JSCE (Journal of System and Computer Engineering)
Vol.7, No.4, October 2026, pp. 439-449
ISSN (online): 2723-1240
DOI : https://doi.org/10.61628/jsce.v7i4.2860

Research Article Open Access (CC–BY-SA) ■

Beyond Accuracy: Demographic Feature Sufficiency in Imbalanced Transmission Preference Prediction

Jeffry [1,a,*]
[1] *Robotics and Artificial Intelligence Engineering, Institut Teknologi Bacharuddin Jusuf Habibie, Jl. Balai Kota No. 1, Parepare 91122, South Sulawesi, Indonesia*
[a] *jeffry@ith.ac.id*
** Corresponding author*

Abstract

Predicting customer preference from limited demographics can appear successful when target classes are highly skewed. This study evaluates whether normalized age and gender indicators are sufficient to predict vehicle transmission preference and whether conventional accuracy masks minority-class failure. A customer dataset of 223,042 records contains MANUAL 178,943, AUTOMATIC 35,810, CVT 8,283, and E-CVT 6. Because E-CVT is too sparse for reliable stratified validation, it is retained for descriptive analysis but excluded from the three-class predictive comparison. We combine association tests, mutual information, an empirical pattern-oracle analysis, and five-fold stratified cross-validation of a majority baseline, multinomial logistic regression, cost-sensitive logistic regression, and class-weighted random forest. Standard logistic regression reproduces the majority baseline, with accuracy 0.8023 but balanced accuracy 0.3333 and MCC 0.0000. Class-weighted random forest lowers accuracy to 0.4946 but improves balanced accuracy to 0.4535, Macro-F1 to 0.3399, and MCC to 0.0969. Age and gender jointly provide only 0.0300 bits of information, approximately 3.5 percent of target entropy. The findings show that high accuracy primarily reflects class prevalence; richer behavioral, product, and transaction features are required for practically useful preference prediction.

Keywords—*customer behavior; imbalanced classification; feature sufficiency; logistic regression; random forest*

1. Introduction

Customer behavior prediction supports product planning, customer relationship management, recommendation, and targeted marketing by translating historical customer records into actionable estimates of future choices. Recent work shows that predictive performance improves when models incorporate behavioral traces, transaction history, browsing patterns, and interaction variables rather than relying only on static demographics (Chen et al., 2021; Kim et al., 2024; Liu et al., 2024). Consumer-behavior studies have also demonstrated that class imbalance can materially distort model learning and that oversampling, cost-sensitive learning, and ensemble methods can improve sensitivity to rare outcomes (Li et al., 2022; Tanvir et al., 2023). At the same time, explainable artificial intelligence is increasingly used to clarify the contribution of customer and transaction variables to predictions, strengthening the interpretability of data-driven business decisions (Predić et al., 2023).

A central difficulty in customer preference classification is that observed outcomes are rarely distributed evenly. Standard learners tend to favor prevalent classes, particularly when minority classes are both relatively and absolutely rare. Recent imbalanced-learning research distinguishes data-level, algorithm-level, cost-sensitive, and ensemble strategies for mitigating

this bias (Chen et al., 2024; Vieira & Rodrigues, 2024). Weighting methods increase the penalty assigned to minority-class errors without duplicating observations (He & Cheng, 2021), while ensemble-based strategies can combine diversity with rebalancing to improve minority recognition (Abedin et al., 2023; Zhao et al., 2021). For multiclass problems, however, imbalance interacts with class overlap and small sample size, so a favorable overall accuracy may coexist with poor performance on specific classes (Naglik & Lango, 2024).

The choice of evaluation metric is therefore not a cosmetic reporting decision. Accuracy is prevalence-sensitive and can appear high when a classifier repeatedly predicts the majority class. Balanced accuracy, class-level recall, Macro-F1, and the Matthews correlation coefficient provide complementary perspectives on discrimination under skewed class distributions (Chicco et al., 2021; Hicks et al., 2022). Recent methodological analyses further caution that aggregate metrics should be selected according to the evaluation objective and interpreted together with the confusion matrix rather than treated as interchangeable measures (Chicco & Jurman, 2023; Opitz, 2024). Cost-sensitive learning is especially relevant when minority errors are consequential, although the value of weighting depends on the information available in the predictors (Araf et al., 2024).

This study extends two previous investigations of automotive customer behavior. Jeffry et al. (2023) analyzed age, gender, and preferred vehicle transmission using ensemble logistic regression and reported accuracy improvements with boosting and bagging. Jeffry et al. (2025) subsequently enriched the feature space with vehicle category, product type, vehicle year, dealer branch, and transaction source, combined class balancing with Bayesian optimization, and used SHAP to support explainability. The methodological progression suggests an unresolved question: how much predictive information is actually present in the original demographic feature set, and to what extent can a high accuracy score be explained simply by the dominance of one transmission class?

