College of Professional Studies: Northeastern University

ALY 6020 – Predictive Analytics

Instructor: Dr. Shahram Sattar

Academic Term: Winter 2025

Module 3: Project

# Understanding Subscription Decline: A Machine Learning Approach

Submitted By:

Sheila Kwartemaa Boateng

January 28, 2025

# Introduction

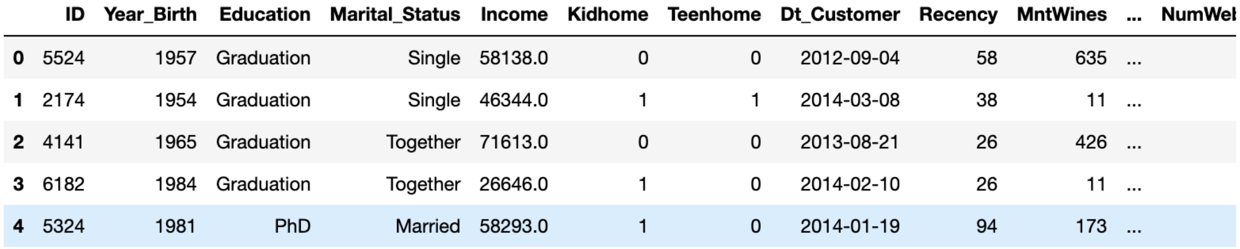
In recent years, the magazine industry has faced significant challenges as subscription rates have fluctuated, especially during periods when consumer behavior shifts. Despite the increased time people spend at home, it was expected that individuals would be more inclined to engage with reading materials such as magazines. However, a noticeable decline in subscriptions suggests that other underlying factors might be at play.

This analysis aims to help the magazine company understand the reasons behind the subscription decline by examining various demographic, behavioral, and transactional factors. The study will first focus on data cleaning, treating outliers, and addressing class imbalances to ensure that the data is well-prepared for modeling. We will then use machine learning models to identify key drivers of subscription behavior. The performance of two models, logistic regression and support vector machine (SVM), will be compared to determine which provides the best insights into subscription trends.

## Data Overview

The data used for the analysis consists of 29 variables and 2240 rows. The variables include demographic information such as age, gender, marital status, and income, along with behavioral data like the number of web visits, catalog purchases, and store visits.

```
# import data
data = pd.read_excel('marketing_campaign.xlsx')
data.head()
```



| | ID | Year_Birth | Education | Marital_Status | Income | Kidhome | Teenhome | Dt_Customer | Recency | MntWines | ... | NumWel |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 5524 | 1957 | Graduation | Single | 58138.0 | 0 | 0 | 2012-09-04 | 58 | 635 | ... | |
| **1** | 2174 | 1954 | Graduation | Single | 46344.0 | 1 | 1 | 2014-03-08 | 38 | 11 | ... | |
| **2** | 4141 | 1965 | Graduation | Together | 71613.0 | 0 | 0 | 2013-08-21 | 26 | 426 | ... | |
| **3** | 6182 | 1984 | Graduation | Together | 26646.0 | 1 | 0 | 2014-02-10 | 26 | 11 | ... | |
| **4** | 5324 | 1981 | PhD | Married | 58293.0 | 1 | 0 | 2014-01-19 | 94 | 173 | ... | |

Upon reviewing the data, there were no duplicate entries, ensuring the integrity of the dataset. However, a small portion of missing values (24) was found in the income column. To address this, the missing values will be filled with the median value of the income column

## Data Preparation

After thoroughly exploring the dataset, several steps were taken to prepare the data for modeling.

- **Missing Values**: Missing values in the `income` column were filled using the median value to maintain the integrity of the dataset.
- **Income Filtering**: Rows with income values greater than $200,000 were removed to avoid any skewing of the analysis due to extreme outliers.
- **Feature Engineering**:
  - The variables pertaining to customer spending (e.g., catalog purchases, store visits) were combined to create a new variable called `Total_Spending`.
  - All the product-related purchase variables were aggregated into a new variable named `TotalPurchases` to represent the overall purchasing behavior of customers.
  - A new variable, `Campaign_Acceptance`, was created to count how many campaigns each customer accepted. This can provide insights into customer engagement.

  - The **tenure** of each customer was calculated as the difference between the maximum `dt_customer` and each `dt_customer` values, providing a sense of how long each customer has been with the company.

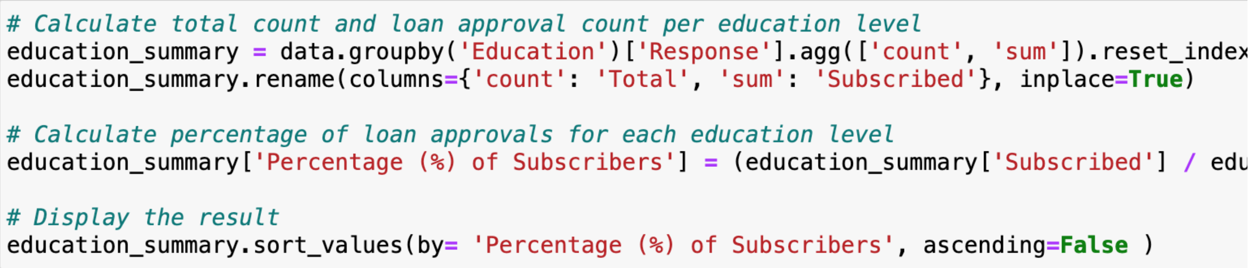
- **Outliers**: Outliers were identified primarily in the columns `income`, `Total_Spending`, `TotalPurchases`, and `NumWebPurchases`. A **log transformation** was applied to these columns to reduce the impact of extreme values and normalize the data.

# Data Exploratory

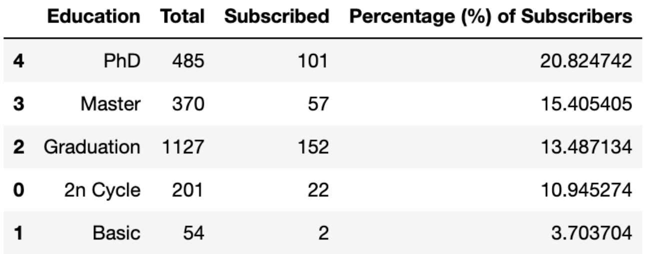
After exploring the data these were the observations made:

## Demographics

**Education**: Individuals with PhD (20.82%) and Master (15.41%) degrees exhibit the highest subscription rates. This is expected, as more educated individuals might have a greater interest in specialized content that the magazine offers.



```
# Calculate total count and loan approval count per education level
education_summary = data.groupby('Education')['Response'].agg(['count', 'sum']).reset_index
education_summary.rename(columns={'count': 'Total', 'sum': 'Subscribed'}, inplace=True)

# Calculate percentage of loan approvals for each education level
education_summary['Percentage (%) of Subscribers'] = (education_summary['Subscribed'] / edu

# Display the result
education_summary.sort_values(by= 'Percentage (%) of Subscribers', ascending=False )
```



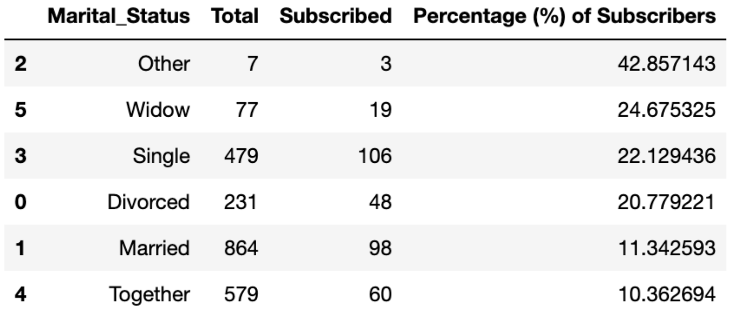
| | Education | Total | Subscribed | Percentage (%) of Subscribers |
|---|---|---|---|---|
| **4** | PhD | 485 | 101 | 20.824742 |
| **3** | Master | 370 | 57 | 15.405405 |
| **2** | Graduation | 1127 | 152 | 13.487134 |
| **0** | 2n Cycle | 201 | 22 | 10.945274 |
| **1** | Basic | 54 | 2 | 3.703704 |

The very low subscription rate for the Basic education group (3.70%) compared to other education levels is notable. This group represents only a small portion of total subscribers (2 out of 54), suggesting either a disinterest in subscriptions or potentially a financial constraint. Overall trend suggests that higher education correlates with higher subscription rates

### Marital Status

Single (22.13%), Divorced (20.78%), and Widow (24.68%) all exhibit significantly higher subscription rates than Married (11.34%) or Together (10.36%). This may suggests that individuals who are not in partnerships may value subscriptions more. Potential reasons include:

