Configuration

In [1]:
```
# Parameters
ENABLE_COLAB = False

PROJECT_NAME = 'ML1010-Group-Project'
EXPERIMENT_NAME = 'ReviewText_Lemma_Bert2 (Random Forest)'
FILE_NAME = '01_ML1010_GP_RF_Bert2'
LOAD_FROM_EXP = False

#Root Machine Learning Directory. Projects appear underneath
GOOGLE_DRIVE_MOUNT = '/content/gdrive'
COLAB_ROOT_DIR = GOOGLE_DRIVE_MOUNT + '/MyDrive/Colab Notebooks'
COLAB_INIT_DIR = COLAB_ROOT_DIR + '/utility_files'

LOCAL_ROOT_DIR = '/home/magni//ML_Root/project_root'
LOCAL_INIT_DIR = LOCAL_ROOT_DIR + '/utility_files'
```

Bootstrap Environment

In [2]:
```
#add in support for utility file directory and importing
import sys
import os

if ENABLE_COLAB:
  #Need access to drive
  from google.colab import drive
  drive.mount(GOOGLE_DRIVE_MOUNT, force_remount=True)

  #add in utility directory to syspath to import
  INIT_DIR = COLAB_INIT_DIR
  sys.path.append(os.path.abspath(INIT_DIR))

  #Config environment variables
  ROOT_DIR = COLAB_ROOT_DIR

else:
  #add in utility directory to syspath to import
  INIT_DIR = LOCAL_INIT_DIR
  sys.path.append(os.path.abspath(INIT_DIR))

  #Config environment variables
  ROOT_DIR = LOCAL_ROOT_DIR

#Import Utility Support
from jarvis import Jarvis
jarvis = Jarvis(ROOT_DIR, PROJECT_NAME)

import mv_python_utils as mvutils
```

```
Wha...where am I?
I am awake now.
```

```
I have set your current working directory to /home/magni/ML_Root/project_root
/ML1010-Group-Project
The current time is 10:30
Hello sir. Extra caffeine may help.
```

Setup Runtime Environment

In [3]:

```
if ENABLE_COLAB:
  #!pip install scipy -q
  #!pip install scikit-learn -q
  #!pip install pycaret -q
  #!pip install matplotlib -q
  #!pip install joblib -q
  #!pip install pandasql -q
  !pip install umap_learn -q
  !pip install sentence_transformers -q
  !pip install spacytextblob -q
  !pip install flair -q
  display('Google Colab enabled')
else:
  display('Google Colab not enabled')

#Common imports
import json
import pandas as pd
import numpy as np
import matplotlib
import re
import nltk
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
from sklearn import metrics
from sklearn.datasets import load_digits
from sklearn.model_selection import train_test_split as tts
#from yellowbrick.classifier import ConfusionMatrix
#from sklearn.linear_model import LogisticRegression
from yellowbrick.target import ClassBalance
from xgboost import XGBClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, confusion_matrix
from sklearn.svm import SVC
from sklearn.ensemble import RandomForestClassifier

nltk.download('stopwords')
%matplotlib inline
```

```
'Google Colab not enabled'
[nltk_data] Downloading package stopwords to /home/magni/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

In [4]:
```
import importlib
import cw_df_metric_utils as cwutils
import DataPackage as dp
import DataPackageSupport as dps
import DataExperiment
import DataExperimentSupport
```

```
2022-01-15 10:30:31.038980: W tensorflow/stream_executor/platform/default/dso
_loader.cc:64] Could not load dynamic library 'libcudart.so.11.0'; dlerror: l
ibcudart.so.11.0: cannot open shared object file: No such file or directory
2022-01-15 10:30:31.039010: I tensorflow/stream_executor/cuda/cudart_stub.cc:
29] Ignore above cudart dlerror if you do not have a GPU set up on your machi
ne.
```

In [23]:
```
importlib.reload(dp)
importlib.reload(dps)
importlib.reload(DataExperiment)
importlib.reload(DataExperimentSupport)
```

Out[23]:
```
<module 'DataExperimentSupport' from '/home/magni/ML_Root/project_root/utilit
y_files/DataExperimentSupport.py'>
```

Load Data

In [5]:
```
#axis_labels=[1,2,3,4,5]
axis_labels=[0,1]
classifier = RandomForestClassifier()
ANALSYSIS_COL = 'reviewText_lemma_bert'
UNIQUE_COL = 'uuid'
TARGET_COL = 'overall_posneg'
```

In [6]:
```
if LOAD_FROM_EXP:
    #start from saved state
    myExp = jarvis.loadExperiment(FILE_NAME)
    myExp.display()

else:
    #start from source file and regenerate
    testDf = pd.read_pickle(jarvis.DATA_DIR_WORK + "/01_NL_ReviewText_All(new

    testDfBert = cwutils.getBertEncodeFrame(df=testDf,
                                            bertColumn=ANALSYSIS_COL,
                                            uniqueColumn=UNIQUE_COL,
                                            otherColumns=[TARGET_COL]
                                            )



    myExp = DataExperiment.DataExperiment(projectName=PROJECT_NAME,
                                          experimentName=EXPERIMENT_NAME,
                                          origData=testDfBert,
                                          uniqueColumn=UNIQUE_COL,
                                          targetColumn=TARGET_COL,
                                          classifier=classifier)
```

