Paula Stefani

Jade Gosar

Machine Learning Final Project

Fall 2023

Professor: Dr. Himanshu Mishra

MKTG 6620-001

Table of Contents

Problem Statement _____ **3**

Methods and Analytical Approach _____ **4-9**

Exploratory Data Analysis _____ **4-5**

Models _____ **6-10**

Results & Conclusion _____ **9-10**

References _____ **13**

Problem Statement

Orange Juice is one of the most consumed items of any American household. According to the USDA August 2003 Report, “on average, at-home orange juice consumption accounts for about 64.1 pounds of oranges per person per year (…), Americans consume 2½ times more orange juice annually than its nearest competitor, apple juice”. The economic impact of orange juice production is of extreme significance, only in 2000/01, “74 percent of Florida's processing crop went to making frozen concentrated orange juice” (Pollack Et al; 2001). When considering the year of 2022, “production of packaged canned orange juice was 354 million gallons (single-strength equivalent basis), generating a total value of $2.182 billion” (Cruz et al.,).

With this in mind, any grocery store that is underselling the targeted goal amounts of orange juice would be losing significant possible revenue. This project is focused on supporting a local grocery store chain by performing data analysis and providing insights on orange juice demand, while identifying how to make the Orange Juice category perform better than what it does currently for the specific store.

Among the two orange juice brands the store has (Citrus Hill and Minute Maid), we were told that Minute Maid has shown higher margins than Citrus Hill. With this in mind, our team has selected two main objectives for this analytical project: identifying what variables influence a customer’s probability of buying Minute Maid (MM) and creating a predictive model in order to predict the probability of a customer purchasing Minute Maid (MM). Our goal is to use the data related to Minute Maid sales to identify what impacts the purchasing of Minute Maid Orange Juice, and come up with business strategies in order to increase the sales and revenue in the orange juice category. We believe this will benefit the grocery store chain since both marketing and sales managers will have valuable insights that will be helpful when selecting and implementing possible business strategies.

Some of the important questions we are trying to answer with this project are: What predictor variables influence the purchase of MM? Are all the variables provided to us in the dataset effective or are some more effective than others? What are the specific possible recommendations to increase demand for MM? Is it possible to create a predictive model that can inform the probability of customers buying MM, and if so, how good is this model?

After conversation with our main stakeholders in this project, we are expected to provide answers for these questions while also implementing data analysis tools and processes to better understand the sales of Minute Maid and come up with valuable insights for the brand and sales managers of the grocery store.

Methods and Analytical Approach

Exploratory Data Analysis (EDA)

For our Exploratory Data Analysis, we started by doing some preprocessing. We changed our target variable Purchase into a factor type and converted the independent integer variables into numeric variables.

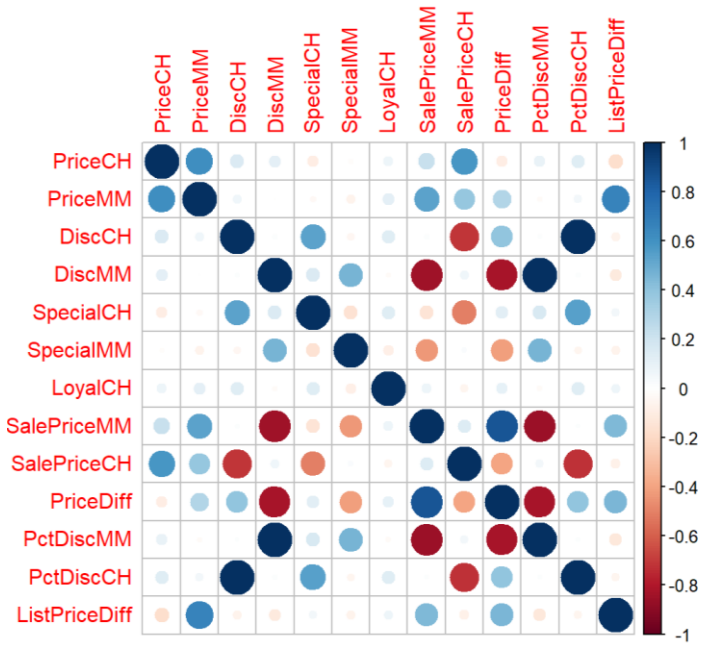
```
# Convert variables to appropriate types
OJ[2:14] <- lapply(OJ[2:14], as.numeric)
OJ$Purchase <- as.factor(OJ$Purchase)

# Check the class of the variables in the df
sapply(OJ,class)
```

