Final Project Report Template

1. Introduction
    1.1. Project overviews
    1.2. Objectives
2. Project Initialization and Planning Phase
    2.1. Define Problem Statement
    2.2. Project Proposal (Proposed Solution)
    2.3. Initial Project Planning
3. Data Collection and Preprocessing Phase
    3.1. Data Collection Plan and Raw Data Sources Identified
    3.2. Data Quality Report
    3.3. Data Exploration and Preprocessing
4. Model Development Phase
    4.1. Feature Selection Report
    4.2. Model Selection Report
    4.3. Initial Model Training Code, Model Validation and Evaluation Report
5. Model Optimization and Tuning Phase
    5.1. Hyperparameter Tuning Documentation
    5.2. Performance Metrics Comparison Report
    5.3. Final Model Selection Justification
6. Results
    6.1. Output Screenshots
7. Advantages & Disadvantages
8. Conclusion
9. Future Scope
10. Appendix
    10.1. Source Code
    10.2. GitHub & Project Demo Link

Data Collection and Preprocessing Phase

Date	8 July 2024
Team ID	SWTID1720369851
Project Title	Ecommerce Shipping Prediction
Maximum Marks	2 Marks

1. INTRODUCTION:

1.1 Project Overview:

Project Name: Ecommerce Shipping Prediction Using Machine Learning

Overview: The Ecommerce Shipping Prediction project aims to enhance the customer experience by providing accurate and reliable delivery estimates for online purchases. Leveraging advanced machine learning algorithms, the system predicts shipping times by analyzing historical delivery data and real-time factors such as traffic, weather, and distance. The solution integrates seamlessly with popular ecommerce platforms to automatically retrieve order details, offering real-time updates and notifications to customers about their delivery status. This project will help ecommerce businesses improve transparency, customer satisfaction, and operational efficiency.

1.2 Objectives:

1. **Accurate Delivery Estimates:**
    - Develop machine learning models that can predict delivery times with high accuracy, considering factors like distance, traffic, weather, and historical delivery data.
2. **Real-Time Updates:**
    - Implement a system that provides customers with real-time updates on their order status, including any changes or delays in the estimated delivery time.
3. **Seamless Integration:**
    - Create a solution that integrates smoothly with major ecommerce platforms (e.g., Shopify, Magento) to automate the retrieval of order details and delivery predictions.
4. **Model Optimization and Training:**

- Continuously train and optimize machine learning models to improve prediction accuracy over time, utilizing both historical and real-time data.

5. **Scalability:**
    - Ensure the system can handle high volumes of orders, providing quick and accurate delivery estimates even during peak shopping periods.
6. **User Experience:**
    - Design user-friendly interfaces for both customers and internal users, allowing easy access to delivery estimates, real-time tracking, and performance monitoring.
7. **Security and Compliance:**
    - Implement robust security measures to protect customer data and ensure compliance with data protection regulations (e.g., GDPR, CCPA).
8. **Performance Monitoring:**
    - Set up tools to monitor the system's performance and accuracy continuously, capturing customer feedback to identify and address areas for improvement.
9. **Customization and Flexibility:**
    - Allow customization of delivery options based on customer preferences and provide flexibility to adapt to various ecommerce platforms and shipping providers.

Project Initialization and Planning Phase

Date	9 July 2024
Team ID	SWTID1720369851
Project Name	Ecommerce Shipping Prediction
Maximum Marks	3 Marks

Define Problem Statements (Customer Problem Statement Template):

Customers often face uncertainty and frustration due to inaccurate or delayed delivery estimates. These issues can stem from various factors such as traffic conditions, weather, and logistical challenges. Additionally, the lack of real-time updates exacerbates the problem, leaving customers in the dark about their order status. There is a pressing need for a system that provides precise delivery time predictions and keeps customers informed with real-time updates.

Customer Problem Statement Template

I am	I'm trying to	But	Because	Which makes me feel
Customer (Employee)	Get the Delivery Date of my order	I am not able to get the accurate delivery estimate	The system does not consider the factors such as traffic and weather conditions	Frustrated
Customer (Student)	Get the current status of my order	I am not able to do so	The system does not provide real time updates on the status	Worried
Customer (Staff)	Retrieve the information of my order from the ecommerce stores	I am not able to do so	The system is not integrated with Ecommerce Platforms	Confused

Problem Statement (PS)	I am (Customer)	I'm trying to	But	Because	Which makes me feel
No Accurate Delivery Estimates	Customer (Employee)	Get the Delivery Details of my order	I am not able to get the accurate delivery estimate	The system does not consider the factors such as traffic and weather conditions	Frustrated
No Real-Time Updates	Customer (Student)	Get the current status of my order	I am not able to do so	The system does not provide real time updates on the status	Worried
Not Intergrated with Ecommerce Platforms	Customer (Staff)	Retrieve the information of my order from the Ecommerce stores	I am not able to do so	The system is not integrated with Ecommerce Platforms	Confused

Project Initialization and Planning Phase

Date	9 July 2024
Team ID	SWTID1720369851
Project Title	Ecommerce Shipping Prediction
Maximum Marks	3 Marks

Project Proposal (Proposed Solution) template

In today's fast-paced ecommerce landscape, customers expect not only a wide variety of products but also reliable and timely delivery of their purchases. Delivery delays or inaccurate delivery estimates can lead to customer dissatisfaction, negatively impacting their overall shopping experience and brand loyalty.

Project Overview	
Objective	To create an application to predict the shipping details of any order.
Scope	To get accurate and real time estimation of the delivery date.
Problem Statement	
Description	Customers often face uncertainty and frustration due to inaccurate or delayed delivery estimates. These issues can stem from various factors such as traffic conditions, weather, and logistical challenges. Additionally, the lack of real-time updates exacerbates the problem, leaving customers in the dark about their order status.
Impact	Enhanced Trust and Satisfaction, Improved Transparency and Streamlined Shopping Experience.
Proposed Solution	
Approach	I am going to use machine learning to solve this type of problem, starting with the data collection and preprocessing of the data, I will be using different models to get the highest accuracy possible.
Key Features	After executing this I will be able to get Accurate Delivery Estimates, Real-Time Updates and Seamless Integration.

Resource Requirements

Resource Type	Description	Specification/Allocation
Hardware		
Computing Resources	CPU/GPU specifications, number of cores	Apple M1, 2 x NVIDIA V100 GPUs, AMD Ryzen 5/7, intel core i5/i7
Memory	RAM specifications	8 GB/ 16GB
Storage	Disk space for data, models, and logs	256 GB/ 512GB/1 TB SSD
Software		
Frameworks	Python frameworks	Flask
Libraries	Additional libraries	Numpy, pandas, scikit-learn, matplotlib, seaborn, scipy
Development Environment	IDE, version control	Jupyter Notebook, Git
Data		
Data	Source, size, format	Kaggle dataset, (10999, 12), csv

Initial Project Planning Template

Date	9 July 2024
Team ID	SWTID1720369851
Project Name	Ecommerce Shipping Prediction
Maximum Marks	4 Marks

Product Backlog, Sprint Schedule, and Estimation (4 Marks)

Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority	Team Members	Sprint Start Date	Sprint End Date (Planned)
Sprint-1	Data Collection & Preparation	ESP-1	Collect the Dataset	2	High	Ayush, Pranshu	9 July	9 July
Sprint-1	Data Collection & Preparation	ESP-2	Data Preparation	1	High	Ayush, Pranshu	9 July	9 July
Sprint-2	Exploratory Data Analysis	ESP-3	Descriptive Statistics	2	Medium	Nikilesh, Parth	9 July	10 July
Sprint-2	Exploratory Data Analysis	ESP-4	Visual Analysis	2	Medium	Parth, Pranshu	10 July	10 July

Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority	Team Members	Sprint Start Date	Sprint End Date (Planned)
Sprint-3	Model Building	ESP-5	Training The Model in Multiple Algorithms	3	High	Nikilesh	10 July	11 July
Sprint-3	Model Building	ESP-6	Testing The Model	4	High	Parth	10 July	11 July
Sprint-4	Performance Testing & Hyperparameter Tuning	ESP-7	Testing Model with Multiple Evaluation Metrics	5	High	Ayush	10 July	11 July
Sprint-5	Model Deployment	ESP-8	Save The Best Model	3	High	Pranshu	11 July	12 July
Sprint-5	Model Deployment	ESP-9	Integrate With Web Framework	6	High	Nikilesh, Ayush	12 July	12 July

Data Collection and Preprocessing Phase

Date	9 July 2024
Team ID	SWTID1720369851
Project Title	Ecommerce Shipping Prediction
Maximum Marks	2 Marks

Data Collection Plan & Raw Data Sources Identification Template

Elevate your data strategy with the Data Collection plan and the Raw Data Sources report, ensuring meticulous data curation and integrity for informed decision-making in every analysis and decision-making endeavor.

