Employee Salary Prediction

Presented by:-
Nensi Patoliya

Introduction

- ❑ Salary prediction is a vital aspect of modern workforce management, bringing the gap between employer expectations and employee compensation.
- ❑ It uses machine learning approach to estimate employee salaries based on various factors like experience, job role, location and education.

Why it is important?

- ❑ Ensures fair compensation and reduce bias.
- ❑ Helps in budgeting and financial planning.
- ❑ Assists job seekers in understanding expected salary ranges.

Problem Statement

01 **Inconsistent salary decisions:-** Salary determination is often subjective, leading to pay disparities for similar roles and experience.

02 **Lack of Standarized Benchmarking:-** Companies struggle to set competitive and fair salaries without data driven insights.

03 **Market Trends and Inflation:-** Changing industry demands and economic factors make it difficult to maintain consistent salary structures.

Project objective

❑ Predict employee salaries using machine learning.

❑ Ensure fair and data-driven salary decisions.

❑ Assist HR in designing competitive pay structures.

❑ Deploy as a web app for real time predictions.

Data Collection & Data Preprocessing

The salary dataset contains **6704 rows** and **6 columns** containing the following data:-

1. Age
2. Gender
3. Education level
4. Job title
5. Years of experience
6. Salary.

Preprocessing of data includes handling info of data, unique values, duplication values in data, finding null values and describe data which shows the total values, min, average and max values of the data.

Libraries Used: Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn, joblib

Model Selection and Training

Model Selection Process:-

❑ **Linear Regression**:

Assumes a linear relationship between input features and salary.

Sensitive to outliers.

❑ **Decision Tree Regression**:

Splits data into nodes based on conditions.

Can lead to overfitting if not pruned properly.

❑ **Random Forest Regression**:

Uses multiple decision trees and averages results.

More robust and less prone to overfitting.

Model Selection and Training

Model Training :-

- Split the dataset into training (80%) and testing (20%) sets.
- Train models using Scikit-learn.
- Use fit() method to train each model on training data.
- Tune hyperparameters using Grid Search or Random Search.

Model Evaluation Metrics

Evaluating Model Performance:-

- ❑ **Mean Absolute Error (MAE):**
  Measures the average absolute difference between actual and predicted values.
  Lower values indicate better performance.

- ❑ **Mean Squared Error (MSE):**
  Measures the average squared differences between actual and predicted values.
  More sensitive to larger errors due to squaring.

- ❑ **R^2 Score (Coefficient of Determination):**
  Indicates how well the model explains variance in salary.
  Closer to 1 means a better fit; negative values indicate poor predictions.

Comparison of Models

Model	MAE	MSE	R^2 Score
Linear Regression	24082.456847	9.300659e+08	0.673828
Decision Tree	6608.343468	1.472971e+08	0.948343
Random Forest	6732.652175	1.335864e+08	0.953152

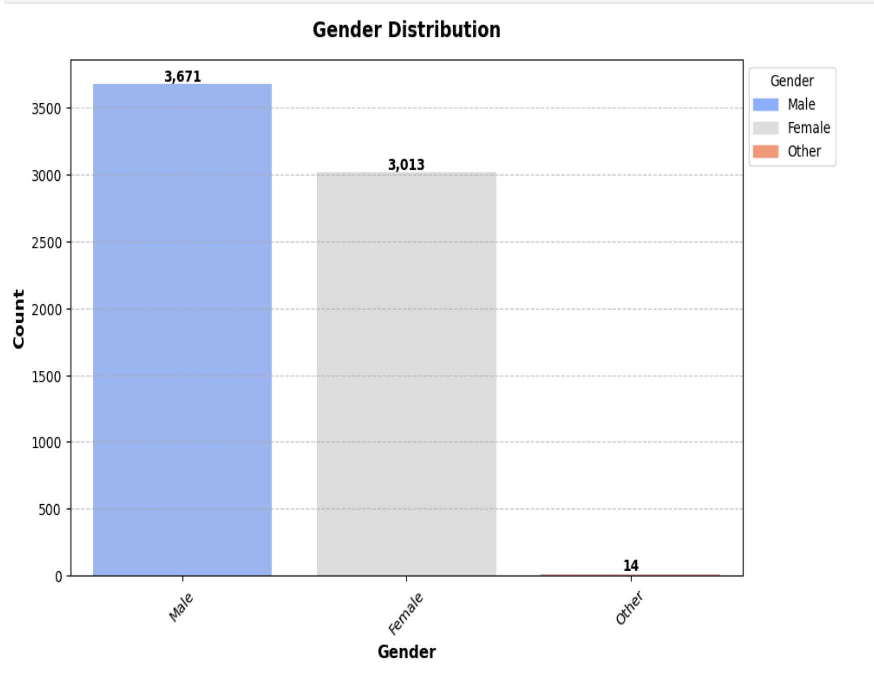
Key Insight:- Random Forest provide the most accurate predictions.

