# Tree-Based Algorithms Approach on Predicting Customer Satisfaction

By: Zikry Adjie Nugraha

# OUTLINE

01 Data Understanding

02 Data Cleaning & Preprocessing

03 Exploratory Data Analysis

04 Feature Engineering

05 Tree-based Machine learning Modelling

06 Conclusion

# DATA UNDERSTANDING

# Data Understanding

**The dataset is about Brazil Public E-Commerce Public dataset by Olist from Kaggle and licensed to be used publicly by its author .**

**The idea of this project is to predict in the future whether customer will give good or bad review based on the predicting review score column that is correlated with other columns.**

**The dataset contains order-related columns, product-related columns, payment-related columns, and specificially review score column.**

# Data Understanding

| Column group | Column name | Description |
|---|---|---|
| Order-related columns | order_status | This is a reference to the order status (delivered or canceled). |
| | order_purchase_timestamp | Displays the timestamp of each item's purchase. |
| | order_delivered_customer_date | Displays the customer's actual order delivery date. |
| | order_estimated_delivery_date | Displays the estimated delivery date that was provided to the customer at the time of purchase. |
| | shipping_limit_date | Displays the seller's shipping limit date for transferring the order to the logistic partner. |
| Payment-related columns | payment_sequential | A customer may pay for an order using multiple payment methods. |
| | payment_type | The customer's preferred method of payment. |
| | payment_installments | The customer's preferred number of payment installments. |

# Data Understanding

| Column group | Column name | Description |
|---|---|---|
| **Payment-related columns** | **payment_value** | The transaction's value. |
| | **price** | The cost of each item. |
| | **freight_value** | The cost of transportation for each item (if an order has more than one item the freight value is split between items). |
| **Product-related columns** | **product_category** | Each item's category. |
| | **product_name_length** | The number of characters extracted from the product name. |
| | **product_description_length** | The number of characters extracted from the product description. |
| | **product_photos_qty** | The number of product photos that have been published. |
| **Review-related columns** | **review_score** | A rating given by a customer on a satisfaction survey ranging from 1 to 5. |

