Detect Shipping Delay of E-commerce

Olivia Tian

Map of presentation

- Introduce topic
- Dataset overview
- Prepare data for machine learning
- Build machine learning model
- Model visulization and evaluation
- Conclusion

Business side

- Online shopping cover everywhere, platforms' competition are high.
- Shipping ontime has positive impact on customer experience, revisit and brand trust.
- Shipping delay involves intangible cost at least.

Understand the topic

Aim: Detect delay item correctly

Worst prediction: False Negative (Recall)

Pay extra cost efficiency: False Positive (Precision)

Model benchmark: Recall

Actual Ontime	Okay	Rush ship-extra cost
Actual Delay	Intangible cost-Loss	Rush ship-extra cost
	Predict Ontime	Predict Delay

Dataset overview

Basic info:

10,999 pieces of data, no missing value

10 features

Warehouse block, Shipment mode, Customer care calls, Customer rating, Cost of Product, Prior purchases, Product importance,Gender, Discount offered, Weight(gms)

Target variable:

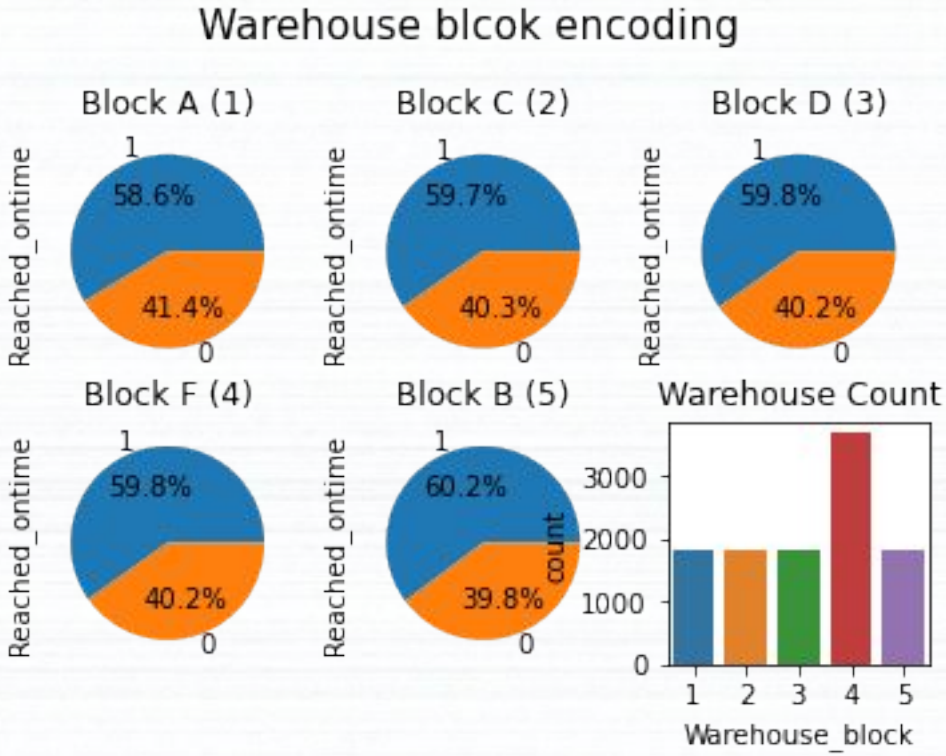
Reached ontime

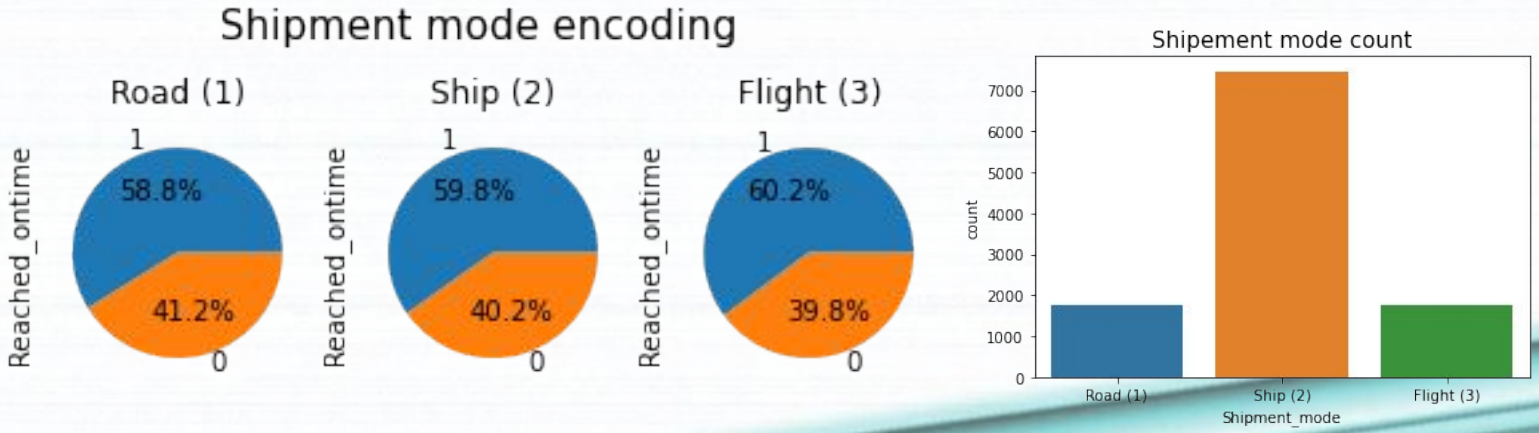
0-Ontime

1-Delay

Prepare data-feature encoding

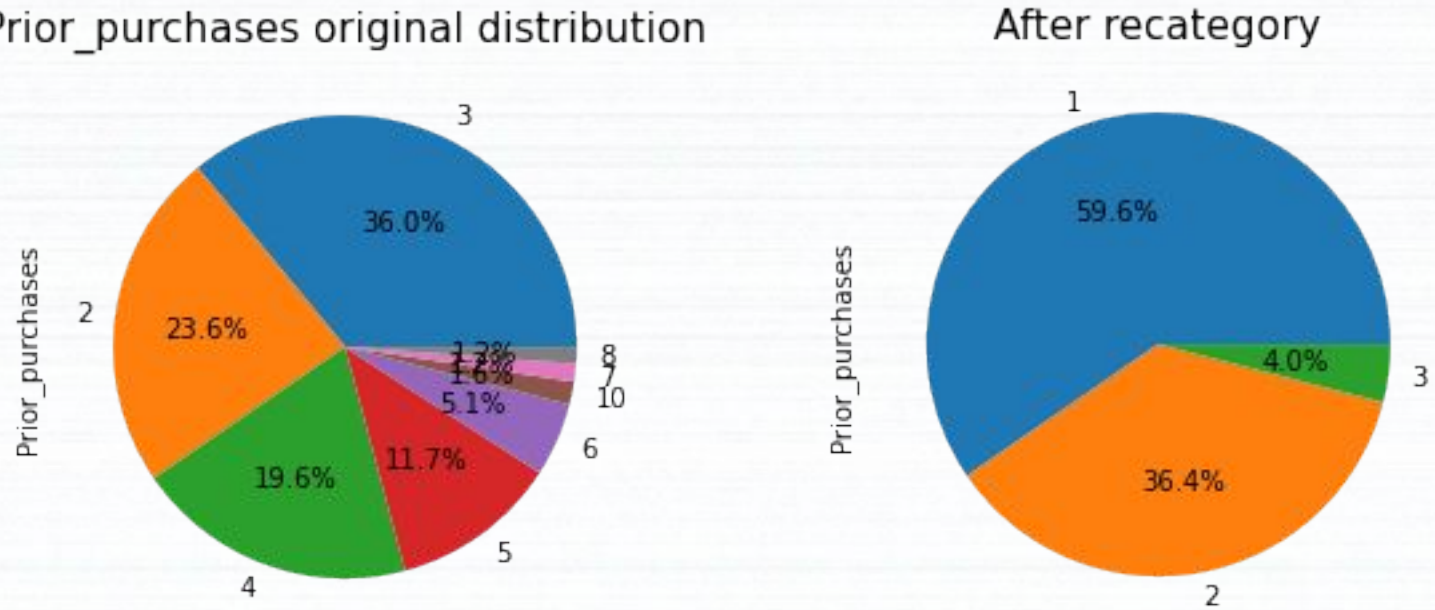
- Warehouse block
- Shipment mode
- Gender
- Product importance)