Accordingly, the present research reframes the task as an audit of demographic feature sufficiency under severe multiclass imbalance. It makes three contributions. First, it quantifies the degree of class imbalance and the strength of the associations between normalized age, encoded gender, and transmission preference. Second, it measures feature sufficiency through mutual information and an empirical pattern-oracle analysis before fitting predictive models. Third, it compares standard and cost-sensitive classifiers under five-fold stratified cross-validation using accuracy, balanced accuracy, Macro-F1, MCC, and class-specific recall. This design separates apparent performance caused by class prevalence from genuine discrimination attributable to the available predictors.

2. Method

2.1 Dataset and Data Audit

The analysis uses a secondary normalized customer dataset from an automotive-customer research line. The related source studies describe customer records obtained from an automotive company/dealership in Makassar, Indonesia; the expanded dataset documented by Jeffry et al. (2025) covers 2008-2023. The normalized export analyzed in the present study contains 223,042 records but does not retain transaction dates, customer identifiers, or raw-age values. Consequently, the exact temporal span of this specific export cannot be independently reconstructed and is not assumed to be identical to the expanded 2025 dataset. No probabilistic sampling was applied to the supplied export: all available complete rows were audited, and all eligible rows were used for modeling after the E-CVT handling rule described below.

The file contains four analytical variables: normalized age (usia), two mutually exclusive binary gender indicators (JKL and JKP), and transmission preference as the target. JKL was interpreted as Male and JKP as Female; internal consistency was verified because JKL + JKP = 1 for every record. No missing values were detected. Age was supplied on a 0-1 normalized scale with 105 distinct values. Because the export does not preserve raw ages, the original

minimum and maximum values, or the preprocessing script, the normalization transform cannot be reconstructed exactly; therefore, age is analyzed only on the supplied normalized scale. Transmission preference is defined as the recorded transmission category associated with the customer's vehicle choice and comprises MANUAL, AUTOMATIC, CVT, and E-CVT. E-CVT contains only six records and is retained for descriptive analysis but excluded from the main three-class predictive comparison, leaving 223,036 modeling records.

Table 1. Dataset and preprocessing specifications.

Item	Spesification
Data source	Secondary customer records from an automotive company/dealership research dataset in Makassar, Indonesia; provenance follows the Jeffry et al. (2023, 2025) research line.
Data period	Transaction dates are not retained in the normalized export. The related expanded dataset is documented as 2008-2023; exact temporal equivalence of this 223,042-row export cannot be verified from the supplied file.
Sampling mechanism	No probabilistic sampling of the supplied export. All available complete records were audited; eligible rows were used in cross-validation.
Initial record count	223,042 customer records.
Variables	Normalized age; JKL (Male); JKP (Female); transmission preference target.
Missing-data cleaning	No missing values were detected in the four analytical variables; no row was removed for missingness.
Gender encoding	JKL=1, JKP=0 denotes Male; JKL=0, JKP=1 denotes Female. Mutual exclusivity was verified for every row.
Age preprocessing	Age was already provided on a 0-1 normalized scale. Raw ages and the original scaling parameters were unavailable, so no back-transformation was performed.
Target definition	Recorded vehicle transmission preference: MANUAL, AUTOMATIC, CVT, or E-CVT.
E-CVT handling	Six E-CVT records retained for descriptive statistics and excluded from predictive cross-validation because stable class-specific estimates are not feasible.
Final modeling sample	223,036 records: MANUAL, AUTOMATIC, and CVT.
Software	Python 3.13.5; pandas 2.2.3; NumPy 2.3.5; SciPy 1.17.0; scikit-learn 1.8.0; Matplotlib 3.10.8. Random seed=42 where applicable.

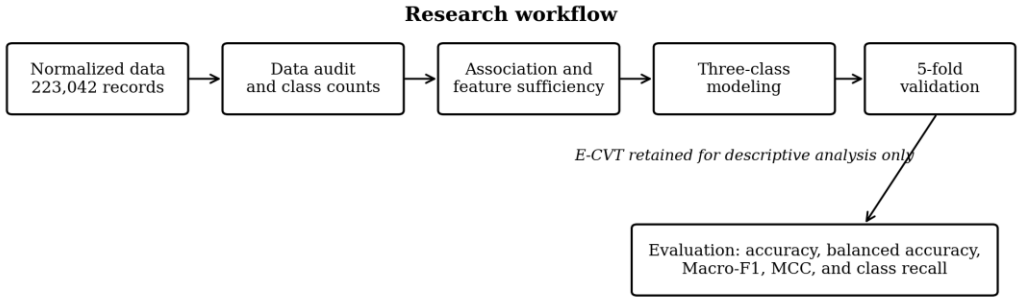


Figure 1. Research workflow used to separate class-prevalence effects from demographic predictive information**.**

