What Determines Employee Salaries: A Regression Analysis Study ?

Neeraj Namani, Srisailam Gitte
University at Albany, State University of New York

INTRODUCTION

This study utilizes multiple regression analysis conducted in R to explore the determinants of employee salaries, focusing on variables such as experience, education, job role and industry. Through meticulous data preprocessing and model selection, we aim to develop a robust predictive model. The objective is to offer actionable insights into a salary structuring and strategies for enhancing employee retention.

How do factors such as experience, education, job role and industry influence employee salaries and how can this information be utilized to inform salary structuring and retention strategies in organizations ?

EDA

- EDA revealed strong positive correlations between salary and key variables such as age and years of experience. The matrix plots and correlation coefficients indicate that years of experience have the highest correlation with salary, suggesting that experience is a significant predictor of salary levels.
- Scatter plots display a distinct trend where salary increases with both age and experience, with density plots and histograms showing the distribution of these variables. Additionally, box plots reveal the spread and central tendency of age, years of experience and salary, suggesting variability in the data that could affect salary predictions.

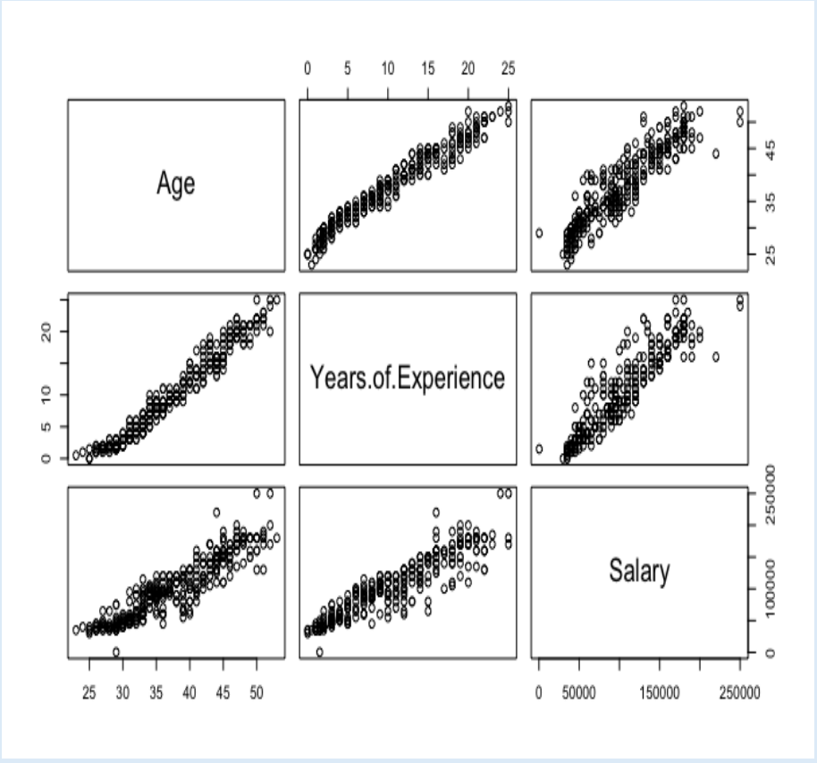


Figure 1: Matrix Plot 1

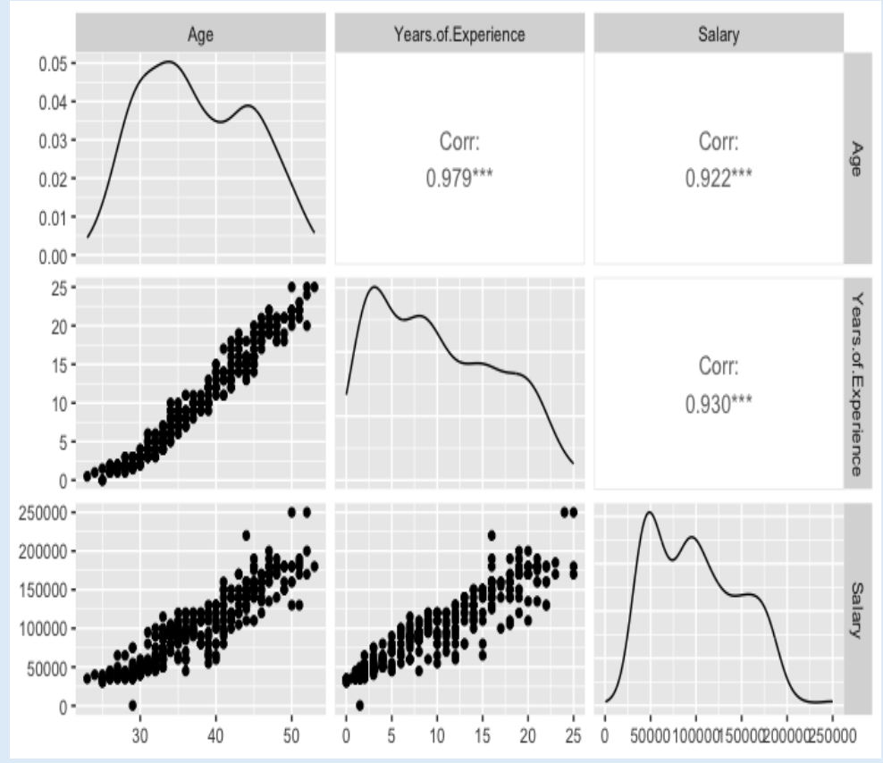


Figure 2: Matrix Plot 2

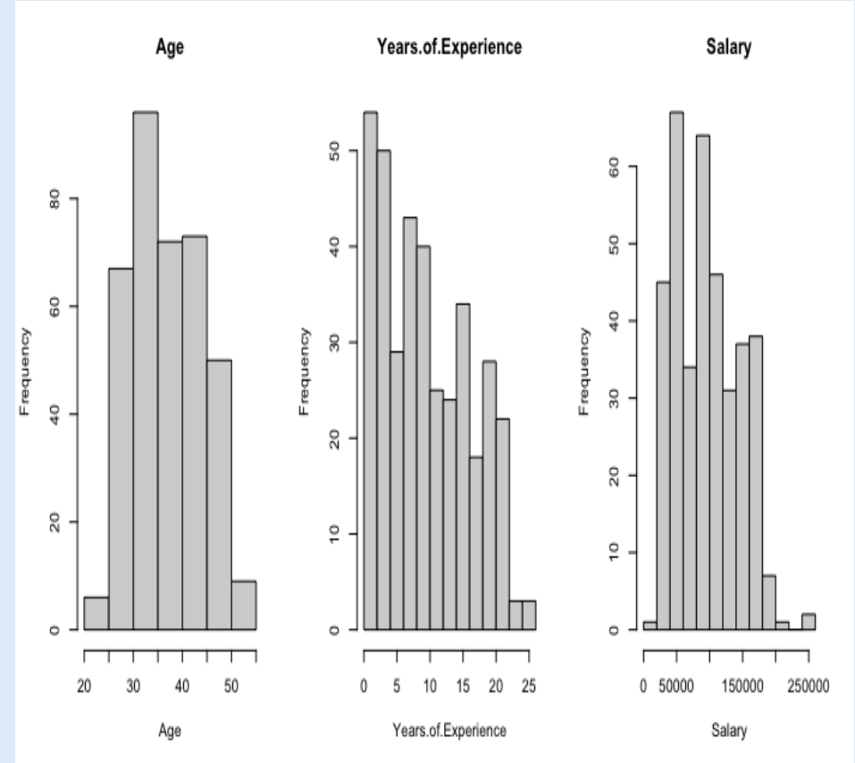


Figure 3: Histogram Plots

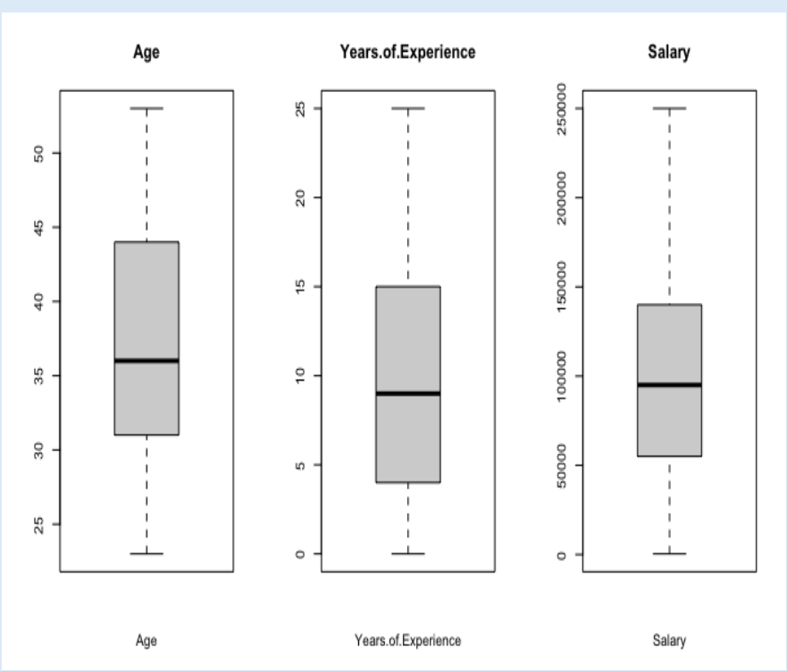


Figure 4: Box Plots

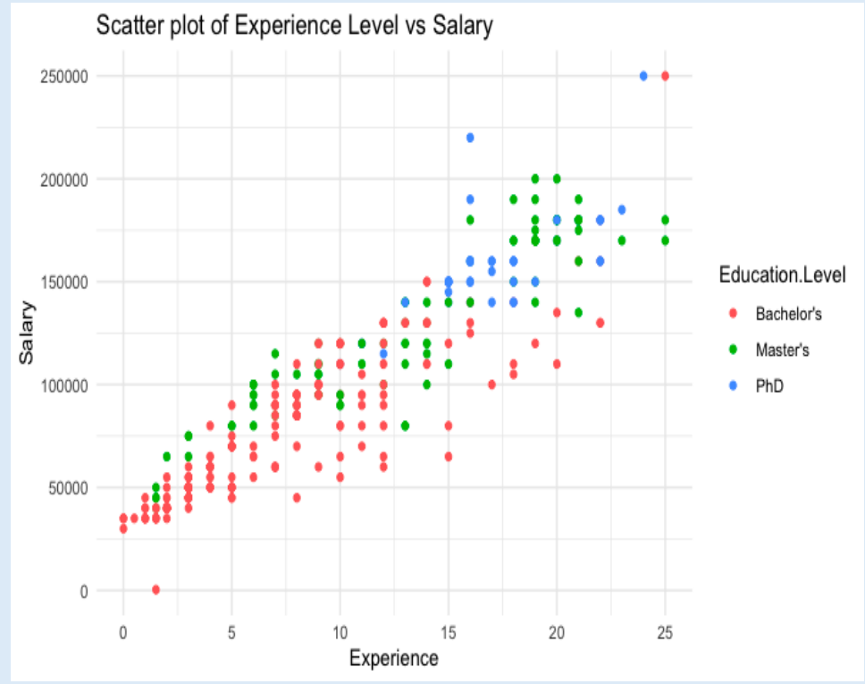


Figure 5: Scatter Plot (Salary vs Experience)

