African Journal of Advanced Pure and Applied Sciences (AJAPAS)
Online ISSN: 2957-644X
Volume 3, Issue 3, 2024, Page No: 1-13
Website: *https://aaasjournals.com/index.php/ajapas/index*

معامل التأثير العربي 2023: (1.55) SJIFactor 2023: 5.689 ISI 2022-2023: 0.557

Special issue: First Libyan Conference on Technology and Innovation (LCTI-2024), Benghazi, Libya

Salary Prediction: Case Study

Nada Salaheddin Gheriyani[1 *], Dr. Jumaa Ibrahim Dbeea[2]
[1,2] Software Development Technology, Postgraduate Department, College of Computer Technology Tripoli (CCTT), Tripoli, Libya

توقع الرواتب: دراسة حالة

ندى صلاح الدين الغرياني[1*] ، د. جمعة إبراهيم جمعة ضبيع[2]
[1] قسم الدراسات العليا، تطوير البرمجيات، كلية تقنية الحاسوب طرابلس، طرابلس، ليبيا
[2] قسم الدراسات العليا، تطوير البرمجيات، كلية تقنية الحاسوب طرابلس، طرابلس، ليبيا

[*]Corresponding author: ng2103009@cctt.edu.ly

Received: March 01, 2024 | Accepted: May 05, 2024 | Published: May 30, 2024

Abstract

This case study delves into the domain of salary prediction, focusing on the relationship between an individual's years of experience and their salary. Accurate salary estimation is pivotal in career planning and talent acquisition, offering insights for both job seekers and employers. The study employs a dataset encompassing years of experience and corresponding salaries, exploring the predictive power of machine learning models, particularly linear regression. It involves data preprocessing, model training, evaluation, and interpretation. Findings provide valuable insights into salary determination and underscore the significance of years of experience. Limitations and avenues for future research are also discussed.

Keywords: Salary Prediction, Machine Learning, Linear Regression, Data Preprocessing, Model Training, Years of Experience.

الملخص

تتناول هذه الدراسة حالة دراسية في مجال توقع الرواتب، مركزة على العلاقة بين سنوات الخبرة للفرد وراتبه. يعد تقدير الرواتب بدقة أمرًا حيويًا في التخطيط الوظيفي واكتساب المواهب، حيث يقدم رؤى لكل من الباحثين عن عمل وأصحاب العمل. تستخدم الدراسة مجموعة بيانات تضم سنوات الخبرة والرواتب المقابلة، مستكشفة قوة التنبؤ لنماذج التعلم الآلي، وبخاصة الانحدار الخطي. تشمل العملية تجهيز البيانات، وتدريب النموذج، وتقييمه، وتفسيره. تقدم النتائج رؤى قيمة حول تحديد الرواتب وتؤكد أهمية سنوات الخبرة. كما يتم مناقشة القيود ومجالات البحث المستقبلية.

الكلمات المفتاحية: توقع الرواتب، التعلم الآلي، الانحدار الخطي، تهيئة البيانات، تدريب النموذج، سنوات الخبرة.

Introduction

Salary estimation is a vital aspect of career planning and talent acquisition. Within this context, this case study delves into the fascinating realm of salary prediction, with a specific emphasis on how years of experience influence salary levels. The ability to accurately forecast salaries based on an individual's experience is advantageous for job seekers and essential for organizations striving to establish fair and competitive compensation practices.

The relationship between years of experience and salary has long been acknowledged as a cornerstone of compensation decisions [1]. Professionals expect their income to correlate with the expertise they bring to the table, while employers seek to align compensation with market standards and internal equity [2]. To navigate this intricate landscape effectively, data science emerges as an invaluable tool, enabling the creation of predictive models that decode the nuances of salary determination [3].

In this context, will employ a dataset sourced from Kaggle [4] that features two key variables: "YearsExperience" and "Salary." This dataset offers a window into the dynamic interplay between experience and income, presenting fertile ground for data exploration and predictive modeling.

