Interpretable Machine Learning for Taxi Tipping Behavior Analysis in Urban Mobility Systems

Hui He, Kefeng Wang*
School of Energy Science and Engineering
Henan Polytechnic University
Jiaozuo 454003, China
*Corresponding author email:kfhere@126.com

***Abstract*—Taxi tipping behavior provides a useful case for studying passenger payment decisions in urban mobility services. Using New York City Yellow Taxi trip records from January to March 2025, this study examines two related tasks: predicting tipping occurrence and estimating tip amount for trips with positive tips. The input feature system includes trip and payment variables, temporal variables, spatial and operational variables, and fare-structure variables. Logistic regression and ordinary least squares regression are used as statistical baselines, while GBDT, XGBoost, LightGBM, CatBoost, and FT-Transformer are compared as tabular prediction models. Models trained on January data are evaluated on a January holdout set and on February and March external test sets. The results show that FT-Transformer achieves the highest AUC for tipping occurrence prediction, while CatBoost achieves the strongest January holdout performance for conditional tip amount regression and remains close to the best external-test results. Significance tests support the classification gains of CatBoost over the logistic regression baseline and the GBDT benchmark, whereas regression gains are more limited. Sensitivity analysis shows that credit-card payment is a strong predictive marker for tipping occurrence but has little additional value for tip amount once tipping has occurred. SHAP analysis further indicates that payment context is more important for tipping occurrence, while trip scale and fare-related variables are more relevant to payment intensity. These findings support a dual-layer interpretation of taxi tipping behavior.**

Keywords-taxi tipping behavior; tabular learning; interpretable machine learning; CatBoost; FT-Transformer; temporal validation

I. INTRODUCTION

Urban taxi services generate trip-level records that describe how passengers travel, how fares are charged, and how payments are made. These records provide a useful basis for studying passenger payment behavior under real operating conditions. Tipping is of particular interest because it is voluntary and varies across trips. A tip may be associated with service experience, payment habit, trip context, and local norms. Previous studies have shown that tipping behavior differs across service settings and cannot be explained only by service quality or by a single trip attribute [1], [2].

Payment context is closely related to tipping. Payment transparency may affect how consumers perceive spending and additional payment [3]. In taxi services, cash, credit-card, and electronic payment modes coexist. Credit-card payment may therefore be strongly associated with whether a passenger gives a tip. This association should be interpreted as predictive rather than causal, because payment mode is not randomly assigned and may be related to location, travel time, fare level, passenger choice, and service context.

(1) The Fundamental Research Funds for the Universities of Henan Province, Project No. NSFRF230425. (2) Soft Science Research Project of Henan Provincial Department of Science and Technology, Project No. 242400410292.

This study analyzes taxi tipping behavior through two prediction tasks. The first task is tipping occurrence, which refers to whether a passenger gives a positive tip. The second task is tip amount, which refers to the amount paid among trips with positive tips. In this paper, payment intensity denotes tip amount conditional on tipping. The term dual-layer interpretation refers to the distinction between tipping occurrence and payment intensity. This distinction is used because the variables associated with whether tipping occurs may differ from those associated with how much is paid after tipping occurs.

Transportation behavior studies often use statistical models because their coefficients are easy to interpret [4], [5]. At the same time, taxi trip records contain nonlinear relationships, categorical fields, spatial identifiers, temporal patterns, and fare-structure information. These data characteristics make tabular machine learning models useful for prediction and comparison. Existing studies have shown that machine learning methods can improve travel behavior prediction in some settings [4]–[6]. Recent tabular-learning studies also suggest that boosted-tree models remain strong baselines, even when deep tabular models are considered [7], [12].

Based on New York City Yellow Taxi trip records from January to March 2025, this study constructs an input feature system that includes trip and payment variables, temporal variables, spatial and operational variables, and fare-structure variables. The empirical analysis compares logistic regression, ordinary least squares regression, GBDT, XGBoost, LightGBM, CatBoost, and FT-Transformer under the same preprocessing and temporal validation setting [8]–[11]. The analysis then uses significance testing, credit-card sensitivity analysis, feature-group analysis, and SHAP interpretation to examine how payment context, trip scale, operating conditions, and fare structure are associated with the two tipping tasks [13], [14].

