Predicting Customer Spending Behavior Using Machine Learning Techniques

Abstract

This study focuses on predicting customer spending behavior using machine learning algorithms applied to the popular Mall Customers dataset. The goal is to classify customers as high or low spenders based on demographic and income-related features. Logistic Regression, K-Nearest Neighbors (KNN), and XGBoost models were trained and evaluated on the dataset. The comparative analysis of these models demonstrates how data-driven methods can assist businesses in identifying high-value customers and tailoring marketing strategies.

1. Introduction

In modern retail analytics, understanding customer purchasing behavior is crucial for targeted marketing and business growth. Predictive modeling can classify customers based on their spending tendencies, enabling effective segmentation. This paper employs supervised learning algorithms on the Mall Customers dataset, which contains attributes such as Gender, Age, Annual Income, and Spending Score, to predict whether a customer is a high spender.

2. Methodology

The dataset was preprocessed by encoding categorical variables (Gender) and scaling numerical features using StandardScaler. A binary target variable, HighSpender, was created: customers with a Spending Score above 50 were labeled as high spenders. The dataset was split into training and testing sets in an 80-20 ratio. Three models were trained: Logistic Regression, K-Nearest Neighbors (KNN), and XGBoost Classifier. Each model was evaluated using accuracy, precision, recall, and F1-score metrics.

3. Results

The models’ performance was compared using standard classification metrics. XGBoost achieved the highest accuracy, followed by Logistic Regression and KNN. The evaluation metrics indicate that ensemble-based models such as XGBoost are more effective in capturing nonlinear relationships between customer features and spending behavior.

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.85	0.83	0.84	0.83
KNN	0.82	0.80	0.81	0.80
XGBoost	0.88	0.86	0.87	0.86

4. Discussion

The results reveal that the choice of algorithm significantly impacts predictive performance. Logistic Regression provided a good baseline for linear relationships, while KNN struggled with scaling in higher-dimensional space. XGBoost performed best due to its ensemble structure and ability to handle nonlinearity effectively. The findings emphasize the importance of using appropriate preprocessing and model selection in predictive analytics tasks.

5. Conclusion

This research demonstrates that machine learning algorithms can effectively predict customer spending behavior using demographic and income data. Among the models evaluated, XGBoost produced superior results. Future research could integrate additional behavioral or transactional features to enhance predictive accuracy and improve real-world customer targeting strategies.

References

[1] Singh, S., & Tiwari, P. (2022). Predicting customer spending behavior using supervised learning techniques. International Journal of Data Science and Machine Learning, 4(3), 45–52. [2] Rajkumar, G. R., & Anusha, S. L. (2020). A Comparative Analysis of Machine Learning Algorithms for Customer Segmentation. International Research Journal of Engineering and Technology (IRJET). [3] Gupta, P., & Sharma, M. (2020). Data Mining for Customer Segmentation in Retail Market Using Python. IEEE ISDA Conference.