



**ANALYZING MACHINE LEARNING TECHNIQUES
FOR PREDICTIVE ANALYTICS TO ENHANCE
DELIVERY OPTIMIZATION IN ECOMMERCE
LOGISTICS**

SUBMITTED BY – SALMA FIROZE

REGISTRATION NO – HNDDS232F-020

HIGHER NATIONAL DIPLOMA IN DATA SCIENCE

NATIONAL INSTITUTE OF BUSINESS MANAGEMENT(NIBM) – NIC

COLOMBO, SRI LANKA

DATE OF SUBMISSION – 28TH OF OCTOBER 2024

“A Project submitted for the partial fulfilment of the requirements of Higher National
Diploma in Data Science (Full Time) Programme”

DECLARATION

I hereby declare that the work presented in this project report was carried out independently by myself and have cited the work of others and given due reference diligently.

.....

Salma Firoze

.....

Date

I certify that the above student carried out her project under my supervision and guidance.

.....

.....

Date

ACKNOWLEDGEMENT

I take this opportunity to convey my heartfelt gratitude to all those who helped me successfully complete this project.

I am highly indebted to our course coordinator Mrs. W.M.S.G.D..C Wanigasekara for providing me with valuable advice, guidance and support throughout the entire project. I am also obliged to my research supervisor Mr. Ashan for providing me with useful insights, guidance and imparting constructive feedback that helped me improve my work.

I would like to acknowledge my family and friends for their unwavering support and constant encouragement.

Thank you for all your invaluable contributions to this project.

EXECUTIVE SUMMARY

The project analyzes an ecommerce shipping dataset to identify the best model which predicts if the delivery was carried out on time or not. The aim is to improve logistics operations with the use of technological knowledge on machine learning.

This project is aimed at determining how predictive analytics could contribute to an increase in the accuracy of delivery times in e-commerce logistics, because timely shipments are among the major factors in customer satisfaction and thus give a competitive advantage. Comparison of different machine learning models is done: Random Forest, Decision Tree, SVM, Logistic Regression, and XGBoost. The performance test of each in predicting on-time delivery is performed, with hyperparameter tuning for each one. Random Forest was leading in performance, though XGBoost was rather close, which outlines the value of the latter for delay prediction.

Correctly estimating the exact delivery time enables a company to proactively create route optimization, resource allocation, and communication with customers to minimize delays and ensure the efficacy of operations and customer trust. The limit on computation prohibited further consideration of deeper learning models, but it laid a platform for future studies on the use of advanced predictive techniques. This work has once more underlined the power of data-driven methods in logistics, providing useful insights that would allow e-

commerce companies to develop appropriate processes and to ensure reliable delivery service.

Contents

CHAPTER 1	1
INTRODUCTION.....	1
1.1 Background	1
1.2 Research Problem	1
1.3 Research Questions	2
1.4 Objectives of the Project.....	2
1.5 Scope of the Research	2
1.6 Justification of the Research	3
1.7 Limitations.....	3
CHAPTER 2	3
LITERATURE REVIEW	3
2.1 Introduction to the Research Theme.....	3
2.3 Findings by other researchers.....	5
2.4 The research gap	6
2.5 Table for variables, their definitions, and sources.....	6
2.6 Chapter Conclusion	7
CHAPTER 3	7
METHODOLOGY	7
3.1 Introduction.....	7

3.2 Population, Sample and Sampling Technique	8
3.3Types of Data to be Collected and Data Source	8
3.4Data Collection Tools and Plan	9
3.5 Conceptual Framework.....	9
3.6 Hypothesis.....	9
3.7 Operationalization Table	10
3.8 Methods of Data Analysis	11
CHAPTER 4	12
DATA ANALYSIS	12
4.1 Data Preprocessing	12
4.2 Descriptive Statistics and Data Analysis	14
4.3 Findings and Interpretation	18
CHAPTER 5	21
DISCUSSION AND RECOMMENDATIONS.....	21
5.1 Discussion	21
5.2 Recommendations	22
5.3 Conclusion	22
References.....	27

CHAPTER 1

INTRODUCTION

1.1 Background

An important factor of ecommerce is efficient, on time delivery; if customers are going to be satisfied then it needs to provide good quality service according to their expectations and deliver the required goods on time. This can be a challenge to many local and global company serving numerous clients nationwide or worldwide. This study focuses on analyzing different predictive models which can contribute to enhance the efficiency of ecommerce logistics which would bring around a competitive advantage in the market by leveraging data driven strategies.

1.2 Research Problem

Delivery operations often encounter various challenges and driving factors which may result in delays. Despite the importance of these factors, there is a lack of comprehensive analysis on how they collectively impact delivery times which in turn can reduce satisfaction of customers which can be looked into otherwise.

1.3 Research Questions

- I. Can a comparative analysis of hyperparameter tuning strategies across multiple machine learning models enhance delivery time prediction accuracy in e-commerce logistics?
- II. How does the application of XGBoost impact the accuracy of delivery time prediction models compared to traditional machine learning algorithms?

1.4 Objectives of the Project

- I. Analyze the performance of different machine learning models on delivery time prediction problems with respect to various hyperparameter tuning strategies.
- II. Compare the performance of the XGBoost model with that of traditional algorithms for high accuracy in estimating delivery time.

1.5 Scope of the Research

The core idea of this research is to predict if shipment reached on time, with a focus on understanding the parameters that contribute to models which perform better. The research focuses on ecommerce shipping logistics, analyzing data across various warehouses, modes of shipment, product importance, prior purchases and more.

