PressAcademia
www.pressacademia.org/journals/jbef
JBEF
ISSN 2146-7943
Journal of Business, Economics and Finance YEAR 2026 VOLUME 15 ISSUE 1

COMPARATIVE PREDICTIVE MODELLING OF TECHNOLOGY-INDUCED LABOUR MARKET DYNAMICS USING XGBOOST AND LIGHTGBM MODELS

DOI: 10.17261/Pressacademia.2026.2030
JBEF- V.15-ISS.1-2026(2)-p.16-25

Lawrence A. Farinola
Rauf Denktas University, Department of Software Engineering, Republic of Northern Cyprus.
Center of Excellence for Interdisciplinary AI and Data Science Research, Rauf Denktas University, Mersin 10 via Türkiye.
lawrence.farinola@rdu.edu.tr , ORCID: 0009-0004-5480-2137

Date Received: January 4, 2026 **Date Accepted:** April 23, 2026

To cite this document
Farinola L. A., (2026). Comparative predictive modelling of technology-induced labour market dynamics using XGBoost and LightGBM Models. Journal of Business, Economics and Finance (JBEF), 15(1), 16-25.
Permanent link to this document: http://doi.org/10.17261/Pressacademia.2026.2030

ABSTRACT

Purpose- Rapid advances in industrial automation, artificial intelligence, and digital production technologies are transforming labour market structures worldwide, intensifying concerns related to job displacement, occupational vulnerability, and regional inequality. This study aims to forecast technology-induced labour market dynamics using an interpretable and policy-relevant machine-learning framework.

Methodology- The study develops an interpretable predictive modelling framework based on a large-scale, harmonized panel dataset comprising 68,882 occupation–region–year observations spanning the period 2010–2023. The dataset integrates labour-force microdata, task-based automation risk indicators, occupational characteristics, and macroeconomic control variables across multiple economies. Two state-of-the-art gradient-boosting algorithms—XGBoost and LightGBM—are trained and evaluated using temporally consistent cross-validation. Model performance is assessed using Root Mean Squared Error (RMSE) and the coefficient of determination (R^2), while model interpretability is achieved through SHapley Additive exPlanations (SHAP).

Findings- Empirical results indicate that XGBoost substantially outperforms LightGBM, achieving a lower RMSE (2,304.76) and a higher R^2 (0.9325), compared to LightGBM's RMSE of 6,017.03 and R^2 of 0.5398. These results demonstrate XGBoost's superior ability to capture nonlinear relationships and heterogeneous automation effects across occupations and regions.

Conclusion- SHAP-based interpretability analysis identifies task repetitiveness, physical proximity, and cognitive complexity as the most influential drivers of automation-related labour market vulnerability. Scenario-based simulations further reveal that targeted policy interventions—such as reskilling programmes and workforce transition support—can significantly reduce projected job displacement, particularly among mid-risk occupations. Overall, the findings confirm that interpretable gradient-boosting models provide a robust and policy-relevant tool for forecasting automation-driven labour market dynamics and supporting evidence-based workforce planning in economies undergoing rapid technological transformation.

Keywords: Industrial automation, labour market dynamics, predictive modeling, XGBoost, LightGBM
JEL Codes: J24, O33, C45

1. INTRODUCTION

Rapid advances in industrial automation, artificial intelligence, and digital production technologies are fundamentally reshaping labour markets worldwide. While automation has the potential to enhance productivity, reduce production costs, and stimulate economic growth, it simultaneously raises concerns regarding job displacement, wage polarization, and widening regional inequalities. Policymakers therefore face an urgent need for reliable, data-driven tools capable of forecasting labour market outcomes under accelerating technological change.

Traditional labour economics approaches—often based on linear regressions or equilibrium models—have struggled to capture the nonlinear and heterogeneous nature of automation's impact across occupations, sectors, and regions. Recent evidence suggests that automation exposure varies substantially within industries and even within occupations, depending on task composition, skill intensity, and institutional context (Frey & Osborne, 2017; Autor & Salomons, 2018). These complexities motivate the use of machine-learning methods, which are better suited to modelling high-dimensional interactions and nonlinear dependencies in large-scale socio-economic data.

In response, this study develops a gradient-boosting-based predictive framework to forecast automation-related labour market outcomes using a large, harmonized, cross-national dataset comprising 68,882 occupation–region–year observations spanning 2010–2023. The analysis focuses exclusively on XGBoost and LightGBM, two state-of-the-art ensemble learning

algorithms that have demonstrated superior performance in structured economic forecasting tasks (Chen & Guestrin, 2016; Ke et al., 2017). By integrating occupational automation probabilities, detailed labour statistics, and macroeconomic controls, the proposed framework aims to generate accurate and interpretable predictions of employment exposure to automation.

Beyond predictive accuracy, interpretability is essential for policy relevance. To this end, the study employs SHapley Additive exPlanations (SHAP) to identify the key drivers of automation-related labour market vulnerability and resilience. This combination of high-performance machine learning and transparent explanation techniques enables evidence-based evaluation of alternative automation trajectories and policy interventions.