Data Collection Plan Template

Section	Description
Project Overview	Customers often face uncertainty and frustration due to inaccurate or delayed delivery estimates. These issues can stem from various factors such as traffic conditions, weather, and logistical challenges. Additionally, the lack of real-time updates exacerbates the problem, leaving customers in the dark about their order status.
Data Collection Plan	The Dataset is collected from the Kaggle official website.
Raw Data Sources Identified	The name of the Dataset is 'E-commerce Shipping Data' with the file name of 'Train.csv'. It has 10999 rows and 12 columns.

Raw Data Sources Template

Source Name	Description	Location/URL	Format	Size	Access Permissions
E-commerce Shipping Data	This Dataset consists an international e-commerce company who wants to discover key insights from their customer database.	https://www.kaggle.com/datasets/prachi13/customer-analytics?select=Train.csv	CSV	440.46 KB	Public

Data Collection and Preprocessing Phase

Date	15 March 2024
Team ID	SWTID1720369851
Project Title	Ecommerce Shipping Prediction
Maximum Marks	2 Marks

Data Quality Report Template

The Data Quality Report Template will summarize data quality issues from the selected source, including severity levels and resolution plans. It will aid in systematically identifying and rectifying data discrepancies.

Data Source	Data Quality Issue	Severity	Resolution Plan
E-commerce Shipping Data	In this Dataset we have 12 columns out of which the columns named as 'Mode of shipment' and 'Weight in gms' are independent variables.	Moderate	We can remove these columns using the drop function.

Data Collection and Preprocessing Phase

Date	10 July 2024
Team ID	SWTID1720369851
Project Title	Ecommerce Shipping Prediction
Maximum Marks	6 Marks

Data Exploration and Preprocessing Template

Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

Section	Description
Data Overview	

```
data.shape
```

```
(10999, 12)
```

```
data.describe()
```

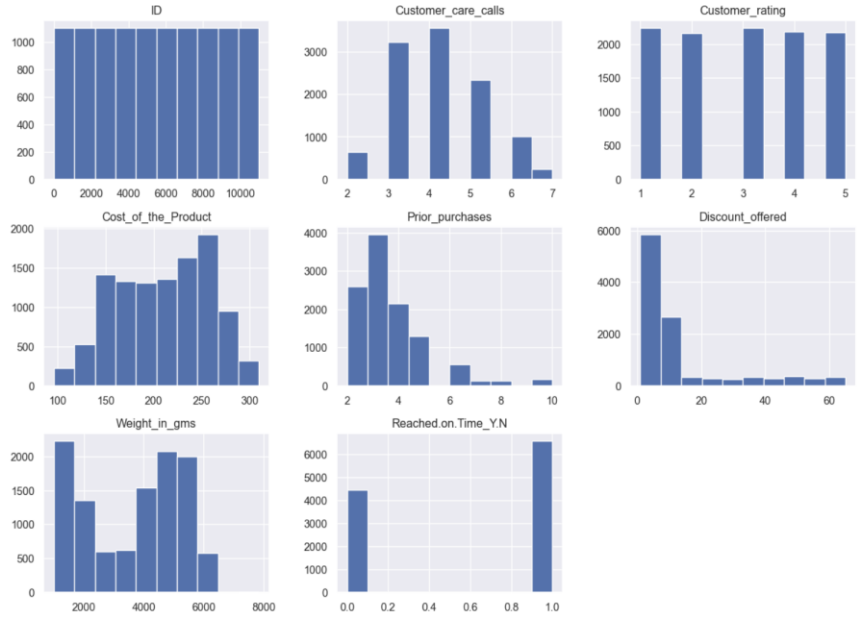
	ID	Customer_care_calls	Customer_rating	Cost_of_the_Product	Prior_purchases	Discount_offered	Weight_in_gms	Reached.on.Time_Y.N
count	10999.00000	10999.000000	10999.000000	10999.000000	10999.000000	10999.000000	10999.000000	10999.000000
mean	5500.00000	4.054459	2.990545	210.196836	3.567597	13.373216	3634.016729	0.596691
std	3175.28214	1.141490	1.413603	48.063272	1.522860	16.205527	1635.377251	0.490584
min	1.00000	2.000000	1.000000	96.000000	2.000000	1.000000	1001.000000	0.000000
25%	2750.50000	3.000000	2.000000	169.000000	3.000000	4.000000	1839.500000	0.000000
50%	5500.00000	4.000000	3.000000	214.000000	3.000000	7.000000	4149.000000	1.000000
75%	8249.50000	5.000000	4.000000	251.000000	4.000000	10.000000	5050.000000	1.000000
max	10999.00000	7.000000	5.000000	310.000000	10.000000	65.000000	7846.000000	1.000000

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10999 entries, 0 to 10998
Data columns (total 12 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   ID                   10999 non-null  int64
 1   Warehouse_block      10999 non-null  object
 2   Mode_of_Shipment     10999 non-null  object
 3   Customer_care_calls  10999 non-null  int64
 4   Customer_rating      10999 non-null  int64
 5   Cost_of_the_Product  10999 non-null  int64
 6   Prior_purchases      10999 non-null  int64
 7   Product_importance   10999 non-null  object
 8   Gender               10999 non-null  object
 9   Discount_offered     10999 non-null  int64
 10  Weight_in_gms        10999 non-null  int64
 11  Reached.on.Time_Y.N  10999 non-null  int64
dtypes: int64(8), object(4)
memory usage: 1.0+ MB
```

Univariate Analysis