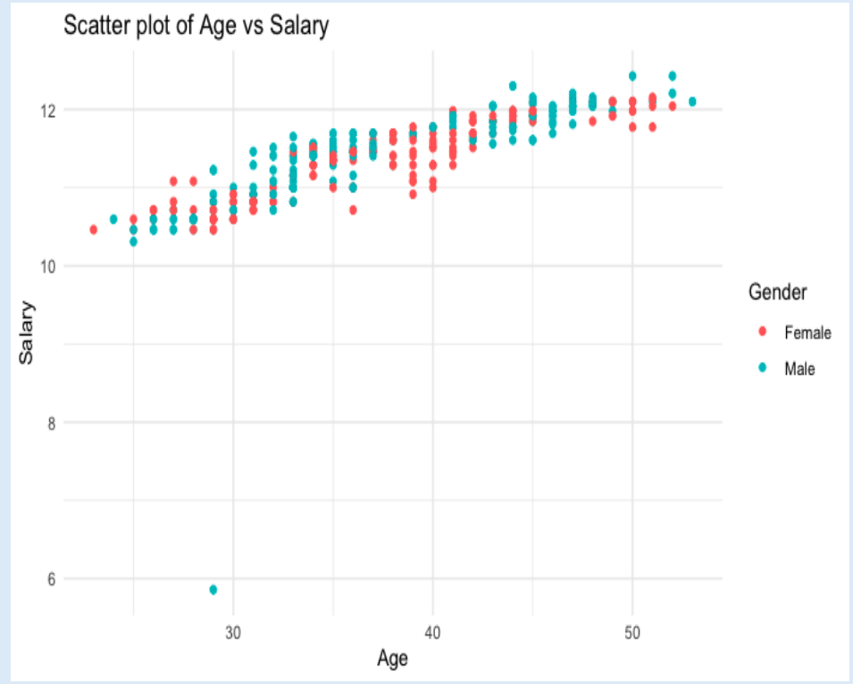
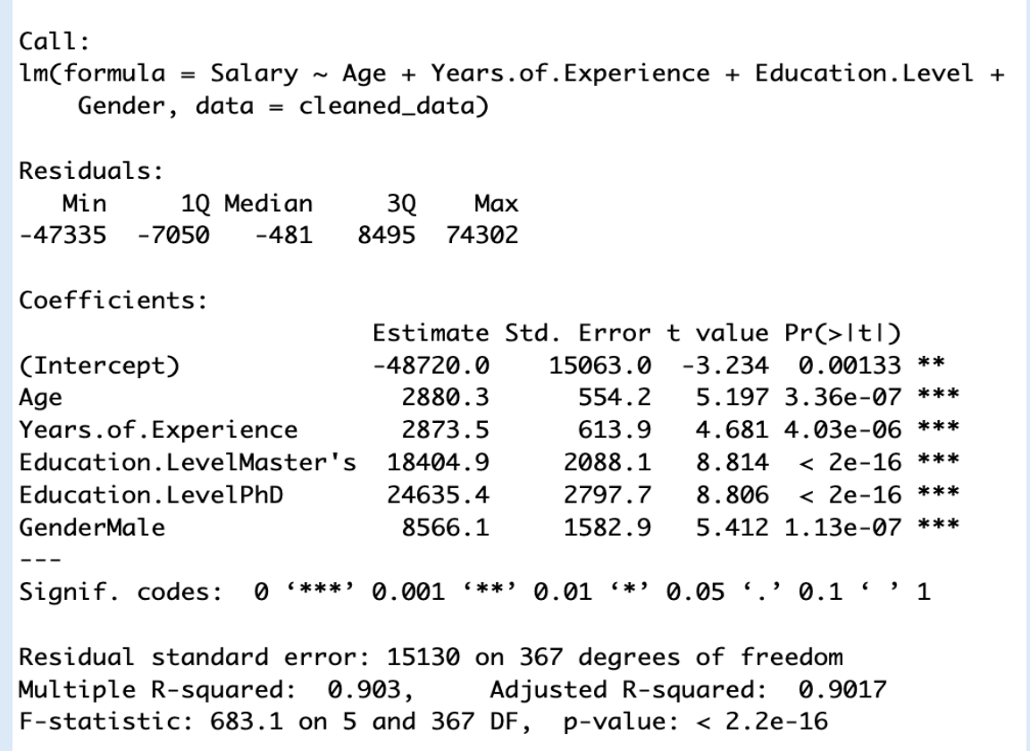


Figure 6: Scatter Plot (Salary vs Age)

