Analysis of Brazilian E-commerce Company Olist on Optimizing Logistics Solutions to Enhance Customers Experience

Big Data Analytics Project

Jiawei Huang, Jingting Xu, Qiaochu Cong, Qiurong Ren

June 12, 2020

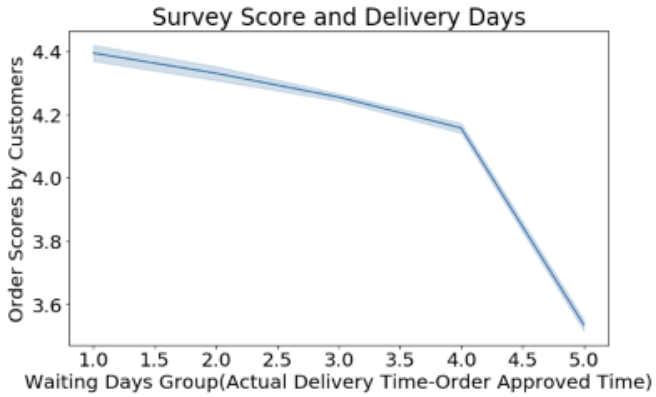
Summary

In this analysis, we've conducted analysis on the Brazilian e-commerce platform and we figured out the importance of logistics related to the customers' experience when they are conducting online shopping. Among the logistics information, the delivery days and the estimated delivery days provided by the platform are generally having large discrepancies. Moreover, we've noticed that different delivery days may induce different evaluation of the customers. Therefore, we've conducted several regression analysis (including linear, regression, decision tree, bagging, random forest, AdaBoost and extra trees regressor) to build a new build to predict more accurate delivery days and compared their performance of the accuracy and of enhancing the customers' rating. As a result, we've provided a model of better predicting the delivery days, that 70% of the new model prediction works better than the estimation provided by the platform. However, in order to conduct a more accurate analysis, we may need extra information and try to avoid potential multicollinearity and single attributes issue in the future.

Problem Description

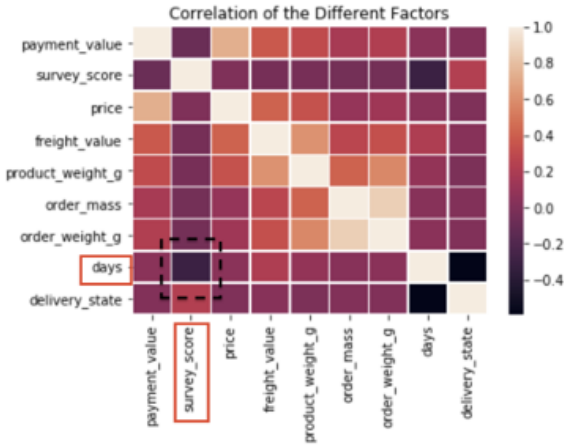
The given eight datasets include 100,000 orders' information of this Brazil e-commerce company range from August 2017 to September 2018. It illustrates the order information (package size, category, purchase date, etc), customer information (payment method, location, etc) and the seller information (city, zip code, etc).

In the case of the e-commerce business, package delivery (logistics) is an important part within online shopping experience, and it is testified by the result after we conducted analysis on the dataset. As graph 1 shows below, the y-axis variable is the order scores evaluated by the customers, which is assumed to represent the customers' satisfaction of the orders. The x-axis variable is the actual waiting days which was generated by using actual delivery time minus the order approved time provided by the given dataset. We then cluster the actual waiting days into 5 groups as graph 2 shows. It's clear that the longer the customers wait for their package to be delivered, the lower the score they will assign to their orders. Another proof of the importance of the package delivery dates to the customer experience is shown on graph 3, that it exits a high correlation between days (waiting days) and the survey score provided by the customers.