```
DataExperiment summary:
---> projectName: ML1010-Group-Project
---> experimentName: ReviewText_Lemma_Bert2 (Random Forest)
---> isDataPackageLoaded: True
---> isBaseModelLoaded: False
---> isBaseModelPredicted: False
---> isBaseModelLearningCurveCreated: False
---> isFinalModelLoaded: False
---> isFinalModelPredicted: False
---> isFinalModelLearningCurveCreated: False
---> isClassifierLoaded: True
RandomForestClassifier()

    DataPackage summary:
    Attributes:
    ---> uniqueColumn: uuid
    ---> targetColumn: overall_posneg
    Process:
    ---> isBalanced: False
    ---> isTrainTestSplit: False
    Data:
    ---> isOrigDataLoaded: True
    ---> isTrainDataLoaded: False
    ---> isTestDataLoaded: False
```

In [7]:
```
myExp.processDataPackage()
```


```
Undersampling data to match min class: 0 of size: 13440
```


```
Completed train/test split (test_size = 0.2):
---> Original data size: 26880
---> Training data size: 21504
---> Testing data size: 5376
---> Stratified on column: overall_posneg
```

In [8]:

```
myExp.display()
```

```
DataExperiment summary:
---> projectName: ML1010-Group-Project
---> experimentName: ReviewText_Lemma_Bert2 (Random Forest)
---> isDataPackageLoaded: True
---> isBaseModelLoaded: False
---> isBaseModelPredicted: False
---> isBaseModelLearningCurveCreated: False
---> isFinalModelLoaded: False
---> isFinalModelPredicted: False
---> isFinalModelLearningCurveCreated: False
---> isClassifierLoaded: True
RandomForestClassifier()

    DataPackage summary:
    Attributes:
    ---> uniqueColumn: uuid
```

```
---> targetColumn: overall_posneg
Process:
---> isBalanced: True
---> isTrainTestSplit: True
Data:
---> isOrigDataLoaded: False
---> isTrainDataLoaded: True
---> isTestDataLoaded: True
```

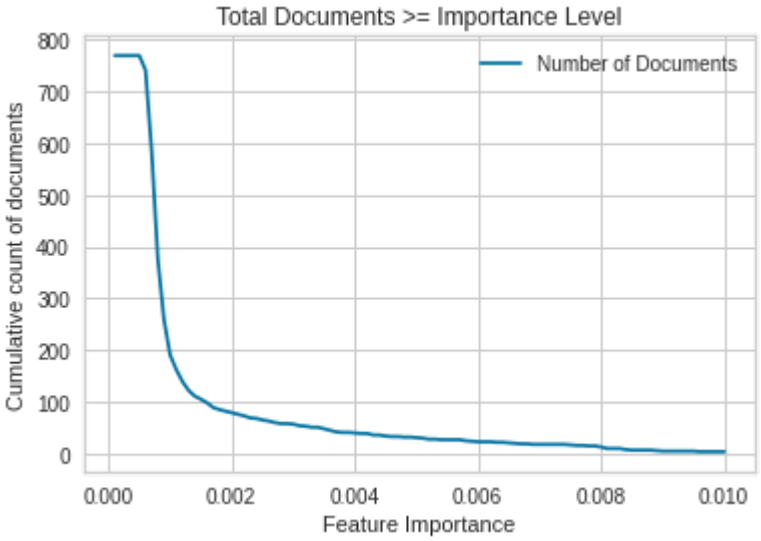
In [9]:
```
myExp.createBaseModel()
```

In [10]:
```
myExp.predictBaseModel()
```

```
Base Model Stats:
Accuracy: 0.8
Precision: 0.8
Recalll: 0.8
F1 Score: 0.8
Cohen kappa:: 0.59
```

In [11]:
```
impFeatures = myExp.analyzeBaseModelFeatureImportance(returnAbove=0.002)
```

```
  0%|          | 0/101 [00:00<?, ?it/s]
Feature Importance Summary:
---> Original feature count: 768
---> Returned feature count: 80
---> Removed feature count: 688
---> Return items above (including): 0.002
```



In [12]:
```
myExp.createFinalModel(featureImportanceThreshold=0.002)
```

```
  0%|          | 0/101 [00:00<?, ?it/s]
  0%|          | 0/101 [00:00<?, ?it/s]
```

In [13]:
```
myExp.display()
```

```
DataExperiment summary:
---> projectName: ML1010-Group-Project
```

```
---> experimentName: ReviewText_Lemma_Bert2 (Random Forest)
---> isDataPackageLoaded: True
---> isBaseModelLoaded: True
---> isBaseModelPredicted: True
---> isBaseModelLearningCurveCreated: False
---> isFinalModelLoaded: True
---> isFinalModelPredicted: False
---> isFinalModelLearningCurveCreated: False
---> isClassifierLoaded: True
RandomForestClassifier()

    DataPackage summary:
    Attributes:
    ---> uniqueColumn: uuid
    ---> targetColumn: overall_posneg
    Process:
    ---> isBalanced: True
    ---> isTrainTestSplit: True
    Data:
    ---> isOrigDataLoaded: False
    ---> isTrainDataLoaded: True
    ---> isTestDataLoaded: True
```

In [14]:

```
myExp.predictFinalModel()
myExp.display()
```

```
Final Model Stats:
Accuracy: 0.8
Precision: 0.8
Recalll: 0.8
F1 Score: 0.8
Cohen kappa:: 0.59
DataExperiment summary:
---> projectName: ML1010-Group-Project
---> experimentName: ReviewText_Lemma_Bert2 (Random Forest)
---> isDataPackageLoaded: True
---> isBaseModelLoaded: True
---> isBaseModelPredicted: True
---> isBaseModelLearningCurveCreated: False
---> isFinalModelLoaded: True
---> isFinalModelPredicted: True
---> isFinalModelLearningCurveCreated: False
---> isClassifierLoaded: True
RandomForestClassifier()