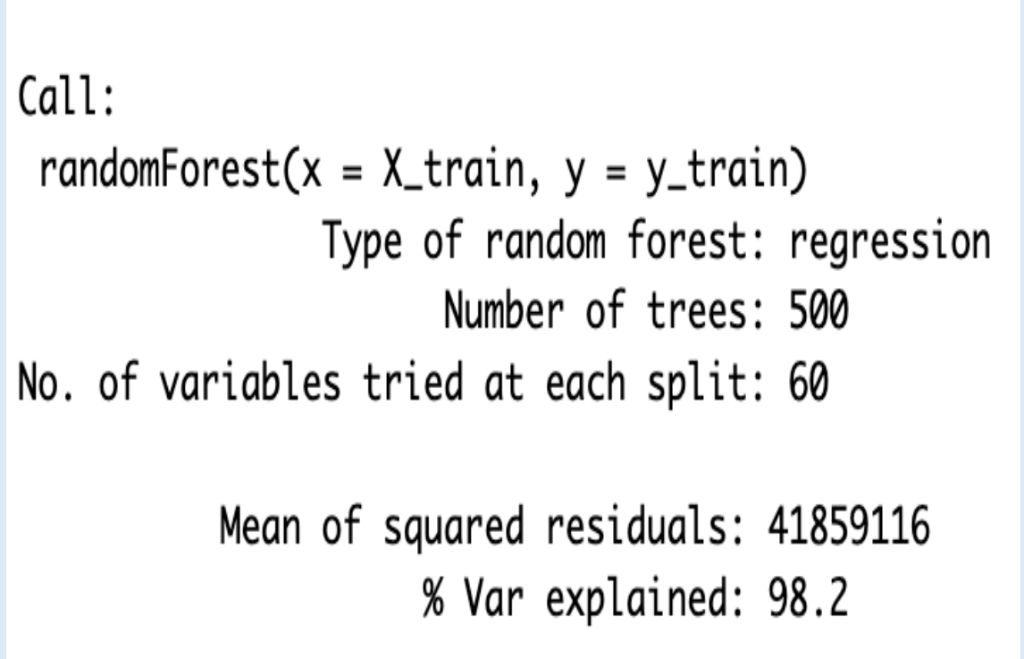
MODELS



```
Call:
lm(formula = Salary ~ Age + Years.of.Experience + Education.Level +
    Gender, data = cleaned_data)

Residuals:
   Min     1Q Median     3Q    Max
-47335  -7050   -481   8495  74302

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                -48720.0    15063.0  -3.234  0.00133 **
Age                          2880.3      554.2   5.197 3.36e-07 ***
Years.of.Experience          2873.5      613.9   4.681 4.03e-06 ***
Education.LevelMaster's     18404.9     2088.1   8.814  < 2e-16 ***
Education.LevelPhD          24635.4     2797.7   8.806  < 2e-16 ***
GenderMale                   8566.1     1582.9   5.412 1.13e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15130 on 367 degrees of freedom
Multiple R-squared:  0.903,     Adjusted R-squared:  0.9017
F-statistic: 683.1 on 5 and 367 DF,  p-value: < 2.2e-16
```

Figure 7: Linear Regression Summary

Age and Years of Experience positively impact salary coefficients of $2880.30 and $2873.50 respectively, indicating higher wages with increased age and experience.



```
Call:
 randomForest(x = X_train, y = y_train)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 60

          Mean of squared residuals: 41859116
                    % Var explained: 98.2
```

Figure 8: Random Forest Summary

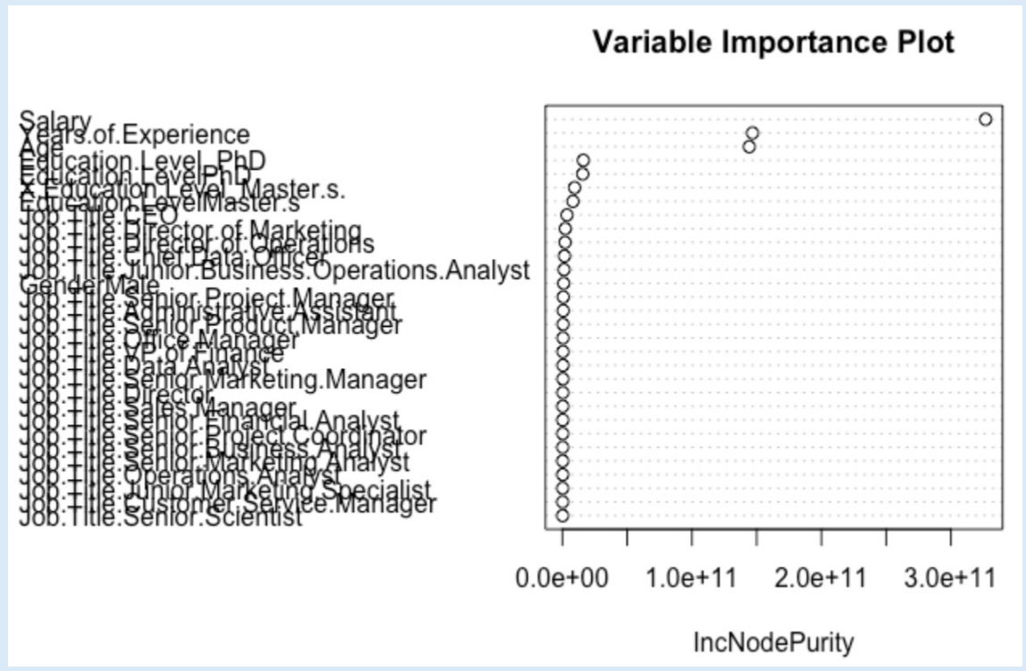


Figure 9: Variable Importance Plot (Random Forest)

Years of Experience and advanced degrees (PhD, Master’s) are the most influential in salary prediction. These variables show higher importance scores based on IncNodePurity, which is a measure of the decrease in node impurity from splitting on the variable, averaged over all trees.

