

A decorative graphic on the left side of the slide consisting of two overlapping parallelograms. The front one is blue and the back one is a light green. They are positioned diagonally, with the blue one partially covering the green one.

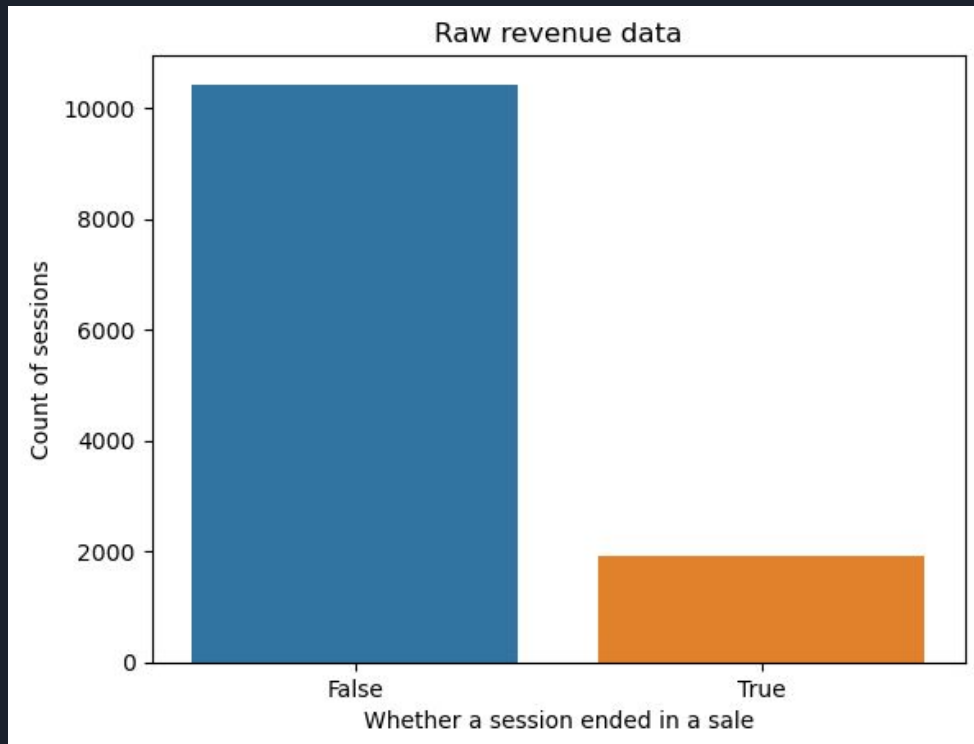
Online Shopping Intention and Customer Classification

An analysis by Kai Tamashiro

Can companies use machine learning to increase revenue from both online sales and passively through ads?



Revenue Classification





Finding the best model for revenue classification

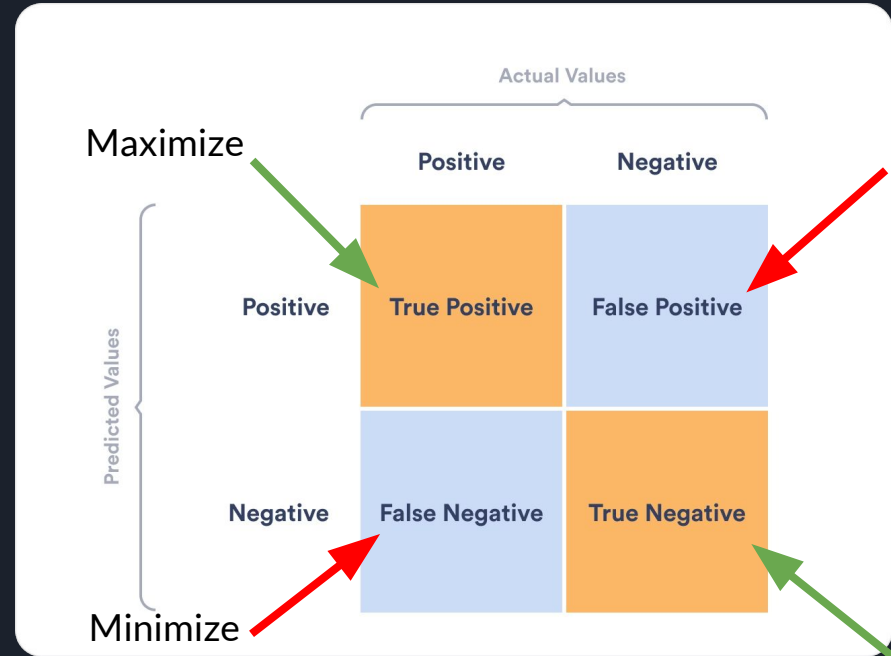
```
sm = SMOTE(random_state=42)
X_trainOS, y_trainOS = sm.fit_resample(X_train, y_train)
Counter(y_trainOS)

Counter({False: 8307, True: 8307})
```

