Machine Learning in E-commerce Review Scoring

C2T4

Our Team

TingYi Lee (5610529)	Task: Data processing, run KNN model Presentation: Data Understanding, Data preparation
Jiayi Li (5630918)	Task: Data processing, run Decision Tree and Logistic Regression model Presentation: Data preparation
HaoDong Liu (5610099)	Task: Data processing, run RandomForest model Presentation: Modeling
Leah Xiao (5615963)	Task: Data processing, run SVM model Presentation: Business Understanding, Evaluation, Deployment

Table of contents

01

Business Understanding

Business Understanding | Data Understanding | Data Preparation | Modeling | Evaluation | Deployment

1. Background

2. Business Objectives

3. Data Overview

A Portuguese e-commerce company wants to use ML to predict customer review scores to identify customers more likely to leave positive reviews. This will allow them to target promotions towards these customers, enhancing their online reputation and increasing sales.

Business Understanding | Data Understanding | Data Preparation | Modeling | Evaluation | Deployment

1. Background

2. Business Objectives

3. Data Overview

- Build a prediction model to predict the customer who likely to leave good review.
- Improve the online reputation of the e-commerce platform by increasing the number of positive customer reviews.
- Use targeted marketing campaigns (emails, special promotions) to encourage satisfied customers to leave reviews.

Business Understanding → Data Understanding → Data Preparation → Modeling → Evaluation → Deployment

1. Background

2. Business Objectives

3. Data Overview

The company has provided 8 datasets, including:

Review Data: review scores, customer comments, timestamps

Customer Data: customer regions, unique customer ID

Order Item Data: products purchased, sellers, pricing, freight value

Payment Data: payment methods, payments sequential, instalment details

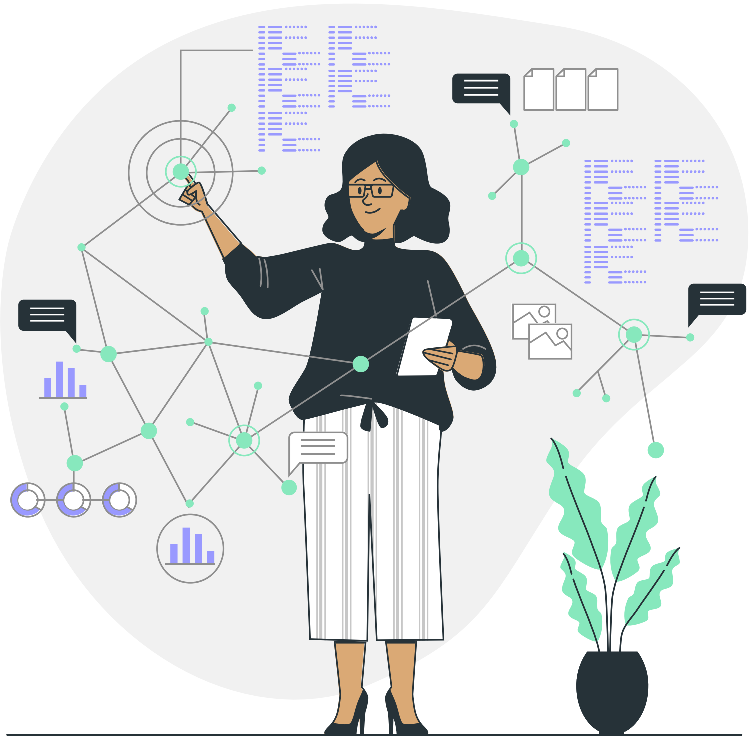
Order Data: order status, delivered carrier date, estimated delivery date

Geolocation Data: customer regions and zip code

Product Data: category name, description length, number of product photos

02

Data Understanding



Business Understanding | Data Understanding | Data Preparation | Modeling | Evaluation | Deployment

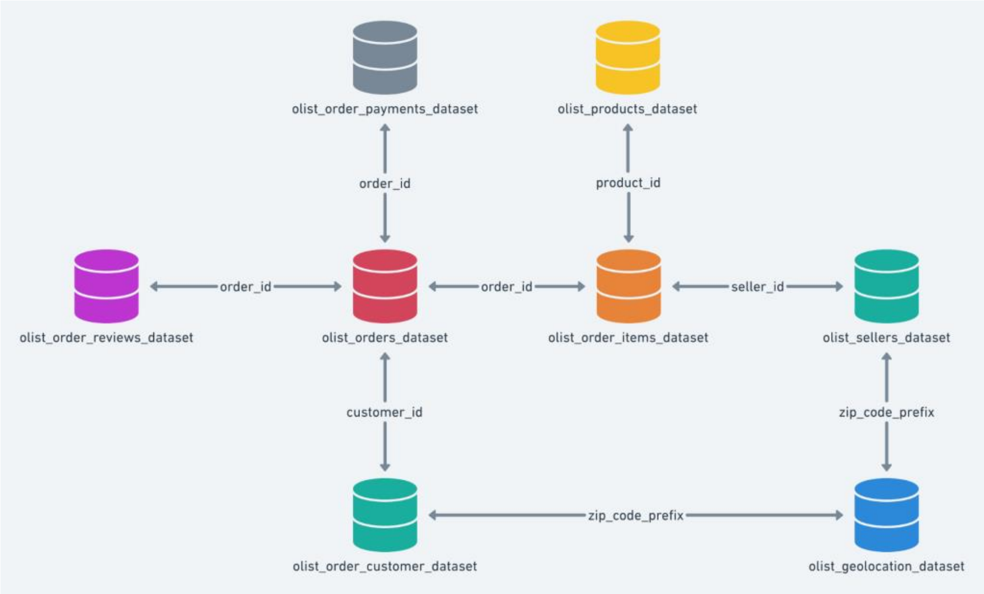
1.1 Acquire necessary data

1. Collect Initial Data

2. Describe Data

3. Explore Data

4. Verify Data Quality



Business Understanding | **Data Understanding** | Data Preparation | Modeling | Evaluation | Deployment

2.1 Number of records

1. Collect Initial Data
2. **Describe Data**
3. Explore Data
4. Verify Data Quality

All data is in CSV format. The data volume is around 100,000 records.

