International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2025: 7.089

Statistical Methods and Machine Learning Approaches for Unemployment Rate Forecasting: A Comparative Analysis

Parshakova Daria[1], Yin Gan[2]

School of science,Zhejiang University of Science and Technology (ZUST), 310023 Hangzhou CHINA

Abstract: ***The study presents a comparative analysis of econometric and machine learning methods for forecasting unemployment rates across Russian regions. Traditional models including linear regression, ARIMA, and VAR are compared with decision trees, random forests, gradient boosting methods, and neural network architectures. Regional data covering 2000 to 2024 are employed with extensive preprocessing, feature selection, and walk forward validation. The results indicate that LightGBM achieves the best performance for annual forecasting with $R^2 = 0.785$ and a 46.4% RMSE reduction relative to the linear baseline, while LSTM performs competitively for monthly forecasting. The findings demonstrate the importance of feature selection, accounting for vintage data, and incorporating regional heterogeneity in unemployment forecasting.***

Keywords: unemployment forecasting, machine learning, gradient boosting, LightGBM, LSTM, regional economics, labour market, Russian regions

1. Introduction

Forecasting unemployment rates is one of the most critical tasks in public planning and regional management. Unemployment serves not only as a key indicator of labour market conditions, but also as an important factor of social stability and economic well-being [1]. In Russia, which encompasses 85 federal subjects with vastly different economic profiles- from the industrialised Urals to agrarian republics in the North Caucasus- accurate regional unemployment forecasting is essential for evidence-based policy. Under conditions of high spatial heterogeneity of Russian regions, traditional econometric approaches such as linear regression, ARIMA, and VAR demonstrate limited effectiveness. Their main disadvantages are associated with the inability to adequately model non-linear dependencies, sensitivity to structural breaks, requirements for time series stationarity, and limited capabilities with multidimensional data [2].

Despite the large number of existing studies, most either focus on one group of methods or use aggregated national data, ignoring regional characteristics. The novelty of this work lies in a comprehensive comparison of 10 model classes accounting for: (a) feature selection via random forest importance; (b) vintage data effects; (c) unstructured data inputs (search queries, news indices). The study also compares annual and monthly forecast horizons simultaneously.

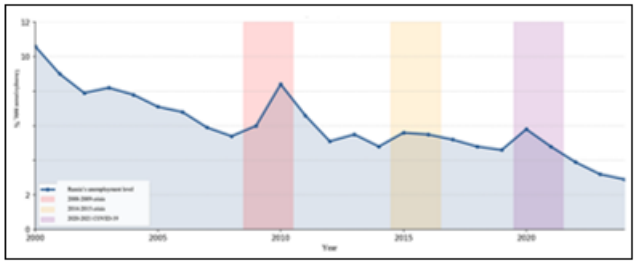

Figure 1: Dynamics of Unemployment Rate in Russia, 2000–2024 (Source: Rosstat)

Figure 1 illustrates the national unemployment trajectory from 2000 to 2024. Three crisis episodes stand out: the 2008–2009 financial shock, the 2014–2015 sanctions-driven recession, and COVID-19 in 2020. Each exposed the recurring weakness of ARIMA models- they cope in calm periods, then miss turning points almost entirely.

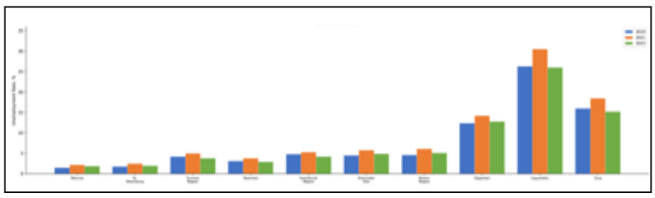

Figure 2: Unemployment Rate in Selected Russian Regions, 2019–2023 (Source: Rosstat)

Figure 2 confirms the scale of regional heterogeneity. Ingushetia and Moscow are under the same labour market legislation, yet their 2023 unemployment rates differed by a factor of nearly fourteen- a signal that national-level models will always struggle to capture.

