Big Data Analytics Project
-Part II-

❑ **Dataset Selection:** Brazilian E-commerce Company Olist from Kaggle

❑ **Project Topic:** Analysis of the E-commerce Business of Olist on Optimizing Logistics Solutions and Increasing Customer Experiences

Group Member:
Jiawei Huang
Jingting Xu
Qiaochu Cong
Qiurong Ren

Contents

1.Overview of Problem

2.Modeling Approaches

3.Challenges & Approaches

4.Analysis Results

5.Insights Gained & Future Work

Overview of Problem

Briefly describe problem, why it is important and interesting

Problem:

Inaccurate estimated delivery date leads lower review scores

Goals:

Develop a model that predict a better and more accurate delivery days.

Modeling Approaches

Describe modeling approaches

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

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, 343

Challenges & Approaches

Describe challenges with modeling and approaches to address challenges

Problem:

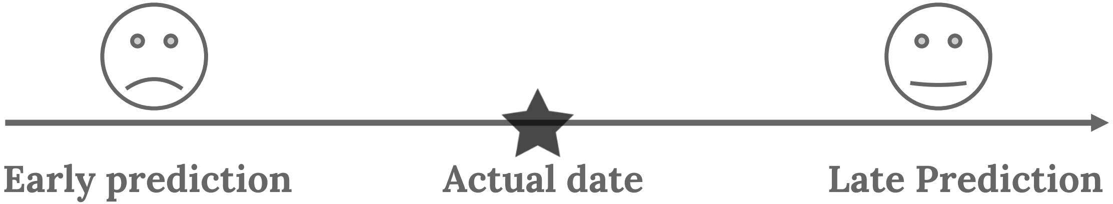
The default model accuracy is not an effective measure of our result. So, method like gridsearch can’t help us to find the best model.

Solution:

Step 1 Two columns are generated: “Better” & “accurate”

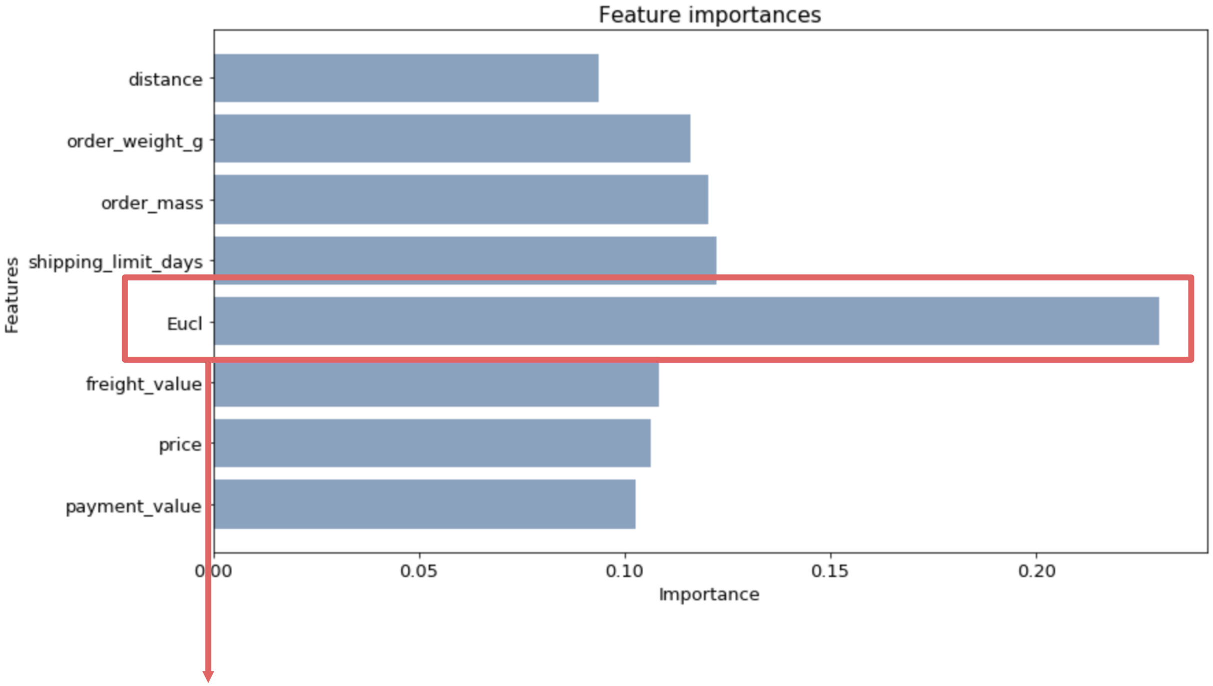
Step 2 Manually adjust parameter

Logic Example:



Analysis Results

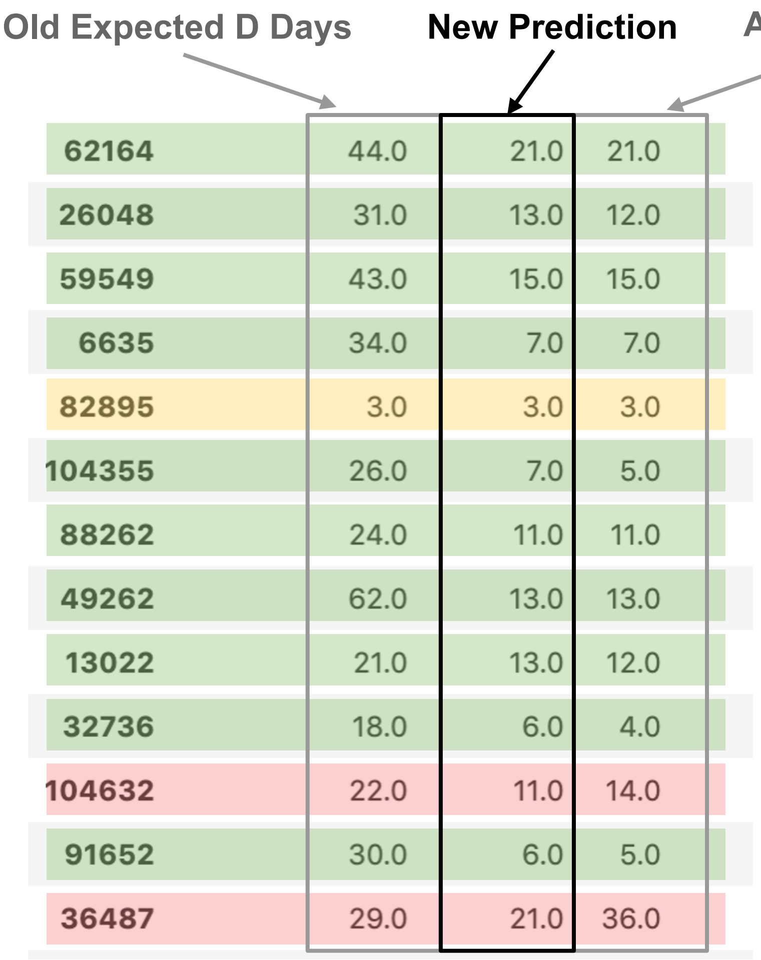
Present analysis results



Euclidean distance between customers and sellers

Analysis Results

Present analysis results



	Old Expected D Days	New Prediction	Actual D Days
62164	44.0	21.0	21.0
26048	31.0	13.0	12.0
59549	43.0	15.0	15.0
6635	34.0	7.0	7.0
82895	3.0	3.0	3.0
104355	26.0	7.0	5.0
88262	24.0	11.0	11.0
49262	62.0	13.0	13.0
13022	21.0	13.0	12.0
32736	18.0	6.0	4.0
104632	22.0	11.0	14.0
91652	30.0	6.0	5.0
36487	29.0	21.0	36.0

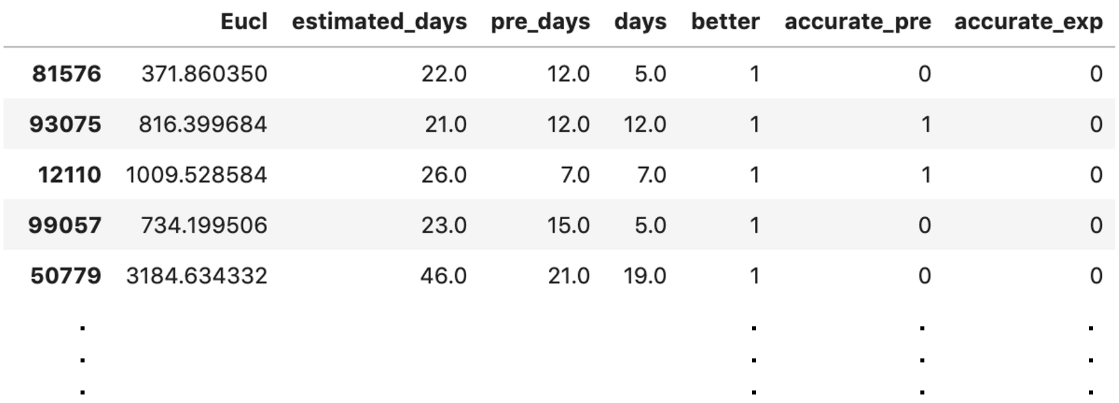
🟩 Our prediction is Better

🟥 Original Estimation is Better

🟨 Same Performance

Analysis Results

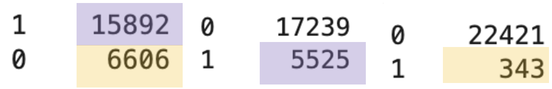
Present analysis results



	Eucl	estimated_days	pre_days	days	better	accurate_pre	accurate_exp
81576	371.860350	22.0	12.0	5.0	1	0	0
93075	816.399684	21.0	12.0	12.0	1	1	0
12110	1009.528584	26.0	7.0	7.0	1	1	0
99057	734.199506	23.0	15.0	5.0	1	0	0
50779	3184.634332	46.0	21.0	19.0	1	0	0
⋮					⋮	⋮	⋮