DataFrame Name	Rows	Columns
order_reviews_df	89999	7
orders_df	99441	8
order_payments_df	103886	5
order_items_df	112650	7
products_df	32951	9
customers_df	99441	4
sellers_df	3095	3
geolocation_df	1000163	4

Business Understanding | **Data Understanding** | Data Preparation | Modeling | Evaluation | Deployment

2.2 Number of missing records

1. Collect Initial Data
2. **Describe Data**
3. Explore Data
4. Verify Data Quality

```
order_reviews_df missing value
review_id                    0
order_id                     0
review_score                 0
review_comment_title     79404
review_comment_message   52429
review_creation_date         0
review_answer_timestamp      0
dtype: int64
```

```
orders_df missing value
order_id                         0
customer_id                      0
order_status                     0
order_purchase_timestamp         0
order_approved_at              160
order_delivered_carrier_date  1783
order_delivered_customer_date 2965
order_estimated_delivery_date    0
dtype: int64
```

```
order_payments_df missing value
order_id                0
payment_sequential      0
payment_type            0
payment_installments    0
payment_value           0
dtype: int64
```

```
order_items_df missing value
order_id               0
order_item_id          0
product_id             0
seller_id              0
shipping_limit_date    0
price                  0
freight_value          0
dtype: int64
```

```
products_df missing value
product_id                   0
product_category_name      610
product_name_lenght        610
product_description_lenght 610
product_photos_qty         610
product_weight_g             2
product_length_cm            2
product_height_cm            2
product_width_cm             2
dtype: int64
```

```
customers_df missing value
customer_id               0
customer_unique_id        0
customer_zip_code_prefix  0
customer_region           0
dtype: int64
```

```
sellers_df missing value
seller_id               0
seller_zip_code_prefix  0
seller_code             0
dtype: int64
```

```
geolocation_df missing value
geolocation_zip_code_prefix  0
geolocation_lat              0
geolocation_lng              0
geolocation_code             0
dtype: int64
```

Business Understanding | Data Understanding | Data Preparation | Modeling | Evaluation | Deployment

3.1 Multiple order_id records

1. Collect Initial Data

2. Describe Data

3. Explore Data

4. Verify Data Quality

- **Order item data**

  ➡ Customer may buy different items in an order.

- **Review data**

  ➡ Same or different review score for different items in an order.

- **Payment data**

  ➡ Multiple payment method for an order, ex. Card + voucher.

Business Understanding | **Data Understanding** | Data Preparation | Modeling | Evaluation | Deployment

3.2 Inconsistent records when merge data

1. Collect Initial Data

2. Describe Data

3. Explore Data

4. Verify Data Quality

Order_review

Order_id	Review_score
Order_1	5
Order_1	4

Order_item

Order_id	product_id
Order_1	product_1
Order_1	product_2
Order_1	product_3

merged

Order_id	product_id	review_score
Order_1	product_1	5
Order_1	product_1	4
Order_1	product_2	5
Order_1	product_2	4
Order_1	product_2	5
Order_1	product_3	4

Grouped by order_id
'review_score': 'mean'
.apply(np.floor)

Order_id	Review_score
Order_1	5
Order_1	4

Order_id	review_score
Order_1	4.5

Order_id	review_score
Order_1	4

Order_id	product_id
Order_1	product_1
Order_1	product_2
Order_1	product_3

Order_id	product_id	review_score
Order_1	product_1	4
Order_1	product_2	4
Order_1	product_3	4

Business Understanding | **Data Understanding** | Data Preparation | Modeling | Evaluation | Deployment

3.3 Duplicate records when merge data

1. Collect Initial Data

2. Describe Data

3. Explore Data

4. Verify Data Quality

Order_review

Order_id	review_score
Order_1	5
Order_1	5

Order_item

Order_id	product_id
Order_1	product_1
Order_1	product_2
Order_1	product_3

merged

Order_id	product_id	review_score
Order_1	product_1	5
Order_1	product_1	5
Order_1	product_2	5
Order_1	product_2	5
Order_1	product_3	5
Order_1	product_3	5

Grouped by order_id
'review_score': 'mean'
.apply(np.floor)

Order_id	review_score
Order_1	5
Order_1	5

Order_id	review_score
Order_1	5

Order_id	review_score
Order_1	5

Order_id	product_id
Order_1	product_1
Order_1	product_2
Order_1	product_3

Order_id	product_id	review_score
Order_1	product_1	5
Order_1	product_2	5
Order_1	product_3	5

Business Understanding | Data Understanding | Data Preparation | Modeling | Evaluation | Deployment

4.1 Quality issues

1. Collect Initial Data

2. Describe Data

3. Explore Data

4. Verify Data Quality

Multiple records

How we merge data between datasets? What is the logic when joining datasets?
Ex. Order_id

Missing Value

How we deal with missing values for each case? By using median to impute or drop the records?
Ex. Some datetime fields

Duplicate Records

Should the duplicate records be dropped? What are the reasons to drop the records. Ex. Duplicate review scores for the same order and item

Invalid Values

How we manage invalid values? Ex. Purchase time is later than delivery time

Datatype issue

When calculate a period of time, the datatime should be transformed.

03

Data Preparation

Business Understanding | Data Understanding | Data Preparation | Modeling | Evaluation | Deployment

