Employee Salary Prediction Using Ensemble Machine Learning Models

Final Group Project – YCBS 273

Group1

1.Introduction

1.1 Background and Motivation

Employee compensation is a cornerstone of human resource management, impacting workforce motivation, retention, and productivity. Traditionally, salary setting depends heavily on managerial discretion, market surveys, and historical benchmarks. However, these approaches risk subjectivity and lagging behind evolving workforce dynamics.

The advent of big data and machine learning technologies offers a promising alternative. By leveraging detailed employee demographic, performance, and tenure data, organizations can develop predictive models to estimate salaries more objectively and accurately. Such models can support equitable pay structures, identify underpaid segments, and inform targeted retention strategies.

1.2 Objectives

This study aims to Develop robust machine learning models predicting employee salaries from multi-dimensional data; Apply advanced feature engineering and selection techniques to enhance model efficiency and interpretability; Compare the predictive performance of three ensemble methods: Gradient Boosting, XGBoost, and Random Forest; Use SHAP explainability to interpret model decisions, providing transparency for HR stakeholders.

2. Literature Review

Classical salary prediction has largely relied on linear regression, which models linear relationships between salary and predictors like education and experience. While straightforward, linear regression fails to capture complex nonlinearities and interaction effects often present in compensation data (Cameron & Trivedi, 2005).

Ensemble methods, particularly tree-based algorithms, offer enhanced flexibility and predictive power. Random Forest aggregates decision trees trained on bootstrapped data, reducing variance (Breiman, 2001). Gradient Boosting sequentially fits trees to residuals, reducing bias and variance (Friedman, 2001). XGBoost optimizes gradient boosting with regularization and parallel processing, excelling in large-scale problems (Chen & Guestrin, 2016).

Complex models pose interpretability challenges, critical in HR contexts. SHAP (Lundberg & Lee, 2017) decomposes predictions into additive feature contributions, offering insights into model behavior and supporting ethical, transparent decision-making.

3. Data Description and Preprocessing

3.1 Dataset Overview

The dataset includes 6,899 anonymized employee records with 31 features spanning demographics (Age, Gender), employment data (YearsAtCompany, JobRole), performance indicators (TrainingPerformance), and compensation (MonthlyIncome).

3.2 Data Cleaning and Missing Value Imputation and Categorical Variable Encoding

Initial data quality assessment revealed missing values in features like NumCompaniesWorked and TotalWorkingYears. We employed KNNImputer with k=5 neighbors, preserving data distribution and avoiding mean imputation biases.Categorical variables were processed as follows: Binary variables encoded as 0/1. Nominal variables (e.g., EducationField) one-hot encoded. Ordinal variables transform preserving order.

3.3 Feature Engineering and selection

Experience was categorized into four binary features (exp_New: <3 years, exp_Mid: 3-7 years, exp_Senior: 7-15 years, exp_Expert: >15 years) to capture nonlinear tenure effects.

Interaction features like JobSatisfaction × EnvironmentSatisfaction were created.

Logarithmic transformation was applied to MonthlyIncome to reduce skewness.

SelectKBest using f_regression, selecting top 20 features.

RFE with XGBoost estimator, recursively eliminating least important features until 25 remained.

Final selected features included Age, YearsAtCompany, EducationLevel_encoded, StockOptionLevel, TrainingPerformance, EnvironmentSatisfaction, and experience-level dummies.

3.4 Data Challenges & Limitations

Despite preprocessing efforts, several data-related challenges remain:

Skewed Distributions: salary and tenure-related variables were highly skewed, requiring log transformation and binning to ensure effective model learning.

Missing and Incomplete Information:key variables such as total working years or number of companies worked had missing values. We used KNN imputation to address these gaps, but uncertainty remains.

Limited Feature Scope: the dataset lacks important qualitative factors such as soft skills, team feedback, or industry-specific context. This may limit the model's ability to fully explain salary decisions.

Multicollinearity Risk: strong correlation was observed between SelfRating and ManagerRating, raising the possibility of redundancy and inflation of feature importance.

Potential Historical Bias: the dataset reflects historical salary decisions, which may encode past inequities (e.g., gender, age, or role-based bias), even if sensitive variables were removed.

Generalizability Limitations: the model is trained on a specific employee population. Applying it to different sectors, geographies, or job types may reduce its accuracy and fairness without retraining.

4.Model Development and Tuning

4.1.Model Selection Rationale

Three ensemble-based regressors were chosen for their proven effectiveness in handling structured datasets with moderate non-linearity and interactions among variables: Gradient Boosting Regressor: Builds models sequentially by minimizing residual errors. Known for high accuracy but sensitive to overfitting if not tuned properly; XGBoost Regressor: An optimized and scalable version of gradient boosting, with regularization terms that make it less prone to overfitting and highly efficient for structured tabular data; Random Forest Regressor: Constructs multiple decision trees and averages their predictions. It is less sensitive to hyperparameter tuning and offers robust generalization.