The primary contributions of this paper are threefold. First, it constructs a large-scale, harmonized labour market panel that integrates task-level automation risk with occupational, regional, and macroeconomic data. Second, it provides a focused empirical comparison of XGBoost and LightGBM in forecasting automation-driven labour market outcomes. Third, it delivers interpretable insights that support policymakers in designing targeted labour market and skills-development strategies. The methodological design underpinning these contributions is detailed in Section 3.

2. LITERATURE REVIEW

2.1. Automation and Labour Market Transformation

A growing body of literature documents the profound effects of automation on employment structures, task composition, and wage distributions. Early task-based frameworks argue that routine and codifiable tasks are particularly susceptible to automation, leading to job polarization and skill-biased technological change (Autor, Levy, & Murnane, 2003). Frey and Osborne (2017) extend this perspective by estimating automation probabilities for detailed occupations, highlighting substantial heterogeneity in technological exposure across the labour market. Subsequent studies confirm that automation effects are uneven across regions and sectors, often amplifying existing inequalities (Autor & Salomons, 2018; Acemoglu & Restrepo, 2020). Recent data-driven policy research further demonstrates that machine-learning-based forecasting frameworks can effectively quantify the socioeconomic consequences of industrial automation, enabling more granular assessments of labour market vulnerability and employment displacement (Farinola, Assogba, & Assogba, 2025).

More recent evidence deepens this understanding by emphasizing task reallocation and regional adjustment dynamics. Acemoglu and Restrepo (2022) show that automation-driven task substitution contributes significantly to wage inequality, while Bajgar et al. (2023) demonstrate that robot adoption reshapes job content rather than uniformly displacing employment. Using regional data from the United States and Europe, Bessen et al. (2023) find that employment growth responses to automation vary substantially across local labour markets, reflecting differences in industrial structure and absorptive capacity. Worker-level analyses further reveal that exposure to robots induces heterogeneous adjustment paths, including occupational mobility and wage effects, rather than simple job loss (Dauth et al., 2022).

Recent studies also highlight the blurred boundary between automation and augmentation. Autor et al. (2023) provide evidence from manufacturing that new technologies often complement human labour by altering task allocation within jobs, while Song et al. (2023) show that automation influences job security and labour market transitions, with policy institutions playing a moderating role. These findings suggest that the labour market consequences of technological change are complex, nonlinear, and context dependent, reinforcing the need for flexible empirical frameworks capable of capturing heterogeneous effects across occupations and regions.

2.2. Machine Learning in Labour Market and Economic Forecasting

Recent advances in machine learning have expanded the methodological toolkit available for labour market and economic analysis. Ensemble tree-based methods, such as Random Forests and gradient boosting, have demonstrated strong predictive performance in settings characterised by nonlinear relationships, high-dimensional feature spaces, and complex interaction effects (Varian, 2014). These properties are particularly relevant for labour market data, where employment outcomes are shaped by interactions among technological, occupational, regional, and macroeconomic factors.

Gradient-boosting models have gained particular prominence due to their flexibility and robustness. XGBoost, introduced by Chen and Guestrin (2016), has been widely adopted in economic forecasting and labour analytics owing to its regularisation mechanisms and computational efficiency. Recent applied studies demonstrate the robustness of gradient-boosting models such as XGBoost and LightGBM in complex, high-dimensional industrial systems, reinforcing their suitability for modelling nonlinear production and employment dynamics (Farinola & Bazarkhan, 2025). LightGBM, proposed by Ke et al. (2017) further improves scalability through histogram-based learning and leaf-wise tree growth, making it well-suited for large, structured panel datasets. Comparative studies in applied economics suggest that these models often outperform traditional econometric approaches when forecasting employment dynamics and wage outcomes (Athey & Imbens, 2019). Large-scale socioeconomic forecasting applications further confirm the capacity of machine-learning models to handle longitudinal national datasets and generate policy-relevant predictions under uncertainty (Farinola & Ayodeji, 2025).

Recent economic research increasingly recognizes the role of machine learning as a complement to structural and reduced-form analysis. Brynjolfsson et al. (2024) argue that machine-learning methods are especially valuable for identifying task-level exposure and productivity effects of new technologies, while Bajgar et al. (2023) highlight the usefulness of flexible models for capturing technology-induced task reallocation. These developments motivate the application of advanced gradient-boosting techniques for forecasting labour market dynamics under rapid technological change.

2.3. Interpretability and Policy-Oriented Machine Learning

Despite their predictive power, machine-learning models have historically faced criticism for limited interpretability, which constrains their usefulness in policy analysis. To address this concern, explainable artificial intelligence techniques have been developed to provide transparent and theoretically grounded explanations of model predictions. Among these, SHapley Additive exPlanations (SHAP) offer a unified framework for attributing feature importance based on cooperative game theory (Lundberg & Lee, 2017).

