Predicting Customer Satisfaction Using Machine Learning: Evidence from a Synthetic Customer Feedback Dataset

Nurana Sadigova

PhD Candidate, Azerbaijan State University of Economics (UNEC), Baku, Azerbaijan
sadiqova.nurane@gmail.com

Magsud Mirzayev

Asst. Prof, Azerbaijan State University of Economics (UNEC), Baku, Azerbaijan
magsud-mirzeyev@unec.edu.az

Abstract: Customer satisfaction is a critical determinant of business performance, customer retention, and long-term competitiveness. As organizations increasingly rely on data-driven decision-making, the ability to accurately predict customer satisfaction has become an important research and managerial objective. This study aims to identify and predict the key determinants of customer satisfaction using a large synthetic customer feedback dataset consisting of 38,444 observations.

The analytical framework combines Exploratory Data Analysis (EDA), Pearson correlation analysis, Multiple Linear Regression (MLR), and Random Forest machine learning techniques. To ensure the robustness of the findings, the dataset was divided into training (80%) and testing (20%) subsets, and model performance was further evaluated using a 5-fold cross-validation procedure. In addition, feature importance, permutation importance, and SHAP (Shapley Additive Explanations) analyses were employed to enhance model interpretability.

The correlation analysis revealed that Service Quality ($r = 0.55$) and Product Quality ($r = 0.55$) exhibit the strongest positive relationships with customer satisfaction, while Income demonstrates a moderate association ($r = 0.25$). The Multiple Linear Regression results confirmed that Service Quality ($\beta = 9.26$) and Product Quality ($\beta = 9.24$) are the most influential predictors of customer satisfaction, followed by Income ($\beta = 4.11$) and Age ($\beta = 2.71$). Furthermore, the Random Forest model outperformed the baseline regression model, achieving an R^2 value of 0.776 on the hold-out test set and a mean cross-validated R^2 of 0.781. Feature importance and SHAP analyses consistently identified Service Quality and Product Quality as the dominant drivers of customer satisfaction.

The findings suggest that perceived service and product quality play a substantially greater role in shaping customer satisfaction than demographic and behavioral characteristics. By integrating traditional statistical techniques with advanced machine learning and explainable artificial intelligence methods, this study contributes to the growing literature on customer analytics and demonstrates the practical value of synthetic data for customer satisfaction prediction. The results provide actionable insights for organizations seeking to enhance customer experience and improve service quality through data-driven strategies.

Keywords: customer satisfaction, machine learning, service quality, synthetic data

1 Introduction

Customer satisfaction has long been one of the most extensively researched topics in the field of marketing. In the classical approach, satisfaction is perceived as a subjective evaluation formed by comparing customer expectations with the service provided. Research conducted in line with this perspective demonstrates that high levels of satisfaction are directly correlated with increased organizational performance and customer loyalty. Subsequently, researchers have proposed numerous models and indices to measure customer satisfaction. Service quality models, particularly the SERVQUAL approach, have been significantly utilized in assessments by taking various service parameters—such as empathy, reliability, and responsiveness—into account. However, due to these models' reliance on static analysis, a need for predictive and dynamic approaches has emerged. Consequently, unlike traditional surveying techniques, recent machine learning (ML) approaches provide a more efficient means of modeling customer behavior through the analysis of big data [1].

The measurement and study of customer satisfaction have entered a new era with the advancement of machine learning and data analytics methods. In recent studies, the use of diverse algorithms to improve customer behavior analysis and predict future satisfaction levels has gained significant momentum. Decision trees, including Random Forest and their ensemble variants, are considered effective in identifying relationships between variables. Findings from several studies indicate that machine learning methods are more efficient and accurate compared to traditional statistical methods. Furthermore, it can be concluded that methods such as Support Vector Machines (SVM) and neural networks prevail in modeling nonlinear structures. Additionally, issues regarding overfitting and model interpretability remain central topics of discussion in this field [5]. Although, organizations face rising demands for research-identifiable records, they are developing innovative strategies to improve data accessibility. One promising solution is the creation of synthetic datasets, which can effectively bypass

traditional hurdles related to data access, user privacy, and confidentiality barriers [11].

The use of synthetic data has also become a subject of debate in recent years. By mimicking the structure of real data, synthetic datasets allow researchers to test models without encountering privacy and ethical concerns. This approach serves as a viable solution, especially when access to real customer data is limited. While synthetic data has been applied in various research contexts, the literature lacks a broad perspective. Most discussions are currently siloed within specific project frameworks or isolated methodologies, particularly in the health sector too [3].

