Analyzing Bike Purchase Predictions Using Machine Learning

Yuzhe Zhang[a]
School of Statistics and Mathematics, Zhejiang Gongshang University, Hangzhou, China

Keywords: Bike, Purchase, Machine Learning.

Abstract: As environmental awareness and the promotion of healthy lifestyles continue to rise globally, understanding the determinants of consumer decisions regarding bicycle purchases has become increasingly important. Using four machine learning models - Random Forest (RF), Decision Tree (DT), eXtreme Gradient Boosting (XGBoost), and Logistic Regression (LR) - this study examines the variables influencing bicycle buying behavior. The experimental results demonstrate that age, income, and car ownership consistently emerged as significant predictors of bicycle purchasing behavior across all models. Among these, RF and XGBoost exhibited the best performance in predicting bicycle purchases, with higher accuracy and robustness. The findings contribute to both theoretical advancements and practical applications, offering valuable insights for businesses and policymakers aiming to promote cycling as a sustainable mode of transportation. Furthermore, this study provides a comprehensive framework for understanding the key factors driving consumer decisions, suggesting that future research should explore hybrid models and additional socioeconomic and cultural variables for more accurate predictions.

1 INTRODUCTION

The rapid increase in environmental awareness and the promotion of healthy lifestyles have led to a significant rise in bicycle purchases worldwide (Gössling, 2020). This trend is not only evident in developed countries, where cycling infrastructure and urban planning are often more supportive of cyclists (Fishman & Cherry, 2016), but also in emerging markets where bicycles are becoming an increasingly popular mode of transport due to their affordability and environmental benefits (Zhao et al., 2022). Understanding the factors that influence consumers' decisions to purchase bicycles is essential for both theoretical advancements and practical applications (ZunigaGarcia et al., 2018).

In order to examine bicycle buying behavior, this study uses four machine learning models: Random Forest (RF), Decision Tree (DT), eXtreme Gradient Boosting (XGBoost), and Logistic Regression (LR) (Sun et al., 2019). These models were chosen for their varied strengths in handling different types of data and providing diverse insights. RF and XGBoost are ensemble methods known for their robustness and accuracy, especially in managing complex datasets with nonlinear relationships. DT models offer interpretability, allowing clear visualization of decision rules, while LR provides a straightforward approach to understanding linear relationships between variables and the target outcome.

This study aims to explore the key factors influencing bicycle purchase decisions by comparing the performance of these four models. The research objective is to identify the most significant predictors of bicycle purchases, providing insights that can inform both market strategies and policy recommendations aimed at promoting cycling as a sustainable mode of transportation.

2 LITERATURE REVIEW

2.1 Research on Bicycle Purchase Behavior Prediction

In recent years, research on predicting bicycle purchase behavior has gained considerable attention, especially in the context of rising environmental awareness and the promotion of healthy lifestyles. Studies show that socioeconomic characteristics,

[a] https://orcid.org/0009-0008-2138-3326

Zhang, Y.
Analyzing Bike Purchase Predictions Using Machine Learning.
DOI: 10.5220/0013214700004568
In *Proceedings of the 1st International Conference on E-commerce and Artificial Intelligence (ECAI 2024)*, pages 274-279
ISBN: 978-989-758-726-9

personal preferences, and environmental factors significantly influence consumer decision-making. For example, Gössling (2020) emphasized the importance of redistributing road space in cities, highlighting the role of policy interventions in reducing car dependency and promoting cycling through better-designed bike lanes. This aligns with Fishman and Cherry (2016), who found that the introduction of e-bikes could significantly alter urban transport patterns.

Similarly, Zhao et al. (2022) explored the relationship between urban expansion and cycling behavior in Chinese cities, noting that as urbanization progresses, cycling becomes more prevalent but is still significantly shaped by urban planning. Zuniga-Garcia et al. (2018) compared high- and low-cycling countries, stressing the importance of infrastructure and sociocultural factors in promoting cycling. These studies suggest that improving infrastructure, adjusting transportation policies, and addressing cultural factors are key to increasing bicycle purchases.

In Beijing, Sun et al. (2019) found that commuting distance, health awareness, and socioeconomic background significantly impact bicycle purchasing behavior. Their research indicated that higher-income groups are more likely to purchase bicycles, particularly as a means of healthy and environmentally friendly transportation. Comprehending these variables is essential for formulating focused market approaches and policy measures.

2.2 Application of Machine Learning Models in Economics

Machine learning techniques have become widely used in economic data analysis due to their ability to handle complex datasets. Feng et al. (2021) applied the RF model to analyze pedestrian injury severity under different weather conditions, demonstrating the model's strong capability in handling multi-dimensional data and capturing non-linear relationships. The study showed that RF excels in capturing complex interactions between features, making it ideal for tasks such as credit risk evaluation and market prediction.

