Transport and Telecommunication, 2025, volume 26, no. 4, 352-361
Transport and Telecommunication Institute, Lauvas 2, Riga, LV-1019, Latvia
DOI 10.2478/ttj-2025-0027

ENHANCING MODE-CHOICE MODELS WITH CONFORMAL PREDICTION: UNCERTAINTY QUANTIFICATION AND DECISION SUPPORT USING TREE-BASED MACHINE LEARNING

Ramin Bohlouli[1], Ken Koshy Varghese[2], Guido Gentile[3], Mohamed Eldafrawi[4]

[1-4]Università di Roma La Sapienza,
Roma, Italy, Via Eudossiana 18 - 00184
[1-4]{ramin.bohlouli, kenkoshy.varghese, guido.gentile, mohamed.eldafrawi}@uniroma1.it

Accurate mode-choice forecasts are vital for effective transportation planning. Transit agencies and city planners rely on precise predictions, but unreliable forecasts can misdirect even the most behaviorally grounded insights. For decades, discrete choice models (DCMs), notably Multinomial Logit (MNL) and Mixed Multinomial Logit (MMNL), have explained why travelers choose particular modes via interpretable parameters, yet they often underperform in forecast accuracy. More recently, machine learning methods (e.g., tree-based algorithms) have come to capture complex, nonlinear patterns, often outperforming DCMs in point-prediction accuracy. However, they lack built-in confidence measures, limiting their use in risk-aware decision making. In this work, we help narrow this gap by wrapping our best ML model in an Inductive Mondrian Conformal Prediction (IMCP) layer with per-mode calibration at 90% nominal coverage. We leverage a survey of approximately 8,000 Italian employees, capturing their socio-economic attributes and travel habits. Using a tailored preprocessing pipeline, we compare XGBoost, Random Forest, and CatBoost, observing that XGBoost performs best on the test set with an overall accuracy of 89.7% and a macro-average F1 score of 83.6%. Our IMCP layer then produces distribution-free prediction sets that contain the true mode at least 90% of the time, both overall and within each individual mode category. Singleton prediction sets can be treated as high-confidence forecast for capacity planning, while multilabel sets (and the occasional empty sets for highly ambiguous cases) highlight where uncertainty is greatest and pinpoint exactly which individuals merit follow-up surveys or targeted incentives.

Keywords: Transportation mode choice, uncertainty quantification, conformal prediction, machine learning in transportation, risk-aware decision-making

1. Introduction

Accurate mode-choice forecasts are fundamental to transportation planning. Transit agencies, city planners, and employers rely on them to make critical decisions, such as estimating demand for new transit routes or parking infrastructure. When forecasts are unreliable, even well-founded behavioral models can lead to misallocated resources and ineffective policies.

Traditionally, this task has been addressed using Discrete Choice Models (DCMs), which rely on Random Utility Theory to represent individual decisions. These models assume that each person selects the mode of transport offering the highest utility, a value typically modeled as a function of observable attributes and an unobservable stochastic component. Notably, models such as the Multinomial Logit (MNL) and the Mixed Multinomial Logit (MMNL) (McFadden, 1973a, 2000b) have been foundational in the field. However, while MNL models benefit from simplicity and tractability, they suffer from restrictive assumptions such as the independence of irrelevant alternatives (IIA) and limited capacity to capture preference heterogeneity. MMNL models address some of these issues by introducing random coefficients, offering a more flexible representation of individual-specific tastes (Ben-Akiva *et al.*, 1985).

The growing use of machine learning (ML) in transportation modeling has brought significant improvements in predictive accuracy (Bishop, 2006; Hagenauer *et al*., 2017; Zhao *et al.*, 2020; Cheng *et al.*, 2019). Tree-based algorithms such as Random Forest, XGBoost, and CatBoost can learn complex patterns from large datasets. However, their adoption in policy-sensitive applications remains limited due to the lack of built-in uncertainty quantification. Unlike DCMs, most ML models do not indicate how confident they are in their predictions (an essential requirement in risk-aware decision-making).

Uncertainty quantification is crucial in fields where predictive errors can have significant consequences, such as in autonomous driving and traffic forecast. Recent research in these domains has begun to incorporate probabilistic frameworks to express model confidence (Ravindran *et al.*, 2022; Chen

et al., 2022; Hasan *et al.,* 2019; Eldafrawi *et al.* 2024). In transportation mode choice, however, efforts to incorporate uncertainty quantification into ML models remain sparse. One notable exception is the use of Bayesian Neural Networks (Zheng *et al.,* 2024), which provide uncertainty estimates but are model-specific and computationally intensive.

