|  |  |
| --- | --- |
| Superstore Data Analysis  Visualization & Prediction Analysis | Abstract  *The article analyzes key factors that affect the performance of Superstore by both visualizations & machine learning algorithms.*  For Prof. Hamidreza Ahady Dolatsara  By TingLei Ruan ,Abbas Ali  BAN5573 Visual Analytics |

Table of Contents

[1. Abstract: 2](#_Toc88394699)

[2. Motivation: 2](#_Toc88394700)

[3. Topic & Project Scope: 2](#_Toc88394701)

[4. Research Question & Significances of study: 2](#_Toc88394702)

[5. Literature Review: 3](#_Toc88394703)

[6. Visualization 3](#_Toc88394704)

[6.1 Sales Dashboard 4](#_Toc88394705)

[6.2 Profit Dashboard 5](#_Toc88394706)

[6.3 Discount Dashboard 6](#_Toc88394707)

[6.4 Shipment Dashboard 7](#_Toc88394708)

[6.5 Sales-Profit -Discount Dashboard 8](#_Toc88394709)

[7. Predictive Analysis 9](#_Toc88394710)

[7.1 Variable Engineering 9](#_Toc88394711)

[7.2 Categorical Variable Encoding 9](#_Toc88394712)

[7.3 Regression 10](#_Toc88394713)

[7.3.1 Decision Tree 10](#_Toc88394714)

[7.3.2 Random Forest 12](#_Toc88394715)

[7.4 Classification 12](#_Toc88394716)

[7.4.1 Binomial Logistic Regression 12](#_Toc88394717)

[7.4.2 Supply Vector Machine for Multi-class Classification 13](#_Toc88394718)

[8. Conclusion 14](#_Toc88394719)

[9. Reference 15](#_Toc88394720)

[10. Appendix 16](#_Toc88394721)

[10.1 R Script 16](#_Toc88394722)

# 1. Abstract:

A superstore is a vast online supermarket in the USA, often selling household goods, clothes, electrical goods, and food. Superstores typically charge anywhere from 15 to 45 percent less than their smaller counterparts. Through this report, the authors delineate the trends of the dataset by drawing visualizations in Tableau and create a model to predict profitability of future transactions by decision trees, logistic regression etc. in R. Finally, the authors present critical thoughts by observing the various output.

# 2. Motivation:

The project objective is to perform Exploratory Data Analysis using TABLEAU & analyze the data by presenting graphical representation, which provides feedback to the management to better understand their customer segment, state performance, region performance, category wise performance, profitability, etc. that helps the management to increase sales, profitability & above all better decision making.

# 3. Topic & Project Scope:

Exploring store data from various perspectives drawing conclusions that help management make corresponding selling , logistic strategies and improve overall performance. This project will primarily focus on what kinds of commodities are preferable online and customers' geographical distribution. All the results will be presented in logical graphs ,dashboards, and clear explanations will be along with that. After that, it will adapt machine learning algorithms to make a predictive analysis mainly on sales. It will finally illustrate insights with a story and provide some suggestions for the management. The store’s management can answer a variety of questions like deciding which region to focus on, what kinds of commodities to prepare, and how many discounts should be placed etc.

# 4. Research Question & Significances of study:

Analyzing data and getting meaningful insights is the key motivation for this project. Especially in retail stores where they have multiple parameters and factors to be considered & each have an impact on the overall health and profitability of the business.

Our research question & significance of performing this research is to analyze the data in terms of shipment mode, segments, region, category, profitability etc. which will help Superstore management to understand the business in depth which will assist them in taking logical decisions ,increasing sales and profitability, avoid wastage of resources , and more importantly decide the future course of action.

Some of the common questions which will be answered are:

· Which shipment mode contributes the most of profit?

· Which segment performs the best?

· Which region contributes and its sales proportion?

· Which category and subcategory performs the best? & In which region?

· Differences of impact on the dependent variable ‘profit’ amongst independent variables.

# 5. Literature Review:

Swasti (2021), and Xiemin (2021) conducted analysis on similar datasets and presented their process and outcomes. Both did well in defining business problems, adopting Python to gain a basic knowledge of the dataset and preprocessing attributes, creating visualizations in Tableau. In addition, they perfectly answered the business problems, such as which region(s) is more profitable, what category or sub-category sells most, by showing the fancy graphs. However, they have also made mistakes and their analysis were not robust enough. Xiemin found the dataset has 19 duplicates and eliminate all of them. While the truth is that those been assumed duplicated observations are not completely identical; they were just similar to each other. In essence, it is normal to see similar observations because one customer purchases many times, and one item can be sold many times as well. Moreover, though they successfully answered all the business problems they pointed out, viewers will have no clue discerning the difference of impact toward profit amongst variables.

Joshua et.al (2019) compared accuracy of different algorithms in predicting annual earning. Their comparison posed that machine learning really improves the ability of predicting the outcome of a target variable, and random forest provides a better accuracy.

Sunil (2015) in his post introduced ways to deal with categorical variables in predictive modeling. He listed problems that caused by categorical variables, such as too many levels, and the categorical variables couldn’t be fitted into a regression model. He then gave some proven ways which help to tackle with categorical variables like converting the variable into numerical values, combining levels, and adopting dummy coding.