Over sampling to
create equal classes

Identifying best model

Scores closer to 1

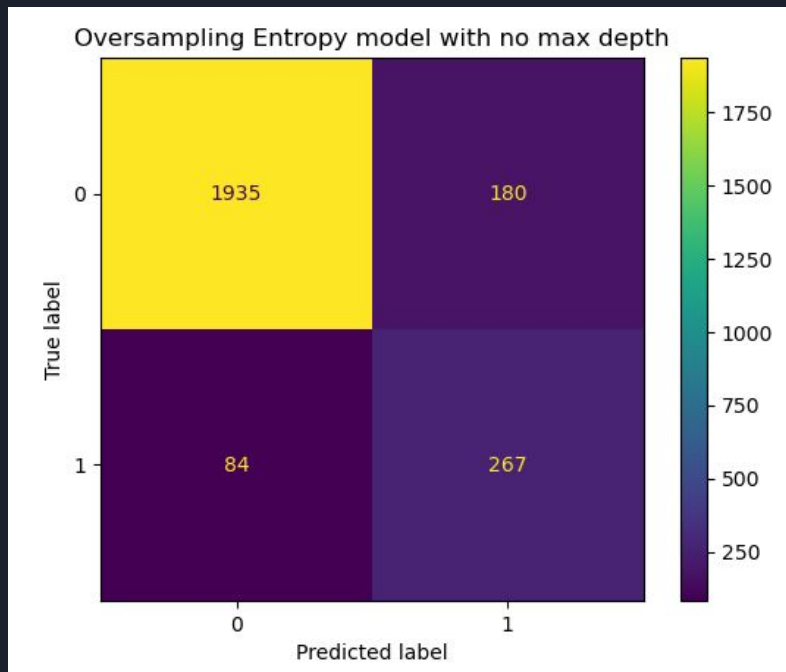


Random Forest: Entropy

Avg precision: .91

Avg recall: .89

F1-score: .90

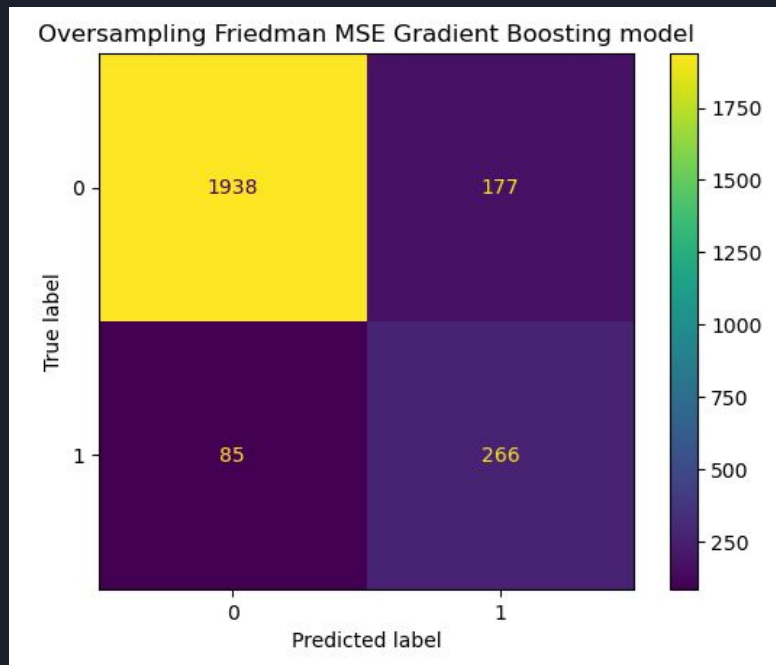


Gradient Boosting

Avg precision: .91

Avg recall: .89

F1-score: .90





Identifying the best model, part 2

```
nm = NearMiss()  
X_trainUS, y_trainUS = nm.fit_resample(X_train, y_train)  
Counter(y_trainUS)  
  
Counter({False: 1557, True: 1557})
```

