***Factors encouraging cycle commuting in Scotland***

by

Taimoor Muzaffar Gondal

taimoormuzaffargondal@gmail.com

CONTENTS

List of Figures III

List of Tables IV

1 Introduction 1

2 Dataset: E-Commerce Shipping Dataset 2

2.1 Aims and Analysis 2

2.2 Key Challenges and Problems with Dataset 4

2.3 Summary Statistics for Dataset 6

3 Unsupervised Analysis 9

3.1 K-Mean Clustering 9

3.2 Summary of K-Mean 11

4 Supervised Analysis 12

4.1 Naive Bayes Method 12

5 Reflections and Conclusion 15

7 Eviroment 17

8 Bibliography 17

LIST OF FIGURES

[Figure 2.1: Overview E-Commerce Shipping Dataset](#_Toc76770364) 3

[Figure 2.2: Identifying any missing values in the dataset](#_Toc76770365) 4

[Figure 2.3: Outliers Inspection for Dataset](#_Toc76770366) 5

[Figure 2.4: Outliers Removal from One Feature of the Dataset](#_Toc76770366) 5

[Figure 2.5: Correlation of complete features of the dataset](#_Toc76770364) 7

[Figure 2.6: Distribution plots for the E-Commerce Dataset](#_Toc76770365) 8

[Figure 3.1: Elbow Methos Results for the E-Commerce Shipping Dataset](#_Toc76770366) 9

[Figure 3.2: K Means Clustering Result for k = 2](#_Toc76770366) 10

[Figure 3.3: K Means Clustering Result for k = 3](#_Toc76770366) 11

[Figure 4.1: Confusion Matrix](#_Toc76770366) 13

[Figure 4.2: RoC for TPR and FPR](#_Toc76770366) 14

**LIST OF TABLES**

[Table 2.1: Statistical Analysis of Important Features of Dataset](#_Toc76770364) 6

[Table 4.1: Score Values for Naïve Buyes Algorithm](#_Toc76770364) 12

**Chapter 1**

**Introduction**

E-commerce shipping includes everything from receiving and processing orders to choosing and packing the purchased item at a warehouse, creating shipping labels, and even handling returns.. It could seem straightforward enough. It might even appear straightforward if you've seen a well-run warehouse. However, the complexity that results from all the moving parts depends on several factors, including the sorts and sizes of the products you sell, the locations to which you'll ship, the delivery options and shipping rates you decide to provide, and more.

The entire e-commerce shipping process, from receiving and processing an order through packing, shipping, and delivering it to the customer's door, is covered. When a company receives an order, staff should first check their inventory levels, then double check the customer's shipping address and any other important details. The items will then be chosen, packed, and ready for shipping.

In this report we are going to perform statistical analysis on the dataset and identify key challenges in the dataset i.e., outliers or anomalies in it. Moreover, Its correlation, data exploration, and data cleaning will be performed. Moreover, Unsupervised and supervised analysis by using machine leaning algorithms, where their correlation, accuracy, F1-score, precision and recall will be calculated and models will be evaluated on those basis. For the case of K Mean, Elbow method will be implemented to choose the best number of clusters in the E-commerce shipping dataset.

**Chapter 2**

**Dataset: E-Commerce Shipping Dataset**

**2.1 Aim and Analysis**

In this dataset, a collection of multiple features has been gathered by a company. In which key insights from a consumer database are sought after by a global e-commerce corporation. They intend to research their customers using some of the most cutting-edge machine learning methodologies. The business sells electronics.

The dataset that was used to create the model had 10999 observations across 12 variables.

The following details are included in the data as its column names:

* **ID:** Customer ID number
* **Warehouse\_block:** The business has a sizable warehouse that is organized into blocks like A, B, C, D, and E.
* **Mode\_of\_shipment:** The business offers a variety of shipping options, including ship, flight, and road.
* **Customer\_care\_calls:** The quantity of phone calls made to inquire about shipments.
* **Customer\_rating:** Every client has given the business a rating. the worst), and five is the highest (Best).
* **Cost\_of\_the\_product:** Product price in US dollars.
* **Prior\_purchases:** The quantity of earlier purchases.
* **Product\_importance:** The corporation has divided their goods into many categories, such as low, medium, and high.
* **Gender:** Male and Female.
* **Discount\_offered:** Given a relief on those goods.
* **Weight\_in\_gms:** In grammes, weights of the products being purchased.
* **Reached\_on\_time:** Basically its the target variable to be achienced, where a value of 1 denotes that the product did not arrive on time and a value of 0 denotes that it did arrive on time.