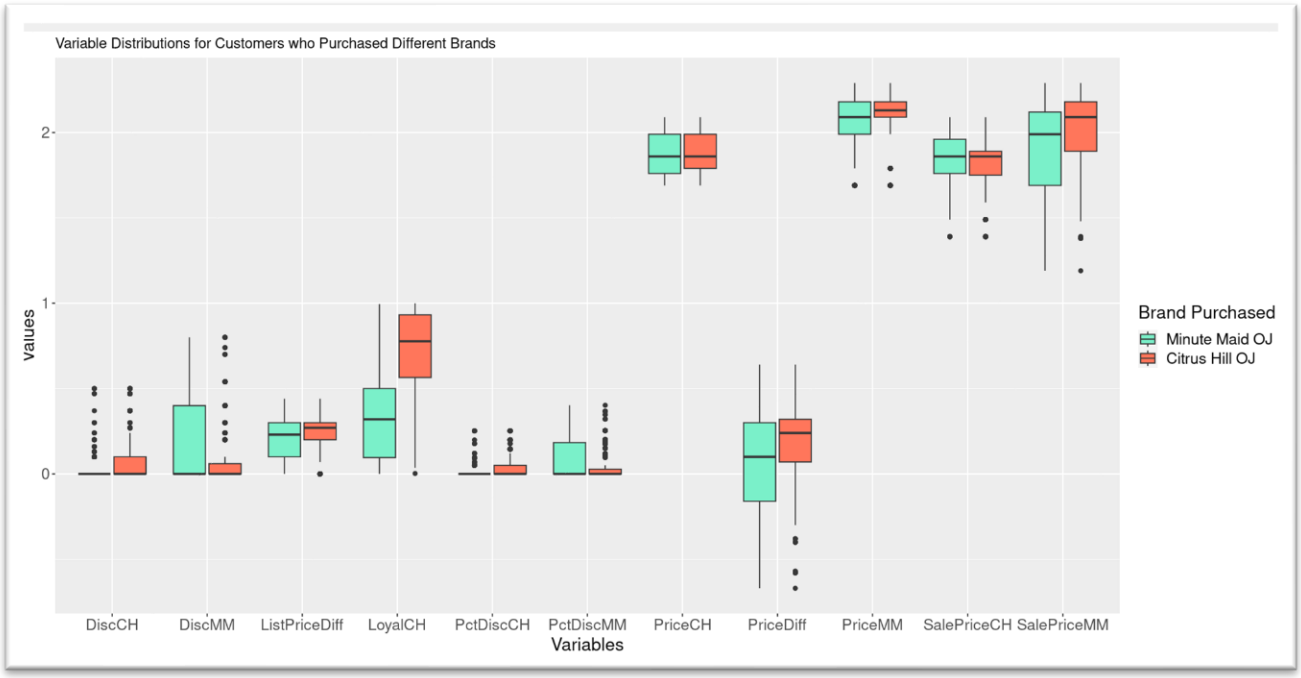
```
##      Purchase       PriceCH       PriceMM        DiscCH        DiscMM
##      "factor"     "numeric"     "numeric"     "numeric"     "numeric"
##     SpecialCH     SpecialMM       LoyalCH   SalePriceMM   SalePriceCH
##     "numeric"     "numeric"     "numeric"     "numeric"     "numeric"
##     PriceDiff     PctDiscMM     PctDiscCH ListPriceDiff
##     "numeric"     "numeric"     "numeric"     "numeric"
```

Next, we review some of the summary statistics (Min, 1st & 3rd Quadrant, Median, Mean and Max) for all the variables in our dataset OJ in order to identify possible differences in means between the groups of customers that purchased Citrus Hill vs Minute Maid. Some of the interesting points we observed with these summary statistics were that MM showed a higher mean value for price charged than CH (2.085 for MM and 1.867 for CH). On the other hand, MM also showed higher mean discounts offered in comparison to CH (0.1234 for MM and 0.0518 for CH).

After initial summary statistics, we performed correlation analysis and created a correlation plot to identify multicollinearity among variables. We were able to see that the variables that presented high positive correlations coefficients were DiscCH with PctDiscCH, and DiscMM with PctDiscMM. Some of the variables that presented high negative correlations were DiscMM with SalePriceMM, DiscMM with PriceDiff, SalePriceMM with PctDiscMM and PriceDiff with PctDiscMM.