II. DATA AND FEATURE DESIGN

A. Data Source and Sample Construction

This study uses public New York City Yellow Taxi trip records released by the New York City Taxi and Limousine Commission. Each record corresponds to one completed taxi trip and contains pickup time, trip distance, fare components, payment type, pickup and drop-off location identifiers, and tip amount.

The analysis uses records from January to March 2025. January 2025 is used as the modeling month. A sample of 200,000 trips is randomly drawn from January using a fixed

random seed and divided into training and holdout subsets. The training subset is used for model fitting and model selection, while the January holdout subset is used for internal evaluation. February and March records are processed with the same rules and used as external test sets. An independent 100,000-trip January sample is also used for a sample-size robustness check.

B. Preprocessing and Target Variables

The same preprocessing rules are applied to all monthly samples. Trips are retained when fare_amount > 0, trip_distance > 0, and tip_amount >= 0. These filters remove trips with non-positive fare values, non-positive travel distance, and negative tip records.

Fare amount, trip distance, and tip amount have right-skewed distributions. These continuous variables are winsorized at the 0.995 quantile to reduce the influence of extreme observations. Missing values such as passenger count are imputed using the median. The same cleaning, winsorization, and imputation rules are applied to January, February, and March data.

Two target variables are defined. Tipping occurrence equals 1 when tip_amount > 0 and 0 otherwise. It is used in the classification task. Tip amount is modeled only for trips with positive tips and is used in the conditional regression task. Payment intensity refers to the level of tip amount after tipping has occurred.

C. Input Feature System and Validation Design

The input feature system is organized into four groups: trip and payment features, temporal features, spatial and operational features, and fare-structure features. Credit-card payment is treated as a predictive feature rather than a causal treatment, because payment mode may be associated with location, time, fare level, passenger preference, and service context. Variables directly derived from the target, such as tip amount and total amount including tips, are not used as predictors.

TABLE I. INPUT FEATURE GROUPS USED IN THE ANALYSIS

Feature group	Variables	Role in the analysis
Trip and payment features	Fare amount; trip distance; extra charge; passenger count; credit-card payment	Basic trip scale and payment mode
Temporal features	Pickup hour; pickup day of week; weekend indicator; night-trip indicator	Time-of-travel context
Spatial and operational features	Pickup zone; drop-off zone; vendor; rate code; store-and-forward flag	Location and operating conditions
Fare-structure features	MTA tax; tolls amount; improvement surcharge; congestion surcharge; airport fee; CBD congestion fee	Pricing and fee context

The validation design follows the time order of the data. Models are trained on the January training subset and evaluated on the January holdout subset. The trained models are then tested on February and March external samples. February and March data are not used for model selection. For the classification task, model performance is evaluated using AUC, accuracy, and F1-score. For the conditional regression task, performance is evaluated using R^2, RMSE, and MAE.

III. MODELS AND EVALUATION STRATEGY

A. Statistical Baselines

Two statistical models are used as interpretable baselines. Logistic regression is used for tipping occurrence prediction, and ordinary least squares regression is used for conditional tip amount prediction [4]–[6].

For the classification task, let x_i denote the input feature vector of trip i, and let $y_i \in \{0,1\}$ denote the tipping occurrence indicator. The logistic regression model estimates the probability that trip i receives a positive tip:

$$P(y_i = 1 | \ x_i) = \frac{1}{1+\exp\left[-\left(\beta_0 + \beta^T x_i\right)\right]} \tag{1}$$

In (1), β_0 is the intercept, and β is the coefficient vector for the input features.

For the conditional regression task, the sample is restricted to trips with positive tips. Let t_i denote the tip amount of tipped trip i. The OLS model is specified as

$$t_i = \alpha_0 + \alpha^T x_i + \varepsilon_i, \quad t_i > 0 \tag{2}$$

In (2), α_0 is the intercept, α is the coefficient vector, and ε_i is the error term. HC3 robust standard errors are used for coefficient inference because fare-related variables may show heteroskedasticity [15].

B. Tabular Prediction Models

This study evaluates GBDT, XGBoost, LightGBM, CatBoost, and FT-Transformer. GBDT, XGBoost, LightGBM, and CatBoost are boosted-tree models. XGBoost improves tree boosting through regularized optimization, LightGBM improves computational efficiency for large-scale tabular data, and CatBoost is designed to handle categorical features effectively [8]–[10]. FT-Transformer is included as a deep tabular model that uses attention-based feature interaction [11].

