*Francesca Marchese, Ian Smith, Emmanuella Acheampong, Andrew Mukurazita*

*BUS-212A- Report 3*

BUS 212A – Advanced Data Analytics

Report 3 – Tree-Based Models

April 05, 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 Report 1, we improved data quality significantly, setting a solid foundation for further analysis, through meticulous preprocessing . In Report 2a, we advanced our investigation with the development of three supervised models: a regression model demonstrating substantial predictive power with an 84.28% variance explanation in wine purchases and a RMSE of 7.43; a logistic regression model showcasing a robust accuracy of 81.9% alongside balanced metrics in classifying customer responses to marketing campaigns, as reflected by an F1 score of 81.9%; and a kNN model with an accuracy of 81.06%, identifying K = 19 as the most effective parameter, evidenced by an F1 score of 80.62%. These advancements provide MarketSphere with actionable insights and underscore the logistic regression model's superior balance in precision and recall, optimizing customer response identification.

In Report 2b, our journey through the data landscape of MarketSphere Inc. took an exploratory turn into unsupervised learning, a technique that enriched our understanding of the customer base. Through the application of Hierarchical and k-Means clustering, we identified hidden patterns within the data, providing a fresh perspective on customer segmentation. This strategic inclusion of clustering variables into our models yielded significant enhancements across various performance metrics:

***The Linear Regression Model*** demonstrated a refined predictive capability for wine purchases, as evidenced by an uptick in the adjusted R-squared value and a reduction in RMSE.

***The Logistic Regression Mode*l** continued to exhibit strong accuracy, fine-tuning its balance of sensitivity and specificity slightly with the integration of clustering insights.

***Our kNN Model*** uncovered that specific combinations of clustering variables significantly improved model precision, accuracy, and recall, shedding light on the predictive nuances of customer responses.

# Introduction to Report 3

In Report 3, we progress to harnessing the power of Tree-Based Models and Ensemble Learning, with the objective of unlocking deeper insights into the intricate patterns of customer behavior at MarketSphere. This report is structured around three core areas:

**Diving into Tree-Based Models:** We will engage in the meticulous development and optimization of both Classification and Regression Trees. Our goal is to deepen our analytical capabilities by employing these models on a variety of target variables and fine-tuning them through hyperparameter adjustments to achieve peak accuracy and performance.

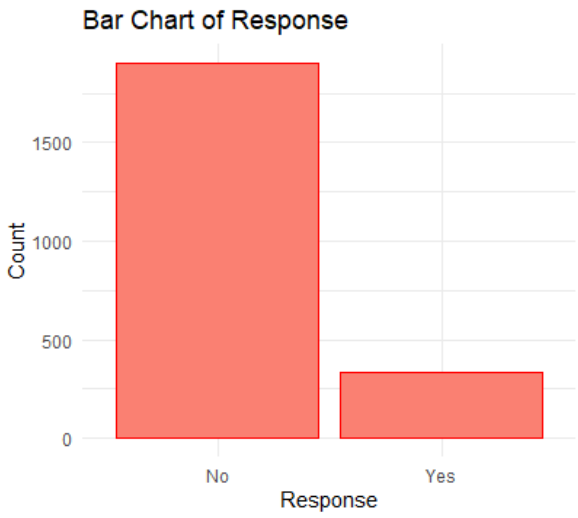
**Embracing Ensemble Learning:** The exploration extends to three advanced ensemble techniques: Bagged Trees, Random Forest, and Boosted Trees. By applying these methods to our dataset, we aim to examine and document the nuances in model accuracy, the influence of hyperparameter tuning, and the significance of various predictors in our models. This segment promises a rich comparative analysis that underscores the strengths and limitations of each ensemble approach.

**Insightful Conclusions and Reflections:** The final section of Report 3 will weave together the insights garnered from both the tree-based and ensemble models. This comparative analysis will shed light on how well we managed to steer clear of underfitting and overfitting, the implications of choosing interpretability over accuracy, and vice versa. Drawing from our findings, we'll articulate strategic recommendations for MarketSphere, emphasizing actionable insights for enhancing marketing strategies and business operations

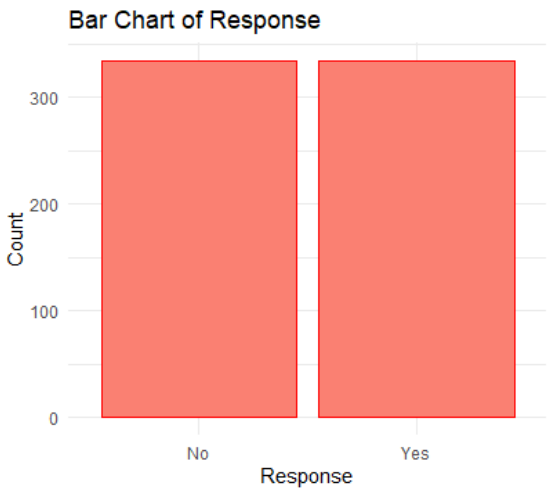
# Part 1: Trees

## Classification Tree

*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.

In the latest phase of our collaboration with MarketSphere Inc., we at DataWise Consultants focused on refining our predictive model using a carefully selected set of features determined to be the most influential in predicting customer response to marketing initiatives. In constructing our classification tree for MarketSphere Inc, the features were meticulously chosen based on an exhaustive search method implemented during our early data preparation stage, ensuring that each variable contributes significantly to the model's predictive capability This dataset comprised 668 observations across the 19 variables selected from the total 34 variables .

Income and expenditures on items like gold and meats offer a lens into purchasing power and product predilections. Household dynamics, indicated by factors such as the presence of teenagers and marital status, inform us about potential responsiveness to campaigns tailored to family-oriented or educational themes. Engagement metrics like last purchase, web visits, and various purchase types (discount, catalog, web, store) reveal how customers interact with the brand across different channels, and prior campaign acceptance rates directly measure marketing receptiveness. Additionally, non-graduate education levels and identified customer clusters through hierarchical and k-Means clustering techniques contribute nuanced understanding of the customer base, enabling precise targeting. These factors collectively serve as robust predictors of the target variable, Response, providing MarketSphere with valuable insights for optimizing future marketing strategies.

## Descriptive Statistics of Selected Variables

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

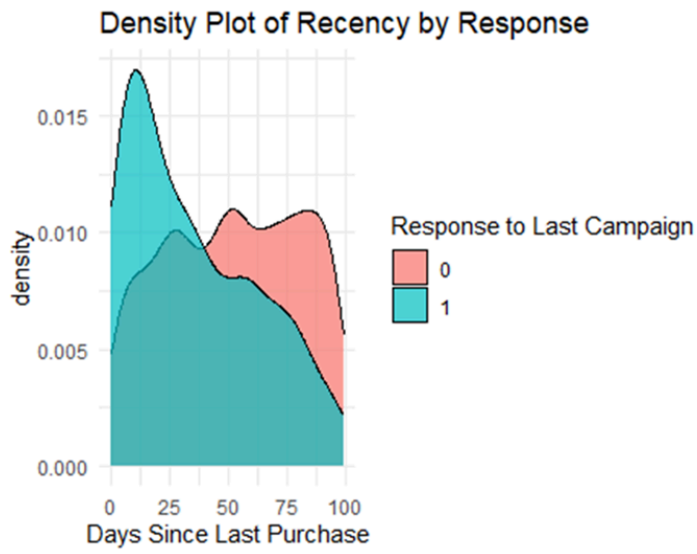



*Figure 33 - Days of last purchase vs. Response*

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*

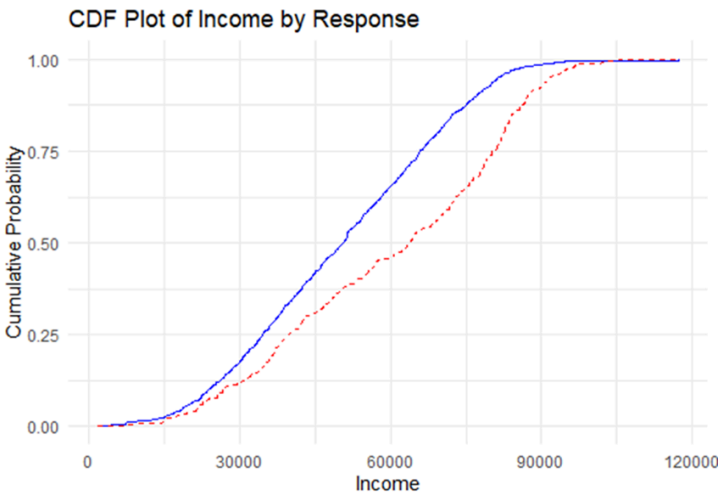



*Figure 34 - Income vs. Response*

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.

