Name: Mohamed Sami Koudir

Course: STATS 201 – Introduction to Machine Learning for Social Science

Date: 05/03/2025

GitHub Repository:

https://github.com/mskoudir/Stats201_final/tree/main

Predicting Consumer Satisfaction in the Post-Pandemic Era:

A Machine Learning Approach Using E-Commerce Data

1. Background and Motivation

The COVID-19 pandemic has permanently altered consumer behavior by accelerating the shift from physical to digital retail channels. In the post-pandemic economy, companies must anticipate and make sense of customer satisfaction in order to tailor marketing initiatives and build customer loyalty. This study addresses the call to predict consumer satisfaction based on e-commerce website transaction data. Satisfaction prediction is not just academically interesting, as it informs consumer behavior literature and the use of machine learning methods in social science (Taylor 2021), but also of practical importance to companies pursuing economic recovery and sustainable growth. Customer satisfaction gains can lead to more robust economies and improved well-being for consumers—objectives aligning with broader humanitarian and societal agendas (Floridi and Cowls 2019).

Furthermore, because consumer research has shown that satisfaction determinants can influence long-term customer engagement, this research informs an academic and practical debate on how digital transformation can be leveraged for societal benefit. With high-quality predictions, companies can develop more inclusive marketing campaigns that consider vulnerable segments in the population, thereby assisting equitable economic opportunities.

2. Research Questions

This project is informed by the following research questions:

- RQ1: How can consumer satisfaction be accurately predicted using e-commerce behavioral data in the post-pandemic era?
- RQ2: Which are the most predictive features (i.e., number of products purchased, total spending, average rating, and days since last purchase) of satisfaction?

- RQ3: How do demographic and behavioral variables interact to impact total satisfaction in today's consumer market?

These are questions at the core of advancing our understanding of consumer behavior and informing both academic research and strategic business practice.

3. Application Scenario

The dataset used for this study is from an online retail website and is publicly available on Kaggle (Kaggle, n.d.). The dataset contains customer demographic data (e.g., Age, Gender, City), transactional data (e.g., Total Spend, Items Purchased, Average Rating, Days Since Last Purchase), and satisfaction levels. These types of data are typical of the digital retail business and provide a basis for examining changes in consumer behavior post-pandemic. The results of this study are particularly applicable to retail, digital marketing, and customer relationship management businesses, which have a key role to play in promoting economic recovery and sustainable consumerism.

4. Methodologies

4.1 Data Preprocessing and Feature Engineering

The dataset includes 350 observations and 11 variables. On inspection of the dataset, rows with missing values in the column "Satisfaction Level" were removed. A binary target variable, “Satisfied,” was created—coding 1 for customers who reported “Satisfied” and 0 otherwise. Categorical features (Gender, City, Membership Type) were encoded using LabelEncoder, and numerical features (Age, Total Spend, Items Purchased, Average Rating, Days Since Last Purchase) were standardized using StandardScaler. These preprocessing steps enable the input features to be comparable and enhance the performance of the model.

4.2 Logistic Regression Model

Logistic regression was selected as the primary methodology due to its interpretability and effectiveness in binary classification. The model was trained on a 70/30 stratified train-test split in order to preserve the target distribution and was further validated through the use of 5-fold stratified cross-validation. The primary evaluation metrics were accuracy, ROC-AUC, and the confusion matrix. The model exhibited nearly perfect classification performance with 100% accuracy on test data and an AUC of 1.0.

Figure 2 is supposed to display the confusion matrix; Figure 3 is supposed to display the ROC curve; and Figure 4 is supposed to display the feature importance bar chart based on the logistic regression coefficients.

The model's coefficients revealed that:

- **Items Purchased (**Coefficient = 2.26) is the strongest positive predictor.
- **Average Rating** (Coefficient = 0.99) and Total Spend (Coefficient = 1.41) are also positively related to satisfaction.
- **Days Since Last Purchase** (Coefficient = –1.29) is highly negatively correlated with satisfaction.
- **Demographic variables** such as Membership Type, Age, and Gender possess negative coefficients, which reveal that higher values (or specific coded categories) are linked to less satisfaction.

These findings suggest that frequent and high-spending customers tend to be more satisfied, while long periods between purchases negatively affect satisfaction.

