Understanding Magazine Subscription Behavior: A Predictive Modeling Approach

Authors: Pinkesh Tandel and Dennis Darko

Northeastern University
College of Professional Studies
Vancouver, BC, Canada
Course: ALY6020: Predictive Analytics
September 29, 2024
Under the guidance of Dr. Nina Rajabi Nasab

Abstract

This study addresses the challenge of declining magazine subscriptions, which defied expectations during a period where people were spending more time at home. Using customer data, we employ logistic regression and support vector machine (SVM) models to predict subscription behavior and identify significant factors influencing subscriptions. Fine-tuning both models improves their performance, and we compare their effectiveness in accurately predicting subscriptions. Based on model evaluation, we provide a final recommendation on which model best serves the company's objectives.

Introduction

The magazine company experienced a decline in subscriptions despite a global shift towards home-based activities. The company believed that increased time at home would lead to more time spent reading, thus expecting a rise in subscriptions. However, this was not the case. This paper aims to analyze and predict customer subscription behavior using machine learning models to identify what factors are driving or inhibiting subscription renewals. By applying logistic regression and support vector machines (SVM), we explore these patterns and recommend a model that can guide the company's future strategies for customer retention.

Data Cleansing and Feature Engineering

Data Overview

The dataset provided by the magazine company contained various demographic and customer engagement features, including age, income, marital status, and past purchases. To improve the quality of the data, we applied multiple data cleansing techniques. Missing values in the `Income` field were replaced with the median income, and outliers were capped at the 99th percentile to minimize the impact of extreme values. Furthermore, a new feature, `Customer_Tenure`, was created to capture the length of a customer's relationship with the company. This feature provided insights into customer loyalty and its effect on subscription renewals.

Encoding and Transformation

Categorical variables such as `Education` and `Marital_Status` were one-hot encoded to make them compatible with machine learning algorithms. This process ensured that these variables were transformed into binary features, thus allowing the models to capture the effect of each category (e.g., education level, marital status) on subscription behavior.

Multicollinearity Analysis

Multicollinearity occurs when two or more independent variables are highly correlated, leading to unstable model coefficients. To detect and address multicollinearity, we calculated the Variance Inflation Factor (VIF) for each feature. Variables with high VIF scores, such as `Year_Birth` and `Income`, were removed to reduce redundancy. For example, `Year_Birth` was replaced by the more meaningful `Customer_Tenure`, which provided a direct measure of customer engagement. By addressing multicollinearity, we improved the interpretability and stability of the models.

Logistic Regression Model

Logistic regression is a well-established method for binary classification tasks, making it suitable for predicting whether a customer will subscribe or not. The model was trained on the cleaned and preprocessed dataset, with the response variable representing subscription behavior (`1` for subscribed, `0` for not subscribed).

Key Findings

- **Customer_Tenure**: The model identified that customers who have been with the company longer are more likely to subscribe. This underscores the importance of fostering long-term relationships with customers.
- **Past Purchases**: Spending on specific products like wine and meat products was positively correlated with subscription behavior, suggesting that customers who make higher-value purchases are more engaged.
- **Marital Status**: Certain marital statuses (e.g., married) were also found to influence subscription behavior positively.

Model Performance

The logistic regression model achieved an accuracy of 88%, with a precision of 0.71 and recall of 0.36 for predicting subscriptions. Although it performed well overall, it struggled with recall for the subscription class, meaning it failed to identify all true subscribers.

Support Vector Machine (SVM) Model

SVM is another robust classification technique, especially useful for datasets with complex decision boundaries. We applied a linear kernel SVM to the same dataset to compare its performance with logistic regression.

Key Findings

The SVM model identified similar patterns to the logistic regression model, with **Customer_Tenure** and **Product Purchases** being the most important features. However, the SVM model is less interpretable than logistic regression in terms of feature importance due to the nature of the decision boundary used in SVM.

