A Causal Inference Study

Impact of Higher Education on Salaries

Seunghyun Park

260686853

Introduction

The hypothesis of my research posits that individuals holding advanced degrees, such as Master's and Doctorates, receive higher salaries compared to their counterparts with bachelor's degrees. To explore this, I divided employees into two groups: those with a bachelor's degree and those with a Master's or Doctorate degree. An initial comparison of the hourly rates between these groups showed no significant differences, leading me to delve deeper using causal inference methods.

To determine the effect of higher education on salaries, I employed causalml library, focusing primarily on two regression methods: the LRSRegressor (Linear Regression) and the XGBTRegressor (XGBoost). Additionally, I explored the T-learner, X-learner, and R-learner models for a comprehensive analysis.

Data EDA and Preprocessing

Histograms

Histograms were generated for all numerical features to understand their distributions. This visualization aids in identifying outliers and the spread of each variable.

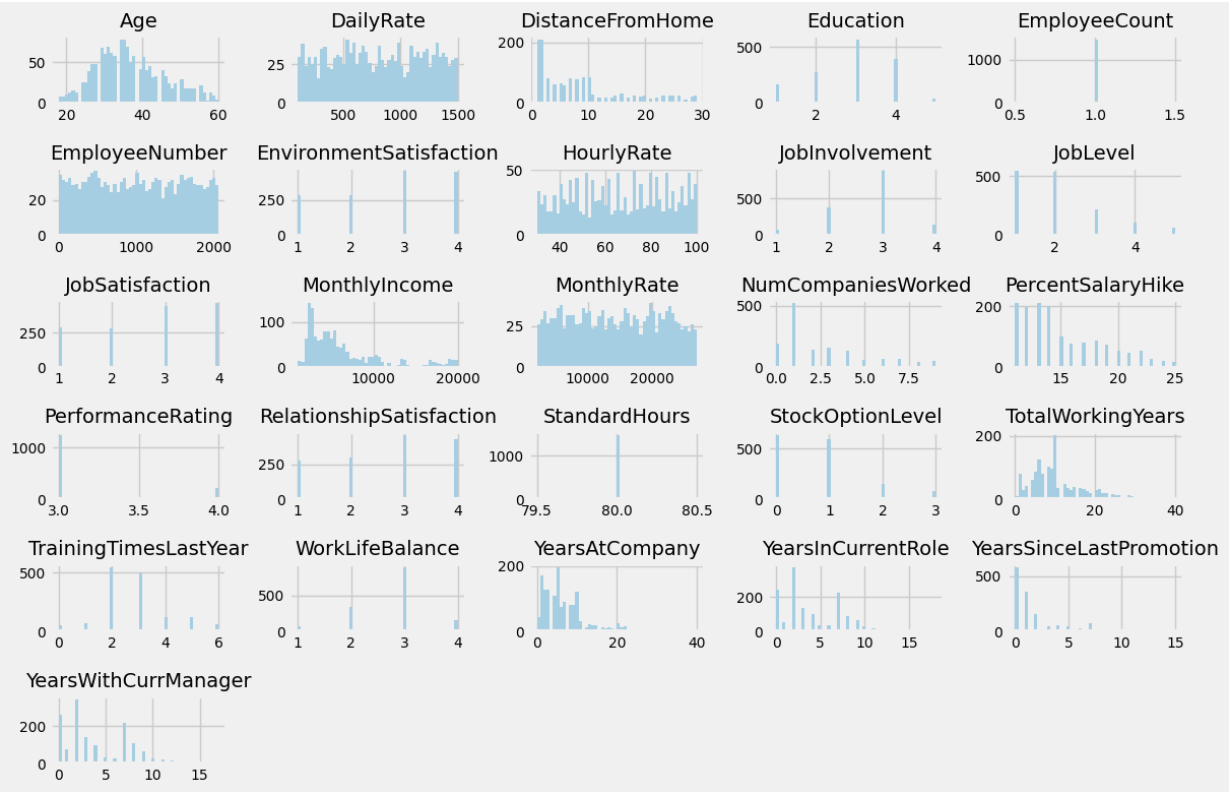

Figure 1: Histograms of Numerical Columns

Pairplot Visualization

Seaborn's pairplot was employed to explore pairwise relationships and distributions.