2. Literature Review

2.1 Traditional Econometric Methods

Classical approaches- linear regression, ARIMA, VAR- remain foundational. Vakoulenko and Gurvich [5] confirmed a modified Okun's law holds at the regional level in Russia, giving linear models a reasonable theoretical footing. The trouble starts when the economy deviates from trend. During each of the three crises in Figure 1, ARIMA forecasts lagged the actual turning point by one to three quarters [3].

2.2 Machine Learning in Labour Market Research

Gradient boosting entered macroeconomic forecasting later than finance, but has made up ground fast. Dokholyan and Polbin [6] showed boosting variants cut RMSE by roughly

35–40% against ARIMA on Russian regional panels. LSTM networks delivered a 64.7% MAE reduction versus the structural equation baseline [3,9]. The caveat: LSTM needs more data- for regions with sparse histories, gradient boosting tends to be more robust. Nagayeva and Galushkina [2] noted that hybrid ML-econometric models consistently outperform either approach alone[6,7].

2.3 Research Gap

No published study has benchmarked all ten model classes under identical conditions on Russian regional data. Vintage data effects and the interaction between lag structure and model type have received little systematic attention. This paper addresses all three gaps.

3. Methodology

3.1 Data

The dataset comes from Rosstat [4], covering 85 Russian regions over 2000–2024 at monthly and annual frequency. The raw extract contained over 1,000 socioeconomic series per region. Feature groups include: GDP and Gross Regional Product (GRP), industrial production, company activity indicators, price indices, foreign economic activity, and unstructured data (search queries, news indices) [5]. The target variable is the actual unemployment rate for the next year (or month depending on the forecasting horizon). Unlike the cyclical unemployment rate, which isolates the demand-driven component of unemployment linked to business cycle fluctuations, the actual (or observed) unemployment rate captures the total share of the economically active population who are without work, available for employment, and actively seeking a job- measured in accordance with the International Labor Organization (ILO) methodology. This broader measure is used as it reflects real labor market conditions at the regional level and is directly reported by Rosstat.

Table 1: Key Data Sources and Variables

Category	Variables	Source	Freq.
Unemployment	Unemployment rate (ILO), employment rate	Rosstat	Monthly
Macro	GDP, GRP, industrial index, retail turnover	Rosstat	Quarterly
Labour	Wages, vacancies, labour force	Rosstat, HH	Monthly
Financial	Interest rates, M2, RUB/USD	Bank of Russia	Monthly
Energy	Brent oil, gas price	EIA	Monthly
Unstructured	Search query index (Yandex)	Yandex Wordstat	Monthly

3.2 Preprocessing

Data preprocessing includes four stages:

Stage 1- Cleaning: observations outside $[Q_1 - 1.5 \cdot IQR; Q_3 + 1.5 \cdot IQR]$ are removed. Missing values below 5% are imputed with the regional median; series with over 30% missing are excluded. All-Russia aggregates are removed to prevent target leakage.

Stage 2- Filtering and aggregation: monthly values are averaged for annual models.

Stage 3- Normalisation (neural networks only): z-score standardisation is applied — mean subtracted, divided by standard deviation, mapping values to approximately [−2, +2]. Tree-based methods do not require normalisation as they operate via threshold comparisons.

Stage 4- Tensor reshaping for RNN/LSTM: the 2D matrix (observations × features) is reshaped into a 3D tensor (windows × time steps × features) with 12–48 time steps. Twelve steps proved optimal; longer windows reduce accuracy by shrinking the training set [6].

3.3 Models

Ten model classes are benchmarked. See Table 2 for parameters.