Graph 1: Relationship Between Survey Score and Delivery Days

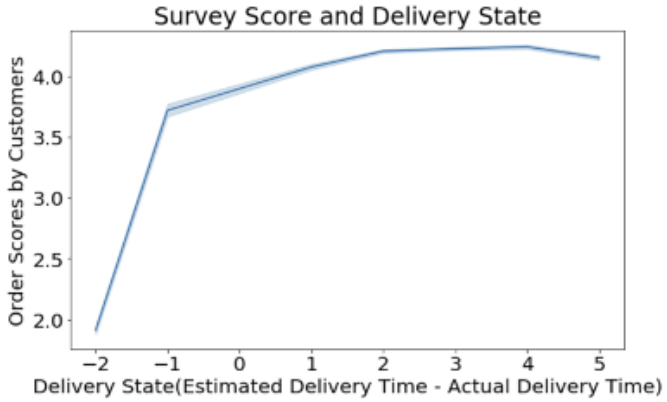
Graph 2: Generating Waiting Days Groups



Graph 3: Correlation Heat-Map among Main Variables

Moreover, since the delivery days are important then we assigned the days into 8 groups as a new variable ‘delivery states’ to help further analysis. As graph 5, it shows how we assigned the days into these 8 groups. During the process, we use the estimated delivery time minusing the actual delivered time to generate the difference between estimated and actual delivered time. When the packages are delayed, that arrived after the estimated time, they will be assigned with negative

values. That's mean, the positive value represents the packages are delivered before the estimated time. Graph 4 shows two scenarios, (i) firstly, when the packages are delayed, the survey scores tend to be very low. What's more, the longer the packages are delayed, the lower the scores are assigned. (ii) Secondly, when the packages were delivered before the estimated time, the scores tend to be higher and then go lower, which means that customers seem to prefer to receive their package before the estimated days but not too early. However, for this situation, may include some multicollinearity solution to help explain. Overall, whether the packages are able to be delivered accurately highly influences customers' shopping experience.



Graph 4: Relationship Between Survey Scores and Delivery State

Graph 5: Generating Delivery State Groups

As the above analysis and proofs shown, it is worthwhile to use the existing information from this brazil e-commerce company's dataset to help predict a more accurate delivery time provided to the customer. By doing so, we expect that a more accurate delivery time will result in higher scores evaluated by the customers of their shopping experience. Besides, we assume that better shopping experiences are able to improve customer stickiness, as well as sales for the company.

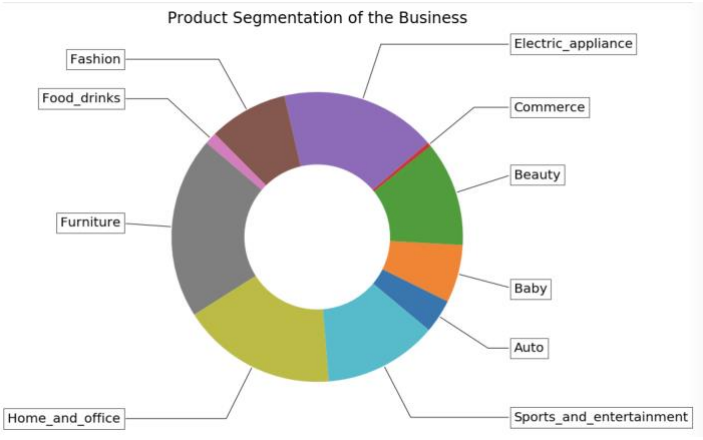
Analysis Task

In order to predict the accurate delivery time, we planned to build feature and label data by processing the original dataset and use new feature data, label data, a regressor model and pyspark

to predict new estimated delivery days. The existing datasets contain region and geolocation of both customers and sellers, product volume, product weight, delivery time, predicted delivery time, time of sellers sending products to logistic stations and order approved time. We are going to generating variables to represent whether sellers and customers are within same zip code, city, state regions, the euclidean distances, the volumes and weights for each order, the difference between actual delivery time and estimated time, the length for orders to be delivered, and the length for sellers sending products to logistic stations. With these new generated variables from the provided dataset, it is worthwhile to perform the analysis and build the model of predicting accurate delivery days.