2.2 Statistical Association and Feature Sufficiency Analysis

Gender and transmission preference were evaluated with the chi-square test of independence, while Cramer's V was used to quantify association strength. Differences in normalized age across the four transmission groups were evaluated with the Kruskal-Wallis test because the objective is to compare distributions without relying on normality assumptions. Epsilon squared

was calculated as an effect-size estimate so that statistical significance could be distinguished from practical magnitude in this very large dataset.

Feature sufficiency was assessed using mutual information between the joint demographic feature pattern X and target Y. The dataset contains 105 distinct normalized age values and two gender states, producing 206 observed demographic patterns. Target entropy and conditional entropy were calculated empirically from the observed frequencies, and mutual information was defined as follows:

$$I(X;Y) = H(Y) - H(Y|X) \quad (1)$$

The ratio I(X;Y) divided by H(Y) is reported as the proportion of target uncertainty removed by the available demographic features. An additional empirical pattern-oracle diagnostic assigns each observed age-gender pattern to its most frequent target class and measures the resulting in-sample accuracy. This is not a formal generalization bound; rather, it provides an optimistic diagnostic of how much deterministic separation is visible in the current feature patterns before model complexity is introduced.

2.3 Classification Models

Four models were evaluated. The majority baseline always predicts the most prevalent class and establishes the accuracy obtainable without learning a relationship between predictors and target. Multinomial logistic regression represents the unweighted linear baseline and maintains continuity with the earlier ensemble-logistic-regression study (Jeffry et al., 2023). Cost-sensitive logistic regression uses inverse-frequency class weights so that errors on minority classes receive larger penalties. Class-weighted random forest provides a nonlinear comparison while retaining the same class-weighting principle. Weighting is a standard algorithm-level approach for imbalanced classification and can improve minority sensitivity without modifying the original observations (He & Cheng, 2021; Araf et al., 2024). The class weight for class c follows:

$$w_c = N / (K N_c) \quad (2)$$

where N is the training-set size, K is the number of classes, and N_c is the number of training observations in class c. Logistic regression used the lbfgs solver with a maximum of 1,000 iterations. Random forest used 50 trees, class_weight set to balanced, parallel processing, and a fixed random seed of 42. Hyperparameter optimization was intentionally not performed because the research question concerns the information sufficiency of the feature set rather than maximizing model complexity.

2.4 Validation and Evaluation Metrics

All three-class models were assessed with five-fold stratified cross-validation using a fixed random seed of 42. Stratification preserves class proportions within each fold and avoids a single holdout split dominating the conclusions. Mean and sample standard deviation across folds are reported. Accuracy is retained for comparability with previous studies, but model interpretation emphasizes balanced accuracy, Macro-F1, MCC, the confusion matrix, and class-specific precision and recall. This multi-metric strategy follows recent guidance that no single aggregate statistic fully describes classification quality, particularly under imbalance (Hicks et al., 2022; Opitz, 2024).

$$\text{Balanced Accuracy} = (1/K) \Sigma \text{Recall}_c \quad (3)$$

$$\text{Macro-F1} = (1/K) \Sigma F1_c \quad (4)$$

Balanced accuracy gives equal weight to each class through mean recall, while Macro-F1 averages class-level F1 scores. MCC summarizes agreement between true and predicted labels and is useful as an additional chance-centered diagnostic; a value near zero indicates little relationship between predictions and ground truth even when raw accuracy is high (Chicco et al., 2021; Chicco & Jurman, 2023).

3. Results And Discussion

3.1 Class Distribution and Demographic Profile

The target distribution is severely imbalanced. MANUAL accounts for 178,943 records, or 80.228 percent, while AUTOMATIC accounts for 35,810 records, CVT for 8,283, and E-CVT for only 6. The imbalance ratio between MANUAL and E-CVT is approximately 29,824 to 1. Figure 2 displays the counts on a logarithmic scale so that the extremely small E-CVT category remains visible. The gender indicators are less imbalanced: Male occurs in 128,455 records and Female in 94,587. Thus, the dominant source of skew is the response variable rather than the demographic coding.