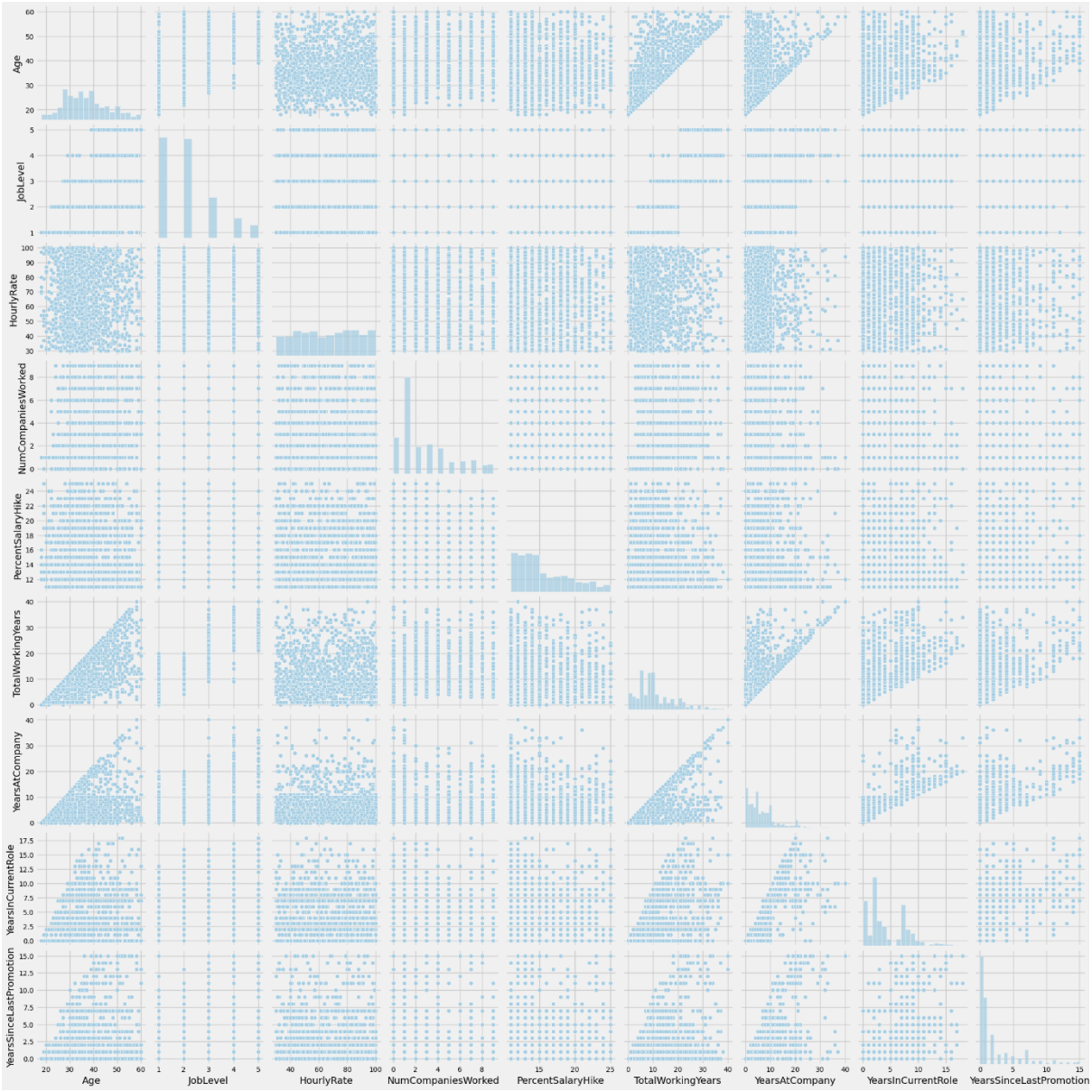

Figure 2: Pairplot of Numerical Columns

Correlation Matrix Heatmap

A heatmap of correlations between numerical variables was created to visualize and assess multicollinearity and potential predictive relationships.

