Sampling&Modelling-SWT

April 2, 2021

[4]:
```
try:
    from collections import OrderedDict
except ImportError:
    from ordereddict import OrderedDict

import pandas as pd

# Array
import numpy as np

# Decompress the file
import gzip

# Visualizations
import matplotlib.pyplot as plt
from matplotlib.colors import ListedColormap
import seaborn as sns
import matplotlib.colors as colors
%matplotlib inline

# Datetime
from datetime import datetime

## Warnings
import warnings
from scipy import stats
warnings.filterwarnings('ignore')

# Large dataset
import dask.bag as db
Cleanreview_df = pd.read_csv('Cleanreview_VideoGames.csv' )
```

[5]:
```
Cleanreview_df.head()
```

[5]:

	Rating	ReviewerID	ProductID	ReviewerName	\
0	5	A1HP7NVNPFMA4N	0700026657	Ambrosia075	
1	4	A1JGAP0185YJI6	0700026657	travis	

```
2       3  A1YJWEXHQBWK2B  0700026657  Vincent G. Mezera
3       2  A2204E1TH211HT  0700026657         Grandma KR
4       5  A2RF5B5H74JLPE  0700026657                jon

                                ProductDescription  Price  \
0  Anno 2070, the newest version of the award-win…  39.99
1  Anno 2070, the newest version of the award-win…  39.99
2  Anno 2070, the newest version of the award-win…  39.99
3  Anno 2070, the newest version of the award-win…  39.99
4  Anno 2070, the newest version of the award-win…  39.99

                        Categories  \
0  [['Video Games', 'PC', 'Games']]
1  [['Video Games', 'PC', 'Games']]
2  [['Video Games', 'PC', 'Games']]
3  [['Video Games', 'PC', 'Games']]
4  [['Video Games', 'PC', 'Games']]

                                          ReviewText RatingClass  ReviewDate  \
0  but when you do it's great. This game is a bit…    positive  2015-10-17
1  But in spite of that it was fun, I liked it I …    positive  2015-07-27
2                              Three Stars ok game.    positive  2015-02-23
3  Two Stars found the game a bit too complicated…    negative  2015-02-20
4  love this game great game, I love it and have …    positive  2014-12-25

                                          CleanText
0               great game bite hard get hang great
1  spite fun like play alright steam bite trouble…
2                                three star ok game
3  two star find game bite complicate not expect …
4       love game great game love play since arrive
```

[6]:
```
Cleanreview_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 497240 entries, 0 to 497239
Data columns (total 11 columns):
 #   Column              Non-Null Count   Dtype
---  ------              --------------   -----
 0   Rating              497240 non-null  int64
 1   ReviewerID          497240 non-null  object
 2   ProductID           497240 non-null  object
 3   ReviewerName        497131 non-null  object
 4   ProductDescription  287372 non-null  object
 5   Price               356582 non-null  float64
 6   Categories          359654 non-null  object
 7   ReviewText          497240 non-null  object
```

```
 8   RatingClass         497240 non-null  object
 9   ReviewDate          497240 non-null  object
 10  CleanText           497187 non-null  object
dtypes: float64(1), int64(1), object(9)
memory usage: 41.7+ MB
```

[7]:
```
#df['date'].dt.year
Cleanreview_df['ReviewDate'] = pd.to_datetime(Cleanreview_df['ReviewDate'])

Cleanreview_df['ReviewYear'] = Cleanreview_df['ReviewDate'].dt.year
Cleanreview_df.head()
```

[7]:

	Rating	ReviewerID	ProductID	ReviewerName	ProductDescription	Price	Categories	ReviewText	RatingClass	ReviewDate	CleanText	ReviewYear
0	5	A1HP7NVNPFMA4N	0700026657	Ambrosia075	Anno 2070, the newest version of the award-win…	39.99	[['Video Games', 'PC', 'Games']]	but when you do it's great. This game is a bit…	positive	2015-10-17	great game bite hard get hang great	2015
1	4	A1JGAP0185YJI6	0700026657	travis	Anno 2070, the newest version of the award-win…	39.99	[['Video Games', 'PC', 'Games']]	But in spite of that it was fun, I liked it I …	positive	2015-07-27	spite fun like play alright steam bite trouble…	2015
2	3	A1YJWEXHQBWK2B	0700026657	Vincent G. Mezera	Anno 2070, the newest version of the award-win…	39.99	[['Video Games', 'PC', 'Games']]	Three Stars ok game.	positive	2015-02-23	three star ok game	2015
3	2	A2204E1TH211HT	0700026657	Grandma KR	Anno 2070, the newest version of the award-win…	39.99	[['Video Games', 'PC', 'Games']]	Two Stars found the game a bit too complicated…	negative	2015-02-20	two star find game bite complicate not expect …	2015
4	5	A2RF5B5H74JLPE	0700026657	jon	Anno 2070, the newest version of the award-win…	39.99	[['Video Games', 'PC', 'Games']]	love this game great game, I love it and have …	positive	2014-12-25	love game great game love play since arrive	2014

[25]:
```
pip install imblearn
```

```
Requirement already satisfied: imblearn in /opt/conda/lib/python3.7/site-
packages (0.0)
Requirement already satisfied: imbalanced-learn in
/opt/conda/lib/python3.7/site-packages (from imblearn) (0.8.0)
Requirement already satisfied: numpy>=1.13.3 in /opt/conda/lib/python3.7/site-
packages (from imbalanced-learn->imblearn) (1.18.4)
Requirement already satisfied: scipy>=0.19.1 in /opt/conda/lib/python3.7/site-
packages (from imbalanced-learn->imblearn) (1.4.1)
Requirement already satisfied: joblib>=0.11 in /opt/conda/lib/python3.7/site-
packages (from imbalanced-learn->imblearn) (0.15.1)
Requirement already satisfied: scikit-learn>=0.24 in
/opt/conda/lib/python3.7/site-packages (from imbalanced-learn->imblearn)
(0.24.1)
Requirement already satisfied: threadpoolctl>=2.0.0 in
/opt/conda/lib/python3.7/site-packages (from scikit-learn>=0.24->imbalanced-
learn->imblearn) (2.1.0)
Note: you may need to restart the kernel to use updated packages.
```

[10]:
```
pip install catboost
```

```
Collecting catboost
  Downloading catboost-0.25-cp37-none-manylinux1_x86_64.whl (67.3 MB)
     |                 | 67.3 MB 93.3 MB/s eta 0:00:01
Requirement already satisfied: scipy in /opt/conda/lib/python3.7/site-
packages (from catboost) (1.4.1)
Collecting graphviz
  Using cached graphviz-0.16-py2.py3-none-any.whl (19 kB)
Requirement already satisfied: pandas>=0.24.0 in /opt/conda/lib/python3.7/site-
packages (from catboost) (1.0.3)
Requirement already satisfied: matplotlib in /opt/conda/lib/python3.7/site-
packages (from catboost) (3.2.1)
Collecting plotly
  Using cached plotly-4.14.3-py2.py3-none-any.whl (13.2 MB)
Requirement already satisfied: six in /opt/conda/lib/python3.7/site-packages
(from catboost) (1.14.0)
Requirement already satisfied: numpy>=1.16.0 in /opt/conda/lib/python3.7/site-
packages (from catboost) (1.18.4)
Requirement already satisfied: python-dateutil>=2.6.1 in
/opt/conda/lib/python3.7/site-packages (from pandas>=0.24.0->catboost) (2.8.1)
Requirement already satisfied: pytz>=2017.2 in /opt/conda/lib/python3.7/site-
packages (from pandas>=0.24.0->catboost) (2020.1)
Requirement already satisfied: cycler>=0.10 in /opt/conda/lib/python3.7/site-
packages (from matplotlib->catboost) (0.10.0)
Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.1 in
/opt/conda/lib/python3.7/site-packages (from matplotlib->catboost) (2.4.7)
Requirement already satisfied: kiwisolver>=1.0.1 in
```

```
/opt/conda/lib/python3.7/site-packages (from matplotlib->catboost) (1.2.0)
Processing ./.cache/pip/wheels/f9/8d/8d/f6af3f7f9eea3553bc2fe6d53e4b287dad18b06a
861ac56ddf/retrying-1.3.3-py3-none-any.whl
Installing collected packages: graphviz, retrying, plotly, catboost
Successfully installed catboost-0.25 graphviz-0.16 plotly-4.14.3 retrying-1.3.3
Note: you may need to restart the kernel to use updated packages.
```