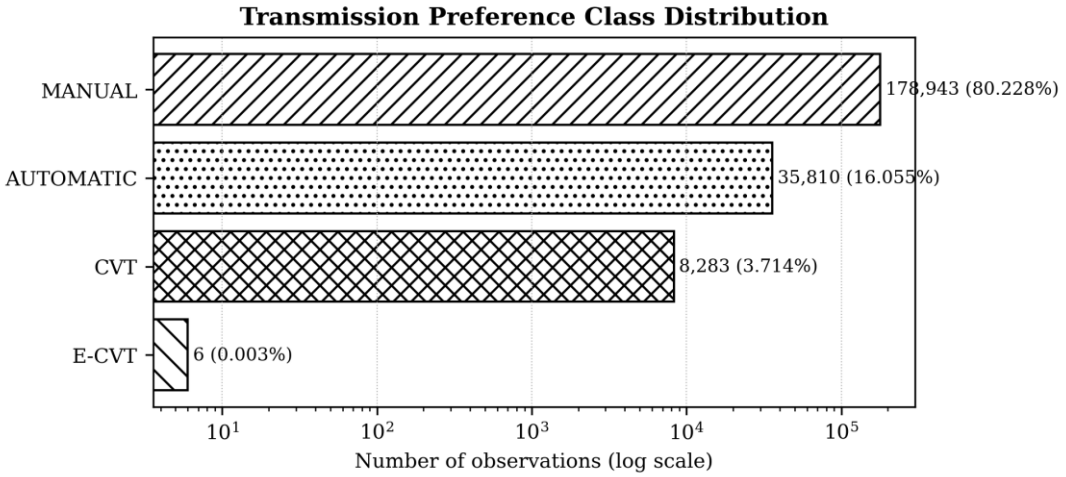


Figure 2. Transmission preference distribution. The logarithmic axis reveals the extreme rarity of E-CVT.

Table 2. Dataset composition by transmission preference.

Class	Count	Share (%)
MANUAL	178,943	80.228
AUTOMATIC	35,810	16.055
CVT	8,283	3.714
E-CVT	6	0.003

3.2 Association Strength and Demographic Information

Gender and transmission preference are statistically associated, with chi-square equal to 3,195.716, three degrees of freedom, and p below 0.001. However, Cramer's V is only 0.1197, indicating that the association is weak in practical terms. This distinction is important because a sample exceeding two hundred thousand observations can make small associations highly significant. The result therefore supports the presence of a population-level relationship without implying strong individual-level predictive power.

Normalized age also differs statistically among transmission classes. The Kruskal-Wallis statistic is 1,276.246 with p below 0.001, but epsilon squared is only 0.0057. The class means are narrowly clustered around 0.40, while the boxplots in Figure 3 show substantial overlap among MANUAL, AUTOMATIC, and CVT. E-CVT appears more dispersed, but its six observations are too few for substantive distributional interpretation. The effect-size result again indicates that statistical significance should not be conflated with strong discrimination.

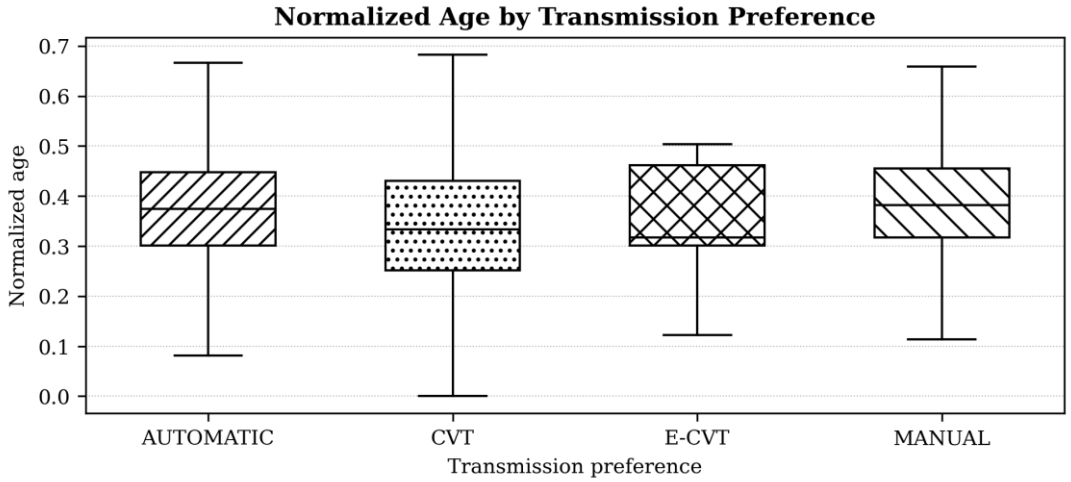


Figure 3. Distribution of normalized age across transmission preference classes; E-CVT contains only six observations**.**