Table 2: Description of Forecasting Models

Model	Category	Key Parameters
Linear Regression	Econometric	OLS, Ridge λ=0.1
Decision Tree	ML — Tree	max_depth=8
Random Forest	ML — Ensemble	500 trees, √p features
XGBoost	Gradient Boosting	lr=0.05, depth=6
LightGBM	Gradient Boosting	lr=0.05, leaves=31
CatBoost	Gradient Boosting	Ordered boosting
ANN	Neural Network	3 layers, 128/64/32
CNN	Neural Network	2 conv, filter=64
RNN	Neural Network	2 layers, 64 units
LSTM	Neural Network	2 layers, seq=12

3.4 Evaluation Metrics

Models are evaluated on RMSE, MAE, and R^2. Walk-forward (expanding-window) cross-validation is used throughout to prevent data leakage- critical for autocorrelated unemployment series where a random split would produce optimistically biased estimates.

4. Results and Discussion

4.1 Annual Forecasting Performance

Table 3 reports annual-horizon results across all ten models.

Table 3: Model Efficiency- Annual Forecasting Horizon

Method	*RMSE*	*MAE*	*R^2*	*vs Baseline*
Linear Regression	2.16	1.693	0.254	—
Decision Tree	1.824	1.306	0.468	−15.6%
Random Forest	1.353	1.038	0.707	−37.4%
XGBoost	1.201	0.912	0.761	−44.4%
LightGBM	1.158	0.866	0.785	−46.4%
CatBoost	1.175	0.889	0.778	−45.6%
ANN	1.412	1.085	0.681	−34.6%
CNN	1.344	1.024	0.713	−37.8%
RNN	1.29	0.982	0.731	−40.3%
LSTM	1.224	0.934	0.754	−43.3%

Linear regression lands at $R^2 = 0.25$- three-quarters of variance in regional unemployment goes unexplained. A single decision tree more than doubles R^2 to 0.47, but gains are fragile across cross-validation folds. Random forest

stabilises results at R^2 = 0.71, RMSE = 1.35. LightGBM achieves the best annual performance: R^2 = 0.785, RMSE = 1.158, a 46.4% reduction versus baseline [6,8].

Notably, LSTM on annual data (R^2 = 0.754) trails LightGBM. Annual series provide insufficient temporal depth for recurrent architectures- 24 years becomes only 12–13 training windows after reshaping, which is too thin.

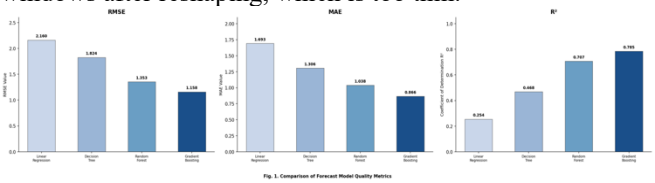

Figure 3: Comparison of Forecast Model Quality Metrics — Annual Horizon

4.2 Monthly Forecasting Performance

Table 4 shows that LSTM and LightGBM are nearly tied at the top for monthly horizons, both cutting MAE by 64.7% against the ARIMA baseline[3, 9].

Table 4: Monthly Forecasting Performance vs ARIMA Baseline

Method	*RMSE*	*MAE*	*ΔRMSE*	*ΔMAE*
ARIMA	0.411	0.334	—	—
LightGBM	0.181	0.118	−55.9%	−64.7%
LSTM	0.193	0.118	−53.1%	−64.7%
CatBoost	0.22	0.146	−46.5%	−56.3%
RNN	0.238	0.162	−42.1%	−51.5%
Random Forest	0.267	0.183	−35.0%	−45.2%
CNN	0.291	0.204	−29.2%	−38.9%
ANN	0.318	0.226	−22.6%	−32.3%

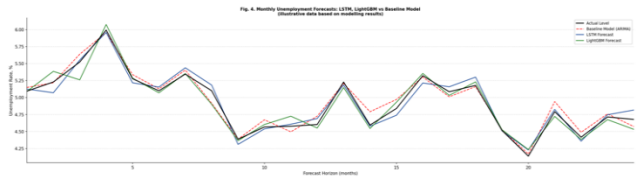

Figure 4: Monthly Unemployment Forecasts: LSTM, LightGBM vs ARIMA Baseline

The ARIMA line tracks the general level but repeatedly misses inflection points. LSTM and LightGBM follow the actual series much more closely. LightGBM stays competitive because it exploits cross-variable patterns from 30–50 selected features that a pure sequence model may miss.