[11]:
```
pip install gensim
```

```
Collecting gensim
  Downloading gensim-4.0.0-cp37-cp37m-manylinux1_x86_64.whl (23.9 MB)
     |                | 23.9 MB 4.7 MB/s eta 0:00:01
Collecting smart-open>=1.8.1
  Downloading smart_open-5.0.0-py3-none-any.whl (56 kB)
     |                | 56 kB 4.4 MB/s  eta 0:00:01
Requirement already satisfied: numpy>=1.11.3 in
/opt/conda/lib/python3.7/site-packages (from gensim) (1.18.4)
Requirement already satisfied: scipy>=0.18.1 in /opt/conda/lib/python3.7/site-
packages (from gensim) (1.4.1)
Installing collected packages: smart-open, gensim
Successfully installed gensim-4.0.0 smart-open-5.0.0
Note: you may need to restart the kernel to use updated packages.
```

[13]:
```
pip install xgboost
```

```
Requirement already satisfied: xgboost in /opt/conda/lib/python3.7/site-packages
(1.3.3)
Requirement already satisfied: numpy in /opt/conda/lib/python3.7/site-packages
(from xgboost) (1.18.4)
Requirement already satisfied: scipy in /opt/conda/lib/python3.7/site-packages
(from xgboost) (1.4.1)
Note: you may need to restart the kernel to use updated packages.
```

[8]:
```
Cleanreview_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 497240 entries, 0 to 497239
Data columns (total 12 columns):
 #   Column              Non-Null Count   Dtype
---  ------              --------------   -----
 0   Rating              497240 non-null  int64
 1   ReviewerID          497240 non-null  object
 2   ProductID           497240 non-null  object
 3   ReviewerName        497131 non-null  object
 4   ProductDescription  287372 non-null  object
 5   Price               356582 non-null  float64
 6   Categories          359654 non-null  object
 7   ReviewText          497240 non-null  object
```

```
 8   RatingClass          497240 non-null  object
 9   ReviewDate           497240 non-null  datetime64[ns]
 10  CleanText            497187 non-null  object
 11  ReviewYear           497240 non-null  int64
dtypes: datetime64[ns](1), float64(1), int64(2), object(8)
memory usage: 45.5+ MB
```

[9]:
```
Cleanreview_df.isnull().sum()

Cleanreview_df=Cleanreview_df.dropna(subset=['CleanText'])
```

[10]:
```
Cleanreview_df.isnull().sum()
```

[10]:
```
Rating                     0
ReviewerID                 0
ProductID                  0
ReviewerName             109
ProductDescription    209843
Price                 140639
Categories            137567
ReviewText                 0
RatingClass                0
ReviewDate                 0
CleanText                  0
ReviewYear                 0
dtype: int64
```

[11]:
```
Cleanreview_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 497187 entries, 0 to 497239
Data columns (total 12 columns):
 #   Column              Non-Null Count   Dtype
---  ------              --------------   -----
 0   Rating              497187 non-null  int64
 1   ReviewerID          497187 non-null  object
 2   ProductID           497187 non-null  object
 3   ReviewerName        497078 non-null  object
 4   ProductDescription  287344 non-null  object
 5   Price               356548 non-null  float64
 6   Categories          359620 non-null  object
 7   ReviewText          497187 non-null  object
 8   RatingClass         497187 non-null  object
 9   ReviewDate          497187 non-null  datetime64[ns]
 10  CleanText           497187 non-null  object
 11  ReviewYear          497187 non-null  int64
dtypes: datetime64[ns](1), float64(1), int64(2), object(8)
memory usage: 49.3+ MB
```

[29]:
```
Cleanreview_df
```

[29]:

	Rating	ReviewerID	ProductID	ReviewerName	ProductDescription	Price	Categories	ReviewText	RatingClass
0	5	A1HP7NVNPFMA4N	0700026657	Ambrosia075	Anno 2070, the newest version of the award-win…	39.99	[['Video Games', 'PC', 'Games']]	but when you do it's great. This game is a bit…	positive
1	4	A1JGAP0185YJI6	0700026657	travis	Anno 2070, the newest version of the award-win…	39.99	[['Video Games', 'PC', 'Games']]	But in spite of that it was fun, I liked it I …	positive
2	3	A1YJWEXHQBWK2B	0700026657	Vincent G. Mezera	Anno 2070, the newest version of the award-win…	39.99	[['Video Games', 'PC', 'Games']]	Three Stars ok game.	positive
3	2	A2204E1TH211HT	0700026657	Grandma KR	Anno 2070, the newest version of the award-win…	39.99	[['Video Games', 'PC', 'Games']]	Two Stars found the game a bit too complicated…	negative
4	5	A2RF5B5H74JLPE	0700026657	jon	Anno 2070, the newest version of the award-win…	39.99	[['Video Games', 'PC', 'Games']]	love this game great game, I love it and have …	positive
…	…	…	…	…	…	…	…	…	…
497235	4	AVECM71LSZLC5	B01HGPUTCA	boris teplitskiy	NaN	NaN	NaN	NaN	NaN
497236	3	A1RS06313BL6WN	B01HH6JEOC	Tom Stopsign	NaN	NaN	NaN	NaN	NaN
497237	3	ACIZ77IGIX2JL	B01HH6JEOC	Era	NaN	NaN	NaN	NaN	NaN
497238	4	A34GG58TJ1A3SH	B01HIZF7XE	seamonkey10	NaN	NaN	NaN	NaN	NaN
497239	2	A6W81WTFK940B	B01HIZGKOE	msam420	NaN	NaN	NaN	NaN	NaN

```
…                                                     …          …
497235      Four Stars not OEM but good replacement parts    positive
497236                           Three Stars Okay stuff.     positive
497237  Only buy on sale. This does add some kids room…     positive
497238  It's Okay, Nothing Profound I think I original…     positive
497239  Not as good as I expected it to be The graphic…     negative

       ReviewDate                                          CleanText  \
0      2015-10-17               great game bite hard get hang great
1      2015-07-27  spite fun like play alright steam bite trouble…
2      2015-02-23                                three star ok game
3      2015-02-20  two star find game bite complicate not expect …
4      2014-12-25       love game great game love play since arrive
…           …                                                    …
497235 2017-07-01       four star not oehem good replacehement part
497236 2018-08-20                            three star okay stuff
497237 2017-08-07  buy sale add kid room things nice right not se…
497238 2018-08-05  okay nothing profound think originally begin p…
497239 2018-03-13  not good expect graphics terrible look like ps…

        ReviewYear
0             2015
1             2015
2             2015
3             2015
4             2014
…              …
497235        2017
497236        2018
497237        2017
497238        2018
497239        2018

[497187 rows x 12 columns]
```

