Salary Prediction in Different Fields Of Data Science Using Modern Machine Learning Methods

Corey Lang, Quan Le, Reece Stammen

Abstract

Using datasets centered around collecting information about those employed in the field of data science, an experiment was conducted using machine learning models to derive relationships from the data. After analyzing related work and applying similar methods to the experiment, results were produced and analyzed for relationships between various independent variables and the dependent variable of salary. After conclusions are drawn about these relationships, strengths and weaknesses of the experiment and models are identified and discussed. The shortcomings of the experiment are directly mentioned along with how they could be addressed in future experiments similar to this one.

Introduction

In today's data-driven world, data science professionals play a crucial role in extracting valuable insights from vast amounts of data to drive informed decision-making and innovations. As the field continues to grow, understanding the salary trends and earning potential of data science professionals becomes increasingly important for both job seekers and employers alike. In this project, we aim to investigate the various factors that influence the salaries of data science professionals, such as job title, years of experience, company size, and geographical location.

The problem we address is significant because it sheds light on the earning potential of different career paths within the data science domain, enabling individuals to make informed decisions about their career trajectories. Furthermore, this analysis can help companies benchmark their compensation packages against industry standards and identify trends that may impact their talent acquisition and retention strategies.

To conduct our investigation, we analyzed a comprehensive dataset “Data Science Fields Salary Categorization: Salaries of Different Data Science Fields in the Data Science Domain” from kaggle.com that contains detailed information on data science professionals across various categories, such as working year, salary, experience, job title, size of the company, and location. Our approach involved using various regression models, graphs and plots to identify patterns and correlations within the data. We also employed machine learning techniques to predict future salary trends based on the available data.

This project fits into the broader context of research on labor economics, human resource management, and data science careers. Prior studies in this area have focused on understanding the factors that drive job satisfaction, skill requirements, and job mobility among data scientists. Our work complements these studies by specifically focusing on salary trends and their underlying determinants.

Through this analysis, we will present key findings on the relationships between various factors and data science professionals' salaries. We will also offer insights into the future trends in terms of earning potential for professionals in this field. By providing a comprehensive understanding of the salary landscape in data science, our project will serve as a valuable resource for individuals and organizations navigating this dynamic and rapidly evolving industry.

Related Work

Quan & Raheem (2022) conducted a comprehensive analysis of various approaches for predicting salaries in the data science field, comparing traditional statistical techniques like multiple linear regression with advanced machine learning methods. The study highlights the significance of skills development over examination scores for accurate salary prediction, the importance of data cleaning and outlier management in regression problems, and the issue of underpayment among engineers affecting the reliability of salary prediction models. A deep learning neural network model integrating contextual data and leveraging natural language processing (NLP) for continuous salary value prediction outperforms other combined models with the lowest mean absolute error (MAE), though it requires longer training times. The authors also propose a backpropagation neural network (BPNN) model with an assisting gradient descent algorithm for enhanced accuracy, finding that increasing the number of neurons and hidden layers improves accuracy up to an optimal point. The study emphasizes the crucial role of skills development and job benefits in salary prediction, calling for further research to identify optimal methodologies for constructing data science salary prediction models.

Gopal et al. (2021) aimed to develop a predictive model for estimating graduate students' salaries based on their profiles, employing data mining techniques such as decision trees, Naïve Bayes, and k-NN. The study analyzes various factors, including department, program, certificate, CGPA, and salary class. The model's performance is assessed using 10-fold cross-validation. In addition, the paper provides a thorough review of relevant literature on data mining and learning patterns among learners. The overarching research objective is to develop an efficient, effective learning system that enables researchers and students to better comprehend learner patterns and derive maximum benefits from the system.