The primary objective of this research is to predict customer satisfaction using machine learning methods based on synthetic customer feedback data. In addition to bypassing the restrictions on accessing real data, the use of synthetic datasets enables models to be tested under various scenarios. As a result of the study, the impact of different variables on satisfaction is evaluated, and the findings are interpreted from the perspective of practical implementation.

2 Literature Review

Customer satisfaction has long been recognized as one of the most important indicators of business success. Organizations that consistently achieve high levels of customer satisfaction are more likely to retain customers, strengthen loyalty, and improve financial performance. For this reason, customer satisfaction has remained a central topic in marketing, service management, and customer relationship management research. Traditionally, researchers relied on survey-based methods and statistical models to measure customer satisfaction. However, the rapid growth of digital technologies, e-commerce platforms, and customer feedback systems has created new opportunities for data-driven approaches that can not only explain satisfaction but also predict it [7].

In recent years, machine learning has emerged as a powerful tool for customer analytics. Unlike traditional statistical techniques, machine learning algorithms are capable of identifying complex and nonlinear relationships within large datasets. As a result, they have become increasingly popular for analyzing customer behavior, predicting satisfaction levels, and supporting managerial decision-making. Recent studies suggest that machine learning models often outperform conventional statistical methods in predictive tasks because they can process large volumes of data and uncover patterns that linear models may not capture. Research by Mozumder et al. (2024) shows that the application of machine learning models offers significant improvements in the measurement and analysis of customer satisfaction, playing a pioneering role in strategic decision-making.

Machine learning has become especially relevant in e-commerce and digital retail contexts, where large-scale customer feedback can be used to predict satisfaction levels and identify key drivers of customer experience. Zaghloul (2024) compared traditional and machine learning approaches for predicting e-commerce customer satisfaction and found that Random Forest achieved strong predictive performance, outperforming several alternative approaches. Similarly, Le (2024) proposed a two-step predictive framework combining deep learning and traditional machine learning methods to analyze Vietnamese e-commerce reviews, showing the usefulness of hybrid models for satisfaction analytics. These studies confirm that machine learning methods are increasingly valuable for transforming customer feedback into predictive business intelligence.

The growing adoption of machine learning methods in customer satisfaction research has enabled organizations to uncover complex patterns within customer reviews and related datasets, providing data-driven insights that strengthen decision-making processes. The strategic adoption of predictive modeling empowers retailers to unlock untapped profit potential, facilitating sustainable business growth and reinforcing competitive resilience amid the growing complexity and volatility of the retail industry [8].

The use of synthetic data has also become increasingly relevant in machine learning research. Real customer data often contain sensitive personal information, which creates privacy, confidentiality, and accessibility challenges. Synthetic data provide an alternative by generating artificial datasets that preserve important statistical patterns without directly exposing real individuals. Rajotte et al. (2022) describe synthetic data generation as a process that captures patterns in real datasets through machine learning methods, while Sanchez-Serrano et al. (2025) emphasize the importance of balancing data utility and privacy protection in synthetic tabular data. Therefore, synthetic customer feedback datasets can be useful for testing predictive models when real customer-level data are unavailable or restricted.

The literature also highlights the importance of rigorous model validation procedures. Contemporary customer analytics studies increasingly employ techniques such as train-test splitting, cross-validation, and multiple performance metrics to ensure the reliability and generalizability of predictive models. Evaluation measures such as R^2, MAE, and RMSE are commonly used in regression-based studies to assess predictive accuracy and compare alternative algorithms. Such validation procedures are particularly important when machine learning models are applied to support business decisions [2].

Overall, the existing literature suggests that customer satisfaction is influenced primarily by service quality and product quality, while demographic and behavioral factors tend to play a secondary role. Furthermore, machine learning methods—particularly ensemble learning algorithms—have demonstrated substantial potential for improving customer satisfaction prediction. Building on these developments, the present study applies both traditional statistical techniques and advanced machine

learning approaches to identify the key determinants of customer satisfaction and evaluate their predictive power using a large synthetic customer feedback dataset.

3 Hypotheses

Based on the theoretical foundations and empirical evidence reviewed above particularly the established roles of perceived service and product quality, and the secondary but non-negligible contribution of demographic and behavioral factors the following hypotheses are proposed for empirical testing:

Hypothesis 1 (H1): Service quality exerts a positive influence on customer satisfaction.

Hypothesis 2 (H2): Product quality is positively associated with levels of customer satisfaction.

Hypothesis 3 (H3): Higher income levels are expected to positively affect overall satisfaction.

Hypothesis 4 (H4): Purchase frequency serves as a positive predictor of customer satisfaction.

Hypothesis 5 (H5): Customer age is significantly associated with customer satisfaction.

These hypotheses are tested in the following sections using correlation analysis, Multiple Linear Regression, and Random Forest based feature importance and interpretability techniques.

