FROM ACCURACY TO CALIBRATION: EVALUATING MACHINE LEARNING AND DISCRETE CHOICE MODELS FOR POLICY SIMULATION IN TRAVEL MODE CHOICE

Cristian Besliu
HEC Lausanne
MSc. in Management (Business Analytics)

Abstract

This study evaluates whether Machine Learning (ML) models can deliver policy-relevant travel mode-choice predictions and behavioural sensitivities, beyond the classical discrete choice modelling (DCM) framework. Using a London passenger travel choice dataset with four modes: public transport, car, cycling, and walking, we estimate baseline DCMs (multinomial and nested logit) and compare them to three ML approaches regarded as best practice for tabular prediction: Random Forest, gradient boosting (XGBoost), and a multi-layer perceptron neural network. A central methodological contribution is to show that model selection based on classification accuracy can be misleading for transport simulation.

Building on these calibrated probability models, we conduct counterfactual policy simulations and compute implied elasticities across multiple scenario magnitudes. Results highlight that ML models can capture non-linear, heterogeneous sensitivities without manually specifying interactions, but they may imply different substitution patterns than DCMs, underscoring the importance of behavioural validation alongside predictive metrics. Overall, the findings support a complementary use case in which DCMs provide interpretable behavioural benchmarks while probability calibrated ML models enhance scenario-based policy evaluation.

1 Introduction

Urban transport systems face increasing pressure from rapid urbanisation, congestion, and the need to reduce greenhouse gas emissions. Sustainable mobility policies implemented across Europe demonstrate that well-designed transport systems can shift behaviour toward public transport, cycling, and walking while improving environmental attributes and social well-being. For example, cities such as Amsterdam, Copenhagen, and Vienna have achieved significant reductions in private car use through integrated cycling networks, car-free zones, and multimodal public transport [1]. At the same time, comparative studies show that modal preferences are shaped by complex interactions between trip characteristics, socio-demographics, and accessibility factors, each operating differently across cities and individuals [4].

In this context, accurately modelling and predicting individual transport mode choice has become central to designing effective mobility policies. Classical DCMs such as the Multinomial Logit (MNL) and Nested Logit (NL) remain the tools at the forefront of behavioural sensitivities and policy impacts estimations.

However, recent advances in ML have opened new possibilities for modelling non-linear dependencies and interactions that classical DCMs may fail to capture. Tree-based ensemble models—including Random Forests (RF) (bagging of decision trees), Gradient Boosting Machines (GBM) (sequential boosting of trees, e.g., XGBoost), as well as Neural Networks (NNs), have shown strong predictive performance on tabular transport datasets, potentially outperforming MNL in accuracy and flexibility. Yet, a systematic comparison between traditional DCMs and contemporary ML approaches, while also ensuring interpretability through elasticities and policy simulations, remains underexplored in academic literature.

This project uses the London Passenger Mode Choice (LPMC) dataset [3], which has details for 81,086 trips taken by 31,954 individuals across 17,616 households from April 2012 to March 2015. By combining classical DCMs with ML methods, the project aims to produce actionable insights for designing sustainable mobility policies in urban environments.

2 Literature Review

2.1 ML vs. DCMs: higher predictive performance, but at a cost

A consistent finding across the recent mode-choice literature is that ML models often outperform classical DCMs, including MNL and NL, when the objective is predictive accuracy.

The core reason is structural. Standard DCMs impose functional-form assumptions (e.g., linear-in-parameters utility, fixed substitution patterns, and Independence of Irrelevant Alternatives (IIA) at the top level), which can under-fit behaviour when mode choice is shaped by non-linearities and interactions. In contrast, ML methods can learn complex decision boundaries and interactions, which tends to translate into better out-of-sample predictive performance [2].

However, the same literature emphasises that the predictive edge of ML comes with costs that matter for transport policy work.

1. Interpretability: DCMs offer parameters that map naturally onto behavioural constructs (marginal utilities of time and cost, implied Value of Time/Willingness to Pay, substitution patterns between alternatives). Many ML models provide probabilities without structural behavioural parameters, creating what the systematic review terms a risk of "behavioural black-boxing" [2].

2. Transferability: The ML literature shows reliance on internal validation only; in contrast, transport applications often require transfer across contexts (new travel routes/links, policy regimes, cities) [2, 4].

In summary the literature highlights that while ML models often outperform in terms of accuracy they lack the interpretability and stability required to robustly simulate scenarios.

2.2 Can ML Uncover Realistic Behavioural Sensitivities?

Against this backdrop, the key methodological question motivating my research is not simply "which model predicts best?", but:

Does training ML mode-choice models for probabilistic accuracy, fit via the Negative Log-Likelihood (NLL; i.e., cross-entropy/log loss), result in behaviourally plausible elasticities, comparable to classical DCMs, when evaluated through multi-scenario simulation?

This question is timely because the principal advantage of ML in mode choice (its ability to capture non-linear patterns) is what makes it advantageous for scenario testing. Non-linear models can produce elasticities without the manual specification of interaction terms. In principle, this matters for policy, because many interventions are not uniform: pricing changes, service changes, and access/egress improvements are expected to have heterogeneous impacts that linear specifications can miss[2, 6].

At the same time, non-linear flexibility raises a risk: elasticities can become unstable. Hassan et al., (2025) note that unlike traditional models, elasticities in ML models are often derived from numerical sensitivity checks and are "not guaranteed to hold under new policy settings", making them structurally unstable for forecasting. Even with SHAP/permutation methods, interpretation is post-fit and can fail to ensure behavioural plausibility [2]. This justifies the expanding direction of behaviourally constrained ML, where predictive training is regularised by penalties for implausible responses.

Sub-Questions:

2. How does the training objective (accuracy vs. NLL) affect probability calibration and aggregate market-share replication—especially for under-represented modes such as cycling?

3, Do ML-derived implied elasticities exhibit economically logical behaviour (correct signs, monotonicity, reasonable magnitudes) and are they stable across scenario magnitudes, relative to elasticities implied by MNL/NL?

