12

Predicting Shipment Delivery Time: The use of Machine Learning Algorithms for decision making in E-commerce Industry

Yunus Alade

Abstract—The purpose of this project is to create a classification model to predict whether a product can get to its destination on time. In real world setting, shipments getting to their destination depend on different factors, in this study, I am going to train a classification model using the determining factors to predict if a shipment will get to its destination on time or not.
The top 6 classification models will be used, and amongst them the most efficient was singled out as the best model for the task. The dataset used for the project contains 11 features and 1 result column. The data was explored to better understand it and visualize it do as to determine if they are fit for this project and are void of duplicate values, missing values, and outliers. Afterwards, the data was split into training and test dataset using the 80:20 quota. This project seeks to determine the effectiveness of Decision Tree, SVM, Random Forest, Logistic Regression, Naive Bayes, and K-Nearest Neighbor models, and their performance were evaluated to determine the best fit for the task.
The result of this project is a trained classification model that can predict whether a shipment will get to its destination on time or not based on the provided features. The performance of the model can be evaluated using various evaluation metrics such as accuracy, precision, recall, and F1 score. The goal is to have a model with high accuracy and precision in predicting on-time deliveries, which can be useful in improving logistics and supply chain management in the e-commerce industry. In this project various classification models were tested but among them, the model that brought back the best result is the Decision Tree Model, with Accuracy of 0.68, Precision of 0.8418657565415245, Recall of 0.5670498084291188, and F1 score of 0.6776556776556777.
Impact statement – Data, when properly analyzed can be used to make accurate predictions and can affect decision making, in the case of this project, e-commerce companies can use the predictions from their data to take precautions if it has been determined that a shipment will not reach its destination of time.

Index Terms—Data Science, Data, Prediction, Classification Model, paper, dataset.

1 Introduction

The movement of products and services across borders and continents is made possible thanks to the logistics and transportation sector. It is essential to have a trustworthy model that can predict the possibility that a shipment will arrive at its destination on time given the volume of shipments and the importance of on-time delivery, this help companies save cost across all production line and build customer loyalty.

The model will be trained on a dataset containing information about shipments, including the shipment's origin, destination, mode of transportation, and other relevant features. The project will explore different classification algorithms such as decision trees, logistic regression, naive Bayes, K-nearest neighbors, and support vector machines to determine the best-performing model. The performance of the selected model will be evaluated using different evaluation metrics such as accuracy, precision, recall, and F1-score. The results of this project will provide insights into the factors that contribute to timely shipment delivery and provide a reliable model that can be used to predict the likelihood of a shipment getting to its destination on time.

- *M. Shell was with the Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, 30332. E-mail: see http://www.michaelshell.org/contact.html*
- *J. Doe and J. Doe are with Anonymous University.*

Manuscript received April 19, 2005; revised August 26, 2015.

No company would want to risk falling behind as supply chain analytics develops quickly. This study will be a step forward in enhancing supply chain operations and maximising profitability by utilising the power of predictive analytics. Predictive analytics will likely become an even more crucial tool for supply chain organisations looking to stay competitive in a market that is changing quickly as the sector continues to expand and change.

2 Literature review/Backgroud

Global product shipments have increased significantly, which has had a huge impact on the e-commerce sector.

In order to meet customers' demands for quick and dependable delivery, businesses are increasing their investment in logistics. Product distribution has become an essential component of the industry. Predicting the timely delivery of supplies to their clients is one difficulty the sector encounters.

Customers' unhappiness and the company's reputation may suffer as a result of late deliveries. In the e-commerce sector, machine learning has become a potent tool for addressing this issue and enhancing the delivery procedure.

The use of machine learning algorithms to forecast cargo delivery times in the e-commerce sector has been explored in a number of research. For instance, a study by Tsai et al. (2019) developed a hybrid model to estimate the delivery

time of e-commerce shipments by combining the benefits of fuzzy logic and neural networks.

A dataset gathered from a significant Taiwanese e-commerce platform was used to test the model's performance, and the findings revealed an improvement in delivery time prediction accuracy. Another strategy based on machine learning was suggested in a different study by Lee et al. (2020) to forecast the delivery status of shipments. The suggested model predicted the delivery status of packages using a number of features, including shipping status, tracking data, and product information.

The model was developed and evaluated using data gathered from a significant South Korean e-commerce platform, and the findings demonstrated increased accuracy in forecasting delivery status.

Additionally, Wang et al. (2020) created a prediction model based on the predicted arrival time of flights for the delivery time of shipments. A dataset of more than 50,000 shipments from a significant Chinese e-commerce platform served as the testing ground for the suggested model, which was based on the random forest method. The outcomes demonstrated that the model considerably increased the predictability of delivery time.

