Francesca Marchese, Ian Smith, Emmanuella Acheampong, Andrew Mukurazita

BUS-212A- Report 2a

BUS 212A – Advanced Data Analytics

Report 2a – Supervised Models

March 3, 2024

Table of Contents

Project Overview

Our project focuses on leveraging data analytics to unlock valuable insights into customer behavior for MarketSphere Inc., a leading retail company in the United States. Our consultancy aims to help MarketSphere in identifying growth opportunities and refining its marketing strategies to better target its customer base.

Progress Update

In our first report, we made significant strides in ensuring the quality of our dataset. Beginning with 2241 unique records, we meticulously refined them down to 2237, by eliminating very extreme outliers which could impact our model's performance. These include three individuals who were over 100 years (born in the 19th century) and the one individual earning above $700,000 because the behaviors of these customers are not likely to be representative of the overall sample.

We also enhanced data completeness and accuracy through rigorous preprocessing. This involved standardizing data types, clarifying column names, and removing redundant variables. Additionally, we addressed missing values through imputation and handled outliers using Winsorization techniques. Through these efforts, we have transformed our dataset into a clean and reliable resource, poised for in-depth analysis and modeling.

Introduction to Report 2a

In continuation of our data consultancy project with MarketSphere Inc., we focus on implementing supervised learning models to gain deeper insights into the company's customer base in Report 2a. The primary objective of this assignment is to apply regression and classification techniques to our dataset, thereby uncovering valuable patterns and relationships that can inform strategic decision-making.

Purpose and Scope:

By leveraging techniques such as Logistic Regression, K-Nearest Neighbors (K-NN), and Multiple Regression, we aim to develop models that neither underfit nor overfit the data.

Part 1. Descriptive, Exploratory Analysis

Regression

Justification for Regression Target Variable Selection ("Wines"):

In the context of Market Sphere's retail analytics, 'Wines' emerged as the primary product of interest due to its status as the most frequently purchased item among their diverse product range with significant contribution to the total revenue and encapsulates varied customer behaviors and preferences. This variable is an excellent candidate for regression due to its continuous numeric nature. Accurate prediction of wine purchases is crucial for Market Sphere as it aids in effective inventory management, targeted marketing strategies, financial planning, and enhances overall customer satisfaction and profitability.

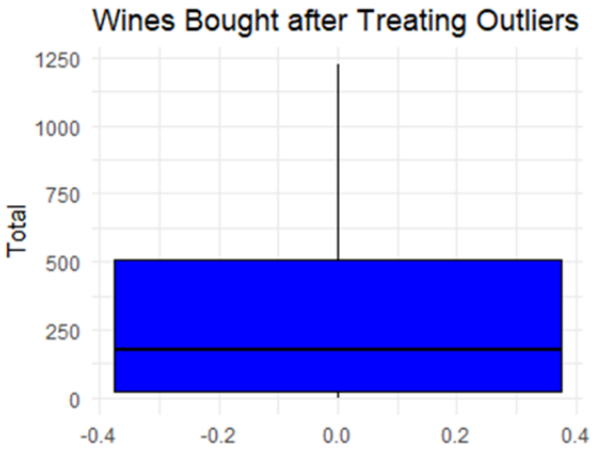
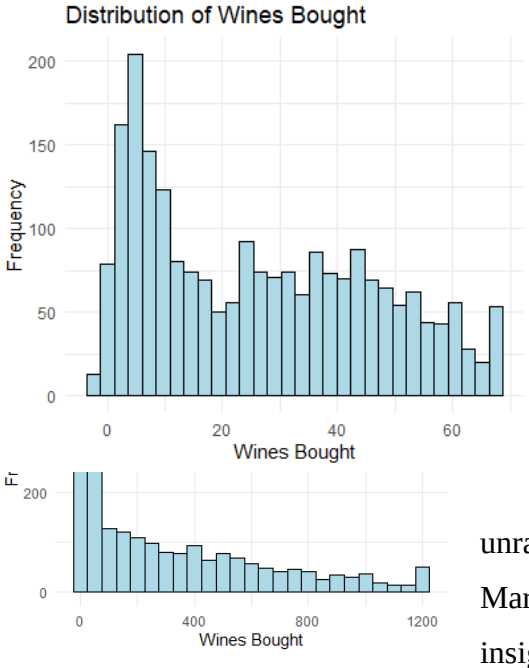


Figure 7 - Total Wines bought without outliers.

The summary statistics for 'Wines' purchases, as exhibited by the dataset encompassing 2,236 data points, demonstrates a broad spectrum of customer spending. Specifically, the data ranges from a minimum of $0 (indicating customers who did not purchase wines) to a maximum of $1,224.6, with a median value of $174. This median, in addition to the mean value of $302.3, illustrates the skewed nature of wine purchasing behavior: a significant number of customers make modest wine purchases, while a smaller, affluent segment contributes to higher sales volumes. The first and third

quartiles, at 24 and 504.2 units, respectively, further highlight the diverse spending patterns among customers.

This skewed distribution is critical for regression analysis. However, linear regression models, the chosen method for this analysis, assume normality in the distribution of the dependent variable. Therefore, the initial skewness in 'Wines' distribution can lead to biased estimates, affecting the model's reliability and accuracy.



To address this, a Box-Cox Transformation was applied to the 'Wines' variable, aiming to normalize the distribution and stabilize variance. This methodological approach is particularly effective in handling skewed data, ensuring that our regression models can produce more accurate and interpretable results.

The histograms (before and after transformation) serve as visual confirmations of the distribution's adjustment. With a focus on 'Wines' as our target variable, we aim to unravel the underlying factors influencing wine purchases at Market Sphere. Understanding these can provide actionable insights into customer preferences, enabling more targeted marketing strategies and optimized inventory management, ultimately fostering enhanced business performance and customer satisfaction.

Feature Selection

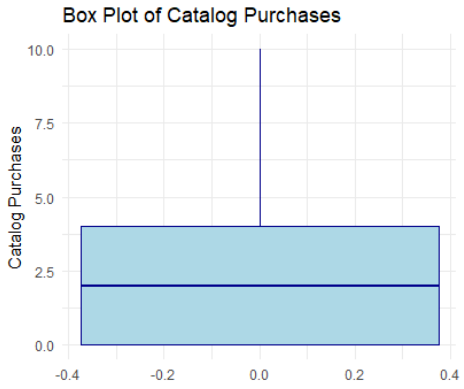
Model Building and Selection

An exhaustive search was conducted to identify the best subset of predictors for our linear regression model. Though about 15 variables were arrived at during the search, the final variables were chosen based on their VIF, relevance and statistical significance. The final model included 'CatalogPurchases', 'StorePurchases', 'Kidhome', and 'Complain' as predictors. These variables were

identified as having a strong correlation with the target variable, while minimizing multicollinearity among themselves with VIFs (1.79, 1.81 and 1.50)

Plots for Visual Correlation and Further Statistics of Predictors

1. CatalogPurchases:



Minimum: 0 (some customers did not make any purchases through catalogs)

Maximum: 10 (some customers made up to ten catalog purchases)

Mean: 2.625 (on average, customers made approximately three catalog purchases)

Median: 2 (half of the customers made two or fewer catalog purchases)

Standard Deviation: The range and interquartile values indicate variability in catalog purchase behavior among customers.

