horizontal line

Summary

|  |  |
| --- | --- |
| **BATCH DETAILS** | **PGPDSEFT CHENNAI NOV’21** |
| **TEAM MEMBERS** | **BALA MURUGAN M**  **KAMALNATH A**  **KIRUBHAKARAN R**  **PRASAD ANANTHA NARAYANAN**  **S VIVEK** |
| **DOMAIN OF PROJECT** | **E-COMMERCE** |
| **PROPOSED PROJECT TITLE** | **PRODUCT SHIPMENT ARRIVAL ON TIME** |
| **GROUP NUMBER** | **CAPSTONE PROJECT GROUP 3** |
| **TEAM LEADER** | **KAMALNATH A** |
| **MENTOR NAME** | **ANJANA AGRAWAL** |

DATE: 07TH April 2022

**SIGNATURE OF THE MENTOR SIGNATURE OF THE TEAM LEADER**

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| SL NO | TOPIC | PAGE NO |
| 1 | Industry Review | 3 |
| 2 | Dataset and Domain | 6 |
| 3 | Data Exploration (EDA) | 10 |
| 4 | Feature Engineering | 23 |
| 5 | Model Building | 24 |
| 6 | Advance ML models | 27 |
| 7 | Summary | 32 |

**1.Industry Review**

**1.1 Current practices**

The E-Commerce industry is one of the most fast growing and booming industry over the last few years capitalizing on the Internet penetration and Digital financial services, which has in turn affected the sales for many brick-and-mortar stores. Online Retail platforms such as Flipkart, Amazon and so on, allow customers to purchase products directly through their websites or mobile applications. Customers now easily buy, return and cancel the products without even going to the stores. E-Commerce websites provide details about the products, real-time reviews and recommendations which in turn helps customers to select products wisely. Product price in online E-Commerce website are comparatively lesser than the offline stores. In most of the cases, we can observe that online platforms drive offline sales through provisions to automatically manage the product inventory. Online Retail provides opportunities for small to large stores and has proven to be one of the most trusted platforms by both buyers and sellers.

* 1. **Background Research**

Acording to a survey conducted by the Interactive Media in Retail Group (IMRG), Online shopping has increased drastically about 5000% when compared to a decade ago in the United States. The consumer's way of shopping and expenditure has changed a lot in the past 10 years. The exponential increase indicates the interest shown by the consumers in the online retail when compared to in-store purchases. Ecommerce shipping encompasses receiving and processing orders, picking and packing the purchased product at a warehouse, printing shipping labels, and even managing returns.  That might sound simple enough. If you’ve seen a well-run warehouse, it might even looksimple. But in fact, all the moving pieces create a level of complexity that depends on the types and sizes of products you sell, to which regions you’ll deliver, the delivery options and shipping speeds you choose to offer and more.

* 1. **Literature Survey**

**1.3.1** A Review of Management and Importance of E-Commerce Implementation in Service Delivery of Private Express Enterprises of China:

China’s EDI started the express delivery market; it continued to expand, and as a result, EDCs grew and gained significant financial results. This industry is developing quite rapidly. At present, four regional express delivery centers have been formed in the eastern region. At the same time, the four major express industries are developing through the rolling and progressive sectors in the central and western regions. Some large cities and megacities have become regional centers for the development of the EDI. A number of express delivery channels based on underlying transport lines have been formed nationwide. The express delivery system of China is centered on the more sensitive transportation demandat thetime. For most EDCs, managers should strive to strengthen enterpriseinformation construction and professional personnel training, unify the management ideas of outlets throughout the country, vigorously promote the standardization of work processes, increase the technological content of equipment, improve the business skills of employees, and provide better customer service (Xiao-Hui, 2014). They should create a private express brand, create more employment opportunities, change people’s lifestyle, improve living standards, and promote national economic development.

For EDCs, the commerce of any business has a key indirect impact, which is a way of buying and selling goods through an electronic network. This new technology encounters certain problems, especially in the logistics areas. Simultaneously, it has helped the development of opportunities in the logistics business, together with the unmanageable development in the logistics. Without e-commerce, we could not be able to reach this emancipation that we all observe today. However, e-commerce logistics are not well developed enough even in thickly populated China. With the regional variation, delivery time also varies, and it is difficult to accomplish on-time delivery in specific areas. The reason behind this is the construction of logistics infrastructures, which are not unified perhaps because of too low efficiency of logistics (Hou, 2014).

**1.3.2** Coordination Analysis of Revenue Sharing in E-Commerce Logistics Service Supply Chain With Cooperative Distribution:

Weltevreden (2008) introduced two types of CDPs: one of which is unattended lock point (ULP) and the other is an attended service point (ASP). ASP is based on the service resources of residential community and campus, such as chain convenience store, campus service center, and property management center. It establishes e-commerce logistics terminal distribution outlets for resource sharing. Comparing the two types, ASP can provide customers with better online shopping experience, including more payment choices and flexible package size, less operational knowledge, and so on. Furthermore, Weltevreden (2008) systematically summarized the literature and used questionnaires and interviews to empirically analyze the application of CDPs in the Netherlands. It was found that the CDPs were very popular and the number grew rapidly, bringing additional benefits to the retail outlets participating in self-financing services.

