Table of Contents

In [ ]:

In [235]:
```
%matplotlib inline
```

In [86]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import OneHotEncoder, StandardScaler
from sklearn import metrics, model_selection
from statistics import mean, median
from time import perf_counter
from sklearn.pipeline import Pipeline
from imblearn.over_sampling import SMOTE
from sklearn.model_selection import GridSearchCV, StratifiedKFold, cross_val_p
redict
from sklearn.metrics import make_scorer
from sklearn.ensemble import IsolationForest
from sklearn.feature_selection import SelectPercentile, f_classif
from sklearn.neighbors import KNeighborsClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.svm import SVC
from sklearn.ensemble import RandomForestClassifier, AdaBoostClassifier, Gradi
entBoostingClassifier, BaggingClassifier
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from sklearn.linear_model import LogisticRegression, PassiveAggressiveClassifi
er, Perceptron, RidgeClassifier, SGDClassifier
from sklearn.neural_network import MLPClassifier
from sklearn.discriminant_analysis import QuadraticDiscriminantAnalysis
from sklearn.gaussian_process import GaussianProcessClassifier
from sklearn.neighbors import NearestCentroid
import imblearn
from imblearn.over_sampling import SMOTE, ADASYN, BorderlineSMOTE, SMOTENC, KM
eansSMOTE, SVMSMOTE
from imblearn.combine import SMOTEENN, SMOTETomek
from sklearn.metrics import make_scorer
from sklearn.ensemble import RandomForestClassifier
import random
from imblearn.ensemble import BalancedBaggingClassifier, BalancedRandomForestC
lassifier
from sklearn.preprocessing import MinMaxScaler
from sklearn.decomposition import PCA
from sklearn.tree import DecisionTreeClassifier
from sklearn.svm import SVC
from sklearn import model_selection
from sklearn import linear_model
from sklearn.feature_selection import RFECV
from sklearn.feature_selection import SelectFromModel
from imblearn.under_sampling import RandomUnderSampler
from imblearn.over_sampling import RandomOverSampler
from sklearn.svm import LinearSVC
from sklearn.feature_selection import SelectFwe, chi2, f_classif
from sklearn.feature_selection import VarianceThreshold
from sklearn.ensemble import ExtraTreesClassifier
```

In [236]:
```
plt.rcParams['figure.figsize'] = [12, 8]
```

Read the file

In [216]:
```
def read_file()->pd.DataFrame:
    df = pd.read_csv('Churn_Modelling.csv')
    df.describe()
    return df
df = read_file()
```

Baseline

Explore the data

In [217]:
```
df.head()
```

Out[217]:

	RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance
0	1	15634602	Hargrave	619	France	Female	42	2	0.00
1	2	15647311	Hill	608	Spain	Female	41	1	83807.86
2	3	15619304	Onio	502	France	Female	42	8	159660.80
3	4	15701354	Boni	699	France	Female	39	1	0.00
4	5	15737888	Mitchell	850	Spain	Female	43	2	125510.82

In [218]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 14 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   RowNumber        10000 non-null  int64
 1   CustomerId       10000 non-null  int64
 2   Surname          10000 non-null  object
 3   CreditScore      10000 non-null  int64
 4   Geography        10000 non-null  object
 5   Gender           10000 non-null  object
 6   Age              10000 non-null  int64
 7   Tenure           10000 non-null  int64
 8   Balance          10000 non-null  float64
 9   NumOfProducts    10000 non-null  int64
 10  HasCrCard        10000 non-null  int64
 11  IsActiveMember   10000 non-null  int64
 12  EstimatedSalary  10000 non-null  float64
 13  Exited           10000 non-null  int64
dtypes: float64(2), int64(9), object(3)
memory usage: 1.1+ MB
```

In [219]:
```
df.describe().T
```

Out[219]:

	count	mean	std	min	25%	50%
RowNumber	10000.0	5.000500e+03	2886.895680	1.00	2500.75	5.000500e+03
CustomerId	10000.0	1.569094e+07	71936.186123	15565701.00	15628528.25	1.569074e+07
CreditScore	10000.0	6.505288e+02	96.653299	350.00	584.00	6.520000e+02
Age	10000.0	3.892180e+01	10.487806	18.00	32.00	3.700000e+01
Tenure	10000.0	5.012800e+00	2.892174	0.00	3.00	5.000000e+00
Balance	10000.0	7.648589e+04	62397.405202	0.00	0.00	9.719854e+04
NumOfProducts	10000.0	1.530200e+00	0.581654	1.00	1.00	1.000000e+00
HasCrCard	10000.0	7.055000e-01	0.455840	0.00	0.00	1.000000e+00
IsActiveMember	10000.0	5.151000e-01	0.499797	0.00	0.00	1.000000e+00
EstimatedSalary	10000.0	1.000902e+05	57510.492818	11.58	51002.11	1.001939e+05
Exited	10000.0	2.037000e-01	0.402769	0.00	0.00	0.000000e+00

Check if there are any missing values

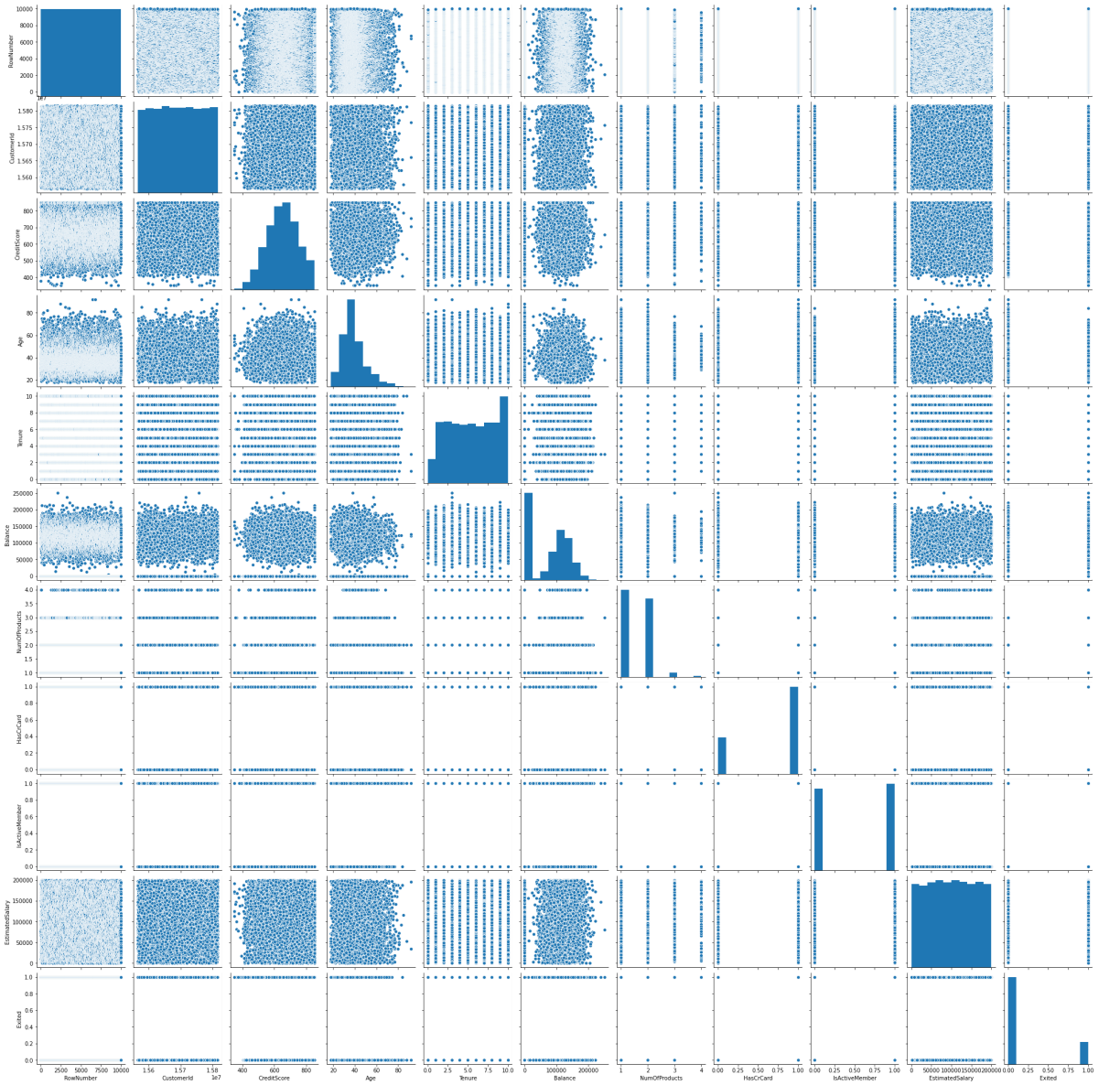
In [220]:
```
df.isnull().sum(axis=0)
```

Out[220]:
```
RowNumber          0
CustomerId         0
Surname            0
CreditScore        0
Geography          0
Gender             0
Age                0
Tenure             0
Balance            0
NumOfProducts      0
HasCrCard          0
IsActiveMember     0
EstimatedSalary    0
Exited             0
dtype: int64
```

There are no missing values that we need to deal with here

Visualize the data

In [221]:
```
sns.pairplot(df)
plt.show()
```



Feature Engineering

Drop columns we know from experience that are irrelevant, One Hot Encoding, Standardization

We can straight away drop Surname, CustomerId, RowNumber without checking anything because we know that these don't have any bearing on the churn rate. Furthermore, including Surname might actually add discrimination based on race etc.

The next step is to one-hot encode Gender and Geography.

The last step is to normalize some of the fields which have numbers.

Scaling is done before outlier handling because once scaling is done, it will allow us to plot the box plots at one go in a single graph. Scaling doesn't affect the frequency distribution and outlier detection.