2. StorePurchases:

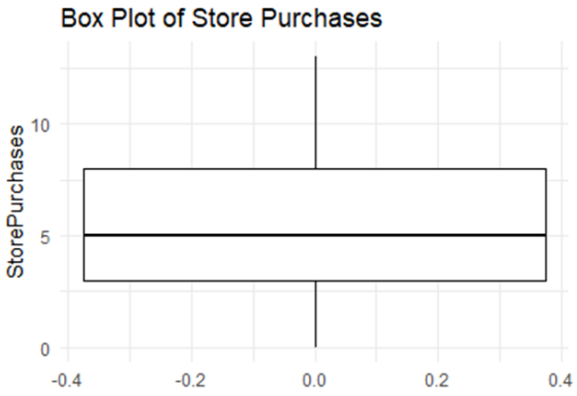


Figure 17 - Store purchases

Minimum: 0 (some customers did not make any in-store purchases)

Maximum: 13 (a few customers made up to thirteen purchases in-store)

Mean: 5.796 (on average, customers made about six in-store purchases)

Median: 5 (half of the customers made five or fewer in-store purchases)

Standard Deviation: Akin to CatalogPurchases, variability is evident given the spread between the minimum and maximum.

3. Kidhome:

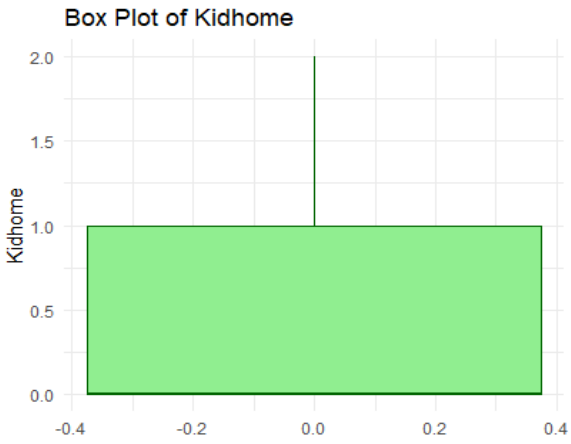
Minimum: 0 (some customers have no children at home)

Maximum: 2 (the highest number of children at home reported by customers)

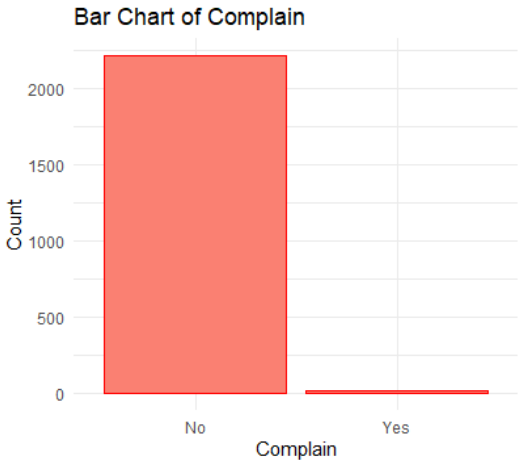
Mean: 0.4441 (indicating that, on average, less than one child is present in customers' homes)

Median: 0 (more than half of the customers reported having no children at home)

Standard Deviation: Indicate a distribution leaning towards fewer children at home.



4. Complain:



Minimum: 0 (indicating that most customers did not register complaints)

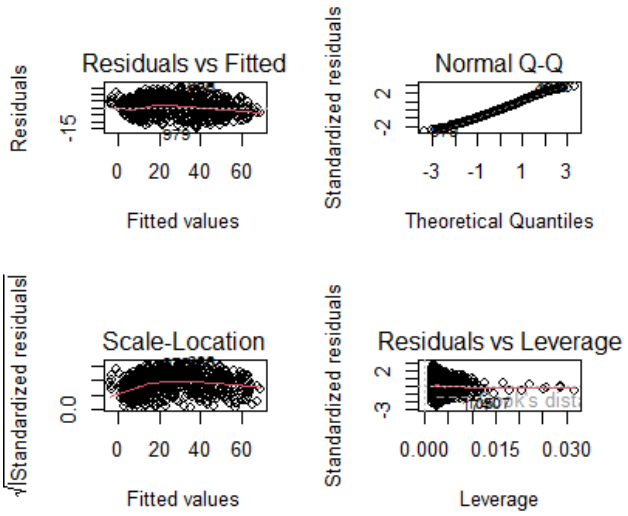
Maximum: 1 (20 complains were made in total the over 2 years)

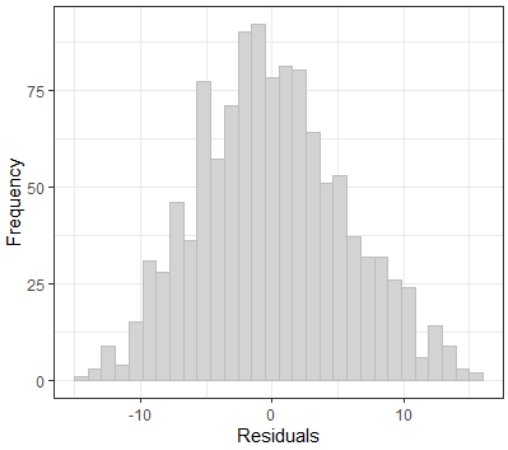
Mean: 0.008944 (very few customers have lodged complaints)

In examining the relationships between 'Wines' and the predictors, we observed that 'Wines' spending positively correlates with 'CatalogPurchases' and 'StorePurchases', indicating that customers who spend more through catalogs or in stores tend to also spend more on wines. Conversely, there is a negative correlation between 'Wines' and 'Kidhome', suggesting that having more children at home may lead to reduced wine expenditures. The variable 'Complain' does not exhibit a clear relationship with 'Wines', implying that customer complaints may not significantly influence wine purchasing habits. Additionally, there doesn't appear to be a relationship between the predictors, indicating that the predictors are identical and independent (i.i.d linear regression assumptions) of each other in terms of customer preference.

