Employee Salary Prediction Using Machine Learning

Data-Driven Insights for Fair and Strategic Compensation

Final Group Project – YCBS 273

Group 1

Fan XIA

Zihan SU

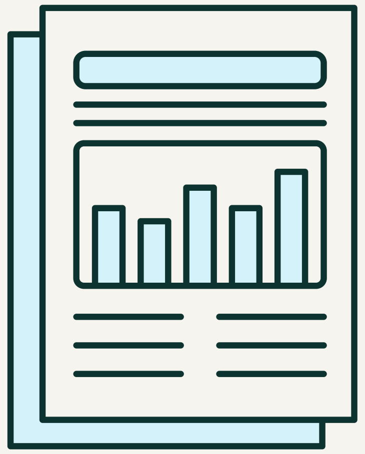
Taoyu MA

Teng ZHANG

Chen TIAN

Business Problem

Inconsistent salary decisions hurt retention and morale



Traditional models lack transparency

Need for explainable, data-driven solution

Project Objectives

- Develop predictive salary models using ensemble machine learning methods
- Compare model performance across Gradient Boosting, XGBoost, and Random Forest
- Apply SHAP to enhance model interpretability and transparency
- Support HR compensation decisions with data-driven, fair recommendations

CRISP-DM Framework Overview

- **Business Understanding:**

Identify compensation inequity and lack of transparency in salary decisions.

- **Data Understanding:**

Analyze 6,899 employee records across demographic, tenure, and performance dimensions.

- **Data Preparation:**

Imputed missing values, encoded categorical variables, engineered nonlinear features.

- **Modeling:**

Built and tuned 3 ensemble models: Gradient Boosting, XGBoost, Random Forest.

- **Evaluation:**

Compared performance using RMSE, R^2, CV R^2; used SHAP for model interpretation.

- **Deployment & Conclusion:**

Insights can inform fair salary reviews and support data-driven HR decision-making.

Data Overview

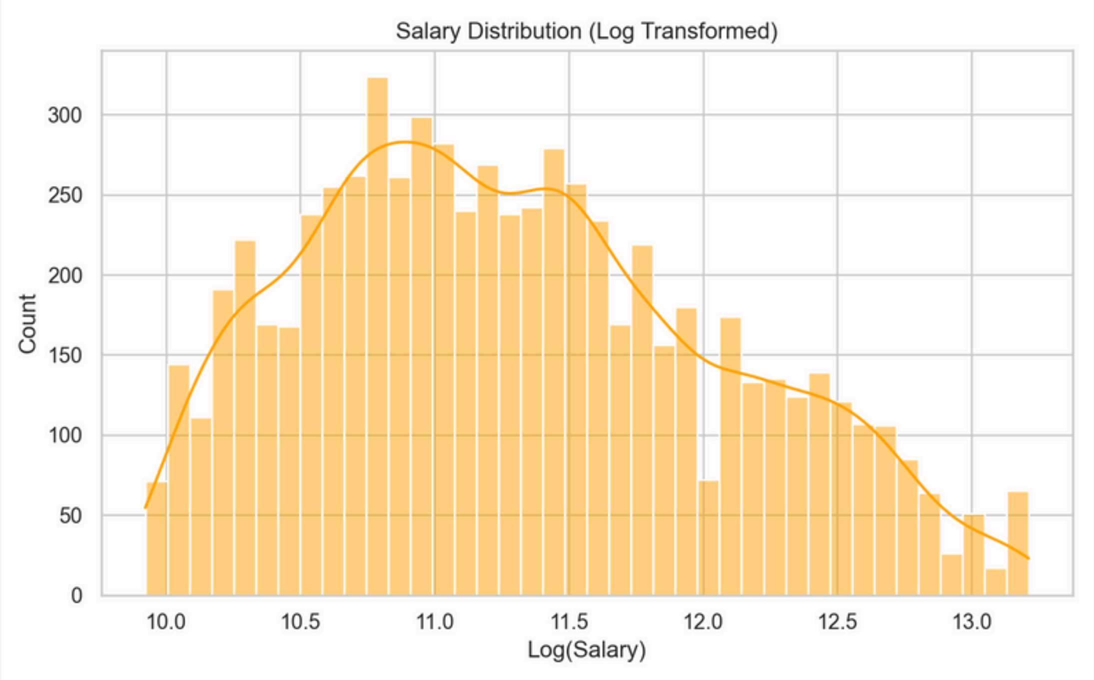
- 6,899 employee records across demographics, tenure, performance, and compensation.
- Key variables: Age, Education, Job Role, Satisfaction Scores, Monthly Income.
- Cleaned and enriched with engineered features (e.g., experience levels, interaction terms).

