Customer Behavior & Wine Purchasing Analysis

Venus Chen
Yuqing Wang
Surapa Kongchan
Yue Li

Agenda

1. Business Understanding
2. Data Understanding
3. Supervised Model
4. Unsupervised Model
5. Prescriptive Analysis Model

Business Understanding

Overall Goal of the Analysis

- Helping a business to better understand its customers
- Targeting the right group of customers who are likely to buy wines
- Providing foundation information a firm can use to modify its marketing strategies based on customers' specific characteristics
    - Allowing the firm to spend the marketing budget most effectively and target specific customers most accurately

Data Understanding

Dataset

- From a third-parties online database: Kaggle
- Containing customers' background information and information about where the customers purchased the products, transactions' information, and related promotions
- There are 2240 observations, with 30 variables/predictors associated with each customers

Data Quality & Preparation

- Key variables includes year of birth, income, numbers of children at home, amount spent on wines, number of store purchases, amount spent on meat, whether or not a customer accept the offer promotion, and number of purchases made with a discount
- We also added an additional variable of age by subtracting the current year (2021) with year of birth

- There are 24 missing values for income
  - We substituted N/A with average income
- Outliers are found in the amount spent on wines and income
  - We dropped the extreme observations that exceeded 95th percentile for amount spent on wines
  - Outliers for income are reasonable
- Incorrect values are found in income (rows containing 666,666)
  - We dropped the incorrect values

- Main variables that will be used are normally distributed

```
    Income
 Min.   :  1730
 1st Qu.: 35303
 Median : 51382
 Mean   : 52247
 3rd Qu.: 68522
 Max.   :666666
 NA's   :24
```

```
    Income
 Min.   :  1730
 1st Qu.: 33494
 Median : 47723
 Mean   : 48985
 3rd Qu.: 64168
 Max.   :162397
```

Data Description

Variable	Type	Description
Income	Numerical	Customer's yearly household income
Age	Numerical	Customer's age
Kidhome	Numerical	Number of children in household
MntWines	Numerical	Amount spent on wine in last 2 years
MntMeatProducts	Numerical	Amount spent on meat in last 2 years
NumDealsPurchases	Numerical	Number of purchases made with a discount
Response	Categorical	1 if customer accepted the offer in the last campaign; 0 otherwise

Supervised Model: Linear Regression 1

Goal of the regression: predicting the amount spent on wines, based on continuous predictors of customers' age and income

- We believed that both predictors have an impact on the amount spent on wines
  - The higher the age and income, the higher amount spent on wines

```
Call:
lm(formula = MntWines ~ AGE + Income, data = marketing_campaign)

Residuals:
     Min       1Q   Median       3Q      Max 
-1227.94  -116.92   -30.21    86.96   676.09 

Coefficients:
              Estimate   Std. Error t value            Pr(>|t|)    
(Intercept) -234.1901798  20.5920457 -11.373 <0.0000000000000002 ***
AGE            0.6859714   0.3710798   1.849              0.0647 .  
Income         0.0090154   0.0002099  42.944 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 199.9 on 2125 degrees of freedom
Multiple R-squared:  0.4792,	Adjusted R-squared:  0.4787 
F-statistic: 977.7 on 2 and 2125 DF,  p-value: < 0.00000000000000022
```

The analysis showed that:

- Age is not significant; the coefficient of age indicates that additional 1 year increase in age is correlated with an increase of $0.69 of wines spending in the last 2 years
- Income is significant; the coefficient of income indicates that additional $1 increase in income is correlated with an increase of $0.009 amount of wines spending in the last 2 years

RMSE for the validation data = 202.7

RMSE for training data = 197.7

Supervised Model: Linear Regression 2

In order to have a better prediction, we improved the linear model by adding more variables to the model: number of children in household and number of purchases made with a discount

- We believed that both predictors have an impact on the amount spent on wines
  - Discount leads to higher spending and having kids might decrease the amount spent on

```
Call:
lm(formula = MntWines ~ AGE + Income + Kidhome + NumDealsPurchases,
    data = train_set)

Residuals:
     Min       1Q   Median       3Q      Max
-1380.09  -104.01   -24.42    65.39   654.27

Coefficients:
                    Estimate  Std. Error t value            Pr(>|t|)
(Intercept)      -103.5479702  32.3510491  -3.201             0.00141 **
AGE                 0.0402653   0.5235081   0.077             0.93870
Income              0.0071901   0.0002963  24.266 < 0.0000000000000002 ***
Kidhome          -143.5667821  11.8756541 -12.089 < 0.0000000000000002 ***
NumDealsPurchases  25.3919760   2.9273340   8.674 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 188.8 on 1215 degrees of freedom
Multiple R-squared:  0.5328,    Adjusted R-squared:  0.5313
F-statistic: 346.4 on 4 and 1215 DF,  p-value: < 0.00000000000000022
```

The analysis showed that:

- KidHome is significant; the coefficient of KidHome indicates that an additional 1 child increase in household is correlated with a decrease of $143.57 amount of wines spending in the last 2 years
- NumDealsPurchases is significant; the coefficient of NumDealsPurchases indicates that every 1 purchases with a discount is correlated with an increase of $25.39 amount of wines spending in the last 2 years

RMSE for the validation data = 178.84

RMSE for training data = 188.45

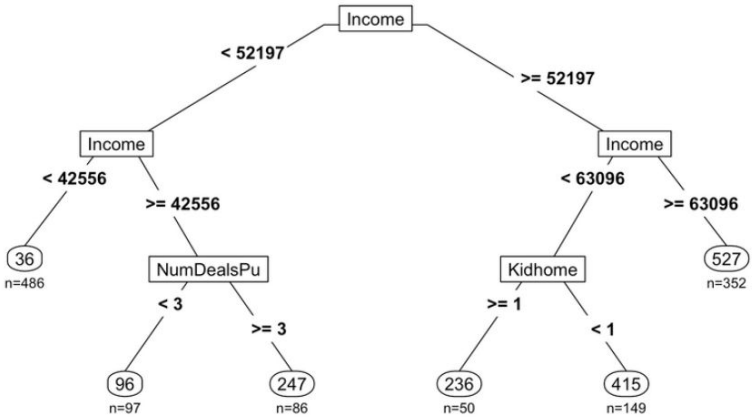
Supervised Model: Regression Tree

Goal of the regression: predicting and visualizing the value of the amount each customer spent on wine, based on the same set of numerical predictors used in linear prediction model

Target Variable: Amount spent on wine in the last 2 years

Predictors: Income, age, number of kids at home, number of purchases made with a discount

```
> rpart.rules(reg_tree_3)
 MntWines
       36 when Income <  42556
       96 when Income is 42556 to 52197                & NumDealsPurchases <  3
      236 when Income is 52197 to 63096 & Kidhome >= 1
      247 when Income is 42556 to 52197                & NumDealsPurchases >= 3
      415 when Income is 52197 to 63096 & Kidhome <  1
      527 when Income >=          63096
```

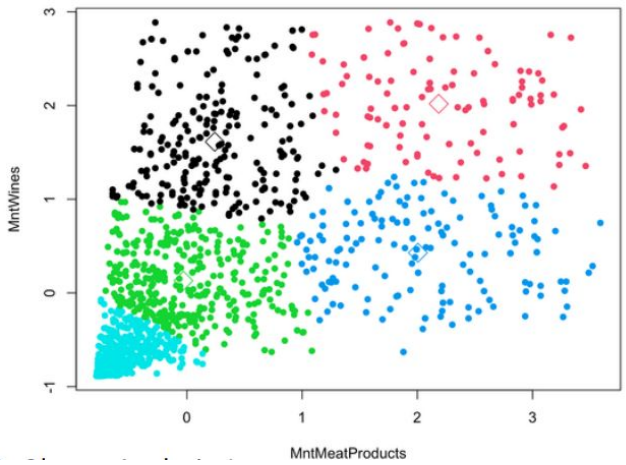


RMSE for the validation data = 173.43
RMSE for training data = 188.46

Unsupervised Model: Cluster Analysis 1

Goal of clustering: forming groups of similar customers, based on their transactions and number of purchases where customers make in one store

K-mean Analysis: group of amount spent on meat products and wines

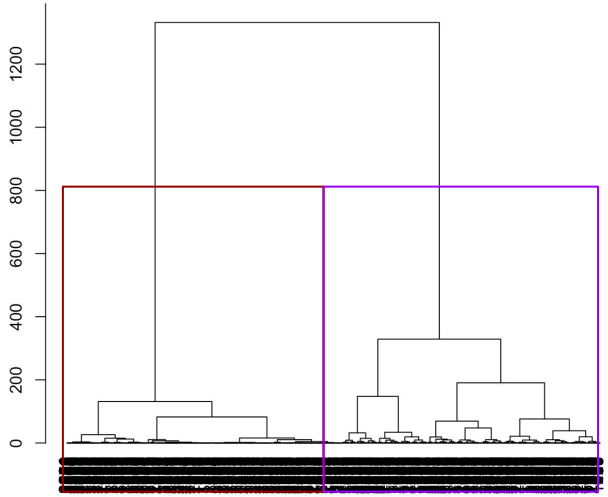


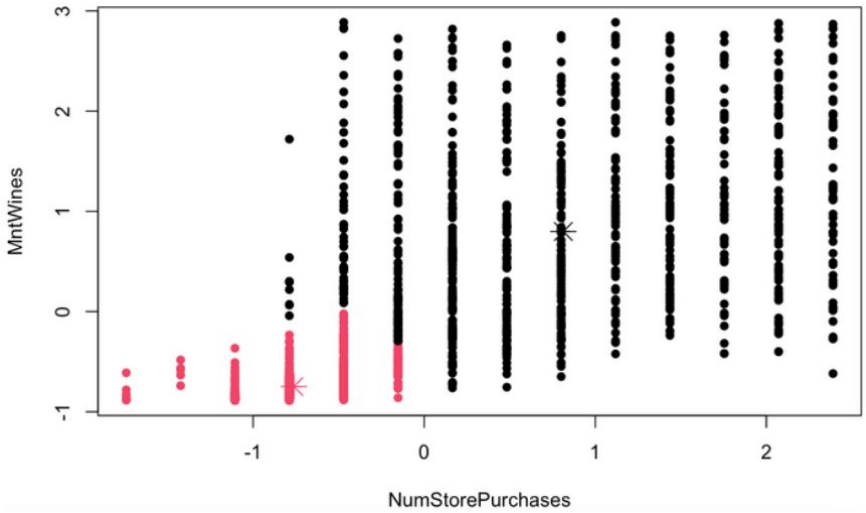
- customers who did not spend much on both meat and wines (1008 observations)
- customers who spent money on meat but not much on wine (165 observations)
- customers who spent a relative amount on meat and wines (96 observations)
- customers who spent much on wine, but not meat (241 observations)
- customers who spent on a lot of both meat and wine (121 observations)

Unsupervised Model: Cluster Analysis 2

Goal of the clustering: forming groups of similar customers, based on their transactions and number of purchases where customers directly make in a store

Ward Method: group of the amount spent on wine and in-store purchases

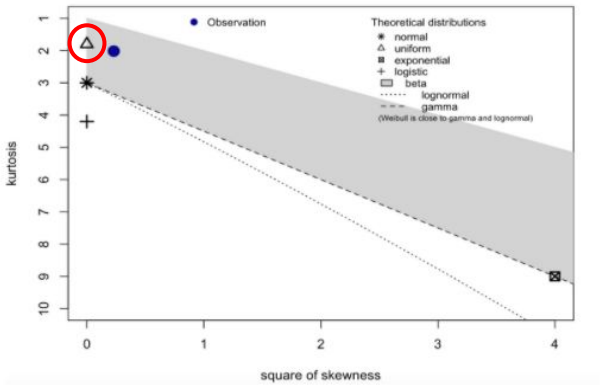




Prescriptive Analysis: Monte Carlo Simulation

Goal of the simulation: to better understand the uncertainty in total spending on wines by a group of new customers who accept the discount promotion

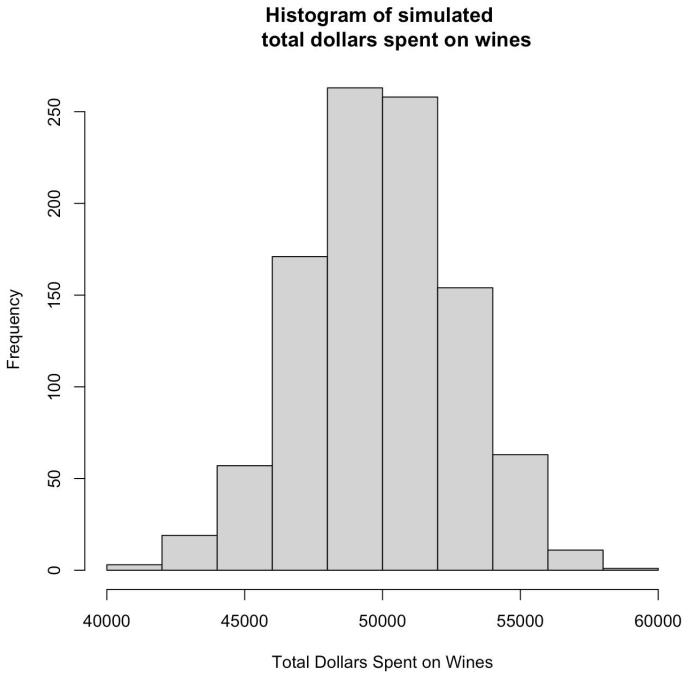
Assumptions: Suppose that the store attracted 100 new customers in a recent campaign. We assumed that the spending on wines from these customers for the next 2 years followed the same distribution pattern as the previous customers who accepted the promotion for the last 2 years



Parameters of the distribution:
Min = 1
Max = 997

Fittest distribution to the average amount of spending each time of the observations is most close to the uniform line

Prescriptive Analysis: Monte Carlo Simulation




```
WineSpent_vec <- c()
nsim <- 1000

for (i in 1:nsim) {
  sim_W <- runif(n = 100, min = 1, max = 997)
  WineSpent_vec[i]<- sum(sim_W)
}
```

Findings:

- The average total spending on wines by this group of new 100 customers would be around $49,911
- The 95% confidence intervals for the expected amount of spending fall into a range of $49,735 to $50,087
- The probability of receiving more than $50,000 in total spending from these 100 new customers is around 48.1%

Q & A