**TOPIC: E-COMMERCE SHIPPING PREDICTION**

Abstract— This project seeks to analyse the predictive e-commerce shipping mechanisms used by supply chain organizations including their product distribution across their several warehouses. It also seeks to analyse detailed transaction orders, inventory and logics from multiple products and warehouses with the objective of analysing if the products ordered, arrived on time or not. A multiple-product demand prediction approach was used to identify specific features identified in the data obtained. This work used 4 machine learning algorithms to analyse the e-commerce shipping data obtained with the objective of predicting weather or not the products ordered arrived on time or not.

1. **Introduction**

**1.1 Background**

Nowadays, technological improvement has made businesses to collect and analyse customers data at every aspect of the organization. Large multinational e-commerce companies like Amazon which is an American technology company that focuses on e-commerce, online advertising, digital streaming, cloud computing and artificial intelligence has advanced technological platforms which combines data from online payment, sales, inventory management and logistics with the objective of obtaining valuable customer, operational and market data of high quality.

**1.2 Objectives**

In Section 2, we examine the relevant literature. Section 3 entails the data cleaning and preparation. Section 4 outlines the data visualization. Section 5 introduces and elaborates on the data analysis and the various machine learning algorithms used. Section 6 addresses the analysis and conclusion Finally; Section 7 wraps up with references used based on the work presented.

1. **Literature Review**

*1.1 Key Forecasting Approaches*

Accurate demand forecasting plays a pivotal role in the effective management of services, operations, production planning, and revenue management, especially in a data-rich environment. Various statistical and machine learning models have been developed for this purpose, including time series models, artificial neural networks, and support vector machines, as discussed in studies such as Carbonneau et al. (2008), Slimani et al. (2015), Chen et al. (2004), Choi et al. (2018), and others.

***1.2 Enhancing Accuracy with Web Data***

Another avenue of research has explored the use of web data, such as search queries and online reviews, to enhance prediction accuracy, as evidenced in studies like Choi and Varian (2012), Cui et al. (2015), Boone et al. (2018), and Chong et al. (2017). While traditional prediction models heavily rely on historical demand data, these findings have encouraged the exploration of new external information sources available in the era of big data. Empirical studies focusing on large e-commerce companies and retailers, which can be found in works such as Horta¸csu and Nielsen (2010), Hwang and Park (2015), Ahire et al. (2015), and Cui et al. (2018).

***1.3 Big Data Triumphs***

The utilization of big data and machine learning methods has proven successful in various domains beyond demand prediction. In recent times, researchers have re-examined various aspects of operations management through the lens of big data. These applications encompass a range of areas, including but not limited to the newsvendor problem, as discussed in Ban and Rudin (2018), and inventory management, as explored in Bertsimas et al. (2016) and Huang and Van Mieghem (2014) have also delved into this area, along with studies on revenue management and pricing conducted by Ferreira et al. (2015) and Aral and Walker (2014).

1. **Methodology**
   1. ***Data Collection***

The dataset for this project was obtained from an international e-commerce company, available on Kaggle. It consists of 10,999 entries with 12 variables, reflecting the company's intent to derive key insights from its customer database using advanced machine learning techniques. The dataset encompasses customer IDs, warehouse block details, shipping methods, customer service interactions, customer ratings, product costs, prior purchases, product importance categorization, customer gender, offered discounts, product weight, and delivery timeliness. This rich dataset serves as a fundamental resource for analysing customer behaviour and optimizing delivery efficiency.

* 1. ***Data Preprocessing***

**Missing Values**: The dataset was checked for any missing values using the is null method followed by the sum function to aggregate the counts. The results confirmed that there were no missing values across all columns, as each column displayed a zero count for missing data.

**Duplicate Rows:** Duplicate rows in the dataset were identified using the duplicated method, followed by sum to count them. It was found that there were no duplicate rows.

**Data Frame Inspection**: data.info() was used to get a concise summary of the DataFrame, ensuring all columns were accounted for and correctly typed with no null values.

**Statistical Summary:** Describe provided a statistical summary for numerical columns, presenting count, mean, standard deviation, minimum, quartiles, and maximum values, which is helpful for understanding the distribution of data.

1. **Data Visualization**

First, we made three bar graphs to understand the relationships, here is the inference of the bar graphs generated:

1. **On-Time Delivery and Ship Mode:**
   * The analysis reveals that the "Not on Time Delivery" is most prominent in a specific ship mode. Further investigation into the factors affecting delivery timeliness in this mode is recommended.
2. **Importance of Products and Timeliness:**
   * Products categorized as having low importance appear to experience delays in delivery. It would be worthwhile to explore the reasons behind this trend and whether there are opportunities for improvement.
3. **Warehouse Block F and Late Deliveries:**
   * Warehouse Block F seems to be associated with delayed deliveries. Investigating the operations and processes within this warehouse block could unveil insights into the causes of these delays.
4. **Customer Calls and Late Deliveries:**
   * There appears to be a correlation between the number of customer calls and late deliveries. This suggests a potential relationship between customer communication and delivery delays, warranting further examination.