Boosted-tree models construct an additive predictor through sequentially fitted trees. Their general form is

$$F_m(x) = F_{m-1}(x) + \eta h_m(x) \tag{3}$$

where $F_m(x)$ is the ensemble prediction after m boosting iterations, $F_{m-1}(x)$ is the previous ensemble prediction, $h_m(x)$ is the newly fitted tree at iteration m, and η is the learning rate. Although these models share this general structure, they differ in optimization, sampling, regularization, and categorical-feature handling. Recent studies show that boosted-tree methods remain strong baselines for many tabular prediction tasks, even when deep tabular models are considered [7], [12].

The main hyperparameter settings are reported in Table II. These settings are fixed before external testing. Tree-based models use pre-specified parameters across all monthly evaluations. FT-Transformer uses an internal validation split from the January training data for early stopping. February and March samples are used only for external testing.

TABLE II. MAIN HYPERPARAMETER SETTINGS OF TABULAR MODELS

Model	Classification	Regression
GBDT	400 trees; max depth = 3; learning rate = 0.05	600 trees; max depth = 3; learning rate = 0.05
XGBoost	600 trees; max depth = 5; learning rate = 0.05; subsample = 0.8; colsample = 0.9; L2 = 1.0	800 trees; max depth = 5; learning rate = 0.05; subsample = 0.85; colsample = 0.9; L2 = 1.0
LightGBM	600 trees; max depth = 5; learning rate = 0.05; subsample = 0.8; colsample = 0.9; L2 = 1.0	800 trees; max depth = 5; learning rate = 0.05; subsample = 0.85; colsample = 0.9; L2 = 1.0
CatBoost	600 iterations; depth = 6; learning rate = 0.05	800 iterations; depth = 6; learning rate = 0.05
FT-Transformer	3 blocks; batch size = 1024; max epochs = 20; early stopping patience = 4	3 blocks; batch size = 1024; max epochs = 20; early stopping patience = 4

The random seed is fixed at 2026 for sampling and model training where applicable. All models use the same target variables, samples, preprocessing rules, and input feature system.

C. Evaluation, Significance Testing, and Interpretation

AUC is used as the main classification metric because it measures discrimination ability across decision thresholds. Accuracy and F1-score are reported as complementary metrics. For conditional tip amount prediction, R^2, RMSE, and MAE are reported.

Significance tests are used to examine whether the main performance differences are statistically supported. For the classification task, AUC differences are evaluated using the DeLong test on the January holdout predictions. For the regression task, paired bootstrap tests with 2000 resamples are used to compare MAE and RMSE between competing models, and empirical two-sided *p*-values are reported. The main text reports MAE-based bootstrap results. These tests are used to support a cautious interpretation of model differences rather than to overstate small numerical gains.

SHAP is used to interpret the CatBoost models because CatBoost performs strongly in both tasks and supports tree-based SHAP analysis. SHAP decomposes a model prediction into additive feature contributions:

$$f(x) = \phi_0 + \sum_j \phi_j \tag{4}$$

In (4), $f(x)$ is the model prediction for input x, ϕ_0 is the baseline prediction, ϕ_j is the SHAP contribution of feature j, and p is the number of input features [13], [14].

A credit-card sensitivity analysis compares model performance with the credit-card payment variable included, with the variable removed, and within the credit-card-only subsample. This design evaluates predictive dependence on payment mode but does not identify a causal effect.

IV. RESULTS AND DISCUSSION

A. Model Performance under the Full Feature System

Table III reports the main performance results. For tipping occurrence prediction, FT-Transformer achieves the highest AUC in all three test sets, with AUC values of 0.960655 on the January holdout set, 0.964027 on the February external set, and 0.960056 on the March external set. CatBoost and LightGBM show similar performance, while XGBoost and GBDT also outperform the logistic regression baseline.

For conditional tip amount prediction, CatBoost achieves the highest R^2 on the January holdout set, with $R^2 = 0.769414$, RMSE = 1.657480, and MAE = 0.964149. On the February and March external sets, its performance remains close to the strongest results. GBDT and OLS remain close to CatBoost, while FT-Transformer performs weaker in this task. This difference suggests that tipping occurrence contains stronger nonlinear and contextual signals, whereas positive tip amount is more directly associated with trip scale and fare components.