These models were ideal given the structure of our dataset, which included both numerical and categorical features and required non-linear modeling capabilities.

4.2. Data Preparation and Feature Engineering

Before model training, significant preprocessing and feature engineering were performed. The key steps included: Handling missing values: Imputed missing categorical values using the mode, and numeric values using the median; Encoding categorical variables: One-hot encoding was applied to categorical variables like 'Education Level' and 'Job Role'; Feature transformation: Skewed variables were log-transformed to normalize distributions (e.g., 'Salary'); New feature creation: Experience was binned into levels (e.g., Junior, Mid, Senior, Expert) and added as four binary features to improve interpretability; Feature selection: Recursive Feature Elimination (RFE) with XGBoost and SelectKBest (f_regression) helped select the most predictive subset of features.

The selected final features included age, experience level, number of trainings, education level, stock option level, performance rating, and length of service.

4.3. Hyperparameter Tuning

To improve model generalization and reduce overfitting, grid search combined with cross-validation was performed. Key hyperparameters tuned were:

Gradient Boosting Regressor: n_estimators: Number of boosting stages, learning_rate: Shrinks contribution of each tree, max_depth: Controls complexity of trees

XGBoost Regressor: n_estimators, max_depth, learning_rate, subsample: Fraction of samples used per tree, colsample_bytree: Fraction of features used per tree

Random Forest Regressor: n_estimators: Number of trees in the forest, max_depth: Maximum tree depth, min_samples_split: Minimum samples to split an internal node

4.4. Performance Evaluation

Models were evaluated using three metrics:

RMSE (Root Mean Squared Error): Penalizes large errors more significantly.

R^2 (R-squared): Indicates proportion of variance explained by the model.

CV_R2_mean: Cross-validated R^2 mean to assess generalizability.

4.5. Model Interpretation with SHAP

To improve the interpretability of the best-performing model (Gradient Boosting), SHAP (SHapley Additive exPlanations) values were used. SHAP provided local and global explanations of feature importance, revealing that Age and Length of Service were positively correlated with salary; Stock Option Level and Performance Rating also had significant predictive power; Number of Trainings had a non-linear effect, with diminishing returns after a certain threshold. The SHAP summary plot and force plot confirmed the logical consistency of the model's decision-making process.

5. Results

5.1 Performance Metrics

Model	RMSE	R^2	CV R^2 mean
Gradient Boosting	58,315.11	0.6426	0.6329
XGBoost	60,290.63	0.6180	0.6348
Random Forest	67,415.94	0.5224	0.5886

Gradient Boosting outperformed other models with significantly lower RMSE and higher R^2, indicating superior predictive accuracy and robustness.

5.2 Predicted vs Actual Analysis

Scatterplots demonstrated tight alignment between Gradient Boosting predictions and actual salaries, especially in mid-to-high salary ranges. XGBoost showed moderate deviation, while Random Forest exhibited larger residuals.

5.3 Feature Importance and SHAP Analysis

Feature importance rankings confirmed Age, YearsAtCompany, and EducationLevel_encoded as the dominant predictors. SHAP summary plots highlighted the magnitude and direction of feature contributions, elucidating complex nonlinear effects and interactions.

6. Discussion and Ethical Considerations

The superior performance of Gradient Boosting aligns with literature on ensemble models' ability to capture complex patterns. The combination of feature selection and hyperparameter tuning enhanced model stability and accuracy.

SHAP interpretability bridges the gap between predictive power and trustworthiness, crucial for HR applications where opaque models can face resistance.

Limitations include lack of external economic variables, absence of soft skill metrics, and static snapshot data rather than longitudinal records. Incorporating these in future work could further improve predictions.

7.Ethical Considerations

Transparent models foster fair salary practices and help identify potential biases embedded in historical data. Ensuring GDPR compliance and employee consent is critical. Predictive models should augment—not replace—human judgment to avoid unintended discrimination.

8. Conclusion

This study develops and evaluates machine learning models for employee salary prediction, demonstrating that Gradient Boosting achieves the best balance of accuracy and interpretability. SHAP analysis provides actionable insights for transparent HR decision-making. The methodology supports data-driven compensation management, promoting fairness and efficiency.

9. Member Contribution

Member	Contribution
Zihan SU	Model training & hyperparameter tuning
Fan XIA	SHAP analysis & business impact
Teng ZHANG	Data preprocessing & EDA
Taoyu MA	Report & presentation design
Chen TIAN	Business understanding & ethics section