[12]:
```
#sliding window sampling

Cleanreview_df = Cleanreview_df.iloc[::-1]
    # The frame can be made into a time series, a numeric index is preserved
'''
    dataframe['Date'] = pd.to_datetime(dataframe.Date)
    dataframe['Year'] = dataframe.Date.dt.year
    dataframe['Month'] = dataframe.Date.dt.month
'''
Cleanreview_df["Num_Index"] = range(1, 497188)
Cleanreview_df = Cleanreview_df.set_index('ReviewDate')
```

[34]:
```
from imblearn.over_sampling import SMOTE
from collections import Counter
from matplotlib import pyplot
from sklearn.preprocessing import LabelEncoder

from sklearn.model_selection import cross_validate
from sklearn.model_selection import train_test_split
from sklearn.model_selection import StratifiedKFold
from sklearn.model_selection import GridSearchCV
from sklearn.model_selection import cross_val_predict
from sklearn.model_selection import cross_val_score
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import learning_curve
from sklearn.ensemble import ExtraTreesClassifier
from sklearn.decomposition import TruncatedSVD
from sklearn.feature_extraction.text import TfidfVectorizer, CountVectorizer,␣
↪HashingVectorizer
from sklearn.pipeline import Pipeline
from sklearn.naive_bayes import MultinomialNB
from catboost import CatBoostClassifier, Pool
from sklearn.naive_bayes import GaussianNB
from sklearn.svm import SVC
from sklearn import metrics
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix
from  sklearn.metrics import precision_recall_fscore_support
from sklearn.ensemble import RandomForestClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import f1_score
from sklearn.metrics import recall_score
from gensim.models import Word2Vec
from tqdm import tqdm
import xgboost as xgb
from xgboost import XGBClassifier
from sklearn.dummy import DummyClassifier
# define the dataset location

X = Cleanreview_df['CleanText']
y = Cleanreview_df['Rating']


from sklearn.feature_extraction.text import TfidfVectorizer
vec = TfidfVectorizer()
X = vec.fit_transform(X)
```

[37]:
```
X.shape
y.shape
```

[37]: (497187, 308634)

[37]: (497187,)

[ ]:

[32]:
```
pip install yellowbrick
```

```
Collecting yellowbrick
  Using cached yellowbrick-1.3.post1-py3-none-any.whl (271 kB)
Requirement already satisfied: cycler>=0.10.0 in /opt/conda/lib/python3.7/site-
packages (from yellowbrick) (0.10.0)
Requirement already satisfied: scikit-learn>=0.20 in
/opt/conda/lib/python3.7/site-packages (from yellowbrick) (0.24.1)
Requirement already satisfied: matplotlib!=3.0.0,>=2.0.2 in
/opt/conda/lib/python3.7/site-packages (from yellowbrick) (3.2.1)
Requirement already satisfied: numpy<1.20,>=1.16.0 in
/opt/conda/lib/python3.7/site-packages (from yellowbrick) (1.18.4)
Requirement already satisfied: scipy>=1.0.0 in /opt/conda/lib/python3.7/site-
packages (from yellowbrick) (1.4.1)
Requirement already satisfied: six in /opt/conda/lib/python3.7/site-packages
(from cycler>=0.10.0->yellowbrick) (1.14.0)
Requirement already satisfied: joblib>=0.11 in /opt/conda/lib/python3.7/site-
packages (from scikit-learn>=0.20->yellowbrick) (0.15.1)
Requirement already satisfied: threadpoolctl>=2.0.0 in
/opt/conda/lib/python3.7/site-packages (from scikit-learn>=0.20->yellowbrick)
(2.1.0)
Requirement already satisfied: kiwisolver>=1.0.1 in
/opt/conda/lib/python3.7/site-packages (from
matplotlib!=3.0.0,>=2.0.2->yellowbrick) (1.2.0)
Requirement already satisfied: python-dateutil>=2.1 in
/opt/conda/lib/python3.7/site-packages (from
matplotlib!=3.0.0,>=2.0.2->yellowbrick) (2.8.1)
Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.1 in
/opt/conda/lib/python3.7/site-packages (from
matplotlib!=3.0.0,>=2.0.2->yellowbrick) (2.4.7)
Installing collected packages: yellowbrick
Successfully installed yellowbrick-1.3.post1
Note: you may need to restart the kernel to use updated packages.
```

[33]:
```
from bs4 import BeautifulSoup
import requests
import pandas as pd
from pandas import concat
import matplotlib.pyplot as plt
from matplotlib import pyplot
%matplotlib inline
```

```
import seaborn as sns
import numpy as np
from math import sqrt

from sklearn.model_selection import TimeSeriesSplit
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error

from sklearn.model_selection import TimeSeriesSplit
from yellowbrick.regressor import ResidualsPlot, PredictionError
from yellowbrick.model_selection import FeatureImportances
```

[23]:
```
tscv = TimeSeriesSplit(max_train_size=80000, n_splits=10)
```

[34]:
```
tscv = TimeSeriesSplit(n_splits=10,test_size=20000, max_train_size=60000)
```

[44]:
```
print(tscv)
```

```
TimeSeriesSplit(gap=0, max_train_size=None, n_splits=10, test_size=None)
```

[35]:
```
prev_train = 0
trained_on = []

for train_index, test_index in tscv.split(X):

#   An array of indices are created that starts
#   at the finish of the previous training set
#   & ends on the start of the current test set

    last_step = test_index[0]
    a_train_index = np.arange(prev_train, last_step)

#   Train & Test portions are then allocated
   # X_train, X_test = X.iloc[a_train_index], X.iloc[test_index]
   # y_train, y_test = y.iloc[a_train_index], y.iloc[test_index]
    X_train, X_test = X[a_train_index], X[test_index]
    y_train, y_test = y[a_train_index], y[test_index]
```

[36]:
```
X_train.shape
```

[36]:
```
(477187, 308634)
```

[25]:
```
y_train.shape
```

[25]:
```
(451989,)
```

[37]:
```
pip install gensim
```

```
Requirement already satisfied: gensim in /opt/conda/lib/python3.7/site-packages
(4.0.0)
Requirement already satisfied: numpy>=1.11.3 in /opt/conda/lib/python3.7/site-
packages (from gensim) (1.18.4)
Requirement already satisfied: smart-open>=1.8.1 in
/opt/conda/lib/python3.7/site-packages (from gensim) (5.0.0)
Requirement already satisfied: scipy>=0.18.1 in /opt/conda/lib/python3.7/site-
packages (from gensim) (1.4.1)
Note: you may need to restart the kernel to use updated packages.
```

[38]:
```
pip install xgboost
```

```
Requirement already satisfied: xgboost in /opt/conda/lib/python3.7/site-packages
(1.3.3)
Requirement already satisfied: numpy in /opt/conda/lib/python3.7/site-packages
(from xgboost) (1.18.4)
Requirement already satisfied: scipy in /opt/conda/lib/python3.7/site-packages
(from xgboost) (1.4.1)
Note: you may need to restart the kernel to use updated packages.
```

[39]:
```
from imblearn.over_sampling import SMOTE
from collections import Counter
from matplotlib import pyplot
from sklearn.preprocessing import LabelEncoder

from sklearn.model_selection import cross_validate
from sklearn.model_selection import train_test_split
from sklearn.model_selection import StratifiedKFold
from sklearn.model_selection import GridSearchCV
from sklearn.model_selection import cross_val_predict
from sklearn.model_selection import cross_val_score
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import learning_curve
from sklearn.ensemble import ExtraTreesClassifier
from sklearn.decomposition import TruncatedSVD
from sklearn.feature_extraction.text import TfidfVectorizer, CountVectorizer,␣
 ↪HashingVectorizer
from sklearn.pipeline import Pipeline
from sklearn.naive_bayes import MultinomialNB
from catboost import CatBoostClassifier, Pool
from sklearn.naive_bayes import GaussianNB
from sklearn.svm import SVC
from sklearn import metrics
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix
```

```
from  sklearn.metrics import precision_recall_fscore_support
from sklearn.ensemble import RandomForestClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import f1_score
from sklearn.metrics import recall_score
from gensim.models import Word2Vec
from tqdm import tqdm
import xgboost as xgb
from xgboost import XGBClassifier
from sklearn.dummy import DummyClassifier
```

