Backorder Prediction

Team Member: Sherrie Liu/ Robin Liu/ Minyao Xu/Phani Uppu

Out of Stock
© Getty Images

Backorder

An order that currently cannot be filled or shipped. Suppliers still allow retailer to order but will ship later when the stock is back to normal

CONTENT

Business Understanding

Data Preparation

Modeling

Evaluation

Deployment

1 Task and Why?

Task: Predict if an order would be a backorder

$11-$15 cost /backorder

According to one study from Fedex, on average, a backorder will cost supplier $11-$15, not even including some further impacts on business relationships and sales

More Time to React

No guarantee retailer will keep the order

Retailer can order in advance

Better Business Planning

Potential to adjust demand forecast and production plan for certain items to decrease chance of backorders

CONTENT

Business Understanding

Data Preparation

Modeling

Evaluation

Deployment

2 Data Description

Data Source

- Kaggle— Can You Predict Product Backorders?

Content

- Training data： 1,687,861 instances
- Test data： 242,076 instances
- Training data： historical data for 8 weeks prior to the week we are trying to predict.

Variables

- Target variable: Product actually went on backorder
- Features: 22 features

Inventory/Transit time/Actual sales over periods/Forecasted sales over period/Previous shipping performance/Risk Flag

2 Data Exploration—Challenges

Missing Values

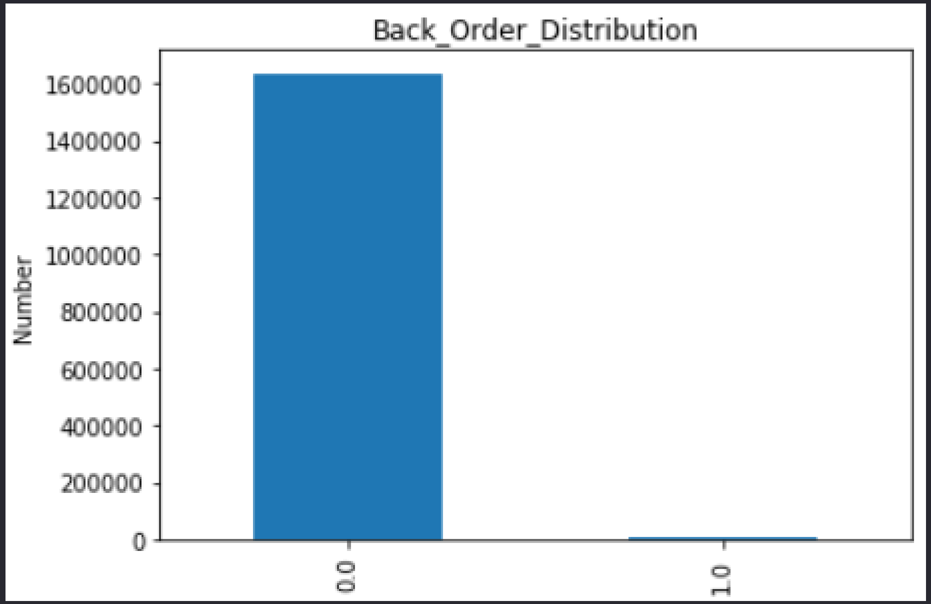
Column: lead_time
Transit time for product
Training set: 6.0%
Test set: 6.1%

Column: perf_6_month_avg
Performance for prior 6 months
Training set: 7.7%
Test set: 7.9%

Column: perf_12_month_avg
Performance for prior 12 months
Training set: 7.2%
Test set: 7.4%

Unbalanced Dataset

Target Variable
No—product did not go on backorder: 99.4%
Yes—product actually went on backorder: 0.6%