Problem Definition and Algorithm

Task Definition

The problem we aim to address in this report is the prediction of salaries for data science professionals based on various attributes. We will utilize two datasets for this task, namely the Data Science Fields Salary Categorization dataset and the Data Science and STEM Salaries dataset. Formally, we define the inputs and outputs as follows:

Inputs

1. Designation (i.e. Title)
2. Experience (i.e. Years at Company, Experience level in the job/Years at Job)
3. Location of the job (e.g., City, State, Country)
4. Additional factors (e.g., Company, Company size, Employment Status)

Outputs

1. Predicted salary for the given data science professional

The objective is to create a machine learning regression model that takes these input attributes and predicts the salary for data science professionals accurately.

This problem is both interesting and important for several reasons. First, by understanding the factors that contribute to the salaries in the data science field, individuals can make informed decisions about their career paths and plan their professional development accordingly. Second, this information can be useful for companies to create competitive compensation packages to attract and retain top talent in the rapidly evolving data science industry. Third, it helps academic institutions and professional training programs to tailor their curricula to better meet the demands of the job market and ensure their graduates are well-prepared for the workforce.

Algorithm/Methodology Definition

For this task, we will employ a Linear Regression model to predict the salaries of data science professionals using each dataset separately. Linear Regression is a simple and interpretable model that serves as a baseline, making it suitable for this initial analysis. We will also use various visualization techniques to better understand the data and model results, such as bar graphs, choropleth maps, seaborn heatmaps, and seaborn hist plots.

The general approach for each dataset will be:

1. Preprocess the data: Encode categorical variables (e.g., Role, Location, Education level), normalize numerical variables, and handle missing values for each dataset.
2. Split the data: Divide each dataset into training and testing sets to evaluate the performance of the Linear Regression model.
3. Train the models: Fit a Linear Regression model to the training data of each dataset.
4. Evaluate the models: Assess the performance of each model using metrics like Mean Squared Error (MSE), and R-squared score on the testing set.
5. Visualize the results: Create visualizations to further understand the data and model results. These visualizations will include:

    A. Bar graphs: Show the distribution of salaries across different roles, locations, or other categorical attributes.
    B. Choropleth maps: Display geographical patterns in the salaries of data science professionals based on job location.
    C. Seaborn heatmap: Visualize the correlation between different input features and the output salaries.
    D. Seaborn Histplot: Illustrate the distribution of salaries in the data science field.

By following these steps, we will apply Linear Regression models to each dataset and gain insights into the salaries of data science professionals based on the given attributes. The visualizations will provide an additional understanding of the patterns and trends in the data, enhancing our interpretation of the model's results.

Experimental Evaluation

Experimental Methodology

In this study, we aim to evaluate the effectiveness of the Linear Regression models in predicting the salaries of data science professionals based on various attributes. Our primary hypothesis is that the Linear Regression models can accurately predict salaries using the given input features. To test this hypothesis, we designed the following experimental methodology:

1. Data selection: We chose two realistic and relevant datasets, the Data Science Fields Salary Categorization dataset and the Data Science and STEM Salaries dataset, for our analysis. These datasets contain a diverse range of attributes, such as job title, years of experience, company size, and geographical location, making them interesting and suitable for our investigation.

2. Independent and dependent variables: The independent variables include Designation (i.e., Title), Experience (i.e., Years at Company, Experience level in the job/Years at Job), Location of the job (e.g., City, State, Country), and Additional factors (e.g., Company, Company size, Employment Status). The dependent variable is the predicted salary for the given data science professional.

3. Data preprocessing: We preprocessed the data by encoding categorical variables, normalizing numerical variables, and handling missing values for each dataset.

4. Data split: We divided each dataset into training and testing sets, typically using an 80/20 or 70/30 split, to evaluate the performance of the Linear Regression models.

5. Model training and evaluation: We fit the Linear Regression models to the training data of each dataset and assessed their performance using metrics like Mean Squared Error (MSE) and R-squared score on the testing set.

6. Performance data collection and analysis: We collected the performance metrics of our models, such as MSE and R-squared scores, and analyzed them to determine the accuracy and reliability of our salary predictions.

