Final Report

Salary Prediction

CS479: Data Science Course Project

Authors: Dawson Burgess, Lucien Lee, Alphonse Crittenden, Nyah Nelson, Bryan Frahm

Professor: Dr. Xiaogang Ma, Associate Professor at Uidaho CS Dept.

Github: https://github.com/dawson-b23/CS479_Final_Proj/tree/main

Table of Contents

Investigation

Data Science Goal

For our investigation, we aimed to understand factors affecting job salary with the goal of developing models to predict a data science job salary based on job description details such as company ratings, founding year, location, job description, etc. This investigation was driven primarily by a curiosity on potential jobs for soon to graduate computer science students, with an emphasis on data science. We observed a jobs dataset with job information from Glassdoor in 2017-2018 due to its relevance to the goal, the richness of features provided, and the wide variety of relations. This dataset is publicly available on Kaggle and was very trivial to export and start modeling on. The dataset is managed in a CSV format, where the file contents were very organized and the file access was very easy for all members of our team.

Dataset and EDA

The dataset comes in a CSV file with 742 records and 33 attributes. This file format allowed for easy access and use with R and Python modeling techniques. Not all attributes in the dataset may be relevant to finding our goal, so we did an exploratory data analysis (EDA) to find some significant relationships before delving into more advanced modeling. We chose the feature 'average salary' as the y value and used all other 32 features as the x value for the EDA.

A majority of the features in this dataset have categorical values which hindered the analysis process due to the final goal of salary prediction, which is a numeric value. Some functions are not compatible with categorical and numeric values, so this was a difficult process to find another method that could still tell us some information of the significance of the relationships. For the categorical attributes, we used ANOVA or analysis of variance to calculate the p-value of the features and the average salary. Based on the low p-values, we found that the attributes industry, seniority, and location all had a significant relationship to the average salary of a job due to their p-values of $2.64\text{x}10^{-16}$, $2\text{x}10^{-16}$, and $2\text{x}10^{-16}$ respectively. For the numeric attributes, we used the cor() function in R to calculate the correlation coefficients. Based on a correlation coefficient closer to 1 or -1, we found that the attributes Python and hourly both had the strongest linear relationship with average salary due to their correlation coefficients of .32 and -.36 respectively. Regardless of the challenge of the types of features, the EDA told us some important information about our initial assumptions and allowed for a more intuitive process when working on the advanced data analysis.

Metadata

The metadata for this dataset is exported as a json file using the Croissant metadata format for ML datasets. All Kaggle datasets use Croissant format, along with Hugging Face and OpenML. This format is particularly useful for ML datasets because of the integration of frameworks including TensorFlow, PyTorch, and JAX that can easily load Croissant datasets. A feature of this metadata is that it makes it easier to search for datasets using keywords, so the process of finding this dataset based on keywords “salary prediction” was very straight-forward. Some key attributes in the metadata that were integral to our analysis were dataset context and description, license, column names and descriptions, creator, and publisher. Our team exported the following metadata for this dataset, which can be accessed via our teams Github page https://github.com/dawson-b23/CS479_Final_Proj/tree/main:

```
{
   "Dataset": {
     "Name": "Salary Dataset from Glassdoor",
     "Description": "This dataset contains information about job listings, including
job titles, company details, job locations, salary estimates, company ratings, and
more. It aims to provide insights into salary trends across different industries and
locations.",
     "Source": "Kaggle",
     "URL":
"https://www.kaggle.com/datasets/thedevastator/jobs-dataset-from-glassdoor",
     "Format": "CSV",
     "Size": "Approximately 10,000 records",
     "License": "Kaggle License",
     "License URL":
"https://www.kaggle.com/datasets/thedevastator/jobs-dataset-from-glassdoor",
     "Columns": [
       { "Name": "Job Title", "Description": "Title of the job listing" },
       { "Name": "Company Name", "Description": "Name of the hiring company" },
       { "Name": "Location", "Description": "Location of the job" },
       { "Name": "Headquarters", "Description": "Headquarters location of the hiring
company" },
       { "Name": "Size", "Description": "Size of the hiring company (e.g., small,
medium, large)" },
       { "Name": "Type of Ownership", "Description": "Ownership type of the hiring
company (e.g., public, private)" },
       { "Name": "Industry", "Description": "Industry of the hiring company" },
       { "Name": "Sector", "Description": "Sector of the hiring company" },
       { "Name": "Revenue", "Description": "Revenue of the hiring company" },
       { "Name": "Competitors", "Description": "Competing companies of the hiring
company" },
       { "Name": "Job Description", "Description": "Description of the job role" },
       { "Name": "Salary Estimate", "Description": "Estimated salary range for the
job" },
       { "Name": "Rating", "Description": "Company rating on Glassdoor" },
       { "Name": "Founded", "Description": "Year the hiring company was founded" }
     ],
     "Missing Values": "Handled during preprocessing",
     "Data Preprocessing": "Preprocessing steps included handling missing values,
extracting salary estimates, encoding categorical variables, and scaling numerical
features.",
     "Use Case": "Suitable for predicting salaries based on various job and company
attributes, exploring salary trends across industries and locations, and analyzing
factors influencing job satisfaction and company ratings."
   }
 }
```

