Prediction of Salary Ranges from Job Postings

Samdeet Khan
Brown University DSI
https://github.com/samdeet-khan/DATA-1030-Final-Project

Introduction

The current job market has experienced an unprecedented level of applications for positions spanning a wide variety of industries. Many postings for these open roles fail to include a salary range, despite this criteria being an important consideration for job applicants. If a candidate has a desired salary range, the process for devoting time to apply for relevant positions can become more streamlined. From the recruiter's perspective, the ability to set a relevant salary range based on certain criteria can be helpful in ambiguous contexts, and also to deal with less potential candidates who filter based on their salary requirements.

This project utilizes certain linear and non-linear machine learning models to address this gap in knowledge by understanding which relevant job posting details (e.g. location of job, size of the company, and required skills) have the most impact on the expected salary range, denoted by an average of the minimum and maximum salary. This is a regression problem due to the prediction of a salary range, a continuous variable, based on other primarily categorical variables in the dataset. The data utilized for this project was sourced from Kaggle, where the author collected data from LinkedIn job postings and organized relevant variables for ease of analysis.

EDA

For the exploratory data analysis, I wanted to understand the nuances of the current job market and to gain some insight into which variables had the greatest influence on average salary outcomes. This began with an examination of location's relevance in this process.

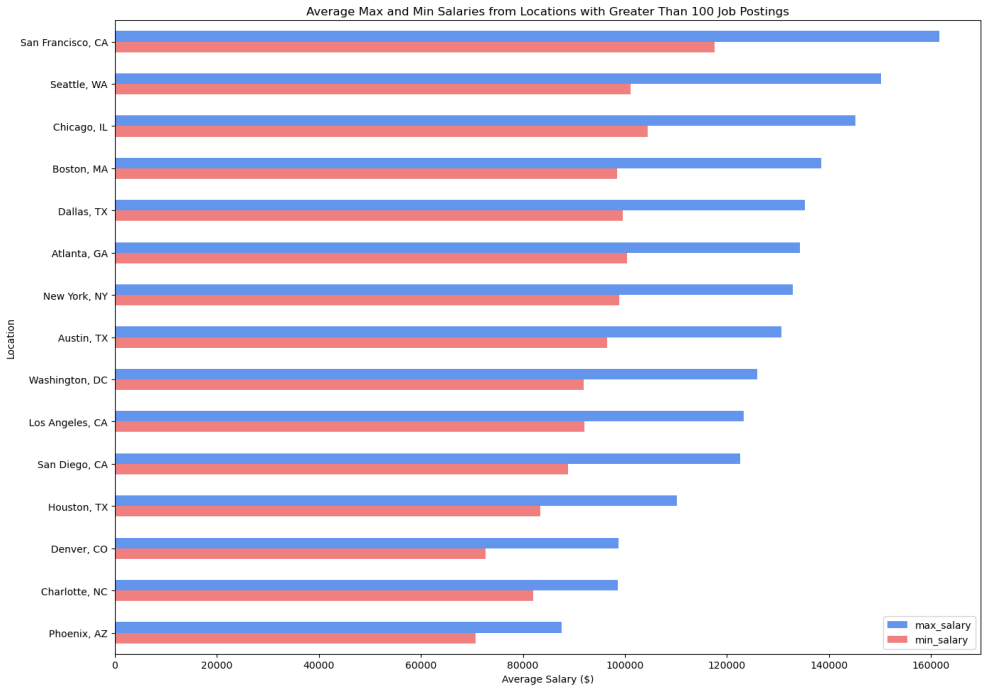

Figure 1. Histogram chart for the average max and min salaries from locations with greater than 100 job postings

Figure 1 shows the cities in the United States that have the most job postings available (greater than 100 in the dataset) on LinkedIn and their relative maximum and minimum salaries. The presence of tech companies and the high cost of living in areas like San Francisco and Seattle reflect their presence in the higher positions, while the presence of certain metropolitan hubs such as Phoenix demonstrates a lower salary range and associated cost of living. After analyzing location, I examined the relevance of skills in determining salary outcomes.

