Buy or Not?

Jason Salazer-Adams

The Challenge

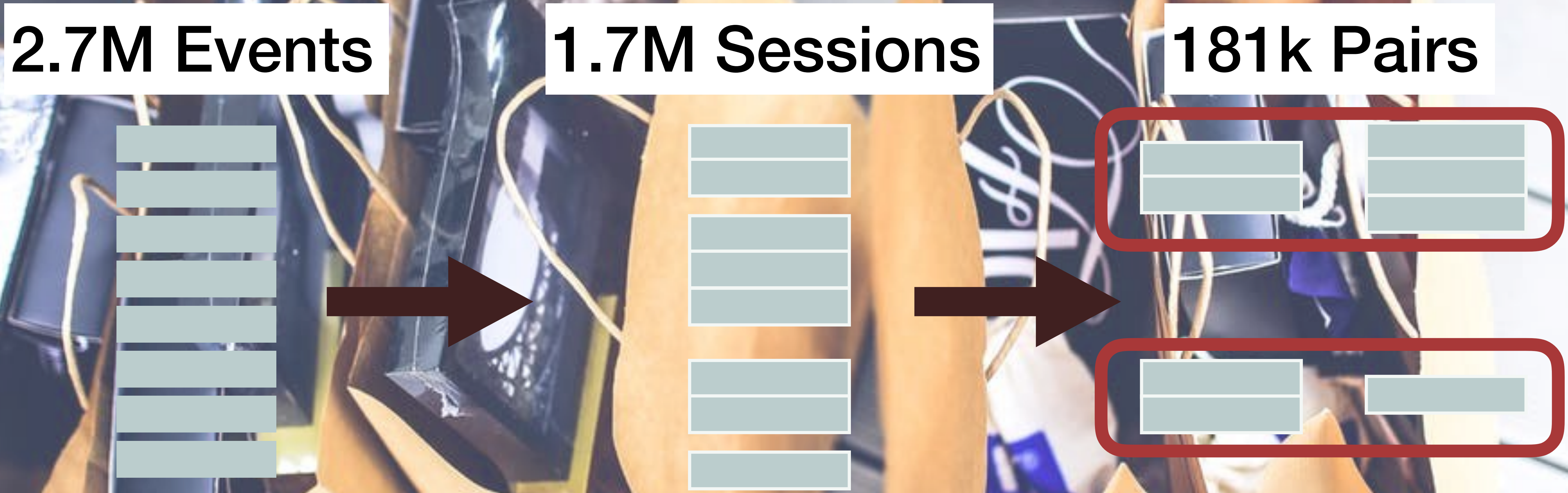
Why?

