CUSTOMER CHURN REDUCTION USING SUPPLY CHAIN ANALYTICS

PGPDSE FT CAPSTONE PROJECT – FINAL REPORT – BATCH 3

PROJECT REPORT SUBMITTED BY

ARUN RICHARD A

BINOY GEORGE

DHANUSH KH

RISHABH BHARDWAJ

SANJAY DEY

GUIDED BY

MR.SUBRAMANIAN P V

CERTIFICATE

Certified that the project report "**CUSTOMER CHURN REDUCTION USING SUPPLY CHAIN ANALYTICS"** is the bonafide work of group3 who carried out the project work under my supervision during the academic year 2021.

SIGNATURE

Mr. SUBRAMANIAN P V

Project Mentor

CONTENTS:

1. ACKNOWLEDGEMENT:

The completion of this undertaking could not have been possible without the participation and assistance of so many people whose names may not all be enumerated. Their contribution are sincerely appreciated and gratefully acknowledged. However, the group would like to express their deep appreciation and indebtedness particularly to the following:

Mr.Subramanian P V, for their endless support, kind and understanding spirit during throughout our learning.

To all relatives, friends and others who in one way or another shared their support, either morally, financially and physically, thank you.

Above all, to the Great Almighty, the author of knowledge and wisdom , for his countless love.

We thank you.

Group 3 Members.

2. PROBLEM STATEMENT:

Company has seen a large drop in sales in the year 2018 after succeeding in previous 3 years due to increase in customer churn. Company needs to reduce customer churn by understanding the reasons for loss of shipment orders.We have observed that the majority of orders were delivered late.so in this project we are taking 'Late_delivery_risk' column as target which is a categorical column with 0 and 1 where if the order is delivered late (1) and (0) if the order is delivered on time. We would be using the Supervised Learning –Classification to predict late delivery risk. The company can use this information to make sure that they deliver the product on time and decrease customer dissatisfaction due to late delivery.

3. OVERVIEW OF THE FINAL PROCESS:

A DataSet of Supply Chains used by the company DataCo Global was used for the analysis. Dataset of Supply Chain, which allows the use of Machine Learning Algorithms and pyhton Software. Areas of important registered activities: Provisioning, Production, Sales, Commercial Distribution. It also allows the correlation of Structured Data with Unstructured Data for knowledge generation. Below is the taxonomy diagram of the dataset to understand the dataset.

There are totally 53 variables and 180519 observations with total of 9567507 data points, there are 29 numerical features and there are 24 categorical features

The name of the columns were as below

'Type', 'Days for shipping (real)', 'Days for shipment (scheduled)','Benefit per order', 'Sales per customer', 'Delivery Status', 'Late_delivery_risk', 'Category Id', 'Category Name', 'Customer City',
'Customer Country', 'Customer Email', 'Customer Fname', 'Customer Id', 'Customer Lname', 'Customer Password', 'Customer Segment','Customer State', 'Customer Street', 'Customer Zipcode', 'Department Id', 'Department Name', 'Latitude', 'Longitude', 'Market','Order City', 'Order Country', 'Order Customer Id', 'order date (DateOrders)', 'Order Id', 'Order Item Cardprod Id', 'Order Item Discount', 'Order Item Discount Rate', 'Order Item Id', 'Order Item Product Price', 'Order Item Profit Ratio', 'Order Item Quantity', 'Sales', 'Order Item Total','Order Profit Per Order', 'Order Region', 'Order State', 'Order Status', 'Order Zipcode', 'Product Card Id', 'Product Category Id','Product Description', 'Product Image', 'Product Name', 'Product Price','Product Status', 'shipping date (DateOrders)', 'Shipping Mode'

'Late_delivery_risk' is our target variable as per our problem statement.

There are couple of columns with have missing values which is as below

- Customer Lname has 8 missing values which .004%
- Customer Zipcope has 3 missing values which is .002%

- Order Zipcode has 155679 missing values which is 86.24%
- Product Description has 180519 missing values which is 100%

Few of the columns which was redundant and which was not giving important information for our problem statement as per our initial data analysis which was removed for our further data analysis was as below

'Category Id','Customer Email','Customer Fname','Customer Lname','Customer Password', 'Customer Street', 'Customer Zipcode','Department Id','Latitude','Longitude', 'Order Item Cardprod Id', 'Order Item Discount','Order Item Profit Ratio','Order Zipcode', 'Product Card Id', 'Product Description', 'Product Image','Product Price', 'Order Customer Id', 'Benefit per order', 'Sales per customer', 'Order Item Id'

After removing the irrelevant variables we were left with 31 variables as below

'Type','Days for shipping (real)','Days for shipment (scheduled)','Delivery Status','Late_delivery_risk','Category Name','Customer City','Customer Country','Customer Id','Customer Segment','Customer State','Department Name','Market','Order City','Order Country','order date (DateOrders)','Order Id','Order Item Discount Rate','Order Item Product Price','Order Item Quantity','Sales','Order Item Total','Order Profit Per Order','Order Region' , 'Order State','Order Status','Product Category Id','Product Name','Product Status','shipping date (DateOrders)','Shipping Mode'

We have scaled few of the numeric variable using standard scalar columns like ' Order Item Discount Rate','Order Item Product Price','Sales','Order Item Total','Order Profit Per Oder'

We have used different classification algorithms to understand the relationship of the target variables with other variables such as decision tree,Logistic Regression, Knn model, Random forest, ada boost, gradient boost based on youden index and hyper parameters/

4. STEP-BY-STEP WALK THROUGH OF THE SOLUTION:

INTRO TO DATASET:

A DataSet of Supply Chains used by the company DataCo Global was used for the analysis. Dataset of Supply Chain, which allows the use of Machine Learning Algorithms and pyhton Software. Areas of important registered activities: Provisioning, Production, Sales, Commercial Distribution. It also allows the correlation of Structured Data with Unstructured Data for knowledge generation. Below is the taxonomy diagram of the dataset to understand the dataset.

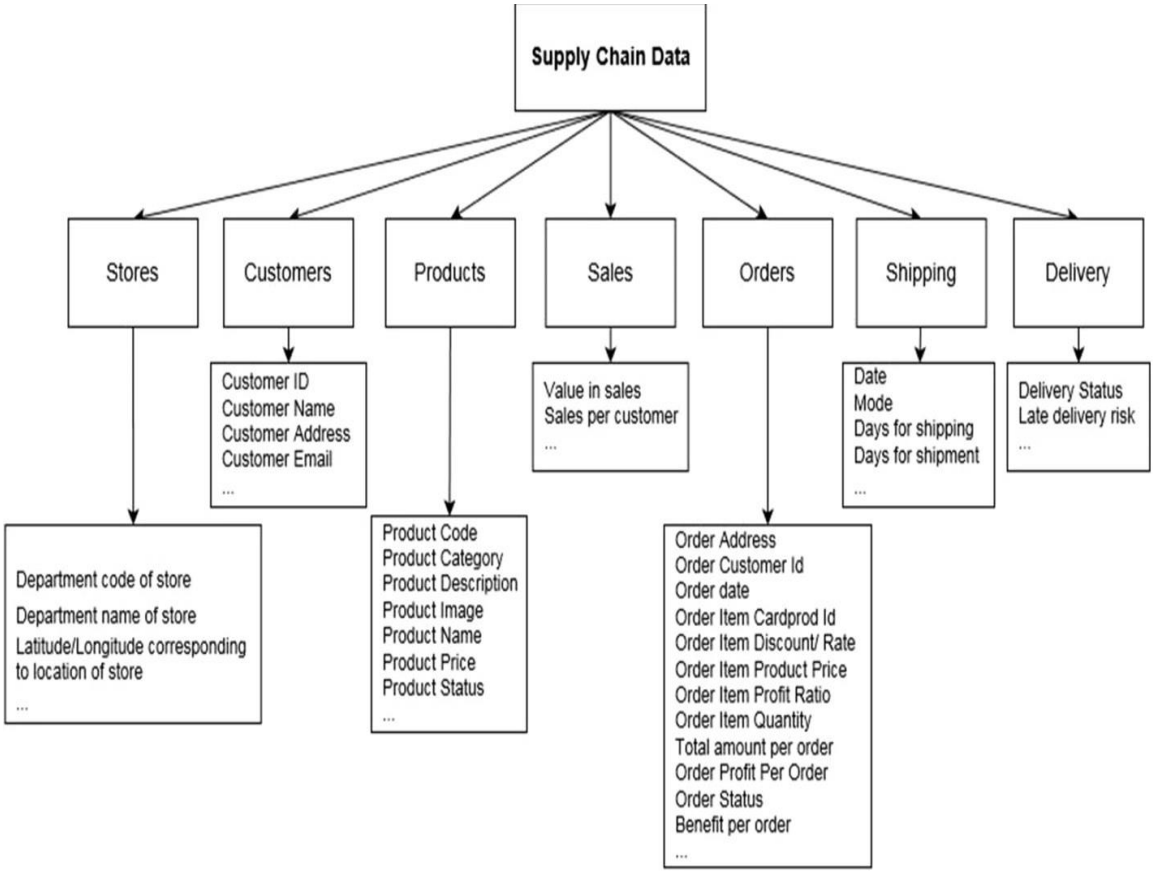


Fig 4.I.1.Taxonomy of the Dataset

Dataset has information about store like Department code, name and there location, details of the customers, details of the product which is been shipped, details about the sales happened, details about the orders, mode of shipping and details on the status of the delivery. Below are the list of columns with its description.

1) Type : Type of transaction made
2) Days for shipping (real) : Actual shipping days of the purchased product
3) Days for shipment (scheduled) : Days of scheduled delivery of the purchased product
4) Benefit per order : Earnings per order placed
5) Sales per customer : Total sales per customer made per customer
6) Delivery Status : Delivery status of orders: Advance shipping , Late delivery , Shipping canceled , Shipping on time
7) Late_delivery_risk : Categorical variable that indicates if sending is late (1), it is not late (0).
8) Category Id : Product category code
9) Category Name : Description of the product category
10) Customer City : City where the customer made the purchase
11) Customer Country : Country where the customer made the purchase
12) Customer Email : Customer's email
13) Customer Fname : Customer name
14) Customer Id : Customer ID
15) Customer Lname : Customer lastname
16) Customer Password : Masked customer key
17) Customer Segment : Types of Customers: Consumer , Corporate , Home Office
18) Customer State : State to which the store where the purchase is registered belongs

19) Customer Street : Street to which the store where the purchase is registered belongs

20) Customer Zipcode : Customer Zipcode

21) Department Id : Department code of store

22) Department Name : Department name of store

23) Latitude : Latitude corresponding to location of store

24) Longitude : Longitude corresponding to location of store

25) Market : Market to where the order is delivered : Africa , Europe , LATAM , Pacific Asia , USCA

26) Order City : Destination city of the order

27) Order Country : Destination country of the order

28) Order Customer Id : Customer order code

29) order date (DateOrders) : Date on which the order is made

30) Order Id : Order code

31) Order Item Cardprod Id : Product code generated through the RFID reader

32) Order Item Discount : Order item discount value

33) Order Item Discount Rate : Order item discount percentage

34) Order Item Id : Order item code

35) Order Item Product Price : Price of products without discount

36) Order Item Profit Ratio : Order Item Profit Ratio

37) Order Item Quantity : Number of products per order

38) Sales : Value in sales

39) Order Item Total : Total amount per order

40) Order Profit Per Order : Order Profit Per Order

41) Order Region : Region of the world where the order is delivered : Southeast Asia ,South Asia ,Oceania ,Eastern Asia, West Asia , West of USA , US Center , West Africa, Central Africa ,North Africa ,Western Europe ,Northern , Caribbean , South America ,East Africa ,Southern Europe , East

of USA ,Canada ,Southern Africa , Central Asia , Europe , Central America, Eastern Europe , South of USA

42) Order State : State of the region where the order is delivered

43) Order Status : Order Status : COMPLETE , PENDING , CLOSED , PENDING_PAYMENT ,CANCELED , PROCESSING ,SUSPECTED_FRAUD ,ON_HOLD ,PAYMENT_REVIEW

44) Order Zipcode : Zipcode of the region where the order is delivered

45) Product Card Id : Product code

46) Product Category Id : Product category code

47) Product Description : Product Description

48) Product Image : Link of visit and purchase of the product

49) Product Name : Product Name

50) Product Price : Product Price

51) Product Status : Status of the product stock :If it is 1 not available , 0 the product is available

52) Shipping date (DateOrders) : Exact date and time of shipment

53) Shipping Mode : The following shipping modes are presented : Standard Class , First Class , Second Class , Same Day

TARGET VARIABLE:

In this project we are taking 'Late_delivery_risk' column as target which is a categorical column with 0 and 1 where if the order is delivered late (1) and (0) if the order is delivered on time

VARIABLE CATEGORIZATION:

There are totally 53 variables and 180519 observations with total of 9567507 data points, there are 29 numerical features and there are 24 categorical features.

PRE PROCESSING DATA ANALYSIS:

There are couple of columns with have missing values which is as below

- Customer Lname has 8 missing values which .004%
- Customer Zipcope has 3 missing values which is .002%
- Order Zipcode has 155679 missing values which is 86.24%
- Product Description has 180519 missing values which is 100%

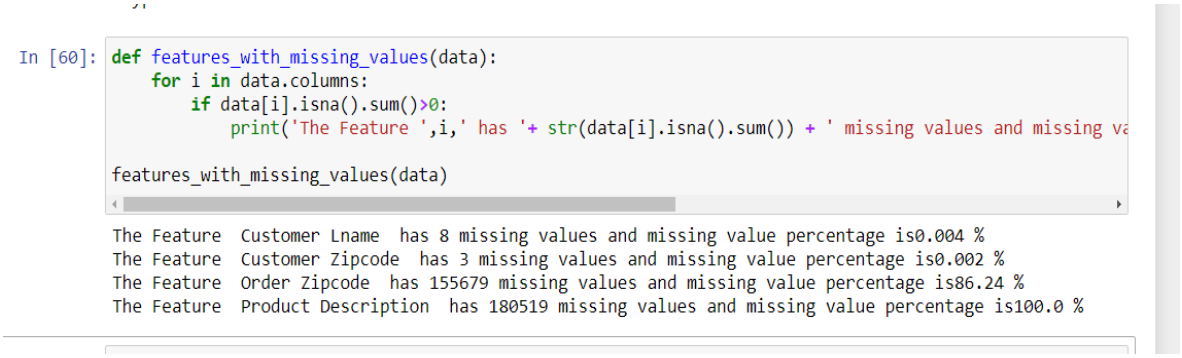


Fig 4.IV.1. Missing Values

Few of the redundant columns are as follows:

- Category Id,Customer Email,Customer Fname,'Customer Lname', 'Customer Password' ,'Department Id', 'Customer Zipcode', 'Product Card Id' - These are all customer specific unique identifier and department identifier which are not helpful for analysis.
- Order Item Id', 'Order Item Cardprod Id', 'Order Zipcode' 'Product Image' - These are all product specific unique identifiers which are not helpful for analysis.
- "Order Customer Id" is exactly same as customer id.
- "Benefit per order" is exactly like 'Order Profit Per Order' so we keep only the later column
- "Sales per customer" is exactly like Order Item Total.
- 'Product Description' is more than 85% of records are null in these variables.

- 'Customer Street' 'Latitude' 'Longitude' *redundent columns which would give similar information as other columns like customer city and order city. *E*ven if we retain these columns, we would need to do feature engineeing operations on these to produce columns with #low number of levels similar to customer country and order country.
- 'Order Item Discount' #derived column from order_item_discount_rate which is not required for our analysis.
- 'Order Item Profit Ratio' *is* derived from Order Profit Per Order and Order Item Total(Order Profit Per Order/Order Item Total) which is not required for our analysis.
- 'Product Price' id redundant column which is like Order Item Product Price

'order date (DateOrders)' and 'shipping date (DateOrders)' these two columns have been used to derive days for shipping(real) and days for shipping(scheduled)

DATA EXPLORATION (EDA):

RELATIONSHIP BETWEEN VARIABLES:

After deleting the redundant columns discussed above we have 30 columns available and there are no null values. Now we have totally 30 variables and 180519 observations with total of 5415570 data points, there are 12 numerical features and there are 18 categorical features.

