STAT 406 - Group Project

Relationship between Amount Spent on Wine and Other Aspects

Jiapan Wang #29322674 Hanliang Liu #86009776

Catherine Cai #52204310 Jack Guo #57752750

Introduction

In the last 70 years, the development of economics has recovered from the second World War. The dramatic increase in this period was never observed in history before due to the rapid development of industry and globalization. Meanwhile, the consumption power of citizens has increased with the development of economics. Therefore, we are interested in finding which set of factors has a great impact on the consumption power of customers, and in this study, we will focus on the amount spent on wine, which represents the consumption power of customers. Our study is observational in nature. It is unfortunately not possible to randomly allocate experiment groups where some countries allocate more resources on certain factors while other countries do not.

Two months ago, there was a tragic car accident near UBC campus late at night, which might have been caused by a possibly drunk driver. This dataset provides the amount spent on wine by customers and other aspects, for example, Birth Year and the Income of customers. It would give help to know the relationship between predictor variables and purchasing power in future careers.

We have chosen the "Customer Personality Analysis" hosted on Kaggle.com. This dataset is a detailed analysis of a company's ideal customers. It helps a business to better understand its customers and makes it easier for them to modify products according to the specific needs, behaviours and concerns of different types of customers. The response variable chosen from this dataset is MntWines, the amount spent on wine in the last 2 years, which is measured in dollars.

Exploratory Data Analysis

In order to have a better understanding of the data, we started by briefly looking at the whole dataset, which has 2240 observations with 29 variables in total. Here is a head of the dataset below:

	ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	Recency	MntWines	MntFruits	MntMeatProducts
1	5524	1957	Graduation	Single	58138	0	0	04-09-2012	58	635	88	546
2	2174	1954	Graduation	Single	46344	1	1	08-03-2014	38	11	1	6
3	4141	1965	Graduation	Together	71613	0	0	21-08-2013	26	426	49	127
4	6182	1984	Graduation	Together	26646	1	0	10-02-2014	26	11	4	20
5	5324	1981	PhD	Married	58293	1	0	19-01-2014	94	173	43	118
6	7446	1967	Master	Together	62513	0	1	09-09-2013	16	520	42	98

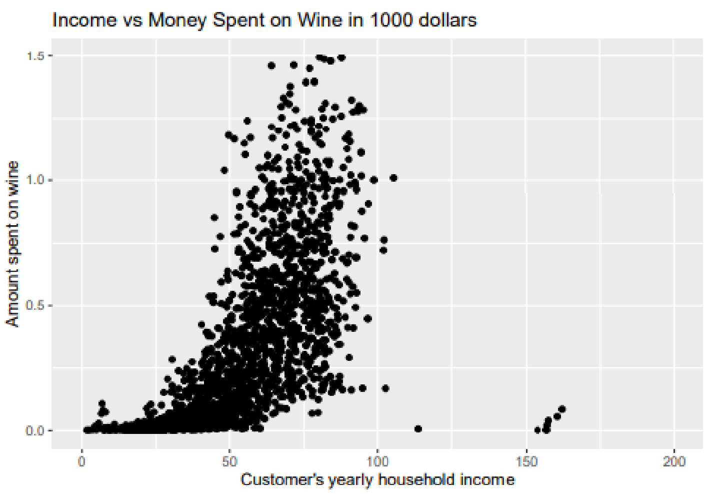
	MntFishProducts	MntSweetProducts	MntGoldProds	NumDealsPurchases	NumWebPurchases	NumCatalogPurchases	NumStorePurchases
1	172	88	88	3	8	10	4
2	2	1	6	2	1	1	2
3	111	21	42	1	8	2	10
4	10	3	5	2	2	0	4
5	46	27	15	5	5	3	6
6	0	42	14	2	6	4	10

	NumWebVisitsMonth	AcceptedCmp3	AcceptedCmp4	AcceptedCmp5	AcceptedCmp1	AcceptedCmp2	Complain	Z_CostContact	Z_Revenue	Response
1	7	0	0	0	0	0	0	3	11	1
2	5	0	0	0	0	0	0	3	11	0
3	4	0	0	0	0	0	0	3	11	0
4	6	0	0	0	0	0	0	3	11	0
5	5	0	0	0	0	0	0	3	11	0
6	6	0	0	0	0	0	0	3	11	0

We explored the relationship between the response variable MntWines, which indicates the amount spent on wine in the last 2 years, and the other 8 explanatory variables. The reasons why we chose these variables are below.

