Proceedings of CONF-CDS 2026 Symposium: Computer Vision-Based Multimodal Cognitive Load Estimation for Adaptive Media Communication
DOI: 10.54254/2755-2721/2026.GL34621

Explainable Machine Learning for E-Commerce Purchase Intention: From Feature Importance to Interaction Effects and User Heterogeneity

Xinyu Wang

School of Computer Science, University of Jinan, Jinan, China
202331222154@stu.ujn.edu.cn

Abstract. Predicting online purchase intention is critical for e-commerce revenue optimization, yet existing approaches often sacrifice interpretability for marginal accuracy gains. This study proposes an interpretable prediction framework combining Random Forest with SHAP post-hoc analysis. Using the UCI Online Shoppers Intention dataset (N=12,330 sessions), Random Forest is benchmarked against Logistic Regression, XGBoost, and LightGBM, with statistical significance validated via McNemar test. Results show that Random Forest achieves the highest classification accuracy and best probability calibration, while gradient boosting models yield marginally higher AUC-ROC and F1-score. SHAP analysis reveals two counter-intuitive findings: (1) PageValues exhibits a saturation effect where marginal contribution to purchase probability plateaus beyond a threshold of approximately 50; (2) high PageValues buffers the negative effect of high ExitRates, suggesting that value-rich content retains users who would otherwise churn. To the best of knowledge in this domain, this is the first study to systematically quantify feature interactions, identify saturation thresholds, and assess user heterogeneity using SHAP on this widely adopted benchmark dataset.

Keywords: Explainable AI, Purchase Intention Prediction, Random Forest, SHAP, E-Commerce Analytics

1. Introduction

E-commerce has grown to over $6 trillion globally [1]. Platforms increasingly deploy machine learning for precision marketing and customer targeting. However, they face a persistent accuracy-interpretability trade-off: deep learning and gradient boosting models achieve high accuracy yet remain black boxes, while traditional statistical methods offer transparency but lack predictive power. This trade-off undermines managerial trust and complicates regulatory compliance for automated decision systems [2].

Existing studies predominantly emphasize prediction accuracy [3, 4]. They rely mainly on gradient boosting or neural networks. These approaches often neglect feature-level mechanisms that drive purchase decisions. While models like XGBoost and LightGBM routinely top rankings, their internal logic remains opaque to marketing practitioners who must justify budget allocation to

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.GL34621

stakeholders. Recent work has begun applying SHapley Additive exPlanations(SHAP) to explain model predictions in e-commerce contexts [5]. However, most studies stop at feature importance ranking without analyzing interaction effects or boundary conditions. Furthermore, many papers compare only two models without rigorous statistical testing [6]. Such limited comparisons cannot confirm whether observed differences are genuine rather than artifacts of a particular train-test split.

This study addresses three gaps and makes corresponding contributions. First, a rigorous four-model comparison is conducted (Logistic Regression, Random Forest, XGBoost, LightGBM) with McNemar statistical testing on an independent hold-out test set. This goes beyond prior work that typically compares only two models without rigorous significance testing [6]. Second, the analysis moves beyond feature importance rankings to explore whether PageValues exhibits non-linear or threshold-driven effects, and to examine interaction patterns between PageValues and ExitRates using SHAP dependence and interaction plots. Third, it is investigated whether model performance and feature explanations vary across user subgroups. These analysis layers have not been employed by prior SHAP-based studies in this domain.

2. Related work

Purchase intention prediction has evolved through three phases. Early approaches relied on logistic regression, association rules, and Recency, Frequency, Monetary (RFM) segmentation [7, 8]. These methods offer inherent interpretability but are limited by linear assumptions that fail to capture complex relationships in user behavior data. The second phase introduced machine learning classifiers, including Support Vector Machine (SVM) and ensemble methods [9]. These approaches progressively improved accuracy. Random Forest [9] emerged as a particularly effective non-linear method. More recently, gradient boosting frameworks have pushed accuracy boundaries further. Deep learning approaches have also contributed to this advancement [10]. Both come at the cost of interpretability and computational efficiency.