Literature which is closely related to this research can be summarized and classified in three broad and diverse categories: the distribution service points of e-commerce logistics, revenue-sharing contract, and game theory in supply chains.In the e-commerce logistics delivery cycle, customers cannot stay home all the daytime, and many packages are not suitable for mailboxes. Thus, the courier’s home delivery service cannot be successful at one time and has to be repeated, thereby increasing the cost of logistics (Weltevreden, 2008). As a result, European and American countries developed a delivery device near the residential area or transportation site, called collection-and-delivery points (CDPs), which supplemented or even replaced home delivery services. Customers can use this equipment to take the package of online shopping and also initiate reverse logistics operation of return. Most recently, this type of CDPs has been widely applied in China’s e-commerce terminal distribution network and has become an important form of China’s sharing economy.

**1.3.3** .A Research on E-Commerce Shipping Platform Models In the Context of “Internet Plus:

Zikun Zhai (2019),In cyber-age, with the creation and growth of e-commerce platforms online causing profound changes to the way of business-doing in traditional sectors, “Internet Plus shipping” is now well predicted as an inevitable trend shipping business. As e-commerce shipping platforms make advertising, publication of information, business communication and transaction easier and more convenient, they are becoming a driving force accelerating the development of shipping industry. In domestic market today, there are already a variety of comparably matured models with which online shipping platforms operate, however, through evaluating their performances; these models have both unique advantages and disadvantages or deficiencies. The paper in the first place gives a summarization to what exactly e-commerce shipping can do, then introduces some most recognizable operational models on current stage, and lastly, from the perspective of “Internet Plus”, offers some proposals with the hope of helping advance its future development.

**1.3.4**.The Impact of E-commerce on Final Deliveries:

(Lierow et al., 2013)According to our analysis, delivery services are amongst the top concerns of both e-shoppers and e-retailers in the EU. E-retailers themselves are considering selectively moving into logistics to offer improved delivery and return experiences. The current challenge is same-day delivery, with direct consequences on the transport market. One example is Amazon, the company is no longer satisfied with merely offering later outbound cut-off times than other e-commerce players (Lierow et al., 2013). It may consider moving into same-day logistics (in Germany) if no logistics partner steps up to offer the service. Recently, DHL is experimenting with same-day deliveries in the Cologne area in cooperation with an online food retailer. In addition, e-retailers are developing sophisticated transport purchasing systems and aggregating increasing volumes of parcels through third-party fulfilment businesses, thus increasing their bargaining power with delivery logistics providers. Some new intermediaries could also emerge: Shutl and Tiramizoo provide back-end services matching an ecommerce order with a courier service for delivery within 90 minutes. This service is currently available in the UKand Germany.

Delivery services such as next-day home delivery are becoming increasingly common. Including the cost of delivery in the cost of the product (rather than having it as a separate line item) and the increasing use of flat-rate subscription models (such as Amazon Prime) are augmenting the need to lower delivery costs. As a result, third-party delivery logistics providers are seeing greater fluctuation in volumes when e-commerce players shift volumes between 3PL providers. Prices for next-day delivery in Germany are hovering around two Euros per parcel, and are only slightly higher in the UK and France .

**2.Dataset and Domain**

**2.1** Data Dictionary

The data contains the following information:

* **ID:** ID Number of Customers.
* **Warehouse block:** The Company have big Warehouse which is divided in to block such as A,B,C,D,E.
* **Mode of shipment:** The Company Ships the products in multiple way such as Ship, Flight and Road.
* **Customer care calls:** The number of calls made from enquiry for enquiry of the shipment.
* **Customer rating:** The company has rated from every customer. 1 is the lowest (Worst), 5 is the highest (Best).
* **Cost of the product:** Cost of the Product in US Dollars.
* **Prior purchases:** The Number of Prior Purchase.
* **Product importance:** The company has categorized the product in the various parameter such as low, medium, high.
* **Gender:** Male and Female.
* **Discount offered:** Discount offered on that specific product.
* **Weight in gms:** It is the weight in grams.
* **Reached on time:** It is the target variable, where 1 Indicates that the product has NOT reached on time and 0 indicates it has reached on time.

**2.2** Variable categorization

There are 8 categorical variables and 3 numerical Variables and there are 11 features and 10999 rows.

The Numerical Features are:

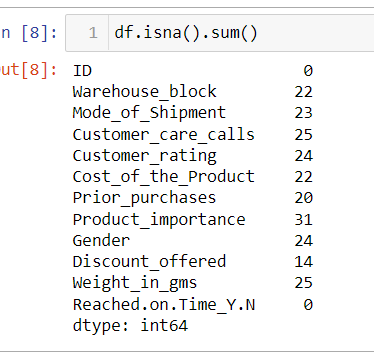
* ID
* Discount\_offered
* Weight\_in\_gms

The Categorical Features are:

* Warehouse\_block
* Mode\_of\_Shipment
* Product\_importance
* Gender
* Customer\_care\_calls
* Customer\_rating
* Cost\_of\_the\_Product
* Prior\_purchases

* 1. Pre Processing Data Analysis
* Identifying the missing value

Below are the number of missing values in each features.