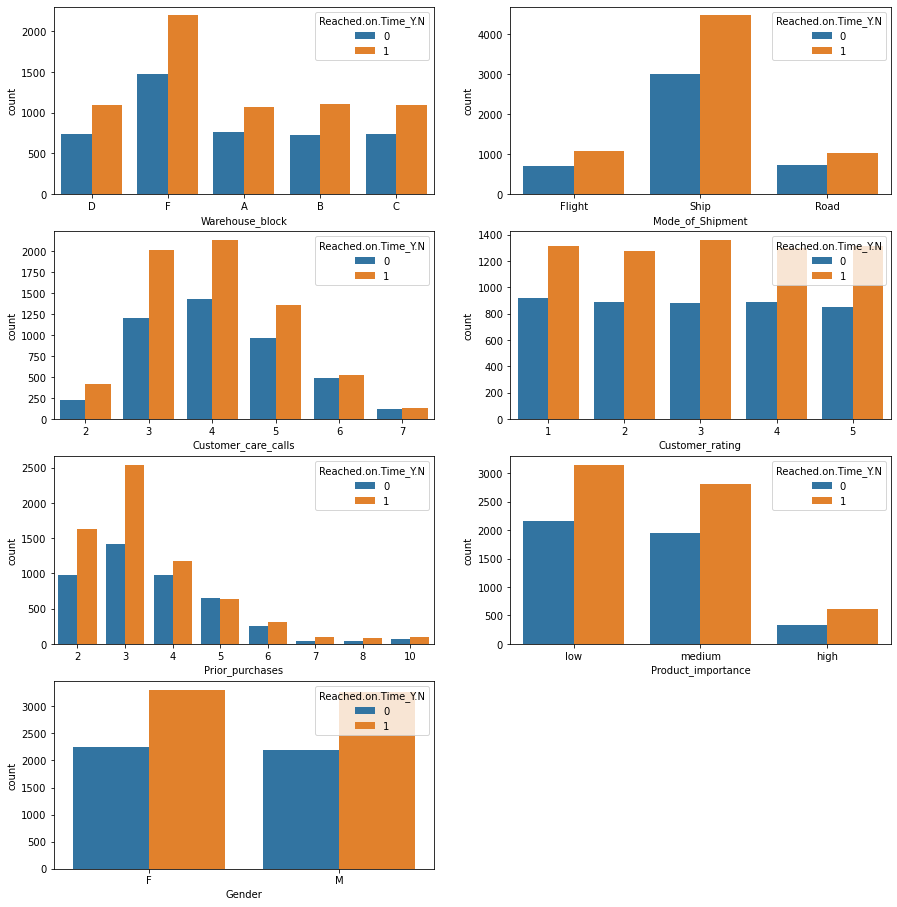
****

Figure 2.1: Overview E-Commerce Shipping Dataset

**2.2 Key Challenges and Problems with Dataset**

Usually, dataset contains many missing values and anomalies due inappropriate data collection points [1]. In this dataset, after processing complete exploratory data analysis technique using pandas, I have check dataset with any missing values. During analysis it has been observed that the dataset contains no missing values as shown in Figure 2.2.

Graphical user interface, application, table, Excel

Description automatically generated

Figure 2.2: Identifying any missing values in the dataset

Meanwhile, as expressed earlier second issue in a dataset is to deal with outliers or anomalies. Anomalies can occur in any dataset due to various issues i.e., human error, data collection error, server error, any kind of cyber-attack and third-party intrusions etc. During dataset analysis, few outliers were identified as shown in Figure 2.3. In this figure, it can be observed that six columns out of twelves have been selected known as call by customer care, ratings provided by customers, total product cost, purchases made by customers on priority basis, discounts being offered by the company to customers, weight of the product. In this figure, it can be observed that boxplot of feature Discount\_offered has shown anomalies or the outliers in the dataset.

