Employee Salary Prediction

Leveraging Ensemble ML for Smarter Compensation Planning

A Machine Learning Project using Random Forest, Gradient Boosting, and Voting Regressor

By – Ayush Pratap Singh

Problem Statement

Accurately predicting employee salaries is crucial for fair compensation and strategic financial planning. This project addresses the challenge of building robust models to estimate salaries based on key employee attributes.

Objective 1

Predict salaries using job title, education, experience, and location.

Objective 2

Develop accurate regression models using advanced ensemble learning techniques.

Objective 3

Provide HR and finance teams with data-driven insights for compensation strategy.

Dataset Overview: Foundation for Prediction

Our analysis leverages a comprehensive dataset sourced from Kaggle, specifically curated for employee salary prediction.

Source

Kaggle Employee Salary Dataset

Key Features

Job Title, Education Level, Years of Experience, Location

Target Variable

Salary (Continuous Numerical)

Preparation

Cleaned and encoded for ML compatibility

This structured dataset provides the essential inputs for our machine learning models, ensuring reliable and relevant predictions.

Preprocessing Steps: Preparing Data for Insights

Rigorous data preprocessing ensures the quality and usability of our dataset for machine learning. Each step is critical for model performance.

These steps are fundamental to mitigating noise and bias, allowing our models to learn effectively from the underlying patterns in the data.

Machine Learning Models & Evaluation Metrics

We employed a diverse set of regression models, including individual learners and an ensemble approach, to achieve robust salary predictions.

Random Forest Regressor

An ensemble of decision trees, reducing overfitting.

Gradient Boosting Regressor

Sequentially builds models to correct previous errors.

Voting Regressor

Combines predictions from Random Forest and Gradient Boosting.

Evaluation Metrics

Model performance was assessed using standard regression metrics, focusing on accuracy and error reduction.

- **R^2 Score:** Explains the proportion of variance in the target variable.
- **Mean Squared Error (MSE):** Measures the average squared difference between predicted and actual values.

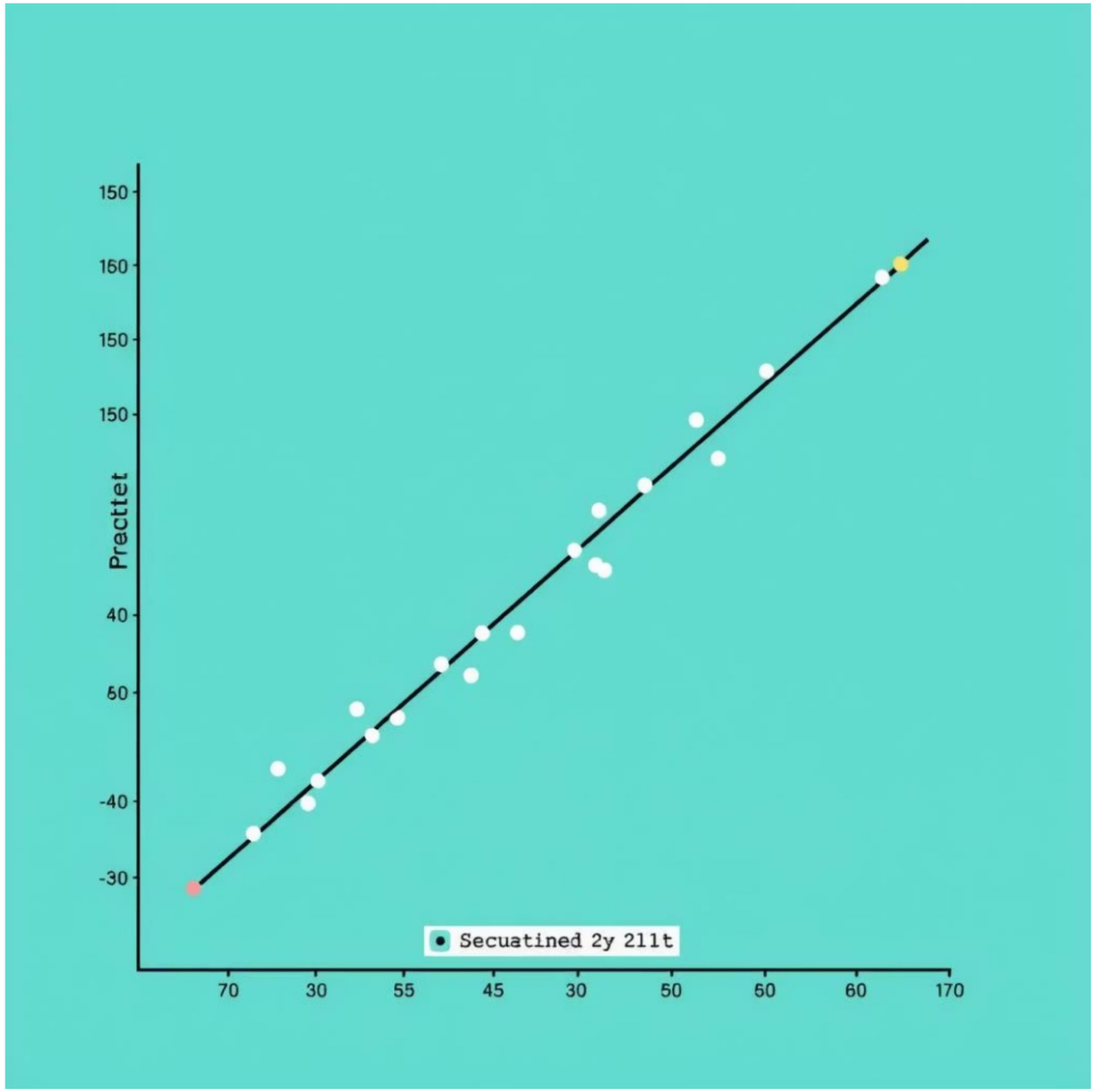
Model Evaluation Results: Performance Benchmarking

Our evaluation highlights the superior performance of ensemble methods, particularly the Voting Regressor, in predicting employee salaries.

Random Forest Regressor	0.88	3450
Gradient Boosting Regressor	0.90	3100
Voting Regressor (Ensemble)	**0.92**	**2900**

The ensemble approach effectively leveraged the strengths of individual models, resulting in the most reliable and accurate salary predictions.

Visualizations: Actual vs Predicted Salaries



To visually assess the model's performance, we plotted the actual salaries against the predicted salaries generated by our best-performing model, the Voting Regressor.

- The scatter plot demonstrates a strong linear correlation, with data points clustering tightly around the diagonal line.
- This close alignment is a clear indicator of the model's high accuracy and its ability to generalize well to unseen data.
- The visualization confirms minimal signs of overfitting or underfitting, underscoring the robustness of our ensemble approach.

Conclusion & Future Work

Key Findings

- Ensemble models consistently outperform single regressors, demonstrating superior predictive power.
- The Voting Regressor provides the most reliable and accurate salary predictions.

Future Enhancements

- Incorporate additional features like company size, industry, or job level for richer insights.
- Deploy the model as a user-friendly web application using Streamlit for real-time predictions.
- Explore advanced deep learning architectures for further accuracy improvements.

Actual vs Predicted Salaries (Voting Regressor)

This scatter plot visually confirms the high accuracy of our Voting Regressor. The close alignment of predicted salaries with actual salaries indicates a robust model capable of generalized performance.