Targeted Marketing

Optimally allocate budget

Prevent profit erosion

Methodology

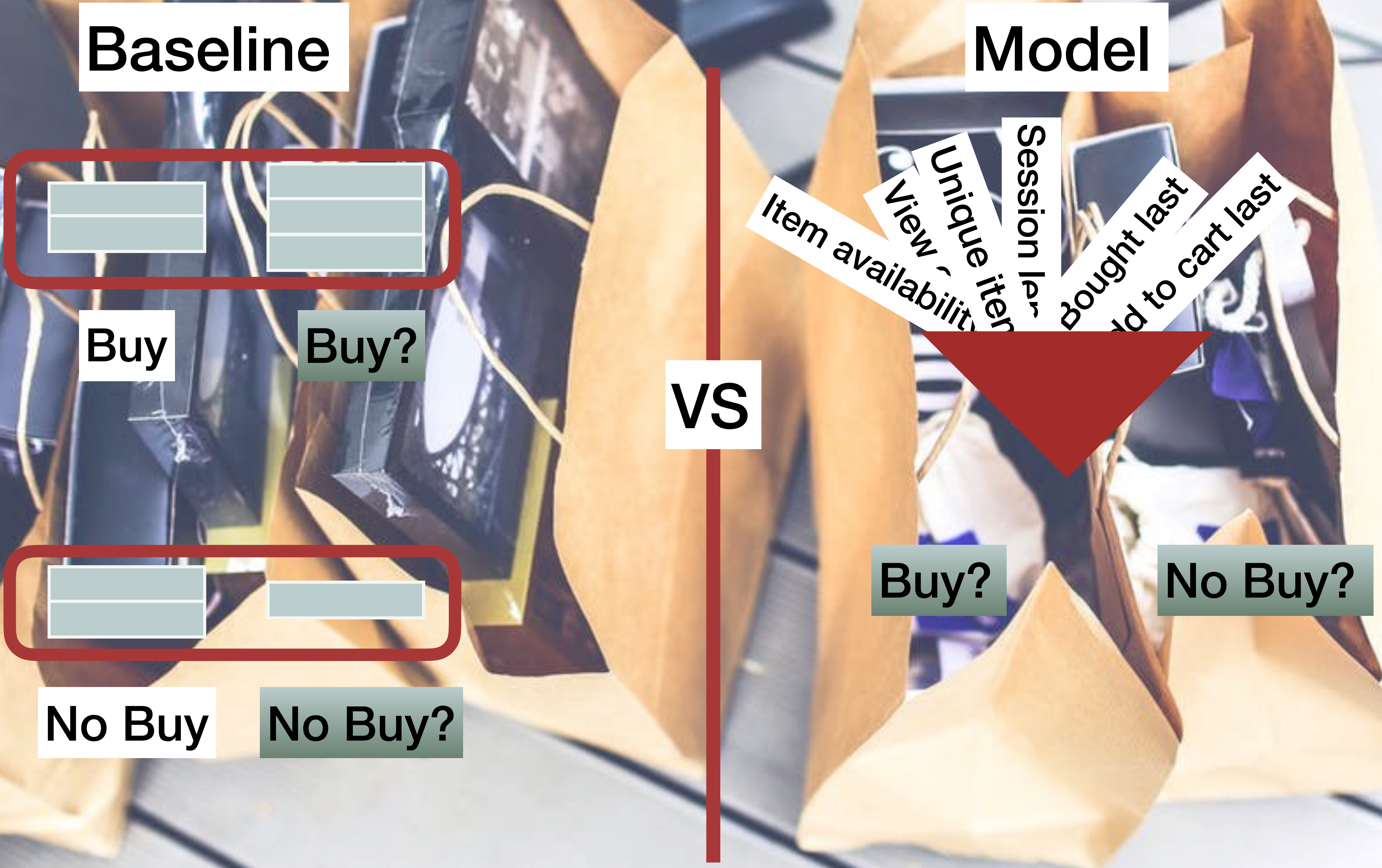


2.7M Events

1.7M Sessions

181k Pairs

Model Comparison



Baseline

Buy

Buy?

No Buy

No Buy?