TABLE III. MODEL PERFORMANCE UNDER THE FULL FEATURE SYSTEM

Task	Model	Jan	Feb	Mar
Classification AUC	Logit	0.954894	0.958482	0.954768
Classification AUC	GBDT	0.958731	0.962078	0.957301
Classification AUC	XGBoost	0.959594	0.963682	0.958636
Classification AUC	LightGBM	0.960155	0.963662	0.959124
Classification AUC	CatBoost	0.960028	0.963924	0.958833
Classification AUC	FT-Transformer	0.960655	0.964027	0.960056
Regression R^2	OLS	0.768845	0.759768	0.766693
Regression R^2	GBDT	0.769249	0.759249	0.769079
Regression R^2	LightGBM	0.767433	0.756118	0.767224
Regression R^2	CatBoost	0.769414	0.759564	0.768869
Regression R^2	FT-Transformer	0.759694	0.751990	0.759424

Figure 1 presents the ROC curves of FT-Transformer for tipping occurrence prediction. The curves for the January holdout set and the February and March external sets remain close, indicating stable discrimination across monthly samples. The inset highlights the upper-left region of the ROC space, where the curves overlap most strongly.

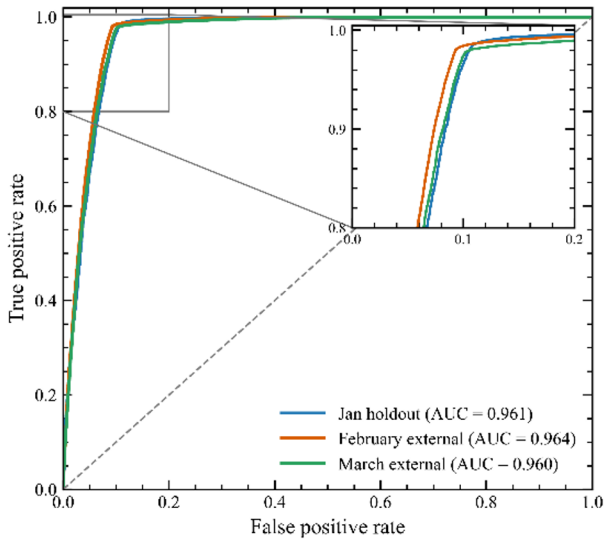

Figure 1. ROC curves of FT-Transformer for tipping occurrence prediction.

B. *Significance and Sensitivity Checks*

Table IV reports the main significance and sensitivity results. For tipping occurrence prediction, CatBoost improves AUC over Logit by 0.005134, with $p \approx 5.49 \times 10^{-11}$. Its AUC gain over GBDT is smaller, at 0.001297, but remains statistically supported with $p \approx 9.89 \times 10^{-4}$. For conditional tip amount prediction, CatBoost improves MAE over OLS by 0.004988, with $p = 0.004$. Its MAE improvement over GBDT is 0.001413, with $p = 0.182$, so the gain over the nonlinear baseline is not clearly significant.

The feature-group check shows that contextual variables provide additional information. Classification AUC increases from 0.951133 under the core trip and payment variables to 0.960028 under the full feature system. The largest increase appears after including spatial and operational variables, where AUC reaches 0.959905. For conditional tip amount prediction, R^2 increases from 0.763233 under the core variables to 0.769414 under the full feature system.

The credit-card sensitivity analysis shows stronger task differences. With credit-card payment included, classification AUC is 0.960028. Removing this variable reduces AUC to 0.784175, and the credit-card-only subsample gives AUC = 0.738129. In contrast, the regression task changes little: $R^2 = 0.769414$ with the variable, $R^2 = 0.769250$ without it, and $R^2 = 0.767580$ in the credit-card-only subsample. Credit-card payment is therefore a strong predictive marker for tipping occurrence, but it has limited additional value for tip amount once tipping has occurred.