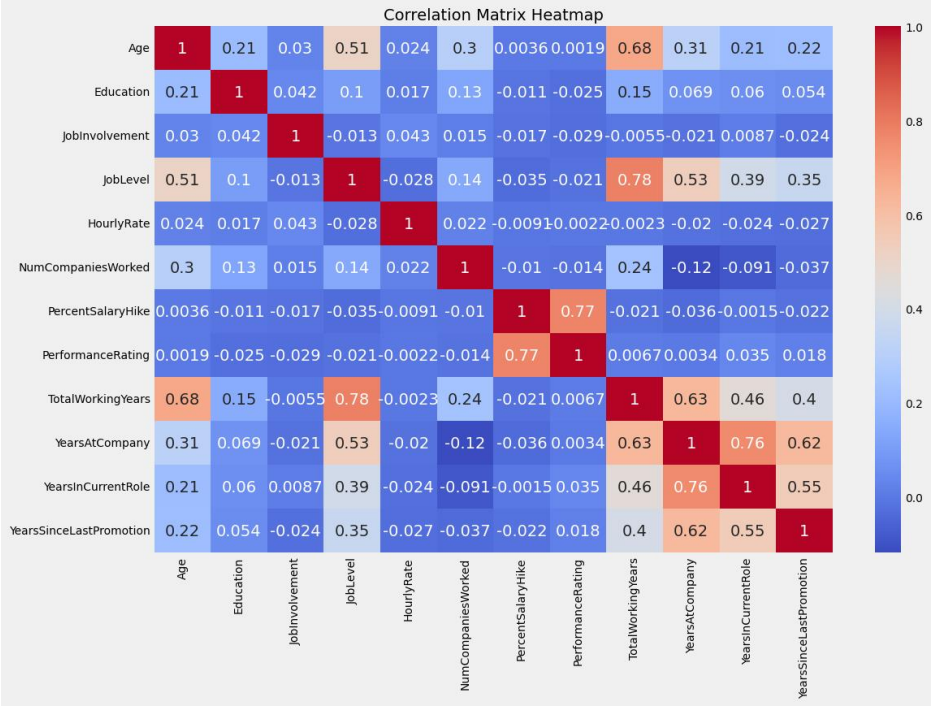

Figure 3: Correlation Matrix Heatmap

Outlier Detection

The Isolation Forest algorithm was utilized for outlier detection, identifying and removing anomalies from the dataset to prevent them from skewing the analysis. This step is crucial for improving model accuracy and generalization.

Encoding Categorical Variables

Categorical variables such as 'EducationField', 'Department', 'Gender', and 'OverTime' were encoded into numerical values using label encoder.

Treatment Effect Analysis Preparation

To investigate the impact of education levels on employee outcomes, an **'Education'** variable was mapped to a binary treatment indicator. Those with a bachelor's degree were mapped as '0' and those with a master's or doctorate degree were mapped as '1'. This categorization facilitates causal inference analysis by distinguishing between treatment and control groups.

Dataset Balancing

The dataset was balanced between the treatment and control groups to ensure an equal representation.

Kernel Density Estimate Plot

Kernel Density Estimate plots for 'HourlyRate' by both group were generated to visually assess the impact of education levels on salaries, further preparing the data for detailed causal inference analysis.

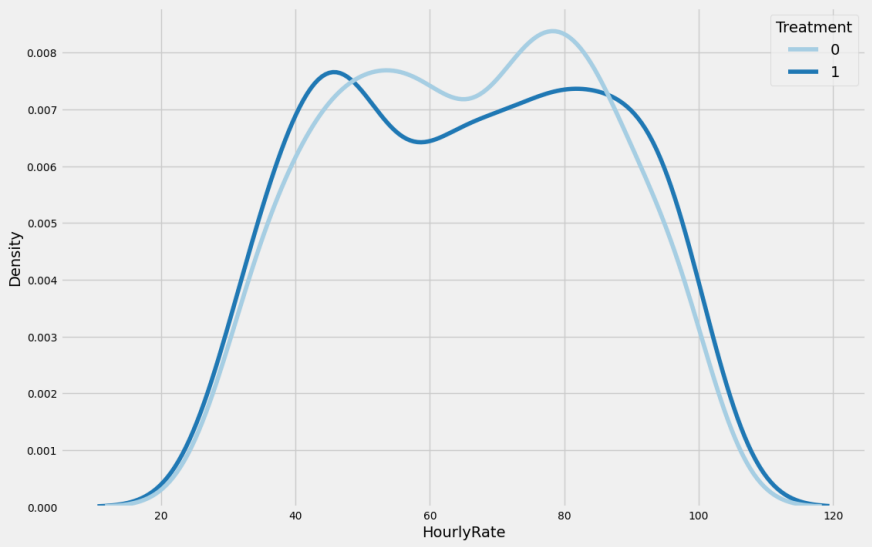

Figure 4: KDE plots for 'HourlyRate'