1.1 orders_review as the target

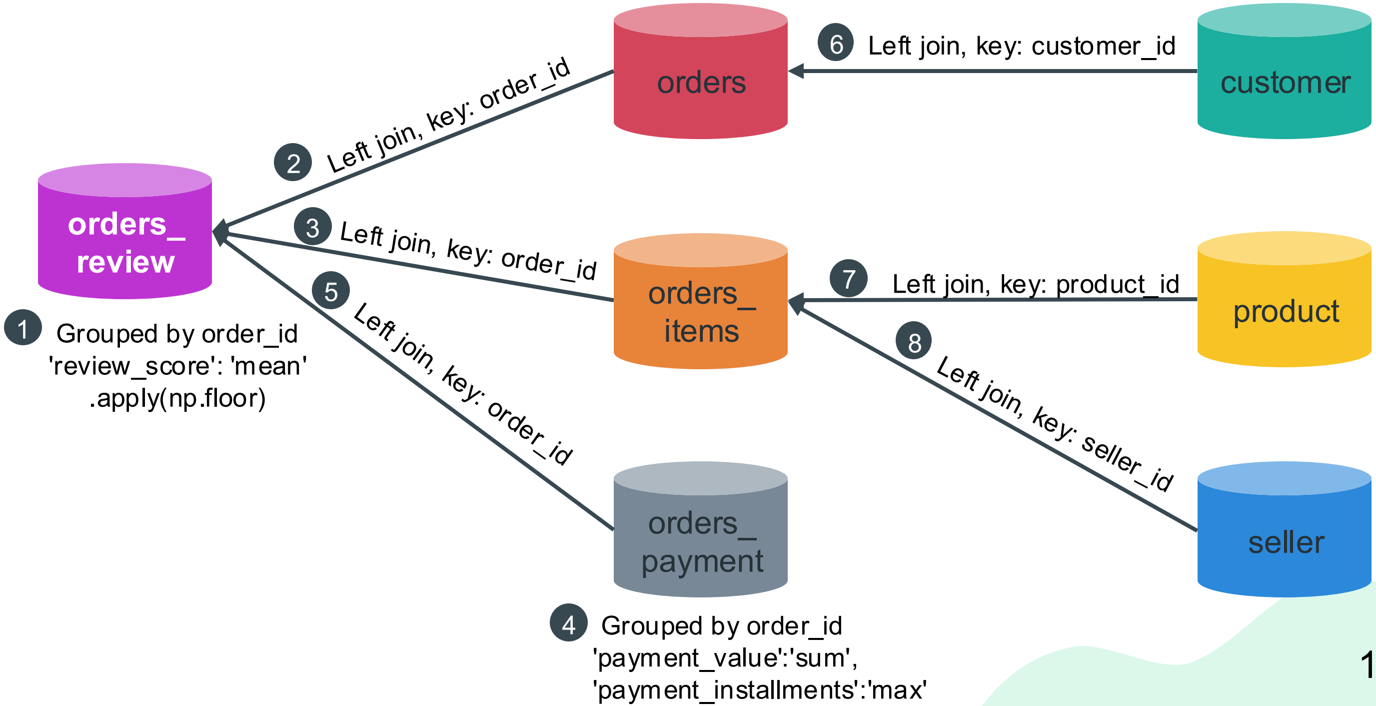
1. Data Integration

2. Data Cleaning

3. Data Aggregation

4. Feature Engineering

5. Data Description



Business Understanding | Data Understanding | Data Preparation | Modeling | Evaluation | Deployment

2.1 Remove record which has at least 5 missing fields

1. Data Integration

2. Data Cleaning

3. Data Aggregation

4. Feature Engineering

5. Data Description

```
(102238, 34)
order_id                         0
review_score                     0
has_effective_comment            0
review_creation_date             0
review_answer_timestamp          0
customer_id                      0
order_status                     0
order_purchase_timestamp         0
order_approved_at              145
order_delivered_carrier_date  1744
order_delivered_customer_date 2886
order_estimated_delivery_date    0
payment_value                    3
payment_installments             3
used_voucher                     3
order_item_id                  699
product_id                     699
seller_id                      699
shipping_limit_date            699
price                          699
freight_value                  699
customer_unique_id               0
customer_zip_code_prefix         0
customer_region                  0
product_category_name         2127
product_name_lenght           2127
product_description_lenght    2127
product_photos_qty            2127
product_weight_g               717
product_length_cm              717
product_height_cm              717
product_width_cm               717
seller_zip_code_prefix         699
seller_code                    699
```

Remove records which has at least 5 missing fields

Removed 749 records

```
(101489, 34)
order_id                         0
review_score                     0
has_effective_comment            0
review_creation_date             0
review_answer_timestamp          0
customer_id                      0
order_status                     0
order_purchase_timestamp         0
order_approved_at               14
order_delivered_carrier_date  1013
order_delivered_customer_date 2154
order_estimated_delivery_date    0
payment_value                    3
payment_installments             3
used_voucher                     3
order_item_id                    0
product_id                       0
seller_id                        0
shipping_limit_date              0
price                            0
freight_value                    0
customer_unique_id               0
customer_zip_code_prefix         0
customer_region                  0
product_category_name         1378
product_name_lenght           1378
product_description_lenght    1378
product_photos_qty            1378
product_weight_g                 1
product_length_cm                1
product_height_cm                1
product_width_cm                 1
seller_zip_code_prefix           0
seller_code                      0
```

Business Understanding | Data Understanding | Data Preparation | Modeling | Evaluation | Deployment

2.2 Setting datetime value as NaN for certain records

1. Data Integration

2. Data Cleaning

3. Data Aggregation

4. Feature Engineering

5. Data Description

```
(101489, 34)
order_id                         0
review_score                     0
has_effective_comment            0
review_creation_date             0
review_answer_timestamp          0
customer_id                      0
order_status                     0
order_purchase_timestamp         0
order_approved_at               14
order_delivered_carrier_date  1013
order_delivered_customer_date 2154
order_estimated_delivery_date    0
payment_value                    3
payment_installments             3
used_voucher                     3
order_item_id                    0
product_id                       0
seller_id                        0
shipping_limit_date              0
price                            0
freight_value                    0
customer_unique_id               0
customer_zip_code_prefix         0
customer_region                  0
product_category_name         1378
product_name_lenght           1378
product_description_lenght    1378
product_photos_qty            1378
product_weight_g                 1
product_length_cm                1
product_height_cm                1
product_width_cm                 1
seller_zip_code_prefix           0
seller_code                      0
```

If one of the datetime is NaN, Set all fields into NaN

Purpose: prevent incorrect data when conducting feature engineering

```
order_id                         0
review_score                     0
has_effective_comment            0
review_creation_date             0
review_answer_timestamp          0
customer_id                      0
order_status                     0
order_purchase_timestamp         0
order_approved_at               14
order_delivered_carrier_date  2155
order_delivered_customer_date 2155
order_estimated_delivery_date 2155
payment_value                    3
payment_installments             3
used_voucher                     3
order_item_id                    0
product_id                       0
seller_id                        0
shipping_limit_date              0
price                            0
freight_value                    0
customer_unique_id               0
customer_zip_code_prefix         0
customer_region                  0
product_category_name         1378
product_name_lenght           1378
product_description_lenght    1378
product_photos_qty            1378
product_weight_g                 1
product_length_cm                1
product_height_cm                1
product_width_cm                 1
seller_zip_code_prefix           0
seller_code                      0
```

Business Understanding | Data Understanding | Data Preparation | Modeling | Evaluation | Deployment

2.3 Filling missing values using median

1. Data Integration

2. Data Cleaning

3. Data Aggregation

4. Feature Engineering

5. Data Description

order_id	0
review_score	0
has_effective_comment	0
review_creation_date	0
review_answer_timestamp	0
customer_id	0
order_status	0
order_purchase_timestamp	0
order_approved_at	14
order_delivered_carrier_date	2155
order_delivered_customer_date	2155
order_estimated_delivery_date	2155
payment_value	3
payment_installments	3
used_voucher	3
order_item_id	0
product_id	0
seller_id	0
shipping_limit_date	0
price	0
freight_value	0
customer_unique_id	0
customer_zip_code_prefix	0
customer_region	0
product_category_name	1378
product_name_lenght	1378
product_description_lenght	1378
product_photos_qty	1378
product_weight_g	1
product_length_cm	1
product_height_cm	1
product_width_cm	1
seller_zip_code_prefix	0
seller_code	0

