International Journal of Artificial Intelligence, Data Science, and Machine Learning
Grace Horizon Publication | Volume 7, Issue 1, 77-81, 2026
ISSN: 3050-9262 | https://doi.org/10.63282/3050-9262.IJAIDSML-V7I1P115

Original Article

Framework for Predicting Customer Channel Preference using Machine Learning

Vaibhav Tummalapalli
Atlanta, USA.

Received On: 18/12/2025 ***Revised On: 20/01/2026*** ***Accepted On: 25/01/2026*** ***Published On: 31/01/2026***

Abstract - *Understanding customer preferences across communication channels is critical for optimizing marketing strategies in the automotive industry. This paper introduces a machine learning framework that predicts customer engagement rates for various channels, including email, SMS, and social media. By linking historical engagement data with future behaviors using observation and performance windows, the framework enables precise channel assignment based on predictive scores. Applications in vehicle purchase campaigns and after-sales promotions highlight the framework's potential to improve marketing efficiency and customer satisfaction. Challenges like data sparsity and interpretability are discussed, along with proposed mitigation strategies*

Keywords - *Machine Learning, Channel Propensity, Optimization, Cohort Structure, Regression, Classification & Regression Metrics, Marketing Analytics.*

1. Introduction

In the automotive industry, optimizing marketing efforts requires a clear understanding of customer engagement across multiple communication channels. Customers interact with businesses through various platforms, including email, SMS, and social media, making it essential to identify their preferred channels. Managing diverse communication channels presents strategic and operational challenges [1], necessitating data-driven frameworks for optimized targeting. As the industry evolves toward omni-channel engagement, predictive frameworks become essential to personalize outreach across touchpoints [2]. This paper presents a machine learning framework for predicting channel preferences by analyzing historical engagement data, transactional behaviors, and demographic-psychographic attributes. The proposed approach structures data into observation and performance windows to link past behaviors with future engagement rates systematically.

By building separate predictive models for each channel and standardizing scores, businesses can allocate marketing resources to maximize ROI and enhance customer satisfaction. Applications of this framework include vehicle purchase campaigns, aftersales promotions, and customer retention strategies, making it a versatile tool for automotive marketing. The paper also addresses key challenges, such as data sparsity and model interpretability, providing actionable solutions to ensure a robust outcome

2. Data Set Up & Structure

Following prior work on channel migration [3], we structure customer history and outcomes using observation and performance windows to model evolving preferences [6]. To accurately predict channel preferences, the data structure is designed to capture the nuances of customer behavior over multiple time periods (cohorts). For each cohort, the data is divided into two key components.

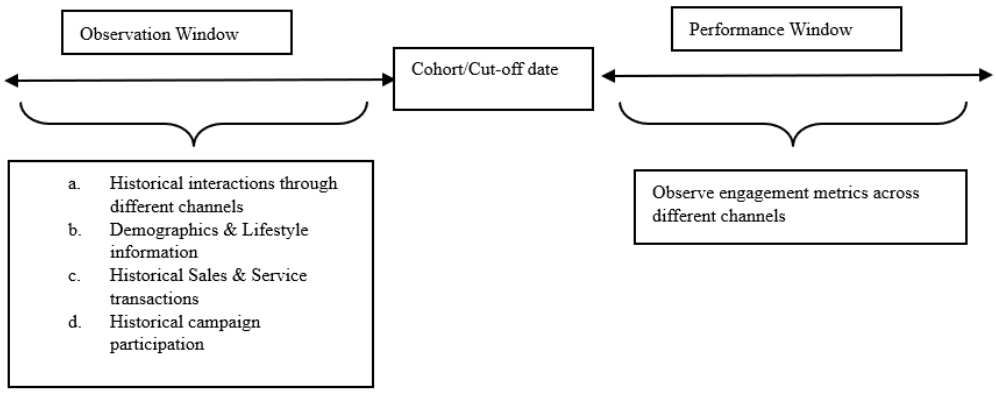

Fig 1: Data Structure/Set Up

2.1. Observation Window

Aggregates all historical customer information to serve as predictors for modeling. This includes:

- Communications: Total interactions across channels, email open rates, SMS response rates, and participation in past campaigns.
- Transactions: Number and types of vehicle purchases, after-sales services, or other relevant transactions.
- Demographics: Key attributes such as age, gender, income, and geographic location.
- Psychographics: Insights into customer lifestyle segments, preferences, and attitudes toward brand or automotive products.

2.2. Performance Window

Measures engagement metrics for each channel, which are used as target variables for the subsequent performance period.

