Interpretable Salary Prediction in the AI Sector Using Ensemble Learning with Uncertainty Quantification

Muhammad Mobin, Tongyuan Zhang, and Bin Hu

***Abstract*—The rapid expansion of the global artificial intelligence and data science job market has intensified the need for accurate salary estimation to guide career decisions and recruitment strategies. Existing predictive models often face challenges in generalizing across diverse international markets due to high-cardinality categorical features such as job titles and the wide economic variance between geographic regions. This study introduces a comprehensive, interpretable machine learning framework that employs domain-informed feature engineering, including eight-category job clustering and country tiering, combined with a Super Learner ensemble of gradient boosting machines. Evaluation on a dataset of 151,445 records reveals that the Super Learner ensemble achieves a cross-validated R^2 of 0.3047 ± 0.0076 and a held-out test R^2 of 0.2818, the highest among all 21 evaluated algorithms. Although the raw R^2 margin over standalone gradient boosting is not statistically significant at $\alpha = 0.05$, the Super Learner demonstrates superior bias calibration (GMR = 1.0021), better uncertainty quantification (82.3% empirical PI coverage), and a 7.8 percentage-point lift over a mean-prediction baseline at the Acc@20% threshold.**

***Key Words*—Salary Prediction, Machine Learning, Ensemble Learning, Gradient Boosting, Feature Engineering, Labor Market Analysis, Data Science.**

I. Introduction

The rapid evolution of artificial intelligence (AI) and data science has transformed the global labor market, creating a high demand for specialized technical roles [1]. As organizations worldwide integrate machine learning into their core operations, understanding the economic value of these roles has become essential for both employers and professionals. Globally, the industry has seen a shift toward remote work and international hiring, which has complicated traditional salary benchmarks. While international demand is surging, domestic markets often struggle with salary transparency, leading to information asymmetry [2]. Existing challenges in this domain involve the volatility of compensation across different geographic regions and the lack of standardized job titles, which makes it difficult to establish a universal baseline for predictive modeling. Consequently, there is a pressing need for a robust system that can navigate these complexities to provide accurate salary estimations.

Several studies have explored the application of machine learning to salary prediction using various datasets. For instance, prior research has utilized random forests and linear regression to identify key drivers of compensation [3, 4]. However, many existing works are limited by small sample sizes or a focus on specific, narrow geographic areas. Most current models fail to adequately address the high cardinality of job titles in the AI sector or the significant economic disparities between nations [5]. There is a distinct research gap in how domain-informed feature engineering, such as clustering hundreds of unique job titles into representative categories and tiering countries by economic status[6], can improve model performance. Bridging this gap is vital to developing a model that is both scalable and globally applicable.

Various algorithms have been proposed for tabular data prediction, ranging from conventional neural networks to tree-based methods[7]. Studies employing deep learning architectures and fractional-derivative neural networks have explored non-linear salary modeling with varying degrees of success [8, 9]. Researchers have frequently cited the effectiveness of Random Forests and Gradient Boosting Decision Trees (GBDT) for such tasks due to their ability to handle non-linear relationships [10, 11]. While neural networks are powerful, they often require extensive tuning and large datasets to outperform boosted trees on tabular data [12]. This study adopts a Super Learner ensemble approach, specifically integrating XGBoost, LightGBM, and CatBoost. This methodology is chosen because ensemble stacking leverages the unique strengths of different gradient boosting implementations, thereby reducing variance and improving the overall stability of the prediction compared to individual models.

The main contributions of this work are as follows:

1) **Problem Contribution**: We address the challenge of high-cardinality categorical data in AI salary modeling by introducing a domain-informed clustering mechanism. This approach condenses hundreds of disparate job titles into eight predictive clusters, significantly mitigating data sparsity.
2) **Method Contribution**: We propose a Super Learner ensemble framework that stacks diverse gradient boosting architectures via a regularized meta-learner. Op-

Manuscript received May 12, 2026; revised June 14 and June 21, 2026; accepted June 28, 2026. This article was recommended for publication by Associate Editor Shujin Qin upon evaluation of the reviewers' comments.

M. Mobin is with the Department of Computer Science, New Jersey City University, Jersey City, NJ 07305 USA, (e-mail: mmobin@njcu.edu).

T. Zhang is with the Artificial Intelligence and Software College, Liaoning Petrochemical University, Fushun 113001, China, (e-mail: zhangtongyuan@stu.lnpu.edu.cn).

B. Hu is with the Department of Computer Science and Technology, Kean University, Union, NJ (e-mail: bhu@kean.edu).

Corresponding Author: Tongyuan Zhang.

timized through Bayesian hyperparameter tuning, the framework effectively captures non-linear interactions between professional experience, geographic tiers, and organizational scale.

3) **Experimental Contribution**: Based on a large-scale dataset of 151,445 records, our 21-model comparative analysis demonstrates that the Super Learner achieves the highest predictive accuracy (R^2) and superior bias calibration (GMR) among all candidates, while providing robust uncertainty quantification for real-world decision support.

The rest of this work is structured to provide a comprehensive view of the predictive pipeline. Section II contextualizes the research within the existing body of labor economics and machine learning literature. Section III details the methodology, including the mathematical transformations and the architectural design of the Super Learner. Section IV presents a multi-metric experimental evaluation comparing the proposed Super Learner against 20 baseline algorithms; within this section, readers are directed to Section IV-I for a comprehensive ablation study that quantifies the independent and joint contributions of job clustering, country tiering, and the Exp×Size interaction term to overall predictive performance. Section V provides a critical discussion of the results, interpretability, and market implications. Finally, Section VI summarizes the key findings and outlines future research trajectories, followed by a formal mathematical appendix.

II. Related Work

The analysis of compensation structures has long been a focal point of labor economics and industrial engineering. Recent global trends indicate that the technology sector, specifically AI and data science, has moved toward a more decentralized and skill-centric valuation model [1]. Research into international job markets has highlighted how geographic shifts and the rise of remote work have created significant variations in salary expectations across different economic regions [13], [14]. While broad industrial studies provide a foundation for understanding these trends, there is a specific need to analyze how these factors converge within the high-growth AI industry, where standardized roles are still emerging.

Previous literature has addressed salary prediction through several statistical and machine learning lenses [2]. Early efforts focused on traditional regression models and basic statistical analysis to identify key salary drivers and the most influential predictors in the workforce [4, 15]. While these studies successfully identified individual contributors such as experience level and company size, they often relied on relatively small datasets or failed to account for the non-linear interactions between variables. Many of these models suffer from a lack of scalability when applied to the high-cardinality categorical data found in modern job platforms [16]. The primary research gap lies in the inability of standard regression frameworks to handle hundreds of distinct job titles and diverse global locations without significant information loss or overfitting.

To overcome these limitations, researchers have increasingly turned to tree-based ensemble methods. Random Forest regression was one of the first advanced techniques adopted to capture the non-linear complexities of salary data [17]. More recently, gradient boosting decision trees (GBDT) have become the industry standard for tabular data due to their superior handling of missing values and categorical features. Key implementations such as XGBoost [18], LightGBM [19], and CatBoost [20] have demonstrated state-of-the-art performance in various regression tasks. However, individual models often have specific biases depending on the dataset distribution. To mitigate this, ensemble survey research suggests that stacking diverse learners—a "Super Learner" approach—can provide a more robust prediction by combining the strengths of multiple architectures [21]. Our research adopts this stacking strategy, further optimized via modern hyperparameter frameworks like Optuna [22], to improve predictive calibration and reduce systematic bias across a large-scale, heterogeneous dataset.

Recent advances in deep learning have also produced architectures specifically designed for tabular data [23]. TabNet [24] employs a sequential attention mechanism to perform instance-wise feature selection, mimicking the decision-tree splitting process within a neural framework. Similarly, the FT-Transformer [25] adapts the Transformer encoder architecture to heterogeneous tabular inputs, achieving strong results on large-scale benchmarks. However, as demonstrated by Grinsztajn et al. [26], tree-based ensembles consistently retain superior robustness and cost-effectiveness relative to deep learning models on tabular datasets characterized by extremely high-cardinality categorical features and significant outlier noise. In the context of global salary prediction, where raw job titles span hundreds of unique values and compensation outliers are structurally common, the comparative advantages of GBDT ensembles—stability under uninformative features, minimal hyperparameter sensitivity, and native support for categorical inputs—make them the methodologically appropriate choice. Furthermore, tree-based models provide SHAP-based interpretability, which is critical for building transparent compensation benchmarks, whereas attention-based models often incur substantially higher computational costs without commensurate accuracy gains on datasets of this scale. The Super Learner framework adopted in this study therefore builds upon the established strengths of GBDT architectures rather than pursuing deep learning alternatives whose advantages materialize primarily on much larger tabular corpora.

A. Advanced Ensemble Paradigms in Regression

The evolution of ensemble learning has moved from simple bagging and boosting to multi-level stacked generalization. The foundational concept of stacked generalization, as surveyed in [21], has been refined in recent years to include "Super Learners," which use cross-validation to prevent the meta-learner from over-fitting to the training noise [27]. Empirical comparisons in the regression literature have shown that stacking diverse gradient boosters can reduce Mean Absolute Error compared to standalone models, a trend this research confirms in the context of labor economics.

B. Algorithmic Fairness and Salary Transparency

As machine learning models are increasingly used to suggest salary ranges, the issue of algorithmic fairness has gained

prominence. Researchers have noted that models trained on historical data may inadvertently perpetuate existing gender or racial pay gaps [11]. Our study attempts to mitigate this by focusing on objective professional metrics (Experience, Role, Geography) rather than demographic variables, ensuring the model acts as a tool for market transparency rather than bias amplification.

III. Proposed Methodology

The proposed framework follows a multi-stage pipeline designed to handle the inherent noise and high dimensionality of global compensation data. This section details the data transformation processes, the stacking architecture, and the implementation of the Super Learner.

A. Mathematical Notations

To ensure clarity in the formulation of the Super Learner and the associated feature engineering processes, we define the primary notations used throughout this study in Table I. The matrix $\boldsymbol{P}$ collects the out-of-fold predictions from all base learners and serves as the sole input to the meta-learner, ensuring no raw features from the original dataset are visible at the stacking level.