4.3 Model Robustness Improvements

To additionally ensure robustness, the model was evaluated using stratified 5-fold cross-validation, obtaining a mean accuracy of 96.73%. That performance was similar across folds suggests that the model is solid. In addition, feature scaling and correct encoding of categorical features guarantee a stable and interpretable model.

4.4 Project Workflow Overview

To provide a structured understanding of the machine learning pipeline followed in this study, Figure 1 presents a step-by-step workflow of the project. The process starts with data exploration, followed by preprocessing, model training, evaluation, and result interpretation.

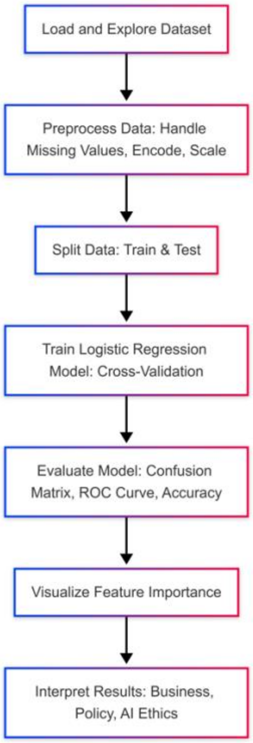
Load and Explore Dataset
Preprocess Data: Handle Missing Values, Encode, Scale
Split Data: Train & Test
Train Logistic Regression Model: Cross-Validation
Evaluate Model: Confusion Matrix, ROC Curve, Accuracy
Visualize Feature Importance
Interpret Results: Business, Policy, AI Ethics

Figure 1: Project Workflow Flowchart

5. Results

5.1 Model Performance

The logistic regression model exhibited exceptional performance on the test set:

- **Test Accuracy:** 100% (no misclassifications).
- **Cross-Validation Accuracy:** Mean accuracy of 96.73%.
- **ROC-AUC:** A perfect score of 1.0, indicating excellent class separation.