A graph of a warehouse block

Description automatically generated with medium confidence

A graph of different colored bars

Description automatically generatedA graph of different colored bars

Description automatically generated

Further, we did KNN Algorithm and Based on our KNN algorithm; here are our interpretations:

1. **Customer Care Calls Impact:**
   * The KNN model suggests that a high number of customer care calls is associated with a specific pattern in the dataset, leading to a prediction of potential delays in delivery.
2. **Warehouse Block F Impact:**
   * The model implies that products from warehouse block F have a distinct influence on the prediction of late deliveries. This could be due to certain characteristics or operational aspects specific to this warehouse block.
3. **Mode of Shipment Consideration:**
   * The KNN model takes into account the mode of shipment, with the example using "Flight." This suggests that the model has learned patterns related to different shipment modes and their potential impact on delivery times.
4. **Product Importance and Delay:**
   * The code indicates that the model considers the importance of the product (low in the example). This inference aligns with the idea that the KNN algorithm has learned relationships between the level of product importance and the likelihood of delayed deliveries.
5. **Real-World Applicability:**
   * The practical application of the model to new data, as demonstrated in the code, highlights the adaptability and real-world usability of the KNN algorithm for making predictions about potential delays in specific shipping scenarios.

**A diagram of a graph

Description automatically generated**

Further, we did Naïve Bayes Algorithm

On the basis, of our algorithm, below is our inferences:

1. **Accuracy Measurement:**
   * The Naive Bayes model achieved an accuracy of approximately [66]% on the test set. This metric represents the proportion of correctly classified instances, indicating the overall effectiveness of the model.
2. **Probabilistic Predictions:**
   * The model provides probabilistic predictions for both the training and test sets using the predict\_proba function. This offers insights into the likelihood of each class, contributing to the transparency of the model's decision-making process.
3. **Class Membership Prediction:**
   * The predict function is utilized to determine the class membership of instances in both the training and test sets. This allows for a direct comparison between predicted and actual classes, aiding in the assessment of model performance.
4. **New Data Prediction:**
   * The Naive Bayes model is applied to new data (new\_data), and the predictions are printed. This demonstrates the adaptability of the model to make predictions on unseen instances, extending its utility beyond the training and test sets.
5. **Alpha Smoothing Parameter:**
   * The Multinomial Naive Bayes model includes an alpha smoothing parameter (alpha=0.01). This parameter helps handle cases where certain feature values have zero probabilities in the training set, contributing to a more robust and generalized model.
6. **Accuracy Comparison:**
   * The accuracy metric, in combination with other evaluation metrics, can be compared with those of other models in your study. This facilitates a comprehensive assessment of the Naive Bayes model's performance relative to alternative algorithms.

Then, for third algorithm, we used Regression Tree, below is the graph for the same:

**A diagram of a company

Description automatically generated**

Based on this, we interpreted the following:

1. **Decision Tree Initialization**:

* Configured the Decision Tree Classifier with a maximum depth of 4, using Scikit-learn, to prevent overfitting and maintain model interpretability. The classifier was trained on the dataset, utilizing the fit method, to create a model capable of making informed predictions.

1. **Tree Visualization:**

* Employed the plot\_tree function to graphically represent the decision tree, offering visual insight into the decision paths and feature importance.

1. **Predictive Performance:**

* Model accuracy was determined to be 68.45% on the test set, reflecting the classifier's ability to correctly predict the target variable. Utilized the accuracy\_score function to quantitatively evaluate the model's performance, ensuring an objective assessment of its predictive accuracy.

1. **Prediction on New Data:**

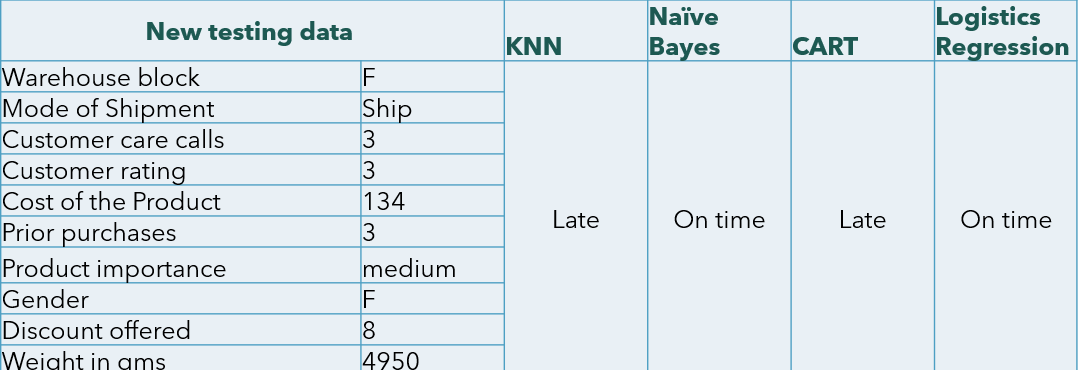
* Demonstrated the model's predictive power on new data inputs, affirming its practical utility beyond the initial dataset.

.

Further, we used Logistics Regression as our fourth algorithm.

1. **Accuracy Measurement:**
   1. The Logistic Regression model achieved an accuracy of approximately [63]% on the test set. This metric represents the proportion of correctly classified instances, indicating the overall effectiveness of the model.
2. **Predictions for New Data:**
   1. The model is used to predict the class of new data (new\_data). This showcases the Logistic Regression model's capability to make predictions on unseen instances, extending its utility beyond the training and test sets.
3. **Binary Classification:**
   1. Logistic Regression is particularly well-suited for binary classification problems. If your target variable consists of two classes (e.g., "Late" or "Not Late" deliveries), Logistic Regression provides probabilities and predicts the class based on a chosen threshold.
4. **Interpretability:**
   1. Logistic Regression models are inherently interpretable. The coefficients assigned to each feature in the model can provide insights into the impact of each predictor variable on the likelihood of a certain outcome.
5. **Evaluation Metrics:**
   1. In addition to accuracy, it's beneficial to explore other evaluation metrics such as precision, recall, and F1 score. The classification report and confusion matrix functions can be used to gain a more nuanced understanding of the model's performance.
6. **Model Fitting:**
   1. The Logistic Regression model is fitted to the training data, adjusting its parameters to learn the patterns present in the training set. This step is crucial for the model to make accurate predictions on new and unseen data.
7. **To select the most suitable model for implementation**

Our comparative analysis extended to a practical scenario where we evaluated the performance of various predictive models on a new set of test data. The data included key customer and order details such as the designated warehouse block, shipping method, customer service interactions, and product weight. Predictions varied across models: KNN and CART indicated a late delivery, whereas Naive Bayes and Logistic Regression predicted the delivery would be on time.



In selecting the most suitable model for implementation, the comparative analysis of model accuracies is crucial. The evaluation encompassed K-Nearest Neighbors (KNN), Naive Bayes, and Logistic Regression algorithms. The Regression Tree emerged as the superior algorithm, exhibiting the highest accuracy at 68.45%. Naive Bayes led with an accuracy of 66.05%, showcasing its strength in probabilistic prediction, followed by KNN and Logistic Regression with accuracies of 63% and 63.5%, respectively. Regression Tree as the optimal model for this application, promising both reliability and valuable insights into the factors influencing on-time delivery.

A graph showing different colored rectangular bars

Description automatically generated

A list of data with black text

Description automatically generated with medium confidence

1. **Conclusion**

To conclude, we used 4 machine learning algorithms to analyse our project such as the K-NN with **63** accuracy, Naive bayes with **66.05** accuracy, logistics regression with **63.5** accuracy and the regression tree having an accuracy of **68.45** with the aim of predicting e-commerce shipment mechanisms using machine learning to determine whether they arrived on time or were been delayed.

The regression tree machine learning algorithm had the highest accuracy of **68.45** and also most shipment arriving on time, therefore the best approach chosen for our project.

**REFERENCES**

Ahire SL, Malhotra MK, Jensen JB (2015) Carton-mix optimization for Walmart. com distribution centers. Interfaces 45(4):341–357.

Aral S, Walker D (2014) Tie strength, embeddedness, and social influence: A large-scale networked experiment. Management Science 60(6):1352–1370.

Ban GY, Rudin C (2018) The big data newsvendor: Practical insights from machine learning. Operations Research.

Berry MJ, Linoff G (1997) Data mining techniques: for marketing, sales, and customer support (John Wiley & Sons, Inc.). Bertsimas D, Kallus N, Hussain A (2016) Inventory management in the era of big data. Production and Operations Management 25(12):2006–2009.

Boone T, Ganeshan R, Hicks RL, Sanders NR (2018) Can google trends improve your sales forecast? Production and Operations Management 27(10):1770–1774.

Boucheron S, Lugosi G, Massart P (2013) Concentration inequalities: A nonasymptotic theory of independence (Oxford

university press). Breiman L (2001) Random forests. Machine learning 45(1):5–32.