E-commerce shipping: Prediction of on time delivery of products using Data Mining

Anika Hossain Bristy
Department of CSE
Daffodil International University
Dhaka, Bangladesh
+8801715703978
anika15-10697@diu.edu.bd

Alimozzaman Durjoy
Department of CSE
Daffodil International University
Dhaka, Bangladesh
+8801795004461
alimozzaman15-11017@diu.edu.bd

Md. Hasibul Hasan
Department of CSE
Daffodil International University
Dhaka, Bangladesh
+8801737900301
hasibul15-11057@diu.edu.bd

ABSTRACT

The goal of this research is to determine the flexibility of eCommerce, since it has become an advantage for any country's economy in recent years. Flexibility is determined by a variety of factors, but on-time product delivery is critical for both the customer and the supplier. That's why we wanted to make a predictive model which will work through a full dataset of any online-based company and from the predictive result the authority will be able to take proper steps to improve the condition. In this paper, we developed a prediction model using machine learning algorithms and applying some techniques of data mining. We obtained the best results for the support vector machine with a prediction accuracy of 69%. The model was based on using data of more than 10000 delivery details of an eCommerce company and it was found that on-time delivery rates are not so high because of some factors (i.e., warehouse block, mode of shipment, product importance, etc.). And from the customer rating, it will be easier to acknowledge the improvements.

Keywords

eCommerce; delivery; data mining; shipment delay; customer rating; reached on-time; machine learning; prediction;

1. INTRODUCTION

Nowadays, the e-commerce business is increasing day by day. People are getting more attracted to e-commerce shopping because of the comfort of shopping. Online shopping has been growing over time. More consumers have begun trusting and getting attracted to online commerce. They have moved a significant part of their shopping online. That's why the competition of e-commerce business is increasing day by day. For e-commerce companies it becomes an important fact to fulfill the needs of customers such as delivering authentic products, hassle-free delivery and payment process, User-friendly website, etc. A customer also expects fast delivery.

But when a product is ordered then it goes through some process to reach the customer. That's why delay occurs often while delivering a product. So, it becomes an important fact to find the cause of the delay and the way of reducing those delays because online or offline, customer satisfaction is very important for any buying and selling company. That's why we have proposed in this paper to analyze what is the behavior or approach of the customers towards online shopping, what is the state of these delivery products, and exactly what can be the reasons for late delivery.

2. LITERATURE REVIEW

Ecommerce started to spread around the time between the early 2000s and the late 1990s. 18 papers were published between 2001 and 2015 (about 2 per year on average) (Mangiaracina, Perego, Seghezzi and Tumino, 2019).

E-commerce has a significant influence on any country's economy since it is a new way of conducting business that is both contemporary and quick (Seyal and Rahman, 2003). This represents the utilization of electronic devices in doing business (Choi et al., 1997) or the method of shopping for and commercialism product or services. The electronic exchange of data, commodities, services, and payments is known as electronic commerce (Harrington and Reed, 1996). It is a smart business process where one can do business by sitting at their home. Every aspect of this new system is very crucial and it is very competitive for the consumers to have a stable position (Sinha and Tanty, 2020). The competitive e-commerce sector depends heavily upon its delivery services for the product to reach a customer as delivery plays a fundamental role in enhancing eCommerce. The delivery process includes several types of intermediate processes and a single failure at any step of the intermediate processes can lead the customer dissatisfaction & failure of customer retention which then affects the consumers to shop with them (Sinha and Tanty,2020). So, it is necessary to know about these processes and identify the reasons. We tried to figure out the sectors of delayed shipping from a company's customer database. The rising congestion from the continued growth of container shipping and the higher frequency of extreme weather events has an effect on delayed deliveries which can arise in the future (Adrian and Stefan, 2020). But there are also some other reasons for this like service of the company, quality products such as if product importance is high and having high rating products are delivered timely or not, customer services, etc. which will be covered in further discussion.

3. METHODOLOGY

In order to study the state of eCommerce shipment, we went for analytical research where we extracted some authentic data of an international eCommerce company from Kaggle, an online community of data science and machine learning practitioners; and went by all the procedures of processing dataset. As a result, we found some tabular results which we converted to visual forms. The traditional data mining process was followed very strictly and we used Python for the process. We went for Data mining with a view to producing some analytical and visual results which will be helpful for further studies. As the whole process was done using Python, so we used libraries like Pandas, to analyze data; NumPy for the numerical values, Label Encoder for turning categorical features into numerical values, and Matplotlib and Seaborn for statistical data visualization. We also used StandardScalar, MinMaxScaler, DecisionTreeClassifier for distribution, classification and some feature extraction methods.