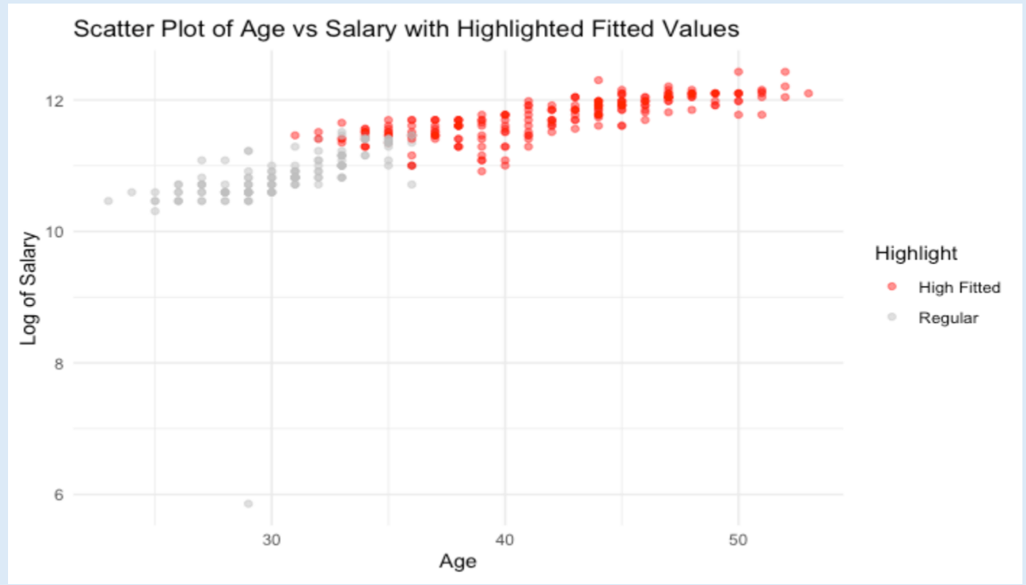


Figure 10: Scatter plot (Salary vs Age Fitted)

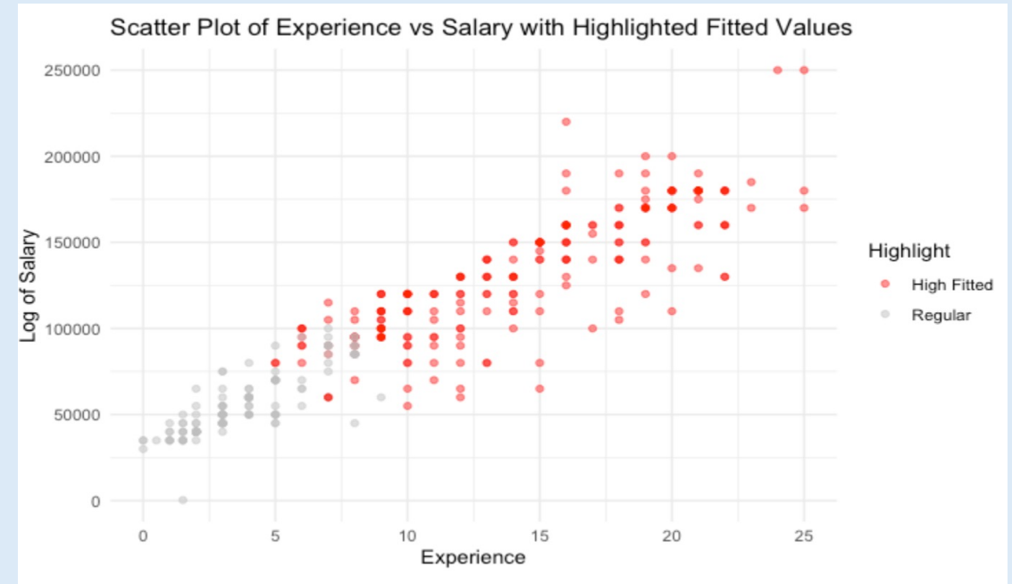
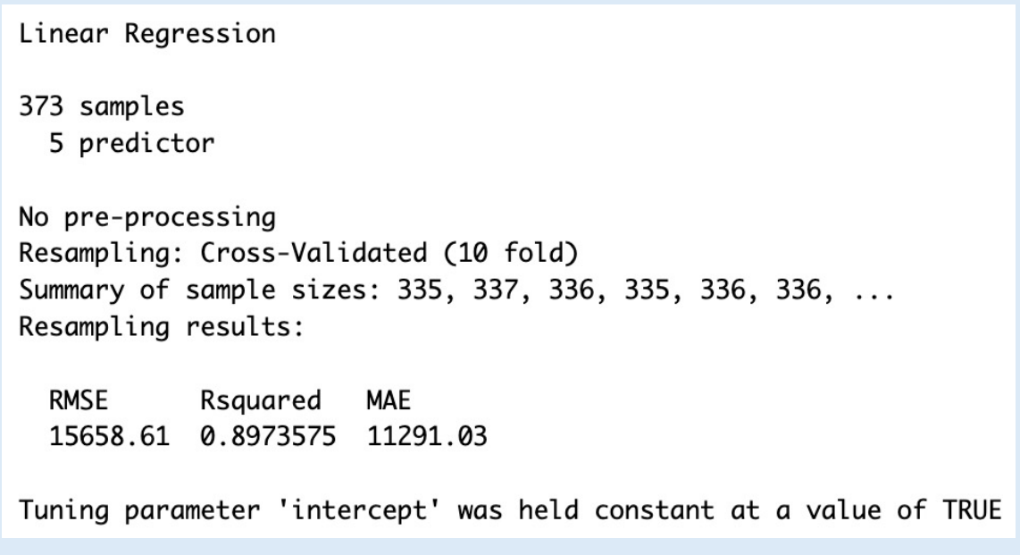
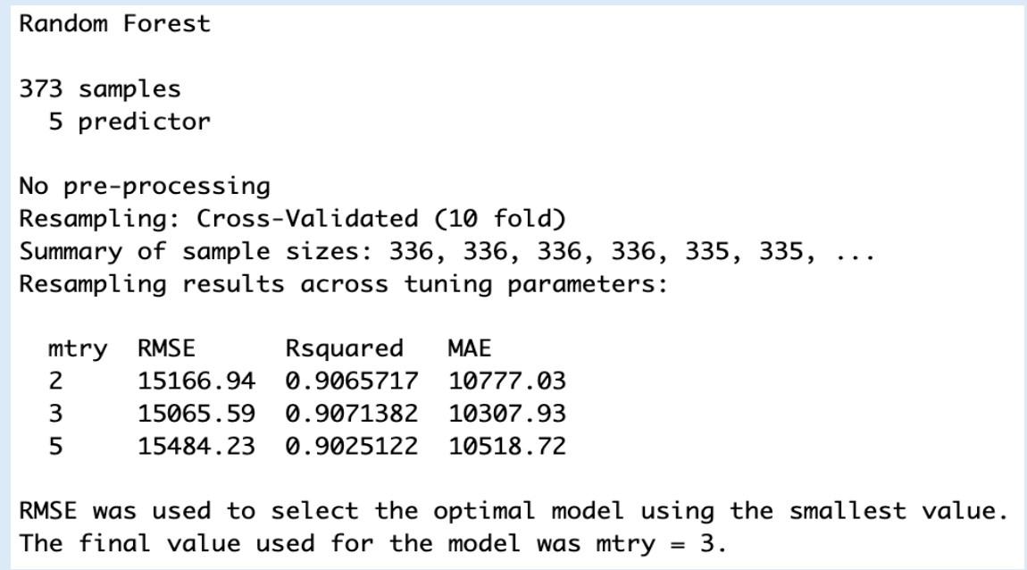


