# Ecommerce Shipment Analysis

R Mukesh Kanna, C Manish, M Harshith

*UG Scholars, School of Computer Science and Technology*

VIT University, Vellore

***Abstract*:**

**It is important for an E-commerce company because customers care about shipping! You can provide the greatest shopping experience, the best prices, and outstanding customer service, but it's your shipping process that ultimately allows your customers to touch your products. While attempts have been made to build classifiers for products , not all algorithms have been tested. In addition to testing algorithms and variations of algorithms to check for highest accuracy for prediction purposes, we are also interested in analysing the data as well to conclude whether the product is delivered or not and the customer ratings is analysed.**

***Keywords: Machine Learning, Prediction, Classification, Artificial Neural Network.***

## I. Introduction

In this day and age, considering a shipping strategy is important. It's not just about choosing a supplier that will do shipping or logistics for fulfilment for your business's e-commerce orders. You need to keep track on the product whether it is delivered or not. As the product reaches on time to the customer, They prefer to place orders even on upcoming times and the ratings given by them will also be good. Customer rating is important of a company as that is the review for the new customers. Customer retention is much important so taking records on those shipment will help them. Forecasting the shipment may not immediately spring to mind as the most crucial part of shipment after all it isn't a tangible as the vessel. However without good forecasting, You could be left with cargo and no vessel to put it on – or vice versa. So accurate forecasting is actually a lot more influential than it may seem.

## II. Literature Review

A model to manage the shipment process was given by TCS in which the machine learning process was used to Introduce and equip the data analytics platform with machine learning and visualization features capable of on boarding huge volumes of out-of-the-box data. Identify the root cause of return orders by combining various data sources such as shipment, transportation, and order booking. Predict whether a shipment is likely to be delivered on-time using invoice, transportation, and shipment data.

The arrival of shipment in a seaway channel used machine learning classifiers and the methods used in it are such as the recorded pressure data are formed into the sample covariance matrices (SCMs) and then vectorized to generate the input data of the classifiers. Divide the preprocessed data into training and test data sets. The labels are designed for the training data using known GPS locations. Train the classifier models on the training data set. Predict the source ranges on test data set using the trained model parameters.

A research paper based on the delivery time estimation for industrial products using machine learning approach Exploratory Data Analysis the first step of our process involved data cleaning and deciding which variables to retain for the final modelling. Data Preparation Before altering the data any further, we wanted to add the necessary features that were not present, such as the distance variable. Data Partitioning and PCA In order to ensure that the models were not only trained correctly but also produced accurate prediction in the test set, on the data was divided into two subsets using a 80/20 split Given the large size of the product line data.

**Problem Statement:**

From the dataset provided by the E-commerce company, They need us to analyze whether the product has been delivered on time or not. And also study the customer ratings to improve customer retention.

## III. Proposed Method:

Data set:

The dataset taken for the shipment prediction of the E-commerce company are taken from the shipment repository. All the input variables may not be relevant to the shipment prediction. Description of the dataset is given below:

a)Input attributes:

1- ID

2- Warehouse block (A,B,C,D)

3- Mode of shipment (Road, Ship, Flight)

4- Customer care calls ( 1 to 6)

5- Customer Rating (1 to 5)

6- Cost of the Product (Price)

7- Prior purchases (1 to 6)

8- Product importance (Low, Medium, High)

9- Gender (Male, Female)

10- Discount Offered (in %)

11- Weight in grams (in grams)

12- Reached on time (1,0)

## IV. Data Processing Methods

For making automated decisions on model selection we need to quantify the performance of our model and give it a score. For that reason, for the classifiers, we are using F1 score which combines two metrics: Precision which expresses how accurate the model was on predicting a certain class and Recall which expresses the inverse of the regret of missing out instances which are misclassified. Since we have multiple classes we have multiple F1 scores.

Splitting for Testing:

The data has been split to train the models. Splitting is done in such a way that all of the data are represented equally.

Algorithms to be tested and trained:

1) Support Vector Machine.

SVM is a system gaining knowledge of set of rules which may be used for category and regression challenges. SVM are generally in category problems. They are truely coordinated of the man or woman observation.

2) Random Forest Classifier.

