Project 2

Adam Gruber

2022-11-09

Introduction

Our company, Grab and Go convenience wants to build a better model for customers purchasing Minute Maid orange juice. Accurately predicting the customers purchases can benefit our company in several ways. Orange juice is a dated product and accurately modeling the sales can prevent excess product from expiring. Accurately modeling purchasing for all our stores can allow us to increase volume purchases, buy in bulk and receive volume discounts from vendors. Accurately modeling the Minute Maid purchases by store can also allow us to partition whole sale purchases from vendors and accurately send product to each store without the need to move inventory from store to store.

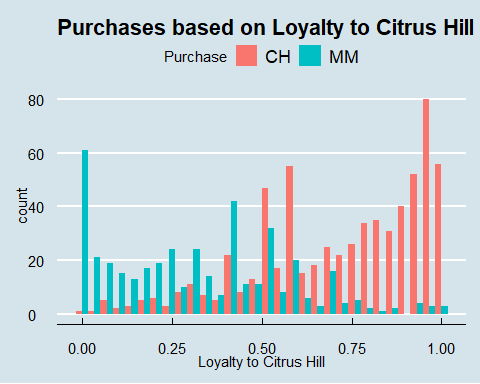
Missed sales of Minute Maid can have several long term effects. It can reduce customer loyalty to Minute Maid brand or Grab and Go stores as customers try other products. Missed sales can interrupt the usual flow of purchases, and make future ongoing modeling more difficult. Some weeks we have many sales, and other we have few to none, making our models unable to accurately predict.

In order to build our model, we first explored the data looking for any patterns, and the key variables. Once they were identified, we built 5 models: A,B,C,D,E. We built the models using logistic regression. We partitioned the data and tested the models, looking to see which ones were most accurate for predicting Minute Maid purchases, Citrus Hill, and then overall. We then used several tests to find which model fit the best. Model A was chosen. We then looked at the odds ratios for each variable to determine which ones are the strongest factors with the most influence. This can help the marketing team determine who are the customers for Minute Maid, and how to target them properly.

Part 1 Exploratory Data Analysis

Our OJ data set should prove very valuable to our model for predicting customer purchases. The data set is a complete data set with no missing values.

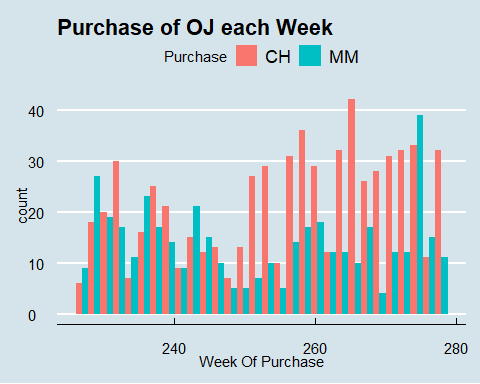
There are several variables that will be used in our model to predict. Some are used to measure numerical data like Price, and others are categorical. They are based on selecting certain categories such as “Yes”, “NO”, or “Store ID”. Each of these will factor into the model in different ways. All of the variables related to price are numerical except “SpecialCH” and “SpecialMM”. Both categorical and yes/no variables. Several variables skew to the left with long tails. One of these is loyalty to Citrus Hill. This is seen in many other ways throughout the data causes a skew to the left, such as Price Difference vs Purchases.



There are several redundant variables that are unneeded for our study. They are redundant because they are captured in other variables already. Too many variables can lead to the model over fitting and needing a larger sample size to make more accurate predictions. Columns I removed were “Store 7”, “Store”, “SPecialCH”, and “SpecialMM” The “Store”, and “Store 7” variables are already captured in the “Store Id”. The “SpecialCH” and “SpecialMM” are captured by the “DiscCH” and “DiscMM”.

Our Customers have a strong preference for Citrus Hill brand vs Minuted Maid. This is measured with brand loyalty. This can also be seen in several different ways. 38.97% of all purchases in the sample were for Minute Maid.