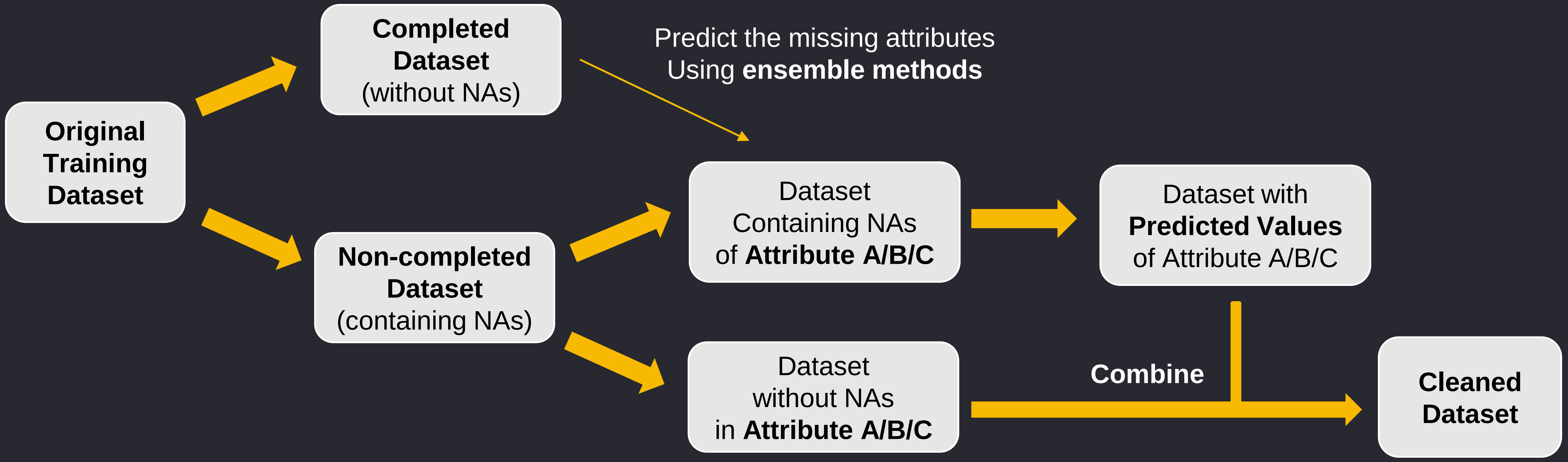


2 Data Exploration—Dealing with Missing Values

Missing Values

Solution: use other variables and regression/classification techniques to predict missing attributes

Reason: Not Missing Completely at Random (MCAR)



Original Training Dataset
Completed Dataset (without NAs)
Non-completed Dataset (containing NAs)
Predict the missing attributes Using ensemble methods
Dataset Containing NAs of Attribute A/B/C
Dataset without NAs in Attribute A/B/C
Dataset with Predicted Values of Attribute A/B/C
Combine
Cleaned Dataset

2 Data Exploration—Dealing with Unbalanced Dataset

Unbalanced Dataset

Problem

- Models tend to predict the majority class all the time, causing meaningless high accuracy.
- When sampling instances from the training set, there’s a good chance no minority class will exist in the sample at all.

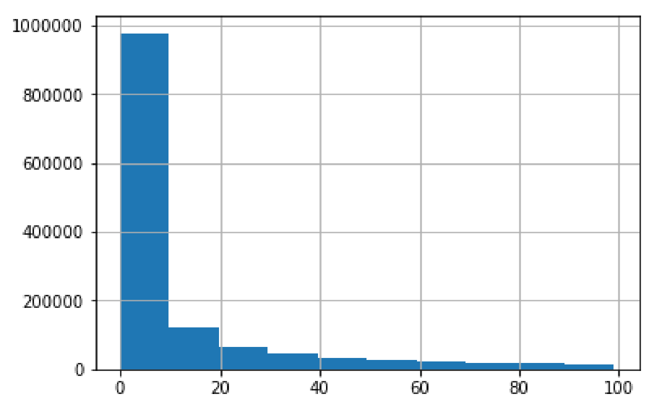
Solution

- Oversampling—Repeat sampling instances in the minority class with replacement.

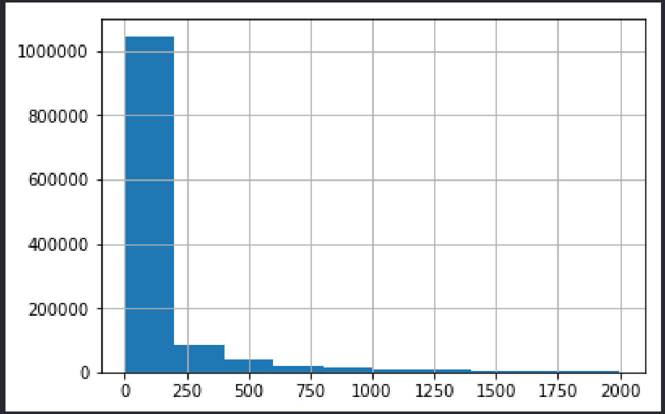
Result

- Adaboost algorithm and neural network work better when trained on a balanced set.
- Recalls are doubled, from 0.04 to 0.1.
- Auc is significantly increased, from 0.8 to 0.92 for adaboost and 0.6 to 0.91 for neural network.