Figure 11: Scatter plot (Salary vs Experience Fitted)



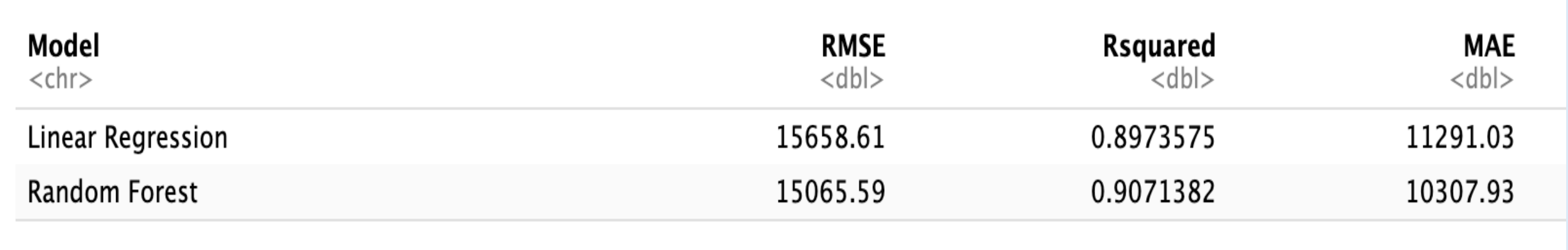
```
Linear Regression

373 samples
  5 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 335, 337, 336, 335, 336, 336, ...
Resampling results:

  RMSE      Rsquared   MAE
  15658.61  0.8973575  11291.03

Tuning parameter 'intercept' was held constant at a value of TRUE
```

Figure 12: Linear Regression (Cross Validation)



```
Random Forest

373 samples
  5 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 336, 336, 336, 336, 335, 335, ...
Resampling results across tuning parameters:

  mtry  RMSE      Rsquared   MAE
  2     15166.94  0.9065717  10777.03
  3     15065.59  0.9071382  10307.93
  5     15484.23  0.9025122  10518.72

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was mtry = 3.
```

Figure 13: Random Forest (Cross Validation)



Model <chr>	RMSE <dbl>	Rsquared <dbl>	MAE <dbl>
Linear Regression	15658.61	0.8973575	11291.03
Random Forest	15065.59	0.9071382	10307.93