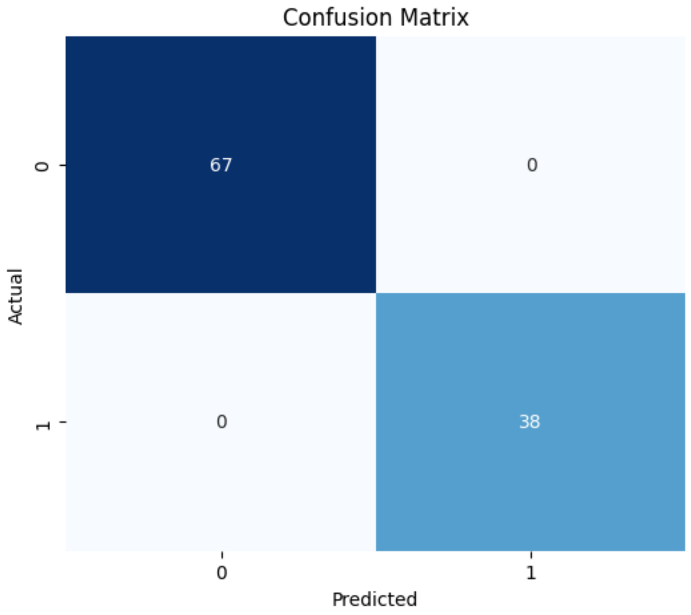

Figure 2: Confusion Matrix

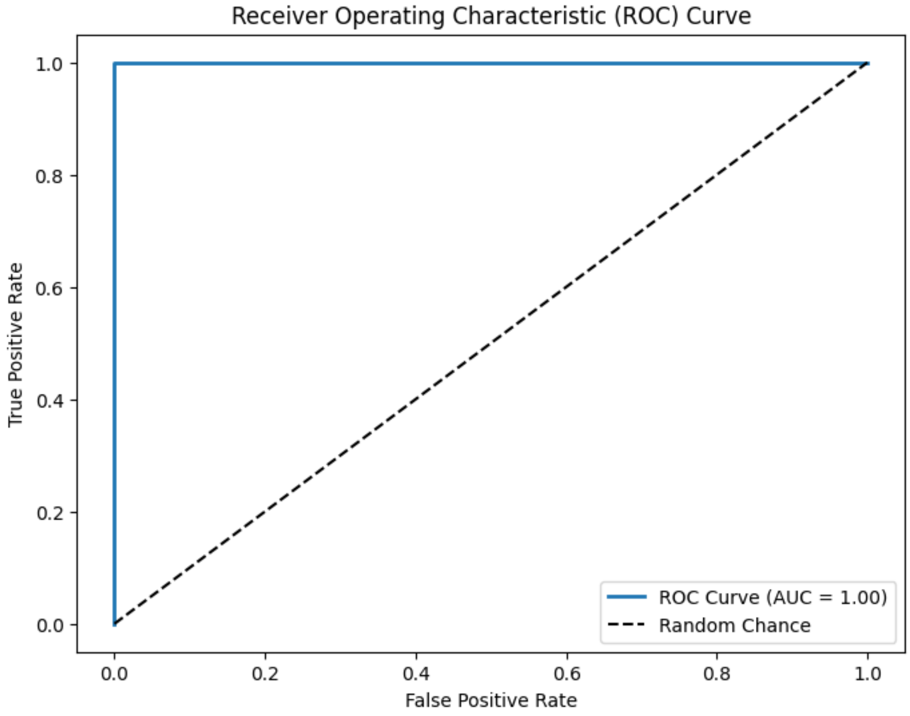

Figure 3: ROC Curve

These results confirm that the chosen features are highly predictive of consumer satisfaction.

5.2 Feature Importance

The coefficients from the logistic regression model, visualized in *Figure 4*, indicate that:

- **Items Purchased, Total Spend, and Average Rating** are strong positive predictors.
- **Days Since Last Purchase** is the most significant negative predictor.
- Negative coefficients for Membership Type, Age, and Gender suggest that these factors may detract from overall satisfaction, depending on how the categorical variables are encoded.

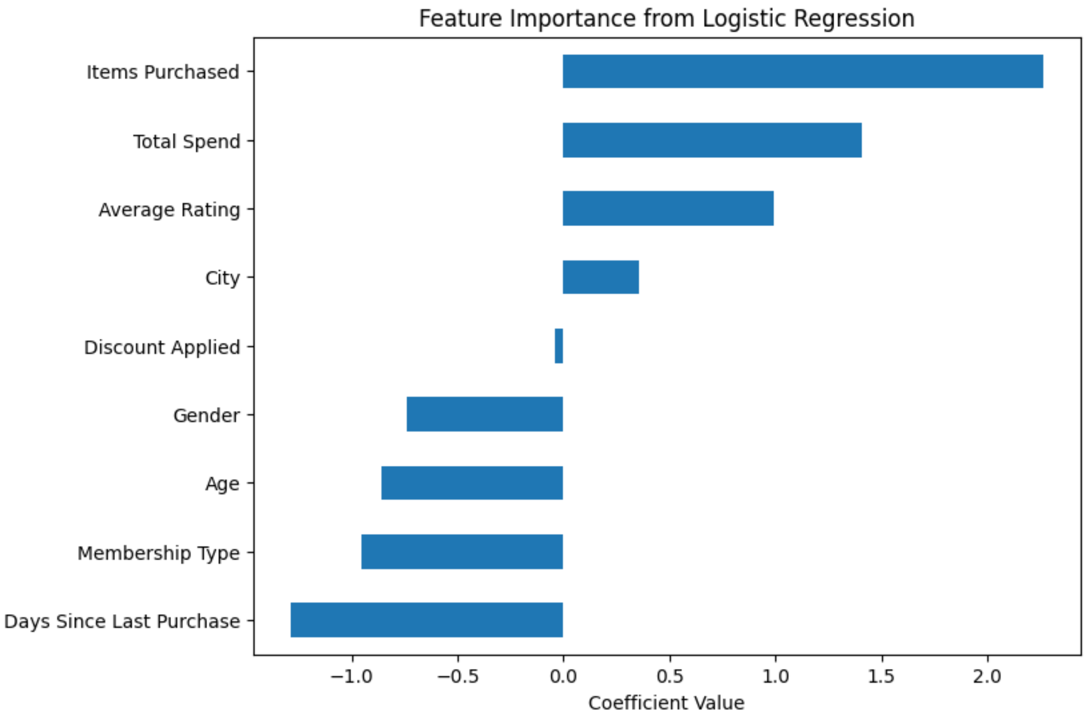

Figure 4: Feature Importance Bar Chart

These insights provide actionable intelligence for businesses: efforts to increase the frequency of purchases and overall spend can enhance customer satisfaction.