3.1 Dataset

The dataset contains more than 10,000 observations of 12 variables. All these variables are expressed in percentages (e.g.: the percentage of individuals who placed their latest orders). The Company have a big Warehouse which is divided into 5 blocks as A, B, C, D, F. The company has 3 product shipment process, they are Ship, Flight, and Road. Products of the company are categorized in 3 way-low, medium, and high. And there were some other variables used like- Customer care calls, Customer rating, Cost, Gender, Discount offer, Reached on Time etc.
Among these Reached on Time is our target variable.

```
<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      10999 non-null  int64
dtypes: int64(8), object(4)
memory usage: 1.0+ MB
```

Figure1: Summary of the dataset

So, we see we had 8 numerical variables and 4 categorical variables which we transformed into numerical values later for work. Then we had to check if there are any missing values in the dataset or garbage values because if there are some, we will have to remove them from the dataset and profile it thoroughly otherwise exacts result can't be calculated properly. So, we checked and didn't find any missing value which means our dataset is perfectly ok for work.

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

Figure2: Checking missing Values of the dataset

Then we applied the traditional data mining processes. Although the dataset is totally alright but we still need to apply some changes for working purpose. We need to verify data, analyze and extract them in a meaningful way. That was main data mining we used here.

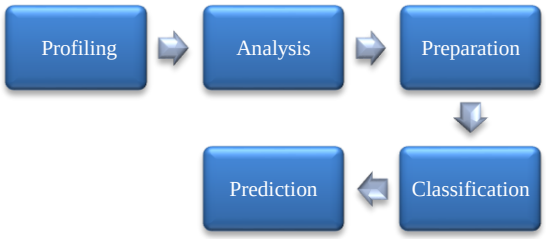

Figure3: Data Mining Techniques

3.1.1. Analysis:

Since our study mainly focused on on-time delivery of products, we tried to figure it out for the complete dataset at first. We found that there was almost 60% of deliveries that didn't reach on time.

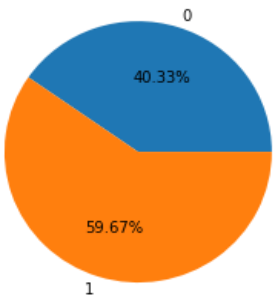

Figure4: Comparison of on-time and delay deliveries

We tried to categorize and analyze the reasons behind this inconvenience later on.

3.1.2. Preparation

Then there were 4 categorical variables in the dataset which we need to focus more for further application and research. So, to make the work easy and fluent we applied Chi square method to find the independence of these variables.
Chi Square test of independence:

We used a statistical hypothesis test to determine the relation among these categorical variables. In this case, we used the Chi-square test of independence. The formula is-

$$x^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$; Where

x^2 = chi squared

O_i = observed value

E_i = expected value

The following formula is employed to calculate the degrees of freedom for the chi-square: df = (r-1)(c-1); wherever r is that the range of rows and c is that the range of columns. The null hypothesis is rejected if the ascertained chi-square takes a look at datum is larger than the crucial price (usually zero.05).

Table1: Chi Square test result

Categorical Variables	Chi-square values
Warehouse Block	0.895
Mode of Shipment	0.689
Product Importance	0.002
Gender	0.636

From the chi-square test, we found the expected feature or value for the further process which was 'product importance' as the calculated value was greater than 0.05 for the rest of the three independent categorical features.

3.2 Working Process

We used here 2 models of machine learning for the comparison of best outcome. They are- the Random Forest classifier and Support Vector Machine (SVM) algorithm. We used 2 algorithms mainly to check which algorithm will be best for our work.

Random Forest classifier: The random forest could be a classification formula consisting of the many choices' trees. It uses fabric and have randomness once building every individual tree to do to make associate unrelated forest of trees whose prediction by committee is additional correct than that of any person tree. The formula for the random forest classifier is-

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(f_i - y_i)^2 \ ;$$

Where,

N = number of data points,

f_i = value returned by the model,

y_i = actual value for data point i.

Support Vector Machine: Image result for SVM formula "Support Vector Machine" (SVM) could be a supervised machine learning algorithmic rule which will be used for each classification and regression challenges. However, it's principally employed in classification issues. however, it's conjointly employed in knowledge prediction.

4. RESULT:

At the terribly initial, it's vital to search out and analyze the explanations behind delay shipping for that company. we have a tendency to found the subsequent results once the analyzation. Around sixty-eight of the delayed deliveries square measure caused once Ships square measure used as a mode of shipments. So, alternate choices like Flight, and Road services may be thought of to cut back the delayed deliveries.