```
data.select_dtypes(include=[np.number]).hist(bins=10, figsize=(15,10))
plt.show()
```



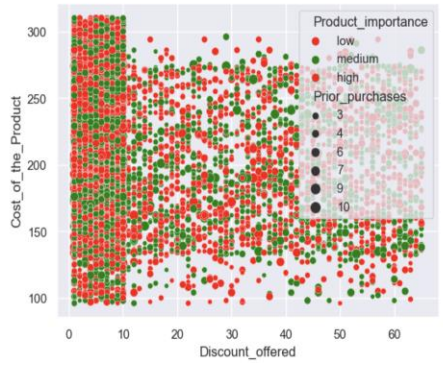
Bivariate Analysis

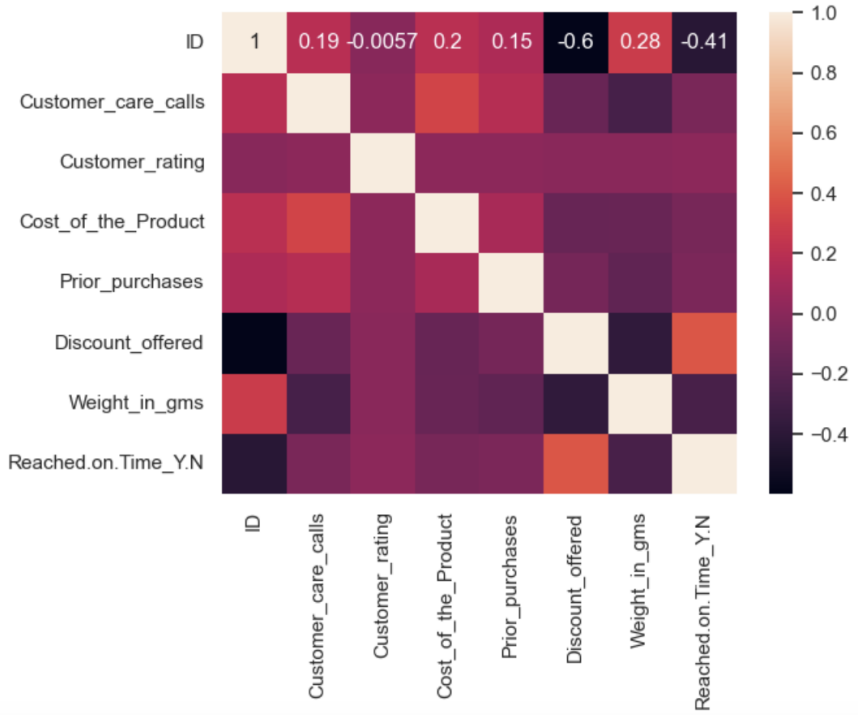
```
data.corr(numeric_only=True)
```

	ID	Customer_care_calls	Customer_rating	Cost_of_the_Product	Prior_purchases	Discount_offered	Weight_in_gms	Reached.on.Tin
ID	1.000000	0.188998	-0.005722	0.196791	0.145369	-0.598278	0.278312	-0.
Customer_care_calls	0.188998	1.000000	0.012209	0.323182	0.180771	-0.130750	-0.276615	-0.
Customer_rating	-0.005722	0.012209	1.000000	0.009270	0.013179	-0.003124	-0.001897	0.
Cost_of_the_Product	0.196791	0.323182	0.009270	1.000000	0.123676	-0.138312	-0.132604	-0.
Prior_purchases	0.145369	0.180771	0.013179	0.123676	1.000000	-0.082769	-0.168213	-0.
Discount_offered	-0.598278	-0.130750	-0.003124	-0.138312	-0.082769	1.000000	-0.376067	0.
Weight_in_gms	0.278312	-0.276615	-0.001897	-0.132604	-0.168213	-0.376067	1.000000	-0.2
Reached.on.Time_Y.N	-0.411822	-0.067126	0.013119	-0.073587	-0.055515	0.397108	-0.268793	1.0

```
# Scatterplot
sns.scatterplot(x='Discount_offered',y='Cost_of_the_Product',data=data, hue='Product_importance', size='Prior_purchases', palette=['red','gree
```

```
<Axes: xlabel='Discount_offered', ylabel='Cost_of_the_Product'>
```

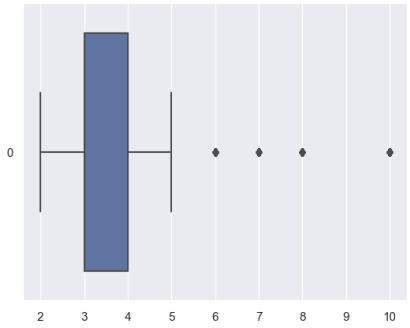


Multivariate Analysis	```# Heat Map
	hm=data.corr(numeric_only=True)
	sns.heatmap(hm, annot=True)```
	```<Axes: >```
	 

Outliers and Anomalies

```
sns.boxplot(data['Prior_purchases'], orient='h')
```

```
<Axes: >
```




```
Q1=data['Prior_purchases'].quantile(0.25)
Q3=data['Prior_purchases'].quantile(0.75)
IQR=Q3-Q1
lower_limit=Q1-1.5*IQR
upper_limit=Q3-1.5*IQR
data[(data.Prior_purchases<lower_limit)|(data.Prior_purchases>upper_limit)]
data=data[(data.Prior_purchases>lower_limit)&(data.Prior_purchases<upper_limit)]
data
```

	ID	Warehouse_block	Mode_of_Shipment	Customer_care_calls	Customer_rating	Cost_of_the_Product	Prior_purchases	Product_importance	Gender
1	2	F	Flight	4	5	216	2	low	M
7	8	F	Flight	4	1	233	2	low	F
10	11	C	Flight	3	4	189	2	medium	M
16	17	C	Flight	3	4	143	2	medium	F
20	21	A	Ship	3	3	161	2	medium	F
...	...	...	...	...	...	...	...	...	...
10892	10893	A	Ship	4	5	167	2	low	F
10897	10898	F	Ship	4	3	242	2	low	M
10898	10899	A	Ship	4	1	222	2	high	M
10901	10902	F	Ship	4	4	201	2	medium	M
10904	10905	A	Ship	4	2	218	2	low	M

2599 rows × 12 columns

Data Preprocessing Code Screenshots

Loading Data

```
# Load the data

data = pd.read_csv("Train.csv")
```

Handling Missing Data	`data.isnull().sum()`
	
	ID 0
	Warehouse_block 0
	Mode_of_Shipment 0
	Customer_care_calls 0
	Customer_rating 0
	Cost_of_the_Product 0
	Prior_purchases 0
	Product_importance 0
	Gender 0
	Discount_offered 0
	Weight_in_gms 0
	Reached.on.Time_Y.N 0
	dtype: int64
	
	`data.duplicated().sum()`
	
	0

Data Transformation	**Encoding the categorical variables**
	
	`le = LabelEncoder()`
	
	`def Label_Enc(col):`
	`  Categorical_col[col] = le.fit_transform(Categorical_col[col])`
	
	`for i in ['Warehouse_block', 'Mode_of_Shipment', 'Product_importance', 'Gender']:`
	`  Label_Enc(i)`
	
	`Categorical_col.head()`
	
	Warehouse_block Mode_of_Shipment Product_importance Gender
	0 3 0 1 0
	1 4 0 1 1
	2 0 0 1 1
	3 1 0 2 1
	4 2 0 2 0

Feature Engineering	`Numerical_col.drop(columns = ["ID"],axis = 1,inplace = True)`
	`Numerical_col.head()`
	
	Customer_care_calls Customer_rating Cost_of_the_Product Prior_purchases Discount_offered Weight_in_gms Reached.on.Time_Y.N
	0 4 2 177 3 44 1233 1
	1 4 5 216 2 59 3088 1
	2 2 2 183 4 48 3374 1
	3 3 3 176 4 10 1177 1
	4 2 2 184 3 46 2484 1

Save Processed Data	`# Save to a CSV file` `data.to_csv('Train(new).csv', index=False)`

Model Development Phase Template

Date	15 March 2024
Team ID	SWTID1720369851
Project Title	Ecommerce Shipping Prediction
Maximum Marks	5 Marks

Feature Selection Report Template

In the forthcoming update, each feature will be accompanied by a brief description. Users will indicate whether it's selected or not, providing reasoning for their decision. This process will streamline decision-making and enhance transparency in feature selection.

Feature	Description	Selected (Yes/No)	Reasoning
ID	ID Number of Customers	No	ID is not the deciding feature whether the product is delivered on time or not.
Warehouse block	The Company have big Warehouse which is divided in to block such as A,B,C,D,E.	No	It cannot be decided by looking into the warehouse if the product is going to be delivered on time.
Mode of shipment	The Company Ships the products in multiple way such as Ship, Flight and Road.	Yes	Mode of shipment is a deciding factor as some mode can be fast while others can be slow.
Customer care calls	The number of calls made from enquiry for enquiry of the shipment.	Yes	If a customer is making more calls to customer care, they may expect their product to be delivered on time.

