Project Summary

Shopping online is currently the need of the hour. Because of this COVID, it's not easy to walk in a store randomly and buy anything you want. We are trying to understand how is the product price varying with sales - Is there any increase in sales with the decrease in price at a day level.

Batch details	April’22 Bangalore
Team members	Krishan Kumar Vaibhav Kumar Mehta Ankit Jhajhria Ravi Ranjan Shaik Aqeeb
Domain of Project	Ecommerce Supply Chain
Proposed project title	Impact of Covid on online shopping
Group Number	Group - 2
Team Leader	Krishan Kumar
Mentor Name	Animesh Tiwari

Date:

Krishan Kumar

Signature of the Mentor

Signature of the Team-Leader

Table of Contents

OVERVIEW

Superstore Systems give your brand the platform to launch into the fast-emerging marketplace arena. The B2C marketplace accounts for more than 53% of online sales in 2018. Your own marketplace gives you ownership of the complete customer journey together with all the data. The $7 trillion B2B market is ripe for disruption as Amazon have shown with their rise in the B2B sector from $1B sales in 2016 to $10B in 2018.

The impact of the COVID-19 outbreak on small businesses has been staggering. All across the county, many businesses have had to reduce their hours and their staff. Some businesses have shut down their operations altogether in an attempt to help stop the spread, seeing huge drop-offs in profits and creating overwhelming uncertainty for their owners and employees.

Business problem statement (GOALS)

- Does product price have any impact on sales?
- IS there any way possible to improve the sales of the super stores?
- What could be the possible strategies that can be opted to cater the needs of Business.

1.Business Problem understanding

The Covid-19 pandemic has changed the buying behaviour of the consumers, there has been significant decline in the number of customers visiting to supermarkets, hypermarkets.
Almost a year since the nationwide lockdown was announced due to Covid -19 the panic buying of daily essentials was at surge. The online players struggled to meet the demands of the customers and local kiranas appeared to be the saviours. With shift due to pandemic the store owners have adopted the new and latest technologies to meet the expectation of customers. Customers got accustomed to the new way of shopping for daily needs and store owner have fast-tracked the push towards digital transformation which impacted the sales of supermarkets immensely.

2.Business Objective

The main business objective or the challenge faced by the hypermarkets in such precedented times is how to minimize the operational expenses and caters the need of the business stability. What should be done to face the challenges aroused due to declining sales because of the pandemic.

3.Approach

Data analysis, cleaning/pre-processing: The pre-processing of the dataset before performing ML functions involves the following:

3.A Structured Based Data Exploration

It is the very first step in EDA which can also be referred to as ***Understanding the MetaData*!** That’s correct, ‘Data about the Data’. It is here that we get the description of the data we have in our data frame.

3.B Descriptive Statistics

Now to know about the *characteristics of the data* set we will use the df. describe () method which by default gives the summary of all the *numerical* variables present in our data frame.

3.C.Handling Duplicates

This involves 2 steps:

1. Detecting duplicates and Removing duplicates.
2. To check for the duplicates in our data

Hereby duplicates mean the exact same **observations** repeating themselves. As we can see that there are no duplicate observations in our data and hence each observation is unique.

3.D.Handling Outliers

What are Outliers? Outliers are the extreme values on the low and the high side of the data. Handling Outliers involves 2 steps: Detecting outliers and Treatment of outliers.

Detecting Outliers

For this we consider any variable from our data frame and determine the upper cut off and the lower cut-off with the help of any of the 3 methods namely :

- Percentile Method
- IQR Method
- Standard Deviation Method

Let's consider the Purchase variable. Now we will be determining if there are any outliers in our data set using the IQR(Interquartile range) Method. What is this method about? You will get to know about it as we go along the process so let's start. Finding the minimum(p0), maximum(p100), first quartile(q1), second quartile(q2), the third quartile(q3), and the iqr(interquartile range) of the values in the Purchase variable.

3.E.Outlier Treatment

Do not worry about the data loss as here we are not going to remove any value from the variable but rather **clip** them. In this process, we replace the values falling outside the range with the lower or the upper cut-off accordingly. By this, the outliers are removed from the data and we get all the data within the range.

3.F. Handling Missing Values

What are Missing Values? Missing Values are the **unknown values** in the data. This involves 2 steps: Detecting the missing values and Treatment of the Missing Values

3.G.Missing Value Treatment

To treat the missing values we can opt for a method from the following :

- Drop the variable
- Drop the observation(s)
- Missing Value Imputation

For variable *Product_Category_2*, 31.56% of the values are missing. We should not drop such a large number of observations nor should we drop the variable itself hence we will go for imputation. Data Imputation is done on the Series. Here we replace the missing values with some value which could be static, mean, median, mode, or an output of a predictive model.

