Machine Learning: Factors Influencing Orange Juice Purchases

Adam Bruce

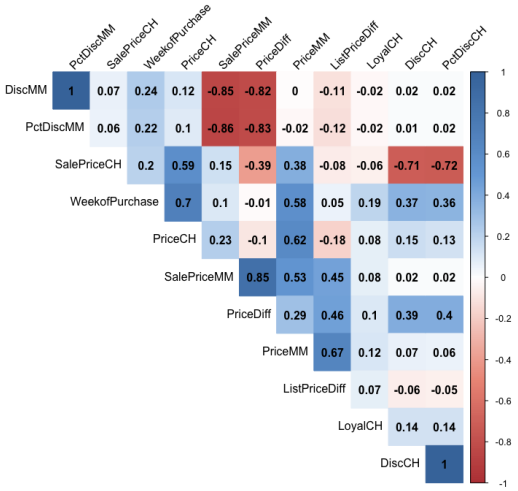
2024-02-21

Introduction

In the grocery industry, understanding customer preference for purchasing an item is of critical importance. The more we understand preferences, the more we can optimize sales through methods like targeted marketing and closeouts. Here, we will use logistic regression to look at both how accurately we can predict and how well we can interpret purchases between two Orange Juice brands: Citrus Hill (hereafter CH) or Minute Maid (hereafter MM). Ultimately, this assessment aims to provide valuable insights into reliably classifying customer purchases and may be adapted to other products of interest for our store.

Exploratory Data Analysis

Variable Selection



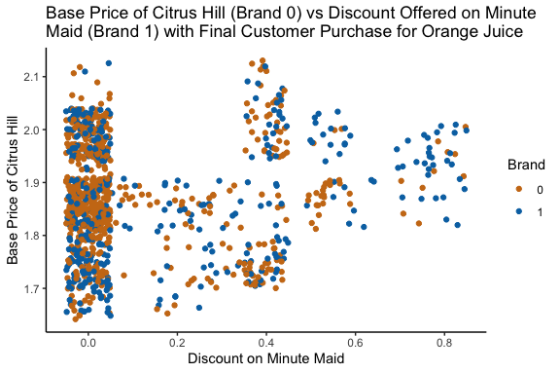
The initial dataset contained 1070 observations for 18 variables and did not contain any missing values. The variables included: Purchase, WeekofPurchase, StoreID, PriceCH, PriceMM, DiscCH, DiscMM, SpecialCH, SpecialMM, LoyalCH, SalePriceMM, SalePriceCH, PriceDiff, Store7, PctDiscMM, PctDiscCH, ListPriceDiff, and STORE. Five were categorical: Purchase, StoreID, SpecialCH, SpecialMM, and STORE. The variable of interest, Purchase, was changed from “CH” and “MM” to 0 and 1 respectively for easier use with logistic regression. Of the 1070 purchases, 653 were CH and 417 were MM.

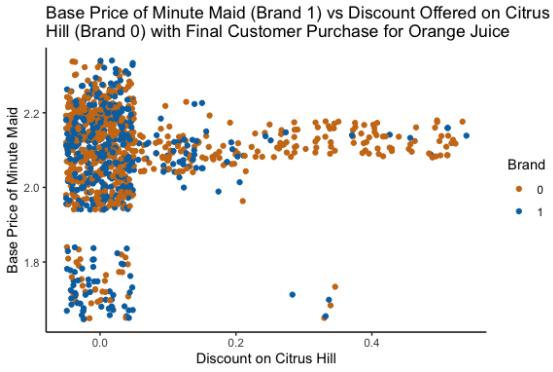
To eliminate redundancy by identifying relationships between variables a correlation matrix was made (above). Blue colors with positive numbers represent positive relationships while red colors with negative numbers represent negative relationships.

