Explaining regional wage disparities with machine learning: A SHAP-based interpretation approach

Andrzej Dudek,[a] Marcin Pełka,[b] Artur Skiba[c]

Abstract. The aim of the study is to provide an explanation for the factors that most influence the differences in wage levels between Polish powiats (equivalent to counties). This study investigates regional wage disparities in Poland by applying machine learning models enhanced by Explanatory Model Analysis techniques. Using powiat-level data from the Local Data Bank (Pol. Bank Danych Lokalnych – BDL) for 2010 and 2023, a neural network framework was developed to predict wage levels based on economic, demographic, infrastructural and environmental variables. To interpret the model, we employed the Variable Importance over Permutation (VIP) and SHapley Additive exPlanations (SHAP) approaches, which provide insights into both the global feature importance and the local contributions of individual variables. The results indicate that the share of the productive population, unemployment rates and social vulnerability remain key determinants of wage differences, although their relative influence shifts significantly over time. The SHAP analysis demonstrates how regional contexts such as Jelenia Góra and Wrocław powiats exhibit distinct factor dynamics, with demographic and infrastructural variables playing varying roles across the studied years. The findings highlight the potential of combining machine learning with explainability methods to uncover complex, nonlinear determinants of wages,

[a] Wroclaw University of Economics and Business, Faculty of Economics and Finance, Department of Financial Investments and Risk Management, ul. Komandorska 118/120, 53-345 Wrocław, Poland, e-mail: andrzej.dudek@ue.wroc.pl, ORCID: https://orcid.org/0000-0002-4943-8703.

[b] Wroclaw University of Economics and Business Branch in Jelenia Góra, Faculty of Economics and Finance, Department of Econometrics and Computer Science, ul. Nowowiejska 3, 58-500 Jelenia Góra, Poland, e-mail: marcin.pelka@ue.wroc.pl, ORCID: https://orcid.org/0000-0002-2225-5229.

[c] Polish Information Processing Society, ul. Solec 38 lok. 103, 00-394 Warszawa, Poland, e-mail: artur.skiba70@gmail.com, ORCID: https://orcid.org/0000-0003-4616-7271.

offering a more transparent analytical basis for understanding evolving regional disparities.

Keywords: deep learning, machine learning, explanatory model analysis, wage disparities

JEL: C15, C45, O150

1. Introduction

Regional wage disparities remain a central topic in labor economics, often explained by the differences in human capital endowments, sectoral structures, and spatial inequalities (Combes et al., 2008; Moretti, 2011). Traditional econometric models have been widely used to quantify these disparities, yet they frequently rely on restrictive assumptions that may not capture complex, nonlinear interactions between the explanatory factors. Recent advances in machine learning provide a powerful alternative by enabling predictive modeling that accommodates high-dimensional and interdependent features without imposing strong functional form restrictions (Mullainathan & Spiess, 2017). However, the opacity of machine learning methods has raised concerns about interpretability, especially in policy-relevant domains such as labor markets, where transparent explanations are crucial.

To address this challenge, methods of Explainable AI (XAI) like SHapley Additive exPlanations (SHAP) have emerged as a robust framework for interpreting complex machine learning models by attributing feature importance based on the principles of the cooperative game theory (Lundberg & Lee, 2017; Masís, 2023; Molnar, 2020). Applying SHAP to wage prediction models allows for a granular understanding of how regional characteristics such as industrial composition, education levels, or urbanization contribute to the observed wage gaps. This approach bridges predictive performance with interpretability, enabling researchers and policymakers to identify the factors that matter most and see how their effects vary across regions. By combining Machine Learning with a SHAP-based interpretation, the analysis of regional wage disparities can advance beyond aggregate statistical associations toward more actionable, fine-grained insights.

2. Analysis of an explanatory model for studies on wage differences

2.1. Analysis of wage differences: A literature review

Recent literature offers numerous analyses of spatial wage differentials across various territorial levels, including Polish powiats and voivodships, and Ukrainian oblasts (Adamczyk et al., 2009; Bolińska & Gomółka, 2018; Dykas et al., 2020; Dykas & Misiak, 2013; Kapela & Kwiatkowski, 2023; Przekota, 2016). Theoretical frameworks typically rely on efficiency wage models. Empirical studies use such indicators as wages, labor productivity, and unemployment rates to estimate wage determinants via regression analysis. Beyond basic metrics, newer models such as those by Kapela and Kwiatkowski (2023) incorporate variables like higher education rates, technological innovation, and patent activity, while also addressing the effects of the 2020 pandemic. The applied methods include least squares, the generalized method of moments, clustering methods, and fixed effects models, which enhance the accuracy of the results. Findings show that factors like proximity to large cities, labour productivity, and human capital play crucial roles in wage disparities, while results regarding capital expenditures and industry output remain ambiguous (Adamczyk et al., 2009; Przekota, 2016).

Wage elasticity relative to unemployment remains a central theme. Many studies confirm that a negative relationship between the two exists, as seen in an earlier work by Phillips (1958) and later by Kaliski (1964), Blanchflower and Oswald (1990), though exceptions occur, such as in South Africa (Kingdon & Knight, 2006) and in some Polish powiat-level (equivalent to county-level) fixed-effects models (Dykas & Misiak, 2013). Modern applications of the Phillips curve continue to show relevance in different national contexts (Bartosik & Mycielski, 2015; Machuca & Cota, 2017). Other important aspects include the growing role of education, innovation, and demographic shifts in explaining wage variation (Combes et al., 2008; Kapela & Kwiatkowski, 2023). Despite the robust research at higher administrative levels, recent powiat-level studies are scarce, with the latest comprehensive analyses dating back to 2014 (Dykas & Misiak, 2013). Consequently, a renewed need emerged to reassess spatial wage dynamics at the powiat level, particularly in light of the post-pandemic developments and ongoing socio-economic changes (c.f. Luśtyk et al., 2024).

2.2. Methods of explanatory model analysis

To address the challenges described in the previous section, newly arisen methods of Explanatory Model Analysis/Explainable AI (see Biecek & Burzykowski, 2021; Masís, 2023; Molnar, 2020), particularly through Variable Importance over Permutation (VIP) and SHAP values, offer significant advantages in analyzing economic phenomena.

VIP enables researchers to assess the relative impact of each predictor by measuring the change in model performance after randomly permuting individual variables. This model-agnostic method provides an intuitive ranking of features, highlighting the most influential economic indicators driving predictive accuracy. It supports a transparent, reproducible evaluation of variable relevance, which is essential for policy analysis and decision-making in complex economic systems.

SHAP values further enhance the explanatory power by attributing prediction contributions to individual features in a theoretically grounded manner based on the cooperative game theory. Unlike aggregate importance scores, SHAP delivers local explanations for each prediction, allowing analysts to understand heterogeneity across economic agents or regions. This granularity is particularly valuable for exploring non-linear interactions and dependencies commonly present in econometric models. Together, VIP and SHAP form a robust framework for interpreting black-box machine learning models, facilitating deeper insights into causal mechanisms and improving the credibility of data-driven economic policy recommendations.

