Shipment Prediction with Machine Learning

Instructor

Neha Ramchandani

Student

Omkar Bhagwat

Chaitravi Anagane

Yogita Mishra

Dr.Mininath Nighot

Project-P112

April 19, 2022

• Introduction

This research project focused on how to apply Machine Learning to make predictions regarding shipping on time, from an international e-commerce company to their customers. Using Random Forest and train/testing method with building more models to predict the shipping, we want to discover key insights from which factors significantly affect 'Reached on time' and predict whether a product will be delivered on time or not.

With the globalisation of trade, transit time reliability has become a critical point in the shipping industry as irregularities will lead to more delays further down the supply chain. The company that sells electronic products provides freight forwarding services to its clients, offering them a complete set of supply chain solutions for shipping their goods across the world.

Using Machine Learning computing, we developed a model capable to predict on time shipping. Our model delivers predictions with a 71% accuracy as early as the transport is booked. The model relies on historical data, for example, cost of the product, product's weight, prior purchase, discount… We researched and tested several Machine Learning algorithms in order to select the one that maximises the prediction accuracy score. This algorithm introduces a prediction component to the output by giving an estimated on time delivery to the customer for a shipment.

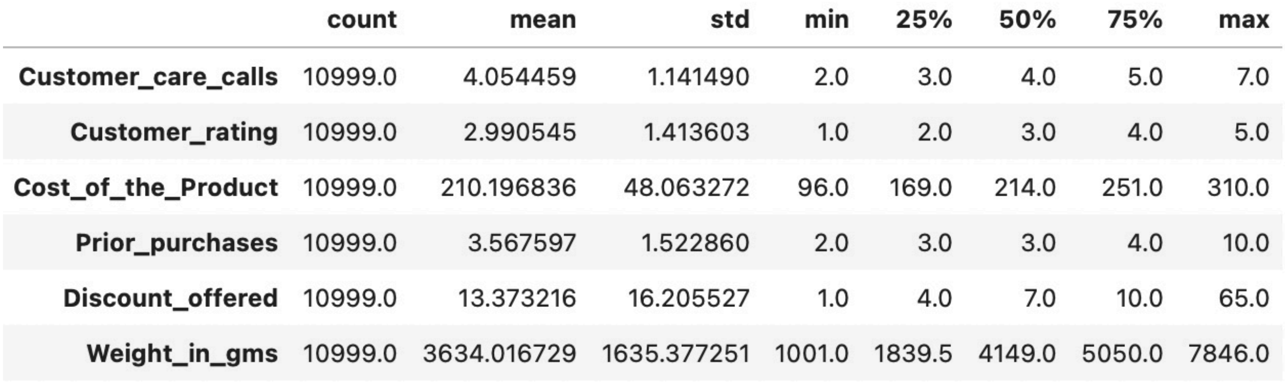
• Data summary

We have our data saved in a CSV file covering a total of 10,999 observation shipments. We first read our dataset into a pandas data frame called data, and then use the head() function to show the first five records from our dataset.