VS

Model

Item availabilit

View c

Unique iten

Session le

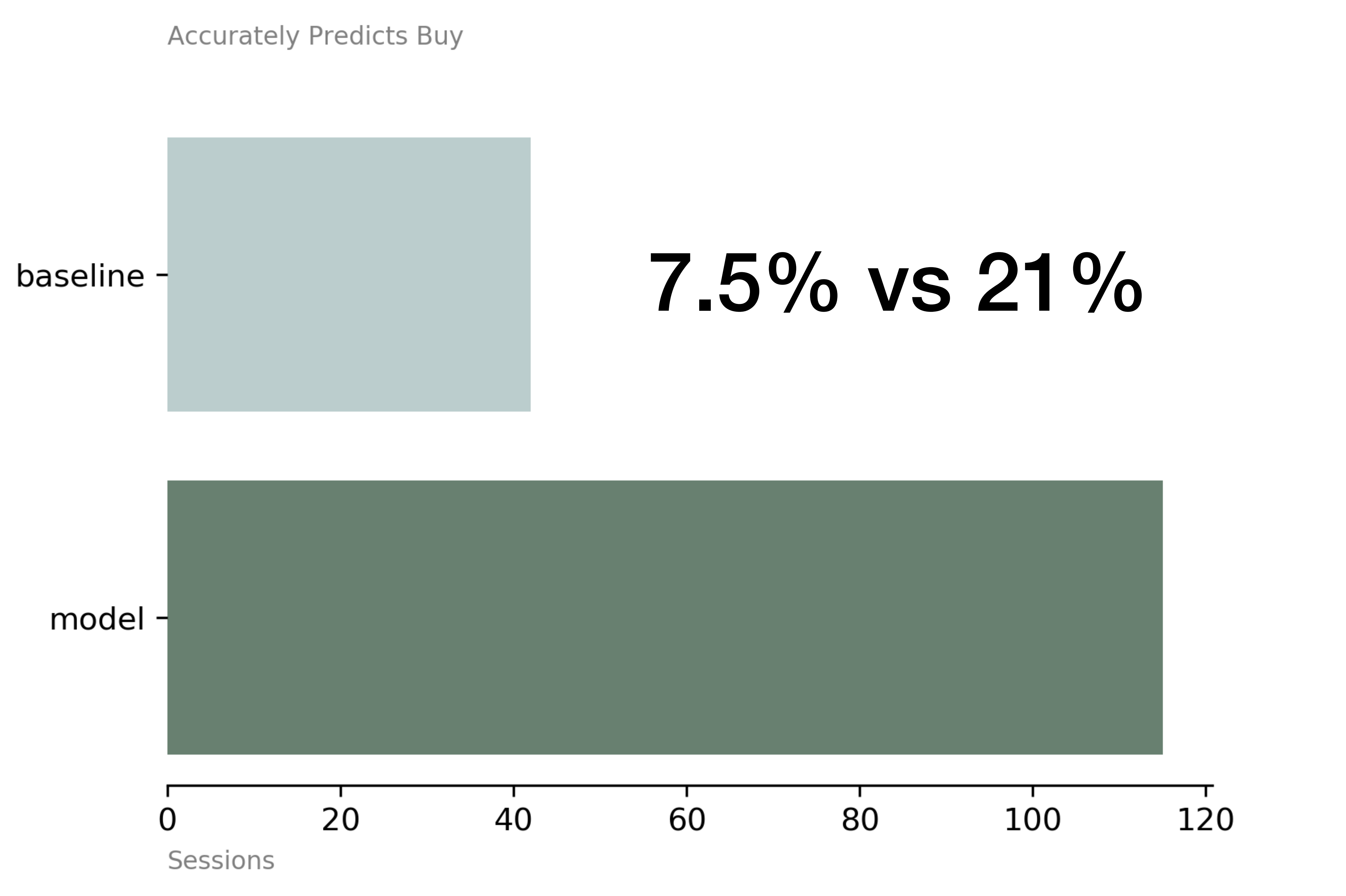
Bought last

dd to cart last

Buy?

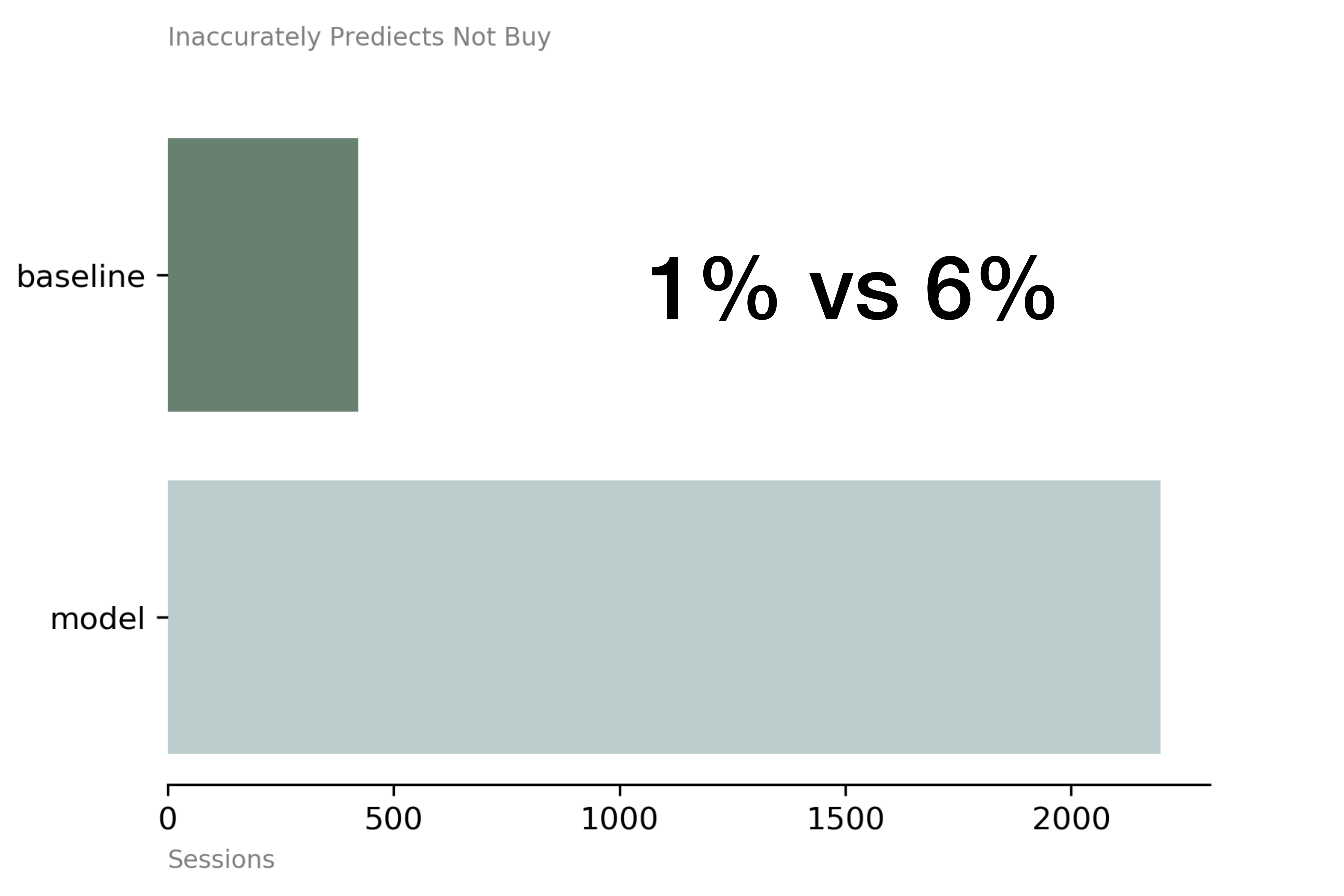
No Buy?