    DataPackage summary:
    Attributes:
    ---> uniqueColumn: uuid
    ---> targetColumn: overall_posneg
    Process:
    ---> isBalanced: True
    ---> isTrainTestSplit: True
    Data:
    ---> isOrigDataLoaded: False
    ---> isTrainDataLoaded: True
    ---> isTestDataLoaded: True
```

In [15]:
```
myExp.createBaseModelLearningCurve(n_jobs=10)
```

```
[learning_curve] Training set sizes: [ 1720  3440  8601 17203]
[Parallel(n_jobs=10)]: Using backend LokyBackend with 10 concurrent workers.
[Parallel(n_jobs=10)]: Done   3 out of  20 | elapsed:    4.1s remaining:   2
3.1s
[Parallel(n_jobs=10)]: Done   9 out of  20 | elapsed:   17.8s remaining:   2
1.8s
[Parallel(n_jobs=10)]: Done  15 out of  20 | elapsed:   38.8s remaining:   1
2.9s
[Parallel(n_jobs=10)]: Done  20 out of  20 | elapsed:  1.3min finished
```

In [16]:
```
myExp.createFinalModelLearningCurve(n_jobs=10)
```

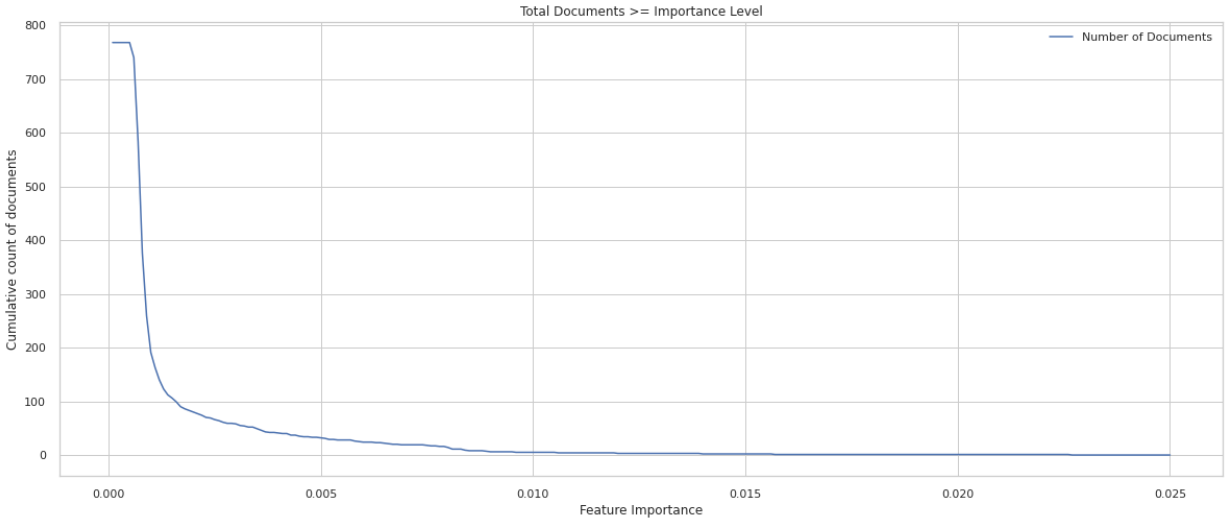
```
[Parallel(n_jobs=10)]: Using backend LokyBackend with 10 concurrent workers.