Essentially, the darker the color, the stronger the relationship between two variables. When a relationship either >= 0.7 or <= -0.7 exists, the variables are redundant. Therefore, one of the variables should be removed. For our variables this included: PriceCH : WeekofPurchase = 0.704, PriceDiff : SalePriceMM = 0.853, DiscMM : SalePriceMM = -0.847, PctDiscMM : SalePriceMM = -0.857, PctDiscCH : SalePriceCH = -0.723, DiscCH : SalePriceCH = -0.711, DiscMM : PriceDiff = -0.824, PctDiscMM : PriceDiff = -0.828, PctDiscCH : DiscCH = 0.999, and PctDiscMM : DiscMM = 0.999. Additionally, ListPriceDiff and PriceMM had a correlation of 067, which was deemed near enough to the threshold.

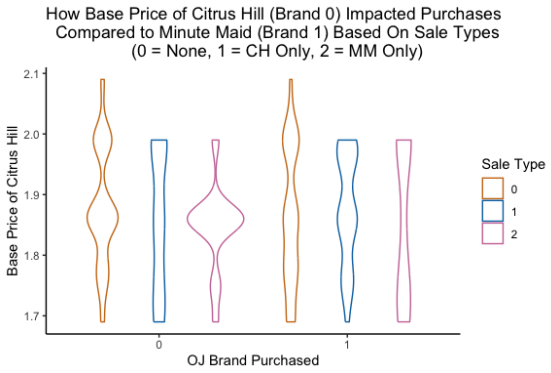
As a result, the variables PctDiscCH/PctDiscMM, SalePriceCH/SalePriceMM, and PriceDiff were removed because we know have discounted price amounts DiscCH/DiscMM. Additionally, WeekofPurchase was removed because PriceCH and ListPriceDiff because of PriceMM. Next, the categorical variables Store7 and STORE were removed because StoreID was indicative of the information they provided. Meanwhile, SpecialCH and SpecialMM were combined into SaleType, which indicated whether neither brand was on special (0), the special was ONLY on CH (1) or the special was ONLY on MM (2). The case where a discount was offered on both brands (3) was discarded because it occurred only 4 times in stores where only MM was on sale! Therefore, this indicated a data entry error and these rows were discarded.

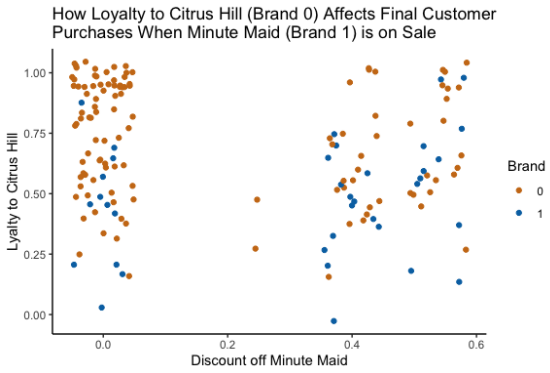
Relationship Analysis





Brand	MaxPrice	MinPrice	MeanPrice	MeanDiscount
Citrus Hill	2.09	1.69	1.87	0.05
Minute Maid	2.29	1.69	2.09	0.12





To investigate our variables impact on the brand of orange juice purchased, and ultimately how they may influence our predictions/interpretations, exploratory tables and plots were made. First, scatterplots of the base listing price of one brand vs discount price of the other brand were made (Top Left & Right). In both cases, when there was no sale on the other brand (x-axis = 0), the opposite brand's base price had no effect on customer purchases. This indicates that when there are no sales, customers likely purchase the brand they prefer. However, when there was a discount on MM (Top Left), customers purchased that brand more often regardless of CH base price, if the discount was greater than around 0.75 cents. Meanwhile, when there was a discount on CH (Top Right), customers purchased that brand more often regardless of MM base prices, if the discount was greater than around 0.20 cents.

This suggests that CH is either viewed as a higher quality product or is sold at a much lower price that any sale extremely impacts its purchase. However, a summary table (middle) provides support to the later theory. The mean sale price for CH is 22 cents cheaper than MM and CH has a maximum base value of only 2.09 compared to 2.29 for MM. Therefore, DiscCH should have a significant impact on our predictions/interpretations.

