

A COMPARATIVE MATHEMATICAL FRAMEWORK FOR BINARY CLASSIFICATION OF ELECTRIC-VEHICLE RANGE ANXIETY

YaJie Liu*, ZhangJingYuan Wang, LiMin Hao

College of Information Engineering, Tarim University, Alaer 843300, Xinjiang, China.

**Corresponding Author: YaJie Liu*

Abstract: Range anxiety has limited the spread of electric vehicles to some extent due to concerns about travel and charging. This paper uses 50,000 observations to explore how well people with range-anxiety scores above 5 can be identified and how nine predictors related to mobility, charging, cognition, experience and demographics affect the fitted boundary. Missing values are filled with medians; otherwise, the IQR rule is used for capping extreme values, and continuous variables are Z-standardised. Regularised logistic regression, random forest and XGBoost are compared via a stratified 70:30 split and five-fold cross-validation; probability thresholds are searched and 100 Bayesian trials are used if necessary. XGBoost has achieved the highest accuracy (85.47%), precision (88.61%) and F1-score (84.56%) on the independent test set. Random Forest has the largest AUC of 0.9006, and XGBoost is only slightly lower at 0.8995. Electric-vehicle knowledge, environmental awareness and technology affinity account for 97.28% of the XGBoost importance and 90.92% of the random-forest importance. The above results separate the performance at a selected threshold from ranking ability across thresholds and also show the effect of non-linear relationships and cases near the score cutoff.

Keywords: Binary classification; Logistic regression; Random forest; XGBoost; Threshold optimization; Range anxiety

1 INTRODUCTION

Battery capacity is just one cause of range anxiety. In daily driving, people are also concerned about whether they can take an unplanned trip; whether charging will be available; how the battery is doing; and whether they feel confident in using an electric vehicle. Recent reviews have pointed out that range anxiety is caused by psychological, human-factor and socio-technical factors jointly [1]. Charging conditions also affect the perceived practicality and willingness to use an electric vehicle, as well as the availability and visibility of stations [2]. With more exposure to electric vehicles, some uncertainty about driving and charging has been reduced [3]. Therefore, high range anxiety is examined together with travel exposure, charging conditions, cognition, previous experience, and demographic background.

The cut-off for high-anxiety is also unlikely to be predicted accurately. Two people on either side of the selected score may have almost identical predictor profiles. Therefore, in evaluating the model, one should check whether it assigns the correct class at a certain probability threshold and how well it ranks high-anxiety cases across all thresholds. A linear model has an easy-to-interpret coefficient direction but cannot capture interactions or non-linear boundaries. Tree ensembles can capture these patterns more easily, although their parameters are less interpretable. All three models have undergone the same preprocessing, sampling, validation and evaluation; therefore, they can be compared under identical conditions.

Analysis uses 50,000 consumer records. A high-range anxiety score is more than 5, and the predictors are travel exposure, charging convenience, electric-vehicle knowledge, battery concern, environmental awareness, technology affinity, previous electric-vehicle experience, and age. Process missing values before modelling, cap extreme values using the interquartile range rule, and standardize continuous variables. Spearman association, a stratified train-test split, and stratified five-fold cross-validation are used next. Regularised logistic regression serves as the linear benchmark; random forest provides a bagged nonlinear model; XGBoost is a sequential boosting model. Accuracy at various probability thresholds is verified, and AUC is employed to assess ranking across all thresholds. Bayesian optimisation is employed for hyperparameter selection of XGBoost.

Determine which model can identify high-anxiety respondents more accurately, and observe how much the various variables affect the fitted boundary. The same preprocessing and validation process are used for all three kinds of models, and coefficients or feature importance are used to interpret the fitted model. It is possible to compare the ranking performance at all thresholds with classification at one selected threshold and also show the uncertainty near the score cutoff. The framework will be employed for predicting and interpreting range anxiety, but the fitted relationships are not considered causal effects on electric-vehicle behaviour.

