LATE DELIVERY RISK PREDICTION MODEL USING NEURAL NETWORK

LITERATURE SURVEY

The dataset used in the analysis was a supply chain dataset of a company called DataCo. The data contained transaction details of purchase orders of products such as clothing, sports, and electronic supplies.

In the analysis, a prediction model is built to analyse the risk associated with late deliveries. The insights drawn from the study can be used to identify the root cause for the risk and can be used to mitigate the risk factors.

The original dataset had 52 columns from which all of it was not required to build the model. So, a new subset called supply_chain_sub with relevant variables were created. It contained 14 columns which are:

ATTRIBUTES	DESCRIPTION
Type	Type of transaction made
Days for shipping (real)	Actual shipping days of the purchased product
Days for shipment (scheduled)	Days of scheduled delivery of the purchased product
Benefit per order	Earnings per order placed
Late_delivery_risk	Categorical variable that indicates if sending is late (1), it is not late (0).
Product Status	Status of the product stock: If it is 1 not available, 0 the product is available
Shipping date (DateOrders)	Exact date and time of shipment
Shipping Mode	The following shipping modes are presented: Standard Class, First Class, Second Class, Same Day
Order Id	Order code
order date (DateOrders)	Date on which the order is made
Order Item Discount	Order item discount value
Order Item Quantity	Number of products per order
Sales	Value in sales
Shipping date (DateOrders)	Exact date and time of shipment

Table 1. Attribute name and detail description of relevant supply chain subset

Apart from the above mentioned, following two attributes were derived:

"Shedule_Mismatch"

- 0 when "days of shipment (scheduled)" >= "days of shipment (real)"
- 1 when "days of shipment (scheduled)" < "days of shipment (real)"

"Order_to_Shipping_duration"

- Duration between the shipping date and order date in days.

CHOICE OF MODEL

- The artificial neural network (ANN) model was used to accurately predict the outcome i.e., the "late_delivery_risk" based on the predictor variables like type, shipping mode, etc.
- The data science pipeline of the solution will be as follows:
  - i. Extract and Load the data.
  - ii. Transform the data (i.e., Clean and enrich the data).
  - iii. Split the data into training and validation sets.
  - iv. Build an ANN model based on the training set.
  - v. Evaluate the efficiency of the model on the validation set.

DATA COLLECTION

- The DataCo supply chain dataset used for the ANN prediction was taken from Mendeley Data Link: https://data.mendeley.com/datasets/8gx2fvg2k6/5
- Each record in the DataCo supply chain dataset corresponds to transaction details of purchase orders.
- The target variable is called late_delivery_risk.

DATA CLEANING PROCESS

The data cleaning process was done by the following steps:

1. HANDLING MISSING VALUES

- First, we checked the number of null values in the dataset and found that the dataset had no null values. So, no further steps were required.

2. DUPLICATE DATA

- In the continuation of our data-cleaning process, we checked for duplicate rows in the subset of the supply chain dataset. Fortunately, we found that there were no duplicate values present.

3. LOW VARIATION DATA

- We also examined categorical variables to determine if any have only one unique value. For which we found that the categorical variable called “Product Status” had only one value, that was zero. Therefore, the variable was removed as it showed low variation in data.

4. IRRELEVANT DATA

To further refine the dataset, we focused on identifying and addressing irrelevant data related to the "Order Status" attribute. It contained data such as: cancelled, closed, complete, on hold, payment review, pending, pending_payment, processing, and suspended fraud. In all this only completed and closed were the only data related to the analysis. Inclusion of other irrelevant data will lead to computational complexity. So, other irrelevant data are eliminated to enhance the model accuracy. This approach ensures that our analysis is not influenced by rare and potentially biased data points.

5. HANDLING INCORRECT DATA

In this step, we focused on identifying and rectifying incorrect data present in the dataset. When calculating summary statistics for the “Days for shipping (real)” and “Days for shipment (scheduled)” variable, we found that certain values were not allowed, which was 0, suggesting that there may have been a data entry error. To correct this issue, records with “Days for shipping (real)” and “Days for shipment (scheduled)” value as 0 were replaced by mean value.

CHALLENGES FACED AND ADDRESSED

The challenge we encountered is to identify values that were clearly incorrect or outside the allowable range, as incorrect data can seriously affect the validity of our conclusions. By correcting unrealistic or unallowable values in the variables, as well as eliminating some zero values that should not exist, we effectively improved the accuracy of our dataset.

6. HANDLING CATEGORICAL DATA

- Encoding the categorical data involves converting categorical attributes into a numerical format, specifically the "Type" and "Shipping_mode" variables. These attributes contain non-numeric values, which need to be converted into a numerical format.

CHALLENGES FACED AND ADDRESSED

- "Type" only had 2 categorical values, which was simply replaced with 0 and 1, the challenge was with “Shipping_mode". We processed the unique values within the " Shipping_mode " attribute to create a consistent and meaningful encoding. By iterating through each unique value, we assigned a numerical code (incrementing by 1) to represent each " Shipping _mode." This encoding scheme maintained the categorical distinction enabling us to work with numerical representations.

