BEER CONSUMPTION REPORT

Supermarket Scanner Data Analysis

Shuhong Cao *scao48@wisc.edu*
Matt Su *csu32@wisc.edu*
Yihui Fan *fan72@wisc.edu*
Xuefeng Xu *xuefeng.xu@wisc.edu*

Econ 690 Machine Learning Course Project

2018-12-13

ABSTRACT

We use unsupervised learning algorithm, Hierarchal Clustering, combined with beer labels' practical meaning to separate the beer into different types. Moreover, we choose supervised learning methods, decision tree and random forest, to evaluate how different factors influence the beer sales of Miller company in the supermarket chain. We find that the location of stores and the type of beer will definitely influence the sales of beer. The promotion of Miller will also affect some types of beer under some specific conditions. According to what we find, we give a detailed promotion plan for Miller company to implement their promotion strategy.

INTRODUCTION

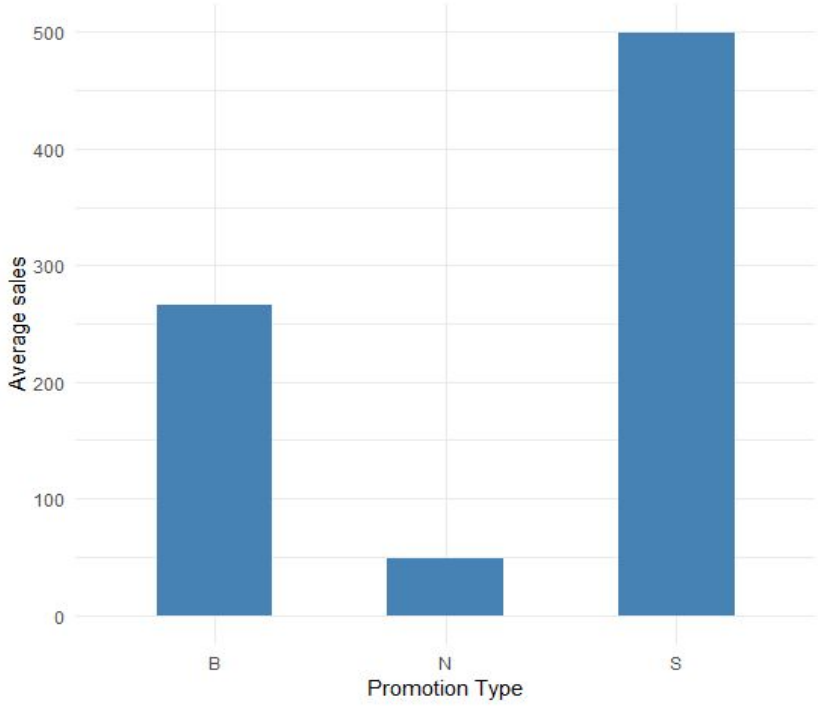
From 1989 to 1994, Chicago Booth and Dominick's Finer Foods coordinated in store-level research about shelf management and pricing, which provided us a categorical data about different goods sold in those two market chains. Among all those products, beer is very likely being influenced by changes of stores' promotion since it is not necessities and people only purchase it on special occasions. We choose the beer industry to analyze the influence of promotion decision to the sales quantity and select the biggest producer in the beer industry, providing a detailed promotion strategy.

Given a large number of types of products and quantities they sold, Miller is the largest beer producer for Chicago Booth and Dominick's Finer Foods from 1989 to 1994. Therefore, we choose miller as our target firm and form a report about the promotion strategy it should implement.

DATA

The data we used comes from Dominick's dataset, designed by James M. Kilts Center, University of Chicago Booth School of Business. There are three tables we use from the