To explicitly deal with these outliers, IQR method has been used. Where the upper limit of this columns has been selected and the outliers were removed by adopting this approach. The Boxplot for the dataset after removing outliers have been shown in the Figure 2.4.

Diagram

Description automatically generated with medium confidence

Figure 2.3: Outliers Inspection for Dataset

Chart

Description automatically generated

Figure 2.4: Outliers Removal from One Feature of the Dataset

**2.3 Summary Statistics for the Dataset**

Brief instructive coefficients known as expressive measurements are utilized to summarize a specific informational collection, which might be an example of a populace or a portrayal of the total populace [2]. Estimations of focal propensity and proportions of fluctuation make up clear measurements (spread). The mean, middle, and mode are focal propensity measurements, while the standard deviation, difference, least and greatest factors are proportions of inconstancy.

In table 2.1, a briefed overview of mean, standard deviation, minimum and maximum values of the important features of the dataset i.e., Call, ratings, costs, purchases, discount, weight and delivery time has been presented. It can be observed that on average approximately 60% of the products were delivered on time to the customers. Similarly, they on average they have the rating of 2.99 out of 5. So the statistical analysis of any dataset plays vital role in dataset on determining the key attributes of the dataset.

**Table 2.1: Statistical Analysis of Important Features of Dataset**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Calls** | **Rating** | **Cost** | **Purchases** | **Discount** | **Weight** | **Reached** |
| **mean** | 4.054459 | 2.990545 | 210.196836 | 3.567597 | 13.373216 | 3634.01673 | 0.596691 |
| **std** | 1.14149 | 1.413603 | 48.063272 | 1.52286 | 16.205527 | 1635.37725 | 0.490584 |
| **min** | 2 | 1 | 96 | 2 | 1 | 1001 | 0 |
| **max** | 7 | 5 | 310 | 10 | 65 | 7846 | 1 |

Another important feature in any dataset is the correlation of all the feature with the dataset is considered. One attribute can be predicted from another via correlation. Positive correlation means that when feature A changes, the attributes of feature B also changes, and vice versa when feature A changes, feature B also changes. Both properties have a linear relationship and move together. If a feature A grows, then a feature B decreases, and vice versa, is said to have a negative correlation. No correlation exists between those two characteristics. Every one of those correlation types has a range of values from 0 to 1, with mildly or strongly positive correlation characteristics being around 0.5 or 0.7. A correlation score value of 0.9 or 1 indicates the outcome when there is a strong and perfect positive correlation. Figure 2.5 provides an overview of the correlation of the dataset among its features.