The third phase has focused on explainability. SHAP has emerged as the leading post-hoc method due to its theoretical foundations in cooperative game theory and unique satisfaction of efficiency, symmetry, dummy, and additivity axioms [11]. Unlike Local Interpretable Model-agnostic Explanations (LIME), which provides local approximations, SHAP offers both global feature importance summaries and local instance-level explanations. Prior studies have applied SHAP to e-commerce datasets. However, they predominantly report feature rankings [5]. Systematic quantification of feature interactions is lacking. The interaction between page value and exit behavior remains particularly underexplored in the purchase intention literature. Table 1 summarizes the positioning of this study relative to the most closely related prior work. While several studies have applied machine learning classifiers to the UCI Online Shoppers Intention dataset and a subset have employed basic SHAP feature importance rankings, none have combined rigorous multi-model comparison with McNemar testing, systematic SHAP interaction analysis, and subgroup heterogeneity assessment. This gap motivates this work.

DOI: 10.54254/2755-2721/2026.GL34621

Table 1. Positioning relative to related studies

Study	Dataset	Models	SHAP Depth	Statistical Test	User Heterogeneity
Chen et al. (2023) [5]	Taobao (custom)	6 models	Summary only	Chi-square	No
Adhikari et al. (2023)	UCI	RF/GB/XGB/LGBM	None	None	No
Baati & Mohsil (2020)	UCI	NB/C4.5/RF	None	None	No
Meka (NCI, 2023)	UCI	LSTM-RF, RF, XGB	Basic importance	None	No
This Study	UCI	LR/RF/XGB/LGBM	Full pipeline	McNemar	Yes

3. Methodology

3.1. Dataset and preprocessing

The UCI Online Shoppers Intention Dataset [12] is used in this study. The dataset contains N=12,330 sessions with 17 raw features, collected from an e-commerce platform between November and December 2017. The target variable Revenue indicates whether a session resulted in a purchase, with a 15.47% positive rate representing moderate class imbalance. Features span page metrics, engagement quality indicators, temporal indicators, and user characteristics.

Preprocessing follows best practices for mixed-type tabular data. The Month and VisitorType categorical variables were one-hot encoded into dummy variables. This deliberately avoids ordinal label encoding that would introduce spurious ordering relationships and bias linear models. Weekend and Revenue were converted to binary integers. Continuous features were standardized for Logistic Regression only. Tree-based models received raw values due to their inherent scale invariance. The final feature space comprises 28 dimensions after encoding.

3.2. Experimental design

A rigorous hold-out evaluation protocol is adopted. Data is partitioned into training (80%, N=9,864) and an independent hold-out test set (20%, N=2,466) using stratified sampling to preserve the 15.47% purchase rate. All hyperparameter tuning is performed via 5-fold cross-validation GridSearchCV on the training set. Final metrics are reported strictly on the never-seen test set to prevent optimistic bias.

Four models are compared to cover the spectrum from linear to ensemble methods. Logistic Regression (LR) with L2 (Ridge) regularization serves as a transparent linear baseline. Random Forest (RF) serves as the primary interpretable non-linear model. XGBoost represents gradient boosting. LightGBM represents histogram-based gradient boosting. All hyperparameters are optimized via 5-fold stratified cross-validation GridSearchCV.

3.3. Evaluation metrics

This study reports six complementary metrics [13, 14]. Area Under the Receiver Operating Characteristic Curve (AUC-ROC) measures ranking quality. Area Under the Precision-Recall Curve (AUC-PR) focuses on minority-class performance for imbalanced data. The harmonic mean of precision and recall (F1-score) provides balanced precision-recall. Matthews Correlation Coefficient (MCC) offers robust classification quality assessment. Brier Score evaluates probability calibration. Cohen's Kappa measures inter-rater agreement. Model differences are validated with the McNemar

DOI: 10.54254/2755-2721/2026.GL34621

test. This test directly assesses whether one model systematically corrects errors that another makes on the same paired test set. It provides stronger evidence of a practical difference than comparing metric confidence intervals.

3.4. SHAP interpretability framework

SHAP values are computed using TreeExplainer [11]. The computation uses the full independent test set (N=2,466). This ensures explanations reflect the model's behavior on unseen data. Global analysis identifies top features by mean absolute SHAP value. Dependence plots reveal the functional form of feature effects, showing whether relationships are linear, saturating, or non-monotonic. Critically, interaction plots quantify how the effect of one feature varies with another. This interaction analysis is the key methodological novelty of this study. It distinguishes our work from prior papers that report only feature rankings [15].

