**Final Project – MKTG 6620**

Biva Sherchan

<https://posit.cloud/content/7032095>

**Business Problem**

Our grocery store chain sells two brands of orange juice Citrus Hill (CH) and Minute Maid (MM). Because the MM gets higher margins than CH, the Brand Manager wants to increase the customer’s probability of buying MM whereas the Sales Manager wants to be able to predict the probability of a customer purchasing MM based on the historical data. Both of the manager’s goals are to make the Orange Juice category perform better than it does currently.

My goal in this project is to 1. find the variables that significantly influence the customer’s probability of buying MM. This will help the Brand manager to develop effective strategies to increase the chances of customers buying MM and hence increase the profit margin of Orange Juice category 2. build a classification predictive model using a supervised machine learning algorithm that will be able to predict the probability of a customer purchasing MM. This predictive model will help the Sales manager to forecast the sales of MM which can be used to efficiently allocate the cash flows and resources, and to plan inventory levels and prevent the bull-whip effect.

**Method**

The dataset used for this project contains 1070 purchase history in which the customer either purchased Citrus Hill (CH) or Minute Maid (MM) Orange juice. Our target variable is “Purchase” which is a factor with levels 0 and 1 indicating whether the customer purchased Citrus Hill (1) or Minute Maid Orange Juice (0), and the dataset has 13 independent variables (listed in Appendix). For data preprocessing I started by switching the binary values of target variable Purchase since we are interested in MM, I updated the value for MM to be 1 and CH to be zero. This is not necessary for the types of models I built and compared for this project but is something I did for my own ease of understanding. Next step was to check if there are any null values that I need to impute but I did not find any. I also checked for the data imbalance by checking the proportion of the target variable and found 39% for 0 value and 61% for 1 value which indicates that the distribution of target variable is quite balanced, so down-sampling of majority class is not required in the training data set. Lastly, I checked for outliers in the dataset using the MCD method, but no data points were classified as an outlier. For the goal of this project, I decided to create different classification models using two different types of supervised machine learning algorithms, Gradient Boosted Trees (XGBoost) and Logistic Regression (glm). I will also use Lasso Regression during the process of variable selection.

Lasso Regression (Least Absolute Shrinkage and Selection Operator) is a regularization technique used in linear regression that introduces a penalty to the regression coefficients. It aims to improve the prediction accuracy and interpretability of the statistical model by simultaneously performing variable selection and regularization. Lasso Regression performs variable selection by shrinking coefficients to zero, resulting in a sparse model, effectively reducing the impact of multicollinearity on the model. However, it does not give any information about the significance of the predictor variables on the target variable, so I use this model mainly just for variable selection process.

Gradient Boosted Trees model implements regularization to prevent overfitting, reducing variance and improve generalization. It also implements tree pruning to avoid overfitting by removing splits that don't add significant value. While Gradient Boosted Trees is a black-box method and is largely opaque which provides predictions without clear or interpretable explanations and does not directly provide the coefficients like generalized linear models. But it allows us to gather some insights about how the different predictor variables affect the target variable and the nature of the relationship, and which played an important role using one of the XAI method, partial dependency plots. Classification trees offer many advantages such as their capability to handle several types of predictor variables, being uninfluenced by outliers, being unaffected by irrelevant predictor variables included in the analysis and being capable of handling multicollinearity among the predictor variables.

Regression model is a transparent model and easily interpretable model that provides the magnitude, direction and significance of influence of each predictor variable on the target variable. However, regression model assumes a linear relationship between features and the log-odds, which might not hold in all cases.

To reduce overfitting, I split the data into training (80%) and test (20%) sets. Training set will be used to train the models and test set will be used for performance evaluation of the models. The performance metrics I will use to compare the different models are Accuracy rate, Sensitivity/Recall/True Positive Rate (TPR) and Specificity/True Negative Rate and roc\_auc curve value. I will also use AIC score to compare the different logistic regression models.

For Gradient Boosted model, I tuned the three hyperparameters trees, tree\_depth and learn\_rate by using a grid search for four levels of each of the three hyperparameters which will result in 4^3=64 combinations of the hyperparameters. To find out which of the 64 possible combinations of the hyperparameters is optimum, I ran a fivefold cross validation so that there will be five partitions of the training data where each of the decision trees will be fit on the four partitions and the performance will be assessed on the fifth partition. The best model is selected based on the highest roc\_auc curve value. The analysis provided the best set of hyperparameter values as trees (number of trees contained in the ensemble) = 667, tree\_depth (the maximum depth of the tree) = 1 and learn\_rate (learning rate) = 0.1. The XGBoost model produced an accuracy rate of 80% and roc\_auc score of 0.87.

When fitting a Logistic Regression model on all the predictors, the model summary gave an error message "Coefficients: (4 not defined because of singularities)" which indicates that the coefficients for some predictor variables could not be estimated due to the presence of perfect or near-perfect multicollinearity. This means that there are multicollinearity issues among predictor variables in the dataset. Per the correlation plot (Fig9.), we see some perfect and some high multicollinearity. Variables like DiscCH and PctDiscCH, DiscMM and PctDiscMM have a perfect collinearity with value of 1. We can also see some high collinearity values between SalePriceMM, PriceDiff, DiscMM and PctDiscMM. We need to remove multicollinearity by removing highly correlated variables because multicollinearity leads to inflated standard errors of regression coefficients, making it difficult to assess the statistical significance of predictors. It also causes coefficients to be sensitive to small changes in the data, making them less reliable and interpretable.