Customer rating	The company has rated from every customer. 1 is the lowest (Worst), 5 is the highest (Best).	Yes	If a customer has a higher rating, they are expected to receive their product before the arrival time.
Cost of the product	Cost of the Product in US Dollars.	Yes	Expensive products are expected to be delivered earlier.
Prior purchases	The Number of Prior Purchase.	Yes	If a customer is a frequent user of the platform, they are going to get their products delivered on time.
Product importance	The company has categorized the product in the various parameter such as low, medium, high.	Yes	If a product is having more importance to the user, they will be getting the product at the earliest.
Gender	Male and Female.	No	This cannot be the deciding feature, as the product delivery is independent of Gender.
Discount offered	Discount offered on that specific product.	Yes	If a product is having more discount so it is expected to be delivered later than usual products.
Weight in gms	It is the weight in grams.	Yes	The weight of the product plays a major role in the delivery time of the product as it can affect the speed of the delivery rate.
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.	Yes	It is used for future reference and to make the prediction whether that delivery location is expected to get their product delivered earlier.

Model Development Phase Template

Date	11 July 2024
Team ID	SWTID1720369851
Project Title	Ecommerce Shipping Prediction
Maximum Marks	6 Marks

Model Selection Report

In the forthcoming Model Selection Report, various models will be outlined, detailing their descriptions, hyperparameters, and performance metrics, including Accuracy or F1 Score. This comprehensive report will provide insights into the chosen models and their effectiveness.

Model Selection Report:

Model	Description	Hyperparameters	Performance Metric (e.g., Accuracy, F1 Score)
Logistic Regression	It's a type of regression analysis that estimates the probability that a given input point belongs to a certain class.	penalty = ['l2','l1','elasticnet'] C = [0.0001, 0.001, 0.002] solver = ['newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga'] hyperparameters = dict(penalty=penalty, C=C , solver=solver)	Confusion Matrix : [[531 377] [437 855]] Accuracy_Score: 63.0 % F1 Score: 67.75 Precision Score: 69.399 Recall Score : 66.176 AUC Score : 62.328

Decision Tree Classifier	It uses a tree-like model of decisions and their possible consequences, including outcomes, resource costs, and utility.	max_depth = [int(x) for x in np.linspace(1, 110, num = 30)] min_samples_split = [2, 5, 10, 100] min_samples_leaf = [1, 2, 4, 10, 20, 50] max_features = ['auto', 'sqrt'] criterion = ['gini', 'entropy'] splitter = ['best', 'random'] hyperparameters = dict(max_depth=max_depth, min_samples_split=min_samples_split, min_samples_leaf=min_samples_leaf, max_features=max_features, criterion=criterion, splitter=splitter)	Confusion Matrix : [[508 400] [392 900]] Accuracy_Score: 64.0 % F1 Score: 69.444 Precision Score: 69.231 Recall Score : 69.659 AUC Score : 62.803

Random Forest	Random Forest is an ensemble learning method that constructs multiple decision trees during training and merges their outputs to improve classification accuracy and control overfitting.	**NA**	Confusion Matrix : [[611 297] [446 846]] Accuracy_Score: 66.227 % F1 Score: 69.487 Precision Score: 74.016 Recall Score : 65.48 AUC Score : 66.385
SVM	Support Vector Machine (SVM) is a powerful supervised learning algorithm used for classification and regression tasks.	**NA**	Confusion Matrix : [[566 342] [402 890]] Accuracy_Score: 66.182 % F1 Score: 70.523 Precision Score: 72.24 Recall Score : 68.885 AUC Score : 65.61
XGBoost	XGBoost (Extreme Gradient Boosting) is a highly efficient and scalable machine learning algorithm for classification and regression tasks.	**NA**	Confusion Matrix : [[567 341] [442 850]] Accuracy_Score: 64.409 % F1 Score: 68.466 Precision Score: 71.369 Recall Score : 65.789 AUC Score : 64.117

KNN	K-Nearest Neighbors (KNN) classifies a data point based on the majority class of its k-nearest neighbors in the feature space.	n_neighbors = list(range(2,30)) p=[1,2] algorithm = ['auto', 'ball_tree', 'kd_tree', 'brute'] hyperparameters = dict(n_neighbors=n_neighbors, p=p, algorithm=algorithm)	Confusion Matrix : [[556 352] [409 883]] Accuracy_Score: 65.409 % F1 Score: 69.885 Precision Score: 71.498 Recall Score : 68.344 AUC Score : 64.789

Model Development Phase Template

Date	11 July 2024
Team ID	SWTID1720369851
Project Title	Ecommerce Shipping Prediction
Maximum Marks	4 Marks

Initial Model Training Code, Model Validation and Evaluation Report

The initial model training code will be showcased in the future through a screenshot. The model validation and evaluation report will include classification reports, accuracy, and confusion matrices for multiple models, presented through respective screenshots.

Initial Model Training Code:

Paste the screenshot of the model training code

Model Validation and Evaluation Report:

Model	Classification Report	Accuracy	Confusion Matrix
Logistic Regression	F1 Score: 67.75 Precision Score: 69.399 Recall Score : 66.176 AUC Score : 62.328	Accuracy_Score: 63.0 %	Confusion Matrix : [[531 377] [437 855]]
Decision Tree Classifier	F1 Score: 69.444 Precision Score: 69.231 Recall Score : 69.659 AUC Score : 62.803	Accuracy_Score: 64.0 %	Confusion Matrix : [[508 400] [392 900]]
Random Forest Classifier	F1 Score: 69.487 Precision Score: 74.016 Recall Score : 65.48 AUC Score : 66.385	Accuracy_Score: 66.227 %	Confusion Matrix : [[611 297] [446 846]]

SVM	F1 Score: 70.523 Precision Score: 72.24 Recall Score : 68.885 AUC Score : 65.61	Accuracy_Score: 66.182 %	Confusion Matrix : [[566 342] [402 890]]
XGBoost	F1 Score: 68.466 Precision Score: 71.369 Recall Score : 65.789 AUC Score : 64.117	Accuracy_Score: 64.409 %	Confusion Matrix : [[567 341] [442 850]]
KNN	F1 Score: 69.885 Precision Score: 71.498 Recall Score : 68.344 AUC Score : 64.789	Accuracy_Score: 65.409 %	Confusion Matrix : [[556 352] [409 883]]

Model Optimization and Tuning Phase Template

Date	11 July 2024
Team ID	SWTID1720369851
Project Title	Ecommerce Shipping Prediction
Maximum Marks	10 Marks

Model Optimization and Tuning Phase

The Model Optimization and Tuning Phase involves refining machine learning models for peak performance. It includes optimized model code, fine-tuning hyperparameters, comparing performance metrics, and justifying the final model selection for enhanced predictive accuracy and efficiency.