3.H. Univariate Analysis

In this type of analysis, we use a *single variable* and plot charts on it. Here the charts are created to see the *distribution* and the *composition* of the data depending on the type of variable namely categorical or numerical.

For Continuous Variables: To see the distribution of data we create Box plots and Histograms.

Topic Survey in brief

1.Problem understanding

The Covid-19 pandemic has changed the buying behaviour of the consumers, there has been significant decline in the number of customers visiting to supermarkets, hypermarkets. Almost a year since the nationwide lockdown was announced due to Covid -19 the panic buying of daily essentials was at surge. The online players struggled to meet the demands of the customers and local kiranas appeared to be the saviours. With shift due to pandemic the store owners have adopted the new and latest technologies to meet the expectation of customers. Customers got accustomed to the new way of shopping for daily needs and store owner have fast-tracked the push towards digital transformation which impacted the sales of supermarkets immensely.

2.Current solution to the problem

Customer retention is extremely critical to the health of any business, regardless ofsize or industry. It is a common representation of a business's ability to keep its existing customers and maximize its revenue. It is important to dig into the factorsthat drive your customer retention rate and generate opportunities to improve yourcustomer success strategy.

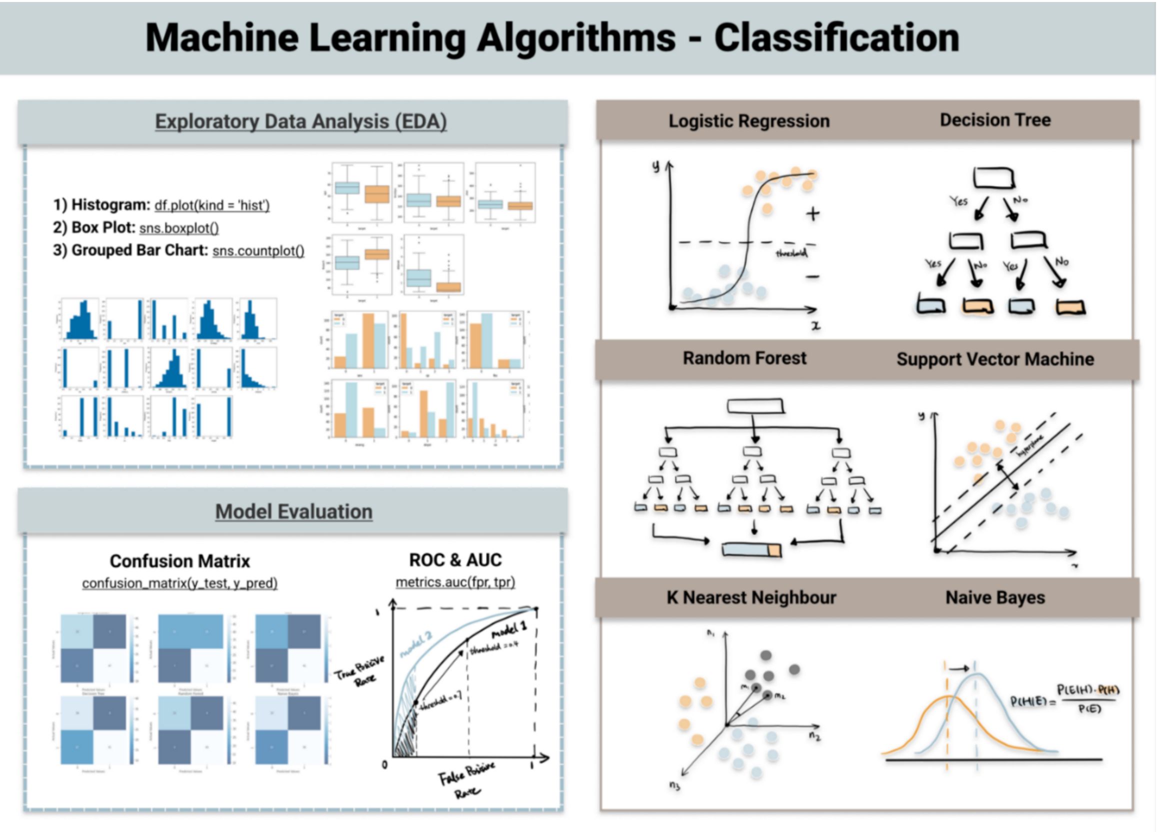
- Competitive Pricing.
- Offer incentives.
- Frequent Feedback from Customers.

- Seamless customer service.
- Wide network coverage.
- Exclusive benefits for existing customers.

3.Proposed solution to the problem

Exploratory Data analysis, Data Visualization, Building ML Models using different Algorithms to drive predictive analysis.