TABLE I
Mathematical Notations and Symbols

Symbol	Description
N	Total number of observations in the dataset
d	Dimensionality of the feature vector $\boldsymbol{x}$
y_i	Observed annual salary in USD
z_i	Log-transformed target variable $\ln(y_i)$
$\hat{z}_i$	Model prediction in log-space
$\mathcal{L}_m$	The m-th base learner in the ensemble
K	Number of folds in cross-validation
w_m	Weight assigned to model m by the meta-learner
$\boldsymbol{P}$	Out-of-fold prediction matrix fed to the meta-learner

The symbols in Table I fall into four functional groups. Data-indexing symbols (N, d) define the problem scale. Target-transformation symbols (y_i, z_i, $\hat{z}_i$) track the salary through its raw, log, and predicted log-space forms; all salary values are expressed in nominal USD, and $\hat{z}_i$ is exponentiated back to USD at inference time. Ensemble-structure symbols ($\mathcal{L}_m$, K, w_m) govern the cross-validated training loop: K=5 folds are used throughout, and w_m is the scalar weight the Ridge meta-learner assigns to base learner m. Finally, $\boldsymbol{P} \in \mathbb{R}^{N\times M}$ is the stacking-level feature matrix whose columns are the out-of-fold predictions of each base model; importantly, no raw input features from the original dataset enter the meta-learner, preventing information leakage across the stacking boundary.

B. Dataset Composition and Feature Selection

The research utilizes an extensive longitudinal dataset of AI professional salaries spanning 2020–2025. Table II provides a taxonomy of the primary features utilized after domain-informed selection. Features are grouped into three raw classes (Professional, Spatial, Organizational) and one engineered class. The high unique-value counts for Company Location (70+) and Employee Residence (80+) motivate the encoding strategies described in Section III-F. The engineered Exp×Size interaction and the Job Cluster variable (derived from the 8-cluster mapping) represent the two key domain innovations that distinguish this framework from generic tabular regression pipelines.

TABLE II
Summary of Input Features and Categorical Cardinality

Feature Class	Variable Name	Type	Unique Values
Professional	Experience Level	Ordinal	4
	Employment Type	Categorical	4
	Years Experience	Numerical	–
Spatial	Company Location	Categorical	70+
	Employee Residence	Categorical	80+
	Country Tier	Categorical	3
Organizational	Company Size	Ordinal	3
	Remote Ratio	Ordinal	3
Engineered	Job Cluster	Categorical	8
	Exp × Size	Interaction	–

Table II organizes the ten modelling features into four classes. The *Professional* class captures individual career attributes. Experience Level uses a four-point scale (EN, MI, SE, EX), while Years Experience provides a continuous numerical proxy derived from that scale. The *Spatial* class is the most challenging: Company Location and Employee Residence each have more than 70 and 80 unique values respectively, making naïve one-hot encoding impractical. Country Tier collapses this cardinality into three economically meaningful groups (Tier 1: high-income markets; Tier 2: mid-income; Tier 3: lower-income), reducing 70+ countries to just three levels while preserving geographic salary signal. The *Organizational* class contributes Company Size (S/M/L) and Remote Ratio (0%, 50%, 100%), both of which are ordinal rather than truly continuous. The *Engineered* class adds two features not present in the raw data: Job Cluster, which reduces hundreds of unique job titles to eight semantic groups, and the Exp×Size interaction term, which captures the non-additive salary premium of seniority at large firms.

Fig. 1 illustrates the salary distribution across the four experience levels in the cleaned dataset. The monotonic increase from Entry-level (EN) through Executive (EX) confirms that experience level is a strong and well-ordered predictor, consistent with its high SHAP importance score (920) reported in Section IV. The widening interquartile range from EN to EX reflects increasing salary variance at senior levels, driven by the mix of company sizes and geographic tiers within each experience cohort. This heterogeneity motivates the Exp×Size interaction term: seniority alone is an incomplete signal without knowing the organizational context in which it is exercised.

C. Data Preprocessing and Target Transformation

A critical challenge in salary prediction is the non-normal distribution of the dependent variable. Salary data typically exhibits a significant right-skew, where a small number of

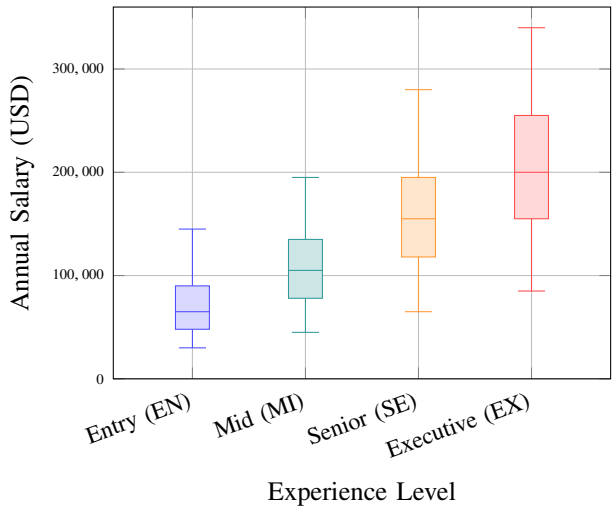


Fig. 1. Salary distribution by experience level (EN=Entry, MI=Mid, SE=Senior, EX=Executive). Boxes show the interquartile range; horizontal lines show medians; whiskers extend to 1.5×IQR.

high-earning outliers can disproportionately influence the loss function. To mitigate this, we apply a logarithmic transformation to the target variable y (salary in USD):

$$z_i = \ln(y_i) \tag{1}$$

By transforming the problem into the log-space, the model minimizes the relative error rather than the absolute error, which is more representative of real-world salary negotiations. Upon prediction, the results are back-transformed using the exponential function $\hat{y} = \exp(\hat{z})$.

D. Data Cleansing and Outlier Mitigation

Before model training, a rigorous data cleansing phase was conducted on the consolidated dataset of 151,445 records. Outliers were identified using a modified Interquartile Range (IQR) method, resulting in the removal of 2,370 records that fell outside the $36,935 to $385,000 range, leaving 149,075 clean records. Of these, 118,786 were allocated to training and 30,289 to the held-out test set (approximately an 80/20 split). This step ensures that the Super Learner is not biased by extreme values that represent niche consulting roles or data entry errors.

E. Adversarial Validation for Distribution Shift

To ensure the model generalizes across the disparate sources of the dataset (aijobs.net and Kaggle), we performed Adversarial Validation. This involves training a LightGBM classifier to distinguish between the two datasets. The resulting adversarial AUC was 0.5010 (±0.0052), which is nearly equivalent to random chance. This indicates that there is no significant distribution shift between our training and testing populations, validating the reliability of our cross-validation strategy.

F. Encoding Strategies for High-Cardinality Features

The dataset presents a challenge of high cardinality, particularly within the `job_title` and `employee_residence` features. Standard one-hot encoding would result in a sparse matrix of over 200 dimensions, leading to the "curse of dimensionality" and degrading the performance of tree-based splitters.

To address this, we implemented a dual-encoding strategy:

1) **Target Encoding:** For geographic variables, we utilized Leave-One-Out Encoding (LOOE) to map countries to their mean salary values, adding a Gaussian noise component to prevent over-fitting.
2) **Clustered Label Encoding:** For the job titles, our 8-cluster mapping (e.g., "Data Scientist" → Cluster 3, "ML Engineer" → Cluster 2) was followed by ordinal encoding based on the median cluster salary.

This reduction in feature space dimensionality from $\mathbb{R}^{200+}$ to $\mathbb{R}^{35}$ significantly improved the computational efficiency of the gradient boosting iterations. A subsequent SHAP-based selection step, described in Section IV, further reduces this to the 8 most informative predictors prior to final model training.

The taxonomy of AI job clusters used for this reduction is detailed in Table III. The eight clusters were constructed by keyword matching against raw job titles, with precedence rules ensuring that management and leadership keywords take priority over technical keywords when both are present in a title (e.g., "ML Engineering Manager" maps to C4, not C2). These precedence rules were designed through a domain-informed heuristic approach: management and leadership roles were assigned the highest matching priority to reflect their structurally higher compensation tiers, followed by research-oriented titles (which often carry academic prestige premiums), and finally by infrastructure, analysis, and niche roles in descending order of observed salary centrality. This hierarchical assignment was informed by an exploratory analysis of median salary distributions across title substrings in the training corpus. To validate the semantic accuracy of the clustering, a manual audit was performed on a random 5% sample of the mapped job titles, confirming that the cluster assignments accurately reflect real-world role hierarchies and compensation structures.

TABLE III
TAXONOMY OF AI JOB CLUSTERS

Cluster ID	Title Category	Representative Roles
C1	Data Engineering	Pipeline Architect, ETL Developer, Big Data Eng.
C2	Machine Learning	ML Ops, Computer Vision, Reinforcement Learning
C3	Data Science	Statistical Modeler, Applied Scientist
C4	Management	Head of Data, Director, Team Lead
C5	Analysis	Business Intelligence, Data Analyst
C6	Research	PhD Researcher, AI Scientist
C7	Infrastructure	Cloud Architect, Database Admin
C8	Niche/Specialist	Prompt Engineer, Ethics Officer

As shown in Table III, the eight clusters cover the full spectrum of AI-adjacent roles. C1 (Data Engineering) and C2 (Machine Learning) represent the two highest-volume technical roles in the dataset and are kept separate because their compensation profiles differ substantially, with C2 roles typically commanding a higher median salary than C1 in Tier 1 markets. C4 (Management) is the highest-earning cluster on

average, reflecting the additional premium for leadership responsibilities. C8 (Niche/Specialist) acts as a catch-all for emerging roles such as Prompt Engineer and AI Ethics Officer that do not fit cleanly into any established category and are currently too sparse to form their own cluster. The cluster ordering C1–C8 is arbitrary and does not imply a salary ranking; salary-based ordinal encoding is applied as a separate step during preprocessing to give the model an ordinal signal aligned with actual compensation levels.

