Data Science Salary Analysis: Insights and Predictive Modeling for Competitive Hiring Strategies

Britt Hoover, Chloe Cramer, Elli Reel, and Dominic Flanagan

1. Executive Summary

This report investigates salary trends within the data science industry to assist TechHire Recruiting Firm in optimizing hiring strategies and establishing competitive pay structures for their clients. By analyzing the "Data Science Salaries 2024" dataset from Kaggle, we explored key factors such as job title, experience level, and employment type, that influence compensation. The data underwent pre-processing and was analyzed using multiple regression models, including Gradient Boosting, Neural Networks, Decision Trees, and Linear Regression, to uncover patterns and relationships. Gradient Boosting was the most accurate model across evaluations (test on train, test on test, and cross-validation), showing its ability to predict salary outcomes. These insights provide TechHire with a foundation for advising clients on market trends and pay strategies, helping them attract and retain talent in the data science field.

2. Problem Description

2.1 Background

The business problem focuses on salary trends in the data science field. TechHire Recruiting Firm, the client, wants to improve their hiring strategies by leveraging insights from salary data. With the increasing demand for data science roles and the varying salary ranges across job titles, experience levels, and geographical locations, understanding salary determinants has become a priority. The Kaggle dataset allows us to identify these trends to give our clients the insights they need. This context highlights the critical need for data-driven approaches to establish competitive salary structures in the tech industry.

2.2 Business Goal and Data Mining Goal

The primary business goal of this project is to provide TechHire Recruiting Firm with valuable insights into salary trends within the data science industry. By understanding the factors that influence compensation, the firm can better support its clients in establishing competitive salary ranges and improving their hiring strategies. Specifically, the business goal is to understand how variables such as experience level, job title, and remote work ratio relate to salary outcomes, allowing TechHire to advise clients effectively on market trends and pay structures.

To achieve this, the data mining goal focuses on identifying patterns and relationships within the dataset that can offer meaningful insights into salary determinants. The emphasis will be on understanding the data, discovering significant trends, and setting the foundation for predictive modeling and other data mining techniques. This goal ensures that the solutions provided align with the business objectives and supply valuable findings for TechHire’s needs.

3. Data Description

3.1 Data

To address TechHire Recruiting Firm’s need to enhance its hiring strategies, we are utilizing the “Data Science Salaries 2024” dataset, sourced from Kaggle. This dataset provides a comprehensive view of salary trends in the data science field and allows us to analyze the key factors of compensation.

The dataset contains a large set of features that capture critical aspects of job roles and compensation, including:

- **Job Title**: A detailed list of various data science roles, such as Data Scientist, Machine Learning Engineer, and Data Analyst.
- **Salary**: The annual compensation associated with each role, serving as the target variable for our analysis.
- **Location**: Geographical information, which allows us to evaluate salary differences based on location.
- **Experience Level**: The required experience for each role, categorized as Entry-Level, Mid-Level, Senior-Level, or Expert.
- **Employment Type**: Employment arrangements, such as full-time, part-time, freelance, or contract roles.
- **Company Size**: An indicator of the organization’s scale (e.g., small, medium, large), which is often correlated with compensation.