[learning_curve] Training set sizes: [ 1720  3440  8601 17203]
[Parallel(n_jobs=10)]: Done   3 out of  20 | elapsed:    1.3s remaining:
7.4s
[Parallel(n_jobs=10)]: Done   9 out of  20 | elapsed:    4.7s remaining:
5.8s
[Parallel(n_jobs=10)]: Done  15 out of  20 | elapsed:   10.3s remaining:
3.4s
[Parallel(n_jobs=10)]: Done  20 out of  20 | elapsed:   16.1s finished
```

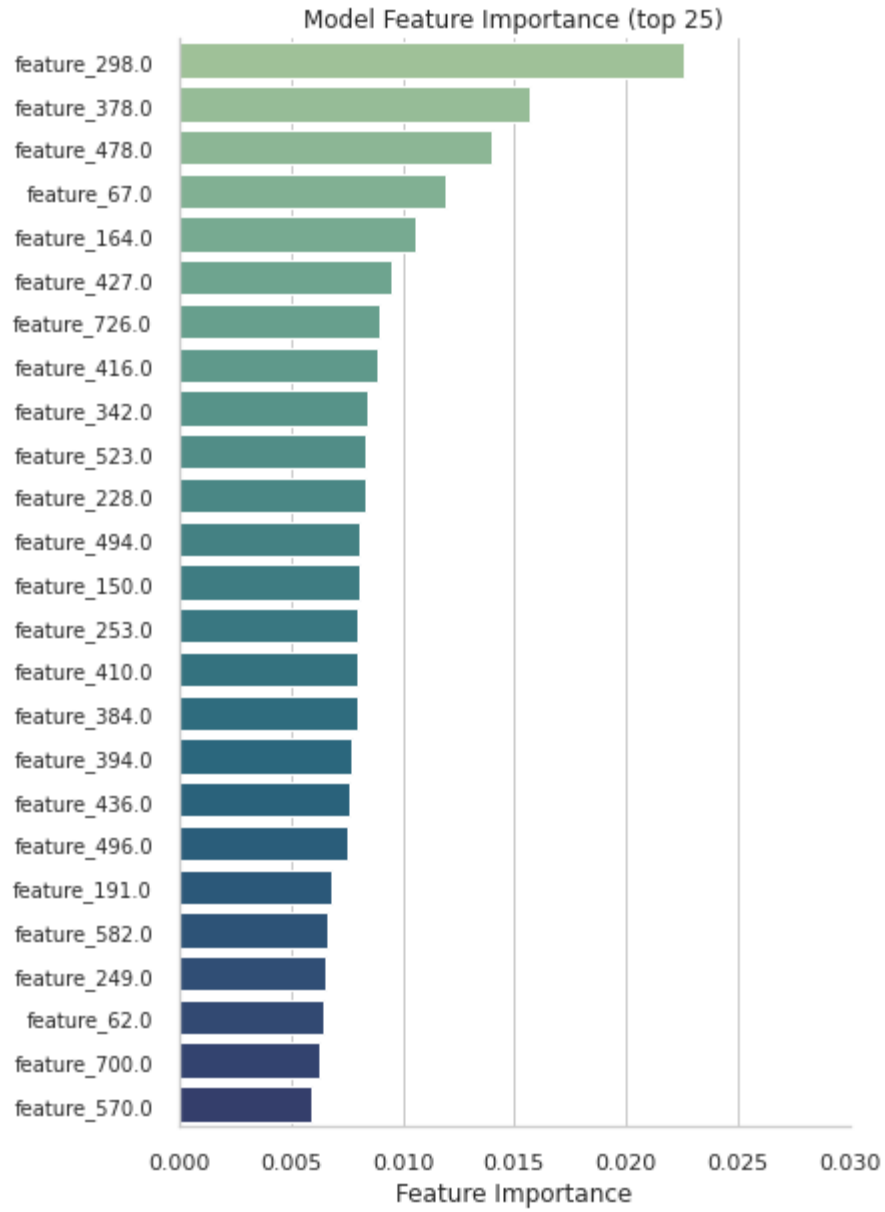
In [36]:
```
importlib.reload(dp)
importlib.reload(dps)
importlib.reload(DataExperiment)
importlib.reload(DataExperimentSupport)
```

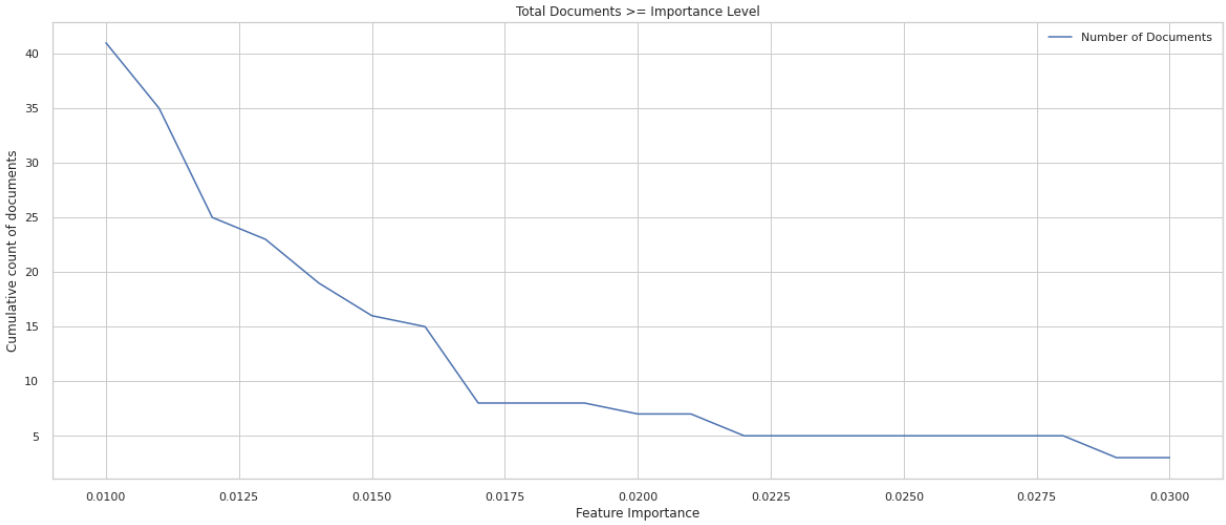
Out[36]:
```
<module 'DataExperimentSupport' from '/home/magni/ML_Root/project_root/utilit
y_files/DataExperimentSupport.py'>
```

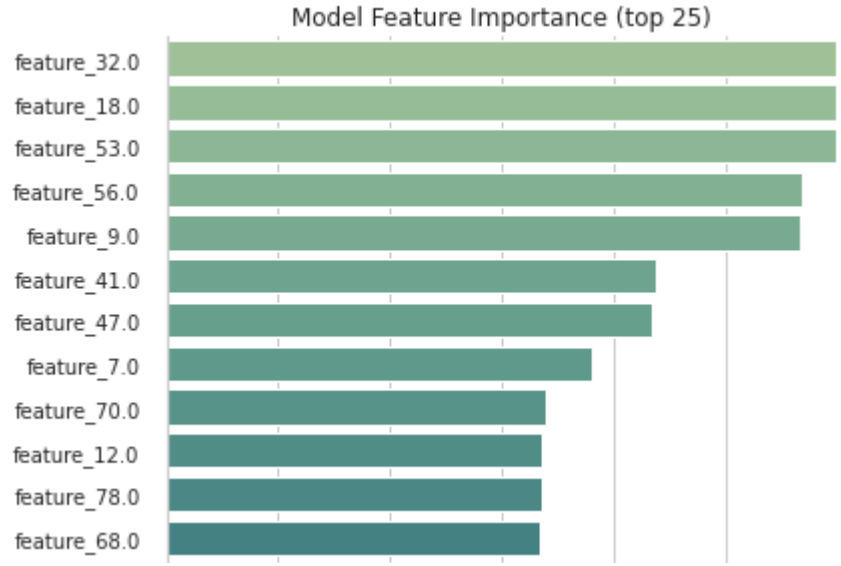
In [34]:
```
myExp.showBaseModelFeatureImportance(upperValue=0.025)
myExp.showFinalModelFeatureImportance(startValue=0.01,
                                      increment=0.001,
                                      upperValue=0.03)
```

