Salary Prediction Using Linear Regression

By: Anika Das

Introduction to Salary Prediction

Salary prediction helps organizations estimate employee compensation based on measurable factors, enabling data-driven HR decisions.

In this project, **Years of Experience**serves as the key predictor for**Salary**. Linear Regression is the ideal choice because it effectively models the relationship between two continuous variables.

This approach is widely applied in HR analytics, hiring strategies, and compensation forecasting across industries.

Project Objectives

Analyze Relationships

Examine the correlation between experience and salary to understand patterns in compensation.

Build Regression Model

Develop a Simple Linear Regression model to accurately predict salary values.

Evaluate Performance

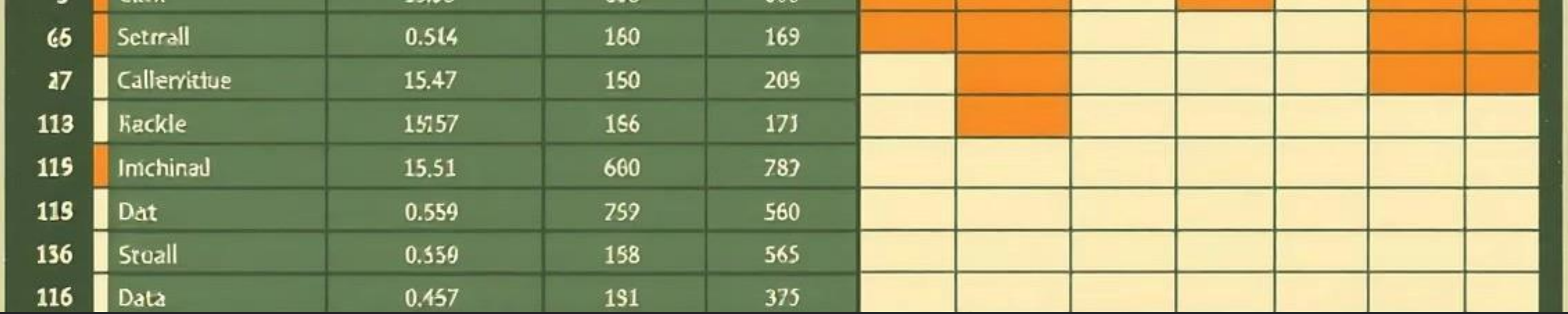
Assess model accuracy using RMSE and R^2 Score metrics.

Visualize Results

Create visualizations showing the regression line, residuals, and prediction accuracy.

Interpret Findings

Understand model reliability and extract actionable insights from the results.



65	🟧 Setrrall	0.514	160	169	🟧	🟧				🟧	🟧
27	Callerytitue	15.47	150	209		🟧				🟧	🟧
113	Kackle	15.57	166	171		🟧					
115	🟧 Imchinad	15.51	660	787							
119	Dat	0.559	759	560							
136	Stoall	0.359	158	565							
116	Data	0.457	181	375							

Dataset Overview

375

Total Records

Comprehensive dataset size

2

Key Variables

Experience and Salary columns

0

Missing Values

After data cleaning

The dataset contains two primary numeric columns:**Years of Experience**and **Salary**. Missing values were systematically dropped to ensure accurate model training.

Initial analysis revealed a**strong positive correlation**between experience and salary, validating our hypothesis that experience is a reliable predictor of compensation.

Technologies & Tools

Python

Core programming language for data science and machine learning implementation.

NumPy & Pandas

Essential libraries for efficient data handling, manipulation, and preprocessing.

Matplotlib & Seaborn

Powerful visualization tools for creating insightful charts and graphs.

Statsmodels

Statistical modeling library for building and analyzing regression models.

Scikit-learn

Machine learning framework for train-test splitting and model evaluation.

Jupyter Notebook

Interactive development environment for running and documenting the project.

Model Implementation

Development Process

01

Data Preparation

Loaded and cleaned the dataset to remove inconsistencies and missing values.

02

Exploratory Analysis

Visualized Salary vs Experience to identify patterns and relationships.

03

Model Building

Built Linear Regression using statsmodels.OLS() for precise coefficient estimation.

