DOI: 10.5281/zenodo.18892764

Model Transparency: Integrating XGBoost with SHAP for Explainable Machine Learning

Andrej Bednařík[1]

Abstract
This paper examines integrating the XGBoost algorithm with SHAP values to balance predictive performance and model interpretability. XGBoost is widely recognized for its high accuracy and efficiency, yet its ensemble structure makes the internal decision-making process difficult to interpret. SHAP offers a theoretically grounded framework based on Shapley values that enables both global and local explanations of model behavior. The paper focuses on identifying key predictors using SHAP summary analysis, exploring variable interactions, and providing detailed explanations of individual predictions through local SHAP visualizations. The results show that combining XGBoost with SHAP creates a robust and transparent modeling framework suitable for domains where explainability is essential. Moreover, SHAP uncovers complex feature relationships that traditional feature-importance methods miss, thereby improving the overall interpretive value of the model.

Keywords
XGBoost; SHAP; model interpretability; feature importance; machine learning

1 Introduction

In recent years, the field of machine learning has increasingly emphasized not only achieving strong predictive performance but also ensuring that models are interpretable and transparent. In application domains such as banking, insurance, and healthcare, understanding why a model makes a particular decision is critical for accountability, regulatory compliance, and user trust. However, the highest predictive accuracy is often achieved by complex models, most notably tree-based ensemble methods, which can obscure the underlying decision mechanism and create a black-box effect. One of the most widely used tools in practice is XGBoost, a powerful gradient-boosting framework capable of handling large-scale data and delivering excellent predictive performance (Chen and Guestrin, 2016). At the same time, XGBoost typically relies on an ensemble of hundreds or even thousands of decision trees, which makes direct interpretation of decision paths difficult. For this reason, in recent years, methods and tools have been developed to make such high-performing models more explainable. A prominent approach is SHapley Additive exPlanations, or SHAP, which is grounded in cooperative game theory and assigns each input feature a contribution to a given model output. SHAP provides a unified framework for explaining predictions, satisfying desirable properties such as local accuracy and symmetry, and enabling consistent comparison across explanation techniques (Lundberg and Lee, 2017). In practice, simpler classical feature importance measures are still commonly used for XGBoost, such as Gain, Cover, or Weight, which produce global rankings of feature relevance. These measures, however, may overlook feature interactions, nonlinear relationships, and the distinction between global and local importance. In contrast, SHAP supports both global and instance-level explanations and can reveal patterns that classical importance metrics may fail to capture (Ponce-Bobadilla et al., 2024). Accordingly, the focus of this article, *Model Transparency: Integrating*

[1]Department of Mathematics and Actuarial Science, Faculty of Economic Informatics, Bratislava University of Economics and Business, Slovakia. Email: andrej.bednarik@euba.sk

XGBoost with SHAP for Explainable Machine Learning, is to integrate predictive performance with interpretability by combining XGBoost with SHAP-based explanations. The goal is to evaluate how SHAP complements or differs from traditional XGBoost feature importance analysis.

2 A search for sources and the current state of the problem

To position this article within the current research landscape, the literature search focused on peer-reviewed and widely cited work on explainable machine learning for tree ensembles, with emphasis on XGBoost and SHAP (Barredo Arrieta et al., 2020). The search strategy combined queries in Scopus, Web of Science, ACM Digital Library, IEEE Xplore, and Google Scholar using keyword sets such as XGBoost, gradient boosted trees, explainable machine learning, explainable AI, SHAP, Shapley values, TreeSHAP, feature importance, and post hoc explanations (Angelov et al., 2021). Priority was given to foundational methodological papers, systematic reviews, and domain-focused surveys in high-impact application areas where transparency is required, especially finance and healthcare (Weber et al., 2024). Across the broader explainability literature, there is strong agreement that predictive performance alone is insufficient in many real-world deployments, because stakeholders often need to understand and justify model behavior (Doshi-Velez and Kim, 2017). This need is especially visible in high-stakes domains, where explanations support accountability, auditing, and error analysis (Barredo Arrieta et al., 2020). Survey papers provide structured taxonomies of explainability and clarify common distinctions such as global versus local explanations and model-specific versus model-agnostic methods (Guidotti et al., 2019). At the same time, critical perspectives argue that explanations do not automatically guarantee safety or validity, and in some settings, inherently interpretable models can be preferable to explaining complex black box systems (Rudin, 2019). This tension motivates the practical focus of this article, namely how to combine high-performing tree boosting with explanation methods in a way that is useful and methodologically defensible (Doshi-Velez and Kim, 2017).

