ITEM-X Project Analysis

Team 5 –

Cenli Han | Dawei Jia | Jinchuan Yang | Siqi Zhi | Yufei Jiang

Roadmaps:

1. Data Preparation (Data plan)

- **Survey data**: customers' attitudes toward attributes M, E, V, C of ITEM-X
- **Insiders data**: customers' purchase behavior related to ITEM-X and its alternatives
- **Other top 30 data**: high-spending customers' purchase behavior related to ITEM-X and its alternatives
- **Customer data**: customers' demographic statistics

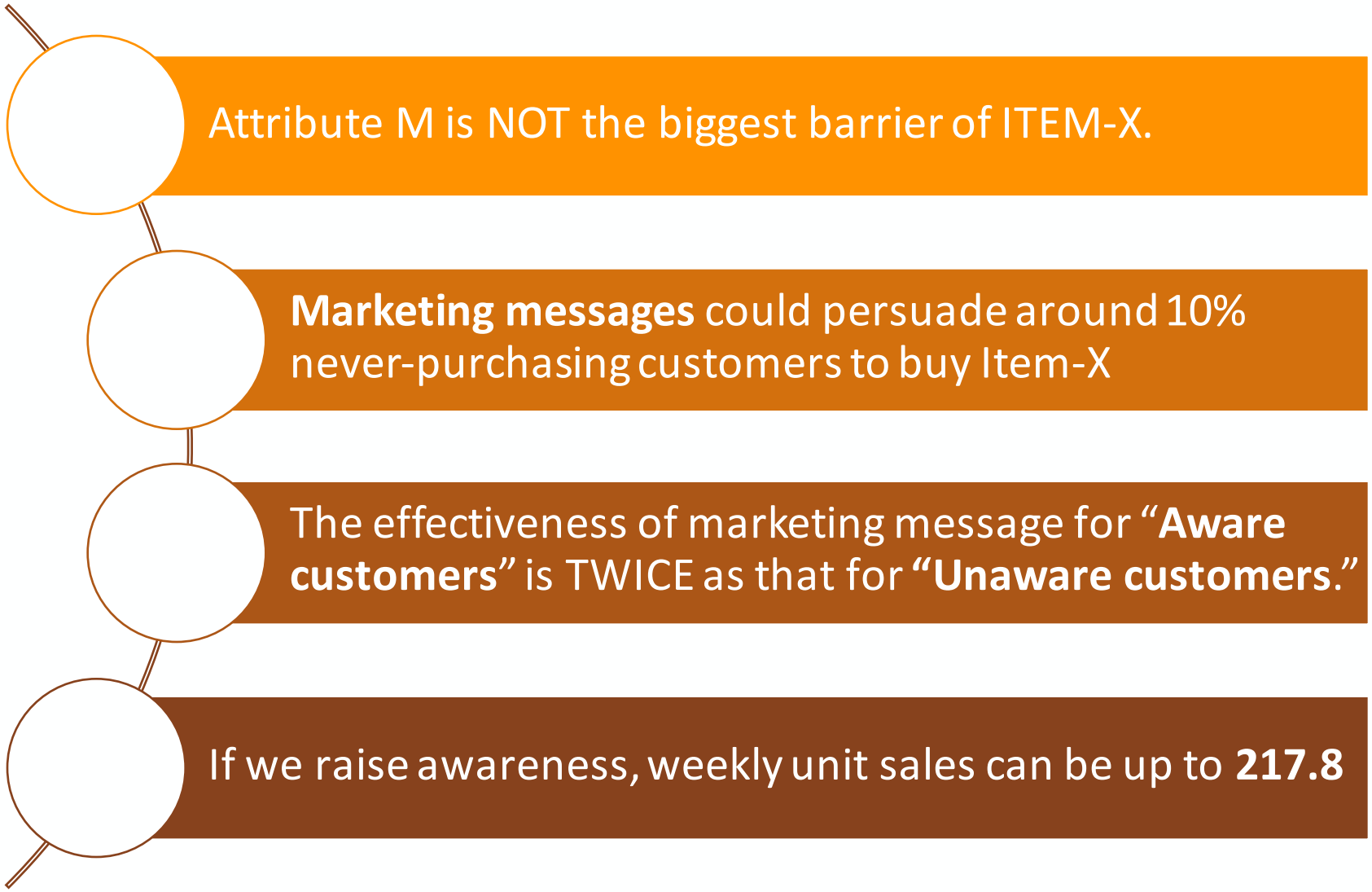
2. Data analysis (Analytics plan)