Recent studies demonstrate the growing relevance of interpretable machine learning in labour economics. Gu and Xiong (2024) apply SHAP-based methods to analyse employment transitions, showing how explainability enhances the identification of policy-relevant drivers of labour market vulnerability. Similarly, Autor et al. (2023) emphasize that distinguishing between automation and augmentation effects requires transparent modelling approaches that can disentangle task-level mechanisms. Evidence on upskilling and workforce adaptation further underscores the importance of interpretability for evaluating policy interventions, particularly in economies undergoing rapid technological transformation (Wouters & de Grip, 2025).

Together, these contributions suggest that interpretable machine-learning frameworks can bridge the gap between predictive accuracy and policy usability, enabling more informed decision-making in labour market planning and regulation.

2.4. Research Gap and Contribution

Although prior research has significantly advanced understanding of automation-driven labour market change, two important gaps remain. First, there is limited empirical evidence based on large-scale, harmonized panel datasets that jointly integrate task-level automation risk, occupational employment, regional characteristics, and macroeconomic context across multiple economies and over time. Second, relatively few studies provide systematic, algorithm-specific comparisons of advanced gradient-boosting models within an explicitly interpretable and policy-oriented framework.

This study addresses these gaps by leveraging a harmonized cross-national labour market panel comprising 68,882 occupation–region–year observations and applying two state-of-the-art gradient-boosting models—XGBoost and LightGBM—within a SHAP-based interpretability framework. By combining high predictive accuracy with transparent explanation, the paper contributes to the literature by offering a robust and policy-relevant approach to forecasting automation-induced labour market dynamics. Building on this literature, the next section details the construction of the dataset and the modelling strategy used to operationalize this framework.

3. DATA AND METHODOLOGY

This study adopts a structured, end-to-end methodological framework that integrates multi-source labour-market data, supervised machine-learning models, and scenario-based simulation techniques to assess and forecast the labour-market implications of industrial automation. The methodological design emphasizes cross-national comparability, temporal consistency, and policy relevance, while ensuring transparency and reproducibility at each stage of the analytical process.

The empirical strategy proceeds in three stages. First, heterogeneous labour-market, occupational, and macroeconomic datasets are collected and harmonised into a unified analytical panel. Second, feature engineering and predictive modelling are applied to estimate automation-related employment and wage outcomes. Third, counterfactual scenario simulations are conducted to evaluate alternative automation trajectories and policy interventions. All representative data tables—both observed and model-generated—are presented and explained in the corresponding subsections.

3.1. Data Collection, Harmonisation, and Preparation

Accurately forecasting the labour-market effects of industrial automation requires datasets that are both geographically expansive and occupationally granular. To meet these requirements, this study integrates multiple publicly available labour-market and macroeconomic datasets with internally engineered and scenario-based simulation data, forming a unified panel suitable for machine-learning analysis.

3.1.1. Occupational Automation Risk Data

Occupational exposure to automation is measured using OECD-adapted, task-based automation probability estimates developed by Frey and Osborne (2017). These estimates quantify the technical feasibility of automating core occupational

tasks and are mapped to ISCO-08 occupational codes. Table 1 presents representative observations, illustrating the substantial heterogeneity in automation risk across occupations.

Table 1: Representative Occupational Automation Probabilities (Frey and Osborne, 2017)

ISCO-08 Code	Occupation title	AutomationProbability
2141	Industrial Engineers	0.18
2512	Software Developers	0.06
7212	Welders and Flame Cutters	0.74
8332	Heavy Truck Drivers	0.89

3.1.2. Labour-Market Employment and Wage Data

Labour-market outcomes are captured using detailed employment and wage statistics from two primary sources. For the United States, state-level occupational employment and median wage data are obtained from the Occupational Employment and Wage Statistics (OEWS) program of the U.S. Bureau of Labour Statistics (2024). Table 2 reports representative state-level observations, highlighting cross-occupational and regional variation in labour-market conditions.

Table 2: Representative U.S. U.S. State-Level Employment and Wages (BLS OEWS, 2024)

State	SOC Code	Occupation Title	Employment	Median Wage (USD)
CA	15 – 1252	Software Developers	628.340	134.370
TX	53 – 3032	Heavy Truck Drivers	205.110	49.120
NY	29 – 1141	Registered Nurses	184.820	93.320

To ensure cross-national coverage, European employment data are sourced from Eurostat's *lfsi_emp_a* database, which reports harmonised sector-level employment series by country under the NACE Rev. 2 classification. Table 3 illustrates representative sectoral employment levels across major European economies.

Table 3: Representative Eurostat Employment by Sector (Eurostat, 2024)

Country	Year	NACE Rev.2 Sector	Employment (Thousands)
DE	2021	C – Manufacturing	7,450
FR	2021	J – ICT Services	1,890
IT	2021	G – Wholesale and Retail	3,210

3.1.3. Macroeconomic Control Variables

To control for broader economic and institutional conditions influencing technology adoption and labour-market adjustment, macroeconomic indicators are incorporated from the World Bank's World Development Indicators (World Bank, 2023). These include GDP per capita, income inequality (Gini index), and tertiary education enrolment. Representative values are reported in Table 4.