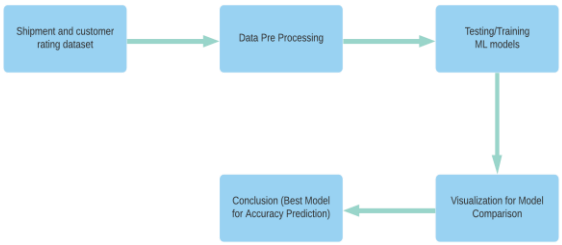
Random Forest Classifier are flexible, smooth to apply system gaining knowledge of set of rules which might be even with out hyper-parameter tuning a huge end result of the time. Random Forest Classifier are greater easy and variety so it's also one of the maximum used set of rules.

3) Artificial Neutral Network.

Artificial neuron simulates identical like how a organic neuron behaves through including collectively the values of the inputs it receives, It sends its very own sign to its output, that is then obtained through different neurons. A neuron doesn't should deal with every of its inputs with identical weight.

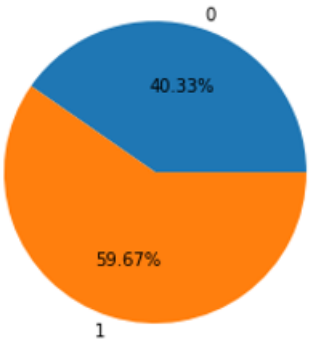
## V. Methodology Implementation

Proposed Flow Diagram:



Outputs:

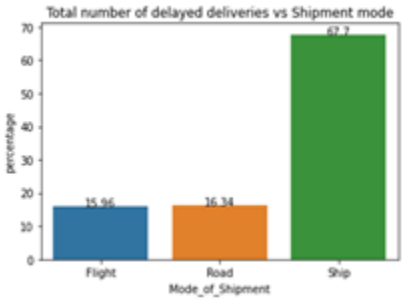
**i) Analysis on whether the product reached on time**




Observation:

- 40% of the deliveries are not reached on time

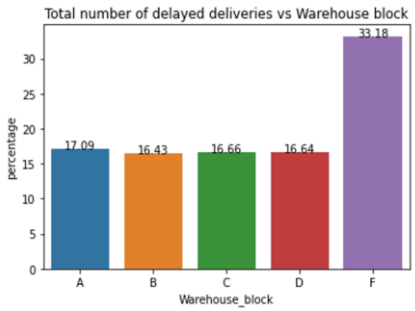
**ii)Total number of delayed deliveries vs the shipment mode**




Observation:

- Around 68% of the delayed deliveries are caused when ships are used as a mode of shipments. So, alternate options like Flight and Road services might be considered to reduce the delayed deliveries.

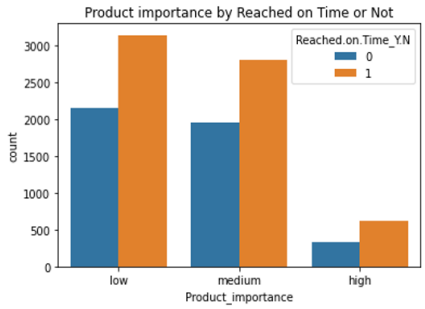
**iii) Total number of delayed deliveries vs the warehouse block**




Observation:

Higher percent of delayed deliveries are recorded in Warehouse Block F. For rest of the block, the percent of delayed deliveries are almost consistent.

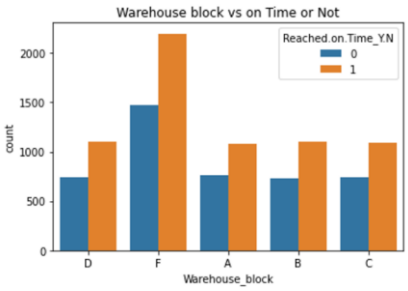
**iv)Product importance by reached on Time or not**




Observation:

Higher number of deliveries falls under low product importance. Very less number of highly importance products delivered. It means customers are ordering more number of low importance products from this ecommerce group

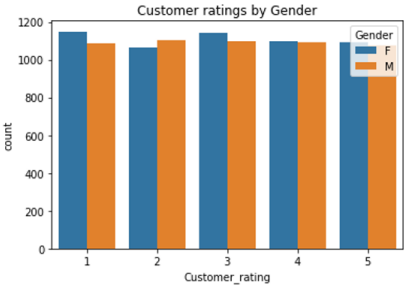
**v) Warehouse block count vs on Time or not**



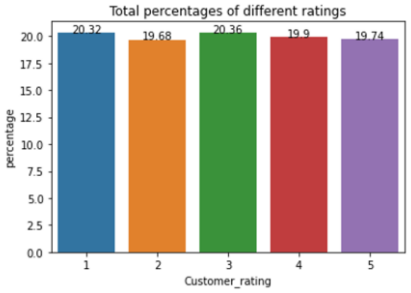

Observation:

- More number of deliveries were from Ware house block 'F'
- For rest of the blocks, the pattern remains same

**vi) Customer ratings by gender**



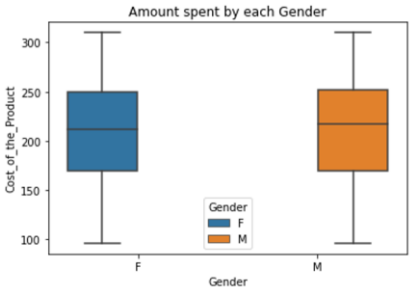
**vii) Total percentage of different rating**




Observation:

The percentage of different ratings given by customers seems to same. Almost 20% of the total deliveries received 5 ratings

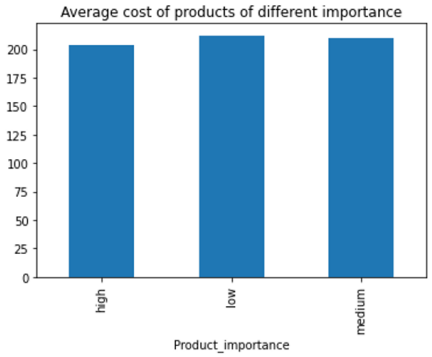
**viii) Amount spent by each gender**



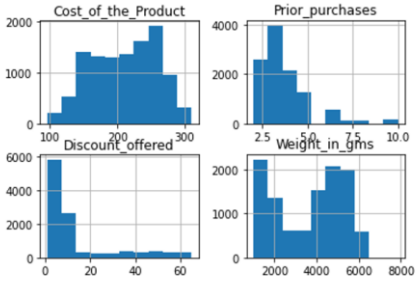

Observation:

Both men and women seem to spend same level of amount on a average while purchasing a product.

**ix) Average cost of products of different importance**




**x) Product details as 'Cost of_the_Product','Prior_purchases','Discount_offered ,'Weight_in_g ms:**




Observation:

- Discount_offered' is positively skewed
- Even 'prior purchase' is also positively skewed

## Results:

### 1) Support Vector Machine

Support vector machine gave us an accuracy of 66%

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.55 | 0.85 | 0.67 | 895 |
| 1 | 0.83 | 0.53 | 0.65 | 1305 |
| accuracy | | | 0.66 | 2200 |
| macro avg | 0.69 | 0.69 | 0.66 | 2200 |
| weighted avg | 0.72 | 0.66 | 0.66 | 2200 |

### 2) Random Forest Classifier

Random Forest Classifier gave us an accuracy of 66%

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.57 | 0.66 | 0.61 | 895 |
| 1 | 0.74 | 0.66 | 0.70 | 1305 |
| accuracy | | | 0.66 | 2200 |
| macro avg | 0.66 | 0.66 | 0.66 | 2200 |
| weighted avg | 0.67 | 0.66 | 0.66 | 2200 |

### 3) Artificial Neutral Network

Artificial Neutral Network gave us an accuracy of 67%

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.56 | 0.90 | 0.69 | 895 |
| 1 | 0.88 | 0.52 | 0.65 | 1305 |
| accuracy | | | 0.67 | 2200 |
| macro avg | 0.72 | 0.71 | 0.67 | 2200 |
| weighted avg | 0.75 | 0.67 | 0.67 | 2200 |

## VI. Conclusion

Based on the performance matrices displayed, we can conclude the fact that for the shipment prediction Artificial Neutral Network has proven to give the best results. Upon estimating the feature that the Artificial Neutral Network takes into account, it can be seen that feature like product delivered on time or not and the customer ratings.

## VII. References


1)Advances in Shipping Data Analysis and Modeling: Tracking and Mapping Maritime Flows in the Age of Big Data (Routledge Studies in Transport Analysis) 1st Edition by César Ducruet.

2)Data Visualization with Python: Create an impact with meaningful data insights using interactive and engaging visuals 28 February 2019 by Mario Dobler.

3)https://dspace.mit.edu/bitstream/handle/1721.1/121280/Jonquais_Krempl_2019.pdf?sequence=1&isAllowed=y Predicting Shipping Time with Machine Learning by Antoine Charles Jean Jonquais

4)https://www.analyticsvidhya.com/blog/2018/04/introduction-to-graph-theory-network-analysis-python-codes/

5)https://www.machinelearningplus.com/plots/top-50-matplotlib-visualizations-the-master-plots-python/