We have checked the variance of the remaining numerical columns and variance of product status show 0. Since there no feature we can get from this column we delete it

```
In [21]: df.std()

Out[21]: Days for shipping (real)            1.623722
         Days for shipment (scheduled)       1.374449
         Late_delivery_risk                  0.497664
         Customer Id                      4162.918106
         Order Id                        21045.379569
         Order Item Discount Rate            0.070415
         Order Item Product Price          139.732492
         Order Item Quantity                 1.453451
         Sales                             132.273077
         Order Item Total                  120.043670
         Order Profit Per Order            104.433526
         Product Status                      0.000000
         dtype: float64
```

Fig 4.V.1. Standard Deviation

FIVE POINT SUMMARY GAVE US BELOW OBSERVATION:

A five-number summary is especially useful in descriptive analyses or during the preliminary investigation of a large data set. A summary consists of five values: the most extreme values in the data set (the maximum and minimum values), the lower and upper quartiles, and the median. These values are presented together and ordered from lowest to highest: minimum value, lower quartile (Q1), median value (Q2), upper quartile (Q3), maximum value.

These values have been selected to give a summary of a data set because each value describes a specific part of a data set: the median identifies the centre of a data set; the upper and lower quartiles span the middle half of a data set; and the highest and lowest observations provide additional information about the actual dispersion of the data. This makes the five-number summary a useful measure of spread.

```
In [57]: df.describe().T
Out[57]:
```

	count	mean	std	min	25%	50%	75%	max
Days for shipping (real)	180519.0	3.497654	1.623722	0.00000	2.000000	3.000000	5.000000	6.000000
Days for shipment (scheduled)	180519.0	2.931847	1.374449	0.00000	2.000000	4.000000	4.000000	4.000000
Late_delivery_risk	180519.0	0.548291	0.497664	0.00000	0.000000	1.000000	1.000000	1.000000
Customer Id	180519.0	6691.379495	4162.918106	1.00000	3258.500000	6457.000000	9779.000000	20757.000000
Order Id	180519.0	36221.894903	21045.379569	1.00000	18057.000000	36140.000000	54144.000000	77204.000000
Order Item Discount Rate	180519.0	0.101668	0.070415	0.00000	0.040000	0.100000	0.160000	0.250000
Order Item Product Price	180519.0	141.232550	139.732492	9.99000	50.000000	59.990002	199.990005	1999.989990
Order Item Quantity	180519.0	2.127638	1.453451	1.00000	1.000000	1.000000	3.000000	5.000000
Sales	180519.0	203.772096	132.273077	9.99000	119.980003	199.919998	299.950012	1999.989990
Order Item Total	180519.0	183.107609	120.043670	7.49000	104.379997	163.990005	247.399994	1939.989990
Order Profit Per Order	180519.0	21.974989	104.433526	-4274.97998	7.000000	31.520000	64.800003	911.799988

Fig 4.V.2 Five point Summary

- ‘Type’- there are 4 types of transactions that happened and majority of them were by using debit card
- ‘Days for shipping (real)’ and ‘Days for shipment (scheduled)’-the average actual shipment time is more than the average scheduled shipment time.
- 50% of the shipment delivery time is more the average schedule shipment time.
- Delivery Status-It is observed that more than 50% of the orders are under late delivery order status only
- ‘Order Profit Per Order’ has negative values which might be either an anomaly or is a major loos bearing order, we have to check accordingly and data is skewed.
- ‘Order Item Total’ also seems like skewed data

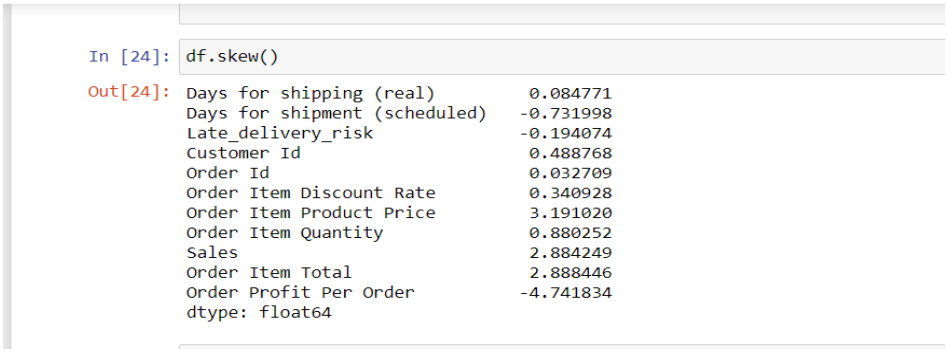
CHECKING SKEWNESS OF OBSERVATIONS:

Skewness refers to a distortion or asymmetry that deviates from the symmetrical bell curve, or normal distribution, in a set of data. If the curve is shifted to the left or to the right, it is said to be skewed.

Skewness can be quantified as a representation of the extent to which a given distribution varies from a normal distribution. A normal distribution has a skew of zero, while a lognormal distribution.

Besides positive and negative skew, distributions can also be said to have zero or undefined skew. In the curve of a distribution, the data on the right side of the curve may taper differently from the data on the left side. These taperings are known as "tails." Negative skew refers to a longer or fatter tail on the left side of the distribution, while positive skew refers to a longer or fatter tail on the right.

The mean of positively skewed data will be greater than the median. In a distribution that is negatively skewed, the exact opposite is the case: the mean of negatively skewed data will be less than the median. If the data graphs symmetrically, the distribution has zero skewness, regardless of how long or fat the tails are.



```
In [24]: df.skew()

Out[24]: Days for shipping (real)         0.084771
         Days for shipment (scheduled)   -0.731998
         Late_delivery_risk              -0.194074
         Customer Id                      0.488768
         Order Id                         0.032709
         Order Item Discount Rate         0.340928
         Order Item Product Price         3.191020
         Order Item Quantity              0.880252
         Sales                            2.884249
         Order Item Total                 2.888446
         Order Profit Per Order          -4.741834
         dtype: float64
```

Fig 4.V.3 Skewness of Numerical features

Data seem to be skewed below is the interpretation

Name of Column	Skew	Interpretation
Days for shipping (real)	0.084771	Data looks Normal
Days for shipment (scheduled)	-0.732	Data looks Left skewed
Late_delivery_risk	-0.19407	Data looks slightly left skewed
Order Item Discount Rate	0.340928	Data looks slightly right skewed
Order Item Product Price	3.19102	Data looks heavily right skewed
Order Item Quantity	0.880252	Data looks slightly right skewed
Sales	2.884249	Data looks heavily right skewed
Order Item Total	2.888446	Data looks heavily right skewed
Order Profit Per Order	-4.7418	Data looks heavily Left skewed

Table 4.V.1 Checking skewness of observation

CHECK FOR MULTI-COLLINEARITY:

Multicollinearity is the occurrence of high intercorrelations among two or more independent variables in a multiple regression model. Multicollinearity can lead to skewed or misleading results when a researcher or analyst attempts to determine how well each independent variable can be used most effectively to predict or understand the dependent variable in a statistical model.

In general, multicollinearity can lead to wider confidence intervals that produce less reliable probabilities in terms of the effect of independent variables in a model. That is, the statistical inferences from a model with multicollinearity may not be dependable. Multicollinearity in a multiple regression model indicates that collinear independent variables are related in some fashion, although the relationship may or may not be casual.

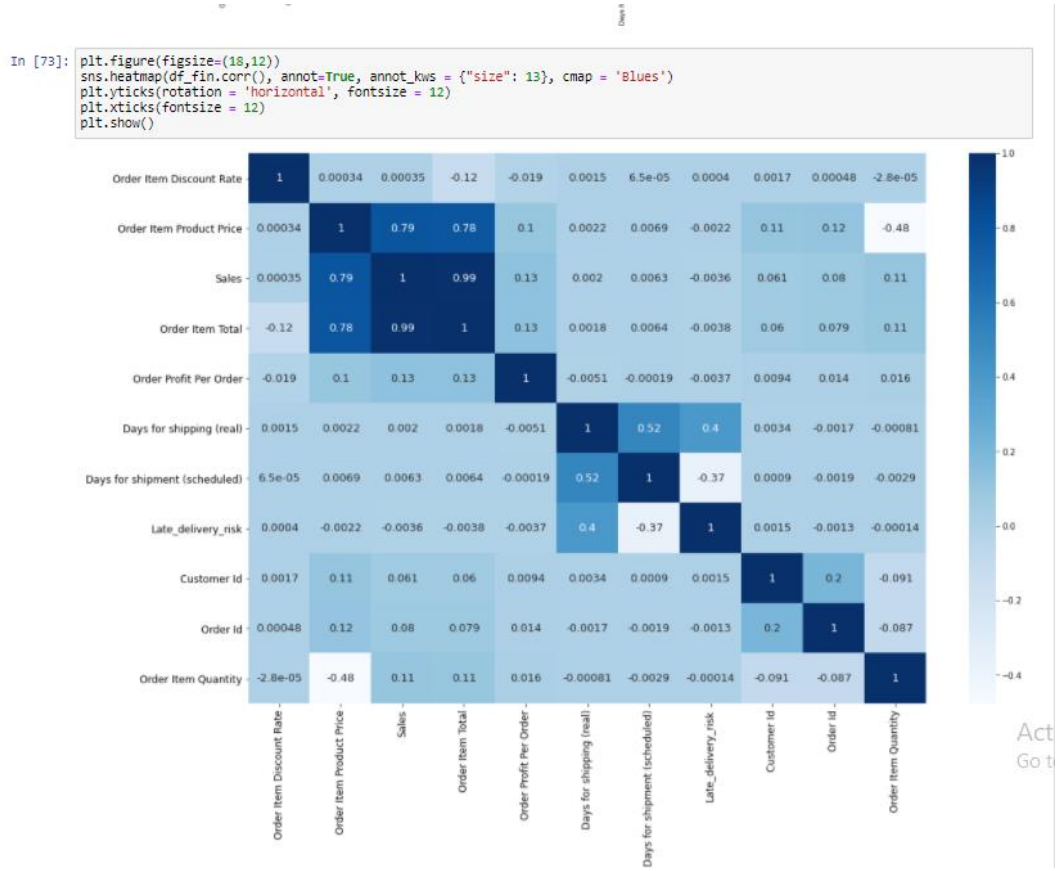


Fig 4.V.4. Correlation of Numerical features

From the above heat map below are the inference

- Days for shipping (real) and Days for shipment (scheduled) has a correlation of .515
- Order Item Product Price and Sales has a correlation of .789
- Order Item Product Price and Order Item Total has a correlation of .781
- Sales and Order Item Total has a correlation of .989 we would be dropping Order Item Total going forward.

DISTRIBUTION OF VARIABLES:

A sample of data will form a distribution, and by far the most well-known distribution is the Gaussian distribution, often called the Normal distribution.

The distribution provides a parameterized mathematical function that can be used to calculate the probability for any individual observation from the sample space. This distribution describes the grouping or the density of the observations, called the probability density function. We can also calculate the likelihood of an observation having a value equal to or lesser than a given value. A summary of these relationships between observations is called a cumulative density function.

A probability distribution is a statistical function that describes all the possible values and likelihoods that a random variable can take within a given range. This range will be bounded between the minimum and maximum possible values, but precisely where the possible value is likely to be plotted on the probability distribution depends on a number of factors. These factors include the distribution's mean (average), standard deviation, skewness, and kurtosis.

Perhaps the most common probability distribution is the normal distribution, or "bell curve," although several distributions exist that are commonly used. Typically, the data

generating process of some phenomenon will dictate its probability distribution. This process is called the probability density function.

Probability distributions can also be used to create cumulative distribution functions (CDFs), which adds up the probability of occurrences cumulatively and will always start at zero and end at 100%.

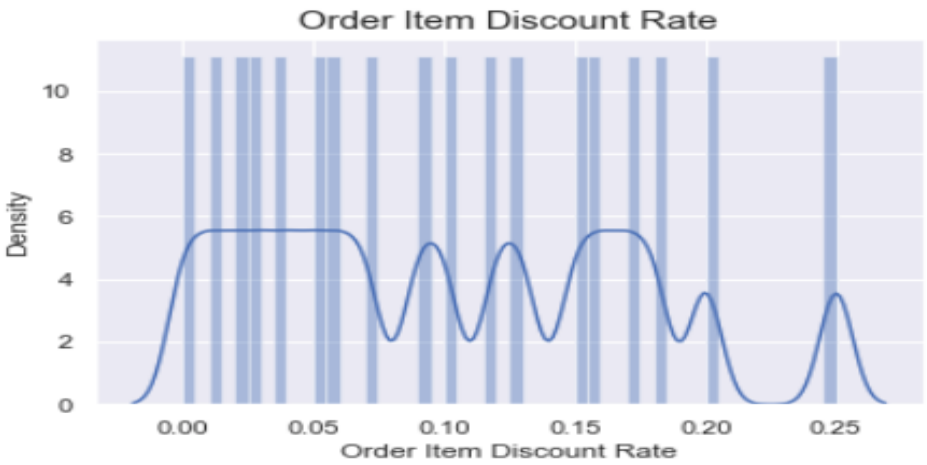


Fig 4.V.5. Distplot of Order Item Discount Rate

The above distribution is almost flat for 5 density which means that the various order discount rates are of equal numbers.

Fig 4.V.6. Distplot of Order Item Product Price

Order item product price is right skewed as there are lot of items of high price which increases the mean of the order item product price. Most of the order items have price between 0 to 500 rupees. Density of the order items around 100 rupees is really high.

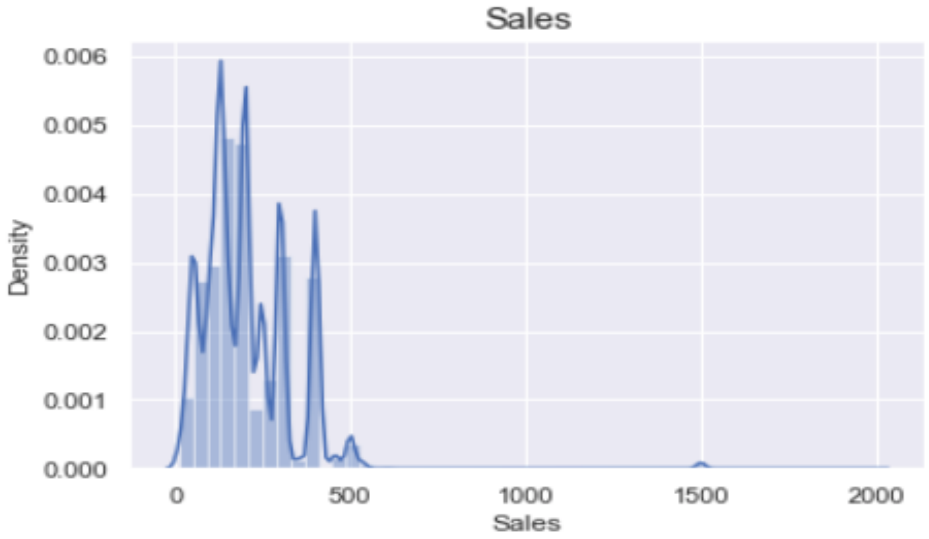


Fig 4.V.7. Distplot of Sales

The sales distribution is also right skewed as there are customers giving larger sales comparatively. Most of the customers are producing sales between 0 to 500 rupees. The mean sales is 203 rupees.

Fig 4.V. 8. Distplot of Order Item Total

Order item total is also right skewed as there are lot of customers ordering items of higher price comparatively. Most of the customers order items with total between 0 to 500. The mean of order item total is 183 rupees which is less than sales mean by 20 rupees. So company needs to reduce discount per order.