Table 3. Age distribution and feature-sufficiency diagnostics**.**

Measure	Result
Mean normalized age: AUTOMATIC	0.3970
Mean normalized age: CVT	0.4016
Mean normalized age: E-CVT	0.4268
Mean normalized age: MANUAL	0.3985
Cramer's V: gender vs. target	0.1197
Kruskal-Wallis epsilon squared	0.0057
Target entropy H(Y)	0.8555 bits
Conditional entropy H(Y|X)	0.8255 bits
Mutual information I(X;Y)	0.0300 bits
Relative entropy reduction	3.50%
Observed demographic patterns	206
Empirical pattern-oracle accuracy	0.8023

The information analysis provides the clearest evidence of limited feature sufficiency. Target entropy is 0.8555 bits, and conditioning on the observed age-gender pattern reduces this value to 0.8255 bits. Mutual information is therefore only 0.0300 bits, equivalent to approximately 3.50 percent of target entropy. Among the 206 observed demographic patterns, MANUAL is the modal target for 201 patterns and AUTOMATIC is modal for only five; CVT and E-CVT are never the modal class for any pattern. Even the optimistic empirical pattern-oracle achieves only 0.8023 accuracy, effectively matching the raw share of MANUAL. This indicates that the feature space itself offers little deterministic separation between transmission preferences.

3.3 Cross-Validated Predictive Performance

The five-fold results in Table 3 expose the accuracy paradox. Both the majority baseline and standard multinomial logistic regression achieve mean accuracy of 0.8023, yet balanced accuracy remains 0.3333, Macro-F1 is 0.2968, and MCC is exactly zero. Their out-of-fold predictions contain only MANUAL. In other words, standard logistic regression does not improve on a classifier that ignores age and gender entirely. The high accuracy is therefore a direct consequence of target prevalence rather than successful discrimination.

Table 4. Five-fold stratified cross-validation results on the three-class dataset.

Model	Accuracy	Balanced Acc.	Macro-F1	MCC
Majority baseline	0.8023 ± 0.0000	0.3333 ± 0.0000	0.2968 ± 0.0000	0.0000 ± 0.0000
Multinomial LR	0.8023 ± 0.0000	0.3333 ± 0.0000	0.2968 ± 0.0000	0.0000 ± 0.0000
Cost-sensitive LR	0.5074 ± 0.0011	0.4021 ± 0.0020	0.2669 ± 0.0005	0.0737 ± 0.0032
Class-weighted RF	0.4946 ± 0.0035	0.4535 ± 0.0045	0.3399 ± 0.0023	0.0969 ± 0.0045

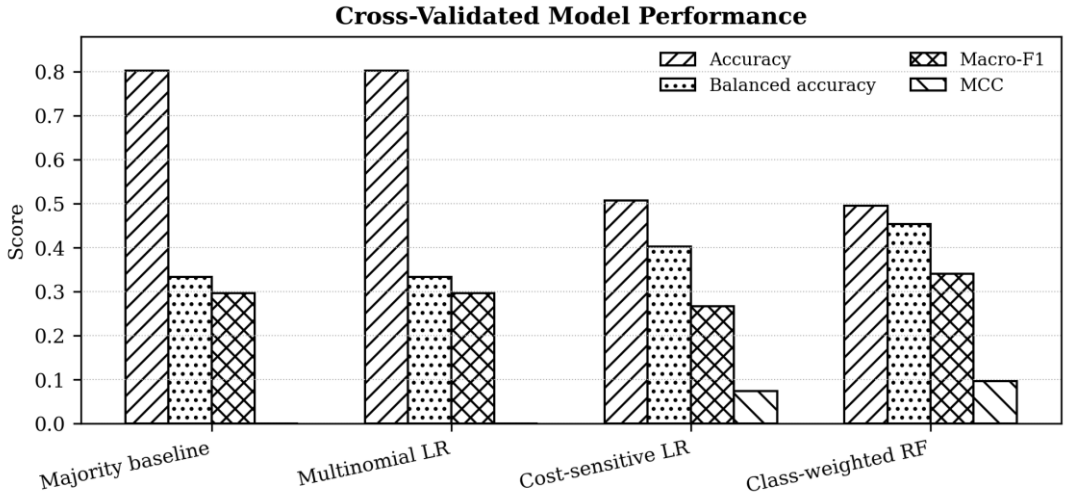


Figure 4. Cross-validated model performance

Cost-sensitive logistic regression changes the decision boundary substantially. Balanced accuracy increases to 0.4021 and MCC becomes positive at 0.0737, but the model fails to produce AUTOMATIC predictions in the aggregated out-of-fold results. Instead, it trades majority-class accuracy for CVT recall. This behavior illustrates a limitation of weighting when the feature space lacks sufficient separation: increasing the penalty on rare-class errors can redistribute predictions without creating information that is absent from the predictors.

