Predicting on-time shipping with Machine Learning

FINC 514 Introduction to Data Science Spring 2021

Professor: Xiaodi Zhu

Students: Tiyanna Jenkins

Erika Leonida

Thi Diem My Nguyen

Anchal Patel

Outline

1. Introduction
2. Data summary
3. Data cleaning process
4. Methodology & Results
5. Model selection
6. Conclusion and future works
7. Q&A

1. Introduction

- Project focused on how to apply Machine Learning to make predictions regarding shipping on-time, from an international e-commerce company to their customers.
- Researched and tested several Machine Learning algorithms in order to select the one that maximizes the prediction accuracy score.
- This algorithm introduces a prediction component to the output by giving an estimated on-time delivering to the customer for a shipment.

1. Introduction

Using Machine Learning to solve the business questions:

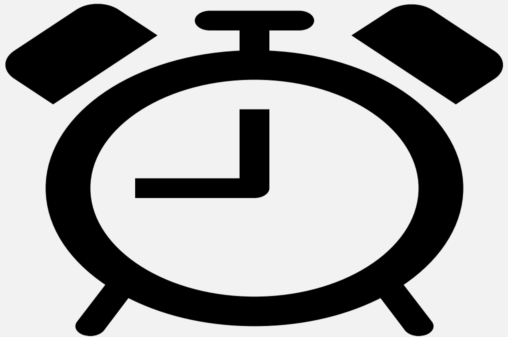
- Discover key insights to figure which factors affected a product reaching on time
- Using logistic regression and train/testing method to build 3 models
- Select the one that maximizes the prediction accuracy score

1. Introduction

Motivation for doing this project

Ensuring on-time delivery to customer indicates:

- The company is doing well in satisfying customers
- The company has direct impact of increased revenues. It is a good starting point would be to have statistical analyses of inside every shipment success and predict accurately about the shipping time



1. Introduction:

The specific data science question we want to analyze

Which factor(s) significantly affect Reached on time?

Predict whether the product will deliver on-time or not?

Type of analysis and model that plan to use for each data science question

For the data science question:

'Which factor(s) significantly affect Reached on time' by using similar data for whether delivered on time.

Type of analysis: Regression Analysis

Model: A Logistic Regression Model

For the data science question:

'Predict whether the product will deliver on-time or not'

Type of analysis: use the training/testing method

Model: Logistic Regression, Decision Tree, and SVM; Run the model and pick the best model

2. Data summary

a. Data source

Data saved in a CSV file covering a total of 10,999 observation shipments, called `Train(2).csv`

First five records from our dataset:

```
[ ] dat.head()
```

	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	Reached.on.Time_Y.N
0	1	D	Flight	4	2	177	3	low	F	44	1233	1
1	2	F	Flight	4	5	216	2	low	M	59	3088	1
2	3	A	Flight	2	2	183	4	low	M	48	3374	1
3	4	B	Flight	3	3	176	4	medium	M	10	1177	1
4	5	C	Flight	2	2	184	3	medium	F	46	2484	1

2. Data summary

Target variable: Reached On Time (1/0)

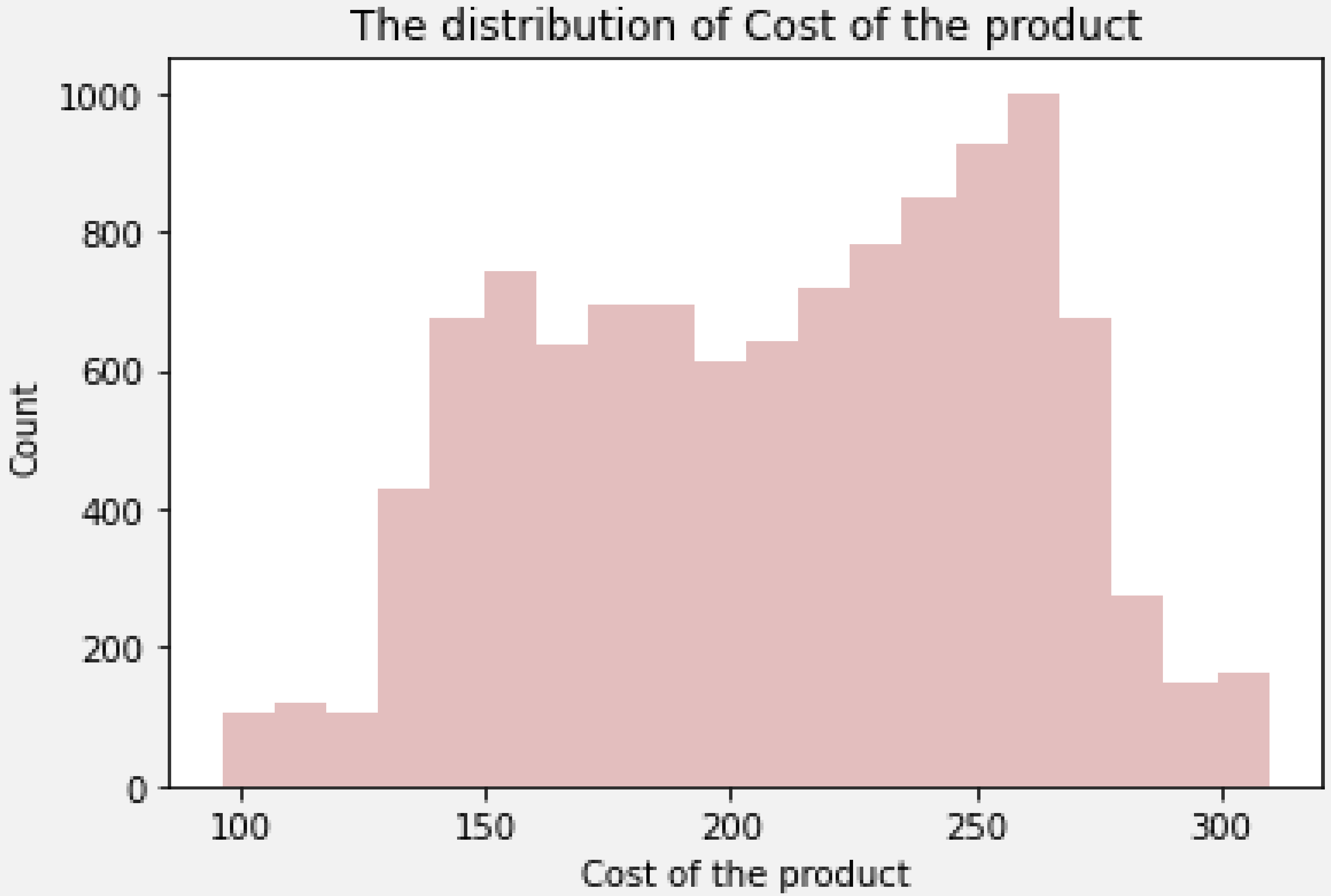
Featured Variables:

