CLARK UNIVERSITY

STAT 4650 - FALL 2021

Members: HO Quoc Ngoc (Louis HO)
Jiayu Sun
Junyang Yan
Xuyao Liu
Instructor: Professor Yue Gao

E-commerce shipping

Contents

Abstract

In this project, we analyzed the factors affecting whether goods will be delivered on time in e-commerce. In order to better classify whether it will be delivered on time, we try to find the best machine learning model to judge on time delivery. We used five models for our dataset: KNN, Gradient Boosting, Random Forest, Logistic Regression and Decision Tree. And analyzed the correlation between each variable and on-time delivery. Through our research, we found that Gradient Boosting is the best machine learning model in this dataset. Through the logistics regression model, it is found that there is a strong correlation between the five conveniences of Customer care calls, Cost of the product, Prior purchases, Discount offered, and Weight in gms and whether they are delivered on time. And logistics regression and decision tree show that Discount offered is the most relevant variable.

Keywords: E-commerce, Classification models, Prediction.

1. Introduction

1.1 Background

In the past few years, e-commerce has become more and more important. Its business areas have touched every industry to the individual. As the leader of an e-commerce company, on-time delivery is one of the best standards to measure the companies' operation system and scale of development.

On-time shipping is the core of leadership in the promotion of electronic trade in the United States, and customs is the center to promote the proliferation of electronic trade[2]. After the "9.11" incident, it was stipulated that the shipping documents of the goods should be notified electronically to the customs 24 hours in advance. At the same time, as a Chinese company, Alibaba has made great progress in the construction of e-commerce in recent years[3]. For example, each major port has its own relatively independent center, but the center is still at risk due to local regulations. In 2019, Alibaba caused about 2.4 billion losses in the spring due to the imperfect land transportation system. In 2017, due to the improvement of warehouse facilities, Amazon's average profit per quarter increased by US$1.33 billion. In 2020, e-commerce websites such as China Taobao paid more attention to on-time services and set up a performance evaluation system for transportation systems. In order to better attract customers, e-commerce companies need to study their customers. on-time rate and percentage directly reflects the departments' attitude to the company's products.

The influencing factors of whether delivery on-time may include: transportation method, discounts offered, product cost, etc. At the same time, it is also necessary to consider whether the customer will conduct multiple consultations because the purchased product is more important. Whether the number of consultations and the importance of the product will affect the on-time shipping.

We plan to use a data set from Kaggle, an international e-commerce company that sells electronic products, to analyze the specific factors that affect the company's customer reviews, so as to help the company's development.

1.2 Hypothesis

1) Is it true that the greater the discount will result in the failure to arrive on time?
2) Does the weight of the products affect whether it arrives on time?
3) Is the more expensive the goods more likely to arrive on time?
4) Does the number of times a customer calls affects arrival on time?

Motivation:

We try to find the most influential factors in e-commerce through data analysis. And by analyzing the results, it helps e-commerce companies to deliver goods on time as much as possible.

1.3 Feature description

- ID: ID Number of Customers.
- Warehouse block: The Company has a big Warehouse which is divided into blocks such as A,B,C,D,E.
- Mode of shipment: The Company Ships the products in multiple ways 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 a rating 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 Purchases.
- Product importance: The company has categorized the product in various parameters 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 original data is collected daily from on Customer Analytics stored in GitHub repository. This data is made available on Github in Data collected data about product shipment to Kagglers.

Link of data: https://www.kaggle.com/prachi13/customer-analytics[1].

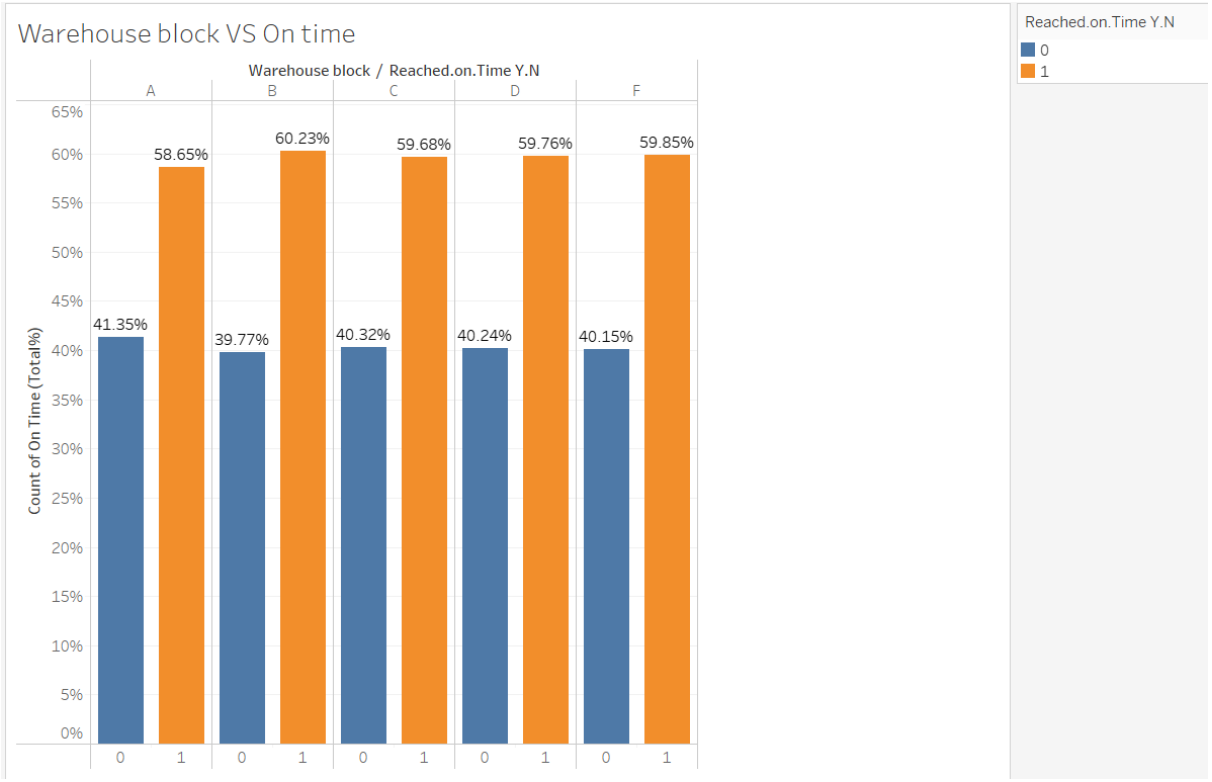
The dataset used for model building contained 10999 observations of 12 variables.