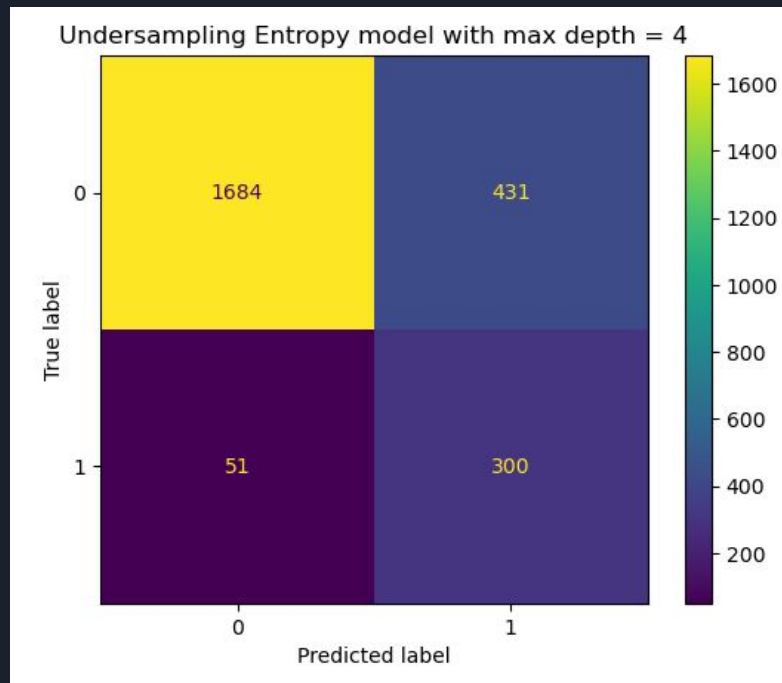
Under sampling to
create equal classes

Random Forest: Entropy with max depth

Avg precision: .89

Avg recall: .80

F1-score: .83

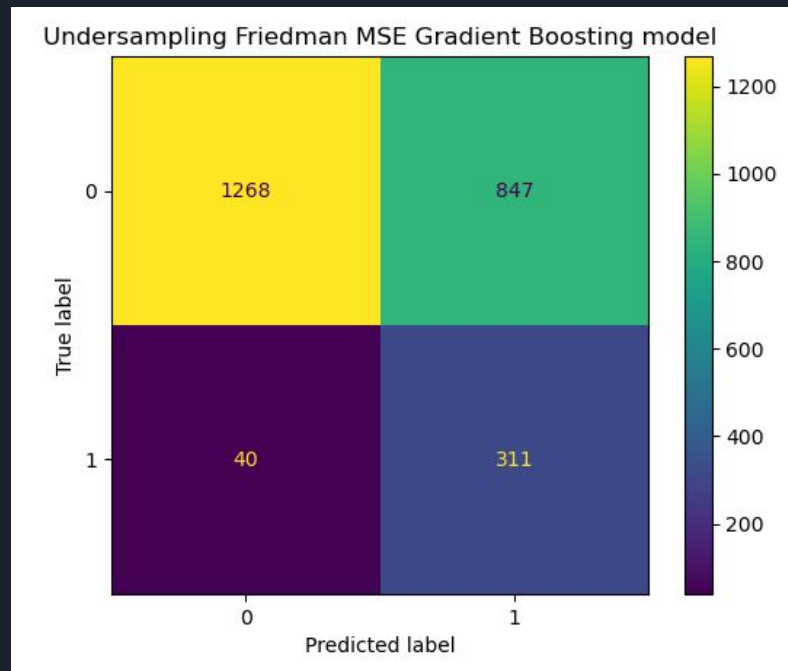


Gradient Boosting, part 2

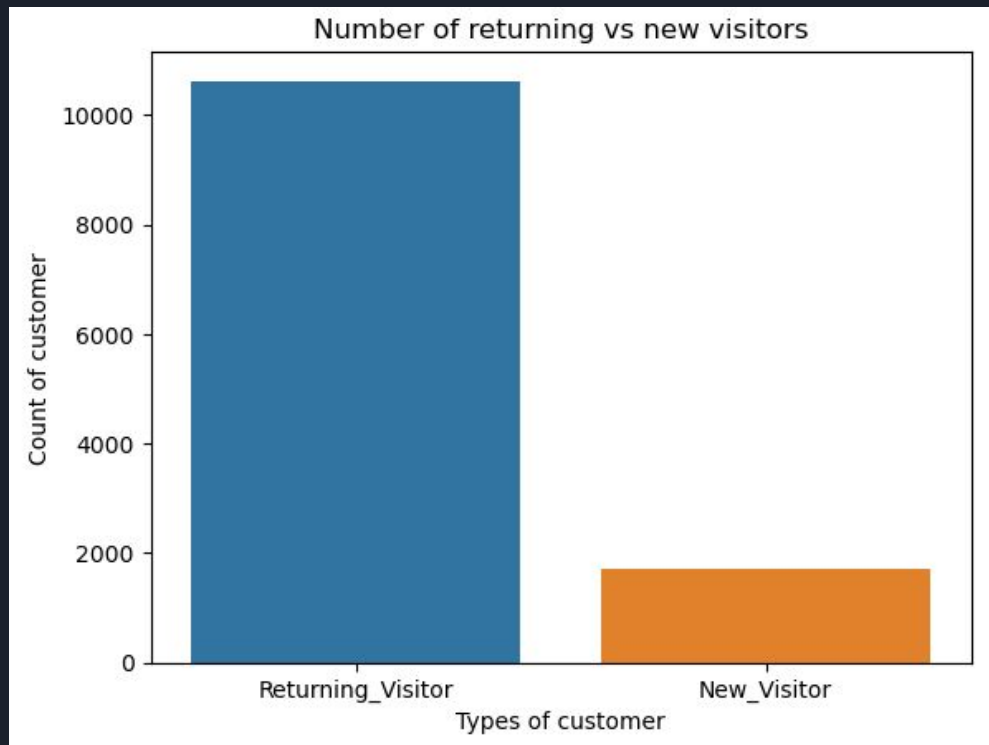
Avg precision: .87

Avg recall: .64

F1-score: .69



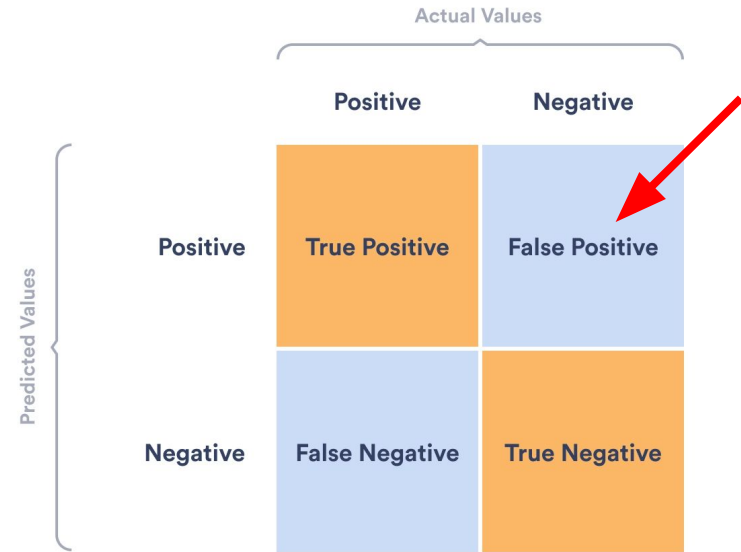
Customer Classification



Finding the best model for customer classification

Same steps as revenue classification

Smallest number of false positives