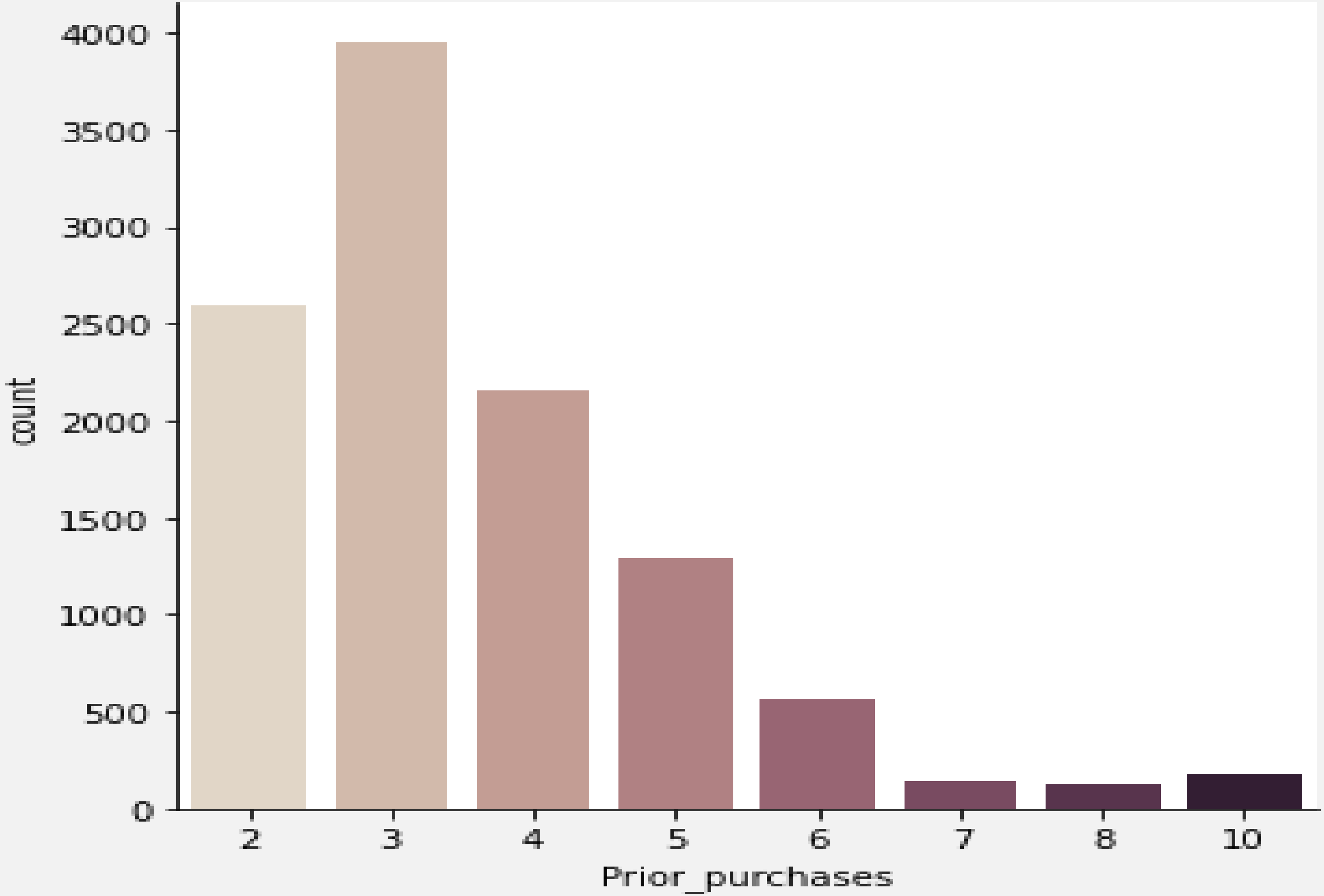
- Warehouse blocks
- Mode of shipment
- Cost of the product
- Prior purchase
- Product importance
- Gender
- Weight in grams.