Within this space, SHAP is one of the most influential frameworks for feature attribution because it is grounded in cooperative game theory and uses a consistent additive form to explain individual predictions (Lundberg and Lee, 2017). For tree-based models, later work introduced efficient algorithms and tooling that make Shapley-style explanations feasible at scale and enable moving from local explanations to global understanding (Lundberg, Erion, et al., 2020). These contributions are particularly relevant for XGBoost, a widely used gradient boosting implementation known for its strong accuracy and scalability (Chen and Guestrin, 2016). In practice, however, many projects still rely on built-in XGBoost importance measures such as gain, cover, or split counts (Chen and Guestrin, 2016). Such summaries can miss interactions, nonlinearities, and differences across instances, leading to a partial or misleading narrative about what drives model outputs (Guidotti et al., 2019). This creates an applied gap because users may obtain different conclusions depending on whether they interpret the model with built-in importance or with SHAP-based attributions (Barredo Arrieta et al., 2020).

Recent research also clarifies limitations that shape the current state of the problem. SHAP attributions can be sensitive to feature dependence because common implementations may rely on assumptions that are violated when predictors are correlated, thereby altering how contributions are allocated (Aas et al., 2021). Other studies question whether Shapley-based importance should be treated as a human-centered explanation rather than a mathematical attribution, noting that stronger explanatory claims often require additional assumptions and, in some cases, causal framing (Kumar et al., 2020). There is also evidence that post hoc explanation methods can be manipulated under certain threat models, which is important when explanations are used for compliance or trust rather than primarily for debugging (Slack et al., 2020). Together, these findings support the view that explainability should be treated as a careful methodological layer rather than a simple visualization add-on (Guidotti et al., 2019).

From an application perspective, systematic reviews and bibliometric studies show rapid

growth of explainable machine learning in regulated sectors (Sharma et al., 2024). In finance, a systematic review documents diverse explainability goals and methods and links adoption to increasing expectations for transparency in risk-sensitive decision support (Weber et al., 2024). In healthcare, bibliometric analysis reports a rapid expansion of empirical work on explainable AI and highlights a concentration of output in leading research countries and its frequent use in prediction and diagnostic settings (Dhiman et al., 2023). These trends indicate that integrating high-performance models, such as XGBoost, with explanation frameworks, such as SHAP, aligns with current research priorities and real-world deployment needs worldwide (Barredo Arrieta et al., 2020).

3 The main findings of the article

The task was formulated as supervised regression, where the target variable was the insurance premium (poistne_pzp). The dataset contained approximately 124,000 records and combined numeric predictors, categorical variables, and date or time information. Before model training, the dataset underwent a structured data processing workflow to improve data quality, reduce noise, and ensure that all variables were represented in a model-compatible form. This preprocessing step included basic consistency checks, handling of missing or invalid entries, and verification of ranges for key numeric variables to prevent extreme values from dominating the learning process. Categorical predictors were standardized to remove label inconsistencies and encoded in a format suitable for gradient-boosted trees, while numeric variables were retained at their original scales because tree-based methods do not require normalization. Date and time variables were transformed into usable representations, such as extracted years or derived duration-based quantities, enabling temporal signals to be incorporated without relying on raw timestamps. Following preprocessing, the final feature set was constructed from a combination of contract-related factors, vehicle attributes, and regional or segment descriptors, namely Datum_pociatku, Datum_storna, Druh_auta, Frekvencia_platenia, Hmotnost, Kanal, Kategoria_vozidla_popis, Objem, Sposob_pouzitia_auta, Vykon, Znacka_vozidla, bonus_malus, okres, vek, and vek_vozidla.

The goal was not only to achieve strong predictive performance in ex post analysis, but also to ensure that the resulting model could be interpreted transparently. For this reason, the modeling workflow combined performance evaluation with a systematic interpretability layer. After training, predictive accuracy was assessed using standard regression metrics, including RMSE, MAE, and R-squared, to capture both typical prediction error and sensitivity to large deviations. In parallel, model explanations were generated using SHAP values, which decompose each prediction into an additive combination of feature contributions relative to a baseline expectation. This enabled two complementary perspectives on the model: a global view describing which variables tend to matter most across the portfolio, and a local view explaining why a particular policy receives a higher or lower predicted premium. In practical terms, this integration supports model validation and communication by enabling analysts to compare classical XGBoost importance rankings with SHAP-based attributions and to identify nonlinear effects, thresholds, and segment-specific behavior that are common in insurance pricing data.

