

A top-down view of various gift boxes and bags scattered on a rustic wooden plank surface. The items include a large red box with a red ribbon, a purple box with an orange ribbon, a small red box with a gold ring inside, a green box with a green ribbon, and several white and red bags. The text "Buy or Not?" is overlaid in a large, bold, black font on a white rectangular background.

# Buy or Not?

Jason Salazer-Adams



# The Challenge

**Why?**

**Targeted Marketing**

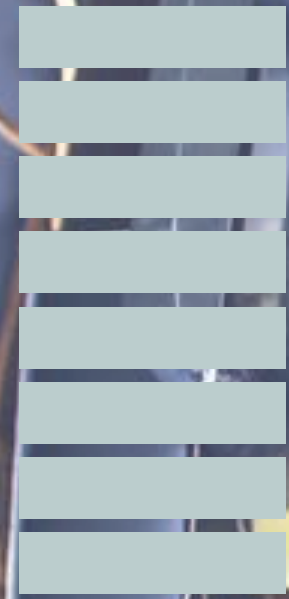
**Optimally allocate budget**

**Prevent profit erosion**

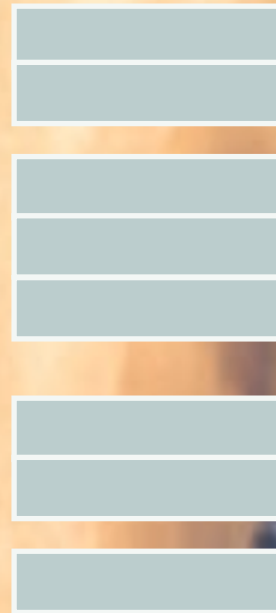


# Methodology

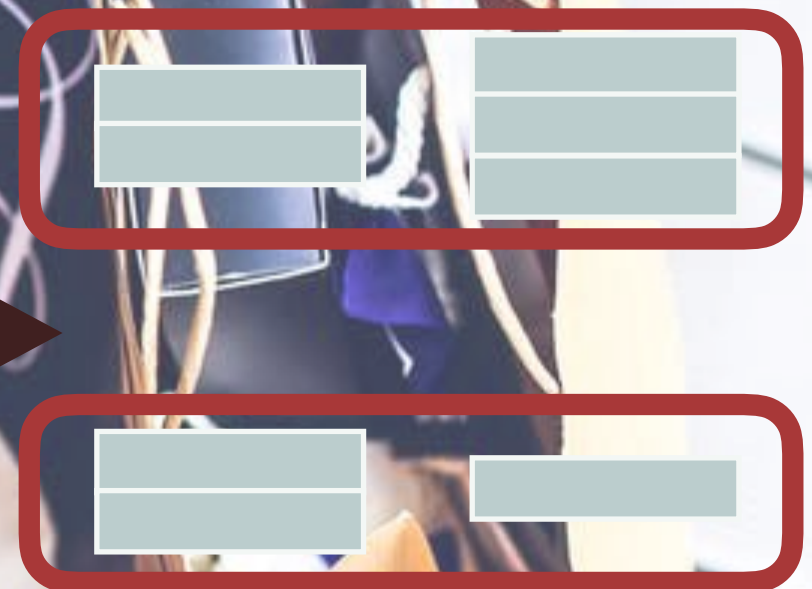
2.7M Events



1.7M Sessions



181k Pairs





# Model Comparison

Baseline

Model

VS

Buy

Buy?