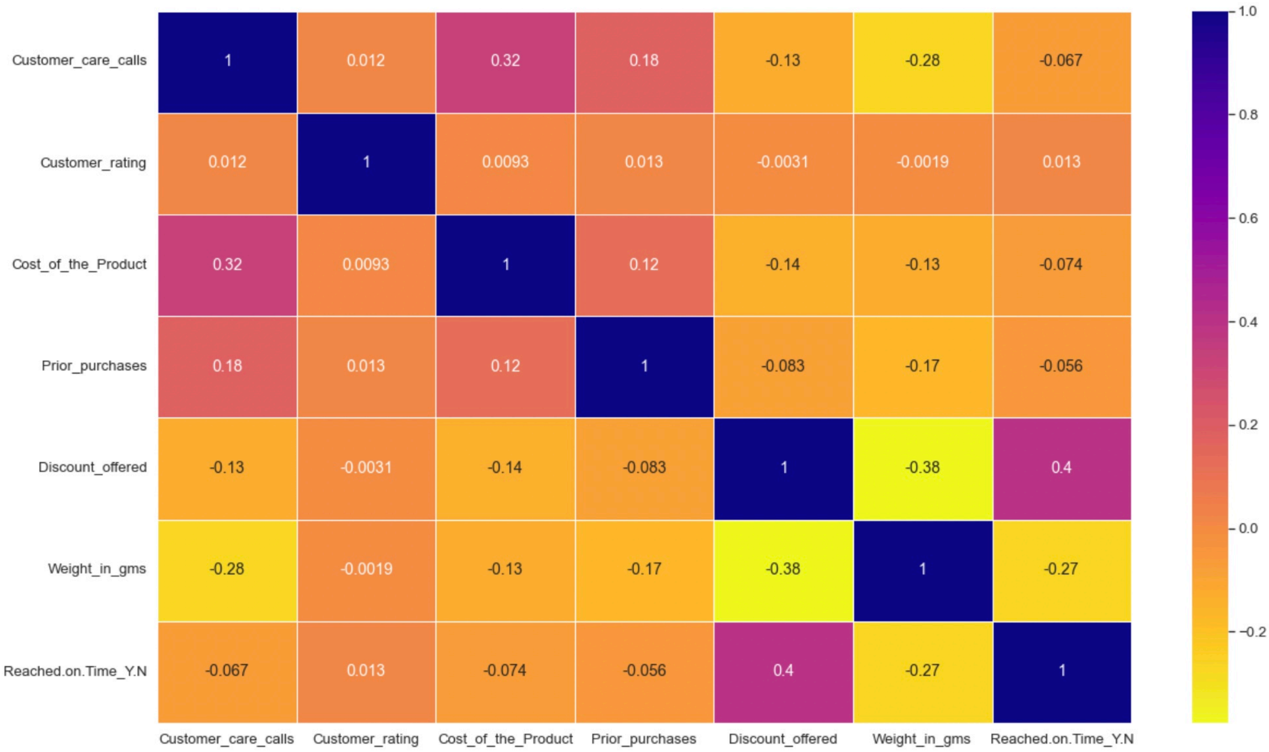
ID	ID Number of Customers.
Warehouse block	The Company have big Warehouse which is divided in to block such as A,B,C,D,F
Mode of shipment	The Company Ships the products in multiple way such as Ship, Flight and Road
Customer care calls	The number of calls made from enquiry for enquiry of the shipment
Customer rating	The company has rated from every customer. 1 is the lowest (Worst), 5 is the highest (Best)
Cost of the product	Cost of the Product in US Dollars
Prior purchases	The Number of Prior Purchase
Product importance	The company has categorized the product in the various parameter such as low, medium, high
Gender	Male and Female
Discount offered	Discount offered on that specific product
Weight in gms	It is the weight in grams
Reached on time	It is the target variable, where 1 indicates that the product has not reached on time and 0 indicates it has reached on time

The dataset contains 12 variables, we use ‘Reach on time’ as a target variable.

- **Statistical summary and Correlation matrix**



	count	mean	std	min	25%	50%	75%	max
Customer_care_calls	10999.0	4.054459	1.141490	2.0	3.0	4.0	5.0	7.0
Customer_rating	10999.0	2.990545	1.413603	1.0	2.0	3.0	4.0	5.0
Cost_of_the_Product	10999.0	210.196836	48.063272	96.0	169.0	214.0	251.0	310.0
Prior_purchases	10999.0	3.567597	1.522860	2.0	3.0	3.0	4.0	10.0
Discount_offered	10999.0	13.373216	16.205527	1.0	4.0	7.0	10.0	65.0
Weight_in_gms	10999.0	3634.016729	1635.377251	1001.0	1839.5	4149.0	5050.0	7846.0