Buyer Identification

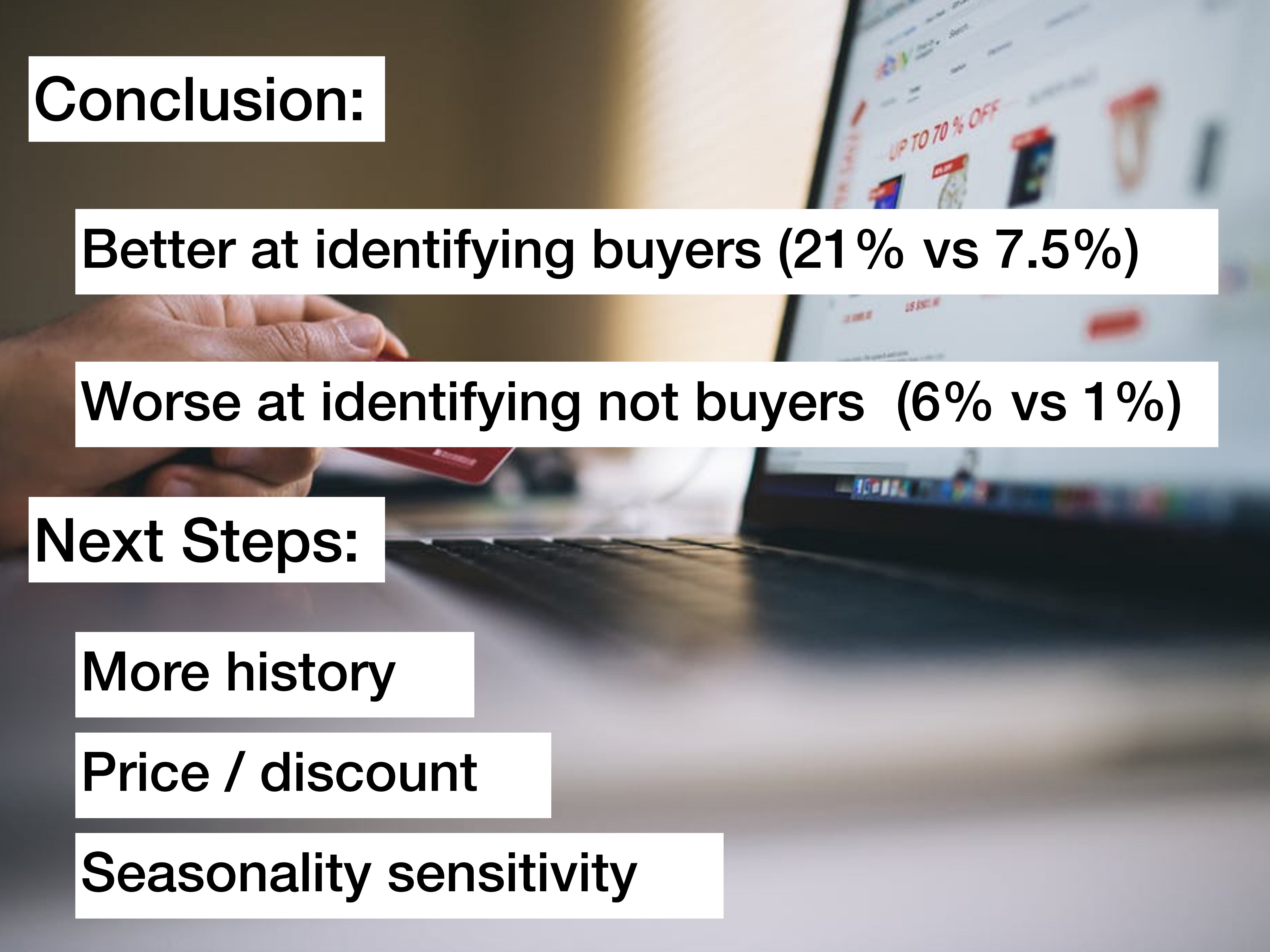


7.5% vs 21%

Not Buyer Identification



1% vs 6%



Conclusion:

- Better at identifying buyers (21% vs 7.5%)
- Worse at identifying not buyers (6% vs 1%)

Next Steps:

- More history
- Price / discount
- Seasonality sensitivity

Thank you

jandlsalazeradams@gmail.com

linkedin.com/in/jason-salazer-adams

github.com/jason-sa

Appendix

Session definition and features

- Session - A session is defined as any time there is a 30 minute gap between events for a visitor.
- Features generated
  - Count of views
  - Session Length
  - Number of unique items viewed
  - Number of add to cart events
  - Number of transaction events
  - Average item availability - If 3 pages were viewed and 2 out of the 3 items were available, then item availability is 66%.

Sampling process

- Preformed stratified sampling for train/test split.
- Tested the following upsampling and downsampling methodologies to tune the model.
    - SMOTE
    - SMOTE then Tomek links
    - SMOTE then Edited Nearest Neighbors - this provided the best CV fit

