SHIPPING MODE PREDICTOR

A comparative study among predictive models to ascertain correct shipping mode as well as choose significant factors

Full Code Available Here

Project Objective:

To build the best predictive model using ML techniques to predict the correct shipping mode for a **Supply Chain and Logistics** company based on the Inventory data in R

Environment Set Up:

- **Some R packages used for this project:** readxl, outliers, psych, DMwR, UBL, ROSE, ROCR, e1071, class, ipred, stringr etc.
- Set up Workind Directory using **getwd()** command.
- Import the dataframe (present in Excel format) using **read_xlsx** command.

Preliminary Analysis:

DataFrame Analysis and Variable Identification:

- **dim():**

```
[1] 7853    8
```

- **str():**

```
tibble [7,853 x 8] (S3: tbl_df/tbl/data.frame)
 $ Order Date         : chr [1:7853] "1/27/2007" "1/27/2007" "1/27/2007" "1/27/2007" ...
 $ Order ID           : num [1:7853] 24544 24544 24544 20422 55937 ...
 $ Order Quantity     : num [1:7853] 31 39 15 30 10 5 11 24 49 38 ...
 $ Product Container  : chr [1:7853] "Medium Box" "Large Box" "Jumbo Drum" "Small Pack" ...
 $ Product Name       : chr [1:7853] "Canon MP41DH Printing Calculator" "Fellowes Neat Ideas® Storage Cubes" "Global Stack Chair without Arms, Black" "Nu-Dell Lea
herette Frames" ...
 $ Product Sub-Category: chr [1:7853] "Office Machines" "Storage & Organization" "Chairs & Chairmats" "Office Furnishings" ...
 $ Sales              : num [1:7853] 6567 1780 578 611 517 ...
 $ Ship Mode          : chr [1:7853] "Express Air" "Regular Air" "Delivery Truck" "Regular Air" ...
```

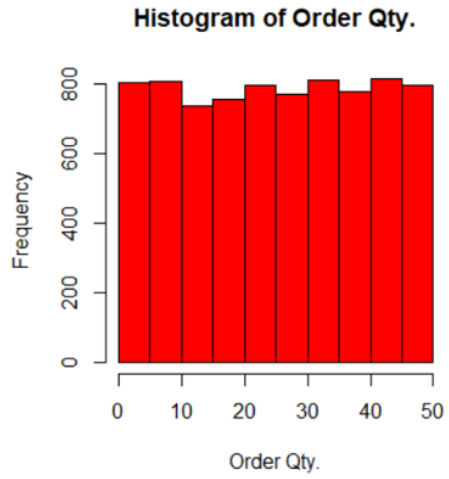
- **summary():**

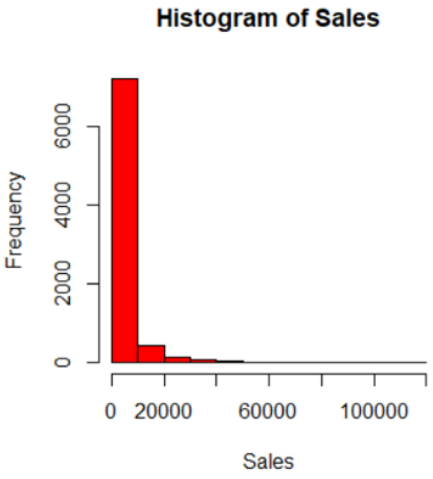
```
  Order Date          Order ID     Order Quantity  Product Container  Product Name       Product Sub-Category     Sales         Ship Mode
 Length:7853        Min.   :    3   Min.   : 1.00   Length:7853        Length:7853        Length:7853         Min.   :    4   Length:7853
 Class :character   1st Qu.:14855   1st Qu.:13.00   Class :character   Class :character   Class :character    1st Qu.:  244   Class :character
 Mode  :character   Median :29637   Median :26.00   Mode  :character   Mode  :character   Mode  :character    Median :  747   Mode  :character
                    Mean   :29861   Mean   :25.59                                                          Mean   : 3044
                    3rd Qu.:44583   3rd Qu.:38.00                                                          3rd Qu.: 2959
                    Max.   :59971   Max.   :50.00                                                          Max.   :114362
```

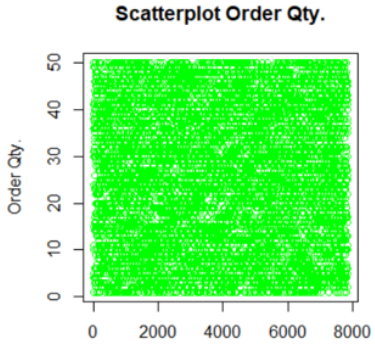
- Variables **Order Date**, **Order ID** and **Product Name** are irrelevant for prediction, so we drop them completely.