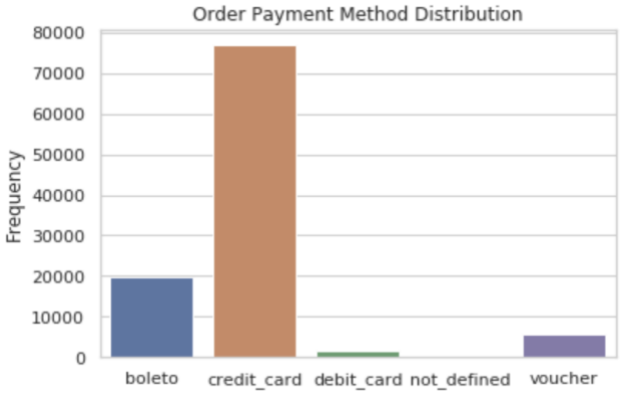
Describe Data

The products are manually labelled into ten categories after browsing the whole product categories classification. A donut chart (Graph 6) was generated to show the market share distribution by product categories. As it is shown in the chart, the furniture, electric appliance and home & office products are the top 3 important elements of the business. In comparison, the auto and food & drinks products account for less proportion among the overall orders. One possible explanation is that food & drinks products can be easily obtained from supermarkets and chain stores. However, products like furniture are hard to carry and online shopping will help customers to transport furniture, which will be a great advantage. Therefore, after analyzing these market shares, it is worthwhile to have further notice on these important categories' products both in logistics situation and the quality, CRM of the products and sellers, since they account for a large portion of the overall business.



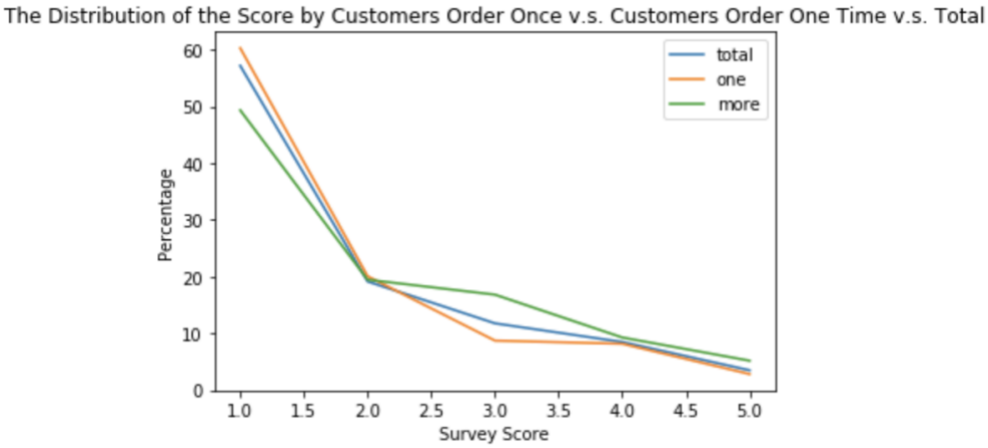
Graph 6: Market share by product category

As of the payment dataset, graph 7 is a bar chart showing the frequency of each payment methods' portion for the overall online ecommerce business. Credit cards rank the most popular payment method that has the highest usage frequency. The second place is taken by boleto which is an official electronic payment method supported by Brazil government. Overall, the credit card wins over the five payment methods.



Graph 7: Market share by product category

As for the score distribution graph 8 shows, they generally range between 1 to 2, which means that the average experience in shopping within this e-commerce company is not too well. Therefore, strategies should be planned and counter methods should be conducted to help ameliorate this situation. For those customers ordered more than once, they generally have higher scores rating at the middle score 3 than the customers ordered only one time. However, it is worth noticing that the customers evaluated of higher score account for only around 10% of the total customers. This situation reinforces the need of sustaining these customers by enhancing their shopping experience.