4. Experiments and results

4.1. Predictive performance

Table 2 presents the comprehensive model comparison on the independent test set. LightGBM achieves the highest AUC-ROC (0.9313) and AUC-PR (0.7438), followed closely by XGBoost. Random Forest attains the highest classification accuracy (89.62%) and best Brier Score (0.0734), indicating superior probability calibration. Logistic Regression trails on all metrics but maintains competitive AUC-ROC (0.8958) despite its linear constraint.

Table 2. Model performance on independent test set (N=2,466)

Model	Accuracy	AUC-ROC	AUC-PR	F1	MCC	Brier Score
Logistic Regression	0.8528	0.8958	0.6259	0.6109	0.5372	0.1245
Random Forest	0.8962	0.9219	0.7265	0.6394	0.5815	0.0734
XGBoost	0.8642	0.9305	0.7465	0.6550	0.5961	0.1007
LightGBM	0.8633	0.9313	0.7438	0.6515	0.5912	0.1003

The pattern reveals an instructive trade-off. Gradient boosting models excel at ranking and threshold-balanced metrics (AUC-ROC, AUC-PR, F1, MCC). Random Forest achieves the highest raw accuracy and best-calibrated probabilities. Practitioners should consider this trade-off when selecting models for specific deployment contexts.

Table 3 presents McNemar statistical significance tests. All pairwise differences are statistically significant ($p<0.001$). RF vs. LR validates that the non-linear ensemble significantly outperforms the linear baseline. RF vs. XGBoost and RF vs. LightGBM confirm that although gradient boosting achieves higher ranking metrics, their classification patterns differ significantly from those of Random Forest. Model selection is not merely choosing the highest AUC.

Table 3. McNemar statistical significance tests

Comparison	RF Correct / Other Wrong	RF Wrong / Other Correct	p-value
RF vs. LR	197	83	<0.001
RF vs. XGBoost	170	91	<0.001
RF vs. LightGBM	169	88	<0.001

DOI: 10.54254/2755-2721/2026.GL34621

4.2. SHAP global interpretability

The SHAP global analysis reveals that PageValues emerges as the dominant predictor (mean |SHAP| = 0.2392), with an influence nearly six-fold greater than the second-ranked feature, ExitRates (0.0404). Figure 1 visualizes this distribution, where the markedly wider horizontal spread of PageValues (colored by feature value) relative to all other features confirms its disproportionate impact. This confirms that the monetary value of pages visited is the overwhelmingly strongest predictor of purchase intention. This finding is consistent with prior literature [10]. This analysis provides greater precision than previous work. The remaining top features are ExitRates (negative effect), Month_Nov (positive seasonal effect associated with holiday shopping), ProductRelated_Duration (engagement depth), and BounceRates (negative engagement quality signal).