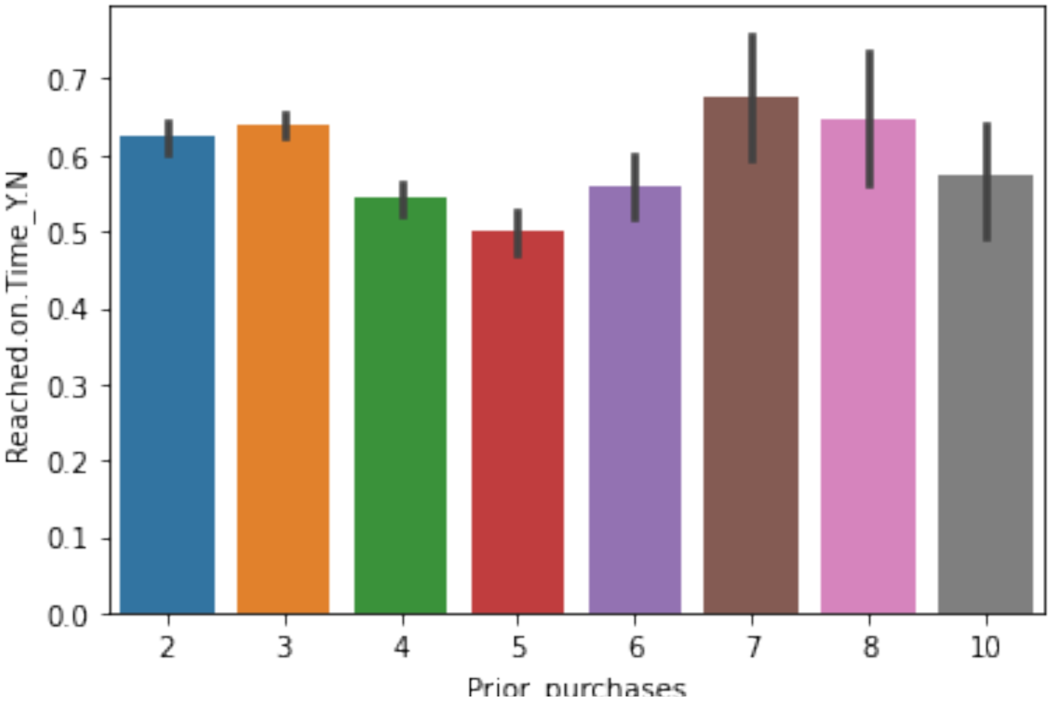


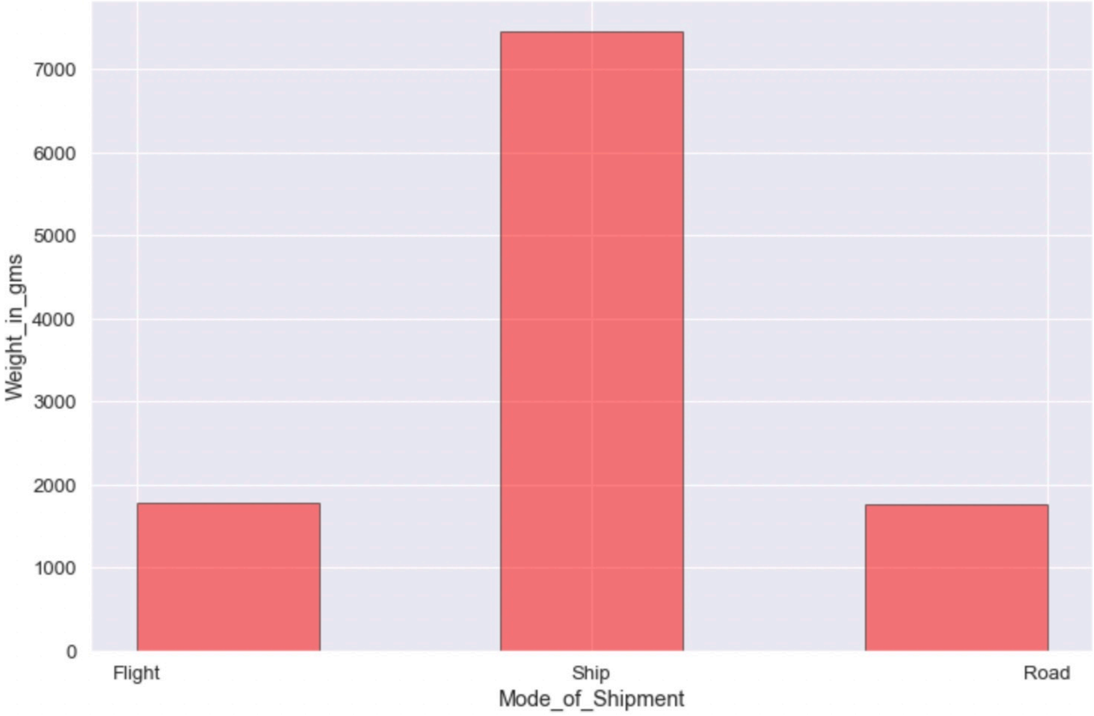
	Customer_care_calls	Customer_rating	Cost_of_the_Product	Prior_purchases	Discount_offered	Weight_in_gms	Reached.on.Time_Y.N
Customer_care_calls	1	0.012	0.32	0.18	-0.13	-0.28	-0.067
Customer_rating	0.012	1	0.0093	0.013	-0.0031	-0.0019	0.013
Cost_of_the_Product	0.32	0.0093	1	0.12	-0.14	-0.13	-0.074
Prior_purchases	0.18	0.013	0.12	1	-0.083	-0.17	-0.056
Discount_offered	-0.13	-0.0031	-0.14	-0.083	1	-0.38	0.4
Weight_in_gms	-0.28	-0.0019	-0.13	-0.17	-0.38	1	-0.27
Reached.on.Time_Y.N	-0.067	0.013	-0.074	-0.056	0.4	-0.27	1

Correlation matrix:

By using codes, we have found out the results for correlation matrix. data.corr () is being used for more information. By correlation matrix, we can find the correlation of every pair of features (and the outcome variable), and visualise the correlations using a heatmap.