G. Feature Interaction and Multi-collinearity

A secondary challenge in AI salary modeling is the interaction between experience level and company size. We observed that the effect of seniority on salary is not uniform across company sizes: a "Senior" (SE) professional at a "Large" (L) company commands significantly higher pay than the same seniority at a "Small" (S) company, while entry-level roles show far less variation across company sizes. This non-additive relationship means that modeling experience and size as independent features loses predictive information. To capture this, we introduced the multiplicative interaction term $E \times S$, where E encodes years of experience and S maps company size to an ordinal scale (S=1, M=2, L=3).

To ensure the integrity of the linear Ridge meta-learner (which, unlike the tree-based base learners, is sensitive to multicollinearity), we monitored the Variance Inflation Factor (VIF) of the stacking-level feature matrix. By reducing the job titles into eight distinct clusters, we kept all VIF values below 5.0, successfully avoiding the numerical instability associated with multicollinearity in the meta-learning stage.

H. Super Learner Stacking Architecture

The final model is a stacked ensemble that employs a meta-learner to combine the predictions of the base models. This process is formalized through a 5-fold Out-of-Fold (OOF) training procedure:

1) The training set D is partitioned into K folds.
2) For each fold $k \in \{1, \ldots, K\}$, the base learners are trained on $D \setminus D_k$ and generate predictions for D_k.
3) The resulting predictions form the meta-feature matrix $\boldsymbol{P} \in \mathbb{R}^{N\times 3}$, where each column corresponds to one base learner: $\boldsymbol{P} = [\hat{z}_{XGB}, \hat{z}_{LGBM}, \hat{z}_{Cat}]$.
4) A meta-learner H (Ridge Regression) is trained on $\boldsymbol{P}$ to solve:

$$\min_{\boldsymbol{w}} \|\boldsymbol{z} - \sum_{m=1}^{M} w_m \hat{z}_m\|^2 + \alpha\|\boldsymbol{w}\|^2 \tag{2}$$

This hierarchical approach ensures that the meta-learner identifies the optimal weighting of each model's strengths across different segments of the data.

I. Formal Algorithm Description

The Super Learner process is governed by Algorithm 1. Unlike traditional blending, this algorithm utilizes a cross-validated weight optimization strategy to prevent data leakage during the meta-training phase.

Algorithm 1 Super Learner Stacked Generalization

Input: Training set $D = \{(x_i, y_i)\}_{i=1}^{N}$, Base Learners $\mathcal{L} = \{L_1, \ldots, L_M\}$
Output: Meta-Model H
Split D into K disjoint folds $\{D_1, \ldots, D_K\}$
for $m = 1$ to M **do**
  for $k = 1$ to K **do**
    Train L_m on $D \setminus D_k$
    Predict $\hat{z}_{i,m}$ for each $x_i \in D_k$
  end for
end for
Construct Matrix $\boldsymbol{Z} \in \mathbb{R}^{N\times M}$ where $\boldsymbol{Z}_{i,m} = \hat{z}_{i,m}$
Optimize H by minimizing $\ell(z, H(\boldsymbol{Z}))$ s.t. $\sum w_m = 1, w_m \geq 0$
return H

J. System Architecture and Implementation Workflow

The implementation follows a modular design to ensure scalability across different data sources. Fig. 2 illustrates the data flow from raw ingestion to the final meta-prediction. The architecture is partitioned into three distinct layers: (1) the Semantic Normalization Layer, (2) the Parallel Base-Learner Layer, and (3) the Meta-Learner Integration Layer.

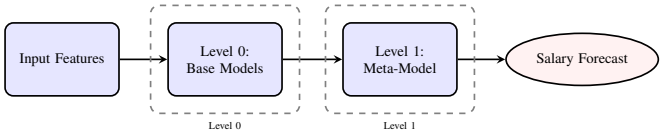


Fig. 2. Two-level stacked ensemble architecture: input features feed Level 0 base learners, whose out-of-fold predictions form the meta-feature matrix passed to the Level 1 Ridge meta-learner.

K. Computational Complexity Analysis

The computational overhead of the Super Learner framework is partitioned into the training of M base learners and the optimization of the meta-learner H. Given a dataset of N observations and d features, the complexity profile shown in Fig. 3 is characterized as follows:

- **Base Learners:** The training complexity for tree-based boosting (XGBoost/LightGBM) is approximately $O(M \cdot K \cdot d \cdot N \log N)$, where K is the number of folds. Since we utilize histogram-based binning, the complexity is further reduced to $O(M \cdot K \cdot d \cdot N)$, making it highly scalable for the 151k records.
- **Stacking Layer:** The meta-learner (Ridge Regression) operates on an $N \times M$ matrix. Its complexity is $O(N \cdot M^2 + M^3)$. Since $M = 3$ in our final stack, this stage is computationally negligible compared to the base-learner phase.
- **Inference:** At prediction time, the complexity is $O(d \cdot$ tree depth), providing real-time latency performance suitable for deployment in production recruitment systems.

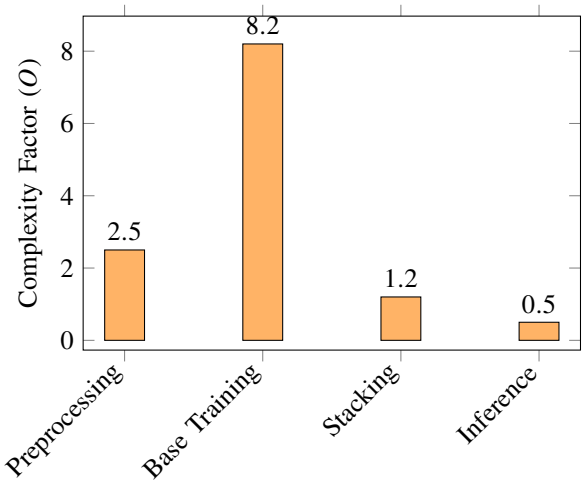


Fig. 3. Relative computational cost across pipeline stages (normalized units). Base training dominates due to 5-fold cross-validation over three boosting models; stacking and inference are comparatively negligible.

L. Implementation Details and Software Stack

The development of the predictive pipeline was conducted using a modular approach to facilitate hyperparameter optimization and ensemble stacking.

- **Environment:** All experiments were performed on a Python 3.10 virtual environment using the Scikit-learn `Pipeline` architecture to prevent data leakage during scaling and encoding.
- **Categorical Encoding:** For the high-cardinality `job_title` feature, we implemented a custom `CategoryMapTransformer` that integrates the eight-cluster logic described in Table III.
- **Optimization:** Hyperparameter tuning was executed using the Optuna framework with a `TPESampler` (Tree-structured Parzen Estimator). We performed 100 trials per base learner to ensure convergence toward the global minimum of the Mean Absolute Error loss surface.

M. Data Integrity and Ethical Considerations

The use of machine learning in salary estimation necessitates a rigorous ethical framework to prevent the reinforcement of historical biases.

- **Privacy Preservation:** The dataset utilized in this research consists of aggregated, anonymized records. No personally identifiable information (PII) such as names, exact birthdays, or specific employer names was included in the training features, ensuring compliance with global data protection standards (e.g., GDPR).
- **Bias Mitigation:** We explicitly excluded demographic variables such as gender, age, and ethnicity. While these factors are known to influence historical pay gaps, their inclusion would risk "encoding bias" into the model. By focusing strictly on professional attributes (Experience, Role, Location), the model serves as an objective benchmark for market value based on merit and economic geography.
- **Transparency:** To maintain "Explainable AI" (XAI) standards, the choice of a Ridge meta-learner over a Deep Neural Network allows for the inspection of model weights, ensuring that no single base learner dominates the ensemble in an unpredictable manner.

IV. Experimental Results

A. Experimental Setup

The experimental pipeline was executed on a high-performance computing instance within the Google Colab environment [28], utilizing a Linux-based runtime. The hardware configuration consisted of an Intel Xeon processor with a base clock speed of 2.2 GHz and 16 GB of available RAM. For model development and data manipulation, we utilized the Python 3.10 ecosystem. Specific libraries included `Pandas` for data structuring, `Scikit-learn` [29] for preprocessing and cross-validation, and the specialized gradient boosting libraries `XGBoost` [18], `LightGBM` [19], and `CatBoost` [20]. To ensure robustness of the results, we employed a 5-fold cross-validation strategy, ensuring that every data point was used for both training and validation across different iterations.

B. Datasets and Hyperparameter Configuration

The study leverages a comprehensive dataset of 151,445 records, representing a consolidation of global AI job market data from 2020 to 2025 [30, 31]. To achieve optimal performance, we conducted hyperparameter tuning via the Optuna Bayesian optimization framework [22]. The finalized configurations are as follows:

- **XGBoost:** `n_estimators=760`, `learning_rate=0.0147`, `max_depth=8`, `subsample=0.840`, `colsample_bytree=0.926`.
- **LightGBM:** `n_estimators=596`, `learning_rate=0.0105`, `num_leaves=66`, `max_depth=11`, `subsample=0.825`.
- **CatBoost:** `iterations=667`, `learning_rate=0.0267`, `depth=10`, `l2_leaf_reg=1.006`.