```
  0%|          | 0/251 [00:00<?, ?it/s]
  0%|          | 0/22 [00:00<?, ?it/s]
```


Total Documents >= Importance Level
Number of Documents
Cumulative count of documents
800
700
600
500
400
300
200
100
0
0.000
0.005
0.010
0.015
0.020
0.025
Feature Importance


Model Feature Importance (top 25)
feature_298.0
feature_378.0
feature_478.0
feature_67.0
feature_164.0
feature_427.0
feature_726.0
feature_416.0
feature_342.0
feature_523.0
feature_228.0
feature_494.0
feature_150.0
feature_253.0
feature_410.0
feature_384.0
feature_394.0
feature_436.0
feature_496.0
feature_191.0
feature_582.0
feature_249.0
feature_62.0
feature_700.0
feature_570.0
0.000
0.005
0.010
0.015
0.020
0.025
0.030
Feature Importance





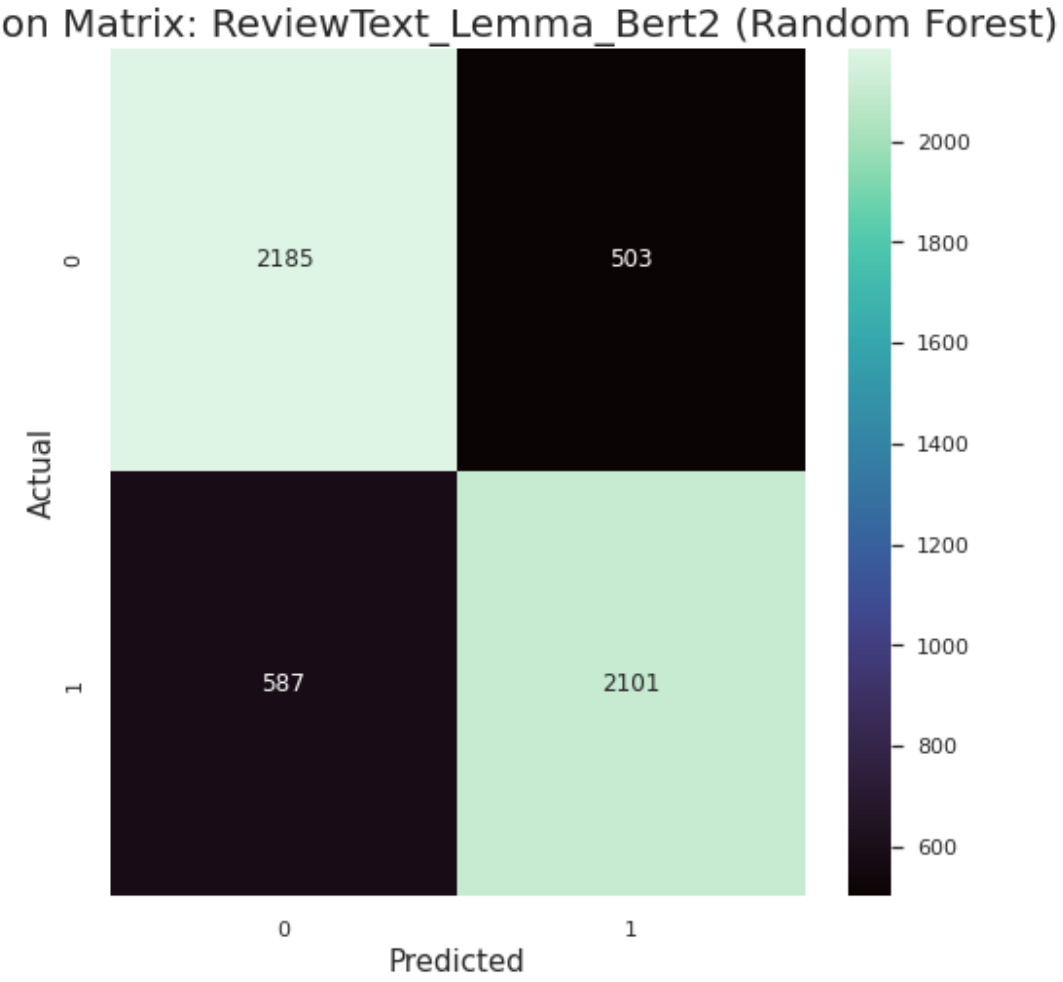
In [18]:
```
myExp.display()
```