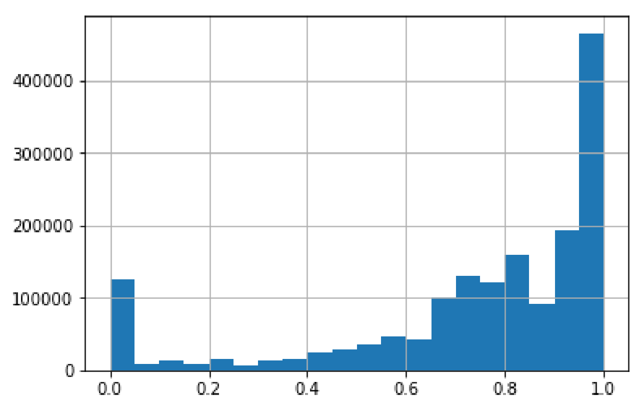
2 Data Exploration—Features Distribution



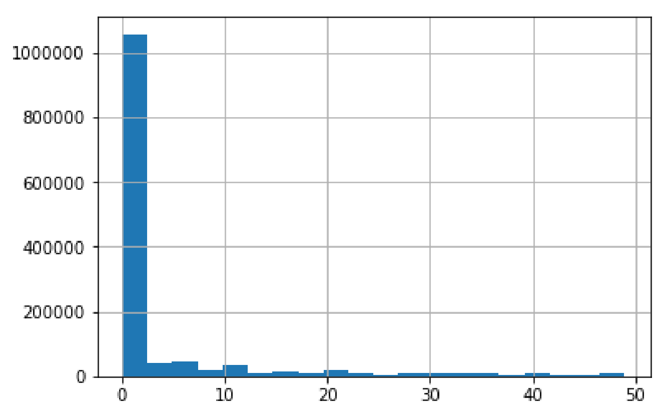
Previous 9 Months Sales
(80% below 100 units)



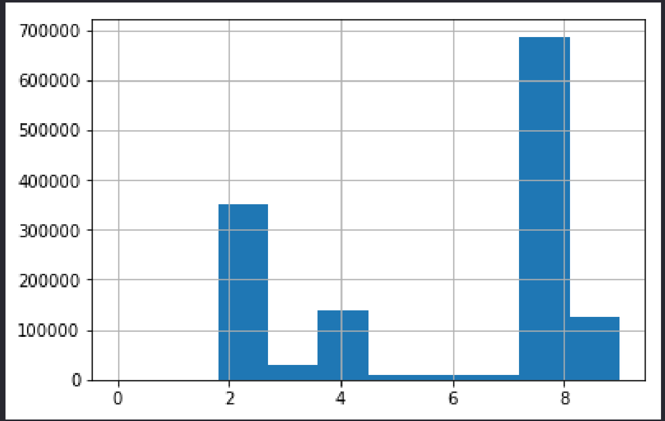
Current National Inventory
(75% are below 2000 units)



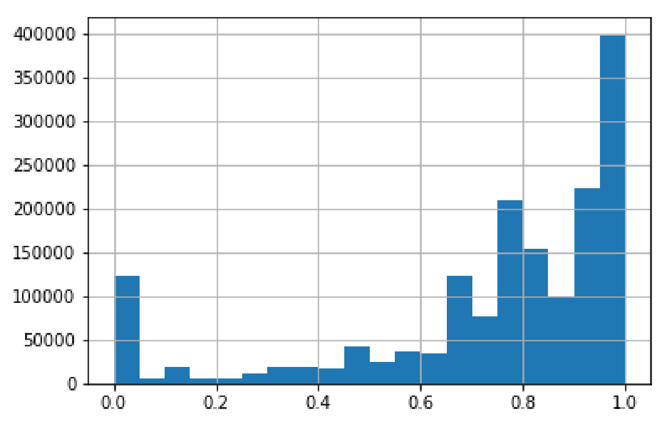
Previous 6 Months Shipping Performance Distribution