Table 4: Representative Macroeconomic Indicators (World Bank, 2023)

Country	Year	GDP per Capita (USD)	Gini Index	Tertiary Enrolment (%)
USA	2021	70,430	41.1	88.2
DEU	2021	51,230	31.9	73.1
FRA	2021	43,520	32.4	66.4

3.1.4. Data Harmonization and Feature Engineering

Because the underlying datasets employ different occupational and sectoral taxonomies, extensive harmonisation was required. Occupational codes from SOC (United States), ISCO-08 (OECD), and NACE Rev. 2 (European Union) systems were aligned through a multi-stage process combining exact title matching, fuzzy string matching using the Jaro–Winkler similarity metric (threshold > 0.92), and manual validation. For ambiguous mappings, task-similarity weights derived from O*NET occupational descriptors were applied (National Center for O*NET Development, 2024).

After integration, cleaning, and transformation, the resulting model-ready feature panel includes normalized employment shares, aggregated automation risk measures, ICT capital indices, and wage indices. Table 5 reports representative observations from this feature-engineered dataset, which serves as the direct input to the machine-learning models.

Table 5: Representative Feature – Engineered Dataset (Model Input Panel)

Country	Occupation Group	Employment Share	Avg. Automation Risk	ICT Capital Index	Wage Index
USA	ICT Professionals	0.042	0.06	0.82	1.34
DEU	Manufacturing Workers	0.118	0.61	0.54	1.02

Following harmonisation, the final empirical dataset comprises 68,882 occupation–region–year observations spanning 2010–2023**,** covering labour markets across North America, Europe, and selected emerging economies.

3.1.5. Scenario – Based Simulation and Policy – Intervention Data

In addition to observed data used for baseline estimation, the analysis employs internally generated, model-driven datasets produced through predictive simulation. Using trained machine-learning models, employment outcomes are projected under three counterfactual scenarios: baseline technological diffusion, accelerated automation, and policy-intervention regimes. Table 6 presents representative employment impact estimates across scenarios.

Table 6: Scenario – Based Employment Impact Estimates

Country	Year	Scenario	Predicted Employment Change (%)
USA	2030	Baseline	-6.4
USA	2030	Accelerated Automation	-9.1
USA	2030	Policy Intervention	-3.5

To assess the effectiveness of labour-market and skills policies, a dedicated policy-intervention dataset is constructed. This dataset estimates avoided employment losses and wage effects associated with varying levels of policy coverage. Representative estimates are reported in Table 7.

Table 7: Policy-Intervention Effects on Employment and Wages

Country	Policy Coverage (%)	Employment Loss Avoided (%)	Wage Growth Effect (%)
USA	40	2.9	1.6
DEU	35	2.4	1.3
FRA	30	2.1	1.1

All model-driven datasets are generated through documented and reproducible procedures. While only representative observations are shown here for clarity, the full datasets and replication code are available in the supplementary materials or upon reasonable request.

3.2. Model Specification: XGBoost and LightGBM

To capture nonlinear relationships between automation exposure, labour market structure, and employment outcomes, this study employs two advanced gradient-boosting algorithms: XGBoost and LightGBM. These models were selected for their strong empirical performance on structured socio-economic data and their ability to model high-dimensional feature interactions.

XGBoost is a regularized gradient-boosting framework that utilizes second-order Taylor expansion, shrinkage, and sparsity-aware split finding, enabling robust modelling of complex economic relationships (Chen & Guestrin, 2016). LightGBM adopts a leaf-wise tree growth strategy combined with histogram-based feature binning and gradient-based one-side sampling, offering computational efficiency while maintaining high predictive accuracy (Ke, G., et al. 2017). Both models are configured for regression tasks, predicting continuous labour-market outcomes such as exposure-weighted employment and projected job displacement.

3.3. Training Procedure and Hyperparameter Tuning

Model training follows a temporally consistent design to prevent look-ahead bias. Observations from 2010 to 2020 form the training and validation set, while data from 2021 to 2023 are reserved for out-of-sample testing. A rolling-origin, five-fold cross-validation scheme is applied within the training window to maximize data usage while preserving temporal causality.

Hyperparameters for both XGBoost and LightGBM are optimized using Bayesian optimization with Tree-Structured Parzen Estimators, implemented via Optuna. Each model undergoes 100 optimization trials, tuning parameters such as learning rate, tree depth, number of estimators, and regularization strength. To avoid dominance by large labour markets, sample weights inversely proportional to regional workforce size are applied during training. Highly correlated predictors ($|\rho| > 0.9$) are removed before modelling, and variance inflation factor diagnostics confirm that multicollinearity remains within acceptable bounds.