Hyperparameter Tuning Documentation (6 Marks):

Model	Tuned Hyperparameters	Optimal Values
Logistic Regression	penalty, C, solver	penalty = ['l2','l1','elasticnet'] C = [0.0001, 0.001, 0.002] solver = ['newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga']
Decision Tree Classifier	max_depth, min_samples_split, min_samples_leaf, max_features, Criterion, splitter	max_depth = [int(x) for x in np.linspace(1, 110, num = 30)]

		min_samples_split = [2, 5, 10, 100] min_samples_leaf = [1, 2, 4, 10, 20, 50] max_features = ['auto', 'sqrt'] criterion = ['gini', 'entropy'] splitter = ['best', 'random']
Random Forest	param_name, param_range, parameter_range, cv	param_name = 'n_estimators', param_range = parameter_range, cv = 5
SVM	NA	NA
XGBoost	NA	NA
KNN	n_neighbors, p, algorithm	n_neighbors = list(range(2,30)) p=[1,2] algorithm = ['auto', 'ball_tree', 'kd_tree', 'brute']

Performance Metrics Comparison Report (2 Marks):

Model	Baseline Metric	Optimized Metric
Logistic Regression	Confusion Matrix : [[531 377] [437 855]] Accuracy_Score: 63.0 % F1 Score: 67.75 Precision Score: 69.399 Recall Score : 66.176 AUC Score : 62.328	Confusion Matrix : [[512 396] [369 923]] Accuracy_Score: 65.227 % F1 Score: 70.701 Precision Score: 69.977 Recall Score : 71.44 AUC Score : 63.914
Decision Tree Classifier	Confusion Matrix : [[508 400] [392 900]] Accuracy_Score: 64.0 % F1 Score: 69.444 Precision Score: 69.231 Recall Score : 69.659 AUC Score : 62.803	Confusion Matrix : [[559 349] [428 864]] Accuracy_Score: 64.682 % F1 Score: 68.982 Precision Score: 71.228 Recall Score : 66.873 AUC Score : 64.218
Random Forest	Confusion Matrix : [[611 297] [446 846]] Accuracy_Score: 66.227 % F1 Score: 69.487 Precision Score: 74.016 Recall Score : 65.48 AUC Score : 66.385	NA

SVM	Confusion Matrix : [[566 342] [402 890]] Accuracy_Score: 66.182 % F1 Score: 70.523 Precision Score: 72.24 Recall Score : 68.885 AUC Score : 65.61	NA
XGBoost	Confusion Matrix : [[567 341] [442 850]] Accuracy_Score: 64.409 % F1 Score: 68.466 Precision Score: 71.369 Recall Score : 65.789 AUC Score : 64.117	NA
KNN	Confusion Matrix : [[556 352] [409 883]] Accuracy_Score: 65.409 % F1 Score: 69.885 Precision Score: 71.498 Recall Score : 68.344 AUC Score : 64.789	Confusion Matrix : [[686 222] [533 759]] Accuracy_Score: 65.682 % F1 Score: 66.784 Precision Score: 77.37 Recall Score : 58.746 AUC Score : 67.148

Final Model Selection Justification (2 Marks):

Final Model	Reasoning

Random Forest	Based on the results of each model algorithm, I decided to use the Random Forest model because this model has a fairly balanced combination of AUC and Recall scores and the results are more interpretable. It is delivering high predictive accuracy by aggregating the results of multiple decision trees, reducing the likelihood of overfitting. It is suitable for large datasets due to its parallel processing capabilities. Balancing the bias and variance by averaging multiple decision trees, leading to a more generalized model. These attributes make Random Forest a reliable and effective choice.

Team ID: SWTID1720369851

Project Title: Ecommerce Shipping Prediction

6. RESULTS:

6.1 Output Screenshots:

```
data.describe()
```

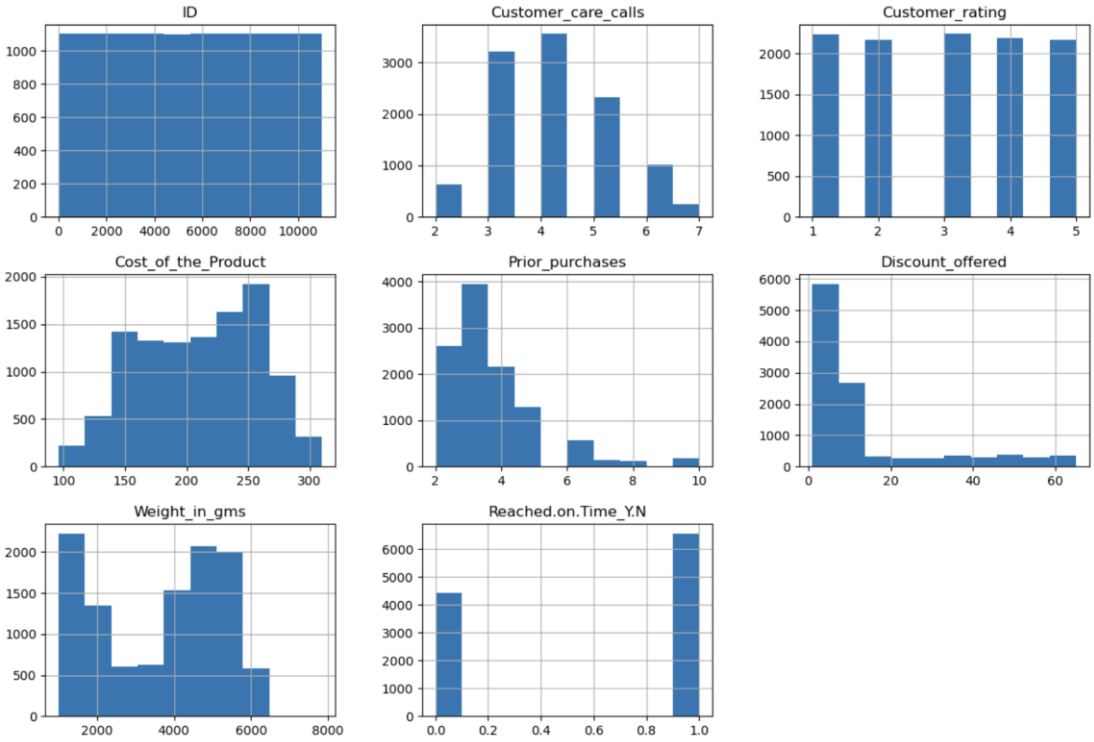
	ID	Customer_care_calls	Customer_rating	Cost_of_the_Product	Prior_purchases	Discount_offered	Weight_in_gms	Reached.on.Time_Y.N
count	10999.00000	10999.000000	10999.000000	10999.000000	10999.000000	10999.000000	10999.000000	10999.000000
mean	5500.00000	4.054459	2.990545	210.196836	3.567597	13.373216	3634.016729	0.596691
std	3175.28214	1.141490	1.413603	48.063272	1.522860	16.205527	1635.377251	0.490584
min	1.00000	2.000000	1.000000	96.000000	2.000000	1.000000	1001.000000	0.000000
25%	2750.50000	3.000000	2.000000	169.000000	3.000000	4.000000	1839.500000	0.000000
50%	5500.00000	4.000000	3.000000	214.000000	3.000000	7.000000	4149.000000	1.000000
75%	8249.50000	5.000000	4.000000	251.000000	4.000000	10.000000	5050.000000	1.000000
max	10999.00000	7.000000	5.000000	310.000000	10.000000	65.000000	7846.000000	1.000000

```
data.info()
```