No Buy

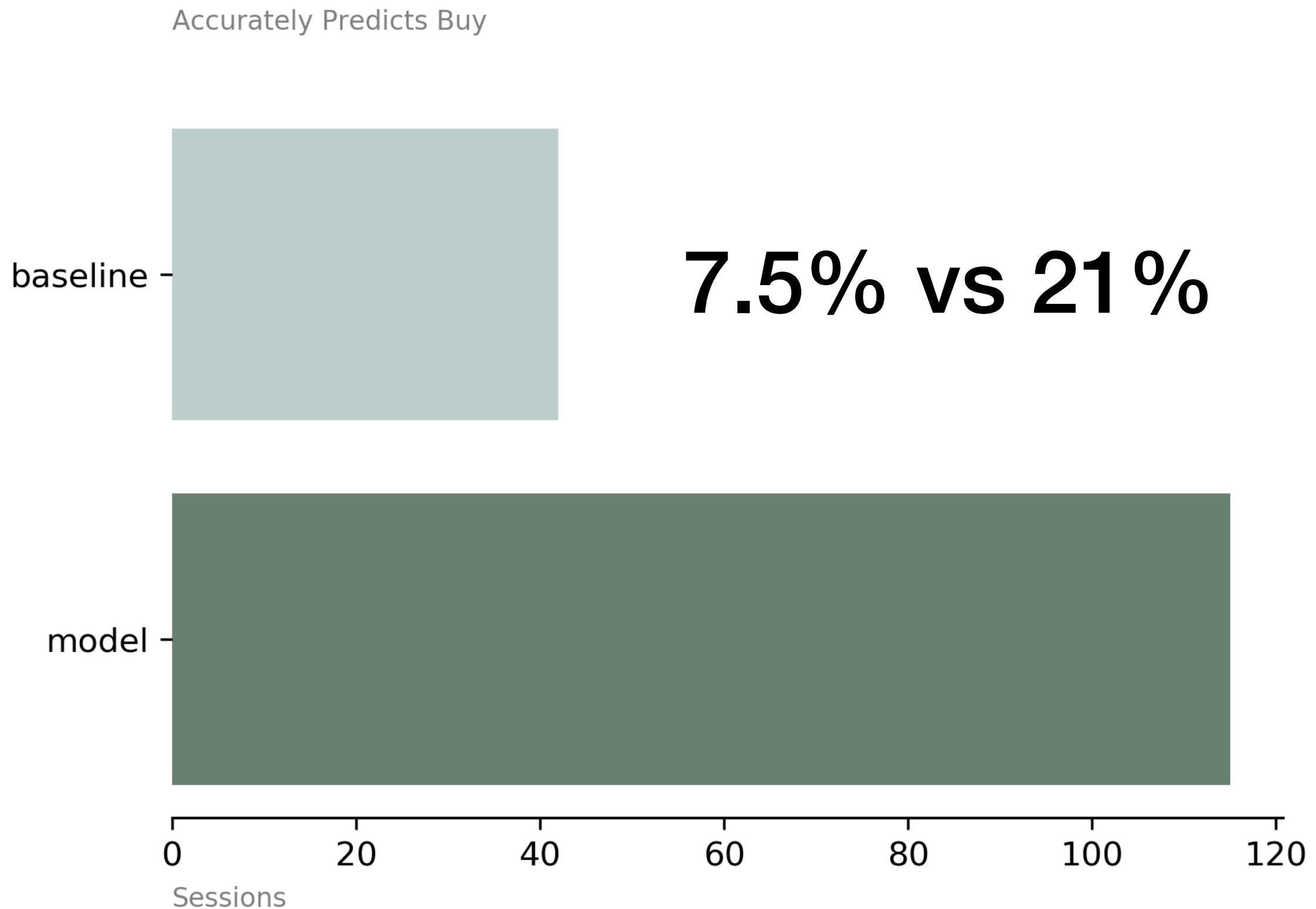
No Buy?

Buy?