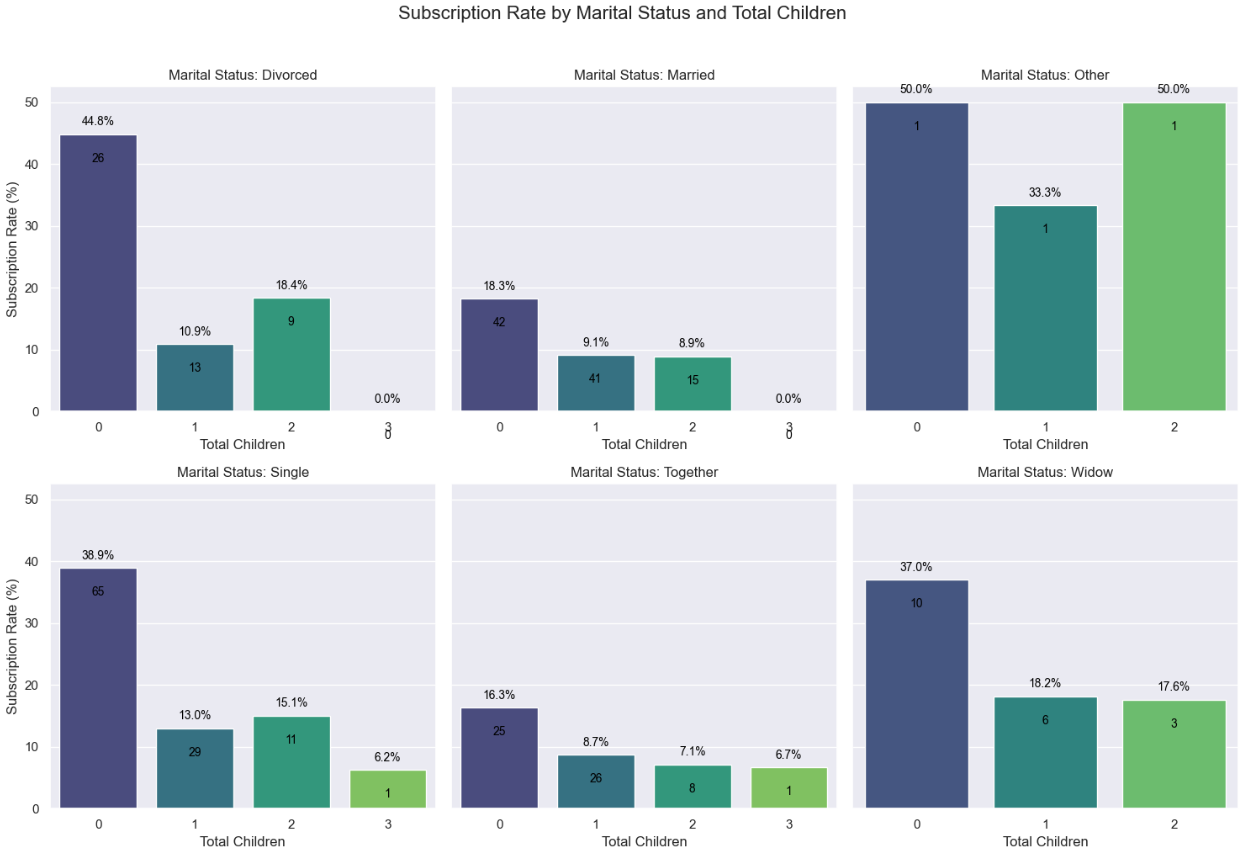
- They might have more time to engage with the magazine.
- The content of the magazine may resonate more with their interests or needs.
- They might prioritize self-enrichment, hobbies, or other personal pursuits, which align with the magazine's offering.



| | Marital_Status | Total | Subscribed | Percentage (%) of Subscribers |
|---|---|---|---|---|
| **2** | Other | 7 | 3 | 42.857143 |
| **5** | Widow | 77 | 19 | 24.675325 |
| **3** | Single | 479 | 106 | 22.129436 |
| **0** | Divorced | 231 | 48 | 20.779221 |
| **1** | Married | 864 | 98 | 11.342593 |
| **4** | Together | 579 | 60 | 10.362694 |

Building on the observation above, the relationship between marital status, the presence of children, and subscription rates further highlights these trends. **Among customers with no children, Single (38.92%), Divorced (44.83%), and Widow (37.03%) individuals consistently show the highest subscription rates**, reinforcing the idea that those not in partnerships or without dependents may have greater disposable income, time, or interest in engaging with the product.

In contrast, **Married (18.26%) and Together (16.34%) customers without children exhibit lower subscription rates**, which decline further as the number of children increases.



**For households with 1 or more children, subscription rates drop significantly across all marital statuses, reaching as low as 0% in some cases for families with 3 children.**

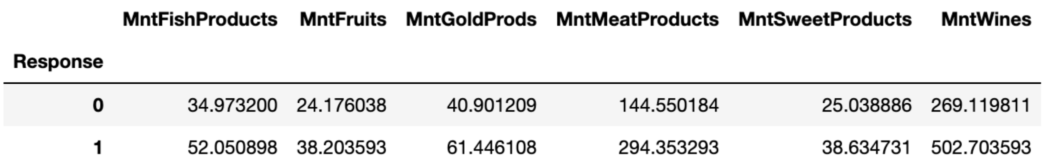
This trend suggests that families, particularly those with multiple children, may face financial constraints or have shifted priorities that reduce the perceived value or feasibility of the subscription. Marketing efforts tailored to these groups could emphasize affordability, family-focused content, or time-saving features to better align with their needs.

## Spending Habit

Subscribers (Response = 1) spend significantly more across all product categories compared to non-subscribers (Response = 0). For instance, spending on wine and meat products is the highest among subscribers, with averages of 502.70 and 294.35, respectively, compared to 269.10 and 144.62 for non-subscribers.

Similarly, while spending on fruits and sweets is relatively lower for both groups, subscribers still spend more in these categories,

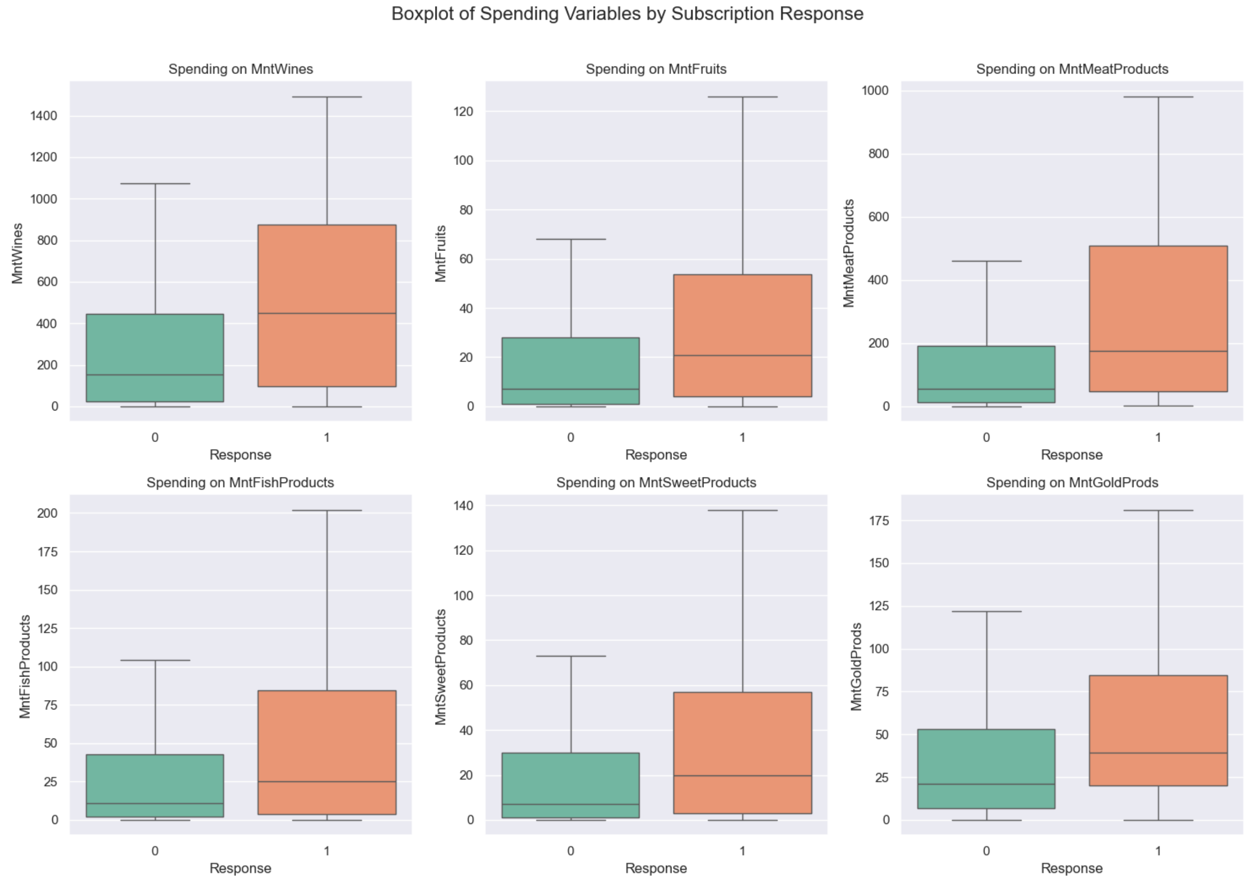
```
data.pivot_table(index='Response',values=['MntWines', 'MntFruits', 'MntMeatProducts',
                                          'MntFishProducts', 'MntSweetProducts', 'MntGoldPr
               aggfunc="mean")
```