```
DataExperiment summary:
---> projectName: ML1010-Group-Project
---> experimentName: ReviewText_Lemma_Bert2 (Random Forest)
---> isDataPackageLoaded: True
---> isBaseModelLoaded: True
---> isBaseModelPredicted: True
---> isBaseModelLearningCurveCreated: True
---> isFinalModelLoaded: True
---> isFinalModelPredicted: True
---> isFinalModelLearningCurveCreated: True
---> isClassifierLoaded: True
RandomForestClassifier()

    DataPackage summary:
    Attributes:
    ---> uniqueColumn: uuid
    ---> targetColumn: overall_posneg
    Process:
    ---> isBalanced: True
    ---> isTrainTestSplit: True
    Data:
```

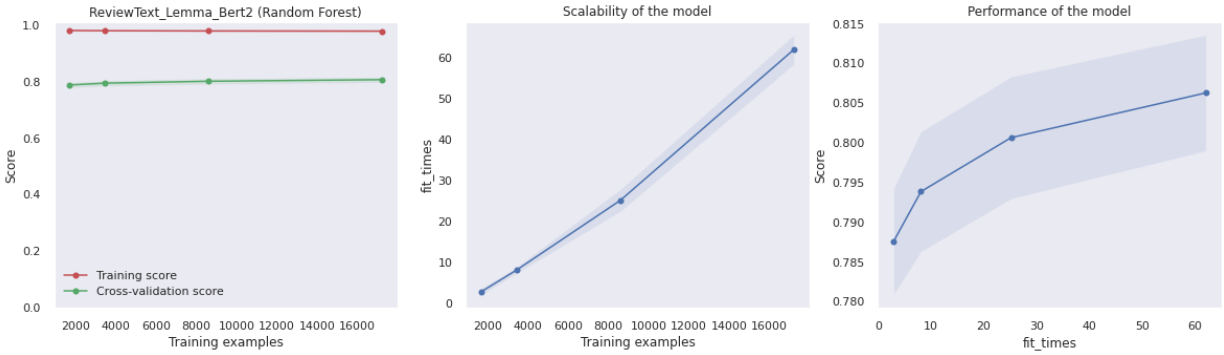
```
    ---> isOrigDataLoaded: False
    ---> isTrainDataLoaded: True
    ---> isTestDataLoaded: True
```

In [38]:
```
myExp.showBaseModelReport(axis_labels)
```

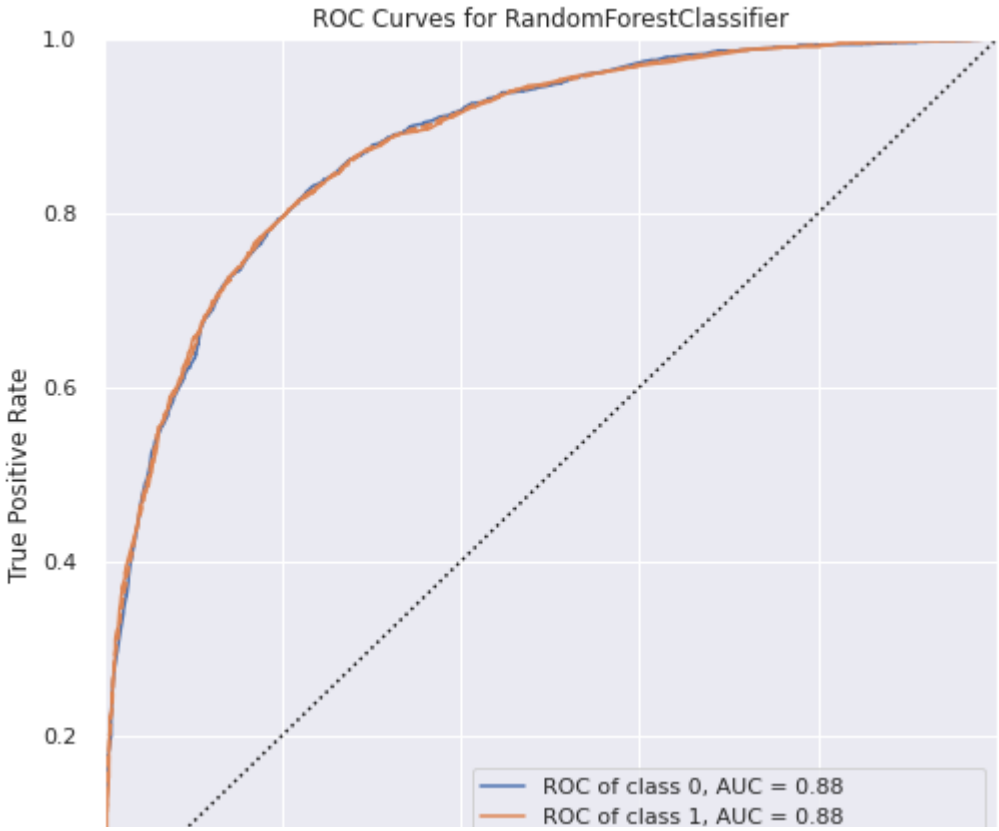
```
Base Model Stats:
Accuracy: 0.8
Precision: 0.8
Recalll: 0.8
F1 Score: 0.8
Cohen kappa:: 0.59
              precision    recall  f1-score   support