🟪 New Prediction

🟨 Old Prediction

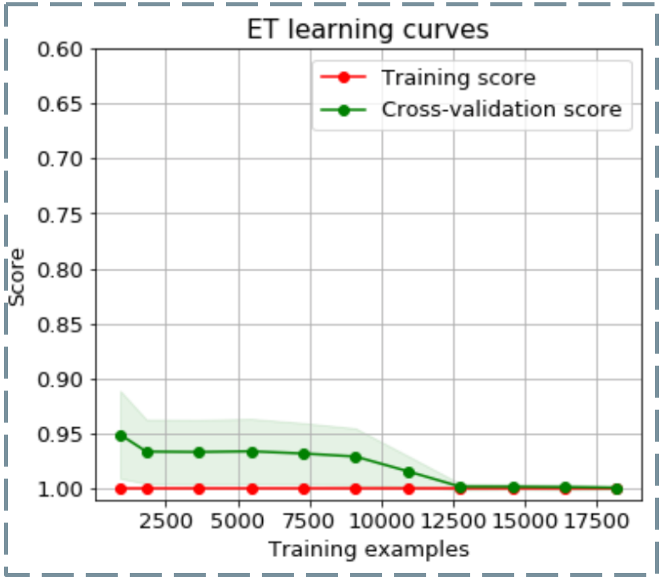


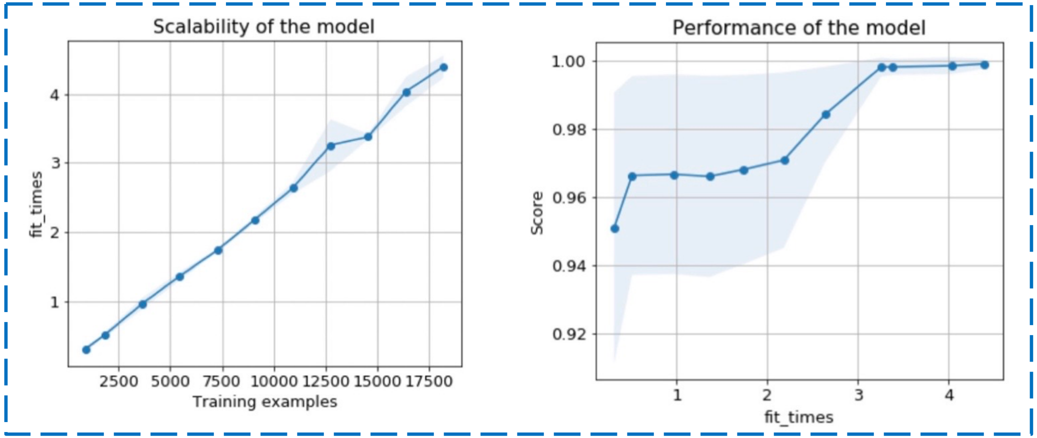
```
1   15892   0   17239   0   22421
0    6606   1    5525   1     343
```

- ❑ Testing dataset length: **22764 rows**
- ❑ **70%** of new prediction works better than old estimation.
- ❑ **5525** predicted result is the exact actual delivery date while the old estimation only accurately predict **343** accurate results.