Next, a violin plot (Bottom Left) shows the base price of CH versus which brand was purchased based on the sale type for that observation. We see that when the base price of CH is high and there are no sales for either brand (Sale Type = 0), customers purchase CH and MM about the same. When he base price is moderately low (below 1.9) and there are no sales, customers perhaps slightly favor MM. However, if the base price of CH is below 1.90 and there is a sale on CH (Sale Type = 1), customers tend to favor buying MM. While this might be unexpected, we know that the base price of CH is on average 22 cents below MM, so when a store runs a sale on CH with an already moderately low base price, customers perhaps view the product as inferior to the more "luxurious" MM. The fact that CH sales are higher than MM when CH base prices are above 1.90 supports this idea. Lastly, when the base price of CH is above 1.80 and there is a sale on MM, buyers slightly favor CH. This might suggest that a sale on MM brings the cost close to the base price of CH when it is above 1.80, where customers purchase CH because it is viewed as a better deal. Overall, we would expect our predictions/interpretations to mirror these findings moving forward.

Finally, a scatterplot of the loyalty to CH versus the discount sale price of MM shows that if the customer loyalty to CH is above 0.75 (75%), they will likely purchase CH regardless of how much MM is discounted. However, a value below that threshold would indicate the customer will likely purchase MM if the product is discounted at all. When no discount exists for MM, a loyalty of only about 0.50 (50%) or greater typically indicates the customer will purchase CH. Thus, high loyalty ratings should lead to predictions/interpretations where CH is preferred, especially in the case where there is not a sale on MM.

Purchasing

Here, we produce a logistic regression model with the goal of accurately predicting whether a customer purchased MM orange juice based on the non-excluded variables outlined in the previous section.