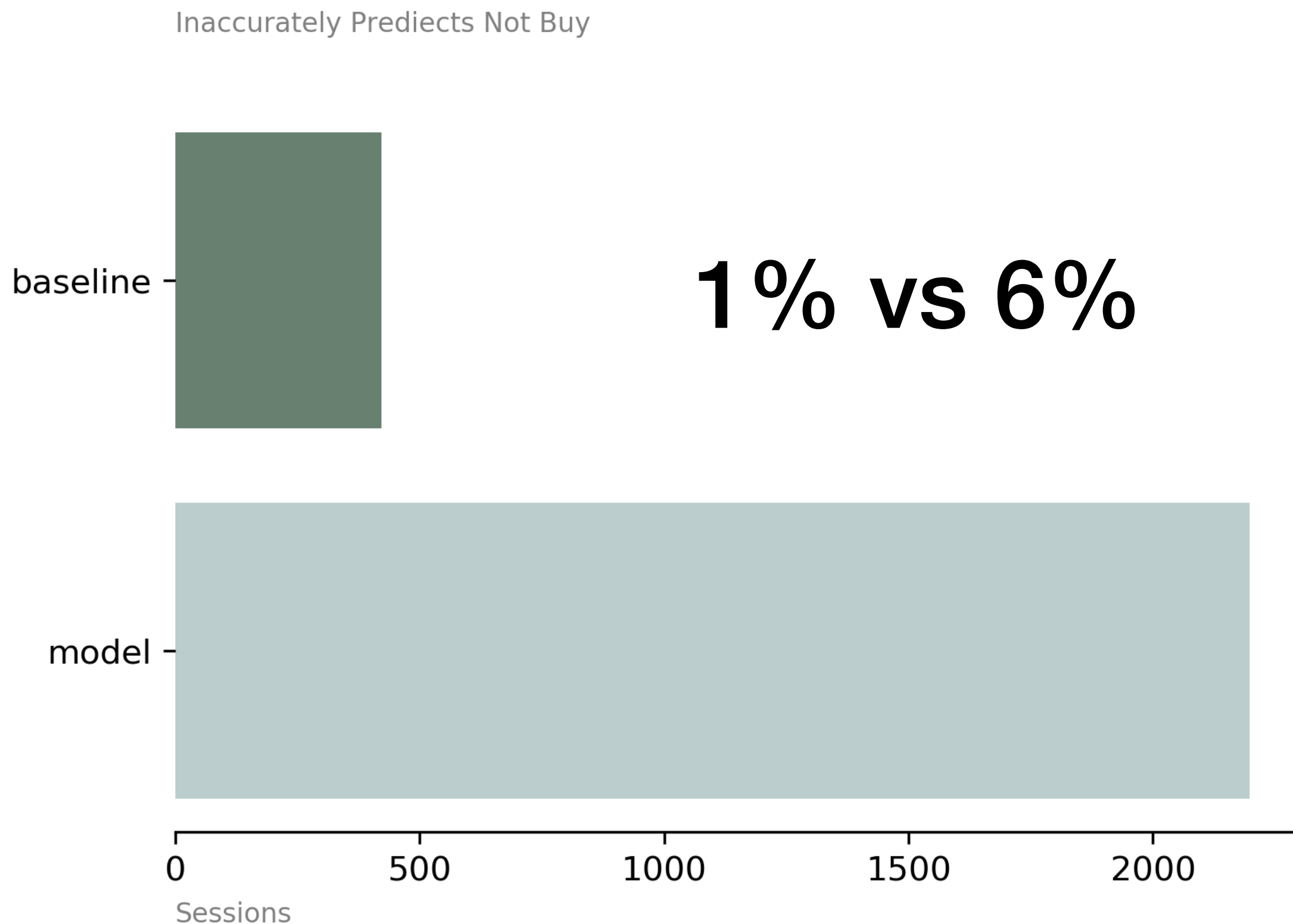
No Buy?

Item availability  
View  
Unique item  
Session last  
bought last  
added to cart last

# Buyer Identification



# Not Buyer Identification





## **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](mailto:jandlsalazeradams@gmail.com)

[linkedin.com/in/jason-salazer-adams](https://www.linkedin.com/in/jason-salazer-adams)

[github.com/jason-sa](https://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