Diagnostics:

The above diagnostic plots were reviewed to assess the validity of the linear regression assumptions.





a) The histogram of residuals indicates a Gaussian distribution, supporting the linear regression model's validity, suggesting that the model's errors are randomly distributed and homoscedastic, thus validating the fit of the model to the data.

b) The *Normal Probability Plot of the Residual* confirms that the residuals are approximately normally distributed, satisfying one of the key assumptions of linear regression.

c) In the *Residuals vs. Fitted Values* plot, the residuals appear randomly scattered around the horizontal line at zero, without any apparent pattern. This randomness suggests that the linear model fits the data well without obvious violations of homoscedasticity or indications of non-linear relationships.

Model Performance:

The best model yielded an adjusted R-squared value of 0.8428, indicating that approximately 84.28% of the variance in wine purchases can be explained by the model. The RMSE (Root Mean Square Error) for the training data was approximately 7.43, indicating the typical deviation from the observed values.

Diagnostic Residual Plots and Outliers:

The diagnostic plots provided evidence that the assumptions of linear regression were largely met. A total of 190 observations, representing 8.5% of the dataset, were identified as outliers and removed based on Cook's distance. This percentage seems acceptable considering the aim was to improve model accuracy without significantly reducing the dataset's size.

Higher-Order Terms:

The inclusion of polynomial terms for 'CatalogPurchases' and 'StorePurchases' did not significantly improve the model's performance. Therefore, the final model excluded these higher-order terms as they contributed to model complexity without clear benefit. The final model did not significantly benefit from higher-order terms, as indicated by similar adjusted R-squared and RMSE values. Hence, the simpler model was preferred for interpretation and application.

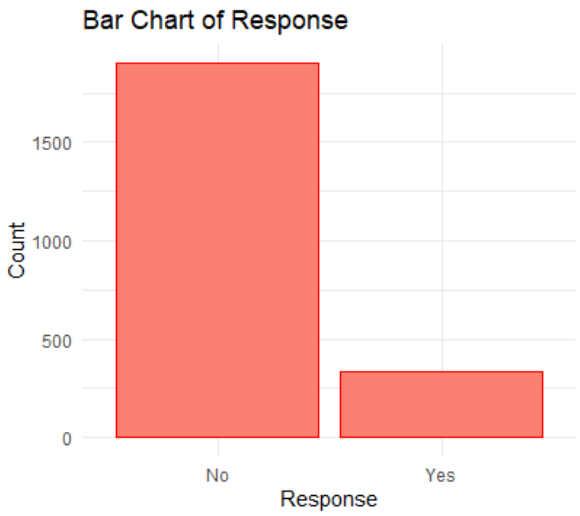
Model Interpretation and Business Insights:

The final model suggests that for every additional catalog purchase, wine purchases increase by approximately 3.68 units, holding other factors constant. Similarly, every additional store purchase is associated with an increase of about 3.04 units in wine purchases. However, having one more child at home is associated with a decrease of approximately 2.91 units in wine purchases.

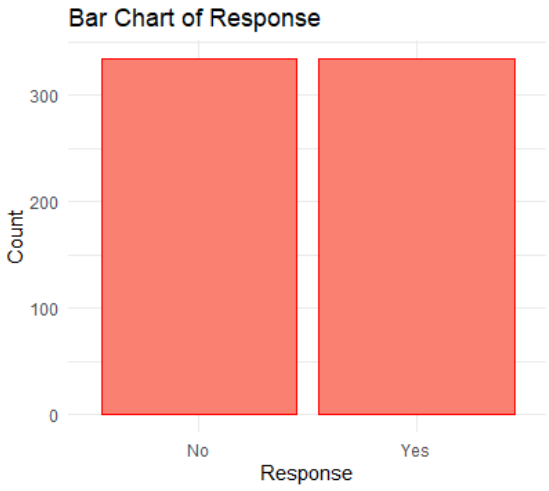
These findings provide valuable insights into customer behavior. Specifically, they highlight the importance of catalog and store channels in driving wine sales and the negative impact of having children on wine purchasing. This information can guide targeted marketing campaigns, inventory planning, and sales strategies, ultimately enhancing profitability and customer satisfaction.

Classification

Justification for Classification Target Variable Selection ("Response")



The selection of "Response" as the target variable for classification analysis is driven by several strategic considerations. The Response variable is a categorical variable which indicates whether a customer responded to the final campaign or not. 1 indicates acceptance, while 0 indicates otherwise. Our exploratory analysis revealed significant correlations between "Response" and various predictor variables, indicating its importance in understanding customer behavior and engagement levels. In a typical business setting, knowing the response rate to marketing campaigns is essential for evaluating campaign effectiveness and optimizing marketing strategies. By accurately classifying customer responses, MarketSphere Inc. can identify target segments that are more likely to engage with marketing initiatives, tailor communication strategies to their preferences, and ultimately enhance campaign ROI and customer satisfaction. The distribution of Response was imbalanced with only 334 out of 2236 customers (representing 14.95%) responding 1(yes) to the last campaign.



In order to prevent this imbalanced distribution from affecting the model's accuracy, we undersampled the majority class (0 responses) to 334 (just as the number of 1 responses) using the RAND function in excel. Finally, we added the sampled points to the minority class and had a dataset of 668 for modeling.

Density Plot of Days Since Last Purchase Vs. Campaign Response:

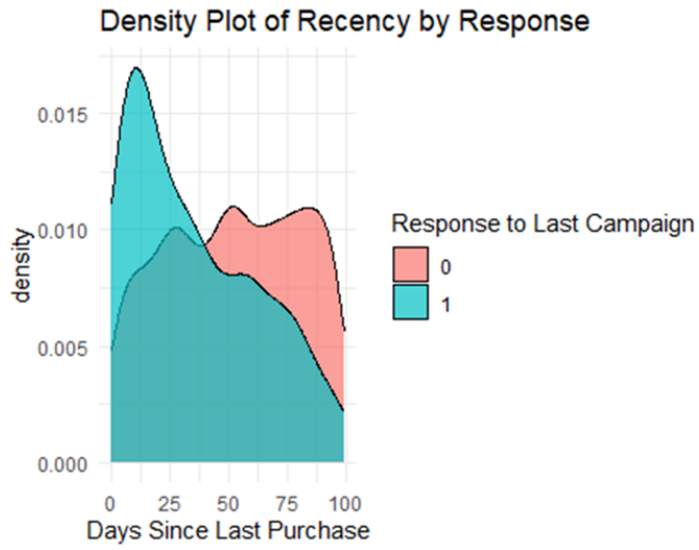


Figure 33 - Days of last purchase vs. Response

The "Density Plot of Recency by Response" illustrates the relationship between the number of days since the last purchase (recency) and the response to the last campaign. Customers with a lower number of days since their last purchase (more recent purchases) have a higher positive response (response '1') to the campaign. Conversely, the density of non-responses (response '0') increases as the number of days since the last purchase grows, implying that those who haven't made a purchase recently are less likely to respond to the campaign. This trend aligns with the idea that more recently engaged customers are more receptive to marketing efforts. For MarketSphere, focusing on customers with recent interactions may yield better campaign response rates.

CDF of Income Vs. Campaign Response

The blue line (response '0') and the red line (response '1') indicate the proportion of individuals at or below certain income levels who did not respond and who responded positively to the campaign, respectively. The red line (response '1') is consistently below the blue line (response '0'), suggesting that individuals with higher incomes are less likely to respond positively to the campaign. Both lines rise steeply in the lower income range and then gradually level off, indicating that most of the population falls within a middle-income bracket. The gap between the two lines narrows as income increases and widens in the

middle, showing that the difference in response rates between lower and higher-income individuals vary at different levels of income.

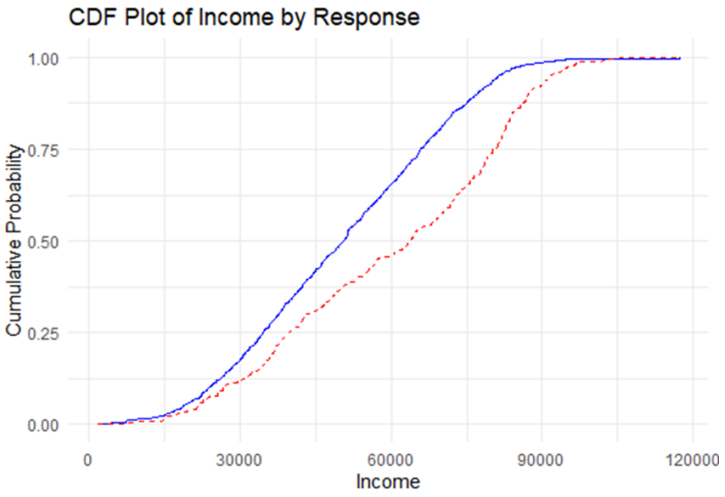


Figure 34 - Income vs. Response

This pattern might imply that campaign messaging or offers are more appealing or relevant to those in the lower income brackets, or that those with higher incomes are less influenced by this particular campaign. This insight could be used to adjust campaign targeting or to tailor messages to be more effective for higher-income brackets.

Logistic Regression

Model Complexity:

```
> ##adding some interaction terms
> data$DiscStore = data$DiscountPurchases * data$StorePurchases
> data$WebCatalog = data$WebPurchases * data$CatalogPurchases  # Another interactio
n term
> data$IncomeSqrd = data$Income^2  # A polynomial term (square of Income)
> data$sqtMeats = sqrt(data$Meats)  #  polynomial term (square root of Wines)
> str(data)
'data.frame':   668 obs. of  23 variables:
 $ Teenhome                          : int  1 1 0 1 1 0 1 0 0 0 ...
 $ Meats                             : int  1 11 168 17 13 24 10 556 403 403 ...
 $ LastPurchase                      : int  1 93 47 24 15 87 43 2 9 9 ...
 $ DiscountPurchases                 : int  1 2 1 2 2 1 2 1 1 1 ...
 $ CatalogPurchases                  : int  0 0 4 0 1 0 1 4 6 6 ...
 $ WebVisitsMonth                    : int  6 4 1 7 4 8 5 8 3 3 ...
 $ WebPurchases                      : int  1 2 3 2 1 2 2 10 7 7 ...
 $ StorePurchases                    : int  2 3 7 3 4 3 4 8 6 6 ...
 $ AcceptedCmp4                      : int  0 0 0 0 0 0 0 0 1 1 ...
 $ AcceptedCmp3                      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ AcceptedCmp5                      : int  0 0 0 0 0 0 0 0 1 1 ...
 $ AcceptedCmp1                      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ AcceptedCmp2                      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Education_non_graduate            : int  1 0 1 1 1 1 1 1 1 1 ...
 $ Marital_Status_Married            : int  0 0 1 0 1 0 1 0 1 1 ...
 $ Marital_Status_Together           : int  1 1 0 0 0 1 0 1 0 0 ...
 $ Recency_Intervals_More than 90 days: int  0 1 0 0 0 0 0 0 0 0 ...
 $ Response                          : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 2
2 2 ...
 $ Income                            : num  34578 42554 59973 38054 42769 ...
 $ DiscStore                         : int  2 6 7 6 8 3 8 8 6 6 ...
 $ WebCatalog                        : int  0 0 12 0 1 0 2 40 42 42 ...
 $ IncomeSqrd                        : num  1.20e+09 1.81e+09 3.60e+09 1.45e+09 1.
83e+09 ...
 $ sqtMeats                          : num  1 3.32 12.96 4.12 3.61 ...
```

Our logistic regression model's complexity is augmented by the inclusion of interaction terms and polynomial transformations to capture the multifaceted nature of the relationships between the predictors and the target variable. These enhancements provide a more nuanced view of the predictors' effects and their synergies.

Interaction terms in the model allow for the assessment of whether the effect of one predictor on the outcome is modified by the level of another predictor. In our model, the interaction between 'DiscountPurchases' and 'StorePurchases' generates a new predictor, *'DiscStore'*. This term is designed to investigate if the combined effect of these two types of purchases on the response is different from what would be expected from their independent effects. Similarly, the *'WebCatalog'* term, created by multiplying 'WebPurchases' and 'CatalogPurchases', is intended to capture any synergistic effect between online browsing and catalog shopping that might influence the response. The rationale behind these terms is grounded in consumer behavior theory, where purchasing patterns are often interrelated and the impact

of a marketing campaign might be amplified or diminished by the interaction between different purchasing channels. For instance, a customer who frequently shops both online and through catalogs might be more receptive to certain types of marketing strategies, which is an effect that would be missed without the inclusion of the *'WebCatalog'* interaction term.

Polynomial terms allow the model to fit non-linear relationships between predictors and the response variable. The term 'IncomeSqrd' is a quadratic transformation of the 'Income' variable, recognizing that the relationship between income and response might not be linear. Our reasoning is that this squared term can capture the diminishing or increasing marginal effect of income on the likelihood of a response. For example, it might be that as income increases, the probability of a positive response increases at a decreasing rate. The transformation *'sqtMeats'*, the square root of the 'Meats' variable, is another polynomial term we included to handle potential non-linearity. The square root transformation is particularly useful when dealing with right-skewed distributions, as it can stabilize variance and make the data conform more closely to the assumptions of the logistic regression model. By transforming the 'Meats' variable in this way, we account for the diminishing influence of meat purchases on the response as the quantity of meat purchased grows.

The use of these complex terms is justified by the likely relationships in the data. Consumer behavior is typically influenced by a range of interacting factors, and purchasing decisions may not always increase proportionally with income or frequency of purchases. The decision to include interaction terms is based on the hypothesis that the influence of certain predictors on the response is contingent on the levels of other predictors. The polynomial terms are included based on the assumption that the relationship between some predictors and the response is non-linear.

To summarize, the inclusion of interaction and polynomial terms in the logistic regression model enhances its predictive capabilities by allowing it to fit more complex relationships in the data. This complexity is necessary for capturing the subtleties and nuances of consumer behavior and response patterns. The model's ability to accommodate these intricacies can lead to more accurate predictions and a deeper understanding of the factors influencing the response variable.

Model Accuracy:

```
>  cm
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 220  49
         1  48 219

               Accuracy : 0.819
                 95% CI : (0.7838, 0.8507)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.6381

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.8209
            Specificity : 0.8172
         Pos Pred Value : 0.8178
         Neg Pred Value : 0.8202
             Prevalence : 0.5000
         Detection Rate : 0.4104
   Detection Prevalence : 0.5019
      Balanced Accuracy : 0.8190

       'Positive' Class : 0

> sensitivity = cm$byClass[1] # also known as recall
> specificity = cm$byClass[2] # also known as precision
> # Let's calculate F1, which combines sensitivity and specificity
> F1 = (2 * sensitivity * specificity) / (sensitivity + specificity)
> cat( "F1 statistic = ", round(F1,3))
F1 statistic =  0.819
```

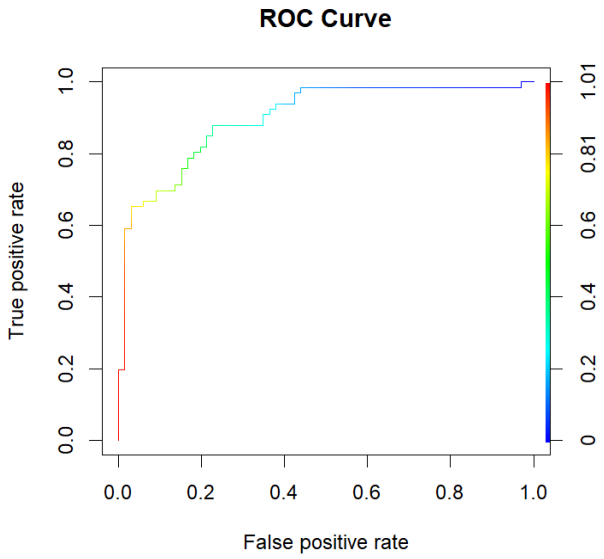
The accuracy metrics derived from the confusion matrix of the final logistic regression model present a robust picture of the model's predictive performance. The model's accuracy stands at 0.819, which is substantially higher than the No Information Rate of 0.5, indicating that the model predictions are significantly better than random chance. This is further supported by a Kappa statistic of 0.6381, which measures inter-rater agreement for categorical items and suggests a good level of agreement beyond chance. The sensitivity, or true positive rate, is 0.821, demonstrating the model's strength in correctly identifying positive instances of the target variable. This is particularly important in scenarios where the cost of false negatives is high. Specificity, or true negative rate, is similarly high at 0.817, indicating the model's efficacy in identifying true negatives. Both sensitivity and specificity are balanced, which is ideal for a model, suggesting that it does not overly favor predicting one class over the other. This balance is essential for maintaining the integrity of predictions across both classes of the target variable.