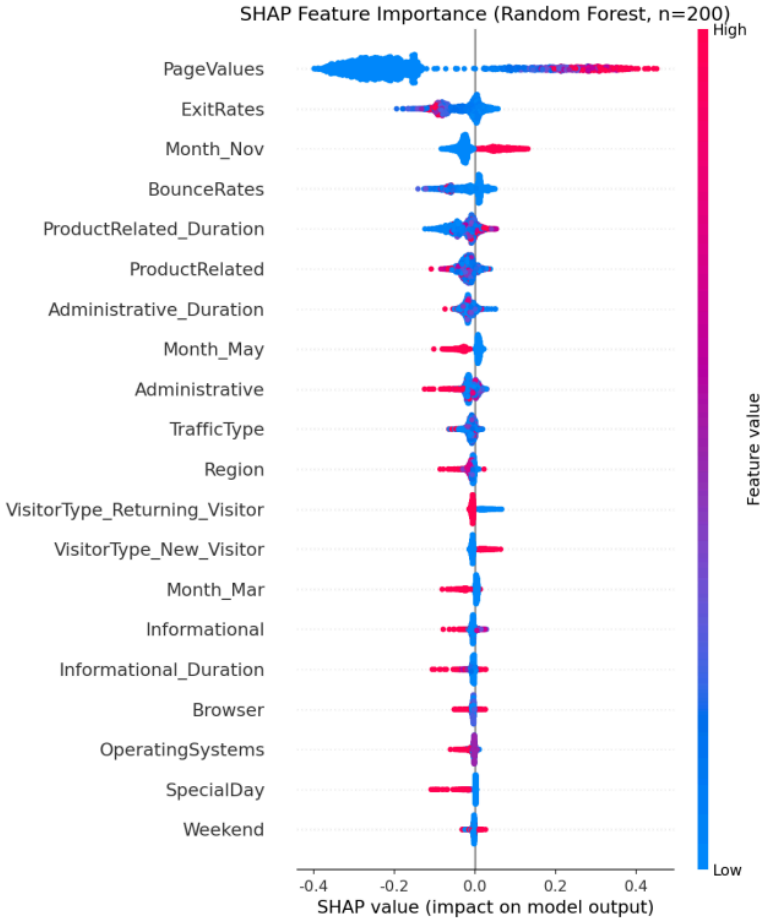

Figure 1. SHAP summary plot

4.3. Saturation effect in PageValues

Figure 2 visualizes this pattern: the SHAP dependence plot for PageValues reveals a saturation effect rather than a linear or inverted-U relationship. To the best of the authors' knowledge, this is the first documentation of such a threshold-driven saturation pattern in the purchase intention literature. When PageValues increases from 0 to approximately 50, the marginal SHAP value rises steeply, indicating a strong positive contribution to purchase probability. Beyond 50, the curve plateaus with diminishing marginal returns. At very high values (>150), the effect stabilizes rather than declining.

DOI: 10.54254/2755-2721/2026.GL34621

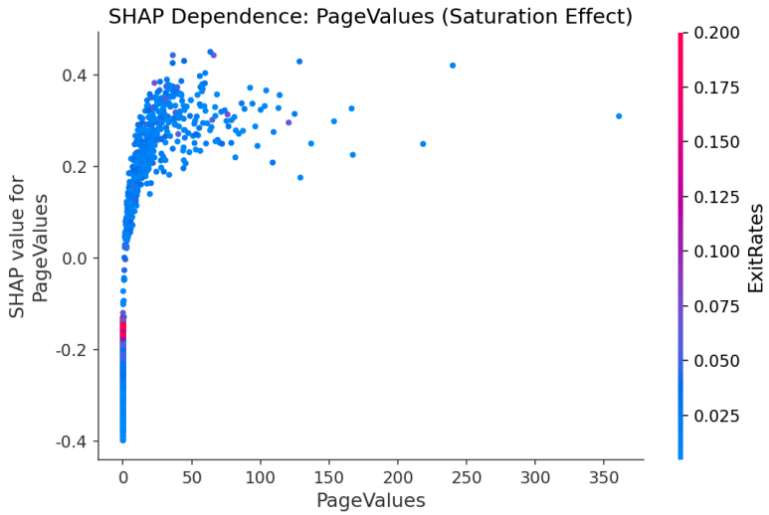

Figure 2. SHAP dependence plot for PageValues

This pattern suggests a threshold mechanism. A minimum page value is required to trigger purchase intent, but additional value beyond the threshold yields progressively smaller gains. This finding has practical implications for e-commerce content strategy. Platforms should focus on ensuring that product pages deliver at least moderate value (PageValues in the 20-50 range), where marginal returns are highest. They should avoid indiscriminately increasing page complexity or promotional content. The plateau challenges naive "more-is-better" assumptions embedded in traditional frameworks [8]. These frameworks typically assume monotonic relationships between monetary proxies and outcomes.

4.4. Buffering interaction: PageValues × ExitRates

Figure 3 presents the SHAP interaction heatmap, revealing a buffering pattern where high PageValues attenuate the negative impact of elevated ExitRates. This finding extends beyond prior SHAP-based studies which typically report only main effects. Among high-PageValues users, the negative impact of elevated ExitRates is attenuated compared to low-PageValues users, where exit behavior strongly predicts non-purchase. This suggests that high-value content compensates for higher exit tendencies. Value-rich pages retain users who would otherwise churn.