Table 1: Model Results for Predicting Training Orange Juice Data

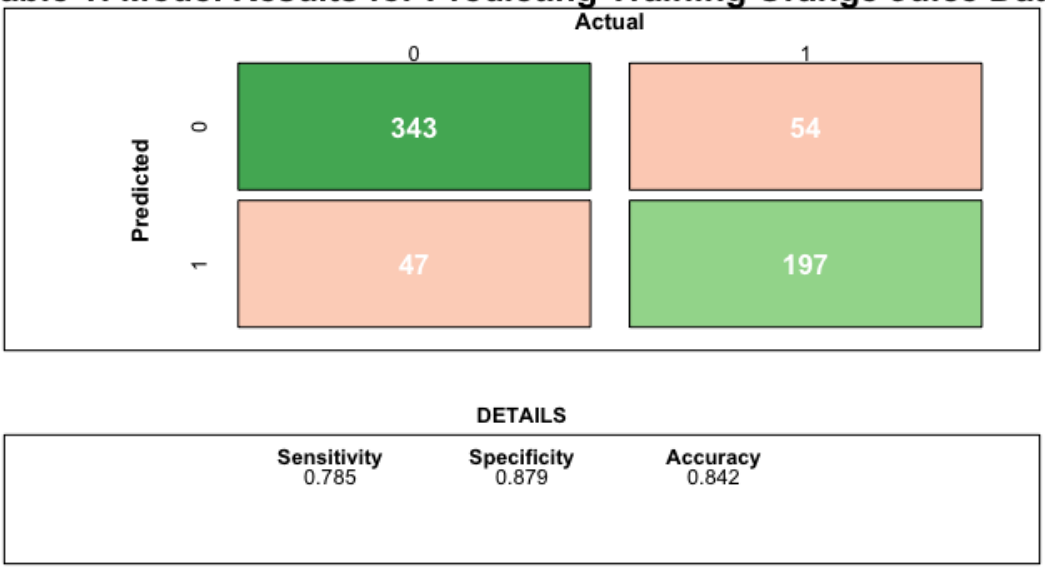
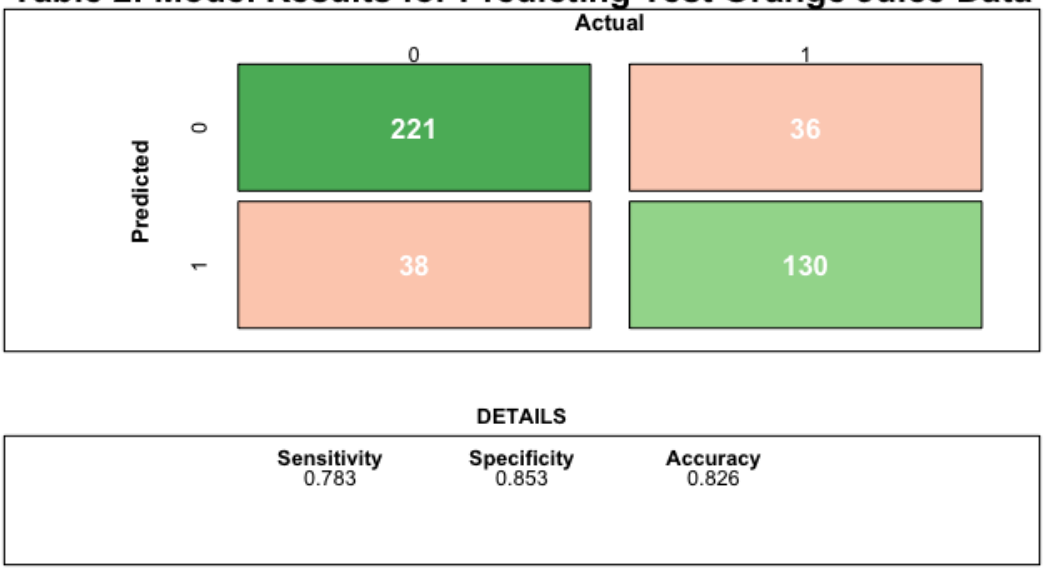


Table 2: Model Results for Predicting Test Orange Juice Data



Logistic regression models work to produce a probability (or percentage if multiplied by 100) of success. Here, a success means a customer purchased MM. In order to do so, they require a binary outcome response (success or failure) where a success is represented as a 1 and failure by a 0. The formula for these models follows what is known as a sigmoid curve, which if plotted on a graph simply looks like an S-Shaped curve between 0 and 1 (y-axis) with probabilities on the x-axis. The formula for creating such a curve is $P = e^{a+bX\ldots bN}/(1 + e^{a+b1\ldots bN})$, where P is the probability of success (P = 1), e is Euler's Constant, a is the slope of the model, and the b1 - bN represents each of the N predictor variables. Using this formula, we can find a probability of MM purchase for each of the observations in the OJ data, and then decide if the probability is high enough to be classified

as an MM or low enough to be a CH purchase. This was done using an optimal cutoff for this analysis. Based on these classifications, we can compare against the actual known response values to get out models accuracy, which we use to determine its viability for future use.

However, if we used all of our 1066 observations to build this model, then we would only be able to assess how accurate the model is on these particular datapoints, not any unseen data. Therefore, a validation approach, which involves taking all our data, splitting it into two groups, building our model based on one of the groups, and then assessing the model on the other group is used to avoid this limitation.

The group used to build the model is known as the training set and the other group is called the test set. Though the amount of data assigned to each is subjective, typically more data is used in the training group. Ultimately, we assigned 65% of our data to the training group (641 observations) and 35% (425 observations) to the test group. Next, a data transformation, known as centering/scaling, was performed to get every value in the training and test sets on the same measurement range before building our model.

Next, a probability of success was calculated for each row of the training data based on the coefficients and an optimal cutoff point was produced. This decides the minimum probability value at or above which a positive is classified. Our model value was 0.479 or 47.9%, so any probability lower than this value was predicted as failure (Customer Purchased CH) and any probability at or higher was predicted as success (Customer Purchased MM).