Figure 14: Statistical Comparisions for two models

MODEL VISUALIZATIONS

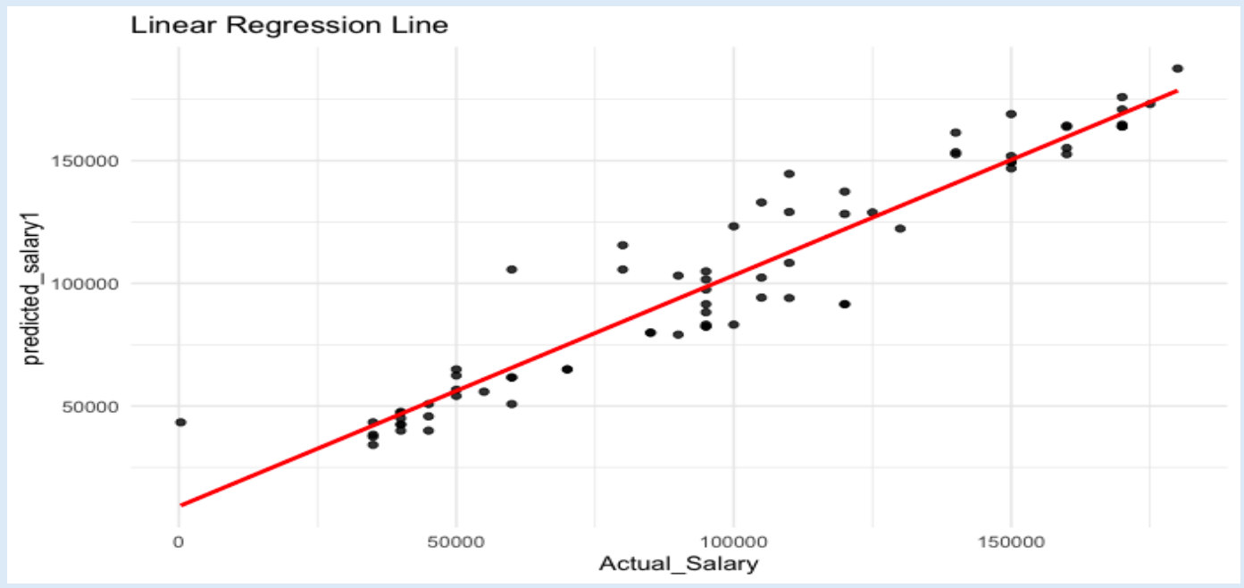


Figure 15: Linear Regression Plot

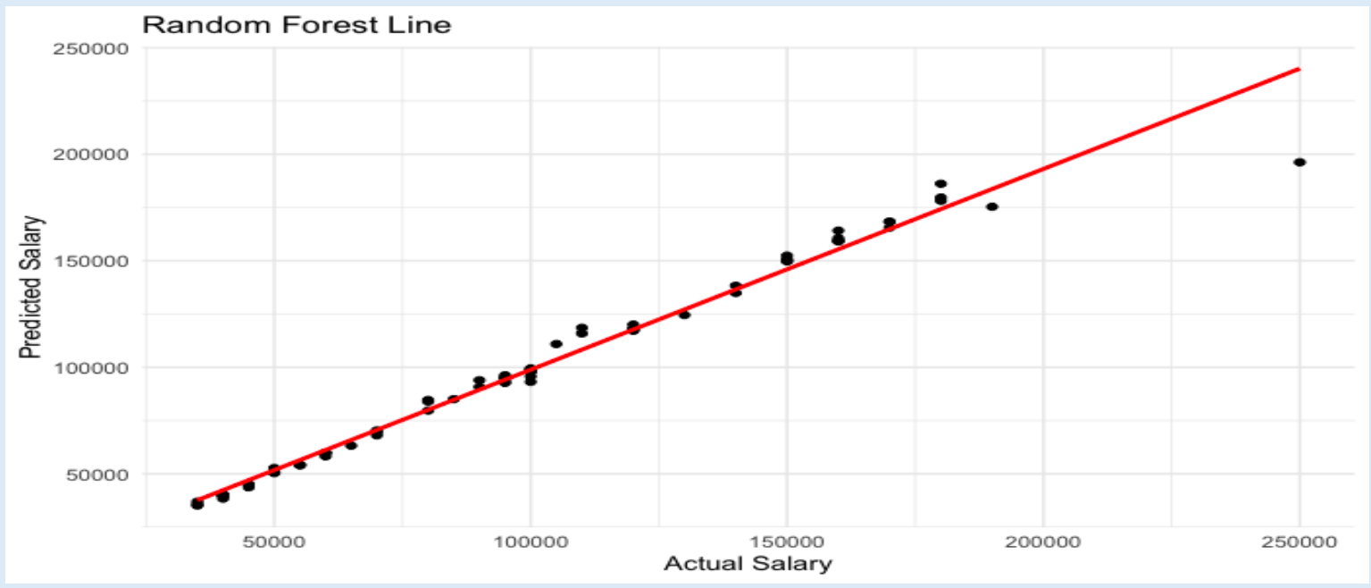
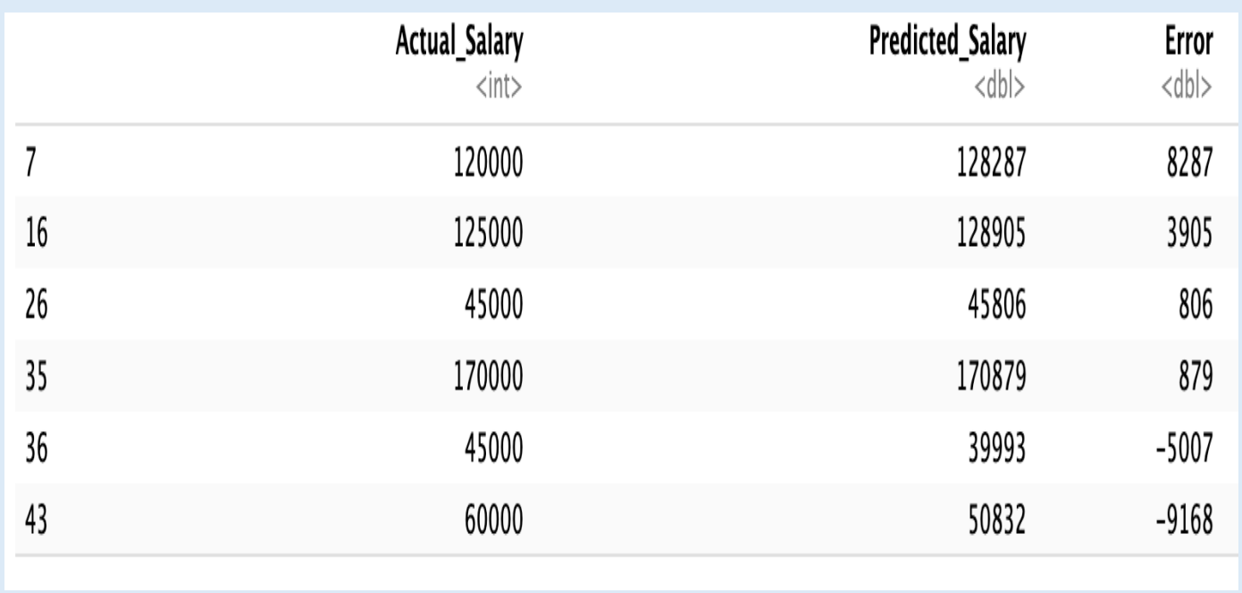
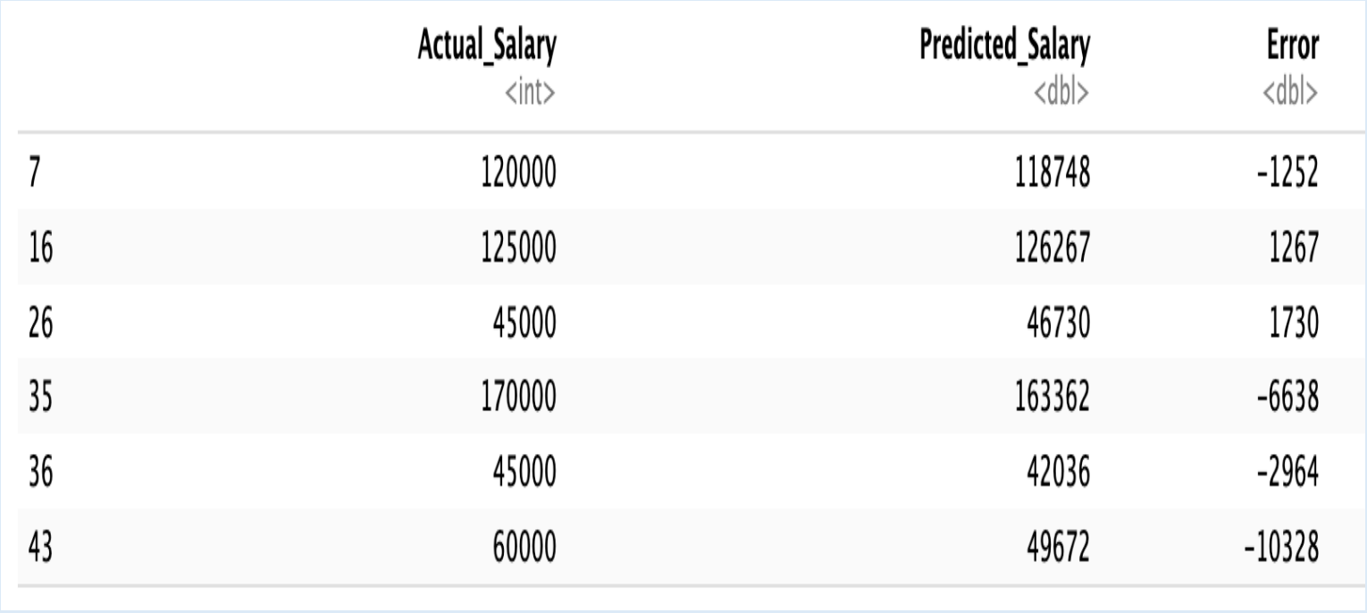


Figure 16: Random Forest Plot



	Actual_Salary <int>	Predicted_Salary <dbl>	Error <dbl>
7	120000	128287	8287
16	125000	128905	3905
26	45000	45806	806
35	170000	170879	879
36	45000	39993	-5007
43	60000	50832	-9168

Figure 17: Predicted vs Actual Salary



	Actual_Salary <int>	Predicted_Salary <dbl>	Error <dbl>
7	120000	118748	-1252
16	125000	126267	1267
26	45000	46730	1730
35	170000	163362	-6638
36	45000	42036	-2964
43	60000	49672	-10328

Figure 18: Predicted vs Actual Salary

- Random Forest performed better than Linear Regression

CONCLUSION

After examining the results of our study alongside the referenced papers. It is evident that both our analysis and the literature concur on the significance of experience and education in predicting employee salaries. Where our study extends these findings is in the deployment of a Random Forest model that outshines Linear Regression in 10-fold cross-validation, offering a more intricate understanding of the determinants of salary.

Our study concludes that experience and advanced education significantly influence salary outcomes, with the Random Forest model providing a more nuanced understanding of these relationships than Linear Regression.

REFERENCES

1. G. Wang, "Employee Salaries Analysis and Prediction with Machine Learning," 2022
Link: https://ieeexplore.ieee.org/document/9943146

2. “Employee Salary Prediction”, 2022 by Tiasa Mukherjee, MS. B. Satyasaivani
Link: https://www.ijariit.com/manuscript/employees-salary-prediction/