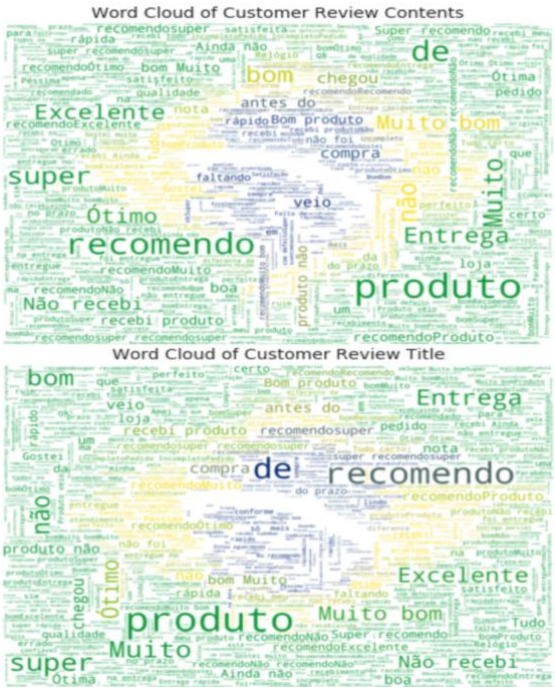


Graph 8: Distribution of score by customers order once v.s. more than one time v.s. total

It is of valuable usage to take a look at the relation between delivery days and the categories since logistics may be varied based on the average size of the products. The radar plot (graph 9) shows the average days the customers received their packages after they confirmed the purchased orders. It shows that they are at average needs 10 days and among all, the category of food and drinks needs less time.

Graph 9: Radar plot of different categories' average arrival days

In case of evaluating the shopping experience of the customer. These below two word clouds (graph 10) shows the overall around 10,000 orders' customers reviews. It seems that generally the customers have a recommended attitude towards this e-commerce platform, while it needs to research for further information about the products to see more real problems and issues behind.



Graph 10: Word clouds of the customer review contents and title

Data Preparation Process

The unit observation of this dataset is each order. There are 8 datasets. Seller id, product id, order id and customer id represent identified id for each seller, product, order and customer. Besides, there is a variable named 'distance' which represents the distance between sellers and customers. We generated this variable by using zip code, city and state information of sellers and customers. This is a categorical variable and there are four categories: within same zip code, same city but not same zip code, same state but not same city, and not same state. In addition, we used the geolocation of sellers and customers to generate the euclidean distance between sellers and customers. By generating these two variables, we are able to know the distance between the sellers and the customers, as well as estimating how long it will take for the delivery.

Next, we have length, width, height and weight for each product. Each order may contain one product to more than one product. We generated the volume of each product by using the length, width and height of each product. Then, we summed the volume and weights of each order and got two new variables. Next, we transformed the variables that represent time into datetime format. And we want to generate 3 variables that represent duration of time. The first one is the actual length of time for the orders to be delivered. We used the actual delivered time to minus the orders approved time, and transformed the variable into numerical data type to represent the length of days.

Besides, as graph 2 shows, if the actual length is less than 3 days, it will be classified into group 1. If the actual length equals or more than 3 days and less than 5 days, it will be classified into group 2, etc. The second variable indicates time durations is the difference between actual delivery time and the estimated delivery time. As graph 5 shows, we use estimated delivery time to minus the actual delivery time. Negative value means the order is delayed.

Then, we also classified this variable into groups as graph 5 shows. The third variable indicates the time length of sellers sending their products to a logistic station. We used the shipping limit data and order approved data to generate a new shipping limit days variable. It means the days that the seller hands over the package to the logistics partner after the customer submitting the order.

Newly Generated Variables Information

New Generated Variables	Meaning of the Variables	Original Variables Used
days	The days customer received the package after they submitted the order	order_customer_delivery_date; order_approved_at
shipping_limit_days	The days the seller hand over the package to the logistics partner after customer submitted the order	shipping_limit_date; order_approved_at
estimated_days	The days customer estimated to receive the package after they submitted the order	order_estimated_delivery_date; order_approved_at
delivery_state	The difference in days that customer estimated to receive the package and actual received days	order_estimated_delivery_date; order_customer_delivery_date
order mass	The overall size of the package	product_length_cm; product_height_cm; product_width_cm
distance	The distance between sellers and customers	region 1; region 2; region3; latitude; longitude
Eucl	Euclidean data	longitude_seller; latitude_seller; longitude_con; latitude_con
category	Listed 10 categories of the products sold	product_category_name
binary_order_times	0 presents customers only purchased one time, while 1 presents customers purchased more than once	customer_id