编码后数据：

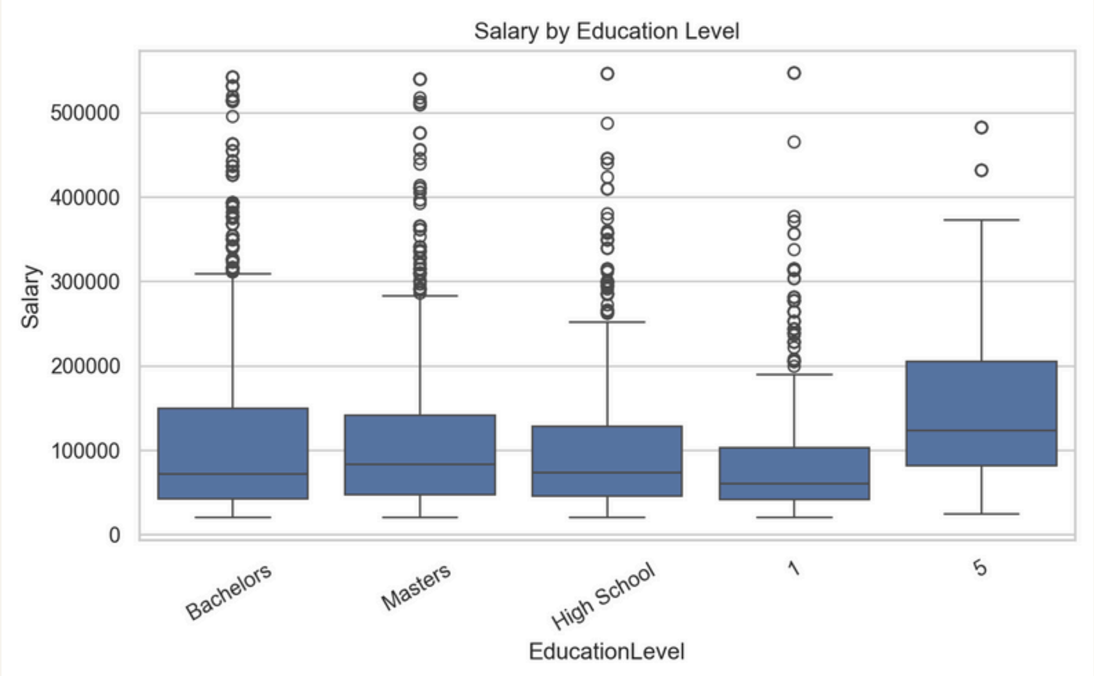
	Salary	Age	StockOptionLevel	YearsAtCompany	EnvironmentSatisfaction	TrainingOpportunitiesWithinYear	TrainingOpportunitiesTaken	SelfRating	ManagerR
0	102059	30	1	10	3.0	3.0	0.0	3.0	
1	102059	30	1	10	4.0	3.0	1.0	3.0	
2	102059	30	1	10	5.0	3.0	0.0	5.0	
3	102059	30	1	10	1.0	3.0	1.0	5.0	
4	102059	30	1	10	3.0	1.0	0.0	4.0	

evel	YearsAtCompany	EnvironmentSatisfaction	TrainingOpportunitiesWithinYear	TrainingOpportunitiesTaken	SelfRating	ManagerRating	EducationLevel_encoded
1	10	3.0	3.0	0.0	3.0	3.0	1
1	10	4.0	3.0	1.0	3.0	2.0	1
1	10	5.0	3.0	0.0	5.0	5.0	1
1	10	1.0	3.0	1.0	5.0	4.0	1
1	10	3.0	1.0	0.0	4.0	3.0	1