In summary, the use of machine learning algorithms has proven effective in predicting the timely delivery of shipments in the e-commerce industry. The studies mentioned above show that machine learning can improve delivery time prediction accuracy, which is crucial for customer satisfaction and business success.

3 METHODOLOGY

The methodology section outlines the steps taken to accomplish the goals of this project. In this section, we will describe the data pre-processing techniques, feature selection and engineering, model selection, and performance evaluation methods used in building the classification model for predicting shipment delivery in the e-commerce industry. The methodology is designed to ensure that the model is trained using appropriate techniques and evaluated rigorously to determine its effectiveness in predicting shipment delivery.

3.1 Data collection

For this project, Dr Haider Raza has been so kind to provide me with a dataset that contains neccesary data to that are useful to create the calssification model needed for this project. The data contained 12 columns and 10999 rows. The columns include, ID, Warehouse block, Mode of Shipment, Customer care calls, Customer rating, Cost of the Product, Prior purchases, Product importance, Gender, Discount offered, Weight in gms, ROT Y. The data was recieved in CSV file format

3.2 Data preprocessing

The next step is to preprocess the by checking for missing values, encoding categorical variables using ordinal encoding, scaling numerical features, checking for outliers and redundant values. While preprocessing the data, histograms were plotted to observe for outliers, some outliers were noticed in "prior purchase" and "discount offered" columns but the histogram was not visually convincing enough so a boxplot was plotted to better visualise the outliers. A manual investigation was done on the data to confirm if the outliers are useful data or noise, and it was confirmed that although they are far from the mean, they were still probable.

3.3 Feature selection

This step involves analyzing the dataset to determine which features are most important for the classification problem, and remove any features that are not useful so as to eliminate noise that could lead to overfitting. In the case of this study, it was determined that the gender column adds no value to the model to be created simply because the gender of a customer does not affect the delivery of a shipment, hence, it is considered as noise. This process reduced the number of columns to 11.

3.4 Splitting the data

Splitting the data is an important step in machine learning projects because it allows you to train your model on one subset of the data, tune hyperparameters on another, and evaluate the final performance of the model on yet another. the data was split into two subsets: a training set, and a test set, labelled as X train, X test, y train, y test. The training set is used to fit the model to the data, , and the test set is used to evaluate the final performance of the model after it has been trained and tuned.

3.5 Model selection

For this project, the The decision tree algorithm is particularly suitable because it is a supervised learning algorithm that can be used for classification tasks, such as predicting whether a shipment will be delivered on time or not. Decision trees are easy to understand and interpret, and can handle both categorical and numerical data, which is important in this project where the data contains both types of variables. Additionally, decision trees can handle missing values, which is another common issue in real-world datasets. After evaluating the performance of carious classification model, the model that brought back the best result is the Decision Tree Model, with Accuracy of 0.68, Precision of 0.8418657565415245, Recall of 0.5670498084291188, and F1 score of 0.6776556776556777. making it the best among the bunch.

Model training: Various models were trained but the decision tree model was chosen using the code snipet:

$\mathrm{dt}_c lassifier = DecisionTreeClassifier(max_depth = 6)$

$\mathrm{dt}_c lassifier.fit(X_train, y_train)$

$\mathrm{y}_p red = dt_c lassifier.predict(X_test)$

from sklearn.metrics import $\mathrm{accuracy}_s core, precision_s core, recall_s core, f1_s core$

$\mathrm{accuracy} = \mathrm{accuracy}_s core(y_test, y_p red) precision = precision_s core(y_test, y_p red) recall = recall_s core(y_test, y_p red) f1 = f1_s core(y_test, y_p red)$

4 RESULTS

: After evaluating the performance of carious classification model, the model that brought back the best result is the Decision Tree Model, with Accuracy of 0.68, Precision of 0.8418657565415245, Recall of 0.5670498084291188, and F1 score of 0.6776556776556777. making it the best among the bunch.

5 CONCLUSION

The conclusion goes here.

6 CONCLUSION

The conclusion goes here.

APPENDIX A
PROOF OF THE FIRST ZONKLAR EQUATION

Appendix one text goes here.

APPENDIX B

Appendix two text goes here.

ACKNOWLEDGMENTS

The authors would like to thank...

REFERENCES

[1] H. Kopka and P. W. Daly, *A Guide to LaTeX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.

PLACE PHOTO HERE

Michael Shell Biography text here.

John Doe Biography text here.

Jane Doe Biography text here.