

# E-Commerce Shipment On-Time Delivery



# Agenda

**01**

**Business Problem**

**02**

**Exploratory Data Analysis**

**03**

**Predictive Modeling**

**04**

**Conclusions**

**05**

**Recommendation**

**06**

**Next Steps**



# 01

## Business Problem

The company has experienced decline in customer satisfaction at one of its major distribution centers due to excessive late deliveries.

We are charged with:

- Analyze the shipping data to find root cause of lateness.
- Create a model to predict if a shipment will be late, so that the company can preemptively take corrective action.



# 02

## Data Analysis

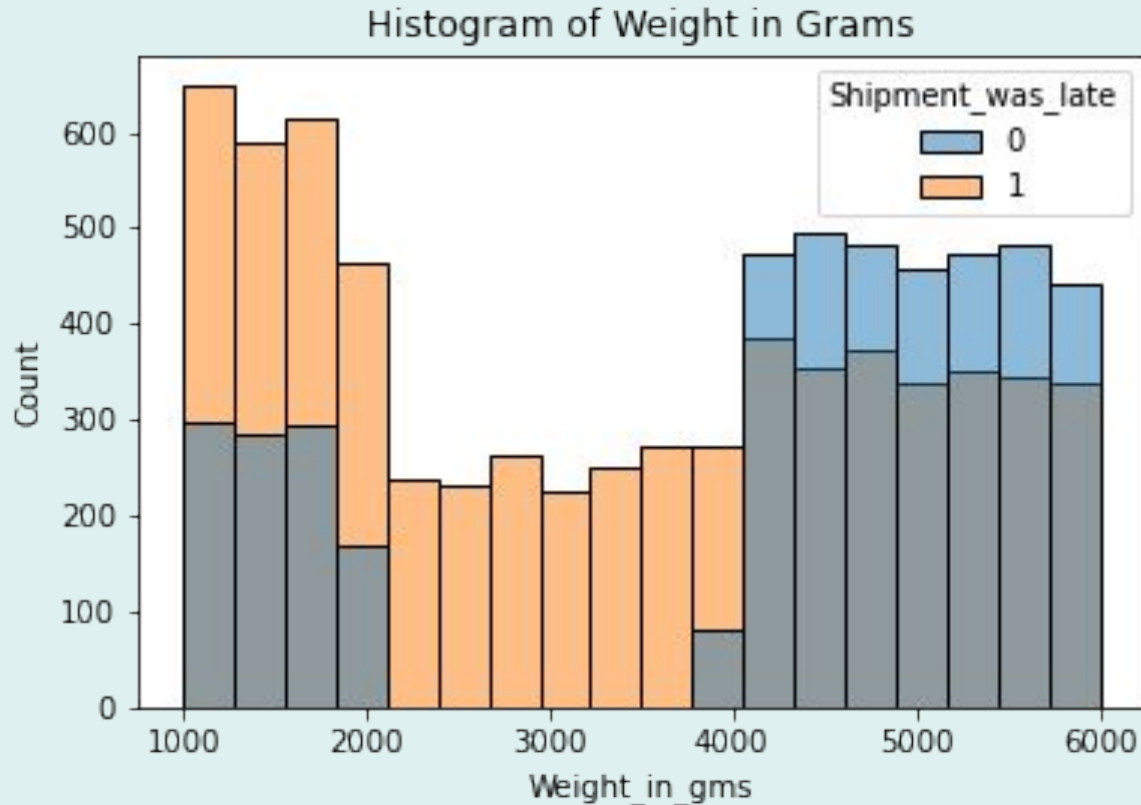
The dataset used in this analysis contained records of 11,000 shipments, including details regarding:

- Warehouse block
- Mode of shipment
- Customer care calls
- Customer rating
- Cost of the product
- Prior purchases
- Product importance
- Gender
- Discount offered
- Weight
- Was the shipment late?