* Missing value Treatment

I. Following features have been imputed using mode since they are categorical variable:

Warehouse\_block, Mode\_of\_Shipment, Customer\_rating, Prior\_purchases, Gender, Product\_importance

II. Following features have been imputed using median since they are continuous variable:

Weight\_in\_gms, Discount\_offered

III. Cost\_of\_the\_Product have been imputed based on Customer\_care\_calls,

Since there is a relationship between them (when average Cost\_of\_the\_Product of the product increases then Customer\_care\_calls also increases ).So we grouped Customer\_care\_calls and took the average of Cost\_of\_the\_Product and then imputed those average cost based on the number of Customer\_care\_calls that specific order has received.

df['Cost\_of\_the\_Product'].fillna(df.groupby('Customer\_care\_calls')['Cost\_of\_the\_Product'].transform('mean'),inplace=True)

IV. Likewise, Customer\_care\_calls have been imputed based on the cost of the product

A= df[df['Customer\_care\_calls'].isna()].index

for i in A:

if df.loc[i,'Cost\_of\_the\_Product']<=189.628931:

df.loc[i,'Customer\_care\_calls']=2

if (df.loc[i,'Cost\_of\_the\_Product']>189.628931) & (df.loc[i,'Cost\_of\_the\_Product']<=198.883786):

df.loc[i,'Customer\_care\_calls']=3

if (df.loc[i,'Cost\_of\_the\_Product']>198.883786) & (df.loc[i,'Cost\_of\_the\_Product']<=206.406780):

df.loc[i,'Customer\_care\_calls']=4

if (df.loc[i,'Cost\_of\_the\_Product']>206.406780) & (df.loc[i,'Cost\_of\_the\_Product']<=214.276916):

df.loc[i,'Customer\_care\_calls']=5

if (df.loc[i,'Cost\_of\_the\_Product']>214.276916) & (df.loc[i,'Cost\_of\_the\_Product']<=246.646825):

df.loc[i,'Customer\_care\_calls']=6

if (df.loc[i,'Cost\_of\_the\_Product']>246.646825) :

df.loc[i,'Customer\_care\_calls']=7

* Checking Duplicate Entries:

There is no duplicate record in the dataset.

* Redundant Columns:

ID can be removed from the further analysis, since it doesn’t give any inference/ or it not useful feature for the analysis.

* 1. Project Justification

Project Statement

An international Retail brand is involved in the sales of unique gifts for all occasions. They have recently launched a website to sell their products through online in order to expand their business and wants to discover key insights from their customer database. The ultimate goal is to find whether the products reach customers on time, whether the customer query is being answered and to find which mode of shipment plays major role in delivering the product on time.

Complexity involved:

* Understanding the relationship between the variables.
* Identifying the pattern for each features.
* Encoding the categorical variables and identifying which Machine Learning model to choose.

Outcomes:

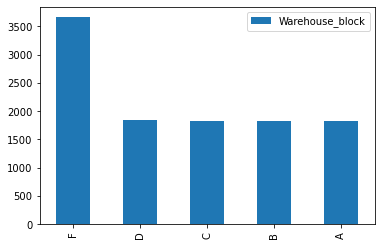
* This will be a business value problem to avoid the loss of client and this delivery part will play vital role on customer faith
* Once we found out the shipment delivery preciseness of products, that might help out to customer retention, higher rating product, sales ratio can be attains to business goals.
* The significant factors affecting the delay of shipment is analysed based on the dataset and is used to create a model
* This will help the E -Commerce company to improve their shipment processs.

**3. Data Exploration (EDA)**

# Understanding each variable:

Understanding how each variable is distributed.

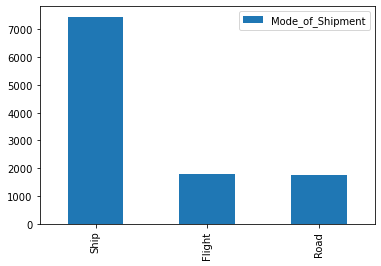
1.Warehouse block



Inference:

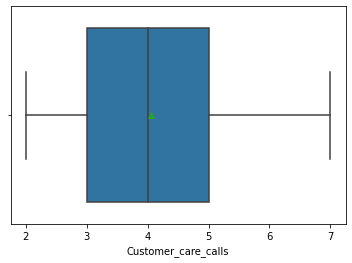
F block has twice the number of products that has been stored in the other in other individual blocks, so we can assume that this block has more storage capacity than the other Warehouse block.

2. Mode of shipment



Inference: 68% of the shipment has been done via ship.

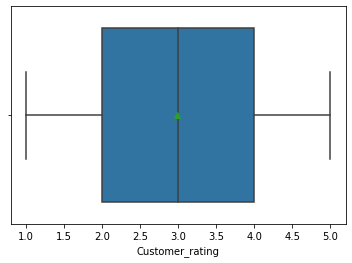
3.Customer care calls



Inference:

On an average each order gets 4 calls and from the data we can say that each order gets at least 2 calls.

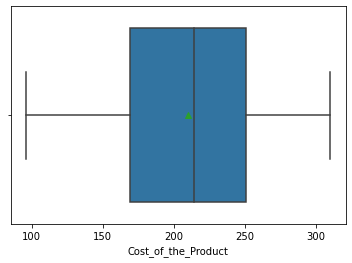
4. Customer rating



Inference:

On an average the customer rating is 3.

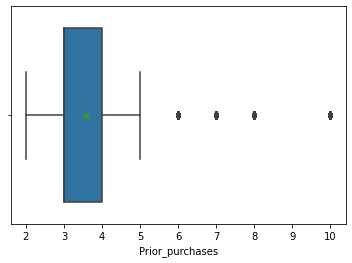
5. Cost of the Product



Inference:

The cost of the product range varies from 96 to 310.

