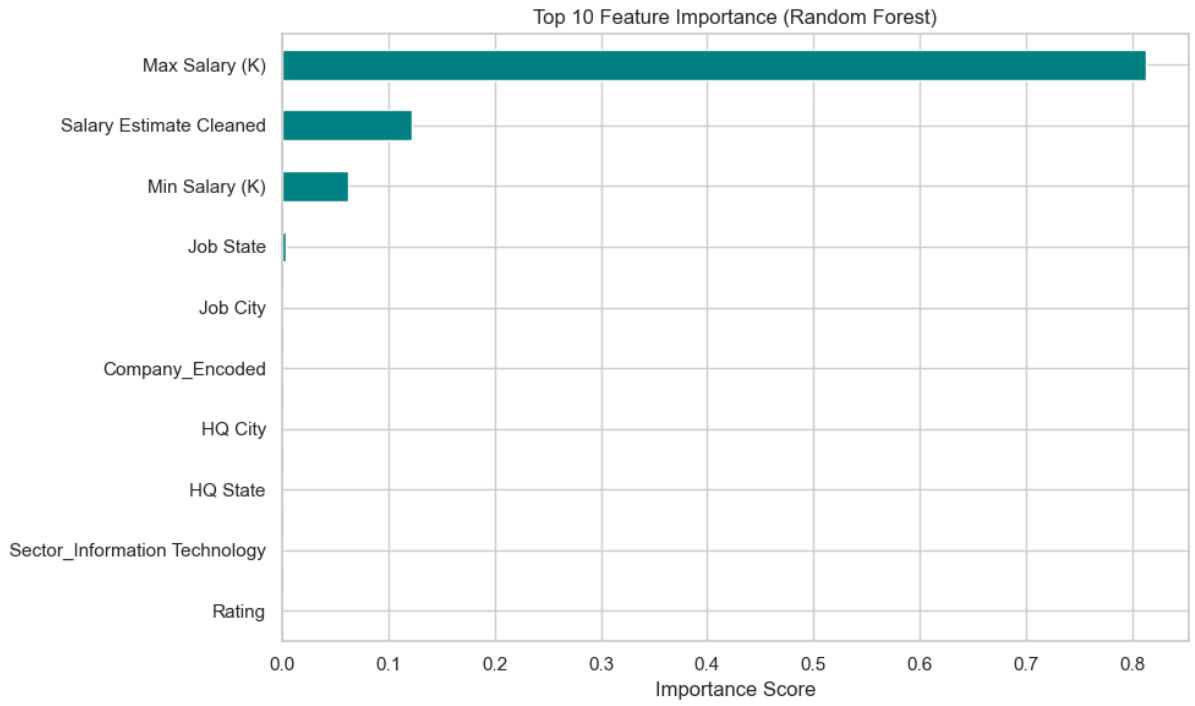

Top 10 Companies Hiring
Staffigo Technical Services, LLC
Diverse Lynx
Lorven Technologies Inc
Kforce
Robert Half
Mondo
Avacend, Inc.
Apple
eTeam Inc.
Apex Systems
Company Name
0
10
20
30
40
50
60
Number of Listings


Company Rating vs Avg Salary
Avg Salary (K)
95
90
85
80
75
70
65
60
0
1
2
3
4
5
Rating


Top Paying Locations
Avg Salary (K)
140
120
100
80
60
40
20
0
Newark, CA
Marin City, CA
Daly City, CA
Los Gatos, CA
Pico Rivera, CA
Whittier, CA
Berkeley, CA
Northfield, IL
Maywood, IL
Glenview, IL
Location


Actual vs Predicted (Linear Regression)
Predicted Values
Actual Values
40
60
80
100
120
140


Model Comparison: R², RMSE & MAE
Score
Model
Metric
R2 Score
RMSE
MAE
0
5
10
15
20
25
RandomForestRegressor
GradientBoostingRegressor
SVR



10. Conclusion

This project delivers critical insights for job seekers and hiring platforms by identifying salary determinants and building a predictive model for compensation analysis. The outcomes highlight the potential of machine learning in interpreting employment trends and enhancing the strategic planning of career paths and HR decisions.

11. Future Work

- Streamlit deployment of the salary prediction tool
- Use of advanced models like XGBoost or LightGBM
- Incorporate NLP techniques to analyze full job descriptions
- Real-time job data integration for dynamic insights