Methodology and Result

Average Treatment Effect

The Average Treatment Effect (ATE), representing the mean outcome difference between the treatment (higher education) and control groups, varied across models. The ATE from the LRSRegressor, based on a linear model, was 0.0278, while the ATE from the XGBTRegressor, which handles complex non-linear relationships more effectively, was -0.9807. These results suggest a slight influence of obtaining a Master's or Doctorate degree on salary, although the effect is relatively minor given the average hourly rate of $66. The ATE estimates from other models consistently indicated a small effect, with maximum value of -2.5. Linear model-based analysis indicated a positive impact of higher education on salaries, whereas the XGBoost model revealed a negative influence, suggesting a complex relationship between educational attainment and salary outcomes.

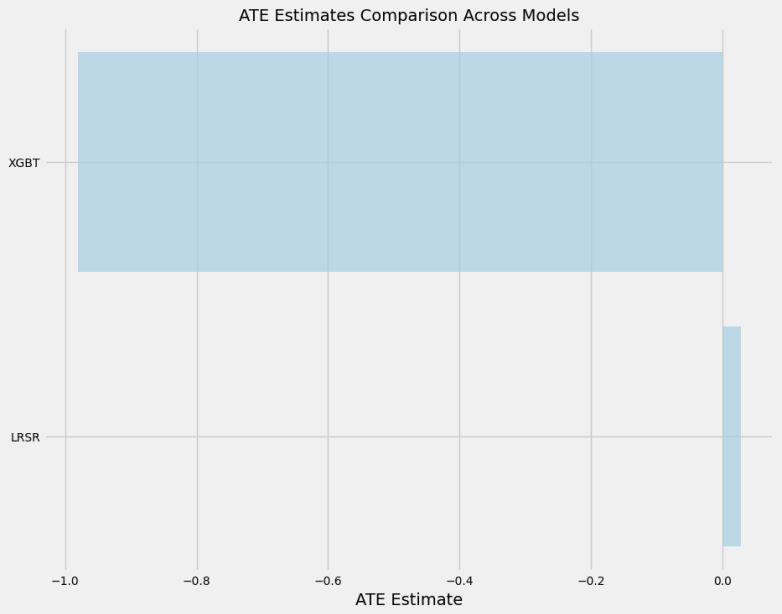

Figure 5: ATE Estimates using XGBT and LRSR

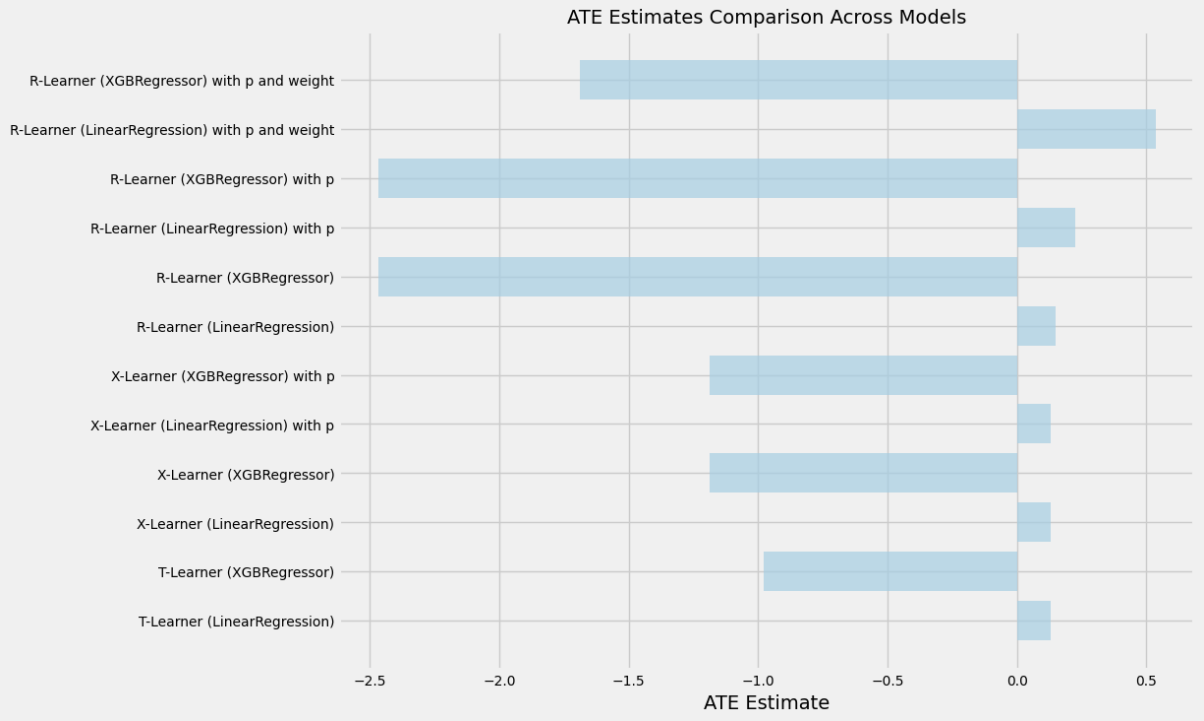

Figure 6: ATE Estimates Comparison Across Models

Individual Treatment Effect

The analysis of the Individual Treatment Effect (ITE/CATE) using XGBT showed that CATE values are predominantly clustered around zero. This distribution suggested that, although there may be some variation at the individual level, the treatment generally exerts minimal to no impact on outcomes. Consequently, while individual responses to the treatment exist, they do not aggregate to a significant effect at the group level. Similarly, the LRSR CATE value was around zero and yielded a minor effect.

