Delivery Delay Prediction for E-Commerce Dataset

Delay or not, just give me an answer!

Group 19

Gursifath Bhasin (gb2760) ● Jonathan Benghiat (jb4653) ● Lukas Wang (bw2712)
Qi Meng (qm2162) ● Siddhant Pravin Mahurkar (sm5129)

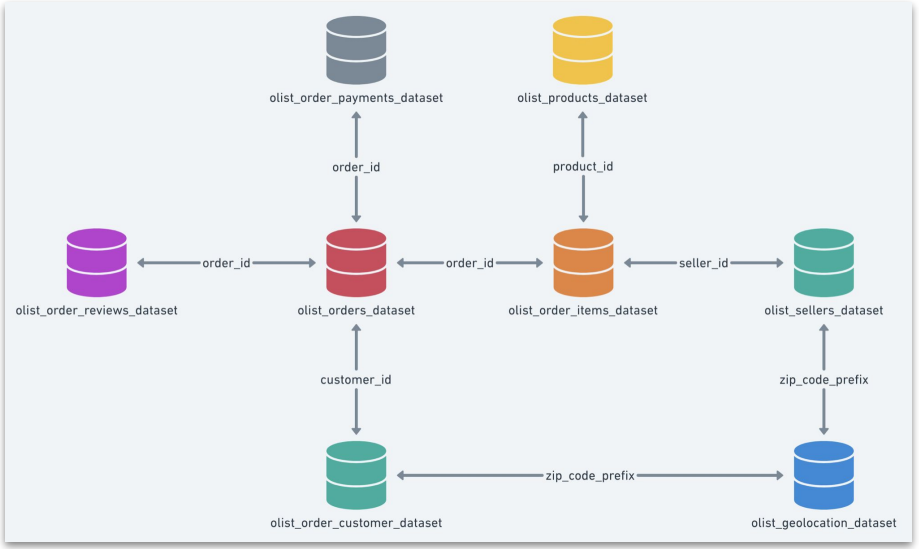
Overview of Dataset

Data Source: Brazilian E-Commerce Public Dataset by Olist

Dataset Context: real-world commerce data from Brazilian marketplaces from 2016 to 2018 with more than 100k records among 71 categories.

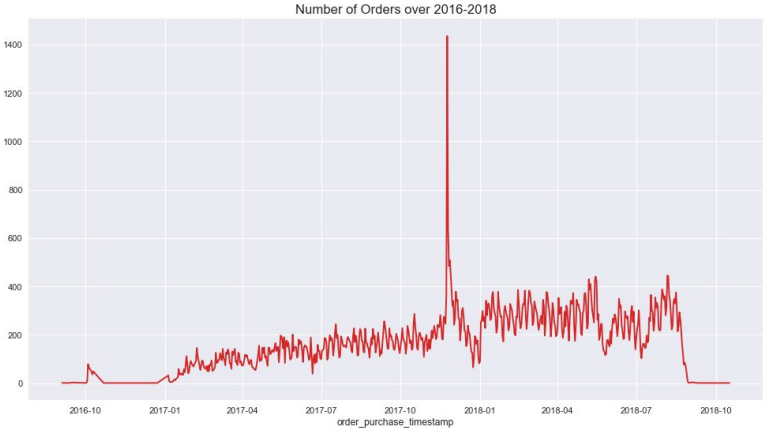
Problem with Original Dataset:

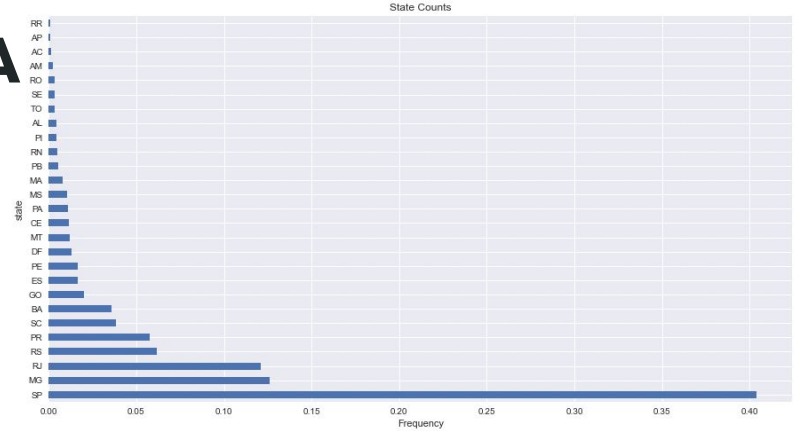
- Too many columns/features (50 columns).
- Many unrelated features.
- Highly imbalanced datasets.