7. Presentation of results: We will present the performance data using tables and visualizations, such as bar graphs, choropleth maps, seaborn heatmaps, and seaborn hist plots, to provide a comprehensive understanding of the patterns and trends in the data and enhance the interpretation of the model's results.

By following this experimental methodology, we aim to rigorously test our hypothesis and evaluate the performance of the Linear Regression models in predicting salaries for data science professionals based on the given attributes.

Results

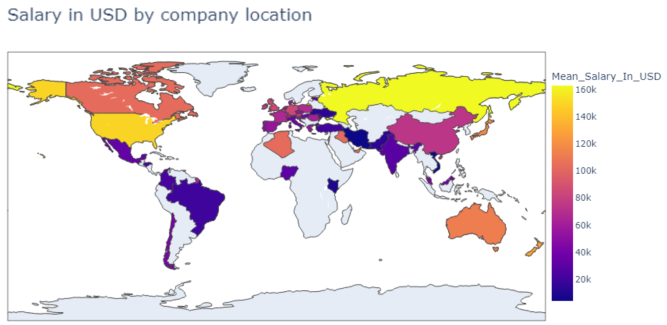

Figure 1.

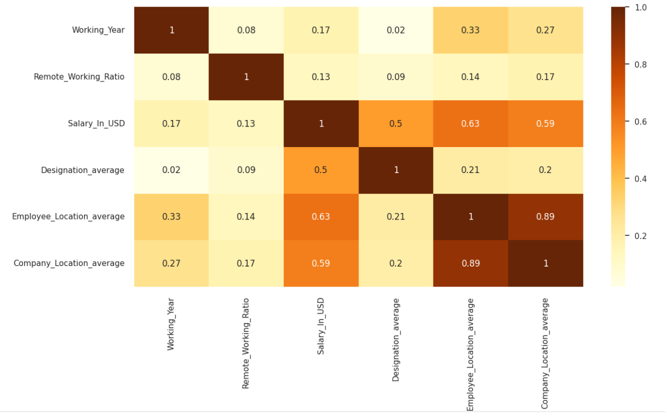

Figure 2.

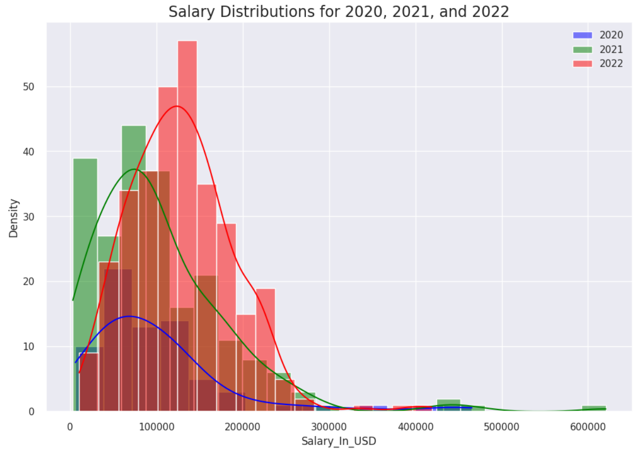

Figure 3.

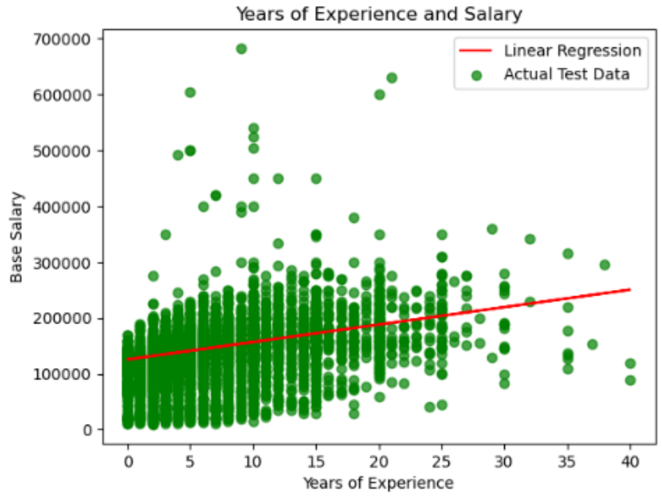

Figure 4.