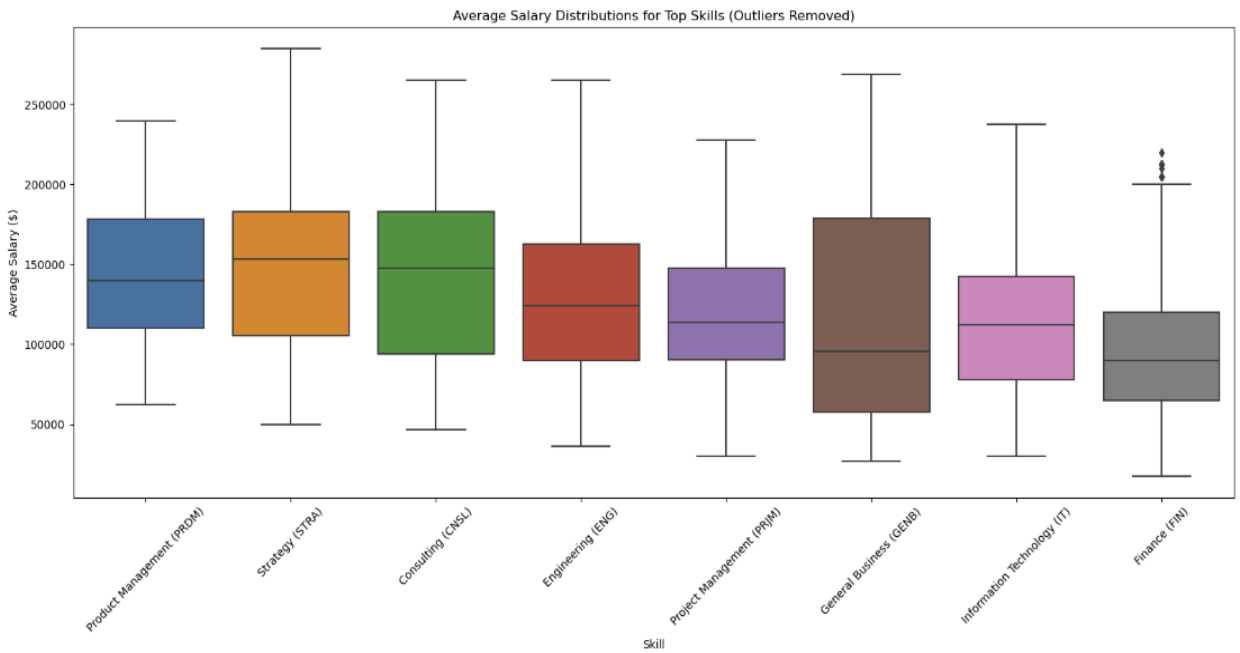

Figure 2. Boxplot range for average salary distributions for top eight skills (outliers removed)

Figure 2 shows the salary distributions for the top eight skills based on associated average salaries. The high interquartile range of the "General Business" skill makes sense given the wide variety of roles at both the entry and executive level. Engineering, on the other hand, has a narrower range of salaries but a higher floor, indicating the greater salary and competition associated with entry level positions.

