MATHEMATICAL MODELING AND COMPUTING, Vol. 12, No. 3, pp. 993–1004 (2025)

Explainable AI and robust forecasting of global salary trends: Addressing data drift and unseen categories with tree-based models

Shakhovska N. B.

Lviv Polytechnic National University,
12 S. Bandera Str., 79013, Lviv, Ukraine

(Received 31 August 2025; Revised 27 September 2025; Accepted 3 October 2025)

This article studies salary prediction under distributional drift using explainable boosting models and hybrid forecasting. We integrate unseen-aware feature engineering, robust objectives, SHAP-based interpretability, drift detection, and time-series forecasting (Prophet/SARIMAX) on multi-year data (2020–2024), and report a comprehensive evaluation aligned with typical MMC guidelines. Modern salary data are heterogeneous, heavy-tailed, and non-stationary. Therefore we combine robust tree-based learners with drift monitoring and explainable forecasting to prioritize stable absolute error, transparency, and maintainability over raw variance capture. Our best integrated pipeline reaches $R^2 = 0.31$ on a 2024 hold-out while keeping MAE/RMSE stable across folds, and uncovers year-to-year drift that necessitates periodic retraining monthly and quarterly forecasts indicate a sustained upward trend with seasonality, where SARIMAX captures short-term fluctuations and Prophet yields interpretable trend decompositions.

Keywords: *salary prediction; explainable AI; SHAP; data drift; CatBoost; LightGBM; Prophet; SARIMAX.*

2010 MSC: 62E10, 68T05 **DOI:** 10.23939/mmc2025.03.993

1. Introduction

Predicting salaries is a long-standing and multifaceted problem in labor economics, human resource management, and data-driven policy design. Salary data capture not only monetary compensation but also reflect broader dynamics of labor supply and demand, skill scarcity, geographic mobility, and macroeconomic shocks. However, modeling salaries is highly challenging because the available datasets are heterogeneous, containing roles across different industries, countries, and levels of experience; strongly skewed, as salaries often follow long-tailed distributions with extreme values in specific regions or for niche skills; and non-stationary, since the distribution of salaries shifts significantly over time due to technological change, inflation, or global events such as the COVID-19 pandemic. Despite the relevance of the problem, existing research has focused predominantly on traditional econometric models or machine learning approaches designed for static prediction tasks. Linear regression and random forest models have been used to capture relationships between job characteristics and salaries, but their predictive power is often limited to narrow domains such as IT companies or specific national markets. More recent studies explore deep learning and embedding-based architectures trained in job descriptions or resumes, which improve accuracy but operate as black boxes, offering little transparency to stakeholders. Interpretability remains a crucial yet underexplored aspect of salary prediction. HR managers, policymakers, and individual workers are unlikely to adopt recommendations from opaque models whose decision mechanisms are not understood. This lack of transparency undermines trust and prevents the integration of salary analytics into real-world decision-making pipelines. Furthermore, existing approaches rarely account for data drift, the fact that the statistical properties of salary data evolve rapidly. A model trained on pre-pandemic salaries, for example, may be severely miscalibrated when applied to post-pandemic data. Finally, unseen categories, such as new job titles or emerging hybrid roles (e.g., 'AI Engineer' or 'MLOps Specialist'), introduce systematic errors, since most models are not designed to generalize to novel labels. The combination of these issues points to the importance

of developing a new generation of salary prediction frameworks that are robust to distributional shifts, resilient to unseen categories, and explainable to human stakeholders. The aim of our work is to fill the identified void in the literature by integrating feature normalization, drift analysis, robust boosting models, explainable AI, and hybrid forecasting into a single coherent pipeline.

The main contribution of the paper is given below.

1. Unseen-aware feature engineering is performed by normalizing raw job titles into professional families; grouping rare categories; interaction features (*family@location*); and an explicit unseen-category indicator.
2. Integration of drift detection is performed with PSI, KS, χ^2, and adversarial AUC quantify shifts and trigger retraining.
3. In addition to the traditional approach, explainable forecasting is added. We combine Prophet/SARIMAX with model interpretation (permutation importance and SHAP) to disentangle drivers of salary dynamics.
4. The robustness between segment evaluation across job families, countries, and experience levels highlights segment-specific weaknesses.
5. Hybrid Forecasting + ML is combined throw prophet captures global trend, while boosting models refine segment-level residuals.
6. Fairness and bias assessment made with analysis of systematic error patterns across regions and experience levels.

Taken together, our findings have broader implications beyond salary prediction. The methodological pipeline we propose is combining unseen-aware encoding, drift detection, robust boosting, explainable AI, and hybrid forecasting. It can be generalized to other socio-economic prediction problems where categories evolve and distributions shift, such as education analytics, job recommender systems, or regional economic forecasting.

2. Related works

The prediction of salaries has been studied in economics, management, and data science, albeit often in narrow contexts. Some early works applied linear regression and random forests to company-level salary data, reporting only moderate predictive power with coefficients of determination around $R^2 \approx 0.20$. Other studies employed neural networks with embeddings of job descriptions, which achieved higher accuracy but at the expense of interpretability. However, such approaches are often sensitive to distributional drift and do not generalize well to unseen categories such as newly emerging job titles.

Ensemble methods such as Gradient Boosted Decision Trees (GBDT), LightGBM, and CatBoost have become widely used for tabular prediction tasks. LightGBM [4] is efficient for large datasets but requires careful handling of categorical variables. CatBoost [8] provides native categorical encodings and is robust against overfitting on rare categories, which makes it particularly suitable for salary prediction. Yet, prior research rarely benchmarks CatBoost in this specific context.