Lastly, we created some EDA analysis and graphs to better understand the distributions and aspects of our variables. In the bloxplot "Variable Distributions for Customers who Purchased Different Brands" we were able to identify some interesting insights about the differences between the distributions of people who purchased Citrus Hill vs Minute Maid. The variables with highest differences in distributions were DiscMM, LoyalCH, and PriceDiff. Some of the variables with similar distributions were PriceCH and SalePriceCH.



Models

In order to perform the analysis for this project, our team decided to use three different model approaches that would complement each other and help us answer the questions from stakeholders while also helping us reach our analytical goal for the project: Logistic Regression, LASSO Penalized Regression and Gradient Boosted Decision Trees.

First, we split the data into train and test sets (70:30) and created a simple logistic regression model that included all variables in order to help us understand the predictors with highest significance in our target variable *Purchase* when considering all variables from the dataset. The reason why we choose logistic regression is because of how it is useful for predicting when the target dependent variable is dichotomous or binary. In this scenario, our target variable *Purchase* was a factor with levels 0 and 1 indicating whether the customer purchased Citrus Hill (1) or Minute Maid Orange Juice (0). We believe logistic regression would be helpful for us to identify the significance of the independent variables in *Purchase*. As it can be seen in the model summary below, the predictors with highest significance in the logistic regression model were PriceMM, LoyalCH, PriceCH, and SpecialMM.

```
Call:
glm(formula = Purchase ~ ., family = binomial, data = trainData_log)

Coefficients: (4 not defined because of singularities)
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    -4.1811     2.1575  -1.938   0.0526 .
PriceCH        -3.6105     1.4139  -2.554   0.0107 *
PriceMM         3.9120     0.9672   4.045 5.24e-05 ***
DiscCH          4.6788    21.3977   0.219   0.8269
DiscMM        -14.1193     9.8797  -1.429   0.1530
SpecialCH      -0.3519     0.3826  -0.920   0.3576
SpecialMM      -0.6302     0.3109  -2.027   0.0426 *
LoyalCH         6.6412     0.4530  14.659  < 2e-16 ***
SalePriceMM         NA         NA      NA       NA
SalePriceCH         NA         NA      NA       NA
PriceDiff           NA         NA      NA       NA
PctDiscMM      25.0597    20.7405   1.208   0.2269
PctDiscCH      -0.1023    40.6273  -0.003   0.9980
ListPriceDiff       NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1146.03  on 856  degrees of freedom
Residual deviance:  653.05  on 847  degrees of freedom
AIC: 673.05

Number of Fisher Scoring iterations: 5
```

After concluding initial logistic regression, we decided to create a LASSO penalized regression model. Our main idea behind this choice was because of how computationally effective LASSO is, and the fact of how LASSO chooses the best predictors variables while reducing the coefficients of the less important predictors. We believe LASSO would be an effective approach in order to identify which predictor variables have biggest influence in the purchase of MM and being able to provide us some comparison model in which we could compare how it would differ from the original logistic regression that included all predictor variables. This would help us answer the stakeholder's question "Are all the variables provided to us in the dataset effective or are some more effective than others?" since it reduces the coefficients of the least important predictor variables. The LASSO model would also be efficient in reducing the multicollinearity among variables identified during EDA process.

Some important points to note are that for our original logistic regression model we did not scaled the variables. However, for the LASSO penalized regression model we scaled the variables in order to compare accuracy and performance results. The idea behind this was to see possible differences in effects from scale vs non-scale variables. We also used cross validation in order to determine the best lambda for our LASSO model, and split the data into train and test set.