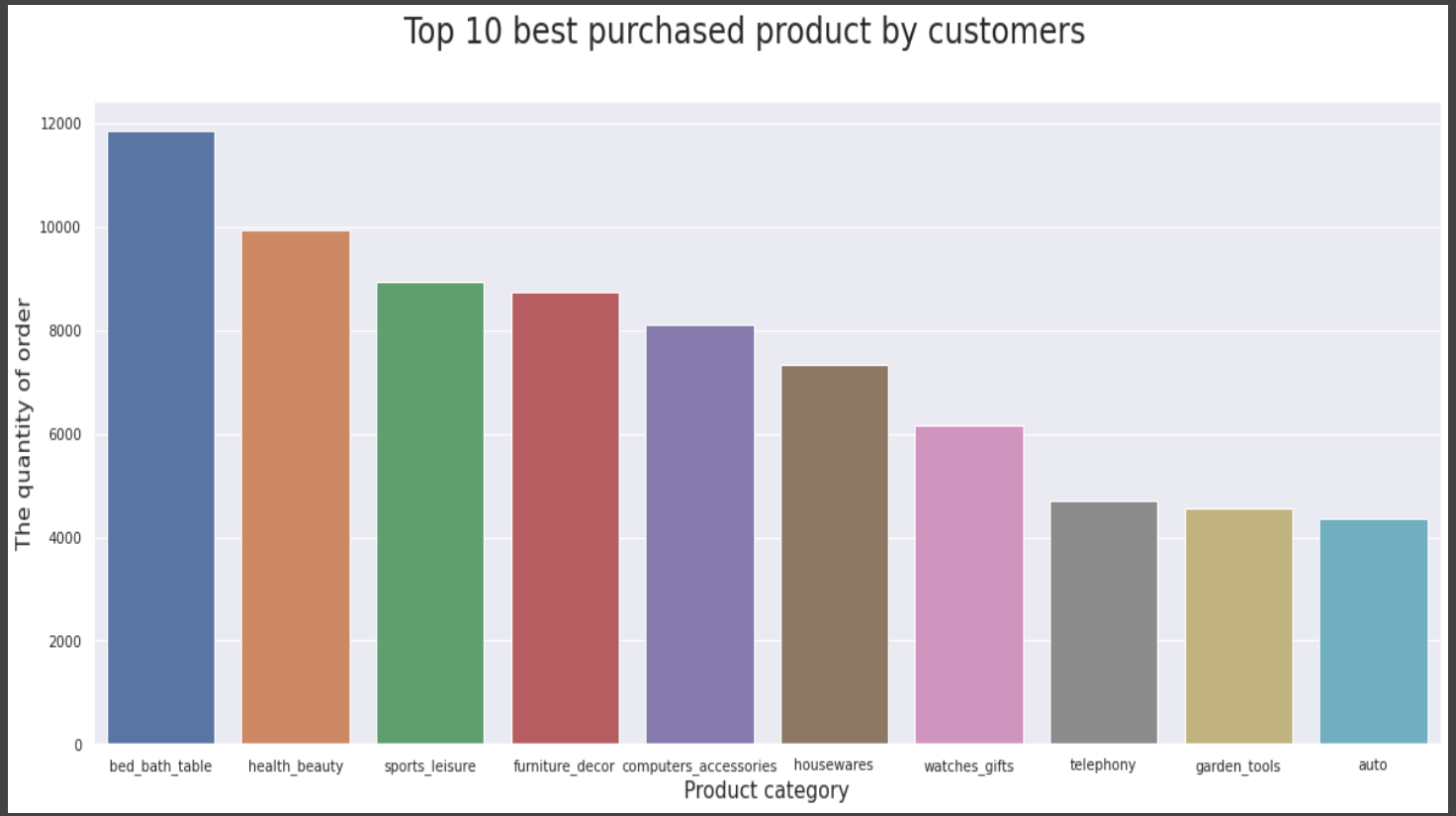
# DATA CLEANING & PREPROCESSING

# Data cleaning & preprocessing

```
order_status                        0
order_purchase_timestamp            0
order_delivered_customer_date    2400
order_estimated_delivery_date       0
shipping_limit_date                 0
payment_sequential                  0
payment_type                        0
payment_installments                0
payment_value                       0
price                               0
freight_value                       0
product_category                    0
product_name_length                 0
product_description_length          0
product_photos_qty                  0
review_score                        0
dtype: int64
```

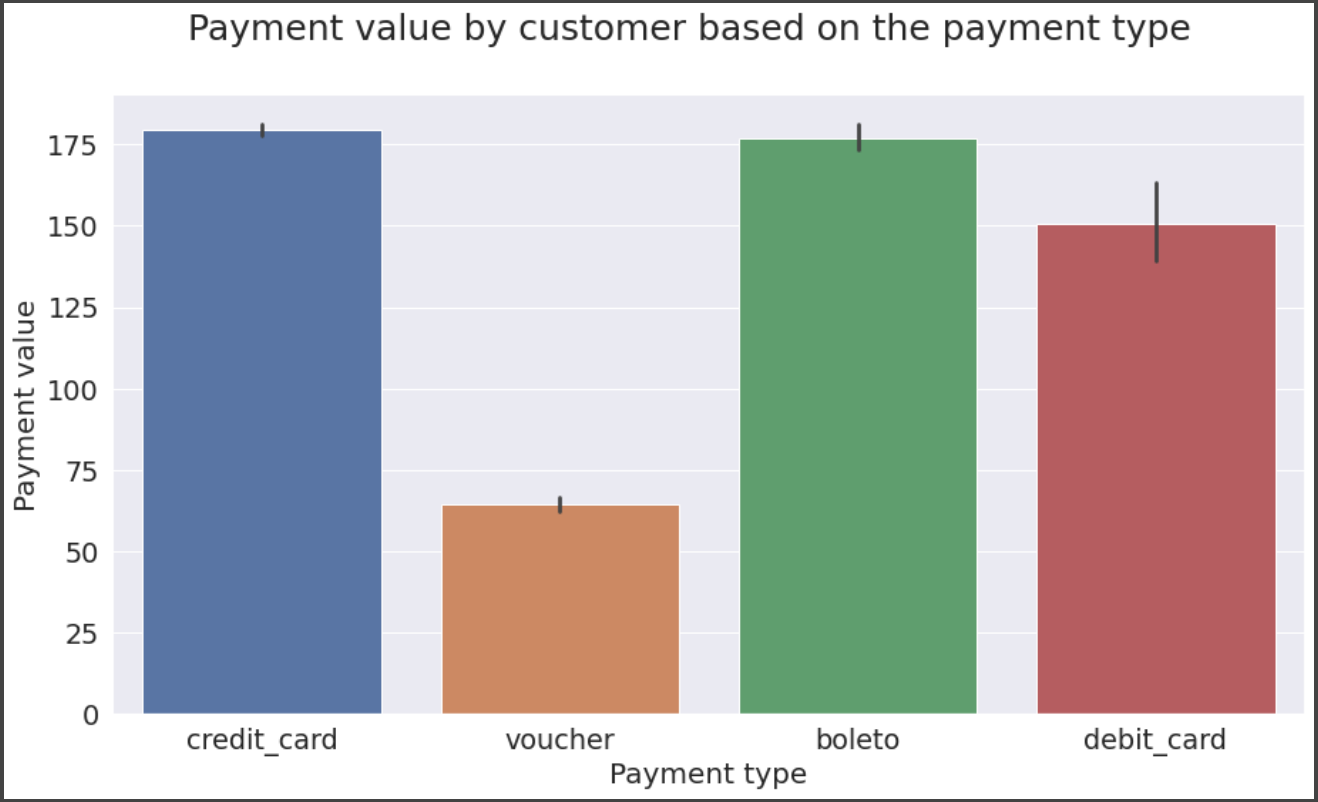
- There are total 2,400 NaN value on order_delivered_customer_date column alone.
- The data cleaning process remove 2.08% NaN value data from 115,609 rows to 113,209 rows.
- The timestamp data from column with date value has been converted to new column with integral data for further analysis on machine learning model.