3.1 What is SHAP

To overcome the limitations of classical feature importance measures in XGBoost-type models, such as weight and gain, SHAP has emerged as an advanced approach. SHAP is grounded in cooperative game theory and assigns each input feature an individual share of the model prediction for a specific observation. As a result, it provides not only a global overview of importance but also the ability to explain predictions locally. A key advantage of SHAP is that it satisfies properties such as symmetry, efficiency, and additivity, which makes it a theoretically

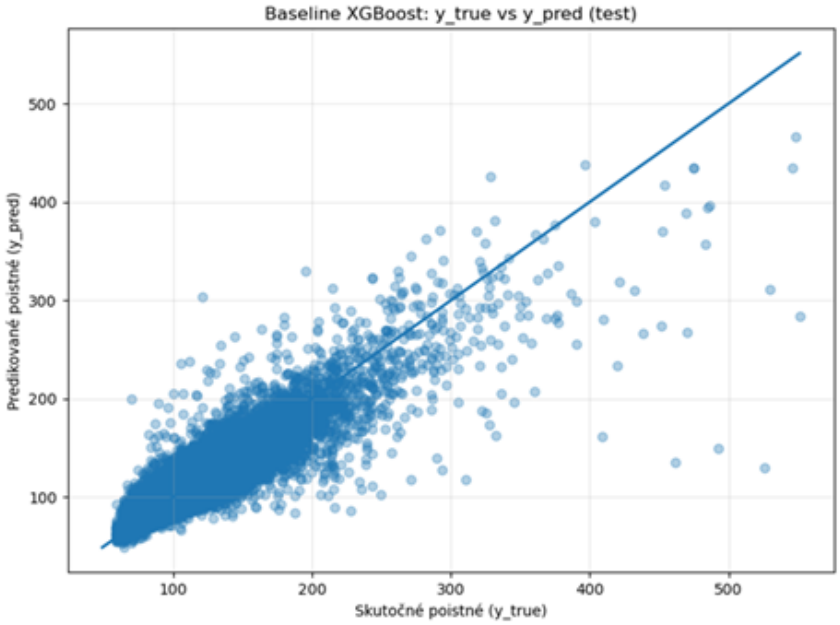

Figure 1. Gain-based feature importance ranking (schematic).

well-founded framework for determining feature importance (Lundberg and Lee, 2017). On a practical level, this means that for every instance in the data, we can obtain a set of SHAP values showing how each variable contributed to shifting the prediction from the baseline value to the final output. SHAP therefore enables the identification of high-impact variables, the determination of whether their contributions are positive or negative, and the detection of interactions or nonlinear effects that may be masked by simpler metrics. In the context of XGBoost, this means that while metrics such as gain or weight provide a certain perspective, SHAP offers a deeper, more detailed interpretation of model behavior across data segments and at the level of individual predictions. From a technical standpoint, SHAP values rely on evaluating all possible combinations of input variables, or an approximation of these combinations, in order to estimate the marginal contribution of each feature, which supports consistent and fair attribution (Lundberg and Lee, 2017). However, using SHAP also incurs higher computational costs and requires careful interpretation. In particular, large SHAP values for a feature do not automatically imply causality; rather, they indicate a strong associative relationship within the trained model. For this reason, integrating SHAP-based interpretations should be accompanied by a thorough understanding of the data, domain requirements, and the broader modeling context.

3.2 Predictive performance of the XGBoost regression model

Using the selected feature set, an XGBoost regression model was trained to predict premium values. Performance was assessed on a held-out test set using standard regression metrics. RMSE captures sensitivity to larger errors, MAE provides an interpretable measure of typical absolute deviation, and R-squared summarizes the proportion of variance explained. The final test performance achieved RMSE = 16.63, MAE = 10.58, and $R^2 = 0.805$. These results confirm that gradient boosted tree ensembles are well-suited for premium prediction in heterogeneous tabular insurance data and can model nonlinear relationships between risk drivers and premium levels.

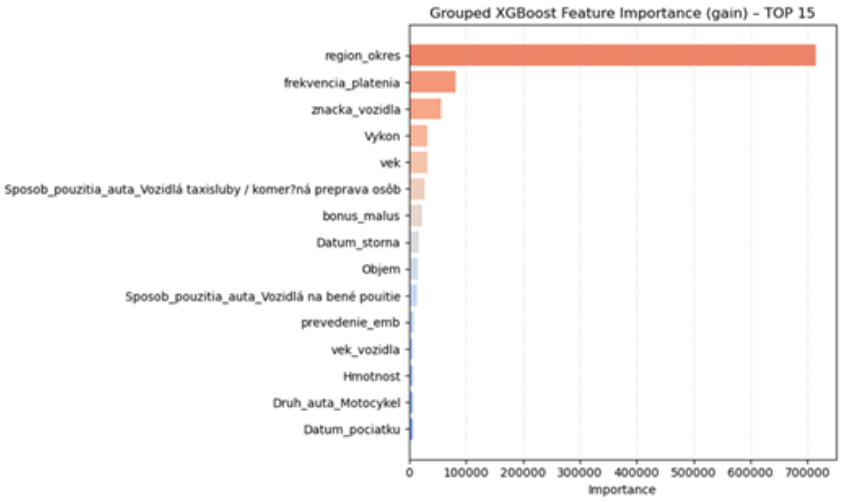

Figure 2. SHAP global importance ranking (mean absolute SHAP) (schematic).