These decisions were then put into a table, known as a confusion matrix, to break down how accurate we were. Table 1 (Above: Top) shows these results and is read as follows:

The top two columns, 0 and 1, are the actual values in the training dataset and the side rows, 0 and 1, are the predicted values in the training dataset.

Summing down from the top we see that 390 observations were purchases of CH (343 + 47), and 251 observations were purchases of MM (54 + 197) in the training set. Likewise by summing the rows, we see that our model predicted 397 observations were CH purchases (343 + 54) and 244 were MM purchases (197 + 47). Based on this, we can find how well we predicted MM purchases by calculating accuracy. This value is simply **the number of predictions we got right divided by the total number of observations in our data**

Following the confusion matrix, we find the total number of correct predictions by lining up the rows and columns. For example, column 0 lines up with the row 0 in Table 1 at the number 343 shown in green. We also see that column 1 and row 1 line up at 197 in green. Adding these two together we get 540, which is the total correct model predictions! By adding up the values in red, we find our total incorrect predictions at 101 (54 + 47). To find accuracy, we then take the correct predictions and divide by the sum of all values to get **540/641 = 0.8424 or 84.24%**. Ultimately, this tells us that for the training data, our model accurately predicts whether or not a customer purchases MM orange juice 84.24 percent of the time.

How do we know if this accuracy value is good? Besides using judgement to say that 84% seems decent, we can actually compare this to a value known as the No Information Rate. Looking at the training data we saw that 390 purchases were CH while only 251 were MM. Therefore, it is more likely that any given purchase was CH. Using this logic, we could randomly predict all of the 641 observations to be CH and 390 of 641 of our guesses would be correct. Therefore, we would have an Accuracy of **390/641 = 0.6084 or 60.84%**, which is the No Information Rate. Based on this, we see that our model predicts purchases accurately 23.4% better than randomly guessing, suggesting it does quite well overall.

Although, every model is built using its own specific observations and thus we need to assess how they perform on new, unseen, data to truly measure performance. This is where a test set comes in handy. Using the same model coefficients from the training data, we now take the 425 test observations and predict the probability of success (MM purchase) again for them. After doing so, we use the same optimal cutoff as before to decide whether each observation was a CH or MM purchase.

We then follow the same procedures as before to produce a confusion matrix and reassess model accuracy. However, there is one caveat to this. Our model was fit with the training data, so we would expect that accuracy on test data would be less than with the training set. Although, if our accuracy is much lower, say 10% or more, we would be at risk of fitting too specifically to the training data. Put simply, our model would fail to account for the fact that every individual occurrence in the real-world is unique. Each shopping trip and customer is different than the last, and thus purchases will change based on this individuality. When models are too specific, they neglect this fact, and thus their accuracy on data not used to build them is very poor.

In our case, the test set confusion matrix can be seen in Table 2 (Above: Bottom). Reading like before, we see that the model predicted 351 of 425 observations correctly for the test set. Ultimately, this resulted in a **Test Accuracy = 0.8269 or 82.69%**, which was less than the **Training Accuracy = 0.8424 or 84.24%** as expected, but not so much less that we are concerned with overfitting. Now, we make the ultimate assessment of how well our model performed by comparing the Test Accuracy to the **Test No Information Rate = 0.6094 or 60.94%**. Again, the model performs quite well, as it predicts the test purchases accurately 21.75% more often than randomly guessing.

