Telerik Academy

Understanding Customer Satisfaction

Team:

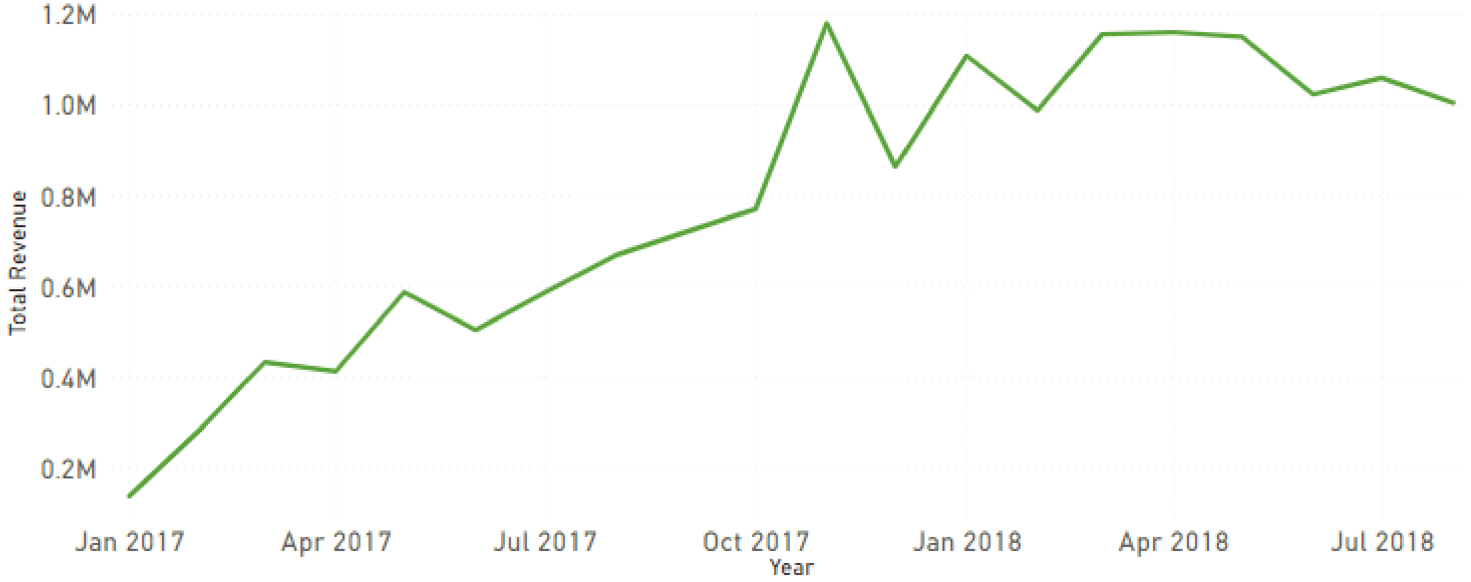
Nelly Dyulgerova

Stiliyana Bachovska

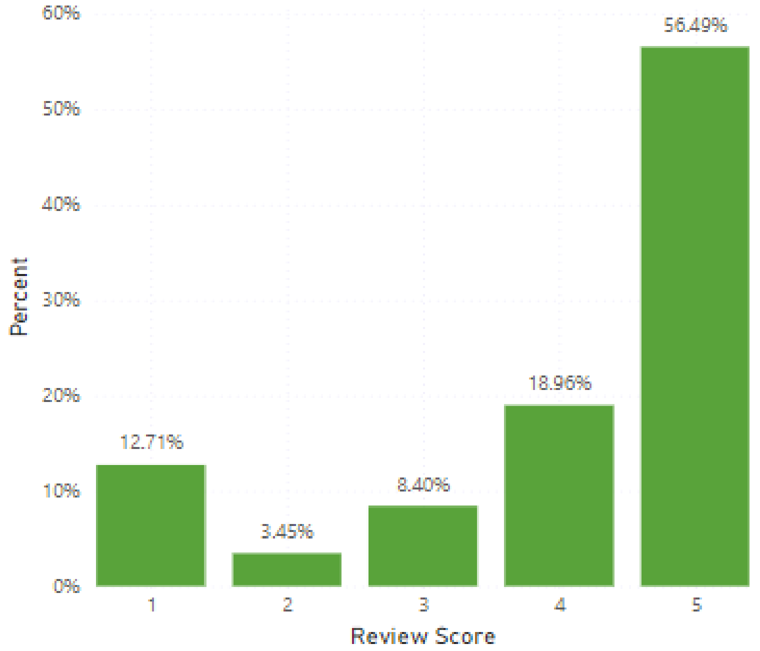
Veselin Georgiev

July 2024

Olist Marketplace

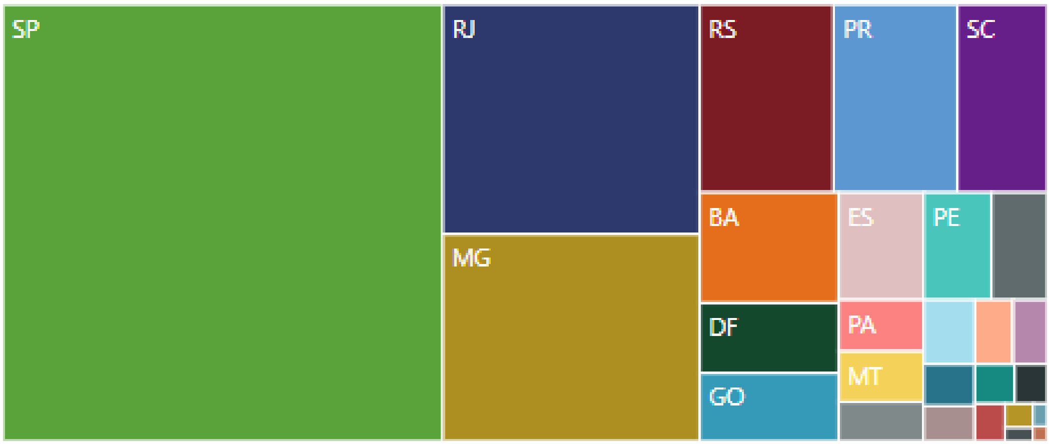

Total Revenue by Year and Month

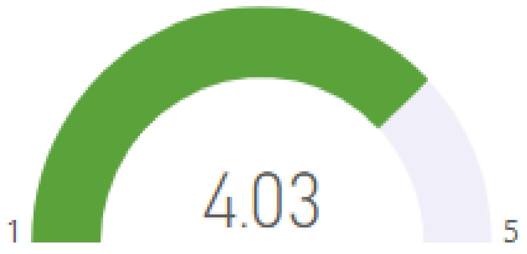
Period included from Jan 2017 to Aug 2018)


Number of Review Scores, as % of total

15.84M

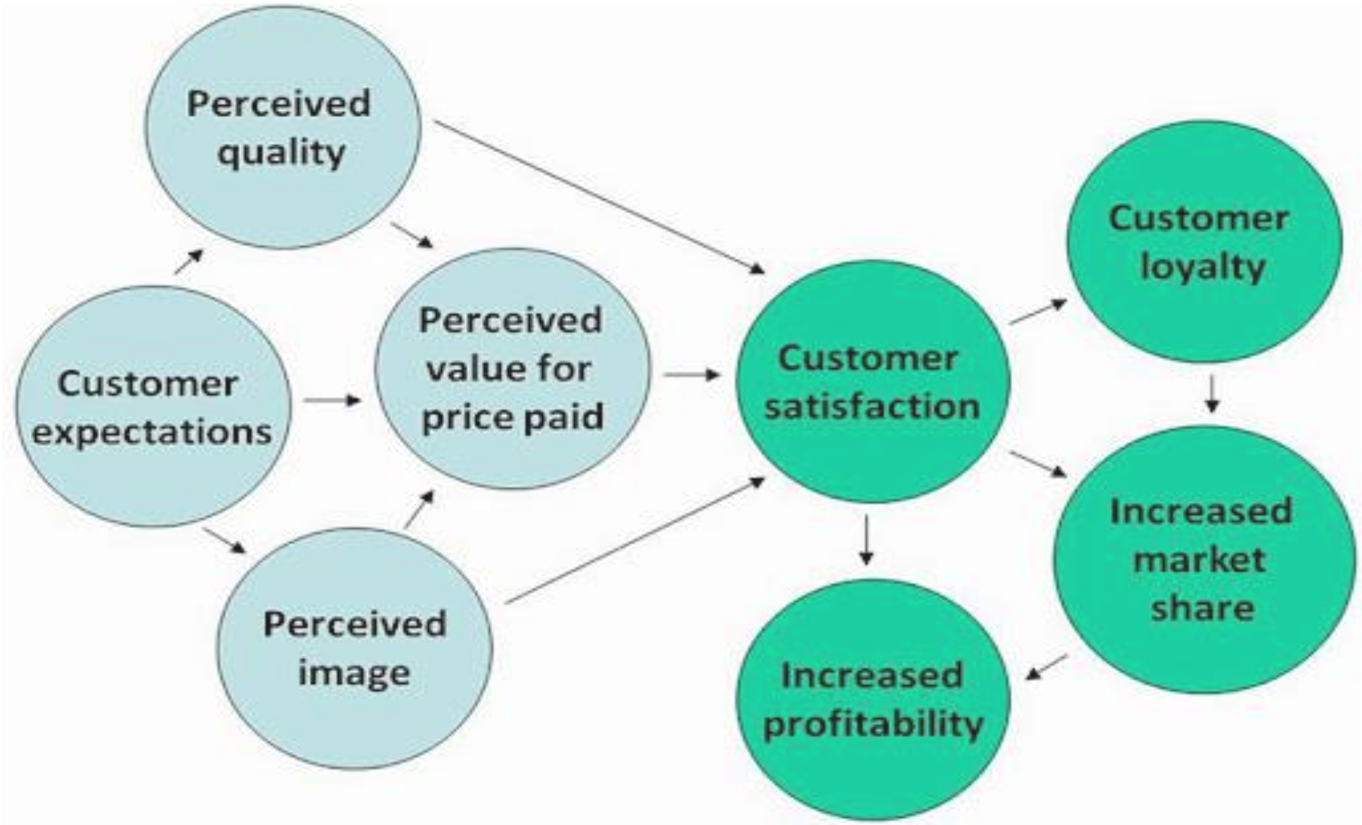
Total Revenue


Customer by State


Average Rating

Telerik Academy

What is Customer Satisfaction (CSAT)?



Common factors:

- product experience
- delivery experience
- price sensitivity
- perceived value
- price
- promotion
- product quality

CSAT is a measurement that determines how happy customers are with a company’s products, services, and capabilities.

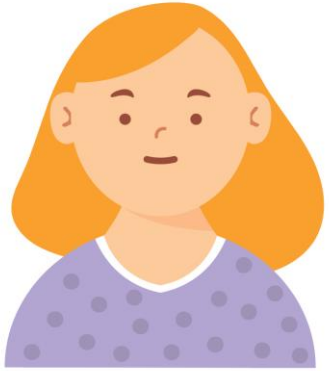
Surveys and ratings can help a company determine how to best improve or changes its products and services

What our customers say when they are not happy

My order came 7 days late and without a notice for the delay

Theo

I received only part of the items I have ordered



Helena

I ordered a mobile phone and I later understood that the price was significantly higher than in similar online stores

Laura

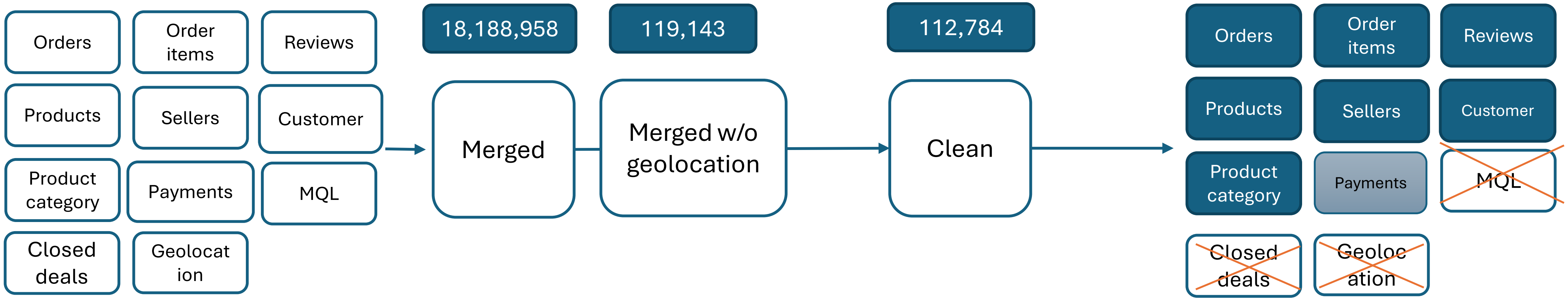
The product takes a long time to arrive!

Miguel

The product came differently than what is shown in the photo, it is not luxury, nor padded as it said.

Bernardo

Data preprocessing: Where and Why



Orders
Order items
Reviews
Products
Sellers
Customer
Product category
Payments
MQL
Closed deals
Geolocation

18,188,958
Merged

119,143
Merged w/o geolocation

112,784
Clean

Orders
Order items
Reviews
Products
Sellers
Customer
Product category
Payments
~~MQL~~
~~Closed deals~~
~~Geolocation~~

Data Quality Issues:

1. Inconsistent data:
 - mismatch (dates <> delivery status, timestamps, status unavailable)
2. Dates are set as string
3. Missing data
4. Typos
5. Not meaningful information
6. Zip code mismatches
7. Review score and comment mismatch

Data Quality Issues:

1. Missing data
2. Creation of duplications

Cleaning

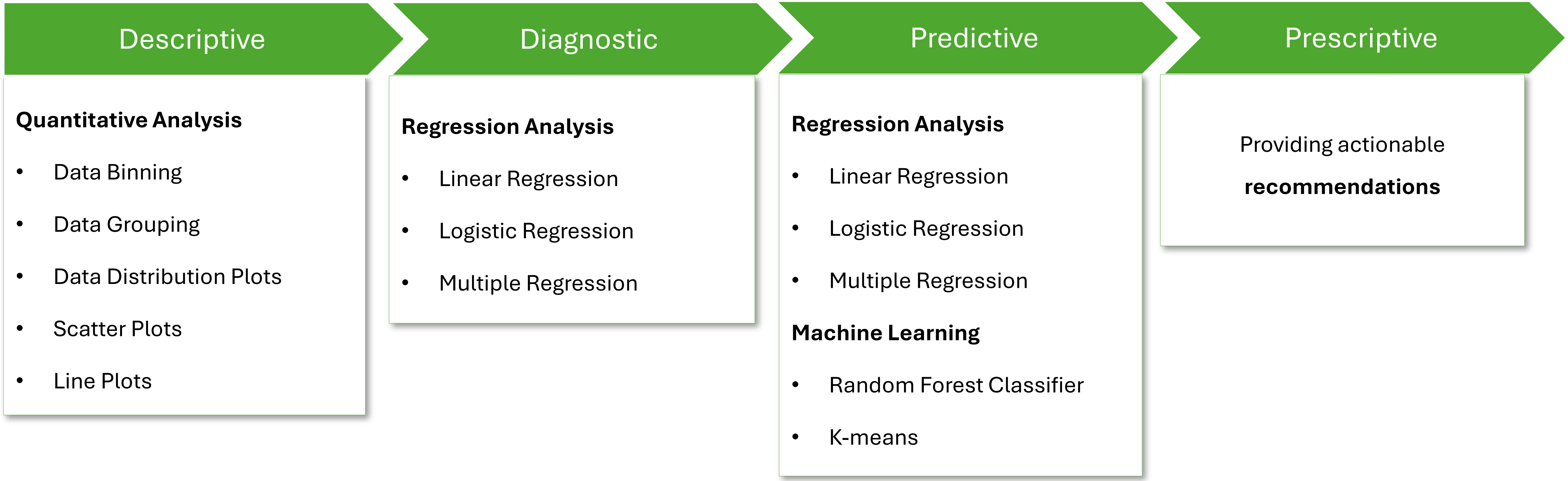
1. Removed geolocation
2. Removed inconsistent data: dates <> delivery status, status 'unavailable'
3. Replaced n/a with 'other' for product category (in SQL)
4. Corrected typos in Python
5. Changed string types as date (in Python)
6. Removed duplications based on review: only last review is kept

