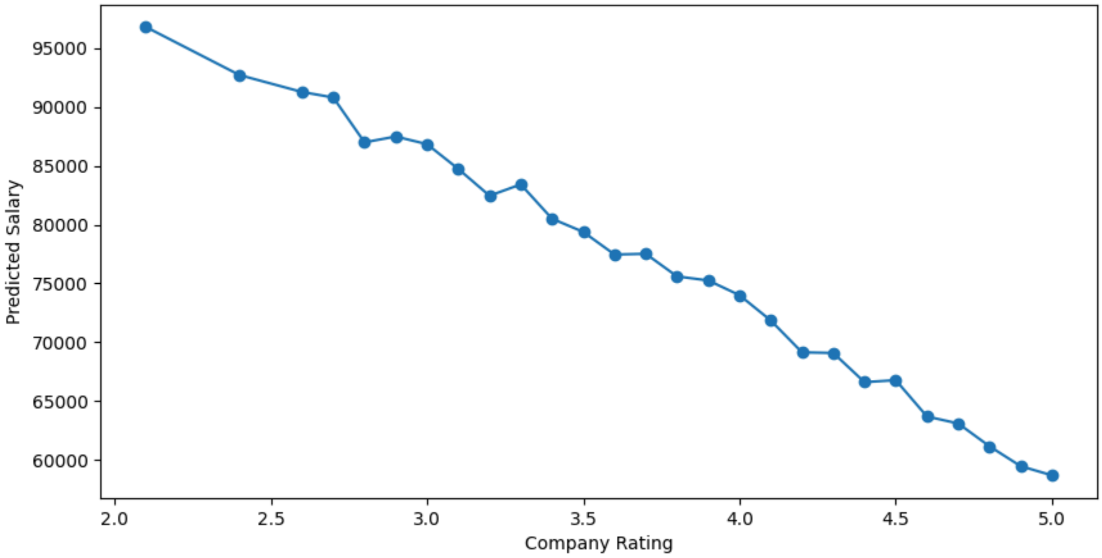
Salary Prediction Report

Introduction

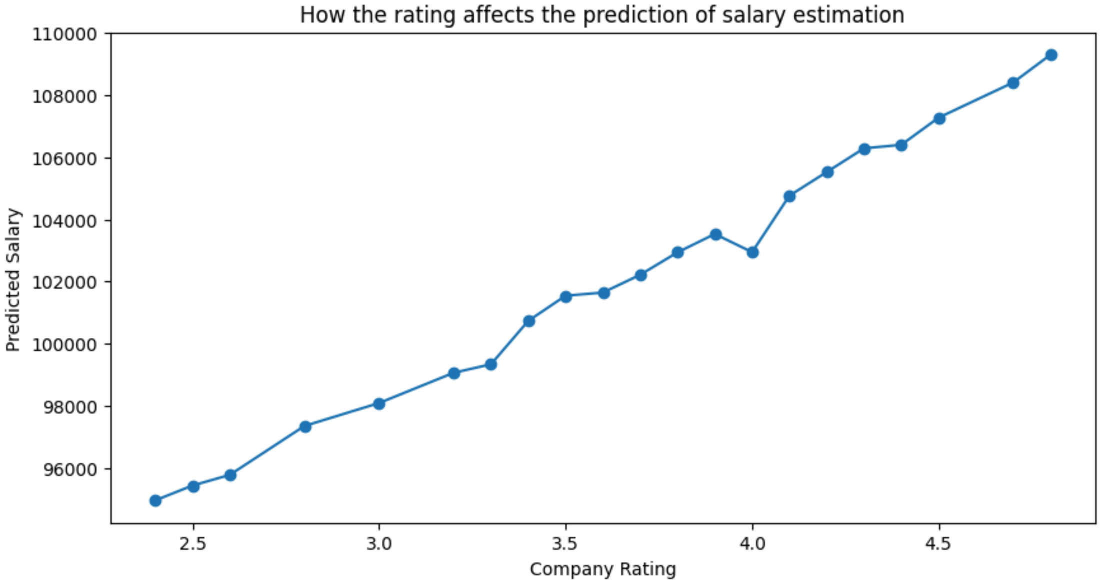
This report presents the development and evaluation of a salary prediction model for data science positions using machine learning techniques. The model is based on Glassdoor job listings, integrating various company-related and job-related factors to explore salary trends.

Data Preprocessing and Feature Engineering

- Parsing Salary Estimate: The salary estimate feature was extracted and processed as the target variable.
- Handling Missing and Invalid Values
- Replacing Invalid Categorical Values: Categorical variables with low variation had their invalid values replaced with the most frequent value.
- Encoding Categorical Features: Applied **label encoding** to selected categorical variables, created dummy variables where necessary.
- Correlation Analysis: Evaluated correlation between numerical features, visualized correlation matrix to detect dependencies.
- Variance Inflation Factor (**VIF**) **Analysis** before model training to ensure that the two selected numerical features (Rating and Founded) were not excessively correlated, preventing multicollinearity issues.
- Outlier Detection and Handling: Utilized the **IQR method** to identify outliers. Visualized distributions of numerical features. Analyzed and handled outliers in each numerical variable separately.
- Detailed Outlier Handling for Salary Estimate: Initial IQR analysis suggested that Salary Estimate had no outliers, yet 264 rows had a salary estimate of zero. Since salary should not be zero, further analysis revealed that companies with zero values included churches, insurance firms, and organizations that might offer unpaid positions. Training the model with these zero values resulted in completely unexpected behavior: higher company ratings corresponded to lower salary estimates.



- After removing these 264 rows, the model exhibited more intuitive behavior and improved stability.



Model Performance and Adjustments

- Model Training and Evaluation

The initial linear regression model showed limitations due to weak correlations between Salary Estimation and input features: Rating and Founded. One key issue was the presence of 264 rows with Salary Estimate = 0, which significantly skewed model predictions. Two different approaches were tested:

1. Retaining Zero Salary Estimates: This led to poor model performance, as the zero values disrupted the prediction accuracy.
2. Removing Zero Salary Estimates: The model improved but still had limited predictive power due to weak feature correlations.

- Regularization Models

Ridge and **Lasso regression** were applied to test if regularization techniques could improve performance. However, the results confirmed that the primary issue was the dataset itself: low correlation between available predictors and salary estimates. .

- SHAP Analysis:

Performed **SHAP** (SHapley Additive Explanations) **analysis** to interpret feature importance. Visualized results using SHAP waterfall plot to understand the impact of different variables on salary predictions.

- Model Evaluation:

Visualized predictions versus actual values to assess model performance. Compared evaluation metrics to determine accuracy and limitations.