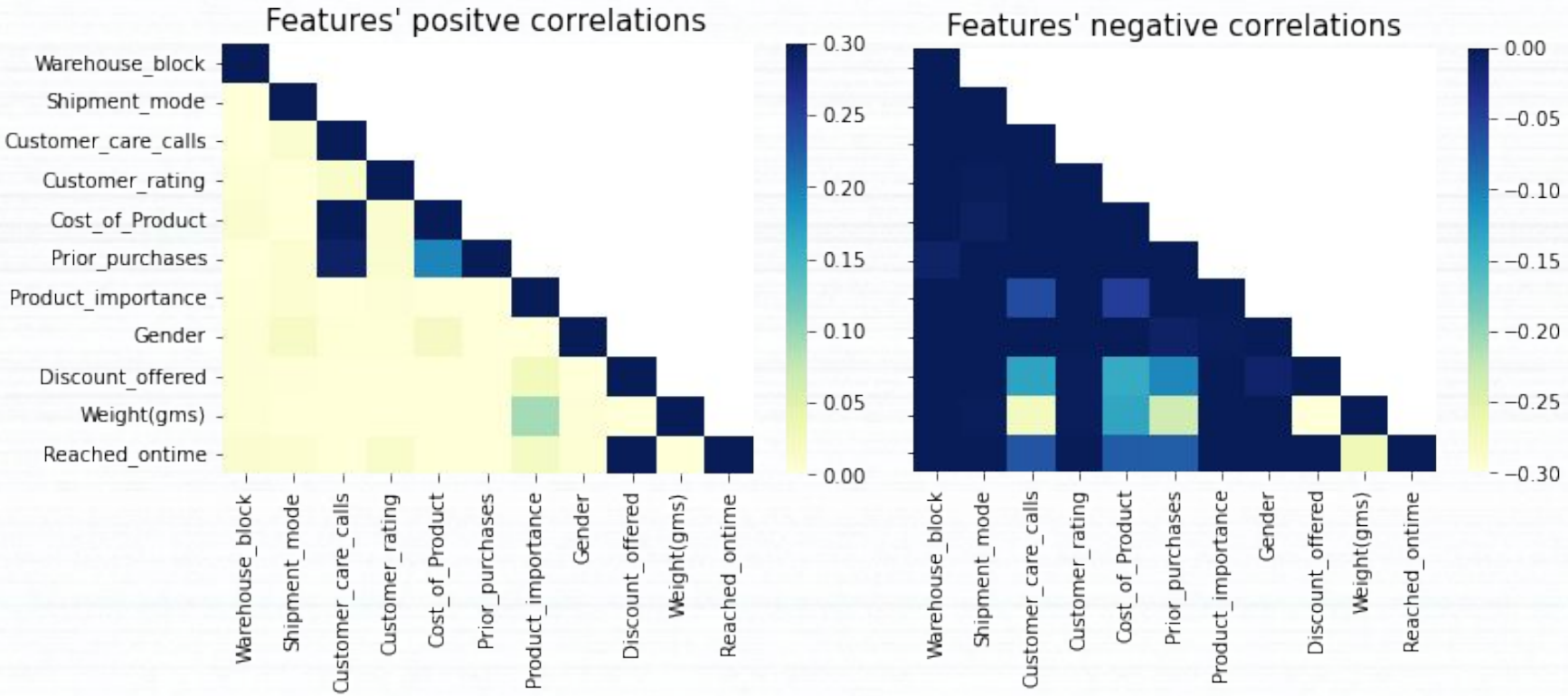
Prepare data-Recategory feature

Prior purchases recatogory by obsevation



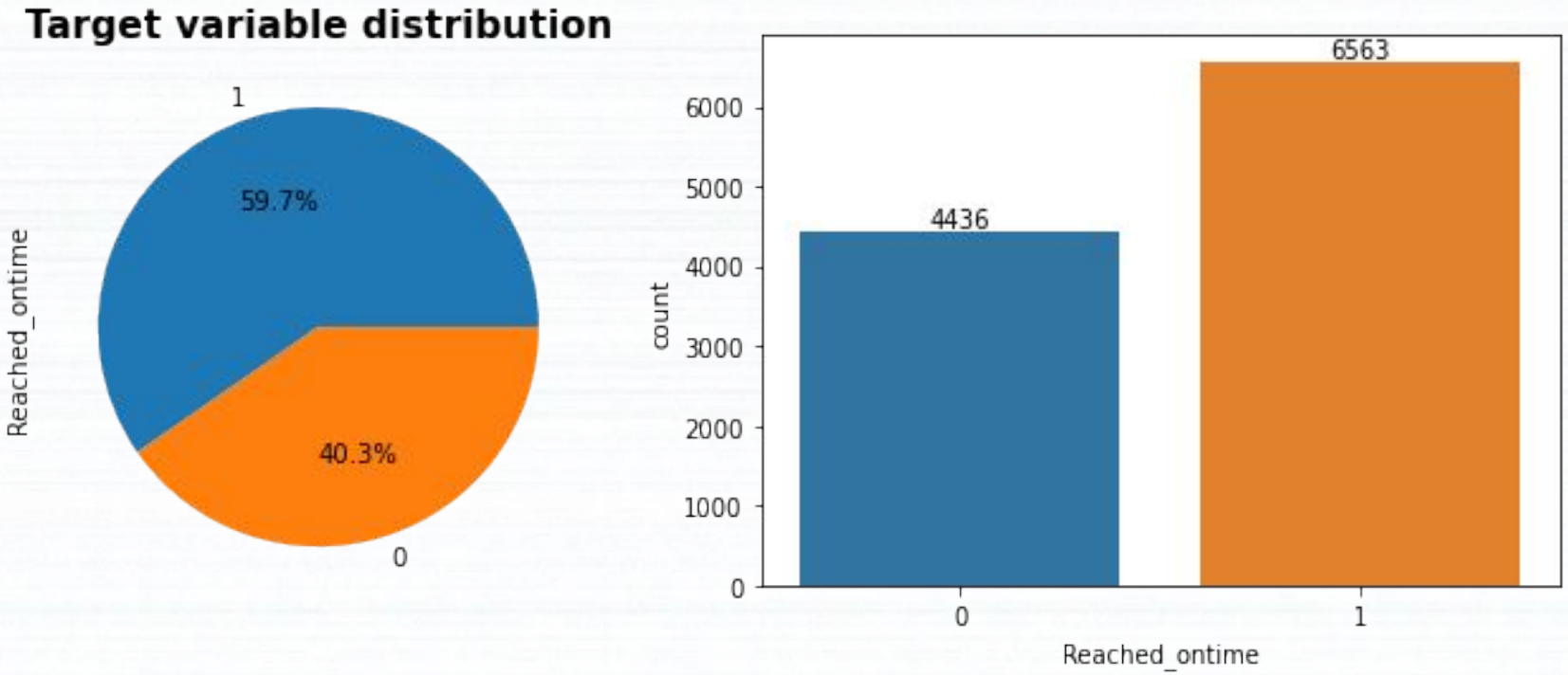
Visulaize features' correlation

Not strong correations(from -0.4 to 0.4) among features

