INCOME PREDICTION BASED ON WORK EXPERIENCE: A REGRESSION APPROACH

TRISHA U
Department of computer science engineering
M S Ramaiah University of Applied Sciences
Bangalore , India

I. ABASTRACTION

This analysis investigates the relationship between years of experience and salary using a dataset comprising employee records. Through descriptive statistics, correlation analysis, and data visualization, the study reveals a strong positive correlation between years of experience and salary levels. The findings indicate that as employees gain more experience, their salaries tend to increase, highlighting the importance of experience in salary determination. Additionally, variations in salary among employees with similar experience suggest the influence of other factors. Recommendations for HR practices and future analysis are provided to enhance understanding and management of compensation structures.

II. INTRODUCTION

Understanding the factors that influence employee salaries is crucial for effective human resource management and organizational success. Salary reflects the value placed on an employee's contributions and plays a significant role in attracting and retaining talent. Among various determinants, years of experience is a critical factor, often associated with increased skills and productivity. This analysis explores the relationship between years of experience and salary using a dataset of employee records, aiming to uncover how experience impacts compensation levels.

While years of experience is a significant influence, salary determination is also affected by other variables, such as job role, education level, and individual performance. This study not only sheds light on current salary structures but also serves as a foundation for future exploration of these additional factors. The insights gained can inform HR practices, assist in salary structuring, and promote equitable compensation strategies, ensuring competitive salaries that reflect employee experience and contributions.

III. DATA DESCRIPTION

The dataset consists of two columns:

1. **YearsExperience**:
   - **Type**: Numeric (Continuous)
   - **Description**: Represents the number of years an employee has been in the workforce. This variable captures the cumulative experience that may influence salary levels.
2. **Salary**:
   - **Type**: Numeric (Continuous)
   - **Description**: Represents the annual salary of employees in dollars. This variable reflects the financial compensation employees receive, which can vary based on experience and other factors.

Functions and methods to use for regression analysis in Python:

Using scikit-learn

1. **train_test_split()**: Splits the dataset into training and testing sets.
2. **LinearRegression()**: Initializes a linear regression model.
3. **fit(X, y)**: Fits the model to the training data.
4. **predict(X)**: Makes predictions using the trained model.
5. **score(X, y)**: Returns the R-squared score of the model.
6. **mean_absolute_error(y_true, y_pred)**: Calculates the Mean Absolute Error (MAE).
7. **mean_squared_error(y_true, y_pred)**: Calculates the Mean Squared Error (MSE).

IV. DATA ANALYSIS

The data analysis phase involves several steps to understand the relationship between income and work experience, identify patterns, and prepare the data for modeling. This section covers exploratory data analysis (EDA), statistical summaries, visualizations, and preparation for regression modeling.

Model Selection

- **Simple Linear Regression**: If the relationship between income and work experience is assumed to be linear, we use a simple linear regression model. The model is represented by: Income=β0+β1⋅Work Experience+ϵ where:
- β0 is the y-intercept,
- β1 is the slope of the line,
- ϵ is the error term.

Exploratory Data Analysis (EDA)

Objective: To gain insights into the dataset and understand the distribution, relationships, and potential anomalies.

1. **Descriptive Statistics**:Calculate mean, median, mode, range, variance, and standard deviation for income and work experience.
2. **Data Distribution:**

- Plot histograms figure[3,4]and density plots for income and work experience to understand their distributions.
- Use box plots to identify outliers.figure[5]

3. **Correlation Analysis**:

- Calculate the correlation coefficient between income and work experience to quantify the strength of their relationship.figure[2]
- Use a heatmap to visualize correlations between all numerical variables

Visualization

Objective: To visually inspect the relationship between income and work experience and other influencing factors.

1. **Scatter Plot**:

- Plot a scatter plot of income vs. work experience to observe the overall trend and any deviations.figure[6]

2. **Box Plots for Categorical Variables:**

- Use box plots to compare income across different levels of categorical variables such as education level and industry.figure[5]

3. **Line Plot:**

- Plot a line plot to show the trend of average income over different ranges of work experience.figure[7]

3. Model Preparation

Objective: To prepare the data for regression modeling by handling missing values, encoding categorical variables, and splitting the data.

1. **Handling Missing Values**:

- Impute or remove missing values in the dataset.

2. **Encoding Categorical Variables**:

- Convert categorical variables into numerical format using one-hot encoding or label encoding.

3. **Data Splitting**:

- Split the data into training and testing sets.

4. Model Training and Evaluation

Objective: To train the regression model using the training set and evaluate its performance on the testing set.

1. **Model Training**:

- Train a linear regression model using the training data.