Analysis Results

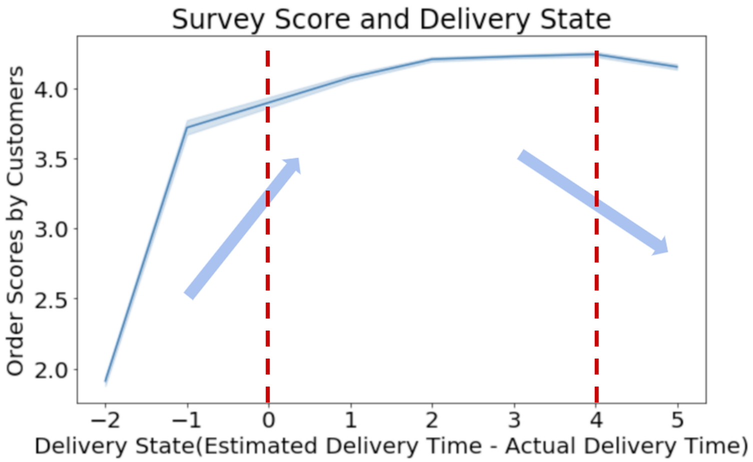
Present analysis results

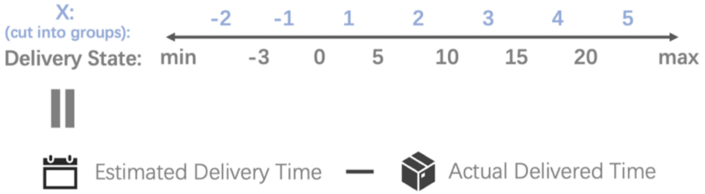




Insights Gained & Future Work

Discuss insights gained and future works




X:
(cut into groups): -2 -1 1 2 3 4 5

Delivery State: min -3 0 5 10 15 20 max

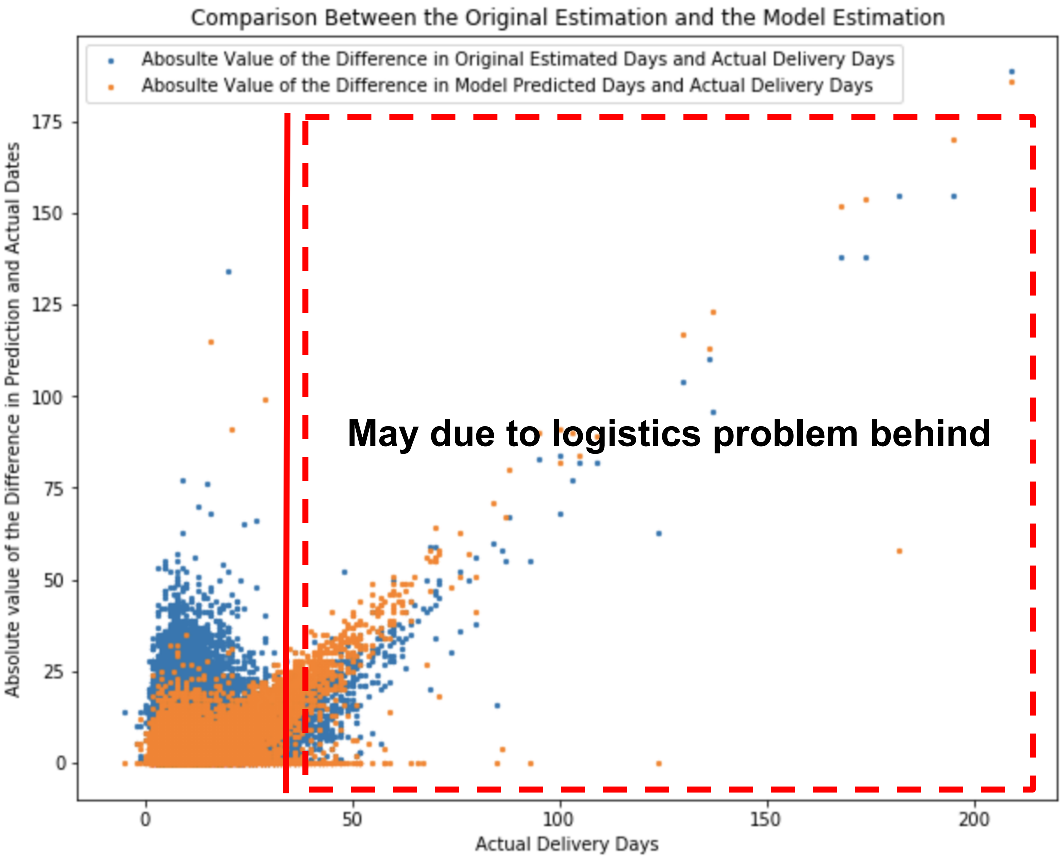
Estimated Delivery Time — Actual Delivered Time

Insights Gained:

- instead of providing an actual delivery date of the package, it is better to **provide a range of the estimated delivery days.** Since it is estimated that customers prefer receive their package late than or at the exact estimated delivery date

Insights Gained
&
Future Work

Discuss insights gained and future works



May due to logistics problem behind

Insights Gained
&
Future Work

Discuss insights gained and future works

Future Work:

- Acquire more considerations on the orders have delayed dates more than 30 days, which may exist logistics problem (like having not enough logistic centers)
- Acquire more feature inputs from the logistic providers (like the real logistics distance between the sellers, logistics center and the customers, instead of the current usage of calculating euclidean distance)
- Sentiment analysis

Sources : https://www.istockphoto.com/vector/global-logistics-network-concept-communications-network-map-brazil-on-the-world-gm1055228636-281963048

Thank You

&

Q&A