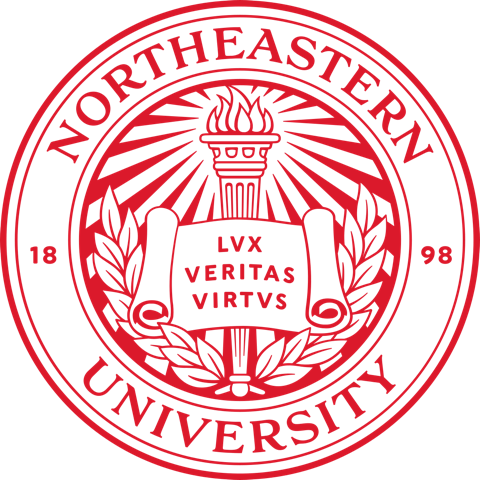
**Module 3 Project**

**Understanding Magazine Subscription Behavior**

Sunil Raj Thota



College of Professional Studies, Northeastern University ALY 6020 – Predictive Analytics

Prof. Justin Grosz

November 20, 2021

**Introduction:**

A magazine publisher is seeking to figure out why subscriptions dropped last year. They assumed that as people spend more time at home, they will spend more time reading. I'll use this data to help the company figure out what's working and what isn't. As a marketing analyst, the Chief Marketing Officer has informed me that recent marketing initiatives have not been as successful as projected. To comprehend this challenge and provide data-driven solutions, you must study the data collection.

**Part 1:**

To begin, we must cleanse the data and identify any outliers (if any) to ensure that we have high-quality data for the model. We will develop a logistic regression model and support vector machine classifier model to estimate customers' responses based on the numerous features in the dataset using optimization approaches to see if we can improve the model's accuracy and compare both models, as stated in the question. After evaluating the findings, we will advise the company on which factors contribute to an increase in acceptance. We've imported all of the packages and libraries we'll need for our initial data exploration. Using the scikit-learn to package, you can create, evaluate, and tune various classification models.

**Data Quality:**

To extract certain information from the dataset, the '?' sign is substituted by "NaN" during data filtering. Since the symbol has been changed with NaN, which stands for a null value, these can now be easily identified and computed to verify the sum of the missing values in the collection. There are 2240 records and 29 columns in the dataset. To get started, we've imported all of the packages that are required for undertaking model analysis. Pandas, NumPy, matplotlib, seaborn, train test split, Logistic Regression, SVM, and preprocessing are all examples of Python libraries.

**Data Cleansing, Preprocessing, and Exploratory Data Analysis:**

We tested for missing values in the dataset after putting it into a Python environment for further analysis. Missing values must be taken into account because they may have an impact on our analysis and AI models.

Let's see if we can convert the column using astype (). The term "outlier" refers to a data point that differs dramatically from the rest of the dataset. Anomaly in the distance between the values, to be precise. This can happen as a result of experimental errors or measurement variability. In the "Income" column, there are 24 null values that can be detected. However, because missing values are sometimes denoted as "Unknown" for categorical data or -1 for numerical data, we must still check for other missing values. There are no more missing values after inspecting each value count in categorical columns. Examining each of the numerical columns There isn't any unusually low or high value. We may deduce that the "Income" column is the only one with missing data. With a few exceptions, most incomes fall between $0 and $100,000. To avoid the influence of outliers on the imputation value, null values will be imputed with the median value. Outliers can be found in a variety of features (see the boxplots below), but Year Birth= 1900 is the only one that is likely to suggest data input problems. The DateTime format of the Dt Customer column should be used.

The sum of 'Kidhome' and 'Teenhome' can be used to calculate the overall number of dependents in the home ('Dependents'). The year that a customer became a customer ('Year Customer') can be calculated using 'Dt Customer'. The sum of all features containing the term 'Mnt' can be used to calculate the total amount spent ('TotalMnt'). The sum of all features containing the term 'Purchases' can be used to calculate total purchases ('TotalPurchases'). The sum of all features containing the keywords 'Cmp' and 'Response' can be used to calculate the total number of campaigns accepted ('TotalCampaignsAcc') (the latest campaign).

We investigate the distribution of mpg more fully because several classification algorithms rely on a logit relationship between features and target. To find patterns, we'll look for feature correlations first. In the clustered heatmap below, positive correlations between characteristics appear red, negative correlations appear blue, and no correlation appears grey.

The following groups of connected features may be seen on this heatmap: The "High Income" cluster consists of the following individuals: The number of purchases ('TotalPurchases' and other 'Num...Purchases' attributes) and the amount spent ('TotalMnt' and other 'Mnt' features) are both positively connected with 'Income'. The number of purchases made in a store, online, or through a catalog ('NumStorePurchases,' 'NumWebPurchases,' and 'NumCatalogPurchases') is positively connected with 'Income.' The 'Have Kids & Teens' cluster: The amount spent ('TotalMnt' and other 'Mnt' features) and the number of purchases ('TotalPurchases' and other 'Num...Purchases' features) are inversely linked with the 'Dependents' feature (with a stronger effect from kids vs. teens)

The number of deals purchased ('NumDealsPurchases') is favorably associated to 'Dependents' (kids and/or teens) and adversely connected to 'Income.' Acceptance of advertising campaigns ('AcceptedCmp' and 'Response') are substantially positively associated in the "Advertising Campaigns" cluster. The "High Income" cluster shows a weak positive connection of advertising campaigns, while the "Have Kids & Teens" cluster shows a weak negative association. Surprisingly, an increase in the number of web purchases ('NumWebPurchases') does not correlate with an increase in the number of website visits in the previous month ('NumWebVisitsMonth'). Instead, the number of web visits per month ('NumWebVisitsMonth') is positively connected with the number of deals purchased ('NumDealsPurchases,' implying that deals are an effective technique to encourage internet purchases.

**Part 2:**

**Logistic Regression Model:**

To perform Logistic Regression analysis, we need to change the categorical variables to hardcoded integers. Here, we have performed this technique for Education and Marital Status. In the same way that linear regression finds an equation that predicts an outcome for a binary variable, Y, from one or more response variables, X, logistic regression (LR) finds a solution that predicts an outcome for a binary variable, Y, from one or more dependent variable, X. The dependent variables, unlike linear regression, can be categorical or continuous, as the model does not require continuous data. When contrasted to other supervised classification methods like kernel SVM or ensemble algorithms, logistic regression is comparatively quick, but its accuracy suffers. For this, let split the dataset into independent and dependent variables. We will be using the Logit function to fit the model with the necessary variables to see the p values and other statistics of each column. I also went ahead and visualized the correlation between the variables using a corr heatmap. Used liblinear as the solver in the Logistic Regression method observed the model fit. And then I have calculated the Accuracy, Precision, Recall, and F1-Scores. The ratio of correctly classified subjects to the total number of subjects is known as accuracy. The ratio of accurately +ve labeled to all +ve labeled is known as precision. A recall is the proportion of those who are in reality to those who are appropriately +ve classified. Precision and recall are both taken into account while calculating the F1 Score. It's the precision and recall's average.

The top 3 significant variables are Marital\_Status\_Single, Education\_PhD, and NumWebVisitsMonth. From this, we can understand that there is more impact by the singles who stay at home due to the pandemic last year. Most of the education was being held online so the number of hits has increased over the months. We can more focus on these 3 customer targets and provide personalized discounts and offers to successfully churn them to paid subscribers. Also, the quality of content should be focused as the websites see more hits from the Ph. D. level educators and single customers.

**Part 3:**

**SVM Classification Model:**

SVMs are a type of machine learning approach that is widely used. They are part of the family of generalized linear models, which use the value of a linear combination of input features to make a classification or regression decision. SVM generates mathematical functions that translate input variables to desired outputs for classification or regression prediction issues using historical data and supervised learning methods. Using the five most frequent kernel functions stated above, we investigate SVM performance. We evaluated the SVM's performance using a variety of hyper-parameters, some of which are unique to the kernel function.