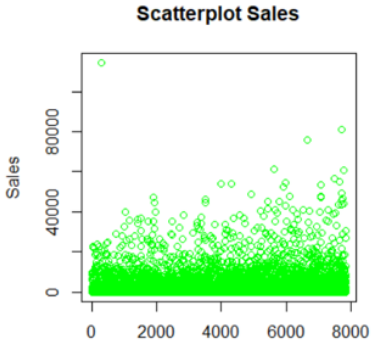
```
> Inventory <- X09_Inventory[-c(1,2,5)]
> str(Inventory)
'data.frame':   7853 obs. of  5 variables:
 $ Order Quantity     : num  31 39 15 30 10 5 11 24 49 38 ...
 $ Product Container  : Factor w/ 7 levels "Jumbo Box","Jumbo Drum",..: 4 3 2 6 5 5 5 1 4 5 ...
 $ Product Sub-Category: Factor w/ 17 levels "Appliances","Binders and Binder Accessories",..: 10 15 4 9 5 8 11 16 9 17 ...
 $ Sales              : num  6567 1780 578 611 517 ...
 $ Ship Mode          : Factor w/ 3 levels "Delivery Truck",..: 2 3 1 3 3 3 3 1 3 3 ...
```

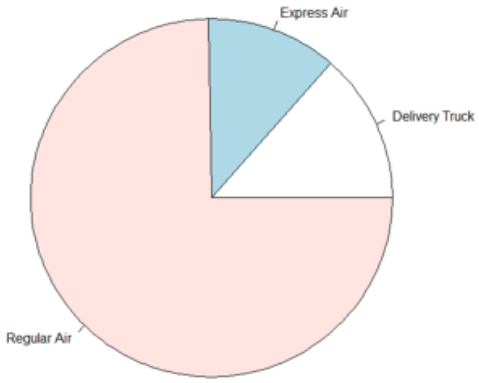
Univariate Analysis:

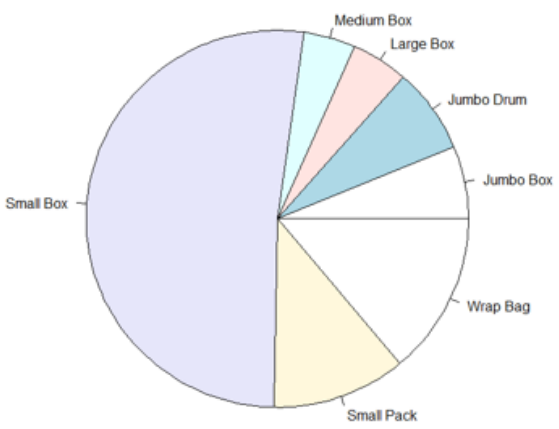


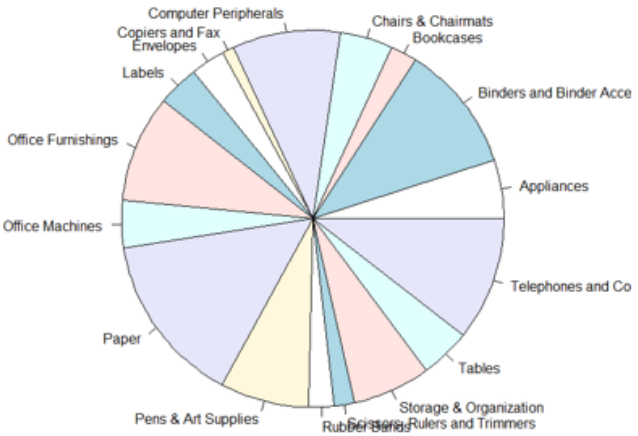


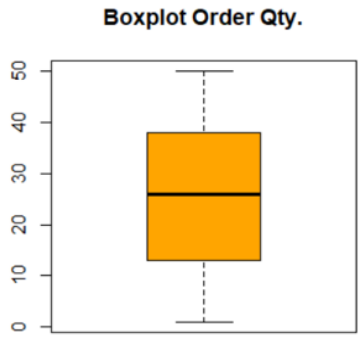


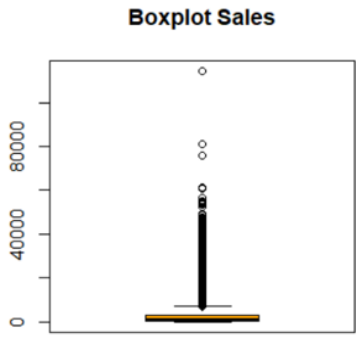




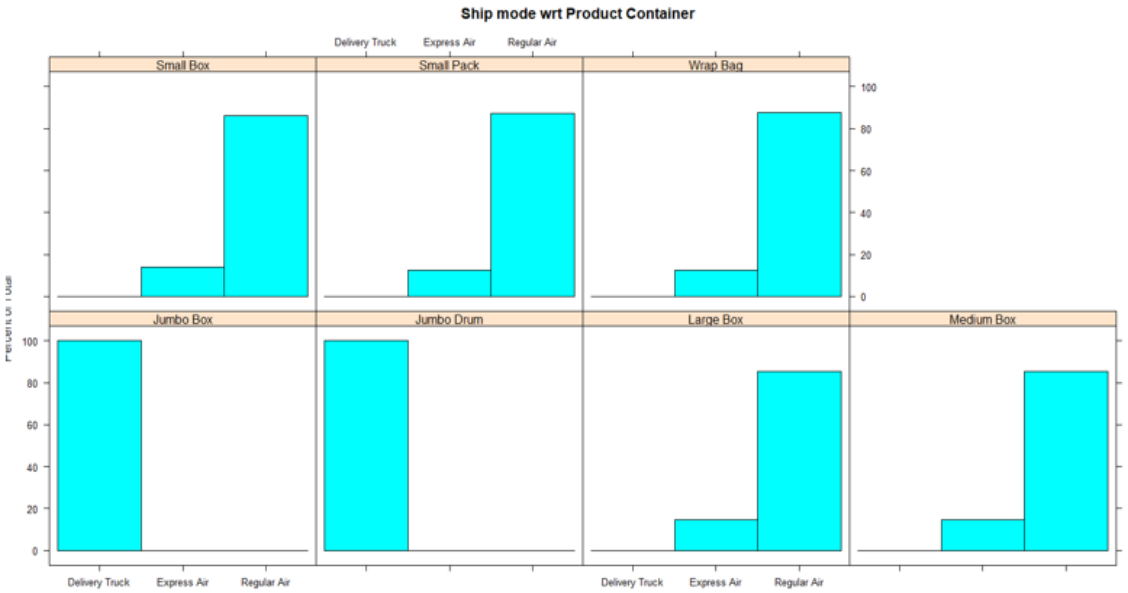


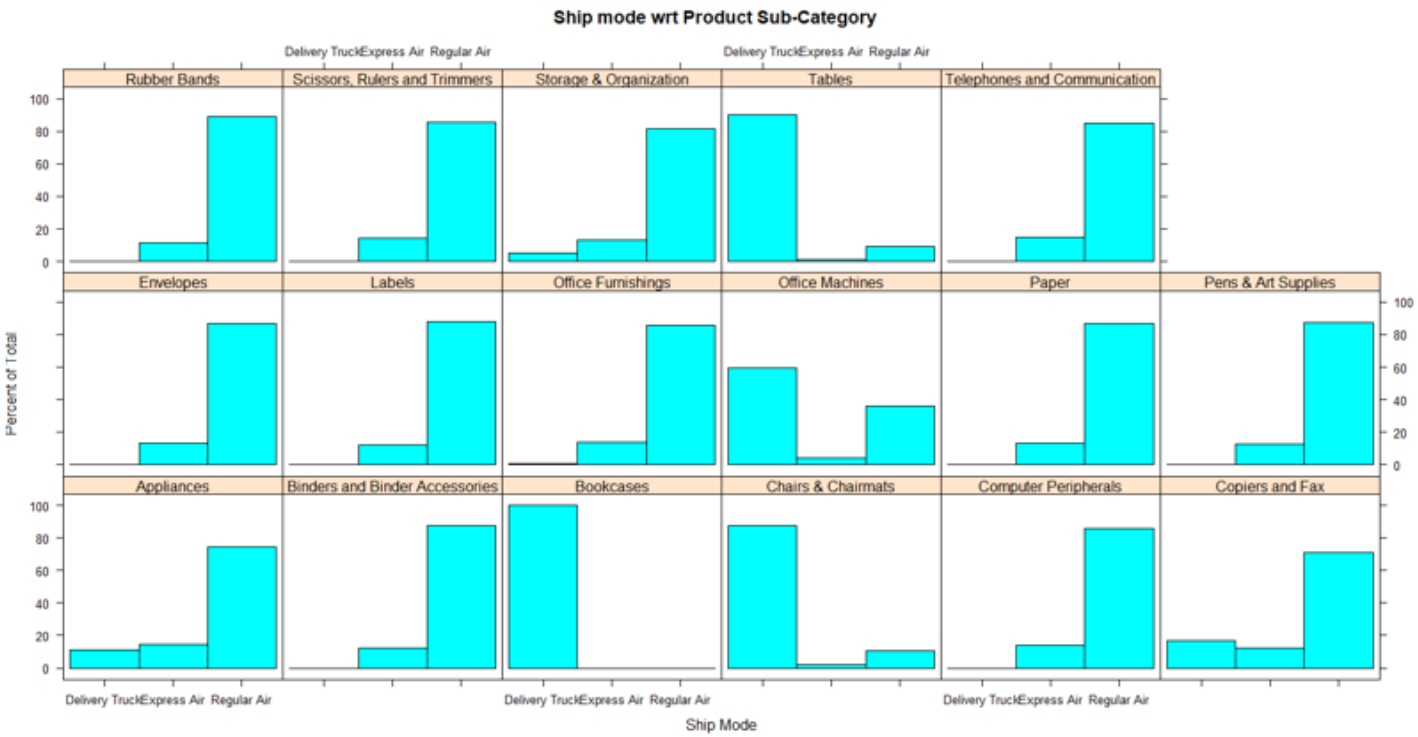




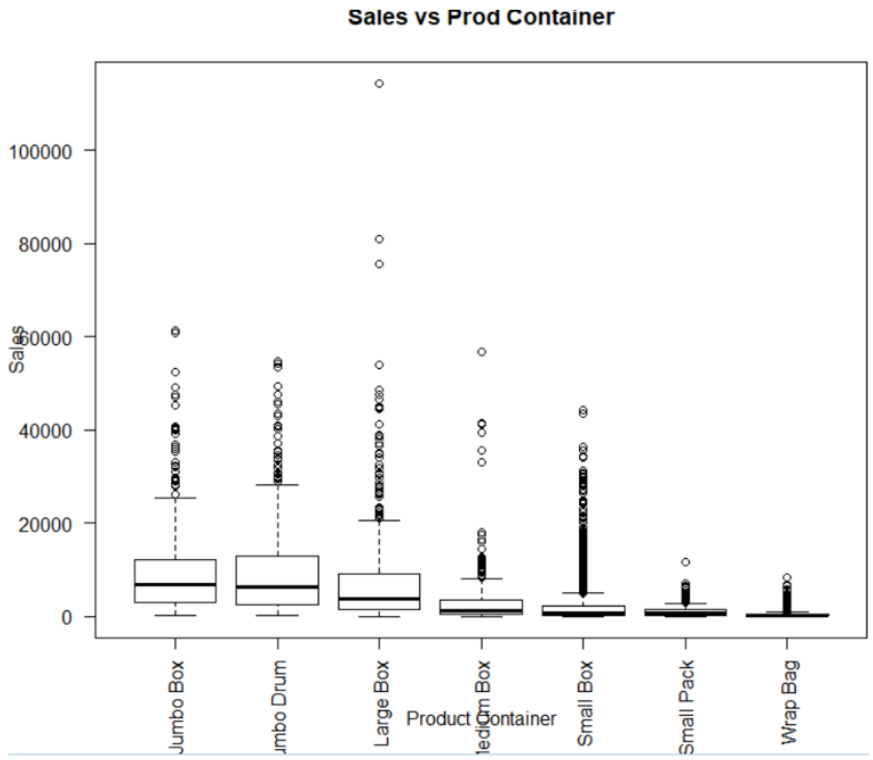


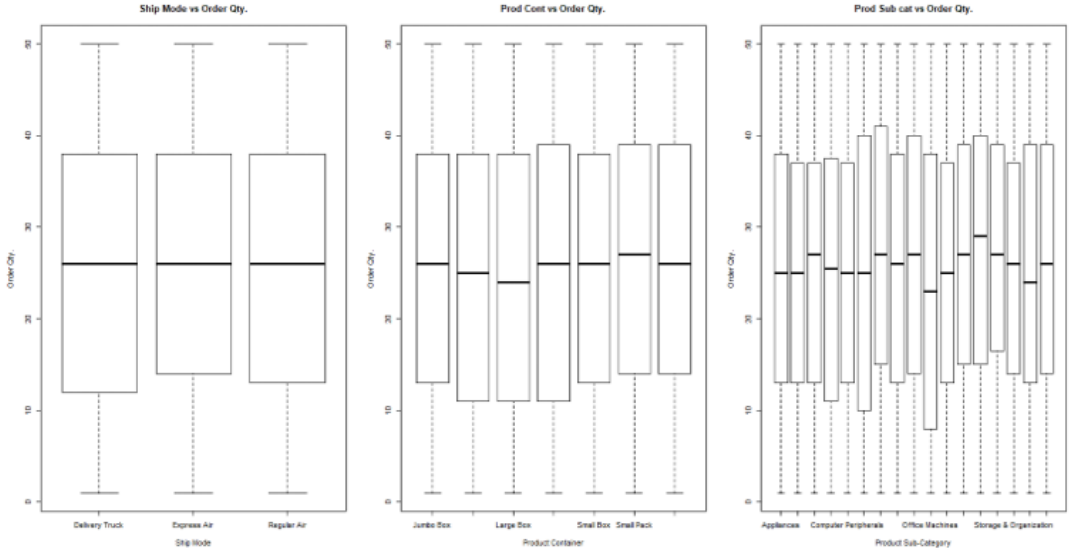
Bivariate Analysis:





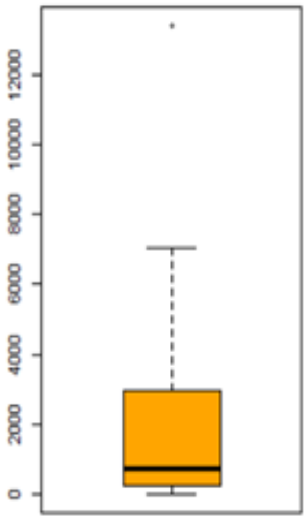

Sales vs Ship Mode
Sales
0
40000
80000
Delivery Truck
Express Air
Regular Air

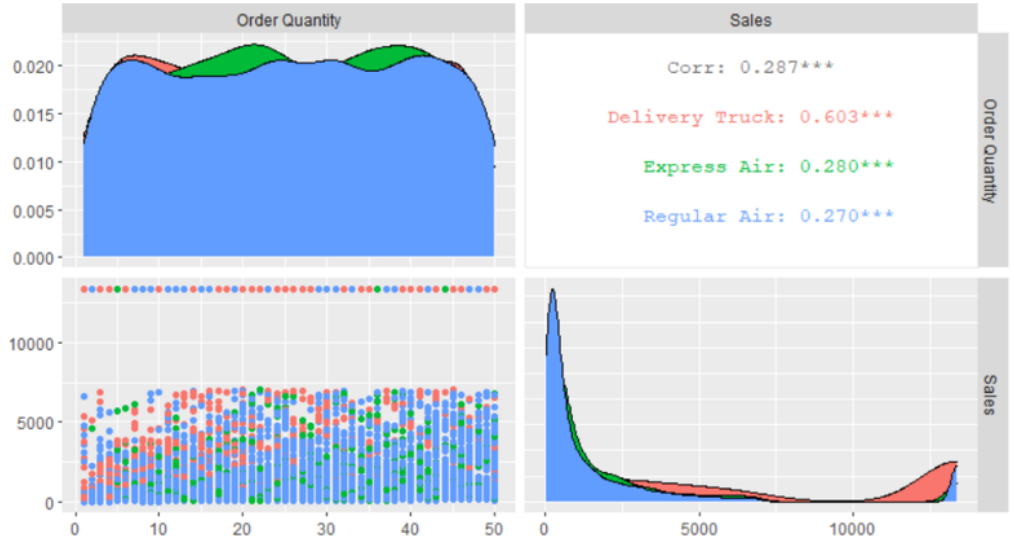

Sales vs Prod Container
Sales
0
20000
40000
60000
80000
100000
Jumbo Box
Large Box
Small Box
Small Pack
Wrap Bag
Product Container


Ship Mode vs Order Qty.
Prod Cont vs Order Qty.
Prod Sub cat vs Order Qty.
Order Qty.
Delivery Truck
Express Air
Regular Air
Ship Mode
Jumbo Box
Large Box
Small Box
Small Pack
Product Container
Appliances
Computer Peripherals
Office Machines
Storage & Organization
Product Sub-Category

Outlier Identification and Multicollinearity:

There seems to be no correlation between independent variables





One Hot Encoding and Imbalance Treatment:

```
> Inventory<-one_hot(as.data.table(Inventory2[,-5]))
Inventory                                  7853 obs. of 27 variables
   Order.Quantity : num 31 39 15 30 10 5 11 24 49 38 ...
   Product.Container_Jumbo.Box : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 2 1 1 ...
   Product.Container_Jumbo.Drum : Factor w/ 2 levels "0","1": 1 1 2 1 1 1 1 1 1 1 ...
   Product.Container_Large.Box : Factor w/ 2 levels "0","1": 1 2 1 1 1 1 1 1 1 1 ...
   Product.Container_Medium.Box : Factor w/ 2 levels "0","1": 2 1 1 1 1 1 1 1 2 1 ...
   Product.Container_Small.Box : Factor w/ 2 levels "0","1": 1 1 1 1 2 2 2 1 1 2 ...
   Product.Container_Small.Pack : Factor w/ 2 levels "0","1": 1 1 1 2 1 1 1 1 1 1 ...
   Product.Container_Wrap.Bag : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
   Product.SubCategory_Appliances : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
   Product.SubCategory_Binders.and.Binder.Accessories: Factor w/ 2 levels "0","1": 1 1 1…
   Product.SubCategory_Bookcases : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
   Product.SubCategory_Chairs.and.Chairmats : Factor w/ 2 levels "0","1": 1 1 2 1 1 1 1 …
   Product.SubCategory_Computer.Peripherals : Factor w/ 2 levels "0","1": 1 1 1 1 2 1 1 …
   Product.SubCategory_Copiers.and.Fax : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1…
   Product.SubCategory_Envelopes : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
   Product.SubCategory_Labels : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 2 1 1 1 1 ...
   Product.SubCategory_Office.Furnishings : Factor w/ 2 levels "0","1": 1 1 1 2 1 1 1 1 …
   Product.SubCategory_Office.Machines : Factor w/ 2 levels "0","1": 2 1 1 1 1 1 1 1 1 1…
   Product.SubCategory_Paper : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 2 1 1 1 ...
   Product.SubCategory_Pens.and.Art.Supplies : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1…
   Product.SubCategory_Rubber.Bands : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
   Product.SubCategory_Scissors.Rulers.and.Trimmers : Factor w/ 2 levels "0","1": 1 1 1 …
   Product.SubCategory_Storage.and.Organization : Factor w/ 2 levels "0","1": 1 2 1 1 1 …
   Product.SubCategory_Tables : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 2 1 1 ...
   Product.SubCategory_Telephones.and.Communication : Factor w/ 2 levels "0","1": 1 1 1 …
   Sales : num 6567 1780 578 611 517 ...
   Ship.Mode : Factor w/ 3 levels "1","2","3": 2 1 3 1 1 1 1 3 1 1 ...
```

We see the data is highly imbalance

```
> prop.table(table(Inventory$Ship.Mode))

        1         2         3
0.7473577 0.1172800 0.1353623
> table(Inventory$Ship.Mode)

   1    2    3
5869  921 1063
> prop.table(table(trainDataLR$Ship.Mode))

        1         2         3
0.7473684 0.1173175 0.1353141
> table(trainDataLR$Ship.Mode)

   1    2    3
4402  691  797
> prop.table(table(testDataLR$Ship.Mode))

        1         2         3
0.7473255 0.1171676 0.1355069
> table(testDataLR$Ship.Mode)

   1    2    3
1467  230  266
```

We increase the ratio of classes 1,2 and 3 from 75:12:14 to 40:27:33

```
> trainDataLR_SMOTE <- ovun.sample(Class ~ ., data = trainDataLR, method =
 "over",p=0.6,seed=1)$data
> prop.table(table(trainDataLR_SMOTE$Ship.Mode))

        1         2         3
0.3994193 0.2742945 0.3262862
```

Model Building and Comparitive Analysis:

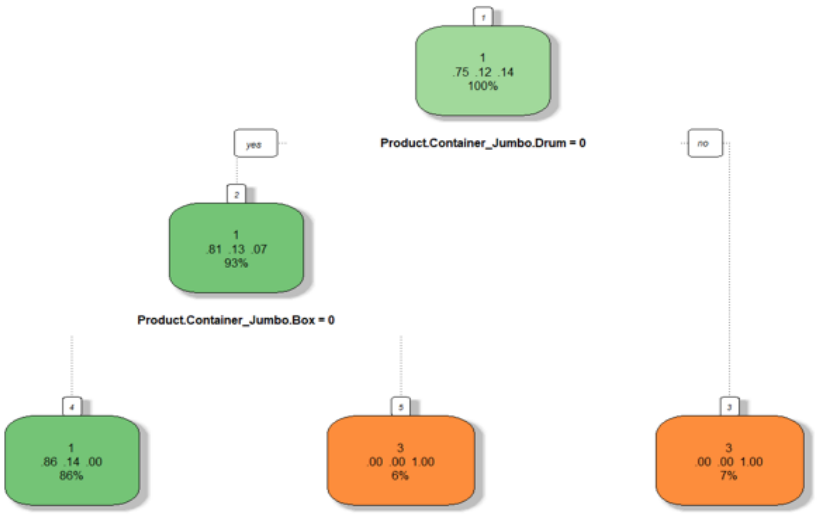
	Accuracy	95% CI	No. of class 2 predicted correctly	No information Rate	P-value (Acc > NIR)	Kappa
Multinomial Logistic Regression	0.8619	(0.8459,0.8769)	6/230	0.7473	< 2.2e-16	0.6085
Support Vector Machine	0.8074	(0.7893,0.8247)	32/320	0.7473	1.684e-10	0.5229
Bagging	0.8828	(0.8678,0.8967)	0	0.7473	<2.2e-16	0.6509
Decision Tree	0.8828	(0.8678,0.8967)	0	0.7473	<2.2e-16	0.6509
Random Forest	0.8248	(0.8072,0.8413)	22/230	0.7473	<2.2e-16	0.5464
Gradient Boosting	0.8686	(0.8528,0.8832)	4/230	0.7473	<2.2e-16	0.6214

Model Comparison:

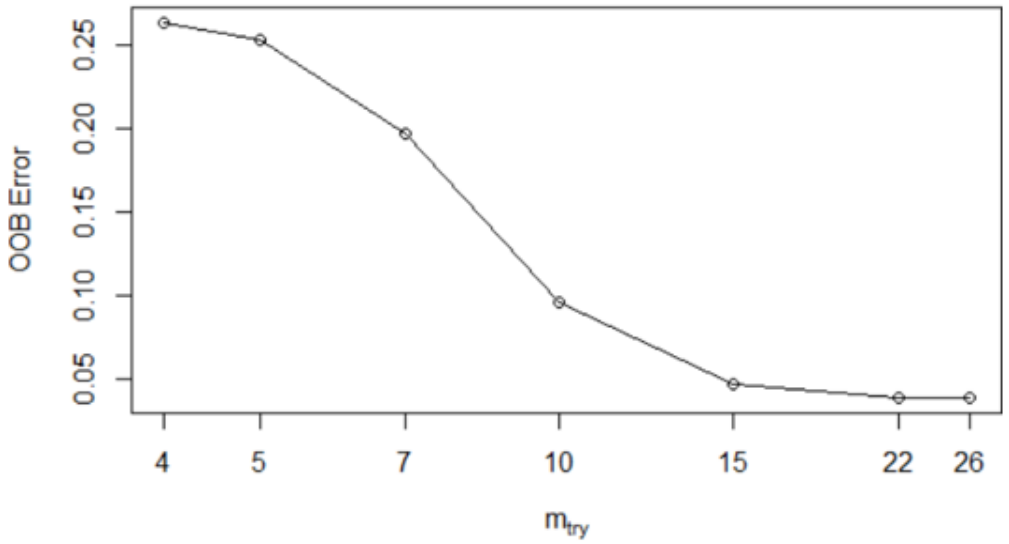
- Bagging and Decision Tree models are outright rejected due to their inability to predict Class 2 samples.
- All the models, including base and ensemble classifiers, have got high predictive power for Class 1 and 3 samples.
- All the models struggled to get a moderate F1 score for Class 2 samples. This again highlights our insight derived from EDA that shipping mode Express Air is difficult to predict because there seems hardly any logic among independent variables. Even complex ML algorithms find it hard to find some pattern w.r.t Express Air.
- Only two models, SVM and RF, could predict some Class 2 samples.
- Out of the above two, we select **SVM as our final predictive model**.
- To increase the number of predictions of Class 2 samples if possible, we can employ Hyper parameter tuning.

Model Plots

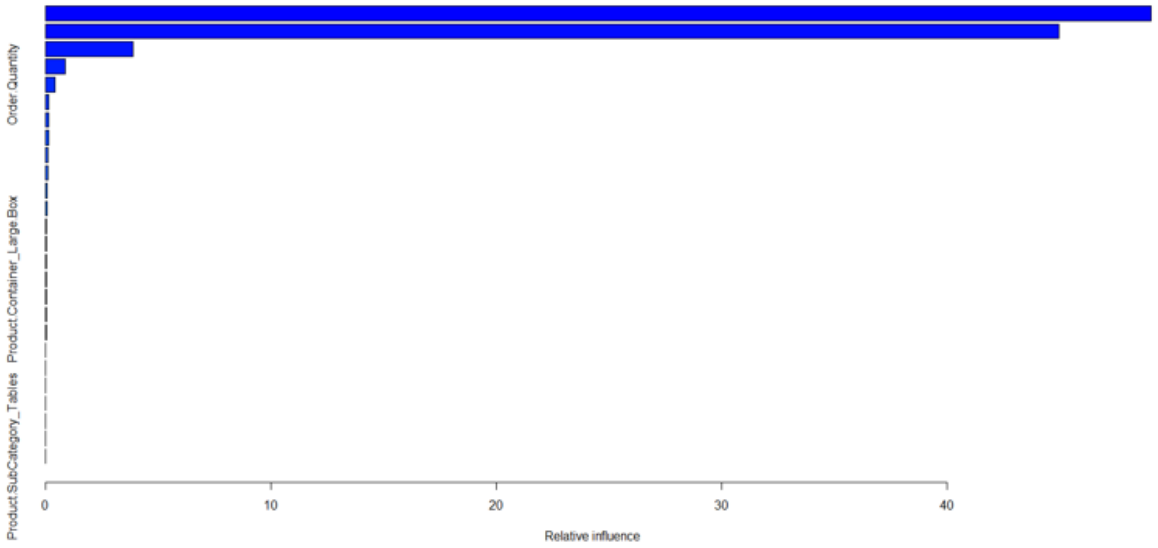
Descion Tree:



Random Forest:



Gradient Boosting:



Insights and Conclusion:

We have built various models to understand the factors which influence the choice of Shipping mode. Using models like RF and Gradient Boosting we found out that the most important factors are:

Sales
Order Quantity
Product Container Type – Jumbo Box
Product Container Type – Jumbo Drum

The model built using SVM multi-class classifier is the best model as testing accuracy is about 81% and it is able to predict 32 out of 230 test samples of Shipping mode Express Air. The model seems quite stable.