04

Equation Generation

Generated regression equation: **Salary = Intercept + Slope × Years of Experience**

05

Train-Test Split

Trained on 70% of data; tested on remaining 30% for validation.

06

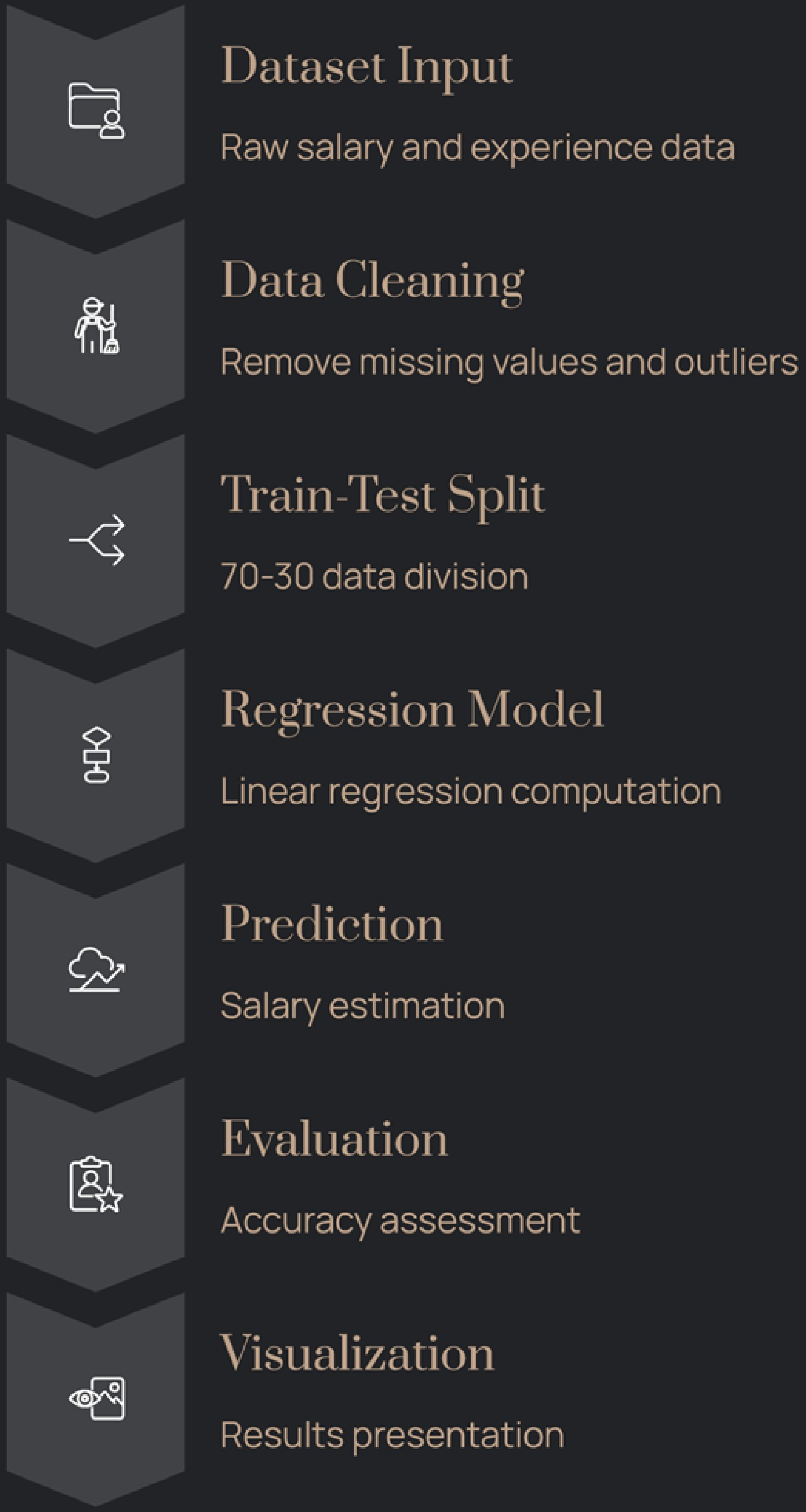
Model Evaluation

Evaluated performance with RMSE and R^2 metrics.