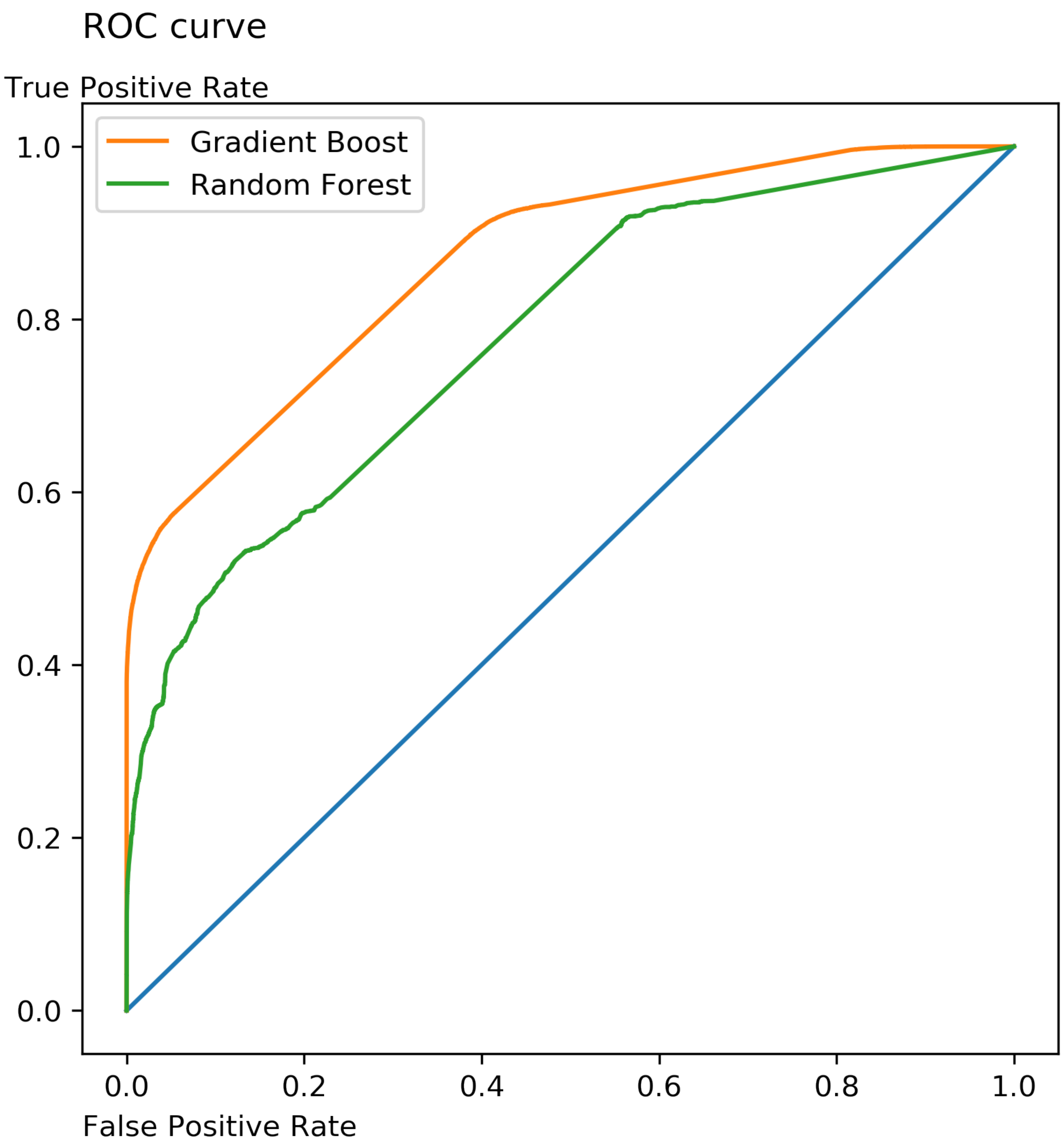
Model selection

- Used BayesSeachCV on Random Forest and Gradient Boost. CV was 10-fold with