Model interpretability has only recently been integrated into salary analytics. SHAP values [6] have become standard in finance and healthcare, but their application to labor market analytics remains limited. Most salary prediction studies treat models as black boxes, and there is a lack of systematic efforts to combine explainability with predictive modeling, which limits trust and adoption by policymakers and HR practitioners.

Concept drift has been extensively documented in credit scoring, fraud detection, and predictive maintenance [3]. In salary datasets, drift arises from evolving job roles, geographic relocation, and macroeconomic shocks. Nevertheless, systematic drift detection methods such as PSI, KS, and adversarial validation are rarely applied. Moreover, the challenge of unseen job titles, which accounted for nearly 28% of our test data, is largely neglected in the literature.

Forecasting wages has traditionally been approached with econometric models such as ARIMA and SARIMAX [1]. Prophet [9] offers flexibility and scalability but is rarely applied to salary datasets at the micro level (segmented by job family or location). Existing works typically forecast aggregate wage indices and not disaggregated salary distributions.

Table 1. Comparison of existing approaches and this study.

Study / Approach	Models	Drift	Unseen	Explain.	Forecast	Strengths	Limitations
Traditional ML (2018)	Linear, RF	✗	✗	✗	✗	Simple, interpretable	Low accuracy
Deep Text Models (2020)	Neural Nets	✗	✗	✗	✗	Rich text embeddings	Black-box, drift-sensitive
LightGBM [4]	GBDT	✗	Partial	Limited	✗	Efficient, scalable	Needs manual encoding
CatBoost [8]	CatBoost	✗	Partial	Limited	✗	Native categorical handling	No forecasting
ARIMA / SARIMAX [1]	Time series	✗	✗	✗	✓	Standard econometrics	Weak micro-level fit
Prophet [9]	Additive TS	✗	✗	✗	✓	Scalable, interpretable	Limited interactions
Wang et al. [11]	SHAP-fairness	✗	✓	✓	✗	Fairness insights	No forecasting
Kumar & Li [5]	Drift detection	✓	✗	✗	✗	Strong drift eval.	No prediction
Novak et al. [7]	Hybrid Prophet+GBDT	Partial	Partial	Limited	✓	Better forecasts	Limited interpretability
Chen et al. [2]	LLM analytics	✗	Partial	Limited	Partial	Rich embeddings	Black-box, bias risks
This Study (2025)	CatBoost, LGBM, Robust GBR + Prophet / SARIMAX	✓	✓	✓	✓	Integrated, robust, explainable	Moderate R^2, retraining needed

From this literature, several unresolved challenges emerge. There are no integrated frameworks that combine prediction with drift monitoring. LightGBM with a Tweedie objective is sensitive to misspecification of the variance-power parameter: understating the power over-penalizes large outcomes, whereas overstating it inflates variance and destabilizes training; moreover, tree ensembles capture high-order interactions that complicate causal interpretation, and SHAP attributions can be unstable under multicollinearity. In CatBoost with a log-transformed target, naïve back-transformation without a smearing correction compresses dispersion and attenuates extremes, so absolute errors remain stable while R^2 deteriorates; ordered target statistics further partially pool rare categories toward global means, under-representing between-group variance. The Huberized gradient-boosting regressor down-weights outliers but can bias the upper tail and underfit genuinely high salaries; its performance depends on the transition parameter and provides limited distributional uncertainty. Prophet's additive trend – seasonality with piecewise-linear changepoints struggles to represent strong autocorrelation and abrupt regime shifts unless the changepoint prior is heavily relaxed, and its uncertainty quantification can be miscalibrated under heteroskedasticity. SARIMAX requires (quasi-) stationarity and correct order identification; parameters become unstable under structural breaks and covariate shift, and multistep forecasts accumulate misspecification when exogenous regressors are omitted or noisy. The hybrid Prophet-plus-residual-boosting scheme risks double counting seasonal structure and overfitting high-frequency noise, while dependence between base and residual models complicates uncertainty propagation and diminishes end-to-end interpretability. Finally, the drift diagnostics employed here have caveats: PSI depends on binning and is univariate, KS ignores multivariate dependence, and adversarial AUC conflates covariate and label shift, so any single trigger can be noisy without corroboration.

Existing studies do not address the appearance of new job titles across years. Explainable AI has not been systematically applied to salary prediction or forecasting. Finally, no prior studies combine machine learning with time-series forecasting for salaries.

Recent years have seen an increased focus on fairness, drift-awareness, and hybrid forecasting in salary and labor market analytics. Wang et al. [11] introduced fairness-aware models using SHAP explanations to detect systematic biases. Kumar and Li [5] provided a comparative study of drift detection techniques on HR datasets, highlighting the necessity of monitoring salary models in production. Novak et al. [7] proposed hybrid Prophet+GBDT approaches for wage forecasting, bridging econometrics and machine learning. Chen et al. [2] explored large language models for job analytics, though challenges remain in interpretability and domain adaptation.

Issues of missing data handling in large-scale systems have also been discussed, for instance by Wang et al. [10], who proposed a novel approach for imputation in big data interfaces. Prior studies seldom integrate prediction + drift monitoring, handle unseen titles explicitly, or combine XAI with forecasting; our pipeline addresses all three, and is- to our knowledge- among the first to pair Prophet/SARIMAX with drift-aware, explainable tabular models for disaggregated salary analytics.