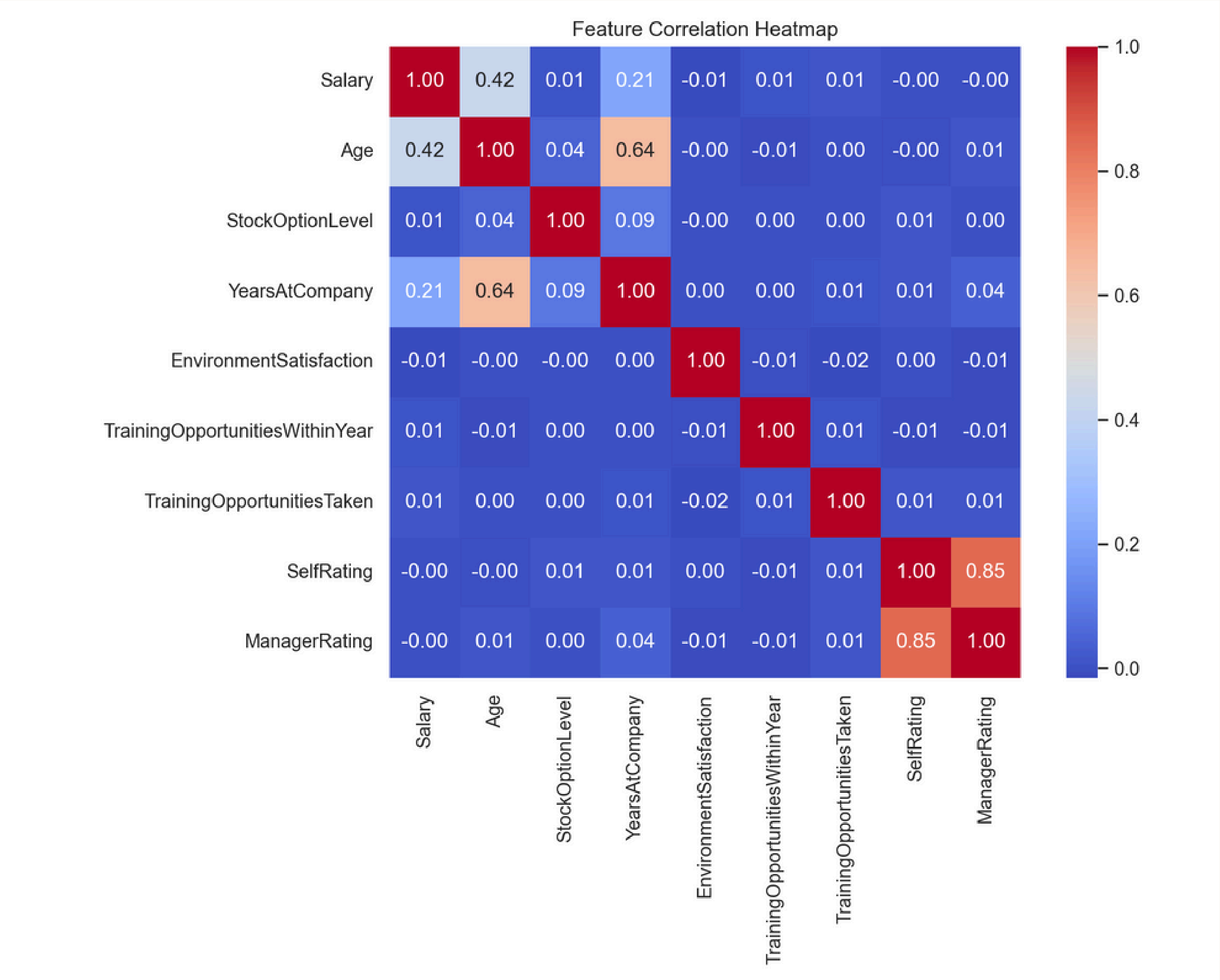
Exploratory Data Analysis – Key Insights



- Salary is highly right-skewed with a long tail of high earners
- Log transformation was applied to normalize the distribution for modeling



- Employees with higher education levels generally earn more
- Outliers exist in all groups, implying education is not the only determinant



- No single variable is sufficient
- Non-linear patterns justify the use of ensemble ML models

Data Cleaning and Feature Engineering

Step	Description
Missing Values	KNN Imputer for numeric features (k=5)
Categorical Encoding	One-hot encoding for nominal; binary and ordinal preserved logically
Feature Engineering	Experience binned into 4 levels (New, Mid, Senior, Expert)
Interaction Features	Combined Job Satisfaction × Environment Satisfaction
Transformation	Log transform on Monthly Income to handle skewness
Feature Selection	SelectKBest + RFE with XGBoost, retained 25 features