Other methods that make explaining black box models possible are partial dependence plots (PDP), which show the marginal effect that one or two variables (features) have on the predicted outcome (Friedman, 2001; Greenwell et al., 2018). PDPs capture only the main effect of the feature and ignore the possible interactions, so it should be used with care.

Accumulated local effects (ALE) plots describe how variables influence the prediction. Moreover, ALE plots are faster than PDPs (Apley & Zhu, 2020). In the ALEs, however, an interpretation of the effect across intervals is not permissible if the features are strongly correlated. ALE effects may differ from coefficients specified in linear regression models when variables interact and are correlated. What is more, ALE plots are not accompanied by Individual Conditional Explanation (ICE) curves and can have many small ups and downs. In this case, when we reduce the number of variables, we not only make the estimates more stable but also smooth out the complexity of the model.

A feature interaction model based on Friedman’s H statistic (Friedman & Popescu, 2008) and variable interaction networks (Hooker, 2004) allow variable interactions to be taken into account in the predictions.

Another way to interpret variable importance is through functional decomposition. It can be done by: functional Analysis of Variance (ANOVA) (Hooker, 2004), generalized functional ANOVA for dependent variables (features) (Hooker, 2007), generalized additive regression modes, or ALE plots.

The permutation feature importance algorithm based on Fisher et al. (2019) measures the increase in the prediction error of the model after the variable’s values are permuted, which breaks the relationship between the variable and the known (true) outcome.

The global surrogate model is another interpretable model that is trained to approximate the prediction of a black box model. The surrogate model uses a much simpler model instead of a complex one (Molnar, 2020).

The local interpretable model-agnostic explanations (LIME) is a technique that approximates any black box machine learning model with a local, interpretable model to explain each individual prediction that is described in the paper by Ribeiro et al. (2016). The main idea is that we perturb (change) the original data points, feed them into the black box model, and then observe the corresponding outcomes. Then the method weighs those new data points as a function of their proximity to the original point. Ultimately, using those sample weights, LIME fits a surrogate model, such as linear regression, on the dataset with the variations. Each original data point can then be explained with the newly trained explanation model.

2.3. VIP and SHAP methods for model explanation

Permutation-based methods like VIP, based on the idea introduced by Breiman (2001), provide a model-agnostic approach to estimating variable importance. This is done by assessing the impact of controlled perturbations in the input data on the predictive performance. Instead of relying on the internal structure of a model, this technique treats the model as a black box and evaluates how the quality of a prediction changes when the values of a given variable are deliberately disrupted. If the variable contributes substantially to the model's predictive mechanism, permuting its values should lead to a notable decline in performance. In contrast, if the variable has little or no influence, prediction quality should remain to a large extent unaffected.

The change in performance – measured through metrics such as mean squared error, accuracy, or alternative loss functions – serves as an inverse proxy for variable importance. A larger degradation in predictive quality implies a higher significance of the variable in the decision-making process. In practice, this procedure is implemented by randomly permuting the values of a selected feature across observations in the dataset and re-evaluating the model's output. Repeating this process for each variable provides a systematic and interpretable measure of feature importance that is independent of the model specification.

This process involves what follows.

Let:

X – a dataset with m explanatory variables and n instances (objects),

Y – column vector of the observed values of the dependent variable,

$\hat{Y}$ – column vector of the predicted values of the dependent variable,

$P(\hat{Y}, X, Y)$ – performance metrics (loss function) for the model.

The procedure then involves the following steps:

1. Training the model;

2. Computing $p_0 = P^0(\hat{Y}, X, Y)$, i.e. the initial value of the loss function;

3. Shuffling (permuting) column vector X_k for given $1 < k < m$. Matrix X after permutation becomes $X^{(*k)}$;

4. Computing model predictions $\hat{Y}^{*k}$ for X^{*k};

5. Computing $p_{*k} = P(\hat{Y}^{*k}, X^{*k}, Y)$;

6. Estimating the importance for variable k in the process of prediction through $vip_k = p_{*k} - p_0$ (alternatively used in the $vip_k = \frac{p_{*k}}{p_0}$ form).

The Shapley values, another technique of Explanatory Model Analysis, originating from the cooperative game theory, provide a rigorous framework for quantifying the joint contribution of explanatory variables to model predictions. In Shapley's (1953) original formulation, the method determined each player's marginal contribution to the overall payoff obtained by a coalition. Transposed into model interpretation, the 'players' are the variables, and the 'payoff' corresponds to the model's prediction. Thus, Shapley values measure how the estimated outcome changes when a specific variable is added to the different subsets of predictors involved in generating the prediction.

The final attribution is obtained as a weighted average of these marginal contributions across all possible subsets. The weighting scheme depends on the size of the subsets: variables added to very small or nearly complete subsets receive higher weights, whereas those added to medium-sized subsets are assigned lower weights. This ensures fairness in attributing contributions across all possible coalitions of variables. The resulting SHAP provides a consistent and theoretically grounded measure of variable importance at both the global (model-wide) and local (instance-specific) levels.

The algorithm for finding the SHAP values for a certain object explained and a certain variable may be stated as follows:

Let:

X – a dataset with m explanatory variables and n instances (objects);

Y – column vector of the observed values of the dependent variable;

$\hat{Y}$ – column vector of the predicted values of the dependent variable;

l – object (instance) index for which the analysis is conducted;

k – feature (variable) index for which the analysis is conducted.

The procedure then involves the following steps:

1. Training the model;

2. Calculating $\widehat{Y_0} = \frac{\sum_{i=1}^{n} \widehat{Y_i}}{n}$, i.e. the average prediction value over the dataset (and initial explanation estimation);

3. Let:

$$V_{-k} = \{1,2,\dots,\mathrm{m}\}\backslash\{\mathrm{k}\} \tag{1}$$

(The set of all variable indices with k excluded);

4. For each s in *0,1,…,m-1*;

5. For all subsets S of V_{-k} of size s, calculating:

- $(\widehat{Y_l})^{*S}$ average prediction for the dataset for which variables' $X_i: i \in S$ values in the whole dataset are set to the values of object X_l;
- $(\widehat{Y_l})^{*S\cup\{k\}}$ be the average prediction for the dataset for which variables' $X_i: i \in S$ and variable's X_k values in the whole dataset are set to the values of object X_l;

and the Shapley value:

$$SHAP_S = \frac{s!\cdot(m-s-1)!}{m!}\left(\widehat{Y_l}^{*S\cup\{k\}} - \widehat{Y_l}^{*S}\right); \tag{2}$$

6. Summing all the $SHAP_S$ values.

The SHAP method was originally introduced by Štrumbelj and Kononenko (2010, 2014) and later popularized by Lundberg and Lee (2017). Its widespread application stems from a solid theoretical foundation and the reliability of its explanatory power.

3. Factors determining wage disparities. Research based on data from the Local Data Bank for 2010 and 2023

The analysis has been conducted on data describing the average compensation level in Polish powiats in the years 2010 and 2023. The data were acquired directly from the Local Data Bank (Pol. Bank Danych Lokalnych – BDL), which is Statistics Poland's official repository, through webservices and contained variables which describe economic (labor market), sociological, demographical, infrastructural and environmental phenomena. The description of dependent and exogenous variables along with BDL identifiers is presented in the Table.