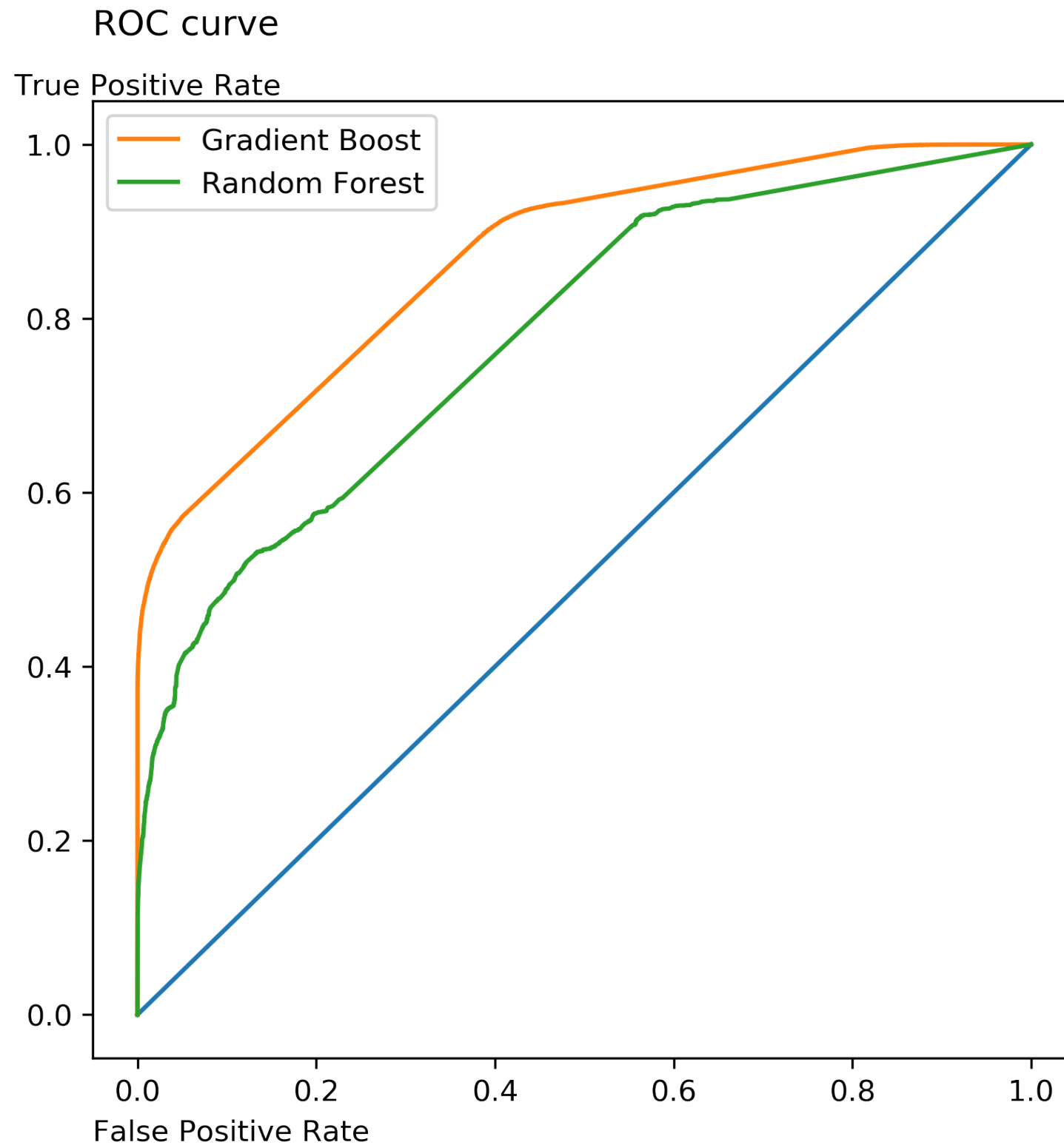
- Preformed stratified sampling for train/test split.
- Tested the following upsampling and downsampling methodologies to tune the model.
  - SMOTE
  - SMOTE then Tomek links
  - SMOTE then Edited Nearest Neighbors - this provided the best CV fit

