­­

Project 1

An in-depth analysis of Orange Juice sales

Dan Lemaire - Gavin Moss - Joni Jensen

Submission Date: June 1st, 2017

Mktg6620 – Machine Learning for Business Applications

Overview 3

Problem Definition 3

Executive Summary 3

Drivers 3

Predictive Model 4

Methods 5

Data Prep 5

Data Cleaning 5

Data Classes 5

Variable Selection 5

Feature Engineering 6

Model Selection 6

Cross-validated Training Scores 6

Confirmation with Bootstrapped Testing Scores 6

Variable importance 6

Logistic regression 6

The problem of collinearity 6

Boruta analysis 6

Recommendations 7

Loyalty 7

Data 7

Pricing 7

Conclusion 7

# Overview

The goal of this analysis is to understand how to increase the revenue from the orange juice category of store. The store sells two brands of orange juice, Minute Maid and Citrus Hill. Since Minute Maid has higher margins than Citrus Hill, this analysis will make recommendations regarding which drivers influence a consumer’s decision to purchase Minute Maid orange juice. This allows our company to leverage those drivers as opportunities to influence Minute Maid sales. It will additionally provide a predictive model for more precise forecasting. This forecasting will be of benefit now, but will be of tremendous benefit later when the company adjusts its marketing to increase Minute Maid sales (since an updated forecast will be required).

## Problem Definition

The stores sell two brands of orange juice, Citrus Hill and Minute Maid. As Minute Maid sales have a lager profit margin we want to find what factors cause customers to purchases Citrus Hill orange juice instead of Minute Maid. We are also interested in building a model that is able to predict with confidence what juice a customer will purchase.