In [222]:
```
def engineer_features(df: pd.DataFrame, donot_scale=False)->tuple:
    ohdf = df.drop(columns=['RowNumber', 'Surname', 'CustomerId'])
    ohdf.describe()
    X, y = ohdf[ohdf.columns[:-1]], ohdf[df.columns[-1]]
    X = pd.get_dummies(X, columns=['Gender', 'Geography'], dummy_na=False)
    scaled_X = X.copy()
    columns_to_transform = ['CreditScore', 'Age', 'Tenure', 'Balance', 'NumOfP
roducts', 'EstimatedSalary']
    if not donot_scale:
        scaled_X[columns_to_transform] = StandardScaler().fit_transform(scaled
_X[columns_to_transform])
    return X, scaled_X, y
X, scaled_X, y = engineer_features(df)
```

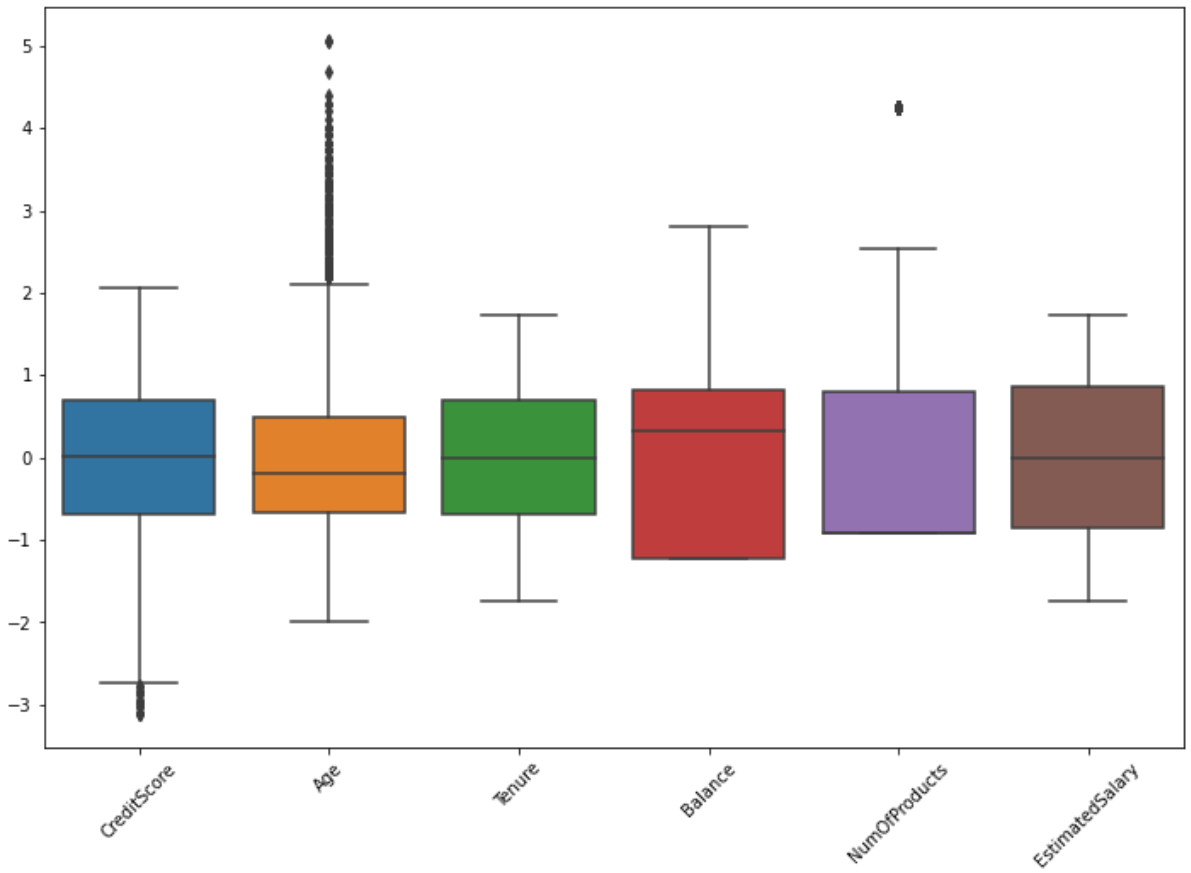
In [223]:
```
def engineer_features2(df: pd.DataFrame)->tuple:
    ohdf = df.drop(columns=['RowNumber', 'Surname', 'CustomerId'])
    X = pd.get_dummies(ohdf, columns=['Gender', 'Geography'], dummy_na=False)
    return X
```

Handle Outliers

If any outliers are present, they must be removed. The way to check for outliers is to

In [224]:
```
ax = sns.boxplot(data=scaled_X[['CreditScore', 'Age', 'Tenure', 'Balance', 'NumOfProducts', 'EstimatedSalary']])
ax.set_xticklabels(ax.get_xticklabels(),rotation=45)
```

Out[224]:
```
[Text(0, 0, 'CreditScore'),
 Text(1, 0, 'Age'),
 Text(2, 0, 'Tenure'),
 Text(3, 0, 'Balance'),
 Text(4, 0, 'NumOfProducts'),
 Text(5, 0, 'EstimatedSalary')]
```

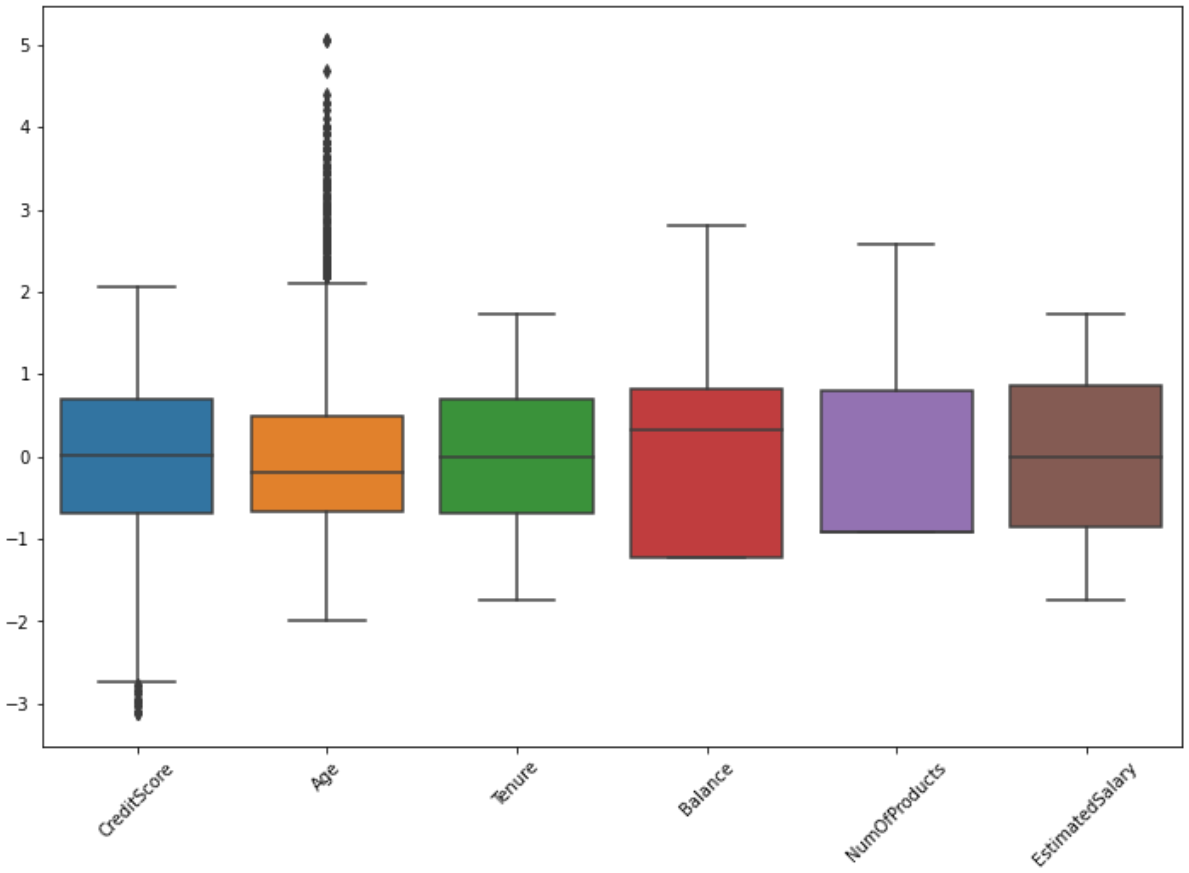


Clip NumOfProducts to [mean - 1.5 *IQR, mean + 1.5* IQR]

NumOfProducts is has outliers which are beyond the whiskers

In [225]:
```
q1 = scaled_X['NumOfProducts'].quantile(0.25)
q3 = scaled_X['NumOfProducts'].quantile(0.75)
avg = scaled_X['NumOfProducts'].mean()
iqr = q3 - q1
themin, themax = avg - 1.5 * iqr, avg + 1.5 * iqr
scaled_X[['NumOfProducts']] = scaled_X[['NumOfProducts']].clip(themin, themax)
ax = sns.boxplot(data=scaled_X[['CreditScore', 'Age', 'Tenure', 'Balance', 'Nu
mOfProducts', 'EstimatedSalary']])
ax.set_xticklabels(ax.get_xticklabels(),rotation=45)
```

Out[225]:
```
[Text(0, 0, 'CreditScore'),
 Text(1, 0, 'Age'),
 Text(2, 0, 'Tenure'),
 Text(3, 0, 'Balance'),
 Text(4, 0, 'NumOfProducts'),
 Text(5, 0, 'EstimatedSalary')]
```