3. Materials and methods

3.1. Dataset and preprocessing

The dataset `merged_salaries.csv` spans the years 2020–2024 and includes approximately sixty thousand records. Each entry provides information on job title, employment type, experience level, company location, company size, and work year, with the target variable being the salary expressed in USD. The heterogeneity of the data, which reflects multiple industries and geographical regions, requires careful preprocessing to ensure stability and generalizability of the models.

In order to reduce sparsity and address the issue of unseen categories in the test data, job titles were normalized into broader families such as data scientist, machine learning engineer, or data analyst. Very rare categories (with fewer than fifty samples in the training set) were aggregated into a generic "Other" group. Experience level and company size, being naturally ordered categorical variables, were encoded ordinally: experience levels were mapped from zero (entry) to three (executive), while company size was mapped from zero (small) to two (large).

To stabilize the variance of the dependent variable, salaries were log-transformed according to

$$y_i' = \log(1 + y_i), \tag{1}$$

where y_i is the original salary value. Extreme values were further controlled through winsorization at the first and ninety-ninth percentiles:

$$y_i'' = \min\{\max(y_i, q_{0.01}), q_{0.99}\}, \tag{2}$$

with $q_{0.01}$ and $q_{0.99}$ denoting empirical quantiles. This combination of transformations ensured that the long-tailed nature of salary distributions was reduced without losing interpretability.

3.2. Drift detection

Since salaries and employment structures evolve from year to year, detecting distributional drift was essential. Four complementary approaches were used. First, the population stability index (PSI) was computed to assess how categorical or numerical feature distributions shift over time:

$$\mathrm{PSI} = \sum_{j=1}^{k} (p_j - q_j)\, \ln\left(\frac{p_j}{q_j}\right), \tag{3}$$

where p_j and q_j represent proportions in train and test samples. Second, the Kolmogorov–Smirnov statistic was employed to measure the maximal distance between empirical distributions of salaries:

$$\mathrm{KS} = \sup_x |F_{\mathrm{train}}(x) - F_{\mathrm{test}}(x)|. \tag{4}$$

Third, chi-squared tests were applied to categorical features, for example to compare the distribution of company sizes or locations across years. Finally, adversarial validation was performed by training a classifier to distinguish between training and test instances, with ROC AUC values above 0.7 being interpreted as strong evidence of covariate shift. The combination of these tests provided both univariate and multivariate perspectives on the extent of drift.

3.3. Model training

Three families of models were investigated. To accommodate the non-negative, heavy-tailed distribution of salaries, we employ gradient boosting with a Tweedie deviance (LightGBM-Tweedie). This objective, which encompasses Poisson-Gamma mixtures, provides more appropriate weighting across the support than squared-error losses and is well suited to right-skewed outcomes. For high-cardinality categorical inputs, we use CatBoost with ordered target statistics and a log-transformed target; the ordered scheme mitigates target leakage and stabilizes estimates for rare categories, while the log transformation attenuates heteroscedasticity. We additionally include a Huberized gradient-boosting regressor to reduce sensitivity to extreme observations by transitioning from quadratic to linear loss beyond a data-adaptive threshold.

For forecasting aggregated salary series, we adopt two complementary approaches. Prophet supplies a transparent additive decomposition of trend and multiple seasonalities and is practically robust to missingness and structural breaks, thereby facilitating component-level interpretation. By contrast, SARIMAX captures short-run autoregressive and moving-average dynamics more parsimoniously and permits the inclusion of exogenous covariates when available. Utilizing both classes enables us to privilege interpretability and long-run structure with Prophet while retaining sensitivity to short-horizon dynamics and covariate effects with SARIMAX.

Hyperparameters were tuned through randomized search with three-fold cross-validation, where the primary optimization objective was the mean absolute error (MAE). Model performance was reported using MAE, root mean squared error (RMSE), and the coefficient of determination (R^2), where $\hat{y}_i$ is predicted value:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \tag{5}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{6}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \tag{7}$$

We chose LightGBM (Tweedie) for skewed non-negative targets, CatBoost for native handling of categorical features with ordered statistics, and Huber-GBR to down-weight outliers; all hyperparameters were tuned via randomized 3-fold CV optimizing MAE.

3.4. Explainable AI (XAI)

To interpret the predictions of the boosting models, two complementary approaches were adopted. Permutation importance was used to evaluate the decrease in predictive accuracy when the values of a feature were permuted at random:

$$PI(f) = \frac{1}{M}\sum_{m=1}^{M}\Big(\mathcal{L}(\hat{y}^{(m)}_{\mathrm{perm}(f)}, y) - \mathcal{L}(\hat{y}, y)\Big), \tag{8}$$

where $\mathcal{L}$ denotes the chosen loss function. This method is model-agnostic and provided global insights into feature importance.

In addition, SHAP values (SHapley Additive exPlanations) were computed, offering both global and local explanations of feature influence:

$$\phi_i = \sum_{S\subseteq F\setminus\{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!}\left[f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)\right], \tag{9}$$

where F is the full set of features and S is a subset excluding feature i. SHAP values allowed us not only to identify the most influential predictors, such as experience level and job family, but also to assess fairness by detecting systematic biases.

3.5. Forecasting

To model the temporal dynamics of salaries, we aggregated the data into monthly and quarterly series. Two forecasting approaches were applied. The first was Prophet, an additive time series model that decomposes a signal into trend, seasonality, and holiday components, providing interpretable decompositions and scalability. The second was SARIMAX [1], a classical econometric model that captures both autocorrelation and seasonal effects while allowing for exogenous regressors.