Stratified sampling in each fold.

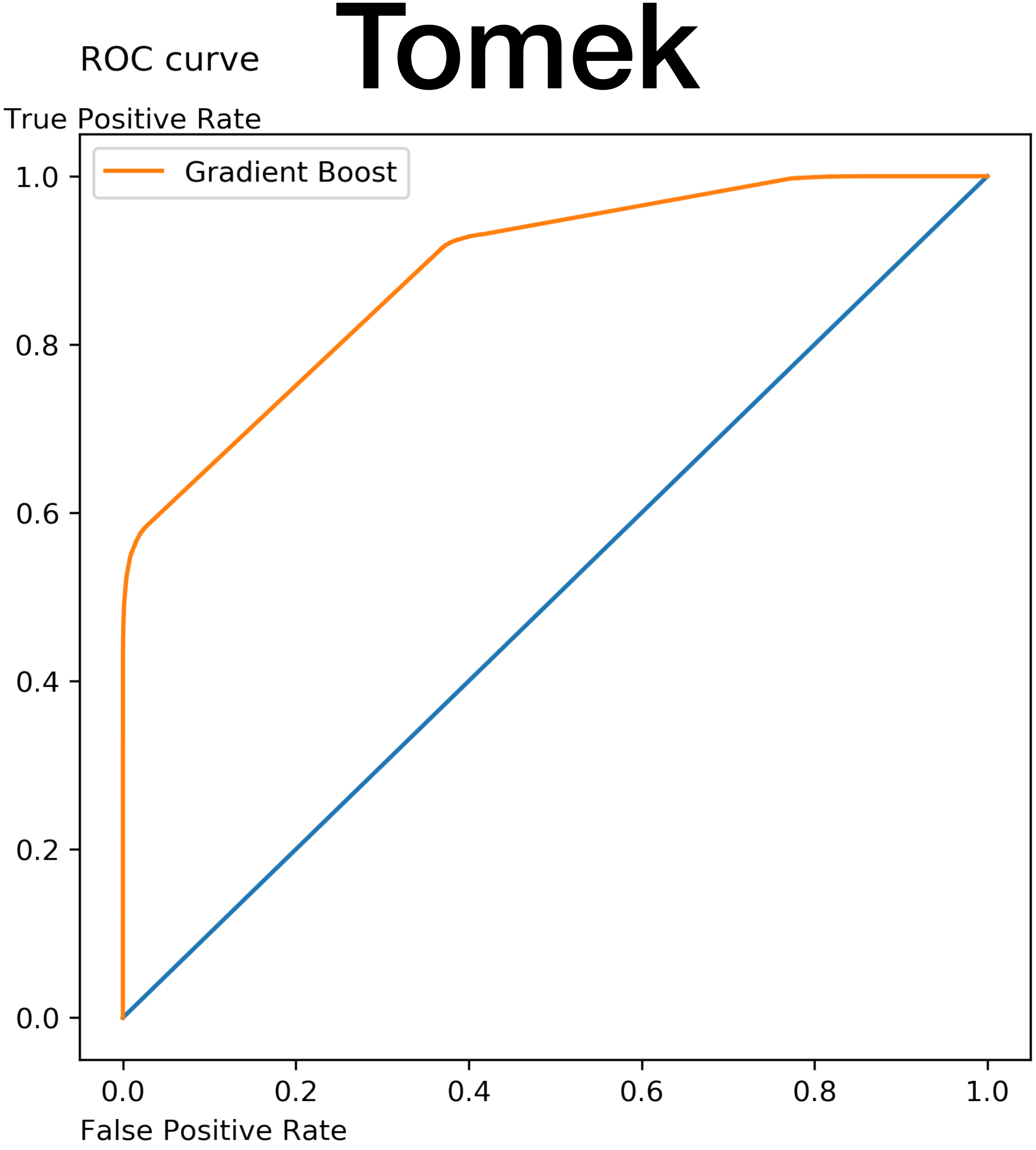
    - Tested Logistic regression, but got poorer results.