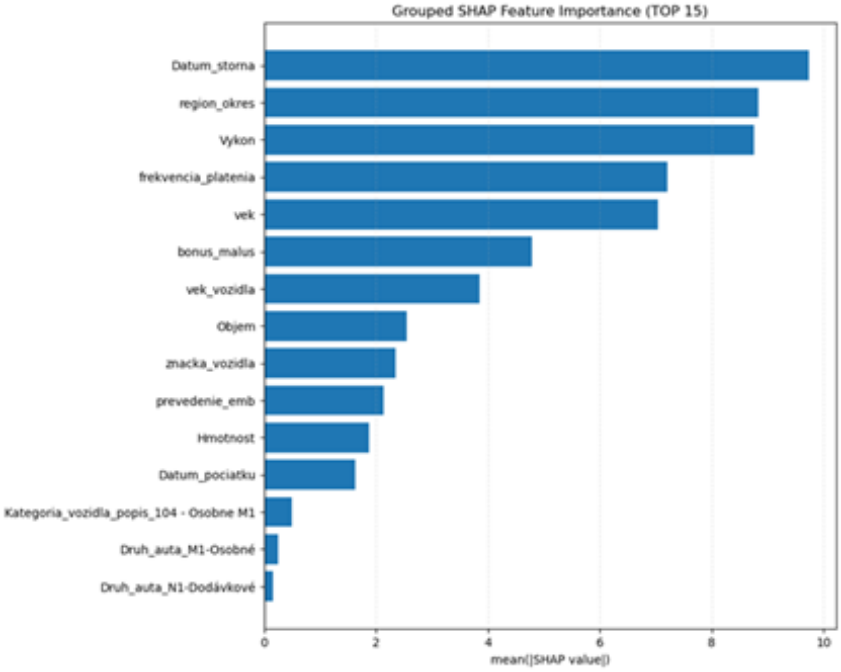

Figure 3. SHAP beeswarm plot (schematic).

3.3 Global importance differences: gain vs SHAP (full text from the paper)

The SHAP-based ranking can differ substantially from the gain-based feature importance because the two metrics capture different concepts: gain reflects how much a feature improves the objective when used in splits during training, whereas mean absolute SHAP reflects how much the feature actually shifts predictions across all observations. As a result, a variable can have high gain due to a small number of very strong splits yet show lower SHAP importance if it affects only a narrow subgroup. Conversely, a variable can have a moderate gain but high SHAP importance if it consistently shifts across many policies. In the context of this study, which is explicitly framed as an ex post analysis, the dominance of Datum_storna is therefore interpreted as evidence that the model captures a strong policy lifecycle and temporal signal present in the historical data-generating process, and that it meaningfully differentiates premium outcomes across the portfolio.