[28]:
```
X_train.shape
```

[28]: (451989, 308634)

[29]:
```
y_train.shape
```

[29]: (451989,)

[24]:
```
X_test.shape
```

[24]: (45198, 308634)

[40]:
```
#Function call for Logistic Regression

logisticRegr = LogisticRegression()

logisticRegr.fit(X_train, y_train)

predictions = logisticRegr.predict(X_test)
```

[41]:
```
print(classification_report(y_test, predictions))
```

```
              precision    recall  f1-score   support

           1       0.54      0.59      0.56       822
           2       0.44      0.21      0.28       822
           3       0.44      0.33      0.38      1762
           4       0.50      0.36      0.42      3792
           5       0.80      0.92      0.86     12802

    accuracy                           0.72     20000
   macro avg       0.55      0.48      0.50     20000
weighted avg       0.69      0.72      0.70     20000
```

[42]:
```
#Modelling using Naive Bayes
# instantiate learning model alpha = optimal_alpha

from sklearn.naive_bayes import MultinomialNB
nb_optimal =  MultinomialNB(alpha = 1.0)

# fitting the model
nb_optimal.fit(X_train, y_train)

# predict the response
```

[42]:
```
MultinomialNB()
```

[43]:
```
predictions = nb_optimal.predict(X_test)
```

[45]:
```
print(classification_report(y_test, predictions))
```

```
              precision    recall  f1-score   support

           1       0.86      0.01      0.01       822
           2       0.00      0.00      0.00       822
           3       0.60      0.00      0.01      1762
           4       0.49      0.05      0.09      3792
           5       0.65      0.99      0.78     12802

    accuracy                           0.65     20000
   macro avg       0.52      0.21      0.18     20000
weighted avg       0.60      0.65      0.52     20000
```

[47]:
```
pip install keras
```

```
Requirement already satisfied: keras in /opt/conda/lib/python3.7/site-packages
(2.4.3)
Requirement already satisfied: numpy>=1.9.1 in /opt/conda/lib/python3.7/site-
packages (from keras) (1.18.4)
Requirement already satisfied: pyyaml in /opt/conda/lib/python3.7/site-packages
(from keras) (5.3.1)
Requirement already satisfied: h5py in /opt/conda/lib/python3.7/site-packages
(from keras) (2.10.0)
Requirement already satisfied: scipy>=0.14 in /opt/conda/lib/python3.7/site-
packages (from keras) (1.4.1)
Requirement already satisfied: six in /opt/conda/lib/python3.7/site-packages
(from h5py->keras) (1.14.0)
Note: you may need to restart the kernel to use updated packages.
```

[2]:
```
pip install tensorflow
```

```
Requirement already satisfied: tensorflow in /opt/conda/lib/python3.7/site-
packages (2.4.1)
Requirement already satisfied: opt-einsum~=3.3.0 in
/opt/conda/lib/python3.7/site-packages (from tensorflow) (3.3.0)
Requirement already satisfied: wheel~=0.35 in /opt/conda/lib/python3.7/site-
packages (from tensorflow) (0.36.2)
Requirement already satisfied: gast==0.3.3 in /opt/conda/lib/python3.7/site-
packages (from tensorflow) (0.3.3)
Requirement already satisfied: h5py~=2.10.0 in /opt/conda/lib/python3.7/site-
packages (from tensorflow) (2.10.0)
Requirement already satisfied: protobuf>=3.9.2 in /opt/conda/lib/python3.7/site-
packages (from tensorflow) (3.11.4)
Requirement already satisfied: absl-py~=0.10 in /opt/conda/lib/python3.7/site-
packages (from tensorflow) (0.12.0)
Requirement already satisfied: flatbuffers~=1.12.0 in
/opt/conda/lib/python3.7/site-packages (from tensorflow) (1.12)
Requirement already satisfied: tensorflow-estimator<2.5.0,>=2.4.0 in
/opt/conda/lib/python3.7/site-packages (from tensorflow) (2.4.0)
Requirement already satisfied: six~=1.15.0 in /opt/conda/lib/python3.7/site-
packages (from tensorflow) (1.15.0)
Requirement already satisfied: tensorboard~=2.4 in
/opt/conda/lib/python3.7/site-packages (from tensorflow) (2.4.1)
Requirement already satisfied: termcolor~=1.1.0 in
/opt/conda/lib/python3.7/site-packages (from tensorflow) (1.1.0)
Requirement already satisfied: numpy~=1.19.2 in /opt/conda/lib/python3.7/site-
packages (from tensorflow) (1.19.5)
Requirement already satisfied: typing-extensions~=3.7.4 in
/opt/conda/lib/python3.7/site-packages (from tensorflow) (3.7.4.2)
Requirement already satisfied: keras-preprocessing~=1.1.2 in
/opt/conda/lib/python3.7/site-packages (from tensorflow) (1.1.2)
Requirement already satisfied: google-pasta~=0.2 in
/opt/conda/lib/python3.7/site-packages (from tensorflow) (0.2.0)
Requirement already satisfied: wrapt~=1.12.1 in /opt/conda/lib/python3.7/site-
packages (from tensorflow) (1.12.1)
Requirement already satisfied: grpcio~=1.32.0 in /opt/conda/lib/python3.7/site-
packages (from tensorflow) (1.32.0)
Requirement already satisfied: astunparse~=1.6.3 in
/opt/conda/lib/python3.7/site-packages (from tensorflow) (1.6.3)
Requirement already satisfied: setuptools in /opt/conda/lib/python3.7/site-
packages (from protobuf>=3.9.2->tensorflow) (46.1.3.post20200325)
Requirement already satisfied: werkzeug>=0.11.15 in
/opt/conda/lib/python3.7/site-packages (from tensorboard~=2.4->tensorflow)
(1.0.1)
Requirement already satisfied: tensorboard-plugin-wit>=1.6.0 in
/opt/conda/lib/python3.7/site-packages (from tensorboard~=2.4->tensorflow)
(1.8.0)
Requirement already satisfied: markdown>=2.6.8 in /opt/conda/lib/python3.7/site-
packages (from tensorboard~=2.4->tensorflow) (3.3.4)
```

```
Requirement already satisfied: google-auth-oauthlib<0.5,>=0.4.1 in
/opt/conda/lib/python3.7/site-packages (from tensorboard~=2.4->tensorflow)
(0.4.4)
Requirement already satisfied: requests<3,>=2.21.0 in
/opt/conda/lib/python3.7/site-packages (from tensorboard~=2.4->tensorflow)
(2.23.0)
Requirement already satisfied: google-auth<2,>=1.6.3 in
/opt/conda/lib/python3.7/site-packages (from tensorboard~=2.4->tensorflow)
(1.16.1)
Requirement already satisfied: importlib-metadata; python_version < "3.8" in
/opt/conda/lib/python3.7/site-packages (from
markdown>=2.6.8->tensorboard~=2.4->tensorflow) (1.6.0)
Requirement already satisfied: requests-oauthlib>=0.7.0 in
/opt/conda/lib/python3.7/site-packages (from google-auth-
oauthlib<0.5,>=0.4.1->tensorboard~=2.4->tensorflow) (1.3.0)
Requirement already satisfied: urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1 in
/opt/conda/lib/python3.7/site-packages (from
requests<3,>=2.21.0->tensorboard~=2.4->tensorflow) (1.25.9)
Requirement already satisfied: idna<3,>=2.5 in /opt/conda/lib/python3.7/site-
packages (from requests<3,>=2.21.0->tensorboard~=2.4->tensorflow) (2.9)
Requirement already satisfied: chardet<4,>=3.0.2 in
/opt/conda/lib/python3.7/site-packages (from
requests<3,>=2.21.0->tensorboard~=2.4->tensorflow) (3.0.4)
Requirement already satisfied: certifi>=2017.4.17 in
/opt/conda/lib/python3.7/site-packages (from
requests<3,>=2.21.0->tensorboard~=2.4->tensorflow) (2020.4.5.2)
Requirement already satisfied: pyasn1-modules>=0.2.1 in
/opt/conda/lib/python3.7/site-packages (from google-
auth<2,>=1.6.3->tensorboard~=2.4->tensorflow) (0.2.8)
Requirement already satisfied: rsa<4.1,>=3.1.4 in /opt/conda/lib/python3.7/site-
packages (from google-auth<2,>=1.6.3->tensorboard~=2.4->tensorflow) (4.0)
Requirement already satisfied: cachetools<5.0,>=2.0.0 in
/opt/conda/lib/python3.7/site-packages (from google-
auth<2,>=1.6.3->tensorboard~=2.4->tensorflow) (4.1.0)
Requirement already satisfied: zipp>=0.5 in /opt/conda/lib/python3.7/site-
packages (from importlib-metadata; python_version <
"3.8"->markdown>=2.6.8->tensorboard~=2.4->tensorflow) (3.1.0)
Requirement already satisfied: oauthlib>=3.0.0 in /opt/conda/lib/python3.7/site-
packages (from requests-oauthlib>=0.7.0->google-auth-
oauthlib<0.5,>=0.4.1->tensorboard~=2.4->tensorflow) (3.0.1)
Requirement already satisfied: pyasn1<0.5.0,>=0.4.6 in
/opt/conda/lib/python3.7/site-packages (from pyasn1-modules>=0.2.1->google-
auth<2,>=1.6.3->tensorboard~=2.4->tensorflow) (0.4.8)
Note: you may need to restart the kernel to use updated packages.
```