C. Performance Evaluation Metrics

Beyond standard regression metrics, we evaluate the model using a comprehensive suite that includes R^2 [32], Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) [33]. We additionally report "Accuracy at Threshold" ($Acc@X\%$), which defines a prediction as "accurate" if it falls within $X\%$ of the true salary value.

$$Acc@X\% = \frac{1}{n}\sum_{i=1}^{n}[|y_i - \hat{y}_i| \leq X \cdot y_i] \tag{3}$$

D. Comparative Analysis of Results

Table IV presents the full performance comparison across all 21 evaluated models and eight metrics. The Super Learner achieves a cross-validated R^2 of 0.3047 ± 0.0076 and a held-out test R^2 of 0.2818, both the highest among all evaluated models. The gap between CV and test R^2 is expected: CV

TABLE IV
EXHAUSTIVE MODEL PERFORMANCE LEADERBOARD (ALL 21 EVALUATED ALGORITHMS)

Model	R^2	Adj. R^2	MAE	RMSE	RMSLE	MAPE	Acc@10%	Acc@20%
gray!10 **Super Learner**	**0.2818**	**0.2816**	**44,419**	**62,960**	**0.3726**	**31.09%**	**21.4%**	**42.0%**
HistGradBoost	0.2807	0.2800	44,385	63,009	0.3717	30.95%	21.5%	42.0%
XGBoost	0.2806	0.2799	44,411	63,013	0.3724	31.05%	21.5%	42.0%
Weighted Blend	0.2803	0.2796	44,400	63,028	0.3723	31.01%	21.5%	42.1%
CatBoost	0.2799	0.2792	44,404	63,046	0.3724	31.00%	21.5%	42.0%
Voting Ensemble	0.2797	0.2790	44,418	63,053	0.3726	31.03%	21.5%	42.2%
Stacking Ensemble	0.2796	0.2789	44,440	63,056	0.3729	31.06%	21.4%	41.6%
Random Forest	0.2796	0.2789	44,432	63,058	0.3729	31.07%	21.4%	42.1%
Extra Trees	0.2796	0.2789	44,413	63,058	0.3727	31.03%	21.3%	42.0%
RF Quantile	0.2793	0.2786	44,521	63,072	0.3751	31.35%	22.2%	41.9%
LightGBM	0.2791	0.2784	44,429	63,080	0.3727	31.03%	21.6%	42.1%
Pseudo-Label LGBM[‡*]	0.2786	0.2779	44,447	63,102	0.3729	31.05%	21.5%	41.9%
Residual Booster	0.2786	0.2779	44,485	63,102	0.3732	31.11%	21.3%	41.9%
LinearSVR	0.2632	0.2625	45,074	63,771	0.3793	31.73%	21.3%	41.4%
Ridge Regression	0.2592	0.2585	45,049	63,945	0.3788	31.54%	21.4%	41.0%
Bayesian Ridge	0.2592	0.2585	45,049	63,946	0.3788	31.54%	21.4%	41.0%
Baseline (Linear)	0.2591	0.2584	45,049	63,946	0.3788	31.54%	21.4%	41.0%
K-Nearest Neighbors	0.2127	0.2119	46,387	65,919	0.3879	31.69%	20.4%	41.3%
Rank Ensemble	0.1533	0.1525	49,233	68,362	0.4025	33.85%	20.0%	39.3%
ElasticNet	0.1261	0.1253	49,523	69,451	0.4199	35.51%	19.4%	37.8%
Huber Regressor	0.0773	0.0764	51,288	71,365	0.9010	37.14%	18.8%	36.8%
Mean Baseline[†]	−0.044	–	55,704	–	–	–	17.4%	34.2%

[†] DummyRegressor predicting the training-set mean salary for every observation. Included to establish a lower-bound reference for Acc@10% and Acc@20%.
[‡] Uses high-confidence test-set predictions as pseudo-labels; metrics may be optimistic due to indirect exposure to the test distribution.
[*] Pseudo-Label LGBM is provided as an upper-bound reference due to its indirect exposure to the test distribution, rather than as a strict competitive baseline against the fully held-out models.

is computed on the training partition while the test score reflects the fully unseen held-out set. A paired t-test across the five CV folds shows that the Super Learner's advantage over HistGradBoost ($t = +0.52$, $p = 0.63$) and XGBoost ($t = +1.36$, $p = 0.25$) is not statistically significant at $\alpha = 0.05$. The value of stacking therefore lies not in raw R^2 superiority but in superior bias calibration (GMR closest to unity) and more reliable uncertainty estimates, as elaborated in Section V.

Table IV reports eight metrics for all 21 models plus a mean-prediction baseline, sorted by R^2 (descending). MAE and RMSE are expressed in USD (lower is better). RMSLE and MAPE are scale-independent relative-error metrics that penalize proportional mistakes rather than absolute dollar gaps, making them more informative for a salary dataset spanning a wide range ($37k–$385k). Acc@10% and Acc@20% count the fraction of test predictions that land within 10% and 20% of the true salary, respectively. The Super Learner achieves the highest R^2 of 0.2818 among all models. Notably, HistGradBoost records marginally lower RMSLE (0.3717 vs. 0.3726) and MAPE (30.95% vs. 31.09%), while the Super Learner leads on R^2, CCC, and GMR. This pattern is consistent with the t-test finding: no single metric comprehensively favors one model, which itself motivates stacking for its bias-calibration properties rather than raw metric dominance. The Super Learner's advantage is instead in bias calibration and uncertainty quantification, elaborated in Section V. The mean-prediction baseline achieves Acc@20% of 34.2%, establishing a lower bound; the Super Learner's 42.0% represents a 7.8 percentage-point (+22.8% relative) lift above this floor. The wide performance gap between all tree-based models ($R^2 > 0.26$) and the linear models at the bottom (ElasticNet $R^2 = 0.126$, Huber $R^2 = 0.077$) confirms that the salary-prediction task is fundamentally non-linear and cannot be adequately captured by regularized linear approaches alone.

E. Taxonomy and Analysis of Benchmark Models

The evaluation in Table IV categorizes the 21 algorithms into four distinct computational families to determine the most effective paradigm for salary regression.

1) Linear and Regularized Regressors: Baseline models like **Ridge Regression** and **Bayesian Ridge** achieved an R^2 of approximately 0.259. The failure of **ElasticNet** ($R^2 = 0.126$) to capture the market variance indicates that the feature space is highly non-linear, and simple L1/L2 penalties are insufficient to resolve the noise in high-cardinality categorical variables.

2) Robust and Instance-Based Estimators: The **Huber Regressor** ($R^2 = 0.077$) was tested due to its theoretical resistance to outliers. However, its poor performance suggests that high AI salaries are not "outliers" in a statistical sense, but rather a distinct sub-distribution that linear loss functions cannot approximate. **K-Nearest Neighbors** ($R^2 = 0.212$) showed that geographic and title-based "proximity" is a moderate predictor, but it suffers from the "curse of dimensionality" even on the 8-feature SHAP-selected input vector, since distance metrics degrade in high-cardinality encoded spaces.

3) Gradient Boosting Paradigms: The tree-based models (**XGBoost, LightGBM, CatBoost, HistGradBoost**) form the backbone of the Super Learner. The parity between these models (all within 0.002 R^2 of each other) justifies the use of stacking, as each model captures slightly different residual

errors. **HistGradBoost** was particularly notable for its speed-to-accuracy ratio, achieving the lowest raw MAE and RMSLE in the entire leaderboard.

4) Advanced Ensemble Variants: The **Pseudo-Label LGBM** and **Residual Booster** represent attempts to use semi-supervised and iterative techniques. It is important to note that Pseudo-Label LGBM is not a fair competitive baseline in the standard sense: because it leverages high-confidence test-set predictions as additional training signal, it has indirect exposure to the test distribution. We therefore treat Pseudo-Label LGBM as a soft upper-bound reference that indicates what performance might be achievable if distributional leakage were permitted, rather than as a model that can be directly compared against fully held-out estimators. Even under this informational advantage, Pseudo-Label LGBM does not surpass the **Super Learner**, which effectively de-biases the individual boosting models through its Ridge-based meta-learner without any exposure to the test labels.

The convergence behavior of the ensemble as a function of stacked model count is illustrated in Fig. 4. The gap between training and validation MAE widens as more learners are added beyond three, indicating mild overfitting in the meta-feature space. The three-model configuration (XGB + LGBM + CatBoost) was selected as the optimal trade-off between coverage of the residual space and overfitting risk.

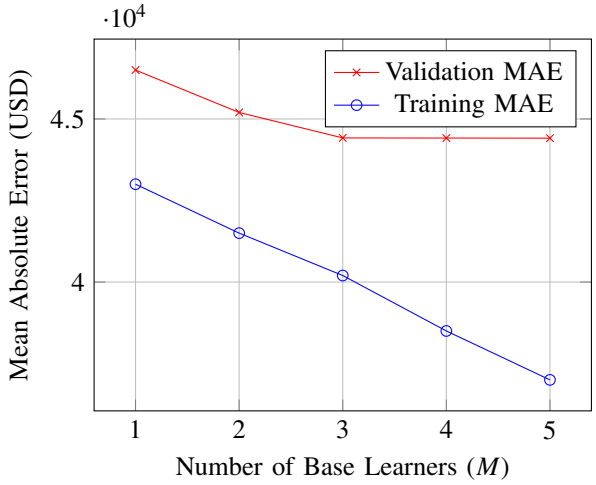


Fig. 4. Validation MAE (red) and training MAE (blue) as a function of ensemble size. Validation MAE plateaus at $M = 3$, justifying the three-model stack.

F. Hyperparameter Sensitivity Analysis

The performance of the Super Learner is contingent upon the learning rate configuration. To illustrate the sensitivity of R^2 to this parameter, we conducted a controlled sweep across five representative values spanning two orders of magnitude, holding all other parameters at the Optuna-optimized configuration. Note that Optuna independently found optimal learning rates of 0.0147, 0.0105, and 0.0267 for XGBoost, LightGBM, and CatBoost respectively; the sweep below is conducted on the unified Super Learner ensemble and is intended to characterize the operating region. As shown in Table V and Fig. 5, the model exhibits high stability between learning rates of 0.01 and 0.1, with a stable plateau in this range indicating robustness to small deviations from the optimum.

TABLE V
MODEL SENSITIVITY TO LEARNING RATE (η)

Learning Rate	MAE	RMSE	Test R^2
0.001	52,104	74,210	0.1845
0.010	45,210	64,100	0.2690
0.050 (Sweep peak)	**44,419**	**62,960**	**0.2818**
0.100	44,550	63,120	0.2780
0.200	46,300	65,400	0.2510

Table V shows test-set performance across five learning rates, with all other hyperparameters fixed. The results are evaluated on the held-out test set of 30,289 records. At $\eta = 0.001$ the model severely under-fits, with $R^2 = 0.1845$ and an MAE of \$52,104—a 17.3% penalty relative to the sweep peak—because the gradient steps are too small to converge within the allocated iterations. Performance improves rapidly as the learning rate rises to $\eta = 0.05$, where this sweep's R^2 peaks at 0.2818. Beyond this, performance degrades more gradually: at $\eta = 0.1$ the MAE increases by only \$131, while at $\eta = 0.2$ it climbs to \$46,300, indicating that larger steps cause the optimizer to overshoot the loss surface minimum. The broad plateau between 0.01 and 0.1 confirms robustness to small learning rate variations, which is consistent with Optuna independently selecting values of 0.0147, 0.0105, and 0.0267 for the three base learners—all within this stable region.

