Magazine Subscription Behavior Analysis Report:

Comparative Study of Logistic Regression and Support Vector Machine Models

Northeastern University

Master of Professional Studies in Data Analytics

Prepared by: Daniel Makala Mackaemba

Date: January 26, 2026

1. Introduction

The magazine publishing industry faces ongoing challenges related to customer acquisition and declining subscription rates. As consumer behavior continues to shift, data-driven approaches are essential for understanding which customers are most likely to respond positively to marketing efforts. This analysis seeks to (1) prepare high-quality data for modeling, (2) compare two classification models for predicting subscription behavior, and (3) generate insights that support improved marketing decision-making.

The dataset includes customer demographics, purchasing behavior, and campaign interaction history, with a binary response variable indicating whether a customer subscribed.

2. Data Preparation and Exploration

To ensure reliable modeling results, the dataset underwent systematic cleaning and preparation. Duplicate records were removed to prevent bias. Missing numeric values were imputed using median values to reduce the influence of extreme observations, while missing categorical values were filled using the most frequent category.

Categorical variables were transformed using one-hot encoding, and feature scaling was applied to standardize numeric variables. Exploratory analysis revealed class imbalance between subscribers and non-subscribers, motivating the use of stratified sampling during the train-test split. Feature distributions and correlations further justified the need for standardization prior to model training.

3. Methodology

Two classification models were selected for comparison. Logistic Regression was chosen for its interpretability and ability to quantify the directional impact of predictors. An SVM with

a radial basis function (RBF) kernel was included to assess whether non-linear decision boundaries could improve predictive performance.

The data were split into training (70%) and testing (30%) sets using stratified sampling. Standard Scaler was applied to ensure consistent feature scaling. Model performance was evaluated using accuracy, precision, and recall, providing a balanced assessment of overall correctness and class-specific prediction quality.

4. Results and Discussion

Both models achieved high accuracy; however, Logistic Regression outperformed SVM in precision and recall, indicating a stronger ability to correctly identify customers likely to subscribe. Confusion matrix analysis showed that Logistic Regression produced fewer false negatives, which is especially important from a business perspective, as missed subscribers represent lost revenue opportunities.

Logistic Regression coefficients also revealed the most influential features associated with subscription behavior, enabling clear interpretation and practical marketing insights. The comparatively weaker recall of the SVM model suggests that its complexity did not translate into improved performance for this dataset.

5. Conclusions and Recommendations

This analysis demonstrates that predictive modeling can effectively support marketing decision-making in subscription-based businesses. While both models performed reasonably well, Logistic Regression is recommended for operational use due to its superior recall, interpretability, and lower complexity.

References

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic Regression (3rd ed.). Wiley.
Foundational reference for logistic regression theory, interpretation, and practical application. Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297.

— Seminal paper introducing Support Vector Machines and kernel-based classification.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An Introduction to Statistical Learning: With Applications in Python (2nd ed.). Springer.

— Practical guidance on preprocessing, feature encoding, and improving predictive performance.