To address this gap, we propose the use of Conformal Prediction (CP) (Shafer *et al.,* 2008), a model-agnostic framework that provides statistically valid prediction sets with formal coverage guarantees. Specifically, we employ Inductive Mondrian Conformal Prediction (IMCP) (Vovk *et al.,* 2005), which calibrates class-conditional prediction thresholds on a separate validation set. This ensures robust uncertainty quantification, especially for imbalanced datasets. We further enhance this by computing Wilson intervals (Wilson, 1927) to correct for finite-sample variability in the model. To the best of our knowledge, this is the first use of IMCP in transportation mode choice modeling. Our methodology combines high-performing ML models with this uncertainty-aware wrapper. We train and tune several classifiers using a stratified cross-validated process, and select XGBoost as the best-performing model. We then apply IMCP with a target coverage of 90%, evaluating both overall and per-class performance. Results show that our framework produces reliable prediction sets, with the majority being singleton predictions and coverage rates closely aligning with theoretical guarantees. Importantly, ambiguous cases, where the prediction set includes multiple plausible modes, offer actionable insights for follow-up surveys or targeted incentives. This study contributes both a novel methodological approach for scholars, and a practical decision-support tool for planners aiming to integrate confidence-aware predictions into transportation policy.

The remainder of this paper is structured as follows: Section 2 reviews the relevant literature. Section 3 describes the dataset and methodological approach. Section 4 presents the simulation study and results. Section 5 concludes with key findings.

As a complementary resource, the code implementing the full methodology presented in this paper is available at https://github.com/RaminBohlouli/ModeChoiceML-With-IMCP.

2. Literature review

Traditional approaches to modeling transportation mode choice have been dominated by discrete choice models (DCMs), such as the Multinomial Logit (MNL) and the Mixed Multinomial Logit (MMNL) models (McFadden, 1973a, 2000b). While widely used, MNL relies on strong assumptions, like the independence of irrelevant alternatives, and struggles to capture complex, non-linear relationships in data. Extensions like MMNL aim to address these limitations by introducing random coefficients to account for unobserved heterogeneity (Ben-Akiva *et al.,* 1985) but often require careful manual specification and offer limited predictive power.

In contrast, the rise of machine learning (ML) has enabled more flexible, data-driven approaches to travel behavior modeling. Recent studies have applied tree-based models (e.g., XGBoost) and neural networks to mode choice tasks, showing higher predictive accuracy than traditional logit models (Bishop, 2006; Hagenauer *et al.,* 2017; Zhao *et al.,* 2020; Cheng *et al.,* 2019). However, most of these models focus exclusively on prediction and overlook interpretability and uncertainty quantification, both of which are essential for risk-aware planning.

In many high-risk fields with high-stakes decisions, such as autonomous driving and medical diagnostics, it is crucial for model users to know the model's confidence levels in their predictions and not to trust them blindly. Uncertainty of model predictions can help users ensure that the model is working as expected, allowing the decision-makers to question or modify model prediction outcomes to alleviate the potential risk when the model exhibits high uncertainty.

Recognizing this gap, researchers have begun to explore techniques to incorporate uncertainty quantification in ML-based models in various fields including autonomous driving (Ravindran *et al.,* 2022), medical image analysis (Chen *et al.,* 2022), security monitoring (Hasan *et al.,* 2019), and road Traffic Accident Severity Assessment (Eldafrawi *et al.,* 2024). However, in the field of transportation mode choice using ML models, the research works focusing on uncertainty quantification are very limited. To our knowledge, the only study addressing uncertainty in ML-based mode choice is Zheng *et al. (*2024), who applied Bayesian Neural Networks (BNNs). While effective, BNNs are not model-agnostic and are limited to specific architectures. This highlights the need for flexible uncertainty quantification tools that can wrap around any predictive model.

Conformal prediction (CP) emerges as a compelling solution to this challenge. CP is a statistically rigorous framework that produces distribution-free prediction sets (rather than single point predictions)

using non-conformity scores under minimal assumptions (Shafer *et al.*, 2008), Its most notable advantage is that it can be applied as a wrapper around any predictive model, without requiring changes to the underlying model. While the original transductive form of CP is computationally intensive (Gammerman *et al.*, 1998), Inductive Conformal Prediction (ICP) (Papadopoulos *et al*., 2002) mitigates this by using a calibration set, enabling scalability. Building on ICP, Inductive Mondrian Conformal Prediction (IMCP) (Vovk *et al.*, 2005), further improves performance in classification tasks by computing conformity scores in a label-conditional manner, partitioning the calibration set by class, thus ensuring more reliable and balanced prediction sets, especially in imbalanced datasets. To the best of our knowledge, our work is the first study that introduces the usage of ICP for transportation mode choice.