AcceptedCmp4, AcceptedCmp5, and NumWebVisitsMonth are the most significant variables in the SVM model. From this, we can understand that there is more impact by these respective campaigns where a greater number of customers are interested to subscribe to the magazine and accept the campaign. To increase more sales, we can focus on creating and establishing customized marketing campaigns for the customers which will highly positively impact the business. As these campaigns were positive and made users visit more often on a recurring monthly basis, it is essential to retain these customers with special offerings and target others.

**Part 4:**

**Comparisons, Findings, and Recommendations:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Speed** | **Accuracy** | **Precision** | **Recall** | **MSE** |
| Logistic Regression | 0:00:0.06 | 0.930 | 0.886 | 0.587 | 0.07 |
| SVM Classfication Model | 0:32:18.2 | 0.866 | 0.727 | 0.1 | 0.133 |

In both, models, NumWebVisitsMonth is one of the majorly contributing to the prediction if a customer has accepted the offer in the last campaign. So, I would recommend this company to focus more on the Previous Campaigns, Education, and Marital Status in providing personalized recommendations and offers. From the above results, I would recommend Logistic Regression as the best model to go with because of its Speed, Accuracy, Recall, and Precision Scores. When coming to the significant variables in both the models, Education\_PhD, Marital\_Status\_Single, NumWebVisitsMonth, AcceptedCmp4, and AcceptedCmp5 are the more significant variables in converting this campaign into a successful one. As these are working, the company has to change its marketing strategy and implement a detailed data-driven decision-making analysis and recommendations to the users.

**Conclusion:**

The most current advertising campaign (column name: Response) was the most successful, and it was especially successful in the United States. Acceptance of advertising campaigns is positively associated with income and adversely connected to having children or teenagers. Create two targeted advertising campaigns, one for high-income persons without children or teenagers and the other for lower-income folks with children or teenagers. Wines and meats are the most popular products (i.e. the average customer spent the most on these items). Concentrate your marketing efforts on increasing sales of the things that aren't as popular.

Deals and catalog purchases are underperforming channels (i.e. the average customer made the fewest purchases via these channels). Web and retail transactions are the most successful channels (i.e., the average client made the most purchases through these channels). To reach more customers, focus your campaigns on the most successful channels.

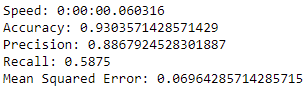
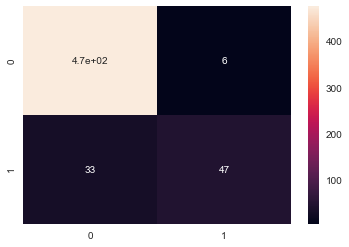
**References:**

Nicolas Bertagnolli. (Oct 12, 2020). How to Get Feature Importances from Any Sklearn Pipeline. *towardsdatascience*.

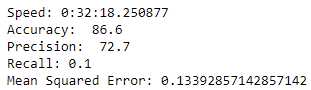
Retrieved from https://towardsdatascience.com/how-to-get-feature-importances-from-any-sklearn-pipeline-167a19f1214

**Appendix:**

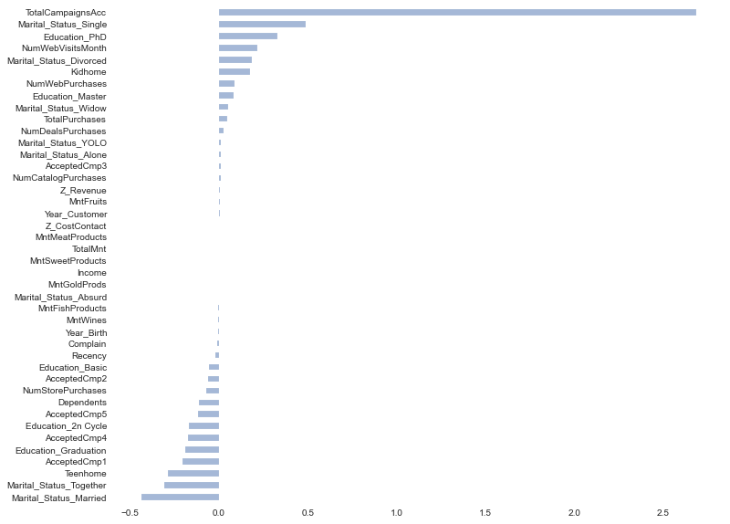
**Figure 1: Logistic Regression Results**