- **Representativeness check**
- **Explanatory analysis**: link insiders customers' attitudes data to their purchase behavior
- **Predictive analysis**: Predict customers' purchase decision by their past purchase and demographic statistics.

3. Conclusion

- Analyze attribute M by using survey and insiders data.
- Predict **potential market** based on **predictive model** built on full set data.

Summary of Wegmans data analysis



Key assumptions:

- When estimating never-purchase customers:
  - Only customers replied "N-" with attribute M are concerned with it.
  - Customers who have a concern and do not purchase ITEM-X within 2 months after survey (between 11/12/17 and 1/20/18) will never purchase.
  - Customers who didn't answer the follow-up question didn't see that message.

- When estimating potential customers:
  - We can raise awareness of every single person in the club.
  - Customers in Insider database are all aware of ITEM-X.
  - Potential customers are all Shoppers Club members and have demographic data.

Limit of Wegmans data analysis

Error rate

- Customers end point choice (intention) sometimes don’t reflects their real awareness.
- “Post units” in transaction data has purchase record for only 2 months.

NAs

- What do customers who have no response to certain questions think about attribute M?
- Do Customers who didn't answer the follow-up question saw that massage?

Predictive model

- Predictive models can’t be used to estimate non-club customers’ buying decision.
- Sample bias and method bias still exist.

Data Plan

A. Is attribute 'M' a major barrier

Database Name	barrierM
Data Source	Survey & insider

Data Design:

- Select HH, Endpoint, followup from survey data where column M = 'N-'
- According to HH, select UNITS, UNITS_POST, and ALTERNATIVE column from insider data

Row	A part of Survey&insider observations
Column	HH , Endpoint, Followup, UNITS, UNITS_POST, ALTERNATIVE

B. Decision Tree (Whether consumer will buy X)

Database Name	whetherbuyX
Data Source	insider & custdata

Data Design:

- Select HH from insider & survey data who have bought item-X, defining whetherT0 = 'pos'
- Select HH from insider & survey data who have not bought item-X, defining whetherT0 = 'neg'
- Combine custdata with whetherT0 data based on HH

Row	A part of custdata observations
Column	HH, DECILE, ZONE_NBR, HOH_AGE, HH_INCOME, HH_ADULTS, HH_CHILDREN, whetherT0"

C. Predictive model (How much will be sold)

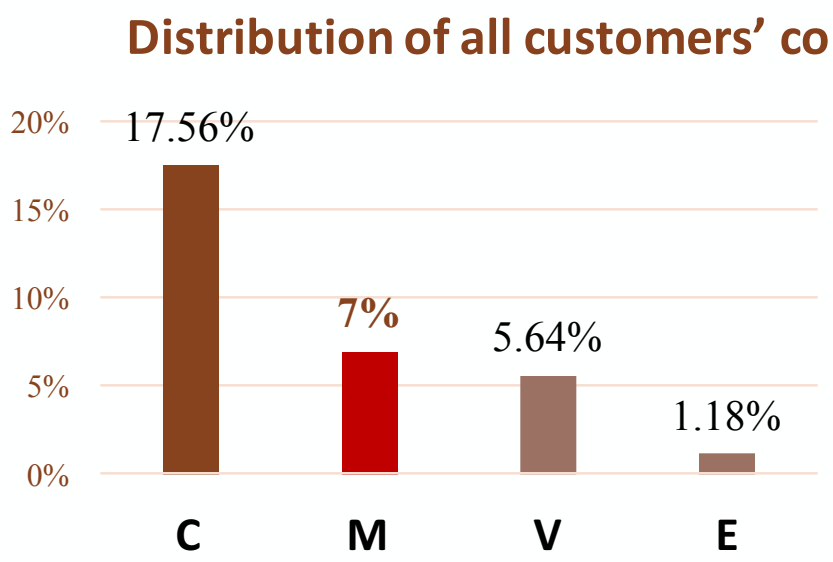
Database Name	unitsell
Data Source	insider & custdata