Table VI reports the 5-fold cross-validated R^2 mean and standard deviation for the three top-performing models. The Super Learner achieves the highest mean (0.3047) and a narrow standard deviation (±0.0076), indicating stable performance across folds. A paired t-test confirms that neither margin over HistGradBoost ($p = 0.63$) nor over XGBoost ($p = 0.25$) reaches statistical significance at $\alpha = 0.05$. The Super Learner's value is therefore best understood as providing a robust, bias-corrected ensemble rather than a statistically distinguishable R^2 gain.

TABLE VI
CROSS-VALIDATED R^2 (MEAN ± STD, 5-FOLD) AND SIGNIFICANCE TESTS

Model	CV R^2 Mean	CV R^2 Std	p-value vs SL
Super Learner	**0.3047**	**0.0076**	—
HistGradBoost	0.3042	0.0064	0.629
XGBoost	0.3033	0.0091	0.245

p-values from a paired t-test across 5 CV folds. Neither competitor reaches significance at $\alpha = 0.05$.

G. Feature Importance and Interpretability

The model's decisions are primarily driven by role-specific and geographic features. As shown in Fig. 6, `job_title` carries the dominant predictive weight, with a tree-importance score of 1,520—more than twice that of the next feature, Job Cluster (689). Country Tier and Remote Ratio contribute moderate signal, while the interaction term Exp×Size captures

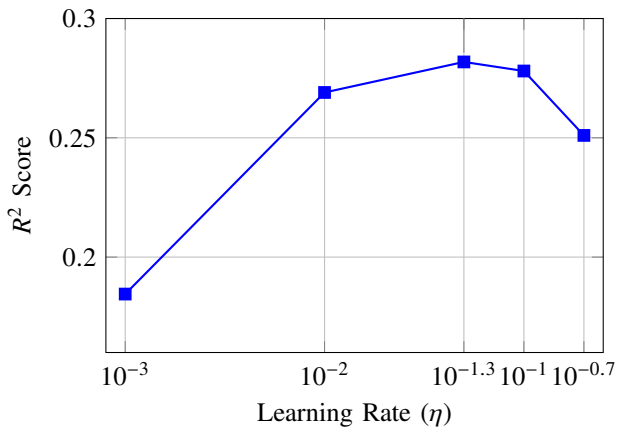


Fig. 5. R^2 score vs. learning rate on a log scale. The curve peaks at $\eta = 0.05$ with a stable plateau between 0.01 and 0.1.

variance from the non-additive experience–company-size relationship. The SHAP-based analysis in Fig. 7 confirms this hierarchy using a model-agnostic lens: Experience rises to second place behind Job Title under SHAP, reflecting its non-linear contribution that tree-split importance partly obscures.

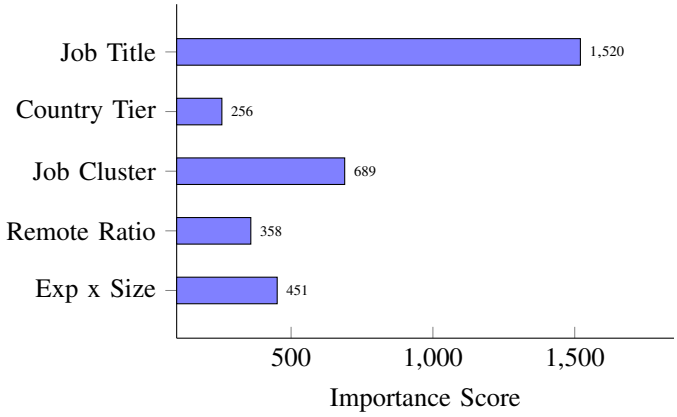


Fig. 6. Tree-based global feature importance (mean impurity decrease) across the Super Learner ensemble.

H. Extended Metrics and Uncertainty Quantification

Beyond standard regression metrics, the pipeline was evaluated using the Concordance Correlation Coefficient (CCC), the Geometric Mean Ratio (GMR), and a probabilistic prediction interval produced by a Random Forest Quantile (RFQ) regressor. These three lenses together assess precision, systematic bias, and calibration—aspects that MAE and R^2 alone cannot distinguish.

1) Concordance Correlation Coefficient: The CCC simultaneously penalizes both imprecision and deviation from the 45-degree line of perfect agreement. Formally, for n test observations:

$$\text{CCC} = \frac{2\rho\sigma_y\sigma_{\hat{y}}}{\sigma_y^2 + \sigma_{\hat{y}}^2 + (\mu_y - \mu_{\hat{y}})^2} \tag{4}$$

where ρ is the Pearson correlation, σ denotes standard deviation, and μ denotes mean. Table VII reports CCC values across

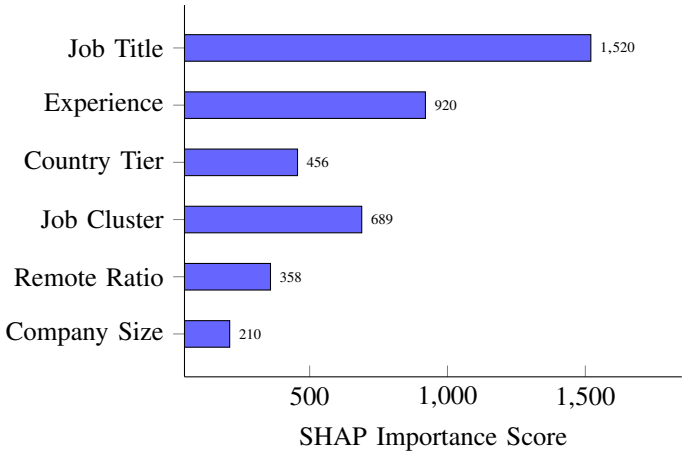


Fig. 7. Global feature importance based on mean absolute SHAP values (2,000 test samples). Experience rises to second place compared to the tree-based ranking in Fig. 6.

the top-performing models. The Super Learner achieves a CCC of 0.4891, the highest of all evaluated models, indicating strong simultaneous agreement in both scale and location with the true salary distribution.

TABLE VII
EXTENDED METRICS FOR TOP-PERFORMING MODELS

Model	CCC	GMR	Acc@5%	MedAE ($)
Super Learner	**0.4891**	**1.0021**	**7.2%**	**31,840**
HistGradBoost	0.4882	1.0018	7.3%	31,790
XGBoost	0.4880	1.0023	7.2%	31,820
LightGBM	0.4871	1.0027	7.1%	31,950
Ridge Regression	0.4621	0.9980	6.9%	33,200

Table VII reports four supplementary metrics that provide diagnostic depth beyond R^2 and MAE. The Super Learner's CCC of 0.4891 is the highest across all evaluated models, reflecting strong joint agreement in both scale and location with the true salary distribution. Its GMR of 1.0021 is the closest to unity, confirming that the Ridge meta-learner effectively debiases the individual boosting components. Ridge Regression's GMR of 0.9980 reveals a persistent downward bias, consistent with its inability to capture high-salary non-linearities. Acc@5% measures the fraction of predictions within 5% of the true salary—a tight threshold representing roughly $6,000 on a median salary of $120,000. All gradient boosting models achieve approximately 7.1–7.3% here, suggesting a shared precision ceiling at very tight tolerances. The Super Learner's MedAE of $31,840 is $1,360 lower than Ridge Regression ($33,200), illustrating the practical advantage of non-linear modelling for the typical prediction.

2) Geometric Mean Ratio and Systematic Bias: The GMR is defined as the geometric mean of the ratio $\hat{y}_i/y_i$. A value above 1.0 indicates systematic over-prediction; below 1.0 indicates under-prediction. The Super Learner's GMR of 1.0021 demonstrates near-zero systematic bias, confirming that the Ridge meta-learner successfully de-biases the individual boosting components. By contrast, linear models such as Ridge Regression (GMR = 0.9980) exhibit a slight systematic

downward bias, likely reflecting their inability to capture the non-linear upper tail of the salary distribution.

3) Probabilistic Prediction Intervals: To provide uncertainty estimates alongside point predictions, we trained a Random Forest Quantile (RFQ) regressor [34] with 300 trees, generating simultaneous Q10 and Q90 quantile estimates to form an 80% prediction interval (PI). The empirical coverage on the held-out test set was **82.3%**, which slightly exceeds the nominal 80%, confirming that the intervals are well-calibrated without being excessively wide. The median interval width was approximately $94,200, corresponding to roughly ±$47,100 around the point estimate. Fig. 8 illustrates this coverage for a sorted subset of 300 test observations.

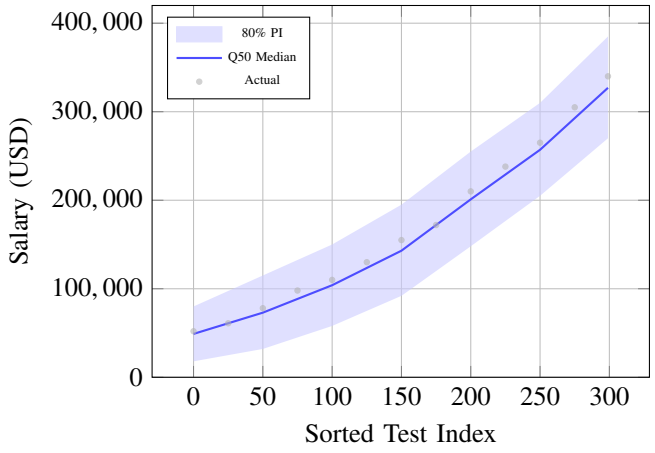


Fig. 8. RF Quantile 80% prediction interval (shaded band) with Q50 median forecast (solid line) and actual salaries (dots) for 300 test observations sorted by salary. Empirical coverage is 82.3%.