4. Under policy scenarios, do ML and DCMs predict different substitution patterns?

2.3 How do Simulation Results Differ by Model Choice?

A distinctive requirement in transport planning is that models are often used in simulation, where trips are assigned to modes probabilistically. This shifts attention away from accuracy and towards metrics that reflect the quality of the predicted probabilities.

Hillel et al., (2018) are explicit on this point. They argue that relying on classification error is misaligned with the main use case of mode choice models in simulation: models should be judged by how well they produce calibrated probabilities that replicate modal shares. They highlight the "accuracy paradox" under class imbalance (rare modes like cycling) and motivate NLL as a more appropriate metric for probabilistic prediction [3].

Mode choice can affect simulation outcomes in at least three ways:

1. Probability calibration: Two models can have similar accuracy but very different probability quality; calibration differences contaminate simulated modal shares [3, 2].

2. Substitution logic under counterfactuals: DCMs impose structured substitution patterns; ML may reproduce observed substitution in-sample, but can behave unpredictably under new policy regimes unless constrained [2].

3. Heterogeneous treatment effects: Non-linear ML may reveal stronger distributional heterogeneity, which can materially change policy conclusions even if aggregate accuracy is similar [6, 2].

3 Methodology

This study adopts a mixed-methods modelling framework that combines classical DCMs with ML approaches to evaluate and predict individual transport modal shares.

3.1 Data and Experimental Design

The first stage of the methodology involves data preparation. All 32 variables in the LPMC dataset are checked for missing values (none found). Categorical variables are one-hot encoded, while continuous features are z-score standardised where required. Although some feature-engineered variables are already available in the dataset, additional variables motivated by the transport economics literature[4] are computed:

Friction variables

- `has_interchange`: Dummy indicating whether the trip includes a public transport interchange (e.g., bus-to-rail).
- `multi_interchange`: Dummy indicating whether the public transport trip includes more than one interchange (≥ 2).
- `is_bus_only`: Dummy indicating whether the public transport trip consists only of bus travel.
- `is_rail_only`: Dummy indicating whether the public transport trip consists only of rail travel.
- `is_mixed_pt`: Dummy indicating whether the public transport trip combines rail and bus segments.
- `pt_in_vehicle`: Total in-vehicle public transport travel time, defined as the bus element duration plus the rail element duration.

The second stage estimates two classical DCMs. MNL serves as the baseline, allowing the estimation of mode-specific coefficients (betas) and behavioural elasticities. This model provides a behavioural structure grounded in utility maximisation but suffers from the restrictive IIA assumption. To address this, a NL is estimated as an extension, which relaxes the IIA assumption.

The third stage implements three ML models selected to represent current best practice in tabular predictive modelling. A RF classifier is first estimated as a baseline capable of capturing non-linearities and interaction effects. This is followed by XGBoost, which typically achieve state-of-the-art performance on structured datasets due to their gradient-boosted decision tree architecture. Finally, a NN is trained to test the potential of deep learning for capturing high-dimensional relationships. To ensure fair comparison, all ML models are trained on the same feature set used in the DCMs.

Hyperparameters are tuned via stratified cross-validation to preserve minority-mode representation in each fold. Importantly, because the objective is policy simulation and market-share replication, models are tuned primarily using NLL. Performance evaluation therefore combines classification metrics—accuracy, per-mode precision/F1, and confusion matrices. The probability- and simulation-relevant diagnostics, include testing the market-share error (the deviation between predicted and observed aggregate mode shares).

Finally, all models are subjected to two policy simulation experiments. These simulations assess how predicted mode shares change when key variables are modified to mirror realistic interventions. For the MNL model, new predicted mode shares are computed using the estimated utility functions. For ML models, predictions are generated under counterfactual inputs, and changes in class probabilities are analysed. For comparability (and due to multiple counterfactual input changes per policy scenario), elasticity estimations for both the MNL and ML models are derived

from the policy simulation results using classical elasticity theory.

3.2 Classical Elasticity Theory

In classical demand theory, elasticity measures the responsiveness of an outcome variable to marginal changes in an explanatory variable. In the context of DCMs, elasticities are defined as partial derivatives of choice probabilities or market shares with respect to an attribute of interest, normalised by the ratio of the attribute level to the outcome level (used to compute the MNL elasticities in Table 4). Formally, the elasticity of the probability of choosing alternative j with respect to attribute x_j is given by

$$\varepsilon_{j,x} = \frac{\partial P_j}{\partial x_j} \cdot \frac{x_j}{P_j}.$$

However, these properties rely on structural assumptions, including separability of attributes, linearity in parameters, and fixed substitution patterns. When these assumptions are relaxed, as in more flexible model classes (e.g., RF, NN, GBM), it motivates the use of scenario-based elasticity measures instead.

3.3 Scenario-based Elasticity Measures

Scenario-based elasticities were computed from calibrated model predictions under counterfactual policy scenarios. For each model and each scenario magnitude m, we first generated a counterfactual dataset by modifying only the policy-relevant attributes (e.g., driving cost and/or travel time for Scenario 1; PT access/wait time and transfer penalty for Scenario 2), while holding all other attributes fixed. The model then produced calibrated choice probabilities $\hat{P}_{n,j}^{(m)}$ for each individual n and alternative j. Predicted market shares were obtained by averaging these probabilities across individuals:

$$s_j^{(m)} = \frac{1}{N}\sum_{n=1}^{N} \hat{P}_{n,j}^{(m)}, \qquad s_j^{(0)} = \frac{1}{N}\sum_{n=1}^{N} \hat{P}_{n,j}^{(0)},$$

where $m = 0$ denotes the baseline (no policy change).

For each alternative j, the scenario impact was summarised by the absolute change in share,

$$\Delta s_j(m) = s_j^{(m)} - s_j^{(0)},$$

and the relative change in share,

$$\%\Delta s_j(m) = \frac{s_j^{(m)} - s_j^{(0)}}{s_j^{(0)}}.$$

When a single attribute x was perturbed by a proportional amount m (e.g., +10% cost), the implied scenario-based elasticity of share with respect to x was computed as the ratio of the percentage change in share to the percentage change in the attribute:

$$\varepsilon_{j,x}^{\text{scen}}(m) = \frac{\%\Delta s_j(m)}{m},$$

This definition yields an elasticity that is explicitly indexed by the scenario magnitude, allowing sensitivities to vary with policy intensity.