2. Visualization

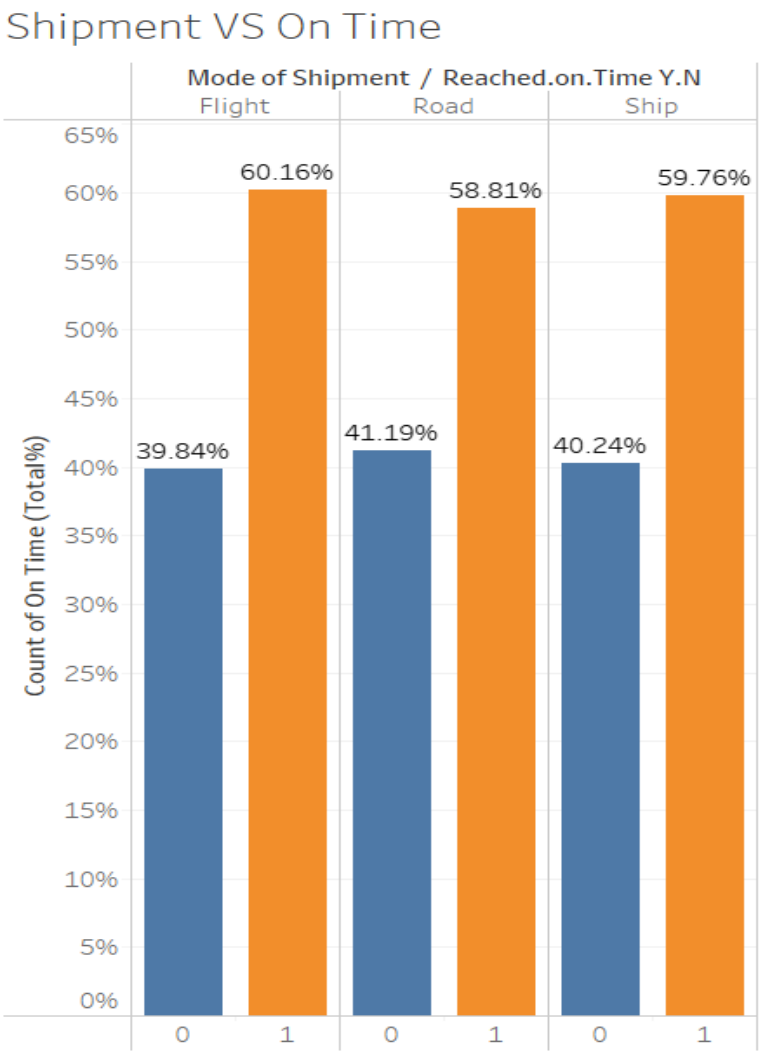
We use Tableau to visualize the dataset.

Reached.on.Time Y.N	
0	4,436
1	6,563

First we can notice the report of each warehouse block having the same distributions of whether the goods delivered on time. nearly 40% of goods were delivered on time and 60% of goods were not. and this result will reflect on our prediction.