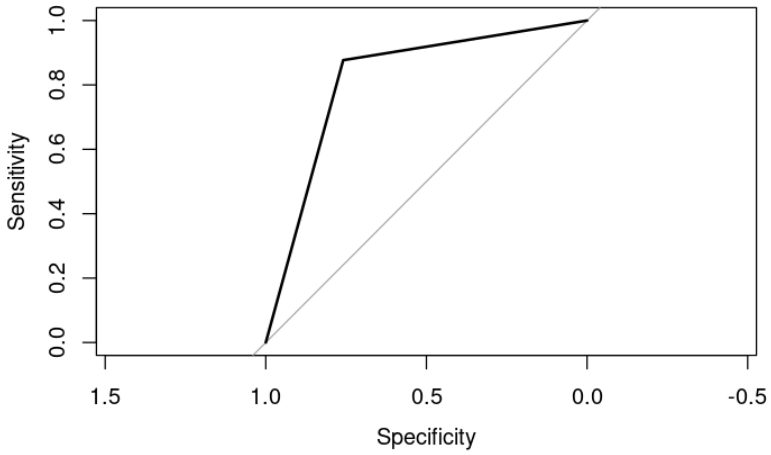
```
14 x 1 sparse Matrix of class "dgCMatrix"
                      s1
(Intercept)    0.8415003
PriceCH        .
PriceMM        .
DiscCH         .
DiscMM         .
SpecialCH      0.0461963
SpecialMM     -0.3637862
LoyalCH        1.9398372
SalePriceMM    .
SalePriceCH   -0.0672893
PriceDiff      0.9593683
PctDiscMM      0.3082719
PctDiscCH      0.0049346
ListPriceDiff  .
```

According to the LASSO model results beside, the predictors with highest influence in the target variable *Purchase* were SpecialCH, SpecialMM, LoyalCH, SalePriceCH, PriceDiff, PctDiscMM, and PctDiscCH.

When compared to the results from the logistic regression model, the variables that were identified as significant in both models were, LoyalCH and SpecialMM.

One interesting analysis we were able to identify when comparing the initial logistic regression model and the LASSO penalized regression model was the difference in accuracy. In the logistic regression that included all independent variables as predictors, accuracy was 0.8216. Once performed the cross validation and creation of LASSO model, accuracy

increased to 0.83099. Even though this is a small increase in accuracy, it shows that the penalized regression model performed better than the original logistic regression model. This is also supported by the ROC curve plot below for the LASSO model, which shows better performance (curve is closer to the top-left corner).

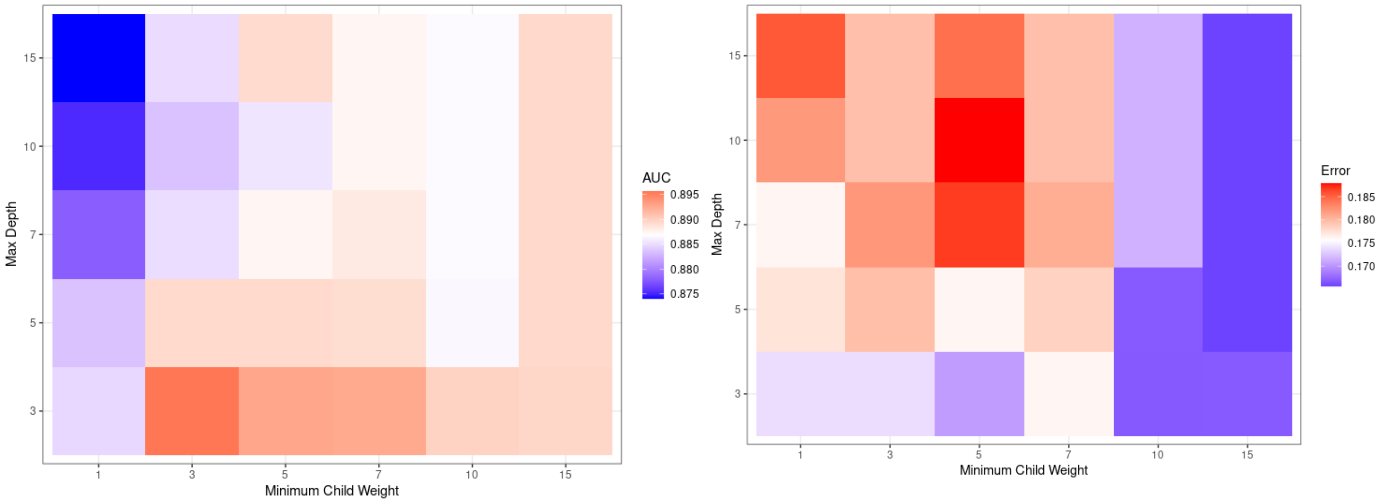


With the analysis performed, we were able to confidently say that we got better results when we did not include some of the variables in the modeling phase. This is due to the fact that some variables in the dataset presented high collinearity. LASSO models reduce multicollinearity by regularization by reducing the coefficients of the features that are multicollinear. This helps to explain why we got better performance metrics for our LASSO model, while also being able to identify which variables were more significant in predicting *Purchase*.