Predictive Approach

How We Approached It

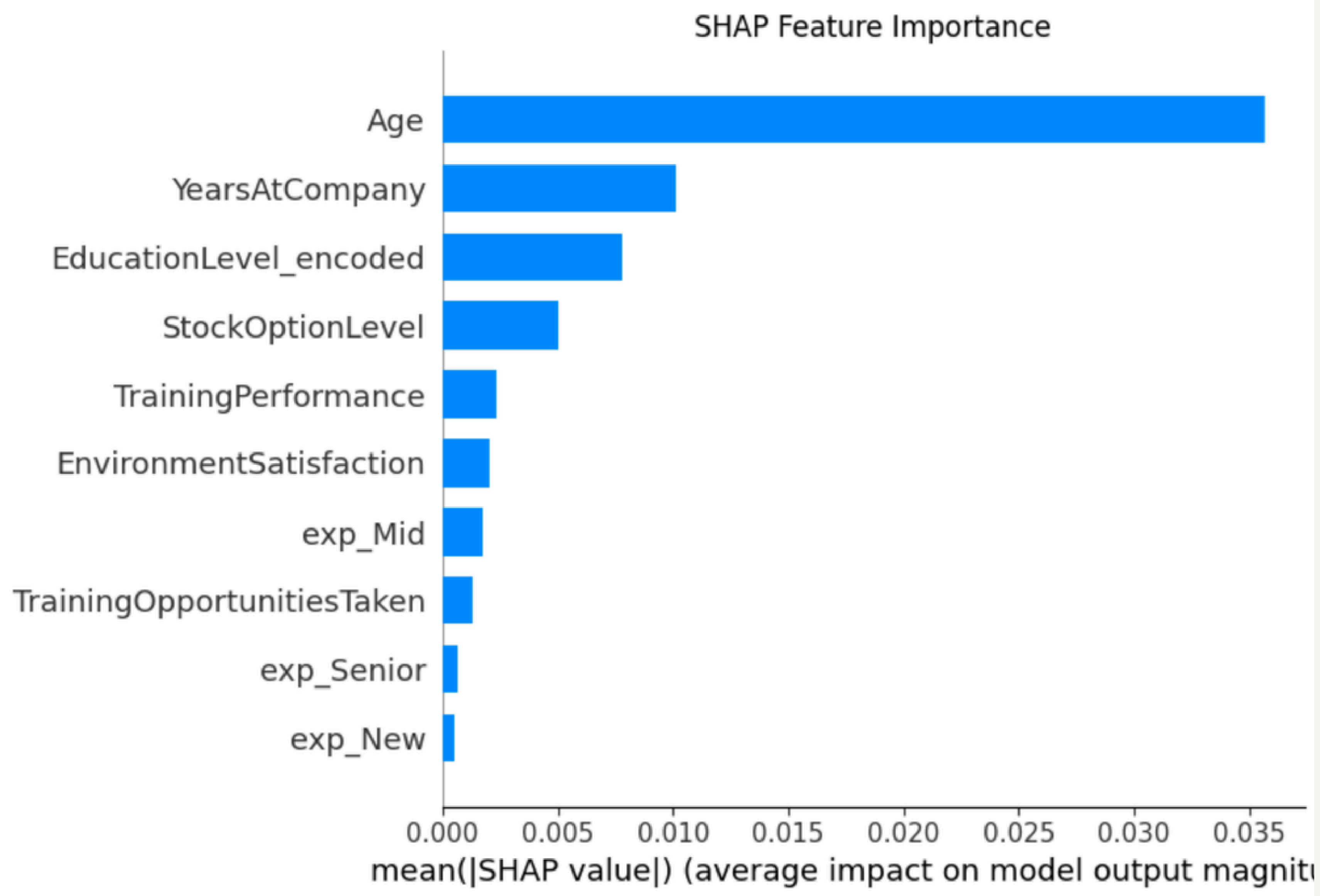
- Applied 3 machine learning models:
    - Gradient Boosting, XGBoost, Random Forest
- Evaluated performance using RMSE & R^2 scores
- Focused on accuracy + interpretability for HR use