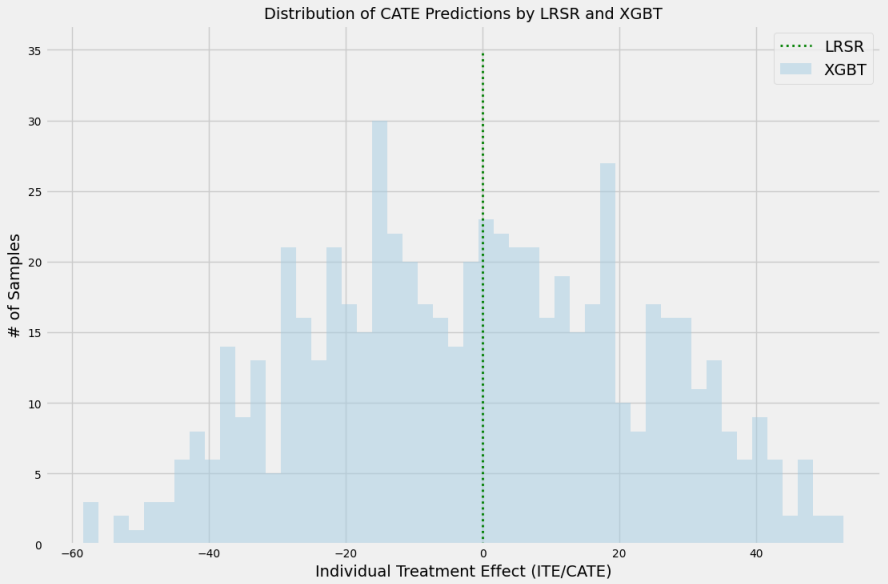

Figure 7: Distribution of CATE Predictions by LRSR and XGBT

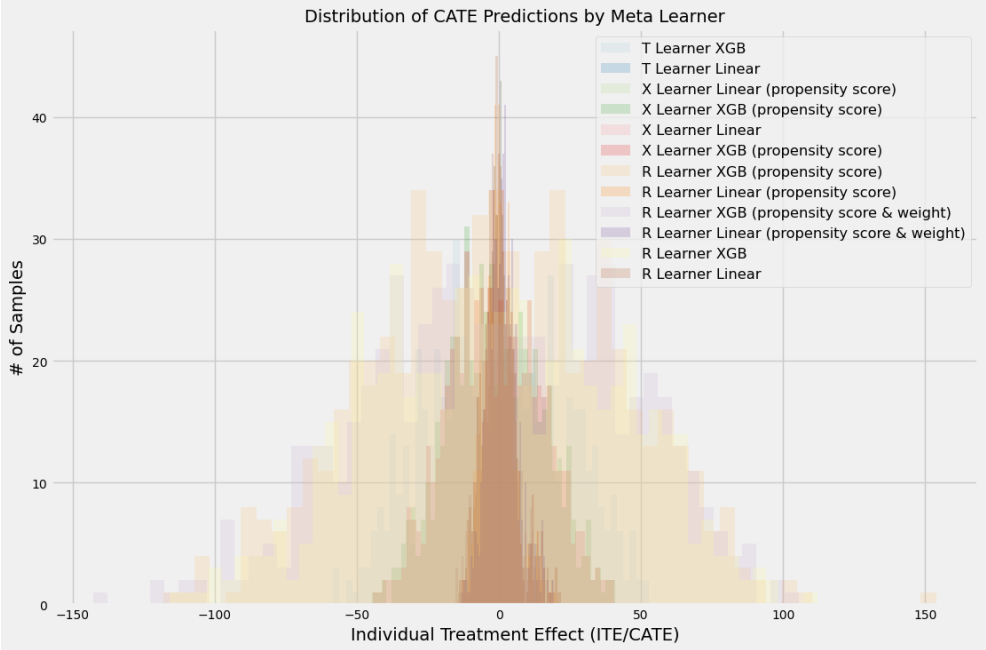

Figure 8: Distribution of CATE Predictions by Meta Learner

Feature Importance

The feature importance analysis for the LRSRegressor yielded NaN (Not a Number) values for all features with warning message, 'there were no meaningful features which satisfy the provided configuration'. This unexpected outcome suggests potential challenges in the model fitting process or data compatibility with the feature importance calculation method.

On the other hand, the XGBTRegressor identified several key factors, with 'Age' being the most significant (0.203414), followed by 'Percent Salary Hike' (0.124994), and 'Total Working Years' (0.120453). Other notable features included 'Years at Company', 'Years Since Last Promotion', and 'Education Field_encoded', among others, indicating the complex factors influencing salary beyond educational attainment.