Filling missing value with median

order_id	0
review_score	0
has_effective_comment	0
review_creation_date	0
review_answer_timestamp	0
customer_id	0
order_status	0
order_purchase_timestamp	0
order_approved_at	14
order_delivered_carrier_date	2155
order_delivered_customer_date	2155
order_estimated_delivery_date	2155
payment_value	3
payment_installments	0
used_voucher	3
order_item_id	0
product_id	0
seller_id	0
shipping_limit_date	0
price	0
freight_value	0
customer_unique_id	0
customer_zip_code_prefix	0
customer_region	0
product_category_name	1378
product_name_lenght	0
product_description_lenght	0
product_photos_qty	0
product_weight_g	1
product_length_cm	1
product_height_cm	1
product_width_cm	1
seller_zip_code_prefix	0
seller_code	0

2.4 Drop unnecessary features

1. Data Integration

2. Data Cleaning

3. Data Aggregation

4. Feature Engineering

5. Data Description

```
order_id                          0
review_score                      0
has_effective_comment             0
review_creation_date              0
review_answer_timestamp           0
customer_id                       0
order_status                      0
order_purchase_timestamp          0
order_approved_at                14
order_delivered_carrier_date   2155
order_delivered_customer_date  2155
order_estimated_delivery_date  2155
payment_value                     3
payment_installments              0
used_voucher                      0
order_item_id                     0
product_id                        0
seller_id                         0
shipping_limit_date               0
price                             0
freight_value                     0
customer_unique_id                0
customer_zip_code_prefix          0
customer_region                   0
product_category_name          1378
product_name_lenght               0
product_description_lenght        0
product_photos_qty                0
product_weight_g                  1
product_length_cm                 1
product_height_cm                 1
product_width_cm                  1
seller_zip_code_prefix            0
seller_code                       0
```

Drop

```
order_id                          0
review_score                      0
has_effective_comment             0
review_creation_date              0
review_answer_timestamp           0
customer_id                       0
order_status                      0
order_purchase_timestamp          0
order_delivered_carrier_date   2155
order_delivered_customer_date  2155
order_estimated_delivery_date  2155
payment_installments              0
used_voucher                      0
order_item_id                     0
product_id                        0
seller_id                         0
shipping_limit_date               0
price                             0
freight_value                     0
customer_unique_id                0
customer_zip_code_prefix          0
customer_region                   0
product_name_lenght               0
product_description_lenght        0
product_photos_qty                0
seller_zip_code_prefix            0
seller_code                       0
```

Business Understanding | Data Understanding | Data Preparation | Modeling | Evaluation | Deployment

1. Data Integration

2. Data Cleaning

3. Data Aggregation

4. Feature Engineering

5. Data Description

new_merge_dataset

df.shape[0] ≠ df['order_id'].nunique()

101489 ≠ 88811

Duplicate order_id

Business Understanding | Data Understanding | Data Preparation | Modeling | Evaluation | Deployment

1. Data Integration

2. Data Cleaning

3. Data Aggregation

4. Feature Engineering

5. Data Description

Different Products in the Same Order

order_id	product_id	order_item_id	review_score	has_effective_comment	price
01144cadcf64b6427f0a6580a3033220	b8a0d73b2a06e7910d9864dccdb0cda2	1.0	3.0	0	59.9
01144cadcf64b6427f0a6580a3033220	9351b1e4334769dc0abe871ee3c7abc3	2.0	3.0	0	62.0

Same Product with Multiple Quantities

order_id	product_id	order_item_id	review_score	has_effective_comment	price
e9d40a10468b79b4c35c82f1bf078545	b114bf337c0626166abe574eee9e3f32	1.0	5.0	0	149.94
e9d40a10468b79b4c35c82f1bf078545	b114bf337c0626166abe574eee9e3f32	2.0	5.0	0	149.94
e9d40a10468b79b4c35c82f1bf078545	b114bf337c0626166abe574eee9e3f32	3.0	5.0	0	149.94
e9d40a10468b79b4c35c82f1bf078545	b114bf337c0626166abe574eee9e3f32	4.0	5.0	0	149.94
e9d40a10468b79b4c35c82f1bf078545	b114bf337c0626166abe574eee9e3f32	5.0	5.0	0	149.94

Business Understanding | Data Understanding | Data Preparation | Modeling | Evaluation | Deployment

1. Data Integration

2. Data Cleaning

3. Data Aggregation

4. Feature Engineering

5. Data Description

```
order_review_df = order_review_df.groupby(['order_id', 'product_id'], as_index=False).agg({
    'review_score': 'first',
    'has_effective_comment': 'first',
    'review_creation_date' : 'first',
    'review_answer_timestamp': 'first',
    'customer_id': 'first',
    'order_status': 'first',
    'order_purchase_timestamp' : 'first',
    'order_delivered_carrier_date' : 'first',
    'order_delivered_customer_date' : 'first',
    'order_estimated_delivery_date' : 'first',
    'payment_value' : 'first',
    'payment_installments':'first',
    'used_voucher': 'first',
    'price': 'sum',
    'freight_value': 'sum',
    'product_id':'first',
    'seller_id': 'first',
    'customer_region':'first',
    'product_category_name':'first',
    'product_description_lenght': 'first',
    'product_photos_qty': 'first',
    'seller_code': 'first',
})
```

Business Understanding → Data Understanding → Data Preparation → Modeling → Evaluation → Deployment

1. Data Integration
2. Data Cleaning
3. Data Aggregation
4. **Feature Engineering**
5. Data Description

Feature Construction → **Encoding Categorical Feature** → **Rescale Data** → **Feature Selection**

Business Understanding → Data Understanding → **Data Preparation** → Modeling → Evaluation → Deployment

Feature Construction

1. Data Integration
2. Data Cleaning
3. Data Aggregation
4. **Feature Engineering**
5. Data Description

```
order_purchase_timestamp          object
order_delivered_carrier_date      object
order_delivered_customer_date     object
order_estimated_delivery_date     object
```

↓ to_datetime

```
order_purchase_timestamp          datetime64[ns]
order_delivered_carrier_date      datetime64[ns]
order_delivered_customer_date     datetime64[ns]
order_estimated_delivery_date     datetime64[ns]
```

Business Understanding | Data Understanding | Data Preparation | Modeling | Evaluation | Deployment

Feature Construction

1. Data Integration
2. Data Cleaning
3. Data Aggregation
4. **Feature Engineering**
5. Data Description

order_purchase_timestamp

order_delivered_carrier_date

→ dispatch_days

order_delivered_customer_date

order_estimated_delivery_date

Business Understanding | Data Understanding | Data Preparation | Modeling | Evaluation | Deployment