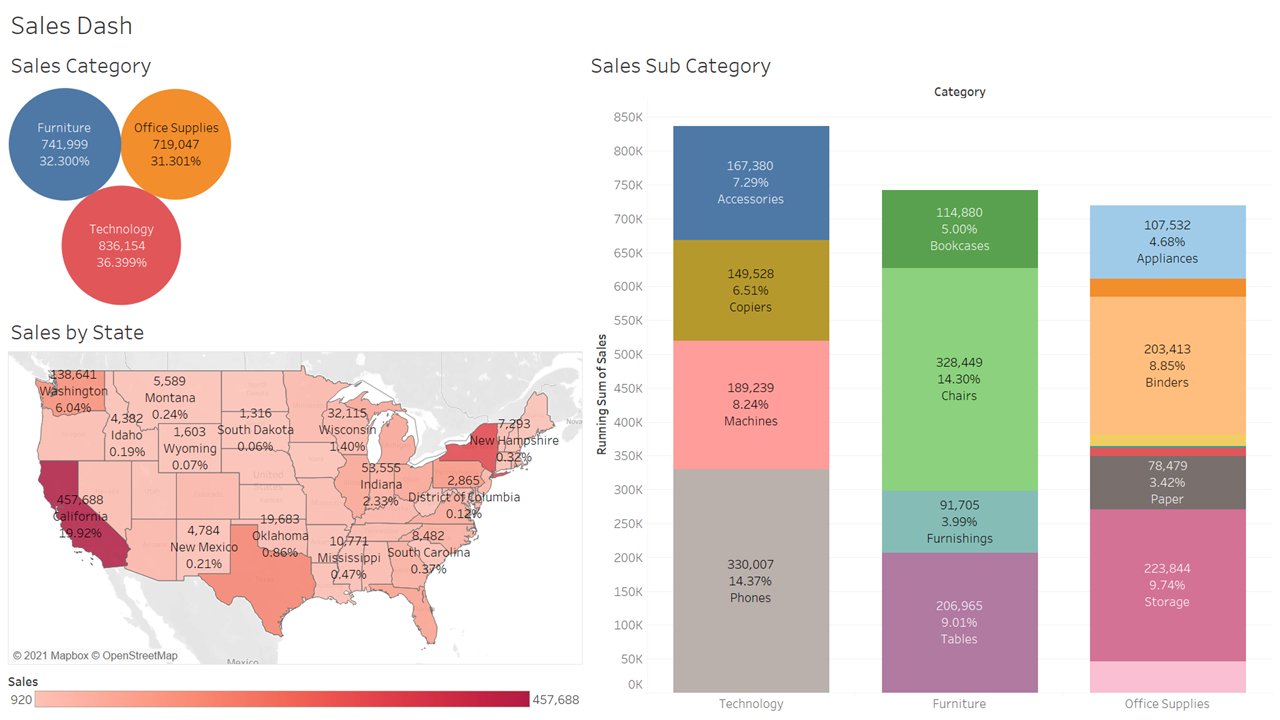
To sum up, some students and scholars have done similar work in analyzing sales of a superstore, but they mostly focused on visualizing variables one by one to find out which value correlates to high sales or profit. Thus, further analysis should pay attention to figure out which variable has a higher impact on sale or profit, and which variable barely effects anything. Since the datasets has many categorical variables like ‘region’, ‘ship mode’, ‘customer segment’, and so on, relying on the principle of dealing with categorical variables is essential for conducting the in-depth machine learning algorithm.

# 6. Visualization

As explained previously the management wants to analyze the data from every angle & perspective to get a quick overview of the business.

To accommodate this request, we have created 5 Dashboards be it Sales, Profit, discount, shipment & lastly all of them together i.e., Sales, profit & discount offered.

## 6.1 Sales Dashboard

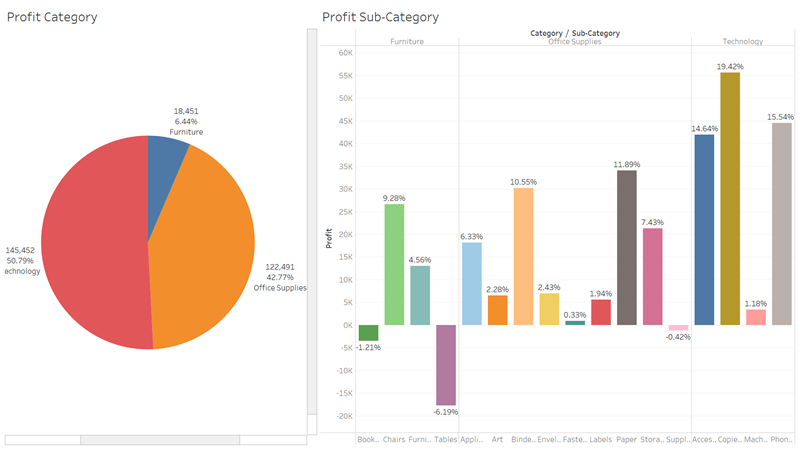


Above Dashboard has 3 parts. As it can be inferred from the diagram sales are nearly having an equal proportion across all categories Furniture, Office Supplies, and Technology with 32.3%, 31.3%, and 36.39% respectively.

Sales by the state have the darker shades reflecting the higher sales and it's evident that California, New York, and Washington are top-selling states with 19.92%, 13.53%, and 6.04% respectively.

When we analyze sales by subcategory, the stack chart gives us a quick overview of how all categories are performing and within those categories how subcategories are performing. Under the technology, category Phones are the best performer, and it is also the leading contributor in the total business with 14.37%. Similarly, the furniture category has chairs with the second-highest contributor in overall business with 14.30%. Office supplier has a few categories like envelopes and labels as the lowest contributor.