Exploratory Data Analysis (EDA)

- ❑ EDA is a data analytics process that aims to understand the data in depth and learn its different characteristics, often using visual means.
- ❑ This allows one to get a better feel for the data and find useful patterns.
- ❑ It helps you prepare your dataset for analysis.
- ❑ Allows a machine learning to predict our dataset better.
- ❑ Give you more accurate results.
- ❑ It also helps us to choose a better machine learning models.

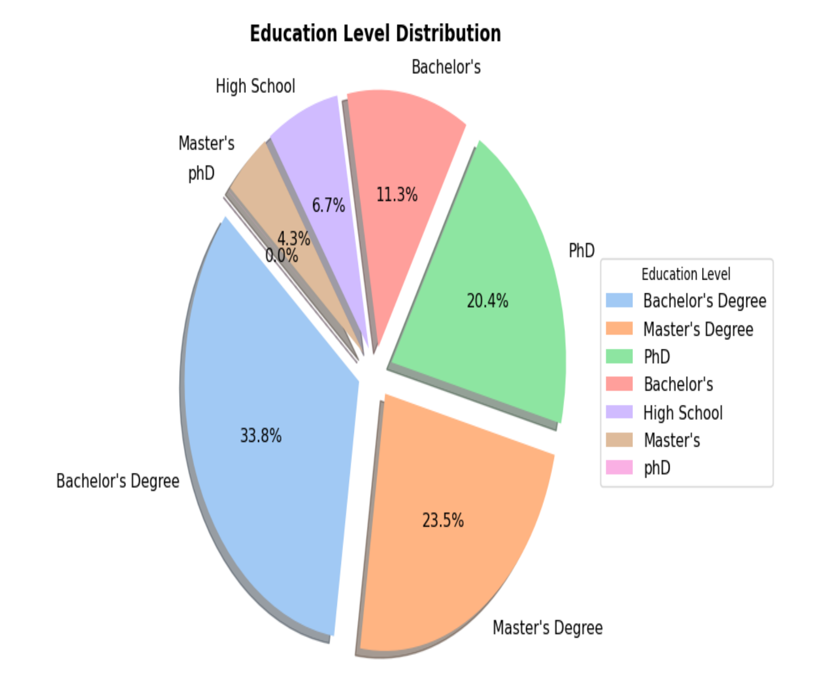
Visuals:- Bar charts, Scatter plots, histogram, lineplot.

Gender Distribution



- ❑ The dataset is imbalanced in terms of gender representation, which could influence model predictions.
- ❑ The low count for the "Other" category may require special handling, such as grouping into another category or using techniques to balance representation.
- ❑ Gender imbalance might reflect industry trends but could also introduce bias in salary predictions.
- ❑ Further analysis may be needed to determine if gender impacts salary predictions significantly.

Education Level Distribution

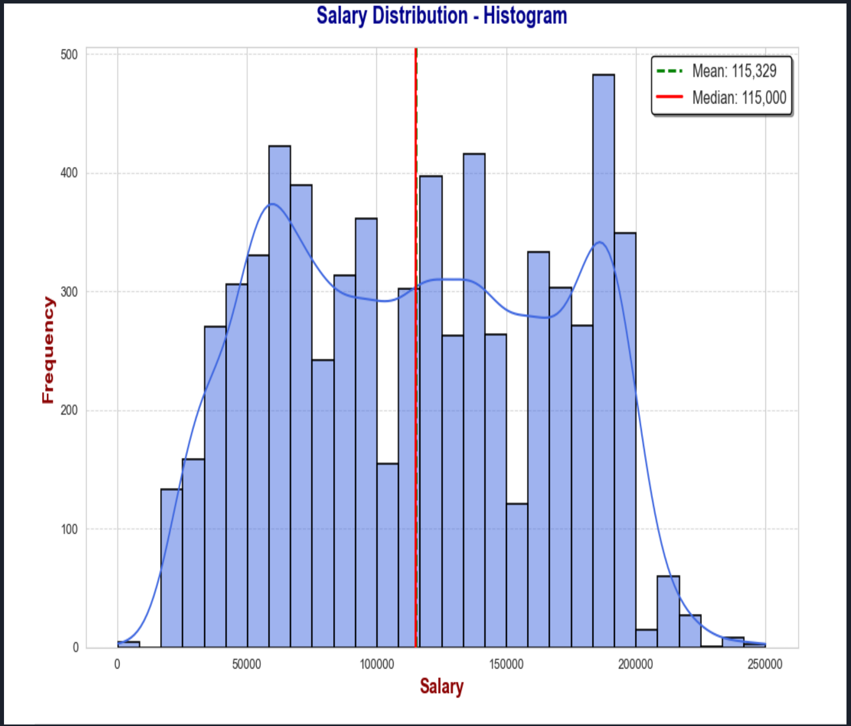


❑ Higher education levels **(Master’s & PhD)** might correlate with **higher salaries**, as they often lead to specialized and well-paying roles.