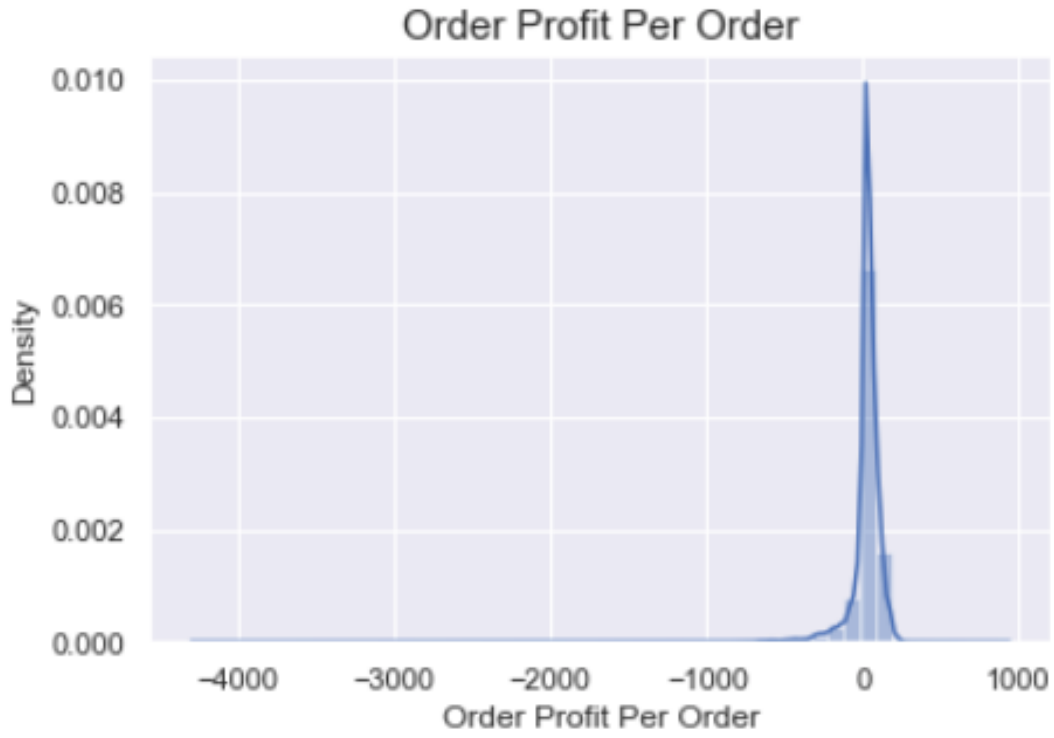


Fig 4.V.9. Distplot of Order Profit Per Order

The order profit per order is left skewed which means there are lot of customers causing loss to the company and the mean order profit per order is around 21. So company needs to make sure that they cut off the customers causing large number of order loss per order.

STATISTICAL SIGNIFICANCE OF VARIABLES:

Statistical significance is a determination by an analyst that the results in the data are not explainable by chance alone. Statistical hypothesis testing is the method by which the analyst makes this determination. This test provides a p-value, which is the probability of observing results as extreme as those in the data, assuming the results are truly due to chance alone. A p-value of 5% or lower is often considered to be statistically significant.

Statistical significance is a determination about the null hypothesis, which hypothesizes that the results are due to chance alone. A data set provides statistical significance when the p-value is sufficiently small.

When the p-value is large, then the results in the data are explainable by chance alone, and the data are deemed consistent with (while not proving) the null hypothesis.

When the p-value is sufficiently small (e.g., 5% or less), then the results are not easily explained by chance alone, and the data are deemed inconsistent with the null hypothesis; in this case, the null hypothesis of chance alone as an explanation of the data is rejected in favor of a more systematic explanation.

Analysis on Categorical Data is show in below table

Categorical Column 1	Categorical Column 2	Type of testing	Inference
Type	Days for shipping (real)	One way Anova testing and post-hoc analysis	That different payment transaction type has different average number of shipping days so there is no connection between them.
Order Status	Category Name	Using crosstab and then chi2_contingency	The highest number of orders cancelled are the ones which ordered the most.
Order Status	Shipping Mode	Using crosstab and then chi2_contingency	The majority of cancelled order are the ones which were ordered under standard shipping mode.
Late_delivery_risk	Market	Using crosstab and then chi2_contingency	That market location has nothing to do with all the late deliveries.
Late_delivery_risk	Customer Segment	Using crosstab and then chi2_contingency	That orders were late irrespective of their segments.
Late_delivery_risk	Department Name	Using crosstab and then chi2_contingency	Late deliveries are same for all different different departments as the p_value is greater than 0.05.
Late_delivery_risk	Order Region	Using crosstab and then chi2_contingency	late delivery and order region are dependent.
Late_delivery_risk	Order Item Quantity	Using crosstab and then chi2_contingency	Number of items ordered is independent of late deliveries
Late_delivery_risk	Customer City	Using crosstab and then chi2_contingency	late delivery is not independent of customer city
Late_delivery_risk	Order City	Using crosstab and then chi2_contingency	order city is not independent of late delivery status.
Late_delivery_risk	Order State	Using crosstab and then chi2_contingency	some non-independence present among order state and late delivery status
Late_delivery_risk	Product Name	Using crosstab and then chi2_contingency	that product names are independent of late delivery status.

Table 4.V.2 Statistical Signficance of Variable

PRESENCE OF OUTLIERS AND ITS TREATMENT:

Most real-world datasets include a certain amount of anomalous values, generally termed as ‘outliers’. These observations substantially deviate from the general trend therefore, it is important to isolate these outliers for improving the quality of original data and reducing the adverse impact they have in the process of analyzing datasets. Practically, nearly all experimental data samples are likely to be contamination by outliers which reduce the efficiency, and reliability of statistical methods. Outliers are analyzed to see if their unusual behavior can be explained. Sometimes outliers have “bad” values occurring as a result of unusual but explainable events. The cause of outliers are not always random or chance. Therefore a study needs to be made before an outlier is discarded.

Detection of Statistical Outliers

Statistical outliers are more common in distributions that do not follow the normal distribution. For example, in a distribution with a long tail, the presence of statistical outliers is more common than in the case of a normal distribution.

The simplest method of identifying whether an extreme value is an outlier is by using the interquartile range. The IQR tells us how spread out the middle half of our data set is.The interquartile range, or IQR, is determined by subtracting the first quartile from the third quartile

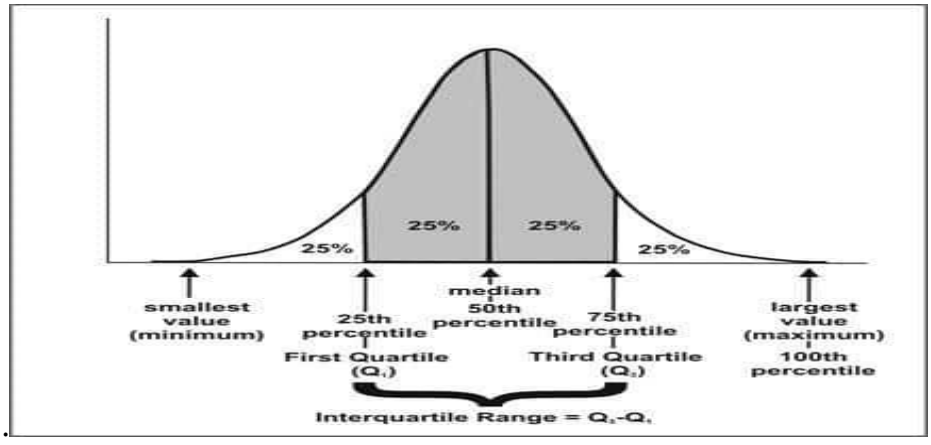


Fig 4.V.10. Plot explaining IQR

We start with the IQR and multiply it by 1.5. Then subtract this number from the first quartile and add this number to the third quartile. These two numbers from our **inner fence**. For the outer fences, we start with the IQR and multiply it by 3. We then subtract this number from the first quartile and add it to the third quartile. These two numbers are our **outer fences**.

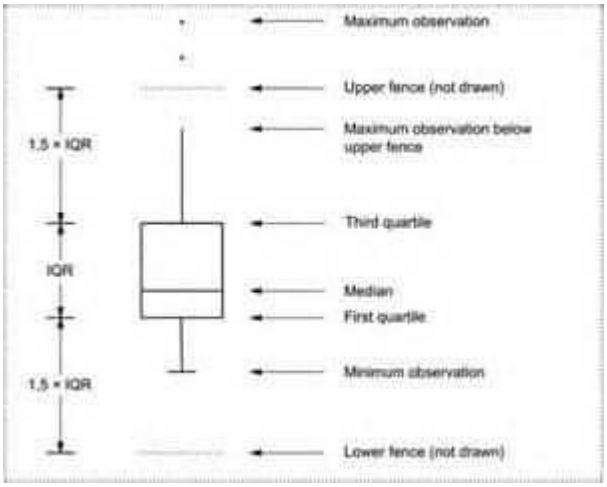


Fig 4.V.11. Explaining IQR with Boxplot

Outliers can now be detected by determining where the observation lies in reference to the inner and outer fences. If a single observation is more extreme than either of our outer fences, then it is an outlier, and more particularly referred to as a **strong outlier**. If our data value is between corresponding inner and outer fences, then this value is a suspected outlier or a **weak outlier**.

Reasons for Identifying Outliers

The presence of outliers indicates errors in measurement or the occurrence of an unexpected and previously unknown phenomenon. It is extremely important to check for outliers in every statistical analysis as they have an impact on all the descriptive statistics, as they are sensitive to them. The mean, standard deviation and correlation coefficient in paired data are just a few of these types of statistics. This could mislead analysts into making incorrect insights as all these statistics get distorted.

Treatment of Outliers:

The treatment of outlier values can be achieved by the following categories of actions that can be taken:

1. **Transformation of Data**: Transformation data is one way to soften the impact of outliers since the most commonly used expressions, square root and logarithms, affect larger numbers to a much greater extent than they do the smaller ones. Transformations may not fit into the theory of the model all the time as they may affect its interpretation. Transforming a variable does more than make a distribution less skewed; it changes the relationship between the variables in the model.
2. **Deletion of Values:** When there are legitimate errors and cannot be corrected, or lie so far outside the range of the data that they distort statistical inferences the outliers should be deleted. When in doubt, we can report model results both with and without outliers to see how much they change. Data transformation and deletion are important tools, but they should not be viewed as an all-out for distributional problems associated with outliers. Transformations and/or outlier elimination should be an informed choice, not a routine task. In some cases, the removal of an outlier value can also induce incorrect inferences made about the data. In such cases, replacing the observation with a measure of central tendency (Mean, Median or Mode), depending on the situation.
3. **Accommodation of Values:** One very effective plan is to use methods that are robust in the presence of outliers. Nonparametric statistical methods fit into this category and should be more widely applied to continuous or interval data. When outliers are not a problem, simulation studies have indicated their ability to detect significant differences is only slightly smaller than corresponding parametric methods. There are also various forms of robust regression models and computer-intensive approaches that deserve further consideration.

Below are the Boxplot for Numerical columns:

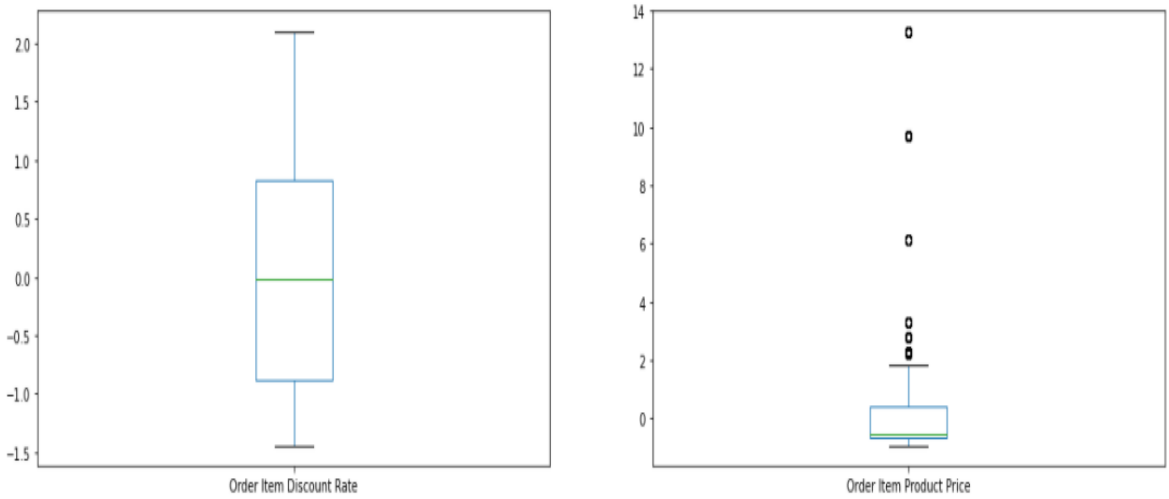


Fig 4.V.12 (a&b) Boxplot of Order Item Discount Rate and Order Item Product Price

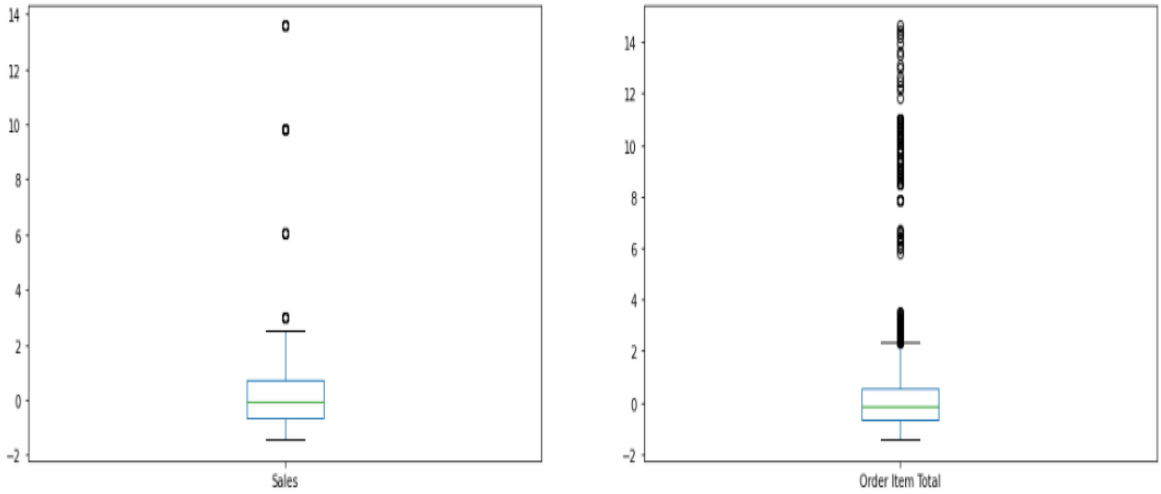


Fig 4.V.13 (a&b) Boxplot of Sales and Order Item Total

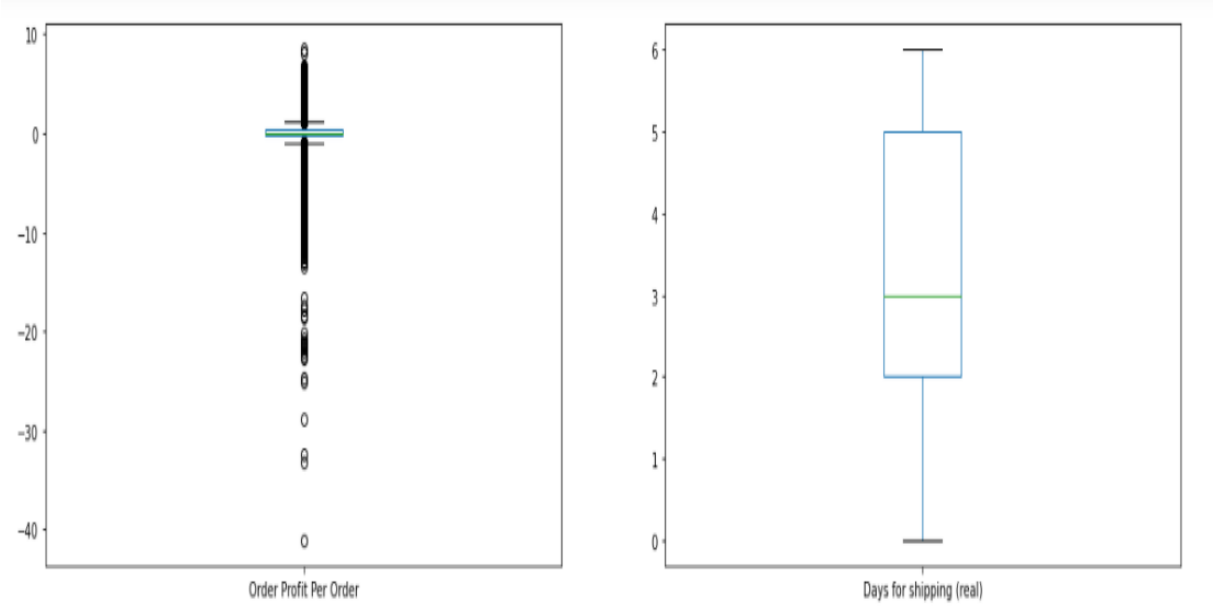


Fig 4.V.14 (a&b) Boxplot of Order Profit Per Order and Days for Shipping(real)