Engagement metrics are critical for assessing how well customers interact with each channel, including:

- Email: Open rates, click-through rates (CTR), number of forwards/shares.
- SMS: Response rates, link clicks, and opt-out rates.
- Social media: Likes, shares, comments, click-through rates.
- Other Channels (e.g., Over-the-Air Updates): App downloads, time spent in apps, and service bookings.

2.3. Engagement Metrics: Selection and Prioritization

Each channel may have multiple engagement metrics, and their selection depends on business objectives and what is deemed critical by stakeholders. For example:

- If the goal is to maximize awareness, email open rates or impressions on social media may be prioritized.
- For direct driving, metrics like click-through rates and conversion rates may be more relevant.
- In aftersales, response rates to SMS campaigns offering discounts for service appointments might be the focus.

2.4. Determining the Most Critical Metrics

To determine the most critical metrics for modeling, consider:

- Business Objectives: Align metrics with campaign goals. For example, if the focus is vehicle purchases, prioritize conversion-related metrics like click-through or response rates.
- Historical Analysis: Examine past campaign performance to identify metrics most correlated with business outcomes (e.g., sales, service bookings).
- Data Availability: Ensure that the selected metrics are consistently available and accurately recorded across time periods and channels.

By tailoring the engagement metrics to the business objectives and carefully selecting the most relevant ones, the model can provide actionable insights. The flexibility of this framework ensures that businesses can adapt their approach depending on their strategic focus, whether it's increasing awareness, boosting conversions, or enhancing customer engagement.

3. Modeling Framework

3.1. Target Variable

The framework begins by defining the **target variable**, which is the engagement metric for a specific channel observed in the performance window. This metric, a continuous variable (e.g., email open rate, SMS click-through rate, or social media interactions), captures customer activity and serves as the basis for model predictions.

3.2. Separate Models for Each Channel

Separate predictive models are built for each channel to account for distinct engagement patterns across different communication mediums. These models leverage aggregated features from the observation window as predictors for the performance window. Predictors include historical interactions (e.g., the number of SMS sent, social media ads viewed), transactional data (e.g., service frequency, purchase history), and customer demographics and psychographics.

3.3. Implementation and Customizing Strategy

After prediction, engagement scores for each channel are standardized to a common scale (e.g., z-scores or min-max scaling). This ensures that scores are comparable across all channels. Once standardized, the channel with the highest engagement score is assigned to the customer. For businesses requiring more flexibility, strategies can include selecting the top channel or using a combination of the top 2–3 channels if their scores are close. These customizations allow the framework to adapt to specific marketing objectives and resource allocations.

4. Data Preparation

4.1. Aggregate Historical Data

The foundation of the framework begins with consolidating a diverse range of historical data sources into a unified dataset for each customer. This includes communication data (e.g., email opens, SMS clicks, and social media interactions), transaction data (e.g., service visit frequency, vehicle purchase history), and rich demographic and psychographic data [7]. The demographic and lifestyle dataset provides a 360-degree view of each prospect, with over 1,500 attributes per individual, spanning critical dimensions such as:

- Demographics: Age, household composition, marital status, ethnic background, and residential information.
- Lifestyle Factors: Interests in areas like books, gardening, travel, sports, and entertainment, providing insights into preferences and hobbies.
- Financial Health: Indicators of income, credit, debt levels, and assets, offering an understanding of purchasing power and financial stability.

- Market Activity: Transaction behaviors, spending trends, and retail engagement, reflecting consumer patterns and preferences.
- Automotive Data: Vehicle ownership details such as brand, model, and age of vehicles, which are especially pertinent for automotive propensity modeling.

Capturing psychographic and experiential data has been shown to enhance engagement modeling [4], particularly in long customer journeys like automotive sales. This comprehensive aggregation enables the creation of robust customer profiles, enhancing segmentation and providing predictive power for identifying channel preferences and engagement patterns.