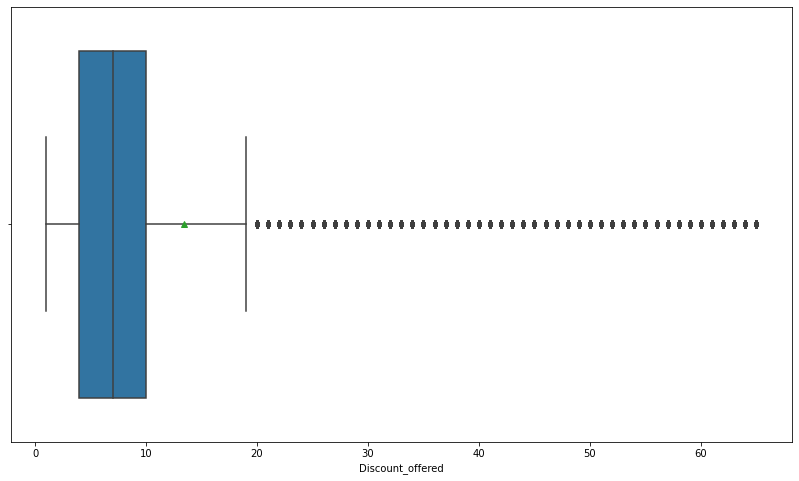
6. Prior purchases



Inference:

Every customer has prior purchases and on an average 3 purchases have been done.

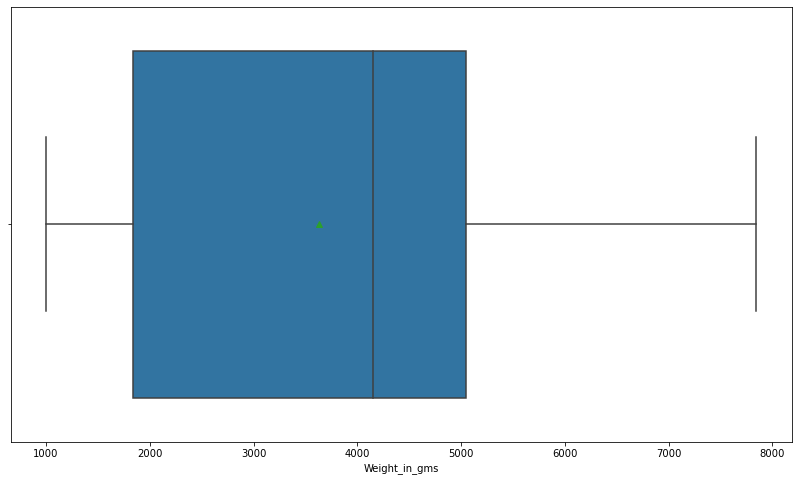
7. Discount offered



Inference:

About 75% of the product has been offered at a discount of less than 10%.

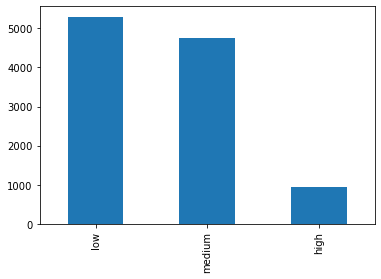
8: Weight in gms



Inference:

Weight ranges between 1kg-8kg.

9. Product importance

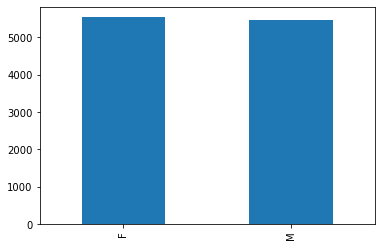
Inference:

48% - low importance products

43% - medium importance products

8% - high importance products

10. Gender



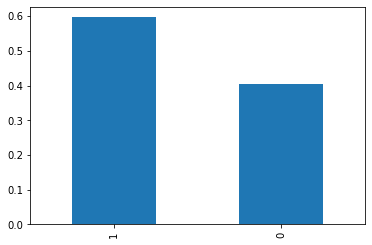
Inference:

We have the same gender proportion.

**Presence of outliers**

We have extreme values in discounts offered and this is not considered to be outliers because these values are important from a business perspective.

**Class imbalance and its treatment**



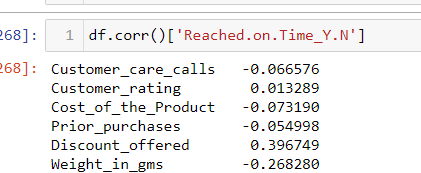
The ratio is 6:4, it is not much imbalanced, so we proceed with the same trend as the training set learns the same pattern.

**multi-collinearity –** There is no multi- collinearity between the independent variables



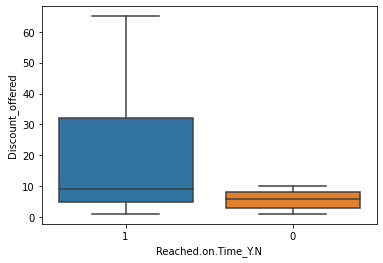
Correlation of target variable with other numerical variable – Except Discount\_offered there is

no Correlation between independent variable and dependent variable.



**RELATIONSHIP BETWEEN VARIABLES**

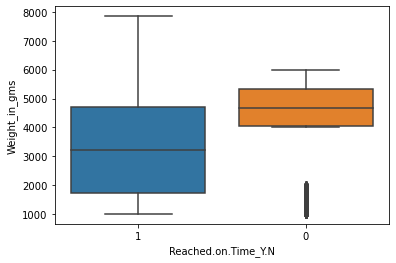
The dataset consists of numerical variables which are significant in determining whether the product reached on time or not. Hence, Bivariate & multivariate techniques are deployed to see the relationship between numerical v/s categorical variables ,numerical v/s numerical and as well as categorical v/s categorical.