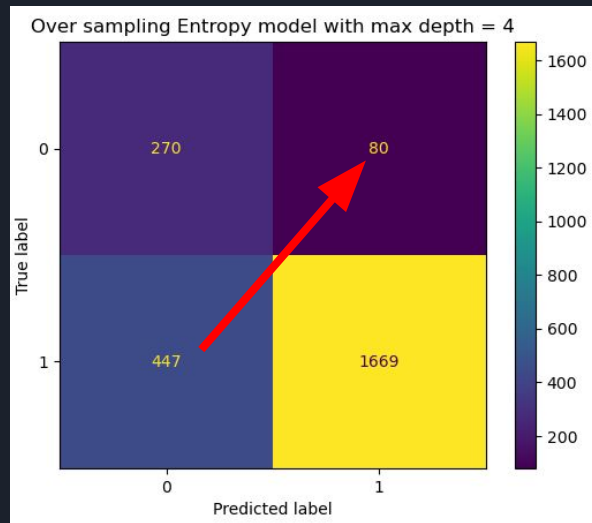
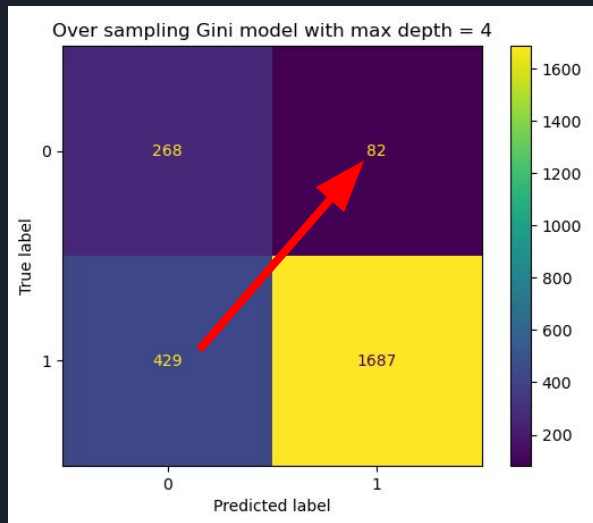
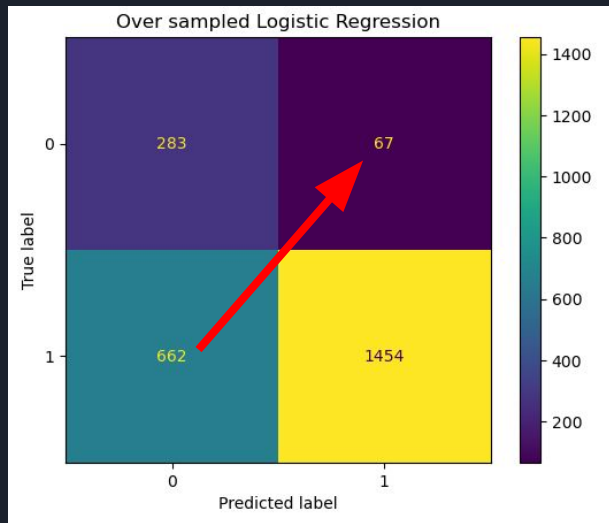


A confusion matrix diagram for customer classification. The matrix is a 2x2 grid. The columns are labeled 'Actual Values' with 'Positive' and 'Negative' headers. The rows are labeled 'Predicted Values' with 'Positive' and 'Negative' headers. The cells are: Top-Left (Positive Predicted, Positive Actual) is orange and labeled 'True Positive'; Top-Right (Positive Predicted, Negative