           0       0.79      0.81      0.80      2688
           1       0.81      0.78      0.79      2688

    accuracy                           0.80      5376
   macro avg       0.80      0.80      0.80      5376
weighted avg       0.80      0.80      0.80      5376
```




```
<Figure size 576x576 with 0 Axes>
```

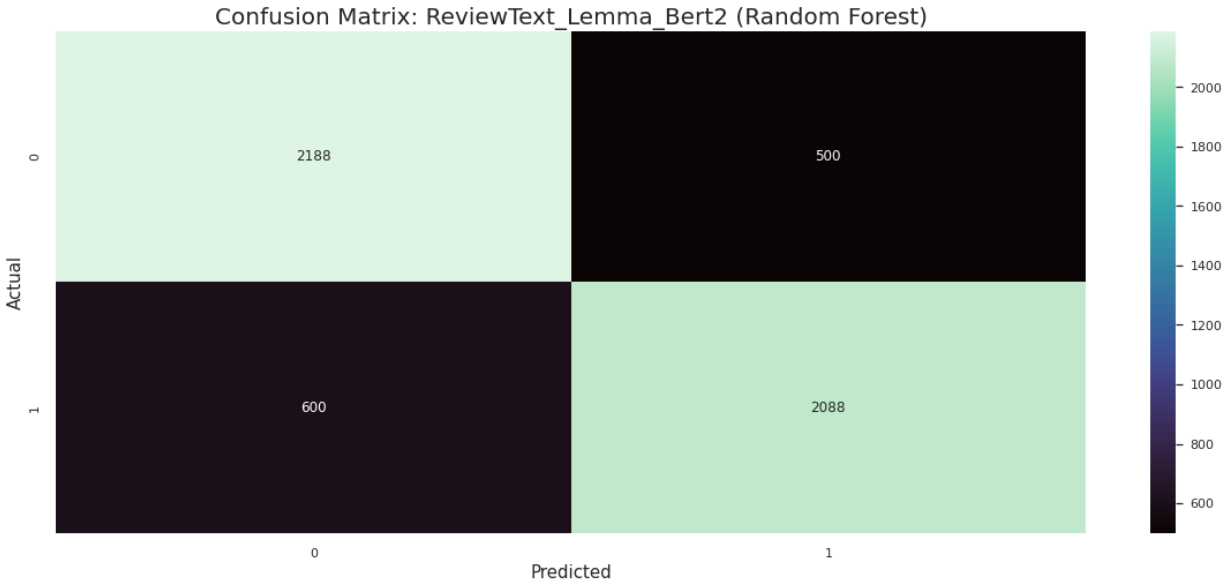


Base model ROCAUC not calculated. Starting now

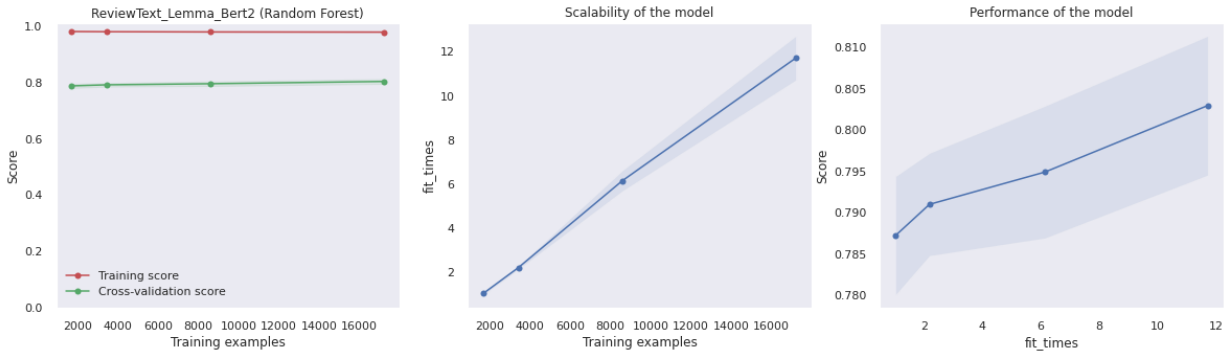


In [20]:
```
myExp.showFinalModelReport(axis_labels)
```

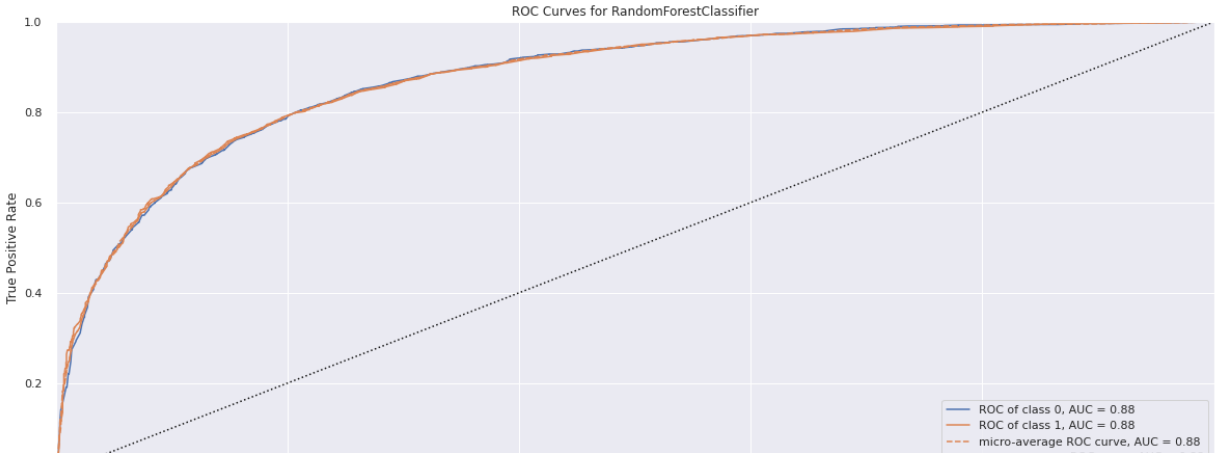
```
Final Model Stats:
Accuracy: 0.8
Precision: 0.8
Recalll: 0.8
F1 Score: 0.8
Cohen kappa:: 0.59
              precision    recall  f1-score   support