1.6 Justification of the Research

As e-commerce continues to expand, optimizing delivery logistics becomes increasingly important. Past studies have focused on using Machine Learning models to predict delivery time on various datasets. This study aims to improve similar models further by testing on other hyperparameters thus leading to better operational efficiency and customer satisfaction when implemented.

1.7 Limitations

The study relies on a secondary dataset so it does not account several factors which may also influence delivery times and the data can be imbalanced. Machine Learning models can also require higher computational resources which makes it hard to test on advanced hyperparameters and large datasets.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction to the Research Theme

Delivery efficiency is a key component of the e-commerce industry as it can significantly influence customer satisfaction and the overall success of several ecommerce companies.

The theme of this research focuses on accurately predicting if shipments reached customers on time which otherwise can be optimized by taking timely measures.

2.2 Theoretical explanation of keywords in the topic

Key Words	Definition/Explanation	Source
Delivery Optimization	Delivery optimization involves the optimization of delivery through betterment in logistical operations, cost-effectiveness, and on-schedule use of advanced algorithms by real-time analysis of data.	Logistics and Supply Chain Management literature
Predictive Analytics	Predictive analytics, also called statistical modeling, involves the creation of a model to predict future events or behaviors using historical trends. It uses techniques such as historical data, statistical algorithms, and machine learning.	Data Science for Business
Logistics	Logistics refers to the overall process that entails managing all of resources' procurement, storage and movement until they reach the customer.	Investopedia Logistics: What it Means and How Businesses Use It

2.3 Findings by other researchers

A dissertation was carried out to predict delivery times of products using an ecommerce shipping dataset with Logistic regression, random forest, support vector machine and decision tree models. Here the random forest algorithm performed best also determining that speed, value, category and location of service highly contribute to the prediction. (Collins & Uchenna, 2024)

Another study which aimed to predict cargo delivery time used Categorical Boosting, Decision Tree, Extreme Learning Machine, Light Gradient Boosting and Support Vector Machine Models out of which SVM outperformed the other models giving a precise day count to customers on when the product will reach them. (Selim & Ceren, 2024)

A distinctive study suggests companies to take a comprehensive approach to implement machine learning algorithms by developing a robust data infrastructure and investing in technology with data analysis experts to overcome delivery challenges and improve customer experience. (Verbytskyi, 2023)

According to a study which used Support Vector Machine to predict delivery time the average Mean Absolute Error obtained with the look back approach was lower than that obtained without the look back approach showing that the look back approach makes the model better. (Oguz & Ece, 2022)

An interesting study which used data from JD Logistics designed a Predict Then Optimize Couriers Allocation framework which resulted in the method outperforming the baseline in task delivery rate and on time delivery rate. (Kaiwen & Li, 2023)

2.4 The research gap

Although various machine learning models have been studied, little research has systematically compared the effects of different hyperparameter tuning strategies on model accuracy especially in the field of logistics. This research focuses on improving accuracy of machine learning models which predicts if ecommerce shipments reached on time or not.

2.5 Table for variables, their definitions, and sources

Variable	Description	Source
ID	ID of customers	Kaggle
Warehouse block	The warehouses of the company	Kaggle
Mode of shipment	The way the product is shipped	Kaggle
Customer care calls	No of calls on shipment inquiry	Kaggle
Customer rating	The rating for the company by customer	Kaggle
Cost of product	Cost of product in US Dollars	Kaggle
Prior purchases	No of times customer has purchased before	Kaggle
Product importance	How important the product is	Kaggle
Gender	Male or Female customer	Kaggle
Discount offered	Discount on specific product	Kaggle
Weight in gms	The weight of product	Kaggle
Reached on time	If the product reaches on time or not	Kaggle

2.6 Chapter Conclusion

Delivery efficiency is a critical component of the e-commerce industry, significantly influencing customer satisfaction and company efficiency. This research focuses on testing different hyperparameters on machine learning models to accurately predict delivered on time status.

CHAPTER 3

METHODOLOGY

3.1 Introduction

- I. As the motive of this research is to test different hyperparameter tuning strategies across multiple machine learning models to enhance delivery time prediction accuracy in e-commerce logistics , I hope to clearly describe the process which will be followed to achieve the expected outcome. In this chapter, all particulars on the types of data to be analyzed, the data collection tool, the conceptual framework for doing the research, the hypothesis to be tested, an operationalization table with additional details about the variables and the method which will be used for data analysis will be discussed.

3.2 Population, Sample and Sampling Technique

The population of this dataset consists of Ecommerce Shipping records. It contains various logistical, product and customer related factors that influence delivery times and customer rating.

The dataset includes records related to 11000 deliveries of products from various warehouses completed within a diverse time period. The complete dataset will be used for analysis after removing null values therefore any sampling technique will not be used.

3.3 Types of Data to be Collected and Data Source

The dataset for this research was obtained from an existing database on the website www.kaggle.com which includes detailed information of Ecommerce shipping logistics.

The type of data collected are as follows:

Numerical data: Customer care calls, Prior Purchases, Weight

Categorical data: Warehouse block, Mode of shipment, Customer rating, Product Importance, Gender, Reached on time

Currency: Cost of product, Discount offered

3.4 Data Collection Tools and Plan

The dataset was obtained from Kaggle as mentioned above. It can be found through [E-Commerce Shipping Data \(kaggle.com\)](https://www.kaggle.com/datasets/ecommerce-shipping-data). As the dataset contains that had been previously collected and processed, there is no need for further collection tools or plans.

3.5 Conceptual Framework

- I. Analyze the performance of different machine learning models on delivery time prediction problems w.r.t. various hyperparameter tuning strategies.