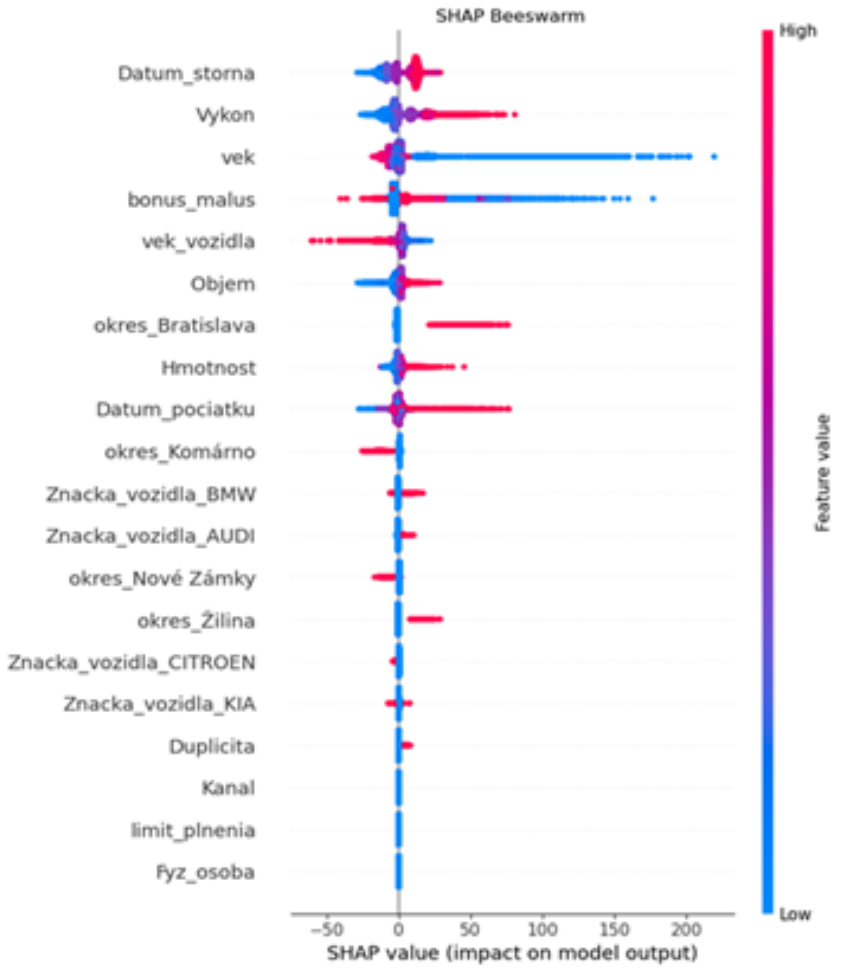

Figure 4. SHAP dependence: bonus_malus vs Vykon (schematic).

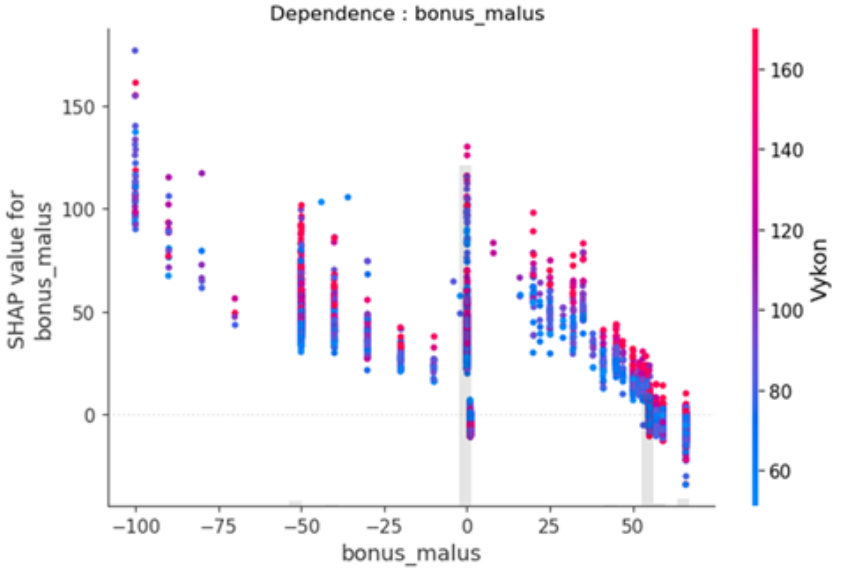

Figure 5. SHAP dependence: vek vs bonus_malus (schematic).

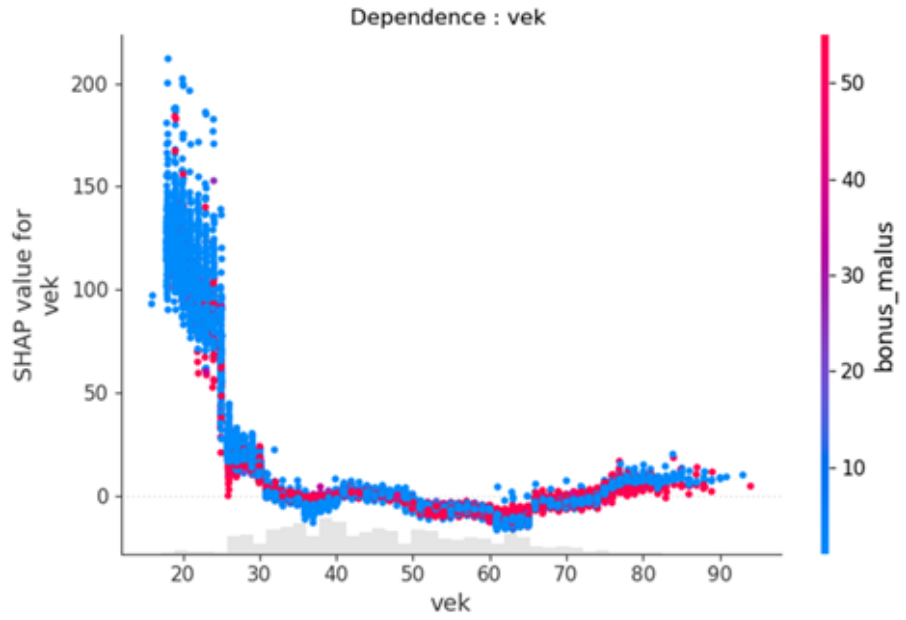

Figure 6. Local interpretability: SHAP waterfall plot (schematic).