We can see in both Table 1 and 2 our accuracy values along with two additional estimates in Sensitivity and Specificity. On some occasions, we might be concerned with how well we can predict a “Yes” case or perhaps how well we can predict a “No” case. For this, we rely on Sensitivity and Specificity. Sensitivity is when we are interested in predicting the “Yes” category. A model is highly sensitive when it predicts the TRUE “Yes” category well, thus avoiding falsely predicting a “No” when the true value was “Yes”. The opposite is true with specificity. A model is highly specific when it predicts the TRUE “No” values well, thus avoiding falsely predicting a “Yes” when the true value was “No”. Typically, there is a tradeoff in sensitivity and specificity. Overall, our model was **more specific at 0.8533 or 85.33 (221/(221 + 38))** than it was sensitive **0.7831 or 78.31% (130/(130+36))** for the test data.

Overall, this model does an adequate job at predicting orange juice purchases for customers in our dataset, and it shows how powerful predictive modeling can be in a real-world setting. Clearly, an approach like this could prove beneficial for our stores by producing purchasing predictions we could use to target specific customers in our advertising.

Marketing

Here, we attempt to produce a Logistic Regression Model that best explains preferences among customers who purchased MM. Therefore, model selection criterions AIC and McFadden's Pseudo R^2 were used to justify decisions among 5 plausible models. Finally, the "best" model was explained by utilizing its coefficients and p-values to explain the relationships, if any, between our predictors and the success response category Purchase = MM.

Plausible Logistic Regression Models Considered for the OJ Dataset

	Model	Structure
m1	Model 1	Purchase ~ LoyalCH
m2	Model 2	Purchase ~ PriceMM + DiscMM
m3	Model 3	Purchase ~ DiscMM + DiscCH + LoyalCH
m4	Model 4	Purchase ~ DiscMM + DiscCH + LoyalCH + SaleType
m5	Model 5	Purchase ~ .

Above, the five Logistic Regression models considered are shown with the format Response ~ Predictors. The first model utilized only customer loyalty to CH as a predictor because LoyalCH showed a very strong relationship to purchases in exploratory analysis. The second model used only variables measuring MM: base price MM and discount MM as it is plausible CH provides no informaton on MM purchases. The third model uses the three numeric predictors that showed a clear relationship with MM during data eploration: DiscMM, DiscCH, and LoyalCH without the categorical predictor SaleType. The fourth model used all variables analyzed in exploratory analysis with the catefoical predictor: DiscMM, DiscCH, LoyalCH, SaleType. Finally, every predictor could be important in explaining an MM purchase, so model five was built with: DiscMM, DiscCH, LoyalCH, SaleType, PriceMM, PriceCH, and StoreID.

Comparison Statistics For Five Plausible OJ Models

	Model	AIC	McFaddens_R^2
m1	Model 1	921.19	0.36
m2	Model 2	1376.98	0.04
m3	Model 3	858.93	0.40
m4	Model 4	859.50	0.41
m5	Model 5	842.95	0.43

Above, Akaike's Information Criterion (AIC) and McFadden's Pseudo R-Squared values used to compare fit to the OJ dataset are shown. In AIC, we observed that models 3, 4 and 5 had a clear separation from models 1 and 2. Their AIC values of 858.93, 859.50 and 842.95 respectively were much lower than the first two model, and thus their prediction error is estimated to be best. In addition, the McFadden Pseuod R^2 estimates for models 3, 4, and 5 were also much higher than the first two models at 0.40, 0.41 and 0.43 respectively. Therefore, models 1 and 2 were dropped from consideration at this point. However, it should be noted model 2 using only variables measuring MM had an extremely high AIC of 1376.98 and low McFadden's R^2 of 0.04, which indicates knowing information on CH is important for predicting/explaining purchases of MM.

When deciding between models 3, 4, and 5, the expectations of explanatory modeling had to be considered. While Model 5 technically had the best overall performance in AIC and R^2 measures, it was considerably more complex than models 3 and 4. As a result, the standard errors could be inflated, indicating the simpler models 3 and 4 would be a better choice. In addition, the improvement in AIC and R^2 were both minor for this model, so it was dropped from consideration. Between the final two models, we observed a lower AIC in model 3 at 858.93 with a lower R^2 of 0.40. Although, the two values were extremely similar to model 4 at 859.50 and 0.41 respectively. Ultimately, model 3 without the categorical predictor, SaleType, was chosen because it decreased the complexity of explanations significantly while providing essentially the same or better information than model 4.

Key Model Outputs for the Best OJ Explanatory Logistic Regression Model

Predictor	Estimate	Standard_Error	P_Value	Significant	Lower_95_CI	Upper_95_CI
Intercept	14.9798	0.2111	2e-16	Yes	10.03	22.958
DiscMM	1.3028	0.4098	1.07e-10	Yes	1.2038	1.4138
DiscCH	0.6774	0.8492	4.5e-06	Yes	0.5694	0.7955
LoyalCH	0.5294	0.3849	2e-16	Yes	0.4895	0.5694

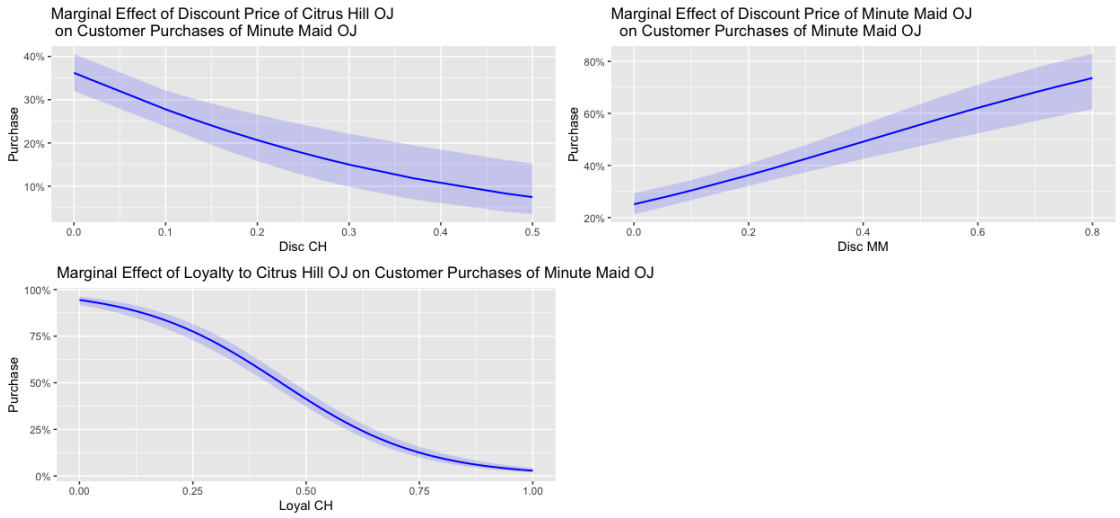
The summary table above breaks down the significant outputs from our best model, and includes a 95% confidence interval lower/upper bound for the coefficient estimates. Though included in the table for reference, the intercept will not be interpreted. Overall, we observe that all three explanatory variables: DiscMM, DiscCH, LoyalCH have extremely significant P-Value estimates at 1.07e-10, 4.50e-06, and 2e-16 respectively. We also see that the standard error estimates for our three predictors are very small relative to the coefficient estimates, which explains why our P-Values are so significant when derived from a Z-Distribution using **Z = estimate/standard error**.

Here, an interpretation of each coefficient estimate and 95% Confidence Interval associated with that estimate will be provided in the context of the predictor and response variables. Note that these variables represent either a cents increase/decrease or a proportion, so we must convert from a "One Unit" change to a more interpretable change like 0.1 ($0.10 cents or 10%) by raising each exponentiated coefficient estimate to the 0.1 power.

DiscMM: For each 0.1 unit increase in discount of Minute Maid OJ, the odds of a customer purchasing MM are multiplied by a factor of 1.302832 given all other predictors are held constant. We are 95% confident that the odds of a customer purchasing MM are multiplied by a factor of between 1.2038 and 1.4138 for each 0.1 unit increase in discount of Minute Maid given all other predictors are held constant.