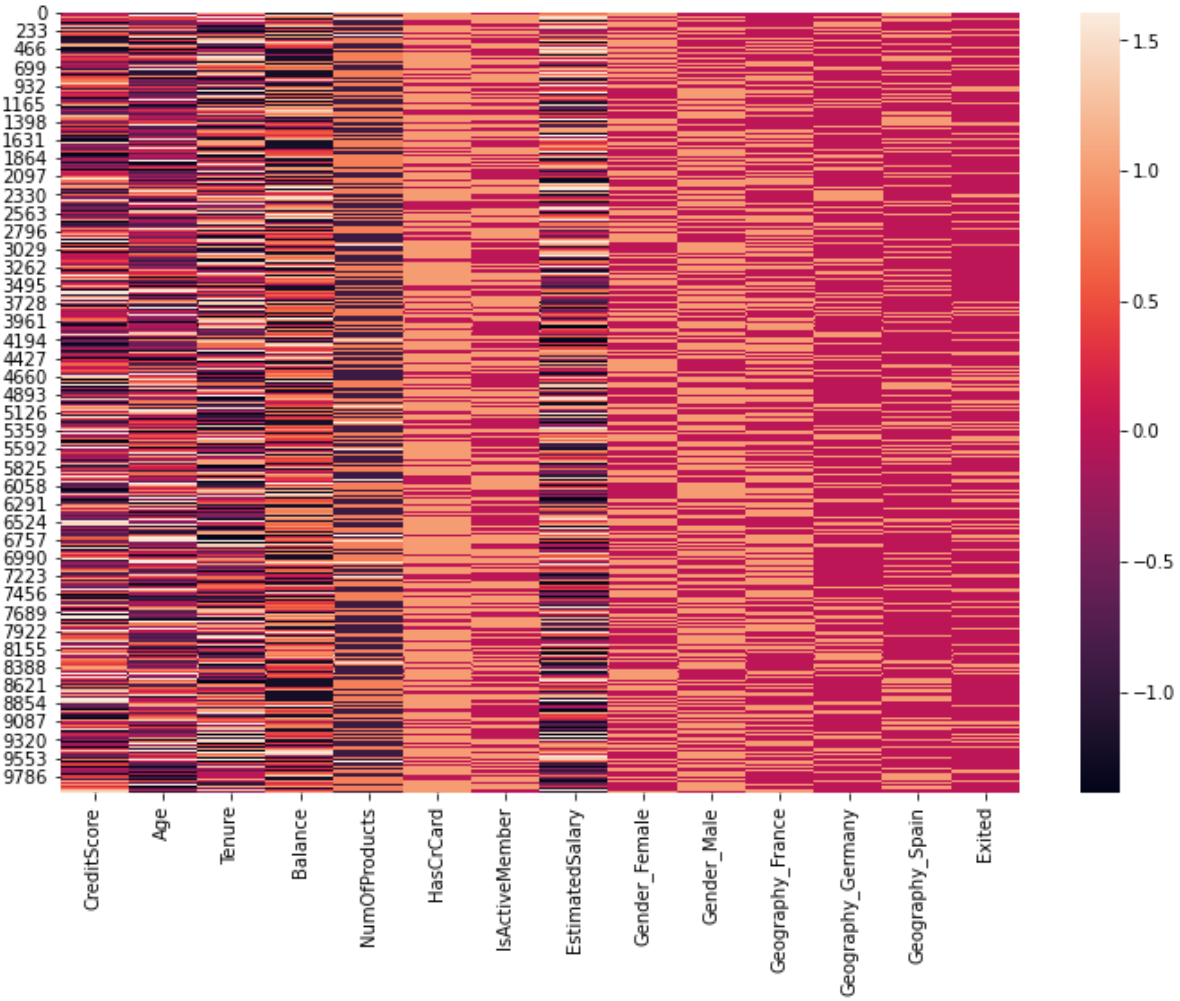


While there are some outliers here in CreditScore and Age, we will choose to let them remain as all these values form a continuum and touch the whiskers.

Visualize the Data as a heatmap

In [226]:
```
sns.heatmap(pd.concat([scaled_X, y], axis=1), robust=True)
```

Out[226]: <AxesSubplot:>



As we can see the Exited field is unbalanced Print the number of instances of each class

In [227]:
```
tc = 0
fc = 0
for i in y==1:
    if True == i:
        tc += 1
    else:
        fc += 1
print(f"Exited: 1 = {tc}, 0 = {fc}")
```

```
Exited: 1 = 2037, 0 = 7963
```

Feature Selection Initial Exploration

Check *univariate feature selection* first

We know that this will be biased against categorical values, but we'll see this anyway

In [228]:
```
def feature_selection_with_select_percentile():
    selector = SelectPercentile(f_classif, percentile=25)
    selector.fit(scaled_X, y)
    print('-' * 120)
    print("%40s ------- %10s" % ('Feature', 'Score'))
    print('-' * 120)
    sorted_indices = np.argsort(selector.scores_)[::-1]
    for i in sorted_indices:
        print("%40s ------- %10.5f" % (scaled_X.columns[i], selector.scores_[i
]))
    print('-' * 120)
    return selector
selector = feature_selection_with_select_percentile()
```

```
------------------------------------------------------------------------------------------
-----------------------------------------
                                 Feature -------      Score
------------------------------------------------------------------------------------------
-----------------------------------------
                                     Age -------  886.06327
                       Geography_Germany -------  310.25838
                          IsActiveMember -------  249.80079
                                 Balance -------  142.47383
                            Gender_Female -------  114.72799
                              Gender_Male -------  114.72799
                           Geography_France -------  111.36052
                             NumOfProducts -------   49.28413
                            Geography_Spain -------   27.80947
                                CreditScore -------    7.34452
                                     Tenure -------    1.96016
                            EstimatedSalary -------    1.46326
                                  HasCrCard -------    0.50940
------------------------------------------------------------------------------------------
-----------------------------------------
```

Also check the *Tree based feature selector* to see which are better

In [15]:
```
def feature_selection_with_random_forest():
    forest = RandomForestClassifier(n_estimators=250, random_state=0)
    forest.fit(scaled_X, y)
    sorted_indices = np.argsort(forest.feature_importances_)[::-1]
    print('-' * 120)
    print("%40s ------- %10s" % ('Feature', 'Importance'))
    print('-' * 120)
    for i in sorted_indices:
        print("%40s ------- %10.5f" % (scaled_X.columns[i], forest.feature_imp
ortances_[i]))
    print('-' * 120)
    return forest
forest = feature_selection_with_random_forest()
```

```
------------------------------------------------------------------------------
----------------------------------------
                                 Feature ------- Importance
------------------------------------------------------------------------------
----------------------------------------
                                     Age -------    0.24002
                         EstimatedSalary -------    0.14694
                             CreditScore -------    0.14569
                                 Balance -------    0.14112
                           NumOfProducts -------    0.12829
                                  Tenure -------    0.08314
                          IsActiveMember -------    0.04043
                        Geography_Germany -------    0.02082
                                HasCrCard -------    0.01836
                         Geography_France -------    0.00993
                          Geography_Spain -------    0.00855
                            Gender_Female -------    0.00854
                              Gender_Male -------    0.00818
------------------------------------------------------------------------------
----------------------------------------
```

Also see *Greedy feature selection* to see what it gives us

In [16]:
```
def feature_selection_with_svc():
    svm = SVC()
    scores = model_selection.cross_val_score(svm, scaled_X, y, cv=10)
    print('Initial Results', scores.mean())

    estimator = linear_model.LogisticRegression(multi_class='auto', solver='lb
fgs')
    rfecv = RFECV(estimator, cv=10)
    rfecv.fit(scaled_X, y)
    sorted_indices = np.argsort(rfecv.ranking_)
    print(rfecv.n_features_)
    print('-' * 120)
    print("%40s ------- %10s" % ('Feature', 'Ranking'))
    print('-' * 120)
    for i in sorted_indices:
        print("%40s ------- %10.5f" % (scaled_X.columns[i], rfecv.ranking_[i
]))
    print('-' * 120)
    return rfecv
rfecv = feature_selection_with_svc()
```

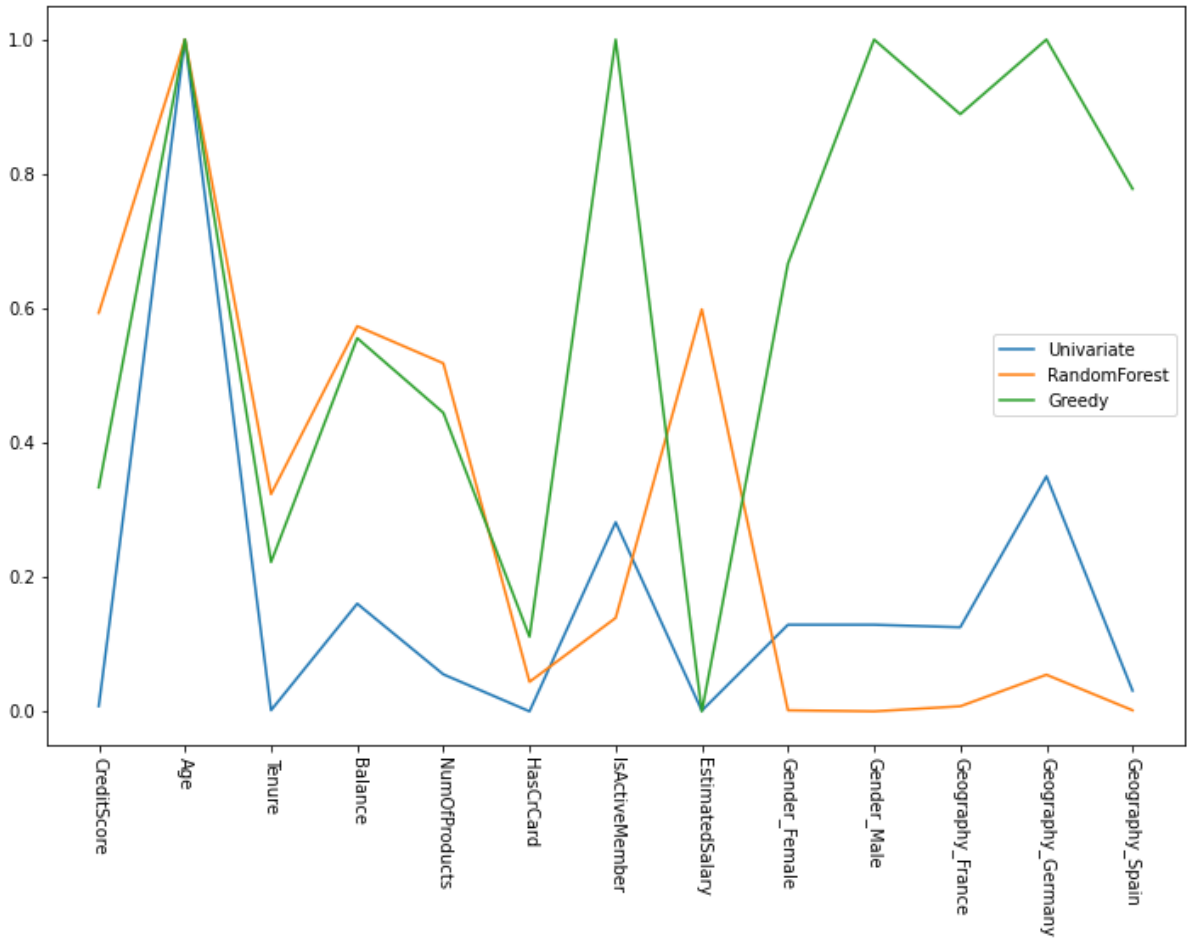
```
Initial Results 0.858
4
------------------------------------------------------------------------------
------------------------------------------
                                 Feature -------    Ranking
------------------------------------------------------------------------------
------------------------------------------
                                     Age -------    1.00000
                          IsActiveMember -------    1.00000
                             Gender_Male -------    1.00000
                       Geography_Germany -------    1.00000
                        Geography_France -------    2.00000
                         Geography_Spain -------    3.00000
                           Gender_Female -------    4.00000
                                 Balance -------    5.00000
                           NumOfProducts -------    6.00000
                             CreditScore -------    7.00000
                                  Tenure -------    8.00000
                               HasCrCard -------    9.00000
                         EstimatedSalary -------   10.00000
------------------------------------------------------------------------------
------------------------------------------
```