2. Data summary

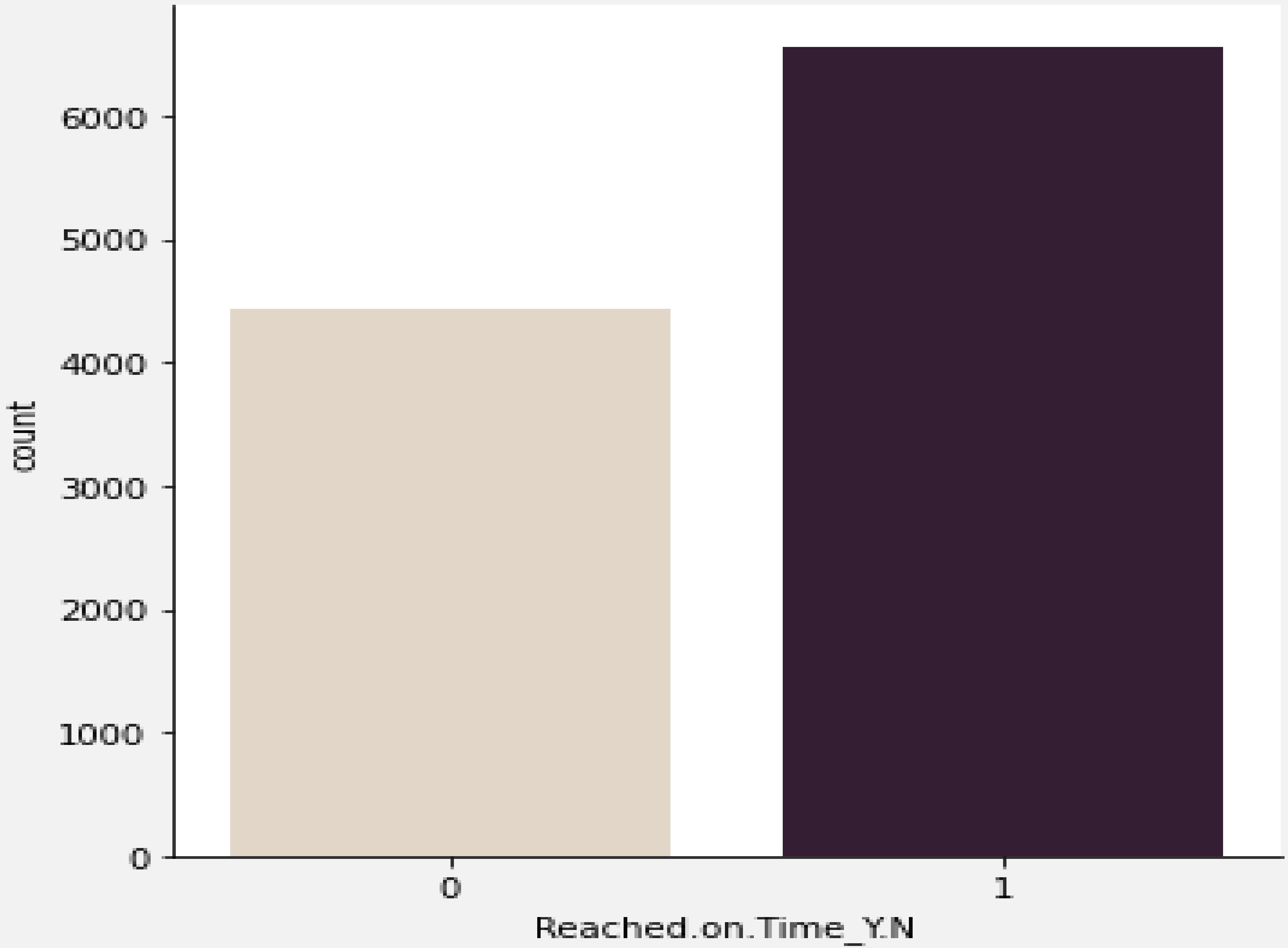
b. Statistical summary and Correlation matrix

Variable	Cost of the product	Discount offered	Weight in gms
Observation	10,999	10,999	10,999
Mean	210	13	3,634
Median	214	7	4,149
Max	310	65	7,846
Min	96	1	1,001
Standard			


The distribution of Cost of the product
Count
0
200
400
600
800
1000
100
150
200
250
300
Cost of the product