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.

Approach

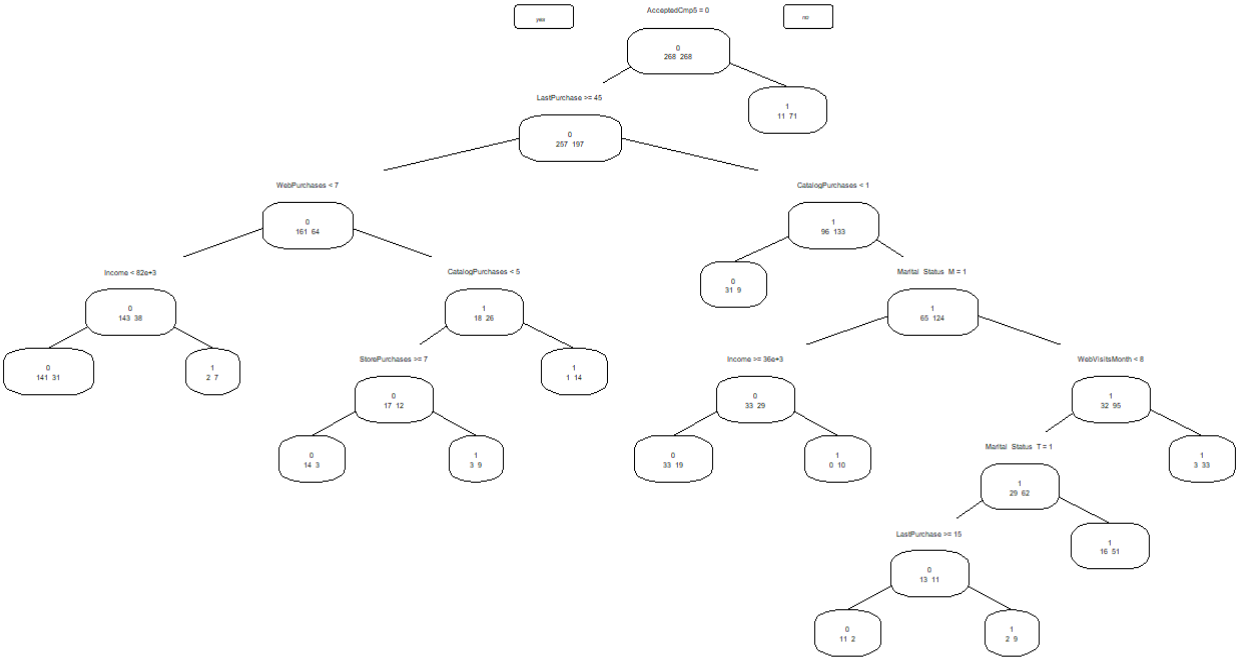
We allocated 80% of our data to training (536 observations) and kept the remaining 20% for validation (132 observations), ensuring a robust data structure for both model training and subsequent evaluation. This division was executed using the createDataPartition function, ensuring a random but representative split.

*Initial Classification Tree*

Our initial classification tree model revealed several key decision points in predicting customer responses for MarketSphere Inc. At the root, the tree first checked whether a customer had accepted the fifth campaign (AcceptedCmp5). For customers not participating in this campaign, the model next evaluates the recency of their last purchase, using a cutoff of 44.5 days to differentiate between recent and less recent interactions. For recent purchasers, online shopping behavior (WebPurchases) is assessed, highlighting the model's focus on digital engagement as a key predictor of response.

Income levels further refine predictions among frequent online shoppers (married customers with an income below $36,352.5 uniformly showed a positive response), suggesting an intersection between financial capacity and likelihood to respond. Conversely, customers with purchases over 44.5 days ago are segmented by their catalog purchasing habits (CatalogPurchases), followed by marital status distinctions, which influence response patterns differently based on familial contexts.

This initial exploration effectively segments customers based on campaign interaction, purchasing recency, online engagement, and demographic factors, setting a foundation for more nuanced analysis in the pursuit of the most effective predictive model. Kindly refer to the R script for a clearer view.

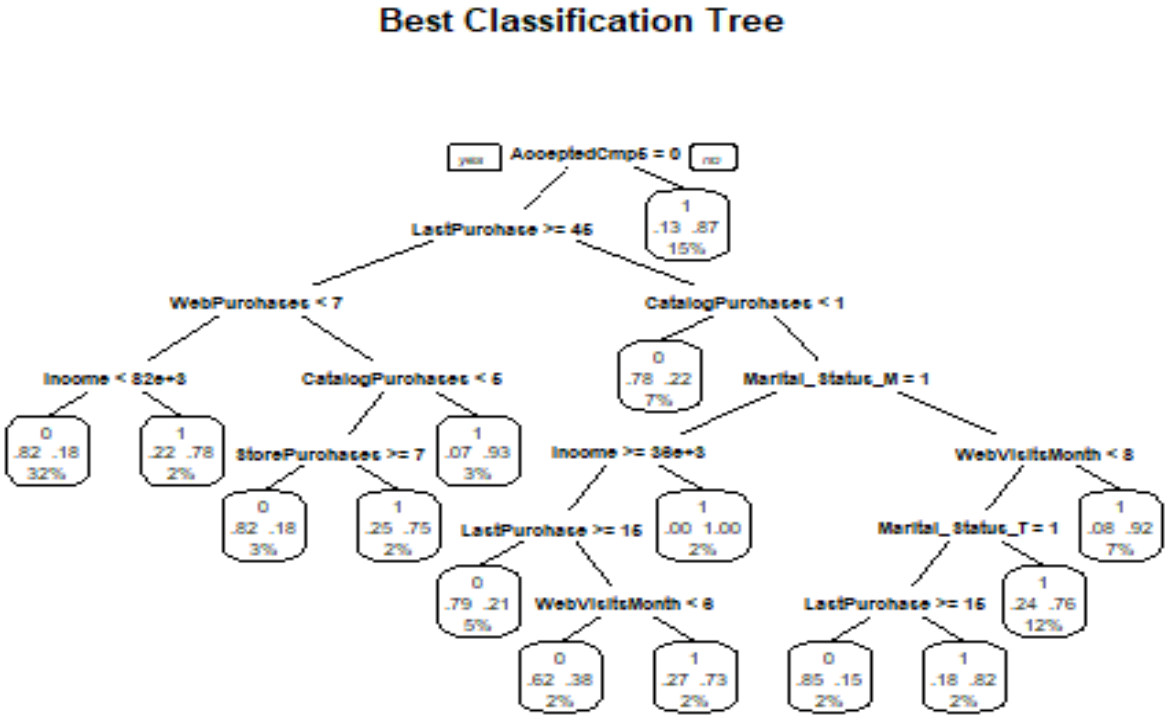


```
> print(initClassTree)
n= 536

node), split, n, loss, yval, (yprob)
      * denotes terminal node

  1) root 536 268 0 (0.50000000 0.50000000)
    2) AcceptedCmp5< 0.5 454 197 0 (0.56607930 0.43392070)
      4) LastPurchase>=44.5 225  64 0 (0.71555556 0.28444444)
        8) WebPurchases< 6.5 181  38 0 (0.79005525 0.20994475)
         16) Income< 81630 172  31 0 (0.81976744 0.18023256) *
         17) Income>=81630 9   2 1 (0.22222222 0.77777778) *
        9) WebPurchases>=6.5 44  18 1 (0.40909091 0.59090909)
         18) CatalogPurchases< 4.5 29  12 0 (0.58620690 0.41379310)
           36) StorePurchases>=6.5 17   3 0 (0.82352941 0.17647059) *
           37) StorePurchases< 6.5 12   3 1 (0.25000000 0.75000000) *
         19) CatalogPurchases>=4.5 15   1 1 (0.06666667 0.93333333) *
      5) LastPurchase< 44.5 229  96 1 (0.41921397 0.58078603)
       10) CatalogPurchases< 0.5 40   9 0 (0.77500000 0.22500000) *
       11) CatalogPurchases>=0.5 189  65 1 (0.34391534 0.65608466)
         22) Marital_Status_Married>=0.5 62  29 0 (0.53225806 0.46774194)
           44) Income>=36352.5 52  19 0 (0.63461538 0.36538462) *
           45) Income< 36352.5 10   0 1 (0.00000000 1.00000000) *
         23) Marital_Status_Married< 0.5 127  32 1 (0.25196850 0.74803150)
           46) WebVisitsMonth< 7.5 91  29 1 (0.31868132 0.68131868)
             92) Marital_Status_Together>=0.5 24  11 0 (0.54166667 0.45833333)
              184) LastPurchase>=14.5 13   2 0 (0.84615385 0.15384615) *
              185) LastPurchase< 14.5 11   2 1 (0.18181818 0.81818182) *
             93) Marital_Status_Together< 0.5 67  16 1 (0.23880597 0.76119403) *
           47) WebVisitsMonth>=7.5 36   3 1 (0.08333333 0.91666667) *
    3) AcceptedCmp5>=0.5 82  11 1 (0.13414634 0.86585366) *
```

*Initial tree's performance*

The initial classification tree model displayed a notable capacity to forecast customer responses for MarketSphere Inc. It secured an accuracy of 74.24%, indicating a strong ability to predict responses correctly in most cases. The model excelled in sensitivity, achieving an 81.82% rate, showing its effectiveness in identifying true positive responses and minimizing false negatives. However, its specificity stood at 66.67%, showing a relatively lower ability to correctly identify true negatives. This performance suggests the model's proficiency in balancing the identification of both responders and non-responders to campaigns, with particular strength in detecting positive responses.

To enhance the predictive accuracy of our classification tree model for MarketSphere Inc., we engaged in hyperparameter tuning using grid search, specifically focusing on the complexity parameter (cp). This process involved exploring a range of cp values from 0.001 to 0.05, incrementing by 0.001, to find the optimal balance between model complexity and generalization. We applied cross-validation with ten folds to ensure a robust evaluation of each model variant, aiming to identify the cp value that yields the highest performance. This systematic approach allowed us to refine our model by preventing overfitting, thus improving its ability to predict customer responses accurately, thus coming up with our best classification tree.



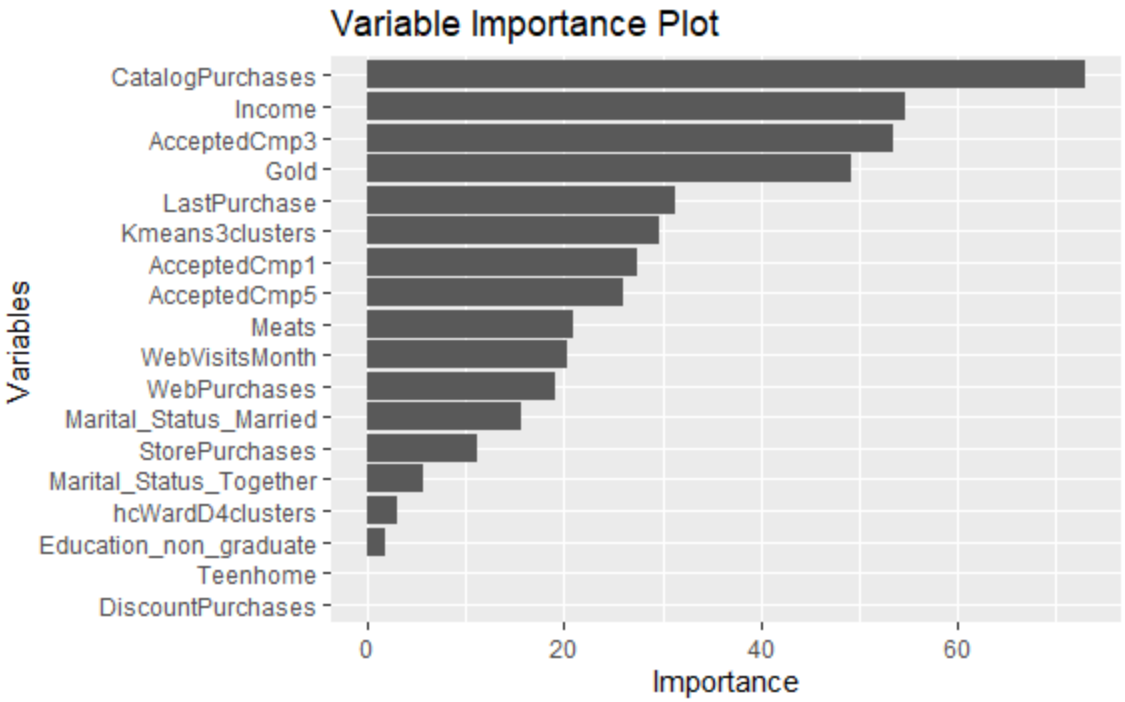

The optimal complexity parameter (cp) was determined to be 0.008, as it achieved the highest accuracy. Please refer to the R script for a clearer view.

```
> print(bestClassTree)
CART

536 samples
 18 predictor
  2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 482, 482, 482, 483, 483, 483, ...
Resampling results across tuning parameters:

  cp     Accuracy   Kappa
  0.001  0.7428721  0.4857127
  0.002  0.7428721  0.4857127
  0.003  0.7372117  0.4743800
  0.004  0.7353599  0.4706763
  0.005  0.7335080  0.4669513
  0.006  0.7335080  0.4669513
  0.007  0.7446890  0.4892407
  0.008  0.7446890  0.4892407
```

The structure of the best classification tree model, obtained after hyperparameter tuning, appears identical to the initial tree model in terms of the splits and variables used. Both trees start with the same root decision based on the acceptance of the fifth campaign (AcceptedCmp5), and follow similar patterns of decision-making based on LastPurchase, WebPurchases, Income, CatalogPurchases, and Marital Status, among others.

However, the selection of cp = 0.008 for the best model suggests that during the tuning process, this particular complexity parameter provided the optimal balance between the tree's depth and its ability to generalize to new data, without overfitting. This indicates that while the structure of the decision tree did not change, the process of tuning and validation likely ensured that the final model was the most robust and predictive version of the tree, given the range of cp values explored.

The fact that the tree's structure remained consistent before and after tuning implies that the variables identified as important in the initial model were indeed the most significant predictors of customer response, according to the data. The hyperparameter tuning process, therefore, refined the model's performance metrics such as accuracy, sensitivity, and specificity, rather than altering the model's fundamental decision-making logic.




Though AcceptedCmp5' was the primary criterion for the initial data split due to its immediate impact in differentiating the responses, we observed that 'CatalogPurchases' emerged as the most significant variable overall. This was deduced from the cumulative reduction in impurity it provided across all the splits it influenced within the tree. This highlights a critical aspect of tree-based models where the importance of a feature is not solely judged by its position as a root node but by its aggregate contribution to enhancing the model's predictive accuracy across various decision nodes.

*Key Takeaways from the Best Classification Tree*

In our refined classification tree model for MarketSphere Inc., key variables and cutoff values provide diagnostic insights into customer behaviors. For instance, the variable 'AcceptedCmp5' is the primary decision node, with a split indicating that those who did not engage with the fifth campaign (< 0.5) exhibited varied responses based on subsequent attributes.

The 'LastPurchase' variable, with a cutoff of approximately 45 days, suggests that customers with more recent interactions are more likely to respond positively. This reflects the importance of timely follow-up actions in marketing strategies, where customers who have recently engaged are likely to still be in a receptive mindset.

Further splits in the tree involve 'WebPurchases', 'Income', and 'CatalogPurchases', emphasizing the role of purchasing channels and economic status in predicting campaign responses. For example, customers with web purchases fewer than 6 and an income below $81,630 were more responsive, pointing towards a segment that is both digitally engaged and possibly more price-sensitive.

The tree also highlighted that for customers with fewer recent purchases, engagement with catalog shopping (with a significant cutoff at 0.5) is a key indicator of response likelihood. This could inform MarketSphere to invest in catalog marketing efforts targeted at less frequent purchasers to re-engage them.

Marital status emerged as a differentiator, with married customers responding differently based on their income levels. Interestingly, income provided a distinct cutoff at $36,352.5, below which married customers uniformly showed positive responses, suggesting that for this income bracket, factors other than earnings might drive responsiveness.

Lastly, the frequency of web visits further segmented customers, particularly unmarried individuals, demonstrating that online engagement is a significant predictor of marketing success in specific demographics.

These diagnostic insights reveal the complex interplay of socioeconomic, behavioral, and engagement factors that can be leveraged to enhance targeted marketing campaigns. For MarketSphere, these findings suggest tailoring strategies that consider the nuanced patterns of customer engagement, channel preference, and responsiveness to past campaigns, aiming to optimize the effectiveness of future marketing endeavors.

*Comparison between the initial classifier and the best classifier*

Upon tuning, the best tree's performance edged out the initial model, with a slight increase in accuracy from 74.24% to 75%. This suggests a more reliable model in distinguishing between customers who would respond to the campaign and those who wouldn't. The trade-offs are notable: while sensitivity decreased from 81.82% to 80.30%, indicating a small drop in the model's ability to correctly identify true positives, specificity increased from 66.67% to 69.70%, meaning the model improved in correctly identifying true negatives.

The F1 score, which balances precision and recall, saw a minor improvement, nudging up to 74.63% from 73.47%. In practical terms, MarketSphere might appreciate this model's improved balance in correctly identifying non-responders, which can be crucial for preventing wasted marketing efforts on uninterested customers.

For a retail business like MarketSphere, accuracy is often considered important because it reflects the overall correctness of the model in identifying responders and non-responders to marketing campaigns. However, in a nuanced business context, sensitivity (recall) can be equally or even more critical. This is because it measures the model's ability to identify all actual responders, which ensures that the marketing efforts reach as many interested customers as possible, maximizing the potential for sales and revenue.

Specificity is also valuable as it informs the business of its capacity to correctly identify those who are unlikely to respond, thus avoiding unnecessary spending on ineffective marketing efforts. Ultimately, the importance of these scores may vary based on the business's objectives and the cost implications of false positives versus false negatives. For example, if missing out on potential customers (false negatives) is costlier than marketing to uninterested ones (false positives), then sensitivity would take precedence. Conversely, if a more conservative approach is preferred to avoid waste, specificity might be emphasized.

*Best Classification Tree vs. Logistic Regression and k-NN Models*

The logistic regression model, with and without clustering, previously achieved accuracies of 81.9% and 82.09%, respectively, both slightly higher than the best classification tree's accuracy of 75%. This suggests that the regression approach was slightly more adept at predicting customer responses for MarketSphere.

The k-NN models also showed a higher accuracy, with the first model at 81.06% and the second at 82.58% when clustering variables were included. The recall rates for these models were particularly high, indicating a strong ability to identify true positives. This was an area where the classification tree, with a sensitivity of 80.30%, did not perform as well.

*Business Implications*

For MarketSphere, these comparisons indicate that logistic regression and k-NN models, particularly with clustering variables, may provide more accurate targeting for marketing initiatives. The improved balance between sensitivity and specificity in the classification tree, however, may offer a better strategy to minimize marketing expenditure on unlikely respondents. The key for MarketSphere lies in choosing the model that aligns best with their campaign goals—whether that's maximizing the capture of potential responders (higher sensitivity) or ensuring marketing efforts are not wasted on non-responders (higher specificity).

In conclusion, while the tuned classification tree presents a reliable model with a balanced approach to predicting customer responses, the logistic regression and k-NN models, particularly with clustering, may still offer more precision for MarketSphere's marketing strategies.

## Part 1: Regression Tree

*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 MarketSphere as it aids in effective inventory management, targeted marketing strategies, financial planning, and enhances overall customer satisfaction and profitability.

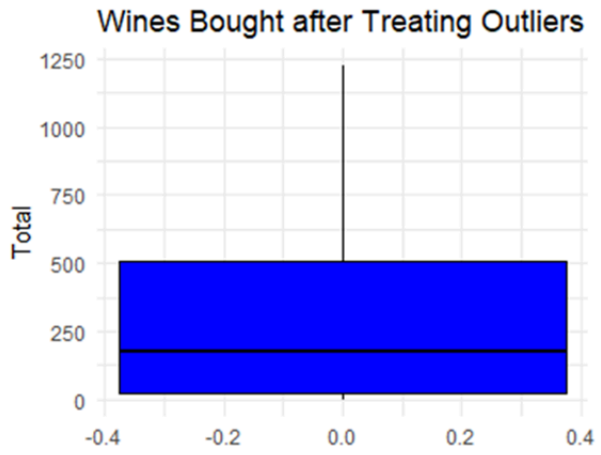
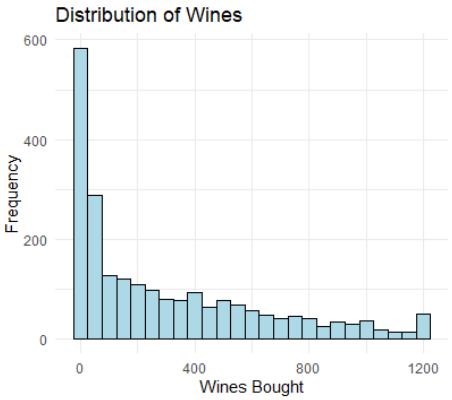


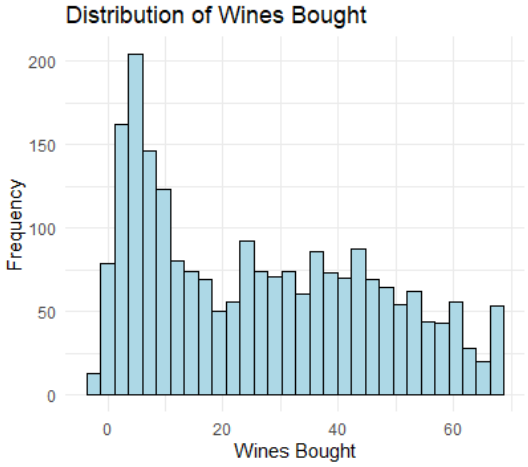

*Figure 7 - Total Wines bought without outliers.*

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 MarketSphere. 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.

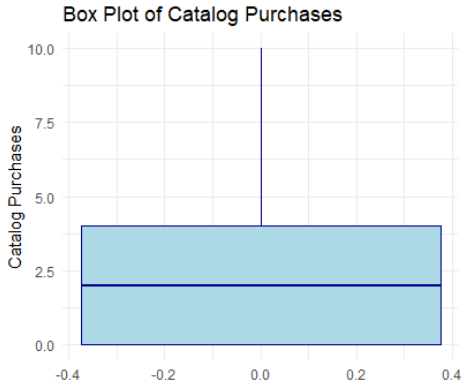
## Feature Selection

*Model Building and Selection:*

For our initial model, 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:

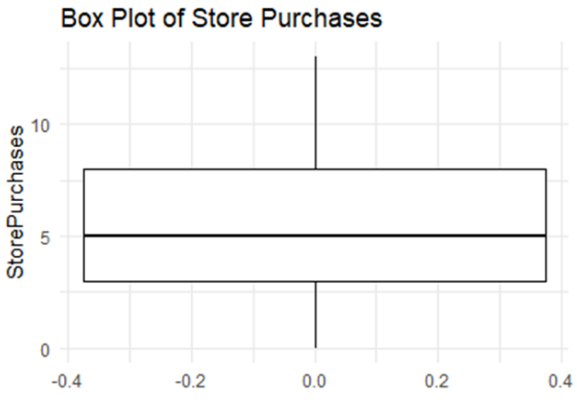



*Figure 17 - Store purchases*

*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:

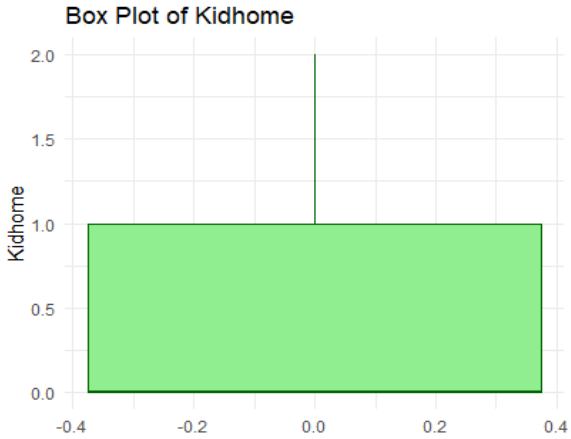
*Minimum*: 0 (some customers have no children at home)

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

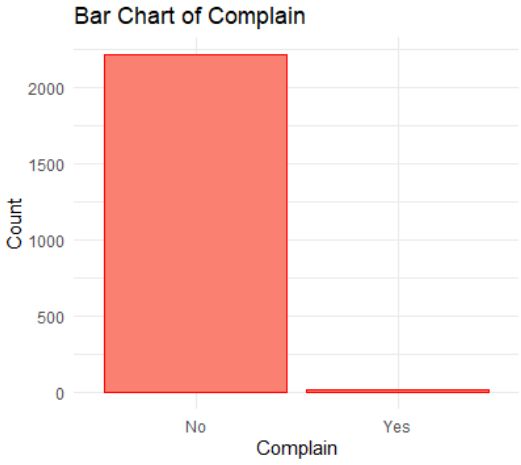
*Mean*: 0.4441 (indicating that, on average, less than one child is present in customers' homes)