# Model selection

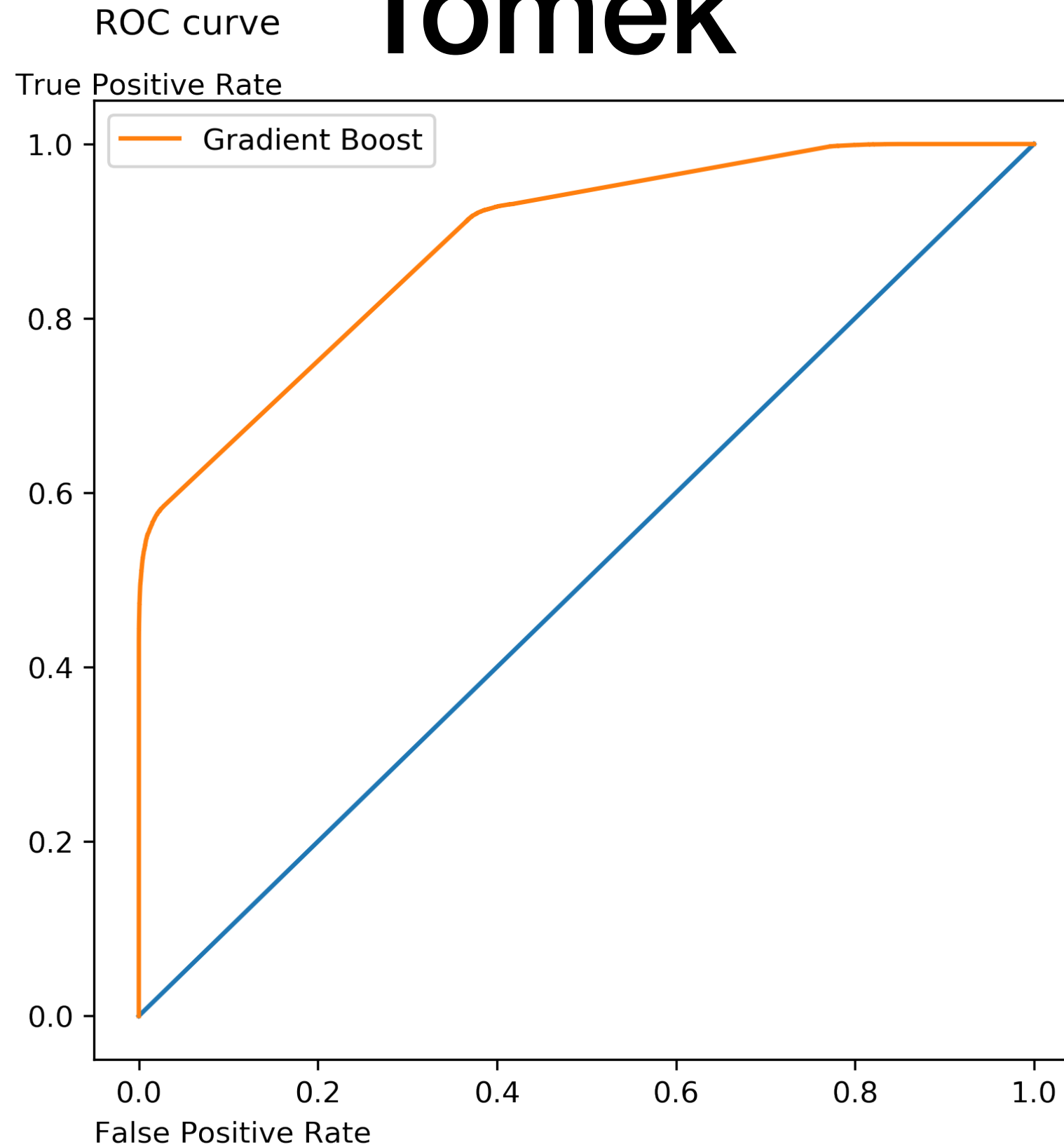
- Used BayesSeachCV on Random Forest and Gradient Boost. CV was 10-fold with Stratified sampling in each fold.
  - Tested Logistic regression, but got poorer results.
- The score function used for selection was AUC