7. HANDLING OUTLIERS

- We created box plots for each relevant numerical attribute to visually identify potential outliers. Based on the Interquartile Range (IQR) method (Gulati,2022), identified and returned a list of outliers and handled outliers by removing records with outlier with the mean of the respective attribute. This approach minimized the impact of outliers on analysis results.

CHALLENGES FACED AND ADDRESSED

- Outliers can arise for a variety of reasons, such as data entry errors or truly extreme observations. There are different ways to deal with different causes. We learned the IQR method to determine which numbers are outliers and what to do about them.

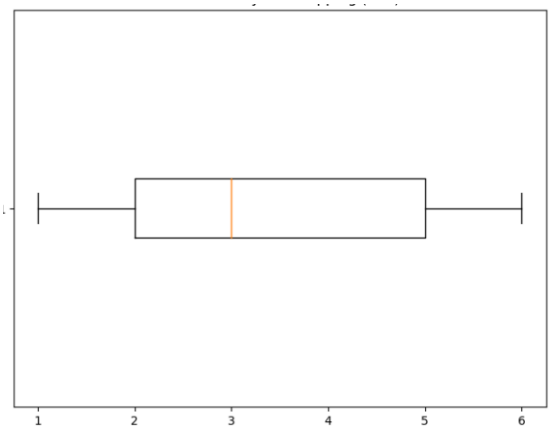
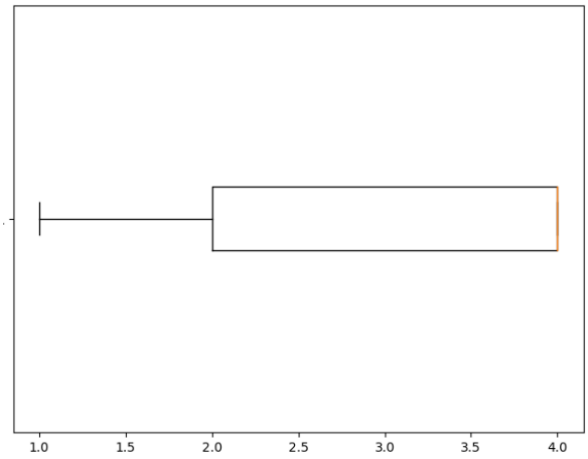


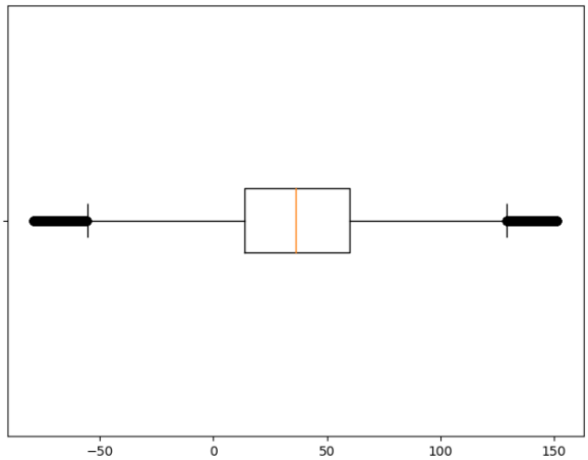
Fig 1: Box plot of Days for shipping (real)



attribute after outlier handling

Fig 2: Box plot of Days for shipping (scheduled) attribute after outlier handling