*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:

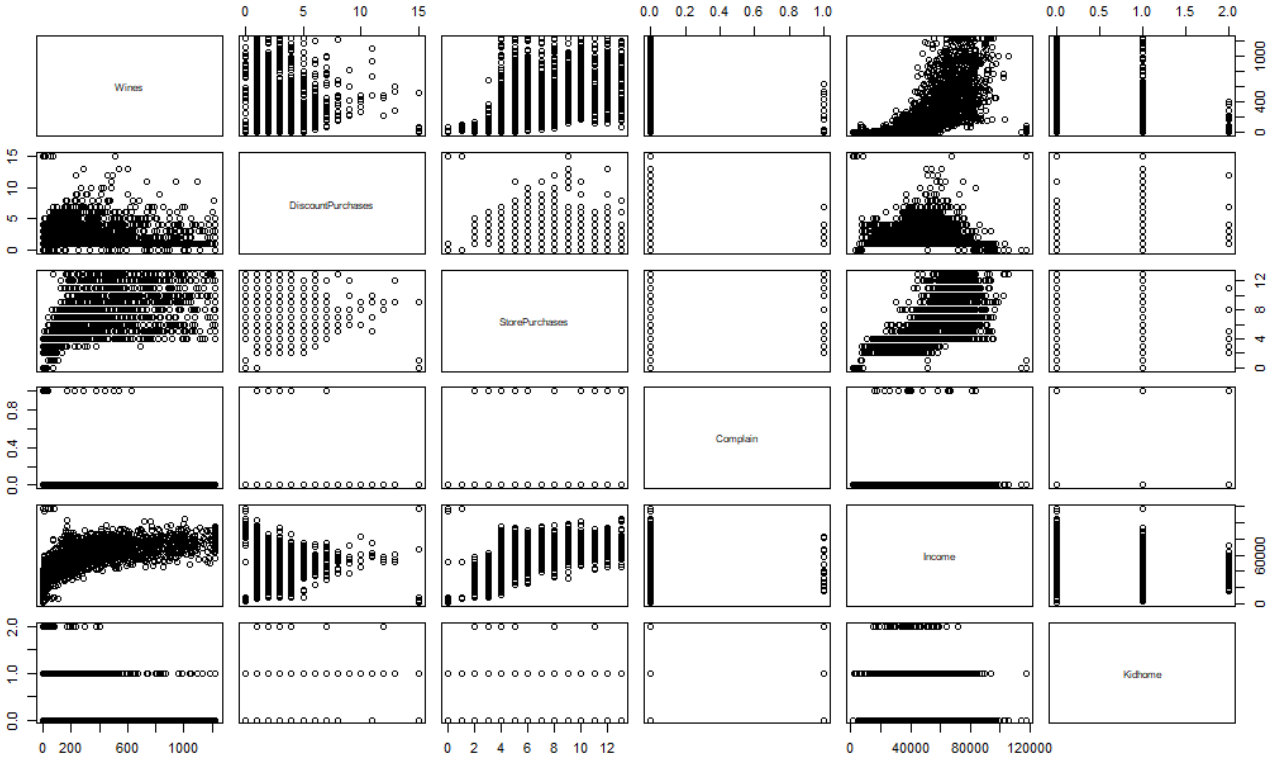



*Minimum*: 0 (indicating that most customers did not register complaints)

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

*Mean*: 0.008944 (very few customers have lodged complaints)

*Scatter Plot of Initially Considered Predictor Variables:*



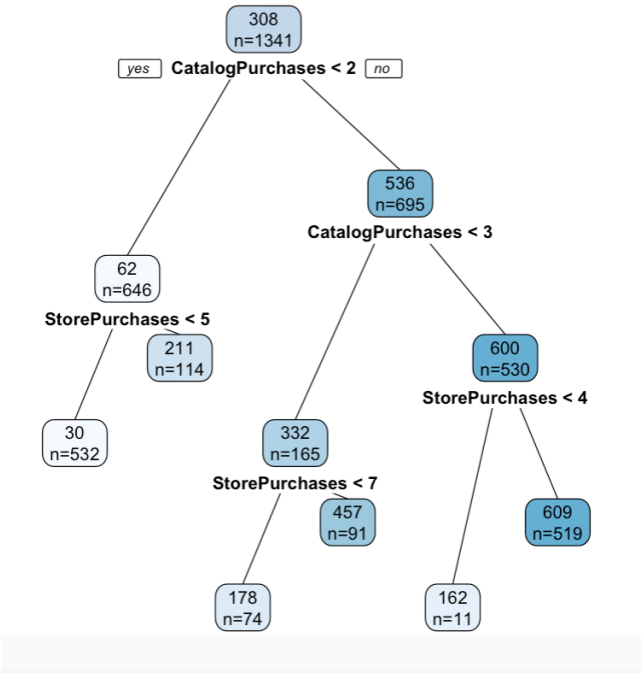

In examining the relationships between 'Wines' and the predictors, we observed that 'Wines' spending positively correlates with 'CatalogPurchases' and 'StorePurchases', indicating that customers who spend more through catalogs or in stores tend to also spend more on wines. Conversely, there is a negative 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. 'Complain' was however not significant to the prediction of Wines bought, hence it was excluded from the model.

*Initial Regression Tree*

Our initial regression tree model remained consistent with our previous regression analysis, using 'Wines' as the target variable and 'CatalogPurchases,' 'StorePurchases,' and 'KidIncome' as the predictor variables; 'Compalin' was excluded, as it proved to be an insignificant predictor of Wine purchases. The tree predicts that for customers with less than 3 catalog purchases and less than 5 store purchases, approximately 34 units of wine will be purchased. For customers with 3 or more catalog purchases and less than 4 store purchases, the regression tree

model predicts around 196 bottles of wine purchased. For customers with less than 3 catalog purchases and 5 to 8 store purchases, the model predicts about 226 bottles of wine purchased. For customers with 8 or more store purchases made and less than 3 catalog purchases predicts that roughly 413 bottles of wine will be purchased; and, lastly, for customers with 3 or more catalog purchases and 4 or more store purchases, the model predicts about 604 bottles of wine purchased. Ultimately, the different combinations of catalog and store purchases allow the model to make predictions on the quantity of wine purchased by customers, allowing us to better serve clients with actionable insights.

We allocated 60% of our data to training and reserved the remaining 40% for validation, ensuring a robust data structure for both model training and subsequent evaluation. This division was executed using the createDataPartition function, ensuring a random but representative split.




The Mean Squared Error (MSE) is a measure used to assess the predictive quality of regression models. Analyzing the MSE of the initial regression tree aids in assessing how well it predicts the wine purchases based on specifications and predictors. The three predictors that we identified as most significant in our previous regression analysis - Catalog Purchases, Store Purchases and Kidhome - will remain consistent in all of our regression tree iterations. The model complexity, tree depth and minimum split sizes, though, will be evaluated and adjusted as necessary in each iteration to improve our tree's predictive capabilities.

With that being said, the training MSE for the initial regression model was 41333.46 and the validation MSE, which had the same parameters, was 44427.80. As a starting point, we chose a maxdepth of 3, which determines the depth of the tree/the longest path from the root node to a leaf node, and we chose a minimum split of 10, which indicates the minimum number of observations necessary to split a node further. Considering both the training and validation MSEs were high, we concluded that our initial regression tree model may be overfitting the data.

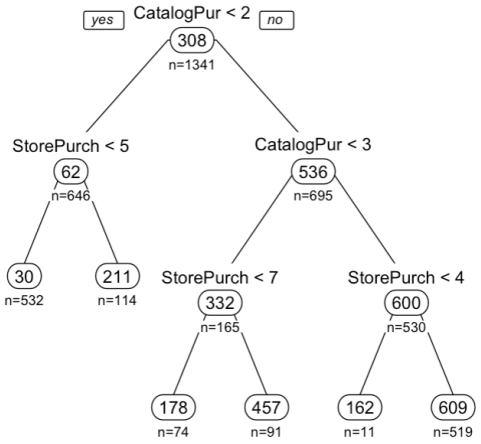
In an attempt to improve the predictive power of our regression tree and avoid overfitting, we implored hyperparameters that would result in the lowest MSE possible. We used the best maxdepth of 3 and the best

minimum split of 5 to refit the model; the regression tree resulted in the same training and validation MSEs as the initial model that had a maxdepth of 3 and a minimum split of 10: those values were 41333.46 and 44427.80, respectively.