# ROC Curve for SMOTE

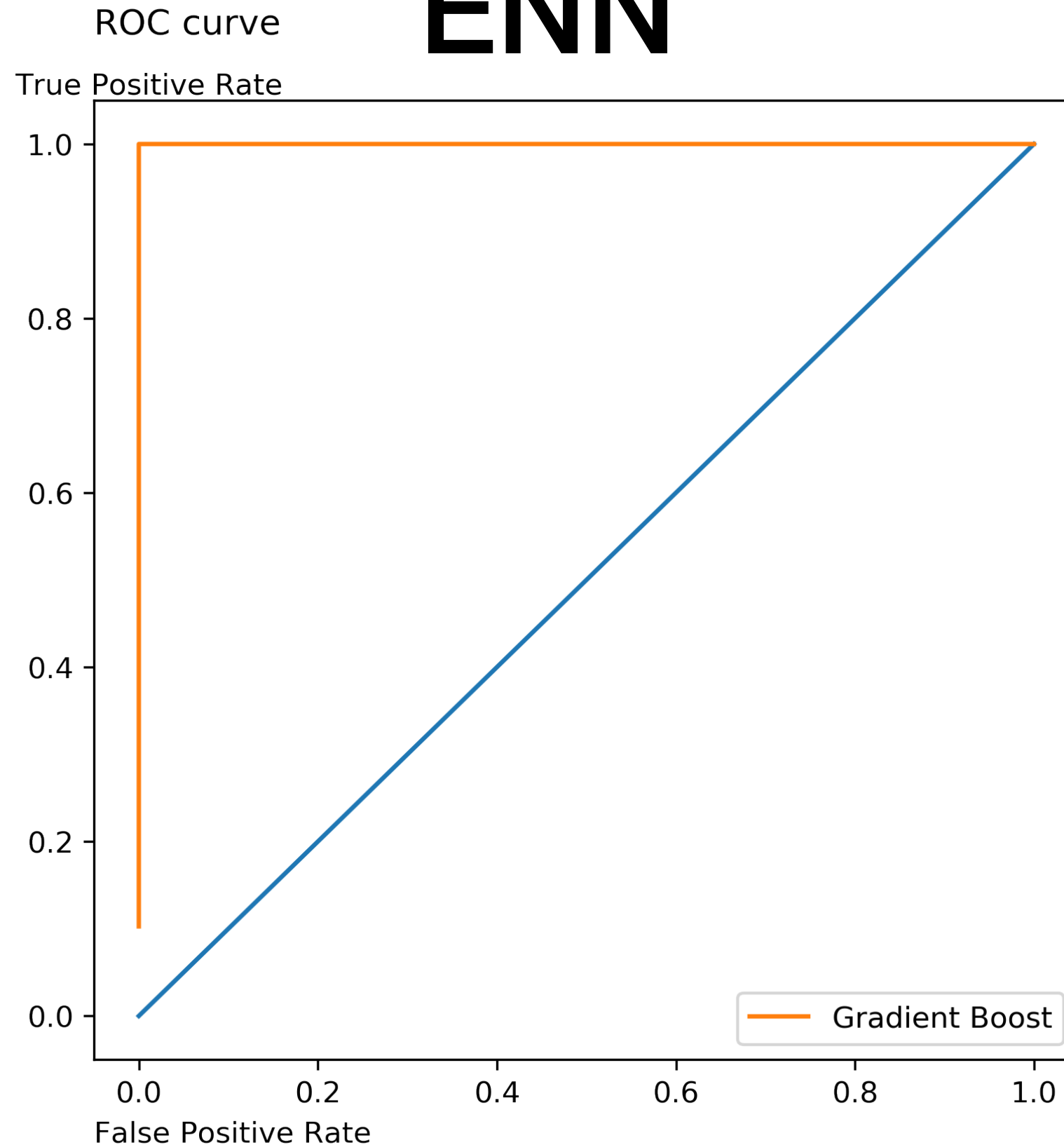


# ROC Curve for SMOTE + Tomek



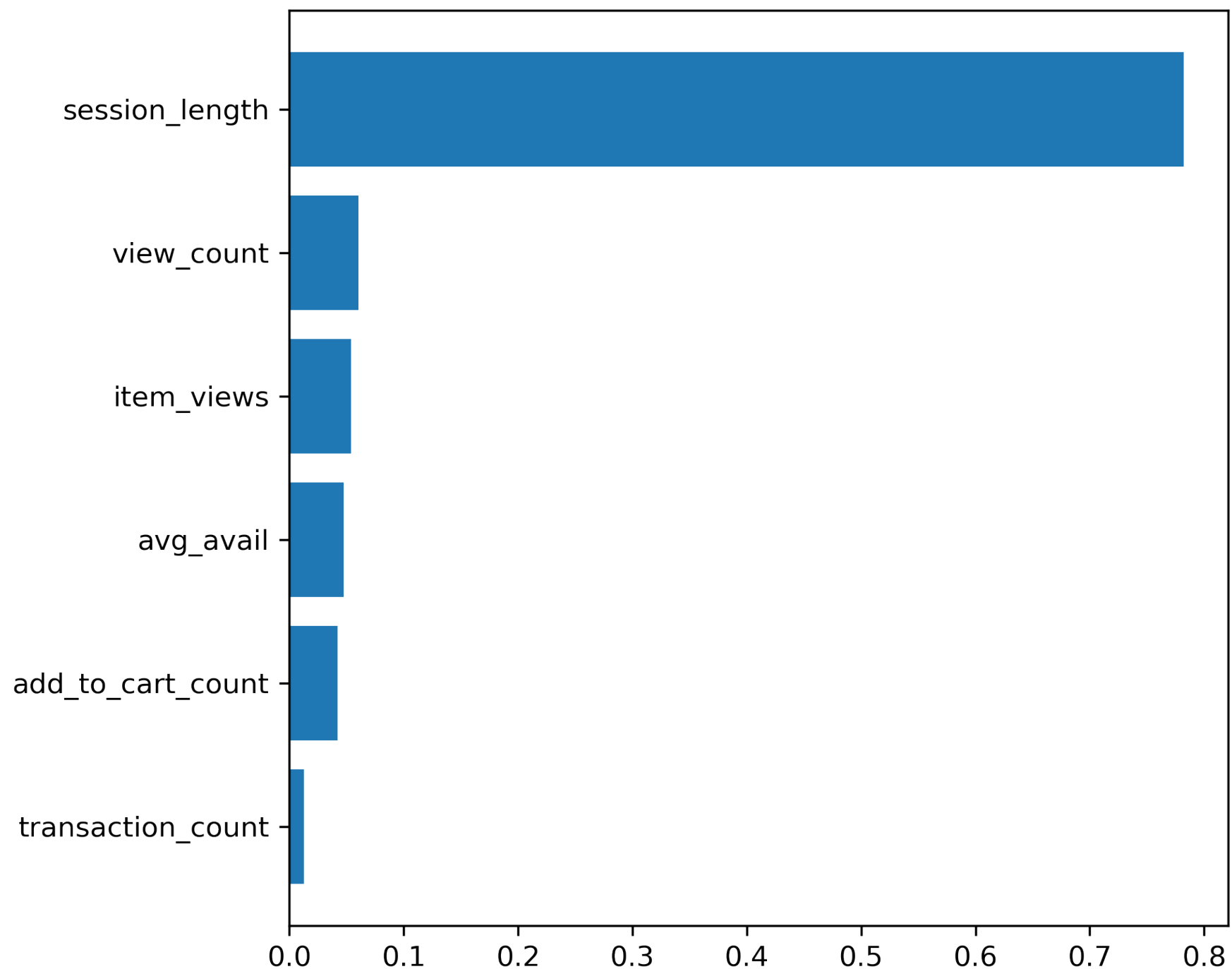


# ROC Curve for SMOTE + ENN



# Feature Importance (SMOTE + ENN)

Gradient Boost Feature Importance



# 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**