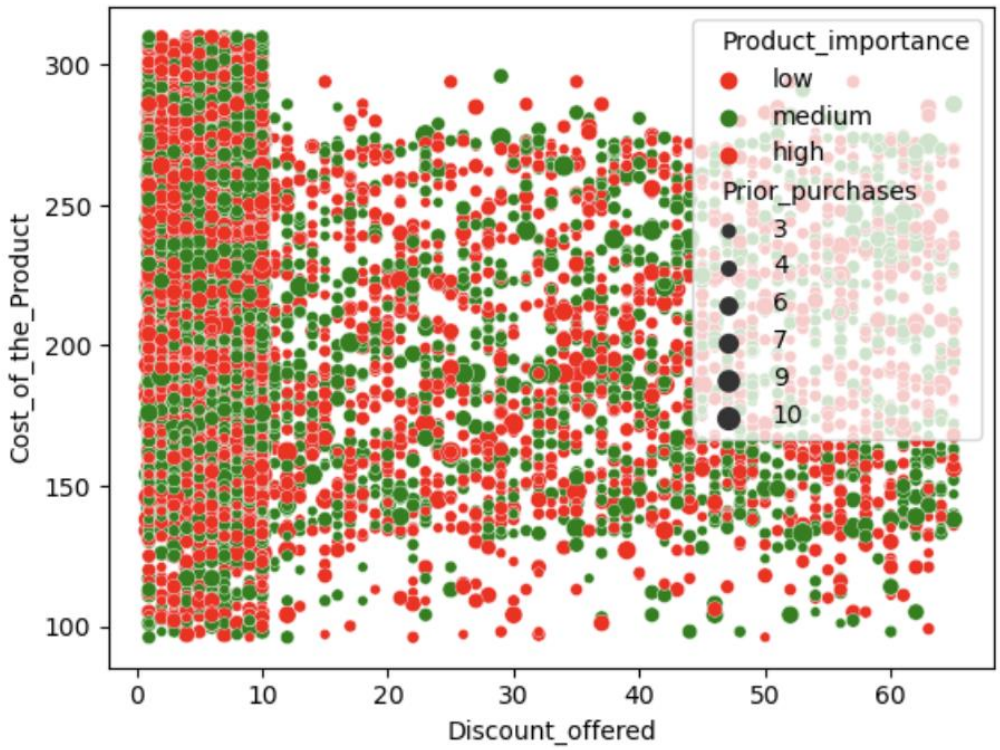
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10999 entries, 0 to 10998
Data columns (total 12 columns):
 #   Column               Non-Null Count  Dtype 
---  ------               --------------  ----- 
 0   ID                   10999 non-null  int64 
 1   Warehouse_block      10999 non-null  object
 2   Mode_of_Shipment     10999 non-null  object
 3   Customer_care_calls  10999 non-null  int64 
 4   Customer_rating      10999 non-null  int64 
 5   Cost_of_the_Product  10999 non-null  int64 
 6   Prior_purchases      10999 non-null  int64 
 7   Product_importance   10999 non-null  object
 8   Gender               10999 non-null  object
 9   Discount_offered     10999 non-null  int64 
 10  Weight_in_gms        10999 non-null  int64 
 11  Reached.on.Time_Y.N  10999 non-null  int64 
dtypes: int64(8), object(4)
memory usage: 1.0+ MB
```

```
data.isnull().sum()
```

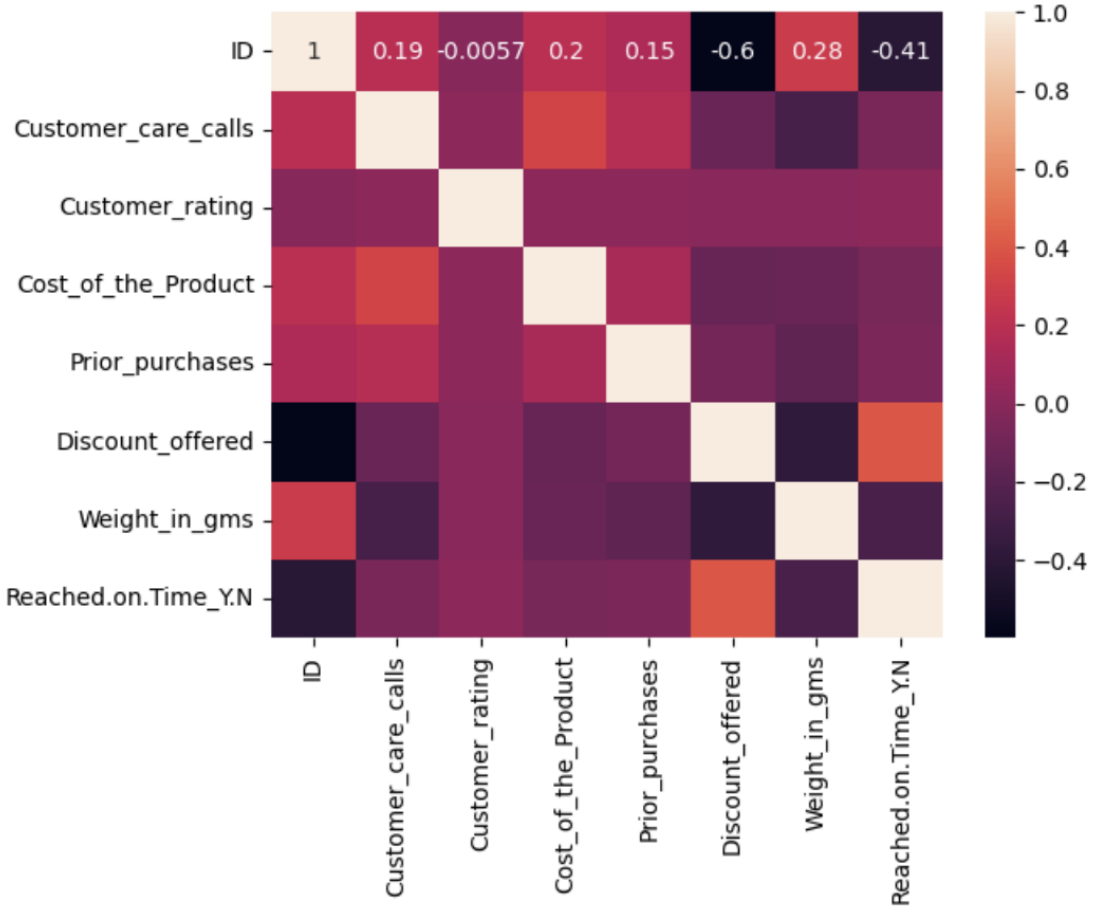
```
ID                     0
Warehouse_block        0
Mode_of_Shipment       0
Customer_care_calls    0
Customer_rating        0
Cost_of_the_Product    0
Prior_purchases        0
Product_importance     0
Gender                 0
Discount_offered       0
Weight_in_gms          0
Reached.on.Time_Y.N    0
dtype: int64
```




```
<Axes: xlabel='Discount_offered', ylabel='Cost_of_the_Product'>
```



<Axes: >

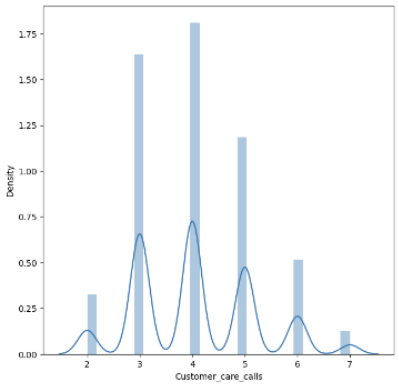


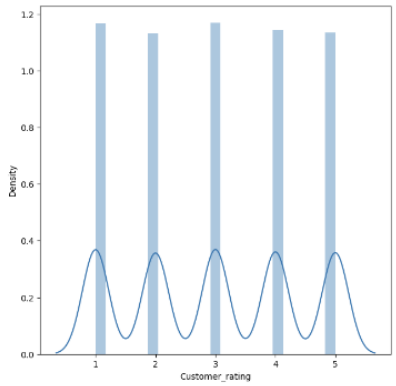