Data Analysis

Questions and Hypotheses Driving Collection

Question 1

Is there a correlation between the company rating and average salary?

Preliminary Analysis: Utilize the provided dataset on company ratings and salaries to visually explore the relationship between company ratings and average salaries through scatter plots and correlation coefficients. Analyze any initial patterns or trends observed in the data.

Full Analysis: Applyrandom forest regression analysis to quantify the correlation between company ratings and average salaries. Train a regression model using company ratings as independent variables and average salaries as the dependent variable. Evaluate the model's performance using the metrics mean squared error and R-squared score.
Post Analysis: Interpret the regression results to determine the strength and significance of the correlation between company ratings and average salaries. Discuss potential implications of the findings, considering factors such as the influence of company reputation on salary negotiations and employee satisfaction.

Answer 1

The Mean Squared Error (MSE) measures the average squared difference between the predicted and actual minimum salaries. In this case, the MSE value of 903.79 indicates that, on average, the predicted minimum salaries are about $903.79 away from the actual minimum salaries.

The R-squared score, which is 0.081, represents the proportion of the variance in the minimum salaries that is predictable from the company ratings. An R-squared value closer to 1 indicates that the model explains a large proportion of the variance in the target variable. However, in this case, the R-squared value is relatively low, suggesting that the company rating alone may not be a strong predictor of minimum salary, as it explains only about 8.1% of the variance.
(See Figure 4)

Question 2

Is there a correlation between certain job descriptions, titles, or words that result in an increased salary?

Preliminary Analysis: Utilize natural language processing techniques to preprocess and analyze job descriptions from the provided dataset. Identify relevant keywords or phrases associated with high-salary positions through frequency analysis or sentiment analysis. Visualize the distribution of salaries based on job titles or keywords.
Full Analysis: Employ regression analysis to quantify the correlation between specific job descriptions, titles, or keywords and salary levels. Train a regression model using job descriptions or title features as independent variables and salary as the dependent variable.

Assess the model's performance and significance of features through metrics such as coefficients and p-values.

Post Analysis: Interpret the regression results to identify significant predictors of increased salary, such as specific skills, responsibilities, or industry sectors mentioned in job descriptions. Discuss potential implications for job seekers in tailoring their resumes and focusing on high-value skills or experiences. Additionally, consider the relevance of these findings for employers in attracting and retaining talent.

Answer 2

Results from linear regression model:

1. **Senior (758.65)**: This indicates that the presence of the word "senior" in the job descriptions has the highest positive impact on the predicted minimum salary. Job descriptions mentioning "senior" likely correspond to positions with higher salaries, as senior roles often require more experience and expertise.
2. **Five (134.28)**: The word "five" appearing in job descriptions may be associated with specific qualifications or requirements, potentially indicating roles with higher levels of responsibility or experience, thus contributing positively to the predicted salary.
3. **Control (122.03)**: Job descriptions mentioning "control" could be related to positions that involve managing or overseeing processes or systems, which may typically command higher salaries due to the level of responsibility.
4. **View (114.35)**: The presence of "view" in job descriptions could be related to aspects such as data visualization or reporting, indicating roles with analytical or technical skills, which are often associated with higher salaries.
5. **Prior (106.81)**: Mention of "prior" suggests that previous experience or knowledge is required for the role, which may correlate with higher salary expectations for positions that demand specific expertise or skills.
6. **Release (105.54)**: This word may be associated with software development or project management roles, indicating involvement in the release process of products or projects, which could be indicative of higher-level roles.