Next 9 Months Sales Forecast
(80% below 50 units)



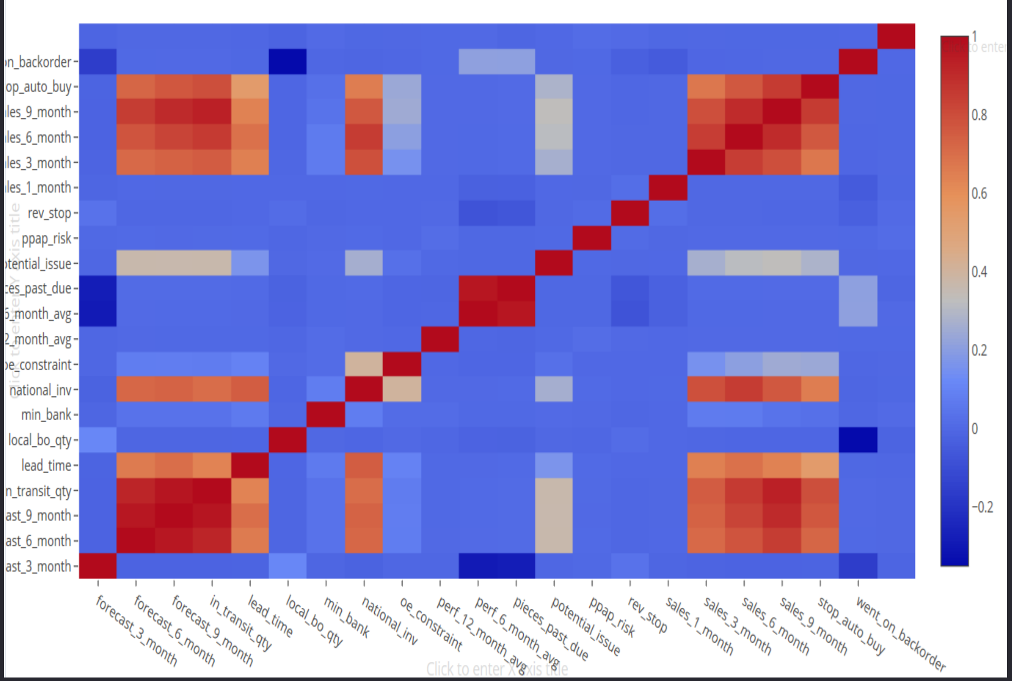
Current Lead Time
(83% are below 10 weeks)



Previous 12 Months Shipping Performance Distribution

2 Data Exploration—Correlation

Correlation Among Variables



Target Variable & Features

- ‘Went on backorder’ is not highly correlated with any of the features.

Among Features

- Amount of quantity in transit from source/ transit time for product/ the forecast sales/ past sales are correlated.

Adaboost and neural networks can handle the problem!

CONTENT

Business Understanding

Data Preparation

Modeling

Evaluation

Deployment

3 Model Building

Split Data

Divide the training dataset into 7 subsets

Because the original dataset is too large and computational hard

Choose Model

Neural network
Adaboost
Random forest

Train Model

Train 3 models on 7 subsets respectively—21 models in total

Do grid search on each model to find the best parameters

Ensemble

Ensemble results of 21 models by calculating their average predicted probability or weighted average probability