DiscCH: For each 0.1 unit increase in discount of Citrus Hill OJ, the odds of a customer purchasing MM are multiplied by a factor of 0.6774023 given all other predictors are held constant. We are 95% confident that the odds of a customer purchasing MM are multiplied by a factor of between 0.5694 and 0.7955 for each 0.1 unit increase in discount of Citrus Hill given all other predictors are held constant.

LoyalCH: For each 0.1 unit increase in the proportion of customer loyalty to Citrus Hill OJ, the odds of a customer purchasing MM are multiplied by a factor of 0.5294 given all other predictors are held constant. We are 95% confident that the odds of a customer purchasing MM are multiplied by a factor of between 0.4895 and 0.5694 for each 0.1 unit increase in the proportion of customer loyalty to CH given all other predictors are held constant.



Above, marginal effect plots for each explanatory variables are shown. These plots allow us to visualize the predicted probabilities (shown as percents) of a customer purchasing MM based on the range of values of each predictor from our model. The plots functionally hold all other predictors constant at their mean. Overall, we see that any discount on CH (Top Left) essentially yields an extremely small chance of a customer purchasing MM, as values after 0% have probabilities less than 0.45. For LoyalCH (Bottom Left), it appears that a 0.50 threshold is an indicator of customers preferring CH to MM. However, any loyalty to CH does lower the chances of an MM purchase. Finally, there is a steady increase in the probability of a customer purchasing MM when it is discounted (Top Right). Around a discount of 0.40, the purchase becomes much more likely to be MM, as the predicted success probability passes 50%.

Ultimately, based on this exploration, we can reasonably say discounts on Minute Maid and Citrus Hill along with customer loyalty to Citrus Hill are key indicators of MM customer

purchasing habits. In an attempt to increase the proportion of customers purchasing MM in our store, we could consider the following marketing approaches: Increase the discount price of MM to a minimum of 0.40 cents, refrain from offering discounts on CH brands altogether, and attempt to steer customer loyalty away from CH and towards MM by perhaps offering rewards or benefits to those who purchase MM.

Summary and Recommendations

In this analysis, we sought to find valuable insights into reliably classifying and interpreting customer purchases of a convenience store product so that we could use a similar approach for products of interest in our Grab-n-Go stores.To find these insights, we used logistic regression modeling for prediction and interpretation based on data regarding the Orange Juice brands Citrus Hill and Minute Maid.

For prediction, we produced a model with 82.69% accuracy when predicting an MM or CH purchase. This model performed much better than randomly guessing each new data point to be the majority class, which yielded an accuracy of 60.94%. Therefore, we confirmed that the model we built was reasonable for prediction. In general, the process of creating such a model involved estimating the probability of a purchase based on all the variables remaining after exploratory data analysis. Going forward, if we wish to predict different product purchases accurately based on customer data, then we should first limit the data to only important variables with such an analysis before fitting the model with the remaining variables. Following this, we suggest using a training/test split to find overall accuracy for justifying a model's use in prediction. If the model performs poorly, then we recommend trying different variable combinations with train/test splits until a justifiable model is reached. Then, our company could take information on new customers, plug them into the model to receive a purchase prediction, and finally build targeted advertisements based on the prediction outcomes.

For interpretation, we observed how we can use model comparison methods to decide which variables are most important to a model. By using these techniques, we were able to narrow in on a "best" model, which we used for interpreting how our variables related to an orange juice purchase of MM. For our Grab-n-Go stores, we recommend using a similar approach when using new customer data to identify purchases for a product of interest. By doing so, we can identify exactly how we expect our customers to act based on the most important variables we are measuring. As a result, we could then further build our targeted advertisements based on these interpretations.

Overall, we showed just how powerful modeling for prediction and interpretation can be in the grocery store industry. For Grab-n-Go, our results should help to maximize profits and increase advertising efficiency in the future.