## 6.2 Profit Dashboard

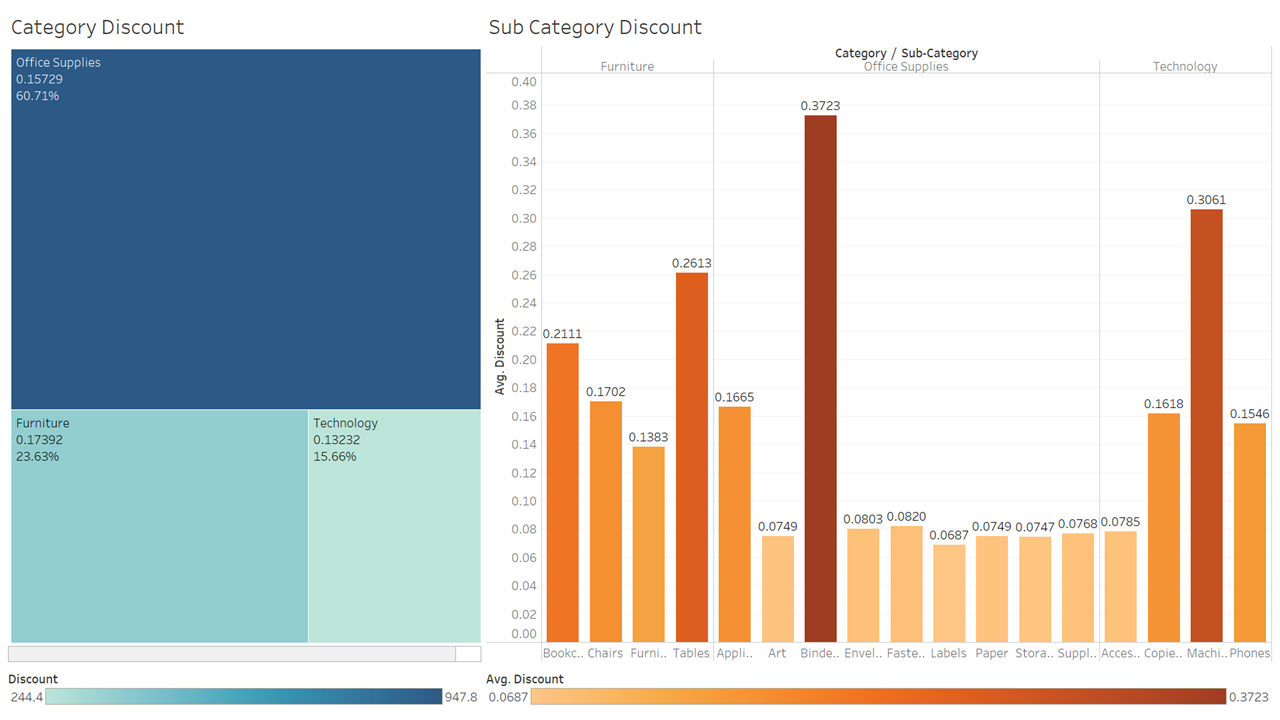


Profit Dashboard being the second in line which also is the major area of management interest, clearly shows that technology is contributing to 50.79% of Superstore's overall business.

Profit by category and subcategory provides us feedback to look deeper into the business of Bookcases and Tables as they are showing a negative profit while chairs are contributing 9.28 % in the overall business.

Binders and Paper are leaders in the office supplies category. Technology has 4 subcategories with all categories doing well but management needs to focus on how to bring the profitability of machines higher.

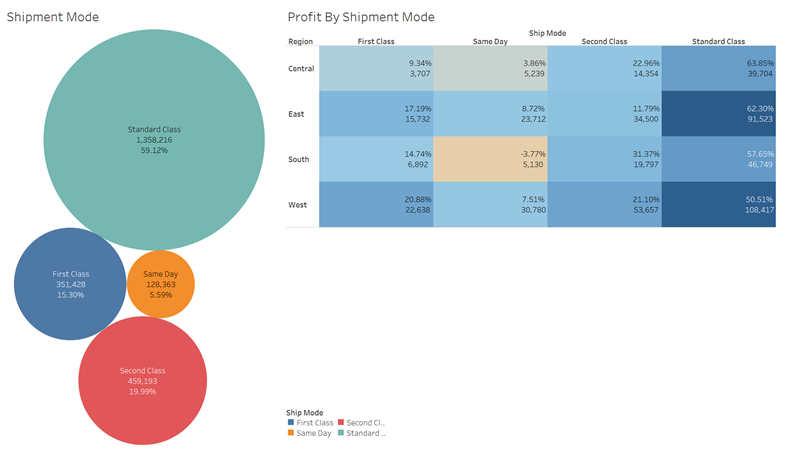
## 6.3 Discount Dashboard



Discount Dashboard depicts that maximum discount in the Superstore business is offered to Office suppliers which is 60.71% but from previous graphs, we have seen that the proportion of sales is the same & profit-wise technology is the highest, but when it comes to offering discounts, technology is the lowest in all 3 categories.

Going to subcategories it's clear the maximum average discount is offered on binders, machines, and tables among their respective categories. Management needs to decide and discuss based on this finding whether it's justified in line with the market- sales, and profit that these category and subcategory generate.