CONTENT

Business Understanding

Data Preparation

Modeling

Evaluation

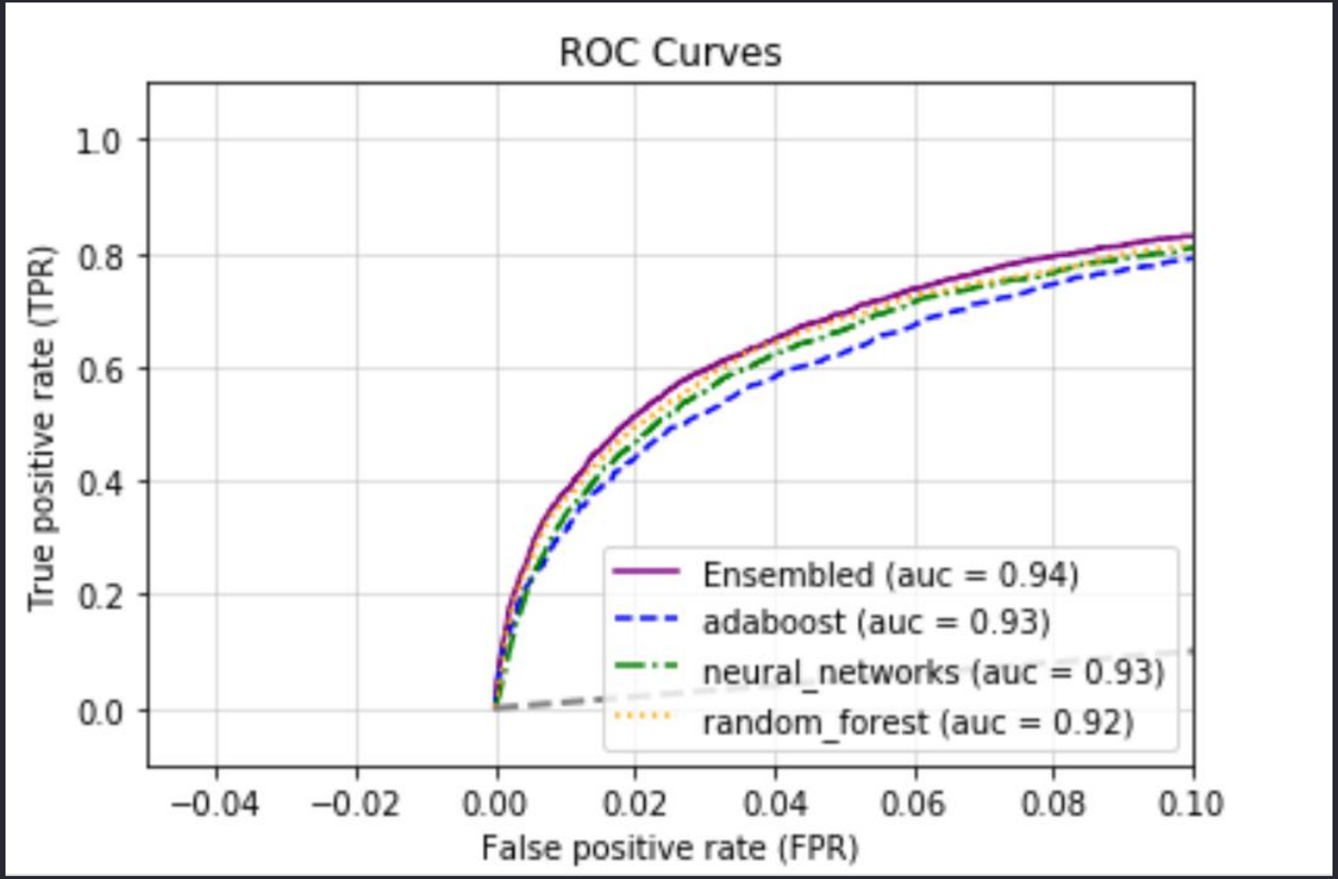
Deployment

4

Model Evaluation—Performance Metrics

	AUC	F1-Score	Recall	Precision	Accuracy
Ensembled	0.94	0.32	0.24	0.50	0.98
Adaboost	0.93	0.20	0.63	0.63	0.94
Neural Network	0.93	0.20	0.69	0.69	0.94
Random Forest	0.92	0.30	0.30	0.31	0.98

4 Model Evaluation—ROC curve



The current Best Performance on Kaggle is around 0.95

The auc of ensembled model is best—0.94

4 Model Evaluation—Cost and Benefits Analysis

Benefits

True Positive

Predicting correctly backorders

Benefits: better reaction time/better production planning /order in advance if necessary