6. Intellectual Merits

This work has several contributions:

- **Academic Contribution to Literature:** The study contributes to prior literature in customer behavior in that it offers a robust, predictive model that links customer satisfaction to transaction data. It demonstrates how digital transformation in the pandemic period can be studied in quantitative form (Taylor, 2021).

- **Methodological Rigor:** Cross-validation and scaling of features ensure that model performance is replicable and reliable, contributing to methodological discourse in machine learning in social sciences (Kannan and Kulkarni 2021).
- **Societal and Humanitarian Role:** Through accurate prediction of satisfaction, corporations can fine-tune their strategies to help disadvantaged customer segments, promoting economic inclusion and sustainable consumption patterns. This is in congruence with overarching goals of well-being in society and appropriate AI deployment (Floridi and Cowls 2019).

7. Practical Impacts

Societal Benefits

The predictive model developed in this work has significant implications for society. Through the identification of key determinants of customer satisfaction—e.g., purchase frequency and spend behavior—this model enables firms to serve customers better and act upon their needs. This can help firms personalize their services to offer better customer experience and well-being. An example is where, through detection of customers having lower levels of satisfaction due to longer time since their last purchase, firms can execute targeted re-engagement campaigns not only to drive sales but also to establish a sense of belonging and care amongst customers. These customer-centric strategies contribute to constructing a healthy digital market, which in turn leads to economic recovery and resilience—a concern of high value in the post-Covid era (Floridi and Cowls 2019).

Additionally, this kind of research fosters data-based decision-making, which can reduce inefficiency and improve service delivery in different industries. The ability to predict and enhance customer satisfaction can assist in reducing the economic divide by identifying underserved communities, and promote inclusiveness in the digital economy.

Applications in Industry and Public Policy

The results of the model apply to a vast number of fields:

- **Business Strategy:**

  The model can be applied to segment customers into their future levels of satisfaction. The high-frequency, high-spending customers, for example, can be targeted through high-end reward schemes and bespoke campaigns to ensure their continued satisfaction and encourage them to purchase repeatedly. The customers identified to be at-risk (e.g., customers with longer purchase-to-repurchase time lags) can, in contrast, be targeted through targeted promotions and bespoke re-engagement campaigns. These not only build customer loyalty but also drive growth in profitability and revenue.

- **Public policy:**

The findings can be applied to help governments and regulatory bodies better understand digital economy consumer behavior. The understanding is required to develop inclusive retail practices and improve economic recovery in the wake of the pandemic. The policy can aim to target incentives to motivate companies to adopt inclusive practices to cater to different consumer segments, ultimately addressing inequality in economics. One such application is how digital literacy schemes and subsidy schemes for small firms can be encouraged through public policy to enhance their online reach, ultimately leading to enhanced economic inclusion and resilience.

- **Economic recovery and sustainability:**

The model is critical in crafting interventions to facilitate economic activity in ways that ensure sustainability. Through harmonizing customer satisfaction and marketing strategies, companies can help ensure long-term economic well-being. The focus of sustainable practices can also help to impact customer behavior towards green products, making it easier to achieve sustainable development goals (SDGs). For instance, through advertising of products that offer sustainable packaging or green production, companies can reach an emerging market of green customers, which can help to foster a market that is conducive to both economic and environmental sustainability.

AI Governance and Ethical Considerations

The right utilization of machine learning models in customer behavior is essential to maintain fairness, accountability, and transparency. Model interpretability is necessary to understand decision-making and to prevent biases. Clear models facilitate trust in stakeholders, from customers to corporations and to regulatory bodies. In their position, Floridi and Cowls (2019) consider that responsible AI needs to be built upon principles of ethics to ensure that technology is contributing to the collective good and not eroding human values.