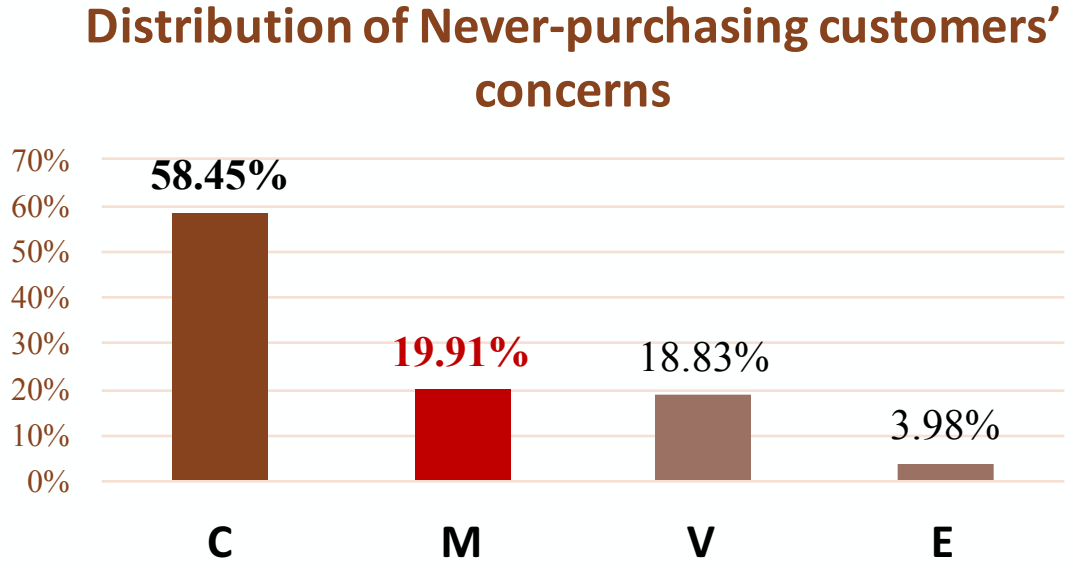
Data Design:

- Calculate units of T0 every HH have bought (UNITS + UNITS_POST = unitbought)
- Combine custdata with unitbought based on HH

Row	A part of custdata observations
Column	HH, DECILE, ZONE_NBR, HOH_AGE, HH_INCOME, HH_ADULTS, HH_CHILDREN, unitbought