7. **Platform (103.97)**: Presence of "platform" in job descriptions could relate to roles involving development or management of software platforms or frameworks, which are typically associated with higher salaries due to the technical expertise required.
8. **Principal (103.92)**: Mention of "principal" may indicate senior or leadership roles within an organization, which often come with higher levels of responsibility and hence higher salaries.
9. **Optimization (103.75)**: Job descriptions mentioning "optimization" may relate to roles in data science, analytics, or operations, indicating a focus on improving efficiency or performance, which could command higher salaries.
10. **Mentor (103.63)**: This suggests that roles involving mentoring or leadership responsibilities are associated with higher salaries, as mentoring often requires significant experience and expertise.

Overall, the coefficients provide insight into how each word contributes to the prediction of minimum salaries based on the job descriptions. Words with higher coefficients are likely to have a stronger influence on the predicted salary.

Results from Random Forest Regression model:

1. **Machine (0.109130)**: This feature (word) has the highest importance score. In the context of job descriptions, the word "machine" appears to be strongly correlated with variations in minimum salary. Jobs that mention "machine" may involve roles related to machine learning, artificial intelligence, or other technical fields, which typically command higher salaries.
2. **Predictive (0.034019)**: The word "predictive" also has a significant importance score. Job descriptions mentioning "predictive" likely involve roles that require skills in predictive analytics or modeling, which are in demand and often associated with higher salaries.
3. **Model (0.028670)**: The term "model" is another important feature. It suggests that job descriptions referencing models, such as predictive models or machine learning models, are correlated with variations in minimum salary.

4. **Learning (0.024588)**: The word "learning" is also relevant, indicating roles related to machine learning or continuous learning and development, which may command higher salaries due to the specialized skills required.
5. **Give (0.023836)**: This feature may represent roles that involve giving presentations, reports, or insights, which could be indicative of higher-level positions with leadership or communication responsibilities.
6. **Infrastructure (0.022845)**: Jobs mentioning "infrastructure" may involve responsibilities related to the development or management of technical infrastructure, which are often associated with higher salaries due to the specialized expertise required.
7. **Scientist (0.021209)**: The term "scientist" suggests roles related to data science or research, which are typically high-paying positions due to the specialized skills and expertise involved.
8. **San (0.020103)**: This feature may represent locations such as San Francisco, which is known for its high cost of living and typically offers higher salaries compared to other locations.
9. **Expression (0.015859)**: Jobs mentioning "expression" may involve roles related to data visualization, storytelling, or communication, which could be indicative of higher-level positions with specialized skills.
10. **Product (0.014615)**: The word "product" suggests roles related to product development or management, which often require specialized skills and experience and may command higher salaries.

Overall, these feature importance scores provide insights into which words or features in job descriptions are most influential in predicting variations in minimum salary. Words with higher importance scores are likely to have a stronger influence on the predicted salary.

Machine Learning Methods

(See GitHub Page: https://github.com/dawson-b23/CS479_Final_Proj/tree/main)

For the data analysis, we utilized a variety of machine learning methods to explore relationships and patterns within the dataset. Initially, we employed linear regression to investigate correlations between company ratings and average salaries, as well as to analyze the impact of certain job descriptions or titles on salary levels. Subsequently, we extended our analysis to include other machine learning techniques such as random forest regression, gradient boosting, support vector machines, decision trees, k-nearest neighbors, ridge regression, and lasso regression. To further extend analysis, our team utilized PyTorch to craft a custom neural network model to train on, test, and analyze the data. These methods allowed us to capture nonlinear relationships and interactions within the data, providing a more comprehensive understanding of the factors influencing salary outcomes.