3. Material and methods

3.1. Data and preprocessing

The dataset used in this study consists of 7,952 employees in Italy, surveyed in 2024 across 41 companies and 177 office locations. It captures detailed information on employees' mobility habits through a diverse set of questions, grouped into two main categories:

1. Employee-answered questions: socioeconomic characteristics and travel behavior.
2. Company-answered questions: organizational policies and office-level characteristics.

The dataset identifies six main transportation modes, with a distribution depicted in Figure. 1.

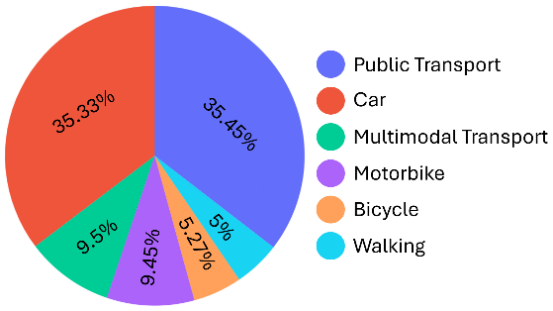

Figure 1. Distribution of Mode Choices among Surveyed Employees

Prior to encoding, the dataset includes 37 features, which are clustered and detailed for both Employee-answered and Company-answered questions in Table 1 and Table 2 respectively.

Table 1. Employee-answered questions

Cluster type	Questions
Socio-Demographic and Contractual	Age, Gender, Contract type, Relationship with company, Number of working days, office
Mobility habits and preferences	Main mode, why this mode, available modes, mode used to reach the closest PT, number of mode changes during work-home commute, departure time from house, office entrance time, work entrance hour rigidity
Cost related	Monthly commute cost, annual PT cost, monthly parking cost
Attitudes, Satisfaction, and Perceptions	Satisfaction level with current mode, Importance rate given to different PT aspects (comfort etc.), Rates given to current available PT, Propensity towards Sharing mobility, carpooling, biking
Private Vehicle and Parking	Fuel type, Emission standard, Parking problem at the office, Access to parking spots at the office, Distance to the closest PT station

Table 2. Company-answered questions

Cluster type	Questions
Infrastructure	Office location, office has changing and shower rooms, Parking availability at the office
Policies	Does company have Bus, Carpooling system, corporate cars?
Employees	Number of employees

Table 3. Encoding Techniques

Question type	Encoding Technique
Multiple Choice (single answer)	One-hot Encoding
Multiple Choice (multiple answers)	Multi-hot encoding
Rating-based	Exploded key–value encodings
Time-based	Cyclical encoding using sine and cosine transformations
Continuous valued	No encoding (Log transformed and Normalized)

In addition to survey questions, several distance and time-based attributes, such as the home-work distance per mode, work-home distance and travel time on foot, were programmatically calculated using GeoServer, based on the geospatial coordinates of employees' residences and their office locations. Given the variety of question types in the dataset, we implemented a pre-processing pipeline to encode the data for ML models (see Table 3). For continuous features, we applied a log transformation to reduce the impact of outliers. To ensure comparability across features and improve model convergence, we performed Z-score normalization using the StandardScaler from the scikit-learn library. The normalization parameters were fit on the training set and then applied to the calibration, test, and cross-validation sets, to avoid any data leakage or contamination between training and evaluation phases.

3.2. Machine learning classifiers

We evaluated three widely-used ensemble models known for their strong performance on structured data: Random Forest (RF), XGBoost, and CatBoost.

Random Forest

This model (Breiman, 2001) builds multiple decision trees using bootstrap samples and random feature subsets, combining their predictions through majority voting. It is robust to noise and overfitting and supports hyperparameter tuning such as tree depth, number of estimators, and minimum sample splits.

XGBoost

This model (Chen *et al.*, 2016) is a gradient boosting algorithm that iteratively fits trees to residual errors, focusing on difficult instances. It supports regularization and fine-grained control via parameters such as learning rate and tree complexity, offering high accuracy and scalability.

CatBoost

This model (Prokhorenkova *et al.*, 2018) extends gradient boosting with native support for categorical features and a unique ordered boosting scheme to prevent overfitting and target leakage. It is particularly effective for tabular datasets with categorical variables.

3.3. IMCP and Wilson Intervals

IMCP

This layer is a probabilistic, class-conditional wrapper around any blackbox classifier (e.g., XGBoost) that outputs prediction sets instead of point predictions. These sets are guaranteed, under minimal distributional assumptions, to contain the true label with a specified confidence level (e.g., 90%).