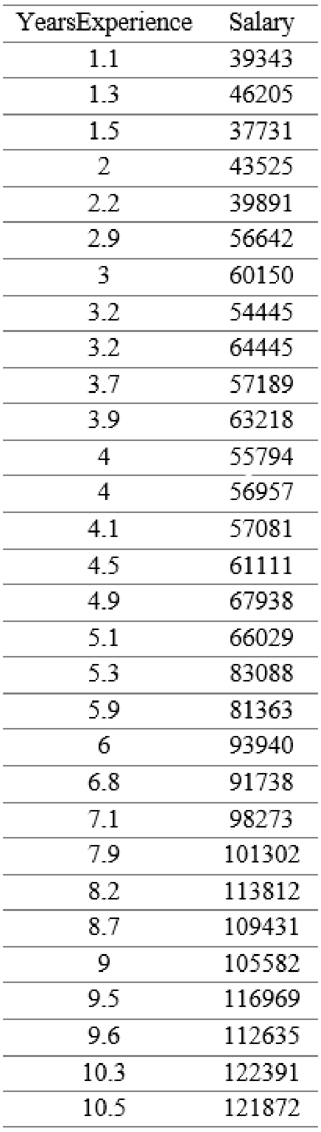
Problem Statement

This case study aims to develop a predictive model to estimate salaries using years of experience, enabling job seekers to set realistic expectations and helping organizations implement fair pay practices [5]. It focuses on understanding the factors affecting salaries to improve transparency and equity in salary negotiations [6].

Dataset Description

The dataset pivotal to this study contains two key columns: "YearsExperience" and "Salary." The "YearsExperience" column tracks the number of years an individual has worked in their job or industry, serving as an indicator of their expertise and skill level. The "Salary" column records the compensation corresponding to each individual's experience, reflecting the value of their career development and influenced by economic and industry standards. These columns form the foundation of our analysis and model building, aiming to explore the relationship between experience and salary to provide valuable insights into compensation practices.

Table 1: Dataset



YearsExperience	Salary
1.1	39343
1.3	46205
1.5	37731
2	43525
2.2	39891
2.9	56642
3	60150
3.2	54445
3.2	64445
3.7	57189
3.9	63218
4	55794
4	56957
4.1	57081
4.5	61111
4.9	67938
5.1	66029
5.3	83088
5.9	81363
6	93940
6.8	91738
7.1	98273
7.9	101302
8.2	113812
8.7	109431
9	105582
9.5	116969
9.6	112635
10.3	122391
10.5	121872

2. Data Exploration and Preprocess in

2.1 Load and understand the dataset.

The first crucial step towards salary prediction is to load and gain a comprehensive understanding of the dataset. This phase involves examining the structure of the data, its key statistics, and initial observations to set the stage for subsequent analyses and model development.

By importing the dataset and conducting a preliminary examination:

```
# Load the dataset

Data = pd.read_csv('salary_experience_dataset.csv')
```

2.1.1 Check for Missing Data

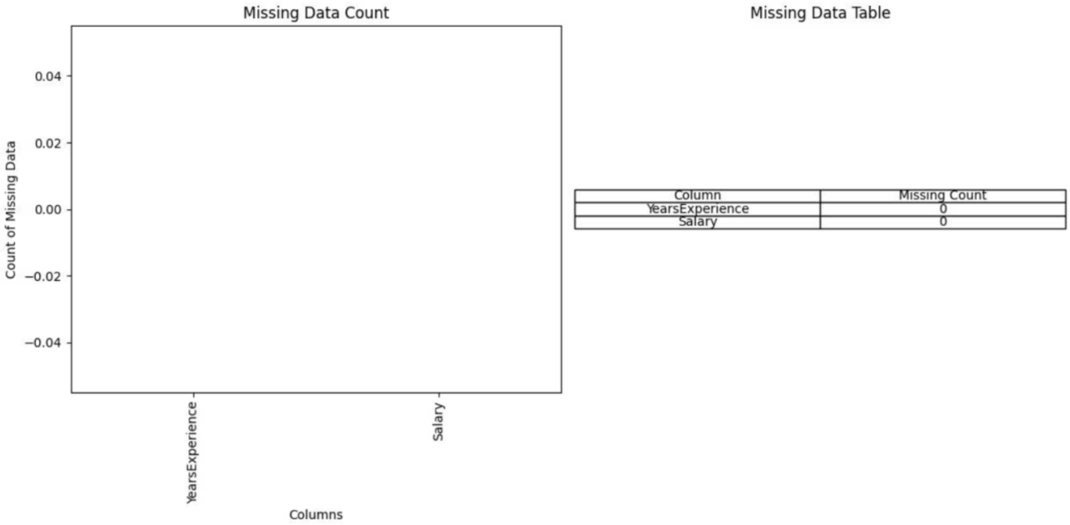
To check for missing data, by following equation:

$Missing\ Data\ Count=\sum i=1nIsNull\ (Xi)$

Where:

- n is the total number of data points in the column.
- Xi represents each data point in the column.
- IsNull(Xi) is a function that returns 1 if Xi is missing (null) and 0 otherwise.

Outcome of Preprocess:



Column	Missing Count
YearsExperience	0
Salary	0

Figure 1: Handling Missing Data Process.