True Negative

Predicting not in backorders correctly

Benefits: No extra costs

Cost

False Positive

Predicting backorders but in reality they are not

Costs: warehousing costs/potentially damaged products

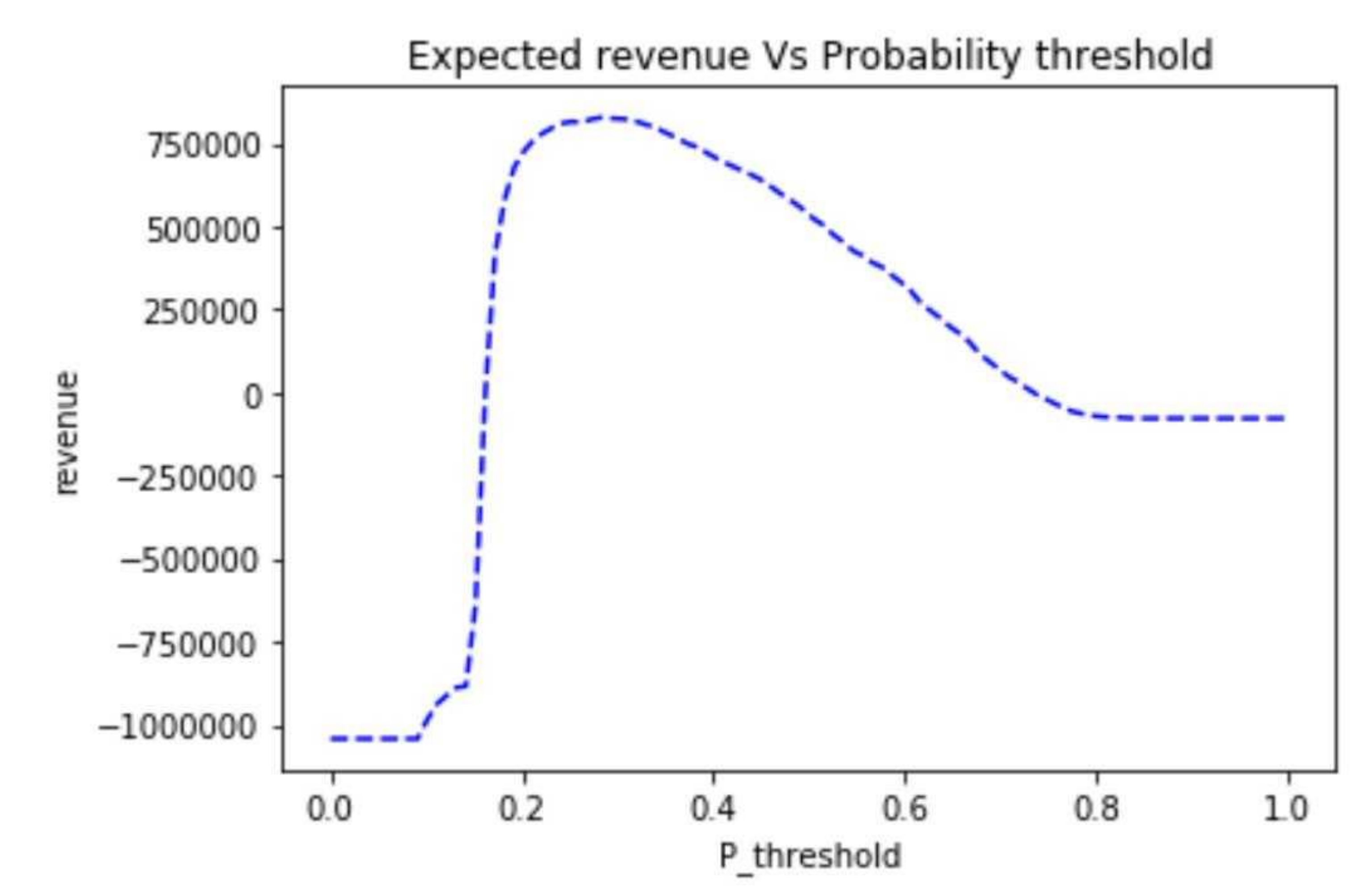
False Negative

Predicting NOT backorders but in reality they are

Costs: warehousing costs/packaging costs/potential sales loss

4

Model Evaluation—Cost and Benefits Analysis



CONTENT

Business Understanding

Data Preparation

Modeling

Evaluation

Deployment

5 Model Deployment

The model can be incorporated into the current ordering system of the supplier and retailer.

With new ordering/ inventory data feed in every day, the prediction of backorder updates daily.

Retailers and suppliers can pick models/classification threshold depending on the following criteria:

- type of retailer(aggressive/conservative)
- type of products(perishable/non-perishable)
- alignment between retailer and suppliers

The alignment sometime can be hard to achieve between retailers and suppliers as different parties have different business priorities and KPIs.

Thank you!
Any Questions?