Actual) is light blue and labeled 'False Positive', with a red arrow pointing to it; Bottom-Left (Negative Predicted, Positive Actual) is light blue and labeled 'False Negative'; Bottom-Right (Negative Predicted, Negative Actual) is orange and labeled 'True Negative'.

Actual Values			
		Positive	Negative
Predicted Values	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

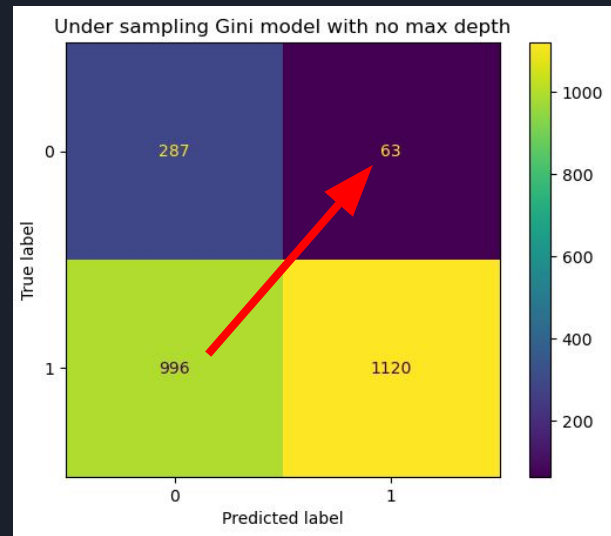
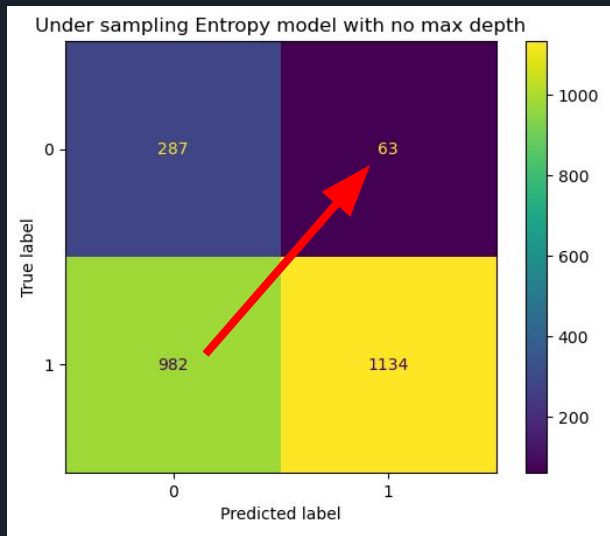
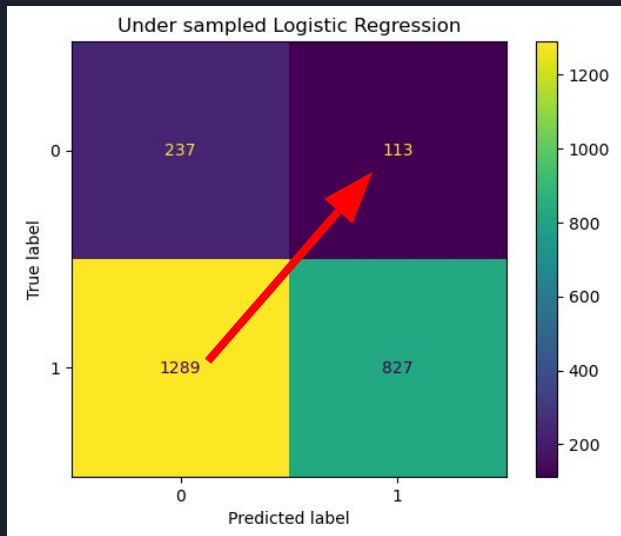
Lowest false positives