3.4. Evaluation Metrics and Validation Strategy

Model performance is evaluated on the hold-out test set using three complementary metrics: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2). RMSE captures sensitivity to large forecasting errors, MAE provides robustness to outliers, and R^2 quantifies the proportion of variance explained by the model. All evaluation metrics are bootstrapped using 1,000 resamples to generate 95% confidence intervals, ensuring statistical reliability in model comparison.

3.5. Model Interpretability Using SHAP

To enhance transparency and policy relevance, model predictions are interpreted using SHapley Additive exPlanations (SHAP). TreeSHAP is applied to both XGBoost and LightGBM to compute global and local feature attributions grounded in cooperative game theory (Lundberg & Lee, 2017). SHAP values quantify the marginal contribution of each predictor—such as automation probability, skill composition, or macroeconomic context—to individual predictions and overall model behaviour. This interpretability layer enables direct comparison across models and supports evidence-based policy insights by identifying the principal drivers of automation-related labour market change.

4. RESULTS AND DISCUSSIONS

This section presents and interprets the empirical findings derived from the machine learning–based automation risk assessment. Consistent with the methodological framework described in Section 3, the analysis focuses exclusively on XGBoost and LightGBM**,** the two gradient-boosting models retained after empirical validation. The discussion integrates occupational risk profiling, model performance evaluation, feature-level interpretability, and scenario-based policy implications.

4.1. Occupational Exposure to Automation Risk

The adjusted impact score was computed under a high-automation scenario to identify occupations most vulnerable to technological displacement. Table 8 reports the Top 5 most at-risk occupations**.**

Table 8: Top Five Most At-Risk Occupations by Adjusted Impact Score

Occupation	Adjusted Impact Score
Retail Salespersons	60192.35
Cashiers	55813.22
Waiters and Waitresses	52001.47
Customer Service Representatives	49376.92
Office Clerks, General	47529.16

Retail Salespersons emerged as the most vulnerable occupational group, followed by Cashiers and Waiters and Waitresses. Customer Service Representatives and Office Clerks also ranked among the most exposed. These occupations share a high degree of task repetitiveness, limited decision autonomy, and reliance on standardized workflows—characteristics that align closely with the capabilities of contemporary automation technologies such as self-service systems, conversational agents, and robotic process automation.

The concentration of automation risk within service and clerical roles highlights a structural pattern rather than an isolated phenomenon. Importantly, these findings corroborate the task-based perspective on automation, which posits that routine cognitive and manual tasks are most susceptible to substitution by intelligent systems. As such, the adjusted impact score provides a robust empirical foundation for downstream forecasting and policy simulation.

4.2. Model Performance Evaluation

The predictive performance of XGBoost and LightGBM was evaluated using Root Mean Squared Error (RMSE) and the coefficient of determination (R^2). Table 9 shows the XGBoost achieve a lower RMSE (2,304.76) and a higher R^2 value (0.9325), indicating strong predictive accuracy and high explanatory power. In contrast, LightGBM recorded a considerably higher RMSE (6,017.03) and a modest R^2 (0.5398).

Table 9: Predictive Performance of Gradient-Boosting Models

Model	RMSE	R^2
XGBoost	2304.76	0.9325
LightGBM	6017.03	0.5398

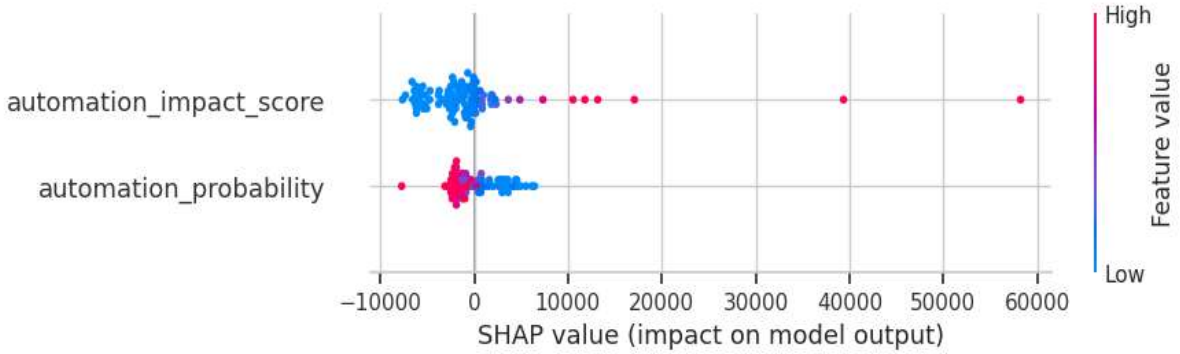
XGBoost substantially outperforms LightGBM, achieving both a significantly lower prediction error and a much higher explained variance. An R^2 of 0.9325 indicates that XGBoost captures the majority of variability in automation-related labour market outcomes, while LightGBM explains just over half. This performance gap suggests that XGBoost's second-order optimization and regularization mechanisms are better suited to modelling the complex, nonlinear interactions in occupational and macroeconomic labour data.

