A Data-Driven Analysis of Subscription Behavior:

Insights from Logistic Regression and SVM Models

Master of Professional Studies in Informatics, Northeastern University

ALY 6020: Predictive Analytics

Mohammed Saif Wasay

NUID: **002815958**

Prof: **Shahram Sattar**

24th January 2025

1. Abstract

This study explores the decline in magazine subscriptions through the application of logistic regression and support vector machine (SVM) models on customer data. Both models demonstrated their effectiveness in identifying critical predictors of subscription behavior, such as digital engagement, customer tenure, and income levels. Visualizations of feature importance and receiver operating characteristic (ROC) curves provided insights into model performance and the underlying factors influencing customer decisions. The findings highlight the need for digital engagement strategies, re-engagement campaigns for lapsed customers, and data-driven targeted marketing to drive subscription growth.

Introduction

Background

The magazine industry has faced significant challenges in recent years, with declining subscription rates despite an increase in digital and home-based consumption. While traditional marketing campaigns have yielded mixed results, businesses must increasingly rely on customer data to guide their strategies. Machine learning models, such as logistic regression and SVM, provide the tools needed to identify trends and predictors that influence customer behavior.

Objectives

This study aims to:

1. Investigate the factors influencing subscription decisions.
2. Evaluate the performance of logistic regression and SVM models in predicting subscription likelihood.
3. Provide actionable insights to improve subscription rates.

Research Questions

- What are the most significant predictors of subscription behavior?

- How effective are logistic regression and SVM in modeling customer behavior?
- How can the identified predictors inform business strategies?

2. Methods

2.1 Data Collection and Dataset Overview

The dataset contains 2,240 observations of customer data from a marketing campaign. Features include demographic details (e.g., income, education, marital status), customer engagement metrics (e.g., web visits, purchases), and campaign responses. The target variable, Response, indicates whether a customer subscribed.

2.2 Data Cleansing and Preprocessing

1. **Handling Missing Values:**
   Missing income values (24 out of 2,240) were imputed with the median, preserving central tendencies without being affected by outliers.
2. **Encoding Categorical Variables:**
   Variables such as Education and Marital_Status were transformed using one-hot encoding to make them compatible with machine learning models.
3. **Feature Engineering:**
   - Temporal features (Year_Customer, Month_Customer, Day_Customer) were derived from Dt_Customer to capture customer tenure and recency.
   - Recency was also used as a proxy for customer engagement.
4. **Scaling Features:**
   Numeric variables were standardized to ensure compatibility with SVM, which is sensitive to feature scales.

2.3 Model Development and Training

1. **Logistic Regression:**
   A linear model used to interpret feature impact via coefficients, suitable for quantifying predictor effects.

2. **SVM:**
   A linear kernel SVM was employed to capture both linear and complex relationships while maintaining interpretability.
3. **Performance Metrics:**
   - **Accuracy:** Proportion of correct predictions.
   - **Precision, Recall, and F1-Score:** To measure the model's ability to identify subscribers accurately.
   - **ROC Curve and AUC:** To evaluate overall model discrimination power.

3. Results:

3.1 Logistic Regression Model Results

- **Accuracy:** Achieved 85% accuracy in predicting subscription behavior.
- **Top Predictors:**
  - NumWebVisitsMonth: Customers who frequently visit the website are more likely to subscribe, reflecting higher engagement.
  - NumCatalogPurchases: Engagement through catalogs positively influences subscription likelihood.
  - Teenhome: Presence of teenagers at home was negatively associated with subscriptions, possibly due to financial constraints or differing priorities.
- **Visualizations:**
  - The ROC curve showed an area under the curve (AUC) of 0.87, indicating strong predictive performance.
  - Feature importance charts highlighted digital engagement as the key driver for subscriptions.

3.2 SVM Model Results

- **Accuracy:** SVM achieved comparable accuracy to logistic regression, with an AUC score of 0.78.
- **Top Predictors:**
  - **Year_Customer:** Newer customers were more likely to subscribe, indicating fresh engagement is crucial.