*Best Regression Tree Model*

Adjusting the cost of complexity parameter (cp) is another way to manage, and ideally improve, the predictive power of decision trees by adjusting the complexity. We began by performing a grid search to obtain the most useful cp value, which was determined to be 0.005. After applying the best cp of 0.005 to the best maxdepth of 3 and the best minimum split of 5, the final training MSE reduced to 40652.29 and the final validation MSE reduced to 43852.944. While our model does have some degree of overfitting, every adjustment made to manage model complexity and predictive capabilities has tradeoffs. Therefore, by adjusting the cp, the MSE values decreased, indicating better performance and increased predictive performance.

Pruning the tree is another great technique to manage a model's complexity and hopefully improve its MSE values, which would, in turn, reduce overfitting. The function 'printcp(cv_ct)' printed the complexity parameter table of the cross-validated regression tree model, highlighting the various metrics associated with performance at each level of pruning. In summary, the table below highlights that the best complexity parameter (cp) is the one that minimizes the cross validation error ('xerror'); it is 0.005. The visualization of the pruned tree is also below.




n= 1341

| | CP | nsplit | rel error | xerror | xstd |
|---|---|---|---|---|---|
| 1 | 0.5088512 | 0 | 1.00000 | 1.00169 | 0.039651 |
| 2 | 0.0607927 | 1 | 0.49115 | 0.50476 | 0.024722 |
| 3 | 0.0214597 | 2 | 0.43036 | 0.42447 | 0.021139 |
| 4 | 0.0207955 | 3 | 0.40890 | 0.42109 | 0.020988 |
| 5 | 0.0145029 | 4 | 0.38810 | 0.40053 | 0.020848 |
| 6 | 0.0061569 | 5 | 0.37360 | 0.38471 | 0.020101 |
| 7 | 0.0050000 | 6 | 0.36744 | 0.37848 | 0.020002 |

Additional grid searches and hyperparameters were applied to the model to analyze the effects and differences on the decision tree's predictive performance, as well as to ensure that the complexity parameter of 0.005 truly did positively impact the model's performance. The output of hyperparameter tuning for a decision tree model using 'tuneGrid' to analyze 5 different values for the cp of the decision tree - 0.005, 0.002, 0.001, 0.0005, and 0.0002 - is displayed below. The output contains the cp value, as well as the Root Mean Squared Error, Rsquared, Mean Absolute Error and a few standard deviation metrics to indicate variability. In summary, the results highlight that a complexity parameter of 0.002 appears to offer the best balance between predictive accuracy (lower RMSE and MAE) and model fit (higher Rsquared). This clearly differs from the previous grid search that identified 0.005 as the best cp value, so we continued to iterate our model to see which holistic approach of features and parameters improved the MSEs of both the training and validation sets the most.

| | cp | RMSE | Rsquared | MAE | RMSESD | RsquaredSD | MAESD |
|---|---|---|---|---|---|---|---|
| 1 | 2e-04 | 210.1559 | 0.6053485 | 133.4949 | 15.898024 | 0.03244322 | 9.640348 |
| 2 | 5e-04 | 210.1975 | 0.6054114 | 133.1888 | 16.561452 | 0.03457960 | 10.311683 |
| 3 | 1e-03 | 210.2415 | 0.6046127 | 134.1944 | 16.184827 | 0.03417574 | 10.079330 |
| 4 | 2e-03 | 205.7317 | 0.6191188 | 134.8349 | 11.634006 | 0.02198627 | 8.939000 |
| 5 | 5e-03 | 207.1426 | 0.6139336 | 135.6249 | 9.373513 | 0.01513897 | 6.063257 |

*Best Regression Tree vs. Linear Regression Models*

The regression tree model, fine-tuned with optimal hyperparameters including a maximum depth of 3 and a minimum split of 5, demonstrated a substantial MSE. When adjusted with a complexity parameter (cp) of 0.002, it showed a training MSE of 41053.399 and a higher validation MSE. This indicates that while the model could capture a significant portion of the variability within the training data, it struggled to generalize effectively to unseen data, reflected in the higher validation MSE.

In contrast, the linear regression model from the previous assignment, with an adjusted R-squared value of 0.8428, suggested a strong ability to explain the variance in wine purchases through the predictors in the model. The RMSE for the training data was around 7.05, showcasing a relatively lower deviation from observed values. The introduction of clustering variables improved the model's performance slightly, with a slight increase in adjusted R-squared and a decrease in RMSE.

The higher MSE and RMSE seen with regression trees compared to linear regression models can be attributed to the trees' complexity and susceptibility to overfitting, which might not generalize well to unseen data. Even with optimal hyperparameters such as maximum depth and minimum split size, they may capture noise in the training data as patterns, leading to higher errors on validation data. Linear regression, on the other hand, excels at capturing

linear relationships and tends to overfit less, resulting in better performance on validation sets. Essentially, if the underlying relationship between variables and target is linear, linear regression models are likely to outperform regression trees due to their ability to more accurately and generally capture these relationships.

# Part 2: Ensemble (Blender) Models

## Classification

For MarketSphere Inc., employing ensemble methods has marked a significant step forward in refining their ability to predict customer responses to marketing campaigns. Through the application of Bagged Trees, Random Forest, and Boosted Trees models, we've embraced a nuanced understanding of customer behavior, each model bringing its unique strengths to the forefront. Below is a detailed report on each of these ensemble methods, including insights from hyperparameter tuning and implications for MarketSphere's marketing strategies.

**Bagged Trees Model**

The Bagged Trees Model, utilizing the 'treebag' method, has demonstrated substantial predictive power with an accuracy of 80.3%. Its ability to achieve a Kappa statistic of 0.6061 signals a strong agreement beyond chance in its predictions. The model's sensitivity and specificity stand at 77.27% and 83.33%, respectively, indicating a balanced capacity to identify both responders and non-responders accurately. This balance is crucial for MarketSphere, ensuring that marketing efforts are not wasted on unlikely prospects while capturing a broad audience of potential customers.

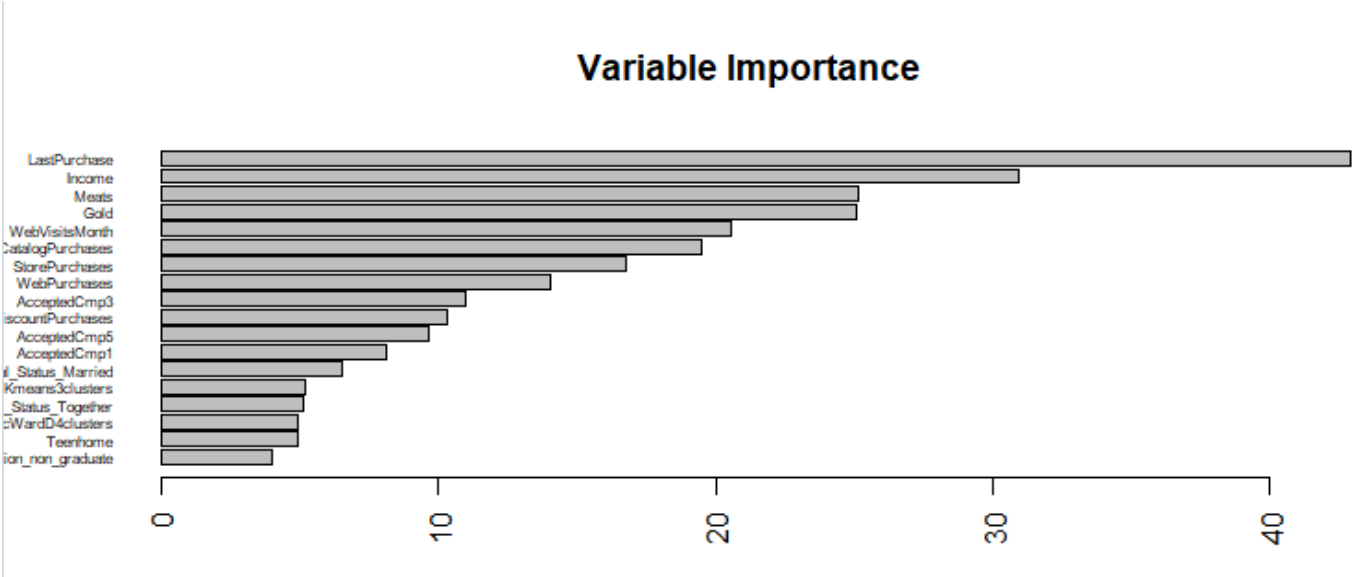
**Hyperparameter Tuning:**

For MarketSphere's analysis, the Bagged Trees Model was employed, a technique that constructs multiple decision trees through bootstrapping, thereby combining various samples of the dataset. This method enhances the robustness of the model and naturally mitigates the risk of overfitting, as it relies on the collective wisdom of numerous trees rather than hyperparameter tuning. To further validate its predictive power, we incorporated a five-fold cross-validation in the training regimen. Here, the data is segmented into five equal parts; in each validation cycle, four segments are used for training and one for validation. Such a rigorous approach not only tests the model's performance across diverse data slices but also safeguards against anomalies, ensuring a more reliable prediction of customer response patterns for MarketSphere.

**Business Implications:** The Bagged Trees Model's balanced performance makes it an invaluable tool for optimizing marketing resources. By accurately segmenting the customer base, MarketSphere can tailor its campaigns more precisely, enhancing customer engagement and conversion rates.

**Random Forest Model**

Enhancing the predictive accuracy to 82.58%, the Random Forest Model introduces an improved framework for customer response prediction. Its specificity rate of 87.88% is particularly noteworthy, suggesting an exceptional ability to correctly identify individuals unlikely to respond to campaigns.




The model's variable importance analysis highlights LastPurchase, Income, and Meats as key predictors, offering strategic insights into customer preferences and behaviors.

**Hyperparameter Tuning:**

```
Random Forest Model
> cat("Accuracy: ", rfCm$overall['Accuracy'], "\n")
Accuracy:  0.8257576
> print(rfGridSearch$bestTune)
  mtry
3    6
```

In the development of our Random Forest Model for MarketSphere, we embarked on a

hyperparameter tuning exercise to identify the optimal configuration for predicting customer responses. The 'mtry' parameter, which determines the number of variables randomly sampled as candidates at each split, was central to our tuning process. Using a grid search approach with cross-validation, we explored a range of 'mtry' values: 2, 4, 6, and 8. This methodical search aimed to find the 'mtry' value that maximizes the model's accuracy while maintaining a balance between complexity and generalizability. The process revealed that an 'mtry' value of 6 resulted in the highest accuracy, indicating that selecting six variables at each split provides the best trade-off between model performance and overfitting risk.

**Business Implications**: With its enhanced specificity, the Random Forest Model allows MarketSphere to minimize marketing misfires, focusing efforts on the most promising leads. The insights into important predictors can guide more personalized and effective marketing strategies, potentially increasing ROI on marketing campaigns.

**Boosted Trees Model**

The Boosted Trees Model stands out with the highest accuracy of 84.09%, indicating its superior ability to generalize predictions. The model's fine-tuning involved adjusting parameters such as the number of trees, interaction depth, shrinkage, and minimum observations in node, which collectively enhanced its predictive performance.

**Hyperparameters and Insights:**

```
Boosted Trees Model
> cat("Accuracy: ", gbmCm$overall['Accuracy'], "\n")
Accuracy:  0.8409091
> # Output the best tuning parameters
> print(gbmGridSearch$bestTune)
   n.trees interaction.depth shrinkage n.minobsinnode
17     100                 3       0.1             10
>
```

The selected optimal model utilized 100 trees, striking a balance between model complexity and minimizing the risk of overfitting, which is crucial for maintaining the model's performance on unseen data. An interaction depth of 3 was determined to be ideal, allowing the model to capture complex interactions among the variables without becoming overly complicated. The shrinkage parameter, or learning rate, was set at 0.1, moderating the pace at which the model learns. This cautious approach helps in avoiding overfitting by taking smaller

steps during gradient descent optimization. Additionally, the selection of 10 for the n.minobsinnode parameter from the tuning grid exemplifies a strategic balance between model complexity and predictive reliability. This choice ensures that each decision node within the model is based on a substantial subset of data, enhancing the model's ability to generalize well to new, unseen customer responses while avoiding overfitting.

**Business Implications:** The Boosted Trees Model's exceptional accuracy and detailed hyperparameter tuning process present MarketSphere with a cutting-edge tool for identifying potential campaign responders. This model's ability to handle complex data interactions suggests that MarketSphere could uncover deeper insights into customer behavior patterns, further refining marketing efforts.

*Comparison and Contrast of Ensemble Models - Classification models*

**Accuracy:** The Boosted Trees Model leads with an 84.09% accuracy, followed closely by the Random Forest Model at 82.58%, and the Bagged Trees Model at 80.3%. This indicates the Boosted Trees Model's superior ability to generalize predictions to unseen data.

**Hyperparameter Complexity:** The Boosted Trees Model required a more detailed hyperparameter tuning process, which, while resulting in higher accuracy, also demands more computational resources and time compared to Bagged Trees and Random Forest models.

**Variable Importance:** In the Random Forest Model, LastPurchase, Income, and Meats emerged as top predictors. The Boosted Trees Model likely relies on a similar set of important predictors, given its performance, but with potentially different weights and interactions due to the model's nature.

Model Interpretability: Random Forest and Bagged Trees offer relatively straightforward interpretations of variable importance and decision rules. In contrast, the Boosted Trees Model, while more accurate, can be more challenging to interpret due to the sequential nature of boosting and interaction effects.

Sensitivity to Overfitting: Boosted Trees, with its sequential correction of errors, can be more sensitive to overfitting compared to Bagged Trees and Random Forests, which rely on averaging and voting mechanisms to improve robustness.

**Business Implications:**

The Boosted Trees Model's high accuracy makes it a valuable tool for MarketSphere in identifying likely responders to marketing campaigns, potentially leading to more efficient allocation of marketing resources.

However, the trade-off in model complexity and interpretability may require additional efforts in model management and explanation to non-technical stakeholders.

The insights drawn from variable importance across models can inform MarketSphere on the critical factors influencing customer responses, guiding more targeted and personalized marketing strategies.

Strategic Recommendations for MarketSphere:

- Leverage the Boosted Trees Model for campaigns where maximum accuracy is critical, and computational resources are available.
- Utilize the Random Forest Model for regular marketing efforts, benefiting from its insights into important predictors and balance between sensitivity and specificity.
- Apply the Bagged Trees Model for broad, initial customer segmentation efforts, where a balanced approach is beneficial.

By integrating these models into their marketing strategies, MarketSphere can enhance the precision of their campaigns, ensuring that resources are allocated efficiently and effectively to maximize customer engagement and ROI.

## Regression

*Ensemble Method Analysis*

The primary purpose of ensemble methods in MarketSphere Inc. is to improve the prediction accuracy of how many bottles of wine customers will purchase, depending upon unique predictive characteristics such catalog purchases, in-store purchases and the number of children they have.

*Bagging*

The *ipred* package was used to create a bagged model using the same data, predictors and outcome as with the regressions trees and prior regression analysis. Our base bagged model produced an MSE of 39425.85 for the training subset and 41988.61 for the validation subset, an improvement form the best regression model

*Boosting*

The *gbm* package was used to construct a boosted model with the regression parameters discussed previously. The initial model produced a training MSE 40269.18 and a validation MSE of 42414.17, performing poorer than the bagged model but still an improvement from our best regression tree model comparing validation. We added different complexities using a grid search, including trying different tree depths, pulling out the best hyperparameters below and a validation RMSE of 205.75

| n.trees | interaction.depth | shrinkage | n.minobsinnode |
|---|---|---|---|
| 150 | 3 | 0.05 | 10 |

*Random Forests*

Trees can be combined into forests in order to achieve both low bias and lower variance, through the combination of ensemble methods that are efficient at reducing one or the other measure of accuracy. Random forests achieve this while also avoiding overfitting with multiple trees by sampling a random subset of features alongside observations in order to build trees. A random forest model was constructed on the same base parameters as the previous ensemble methods, producing an initial training MSE of 38924.03 and validation MSE of 43007.35. We applied a grid search similar to the boosted tree model, producing a final validation RMSE of 209.85.

In all, application of the three ensemble methods produced varying results in terms of the improved accuracy observed in transitioning from the training to validation subset, as well as the final MSE for both subsets. However, each model succeeded in generating a lower final MSE than the initial regression tree model, with our bagged model producing the lowest validation MSE of 41988.61 the random forest model producing the lowest initial training MSE of 38924.03.