We are classifying the customers in different classes based on their behavior. That will help to analyze the patterns in consumer behavior.



Machine Learning Algorithms - Classification

Exploratory Data Analysis (EDA)

1) Histogram: df.plot(kind = 'hist')
2) Box Plot: sns.boxplot()
3) Grouped Bar Chart: sns.countplot()

Model Evaluation

Confusion Matrix
confusion_matrix(y_test, y_pred)

ROC & AUC
metrics.auc(fpr, tpr)

Logistic Regression

Decision Tree

Random Forest

Support Vector Machine

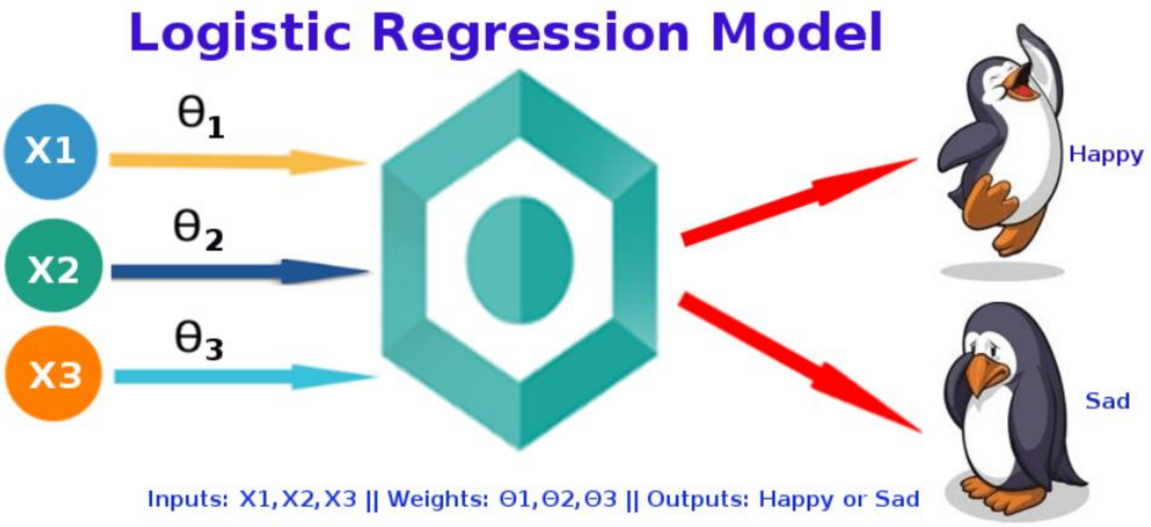
K Nearest Neighbour

Naive Bayes

Logistic regression

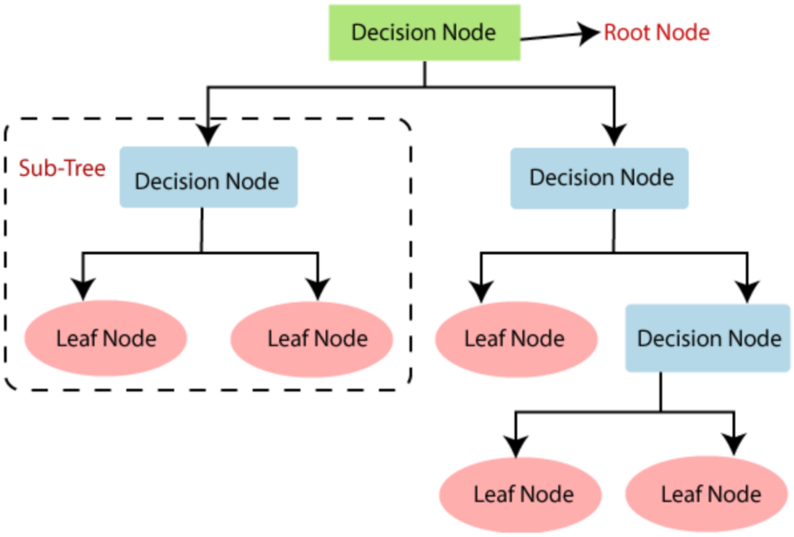
is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

Logistic Regression is another statistical analysis method borrowed by Machine Learning. It is used when our dependent variable is dichotomous or binary. It just means a variable that has only 2 outputs, for example, A person will survive this accident or not, The student will pass this exam or not. The outcome can either be yes or no (2 outputs). This regression technique is similar to linear regression and can be used to predict the **Probabilities** for classification problems.


Logistic Regression Model
X1
X2
X3
θ1
θ2
θ3
Happy
Sad
Inputs: X1,X2,X3 || Weights: Θ1,Θ2,Θ3 || Outputs: Happy or Sad

Decision Tree Algorithm

- is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.
- In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.
- The decisions or the test are performed on the basis of features of the given dataset.
- It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.
- It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.
- In order to build a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm.
- A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees.