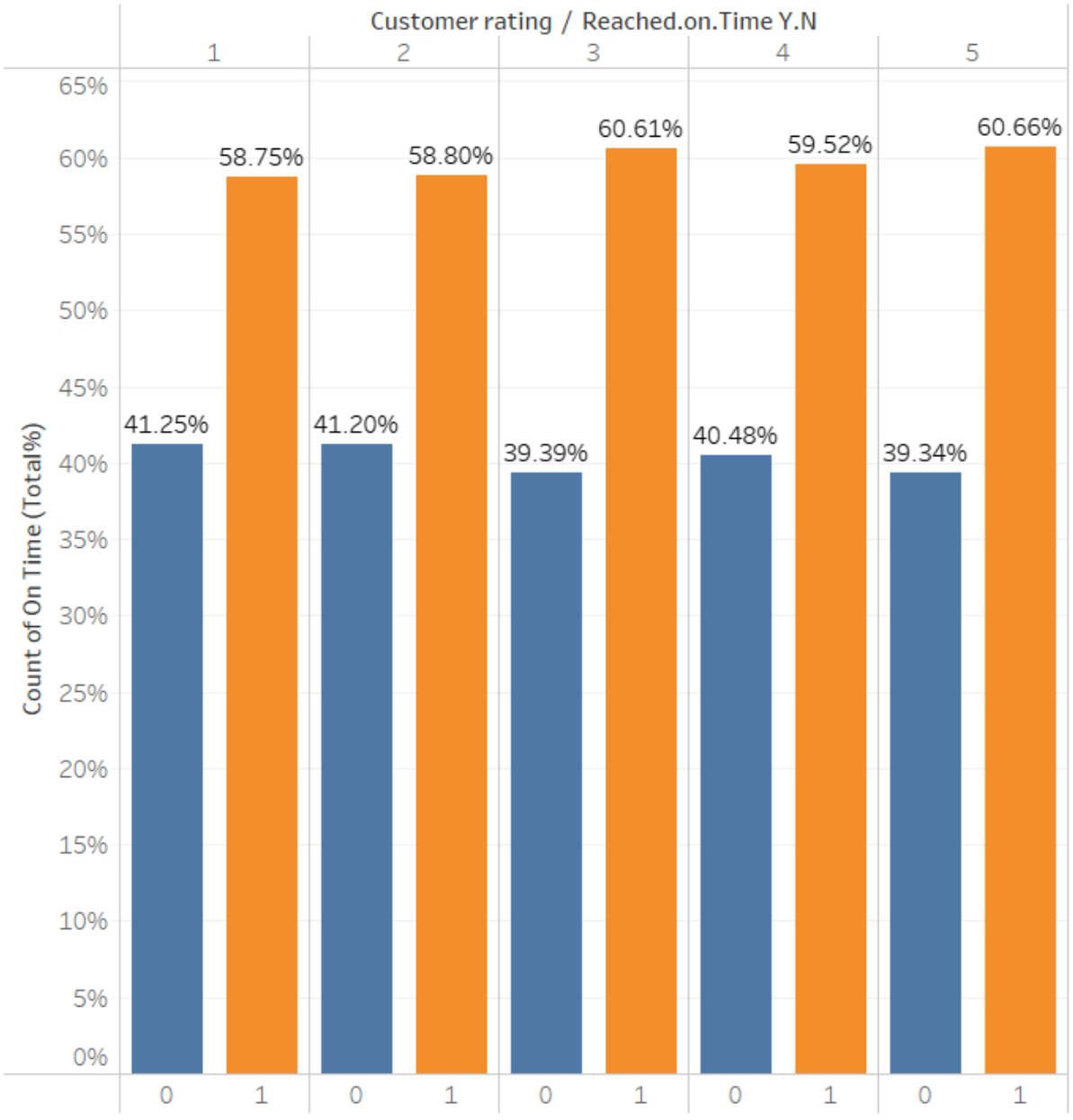


The figure above depicts the relationship between different warehouses and on-time delivery ratios. We found that in different warehouses, the proportion of goods being delivered on time was roughly the same, around 60%. Therefore, we believe that the warehouse has little effect on the on-time delivery.

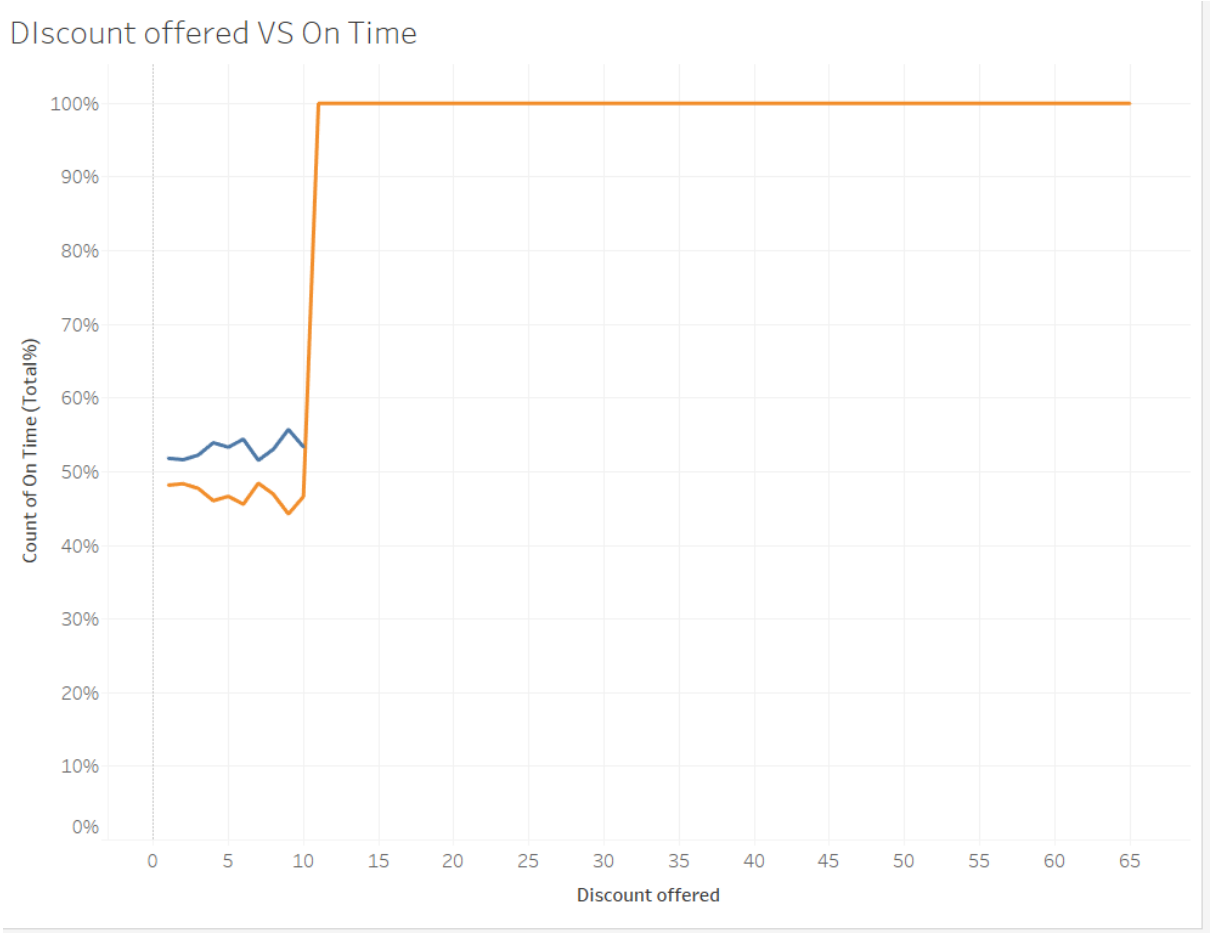


The figure above depicts the relationship between different transportation methods and the on-time delivery ratio. We found that in the three modes of transportation, the proportion of goods being delivered on time was roughly the same, at about 60%. Therefore, we believe that the mode of transportation hardly affects whether the goods are delivered on time.