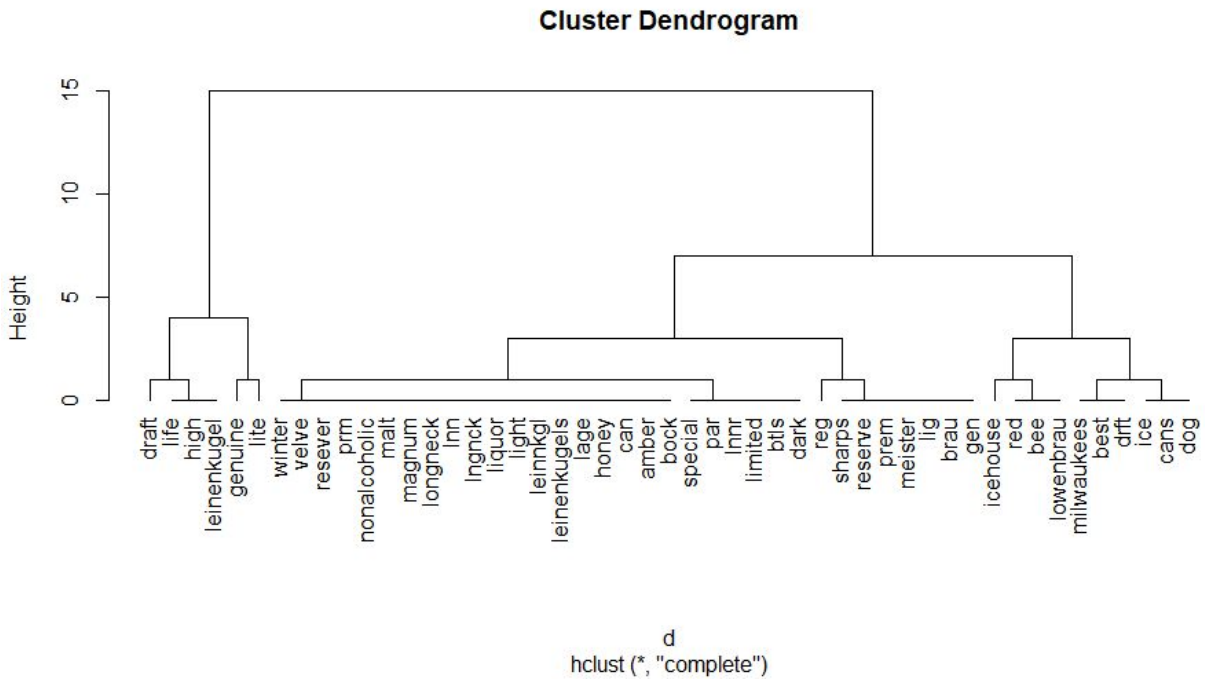
dataset: first is the category-specific movement level data, which contains the sales information at the store level for each upc in a category and is stored on a weekly basis; second is the category-specific upc level data, which contains a description of each UPC in a category; the third is the store-specific store level data, which contains demographics file consists of store-specific demographic data as income, age distribution and car ownership.



We merge all those three tables together using upc code and store number. We could clearly see that using promotion will increase the sales quantity for sure. As the plot shows below, B and S indicate “Bonus Buy” and “price reduction” separately, while N indicates no promotion. The plot shows that with bonus and reduction, the average sales will increase. At the meantime, price reduction seems to have much larger influence compared with bonus.

We summarize the types of miller beer by its description. We first run Hierarchal Clustering and then

according to the results and the real meaning of beer label to decide the final clustering. The word cloud map besides shows the highest frequency descriptions that show on the beer's label. The dendrogram below shows how unsupervised learning classify the description. However, we still need to consider the actual meaning of classification, like *MILLER LITE 22OZ NR* and *MILLER LITE BEER* should all be the same lite beer type. After adjusting the results, we get 15 categories: *Highlife, Lite, Milwaukee, Magnum, Meister, Red dog, Sharp, Leinenkugel, Genuine, Icehouse, Non-alcoholic, Reserve, Lowenbrau, Reg, Lite ice*.