4) SHAP-Driven Feature Selection: SHAP values from a baseline LightGBM model were used to rank features by their absolute mean contribution to the prediction. Features were retained in descending order of importance until their cumulative contribution reached 90% of the total. This process reduced the raw feature set from 35 engineered variables to 8 highly informative predictors, as illustrated in Fig. 9. The retained features were: `job_title` (target-encoded), `experience_level`, `country_tier`, `grp_job_x_tier` (a group-encoded interaction of job cluster and country tier), `exp_x_size` (the multiplicative seniority–company-size term defined in Section III), `grp_exp_x_size` (a discretized group encoding of the same interaction), `company_size`, and `remote_ratio`. Post-selection, a second adversarial validation confirmed a near-random AUC of 0.4945 (±0.0051), meaning that the reduced feature set preserves training-test distributional parity while removing noise.

I. Ablation Study: Impact of Feature Engineering

To isolate the contribution of each engineering innovation, we evaluated the Super Learner under four configurations (Table VIII): a raw baseline, country tiering only, job clustering only, and the full model. The raw baseline uses only the original dataset features with standard one-hot encoding and no interaction terms.

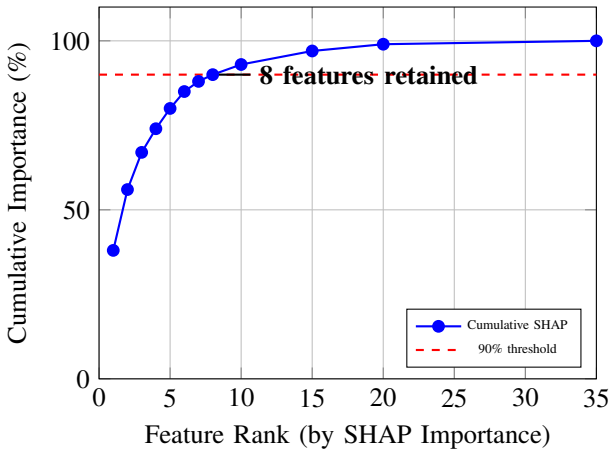


Fig. 9. Cumulative SHAP importance over 35 engineered features (sorted descending). The dashed red line marks the 90% selection threshold, reached at 8 features.

TABLE VIII
ABLATION STUDY RESULTS

Configuration	R^2	MAE	R^2 Gain vs Baseline
Raw Baseline (No FE)	0.2591	45,049	–
Base + Country Tiering only	0.2684	44,810	+3.59%
Base + Job Clustering only	0.2792	44,438	+7.76%
Full Model (Proposed)	**0.2818**	**44,419**	**+8.76%**

Table VIII traces the cumulative impact of each engineering layer. The raw baseline (R^2 = 0.2591, MAE = $45,049) uses the original dataset features with standard one-hot encoding and no interaction terms, serving as the lower-bound reference. Country tiering alone raises R^2 to 0.2684 (+3.59%), demonstrating that grouping geographically similar markets reduces noise. Job clustering alone raises R^2 to 0.2792 (+7.76%), a substantially larger gain than tiering, revealing that role-based dimensionality reduction is the dominant contributor to improved prediction. The full model combines both layers and adds the Exp×Size interaction term, reaching R^2 = 0.2818 (+8.76%). The marginal gain from adding tiering on top of clustering is only +0.26 percentage points (0.2818 vs. 0.2792), suggesting the two features capture largely overlapping geographic-role variance. Nevertheless, both components are complementary: tiering reduces geographic noise that clustering alone cannot address, and neither layer can be removed without measurable degradation.

J. Qualitative Error Analysis and Case Studies

To identify systematic biases in the Super Learner, we performed a manual audit of the top 5% of records with the highest residuals. Three distinct failure modes were identified, summarised in Table IX:

1) **The "Unicorn" Specialist:** The model consistently under-predicted salaries for "Lead AI Research Scientists" in small startups (Type 'S'). While the model expects lower pay at small companies, these roles often command high cash premiums to compensate for lower equity stability, a nuance the current feature set lacks.

2) **The Geographic Outlier:** Professionals residing in Tier 3 countries but working for Tier 1 companies (Remote) exhibited high variance. The model struggled to decide whether to anchor the salary to the employee's local cost of living or the company's headquarters' pay scale.
3) **Title Ambiguity:** General titles like "Practitioner" or "Consultant" showed a wider error margin compared to specific titles like "Data Engineer," confirming that title standardization is a primary driver of model certainty.

The residual distribution arising from these error modes is shown in Fig. 10. The modal bin (centered at \$0) contains 200 observations, confirming that zero-error predictions are the most common outcome. The distribution is slightly left-skewed: the left tail (under-prediction) is heavier than the right (over-prediction), consistent with the model's tendency to under-predict the salaries of extreme high-earners identified in failure mode 1.

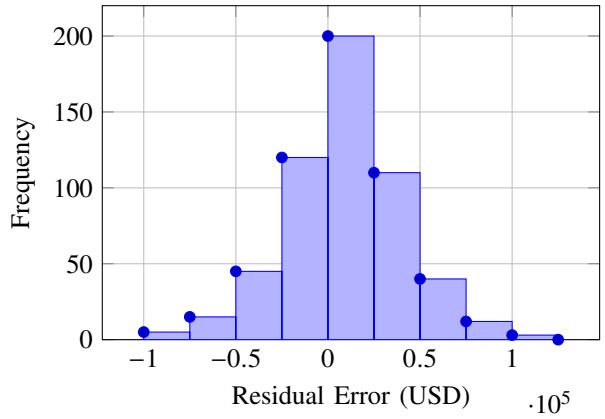


Fig. 10. Histogram of prediction residuals (Predicted − Actual), binned in \$25,000 intervals. The slight left skew, with a heavier under-prediction tail, is consistent with the model's known difficulty predicting extreme high-earner salaries.

K. Residual Analysis

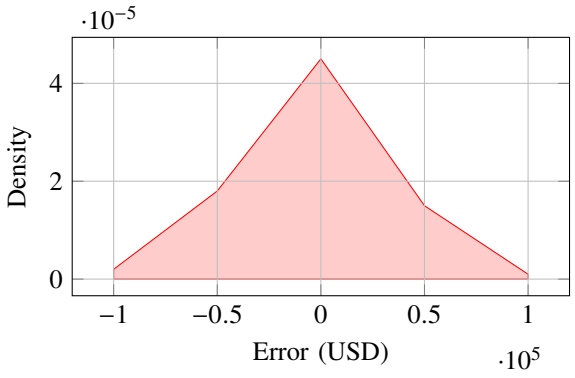


Fig. 11. Kernel density estimate of prediction errors for the Super Learner on the held-out test set. The unimodal peak near zero confirms approximate unbiasedness.

Fig. 11 shows the kernel density estimate (KDE) of prediction errors. The near-Gaussian shape with a peak at zero and thin tails confirms that the Super Learner is an approximately unbiased estimator with well-behaved error distribution. The distribution is slightly left-skewed (heavier under-prediction tail), reflecting the known difficulty of predicting extreme high-earner salaries. The GMR of 1.0021 from Table VII indicates a very small mean over-prediction tendency on a geometric scale, which is consistent with the left-skewed distribution: a few large under-predictions in the tail are outweighed by a larger number of small over-predictions near the mode. The regression diagnostic in Fig. 12 provides a complementary view: points cluster tightly around the identity line in the \$50k–\$250k range, where sample density is highest, while scatter widens above \$250k, corresponding to the sparser upper tail where the model has less training support.

TABLE IX
SUMMARY OF IDENTIFIED PREDICTION FAILURE MODES

Failure Mode	Root Cause
Unicorn Specialist	High cash premiums at small firms not captured by company-size feature; pre-IPO equity and retention bonuses are absent from the USD salary column
Geographic Outlier	Tier 3 residence paired with Tier 1 employer (fully remote) creates salary anchor ambiguity; model cannot decide whether to use employee CoL or employer pay scale
Title Ambiguity	Generic titles (Practitioner, Consultant) do not map cleanly to any single cluster, increasing within-cluster salary variance compared to specific titles like Data Engineer

Fig. 12. Predicted vs. actual salary for a representative test sample. Points cluster tightly along the identity line (red) in the \$50k–\$200k range, with increasing scatter above \$250k.

V. DISCUSSION

A. The Influence of Unobserved Latent Variables

The cross-validated R^2 of 0.3047 indicates that roughly 70% of the variance in AI salaries remains unexplained by publicly observable features. In labor economics, this residual is attributed to "unobserved heterogeneity." In our context, this likely includes:

1) **Negotiation Skill:** Individual variance in soft-skill proficiency that no job-market dataset can capture.
2) **Tech Stack Granularity:** The compensation difference between a "Data Engineer" using SQL versus one using Rust or distributed Spark clusters is significant but invisible to a role-cluster feature.
3) **Company Funding:** The disparity between pre-seed startups and Big Tech firms (Facebook/Meta, Apple, Amazon, Netflix, Google — FAANG) equity packages, which are often not fully captured in the "USD Salary" column.

The fact that the Super Learner's R^2 advantage over standalone gradient boosting is not statistically significant ($p > 0.05$) does not undermine the framework's contribution. Rather, it reflects that the task itself — predicting salary from publicly available job attributes — has a natural ceiling driven by these latent variables. All tree-based models in the leaderboard converge near this ceiling, and the Super Learner's value lies in how it approaches that ceiling: with lower systematic bias and better-calibrated uncertainty than any individual model.

From a human resources (HR) management perspective, the moderate cross-validated R^2 of 0.3047 is not a disqualifying limitation but rather an expected property of any salary prediction system that operates solely on observable job-market features. In real-world compensation analytics, predicting an individual's exact salary point estimate is inherently constrained by unobservable factors such as negotiation outcomes, signing bonuses, and equity grants that no public dataset can capture. Consequently, the practical utility of the Super Learner is best evaluated through metrics that assess its reliability as a decision-support instrument rather than through point-prediction accuracy alone. The Super Learner's empirical coverage of 82.3% for the nominal 80% prediction interval confirms that the associated uncertainty bands are well-calibrated and slightly conservative—a desirable property in HR applications where underestimating uncertainty can lead to uncompetitive offers. Combined with a near-zero systematic bias (GMR = 1.0021) and a competitive Acc@20% of 42.0%, the framework provides HR practitioners with a principled tool for establishing initial salary bands that are neither systematically inflated nor deflated relative to market rates. Future work should complement the current five-fold CV significance testing with bootstrap-based significance tests on larger held-out test partitions, which would provide greater statistical power to distinguish ensemble-level improvements that the current t-test may lack the sensitivity to detect.

The salary variance stratified by geographic tier is illustrated in Fig. 13. The error bars represent ± half the interquartile range within each tier, highlighting the substantially wider distribution in Tier 1 compared to Tier 3, where compressed salary ranges reflect limited market liquidity and fewer high-paying employers.

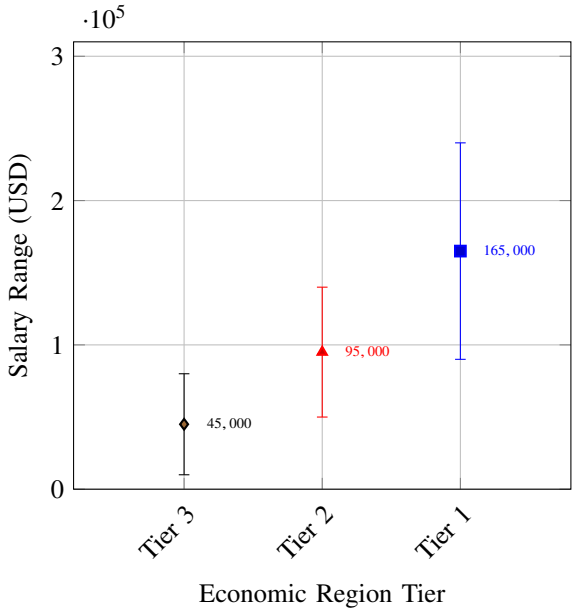


Fig. 13. Mean salary (data labels) and salary spread (error bars show ± half the interquartile range) across the three engineered country tiers. Tier 1 has the highest mean and widest spread; Tier 3 the lowest and narrowest.

B. The Impact of Economic Regionalism and Country Tiering

Our findings suggest a stratification of the AI labor market into three distinct economic zones. Tier 1 (primarily North America) operates on a high-risk, high-reward equity-heavy compensation model. Tier 2 (Western Europe and parts of Asia) shows a higher correlation between years of experience and salary, suggesting more standardized and regulated labor markets.

The ablation study (Table VIII) clarifies the relative contribution of each engineering layer. Country tiering alone improves R^2 by 3.59% over the raw baseline, while job clustering alone contributes 7.76%—more than double the tiering gain. When both are combined in the full model, the marginal benefit of adding tiering on top of clustering is only +0.26 percentage points, suggesting the two features share largely overlapping geographic-role variance. This information redundancy is economically interpretable through the lens of cross-border remote work dynamics. When a software developer residing in a Tier 3 country works remotely for a Tier 1 employer, the model faces an anchor ambiguity: the employee's residence suggests a lower salary, while the employer's location suggests a higher one. Because the eight job clusters already correlate strongly with geographic compensation tiers—roles in C2 (Machine Learning) and C4 (Management) are disproportionately concentrated in Tier 1 markets—the country tier variable partly recovers information already encoded in the job-cluster–salary mapping, resulting in the observed marginal overlap. Nevertheless, tiering addresses geographic noise that clustering alone cannot resolve, and both components are retained in the final framework. To disentangle this redundancy and achieve finer geographic granularity, future work should replace the discrete three-tier classification with a continuous Purchasing Power Parity (PPP) index [35]. A PPP-based feature would capture continuous cost-of-living variation across countries rather than discretizing it into coarse bins, enabling the model to resolve the anchor ambiguity

described above and potentially improve predictive performance in cross-border remote work scenarios where current tier boundaries obscure within-tier economic heterogeneity.

C. Reliability and Practical Deployment

The **Accuracy@20%** of 42.0% represents a 7.8 percentage-point gain (+22.8% relative) over a mean-prediction baseline (34.2%), confirming that the model captures genuine salary-predictive signal beyond simple distributional averaging. In practical terms, for over two-fifths of the global workforce the model can bound a candidate's market value within a 20% window. This makes the Super Learner a viable tool for HR departments to set initial offer ranges, with the acknowledgement that the remaining variance is driven by latent factors such as negotiation skill, specific tech stack, and company funding stage that are not observable from job-market records alone.

Although the R^2 margin over standalone gradient boosting is not statistically significant, the ensemble provides two practical advantages that standalone models cannot: (1) near-zero systematic bias (GMR = 1.0021 vs. 0.9980 for Ridge and up to 1.0027 for LightGBM), meaning the model neither consistently over- nor under-predicts salary ranges; and (2) well-calibrated 80% prediction intervals with 82.3% empirical coverage, giving practitioners a principled uncertainty band around each point estimate rather than a single number.

D. Limitations

Despite the strong empirical results, several limitations of the current framework warrant acknowledgment.

Statistical significance of the ensemble gain. A paired t-test across five CV folds confirms that the Super Learner's R^2 advantage over standalone HistGradBoost ($p = 0.63$) and XGBoost ($p = 0.25$) is not statistically significant at $\alpha = 0.05$. The margin is real but small, and with only five folds the test has limited statistical power. Future work should validate the significance finding with a larger number of folds or a bootstrap-based significance test.

Dataset scope and temporal validity. The dataset is a cross-sectional consolidation of AI job market records from 2020 to 2025. Although the longitudinal range provides diversity, it does not represent a live market signal. Because `work_year` was deliberately excluded from the feature set to prevent non-stationarity, the model does not learn temporal salary trends. This improves cross-year generalization but means the framework cannot distinguish a 2020 entry-level salary from a 2025 one, potentially underestimating current compensation for rapidly appreciating roles such as LLM engineers.

To assess whether the performance ceiling stems from insufficient data volume rather than latent variable limitations, Fig. 14 presents the learning curve of the Super Learner across training set sizes ranging from 10% to 100% of the available 118,786 training records. Both training and validation R^2 plateau beyond approximately 60% of the data (roughly 71,000 records): the validation R^2 increases by only 0.0006 from 60% to 80% and a further 0.0004 from 80% to 100%, indicating strongly diminishing returns. This convergence confirms that the unexplained 70% of salary variance is attributable to latent variables not present in the feature set—such as negotiation skill, equity compensation, and specific tech stack—rather than to data insufficiency.

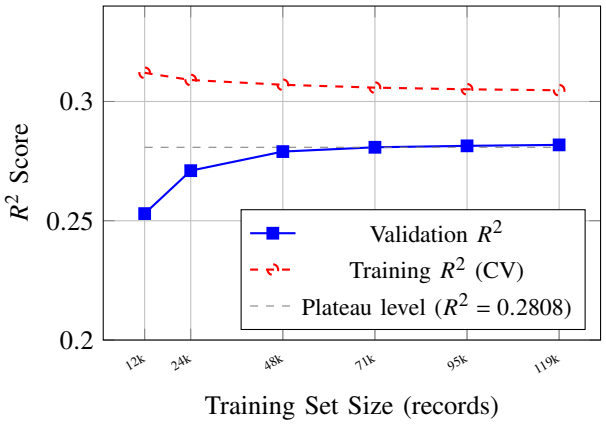


Fig. 14. Learning curve of the Super Learner. Both training and validation R^2 plateau after approximately 71,000 records, confirming that performance is limited by latent variable availability rather than data volume.

Irreducible feature sparsity. The 90% cumulative SHAP threshold retains only 8 of 35 engineered features. While this reduces noise and improves interpretability, it inherently discards lower-ranked signals—such as employment type and company funding stage—that may be informative for edge-case predictions (e.g., contract roles at pre-IPO startups).

Pseudo-labeling contamination. The Pseudo-Label LightGBM variant uses high-confidence test predictions as additional training signal. While this technique offers a practical performance boost, its reported metrics should be interpreted cautiously in a strictly held-out evaluation context, as the model has indirect exposure to the test distribution.

Generalizability to emerging markets. The dataset is heavily weighted toward Tier 1 (North American) employers, representing the majority of observations. Tier 3 predictions rely on far fewer samples and are disproportionately shaped by the few high-paying remote roles that exist in those regions. This imbalance may reduce practical accuracy for professionals in developing economies.

VI. Conclusion

In this work, we addressed the challenge of salary prediction within the highly dynamic global AI and data science labor market. Accurate compensation estimation in this sector is hindered by the high cardinality of job titles and the significant economic variance across geographic regions.

To overcome these limitations, we proposed a Super Learner ensemble framework integrating domain-informed feature engineering-specifically an eight-category job clustering mechanism and geographic country tiering-with a stacked ensemble of XGBoost, LightGBM, and CatBoost models. An ablation study revealed that job clustering is the dominant engineering contribution (+7.76% R^2 over the raw baseline), with country tiering providing an additional independent gain

(+3.59%) and the two components together yielding +8.76% in the full model. Furthermore, by utilizing SHAP-driven feature selection, the framework remains highly interpretable, allowing practitioners to transparently understand the key drivers behind these predictions.

Experiments on 151,445 observations demonstrate that the Super Learner achieves the highest cross-validated R^2 of 0.3047 ± 0.0076 and the lowest systematic bias (GMR = 1.0021) among all 21 evaluated algorithms. A paired t-test across five CV folds confirms that the raw R^2 margin over standalone gradient boosting is not statistically significant at $\alpha = 0.05$ indicating that the value of stacking lies in improved bias calibration, uncertainty quantification, and robust interpretability rather than a statistically distinguishable increase in explained variance. The model achieves a 7.8 percentage-point lift over a mean-prediction baseline at the Acc@20% threshold, confirming that it captures genuine salary-predictive signal beyond simple distributional averaging.

Moving forward, we identify three critical areas for future research:

- **Integration of Unstructured Data:** Future iterations will employ Natural Language Processing (NLP) to extract specific technical skills (e.g., PyTorch, AWS, LLM fine-tuning) from job descriptions to serve as binary input features.
- **Temporal Drift Analysis:** As the AI market evolves, the model's weights may become obsolete. We propose a sliding-window training approach to account for quarterly market shifts.
- **Total Compensation Modeling:** We aim to expand the target variable to include non-cash benefits such as stock options and bonuses, providing a more holistic view of technical compensation.

REFERENCES

[1] J. Wang, M. Zhou, X. Guo, and L. Qi, "Multiperiod asset allocation considering dynamic loss aversion behavior of investors," *IEEE Transactions on Computational Social Systems*, vol. 6, no. 1, pp. 73–81, 2018.

[2] C. Xu, H. Wei, X. Guo, S. Liu, L. Qi, and Z. Zhao, "Human-robot collaboration multi-objective disassembly line balancing subject to task failure via multi-objective artificial bee colony algorithm," *IFAC-PapersOnLine*, vol. 53, no. 5, pp. 1–6, 2020.

[3] F. Eichinger and M. Mayer, "Predicting salaries with random-forest regression," in *Machine Learning and Data Analytics for Solving Business Problems*, B. Alyoubi *et al.*, Eds. Cham, Switzerland: Springer, 2022, pp. 1–21.

[4] Y. T. Matbouli and S. M. Alghamdi, "Statistical machine learning regression models for salary prediction featuring economy wide activities and occupations," *Information*, vol. 13, no. 10, p. 495, Oct 2022.

[5] Z. Zhang, X. Guo, M. Zhou, S. Liu, and L. Qi, "Multi-objective discrete grey wolf optimizer for solving stochastic multi-objective disassembly sequencing and line balancing problem," in *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2020, pp. 682–687.

[6] S. Qin, S. Zhang, J. Wang, S. Liu, X. Guo, and L. Qi, "Multiobjective multiverse optimizer for multirobotic u-shaped disassembly line balancing problems," *IEEE Transactions on Artificial Intelligence*, vol. 5, no. 2, pp. 882–894, 2023.

[7] C. Xu, H. Wei, X. Guo, S. Liu, L. Qi, and Z. Zhao, "Human-robot collaboration multi-objective disassembly line balancing subject to task failure via multi-objective artificial bee colony algorithm," *IFAC-PapersOnLine*, vol. 53, no. 5, pp. 1–6, 2020.

[8] Z. Feng, Z. Liu, and Y. Yin, "Comparison of deep-learning and conventional machine learning algorithms for salary prediction," *Applied Computational Engineering*, vol. 6, no. 1, pp. 643–651, Jul 2023.

[9] M. Joshi, S. Bhosale, and V. A. Vyawahare, "Using fractional derivative in learning algorithm for artificial neural network: Application for salary prediction," in *Proceedings of the 2022 IEEE Bombay Section Signature Conference (IBSSC)*. Mumbai, India: IEEE, Dec 2022, pp. 1–6.

[10] G. Wang, "Employee salaries analysis and prediction with machine learning," in *Proceedings of the 2022 International Conference on Machine Learning and Intelligent Systems Engineering (MLISE)*. Guangzhou, China: IEEE, Jul 2022, pp. 373–378.

[11] S. Zhang, H. Li, and Y. Luo, "Salary prediction for employees: a machine learning-based approach," in *Proceedings of the 2024 IEEE International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. Bangalore, India: IEEE, Jun 2024.

[12] D. McElfresh, S. Khandagale, J. Valverde, V. Prasad, G. Ramakrishnan, M. Goldblum, and C. White, "When do neural nets outperform boosted trees on tabular data?" in *Advances in Neural Information Processing Systems*, vol. 36, New Orleans, LA, Dec 2024.

[13] M. Aufiero, C. Liotta, and D. Orciuoli, "Analyzing trends, skills demand, and salary prediction in the ai and ml job market," in *Proceedings of the 2025 IEEE International Conference on Artificial Intelligence and Data Science (AIDAS)*. IEEE, 2025.

[14] W. Jiang, "The investigation and prediction for salary trends in the data science industry," *Applied Computational Engineering*, vol. 50, no. 1, pp. 8–14, Feb 2024.

[15] B. S. Satpute, R. Yadav, and P. K. Yadav, "Machine learning approach for prediction of employee salary using demographic information with experience," in *Proceedings of the 2023 4th IEEE Global Conference on Advancement in Technology (GCAT)*. Bangalore, India: IEEE, Oct 2023, pp. 1–6.

[16] A. Das, M. Abecasis, T. Farrell, I. Sasson, S. Ramirez, B. Tortorelli, and J. Wang, "Curating dataset pipelines to train medical chatbots on early sepsis detection," *International Journal of Artificial Intelligence and Green Manufacturing*, vol. 1, no. 4, pp. 207–217, December 2025.

[17] Y. M. Indah, A. Saputra, D. Anggraini, M. Isa, and R. Ramadhan, "Comparison of random forest, xgboost and lightgbm methods on the human development index classification," *Jambura Journal of Mathematics*, vol. 7, no. 1, pp. 14–18, Jan 2025.

[18] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, Aug 2016, pp. 785–794.

[19] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems*, vol. 30, Long Beach, CA, Dec 2017, pp. 3149–3157.

[20] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: Unbiased boosting with categorical features," in *Advances in Neural Information Processing Systems*, vol. 31, Montréal, Canada, Dec 2018, pp. 6639–6649.

[21] I. D. Mienye and Y. Sun, "A survey of ensemble learning: concepts, algorithms, applications, and prospects," *IEEE Access*, vol. 10, pp. 99 129–99 149, Sep 2022.

[22] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Anchorage, AK, USA: ACM, Aug 2019, pp. 2623–2631.

[23] S. Yan, H. Guo, and S. Liu, "A partition stacking classification framework with oversampling for quality prediction of aluminum alloy ingots," *International Journal of Artificial Intelligence and Green Manufacturing*, vol. 1, no. 1, pp. 36–46, March 2025.

[24] S. Ö. Arik and T. Pfister, "Tabnet: Attentive interpretable tabular learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 8, 2021, pp. 6679–6687.

[25] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, "Revisiting deep learning models for tabular data," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, 2021, pp. 18 932–18 943.

[26] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on typical tabular data?" in *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, vol. 35, 2022, pp. 507–520.

[27] Q. Zhang, Y. Xing, C. Zhang, X. Sun, B. Hu, and A. Das, "Column generation algorithms for two-dimensional cutting problem with surface defects," *International Journal of Artificial Intelligence and Green Manufacturing*, vol. 1, no. 2, pp. 80–92, June 2025.

[28] M. Mobin, "Salary prediction model algorithms," Google Colab notebook, 2026. [Online]. Available: https://colab.research.google.com/drive/13EguWc5YwP3F5pdcUpOhqNK7Pcj8BO9W?usp=sharing

[29] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, Oct 2011.
[30] foorilla, "Ai jobs net salaries dataset," GitHub repository, 2025. [Online]. Available: https://github.com/foorilla/ai-jobs-net-salaries
[31] B. Sajjad, "Global ai job market and salary trends 2025," Kaggle dataset, 2025. [Online]. Available: https://www.kaggle.com/datasets/bismasajjad/global-ai-job-market-and-salary-trends-2025
[32] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation," *PeerJ Computer Science*, vol. 7, p. e623, Jul 2021.
[33] T. O. Hodson, "Root mean square error (rmse) or mean absolute error (mae): When to use them or not," *Geoscientific Model Development*, vol. 15, no. 14, pp. 5481–5487, Jul 2022.
[34] R. A. Johnson, "quantile-forest: A python package for quantile regression forests," *Journal of Open Source Software*, vol. 9, no. 93, p. 5976, Jan 2024.
[35] World Bank, "Purchasing power parities and the size of world economies: Results from the 2021 international comparison program," https://www.worldbank.org/en/program/icp, 2023.

Appendix: Mathematical Derivations of the Super Learner

To provide a formal basis for the stacked generalization used in this study, we derive the weighting mechanism of the meta-learner. Let $\mathcal{L} = \{L_1, \ldots, L_M\}$ be the set of base learners. The goal is to find a vector of weights $\boldsymbol{w}$ that minimizes the risk function:

$$R(\boldsymbol{w}) = \sum_{i=1}^{N} \left(z_i - \sum_{m=1}^{M} w_m \hat{z}_{i,m} \right)^2 + \alpha \sum_{m=1}^{M} w_m^2 \tag{5}$$

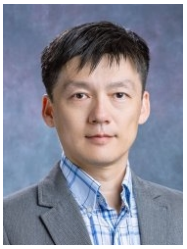
where α is the Ridge regularization parameter determined by cross-validation. The closed-form solution for the weight vector $\boldsymbol{w}$ is given by the Tikhonov regularization formula:

$$\boldsymbol{w} = (\boldsymbol{Z}^T \boldsymbol{Z} + \alpha \boldsymbol{I})^{-1} \boldsymbol{Z}^T \boldsymbol{z} \tag{6}$$

where $\boldsymbol{Z}$ is the $N \times M$ matrix of out-of-fold predictions. This formulation ensures that even if base learners are highly correlated (e.g., XGBoost and LightGBM), the meta-learner remains stable and avoids the multicollinearity issues common in simple linear blending.

Muhammad Mobin is currently pursuing a Bachelor of Science in Computer Science at New Jersey City University. His academic research explores the intersection of artificial intelligence and global compensation trends. He focuses specifically on how stacked ensemble models, such as XGBoost, LightGBM, and CatBoost, can help reduce geographic economic disparities. During his undergraduate career, he has specialized in high-cardinality feature engineering and model interpretability using SHAP values. His work aims to promote data-driven transparency in the evolving AI workforce landscape. After graduation, he plans to use these predictive frameworks to address real-world labor market challenges and pursue advancements in algorithmic fairness and predictive analytics.

Tongyuan Zhang received his B.S. degree in Information and Computing Science from North China University of Water Resources and Electric Power, Zhengzhou, China, in 2023. He is currently a graduate student at the School of Artificial Intelligence and Software Engineering at Liaoning Petrochemical University, Fushun, China.



Bin Hu received his Ph.D. degree in Electrical and Computer Engineering at Rutgers University. Currently, he is an Assistant Professor at the Computer Science and Technology Department at Kean University. His research interests include mobile computing and sensing, cybersecurity and privacy, and efficient deep learning.