Final Datasets

1.Denormalized Table (7 datasets)
2.Denormalized Table (2 datasets)

Telerik Academy

Methodology: Approach



Descriptive

Quantitative Analysis

- Data Binning
- Data Grouping
- Data Distribution Plots
- Scatter Plots
- Line Plots

Diagnostic

Regression Analysis

- Linear Regression
- Logistic Regression
- Multiple Regression

Predictive

Regression Analysis

- Linear Regression
- Logistic Regression
- Multiple Regression

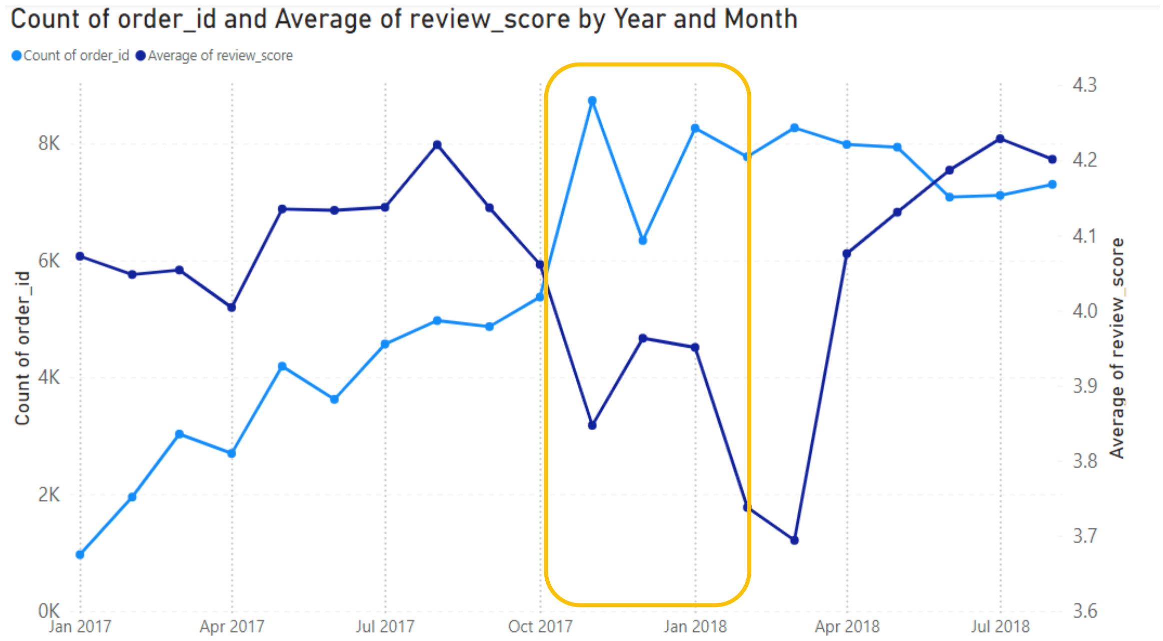
Machine Learning

- Random Forest Classifier
- K-means

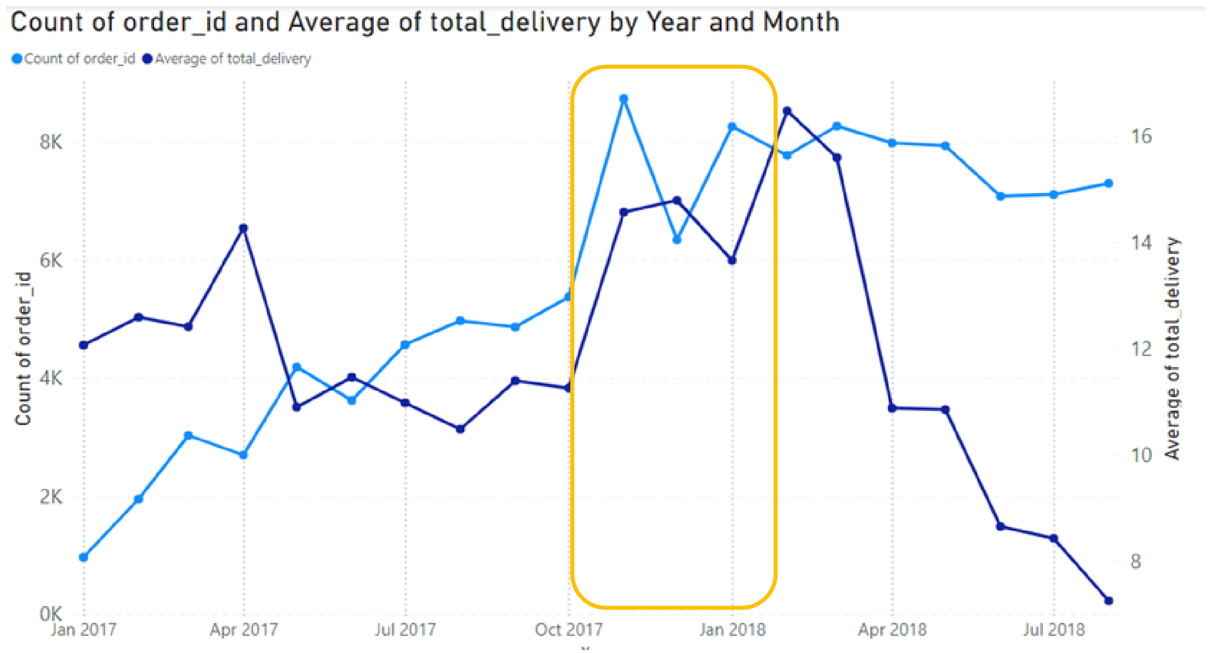
Prescriptive

Providing actionable **recommendations**

Olist Orders



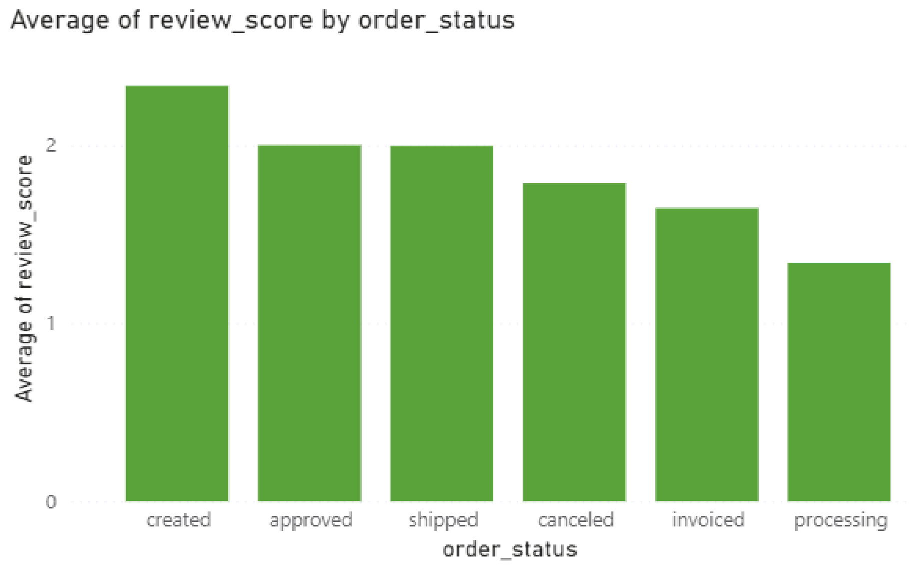
*outliers have been removed (period included from Jan 2017 to Aug 2018)

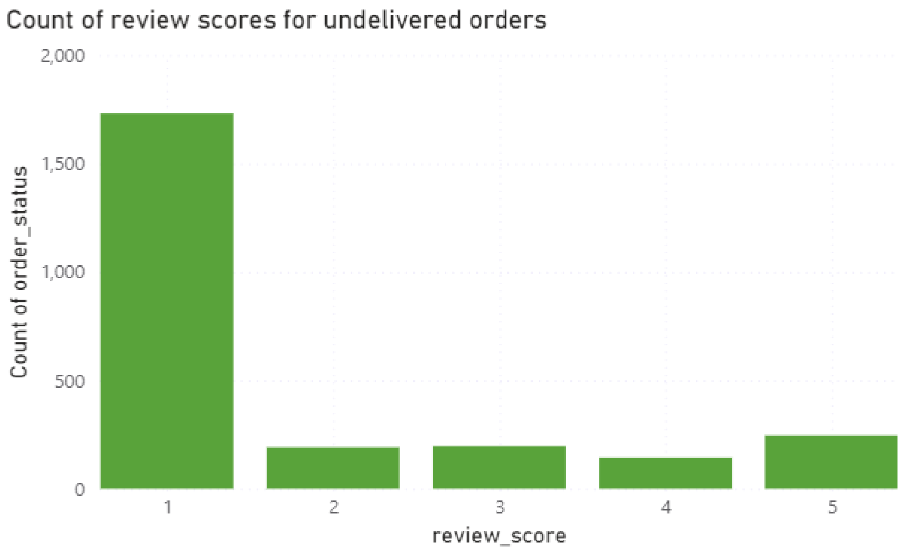


- Business Growth Effect

- Peak in the ordered items around Christmas and New Years (24 Dec – 1 Jan) holidays and Carnival (February)
  - **Reviews** are one of the **lowest** and **delivery time** is one of the **highest**.

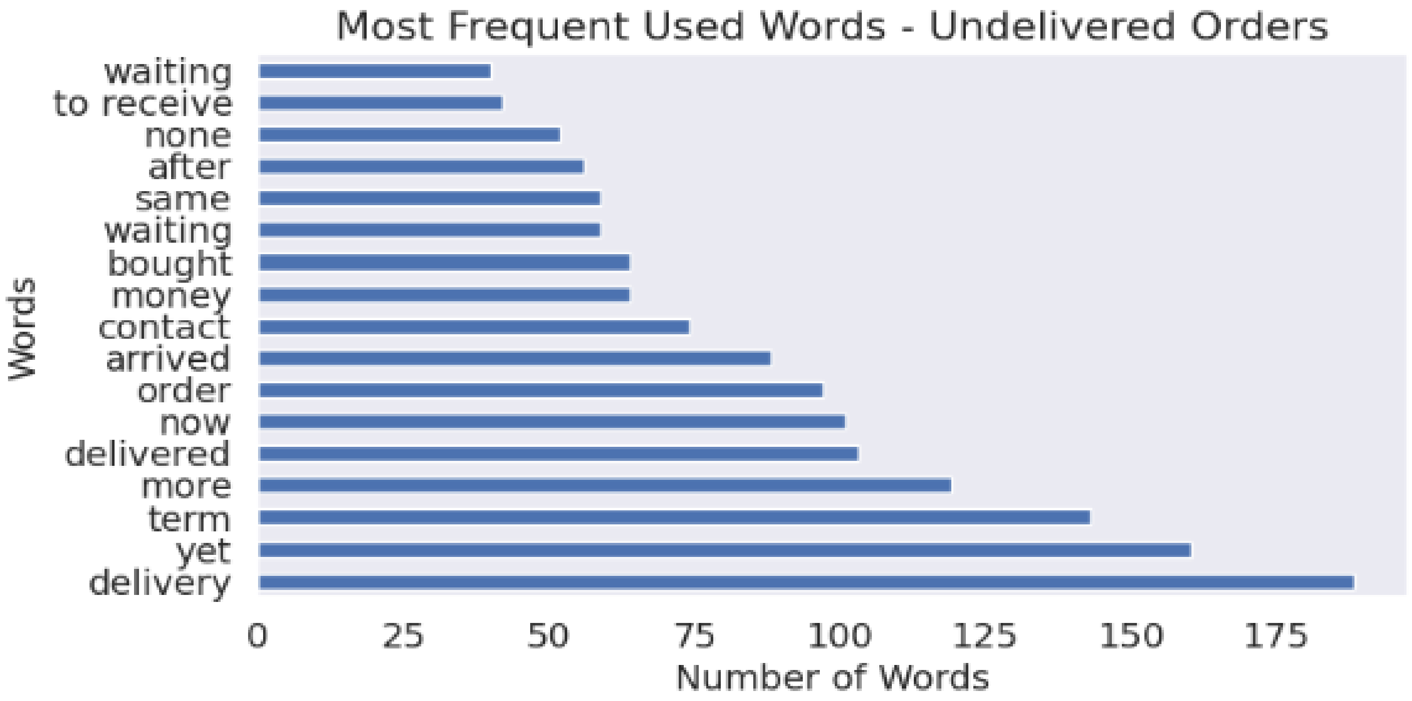
Olist Order Status





- **85% of undelivered** orders are with review score less than 4.

- In **66%** of the **review comments** contain the words “delivery, waiting, arrived, to receive, more”



Methodology: Assumptions

- Review score is transformed in binary format **0/1**
  - **>= 4** ⟹ "**1**" (satisfied)
  - **< 4** ⟹ "**0**" (not satisfied)
- Calculations of days:
  - **Order approval** time = Order approved at – Order purchase timestamp
  - **Carrier delivery** time = Order delivered carrier date – Order approved at
  - **Customer delivery** time = Order delivered customer date – Order delivered carrier date
  - **Total delivery** time = Order delivered customer date – Order purchase timestamp
  - **Delivery delay** = Order delivered customer date – Order estimated delivery date

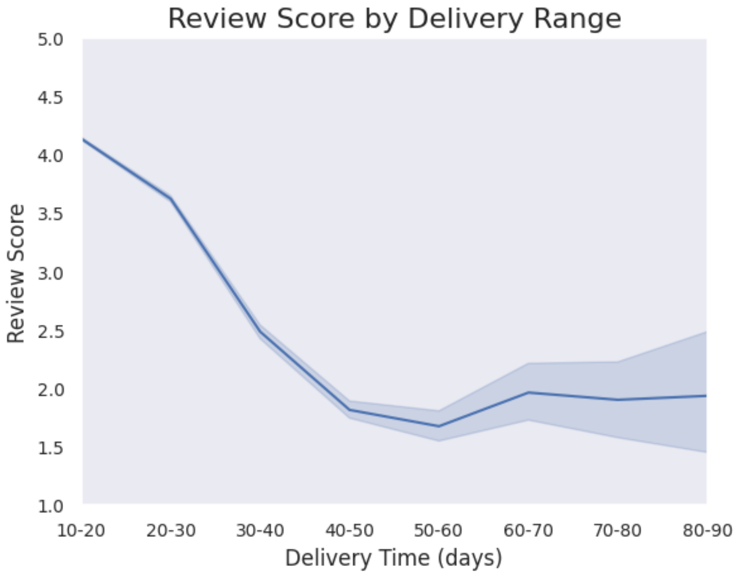
Methodology: Assumptions

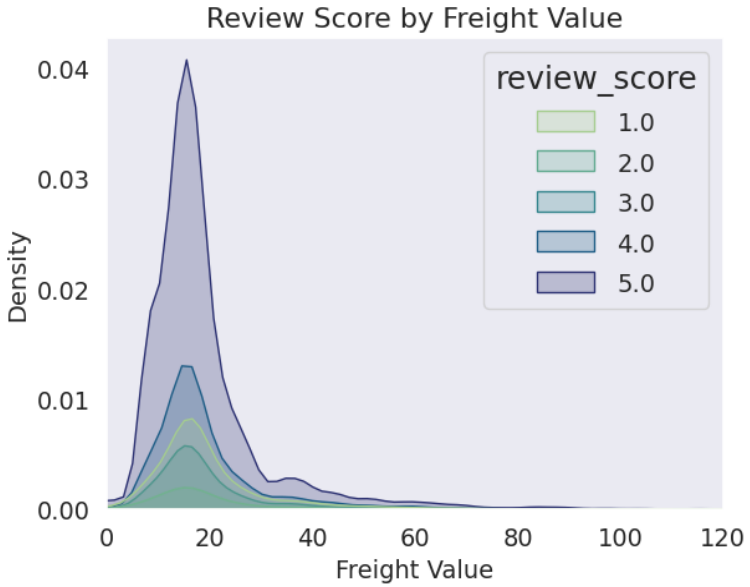
Seller and Product Categories Total Scores are created by rating the category against various criteria, which are ranked and rated on a scale of 1 to 5.