- The score function used for selection was AUC

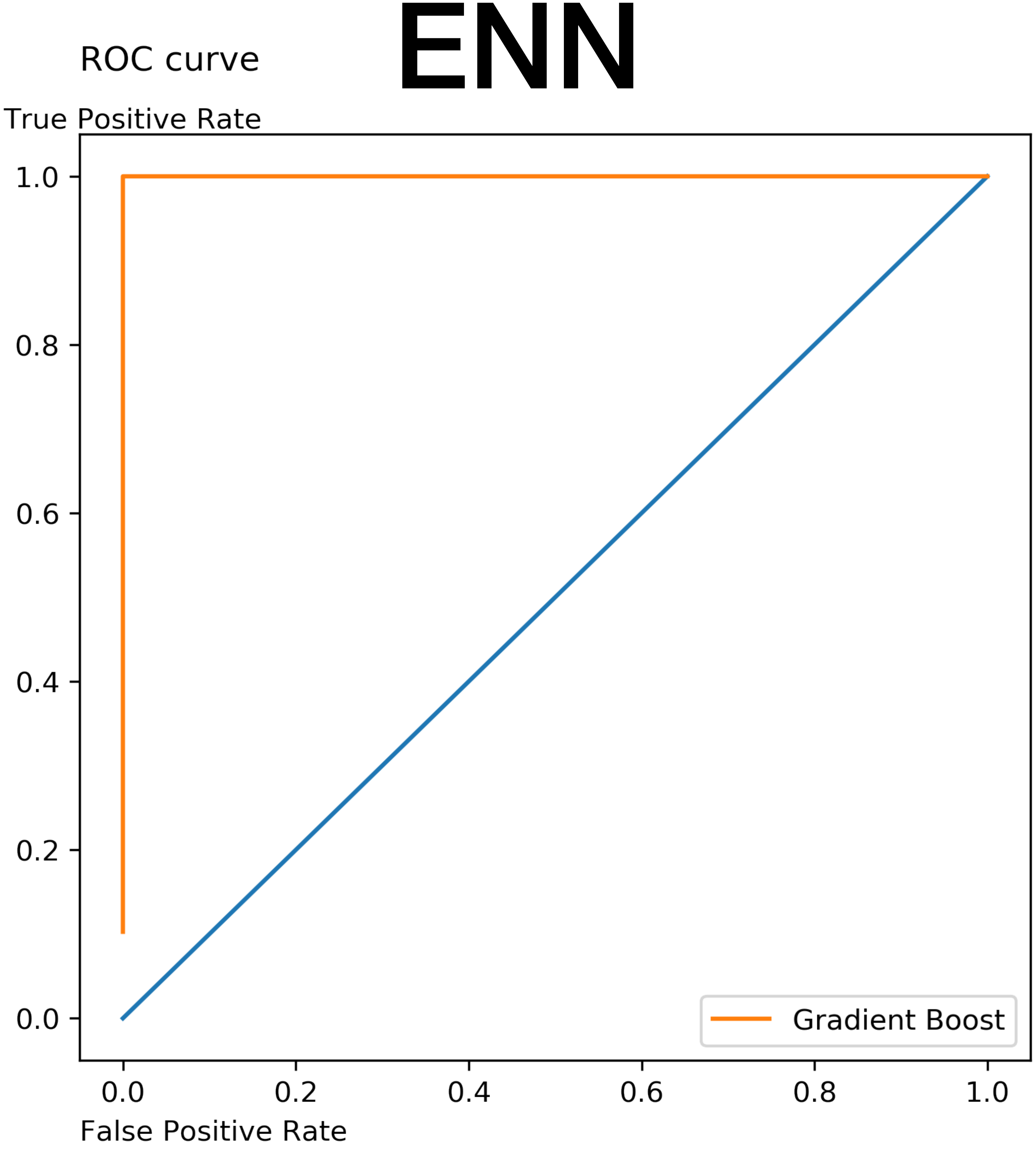
ROC Curve for SMOTE



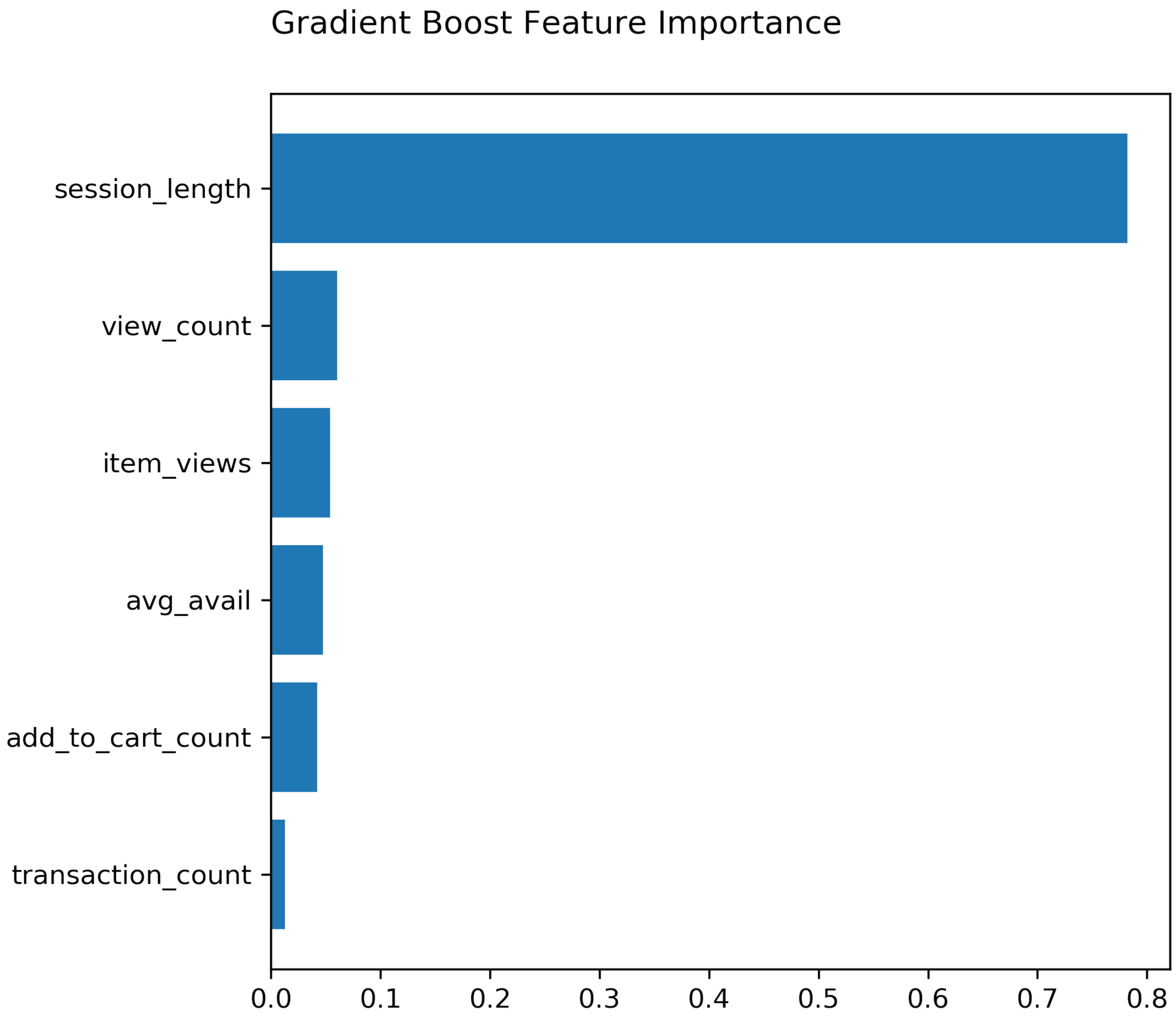
ROC Curve for SMOTE + Tomek



ROC Curve for SMOTE + ENN



Feature Importance (SMOTE + ENN)



Confusion Matrix

Model

	Predicted No	Predicted Yes
Actual No	32612	2197
Actual Yes	440	115

Recall: 0.21

Precision: 0.05

F1: 0.08

Baseline

	Predicted No	Predicted Yes
Actual No	34388	421
Actual Yes	513	42

Recall: 0.076

Precision: 0.091

F1: 0.083