3.4 SHAP beeswarm plot (full text from the paper)

Figure **??** presents a SHAP beeswarm plot providing a global summary of feature effects by displaying the distribution of SHAP values for each variable across all observations. Features are ordered from top to bottom by overall importance, typically measured by the mean absolute SHAP value, so variables at the top have the largest average impact on the predicted premium. Each point represents one policy in the dataset, positioned on the x-axis according to its SHAP value, which indicates how much that feature pushes the prediction above or below the model baseline. The color encodes the actual feature value, with blue indicating low values and red indicating high values, allowing the direction of the relationship to be interpreted visually.

The plot confirms that the strongest global drivers are Datum_storna, Vykon, and vek. For Datum_storna, higher values are associated with positive SHAP contributions, meaning that later cancellation-related values tend to increase predicted premiums relative to the baseline, while lower values tend to have smaller or negative effects. For Vykon, the color pattern shows that higher engine power values are concentrated on the positive SHAP side, indicating that more powerful vehicles systematically increase predicted premiums, whereas lower power values are associated with negative or near-zero contributions. For vek, the spread is asymmetric, with younger ages predominantly on the positive SHAP side, indicating that younger customers receive higher predicted premiums, while older ages tend to cluster near zero or slightly negative values, consistent with the dependence plot, which shows the age effect becoming milder after early adulthood.

For bonus_malus, the beeswarm shows a wide spread of SHAP contributions, indicating that this variable can strongly shift predictions in both directions depending on its level. The color distribution suggests that different bonus malus values correspond to distinct premium adjustments rather than a single smooth trend, which aligns with the discrete bands seen in the dependence plot and reflects stepwise segmentation learned by the tree ensemble. The feature vek_vozidla also shows a clear directional pattern, where higher vehicle age tends to be associated with negative SHAP values, indicating that older vehicles reduce the predicted premium relative to the baseline in this dataset, while newer vehicles contribute positively. Objem and Hmotnost contribute with smaller but still visible effects, and their spreads indicate that these technical vehicle characteristics matter primarily in certain ranges rather than uniformly

across all observations.

Several one-hot-encoded regional indicators, such as okres_Bratislava, okres_Komárno, okres_Nové Zámky, and okres_Žilina, are identified as important contributors, confirming that geographic segmentation is embedded in the model and that specific districts systematically shift predicted premiums. Because these are binary indicators, the color typically distinguishes between the presence and absence of the category, and the horizontal spread reflects the magnitude of the premium shift when the policy belongs to that district. Brand indicators such as Znacka_vozidla_BMW, Znacka_vozidla_AUDI, Znacka_vozidla_CITROEN, and Znacka_vozidla_KIA appear lower in the ranking, with narrower SHAP spreads, suggesting that brand effects are present but less pronounced than the dominant drivers such as cancellation timing, power, age, and region. Finally, variables such as Duplicita, Kanal, limit_plnenia, and Fyz_osoba show SHAP values clustered tightly around zero, indicating minimal average contribution to the model output, either because their effect is weak in this dataset or because their information is largely captured by other correlated predictors.

3.5 Nonlinear effects and segment-specific behavior revealed by SHAP

A key advantage of SHAP is that it supports deeper diagnostics beyond a single ranking. Dependence plots can be used to examine how predicted premiums change across the range of a variable and to identify thresholds or nonlinearities typical for tree-based models. For example, continuous vehicle attributes such as Objem, Hmotnost, and Vykon often exhibit piecewise patterns, with their impact increasing after certain breakpoints. Similarly, contract-related variables such as Frekvencia_platenia can shift predicted premiums in a structured way that reflects the learned segmentation of payment patterns. Categorical variables such as Znacka_vozidla, Druh_auta, Kategoria_vozidla_popis, or Kanal can be investigated through the distribution of their SHAP contributions, which highlights whether certain categories systematically increase or decrease premiums.

3.6 Bonus_malus vs Vykon (full text from the paper)

Figure 4 presents a SHAP dependence plot illustrating how bonus_malus affects the model prediction through its SHAP contribution. The x-axis shows the bonus malus value, while the y-axis shows the SHAP value for bonus_malus, interpreted as the marginal impact of bonus malus on the predicted premium relative to the model baseline. Positive SHAP values indicate that the given bonus malus level increases the predicted premium, whereas values near zero or negative indicate little effect or a decrease relative to the baseline. Points are colored by Vykon (engine power), which helps reveal whether the effect of bonus malus changes systematically with vehicle power.

A clear overall pattern is visible: as bonus_malus increases into higher positive values, the SHAP contribution generally declines toward zero and can even become negative for the highest levels. This means the model associates higher bonus malus values in this dataset with smaller premium increases, and in some cases with a reduction relative to the baseline. In contrast, strongly negative bonus malus values are associated with large positive SHAP values, indicating substantial premium increases for those policies. The plot also shows distinct vertical bands, suggesting that bonus_malus takes on discrete or heavily clustered values in the dataset, so the model learns stepwise adjustments rather than a smooth, continuous relationship.