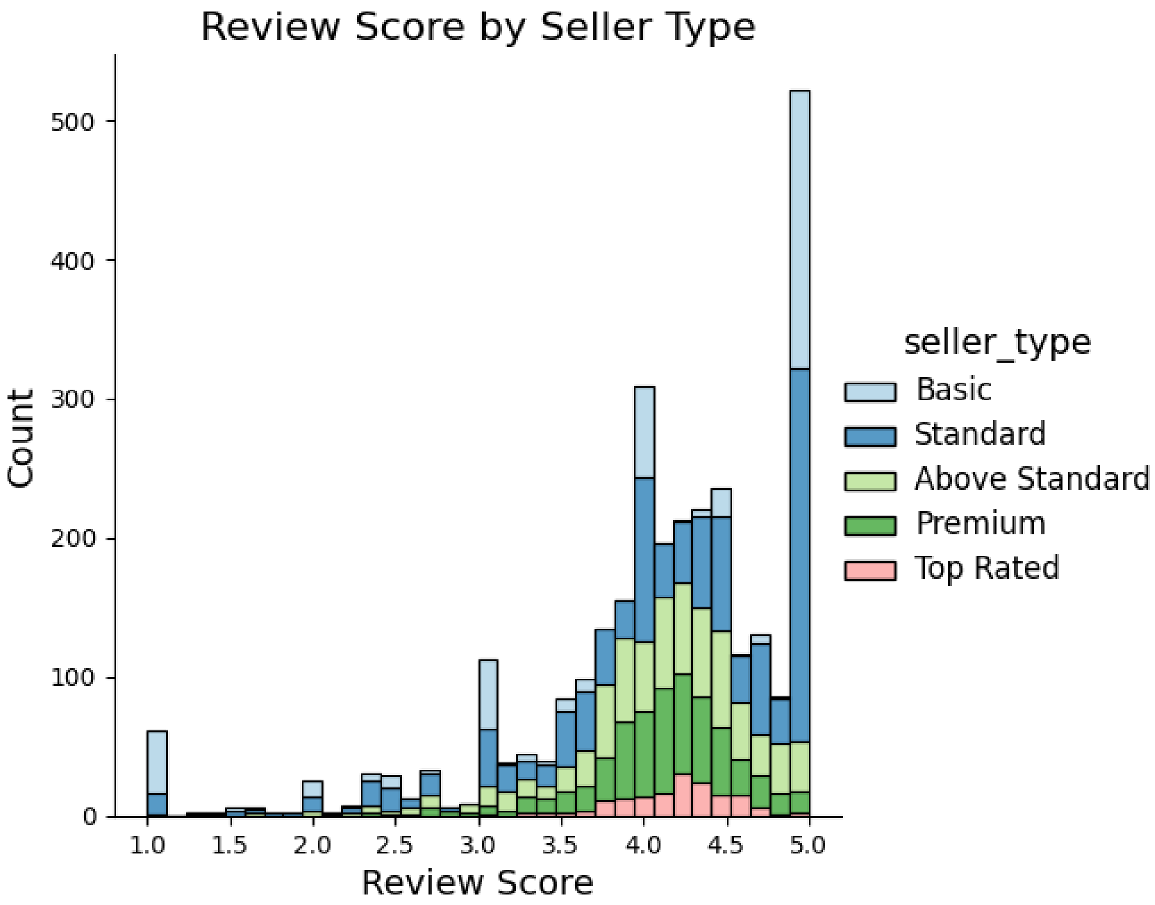
Seller id	Order amount score	Revenue score	Delivery score	Time as seller score	**Seller total score**	Seller type
0015a82c2db000af6aaaf3ae2ecb0532	2	5	3	2	**3.00**	**Above Standard**
001cca7ae9ae17fb1caed9dfb1094831	5	4	2	5	**4.00**	**Premium**

Product Category Name	Order amount score	Product description score	Product photos score	Revenue score	**Product Category total score**	Category Type
air conditioning	3	4	3	5	**3.75**	**Premium**
art	2	4	2	3	**2.75**	**Standard**

Quantitative Analysis: Other factors


Review Score by Delivery Range

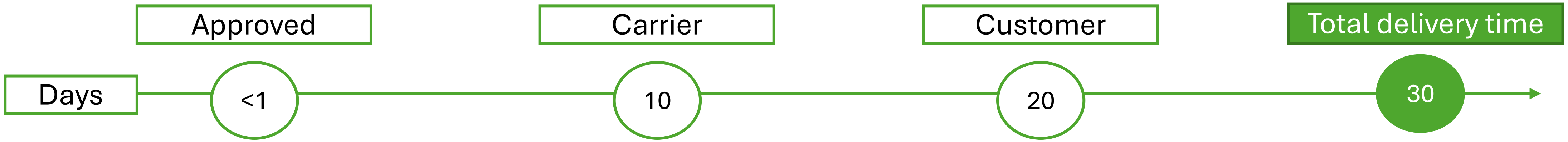

Review Score by Freight Value


Review Score by Seller Type

Factors that could be further analyzed:

- Total Delivery Time
- Delay days
- Order Size
- Freight Value
- Seller Type
- Product Category Type
- Product Weight
- Price

Regression Analysis - Total Delivery Time



Approved

Carrier

Customer

Total delivery time

Days

<1

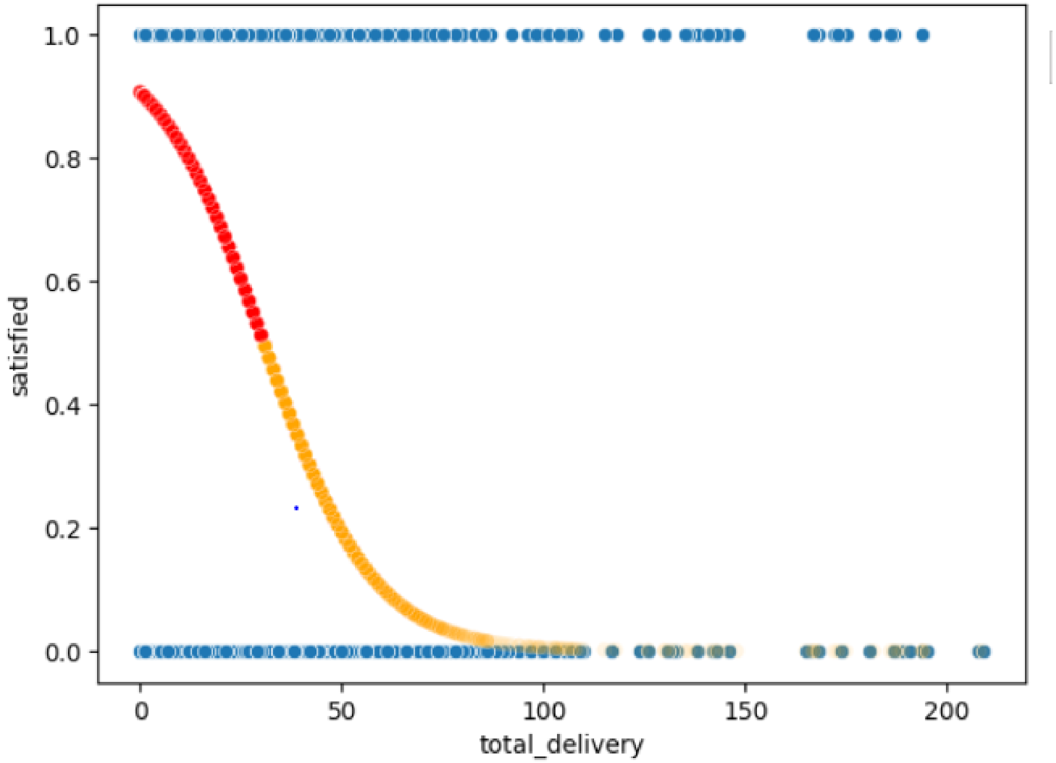
10

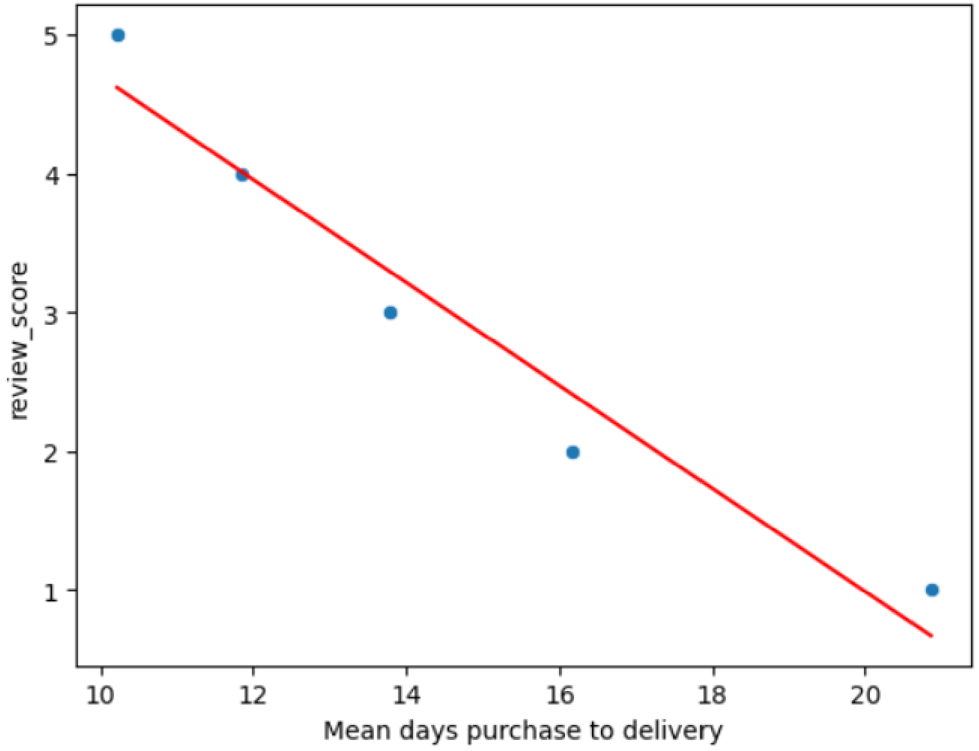
20

30

- After **~30 days** total delivery time, the customer will become unsatisfied.

- Each additional delivery day reduces the expected review score by **0.37 points**





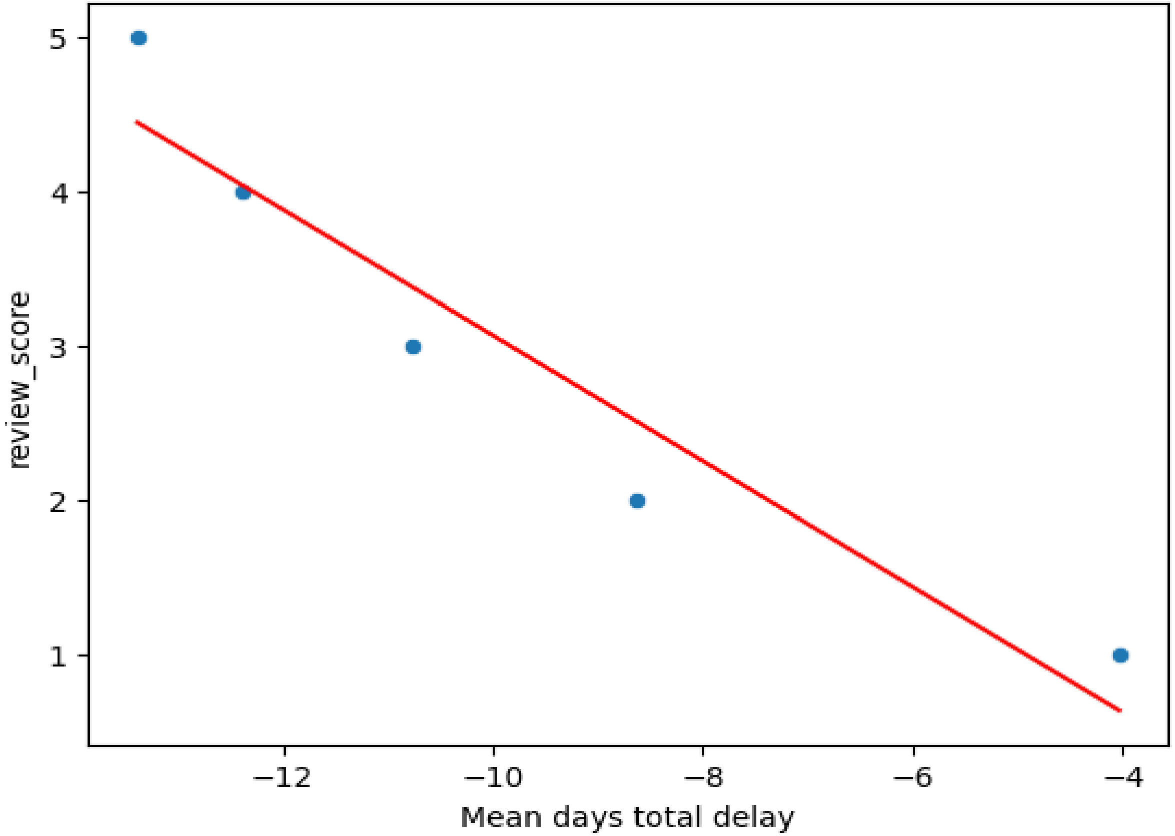
***Red/Orange Colors** - 'satisfied' / 'unsatisfied' customers*

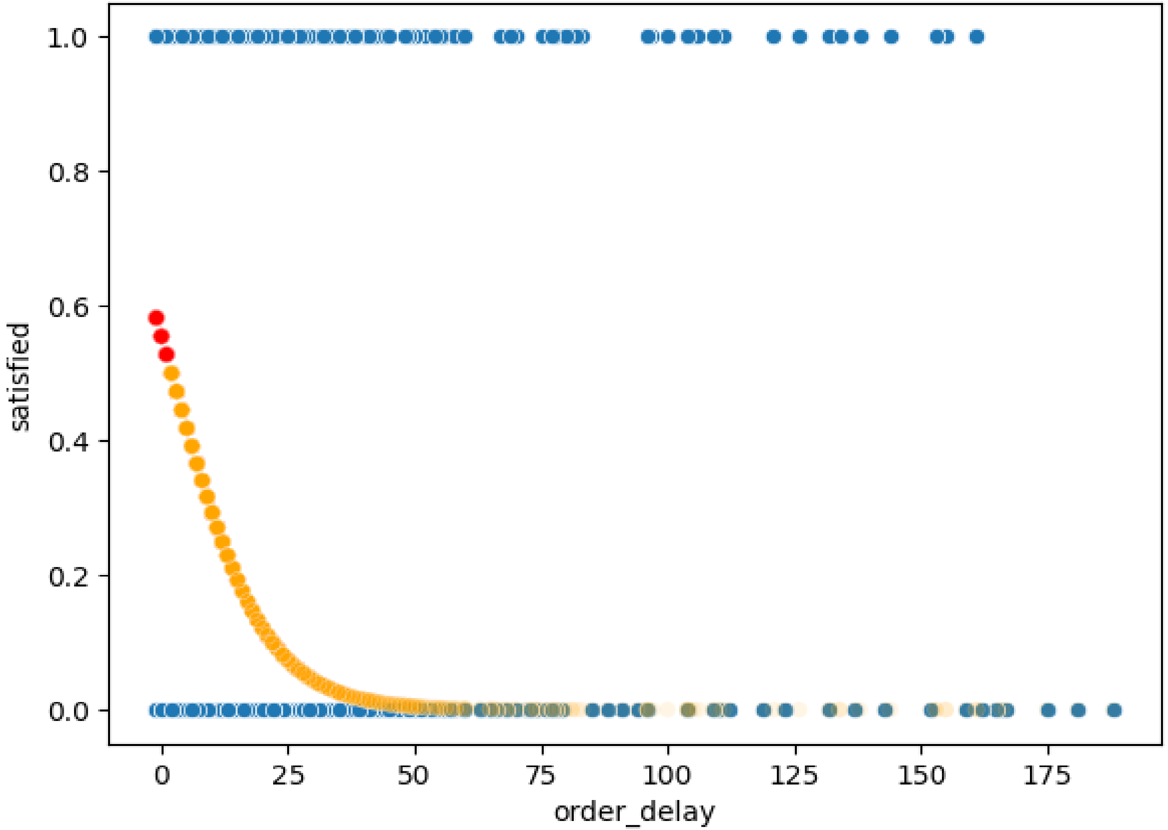
Telerik Academy

Regression Analysis - Delivery Delay

- Each day additional delay reduces the review score by **0.76 points.**

- After **~5 days** of delay, customer satisfaction starts dropping



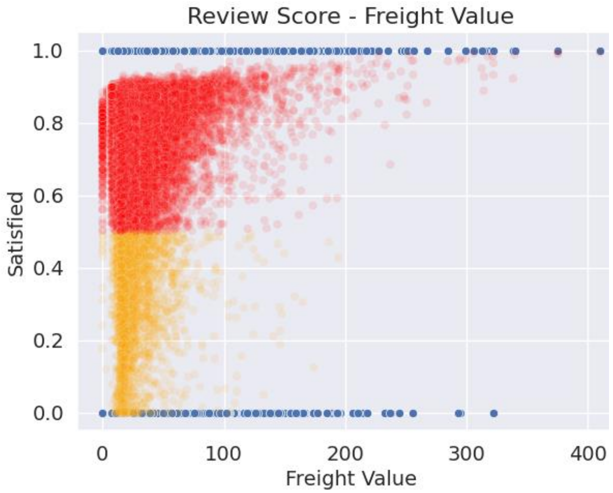


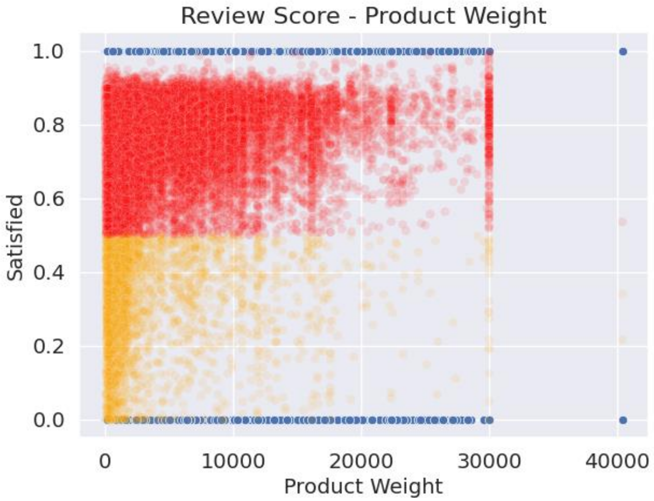
***Red/Orange Colors** - 'satisfied' / 'unsatisfied' customers*