Class-weighted random forest provides the strongest imbalance-aware result. Its mean accuracy is lower at 0.4946, but balanced accuracy rises to 0.4535, Macro-F1 to 0.3399, and MCC to 0.0969. The out-of-fold confusion matrix in Figure 5 shows that the model detects 54.6 percent of CVT and 28.0 percent of AUTOMATIC observations, while MANUAL recall declines to 53.5 percent. CVT precision is only 8.0 percent, confirming that minority sensitivity is obtained at the cost of many false positives. The small MCC further shows that the overall relationship between predictions and true labels remains weak.

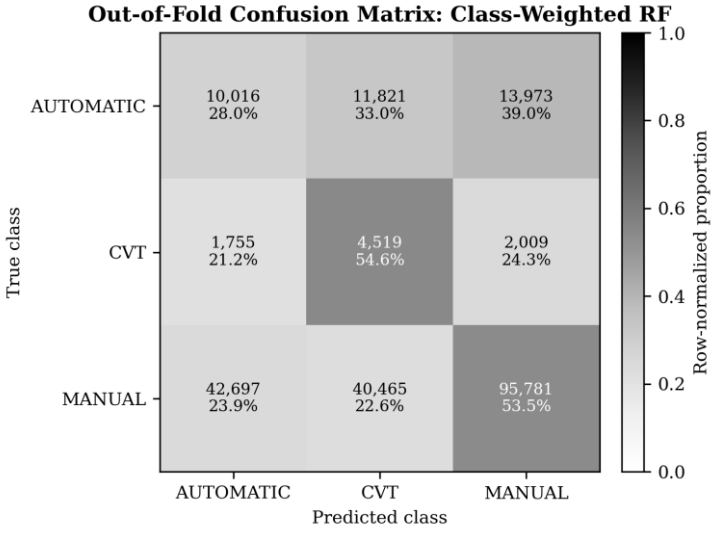


Figure 5. Row-normalized out-of-fold confusion matrix for the class-weighted random forest**.**

Table 5. Out-of-fold class-level performance of the class-weighted random forest.

Class	Precision	Recall	F1-score	Support
AUTOMATIC	0.1839	0.2797	0.2219	35,810
CVT	0.0796	0.5456	0.1389	8,283
MANUAL	0.8570	0.5353	0.6590	178,943

3.4 Principal Findings

The principal result is that the apparent predictive success of demographic-only modeling is largely an artifact of target prevalence. The majority baseline and standard multinomial logistic regression both achieve 0.8023 accuracy, yet both have balanced accuracy of 0.3333 and MCC of 0.0000 because their out-of-fold predictions contain only MANUAL. In contrast, class weighting increases minority-class detection, but the best imbalance-aware model, class-weighted random forest, still reaches only 0.4535 balanced accuracy, 0.3399 Macro-F1, and 0.0969 MCC. The information-theoretic analysis reaches the same conclusion from a different direction: age and gender jointly reduce target uncertainty by only 0.0300 bits, or approximately 3.5% of the target entropy. Thus, the major finding is not that one classifier solves the problem, but that the available demographic feature space contains insufficient information for reliable individual-level transmission-preference prediction.

3.5 Explanation of the Findings

The results can be explained by the interaction of severe class imbalance and weak feature separation. MANUAL represents more than four-fifths of the sample, so a classifier can obtain high raw accuracy by reproducing the dominant prevalence pattern. At the same time, gender has only a weak association with transmission preference (Cramer's V=0.1197), and normalized age shows a statistically significant but very small effect size (epsilon squared=0.0057). Class weighting changes the loss function and therefore shifts the decision boundary toward minority classes, but it cannot create discriminatory information that is absent from the predictors. This explains why minority recall rises while precision remains low, particularly for CVT. The result supports interpreting class weighting as an error-cost adjustment rather than as a substitute for informative variables.

3.6 Comparison with Prior Work

The findings refine, rather than contradict, the earlier automotive studies. Jeffry et al. (2023) reported that boosting and bagging improved logistic-regression accuracy to 76% and 75%, respectively, using demographic variables and transmission preference. Direct numerical ranking against the present results is inappropriate because the data preparation, class structure, and validation procedures differ. The present analysis instead demonstrates why accuracy should be accompanied by prevalence-aware metrics: an 80.23% accuracy can be achieved here without identifying any minority-class observation. The subsequent study by Jeffry et al. (2025) expanded the feature space with vehicle category, product type, vehicle year, dealer branch, and transaction source, while adding balancing, Bayesian optimization, and SHAP. The current feature-sufficiency analysis provides a quantitative rationale for that evolution because demographic information alone removes only about 3.5% of target uncertainty. This is also consistent with recent customer analytics research showing that behavioral, transactional, and interaction features contribute information that static demographics cannot capture

