

# Dataset 3: Amazon Reviews for Sentiment Analysis

Dataset: This dataset consists of 0.4 million Amazon customer reviews (input text) and “positive”/ “negative” class labels for learning. The dataset has almost equal distribution of data for each class label.

I used random sampling implemented as stratified random sampling to take 10000 records as dataset, having randomly selected 5000 positive labelled records and randomly selected 5000 negative labelled records. Then, used 4000 from each class category in training set and remaining in test set. Thus, total stratified random sampled dataset with 8000 records in training set and 2000 records in test set.

Problem: The problem is to find a classifier for this dataset that has high accuracy and left and top corner approaching ROC plot value.

# Machine learning techniques used:

- Decision Trees: Random Forest

I used Random forest amongst decision tree algorithms because it is an ensemble algorithm of decision tree category of classifiers.

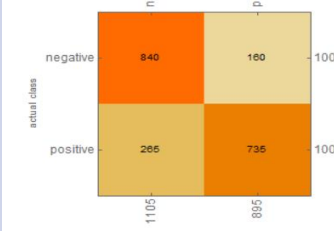
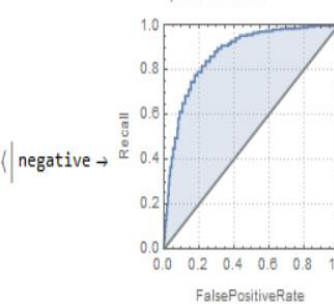
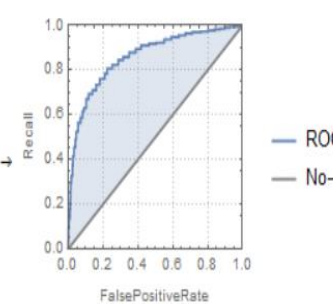
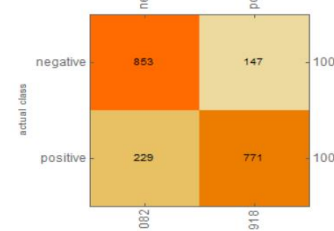
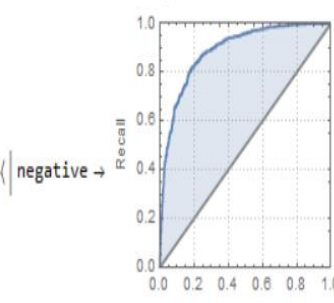
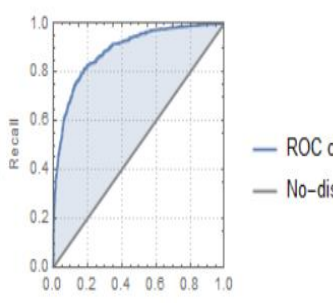
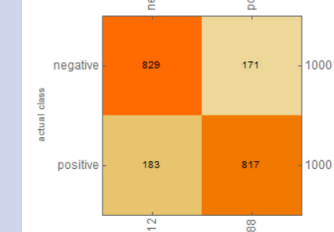
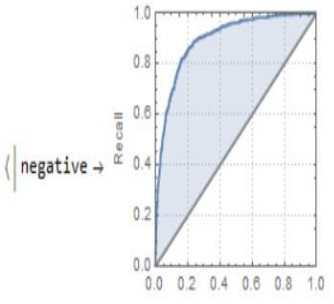
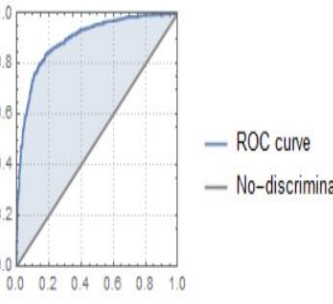
- Neural Networks: NeuralNetwork of Classifier of Mathematica

I used this classifier amongst Neural Networks because this algorithm in Mathematica is based on deep learning and deep learning performs well on text and image analysis.

- Support Vector Machine:

I used Support Vector machine amongst others because it is a non-probabilistic based classifier and it avoids overfitting on data with a wide margin between classes. I chose this after observing results on first two algorithms.

# Visualizations

Algorithm	Accuracy	Precision	Recall	Confusion Matrix	ROC curve
Random Forest	78.75 %	Negative: 76.02% Positive: 82.12%	Negative: 84% Positive: 73.5%	<p>0.7875 &lt;  negative → 0.760181, positive → 0.821229  &gt; &lt;  negative → 0.84, positive → 0.735  &gt;</p>  <p>actual class</p> <p>predicted class</p>	 <p>negative →</p> <p>ROC curve</p> <p>No-discrimination line</p> <p>positive →</p>  <p>ROC curve</p> <p>No-discrimination line</p>
Neural Network	81.2 %	Negative: 78.83 % Positive: 83.98%	Negative: 85.3% Positive: 77.1%	<p>0.812 &lt;  negative → 0.788355, positive → 0.839869  &gt; &lt;  negative → 0.853, positive → 0.771  &gt;</p>  <p>actual class</p> <p>predicted class</p>	 <p>negative →</p> <p>ROC curve</p> <p>No-discrimination line</p> <p>positive →</p>  <p>ROC curve</p> <p>No-discrimination line</p>
Support Vector Machine	82.3 %	Negative: 81.9% Positive: 82.7%	Negative: 82.9% Positive: 81.7%	<p>0.823 &lt;  negative → 0.81917, positive → 0.826923  &gt; &lt;  negative → 0.829, positive → 0.817  &gt;</p>  <p>actual class</p> <p>predicted class</p>	 <p>negative →</p> <p>ROC curve</p> <p>No-discrimination line</p> <p>positive →</p>  <p>ROC curve</p> <p>No-discrimination line</p>