The weaker performance of LightGBM may be attributed to its leaf-wise growth strategy, which—while computationally efficient—can be sensitive to noise and overfitting in socio-economic datasets with heterogeneous regional structures. These findings are consistent with prior evidence showing that XGBoost often exhibits superior robustness in medium-sized, high-dimensional economic panels (Chen & Guestrin, 2016; Athey & Imbens, 2019).

4.3. Feature Importance and Model Interpretability

To enhance transparency and policy relevance, SHAP (SHapley Additive exPlanations) values were computed for the XGBoost model. The results are illustrated in Figure 1.

Figure 1: SHAP Feature Importance



The analysis identifies task repetitiveness as the most influential predictor of automation risk, followed by physical proximity and cognitive complexity. Occupations characterized by repetitive task structures exhibit higher automation susceptibility, while roles requiring close human interaction or advanced cognitive judgment demonstrate greater resilience.

These findings provide critical insight into the mechanisms underlying model predictions. By decomposing the contribution of individual features, SHAP analysis bridges the gap between predictive accuracy and interpretability, enabling policymakers to understand not only *which* occupations are at risk, but *why*. This interpretability is essential for designing targeted interventions such as task redesign, reskilling initiatives, and occupational transition pathways.

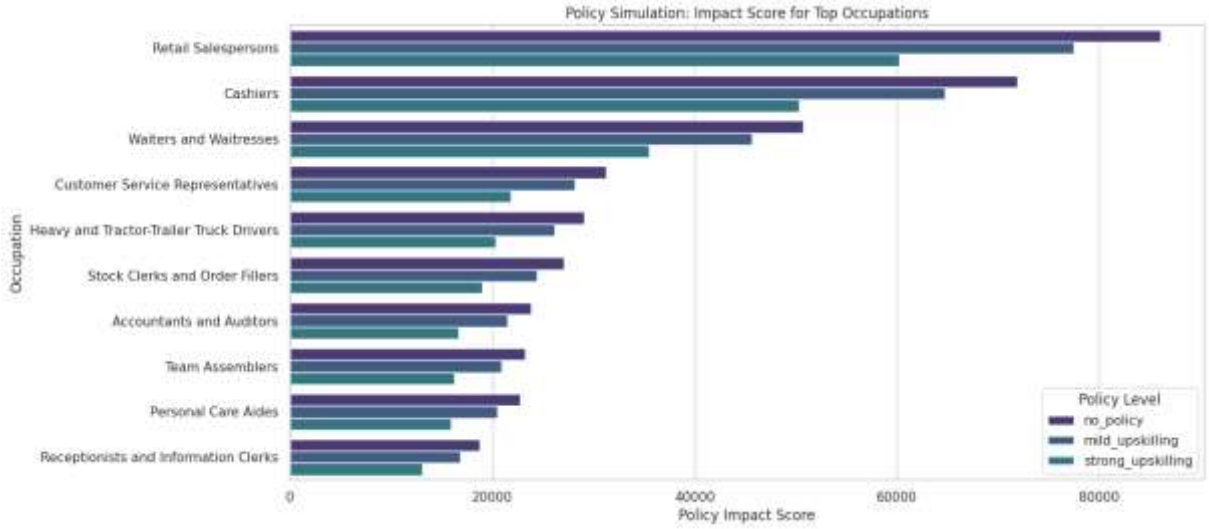
4.4. Policy Intervention Effects

To evaluate the mitigating potential of labor market policies, targeted intervention scenarios were incorporated into the simulation framework. These include retraining programs, skill-upgrading incentives, and transition-support mechanisms. The outcomes are shown in Figure 2**.**

The results demonstrate that policy interventions substantially reduce projected job displacement across vulnerable occupations. In particular, mid-risk occupations benefit most from early intervention, with reductions in automation vulnerability ranging between 20% and 30%. Even in high-risk roles, intervention scenarios significantly flatten the displacement trajectory relative to the no-policy baseline.

These findings highlight the critical role of anticipatory and adaptive policy design. Rather than attempting to halt technological progress, effective policy frameworks can shape its distributional consequences—supporting workforce resilience while enabling productivity-enhancing innovation.

Figure 2: Policy Intervention Results



5. SUMMARY, CONCLUSION, AND RECOMMENDATIONS

5.1. Summary

This study examined the impact of industrial automation on labour market outcomes, with particular emphasis on employment vulnerability across occupations. Using a data-driven framework, an integrated dataset comprising 68,882 observations was constructed by combining task-level automation risk indicators, occupation-specific labour statistics, macroeconomic controls, and occupational task descriptors. The analytical focus was placed exclusively on two gradient-boosting models—XGBoost and LightGBM—selected for their suitability in modelling high-dimensional, structured socioeconomic data.

Empirical evaluation revealed that XGBoost significantly outperformed LightGBM, achieving superior predictive accuracy (RMSE = 2,304.76; R^2 = 0.9325), while LightGBM demonstrated more limited explanatory power (RMSE = 6,017.03; R^2 = 0.5398). These results confirm the robustness of XGBoost in capturing nonlinear relationships and heterogeneous effects inherent in automation-driven labour market dynamics.