# EXPLORATORY DATA ANALYSIS


Top 10 best purchased product by customers
The quantity of order
12000
10000
8000
6000
4000
2000
0
bed_bath_table
health_beauty
sports_leisure
furniture_decor
computers_accessories
housewares
watches_gifts
telephony
garden_tools
auto
Product category

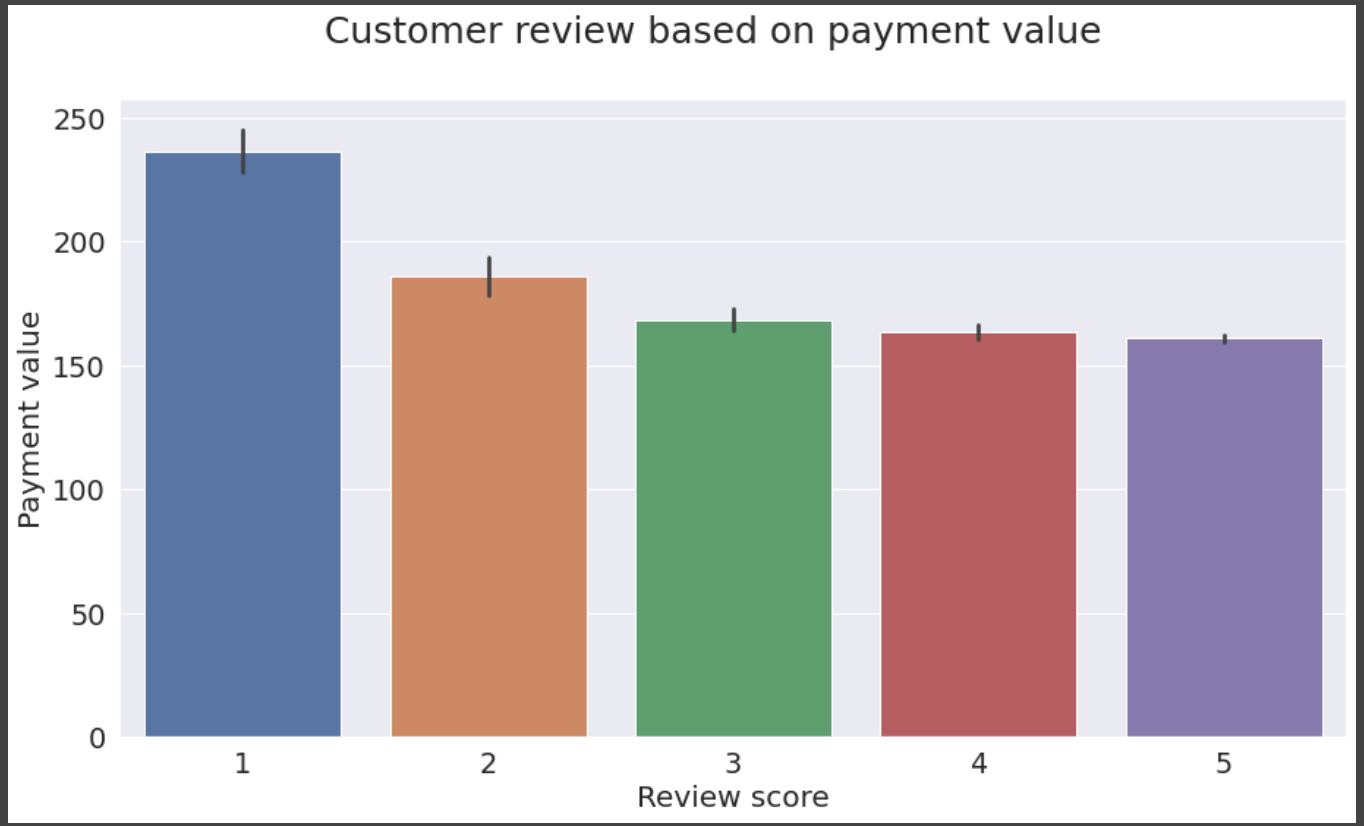
## Business insight on the top ten most purchased products:

1. The top ten most purchased products are from the product categories of bed bath table, health beauty, sport leisure, furniture decoration, computer accessories, housewares, watches gifts, telephony, garden tools, and auto.
2. Customers' most popular product is the bed bath table, which has received over 10,000 orders.
3. More than 8,000 orders have been placed in the categories of health and beauty, sport and leisure, furniture and decoration, and computer accessories.
4. More than 6,000 orders were placed for housewares and watches gifts.
5. More than 3,000 orders have been placed in the categories of telecommunications, garden tools, and auto.
6. These top ten products played an important role in determining customer satisfaction, and they will be used as the parameter to do the one-hot encoding process later on during the feature engineering process.


Payment value by customer based on the payment type
Payment value
175
150
125
100
75
50
25
0
credit_card
voucher
boleto
debit_card
Payment type

# Business insight into customer payment types:

1. Credit cards have the highest payment value, followed by boleto, debit cards, and vouchers.
2. Both credit card and boleto payments have a payment value of more than 175.
3. The payment value using debit card is lower than both credit card and boleto as it has the value of 150.
4. Voucher has the lowest payment value among the others because half of the actual price of the product can be paid by customers using redeemed voucher.


Customer review based on payment value
Payment value
250
200
150
100
50
0
1
2
3
4
5
Review score

## Business insight on the customer review based on payment value:

1. Review score is increase slightly as the payment value is decrease.
2. Review score with value of 1 occurs when the payment value is more than 200.
3. Review score with value of 2 occurs as the payment value decrease into around 175.
4. Review score with value of 3, 4, and 5 occurs when the payment value is around 160 which makes the payment value of 160 will be the best option for making customer make high review score.


Customer review based on Price
Price
0
20
40
60
80
100
120
1
2
3
4
5
Review score

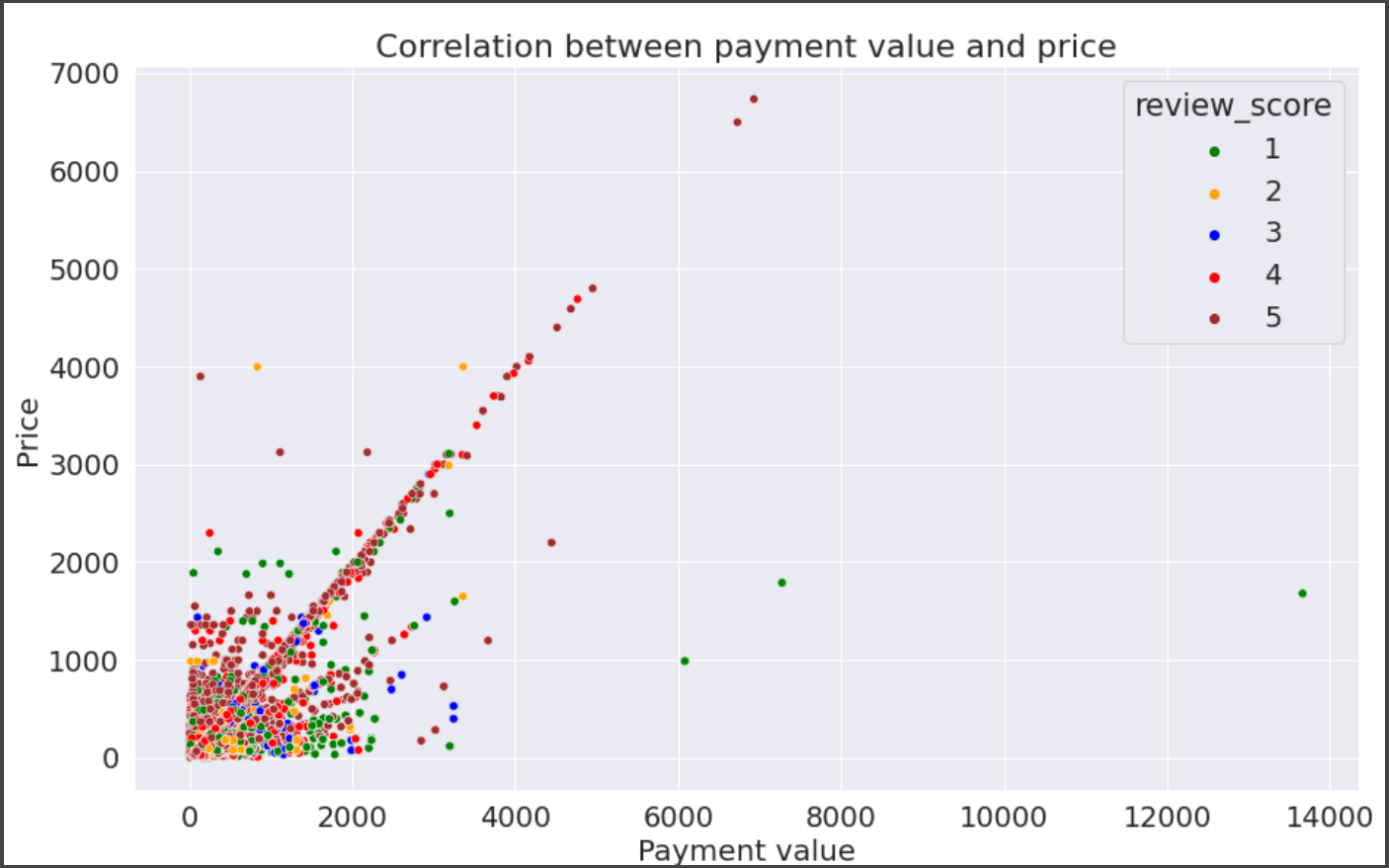
## Business insight on the customer review based on price:

1. The lower price value has no discernible effect on the review score.
2. A product with a price tag of more than 120 can elicit both positive and negative feedback from customers.
3. The review score of 1 can occur when a product has a high price but the quality of the product does not meet the customer’s expectations (The high price can also mean high quality product).
4. When a product has a high price and the quality of the product meets the customer’s expectations, a review score of 5 is possible.
5. The review score of 3 has the lowest price among the others, with a value of around 110.


Customer review based on freight value
Freight value
0
5
10
15
20
1
2
3
4
5
Review score

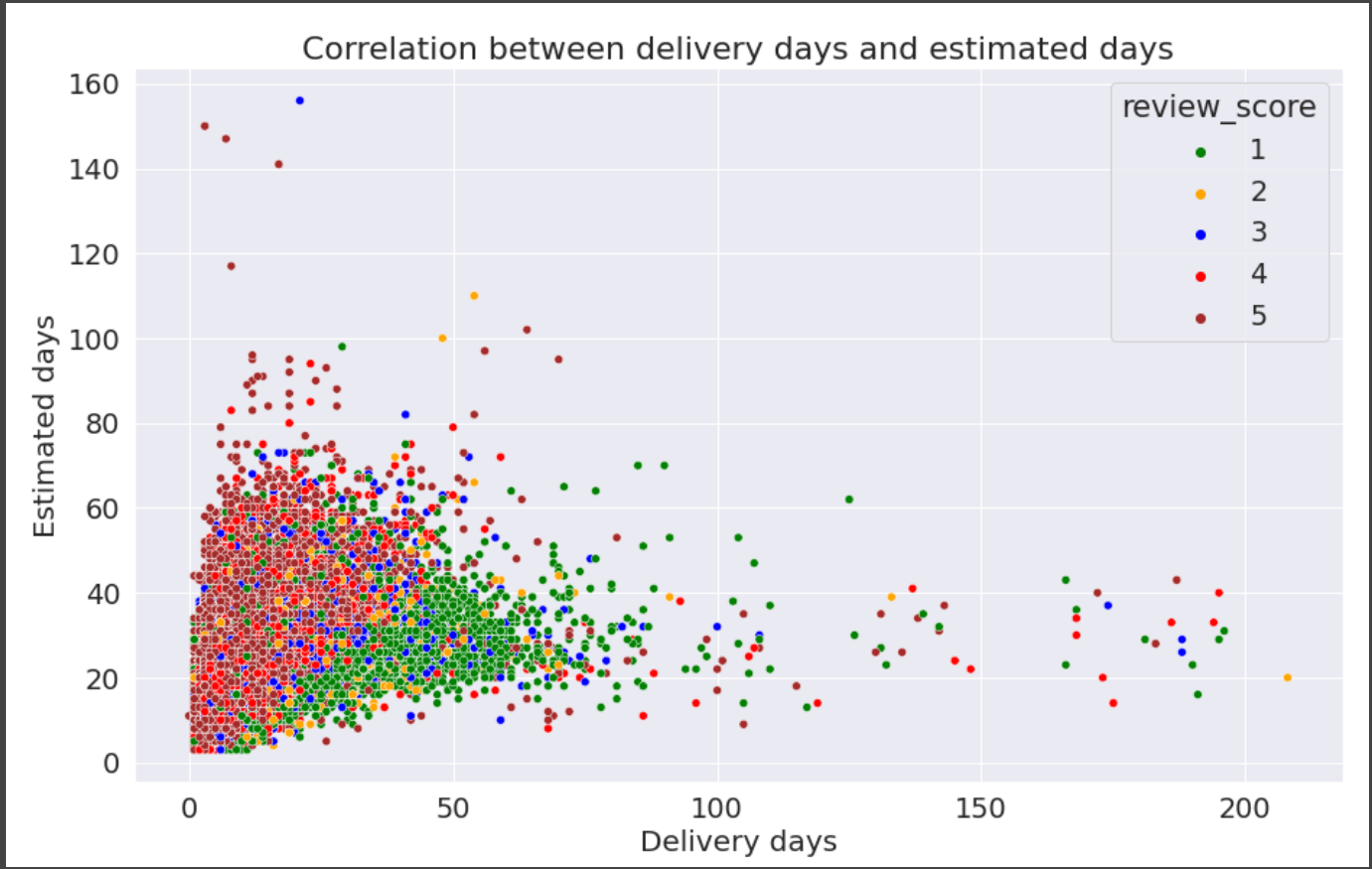
# Business insight on the customer review based on freight value:

1. Customer will likely to give better review score as the freight value decrease.
2. Freight value of more than 20 gives the variety of customer review from 1 to 4.
3. Customer will give the best review score of 5 when the freight value is less than 20.
4. The lowest freight value makes the best review score because customers will be happy when they only have to pay a little for the additional payment of freight value.


Correlation between payment value and price
Price
Payment value
0
1000
2000
3000
4000
5000
6000
7000
0
2000
4000
6000
8000
10000
12000
14000
review_score
1
2
3
4
5

## Business insight on the customer review based on correlation of payment value and price:

1. Most of the good review score of 4 and 5 occurs when the price and payment value has the same value.
2. The correlated review score of the same value between price and payment value occurs when the customer only has to pay the price of a product without any additional value that will increase the payment value.
3. Most of the review score of 5 also happened when the payment value is lower than the actual price of the products.
4. Most of the review score of 1 happened when the payment value is higher than the actual price of the products.


Correlation between delivery days and estimated days
Estimated days
Delivery days
0
20
40
60
80
100
120
140
160
0
50
100
150
200
review_score
1
2
3
4
5

# Business insight on the customer review based on correlation of estimated days and delivery days of product arrival:

1. The majority of the good review scores of 4 and 5 have long estimated days, but the actual delivery days of the product arrive early or on time, making the product arrive at the customer’s hand faster than the expected time.
2. The majority of the negative review scores of 1 and 2 have short estimated delivery days, but the actual delivery days of the product are longer than the expected time.

# FEATURE ENGINEERING

# Feature Engineering

## New feature column

### arrival_time

Define how many days the product need to arrive at customer according to estimated days and delivery days.

### delivery_arrival

Define whether the arrival is on time or late based on the arrival time (if negative value on arrival time means late and vice versa).

### score

Define whether customer gives good or bad review (0 for review_score = 1-2, 1 for review_score = 4-5, and neutral value of review_score 3 is remove.

# Feature Engineering

## Label and one hot encoding on categorical feature

### Column with 2 distinct value

Convert the negative value to 0 and the positive value to 1.

### Column with 2+ distinct value

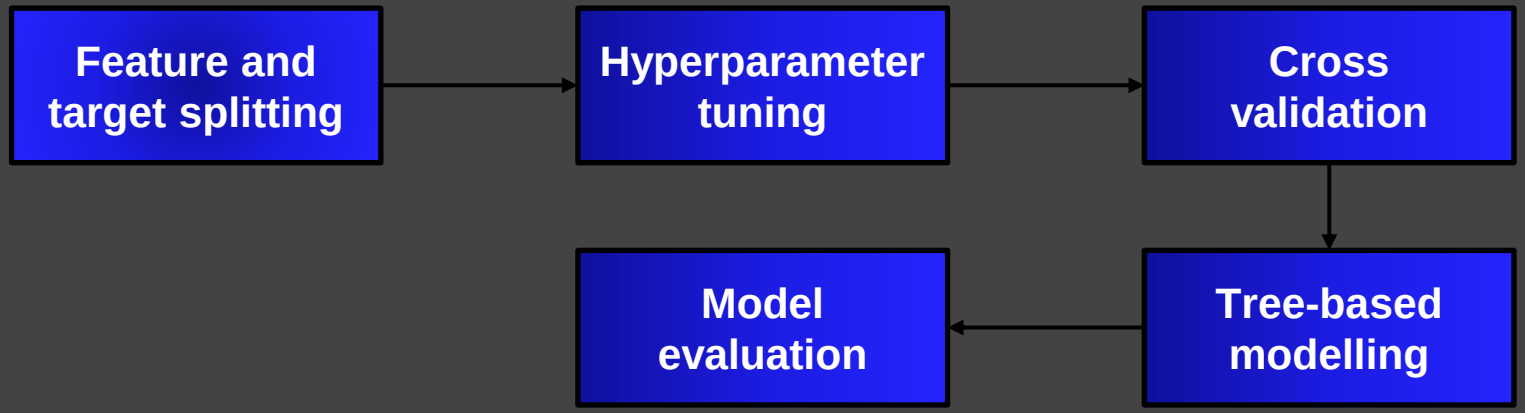
One hot encoding process using pd.get_dummies function on specific column.