Attribute M is not the biggest concern, compared with attribute C

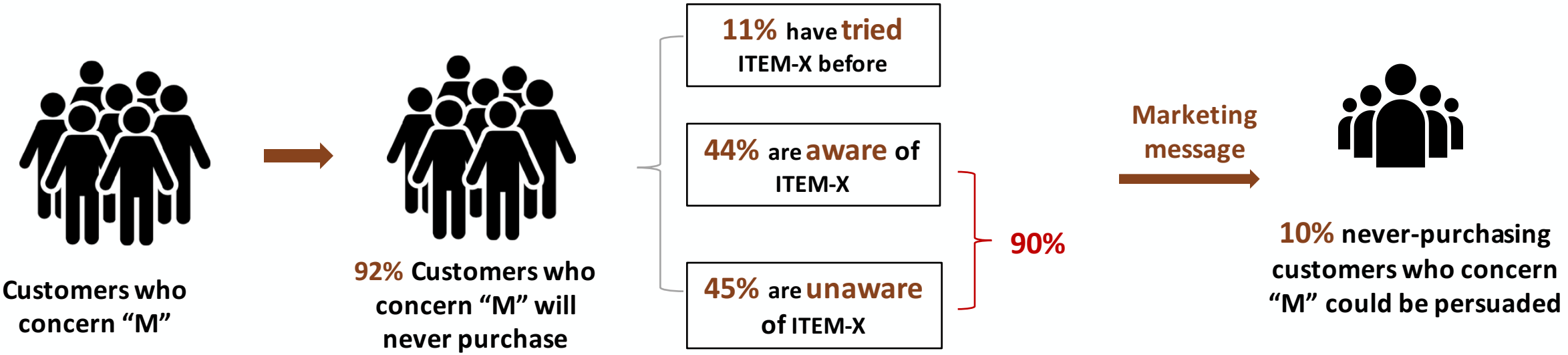




Conclusion

- Around **7%** of all the customers have a concern with attribute "M", which is not a significantly big percentage.
- Over **half of the never-purchasing customers** have a concern with **attribute C**, which is more likely to be considered as a major barrier instead of attribute M.

Marketing messages could persuade around 10% never-purchasing customers who concern with "M"



Conclusion

- Around **92%** customers who have a concern with attribute "M" will never purchase, and 10% of them could be persuaded by marketing messages.
- **Over 90%** of never-purchasing customers who concern "M" **haven't tried ITEM-X before,** and they are the **primary target group** of the **marketing message**.

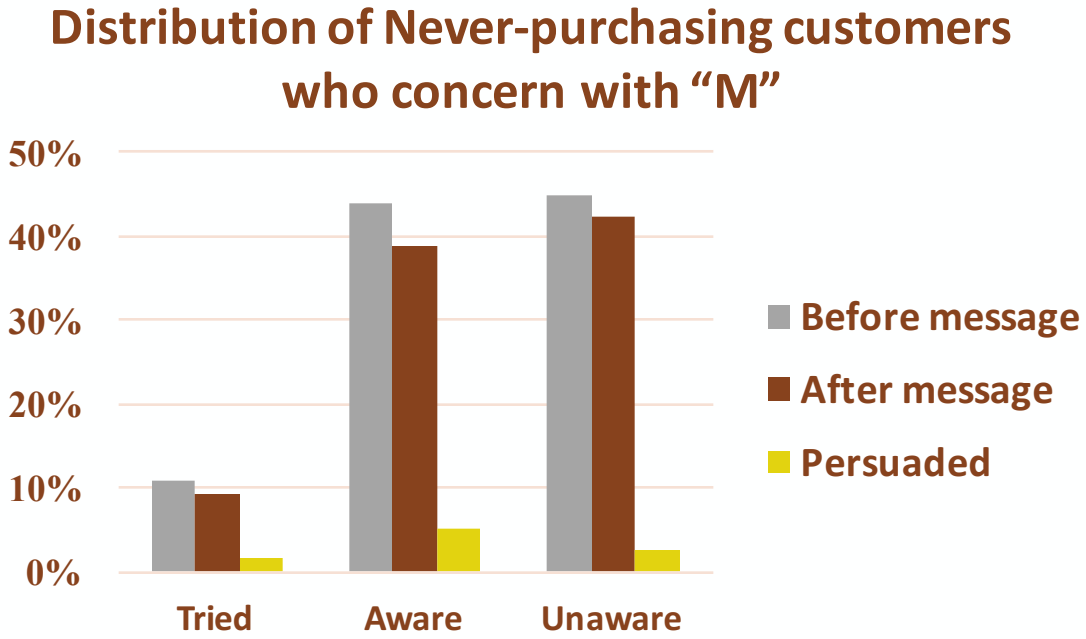
Marketing messages are much more effective for “Aware customers” than “Unaware customers”

Customers persuaded by marketing message

Tried customers
16% could be persuaded

Aware customers
12% could be persuaded

Unaware customers
6% could be persuaded



Conclusion

- **The effectiveness of marketing message** for “Aware customers” is **twice** as that for “Unaware customers.”
- Marketing message could **get a better result** if Wegmans could **increase ITEM-X’s overall awareness** to make more “Unaware customers” become “Aware customers” at first.