Coustomer rating VS On Time

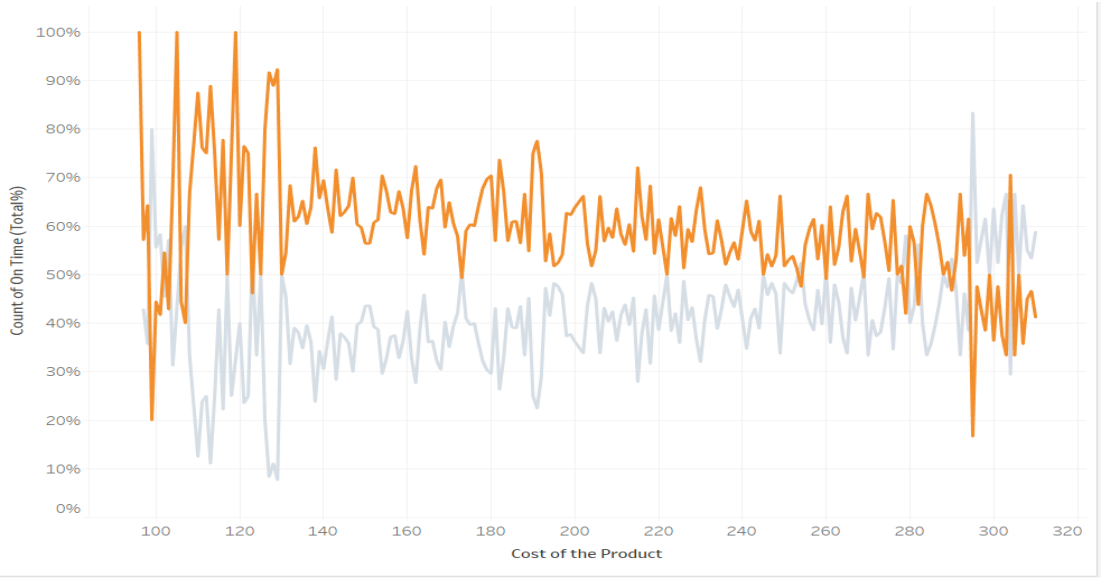


Then we find the relationship between on-time percentage to the customer rating,which is similar to the previous result. So we then set a hypothesis that the reached on time is not highly correlated to warehouse blocking, mode of shipment and customer rating.

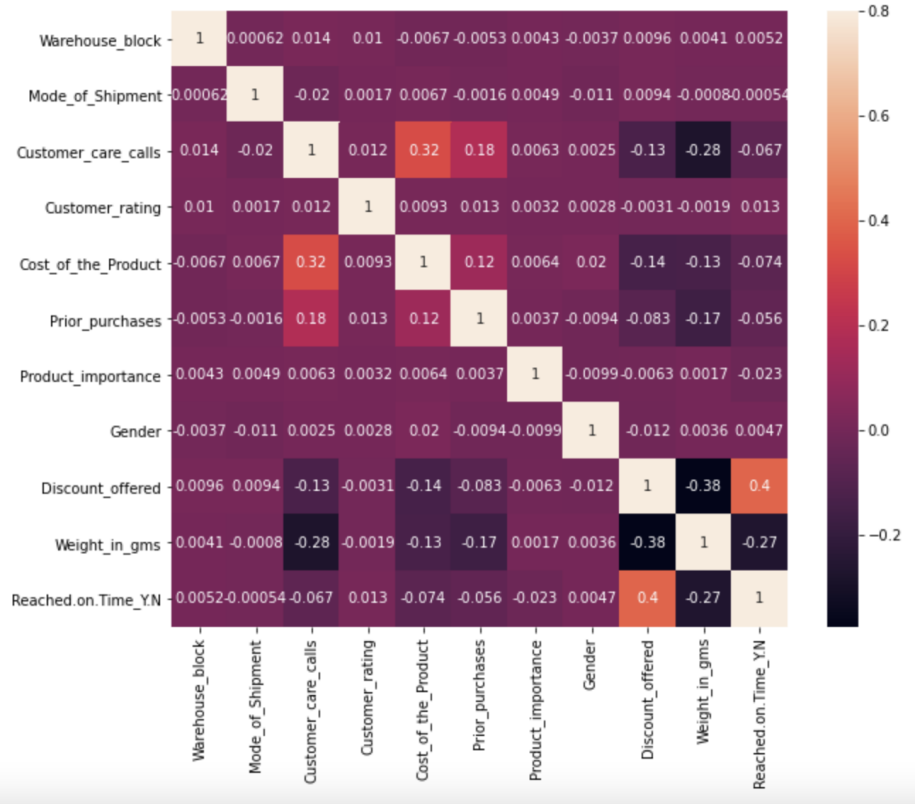


The line graph above describes the relationship between discount and on-time delivery percentage. We found that when the discount is less than 10, the proportion of not being delivered on time is higher. But after the discount is greater than 10, all the goods are

delivered on time.



The above line graph describes the relationship between the percentage of the price of the goods delivered on time. We only observe the lines that are delivered on time, and we find that the overall trend of volatility is decreasing. When the price of goods increases, the proportion of on-time delivery decreases.



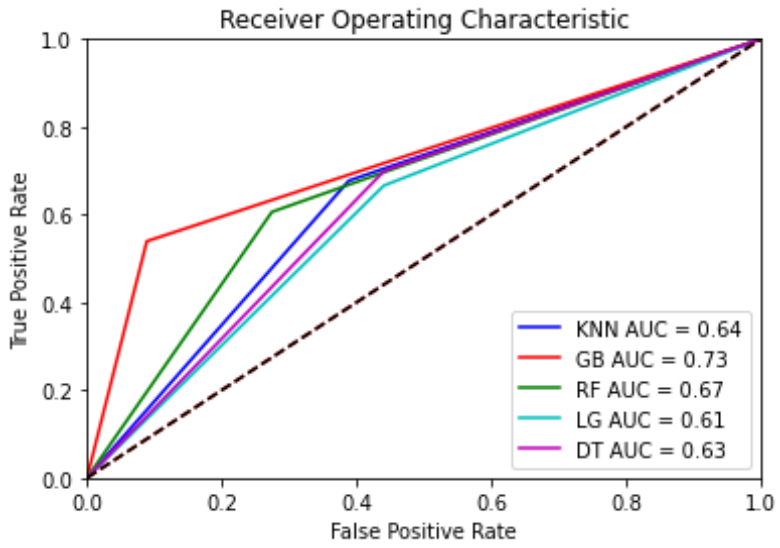
From the heat map, we can find that there is a medium or weak correlation between the variables. This means that there is no multicollinearity between the variables. The relationship between the variables are also weak and some negative correlation. More detail, the variable Customer_care_calls has a strong correlation between the Prior_purchases and Cost_of_the_Product variables compared to other variables. Besides, the variable Customer_rating is not related with other variables in this data. The variables Weight_in_gms has negative relationship between the variables as Reach.on.time_YN, Discount_offerred, Prior_purchases, Cost_of_the_Product and Customer_care_calls, it means the weight_im_gms has a negative affect with others.

3. Methodology

We have tried five models for our dataset: KNN, Gradient Boosting, Random Forest, Logistic Regression and Decision Tree. Here is the calculation result of these five models.