Target Variable: Arrived on Time
- II. Compare the performance of the XGBoost model with that of traditional algorithms for high accuracy in estimating delivery time.

3.6 Hypothesis

- I. Analyze the performance of different machine learning models on delivery time prediction problems w.r.t. various hyperparameter tuning strategies.

Null Hypothesis: There is no significant difference in the accuracies of delivery time predictions among different hyperparameter tuning strategies applied to machine learning models.

Alternate Hypothesis: There is a significant difference in the delivery

time prediction accuracy amongst various hyperparameter tuning strategies applied on the machine learning models.

- II. Assess the effectiveness of XGBoost compared to traditional algorithms in accurately predicting delivery times

Null Hypothesis: XGBoost application will lead to no significant difference in delivery time prediction accuracy compared to traditional machine learning algorithms.

Different Hypothesis: The application of XGBoost significantly improved the accuracy in the prediction of delivery time from that obtained through the application of traditional machine learning algorithms.

3.7 Operationalization Table

Variable	Indicators	Measure
ID	ID of customers	By unique numbers
Warehouse block	The warehouses of the company	(Categorical) A B C D E
Mode of shipment	The way the product is shipped	(Categorical) Flight Ship Road
Customer care calls	No of calls on shipment inquiry	By number

Customer rating	The rating for the company by customer	By scores
Cost of product	Cost of product in US Dollars	By currency US dollar
Prior purchases	No of times customer has purchased before	By number
Product importance	How important the product is	(Categorical) High, Medium, Low
Gender	Male or Female customer	Male or Female
Discount offered	Discount on specific product	By dollar value
Weight in gms	The weight of product	In grams
Reached on time	If the product reaches on time or not	By 1 – Yes or 0 – No

3.8 Methods of Data Analysis

The research objectives will be addressed using various statistical methods, machine learning approaches and visualizations in Python.

- I. Analyze how different hyperparameter tuning strategies affect the performance of various machine learning models in predicting delivery times

Models: Random Forest, Decision Trees, Logistic regression, Support Vector Machine

- II. Assess the effectiveness of XGBoost compared to traditional algorithms in accurately predicting delivery times,

Model: XGBoost

The hypotheses will be tested based on the results from these analyses, leading to conclusions that can help optimize ecommerce shipping logistics.

CHAPTER 4

DATA ANALYSIS

4.1 Data Preprocessing

The first step carried out was exploring the dataset. It contains 10998 records and 11 columns out of which 4 are categorical and the rest numerical. There are no null values which is an added advantage.

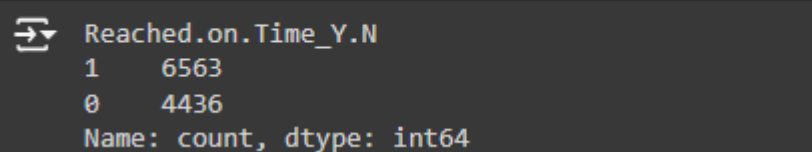
```
[ ] print(data.info())  
  
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 10999 entries, 0 to 10998  
Data columns (total 12 columns):  
#   Column                                Non-Null Count  Dtype  
---  ---                                -  
0   ID                                    10999 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  
None
```

The column named ID was then dropped as it does not contribute to any predictions but is present just as an unique identifier.

```
[ ] data.drop(['ID'], axis=1, inplace=True)
```

The dataset was then checked for any imbalances. There were 6563 records of which shipments were delivered on time and 4436 of that which was not delivered on time. This shows that there is a slight imbalance in data. It was adjusted using undersampling technique.

```
[ ] print(df['Reached.on.Time_Y.N'].value_counts())
```



Reached.on.Time_Y.N	
1	6563
0	4436

Name: count, dtype: int64

The categorical variables were converted to a single integer using label encoder to make it suitable to be processed by machine learning algorithms.

```
[ ] label_encoder = LabelEncoder()

data['Warehouse_block'] = label_encoder.fit_transform(data['Warehouse_block'])
data['Mode_of_Shipment'] = label_encoder.fit_transform(data['Mode_of_Shipment'])
data['Product_importance'] = label_encoder.fit_transform(data['Product_importance'])
data['Gender'] = label_encoder.fit_transform(data['Gender'])
```

Standard scalar is used to help the models work better and faster by putting all features on the same scale.

```
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
```

4.2 Descriptive Statistics and Data Analysis

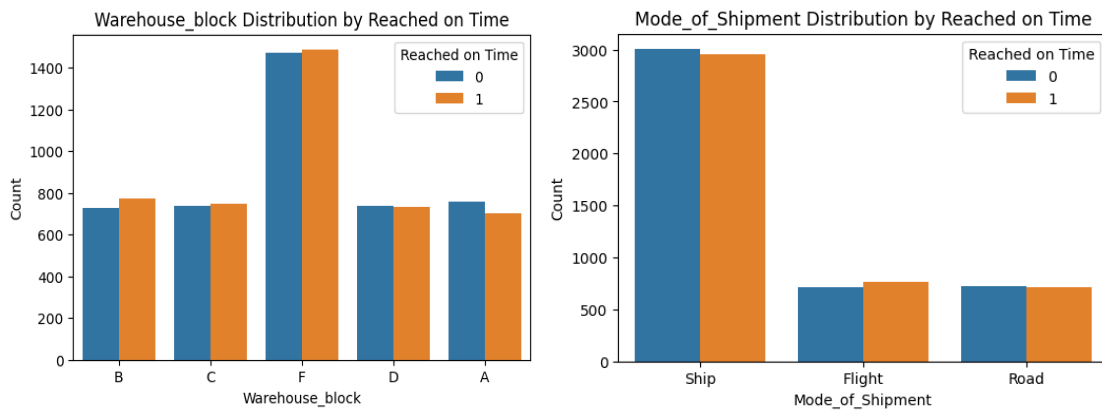
The summary of the variables which are included in the research computed using colab notebook is presented in this section to offer significant understanding on its range, spread and central tendency.