- **Relation Between Prior purchases & Reached on time:**

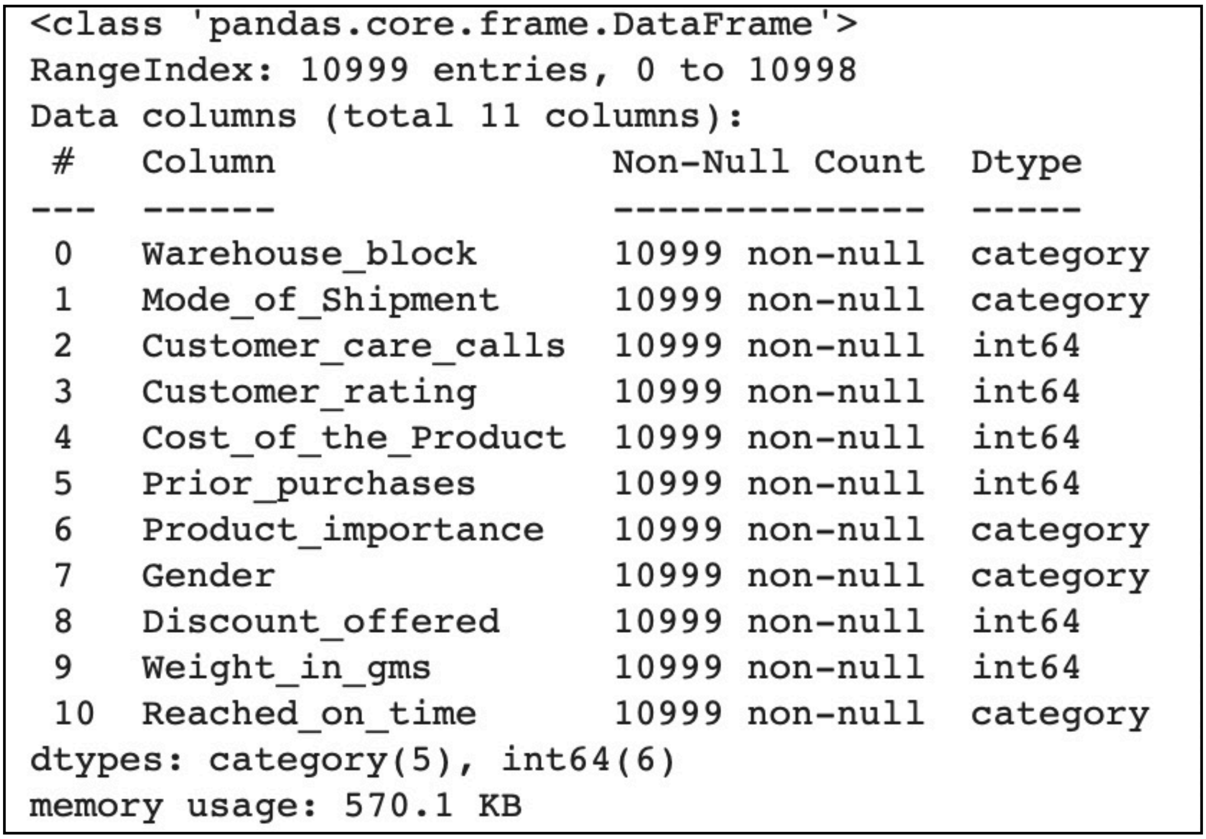




The above histogram is based on the mode of the shipment. According to the Figure Three modes of shipment are being used, i.e., Flight, Ship, Road. The Ship is the largest in use to transport goods from one place to another