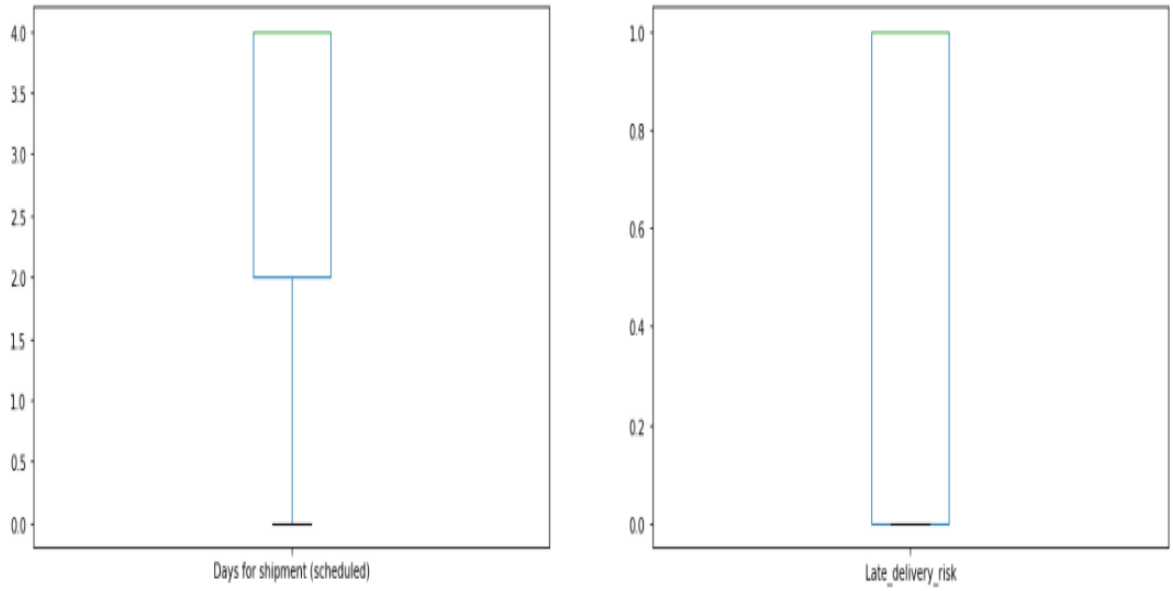


Fig 4.V.15 (a&b) Boxplot of Days for shipment (scheduled) and Late_delivery_risk

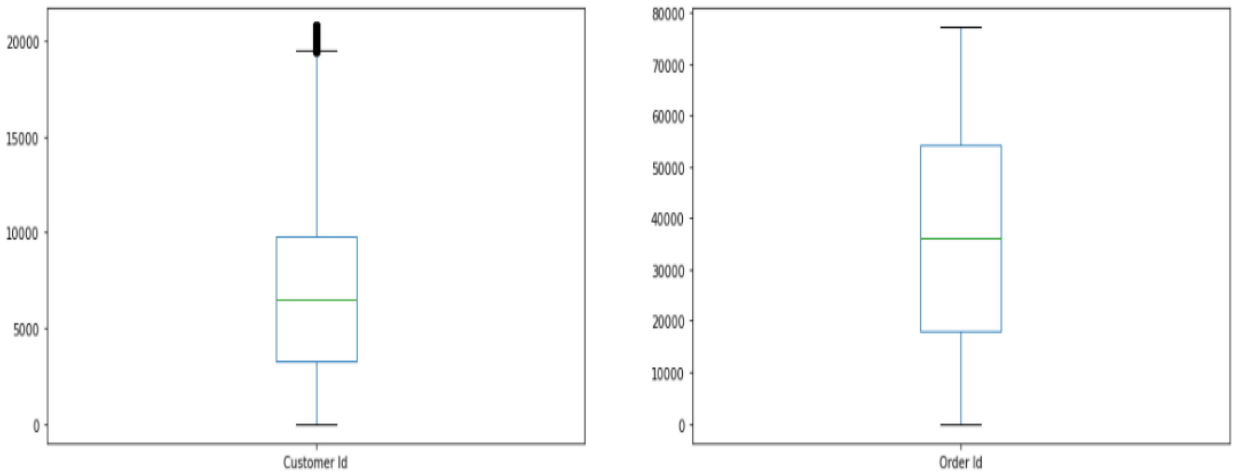


Fig 4.V.16 (a&b) Boxplot of Customer Id and Order id

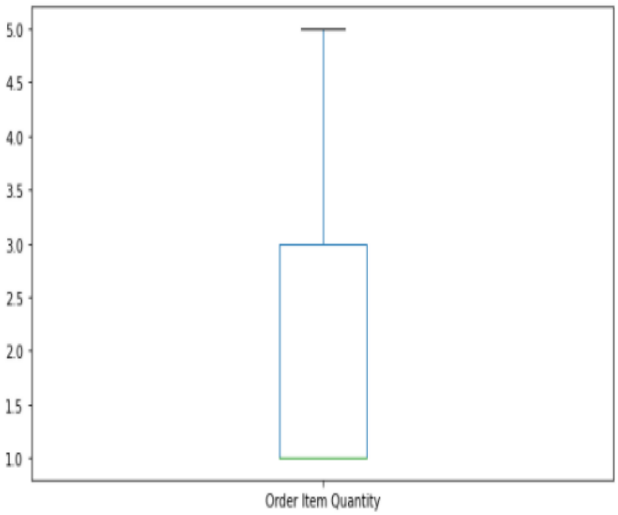


Fig 4.V.17 Boxplot of Order item Quantity

From above boxplot we can understand below numerical columns have outliers.

- Order Item Product
- Sales
- Order Item Total
- Order Profit Per Order
- Customer Id

Here, we have used IQR(Inter Quartile Range) method to treat the outlier in our dataset.

CLASS IMBALANCE AND ITS TREATMENT

Classification predictive modeling involves predicting a class label for a given observation.

An imbalanced classification problem is an example of a classification problem where the distribution of examples across the known classes is biased or skewed. The distribution can vary from a slight bias to a severe imbalance where there is one example in the minority class for hundreds, thousands, or millions of examples in the majority class or classes.

Imbalanced classifications pose a challenge for predictive modeling as most of the machine learning algorithms used for classification were designed around the assumption of an equal number of examples for each class. This results in models that have poor predictive performance, specifically for the minority class. This is a problem because typically, the minority class is more important and therefore the problem is more sensitive to classification errors for the minority class than the majority class.

Below is the count plot to understand class imbalance. There is a small imbalance between the class we would decide the treatment of it after building base model.

```
In [88]: df['Late_delivery_risk'].value_counts(normalize=True)*100

Out[88]: 1    54.829132
         0    45.170868
         Name: Late_delivery_risk, dtype: float64

In [96]: sns.countplot(df['Late_delivery_risk']);
         plt.text(x=.9, y= df['Late_delivery_risk'].value_counts()[1]/2,
                  s = '54.82 %');
         plt.text(x=-.1, y= df['Late_delivery_risk'].value_counts()[0]/2,
                  s = '45.17%');
         plt.title('Count Plot for Target Variable (Late_delivery_risk)', fontsize = 12)
         plt.xlabel('Target Variable', fontsize = 10)
         plt.ylabel('Count', fontsize = 10)

Out[96]: Text(0, 0.5, 'Count')
```

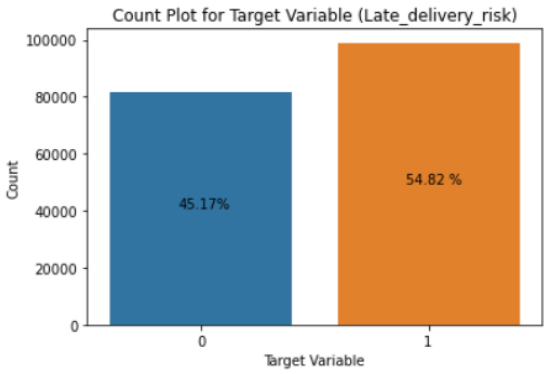


Fig 4.V.18 Countplot showing class imbalance

FEATURE ENGINEERING:

Feature engineering is the process of transforming raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy on unseen data.

Date and Time

The common data type group are all different formats of dates and time. The problem here is that it can vary by the format a lot. It could easily lead to some misunderstandings or an underperforming model if the formats DD/MM/YYYY and MM/DD/YYYY were put in the same dataset as simple strings. Once again, the problem is that data is not a straightforward numerical data. It cannot be directly fed into a machine learning model. The easiest way is to split the data into three integer features representing the day, the month and the year. We can also construct some cultural-related

features. For example, whether the day is a day of the weekend or it is a holiday. Other options are time or days elapsed from a certain event or intervals between consecutive events.

The Order date column has been split into two features which are Month and Year. These two columns will be included in the model building while eliminating the Order date column.

One-Hot Encoding

Though label encoding is straight but it has the disadvantage that the numeric values can be misinterpreted by algorithms as having some sort of hierarchy/order in them. This ordering issue is addressed in another common alternative approach called 'One-Hot Encoding'. In this strategy, each category value is converted into a new column and assigned a 1 or 0 (notation for true/false) value to the column. Though this approach eliminates the hierarchy/order issues but does have the downside of adding more columns to the data set. It can cause the number of columns to expand greatly if you have many unique values in a category column.

One-Hot Encoding is done on below variables.

1. Order Item Quantity
2. Category Name
3. Customer Country
4. Customer Segment
5. Customer State
6. Department Name
7. Market
8. Order Country
9. Order Region
10. Product Name
11. Shipping Mode
12. Month
13. Year

FEATURE TRANSFORMATION:

It means transforming our original feature to the functions of original features. Scaling, discretization, binning and filling missing data values are the most common forms of data transformation. To reduce right skewness of the data,we use log.No transformation operations were performed on any attribute.

MISSING VALUES:

In the real world, it is sometimes impossible to acquire some data. Or the data is lost somewhere in the processing pipeline. Due to that, there are usually some missing values in our data. Handling them is an art in itself. This part of the data processing is called data cleaning and often is considered to be a separate process.

Nonetheless, when creating some new features, we need to remember about it as the missing values can be hidden under different names and values. Some programming languages and libraries have a special object for such values. It is usually represented by "NaN" - not a number, but often some arbitrarily chosen values can be used instead.The missing values can be replaced with "0" what enables one to calculate the sum without complications, but it prevents us from generating a new feature with dividing by the value.

One more common option is filling the missing values with the mean or the median calculated from the present values. But once again, if we compute the average again we will get a different value, so there will be a serious difference between a new feature based on the true mean and the miscalculated. These examples show the never changing truth - know your data! And this is important during the feature engineering as well.

Attributes having more than 60% of missing values were removed as they were of no importance while explaining the variance in the model.

Our dataset has 53 columns and 180519 Rows. There are 4 columns that have missing values.

The percentage of missing values exceeds 80% for the two variables, "Product Description" & "Order Zipcode" and hence we need to drop them.The variables, Customer Lname and Customer Zipcode has less than 10 missing values. We can remove the Customer Lname and Customer Zipcode as these do not contribute to our analysis.

SCALING THE DATA:

STANDARD NORMALIZATION:

In nature and in the human society, many things are governed by the normal (Gaussian) distribution. That is why we introduce normalization characteristic to the distribution. It is given by the following equation:

$$\tilde{X} = \frac{X - \mu}{\sigma}$$

Here the X is our new feature. It is acquired by subtracting the mean value of the old feature from every sample of the old feature and then dividing it by the standard deviation which tells us about how wide is the spread of values of the feature. This brings the value of X into an interval of [-1,1].

Different models require different normalization in order to work efficiently. For example, in the case of the *k*-nearest neighbours, the scale of a particular feature plays a role of a weight. The bigger are the values the more important is the feature. On the other hand, Naïve Bayes and the Decision Tree based algorithms neither benefit, nor get hurt by the normalization.

Data scaling was done on continuous data to reduce the effect of outliers on result of KNN and logistic regression models. Also to make sure that the effect of unit was taken out.

These are the variables which were subjected to scaling.

1. Order Item Discount Rate
2. Order Item Product Price
3. Sales
4. Order Item Total
5. Order Profit Per Order

FEATURE SELECTION

FEATURE IMPORTANCE

This can be helpful as a pre-cursor to selecting features. Features are allocated scores and can then be ranked by their scores. Those features with the highest scores can be selected for inclusion in the training dataset, whereas those remaining can be ignored.Feature importance scores can also provide you with information that you can use to extract or construct new features, similar but different to those that have been estimated to be useful.

A feature may be important if it is highly correlated with the dependent variable (the thing being predicted). Correlation coefficients and other univariate (each attribute is considered independently) methods are common methods.

More complex predictive modelling algorithms perform feature importance and selection internally while constructing their model. Some examples include MARS, Random Forest and Gradient Boosted Machines. These models can also report on the variable importance determined during the model preparation process.Feature selection was done with the help of decision tree's feature importance operation for the tunned models.

FEATURE SELECTION:

Those attributes that are irrelevant to the problem need to be removed. There will be some features that will be more important than others to the model accuracy. There will also be features that will be redundant in the context of other features.

Feature selection addresses these problems by automatically selecting a subset that are most useful to the problem. Feature selection algorithms may use a scoring method to rank and choose features, such as correlation or other feature importance methods.

More advanced methods may search subsets of features by trial and error, creating and evaluating models automatically in pursuit of the objectively most predictive sub-group of features. There are also methods that bake in feature selection or get it as a side effect of the model. Stepwise regression is an example of an algorithm that automatically performs feature selection as part of the model construction process.

Regularization methods like LASSO and ridge regression may also be considered algorithms with feature selection baked in, as they actively seek to remove or discount the contribution of features as part of the model building process.

Some of the commonly used feature scoring functions are:

- F-score,
- mutual information score,
- Chi-square score.

F-score can find the linear relation between feature and target columns, and create scores accordingly. Using scores for each feature, we can eliminate the ones with a lower F-score. Similarly, mutual information score can capture both linear and non-linear relationships between feature and target column, but needs more samples.

Chi square is a test used in statistics to test the independence of two events. A lower value of chi square suggests that the two variables(feature and target) are independent. Higher values for two variables means dependent hence important features.

ITERATIVE PROCESS OF FEATURE ENGINEERING:

Knowing where feature engineering fits into the context of the process of applied machine learning highlights that it does not standalone. It is an iterative process that interplays with data selection and model evaluation, again and again, until we run out of time on our problem.

The process might look as follows:

- Brainstorm features: Really get into the problem, look at a lot of data, study feature engineering on other problems and see what you can steal.
- Devise features: Depends on your problem, but you may use automatic feature extraction, manual feature construction and mixtures of the two.
- Select features: Use different feature importance scorings and feature selection methods to prepare one or more "views" for your models to operate upon.
- Evaluate models: Estimate model accuracy on unseen data using the chosen features.

You need a well-defined problem so that you know when to stop this process and move on to trying other models, other model configurations, ensembles of models, and so on.

INFORMATION GAIN:

Information gain calculates the reduction in entropy from the transformation of a dataset. It can be used for feature selection by evaluating the Information gain of each variable in the context of the target variable.