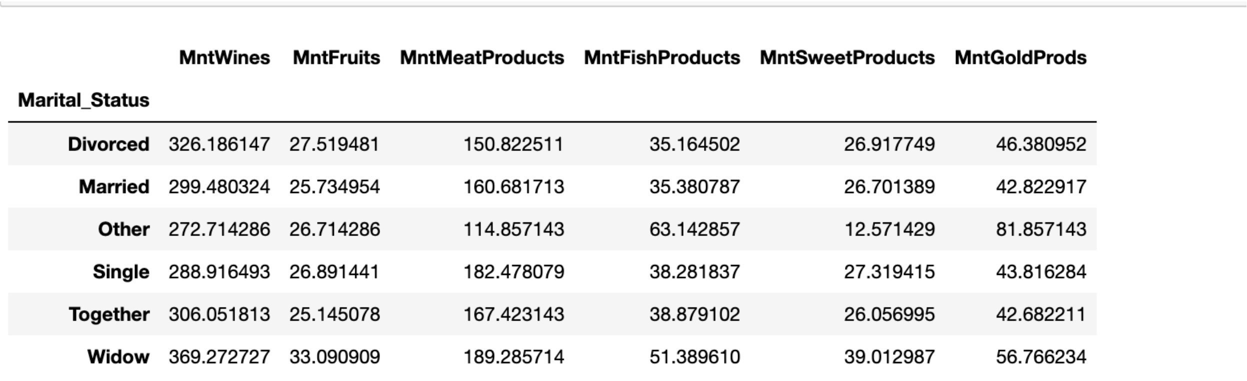
| | MntFishProducts | MntFruits | MntGoldProds | MntMeatProducts | MntSweetProducts | MntWines |
|---|---|---|---|---|---|---|
| **Response** | | | | | | |
| **0** | 34.973200 | 24.176038 | 40.901209 | 144.550184 | 25.038886 | 269.119811 |
| **1** | 52.050898 | 38.203593 | 61.446108 | 294.353293 | 38.634731 | 502.703593 |

This trend is further illustrated by the boxplot below, where all the medians for subscribers across product categories are visibly higher than those for non-subscribers. This suggests that customers who spend more are more likely to subscribe, indicating a strong correlation between higher spending behavior and subscription status.

This makes sense, as customers with higher spending might also have greater financial means, allowing them to prioritize magazine subscriptions as part of their lifestyle. These individuals may view the magazine's content as a valuable complement to their spending habits, especially in categories such as wines and meat products.



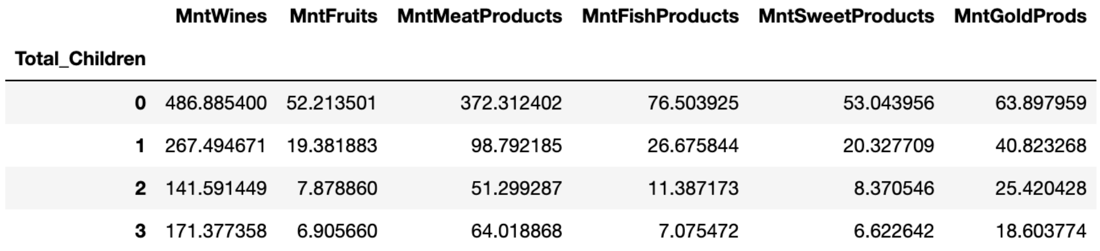
When we viewed the relation between spending vs marital status and number of children, it appears that customers without dependents, such as those who are single, divorced, or widowed, generally show higher spending levels across both luxury items (e.g., wine, gold) and food products. This suggests that individuals with fewer financial responsibilities are more likely to invest in discretionary items like subscriptions, as they may have more disposable income.



| | MntWines | MntFruits | MntMeatProducts | MntFishProducts | MntSweetProducts | MntGoldProds |
|---|---|---|---|---|---|---|
| **Marital_Status** | | | | | | |
| **Divorced** | 326.186147 | 27.519481 | 150.822511 | 35.164502 | 26.917749 | 46.380952 |
| **Married** | 299.480324 | 25.734954 | 160.681713 | 35.380787 | 26.701389 | 42.822917 |
| **Other** | 272.714286 | 26.714286 | 114.857143 | 63.142857 | 12.571429 | 81.857143 |
| **Single** | 288.916493 | 26.891441 | 182.478079 | 38.281837 | 27.319415 | 43.816284 |
| **Together** | 306.051813 | 25.145078 | 167.423143 | 38.879102 | 26.056995 | 42.682211 |
| **Widow** | 369.272727 | 33.090909 | 189.285714 | 51.389610 | 39.012987 | 56.766234 |

On the other hand, customers who are married or in partnerships with children tend to spend less, likely due to the financial responsibilities associated with raising children, which could explain their lower subscription rates.

In essence, individuals without dependents or with fewer children not only show a higher propensity to subscribe but also demonstrate a higher level of spending across categories. This pattern reinforces the idea that those with fewer financial constraints are more likely to engage with both subscriptions and discretionary spending.



| | MntWines | MntFruits | MntMeatProducts | MntFishProducts | MntSweetProducts | MntGoldProds |
|---|---|---|---|---|---|---|
| Total_Children | | | | | | |
| 0 | 486.885400 | 52.213501 | 372.312402 | 76.503925 | 53.043956 | 63.897959 |
| 1 | 267.494671 | 19.381883 | 98.792185 | 26.675844 | 20.327709 | 40.823268 |
| 2 | 141.591449 | 7.878860 | 51.299287 | 11.387173 | 8.370546 | 25.420428 |
| 3 | 171.377358 | 6.905660 | 64.018868 | 7.075472 | 6.622642 | 18.603774 |

## Income

The results show that, on average, individuals who subscribed (Response = 1) have a higher income ($60, 209.68) compared to those who did not subscribe (Response = 0), with an average income of $50,496.58.

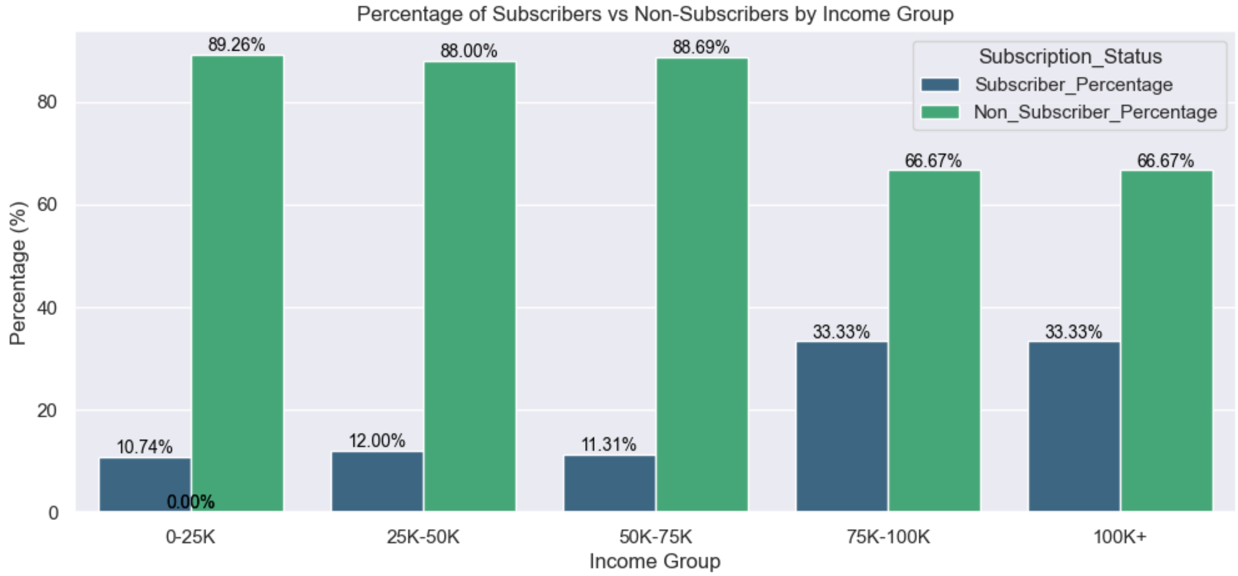
This suggests that:

- Higher Income and Subscription Likelihood: People with higher incomes seem to be more likely to subscribe to the magazine, which aligns with the observation that individuals with more disposable income might be more willing to spend on subscriptions, including magazines.

```
Response
0    50496.576370
1    60209.675676
Name: Income, dtype: float64
```

When we categorized income into groups, we deduce that:

- The highest subscription percentages are observed in the higher income groups: 75K-100K and 100K+ income groups both have 33.33% subscribers, which is significantly higher compared to the lower income groups.
- On the other hand, the 0-25K, 25K-50K, and 50K-75K income groups have lower subscription rates, ranging from 10.74% to 12.00% for subscribers.
- **Lower Income Groups (0-50K):** In the 0-25K and 25K-50K income groups, most customers do not subscribe, with 89.26% and 88.00% respectively falling into the non-subscriber category. Customers in these income groups are more likely to be non-subscribers, which may indicate budget constraints or lower disposable income for non-essential purchases like magazine subscriptions.
- **Middle and Higher Income Groups (50K-100K and 100K+):** Subscription rates are higher in these groups. The 75K-100K and 100K+ groups show more balanced distribution between subscribers and non-subscribers, with 33.33% of customers subscribing and 66.67% not subscribing. This indicates that higher-income groups have more disposable income, and as a result, may be more open to spending on non-essential services like subscriptions.



Non-subscriber percentages are inversely related to income levels. As income increases, the proportion of non-subscribers decreases. Hence we can conclude that Income plays a significant role in determining subscription behavior. Lower-income groups are much less likely to subscribe, whereas higher-income groups show a higher likelihood of subscribing.

**However**, even in higher-income groups, a considerable portion of people still do not subscribe. The 100K+ group still has a significant 66.67% non-subscriber rate, suggesting that factors other than income, such as interest in the magazine's content or the perceived value of the subscription, could also influence the decision to subscribe.

## Campaign Effectiveness

### High Non-Acceptance Rates Across Campaigns:

The majority of customers (~92-98%) did not accept the campaigns (Acceptance = 0), indicating that the campaigns were not highly successful in driving engagement for most customers.

Interestingly, despite a small percentage of customers accepting the campaigns (6-7% for most campaigns), a significant proportion of these customers end up subscribing. For example, in AcceptedCmp1, 55.63% of those who accepted the campaign went on to subscribe. This highlights that customers who engage with the campaigns have a strong likelihood of subscribing, making those acceptors valuable for driving subscriptions.

Across all campaigns, the proportion of subscribers among those who accepted a campaign (Acceptance = 1) is significantly higher than among non-acceptors (Acceptance = 0). For instance, in Campaign AcceptedCmp1, 55.63% of acceptors subscribed, compared to only 12.27% of non-acceptors. This reinforces the idea that engagement with the campaign is a strong indicator of subscription potential.

```
Percentage Summary
Campaign AcceptedCmp1:

   Acceptance  Not_Subscribe  Subscribed
0    93.58047      87.729469   12.270531
1     6.41953      44.366197   55.633803
--------------------------------------------------
Campaign AcceptedCmp2:

   Acceptance  Not_Subscribe  Subscribed
```

```
0   98.643761      85.655362   14.344638
1    1.356239      33.333333   66.666667
--------------------------------------------------
Campaign AcceptedCmp3:

   Acceptance  Not_Subscribe  Subscribed
0   92.631103      87.506101   12.493899
1    7.368897      52.760736   47.239264
--------------------------------------------------
Campaign AcceptedCmp4:

   Acceptance  Not_Subscribe  Subscribed
0   92.585895      86.767578   13.232422
1    7.414105      62.195122   37.804878
--------------------------------------------------
Campaign AcceptedCmp5:

   Acceptance  Not_Subscribe  Subscribed
0   92.721519      88.200878   11.799122
1    7.278481      43.478261   56.521739
--------------------------------------------------
```

**Missed Opportunity Among Non-Acceptors:**

Despite the high non-acceptance rates, a substantial number of non-acceptors still subscribe to the service (e.g., 11.48% in AcceptedCmp1). This suggests that while they didn't engage with the campaign, these customers are still likely to be influenced by other factors, possibly through different marketing channels or internal motivations. This indicates a missed opportunity for targeting these non-acceptors more effectively, as they represent a group that could potentially be converted with improved or alternate strategies.

**Campaign Performance Insights:**

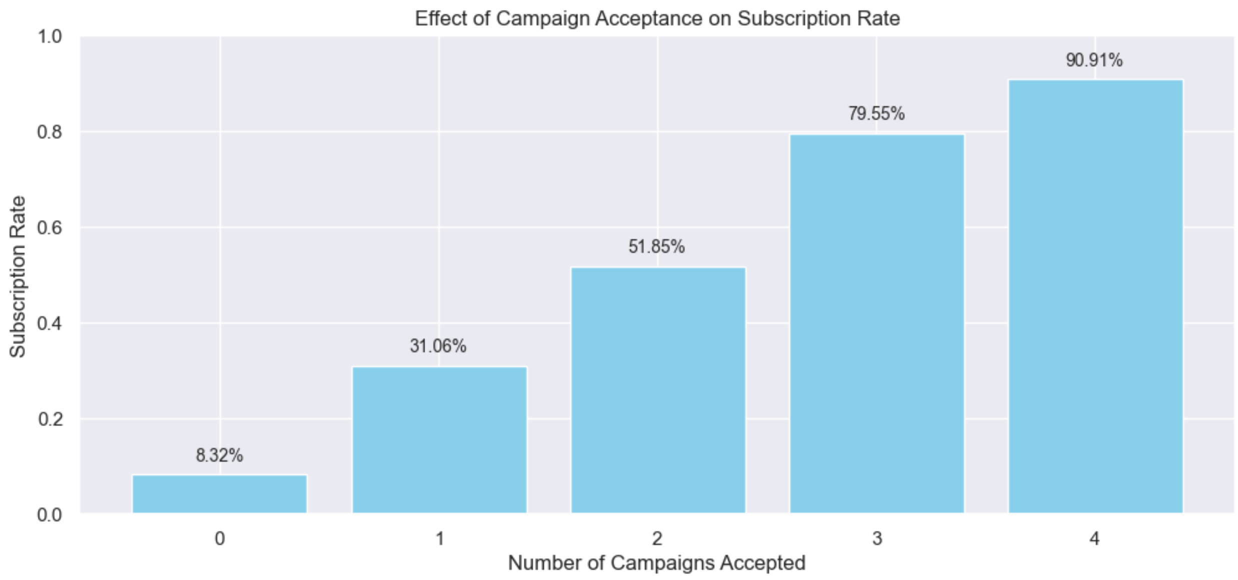
Campaign AcceptedCmp3, 4, and 5 had the highest acceptance rates, while Campaign AcceptedCmp2 had the lowest acceptance rate. However, among the three highest-performing campaigns, Campaign 5 stood out with the highest subscription rate among acceptors, followed by Campaign 3, and then Campaign 4.

In conclusion, there is significant potential for the campaigns to drive higher subscription rates. To fully capitalize on this opportunity, the company should focus on developing more targeted campaigns that effectively engage the right audience.