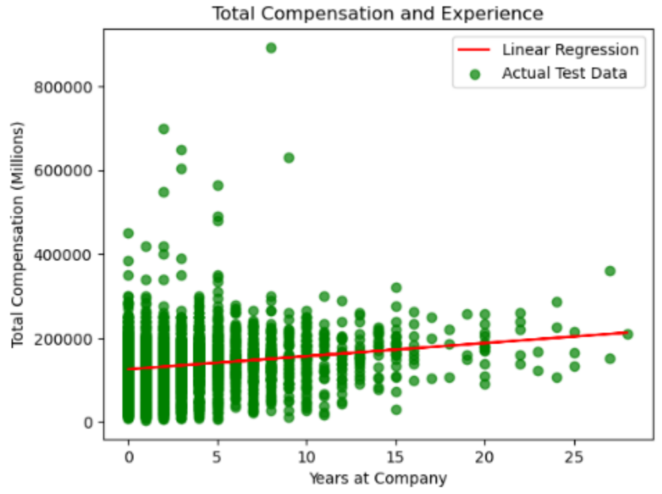

Figure 5.

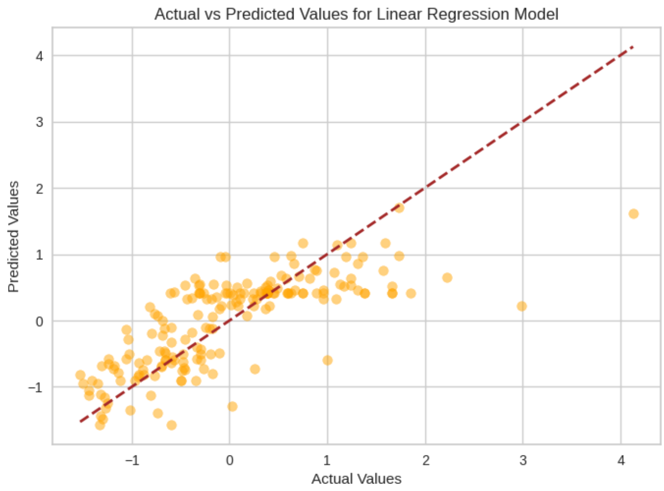

Figure 6.

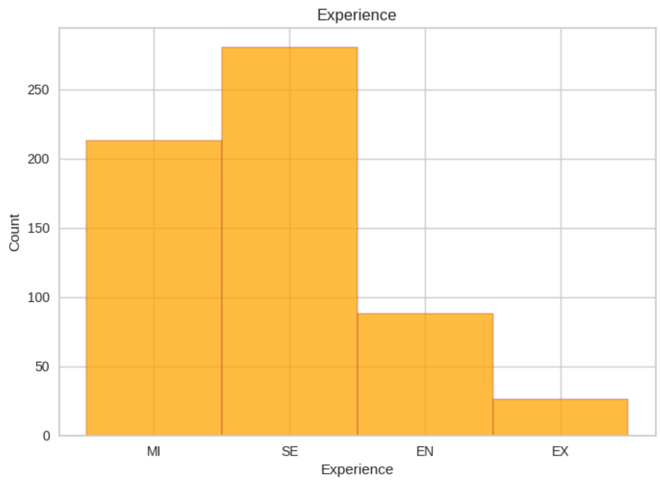

Figure 7.