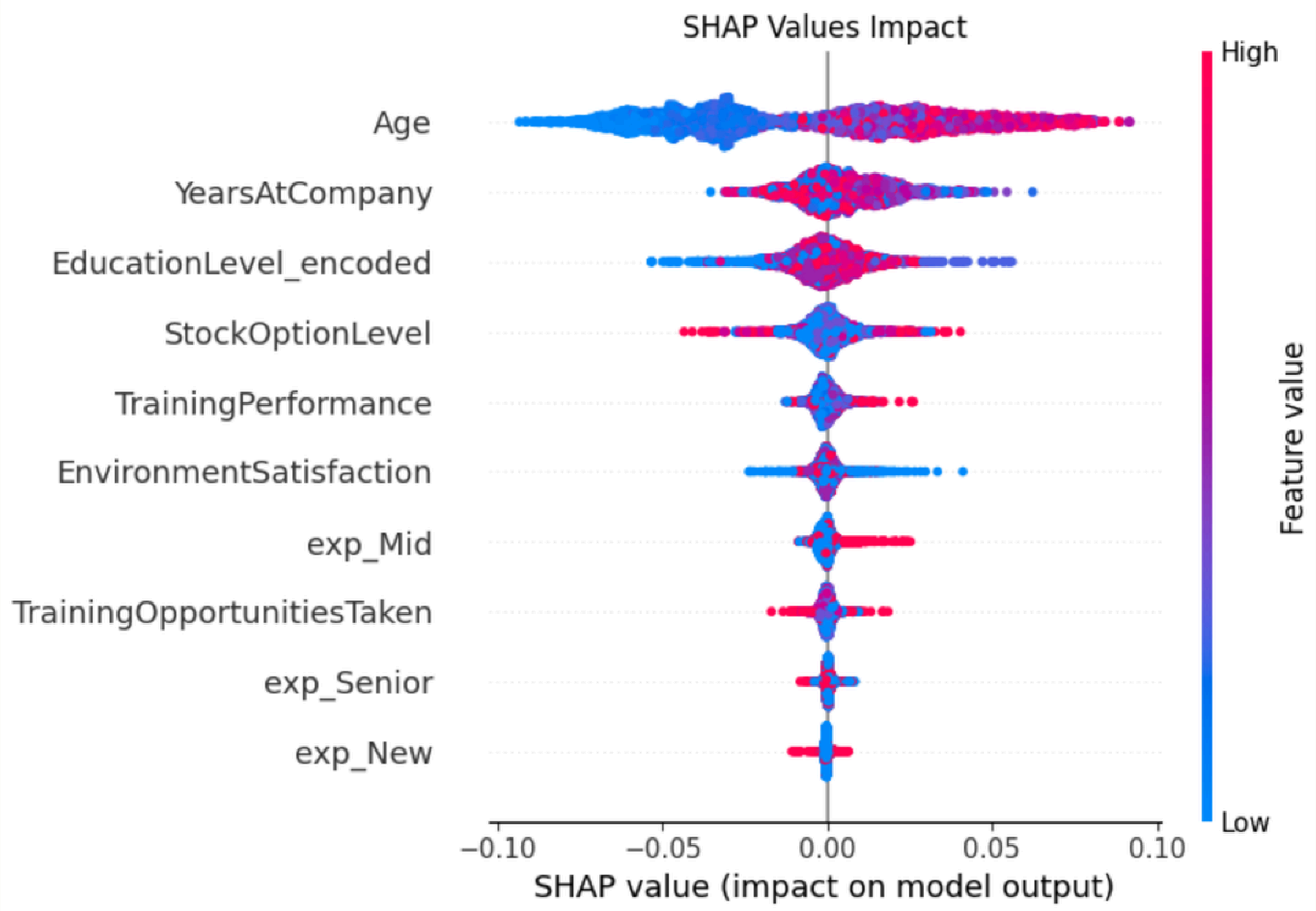
Model Performance Comparison

- Gradient Boosting performed best: lowest error, highest accuracy
- Key drivers of salary: Age, Years at Company, Education Level, Job Satisfaction

Model	RMSE	R^2	CV R^2 mean	Performance Summary
Gradient Boosting	58,315.11	0.6426	0.6329	showed the best balance of accuracy and generalization
XGBoost	60,290.63	0.6180	0.6348	showed consistent cross-validation performance
Random Forest	67,415.94	0.5224	0.5886	had weaker results and higher residuals

SHAP – Model Interpretability





- Age is the most important factor in salary prediction
- Years at Company and Education Level also play strong roles
- Stock Options and Training Performance contribute to salary increases
- High Age, Tenure, and Education → Higher predicted salary
- SHAP improves transparency and fairness in HR-related AI decisions

Business Impact

Promotes pay equity and consistency across departments

Identifies potential underpaid employees for targeted review

Supports budgeting, hiring, and retention strategies with data-driven insight

Ethical Consideration

Fairness

avoid historical bias in salary data

Compliance

ensure GDPR and employee data privacy

Human in the loop

model aids, not replaces, compensation decisions

Conclusion & Recommendation

Conclusion

- Gradient Boosting outperformed other models, achieving the best balance of accuracy and interpretability
- The model provides actionable insights to support fair and consistent compensation decisions
- SHAP analysis enhances transparency, making the model suitable for HR use cases where trust and auditability matter

Next Step

- Integrate the model into HR salary review workflows to assist with data-driven compensation adjustments.
- Expand the dataset by incorporating external economic indicators and performance metrics
- Monitor model fairness and continuously assess for bias and compliance with GDPR and internal policies

Thank You