4.2. Clean and Transform Data

Once the data is aggregated, it undergoes rigorous cleaning and transformation to ensure accuracy and usability for modeling. The key steps include:

- Data Partition: Split the data into training and validation sets to evaluate model performance. Use a typical split ratio such as 70-30 or 80-20, ensuring that both sets are representative of the overall population [12].
- Outlier Detection & Treatment: Detect and handle extreme values in numeric features based on their distribution. If the column has a Gaussian Distribution, use the standard deviation method (e.g., cap and floor at ±3σ) and if the distribution is non-Gaussian apply Interquartile Range (IQR) or Median Absolute Deviation (MAD) methods to cap and floor outliers. For fields with known limits (e.g., payment values), use predefined thresholds or replace outliers with business-defined values [10].
- Handle Missing Values: Choose the imputation technique based on the percentage and pattern of missing data [8] [9]:
- Low Missing Rates (e.g., <10%): Use mean, median, or mode imputation.
- High Missing Rates: Consider advanced techniques such as:
- KNN Imputation: Replaces missing values using the nearest neighbors in the feature space.
- Multivariate Imputation by Chained Equations (MICE): Models missing values based on other variables.
- Create Missing Indicators: For certain variables, a binary flag indicating whether data is missing may add predictive value.
- Encode Categorical Variables: Use appropriate encoding strategies based on the variable type [11]
- One-Hot Encoding: For categorical variables with a small number of distinct levels.
- Ordinal Encoding: For variables with a natural order, such as low, medium, and high.
- High-Cardinality Variables: Club low-frequency categories together based on domain knowledge or create an "Other" category to reduce sparsity.
- Transform and Normalize Features:Apply mathematical transformations (e.g., log, square root) to normalize skewed numeric variables.
- Standardize variables to have a mean of 0 and a standard deviation of 1, especially for models like Logistic Regression or SVM that are sensitive to scaling.
- For binning or discretization, use techniques such as optimal binning based on Chi-Square values to improve interpretability and model performance [11].

By combining the richness of aggregated historical data with meticulous data preparation, this framework ensures the creation of clean, structured, and insightful datasets. These datasets not only enhance the accuracy of predictive models but also ensure that the derived insights are actionable and aligned with business goals.

5. Model Development & Evaluation

5.1. Model Training

To predict engagement metrics across different communication channels, the framework employs separate regression models tailored to each channel. This ensures that each model captures unique engagement patterns specific to the channel, such as email open rates, SMS click-through rates, or social media interactions. The steps include:

Regression Models:

- Algorithms like Gradient Boosting Machines (GBMs, XG Boost, CAT Boost), Random Forests, and Linear Regression are employed to predict engagement metrics.
- GBMs and Random Forests are particularly effective for handling non-linear relationships and interactions among features, making them suitable for complex datasets like those used in this framework.
- For each channel, historical engagement metrics, transactional data, demographics, and psychographics serve as predictors, while the engagement metric (e.g., email open rate) in the performance window acts as the target variable.

5.2. Evaluation

Once the models are trained, their performance is evaluated using standard regression metrics to ensure they effectively predict the engagement metric. These metrics include:

- **Root Mean Square Error (RMSE)**:
  - Measures the average magnitude of error between predicted and actual engagement metrics.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(\widehat{y_i}-y_i)^2}{n}}$$

  Where $\widehat{y_i}$ is the predicted value for observation i and y_i is the actual value for observation i.
  - Lower RMSE values indicate better model performance. RMSE penalizes larger errors

more heavily, making it suitable for use cases where large deviations are critical.

- **Mean Absolute Error (MAE)**:
  - Measures the average magnitude of errors without considering their direction.

$$\text{MAE} = \frac{\sum_{i=1}^{n}|\hat{y}_i - y_i|}{n}$$

MAE provides an intuitive measure of the average prediction error, making it easier to interpret.

- **Coefficient of Determination (R^2)**:
  - Evaluates how well the model explains the variance in the target variable.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

R^2 Squared ranges from 0 to 1, with values closer to 1 indicating that the model explains most of the variance in the target variable.

- **Channel-Specific Evaluation**
  - For each channel-specific model, evaluate RMSE and MAE to measure the accuracy of predicted engagement metrics [5].
  - Use R^2 to determine how well the model captures the variability in engagement for that channel.
  - Compare metrics across training and validation datasets to ensure the model generalizes well and is not overfitting.

By employing these evaluation metrics, businesses can identify the best-performing models for each channel and ensure they provide reliable predictions. Poorly performing models can be revisited to refine features, algorithms, or hyperparameters to enhance their accuracy and robustness.

6. Score integration & Implementation

Each channel-specific model generates predicted engagement scores, such as open rates for email, click-through rates for SMS, or interaction rates for social media. Since these scores are inherently derived from separate models, they may have different ranges and distributions. Standardization ensures comparability across all channels. Common techniques include:

6.1. Min-Max Scaling:

- Rescale scores to a fixed range (e.g., 0 to 1)

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

Where X is the original score
$X_{\min}$: Minimum score for the channel
$X_{\max}$: Maximum score for the channel
X' : Rescaled score for the channel

Ensures that all scores fall within the same range, making them directly comparable.

6.2. Z-Score Standardization:

- Convert scores to standard deviations from the mean.

$$Z = \frac{X - \mu}{\sigma}$$

Where X is the original score

- μ : Mean score for the channel
- σ : Standard deviation of score for the channel
- Z: Standardized score for the channel

Useful when the score distributions differ significantly across channels.

6.3. Decision Making

After standardizing the scores, customers are assigned to channels based on their predicted engagement potential. The assignment can be customized based on business objectives:

- Top-Performing Channel:
  - Assign the channel with the highest standardized score to each customer.
  - This strategy is ideal for campaigns with limited budgets or a single-channel focus.
- Multi-Channel Allocation:
  - Identify the top 2–3 channels with the highest scores and include these in the campaign strategy.
  - Example: If email, SMS, and social media scores for a customer are closely ranked, including all three channels to maximize reach.
- Threshold-Based Assignment:
  - Assign channels only if the standardized score exceeds a certain threshold.
  - Example: If the email score is above 0.7 (on a scale of 0 to 1), the customer is included in the email campaign; otherwise, they are excluded.
- Score Difference Analysis:
  - If the difference between the top two channel scores is small, allocate both channels.
  - If one channel's score is significantly higher than others, prioritize it as the primary communication medium.

6.4. Applications

This framework supports personalized marketing campaigns in the automotive industry, such as:

- Vehicle Purchase Campaigns: Predict which channel customers are most likely to engage with when planning their next vehicle purchase.
- Aftersales Engagement: Identify the best channel to promote services like maintenance packages or extended warranties.
- Retention Campaigns: Use preferred channels to re-engage customers at risk of churn.

7. Conclusion

This framework offers a systematic approach to predicting customer channel preferences using machine learning. By leveraging historical and psychographic data, it ensures personalized, data-driven marketing strategies that optimize engagement and resource allocation. Future

enhancements could include real-time engagement data and advanced neural network architecture for improved predictions.

References

[1] S. A. Neslin, D. Grewal, R. A. Leghorn, V. Shankar, M. L. Teerling, J. S. Thomas, and P. C. Verhoef, "Challenges and opportunities in multichannel customer management," *Journal of Service Research*, vol. 9, no. 2, pp. 95–112, 2006.

[2] P. C. Verhoef, P. K. Kannan, and J. J. Inman, "From multi-channel retailing to omni-channel retailing: Introduction to the special issue on multi-channel retailing," *Journal of Retailing*, vol. 91, no. 2, pp. 174–181, 2015.

[3] A. Ansari, C. F. Mela, and S. A. Neslin, "Customer channel migration," *Journal of Marketing Research*, vol. 45, no. 1, pp. 60–76, 2008.

[4] K. N. Lemon and P. C. Verhoef, "Understanding customer experience throughout the customer journey," *Journal of Marketing*, vol. 80, no. 6, pp. 69–96, 2016.

[5] R. T. Rust and M. H. Huang, *Handbook of Service Marketing Research*, Edward Elgar Publishing, 2014.

[6] Tummalapalli Vaibhav. (2025). Stratified sampling in Cohort-based data for Machine learning Model development. International Scientific Journal of Engineering and Management. 04. 1-8. 10.55041/ISJEM03377

[7] V. Tummalapalli, "Feature Engineering for Building Machine Learning Models in Automotive Industry," International Scientific Journal of Engineering and Management, vol. 4, no. 8, pp. 1–9, 2025. doi: 10.55041/ISJEM04985

[8] V. Tummalapalli, "Comprehensive study of data imputation techniques for machine learning models," International Journal of Innovative Research in Engineering & Multidisciplinary Physical Sciences, vol. 13, no. 4, 2025, doi: 10.37082/IJIRMPS.v13.i4.232674

[9] V. Tummalapalli, "Understanding distance metrics in KNN imputation: Theoretical insights and applications," Journal of Mathematical & Computer Applications, vol. 4, no. 4, pp. 1–4, 2025. doi: 10.47363/JMCA/2025(4)208

[10] Vaibhav Tummalapalli. (2025). Outlier Detection & Treatment for Machine Learning Models. International Journal of Innovative Research and Creative Technology, 11(3), 1–8. https://doi.org/10.5281/zenodo.16500050

[11] V. Tummalapalli, "Machine learning pipeline for automotive propensity models," International Journal of Core Engineering & Management, vol. 8, no. 3, 2025, ISSN 2348-9510.