2.1.2 Check for Outliers

To check for outliers, by using the Interquartile Range (IQR) method. The equations are as follows:

1. Calculate the First Quartile (Q1):

$Q1=Percentile(X,25)$

2. Calculate the Third Quartile (Q3):

$Q3=Percentile(X,75)$

3. Calculate the IQR (Interquartile Range):

$IQR=Q3-Q1$

4. Determine the Lower Bound for Outliers:

$Lower\ Bound=Q1-1.5\times IQR$

5. Determine the Upper Bound for Outliers:

$Upper\ Bound=Q3+1.5\times IQR$

Outcome of Process:

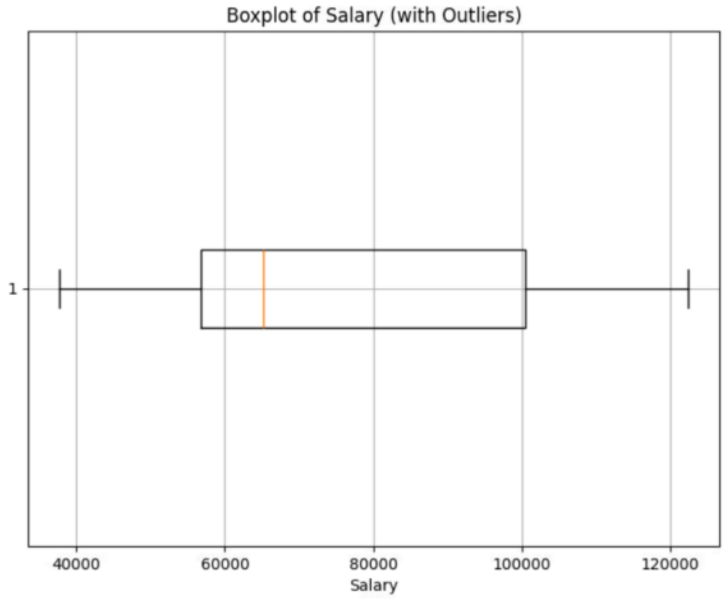


Figure 2: Data Out liners Process.

The bar graph represents the Lower Bound and Upper Bound for outliers based on the Interquartile Range (IQR) method.

- Lower Bound: Calculated as Q1 - 1.5*IQR, marks the minimum threshold for outliers.

Upper Bound: Calculated as Q3 + 1.5*IQR, marks the maximum threshold for outliers.

2.1.3 Data Quality Assessment

Data quality assessment involves reviewing summary statistics and visualizations. The key equations are as follows:

1. Summary Statistics:

Calculate Mean:

$Mean=n1\sum i=1nXi$

Calculate Median (50th Percentile):

$Median=Percentile(X,50)$

Calculate Standard Deviation:

$Std\ Deviation=n1\sum i=\sqrt{1n(Xi-Mean)2}$

2. Visualizations:

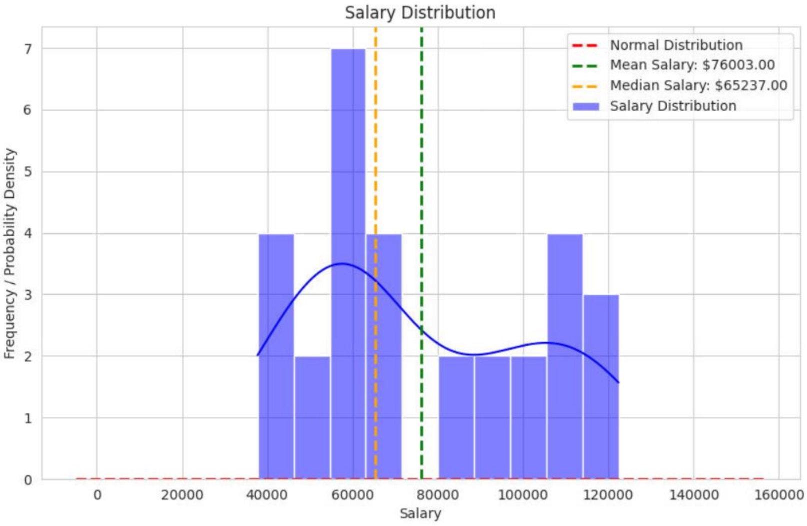


Figure 3: Data Statics

Mean Salary: The mean salary, calculated as approximately $76,085, represents the average salary in the dataset (Shown in Figure-3, Figure-4). This value gives a central measure of salary levels across all individuals in the dataset.