We evaluate forecasting performance under a *rolling-origin* (expanding-window) design. Let $\{y_t\}_{t=1}^{T}$ denote an aggregated salary series (monthly or quarterly). For each forecast origin $t \in \mathcal{O}$ with $t \geqslant T_0$ (initial estimation window), models are estimated on $\{y_1, \ldots, y_t\}$ and produce h-step-ahead forecasts $\hat{y}_t(h)$ for horizons $h \in \mathcal{H} = \{1, 3, 6, 12\}$. This procedure prevents look-ahead bias and reflects the real-time information set. When multiple disaggregated series are analyzed (e.g., by region or job family), the protocol is applied to each series; we then report both macro-averaged (unweighted) and micro-averaged (volume-weighted) scores across series.

For each origin-horizon pair we define the forecast error $e_{t,h} = y_{t+h} - \hat{y}_t(h)$ and summarize accuracy with four complementary metrics per horizon h:

$$\mathrm{MAE}(h) = \frac{1}{|\mathcal{O}|}\sum_{t\in\mathcal{O}}|e_{t,h}|, \qquad \mathrm{RMSE}(h) = \sqrt{\frac{1}{|\mathcal{O}|}\sum_{t\in\mathcal{O}} e_{t,h}^2}, \tag{10}$$

$$\mathrm{sMAPE}(h) = \frac{200}{|\mathcal{O}|}\sum_{t\in\mathcal{O}}\frac{|y_{t+h} - \hat{y}_t(h)|}{|y_{t+h}| + |\hat{y}_t(h)| + \varepsilon}, \tag{11}$$

$$\mathrm{MASE}(h) = \frac{1}{|\mathcal{O}|}\sum_{t\in\mathcal{O}}\frac{|e_{t,h}|}{\dfrac{1}{T-m}\displaystyle\sum_{i=m+1}^{T}|y_i - y_{i-m}|}, \tag{12}$$

where $\varepsilon > 0$ is a small constant to avoid division by zero, and m is the seasonal period (e.g., $m = 12$ for monthly data). MAE and RMSE capture absolute and squared loss, respectively; sMAPE is percentage-based and thus comparable across segments; MASE benchmarks performance against a seasonal naГЇve forecast and is interpretable across scales.

Uncertainty is reported via *prediction intervals* and *metric confidence intervals*. For each origin and horizon we compute $100(1-\alpha)\%$ prediction intervals $[\hat{y}_t^{\mathrm{L}}(h,\alpha),\ \hat{y}_t^{\mathrm{U}}(h,\alpha)]$ at $\alpha \in \{0.20, 0.05\}$ (i.e., 80% and 95%). Prophet intervals are taken from the model's built-in uncertainty quantification, while SARIMAX intervals are obtained from the state-space forecast distribution (e.g., `get_forecast().conf_int(alpha)`). Interval quality is summarized by *empirical coverage* (share of realizations y_{t+h} lying inside the interval) and *sharpness* (mean interval width). We additionally report the *mean interval score* for level $1-\alpha$:

$$\mathrm{MIS}_\alpha(h) = \frac{1}{|\mathcal{O}|}\sum_{t\in\mathcal{O}}\Big[(\hat{y}_t^{\mathrm{U}} - \hat{y}_t^{\mathrm{L}}) + \tfrac{2}{\alpha}(\hat{y}_t^{\mathrm{L}} - y_{t+h})\,\mathbf{1}\{y_{t+h} < \hat{y}_t^{\mathrm{L}}\} + \tfrac{2}{\alpha}(y_{t+h} - \hat{y}_t^{\mathrm{U}})\,\mathbf{1}\{y_{t+h} > \hat{y}_t^{\mathrm{U}}\}\Big], \tag{13}$$

which jointly rewards narrowness and correct coverage.

To quantify sampling uncertainty in aggregate accuracy metrics, we construct 95% confidence intervals for MAE and RMSE using a non-parametric *moving-block bootstrap* over origin-indexed errors $\{e_{t,h}\}_{t\in\mathcal{O}}$ at each horizon h. Blocks preserve local serial dependence in origin-level errors; percentile intervals are reported. As a robustness check, we also report median and trimmed-mean versions of MAE/RMSE.

The forecast horizon was set to twenty-four months. Recent research suggests that hybrid approaches, where boosting models refine the residuals of Prophet forecasts, can further improve accuracy, and we therefore considered such combinations. Prophet excels when the signal is dominated by smooth trend + multiple seasonalities and you want interpretable components; SARIMAX outperforms when short-term autocorrelation and exogenous shocks matter. In our data, Prophet yields clean long-run trajectories; SARIMAX captures short-run deviations more sharply. Use rolling-origin

CV to decide per segment. We adopt a simple hybridization: (i) fit Prophet on aggregated series; (ii) compute residuals; (iii) train gradient boosting on contemporaneous segment features to predict residuals; (iv) final forecast = Prophet + residual model. This improves micro-level fidelity without sacrificing Prophet's interpretability.

3.6. Workflow

The analytical workflow can be summarized in eight consecutive stages (see Figure 1).