In terms of DT models, Wibowo and Wibowo (2020) investigated their performance in classifying credit scores. DT models, known for their simplicity and interpretability, are frequently used in economic decision support systems, especially in scenarios requiring transparent decision paths, such as customer segmentation and fraud detection.

Due to its effectiveness in managing complicated feature interactions and massive amounts of data, XGBoost has become more and more well-known in recent years. In a comparative study of XGBoost, Agrawal and Maurya (2020) discovered that the model is very good at handling non-linear features and avoids overfitting, which makes it ideal for applications like credit scoring and financial market prediction.

3 MODEL DESCRIPTION

3.1 Random Forest

Random Forest (RF) is an ensemble learning technique (Feng et al., 2021) that generates several decision trees during the training phase. For classification tasks, the output is the most frequent class prediction, while for regression tasks, the output is the average prediction.

According to Feng et al. (2021), RF is an ensemble learning technique that builds several DTs during training and produces the mean prediction (regression) or mode of the classes (classification) of the individual trees. The classification task is shown in Equation (1).

$$\hat{y} = mode\{T_b(x)\}_{b=1}^{B} \quad (1)$$

The regression task is shown in Equation (2).

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \quad (2)$$

Here $T_b(x)$ rdenotes the prediction from the b-th tree for input x, while B is the total number of trees. b is the index ranging from 1 to B, representing each individual tree in the RF.

The RF algorithm, introduced by Breiman, is designed to improve the accuracy and robustness of DTs by aggregating the results of multiple trees, thereby reducing overfitting and enhancing generalization. RFs are particularly effective for large datasets with high dimensionality and complex feature interactions, making them a popular choice in economics for tasks such as credit risk evaluation, stock market prediction, and consumer behavior analysis. The capacity of RFs to handle both numerical and categorical variables and to provide feature importance estimates, which aid in the

interpretation of model results, is one of their main advantages.

3.2 Decision Tree

According to the value of the input features, a Decision Tree (DT) model divides the data into subsets (Wibowo. A & Wibowo. P. B, 2020). The Gini impurity, which calculates the likelihood that an element selected at random would receive an inaccurate label if its label were assigned at random based on the dataset's label distribution is defined as Equation (3).

$$Gini = 1 - \sum_{i=1}^{n} p_i^2 \quad (3)$$

p_i represents the proportion of observations belonging to class i, and n is the number of classes.

DTs are one of the simplest yet powerful predictive modeling approaches. They are highly interpretable because they represent decisions and their possible consequences in a tree-like structure. This method is particularly useful when the primary goal is to gain insights into the decision-making process rather than just prediction. Due to its capacity to manage intricate relationships between variables and produce precise, actionable rules, DT is frequently employed in economic applications for purposes such as consumer segmentation, fraud detection, and identifying critical drivers of economic outcomes.

3.3 Extreme Gradient Boosting

A gradient-boosting system called Extreme Gradient Boosting (XGBoost) creates DTs in a sequential fashion with the goal of fixing mistakes committed by earlier trees. Equation (4) provides a definition for the objective function.

$$L(\phi) = \sum_{i=1}^{n} l\,(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega\,(f_k) \quad (4)$$

Define $l(y_i, \hat{y}_i)$as the loss for prediction $\hat{y}_i$, and $\Omega(f_k)$ as the regularization term for tree k. It was developed by Chen and Guestrin (2020), XGBoost is known for its efficiency, speed, and scalability (Lundberg & Lee, 2020). It has become one of the most popular machine learning algorithms in recent years, particularly in data science competitions and realworld applications that require high predictive accuracy (Agrawal & Maurya, 2020). XGBoost's ability to handle missing data, capture complex patterns, and prevent overfitting through regularization makes it ideal for economic modeling tasks, such as predicting financial market trends, credit scoring, and analyzing large-scale economic data (Chen & Guestrin, 2020).

3.4 Logistic Regression

Logistic Regression (LR) is a linear model for binary classification. The logistic function, which models the probability that a given input point belongs to a particular class, is given by Equation (5).

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}} \quad (5)$$

β_0 is the intercept, $\beta_1 \ldots \beta_n$ are the coefficients, and x_i are the predictor variables.

Because of its ease of use and efficiency in binary classification problems, LR is one of the most popular statistical models. It is easily interpreted and simple to apply because it assumes a linear relationship between the input variables and the response variable's log odds. In economic research, LR is often employed to model binary outcomes such as purchase decisions, default risk, and employment status. Its ability to provide insights into the influence of predictor variables and its relatively low computational cost makes it a preferred choice for many applied econometric analyses.