Feature Construction

1. Data Integration
2. Data Cleaning
3. Data Aggregation
4. Feature Engineering
5. Data Description

order_purchase_timestamp

order_delivered_carrier_date

order_delivered_customer_date

order_estimated_delivery_date

genuine_delivery_days

Business Understanding | Data Understanding | Data Preparation | Modeling | Evaluation | Deployment

Feature Construction

1. Data Integration

2. Data Cleaning

3. Data Aggregation

4. Feature Engineering

5. Data Description

order_purchase_timestamp

order_delivered_carrier_date

order_delivered_customer_date

order_estimated_delivery_date

delivery_diff_days (+ /−)

whether_exceed_estimated (0 /1)

Business Understanding | Data Understanding | Data Preparation | Modeling | Evaluation | Deployment

Feature Construction

1. Data Integration
2. Data Cleaning
3. Data Aggregation
4. Feature Engineering
5. Data Description

order_purchase_timestamp	order_delivered_carrier_date
2017-07-25 18:57:58	2017-07-26 17:43:33
2018-07-26 22:42:32	2018-07-27 14:34:00

Raw Time Calculation

dispatch_days = 0 (< 24 h) ❌

.dt.normalize()

Day-Based Adjustment

dispatch_days = 1 ✅

Business Understanding → Data Understanding → Data Preparation → Modeling → Evaluation → Deployment

Feature Construction

1. Data Integration
2. Data Cleaning
3. Data Aggregation
4. **Feature Engineering**
5. Data Description

review_answer_timestamp, review_creation_date → review_response_time

price, freight_value $\xrightarrow{\div}$ freight_ratio

seller_code, customer_region → same_region (0 /1)

Business Understanding | Data Understanding | Data Preparation | Modeling | Evaluation | Deployment

Feature Construction

1. Data Integration
2. Data Cleaning
3. Data Aggregation
4. **Feature Engineering**
5. Data Description

dispatch_days	1961
genuine_delivery_days	1961
delivery_diff_days	1961
review_response_time_h	0
freight_ratio	0
same_region	0

delivery_diff_days
90276.000000
-12.006635
10.180449
-147.000000
[illegible]
[illegible]
[illegible]
[illegible]

✅ **Missing values imputed using the median**

dispatch_days	0
genuine_delivery_days	0
delivery_diff_days	0
review_response_time_h	0
freight_ratio	0
same_region	0

same_region
[illegible]2237.000000
0.359747
0.479929
0.000000
0.000000
0.000000
1.000000
1.000000

Business Understanding | Data Understanding | Data Preparation | Modeling | Evaluation | Deployment

Encoding Categorical Feature

1. Data Integration
2. Data Cleaning
3. Data Aggregation
4. **Feature Engineering**
5. Data Description

	product_category_name_english	product_category_name
Electronics & Accessories	computers_accessories	informatica_acessorios
	tablets_printing_image	tablets_impressao_imagem
	fixed_telephony	telefonia_fixa
	telephony	telefonia
	consoles_games	consoles_games
	audio	audio
	electronics	eletronicos
Home & Furniture	furniture_decor	moveis_decoracao
	bed_bath_table	cama_mesa_banho
	kitchen_dining_laundry_garden_furniture	moveis_cozinha_area_de_servico_jantar_e_jardim
	housewares	utilidades_domesticas
	home_comfort	
	home_comfort_2	casa_conforto_2
	home_appliances	eletrodomesticos
	home_appliances_2	eletrodomesticos_2
	small_appliances	eletroportateis

Business Understanding | Data Understanding | Data Preparation | Modeling | Evaluation | Deployment

Encoding Categorical Feature

1. Data Integration
2. Data Cleaning
3. Data Aggregation
4. Feature Engineering
5. Data Description

```
product_category_name_Electronics & Accessories     0
product_category_name_Entertainment & Hobbies       0
product_category_name_Fashion & Personal Care       0
product_category_name_Food & Daily Essentials       0
product_category_name_Home & Furniture              0
product_category_name_Industry & Construction       0
product_category_name_Sports & Outdoor              0
product_category_name_unknown                       0
```

5 Industry & Construction

6 Entertainment & Hobbies

7 Sports & Outdoor

8 Unknown

- ✅ Reduce dimensionality
- ✅ Enhance computational efficiency

Business Understanding | Data Understanding | Data Preparation | Modeling | Evaluation | Deployment

Encoding Categorical Feature

1. Data Integration

2. Data Cleaning

3. Data Aggregation

4. Feature Engineering

5. Data Description

Order Status	Count	Average_review_score
delivered	99,335	4.067
approved	3	2.000
canceled	459	1.627
invoiced	309	1.654
processing	299	1.344
shipped	1,077	1.982
unavailable	7	1.571

Encoding Categorical Feature

1. Data Integration

2. Data Cleaning

3. Data Aggregation

4. Feature Engineering

5. Data Description

Order status
delivered 1
others 0

Binary Encoding

✅ Mitigates data imbalance issues

✅ Improves model efficiency

Business Understanding | Data Understanding | Data Preparation | Modeling | Evaluation | Deployment

Encoding Categorical Feature

1. Data Integration

2. Data Cleaning

3. Data Aggregation

4. Feature Engineering

5. Data Description

Payment Installments

1	44,291
2	11,410
3	9,665
4	6,567
5	4,855
6	3,686
7	1,531
8	4,082
9	598
10	5,210
11	24
12	140
13	16
14	11
15	73
16	6
17	6
18	27
20	16
21	3
22	1
24	19

Mode Imputation →

1	44,291
2	11,410
3	9,665
4	6,567
5	4,855
6	3,686
7	1,531
8	4,082
9	598
10	5,210
11	24
12	140
13	16
14	11
15	73
16	6
17	6
18	27
20	16
21	3
22	1
24	19

Business Understanding | Data Understanding | Data Preparation | Modeling | Evaluation | Deployment

Encoding Categorical Feature

1. Data Integration

2. Data Cleaning

3. Data Aggregation

4. Feature Engineering

5. Data Description

Payment Installments

1	44,291
2	11,410
3	9,665
4	6,567
5	4,855
6	3,686
7	1,531
8	4,082
9	598
10	5,210
11	24
12	140
13	16
14	11
15	73
16	6
17	6
18	27
20	16
21	3
22	1
24	19

Full Payment

Half-Year Plan

One-Year Plan

Two-Year Plan

One – Hot !