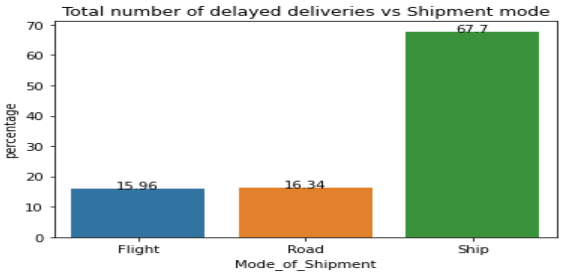

Figure5: Graphical representation of delay VS shipment mode

Higher percent of delayed deliveries were recorded in Warehouse block F. For rest of the block, the percent of delayed deliveries were almost same.

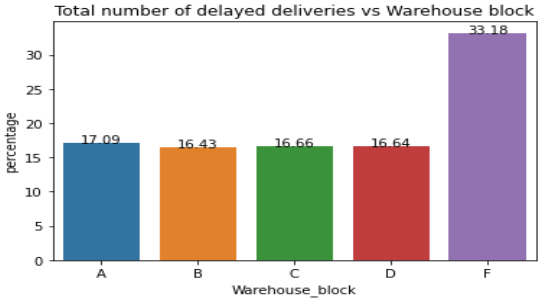

Figure6: Graphical representation of delayed delivaries VS Warehouse

We also found that more numbers of deliveries were also from warehouse block F so the delayed deliveries are

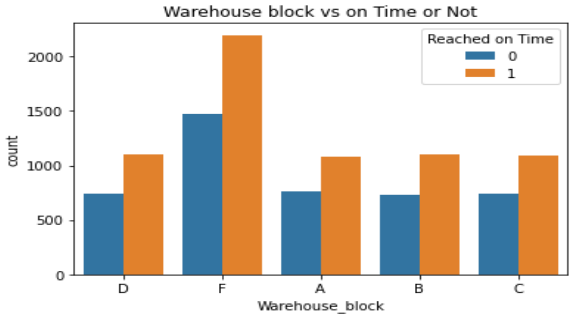

Figure6: Comparison of deliveries on-time or not from the warehouses

We found that the number of deliveries of the company was much more for their low importance products than medium and high importance products. So, the higher number of delay deliveries were also for low important products, a smaller number of high importance products were delivered but the delay also seemed to be less.

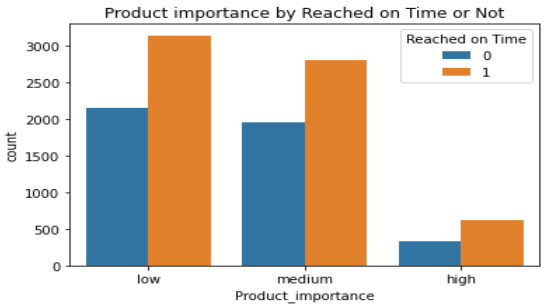

Figure6: Graphical representation of deliveries by product importance

In the company, the average cost of the products seemed not to be perfect. They have higher cost for the low important products where high important products are less costly. However, there was little variance overall.

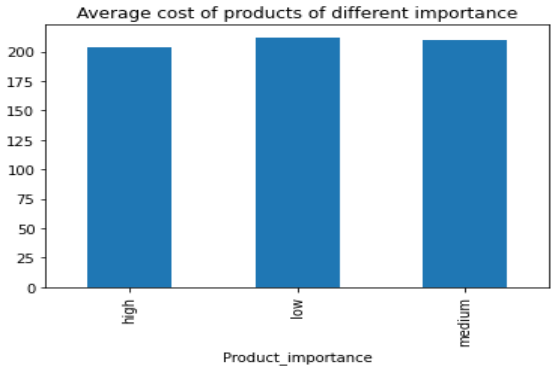

Figure7: Graphical representation of average cost of the products

Customer rating for the company is also seemed not so good. The company has been rated by every customer and the rating was defined on a scale of 1 to 5 where 1 is the lowest (worst) and 5 is the highest (best). The percentage of different ratings given by customers seems to be same. Almost 20% of the total deliveries received 5 ratings.

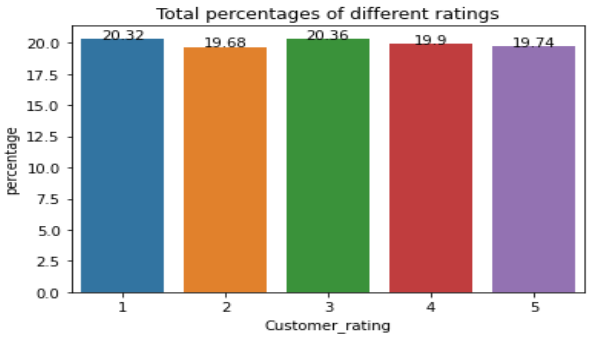

Figure8: Graphical representation of different customer ratings

Figure9: Comparison of Customer Rating by Gender

80% of data of the total dataset was kept for training the dataset and the remaining 20% was for testing the algorithms. The prediction result will be in two terms, 0 and 1, where 0 indicates the product has reached on time and 1 indicates that the product has not reached on time. The classification result was 65% for our dataset. And the prediction result we found, is shown below:

Table1: Prediction result of SVM Classifier

Class label	Precision	Recall	f-1 Score	Support
0	0.57	0.97	0.72	895
1	0.95	0.50	0.66	1305
Accuracy			0.69	2200

Table2: Prediction result of Random Forest Classifier

Class label	Precision	Recall	f-1 Score	Support
0	0.58	0.71	0.64	895
1	0.77	0.65	0.70	1305
Accuracy			0.68	2200

So, we see that the SVM algorithm is giving more accurate result than the Random Forest classifier and the classification result is around 65%.