The coloring by Vykon indicates that engine power contributes additional differentiation within the same bonus malus levels. For many bonus malus values, observations with higher Vykon tend to have higher SHAP values than those with lower power, suggesting that the model combines risk experience captured by bonus malus with vehicle performance when setting premiums. This interaction is particularly visible in the mid-range of bonus malus values, where there is noticeable spread in SHAP contributions at the same x value, and the color gradient

indicates that part of this spread is explained by differences in engine power. Overall, the figure demonstrates that the model treats bonus malus as a strong, structured risk signal, but its effect is not purely linear and is moderated by other risk-related variables, such as vehicle power.

3.7 Vek vs Bonus_malus (full text from the paper)

Figure 5 presents a SHAP dependence plot showing how the feature vek (customer age) influences the model prediction through its SHAP contribution. The x-axis represents age, while the y-axis shows the SHAP value for vek, interpreted as the marginal effect of age on the predicted premium relative to the model baseline: positive SHAP values indicate that age increases the predicted premium, while negative SHAP values indicate that age decreases it. The points are colored by bonus_malus, which allows the plot to reveal potential interaction effects between age and the bonus malus level.

The most prominent pattern is a very large positive SHAP contribution at young ages, roughly 28-30, where SHAP values are strongly positive and highly dispersed, in some cases exceeding 200. This indicates that the model assigns substantially higher predicted premiums to younger policyholders and that the effect varies across individuals, suggesting the presence of additional interacting factors. Immediately after this region, there is a sharp drop, and from approximately the early 30s onward, the SHAP values cluster close to zero, indicating that, for the majority of middle-aged policyholders, age alone contributes only modestly to premium adjustments relative to other drivers.

A second, weaker pattern appears at higher ages, where SHAP values become slightly positive again in older segments, indicating that the model begins to increase predicted premiums for elderly policyholders, although the magnitude is far smaller than for very young customers. Between roughly 40 and 65, many points fall slightly below zero, suggesting a mild discount effect for these ages relative to the baseline, consistent with a lower-risk segment identified by the model. The color gradient indicates how bonus_malus interacts with age. While the overall shape of the age effect is present across the full range of bonus malus values, the dispersion and extremes in the youngest segment show that bonus malus levels contribute to additional differentiation inside that group. In practice, this means the model does not treat young age as a single category; instead, it combines age with a bonus-malus to produce more granular premium adjustments, a typical behavior of tree-based ensembles that capture nonlinearities and interactions.

3.8 Local explanations of individual premium predictions

To support transparency on the level of single policies, local explanations were produced for selected observations using SHAP waterfall-style decompositions. Each local explanation breaks down a predicted premium into a baseline expected value and feature-specific contributions that push the prediction upward or downward. This makes it possible to explain why two policies with similar characteristics can still receive different premiums, for example, due to interactions between bonus_malus and vehicle attributes, or because certain categorical segments, such as okres or Znacka_vozidla, shift the model output even when numeric risk factors are comparable. Such local explanations are useful for case-based validation, error analysis, and communication with non-technical stakeholders, because they translate a complex ensemble decision into an additive narrative while remaining faithful to the trained model (Lundberg and Lee, 2017).

3.9 Local interpretability (full text from the paper)

Figure **??** presents a SHAP waterfall plot for a single observation. It decomposes the model's prediction into a baseline value and a sequence of feature contributions that push the prediction up or down. The starting point on the x-axis is the expected model output, $E[f(x)] = 109.547$,

and the final predicted premium for this particular policy is $f(x) = 152.678$. Red bars indicate features that increase the prediction, while blue bars indicate features that decrease it, and the length of each bar corresponds to the magnitude of the contribution. In this case, the strongest positive driver is frekvencia_platenia, which increases the prediction by approximately +23.08. It is important to note that payment frequency was encoded as a grouped categorical feature, combining the original four categories into a single grouping. Because of this grouping, the SHAP contribution shown here reflects the cumulative effect of belonging to the specific grouped payment frequency category for this policy. In other words, the model does not attribute the effect to separate dummy levels in this visualization; instead, the displayed impact aggregates the contribution of the relevant category level, which is why frekvencia_platenia appears with a large single contribution.

