CE477 - Project

Yankı Omaç
Electrical and Electronics Engineering
İzmir University of Economics
İzmir, Turkey
Email: yanki.omac@std.izmirekonomi.edu.tr

Recep Atabey Demir
Computer Engineering
İzmir University of Economics
İzmir, Turkey
Email: atabey.demir@std.izmirekonomi.edu.tr

Bersay Yakıcı
Software Engineering
İzmir University of Economics
İzmir, Turkey
Email: bersay.yakici@std.ieu.edu.tr

Halil Arda Aşılıoğlu
Software Engineering
İzmir University of Economics
İzmir, Turkey
Email: arda.asilioglu@std.izmirekonomi.edu.tr

***Abstract*—**

An era where e-commerce is booming, the ability to understand and optimize customer experience is paramount for businesses aiming to thrive. An international e-commerce company, specializing in electronic products, has embarked on an ambitious project to delve deep into their customer database to uncover vital insights that could revolutionize their operations. Leveraging advanced machine learning techniques, the company aims to dissect the complex dynamics of customer interactions and product shipments to enhance satisfaction and efficiency.

I. Introduction

In today's thriving e-commerce landscape, understanding and enhancing the customer experience are essential for business success. Our international e-commerce company, specializing in electronic products, is embarking on a data-driven journey to revolutionize its operations. Leveraging advanced machine learning techniques, we aim to dissect customer interactions and product shipments, with the goal of boosting satisfaction and efficiency. Our robust dataset of 10,999 observations across 12 variables provides insights into the entire customer journey from purchase to delivery. Existing literature emphasizes the importance of data-driven approaches in customer segmentation and recommendation systems.Ms. Stuti M. Meshram and Dr. Neeraj Sahu (2023) showed understanding customer sentiments is crucial for enhancing the overall shopping experience. Also this paper Saxena, Vinod Mahor (2022) focuses on customer behavior analysis using decision tree machine learning and shows decision trees are powerful tools for understanding customer journeys and optimizing strategies. Timely delivery plays a crucial role in customer satisfaction. Reviews often highlight the importance of prompt and reliable delivery. Our project aims to address this by optimizing logistics and predicting delivery times. While existing research provides valuable insights, gaps remain. Our project focuses on last-mile delivery optimization and personalized communication with customers. Our project seeks to empower the enterprise with analytical tools for navigating the e-commerce landscape. By combining data-driven strategies with insights from the literature, we aim to surpass customer expectations and improve operational efficiency.

[1] [2] [3]

II. Data Set Description

The foundation of this analytical venture is a robust dataset comprising 10,999 observations across 12 meticulously curated variables. These variables provide a comprehensive overview of the customer journey, from the initial purchase to the final delivery. Data set is a clean data set and it has no null or missing values.

ID: A unique identifier for each customer, ensuring precise tracking and personalized insights.

Warehouse Block: With the company's expansive warehouse segmented into blocks A through E, this variable helps in logistics optimization and inventory management.

Mode of Shipment: Understanding the impact of different shipment methods (Ship, Flight, Road) on customer satisfaction and delivery efficiency.

Customer Care Calls: The frequency of customer inquiries serves as an indicator of service quality and customer engagement.

Customer Rating: A direct measure of customer satisfaction, with ratings ranging from 1 (lowest) to 5 (highest).

Cost of the Product: This financial metric is crucial for pricing strategies and profitability analysis.

Prior Purchases: Tracking customers' purchase history aids in predicting future buying behavior and personalizing marketing efforts.

Product Importance: Categorizing products based on their importance (low, medium, high) enables tailored handling and prioritization.

Gender: Analyzing shopping patterns and preferences across genders.

Discount Offered: Examining the impact of discounts on sales volume and customer acquisition.

Weight in Grams: The logistical aspect of shipping, influencing costs and delivery methods.

Reached on Time: The critical outcome variable indicating whether a product was delivered within the expected timeframe, serving as a benchmark for operational efficiency. 0 indicates it reached on time and 1 indicates that its not reached on time.

III. Preprocessing

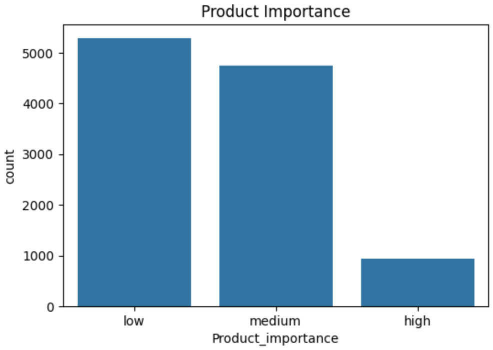

Fig. 1. Histogram of Product Importance

Figure 1, illustrates the distribution of product importance across different categories. The counts for each category are as follows:

Low Importance: Over 5000 occurrences. Medium Importance: Between 4500 and 5000 occurrences. High Importance: Between 750 and 1000 occurrences. The analysis suggests that a significant proportion of products fall into the low importance category, while fewer products are classified as high importance.