Model Performance

The SVM model achieved an accuracy of 87%, with a precision of 0.73 and recall of 0.28 for predicting subscriptions. Like logistic regression, it struggled with recall for the subscription class, missing several true positives.

Fine-Tuning and Model Improvement

To further improve both models, we applied hyperparameter tuning using grid search. For logistic regression, we adjusted the regularization parameter `C` and tested both `l1` and `l2` penalties. For SVM, we fine-tuned the `C` parameter, kernel choice, and `gamma` value.

Improvement Results

- **Logistic Regression**: After fine-tuning, the F1-score for predicting subscribers improved, with more balanced precision and recall.
- **SVM**: The tuned SVM model also saw improvements in precision and recall, though it still lagged behind logistic regression in terms of recall for the subscription class.

Model Comparison and Final Recommendation

After evaluating both models, we compared their performance across key metrics:

- **Logistic Regression**: Accuracy (88%), Precision (0.71), Recall (0.36)
- **SVM**: Accuracy (87%), Precision (0.73), Recall (0.28)

While both models performed similarly in terms of accuracy, logistic regression provided a better balance between precision and recall. Given the company's goal of understanding and improving subscription behavior, **logistic regression** is the recommended model due to its interpretability and its ability to identify subscribers more effectively than SVM.

Conclusion

By applying logistic regression and SVM models to predict subscription behavior, we gained insights into the factors driving subscriptions. The logistic regression model, in particular, provided useful interpretability and better recall, making it the recommended choice for the company. The company can leverage these findings to focus on customer retention strategies, particularly targeting long-term customers and those with high engagement in specific product categories.

References

Field, A. (2017). *Discovering statistics using IBM SPSS Statistics* (5th ed.). Sage Publications.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning with applications in R*. Springer.

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. MIT Press.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825-2830.

Vapnik, V. (1995). *The nature of statistical learning theory*. Springer.

Appendix

Figure 1: Logistic Regression Confusion Matrix

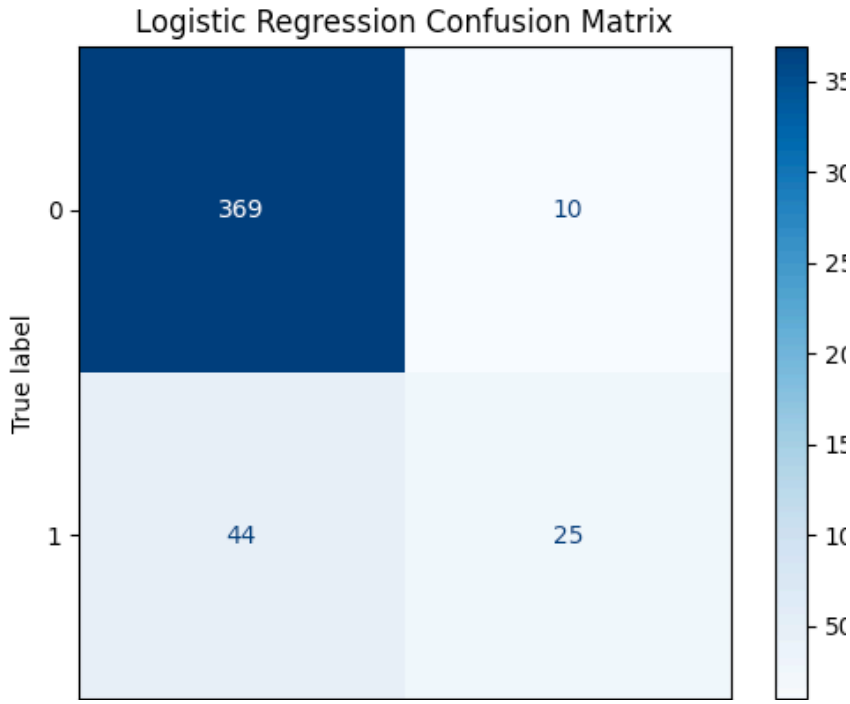

Figure 2: SVM Confusion Matrix