1. Reached on time vs Discount offered

Inference:

Products with higher discounts(more than 10) are not reaching on time. Products whose offer is less than or equal to 10% has a high chance of getting delivered on time.

2. Reached on time vs Weight in gms

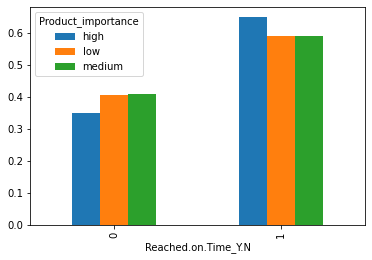


Inference:

Products weighing between 1000gms-4000gms and 6000gms-8000gms are not reaching on time except few products in range 1000gms-4000gms

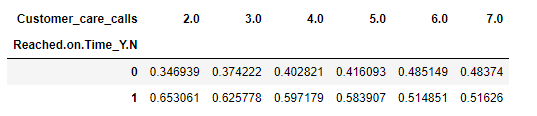
Products weighing between 4000 gms-6000 gms have a high chance of getting delivered on time.

3. Reached on time vs Product importance

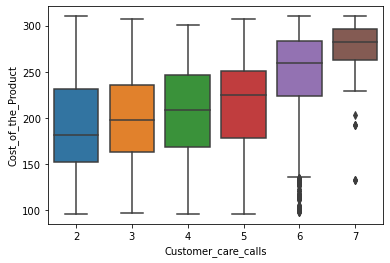
Inference:

65% of the product with high importance has not reached on time, whereas other products have a low ratio compared to it.

4. Reached.on.Time\_Y.N vs Customer care calls



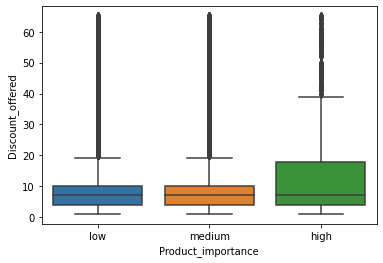
In all the number of customer care calls made, the percentage of not reaching on time is high compared with reaching on time but the trend is increasing with 0 and decreasing with 1.

5. Customer care calls vs Cost of the product

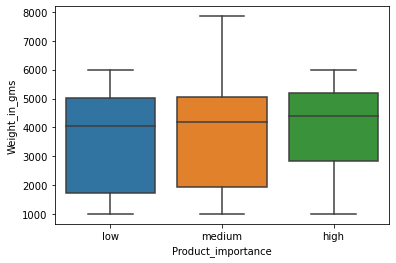
Inference:

With increase in cost of the product the number of customer care calls as increased

6. Product importance vs Discount offered

Inference:

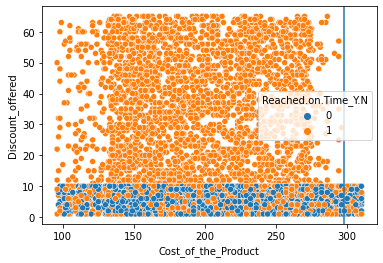
More than 25% of the high importance products have a discount of more than 10%. While others have only 25% of the products that has a discount of more than 10%

7. Product importance vs Weight in gms

Inference:

Only medium importance products have a weight of 6-8kg.

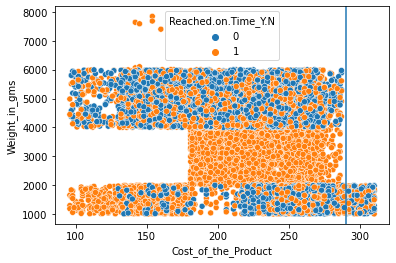
8. Cost of the Product vs Discount offered vs Reached on time



Inference:

More products with discounts between the range 0% to 10% are delivered on time irrespective of the cost. When the cost of the product is more than 300, then the discount was not more than 10%

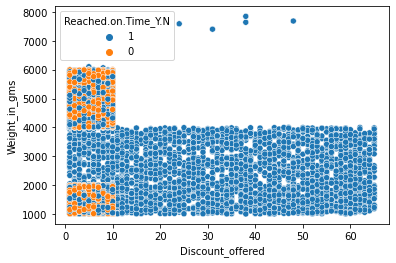
9. Cost of the Product vs Weight in grams vs Reached on time



Inference:

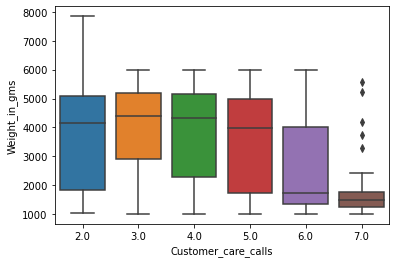
When the cost of the product is more than 290, then the weight of the product more than 2000 gms

10. Discount\_offered vs Weight in gms vs Reached on time

Inference:

The discount offered is in the range of 0% - 10%, when weight of the product is more than 4000 gms