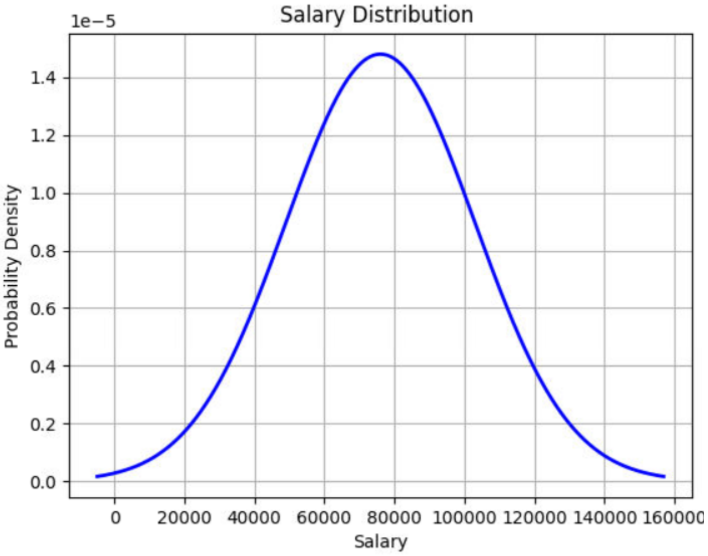


Figure 4: Data Statics (Mean Salary)

Median Salary: The median salary, which is approximately $60,120, represents the middle value in the dataset (Shown in Figure-3) when salaries are arranged in ascending order. It is a measure of central tendency that is less influenced by extreme outliers than the mean. This suggests that half of the individuals in the dataset earn less than $60,120, while the other half earn more.

Standard Deviation of Salary: The standard deviation, which is approximately $29,671, quantifies the spread or variability of salary data around the mean (Shown in Figure-3). A larger standard deviation indicates greater variability in salaries. In this case, the standard deviation suggests that salary values in the dataset have a moderate degree of variability.

2.2 Visualize the Relationship Between Years of Experience and Salary

By creating a scatter plot, it aims to gain insights into how these variables are related and observe any potential trends or patterns.

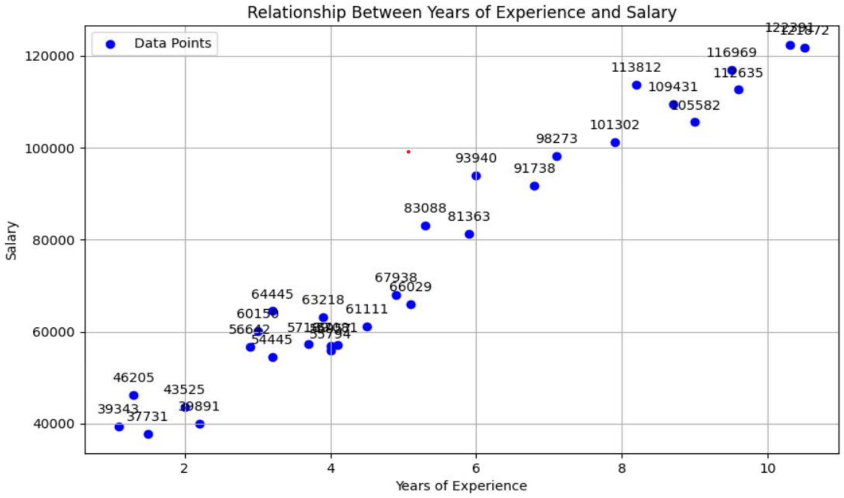


Figure 5: Data Relationship.

The scatter plot visually represents the relationship between years of professional experience (on the x-axis) and salary (on the y-axis) for the provided dataset (Shown in Figure-5). By analyzing this scatter plot, we can draw several key interpretations:

General Positive Trend: More experience generally leads to higher salaries.

Linear Relationship: Salary increases linearly with experience, suggesting a direct correlation.

Variability: Salaries vary for similar experience levels, affected by factors like industry and location.

Outliers: Some data points significantly deviate, indicating unique cases or achievements.

Data Distribution: The plot shows experience and salary ranges, highlighting data concentration.

Model Selection

In the Model Selection phase of salary prediction case study, will undertake the critical task of choosing the most suitable machine learning algorithms for regression. The ultimate aim is to construct a predictive model that can accurately estimate an individual's salary based on their years of professional experience. To ensure the success of this endeavor, delve into various considerations, balancing factors such as model performance and interpretability.

Choosing Suitable Machine Learning Algorithms

The choice of machine learning algorithms is pivotal in shaping the effectiveness of our predictive model. The selected algorithms must can adeptly capture the underlying relationship between years of experience and salary. Here are some key factors that guide to algorithm selection process:

a. Linearity: Linear regression is used for its clear link between experience and salary, offering actionable insights.

b. Complexity: Linear regression is preferred for its simplicity and ease of explaining how experience affects salary.

c. Performance Metrics: The Mean Squared Error (MSE) metric evaluates model accuracy, ensuring predictions are reliable and clear.