In [17]:

```
def get_normalized(a: list):
    """
    Normalizes a list to [0,1]
    """
    themin = min(a)
    themax = max(a)
    ret = []
    print(a, type(a))
    for i in a:
        ret.append((i - themin) / (themax - themin))
    return ret

print('-' * 120)
print("%40s ------------------ %10s %10s %10s" % ('Feature', 'Score', 'Importa
nce', 'Ranking'))
print('-' * 120)
for i, j, k, l in zip(scaled_X.columns, selector.scores_, forest.feature_impor
tances_, rfecv.ranking_):
    print("%40s ------------------ %10.1f %10.3f %10d" % (i, j, k, l))
print('-' * 120)
fig, ax = plt.subplots()
ax.plot(get_normalized(selector.scores_), label='Univariate')
ax.plot(get_normalized(forest.feature_importances_), label='RandomForest')
rf_ranking_inv = [max(rfecv.ranking_) - i for i in rfecv.ranking_]
ax.plot(get_normalized(rf_ranking_inv), label='Greedy')
plt.xticks(np.arange(0, len(scaled_X.columns), 1.0))
ax.set_xticklabels(scaled_X.columns, rotation=270)
ax.legend()
```

```
-------------------------------------------------------------------------------
----------------------------------------
                           Feature ----------------      Score Import
ance    Ranking
-------------------------------------------------------------------------------
----------------------------------------
                       CreditScore ----------------        7.3
0.146          7
                               Age ----------------      886.1
0.240          1
                            Tenure ----------------        2.0
0.083          8
                           Balance ----------------      142.5
0.141          5
                     NumOfProducts ----------------       49.3
0.128          6
                        HasCrCard ----------------        0.5
0.018          9
                    IsActiveMember ----------------      249.8
0.040          1
                   EstimatedSalary ----------------        1.5
0.147         10
                     Gender_Female ----------------      114.7
0.009          4
                       Gender_Male ----------------      114.7
0.008          1
                  Geography_France ----------------      111.4
0.010          2
                 Geography_Germany ----------------      310.3
0.021          1
                   Geography_Spain ----------------       27.8
0.009          3
-------------------------------------------------------------------------------
----------------------------------------
[7.34452216e+00 8.86063275e+02 1.96016363e+00 1.42473832e+02
 4.92841330e+01 5.09401035e-01 2.49800794e+02 1.46326192e+00
 1.14727989e+02 1.14727989e+02 1.11360517e+02 3.10258384e+02
 2.78094682e+01] <class 'numpy.ndarray'>
[0.14568742 0.24001543 0.08314233 0.14111972 0.12828553 0.01836008
 0.04043115 0.14693745 0.00854344 0.00817559 0.00993053 0.02081914
 0.00855218] <class 'numpy.ndarray'>
[3, 9, 2, 5, 4, 1, 9, 0, 6, 9, 8, 9, 7] <class 'list'>
```

Out[17]: <matplotlib.legend.Legend at 0x1ec48994400>



With all the features, using the three different methods to see which features are noisy, we don't get any clear indication about which features are important and which are not.

HasCrCard is probably not important, but even that is not very clear, so we'll let things be.

On the other hand, we know that Age, Balance, and IsActiveMember are probably very important.

Compare Basic Models

In [ ]:

```
import warnings; warnings.simplefilter('ignore')


def compare_basic_models(run_fn, test_models=None, no_sort=False)->None:
    """
    Run models against the data, compare them according to F1 and Accuracy
    """
    basic_models = [
        ('DecisionTreeClassifier', lambda: DecisionTreeClassifier(random_state
=1)),
        ('KNeighborsClassifier', lambda: KNeighborsClassifier()),
        ('NearestCentroid', lambda: NearestCentroid()),
        ('NaiveBayes', lambda: GaussianNB()),
        #('Gaussian Process', lambda: GaussianProcessClassifier(random_state=
1)),
        ('Support Vector Machines', lambda: SVC(random_state=1)),
        ('Random Forest', lambda: RandomForestClassifier(random_state=1)),
        ('Ada Boost', lambda: AdaBoostClassifier(random_state=1)),
        ('Gradient Boost', lambda: GradientBoostingClassifier()),
        ('Linear Discriminant Analysis', lambda: LinearDiscriminantAnalysis
()),
        ('Quadriatic Discriminant Analysis', lambda: QuadraticDiscriminantAnal
ysis()),
        ('Logistic Regression', lambda: LogisticRegression(random_state=1)),
        ('Ridge Classifier', lambda: RidgeClassifier(random_state=1)),
        ('Bagging Classifier', lambda: BaggingClassifier(random_state=1)),
        ('SGD Classifier', lambda: SGDClassifier(random_state=1)),
        ('Passive Aggressive Classifier', lambda: PassiveAggressiveClassifier(
random_state=1)),
        ('Perceptron', lambda: Perceptron(random_state=1)),
        ('Multi-Layer Perceptron', lambda: MLPClassifier(random_state=1))
    ]
    if None != test_models:
        basic_models = test_models
    accuracy_scores = []
    f1_scores = []
    models = []
    run_times = []
    for name, fn in basic_models:
        print(name, fn)
        t1 = perf_counter()
        acc, f1 = run_fn(scaled_X, y, fn)
        run_times.append(perf_counter() - t1)
        accuracy_scores.append(acc)
        f1_scores.append(f1)
        models.append(name)
    fig, ax = plt.subplots()
    min_times = min(run_times)
    max_times = max(run_times)
    if len(run_times) > 1:
        for i in range(len(run_times)):
            run_times[i] = (run_times[i] - min_times) / (max_times - min_times
)
    sorted_indices = np.argsort(f1_scores) # Sort by f1 score
    if no_sort:
        sorted_indices = [i for i in range(len(f1_scores))]
```

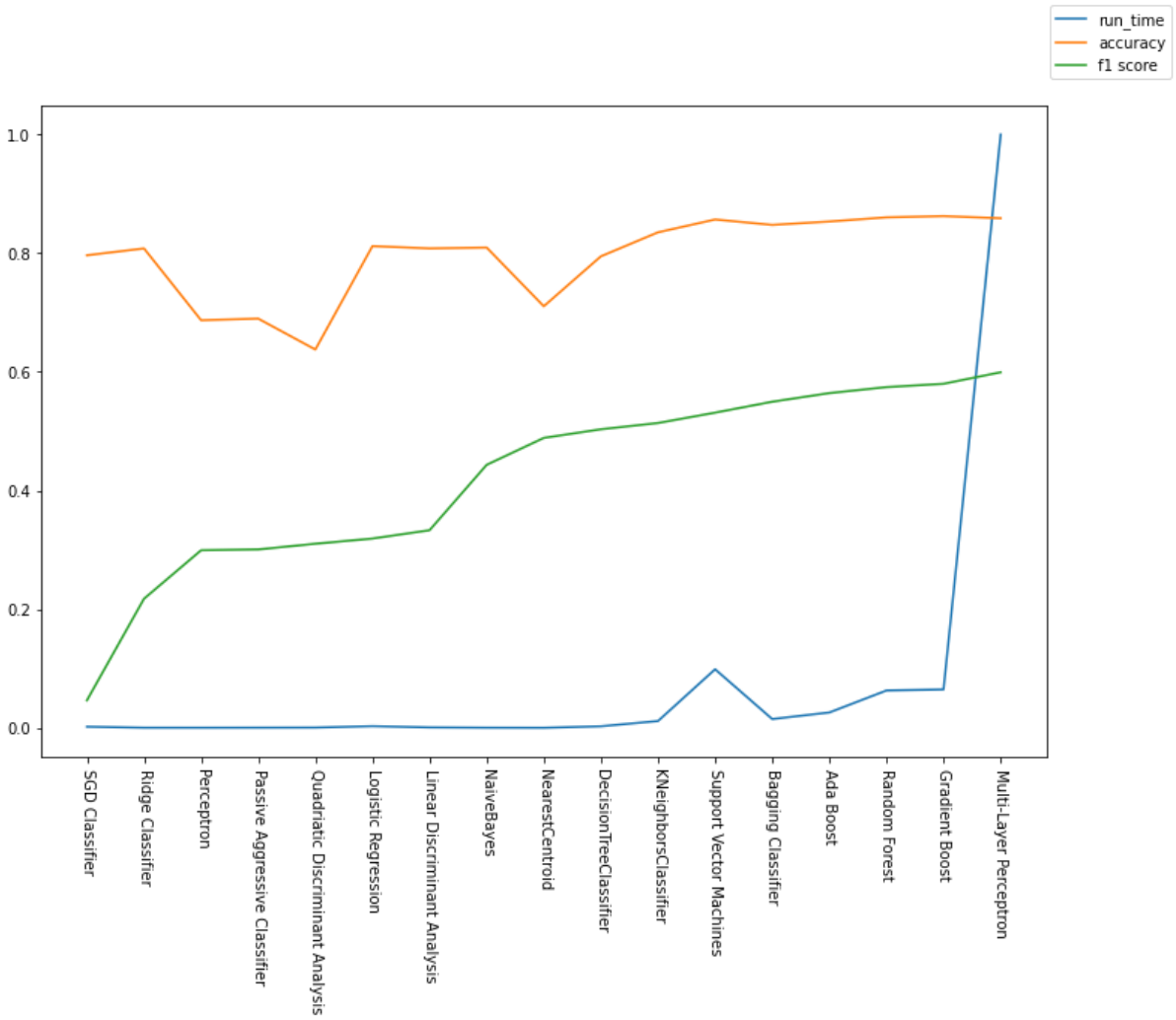
```
    ax.plot(range(len(run_times)), [run_times[i] for i in sorted_indices], lab
el="run_time")
    ax.plot(range(len(accuracy_scores)), [accuracy_scores[i] for i in sorted_i
ndices],  label="accuracy")
    ax.plot(range(len(f1_scores)), [f1_scores[i] for i in sorted_indices],  la
bel="f1 score")
    plt.xticks(np.arange(0, len(accuracy_scores), 1.0))
    ax.set_xticklabels([models[i] for i in sorted_indices], rotation=270)
    fig.legend()

    print('-' * 120)
    print("%40s  ----   %10s %10s %10s" % ('Model', 'Accuracy', 'F1 Score', 'R
un Time'))
    print('-' * 120)
    for i in sorted_indices:
        print("%40s  ----   %10.3f %10.3f %10.3f" % (models[i], accuracy_score
s[i], f1_scores[i], run_times[i]))
    print('-' * 120)
    return None
```