Analysis Approaches

The feature we used can be described into three categories: (i) geolocation features, (ii) time features, and (iii) order features. Geolocation features describe if the seller and consumers of the order came from the same regions. The geolocation dataset we use distributes locations into region 1 (state), region 2 (municipalities) and region 3 (Statistical Areas). The geolocation features also include the euclidean distance of the sellers and consumers. Time feature contains the days of the seller handing the order to logistics partners. The order features include order mass and order volume.

The actual order shipping days are the label of our analysis. Again, our goal is to predict better estimated delivery days for the consumer. To prevent overfitting problems, we divide the features and labels into training and testing sets (training: 80%, testing 20%). When selecting the best models, we select the model that has the highest performance in the test dataset. The criteria we set up for selecting the best model is not only based on overall accuracy. As mentioned in the previous part, the customer prefers to see an expected delivery date that is exact the true delivery date or a litter later than the delivery date. Based on that, we develop these criteria below:

1) If both original expected delivery days and our prediction is less than the actual delivery days, the closer to actual delivery days the better.

2) If both original expected delivery days and our prediction is larger than the actual delivery days, the closer to actual delivery days the better.

3) If one result is larger and one result is smaller than the actual delivery days, the larger one is better.

To achieve this goal, we make a demo to see the performance of different regressors (with default parameters) in predicting shipping days. The result of the demo is listed in the analysis results section. After selecting which is the best regressor for this project, we tune some parameters to achieve higher performance.

Describe challenges and solutions

(1) Data cleaning

In the process of cleaning data, several data quality problems appear. First, in the geolocation dataset, some of the cities have names which are both in English and Portuguese. This problem causes a situation when we group the data, some data in the same city are splitted into different groups. Our solution is purchasing a Brazil zip codes dataset and using this dataset to regenerate a clean dataset with unique city names. Second, in the products dataset, there are too many detailed categories which make it hard to analyze. In this case, we manually classify the products into different categories which are clear and easy to understand. Third, in the zip code dataset, after merging it to the main dataset and comparing the boundary of Brazil to the geographic range of our dataset, we found that there were many outliers, which either lack longitude and latitude or are outside Brazil. So, to solve this problem, the outliers are just dropped since they are wrong data. Fourth, wrong data also appears in delivery dates. Some of the data have actual delivery dates earlier than the time when the orders were approved, which does not make sense. These data are also dropped to make the dataset to be more reasonable.

(2) Model evaluation

The main problem we face in the model evaluation part is that we can’t solely rely on the accuracy to evaluate our result. We want to know whether our model has a better prediction than the existing estimation but what we can get from the built-in evaluation of the model is just the fitness of actual days. So, we need to build up our own standard. Also, the built-in auto-tuning method like grid search won't work in this situation. Manual tuning is the basic method. To show the result of a better prediction, two columns called “better” and “accuracy” are generated. “Better”

column shows a boolean test. '1' means our model is better according to our new standard. "Accurate" column also shows a boolean test. In this case, '1' means the predicted or estimated delivery days equals to the actual delivery days.

Analysis Results and Insights Gained

(1) Model Selection

To choose the best model, we set up two criteria: 1) is our prediction better than the original expectation delivery date? 2) is our prediction more accurate than the original expectation? The first criteria is calculated based on the logic in the previous section. The result is the percentage of situations where our model prediction is better among all predictions in testing datasets. The second criteria is created by comparing the accuracy rate of our own model prediction and the original expected delivery days. The first number is the number of situations where our model prediction accurately predicts the actual delivery days and the second number is the accuracy of original expected delivery days. The results are presented below. Based on the comparison above, we choose the extra trees regressor as it has the best performance in both criteria.