IMCP uses a calibration set to estimate uncertainty class-conditionally by computing separate thresholds for each class. This is crucial when dealing with imbalanced classes or varying uncertainty across labels. However, IMCP works in expectation (over the randomness of the data and calibration sample) and under two main assumptions: 1 - exchangeability within each class, meaning that the calibration set and the test data must be exchangeable conditional on the class label. 2 - calibration set must be large enough to give meaningful empirical quantiles per class. The workflow of IMCP after the classifier is trained is as follows:

For each calibration example (x_i, y_i), based on the probability vector $\hat{p}^{(i)} = f(x_i)$, the so-called nonconformity score is computed as

$$s_i = 1 - \hat{p}^{(i)}(y_i). \quad (1)$$

This score measures how unlikely the true label appears according to the model. The nonconformity scores are grouped by class l and sorted in ascending order:

$$Scores_{\ l} = \{ s_i \mid y_i = l \}. \quad (2)$$

Quantile thresholds π_l are computed for each class l at confidence level $1-\alpha$, where α is the miscoverage rate. If n_l is the number of calibration examples with label l, then

$$k_l = \lfloor (n_l + 1)(1-\alpha) \rfloor. \quad (3)$$

And π_l is defined as the k_l-th smallest value in (2). For each test instance $x \in X_{test}$ the classifier outputs a probability vector $\hat{p}(x)$. For each class l, the nonconformity score is

$$s_l(x) = 1 - \hat{p}_l(x). \quad (4)$$

Finally, the prediction set $\Gamma(x)$ is

$$\Gamma(x) = \{ l : s_l(x) \leq \pi_l \}. \quad (5)$$

After generating the prediction sets, to evaluate the performance IMCP, the two following metrics are generated:

- Coverage: Proportion of test examples where the true label is included the prediction set:

$$Mondrian\ Coverage = \frac{1}{n_l} \sum_{i:\, y_i = l} 1\{ y_i \in \Gamma(x_i) \}. \quad (6)$$

- Average Set Size: Mean number of labels per prediction set, reflecting efficiency.

To get the overall or marginal prediction sets we should follow the abovementioned steps as well, but with the difference that the classes are processed all together rather than individually.

Wilson Intervals

Given that IMCP works in expectation, we need to consider the finite sample variation effect. This phenomenon is common especially in scenarios where the classes are imbalanced and with it a finite calibration set, even a large one, the empirical coverage will fluctuate around the target level $1-\alpha$. To account for this sampling variability, we compute the Wilson score intervals (Wilson, 1927) around the observed class-wise coverage. This method is preferable to a normal approximation, especially for proportions near 0 or 1, as it ensures bounds remain within [0, 1]:

$$p \in (w^-, w^+) = \frac{1}{1+\frac{z_\beta^2}{n}} \left(\hat{p} + \frac{z_\beta^2}{2n} \pm \frac{z_\beta}{2n} \sqrt{4n\hat{p}(1-\hat{p}) + z_\beta^2} \right), \quad (7)$$

where *n* is the number of samples, β is the desired confidence level, $\hat{p}$ is the coverage, and z_β is the standard-normal interval half-width corresponding to the desired confidence.

3.4. Evaluation metrics

The evaluation metrics for classification include accuracy, precision, recall, and F1-score (Japkowicz *et al*., 2011). Let n_{ij} denote the number of instances whose true class is *i* and that are predicted as class *j*, for a total of *J* classes. The value n_{ii} therefore represents the number of correctly classified instances for class *i*. Accuracy is then defined as:

$$Accuracy = \frac{\sum_{i=1}^{J} n_{ii}}{\sum_{i=1}^{J} \sum_{j=1}^{J} n_{ij}}. \quad (8)$$

However, in imbalanced datasets, accuracy can be misleading. Precision (correctness of positive predictions) and recall (coverage of actual positives) are more informative. Their harmonic mean, the F1-score, balances both:

$$F1 - Score = \frac{2PR}{P+R}. \quad (9)$$

For multi-class tasks, metrics can be aggregated via:

- Macro-average: unweighted mean across classes;
- Weighted average: mean weighted by class frequency.

These strategies ensure a more nuanced performance assessment, especially when class distributions vary.

4. Simulation study

4.1. Simulation setup

After applying a comprehensive preprocessing pipeline for data cleaning and outlier handling, we split the dataset into training + cross-validation (60%), test (20%) and calibration sets (20%). Given that the dataset is imbalanced, we used stratified sampling to ensure that the distribution of the target labels is preserved across all subsets. This is crucial to avoid the under-representation of minority classes during training and evaluation.