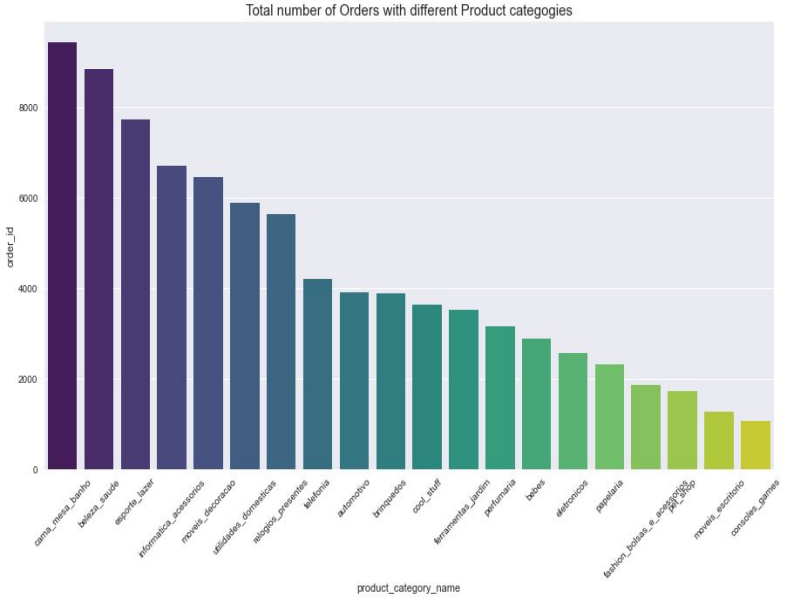


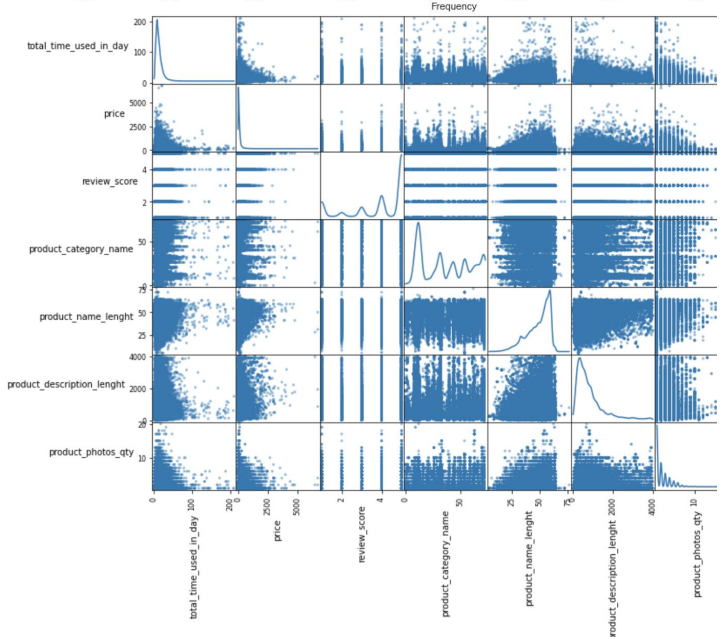
Schema of Dataset

Initial EDA

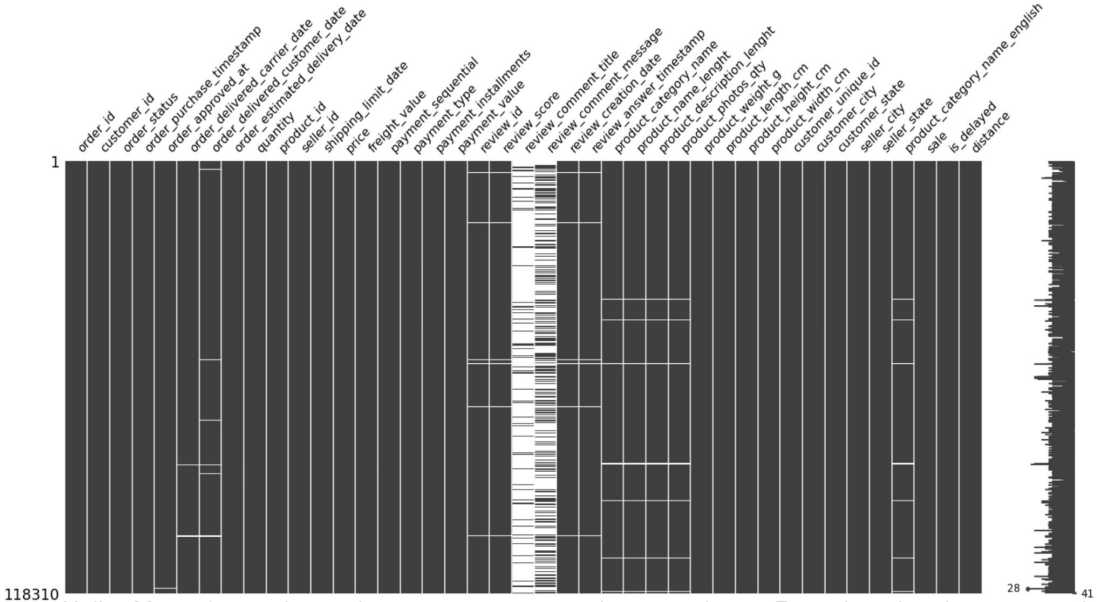

Number of Orders over 2016-2018
order_purchase_timestamp


State Counts
state
Frequency


Total number of Orders with different Product categogies
order_id
product_category_name


total_time_used_in_day
price
review_score
product_category_name
product_name_lenght
product_description_lenght
product_photos_qty

Missing Data Analysis



Nullity Matrix for quick visualization on patterns in data completion. Eg: order_id and customer_id seem to be completely populated while review_comment_title and review_comment_message appear to have the maximum missing values.