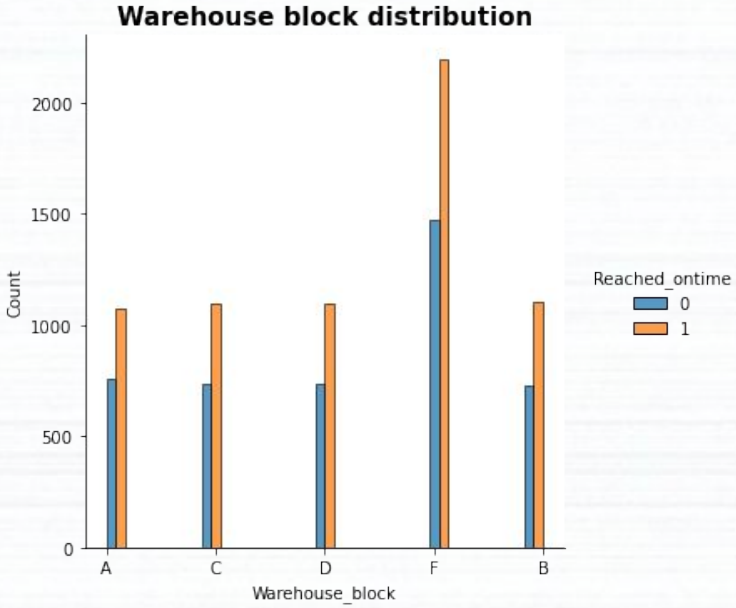


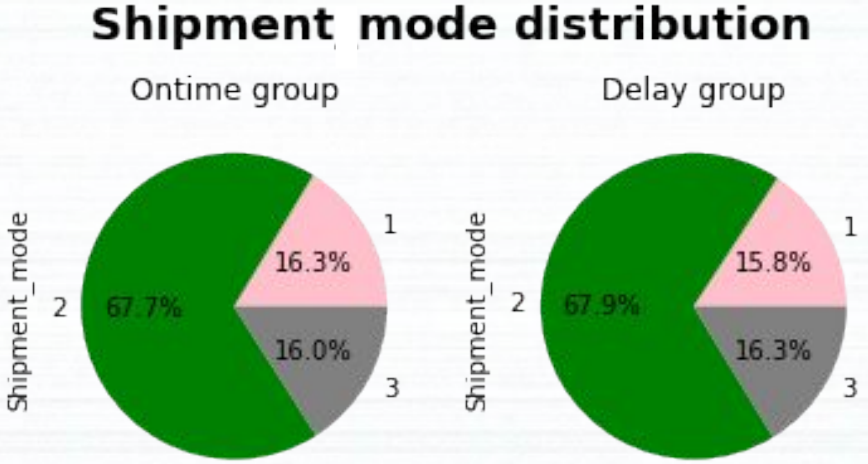
EDA

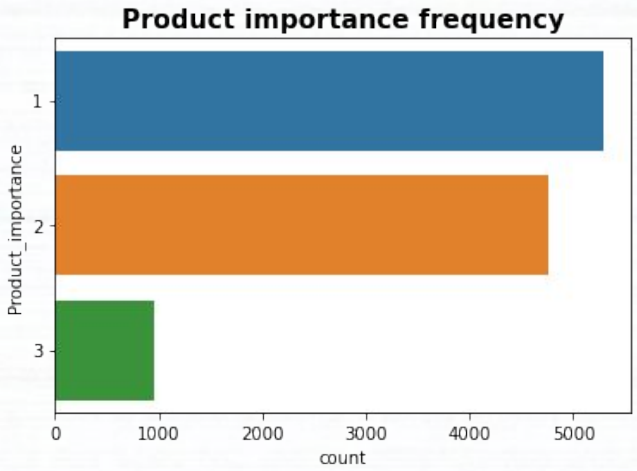
Target variable is balanced

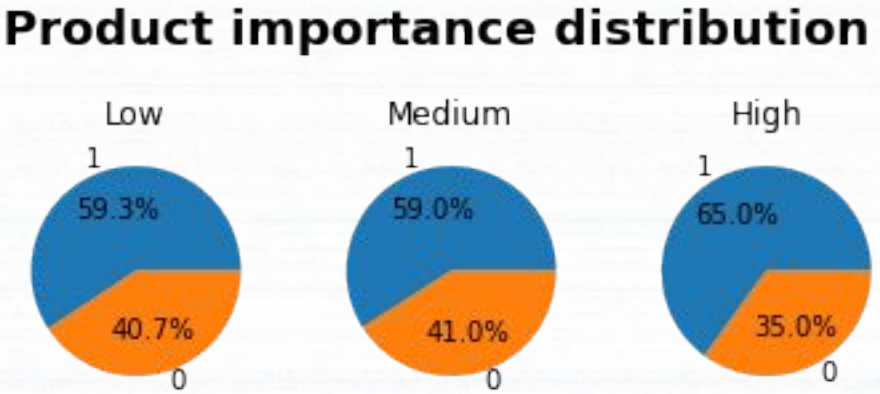


Feature variables