4 EMPIRICAL ANALYSIS

4.1 Model Results Comparison

As shown in Table 1, in this study, four machine learning models - RF, DT, XGBoost, and LR - were employed to analyze bicycle purchase behavior. The models predict whether a user will purchase a bicycle, with the positive class representing "purchase" and the negative class representing "no purchase." The performance of each model is summarized in terms of accuracy, precision, recall, and F1-score.

Among the models, RF demonstrated the highest overall performance, with an accuracy of 71%, followed closely by XGBoost at 69.5%. In contrast, the DT and LR models performed less effectively, achieving accuracies of 63% and 56%, respectively. It is worth noting that RF and XGBoost performed better in both precision and recall, particularly for the positive class (bicycle purchase), indicating their superior ability to identify actual purchasing

Table 1: Model Performance Comparison.

Model	Accuracy	Precision (0)	Precision (1)	Recall (0)	Recall (1)	F1 (0)	F1 (1)
RF	0.7100	0.7182	0.7000	0.7453	0.6702	0.7315	0.6848
DT	0.6300	0.6455	0.6111	0.6698	0.5851	0.6574	0.5978
XGBoost	0.6950	0.7228	0.6667	0.6887	0.7021	0.7053	0.6839
LR	0.5600	0.5833	0.5326	0.5943	0.5213	0.5888	0.5269

Table 2: Feature Importance Across Models.

Feature	RF	DT	XGBoost	LR
Income	0.1220	0.1404	0.0335	0.4216
Children	0.0959	0.0786	0.0309	-0.2141
Cars	0.0975	0.0976	0.0581	-0.6129
Age	0.2078	0.1982	0.0356	0.1013
Martial Status_Single	0.0454	0.0538	0.0327	0.3751
Gender_Male	0.0507	0.0493	0.0311	0.0844
Education_Graduate Degree	0.0168	0.0223	0.0397	-0.1863
Education_High School	0.0220	0.0320	0.0324	0.0791
Education_Partial College	0.0235	0.0126	0.0193	-0.0255
Education_Partial High School	0.0148	0.0088	0.1208	-0.1024
Occupation_Management	0.0117	0.0093	0.0137	-0.0753
Occupation_Manual	0.0125	0.0054	0.0251	-0.0827
Occupation_Professional	0.0198	0.0188	0.0371	0.1073
Occupation_Skilled Manual	0.0195	0.0219	0.0341	-0.0659
Home Owner_Yes	0.0358	0.0355	0.0341	0.1844
Commute Distance_1-2 Miles	0.0333	0.0419	0.0431	-0.0991
Commute Distance_2-5 Miles	0.0309	0.0409	0.0336	-0.0855
Commute Distance_5-10 Miles	0.0311	0.0414	0.0439	-0.3285
Commute Distance_More than 10 Miles	0.0184	0.0084	0.0292	-0.3770
Region_North America	0.0271	0.0119	0.0351	0.0217
Region_Pacific	0.0222	0.0350	0.0938	0.3419
Age Brackets_Middle Age	0.0255	0.0301	0.1432	0.2280
Age Brackets_Old	0.0157	0.0059	0.0000	0.0098

behavior. Although LR provides strong linear interpretability, its performance is hindered by its inability to capture complex relationships between variables, resulting in lower accuracy and recall rates.

4.2 Feature Importance Analysis

As shown in Table 2, in this study, based on the model results, this paper conducted a detailed analysis of the key features influencing bicycle purchases. Below, this paper focuses on the most important features, considering both model performance and their real-world implications.

The positive category represents users who buy bicycles, and the negative category represents users who do not buy bicycles. Understanding which characteristics make users more inclined to buy or not buy bicycles, can provide relevant policymakers and enterprises with a reference to help them better promote bicycle sales or develop related marketing strategies.

Income emerged as a critical factor in determining bicycle purchases across all models. In the LR model, the income coefficient is 0.421572, indicating a strong positive correlation between higher income levels and the likelihood of purchasing a bicycle. This aligns with real-world consumer behavior, where higher-income individuals are more likely to view bicycles as part of a healthy lifestyle or as an environmentally friendly mode of transport. Furthermore, higher income levels enable the purchase of higher-end bicycles, which could contribute to this trend.

The number of cars owned by a household or individual showed a negative correlation with bicycle purchases in multiple models. For instance, in LR, the coefficient for car ownership is 0.612922, suggesting that owning more cars significantly decreases the likelihood of purchasing a bicycle. This is consistent with the notion that car ownership reduces the need for bicycles as a primary mode of transport. Policies aimed at reducing car dependency or promoting

bicycles as a viable transportation alternative may help increase bicycle sales.