In [ ]:
```
def run_basic_model(X:pd.DataFrame, y:pd.DataFrame, model_gen_fn):
    kf = model_selection.StratifiedKFold(n_splits=6, shuffle=True, random_stat
e=5)
    accuracy_scores = []
    f1_scores = []
    for train_index, val_index in kf.split(X, y):
        clf = model_gen_fn()
        clf.fit(X.to_numpy()[train_index], y.to_numpy()[train_index])
        y_pred = clf.predict(X.to_numpy()[val_index])
        y_true = y.to_numpy()[val_index]
        accuracy_scores.append(metrics.accuracy_score(y_true, y_pred))
        f1_scores.append(metrics.f1_score(y_true, y_pred))
    #print(median(accuracy_scores), median(f1_scores))
    return mean(accuracy_scores), mean(f1_scores)
```

In [230]:
```
compare_basic_models(run_basic_model)
```

```
DecisionTreeClassifier <function compare_basic_models.<locals>.<lambda> at 0x
000001D921C8B1F0>
KNeighborsClassifier <function compare_basic_models.<locals>.<lambda> at 0x00
0001D921C8B3A0>
NearestCentroid <function compare_basic_models.<locals>.<lambda> at 0x000001D
85A6FC9D0>
NaiveBayes <function compare_basic_models.<locals>.<lambda> at 0x000001D85C11
3B80>
Support Vector Machines <function compare_basic_models.<locals>.<lambda> at 0
x000001D85C0DA940>
Random Forest <function compare_basic_models.<locals>.<lambda> at 0x000001D85
C0DAA60>
Ada Boost <function compare_basic_models.<locals>.<lambda> at 0x000001D85FCC9
040>
Gradient Boost <function compare_basic_models.<locals>.<lambda> at 0x000001D9
21DB81F0>
Linear Discriminant Analysis <function compare_basic_models.<locals>.<lambda>
at 0x000001D921DB8040>
Quadriatic Discriminant Analysis <function compare_basic_models.<locals>.<lam
bda> at 0x000001D921DB80D0>
Logistic Regression <function compare_basic_models.<locals>.<lambda> at 0x000
001D921DB8160>
Ridge Classifier <function compare_basic_models.<locals>.<lambda> at 0x000001
D921DB83A0>
Bagging Classifier <function compare_basic_models.<locals>.<lambda> at 0x0000
01D921DB8430>
SGD Classifier <function compare_basic_models.<locals>.<lambda> at 0x000001D9
21DB84C0>
Passive Aggressive Classifier <function compare_basic_models.<locals>.<lambda
> at 0x000001D921DB8550>
Perceptron <function compare_basic_models.<locals>.<lambda> at 0x000001D921DB
85E0>
Multi-Layer Perceptron <function compare_basic_models.<locals>.<lambda> at 0x
000001D921DB8670>

[0.7946989833725933, 0.8352004269014144, 0.7105023773156534, 0.80950130502110
86, 0.8568007887018034, 0.8603999488217603, 0.8534005683857226, 0.86249994898
97939, 0.8081003247129686, 0.637607952599156, 0.8119003450210318, 0.808000224
6849549, 0.8476987075574081, 0.796500483816802, 0.6897167985370513, 0.6870214
792575698, 0.8589005488218083]

[0.5031704409586167, 0.5137007304064225, 0.4886215058787005, 0.44326044155491
995, 0.5311815256147012, 0.5742653346426003, 0.5641624471239775, 0.5798795450
56633, 0.33304487433477115, 0.3102288048496446, 0.3190099885747713, 0.2175284
9103995742, 0.5496146844524826, 0.046178375883008924, 0.3006348728021189, 0.2
992211692260719, 0.5993378896652194]

--------------------------------------------------------------------------------
-----------------------------------------
                                  Model  ----     Accuracy   F1 Score   Run
Time
--------------------------------------------------------------------------------
-----------------------------------------
                         SGD Classifier  ----        0.797      0.046
0.002
                       Ridge Classifier  ----        0.808      0.218
0.000
```

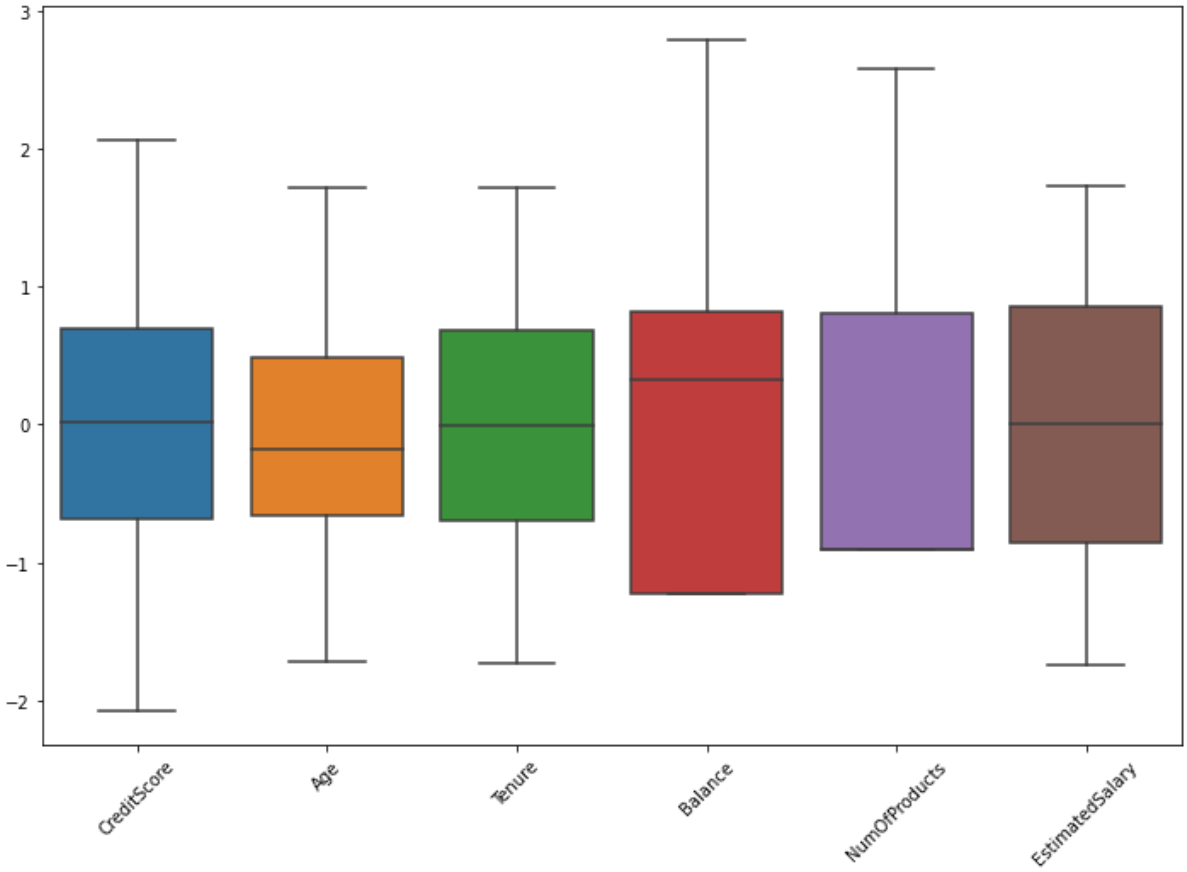
```
                           Perceptron  ----        0.687      0.299
0.000
          Passive Aggressive Classifier  ----        0.690      0.301
0.000
       Quadriatic Discriminant Analysis  ----        0.638      0.310
0.000
                  Logistic Regression  ----        0.812      0.319
0.003
           Linear Discriminant Analysis  ----        0.808      0.333
0.001
                           NaiveBayes  ----        0.810      0.443
0.000
                       NearestCentroid  ----        0.711      0.489
0.000
                DecisionTreeClassifier  ----        0.795      0.503
0.003
                  KNeighborsClassifier  ----        0.835      0.514
0.011
               Support Vector Machines  ----        0.857      0.531
0.099
                    Bagging Classifier  ----        0.848      0.550
0.015
                             Ada Boost  ----        0.853      0.564
0.026
                         Random Forest  ----        0.860      0.574
0.063
                         Gradient Boost  ----        0.862      0.580
0.065
                  Multi-Layer Perceptron  ----        0.859      0.599
1.000
--------------------------------------------------------------------------------
-----------------------------------------
```