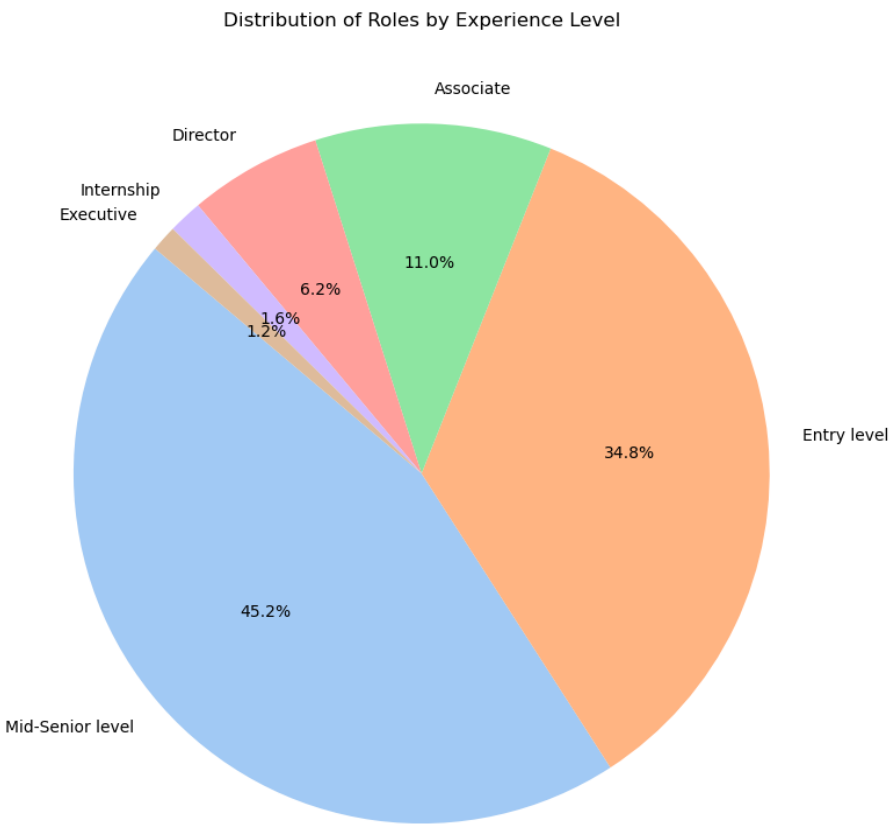

Figure 3. Pie chart distribution of roles by experience level

Figure 3 shows the distribution of roles by experience level, demonstrating the large presence of entry level and mid-senior level positions in comparison to the more executive level and director roles. This outcome can be attributed to the reliance on internal promotions for such roles and the generally greater number of entry to mid level positions that exist within a company.

Methods

The relevant features in the dataset are the following:

- company_id: categorical variable
- location: categorical variable
- formatted_work_type: categorical variable
- formatted_experience_level: categorical variable
- skill_abr: categorical variable
- Remote_allowed (0 - no, 1 - yes): binary variable
- Company_size (1 - 7): ordinal variable
- mean(min_salary and max_salary): target variable

Due to the small size of the dataset and its fulfillment of the i.i.d. assumption, the k-fold cross validation method became the preferred choice to split the dataset into five separate folds. The categorical variables required one-hot encoding while the binary and ordinal variables did not require further processing. The target variable is the average of the minimum and maximum salary for a job due to the convenience of interpreting a single value to represent the range.

The dimension of the dataset prior to the preprocessing stage was (7234,7); following the preprocessing steps, this grew to (7234, 1794). The increase in the feature size can be attributed to the categorization of specific values as features in the dataset (e.g. location_phoenix, company_size_7). The formatted_experience_value column had about 28% of its values missing, so an imputer was applied to add a placeholder "N/A" value. This made the corresponding features usable for the machine learning models.

The four ML models utilized in this project are the Ridge Regression, Lasso Regression, Random Forest, and XGBoost. The combination of linear regression and nonlinear methods (Random Forest and XGBoost) provides a comprehensive approach to determining the best model for analyzing the data further. Prior to any hyperparameter tuning, I wanted to understand the test scores for the models when applied to the preprocessed dataset. This yielded the following results:

	Ridge Regression	Lasso Regression	Random Forest	XGBoost
MSE	657,620,026.19	1,085,318,427.50	1,180,352,355.20	1,357,245,364.64
R-squared	0.83	0.72	0.69	0.64

Figure 4. Model performances with MSE (pre-hyperparameter tuning)

The large MSE values concerned me until I understood that they represented the squared values of the difference between the mean and predicted values, which were generally in the tens of thousands of dollars. To make the test score more interpretable in the context of my problem and for better comparison among the different models, I decided to change from MSE to RMSE for the test score. This generated the following results:

	Ridge Regression	Lasso Regression	Random Forest	XGBoost
RMSE	25,644.10	32,944.17	34,356.26	36,840.81
R-squared	0.83	0.72	0.69	0.64

Figure 5. Model performances with RMSE (pre-hyperparameter tuning)

Based on the various R-squared and RMSE values, the Ridge regression model seemed to perform best prior to the application of any hyperparameter tuning. With this assumption, I proceeded to optimize the parameters for each machine learning model. The following values were tested for each parameter, with the bold values representing those that yielded the best test scores.