### Column with 10+ distinct value

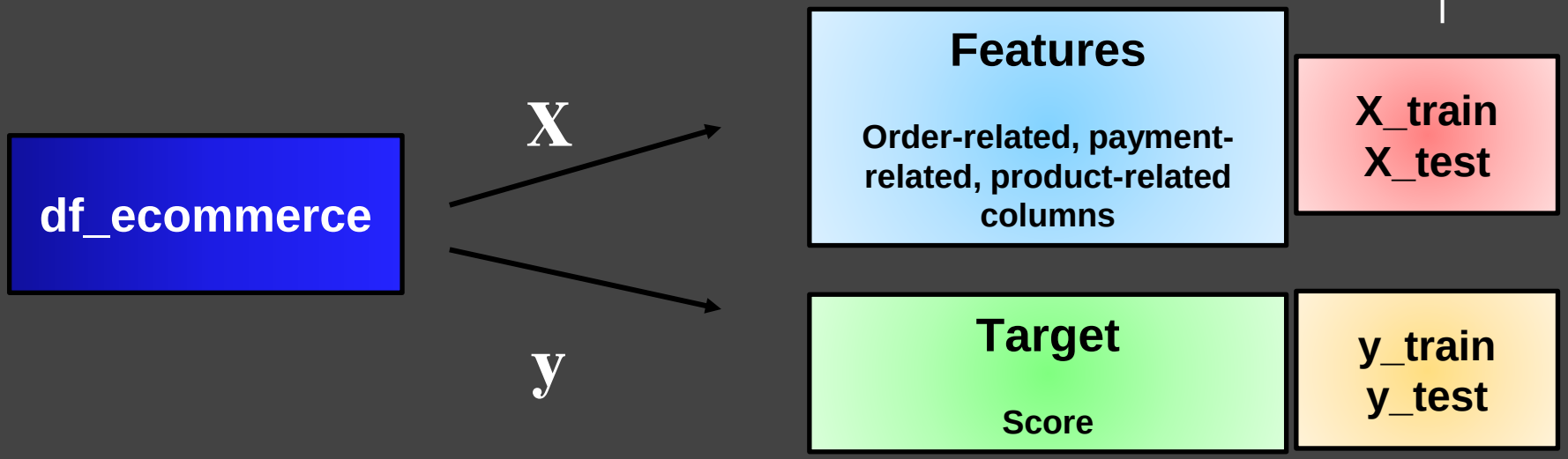
Define the top 10 most appeared value and then do the one hot encoding to those specific values.