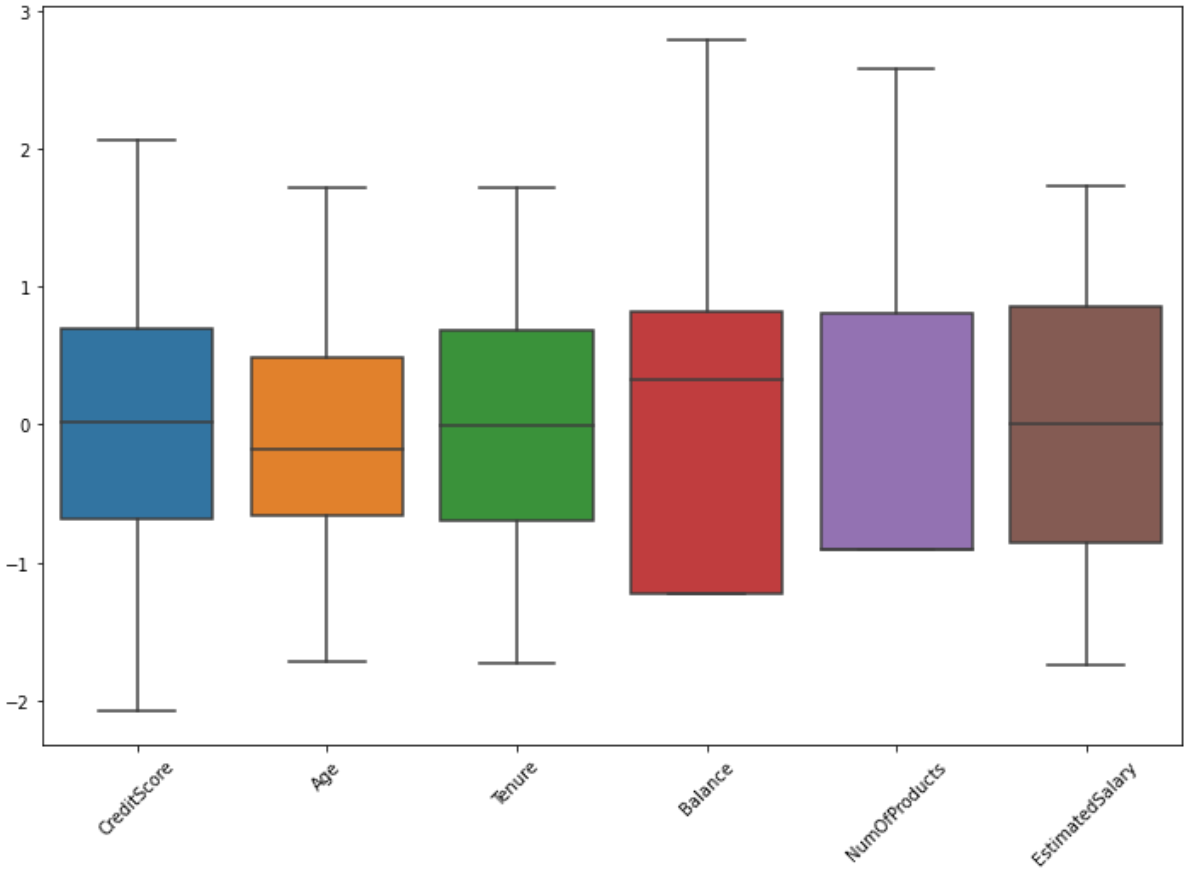


Does clipping CreditScore and Age give better results?

In [231]:

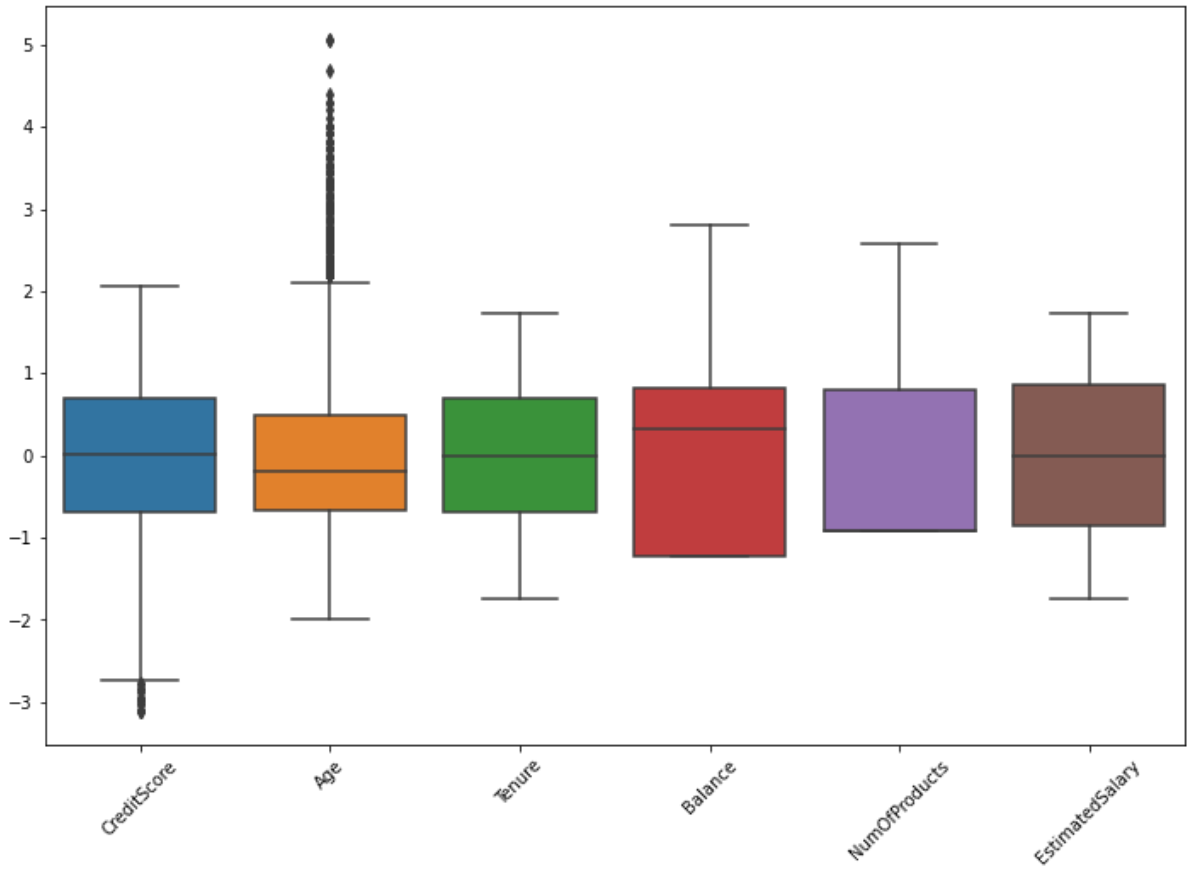
```
ax = sns.boxplot(data=temp_scaled_X[['CreditScore', 'Age', 'Tenure', 'Balance'
, 'NumOfProducts', 'EstimatedSalary']])
ax.set_xticklabels(ax.get_xticklabels(),rotation=45)
plt.show()
q1 = scaled_X['Age'].quantile(0.25)
q3 = scaled_X['Age'].quantile(0.75)
avg = scaled_X['Age'].mean()
iqr = q3 - q1
themin, themax = avg - 1.5 * iqr, avg + 1.5 * iqr
temp_scaled_X = scaled_X.copy()
temp_scaled_X[['Age']] = scaled_X[['Age']].clip(themin, themax)
q1 = scaled_X['CreditScore'].quantile(0.25)
q3 = scaled_X['CreditScore'].quantile(0.75)
avg = scaled_X['CreditScore'].mean()
iqr = q3 - q1
themin, themax = avg - 1.5 * iqr, avg + 1.5 * iqr
temp_scaled_X[['CreditScore']] = scaled_X[['CreditScore']].clip(themin, themax
)
ax = sns.boxplot(data=temp_scaled_X[['CreditScore', 'Age', 'Tenure', 'Balance'
, 'NumOfProducts', 'EstimatedSalary']])
ax.set_xticklabels(ax.get_xticklabels(),rotation=45)
plt.show()
ax = sns.boxplot(data=scaled_X[['CreditScore', 'Age', 'Tenure', 'Balance', 'Nu
mOfProducts', 'EstimatedSalary']])
ax.set_xticklabels(ax.get_xticklabels(),rotation=45)

saved_scaled_X = scaled_X
scaled_X = temp_scaled_X

compare_basic_models(run_basic_model)

scaled_X = saved_scaled_X
```

