Orange Juice Sales at Wasatch Grocery Chain

Identification of Significant Predictor Variables and Predictive Modelling of Customer
Preferance in Minute Maid Sales

Chris Gearheart and Chris Porter

2022-11-15

Introduction

Wasatch Grocery Chain (WGC) is a regional grocery chain operating in the Intermoutain West of the US. WGC sells two brands of orange juice in its stores, Citrus Hill (CH) and Minute Maid (MM) of which MM is the more profitable to the company. This report will identify what customer factors within available data contribute to purchase of MM over CH, as well as to what degree these factors influence customer choice. In addition, a predictive model is created that will allow the Sales Department to identify other customers within our customer base that are more likely to purchase Minute Maid brand orange juice, thus driving profitability across the company.

```
set.seed(1234)
df <- read.csv(url("http://data.mishra.us/files/project/OJ_data.csv"))
df[-1] <- lapply(df[-1],as.numeric)
df$Purchase <- as.factor(df$Purchase)
purchase_testtrain <- initial_split(df, prop = 0.75, strata = Purchase)
train <- training(purchase_testtrain)
test <- testing(purchase_testtrain)
```

Available Data

The data set used in this report contains 13 possible predictor variables as well as 1 outcome variable, Purchase, which records whether or not a customer purchased MM. There are a total of 1070 observations in the data set. The data set was further partitioned into a **training** data set, containing 801 observations, and a validation **testing** data set containing 269 observations.

Methods

Results

```
set.seed(1234)
recipe_oj <- recipe(Purchase ~ ., train)

model_oj_bt <- boost_tree(trees = tune(), tree_depth = tune(), learn_rate = tune()) %>%
  set_engine('xgboost', verbosity = 0) %>%
  set_mode('classification')

hyperparameter_grid <- grid_regular(trees(), tree_depth(), learn_rate(), levels = 5)

purchase_folds <- vfold_cv(train, v=4) # 4-fold Cross validation

oj_workflow <- workflow() %>% add_model(model_oj_bt) %>% add_recipe(recipe_oj) #Set Workflow

# Tune Hyper-parameters
oj_tune <- oj_workflow %>% tune_grid(resamples = purchase_folds,
                                     grid = hyperparameter_grid,
                                     metrics = metric_set(accuracy))

best_bt_model <- oj_tune %>% select_best('accuracy') #Select best Hyper-parameters from grid

best_bt_model
```

```
## # A tibble: 1 x 4
##   trees tree_depth learn_rate .config
##   <int>      <int>      <dbl> <chr>
## 1  1000          1        0.1 Preprocessor1_Model023
```
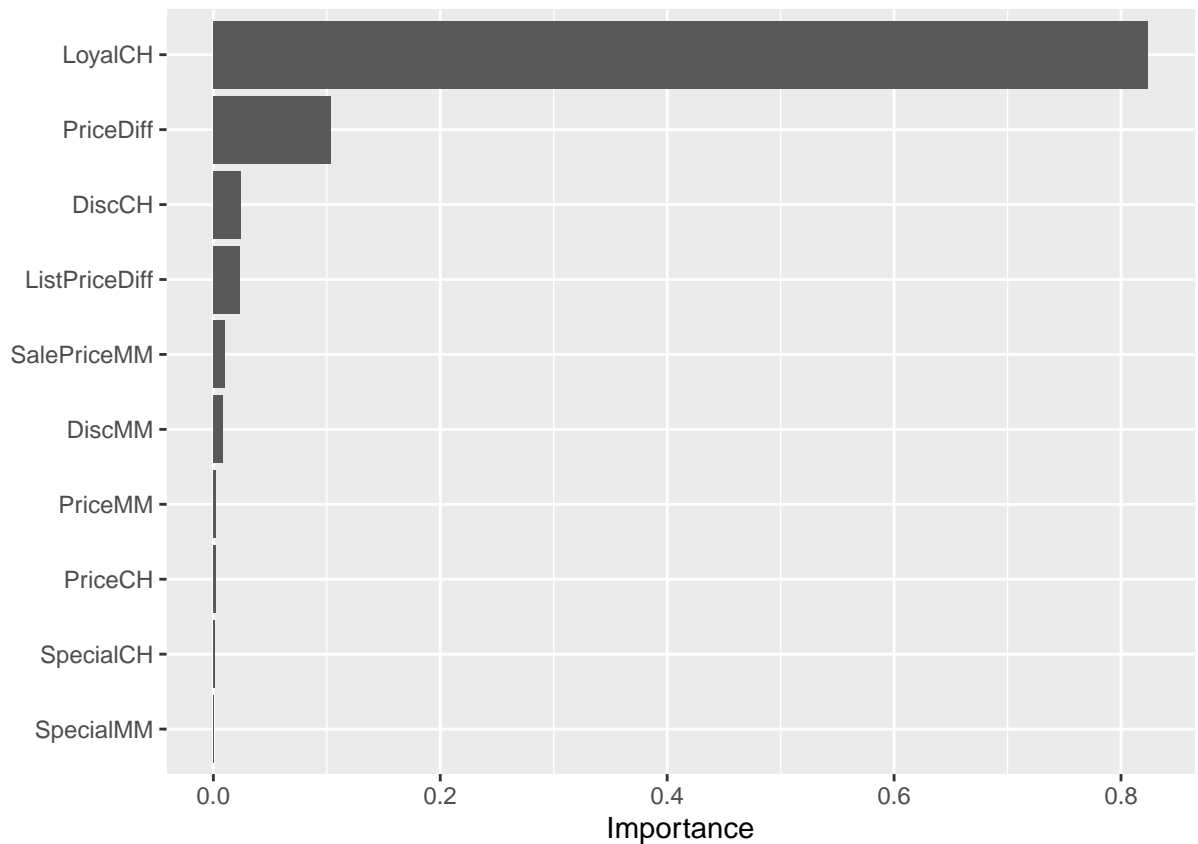
```
oj_final_workflow <- oj_workflow %>% finalize_workflow(best_bt_model) # Create Final Workflow based up

final_fit <- oj_final_workflow %>% last_fit(split = purchase_testtrain) # Final Fit Model

final_fit %>% collect_metrics()
```

```
## # A tibble: 2 x 4
##   .metric  .estimator .estimate .config
##   <chr>    <chr>          <dbl> <chr>
## 1 accuracy binary         0.799 Preprocessor1_Model1
## 2 roc_auc  binary         0.892 Preprocessor1_Model1
```

```
oj_final_workflow %>% fit(data = train) %>% extract_fit_parsnip() %>% vip(geom = 'col') #Plot most impo
```

```
vi_values <- oj_final_workflow %>% fit(data = train) %>% extract_fit_parsnip() %>% vi()
vi_gt_1 <- vi_values %>% filter(Importance >= 0.01)
vi_gt_1
```

```
## # A tibble: 5 x 2
##   Variable     Importance
##   <chr>             <dbl>
## 1 LoyalCH          0.824
## 2 PriceDiff        0.104
## 3 DiscCH           0.0239
## 4 ListPriceDiff    0.0232
## 5 SalePriceMM      0.0104
```

the most important variable is LoyalCH with a 82.3503048%

```r
model_fitted <- oj_final_workflow %>% fit(data = train)

explainer_rf <- explain_tidymodels(model_fitted,
                                   data = train[,-1],
                                   y = train$Purchase,
                                   type = "pdp",verbose = FALSE)

pdp_LoyalCH <- model_profile(explainer_rf,
                             variables = "LoyalCH",
                             N=NULL)
pdp_PriceDiff <- model_profile(explainer_rf,
                               variables = "PriceDiff",
                               N=NULL)
pdp_DiscCH <- model_profile(explainer_rf,
                            variables = "DiscCH",
                            N=NULL)
pdp_ListPriceDiff <- model_profile(explainer_rf,
                             variables = "ListPriceDiff",
                             N=NULL)
pdp_SalePriceMM <- model_profile(explainer_rf,
                             variables = "SalePriceMM",
                             N=NULL)
pdp_DiscMM <- model_profile(explainer_rf,
                            variables = "DiscMM",
                            N=NULL)

plot(pdp_LoyalCH)
```
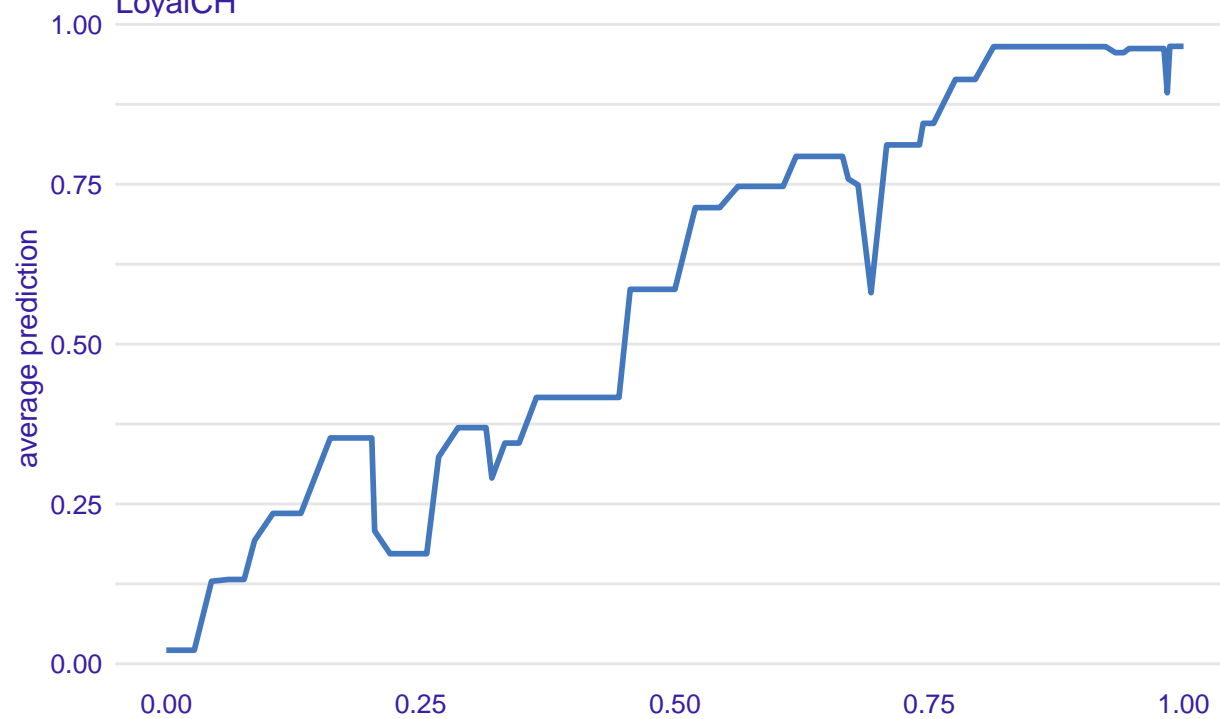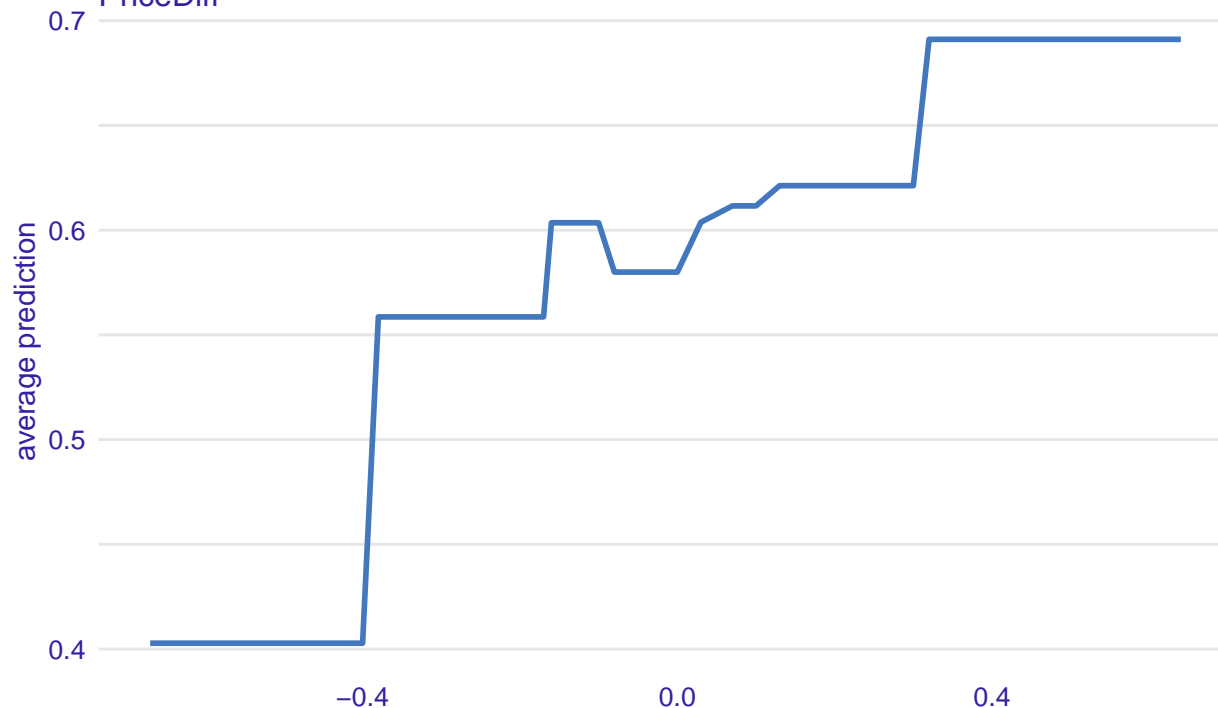
Partial Dependence profile

Created for the workflow model
LoyalCH


```
plot(pdp_PriceDiff)
```

Partial Dependence profile
Created for the workflow model
PriceDiff


```
#plot(pdp_DiscCH)
#plot(pdp_ListPriceDiff)
#plot(pdp_SalePriceMM)
#plot(pdp_DiscMM)
```

Conclusions and Recommendations

Brand

Sales

Appendix 1: Data Characteristics

```
summary(df)
```

```
##  Purchase    PriceCH         PriceMM         DiscCH           DiscMM
##  0:417    Min.   :1.690   Min.   :1.690   Min.   :0.00000   Min.   :0.0000
##  1:653    1st Qu.:1.790   1st Qu.:1.990   1st Qu.:0.00000   1st Qu.:0.0000
##           Median :1.860   Median :2.090   Median :0.00000   Median :0.0000
##           Mean   :1.867   Mean   :2.085   Mean   :0.05186   Mean   :0.1234
##           3rd Qu.:1.990   3rd Qu.:2.180   3rd Qu.:0.00000   3rd Qu.:0.2300
##           Max.   :2.090   Max.   :2.290   Max.   :0.50000   Max.   :0.8000
##    SpecialCH        SpecialMM         LoyalCH          SalePriceMM
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.000011   Min.   :1.190
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.325257   1st Qu.:1.690
##  Median :0.0000   Median :0.0000   Median :0.600000   Median :2.090
##  Mean   :0.1477   Mean   :0.1617   Mean   :0.565782   Mean   :1.962
```

6

```
##    3rd Qu.:0.0000    3rd Qu.:0.0000    3rd Qu.:0.850873    3rd Qu.:2.130
##    Max.    :1.0000    Max.    :1.0000    Max.    :0.999947    Max.    :2.290
##     SalePriceCH        PriceDiff         PctDiscMM          PctDiscCH
##    Min.    :1.390    Min.    :-0.6700    Min.    :0.0000    Min.    :0.00000
##    1st Qu.:1.750    1st Qu.: 0.0000    1st Qu.:0.0000    1st Qu.:0.00000
##    Median :1.860    Median : 0.2300    Median :0.0000    Median :0.00000
##    Mean    :1.816    Mean    : 0.1465    Mean    :0.0593    Mean    :0.02731
##    3rd Qu.:1.890    3rd Qu.: 0.3200    3rd Qu.:0.1127    3rd Qu.:0.00000
##    Max.    :2.090    Max.    : 0.6400    Max.    :0.4020    Max.    :0.25269
##    ListPriceDiff
##    Min.    :0.000
##    1st Qu.:0.140
##    Median :0.240
##    Mean    :0.218
##    3rd Qu.:0.300
##    Max.    :0.440
```

summary(test)
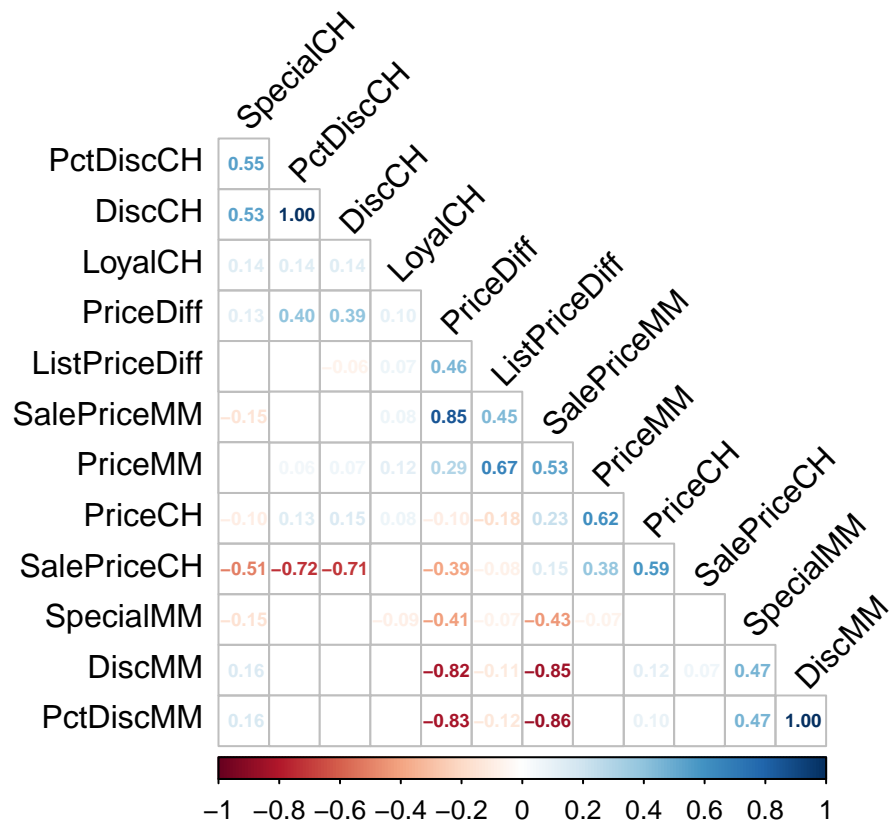
```
##    Purchase    PriceCH          PriceMM          DiscCH            DiscMM
##    0:105    Min.    :1.690    Min.    :1.690    Min.    :0.00000    Min.    :0.0000
##    1:164    1st Qu.:1.790    1st Qu.:1.990    1st Qu.:0.00000    1st Qu.:0.0000
##             Median :1.860    Median :2.090    Median :0.00000    Median :0.0000
##             Mean    :1.874    Mean    :2.079    Mean    :0.05167    Mean    :0.1094
##             3rd Qu.:1.990    3rd Qu.:2.180    3rd Qu.:0.00000    3rd Qu.:0.2000
##             Max.    :2.090    Max.    :2.290    Max.    :0.50000    Max.    :0.8000
##     SpecialCH        SpecialMM         LoyalCH           SalePriceMM
##    Min.    :0.0000    Min.    :0.0000    Min.    :0.000011    Min.    :1.190
##    1st Qu.:0.0000    1st Qu.:0.0000    1st Qu.:0.384000    1st Qu.:1.780
##    Median :0.0000    Median :0.0000    Median :0.635200    Median :2.090
##    Mean    :0.1264    Mean    :0.1413    Mean    :0.595184    Mean    :1.969
##    3rd Qu.:0.0000    3rd Qu.:0.0000    3rd Qu.:0.875808    3rd Qu.:2.130
##    Max.    :1.0000    Max.    :1.0000    Max.    :0.999870    Max.    :2.290
##     SalePriceCH        PriceDiff         PctDiscMM          PctDiscCH
##    Min.    :1.390    Min.    :-0.6700    Min.    :0.00000    Min.    :0.00000
##    1st Qu.:1.750    1st Qu.: 0.0000    1st Qu.:0.00000    1st Qu.:0.00000
##    Median :1.860    Median : 0.2300    Median :0.00000    Median :0.00000
##    Mean    :1.823    Mean    : 0.1468    Mean    :0.05231    Mean    :0.02709
##    3rd Qu.:1.890    3rd Qu.: 0.3000    3rd Qu.:0.09569    3rd Qu.:0.00000
##    Max.    :2.090    Max.    : 0.6400    Max.    :0.40201    Max.    :0.25269
##    ListPriceDiff
##    Min.    :0.0000
##    1st Qu.:0.1000
##    Median :0.2400
##    Mean    :0.2045
##    3rd Qu.:0.2900
##    Max.    :0.4400
```

summary(train) *#need to equalize the 0/1 split in train data set*

```
##    Purchase    PriceCH          PriceMM          DiscCH            DiscMM
##    0:312    Min.    :1.690    Min.    :1.690    Min.    :0.00000    Min.    :0.0000
##    1:489    1st Qu.:1.790    1st Qu.:1.990    1st Qu.:0.00000    1st Qu.:0.0000
##             Median :1.860    Median :2.090    Median :0.00000    Median :0.0000
##             Mean    :1.865    Mean    :2.088    Mean    :0.05192    Mean    :0.1281
```

```
##             3rd Qu.:1.990   3rd Qu.:2.180   3rd Qu.:0.00000   3rd Qu.:0.2400
##             Max.   :2.090   Max.   :2.290   Max.   :0.50000   Max.   :0.8000
##    SpecialCH       SpecialMM       LoyalCH         SalePriceMM
##  Min.   :0.0000  Min.   :0.0000  Min.   :0.000014  Min.   :1.19
##  1st Qu.:0.0000  1st Qu.:0.0000  1st Qu.:0.320000  1st Qu.:1.69
##  Median :0.0000  Median :0.0000  Median :0.585435  Median :2.09
##  Mean   :0.1548  Mean   :0.1685  Mean   :0.555908  Mean   :1.96
##  3rd Qu.:0.0000  3rd Qu.:0.0000  3rd Qu.:0.836160  3rd Qu.:2.18
##  Max.   :1.0000  Max.   :1.0000  Max.   :0.999947  Max.   :2.29
##    SalePriceCH      PriceDiff       PctDiscMM         PctDiscCH
##  Min.   :1.390   Min.   :-0.6700  Min.   :0.00000   Min.   :0.00000
##  1st Qu.:1.750   1st Qu.: 0.0000  1st Qu.:0.00000   1st Qu.:0.00000
##  Median :1.860   Median : 0.2400  Median :0.00000   Median :0.00000
##  Mean   :1.813   Mean   : 0.1464  Mean   :0.06164   Mean   :0.02739
##  3rd Qu.:1.890   3rd Qu.: 0.3200  3rd Qu.:0.11834   3rd Qu.:0.00000
##  Max.   :2.090   Max.   : 0.6400  Max.   :0.40201   Max.   :0.25269
##  ListPriceDiff
##  Min.   :0.0000
##  1st Qu.:0.1400
##  Median :0.2400
##  Mean   :0.2225
##  3rd Qu.:0.3000
##  Max.   :0.4400
```

```r
corr <- cor(df[-1]) #correlogram of numeric variables, excluding outcome variable
testDf <- cor.mtest(df[-1], conf.level = 0.95) #compute significance of correlation
# Plot correlogram
corrplot(corr, p.mat = testDf$p, method = 'number', type = 'lower', insig='blank',
         addCoef.col ='black', number.cex = 0.6, order = 'AOE', diag=FALSE, tl.srt = 45, tl.col = 'blacl
```

```
# TEST TEST TEST Chris gearheart's test
```