For bundled scenarios in which two attributes were changed simultaneously, elasticities were computed separately for each attribute. Specifically, for a bundled scenario involving proportional changes m_1 and m_2 in attributes x_1 and x_2, two attribute-specific elasticities were reported:

$$\varepsilon_{j,x_1}^{\text{scen}}(m) = \frac{\%\Delta s_j(m)}{m_1}, \qquad \varepsilon_{j,x_2}^{\text{scen}}(m) = \frac{\%\Delta s_j(m)}{m_2}.$$

These elastic should be interpreted as normalised sensitivities by scenario rather than partial derivatives, as they capture the total response of market shares to joint changes while attributing that response to each policy component individually.

4 Models

Having defined the dataset, evaluation metrics, and the counterfactual simulation framework, we now turn to the choice models used in the empirical comparison. Section 4 describes the classical DCM baselines and the ML models, along with their estimation and tuning procedures.

4.1 Classical DCMs

Classical DCMs such as MNL and NL form the empirical foundation of transport mode choice analysis and are grounded in Random Utility Theory (RUT). Under RUT, each individual n associates a latent utility $U_{n,m}$ with every available mode m and chooses the mode with the highest utility. The utility function is expressed as:

$$U_{n,m} = V_{n,m} + \varepsilon_{n,m},$$

where $V_{n,m}$ is the systematic (observable) component and $\varepsilon_{n,m}$ is a random error term. Assuming that the error terms are independently and identically Gumbel-distributed leads to the MNL model. The choice probability for mode m is given by:

$$P_n(m) = \frac{\exp(V_{n,m})}{\sum_{j\in\mathcal{M}} \exp(V_{n,j})},$$

where $\mathcal{M}$ denotes the set of available modes.

In this study, we specify classical utility functions for the four relevant travel modes: Public Transport (PT), driving, cycling, and walking. Mode utilities are constructed using travel time, monetary cost (for PT and driving), and detailed PT components such as in-vehicle time (bus and rail), access walking time, transfer walking and waiting times, and the number of interchanges. This decomposition is consistent with empirical transport literature[5], where different time components are assumed to carry different behavioural weights (e.g., waiting time and transfers are perceived more negatively than in-vehicle time). The

utility functions have been specified additively following a forward-selective approach. Both the MNL and NL model was specified and estimated using the Biogeme software in python.

$$
\begin{aligned}
V_{\text{PT}} &= \text{ASC}_{\text{PT}} \\
&+ B_{\text{COST}} \cdot c_{\text{transit}} \\
&+ B_{\text{VEH\_TIME}} \cdot (dur\_pt\_rail + dur\_pt\_bus) \\
&+ B_{\text{WALK\_TIME}} \cdot (dur\_pt\_access + dur\_pt\_int\_walking) \\
&+ B_{\text{WAIT\_TIME}} \cdot dur\_pt\_int\_waiting \\
&+ B_{\text{TRANSFERS}} \cdot pt\_n\_interchanges \\
&+ B_{\text{AGE\_18\_34}} \cdot age\_18\_34 \\
&+ B_{\text{DAY\_OF\_WEEK\_PT}} \cdot day\_of\_week
\end{aligned} \tag{4.1}
$$

$$
\begin{aligned}
V_{\text{CAR}} &= \text{ASC}_{\text{CAR}} \\
&+ B_{\text{TIME}} \cdot dur\_driving \\
&+ B_{\text{COST}} \cdot c\_driving\_total \\
&+ B_{\text{COST\_X\_AGE}} \cdot c\_driving\_total \cdot age \\
&+ B_{\text{TRAFFIC}} \cdot driving\_traffic\_percent \\
&+ B_{\text{AGE\_35\_54}} \cdot age\_35\_54 \\
&+ B_{\text{FEMALE}} \cdot female \\
&- B_{\text{N\_HOUSEHOLD\_2\_5}} \cdot n\_household\_2\_5 \\
&+ B_{\text{CAR\_OWNERSHIP}} \cdot car\_ownership \\
&+ B_{\text{DRIVING\_LICENSE}} \cdot driving\_license \\
&+ B_{\text{START\_TIME\_LINEAR\_CAR}} \cdot start\_time\_linear \\
&+ B_{\text{DAY\_OF\_WEEK\_CAR}} \cdot day\_of\_week
\end{aligned} \tag{4.2}
$$

$$
\begin{aligned}
V_{\text{BIKE}} &= \text{ASC}_{\text{BIKE}} \\
&+ B_{\text{TIME}} \cdot dur\_cycling \\
&+ B_{\text{DISTANCE\_CYCLE}} \cdot distance
\end{aligned} \tag{4.3}
$$

$$
\begin{aligned}
V_{\text{WALK}} &= \text{ASC}_{\text{WALK}} \\
&+ B_{\text{TIME}} \cdot dur\_walking \\
&+ B_{\text{DISTANCE\_WALK}} \cdot distance
\end{aligned} \tag{4.4}
$$

4.2 NL Model

While the MNL model provides a useful baseline, it relies on the restrictive IIA assumption. IIA implies that the relative odds of choosing between any two transport modes are unaffected by the presence or attributes of other alternatives. In practice, this assumption may be violated when certain modes are closer substitutes due to shared unobserved attributes.

The NL model relaxes the IIA assumption by allowing alternatives within a nest to share correlated unobserved utility components. Choice probabilities are decomposed into the probability of selecting a nest and the conditional probability of selecting an alternative within that nest.

In this project, a theoretically motivated nesting structure was specified to capture potential correlation among sustainable transport modes. Specifically, PT, walking, and cycling were grouped into a *non-car (sustainable)* nest, while car travel was treated as a singleton alternative:

- **Non-car (Sustainable) Nest**: {PT, walking, cycling},
- **Car**: {driving}.

Estimating the NL model allows the dissimilarity parameter μ_k associated with each nest to be tested empirically. Under Random Utility Maximisation (RUM), the NL model requires $0 < \mu_k \leq 1$. Values of μ_k significantly below one indicate positive correlation in unobserved utility within the nest and justify the nested structure. In contrast, values equal to one imply that the nest collapses to the standard MNL model, indicating no additional correlation beyond what is already captured by observed attributes.