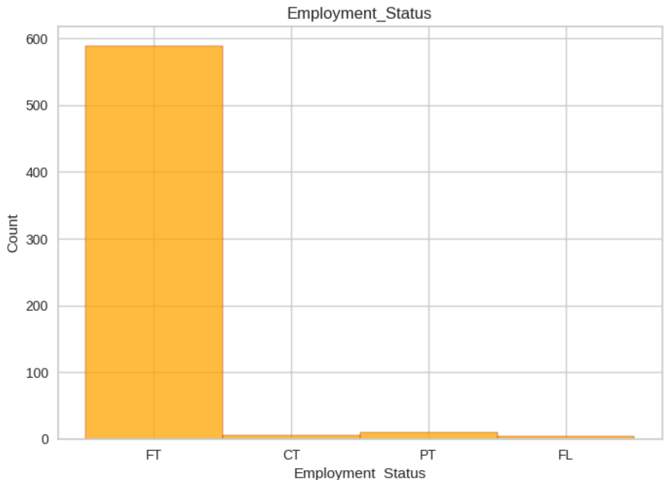

Figure 8.

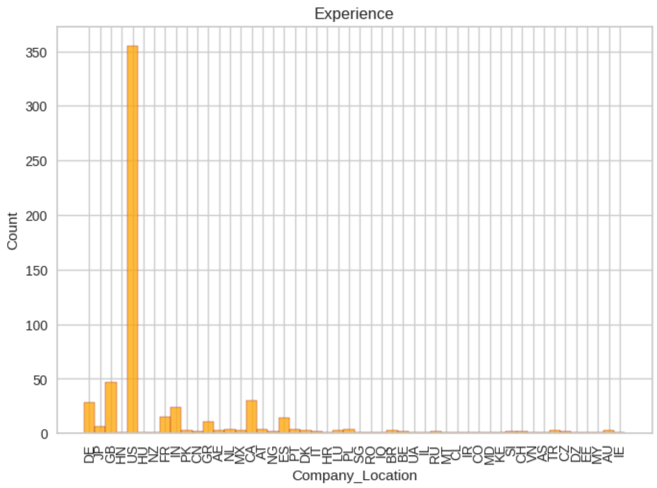

Figure 9.

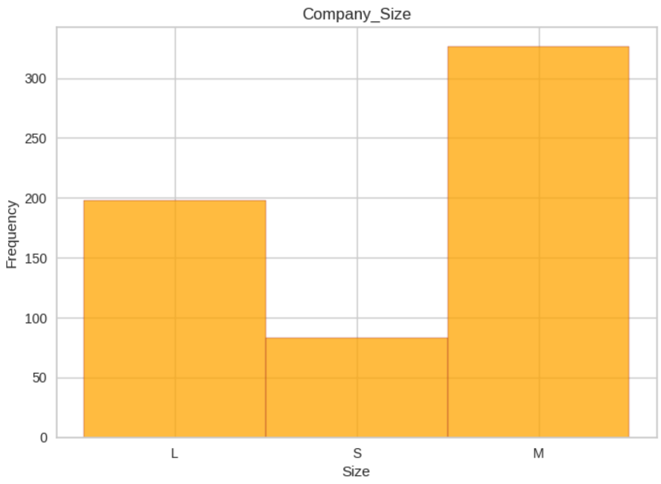

Figure 10.

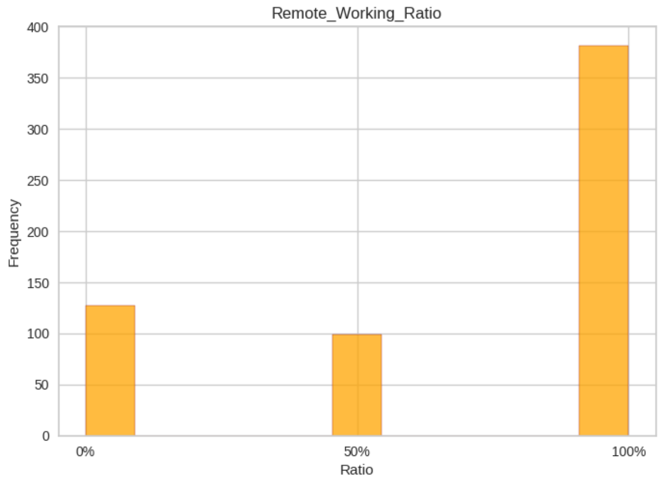

Figure 11.