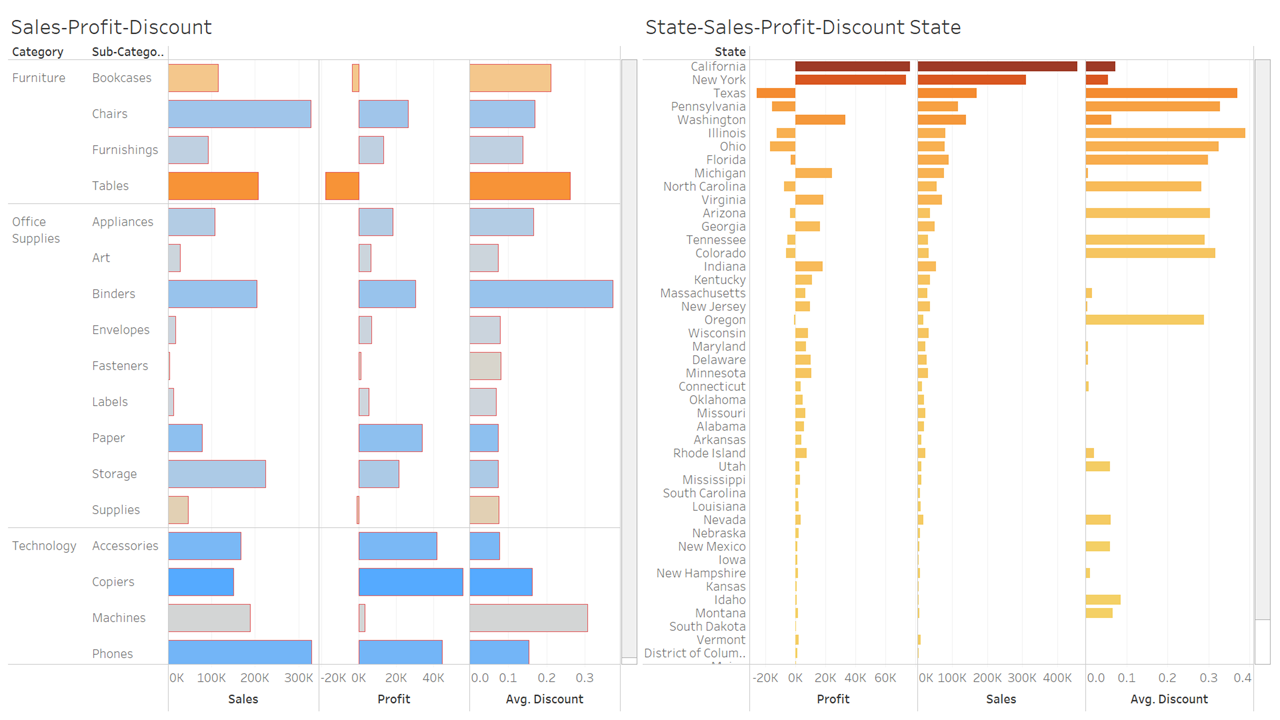
## 6.4 Shipment Dashboard



The above graph explains the contribution of the shipment mode in terms of sales where Standard class is the highest contributor followed by first, second and about 5.59% customers prefer same-day delivery.

The heat map shows the profitability of the region in terms of shipment class and it's evident that standard class is the most profitable, but management should focus on same day delivery in the southern region which is showing negative profit and is a matter of concern which should be addressed on a priority basis.

## 6.5 Sales-Profit -Discount Dashboard



Above all parameters are seen individually and in isolation but the above 2 diagrams give a quick overview in terms of sales, profit, and discount. Despite tables being good in sales and have been offering decent discount still the profit is in negative which should be addressed, also bookcases are in a similar situation.

Machines are showing high sales and discount offerings, but their profit is not in the same proportion. Management should further drill down and investigate how they can improve the profit in machines.

Already explained in the map but to see the performance of the states from all dimensions and make a meaningful analysis we have the next graph. California and New York have the highest sales but the as we can see Texas is showing negative profit which is a matter of concern along with many other states like Pennsylvania, Illinois, Ohio, Florida, North Carolina, etc. which calls for management attention.

Further to that good average discount is offered in many states like Texas, Pennsylvania,Illinois etc but the sales and profit are not very healthy and most cases profit is negative which needs to be addressed.

# 7. Predictive Analysis

## 7.1 Variable Engineering

* Adding the variable ‘unit\_price’ by dividing ‘Sales’ by ‘Quantity' because unit price can better reflect whether the product is expensive or cheap.
* Adding the variable ‘profit\_ratio’ by dividing Profit by Sales, which better indicates the extent of profit of a transaction.
* Creating a two-level variable ‘pro\_cate’ that lables a transaction is whether profitable or not.
* Based on the value of ‘profit\_ratio’, classify transactions into six categories of profit: Extremely Profitable (Profit\_ratio > 0.4), Somewhat Profitable (0.2<Profit\_ratio<0.4), Slightly Profitable (0 < Profit\_ratio < 0.2), Slightly loss (-0.3 < Profit\_ratio < 0), Somewhat loss (-0.6 < Profit\_ratio < -0.3), Huge loss ( Profit\_ratio < -0.6)

## 7.2 Categorical Variable Encoding

For variables that have ordinal relationships, encoding them to ordinal numbers like 1,2,3; for nominal categorical variables, encoding them by one-hot method.

* Encoding variable ‘shipmode’ by labels 1,2,3,4 because the four values of ‘shipmode’ have intrinsic orders. That is, the standard class is the most elementary type of shipping, second class and first class are better, and the same day is the most expensive type of shipping.
* Encoding variable ‘Profit\_Category’ to six labels (1,2,3,4,5,6), 1 is on behalf of Huge loss, while 6 represents extremely profitable.
* Encoding variable ‘State’ to five levels (1,2,3,4,5) by referring to the list of GDP per capita. The most wealthy states (Over 70,000$) are coded as 5, yet states that have a GDP per capita less than 40,000$ are coded as 1.
* Encoding variables ‘Segment’ and ‘Category’ by one-hot encoding.