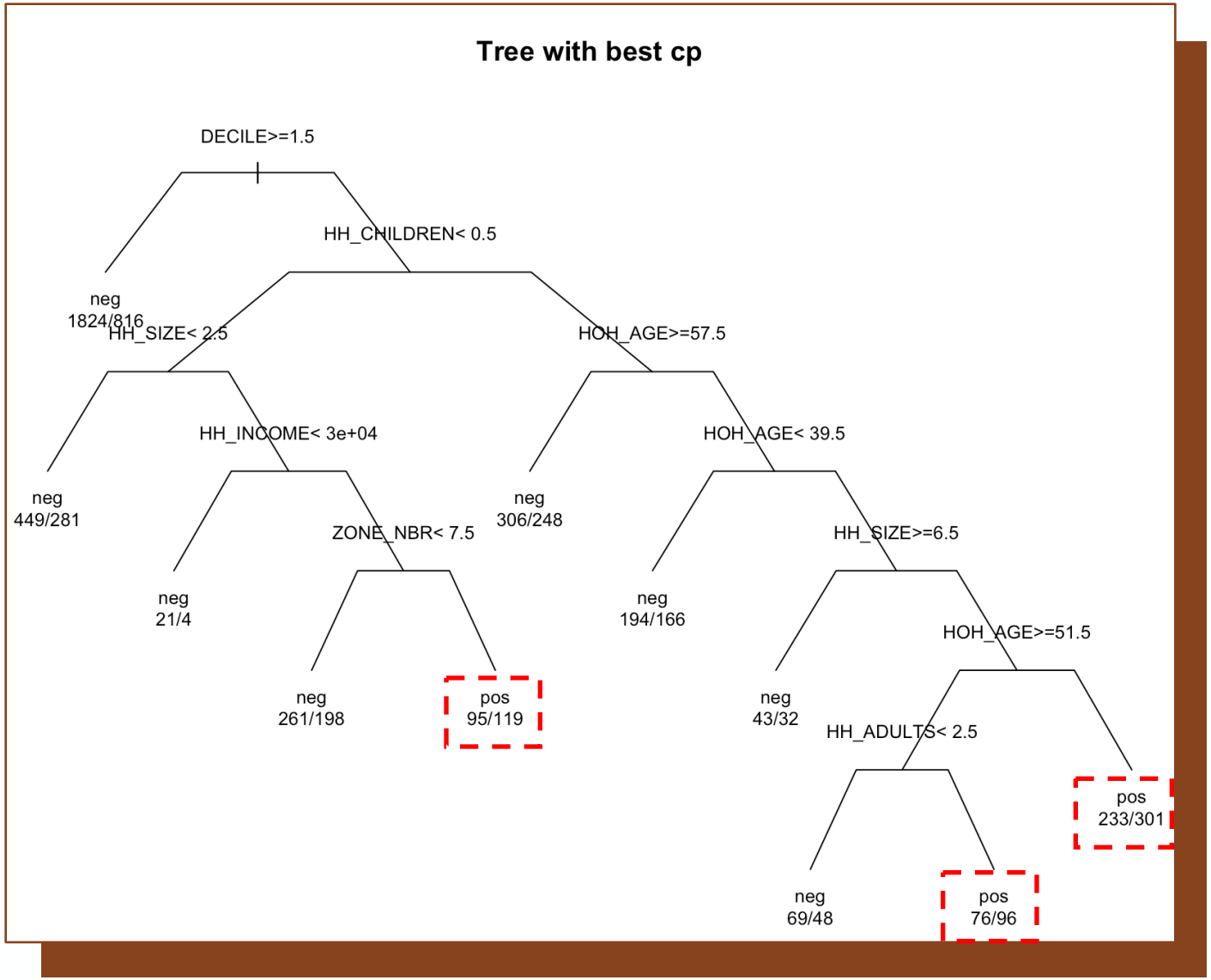
Size of Prize: Who will buy? -- Decision Tree

Database and Variables

- Insider data join customer data.
- Remove HH column when running the model.
- A new variable WhetherT0 with **pos** and **neg**. The criteria is: If *ALTERNATIVE* is *T0* then WhetherT0 is pos. If it is not then WhetherT0 is neg.
- Separate the whole dataset into training and testing datasets. The size is 3:1.