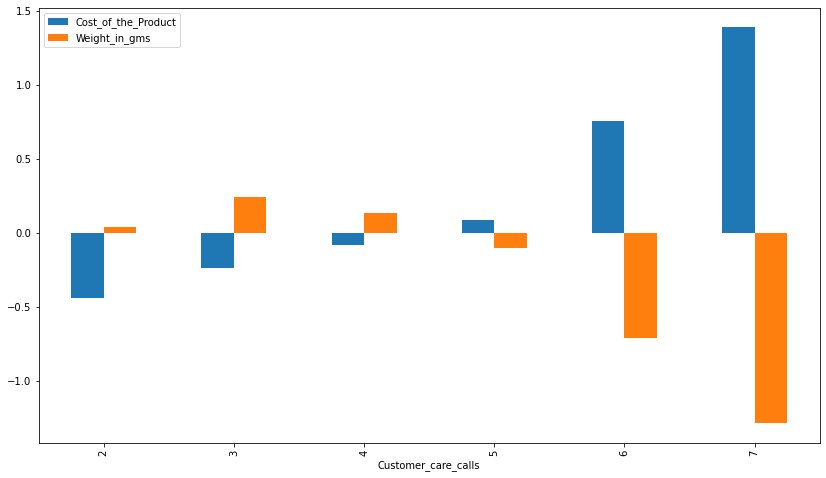
11. Customer care calls vs Weight in gms



Inference:

Lower weight products are getting the most number of calls.

12. Customer care calls vs Weight in gms vs Cost of the product



Inference:

The number of customer care call increases with increase in cost of the product and with decrease in weight

**Statistical Significance of Variables**

**For cat vs cat we use chi2 test**

P value with correspond to dependant variable are:

Warehouse\_block is 0.9007206555567301

Mode\_of\_Shipment is 0.6728059239062216

Product\_importance is 0.002760734726303322

Gender is 0.5597150305785878

Customer\_care\_calls is 2.571118384515917e-10

Customer\_rating is 0.49539819318961187

Prior\_purchases is 2.529093788969353e-23

**For numerical vs cat we use mann whitney u(non parametric test)**

P value with correspond to dependant variable are:

Cost\_of\_the\_Product is 0.0

Discount\_offered is 0.0

Weight\_in\_gms is 0.0

The significant variables are Product\_importance, Customer\_care\_calls, Prior\_purchases, Cost\_of\_the\_Product, Discount\_offered, Weight\_in\_gms.

**4. Feature Engineering**

* Scaling : The data is scaled using Z / StandardScaler transformation. StandardScaler is useful for the features that follow a Normal distribution. It will making data points generalized so that the distance between them will be lower. StandardScaler removes the mean and scales each feature/variable to unit variance.
* Feature selection : Performed Label encoding for Warehouse\_block, Mode\_of\_Shipment, Gender and performed Ordinal encoding for Product\_importance.

**5. Model building**

* 1. **Logistic Regression as Base Model**

**Assumption**

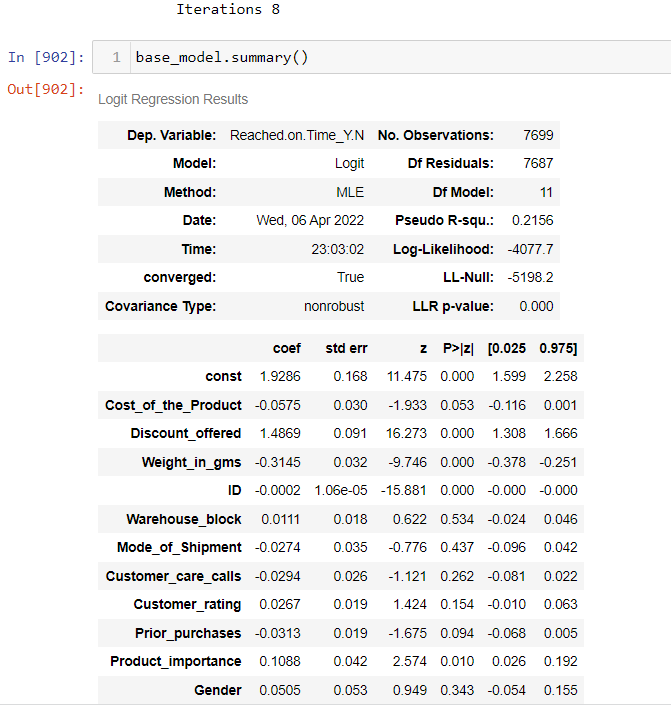
1)Independence of error ,whereby all sample group outcomes are separate from each other (i.e there is no duplicate responses).

2)Absence of multicollinearity.

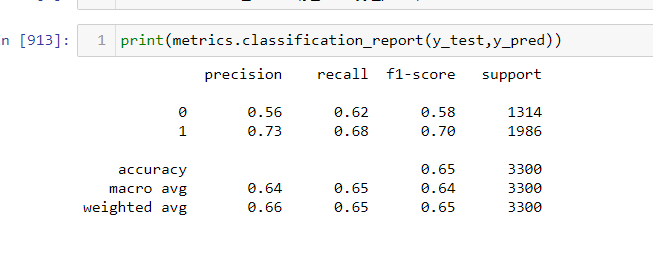
3)Lack of strongly influential outliers.

4)On top of it, since our target variable is categorical we selected logistic regression as base model.

**Outcome**



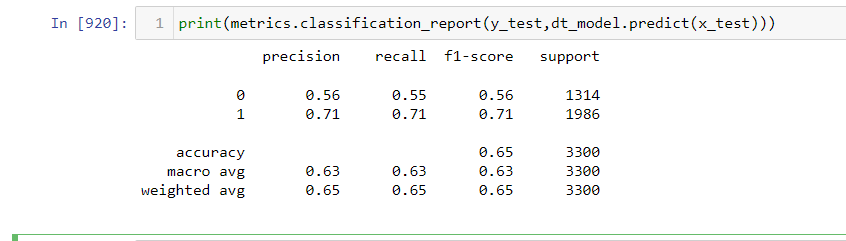
**Classification Report -** To get even more insight into model performance, we should examine other metrics like precision, recall, and F1 score.