	Model	Training Score	Test Score (Accuracy)	Precision	Recall	F1 Score
0	K-Nearest Neighbors	0.782362	0.651364	0.651364	0.651364	0.651364
1	Gradient Boosting	0.715081	0.685909	0.685909	0.685909	0.685909
2	Random Forest	1	0.661364	0.661364	0.661364	0.661364
3	LogisticRegression	0.637913	0.624091	0.624091	0.624091	0.624091
4	DecisionTree	1	0.660909	0.660909	0.660909	0.660909

All five models show a high accuracy rate, around 65%, which is higher than random guessing. Among them, Gradient Boosting has the highest accuracy rate of 68.59%, which means that the model is relatively careful to avoid marking that it is not delivered on time as delivered on time.



The above figure shows the ROC curve of five models. From this figure, we find that Gradient Boosting has the best ROC curve with an AUC area of 0.73, which is relatively high.

3.1 LogisticRegression

In order to find the correlation between each variable and whether it was delivered on time, we established a logistic regression model. In this model we use 6 variables, namely: Customer care calls, Customer rating, Cost of the product, Prior purchases, Discount offered, Weight in gms.

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[4].

```
Optimization terminated successfully.
         Current function value: 0.547883
         Iterations 8
                           Logit Regression Results
==============================================================================
Dep. Variable:                On_Time   No. Observations:                 8799
Model:                          Logit   Df Residuals:                     8792
Method:                           MLE   Df Model:                            6
Date:                Wed, 01 Dec 2021   Pseudo R-squ.:                  0.1886
Time:                        21:12:33   Log-Likelihood:                -4820.8
converged:                       True   LL-Null:                       -5941.3
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
                         coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const                  1.2404      0.207      6.001      0.000       0.835       1.646
Customer_care_calls   -0.0966      0.024     -4.034      0.000      -0.144      -0.050
Customer_rating        0.0298      0.017      1.728      0.084      -0.004       0.064
Cost_of_the_Product   -0.0021      0.001     -3.704      0.000      -0.003      -0.001
Prior_purchases       -0.0801      0.017     -4.648      0.000      -0.114      -0.046
Discount_offered       0.1147      0.005     22.814      0.000       0.105       0.125
Weight_in_gms         -0.0002   1.79e-05    -12.349      0.000      -0.000      -0.000
==============================================================================
```

Logistic Regression results show that Customer care calls, Cost of the product, Prior purchases, Discount offered, Weight in gms. The P values of these five variables are all small, which means that these five variables are significantly related to the delivery on time. The P value of Customer rating is 0.084 greater than 0.05, which means that there is no obvious correlation between Customer rating and delivery on time. In the LogisticRegression result table, we can find the coefficients of each table variabyile, from which we have established a Logistic Regression model:

$$\hat{P}(X) = \frac{e^{1.2404-0.0966x_1+0.0298x_2-0.0021x_3-0.0801x_4+0.1147x_5-0.0002x_6}}{1+e^{1.2404-0.0966x_1+0.0298x_2-0.0021x_3-0.0801x_4+0.1147x_5-0.0002x_6}}$$

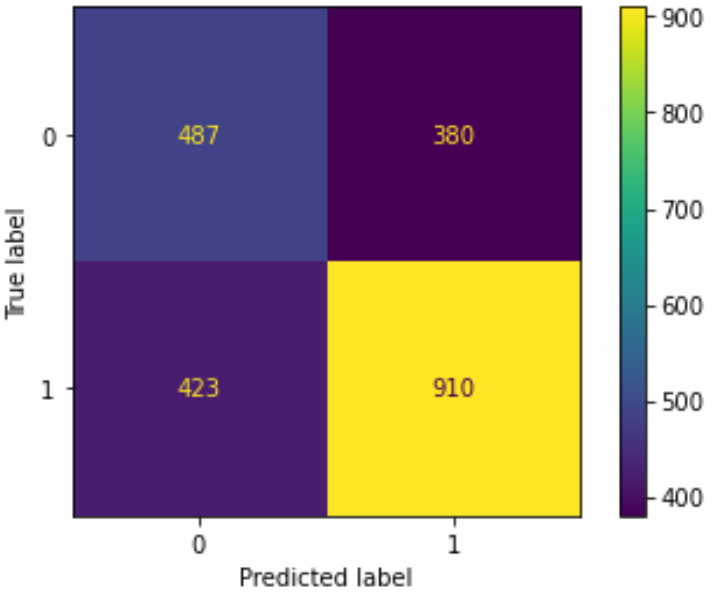
Among them, X1 is Customer care calls, X2 is Customer rating, X3 is Cost of the product, X4 is Prior purchases, X5 isDiscount offered, and X6 is Weight in gms.Since the P value of Customer rating is relatively large, we modify the model to:

$$\widehat{P}(X) = \frac{e^{1.2404-0.0298X_2-0.0021X_3-0.0801X_4+0.1147X_5-0.0002X_6}}{1+e^{1.2404-0.0298X_2-0.0021X_3-0.0801X_4+0.1147X_5-0.0002X_6}}$$