Business Understanding | Data Understanding | Data Preparation | Modeling | Evaluation | Deployment

Rescale Data

1. Data Integration
2. Data Cleaning
3. Data Aggregation
4. Feature Engineering
5. Data Description

```
review_response_time_h        float64
dispatch_days                 float64
genuine_delivery_days         float64
price                         float64
freight_value                 float64
product_description_lenght    float64
product_photos_qty            float64
delivery_diff_days            float64
```

Normalization (Min-Max)

1.rescale absolute values

2.restore original sign

⚠ **Rescaling is applied after dataset splitting**

✅ Retains early/late delivery distinction in normalized form

Feature Selection

1. Data Integration

2. Data Cleaning

3. Data Aggregation

4. Feature Engineering

5. Data Description

Original Features
➢ Order_status ➢ Price ➢ freight_value ➢ product_description_length ➢ product_photos_qty

Engineered Features
➢ Used_voucher ➢ review_response_time_h ➢ dispatch_days ➢ genuine_delivery_days ➢ delivery_diff_days ➢ whether_exceed_estimated ➢ freight_ratio ➢ same_region

Business Understanding | Data Understanding | Data Preparation | Modeling | Evaluation | Deployment

1. Data Integration

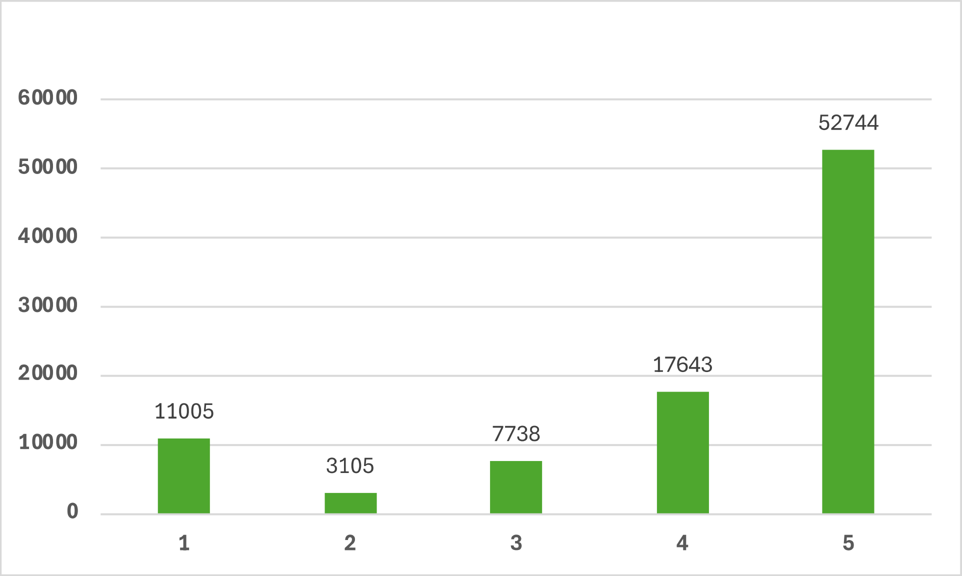
2. Data Cleaning

3. Data Aggregation

4. Feature Engineering

5. Data Description

Distribution of Review Scores



Review Scores

04

Modeling

Business Understanding | Data Understanding | Data Preparation | **Modeling** | Evaluation | Deployment

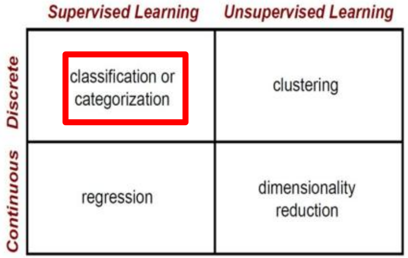
1. Rationale for Model Selection

2. Issue with Five-Class

3. Hyperparameter Tuning & Feature Selection

4. Model Performance Analysing

We Choose KNN, SVM, Random Forest, Logistic Regression, and Decision Tree as our models.



1. Determine the type of task

Labeled Data → Supervised Learning
User Ratings (1-5 stars) → Discrete Target Variable

2. Select appropriate models Our selection is based on three key factors:

(1) Model Diversity for Robustness

- Captures both linear and non-linear patterns
  - Linear models: Logistic Regression, SVM (with linear kernel)
  - Non-linear models: Decision Tree, Random Forest, KNN, SVM
- Cover both simple and complex decision boundaries
  - Simple models: Logistic Regression, Decision Tree
  - Complex models: Random Forest, SVM, KNN

We Choose KNN, SVM, Random Forest, Logistic Regression, and Decision Tree as our models.

2. Select appropriate models

(2) Performance on Structured E-Commerce Data

E-commerce data contains a mix of nominal and numerical features, requiring models that handle different data types efficiently

(3) Interpretability vs. Predictive Power Trade-Off

Highly interpretable: Decision Tree, Logistic Regression

Strong predictive power, but less interpretable: Random Forest, SVM & KNN

----This selection ensures that we explore different modeling approaches to find the best fit for our e-commerce rating prediction.

Business Understanding | Data Understanding | Data Preparation | **Modeling** | Evaluation | Deployment

1. Rationale for Model Selection

2. Issue with Five-Class

3. Hyperparameter Tuning & Feature Selection

4. Model Performance Analysing

Classification Report Summary

Random Forest (61.51%)

	precision	recall	f1-score	support
1	0.67	0.41	0.51	2201
2	**1.00**	**0.00**	**0.01**	621
3	**0.73**	**0.01**	**0.01**	1548
4	**0.42**	**0.01**	**0.01**	3528
5	0.61	0.99	0.75	10549

KNN (59.84%)

	precision	recall	f1-score	support
1	0.59	0.40	0.47	2201
2	**0.00**	**0.00**	**0.00**	621
3	**0.12**	**0.01**	**0.01**	1548
4	**0.20**	**0.05**	**0.08**	3528
5	0.61	0.93	0.74	10549

SVM (60.87%)

	precision	recall	f1-score	support
1	0.60	0.40	0.48	2201
2	**0.00**	**0.00**	**0.00**	621
3	**0.00**	**0.00**	**0.00**	1548
4	**0.33**	**0.00**	**0.00**	3528
5	0.61	0.98	0.75	10549

Decision Tree (61.15%)

	precision	recall	f1-score	support
1	0.67	0.39	0.49	2201
2	**0.00**	**0.00**	**0.00**	621
3	**0.00**	**0.00**	**0.00**	1548
4	**0.00**	**0.00**	**0.00**	3528
5	0.61	0.99	0.75	10549

Business Understanding | Data Understanding | Data Preparation | Modeling | Evaluation | Deployment