Additional positive contributions come from Vykon (+9.7), Datum_storna (+9.11), vek_vozidla (+7.42), and znacka_vozidla (+6.2), indicating that the vehicle's power, cancellation-related timing, vehicle age, and brand segment all increase the predicted premium relative to the baseline. Smaller positive effects are also visible for Objem (+1.78), Hmotnost (+1.57), and prevedenie_emb (+0.89). On the negative side, vek decreases the prediction by about −12.98, suggesting that the customer's age places this observation in a segment with lower premiums than the average baseline. Further modest downward adjustments are attributed to bonus_malus (−1.67), Datum_pociatku (−1.02), zmluva (−0.64), and several minor categorical indicators, while the remaining features collectively contribute only a small additional negative shift. Overall, the plot provides a transparent case-level explanation: the predicted premium is substantially above the baseline primarily because the grouped payment frequency category contributes a strong upward adjustment, reinforced by power and temporal lifecycle-related features, while age offsets part of the increase with a notable negative contribution.

4 Conclusion

SHAP adds a substantial interpretability layer beyond standard XGBoost feature importance, explaining not only which variables matter but also how they influence predictions. Unlike gain-based importance, which provides a global ranking without indicating whether a feature increases or decreases the model's output, SHAP quantifies both the direction and magnitude of each feature's impact on the model's output. This enables clear local explanations at the level of individual policies, where each prediction can be decomposed into a baseline value and feature contributions that push the premium upward or downward. When aggregated across the dataset, SHAP also provides more reliable global insights because it reflects how features actually change predictions across the portfolio, rather than how often they are used in splits during training. In addition, SHAP reveals nonlinear relationships, threshold effects, and interactions between predictors, which are common in insurance pricing data and may remain hidden under classical importance metrics. Overall, integrating SHAP with XGBoost improves model transparency and supports more defensible interpretation and communication of model behavior in domains where explainability is critical.

Resources

Aas, K., M. Jullum, and A. Løland (2021). "Explaining individual predictions when features are dependent: More accurate approximations to Shapley values". In: *Artificial Intelligence* 298, p. 103502. DOI: 10.1016/j.artint.2021.103502.

Angelov, P. P., E. A. Soares, R. Jiang, N. I. Arnold, and P. M. Atkinson (2021). "Explainable artificial intelligence: An analytical review". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 11.5, e1424. DOI: 10.1002/widm.1424.

Barredo Arrieta, A., N. Diaz Rodriguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera (2020). "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". In: *Information Fusion* 58, pp. 82–115. DOI: 10.1016/j.inffus.2019.12.012.

Chen, Tianqi and Carlos Guestrin (2016). "XGBoost: A scalable tree boosting system". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, pp. 785–794. DOI: 10.1145/2939672.2939785.

Dhiman, P., A. Bonkra, A. Kaur, Y. Gulzar, Y. Hamid, M. S. Mir, and A. B. Soomro (2023). "Healthcare trust evolution with explainable artificial intelligence: Bibliometric analysis". In: *Information* 14.10, p. 541. DOI: 10.3390/info14100541.

Doshi-Velez, Finale and Been Kim (2017). *Towards a rigorous science of interpretable machine learning*. arXiv: 1702.08608. URL: https://arxiv.org/abs/1702.08608.

Guidotti, Riccardo, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi (2019). "A survey of methods for explaining black box models". In: *ACM Computing Surveys* 51.5, p. 93. DOI: 10.1145/3236009.

Kumar, I. E., S. Venkatasubramanian, C. Scheidegger, and S. Friedler (2020). "Problems with Shapley value based explanations as feature importance measures". In: *Proceedings of the 37th International Conference on Machine Learning*. PMLR, pp. 5491–5500.

Lundberg, Scott M., Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee (2020). "From local explanations to global understanding with explainable AI for trees". In: *Nature Machine Intelligence* 2.1, pp. 56–67. DOI: 10.1038/s42256-019-0138-9.

Lundberg, Scott M. and Su-In Lee (2017). "A unified approach to interpreting model predictions". In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, pp. 4768–4777.

Ponce-Bobadilla, A. V., V. Schmitt, C. S. Maier, S. Mensing, and S. Stodtmann (2024). "Practical guide to SHAP analysis: Explaining supervised machine learning model predictions in drug development". In: *Clinical and Translational Science* 17.11, e70056. DOI: 10.1111/cts.70056.

Rudin, Cynthia (2019). "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead". In: *Nature Machine Intelligence* 1.5, pp. 206–215. DOI: 10.1038/s42256-019-0048-x.

Sharma, C., S. Sharma, K. Sharma, and G. K. Sethi (2024). "Exploring explainable AI: A bibliometric analysis". In: *Discover Applied Sciences* 6, p. 615. DOI: 10.1007/s42452-024-06324-z.

Slack, Dylan, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju (2020). "Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods". In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. Association for Computing Machinery, pp. 180–186. DOI: 10.1145/3375627.3375830.

Weber, P., K. V. Carl, and O. Hinz (2024). "Applications of explainable artificial intelligence in finance: A systematic review of finance, information systems, and computer science literature". In: *Management Review Quarterly* 74, pp. 867–907. DOI: 10.1007/s11301-023-00320-0.