Keeping above insights in mind and those from EDA, we:

1. dropped certain attributes (columns) which don't contribute meaningfully to our model.
2. calculated new feature from the existing features which seem to contribute to the delay
3. dropped rows with small number of missing values.

Column Name	# of missing values	% of missing values
review_comment_title	104418	88.26
review_comment_message	68628	58.01
order_delivered_customer_date	2588	2.19
product_category_name_english	1734	1.47
product_photos_qty	1709	1.44
product_description_length	1709	1.44
product_name_length	1709	1.44
product_category_name	1709	1.44
seller_city	0	0.00
customer_id	0	0.00
price	0	0.00

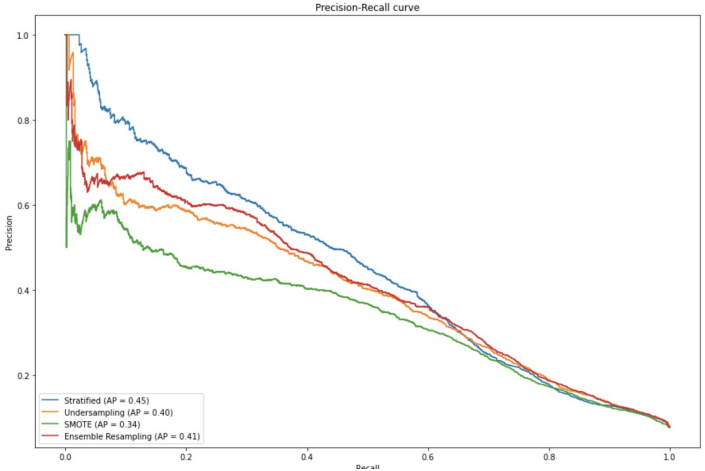
Snippet of missing values per data column