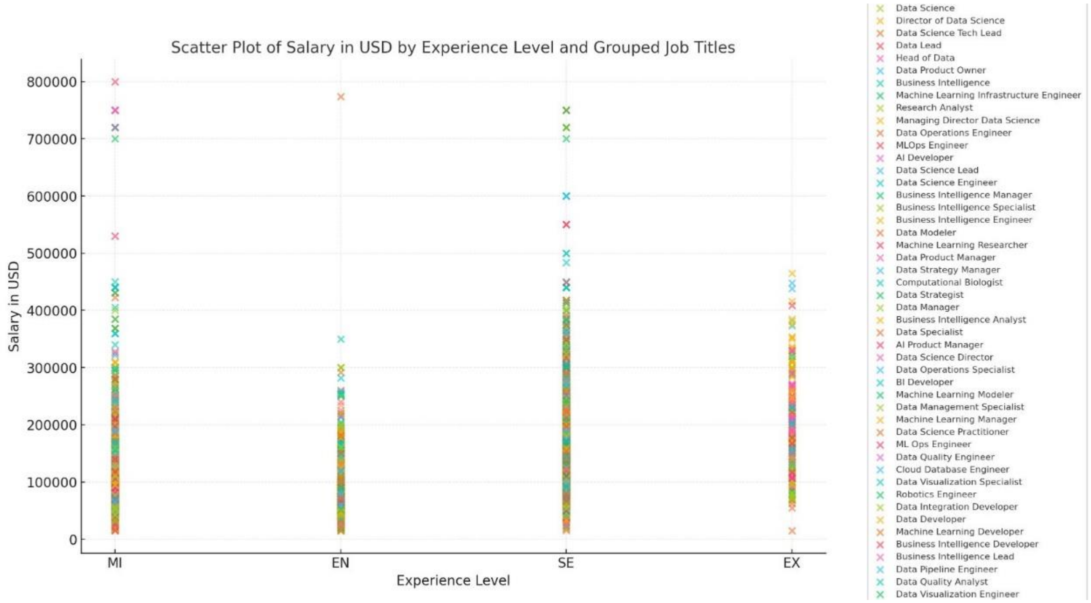
This dataset is critical for exploring the salary dynamics in the data science industry and identifying patterns that could inform TechHire’s strategies. By analyzing these features, we aim to provide actionable insights to help the firm design competitive salary packages and attract top talent.

3.2 Exploratory Analysis

In our exploratory analysis we valued three different charts that efficiently visualized our data mining problem.

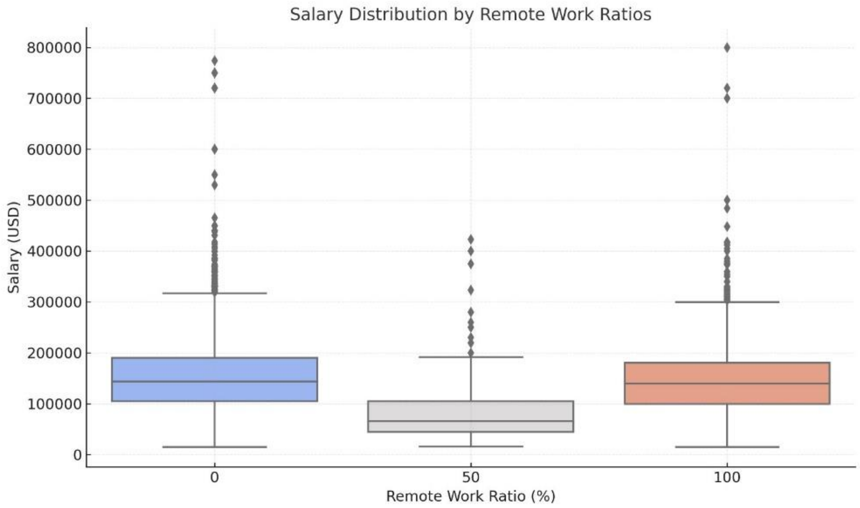
These are the scatter plot, identified in Table 1.1, the box plot in Table 1.2, and the bar chart in Table 1.3.

Table 1.1: Scatter Plot



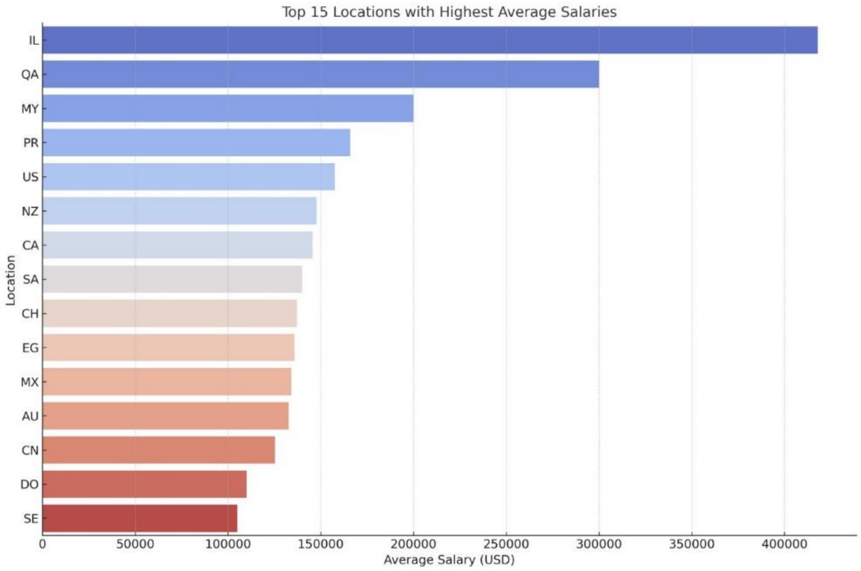
The scatter plot highlights the relationship between salaries, experience levels (Entry-Level, Mid-Level, Senior-Level, and Executive), and grouped job titles. As expected, senior-level roles (SE) generally command higher salaries, with certain job titles such as Data Scientist, Machine Learning Engineer, and Research Scientist reaching the highest salary ranges, often exceeding $400,000. Entry-level roles (EN) and mid-level roles (MI) display lower salary distributions, reflecting a typical progression in compensation with experience. Interestingly, executive roles (EX) show significant variability in salaries, with some overlapping the senior-level range, indicating that not all executive positions are uniformly high-paying.