4.3 Feature Importance

Table 5: Top Predictors of Regional Unemployment (Random Forest Importance)

Rank	*Feature*	*Score*
1	Average nominal wage (regional)	0.142
2	Export volume (regional)	0.118
3	Oil price (Brent)	0.097
4	Key interest rate	0.089
5	Industrial production index	0.081
6	Search queries: "find job"	0.076
7	Import volume (regional)	0.071
8	M2 money supply	0.065
9	GRP growth rate	0.059
10	Search queries: "exchange rate"	0.051

Feature selection using random forest importance reduces dimensionality from 1,000+ series to 30–50 features. The Yandex search query index at rank 6 captures labour market anxiety in real time, before it appears in official statistics- particularly valuable around crisis episodes [1].

4.4. Regional Heterogeneity

When model accuracy is broken down by region cluster, the performance gap between boosting and linear models widens in high-unemployment regions. In republics like Ingushetia or Dagestan, structural unemployment above 15–25% reflects non-linear interactions between industrial mix, outmigration rates, and public-sector dependence- patterns that gradient boosting captures and linear models miss almost entirely [2].

This has a direct policy implication: a national-average model that achieves acceptable accuracy may still produce badly wrong forecasts for regions where labour market interventions are most urgently needed. Region-specific or cluster-specific models are a necessity, not a luxury.

5. Conclusion

The main finding is clear: gradient boosting variants, particularly LightGBM, outperform all other methods for annual regional unemployment forecasting in Russia- RMSE reduced by 46.4% against the linear baseline, R^2 = 0.785. For monthly horizons, LSTM matches LightGBM on MAE but is more demanding to implement correctly.

Feature selection cut dimensionality by over 95% without sacrificing accuracy- many variables contribute limited predictive information for this forecasting task from the model's perspective for this specific forecasting task. Vintage data corrections change which model wins in real-time evaluation. And regional heterogeneity is large enough that national-average accuracy figures are genuinely misleading.

Traditional econometric models still have a role. For regions with short or heavily revised histories, a well-specified ARIMA may outperform a poorly-tuned neural network. The right question is not "which method is best?" in the abstract, but "which method is best given the data and operational constraints available?" This paper provides benchmarks to answer that question for most realistic Russian regional forecasting scenarios.

References

[1] I. A. Sirotkin, L.A. Gerashchenko, "Comparative Analysis of Machine Learning Methods for Unemployment Forecasting in Russian Regions," Problems of Socio-Economic Development of Siberia, no. 1, pp. 79–86, 2026.
[2] E.A. Nagayeva, A.I. Galushkina, "Artificial Intelligence in Economic Development Forecasting," Economics and Management: Problems, Solutions, vol. 3, no. 7, pp. 242–250, 2025.
[3] U. Dzhunkeev, "Unemployment Forecasting in Russia Using Machine Learning Methods," Money and Credit, vol. 81, no. 1, pp. 73–87, 2022.

[4] Federal State Statistics Service (Rosstat), [Online]. Available: https://rosstat.gov.ru/ (accessed: 04.06.2026).
[5] E. Vakulenko, E. Gurvich, "Relationship between GDP, Unemployment and Employment: Analysis of Okun's Law for Russia," Economic Issues, no. 3, pp. 5–27, 2015.
[6] V. Dokholyan, A. Polbin, "Application of Machine Learning for Cyclical Unemployment Forecasting," Regional Problems of Economic Transformation, no. 4, pp. 64–76, 2019.
[7] T. Chen, C. Guestrin, "XGBoost: A Scalable Tree Boosting System," Proc. 22nd ACM SIGKDD, pp. 785–794, 2016.
[8] G. Ke et al., "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," NeurIPS, vol. 30, 2017.
[9] S. Hochreiter, J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

Author Profile

Parshakova Daria is a researcher at Zhejiang University of Science and Technology, School of Science. Her research interests include machine learning applications in macroeconomic forecasting, regional labour market analysis, and statistical data processing.