❑ the relatively small number of employees with **only High School education** could indicate a dataset bias towards **degree holders**.

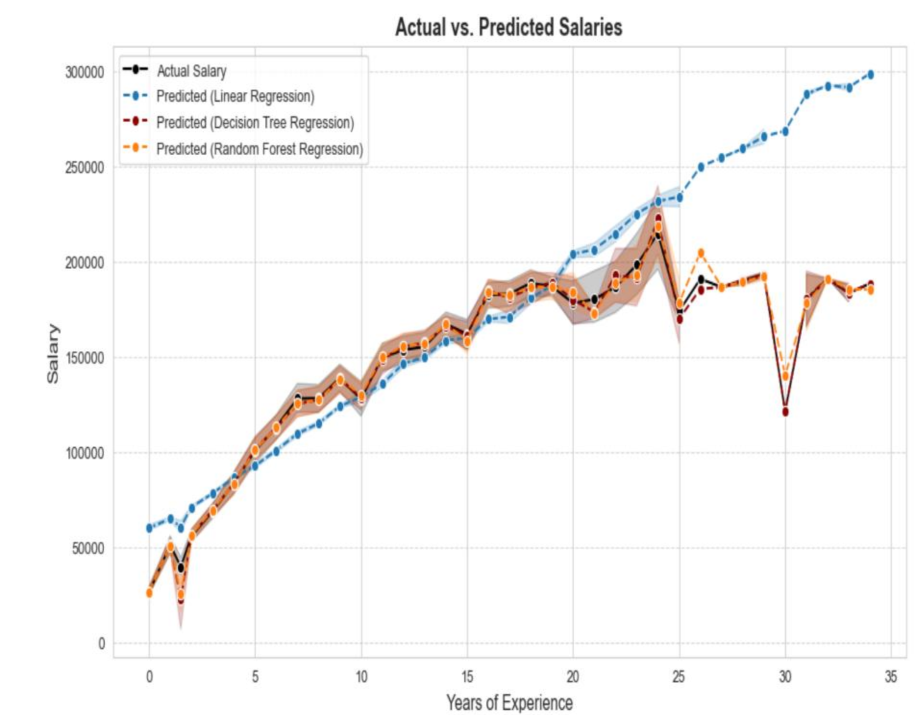
❑ The dataset might **lack representation** of lower education levels, which could impact the salary prediction model’s ability to generalize across different groups.

Salary Distribution Analysis



- ❑ The histogram represents salary distribution, showing the frequency of different salary ranges.
- ❑ **Mean Salary:** 115,329 (green dashed line)
- ❑ **Median Salary:** 115,000 (red solid line).
- ❑ The distribution appears **bimodal**, indicating two peaks, suggesting the presence of two salary groups.
- ❑ Some skewness is present, but the mean and median are close, implying a relatively balanced distribution.

Comparing Model Predictions to Actual Salaries



- ❑ The graph compares actual vs. predicted salaries using Linear Regression, Decision Tree, and Random Forest models.
- ❑ Linear Regression underestimates higher salaries, Decision Tree captures variations but overfits, while Random Forest provides the best balance.
- ❑ Random Forest is the most reliable model, offering accurate predictions with minimal fluctuations and better confidence.

Model Deployment

Model Saving:

- ❑ Trained models are saved using **Joblib** for future predictions without retraining.
- ❑ Ex. joblib.dump(model, "random_forest.pkl")

Deployment Strategy:

- ❑ **Platform:** Deploying the model using **Streamlit** for an interactive web application.
- ❑ **User Input:** Users can enter their age, experience, education level, and gender to predict salary.
- ❑ **Backend:** The saved model will load and make predictions in real-time.

Conclusion & Future Enhancements

Key Takeaways:-

- ❑ Machine learning models can accurately predict salaries based on multiple factors.
- ❑ Feature selection & data preprocessing play a crucial role in prediction accuracy.
- ❑ A deployed web app can provide salary estimates to HR professionals and job seekers.

Future Improvements:-

- ❑ **Feature Engineering:** Introduce additional variables like industry, job role, and location to improve predictions.

- ❑ **Hyperparameter Tuning:** Further optimize models using **GridSearchCV** and **Bayesian Optimization**.

- ❑ **Cloud Integration:** Deploy models on **AWS/GCP** with an API for scalability and real-time data processing.

Thank you!