Ridge
- alpha ([0.01, **0.1**, 1, 10])
- GridSearchCV

Lasso
- alpha ([0.01, 0.1, **1**, 10])
- RandomizedSearchCV

XGBoost
- regressor__n_estimators: [100, 200, **500**]
- regressor__learning_rate: [0.01, 0.05, **0.1**]
- regressor__max_depth: [3, 5, **7**]
- regressor__min_child_weight: [**1**, 3, 5]
- regressor__gamma: [0, **0.1**, 0.2]
- regressor__subsample: [0.7, **0.8**, 1.0]
- regressor__colsample_bytree: [**0.7**, 0.8, 1.0]

Random Forest
- max_depth: [3, **5**, 10]
- min_samples_split: [2, 5, **10**]
- min_samples_leaf: [1, 2, **4**]

Results

After hyperparameter tuning the models, each method yielded better performances in terms of lower RMSE and higher R-squared values, but the Ridge Linear Regression still outperformed the others, including the baseline model which was established from the existing min_salary and max_salary values in the dataset:

	Ridge	Lasso	Random Forest	XGBoost	Baseline
RMSE	25,405.15	25,613.24	33,006.17	35,260.68	40,806.26
R-squared	0.86	0.86	0.70	0.72	N/A

Figure 6. Model performances with RMSE (post-hyperparameter tuning)

The XGBoost and Random Forest methods are nonlinear and can capture non-linear relationships in data more effectively than the linear models like Lasso and Ridge regression. However, their relatively poorer performance in this context might be due to the variables in the data demonstrating more linear relationships. For example, a linear model might better capture how a certain experience level, such as 'director', directly impacts the salary range. Despite this, each model performs better than the baseline in terms of the RMSE test score, indicating their stronger performance in predicting the average salary values. The relative performances of these models is also demonstrated in the following graph:

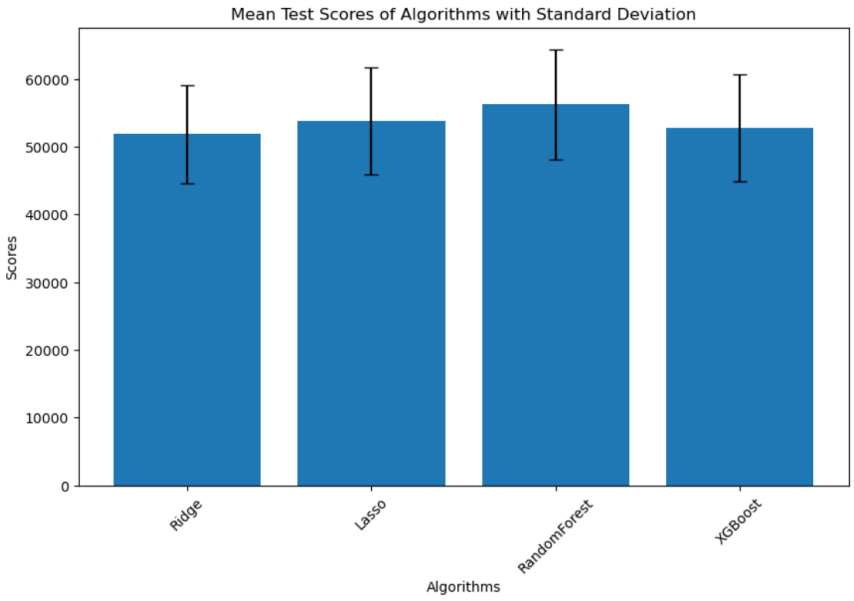

Figure 7. Mean test scores of machine learning models with standard deviations

Figure 7 shows the mean test scores of each machine learning model with the corresponding standard deviation error bars. A lower test score, or RMSE value, indicates a stronger performance, so the Ridge model still demonstrates its ability to outperform the other models for the given dataset. Based on these outcomes, the data pipeline was established using the Ridge model, represented by Figure 8.

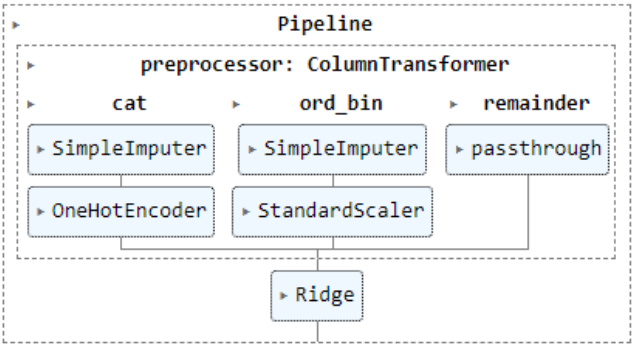

Figure 8. Data pipeline architecture

Figure 9 compares the true average salary to the predicted values from our Ridge Regression model and a simple baseline model. The x-axis shows the real salaries, while the y-axis shows what our models think those salaries should be. The red line is where the predicted salaries perfectly match the real salaries:

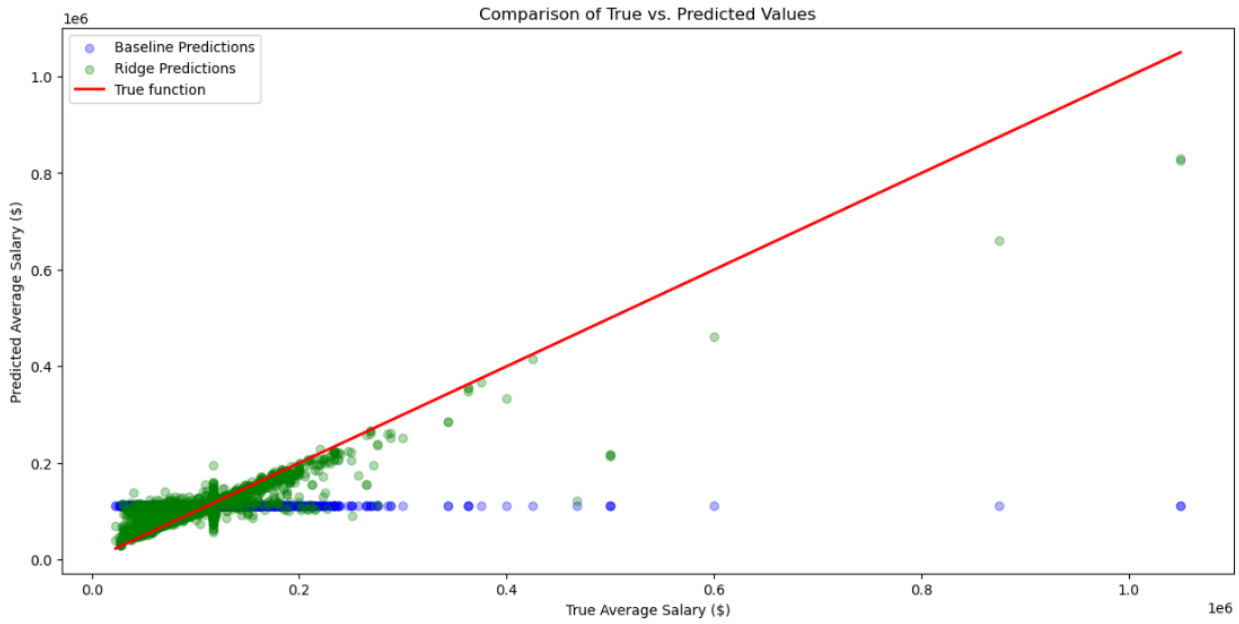

Figure 9. Comparison of Ridge model (green) vs baseline model (blue) to the true function (red)

The green dots from the Ridge model are closer to the red line than the blue baseline dots, which means Ridge is predicting more accurately. A few green dots are too high or too low, showing where Ridge missed the mark, but not as many exist as with the baseline model. The Ridge model is doing a better job than the baseline, with predictions that are usually closer to the actual average salaries. This matches the outcomes from before; Ridge gave the best test score results prior to any fine-tuning of the machine learning models.

Understanding its strength with the dataset comprehensively, I proceeded to calculate the SHAP values for global feature importance. For a small dataset, it seemed more appropriate to focus on the general trends and variables with high impact on the value for the average of min_salary and max_salary.