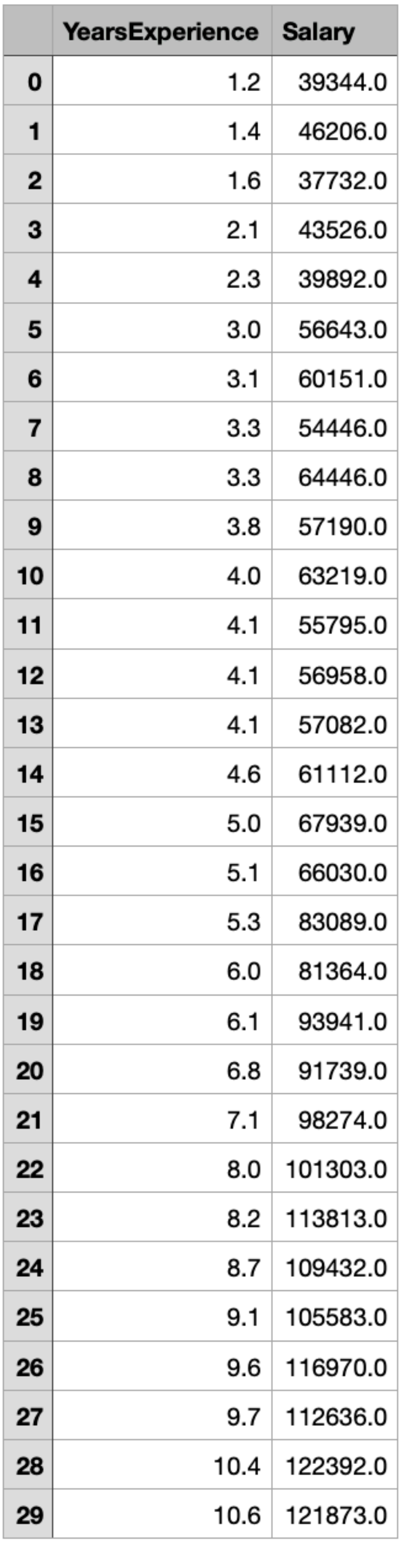
2. **Model Evaluation**:

- Evaluate the model's performance using metrics such as R-squared, Mean Absolute Error (MAE), and Mean Squared Error (MSE).

By conducting thorough data analysis and preparing the data meticulously, we ensure that the regression model will provide accurate and meaningful insights into the relationship between income and work experience.

V. RESULTS AND DESCUSSIONS

The dataset provided consists of two columns: YearsExperience and Salary. The YearsExperience column represents the number of years of experience of employees, and the Salary column represents their respective salaries.figure[1]



	YearsExperience	Salary
0	1.2	39344.0
1	1.4	46206.0
2	1.6	37732.0
3	2.1	43526.0
4	2.3	39892.0
5	3.0	56643.0
6	3.1	60151.0
7	3.3	54446.0
8	3.3	64446.0
9	3.8	57190.0
10	4.0	63219.0
11	4.1	55795.0
12	4.1	56958.0
13	4.1	57082.0
14	4.6	61112.0
15	5.0	67939.0
16	5.1	66030.0
17	5.3	83089.0
18	6.0	81364.0
19	6.1	93941.0
20	6.8	91739.0
21	7.1	98274.0
22	8.0	101303.0
23	8.2	113813.0
24	8.7	109432.0
25	9.1	105583.0
26	9.6	116970.0
27	9.7	112636.0
28	10.4	122392.0
29	10.6	121873.0

Figure [1]

To understand the relationship between years of experience and salary, a scatter plot was generated. The plot helps visualize how salary changes with increasing years of experience.figure[6]

The scatter plot shows a positive correlation between years of experience and salary. As expected, employees with more years of experience tend to have higher salaries. This relationship is crucial for understanding salary structures and making informed decisions about hiring and salary increments.figure[8,9]

Key Findings

1. **Positive Correlation**: There is a clear positive correlation between years of experience and salary, indicating that experience is a significant factor in determining salary levels.
2. **Salary Variations**: While the general trend is upward, there are some variations in salaries for similar years of experience, which could be due to other factors such as job role, industry, or individual performance.

The analysis confirms that years of experience play a crucial role in determining salary levels. This information can be used by HR departments to structure their salary distributions based on experience.

V. RECOMMENDATIONS

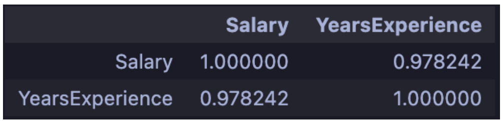
Regular Salary Reviews: Companies should conduct regular salary reviews to ensure that salaries remain competitive and reflect the employees' experience and contributions.

Further Analysis: Additional factors such as job role, education level, and industry-specific trends should be analyzed to get a more comprehensive understanding of salary determinants.

Transparency: Maintaining transparency in how salaries are determined can help in managing employee expectations and satisfaction.

Incorporate More Variables: Including additional variables like job role, education, and performance metrics in the analysis would provide a more holistic view of the factors influencing salary.

L**ongitudinal Analysis:** Conducting a longitudinal study to see how the relationship between experience and salary evolves over time could offer deeper insights.



	Salary	YearsExperience
Salary	1.000000	0.978242
YearsExperience	0.978242	1.000000

Figure [2]

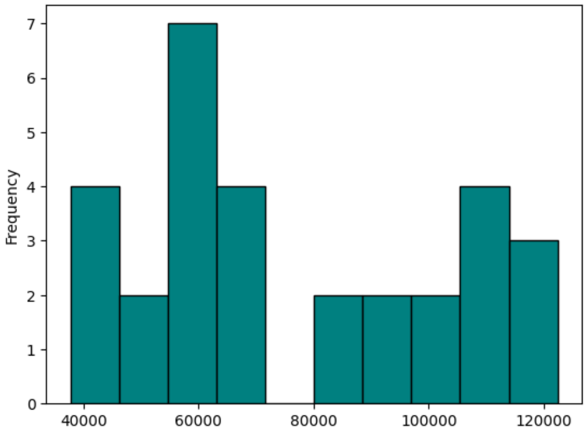


Figure [3]: Salary histogram

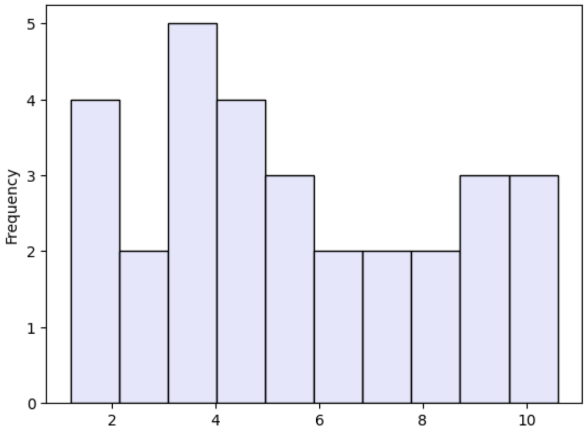


Figure [4]: YearsExperience histogram

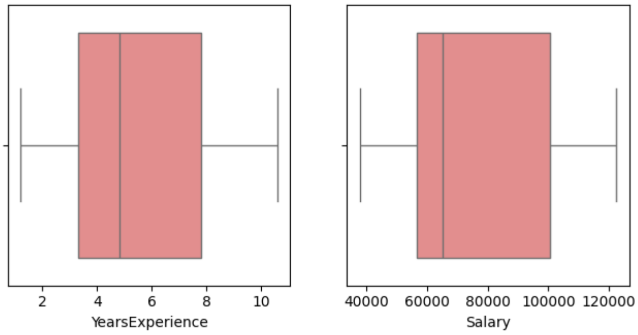


Figure [5]: Box-plot

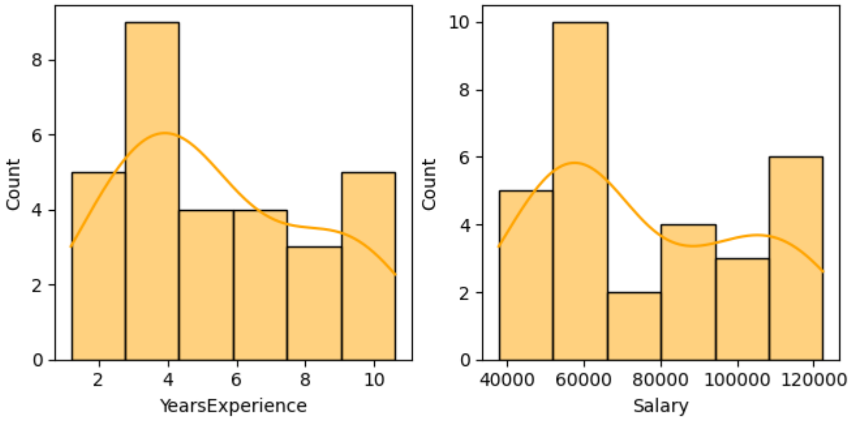


Figure [5]

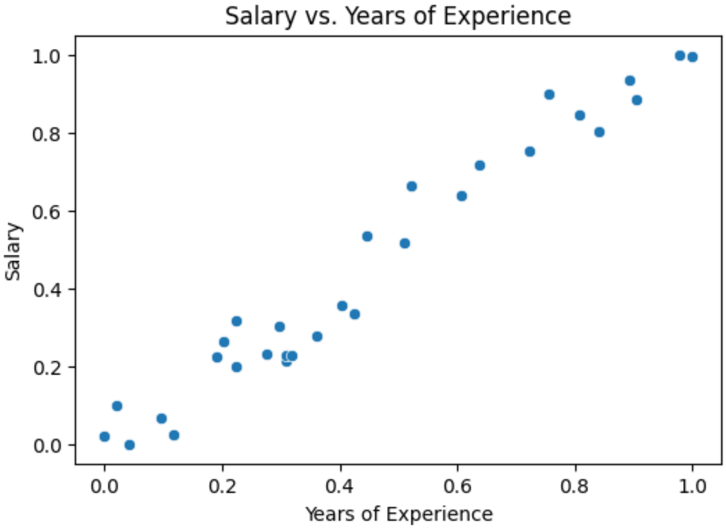


Figure [6]