5. DISCUSSION

The number of e-commerce business companies is increasing day by day. The success of a company depends upon how the company provides services to its customers. Customers always want a satisfactory level of service. And it depends on different parameters. Smooth and on-time delivery is one of them for an e-commerce business company. Company authorities often try to find out the problems of their service and rebuild or fix the problems. To find the cause of delivery problems, our model is the best one which will show and measure different facts related to delivering products by analyzing a full dataset of a company. Using this model an e-commerce company will be able to find out their delivery errors and will fix the problems.

6. ACKNOWLEDGMENTS

We would wish to convey our supervisor, faculty member Dr. Fizar Ahmed, whose experience was valuable in formulating the analysis queries and methodology. Your perceptive feedback pushed North American nation to sharpen our thinking and brought our work to the next level.

7. REFERENCES

[1] Han, J, and Kamber, M., Data Mining: Concepts and Techniques, San Francisco, CA: Morgan Kaufmann Publishers, 2001.

[2] Kohavi, R., Mason, L., Parekh, R., & Zheng, Z. (2004). Lessons and challenges from mining retail e-commerce data. Machine Learning, 57(1), 83-113.

[3] Ansari, S., Kohavi, R., Mason, L., & Zheng, Z. (2001). Integrating E-commerce and data mining: Architecture and challenges. In Proceedings of the IEEE International Conference on Data Mining (ICDM'2001). IEEE.

[4] Viellechner, A., & Spinler, S. (2020, January). Novel Data Analytics Meets Conventional Container Shipping: Predicting Delays by Comparing Various Machine Learning Algorithms. In Proceedings of the 53rd Hawaii International Conference on System Sciences.

[5] Mangiaracina, R., Perego, A., Seghezzi, A., & Tumino, A. (2019). Innovative solutions to increase last-mile delivery efficiency in B2C e-commerce: a literature review. International Journal of Physical Distribution & Logistics Management.

[6] Wei, L., Kapuscinski, R., & Jasin, S. (2020). Shipping Consolidation Across Two Warehouses with Delivery Deadline and Expedited Options for E-commerce and Omni-channel Retailers. Manufacturing & Service Operations Management.

[7] Kimball, R., Reeves, L., Ross, M., & Thornthwaite, W. (1998). The data warehouse lifecycle toolkit: Expert methods for designing, developing, and deploying data warehouses. John Wiley & Sons.

[8] Akter, S., & Wamba, S. F. (2016). Big data analytics in E-commerce: a systematic review and agenda for future research. Electronic Markets, 26(2), 173-194.

[9] SEGHEZZI, A. (2021). Innovative solutions to increase last-mile delivery efficiency in B2c e-commerce.

[10] Lin, Y. S., Zhang, Y., Lin, I. C., & Chang, C. J. (2018, March). Predicting logistics delivery demand with deep neural networks. In 2018 7th International Conference on Industrial Technology and Management (ICITM) (pp. 294-297). IEEE.

[11] Lee, S., Lee, S., & Park, Y. (2007). A prediction model for success of services in e-commerce using decision tree: E-customer's attitude towards online service. Expert Systems with Applications, 33(3), 572-581.

[12] Pratama, P. Y. ON TIME DELIVERY IMPROVEMENT AT PT. POS INDONESIA.

[13] Sinha, S. N. (2020). E-COMMERCE ADAPTABILITY WITH REFERENCE TO DELIVERY OF PRODUCTS. PalArch's Journal of Archaeology of Egypt/Egyptology, 17(9), 123-130.

[14] Choi, S. Y., Stahl, D. O., & Whinston, A. B. (1997). The economics of electronic commerce (p. 626). Indianapolis, IN: Macmillan Technical Publishing.

[15] J. Yu, G. Tang, X. Song, X. Yu, Y. Qi, D. Li, and Y. Zhang, "Ship arrival prediction and its value on daily container terminal operation," Ocean Engineering, vol. 157, pp. 73–86, 2018.

[16] NICHE, A. S. T. P. INTERNATIONAL, E-COMMERCE: A SOLUTION TO PENETRATING NICHE.