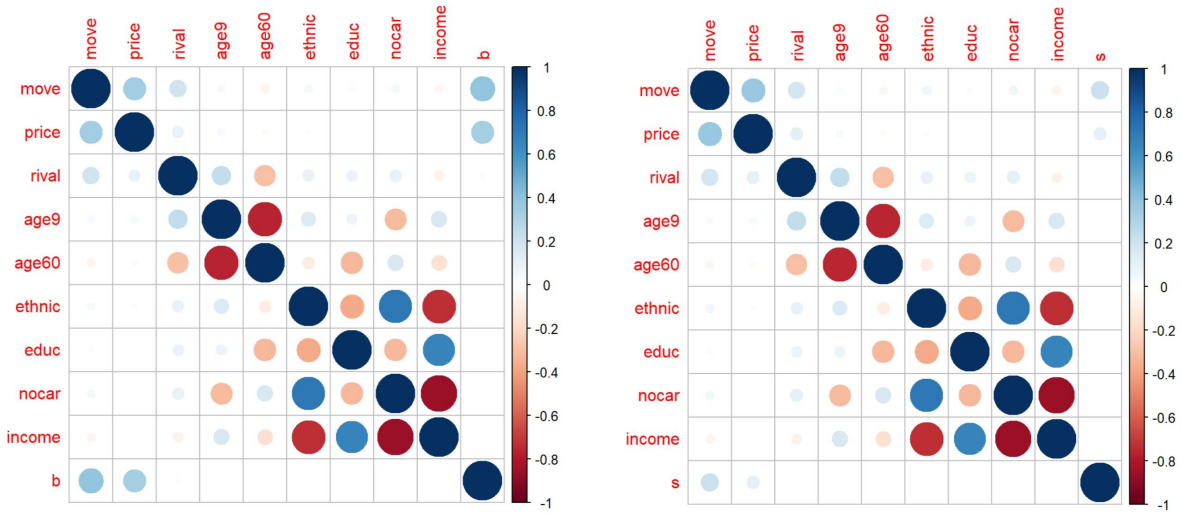


Besides the type of miller beer, we also include other variables and the promotion type. We also choose a rival: Budweiser, who is the second largest producer among those supermarket chains. The sold item of Budweiser may reflex the competition among beer producers. Describe as the table:

Variable	Description	Remarks
move	The Number of beer Sold	The number of sold items in the original table means the package sold. We multiple it with the size of the package
price	The Price of Beer	Package price sold in the store

rival	The Budweiser Sales	The bottles sold by Budweiser
age9	Population under Age 9	Percentage and area specific
age60	Population over Age 60	Percentage and area specific
ethnic	Blacks & Hispanics	Percentage and area specific
educ	College Graduates	Percentage and area specific
nocar	Population With No Vehicles	Percentage and area specific
income	Local Median Income	Use log form and area specific
prom	Promotion Type	Whether the product was sold on a promotion that week
holiday	Holiday week	Whether One or several days of the week are holidays
zone	Store Location group	The geological area that the store located in

Trying to get if both types of promotion have relation with sales, we separate the table into bonus table and price reduction table so we can compare the correlation (i.e we put all non-promotion data and bonus data into a file, the non-promotion data and price reduction data into another file). As the correlation plots showing below: price, rival and two promotion types (*b* and *s* in the table) are highly related to the sales quantity. The other variables could use as control variables.



METHODOLOGY

LEARNING ALGORITHM

Our target is to make a promotion strategy for the beer firm Miller, i.e. to find out the best way to sell their beer based on variables listed above. Besides price which we already know has a close and inverse relationship with quantity, we have no idea about how does other factors such as education attainment and income level affect sales of beer. Hence we need a non-parametric supervised machine learning method. The technique we used is the random forest, which is a robust algorithm because it grows multiple trees rather than a single one, and then merge them together to get a stable and accurate result.

At the very first, we use a decision tree since it provides us a direct and easy-interpreting tree graph with deciding variables on the branches. Decision tree gives us a rough idea about how different variables work in different scenario and influence the final sales. In order to make it easy to compare, we select data from miller's beer consumption and split it into two datasets: all non-promotion data and bonus data into a file, the non-promotion data and price reduction data into another file. With this kind of setup, we could compare the treatment effect of using bonus (B) and price reduction (S), comparing them with each other and decide which one get the largest treatment effect. Decision tree is intuitive but nevertheless easy being affected by variation level. Hence for more robust results, we turn into random forest.

CASE STUDY

We choose five different cases to fit into the random forest model in order to get the predicted promotion outcome for each specific case that maximize the sales. This five cases are representative and very distinct from each other. Exploring those scenarios will give Miller company a detailed promotion plan based on practical situations.