[17]:
```
pip install nltk
```

```
Processing ./.cache/pip/wheels/45/6c/46/a1865e7ba706b3817f5d1b2ff7ce8996aabdd0d0
```

```
3d47ba0266/nltk-3.5-py3-none-any.whl
Requirement already satisfied: click in /opt/conda/lib/python3.7/site-packages
(from nltk) (7.1.2)
Requirement already satisfied: joblib in /opt/conda/lib/python3.7/site-packages
(from nltk) (0.15.1)
Collecting regex
  Using cached regex-2021.3.17-cp37-cp37m-manylinux2014_x86_64.whl (721 kB)
Requirement already satisfied: tqdm in /opt/conda/lib/python3.7/site-packages
(from nltk) (4.45.0)
Installing collected packages: regex, nltk
Successfully installed nltk-3.5 regex-2021.3.17
Note: you may need to restart the kernel to use updated packages.
```

[18]:
```
pip install plotly==4.14.3
```

```
Requirement already satisfied: plotly==4.14.3 in /opt/conda/lib/python3.7/site-
packages (4.14.3)
Requirement already satisfied: six in /opt/conda/lib/python3.7/site-packages
(from plotly==4.14.3) (1.15.0)
Requirement already satisfied: retrying>=1.3.3 in /opt/conda/lib/python3.7/site-
packages (from plotly==4.14.3) (1.3.3)
Note: you may need to restart the kernel to use updated packages.
```

[20]:
```
pip install chart_studio
```

```
Requirement already satisfied: chart_studio in /opt/conda/lib/python3.7/site-
packages (1.1.0)
Requirement already satisfied: retrying>=1.3.3 in /opt/conda/lib/python3.7/site-
packages (from chart_studio) (1.3.3)
Requirement already satisfied: plotly in /opt/conda/lib/python3.7/site-packages
(from chart_studio) (4.14.3)
Requirement already satisfied: requests in /opt/conda/lib/python3.7/site-
packages (from chart_studio) (2.23.0)
Requirement already satisfied: six in /opt/conda/lib/python3.7/site-packages
(from chart_studio) (1.15.0)
Requirement already satisfied: idna<3,>=2.5 in /opt/conda/lib/python3.7/site-
packages (from requests->chart_studio) (2.9)
Requirement already satisfied: chardet<4,>=3.0.2 in
/opt/conda/lib/python3.7/site-packages (from requests->chart_studio) (3.0.4)
Requirement already satisfied: certifi>=2017.4.17 in
/opt/conda/lib/python3.7/site-packages (from requests->chart_studio)
(2020.4.5.2)
Requirement already satisfied: urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1 in
/opt/conda/lib/python3.7/site-packages (from requests->chart_studio) (1.25.9)
Note: you may need to restart the kernel to use updated packages.
```