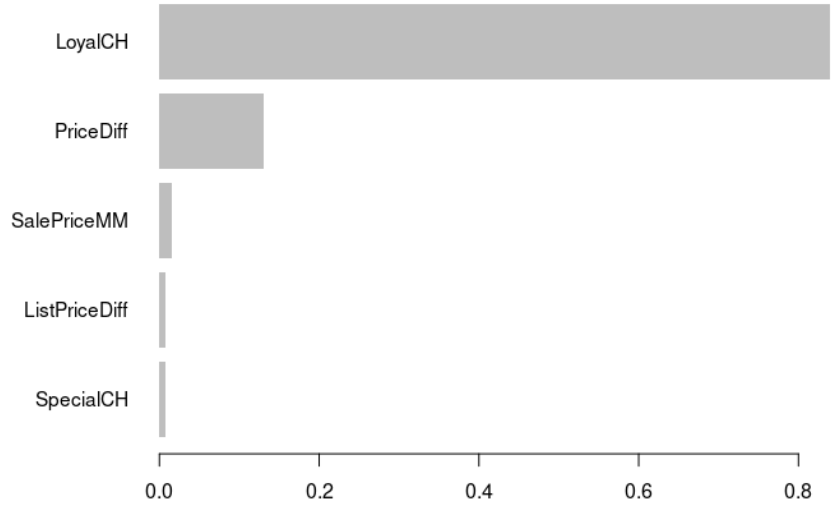
Finally, our team created a Gradient Boosted Decision Tree model in order to help us predict the probability of a customer purchasing MM. We decided to use a Boosted Decision Tree model since it is able to handle both categorical and numerical features, as well as non-linear relationships between features and the target variable. Besides, we decided to include the Boosted Decision Tree model besides the logistic regression and LASSO regression models since decision trees are able to handle missing values and outliers in the data much better than logistic regressions. We believe the addition of a Boosted Decision Tree model was helpful in supporting our analysis and helping us answer the questions from Sales management.

Some of the efforts used during modeling phase to reduce overfitting were the split into test and train datasets for all the models created, the use of 5-fold cross validation in both LASSO penalized regression model and the Gradient Boosted Decision Trees model. The Gradient Boosted Decision Tree model performed better on the train set than the other models, with an accuracy of .903 after tuning hyperparameters. We tuned various hyperparameters to their optimal combinations, including the optimal number of iterations, max depth values, minimum number of samples in node to split, and gamma parameter. We determined the optimal combinations of max depth values and minimum

child weight by providing a vector of values for each and running an xgboost model, using cross validation, to determine the combination of parameters that provided the highest AUC with the lowest error. All the combinations that were tested are shown in the visualizations below which we used to compare the performance of the models when it came to error and AUC. The optimal combination of these values was determined to be 3 for both max_depth and minimum_child_weight, which is confirmed by a very high AUC value with a low error metric shown in the graphs below. We followed a similar tuning process to find the optimal gamma value for the final mode, which was .10 as it returned an accuracy score of 0.89724 with the lowest error metric score.



Our final XGBoost model utilized these optimal hyperparameters to create predictions on the same test set that was used to measure the performance of the Logistic and LASSO regression models. When using performance metrics such as accuracy and ROC-AUC, all of these models had very similar performances on the holdout set. The final XGBoost model had an accuracy of 0.836 and an AUC of 0.817. This means that the XGBoost model performed the best according to accuracy but had a slightly lower value for AUC, meaning these models had very comparable performance when predicting purchase behavior in the test set. This model also showed similar significant variables, as is shown in the importance matrix below that represents the level of influence of the top 5 most significant variables in the model.



Overall, we believe that the creation of multiple methods was essential in supporting our analysis for this project since the LASSO penalized regression and logistic regression models were helpful in answering the questions made by the Brand manager regarding which predictor variables ad biggest influence on purchasing of Minute Maid while the Gradient Boosted Decision Tree was helpful in answering the questions from the Sales manager related to predicting probability of customers buying Minute Maid. We believe the use of different models complemented each other for the creation of better analysis.