TABLE IV. SIGNIFICANCE, FEATURE-GROUP, AND CREDIT-CARD SENSITIVITY RESULTS

Analysis	Comparison or setting	Result
Classification significance	Logit vs. CatBoost	**Δ**AUC = 0.005134, $\boldsymbol{p \approx 5.49 \times 10^{-11}}$
Classification significance	GBDT vs. CatBoost	**Δ**AUC = 0.001297, $\boldsymbol{p \approx 9.89 \times 10^{-4}}$
Regression significance	OLS vs. CatBoost	MAE improvement = 0.004988, $\boldsymbol{p = 0.004}$
Regression significance	GBDT vs. CatBoost	MAE improvement = 0.001413, $\boldsymbol{p = 0.182}$
Feature-group check	Core variables to full feature system	Classification AUC: 0.951133 to 0.960028; Regression $\boldsymbol{R^2}$: 0.763233 to 0.769414
Credit-card sensitivity	With vs. without credit-card payment	Classification AUC: 0.960028 to 0.784175; Regression $\boldsymbol{R^2}$: 0.769414 to 0.769250

These findings should not be interpreted as causal effects of payment mode. Credit-card payment is selected by passengers and may be related to time, location, fare level, trip purpose, and other unobserved factors. In addition, recorded tip amounts in taxi trip data may capture card-based tips more completely than cash tips, which further supports interpreting credit-card payment as a predictive marker rather than a causal driver. The sensitivity results support a predictive interpretation: payment mode is highly informative for whether tipping occurs, but it explains little variation in tip amount among tipped trips.

C. *SHAP-Based Interpretation*

SHAP analysis is conducted with CatBoost to compare feature roles across the two tasks. For tipping occurrence prediction, credit-card payment has the largest contribution. Fare amount, extra charge, vendor information, drop-off zone, congestion surcharge, trip distance, passenger count, pickup zone, and rate code also contribute to the model. This pattern indicates that tipping occurrence is strongly associated with payment mode, while spatial, operational, and fare-structure variables provide additional information.

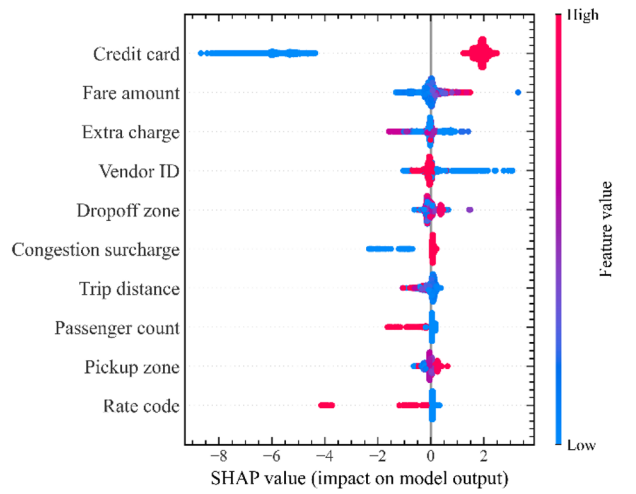

Figure 2. SHAP summary plot of CatBoost for tipping occurrence prediction.

For conditional tip amount prediction, the interpretation focuses on payment intensity among trips with positive tips. Fare amount, trip distance, and fee-related variables play a larger role than credit-card payment after tipping has occurred. This pattern is consistent with the regression results and the credit-card sensitivity analysis. The SHAP results support the dual-layer interpretation of taxi tipping behavior: tipping occurrence is mainly associated with payment context and operating conditions, while tip amount is more closely related to trip scale and fare structure.

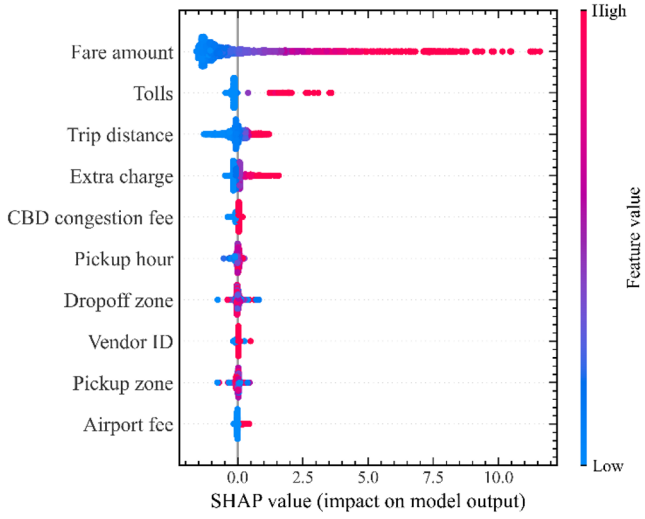

Figure 3. SHAP summary plot of CatBoost for conditional tip amount prediction.

