Predicting Employee Salaries Using Ensemble Machine Learning

Leveraging advanced ML techniques to accurately forecast compensation

Project Overview and Goals

This presentation outlines our approach to predicting employee salaries using an ensemble of machine learning models. Our goal is to leverage various data points to build a robust and accurate predictive tool.

Predict Salaries

Utilize features like job title, experience, education, and location to estimate employee salaries.

Kaggle Dataset

Source and process comprehensive salary datasets from Kaggle for model training.

Ensemble Models

Implement Random Forest, Gradient Boosting, and Voting Classifiers for enhanced accuracy.

Dataset Overview and Feature Description

Our model is built upon a detailed dataset from Kaggle, comprising 375 rows with various employee attributes. These features are critical for capturing the nuances that influence salary.

- **Job Title:** Specific role within the organization.
- **Years of Experience:** Total professional experience.
- **Education Level:** Highest degree attained (e.g., Bachelor's, Master's, PhD).
- **Location:** Geographic region affecting compensation.
- **Age:** Employee's age, potentially correlated with experience.
- **Industry:** Sector in which the job is located.

Data Preprocessing and Feature Engineering

To ensure model robustness, rigorous preprocessing was essential. This involved handling missing data, transforming categorical variables, and scaling numerical features.

Handle Missing Values

Imputation (mean, median) or removal of rows with incomplete data to maintain data integrity.

Encode Categorical Variables

Applied Label Encoding or One-Hot Encoding to convert job titles, education, and locations into numerical formats suitable for ML algorithms.

Scale Numerical Features

Used StandardScaler to normalize numerical attributes like experience and age, preventing features with larger scales from dominating the model.

Feature Selection

Identified and removed irrelevant or redundant features to reduce noise and improve model efficiency.

Exploratory Data Analysis (EDA)

Our EDA revealed key insights into the dataset, informing our feature engineering and model selection. Visualizations played a crucial role in understanding the data patterns.

- **Salary Distribution:** Observed a right-skewed distribution, necessitating a log transformation for normalization.
- **Correlation Analysis:** Identified strong positive correlations between salary and features like years of experience and education level.
- **Visualizations:** Utilized boxplots to compare salaries across different job titles and heatmaps to visualize feature correlations, uncovering hidden relationships.

Model Selection: Ensemble Techniques

We selected three powerful ensemble models to enhance prediction accuracy and robustness. Each model brings unique strengths to the task of salary prediction.

Random Forest

An ensemble of decision trees, effective for handling high-dimensional data and reducing overfitting by averaging predictions.

Gradient Boosting

A sequential learning approach (e.g., XGBoost) that iteratively builds new models to correct errors from previous ones, improving overall accuracy.

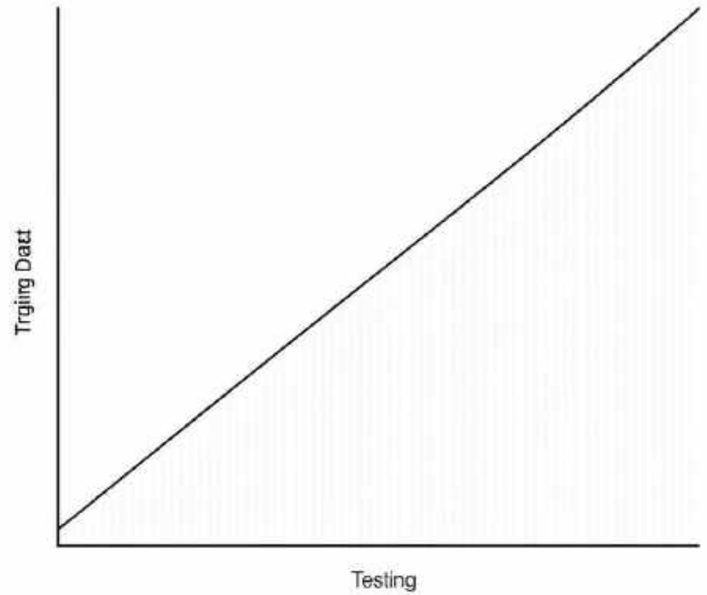
Voting Classifier

Combines predictions from multiple individual models (Random Forest and Gradient Boosting) to produce a more stable and accurate final prediction, leveraging collective intelligence.

Model Training and Hyperparameter Tuning

Our training regimen focused on optimizing model performance through strategic data splitting and fine-tuned hyperparameter adjustments for each ensemble method.

The dataset was split into an 80/20 train/test ratio, with stratification on salary brackets to ensure representative samples across all ranges. This prevents bias towards certain salary levels.



- **Random Forest:** Initialized with 100 estimators; max_depth was optimized using GridSearchCV to prevent overfitting and capture optimal complexity.
- **Gradient Boosting:** Key hyperparameters like learning_rate and n_estimators were optimized through iterative tuning to achieve peak performance.
- **Voting Classifier:** Implemented with weighted soft voting, allowing us to assign different importance to Random Forest and Gradient Boosting models based on their individual performance.

Evaluation Metrics and Performance Results

Our models were rigorously evaluated using standard regression metrics. The Voting Classifier consistently delivered superior performance, demonstrating the power of ensemble methods.

R^2 Score	0.85	0.88
MAE	3200	2900
RMSE	4500	4100

While both Random Forest and Gradient Boosting performed well, Gradient Boosting showed a notable improvement in R^2 and MAE. The Voting Classifier further enhanced stability and overall accuracy, providing the most robust predictions.

Feature Importance and Key Insights

Understanding which features drive salary predictions is crucial for actionable insights. Our analysis highlights the most influential factors.

The feature importance analysis, particularly from the Random Forest model, validated our initial hypotheses. These insights are invaluable for HR departments and recruiters, enabling them to make data-driven decisions regarding compensation strategies and talent acquisition.

Deployment and Future Work

Our project aims to provide practical solutions and continuous improvement for salary prediction.

Deployment Options

- Develop a user-friendly Flask or Streamlit web application for real-time salary predictions.
- Allow users to input employee features and receive instant salary estimates.

Future Enhancements

- Incorporate more granular features like specific skills or performance metrics.
- Explore advanced deep learning models for even greater accuracy.
- Develop interactive dashboards for deeper insights into salary trends.

Ethical Considerations

- Implement fairness metrics to mitigate biases in salary predictions.
- Ensure transparency in model decisions to promote equitable compensation practices.

BY SHUBHANGI SRIVASTAVA, BTECH CSE-2ND YR

GREATER NOIDA INSTITUTE OF TECHNOLOGY (GNIOT),GREATER NOIDA