The extra trees model implements a meta-estimator that matches a large number of random decision trees (also known as extra trees) on each subsample of the data set and USES averages to improve prediction accuracy and control overfitting. From a statistical point of view, the use of irrelevant features can help reduce variance while increasing the bias of the training model. If too many irrelevant (noise) features are selected for the split point, bias and variance are both too high, which is also a weakness of Extra Trees.

Model Selection Demo

Regressor	Better?	Accurate?
Linear Regression	0.6337	1392, 343
Decision tree regressor	0.5742	5890,343
AdaBoost regressor	0.7007	880, 343
Bagging regressor	0.6528	3158, 343
Extra-trees regressor	0.7076	5495,343
Random forest regressor	0.6835	3066,343

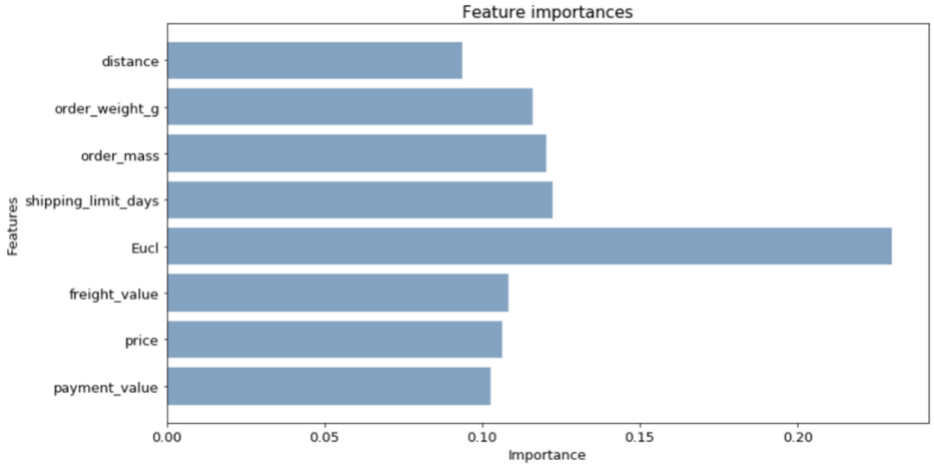
(2) Description of the Model

After selecting the model, we tuned some parameters to see which set of parameters reuters higher performance. Since we are not judging model performance solely based on overall accuracy, we are not able to use grid search to return a single set of parameters. We calculate the two criteria by manually assigning the parameters. The results of the performance tuning are listed below. Based on the comparison above, we select the extra-trees parameters with a number of estimators of 100, nodes are expanded until all leaves are pure and minimum sample split of 2.

Parameter Tuning Results

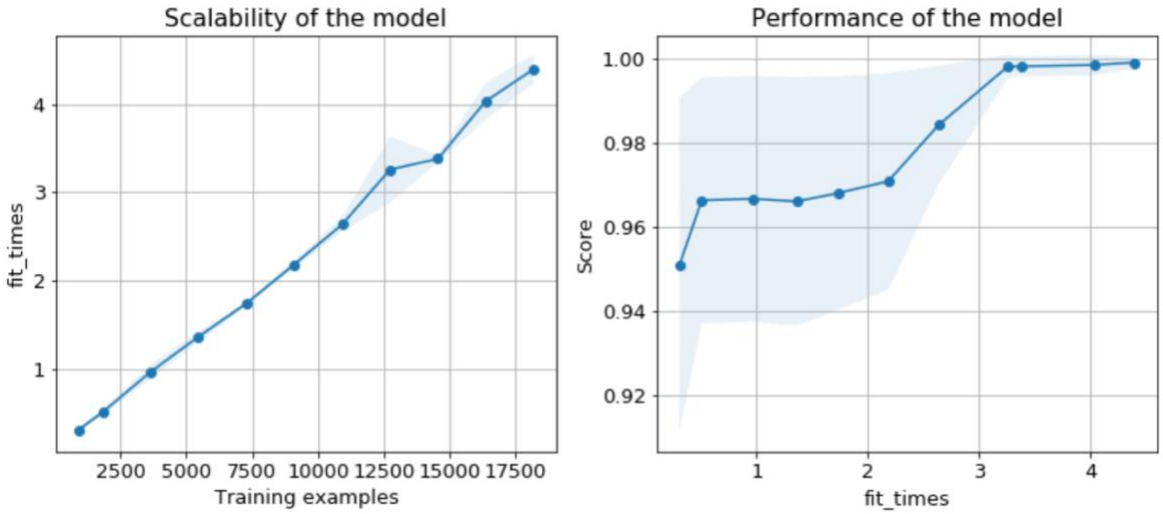
n_estimators	max_depth	min_samples_ split	Better ?	Accurate ?
100	10	2	0.65	1526, 343
100	20	2	0.65	1528, 343
100	None	2	0.707	5522, 343
200	None	2	0.706	5495, 343
200	None	3	0.704	4805, 343
500	10	2	0.65	2186, 343
500	20	2	0.66	2206, 0.665