Confusion Matrix Analysis

1. Rationale for Model Selection

2. Issue with Five-Class

3. Hyperparameter Tuning & Feature Selection

4. Model Performance Analysing

Random Forest

	1	2	3	4	5
1	905	0	0	5	**1291**
2	114	2	0	2	**503**
3	125	0	8	7	**1408**
4	78	0	1	20	**3429**
5	121	0	2	14	10412

KNN

	1	2	3	4	5
1	163	4	17	109	**1908**
2	32	11	2	31	**545**
3	42	1	27	94	**1384**
4	41	1	20	191	**3275**
5	125	3	36	466	9919

SVM

	1	2	3	4	5
1	900	0	0	0	**1349**
2	136	0	0	0	**485**
3	135	0	0	0	**1355**
4	125	0	0	1	**3416**
5	216	0	0	2	10327

Dicision Tree

	1	2	3	4	5
1	873	0	0	0	**1376**
2	121	0	0	0	**500**
3	99	0	0	0	**1391**
4	81	0	0	0	**3461**
5	138	0	0	0	10407

Issues with Five-Class Classification

- Classes 2, 3, 4 are almost always misclassified as class 5.
- Due to deep data imbalance, the model heavily favors predicting 5-star ratings.

Why Convert to Binary Classification?

- Business Justification: The goal is to identify potential positive reviewers (5-star ratings).
- Data Structure Optimization: Reducing class imbalance and improving model generalization. (a closer data volume)

Hyperparameter Tuning ---Using GridSearchCV

1. Rationale for Model Selection

2. Issue with Five-Class

3. Hyperparameter Tuning & Feature Selection

4. Model Performance Analysing

Decision Tree
```
param_grid = {
    'max_depth': np.arange(1, 8),               2
    'min_samples_split': np.arange(2, 8),       2
    'min_samples_leaf': [5,10,15]               5
}
```

SVM
```
param_grid = {
    'C':  [1,10],                              1
    'kernel':  ['linear', 'rbf', 'poly' ],     rbf
    'degree': [5,10,15],                       5
    'gamma': [0.01,0.1]                        0.01
}
```

KNN
```
param_grid = {
    'n_neighbours': list(range(3,16)),          15
    'weights': ['uniform', 'distance'],         distance
    'metric': ['euclidean', 'manhattan']        manhattan
}
```

Random Forest
```
param_grid = {
    'classifier__n_estimators': [200, 300, 500],       500
    'classifier__max_depth': [20, 25, 30],             25
    'classifier__min_samples_split': [5, 8, 12],       8
    'classifier__min_samples_leaf': [1, 2, 4]          2
}
```

Business Understanding | Data Understanding | Data Preparation | **Modeling** | Evaluation | Deployment

1. Rationale for Model Selection
2. Issue with Five-Class
3. Hyperparameter Tuning & Feature Selection
4. Model Performance Analysing

Feature Selection Optimization

Step 1 : Drop the features that we can't use

- 'review_score' ---- the variable that we need to classify

Step 2 : Run the model with the remaining features

- Evaluating changes in model performance after changing features
  ----If there is no significant performance degradation after removal, retain it

Step 3 : Get feature_importances/feature_coefficient from model

- Keep features with high importance/ high coefficient
- Remove redundant features (low importance or business irrelevant features)
- Repeat Step 2

Step 4 : Eventually get the conbination of features that enable the model to perform best

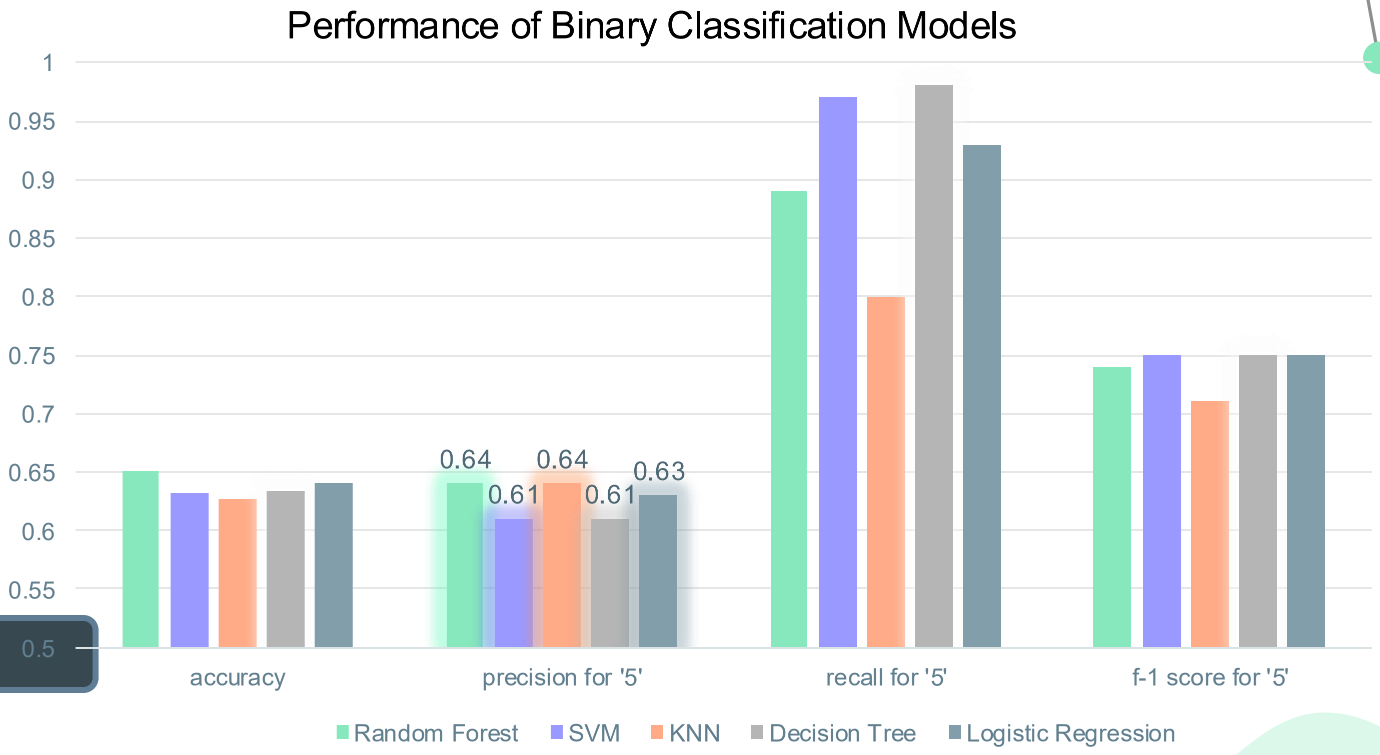
Business Understanding | Data Understanding | Data Preparation | **Modeling** | Evaluation | Deployment