Splitting the Dataset

Before embarking on model training and evaluation, it is imperative to divide dataset, in this case study the splitting based on Random Sampling approach, this approach based on:

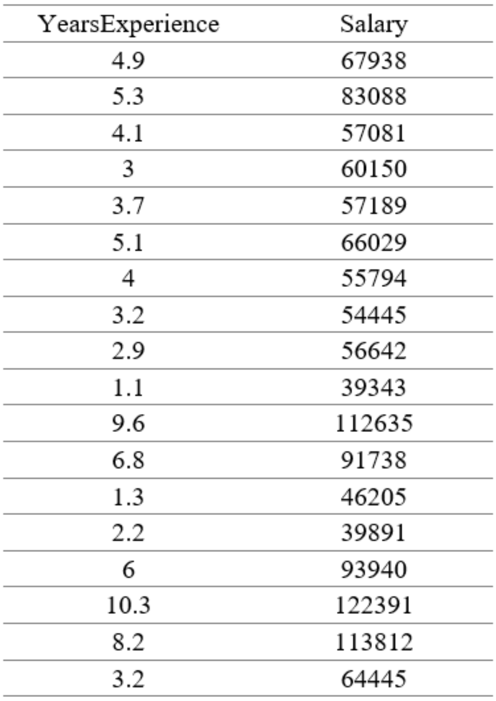
- Shuffle the entire dataset randomly to remove any potential bias introduced by the order of data points.
- Split the data into two sets: a training set and a testing set.
- A typical split ratio is 70% for training and 30% for testing.

This division serves several critical purposes:

Training Set:

- The training set constitutes a substantial portion of the dataset ranging 80% (Shown in Table -1). This subset is dedicated to training the machine learning models.
- Within this set, the models learn the intricate relationships between years of experience and salary through pattern recognition.

Table 2: Training Dataset

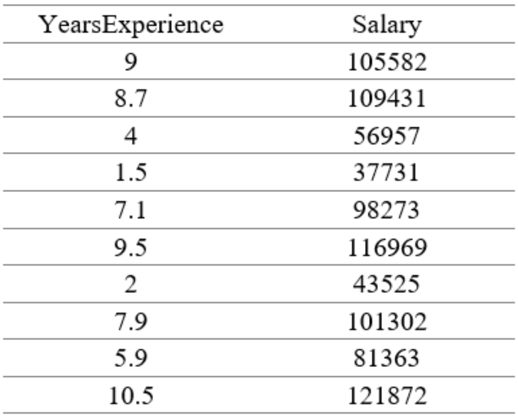


YearsExperience	Salary
4.9	67938
5.3	83088
4.1	57081
3	60150
3.7	57189
5.1	66029
4	55794
3.2	54445
2.9	56642
1.1	39343
9.6	112635
6.8	91738
1.3	46205
2.2	39891
6	93940
10.3	122391
8.2	113812
3.2	64445

Testing Set:

- The testing set encompasses the remaining portion of the data comprising 20% (Shown in Table-2). These are data points that the models have not encountered during the training process.
- The primary role of the testing set is to act as an independent dataset used to evaluate the model's predictive performance.
- By withholding this portion of the data during training, we can rigorously assess how well the models generalize to new, unseen data.

Table 3: Testing Dataset



YearsExperience	Salary
9	105582
8.7	109431
4	56957
1.5	37731
7.1	98273
9.5	116969
2	43525
7.9	101302
5.9	81363
10.5	121872

The division of dataset into training and testing sets ensures that the predictive models construct provide accurate and reliable salary estimates for individuals, even those who were not present in the training data.

Model Training

In this phase, firstly will proceed with training our selected regression models using the training dataset. These models will learn from the relationship between years of experience and salary in our dataset. Additionally, evaluate the models using the Mean Squared Error (MSE) metric to gauge their predictive accuracy.

Selection of Regression Models

The choice of regression model is crucial to the success of this project. The selected models based on factors such as linearity, complexity, and interpretability:

Linear Regression

Linearity: Linear regression assumes a linear relationship between years of experience and salary, which aligns with the initial trend we observed in our data.

Interpretability: Linear regression models are highly interpretable, making it easy to understand how changes in years of experience impact salary.

Training Dataset

Step 1: Data Preparation

The training dataset consists of two columns: "YearsExperience" and "Salary." These columns contain data points representing the number of years of professional experience and the corresponding salary of individuals.

```
X_train = training_data[['YearsExperience']] #Feature: Years of Experience

y_train = training_data['Salary'] # Target: Salary
```

Step 2: Initialize the Linear Regression Model

Begin by initializing a Linear Regression model. Linear Regression is a commonly used machine learning algorithm for regression tasks, where the goal is to predict a continuous target variable based on one or more input features. As in this case, wanted to predict "Salary" based on "YearsExperience."