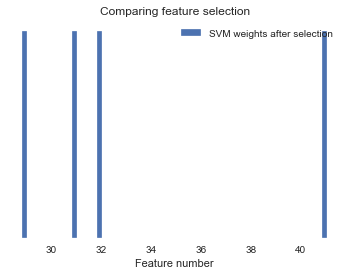
**Figure 2: SVM Classification Results**

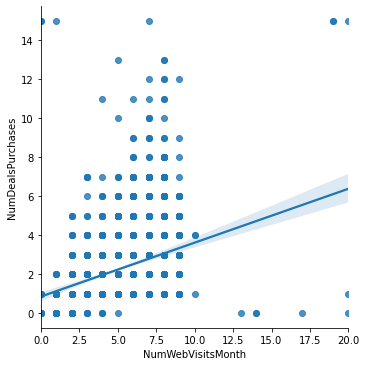
**Figure 3: Logistic Regression Feature importances**

****

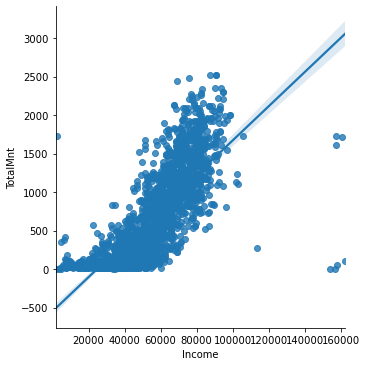
**Figure 4: SVM Classification Feature importances**

****

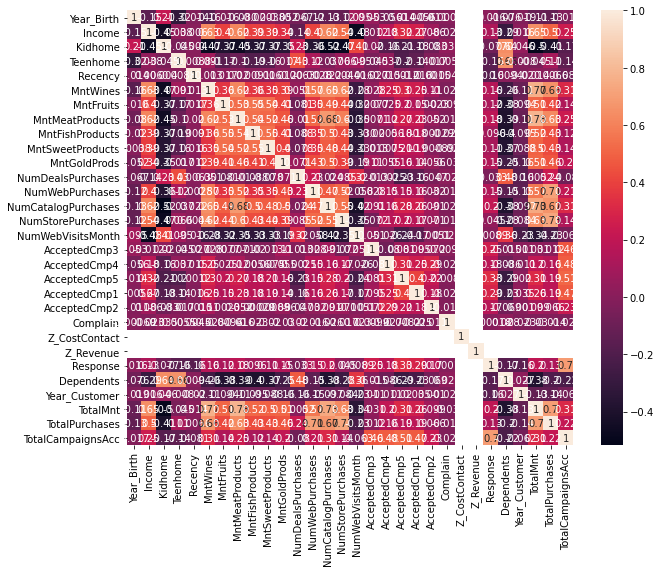
**Figure 5: NumWebVisitsMonth vs NumDealsPurchases**

****

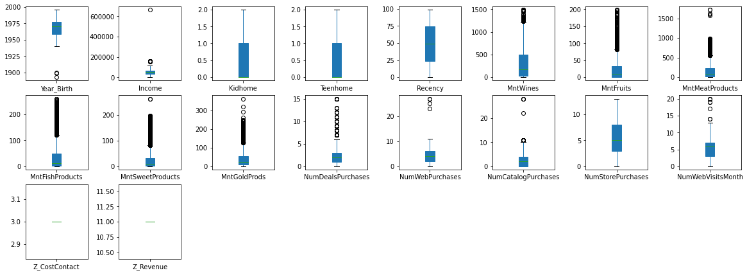
**Figure 6: Income vs TotalMnt**

****

**Figure 7: Correlation Plot**

****

**Figure 8: Columnwise Plots**

****