*Comparative Analysis of the ensemble methods*

**Accuracy and MSE:** All three ensemble methods surpassed the performance of the initial regression tree model, with bagging showing the lowest validation MSE, suggesting its superior generalizability. Random Forests led in reducing the training MSE, highlighting its efficacy in learning from the training dataset.

**Model Complexity and Tuning:** Boosting and Random Forests involve more intricate hyperparameter tuning, which, while allowing for finely-tuned models, also increases the risk of overfitting if not carefully managed. Bagging stands out for its straightforward approach, emphasizing variance reduction with less emphasis on hyperparameter complexity.

**Trade-offs:** The choice between these methods hinges on the specific demands of the dataset and the predictive task at hand. Bagging offers a more generalized approach with minimal tuning, Boosting focuses on sequentially improving predictions, and Random Forests provide a balance between bias and variance through feature randomness.

# Part 3: Final Analysis

## Classification

Comparing the classification trees from Part 1 with the ensemble models from Part 2, we observe notable differences in performance and complexity.

The single classification trees provided a foundation with accuracies around 75%. They offered straightforward interpretations of the data through clear-cut decision rules. However, they lacked the robustness provided by ensemble methods, which brought together multiple models to improve prediction accuracy and combat overfitting.

The ensemble models outperformed the single trees with the Bagged Trees model achieving an 80.3% accuracy, the Random Forest model 82.58%, and the Boosted Trees model leading with 84.09%. These methods are particularly effective against overfitting—Bagged Trees through aggregation of various decision trees, Random Forest by introducing randomness in feature selection, and Boosted Trees through sequential improvement of predictions.

*Underfitting and Overfitting*

The ensemble models are less prone to overfitting due to their aggregated nature, which averages out biases and reduces variance. Hyperparameter tuning played a vital role in avoiding overfitting by optimizing model parameters to generalize well to new data. Cross-validation further ensured that the models were not tuned to the idiosyncrasies of the training data alone, thereby avoiding underfitting.

*Interpretability vs. Accuracy Tradeoff*

While ensemble models improved accuracy, they did so at the cost of interpretability. Single trees are much easier to understand and communicate to stakeholders, as they provide clear decision paths and rules. Ensemble models, however, operate as a black box, making them more complex and harder to dissect.

## Regression

*Regression Ensemble Models vs. Simple Regression Trees: A Comparative Analysis*

Model Complexity and Predictive Accuracy:

Our journey through predictive modeling for MarketSphere commenced with simple regression trees (Part 1) and evolved into the exploration of ensemble models (Part 2). Simple regression trees offered an initial understanding with a training MSE of 41,053.399 and a validation MSE of 44,428.08. These models, while easily interpretable, were constrained by their simplistic approach to complex data patterns. In contrast, ensemble methods—Bagging, Boosting, and Random Forests—introduced a more sophisticated analysis framework. For instance, the Bagging approach reduced validation MSE to 41,988.61, showcasing its effectiveness in mitigating errors through model aggregation. Boosting further refined the model's accuracy, achieving a validation MSE of 42,414.17 by iteratively correcting previous errors. Random Forests stood out by achieving the lowest initial training MSE of 38,924.03, attributed to its strategy of combining diverse trees built on varied data and feature subsets.

**Balancing Bias and Variance**

The shift to ensemble methods marked a strategic move towards balancing bias and variance, crucial for avoiding overfitting and underfitting. Ensemble models excel in this aspect by leveraging multiple models to average out individual biases and reduce variance. For example, Random Forests utilize randomness in feature selection to ensure model diversity, thus reducing the risk of overfitting. Our tuning of hyperparameters, such as the mtry value in Random Forests (optimal at 6), played a pivotal role in optimizing model performance, ensuring they generalize well to unseen data without sacrificing complexity for accuracy.

**Interpretability vs. Predictive Power**

A notable trade-off emerged between model interpretability and predictive power. Simple regression trees provide clear, straightforward insights into how features like CatalogPurchases and StorePurchases impact wine purchases. However, their simpler nature often limits predictive accuracy. Conversely, ensemble models, with their layered, complex structures, significantly enhance predictive accuracy but at the expense of straightforward interpretability. For instance, while Random Forests offer valuable insights through variable importance measures, the collective decision-making process within these models complicates the direct interpretation of feature impacts.

**Implications for MarketSphere**

For MarketSphere, the choice between simple regression trees and ensemble models depends on the specific business needs—whether the emphasis is on obtaining the highest predictive accuracy to inform marketing strategies or maintaining model simplicity for easier stakeholder communication. While ensemble models promise improved predictive performance, essential for identifying potential customer responses accurately, they require careful consideration of their complexity and the challenges it poses for interpretation.

*Conclusion and Insights for stakeholders*

Each approach offers distinct advantages and drawbacks, emphasizing the importance of aligning model selection with business objectives and the ability to communicate findings effectively within MarketSphere.

For MarketSphere, choosing between these models involves balancing the need for accuracy with the ability to explain decisions to stakeholders. If the highest prediction accuracy is paramount, and the complexity of the models can be managed, ensemble methods are the way forward. However, if interpretability is crucial for implementation and stakeholder buy-in, single classification trees may be more appropriate, despite their slightly lower performance.

The exploration of classification and regression models, from foundational techniques to complex ensemble methods, reveals critical insights for various business sectors. These insights emphasize the importance of leveraging predictive analytics to improve operational efficiency, optimize marketing strategies, and enhance financial risk management. Leaders are advised to balance model complexity with interpretability to ensure insights remain actionable, prioritize data literacy to foster a data-driven culture, and stay updated with data science advancements to sustain a competitive advantage. Ethical considerations and mitigating biases in modeling are imperative to ensure fairness and transparency. Ultimately, integrating data science into business strategies not only unlocks new opportunities for innovation but also necessitates responsible management and ethical consideration in its application.

## *Reflection*

Reflecting on the entirety of this analytical project, spanning from the intricacies of classification to the depths of regression, one can't help but marvel at the journey undertaken. It was a journey that moved from the clear, interpretable simplicity of single decision trees to the nuanced accuracy of ensemble methods, each step revealing new insights and complexities within the data.

The project began with a focus on classification, where initial models laid bare the factors influencing customer responses. This phase highlighted the value of interpretability, with single trees providing a straightforward view into the dataset. Yet, it also underscored their limitations, setting the stage for the introduction of ensemble methods. Bagged Trees, Random Forests, and Boosted Trees each brought a unique lens to the analysis, enhancing predictive accuracy and combating overfitting through their collective approach. Hyperparameter tuning and cross-validation were pivotal, refining the models to ensure they captured the essence of the data without being swayed by its idiosyncrasies.

Transitioning to regression, the narrative expanded to encompass the prediction of continuous outcomes. Here again, the progression from basic linear regression to ensemble methods like bagging, boosting, and random forests unfolded. The exploration through regression revealed the complex interplay of variables and the critical role of hyperparameter tuning in optimizing model performance. Despite the increased predictive accuracy of ensemble methods, the challenge of balancing model complexity with interpretability persisted.

Ultimately, it is quite common for models to have some degree of overfitting, especially when dealing with complex datasets. While many attempts were made to manage the regression model's complexity, specifically controlling its depth size, minimum split size and complexity parameter, the MSE of the validation set remained higher than the training MSE. This, though, indicates a valiant attempt at balancing model complexity and we will look to improve this to further reduce the model's overfitting tendencies and improve the model's ability to predict underlying patterns. Additionally, our complex dataset which we used to develop our regression tree model and ensemble models was originally collected for clustering purposes; therefore, tailoring and/or restructuring our dataset to be better suited for regression analysis would likely improve its predictive capabilities, reduce overfitting and give us more flexibility in managing the model's complexity.

Across both classification and regression analyses, the significance of variables, the impact of performance metrics, and the comparisons between models painted a comprehensive picture of predictive modeling. It was a journey that not only aimed at enhancing model accuracy but also at understanding the underlying dynamics of the data.

The business implications of these insights cannot be overstated. For manufacturing, marketing, finance, and other business functions, the models offer a pathway to predictive maintenance, quality control, customer segmentation, credit scoring, fraud detection, and supply chain optimization. Yet, as models grow in complexity, the trade-off between accuracy and interpretability becomes increasingly pronounced, presenting a conundrum for business leaders and decision-makers.

This project has been a testament to the art and science of data analytics. It showcased the delicate balance between embracing model complexity for accuracy's sake and striving for the simplicity that fosters understanding and actionable insights. For leaders across industries, the journey underscores the importance of fostering a data-driven culture, one that values not just the insights gleaned from sophisticated models but also the clarity in communication and decision-making that comes from interpretable analyses.

# 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