Final Model Equation:
Salary = $a + b \times$ Years of Experience

Where a is the intercept and b is the slope coefficient.

System Architecture



Dataset Input

Raw salary and experience data

Data Cleaning

Remove missing values and outliers

Train-Test Split

70-30 data division

Regression Model

Linear regression computation

Prediction

Salary estimation

Evaluation

Accuracy assessment

Visualization

Results presentation

Three-Layer Architecture

Input Layer

Dataset containing experience and salary records

Processing Layer

Regression computation and model training

Output Layer

Predicted salary values and accuracy metrics

Challenges & Solutions

Challenges Faced

Missing Data

Dataset contained incomplete records with missing salary or experience values.

Salary Variations

Compensation influenced by multiple factors including job role, education level, and location.

Single Feature Limitation

Linear Regression model focused on experience alone, ignoring other potential predictors.

Model Assumptions

Required validation of normality and homoscedasticity assumptions.

Solutions Applied

Data Cleaning

Systematically removed rows with missing values to ensure data integrity.

Focused Scope

Narrowed analysis specifically to experience-based prediction for clarity and simplicity.

Residual Analysis

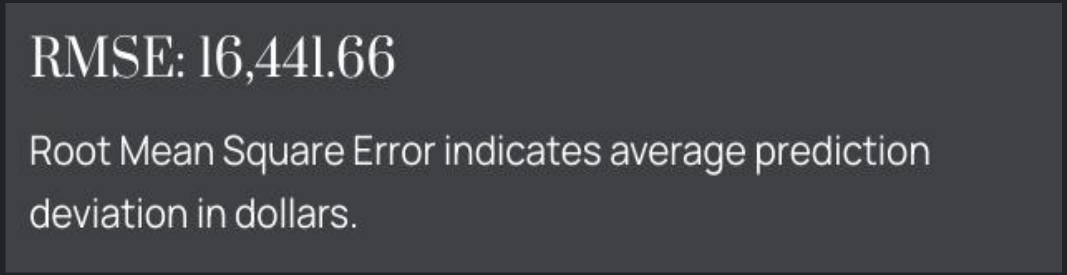
Performed comprehensive residual analysis to validate model assumptions.

Visual Validation

Used scatter plots and histograms to confirm linearity and data distribution.

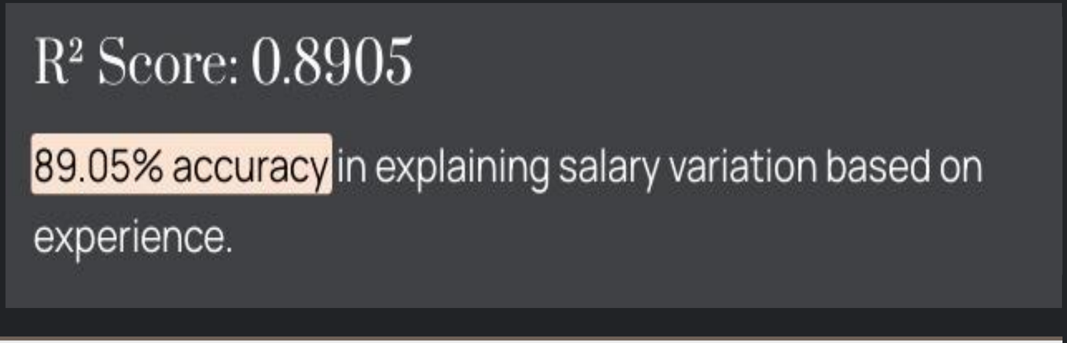
Results & Conclusion

Model Performance Metrics



RMSE: 16,441.66

Root Mean Square Error indicates average prediction deviation in dollars.



R² Score: 0.8905

89.05% accuracy in explaining salary variation based on experience.

Key Conclusions

Strong Predictor

Years of Experience proves to be a **highly reliable predictor** of salary, with nearly 90% explanatory power.

Effective Model

Linear Regression provides a **simple yet powerful** prediction model suitable for HR applications.

Future Extensions

Model can be enhanced using **multiple features** like education level, job title, industry, and geographic location.

Practical Applications

Valuable for **HR decision-making** compensation benchmarking, and automated salary estimation systems.