V. CONCLUSION

This study analyzed taxi tipping behavior using New York City Yellow Taxi trip records from January to March 2025. The analysis separated tipping behavior into two tasks: tipping occurrence prediction for all valid trips and conditional tip amount prediction for trips with positive tips. A unified input feature system was constructed from trip and payment variables, temporal variables, spatial and operational variables, and fare-structure variables. Statistical baselines, boosted-tree models, and FT-Transformer were evaluated under the same preprocessing rules and temporal validation design.

The results show different modeling patterns across the two tasks. For tipping occurrence prediction, FT-Transformer achieves the highest AUC on the January holdout set and the February and March external test sets. For conditional tip amount prediction, CatBoost achieves the strongest January holdout result and remains close to the best external-test results, but its advantage over GBDT is small and not clearly significant. This contrast indicates that tipping occurrence contains stronger nonlinear and contextual signals, while positive tip amount is more closely related to trip scale and fare-structure variables.

The sensitivity and SHAP analyses support a cautious interpretation. Credit-card payment is a strong predictive marker for tipping occurrence, but it has limited additional value for predicting tip amount once tipping has occurred. Because payment mode is selected by passengers and may be associated with location, time, fare level, and unobserved trip context, this result should be interpreted as predictive association rather than causal evidence. Future work can extend the analysis to longer periods, multiple cities, and richer contextual data, and causal designs may be needed to separate payment-mode effects from selection into payment methods.

ACKNOWLEDGMENT

The authors would like to thank the New York City Taxi and Limousine Commission (NYC TLC) for providing the publicly available trip record data used in this study.

REFERENCES

[1] D. Elliott, M. Tomasini, M. Oliveira, and R. Menezes, “Tippers and stiffers: An analysis of tipping behavior in taxi trips,” in Proc. IEEE Int. Conf. Data Mining Workshops (ICDMW), New Orleans, LA, USA, 2017, pp. 934–941.

[2] M. Lynn and M. McCall, “Gratitude and gratuity: A meta-analysis of the service-tipping relationship,” Journal of Socio-Economics, vol. 29, no. 2, pp. 203–214, 2000.

[3] D. Soman, “The effect of payment transparency on consumption: Quasi-experiments from the field,” Marketing Letters, vol. 14, no. 3, pp. 173–183, 2003.

[4] X. Zhao, X. Yan, A. Yu, and P. Van Hentenryck, “Prediction and behavioral analysis of travel mode choice: A comparison of machine learning and logit models,” Travel Behaviour and Society, vol. 20, pp. 22–35, 2020.

[5] M. T. Kashifi, A. Jamal, M. S. Kashefi, M. Almoshaogeh, and S. M. Rahman, “Predicting the travel mode choice with interpretable machine learning techniques: A comparative study,” Travel Behaviour and Society, vol. 29, pp. 279–296, 2022.

[6] S. Mullainathan and J. Spiess, “Machine learning: An applied econometric approach,” Journal of Economic Perspectives, vol. 31, no. 2, pp. 87–106, 2017.

[7] R. Shwartz-Ziv and A. Armon, “Tabular data: Deep learning is not all you need,” Information Fusion, vol. 81, pp. 84–90, 2022.

[8] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, San Francisco, CA, USA, 2016, pp. 785–794.

[9] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, “LightGBM: A highly efficient gradient boosting decision tree,” in

Advances in Neural Information Processing Systems, vol. 30, 2017, pp. 3146–3154.

[10] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: Unbiased boosting with categorical features," in Advances in Neural Information Processing Systems, vol. 31, 2018, pp. 6639–6649.

[11] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, "Revisiting deep learning models for tabular data," in Advances in Neural Information Processing Systems, vol. 34, 2021, pp. 18932–18943.

[12] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on typical tabular data?" in Advances in Neural Information Processing Systems, vol. 35, pp. 507–520, 2022.

[13] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems, vol. 30, 2017.

[14] X. Kong, Y. Zhang, W. L. Eisele, and X. Xiao, "Using an interpretable machine learning framework to understand the relationship of mobility and reliability indices on truck drivers' route choices," IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 8, pp. 13419–13428, 2022.

[15] M. A. Mansournia, M. Nazemipour, A. I. Naimi, and D. G. Altman, "Reflection on modern methods: Demystifying robust standard errors for epidemiologists," International Journal of Epidemiology, vol. 50, no. 1, pp. 346–351, 2021.