To optimize the performance of each ML model, we applied hyperparameter tuning via the following process:

1. Define a grid of possible hyperparameter values for each model.
2. Use GridSearchCV to evaluate all combinations of hyperparameters.
3. For each combination, perform 5-fold stratified cross-validation on the training set.
4. Train on 4 folds and validate on the remaining one, rotating the validation fold.
5. Compute the average accuracy across the 5 folds.
6. Select the hyperparameter combination that yields the highest average accuracy.
7. Refit the best model on the full training set using the selected hyperparameters.

All models were trained on a machine with an AMD Ryzen 7 4700G CPU and 16 GB of RAM. The implementation used the Scikit-learn and XGBoost libraries in Python 3.11.9.

4.2. Performance comparison of ML models

Once training was complete, we evaluated each model using the metrics defined in Section 3.4. The results are summarized in Table 4.

While both XGBoost and CatBoost achieved the same F1 score, XGBoost outperformed CatBoost in terms of macro average, which is especially important in imbalanced classification settings. Therefore, XGBoost was selected as the best model, with its hyperparameters listed in Table 5.

Table 4. Comparison of ML classifiers

Model	F1-Score	Macro-Average	Weighted-Average	Training Time
RF	0.864	0.8	0.86	283.4 s
XGBoost	0.879	0.836	0.879	432.6 s
Catboost	0.879	0.829	0.879	3113.5 s

Table 5. Fine-tuned XGBoost Hyperparameters

Hyperparameter	Maximum Depth	Number of Estimators	Alpha	Lambda	Learning rate
Value	4	200	0.1	0.1	0.1

4.3. Wrapping the best model with IMCP and Wilson Intervals

To quantify predictive uncertainty, we wrapped our best-performing classifier with an IMCP layer. We set a target confidence level of 90%, corresponding to a miscoverage rate of $\alpha = 0.1$.

In conformal prediction, there is a fundamental trade-off between coverage and prediction set size; Higher coverage typically comes at the expense of larger, less informative prediction sets. As depicted in Figure. 2, in our experiment:

- At 99% coverage, the average prediction set size was almost 6, which includes all labels and is thus uninformative.
- At 90% coverage, the average set size was 1.10, which is far more useful for decision-making while still maintaining strong confidence.

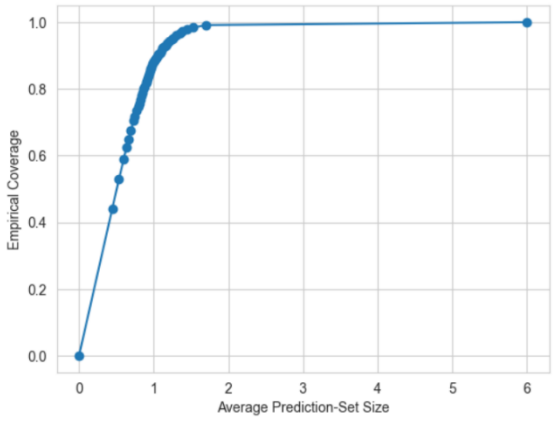

Figure 2. Prediction Set Size vs Average coverage Tradeoff

Ideally, conformal prediction sets should have an average size close to 1, indicating that the model can make precise predictions in most cases. Also, the global Mondrian coverage should be over 90% since we have opted for this level. In our case, these two objectives were achieved: as seen in Figure. 3, most prediction sets were singleton, meaning the model made confident, unambiguous predictions and also the global Mondrian coverage was 90.9%.

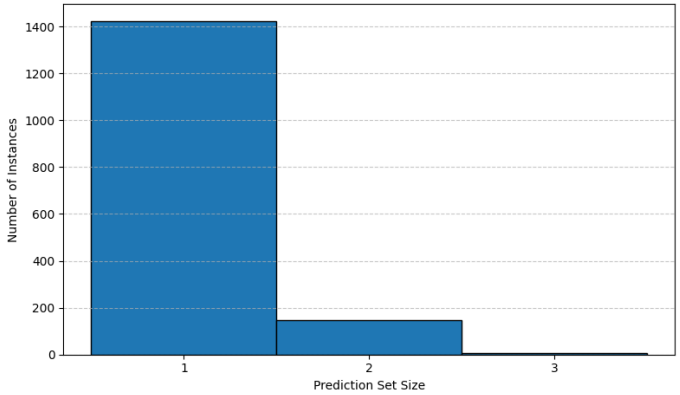

Figure 3. Prediction Set Sizes Histogram (IMCP, 90% Confidence)

As stated before, IMCP extends standard inductive conformal prediction by conditioning on the true class label, ensuring that the coverage guarantee holds per class. We therefore evaluated coverage class-wise using a confidence level of 90%. Then to account for the effect of finite sample variations discussed in Section 3.3, we further enhanced our uncertainty quantification layer with the Wilson intervals to make sure that the calculated coverages are valid on average. The confidence level used for the Wilson interval calculations was the standard β = 95% which corresponds to z = 1.96.