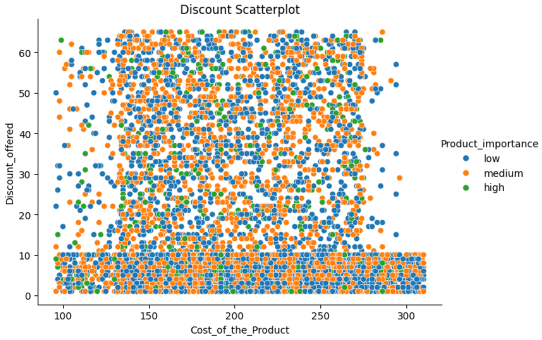

Fig. 2. Scatter plot of Cost and Discount

Figure 2, illustrates the relationship between the cost of a product and the discount offered. Each data point represents a product, color-coded by its importance level (low, medium, or high).

X-axis: Represents the cost of the product (ranging from 0 to 300). Y-axis: Represents the discount offered (ranging from 0 to 60). Observations:

Data points are widely scattered, indicating variability in discounts across different product costs and importance levels. No clear linear trend is visible; the distribution is diverse. The dense concentration of points occurs across all levels of product importance, especially in the mid-range of cost and discount. This scatter plot provides insights into the complex relationship between cost and discount, suggesting that discounts are not consistently tied to either product cost or importance.

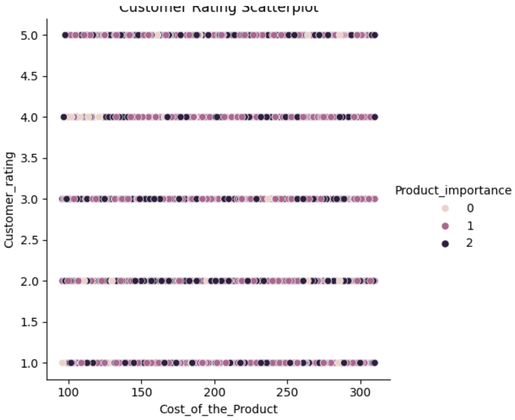

Fig. 3. Scatter plot of Cost and Customer

Figure 3, represents customer ratings in relation to the cost of the product. The x-axis displays the cost of the product, ranging from 100 to 300. The y-axis represents customer ratings, which vary from 1.0 to 5.0. Regardless of the cost and importance level, products consistently receive a wide range of ratings. Data points are distributed evenly across all levels of customer ratings.

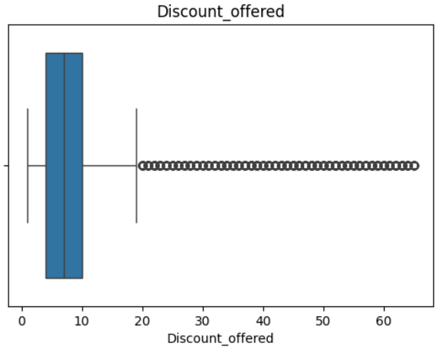

Fig. 4. Box plot of Discount

Figure 4, illustrates the discounts provided by a certain entity. The x-axis represents the percentage of discount, ranging from 0 to 60. Notably, a prominent blue bar extends upwards around the 10 mark on the x-axis. This suggests that a significant number of offers fall within this range. The presence of error bars on both sides of the blue bar indicates variability or a range in which these discounts are typically offered.There are no lower outliers, indicating that extremely low discounts are not common. However, the upper end of the data exhibits a cluster of outliers. These outliers extend beyond the IQR, reaching up to nearly 60 discount. While such high discounts are infrequent, they do exist.