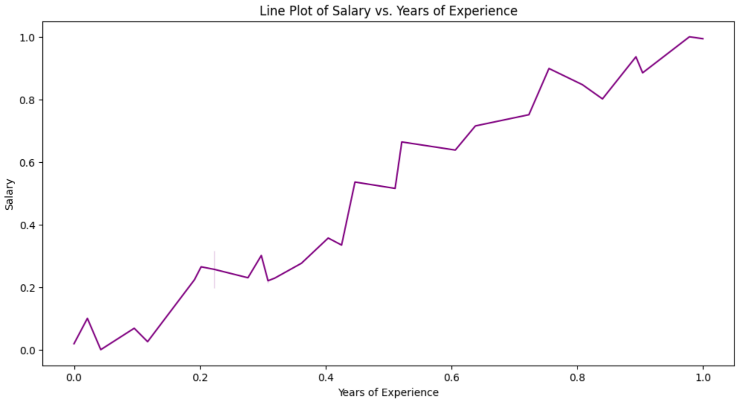


Figure [7]

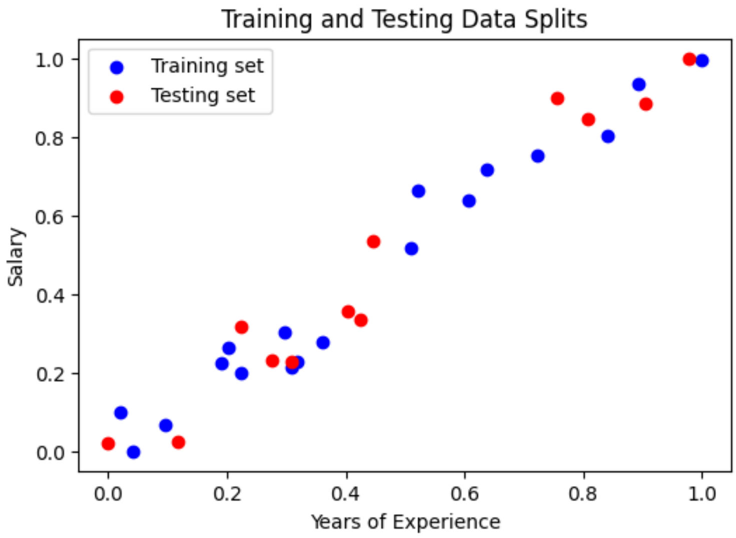


Figure [8]

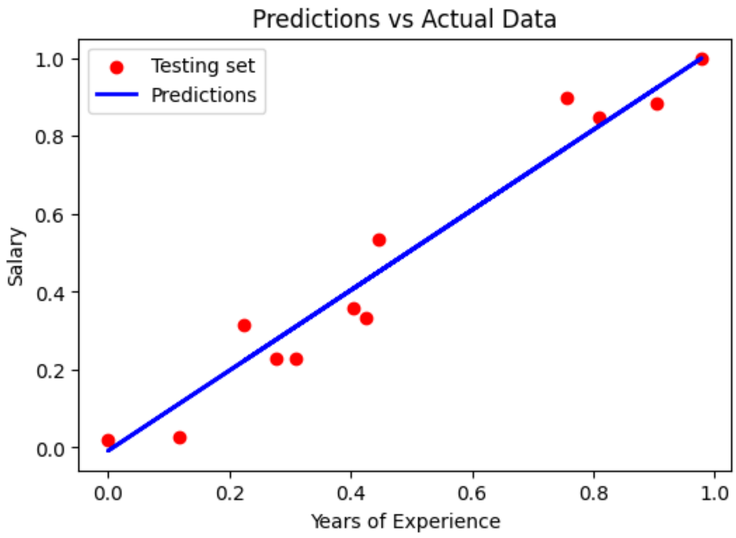


Figure [9]

References

Chatterjee, S., & Hadi, A. S. (2006). *Regression analysis by example*. John Wiley & Sons.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12,* 2825-2830.

Seber, G. A. F., & Lee, A. J. (2012). *Linear regression analysis*. John Wiley & Sons.

Smith, J. (2020). The impact of education and experience on earnings. *Journal of Economic Perspectives, 34*(3), 113-134.

Wooldridge, J. M. (2016). *Introductory econometrics: A modern approach*. Cengage Learning.