Data Sampling Techniques

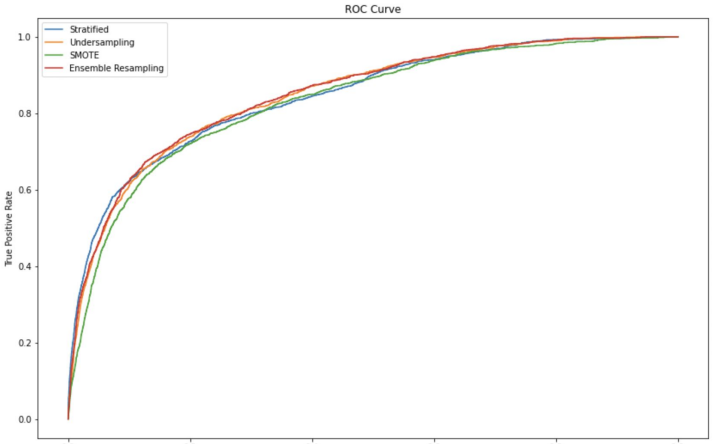
The original data is highly imbalanced where the minority class (order is delayed) represents roughly 7.5% of the total dataset.

We plan to try out four different sampling techniques in order to tackle the issue imbalanced dataset which are:

- Stratified Sampling
- Undersampling
- SMOTE
- Ensemble Resampling



PR curve of preliminary sampling experiments on random forest



ROC curve of preliminary sampling experiments on random forest

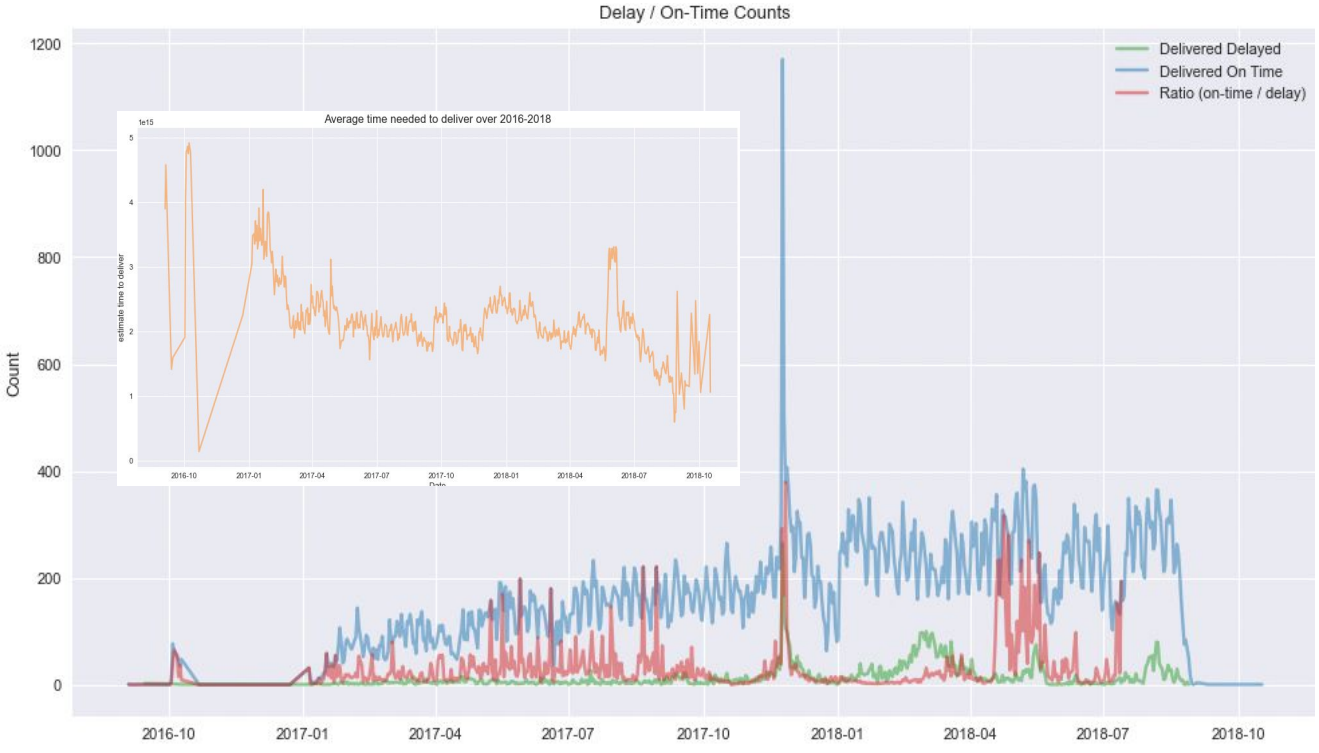
Does Time / Holiday Matter?