- Case 1 is *miller lite beer* at zone 2 during holiday weeks for the highest income level. The reason is that "miller lite beer" is top one product for selling from the dataset, and as from the map of the data manual, zone 2 is near to the highway 90 and 94. It is clearly to anticipate that those area will more likely to sell more goods even without make any promotion. By calculation, the mean of the total number for holiday weeks under this condition is higher than non-holidays', so that implies more people are willing to come around that area during holidays. Since

the store in this area will more likely sell more goods, so the income level we consider is the highest one among all of the income level, which is 10.92237.

- Case 2 is *miller genuine draft* at zone 1 during holiday weeks for those income level less or equal to average 10.52109. Since zone 1 is closed to University of Chicago and due to the existence of fraternity, the consumption of beer should be considerable, especially for holiday weeks. The store at zone 1 will mainly focus on students, which indicate the income should be a comparingly lower level and the price will be expected not too crazy, which means the price is affordable to student but still can make a certain profit. Since among the mean price of all those five types beer, "miller genuine draft" is the middle one, so it is chosen for sample 2.

- Case 3 is *miller genuine drft* at zone 7 during non holiday weeks for those income level greater or equal to 10.39542. Since zone 7 is very closed to Chicago and has lower consumption level than expect, zone 7 would be interesting to figure out what kind of promotion should apply. Comparing mean prices of holiday weeks and non holiday weeks, the difference between them is smallest among the sample we have above. Since it's commonly assumed that the cost of metropolis is higher than rural area, the income we choose will be higher than the mean income 10.39542 and the type of beer we choose, "miller genuine drft", will relatively has the highest mean price.

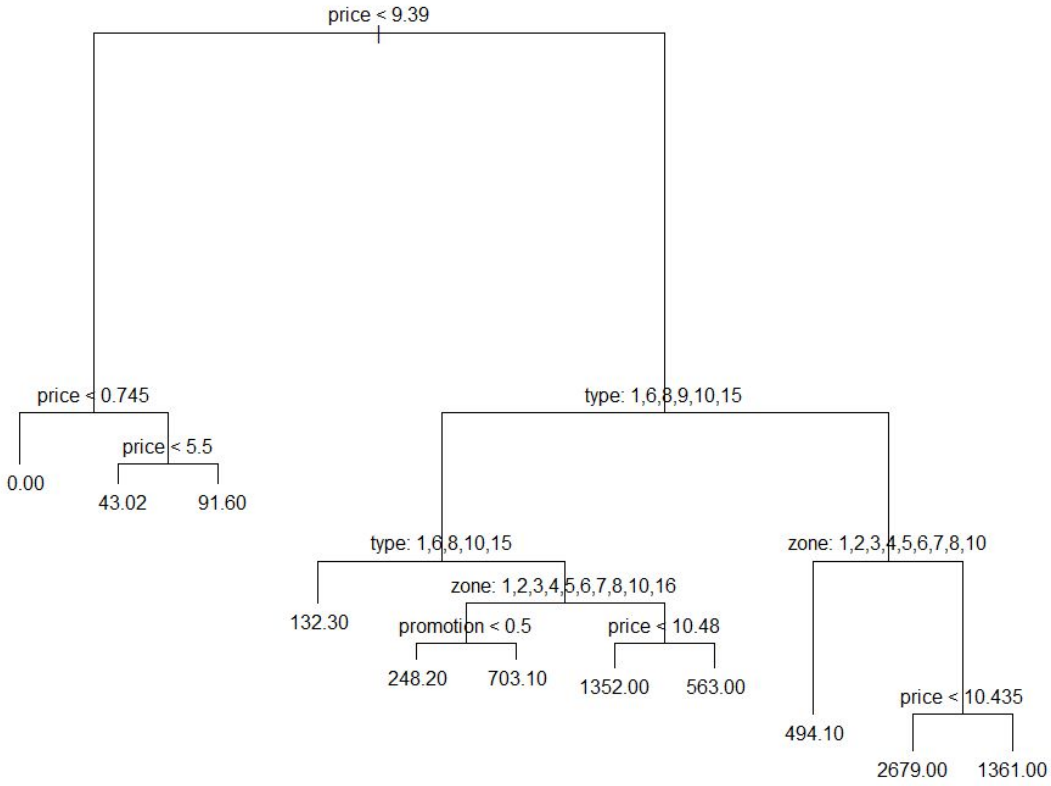
- Case 4 is *miller gen drft lnnr* at zone 15 during non holiday weeks for those income level less than or equal to 10.56658. Since zone 15 is the area that is far away from Chicago and the main area of the stores, it can be counted as rural area or area without too much people live in or travel around. The reason for choosing "miller gen drft lnnr" is that it has not only the lowest mean price but also the lowest selling among those five products. And since we treat zone 15 as rural area, the income level should less than the mean level of this area and the selling should not very too much between holiday weeks and non holiday weeks.

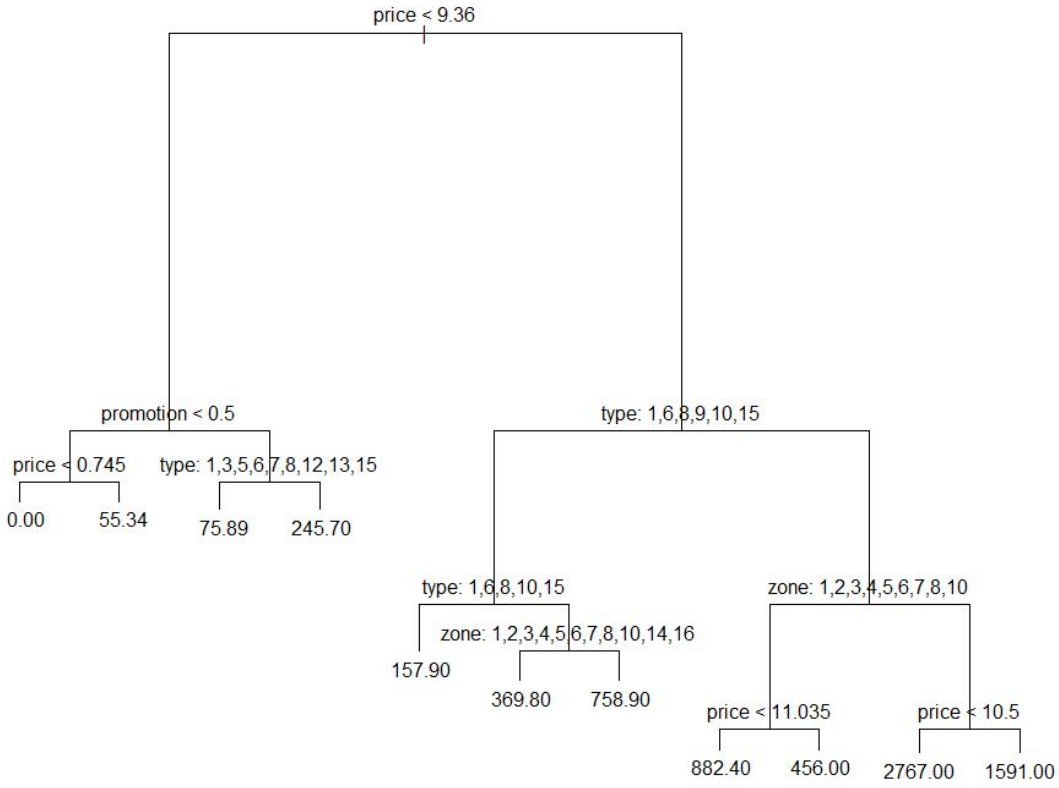
- Case 5 is *miller genuine draft* at zone 15 during non holiday weeks for those income levels are the lowest, 10.49385. Since holiday is insignificant when we regress the total number on holiday, so we choose the one with more observations. From the regression we have before, the zone 15 has the biggest coefficient among 16 zones, so the zone 15 is chosen. The reason for choosing "miller genuine draft" is that comparing the top five products, this one is the second one and relatively close to the top one. By comparing the total number sold of different income levels, we choose the highest income level 10.49385 with the smallest total number sold so that there are chances to improve it.

RESULTS

1.DECISION TREE

The results of the best-fit tree we get is showing below: the first tree is bonus treatment effect and the second tree is price reduction treatment effect. The efficacy of each decision is decided by the number of beer was sold (*sales*).





Firstly, we noticed there are some interesting points here:

- The type of product does affect the sales. We have 15 types of beer and it turns out that certain kinds of beer are very popular in the supermarket. We could list all the types we select:

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Highlife	Lite	Milwaukee	Magnum	Meister	Red Dog	Sharp	Leinenkugel
9	**10**	**11**	**12**	**13**	**14**	**15**	
Genuine	Ice House	Reg	Reserve	LowenBrau	Non-alcoholic	Lite Ice	

- Surprisingly, the time of the weeks does not influence the beer consumption according to the tree. We predict in some holiday week people are more likely to buy more beer. However, it shows no sign that holiday could influence the beer consumption. We guess people in this area are maybe more into using other

alcoholic drink to celebrate and keep the beer consumption the same.

- The sales of beer are influenced by the location of the store. In some area, people are willing to buy more beers and also more easily to get influenced by promotions. Stores have been split into different zones according to their geological locations. Although visible factors like income and age distribution seem to have limited impact on sales, other factors related to location zone do have an impact even though we could not separate them into variables.

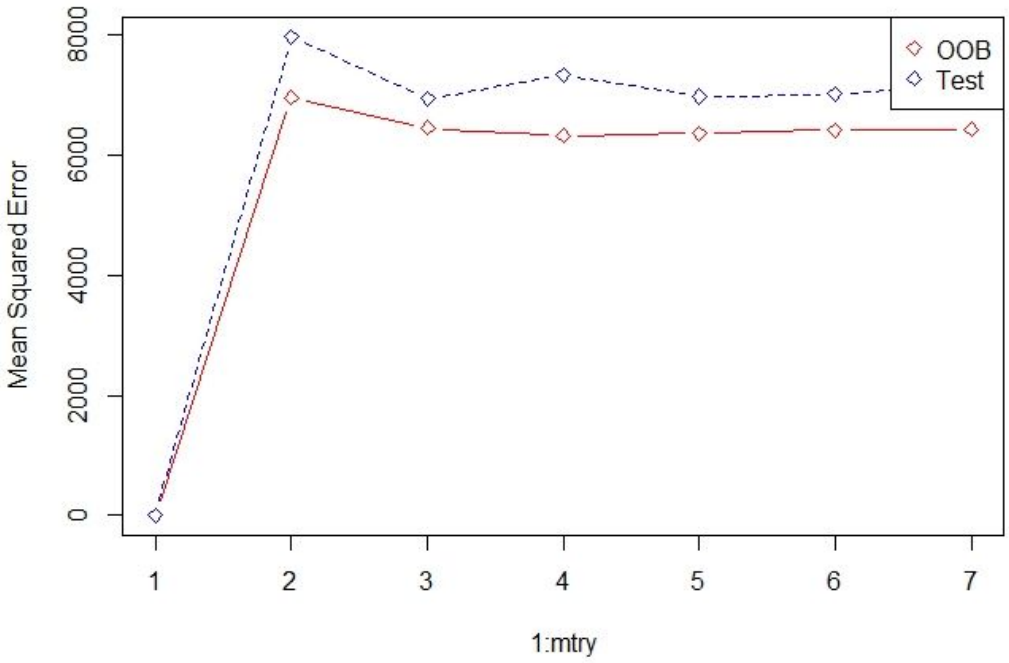
It shows in some branches of our tree, the promotion will not have significant influences on the beer sales. However, for beer types of Highlife, Red Dog, Leinenkugel, Ice House, Lite Ice in zones of 1, 2, 3, 4, 5, 6, 7, 8, 10, 16, using bonus promotion will increase the beer sales for (703 - 248 =) 455 bottles. For beer price is lower than $9.39, using price reduction will also increase the beer consumption, especially for Highlife, Milwaukee, Meister, Red Dog, Sharp, Leinenkugel, Reserve, LowenBrau, Lite Ice, increasing (245 - 55 =) 190 bottles. Even for beer not in those types, the price reduction will increase the sales for (76 - 55 =) 21 bottles.

2. RANDOM FOREST

By adding variables one by one in a random forest test and then checking the lowest MSE of all models, we reselect our final variables for our random forest: promotion type *(prom)*, 6 classes of size *(size_class)*, population under age 9 *(age9)*, college graduates *(educ)*, containing holiday in a week *(holiday)*, location *(zone)*, 5 best selling types *(best_type)*. We excluded price here because it is believed to be influenced by promotion type and size of the product. We replaced type with *best_type* since the best selling 5 types which almost take up 75% of the whole sales. For the relation of education level and alcohol consumption, we find that the education level will negatively affect the alcohol consumption (Crum and Helzer, 1993), which indicates the variable education should be considered as one of the main final variables. Although there is no direct articles can prove the relation between number of kids and amount of alcohol consumption, the common sense is that kids cannot drink until 21, which implies the alcohol consumption should decrease comparing with those area with smaller population under age 9.

After selecting variables we started to train our random forest. The most important parameter we here concerned about is *m*, meaning that each time the tree comes to split a node , *m* variables would be selected randomly and then one of those *m* variables

would be selected in the next node. To find out which *m* we should use, we tried all the possible values:



Result shows that the model has the lowest MSE when m is 3 or 4.

We created 3 separate random forest for different promotion types: non-promotion, bonus promotion and price reduction promotion. In this case, we can compare the different outcomes under among promotion types with the same control variables. Growing 100 trees, the non-promotion random forest has the mean of squared residuals 8219.778 and about 69% of the variance has been explained. The random forest of Bonus promotion has the mean of squared residuals 105861 and about 65% of variance is explained. The random forest of price reduction promotion has MSE 251087 and 45% of the variance is explained. The size (number of bottles per unit), the type of beer and the zone are three most important features in random forest.

Finishing building model, we put in our five cases and see for each case, what kind of promotion could facilitate the largest sales. Results are shown in conclusion part.

CONCLUSION

We use unsupervised learning algorithm, Hierarchal Clustering, combined with beer labels' practical meaning to separate the beer into different types. Moreover, we choose decision tree and random forest to evaluate how different factors influence the beer sales of Miller company in supermarket chains. We find that the location of store and the type of beer will definitely influence the sales of beer. The promotion of Miller will also affect some types of beer in some specific conditions.

Finally, given by the outcome from our random forest model, we suggest Miller company implement the following promotion strategy for each case:

- For case 1: Sell miller lite beer of 24 units in a package in store 5 on holidays using reduction.
- For case 2: Sell miller genuine draft of 24 units in a package in store 68, 71, 93, 95, 111, 123, 124, 130 on non-holidays using reduction.
- For case 3: Sell miller genuine draft of 12 or 24 units in a package in store 53, 109 on non-holidays using reduction.
- For case 4: Sell miller genuine draft of 1 or 8 or 12 or 24 or 30 units in each package in store 102, 103 on non-holidays using reduction.
- For case 5: Sell miller genuine draft lnnr of 1 or 8 or 12 or 30 unites in each package in store 102 on non-holidays using reduction.

type	holiday	zone	income	size_ class	reduct ion	bonus	no_pro motion	Choice
MILLER LITE BEER	1	2	11.66	24	829	633	273	S
MILLER GENUINE DRAFT	1	1	10.05	24	645	561	75	S
MILLER GENUINE DRAFT	0	7	10.56608	12	708	319	103	S
MILLER GENUINE DRAFT	0	7	10.56608	24	730	471	79	S
MILLER GENUINE DRFT	0	15	10.90857	1	545	249	8	S
MILLER GENUINE	0	15	10.90857	8	545	249	101	S

DRFT								
MILLER GENUINE DRFT	0	15	10.90857	12	514	183	111	S
MILLER GENUINE DRFT	0	15	10.90857	24	567	319	439	S
MILLER GENUINE DRFT	0	15	10.90857	30	545	249	378	S
MILLER GEN DRFT LNNR	0	4	10.06608	1	579	373	22	S
MILLER GEN DRFT LNNR	0	4	10.06608	8	579	373	72	S
MILLER GEN DRFT LNNR	0	4	10.06608	12	585	472	73	S
MILLER GEN DRFT LNNR	0	4	10.06608	30	579	373	168	S

APPENDIX

1. Code and other materials are included in a separate file.
2. Crum, R M and Helzer, J E and Anthony, J C. Level of education and alcohol abuse and dependence in adulthood: a further inquiry. *Am J Public Health*. 1993; 83:830 - 837