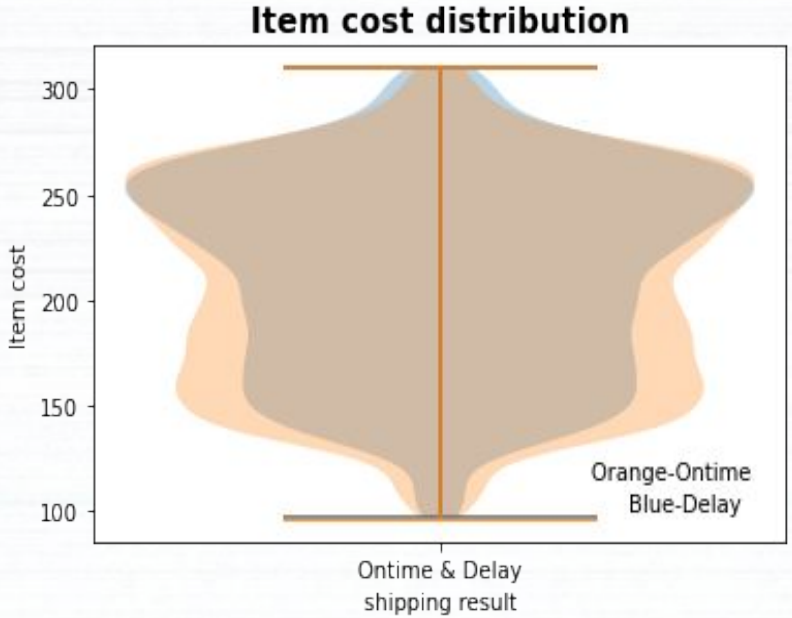


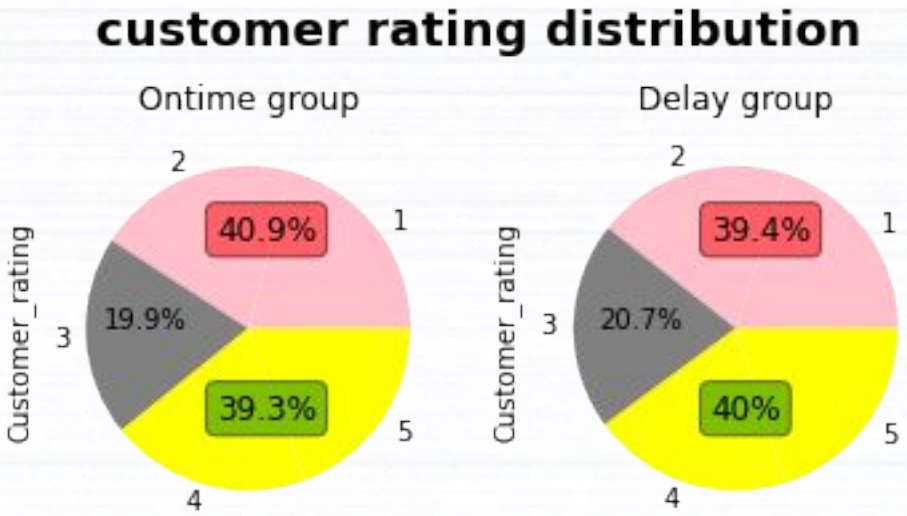


Feature variables

When cost is around $150,
Higher posibility to be ontime

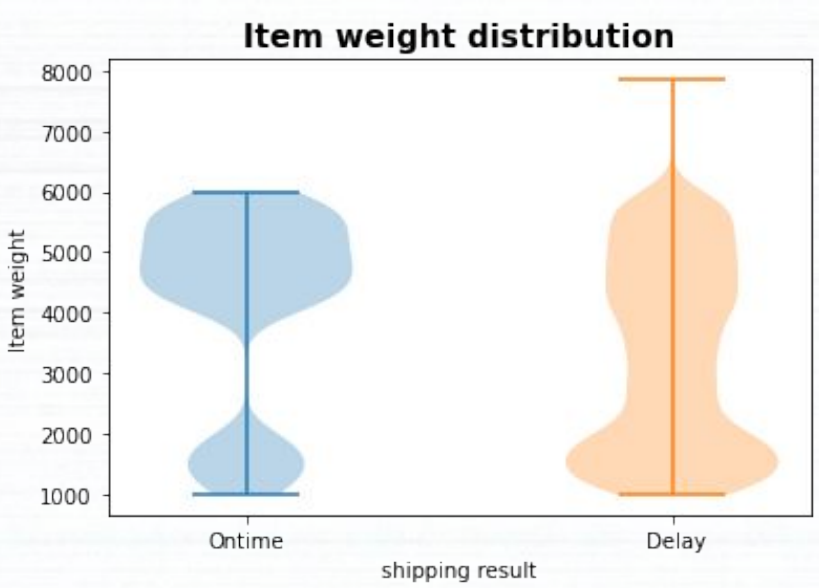
Ontime items has higher chance to be low customer rating

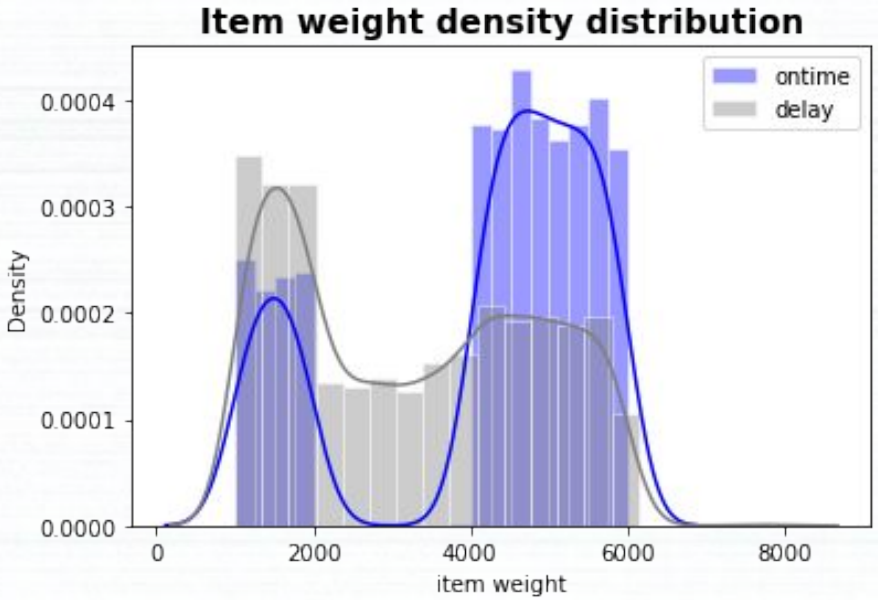




Feature variables

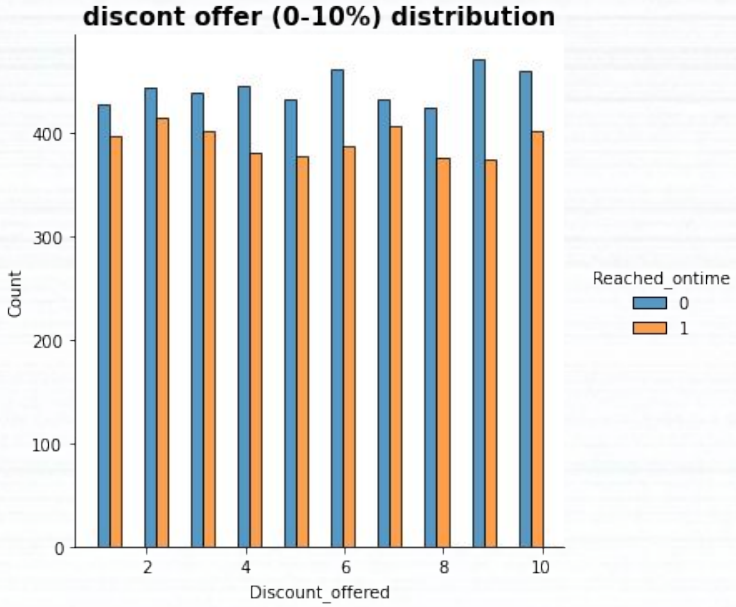
When item weight is 1-2 kg, higher chance ship delay
When item weight is 4-6 kg, higher chance ship ontime





Feature variables

When item discount offer >10, all shipping delay
When item discount offer <=10, more items ship ontime



Feture Engineer in Model

- Reduce features to 7 by RFE feature selection
- Supported evidence from PCA

Principal Component Analysis			
Selected features	Component explained ratio		