Discussion

Our initial hypothesis stated that there was a strong relationship between the inputs (experience, years at a company, company size, etc.) and the output, which was the salary attached to a position. Years of experience, years at a company, and location all displayed a strong relationship when predicting salary. Other independent variables (remote working ratio, company size, education level) didn't show as much correlation to salary as we had hoped they would.

After interpreting some of the results of the other models we created, such as decision trees and SVMs, we concluded that noise was too present in our data for some of these models to be effective. We should have spent more time cleaning the data to get a set that would've more accurately predicted relationships between some of the other variables. Because location has such a strong impact on salaries in the profession due to the large variance in international economies, it made it hard to portray other relationships. Our datasets included salaries collected from companies and employees around the world, but we should've limited cleaned the data to where the salaries were only collected from one state or region in the U.S. to decrease the variability due to location.

Regression models were much more effective at predicting salary as a continuous, dependent variable than our classification models that attempted to predict remote working ratio or company size of a single employee. This was most likely again due to the noise from the dataset than can be attributed to location and the variance of global economies.

Future Work

Our methodology for predicting salaries in the data science domain has several shortcomings:

1. Challenges in learning and implementing new machine models in class: We propose allocating dedicated time for understanding and experimenting with newly learned models, and seeking guidance from instructors and peers for proper implementation.

2. Balancing model complexity and predictive accuracy to avoid overfitting/underfitting: We suggest using techniques such as cross-validation, regularization, and pruning, along with monitoring performance metrics and model complexity during training.

3. Ensuring accurate salary predictions: We recommend incorporating additional relevant features and experimenting with various machine learning algorithms, as well as refining the model through hyperparameter tuning and feature selection.

4. Difficulty in integrating and correlating different variables: We propose employing advanced statistical techniques like multivariate regression analysis or dimensionality reduction methods, such as PCA, to uncover hidden relationships between variables and gain a deeper understanding of underlying patterns.

5. Limited numerical values for in-depth analysis: To overcome this limitation, we suggest augmenting the dataset with additional data sources or employing techniques like data imputation or resampling to derive more meaningful insights from the available data.

Conclusion

By using both linear and multiple regression models with various inputs, we found that location has perhaps the greatest impact on the salaries of data science professionals, followed by years of experience and years at a company. Salary can vary greatly due to location and differences in the strength of a country's or region's economy when compared to others. There most likely is some level of correlation between other variables and salary, but noise in the dataset must be further reduced to bring those relationships to light. To effectively create models that can display these relationships, data must be collected from one specific geographical location, such as a large city or single state, and analyzed independently of other geographical locations.

By analyzing these results, others can easily understand that geographic location is the strongest indicator of salary in the broadest sense, but once the scope of a dataset is cut down into individual geographic locations that have consistent economic characteristics, more relationships between salary and other variables, such as remote working ratio and company size, will become more apparent.

REFERENCES

Chauhan, A. (2022). Data Science Fields Salary Categorization. www.kaggle.com. https://www.kaggle.com/datasets/whenamancodes/data-science-fields-salary-categorization

Gopal, K., Singh, A., & Sagar, S. (2021). Salary Prediction Using Machine Learning. INTERNATIONAL JOURNAL of INNOVATIVE RESEARCH in TECHNOLOGY, 8(1), 380–383. https://ijirt.org/master/publishedpaper/IJIRT151548_PAPER.pdf

Ogozaly, J. (2021). Data Science and STEM Salaries. www.kaggle.com. https://www.kaggle.com/datasets/jackogozaly/data-science-and-stem-salaries

Quan, T. Z., & Raheem, M. (2022). Salary Prediction in Data Science Field Using Specialized Skills and Job Benefits – A Literature Review. *Journal of Applied Technology and Innovation*, *6*(2600-7304), 70–74.