           0       0.78      0.81      0.80      2688
           1       0.81      0.78      0.79      2688

    accuracy                           0.80      5376
   macro avg       0.80      0.80      0.80      5376
weighted avg       0.80      0.80      0.80      5376
```



<Figure size 1440x576 with 0 Axes>



Final model ROCAUC not calculated. Starting now



In [21]:
```
myExp.display()
```

```
DataExperiment summary:
---> projectName: ML1010-Group-Project
---> experimentName: ReviewText_Lemma_Bert2 (Random Forest)
---> isDataPackageLoaded: True
---> isBaseModelLoaded: True
---> isBaseModelPredicted: True
---> isBaseModelLearningCurveCreated: True
---> isFinalModelLoaded: True
---> isFinalModelPredicted: True
```

```
---> isFinalModelLearningCurveCreated: True
---> isClassifierLoaded: True
RandomForestClassifier()

    DataPackage summary:
    Attributes:
    ---> uniqueColumn: uuid
    ---> targetColumn: overall_posneg
    Process:
    ---> isBalanced: True
    ---> isTrainTestSplit: True
    Data:
    ---> isOrigDataLoaded: False
    ---> isTrainDataLoaded: True
```

Save Experiment

In [22]:
```
jarvis.saveExperiment(myExp, FILE_NAME)
```

```
[CV] END ..................., score=(train=0.978, test=0.806) total time=  2
9.3s
[CV] END ..................., score=(train=0.978, test=0.799) total time=
6.5s
[CV] END ..................., score=(train=0.979, test=0.807) total time= 1.1
min
[CV] END ..................., score=(train=0.979, test=0.793) total time=
6.7s
[CV] END ..................., score=(train=0.980, test=0.791) total time=
7.1s
[CV] END ..................., score=(train=0.980, test=0.787) total time=
9.5s
[CV] END ..................., score=(train=0.979, test=0.812) total time=  2
0.8s
[CV] END ..................., score=(train=0.979, test=0.789) total time=
1.1s
[CV] END ..................., score=(train=0.980, test=0.790) total time=
6.3s
[CV] END ..................., score=(train=0.980, test=0.790) total time=
2.4s
[CV] END ..................., score=(train=0.980, test=0.796) total time=  2
5.6s
[CV] END ..................., score=(train=0.979, test=0.792) total time=
2.3s
[CV] END ..................., score=(train=0.979, test=0.785) total time=
6.3s
[CV] END ..................., score=(train=0.977, test=0.810) total time= 1.0
min
[CV] END ..................., score=(train=0.980, test=0.793) total time=
2.3s
[CV] END ..................., score=(train=0.980, test=0.782) total time=
2.3s
[CV] END ..................., score=(train=0.979, test=0.808) total time=
5.3s
[CV] END ..................., score=(train=0.979, test=0.789) total time=
3.1s
[CV] END ..................., score=(train=0.982, test=0.776) total time=
```

```
4.3s
[CV] END ....................., score=(train=0.979, test=0.791) total time=  2
5.5s
[CV] END ....................., score=(train=0.977, test=0.810) total time=  1
2.4s
[CV] END ....................., score=(train=0.982, test=0.786) total time=
2.3s
[CV] END ....................., score=(train=0.978, test=0.805) total time= 1.1
min
[CV] END ....................., score=(train=0.979, test=0.800) total time=  1
2.7s
[CV] END ....................., score=(train=0.979, test=0.798) total time=  2
5.5s
[CV] END ....................., score=(train=0.980, test=0.794) total time=
1.2s
[CV] END ....................., score=(train=0.978, test=0.798) total time=  1
1.7s
[CV] END ....................., score=(train=0.979, test=0.797) total time=
8.1s
[CV] END ....................., score=(train=0.979, test=0.793) total time= 1.1
min
[CV] END ....................., score=(train=0.982, test=0.783) total time=
1.2s
[CV] END ....................., score=(train=0.982, test=0.775) total time=
1.0s
[CV] END ....................., score=(train=0.979, test=0.792) total time=  1
2.2s
[CV] END ....................., score=(train=0.980, test=0.787) total time=
8.1s
[CV] END ....................., score=(train=0.980, test=0.797) total time=
3.2s
[CV] END ....................., score=(train=0.980, test=0.807) total time=
8.9s
[CV] END ....................., score=(train=0.978, test=0.815) total time=  5
5.8s
[CV] END ....................., score=(train=0.980, test=0.787) total time=
2.4s
[CV] END ....................., score=(train=0.980, test=0.795) total time=
1.1s
[CV] END ....................., score=(train=0.980, test=0.800) total time=
2.2s
[CV] END ....................., score=(train=0.978, test=0.815) total time=
0.9s
```

Scratchpad

In [ ]: