# Unveiling the Orange Juice Market Landscape

by
Abhiram Mannam
Kushal Ram Tayi

# Table of Contents

# Problem Statement

**Background**

The grocery store chain aims to elevate the performance of its orange juice category, focusing on improving sales of both Citrus Hill (CH) and Minute Maid (MM) brands. In the grocery store chain, the Minute Maid (MM) gets higher margin than Citrus Hill (CH) in terms of returns on per unit sale.

**Sales Manager challenges and objective**

The Sales Manager is facing with the challenge of predicting customer behavior within the orange juice category, specifically focusing on Minute Maid (MM) sales.
The Sales Manager expects the data scientist to construct a predictive model that reliably estimates the probability of customers choosing for Minute Maid.
The model should be accurate and capable of making actionable predictions to drive MM sales growth.

**Brand Manager challenges and objective**

The Brand Manager is focused on understanding variables influencing the choice between Citrus Hill (CH) and Minute Maid (MM).
The main challenge of Brand Manager is to understand which factor significantly influence customers' inclination towards purchasing MM over CH is crucial.
The Brand Manager anticipates the data scientist to uncover influential variables impacting customers' preference for MM and provide actionable recommendations to increase the likelihood of customers selecting Minute Maid (MM) over Citrus Hill (CH).

**Data Scientist Role**

The data scientist plans to use machine learning models based on each manager's objectives. For the Brand Manager's challenges, techniques like logistic regression will be developed to identify influential variables. Meanwhile, predictive modeling approaches like boosted trees model will be utilized to fulfill the Sales Manager's objective of creating an accurate predictive model.

Additionally, careful consideration will be given to the dataset's features and their impact on MM purchases. Measures to handle multicollinearity, feature selection, and model interpretation will be implemented to ensure strategizing sales tactics and explainability in the analysis.

# Methods

**Dataset**

Given dataset contains 1070 purchases in which the customer either purchased Citrus Hill (CH) or Minute Maid (MM) Orange Juice.
The target variable is Purchase" which indicates whether the customer chose Minute Maid (0) or Citrus Hill (1).

**Preprocessing**

The dataset contains 83 rows of duplicate values, and having duplicates in the dataset doesn't add any new information to the model as the variability of the duplicate data is shown by other set of duplicates. So, dropping the duplicate values increases the data quality.

We have no missing data in the dataset and all columns have no NAs in them.

We converted the columns 1 through 13 from the original dataset (data) to numeric values, and the target variable 'Purchase' to factor.

**Class Imbalance**

The target variable has class imbalance as one class (Citrus Hill (1)) 608 count whereas the other (Minute Maid (0)) has 379 counts. This imbalance may affect the model and result in biased predictions towards the majority class. This shows that Citrus Hill (1) has 61.6 % of the target values whereas Minute Maid (0) has 38.39%.
To address this class imbalance problem, we have done under sampling from the 'Imblearn' library - under sampling method in python to get the same class values.
After under sampling the target variable contains an equal number of instances '379' for both classes (0 and 1).

**Scale/standardize variables**

Certain models exclusively utilize standardized or scaled data for efficient interpretation. We applied Standard Scaling using the Python library 'StandardScaler' to preprocess the dataset, generating scaled data. This approach is good for interpretability by providing a clear understanding of the predictors' influence on the target outcome.

**Cross-Validation**

Cross_validation of the dataset into X_Train (Train dataset) and X_valid (validation dataset) with 80%-20% respectively for the validation of the model and to know about overfitting. Also, done for target variable into 80-20% split for cross validation.

**Logistic Regression**

Logistic regression is a statistical model used for binary classification. It calculates the probability of an observation belonging to a particular class based on input features and applies a logistic function to map the output in range between 0 and 1. The class with the highest probability is predicted as the outcome.

First, we used all the variables to fit the data into the logistic model with 'Purchase' as the target variable and here we get the results detailed in the figure 1.

Logit Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | Purchase | No. Observations: | 606 |
| Model: | Logit | Df Residuals: | 596 |
| Method: | MLE | Df Model: | 9 |
| Date: | Mon, 20 Nov 2023 | Pseudo R-squ.: | 0.4359 |
| Time: | 05:41:12 | Log-Likelihood: | -236.89 |
| converged: | False | LL-Null: | -419.93 |
| Covariance Type: | nonrobust | LLR p-value: | 2.338e-73 |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 0.0799 | 0.117 | 0.685 | 0.493 | -0.149 | 0.309 |
| x1 | -0.4408 | 1.36e+06 | -3.24e-07 | 1.000 | -2.67e+06 | 2.67e+06 |
| x2 | -0.4454 | 3.38e+06 | -1.32e-07 | 1.000 | -6.62e+06 | 6.62e+06 |
| x3 | -0.9738 | 4.58e+06 | -2.13e-07 | 1.000 | -8.98e+06 | 8.98e+06 |
| x4 | -1.7503 | nan | nan | nan | nan | nan |
| x5 | -0.0619 | 0.146 | -0.425 | 0.671 | -0.347 | 0.224 |
| x6 | -0.1336 | 0.139 | -0.960 | 0.337 | -0.407 | 0.139 |
| x7 | 2.0884 | 0.166 | 12.585 | 0.000 | 1.763 | 2.414 |
| x8 | 1.2714 | 1.87e+06 | 6.8e-07 | 1.000 | -3.66e+06 | 3.66e+06 |
| x9 | 0.4466 | 4.56e+06 | 9.8e-08 | 1.000 | -8.93e+06 | 8.93e+06 |
| x10 | 0.9577 | nan | nan | nan | nan | nan |
| x11 | 3.1609 | 2.500 | 1.265 | 0.206 | -1.738 | 8.060 |
| x12 | 1.3408 | 2.554 | 0.525 | 0.600 | -3.665 | 6.347 |
| x13 | -0.1447 | 2.07e+06 | -6.99e-08 | 1.000 | -4.05e+06 | 4.05e+06 |

Figure: 1. Logistic model summary with all variables.

Here we could see that there are some 'nan' in the summary which could mean that there is multi collinearity between the columns in the dataset. Logistic regression is susceptible to multicollinearity. So, to view the multicollinearity between the columns in the dataset we use Variance Inflation Factor (VIF).

**Model metrics**

Here the AIC for this logistic model is 501.779.
Accuracy: 0.84
RMSE: 0.41
ROC-AUC Score: 0.8371080139372823
Confusion Matrix:
[[67 15]
 [10 60]]
True Positives: 60
True Negatives: 67
False Negatives: 10
False Positives: 15
Sensitivity (Recall): 0.8571428571428571
Precision: 0.8
Specificity: 0.8170731707317073

**Multicollinearity**

VIF (Variance Inflation Factor) calculates the multicollinearity between the variables in the train dataset. VIF values range between -inf to +inf with high collinearity has VIF value of greater than 5. The multicollinearity values for the dataset are calculated using the 'variance_inflation_factor' from 'statsmodels' library in python.

| Variable | VIF |
|---|---|
| PriceCH | inf |
| PriceMM | inf |
| DiscCH | inf |
| DiscMM | inf |
| SpecialCH | 1.920833 |
| SpecialMM | 1.874833 |
| LoyalCH | 4.084527 |
| SalePriceMM | inf |
| SalePriceCH | inf |
| PriceDiff | inf |
| PctDiscMM | 572.549034 |
| PctDiscCH | 602.048674 |
| ListPriceDiff | inf |

Table 1: VIF values.

The problem of multicollinearity is due to the fact that some variable values are dependent on the other variable values for example, the SalePriceMM is dependent on PriceMM and DIscMM since SalePriceMM = PriceMM-DiscMM%.

Multicollinearity has negative impacts on the logistic regression. Here from the table 1, the dataset columns PriceCH, PriceMM, DiscCH, DiscMM, SalePriceMM, SalePriceCH, PriceDiff have VIF values of infinity which means those columns values are explained by the other columns in the dataset and columns like PctDiscMM and PctDiscCH which has high values of around 600 which is high, so we need to eliminate the variables that has high collinearity by doing penalized regression. The best penalized regression is LASSO.

**LASSO**

Lasso is a technique that makes some of the features in a model equal to zero, effectively removing them. It's used to simplify and select the most important features in a machine learning model, helping prevent overcomplicated models.

In LASSO we calculate the alpha value using hyperparameter tuning. We used five-fold cross validation to evaluate the best alpha from the values of [0.001,0.01,0.1,1 and 10] and performed grid search to find 0.01 as the best alpha.

When used the alpha as 0.01 and perform LASSO model on the dataset we find that some variables are not good predictors for the model and their coefficients are dropped to zero.

| Variable | Coefficients |
|---|---|
| PriceCH | 0.00 |
| PriceMM | 0.005838 |
| DiscCH | 0.00 |
| DiscMM | 0.00 |
| SpecialCH | 0.00 |
| SpecialMM | -0.013485 |
| LoyalCH | 0.314581 |
| SalePriceMM | 0.00 |
| SalePriceCH | 0.00 |
| PriceDiff | 0.077528 |
| PctDiscMM | 0.00 |
| PctDiscCH | 0.00 |
| ListPriceDiff | 0.007648 |

Table 2. LASSO Coefficients

Here from table 2, we can see that PriceCH, DiscCH, DiscMM, SpecialCH, SalePriceCH, SalePriceCH, PctDiscMM and PctDiscCH have coefficients as 0 so the LASSO is removing the coefficients with high multicollinearity and improves the model performance.

**Logistic model (LASSO predictors)**

Now let's use logistic model only using the LASSO predictors with the 'Purchase' as target variable.
The output of the LASSO model is in figure 2.

Logit Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | Purchase | No. Observations: | 606 |
| Model: | Logit | Df Residuals: | 600 |
| Method: | MLE | Df Model: | 5 |
| Date: | Tue, 21 Nov 2023 | Pseudo R-squ.: | 0.4324 |
| Time: | 01:56:38 | Log-Likelihood: | -238.36 |
| converged: | True | LL-Null: | -419.93 |
| Covariance Type: | nonrobust | LLR p-value: | 2.588e-76 |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | -4.0756 | 2.081 | -1.958 | 0.050 | -8.155 | 0.004 |
| PriceMM | 0.1176 | 1.111 | 0.106 | 0.916 | -2.059 | 2.295 |
| SpecialMM | -0.2156 | 0.324 | -0.665 | 0.506 | -0.851 | 0.420 |
| LoyalCH | 6.5943 | 0.518 | 12.742 | 0.000 | 5.580 | 7.609 |
| PriceDiff | 2.3583 | 0.564 | 4.181 | 0.000 | 1.253 | 3.464 |
| ListPriceDiff | 0.7589 | 1.606 | 0.472 | 0.637 | -2.389 | 3.907 |

Figure: 2. Logistic model summary with all LASSO variables.

**Model Metrics**

The AIC is 488.7135399181619
Accuracy: 0.84
RMSE: 0.41
ROC-AUC Score: 0.8371080139372823
Confusion Matrix:
[[67 15]
 [10 60]]
True Positives: 60
True Negatives: 67
False Negatives: 10
False Positives: 15
Sensitivity (Recall): 0.8571428571428571
Precision: 0.8
Specificity: 0.8170731707317073

Here we can see that while we removed the multicollinearity from the dataset, we see the improvement in the AIC of the model from 501 to 488 which shows the model performance is improved.

**Interpretation**

From the model summary output, we can see that the LoyalCH has a high positive influence on the purchase probability of the Citrus Hill (CH) and PriceDiff has a positive influence on the purchase probability of the Citrus Hill (CH).

PriceDiff which is the difference in the Sale price of MM less sale price of CH has 2.3583 log odds of Purchase of Citrus Hill for one unit increase in Price Diff. This coefficient effect is also significant as p-values is less than 0.05 (industry standard).

The LoyalCH which is Customer brand loyalty for CH has 6.5943 log odds of Purchase of Citrus Hill for one unit increase in the brand loyalty of Citrus Hill. And also, the effect is significant as the p-value is less than 0.05.

PriceMM and ListPriceDiff has positive influence of 0.1176 and 0.7889 log odds respectively on the purchase of Citrus Hill but the influence is not significant as p-value is greater than 0.05 whereas the SpecialMM has negative influence of 0.2156 log odds on the purchase of Citrus Hill but the influence is not significant as the p-value is greater than 0.05.

**Gradient Boosted Trees**

Gradient Boosting is a machine learning algorithm that combines the predictions of multiple decision trees in an iterative manner. It optimizes model performance by minimizing a loss function using gradient boosting, employing techniques like regularization, and handling missing data efficiently. Also known for its speed, accuracy, and effectiveness in various machine learning tasks, especially in structured data problems.

We used Gradient boosted trees method for the train data to predict purchase behaviors. We used hyperparameters tuning with five-fold cross validation on the following metrics:
‘n_estimators': [50, 100, 150, 170], The trees to be built for the model.
'learning_rate': [ 0.01,0.05,0.1, 0.5,1], The rate at which the model learn.
'max_depth': [1,2,3,4,5], Maximum depth of the tree.

We get the best parameters from the grid search.as ‘learning_rate': 0.1, 'max_depth': 1, 'n_estimators': 150. By using these values into the model, we define and build the boosted trees model.

**Model Metrics**

Accuracy on the test set: 0.84
RMSE: 0.41
ROC-AUC Score: 0.8339721254355401
Confusion Matrix:
[[70 12]
 [13 57]]
True Positives: 57
True Negatives: 70
False Negatives: 13
False Positives: 12
Sensitivity (Recall): 0.8142857142857143
Precision: 0.8260869565217391
Specificity: 0.8536585365853658

The gradient boost trees improve the precision and specificity of the model when compared to the logistic regression model. We can get the precision and specificity of the model using the confusion matrix.

**XAI (eXplainable Artificial Intelligence)**

Gradient Boost provides a plot_importance function that allows you to visualize the importance of each feature in your model. This can give you a high-level understanding of which features are contributing the most to the model's predictions.

We can see from the figure 3, that the LoyalCH has a high effect and feature importance on the purchase pattern of the customers.
PriceDiff also has a good feature importance on the purchase pattern of the customers.

This is consistent with the interpretation outputs of the logistic regression output as the LoyalCH and PriceDiff has the high and significant influence on the purchase.

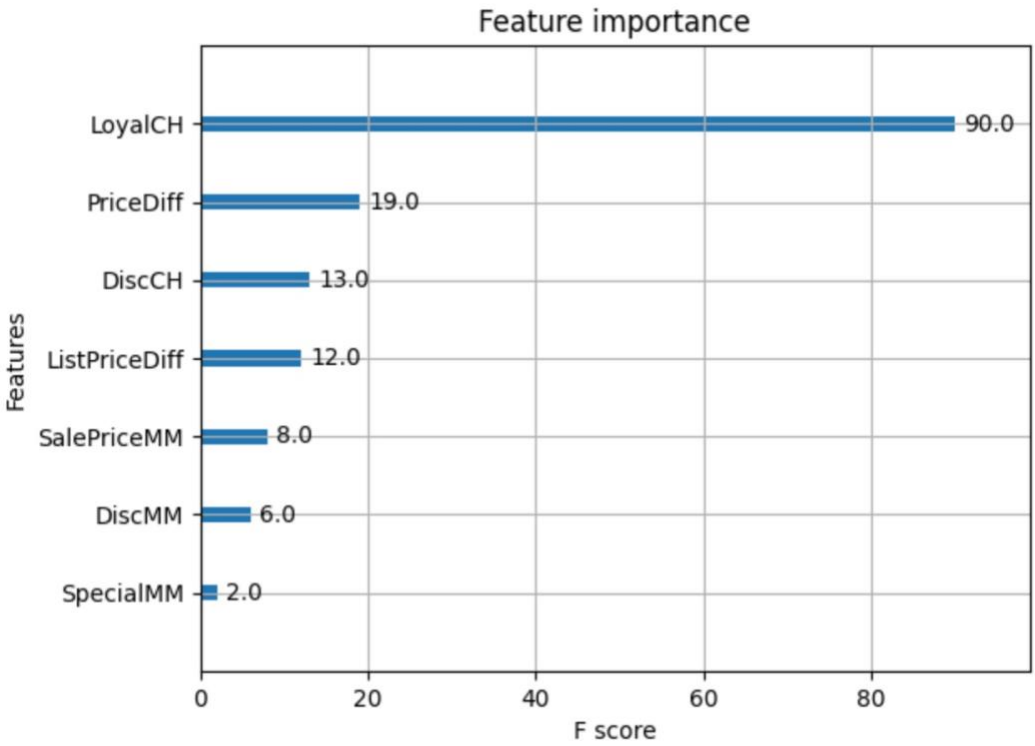


Figure 3: Feature importance in the Gradient Boosted Trees.

# Results

| Metric | Logistic Value | Boosted Trees Value |
|---|---|---|
| Accuracy | 0.84 | 0.84 |
| ROC-AUC | 0.8371 | 0.8339 |
| RMSE | 0.41 | 0.41 |
| True Positives | 60 | 57 |
| True Negatives | 67 | 70 |
| False Negatives | 10 | 13 |
| False Positives | 15 | 12 |
| Sensitivity (Recall) | 0.857 | 0.814 |
| Precision | 0.8 | 0.826 |
| Specificity: | 0.817 | 0.853 |

Table 3. Comparison of Metrics

From the table 3, we can see that the metrics of the both logistic and boosted trees model are on par with each other.

All predictor variables influence the purchase pattern of MM but only some of them are significant while others are not.

If we need to know the influence of the predictors on the outcome variable purchase then logistic model is a good model. The results of the logistic model provide around 0.84 accuracy and AIC of 488, so this is considered as good model.

In summary, the analysis of the factors influencing the purchase of Citrus Hill reveals the following:

PriceDiff: The difference in sale prices between MM and CH, with a log odds coefficient of 2.3583 for one unit increase, has a significant positive effect on the purchase of Citrus Hill. The significance is confirmed by a p-value less than 0.05.

LoyalCH (Customer Brand Loyalty for CH): The log odds coefficient of 6.5943 indicates a substantial positive impact on the purchase of Citrus Hill for every one-unit increase in customer brand loyalty for CH. This effect is also deemed significant, supported by a p-value less than 0.05.

PriceMM and ListPriceDiff: While both PriceMM and ListPriceDiff show positive influences of 0.1176 and 0.7889 log odds, respectively, on Citrus Hill purchase, these effects are not considered significant. This is because the associated p-values are greater than the standard threshold of 0.05.

SpecialMM: SpecialMM, with a negative influence of 0.2156 log odds, does not have a significant impact on the purchase of Citrus Hill. The p-value is greater than 0.05, indicating that the observed effect may be due to random variability.

In conclusion, PriceDiff and LoyalCH are identified as significant factors positively influencing the purchase of Citrus Hill, while PriceMM, ListPriceDiff, and SpecialMM do not show statistically significant effects in this analysis.

In summary, the analysis suggests that gradient boost trees outperform logistic regression in terms of precision and specificity. The evaluation is based on the confusion matrix, which provides a detailed assessment of the model's performance.

Furthermore, the eXplainable Artificial Intelligence (XAI) capabilities of the gradient boost model are leveraged through the plot_importance function. This function allows for the visualization of feature importance, providing insights into which features significantly contribute to the model's predictions.

Figure 3 highlights that LoyalCH and PriceDiff are particularly influential in shaping the purchase patterns of customers. The high effect and feature importance of LoyalCH are consistent with the logistic regression outputs, where LoyalCH was identified as having a significant positive impact on the purchase of Citrus Hill. Similarly, PriceDiff, which was also identified as a significant factor in logistic regression, is reaffirmed by its substantial feature importance in the gradient boost model.

# Conclusion

The outcome of the convergence of results across models strengthens the understanding that LoyalCH and PriceDiff play crucial roles in influencing customer purchase patterns, providing a consistent and interpretable narrative in both logistic regression and gradient boost analyses.

From the objective of the brand manager who is into the influence of the variables on the Purchase of MM, I would recommend using the logistic model as the model has more on to the interpretation methods of influences on the target outcome.

I would recommend the brand manager to reduce the price of the MM as the price difference has a significant impact on the sales of the MM and CH. If CH has low price, then it has positive influence on the purchase of CH.

I would recommend the Sales Manager to focus more on the building predictive model using the gradient boost trees as it is more of black box model which only provides probabilities of the outcome to purchase the MM or CH rather than the influence of the predictors.

# Appendix

**Link to code**

https://colab.research.google.com/drive/1jg6RVr77C02fae-yl9IUTcvx64ovCRAP?usp=sharing