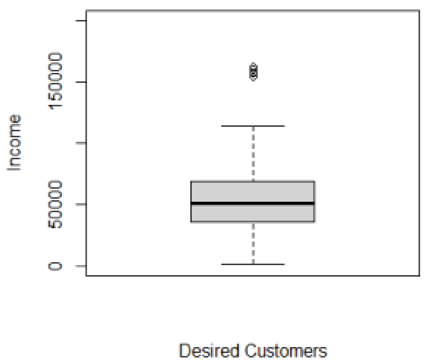
Since the response variable is MntWines, we intuitively considered the consumer's personal information, such as income, education level, etc., and the consumer's family information, such as family size, marital status, etc. In addition, the consumption of other items, like fruits, meat, fish, etc., may have an impact on our response variable, and we also reserved them. We have discarded the remaining variables since they respond to the employment situation of consumers and different companies, which are related to the target of the original report source but not within the scope of our research.

To start with, we plotted MntWines against every single explanatory variable, and finally reserved two graphs as follows, which have good interpretability:

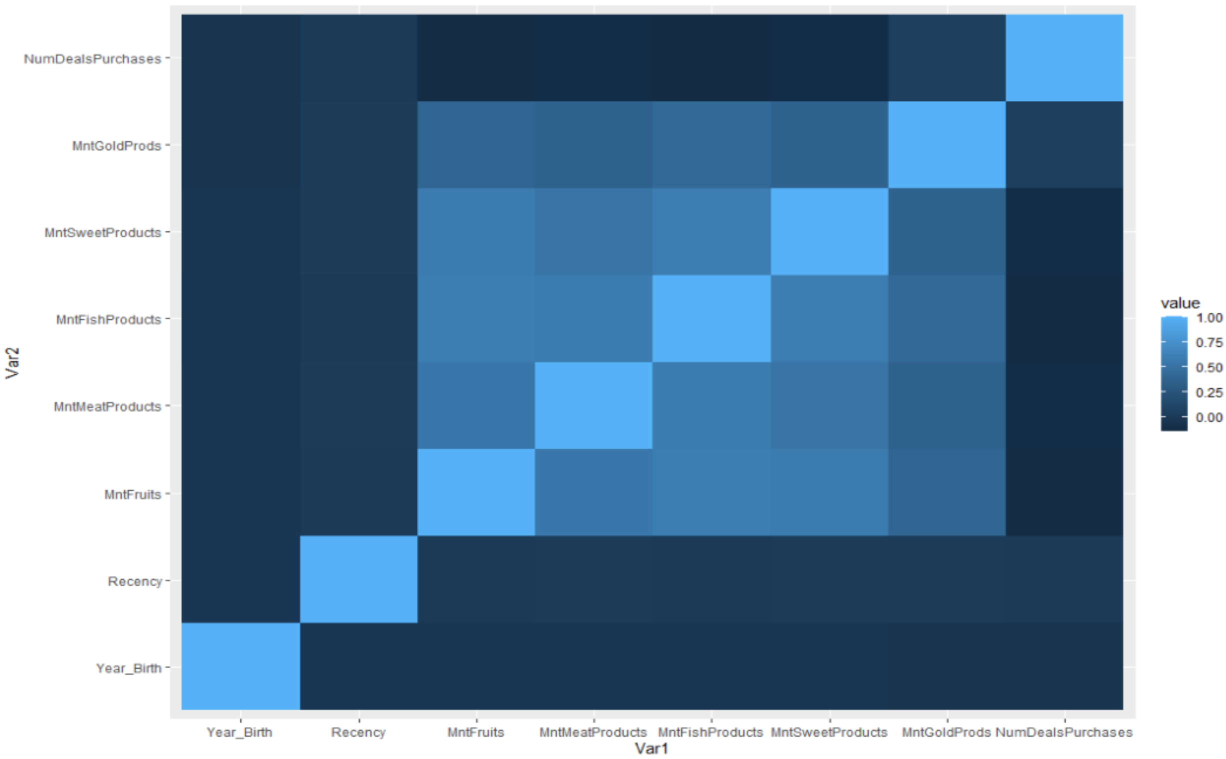


As can be seen, the plot above describes that MntWines is non-linearly positively correlated with customers' income. The plot below illustrates MntWines seems to have a linear relation with money spent on meat products.

Below is a rough plot of customers' income, and we would say customers' income is roughly evenly distributed.



Here is a correlation heatmap for the variables below. The lighter the color of the block, the larger the correlation of its corresponding variables. As can be seen, variables that measure the amounts spent on different products have a higher correlation than others, especially for fish and meat products, which is plausibly reasonable.



Also, the potential risk are represented by Cp and AIC in different models below:

```
##
##                                Stepwise Selection Summary
## ---------------------------------------------------------------------------------------------
##                                  Added/                   Adj.
## Step        Variable            Removed     R-Square    R-Square     C(p)         AIC          RMSE
## ---------------------------------------------------------------------------------------------
##    1         Income             addition       0.335       0.335    782.3200    31189.1438    275.1787
##    2    MntMeatProducts         addition       0.416       0.415    420.7530    30904.2891    257.9906
##    3         Kidhome            addition       0.460       0.459    224.5120    30732.6756    248.1359
##    4        Education           addition       0.481       0.479    131.0650    30651.9723    243.4392
##    5      MntGoldProds          addition       0.501       0.499     42.8000    30566.6858    238.7457
##    6    NumDealsPurchases       addition       0.511       0.509      1.4850    30525.5484    236.4868
##    7    MntSweetProducts        addition       0.511       0.509      0.0260    30524.0667    236.3547
## ---------------------------------------------------------------------------------------------
```

According to the result from the graph above we can get the risk of different possible models from subsets regression.