Black Friday **increases** the on-time / delay ratio!

WHY?

- System increases the estimated delivery date (**Wrong** → orange graph)
  - Delivery service is awesome (Large capacity)

So, time / holiday matters!



Review Matters?

Imbalanced

Over 77% of customers leave 4 or 5 rating in review

Sparse

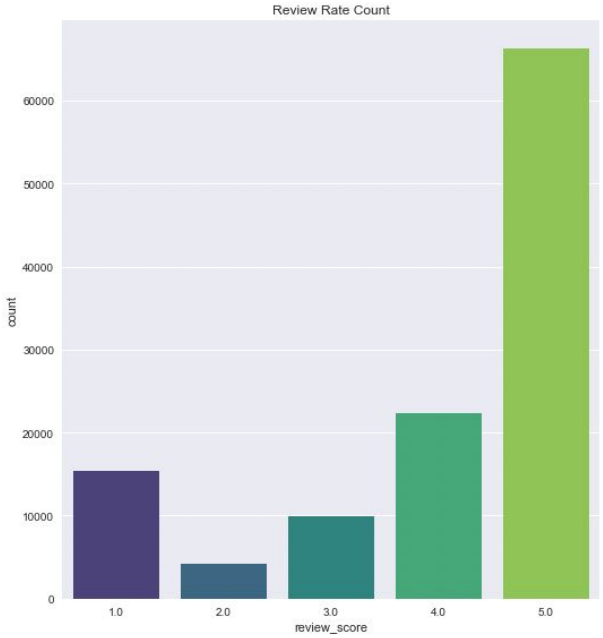
Over 58% of customer don't leave comments