Table 6. Class-conditional Coverage values enforced with Wilson intervals

Class	**Car**	**Multimodal**	**Motorbike**	**Bicycle**	**PT**	**Walking**
Sample size	562	151	150	84	564	79
Coverage	0.915	0.927	0.94	0.9	0.886	0.987
Intervals	(0.89, 0.93)	(0.87, 0.95)	(0.84, 0.94)	(0.87, 0.97)	(0.86, 0.91)	(0.93, 0.99)

Table 6 shows that for all the classes the computed coverage falls within the Wilson intervals. Although the nominal coverage for PT, falls short than the defined desired coverage level (90%), it is still within the defined intervals and hence is acceptable. To ensure that this slight shortfall does not contradict the theoretical guarantees of Mondrian conformal prediction we performed two diagnostic checks:

1. Calibration Score Distribution: Figure. 4 shows the distribution of nonconformity scores for PT in the calibration set. The red vertical line indicates the empirical quantile level (approximately ≈ 0.902)

used for threshold selection with α = 0.1, while the green dashed line marks the class-specific threshold π_3. The distribution was smooth and monotonic near the threshold, indicating a stable and well-behaved quantile estimate.

2. Calibration vs. Test Distribution Check: We also compared the nonconformity score distributions between calibration and test sets for PT using a Kolmogorov–Smirnov test (Frank and Massey, 1951). The result (KS statistic = 0.0284, p-value = 0.9773) confirms no significant distributional shift, ruling out calibration-test mismatch as the source of undercoverage (See Figure 5).

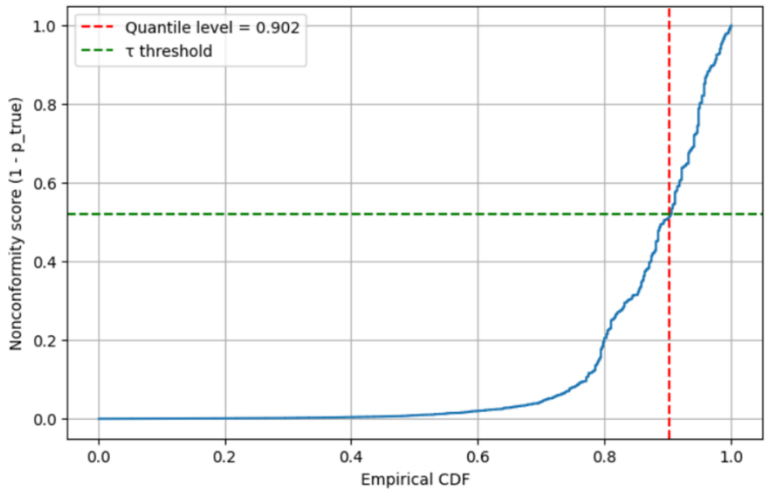

Figure 4. Non-conformity distribution for PT

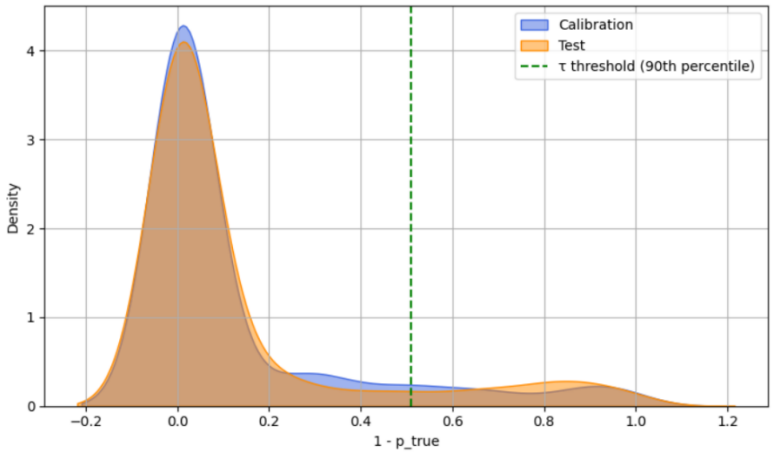

Figure 5. Non-conformity score distribution shift

Given that threshold estimation is stable and no distributional mismatch was found, the deviation is most plausibly due to finite-sample effects which we addressed by defining Wilson intervals.