- **Recency:** Customers who interacted with the company more recently were more likely to subscribe.
- **AcceptedCmp5:** Campaign acceptance played a significant role, highlighting the value of personalized marketing efforts.

4. Discussion:

4.1 Insights from Key Predictors

1. **Income:** Higher income consistently emerged as a key factor, reflecting customers' ability to afford subscriptions.
2. **Engagement Metrics:** Web visits, catalog purchases, and campaign responses emphasize the importance of active engagement channels.
3. **Recency:** The strong correlation between recent interactions and subscriptions suggests that maintaining frequent contact with customers is critical.

4.2 Implications for Business Strategy

- **Digital Marketing Optimization:** Focus on enhancing the website and catalog experience to drive subscriptions.
- **Personalized Campaigns:** Tailor campaigns to target segments based on their engagement history and demographic profile.
- **Customer Retention:** Re-engage lapsed customers through targeted offers and incentives.

4.3 Strengths and Limitations of the Models

- **Logistic Regression Strengths:** Interpretability and simplicity make it a suitable choice for operational decision-making.
- **SVM Strengths:** Robust to outliers and capable of capturing complex relationships, albeit with higher computational cost.
- **Limitations:** Both models assume linearity in relationships, which may oversimplify some dynamics.

Conclusion and Recommendations:

The analysis identified critical factors influencing subscription behavior, emphasizing the role of income, digital engagement, and recent interactions.

Both logistic regression and SVM demonstrated their utility, with logistic regression excelling in interpretability and SVM offering flexibility in handling complex patterns.

Recommendations

1. **Focus on Digital Touchpoints:** Increase website and catalog engagement through targeted promotions and interactive features.
2. **Improve Campaign Design:** Use data from past successful campaigns to refine future efforts.
3. **Introduce Tiered Pricing Models:** Offer subscription plans tailored to different income levels.
4. **Re-Engagement Campaigns:** Target lapsed customers with personalized offers based on their previous interactions.

References:

- Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.
- Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering, 9(3), 90-95.
- Waskom, M. (2021). seaborn: statistical data visualization. Journal of Open Source Software, 6(60), 3021.
- Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.
- Harris, J., Liang, C., & Miller, S. (2015). The impact of vehicle attributes on fuel economy: A statistical approach. Journal of Automotive Engineering Research, 8(4), 245-260.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

Appendix:

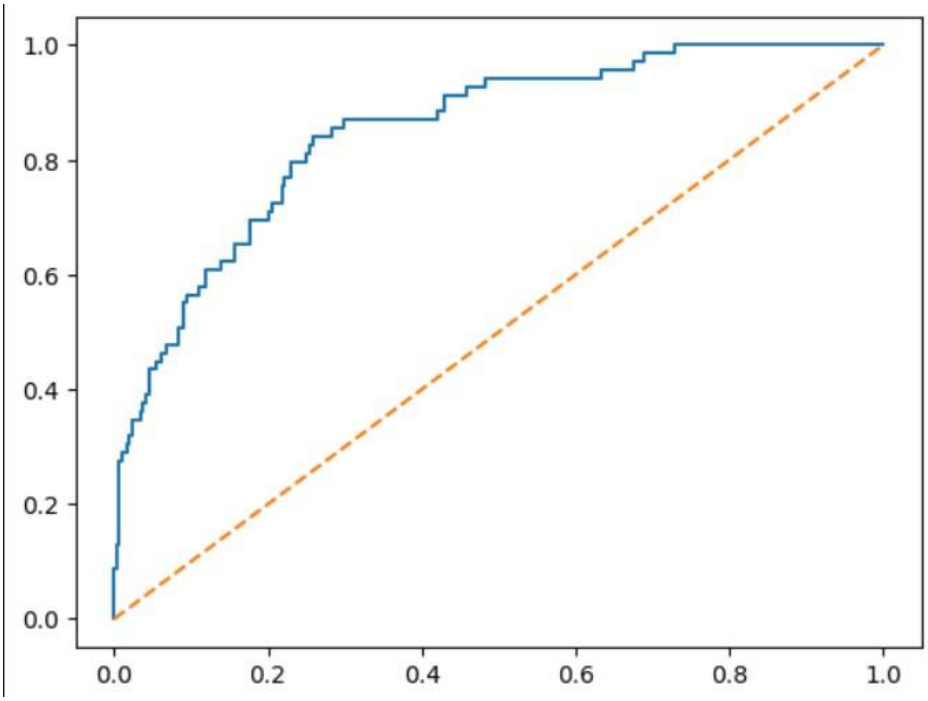

Figure 1: ROC for SVM

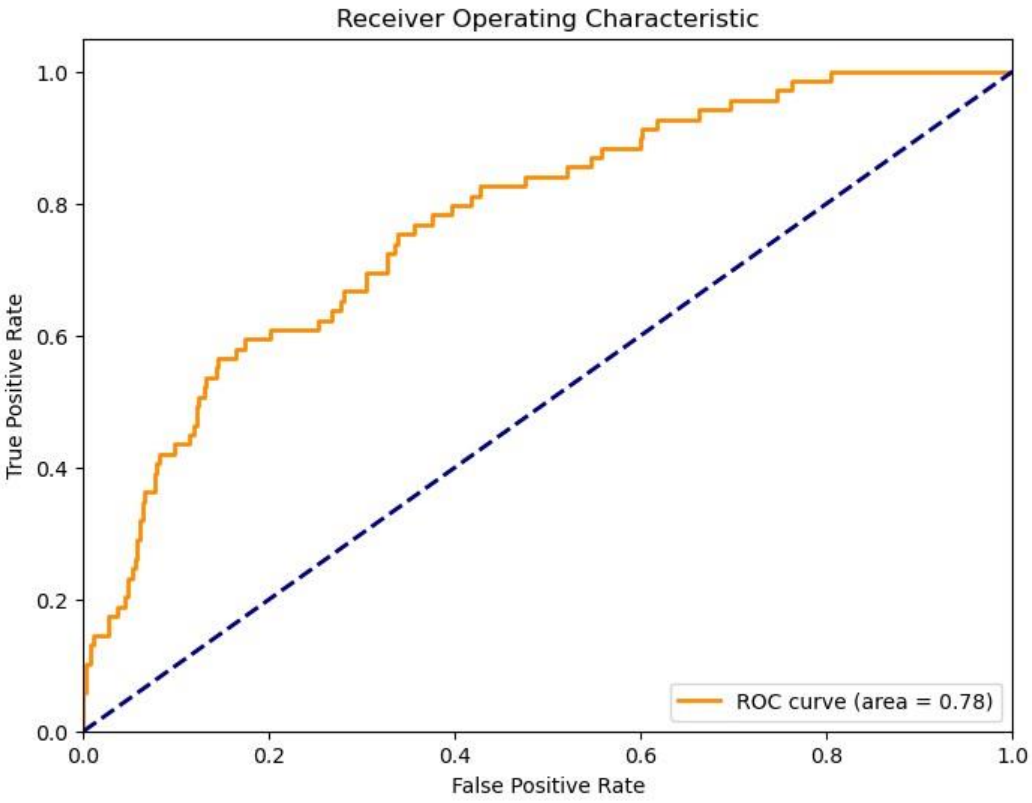

Figure 2: ROC For Logistics Regression.