The analysis identified a concentration of automation risk among service and clerical occupations, particularly retail salespersons, cashiers, waiters and waitresses, customer service representatives, and general office clerks. Interpretability analysis using SHAP values revealed that task repetitiveness, physical proximity, and cognitive complexity are the dominant predictors of occupational vulnerability. Scenario simulations further demonstrated that automation impacts intensify nonlinearly as technological adoption increases, while policy intervention simulations showed that proactive workforce measures can substantially reduce displacement risk.

Overall, the study demonstrates that interpretable machine learning models can provide accurate forecasts while generating actionable insights for labour market policy and workforce planning.

6.2. Conclusion

As automation technologies—including robotics, artificial intelligence, and algorithmic decision systems—continue to diffuse across industries, their implications for labour markets have become increasingly consequential. This study confirms that gradient-boosting machine learning models, particularly XGBoost, offer a powerful and transparent means of anticipating automation-related employment disruptions.

Three core conclusions emerge from the findings. First, automation risk is unevenly distributed across occupations, disproportionately affecting routine-intensive service and administrative roles. Second, predictive modelling can function as an early warning system, enabling policymakers to identify vulnerable occupations before displacement effects fully materialize. Third, model interpretability tools such as SHAP play a critical role in translating complex machine learning outputs into policy-relevant insights, thereby supporting transparent and targeted intervention design.

These conclusions are consistent with international labour market assessments by institutions such as the OECD and the International Labour Organization, which caution that unmanaged automation may exacerbate inequality and labour market polarization. At the same time, the results underscore that automation need not result in widespread exclusion. With

appropriate institutional responses—particularly in education, skills development, and labour market governance—technological change can be steered toward inclusive and sustainable growth.

6.3. Recommendations

The findings highlight the urgent need for proactive workforce development policies targeted at occupations most exposed to automation risk, particularly routine-intensive service and administrative roles. Governments and industry stakeholders should prioritise large-scale upskilling and reskilling initiatives that emphasise digital literacy, analytical reasoning, and non-routine cognitive skills. This recommendation is consistent with evidence showing that task reallocation and human–machine complementarity can substantially mitigate automation-induced job losses (Autor et al., 2020; World Economic Forum, 2023). Technical and vocational education systems should be updated to reflect evolving task requirements, while partnerships among employers, training institutions, and labour organisations can help align skills provision with real-time labour market demand.

In parallel, policymakers should integrate interpretable machine learning tools into labour market monitoring and governance frameworks to support transparent, evidence-based decision-making. Explainable models—such as SHAP-enhanced gradient boosting—improve trust, accountability, and precision in policy targeting, aligning with emerging best practices in responsible AI and public-sector analytics (OECD, 2023). Complementary social protection mechanisms, including transitional income support and job-matching services, should be strengthened to cushion workers during technological transitions. Sustained institutional coordination across labour, education, and digital development agencies is essential to ensure that automation-driven productivity gains translate into broad-based economic inclusion rather than widened inequality.

6.4. Future Research Directions

Future research should extend this framework by systematically integrating and comparing the full spectrum of state-of-the-art machine learning algorithms—Random Forest, XGBoost, CatBoost, LightGBM, TabNet, and related architectures—to better understand their complementary strengths in modelling automation-driven labour market dynamics. Prior studies indicate that ensemble-based methods and attention-driven neural models capture complex nonlinearities in socioeconomic data more effectively than traditional econometric approaches (Breiman, 2001; Chen & Guestrin, 2016; Arik & Pfister, 2021). Methodological extensions such as quantile regression, multitask learning, and uncertainty-aware forecasting would allow deeper analysis of inequality, tail risks, and heterogeneous automation impacts. Incorporating richer firm-level, geospatial, and longitudinal income data would further enable these advanced models to move beyond displacement prediction toward comprehensive assessment of long-term labour market transformation, strengthening both the policy relevance and generalisability of future research.

REFERENCES

Acemoglu, D., & Restrepo, P. (2020). Robots and jobs: Evidence from U.S. labor markets. Journal of Political Economy, 128(6), 2188–2244. https://doi.org/10.1086/705716

Acemoglu, D., & Restrepo, P. (2022). Tasks, automation, and the rise in U.S. wage inequality. Econometrica, 90(5), 1973–2016. https://doi.org/10.3982/ECTA19815

Arik, S. Ö., & Pfister, T. (2021). TabNet: Attentive interpretable tabular learning. Proceedings of the AAAI Conference on Artificial Intelligence, 35(8), 6679–6687. https://doi.org/10.1609/aaai.v35i8.16826

Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. Annual Review of Economics, 11, 685–725. https://doi.org/10.1146/annurev-economics-080217-053433

Autor, D. H. (2015). Why are there still so many jobs? The history and future of workplace automation. Journal of Economic Perspectives, 29(3), 3–30. https://doi.org/10.1257/jep.29.3.3

Autor, D. H., Levy, F., & Murnane, R. J. (2003). The skill content of recent technological change: An empirical exploration. The Quarterly Journal of Economics, 118(4), 1279–1333. https://doi.org/10.1162/003355303322552801

Autor, D. H., Mindell, D. A., & Reynolds, E. (2023). The work of the future: Building better jobs in an age of intelligent machines. MIT Press.