Analyses and Results

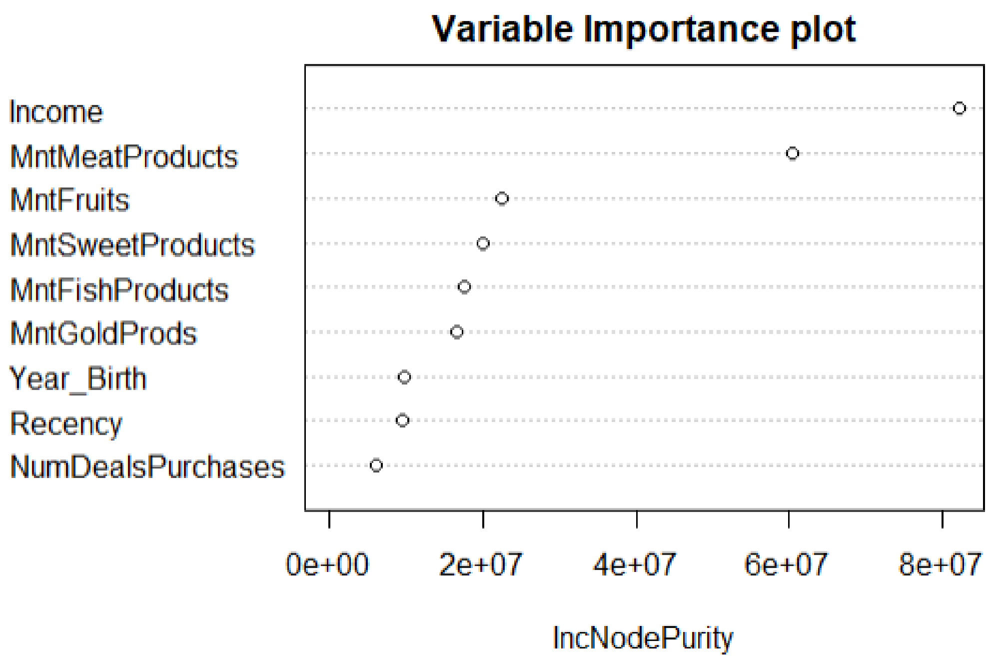
Random Forest

Random Forest is a supervised method for Machine Learning, which is commonly used to solve classification issues. It creates decision trees from various samples, using the majority vote for classification.

We applied a random forest method, and obtained that Income and MntMeatProducts are most useful on the Variable Importance plot, which is consistent with the analysis above.

```
Call:
 randomForest(formula = MntWines ~ ., data = newwine, ntree = ntrees,
type = classification, na.action = na.omit)
               Type of random forest: regression
                     Number of trees: 200
No. of variables tried at each split: 3

          Mean of squared residuals: 35092.96
                    % Var explained: 69.15
```



Model Selection / Evaluation

Next, we applied a subset model selection method and obtained the results below:

Best Subsets Regression

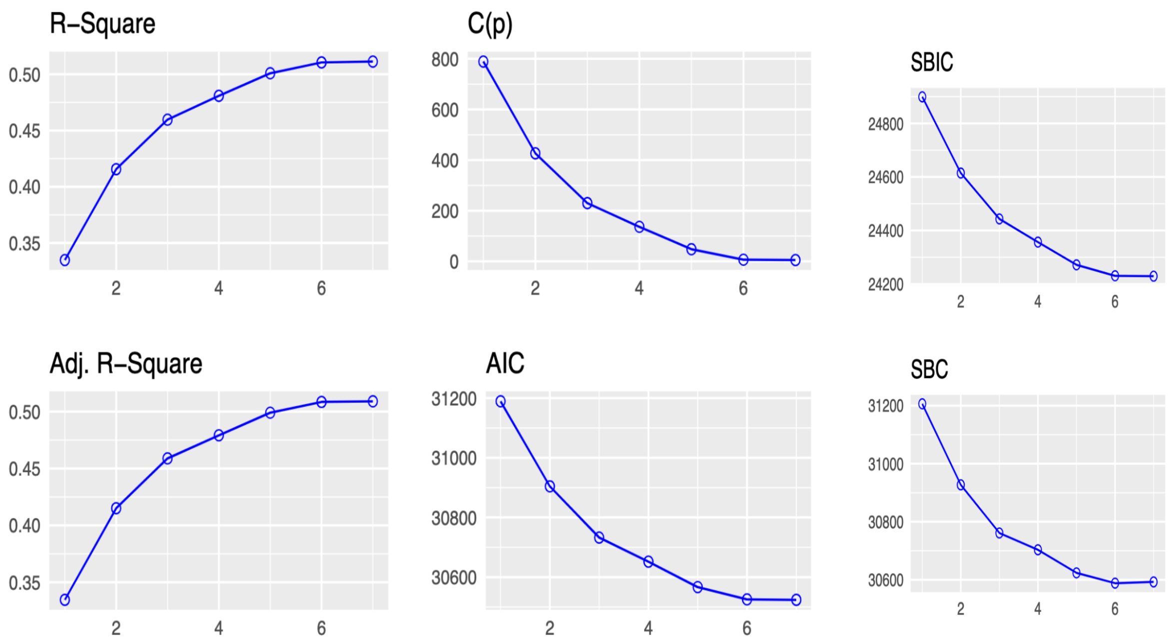
Model Index	Predictors
1	Income
2	Income MntMeatProducts
3	Income MntMeatProducts Kidhome
4	Income MntMeatProducts Kidhome Education
5	Income MntMeatProducts Kidhome Education MntGoldProds
6	Income MntMeatProducts Kidhome Education MntGoldProds NumDealsPurchases
7	Income MntMeatProducts Kidhome Education MntGoldProds NumDealsPurchases MntSweetProducts

Subsets Regression Summary

Model	R-Square	Adj. R-Square	Pred R-Square	C(p)	AIC	SBIC	SBC	MSEP
1	0.3348	0.3345	0.2447	789.0891	31189.1438	24899.2249	31206.2541	167802910.1422
2	0.4156	0.4151	0.3264	426.7002	30904.2891	24614.5453	30927.1030	147495118.0245
3	0.4596	0.4589	0.3881	230.0115	30732.6756	24443.1949	30761.1929	136442296.1437
4	0.4808	0.4792	0.413	136.3489	30651.9723	24356.6915	30703.3035	131147893.5451
5	0.5009	0.4991	0.4399	47.8790	30566.6858	24271.7585	30623.7204	126139572.3819
6	0.5105	0.5085	0.4462	6.4671	30525.5484	24230.8605	30588.2864	123763867.1762
7	0.5113	0.5091	0.4447	5.0000	30524.0667	24229.4109	30592.5083	123625559.9795

AIC: Akaike Information Criteria
SBIC: Sawa's Bayesian Information Criteria
SBC: Schwarz Bayesian Criteria
MSEP: Estimated error of prediction, assuming multivariate normality

From the EDA section, we found that the "Income" explanatory variable has the largest prediction risk since it has the largest Cp and AIC and "MntSweetProducts" has the smallest prediction risk since it has the lowest Cp and AIC in different possible models. Taking the above analysis results into consideration, a combination of Income and MntMeatProducts is the best model at present. In order to represent the risk of different models, we plotted R-square, Cp, AIC, SBC and SBIC.



From 6 graphs above, we conclude that the model with explanatory variables of Income, MntMeatProducts, KidHome, Education, MntGoldProds, NumDealsPurchases and MntSweetProducts, is the best choice due to its low Cp and AIC.

Discussion

From our methods in the study, we built several models to predict the amount spent on wine based on variables. The model with the lowest risk(represented by Cp and AIC) and the model we got from the random forest are significantly different. However, both of these two strategies have their own strengths and limitations.

Once the assumptions of AIC (or AICc) have been met, the biggest advantage of using AIC/AICc is that our models do not need to be nested for the analysis to be valid, unlike other single-number measurements of model fit like the likelihood-ratio test. AIC is low for models with high log-likelihoods (the model fits the data better, which is what we want), but adds a penalty term for models with higher parameter complexity, since more parameters mean a model is more likely to overfit to the training data. Nevertheless, the main limitation of AIC is that the AIC makes assumptions that we are using the same data between models and we are measuring the same outcome variable between models, and have a sample of infinite size. The last assumption implies that the sample size in our study might cause huge errors.

Compared to models predicted by AIC, Random Forest is based on the bagging algorithm and uses Ensemble Learning technique. It creates as many trees on the subset of the data and combines the output of all the trees. In this way, it reduces overfitting problems in decision trees and also reduces the variance and therefore improves the accuracy. Random Forest can be used to solve both classifications as well as regression problems, which is more generally used than AIC. However, the main disadvantage of Random Forest is that it creates a lot of trees (unlike only one tree in case of decision tree) and combines their outputs. By default, it creates 100 trees in the Python sklearn library. To do so, this algorithm requires much more computational power and resources. On the other hand, the decision tree is simple and does not require so many computational resources.

Conclusion

The models from two methods almost have the same error . To improve the accuracy in the future analysis, we may try using the K-fold validation method to split the data set. Moreover, we may need more experiment units or consider more variables into consideration. Since there are much more limitations when it comes to real life problems. For the random forest, we have only a few explanatory variables, and it would be more accurate for prediction if using more variables. For the AIC/Cp model, we ignore the correlation between the variables, which may lead to accuracy loss.