**Combine Acceptance Campaign**

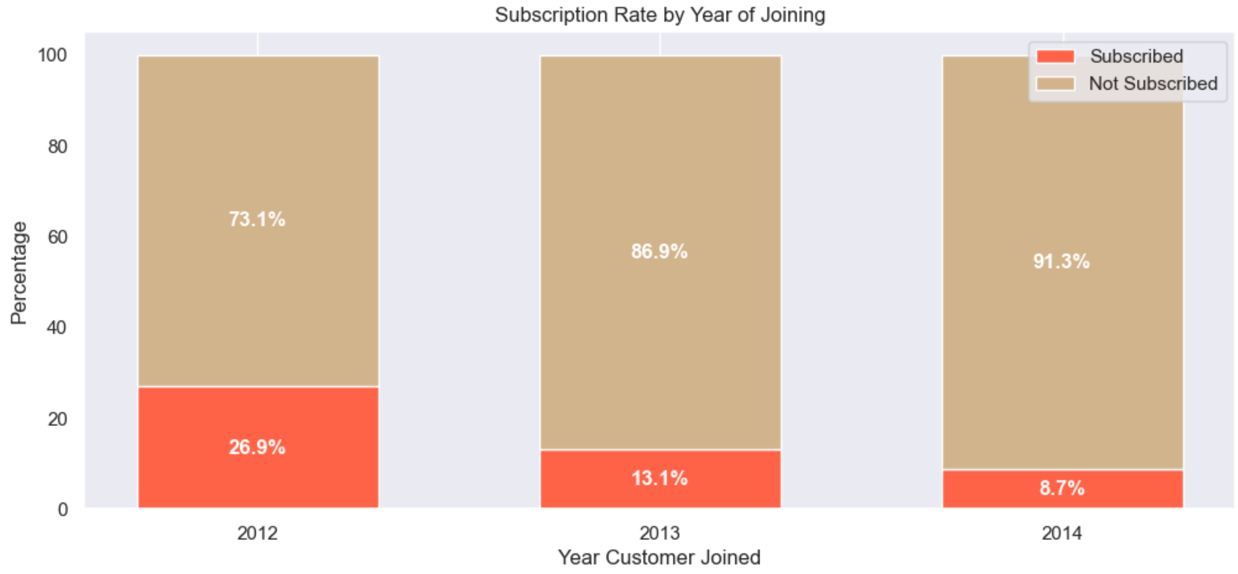
Summing all acceptance campaign show that as customers engage with more campaigns, their chances of subscribing increase significantly. Specifically, customers who accept more campaigns have a notably higher subscription rate. For instance, those who did not accept any campaign have a subscription rate of 8.32%, those who accepted 1 campaign have a subscription rate of 31.06%, and those who accepted 2 campaigns have a subscription rate of 51.85%. As engagement increases further, customers who accepted 3 campaigns show a subscription rate of 79.55%, and those who accepted 4 campaigns exhibit a subscription rate of 90.91%.

This demonstrates that customer engagement with campaigns plays a crucial role in boosting subscription behavior.



## Tenure & Year Joined

The data reveals a declining trend in subscription rates over the years. Customers who joined in 2012 had the highest subscription rate, with 26.94% subscribing, while those who joined in 2013 had a lower rate of 13.08%. The trend continues into 2014, where only 8.70% of customers subscribed, meaning over 91% did not subscribe. This suggests that newer customers are less likely to subscribe compared to earlier ones. Possible reasons could include changes in marketing strategies, customer preferences, or differences in engagement levels between early and later joiners.



The average tenure (number of days since joining) for customers who did not subscribe is 336.99 days, whereas for those who subscribed, it is 448.08 days. This suggests that customers who have been with the company longer are more likely to subscribe.

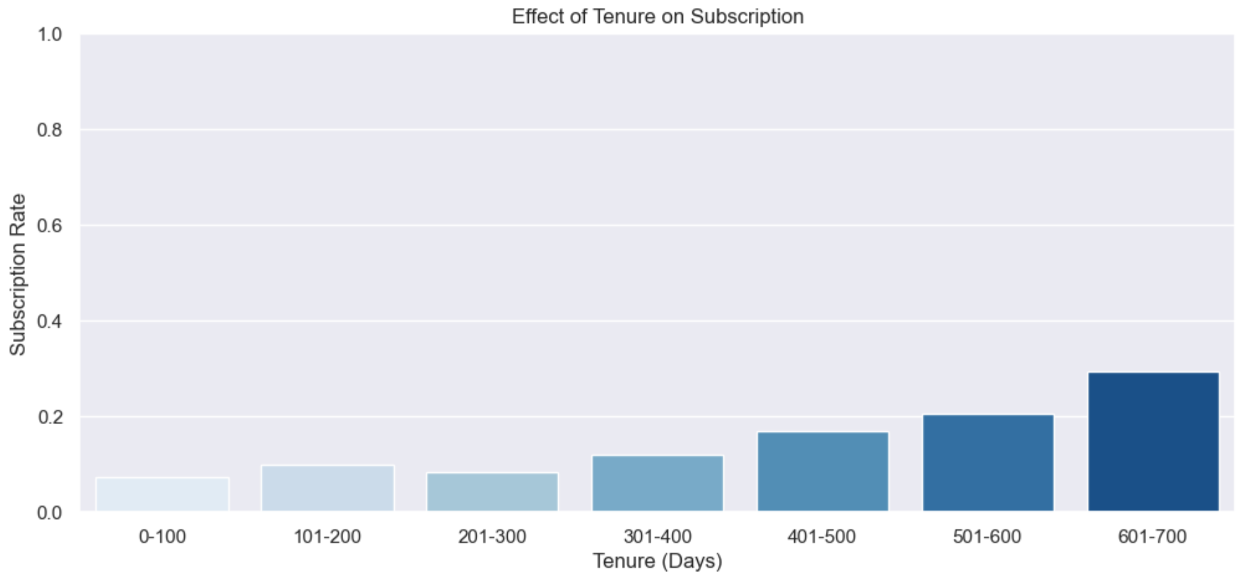
The median tenure for customers who subscribed is higher than those who did not. This reinforces the idea that longer-tenured customers are more engaged and more likely to subscribe.

```
Response
0    336.990420
1    448.081081
```

Subscription Rates by Tenure Bins futher support this observation:

- Customers with tenure between 0-100 days have the lowest subscription rate (7.23%).
- Subscription rates gradually increase across the tenure bins, reaching 29.48% for those with tenure between 601-700 days.
- This suggests that longer-tenured customers are more likely to subscribe.

The steady increase in subscription rates across tenure bins highlights the importance of customer retention strategies.



Based on the analysis, it seems that the customers who are more likely to subscribe are long-term or existing customers, rather than new ones. This indicates that newer customers may not be as engaged or responsive to subscription offers, which could lead to lower conversion rates among them.

Other Vairiables

For the variable complaints, there was very little difference in the subscription rates between customers who have complaints (15.00%) and those who do not (15.05%). This suggests that complaints do not significantly affect the likelihood of subscribing, at least in this dataset. The subscription rate is almost identical between the two groups, indicating that other factors might be more influential in determining whether a customer subscribes.

```
complaints_response_rate = data.groupby('Complain')['Response'].mean()
complaints_response_rate
```

```
Complain
0    0.150547
1    0.150000
Name: Response, dtype: float64
```

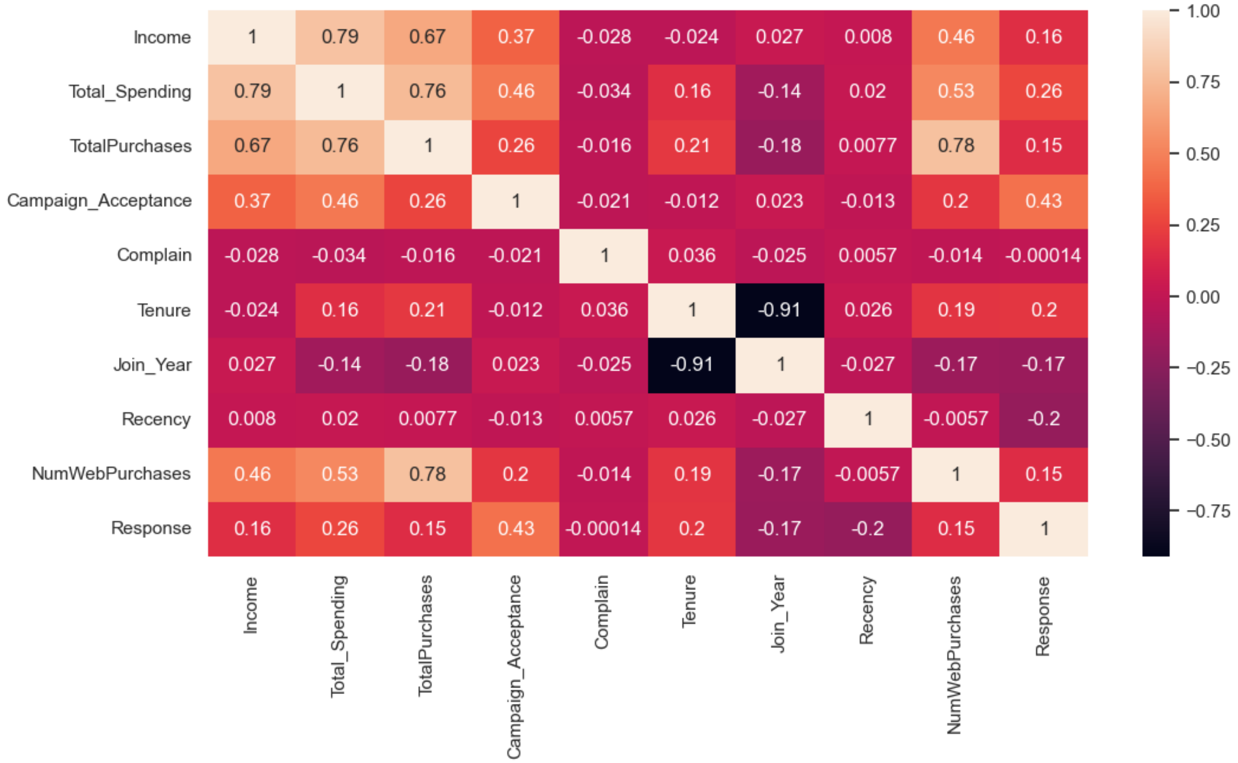
```
from scipy.stats import ttest_ind
# Separate the groups
subscribe = data[data['Response'] == 1]['Complain']
not_subscribe = data[data['Response'] == 0]['Complain']

# Perform a t-test
t_stat, p_value = ttest_ind(subscribe, not_subscribe, equal_var=False)

print(f"T-Statistic: {t_stat}, p-value: {p_value}")
```

```
T-Statistic: -0.006816320260851771, p-value: 0.994564374706362
```

The t-test shows a p-value of $0.9946 > 0.05$. Hence, we fail to reject the null hypothesis, meaning that complaints do not appear to have a significant effect on subscription behavior in this dataset. Therefore, there is no strong evidence to suggest that customer complaints are associated with a change in subscription rates.