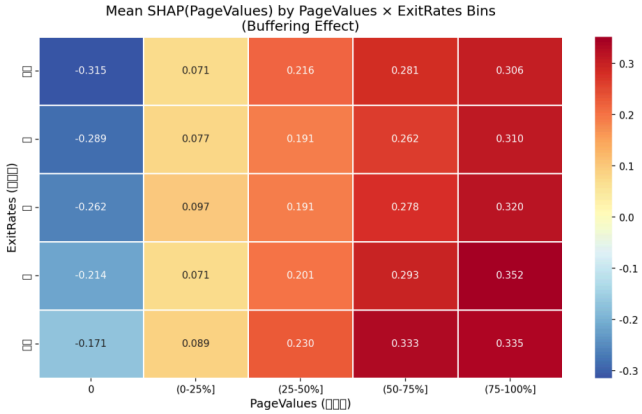

Figure 3. SHAP interaction plot for PageValues × ExitRates

DOI: 10.54254/2755-2721/2026.GL34621

This interaction carries actionable implications. Platforms should recognize that content quality and user engagement stability interact non-trivially. For high-value pages, moderate exit behavior does not necessarily imply lost conversions. For low-value pages, even low exit rates may not salvage conversion. This suggests a differentiated segmentation strategy. High-PageValue sessions warrant retention campaigns and personalized follow-up even if ExitRates are elevated. Low-PageValue sessions with rising ExitRates should be deprioritized.

4.5. Local explanation: A purchase case

Figure 4 illustrates this granularity through a waterfall plot for a correctly predicted purchaser (predicted probability=0.92, actual label=Purchase), showing SHAP's contribution breakdown at the individual session level. The baseline probability is 0.50 (adjusted for class balancing). PageValues=56.6 contributes +0.33, accounting for the majority of the upward shift from baseline to prediction. ExitRates=0.014 (very low) adds +0.03. BounceRates=0 adds +0.04. Month_Nov (holiday shopping season) contributes +0.03.

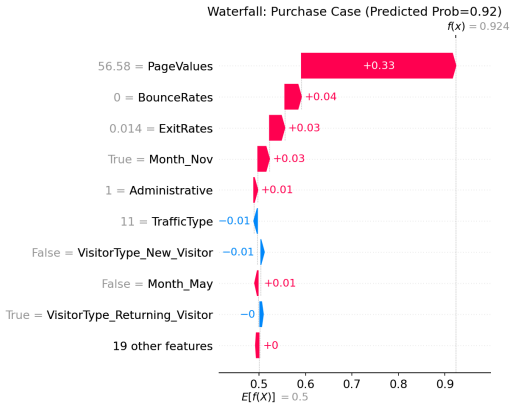

Figure 4. SHAP waterfall plot for a predicted purchaser (prob=0.92)

Interestingly, being a Returning Visitor contributes +0.00. This suggests that session-level page value and engagement metrics outweigh visitor type in driving purchase decisions. This challenges the conventional wisdom that returning visitors universally outperform new visitors. For marketing teams, this implies that onboarding and first-session experience optimization may yield higher marginal returns than exclusively targeting returning customer segments.

4.6. User heterogeneity analysis

To examine whether model effectiveness varies across user segments, AUC scores were computed for subgroups divided by median values of key features on the independent test set (Table 4). Results reveal substantial performance heterogeneity. AUC ranges from 0.754 (high-PageValues users) to 0.950 (non-November users), a gap of 0.196. Notably, the model performs substantially worse during the November shopping season (AUC=0.824) compared to other months (AUC=0.950). This suggests that promotional-season user behavior is inherently more unpredictable. Such unpredictability likely stems from impulse purchases and deal-seeking that deviates from normal browsing patterns. Counter-intuitively, high-PageValues users are harder to predict (AUC=0.754) than low-PageValues users (AUC=0.842). This may be because high-value browsers exhibit more heterogeneous purchase intentions. These findings confirm that a one-size-

DOI: 10.54254/2755-2721/2026.GL34621

fits-all deployment strategy is suboptimal. Platforms should apply differentiated prediction thresholds or segment-specific models, particularly for promotional seasons.