To manage the results, we stored the performance metrics and predictions generated by each machine learning model in a structured format, such as dictionaries or pandas DataFrames. This facilitated easy comparison and interpretation of the results across different models. Additionally, we visualized the results using various plots, such as bar plots for comparing mean squared errors and R-squared scores, and scatter plots for visualizing actual versus predicted salary values. These visualizations helped in identifying the strengths and weaknesses of each model and interpreting the overall performance of the machine learning analysis.

Overall, the combination of multiple machine learning methods, along with careful management and visualization of results, allowed us to conduct a comprehensive data analysis and gain valuable insights into the factors influencing salary predictions in the dataset.

Analysis and Validation

(See GitHub Page: https://github.com/dawson-b23/CS479_Final_Proj/tree/main)

To ensure the validity of our analysis, we followed a structured approach with several validation steps. Firstly, we split the dataset into training and testing sets using cross-validation techniques to assess model performance on unseen data. We also applied feature scaling and preprocessing methods consistently across all models to maintain consistency and comparability. Additionally, we employed multiple evaluation metrics such as mean squared error, R-squared score, and visual inspection of actual versus predicted values to assess model accuracy and generalization. Moreover, we conducted sensitivity analysis by varying hyperparameters and model configurations to assess robustness. Finally, we validated our findings by comparing results across different machine learning methods and interpreting the consistency of trends and patterns observed. This rigorous validation process ensured the reliability and trustworthiness of our analysis outcomes.

Visualization

Presentation of Data

For the data presentation, we created visualizations such as histograms, box plots, and correlation matrices to explore the distribution and relationships among variables in the dataset. These visualizations provided insights into the characteristics of the data, including salary distributions, company ratings, and other relevant features.

For the analysis results, we generated visualizations such as bar plots comparing mean squared errors and R-squared scores of different machine learning models. Additionally, scatter plots were used to visualize actual versus predicted salary values for each model, allowing for a qualitative assessment of model performance. These visualizations aided in the interpretation and comparison of results across different machine learning techniques, enabling users to make informed decisions based on the analysis outcomes.
(See Figures 1, 2, and GitHub Page:
https://github.com/dawson-b23/CS479_Final_Proj/tree/main)

Documentation

(See GitHub Page: https://github.com/dawson-b23/CS479_Final_Proj/tree/main)
The presentation and visualization products were managed using a combination of tools and techniques to ensure organization and accessibility. We stored the visualizations in various formats, such as image files or interactive plots, depending on the specific requirements. Metadata such as titles, labels, and descriptions were created and included to provide context and enhance understanding for users seeking to access the visualizations. Additionally, version control systems were utilized to track changes and revisions, ensuring the integrity and consistency of the presentation materials throughout the analysis process.

Visualizations meeting project goals

The presentation and visualization products effectively meet the goal of the investigation by providing clear and insightful representations of the data analysis results. Through histograms, box plots, and correlation matrices, users gain a comprehensive understanding of the dataset's characteristics and relationships between variables. Additionally, bar plots and scatter plots allow for the comparison and evaluation of different machine learning models, enabling users to identify the most effective approach for predicting salary outcomes. Overall, these visualizations add significant value by facilitating informed decision-making and guiding future actions based on the insights gained from the analysis.

Data Management Plan

Overall Plan

Creation of Logical Connections

The data management plan involves establishing logical connections between different datasets, ensuring that relevant variables and relationships are properly linked to facilitate analysis and interpretation.

Physical Data Handling

Data will be stored securely in appropriate physical locations, with backup and redundancy measures in place to prevent data loss and ensure data integrity.

Interoperability Support

Data will be formatted using standard protocols and formats to ensure compatibility with various software tools and platforms, facilitating interoperability and data exchange between different systems.

Security Support

Robust security measures, including encryption, access controls, and authentication mechanisms, will be implemented to safeguard data against unauthorized access, breaches, or tampering.

Metadata collection, management, and access

Metadata was exported as a json file using Croissant metadata format for ML datasets. Our team also created a basic JSON metadata file for users to be able to conduct a quick overview of the dataset/data before deciding whether or not to make use of it.

Persistence

Data will be stored persistently in reliable and scalable storage systems, ensuring its availability and accessibility over time, even in the face of hardware failures or system upgrades.