The below plot shows the variables with different information gain. Attributes with high information gain will result in high drop in entropy will result in better prediction of the target variable comparatively

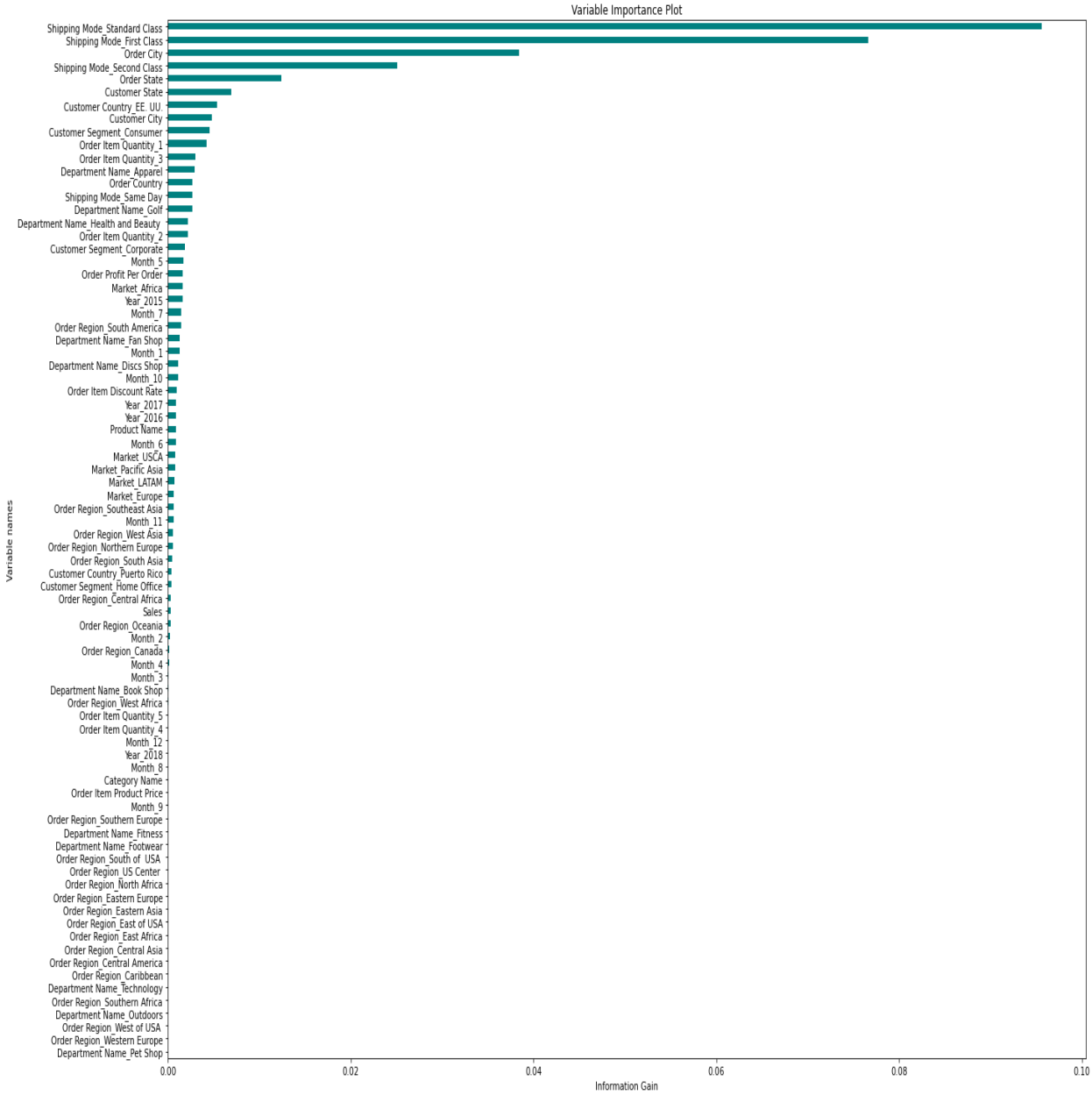


Fig 4.V.19 Variable vs Information gain

DIMENSIONALITY REDUCTION:

FEATURE EXTRACTION:

Feature extraction is the process of making new features which are composite of the existing ones. One of the great example of Feature Extraction is dimensionality reduction. There can be millions of features in a dataset with audio, images, or even a tabular one. While a lot of them can be redundant, there is also the problem of model complexity.

For some machine learning algorithms, the training time complexity increases exponentially as the number of features grows. In this case, we use feature extraction or dimensionality reduction.

There are algorithms like PCA, TSNE, and others that can be used to reduce feature dimensionality. They aggregate different features by using mathematical operations, while trying to keep the information intact.

Dimensionality reduction techniques were not used on our attributes.

BASE MODEL:

DECISION TREE:

Decision tree algorithm is one of the most versatile algorithms in machine learning which can perform both classification and regression analysis. It is very powerful and works great with complex datasets. Apart from that, it is very easy to understand and read. That makes it more popular to use. When coupled with ensemble techniques – which we will learn very soon- it performs even better.

As the name suggests, this algorithm works by dividing the whole dataset into a tree-like structure based on some rules and conditions and then gives prediction based on those conditions. Decision tree has no assumptions unlike some other algorithms.

- Decision tree algorithm falls under the category of supervised learning. They can be used to solve both regression and classification problems.
- Decision tree uses the tree representation to solve the problem in which each leaf node corresponds to a class label and attributes are represented on the internal node of the tree.

We can represent any boolean function on discrete attributes using the decision tree

- At the beginning, we consider the whole training set as the root.
- Feature values are preferred to be categorical. If the values are continuous then they are discretized prior to building the model.
- On the basis of attribute values records are distributed recursively.
- We use statistical methods for ordering attributes as root or the internal node.

In Decision Tree the major challenge is to identification of the attribute for the root node in each level. This process is known as attribute selection. We have two popular attribute selection measures:

- Information Gain
- Gini Index

Information gain:

we use a node in a decision tree to partition the training instances into smaller subsets the entropy changes. Information gain is a measure of this change in entropy.

$$Gain(S, A) = Entropy(S) - \sum_{v \epsilon Values(A)} \frac{|S_v|}{|S|}.Entropy(S_v)$$

Entropy:

Entropy is the measure of uncertainty of a random variable, it characterizes the

impurity of an arbitrary collection of examples. The higher the entropy more the information content.

Building Decision Tree using Information Gain

The essentials:

- Start with all training instances associated with the root node
- Use info gain to choose which attribute to label each node with
- *Note:* No root-to-leaf path should contain the same discrete attribute twice
- Recursively construct each subtree on the subset of training instances that would be classified down that path in the tree.

Gini Index

- Gini Index is a metric to measure how often a randomly chosen element would be incorrectly identified.
- It means an attribute with lower Gini index should be preferred.
- Sklearn supports "Gini" criteria for Gini Index and by default, it takes "gini" value.
- The Formula for the calculation of the of the Gini Index is given below

$$GiniIndex = 1 - \sum_j p_j^2$$

Advantages of Decision Tree:

- It can be used for both Regression and Classification problems.
- Decision Trees are very easy to grasp as the rules of splitting is clearly mentioned.
- Complex decision tree models are very simple when visualized. It can be understood just by visualising.
- Scaling and normalization are not needed.

Disadvantages of Decision Tree:

- A small change in data can cause instability in the model because of the greedy approach.
- Probability of overfitting is very high for Decision Trees.
- It takes more time to train a decision tree model than other classification algorithms.

Base model was built after the feature selection by using train test split and decision tree algorithm with pruning. Maximum depth was used to prune the Decision Tree.

The Classification report below shows the performance of the base model

```
print(classification_report(y_test,y_test_pred))
```

	precision	recall	f1-score	support
0	0.61	0.76	0.68	16294
1	0.75	0.60	0.67	19810
accuracy			0.67	36104
macro avg	0.68	0.68	0.67	36104
weighted avg	0.69	0.67	0.67	36104

```
print(classification_report(y_train,y_train_pred))
```

	precision	recall	f1-score	support
0	0.63	0.79	0.70	65248
1	0.78	0.62	0.69	79167
accuracy			0.69	144415
macro avg	0.70	0.70	0.69	144415
weighted avg	0.71	0.69	0.69	144415

Fig 4.V.20 Performance report of the base model

The base model seems to be under fit. For our problem statement, the model is sensitive towards recall(1).The model is supposed to have higher Recall(1) so as to predict late deliveries more accurately. As can be seen by the recall value of 1 which is

67% only, there is a need to fine tune the model to improve the performance of the base model.

The model produces lot of False Negatives and the aim is to reduce this and increase the true positives so as to not miss out on the late delivery predictions of orders which might end being late delivered and hence lead to customer dissatisfaction.

FINAL MODEL:

Boosting:

Boosting is an ensemble approach (meaning it involves several trees) that starts from a weaker decision and keeps on building the models such that the final prediction is the weighted sum of all the weaker decision-makers.

The weights are assigned based on the performance of an individual tree.

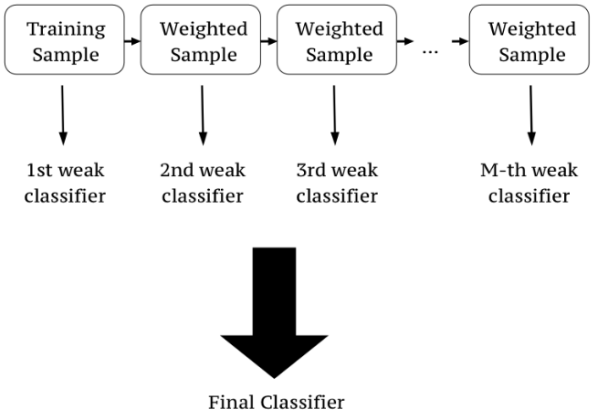


Fig4.V.21 Boosting Technique

Ensemble parameters are calculated in **stagewise way** which means that while calculating the subsequent weight, the learning from the previous tree is considered as well.

Gradient Boosting:

Gradient boosting Algorithm involves three elements:

- A loss function to be optimized.
- Weak learner to make predictions.
- An additive model to add weak learners to minimize the loss function.

a. Loss Function

- The loss function used depends on the type of problem being solved.
- It must be differentiable. Although, many standard loss functions are supported and you can define your own.

b. Weak Learner

- We use decision trees as the weak learner in gradient boosting algorithm.
- Specifically, we use regression tree that output real values for splits. And whose output can be added together. It allows next models outputs to be added and "correct" the residuals in the predictions.
- Trees need to construct in a greedy manner. It helps in choosing the best split points based on purity scores like Gini or to cut the loss.
- Initially, such as in the case of AdaBoost. Also, we use very short decision trees that only had a single split, called a decision stump.
- Generally, we use larger trees with 4-to-8 levels.
- It is common to constrain the weak learners in specific ways. Such as a maximum number of layers, nodes, splits or leaf nodes.
- This is to ensure that the learners remain weak, but can still need to construct in a greedy manner.

c. Additive Model

- Trees need to add one at a time, and existing trees in the model need not change.
- We use a gradient descent procedure to minimize the loss when adding trees.

- Traditionally, we use gradient tree to cut a set of parameters. Such as the coefficients in a regression equation or weights in a neural network. After calculating error or loss, the weights need to be update to minimize that error.
- Instead of parameters, we have weak learner sub-models or more specifically decision trees. After calculating the loss, to perform the gradient descent procedure. We must add a tree to the model that reduces the loss.
- We do this by parameterizing the tree. Then change the parameters of the tree and move in the right direction by (reducing the residual loss.)

5. MODEL EVALUATION:

Different classification models as shown in table were built with keeping in mind that F1-score and eventually Recall (1) of the model should be optimum.

After building base model for unpruned Decision Tree, it was pruned using hyperparameter tuning and the results were tabulated, next the Youden's cutoff index was found out to set a threshold cutoff for probability to improve the model performance. There was no change in the performance metric when the tree was pruned but the performance of the model did reduce when the threshold was changed as it was synthetically classifying false negatives into True positives.

The same Approach was used to build different models and were tabulated simultaneously to compare F1-scores. The best model after different tuning was GradientBoostingClassifier which also gave negligible variance error for the target variable.

	Model	Training Precision_1	Testing Precision_1	Training Recall_1	Testing Recall_1	Training f1_weighted	Testing f1_weighted	Training Accuracy	Testing Accuracy	AUROC Train	AUROC Test
0	Unpruned decision Tree	1.00	0.84	1.00	0.84	1.000000	0.820000	1.00	0.82	1.00	0.82
1	depth tuned decision Tree	0.99	0.83	0.99	0.83	0.990000	0.820000	0.99	0.81	0.99	0.82
2	cutoff tuned decision tree	0.89	0.79	0.99	0.80	0.980000	0.790000	0.99	0.81	0.99	0.82
3	Logistic regression without tuned cutoff	0.56	0.56	0.87	0.87	0.490000	0.490000	0.55	0.55	0.52	0.52
4	Logistic regression with tuned cutoff	0.48	0.48	0.72	0.72	0.576000	2.000000	0.53	0.52	0.55	0.54
5	KNN Model untuned	0.75	0.67	0.79	0.71	0.740000	0.650000	0.74	0.65	0.73	0.64
6	KNN Model tuned	0.85	0.72	0.86	0.74	0.840000	0.700000	0.84	0.70	0.84	0.70
7	KNN Model tuned with cutoff	0.70	0.59	0.95	0.85	0.806061	0.696528	0.79	0.67	0.81	0.69
8	Random Forest untuned	1.00	0.85	1.00	0.68	1.000000	0.780000	1.00	0.76	1.00	0.77
9	Random Forest tuned	1.00	0.85	1.00	0.68	1.000000	0.760000	1.00	0.76	1.00	0.77
10	Random Forest tuned with cutoff	0.99	0.65	1.00	0.90	0.994975	0.754839	0.99	0.74	1.00	0.75
11	AdaBoost untuned	0.85	0.85	0.54	0.68	0.690000	0.760000	0.70	0.76	0.71	0.76
12	AdaBoost tuned	0.85	0.85	0.54	0.54	0.690000	0.690000	0.70	0.70	0.71	0.71
13	GradientBoost untuned	0.84	0.84	0.55	0.55	0.690000	0.690000	0.70	0.70	0.71	0.71
14	GradientBoost tuned with cutoff	0.80	0.77	0.84	0.70	0.819512	0.733333	0.84	0.77	0.83	0.76
15	GradientBoost Final Tuned Model	1.00	0.92	0.99	0.70	0.990000	0.880000	0.99	0.88	0.99	0.89

Fig 5.1 Scores of Different Approach

The below plot explains the recall score for various models like Logistic Regression ,K-Nearest Neighbors, Decision Tree, Navie Bayers, Random forest. From the plot we can infer that the decision trees perform better comparatively.

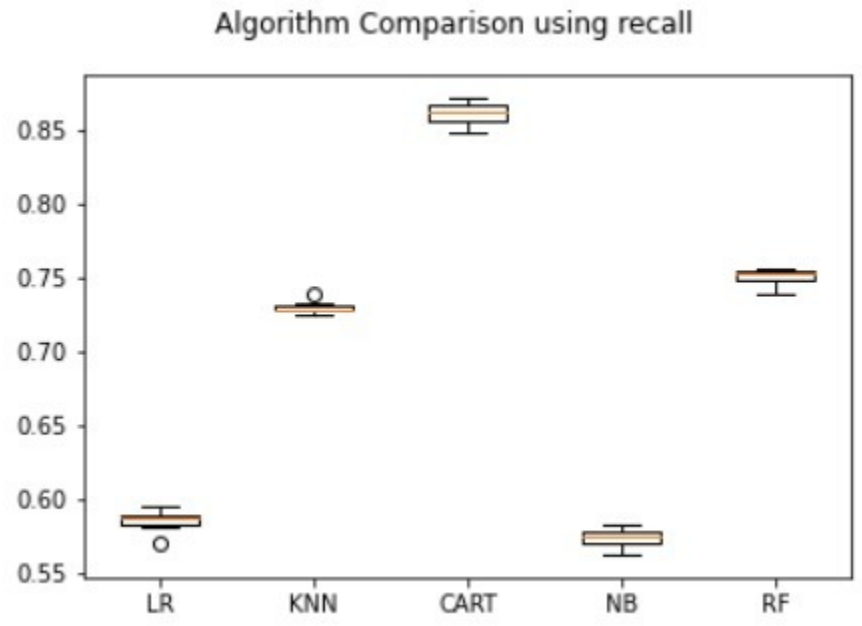


Fig 5.2 Algorithm comparison using recall

After fitting Decision tree classifier as base model, boosting models like AdaboostClassifier, GradientboostClassifier and XGBclassifier were fit to the data to reduce both variance and bias. GradientboostClassifier managed to produce the best results comparatively.

The below plot shows the feature importance's of each feature. The top 20 features out of these were taken to build the final model.