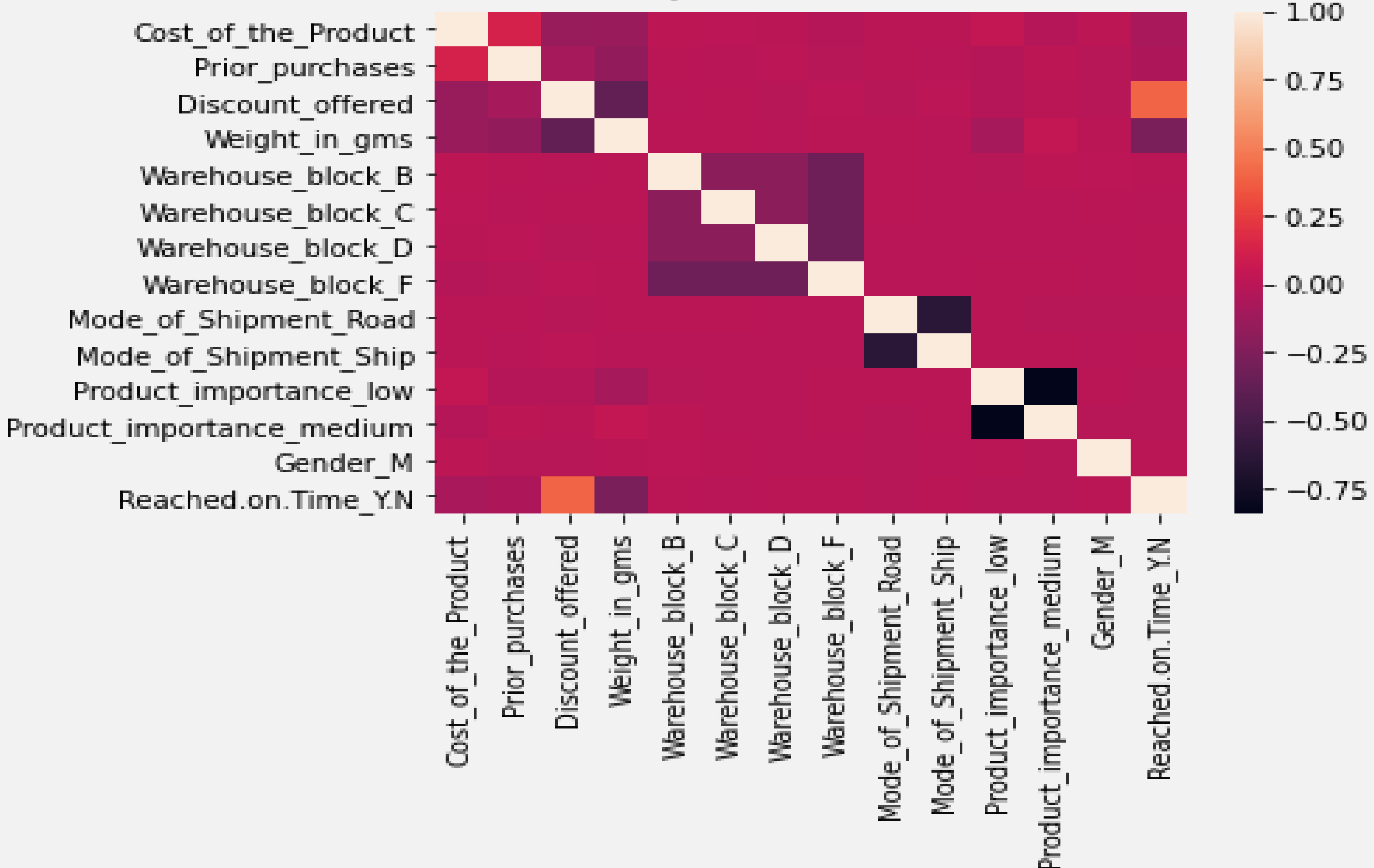


count
4000
3500
3000
2500
2000
1500
1000
500
0
2
3
4
5
6
7
8
10
Prior_purchases

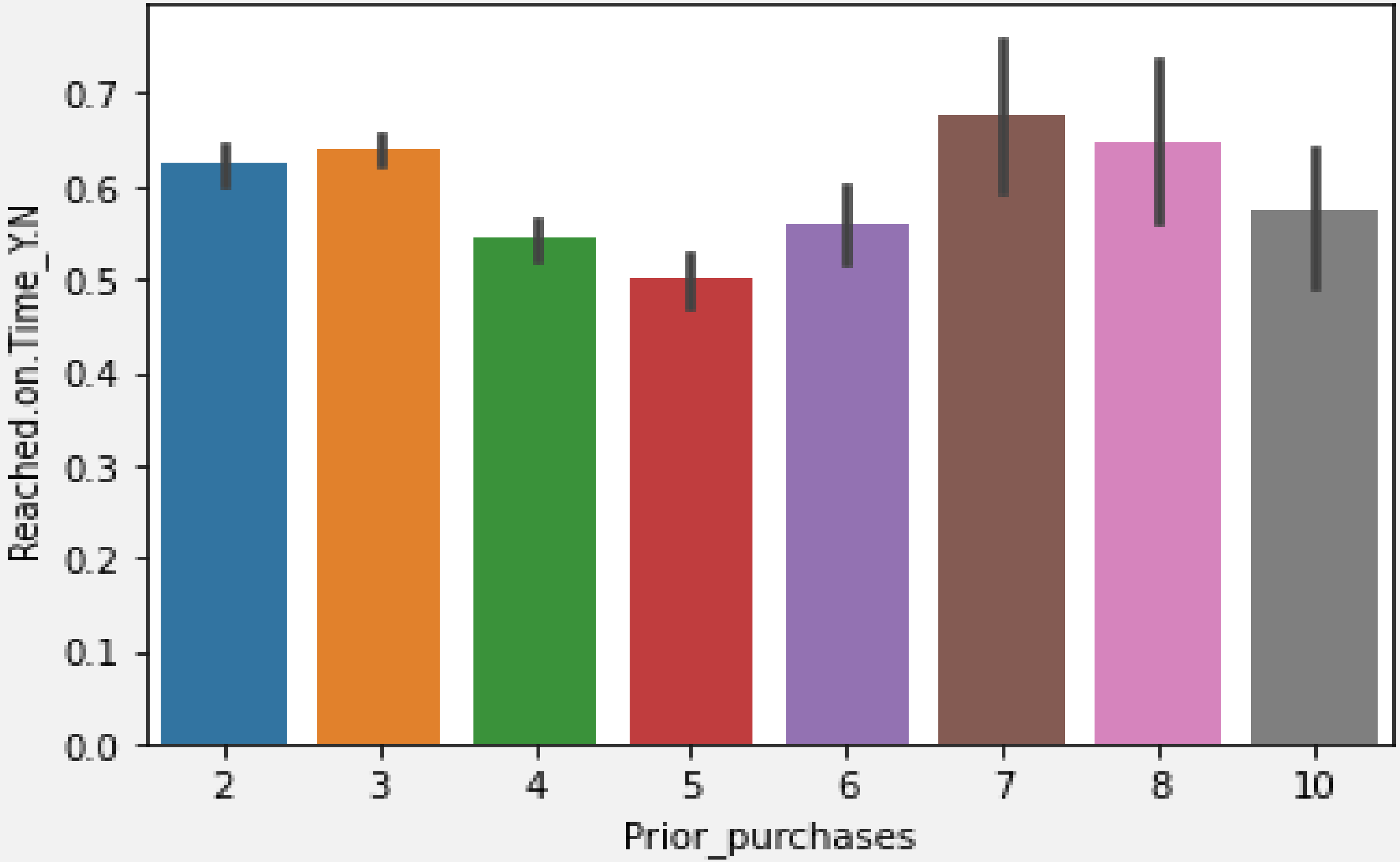

6000
5000
4000
3000
2000
1000
0
count
0
1
Reached.on.Time_Y.N

2. Data summary

Visualize the correlations using a heatmap



2. Data summary
Relation between Prior purchases and Reached on time.



3. Data cleaning process

- The data has no null value so we pass the cleaning process.

```
subdat.info()
#There are no missing values
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10999 entries, 0 to 10998
Data columns (total 9 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   Warehouse_block      10999 non-null  object
 1   Mode_of_Shipment     10999 non-null  object
 2   Cost_of_the_Product  10999 non-null  int64
 3   Prior_purchases      10999 non-null  int64
 4   Product_importance   10999 non-null  object
 5   Gender               10999 non-null  object
 6   Discount_offered     10999 non-null  int64
 7   Weight_in_gms        10999 non-null  int64
 8   Reached.on.Time_Y.N  10999 non-null  int64
dtypes: int64(5), object(4)
memory usage: 773.5+ KB
```

4. Methodology & Results

To answer the research question about which factor(s) significantly affect reach on-time:

- Used Logistic Regression model
- Reached on time as a dependent variable Y
- Warehouse blocks, mode of shipment, cost of the product, prior purchases, product importance, gender, and weight in grams as independent variables X

```
X_dat = sm.add_constant(X_dat)
logit = sm.Logit(y_dat, X_dat)
logit.fit().summary()
```

The outcome allows to estimate the parameters of the logistic regression model, the values of which are presented in table.

```
Current function value: 0.547044
Iterations 8
```

Logit Regression Results

Dep. Variable:	Reached.on.Time_Y.N	**No. Observations:**	10999
Model:	Logit	**Df Residuals:**	10985
Method:	MLE	**Df Model:**	13
Date:	Wed, 05 May 2021	**Pseudo R-squ.:**	0.1888
Time:	00:57:38	**Log-Likelihood:**	-6016.9
converged:	True	**LL-Null:**	-7417.0
Covariance Type:	nonrobust	**LLR p-value:**	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	1.2915	0.191	6.756	0.000	0.917	1.666