	First	Second	Third
10	0.175	0.134	0.102
7	**0.250**	**0.192**	**0.144**

Selected 7 feautres are: Customer care calls, Customer rating, Product Cost, Prior purchases, Product importance, Discount offered, Weight(gms)

Exploring model process

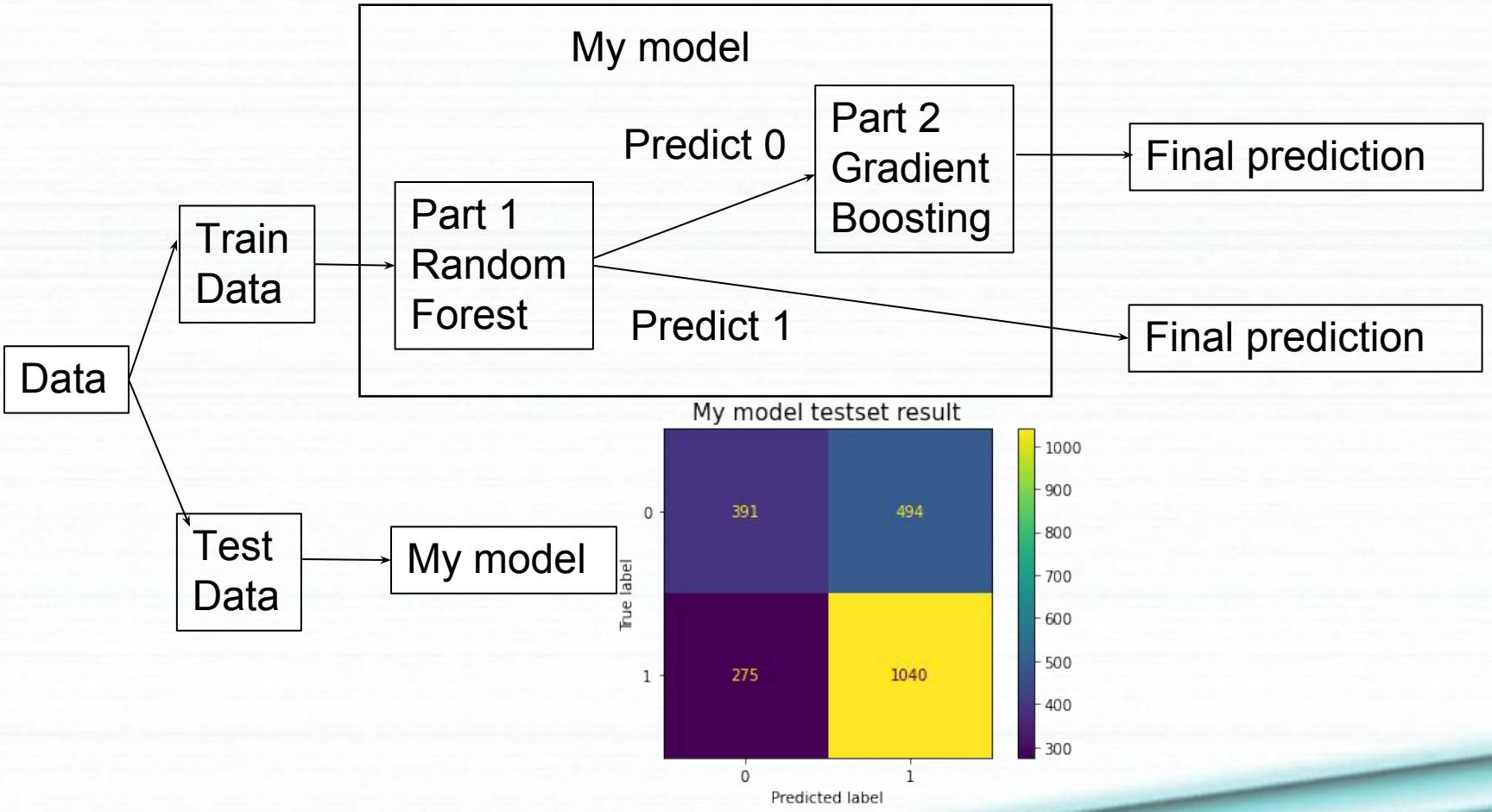
Model	Recall	Precision	Accuracy
Logistic Regression	0.67	0.70	0.63
SVC	0.62	0.78	0.67
Decision Tree	0.65	0.77	0.68
Random Forest	0.65	0.78	0.68
Ada Boosting	0.65	0.77	0.68
Gradient Boosting	0.72	0.71	0.66
GaussianNB	0.43	0.98	0.65
KNN	0.69	0.70	0.64

First mind from results:
models have low recall, high precision, similar accuracy in general

Build my model

Part 1: Random Forest

Part 2: Gradient Boosting



Visulize shipping delay cost

Assume:
Rush shipping is $1 per ship to make it from delay to on time
Delay shipping has intangible cost(assume $10 and $20)

Actual 0 Ontime	Okay	Rush ship-extra cost $1
Actual 1 Delay	Intangible cost-Loss	Rush ship-extra cost $1
	Predict 0 Ontime	Predict 1 Delay

Evaluation- Cost on testset

Model	Recall	Precision	Accuracy	Total Cost $10	Total Cost $20
Logistic Regression	0.67	0.70	0.63	5,624	9,994
SVC	0.62	0.78	0.67	6,034	11,014
Decision Tree	0.65	0.95	0.68	5,669	10,299
Random Forest	0.65	0.78	0.68	5,665	10,225
Ada Boosting	0.65	0.77	0.68	5,704	10,304
GradientBoosting	0.72	0.71	0.66	5,031	8,741
GaussianNB	0.43	0.98	0.65	8,088	15,598
KNN	0.69	0.70	0.64	5,401	9,511
My model	0.79	0.68	0.65	4,284	7,034
Stacking (part1&2)	0.73	0.73	0.67	4,880	8,440
No model	NA	NA	NA	13,150	26,300

Find intangible loss breakeven

No model VS My model

1315* B >275 *B +(494+1040)*1

B> 1.475

Conclusion:

1. When rush shipping cost $1/order, if set intangible cost more than $1.475/order, use my model will get less total cost. Also, rush shipping: Delay cost ratio (1:1.475) suits all conditions.
2. When set delay shipping intangible cost higher, my model is more benifical to company.

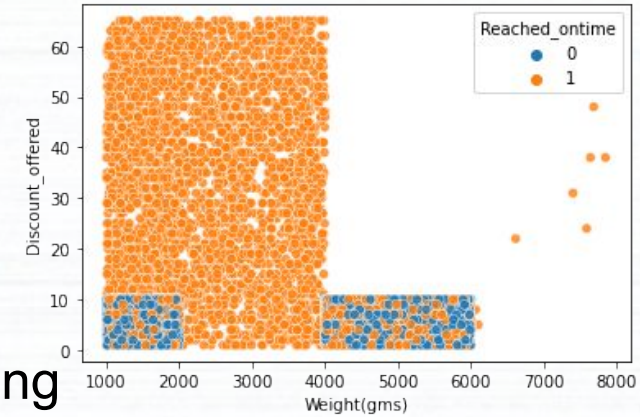
Further development

- Data source bugs:

Prior purchase (=0, 1)data missing

Deliver ontime item weight (2kg-4kg) missing

Deliver delay item discount offered >10% missing



- Shipping delay intangible cost is not given, difficult to show more details

Thanks

Data source:
https://www.kaggle.com/datasets/prachi13/customer-analytics