- **Data cleaning process**

The data is clean (has no null value) so we don’t need to do the cleaning process.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10999 entries, 0 to 10998
Data columns (total 11 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   Warehouse_block      10999 non-null  category
 1   Mode_of_Shipment     10999 non-null  category
 2   Customer_care_calls  10999 non-null  int64
 3   Customer_rating      10999 non-null  int64
 4   Cost_of_the_Product  10999 non-null  int64
 5   Prior_purchases      10999 non-null  int64
 6   Product_importance   10999 non-null  category
 7   Gender               10999 non-null  category
 8   Discount_offered     10999 non-null  int64
 9   Weight_in_gms        10999 non-null  int64
 10  Reached_on_time      10999 non-null  category
dtypes: category(5), int64(6)
memory usage: 570.1 KB
```

To answer the research question about how to use Machine Learning to predict whether a shipment will be delivered on time or not, we apply the training/testing method with more models but here we are taking Logistic Regression, Random Forest, and SVM. After running the models, we will choose and pick the best model base on the accuracy score.

We selected statistically significant variables to tie our model. To test and validate the performance of the different algorithms, we split the data in a training set and a test or validation set. All algorithms were trained and tested on the same training and test set so that we could compare their performance.

We used Python programming language to write the code for our models. To create, train and use the final models for prediction, there are three relevant scripts in the final program. The first script is used to clean the data, remove missing values and transform all columns to the right format, our dataset has no missing value, so we don’t do this part. The second script automatically reads in all the relevant data, trains (on new data or adding data to an existing model), and saves the final model. The third script is used for prediction in a production environment.

To predict on time shipping, we apply on three model’s Logistic regression, Random Forest, SVM.

1. Logistic Regression:

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, Logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval, or ratio-level independent variables.

Accuracy Score for LR model: **63.56%**

2. Random Forest:

Random forest is a commonly-used machine learning algorithm, which combines the output of multiple decision trees to reach a single result. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree. Its ease of use and flexibility have fueled its adoption, as it handles both classification and regression problems.

Accuracy Score for Random Forest model: **71.87%**

3. SVM:

A support vector machine (SVM) is a supervised machine learning model that uses classification algorithms for two-group classification problems. After giving an SVM model set of labeled training data for each category, they can categorize new text. Appendix J describes the code in detail we have used to fit the model.

Accuracy Score for SVM model: **65.49%**

• Model selection

Sr. no	Model	Accuracy Score
1	Logistic Regression	64%
2	**Random Forest**	**72%**
3	SVM	65%

Compare three models, we see that Random Forest model has the highest accuracy score. We will choose this model to predict the on time shipping on our dataset.

- # Conclusion and future works

1. Results and Analysis

*Relationship between feature variables and target variables.

Cost of the product, Prior purchases, Discount offered, Weight in gram and Product important function have a significant influence on the model, especially Cost of the product and Prior purchases. It is good to see our machine learning model match what we have been assuming.

Cost of the product has a negative influence on the prediction, higher Cost of the product is correlated with a not on time shipment, higher on Cost of the product, less on not on time shipment.

- Predict on time shipment:

After comparing the results yielded by three different algorithms (Random Forest, Logistic Regression, and SVM), we decided to build the final prediction model with a Random Forest algorithm. We selected the set of random forest model as they result in the highest accuracy score (71%) depending on the model segments.

2. Limitations of our approach

Machine Learning is a method that is rather data-intensive in two ways: It requires a lot of data in order to kick-start the analysis and modelling and it requires that this data be of at least moderate and at best great quality. For great quality to be achieved, this means there should be no missing or wrong data points in the dataset, as well as consistent and useable formatting of the data. If not, the length of the data cleaning part of the project might be considerably extended, which could pose a timing problem, especially if the project is bound in time.

Furthermore, developing such a model demands analytics and coding skills. These two skills, even if required, are not enough: having subject-matter experts providing input on the industry practices and interpreting results and data is crucial to the success of such an endeavour.

Looking further down the road, since the end goal of a project like ours is to supply a functioning tool, this means the company will have to work on the models to make them a part of their computing environment. For example, the models we are delivering draw their data from several Excel files but in the future, we can imagine a direct link to a database instead.

3. Potential future works

In this study, we limited our scope to electronic products, but the same models developed for this purpose could be used for any products, given that the models are trained and supplied with the appropriate data. Consequently, the relevant factors we have identified as impactful for predicting on time shipping, such as cost of the product or prior purchases, in these models are relevant for any other products.

Furthermore, another insight of our study is that our approach could be applied to any shipping industry. We could envision a similar project being conducted for the other industry for example. As long as the necessary data is available and the impactful factors can be identified, the method can be adapted for any field of industry.

In the future, it would be worth investigating the possibility of including weather patterns in the model to improve its accuracy, if access to reliable data can be guaranteed. Another way to improve the model would be to know the area of customers. This is the area of the customers where products will be delivered to. We are confident that this would increase the accuracy of the model even more.