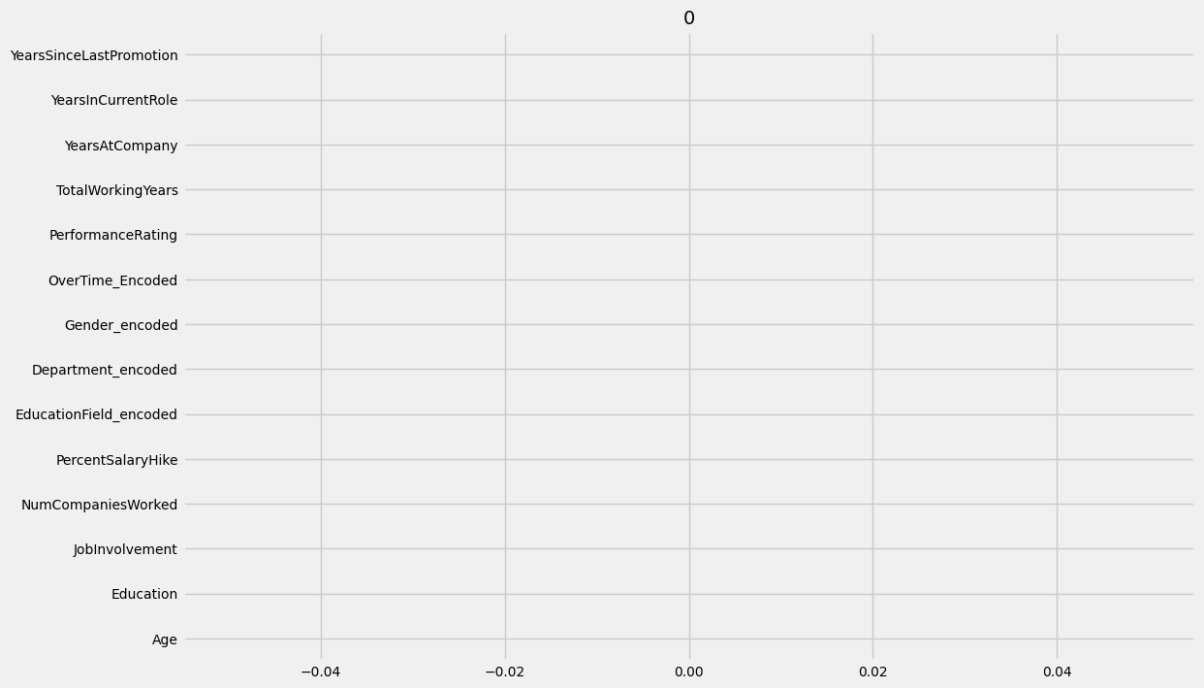

Figure 9: Feature Importance by LRSR

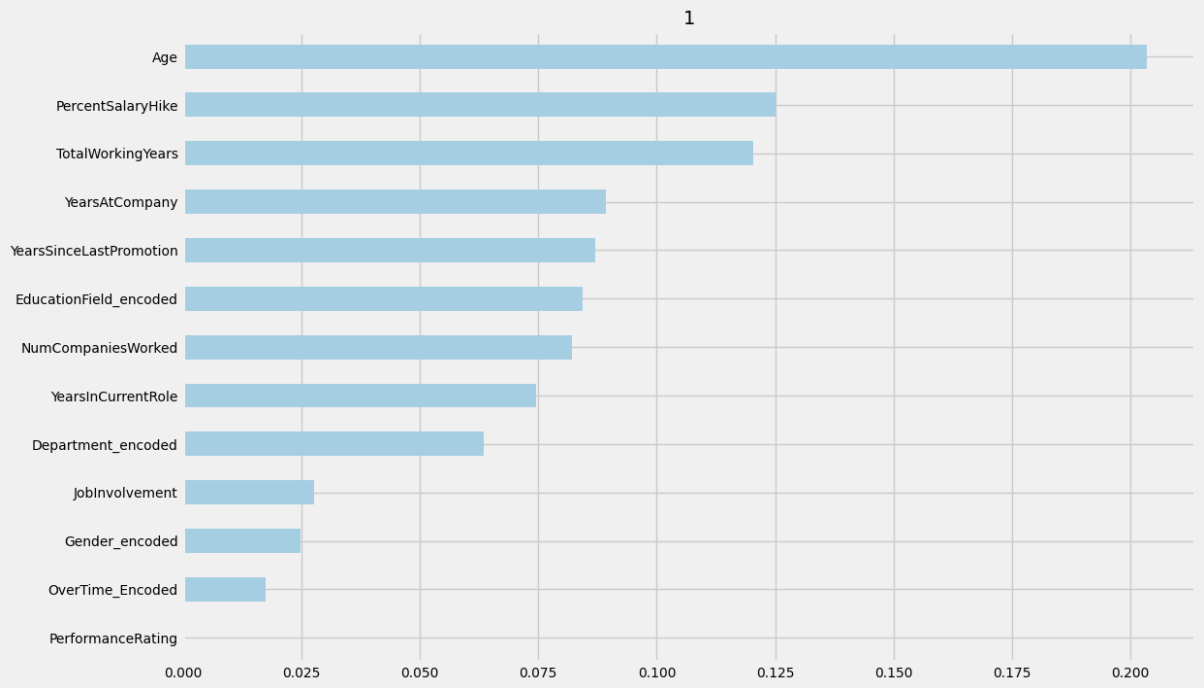

Figure 10: Feature Importance by XGBT

Feature Name	**Feature Importance**	
	LRSRegressor	**XGBTRegressor**
Age	NaN	0.203414
JobInvolvement	NaN	0.027720
NumCompaniesWorked	NaN	0.082039
PercentSalaryHike	NaN	0.124994
EducationField_encoded	NaN	0.084371
Department_encoded	NaN	0.063602
Gender_encoded	NaN	0.024895
OverTime_Encoded	NaN	0.017520
PerformanceRating	NaN	0.000000
TotalWorkingYears	NaN	0.120453
YearsAtCompany	NaN	0.089329
YearsInCurrentRole	NaN	0.074611
YearsSinceLastPromotion	NaN	0.087052

Figure 11: Feature Importance Value Table

Conclusion

In conclusion, although advanced degrees are commonly associated with higher earning potential, my analysis suggests that their impact on salary is minimal and that other professional and demographic factors play a more significant role. This challenges the traditional view and highlights the importance of considering a broader range of factors when evaluating the benefits of further education for career advancement.