4 Methodology

4.1 Data Source and Variables

The dataset consists of 38,444 individual customer observations. Key variables analyzed include demographic traits (age, gender, country, income), behavioral patterns (purchase frequency), and perceptual measures (product and service quality), all of which are assessed in relation to the satisfaction score. Synthetic data were employed to overcome privacy and accessibility limitations associated with real customer information while preserving the statistical characteristics necessary for predictive analytics.

4.2 Analytical Procedure

The analytical framework was designed to provide both robust predictive performance and meaningful insights into the determinants of customer satisfaction. The analysis was conducted in several sequential stages.

First, Exploratory Data Analysis (EDA) was performed to assess data quality, examine variable distributions, and identify potential anomalies. Descriptive statistics and visualization techniques were used to ensure the suitability of the dataset for subsequent analysis.

Second, Pearson correlation analysis was employed to explore the relationships between the independent variables and customer satisfaction. This step provided an initial understanding of the strength and direction of associations among the variables.

Third, Multiple Linear Regression (MLR) was utilized as a baseline statistical approach to test the proposed hypotheses and estimate the influence of explanatory variables on customer satisfaction. The model's explanatory power was evaluated using regression coefficients and the coefficient of determination (R^2).

To capture potential nonlinear relationships and improve predictive accuracy, a Random Forest regression model was subsequently developed. Random Forest was selected because of its ability to model complex interactions among variables while reducing the risk of overfitting.

For model validation, the dataset was randomly divided into training (80%) and testing (20%) subsets. Predictive performance was assessed using the coefficient of determination (R^2), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE). In addition, a 5-fold cross-validation procedure was conducted to evaluate the stability and generalizability of the models across different data partitions.

To enhance model interpretability, several complementary explainability techniques were applied. Random Forest Feature Importance (Mean Decrease in Impurity) was used to identify the most influential predictors, while Permutation Importance Analysis assessed the impact of each variable on model performance by measuring the decrease in predictive accuracy after random shuffling. Furthermore, SHAP (Shapley Additive Explanations) analysis was employed to quantify both the magnitude and direction of each variable's contribution to customer satisfaction predictions.

Finally, Partial Dependence Plots (PDPs) were generated for the most influential predictors to visualize their marginal effects on customer satisfaction while holding all other variables constant.

This multi-method analytical framework combines traditional statistical techniques with advanced machine learning and explainable artificial intelligence methods,

providing a comprehensive and reliable assessment of the factors influencing customer satisfaction.

5 Findings and Discussion

The statistical analysis of the synthetic customer data under study indicates significant relationships between customer satisfaction and diverse variables. The visual findings are interpreted as follows:

The coefficient of determination (R^2 = 0.7001)

	Factors	Effect size (Beta)
3	ServiceQuality	9.261248
2	ProductQuality	9.237983
1	Income	4.107652
0	Age	2.709117
4	PurchaseFrequency	1.857789
5	Gender_num	0.072845
7	Loyalty_num	0.001522
6	Feedback_num	-0.022677

Regression analysis

The conducted multivariate linear regression and correlation analyses have enabled the identification of the key factors influencing customer satisfaction (Satisfaction Score). The coefficient of determination for the developed model is R^2= 0.7001, indicating that approximately 70% of the variation in customer satisfaction is explained by the independent variables included in the model.

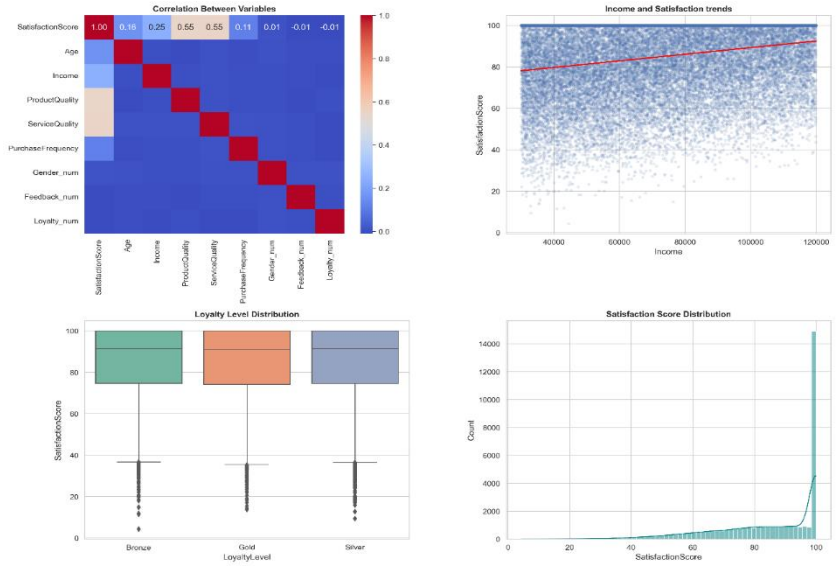
The empirical findings provide strong support for the proposed hypotheses. Specifically, the high beta coefficients for Service Quality and Product Quality validate H1 and H2, establishing them as primary drivers of satisfaction. These findings align with A. Parasuraman's service quality framework, further validating that customer experience is a critical determinant of overall satisfaction. The positive effect sizes for Income (H3) and Purchase Frequency (H4) further confirm that economic and behavioral factors significantly enhance satisfaction levels. Additionally, the significant effect size for Age (beta = 2.71) supports H5, confirming its role as a relevant demographic determinant in this context.