[21]:
```
pip install cufflinks
```

```
Processing ./.cache/pip/wheels/e1/27/13/3fe67fa7ea7be444b831d117220b3b586b872c9a
cd4df480d0/cufflinks-0.17.3-py3-none-any.whl
Requirement already satisfied: ipython>=5.3.0 in /opt/conda/lib/python3.7/site-
packages (from cufflinks) (7.14.0)
Requirement already satisfied: pandas>=0.19.2 in /opt/conda/lib/python3.7/site-
packages (from cufflinks) (1.0.3)
Requirement already satisfied: plotly>=4.1.1 in /opt/conda/lib/python3.7/site-
packages (from cufflinks) (4.14.3)
Requirement already satisfied: six>=1.9.0 in /opt/conda/lib/python3.7/site-
packages (from cufflinks) (1.15.0)
Collecting colorlover>=0.2.1
  Using cached colorlover-0.3.0-py3-none-any.whl (8.9 kB)
Requirement already satisfied: numpy>=1.9.2 in /opt/conda/lib/python3.7/site-
packages (from cufflinks) (1.19.5)
Requirement already satisfied: setuptools>=34.4.1 in
/opt/conda/lib/python3.7/site-packages (from cufflinks) (46.1.3.post20200325)
Requirement already satisfied: ipywidgets>=7.0.0 in
/opt/conda/lib/python3.7/site-packages (from cufflinks) (7.5.1)
Requirement already satisfied: prompt-toolkit!=3.0.0,!=3.0.1,<3.1.0,>=2.0.0 in
/opt/conda/lib/python3.7/site-packages (from ipython>=5.3.0->cufflinks) (3.0.5)
Requirement already satisfied: backcall in /opt/conda/lib/python3.7/site-
packages (from ipython>=5.3.0->cufflinks) (0.1.0)
Requirement already satisfied: pygments in /opt/conda/lib/python3.7/site-
packages (from ipython>=5.3.0->cufflinks) (2.6.1)
Requirement already satisfied: traitlets>=4.2 in /opt/conda/lib/python3.7/site-
packages (from ipython>=5.3.0->cufflinks) (4.3.3)
Requirement already satisfied: jedi>=0.10 in /opt/conda/lib/python3.7/site-
packages (from ipython>=5.3.0->cufflinks) (0.17.0)
Requirement already satisfied: pickleshare in /opt/conda/lib/python3.7/site-
packages (from ipython>=5.3.0->cufflinks) (0.7.5)
Requirement already satisfied: pexpect; sys_platform != "win32" in
/opt/conda/lib/python3.7/site-packages (from ipython>=5.3.0->cufflinks) (4.8.0)
Requirement already satisfied: decorator in /opt/conda/lib/python3.7/site-
packages (from ipython>=5.3.0->cufflinks) (4.4.2)
Requirement already satisfied: python-dateutil>=2.6.1 in
/opt/conda/lib/python3.7/site-packages (from pandas>=0.19.2->cufflinks) (2.8.1)
Requirement already satisfied: pytz>=2017.2 in /opt/conda/lib/python3.7/site-
packages (from pandas>=0.19.2->cufflinks) (2020.1)
Requirement already satisfied: retrying>=1.3.3 in /opt/conda/lib/python3.7/site-
packages (from plotly>=4.1.1->cufflinks) (1.3.3)
Requirement already satisfied: nbformat>=4.2.0 in /opt/conda/lib/python3.7/site-
packages (from ipywidgets>=7.0.0->cufflinks) (5.0.6)
Requirement already satisfied: ipykernel>=4.5.1 in
/opt/conda/lib/python3.7/site-packages (from ipywidgets>=7.0.0->cufflinks)
(5.3.0)
Requirement already satisfied: widgetsnbextension~=3.5.0 in
/opt/conda/lib/python3.7/site-packages (from ipywidgets>=7.0.0->cufflinks)
(3.5.1)
```

```
Requirement already satisfied: wcwidth in /opt/conda/lib/python3.7/site-packages
(from prompt-toolkit!=3.0.0,!=3.0.1,<3.1.0,>=2.0.0->ipython>=5.3.0->cufflinks)
(0.1.9)
Requirement already satisfied: ipython-genutils in
/opt/conda/lib/python3.7/site-packages (from
traitlets>=4.2->ipython>=5.3.0->cufflinks) (0.2.0)
Requirement already satisfied: parso>=0.7.0 in /opt/conda/lib/python3.7/site-
packages (from jedi>=0.10->ipython>=5.3.0->cufflinks) (0.7.0)
Requirement already satisfied: ptyprocess>=0.5 in /opt/conda/lib/python3.7/site-
packages (from pexpect; sys_platform != "win32"->ipython>=5.3.0->cufflinks)
(0.6.0)
Requirement already satisfied: jupyter-core in /opt/conda/lib/python3.7/site-
packages (from nbformat>=4.2.0->ipywidgets>=7.0.0->cufflinks) (4.6.3)
Requirement already satisfied: jsonschema!=2.5.0,>=2.4 in
/opt/conda/lib/python3.7/site-packages (from
nbformat>=4.2.0->ipywidgets>=7.0.0->cufflinks) (3.2.0)
Requirement already satisfied: tornado>=4.2 in /opt/conda/lib/python3.7/site-
packages (from ipykernel>=4.5.1->ipywidgets>=7.0.0->cufflinks) (6.0.4)
Requirement already satisfied: jupyter-client in /opt/conda/lib/python3.7/site-
packages (from ipykernel>=4.5.1->ipywidgets>=7.0.0->cufflinks) (6.1.3)
Requirement already satisfied: notebook>=4.4.1 in /opt/conda/lib/python3.7/site-
packages (from widgetsnbextension~=3.5.0->ipywidgets>=7.0.0->cufflinks) (6.0.3)
Requirement already satisfied: pyrsistent>=0.14.0 in
/opt/conda/lib/python3.7/site-packages (from
jsonschema!=2.5.0,>=2.4->nbformat>=4.2.0->ipywidgets>=7.0.0->cufflinks) (0.16.0)
Requirement already satisfied: attrs>=17.4.0 in /opt/conda/lib/python3.7/site-
packages (from
jsonschema!=2.5.0,>=2.4->nbformat>=4.2.0->ipywidgets>=7.0.0->cufflinks) (19.3.0)
Requirement already satisfied: importlib-metadata; python_version < "3.8" in
/opt/conda/lib/python3.7/site-packages (from
jsonschema!=2.5.0,>=2.4->nbformat>=4.2.0->ipywidgets>=7.0.0->cufflinks) (1.6.0)
Requirement already satisfied: pyzmq>=13 in /opt/conda/lib/python3.7/site-
packages (from jupyter-client->ipykernel>=4.5.1->ipywidgets>=7.0.0->cufflinks)
(19.0.1)
Requirement already satisfied: prometheus-client in
/opt/conda/lib/python3.7/site-packages (from
notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidgets>=7.0.0->cufflinks)
(0.8.0)
Requirement already satisfied: jinja2 in /opt/conda/lib/python3.7/site-packages
(from notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidgets>=7.0.0->cufflinks)
(2.11.2)
Requirement already satisfied: Send2Trash in /opt/conda/lib/python3.7/site-
packages (from
notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidgets>=7.0.0->cufflinks)
(1.5.0)
Requirement already satisfied: nbconvert in /opt/conda/lib/python3.7/site-
packages (from
notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidgets>=7.0.0->cufflinks)
```

```
(5.6.1)
Requirement already satisfied: terminado>=0.8.1 in
/opt/conda/lib/python3.7/site-packages (from
notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidgets>=7.0.0->cufflinks)
(0.8.3)
Requirement already satisfied: zipp>=0.5 in /opt/conda/lib/python3.7/site-
packages (from importlib-metadata; python_version <
"3.8"->jsonschema!=2.5.0,>=2.4->nbformat>=4.2.0->ipywidgets>=7.0.0->cufflinks)
(3.1.0)
Requirement already satisfied: MarkupSafe>=0.23 in
/opt/conda/lib/python3.7/site-packages (from jinja2->notebook>=4.4.1->widgetsnbe
xtension~=3.5.0->ipywidgets>=7.0.0->cufflinks) (1.1.1)
Requirement already satisfied: defusedxml in /opt/conda/lib/python3.7/site-
packages (from nbconvert->notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidgets
>=7.0.0->cufflinks) (0.6.0)
Requirement already satisfied: testpath in /opt/conda/lib/python3.7/site-
packages (from nbconvert->notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidgets
>=7.0.0->cufflinks) (0.4.4)
Requirement already satisfied: bleach in /opt/conda/lib/python3.7/site-packages
(from nbconvert->notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidgets>=7.0.0->
cufflinks) (3.1.5)
Requirement already satisfied: pandocfilters>=1.4.1 in
/opt/conda/lib/python3.7/site-packages (from nbconvert->notebook>=4.4.1->widgets
nbextension~=3.5.0->ipywidgets>=7.0.0->cufflinks) (1.4.2)
Requirement already satisfied: mistune<2,>=0.8.1 in
/opt/conda/lib/python3.7/site-packages (from nbconvert->notebook>=4.4.1->widgets
nbextension~=3.5.0->ipywidgets>=7.0.0->cufflinks) (0.8.4)
Requirement already satisfied: entrypoints>=0.2.2 in
/opt/conda/lib/python3.7/site-packages (from nbconvert->notebook>=4.4.1->widgets
nbextension~=3.5.0->ipywidgets>=7.0.0->cufflinks) (0.3)
Requirement already satisfied: webencodings in /opt/conda/lib/python3.7/site-
packages (from bleach->nbconvert->notebook>=4.4.1->widgetsnbextension~=3.5.0->ip
ywidgets>=7.0.0->cufflinks) (0.5.1)
Requirement already satisfied: packaging in /opt/conda/lib/python3.7/site-
packages (from bleach->nbconvert->notebook>=4.4.1->widgetsnbextension~=3.5.0->ip
ywidgets>=7.0.0->cufflinks) (20.4)
Requirement already satisfied: pyparsing>=2.0.2 in
/opt/conda/lib/python3.7/site-packages (from packaging->bleach->nbconvert->noteb
ook>=4.4.1->widgetsnbextension~=3.5.0->ipywidgets>=7.0.0->cufflinks) (2.4.7)
Installing collected packages: colorlover, cufflinks
Successfully installed colorlover-0.3.0 cufflinks-0.17.3
Note: you may need to restart the kernel to use updated packages.
```

[22]:
```
#LSTM modelling
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

```
import seaborn as sns
from keras.preprocessing.text import Tokenizer
from keras.preprocessing.sequence import pad_sequences
from keras.models import Sequential
from keras.layers import Dense, Embedding, LSTM, SpatialDropout1D
from sklearn.model_selection import train_test_split
from keras.utils.np_utils import to_categorical
from keras.callbacks import EarlyStopping
from keras.layers import Dropout
import re
from nltk.corpus import stopwords
from nltk import word_tokenize
STOPWORDS = set(stopwords.words('english'))
from bs4 import BeautifulSoup
import plotly.graph_objs as go
#import plotly.plotly as py
import chart_studio.plotly as py
import cufflinks
from IPython.core.interactiveshell import InteractiveShell
import plotly.figure_factory as ff
InteractiveShell.ast_node_interactivity = 'all'
from plotly.offline import iplot
cufflinks.go_offline()
cufflinks.set_config_file(world_readable=True, theme='pearl')
```

[23]:
```
# The maximum number of words to be used. (most frequent)

'''
X = Cleanreview_df['CleanText']
y = Cleanreview_df['Rating']
'''
MAX_NB_WORDS = 50000
# Max number of words in each Review.
MAX_SEQUENCE_LENGTH = 250
# This is fixed.
EMBEDDING_DIM = 100
#, lower=True
tokenizer = Tokenizer(num_words=MAX_NB_WORDS, filters='!"#$%&()*+,-./:;<=>?
 ↪@[\]^_`{|}~', lower=True
                      )
tokenizer.fit_on_texts(Cleanreview_df['CleanText'].values)
word_index = tokenizer.word_index
print('Found %s unique tokens.' % len(word_index))
```

[23]:
```
"\nX = Cleanreview_df['CleanText']\ny = Cleanreview_df['Rating']\n"
```

```
Found 308309 unique tokens.
```

[24]:
```
X = tokenizer.texts_to_sequences(Cleanreview_df['CleanText'].values)
X = pad_sequences(X, maxlen=MAX_SEQUENCE_LENGTH)
print('Shape of data tensor:', X.shape)
```

```
Shape of data tensor: (497187, 250)
```

[26]:
```
Y = pd.get_dummies(Cleanreview_df['Rating']).values
print('Shape of label tensor:', Y.shape)
```

```
Shape of label tensor: (497187, 5)
```

[13]:
```
from sklearn.model_selection import TimeSeriesSplit
from sklearn.utils import indexable
from sklearn.utils.validation import _num_samples

class TimeSeriesSplitCustom(TimeSeriesSplit):
    def __init__(self, n_splits=5, max_train_size=None,
                 test_size=1,
                 min_train_size=1):
        super().__init__(n_splits=n_splits, max_train_size=max_train_size)
        self.test_size = test_size
        self.min_train_size = min_train_size

    def overlapping_split(self, X, y=None, groups=None):
        min_train_size = self.min_train_size
        test_size = self.test_size

        n_splits = self.n_splits
        n_samples = _num_samples(X)

        if (n_samples - min_train_size) / test_size >= n_splits:
            print('(n_samples -  min_train_size) / test_size >= n_splits')
            print('default TimeSeriesSplit.split() used')
            yield from super().split(X)

        else:
            shift = int(np.floor(
                (n_samples - test_size - min_train_size) / (n_splits - 1)))

            start_test = n_samples - (n_splits * shift + test_size - shift)

            test_starts = range(start_test, n_samples - test_size + 1, shift)

            if start_test < min_train_size:
                raise ValueError(
                    ("The start of the testing : {0} is smaller"
```

```
                    " than the minimum training samples: {1}.").
↪format(start_test,
                                                                  ␣
↪min_train_size))

            indices = np.arange(n_samples)

            for test_start in test_starts:
                if self.max_train_size and self.max_train_size < test_start:
                    yield (indices[test_start - self.max_train_size:test_start],
                           indices[test_start:test_start + test_size])
                else:
                    yield (indices[:test_start],
                           indices[test_start:test_start + test_size])
```

[15]:
```
tscv=TimeSeriesSplitCustom(n_splits=10, test_size=20000, min_train_size=80000)
```

[16]:
```
prev_train = 0
trained_on = []

for train_index, test_index in tscv.split(X):

#   An array of indices are created that starts
#   at the finish of the previous training set
#   & ends on the start of the current test set

    last_step = test_index[0]
    a_train_index = np.arange(prev_train, last_step)

#   Train & Test portions are then allocated
   # X_train, X_test = X.iloc[a_train_index], X.iloc[test_index]
   # y_train, y_test = y.iloc[a_train_index], y.iloc[test_index]
    X_train, X_test = X[a_train_index], X[test_index]
    y_train, y_test = Y[a_train_index], Y[test_index]
```

[17]:
```
print(X_train.shape,y_train.shape)
print(X_test.shape,y_test.shape)
```

```
(477187, 250) (477187, 5)
(20000, 250) (20000, 5)
```

[18]:
```
x_train, x_val, y_train, y_val = train_test_split(X_train,␣
↪y_train,train_size=90000,
                                                  test_size = 30000,
                                                  random_state=12)
```

[19]:
```
print(x_train.shape,y_train.shape)
print(x_val.shape,y_val.shape)
```

```
(90000, 250) (90000, 5)
(30000, 250) (30000, 5)
```

[47]:
```
model = Sequential()
model.add(Embedding(MAX_NB_WORDS, EMBEDDING_DIM, input_length=X.shape[1]))
model.add(SpatialDropout1D(0.2))
model.add(LSTM(100, dropout=0.2, recurrent_dropout=0.2))
model.add(Dense(5, activation='softmax'))
model.compile(loss='categorical_crossentropy', optimizer='adam',␣
 ↪metrics=['accuracy'])
print(model.summary())
```

```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_1 (Embedding)      (None, 308634, 100)       5000000
_________________________________________________________________
spatial_dropout1d_1 (Spatial (None, 308634, 100)       0
_________________________________________________________________
lstm_1 (LSTM)                (None, 100)               80400
_________________________________________________________________
dense_1 (Dense)              (None, 5)                 505
=================================================================
Total params: 5,080,905
Trainable params: 5,080,905
Non-trainable params: 0
_________________________________________________________________
None
```

[28]:
```
epochs = 5
batch_size = 64

history = model.fit(x_train, y_train, epochs=epochs,␣
 ↪batch_size=batch_size,validation_split=0.
 ↪1,callbacks=[EarlyStopping(monitor='val_loss', patience=3, min_delta=0.
 ↪0001)])
```

```
      ␣
↪---------------------------------------------------------------------------

       NameError                                 Traceback (most recent call␣
↪last)
```

```
        <ipython-input-28-050c99987b68> in <module>
          2 batch_size = 64
          3
    ----> 4 history = model.fit(x_train, y_train, epochs=epochs,␣
↪batch_size=batch_size,validation_split=0.
↪1,callbacks=[EarlyStopping(monitor='val_loss', patience=3, min_delta=0.0001)])


        NameError: name 'x_train' is not defined
```

[27]:
```
from sklearn.metrics import classification_report