Table. Description of variables used in the research

Variable ID	Internal Name	Type of variable	Description (English)
64428	compensation_level	Dependent variable (economic)	Average gross monthly wages in PLN
60530	regon_entities_ratio	Labor market	Business entities with registered REGON per 10,000 population
60270	unemployment_ratio	Labor market	Registered unemployment rate (overall)
458700	social_care_ratio	Sociological	Beneficiaries of social assistance by place of residence as the percentage of the total population
60566	productive_population_ratio	Demographical	The percentage share of the working-age population in the total population
450551	birthrate	Demographical	Natural increase (births minus deaths) per 1,000 population
450543	marriages_ratio	Demographical	Marriages per 1,000 population
60300	hotels_beds_ratio	Touristic	Bed places per 1,000 population

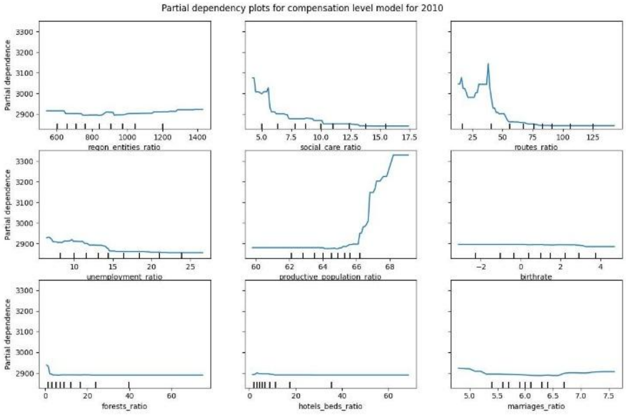
395404	routes_ratio	Infrastructural	Gmina (Polish equivalent to municipality) and powiat hard surface roads in km per 10,000 population
1646059	forests_ratio	Environmental	Municipal forest area in m2 *per capita*

Source: Local Data Bank (https://bdl.stat.gov.pl).

To find the most influenced factors for wages level modelling, we have built the eXtreme Gradient Boosting (Chen & Guestrin, 2016) model based on 319 objects describing powiats. The distinct models have been built for both studied years.

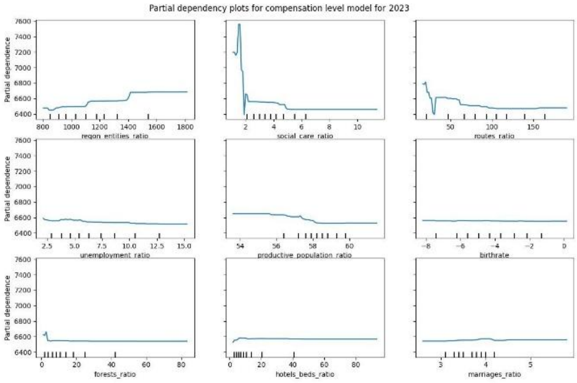
The python code implementing this procedure is included in Appendix 1. The full text results are presented in Appendix 2. The partial dependency plots presented in Figures 1 and 2 demonstrate that both models' convergence is stable.

Figure 1. Partial dependency plots for explanatory variables for wage levels in powiats in the 2010 model



Source: authors' calculations (code presented in Appendix 1).

Figure 2. Partial dependency plots for explanatory variables for wage levels in powiats in the 2023 model

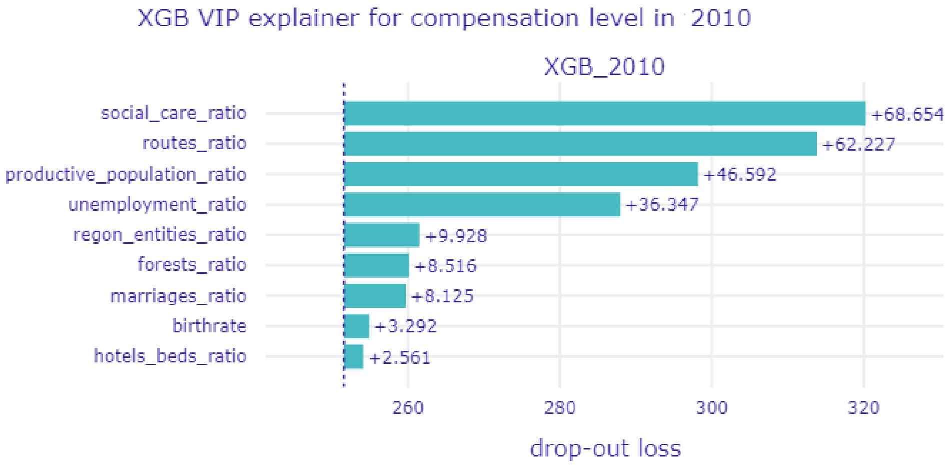


Source: authors' calculations (code presented in Appendix 1).

The model shows solid learning on training data ($R^2 = 0.626$). The test performance is positive and reasonable ($R^2 = 0.290$), indicating it captures useful predictive relationships. The gap between 0.626 and 0.290 suggests some degree of overfitting, but not severe, which is typical and acceptable for many socioeconomic datasets. The model generalizes moderately well and is reliable enough to proceed with interpretation (VIP, SHAP).

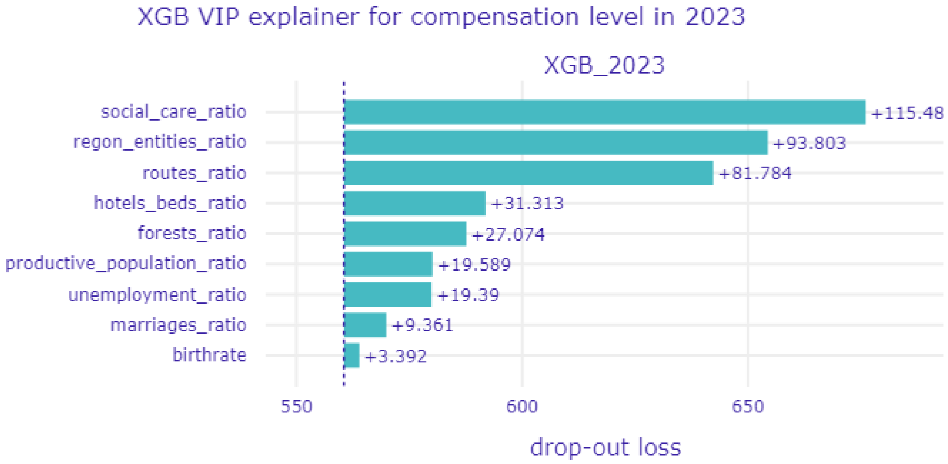
The VIP method is used to evaluate the influence of explanatory variables on the explained phenomena (wage level in powiats). The results for the models for 2010 and 2023 are presented in Figures 3 and 4.

Figure 3. Variable importance plot for exogenous variables for wage levels in powiats in 2010



Source: authors' calculations (code presented in Appendix 2).

Figure 4. Variable importance plot for exogenous variables for wage levels in powiats in 2023



Source: authors' calculations (code presented in Appendix 1).

The VIP results for 2010 indicate that the most influential variable is the *social_care_ratio*, with the highest dropout loss equal to 320.23. Thich means that removing this variable causes the strongest deterioration in model performance, suggesting that the social-assistance burden was a key structural determinant of compensation levels in 2010. The next highly influential variables are the *routes_ratio* (313.80) and *productive_population_ratio*

(298.16), both of which significantly worsen prediction when excluded, showing that transportation accessibility and the working-age population share are critical factors.

Further in the ranking, variables such as the *unemployment_ratio* (287.92), *regon_entities_ratio* (261.50), and *forests_ratio* (260.09) still contribute substantially to model accuracy, but their influence is more moderate. Their dropout losses imply that labor-market structure, business density, and environmental context affect compensation prediction, but to a lesser degree than factors related to social services and transport. These mid-ranked variables form a secondary explanatory layer that stabilizes the model.

At the lower end of the importance distribution, the predictors with the smallest dropout losses, namely the *marriages_ratio* (259.70), *birthrate* (254.86), and the *hotels_beds_ratio* (254.13) exerted the least influence in 2010. Removing them increases error only slightly, suggesting they contain comparatively limited independent information for determining compensation differences. In this year, demographic and tourism indicators appear marginal relative to the socioeconomic structure and accessibility.

The VIP analysis of the 2023 wage prediction model for Polish powiats highlights the relative strength of diverse structural, demographic, and environmental determinants.

In 2023, the variable importance structure shifts noticeably, with the *social_care_ratio* again emerging as the most influential predictor. This time, it shows an even higher dropout loss of 676.01, making it the dominant factor in the model. The next influential variables are the *regon_entities_ratio* (654.33) and *routes_ratio* (642.31), both showing large performance drops when removed. This highlights the growing importance of business density and transportation infrastructure for explaining compensation levels in 2023.

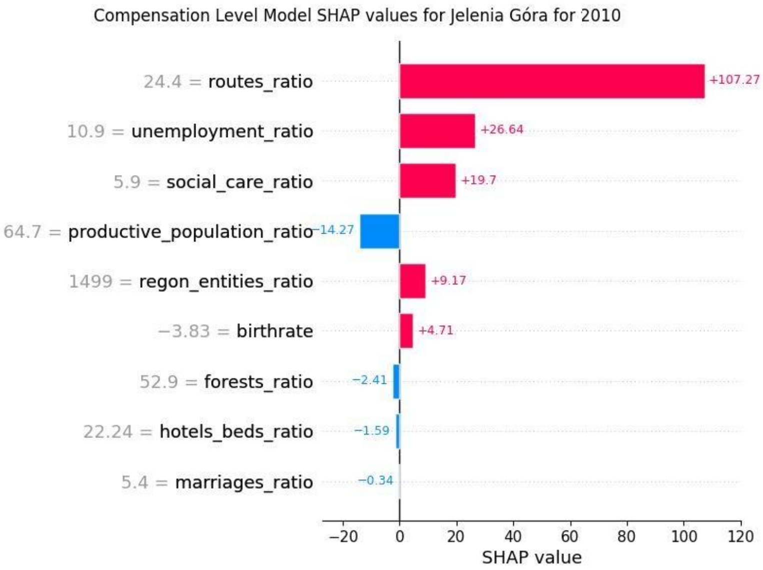
The middle tier of variables, including the *hotels_beds_ratio* (591.84), *forests_ratio* (587.60), and *productive_population_ratio* (580.12) also carry substantial explanatory weight. Their dropout losses show that tourism capacity, environmental features, and demographic composition meaningfully support model predictions. Compared to 2010, these secondary predictors become more informative, suggesting a more complex structure of the determinants.

The least influential predictors are the *unemployment_ratio* (579.92), *marriages_ratio* (569.89), and *birthrate* (563.92), whose dropout losses are closer to the full model but still in the lower range of importance. Although still impactful, the demographic and labor-market indicators exert smaller marginal effects compared with structural and institutional features. The 2023 importance pattern therefore portrays a landscape where social-service load, enterprise density, and infrastructure dominate compensation prediction, while demographic variables play a supportive yet reduced role.

The explanatory model analysis method allows a deeper insight into factors determining the analyzed phenomenon (compensation level). The analysis covers not only general model explanation but also most influential factors in individual cases.

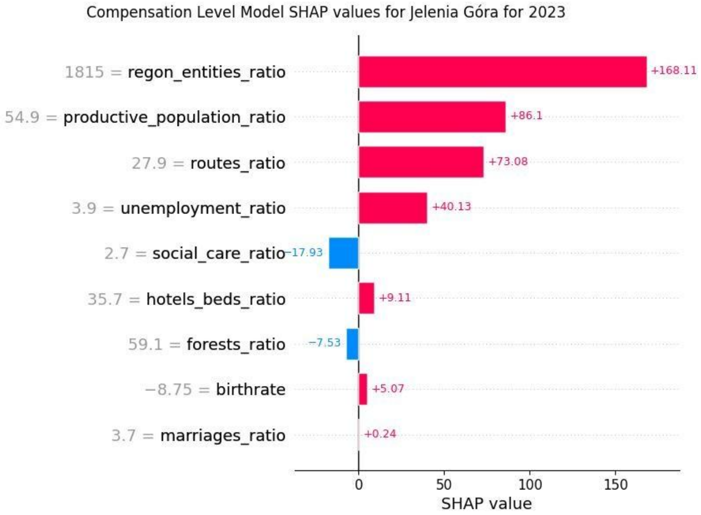
To better understand the influence of the given phenomena on overall compensation differences at local level, a SHAP analysis is conducted. The SHAP values for the 2010 and 2023 models for the Jelenia Góra powiat are presented in Figures 5 and 6.

Figure 5. SHAP plot for exogenous variables for wage levels in the Jelenia Góra powiat in 2010



Source: authors' calculations (code presented in Appendix 1).

Figure 6. SHAP plot for exogenous variables for the wage levels in the Jelenia Góra powiat in 2023



Source: authors' calculations (code presented in Appendix 2).

For the Jelenia Góra powiat, in 2010, the strongest SHAP contributor was the *routes_ratio*, with a positive effect of 107.27 at a value of 24.40. This highlights the powiat's relative transport accessibility as a major factor supporting its compensation prediction. The next significant variables are the *unemployment_ratio* (+26.64 at 10.90) and *social_care_ratio* (+19.70 at 5.90), indicating that despite relatively high unemployment and social-care indicators, these conditions still contribute positively within the model structure.

Negative contributions also proved to play an essential role. The *productive_population_ratio* (–14.27 at 64.70) pulls the prediction downward, suggesting demographic or economic strain associated with the powiat’s working-age population share. The *forests_ratio* (–2.41), *hotels_beds_ratio* (–1.59), and *marriages_ratio* (–0.34) also reduce the prediction slightly, implying that environmental and tourism indicators contribute less positively for Jelenia Góra compared to other powiats.

A few variables exert small positive influences. The *regon_entities_ratio* (+9.17 at 1,499) and *birthrate* (+4.71 at –3.83) add a marginal upward pressure on salaries. The overall SHAP structure for 2010 reflects a mix of strong transport infrastructure effects and modest socioeconomic constraints, with demographic features moderating the powiat’s predicted compensation level.