The estimated dissimilarity parameter for the non-car nest, $\mu_{\text{non-car}}$, is statistically significant and takes a value of 1.032 (see $\mu_{\text{non-car}}$ in Table 5). Although statistically different from unity, this estimate exceeds the theoretical upper bound imposed by RUT. A value of $\mu_k > 1$ implies negative correlation in unobserved utilities within the nest, which is inconsistent with the behavioural interpretation of NL models and violates the underlying random utility framework.

Given this finding, the NL specification is rejected, and the analysis proceeds using the standard MNL model. This choice ensures consistency with RUT.

4.3 Tree-Based Methods

Tree-based ensemble models provide a non-parametric alternative to classical DCMs by capturing non-linear relationships between variables. Among these, the RF classifier is suitable for tabular datasets such as the LPMC, as it incorporates built-in regularisation through bootstrapping and feature subsampling.

We tune a RandomForestClassifier using grid search with 3-fold cross-validation. The optimal hyperparameters were:

Table 1: Random Forest Hyperparameters

Hyperparameter	**Value**
max_depth	20
max_features	$\sqrt{\text{features}}$
min_samples_split	2
n_estimators	150

These hyperparameters result in a cross-validation NLL of 0.5903, indicating strong generalisation relative to the untuned model.

4.4 Gradient Boosting Models

We benchmarked DCMs against GBMs using XGBoost (XGB), which sequentially builds shallow trees to minimise a differentiable loss. Following recent evidence that probability-based objectives are essential when models are used for market-share prediction and policy simulation, hyperparameters were selected by minimising NLL rather than maximising accuracy, in line with the probability-

calibration focus recommended in the mode-choice ML literature [2].

A Bayesian hyperparameter search with scoring set to NLL, a learning rate of 0.01 (as recommended by Hassan et al., 2025), and 3-fold cross-validation (literature recommends 10-fold cross-validation [2], however 3-fold was selected to optimise processing time) yielded the following best configuration:

Table 2: Best XGB hyperparameters (Bayesian optimisation).

Hyperparameter	Value
`colsample_bylevel`	1.0
`colsample_bytree`	0.5
γ	0.001
`max_delta_step`	10.0
`max_depth`	14
`min_child_weight`	1.0
α	0.001
λ	4.0
`subsample`	1.0

achieving a best cross-validation NLL of 0.4946. The model was then retrained with early stopping (patience of 50 rounds), setting a high upper bound on the number of boosting rounds and letting validation NLL determine the effective ensemble size. Early stopping selected 1595 boosting iterations (from a maximum of 4000), indicating that additional trees beyond this point degraded out-of-sample probability fit.

Initially, the model produced below par aggregate probability predictions for the minority cycling class (1.54% vs approximately 2.9% for the NN and RF), however the aggregate market share predictions were optimised by increasing the `min_child_weight`, which prevents the model from creating highly specific leaves that may lead to overconfident probabilities, from 1 to 30. This helped produce highly accurate aggregate market share predictions for the XGB.

4.5 Neural Network and Multilayer Perceptron

The NN model used in this study is a MLP, i.e., a feed-forward neural network composed of stacked fully connected layers. Each hidden layer applies an affine transformation followed by a non-linear activation function, producing increasingly abstract feature representations. A final softmax layer converts the resulting latent scores into a valid probability distribution across the four alternatives.

4.5.1 Neural Architecture Search

A systematic architecture search was conducted over alternative depths and layer widths (*Small, Medium, Deep, Large, Wide*) and activation functions (`relu`, `tanh`, `swish`), while keeping batch normalisation and dropout as regularisation components. Architectures were trained with early stopping on validation loss and evaluated out-of-sample using categorical cross-entropy, which is equivalent to the NLL in multi-class settings. This criterion is therefore aligned with the objective of obtaining well-calibrated choice probabilities, rather than maximising accuracy alone.

Across the candidate specifications, the lowest test loss was obtained by a wide two-layer network (512–256) with `swish` activation (test loss ≈ 0.635). However, the *Small* specification (64–32) with `swish` achieved a very similar test loss (test loss ≈ 0.643) at a fraction of the computational cost (4,964 parameters versus 154,372 for the wide network, and substantially lower training time) (see Table 3 for the full NAS results). Given the marginal performance difference relative to the efficiency gains and the reduced risk of overfitting, the *Small (64–32) + swish* network was selected for all subsequent scenario simulations and elasticity analysis.

The hidden layers use the `swish` activation, which improved convergence relative to ReLU. Batch normalization stabilised training, and dropout (0.2) was required to reduce overfitting. A softmax output layer generates class probabilities for the four transport modes.

4.5.2 Training Procedure

The model was trained using categorical cross-entropy with one-hot encoded labels for the four alternatives. Input features were standardised using zero-mean and unit-variance scaling. Training was performed using the Adam optimizer with a learning rate of 0.01, consistent with previous literature [3].

Early stopping patience of 10 was applied based on validation loss to prevent over-fitting and to ensure comparable probability calibration across candidate architectures. The best validation loss when training the NN was achieved with a total of 53 epochs before training was automatically stopped.

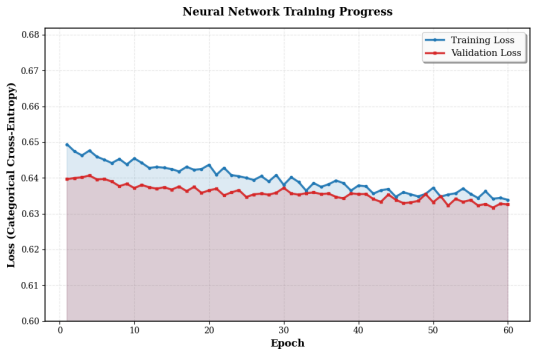

Fig. 1: Neural Network Training Progress

4.5.3 Results

Under the selected *Small (64–32) + swish* specification, the neural network achieved a test categorical cross-entropy of approximately 0.643. The corresponding overall test accuracy for this configuration was 0.751.

5 Model Comparison

This section compares classical DCM and ML approaches along two dimensions that matter for policy analysis. First, we evaluate predictive performance using both hard-class metrics (e.g., accuracy) and probability-based criteria (e.g., NLL) to distinguish models that merely classify well from those that produce simulation-ready probabilities.

5.1 Why Training on Accuracy Looked "Great" but Produced Bad Market Shares

When mode choice models are used for planning and policy, the output of interest is rarely the single most likely mode for each trip. Instead, transport simulation typically needs quality choice probabilities so that aggregate market shares are reproduced and scenario-induced shifts in probabilities can be interpreted as behavioural response. This distinction explains why models trained to maximise classification accuracy can perform well on a standard test set yet fail at the level that matters for simulation.

Accuracy optimizes a 0–1 decision rule (argmax class). That objective is indifferent to probability quality as long as the top-ranked class is correct. Under class imbalance, this leads to a predictable failure mode: the model is rewarded for being "confident" in majority classes, while minority classes (i.e., cycling) contribute little to the objective. The result can be very high accuracy while the probability mass assigned to minority classes is too low, producing poor market-share replication.

For example, when tuning the XGB using accuracy as the objective, we observed a clear calibration failure for minority modes. In particular, the model predicted a cycling market share of 0.4%, compared with 2.9% observed—a substantial underestimation relative to the other models. In our case, the accuracy-optimised configuration selected a much smaller ensemble (e.g., n_estimators = 300, learning_rate = 0.1) than the NLL-optimized configuration (e.g., n_estimators = 1565, learning_rate = 0.01), yielding accurate class predictions overall but poorly calibrated probabilities.

5.2 Market-share Calibration vs. Class-wise Predictive Performance

Figure 2 compares observed and predicted market shares across the four travel modes. ML models were tuned using the NLL objective. After NLL-based optimisation, all models closely reproduce the aggregate market shares, including for the under-represented *Cycle* class (observed share: 2.9%).

While the market-share results indicate that all models

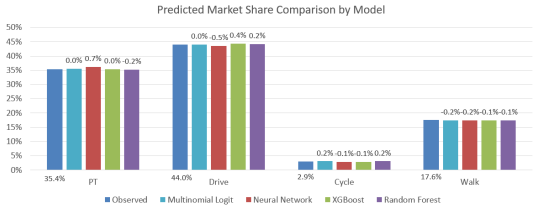

Fig. 2: Market Share Projections by Model

are suitable for aggregate forecasting, Figure 3 highlights a different aspect of performance: the ability to correctly identify rare choices at the individual level. The class-wise F1-score is based on hard class assignments (e.g., arg max predicted probability) and therefore penalises models that poorly select a rare class as the most likely outcome, even if they assign it non-trivial probability. This is visible for cycling: both the MNL and NN achieve an F1-score of 0% for cycling, implying that cycling is almost never predicted as the top class. In contrast, RF and XGB achieve non-zero F1-scores for cycling (9% and 5% respectively), suggesting that the tree-based models capture some of the non-linear patterns and interactions that separate cyclists from the other modes.

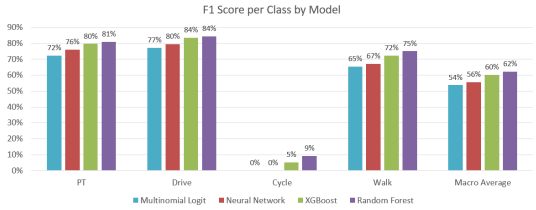

Fig. 3: F1 Score per Class by Model

Good aggregate market shares can be obtained even when a rare mode is seldom the top predicted class. Nevertheless, individual-level discrimination remains important for policy simulation because scenario impacts depend on whether probability mass is concentrated on the trips and travellers who are realistically near the switching margin; diffuse or misplaced probability mass can reproduce baseline shares while producing misleading scenario-induced shifts and segment-level conclusions.

6 Policy Scenarios and Simulation

The next section moves beyond predictive performance and focuses on what matters for transport policy: the implied behavioural responses. We compare elasticities and scenario-based simulations across modelling approaches by applying counterfactual changes to key attributes and examining the resulting changes in predicted market shares.

6.1 Policy Scenario 1

Policy Scenario 1 simulates a congestion-pricing–type intervention in which driving costs increase by 10%, 20%, and 30%, accompanied by proportional increases in driving time of 2%, 5%, and 8%, respectively. This joint shock captures a short-run policy environment where higher monetary charges coexist with congestion-related travel time penalties. The scenario is evaluated using calibrated market share predictions from the MNL, NN, XGB, and RF models.

Across all model classes, the policy produces a consistent directional response: driving mode share decreases monotonically as policy intensity increases, while PT gains market share. This consistency suggests that all models capture the basic economic intuition that higher generalised driving costs reduce car attractiveness. However, the substitution patterns differ across models (see Figure 4).

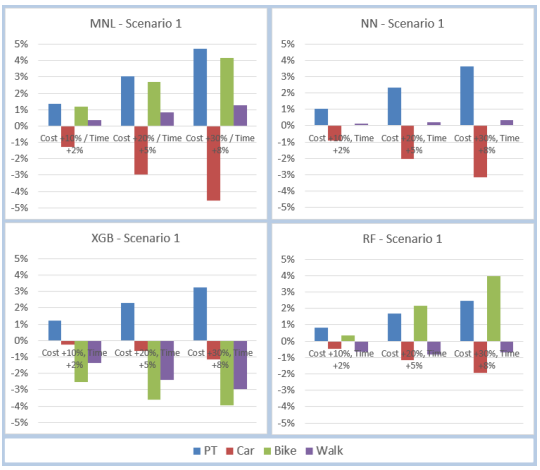

Fig. 4: Scenario 1 - Predicted Market Shares by Model

The MNL model exhibits a proportional decline in car share across policy magnitudes, with corresponding linear gains in PT. This pattern reflects the linear-in-parameters utility specification and the absence of endogenous interaction effects, resulting in relatively stable implied elasticities across scenarios.

Interestingly, similarly to the MNL, the NN displays a flatter response at higher policy intensities (although lower in magnitude), with diminishing marginal reductions in car share beyond the initial cost–time increase (see Figure 6). This saturation effect suggests that the NN captures non-linear thresholds in mode switching behaviour, whereby further increases in driving disutility yield smaller behavioural adjustments once a subset of highly car-dependent individuals remains insensitive to additional penalties.

The tree-based models exhibit unique substitution mechanisms. The RF predicts a monotonic but moderately non-linear response, with small yet consistent gains in cycling alongside PT, indicating that the model captures interaction effects without enforcing global smoothness. XGB, by contrast, reallocates demand more strongly toward PT while simultaneously reducing walking and cycling shares, suggesting more complex cross-elasticities between non-car modes.

6.2 Policy Scenario 2

Policy Scenario 2 simulates a PT service improvement package combining reductions in access and waiting time with proportional reductions in transfer penalties. Specifically, PT access/wait times and transfer penalties are jointly reduced by 10%, 20%, and 30%. This scenario captures realistic operational improvements such as increased service frequency.

Across all models, the policy generates a strong and monotonic increase in PT mode share, accompanied by a substantial decline in car usage. Compared to Policy Scenario 1, the magnitude of the response is larger (see Figure 5), indicating that improvements in PT level of service generate stronger demand shifts than moderate increases in driving disutility.

Fig. 5: Scenario 2 - Predicted Market Shares by Model

While all models agree on the direction of effects, they imply different substitution mechanisms and degrees of non-linearity. The MNL model predicts the strongest and most linear response, whereas the ML models—particularly NN and XGB, display similar effects to the MNL at a smaller magnitude. The RF exhibits a stronger car substitution pattern, with cycling and walking remaining unaffected, which suggests endogenous interaction effects and non-linear substitution patterns.

6.3 Policy Scenarios Conclusion and Next Steps

Taken together, these differences imply that while the MNL yields interpretable elasticities, the ML models produce responses in which substitution patterns evolve with

policy intensity. As a result, the concept of a constant elasticity becomes less informative for ML-based models. Instead, full counterfactual simulations are required to capture the non-linear behavioural responses implied by flexible predictive architectures.

To further interpret the internal mechanisms driving these responses, post-hoc interpretability tools can provide complementary insight. Partial Dependence Plots (PDPs) summarise the average marginal effect of a given feature on the predicted outcome by integrating out the influence of all other variables, thereby offering a global view of nonlinearity learned by the model.

In addition, Shapley additive explanation (SHAP) plots decompose individual predictions into additive feature contributions, allowing both global importance patterns and local, observation-specific effects to be examined. When analysed jointly with interaction-aware SHAP visualisations, these plots can reveal how combinations of attributes jointly influence mode choice probabilities.

7 Results and Discussion

This section evaluates (i) predictive and simulation performance, and (ii) behavioural implications under counterfactual policy scenarios. Since the primary use case is policy simulation, emphasis is placed on probability quality and the resulting aggregate market shares, rather than hard-label accuracy alone.

7.1 Market Share and Simulation Suitability

Across the ML models tuned on NLL, predicted probabilities aggregate to market shares that closely match observed modal splits, including minority modes such as cycling. This stands in sharp contrast to accuracy-optimised specifications, which can achieve excellent classification accuracy while allocating insufficient probability mass to under-represented alternatives, thereby producing distorted market-share forecasts. The results support the notion that accuracy is not an adequate objective for policy-facing mode-choice modelling when decisions are based on simulated probabilities rather than deterministic class assignment.

7.2 Implied Elasticities: MNL vs. ML

Counterfactual simulations were used to compute implied elasticities with respect to key policy levers (e.g., travel time and monetary cost). Overall, implied elasticities from ML models are of a similar order of magnitude to those obtained from the MNL benchmark (MNL benchmark cross-checked using the direct elasticities for the MNL in Table 4 which were estimated using elasticity theory in Section 3.2), suggesting that ML can recover broadly reasonable sensitivity to time and cost when probabilities are appropriately trained and calibrated. However, a consistent pattern emerges: the MNL generally exhibits larger-magnitude behavioural responses than the ML models.

This difference is plausible given the structural properties of MNL versus flexible ML predictors. In the MNL, utility is linear in attributes and probability changes follow a tightly constrained logit form, yielding smooth and often relatively strong marginal effects that apply globally unless interactions are explicitly added. In contrast, ML elasticities are inherently local: the marginal effect of cost or time can vary across the feature space, and regularisation/ensemble averaging can dampen extreme responses. As a result, ML models may imply more moderate average elasticities even while capturing heterogeneous responses across individuals and contexts.

7.3 Substitution Patterns and Cross-effects

While elasticities are comparable in magnitude, the models can imply different substitution patterns (cross-effects) under policy scenarios. This is expected: substitution in MNL is governed by the logit structure, whereas substitution in ML emerges from learned non-linear decision boundaries and implicit high-order interactions. Consequently, two models can produce similar baseline market shares and similar average elasticities, yet differ meaningfully in where demand reallocates when attributes change.

These differences motivate model introspection beyond aggregate outcomes. In particular, PDPs help reveal the shape and monotonicity of the learned response to key policy variables, while SHAP main effects and interaction plots can diagnose which covariate interactions drive non-intuitive substitution. In practice, this interpretability layer is essential for translating ML simulation outputs into credible behavioural responses.

8 Conclusion

This paper asked whether training ML mode-choice models for probabilistic accuracy via NLL yields plausible elasticities comparable to classical DCMs under multi-scenario simulation. Overall, the answer is a qualified yes: NLL-trained ML models produce simulation-suitable probabilities and implied elasticities that are broadly comparable in sign and order of magnitude to MNL/NL, while still implying different substitution patterns that ask for behavioural validation.

(2) Accuracy vs. NLL for simulation

Accuracy-optimised models can look excellent on classification metrics yet generate miscalibrated probabilities, especially for minority modes. In our results, accuracy-tuned gradient boosting underpredicted cycling (0.4% vs. 2.9% observed). Re-tuning on NLL markedly improved probability calibration and market-share replication across RF, XGB, and NN. For policy simulation, NLL is therefore the appropriate objective.

(3) Behavioural plausibility of implied elasticities

Across counterfactual perturbations, ML-derived implied elasticities display economically plausible patterns in

the aggregate and are of similar magnitude to MNL/NL, though MNL/NL typically imply larger average responses. ML elasticities are inherently local, reflecting non-linearities and implicit interactions; consequently, they should be checked for stability across scenario magnitudes, particularly under larger shocks [2].

(4) Substitution under policy scenarios

Even with comparable baseline shares, ML and DCMs can predict different substitution patterns because substitution is structural in MNL/NL but emergent in ML. These differences motivate model analysis (e.g., PDPs and SHAP interaction diagnostics) to identify the high-order interactions driving scenario responses and to assess behavioural plausibility.

In sum, NLL-trained ML models are suitable for policy simulation and can recover plausible sensitivities without manual interaction specification, offering a useful complement to DCM benchmarks, provided that elasticity stability and substitution patterns are explicitly validated [2, 6].

REFERENCES

[1] U. O. Abdullahi and A. Adnan, *Sustainable Urban mobility: Lessons from European Cities*, Global Journal of Engineering and Technology Advances, 21 (2024), pp. 157–170, https://doi.org/10.30574/gjeta.2024.21.2.0210, https://gjeta.com/content/sustainable-urban-mobility-lessons-european-cities (accessed 2025-11-23). Last Modified: 2024-12-05T09:24+05:30 Publisher: Global Journal of Engineering and Technology Advances.

[2] M. Hassan, M. E. Kabir, S. T. Akter, S. S. Shraban, K. S. Basaruddin, and M. A. Islam, *Machine learning in travel mode choice studies: A systematic literature review of applications, methods, and challenges*, Results in Engineering, 28 (2025), p. 108140, https://doi.org/10.1016/j.rineng.2025.108140, https://www.sciencedirect.com/science/article/pii/S2590123025041866 (accessed 2025-12-23).

[3] T. Hillel, M. Z E B Elshafie, and Y. Jin, *Recreating passenger mode choice-sets for transport simulation: A case study of London, UK*, Proceedings of the Institution of Civil Engineers - Smart Infrastructure and Construction, 171 (2018), pp. 29–42, https://doi.org/10.1680/jsmic.17.00018, https://www.sciencedirect.com/org/science/article/pii/S239787591800008X (accessed 2025-12-30).

[4] A. Pijoan, O. Kamara-Esteban, A. Alonso-Vicario, and C. E. Borges, *Transport Choice Modeling for the Evaluation of New Transport Policies*, Sustainability, 10 (2018), p. 1230, https://doi.org/10.3390/su10041230, https://www.mdpi.com/2071-1050/10/4/1230 (accessed 2025-11-23). Publisher: Multidisciplinary Digital Publishing Institute.

[5] M. WARDMAN, *PUBLIC TRANSPORT VALUES OF TIME*, World Transit Research, (2004), https://www.worldtransitresearch.info/research/1505.

[6] T. Yanar, *Understanding the choice for sustainable modes of transport in commuting trips with a comparative case study*, Case Studies on Transport Policy, 11 (2023), p. 100964, https://doi.org/10.1016/j.cstp.2023.100964, https://www.sciencedirect.com/science/article/pii/S2213624X23000184 (accessed 2025-12-23).

Appendix A. Use of Computational and Generative Tools

Specifically, the following tools were used:

- **GPT-5.2** was used for drafting support, LaTeX formatting assistance, and clarification of technical implementation details.
- **GitHub Copilot**, with **Claude Haiku 4.5** as the underlying model, was used within Visual Studio Code to assist with code autocompletion and syntax checking.

All analytical decisions, model specifications, experimental design choices, and conclusions remain the sole responsibility of the author. No tool was used to generate data, fabricate results, or substitute for substantive analytical judgment.

Appendix B. NN Architecture Search

Table 3: Neural Network Architecture and Activation Search Results (sorted by test loss)

Architecture	**Activation**	**Params**	**Epochs**	**Test NLL**	**Test Loss**
Wide (512–256)	Swish	154,372	30	0.7519	**0.6350**
Small (64–32)	Swish	4,964	30	0.7494	0.6431
Wide (512–256)	ReLU	154,372	10	0.7368	0.6812
Wide (512–256)	Tanh	154,372	10	0.7235	0.7025
Large (256–128–64–32)	Swish	54,756	10	0.7309	0.7216
Large (256–128–64–32)	ReLU	54,756	10	0.7217	0.7330
Small (64–32)	Tanh	4,964	10	0.7225	0.7374
Small (64–32)	ReLU	4,964	10	0.7201	0.7394
Large (256–128–64–32)	Tanh	54,756	10	0.7184	0.7419
Deep (128–128–64–64–32)	Tanh	37,540	10	0.7185	0.7491
Deep (128–128–64–64–32)	Swish	37,540	10	0.7202	0.7502
Deep (128–128–64–64–32)	ReLU	37,540	10	0.7090	0.7606
Medium (128–64–32–16)	Tanh	16,628	10	0.7174	0.7622
Medium (128–64–32–16)	Swish	16,628	10	0.7196	0.7666
Medium (128–64–32–16)	ReLU	16,628	10	0.7133	0.7701

Appendix C. Scenario Elasticities

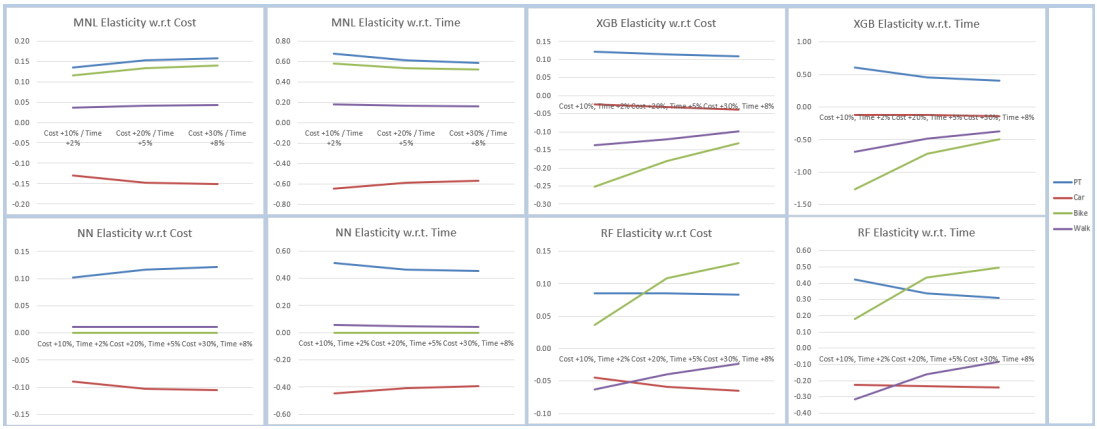

Fig. 6: Elasticities Scenario 1

Fig. 7: Elasticities Scenario 2

Appendix D. Direct Elasticities from the MNL

Table 4: Direct Elasticities from the Multinomial Logit Model

Mode	**Variable**	**Direct Elasticity**
Public Transport	Cost	−0.136
	Vehicle Time	−0.331
	Walking Time	−0.495
	Wait Time	−0.271
	Transfers	0.194
	Age 18–34	0.019
Car	Time	−0.795
	Cost	−0.228
	Traffic Percent	−0.466
	Age 35–54	−0.017
	Female	0.001
	Household 2–5	−0.106
Bike	Time	−1.642
Walk	Time	−5.025

Appendix E. NL Coefficients

Table 5: Estimated NL parameters (robust standard errors)

Parameter	Value	Rob. Std. err	Rob. t-test	Rob. p-value
ASC_BIKE	-2.362386	0.067145	-35.183461	0.000000×10^{0}
ASC_CAR	-2.140398	0.062955	-33.998888	0.000000×10^{0}
ASC_WALK	2.200493	0.071533	30.762023	0.000000×10^{0}
B_AGE_18_34	0.084532	0.025283	3.343457	8.274162×10^{-4}
B_AGE_35_54	-0.051310	0.024586	-2.086938	3.689375×10^{-2}
B_CAR_OWNERSHIP	1.274133	0.015636	81.489146	0.000000×10^{0}
B_COST	-0.154311	0.009056	-17.038846	0.000000×10^{0}
B_COST_X_AGE	0.001066	0.000208	5.132928	2.852683×10^{-7}
B_DAY_OF_WEEK_CAR	0.062863	0.007018	8.957527	0.000000×10^{0}
B_DAY_OF_WEEK_PT	-0.036756	0.006745	-5.449112	5.062206×10^{-8}
B_DISTANCE_CYCLE	-0.000032	0.000010	-3.221931	1.273298×10^{-3}
B_DISTANCE_WALK	-0.000871	0.000042	-20.796224	0.000000×10^{0}
B_DRIVING_LICENSE	1.053266	0.025212	41.776679	0.000000×10^{0}
B_FEMALE	0.043074	0.020559	2.095138	3.615871×10^{-2}
B_N_HOUSEHOLD_2_5	-0.529777	0.021842	-24.254516	0.000000×10^{0}
B_START_TIME_LINEAR_CAR	0.029621	0.002338	12.672228	0.000000×10^{0}
B_TIME	-4.715272	0.120977	-38.976720	0.000000×10^{0}
B_TRAFFIC	-2.447247	0.073255	-33.407146	0.000000×10^{0}
B_TRANSFERS	-0.030017	0.055573	-0.540136	5.891031×10^{-1}
B_VEH_TIME	-2.136830	0.083046	-25.730718	0.000000×10^{0}
B_WAIT_TIME	-3.892656	0.507742	-7.666597	1.776357×10^{-14}
B_WALK_TIME	-4.463429	0.130408	-34.226645	0.000000×10^{0}
mu_non_car	1.032208	0.027294	37.818456	0.000000×10^{0}

Appendix F. MNL Coefficients

Table 6: Estimated MNL parameters (robust standard errors)

Parameter	Value	Rob. Std. err	Rob. t-test	Rob. p-value
ASC_BIKE	-2.417174	0.053152	-45.476402	0.000000×10^{0}
ASC_CAR	-2.113178	0.058021	-36.420724	0.000000×10^{0}
ASC_WALK	2.266432	0.050327	45.034512	0.000000×10^{0}
B_AGE_18_34	0.090786	0.024555	3.697183	2.180052×10^{-4}
B_AGE_35_54	-0.051740	0.024693	-2.095359	3.613905×10^{-2}
B_CAR_OWNERSHIP	1.275336	0.015663	81.424470	0.000000×10^{0}
B_COST	-0.152912	0.008912	-17.157069	0.000000×10^{0}
B_COST_X_AGE	0.001046	0.000207	5.059070	4.213071×10^{-7}
B_DAY_OF_WEEK_CAR	0.062394	0.007093	8.796703	0.000000×10^{0}
B_DAY_OF_WEEK_PT	-0.037712	0.006862	-5.496139	3.881962×10^{-8}
B_DISTANCE_CYCLE	-0.000032	0.000010	-3.178872	1.478492×10^{-3}
B_DISTANCE_WALK	-0.000905	0.000037	-24.521539	0.000000×10^{0}
B_DRIVING_LICENSE	1.056748	0.025030	42.219392	0.000000×10^{0}
B_FEMALE	0.043452	0.020605	2.108781	3.496352×10^{-2}
B_N_HOUSEHOLD_2_5	-0.530412	0.021899	-24.220627	0.000000×10^{0}
B_START_TIME_LINEAR_CAR	0.029687	0.002342	12.674602	0.000000×10^{0}
B_TIME	-4.752854	0.116642	-40.747268	0.000000×10^{0}
B_TRAFFIC	-2.456812	0.073343	-33.497384	0.000000×10^{0}
B_TRANSFERS	-0.035731	0.055993	-0.638127	5.233911×10^{-1}
B_VEH_TIME	-2.123120	0.083059	-25.561733	0.000000×10^{0}
B_WAIT_TIME	-3.875675	0.510962	-7.585061	3.330669×10^{-14}
B_WALK_TIME	-4.459136	0.131941	-33.796393	0.000000×10^{0}