Correlation Analysis

	SatisfactionScore
SatisfactionScore	1.000000
ServiceQuality	0.553614
ProductQuality	0.547690
Income	0.245129
Age	0.157510
PurchaseFrequency	0.113018
Gender_num	0.007579
Loyalty_num	-0.006690
Feedback_num	-0.008227

Correlation analysis

Based on the inter-variable correlation matrix, it was determined that 'Product Quality' (0.55) and 'Service Quality' (0.55) have a moderate positive correlation with 'Satisfaction Score'. This confirms that product and service quality are key determinants of customer satisfaction. The correlation of 'Age' (0.16) and 'Income' (0.25) with satisfaction is relatively weak, while other variables (gender, loyalty, etc.) exhibit even lower predictive power."



Income and Satisfaction Relationship

Based on the scatter plot and correlation analysis, a weak yet positive trend is observed between Income and Satisfaction Score. The regression line indicates that as income levels rise, satisfaction increases marginally; however, the significant dispersion of data points suggests that this relationship is not particularly robust.

This pattern implies that satisfaction is influenced by various underlying factors beyond financial capacity alone. Ultimately, while income plays a role, the primary drivers of satisfaction remain Service Quality and Product Quality, which exhibit substantially higher effect sizes and correlations.

Loyalty Level Analysis

The "Loyalty Level Distribution" analysis reveals that satisfaction scores are distributed remarkably similarly across the Bronze, Gold, and Silver loyalty tiers. For all three categories, the median satisfaction values remain consistently high, converging around the 90+ range. However, the presence of outliers in the lower quartiles of the boxplot indicates that a specific subset of customers reports significantly lower satisfaction levels, regardless of their loyalty status.

Distribution of Satisfaction Scores

The histogram results indicate that satisfaction scores are primarily concentrated within the 80–100 range. The distribution exhibits a right-concentrated (negatively skewed) structure, revealing an exceptionally high level of satisfaction within this synthetic dataset. This concentration suggests that the vast majority of customers are categorized as "highly satisfied," which poses a potential risk of model bias during the training phase.

Furthermore, such a heavily skewed distribution serves as a precursor to class imbalance issues in future machine learning models. This lack of variance in the lower satisfaction ranges may limit the model's ability to accurately identify and learn the drivers of dissatisfaction, potentially overestimating the positive impact of features like Service Quality and Product Quality.

Model Validation

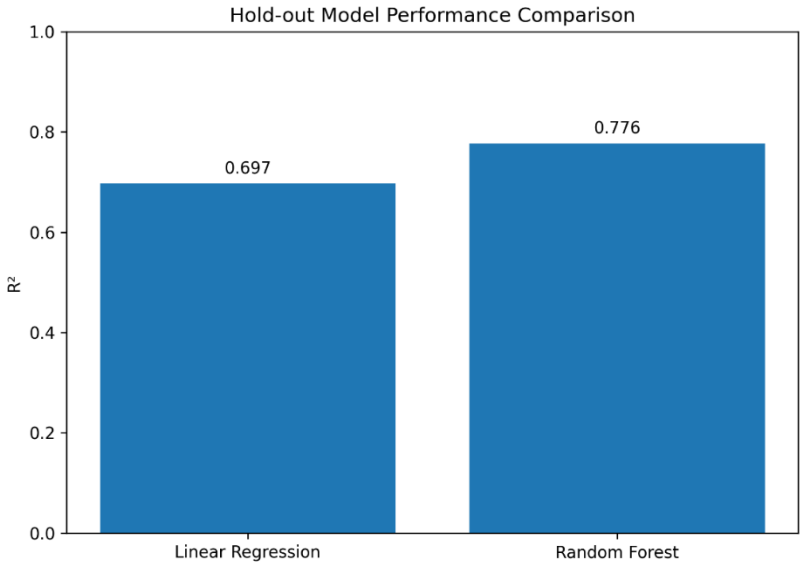
The dataset was divided into training and testing subsets using an 80/20 split. Linear Regression was used as a baseline model, while Random Forest was used as an ensemble learning method capable of capturing nonlinear patterns. Model performance was evaluated using R^2, MAE, and RMSE.