Cost_of_the_Product	-0.0026	0.000	-5.267	0.000	-0.004	-0.002
Prior_purchases	-0.0823	0.015	-5.439	0.000	-0.112	-0.053
Discount_offered	0.1138	0.004	25.664	0.000	0.105	0.122
Weight_in_gms	-0.0002	1.52e-05	-14.107	0.000	-0.000	-0.000
Warehouse_block_B	0.0852	0.075	1.132	0.258	-0.062	0.233
Warehouse_block_C	0.0522	0.075	0.693	0.488	-0.095	0.200
Warehouse_block_D	0.0621	0.075	0.826	0.409	-0.085	0.209
Warehouse_block_F	0.0422	0.065	0.646	0.518	-0.086	0.170
Mode_of_Shipment_Road	-0.0322	0.077	-0.420	0.674	-0.182	0.118
Mode_of_Shipment_Ship	-0.0150	0.060	-0.249	0.803	-0.133	0.103
Product_importance_low	-0.3566	0.084	-4.268	0.000	-0.520	-0.193
Product_importance_medium	-0.3417	0.084	-4.076	0.000	-0.506	-0.177
Gender_M	0.0528	0.044	1.211	0.226	-0.033	0.138

Variable	Coefficient	P-value
Cost_of_the_Product	-0.0026	.000***
Prior_purchases	-0.0823	.000***
Discount_offered	0.1138	.000***
Weight_in_gms	-0.0002	.000***
Warehouse_block_B	0.0852	0.258
Warehouse_block_C	0.0522	0.488
Warehouse_block_D	0.0621	0.409
Warehouse_block_F	0.0422	0.518
Mode_of_Shipment_Road	-0.0322	0.674
Mode_of_Shipment_Ship	-0.015	0.803
Product_importance_low	-0.3566	.000***
Product_importance_medium	-0.3417	.000***
Gender_M	0.0528	0.226

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Parameters of the logistic regression model and their assessment:

+ Parameters turned out to be statistically significant:

- Cost of the product
- Prior purchase
- Discount offered
- Weight in gms
- Product importance

+ Parameters didn’t turn out to be statistically significant:

- Warehouse block
- Mode of shipment
- Gender

4. Methodology & Results

To answer the research question about how to use Machine Learning to predict whether a shipment will be delivered on-time or not

- Apply the training/testing method with three models Logistic Regression, Decision Tree, and SVM
- Pick the best model base on the accuracy score
- Selected statistically significant variables to tie our model
- Algorithms were trained and tested on the same training and test set so that we could compare their performance

Apply Logistics regression model

Apply Logistics regression model

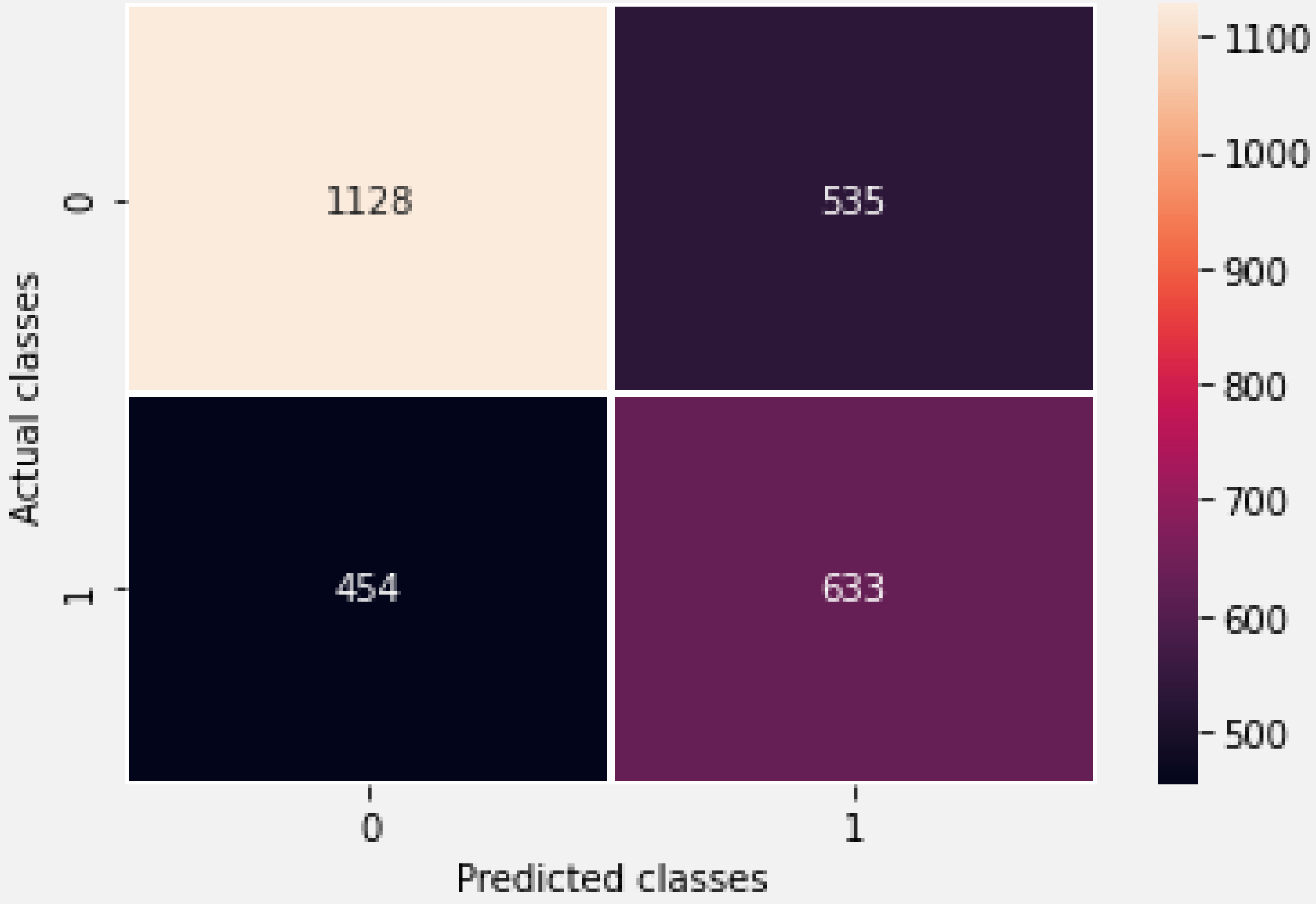
- About logistic regression: Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary),
- Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal,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

The code used to fit the model:

```
lr = LogisticRegression()
lr.fit(trainX, trainY)
```

Apply Logistics regression model

Accuracy Score for LR model: 0.6403636363636364. The model makes 64% accuracy rate on testing dataset.



Apply Decision Tree model

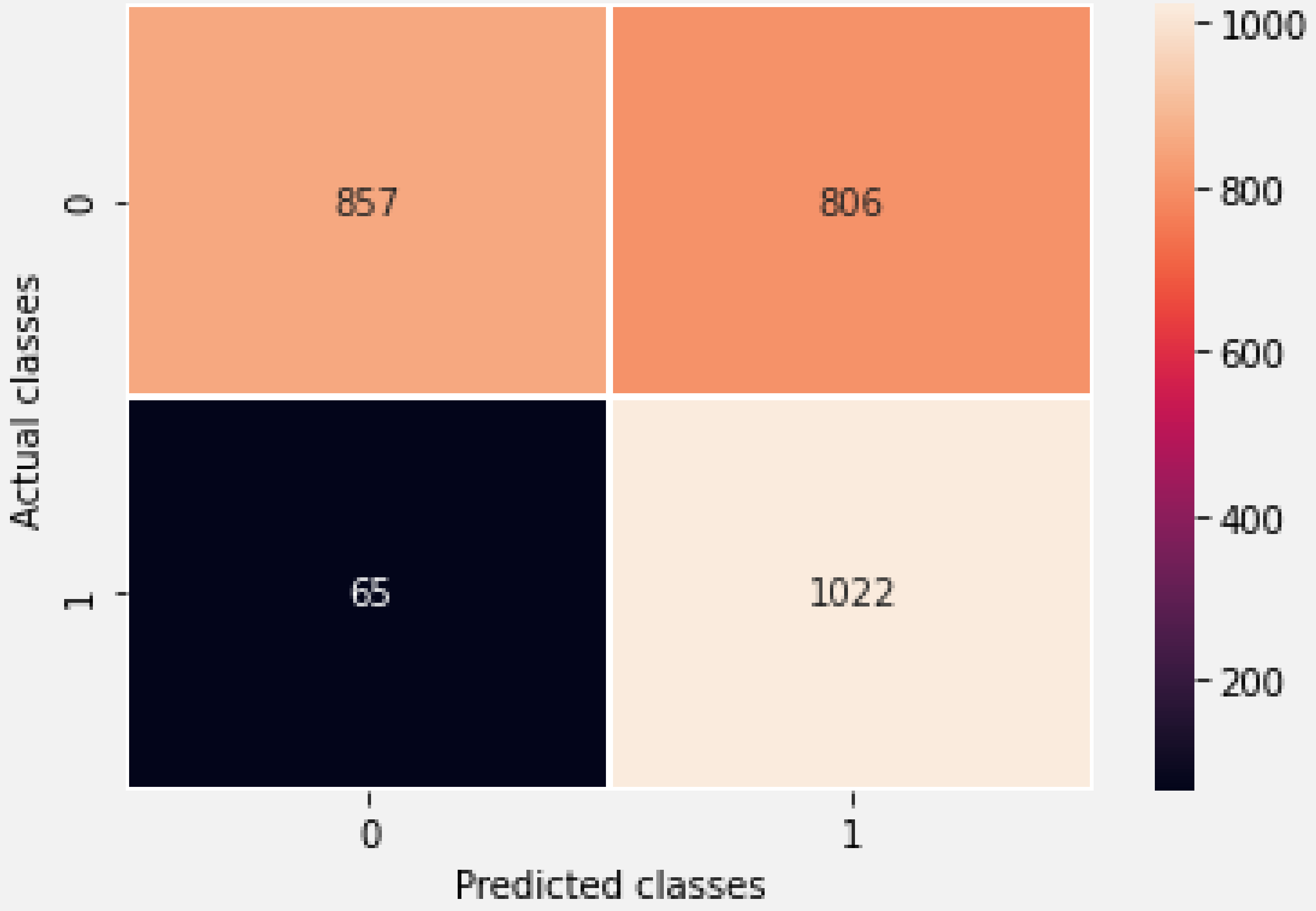
A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences including