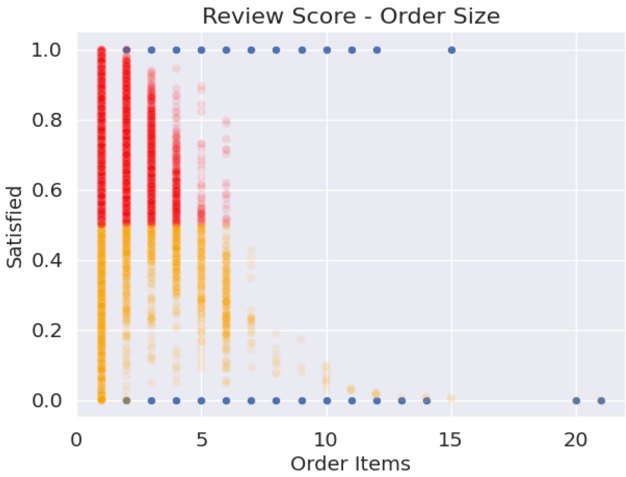
Telerik Academy

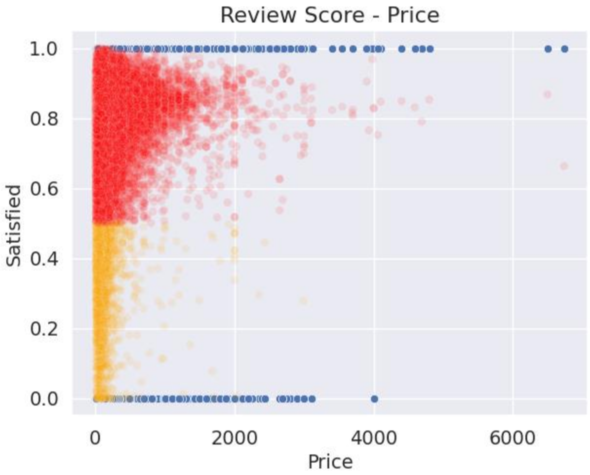
Regression Analysis (Logistic): Multiple factors

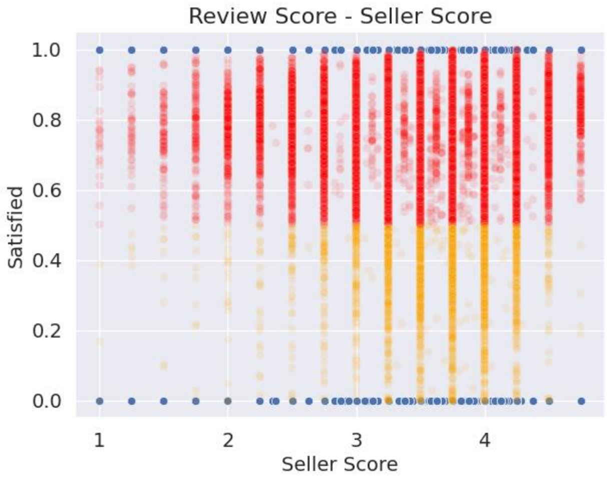
- Results **do not show strong correlation** between the dependent variable and the studied factors
- Signs of **correlation** are visible and **further research is needed**

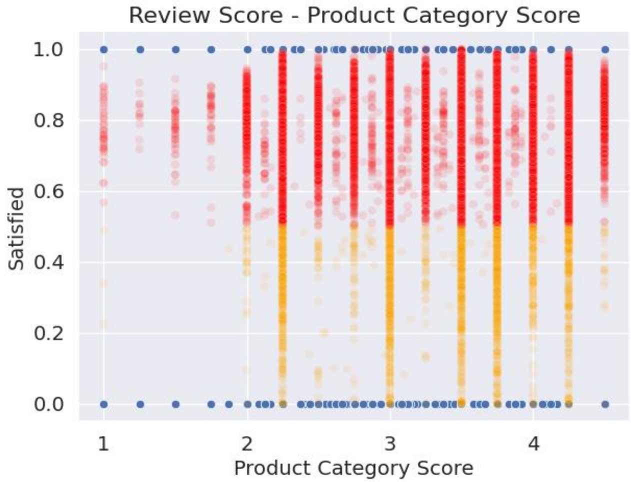

Review Score - Freight Value


Review Score - Product Weight


Review Score - Order Size


Review Score - Price


Review Score - Seller Score

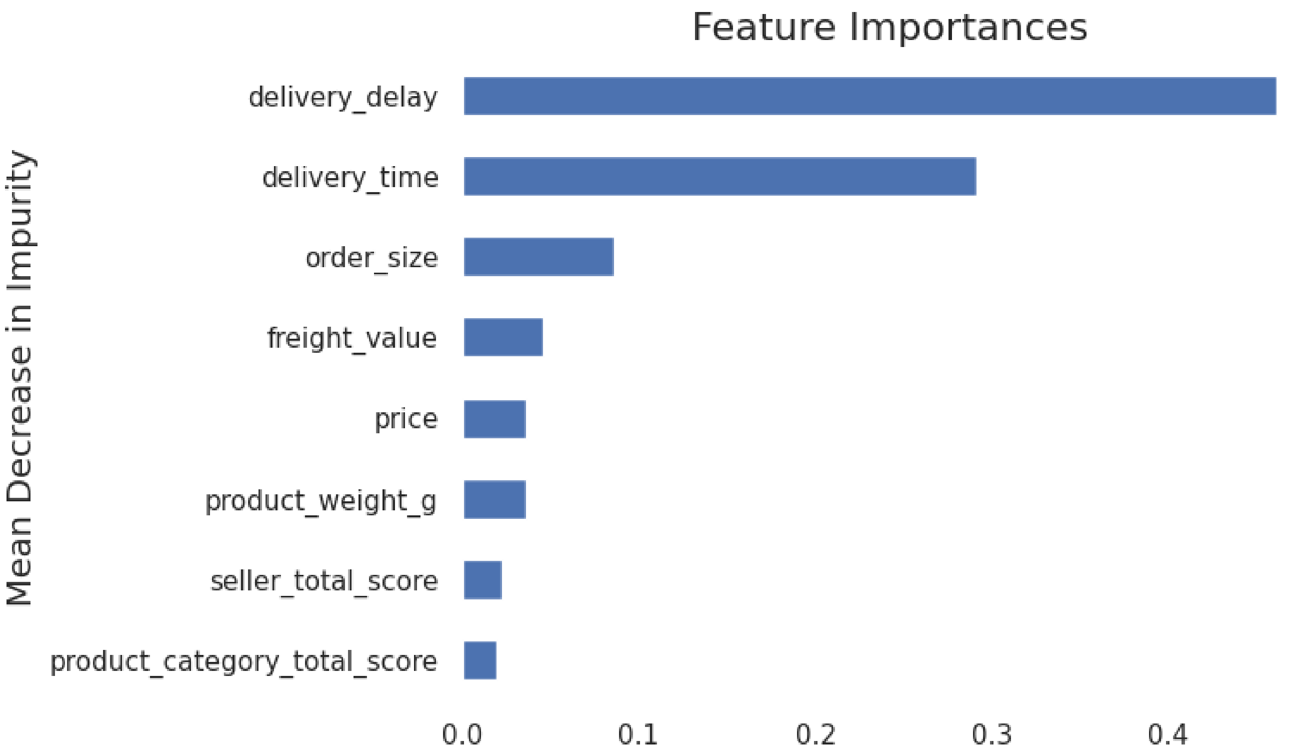

Review Score - Product Category Score

***Red/Orange Colors** - predictions threshold - 'satisfied' / 'unsatisfied' customers

Telerik Academy

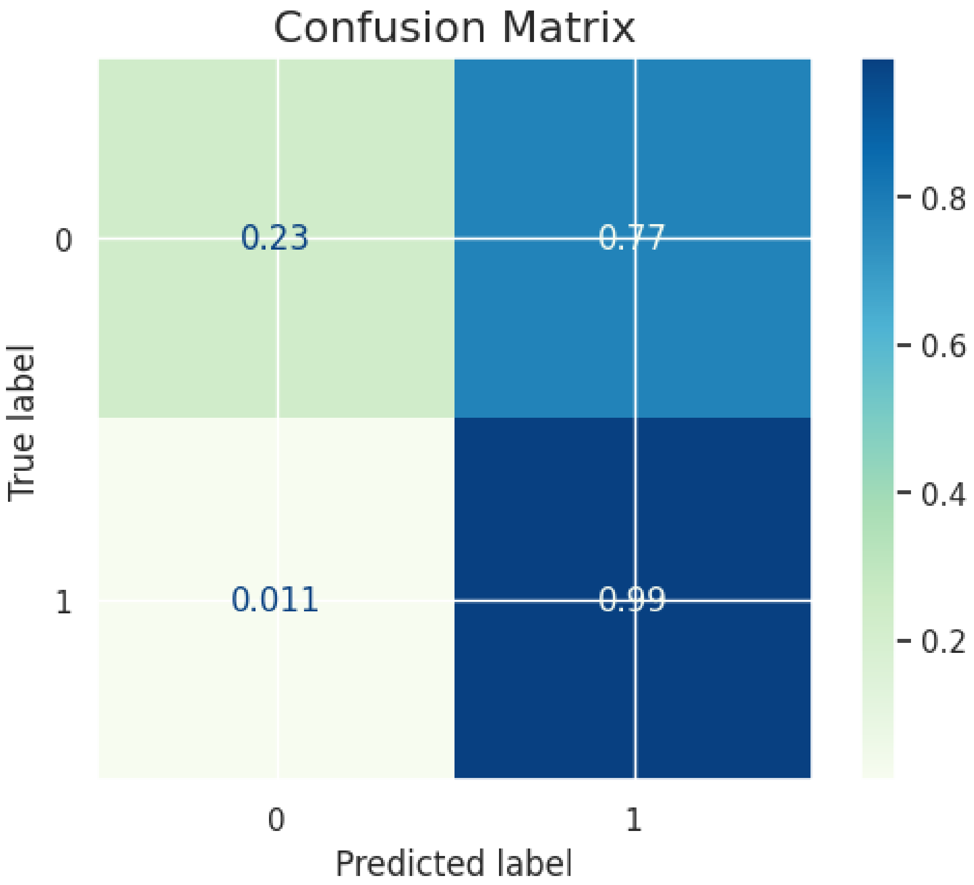
Machine Learning: Random Forest Classifier

RFC Accuracy Score: **0.83**

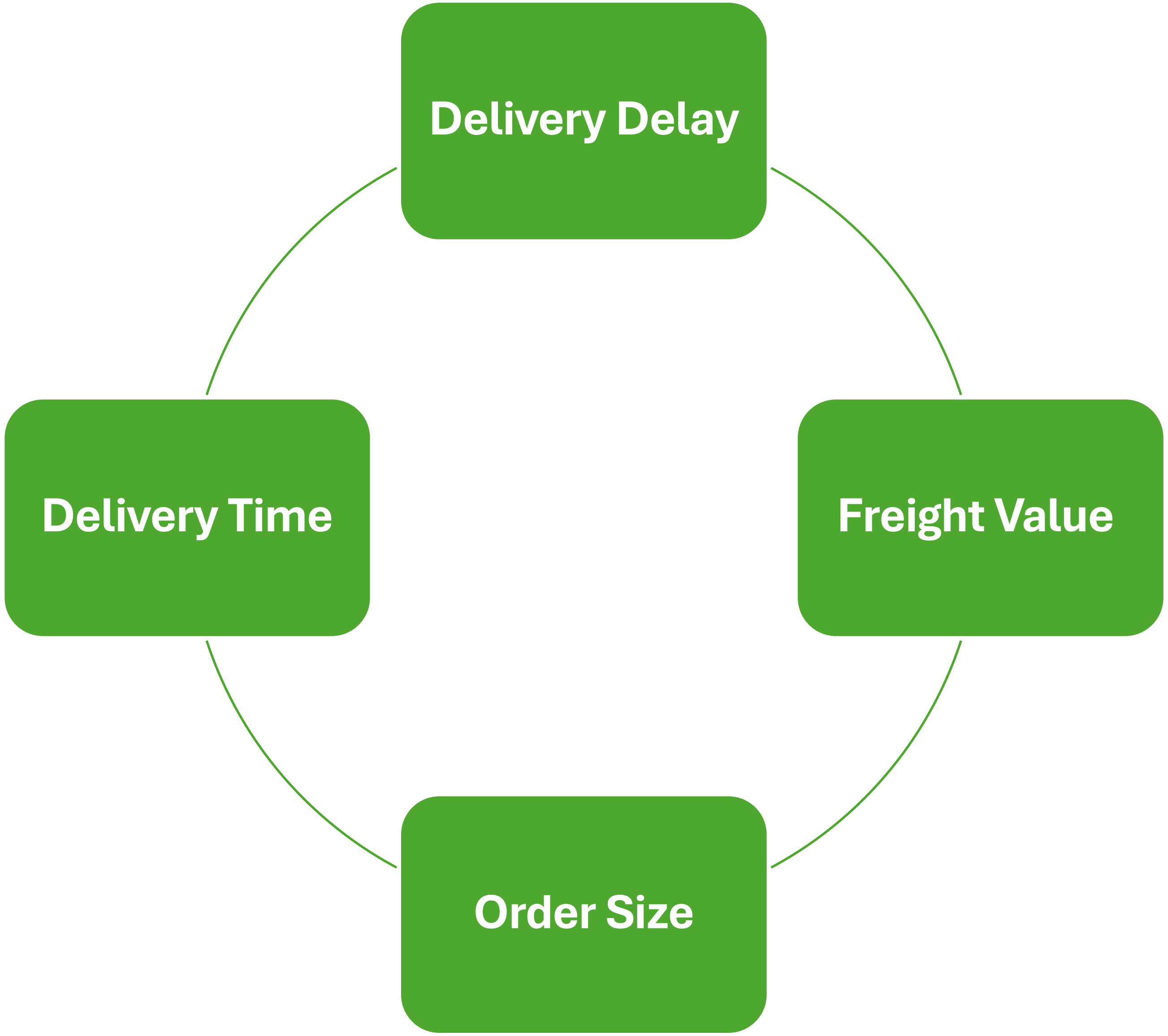


*Binary variable 0/1 is used for satisfied/unsatisfied client (review score >= 4)

61% chance for a satisfied/unsatisfied customer to be **classified properly**.