Model	**R^2**	**MAE**	**RMSE**
Linear Regression	0.697	7.38	9.22
Random Forest	0.776	5.53	7.92

Table 1.

Hold-out model performance comparison

The hold-out results show that Random Forest achieved R^2=0.776, MAE=5.53, and RMSE=7.92. Compared with Linear Regression, which achieved R^2=0.697, the Random Forest model demonstrated stronger predictive performance.

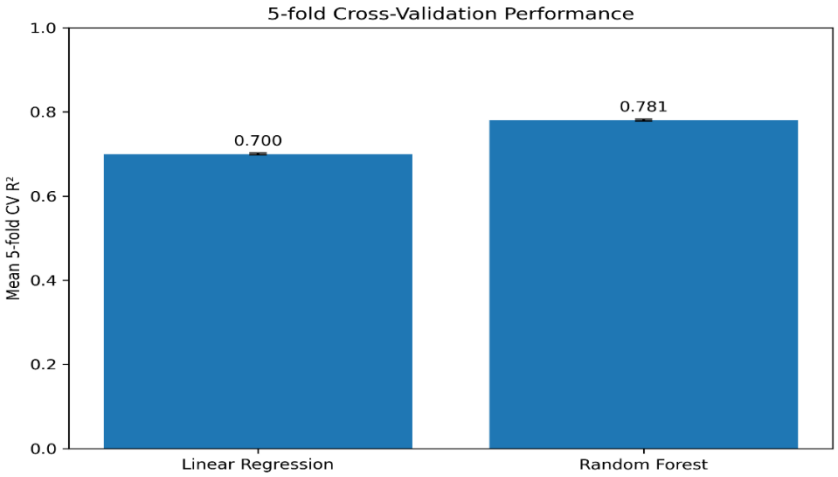


Model	**Mean R²**	**SD R²**	**Mean MAE**	**SD MAE**	**Mean RMSE**	**SD RMSE**
Linear Regression	0.700	0.002	7.40	0.03	9.26	0.03
Random Forest	0.781	0.002	5.51	0.03	7.91	0.04

Table 2.

Five-fold cross-validation performance

The 5-fold cross-validation results further confirm model stability. Random Forest achieved a mean cross-validated R^2 of 0.781 (SD=0.002), while Linear Regression achieved a mean R^2 of 0.700. The low standard deviations indicate that the results are consistent across folds and are not dependent on a single random split.

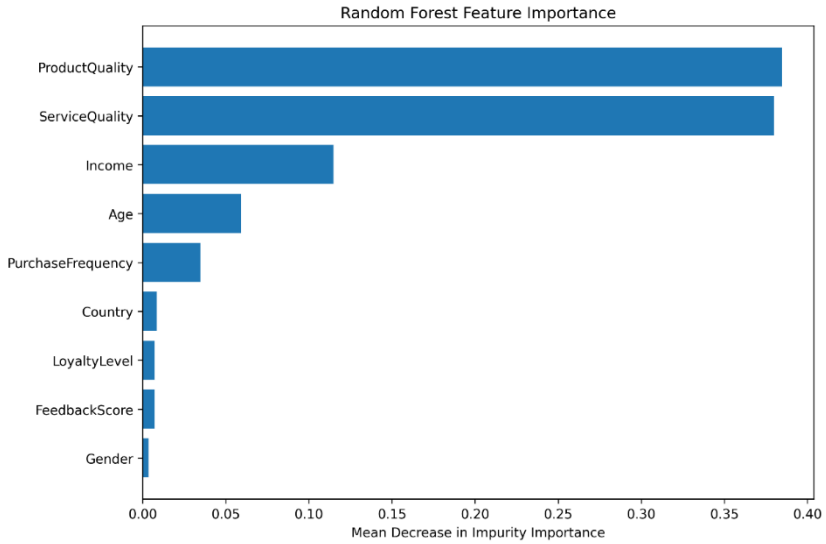


Random Forest Feature İmportance

Impurity-based Random Forest feature importance was calculated to identify the variables that most strongly contributed to predictive performance. ProductQuality and ServiceQuality were the dominant predictors, followed by Income, Age, and PurchaseFrequency.

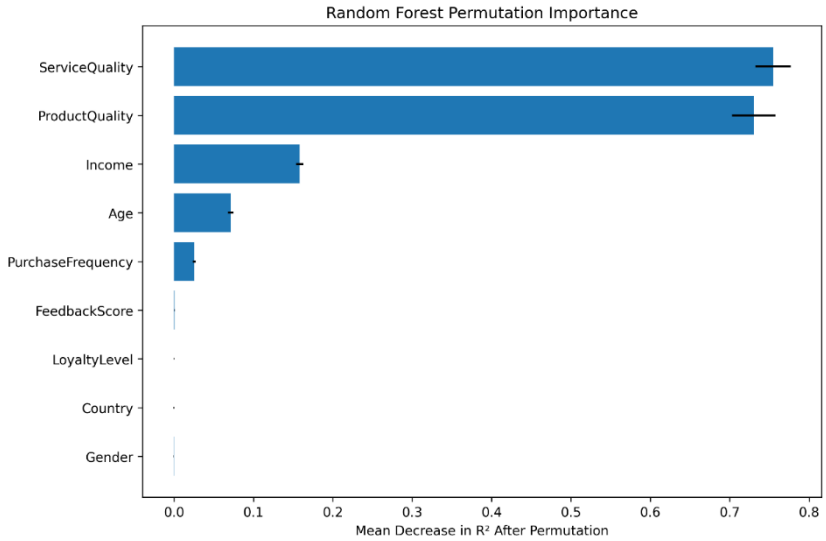
Rank	Variable	Importance
1	ProductQuality	0.385
2	ServiceQuality	0.380
3	Income	0.115
4	Age	0.059
5	PurchaseFrequency	0.035
6	Country	0.009
7	LoyaltyLevel	0.007
8	FeedbackScore	0.007
9	Gender	0.003

Table 3.
Random Forest impurity-based feature importance



Permutation importance

Permutation importance was also applied because it evaluates the decrease in model performance after randomly shuffling each predictor. This method provides a more validation-oriented interpretation than impurity-based importance.



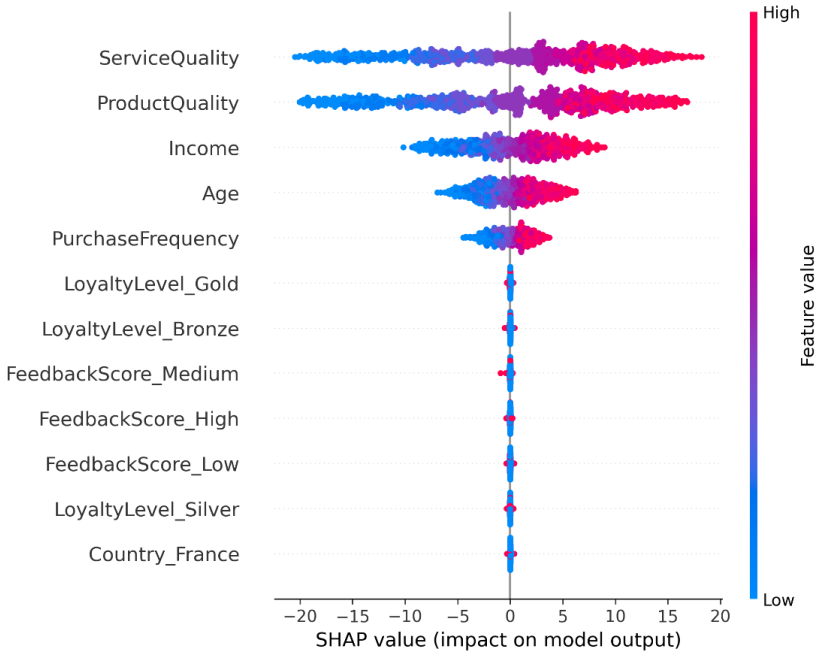
Rank	Variable	Mean decrease in R^2	SD
1	ServiceQuality	0.755	0.022
2	ProductQuality	0.731	0.027
3	Income	0.158	0.005
4	Age	0.071	0.004
5	PurchaseFrequency	0.025	0.002
6	FeedbackScore	0.001	0.000
7	LoyaltyLevel	0.000	0.000
8	Country	-0.000	0.000
9	Gender	-0.000	0.000