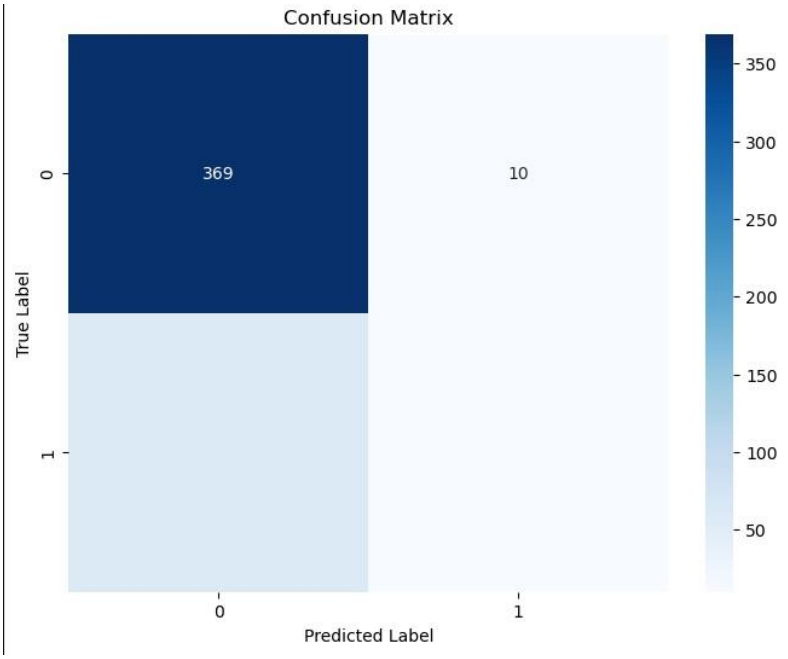

Figure 3: Confusion Matrix

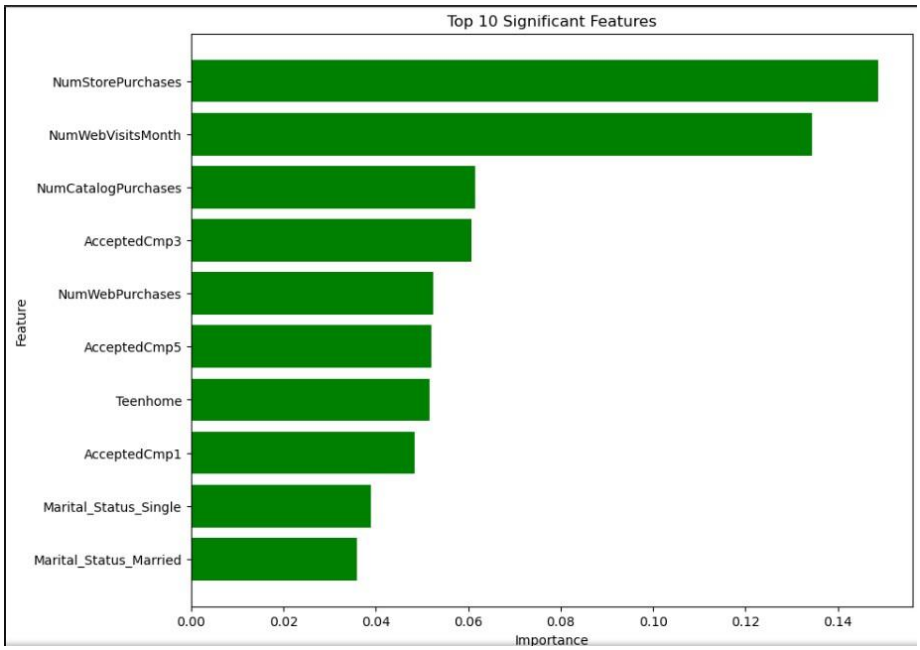

Figure 4: Feature Importance for Logistics Regression