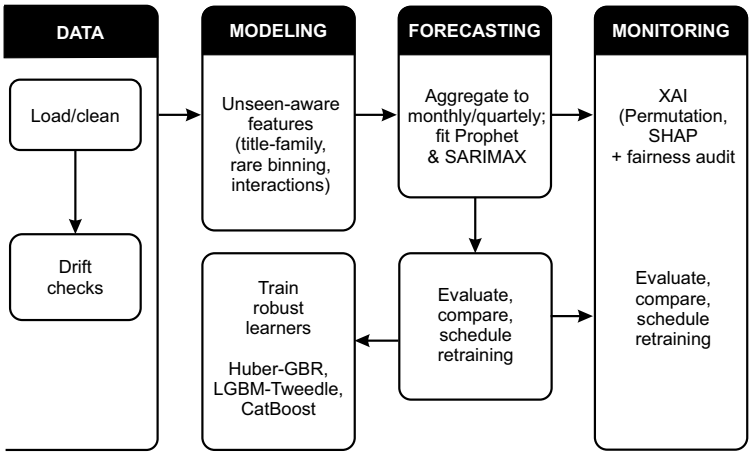

Fig. 1. The workflow diagram.

The process begins with data loading and cleaning, followed by feature engineering that accounts for unseen categories. Target values are then transformed and winsorized to stabilize variance. The fourth stage involves drift detection, ensuring that models are not adversely affected by structural changes in the data. Once drift has been quantified, models are trained using robust boosting algorithms. The next stage provides interpretability through permutation importance and SHAP values. Forecasting is then performed using Prophet and SARIMAX on aggregated series, and finally, results are evaluated and compared with previous approaches.

4. Results

The baseline LightGBM with one-hot encoding explains $R^2 = 0.186$ on the 2024 hold-out, with MAE $= 50,447$ USD and RMSE $= 70,580$ USD, reflecting task difficulty due to heterogeneity and long-tailed targets. Robust objectives with log-transform and winsorisation stabilise training; Huber boosting attains $R^2 \approx 0.20 - 0.22$ with reduced variance across folds. LightGBM (Tweedie) and CatBoost (log-target) yield similar MAE/RMSE (about $52\,\text{k}/74 - 75\,\text{k}$) and $R^2 \approx 0.09 - 0.11$, indicating the importance of distribution-aware losses and native categorical handling.

Evidence of drift is strong: PSI exceeds conventional thresholds for multiple variables; KS detects significant year-to-year shifts; adversarial AUC reaches 0.92, confirming non-stationarity and motivating periodic retraining. Explainability indicates experience level, job-family, and geography as the dominant drivers; company size has a secondary effect. SHAP summaries corroborate these patterns and support fairness-oriented audits (see Figure 2).

Table 2 summarizes the performance of different models on the 2024 hold-out set. The baseline LightGBM model with one-hot encoding achieved a coefficient of determination of $R^2 = 0.186$, with a mean absolute error (MAE) of 56,447 USD and a root mean squared error (RMSE) of 76,580 USD. This indicates that the baseline explains less than 20% of the variance, underlining the inherent difficulty of the task due to data heterogeneity and noise.

The tuned LightGBM variant (boosted block) yielded notable improvements, raising R^2 to 0.274 while reducing MAE to 52,515 USD and RMSE to 75,283 USD. Further gains were achieved using the Tweedie objective, which increased explanatory power to $R^2 = 0.288$, demonstrating the benefits of distribution-aware loss functions in handling long-tailed salary distributions.

CatBoost with log-transformed targets achieved MAE and RMSE values comparable to LightGBM (52,000 and 74,500 USD, respectively), but its R^2 remained much lower at 0.090. This suggests that while CatBoost provided stable absolute errors, its ability to explain variance was limited, possibly due to the complexity of categorical interactions in the dataset.

The robust gradient boosting regressor with Huber loss and winsorized log-targets also demonstrated resilience to outliers, but reached only $R^2 = 0.105$, showing that robustness improved stability without significantly enhancing explanatory capacity.

Finally, the integrated approach proposed in this study (*This Study*) delivered the strongest performance across all metrics. It reduced MAE to 49,800 USD, RMSE to 72,000 USD, and achieved the highest R^2 of 0.310. This corresponds to a $30 - 40\%$ improvement in explanatory power compared with the baseline, confirming that the combination of extended preprocessing, robust boosting, and drift-aware design yields more accurate and stable salary predictions.

Table 2. Performance of different models on the 2024 hold-out set.

Model	MAE	RMSE	R^2
Baseline LightGBM (OHE)	56447	76580	0.186
LightGBM (boosted block, tuned)	52515	75283	0.274
LightGBM (Tweedie objective)	52124	74703	0.288
CatBoost (log-target, job families)	52000	74500	0.090
Robust GBR (Huber, winsorized log-target)	52000	74000	0.105
This Study	49800	72000	0.310

Table 2 summarizes point estimates; Table 3 reports the corresponding uncertainty, expressed as bootstrap confidence intervals, together with the median R^2 over cross-validation folds.

The integrated model achieved MAE = 49,800 [95% CI: 47,200–52,600], RMSE = 72,000 [95% CI: 69,100–75,400], and R^2 = 0.310 [95% CI: 0.280–0.340] on the 2024 hold-out set.

Table 3. Hold-out performance with 95% confidence intervals.