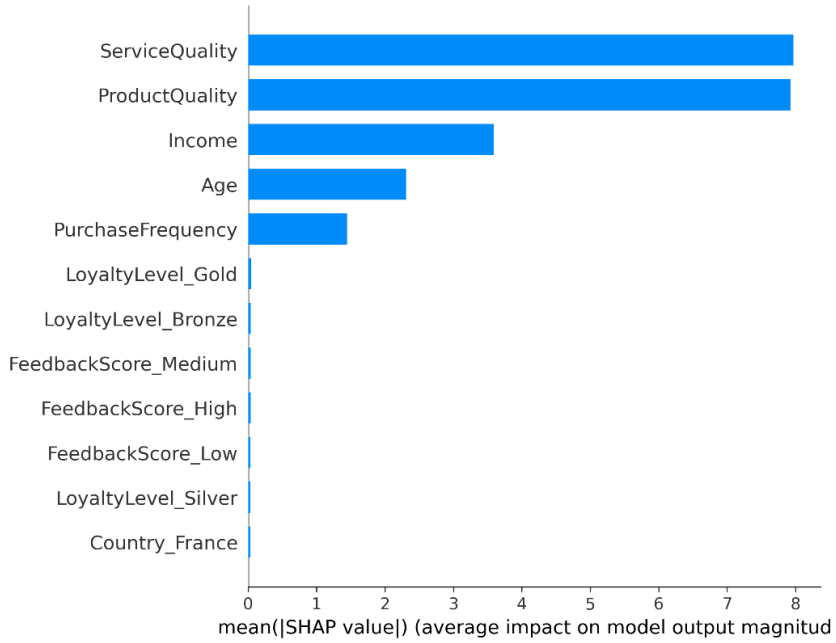
Table 4.

Random Forest permutation importance

SHAP Interpretability analysis

SHAP analysis was conducted to improve the interpretability of the Random Forest model. Unlike standard feature importance, SHAP values show both the magnitude and direction of each predictor's contribution to the predicted satisfaction score. The SHAP summary plot indicates that higher ServiceQuality and ProductQuality values increase predicted customer satisfaction, whereas lower values decrease the model output.





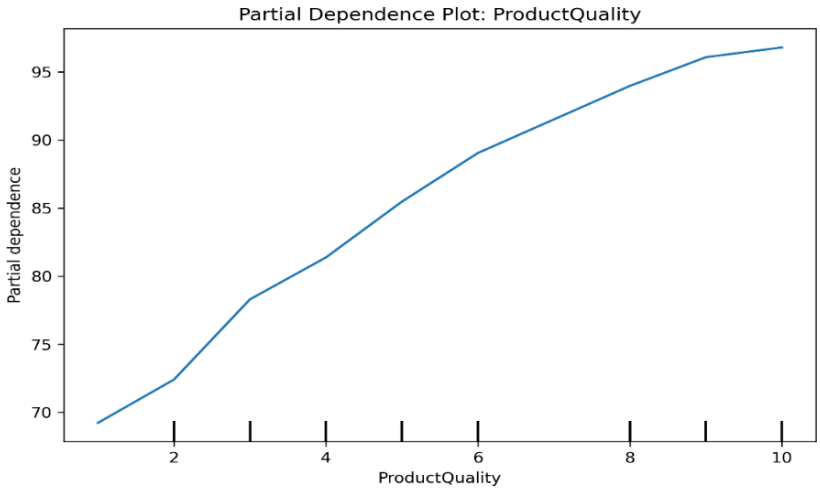
Rank	Variable	Mean |SHAP|
1	ServiceQuality	7.968
2	ProductQuality	7.929
3	Income	3.592
4	Age	2.310
5	PurchaseFrequency	1.445
6	Country	0.151
7	LoyaltyLevel	0.114
8	FeedbackScore	0.110
9	Gender	0.053

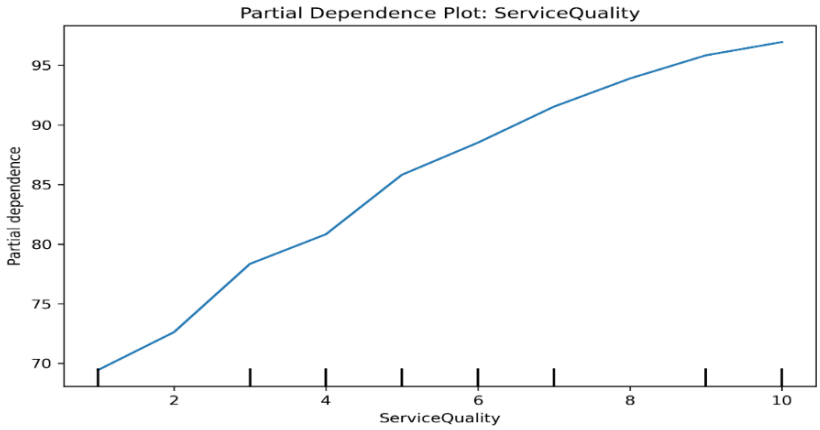
Table 5

Aggregated mean absolute SHAP values

Partial Dependence Interpretation

Partial dependence plots were produced for the two most influential predictors. The plots show a clear positive relationship: as ProductQuality and ServiceQuality increase, the predicted satisfaction score also increases. This finding supports the hypotheses that product and service quality are the main drivers of customer satisfaction.