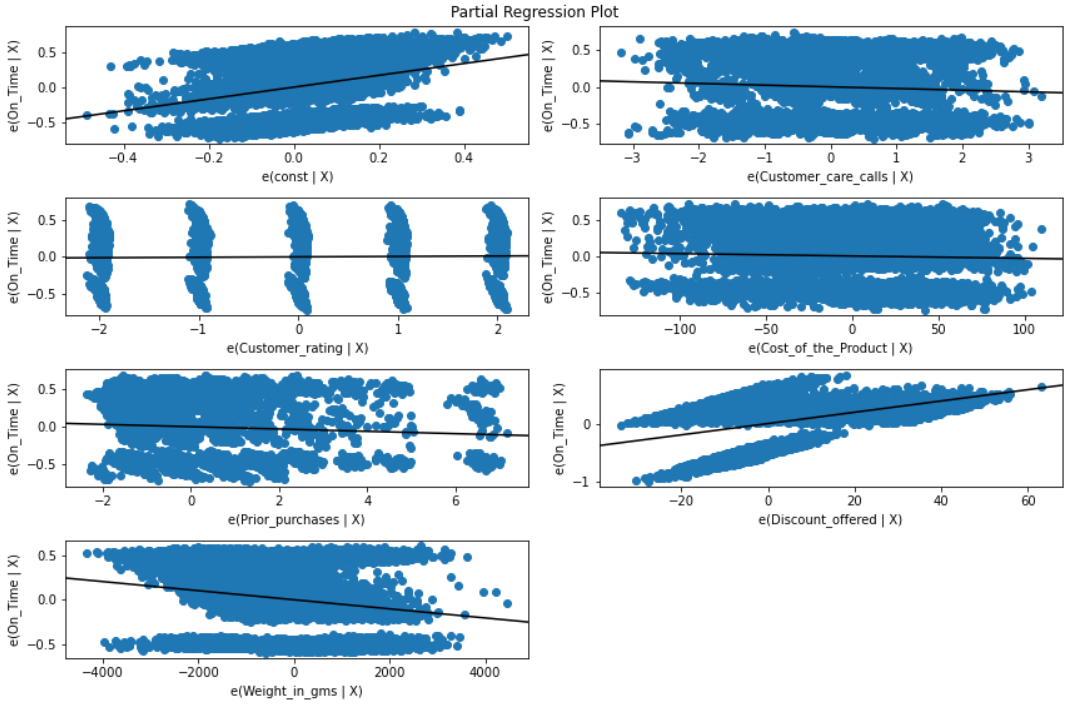
The correlation coefficient of Discount Offer is the highest, and Discount Offer can be regarded as the most important factor affecting on-time delivery. Among the coefficients of these five variables, the coefficient of Discount Offer is positive, and the relationship between Discount Offer and on-time delivery is positively correlated, meaning that the larger the discount, the more likely it is to deliver on time. The coefficients of the other variables are negative, and the other four variables are negatively related to the probability of delivery on time.

```
Accuracy: 0.635
              precision    recall  f1-score   support

           0       0.54      0.56      0.55       867
           1       0.71      0.68      0.69      1333

    accuracy                           0.64      2200
   macro avg       0.62      0.62      0.62      2200
weighted avg       0.64      0.64      0.64      2200
```

The accuracy of this model is 0.635, and the correct number of predictions accounts for 64% of the total sample size. Through the confusion matrix below, we can also get that 487+910 samples out of 2200 samples are correctly classified.



The scatter plot shows the relationship between each variable and arrival on time. It is consistent with the correlation between our logistic regression results.

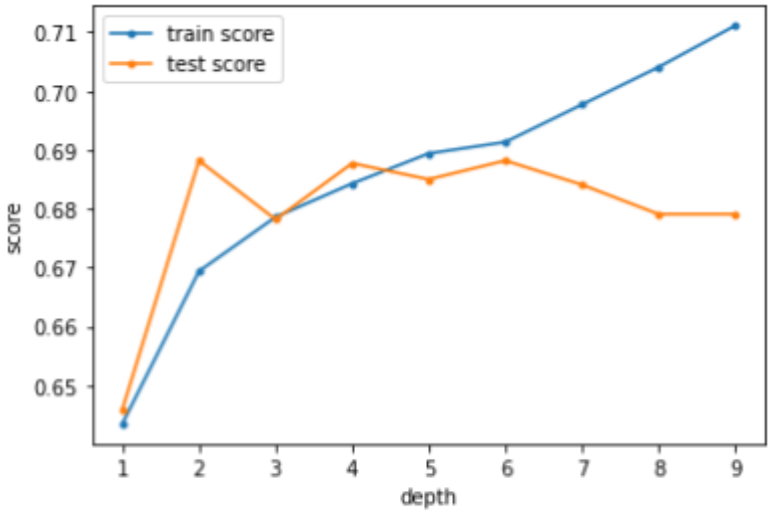


3.2 Decision trees

We try to use a decision tree to predict the relationship between each variable and the delivery on time.

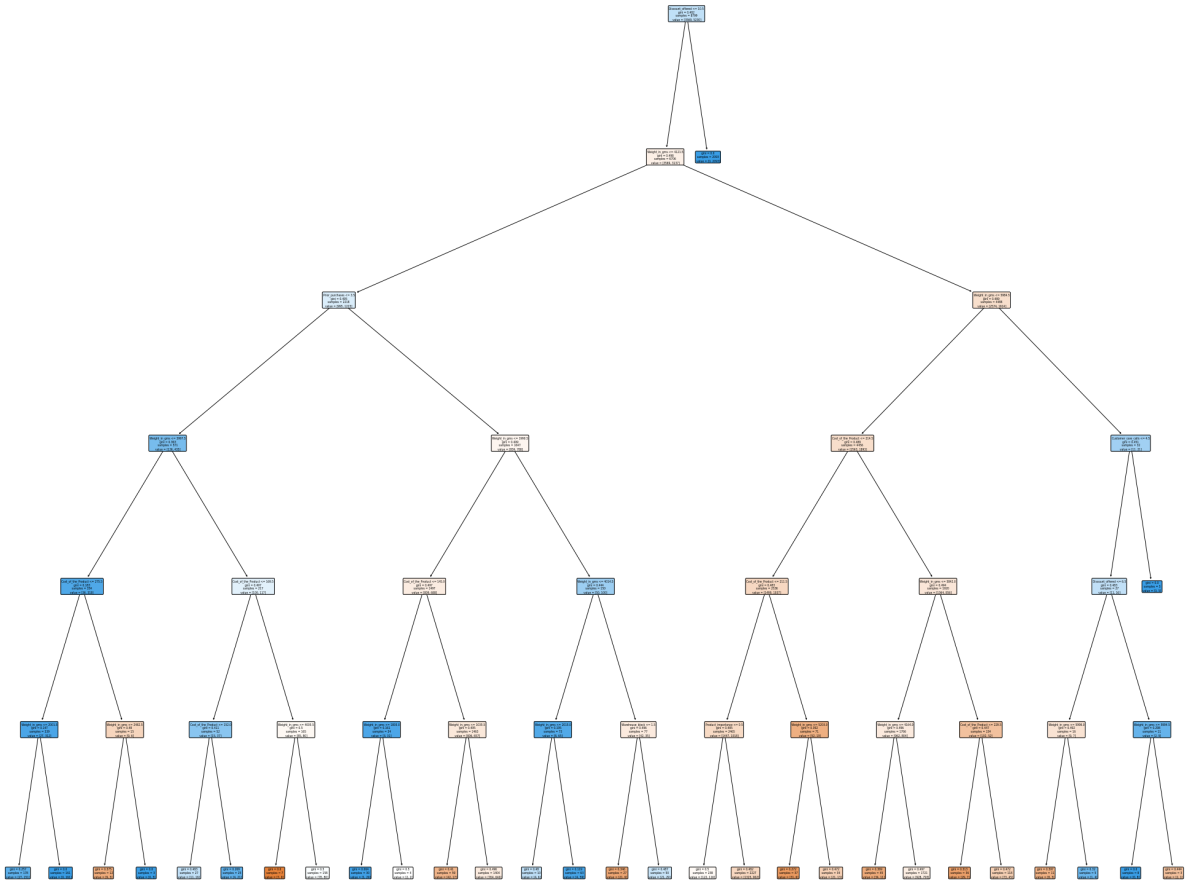
A decision tree is a tree-like model that acts as a decision support tool, visually displaying decisions and their potential outcomes, consequences, and costs. From there, the "branches" can easily be evaluated and compared in order to select the best courses of action[5].