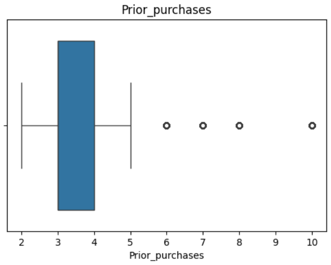

Fig. 5. Box plot of Prior Purchases

The blue rectangular box in the Figure 5 represents the interquartile range (IQR). This range encapsulates the middle half of the observed prior purchase values. Specifically, the lower edge of the box corresponds to approximately 3 on the x-axis, while the upper edge aligns with around 4. This implies that half of the customers made prior purchases within this range. The line inside the blue box represents the median value. In this case, the median is close to 3 on the x-axis. The median serves as a robust measure of central tendency, indicating that half of the customers made three or fewer prior purchases. Beyond the upper whisker of the box, we observe several outliers represented by circles. These outliers correspond to individual observations that deviate significantly from the bulk of the data. Specifically, these customers have made notably more purchases, ranging from approximately 6 to nearly 10.

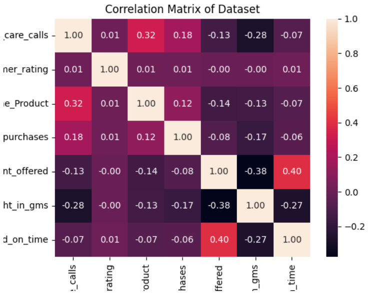

Fig. 6. Heatmap of Numerical data

Figure 6, illustrates the correlation coefficients between various numerical variables. The color gradient on the right side of the heatmap corresponds to the range of correlation coefficients. It spans from -0.4 (depicted in dark purple) to 1.0. Warmer colors (reds and oranges) indicate positive correlations, while cooler colors (purples) represent negative correlations. Each cell in the matrix displays a numerical value representing the correlation coefficient between a pair of variables. Notably, there is a positive correlation of 0.40 between Discount offered and Reached on time. This suggests that as the discount offered increases, the likelihood of orders being delivered on time also rises. Additionally, Customer care calls and Cost of the Product exhibit a positive correlation of 0.32. Conversely, there is a negative correlation of -0.38 between Discount offered and Weight in gms. This implies that higher discounts may be associated with lighter products.

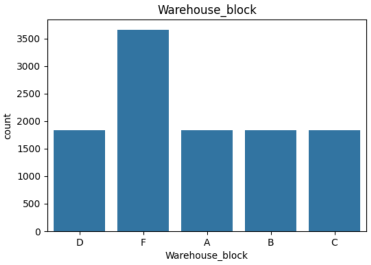

Fig. 7. Distribution of Warehouses

Figure 7, illustrates the distribution of warehouses across different blocks labeled D, F, A, B, and C. Block F has the highest count of warehouses, exceeding 3500. It represents the most densely populated area in terms of warehouses. Other blocks have similar counts of warehouses, both slightly below 2000. While they don't have as many warehouses as Block F, they still contribute significantly to the overall distribution.

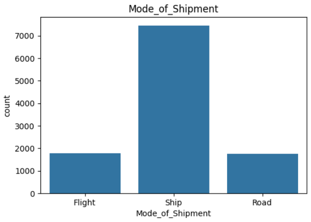

Fig. 8. Distribution of Shipment

Figure 8, illustrates the distribution of shipments across three primary modes of transport: flight, ship, and road. Each mode is represented by a distinct blue bar. Notably, shipping emerges as the dominant choice, accounting for over 7,000 shipments. In contrast, both flight and road transport exhibit significantly lower counts, hovering just above 1,000 each. This data underscores the reliance on maritime transport within the context of our study.

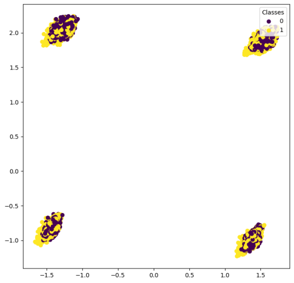

Fig. 9. 2D Visualization of PCA

Figure 9 and Figure 10 illustrates the visualization of PCA being applied to data set. Target class is Reached on time

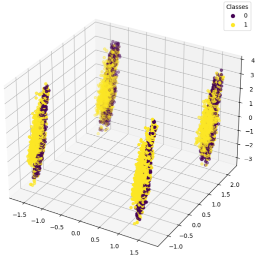

Fig. 10. 3D Visualization of PCA

which has either 0 or 1 values. Thats why it is a 2-class PCA. This PCA graph may not be that useful since both 0's and 1's are intertwined in their relative clusters. There is no obvious distinction.

IV. Classification

Various classification algorithms are being trained. For choosing ideal parameters, a grid-search block is implemented, which is an exhaustive search method to find best parameters for highest results.

Model	Accuracy
LightGBM	0.6727
Logistic Regression	0.6427
Naive Bayes	0.6441
Random Forest	0.6768
KNN	0.6771
Decision Tree	0.6809

All the models above are being trained with using a 5-fold cross-validation grid search mechanism. Target variable was 'Reached on Time'. If delivery reaches on time its value is 1, if not then its 0.