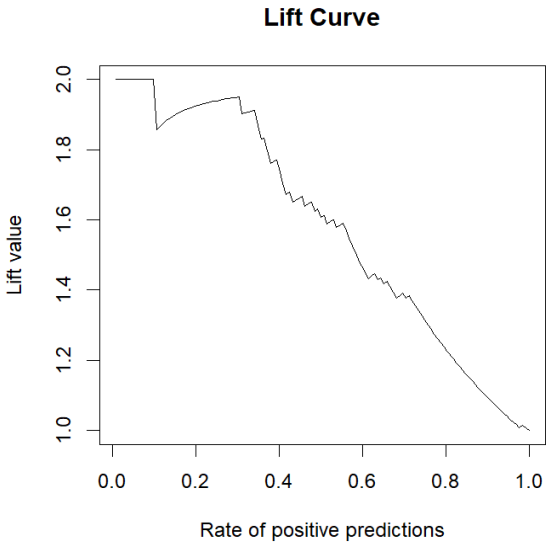
The F1 statistic, which is the harmonic mean of precision and recall, stands at 0.819, further indicating that the model maintains a good balance between sensitivity and specificity. This metric is particularly useful when the cost of false positives and false negatives is high or when classes are imbalanced.

```
                                        Pr(>|z|)
(Intercept)                             0.072622 .
Teenhome                                0.006997 **
Meats                                   0.900204
LastPurchase                            7.36e-08 ***
DiscountPurchases                       0.645818
CatalogPurchases                        0.085159 .
WebVisitsMonth                          2.20e-06 ***
WebPurchases                            0.462638
StorePurchases                          0.001353 **
AcceptedCmp4                            0.408290
AcceptedCmp3                            0.000131 ***
AcceptedCmp5                            5.55e-05 ***
AcceptedCmp1                            0.003729 **
AcceptedCmp2                            0.163109
Education_non_graduate                  0.002254 **
Marital_Status_Married                  1.23e-05 ***
Marital_Status_Together                 8.29e-07 ***
`\\`Recency_Intervals_More than 90 days\\`` 0.378494
Income                                  0.455037
DiscStore                               0.701426
WebCatalog                              0.981636
IncomeSqrd                              0.443797
sqtMeats                                0.289311
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

In the logistic regression model, the p-values help assess the statistical significance of predictor variables. Variables such as 'Teenhome', 'LastPurchase', 'WebVisitsMonth', 'StorePurchases', and several campaign acceptance indicators ('AcceptedCmp3', 'AcceptedCmp5', 'AcceptedCmp1') show p-values less than 0.05, denoting strong statistical significance. Particularly, 'AcceptedCmp3', 'AcceptedCmp5', along with 'Marital_Status_Married' and 'Marital_Status_Together' are highly significant with p-values less than 0.001, indicating robust evidence against their corresponding null hypotheses. Conversely, predictors like 'Meats', 'DiscountPurchases', and 'WebPurchases', among others, have p-values greater than 0.1, suggesting they may not contribute significantly to the model. It's crucial to consider that while p-values indicate the presence of an effect, they don't reflect its magnitude, and the relevance of statistically significant predictors should be carefully interpreted in the context of the domain and practical significance.

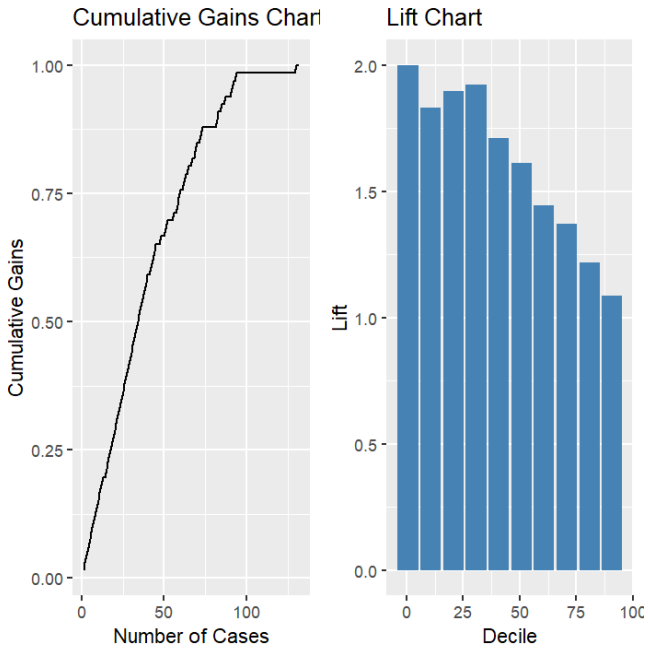
Plots:





ROC Curve: The True Positive Rate (TPR) on the ROC Curve's y-axis indicates the model's ability to correctly identify actual positives, while the False Positive Rate (FPR) on the x-axis shows how many actual negatives were incorrectly labeled as positive. The curve represents different sensitivity/specificity pairs for various thresholds, with the curve's proximity to the top-left corner denoting better model performance in class differentiation. The Area Under the Curve (AUC) quantifies this performance, where a value of 0.5 means no better than random chance, and 1.0 is perfect discrimination. The color gradient represents threshold changes, and selecting a point on the curve involves a trade-off between capturing true positives and avoiding false positives, dependent on the consequences of misclassifications. Overall, a quick rise of the curve towards a high TPR with a low FPR suggests good predictive power of our model.

Lift Curve: The lift curve shows that our model is highly effective at the outset, providing significant improvement over random selection by achieving a high rate of success in the top percentage of cases. As more cases are included, the effectiveness diminishes, converging toward a lift of 1, which is equivalent to a random guess. This is a common pattern where initial predictions are very productive, but less so as one reaches deeper into the pool of cases. The decline in lift suggests that there is a point at which targeting additional subjects by MarketSphere Inc. will become less cost-effective, and this can be used to determine an optimal cutoff for its marketing campaigns. In evaluating the model, the initial steepness of the lift curve is key; the model is considered more effective if it maintains a high lift over a larger segment of the population.



Cumulative Gains Chart: The cumulative gains curve illustrates the model's ability to correctly identify positive outcomes. On the y-axis, it shows the cumulative percentage of correctly predicted positives relative to the total actual positives. A sharp initial ascent indicates the model's high effectiveness, capturing a substantial share of positive outcomes within a small fraction of cases. As the curve extends to the right and begins to plateau, it reflects the diminishing returns of including more cases, as fewer remaining positives can be captured. The optimal model would display a step function-like steep initial rise if it could perfectly distinguish all positive outcomes from the outset.

Lift Chart: The lift chart, plotted with lift values on the y-axis against population deciles on the x-axis, illustrates the efficiency of the predictive model in distinguishing positive outcomes compared to random guessing. High lift values in the initial deciles demonstrate our model's strength in accurately predicting positive outcomes among the top-ranked individuals. However, as we progress through the deciles towards the right of the chart, we observe a natural decrease in lift, indicating a gradual reduction in the model's ability to identify additional positive outcomes as effectively, highlighting its predictive prioritization capability within the initial segments of the population.

Overall, the performance metrics suggest that the final logistic regression model is a reliable predictor of the target variable, with statistically significant predictors and high accuracy, sensitivity, and specificity.

Interpretation of Odds Ratios:

```
                                                Pr...z..    odds
(Intercept)                                      0.07262 0.13142
Teenhome                                         0.00700 0.39431
Meats                                            0.90020 0.99945
LastPurchase                                     0.00000 0.97432
DiscountPurchases                                0.64582 1.10218
CatalogPurchases                                 0.08516 1.28337
WebVisitsMonth                                   0.00000 1.60078
WebPurchases                                     0.46264 1.08301
StorePurchases                                   0.00135 0.75492
AcceptedCmp4                                     0.40829 1.51438
AcceptedCmp3                                     0.00013 4.59338
AcceptedCmp5                                     0.00006 9.38602
AcceptedCmp1                                     0.00373 6.38006
AcceptedCmp2                                     0.16311 5.66205
Education_non_graduate                           0.00225 2.18837
Marital_Status_Married                           0.00001 0.27090
Marital_Status_Together                          0.00000 0.20274
`\\`Recency_Intervals_More than 90 days\\``      0.37849 0.53306
Income                                           0.45504 0.99997
DiscStore                                        0.70143 1.01006
WebCatalog                                       0.98164 0.99950
IncomeSqrd                                       0.44380 1.00000
sqtMeats                                         0.28931 1.15220
>
```

In the logistic regression model, the odds ratios for the variables 'AcceptedCmp3' and 'AcceptedCmp5' are 4.59338 and 9.38602, respectively, indicating strong positive associations with the likelihood of a response. Specifically, customers accepting campaign 3 or 5 have 4.6 and 9.4 times higher odds, respectively, of a positive response than those who do not, ceteris paribus. Conversely, the marital status of customers significantly influences the response negatively; those who are married or living together have 73% and 80% lower odds of responding positively compared to their counterparts. For 'Income', the odds ratio is nearly 1, suggesting an insignificant impact on response likelihood, but the presence of 'IncomeSqrd' suggests the relationship may be non-linear. Lastly, an increase by one unit in the square root of the Meats variable, 'sqtMeats', increases the odds of a positive response by about 15.2%, controlling for other factors.

Interpretation of Confusion Matrix:

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 220  49
         1  48 219

               Accuracy : 0.819
                 95% CI : (0.7838, 0.8507)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.6381
```

The R output above presents our logistic regression model confusion matrix and its statistics, revealing its robust predictive performance. The model correctly identifies the majority of both negative and positive classes, with 220 true negatives and 219 true positives, and only 49 false negatives and 48 false positives, indicating a well-balanced prediction capability. With an accuracy of 0.819—significantly higher than the No Information Rate of 0.5—the model's predictions are statistically validated as significantly better than chance, as shown by the P-Value [Acc > NIR] of less than 2e-16. The 95% confidence interval for accuracy ranges between 78.38% and 85.07%, suggesting consistent performance across similar datasets. The Kappa statistic of 0.6381 further confirms the model's reliability, indicating a substantial level of agreement beyond random chance. Collectively, these metrics underscore the model's efficacy in delivering accurate and dependable classifications significantly surpassing random predictions.