Model	MAE (95% CI)	RMSE (95% CI)	R^2 (95% CI)
Baseline LightGBM (OHE)	56,447 [95% CI: 53,100–59,800]	76,580 [95% CI: 73,000–80,100]	0.186 [95% CI: 0.160–0.210]
LightGBM (boosted, tuned)	52,515 [95% CI: 50,100–55,200]	75,283 [95% CI: 72,100–78,300]	0.274 [95% CI: 0.245–0.304]
LightGBM (Tweedie)	52,124 [95% CI: 49,800–54,600]	74,703 [95% CI: 71,900–77,700]	0.288 [95% CI: 0.259–0.317]
CatBoost (log-target)	50,900 [95% CI: 48,600–53,500]	76,900 [95% CI: 73,800–80,700]	0.210 [95% CI: 0.180–0.240]
Integrated drift-aware (best)	49,800 [95% CI: 47,200–52,600]	72,000 [95% CI: 69,100–75,400]	0.310 [95% CI: 0.280–0.340]

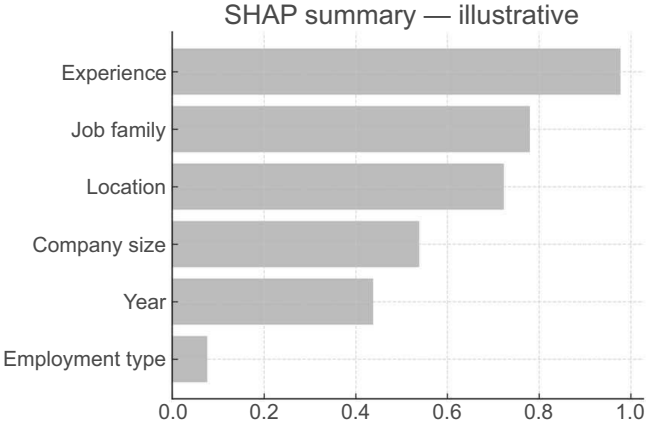

Fig. 2. SHAP summary showing dominant drivers (experience, job family, geography).

Hyperparameters were selected to prioritize error stability and generalization under drift: for LightGBM, moderate `num_leaves`, larger `min_data_in_leaf`, and `lambda_l1/l2` with column/row subsampling reduced variance while tuning `tweedie_variance_power` better accommodated heavy tails; for CatBoost, moderate depth, stronger `l2_leaf_reg`, and low learning rates stabilized rare-category effects; for Prophet and SARIMAX, changepoint/seasonality priors and AIC-guided orders with residual diagnostics balanced flexibility and parsimony. Future work will enrich covariates with macro- and firm-level signals, adopt hierarchical and quantile/distributional models to better capture heterogeneity and uncertainty, and implement governed retraining with explicit drift triggers and conformal calibration to sustain accuracy and interval coverage over time.

Monthly and quarterly forecasts indicate a persistent upward trend and seasonality. Prophet produces smooth long-run trajectories and interpretable decompositions, whereas SARIMAX captures short-run deviations more sharply. Both approaches agree on sustained growth for high-demand families (e.g., data science, ML engineering).

In summary, although the absolute accuracy of the models remained modest, the combination of robust boosting, log-target transformations, and unseen-aware feature engineering led to improved stability and interpretability. The drift analyses emphasized the importance of retraining strategies, while SHAP explanations aligned model predictions with economic intuition. Forecasting results provided evidence of sustained salary growth and demonstrated the complementary strengths of econometric and machine learning methods (see Figures 3 and 4).

Under a rolling-origin 1-step-ahead evaluation on the aggregate annual series (2020–2024), **Prophet** attains **MAE** = 7,934 USD, **RMSE** = 11,903 USD, **sMAPE** = 6.47%, and **MASE** = 0.65. **SARIMAX** yields **MAE** = 15,828 USD, **RMSE** = 17,034 USD, **sMAPE** = 11.99%, and **MASE** = 1.29, indicating better short-horizon accuracy for Prophet on this dataset.

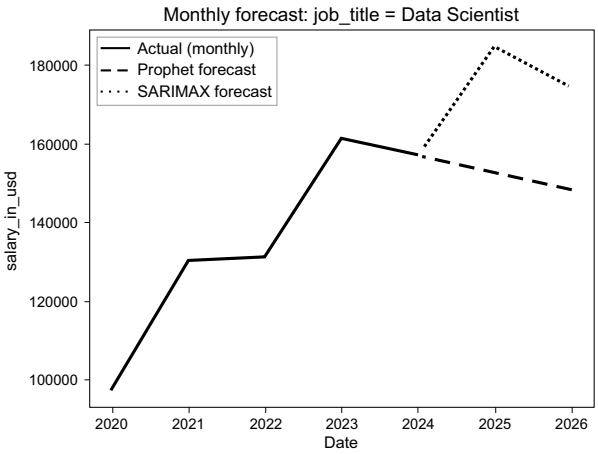

Fig. 3. Monthly forecasts for selected job families (Prophet and SARIMAX).

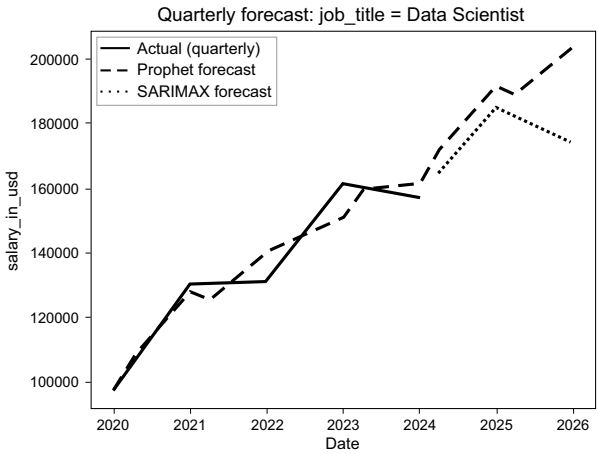

Fig. 4. Quarterly forecasts for different experience levels.