```
Categorical_col = data[cat_col]
Categorical_col.head()
```

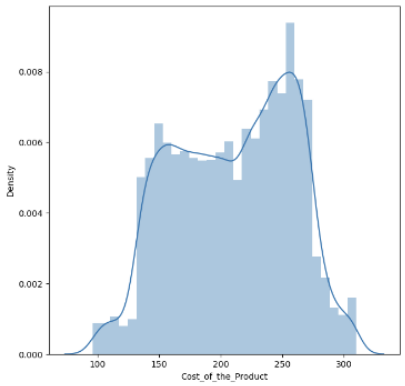
	Warehouse_block	Mode_of_Shipment	Product_importance	Gender
0	D	Flight	low	F
1	F	Flight	low	M
2	A	Flight	low	M
3	B	Flight	medium	M
4	C	Flight	medium	F

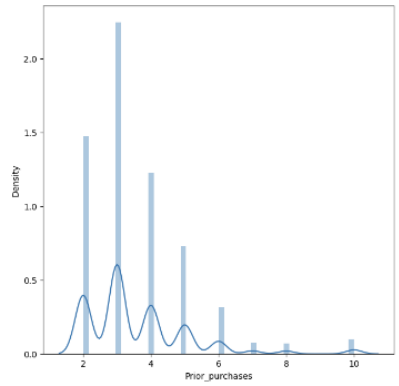
```
Categorical_col.head()
```

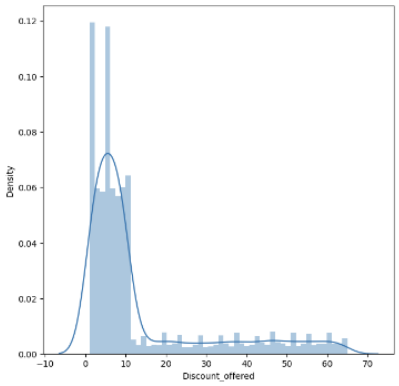
	Warehouse_block	Mode_of_Shipment	Product_importance	Gender
0	3	0	1	0
1	4	0	1	1
2	0	0	1	1
3	1	0	2	1
4	2	0	2	0

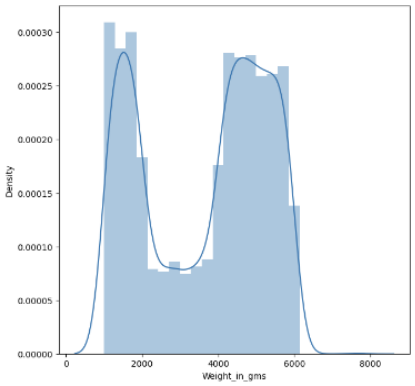

Density
Customer_care_calls

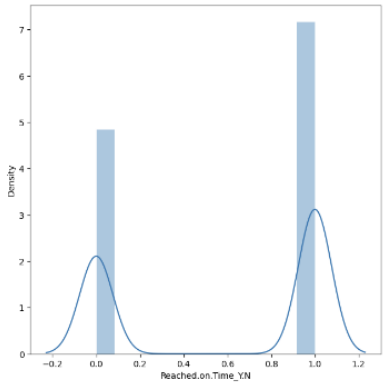

Density
Customer_rating


Density
Cost_of_the_Product

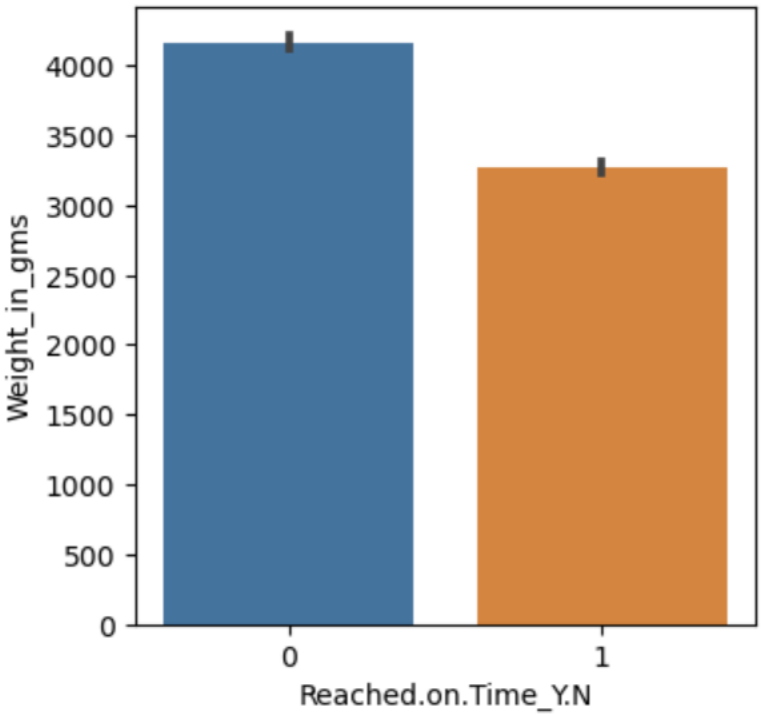

Density
Prior_purchases

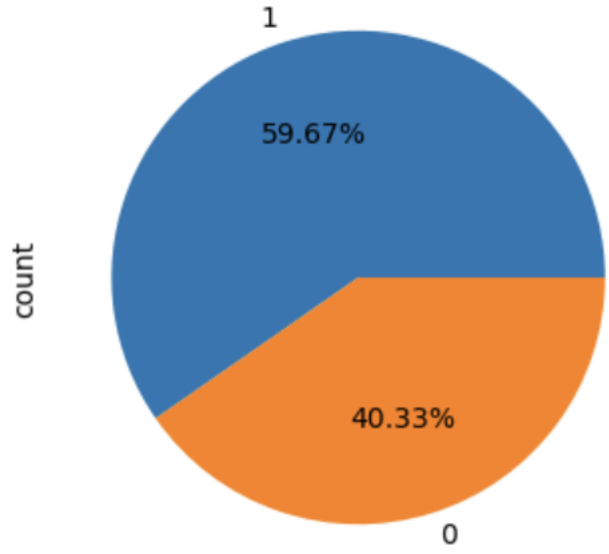

Density
Discount_offered

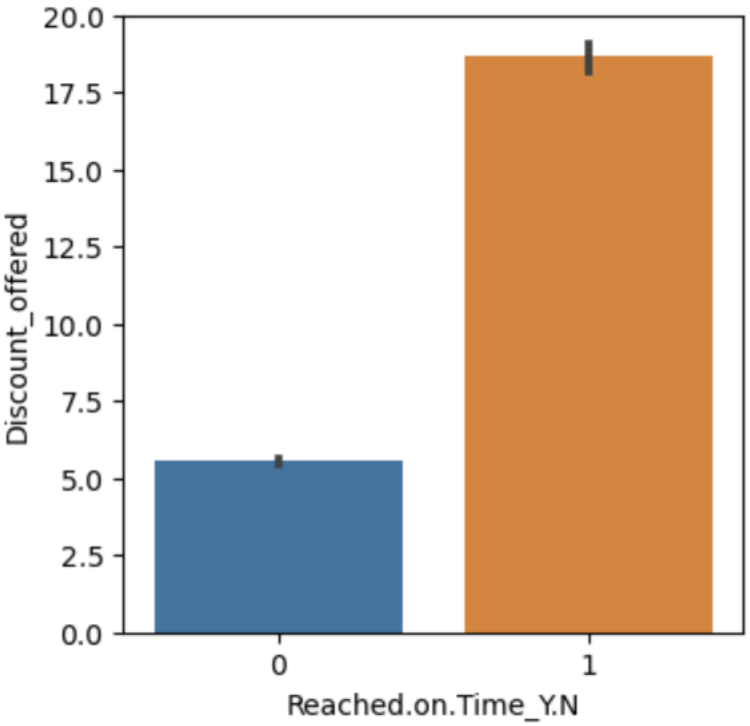

Density
Weight_in_gms

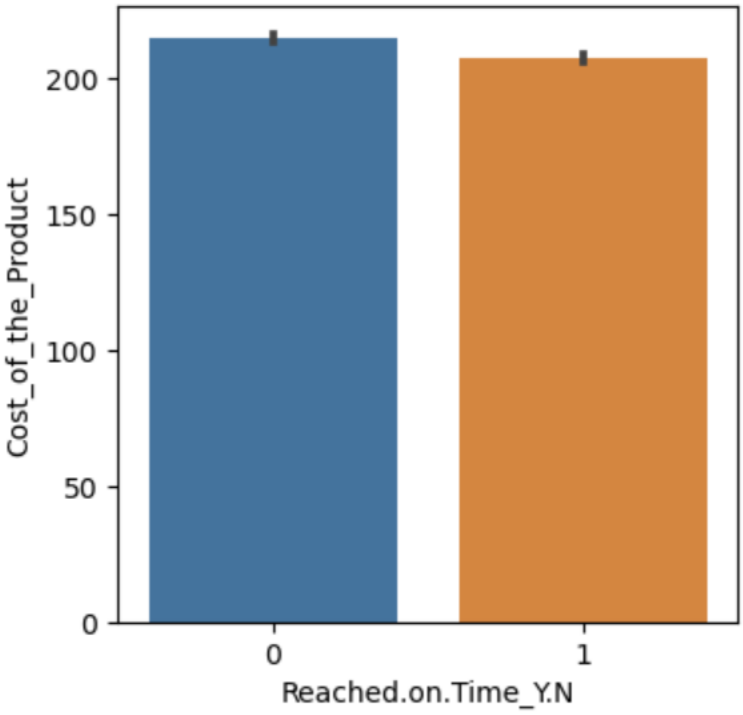

Density
Reached.on.Time_Y.N

[ ]


4000
3500
3000
2500
2000
1500
1000
500
0
Weight_in_gms
0
1
Reached.on.Time_Y.N


1
59.67%
40.33%
0
count


20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
Discount_offered
0
1
Reached.on.Time_Y.N


200
150
100
50
0
Cost_of_the_Product
0
1
Reached.on.Time_Y.N

Predicting whether the product is getting delivered on time or not.

Select Warehouse | Select Mode of Shipment | Select Product Importance | Select Gender

Product ID | Customer Care Calls | Customer Rating | Product Cost | Prior Purchase | Discount Offered | Product Weight (gms) | Predict

Team ID: SWTID1720369851

Project Title: Ecommerce Shipping Prediction

E-commerce shipping prediction refers to the use of data analysis and machine learning to estimate shipping times and optimize logistics. Here are some advantages and disadvantages of this technology:

7.Advantages

1. Improved Customer Experience

- ***Accurate Delivery Estimates:*** Providing precise delivery times enhances customer satisfaction and trust.
- ***Proactive Communication:*** Customers can be informed of delays, improving transparency.

2. Operational Efficiency

- ***Optimized Routing:*** Algorithms can find the best routes, reducing shipping times and costs.
- ***Inventory Management:*** Predictive analytics can help manage stock levels, reducing overstock and stockouts.

3. Cost Reduction

- ***Reduced Shipping Costs:*** Optimizing routes and delivery schedules can lower transportation expenses.
- ***Labor Efficiency:*** Predicting busy periods helps in better workforce planning.

4. Competitive Advantage

- ***Customer Retention:*** Reliable delivery times can enhance customer loyalty.
- ***Market Differentiation:*** Advanced shipping prediction capabilities can distinguish a business from competitors.

5. Environmental Benefits

- ***Reduced Carbon Footprint:*** Optimized routes and fewer delivery attempts contribute to lower emissions.
- ***Efficient Resource Use:*** Better prediction leads to more efficient use of vehicles and other resources.

Disadvantages

1. Implementation Costs

- ***High Initial Investment:*** Developing and integrating predictive systems can be expensive.
- ***Ongoing Maintenance:*** Continuous updates and maintenance of the system add to operational costs.

2. Data Dependency

- ***Data Quality:*** Accurate predictions require high-quality, extensive data.
- ***Data Privacy:*** Handling sensitive customer and operational data involves compliance with privacy regulations.

3. Complexity

- ***Technical Expertise:*** Implementing and managing predictive systems requires specialized knowledge.
- ***System Integration:*** Integrating predictive models with existing logistics and inventory systems can be complex.

4. Risk of Inaccuracy

- ***Model Limitations:*** Predictive models can sometimes fail to account for unexpected variables, leading to inaccurate estimates.
- ***External Factors:*** Weather, traffic conditions, and other unforeseen events can affect delivery times, regardless of predictions.

5. Dependence on Technology

- ***System Downtime:*** Reliance on technology means that system failures can disrupt operations.
