Sales and Brand Analysis Report

Daryle Bilog

Cover Page

Daryle Bilog

Title: Sales and Brand Analysis Report

Define the Problem

As a new data scientist, the Brand Manager and Sales Manager express their interest in improving the performance of the orange juice category. Based on the historical data, Minute Maid (MM) yields higher margins compared to Citrus Hill (CH), thus, both managers aim to increase the sales of this brand. There are two sections of problems that we need to solve for this project. Below are the detailed problem statements for each section:

Brand Manager's Questions: The Brand Manager is interested in understanding the significant variables that affects the probability of a customer to buy MM. They specifically want to know the effectiveness of the variables indicated on the dataset provided. In addition, the Brand Manager seeks for recommendations to increase the sale of MM.

Sales Manager's Questions: On the other hand, the Sales Manager wants to build a predictive model that can inform them about the probability of customer on buying MM.

The confidence for each of the recommendations and predictive model should be stated by the end of the project to inform both the Brand Manager and Sales Manager how confident the solutions will be.

Overall, the goal of the project is to address the specific needs of both the brand manager and the sales manager using appropriate statistical analyses and provide actionable recommendations based on the findings.

Define Methods

Did you scale/standardize variables? What type of preprocessing did you perform?

-Yes, I scaled/standardized the variables. Standardizing variables is a common preprocessing step, especially when using models like logistic regression and boosted trees, as it ensures that all variables contribute equally to the analysis. I used the preprocess function from the recipes package to scale numerical predictors.

What efforts did you make to reduce overfitting (i.e., train/test split, cross-validation, etc.)?

-To reduce overfitting, I employed a train/test split. I used the initial_split function from the rsample package to divide the dataset into training and testing sets. Additionally, I used cross-validation during the

hyperparameter tuning phase for the boosted tree model to ensure the generalization of the model to new data.

Describe the data and variables that you used in your analyses.

-The dataset (OJ) contains information on various factors influencing sales, including pricing, discounts, special offers, and customer loyalty. I used predictors such as PriceCH, PriceMM, DiscCH, DiscMM, SpecialCH, SpecialMM, LoyalCH, and others. The target variable is Purchase, indicating whether a purchase was made.

What happens when you include all variables?

-When including all variables, the logistic regression model and boosted tree model have a larger feature space. This may lead to overfitting, increased complexity, and potential issues with multicollinearity. Including all variables might result in a less interpretable and less generalizable model.

Is it better to not include some variables? Why? Provide methodological rationale for your responses and any criteria you used to include or exclude variables. Did multicollinearity play a role in your decision to exclude some variables.

-Yes, it is often better to exclude some variables. I used variable importance measures, correlation analysis, and variance inflation factor (VIF) to identify and exclude variables contributing less information or causing multicollinearity. Excluding variables can lead to a more parsimonious model, reducing overfitting and enhancing interpretability.

Was the predictive performance of the Logistic and Boosted Trees model comparable when using metrics such as accuracy, ROC-AUC?

-The predictive performance of the models was assessed using metrics such as accuracy and ROC-AUC. Both models were tuned and evaluated using the same metrics. The final models demonstrated comparable performance; however, the choice between them might depend on the specific requirements of the stakeholders and the trade-off between interpretability (logistic regression) and predictive accuracy (boosted trees).

When you use the XAI method PDP with the Gradient Boosted Trees, did you observe the same pattern of influence of predictor variables on the outcome as you see in the logistic regression (e.g., PDP shows predictor variable X positively affects Outcome Y. Does Logistic regression also show a positive influence of X on Y?).

-The PDPs from the Gradient Boosted Trees model might reveal different patterns of variable influence compared to logistic regression. Boosted trees can capture complex interactions, whereas logistic regression assumes linear relationships. While some variables may exhibit consistent effects, the patterns might not be identical due to differences in model assumptions and structures.

Explain in detail the analyses you conducted and any assumptions you made.

-I conducted exploratory data analysis, assessed multicollinearity, built logistic regression and boosted tree models, and performed hyperparameter tuning using cross-validation. Assumptions included linearity in logistic regression, the absence of multicollinearity, and the appropriateness of the chosen hyperparameters for the boosted tree model. Interpretations were made based on variable importance, coefficient estimates, and PDPs.

Analysis Methods

I utilized logistic regression and boosted tree models to analyze the data. Logistic regression helps identify significant predictors for the outcome, while boosted trees offer a more complex, non-linear approach. I chose these methods for their interpretability and predictive power.

Preprocessing

- **Scaling/Standardization:** Variables were scaled for logistic regression.
- **Data Description:** The dataset includes information on prices, discounts, loyalty, and other factors influencing sales.
- **Variable Selection:** Considered multicollinearity and removed variables with high VIF.

Model Comparison

- **Logistic Regression:** Emphasizes linear relationships.
- **Boosted Trees:** Captures complex, non-linear patterns.

Results and Conclusion

Logistic Regression Results

- Identified significant predictors: "PriceCH," "PriceMM," "DiscMM," "LoyalCH," and "PctDiscMM."
- Achieved an accuracy of 82.5% on the test set.

Boosted Tree Results

- Tuned hyperparameters: trees = 150, tree depth = 3, learning rate = 0.1.
- Achieved an accuracy of 19.53% on the test set.

Model Comparison

- Logistic and boosted tree models differed significantly in predictive performance.
- Logistic regression outperformed the boosted tree in accuracy.

XAI Method - Partial Dependence Plots (PDP)

- Examined PDP with boosted trees.
- Compared the influence of predictor variables on outcomes between logistic regression and boosted trees.

Recommendations

Brand Manager Analysis

Logistic Regression Model:

- **Objective**: Identify significant variables influencing MM purchases.
- **Method**: Employ logistic regression for variable selection and interpretability.
- **Preprocessing**: Address multicollinearity through variable removal.
- **Significant Variables**: PriceCH, PriceMM, DiscMM, LoyalCH, PctDiscMM.

Partial Dependence Plots (PDP):

- **Action**: Analyze PDPs for key variables (LoyalCH, DiscMM).
- **Insight**: Understand how changes in predictors impact the probability of purchasing MM.

Brand Manager Recommendations:

Optimize Pricing Strategies:

- **Insight**: PriceMM has a positive influence, while PriceCH has a negative impact.
- **Strategy**: Adjust PriceMM to maximize profitability and consider promotions on PriceCH strategically.

Leverage Customer Loyalty:

- **Insight**: LoyalCH significantly influences MM purchases.
- **Strategy**: Implement loyalty programs or targeted promotions for MM to enhance customer retention.

Promotional Strategies:

- **Insight**: DiscMM positively affects MM purchases.
- **Strategy**: Implement targeted discounts or promotions for MM to boost sales.

Sales Manager Analysis

Gradient Boosted Trees Model:

- **Objective**: Build a predictive model for MM purchases.
- **Method**: Utilize boosted trees for capturing complex relationships.
- **Preprocessing**: Tune hyperparameters for optimal performance.

Optimize Predictive Model:

- **Insight**: Achieve an accuracy of 80.9% and ROC AUC of 88.4%.
- **Strategy**: Implement the boosted tree model for predicting MM purchases.

Monitoring and Refinement:

- **Insight**: Sensitivity (True Positive Rate) is low.
- **Strategy**: Regularly monitor model performance and refine as needed to improve predictions.

Promotional Campaigns:

- **Insight**: Use the ROC Curve to adjust the model threshold for promotional campaigns.
- **Strategy**: Identify an optimal threshold for targeted promotional campaigns based on the ROC Curve.

Conclusion

The analysis provides valuable insights for both managers. The choice of logistic regression for the Brand Manager's objective of understanding the variables that influence a customer's probability of buying Minute Maid (MM) is based on several considerations like the interpretability of the coeffecients that this model provides on the analysis. These coeffecients represents the log-odds of each of teh variables which makes it easy to interpret the impact of each predictor. The analysis also considers the testing of statistical significance of each coeffecient which is important for the Brand Manager to focus on the most influential factors.

Moreover, logistic regression is less prone to multicollinearity issues compared to some other models. Multicollinearity can make it challenging to interpret the individual impact of predictors, and logistic regression is relatively robust in this regard.

On the other hand, Gradient Boosted Trees (GBT) is used for the Sales Manager's objective on building predictive model to estimate the probability of someone buying Minute Maid (MM). This can be justified because GBT is well-suited for more complex interactions between variables. The model that was created achieved an accuracy of 80.9% and an ROC AUC of 88.4%, indicating a strong ability to discriminate between MM purchases and non-purchases.

The ROC Curve analysis allows for the adjustment of the model's threshold based on the trade-off between sensitivity and specificity. This is crucial for customizing the model to the Sales Manager's specific needs, such as optimizing promotional campaigns.

Overall, GBT offers a robust approach on building predictive model that the Sales Manager needs for the objective of this analysis while Logistic Regression was chosen as the model for the Brand Manager due to its interpretability, simplicity, and sustainability which can provide a clear insight into variables influencing MM purchase probabilities.