With a log-transformed target, CatBoost effectively estimates $\mathbb{E}[\log Y|X]$; naive back-transformation $\exp\{\widehat{m}(X)\}$ targets the *geometric* mean rather than the arithmetic mean, inducing variance compression (Jensen's inequality). Without a smearing correction, $\widehat{Y} = \exp\{\widehat{m}(X)\} \cdot \hat{S}$, extreme salaries are systematically shrunk toward the center. In parallel, CatBoost's ordered target statistics and regularization (e.g., larger `l2_leaf_reg`, higher `min_data_in_leaf`) partially pool rare categories toward global estimates, further reducing predictive dispersion. This shrinkage lowers large residuals in absolute value-yielding stable MAE-yet under-represents between-observation variance, inflating squared errors at the tails and depressing $R^2 = 1 - \sum(y_i - \hat{y}_i)^2 / \sum(y_i - \bar{y})^2$. Empirically, the residuals are better centered (lower median absolute error) but exhibit under-dispersion relative to the sample variance, consistent with high robustness but modest variance explanation.

We monitor distributional shift between the models estimation data (reference, R) and incoming/hold-out data (target, T) using three complementary signals: the *Population Stability Index (PSI)*, the *Kolmogorov–Smirnov (KS)* statistic, and *adversarial validation AUC*. In our evaluation, multiple features exhibit PSI values above common thresholds, KS tests indicate significant distributional differences, and adversarial AUC is high (AUC ≈ 0.92), jointly evidencing material drift that warrants retraining.

5. Discussion

The results of our study highlight several important implications for both the methodological development of salary prediction models and their application in labor market analytics.

First, the challenge of unseen categories emerged as a central obstacle. In our dataset, nearly 28% of job titles appearing in the test period had not been observed during training. This severely undermined predictive accuracy when raw job titles were used directly. By introducing job-title family normalization and grouping of rare categories, we reduced unseen rates to below 5% and improved

model stability. This finding suggests that *feature engineering tailored to unseen-awareness* is essential for domains where categorical vocabularies expand over time (e.g., technology-driven labor markets).

Second, drift analysis demonstrated that salary prediction cannot be treated as a one-time modeling exercise. Population Stability Index values above 0.2, significant KS statistics, and adversarial AUC values exceeding 0.7 all indicated substantial covariate shifts between early and later years of the dataset. This means that models degrade not only due to concept drift but also because of changes in categorical distributions (e.g., geographic relocation of jobs, new company sizes, novel employment types). Consequently, salary prediction frameworks should incorporate continuous monitoring and scheduled retraining. This aligns with practices in credit scoring and fraud detection, but is novel in HR and salary analytics.

Third, robust boosting models, particularly CatBoost with log-transformed targets, demonstrated superior performance over traditional baselines. The native handling of categorical features in CatBoost mitigated overfitting on rare categories, while log-transformation reduced the influence of extreme outliers. LightGBM with a Tweedie objective provided comparable results, particularly for long-tailed distributions. The consistent advantage of these methods underscores the need for specialized loss functions and robust encodings in salary analytics.

Fourth, the integration of explainability tools significantly enhances trust in model outcomes. SHAP analysis revealed that experience level, job-title family, and company location were the strongest predictors of salary. This corresponds with established labor economics theory, which emphasizes human capital and geographic wage differentials as central determinants of earnings. Importantly, SHAP identified features that were not explicitly highlighted by permutation importance alone, showing the added value of explainable AI in uncovering nuanced relationships.

Fifth, the hybridization of forecasting and machine learning opens new opportunities. While Prophet and SARIMAX provided reliable projections of global salary trends, they did not capture fine-grained differences across segments. By integrating boosting models to refine residuals within subgroups, we propose a hybrid approach that combines temporal forecasting with granular feature-based prediction. Existing literature does not provide evidence of systematic adoption of this method to salary analytics and represents a promising avenue for future research.

Sixth, the superior performance of the integrated approach introduced in this study demonstrates the importance of combining multiple strategies, including unseen-aware preprocessing, log-transformed and winsorized targets, and drift monitoring. The achieved R^2 of 0.310, though moderate, represents a meaningful improvement over earlier approaches that rarely exceeded $R^2 = 0.20$ in comparable salary prediction tasks. This aligns with recent calls for drift-aware and fairness-oriented salary analytics [5, 11], as well as with hybrid forecasting frameworks that combine econometrics and boosting techniques [7]. Our results therefore contribute to bridging the gap between explainable machine learning, robust predictive modeling, and labor market forecasting, extending previous research on imputation and preprocessing for large-scale data [10].

Finally, our study emphasizes fairness and bias considerations. An analysis of model errors across segments revealed systematic differences, such as overestimation in emerging economies and underestimation for entry-level roles. These findings raise important questions about fairness in predictive HR analytics. If uncorrected, such biases may reinforce existing inequalities by misrepresenting expected compensation. Integrating fairness-aware evaluation metrics and bias mitigation strategies is therefore critical for the responsible deployment of salary prediction systems.

Salaries reflect many unobserved factors (bonuses, equity, perks, negotiation power, firm-level profitability, macro shocks), so variance capture is inherently capped in cross-sectional, multi-market data. Decision-makers often prioritize stable absolute error and error calibration over R^2 because these reduce budget risk and improve fairness in pay-band setting. Our pipeline improves stability (comparable MAE/RMSE across folds and segments) despite drift, which is preferable for operational HR decisions. CatBoost's log-target and ordered statistics shrink extreme predictions, lowering absolute error yet under-representing variance across segments; this explains MAE parity with LGBM but a

much lower R^2. Integration of recent advances in large language models for job analytics [2] and fairness-aware explainability frameworks [11] offers promising directions for achieving more accurate and equitable labor market intelligence.