Results: Main drivers



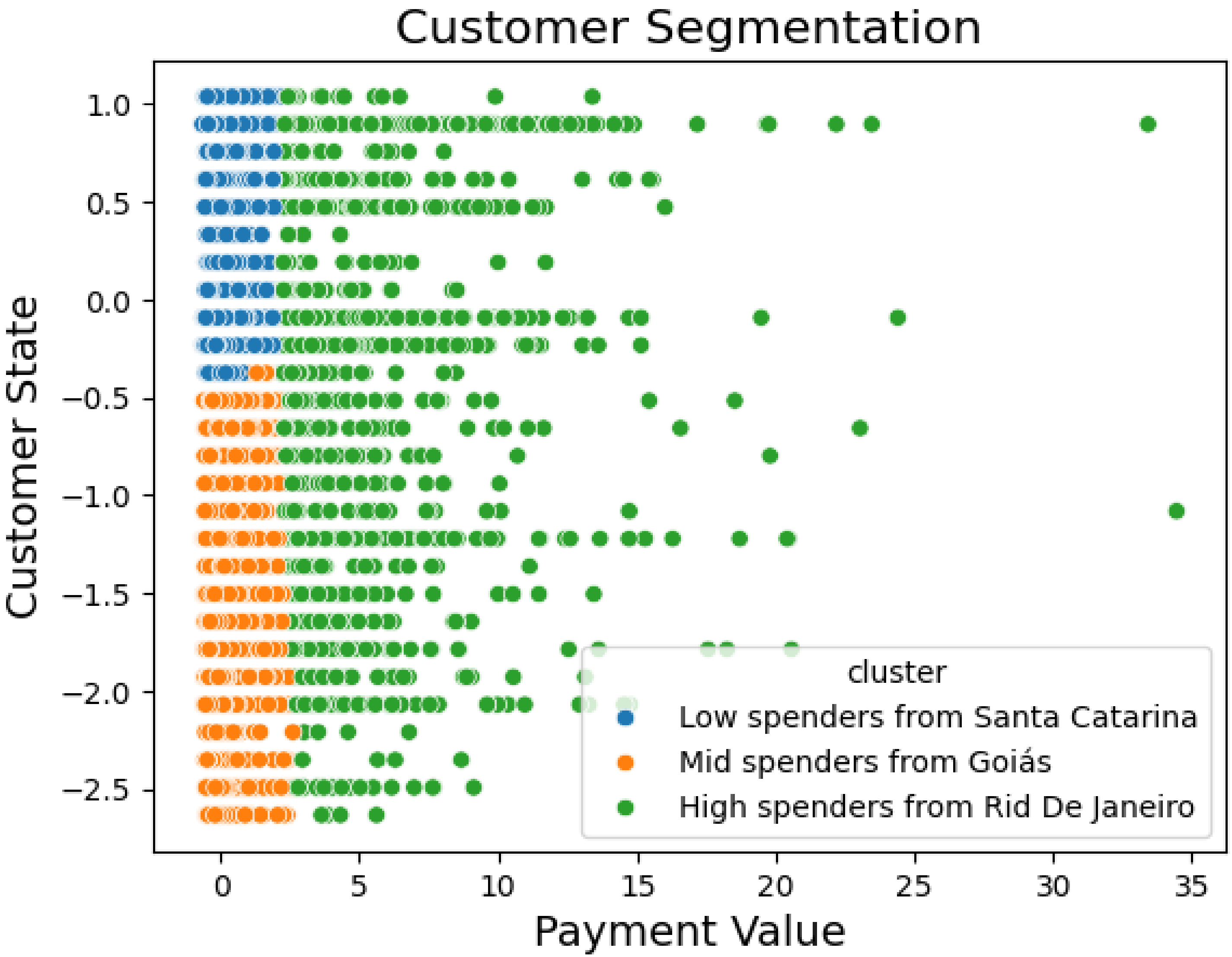
Delivery Delay

Freight Value

Order Size

Delivery Time

Customer Segmentation



What else we tried?

- **Analysis based on other product features -** length, width, photos, description showed that there is no statistical significance.
- **Distance:** review score is below 4 for orders where seller and customer are in different states and more than 4 where there are in the same state
- **Payment:** Analysis showed no statistical significance
- **Price:** Analysis showed no statistical significance

Recommendations

Delivery Time, Delivery Delay and Order Status

- Create express delivery option
- Avoid delays; customer to be contacted in advance when a delay is expected
- Specify dispatch and shipping time in the listing, add more clarity and transparency on every step
- Provide information and delivery confirmations to the customers (eliminate delivery status uncertainty)
- Plan for the busiest periods throughout the year
- Hire a fulfilment service for processing orders

Order Size

- Create detailed tracking data for each order item and have separate timestamps and notifications

Freight

- Freight value should be a fixed sum but a combination of several factors that reflect the type of products being transported (like weight, height, volume, type, etc.)

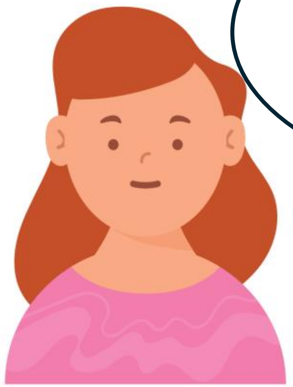
Turn your weaknesses into strengths

Everything went as planned and delivery was made before the deadline

Theo

Although the product was out of stock, the company contacted us to notify us and cancelled the order.

Helena



Laura

Very good quality product and excellent price, arrived before the promised deadline

Product delivered as requested, and very quickly.

Miguel

I loved it, it was beautiful, very delicate, I loved it

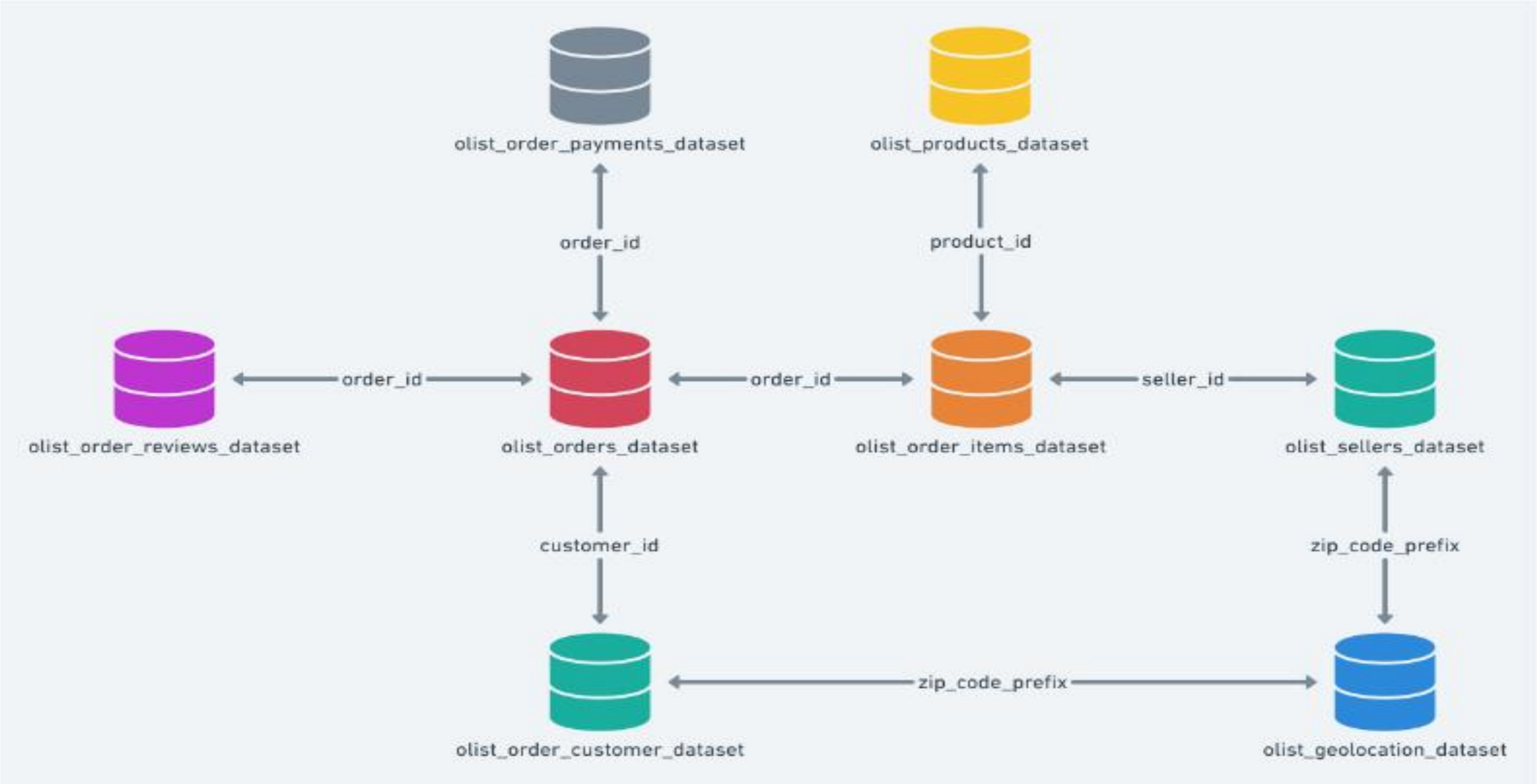
Bernardo

Thank you!

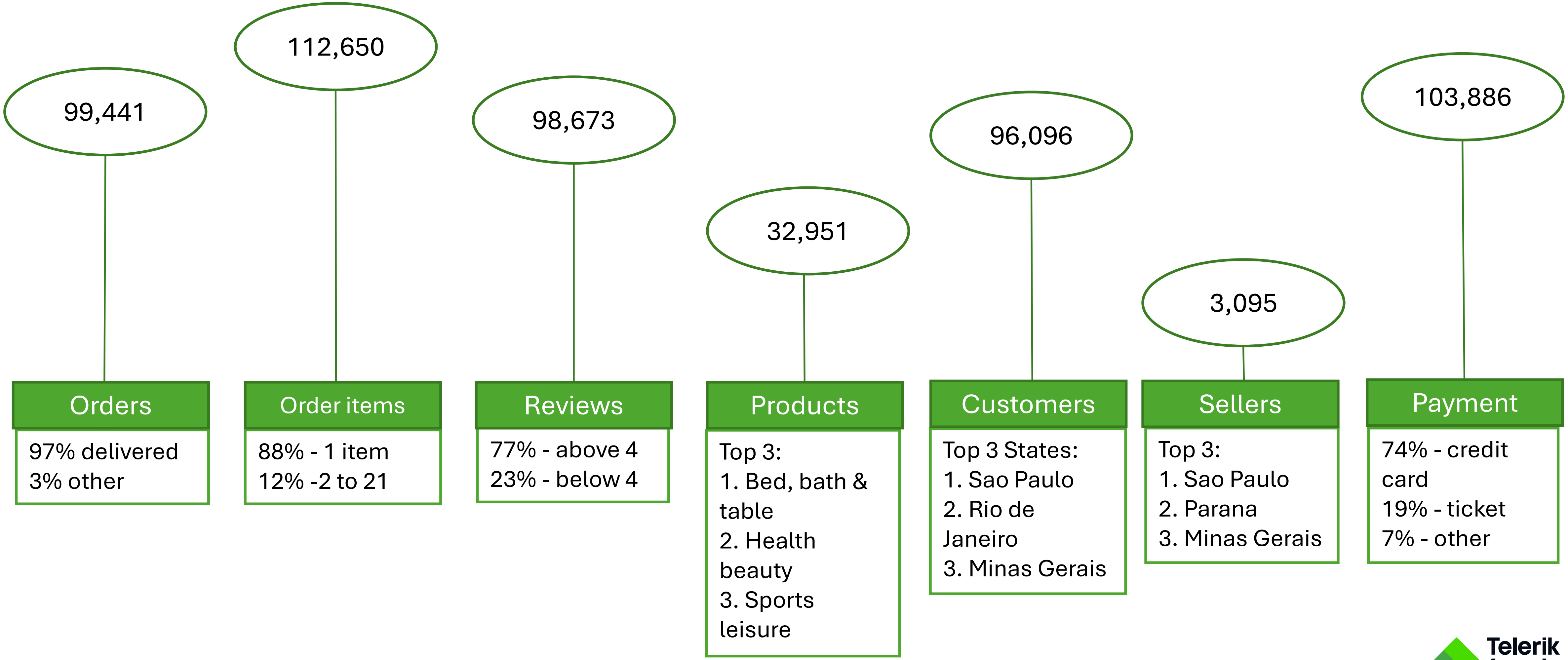
Appendix

- Data Preprocessing
- Regression Analysis (Linear): Delivery time components
- Regression Analysis (Logistics): Delivery time components

Data preprocessing: Available data



Data preprocessing: Available data



99,441

Orders

97% delivered
3% other

112,650

Order items

88% - 1 item
12% -2 to 21

98,673

Reviews

77% - above 4
23% - below 4

32,951

Products

Top 3:
1. Bed, bath & table
2. Health beauty
3. Sports leisure

96,096

Customers

Top 3 States:
1. Sao Paulo
2. Rio de Janeiro
3. Minas Gerais

3,095

Sellers

Top 3:
1. Sao Paulo
2. Parana
3. Minas Gerais

103,886

Payment

74% - credit card
19% - ticket
7% - other

Telerik Academy

Data preprocessing: Orders

Data issues

- Inconsistent data (mismatch between delivery status and dates):
  - 14 orders which are not approved are delivered;
  - 2 delivered order which are with missing carrier date;
  - 7 orders are delivered but are with missing delivered customer date;
  - 6 orders are with status cancelled but are delivered to the customer,;
  - 609 orders with status unavailable.

2. Inconsistent data (mismatch between the dates):
   - 359 orders which are delivered at the carried before being approved;
   - 23 orders which are delivered to the customer before being delivered to the carrier;
   - 2954 orders for which the review survey was sent before the delivery or the estimated delivery date.

3. Format discrepancies:
   - Dates are set as string.

Data preprocessing: Orders (2)

Normalized data

Missing data:

- 775 orders (0.8%) have no order items details.

Approach

1. Orders with mismatch between delivery status and dates are excluded from the dataset (point 1 from data issues);
2. Dates which are set as string are converted to datetime in Python.

Data preprocessing: Products

Data issues

1. Missing data:

- 610 products (1.9%) don't have product category, length, description, photos quantity;
- 2 products don't have product weight, length, height, width.

2. Typos:

- Some of the English categories are not input correctly (e.g. “fashio_female_clothing”, “home_confort”).

Normalized data

Missing data

- 1.6% of the orders don't have product category, length, description, photos quantity;
- 0.2% of the orders don't have product weight, length, height, width.

Approach

1. Product category changed to "Other" in SQL for the load in Python;
2. Missing items will be ignored in the analysis.

Data preprocessing: Reviews

Data issues

1. Missing data:

- 87,658 reviews (88%) don't have title;
- 58,274 reviews (59%) don't have message.

2. Not meaningful information:

- Some of the comments are filled as "xxx","-","*“, etc.

3. Mismatch between review score and comment:

For some of the reviews the review score is high but the comment is negative.

Normalized data

1. Missing data:

- 0.8% don't have reviews;
- 88% don't have comment title;
- 58% don't have review message.

2. Duplicate data:

- For one order there might be more than one review.

Approach

- For the duplicated items, in Python there are removed as the last review is kept.

Data preprocessing: Customers & Sellers

Data issues

1. Zip code mismatches:

- Customers: 39 zip codes corresponds to different city or state;
- Sellers: 49 zip codes corresponds to different city or state.