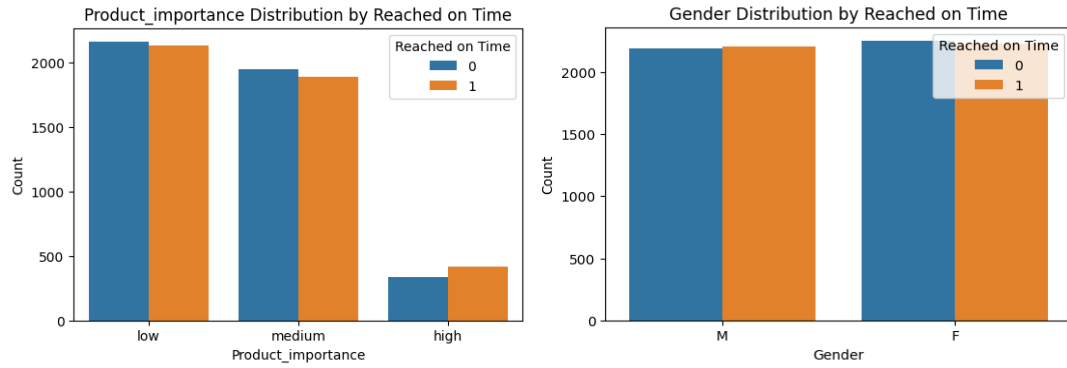
```
[ ] print(data.describe())
```

	ID	Customer_care_calls	Customer_rating	Cost_of_the_Product	\
count	10999.00000	10999.000000	10999.000000	10999.000000	
mean	5500.00000	4.054459	2.990545	210.196836	
std	3175.28214	1.141490	1.413603	48.063272	
min	1.00000	2.000000	1.000000	96.000000	
25%	2750.50000	3.000000	2.000000	169.000000	
50%	5500.00000	4.000000	3.000000	214.000000	
75%	8249.50000	5.000000	4.000000	251.000000	
max	10999.00000	7.000000	5.000000	310.000000	

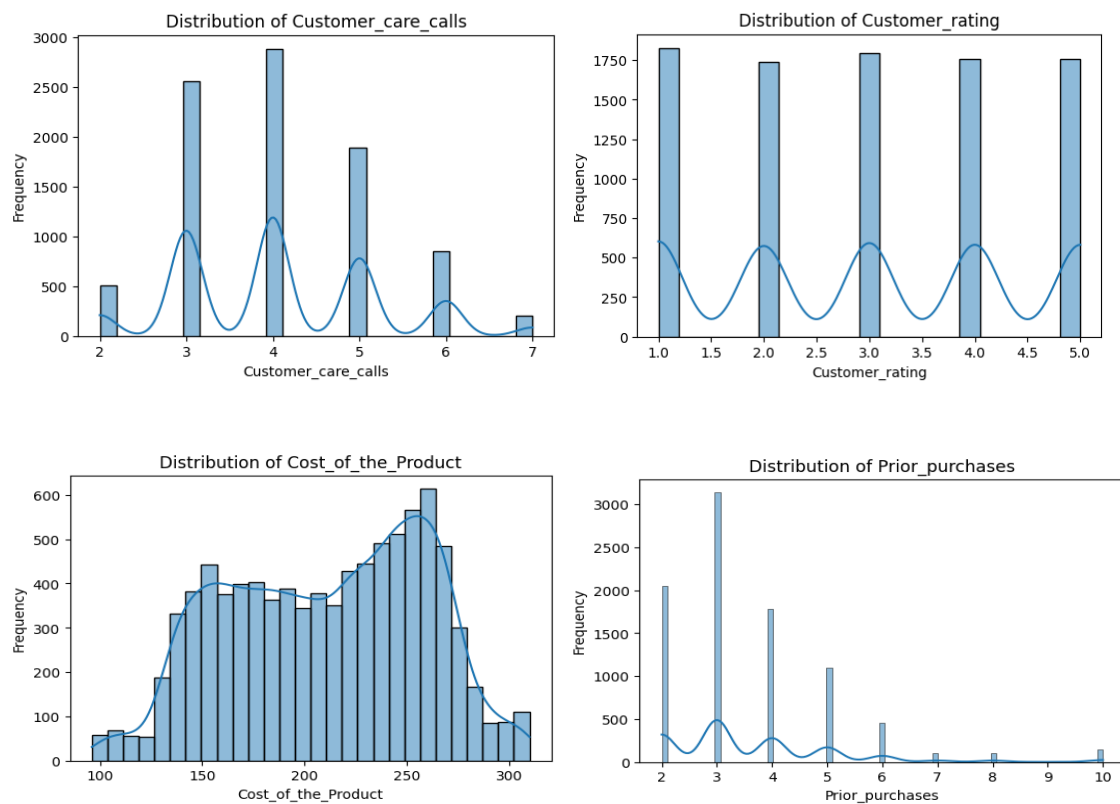
	Prior_purchases	Discount_offered	Weight_in_gms	Reached.on.Time_Y.N
count	10999.000000	10999.000000	10999.000000	10999.000000
mean	3.567597	13.373216	3634.016729	0.596691
std	1.522860	16.205527	1635.377251	0.490584
min	2.000000	1.000000	1001.000000	0.000000
25%	3.000000	4.000000	1839.500000	0.000000
50%	3.000000	7.000000	4149.000000	1.000000
75%	4.000000	10.000000	5050.000000	1.000000
max	10.000000	65.000000	7846.000000	1.000000