Fig 3: Box plot of Benefit per order attribute after outlier handling



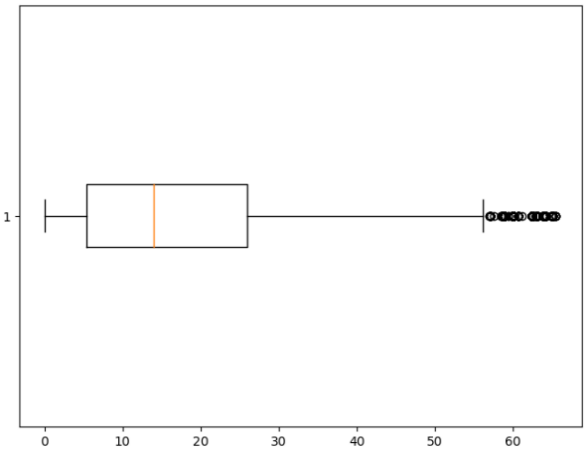


Fig 4: Box plot of Order Item Discount attribute after outlier handling

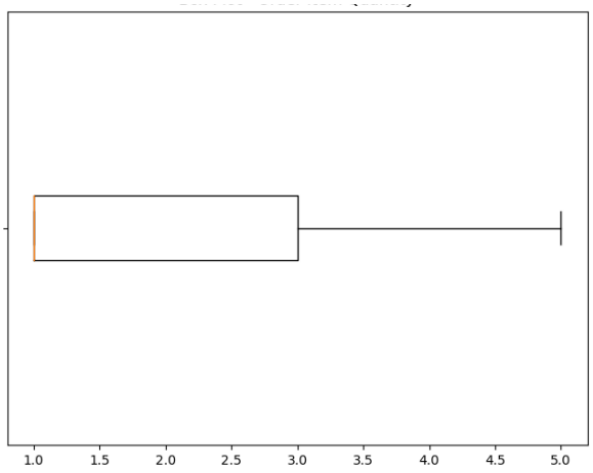


Fig 5: Box plot of Order Item Quantity attribute after outlier handling

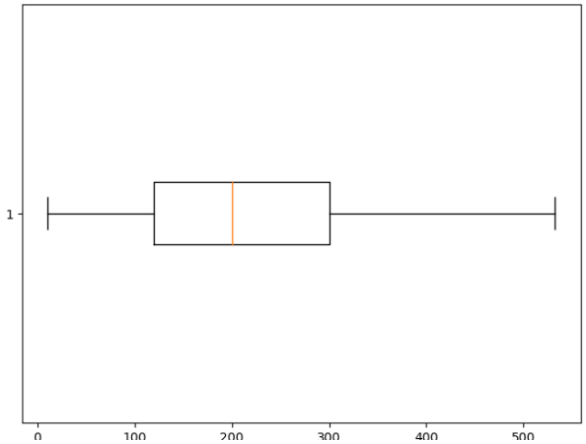


Fig 6: Box plot of Sales attribute after outlier handling

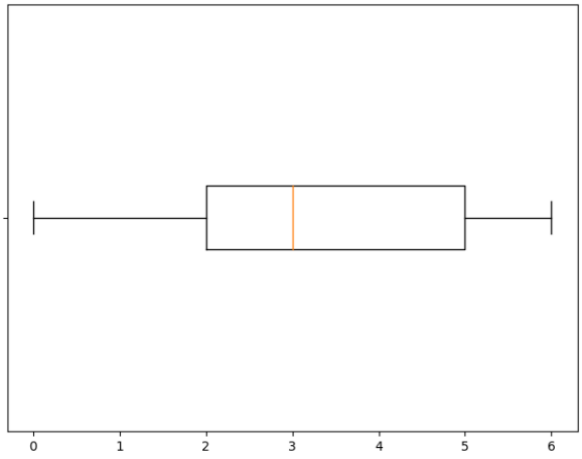


Fig 7: Box plot of Order_to_Shipping_duration attribute after outlier handling

EXPLORATORY DATA ANALYSIS & VALIDATION SPLIT

- After cleaning the data, we use a plot histogram to identify the distribution of values and prepare the data for ANN analysis by defining the predictor and outcome variables, and then split the data into a training set and a validation set.

MACHINE LEARNING MODEL DEVELOPMENT

- The solution to the business problem involves the prediction of a numerical outcome variable "late_delivery_risk" based on a set of supply chain predictor variables. The following machine learning algorithms are being deployed on the training data and their efficiencies are evaluated.
    - ➢ Artificial Neural Network

a. ARTIFICAL NEURAL NETWORK

- **SUITABILITY OF MODEL**: The aim is to predict a categorical variable based on

values of a list of categorical and numerical variables hence Artificial Neural Network seems to be an ideal model for the solution.

- The Artificial Neural Network prediction model is built using the training dataset and the model performance is evaluated against the validation dataset.
- Extreme values in the correlation matrix of predictor variables clearly indicate that there is multicollinearity among predictor variables and suitable predictor reduction techniques must be used to optimize the regression model.
- The regression models are built and evaluated using the different sets of number of hidden layer values and the optimal value is determined.

MODEL EVALUATION

- The model evaluation results for the Artificial Neural Network on model on the training set and validation set are as follows:

	Prediction	
Actual	0	1
0	20313	0
1	0	27151

Table 2. Confusion Matrix for results of model evaluation on training set

	Prediction	
Actual	0	1
0	13486	0
1	0	18157

Table 2. Confusion Matrix for results of model evaluation on training set

RESULT

The model achieved an accuracy of 100% on training and validation set with number of hidden layers as 3 hence the ANN model seems to be best fir for this application.

LIMITATION

Accuracy of 100% may be indication of overfitting. Therefore, there are chances that the model may perform poorly on test data however it can be confirmed only extensive testing.

REFERENCE

1. Kumar, Ajitesh. “Python - Replace Missing Values with Mean, Median & Mode.” Data Analytics, 26 Mar. 2023, https://vitalflux.com/pandas-impute-missing-values-mean-median-mode/#:~:text=Mean%20imputation%20is%20often%20used,to%20outliers%20than%20the%20mean

2. Cloud, Saturn. “How to Remove Entries Based on the Number of Occurrences Using Python Pandas | Saturn Cloud Blog.” Saturncloud.io, 19 June 2023, https://saturncloud.io/blog/how-to-remove-entries-based-on-the-number-of-occurrences-using-python-pandas/

3. Gulati, Aman Preet. “Dealing with Outliers Using the IQR Method.” *Analytics Vidhya*, 13 Sept. 2022, www.analyticsvidhya.com/blog/2022/09/dealing-with-outliers-using-the-iqr-method/