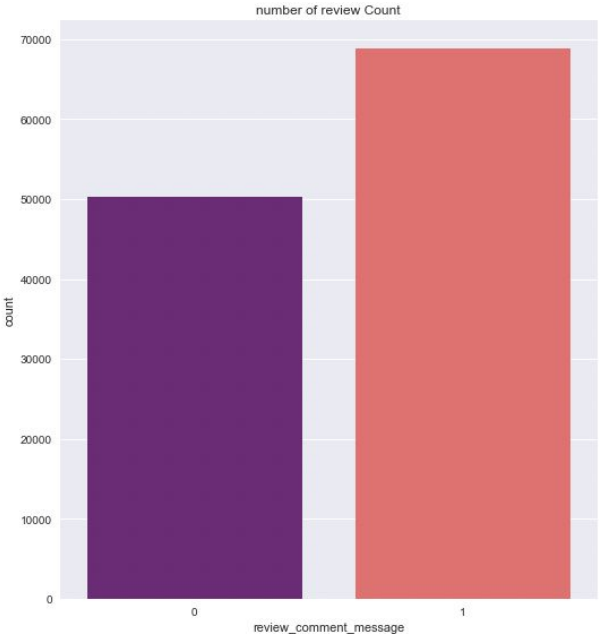
Hard to extract

Comment not in English

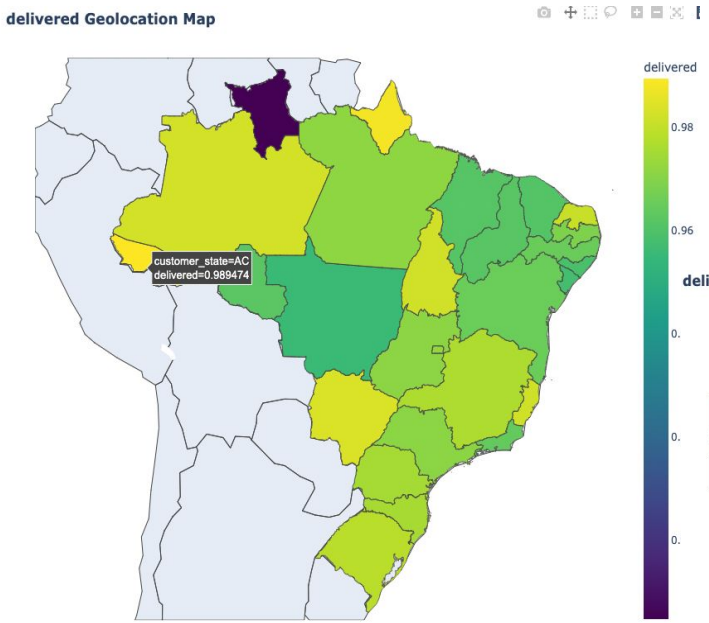
Basically the same

So, review does not matter.





Geolocation Matters?

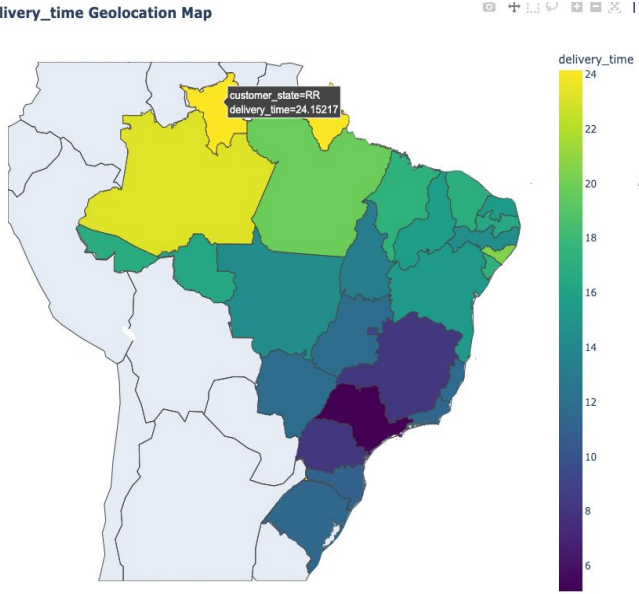


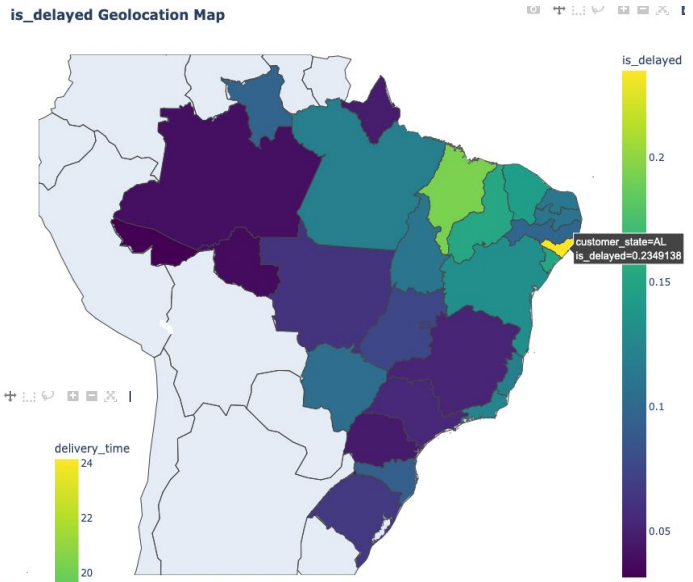
Delivered Rate

RR has lowest delivered rate, may be due to its position

Delivery time

RR is northeast so may need more time to deliver





Delay Rate

AL is highest and northeastern part tend to has larger delay rate

So, geolocation matters!

Do Numerical columns matter?

Payment?

Payment sequential seems to have some relation (basic outliers)

Review Score?

YES!!!