Critical Assessment of Topic Survey

1. key area, gaps identified in the topic survey where the project can add value to the customers and business

The main business objective or the challenge faced by the hypermarkets in such precedented times is how to minimize the operational expenses and caters the need of the business stability. What should be done to face the challenges aroused due to declining sales because of the pandemic.

2.Key Gaps trying to Solve

Customers Analysis

Profile the customers based on their frequency of purchase - calculate frequency of purchase for each customer

Do the high frequent customers are contributing more revenue

Are they also profitable - what is the profit margin across the buckets

Which customer segment is most profitable in each year.

How the customers are distributed across the countries.

Product Analysis

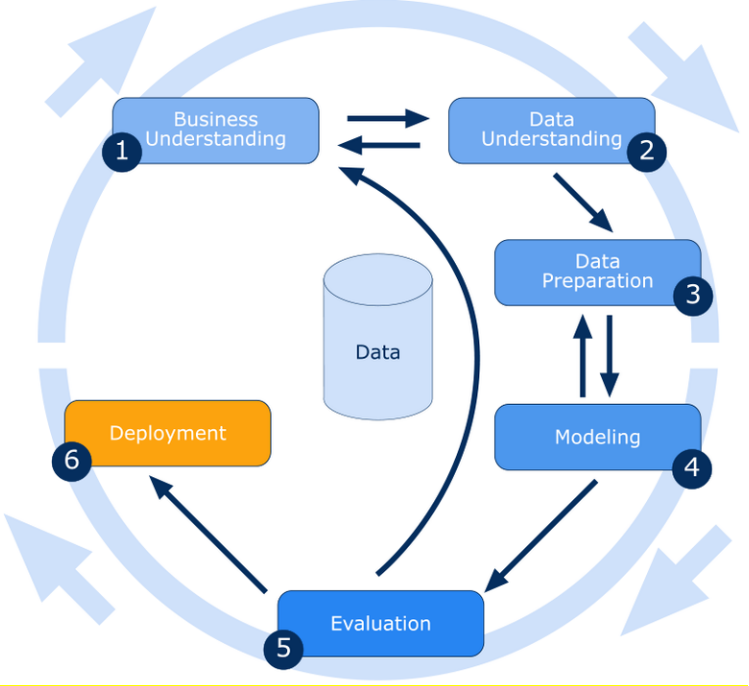
Which country has top sales?

Which are the top 5 profit-making product types on a yearly basis

How is the product price varying with sales - Is there any increase in sales with the decrease in price at a day level

What is the average delivery time across the counties

METHODOLOGY TO BE FOLLOWED



• Business Understanding:

It's all about understanding the overview, the aspects of business activities & the necessary problems which the business is facing.

• Data understanding:

It involves study of data, shape, datatypes, number of rows and columns, type of columns and categories them into numerical and categorical data.

• Data preparation:

This involves Preprocessing of Data

- Access the data
- Ingest (or fetch) the data
- Cleanse the data
- Format the data
- Combine the data
- And finally Analyze the data

- **Modeling**

Based on the observation of Descriptive & Inferential Statistic &

recognizing the right model.

- **Evaluation**

Uses some metric or combination of metrics to "measure" objective

performance of model. Test the model against previously unseen data.

- **Deployment**

Applying the Data to the mode

REFERENCES

The references can be blogs, articles or even social media news relevant to explain the importance of the projects.

https://www.researchgate.net/post/What_is_the_major_meaning_of_PCs_in_Principal_Component_Analysis

https://onlinelibrary.wiley.com/doi/10.1111/poms.13717

https://www.researchgate.net/publication/353923375_The_Impact_of_Covid-19_Lockdown_on_General_Trade_Grocery_Stores_Kirana_Stores_and_Independent_Self_Service_Stores_in_India

https://www.census.gov/library/stories/2022/04/ecommerce-sales-surged-during-pandemic.html

https://www.analyticsvidhya.com/blog/2022/02/exploratory-data-analysis-in-python/#h2_10

https://towardsdatascience.com/top-machine-learning-algorithms-for-classification-2197870ff501

https://www.analyticsvidhya.com/blog/2021/08/conceptual-understanding-of-logistic-regression-for-data-science-beginners/

https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm

Notes For Project Team

Original owner of data	Kaggle
Data set information	Global_Superstore2
Any past relevant articles using the dataset	NA
Reference	
Link to web page	https://www.kaggle.com/datasets/apoorvaappz/global-super-store-dataset?select=Global_Superstore2.xlsx

****