Source: [E-Commerce Shipping Data \(kaggle.com\)](https://www.kaggle.com/datasets/ashishpatel26/e-commerce-shipping-data)



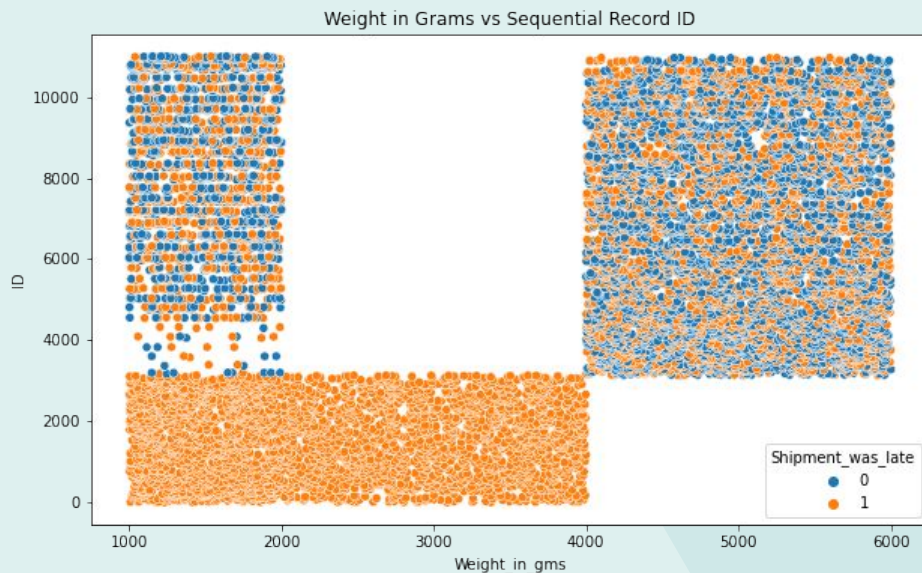
# Problems with the Data: Shipment Weight



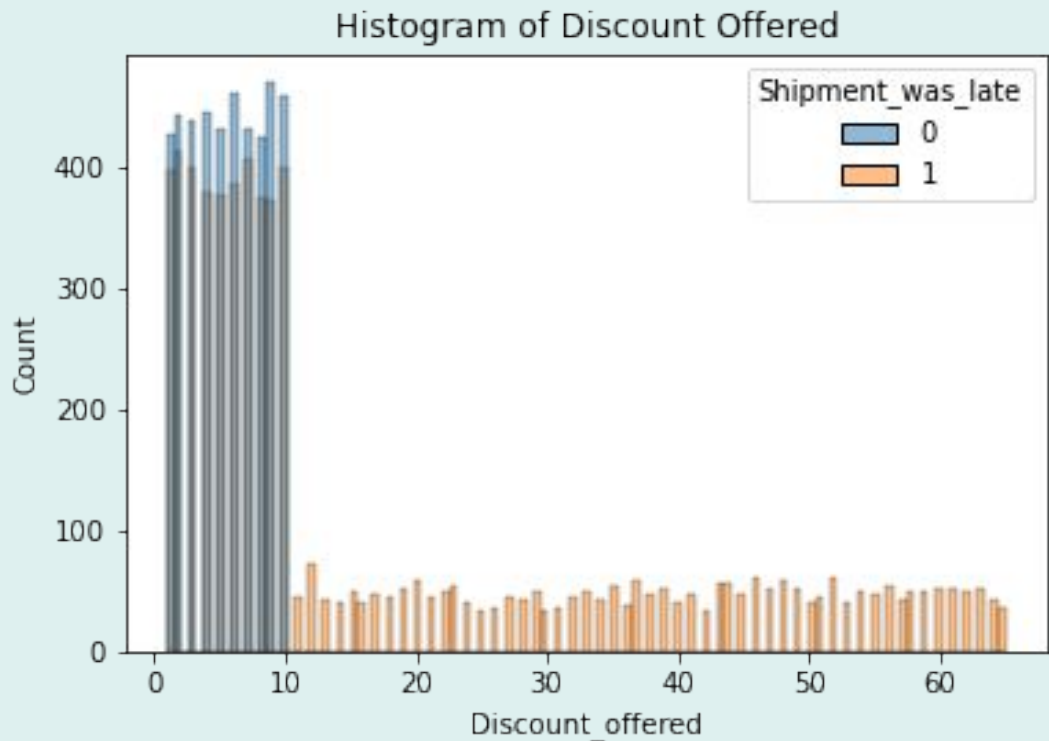
# Problems with the Data: Shipment Weight



- Strange segmentation in scatterplots
- Further investigation is required



# Problems with the Data: Discount Offered



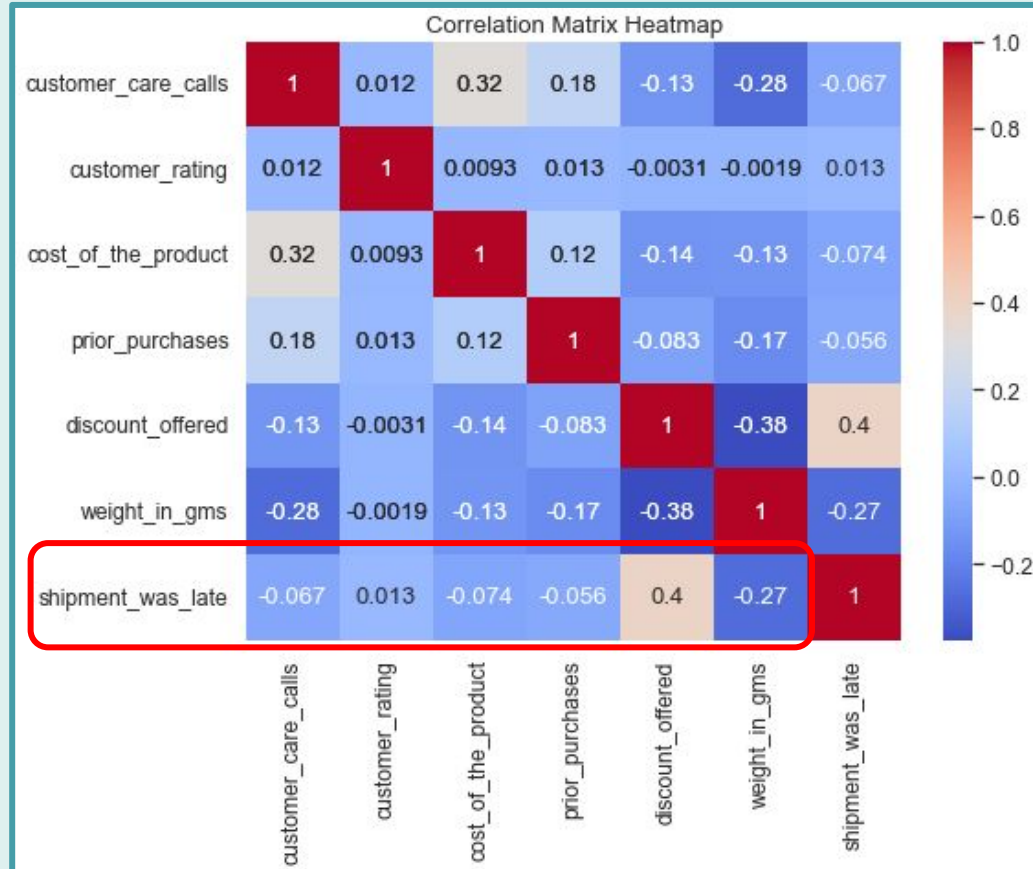
Leading Theory:

A discount of more than \$10 is only offered if a shipment is late.

If true, we cannot use Discount offered as a predictor of whether a shipment will be late!

further investigation is required.

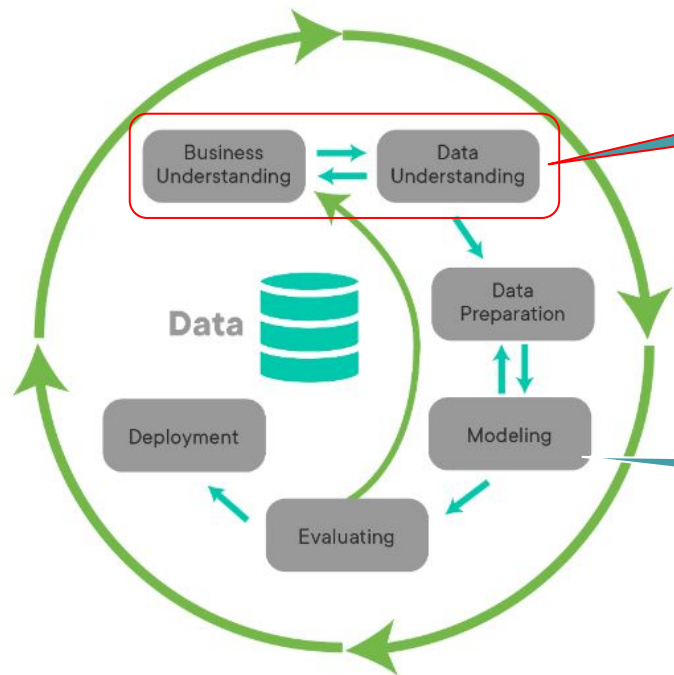
# Problems with the Data: Little Correlation





# Where We Are

## Cross-Industry Standard Process for Data Mining (CRISP-DM)



Due to unanswered questions about the data, we are still here

However, we can offer preliminary stop-gap predictive models

# 03

## Predictive Modeling

Business Goal: Predict if a shipment will be late

Possible prediction outcomes:

- We correctly predict a shipment will be late (True Positive)
- We correctly predict a shipment will be on-time (True Negative)
- We incorrectly predict a shipment will be late (False Positive)
- We incorrectly predict a shipment will be on-time (False Negative)

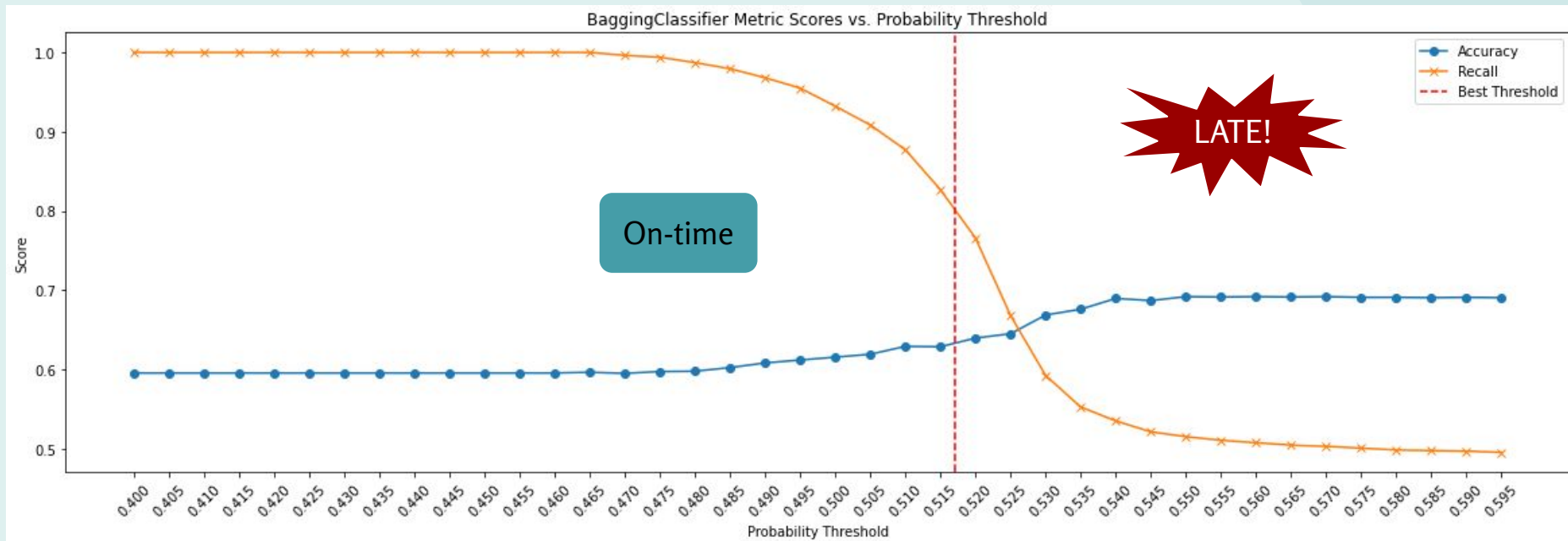
**BAD!!!**

Optimize for Recall Score



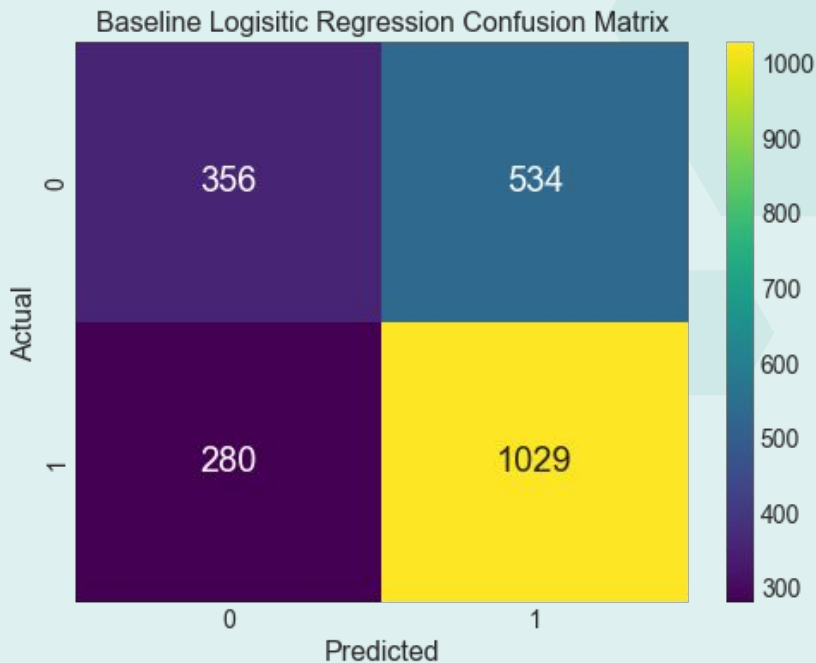
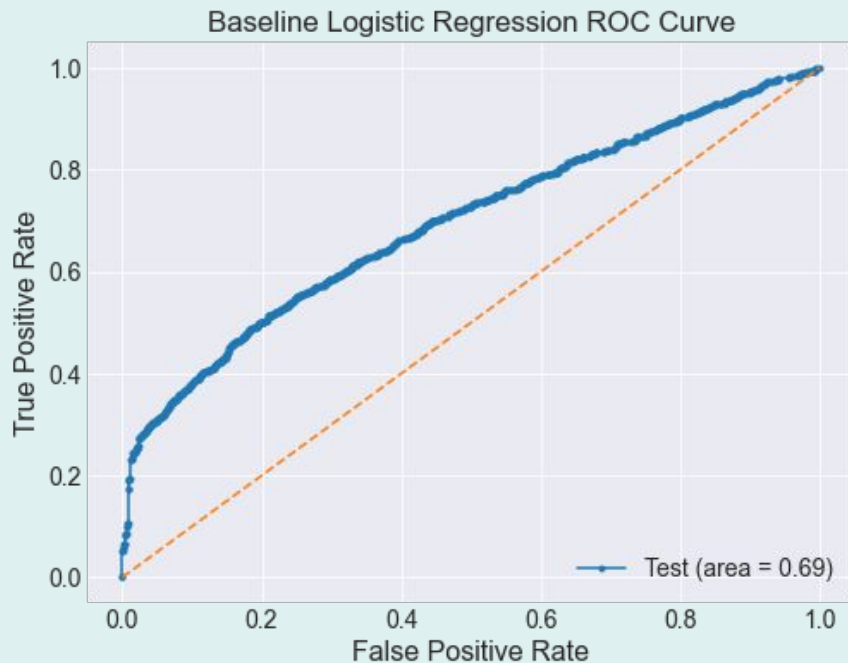
# Modeling Process Methodology

- Feature Selection, outlier removal, create baseline model
- Key hyperparameter grid search with 5-fold cross-validation (optimize Recall)
- Probability threshold tuning of the final models (Maximize Recall, holding Accuracy @ ~63%)



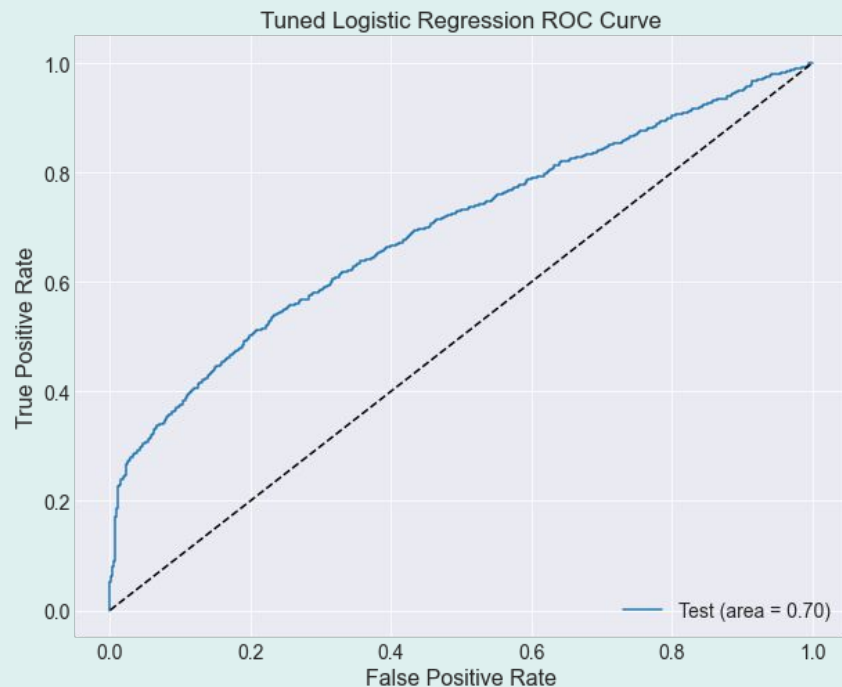
# Logistic Regression Baseline Model

Precision	F1	Accuracy	Recall
0.63	0.72	0.63	0.79



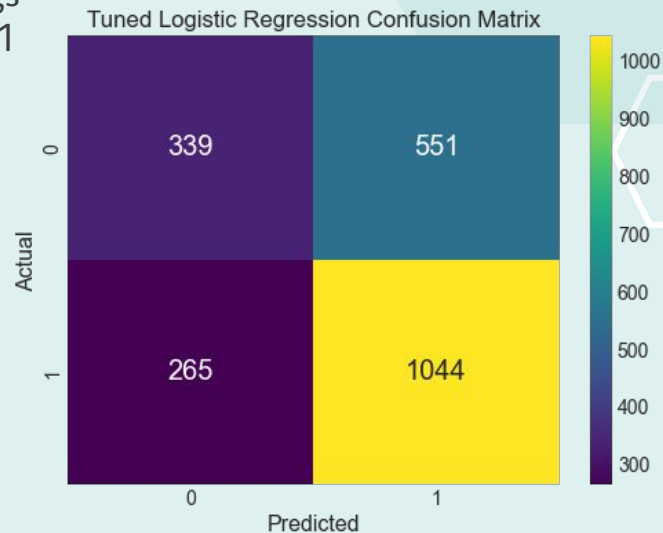
# Logistic Regression Hyperparameter Tuned

Precision	F1	Accuracy	Recall
0.65	0.72	0.63	0.80



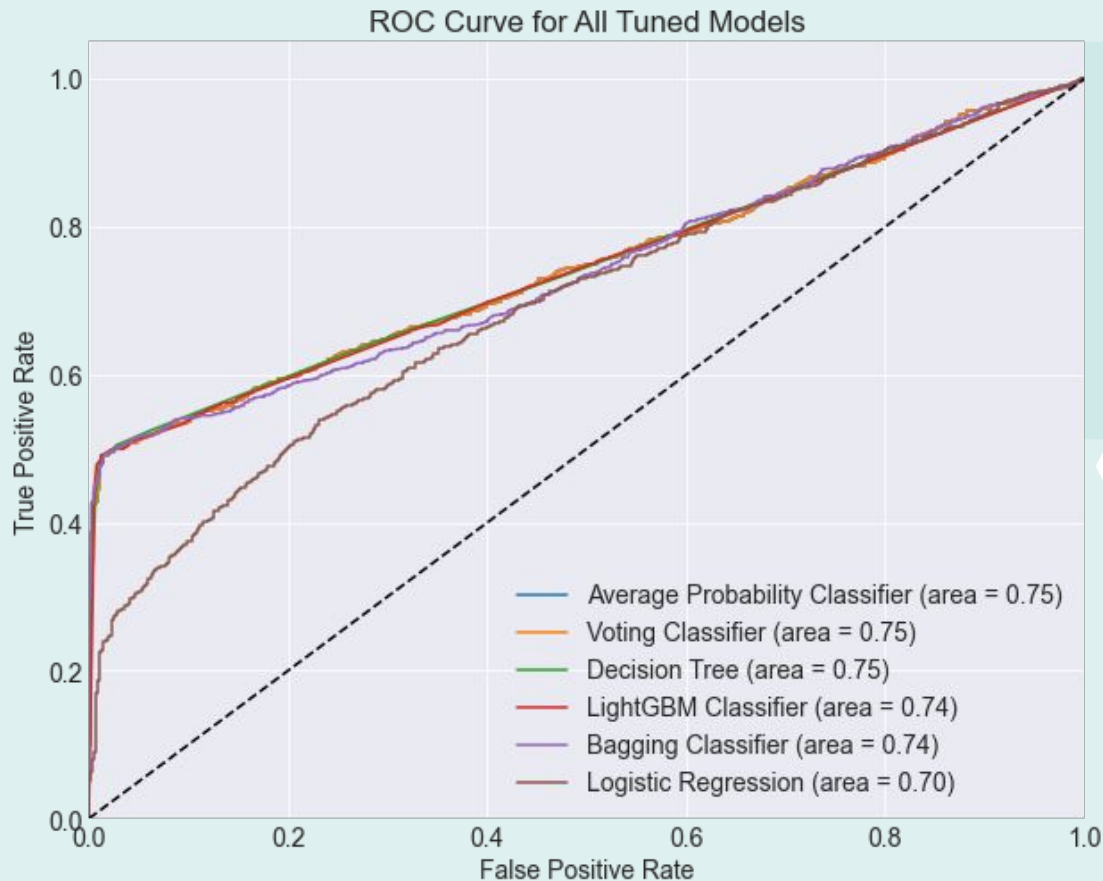
## Best Hyperparameters:

- GridSearch: (CV: 5, Scoring: Recall)
- C: 0.01
- Class\_weight: None
- Max\_iter: 100
- Penalty: l2
- Solver: lbfgs
- Tol: 0.0001

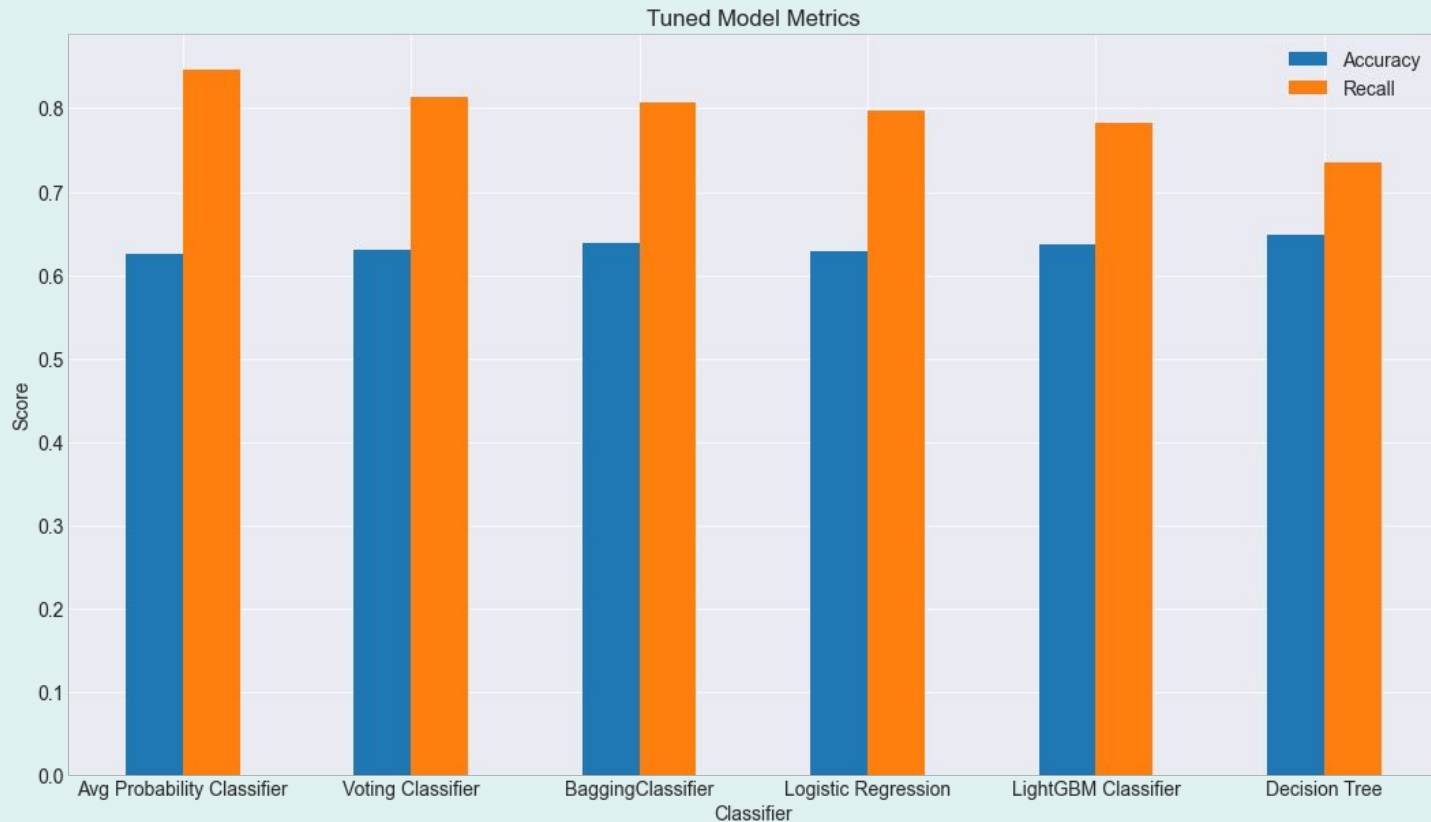


# Performance of All Models

Classifier Models
Average Probability Classifier
Voting Classifier
Bagging Classifier
Logistic Regression
LightGBM Classifier
Decision Tree



# All Tuned Models Final Scoring



Model Selected: Average Probability Classifier

# 04

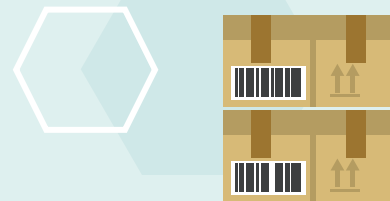
## Conclusion

Best Model: Average Probability Classifier

- Highest Primary Metric: Recall
- Accuracy was not the highest, but comparable to other models
- Combines predictions from multiple models

Final Metrics:

Precision	F1	Accuracy	Recall
0.64	0.73	0.63	0.85





# 05

## Recommendation



1. Back to square 1: Business Understanding <-> Data Understanding
2. Improve data acquisition methods to increase performance of future models
3. Implement the ***Average Probability Classification*** model as a stop-gap to avoid late shipments until items 1 and 2 can be addressed

Final Metrics:

Precision	F1	Accuracy	Recall
0.64	0.73	0.63	0.85

# 06

## Next Steps



1. Prioritize focus on **Business/Data Understanding**; answer:
  - Are discounts offered over \$10 due to prior knowledge lateness?
  - Root cause(s) for segmentation in shipment weight data
  - Root cause(s) for segmentation in ID... is the data time sequential?
2. Investigate improving data acquisition:
  - What data can we gather that will help us better predict lateness?
3. Deployment of ***Average Probability Classification***, predict late shipments:
  - Flag potential late shipments for expediting delivery
  - If lateness cannot be corrected, preemptively engage with customer to inform, offer discount, etc.





# Thanks!

Do you have any questions?

Dale DeFord  
[daledeford@gmail.com](mailto:daledeford@gmail.com)



Brian Woo  
[brianhwoo@gmail.com](mailto:brianhwoo@gmail.com)