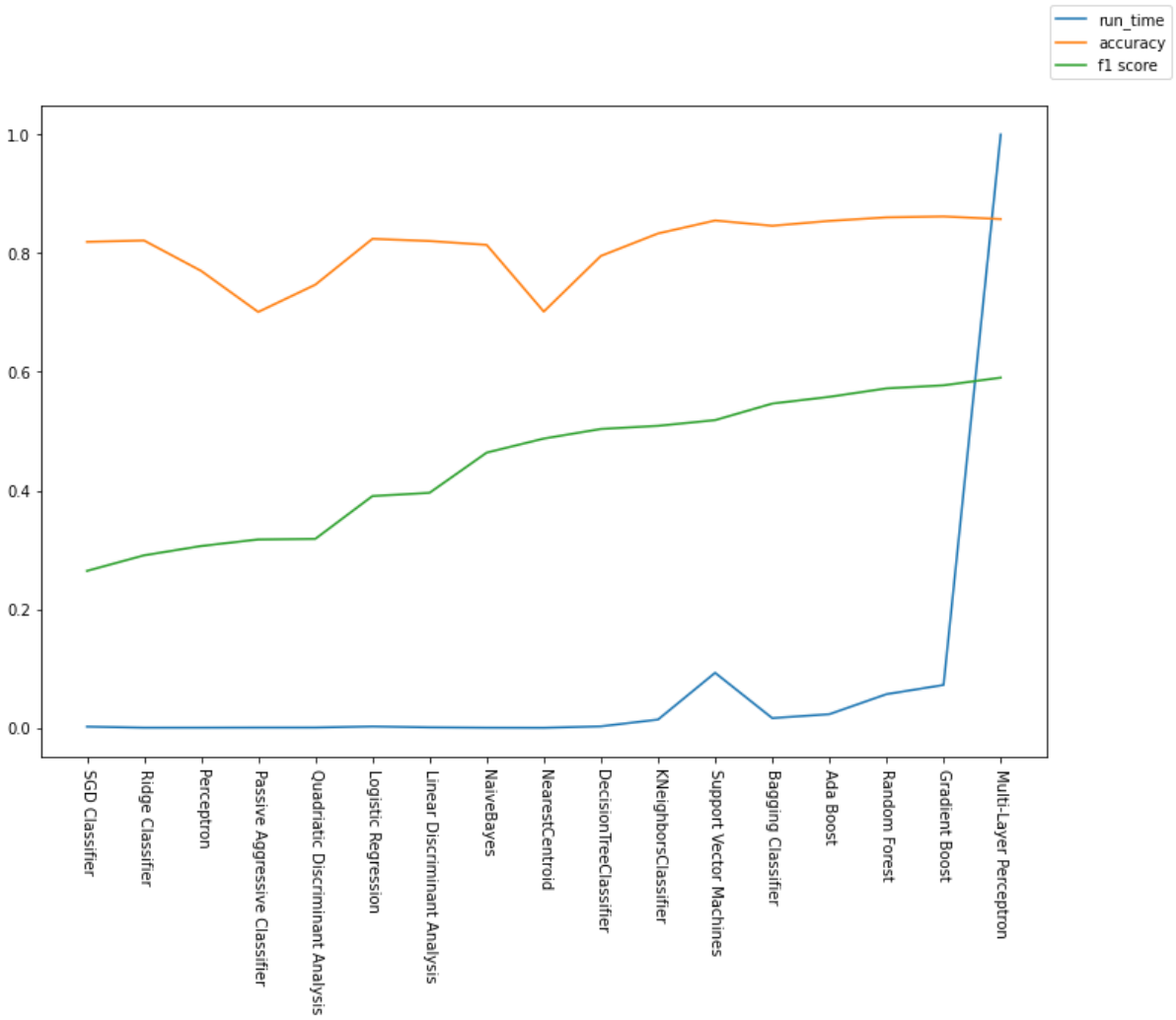

3
2
1
0
−1
−2
CreditScore
Age
Tenure
Balance
NumOfProducts
EstimatedSalary


3
2
1
0
−1
−2
CreditScore
Age
Tenure
Balance
NumOfProducts
EstimatedSalary

```
DecisionTreeClassifier <function compare_basic_models.<locals>.<lambda> at 0x
000001D921ED7550>
KNeighborsClassifier <function compare_basic_models.<locals>.<lambda> at 0x00
0001D85C0DAA60>
NearestCentroid <function compare_basic_models.<locals>.<lambda> at 0x000001D
92285CEE0>
NaiveBayes <function compare_basic_models.<locals>.<lambda> at 0x000001D92285
CE50>
Support Vector Machines <function compare_basic_models.<locals>.<lambda> at 0
x000001D92285CF70>
Random Forest <function compare_basic_models.<locals>.<lambda> at 0x000001D92
6084160>
Ada Boost <function compare_basic_models.<locals>.<lambda> at 0x000001D926084
1F0>
Gradient Boost <function compare_basic_models.<locals>.<lambda> at 0x000001D9
26084280>
Linear Discriminant Analysis <function compare_basic_models.<locals>.<lambda>
at 0x000001D926084310>
Quadriatic Discriminant Analysis <function compare_basic_models.<locals>.<lam
bda> at 0x000001D9260843A0>
Logistic Regression <function compare_basic_models.<locals>.<lambda> at 0x000
001D926084430>
Ridge Classifier <function compare_basic_models.<locals>.<lambda> at 0x000001
D9260844C0>
Bagging Classifier <function compare_basic_models.<locals>.<lambda> at 0x0000
01D926084550>
SGD Classifier <function compare_basic_models.<locals>.<lambda> at 0x000001D9
260845E0>
Passive Aggressive Classifier <function compare_basic_models.<locals>.<lambda
> at 0x000001D926084670>
Perceptron <function compare_basic_models.<locals>.<lambda> at 0x000001D92608
4700>
Multi-Layer Perceptron <function compare_basic_models.<locals>.<lambda> at 0x
000001D926084790>

[0.7954008237968253, 0.8332002867133657, 0.701602236575494, 0.81400142540519
5, 0.855100828573781, 0.8604002488817963, 0.8544001283776859, 0.8619000689657
987, 0.8205006057612008, 0.7469090215570331, 0.8242003460052287, 0.8214003057
731791, 0.8461982273413264, 0.8190986052489382, 0.700916419357185, 0.77020796
32092789, 0.857499508501661]

[0.5037048797982363, 0.5089879804640534, 0.48752384058352977, 0.4638115851373
148, 0.5186137509635543, 0.5720896277616712, 0.5578547353621756, 0.5772606442
819923, 0.3960868197866976, 0.3185031625042545, 0.3906023043257644, 0.2906881
081709625, 0.5464874727535429, 0.2644552407389587, 0.3174896123028212, 0.3063
8731279514375, 0.590189636921225]

--------------------------------------------------------------------------
-----------------------------------------
                                  Model  ----     Accuracy   F1 Score   Run
Time
--------------------------------------------------------------------------
-----------------------------------------
                          SGD Classifier  ----        0.819      0.264
0.002
                        Ridge Classifier  ----        0.821      0.291
0.000
```

```
                           Perceptron  ----        0.770      0.306
0.000
          Passive Aggressive Classifier  ----        0.701      0.317
0.000
       Quadriatic Discriminant Analysis  ----        0.747      0.319
0.000
                  Logistic Regression  ----        0.824      0.391
0.002
           Linear Discriminant Analysis  ----        0.821      0.396
0.001
                           NaiveBayes  ----        0.814      0.464
0.000
                       NearestCentroid  ----        0.702      0.488
0.000
                DecisionTreeClassifier  ----        0.795      0.504
0.002
                  KNeighborsClassifier  ----        0.833      0.509
0.014
               Support Vector Machines  ----        0.855      0.519
0.093
                    Bagging Classifier  ----        0.846      0.546
0.016
                             Ada Boost  ----        0.854      0.558
0.023
                         Random Forest  ----        0.860      0.572
0.057
                         Gradient Boost  ----        0.862      0.577
0.072
                Multi-Layer Perceptron  ----        0.857      0.590
1.000
--------------------------------------------------------------------------------
-----------------------------------------
```


5
4
3
2
1
0
−1
−2
−3
CreditScore
Age
Tenure
Balance
NumOfProducts
EstimatedSalary



No, Clipping these columsn produces worse results.

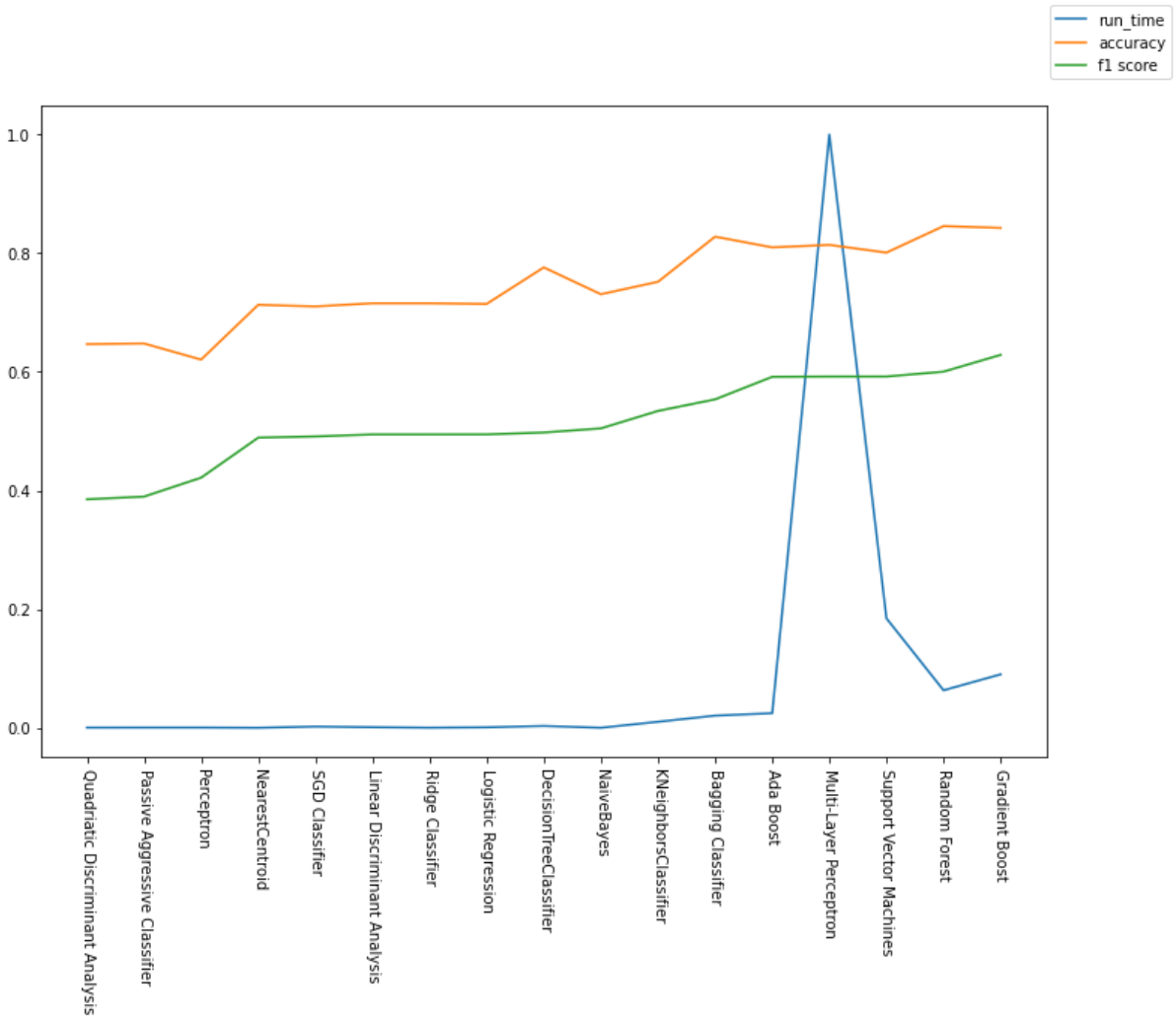
Evaluate the same models with SMOTE

In [ ]:
```
def run_basic_model_with_smote(X, y, model_gen_fn):
    accuracy_scores = []
    f1_scores = []
    kf = model_selection.StratifiedKFold(n_splits=6, shuffle=True, random_stat
e=2)
    for train_index, val_index in kf.split(scaled_X, y):
        model = model_gen_fn()
        smote = SMOTE(random_state=1)
        p = imblearn.pipeline.Pipeline([('smote', smote), ('model', model)])
        p.fit(X.to_numpy()[train_index], y.to_numpy()[train_index])
        y_pred = p.predict(X.to_numpy()[val_index])
        y_true = y.to_numpy()[val_index]
        accuracy_scores.append(metrics.accuracy_score(y_true, y_pred))
        f1_scores.append(metrics.f1_score(y_true, y_pred))
    return mean(accuracy_scores), mean(f1_scores)
```