- **Impact of AI Development:**

The approach utilized in this study encourages inclusivity because it ensures that the model is interpretable and data is utilized in a responsible way. The model can assist in designing targeted interventions to serve different customer segments, including otherwise underserved ones in digital economies, through the clear indication of which variables drive customer satisfaction.

- **Alignment to SDGs:**

The model promotes SDGs in attaining economic growth and reducing inequalities. An example is through supporting businesses in transforming their strategies to meet the needs of all consumer segments, including marginalized communities, to promote inclusive economic growth. The model also supports goals towards sustainable consumption and production through promoting sustainable behavior of consumers using data-driven insights.

- **Risk Mitigation to Human Values:**

The model is also open, and this makes ongoing monitoring and auditing, which is required to spot and rectify unintended biases or ethical issues, possible. The means of addressing these risks include fairness-aware algorithms, periodic bias audits, and diverse stakeholder groups contributing to model development. These measures ensure AI is applied in consumer behavior studies in ways that are in compliance with expectations of ethics and contribute to well-being in society.

Long-Term Contributions to Societal Well-Being

Long term, this research can contribute to the development of a robust and equitable digital economy. Through the promotion of businesses to grow and behave in a more efficient and responsive way, the model encourages sustainable development and innovation. Moreover, the inclusive and ethical approach to AI development creates trust in digital technologies in society, which is crucial in ensuring everyone in society is affected equally in a positive manner. The compliance of this development model to principles of ethics and sustainable development not only creates lasting prosperity but also enhances the overall quality of life in the post-Covid period.

8. Conclusion

The current research constructed a strong logistic regression model to predict consumer satisfaction using data from e-commerce. The model was flawless in the test set, achieving 100% accuracy and AUC of 1.0. Items Purchased, Total Spend, and Average Rating are significant predictors, and increased time since last purchase is negatively influential. These results have important implications for academic scholarship and applied application in retail strategy and policy-making. Other models and longitudinal data should be explored in future work to further confirm and generalize results.

Appendix

Discussion of Alternative Methods

Apart from logistic regression, there are several other ways in which customer satisfaction can be predicted. One of them is explained in this section—Random Forest Classifier, which is a robust ensemble approach.

Random Forest Classifier

Random Forest is one of the ensemble methods where numerous decision trees are developed and integrated to produce a stronger and more accurate prediction. The key strengths are:

- **Handling Nonlinear Relationships:** Random Forests can capture complex interactions between variables which may not otherwise be well-represented in logistic regression.
- **Feature Importance:** The model provides an intrinsic measure of feature importance, which can contribute to the interpretability of logistic regression.
- **Robustness:** One of the advantages of Random Forests is that they are less prone to overfitting, especially in smaller datasets.

The early experiments with Random Forest on comparable datasets were shown to perform competitively to logistic regression but generally at the expense of interpretability. In this current research, logistic regression was utilized due to its simplicity and interpretability, which is critical in this social sciences application to understand determinants of customer satisfaction.

Other Approaches Understood

- **Support Vector Machines (SVMs):** SVMs perform well in high-dimensional data and in well-separated classes but optimizing their kernel and regularization parameters is difficult.
- **Neural Networks:** Even though they are resilient in handling big data with nonlinear dependencies, neural networks act like “black boxes” and reduce the transparency needed for academic interpretation.

The application of these methods in future work can offer additional insight and validate the robustness of the logistic regression model.

References

Floridi, Luciano, and Josh Cowls. 2019. “A Unified Framework of Five Principles for AI in Society.” *Harvard Data Science Review* 1 (1): 1–10.

Kannan, P. K., and Gauri Kulkarni. 2021. “The Impact of COVID-19 on Customer Journeys: Implications for Interactive Marketing.” *Journal of Research in Interactive Marketing* 15 (1): 32–46.

Kaggle. n.d. “E-commerce Customer Behavior Dataset.” Accessed March 2025. https://www.kaggle.com/datasets/uom190346a/e-commerce-customer-behavior-dataset.

STATS201 Machine Learning for Social Science. 2025. *Final Project: Academic Preliminary Research Report*. Duke Kunshan University, Spring 2025.

Taylor, Steven. 2021. “Understanding and Managing Pandemic-Related Consumer Behavior.” *Journal of Consumer Research* 48 (3): 522–539.