Age was identified as one of the most important predictors in the RF and DT models. Older individuals were more likely to purchase bicycles, a finding that may be explained by health considerations. As people age, they often become more concerned with maintaining physical health, and cycling is increasingly viewed as an accessible and beneficial form of exercise. From a practical perspective, marketing campaigns targeting older demographics that emphasize the health benefits of cycling could further stimulate bicycle sales in this segment.

4.3 Key Factors in Bicycle Purchase Behavior

The results of this study highlight the key factors influencing bicycle purchase behavior as identified by the four models. Age, income, and car ownership consistently emerged as significant predictors regardless of the model used. These insights provide valuable information for businesses seeking to target potential customers, as well as policymakers looking to promote cycling as a sustainable transportation option.

For example, RF and XGBoost both identified age as a crucial factor, indicating that older individuals are more likely to purchase bicycles. This finding suggests that marketing strategies could be designed specifically to appeal to older demographics by emphasizing the health benefits of cycling and positioning bicycles as an effective means of maintaining physical well-being.

4.4 Policy and Business Implications

Similarly, income played a vital role in predicting bicycle purchases across all models. Higher-income individuals are more likely to buy bicycles, possibly due to their greater financial capacity to invest in high-quality bicycles and a lifestyle that promotes sustainable transportation. Businesses could focus on premium pricing strategies or offer advanced bicycle features targeting affluent consumers.

Car ownership, which showed an inverse relationship with bicycle purchases, highlights an area of potential policy intervention. Reducing car dependency and offering incentives for cycling - such as building more cycling infrastructure or providing tax breaks for bicycle purchases - could increase bicycle sales. This insight helps policymakers create effective strategies for promoting healthier, more sustainable urban environments.

5 CONCLUSIONS

The rapid rise in environmental awareness and the promotion of healthier lifestyles have significantly contributed to the increase in bicycle purchases globally. Understanding the factors driving consumer decisions to purchase bicycles is crucial for both theoretical research and practical applications, as it offers valuable insights for businesses and policymakers alike. This study applied four machine learning models—RF, DT, XGBoost, and LR—to analyze bicycle purchase behavior. Each model demonstrated unique advantages, with RF and XGBoost showing the most robust predictive performance. Key features such as age, income, and car ownership were identified as significant predictors, reflecting real-world consumer trends and socioeconomic factors. This study is not without limits, though. Initially, the models employed were restricted to a particular group of variables, which might not adequately represent the intricacy of customer behavior. Additionally, the data used for model training may not account for regional or cultural differences in bicycle purchasing patterns. These limitations suggest that further research should incorporate additional variables, such as lifestyle choices, environmental policies, and urban infrastructure, to better understand consumer behavior. Future studies should investigate hybrid models, which combine the best features of various machine learning approaches to increase interpretability and accuracy. Expanding the scope of analysis to include broader socioeconomic and cultural factors will allow for a more comprehensive understanding of bicycle purchase behavior. Additionally, investigating the role of infrastructure development, such as bike lanes and public cycling initiatives, could provide further insights into how urban planning can promote cycling as a primary mode of transportation.

REFERENCES

Agrawal, S., Maurya, A. K., 2020. A comparative analysis of XGBoost. International Journal of Scientific & Technology Research, 9(2), 132-136.

Chen, T., Guestrin, C., 2020. XGBoost: A Scalable Tree Boosting System. ACM Transactions on Knowledge Discovery from Data, 14(1), 1-24.

Feng, J., Wu, Y., Xu, Z., 2021. RF-based pedestrian injury severity analysis for different weather conditions. Accident Analysis & Prevention, 149, 105870.

Fishman, E., Cherry, C., 2016. E-bikes in the mainstream: Reviewing a decade of research. Transport Reviews, 36(1), 72-91.

Gössling, S., 2020. Why cities need to take road space from cars—and how this could be done. Journal of Urban Design, 25(4), 443-448.

Lundberg, S. M., Lee, S. I., 2020. A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems, 33.

Sun, G., Li, W., Zeng, C., 2019. Bicycle commuting behavior in Beijing: Lessons from the commuters. Transport Policy, 82, 58-67.

Wibowo, A., Wibowo, P. B., 2020. Performance of Decision Tree algorithms in classifying credit scoring. Journal of Physics: Conference Series, 1566(1), 012033.

Zhao, P., Lu, B., Ge, J., 2022. Urban expansion and cycling behavior: Evidence from Chinese cities. Journal of Transport Geography, 100, 103284.

ZunigaGarcia, N., Nazelle, A., Nieuwenhuijsen, M. J., 2018. Cycling in cities: Key lessons learned from high- and low-cycling countries. International Journal of Environmental Research and Public Health, 15(4), 556.