Dataset used for prediction analysis

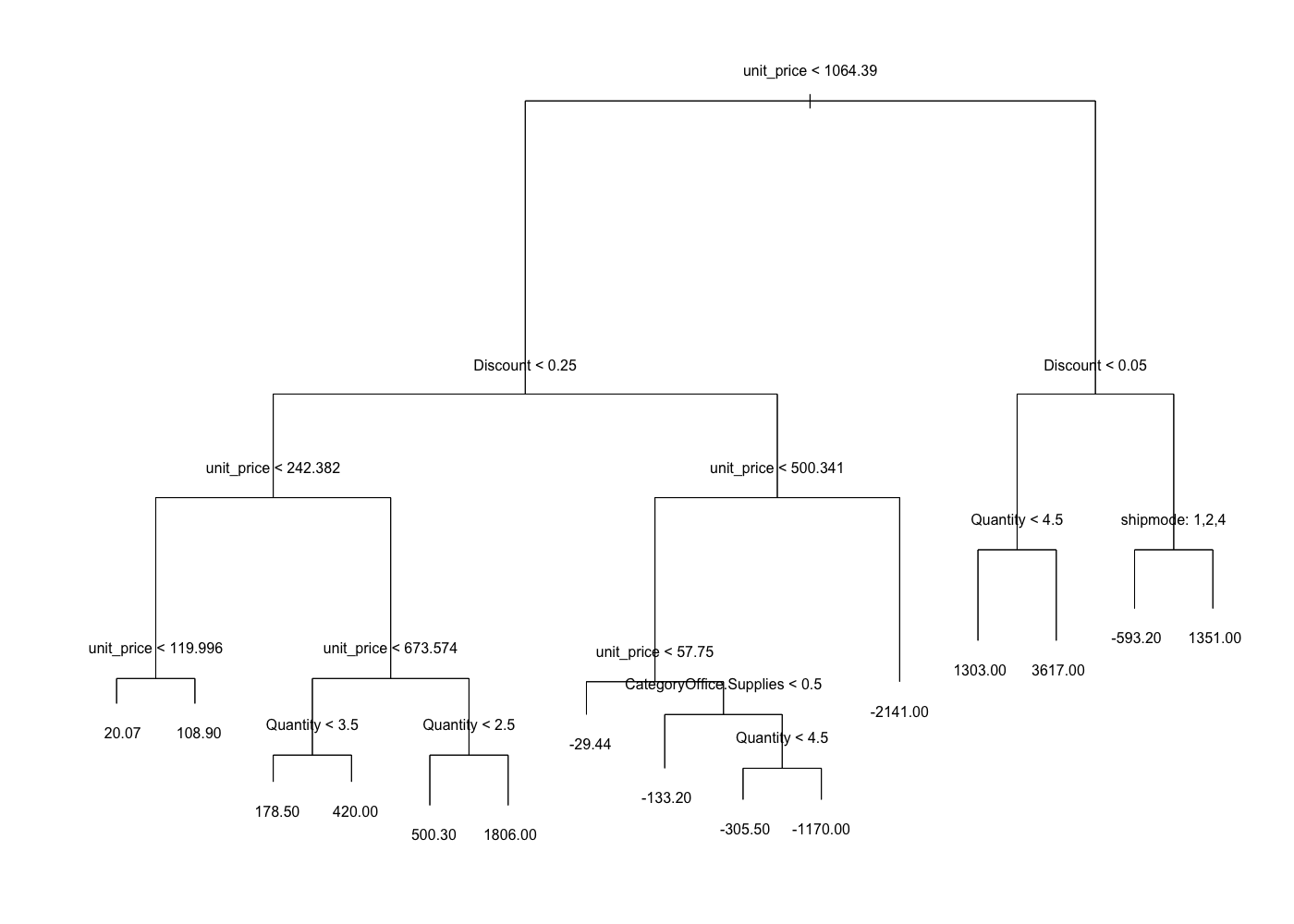
|  |  |  |
| --- | --- | --- |
| Variable\_name | Values | Description |
| shipmode | 1,2,3,4 | Shipping mode of the transaction |
| Quantity | 1,2,3,4,etc. | Quantity of the product in the transaction |
| Discount | 0,0.1,0.2, etc. | Discount placed on the product of the transaction |
| Profit | Continuous values | Total profit of the transaction |
| Unit price | Continuous values | Per item price of the products of the transaction |
| state | 1,2,3,4,5 | Classifiers of state, 5 refers to the most wealthy states, lower numbers refer to less wealthy states |
| SegmentConsumer | 0,1 | 1 refers to the transaction lying in segment ‘consumer’, 0 means not. |
| SegmentCorporate | 0,1 | 1 refers to the transaction lying in segment corporate, 0 means not. |
| SegmentHome.Office | 0,1 | 1 refers to the transaction lying in segment ‘home.office’, 0 means not. |
| CategoryFurniture | 0,1 | 1 refers to the transaction lying in category ‘furniture’, 0 means not. |
| CategoryOffice.Supplies | 0,1 | 1 refers to the transaction lying in category ‘office.supplies’, 0 means not. |
| CategoryTechnology | 0,1 | 1 refers to the transaction lying in category ‘technology’, 0 means not. |
| pro\_cate | 0,1 | 1 indicates the transaction is profitable, 0 means that the transaction deficits. |

## 7.3 Regression

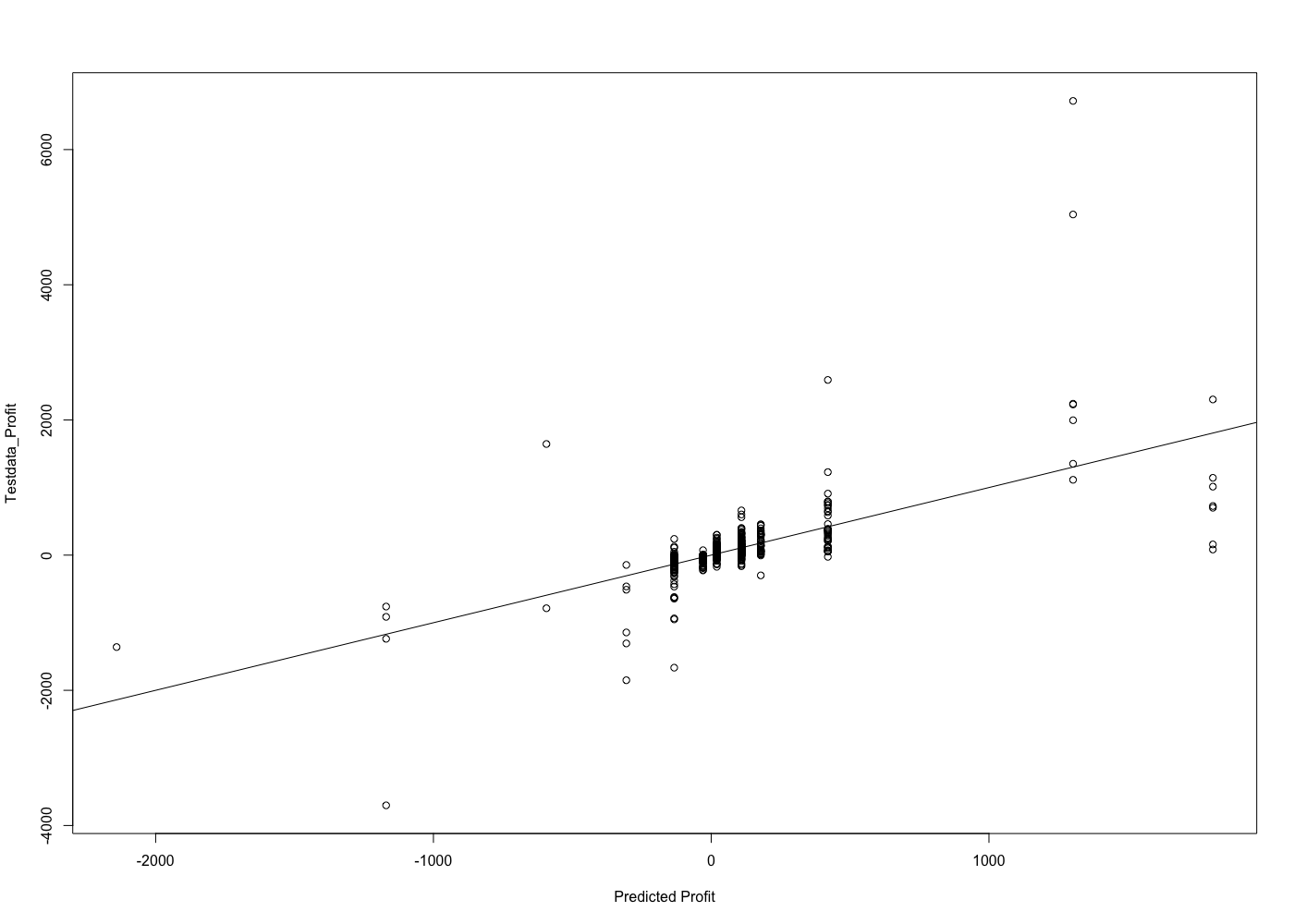
### 7.3.1 Decision Tree

Variables that are actually used in tree construction: unit\_price, Discount, Quantity, CategoryOffice.Supplies, shipmode.

Residual Mean Deviance**:** 21810

****

The tree above shows some meaningful information. In general, transactions that were offered a high discount tend to be lost. Given a relatively low discount placed, transaction is more profitable when the unit price of the product is high. In contrast, if the discount is over 0.25, then the transaction will be more or less unprofitable. Especially noting that the most unprofitable branch under the tree is discount > 0.25 & unit\_price > 500.341), while the most profitable branch is unit\_price > 1064.39 & Discount < 0.05 & Quantity >= 5. Products of transactions within these two groups are mostly machines, some were sold with huge discounts and caused miserable loss, but the others were sold at a normal price and harvested a good return. Furthermore, the second highest loss branch of the tree is the ‘-1170’ one. By slicing the dataset according to the information of the branches, it appears that most of the products are ‘Binders’. Next, we searched in the Amazon and found that products that belong to binders with an unit price higher than 58 are machines like coil binding machine. Thus, these highly lost office supplies are practically machines. These uncover the fact that selling machines is risky, though they have a high marginal return. However, If they were kept too long in the warehouse, the merchant has to lower down the prices and sell them quickly. The tree also shows that ‘Quantity’ and ‘shipmode’ have a little impact on the final value of profit, let alone variables that were not included in the tree.

Applying the tree model to the test dataset indicates a squared MSE(Mean Squared Error) of 212, which is far away from the mean of Profit (28), and the plot that shows the accuracy is:

The squared MSE is too high, and the plot shows many unmatched values. The reason is that there are outliers in the dataset, of which the tree model can’t predict accurately.

### 7.3.2 Random Forest

By applying the same variables and train data to Random Forest results in a squared MSE of 159, which is better than the previous one. Nevertheless, the error is still not that acceptable. Therefore, classification models are used for a better prediction.

## 7.4 Classification

### 7.4.1 Binomial Logistic Regression

Since the logistic regression used for predicting whether a transaction loss or not is the simple, binomial one. That means, its target variable can’t exceed two levels. Thus, the target variable used for this classification is ‘pro\_cate’.

Coefficient Estimated Value and Significance table:

|  |  |  |
| --- | --- | --- |
| Coefficients | Estimate | P-Value |
| shipmode | 0.116 | 0.041 |
| Quantity | 0.032 | 0.180 |
| Discount | -2.396 | 2e-16 |
| unit\_price | 0.0004 | 0.366 |
| state | 0.060 | 0.204 |
| SegmentConsumer | 0.242 | 0.101 |
| SegmentCorporate | 0.060 | 0.709 |
| SegmentHome.Office | NA | NA |
| CategoryFurniture | -1.148 | 2e-16 |
| CategoryOffice.Supplies | 0.427 | 0.003 |
| CategoryTechnology | NA | NA |

The table above shows that variable ‘shipmode’ has a slight influence on pro\_cate, meaning that products that were shipped by express-level modes tend to be profitable. However, a possible explanation is that the superstore tends to ship their profitable products by high-level shipping modes. Thus, without further analysis, it is only confidence to say that shipping mode has a relation to pro\_cate, but not that the two have an interactive correlation.

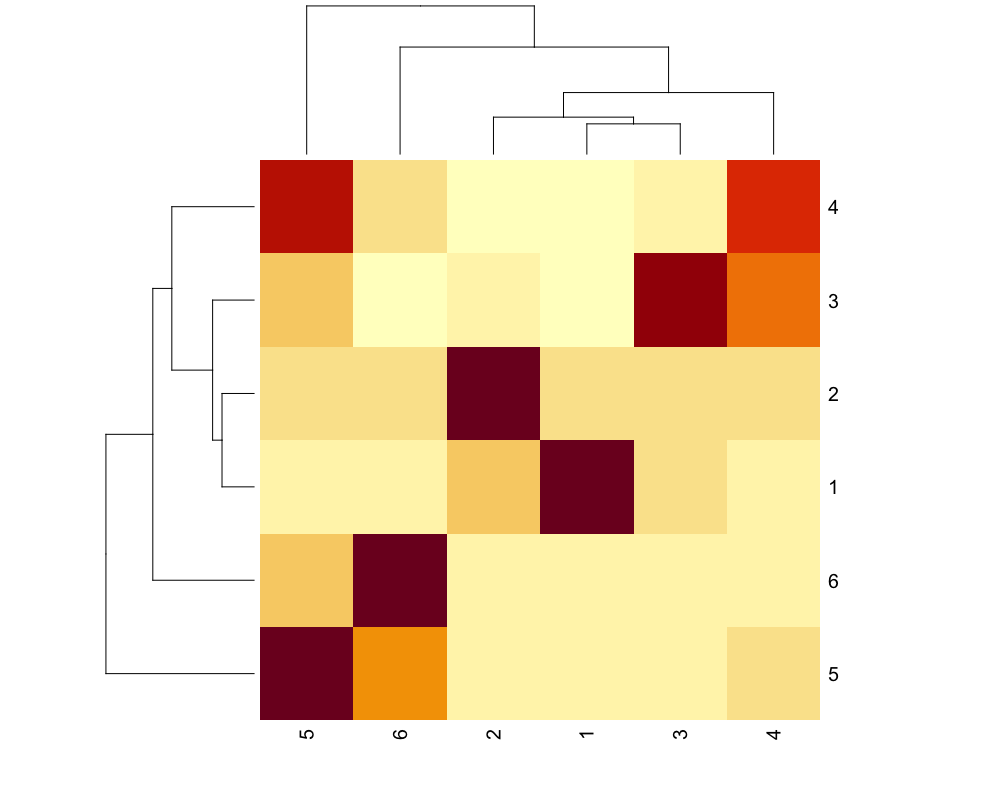
Testing the model on the test data results in an accuracy of 0.942, so this model is strong enough to predict whether a transaction will be profitable or not. Nevertheless, only profitable or not is not enough for accurately understanding transactions because some transactions lose a lot while some earn a lot. Therefore, applying a multi-level classification model is necessary to make a more precise prediction.

### 7.4.2 Supply Vector Machine for Multi-class Classification

Defining Profit\_Category as the target variable for this classification. By playing with the two arguments ‘gamma’ and ‘cost’ of the function svm, and adding or deleting variables from the formula, the model finally turns out a decent accuracy, 0.64. The adding/deleting process testifies that variables state/unit\_price/Quantity/shipmode barely affect the accuracy, while variable Discount evidently impacts the result.

**Accuracy Table**:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **SVM\_Prediction** | | | | | |
| **Test Data Value** | **1** | **2** | **3** | **4** | **5** | **6** |
| **1** | **56** | **8** | **1** | **0** | **0** | **0** |
| **2** | **0** | **113** | **4** | **0** | **0** | **0** |
| **3** | **0** | **14** | **98** | **61** | **36** | **6** |
| **4** | **0** | **0** | **15** | **183** | **213** | **39** |
| **5** | **0** | **0** | **0** | **69** | **497** | **195** |
| **6** | **0** | **0** | **0** | **2** | **59** | **330** |



The table reveals that the model has a pleasing prediction over profit category 1,2,6, and the prediction toward level 5 is also decent, whereas it only has about 40% of accuracy in predicting level 3,4. A reasonable illustration to this phenomenon is that the features of transactions’ profit that lies between -.30 ~ .20 might vary a lot, thereby causing the inaccuracy.

# 8. Conclusion

Exploring the dataset through the ways above discovers several facts. ‘Discount’ accounts for most of the variable of profit of a transaction; Amongst different categories of products, technological products have a better return than office supplies than furniture; ‘unit\_price’ will not determine whether a transaction earns or not, but it has a positive relation with the size of the figure. GDP per capita of a ‘state’, ‘Quantity’, ‘Segment’, and ‘shipmode’ does not impact the profitability of a transaction.

# 9. Reference

Khurana, S. (2021, January 18). *Superstore sales use case data analytics and visualization*.

Medium. Retrieved October 21, 2021, from [https://medium.com/clique-org/superstore- sales-use-case-data-analytics-and-visualization-62afacd0777](https://medium.com/clique-org/superstore-%20sales-use-case-data-analytics-and-visualization-62afacd0777).

Xiemin, K. H. (2021, July 20). *Exploratory Data Analysis: Super Store*. Medium. Retrieved

October 21, 2021, from https://medium.com/analytics-vidhya/exploratory-data-analysis-super-store-cb91c37bcb06

Sunil, R(2020, December 23). *How to deal with categorical variable in Predictive modeling*.

Analytics Vidhya. Retrieved October 21, 2021, from

<https://www.analyticsvidhya.com/blog/2015/11/easy-methods-deal-categorical-variables-predictive-modeling>.

Joshua, H et.al, (2019, December). Improving Earnings Predictions with Machine learning.

Retrieved October 21, 2021, from<https://zicklin.baruch.cuny.edu/wp-content/uploads/sites/10/2019/12/Improving-Earnings-Predictions-with-Machine-Learning-Hunt-Myers-Myers.pdf>

# 10. Appendix

## 10.1 R Script

# Library Packages

library(readr)

library(dplyr)

library(caret)

library(tree)

library(randomForest)

library(varhandle)

library(e1071)

library(plot.matrix)

# Read Data

df\_b <- read\_csv("~/Desktop/BAN5573/Project\_df\_SampleSuperstore.csv")

View(df\_b)

# Adding variables

df\_b$unit\_price <- df\_b$Sales / df\_b$Quantity

df\_b$profit\_ratio <- (df\_b$Profit / df\_b$Sales) \* 100

# Rename Variables

df\_b <- rename(df\_b, shipmode = 'Ship Mode')

# transform dependent variable from numerical values to categorized values.

for (i in 1:nrow(df\_b)){

if (df\_b$profit\_ratio[i] > 40){

df\_b$Profit\_Category[i] <- 'Extremely Profitable'

} else if (df\_b$profit\_ratio[i] > 20 && df\_b$profit\_ratio[i] <= 40) {

df\_b$Profit\_Category[i] <- 'Somewhat Profitable'

} else if (df\_b$profit\_ratio[i] > 0 && df\_b$profit\_ratio[i] <= 20) {

df\_b$Profit\_Category[i] <- 'Slightly Profitable'

} else if (df\_b$profit\_ratio[i] > -50 && df\_b$profit\_ratio[i] <= 0){

df\_b$Profit\_Category[i] <- 'Slightly loss'

} else if (df\_b$profit\_ratio[i] > -100 && df\_b$profit\_ratio[i] <= -50){

df\_b$Profit\_Category[i] <- 'somewhat loss'

} else {

df\_b$Profit\_Category[i] <- 'Huge loss'

}

}

# Data filtering

df\_b <- df\_b %>% select(-'Country',-City,-Sales,-'Postal Code')

View(df\_b)

# Features engineering ####

# Encoding variable 'ship mode' to an ordinal array 1,2,3,4

df\_b$shipmode <- factor(df\_b$shipmode,

levels = c('Standard Class','Second Class','First Class','Same Day'),

labels = c(1,2,3,4))

df\_b$Profit\_Category <- factor(df\_b$Profit\_Category,

levels = c('Huge loss','somewhat loss','Slightly loss','Slightly Profitable','Somewhat Profitable','Extremely Profitable'),

labels = c(1,2,3,4,5,6))