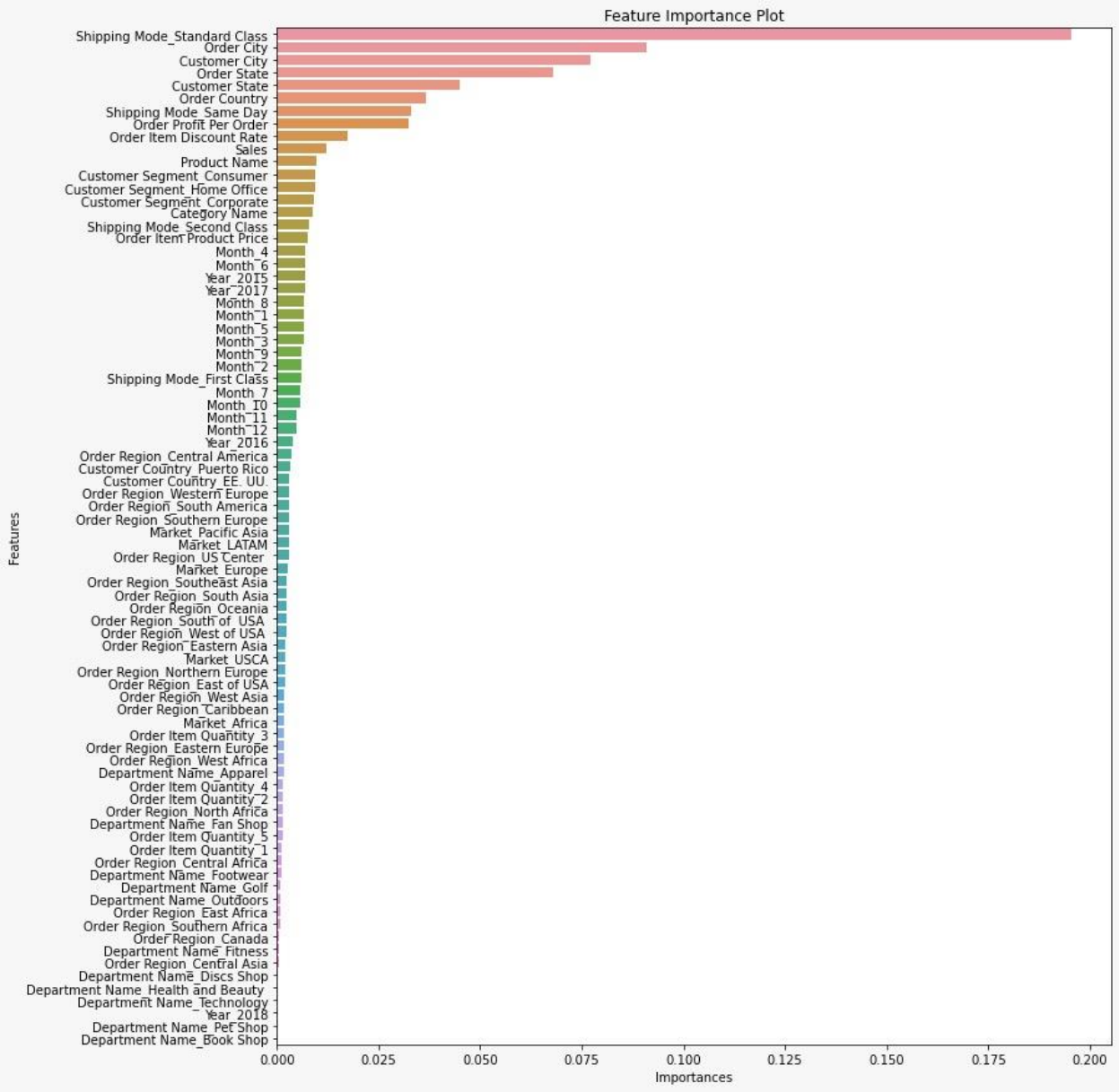


Fig 5.3 Features with feature importance score

The bias error is around 14% which is the least we could get for the given data and variance error is 0.4% which means the model has great consistency. By observing both bias error and variance error we could say that the model is perfectly fit with the below scores.

```
[0.8444272  0.85519055 0.85279831 0.8523097  0.85818006]
Bias error: 0.14741883666119193
Variance error: 0.005368913259793521
```

Fig 5.4 bias and variance error for Gradient boosting classifier with max_depth =60

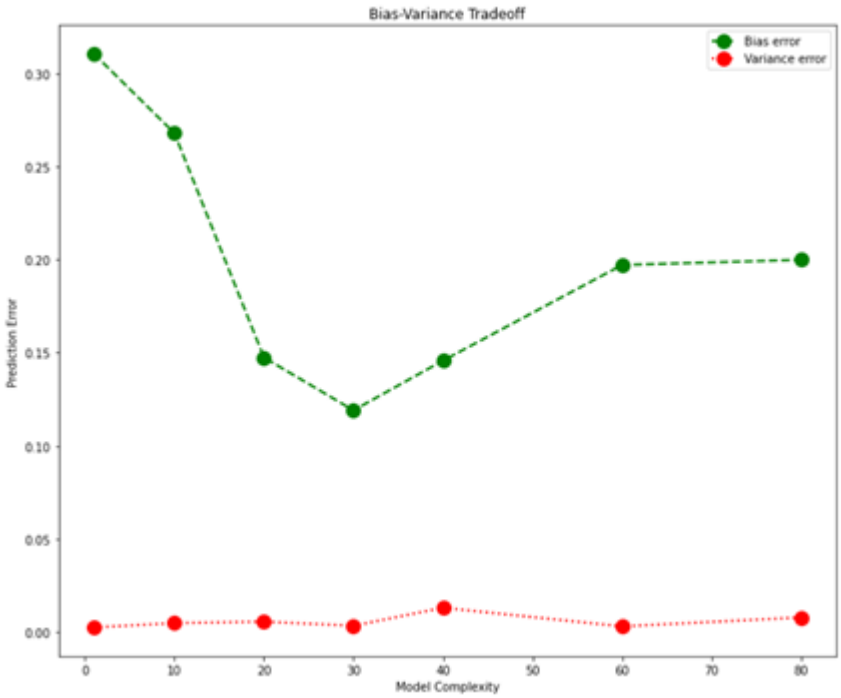


Fig 5.5 Bias-variance Tradeoff

	precision	recall	f1-score	support
0	0.85	0.91	0.88	14523
1	0.92	0.86	0.89	17514
accuracy			0.88	32037
macro avg	0.88	0.89	0.88	32037
weighted avg	0.89	0.88	0.88	32037

Fig 5.6 Performance report of the final Model

The above classification report shows that the final model which we have fit using gradient boosting classifier has resulted in 89% f1-score(1) and 88% f1-score(0) along with 86% recall(1) and 91% recall(0). This means we have managed to increase the recall (1) from 60% to 86% , recall(0) from 76% to 91% , F1-score(1) from 67% to 89% and F1-score(0) from 68% to 88%.

Since our model is sensitive to recall (1), we had to make sure that the recall (1) will be increased from 60% to more than 85%. By increasing the recall (1) we have managed to reduce false negatives (i.e. predicting late delivery orders with high late delivery risk as orders with low late delivery risk) and increase True positive (i.e. predicting late delivery orders with high late delivery risk as orders with high late delivery risk).

Shipping Mode-

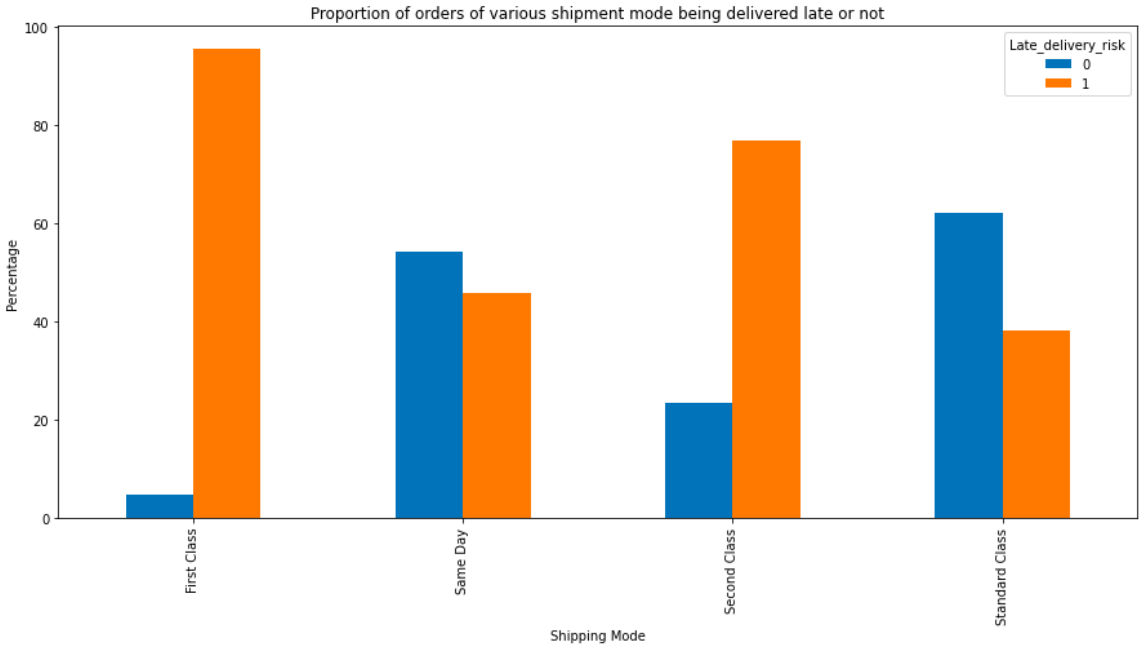


Fig 5.7 Proportion of orders of various shipment mode being delivered late or not

ORDER CITY-

Late_delivery_risk	Order_city	Order_city	Order_city	Order_city
1	Almaty	Ercolano	Madero	Rennes
1	Ancona	Getafe	Mantes-la-Jolie	Rotherham
1	Apucarana	Glasgow	Mersin	Salamanca
1	Araçatuba	Gradignan	Milan	San Francisco de Macorís
1	Bhopal	Gujranwala	Naples	Santa Cruz do Sul
1	Bologna	Guzmán	Pamiers	Shiraz
1	Carcassonne	Gy?r	Pekín	São Miguel dos Campos
1	Chesterfield	Huixquilucan	Pierrefitte-sur-Seine	Thonon-les-Bains
1	Clarksville	Istres	Pirapora	Villahermosa
1	Cuajimalpa	Itapecuru Mirim	Port Macquarie	Villeurbanne
1	Cuscatancingo	Ivano-Frankivs'k	Posadas	Wiesbaden
1	Cárdenas	Kingswood	Recife	Yantai

Table 4.V.3 Significance Order cities with High Late Deliveries

CUSTOMER CITY-

Customer_city	Adj. Res.	Adj. Res.
1	Apopka	Jamaica
1	Ballwin	Madera
1	Baltimore	Medina
1	Chula Vista	Pico Rivera
1	Denton	Plano
1	Denver	Roseville
1	Ewa Beach	San Bernardino
1	Fayetteville	Troy
1	Flushing	Tustin
1	Garland	Waukegan
1	Hamilton	

Table 4.V.4 Significance Customer cities with Late Deliveries

6. COMPARISION TO BENCHMARK:

We had kept 90% as the benchmark f1_weighted score, but we were only able to build a model with 85% f1_weighted score as our data was real time data which was unfiltered and large so the independent variables of our data could only able to explain 85% of variance present in the dependent variable. Most of the independent categorical variables were of high cardinality, so encoding those variables was a difficult task.

VISUALIZATION:

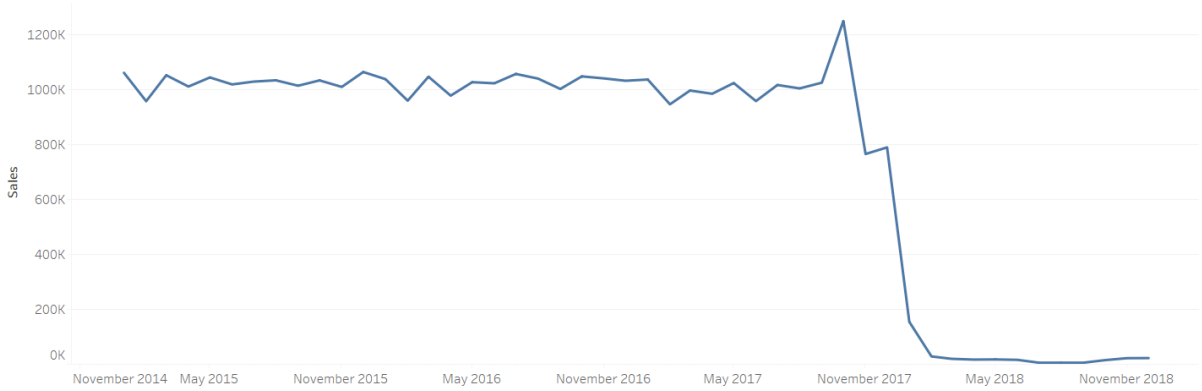


Fig 7.1.Plot to show drop in sales after September 2017

From the above graph we can infer that there is a drop in sales of company from September 2017 and which is very steep drop in sales. There might be various reason for it one of reason we feel might be increase in customer churn.

Fig 7.2. Count plot of Delivery Status

From the above graph we can infer that majority of the products were delivered late which might also be the cause of increase in customer churn.

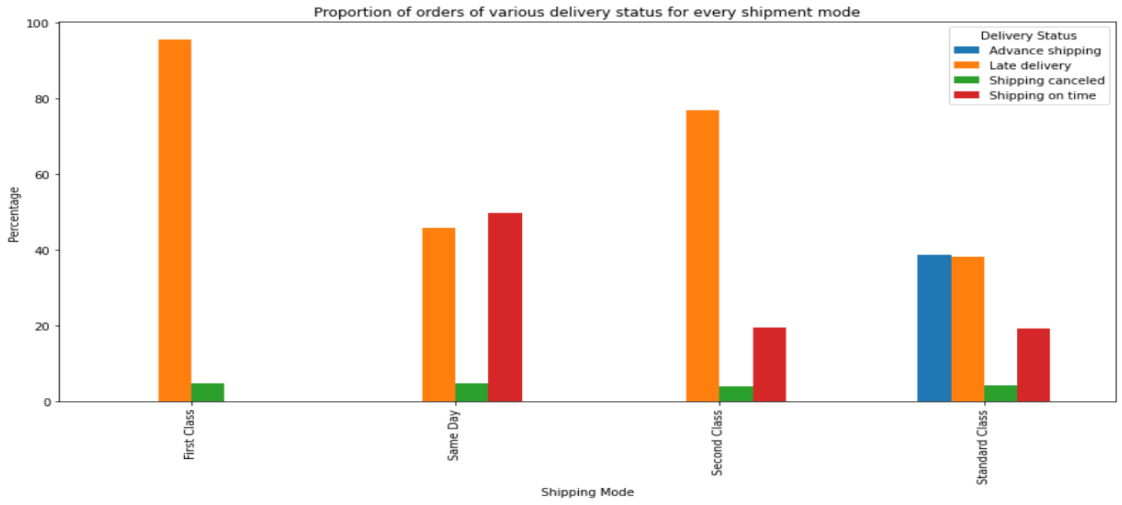


Fig 7.3. Countplot for proportion of orfers of various delivery status for every shipment mode

- First class- company schedules the delivery of the Product in a day.

- Second class- company schedules the delivery of the Product in two days.
- Standard class- company schedules the delivery of the Product in a four days.
- Same day- company schedules the delivery of the Product on the same day.
- 95% of the first class deliveries were late delivered and the remaining were cancelled whereas only 50% of the same day deliveries were late delivered.
- Company is concentrating more towards same day deliveries compared to first class deliveries.
- 75% of the second class deliveries were late delivered.

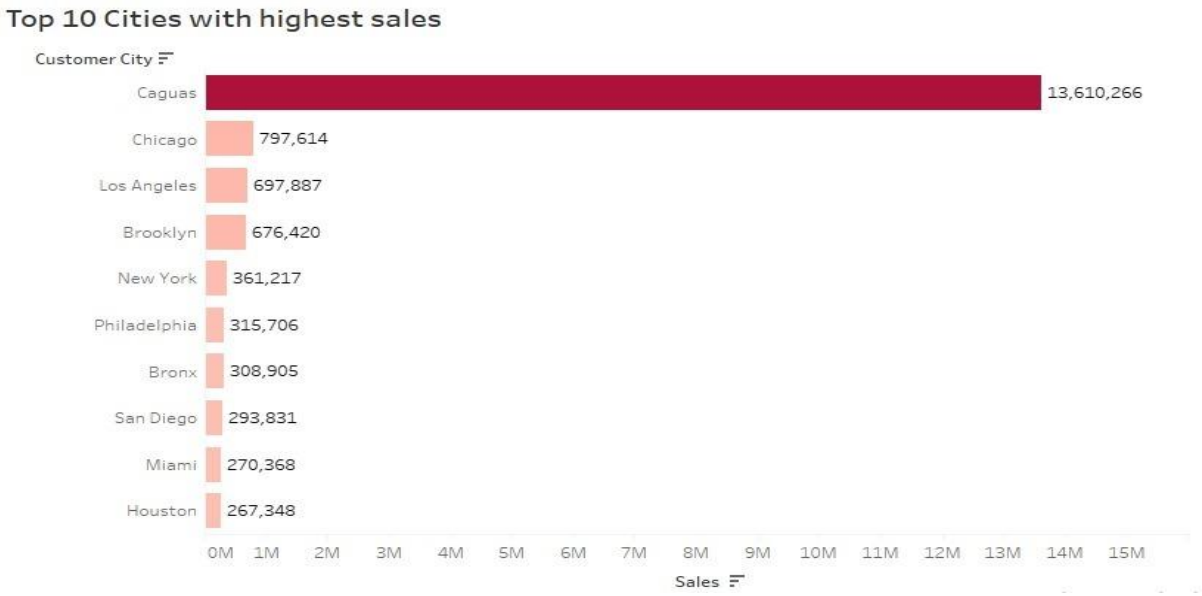


Fig 7.4.Graph showing top 10 cities with highest sales

From the above graph we can infer that Caguas is a city were company is doing really well in terms of sales. It would be advisable that the company should look into what different are they doing in this particular city for such drastic difference and try using the same in other cities to increase their sales.

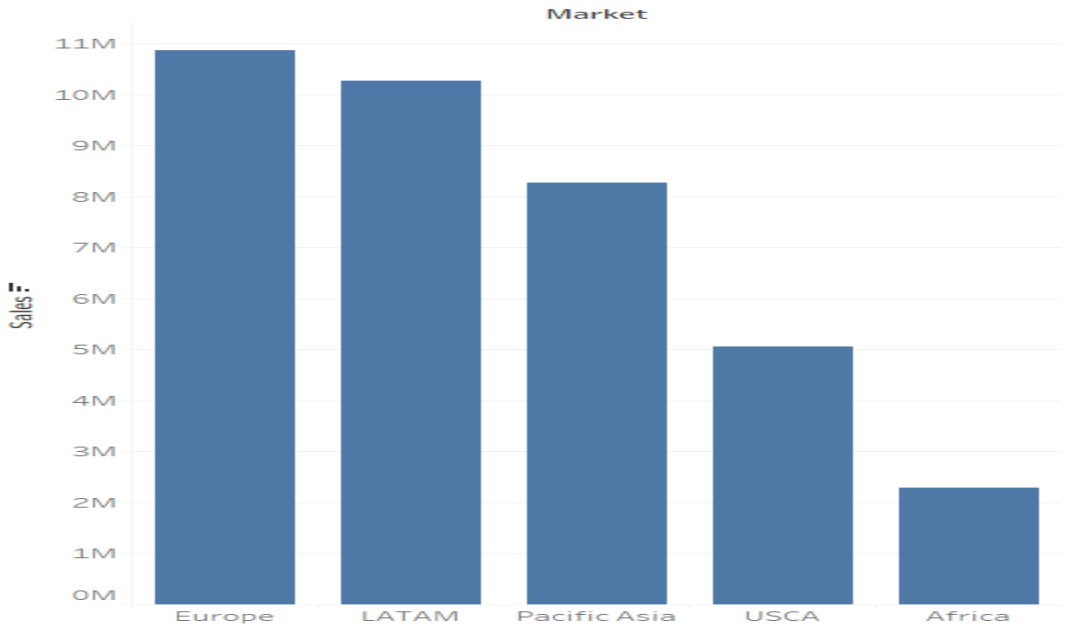


Fig 7.5.Graph showing Sales v/s Market place

From the above graph we can infer that they are doing well in Markets like Europe and LATAM compare to other markets. They need to concentrate little more in these markets probably increase the advertising or offices in these market to make more gain.

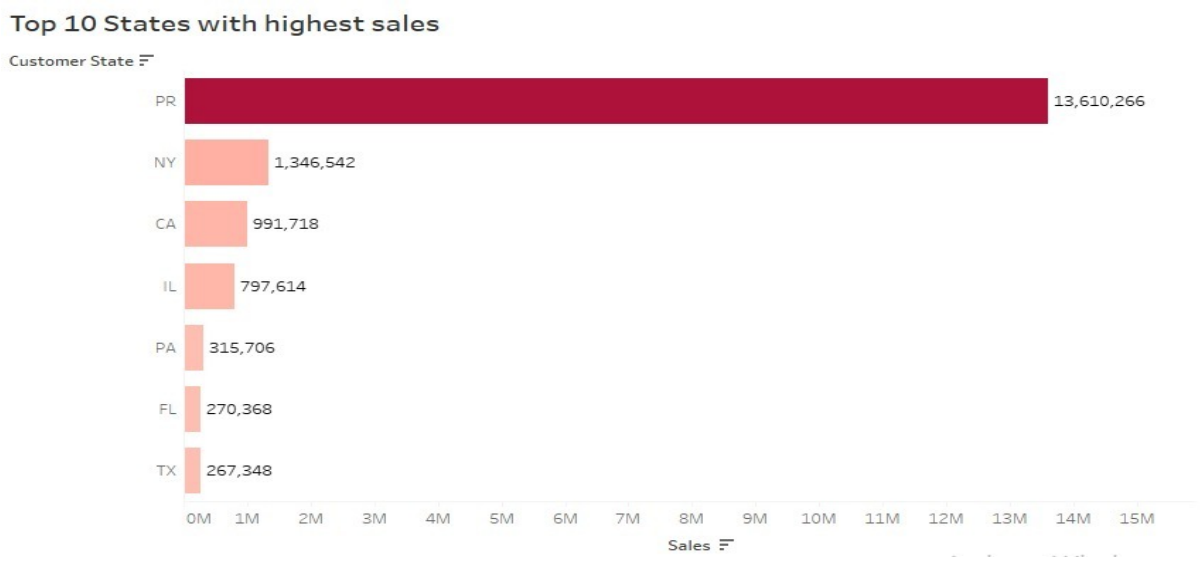


Fig 7.6. Graph showing Top 10 States with highest sales

From the above graph we can infer that PR is the state were company is doing really well in terms of sales. It would be advisable that the company should look into what

different are they doing in this particular city for such drastic difference and try using the same in other cities to increase their sales.

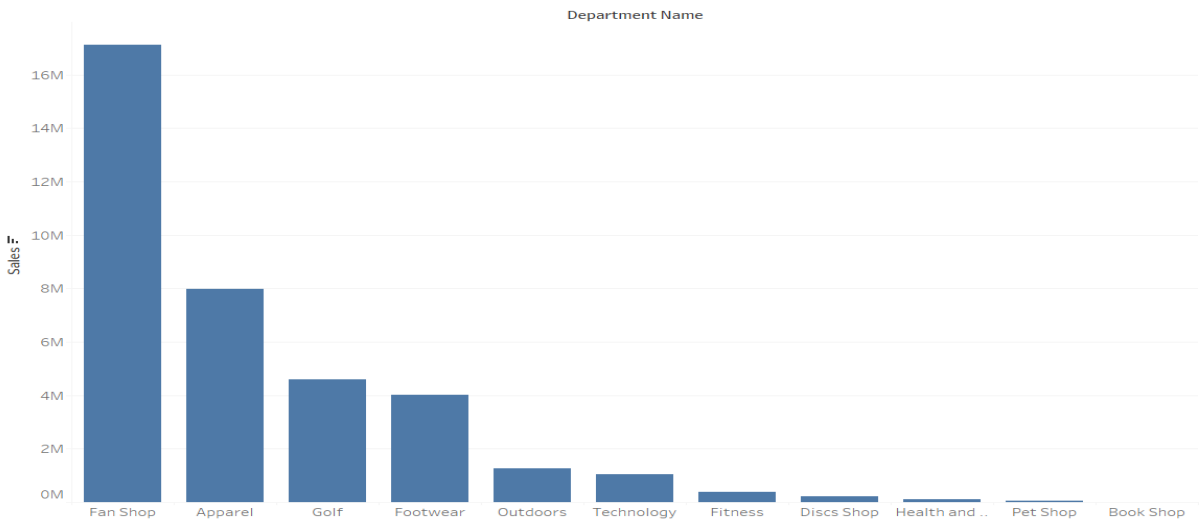


Fig 7.7 Graph showing Sales v/s Department Name

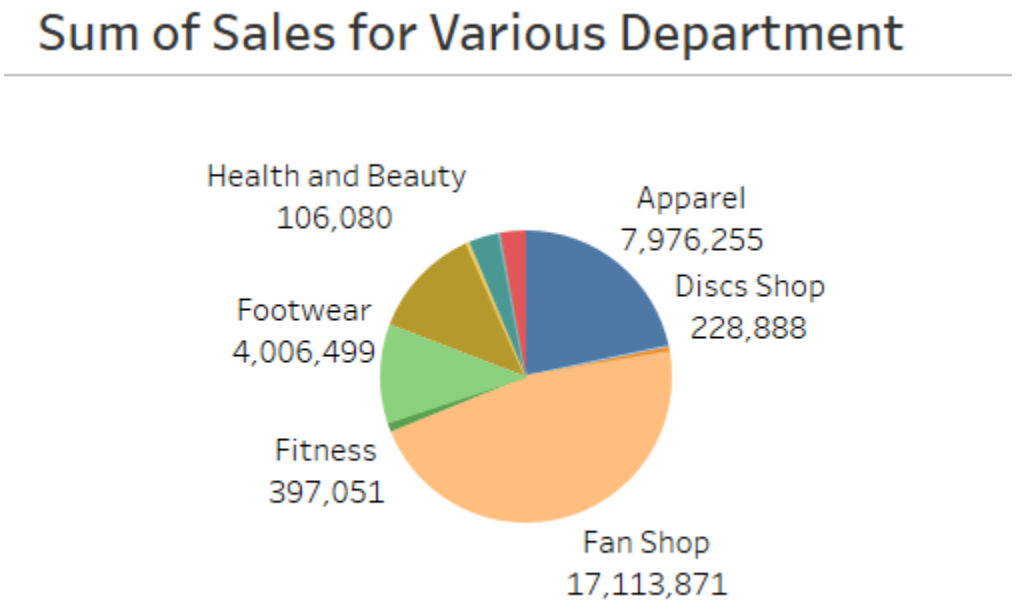


Fig 7.8. Pie plot showing sum of sales for various department

From the above 2 graph we can infer that Fan Shop is the department which has better sales than other department this might be because they might be have more clients who deal with Fan shop products and comfortable with them than other department. They should try reaching to more clients from other departments and try building a confident that they have expertise in shipping there products.

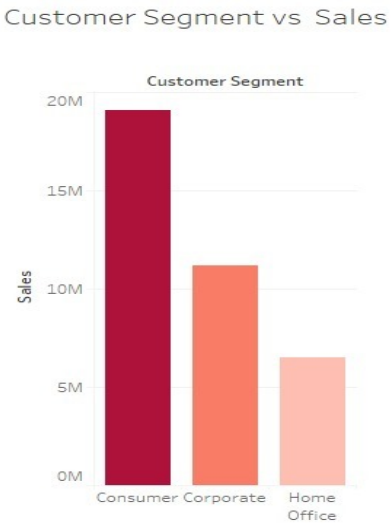


Fig 7.9 Graph showing Customer Segment v/s Sales

From the above graph we can infer that the company has more clients who are consumers they should try getting more customers from Corporate and Home office

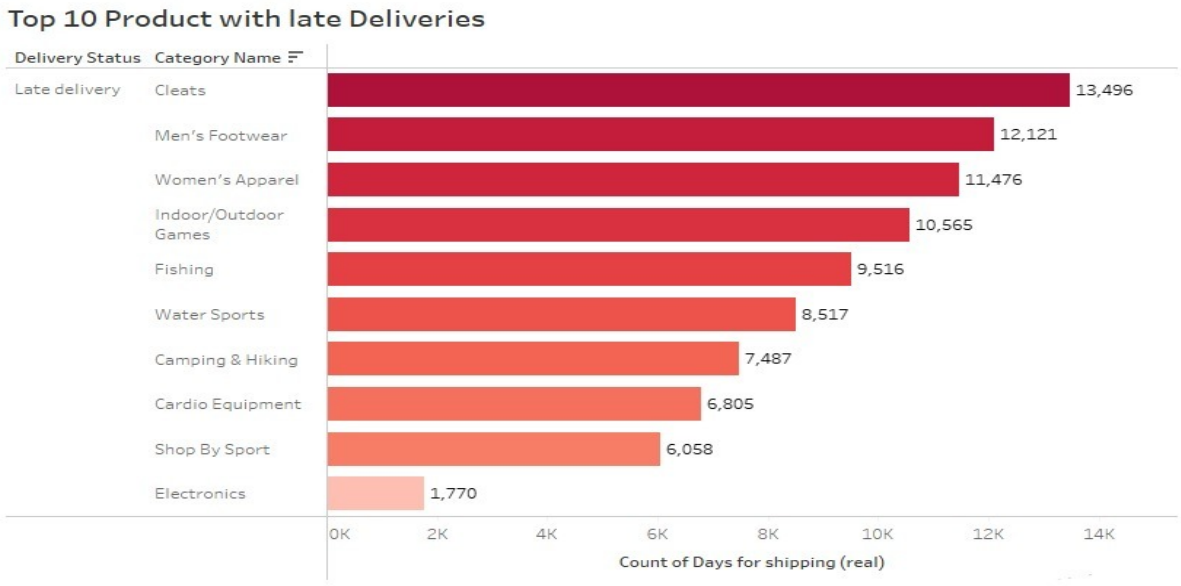


Fig 7.10 Graph showing Top 10 product with late Deliveries

From the above graph we can infer Cleats, Men's Footwear, Women's Apparel, Indoor/Outdoor Games, Fishing, Water Sports, Camping & Hiking, Cardio Equiment, Shop by sport are the products which are delivered late

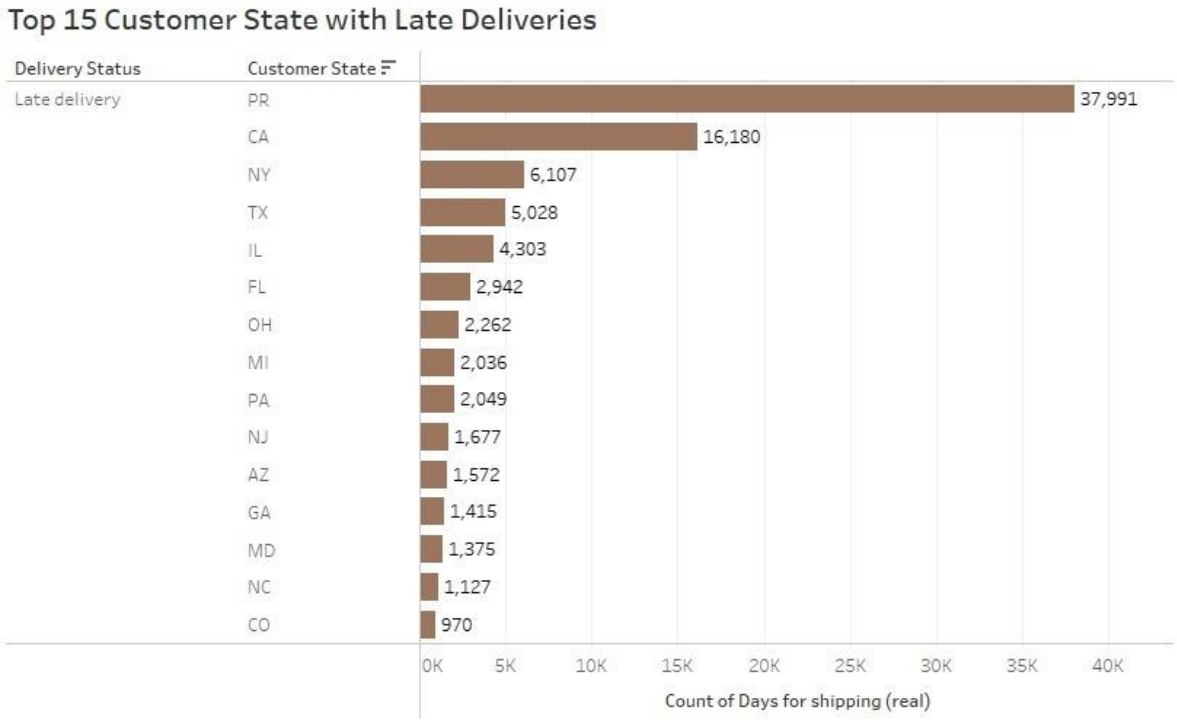


Fig 7.11 Graph showing Top 15 Customer State with Late Deliveries

From the above graph we can infer that PR and CA are the two states with most late deliveries. So company should take necessary measures to reduce late deliveries in these states.