- ***Cybersecurity Risks:*** Increased digitalization brings heightened risk of cyber-attacks.

E-commerce shipping prediction, when implemented effectively, can offer significant benefits in terms of efficiency, cost savings, and customer satisfaction. However, the challenges and risks associated with its implementation must be carefully managed to fully realize its potential.

Team ID: SWTID1720369851

Project Title: Ecommerce Shipping Prediction

8.Conclusion

E-commerce shipping prediction presents a transformative opportunity for online businesses by leveraging data analytics and machine learning to enhance delivery accuracy and operational efficiency. While the advantages are substantial—ranging from improved customer satisfaction and reduced shipping costs to environmental benefits—the challenges and risks must be navigated carefully. High initial investments, data quality concerns, and the complexity of implementation can pose significant hurdles. However, with strategic planning, continuous improvement, and a focus on integrating robust systems, businesses can effectively mitigate these challenges. Ultimately, the successful adoption of shipping prediction technologies can provide a competitive edge, fostering customer loyalty and operational excellence in the dynamic e-commerce landscape.

Team ID: SWTID1720369851

Project Title: Ecommerce Shipping Prediction

9.Future Scope

The future scope of ecommerce shipping prediction is a broad and evolving topic. Here's a concise overview of some key areas:

1. AI and machine learning advancements:

AI and ML algorithms will become more sophisticated in predicting shipping times, optimizing routes, and forecasting demand. This will lead to more accurate delivery estimates and efficient resource allocation.

2. Real-time tracking and predictive analytics:

Advanced tracking systems will provide more granular, real-time updates on package locations. Predictive analytics will help identify potential delays or issues before they occur, allowing for proactive problem-solving.

3. Automation and robotics in warehouses and logistics:

Increased use of robots and automated systems in warehouses will speed up order fulfillment and reduce errors. This includes automated sorting systems, self-driving forklifts, and robotic arms for picking and packing.

4. Sustainable and eco-friendly shipping solutions:

As environmental concerns grow, there will be a greater focus on reducing the carbon footprint of shipping. This may include electric delivery vehicles, optimized packaging to reduce waste, and carbon-neutral shipping options.

5. Last-mile delivery innovations:

New solutions for the most expensive part of shipping - the last mile - will emerge. This could include strategies like micro-fulfillment centers, crowdsourced delivery, and smart lockers for package pickup.

6. Integration of IoT devices for better tracking and condition monitoring:

Internet of Things (IoT) sensors will provide real-time data on package conditions (temperature, humidity, shock) and location, ensuring quality control and enabling more precise tracking.

7. Personalized delivery experiences:

Customers will have more control over their deliveries, with options to change delivery times or locations on the fly, and receive personalized updates based on their preferences.

8. Blockchain for supply chain transparency:

Blockchain technology will be used to create tamper-proof records of a product's journey, enhancing transparency and traceability in the supply chain.

9. Drone and autonomous vehicle deliveries:

As regulations evolve, drone deliveries and self-driving vehicles will become more common, especially for last-mile delivery in urban areas or hard-to-reach locations.

10. Predictive demand forecasting:

Advanced analytics will improve the accuracy of demand forecasting, allowing retailers and logistics companies to better prepare for fluctuations in shipping volume.

Team ID: SWTID1720369851

Project Title: Ecommerce Shipping Prediction

Video Link

https://drive.google.com/drive/folders/1OLm3e3tD7TPb9cFztZv5LWU8gOWM78V-?usp=sharing