Discovery

Data will be organized and indexed to enable easy discovery and retrieval, allowing users to quickly locate relevant datasets and information for analysis or decision-making purposes.

Dissemination

Results of the analysis will be disseminated through various channels, including reports, presentations, and visualizations, to communicate findings effectively to users and facilitate knowledge sharing.

Data ownership

Clear ownership and responsibility for the data will be established, ensuring accountability and adherence to data governance policies throughout its lifecycle.

Figures

Figure 1: R-squared scores for varying models

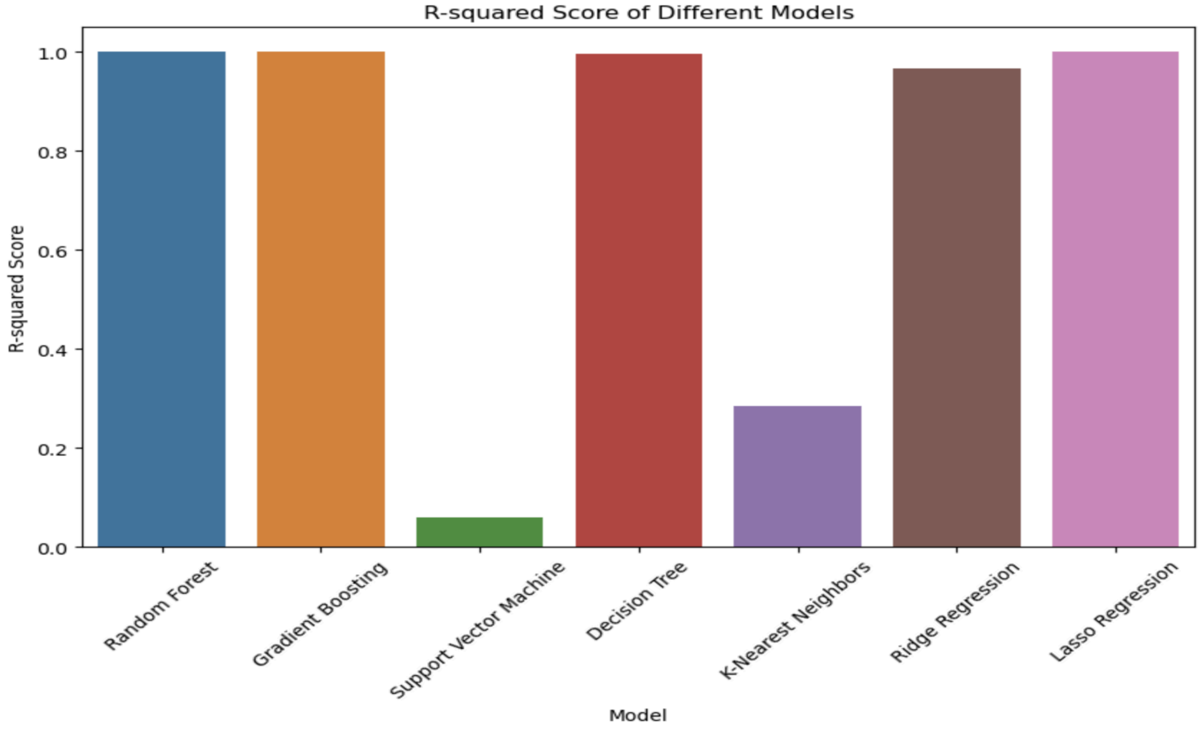


Figure 2: Mean squared error for varying models

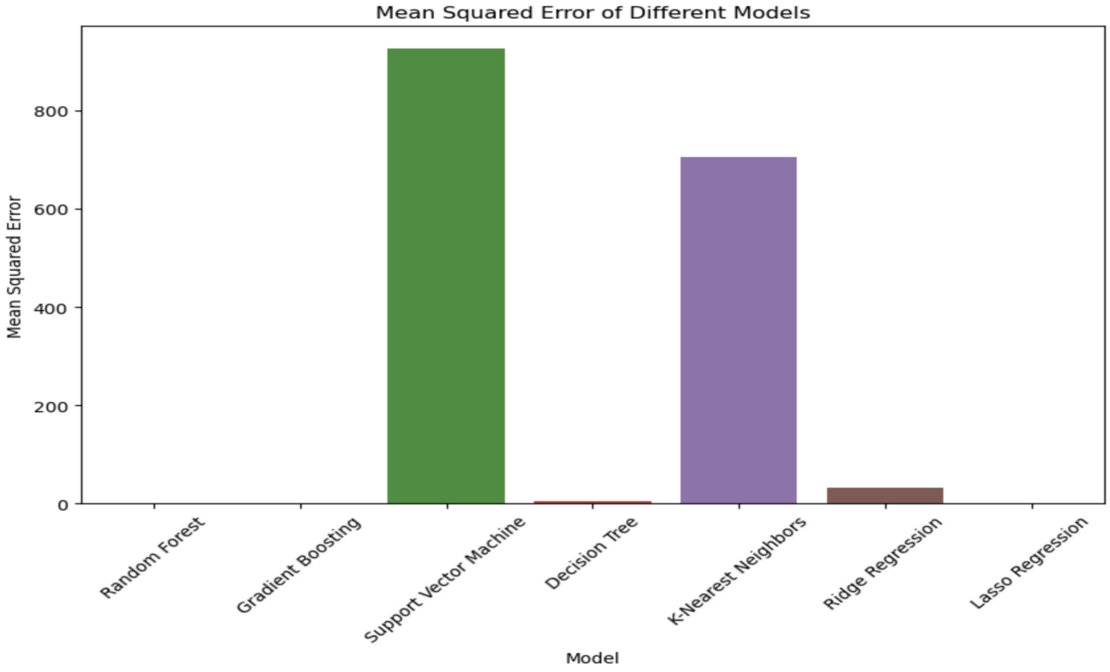
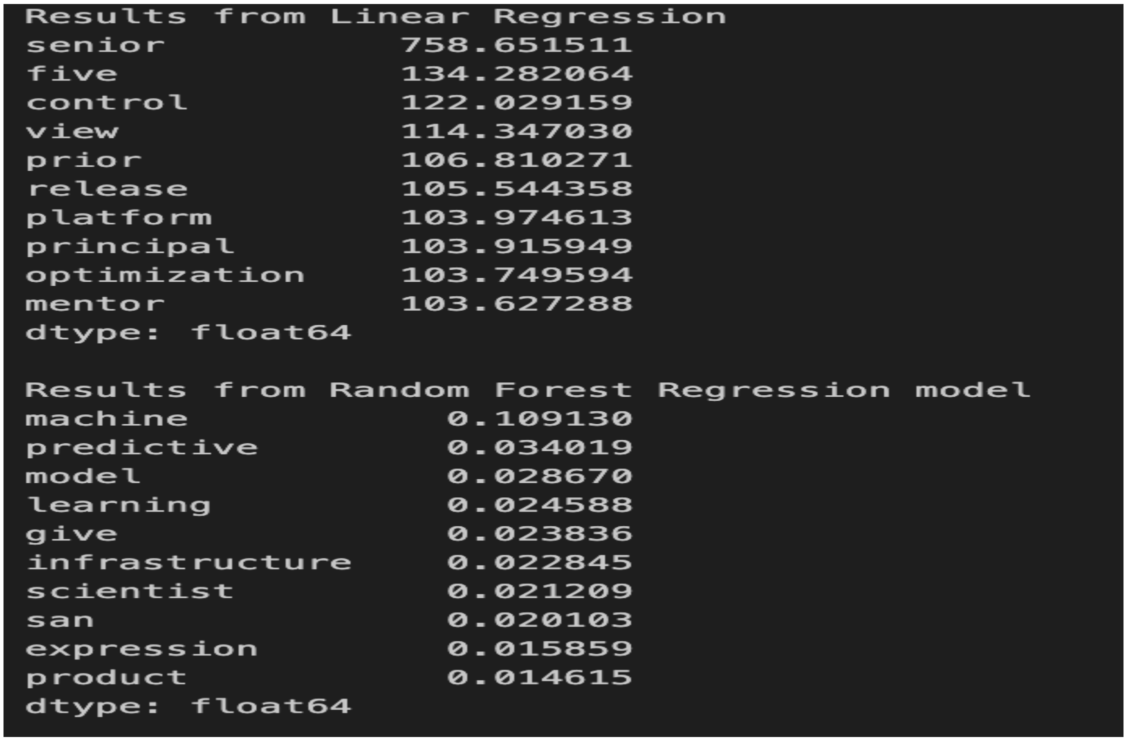