From the below graph 11, it shows that the eight features we used to help build up the model, 'Eucl' ranks the most important. That is, the distance between the sellers and the customers rank the most important element when predicting the delivery days of the package to arrive.



Graph 11: Feature Importances

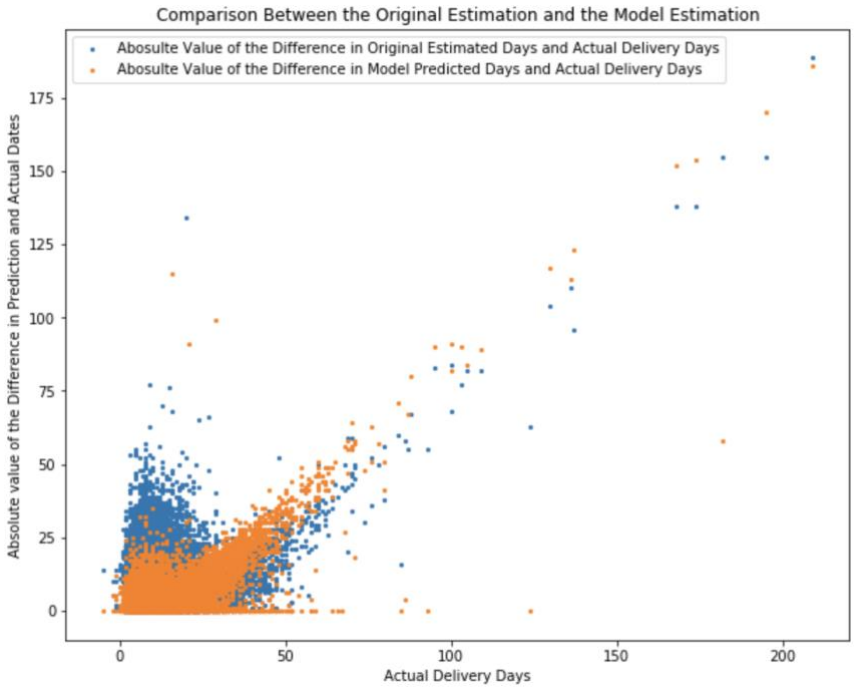
The graph 12 showing the scalability and performance of the model. As the graph shown, the more the training examples, the more time it needs to fit the model, that the relation between the training examples number and fit time are positive and they increase proportionally. Moreover, the higher the time needs to fit the model, the higher the score it has, which means that when the fit time is around 3.2, the score will reach to the highest level.



Graph 12: Scalability and Performance of the Model

(3) Model Result

By counting the number of 0 and 1s of the column we created for evaluation, we noticed that about 70% of the rows, our new model prediction works better than the e-commerce platform's predictions. Also, 5525 of the newly predicted result is the exact actual delivery date while the original estimation only accurately predicted 343 accurate results. A more direct presentation of the comparison of our model prediction and the original model is shown below (graph 11).The x-axis is the actual days the customers waited for their package to arrive, while the y-axis is the absolute value of the difference in the predicted delivery days and the actual waiting days. It shows that our model prediction are performing well than the original model prediction, since the absolute value of the difference are generally lower than the original model.