For Jelenia Góra in 2023, the *regon_entities_ratio* became the strongest positive contributor, with a SHAP value of +168.11 at 1,815 entities. This signals the increasing importance of local business density for salary levels. The *productive_population_ratio* (+86.10 at 54.90) and *routes_ratio* (+73.08 at 27.90) also strongly elevate the prediction, with transportation accessibility remaining a key structural advantage.

Additional positive contributions derive from the *unemployment_ratio* (+40.13 at 3.90) and *hotels_beds_ratio* (+9.11 at 35.70), indicating that tourism infrastructure played a more supportive role in 2023 than in 2010. Meanwhile, the *social_care_ratio* shows a negative impact (–17.93), which suggests an increasing sensitivity of the model to social-assistance burdens.

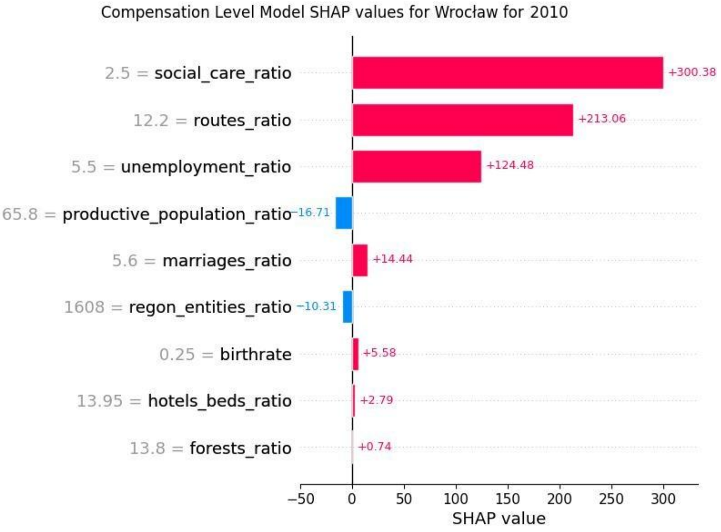
The *forests_ratio* also contributes negatively (–7.52), moderating the positive effects of other variables.

Smaller contributions come from *birthrate* (+5.07) and the *marriages_ratio* (+0.24), which have a limited influence. Overall, the SHAP profile for 2023 indicates that Jelenia Góra's salary structure is shaped by a combination of economic density, demographic composition, and improved labor-market indicators, with structural accessibility continuing to reinforce compensation predictions.

For Jelenia Góra, the SHAP comparison between 2010 and 2023 shows a clear shift in the structure of factors influencing compensation levels. In 2010, the main positive driver was the *routes_ratio* (+107.27 at 24.40), supported by the *unemployment_ratio* (+26.64) and *social_care_ratio* (+19.70), while the the *productive_population_ratio* (–14.27) exerted a negative influence and the remaining variables had only small effects. In 2023, however, the leading factor becomes the *regon_entities_ratio* (+168.11 at 1815), accompanied by strong positive contributions from the *productive_population_ratio* (+86.10) and *routes_ratio* (+73.08). This indicates a transition from an 'infrastructure-driven' model to a more 'economic-demographic' one. The role of the *social_care_ratio* also changes, from a small positive effect in 2010 (+19.70) to a clearly negative effect in 2023 (–17.93), suggesting the model became more sensitive to social-assistance burdens.

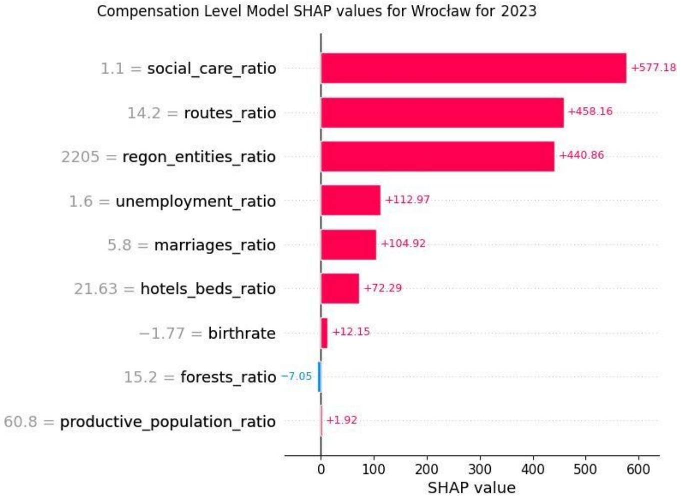
Jelenia Góra is a representative of medium-sized powiats. To broaden the analysis, the SHAP values have been estimated for a representative of larger powiats, like the Wrocław powiat with results presented in Figures 7 and 8.

Figure 7. SHAP plot for exogenous variables for wage levels in the Wrocław powiat in 2010



Source: authors' calculations (code presented in Appendix 1).

Figure 8. SHAP plot for exogenous variables for the wage levels in the Wrocław powiat in 2023



Source: authors' calculations (code presented in Appendix 1).

In the Wrocław powiat (2010), the SHAP analysis highlights the *social_care_ratio* as the dominant positive driver, contributing 300.38 units to the prediction at a value of 2.50. This indicates that Wrocław's low social-care burden is interpreted by the model as strongly favorable for compensation levels. Similarly, the *routes_ratio* (SHAP = 213.06, value = 12.20) exerts a substantial positive impact, reflecting Wrocław’s well-developed transport networks.

Another strong contributor is the *unemployment_ratio*, adding 124.48 units at a relatively low level of 5.50, suggesting that lower unemployment aligns with higher predicted salaries. The *marriages_ratio* also shows a smaller but positive impact (+14.44), hinting at demographic vitality. In contrast, the *regon_entities_ratio* (SHAP = –10.31 at 1,608 entities) slightly reduces the prediction, which may reflect saturation or diminishing marginal returns in areas with very high business density.

Most remaining variables contribute modestly. *Birthrate* (+5.58), the *hotels_beds_ratio* (+2.79), and *forests_ratio* (+0.74) collectively reinforce the positive prediction but with relatively small effects. Their limited magnitude suggests that Wrocław’s compensation structure in 2010 was driven far more by social infrastructure, transportation connectivity, and labor-market conditions than by tourism capacity or environmental features.

In 2023, the Wrocław powiat showed significantly larger SHAP magnitudes than in 2010. The strongest contributor was still the *social_care_ratio*, this time with an even more extreme value of +577.18 at a feature value of 1.10, reinforcing the model’s interpretation of a low social-care burden as a strong positive salary determinant. The *routes_ratio* follows with 458.16 at 14.20, highlighting substantial benefits from transport connectivity.

A major upward contribution also comes from the *regon_entities_ratio*, adding 440.86 at a high value of 2,205, implying that in 2023, business density exerted a far stronger positive effect than in 2010. The *unemployment_ratio* (+112.97) and *marriages_ratio* (+104.92) further elevated the compensation prediction, linking favorable labor-market and demographic conditions to higher wages.