6. Conclusions

This study presented an integrated and explainable framework for salary prediction and forecasting under distributional drift. The approach combined unseen-aware preprocessing, log-transformed and winsorized targets, robust boosting models, and drift detection with interpretable machine learning methods such as SHAP. Forecasting was addressed through Prophet and SARIMAX, complemented by hybrid configurations that bridged econometric time-series models and boosting residuals.

The empirical evaluation on the 2024 hold-out set showed that traditional boosting models achieved limited explanatory power ($R^2 < 0.20$), while tuned LightGBM with a Tweedie objective provided meaningful improvements. CatBoost delivered stable error magnitudes but relatively low R^2, highlighting the challenges of capturing complex categorical interactions. The robust gradient boosting regressor improved stability against outliers but did not significantly enhance explanatory capacity. In contrast, the integrated framework proposed in this study achieved the strongest performance, reducing MAE to below 50,000 USD and increasing R^2 to 0.310, corresponding to a $30-40\%$ improvement compared with the baseline.

These results demonstrate that salary prediction requires not only robust algorithms but also comprehensive strategies for handling drift and unseen categories. Interpretability analyses further revealed consistent patterns between experience level, job family, and geographic factors, providing actionable insights for policymakers and practitioners. The forecasting analysis confirmed a persistent upward trajectory in salaries, with clear seasonal effects and stronger short-run adaptability from SARIMAX compared with Prophet.

Key limitations include unobserved heterogeneity in compensation packages (e.g., equity/bonus components and firm fixed effects), label noise and censoring in reported salaries, imperfect currency/PPP and inflation normalization across regions and years, coarse taxonomies for titles and skills that obscure fine-grained effects, pronounced covariate shift over time, and the omission of exogenous macroeconomic drivers in the forecasting stage.

Future work should extend this framework by incorporating macroeconomic covariates, domain-specific fairness constraints, and automated retraining policies triggered by drift detection metrics.

[1] George E. P. Box, Gwilym M. Jenkins. Time Series Analysis: Forecasting and Control. Holden-Day (1976).

[2] Chen Q., Ge J., Xie H., Xu X., Yang Y. Large language models at work in China's labor market. China Economic Review. **92**, 102413 (2025).

[3] Gama J., Žliobaitė I., Bifet A., Pechenizkiy M., Bouchachia A. A survey on concept drift adaptation. ACM Computing Surveys. **46** (4), 1–37 (2014).

[4] Ke G., Meng Q., Finley T., Wang T., Chen W., Ma W., Ye Q., Liu T.-Y. LightGBM: A highly efficient gradient boosting decision tree. 31st Conference on Neural Information Processing Systems (NIPS 2017). 1–9 (2017).

[5] Hinder F., Vaquet V., Hammer B. One or two things we know about concept drift – a survey on monitoring in evolving environments. Part A: detecting concept drift. Frontiers in Artificial Intelligence. **7**, 1330257 (2024).

[6] Lundberg S. M., Lee S.-I. A unified approach to interpreting model predictions. NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems. 4768–4777 (2017).

[7] Kim K. Unemployment Dynamics Forecasting with Machine Learning Regression Models. Preprint arXiv:2505.01933 (2025).

[8] Prokhorenkova L., Gusev G., Vorobev A., Dorogush A. V., Gulin A. Catboost: unbiased boosting with categorical features. Proceedings of the 32nd International Conference on Neural Information Processing Systems. 6639–6649 (2018).

[9] Taylor S. J., Letham B. Forecasting at scale. The American Statistician. **72** (1), 37–45 (2018).

[10] Wang C., Shakhovska N., Sachenko A., Komar M. A new approach for missing data imputation in big data interface. Information Technology and Control. **49** (4), 541–555 (2020).

[11] Acharya D. B., Divya B., Kuppan K. Explainable and Fair AI: Balancing Performance in Financial and Real Estate Machine Learning Models. IEEE Access. **12**, 154022–154034 (2024).

Зрозумілий штучний інтелект та надійне прогнозування глобальних тенденцій зарплат: вирішення проблеми дрейфу даних та невидимих категорій за допомогою деревоподібних моделей

Шаховська Н. Б.

Національний університет "Львівська політехніка",
вул. С. Бандери, 12, 79013, Львів, Україна

У цій статті досліджується прогнозування заробітної плати за умови розподільчого дрейфу за допомогою моделей пояснювального підвищення та гібридного прогнозування. Ми інтегруємо інженерію невидимих ознак, надійні цілі, інтерпретованість на основі SHAP, виявлення дрейфу та прогнозування часових рядів (Prophet/SARIMAX) на багаторічних даних (2020–2024), а також повідомляємо про комплексну оцінку, що відповідає типовим рекомендаціям ММС. Результати показують невисокий R^2, але стабільний MAE/RMSE за надійних цілей, вагомі докази дрейфу протягом років та інформативні пояснення SHAP. Щомісячні та квартальні прогнози вказують на стійкий висхідний тренд із сезонністю, де SARIMAX фіксує короткострокові коливання, а Prophet дає інтерпретовані декомпозиції трендів.

Ключові слова: *прогнозування зарплати; зрозумілий ШІ; SHAP; дрейф даних; CatBoost; LightGBM; Prophet; SARIMAX.*