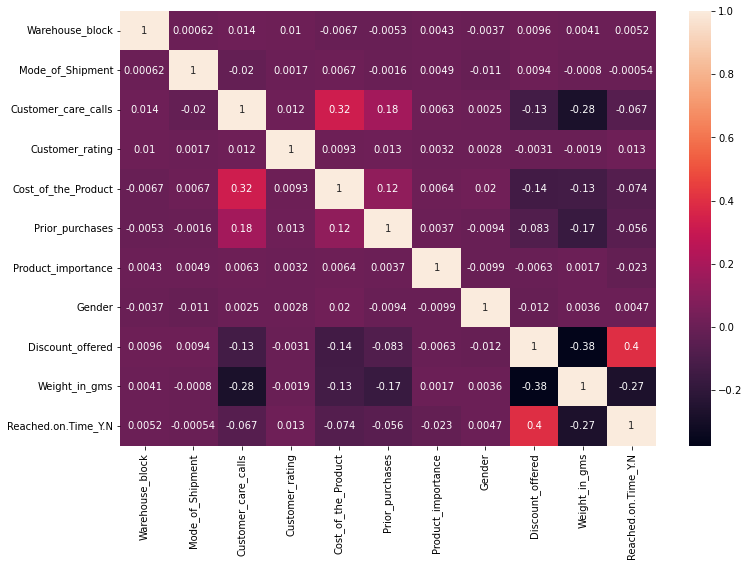


Figure 2.5: Correlation of complete features of the dataset

From this figure correlation for each feature with each other can be seen. Moreover, it can also be confirmed that the correlation of a column with itself is always 1. Meanwhile from figure it can bee estimated that few feature has negative correlation and few feature posses the positive correlation. In this dataset, zero correlation cannot be observed. Apart from the correlation, another important statistical analysis feature is distribution plots for the any dataset. By differentiating the observational dissemination of the information with the hypothetical qualities expected from a specific circulation, conveyance plots give a visual evaluation of the dispersion of test information. It highlights the impact of altering distributions and parameter values, displays where goal values sit within a distribution, and allows users to see the proportions connected to key areas. Figure 2.6 shows the distribution plot for data under consideration.