Table 4. Model performance across user subgroups

User Segment	Criterion	AUC	Group Size
High PageValues	> median	0.754	537
Low PageValues	<= median	0.842	1,929
Low ExitRates	<= median	0.912	1,237
High ExitRates	> median	0.917	1,229
November shoppers	Month_Nov=1	0.824	638
Non-November	Month_Nov=0	0.950	1,828

5. Conclusion

This study advances e-commerce purchase intention prediction by systematically quantifying feature interactions, saturation thresholds, and user heterogeneity. These dimensions have not been explored by prior SHAP-based studies in this domain. The core findings are threefold. First, four-model comparison with McNemar testing reveals trade-offs between ranking quality, calibration, and interpretability, rather than a single optimal choice. Second, SHAP dependence analysis identifies a saturation effect in PageValues with a threshold around 50. This challenges monotonicity assumptions in traditional RFM frameworks. Third, SHAP interaction analysis reveals a buffering pattern between PageValues and ExitRates. Subgroup AUC varies by up to 0.196. This variation is most notable between promotional-season and non-promotional users.

For platform operators, it is recommended to maintain PageValues in the 20-50 range for maximum marginal return. Promotional-period users should be treated as a distinct cohort requiring specialized models. SHAP waterfall plots can enable transparent real-time session scoring.

This study relies on a single public dataset without causal inference. Future work can integrate the DoWhy causal framework to validate the causality of SHAP relationships. Cross-cultural validation through multi-country datasets is also planned.

References

[1] Lemon, K. N., Verhoef, P. C. Understanding Customer Experience Throughout the Customer Journey [J]. Journal of Marketing, 80(6), 2016: 69-96.
[2] European Parliament. Regulation (EU) 2016/679 of the European Parliament and of the Council (General Data Protection Regulation) [EB/OL]. 2016. https: //eur-lex.europa.eu/eli/reg/2016/679/oj
[3] Chen, T., Guestrin, C. XGBoost: A Scalable Tree Boosting System [C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016: 785-794.
[4] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree [C]//Advances in Neural Information Processing Systems, Vol. 30. Curran Associates, 2017: 3149-3157.
[5] Chen, Y., Liu, H., Wen, Z., Lin, W. How Explainable Machine Learning Enhances Intelligence in Explaining Consumer Purchase Behavior: A Random Forest Model with Anchoring Effects [J]. Systems, 11(6), 2023: 312.
[6] Dietterich, T. G. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms [J]. Neural Computation, 10(7), 1998: 1895-1923.
[7] Hosmer, D. W., Lemeshow, S. Applied Logistic Regression [M]. 2nd ed. John Wiley & Sons, 2000.
[8] Fader, P. S., Hardie, B. G., Lee, K. L. RFM and CLV: Using Iso-Value Curves for Customer Base Analysis [J]. Journal of Marketing Research, 42(4), 2005: 415-430.
[9] Breiman, L. Random Forests [J]. Machine Learning, 45(1), 2001: 5-32.

DOI: 10.54254/2755-2721/2026.GL34621

[10] Sakar, C. O., Polat, S. O., Katircioglu, M., Kastro, Y. Real-Time Prediction of Online Shoppers' Purchasing Intention Using Multilayer Perceptron and LSTM Recurrent Neural Networks [J]. Neural Computing and Applications, 31, 2019: 6893-6908.
[11] Lundberg, S. M., Lee, S.-I. A Unified Approach to Interpreting Model Predictions [C]//Advances in Neural Information Processing Systems, Vol. 30. Curran Associates, 2017: 4768-4777.
[12] UCI Machine Learning Repository. Online Shoppers Purchasing Intention Dataset [DB/OL]. 2018. https: //doi.org/10.24432/C5F88Q
[13] Saito, T., Rehmsmeier, M. The Precision-Recall Plot Is More Informative Than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets [J]. PLOS ONE, 10(3), 2015: e0118432.
[14] Chicco, D., Jurman, G. The Advantages of the Matthews Correlation Coefficient (MCC) over F1 Score and Accuracy in Binary Classification Evaluation [J]. BMC Genomics, 21, 2020: 6.
[15] Molnar, C. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable [EB/OL]. 2nd ed. 2022. https: //christophm.github.io/interpretable-ml-book/