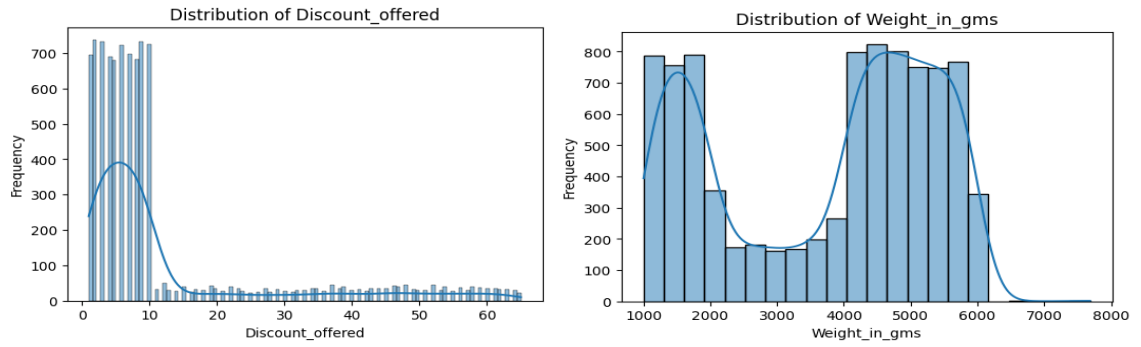
In addition to the descriptive statistics, the graphs given below will illustrate the relationships and trends among selected variables.





There is a slightly higher number of shipments reaching on time through flight when compared to roads and ships. Also shipments of higher importance have reached on time more in percentage than the others.



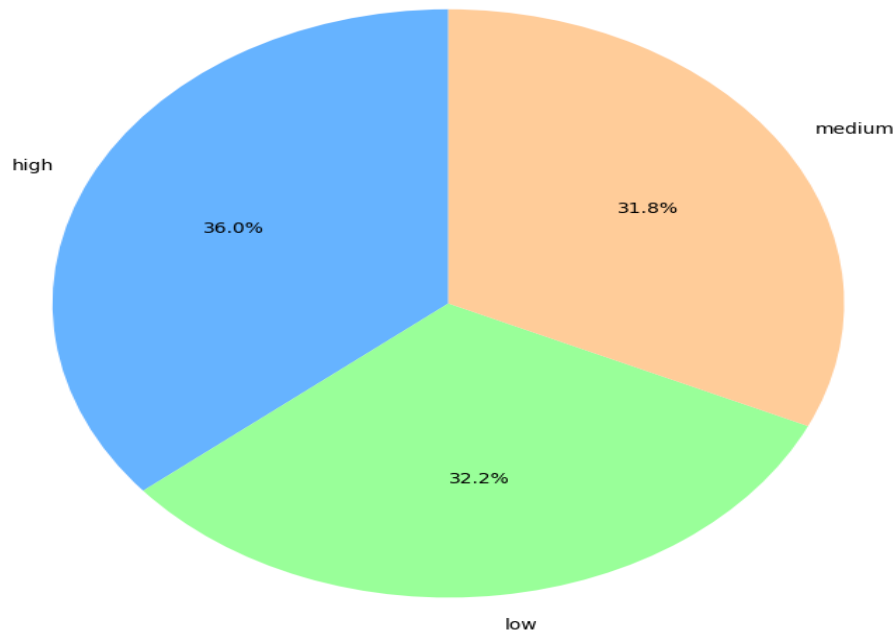


The distribution of all numerical variables are as given above.

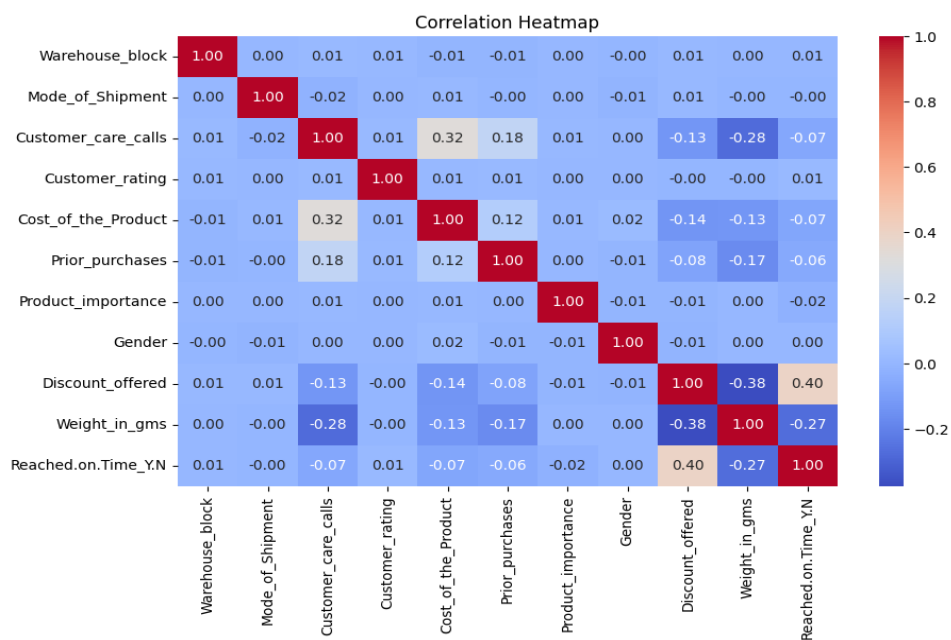


Percentage of on time deliveries by flight is comparatively higher than the other two modes. The warehouse A displays a lower percentage of on time deliveries by all modes.

Distribution of On-Time Deliveries by Product Importance



Products with higher importance have been delivered on time mostly compared to that of medium and low importance.



According to the correlation map above Discount offered has a higher correlation with reached on time, cost of the product with customer care calls.

4.3 Findings and Interpretation

The machine learning models that were run to achieve the set objectives are presented and discussed below.

- I. Analyze how different hyperparameter tuning strategies affect the performance of various machine learning models in predicting if the shipment reached on time.

Model	Hyperparameter Values					Best Parameters	ROC-AUC score	Accuracy
	n-estimators	max-depth	min-sample-split	min sample leaf	max features	cv		
Random Forest	50,100	5,10,20				5 100,5	0.728578	0.73
	250,260	10,15	3,5	1,2	None	5 260,10,5,2	0.728969	0.73
	300,350	10,15	3,5		1 sqrt	5 300,10,3,1	0.728961	0.73
Decision Trees		5,10,20	2,5,10			5 5,2	0.72744	0.73
		2,3,5	1,3,10	1,2		5 3,3,1	0.72068	0.73
		12,15		3 1,2		5 12,3,1	0.70691	0.71
Logistic Regression	c	solver						
	0.1,1,10					5 0.1	0.66904	0.67
	0.5,1,5	liblinear				5 0.5	0.66904	0.67
SVM	0.01,0.1	liblinear				5 0.01	0.672	0.67
	c	kernel						
	0.01,0.1,1	linear,rbf				5 0.01,rbf	0.69468	0.7
	1,10	rbf				5 1	0.7078	0.71
	0.1,1,10	linera or rbf				5 0.1,rbf	0.7037	0.7

The table presents the results of hyperparameter tuning across four machine learning models; Random Forest, Decision Trees, Logistic Regression, and SVM with a focus on improving accuracy and ROC AUC scores.

1. Random Forest

The highest score was achieved with n-estimators=260 and max depth=10 increasing or decreasing values beyond this showed a decrease even with min-sample split,min sample leaf and max features.

2. Decision Trees

The highest accuracy was achieved with max depth=5, min samples split=2 and here too decreasing or increasing values beyond that brought the ROC scores down.

3. Logistic Regression

A better accuracy was observed when $c=0.01$ and solver=liblinear. Values more than 0.01 for c decreased the accuracy.

4. SVM

Accuracy of SVM model was high when $c=1$ and kernel=rbf. Higher and lower values for c or different kernels reduced its accuracy.

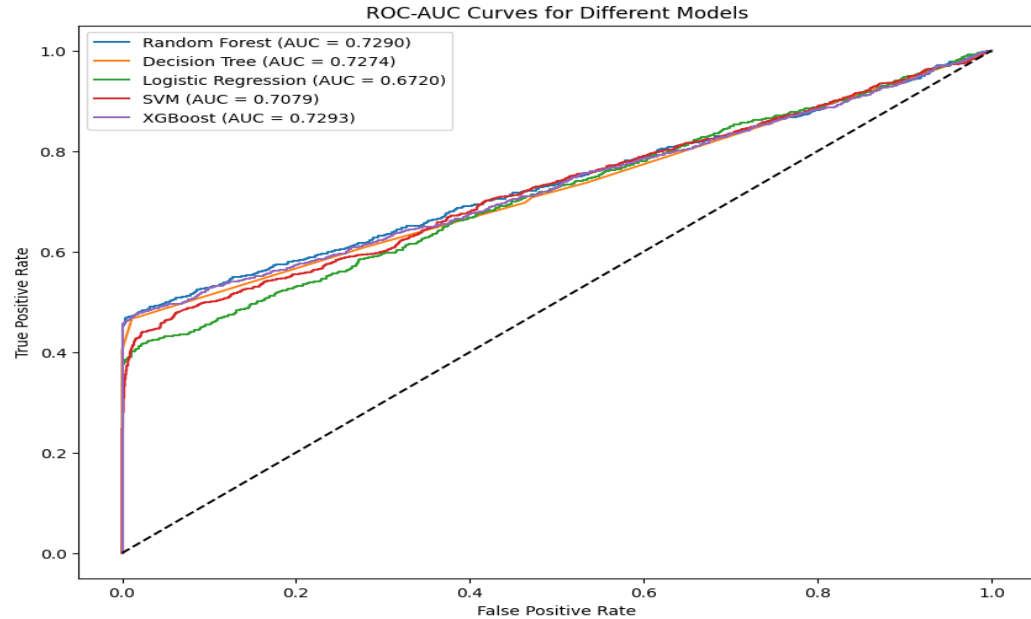
Comparing all 4 models Random Forest performed best with a ROC score of 0.728, Decision Trees also performed closely well with a score of 0.727 while SVM performed moderately with a score of 0.707 while Logistic Regression performed least with a score of 0.672.

While these results show an improvement further tuning of parameters might yield additional performance.

II. Assess the effectiveness of XGBoost compared to traditional algorithms in accurately predicting if the shipment reached time.

	n-estimators	max-depth	learning-rate	subsample	colsample_tree				
XGBoost	50,100,200	3,5,10	0.01,0.1,0.2			5	100,3,0.01	0.7248	0.73
	200,250	5,7	0.1,0.15	0.8	1	5	200,5,0.015	0.729	0.73
	250,300	5,7	0.01,0.015,0.05	0.8	0.8,1	5	250,5,0.01,1	0.728	0.73

XGBoost showed performed well with n-estimators=200, max-depth=5, learning rate=0.015 and colsample-tree=1. Values higher or lower than that reduced its performance.



By comparing the ROC scores of all the best models it is evident that XGBoost has performed relatively better than the traditional models indicating strong predictive performance is identifying on-time and delayed shipments.

Random Forest and Decision Tree curves are slightly below XGBoost but still demonstrates a strong predictive ability. SVM performed slightly lower while Logistic Regression had the lowest score

ROC-AUC is preferred in this case over Accuracy because it is a well-balanced metric for any model performance, which is very relevant in an imbalanced dataset of on-time and delayed deliveries. The accuracy can mislead and support one predominant class, while ROC-AUC actually looks at the ability of the model in distinguishing classes over all threshold levels, thus offering a more comprehensive view of predictive effectiveness in logistics.

CHAPTER 5

DISCUSSION AND RECOMMENDATIONS

5.1 Discussion

In this section a brief out on the interpretation will be discussed while narrowing down the field of study.

This study that tested 4 different machine learning models with various hyperparameters namely Random Forest, Decision Trees, SVM and Logistic Regression to predict on time and delayed shipments indicates that Random Forest model performed best while Logistic Regression performed least. Additionally the XGBoost model too performed quite better than the traditional machine learning models.

The results were better than the prevailing finding by (Collins & Uchenna, 2024) on the same study which was carried out to predict delivery times of products using the ecommerce shipping dataset. Here too random forest algorithm performed best yet with lesser scores than this which shows that the right hyperparameters can bring a difference in model scores.

The main limitation to this study is the computational time that complex hyperparameter and larger values take therefore these models can be improved further if such requirements are met.

Future research can experiment with with other Machine Learning models like Gradient Boosting Machine (GBM), CatBoost or ensemble methods like blending different models

(eg combining LightGBM, RandomForest and CatBoost). Also with better computational resources researchers can opt Deep Learning models like Neural Networks.

5.2 Recommendations

Investing in high performance computational resources will help future studies to test complex models and hyperparameter grids which would in turn lead to significant performance gains in delivery time prediction.

While this study focused on XGBoost and four traditional methods adding Gradient Boosting Machines, CatBoost or ensemble techniques could yield further insights.

Feature Engineering methods like PCA or Recursive Feature elimination may give new insights to predictions impacting delivery time without significantly increasing computational costs.

By predicting whether shipments will reach on time, companies can take proactive steps to increase customer satisfaction and streamline logistics. This includes notifying customers of expected delays, optimizing delivery routes and reallocating resources to prevent delays. Through these strategies the company can improve delivery reliability, enhance customer trust and create a more efficient supply chain.

5.3 Conclusion

This project focused on the enhancement of the prediction of delivery time in e-commerce logistics, investigating and comparing several machine learning models, such as Random Forest, Decision Tree, SVM, Logistic Regression, and XGBoost. By using

systematic hyperparameter tuning, it can be observed that, among the traditional algorithms, the best performance belonged to the Random Forest model, while the XGBoost performed competitively with state-of-the-art algorithms.

The study has identified the potentiality of Predictive analytics for Operational efficiency in logistics using on-time and delayed delivery forecasts. Automatically, this allows companies to be proactive in such measures that would involve route optimization in delivery, re-allocation of resources, and development of better communication with customers for their satisfaction with a competitive advantage. Besides, this research has given a careful selection of hyperparameters as one of the most crucial factors for maximizing performance. Carefully tuned parameters can make a lot of difference in predictive accuracy.

These findings could be a starting point in exploring other machine learning models and ensemble techniques to improve these predictions using high-performance computing. The study provides, therefore, a strong basis upon which predictive analytics could be integrated within the context of e-commerce logistics for promising applications in real-time decision-making and continuous improvement in delivery performance.

APPENDICES

```
rf2 = RandomForestClassifier(random_state=42)
param_grid_rf2 = {
    'n_estimators': [250, 260],
    'max_depth': [10, 15],
    'min_samples_split': [3, 5],
    'min_samples_leaf': [1, 2],
    'max_features': [None]
}

grid_rf2 = GridSearchCV(rf2, param_grid_rf2, scoring='roc_auc', cv=5)
grid_rf2.fit(X_train, y_train)

rf2_best = grid_rf2.best_estimator_

rf2_pred = rf2_best.predict(X_test)
print("Random Forest Best Parameters:", grid_rf2.best_params_)
print(classification_report(y_test, rf2_pred))
print(confusion_matrix(y_test, rf2_pred))
print(f"Random Forest ROC-AUC: {roc_auc_score(y_test, rf2_pred)}")
```

Random Forest Best Parameters: {'max_depth': 10, 'max_features': None, 'min_samples_leaf': 2, 'min_samples_split': 5, 'n_estimators': 260}

	precision	recall	f1-score	support
0	0.65	0.98	0.78	1334
1	0.95	0.48	0.64	1328
accuracy			0.73	2662
macro avg	0.80	0.73	0.71	2662
weighted avg	0.80	0.73	0.71	2662

```
[[1302  32]
 [ 688 640]]
Random Forest ROC-AUC: 0.728969852423186
```

```
[ ] dt = DecisionTreeClassifier(random_state=42, class_weight='balanced')
    param_grid_dt = {
        'max_depth': [5, 10, 20],
        'min_samples_split': [2, 5, 10]
    }
    grid_dt = GridSearchCV(dt, param_grid_dt, cv=5, scoring='roc_auc')
    grid_dt.fit(X_train, y_train)
    dt_best = grid_dt.best_estimator_

    dt_pred = dt_best.predict(X_test)
    print("Decision Tree Best Parameters:", grid_dt.best_params_)
    print(classification_report(y_test, dt_pred))
    print(confusion_matrix(y_test, dt_pred))
    print(f"Decision Tree ROC-AUC: {roc_auc_score(y_test, dt_pred)}")
```

Decision Tree Best Parameters: {'max_depth': 5, 'min_samples_split': 2}

	precision	recall	f1-score	support
0	0.65	0.99	0.78	1334
1	0.97	0.47	0.63	1328
accuracy			0.73	2662
macro avg	0.81	0.73	0.71	2662
weighted avg	0.81	0.73	0.71	2662

```
[[1316  18]
 [ 706 622]]
Decision Tree ROC-AUC: 0.7274401203012952
```

```

lr3 = LogisticRegression(random_state=42)
param_grid_lr3 = {
    'C': [0.01, 0.1],
    'solver': ['liblinear']}

grid_lr3 = GridSearchCV(lr3, param_grid_lr3, scoring='roc_auc', cv=5)
grid_lr3.fit(X_train, y_train)

lr3_best = grid_lr3.best_estimator_

lr3_pred = lr3_best.predict(X_test)
print("Logistic Regression Best Parameters:", grid_lr3.best_params_)
print(classification_report(y_test, lr3_pred))
print(confusion_matrix(y_test, lr3_pred))
print(f"Logistic Regression ROC-AUC: {roc_auc_score(y_test, lr3_pred)}")

```

Logistic Regression Best Parameters: {'C': 0.01, 'solver': 'liblinear'}

	precision	recall	f1-score	support
0	0.63	0.86	0.72	1334
1	0.77	0.48	0.60	1328
accuracy			0.67	2662
macro avg	0.70	0.67	0.66	2662
weighted avg	0.70	0.67	0.66	2662

```

[[1146 188]
 [ 684 644]]
Logistic Regression ROC-AUC: 0.6720051119018804

```

```

svm2 = SVC(random_state=42, probability=True)
param_grid_svm4 = {
    'C': [1,10],
    'kernel': ['rbf']}

grid_svm2 = GridSearchCV(svm2, param_grid_svm4, scoring='roc_auc', cv=5)
grid_svm2.fit(X_train, y_train)

svm2_best = grid_svm2.best_estimator_

svm2_pred = svm2_best.predict(X_test)
print("SVM Best Parameters:", grid_svm2.best_params_)
print(classification_report(y_test, svm2_pred))
print(confusion_matrix(y_test, svm2_pred))
print(f"SVM ROC-AUC: {roc_auc_score(y_test, svm2_pred)}")

```

SVM Best Parameters: {'C': 1, 'kernel': 'rbf'}

	precision	recall	f1-score	support
0	0.64	0.98	0.77	1334
1	0.95	0.44	0.60	1328
accuracy			0.71	2662
macro avg	0.79	0.71	0.69	2662
weighted avg	0.79	0.71	0.69	2662

```

[[1303 31]
 [ 745 583]]
SVM ROC-AUC: 0.7078838216433951

```



```

xgb2 = XGBClassifier(random_state=42)
param_grid_xgb2 = {
    'n_estimators': [200, 250],
    'max_depth': [5, 7],
    'learning_rate': [0.1, 0.015],
    'subsample': [0.8],
    'colsample_bytree': [1.0]
}

grid_xgb2 = GridSearchCV(xgb2, param_grid_xgb2, scoring='roc_auc', cv=5)
grid_xgb2.fit(X_train, y_train)

xgb2_best = grid_xgb2.best_estimator_

xgb2_pred = xgb2_best.predict(X_test)
print("XGBoost Best Parameters:", grid_xgb2.best_params_)
print(classification_report(y_test, xgb2_pred))
print(confusion_matrix(y_test, xgb2_pred))
print(f"XGBoost ROC-AUC: {roc_auc_score(y_test, xgb2_pred)}")

```

XGBoost Best Parameters: {'colsample_bytree': 1.0, 'learning_rate': 0.015, 'max_depth': 5, 'n_estimators': 200, 'subsample': 0.8}

	precision	recall	f1-score	support
0	0.65	0.98	0.78	1334
1	0.96	0.48	0.64	1328
accuracy			0.73	2662
macro avg	0.81	0.73	0.71	2662
weighted avg	0.81	0.73	0.71	2662

```

[[1309  25]
 [ 694 634]]
XGBoost ROC-AUC: 0.7293345044345297

```

References

- Collins, N., & Uchenna, N. (2024). The Role of Predictive Analysis in Ecommerce Logistics: Comparative Analysis of Four Machine Learning Algorithms for Delivery Optimization. *Journal of Shipping and Logistics*, 1-55.
- Kaiwen, X., & Li, L. (2023). A Predict Then Optimize Couriers Allocation Framework for Emergency Last-mile Logistics. 5237-5248.
- Oguz, E., & Ece, N. (2022). Delivery Time Prediction Using Support Vector Machine Combined with Look-back Approach. *Alntelia*, 33-38.
- Selim, H., & Ceren, U. (2024). Development of Cargo Delivery Time Prediction Models. *Cukurova University Journal of Natural and Applied Sciences*, 31-35.
- Verbytskyi, Y. (2023). Delivery Routes Optimization Using Machine Learning Algorithms. *СХІДНА ЄВРОПА: ЕКОНОМІКА, БІЗНЕС ТА УПРАВЛІННЯ*, 85-89.