From the VIP plot (Fig1.), we can see that the variables LoyalCH, PriceDiff, and ListPriceDiff are the three most important predictors, LoyalCH being the most important variable and has a pretty high importance value. PriceDiff is the second most important variable, but the importance is much lower than LoyalCH. To examine how these predictors influenced the target variable Purchase, I used the partial dependency plots. From the PDP plots we can see that predictors LoyalCH (Fig2.), PriceDiff (Fig3.) and DiscCH (Fig4.) have pretty linear negative relationship with the target variable Purchase indicating that the increase in the values of these predictors i.e., the higher the probability of customer buying CH (over MM), or the higher the sale price difference of Minute Maid Orange vs Citrus Hill juice, or the higher the discount offered for Citrus Hill juice, we will see lower chances of customers buying Minute Maid orange juice. The PDP plots for PctDiscMM(Fig5.), PctDiscCH (Fig6.), PriceMM (Fig7.) and SalePriceCH (Fig8.) are flat indicating that these variables do not have any effect on the target variable Purchase which is what we see later that Lasso regression shrunk the coefficients of these variables to zero.

For variable selection, I ran Lasso regression model to analyze the variables the model did not shrink to zero which were DiscCH, LoyalCH, PriceDiff and ListPriceDiff. Per the correlation plot (Fig10.), we see that these set of variables did not have high correlation and VIF values were all below 2. These are also 4 of the 5 most important variables in the VIP plot (Fig1.). The 4th important variable in VIP plot SalePriceMM was found to have high collinearity with PriceDiff (Fig7.) and hence was not selected as part of the final predictor variables. Lasso regression also had shrunk the coefficient of SalePriceMM to zero. Hence, my final selected variables are DiscCH, LoyalCH, PriceDiff and ListPriceDiff which I used to build Logistic Regression model (glm\_model2). The model produced an AIC score of 646.1, accuracy rate of 80%, sensitivity of 82.44% and specificity of 76.19%. Sensitivity measures the proportion of actual positive cases that are correctly identified by the model. The model can correctly identify positive cases i.e., chances of customer buying MM over CH 82% of the time. Specificity measures the proportion of actual negative cases that are correctly identified by the model. The model can correctly identify negative cases i.e., chances of customer buying CH over MM 76% of the time.

From the Logistic regression model (glm\_model2) summary (FigB.), we can see that only two predictor variables have a statistically significant effect on the target variable which are LoyalCH and PriceDiff. LoyalCH has the most statistically significant effect on the target variable with p-value <2e-16, the second one being PriceDiff with p-value 1.64e-08. Every unit increase in the customer brand loyalty for CH (i.e., the probability of customer buying CH over MM), decreases the log odds of buying MM vs buying CH by 6.79. Every unit increase in the sale price of MM less sale price of CH, decreases the log odds of buying MM vs buying CH by 2.71. This is consistent with the result we found from the VIP and PDP plots. I decided to create one more Logistic Regression model using only the two significant variables LoyalCH and PriceDiff (glm\_model3). The model produced the same results for accuracy, sensitivity and specificity as that of glm\_model2 but had a higher AIC score of 647.02 indicating that glm\_model2 fit the training data better. Both the logistic regression models glm\_model2 and glm\_model3 produced a roc\_auc score of 0.86 (FigC.) which is slightly lower than that of XGBoost model which is 0.87 which indicates that XGBoost model is a better performing model and can better distinguish positive and negative classes.