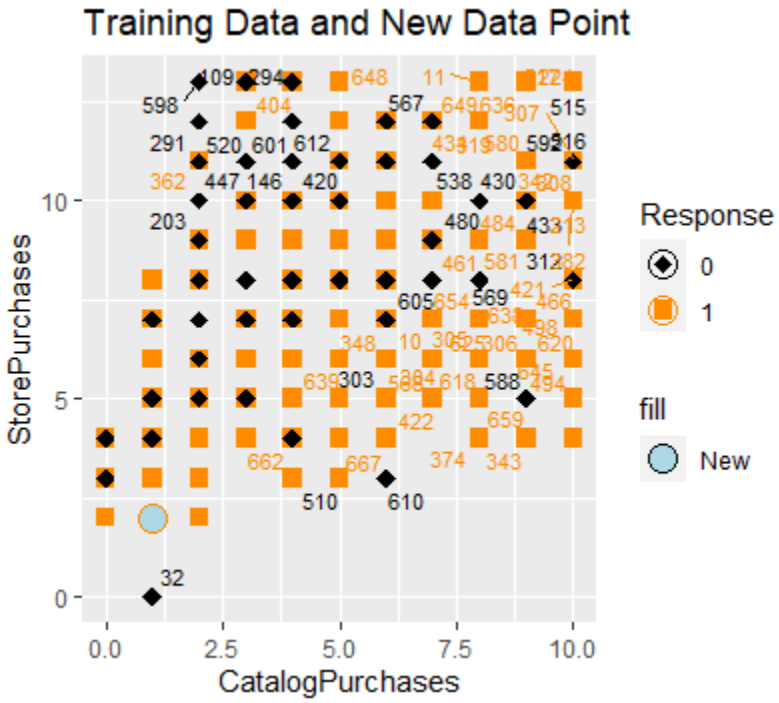
KNN Model

Variable Selection:

An exhaustive search approach was employed to determine the most impactful predictors for the response variable. This method systematically evaluated all possible combinations of predictors, optimizing based on criteria such as the smallest prediction error (Cp statistic). This rigorous variable selection process aimed to balance model complexity with predictive power, ultimately identifying a subset of variables contributing most significantly to predicting customer response.

Model Development:

The model was trained with normalized data to account for the varying scales of different features, crucial for the distance calculation inherent in k-NN. An exhaustive parameter tuning was conducted via Leave-One-Out Cross-Validation (LOOCV), optimizing the 'k' parameter to enhance model accuracy while preventing overfitting.



Methodology

was applied with k=3 to predict the response of a new customer (new neighbor), based on the proximity to existing data points in the feature space. The process involved scaling the dataset to normalize the range of different features, which is crucial for the distance-based K-NN algorithm. Subsequently, cross-validation was employed to identify the optimal value of k, enhancing the model's predictive accuracy and generalizability. The best performance was achieved with k=15, yielding an accuracy of 78.4%.

Results and Analysis

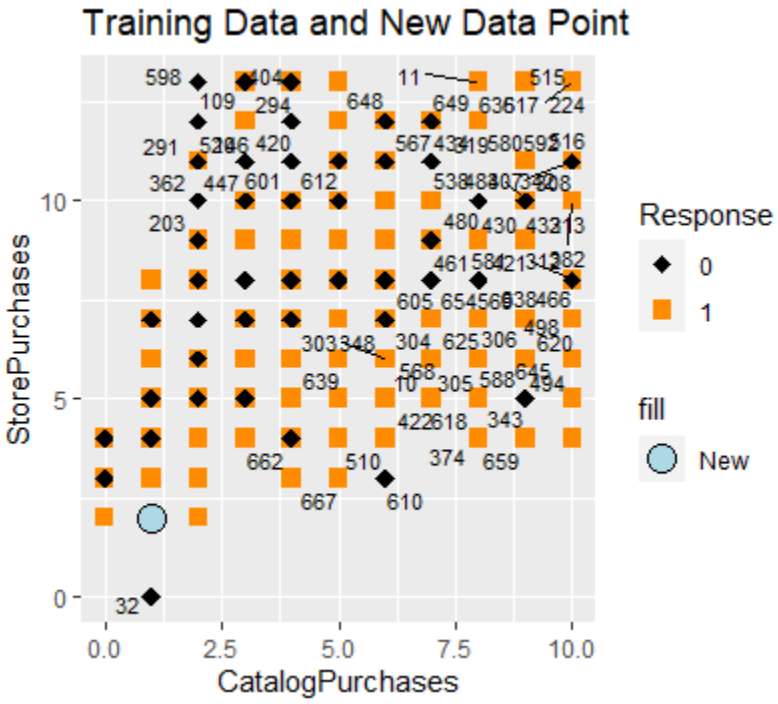
Distribution of Purchases:

The scatter plots display diverse customer engagement levels, as evident from their purchasing behaviors across 'CatalogPurchases' and 'StorePurchases'. This variety indicates no one-size-fits-all approach, emphasizing the need for personalized marketing strategies.

Response Pattern:

The differentiation between responders (1) and non-responders (0) in the campaign reveals potential trends where certain purchasing behaviors might correlate with a higher likelihood of positive responses. Notably, areas densely populated with responders (orange squares) could identify target segments for future campaigns.

New Data Point Representation



In both charts, the 'New' data point's position provides insights into its potential behavior based on the established customer patterns. Its analysis helps gauge the likely response to marketing efforts, based on similarity to existing customer groups.

Comparative Analysis between Initial and Optimized Models:

The initial model (k=3) and the optimized model (k=15) generated two distinct plots. Despite maintaining consistent scales and categories, the transition from k=3 to k=15 offers a nuanced understanding of the model's sensitivity to the choice of k and its impact on the prediction landscape.

Discussion:

The transition from k=3 to k=15, guided by cross-validation, underscores the importance of model tuning in predictive analytics. The increase in accuracy from the initial to the optimized model highlights the nuanced complexities within customer data and the significance of selecting an appropriate k value.

The consistent distribution across both charts suggests that the selected features ('CatalogPurchases' and 'StorePurchases') are significant indicators of customer responses, supporting their choice based on the exhaustive search for optimal predictors.

How Overfitting was avoided throughout the project

To avoid, manage, or minimize overfitting in this project:

Scaling Predictors: Features were scaled to ensure uniformity; this step is crucial as it prevented variables with larger scales from disproportionately influencing the model, a common pitfall that can lead to overfitting.

Categorical Variable Transformation: We transformed categories that skewed distribution into binary variables, such as converting 'Education' into 'graduate' and 'non-graduate'. This approach reduces complexity and helps models focus on significant distinctions.