4.4. Discussion

Based on the prediction set sizes the following scenarios can be interpreted:

- Prediction Set Size = 1: Reflects high-confidence predictions, where the model identified a unique class label within the conformal prediction set. This suggests that the features available for those instances were sufficiently informative and the underlying classifier was confident in its output.
- Prediction Set Size ≥ 2: Signals uncertainty, where the model could not conclusively differentiate between multiple plausible class labels. From a model perspective, these instances represent areas of ambiguity. Possible causes include overlapping feature distributions, insufficient training signal, or intrinsic class similarity.

While these ambiguous predictions are rare, they are particularly interesting from an applied perspective. For example, if the prediction set contains both "Car" and "PT" for a particular individual, this may indicate behavioral ambivalence, suggesting that the person is potentially open to shifting modes given the right external conditions (e.g., incentives or policy nudges). Although such interpretations go beyond the formal guarantees of IMCP, which only ensures valid marginal and class-conditional coverage, they open up practical opportunities for targeted intervention. Specifically:

- **Data-level improvements:** Collecting more samples for borderline cases, engineering more discriminative features, or refining the labeling process.
- **Survey follow-ups:** Deploying clarifying questionnaires to individuals with ambiguous predictions could enhance label certainty and model performance.
- **Policy design:** Using these ambiguous cases to identify candidates for mode-shifting incentives in behavioral transportation studies.

While speculative, this perspective highlights the potential value of prediction set size not only as a measure of uncertainty, but also as a tool for strategic decision-making in mobility policy.

5. Conclusion

This study addressed the lack of uncertainty quantification in machine learning-based mode choice models, a key limitation for risk-aware transportation planning. We first compared the performance of three tree-based ML algorithms (RF, XGBoost, and CatBoost), and then wrapped the best-performing model, XGBoost, with Inductive Mondrian Conformal Prediction (IMCP), enhanced with Wilson intervals to account for finite-sample variability. This approach produced high-accuracy forecasts with formal confidence guarantees, achieving over 90% coverage with mostly singleton prediction sets.

Our findings suggest that conformal prediction can make ML models not only accurate but also more actionable, helping identify both confident and uncertain predictions that may inform further investigation or policy considerations. The model-agnostic nature of the conformal layer adds to its flexibility and practical appeal. Future research could explore integrating SHAP-based explanations with conformal prediction to provide forecasts that are both interpretable and uncertainty-aware at the individual level. This combination would offer a more nuanced understanding of mode choice behavior, highlighting not only likely outcomes but also the features driving confidence or ambiguity. Such insights could support targeted interventions, subgroup fairness analysis, and more efficient data collection, aligning with broader goals in explainable and trustworthy AI for transportation planning.

Acknowledgments

This research was funded by the European Union – Next Generation EU, under the Italian National Recovery and Resilience Plan (PNRR), Mission 4, Component 1, CUP: B53C23002470006.

Declaration of Generative AI and AI-assisted technologies in the writing process

The author(s) acknowledge the use of generative artificial intelligence (AI) and AI-assisted technologies in the preparation of this manuscript. Specifically, OpenAI's ChatGPT was used for language polishing, paraphrasing, and spelling correction purposes. The author(s) reviewed and approved all generated content to ensure accuracy and intellectual integrity. No AI tool was used for data analysis, result generation, or the formulation of scientific conclusions.

References

1. Ben-Akiva, M. E. B., Lerman, S.R. (1985) *Discrete choice analysis: theory and application to predict travel demand.* Cambridge London: MIT Press.
2. Bishop, C. M. (2006) *Pattern recognition and machine learning.* Springer Science+Business Media, LLC.
3. Breiman, L. (2001) Random Forests. *Machine Learning*, 45, 5-32. DOI: 10.1023/A:1010950718922.
4. Chen, C., Qin, C., Ouyang, C., Li, Z., Wang, S., Qiu, H., Chen, L., Tarroni, G., Bai, W., Rueckert, D. (2022) Enhancing MR image segmentation with realistic adversarial data augmentation. *Medical Image Analysis*, 82,102597, DOI: 10.1016/j.media.2022.102597.
5. Chen, T., Guestrin, C. (2016) XGBoost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (KDD '16). Association for Computing Machinery, 785–794. DOI:10.1145/2939672.2939785.