## Correlation Matrix

The correlation plot reveals that the variable **Campaign_Acceptance** has the most positive correlation with the target variable **Response** (0.42), followed by **Total_Spending** (0.26), **Tenure** (0.20), and **Income** (0.16), among others. This suggests that customers who accepted more campaigns tend to have a higher likelihood of responding positively, while factors such as total spending, tenure, and income also have a moderate positive impact on the likelihood of subscription.

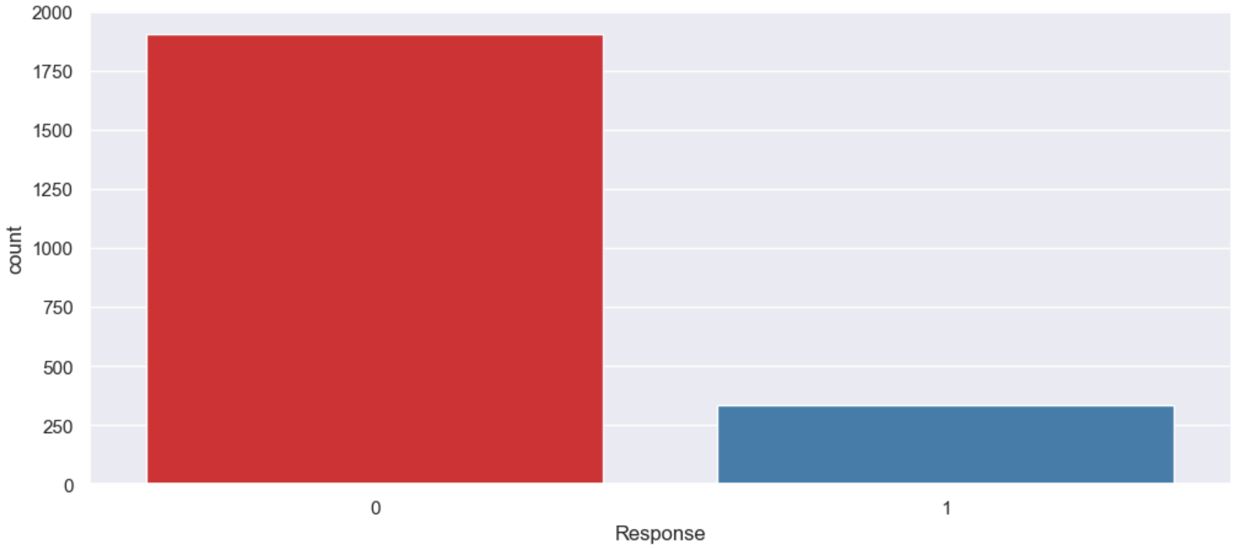


# Data Modeling

The dataset was split into an 80% training set and a 20% test set. Numeric variables were scaled using standard scaling to ensure they were on the same scale. We then used both **Logistic Regression** and **Support Vector Machine (SVM)** models to predict the target classes and evaluate their performance.

### Balancing the Data

The dataset in question exhibits a significant class imbalance, with 1,879 instances of non-subscribers (class 0) and only 333 instances of subscribers (class 1).



This imbalance can lead to biased model performance, particularly in classification tasks, where the model tends to predict the majority class more often, neglecting the minority class. Before addressing the class imbalance, both the Logistic Regression and SVM models failed to predict subscribers (class 1) effectively. The precision and recall for Logistic Regression were 0.74 and 0.37, respectively, while for SVM, they were 0.81 and 0.31, respectively.

Given the importance of accurately predicting subscribers (class 1) for a magazine company trying to understand and maintain its subscriber base, addressing class imbalance is crucial. This is because:

#### Impact on Business Decisions

- **Subscribers (Class 1)** are crucial for the magazine's growth. Accurately predicting who is likely to subscribe or cancel helps the company improve customer retention, create targeted marketing campaigns, and prevent churn. Missing these predictions can lead to lost revenue and missed opportunities.

#### Increased Cost of Misclassification

- Misclassifying subscribers (false negatives) as non-subscribers wastes marketing resources on the wrong customers, while ignoring those who need retention efforts. Misclassifying non-subscribers (false positives) can also result in unnecessary spending, offering subscriptions to people who aren't interested.

### Balanced Performance Using Balancing Techniques

We applied three balancing techniques—SMOTE, RandomUnderSampler, and SMOTETomek—to address the class imbalance in the dataset.

Among these techniques, **SMOTETomek** provided the best balance between precision and recall for both classes, resulting in improved model performance. Thus, the SMOTETomek approach was chosen.

**Balanced Performance:** SMOTETomek offers a good compromise between identifying subscribers (recall) and minimizing false positives (precision), ensuring a more accurate prediction of potential subscribers.

**Improved Recall:** While not as high as RandomUnderSampler, the recall of 0.63 represents a significant improvement over the original model, enabling the company to identify more potential subscribers.

**Manageable False Positives:** The precision of SMOTETomek is similar to RandomUnderSampler, but the slightly lower recall means fewer false positives to manage, helping to optimize resource allocation.

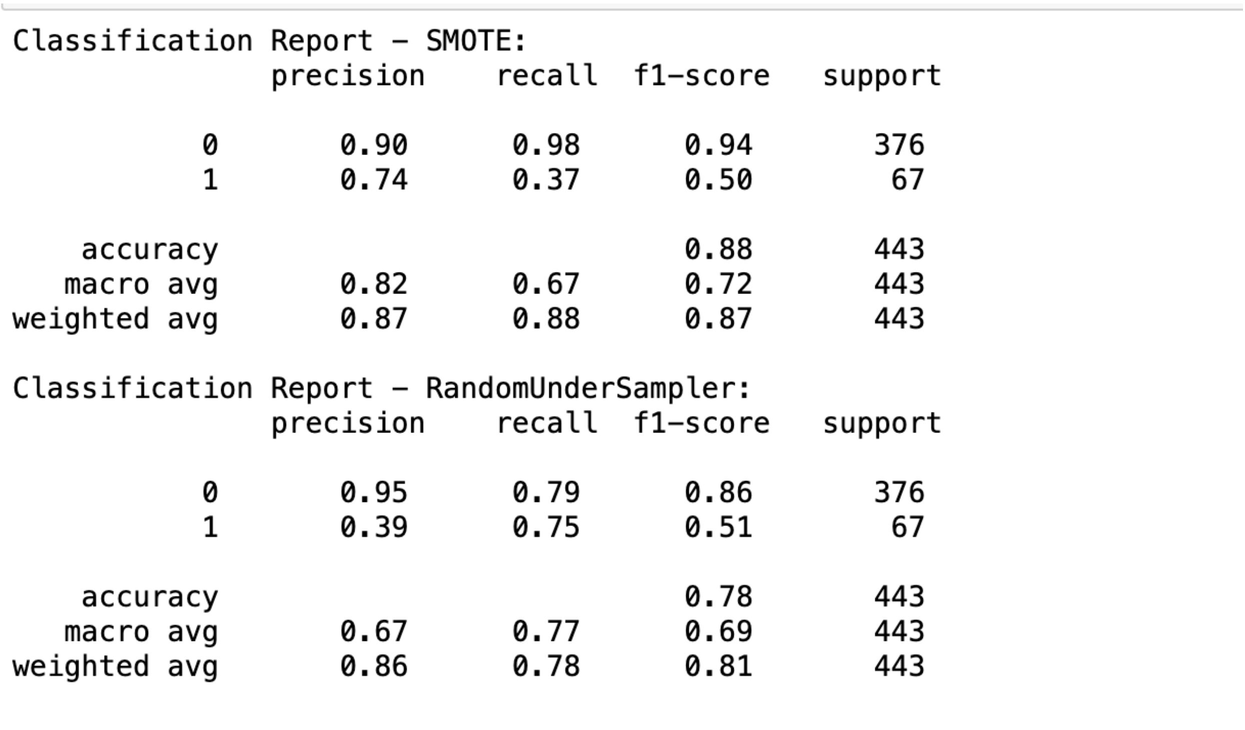
**Overall Accuracy:** SMOTETomek maintains a good overall accuracy, which is essential for model reliability and consistency in predictions.