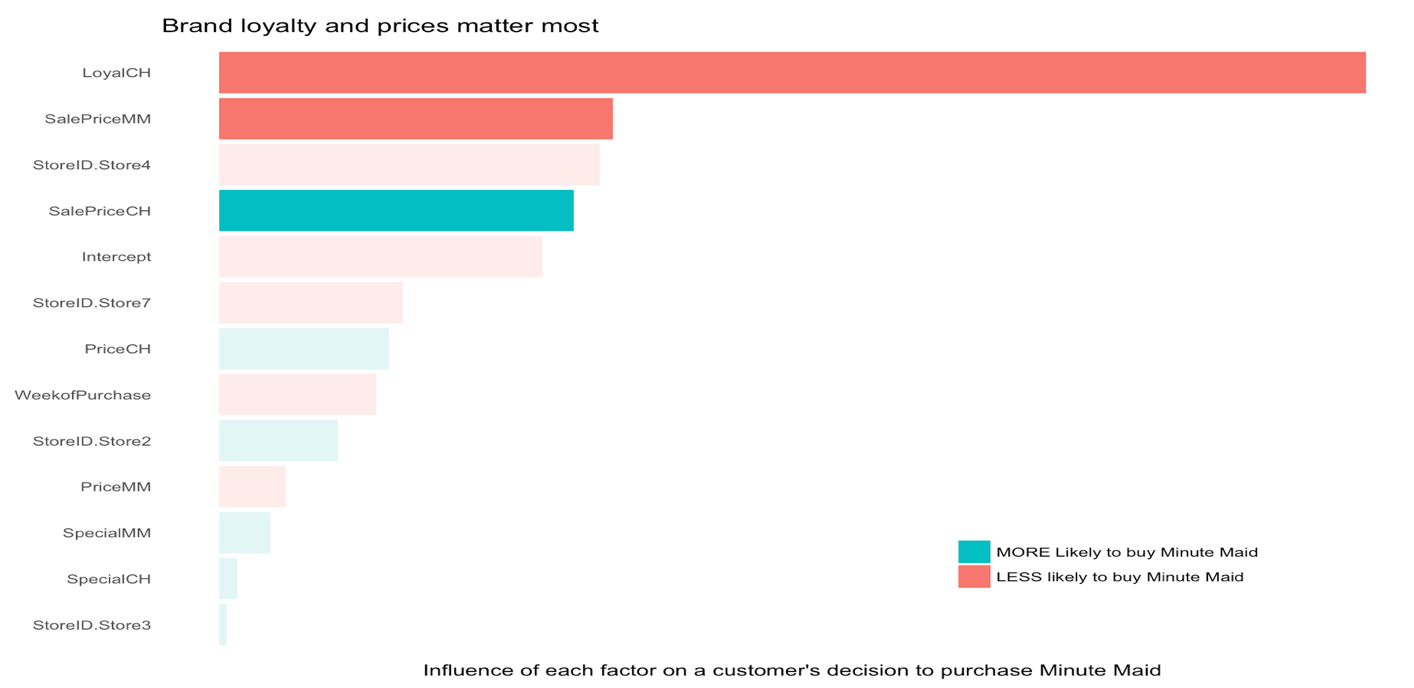
## Executive Summary

By using Logistic Regression, we have arrived at three recommendations given the goals and assumptions of this analysis:

* Erode customer loyalty to Citrus Hill to improve Minute Maid sales
* Test the effect of increased prices on Citrus Hill
* Collect more information to test for seasonality

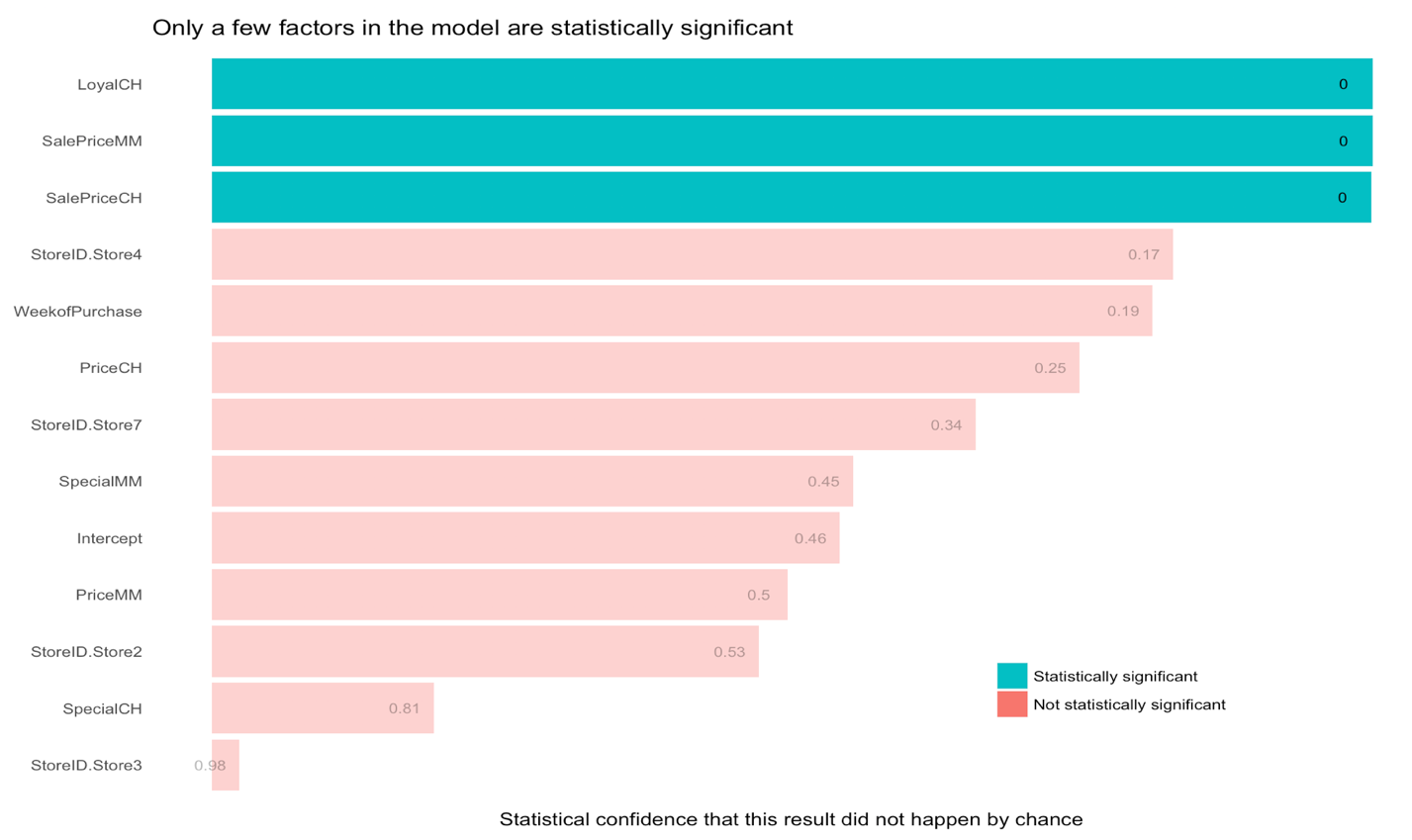
# Drivers

The following chart shows the magnitude of the effect different drivers have on a customer’s propensity to buy Minute Maid and Citrus Hill orange juice.



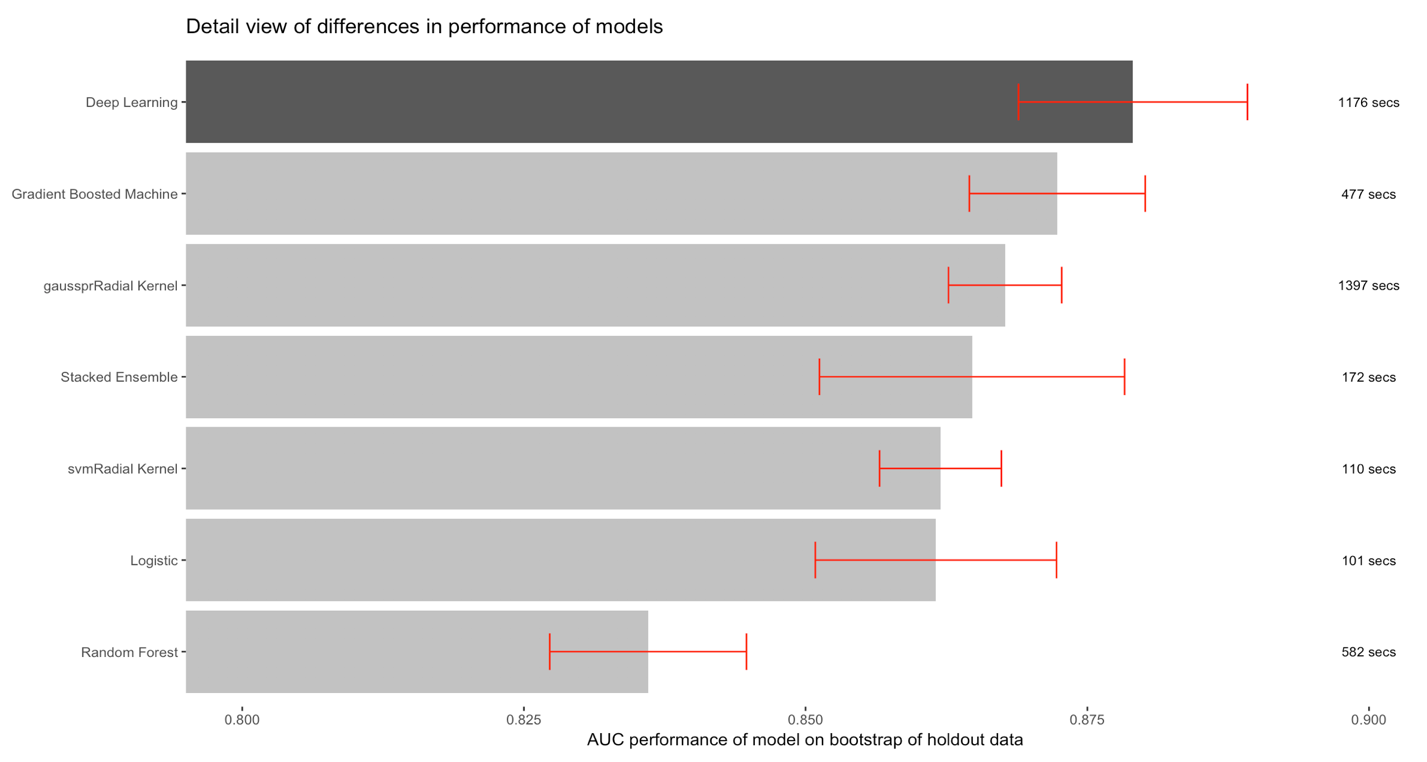
The most prominent drivers of the purchase decision are brand loyalty and the price of each orange juice. Increased brand loyalty to Citrus Hill decreases the likelihood that someone will purchase Minute Maid, as does a higher price on Minute Maid. A higher price on Citrus Hill increase the likelihood that a customer will purchase Minute Maid.

Other effects are not statistically significant at a confidence level of 95%, as can be seen in the chart below.



# Predictive Model

For the purpose of prediction, we evaluated 7 models with a wide range of tuning enhancements. The chart below shows how each of these models scores. All models performed well with reliable predictive power and different strengths and weaknesses. The basic logistic regression takes minimal computing resources and tends to generalize well. A deep learning model performs the best of all, but takes the most computing resources. If this model is likely to be used in the future and the resources to re-train the model are not a factor, then the deep learning model will continue to improve as it collects more information. However, if the intent is to run a simple model on a regular basis with minimal resources, then the logistic regression is cheap and “good enough”. The model choice relies on how the model will be implemented and if the run time is a concern.



# Methods

## Data Prep

### Data Cleaning

This dataset was already very clean with no missing values and no outliers. We found 24 duplicated rows (out of 1070 observations) and found that this did effect the training results of the models. These duplicates were removed.

### Data Classes

With some variables (such as the Week of Purchase), we could treat the variable as a factor or a numeric variable. First, we converted the week numbers to 1 through 52 for clarity. Next, we tested the models with these variables as factors and as continuous. A factor will allow the model to find non-linear relationships (such as certain months being different than other months) but is prone to overfitting. In this data, the numeric variable was much less prone to overfitting and produces the best results, though data with more than one year of observations may yield different results. If more than one year were present, autocorrelation (times-series approach) would be preferred.

### Variable Selection

We noticed that several of the columns in this dataset were simply linear combinations of other columns. We removed columns that were linear combinations of other columns and tested the performance of the model to confirm. We used the Caret function findLinearCombos(). We used variance inflation factor analysis to test for strong multicollinearity and found none (with the exception of an engineered feature that will be discussed in the next section).

### Feature Engineering

For the purpose of understanding which drivers influence a customer to purchase Minute Maid brand, we did not add any additional features to keep the message strong and simple. However, we found that by engineering a “MonthofPurchase” feature to supplement the “WeekofPurchase” variable, we were able to improve the predictive power of the model over keeping only one of the features. There must be a signal in Month that is hidden in the noise of Week that improving the model.

## Model Selection

### Cross-validated Training Scores

To score each model, we used the area under the ROC curve as our metric. Each model was trained with 10-fold cross validation to arrive at a confidence interval of expected performance against new data. The deep learning model performed the best, and its increased performance over the parsimonious logistic regression was statistically significant. We recognized that there may be additional user requirements (such as model training resources, frequency of re-training improvements, and parsimony) that will guide the choice of predictive model. The deep learning model performed the best and would improve the most with more data, but could not be interpreted for inference and consumed the most resources to train. The support vector machine with the Gaussian Process Radial kernel performed as well as the Deep Learning model, but with a narrower confidence interval so could be interpreted as more reliable (it took nearly as long to train). The logistic regression performed noticeably worse but the difference in performance may not be noteworthy and it trained in the shortest runtime.

### Confirmation with Bootstrapped Testing Scores

To report the most accurate expected performance of each model against new data, we looked at the performance of each model on a holdout dataset that we called “test”. We bootstrapped the test data to arrive at a 95% confidence interval and all figures were within the range of the expected performance scores from the training set, confirming that our choices and analysis did not have surprising results against new data.

## Variable importance

### Logistic regression

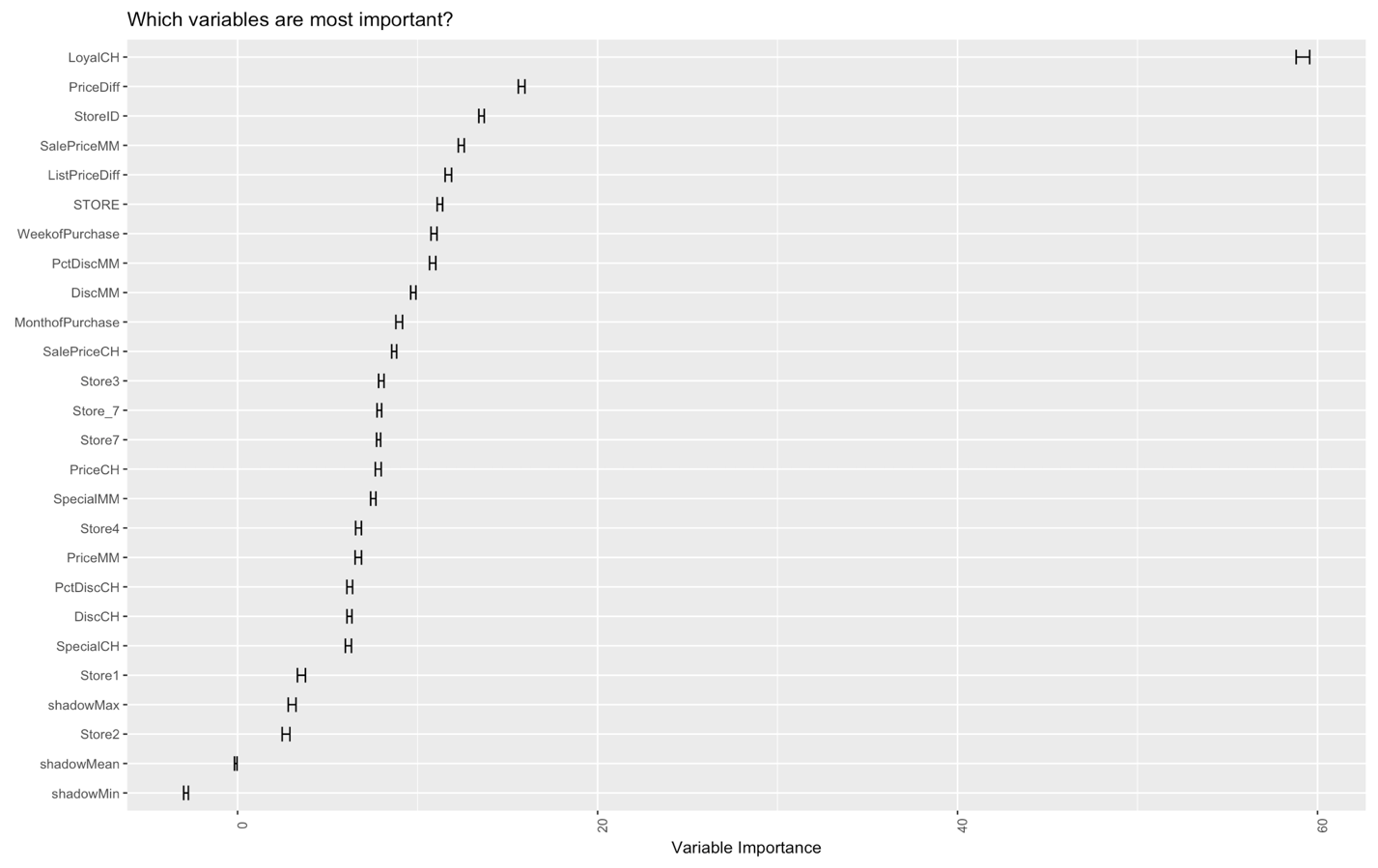
To understand what drives a consumer’s decision to purchase each brand or orange juice, we considered the standardized coefficients of the model. Since the ranges of each variable were different, it was necessary to standardize the variable importances to be able to compare apples to apples (or oranges to oranges in this case). We could be statistically confident at 95% in only three of the variables: customer loyalty to Citrus Hill, and the prices of each brand of orange juice at checkout.

### The problem of collinearity

Unfortunately, some collinearity does exist between the features that were kept, and linear combinations exist among the features that were removed. We wanted to ask more specific questions like: “How does the difference in prices matter compared to the prices themselves?” To do this, we applied a Boruta analysis to understand each specific driver.

### Boruta analysis

Boruta analysis assesses variable importance by adding randomized features called “shadows” and applying repeated random forests to them. Over several iterations, one can measure how much each specific variable impacts the total performance of a model apart from its correlation with other variables. The chart below should be used only as a reference to supplement understanding of the variable importances noted previously. A 95% confidence interval for each variable importance is included.



# Recommendations

## Loyalty

## Data

## Pricing

# Conclusion