Lesser yet notable effects included the *hotels_beds_ratio* (+72.29), *birthrate* (+12.15), and a small negative influence from the *forests_ratio* (–7.05). The *productive_population_ratio* contributed only +1.92, indicating minimal effect. Overall, the SHAP profile revealed that in 2023, Wrocław's compensation structure was strongly shaped by socioeconomic advantage, business density, and infrastructure, with demographic indicators reinforcing but not dominating the signal.

For Wrocław, the comparison of 2010 and 2023 reveals an increase in the strength of the main predictive factors and a shift in the importance of several of them. In 2010, the model was dominated by the *social_care_ratio* (+300.38 at 2.50) and *routes_ratio* (+213.06), with a notable but smaller effect from the *unemployment_ratio* (+124.48), while the *regon_entities_ratio* was even slightly negative (–10.31). In 2023, all major 2010 factors remained influential: the *social_care_ratio* (+577.18), *routes_ratio* (+458.16), and especially the *regon_entities_ratio* (+440.86 at 2205), indicating that business density became a key advantage for the city. At the same time, the *marriages_ratio* (+104.92) and *hotels_beds_ratio* (+72.29) gained significantly more importance than in 2010, while the effect of the *productive_population_ratio* decreased and became nearly neutral (+1.92). This shows that in 2023, compensation levels in Wrocław were primarily shaped by a combination of institutional-infrastructural strengths and high economic activity.

4. Conclusions

The results demonstrate that machine learning, when combined with interpretability methods, can capture the complexity of regional wage disparities beyond the scope of traditional econometric approaches. While labor market and demographic indicators consistently emerge as the strongest determinants, their relative importance evolves in response to broader socio-economic changes. The observed shifts between 2010 and 2023 underline the dynamic nature of regional wages formation, where structural conditions such as productive population ratios and enterprise density interact with local demographic and infrastructural contexts in non-linear ways.

Importantly, a SHAP-based analysis allows for a nuanced understanding of these dynamics by revealing how the same variable can contribute differently across powiats and time periods. This local interpretability enhances the practical value of predictive modeling for policymakers, offering insights that extend beyond aggregate associations. The findings suggest that data-driven approaches, when complemented with robust explanatory tools, provide not only accurate predictions but also meaningful guidance for regional development strategies aimed at mitigating wage inequalities.

References

Adamczyk, A., Tokarski, T., & Włodarczyk, R. W. (2009). Regional Wage Differences in Poland. *Gospodarka Narodowa. The Polish Journal of Economics*, *234*(9), 87–108. https://doi.org/10.33119/GN/101248.

Apley, D. W., & Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *82*(4), 1059–1086. https://doi.org/10.1111/rssb.12377.

Bartosik, K., & Mycielski, J. (2015). *Dynamika płac a długotrwałe bezrobocie w polskiej gospodarce* (INE PAN Working Paper Series, no 38). https://www.inepan.pl/images/pliki/Working_Papers/WorkingPapers_38.pdf.

Biecek, P., & Burzykowski, T. (2021). *Explanatory Model Analysis: Explore, Explain, and Examine Predictive Models*. CRC Press.

Blanchflower, D. G., & Oswald, A. J. (1990). The Wage Curve. *The Scandinavian Journal of Economics*, *92*(2), 215–235. https://doi.org/10.2307/3440026.

Bolińska, M., & Gomółka, A. (2018). Determinanty przestrzennego zróżnicowania płac w obwodach Ukrainy Zachodniej w latach 2004–2015. *Modern Management Review*, *23*, 31–44. https://doi.prz.edu.pl/pl/publ/zim/341.

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

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In B. Krishnapuram & M. Shah (Eds.), *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). https://doi.org/10.1145/2939672.2939785.

Combes, P.-P., Duranton, G., & Gobillon, L. (2008). Spatial Wage Disparities: Sorting Matters!. *Journal of Urban Economics*, *63*(2), 723–742. https://doi.org/10.1016/j.jue.2007.04.004.

Dykas, P., & Misiak, T. (2013). Determinanty przestrzennego zróżnicowania wybranych zmiennych makroekonomicznych. In M. Trojak & T. Tokarski (Eds.), *Statystyczna analiza przestrzennego zróżnicowania rozwoju ekonomicznego i społecznego Polski* (pp. 67–80). Wydawnictwo Uniwersytetu Jagiellońskiego.

Dykas, P., Misiak, T., & Tokarski, T. (2020). Determinants of spatial differentiation of labour markets in Ukraine. *Przegląd Statystyczny. Statistical Review*, *67*(1), 33–50. https://doi.org/10.5604/01.3001.0014.1784.

Fisher, A., Rudin, C., & Dominici, F. (2019). All Models are Wrong, but Many are Useful: Learning a Variable’s Importance by Studying an Entire Class of Prediction Models Simultaneously. *Journal of Machine Learning Research*, *20*, 1–18. https://jmlr.org/papers/volume20/18-760/18-760.pdf.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistic*s, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451.

Friedman, J. H., & Popescu, B. E. (2008). Predictive learning via rule ensembles. *The Annals of Applied Statistics*, *2*(3), 916–954. https://doi.org/10.1214/07-AOAS148.

Greenwell, B. M., Bradley, C. B., & McCarthy, A. J. (2018). *A simple and effective model-based variable importance measure*. https://doi.org/10.48550/arXiv.1805.04755.

Hooker, G. (2004). Discovering additive structure in black box functions. In W. Kim, R. Kohavi, J. Gehrke & W. DuMouchel, *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 575–580). Association for Computing Machinery. https://doi.org/10.1145/1014052.1014122.

Hooker, G. (2007). Generalized functional ANOVA diagnostics for high-dimensional functions of dependent variables. *Journal of Computational and Graphical Statistics, 16*(3), 709–732. https://doi.org/10.1198/106186007X237892.

Kaliski, S. F. (1964). The Relation Between Unemployment and the Rate of Change of Money Wages in Canada. *International Economic Review, 5*(1), 1–33. https://doi.org/10.2307/2525631.

Kapela, M., & Kwiatkowski, E. (2023). Regional Wage Differentiation and Qualitative Determinants of Economic Development: Evidence from Poland. *Zeszyty Naukowe Uniwersytetu Ekonomicznego w Krakowie. Cracow Review of Economics and Management*, (3), 47–65. https://doi.org/10.15678/ZNUEK.2023.1001.0303.

Kingdon, G. G., & Knight, J. (2006). How Flexible Are Wages in Response to Local Unemployment in South Africa?. *ILR Review, 59*(3), 471–495. https://doi.org/10.1177/001979390605900308.

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (pp. 4765–4774). https://papers.nips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

Luśtyk, A., Połeć, A., & Voznyuk, I. (2024). Wage Differences in Poland at the County Level and their Determinants. *Central European Economic Journal*, *11*(58), 447–460. https://doi.org/10.2478/ceej-2024-0028.

Machuca, J. A. L., & Cota, J. E. M. (2017). Salarios, desempleo y productividad laboral en la industria manufacturera mexicana. *Ensayos Revista de Economía, 36*(2), 185–228. https://ensayos.uanl.mx/index.php/ensayos/issue/view/10/17.

Masís, S. (2023). *Interpretable machine learning with Python*. Packt Publishing.