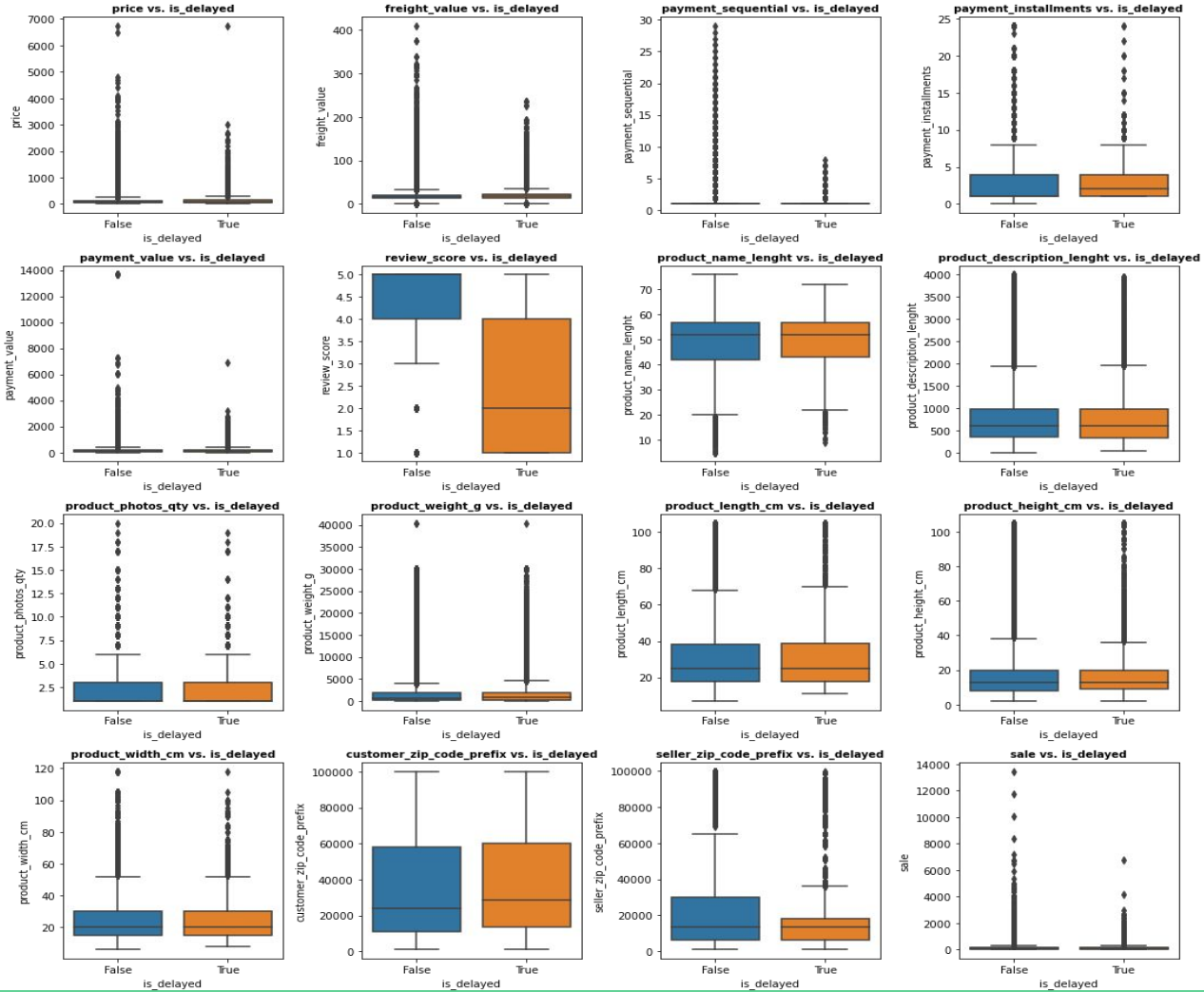
Product Size?

Can’t see much correlation.

Sales?

The basic outliers are related.

So, numerical data matters!



Proposed Method

Base Models:

Pipelining several traditional machine learning methods, and choose the one with the best performances.

Hyperparameter Tuning:

Grid Search & Random Search

Model Evaluation:

Recall & Precision & F1 Score & PR Curve & ROC Curve

Model Selection:

Stratified K-fold

Logistic Regression
SVM
Decision Tree
Random Forest
Naive Bayes
LDA
Gradient Boosting
XGBoost
LightGBM

Sample Methods