The results demonstrate that Random Forest outperformed Linear Regression. In the hold-out validation, Linear Regression achieved R^2=0.697, MAE=7.38, and RMSE=9.22. In contrast, Random Forest achieved R^2=0.776, MAE=5.53, and RMSE=7.92. These results suggest that customer satisfaction is influenced by nonlinear patterns that are more effectively captured by ensemble learning methods.

The 5-fold cross-validation results further support this conclusion. Random Forest obtained a mean cross-validated R^2 of 0.781, compared with 0.700 for Linear Regression. The relatively small standard deviations across folds indicate that the predictive results are stable and generalizable within the synthetic dataset.

The interpretability analysis confirms the dominant role of service and product quality. Both impurity-based feature importance and permutation importance

identify ServiceQuality and ProductQuality as the strongest predictors. SHAP analysis provides additional evidence by demonstrating that higher values of these variables increase predicted satisfaction scores. Income, Age, and PurchaseFrequency contribute to the model, but their effects are considerably weaker. Overall, the results confirm that customer satisfaction is primarily shaped by perceived product and service quality, while demographic and behavioral factors play a secondary role.

Conclusion

This study demonstrates the effectiveness of machine learning techniques in predicting customer satisfaction using a synthetic customer feedback dataset comprising 38,444 observations. The findings consistently identify Service Quality and Product Quality as the strongest determinants of customer satisfaction, while Income exhibits a moderate effect and demographic variables contribute relatively less.

In addition to Multiple Linear Regression, a Random Forest model was employed to improve predictive performance. The results indicate that Random Forest outperformed the baseline regression model, achieving an R^2 value of 0.776 on the hold-out test set and a mean cross-validated R^2 of 0.781. Furthermore, Feature Importance, Permutation Importance, and SHAP analyses consistently confirmed the dominant role of Service Quality and Product Quality in shaping customer satisfaction.

From a managerial perspective, the findings suggest that organizations should prioritize service quality improvement, product quality enhancement, and customer experience management to increase customer satisfaction. Although the study is based on synthetic data, it demonstrates the practical value of machine learning and explainable artificial intelligence for customer analytics. Future research may validate the proposed framework using real customer data and additional machine learning algorithms.

The findings provide valuable insights for both researchers and practitioners and highlight the potential of combining machine learning and explainable AI techniques for customer satisfaction analytics.

References

[1] Chang L.A.: “Comparison of Three Methods to Generate Synthetic Datasets for Social Science”. Journal of Systemics, Cybernetics and Informatics, 2025 23(7), 39-44

[2] Dos Santos M.A., Fernández J.G., Fuentes-Solis R., Zarco C.: “Machine learning insights into recommendation intention: evidence from the fitness industry”. Journal of Big Data, 2026.

[3] Gonzales A., Guruswamy G., Smith S.R.: “Synthetic data in health care: A narrative review”. PLOS Digit Health 2(1): e0000082.

[4] Le, H. S., Huynh Do T., Nguyen M.H., Tran H., Pham T.T., Nguyen N.T., Nguyen V.: “Predictive model for customer satisfaction analytics in e-commerce reviews”. International Journal of Information Management Data Insights. Volume 4, Issue 2, 2024

[5] M. A. A. et al.: "Customer Satisfaction Prediction Using Machine Learning, “Towards Data Science”, 2025.

[6] Mozumder A.S., Nguyen T.N., Devi S. et al.: “Enhancing Customer Satisfaction Analysis Using Advanced Machine Learning Techniques in Fintech Industry”. Journal of Computer Science and Technology Studies, 2024, pp: 35-41

[7] Mustak M., Hallikainen H., Laukkanen T., Loïc Plé, Hollebeek L.D., Aleem M.: “Using machine learning to develop customer insights from user-generated content”. Journal of Retailing and Consumer Services, volume 81, 2024

[8] Rahman M.A., Modak C., Mozumder A.S., Miah M.N.I. et al.: “Advancements in Retail Price Optimization: Leveraging Machine Learning Models for Profitability and Competitiveness”. Journal of Business and Management Studies, 2024, pp: 103-110

[9] Rajotte, J. F., et al.: “Synthetic data as an enabler for machine learning applications”. iScience, Volume 25, Issue 11, 2022

[10] Sanchez-Serrano P., et al.: “A decision framework for privacy-preserving synthetic data generation”. Computers & Industrial Engineering, Volume 126, 2025

[11] Surendra H, Mohan H., “A Review of Synthetic Data Generation Methods for Privacy Preserving Data Publishing”. J Sci Technol Res. 2017;6(03):95–101.

[12] Zaghloul, M., Barakat S., Rezk A.: “Predicting e-commerce customer satisfaction: Traditional and machine learning approaches”. Journal of Retailing and Consumer Services, Volume 79, 2024