**Business situation**

This is your first day as data scientist at a grocery store chain. You look at your schedule

and find that you have a meeting with both the brand manager and sales manager. At

the meeting they tell you that they want to know how to make the Orange Juice category

perform better than what it does currently. You learn that the grocery store chain sells

two brands of orange juice Citrus Hill (CH) and Minute Maid (MM). MM gets higher

margins than CH.

Listening to their description of the situation and their requirements it seems like the

Brand Manager is interested in finding out what variables influence a person’s

probability of buying MM. In other words, what can he do to increase customers’

probability of buying MM? On the other hand, the sales manager is interested in having

a predictive model where he can simply predict the probability of a customer

purchasing MM.

They look to you for providing a solution to their specific problems. You tell them that

although both relate to MM the nature of questions, they need answered makes the

analysis process very different. They tell you to use different analyses if needed but

inform them what you used to answer their question, why this method is better, and

importantly what are your specific recommendations. They ask you to present a written

report in which you clearly explain the method and your recommendations in a user friendly

manner.

You are given a dataset that contains 1070 purchases where the customer either

purchased Citrus Hill (CH) or Minute Maid (MM) Orange Juice [see appendix for details

of this dataset].

Considering the different goals of the brand and sales manager, you first need to figure

out whether using the same method can help you answer both their questions or would

you have to use different methods. Although you would prefer to use the same method, you realize that answering their specific queries as completely as possible is more

important. To understand the problem better you write down the questions asked by

each manager.

**Brand manager**

1. What predictor variables influence the purchase of MM?

2. Are all the variables in the dataset effective or are some more effective than

others?

3. How confident are you in your recommendations?

4. Based on your analysis what are specific recommendations you have for the

brand manager?

**Sales manager**

1. Can you provide him a predictive model that can tell him the probability of

customers buying MM?

2. How good is the model in its predictions?

3. How confident are you in your recommendations?

**Data Dictionary**

OJ is a data frame with 1070 observations on the following 18 variables.

**Purchase:** A factor with levels CH and MM indicating whether the customer purchased

Citrus Hill or Minute Maid Orange Juice

**WeekofPurchase:** Week of purchase. Here week 227 is week 1 of a year (i.e., January first week)

**StoreID:** Store ID

**PriceCH:** Price charged for CH. Also called List Price for CH

**PriceMM:** Price charged for MM. Also called List Price for MM

**DiscCH:** Discount offered for CH

**DiscMM:** Discount offered for MM

**SpecialCH:** Indicator of special on CH. Special can be a free gift, loyalty points etc.

**SpecialMM:** Indicator of special on MM. Special can be a free gift, loyalty points etc.

**LoyalCH:** Customer brand loyalty for CH. That is, probability to buy CH (over MM) based

on prior purchase behavior.

**SalePriceMM:** Sale price for MM. This is the difference between the list price and discount.

**SalePriceCH:** Sale price for CH. This is the difference between the list price and discount.

**PriceDiff:** Sale price of MM less sale price of CH

**Store7:** A factor with levels No and Yes indicating whether the sale is at Store 7

**PctDiscMM:** Percentage discount for MM

**PctDiscCH:** Percentage discount for CH

**ListPriceDiff:** List price of MM less list price of CH

**STORE:** Which of 5 possible stores the sale occurred at