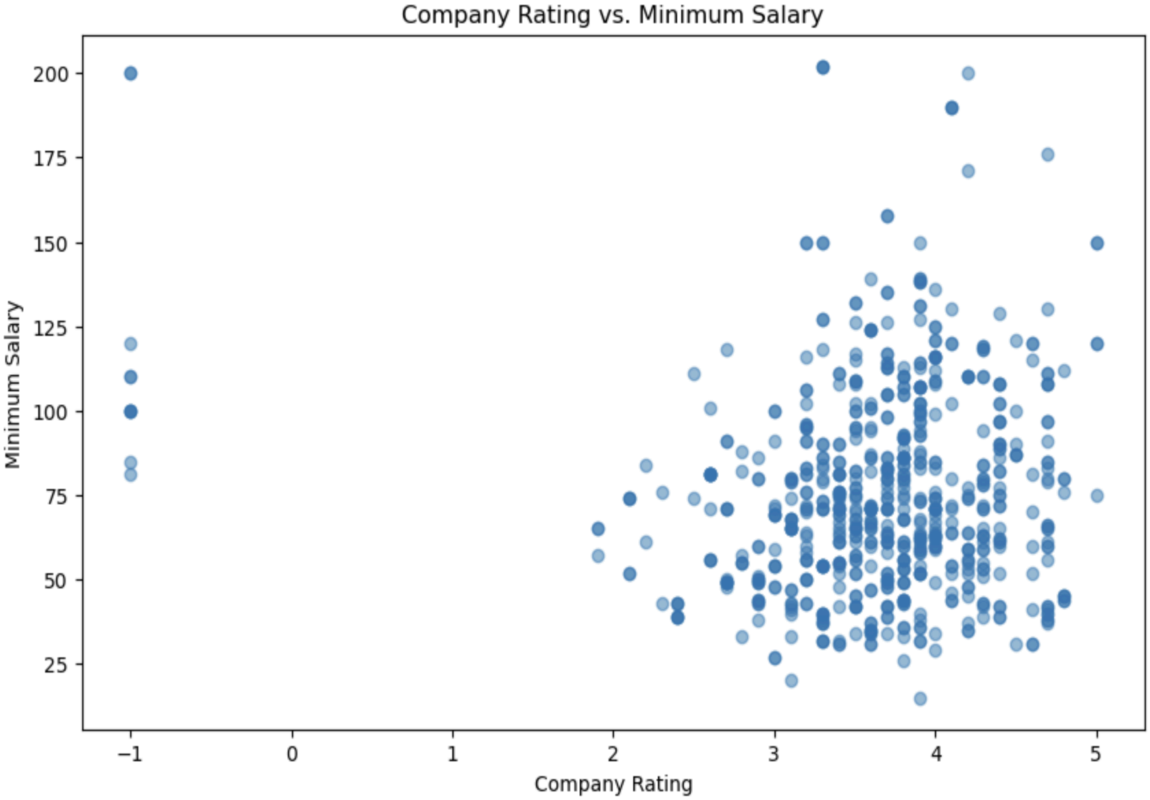
Figure 3: Results from semantic analysis



```
Results from Linear Regression
senior              758.651511
five                134.282064
control             122.029159
view                114.347030
prior               106.810271
release             105.544358
platform            103.974613
principal           103.915949
optimization        103.749594
mentor              103.627288
dtype: float64

Results from Random Forest Regression model
machine               0.109130
predictive            0.034019
model                 0.028670
learning              0.024588
give                  0.023836
infrastructure        0.022845
scientist             0.021209
san                   0.020103
expression            0.015859
product               0.014615
dtype: float64
```

Figure 4: Company Rating vs. Minimum Salary


Company Rating vs. Minimum Salary
Minimum Salary
Company Rating