Results & Conclusion

Our analysis suggests that there are multiple variables that are significant and can be used when predicting customer purchase behavior as it relates to buying either Citrus Hill or Minute Maid Orange Juice. Due to the fact that the brand and sales managers were interested in answering different questions, we focused on providing insights from the Logistic Regression and LASSO models to the Brand Manager as the model outputs represent predictor variables that are influential in the purchase of Minute Maid OJ as well as ones that are more significant than others. As for the Sales manager, we created and interpreted the Gradient Boosted Tree Model because it has better predictive power when it comes to accuracy and can be used to inform the manager of the probability of customers buying Minute Maid over Citrus Hill.

- **Recommendations for the Brand Manager**

After performing the creation of the Logistic Regression and LASSO models we were able to identify important insights that can be helpful for the Brand Manager to create better marketing strategies and help increase probability of buying Minute Maid.

According to our LASSO model, the predictor variables with highest influence in the purchase of Minute Maid were SpecialCH, SpecialMM, LoyalCH, SalePriceCH, PriceDiff, PctDiscMM, and PctDiscCH. However, among these variables, the ones that both the logistic regression and LASSO models assigned as relevant predictors were LoyalCH (probability of a customer buying CH based on previous purchasing behavior) and SpecialMM (Specials like free gift, loyalty points, etc in MM). This suggests that the Brand department should focus on strategies that increase customer loyalty by mainly supporting specials like loyalty points when buying the specific brand, free gifts, coupons, rewards programs, loyalty spreads, etc. We believe this can be helpful in increasing the probability of a customer to buy Minute Maid while also increasing the loyalty of the customer with the brand.

Some of the most effective aspects that the Brand Manager should focus to support Minute Maid sales are the percentage of discounts for MM, the difference in sales price between MM and CH in the grocery store and the loyalty of customers that are loyal to CH and MM. Trying to shift the loyalty from CH to MM using specific sales and marketing strategies while also reducing the difference in price between MM and CH would be helpful in increasing sales of Minute Maid.

- **Recommendations for the Sales Manager**

The predictive model that was created to determine the probability of customers buying Minute Maid was a Gradient Boosted Tree Model that had an accuracy of .836, slightly outperforming the other models in this metric. This is a fairly high accuracy score for a predictive model and, due to the fact that it could be further tuned with additional hyperparameters if requested, allows for the most flexibility in implementing a predictive model going forward. The variables that have the most influence over predicting probability of customers buying Minute Maid are whether the customer is loyal to Citrus Hill, the price difference between the two brands, the sale price of Minute Maid, the difference in the list price between the two, and whether Citrus Hill was on a special or not. Due to the fact that whether a customer is loyal to Citrus Hill or not has a significant influence on their likelihood to buy Minute Maid, we recommend that the Sales Manager target customers that are not already loyal to Citrus Hill to build up the loyalty for the Minute Maid brand. It could also be worthwhile to explore what price difference between the two products causes the customer to change their preferences from Citrus Hill to Minute Maid, as the price difference is also an important variable in the prediction model.

Overall, all of these models could be used to inform the Sales Manager of what variables are most influential in a customer's purchase of Minute Maid over Citrus Hill, but the Gradient Boosted Tree Model has the best accuracy score when it comes to predicting the probability of customers buying Minute Maid. For this reason, the final tuned XGBoost model should be used by the Sales Manager while the other regression models suit the needs of the Brand Manager better. All in all, the model for the Sales Manager represents the data in the train and test set the best when it comes to accuracy of predictions and therefore we are

confident that it will meet the needs of predicting the probability of customers buying Minute Maid.

References:

Cruz, J., Ferreira, J., & Court, C. D. (2023, February 15). *2020-2021 Economic Contributions of the Florida Citrus Industry*. Food and Resource Economics Department - University of Florida, Institute of Food and Agricultural Sciences - UF/IFAS. https://fred.ifas.ufl.edu/media/fredifasufledu/economic-impact-analysis/reports/FRE_Economic_Contributions_Florida_Citrus_Industry_Report_2020_21_WEB-(2).pdf

Pollack, S. L., Lin, B. H., & Allshouse, J. (2003, August). *Characteristics of U.S. Orange Consumption*. USDA Economic Research Service. https://www.ers.usda.gov/webdocs/outlooks /37012/50262_ fts30501.pdf?v=1030

Willig, G. (2023, January 16). *Decision tree vs logistic regression*. Medium. https://gustavwillig.medium.com/decision-tree-vs-logistic-regression-1a40c58307d0#

Yıldırım, S. (2020, February 17). *Gradient boosted decision trees-explained*. Medium. https://towardsdatascience.com/gradient-boosted-decision-trees-explained-9259bd8205af