Molnar, C. (2020). *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. Leanpub. https://originalstatic.aminer.cn/misc/pdf/Molnar-interpretable-machine-learning_compressed.pdf.

Moretti, E. (2011). Local Labor Markets. In O. Ashenfelter & D. Card (Eds.), *Handbook of Labor Economics* (Vol. 4B, pp. 1237–1313). Elsevier. https://doi.org/10.1016/S0169-7218(11)02412-9.

Mullainathan, S., & Spiess, J. (2017). Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives, 31*(2), 87–106. https://doi.org/10.1257/jep.31.2.87.

Phillips, A. W. (1958). The Relation Between Unemployment and the Rate of Change of Money Wage Rates in the United Kingdom, 1861–1957. *Economica*, *25*(100), 283–299. https://doi. org/10.1111/j.1468-0335.1958.tb00003.x.

Przekota, G. (2016). Ocena poziomu i przyczyn zróżnicowania wynagrodzeń w Polsce. *Roczniki Ekonomiczne Kujawsko-Pomorskiej Szkoły Wyższej w Bydgoszczy*, (9), 386–403. https://kpsw.edu.pl/pobierz/wydawnictwo/re9/przekota2.pdf.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?". Explaining the Predictions of Any Classifier. *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778.

Shapley, L. S. (1953). A value for n-person games. In H. W. Kuhn, A. & W. Tucker (Eds.), *Contributions to the Theory of Games* (Vol. 2, pp. 307–317). Princeton University Press. https://doi.org/10.1515/9781400881970-018.

Štrumbelj, E., & Kononenko, I. (2010). An Efficient Explanation of Individual Classifications using Game Theory. *Journal of Machine Learning Research*, *11*(1), 1–18. https://doi.org/10.1145/1756006.1756007.

Štrumbelj, E., & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, *41*(3), 647–665. https://doi.org/10.1007/s10115-013-0679-x.

Appendix 1.

The code used in the research study is presented below. The data were acquired directly from BDL through the webservices. To repeat the analysis for years other than 2010 and 2023 (assuming that data are available in the repository for the chosen years), the only line that requires change is 'for YEAR in [2010,2023]:'.

```
import requests
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import shap
import dalex as dx

from sklearn.model_selection import train_test_split,
RepeatedKFold, cross_validate
from sklearn.preprocessing import StandardScaler
# from sklearn.neural_network import MLPRegressor
from sklearn.inspection import PartialDependenceDisplay
from sklearn.metrics import r2_score, mean_squared_error
from xgboost import XGBRegressor

base_url = "https://bdl.stat.gov.pl/api/v1/data/by-
variable/"
params = {
 'format': 'jsonapi',
 'unit-level': 5,
 'page-size': 100,
}

def get_data_by_variable(variable_id, variable_name,
year):
 ids = []
 values = []

 for page in range(4):
 params['page'] = page
 params['year'] = year
 response = requests.get(f"{base_url}{variable_id}",
params=params)
 data = response.json()

 for item in data['data']:
 attributes = item['attributes']
 id_ = item['id']
 val_data = attributes['values']

 if val_data:
 val = val_data[0]['val']
 ids.append(id_)
 values.append(val)

 return pd.DataFrame({variable_name: values}, index=ids)


for YEAR in [2010, 2023]:
 df_vars = {
 64428: 'compensation_level',
 60530: 'regon_entities_ratio',
 458700: 'social_care_ratio',
 395404: 'routes_ratio',
 60270: 'unemployment_ratio',
 60566: 'productive_population_ratio',
 450551: "birthrate",
 1646059: "forests_ratio",
 60300: "hotels_beds_ratio",
 450543: "marriages_ratio"
 }

 df = None
 for key, val in df_vars.items():
 df_current = get_data_by_variable(key, val, YEAR)
 if df is None:
 df = df_current
 else:
 df = df.join(df_current)

 # Basic dataset summary
 X = df.drop(columns=['compensation_level'])
 y = df['compensation_level']
 n_obs, n_features = X.shape
 print(f"\n=== YEAR {YEAR} ===")
 print(f"Number of observations: {n_obs}")
 print(f"Number of predictors: {n_features}")
 print(f"Observation-to-predictor ratio: {n_obs / n_features:.2f}")

 # Train-test split BEFORE scaling to avoid leakage
 X_train, X_test, y_train, y_test = train_test_split(
 X, y, test_size=0.05, random_state=42
 )

 xgb = XGBRegressor(
 n_estimators=200,
 learning_rate=0.01,
 max_depth=3,
 subsample=0.8,
 colsample_bytree=0.8,
 reg_lambda=1.0,
 random_state=42,
 objective="reg:squarederror"
 )

 cv = RepeatedKFold(n_splits=5, n_repeats=5, random_state=42)
 cv_results = cross_validate(
 xgb,
 X_train,
 y_train,
 cv=cv,
 scoring=['r2', 'neg_root_mean_squared_error'],
 return_train_score=True
 )

 print(f"CV mean test R²:
{np.mean(cv_results['test_r2']):.3f}")
 print(f"CV mean test RMSE: {-
np.mean(cv_results['test_neg_root_mean_squared_error']):
.3f}")

 xgb.fit(X_train, y_train)
 y_pred_train = xgb.predict(X_train)
 y_pred_test = xgb.predict(X_test)
 train_r2 = r2_score(y_train, y_pred_train)
 test_r2 = r2_score(y_test, y_pred_test)
 train_rmse = mean_squared_error(y_train,
y_pred_train)**.5
 test_rmse = mean_squared_error(y_test, y_pred_test)**.5

 print(f"Train R²: {train_r2:.3f}, RMSE:
{train_rmse:.3f}")
 print(f"Test R²: {test_r2:.3f}, RMSE: {test_rmse:.3f}")

 model_explainer = dx.Explainer(xgb, X, y,
label=f"XGB_{YEAR}")
 vi = model_explainer.model_parts(N=10000,
random_state=42)
 print(vi.result)
 fig_vi = vi.plot(show=False, title=f"XGB VIP explainer
for compensation level in year {YEAR}")
 fig_vi.write_image(f"vip_plot_{YEAR}.png")
 # SHAP analysis
 explainer = shap.Explainer(xgb.predict, X,
feature_names=X.columns)
 shap_values = explainer(X)

 # Jelenia Góra
 index_jg = df.index.get_loc("030210161000")
 shap_df_jg = pd.DataFrame({
 'Feature': X.columns,
 'SHAP Value': shap_values[index_jg].values,
 'Feature Value': shap_values[index_jg].data
 })
 print("SHAP values for county Jelenia Góra
(030210161000):")
 print(shap_df_jg.sort_values(by='SHAP Value', key=abs,
ascending=False).head(10))

 plt.figure(figsize=(18, 6))
 plt.suptitle(f"Compensation Level Model SHAP values for
Jelenia Góra for year {YEAR}")
 shap.plots.bar(shap_values[index_jg], max_display=10,
show=False, show_data=True)
 plt.tight_layout(pad=1.0)
 plt.savefig(f"Jelenia_shap_{YEAR}.jpg")
 plt.show()

 # Wrocław
 index_wr = df.index.get_loc("030210564000")
 shap_df_wr = pd.DataFrame({
 'Feature': X.columns,
 'SHAP Value': shap_values[index_wr].values,
 'Feature Value': shap_values[index_wr].data
 })
 print("SHAP values for county Wrocław (030210564000):")
 print(shap_df_wr.sort_values(by='SHAP Value', key=abs,
ascending=False).head(10))

 plt.figure(figsize=(12, 6))
 shap.plots.bar(shap_values[index_wr], max_display=10,
show=False, show_data=True)
 plt.suptitle(f"Compensation Level Model SHAP values for
Wrocław for year {YEAR}")
 plt.tight_layout(pad=1)
 plt.savefig(f"Wroclaw_shap_{YEAR}.jpg")
 plt.show()

 # PDP plots (using scaled data from final model)
 fig, ax = plt.subplots(figsize=(12, 8))
 PartialDependenceDisplay.from_estimator(
 xgb,
 X,
 features=list(range(X.shape[1])),
 feature_names=X.columns,
 ax=ax
 )
 plt.suptitle(f"Partial dependency plots for compensation
level model for year {YEAR}")
 plt.tight_layout()
 plt.savefig(f"PDP_{YEAR}.jpg")
 plt.show()
```