# predict
pred = model.predict(X_test, batch_size = 32)
#pred = np.argmax(predictions, axis=1)
# label
y_train = np.argmax(y_test, axis=1)

print(y_train.shape, pred.shape)
print(y_train[:5], pred[:5])
```

```
(20000,) (20000, 5)
[3 4 4 4 3] [[1.48051313e-05 7.55738120e-06 1.80945048e-04 9.97719586e-01
  2.07707332e-03]
 [2.84047492e-05 5.74448995e-06 1.05943036e-04 3.43446329e-04
  9.99516487e-01]
 [1.61807242e-04 4.79152959e-06 7.49792307e-05 9.00321538e-05
  9.99668360e-01]
 [4.39545140e-03 8.43503326e-03 4.39227298e-02 1.48705587e-01
  7.94541121e-01]
 [1.43229379e-04 5.61881345e-04 2.70361025e-02 7.20617652e-01
  2.51641035e-01]]
```

[28]:
```
print(classification_report(y_train, np.argmax(pred, axis = 1)))
```

```
              precision    recall  f1-score   support

           0       0.52      0.51      0.51       822
           1       0.31      0.24      0.27       822
           2       0.38      0.38      0.38      1762
           3       0.43      0.42      0.43      3792
           4       0.83      0.85      0.84     12802

    accuracy                           0.69     20000
   macro avg       0.49      0.48      0.49     20000
weighted avg       0.68      0.69      0.69     20000
```

[29]:
```
from sklearn.metrics import classification_report

# predict
pred = model.predict(x_val, batch_size = 32)
#pred = np.argmax(predictions, axis=1)
# label
y_val = np.argmax(y_val, axis=1)

print(y_val.shape, pred.shape)
print(y_val[:5], pred[:5])
```

```
(30000,) (30000, 5)
[3 3 0 4 3] [[9.5525039e-03 5.6797128e-02 4.2761543e-01 3.7993366e-01
1.2610130e-01]
 [3.0915282e-04 3.0044079e-04 4.4510472e-03 3.4614924e-01 6.4879006e-01]
 [9.3707633e-01 1.3723333e-03 1.7779259e-03 1.2419300e-02 4.7354162e-02]
 [9.3280029e-04 6.8879209e-04 4.7580800e-03 2.3685062e-01 7.5676978e-01]
 [2.4035696e-03 1.2138725e-02 4.2486575e-02 7.6753688e-01 1.7543422e-01]]
```

[30]:
```
print(classification_report(y_val, np.argmax(pred, axis = 1)))
```

```
              precision    recall  f1-score   support

           0       0.61      0.60      0.61      1812
           1       0.41      0.33      0.37      1520
           2       0.48      0.46      0.47      3031
           3       0.52      0.48      0.50      5586
           4       0.84      0.88      0.86     18051

    accuracy                           0.72     30000
   macro avg       0.57      0.55      0.56     30000
weighted avg       0.71      0.72      0.71     30000
```

[30]:
```
pip install mlxtend
```

```
Collecting mlxtend
  Downloading mlxtend-0.18.0-py2.py3-none-any.whl (1.3 MB)
     |                  | 1.3 MB 5.6 MB/s eta 0:00:01
Requirement already satisfied: pandas>=0.24.2 in
/opt/conda/lib/python3.7/site-packages (from mlxtend) (1.0.3)
Requirement already satisfied: joblib>=0.13.2 in /opt/conda/lib/python3.7/site-
packages (from mlxtend) (0.15.1)
Requirement already satisfied: scikit-learn>=0.20.3 in
/opt/conda/lib/python3.7/site-packages (from mlxtend) (0.24.1)
Requirement already satisfied: numpy>=1.16.2 in /opt/conda/lib/python3.7/site-
packages (from mlxtend) (1.19.5)
Requirement already satisfied: setuptools in /opt/conda/lib/python3.7/site-
```

```
packages (from mlxtend) (46.1.3.post20200325)
Requirement already satisfied: scipy>=1.2.1 in /opt/conda/lib/python3.7/site-
packages (from mlxtend) (1.4.1)
Requirement already satisfied: matplotlib>=3.0.0 in
/opt/conda/lib/python3.7/site-packages (from mlxtend) (3.2.1)
Requirement already satisfied: python-dateutil>=2.6.1 in
/opt/conda/lib/python3.7/site-packages (from pandas>=0.24.2->mlxtend) (2.8.1)
Requirement already satisfied: pytz>=2017.2 in /opt/conda/lib/python3.7/site-
packages (from pandas>=0.24.2->mlxtend) (2020.1)
Requirement already satisfied: threadpoolctl>=2.0.0 in
/opt/conda/lib/python3.7/site-packages (from scikit-learn>=0.20.3->mlxtend)
(2.1.0)
Requirement already satisfied: kiwisolver>=1.0.1 in
/opt/conda/lib/python3.7/site-packages (from matplotlib>=3.0.0->mlxtend) (1.2.0)
Requirement already satisfied: cycler>=0.10 in /opt/conda/lib/python3.7/site-
packages (from matplotlib>=3.0.0->mlxtend) (0.10.0)
Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.1 in
/opt/conda/lib/python3.7/site-packages (from matplotlib>=3.0.0->mlxtend) (2.4.7)
Requirement already satisfied: six>=1.5 in /opt/conda/lib/python3.7/site-
packages (from python-dateutil>=2.6.1->pandas>=0.24.2->mlxtend) (1.15.0)
Installing collected packages: mlxtend
Successfully installed mlxtend-0.18.0
Note: you may need to restart the kernel to use updated packages.
```

[41]:
```
x_train, x_val, y_train, y_val = train_test_split(X, y,train_size=1000,
                                                  test_size = 300,
                                                  random_state=12)
```

[43]:
```
x_train.shape
```

[43]:
```
(1000, 308634)
```

[46]:
```
#Hypothetical tests for comparing 2 algorithms

# use 5x2 statistical hypothesis testing procedure to compare two machine␣
 ↪learning algorithms
from numpy import mean
from numpy import std
from sklearn.datasets import make_classification
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import RepeatedStratifiedKFold
from sklearn.linear_model import LogisticRegression
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from mlxtend.evaluate import paired_ttest_5x2cv
# define dataset
#X, y = make_classification(n_samples=1000, n_features=10, n_informative=10,␣
 ↪n_redundant=0, random_state=1)
```

```
# evaluate model 1
#X=np.argmax(X, axis=1)
#Y = np.argmax(Y, axis=1)
model1 = LogisticRegression()
cv1 = RepeatedStratifiedKFold(n_splits=10, n_repeats=3, random_state=1)
scores1 = cross_val_score(model1, x_train, y_train, scoring='accuracy', cv=cv1,␣
↪n_jobs=1)
print('LogisticRegression Mean Accuracy: %.3f (%.3f)' % (mean(scores1),␣
↪std(scores1)))
# evaluate model 2
#nb_optimal =  MultinomialNB(alpha = 1.0)

model2 = MultinomialNB(alpha = 1.0)
cv2 = RepeatedStratifiedKFold(n_splits=10, n_repeats=3, random_state=1)
scores2 = cross_val_score(model2, x_train, y_train, scoring='accuracy', cv=cv2,␣
↪n_jobs=1)
print('Naive Bayes  Mean Accuracy: %.3f (%.3f)' % (mean(scores2), std(scores2)))
# check if difference between algorithms is real
t, p = paired_ttest_5x2cv(estimator1=model1, estimator2=model2, X=x_train,␣
↪y=y_train, scoring='accuracy', random_seed=1)
# summarize
print('P-value: %.3f, t-Statistic: %.3f' % (p, t))
# interpret the result
if p <= 0.05:
        print('Difference between mean performance is probably real')
else:
        print('Algorithms probably have the same performance')
```

```
LogisticRegression Mean Accuracy: 0.640 (0.018)
LinearDiscriminantAnalysis Mean Accuracy: 0.605 (0.005)
P-value: 0.041, t-Statistic: 2.744
Difference between mean performance is probably real
```

[ ]: