 

ASSIGNMENT – SCLA

Post-graduation in Data Science & Business Analytics

**By**

**Jithesh Janardhanan**

Contents

[1. Problem statement and Objective 2](#_Toc82654588)

[Variables used in this study 2](#_Toc82654590)

[Objective 2](#_Toc82654592)

[Assumptions: - 2](#_Toc82654593)

[2. Exploratory Data Analysis 2](#_Toc82654594)

[Univariate and Bivariate Analysis 3](#_Toc82654595)

[Highlights 3](#_Toc82654596)

[Bivariate Analysis 4](#_Toc82654597)

[Highlights 5](#_Toc82654598)

[Data Engineering/cleaning 5](#_Toc82654599)

[3. Model Building 6](#_Toc82654600)

[Multinomial Logistic regression 6](#_Toc82654601)

[Model Outputs 6](#_Toc82654602)

[Support Vector Machine 6](#_Toc82654603)

[Model outputs 7](#_Toc82654604)

[Decision Tree 7](#_Toc82654605)

[Model Outputs 8](#_Toc82654606)

[Random Forest 8](#_Toc82654607)

[4. Model Comparison 10](#_Toc82654608)

[5. Conclusion and findings 10](#_Toc82654609)

# Problem statement and Objective

# The data contain close to 8000 transactions of different products being shipped in different product containers of different sizes. Our objective here is to predict the appropriate shipping mode for different product subcategories based on the price, size, quantity, and product sub-category

## Variables used in this study

## The variables in this problem are order quantity and Sales as quantitative variables and Product name, Product container, Product sub-category and shipping mode as qualitative variables. We are considering shipping mode as the Dependent variable and all others as the predictor variables.

## Objective

The objective of this assignment is to build a model to predict the appropriate shipping mode for the any future transaction.

## Assumptions: -

* All the observations are independent from each other
* Samples are random
* All the measurements are accurate with respect to the data

# Exploratory Data Analysis

1. The given inventory data contains 7853 records with 8 columns
2. Below given is the structure and summary output of the data

Text, letter

Description automatically generated

1. We convert the categorical variables identified as characters in the summary into factors for analysis
2. Excluding Order date and Order ID, there are 2 quantitative variables and 4 categorical variables in the data.
3. There are no missing values in the dataset
4. We exclude the irrelevant variables such as Order ID, Order date and Product name from the analysis
5. We also replace the extreme values in the data set using the general principle of IQR

## Univariate and Bivariate Analysis

Below are the charts from univariate analysis

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Chart, pie chart, radar chart

Description automatically generated

Chart, pie chart

Description automatically generated

### Highlights

* Order Quantity looks like a uniform distribution.
* Sales follows a right skewed distribution with extreme values on both sides.
* Regular air contributes majority of the shipping mode
* More than half of the containers are small boxes

## Chart, box and whisker chart Description automatically generatedChart, box and whisker chart Description automatically generatedChart, bar chart Description automatically generatedBivariate Analysis

Timeline

Description automatically generated

Chart, box and whisker chart

Description automatically generatedChart

Description automatically generated

### Chart, box and whisker chart Description automatically generatedChart, box and whisker chart Description automatically generatedHighlights

* The dependent variable, Ship mode is analyzed against other variables in the data
* Regular air is the most preferred shipping mode except for Jumbo box and Jumbo drum where Delivery truck is preferred
* If we look at the product subcategories specifically, regular air is majorly the most preferred shipping mode
* Delivery truck sales are with the highest median value, while Regular air and Express air shipping modes have similar median value
* As we can expect, Jumbo box and Jumbo drum amount to big chunk of sales, as we see from the median value sales box plot of these containers
* Amongst the product subcategories, Copiers and Fax, Office Machines, Tables have high median sales values compared to others
* Order quantity does not differ much for different shipping modes. Same goes for container types and product categories as well

## Data Engineering/cleaning

* The outliers in the sales variable are capped to 5th and 95th percentile values
* The correlation amongst the independent variable seems to be minimal from the correlation values and plots

Graphical user interface, text, application

Description automatically generatedA picture containing diagram

Description automatically generated

* One hot encoding is done on independent categorical variables for analysis. For target variable we do encoding as 1 stand for Regular Air, 2 for Express Air and 3 for Delivery Truck.
* We treat the data imbalance in the dependent variable, using ROSE package to over sample the minority class (Express air and Delivery truck
* The data is split into train and test in the ratio 70:30 to build the model on Train data and predict on test data to check the model performance measures

# Model Building

## Multinomial Logistic regression

Here we use multinomial logistic regression. The algorithm allows us to predict a categorical dependent variable which has more than two levels. Like any other regression model, the multinomial output can be predicted using one or more independent variable.

The independent variable, Ship Mode is multilevel. 1 is for Regular Air, 2 for Express Air and 3 for Deliver Truck shipping mode. We are not scaling the data here as it is not creating any additional impact on the result

### Text Description automatically generatedText, letter Description automatically generatedModel Outputs

* Overall, the model looks fine with no over fitting
* The model predicts class 1 and class 3 with high precision, but the class 2 prediction accuracy is low (Low F1 score for class 2 prediction)
* The overall model accuracy is 83% which is decent.

## Support Vector Machine

We use support vector machine algorithm to build a model on the data set and predict to see if it is giving better results

### Text, letter Description automatically generatedModel outputs

A picture containing graphical user interface

Description automatically generated

* The model outputs better results than the LR method
* The overall model accuracy is 83%
* Class 1 and 3 has better F1 score, but class 2 F1 score is again low here, but higher than that of LR method
* Overall, the model looks fine with no over fitting

## Decision Tree

Decision tree is a supervised machine learning algorithm each node represents a predictor variable (feature), the link between the nodes represents a Decision and each leaf node represents an outcome (response variable)

A decision tree model is built on the imbalanced data set and tested on the test data to obtain the following results

### Model Outputs

Diagram, schematic

Description automatically generated

Text

Description automatically generatedText, letter

Description automatically generated

* No overfitting observed
* The model accuracy is pretty good which is 88%
* The model has got good F1 score for both class 1 and class 3 but if fails to predict the Class 2.
* Can’t consider this model because of its inefficiency to predict the class 2

## Random Forest

We build with all the variables on the ROSE data and test the model on train and test data

Random forest is an ensemble model of large number of decision trees. It follows the voting principle to select the class in a classification model and let us see the model outputs to interpret

Chart, line chart

Description automatically generatedText

Description automatically generated

Table

Description automatically generated

* Some bit of overfitting is observed in the train data set class prediction with very high accuracy
* The model looks good with 82% accuracy overall
* Like SVM model, the RF model predicts the class1 and class 3 with more precision and class 2 better than any other model we tried

# Model Comparison

* The following table gives the model performance measures for different models we ran so far
* Almost all the model we ran, were good in predicting the class1 and class3 with high accuracy and precision
* The models seem to be inefficient in predicting the calss2 with very low accuracy and F1 score
* This ties back to the initial EDA we did as the Express Air shipping mode was not clearly differentiated by the predictor variables
* Decision tree model is rejected because of its inefficiency to predict the class 2 at all
* SVM and RF model predicts the class 2 better than other models we ran but since we have more class 2 predicted correctly with the RF model, we go with the **RF modelling** on this data set for predicting the shipping mode
* Even with oversampling, it was difficult to predict the class 2, as it’s not clearly differentiated by the predictor variables

# Conclusion and findings

* As we ran various models to predict the accurate shipping mode for different transactions, we arrived at a conclusion that, Support Vector Machine learning algorithm was the best in differentiating different classes of shipping modes for different types of transactions

The important factors in predicting the shipping mode with RF modelling are

* Product Container Jumbo Drum
* Product Container Jumbo Box
* Sales
* Order Quantity

The modelling exercise we did predicts the Regular air and Delivery Truck shipping modes with high accuracy and precision

All the models we ran fails to predict the class 2 as the important factors that came out of our analysis doesn’t differentiate the Express Air shipping mode well.

We might need to collect more data to predict the class 2 more accurately. Also, the management need to wisely strategize some criteria for the Express air shipping mode. This would enhance the prediction of shipping more using a machine learning algorithm.