Exhaustive Search for Predictors: By using an exhaustive search, we systematically evaluated combinations of predictors to identify the most impactful ones, thereby avoiding the inclusion of irrelevant features that could lead to overfitting.

Variance Inflation Factor (VIF) and Pair Plots: We used VIF to identify and eliminate multicollinearity between variables, and pair plots to assess relationships and redundancies among predictors, further refining our feature set.

Handling Outliers: The elimination of outliers from the dataset ensured that our models were not swayed by anomalous data points, which can distort model training and lead to overfitting.

Balanced Dataset: For classification tasks, we used a balanced dataset by undersampling the majority class, which prevents the model from being biased toward the more prevalent class and helps in generalizing better to unseen data.

Additional Measures and Insights:

Cross-Validation: While not explicitly mentioned earlier, cross-validation techniques, particularly Leave-One-Out Cross-Validation (LOOCV) used in k-NN model optimization, are instrumental in evaluating the model's generalizability and thus combating overfitting.

Comparison and Contrast between Logistic Regression and K-NN:

Foundational Differences:

Nature: Logistic Regression is a parametric model, meaning it assumes a specific form for the relationship between predictors and the probability of the target outcome. It is based on the logistic function to model the probability that a given input belongs to a particular category. On the other hand, k-Nearest Neighbors (k-NN) is a non-parametric, instance-based learning method where predictions for new instances are made based on the closest training examples in the feature space.

Model Interpretation: Logistic Regression provides insights into the relationship between the independent variables and the outcome through its coefficients, which can be interpreted in terms of odds ratios. Conversely, k-NN does not provide such interpretable parameters as it makes predictions based on the similarity and majority voting from the nearest neighbors.

Model Complexity and Training:

Training Process: Logistic Regression involves finding the best parameters for the logistic function using methods like Maximum Likelihood Estimation. This process is typically quicker and less computationally intensive than k-NN, which requires no training phase but stores the entire training dataset and uses it during the prediction phase, making it memory-intensive and slower for predictions.

Feature Scaling: Logistic Regression can be relatively robust to small changes in the dataset, but it still requires careful feature selection and can be affected by irrelevant or highly correlated features. In contrast, k-NN is highly sensitive to the scale of the data, as it relies on distance measurements. Thus, feature scaling (normalization or standardization) is crucial for k-NN to perform well.

Performance and Generalization:

Handling Non-linear Data: Logistic Regression is typically used for linearly separable data. It can struggle with complex, non-linear boundaries unless feature engineering is applied. k-NN, however, can adapt to any data shape given a suitable value of k and distance metric, making it more flexible in capturing non-linear relationships.

Overfitting and Underfitting: Logistic Regression can be prone to overfitting especially when the data is high-dimensional; regularization techniques are often applied to avoid this. k-NN can also suffer from overfitting, particularly with a small value of k. However, its susceptibility to overfitting increases with the dimensionality of the data (curse of dimensionality).

Usability and Scenarios of Application:

Reflection

The project was a profound learning experience, emphasizing the importance of thorough data preprocessing, the value of exhaustive feature selection, and the nuances of model tuning. The challenges encountered reinforced the need for a meticulous approach to data science tasks, from understanding the underlying data to interpreting and validating model outputs effectively.

In order to ensure we had a balanced dataset for correlation, we had to undersample the majority class (Response = 0) and this led to us maintaining 2 different datasets for our models. The original dataset (updated_cleaned_dataset.csv) was used for regression, while the balanced one was used for classification (updated_cleaned_dataset_balanced.csv).

On logistic regression, the model's findings, particularly around campaign acceptance and demographic factors such as marital status, offer actionable insights that MarketSphere Inc. can leverage to refine their

marketing initiatives. Certainly, logistic regression has served as an invaluable predictive analytics tool, enabling MarketSphere Inc. to gain a deeper understanding of their customer base and to optimize their marketing strategies accordingly.

Ultimately, ideating our regression and classification models was challenging. As we began the process, we quickly realized that our data was not normally distributed, which is not only a requirement for logistic regression, but it also made it difficult to predict outcomes due to the skewed distribution. The variable distributions had large variations too. Attempts to scale them resulted in the introduction of negatives and infinity values which proved challenging to work with. Also, attempts to add high order terms for most of the variables resulting in high VIFs and poor model performance for regression, surprisingly.

Another discovery we consider surprising was how variables that showed strong correlation with the target variable Wines (such as Income and Recency) were not included by the exhaustive search for subset variables. For curiosity sake, we attempted to introduce these eliminated variables into the model and realized though they were correlated with Wines, they were not significant in predicting Wines bought and eventually had to drop them.

To properly adjust our data, we unfortunately had to drop a fair amount of observational points to ensure our data was normally distributed; while less than ideal, this adjustment allowed us to confidently and accurately move forward, assuring us that our model was not only successful on our training subset, but also on our holdout data.

Another slight challenge we faced was viewing and sharing code with one another, as each group member has different models of PCs and Macs; therefore, our versions of R vary. We were able to effectively work around this, though, to ensure we could collaborate, ideate and ultimately run the same models with ease. This project was certainly a team effort, as improving the accuracy of our regression and classification models required an open mind and fresh eyes to identify areas of improvement and statistically significant predictors, particularly in our holdout data.

Other challenges such as extracting performance metrics and addressing data point overlaps in visualizations were faced. These issues underscored the importance of meticulous data inspection and the need for clarity in interpreting results.

The visualizations, despite their challenges, provided invaluable insights into the data's structure and the model's behavior, underscoring the importance of graphical representations in data analysis. The journey from data preparation through to model evaluation highlighted the critical balance between statistical techniques and practical insights, underscoring the importance of a holistic approach to predictive modeling.

Overall, the project was not just a technical task but a comprehensive analytical process, blending data manipulation, statistical reasoning, and visual storytelling to uncover the patterns within the data and predict customer responses effectively. This experience has underscored the multifaceted nature of data science and the continuous interplay between theory and practice in deriving meaningful insights from data.

Citations

Boudet, J., Brodherson, M., Robinson, K., & Stein, E. (2023, June 26). *Beyond belt-tightening: How marketing can drive resiliency during uncertain times.* McKinsey & Company. https://www.mckinsey.com/capabilities/growth-marketing-and-sales/our-insights/beyond-belt-tightening-how-marketing-can-drive-resiliency-during-uncertain-times#/

Checa, A., Heller, C., Stein, E., & Wilkie, J. (2023, April 5). *Modern marketing: Six capabilities for multidisciplinary teams.* McKinsey & Company. https://www.mckinsey.com/capabilities/growth-marketing-and-sales/our-insights/modern-marketing-six-capabilities-for-multidisciplinary-teams

Customer Personality Analysis. (n.d.). Www.kaggle.com. https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis/data

Appendix Back to Top