The Decision Tree

- Use Rpart package in R to build the tree and find the best CP to prune the tree.
- According to the tree, we can know three kinds of people will buy T0.
- **Group 1**: Decile < 1.5; Children >= 0.5; Age <51.5; Age >= 39.5; HH size <6.5;
- **Group 2**: Decile < 1.5; Children >= 0.5; Age <57.5; Age >= 51.5; HH size <6.5; Adults >=2.5;
- **Group 3**: Deciel <1.5; Children <0.5; HH size >=2.5; Income >=3000; Zone number >=7.5;



Size of Prize: Who will buy? – Analysis of Decision Tree

Analysis of the Tree

- Confusion Matrix shows the distribution of predicted values and actual values.
- The accuracy of our prediction is **60%.** The 95% confidence interval is (0.5779, 0.6137).
- The true positive rate (Sensitivity) is 0.18190.
- The true negative rate (Specificity) is 0.86153

Confusion Matrix and Statistics

		Actual Situation	
		Neg(won't buy)	Positive(will buy)
Predicted Situation	Neg(won't buy)	1543	940
	Positive(will buy)	248	209

* The predicted value is calculated from test dataset

Bias Analysis of the Tree

- The decision tree model is built on the joint dataset of insider and survey data, where people in the dataset are all aware of itemX. Furthermore, we divide the dataset into training and testing sets. Thus, the size of training set is 5880, which is smaller than the original set.
- The model has its own accuracy rate, which is 60% in this case. This is because the model consider every root as only positive or negative. Thus, it could influence our prediction.

Next step: estimate the quantity

- To predict the specific quantity of itemX using predictive model, we create a dataset includes only people will buy itemX using demographic attributes of three groups of people based on the insider dataset.
- The size of this dataset is **2305**.
- We predict that the percent of buying itemX if all people are aware of it is 2305/18850 = **12.2%**

Size of Price: How many units each customer will buy in the following thirteen months? -- Predictive Modeling

Data Preparation

1. Insider join custdata by unique identifier for the household(HH).
2. Create a new variable named "unitbought(the sum of Units and Units_post).
3. Using 10 K-Folds to create training datasets and validation datasets

Build Predictive Models

1. **Linear Regression model:** choose 12 transformations and interactions of variables to build over 8000 models.
2. **MARS model:** Use different transformations of variables, trace, thresh and degree to build 11 models.
3. **Neural Networks model:** choose different variables and size to build over 8000 models in total.

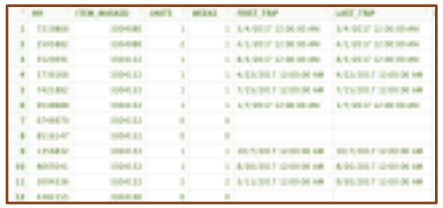
Best Model Type: MARS

Use the buying-customer dataset generated by decision tree to predict the units each customers will buy in the following 13 months.

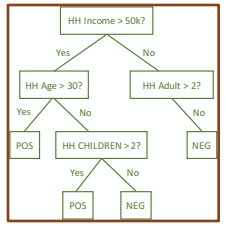
Cross Validation

Model Type	Best Model	MSE
Linear Regression	unitbought~DECILE+factor(ZONE_NBR)+HOH_AGE+HH_INCOME+HH_SIZE+HH_ADULTS+HH_CHILDREN	5.435854
MARS	**Basic fit: unitbought~log(DECILE)+log(ZONE_NBR)+log(HOH_AGE)+log(HH_INCOME)+log(HH_SIZE)+log(HH_ADULTS)+log(HH_CHILDREN)**	**5.411643**
Neural Networks	unitbought~DECILE+ZONE_NBR+HOH_AGE+HOH_AGE+HH_INCOME+HH_SIZE+HH_ADULTS+HH_CHILDREN, size=2,maxit=10000)	5.462437