From the accuracy and depth graphs, we think that the maximum depth of 6 is a good number. For this decision tree model, a high accuracy rate can be obtained, which is 68.82%

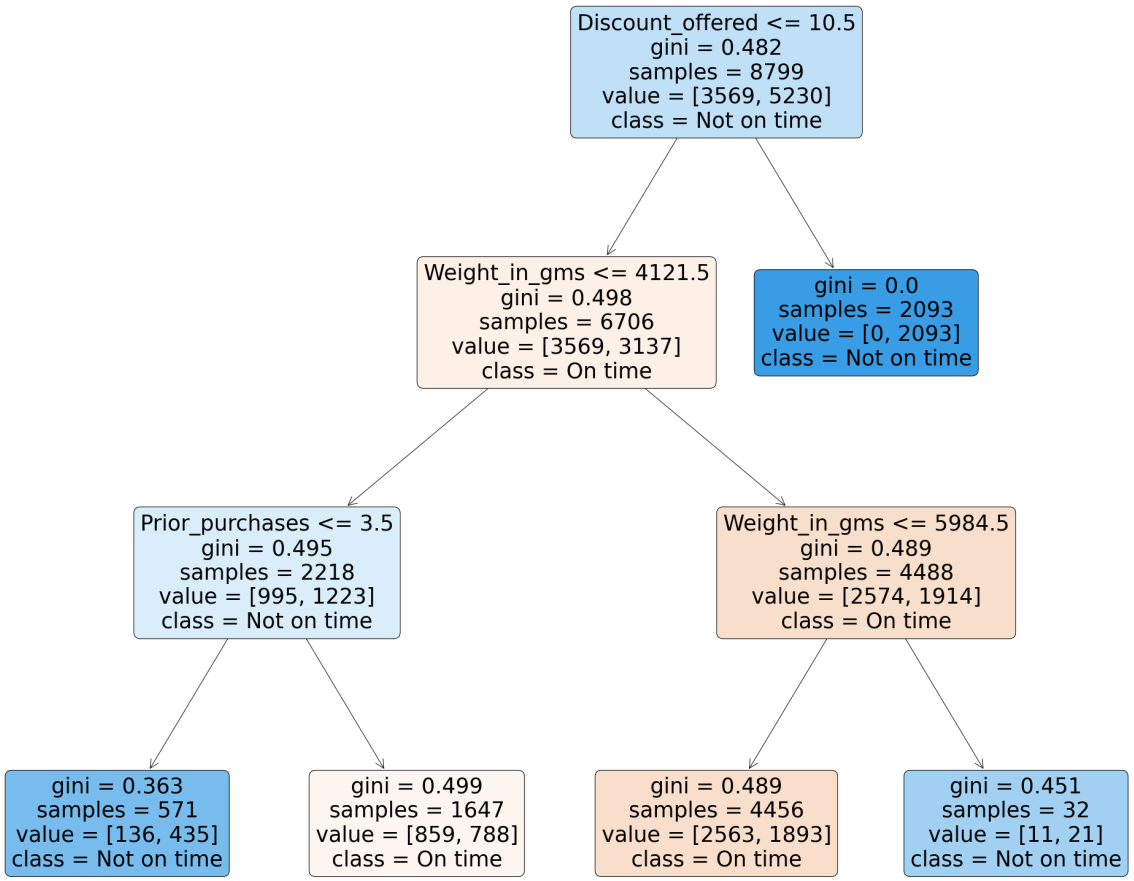


The accuracy of the testing data with max depth is 6 : 68.82%

Next, we created a decision tree with a maximum depth of 6.



Since we use Python, the decision tree with a maximum depth of 6 is not very clear, so we also created a decision tree with a maximum depth of 3 for analysis.




```
Accuracy on Training data   : 67.85998408910103
The accuracy of the DT model: 67.82%
```

We obtained a decision tree with a maximum depth of 3, and the accuracy rate was 67.82%. discount is the most important factor in determining the probability of delivery on time. In the first layer of the Decision tree, Gini Index is 0.482, which means it is delivered on time or not distributed relatively evenly. From the above visualization, we know that the smaller amount of data is the smaller amount that did not arrive on time. In the decision tree model, we found that in our sample 3569 data were not delivered on time, and 5230 data were

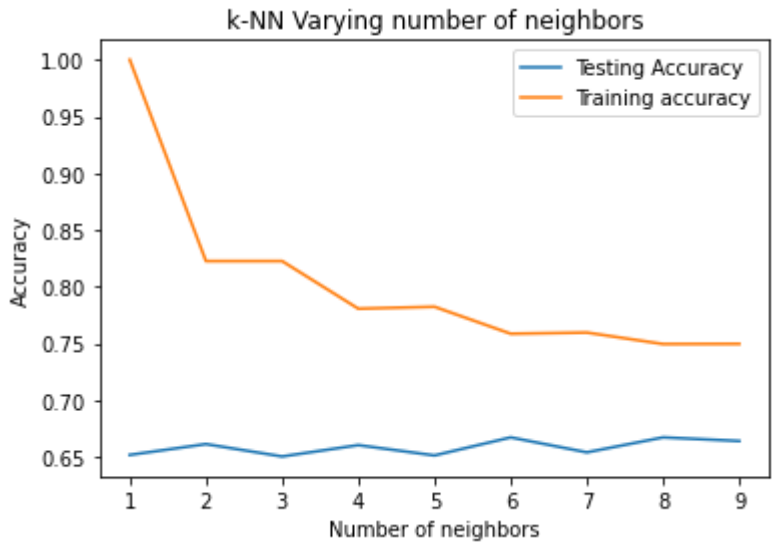
delivered on time. All discounts greater than 10.5 were delivered on time, which is consistent with our visual display results. This could be because when the discount increases, e-commerce companies are doing promotional activities, they will increase the stock of goods and the company's employees in advance to prevent the goods from being delivered on time.