3.7 Practical Implications and Recommended Advanced Features

For operational modeling, demographic-only targeting should not be used as the primary basis for customer-level transmission recommendations. The next model should be designed around feature enrichment before additional hyperparameter tuning. Table 6 provides a concrete data-collection framework. The recommended variables are selected because they represent product fit, purchasing capacity, interaction context, and behavioral intent, all of which are closer to the mechanism that produces a vehicle-choice decision than age or gender alone. Where privacy and governance requirements permit, these features should be captured at transaction or interaction level with consistent timestamps so that temporally separated validation can be performed. Model selection should then be based on the business cost of false positives and false negatives, with class-level calibration and monitoring rather than accuracy alone.

3.8 Limitations

Several limitations define the scope of the conclusions. First, the analyzed file is a normalized secondary export and does not retain transaction dates, raw age, or the original normalization parameters; therefore, the exact temporal span and age transformation cannot be reconstructed from the supplied file alone. Second, E-CVT contains only six observations and cannot support stable out-of-sample estimation. Third, the analysis intentionally restricts predictors to normalized age and gender indicators; poor performance should therefore be interpreted as evidence about this feature set, not as evidence that transmission preference is intrinsically unpredictable. Fourth, the empirical pattern-oracle is descriptive and optimistic because it is computed on the observed patterns. Fifth, the current study lacks an external or temporally separated validation set. Future work should re-link the modeling table to the source transaction database, preserve the original preprocessing metadata, collect the advanced features summarized in Table 6, and validate the resulting model on a later time period or an independent dealership sample

Table 6. Recommended advanced feature framework for the next predictive model.

Feature domain	Recommended variables	Expected analytical value
Vehicle/product attributes	Model, body type, category, engine/fuel type, transmission availability, model year, trim	Captures product compatibility and the choice set actually available to the customer.
Purchasing history	Prior vehicle purchases, repeat-customer status, recency, frequency, monetary value, trade-in history	Represents revealed preference, loyalty, and repeat-purchase behavior.
Price and financing	Listed price, discount, down payment, installment amount, tenor, financing type, approved credit range	Measures affordability constraints and price sensitivity that may shape transmission choice.
Dealer interaction	Branch, salesperson, test-drive activity, quotation count, follow-up count, visit frequency	Captures local inventory, sales influence, and strength of purchase intent.
Transaction source	Walk-in, referral, website, social media, campaign, marketplace, returning customer	Identifies acquisition channel and context in which the purchase decision was initiated.
Customer engagement	Website/product views, brochure downloads, inquiry topics, campaign response, contact-channel interactions	Provides behavioral intent signals closer to the decision than static demographics.
Temporal behavior	Transaction date, inquiry-to-purchase duration, season/month, time since previous purchase, sequence of interactions	Enables trend, seasonality, recency, and temporally separated validation.

4. Conclusions

This study demonstrates that high accuracy can be misleading when vehicle transmission preference is severely imbalanced and the available predictors contain limited discriminative information. In 223,042 records, MANUAL represents 80.228% of observations, allowing a majority classifier and standard multinomial logistic regression to achieve 0.8023 accuracy while producing no minority-class predictions, reflected by balanced accuracy of 0.3333 and MCC of zero. Cost-sensitive learning increases minority sensitivity, and class-weighted random forest provides the strongest imbalance-aware performance with balanced accuracy of 0.4535, Macro-F1 of 0.3399, and MCC of 0.0969, but its low minority precision prevents strong customer-level recommendations. Statistical significance in age and gender relationships is accompanied by small effect sizes, while the joint demographic features provide only 0.0300 bits of mutual information, approximately 3.5% of target entropy. The main implication is that future gains should prioritize a reproducible data pipeline and feature enrichment over algorithmic tuning alone. Specifically, future datasets should preserve source dates and

preprocessing metadata and should add vehicle/product attributes, purchasing history, price and financing, dealer interaction, transaction source, customer engagement, and temporal behavior, followed by temporally separated or external validation and explainability analysis.