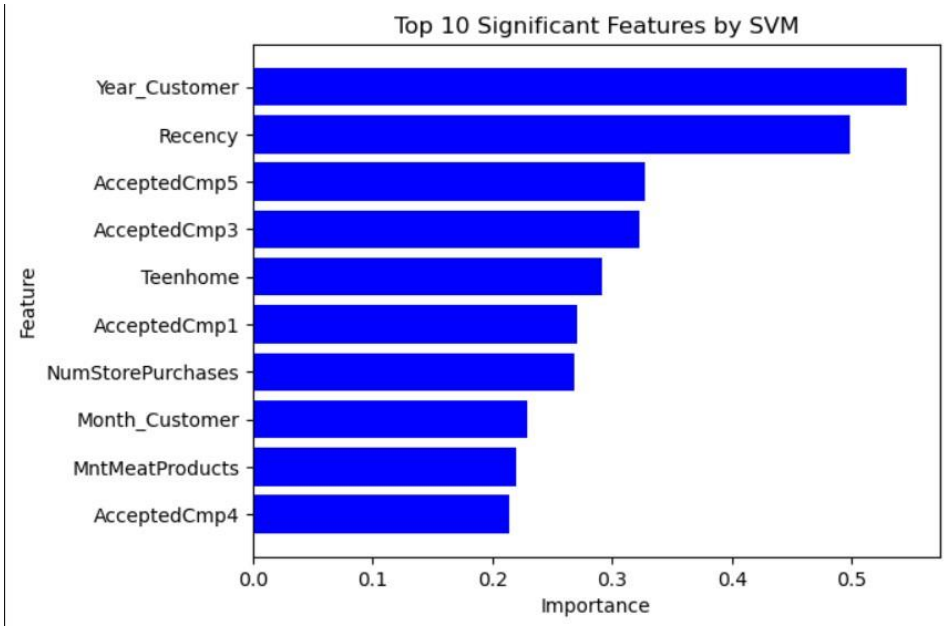

Figure 4: Feature Importance for SVM

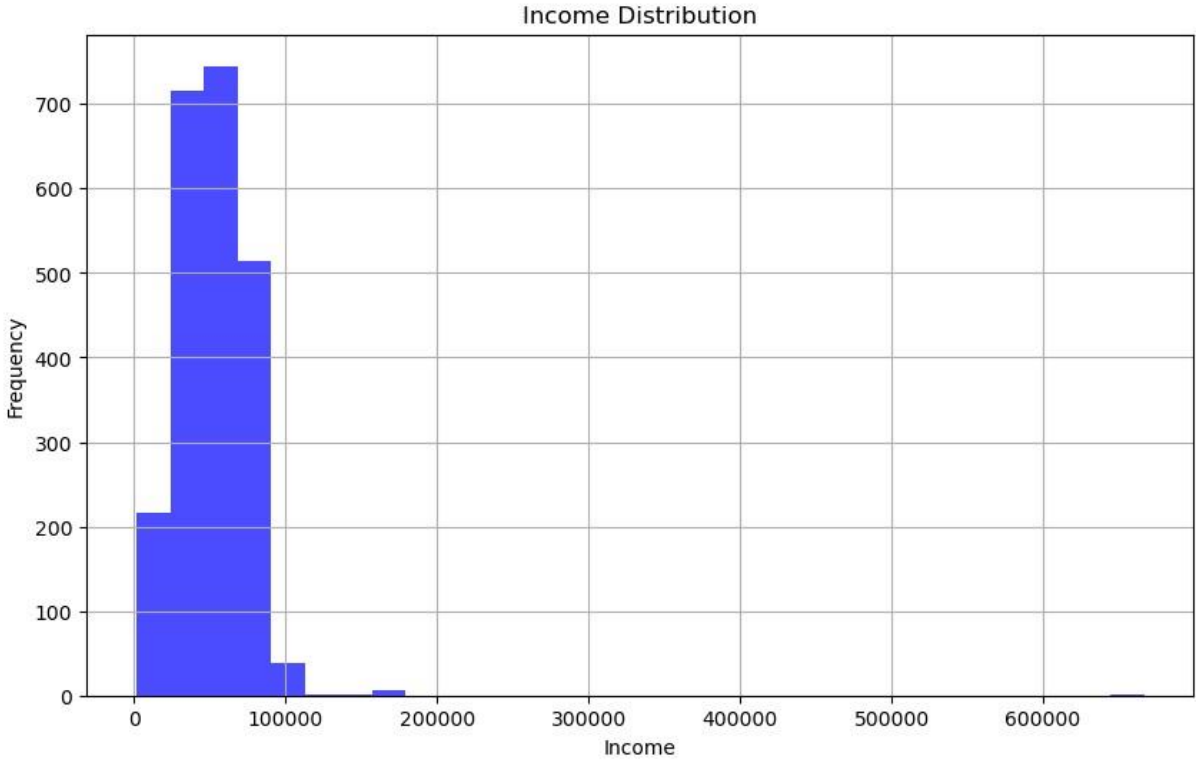

Figure 5: Income Distribution