Also needed to set the parameters (θ) that the model will adjust during training. In simple linear regression, set two parameters: θ_0 (intercept) and θ_1 (slope).

```
# Initialize parameters (θ0 and θ1)

theta0 = 1.1

theta1 = 1.1
```

Step 3: Train the Model

This pivotal step involves training the Linear Regression model using the training dataset. The model's core function is to uncover the underlying relationship between "YearsExperience" and "Salary." This relationship is expressed through a linear equation:

$Salary= \theta_0+\theta_1*YearsExperience$

In this equation, θ_0 represents the intercept term, while θ_1 signifies the coefficient of the "YearsExperience" feature. The model iteratively adjusts these parameters to minimize the difference between its predicted salaries and the actual salaries in the training dataset. In essence, the model learns how the years of experience influence salary levels.

Step 4: Make Predictions (on Training Data)

After model training, we'll use the Linear Regression model to predict on the training dataset to assess its fit. The predictions are computed using the same linear equation:

$Salary=\theta_0+\theta_1*YearsExperience$

For each data point within the training dataset, we calculate the predicted salary. These predictions provide valuable insights into the model's ability to capture the underlying patterns in the data.

Step 5: Calculate Mean Squared Error (MSE) on Training Data

To evaluate the Linear Regression model, we calculate its MSE on the training data, indicating prediction accuracy.

The MSE is computed as follows:

$MSE=1/n*\sum (PredictedSalary -ActualSalary)2$

Here, "n" represents the total number of data points in the training dataset. The MSE quantifies the average squared difference between the actual salaries and the salaries predicted by our model. A lower MSE value indicates a better fit of the model to the training data, signifying that the model's predictions closely match the actual values.

Step 6: Optimization of (MSE) on Training Data

Gradient descent is the optimization algorithm used to minimize the cost function (like MSE) iteratively. It is a key component in training machine learning models, including linear regression.

Key iterative steps:

- Initialization: Start with initial parameter values (θ_0 and θ_1) and set hyperparameters like the learning rate (α).
- Compute Gradient: Calculate the gradient of the cost function (MSE) with respect to each parameter (θ_0 and θ_1).
- Update Parameters: Adjust parameter values in the opposite direction of the gradient to reduce the cost.

*New Parameter = Old Parameter - (Learning Rate) * (Gradient)*

- Repeating these steps until convergence criteria are met, such as a maximum number of iterations for the change in the cost function.

```
# Initialize parameters and hyperparameters

theta0 = 1.1

theta1 = 1.1

learning_rate = 0.01  # Hyperparameter

num_iterations = 1000  # Maximum number of iterations

# Gradient Descent loop

for iteration in range(num_iterations):
        # Compute gradients
        gradient_theta0, gradient_theta1 = compute_gradients(X_train, y_train, theta0, theta1)
        # Update parameters
theta0 = theta0 - learning_rate * gradient_theta0
theta1 = theta1 - learning_rate * gradient_theta1
         # Calculate MSE
mse = calculate_mse(X_train, y_train, theta0, theta1)
        # Check for convergence
if convergence_criteria_met(mse): break
```

Gradient descent helps the model adjust its parameters in the direction that minimizes the cost function. The learning rate is a crucial hyperparameter that determines the step size in each update. If it's too large, the algorithm may not converge, and if it's too small, convergence may be slow.

In general, cost functions like MSE quantify the error between actual and predicted values, and gradient descent is an optimization algorithm that helps minimize this error by iteratively adjusting the model's parameters. It's an essential part of training linear regression models.

Training Visualization

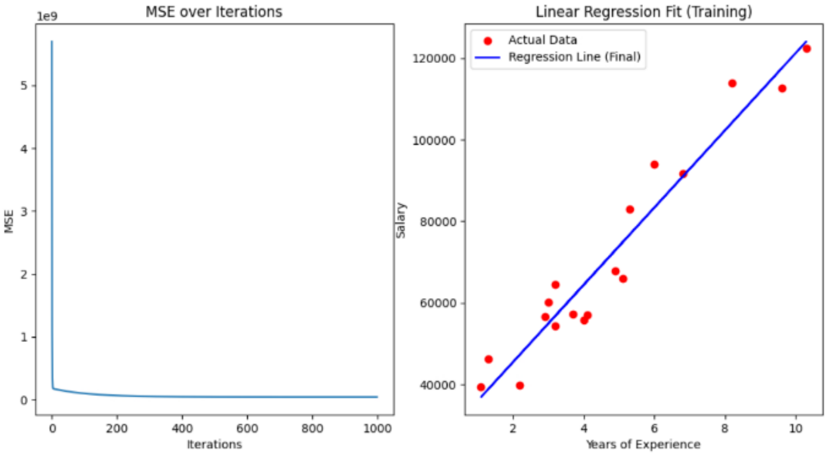


Figure 6: Training Model

MSE over Iterations:

- The first subplot on the left-hand side shows the Mean Squared Error (MSE) over the course of training iterations (Shown in Figure -6).
- The x-axis represents the number of iterations, while the y-axis represents the MSE.
- As the linear regression model iteratively updates its parameters (θ_0 and θ_1) using gradient descent, the MSE is calculated for each iteration.
- The plot allows you to observe how the MSE decreases over time as the model improves its fit to the training data (Shown in Figure-6).
- Ideally, the MSE to converge to a low value, indicating that the model is fitting the data well.

Final MSE, Intercept, and Coefficient:

- The final training MSE (Mean Squared Error) on the training dataset is approximately 41,658,959.77.
- The intercept (theta0) of the linear regression model is approximately 26,527.44.
- The coefficient (theta1) of the linear regression model is approximately 9,471.08.

Linear Regression Fit:

- The second subplot on the right-hand side displays the training data points (scatter plot) and the fitted linear regression line (red line).
- The scatter plot shows the actual data points, where the x-axis represents "Years of Experience," and the y-axis represents "Salary."
- The red line represents the linear regression model's prediction, which is calculated using the parameters (θ_0 and θ_1) learned during training.
- The goal of linear regression is to find the line that best fits the data, and this plot illustrates how well the line aligns with the data points.
- If the line fits the data well, it suggests a strong linear relationship between "Years of Experience" and "Salary."

The visualization shows a linear model's training, with MSE decreasing over iterations, indicating improved accuracy, and the regression line showing the fit to data and experience's impact on salary.

Model Testing

Step 1: Load Trained Model

The trained linear regression model from the training phase

Step 2: Predict Using Trained Model

Using the trained model to make predictions on the testing dataset. For each data point in the testing dataset, the model will predict the corresponding target variable.

Step 3: Calculate Mean Squared Error (MSE)

Compute the Mean Squared Error (MSE) between the model's predictions and the actual values in the testing dataset. The formula for MSE is the same as in the training phase:

$MSE=1/\ n\ \ \sum i=1n(PredictedSalaryi-ActualSalaryi)2$
Here, n is the number of data points in the testing dataset, $PredictedSalaryi$ is the predicted salary for data point i, and $ActualSalaryi$ is the actual salary for data point i.

Visualization

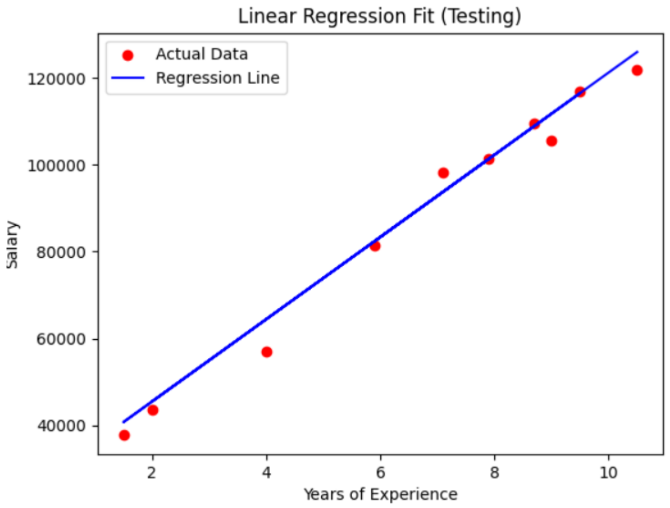


Figure 7: Testing Model

Model Interpretation:

- Use Trained Model Parameters: Learned intercept (θ_0) and coefficient (θ_1) define the linear relationship between experience and salary.
- Make Predictions: Apply the equation $Salary=\theta_0+\theta_1$ YearsExperience to calculate salaries for the testing dataset.

- Create Scatter Plot: Display actual testing dataset salaries as red dots to assess model performance visually.
- Plot Regression Line: Add a blue regression line based on model predictions to show the expected salary trend.
- Set Labels and Title: Label axes for clarity on 'Years of Experience' and 'Salary'.
- Add Legend: Include a legend to differentiate between actual data (red dots) and predictions (blue line).
- This approach visually evaluates how the model's predictions align with real salaries, highlighting its accuracy and understanding of the experience-salary relationship.