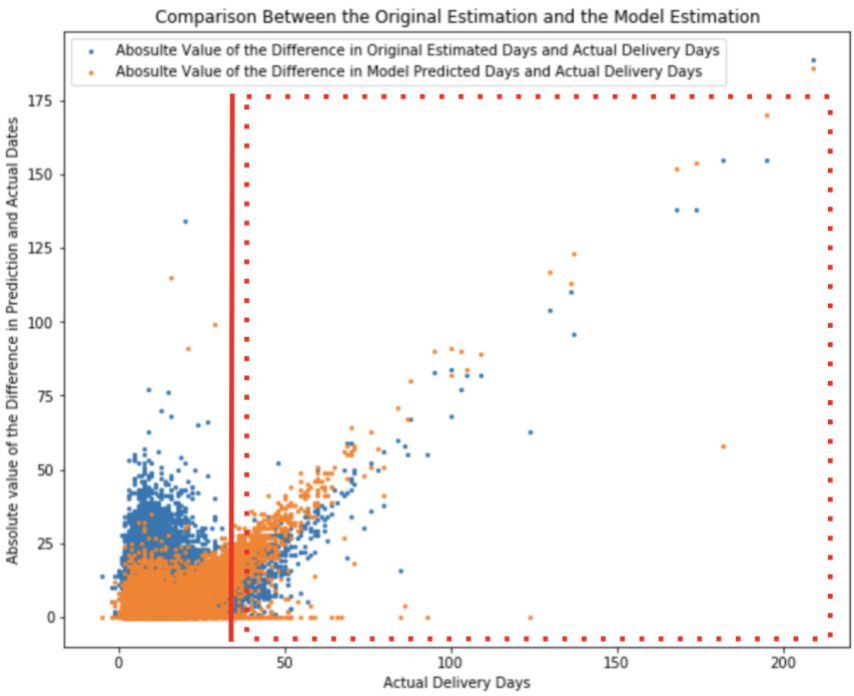


Graph 13: Comparison Between the Original Estimation and the Model Estimation

Since our prediction cannot accurately predict all samples in the testing dataset. Specifically, it might predict a delivery date that is before the actual delivery date, which reduces customers' satisfaction largely. Therefore, instead of providing a specific date, we could provide a range of dates to provide customers a flexible expectation toward their package delivery. From this project, we noticed that when we cannot completely predict the delivery date, giving a date that is later than the actual delivery date is better than an earlier date. The criteria of testing our prediction is not solely based on the accuracy but the difference between the expected and actual delivery date.

Future Work

Although our prediction performs better than the old estimation in general, it is not 100% better than the old estimation. As the below graph shows, after the actual days need to receive the package of around 30 days, the original model performs better prediction than our model. Thus, it may need to conduct further research on the actual logistics situation.



Graph 14: Problem on the Comparison Between the Origin and Model Estimation

To gain a more accurate model, we would: first, acquire more feature inputs from the logistic providers. For instance, we may need real logistics distance between the sellers, logistics center and the customers instead of using euclidean distance to conduct the distance analysis between the sellers and the customers. Second, we would need to acquire more information from local logistic services to help perform a better understanding of the actual logistics situation. For instance, how many distribution centers are there in the region of the consumer and the necessity of establishing more logistics centers in specific areas. Third, make good use of the customers review data to conduct sentiment analysis to analyze how the logistics issue exists in different customers. Moreover, by conducting sentiment analysis, it is possible to perform a more specific overview of the customer's thoughts toward their shopping experience with this e-commerce platform. Last but not least, it should be noticed that even though we consider the logistics is an important element in deciding the customers' experience evaluation, we are not assigning single attributes here. Besides, other elements like the product quality, services level are also important during the evaluation of the customers toward this e-commerce platform. Therefore, what we provide here is a better prediction of the delivery days based on the analysis that the inaccurate estimated days have a tendency to decrease the score evaluated by the customers to some extent.