| Purchase | Quantity |
| --- | --- |
| CH | 653 |
| MM | 417 |



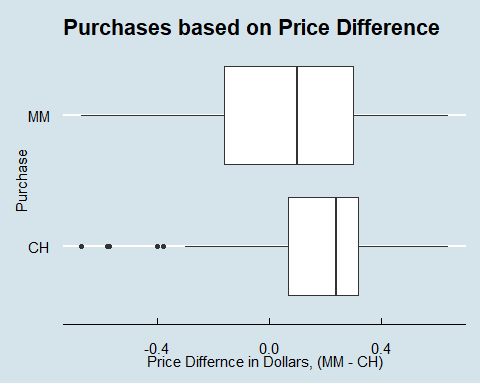
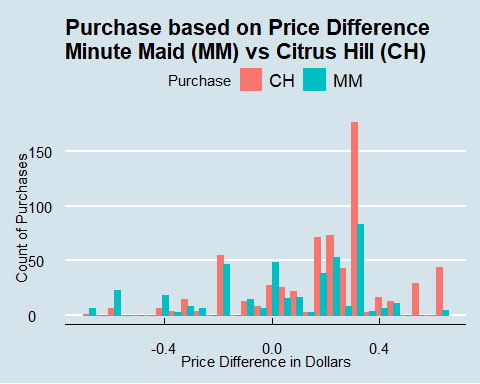
When comparing Price Averages, Minute Maid has a higher average price. It also had a average higher discount. This helped compensate for a larger initial price. The Sale Price after discount was still higher for Minute Maid on Average compared to Citrus Hill. This makes sense with the average Average Price Difference (Minute Maid - Citrus Hill). There are times where Minute Maid has a lower Sale Price compared to Citrus Hill.

Price Averages

| Price.CH | Price.MM | Sale.Price.MM | Sale.Price.CH | Price.Difference |
| --- | --- | --- | --- | --- |
| 1.867 | 2.085 | 1.962 | 1.816 | 0.1465 |

There are several outliers in Discount for Minute Maid Discounts. This translates to outliers in Sale Price of Minute and then Outliers in Price Difference. This should not be a concern in our model. Minute Maid was on discount for 16.2% of the sample observations. Citrus Hill was on discount for 14.8% of observations.

Our box plot shows customers choose Minute Maid when the average price difference is lower, meaning Minute Maid costs the same or less. This is because our customers either prefer or Citrus Hill as their drink of choice or prefer it lower price on average. The Outliers on the Box Plot come from customers that are still purchasing Citrus Hill despite it being more expensive than Minute Maid.

Part 2 Purchasing Model

In order to build our model and test it we partitioned our data into two different sets. The 60% of the data was used as training data to build the model. 40% was used as testing data to see how accurate the model was. In order to build the most accurate model, 3 different models were created using different variables. This was done to find out how much each variable affects the accuracy and which ones may cause over fitting. Too many variables can cause the model to mispredict because it looks at too many different factors when trying to predict the purchases.

Model A used Week, store id, Price of Citrus Hill and Minute Maid, Discount of Citrus Hill and Minute Maid, Loyalty to Citrus Hill, Sale Price for both, and Price Difference.

Model B used all those variables less store id.

Model C use all variables from model B less week.

Some data sets that have numeric variables that have wide range such as 10,000 for one variable and 5 for another may require he data to be standardized or normalized. This data set was had a very small range of numeric values from -1 to 2.50. Our model has a high accuracy as is, without needing to be transformed. In order to minimize errors, when building the model an opitmal cutoff was generated to reduce missclassification.

All 3 models had the same base level of Prevalence since that is base levels odds of picking Minute Maid. In our total data set, Minute Maid was 38.97% of purchases. In our models, prevalence was 39.04%, which was based on the number of Minute maid purchases in the training data.

Sensitivity refers to how likely our model is to predict the customer purchases Minuted Maid. Specificity refers to how likely the model is predicts the purchase of Citrus Hill which is the expected and most common value. Overall when looking at our Training data it appears Model A is most accurate. Often times, having too many variables will cause over fitting but here using the variables adds to the accuracy.

Minute Maid Training Data

| Training.Data | Accuracy | Sensitivity | Specificity |
| --- | --- | --- | --- |
| Model A | 0.8647 | 0.8247 | 0.8903 |
| Model B | 0.8538 | 0.7809 | 0.9005 |
| Model C | 0.8523 | 0.7888 | 0.8929 |

IN order to find the accuracy of our model, we need to test it using the test data that was partitioned. When testing we looked at all 3 models based on their changing variables. All models had a decreased level of accuracy as seen and that is expected when testing our data set. There is only a slight drop in accuracy. Accuracy when looking at our training Model A compared to testing for Model A shows only a slight drop. Accuracy dropped from .8647 to .8244. This is a sign our model is not over fitting. This refers to over calculating errors.

When building a model, the goal is always to minimize error, without over classifying or under classifying. This is why several variables were removed from the initial data sample. Those extraneous variables would cause issues with main variables, such ones that had a one to one relationship with over variables such as store and store id. Some variables were unneeded when building the overall model for determining purchases of Minute Maid at all our stores, such as Store7 variable.

Sensitivity, predicting for Minute Maid purchases, dropped to .7651 from .8247 Specificity, predicting Citrus Hill purchases, dropped to .8621 from .8903 in Model A. This means we can predict for 76.51% accuracy the purchase of Minute Maid compared to the no information rate of 61.12%. The no information rate refers to random chance.

Minute Maid Testing Data

| Testing.Data | Accuracy | Sensitivity | Specificity |
| --- | --- | --- | --- |
| Model A | 0.8173 | 0.7892 | 0.8352 |
| Model B | 0.8197 | 0.8012 | 0.8314 |
| Model C | 0.8150 | 0.7470 | 0.7582 |
| Model D | 0.6557 | 0.4337 | 0.7969 |
| Model E | 0.8197 | 0.8012 | 0.8314 |

Part 3 Marketing

In order to find the best fit and best prediction for our model, we created 5 different models based off the key variables. When summarizing model A, I found that several variables returned NA values for intercept. This is a sign, those variables are linearly dependent on other variables. These variables that are dependent in model A are Sale Price MM, Sale Price CH, and Price Diff. These values are captured in the price of other variables, Price of Ch, and Price of MM. I removed those variables from model C,D,E. These are the predictors for each model.

Model A has: WeekofPurchase, StoreID, PriceCH, PriceMM, DiscCH, DiscMM, LoyalCH, SalePriceMM, SalePriceCH, PriceDiff

Model B has: StoreID, PriceCH, PriceMM, DiscCH, DiscMM, LoyalCH, SalePriceMM, SalePriceCH, PriceDiff

Model C has: PriceCH, PriceMM, PriceDiff

Model D has: just PriceDiff

Model E has: StoreID, PriceCH, PriceMM, DiscCH, DiscMM, LoyalCH

Model E was constructed looking at the best predictors that were most significant according to their P value, and looking at metrics for the best fit using McFadden and AIC test. The extraneous variables were also removed for Model E. McFadden test looks for the largest value. That is model A and B. AIC test looks for smallest value which is Model B and A. Both are very similar.

Model Fit Metrics

| Predictive.Models | McFAdden | AIC |
| --- | --- | --- |
| Model A | 0.451960 | 493.4278 |
| Model B | 0.451850 | 491.5268 |
| Model C | 0.062084 | 814.8099 |
| Model D | 0.056115 | 815.9444 |
| Model E | 0.451850 | 491.5268 |

Model A is very similar in AIC, and McFadden score to Model B and E. The accuracy, sensitivity, and specificity are still better as seen in our Minute Maid Testing Data table.

The Summary of Model A shows all the Predictors and their coefficients. It shows the linear variables of SalePrice for MM, CH and Price Diff. Those are explained by other variables and show NA instead. Several of the predictors have negative values like LoyaltyCH. This means they are negative predictor if a customer will purchase Minute Maid. This customers that are more loyal to CH are less likely to purchases Minute Maid.

A P value less than .05 are considered significant. This means they are strong indicators for whether a customer will purchase Minute Maid or not. The strongest positive factor for predicting purchase of Minute Maid is the Price of Citrus Hill. This can be seen in the following graphs.

Model A Coefficients

## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) 2.769500050 1.77692368 1.5585926 1.190928e-01  
## WeekofPurchase -0.007217003 0.01122558 -0.6429069 5.202845e-01  
## StoreID2 0.119706472 0.27717709 0.4318772 6.658307e-01  
## StoreID3 -0.142506479 0.38100646 -0.3740264 7.083847e-01  
## StoreID4 -0.457335859 0.41061038 -1.1137952 2.653671e-01  
## StoreID7 -0.658067777 0.28005355 -2.3497926 1.878387e-02  
## PriceCH 4.655561226 1.84454629 2.5239601 1.160411e-02  
## PriceMM -3.262464532 0.87600412 -3.7242571 1.958912e-04  
## DiscCH -3.410941743 1.01502880 -3.3604384 7.781889e-04  
## DiscMM 2.571623196 0.44862672 5.7322114 9.912962e-09  
## LoyalCH -6.185117541 0.40585069 -15.2398841 1.922109e-52

The explanatory variables can have the odds for each variable calculated towards purchasing Minute Maid. The graph shows log odds each variable for purchasing Minute Maid. Log transformation was chosen for easy comparison of variables.

The price of Citrus Hill is the strongest indicator followed by Discount for Minute Maid. This makes sense because as price of Citrus Hill increases, purchases will decrease. Since we assume binary choices, our best predictor for customers purchasing Minute Maid is the price of the competition, Citrus Hill, and if we are running a Discount on Minute Maid. The graph also shows the confidence interval for each explanatory variable. This shows that we are 95% confident the odds ratio falls between 1.06 and 8.29 for PriceCH. The odds of purchasing Minute Maid increase by a factor of 105.17 for each unit increase in Price of Citrus Hill. This also applies the other variables in our model. The odds of purchasing Minute Maid decrease by a factor of .05 for each unit increase in the Discount of Citrus Hill.

Chart, scatter chart

Description automatically generated

Chart, scatter chart

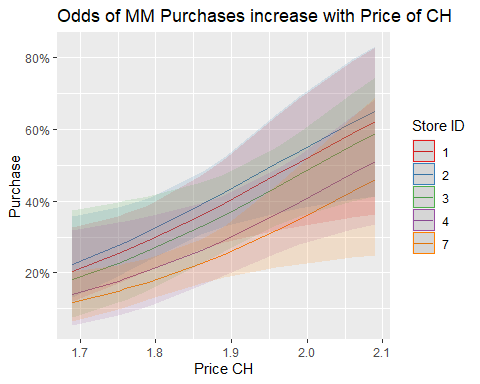
Description automatically generated

Model A Confidence Intervals for Explanatory Variables

## 2.5 % 97.5 %  
## (Intercept) -0.70669701 6.26670987  
## WeekofPurchase -0.02928973 0.01477032  
## StoreID2 -0.42391941 0.66393901  
## StoreID3 -0.89234598 0.60308176  
## StoreID4 -1.26929128 0.34274617  
## StoreID7 -1.21046561 -0.11120593  
## PriceCH 1.05617314 8.29554853  
## PriceMM -4.99998478 -1.56137127  
## DiscCH -5.47346408 -1.47675982  
## DiscMM 1.70493686 3.46583002  
## LoyalCH -7.00868846 -5.41567750

Our Linear variables have are shown as having no confidence interval and odds ratio because they are already captured in other variables.

The odds of purchasing Minute Maid can be graphed using the strongest predictor, the price of Citrus Hill. All stores see an increase in the odds of purchasing minute maid as the price of citrus hill increases. The store 3 has the highest odds and store 7 the lowest odds. This can also be seen in the Odds of Minute Maid Purchase graph. This shows the store 7 as having a negative odds ratio for purchase of Minute Maid.



In order for Grab, and Go convenience to increase the purchase of Minute Maid OJ, we need to increase the price of Citrus Hill OJ, or increase the discounts on Minute Maid OJ to make the prices more comparable for the consumer. The Price of Minute Maid was typically higher in our sample data, but as the price of Citrus Hill increased one unit, the odds of purchasing Minute Maid increased by a factor of 105.17. The odds of purchasing Minute Maid also increased by a factor of 13.09 for each unit increase in Minute Maid discount. Reducing/balancing prices in stores 1,2,3,4 will see biggest increase in sales of Minute Maid.

Brand Loyalty to Citrus Hill is the largest explanatory variable for decreasing the odds of purchasing Minute Maid. This factor may not be controllable or easily measured. It is possible though our model may need to be re-evaluated if their is negative publicity. For Example in 1991 Citrus Hill Brand had 12,000 gallons of OJ seized by the FDA for false advertising. This kind of negative publicity on tv can erode consumer confidence and cause our model to be inaccurate.

Summary and Recommendations

Our model A was chosen as it was considered the most even model when considering all the tests. Model A had an overall accuracy of 81.73% for all purchases. It predicted Minute Maid purchases accurately at a 78.92% rate. Model A had the largest McFadden test value and second smallest AIC value. Model A had the most significant variables of any model, looking at p value of less than 0.05. It also had the most variables of any model.

Our marketing team should look at the Price of Citrus Hill as the biggest factor determining sales of Minute Maid. The odds of purchasing Minute Maid increase by a factor of 105.17 for each unit increase in Price of Citrus Hill. Our next biggest factor is Discount of Minute Maid. Typically price of Minute Maid was higher than Citrus Hill. As prices became comparable, or even lower for Minute Maid, the odds of purchase increased. Our set by our marketing department are the biggest factors for predicting future sales.

Our model will need to be continuously re-evaluated as the factors might change over time. Loyalty was the biggest factor to reduce the odds of purchase. Brand Loyalty can change over time. Poor in-stock, poor quality products, and changing customer tastes can all affect brand loyalty over time. Prices of fresh products can fluctuate often with seasons and production levels. This may need re-evaluation every couple months. Over time we can build accurate models for various seasons as well, such as summer, spring, flu season, or orange harvest season. They are all external, seasonal factors that can affect the purchases of Minute Maid OJ but can be added into our model to make more accurate predictions.