References

Abedin, M. Z., Guotai, C., Hajek, P., & Zhang, T. (2023). Combining weighted SMOTE with ensemble learning for the class-imbalanced prediction of small business credit risk. Complex & Intelligent Systems, 9, 3559–3579. https://doi.org/10.1007/s40747-021-00614-4

Araf, I., Idri, A., & Chairi, I. (2024). Cost-sensitive learning for imbalanced medical data: A review. Artificial Intelligence Review, 57(4), 80. https://doi.org/10.1007/s10462-023-10652-8

Chen, S.-X., Wang, X.-K., Zhang, H.-Y., & Wang, J.-Q. (2021). Customer purchase prediction from the perspective of imbalanced data: A machine learning framework based on factorization machine. Expert Systems with Applications, 173, 114756. https://doi.org/10.1016/j.eswa.2021.114756

Chen, W., Yang, K., Yu, Z., Shi, Y., & Chen, C. L. P. (2024). A survey on imbalanced learning: Latest research, applications and future directions. Artificial Intelligence Review, 57, 137. https://doi.org/10.1007/s10462-024-10759-6

Chicco, D., & Jurman, G. (2023). The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. BioData Mining, 16, 4. https://doi.org/10.1186/s13040-023-00322-4

Chicco, D., Tötsch, N., & Jurman, G. (2021). The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. BioData Mining, 14, 13. https://doi.org/10.1186/s13040-021-00244-z

He, J., & Cheng, M. X. (2021). Weighting methods for rare event identification from imbalanced datasets. Frontiers in Big Data, 4, 715320. https://doi.org/10.3389/fdata.2021.715320

Hicks, S. A., Strümke, I., Thambawita, V., Hammou, M., Riegler, M. A., Halvorsen, P., & Parasa, S. (2022). On evaluation metrics for medical applications of artificial intelligence. Scientific Reports, 12, 5979. https://doi.org/10.1038/s41598-022-09954-8

Jeffry, J., Usman, S., & Aziz, F. (2023). Analisis perilaku pelanggan menggunakan metode ensemble logistic regression. Jurnal Teknologi dan Ilmu Komputer Prima (JUTIKOMP), 6(2), 90–97. https://doi.org/10.34012/jutikomp.v6i2.4258

Jeffry, J., Azis, A. I. S., & Kandakon, E. T. J. (2025). Integrating Bayesian Optimization into Ensemble Logistic Regression for Explainable AI-Based Customer Behavior Analysis. Sinkron: Jurnal dan Penelitian Teknik Informatika, 9(4), 1900–1911. https://doi.org/10.33395/sinkron.v9i4.15219

Khatri, J., Marín-Morales, J., Moghaddasi, M., Guixeres, J., Giglioli, I. A. C., & Alcañiz, M. (2022). Recognizing personality traits using consumer behavior patterns in a virtual retail store. Frontiers in Psychology, 13, 752073. https://doi.org/10.3389/fpsyg.2022.752073

Kim, S., Shin, W., & Kim, H. W. (2024). Predicting online customer purchase: The integration of customer characteristics and browsing patterns. Decision Support Systems, 177, 114105. https://doi.org/10.1016/j.dss.2023.114105

Li, Y., Jia, X., Wang, R., Qi, J., Jin, H., Chu, X., & Mu, W. (2022). A new oversampling method and improved radial basis function classifier for customer consumption behavior prediction. Expert Systems with Applications, 199, 116982. https://doi.org/10.1016/j.eswa.2022.116982

Liu, Y., Li, B., Yang, S., & Li, Z. (2024). Handling missing values and imbalanced classes in machine learning to predict consumer preference: Demonstrations and comparisons to

prominent methods. Expert Systems with Applications, 237, 121694. https://doi.org/10.1016/j.eswa.2023.121694

Naglik, I., & Lango, M. (2024). GMMSampling: A new model-based, data difficulty-driven resampling method for multi-class imbalanced data. Machine Learning, 113, 5183–5202. https://doi.org/10.1007/s10994-023-06416-8

Opitz, J. (2024). A closer look at classification evaluation metrics and a critical reflection of common evaluation practice. Transactions of the Association for Computational Linguistics, 12, 820–836. https://doi.org/10.1162/tacl_a_00675

Predić, B., Ćirić, M., & Stoimenov, L. (2023). Business purchase prediction based on XAI and LSTM neural networks. Electronics, 12(21), 4510. https://doi.org/10.3390/electronics12214510

Tanvir, A.-A., Khandokar, I. A., Islam, A. K. M. M., Islam, S., & Shatabda, S. (2023). A gradient boosting classifier for purchase intention prediction of online shoppers. Heliyon, 9(4), e15163. https://doi.org/10.1016/j.heliyon.2023.e15163

Vieira, P. M., & Rodrigues, F. (2024). An automated approach for binary classification on imbalanced data. Knowledge and Information Systems, 66, 2747–2767. https://doi.org/10.1007/s10115-023-02046-7

Zhao, D., Wang, X., Mu, Y., & Wang, L. (2021). Experimental study and comparison of imbalance ensemble classifiers with dynamic selection strategy. Entropy, 23(7), 822. https://doi.org/10.3390/e23070822.