1. Rationale for Model Selection

2. Issue with Five-Class

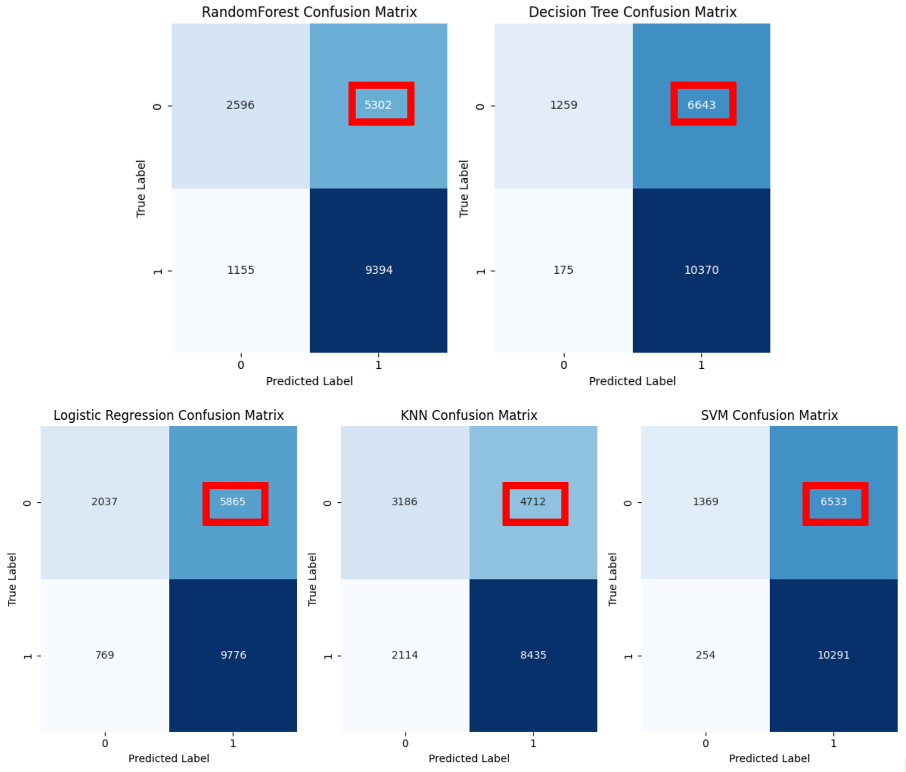
3. Hyperparameter Tuning & Feature Selection

4. Model Performance Analysing



Business Understanding | Data Understanding | Data Preparation | **Modeling** | Evaluation | Deployment

Confusion Matrix



1. Rationale for Model Selection

2. Issue with Five-Class

3. Hyperparameter Tuning & Feature Selection

4. Model Performance Analysing

05

Evaluation

Business Understanding | Data Understanding | Data Preparation | Modeling | **Evaluation** | Deployment

1. Indicator Explanation

2. Choose Model

Precision and Recall

- **Core business goal:** Accurately find the "right" customers to leave positive reviews, rather than asking all customers to leave reviews, to optimize brand image and market performance.
- **Focus on positive reviewers:** If the company wants to target users who are likely to give good reviews for marketing, focus on precision can ensure that a higher percentage of users predicted to give positive reviews give positive reviews.
- **Avoid false positives:** Mistakenly identifying users who are likely to give negative reviews as positive reviewers (i.e., false positives) can result in wasted resources and even damage to brand reputation.

1. Indicator Explanation

2. Choose Model

Random Forest

Accuracy = 0.6500

Precision = 0.64

Confusion Matrix = $\begin{bmatrix} 2596 & 5302 \\ 1155 & 9394 \end{bmatrix}$

- TP (9394): predicted good score and actually good score
- TN (2596): predicted bad score and actually bad score
- FP (5302): predicted good score but actually bad score (waste of resources and damage brand reputation)
- FN (1155): predicted bad score but actually good score (loss of opportunity)

Business Understanding | Data Understanding | Data Preparation | Modeling | Evaluation | Deployment

1. Indicator Explanation

2. Choose Model

- Recall = $\frac{TP}{(TP+FN)}$ = *0.89*

All the samples that were actually good score, 89% were correctly predicted as good score by the model, 11% of the positive examples were misclassified as bad score.

- Precision = $\frac{TP}{(TP+FN)}$ = 0.64

All the samples predicted by the model as good score, 64% are actually good score, which means that 36% of the predicted good score examples are bad score.

- F1 Score = $2 \times \frac{precision \times recall}{precision + recall}$ = 0.74

Indicates that the model is relatively stable and will not ignore too many real customers with positive reviews, nor will it introduce too many false positives

06

Deployment

1. Suggestion

2. Improvement

- Although the random forest model currently selected performs best, there is still room for improvement in accuracy and precision.
- Low precision means that there are still many false positives (FP), which means that promotional information may be sent to some customers who actually give low score, bringing potential risks.
- It is recommended to gradually promote it based on small-scale testing to reduce negative impacts.

1. Suggestion

2. Improvement

Engineer additional features

Such as sentiment analysis from customer comments, to enhance model predictions.

Introducing PCA

After feature engineering, we have many derived features (e.g. genuine_delivery_days and delivery_diff_days). Using them directly may lead to redundancy and increased computational cost.

Label Redefinition and Balancing the Dataset

Classify users with 5 score and comments as positive, and others as negative, to make the model more consistent with business goals. Use methods such as category weighting to handle imbalance and ensure effective model learning.

References

***** Different links are contributed by different teammates, which may not directly run in sequence. But we can make sure that all our data are consistent when running model.*****

- Data processing: https://colab.research.google.com/drive/1f578sqnGRiY3mXR1xS1jD6W5XBr8vSqz?usp=sharing

- Modeling
  - KNN Model: https://colab.research.google.com/drive/1jmECiPzqeFOtxW1AulVgFvtP_3N4kxYY?usp=sharing
  - Random forest: https://colab.research.google.com/drive/1sDJdJHfocOFeFwSwFAYUmMRydoFD1Bdn?usp=sharing
  - SVM: https://colab.research.google.com/drive/1hsV8QC5iJzg_0fUJFbOH6EMFNcW3GUWa?usp=sharing
  - Decision Tree: https://colab.research.google.com/drive/1oeybIkDwjSa4GKvnCS1bI-krnVWv4A_C?usp=sharing
  - Logistic Regression: https://colab.research.google.com/drive/1VBBsOa-1HnRQAeB4wjxK83nn0oM2VYFI?usp=sharing

Thanks!