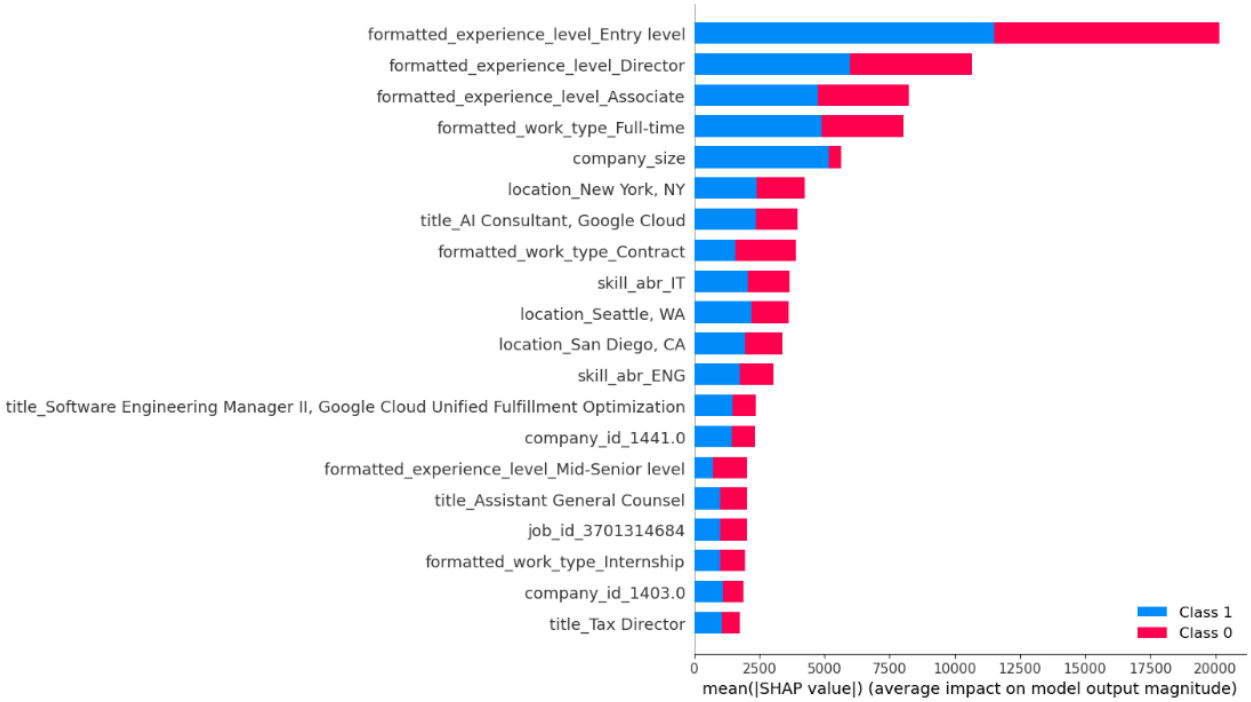

Figure 10. SHAP value summary plot

In Figure 10, “formatted_experience_level”, “location”, and “company_size” demonstrate the most presence in terms of impact on the model’s output, or the average salary value. Based on this insight, I calculated the global feature importance values for the three variables and obtained the following metrics:

- Importance of location: 13,582,569.81
- Importance of formatted_experience_level: 69,402.58
- Importance of company_size: 1,836.09

The values show that location plays the most significant role in the determination of salary outcomes, which makes sense given that the cost of living and the concentrated presence of tech companies in certain areas make the expected salary range higher. The experience level is also significant and should be considered when factoring the expected salary range. An entry-level role is unlikely to demand the same salary as a director or associate, but the industry of the position is also relevant. For example, as I mentioned in my EDA when examining the boxplot of skills, the floor for engineering salaries is higher than that for business, although business skills tend to favor careers with higher ceilings due to their relevance at the management level. Company size contributes, but not as significantly as location and the experience level, indicating the willingness for smaller companies to pay competitive salaries in order to attract top talent for intensive, specialized work.

Outlook

Although my analysis provided results that one might expect when evaluating factors that play a role in salary determination, the numbers helped to understand and relatively weigh which factors play a more significant role. By focusing on specific variables like location, experience level of the position, and size of the organization offering the job, an applicant may have a rough idea of the salary range they can expect to apply for, even if the salary range is not readily available within the job posting.

The computational limitations hindered my analyses in certain steps. For example, when optimizing the various machine learning models, I opted for a RandomizedSearchCV rather than GridSearchCV for the Lasso model. Although the Ridge linear regression performed well, the Lasso model may have outperformed it with more intensive hyperparameter tuning on the alpha value. With further time, I would also like to analyze specific features of the dataset using SHAP values rather than take a global feature approach. This would allow me to understand the nuances of specific interactions between variables on the prediction, such as how a combination of certain variables like location and size of company can correlate with lower or higher salaries.

Optimizing a more effective baseline model that goes beyond simply averaging the existing min_salary and max_salary values in the dataset may yield more effective results. This can help to build a more robust ML model for comparison and set a higher benchmark for the ML models to surpass, offering a clearer insight into the value added by approaches like Ridge and Lasso.

References

Kon, A. 2023. "LinkedIn Job Postings Dataset." Kaggle. Accessed on October 5, 2023. Available at: [https://www.kaggle.com/datasets/arshkon/linkedin-job-postings/]