6. Cheng, L., Chen, X., De Vos, J., Lai, X., Witlox, F. (2019) Applying a random forest method approach to model travel mode choice behavior. *Travel Behaviour and Society*, 14, 1-10, DOI: 10.1016/j.tbs.2018.09.002.
7. Eldafrawi, M., Varghese, K. K., Afsari, M., Babapourdijojin, M. (2024) Machine learning-enhanced conformal prediction approach for road traffic accident severity assessment: A case study of Rome. DOI: 10.2139/SSRN.4679159.
8. Forman, G., Scholz, M. (2010) Apples-to-apples in cross-validation studies: Pitfalls in classifier performance measurement. *ACM SIGKDD Explorations Newsletter*, 12(1), 49–57. DOI: 10.1145/1882471.1882479.
9. Frank, K., Massey, JR. (1951) The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253), 68–78. DOI:10.2307/2280095.
10. Gammerman, A., Vovk, V., Vapnik, V. (1998) Learning by transduction. In: *Proceedings of Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI '98)*, Madison, July 1998. San Francisco: Morgan Kaufmann Publishers Inc, 148–155. arXiv:1301.7375.
11. Hagenauer, J., Helbich, M. (2017) A comparative study of machine learning classifiers for modeling travel mode choice. *Expert Systems with Applications*, 78, 273–282. DOI: 10.1016/j.eswa.2017.01.057.
12. Hasan, M., Dipto, A. Z., Islam, M. S., Sorwar, A., Alam, S. (2019) A smart semi-automated multifarious surveillance bot for outdoor security using thermal image processing. *Advances in Networks*, 7(2), 21–28. DOI: 10.11648/j.net.20190702.12.
13. Jahangiri, A., Rakha, H. A. (2015) Applying machine learning techniques to transportation mode recognition using mobile phone sensor data. *IEEE Transactions on Intelligent Transportation Systems*, 16(5), 2406—2417. DOI:10.1109/TITS.2015.2405759.
14. Japkowicz, N., Shah, M. (2011) *Evaluating learning algorithms: A classification perspective*. Cambridge University Press. DOI: 10.1017/CBO9780511921803.
15. Kohavi, R. (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *Proceedings of the 14th international joint conference on Artificial intelligence*, 2, 1137–1143. San Francisco: Morgan Kaufmann Publishers Inc. DOI: 10.5555/1643031.1643047.
16. McFadden, D. (1973) *Conditional logit analysis of qualitative choice behavior*. Available at: https://eml.berkeley.edu/reprints/mcfadden/zarembka.pdf.
17. McFadden, D., Train, K. (2000) Mixed MNL models for discrete response. *Journal of Applied Econometrics*, 15(5), 447–470. DOI:10.1002/1099-1255(200009/10)15:5<447::AID-JAE570>3.0.CO;2-1.
18. Papadopoulos, H., Proedrou, K., Vovk, V., Gammerman, A. (2002) Inductive confidence machines for regression. In: *Proceedings of the 13th European Conference on Machine Learning (ECML'02*). Springer, 45-–356. DOI:10.1007/3-540-36755-1_29.
19. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., Gulin, A. (2018) CatBoost: Unbiased boosting with categorical features. In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, Montréal, December 2018. New York: Curran Associates Inc., 6639–6649. arXiv:1706.09516.
20. Ravindran, R., Santora, M. J., Jamali, M. M. (2022) Camera, LiDAR, and Radar Sensor Fusion Based on Bayesian Neural Network (CLR-BNN). *IEEE Sensors Journal*, 22(7), 6964–6974. DOI: 10.1109/jsen.2022.3154980.
21. Salas, P., De la Fuente, R., Astroza, S., Carrasco, J.A. (2022) A systematic comparative evaluation of machine learning classifiers and discrete choice models for travel mode choice in the presence of response heterogeneity. *Expert Systems with Applications*, 193(1784), 116253. DOI:10.1016/j.eswa.2021.116253.
22. Shafer, G., Vovk, V. (2008) A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9, 371–421. arXiv:0706.3188.
23. Vovk, V., Gammerman, A., Shafer, G. (2005) *Algorithmic learning in a random world.* 1st edition. Berlin: Springer Verlag. DOI:10.1007/b106715.
24. Wilson, E. B. (1927) Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158), 209–212. DOI: 10.2307/2276774.
25. Zhao, X., Yan, X., Yu, A., Van Hentenryck, P. (2020) Prediction and behavioral analysis of travel mode choice: A comparison of machine learning and logit models. *Travel Behaviour and Society*, 20, 22–35. DOI:10.1016/j.tbs.2020.02.003.
26. Zheng, S., Fang, Z., Zhao, L. (2024) Incorporating uncertainty quantification into travel mode choice modeling: A Bayesian neural network (BNN) approach and an uncertainty-guided active survey framework. DOI:10.48550/arXiv.2406.10948.