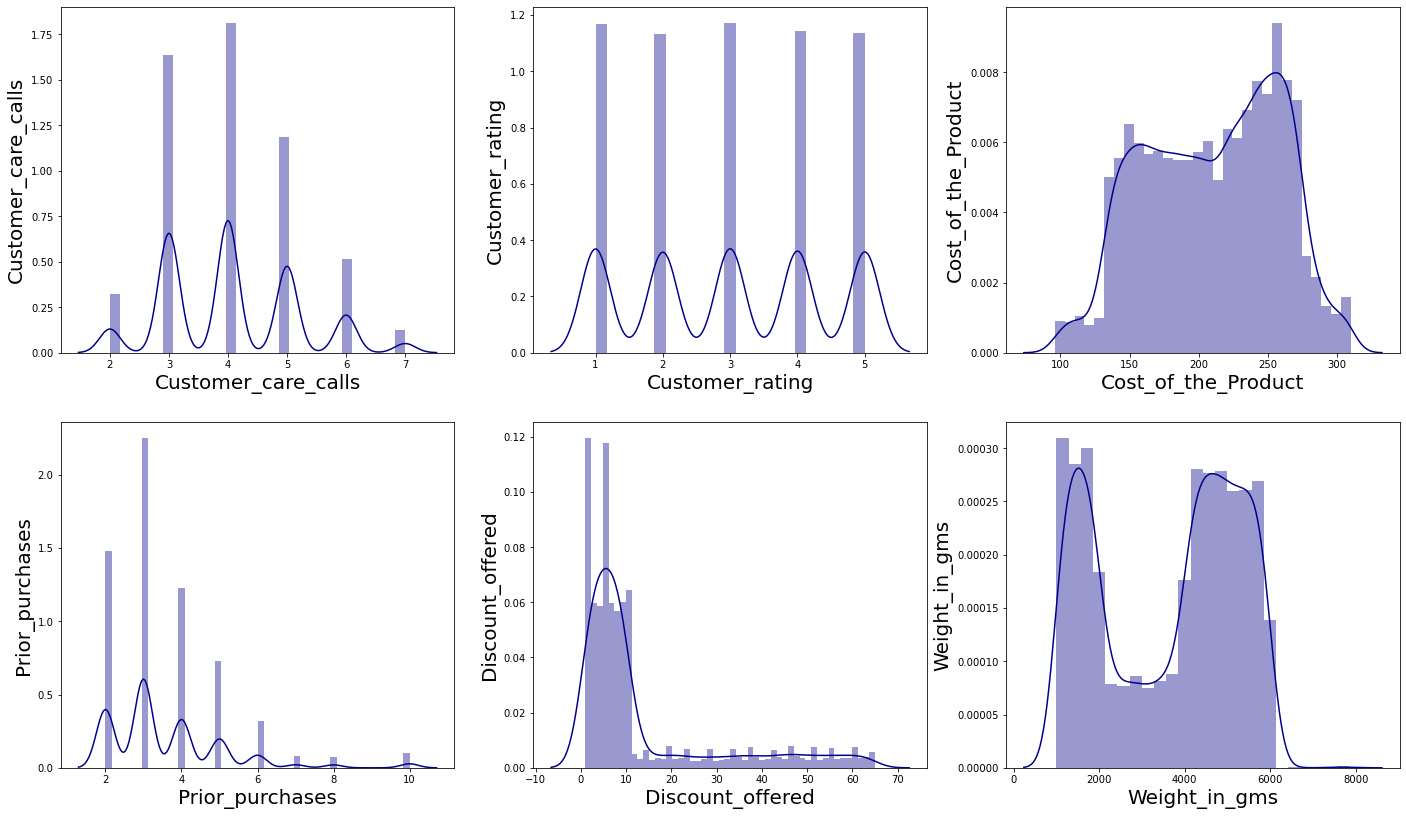


Figure 2.6: Distribution plots for the E-Commerce Dataset

From figure 2.6, we can depict the distribution of each feature. It can be seen that discount price being offered on the purchased products has a lower distribution trend. Similarly, costs of the products have overall distributed with higher volume. While summarizing the statistical analysis section, it can be concluded that there are mainly correlation, distribution plots and statistical summary features such as mean, variance and standard deviation etc. All these parameters play a vital role in determining the key role of features in predictions.

**Chapter 3**

**Unsupervised Analysis**

Clustering is used in datasets with two or more variable quantities to identify collections of similar objects. In actuality, this information might be acquired from a range of databases, such as those for geography, biology, or marketing. In this dataset, as we are analyzing the E-commerce shipping dataset. So the goal here is to find clusters in data that is not well-documented, K-means clustering method is used. This can be used to confirm structural assumptions about the types of groups that exist or to find unknown groups in large and complex data sets.

**3.1 K Means Clustering**

K-means employs an iterative process to find a final clustering depending on the number of clusters defined by the user by searching for a predetermined number of clusters within an unlabeled dataset. In this dataset, to observe the homogeneity and completeness score, I have performed KNN models with K=2, and it shows the Inertia values of 778.3247436724189.

Chart, line chart

Description automatically generated

Figure 3.1: Elbow Methos Results for the E-Commerce Shipping Dataset

Here, we cannot determine the impact or the performance of the model. To observe this performance, Elbow Method was applied on the selected features of the dataset to determine the best number of clusters to opted with K Means. Figure 3.1 shows the results of Elbow Methos. The elbow method is a heuristic approach which used in cluster analysis in machine learning problems to estimate the number of clusters in any data set. Considering the variable being described as a function of the total number of clusters, here the method is to select the curve of the system as the appropriate number of clusters. The graph demonstrates that k=3 is a reasonable choice. Because the curve is monotonically declining, it can be challenging to determine how many clusters to utilize because there may be no elbow or a clear point where the curve begins to flatten out.

A picture containing graphical user interface

Description automatically generated

Figure 3.2: K Means Clustering Result for k = 2

Here lets us observe the performance of K Means with k = 2. Figure 3.2 presented the results of K Means with K=2. Similarly, Figure 3.3, presents the results with K = 3. K-means clustering seeks to generate groups out of comparable types of items. It determines whether two objects are similar to one another and clusters them. The three steps that the K-means clustering method takes. Finding clusters where the observations within each cluster are more similar than the clusters themselves is a sign of an effective clustering solution.

Chart

Description automatically generated

Figure 3.3: K Means Clustering Result for k = 3

Here we can see that with the clusters value k = 3, our model proposed exceptionally well with the inertia value, 462.2968436984273.

**3.2 Summary of K Means:**

E-commerce shipping data performed exceptionally well with K = 3, with inertia value 462.2968436984273, homogeneity score = 0.10673791589138258 and completeness score of 0.5183662098362855.

**Chapter 4**

**Supervised Analysis**

Using supervised approaches necessitates that the labels used to identify a preceding classification be attached to the genes and associated circumstances. The classification name may be related to a disease subtype. In this instance, the supervised mode analysis of microarray data is used to classify the samples. In this chapter I am going to implement Naïve Buyes. Naïve Bayes algorithm is a supervised learning method for classification problem based on Bayes method. It is often used in word processing and extensive training. Naive Bayes classification is one of the simplest and most effective classification techniques available today. It helps in the development of machine learning software that can make accurate predictions. As a practitioner, you make predictions based on what will happen.

**4.1 Naïve Buyes Method**

In this dataset, the company wanted to estimate specific objectives as depicted in the previous section. After applying Naïve Buyes supervised learning model the 65% accuracy was achieved with the precision, recall, f1-score and support shown in Table 4.1.

