Salary Prediction Model Analysis

Introduction

This project aims to build a predictive model for estimating the **total yearly compensation** based on various features such as **stock grant value**, **years of experience**, **bonus**, **job title**, **location**, and **education level**. The dataset contains information about employees' compensation, roles, and qualifications, and we aim to predict the salary (total yearly compensation) using different machine learning techniques.

Data Overview

The dataset includes the following features:

- **totalyearlycompensation**: Target variable (salary).
- **stockgrantvalue**: Stock grant value.
- **yearsofexperience**: Number of years of experience.
- **bonus**: Bonus received.
- **job title**: Various titles such as Product Designer, Software Engineer, etc.
- **education**: Education level, such as Bachelor's, Master's, or PhD.
- **location**: Geographical location (e.g., New York, San Francisco).

Data Preprocessing

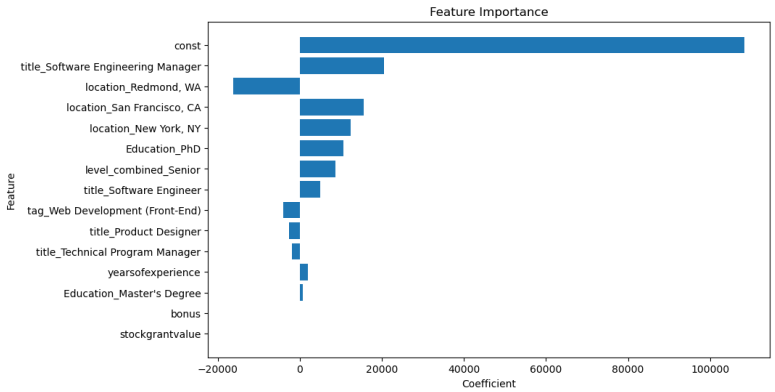
Several preprocessing steps were taken to prepare the dataset:

- Categorical features such as **job title** and **location** were one-hot encoded.
- Null values were handled by appropriate methods.
- Features with high multicollinearity were identified and reduced to avoid overfitting.

Feature Selection

To enhance model performance and reduce complexity, feature selection methods were applied:

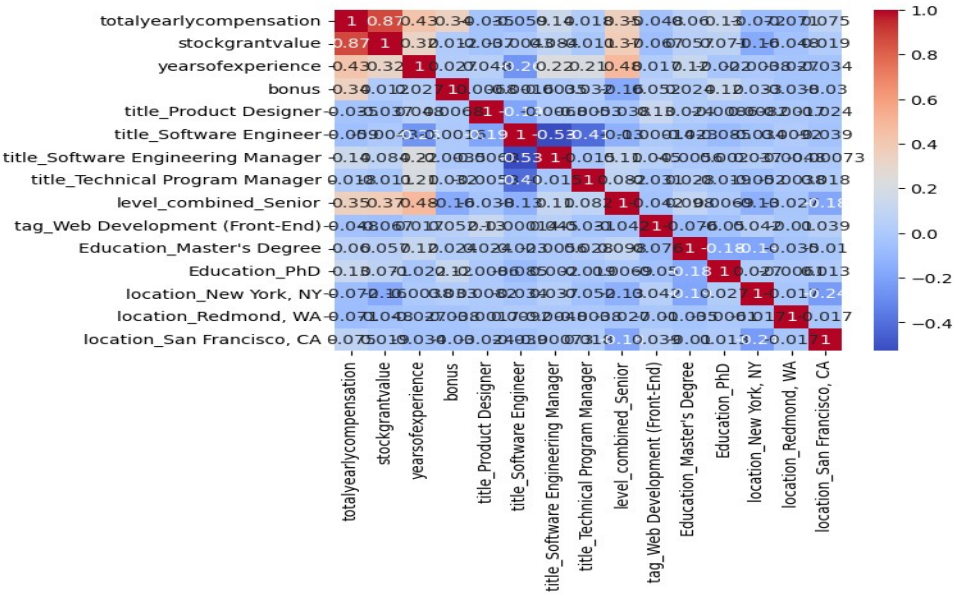
- **Variance Threshold**: Removed features with low variance.
- **Correlation Analysis**: Features that were highly correlated with each other were eliminated.
- **Statistical Tests (SelectKBest)**: The most relevant features were selected based on statistical significance.
- **Forward Selection**: Linear regression was used to add features sequentially.
- **Recursive Feature Elimination (RFE)**: A decision tree regressor was used to eliminate non-important features.



Correlation Analysis

A correlation matrix was generated to examine the relationships between the target variable and features. Some key observations include:

- **Stock grant value** and **total yearly compensation** had a high positive correlation (0.87).
- **Years of experience** had a moderate positive correlation (0.43) with the target variable.
- **Job titles** such as **Software Engineering Manager** and **Technical Program Manager** showed moderate correlations with total compensation.



Model Development

Multiple machine learning models were developed to predict **total yearly compensation**. These models were evaluated using metrics such as **R-squared (R^2)**, **Mean Squared Error (MSE)**, and **Root Mean Squared Error (RMSE)**.

1. OLS (Ordinary Least Squares) Regression

The OLS model was built to understand the linear relationships between the features and the target variable. The results show:

- **R-squared (R^2)**: 0.912, indicating that the model explains 91.2% of the variance in total yearly compensation.
- Significant predictors: **stockgrantvalue**, **years of experience**, **bonus**, and **location**.
- **P-values** for most features were below 0.05, suggesting they are statistically significant.

2. Lasso Regression

Lasso regression was used to regularize the model and prevent overfitting by penalizing large coefficients:

- **R^2**: 0.9015, indicating good predictive accuracy.

- **Mean Squared Error (MSE)**: 334,948,969, showing the error between predicted and actual values.
- The model coefficients show how each feature contributes to the prediction, with **stockgrantvalue**, **years of experience**, and **bonus** being the most impactful.

Model Evaluation

The performance of the models was assessed using both the training and test datasets:

- **Training R^2**: 0.9144, indicating the model fits the training data very well.
- **Test R^2**: 0.9015, which is slightly lower, showing good generalization to new data.
- **Training MSE**: 270,059,102, which is the error when predicting on training data.
- **Test MSE**: 334,953,490, indicating a slight increase in error when applied to unseen data.

This demonstrates that the model performs robustly but may slightly overfit the training data, as seen from the gap between training and test performance.

Conclusion

- The predictive model for total yearly compensation provides strong performance with an **R^2 of 0.91**.
- **Stock grant value**, **years of experience**, and **bonus** emerged as the most important features.
- The model can be improved further by addressing the slight overfitting, possibly through more advanced regularization techniques or feature engineering.

Key Skills and Techniques Demonstrated:

- Data preprocessing and feature engineering
- Feature selection and dimensionality reduction
- Application of statistical tests and machine learning models (OLS, Lasso)
- Evaluation of model performance using R^2, MSE, and RMSE

This project demonstrates my ability to build, evaluate, and refine predictive models, making use of modern machine learning techniques. The insights gained from this model can be useful for HR departments and financial analysts in determining fair compensation based on employee characteristics.