A graph with black dots

Description automatically generated

Fig 1. VIP Plot

A graph showing a line

Description automatically generated with medium confidence

Fig 2. PDP Plot for LoyalCH

A graph showing a line

Description automatically generated with medium confidence

Fig 3. PDP Plot for PriceDiff

A graph with a line

Description automatically generated

Fig 4. PDP Plot for DiscCH

A graph with blue lines

Description automatically generated

Fig 5. PDP Plot for PctDiscMM

A graph with blue lines

Description automatically generated

Fig 6. PDP Plot for PctDiscCH

A graph with blue lines

Description automatically generated

Fig 7. PDP Plot for PriceMM

A graph with blue lines

Description automatically generated

Fig 8. PDP Plot for SalePriceCH

A graph with numbers and letters

Description automatically generated

Fig 9. Correlation Plot

A graph with different colored squares

Description automatically generated

Fig 10. Correlation Plot of selected variables

**Results and Conclusion**

From the summary (FigB.) of Logistic regression model built using the selected predictor variables LoyalCH, PriceDiff, ListPriceDiff and DiscCH (glm\_model2), we can see that only two predictor variables have a statistically significant effect on the target variable which are LoyalCH and PriceDiff. LoyalCH has the most statistically significant effect on the target variable with p-value <2e-16, the second one being PriceDiff with p-value 1.64e-08. Every unit increase in the customer brand loyalty for CH (i.e., the probability of customer buying CH over MM), decreases the log odds of buying MM vs buying CH by 6.79. Every unit increase in the sale price of MM less sale price of CH, decreases the log odds of buying MM vs buying CH by 2.71. This is consistent with the relationship we found from the VIP and PDP plots. The fact that the PDP plot of PriceMM is flat shows that the probability of a customer buying MM is not affected by just the price of MM but how the price of MM compares to that of CH i.e., PriceDiff. This information will help the Brand manager to develop effective strategies to increase the customer’s probability of buying Minute Maid (MM) orange juice and hence increase the profit margin of Orange Juice category. Since the customer brand loyalty for CH i.e., Loyal CH is a significant variable in determining the chances of a customer buying MM vs CH, we can use this information to segment customers based on loyalty level and tailor marketing campaigns like messages, promotions, or discounts directed at customers with low level of loyalty for CH as they are more likely to buy MM over CH than the customers with high level of loyalty for CH. More analysis can be done to find what are the factors that make a customer loyal to CH and how we can set the optimal price difference between MM and CH in such a way that customers will switch to MM.

The XGBoost model produced an accuracy rate of 80% and roc\_auc score of 0.87. The model built using the selected variables (glm\_model2) produced an AIC score of 646.1, accuracy rate of 80%, sensitivity of 82.44% and specificity of 76.19%. Sensitivity measures the proportion of actual positive cases that are correctly identified by the model. The model can correctly identify positive cases i.e., chances of customer buying MM over CH 82% of the time. Specificity measures the proportion of actual negative cases that are correctly identified by the model. The model can correctly identify negative cases i.e., chances of customer buying CH over MM 76% of the time. The model built using the two significant predictor variables LoyalCH and PriceDiff (glm\_model3) produced the same results for accuracy, sensitivity and specificity but had a higher AIC score of 647.02 indicating that glm\_model2 fit the training data better. Both the logistic regression models glm\_model2 and glm\_model3 produced a roc\_auc score of 0.86 (FigC.) which is slightly lower than that of XGBoost model 0.87 which indicates that XGBoost model is a better performing model and can better distinguish positive and negative classes. ROC curve is a graphical representation that illustrates the trade-off between the true positive rate (sensitivity) and the false positive rate (1 - specificity) at various thresholds. The curve plots the true positive rate against the false positive rate, showing how well the model distinguishes between classes as the discrimination threshold changes. The AUC value quantifies the overall performance of the model across all possible classification thresholds. It represents the area under the ROC curve, providing a single scalar value to measure the model's ability to distinguish between classes. Hence, my final prediction model is the XGBoost model which the sales manager can use to predict the probability of a customer purchasing MM. This predictive model will give the output of probabilities of customer buying MM. We can set a specific threshold value depending on the company’s tolerance for False Positive Rate and True Positive Rate to convert the probabilities into binary outcomes i.e., customer will buy MM vs will not buy MM. This prediction model will help the Sales manager to forecast the sales of MM which can be used to efficiently allocate the cash flows and resources, and to plan inventory levels to reduce the bull-whip effect. The sales manager can allocate resources effectively by segmenting the customer base based on their likelihood to purchase MM and focusing on customers with a higher predicted probability of purchasing MM. This includes developing specific strategies for each segment, directing sales efforts, promotions, and resources towards these segments to maximize returns and ensuring a more personalized and effective approach to cater to their needs and preferences.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **XGBoost Model** | **Logistic Regression Models** | |
| **Metrics** |  | glm\_model2 | glm\_model3 |
| **AIC** |  | 646.1 | 647.02 |
| **Accuracy on test set** | 80% | 80% | 80% |
| **Sensitivity/Recall/True Positive Rate** |  | 82.44% | 82.44% |
| **Specificity/True Negative Rate** |  | 76.19% | 76.19% |
| **ROC\_AUC** | 0.87 | 0.86 | 0.86 |

FigA.

A screenshot of a computer program

Description automatically generated

FigB.

A graph with a blue and red line

Description automatically generated

FigC.

**Appendix**

OJ is a data frame with 1070 observations for the following 14 variables.

Purchase: A factor with levels 0 and 1 indicating whether the customer purchased Citrus Hill (1) or Minute Maid Orange Juice (0)

PriceCH: Price charged for CH. Also called List Price for CH

PriceMM: Price charged for MM. Also called List Price for MM

DiscCH: Discount offered for CH

DiscMM: Discount offered for MM

SpecialCH: Indicator of special on CH. Special can be a free gift, loyalty points etc.

SpecialMM: Indicator of special on MM. Special can be a free gift, loyalty points etc.

LoyalCH: Customer brand loyalty for CH. That is, the probability to buy CH (over MM) based on prior purchase behavior.

SalePriceMM: Sale price for MM. This is the difference between the list price and discount. SalePriceCH: Sale price for CH. This is the difference between the list price and discount. PriceDiff: Sale price of MM less sale price of CH

PctDiscMM Percentage discount for MM PctDiscCH: Percentage discount for CH

ListPriceDiff: List price of MM less list price of CH