LightGBM is an optimized Gradient Boosting method that performed with 0.6727 accuracy score. Logistic Regression which is a binary classification method performed slightly worse with 0.6427. Naive Bayes comes next with an in-between result such as 0.6441. Random Forest outperformed all the models so far with 0.6768 accuracy and KNN model had slightly better score with 0.6771.

Best accuracy score is Decision Trees so far with 0.6809. Since our data contains so many parameters, its rule-based approach is better for classifying our target variable and eliminating redundant features.

Figure 11 illustrates the Decision Tree. Like any other tree structure, decision tree also splits up to two new nodes every time it can. It contains several rules and it will go to that specific branch of the tree. Currently tree is very symmetric and gives best metric results in terms of accuracy.The wieght in grams is the mostly used feature in the decision and it is considered as the most important feature in case of product arrival.

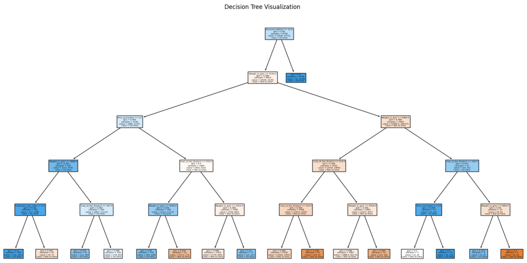

Fig. 11. Decision Tree Visualization

V. Regression

For regression part of the problem, Product cost is chosen as our target variable. Like classification part, grid search algorithm is also used and benefited to optimize hyperparameters and to get best results. Two selected models are KNN and Linear Regression.

Model	MAPE	RMSE
KNN	83.59	0.4284
Linear Regression	91.03	0.4472

KNN is chosen because of its success in classification and Linear Regression is chosen due to its easy implementation. Linear regression did not worked out as intended meaning that data cannot be explained with a linear relationship. KNN worked slightly better but its still not enough to explain data too. This can lead to a conclusion where either these models are not suitable for this task or dataset is not enough to learn product cost from other variables. 83.59 and 91.03 percent mean absolute percentage errors are too much to rely on these models.

VI. Clustering

In this section we will share our findings with clustering methods. We used 3 different clustering methods which are K-means clustering, Dendogram clustering and DBSCAN clustering.

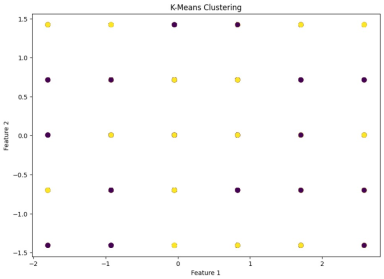

Fig. 12. K-Means

In K-Means clustering shown in Figure 12, data has been divided into 2 clusters with similar distributions. K-Means did an alright job of dividing clusters. Although K-Means identifies distinct clusters well, non-uniformity of data distribution may affect the performance of the algorithm.

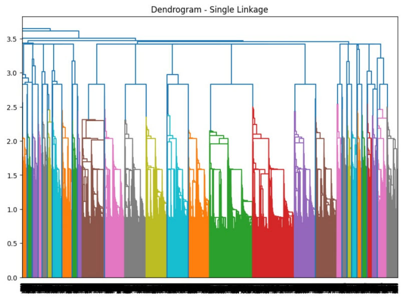

Fig. 13. Single Link Dendrogram

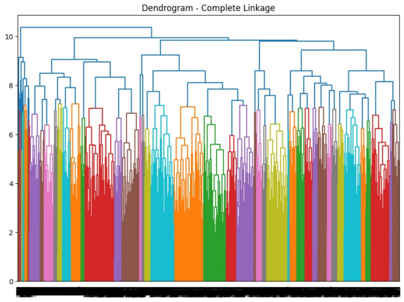

Fig. 14. Complete Link Dendrogram

Figure 13 and 14 shows Dendrogram based clusters with both single link and complete link. Single link provides elongated clusters and complete link provides more balanced clusters. It is advantageous to form compact clusters, but may miss some natural structures.

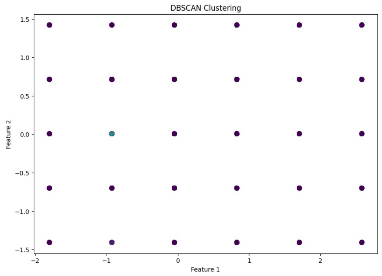

Fig. 15. DBSCAN

Figure 15 shows DBSCAN clustering which is our choice for density-based clustering.Most data points fell into a single cluster, with few noise points identified in this type of clustering.