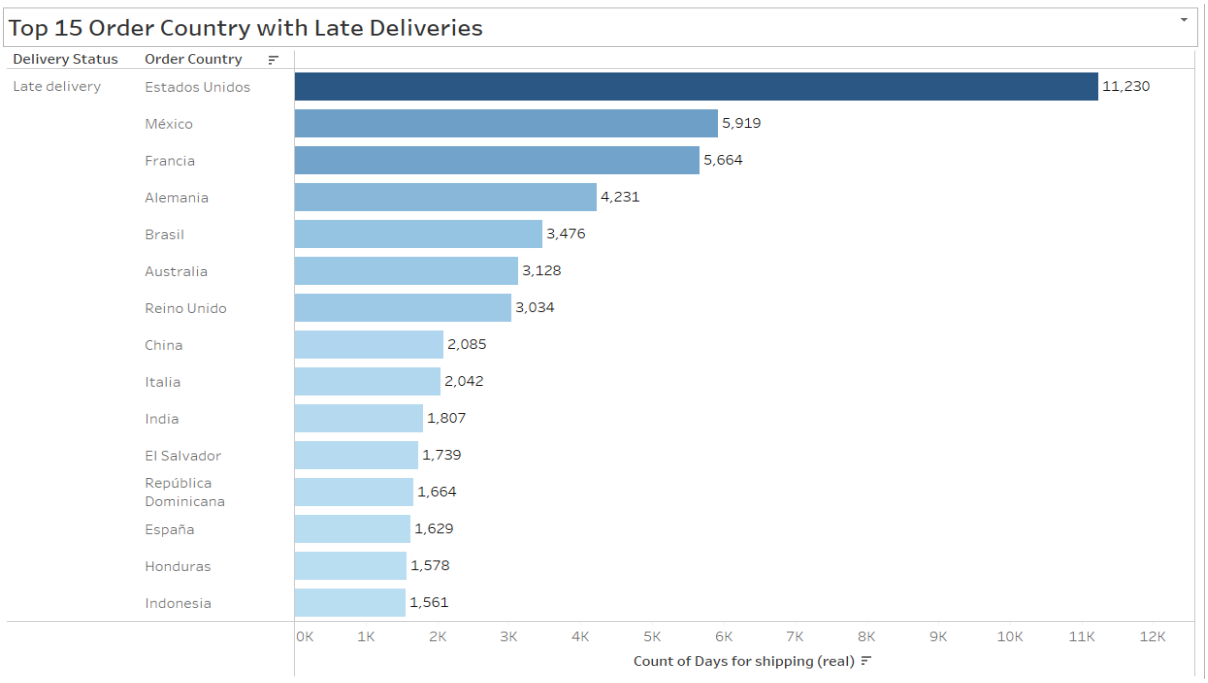


Fig 7.12. Graph showing Top 15 Order Country with Late Deliveries

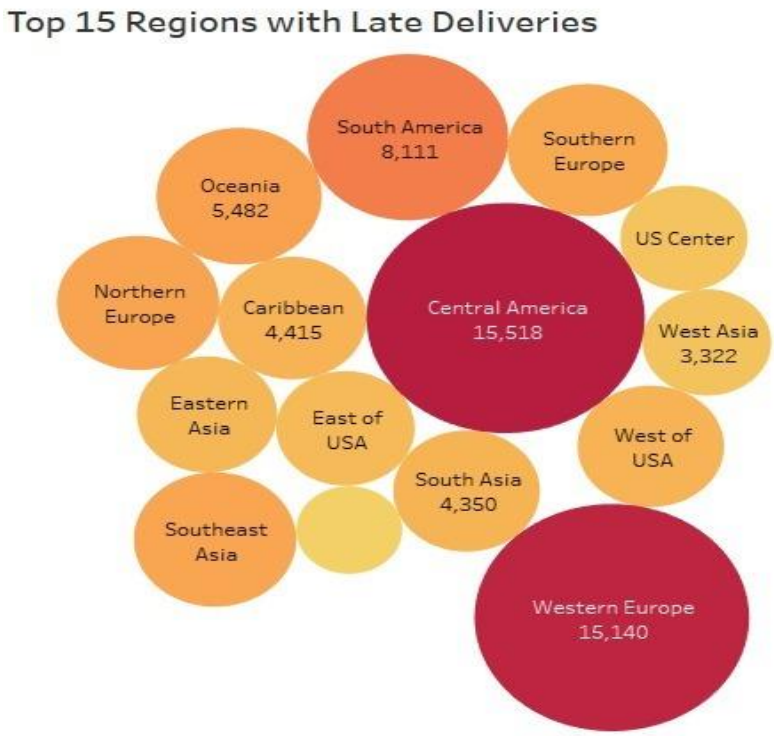


Fig 7.13. Graph showing Top 15 Region with Late Deliveries

Central America and Western Europe are the two regions with most late deliveries. So company should take necessary measures to reduce late deliveries in these regions.

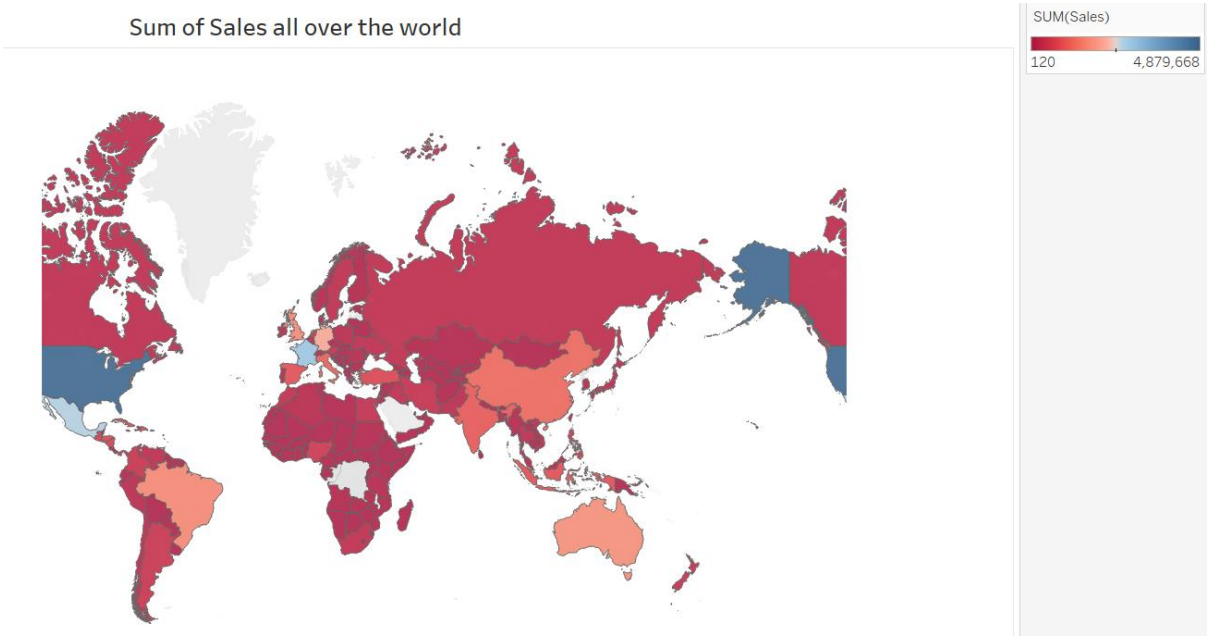


Fig 7.14. Graph showing Sum of Sales all over the world

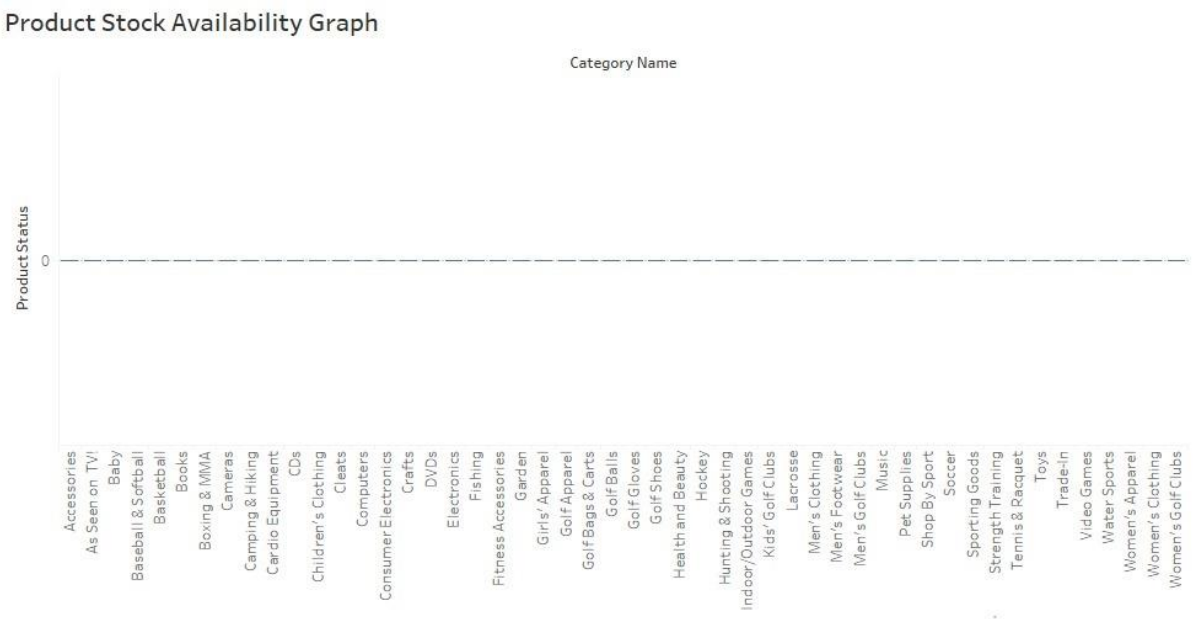


Fig 7.15. Graph showing Product Stock Availabilty

Product status defines the availability of the product.

- 0 Defines the availability of the product.
- 1 Defines non-availability of the product.

From the above graph, one can clearly observe that all product stocks are available as there is no 1 value.

So the problem of sales drop is definitely not with the availability of stocks.

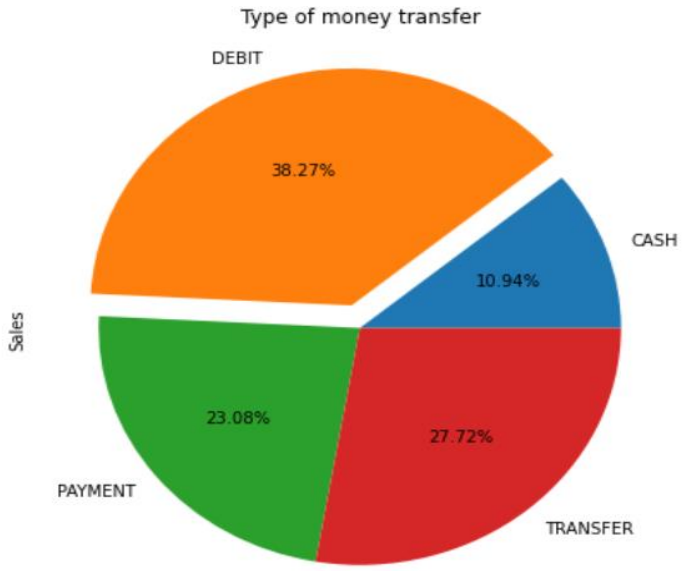


Fig 7.16. Pie graph showing Mode of transfers

Since the debit transfer is more compared to other transfers, the company can make an agreement with some of the major debit card companies to attract customers.

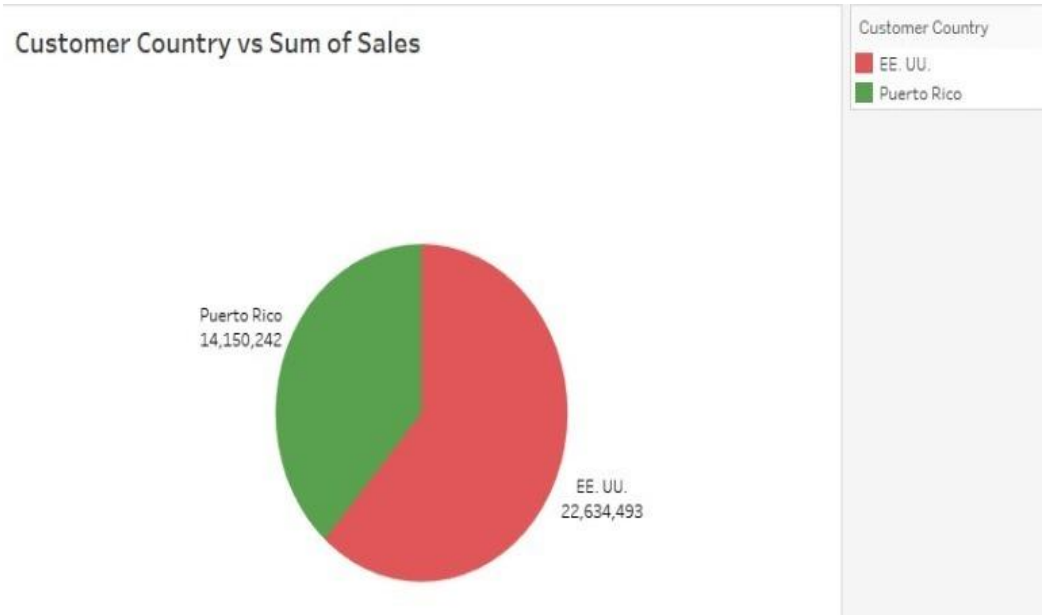


Fig 7.17. Pie graph showing Customer Country v/s Sum of Sales

Customers are only from two country whereas orders are being delivered to 164 countries all over the world. So the company should advertise itself to reach larger range of customers all over the world in order to increase sales.

Overall inference of EDA :

Problem- The a plot explains the major drop in sales around September 2017 which is due to the increase in customer churn.

Reason-Majority of the orders were delivered late which leads to decrease in customer satisfaction and could be one of the reasons for the drop in sales and for the increase in customer churn which is explained in b plot.

Solution-We have decided to build a model to predict the late delivery risk so that the company can make sure that they would deliver the product with late delivery risk on time which would increase customer satisfaction resulting in decrease of customer churn. While the remaining plots would lead to general suggestions to decrease customer churn which are mentioned below the respective plots.

8. IMPLICATION:

- Our solution is instrumental in predicting the late deliveries to the customers due to the various factors affecting it.
- By looking at the important customer cities, the company can plan well before the deliveries to overcome any challenges that is hampering the delivery time.
- This in turn will lead to customer satisfaction, less churn and ultimately yields more profits to the firm.

9. LIMITATIONS:

Various limitations are stated as below:

- The dataset for a supply chain domain is generally very large in terms of records and its various factors.
- There are a lot of external factors for late delivery apart from the ones which can be recorded like the transportation laws across borders, inter/intra state disputes resulting etc.

- Man-made errors are also a major reason for struggle that this industry like man-handling the orders, incorrect data entry into the system etc**.**

10. CLOSING REFLECTION:

Since the supply chain dataset which we had considered was large and unprocessed. We had to deal with all the cleaning techniques like missing values, outlier treatment, insignificant variables, redundant variables, multicollinearity etc.

We also used variance of variables for feature selection along with chi2 , anova to check the dependency of dependent variable with the independent features and its classes. We had also considered information gain of every feature for feature selection.

The major challenge was dealing categorical variables with high cardinality which we managed by encoding them by using one hot encoding and label encoding.

This above combination of one hot and label encoding managed to produce data which was able to explain only 85% of variance present in the target variable through GradientboostingClassifier.

Next time we would like to learn various techniques to deal with categorical variables with high cardinality**.**

INFERENCE AND RECOMMENDATION:

From our best model we observed that, Features like Shipping Mode, Customer city, order city, order state customer state and order country are the features having the highest impact on the delivery timing i.e either it is late or not.

- Referring to **Fig 5.7** Shipping mode like standard class has shown that in this mode, the orders are not late but other shipping mode like first class and second class have comparatively higher number of orders being late which suggests that in these shipping mode the promised timeline for delivery is over committed which is unrealistic. So company might need to

change the delivery timelines for these modes so as to deliver it on stipulated time.

- Referring to table **Table 4.V.3** Order city is another important feature where cities like Bhopal, chesterfield, Glasgow etc are quite significant with late deliveries which is confirmed by chi2 test as well as shown above. So, whenever a order is placed to be delivered to above mentioned cities, the company should be more cautious and look into the transportation rules, exact location and try to mitigate the delivery challenges at these locations
- Referring to table **Table 4.V.4** Customer city, meaning the cities from where the orders are placed is another important factor for delivery promptness. Cities like Jamaica(NY) are very significant and the company should look into the above customer city as to what special instructions while placing an order is there so that they can avoid any late deliveries for customer belonging to these cities.
- Order states like Albany, Agra, Belfort etc have significant effect on late deliveries and the company has to give priorities to these states for delivery
- Colorado is the only state from where when an order is placed, it is being delivered late which is shown by the statistical test as well. So, if the order is being placed from Colorado, it is possible that the product being order is exclusively available and the location these customers want to deliver are also very remote.
- Countries like Cuba, Ecuador, Estonia and togo are the countries where there have been lot of late deliveries. The custom rules in these countries are very time consuming and strict and hence to overcome this, there should be proper paper works duly signed and authorized so as avoid any such man-made delays.