# Recategorize variable 'state' according to state gdp per capita

high\_income <- c('District of Columbia','Massachusetts','New York','Alaska','North Dakota','California')

relatively\_high\_income <- c('Connecticut','Washington','Wyoming','Delaware','New Jersey','Maryland','Illinois','Texas','Colorado','Minnesota')

mid\_income <- c('Nebraska','Hawaii','New Hampshire','Virginia','Pennsylvania','Iowa','Kansas','South Dakota','Oregon','Ohio','Wisconsin','Rhode Island','Louisiana','Utah','Oklahoma','Georgia','Nevada')

relatively\_low\_income <- c('Indiana','Vermont','North Carolina','Tennessee','Michigan','Missouri','New Mexico','Florida','Arizona','Montana','Maine','Kentucky','South Carolina','Alabama','Idaho','West Virginia')

low\_income <- c('Arkansas','Mississippi')

for (i in 1:nrow(df\_b)){

if (df\_b$State[i] %in% high\_income) {

df\_b$state[i] <- 5

} else if(df\_b$State[i] %in% relatively\_high\_income){

df\_b$state[i] <- 4

} else if (df\_b$State[i] %in% mid\_income) {

df\_b$state[i] <- 3

} else if (df\_b$State[i] %in% relatively\_low\_income){

df\_b$state[i] <- 2

} else if (df\_b$State[i] %in% low\_income) {

df\_b$state[i] <- 1

}

}

# Encoding variable 'Segment' and 'Category'

dmy <- dummyVars("~Segment + Category",data = df\_b)

df\_b1 <- data.frame(predict(dmy,newdata = df\_b))

df\_b2 <- cbind(df\_b,df\_b1)

#Create a new categorical variable seperating profitable and unprofitable transactions.

for (i in 1:nrow(df\_b2)) {

if (df\_b2$profit\_ratio[i] >= 0) {

df\_b2$pro\_cate[i] = 1

} else {

df\_b2$pro\_cate[i] = 0

}

}

# Filter for regression decision tree

df\_b2\_tree <- df\_b2 %>% select(-State,-Segment,-Region,-Category,-`Sub-Category`,-profit\_ratio,-Profit\_Category)

View(df\_b2\_tree)

# Modeling #######

# Regression ======

# Regression Decision Tree -------------------------

# Fitting Regression Trees

set.seed(71)

train\_ind <- sample(seq\_len(nrow(df\_b2\_tree)), size = 0.8\*nrow(df\_b2\_tree))

train <- df\_b2\_tree[train\_ind, ]

test <- df\_b2\_tree[-train\_ind, ]

View(train)

tree.df <- tree(Profit ~ .-pro\_cate,train)

summary(tree.df)

plot(tree.df)

text(tree.df, pretty = 0)

# Testing decision tree model

yhat1 <- predict(tree.df, newdata = test)

df.test <- test[,"Profit"]

plot(yhat1,df.test,xlab = 'Predicted Profit',ylab = 'Testdata\_Profit')

abline(0,1)

MSE\_tree <- mean((yhat1 - df.test)^2)

sqrt(MSE\_tree)

# Random Forest ------------

rf.df <- randomForest(Profit ~.-pro\_cate, data = train)

yhat.rf <- predict(rf.df, newdata = test)

MSE\_rf <- mean((yhat.rf - df.test)^2)

sqrt(MSE\_rf)

importance(rf.df)

# Part Conclusion --------

# The result shows that random forest yields an improvement over regression

# decision tree, but the squared MSE still maintains around 160,

# Which is substantially higher than the mean profit.

# Therefore, we would adopt classification models to get a better fit.

# Classification Models ==========

df\_b2\_classification <- df\_b2 %>% select(-State,-Segment,-Region,-Category,-`Sub-Category`,-Profit,-profit\_ratio,-Profit\_Category)

df\_b2\_classification$shipmode <- as.numeric(as.character(df\_b2\_classification$shipmode))

str(df\_b2\_classification)

View(df\_b2\_classification)

train\_cla <- df\_b2\_classification[train\_ind, ]

test\_cla <- df\_b2\_classification[-train\_ind, ]

# Logistic Regression --------

logistic.fits <- glm(pro\_cate~ ., data = train\_cla, family = binomial)

summary(logistic.fits)

logistic.probs <- predict(logistic.fits, test\_cla, type = "response")

glm.pred <- rep(0,1999)

glm.pred[logistic.probs > .5] <- 1

table(glm.pred,test\_cla$pro\_cate)

accuracy\_of\_LR <- mean(glm.pred == test\_cla$pro\_cate)

accuracy\_of\_LR

# Supply Vector Machine for Multi-class Classification -----

# SVM with variable 'Profit\_Category'

df\_b2\_svm <-df\_b2 %>% select(-State,-Segment,-Region,-Category,-`Sub-Category`,-Profit,-profit\_ratio,)

df\_b2\_svm$shipmode <- as.numeric(as.character(df\_b2\_svm$shipmode))

View(df\_b2\_svm)

train\_svm <- df\_b2\_svm[train\_ind, ]

test\_svm <- df\_b2\_svm[-train\_ind, ]

svm1 <- svm(Profit\_Category ~.-pro\_cate-Quantity, data = train\_svm, method = "C-classification", kernel = "radial",gamma = 0.1, cost = 5)

summary(svm1)

svm1$SV

# Testing SVM

svm\_prediction <- predict(svm1,test\_svm)

svm\_tab <- table(test\_svm$Profit\_Category, svm\_prediction)

svm\_tab

overall\_accuracy <- mean(test\_svm$Profit\_Category == svm\_prediction)

overall\_accuracy

# Visualizing Results:

plot(svm\_tab)

svm\_file <- data.frame(test\_svm$Profit\_Category,svm\_prediction)

write.csv(svm\_file, file = "svm\_file.csv")

heatmap(svm\_tab)

.