Logistic Regression?



Under sampling to test flexibility

Random Forests Triumph





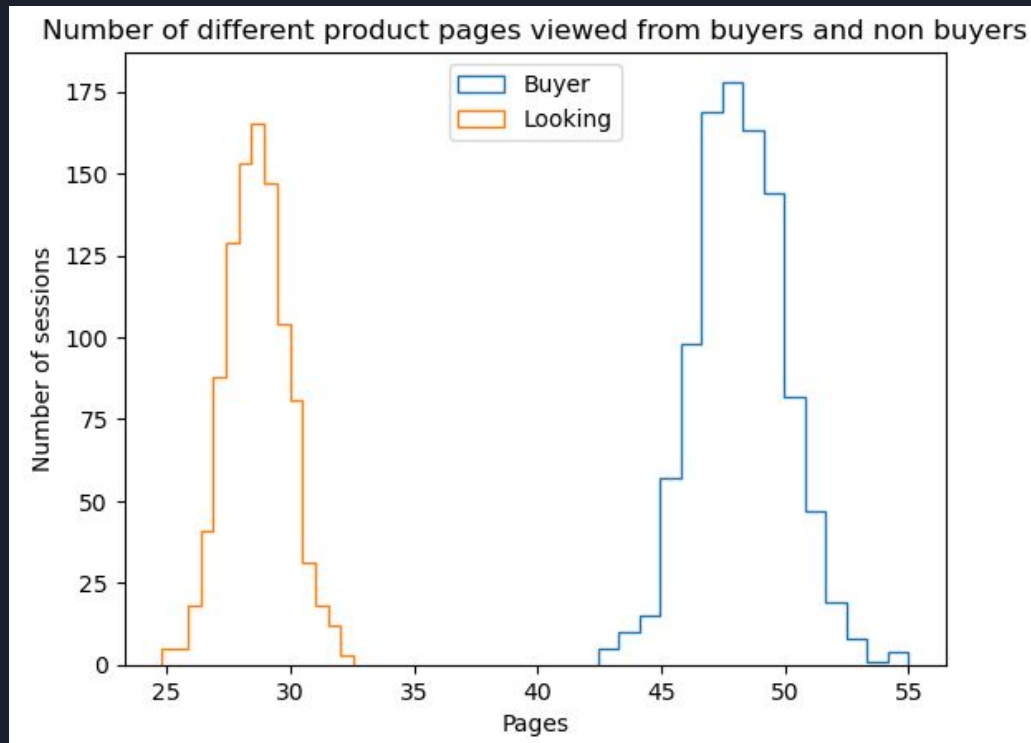
Implement Entropy Random Forest Model

Tweak model to
appropriate max depth

Push pop-ups to
customers as they are
browsing the site



Utilize pages viewed



Invest in premium ad network

Passive revenue from ad impressions