# TREE-BASED MACHINE LEARNING MODELLING

# Modelling Process

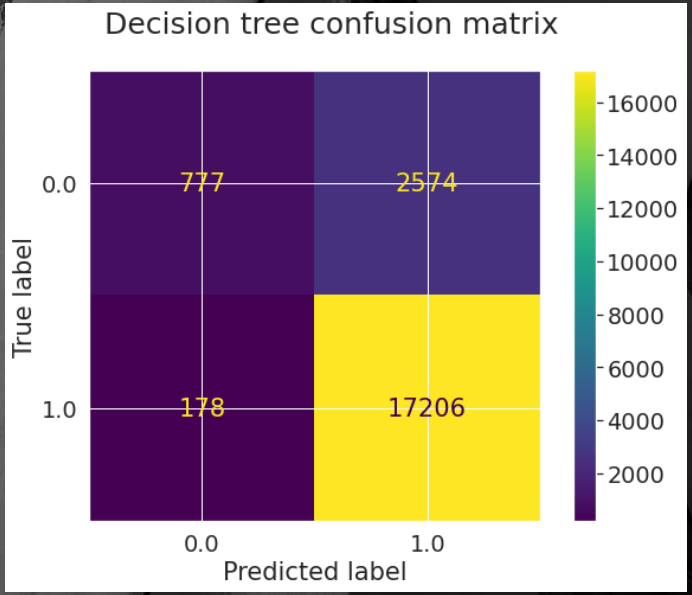


# Feature and target splitting



# Decision tree

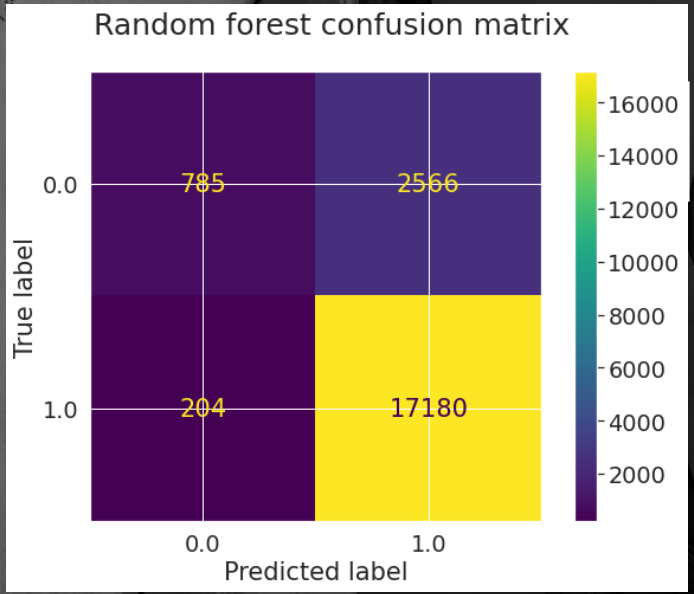
- Hyperparameter tuning process of cross validation only takes around 15 seconds.
- Train data accuracy is 0.8714% while the test data accuracy is 0.8673%.
- f1-score of this modelling is 0.93%.




```
Decision tree model
Accuracy Training Data: 0.8714251265975403
Accuracy Test Data: 0.8672775500361707

Decision tree model
              precision    recall  f1-score   support

         0.0       0.81      0.23      0.36      3351
         1.0       0.87      0.99      0.93     17384

    accuracy                           0.87     20735
   macro avg       0.84      0.61      0.64     20735
weighted avg       0.86      0.87      0.83     20735
```

# Random forest

- Hyperparameter tuning process of cross validation takes around 2 minutes.
- Train data accuracy is 0.8701% while the test data accuracy is 0.8664%.
- f1-score of this modelling is 0.93% same as the decision tree modelling.



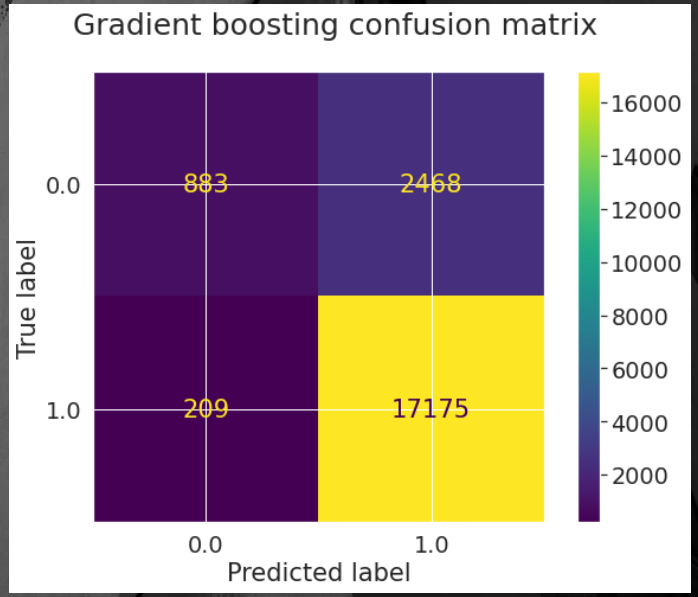

```
Random forest model
Accuracy Training Data: 0.870122980467808
Accuracy Test Data: 0.8664094526163492

Random Forest Model
              precision    recall  f1-score   support

         0.0       0.79      0.23      0.36      3351
         1.0       0.87      0.99      0.93     17384

    accuracy                           0.87     20735
   macro avg       0.83      0.61      0.64     20735
weighted avg       0.86      0.87      0.83     20735
```

# Gradient boosting

- Hyperparameter tuning process of cross validation takes more than 11 minutes, the longest than the other modelling.

- Train data accuracy is 0.8775% while the test data accuracy is 0.8709% making this model is the most fitted than the other modelling.

- f1-score of this modelling is 0.93% same as decision tree and random forest modelling.




```
Gradient boosting model
Accuracy Training Data: 0.8775259223535086
Accuracy Test Data: 0.8708946226187605

Gradient boosting model
              precision    recall  f1-score   support

         0.0       0.81      0.26      0.40      3351
         1.0       0.87      0.99      0.93     17384

    accuracy                           0.87     20735
   macro avg       0.84      0.63      0.66     20735
weighted avg       0.86      0.87      0.84     20735
```

# CONCLUSION

- Hyperparameter tuning and cross validation are used to get the best result on the modelling process by using the best parameters.
- f1-score of each tree-based modelling has a score of 0.93%.
- Gradient boosting modelling has the most fitted data result but also the longest hyperparameter cross validation process.
- It is suggested to use decision tree modelling as it has the most quickest process along with the high f1-score on predicting customer satisfaction.

# THANK YOU!

# ANY QUESTION?