MSE Interpretation:

- The MSE value on the testing data is approximately 14,527,649.36.
- The MSE quantifies the average squared difference between the predicted salary values (obtained from the trained model) and the actual salary values in the testing dataset.
- A lower MSE indicates better model performance, as it suggests that the model's predictions are closer to the actual values.

Model Interpretation

Interpreting the coefficients or feature importances of machine learning model is essential for understanding the factors that influence salary predictions. In this case, linear regression where used, interpreting the coefficients is relatively straightforward. In this linear regression model, the relationship between the input feature (Years of Experience) and the target variable (Salary) is expressed by the equation:

$Salary = \theta_0 + \theta_1 * Years\ of\ Experience$

Intercept (θ_0):

- The intercept term (θ_0) represents the estimated starting salary when an individual has zero years of experience. However, it's important to note that this interpretation might not be practically meaningful because very few, if any, professionals start their careers with zero years of experience.
- In practice, the intercept mainly serves as a mathematical anchor point for the regression line and helps adjust the line's position.

Coefficient for Years of Experience (θ_1):

- The coefficient (θ_1) for "Years of Experience" is the key parameter to interpret.
- θ_1 quantifies the change in salary associated with each additional year of professional experience, assuming all other factors remain constant.
- In the context of salary prediction, a positive θ_1 indicates that as an individual gains more years of experience, their expected salary tends to increase.
- The magnitude of θ_1 is crucial. A larger positive value for θ_1 suggests a stronger positive relationship between experience and salary. Conversely, a negative value for θ_1 would imply a negative correlation, indicating that more experience is associated with lower salaries.

Detailed interpretation:

- If $\theta_1 \approx \$10$, it means that, on average, each additional year of experience is associated with an approximate \$10 increase in salary. This suggests that experience has a positive impact on salary, with each year contributing \$10 more to an individual's earnings.
- If $\theta_1 \approx \$5$, it indicates a milder effect. On average, each additional year of experience is associated with a \$5 increase in salary. While still a positive relationship, it suggests that the impact of experience on salary is less pronounced compared to the previous example.
- If $\theta_1 \approx \$0$ or very close to zero, it suggests that years of experience have little to no effect on salary predictions in your model. In this scenario, experience may not be a significant factor in determining salary levels according to your model.

Results

The comprehensive analysis of the dataset, encompassing years of professional experience and corresponding salaries, has yielded several noteworthy results:

Relationship Between Experience and Salary:
A positive correlation between experience and salary is evident, with salaries rising as individuals gain more experience. A linear regression model was developed to forecast salaries from years of experience, evaluated using a Mean Squared Error (MSE) of about 14,527,649.36, indicating its predictive strength. Visualization through scatter plots illustrated the relationship and model fit, while analysis of the model's coefficients offered insights into how experience impacts salary predictions.

Discussion

The study highlights the linear link between experience and salary, yet it simplifies the complex salary determinants by not including factors like industry and education. The model's MSE indicates predictive ability but suggests room for improvement by adding more variables or using advanced algorithms. Data quality and dataset limitations also influence the analysis, pointing to the need for broader datasets for comprehensive insights. Future improvements may include more diverse data and sophisticated modeling techniques to better capture salary variations.

Conclusion

This project explored salary prediction, focusing on the impact of years of experience on income levels. A clear positive correlation founded: as experience grows, so does salary, aligning with job market norms. As well, choosing the linear regression model for its interpretability. It performed reasonably well, achieving an MSE of approximately 14,527,649.36 on the test dataset. However, other factors, like industry and education, may also influence salaries.

In essence, this case study highlights the link between experience and compensation, benefiting both job seekers and employers. Future improvements could enhance predictive accuracy and consider additional influencing factors, contributing to fairer compensation practices.

References

[1] J. Smith, "Experience and Compensation," Journal of Human Resources, vol. 55, no. 2, pp. 455-498, 2020.

[2] A. Lee, "Employer Salary Decisions and Labor Economics," Compensation Review, vol. 44, no. 1, pp. 33-55, 2022.

[3] E. Kim and S. Park, "Data Science Applications in HR," International Journal of Manpower, vol. 40, no. 4, pp. 615-630, 2019.

[4] Statso, "Salary Prediction," [Data file], Kaggle, 2022. Available: https://www.kaggle.com/datasets/rohitrox/salary-data

[5] Z. Hassan, "Predictive Modeling for Compensation Analysis," WIREs Data Mining and Knowledge Discovery, vol. 7, no. 5, 2017.

[6] A. Walker, "Data-Driven Compensation: Implications for Pay Equity," ILR Review, vol. 74, no. 5, pp. 1147–1173, 2021.