Normalized data

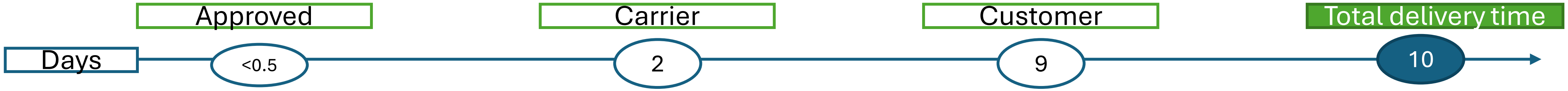
Missing data

- 1.6% of the orders don't have product category, length, description, photos quantity;
- 0.2% of the orders don't have product weight, length, height, width.

Approach

- Only the state is used for the analysis.

Regression Analysis (Linear): Total delivery time



Approved

Carrier

Customer

Total delivery time

Days

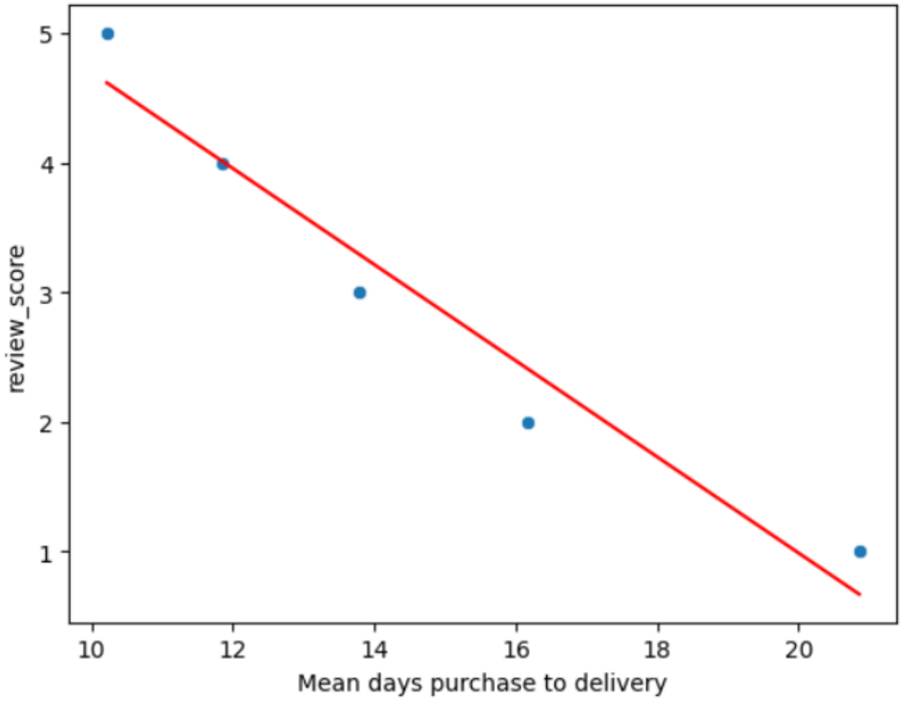
<0.5

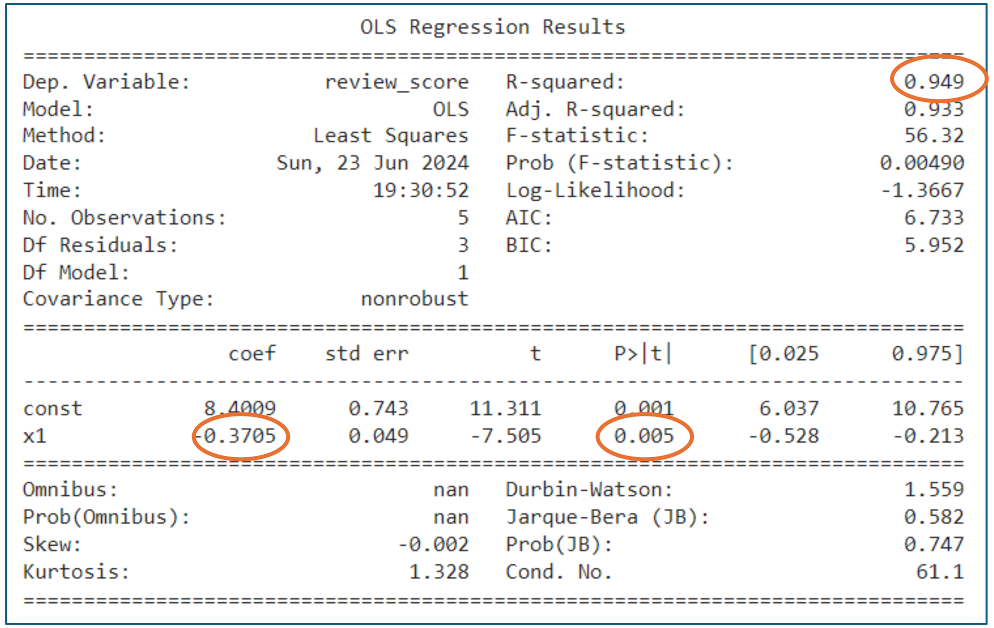
2

9

10

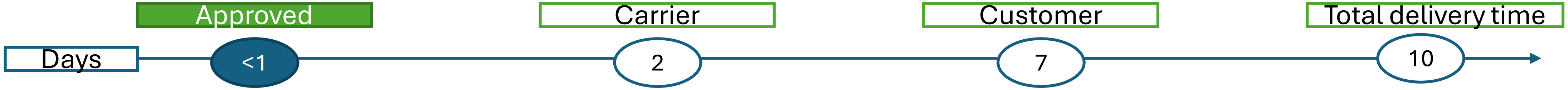
- The regression summary shows strong significance between the delivery time and review score.
- When total delivery time is more than **10 days**, customer reviews starts dropping.
- Each day additional delivery day reduces the review score by **0.37 points**



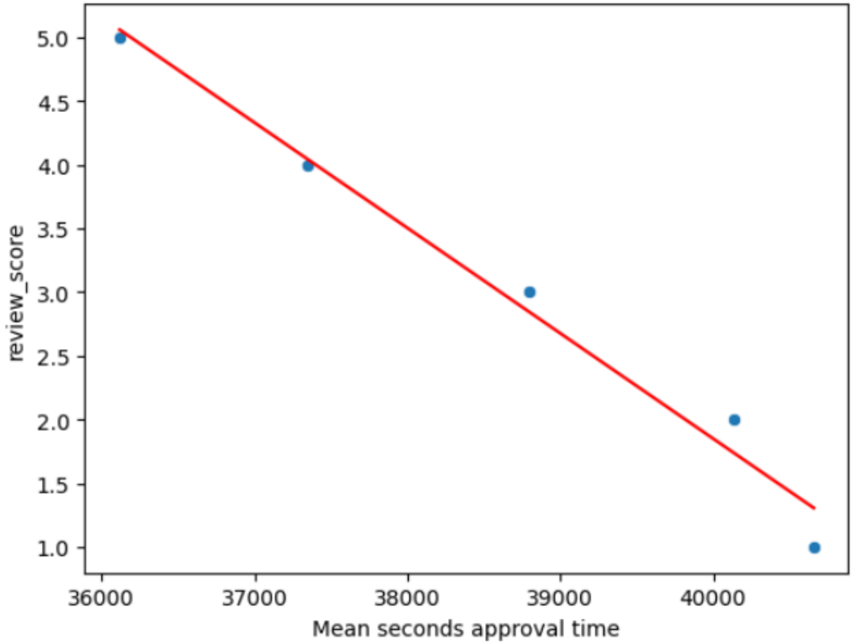


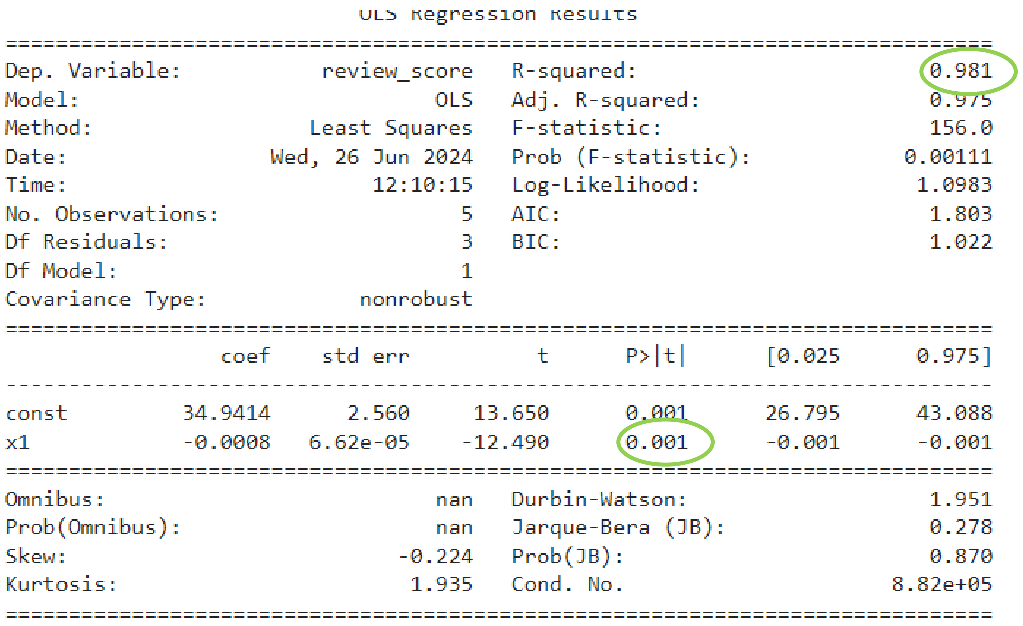
```
                            OLS Regression Results
==============================================================================
Dep. Variable:           review_score   R-squared:                       0.949
Model:                            OLS   Adj. R-squared:                  0.933
Method:                 Least Squares   F-statistic:                     56.32
Date:                Sun, 23 Jun 2024   Prob (F-statistic):            0.00490
Time:                        19:30:52   Log-Likelihood:                -1.3667
No. Observations:                   5   AIC:                             6.733
Df Residuals:                       3   BIC:                             5.952
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          8.4009      0.743     11.311      0.001       6.037      10.765
x1            -0.3705      0.049     -7.505      0.005      -0.528      -0.213
==============================================================================
Omnibus:                          nan   Durbin-Watson:                   1.559
Prob(Omnibus):                    nan   Jarque-Bera (JB):                0.582
Skew:                          -0.002   Prob(JB):                        0.747
Kurtosis:                       1.328   Cond. No.                         61.1
==============================================================================
```

Regression Analysis (Linear): Delivery time (Approval)



- When approval time is more than 10 hours, customer reviews starts dropping.
- The regression summary shows strong significance between the approval time and review score.

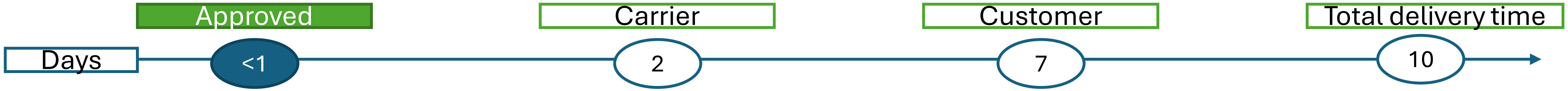




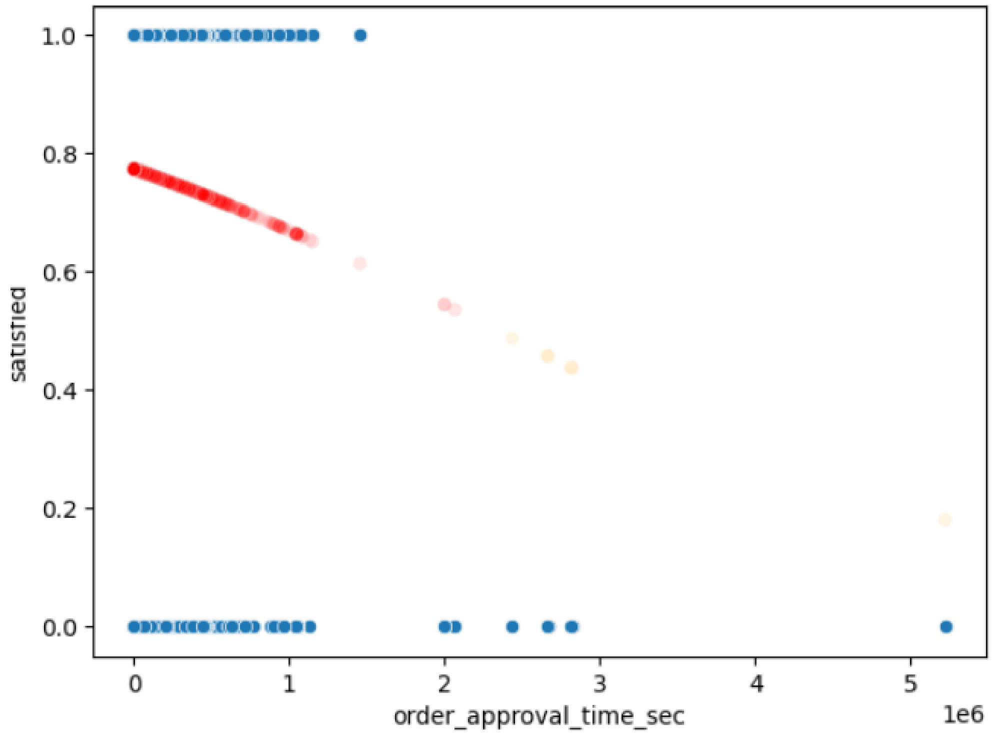
```
                            OLS Regression Results
==============================================================================
Dep. Variable:           review_score   R-squared:                       0.981
Model:                            OLS   Adj. R-squared:                  0.975
Method:                 Least Squares   F-statistic:                     156.0
Date:                Wed, 26 Jun 2024   Prob (F-statistic):            0.00111
Time:                        12:10:15   Log-Likelihood:                 1.0983
No. Observations:                   5   AIC:                             1.803
Df Residuals:                       3   BIC:                             1.022
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         34.9414      2.560     13.650      0.001      26.795      43.088
x1            -0.0008   6.62e-05    -12.490      0.001      -0.001      -0.001
==============================================================================
Omnibus:                          nan   Durbin-Watson:                   1.951
Prob(Omnibus):                    nan   Jarque-Bera (JB):                0.278
Skew:                          -0.224   Prob(JB):                        0.870
Kurtosis:                       1.935   Cond. No.                     8.82e+05
==============================================================================
```

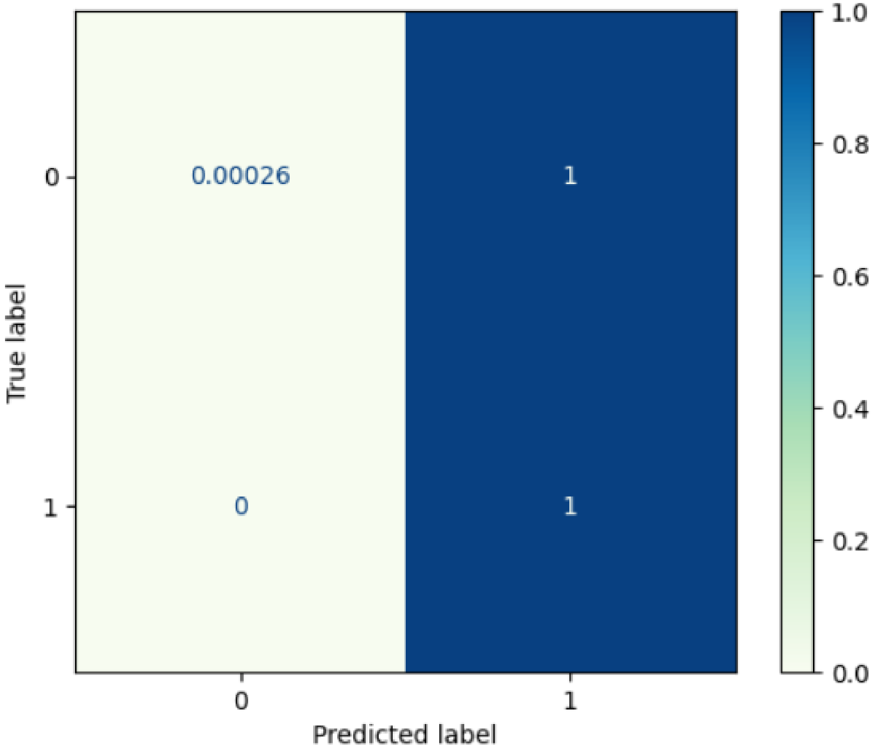
Telerik Academy

Regression Analysis (Logistics): Delivery time (Approval)

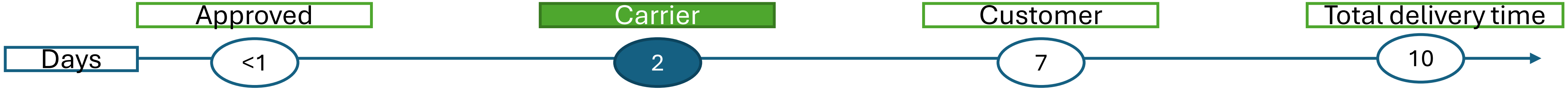


- After 0.5 days approval time, the customer will become unsatisfied.
- There is 50% chance that when a customer is satisfied/unsatisfied it will be properly classified.