****

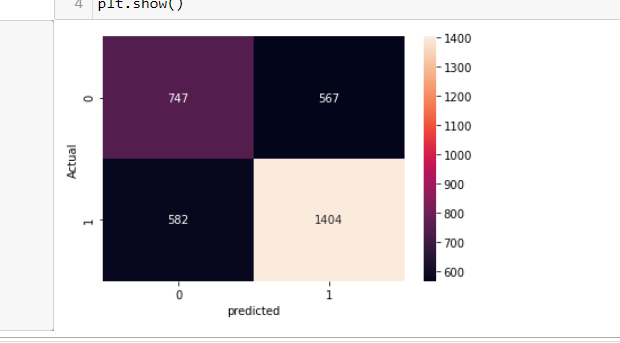
**5.2 Decision Tree Model**

* Decision trees are commonly used in **operations research, specifically in decision analysis**, to help identify a strategy most likely to reach a goal, but are also a popular tool in machine learning.
* The main advantage of the decision tree classifier is its ability to using different feature subsets and decision rules at different stages of classification
* Decision tree using greedy algorithm and it makes most optimal decision and recursive meaning it splits the larger question into smaller questions and resolve them the same way.
* In our business problem to predict the model based on Shipment delivered on time or not. We have to identify the Bivariate target value. Decision tree method is most familiar and simple to understand and interpret into business people.

**Classification Report -** To get even more insight into model performance, we should examine other metrics like precision, recall, and F1 score



**Confusion Matrix -** confusion matrix is a table that shows how well a classification model (or "classifier") performs on a set of test data for which the true values are known.

****

6.Advance ML Models

After building two base models, We built other classification and boosting models. In this Business case scenario, we can consider the Precision score of the majority class to justify the model prediction. From all models that we got the precision score of

**We have chosen precision score for measuring the performance as we were more concerned about predicting the actual true positive**

**MODEL** **Precision score**

Logistic Regression 0.543648

Naive Bayes 0.530057

K-nearest neighbours 0.540268

Random Forest 0.561006

Adaboost 0.556310

Gradientboosting 0.557868

Xgboost 0.563969

From the above results, we can select the most highest score is from Xgboost model, But that is also gives lower score of 56%. We have to take other steps to improve the precision score of all the models once again.

**Feature Selection**

In feature selection, we used to get important features from all the models and mostly used features of top 3 from all the models and append it in a list to find out what are all the most effective features would used to build the models

From that we got 5 most important features that should impact the model building are

1.Weight\_in\_gms

2.Cost\_of\_the\_Product

3.Customer\_care\_calls

4.Discount\_offered

5.Prior\_purchases

Choosing this features in Independent variables to train and test the dataset and Build the model by important features.

**Important Features Model Score**

After Building with only those five important features we got the scores of

**MODEL** **Precision score**

Logistic Regression 0.536011

Naive Bayes 0.530779

K-nearest neighbours 0.551653

Random Forest 0.569251

Adaboost 0.559385

Gradientboosting 0.560999

Xgboost 0.546826

By Selecting important features and Build the model thus we got the highest score of 56.9% from the Random Forest model. This score is also Low for build a good model, Then we will try out next steps to improve the score of precision and that will help us to find out the True positive predictions from actual dataset.

Club and trasnformation

**Hyperparameter Tuning**

When creating a machine learning model, you'll be presented with design choices as to how to define your model architecture. Often times, we don't immediately know what the optimal model architecture should be for a given model, and thus we'd like to be able to explore a range of possibilities. In true machine learning fashion, we'll ideally ask the machine to perform this exploration and select the optimal model architecture automatically. Parameters which define the model architecture are referred to as **hyperparameters** and thus this process of searching for the ideal model architecture is referred to as hyperparameter tuning*.*

These hyperparameters might address model design questions such as:

* What degree of [polynomial features](https://www.jeremyjordan.me/polynomial-regression/) should I use for my [linear model](https://www.jeremyjordan.me/linear-regression/)?
* What should be the maximum depth allowed for my [decision tree](https://www.jeremyjordan.me/decision-trees-for-classification/)?
* What should be the minimum number of samples required at a leaf node in my decision tree?
* How many trees should I include in my [random forest](https://www.jeremyjordan.me/ensemble-learning/)?
* What should I set my learning rate to for gradient descent?

**Grid Search CV**

From Grid Search CV, we can find out the most important parameters used for each and individual models. There after once again to build the all models with tuned hyper parameters, there may be chance of increasing the precision score as well the overall accuracy of the model.

In this case, Precision score has not improved for various Models which has been built using hyperparameter tuning and there is only a slight difference when compared with base models. so we can conclude that hyperparameter tuning is not giving good results for our data

**Transformation and Clubbing the Data**

Simply put, data transformation makes your data useful. Data transformation is the process in which you take data from its raw, siloed and normalized source state and transform it into data that’s joined together, dimensionally modeled, de-normalized, and ready for analysis. Building and training models to process data is a brilliant concept, and more enterprises have adopted, or plan to deploy, machine learning to handle many practical applications. But for models to learn from data to make valuable predictions, the data itself must be organized to ensure its analysis yield valuable insights.