**Representation of Both Classes:** SMOTETomek both oversamples the minority class (subscribers) and cleans the class boundary, potentially providing a more representative dataset for the company's analysis.

Hence for the magazine company's goal of understanding and retaining subscribers, the SMOTETomek approach offers:

- A well-balanced approach to identifying potential subscribers without overwhelming the company with false positives.
- An improved ability to detect potential subscription issues compared to the original, imbalanced model.
- A more nuanced view of the data, potentially revealing insights that were previously obscured by the class imbalance.

While RandomUnderSampler showed a higher recall, the more balanced approach of SMOTETomek is likely to offer more reliable and actionable insights, which are crucial for the company's strategy to address subscription declines effectively.



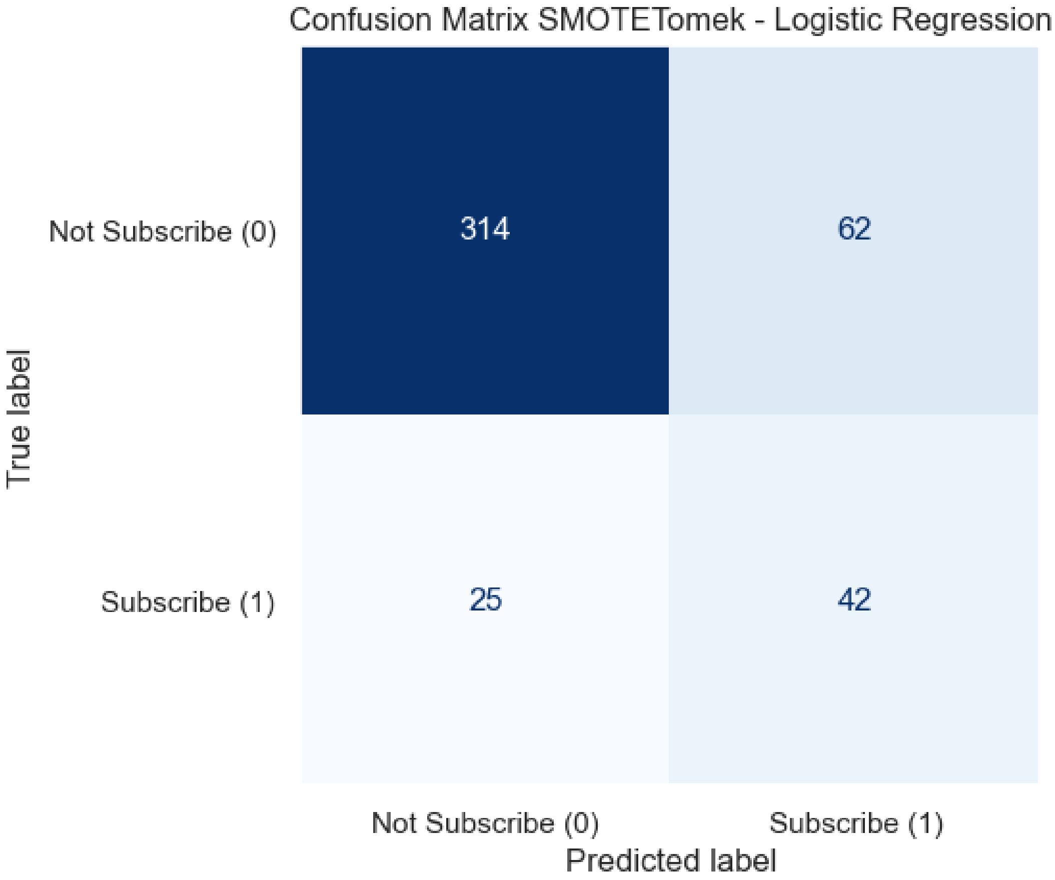
```
Classification Report - SMOTE:
              precision    recall  f1-score   support

           0       0.90      0.98      0.94       376
           1       0.74      0.37      0.50        67

    accuracy                           0.88       443
   macro avg       0.82      0.67      0.72       443
weighted avg       0.87      0.88      0.87       443

Classification Report - RandomUnderSampler:
              precision    recall  f1-score   support

           0       0.95      0.79      0.86       376
           1       0.39      0.75      0.51        67

    accuracy                           0.78       443
   macro avg       0.67      0.77      0.69       443
weighted avg       0.86      0.78      0.81       443
```

Figure 1: Logistic Regression Balancing the Data

### Logistic Regression

The Logistic Regression Model using the SMOTETomek technique achieved an accuracy of 0.80. The precision for predicting non-subscribers (class 0) is high at 0.93, indicating few false positives. However, the recall for subscribers (class 1) is 0.63, suggesting the model is capturing a significant portion of potential subscribers. The F1-score of 0.49 for class 1 reflects a trade-off between precision and recall. While the overall accuracy is solid, the model still struggles with identifying subscribers, as shown by the lower precision and recall for class 1. The weighted average F1-score of 0.82 indicates a decent balance overall.

```
Model Accuracy: 0.8036
SMOTETomek - Logistic Regression
              precision    recall  f1-score   support

           0       0.93      0.84      0.88       376
           1       0.40      0.63      0.49        67

    accuracy                           0.80       443
   macro avg       0.67      0.73      0.68       443
weighted avg       0.85      0.80      0.82       443
```



The logistic regression results indicate several significant variables impacting subscription behavior. Higher Total Spending and Campaign Acceptance positively influence the likelihood of subscription, suggesting that targeted campaigns to high-spending customers could increase subscriptions. Tenure also plays a key role, with longer customer relationships boosting the probability of subscribing. On the other hand, higher Income, more Recent Interactions, and Basic Education are associated with a decreased likelihood of subscribing. This is very interesting, especially since our exploratory analysis established that higher income is typically associated with a higher likelihood of subscribing.

Furthermore, **Marital Status** significantly influences subscription likelihood, with **married**, **single**, and **widowed** individuals showing lower chances of subscribing, which is an unexpected finding. This could suggest that family or relationship status, or associated financial priorities, might play a role in subscription behavior.

Among all the input variables only Marital_Status_Other seem not to be significant since p-value >0.05.

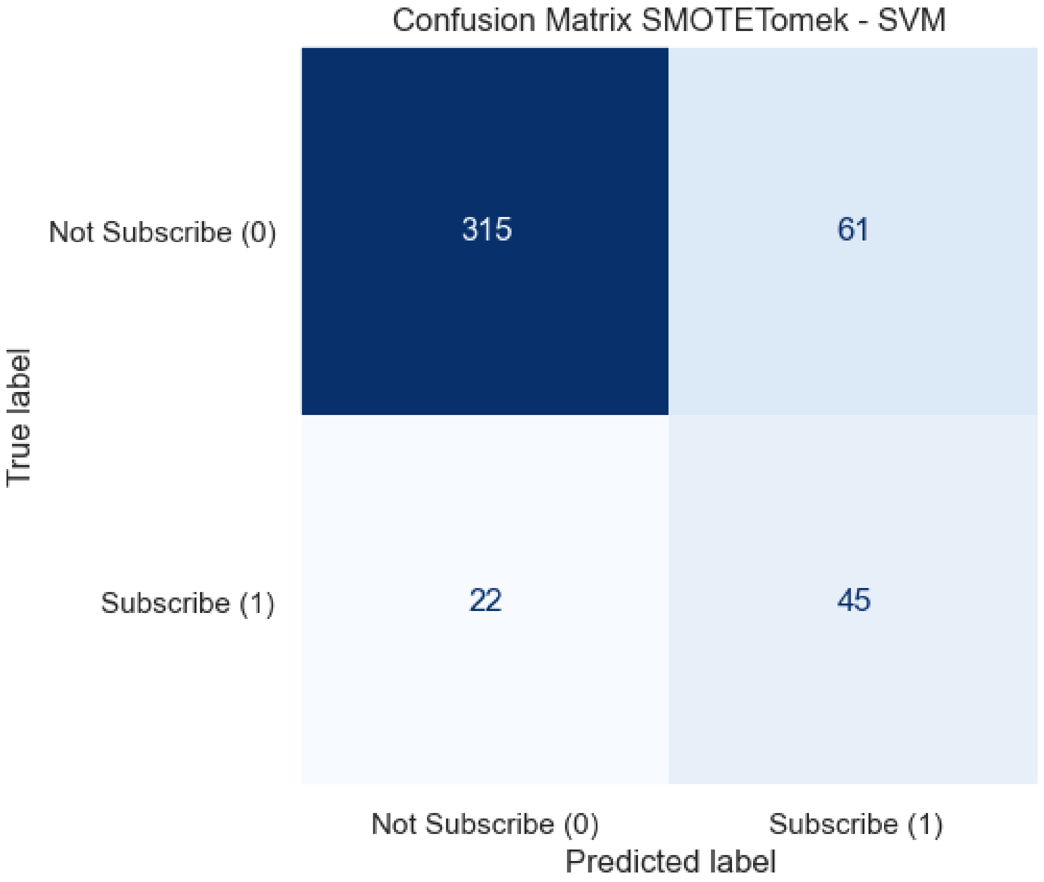
```
===========================================================================================
                               coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------------
const                        1.8067      0.186      9.737      0.000       1.443       2.170
Income                      -0.3859      0.101     -3.822      0.000      -0.584      -0.188
Total_Spending               1.2135      0.178      6.807      0.000       0.864       1.563
TotalPurchases              -1.0464      0.175     -5.996      0.000      -1.389      -0.704
Campaign_Acceptance          1.0019      0.062     16.152      0.000       0.880       1.124
Tenure                       0.9753      0.066     14.864      0.000       0.847       1.104
Recency                     -1.0714      0.063    -16.984      0.000      -1.195      -0.948
NumWebPurchases              0.4240      0.110      3.858      0.000       0.209       0.639
Education_Basic             -3.5547      0.807     -4.403      0.000      -5.137      -1.972
Education_Graduation        -1.3546      0.163     -8.329      0.000      -1.673      -1.036
Education_Master            -1.5307      0.206     -7.432      0.000      -1.934      -1.127
Education_PhD               -0.6685      0.183     -3.655      0.000      -1.027      -0.310
Marital_Status_Married      -2.3545      0.154    -15.273      0.000      -2.657      -2.052
Marital_Status_Other        -2.8349      1.489     -1.903      0.057      -5.754       0.084
Marital_Status_Single       -1.4737      0.163     -9.052      0.000      -1.793      -1.155
Marital_Status_Together     -2.8846      0.182    -15.878      0.000      -3.241      -2.528
Marital_Status_Widow        -2.6844      0.378     -7.103      0.000      -3.425      -1.944
===========================================================================================
```