Regression Analysis (Linear): Delivery time (Carrier)



Days

Approved

<1

Carrier

2

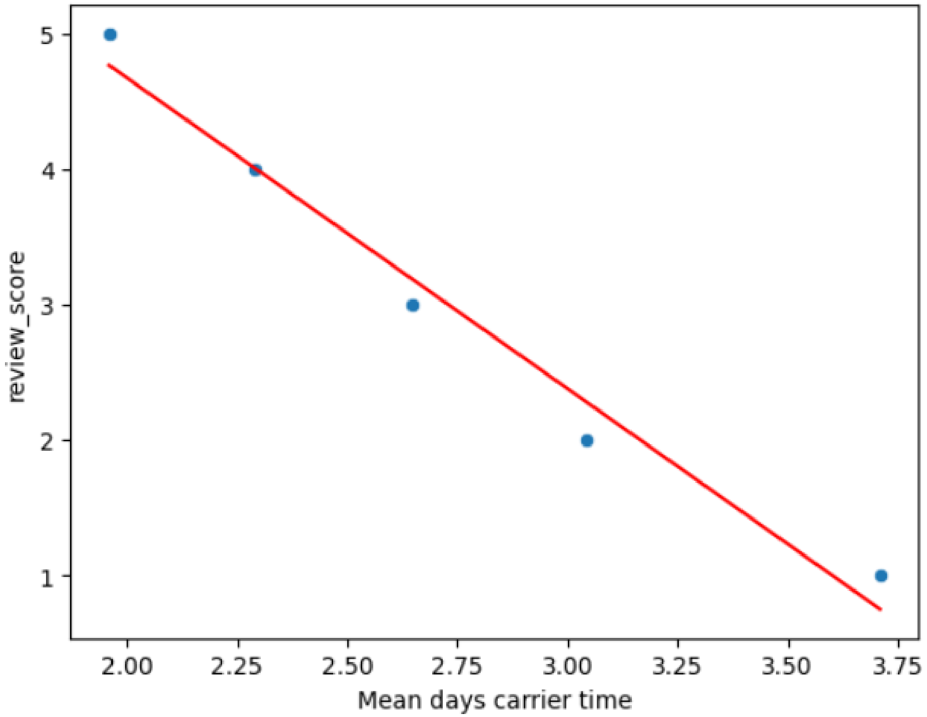
Customer

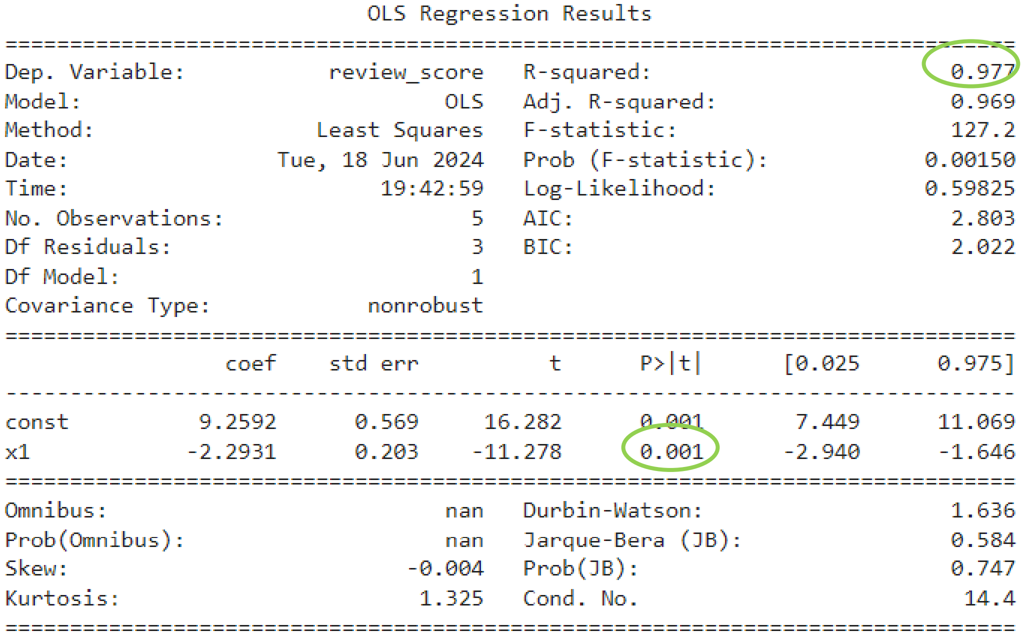
7

Total delivery time

10

- When total dispatch time to carrier is more than 2 days, customer reviews starts dropping.
- The regression summary shows strong significance between the dispatch time to carrier and review score.

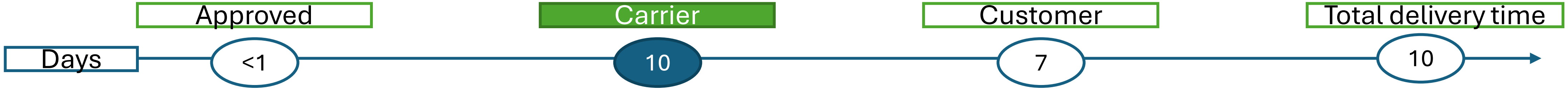




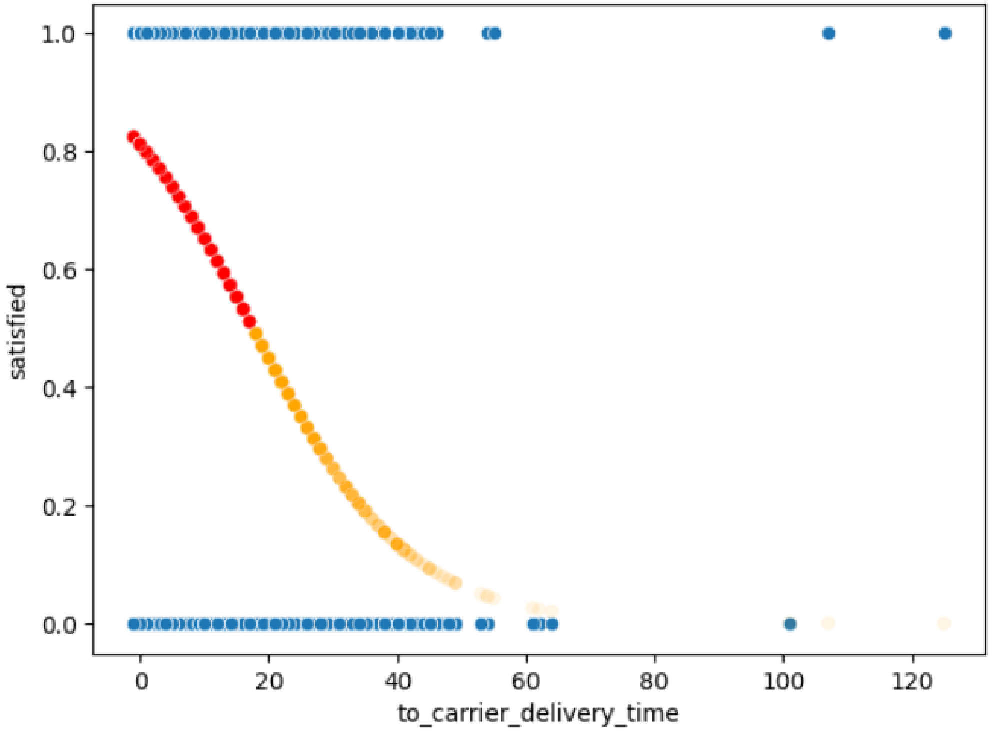
```
                            OLS Regression Results
==============================================================================
Dep. Variable:           review_score   R-squared:                       0.977
Model:                            OLS   Adj. R-squared:                  0.969
Method:                 Least Squares   F-statistic:                     127.2
Date:                Tue, 18 Jun 2024   Prob (F-statistic):            0.00150
Time:                        19:42:59   Log-Likelihood:                0.59825
No. Observations:                   5   AIC:                             2.803
Df Residuals:                       3   BIC:                             2.022
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          9.2592      0.569     16.282      0.001       7.449      11.069
x1            -2.2931      0.203    -11.278      0.001      -2.940      -1.646
==============================================================================
Omnibus:                          nan   Durbin-Watson:                   1.636
Prob(Omnibus):                    nan   Jarque-Bera (JB):                0.584
Skew:                          -0.004   Prob(JB):                        0.747
Kurtosis:                       1.325   Cond. No.                         14.4
==============================================================================
```

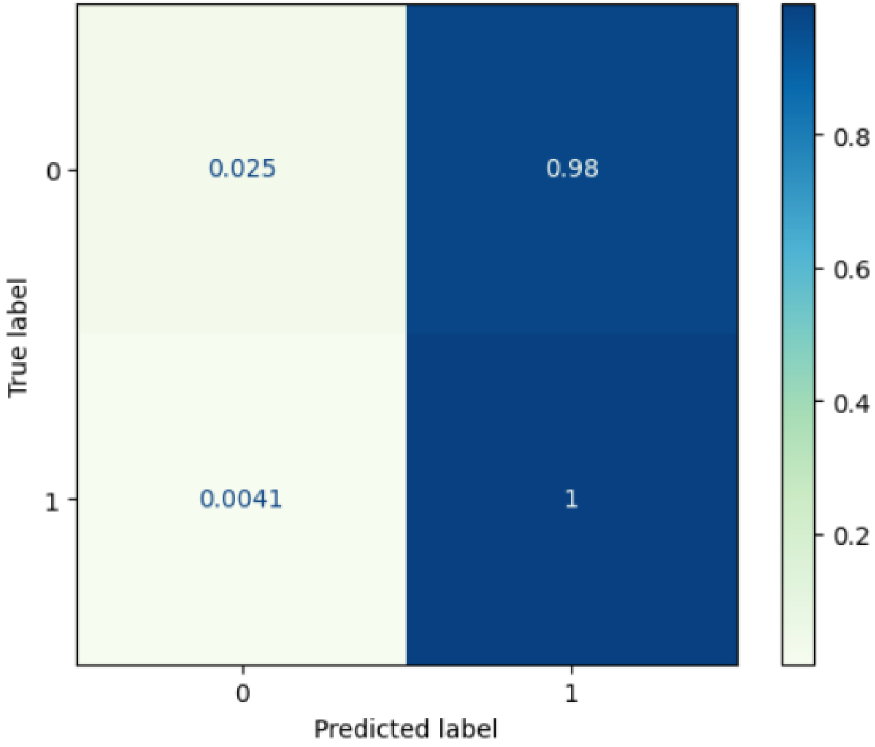
Telerik Academy

Regression Analysis (Logistics): Delivery time (Carrier)



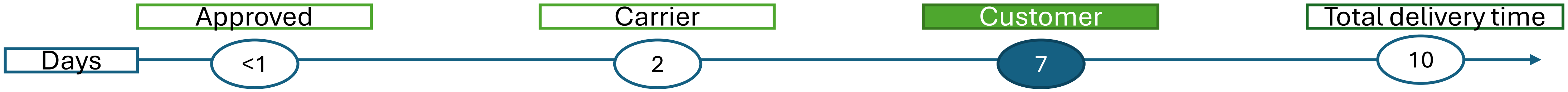
- After ~10 days dispatch time to carrier, the customer will become unsatisfied.
- There is 50% chance that when a customer is satisfied/unsatisfied it will be properly classified.



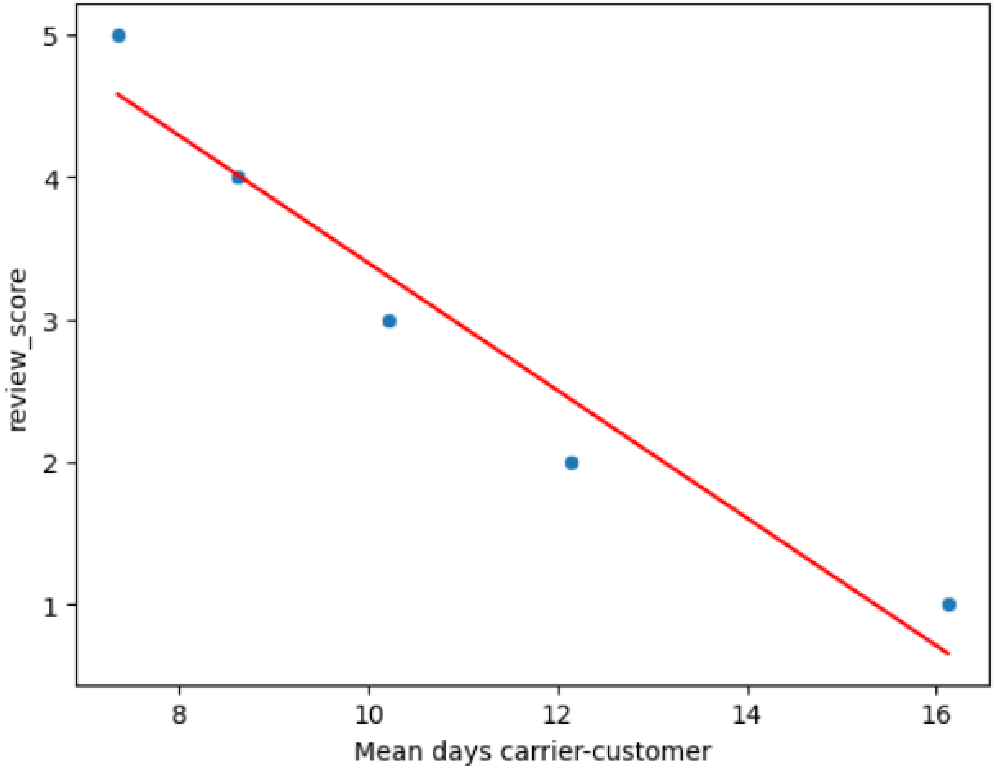


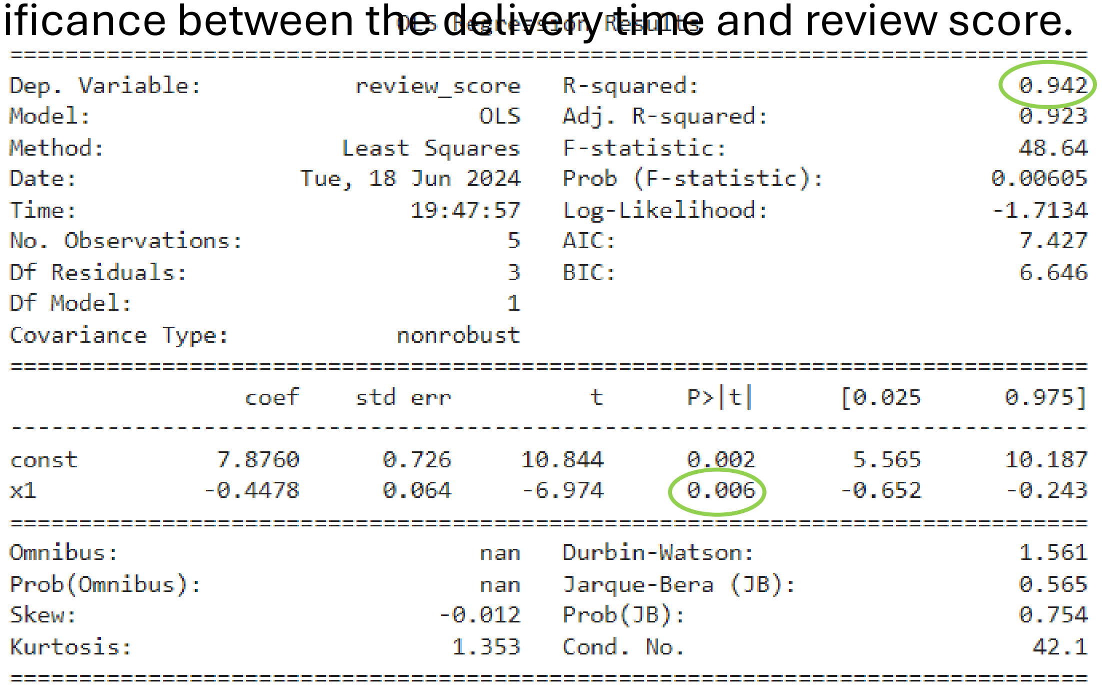
Telerik Academy

Regression Analysis (Linear): Delivery time (Customer)