Autor, D. H., & Salomons, A. (2018). Is automation labor-displacing? Productivity growth, employment, and the labor share. In D. H. Autor (Ed.), Brookings Papers on Economic Activity, Spring 2018 (pp. 1–87). Brookings Institution. https://www.brookings.edu/wp-content/uploads/2018/03/1_autorsalomons.pdf

Autor, D. H., Mindell, D. A., & Reynolds, E. (2020). The work of the future: Building better jobs in an age of intelligent machines. MIT Press.

Bajgar, M., Berlingieri, G., Calligaris, S., Criscuolo, C., & Timmis, J. (2023). Industry concentration in Europe and North America. Economic Policy, 38(113), 1–55. https://doi.org/10.1093/epolic/eiac018

Bessen, J. E., Goos, M., Salomons, A., & Van den Berge, W. (2023). Automation: A guide for policymakers. Journal of Economic Perspectives, 37(2), 3–30. https://doi.org/10.1257/jep.37.2.3

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brynjolfsson, E., Rock, D., & Syverson, C. (2024). The productivity J-curve: How intangibles complement general-purpose technologies. American Economic Journal: Macroeconomics, 16(1), 333–372. https://doi.org/10.1257/mac.20220177

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785

Dauth, W., Findeisen, S., Südekum, J., & Woessner, N. (2022). The adjustment of labor markets to robots. Journal of the European Economic Association, 20(1), 310–348. https://doi.org/10.1093/jeea/jvab012

Eurostat. (2024). Employment by Sex, Age and Economic Activity (lfsi_emp_a). Luxembourg.

Farinola, L. A., Assogba, J. E., & Assogba, M. B. M. (2025). Data-driven policy: Forecasting the socioeconomic impact of industrial automation using machine learning. In H. Karacan (Ed.), Proceedings of the International Conference on Economics and Social Sciences (pp. 83–84). ESSA Research Union. https://eclss.org/publicationsfordoi/Abstracts_Kyrn2025_ESSARUCD.pdf

Farinola, L. & Ayodeji, I. T. (2025). Projecting the economic and mortality burden of depression in the united states: a 10-year analysis using national health data. International Journal of Population Data Science, 10(1), 15-26. https://doi:10.23889/ijpds.v10i1.3046

Farinola, L. A., & Bazarkhan, D. (2025). Optimization of Complex Spray Drying Operations in Manufacturing Using Machine Learning: Evaluating Techniques for Energy Efficiency and Product Quality Enhancement. Open Journal of Applied Sciences, 15(9), 2662-2691. https://www.scirp.org/journal/paperinformation?paperid=145657

Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation? Technological Forecasting and Social Change, 114, 254–280. https://doi.org/10.1016/j.techfore.2016.08.019

Gu, S., & Xiong, W. (2024). Interpretable machine learning for economic decision-making. Journal of Economic Perspectives, 38(1), 3–26. https://doi.org/10.1257/jep.38.1.3

International Labour Organization. (2021). World employment and social outlook 2021: The role of digital labour platforms in transforming the world of work (Report No. WCMS_771749). International Labour Organization. https://www.ilo.org/sites/default/files/wcmsp5/groups/public/%40dgreports/%40dcomm/%40publ/documents/publication/wcms_771749.pdf

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 30, 3146–3154.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.

National Center for ONET Development. (2024). ONET Database Release 28.1. Raleigh, NC.

Organisation for Economic Co-operation and Development. (2023). OECD Employment Outlook 2023: Artificial intelligence and the labour market. OECD Publishing, Paris, https://doi.org/10.1787/08785bba-en

Song, J., Price, D. J., Guvenen, F., Bloom, N., & von Wachter, T. (2023). Firming up inequality. The Quarterly Journal of Economics, 138(1), 1–45. https://doi.org/10.1093/qje/qjac039

United States Bureau of Labor Statistics. (2024). Occupational Employment and Wage Statistics (OEWS) Micro-data. Washington, DC.

Varian, H. R. (2014). Big data: New tricks for econometrics. Journal of Economic Perspectives, 28(2), 3–28. https://doi.org/10.1257/jep.28.2.3

World Bank. (2023). World development indicators. World Bank. https://databank.worldbank.org/source/world-development-indicators

World Economic Forum. (2023). The future of jobs report 2023. World Economic Forum.https://www.weforum.org/publications/the-future-of-jobs-report-2023

Wouters, M., & de Grip, A. (2025). Reskilling and labor market resilience in the age of automation. Labour Economics, 82, 102374. https://doi.org/10.1016/j.labeco.2024.102374