## SVM

The SMOTETomek technique with SVM achieved an overall accuracy of 81%. It performed well with non-subscribers (class 0), with high precision (0.93) and recall (0.84), indicating accurate predictions. However, for subscribers (class 1), the precision (0.42) was lower, indicating some false positives, while recall (0.67) showed decent identification of actual subscribers. The F1-score for subscribers (0.52) reflects a moderate balance between precision and recall.

```
SMOTETomek - SVM
Accuracy: 0.8126410835214447
              precision    recall  f1-score   support

           0       0.93      0.84      0.88       376
           1       0.42      0.67      0.52        67

    accuracy                           0.81       443
   macro avg       0.68      0.75      0.70       443
weighted avg       0.86      0.81      0.83       443
```

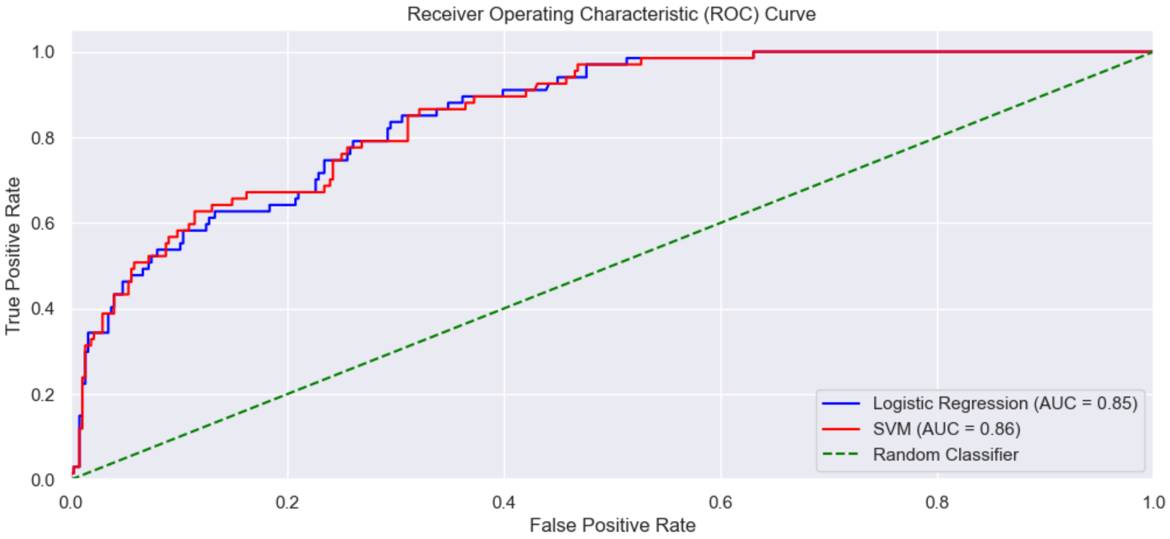
The confusion matrix shows that the model correctly predicted 315 non-subscribers (true negatives) and 45 subscribers (true positives). However, it misclassified 61 non-subscribers as subscribers (false positives) and 22 subscribers as non-subscribers (false negatives). While the model is effective at identifying non-subscribers, it faces challenges in accurately predicting subscribers, which is crucial for the business to target potential customers and prevent churn.



### Comparing Model

Among the two models, **SMOTETomek with SVM** appears to be the better choice for the business context of a magazine company. While both models show similar accuracy, the key differentiators are **recall** and **F1-score**.

- SMOTETomek with SVM has a higher recall of 0.67 compared to 0.63 with Logistic Regression, which is crucial for identifying subscribers (Class 1) and minimizing missed opportunities.
- Additionally, the F1-score for SVM (0.52) is higher than that of Logistic Regression (0.49), indicating a more balanced performance between precision and recall.
- **Precision**: Both models have similar precision for identifying subscribers, but SVM is marginally better (0.42) compared to Logistic Regression (0.40). This means SVM is a bit better at minimizing false positives.
- The AUC (Area Under the Curve) scores for Logistic Regression (0.8547) and SVM (0.8565) are very close, indicating that both models perform similarly well in distinguishing between the two classes. These scores, being above 0.85, suggest strong predictive power for both models. The slightly higher AUC for SVM (0.8565) implies it may have a marginal edge in overall classification performance



Given that the goal is to effectively target and retain subscribers, **SMOTETomek with SVM** is more adept at identifying potential subscribers with fewer false positives, making it the more reliable choice for the company's strategy.

The **SVM model** identifies several key variables influencing subscription behavior. **Total Spending** (1.02) and **Campaign Acceptance** (0.65) are positively associated with a higher likelihood of subscribing, suggesting that customers who spend more and engage with campaigns are more likely to convert. In contrast, **Income** (-0.25) and **Recency** (-0.77) show negative relationships, indicating that wealthier customers or those with more recent interactions are less likely to subscribe, which aligns with unexpected findings from the logistic regression. Educational background, especially **Basic Education** (-2.10), plays a significant role, suggesting that lower education levels may reduce the likelihood of subscription. Additionally, **Marital Status** variables, such as **Married**, **Single**, and **Widow**, are linked to a lower probability of subscribing.

# Conclusion

In conclusion, the decline in subscriptions may be influenced by factors such as recency, spending, and tenure. To address this, the magazine company should focus on building long-term relationships with customers, as longer tenure correlates with higher subscription likelihood. Targeted campaigns for high-spending customers and those accepting offers could also drive subscriptions. Additionally, reconsidering the timing of subscription offers may help mitigate the negative impact of recent interactions. Personalizing marketing efforts based on factors like education and marital status can further improve subscription rates and retention.

# Reference

National Bureau of Economic Research. (2019). Impact of customer behavior on subscription services. https://www.nber.org/research-impact-customer-behavior-subscription-services

OpenAI. (2025). *ChatGPT* [Large language model]. https://chatgpt.com

Smith, J. A., & Brown, M. L. (2020). Factors affecting subscription behavior in digital media markets. Journal of Media Economics, 33(2), 123-139. https://doi.org/10.1080/08997764.2020.1754282

# Appendix

```
    Feature Coefficient- SVM
0   Income  -0.250120
1   Total_Spending  1.019381
2   TotalPurchases  -0.853484
3   Campaign_Acceptance 0.648533
4   Tenure  0.679762
5   Recency -0.767351
6   NumWebPurchases 0.301450
7   Education_Basic -2.100477
8   Education_Graduation    -0.969176
9   Education_Master    -1.093992
10  Education_PhD   -0.401113
11  Marital_Status_Married  -1.671088
12  Marital_Status_Other    -1.000000
13  Marital_Status_Single   -0.857380
14  Marital_Status_Together -2.047686
15  Marital_Status_Widow    -1.565923
```