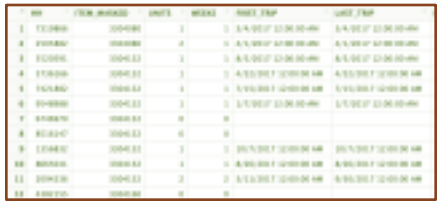
Size of Prize: How many weekly units we sell?



input



output



input

MA
RS

output

Total customers data from Loyalty Card Club Database

The Decision Tree

A vector of predicted Household who will purchase

Predictive model

Weekly units sales = Sum(Q(HH1) +Q(HH2)+Q(HH3)+……+Q(HHn))/(13 months*4 weeks)

Weekly Units Sales:

217.7568

Bias Analysis

1. **Sample bias:** According to the chi-square test on the right, it shows Insider data is similar with othertop30 data. In other words, customers in our model is also high-spending group of people. Thus, we could overestimate the weekly sales.
2. **Method bias:** the criteria of choosing predictive model is mean square error. We cannot avoid error but only try to make it smaller. Thus, predictive model also has bias.

Chi-Square Test

Chi-square test:
Insider spending ~ othertop30 spending
P-value: 0
Insider data is similar with othertop30 data.

Thank you!

Please refer to our appendix

in case of any questions

Appendix

Part1: Is attribute “M” a major barrier to purchase?

Distribution of all customers’ concern	
All customers in survey	9659
Attribute	% of customers’ concern
C	(1692/9659) * 100% = 17.56%
M	(682/9659) * 100% = 7.06%
V	(545/9659) * 100% = 1.18%
E	(113/9659) * 100% = 5.64%

Distribution of never-purchasing customers’ concern	
Never-purchasing customers in survey	2981
Attribute	% of never-purchasing customers’ concern
C	(1692/2981) * 100% = 56.9%
M	(682/2981) * 100% = 22.9%
V	(545/2981) * 100% = 18.3%
E	(113/2981) * 100% = 3.8%

Appendix

Part1: Is attribute “M” a major barrier to purchase?

92% customers who have concern with attribute “M” will never purchase

- Customers who have concern with attribute “M” = 628
- Customers who have concern with attribute “M” will buy (Alternative=“T0 & UNITS_POST != 0) = 51
- Customers who have concern with ”M” will never purchase = 628-51 = 577
- % of customers who have concern with attribute “M” will never purchase =(577/628) * 100% = 92%

Customers who have concern with attribute “M” will never purchase
- Segment by “Tried”, “Aware” and “Unaware”

- **Tried customers:** who have tried ITEM-X before (EndPoint= “E2”) =63 (11% of never-purchasing customers who concern with “M”)
- **Aware customers**: who haven’t tried before but aware of ITEM-X (EndPoint = “E4”) = 254 (44% of never-purchasing customers who concern with “M”)
- **Unaware customers:** who haven’t tried before and aware of ITEM-X (EndPoint = “E6”) = 260 (45% of never-purchasing customers who concern with “M”)

Appendix

Part1: Is attribute "M" a major barrier to purchase?

10% never-purchasing customers who have concern with attribute "M" will be persuaded by marketing messages

- Never-purchasing customers who have concern with attribute "M" and receive the message (receive the message: follow-up question has been answered)= 377
- Persuaded customers (Alternative="T0" & UNITS_POST != 0 & (EndPoint = "E2" | "E4" |"E6")) = 37
- % of customers who have concern with attribute "M" will never purchase =(37/377)*100% = 10%

Marketing messages' effectiveness for "Tried", "Aware" and "Unaware" customers

- **Tried customers**
  - Concerned with attribute "M" will never purchase = 31
  - Persuaded by marketing message = 5
  - % of persuaded = 5/31*100% = 16%
- **Aware customers**
  - Concerned with attribute "M" will never purchase = 174
  - Persuaded by marketing message = 22
  - % of persuaded = 22/174*100% = 12%
- **Unaware customers**
  - Concerned with attribute "M" will never purchase = 172
  - Persuaded by marketing message = 10
  - % of persuaded = 10/172*100% = 6%