Table 1.2: Box plot



The box plot reveals that fully remote roles (100% remote) tend to offer higher and more consistent salaries, suggesting that companies may provide competitive pay to attract remote talent. Fully on-site roles (0% remote) exhibit a wider range of salaries, reflecting location-dependent factors or variability in job types. Partial remote roles show moderate salaries with less fluctuation, indicating a balance between flexibility and compensation.

Table 1.3: Bar Chart



The bar chart of locations reveals that countries like Israel (IL), Qatar (QA), and Malaysia (MY) offer the highest average salaries in the data science field, with salaries significantly exceeding those in other regions. This highlights regional variations in compensation, likely influenced by factors such as demand for talent, cost of living, and the maturity of the tech industry in these locations.

3.3 Data Pre-processing

To prepare the dataset for analysis, several preprocessing steps were implemented using Orange's pre-processing tools. First, categorical variables were processed using the "One feature per value" ensuring compatibility with machine learning models. For numerical features, normalization was performed to scale the data to an interval of [0, 1], which improves the comparability of features and supports model performance. These preprocessing steps confirmed the dataset was ready for further analysis and modeling.

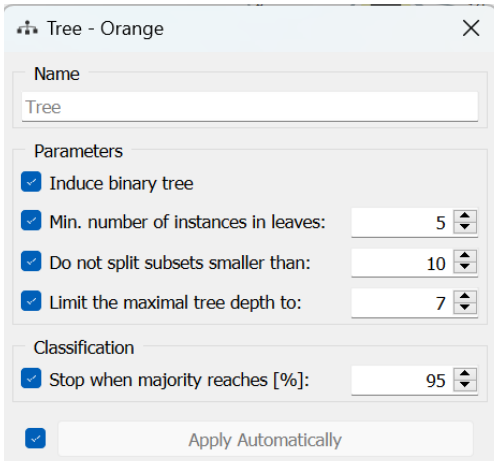
4. Data Mining Solution

4.1 Models

Our target variable is Salary in USD, which represents a supervised learning problem, specifically a regression task, because the target variable is continuous. We explored a variety of predictive models to find the model that best suits our data mining problem. We started with simpler models like linear regression and then progressed to more complex models like decision trees and neural networks. Through our analysis, we found that models with higher complexity, such as gradient boosting and neural networks, performed best with our dataset, so we focused our efforts on these models.

To optimize the performance of these models, we carefully tuned their hyperparameters which you can see in the images below.

Decision Tree:



Tree - Orange

Name

Tree

Parameters

☑ Induce binary tree

☑ Min. number of instances in leaves: 5

☑ Do not split subsets smaller than: 10

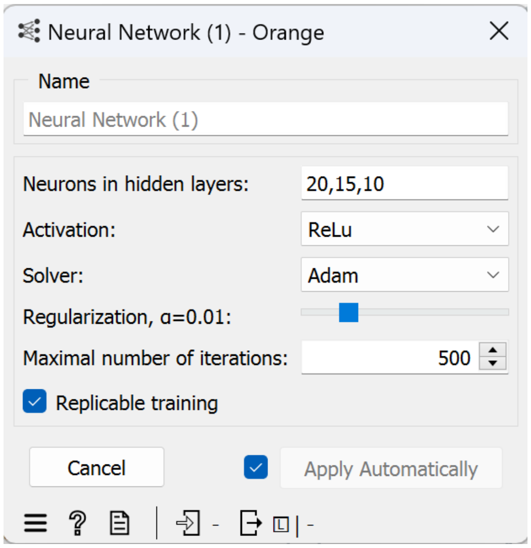
☑ Limit the maximal tree depth to: 7

Classification

☑ Stop when majority reaches [%]: 95

☑ Apply Automatically

Neural Network:



Neural Network (1) - Orange

Name: Neural Network (1)

Neurons in hidden layers: 20,15,10

Activation: ReLu

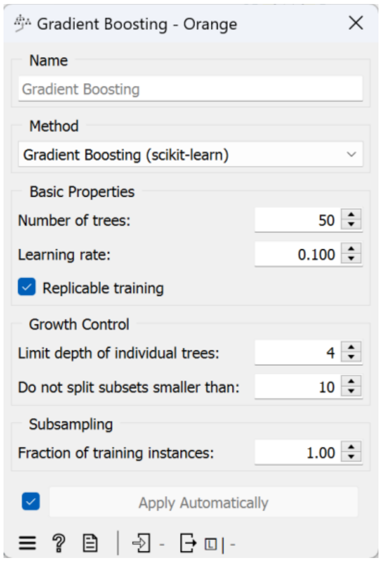
Solver: Adam

Regularization, α=0.01:

Maximal number of iterations: 500

☑ Replicable training

Gradient Boosting:



Gradient Boosting - Orange

Name: Gradient Boosting

Method: Gradient Boosting (scikit-learn)

Basic Properties

Number of trees: 50

Learning rate: 0.100

☑ Replicable training

Growth Control

Limit depth of individual trees: 4

Do not split subsets smaller than: 10

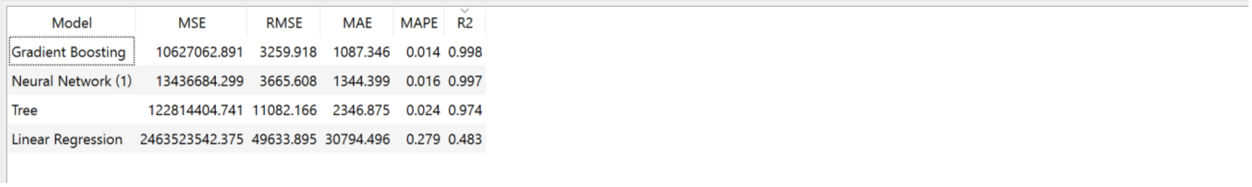
Subsampling

Fraction of training instances: 1.00

By carefully selecting and tuning the hyperparameters for each model, we made sure the models fit the data well and could capture the important patterns needed to predict salaries accurately.

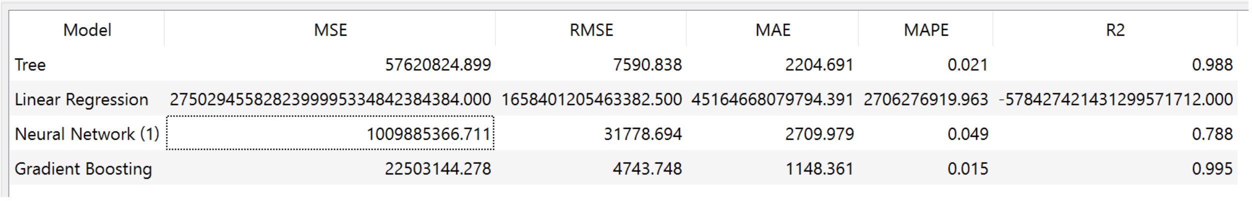
4.2 Performance Evaluation

Test on Train:



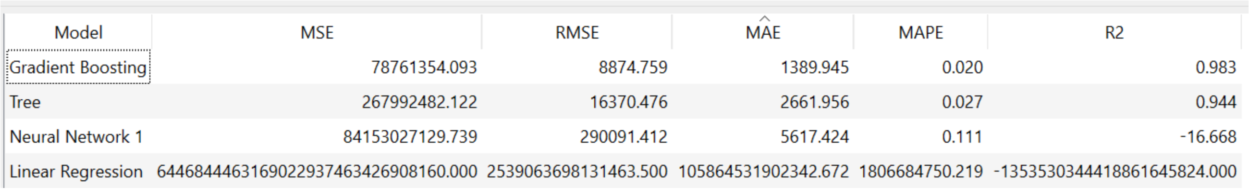
Model	MSE	RMSE	MAE	MAPE	R2
Gradient Boosting	10627062.891	3259.918	1087.346	0.014	0.998
Neural Network (1)	13436684.299	3665.608	1344.399	0.016	0.997
Tree	122814404.741	11082.166	2346.875	0.024	0.974
Linear Regression	2463523542.375	49633.895	30794.496	0.279	0.483

Test on Test:



Model	MSE	RMSE	MAE	MAPE	R2
Tree	57620824.899	7590.838	2204.691	0.021	0.988
Linear Regression	2750294558282399995334842384384.000	1658401205463382.500	45164668079794.391	2706276919.963	-578427421431299571712.000
Neural Network (1)	1009885366.711	31778.694	2709.979	0.049	0.788
Gradient Boosting	22503144.278	4743.748	1148.361	0.015	0.995

Cross Validation:



Model	MSE	RMSE	MAE	MAPE	R2
Gradient Boosting	78761354.093	8874.759	1389.945	0.020	0.983
Tree	267992482.122	16370.476	2661.956	0.027	0.944
Neural Network 1	84153027129.739	290091.412	5617.424	0.111	-16.668
Linear Regression	6446844463169022937463426908160.000	2539063698131463.500	105864531902342.672	1806684750.219	-135353034418861645824.000

Across the three evaluations, test on train, test on test, and cross-validation, Gradient Boosting performed the best, with the lowest error metrics (MSE, RMSE, MAE, MAPE) and the highest R² values. This shows how well the Gradient boosting model handled the data. Neural Network 1 had okay results but higher errors, making it less reliable than Gradient Boosting. The Decision Tree model did fairly well but wasn't as accurate. Linear Regression consistently had the worst results with high errors and poor R² values, showing it's not a good fit for this dataset. Overall, Gradient Boosting is the best choice for predicting salary trends.

5 Conclusion

5.1 Recommendations

Based on our analysis, here are some recommendations for TechHire Recruiting Firm:

1. Offer Competitive Salaries for High-Demand Roles:
    - Roles like Research Scientist, Applied Scientist, and Research Engineer have the highest salaries, averaging over $190,000. TechHire should advise clients to offer competitive pay for these roles to attract top talent.
    - For lower-paying roles like Data Analyst, focus on offering other benefits, such as career growth opportunities or flexible work options.
2. Consider Location When Setting Salaries:
    - Salaries vary greatly by location. Clients should tailor their salary ranges to match the competitiveness of each region.
3. Reward Experience:
    - Senior and expert-level roles earn significantly more than entry-level roles. Clients should create clear career progression paths and adjust salaries based on experience to retain skilled professionals.

4. Prioritize Skills for High-Paying Roles:
   - Specialized roles like Machine Learning Engineer and Data Architect have higher salaries. Clients should focus on hiring candidates with these in-demand skills and highlight them in job descriptions.
5. Support Remote Work:
   - Remote work ratios impact salaries. Promoting remote or hybrid work options may help attract top candidates, especially in highly competitive regions.

5.2 Limitations

While our analysis provides useful insights, there are a few limitations to keep in mind:

- Missing Features: The dataset doesn't include factors like employee performance or team size, which could improve salary predictions.
- Uneven Data: Some roles and locations have fewer data points, which might affect accuracy.
- External Factors: Trends like economic changes or shifts in work culture (e.g., remote work) are not fully captured in the dataset.

5.3 Future Work

To improve and expand on this analysis, we recommend:

- Adding More Data: Include details like performance metrics, team sizes, and industry trends to refine the results.
- Tracking Salary Trends Over Time: Analyze how salaries change over time to help predict future trends for high-demand roles.
- Creating Interactive Dashboards: Build easy-to-use tools that let TechHire explore salary trends and present insights to their clients.

These steps will help TechHire refine its hiring strategies and offer stronger, data-backed advice to its clients.