2 DATA, LABEL FUNCTION, AND STATISTICAL PREPROCESSING

2.1 Binary Response and Predictor Space

Analysis uses the original 50,000 consumer records and adds no external observations during model development. The binary label is based on whether the recorded range-anxiety score r_i exceeds 5; otherwise, it is placed in the low-anxiety category.

$$y_i = \mathbb{I}(r_i > 5), y_i \in \{0, 1\} \quad (1)$$

The predictors are: daily commuting distance, weekly travel distance, charging-station convenience, electric-vehicle knowledge, concern about battery replacement, environmental awareness, technology affinity, previous electric-vehicle experience and age. Stratified random sampling places 35,000 observations in the training set and 15,000 in the test set.

2.2 Robust Cleaning and Standardization

Missing numerical values are filled in with the median of the training sample. Values outside the interquartile fences are winsorized instead of being deleted; thus, extreme values are reduced, and all observations are retained. Then, each continuous variable is converted into

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \quad (2)$$

The Location and scale terms are only calculated using training data. Standardisation will make it easier to compare the magnitudes of the coefficients, and travel variables with large numbers will not have too much impact on the optimiser.

2.3 Rank Association

Spearman correlation is used because the anxiety score and several predictors need not follow normal distributions. For rank differences d_i , the coefficient is

$$\rho = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2-1)} \quad (3)$$

Daily commuting distance and weekly travel distance have positive coefficients of 0.42 and 0.40 . Electric-vehicle knowledge has a negative coefficient of -0.38 , charging accessibility has -0.31 , and distance to the nearest station has 0.28 . These marginal relations justify retaining both travel exposure and cognitive-access variables in the nonlinear models. The correlation ranking and grouped feature distributions are shown in Figure 1.

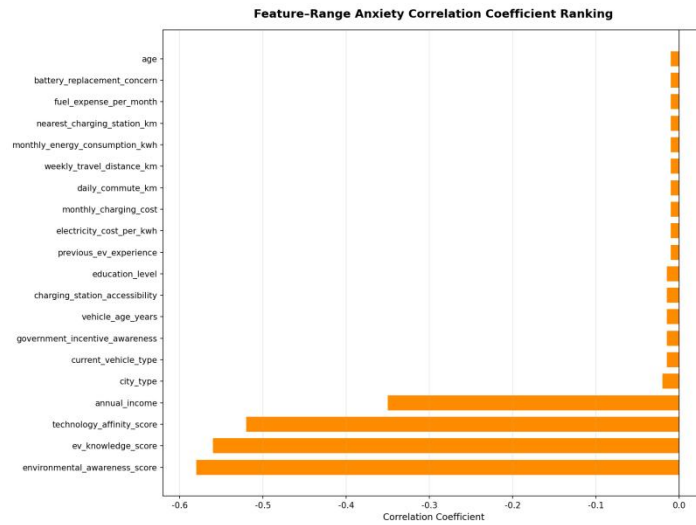


Figure 1 Correlation Ranking and Feature Distributions for Range-Anxiety Classification

3 REGULARIZED LOGISTIC CLASSIFICATION

3.1 Linear Score and Probability Map

Logistic regression defines an interpretable linear reference surface in the standardized nine-dimensional space. For observation i , the score is

$$\eta_i = \beta_0 + \sum_{j=1}^9 \beta_j z_{ij} \quad (4)$$

and the posterior probability of high anxiety is

$$p_i = P(y_i = 1 | z_i) = \frac{1}{1 + e^{-\eta_i}} \quad (5)$$

3.2 Penalized Convex Objective

The coefficients are obtained by minimizing binary cross-entropy with an L2 penalty. The regularized objective is

$$J(\beta) = -\sum_{i=1}^n [y_i \ln p_i + (1 - y_i) \ln(1 - p_i)] + \frac{\lambda}{2} \sum_{j=1}^9 \beta_j^2 \quad (6)$$

The objective is convex and is solved by L-BFGS. Five-fold stratified cross-validation gives a mean accuracy of 83.55% with a standard deviation of 0.0047; training accuracy is 83.60%. A probability threshold is searched from 0.10 to 0.90 in increments of 0.01. The optimal training threshold is 0.53, producing 83.67% training accuracy.

3.3 Coefficient Geometry

Table 1 Estimated Standardized Logistic-Regression Coefficients

Predictor	Coefficient	Predictor	Coefficient
Environmental awareness	-1.3260	Battery concern	0.0021
Technology affinity	-1.0057	Charging convenience	0.0030
EV knowledge	-0.0406	Age	0.0082
Daily commute	-0.0157	Weekly travel	0.0381
Previous EV experience	-0.0053		

The fitted standardized coefficients are summarized in Table 1. The two largest absolute coefficients are negative. Under the high-anxiety coding, greater environmental awareness and technology affinity are associated with lower predicted high-anxiety probability after conditioning on the other inputs. The coefficients describe the fitted classification surface and are not interpreted causally.

4 NONLINEAR ENSEMBLE MODELS

4.1 Random Forest

Random forest draws bootstrap samples and grows decorrelated trees. A bootstrap sample contains approximately 63.2% unique training observations, leaving approximately 36.8% for out-of-bag assessment. At each split, three of the nine predictors are randomly considered. For K trees, majority voting gives

$$\hat{y}(x) = \text{mode}\{h_k(x): k = 1, \dots, K\} \quad (7)$$

The ensemble error is related to average tree correlation and mean tree strength through

$$PE^* \leq \frac{\bar{\rho}(1-s^2)}{s^2} \quad (8)$$

where the bound decreases as strength increases and increases with tree correlation. The fitted forest uses 300 trees, maximum depth 12, minimum split size 5, and minimum leaf size 2. Its cross-validation accuracy is 86.10% with standard deviation 0.0050. Training accuracy is 86.56%, and a threshold of 0.27 maximizes training accuracy at 91.02%.

4.2 Regularized XGBoost

XGBoost constructs an additive score from regression trees. At iteration t , the model minimizes empirical loss and tree complexity:

$$L^{(t)} = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \quad (9)$$

$$\Omega(f) = \gamma T + \frac{\lambda}{2} \sum_{j=1}^T w_j^2 + \alpha \sum_{j=1}^T |w_j| \quad (10)$$

Optuna performs 100 tree-structured Parzen-estimator trials with five-fold stratified accuracy as the target. The optimum uses 350 trees, depth 3, learning rate 0.0222, row subsampling 0.850, feature subsampling 0.778, L1 coefficient 0.902, L2 coefficient 2.316, minimum child weight 3, and split penalty 0.260. The optimal cross-validation accuracy is 86.10%; the retained probability threshold is 0.50 and training accuracy is 86.43%.

5 COMPARATIVE CLASSIFICATION RESULTS

5.1 Test Metrics

All three fitted models are evaluated on the same 15,000-observation test set. Accuracy and F1 are threshold dependent, whereas AUC measures ranking discrimination over all thresholds. Both kinds of measures will not be used to assess the model's quality. Table 2 shows the results of the independent test set.

Table 2 Results of the Three Classifiers on the Independent Test Set

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	AUC
Logistic regression	82.88	87.17	76.47	81.47	0.8674
Random forest	83.45	83.01	83.44	83.22	0.9006
XGBoost	85.47	88.61	80.86	84.56	0.8995

XGBoost has the highest accuracy, precision and F1-score on the test set. Random forests have good recall and AUC; however, their AUC is only 0.0011 higher than that of XGBoost. Logistic regression still has relatively high precision, but its recall is lower, which is consistent with the difficulty of representing interacting effects with one linear boundary. The prediction results of the three models are shown in Figure 2.

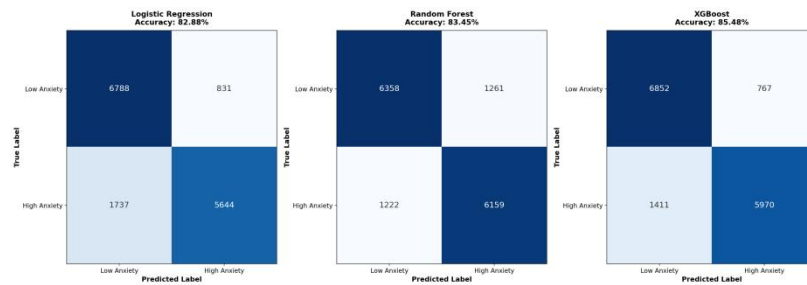


Figure 2 Confusion Matrices for the Three Models on the Test Set

5.2 Class-Specific XGBoost Performance

Table 3 XGBoost Results for Each Class on the Test Set

Class	Precision	Recall	F1-score
Low anxiety	0.83	0.90	0.86
High anxiety	0.89	0.81	0.85

Table 3 shows a different balance for the high-anxiety class: relatively high precision, but low recall. That is to say, some of the observations that actually belong to the high-anxiety class are predicted as low anxiety. Their predictor profiles overlap with those of the other class near the score cutoff of 5, and self-reported anxiety labels may introduce some additional noise.

6 FEATURE CONTRIBUTION AND MATHEMATICAL INTERPRETATION

6.1 Ensemble Importance

Most of the predictive signal in both ensembles is carried by three cognitive variables, although the models divide that signal differently. Random forest accumulates impurity reductions across bootstrapped trees; XGBoost records gains while successively correcting the current loss. Table 4 lists the importance values obtained from the fitted models.

Table 4 Importance Values from the Fitted Random-Forest and XGBoost Models

Predictor	Random forest	XGBoost
EV knowledge	0.2815	0.4061
Environmental awareness	0.3005	0.2873
Technology affinity	0.3272	0.2794
Previous EV experience	0.0020	0.0058
Daily commute	0.0193	0.0050
Weekly travel	0.0205	0.0047
Age	0.0151	0.0044
Charging convenience	0.0171	0.0043
Battery concern	0.0169	0.0031

Together, the three leading variables make up 90.92% of random-forest importance and 97.28% of XGBoost importance. Travel exposure, charging convenience, age, battery concern, and earlier electric-vehicle experience remain available to the models, but add relatively little split gain. Figure 3 places the two importance profiles side by side.

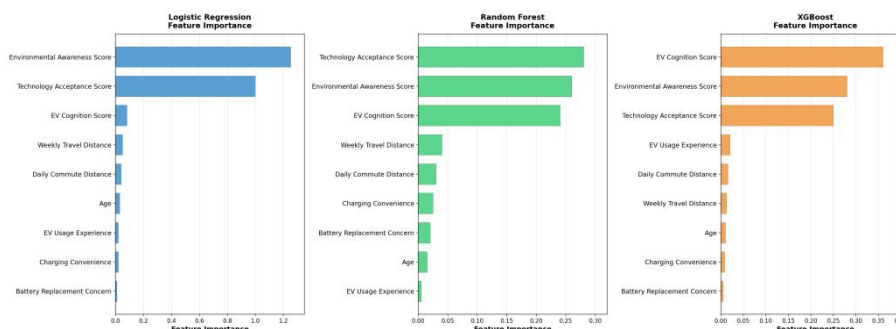


Figure 3 Feature-Importance Profiles for the Fitted Models

6.2 Additive Local Attribution

SHAP expresses the fitted XGBoost score as a sum of local feature contributions. For observation i , this decomposition is

$$g(x_i) = \phi_0 + \sum_{j=1}^9 \phi_{ij} \quad (11)$$

The baseline is the model's expected output, and each added term records the contribution of one feature for that observation. The direction of the SHAP results is consistent with the large negative logistic coefficients: higher environmental awareness and technology affinity tend to shift the fitted score away from high anxiety. These are predictive relationships and should not be read as causal effects.

7 MATHEMATICAL DISCUSSION

7.1 Linear and Nonlinear Decision Boundaries

Logistic regression employs a standardised input and divides the two classes with an affine surface. The penalized convex objective has a unique optimum, and although it is simple to determine whether the coefficients are positive or negative, they cannot express all conditional interactions. Random forests average decorrelated, axis-aligned trees built from bootstrap samples to reduce variance. XGBoost gradually adds shallow trees to learn patterns that have been missed by previous trees. Shrinkage, row and feature subsampling, L1 and L2 penalties, and split penalties are used to limit model complexity.

The results are in line with the above differences among the models. Logistic regression has the lowest accuracy and AUC; the two tree models can capture interactions between cognition, travel and charging conditions to enhance recall or F1. XGBoost performs best after setting a decision threshold; however, random forest is slightly better at probability ranking across thresholds. Gradient boosting has been used previously to address range anxiety based on charging behaviour, as shown in [4].

7.2 Domain Interpretation and Error Sources

Earlier studies have shown that drivers trade off the range of their vehicles and charging facilities [5], and access to charging stations also affects purchase intentions [6]. Research on range anxiety has identified problems with infrastructure, behaviour and psychology [7]. Recently, some studies have begun to use actual data to quantify range anxiety and its causes via mathematics and machine learning [8]. The above predictors are supported by these studies; however, the coefficients and importance values reported in this paper apply only to the current sample.

Most classification errors occur near the cutoff used to divide the two classes. A respondent with a score of exactly 5 may be very similar to one with a score just above 5, but they receive different labels. Dividing a continuous score into two classes also loses differences among the classes. Therefore, part of the overlap is due to label definition rather than failed optimisation. Adjust the probability threshold to change the trade-off between precision and recall; however, information lost in converting the original score to two classes cannot be recovered.

7.3 Limitations

Fitted relationships cannot determine causality because the data are cross-sectional. Test performance is also affected by the cutoff of 5 and the class distribution in this sample. Importance values are produced by specific models; therefore, they should not be considered fixed structural parameters. Studies of charging decisions and range-related anxiety have also focused on interpretable prediction rather than causal identification [9-10]. The fitted boundary needs to be re-examined when there are changes in the study period, area or vehicle population.

8 CONCLUSION

This study is a binary-classification process that states the label rule, preprocesses data, uses stratified validation, includes regularisation in the model objective, checks probability thresholds, and interprets the fitted models. XGBoost has a high accuracy and F1-score; Random Forest has a high AUC. Three cognitive variables account for most of the importance in both tree models, and many remaining errors are due to overlap near the anxiety-score cutoff. The assumptions, model settings, comparisons and main sources of uncertainty are listed below for reference in the analysis. In the future, we will add charging conditions that vary over time and more information on distance-related anxiety, and continue to report ranking performance separately from classification at a selected threshold.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

- [1] Rainieri G, Buizza C, Ghilardi A. The psychological, human factors and socio-technical contribution: A systematic review towards range anxiety of battery electric vehicles' drivers. *Transportation Research Part F: Traffic Psychology and Behaviour*, 2023, 99: 52-70.
- [2] White L V, Carrel A L, Shi W, et al. Why are charging stations associated with electric vehicle adoption? Untangling effects in three United States metropolitan areas. *Energy Research & Social Science*, 2022, 89: 102663.
- [3] Jaiswal D, Kant R, Singh P K, et al. Investigating the role of electric vehicle knowledge in consumer adoption: Evidence from an emerging market. *Benchmarking: An International Journal*, 2021, 29(3): 1027-1045.
- [4] Song Y, Hu X. Learning electric vehicle driver range anxiety with an initial state of charge-oriented gradient boosting approach. *Journal of Intelligent Transportation Systems*, 2023, 27(2): 238-256.
- [5] Rosales-Tristancho A, Carazo A F, Brey R, et al. Meta-analysis of consumers' trade-offs between range and charging infrastructure for electric vehicles. *Transportation Research Part D: Transport and Environment*, 2024, 134: 104342.
- [6] Kristoffersson I, Pyddoke R, Kristoffersson F, et al. Access to charging infrastructure and the propensity to buy an electric car. *Transportation Research Part D: Transport and Environment*, 2025, 139: 104588.
- [7] Pamidimukkala A, Kermanshachi S, Rosenberger J M, et al. Barriers and motivators to the adoption of electric vehicles: A global review. *Green Energy and Intelligent Transportation*, 2024, 3(2): 100153.
- [8] Ma X, Xie R, Guo L, et al. Range anxiety of battery electric vehicles: Quantification and determinants using real-world data. *Transportation Research Part D: Transport and Environment*, 2025, 146: 104837.
- [9] Bas J, Cirillo C, Cherchi E. Classification of potential electric vehicle purchasers: A machine learning approach. *Technological Forecasting and Social Change*, 2021, 168: 120759.
- [10] Ovchinnikov A, Forestell D. From range anxiety to charge anxiety: Operations scholars' reflections on the state of electric vehicles' public charging infrastructure. *Service Science*, 2026, 18(1): 55-78.