**Table 4.1: Score Values for Naïve Buyes Algorithm**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1-Score** | **Support** |
| 0 | 0.54 | 0.99 | 0.70 | 895 |
| 1 | 0.98 | 0.43 | 0.60 | 1305 |
| Accuracy | - | - | 0.66 | 2200 |
| Macro Avg | 0.76 | 0.71 | 0.65 | 2200 |
| Weighted Avg | 0.80 | 0.66 | 0.64 | 2200 |

Here, Precision is the ratio of the number of relevant documents found during a search to the total number of documents found during that search, whereas recall is the number of relevant documents found during a search divided by the total number of relevant documents already present. In the Meanwhile, the exactness of a model on a dataset is checked by the F-score, otherwise called the F1-score. It's utilized to survey paired classification conspires that mark models as "positive" or "negative." The consonant mean of the model's accuracy and review is known as the F-score, which is a strategy for joining the model's accuracy and review. The F-score is as often as possible used to evaluate AI models, especially those utilized in normal language handling, as well as data recovery frameworks like web crawlers [3]. Another Important matrix is the correlation in case of Naïve Buyes algorithm. For this specific dataset the obtained confusion matrix is given in the figure 4.1 is given below.

Chart, treemap chart

Description automatically generated

Figure 4.1: Confusion Matrix

In the event that your dataset has multiple classes or on the other hand, assuming each class has an inconsistent measure of perceptions, grouping exactness alone might misdirect. You can acquire a better understanding of the categorization model's successes and failures by calculating a confusion matrix. Similarly, the region of convergence for this case is given in the figure 4.2 below.

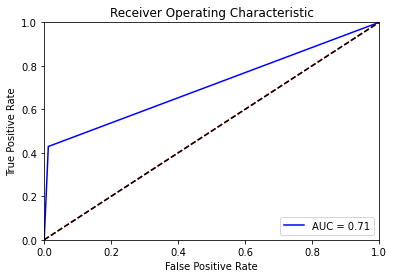


Figure 4.2: RoC for FPR and TPR

**Chapter 5**

**Reflection and Conclusion**

In this report I have explored an E-commerce shipping dataset which was gathered by a company to get insights about their potential buyers and products being purchased with respect to its ratings. In this report, I have performed data exploratory analysis to analyze the dataset with respect to statistical analysis and outliers identifications. Upon the identification of outliers, those were removed and data cleaning process have been performed. After successful data cleaning, Unsupervised and Supervised Analysis was carried out by using K Mean and Naïve Buyes algorithm. During the implementation of K Mean algorithm three clusters were selected by using the elbow method. In which it was observed that the dataset analysis was improved while selecting three clusters with better inertia values. Similarly, it completeness score and homogeneity score was also determined. Afterwards, during supervised learning analysis, Naïve Buyes approach was used and its RoC, Correlation matrix, Precision, Recall, Accuracy and Support. It has been observed that it performed improved performance. For more improved performance other methods of supervised learning can be used for this purpose.

**Appendix A**

**Environment**

Language: Python

IDE: Jupyter Notebook

System Specs: i5, 6th Gen, 8 GB RAM

**References**

[1] Bar-Ilan, Judit. "Data collection methods on the Web for infometric purposes—A review and analysis." *Scientometrics* 50, no. 1 (2001): 7-32.

[2] Oprescu, Andreea M., Gloria Miró-Amarante, Lutgardo García-Díaz, Victoria E. Rey, A. Chimenea-Toscano, Ricard Martínez-Martínez, and M. C. Romero-Ternero. "Towards a data collection methodology for Responsible Artificial Intelligence in health: A prospective and qualitative study in pregnancy." Information Fusion 83 (2022): 53-78.

[3] Limbong, Josua Josen Alexander, and Irwan Sembiring. "Analisis Klasifikasi Sentimen Ulasan Pada E-Commerce Shopee Berbasis Word Cloud Dengan Metode Naive Bayes Dan K-Nearest Neighbor." Jurnal Teknologi Informasi dan Ilmu Komputer 9, no. 2 (2022): 347-356.