In conclusion, K-Means identifies distinct clusters well, Hierarchical Clustering provides hierarchical analysis of the data structure, and DBSCAN is successful in identifying noise points. The strengths and weaknesses of each algorithm should be evaluated based on the data set and the purpose of the analysis.

VII. Ensemble

In this section we will share our findings with bagging and boosting methods. AdaBoost and Bagging methods are being trained in this section. Results are in the table below.

Model	Accuracy
AdaBoost	0.6373
Bagging	0.6809

Both models used Decision Tree as their estimator and with parameter grid earned from grid search. AdaBoost performed worse than base model and Bagging performed the same. This leads to a conclusion that applying these models does not improve test accuracy in an important manner.

VIII. Association Mining

Apriori Algorithm is applied to our data set. Results are shown in the table below.

Antecedents	Consequents	Confidence
3	1	0.712007
F	1	0.714655
M	1	0.709571
Ship	1	0.712946
low	1	0.707948
medium	1	0.708877
F, 3	1	0.716827
Ship, 3	1	0.709797
Ship, F	1	0.712715

The table has items like F and M which corresponds to the gender of the customer, ship corresponds to the shipment vehicle,low and medium corresponds to the product importance and the 1 in the consequent refers if the product arrived on time or not, but we couldn't find a way to understand which columns do the numbers like 3 come from so we cannot explain the rules containing these numbers.

The results show that products which are shipped via ships have more chance to arrive on time and Products with low or medium product importance have higher chances of arriving on time. The genders of the customers are not taken as an important factor.

IX. Discussion

This study aims to evaluate the machine learning models for several tasks, including classifying whether products arrived on time and predicting the cost of a product. Various machine learning algorithms were applied to the dataset, with the ”reached on time” attribute targeted for classification and the cost of the product attribute targeted for regression.

For classification, six different models underwent a 5-fold cross-validation grid search mechanism with a range of parameters defined for each. These models include LightGBM, Logistic Regression, Naive Bayes, Random Forest, KNN, and Decision Tree. Among them, the Decision Tree performed the best with an accuracy score of 0.6809. This finding suggests that the Decision Tree algorithm best represents the dataset, indicating its effectiveness in discerning patterns related to delivery punctuality.

In regression analysis, two models were trained: KNN and Linear Regression. Out of these, KNN demonstrated better performance compared to Linear Regression. The superiority of KNN implies that the dataset may lack a linear relationship between predictors and the target variable, indicating its complexity. However, KNN's performance fell short of expectations, suggesting either the regression task's unsuitability for the dataset or the necessity for additional models to achieve improved results.

Furthermore, ensemble learning techniques were explored in this study. AdaBoost and Bagging methods, employing Decision Trees as base estimators, were trained and evaluated. Despite optimization through grid search, both AdaBoost and Bagging failed to significantly enhance test accuracy. This indicates that the application of these ensemble methods did not yield substantial improvements in predictive performance.

Additionally, association mining using the Apriori Algorithm provided insights into the relationships between various factors affecting delivery punctuality. Results revealed that products shipped via ships and those categorized as low or medium importance were more likely to arrive on time.

Cluster analysis was performed to gain further insights into the dataset structure. The K-Means algorithm, with K = 2, effectively divided the data into two clusters. Hierarchical Clustering, employing both Single and Complete Linkage methods, provided hierarchical analysis of the data structure, with Single Linkage forming elongated clusters and Complete Linkage forming more balanced clusters. Moreover, DBSCAN successfully identified noise points and irregularly shaped clusters based on density, although parameter selection was crucial.

In conclusion, the study comprehensively evaluated various machine learning models and clustering algorithms for analyzing e-commerce data. While Decision Tree excelled in classifying delivery punctuality, KNN showed promise in regression analysis. Ensemble methods did not significantly enhance predictive performance, and association mining provided valuable insights into factors affecting delivery punctuality. The selection of appropriate algorithms should be based on the dataset characteristics and analytical objectives.

REFERENCES

[1] W. O. Gibin, "On-time delivery dataset," 2024. [Online]. Available: https://www.kaggle.com/datasets/willianoliveiragibin/on-time-delivery/code

[2] P. V. M. Ankit Saxena, "Customer behavior analysis in e-commerce using machine learning approach," 2022.

[3] D. N. S. Ms. Stuti M. Meshram, "Sentiment analysis of e-commerce product review through machine learning," 2023.