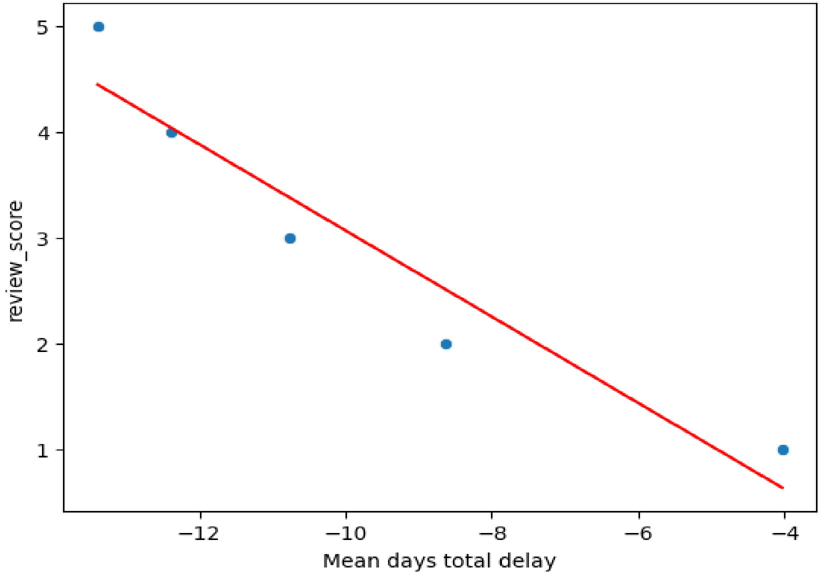
- When dispatch time from carrier to customer is more than 7 days, customer reviews starts dropping.
- gnificance between the delivery time and review score.

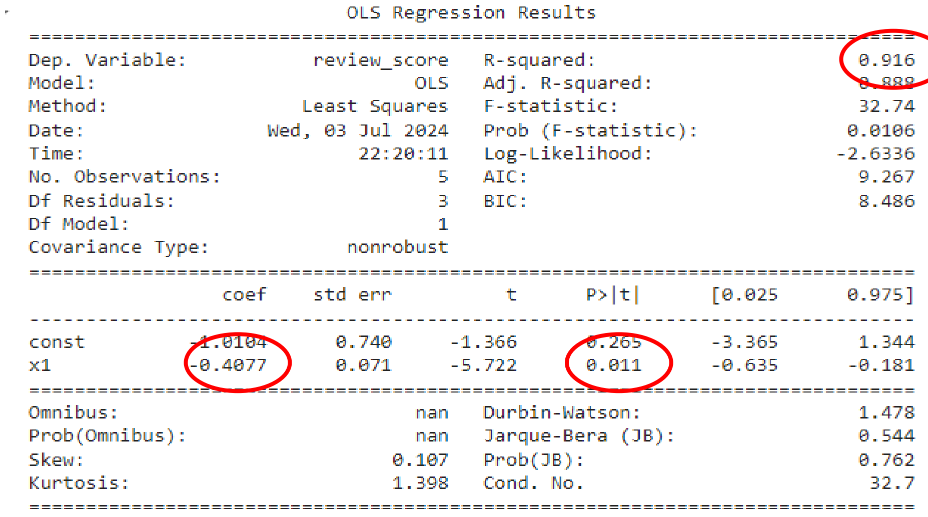




```
                            OLS Regression Results
==============================================================================
Dep. Variable:           review_score   R-squared:                       0.942
Model:                            OLS   Adj. R-squared:                  0.923
Method:                 Least Squares   F-statistic:                     48.64
Date:                Tue, 18 Jun 2024   Prob (F-statistic):            0.00605
Time:                        19:47:57   Log-Likelihood:                -1.7134
No. Observations:                   5   AIC:                             7.427
Df Residuals:                       3   BIC:                             6.646
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          7.8760      0.726     10.844      0.002       5.565      10.187
x1            -0.4478      0.064     -6.974      0.006      -0.652      -0.243
==============================================================================
Omnibus:                          nan   Durbin-Watson:                   1.561
Prob(Omnibus):                    nan   Jarque-Bera (JB):                0.565
Skew:                          -0.012   Prob(JB):                        0.754
Kurtosis:                       1.353   Cond. No.                         42.1
==============================================================================
```

Regression Analysis (Linear): Delivery delay

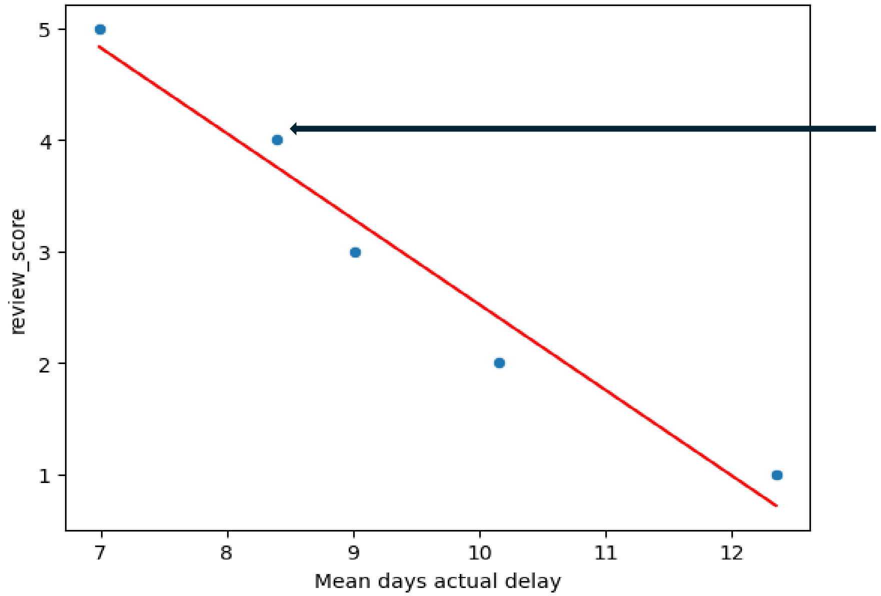




```
                            OLS Regression Results
==============================================================================
Dep. Variable:           review_score   R-squared:                       0.916
Model:                            OLS   Adj. R-squared:                  0.888
Method:                 Least Squares   F-statistic:                     32.74
Date:                Wed, 03 Jul 2024   Prob (F-statistic):             0.0106
Time:                        22:20:11   Log-Likelihood:                -2.6336
No. Observations:                   5   AIC:                             9.267
Df Residuals:                       3   BIC:                             8.486
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -1.0104      0.740     -1.366      0.265      -3.365       1.344
x1            -0.4077      0.071     -5.722      0.011      -0.635      -0.181
==============================================================================
Omnibus:                          nan   Durbin-Watson:                   1.478
Prob(Omnibus):                    nan   Jarque-Bera (JB):                0.544
Skew:                           0.107   Prob(JB):                        0.762
Kurtosis:                       1.398   Cond. No.                         32.7
==============================================================================
```

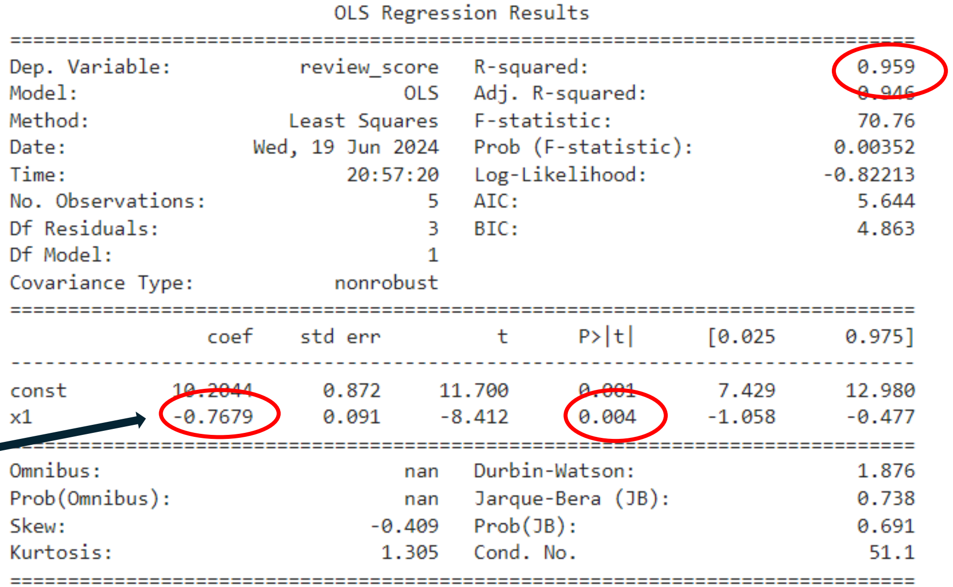
Strong significance between the delay time and review score

*outliers are removed - only orders with delay are analyzed



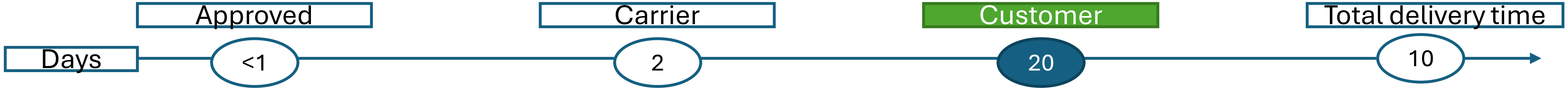
Customer reviews starts dropping after **8 days of delay**

Review score is dropping by **0.76 points for each day of delay**



```
                            OLS Regression Results
==============================================================================
Dep. Variable:           review_score   R-squared:                       0.959
Model:                            OLS   Adj. R-squared:                  0.946
Method:                 Least Squares   F-statistic:                     70.76
Date:                Wed, 19 Jun 2024   Prob (F-statistic):            0.00352
Time:                        20:57:20   Log-Likelihood:               -0.82213
No. Observations:                   5   AIC:                             5.644
Df Residuals:                       3   BIC:                             4.863
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         10.2044      0.872     11.700      0.001       7.429      12.980
x1            -0.7679      0.091     -8.412      0.004      -1.058      -0.477
==============================================================================
Omnibus:                          nan   Durbin-Watson:                   1.876
Prob(Omnibus):                    nan   Jarque-Bera (JB):                0.738
Skew:                          -0.409   Prob(JB):                        0.691
Kurtosis:                       1.305   Cond. No.                         51.1
==============================================================================
```

Regression Analysis (Logistics): Delivery time (Customer)



Approved

Carrier

Customer

Total delivery time

Days

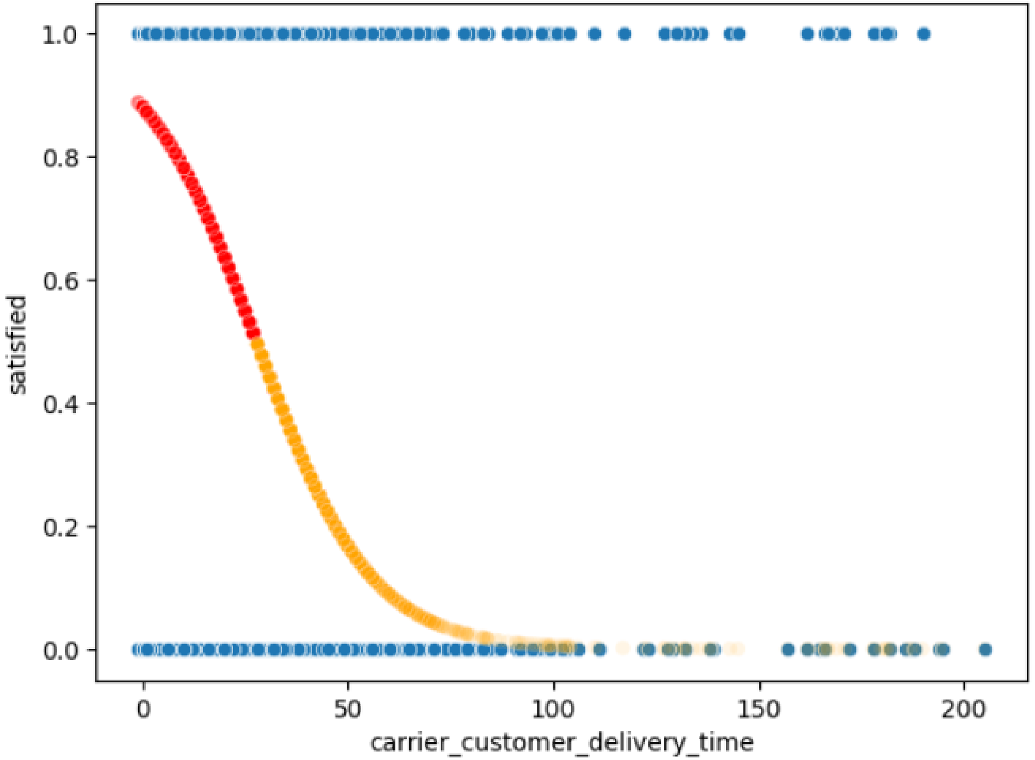
<1

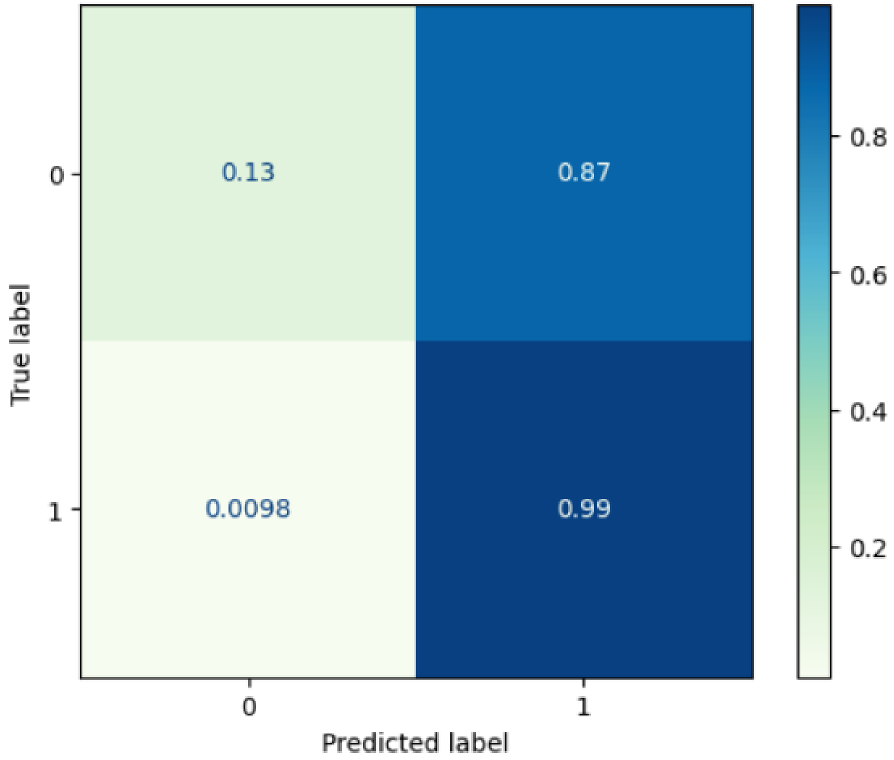
2

20

10

- After ~20 days delivery between carrier to customer, the customer will become unsatisfied.
- There is 50% chance that when a customer is satisfied/unsatisfied it will be properly classified.


satisfied
carrier_customer_delivery_time


True label
Predicted label