**Why need data transformation?**

* the algorithm is more likely to be biased when the data distribution is skewed
* transforming data into the same scale allows the algorithm to compare the relative relationship between data points better

There are different types of transformation available such as MinMax, Standard Scaler , Log transformation etc.

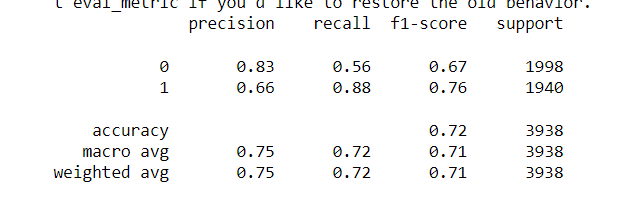
According to our dataset we have used Log transformation to build model. When the data sample follows the **power law distribution**, we can use log scaling to transform the right skewed distribution into normal distribution. To achieve this, simply use the np.log() function. After the log transformation, features have become more normally distributed.

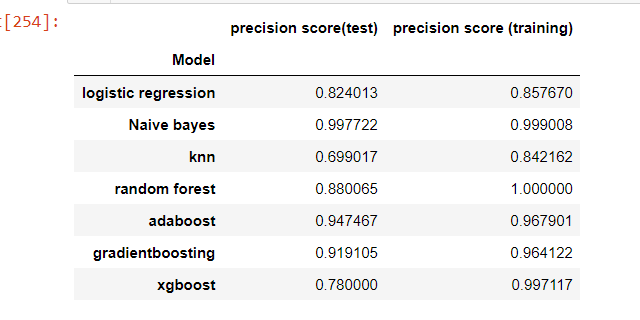
**Data Clubbing:** For few categorical variable such as Prior\_purchases, Customer\_care\_calls we have clubbed the data /replaced , say for example if Prior\_purchases is more than 6 ,then we have replaced those values as 6, like wise if Customer\_care\_calls is more than 6, then we have replaced those values as 6 and built model using Transformed data and clubbed data but still the precision sore has not improved

**Random Oversampling Imbalanced Datasets**

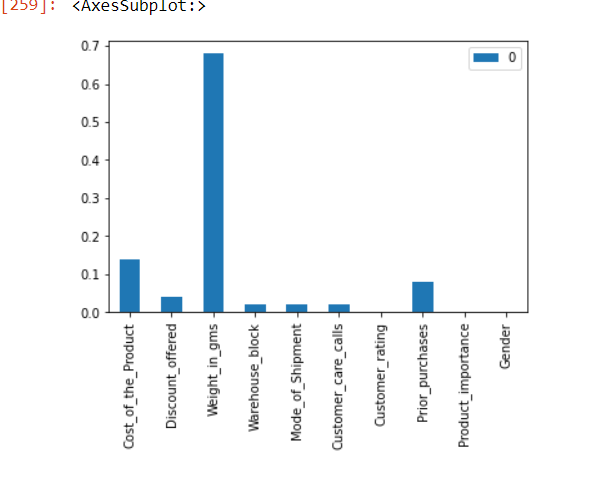
Random oversampling involves randomly duplicating examples from the minority class and adding them to the training dataset. Examples from the training dataset are selected randomly with replacement. This means that examples from the minority class can be chosen and added to the new “more balanced” training dataset multiple times; they are selected from the original training dataset, added to the new training dataset, and then returned or “replaced” in the original dataset, allowing them to be selected again. This technique can be effective for those machine learning algorithms that are affected by a skewed distribution and where multiple duplicate examples for a given class can influence the fit of the model. This might include algorithms that iteratively learn coefficients, like artificial neural networks that use stochastic gradient descent. It can also affect models that seek good splits of the data, such as support vector machines and decision trees. It might be useful to tune the target class distribution. In some cases, seeking a balanced distribution for a severely imbalanced dataset can cause affected algorithms to overfit the minority class, leading to increased generalization error. The effect can be better performance on the training dataset, but worse performance on the holdout or test dataset. The increase in the number of examples for the minority class, especially if the class skew was severe, can also result in a marked increase in the computational cost when fitting the model, especially considering the model is seeing the same examples in the training dataset again and again.

Below is the classification report of XGB classifier. We can clearly see that precision score has rapidly increased after oversampling the target variable



Now, replacing not reached on time as minority (replacing 1 to 0 and 0 to 1 in target variable ) and building model using 3000 random samples. Below are the precision score for various models . Out of which all the models we can clearly see that Adaboosting classifier producing better Precision score when compared with other models.

Feature importance obtained from AdaBoost Classifier.



**7. Summary**

* Weight of the product, Cost of the product, Discount offered, Prior purchases are the features that has a high impact on products which are not reached on time.
* From EDA we understood that products with higher discounts is not reaching on time so the products with higher discounts has to be focussed and has to delivered on time.
* Products lighter in weight is not reaching on time so ways has to be created in order to make lighter weight products to reach on time.
* We could further improve the precision score/make the model better if we have other details such as Purchased date, delivered date, distance, shipment partners, vendor details, country manufactured.

|  |  |
| --- | --- |
| ORIGINAL OWNER OF DATA | PRACHI GOPALANI |
| DATA SET INFORMATION | E-COMMERCE SHIPPING DATA |
| ANY PAST RELEVANT ARTICLES USING THE DATASET |  |
| REFERENCE |  |
| LINK TO WEB PAGE | https://www.kaggle.com/datasets/prachi13/customer-analytics |

### 

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*