**ALY6020 Predictive Analytics**

Dr. Thomas Goulding

Module 3

Marketing Campaign Project

Jeff Hackmeister

6/12/2025

**INTRODUCTION**

This study is working with magazine subscriber data assembled by a marketing team to decide the main causes of a recent year-over-year decline in subscribers. The provided dataset contained observations on 2,240 customers with information on their recent purchasing behavior as well as several household demographics. From this data, two different classification models were created with the goal of determining characteristics correlated with subscribers and to guide future marketing decisions in targeting those most likely to subscribe.

**EXPLORATORY DATA ANALYSIS AND CLEANING**

To begin with, the needed Python packages were imported and the data read in from the provided Excel spreadsheet.

A screenshot of a computer

AI-generated content may be incorrect.

After exploring the basic shape of the dataset, 24 null values were found in the Income variable. Given that this is a very small percentage of the dataset (roughly 1% of all observations) those 24 observations were dropped. This left a dataset of 2,216 observations of 29 variables. To explore the distribution of the included variables, a series of histograms was produced (see Image 1 in the Appendix). From this exploration, the 5 binary variables for accepted comps along with customer cost and revenue were dropped from the study (see Image 2).

Next, to evaluate potential outlier values, a series of boxplots were produced (see Image 3). While many of the variables contained reasonable outlier values, the outliers seen in Year\_Birth and Income were singled out.

A diagram of a number of numbers and a number of numbers

AI-generated content may be incorrect.

After further exploration, birth years before 1940 were dropped along with incomes over $200,000. The variables for Education and Marital\_Status were also cleaned to ensure clear labels and clean coding for each category. For Education, this meant applying numeric values to the categories.

A computer screen shot of a computer screen

AI-generated content may be incorrect.

While for marital status, similar concepts were combined for clarity and then numerically coded using one hot method. The “divorced” option was dropped so that it could serve as our reference point to avoid issues with multicollinearity later in the analysis.

A screen shot of a computer code

AI-generated content may be incorrect.



Further data cleaning techniques were used to prepare for the modeling phase. The Dt\_customer variable (the year a customer first subscribed) was combined with the Year\_Birth variable to calculate the age (in years) a customer first subscribed, those two input variables were dropped to also combat multicollinearity.

**A screenshot of a computer code

AI-generated content may be incorrect.**

A correlation heatmap was then composed (see Image 4). High levels of correlation were seen with the various categories of historical purchases, as well as between the count of kids and teens in the home. Therefore, the decision was made to combine these similar variables into composite variables named Total\_Spend and Family\_Size.

****

Then a new correlation heatmap was produced.

A screenshot of a computer screen

AI-generated content may be incorrect.

**LOGISTIC REGRESSION MODEL**

With a properly cleaned and prepped data set, the next task was to create distinct test and training sets with an 80/20 split and then to fit a logistic regression model to the training set.

**A computer code with black text

AI-generated content may be incorrect.**

Using the model.score() function, an accuracy of 85.47% was calculated.

****

Then predictions could be made on the testing data with this model. The accuracy achieved on the test data was 85.53%.

A screenshot of a computer code

AI-generated content may be incorrect.

The confusion matrix from these predictions illustrates both the strengths as weaknesses of the model.

A blue squares with white text

AI-generated content may be incorrect.

There were 362 true negatives from the 443 predictions, demonstrating the predictive accuracy of the model but also the high bias towards negative results. This can be further illustrated in the classification report.

A screenshot of a computer

AI-generated content may be incorrect.

Where the diagnostic scores for the predictions of not subscribing are quite good while the scores for the predicted subscribers are much lower. This is an indication of overfitting in the model but also a reflection of the high rates of non-subscribers in the original data.

Looking at the most predictive variables, we can see the store purchases, married status, recency of offer and customer tenure had the largest negative impact of subscriptions.

A graph with blue squares

AI-generated content may be incorrect.

The performance of the web visits and purchasing variables along with education and single status should provide the marketing department with a customer profile for potential targets of single, well-educated consumers who do most of their shopping online.

**SUPPORT VECTOR MACHINE**

To further our study of this dataset, a second predictive model will be fit to the data, a support vector machine (SVM).

Using the same training and testing data was used for the logistic regression model, a SVM model was fit and then used to make predictions from the testing data. Given the nature of the SVM model, a scaler was first used for all the variables in the dataset. This ensures features with larger numerical values, like income, do not overshadow variables like family size with lower values but potentially greater predictive value.

A computer code with black text

AI-generated content may be incorrect.

A screenshot of a computer program

AI-generated content may be incorrect.

Then, a similar classification report was produced.

A screenshot of a computer screen

AI-generated content may be incorrect.

While the SVM shows more balanced precision/recall scores, the overall discriminative power (AUC: 0.650) is lower than logistic regression (AUC: 0.847)

Looking at the predictive significance of each variable, we can see that family size has the largest impact on the model.

A screenshot of a computer

AI-generated content may be incorrect.

Family size and total spend were also highly influential in this model.

A graph of a bar graph

AI-generated content may be incorrect.

**INTERPRETATION AND RECOMMENDATIONS**

The two different modeling techniques each produced highly accurate predictive models for determining whether a customer would positively respond to a marketing campaign for magazine subscription. While each model has some unique nuances, the guidance to the marketing department is clear. Wealthier, better educated, single consumers are much more promising candidates for future marketing campaigns. Given the data provided for this study, that is the demographic that should be targeted in future efforts.

Given the high rate of negative response in the dataset, the next steps for this process should be to gather a new dataset of similar variables on all existing customers to level set the customer profile. Ideally, then the marketing department could conduct an A/B test for a new campaign with one targeted directly at the identified demographic and another more general campaign and the results could be compared for return on investment.

**REFERENCES**

[1] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning : Data Mining, Inference, and Prediction, Second Edition*. Springer New York.

[2] OpenAI. (2024). ChatGPT (Claude 4) [Large language model]. https://claude.ai

‌[3] Steele, B., Chandler, J., & Reddy, S. (2016). *Algorithms for Data Science*. Springer.

**IMAGE APPENDIX**

Image 1

A screenshot of a graph

AI-generated content may be incorrect.

Image 2

A screenshot of a graph

AI-generated content may be incorrect.

Image 3

A chart of a graph

AI-generated content may be incorrect.

Image 4

A screenshot of a chart

AI-generated content may be incorrect.