Appendix 2.

The full results obtained after the execution of the code presented in Appendix 1 are as follows:

```
=== YEAR 2010 ===
Number of observations: 379
Number of predictors: 9
Observation-to-predictor ratio: 42.11
CV mean test R²: 0.299
CV mean test RMSE: 340.347
Train R²: 0.629, RMSE: 254.324
Test R²: 0.369, RMSE: 192.135
Preparation of a new explainer is initiated

 -> data : 379 rows 9 cols
 -> target variable : Parameter 'y' was a pandas.Series. Converted to
a numpy.ndarray.
 -> target variable : 379 values
 -> model_class : xgboost.sklearn.XGBRegressor (default)
 -> label : XGB_2010
 -> predict function : <function yhat_default at 0x000002C41C6A75B0>
will be used (default)
 -> predict function : Accepts pandas.DataFrame and numpy.ndarray.
 -> predicted values : min = 2.72e+03, mean = 2.89e+03, max = 4.11e+03
 -> model type : regression will be used (default)
 -> residual function : difference between y and yhat (default)
 -> residuals : min = -6.88e+02, mean = 5.32, max = 1.9e+03
 -> model_info : package xgboost

A new explainer has been created!
 variable dropout_loss label
0 _full_model_ 251.572326 XGB_2010
1 hotels_beds_ratio 254.133193 XGB_2010
2 birthrate 254.864096 XGB_2010
3 marriages_ratio 259.696907 XGB_2010
4 forests_ratio 260.088788 XGB_2010
5 regon_entities_ratio 261.499846 XGB_2010
6 unemployment_ratio 287.919198 XGB_2010
7 productive_population_ratio 298.164109 XGB_2010
8 routes_ratio 313.799260 XGB_2010
9 social_care_ratio 320.226454 XGB_2010
10 _baseline_ 468.653795 XGB_2010
ExactExplainer explainer: 380it [00:52, 7.25it/s]
SHAP values for county Jelenia Góra (030210161000):
 Feature SHAP Value Feature Value
2 routes_ratio 107.268913 24.40
3 unemployment_ratio 26.644887 10.90
1 social_care_ratio 19.697117 5.90
```

```
4 productive_population_ratio -14.265568 64.70
0 regon_entities_ratio 9.167773 1499.00
5 birthrate 4.711150 -3.83
6 forests_ratio -2.413097 52.90
7 hotels_beds_ratio -1.588051 22.24
8 marriages_ratio -0.335435 5.40
SHAP values for county Wrocław (030210564000):
 Feature SHAP Value Feature Value
1 social_care_ratio 300.382093 2.50
2 routes_ratio 213.063043 12.20
3 unemployment_ratio 124.481779 5.50
4 productive_population_ratio -16.713112 65.80
8 marriages_ratio 14.444688 5.60
0 regon_entities_ratio -10.308929 1608.00
5 birthrate 5.582377 0.25
7 hotels_beds_ratio 2.794502 13.95
6 forests_ratio 0.744744 13.80

=== YEAR 2023 ===
Number of observations: 380
Number of predictors: 9
Observation-to-predictor ratio: 42.22
CV mean test R²: 0.251
CV mean test RMSE: 718.596
Train R²: 0.564, RMSE: 562.235
Test R²: -0.034, RMSE: 527.064
Preparation of a new explainer is initiated

 -> data : 380 rows 9 cols
 -> target variable : Parameter 'y' was a pandas.Series. Converted to
a numpy.ndarray.
 -> target variable : 380 values
 -> model_class : xgboost.sklearn.XGBRegressor (default)
 -> label : XGB_2023
 -> predict function : <function yhat_default at 0x000002C41C6A75B0>
will be used (default)
 -> predict function : Accepts pandas.DataFrame and numpy.ndarray.
 -> predicted values : min = 4.5e+03, mean = 6.56e+03, max = 9.46e+03
 -> model type : regression will be used (default)
 -> residual function : difference between y and yhat (default)
 -> residuals : min = -4.5e+03, mean = 2.39, max = 3.35e+03
 -> model_info : package xgboost

A new explainer has been created!
 variable dropout_loss label
0 _full_model_ 560.529352 XGB_2023
1 birthrate 563.920908 XGB_2023
2 marriages_ratio 569.890669 XGB_2023
3 unemployment_ratio 579.918983 XGB_2023
4 productive_population_ratio 580.117875 XGB_2023
5 forests_ratio 587.603146 XGB_2023
6 hotels_beds_ratio 591.842447 XGB_2023
7 routes_ratio 642.313577 XGB_2023
8 regon_entities_ratio 654.331861 XGB_2023
9 social_care_ratio 676.012299 XGB_2023
10 _baseline_ 935.037314 XGB_2023
ExactExplainer explainer: 381it [00:34, 8.13it/s]
```

```
SHAP values for county Jelenia Góra (030210161000):
 Feature SHAP Value Feature Value
0 regon_entities_ratio 168.111842 1815.00
4 productive_population_ratio 86.101887 54.90
2 routes_ratio 73.077743 27.90
3 unemployment_ratio 40.129574 3.90
1 social_care_ratio -17.931182 2.70
7 hotels_beds_ratio 9.112820 35.70
6 forests_ratio -7.528812 59.10
5 birthrate 5.067187 -8.75
8 marriages_ratio 0.236881 3.70
SHAP values for county Wrocław (030210564000):
 Feature SHAP Value Feature Value
1 social_care_ratio 577.179101 1.10
2 routes_ratio 458.159322 14.20
0 regon_entities_ratio 440.862895 2205.00
3 unemployment_ratio 112.970279 1.60
8 marriages_ratio 104.923334 5.80
7 hotels_beds_ratio 72.285200 21.63
5 birthrate 12.149382 -1.77
6 forests_ratio -7.046792 15.20
4        productive_population_ratio         1.921097         60.804
productive_population_ratio 3.846133 2.033800
```