- Chance event outcomes
- Resource costs
- Utility

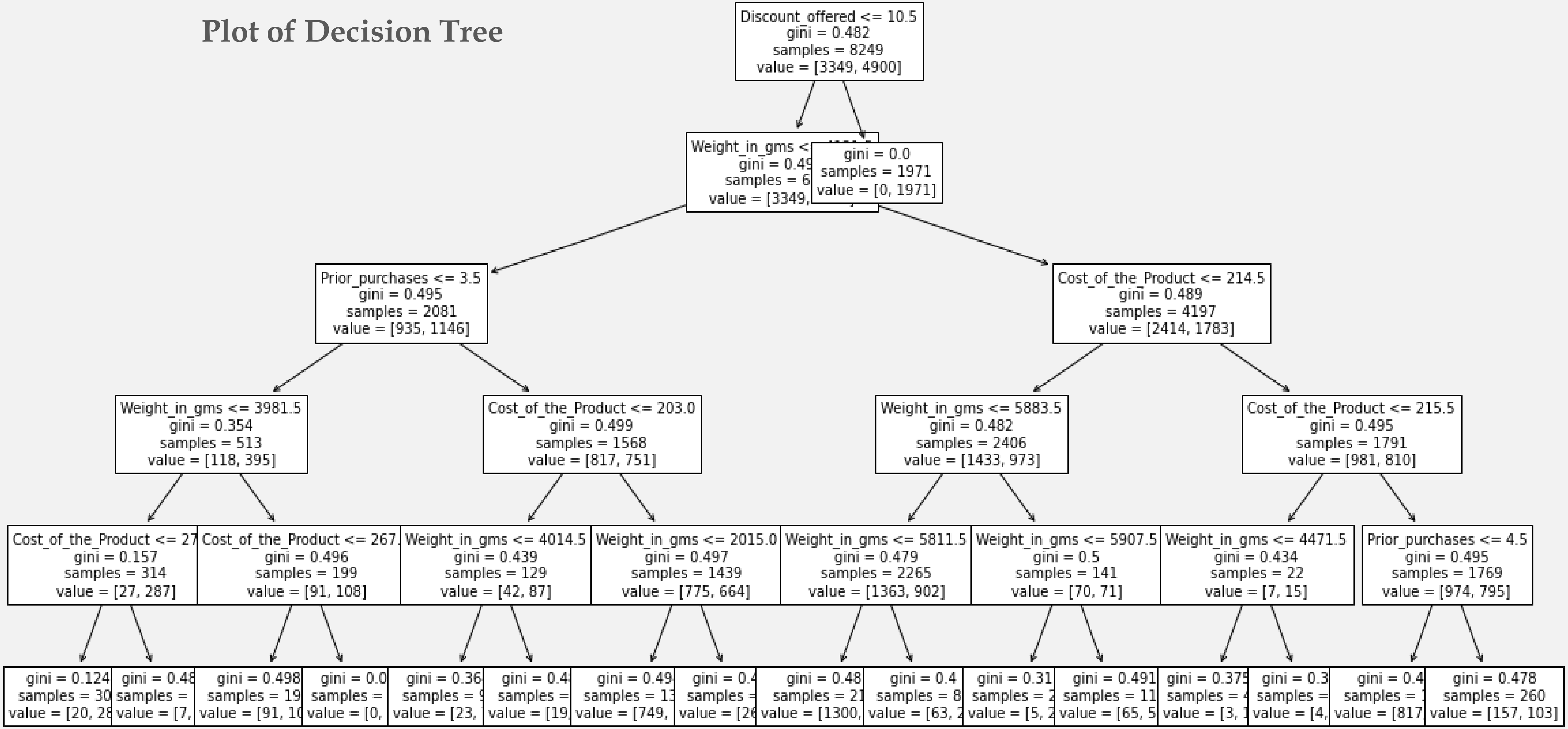
In our decision tree model, we have 5 classes with 6 input variables, and we use Reached on time as output variable. This is the code we use to fit the model.

```
# Decision Tree - training
from sklearn.tree import DecisionTreeClassifier
dt = DecisionTreeClassifier(max_depth=5)
dt.fit(trainX, trainY)
```

Accuracy Score for Decision Tree model: 0.6832727272727273



Plot of Decision Tree



Apply SVM model

About SVM:

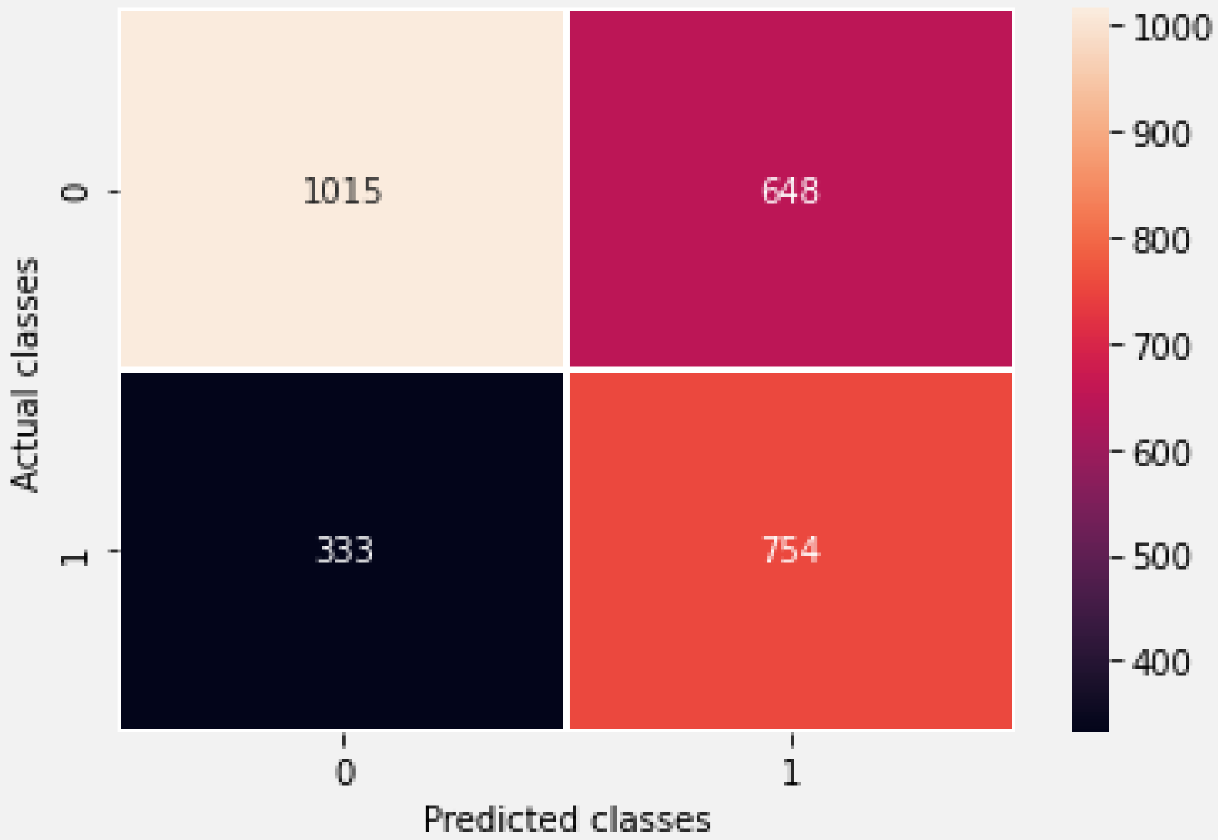
- A support vector machine (SVM) is a supervised machine learning model that uses classification algorithms for two-group classification problems.
- After giving an SVM model sets of labeled training data for each category, they're able to categorize new text.

The code used to fit the model:

```
svm = SVC(kernel='linear')
svm = svm.fit(trainX, trainY)
```

Accuracy Score for SVM model: 0.6432727272727272

The model makes 64% accuracy rate on testing dataset



5. Model selection

Model	Accuracy Score	Type I Error
Logistic Regression	64%	454
Decision Tree	68%	65
SVM	64%	333

Decision Tree model has the highest accuracy score and lowest type I error. We will choose this model to predict the on-time shipping on our dataset

6. Conclusion and future works

a. Results and Analysis:

✓ Relationship between feature variables and target variable

- Cost of the product, Prior purchases, Discount offered, Weight in grams and Product important function have significant influence on the model, specially Cost of the product and Prior purchases

- 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 or heigher in on-time shipping

6. Conclusion and future works

Results and Analysis

✓ **Predict on-time shipment:**

- Built the final prediction model with a Decision Tree algorithm as they result in the highest accuracy score (68%) and the least type I error (65), depending on the model segments.

6. Conclusion and future works

b. Limitations of our approach

- It requires a lot of data in order to kick-start the analysis and modelling and the data should be at best great quality
- A model demands analytics and coding skills
- A project only supply a functioning tool, this means the company will have to work on the models in order to make them a part of their computing environment

6. Conclusion and future works

c. Potential future works

- Limited our scope to electronic products but the same models developed for this purpose could be used for any products.
- Insight of our study is that our approach could be applied to any shipping industry.
- In the future, it would be worth investigating the possibility of including weather and customer's area patterns in the model as a way to improve its accuracy, if access to reliable data can be guaranteed.

7. Reference

Antoine Charles Jean Jonquais and Florian Krempl. “Predicting Shipping Time with Machine Learning.” The Program in Supply Chain management, May 10 2019.

Borucka, Anna. "Logistic regression in modeling and assessment of transport services." Open Engineering 10.1 (2020): 26-34.

Prachi Gopalani. “Product Shipment Delivered on time or not? To Meet E-Commerce Customer Demand.” https://www.kaggle.com/prachi13/customer-analytics, Feb 23 2021

Susan Li. “Building A Logistic Regression in Python, Step by Step.” https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8, Sep 28, 2017.

De Gruyter. “Logistic regression in modeling and assessment of transport services”, Jan 31 2020.

Q&A