3.3 K Nearest Neighbors

We also used KNN. The k-nearest neighbors algorithm, or kNN, is one of the simplest machine learning algorithms. Usually, k is a small, odd number - sometimes only 1. The larger k is, the more accurate the classification will be, but the longer it takes to perform the

classification[6].

	K	accuracy		precision	recall	f1-score	support	_	_precision	_recall	_f1-score	_support	confusion_matrix
0	1	0.651818	0	0.555556	0.582468	0.568694	867	accuracy			0.651818	2200	[[505, 362], [404, 929]]
1	1	0.651818	1	0.719597	0.696924	0.708079	1333	macro avg	0.637576	0.639696	0.638386	2200	
2	1							weighted avg	0.65495	0.651818	0.653149	2200	
3	2	0.661364	0	0.548259	0.799308	0.650399	867	accuracy			0.661364	2200	[[693, 174], [571, 762]]
4	2	0.661364	1	0.814103	0.571643	0.671662	1333	macro avg	0.681181	0.685475	0.66103	2200	
5	2							weighted avg	0.709336	0.661364	0.663282	2200	
6	3	0.650455	0	0.553611	0.583622	0.56822	867	accuracy			0.650455	2200	[[506, 361], [408, 925]]
7	3	0.650455	1	0.719285	0.693923	0.706376	1333	macro avg	0.636448	0.638773	0.637298	2200	
8	3							weighted avg	0.653994	0.650455	0.65193	2200	
9	4	0.660455	0	0.549751	0.764706	0.639653	867	accuracy			0.660455	2200	[[663, 204], [543, 790]]
10	4	0.660455	1	0.794769	0.592648	0.678986	1333	macro avg	0.67226	0.678677	0.659319	2200	
11	4							weighted avg	0.698209	0.660455	0.663485	2200	
12	5	0.651364	0	0.552083	0.611303	0.580186	867	accuracy			0.651364	2200	[[530, 337], [430, 903]]
13	5	0.651364	1	0.728226	0.677419	0.701904	1333	macro avg	0.640155	0.644361	0.641045	2200	
14	5							weighted avg	0.65881	0.651364	0.653936	2200	
15	6	0.667273	0	0.556866	0.762399	0.643622	867	accuracy			0.667273	2200	[[661, 206], [526, 807]]
16	6	0.667273	1	0.796644	0.605401	0.68798	1333	macro avg	0.676755	0.6839	0.665801	2200	
17	6							weighted avg	0.702149	0.667273	0.670499	2200	
18	7	0.654091	0	0.554303	0.623991	0.587086	867	accuracy			0.654091	2200	[[541, 326], [435, 898]]
19	7	0.654091	1	0.73366	0.673668	0.702386	1333	macro avg	0.643982	0.64883	0.644736	2200	
20	7							weighted avg	0.662977	0.654091	0.656947	2200	
21	8	0.667273	0	0.556866	0.762399	0.643622	867	accuracy			0.667273	2200	[[661, 206], [526, 807]]
22	8	0.667273	1	0.796644	0.605401	0.68798	1333	macro avg	0.676755	0.6839	0.665801	2200	
23	8							weighted avg	0.702149	0.667273	0.670499	2200	
24	9	0.664091	0	0.562257	0.666667	0.610026	867	accuracy			0.664091	2200	[[578, 289], [450, 883]]
25	9	0.664091	1	0.753413	0.662416	0.70499	1333	macro avg	0.657835	0.664541	0.657508	2200	
26	9							weighted avg	0.67808	0.664091	0.667566	2200	



We calculated some scores for K from 1 to 9, including accuracy, F1 score, confusion matrix

and other values. We found that K is 6 and 8 have the best F1 scores, recall values and so on. Meanwhile , we drew a line graph of Testing Accuracy and Training Accuracy, after K=6, it is become stable

4. Conclusion

4.1 Result

In the above analysis, we found that 60 products were delivered on time. We believe that the three factors of warehouse location, transportation method and customer ratings have little effect on on-time delivery. Gradient Boosting can be used for this data as a module, and we obtained an accuracy rate of 68.6%. We found that the discount of the goods is the most important factor affecting on-time delivery. The larger the discount, the easier it is to deliver on time.

4.2 Limitation

Since we are using the decision tree model derived from Python, the decision tree diagram of python gives more information, but when the decision tree has more branches, the diagram becomes fuzzy and it is difficult to see the information clearly. In the future, we can try to use the R language to create a decision tree model.

5.3 Suggestion

- E-commerce companies can recruit temporary workers to avoid shortage of employees to deliver goods on time during the period of big discounts.
- After receiving calls from customers many times,it means that the product may not arrive on time. The e-commerce company can contact the transportation company to increase the possibility of delivery on time.
- E-commerce companies can make preparations for production and transportation of pre-sale products to ensure that they can be delivered on time.

Reference

1. Prachi Gopalani. E-Commerce Shipping Data. Retrieved September 18, 2021. From Kaggle. Website:https://www.kaggle.com/prachi13/customer-analytics
2. Khalid, B. (2008). *Security and risk-based models in shipping and ports: Review and critical analysis.* Retrieved 21 November 2021, from http://www.internationaltransportforum.org/jtrc/DiscussionPapers/DP200820.pdf
3. Cacho, J., Marques, L., & Nascimento, Á. (2020). *Customer-Oriented Global Supply Chains: Port Logistics in the Era of Globalization and Digitization*. Retrieved 18 November 2021, from https://www.igi-global.com/chapter/customer-oriented-global-supply-chains/254702
4. *What is Logistic Regression?.* Retrieved November 21, 2021. From: https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/what-is-logistic-regression/
5. *Decision Tree Analysis.* Retrieved November 21, 2021. From: https://www.omnisci.com/technical-glossary/decision-tree-analysis
6. *KNN.* Retrieved November 21, 2021. From DeepAI. Website: https://deepai.org/machine-learning-glossary-and-terms/kNN