In [20]:
```
compare_basic_models(run_basic_model_with_smote)
```

```
DecisionTreeClassifier <function compare_basic_models.<locals>.<lambda> at 0x
000001EC48AF5E50>
KNeighborsClassifier <function compare_basic_models.<locals>.<lambda> at 0x00
0001EC48AF5C10>
NearestCentroid <function compare_basic_models.<locals>.<lambda> at 0x000001E
C48AF5B80>
NaiveBayes <function compare_basic_models.<locals>.<lambda> at 0x000001EC48AF
5F70>
Support Vector Machines <function compare_basic_models.<locals>.<lambda> at 0
x000001EC48AF5DC0>
Random Forest <function compare_basic_models.<locals>.<lambda> at 0x000001EC4
8AF5CA0>
Ada Boost <function compare_basic_models.<locals>.<lambda> at 0x000001EC48AF5
AF0>
Gradient Boost <function compare_basic_models.<locals>.<lambda> at 0x000001EC
48AF5A60>
Linear Discriminant Analysis <function compare_basic_models.<locals>.<lambda>
at 0x000001EC48AF58B0>
Quadriatic Discriminant Analysis <function compare_basic_models.<locals>.<lam
bda> at 0x000001EC48E015E0>
Logistic Regression <function compare_basic_models.<locals>.<lambda> at 0x000
001EC48E01550>
Ridge Classifier <function compare_basic_models.<locals>.<lambda> at 0x000001
EC48E014C0>
Bagging Classifier <function compare_basic_models.<locals>.<lambda> at 0x0000
01EC48E01040>
SGD Classifier <function compare_basic_models.<locals>.<lambda> at 0x000001EC
48E010D0>
Passive Aggressive Classifier <function compare_basic_models.<locals>.<lambda
> at 0x000001EC48E011F0>
Perceptron <function compare_basic_models.<locals>.<lambda> at 0x000001EC48E0
1280>
Multi-Layer Perceptron <function compare_basic_models.<locals>.<lambda> at 0x
000001EC48E013A0>

[0.7760009822765339, 0.7517983198078272, 0.7129967163830139, 0.73080035853573
59, 0.8009976636125355, 0.8456989874534097, 0.8097986645167965, 0.84259942729
34129, 0.715598536955274, 0.6467847486925185, 0.7144985768752612, 0.715598536
955274, 0.8279003262972856, 0.7103015171275469, 0.6477943907016915, 0.6205905
877648, 0.8139980647327918]

[0.49773909368697317, 0.5340905579876735, 0.4890381273390903, 0.5046504045580
438, 0.5921478562351585, 0.6002998630496208, 0.5916159095475583, 0.6285986183
281258, 0.4944645757888678, 0.38513165717897785, 0.4945295530508275, 0.494464
5757888678, 0.5536137536279351, 0.49110946471286265, 0.3896596518307815, 0.42
158509394340804, 0.5920796851762115]

------------------------------------------------------------------------------
-----------------------------------------
                                  Model  ----     Accuracy   F1 Score   Run
Time
------------------------------------------------------------------------------
-----------------------------------------
        Quadriatic Discriminant Analysis  ----        0.647      0.385
0.000
           Passive Aggressive Classifier  ----        0.648      0.390
0.000
```

```
                           Perceptron  ----        0.621      0.422
0.000
                       NearestCentroid  ----        0.713      0.489
0.000
                        SGD Classifier  ----        0.710      0.491
0.002
          Linear Discriminant Analysis  ----        0.716      0.494
0.001
                      Ridge Classifier  ----        0.716      0.494
0.000
                  Logistic Regression  ----        0.714      0.495
0.001
               DecisionTreeClassifier  ----        0.776      0.498
0.003
                           NaiveBayes  ----        0.731      0.505
0.000
                 KNeighborsClassifier  ----        0.752      0.534
0.010
                  Bagging Classifier  ----        0.828      0.554
0.021
                           Ada Boost  ----        0.810      0.592
0.025
               Multi-Layer Perceptron  ----        0.814      0.592
1.000
             Support Vector Machines  ----        0.801      0.592
0.184
                       Random Forest  ----        0.846      0.600
0.063
                       Gradient Boost  ----        0.843      0.629
0.090
--------------------------------------------------------------------------------
-----------------------------------------
```



Hyper Parameter Optimization

Test GradientBoost with GridSearchCV

In [42]:
```
def grid_search_gradient_boosting(X, y):
    sm = SMOTE(random_state=1)
    gbc = GradientBoostingClassifier(random_state=1)
    p = imblearn.pipeline.Pipeline([('sm', sm), ('gbc', gbc)])
    param_grid = {
        'gbc__n_estimators': [10, 100, 1000],
        #'gbc__loss': ['deviance', 'exponential'],
        'gbc__max_depth': [3, 10],
        'gbc__min_samples_split': [i for i in range(2, 210, 40)],
        'gbc__max_features': [None, 'auto', 'sqrt', 'log2'],
    }
    scorer = make_scorer(metrics.f1_score)
    cv1 = StratifiedKFold(n_splits=6, random_state=10, shuffle=True)
    search = GridSearchCV(p, param_grid, n_jobs=7, cv=cv1, scoring=scorer)
    search.fit(X, y)
    print("GRADIENT BOOSTING")
    print(f"Best Params = {search.best_params_}")
    print(f"Best score = {search.best_score_}")
    print('-' * 120)
    return search.best_estimator_

random.seed(22)
gbc = grid_search_gradient_boosting(scaled_X.to_numpy(), y.to_numpy())
```

```
GRADIENT BOOSTING
Best Params = {'gbc__max_depth': 3, 'gbc__max_features': None, 'gbc__min_samp
les_split': 42, 'gbc__n_estimators': 100}
Best score = 0.6311883955640584
------------------------------------------------------------------------------
----------------------------------------
```

In [176]:
```
def grid_search_gradient_boosting2(X, y):
    sm = SMOTE(random_state=1)
    gbc = GradientBoostingClassifier(random_state=1)
    p = imblearn.pipeline.Pipeline([('sm', sm), ('gbc', gbc)])
    param_grid = {
        'gbc__n_estimators': [10, 100, 1000, 10000],
        #'gbc__loss': ['deviance', 'exponential'],
        'gbc__max_depth': [3],
        'gbc__min_samples_split': [42],
        'gbc__max_features': [None],
        'gbc__learning_rate': [0.01, 0.05, 0.1, 0.15, 0.2]
    }
    scorer = make_scorer(metrics.f1_score)
    cv1 = StratifiedKFold(n_splits=6, random_state=10, shuffle=True)
    search = GridSearchCV(p, param_grid, n_jobs=7, cv=cv1, scoring=scorer)
    search.fit(X, y)
    print("GRADIENT BOOSTING")
    print(f"Best Params = {search.best_params_}")
    print(f"Best score = {search.best_score_}")
    print('-' * 120)
    return search.best_estimator_

random.seed(22)
gbc = grid_search_gradient_boosting2(scaled_X.to_numpy(), y.to_numpy())
```

```
GRADIENT BOOSTING
Best Params = {'gbc__learning_rate': 0.1, 'gbc__max_depth': 3, 'gbc__max_feat
ures': None, 'gbc__min_samples_split': 42, 'gbc__n_estimators': 100}
Best score = 0.6311883955640584
------------------------------------------------------------------------------
----------------------------------------
```

Test RandomForest with GridSearchCV

In [36]:
```
def grid_search_random_forest(X, y):
    sm = SMOTE(random_state=1)
    rfc = RandomForestClassifier(random_state=1)
    p = imblearn.pipeline.Pipeline([('sm', sm), ('rfc', rfc)])
    param_grid = {
        'rfc__n_estimators': [10, 100, 1000],
        'rfc__criterion': ['gini', 'entropy'],
        'rfc__min_samples_split': [i for i in range(20, 200, 40)],
        'rfc__max_depth': [None, 5],
        'rfc__class_weight': ['balanced', 'balanced_subsample']
    }
    #scoring = {'F1': make_scorer(metrics.f1_score), 'ACC': make_scorer(metric
s.accuracy_score)}
    scorer = make_scorer(metrics.f1_score)
    cv1 = StratifiedKFold(n_splits=6, random_state=10, shuffle=True)
    search = GridSearchCV(p, param_grid, n_jobs=7, cv=cv1, scoring=scorer)
    search.fit(X, y)
    print("RANDOM FOREST")
    print(f"Best Params = {search.best_params_}")
    print(f"Best score = {search.best_score_}")
    print('-' * 120)
    return search.best_estimator_

random.seed(22)
rfc  = grid_search_random_forest(scaled_X.to_numpy(), y.to_numpy())
```

```
RANDOM FOREST
Best Params = {'rfc__class_weight': 'balanced_subsample', 'rfc__criterion':
'entropy', 'rfc__max_depth': None, 'rfc__min_samples_split': 100, 'rfc__n_est
imators': 1000}
Best score = 0.624375225286892
------------------------------------------------------------------------------
------------------------------------------
```

In [172]:
```
def grid_search_random_forest2(X, y):
    sm = SMOTE(random_state=1)
    rfc = RandomForestClassifier(random_state=1)
    p = imblearn.pipeline.Pipeline([('sm', sm), ('rfc', rfc)])
    param_grid = {
        'rfc__n_estimators': [1000, 10000],
        'rfc__criterion': ['entropy'],
        'rfc__min_samples_split': [42],
        'rfc__max_depth': [None, 5],
        'rfc__class_weight': ['balanced_subsample'],
        'rfc__min_impurity_decrease': [0, 0.00001, 0.0001, 0.001, 0.01, 0.1]
    }
    #scoring = {'F1': make_scorer(metrics.f1_score), 'ACC': make_scorer(metric
s.accuracy_score)}
    scorer = make_scorer(metrics.f1_score)
    cv1 = StratifiedKFold(n_splits=6, random_state=10, shuffle=True)
    search = GridSearchCV(p, param_grid, n_jobs=7, cv=cv1, scoring=scorer)
    search.fit(X, y)
    print("RANDOM FOREST")
    print(f"Best Params = {search.best_params_}")
    print(f"Best score = {search.best_score_}")
    print('-' * 120)
    return search.best_estimator_

random.seed(22)
rfc  = grid_search_random_forest2(scaled_X.to_numpy(), y.to_numpy())
```

```
RANDOM FOREST
Best Params = {'rfc__class_weight': 'balanced_subsample', 'rfc__criterion':
'entropy', 'rfc__max_depth': None, 'rfc__min_impurity_decrease': 0.0001, 'rfc
__min_samples_split': 42, 'rfc__n_estimators': 1000}
Best score = 0.6212218073500146
------------------------------------------------------------------------------
----------------------------------------
```

Test SVC with GridSearchCV

In [217]:
```
def grid_search_svc(X, y):
    sm = SMOTE(random_state=1)
    svc = SVC(random_state=1)
    p = imblearn.pipeline.Pipeline([('sm', sm), ('svc', svc)])
    param_grid = {
        'svc__kernel': ['linear', 'poly', 'rbf', 'sigmoid', 'precomputed'],
        'svc__class_weight': [None, 'balanced'],
        'svc__gamma': ['scale', 'auto', 0.001],
        'sm__k_neighbors': [5, 20],
    }
    scorer = make_scorer(metrics.f1_score)
    cv1 = StratifiedKFold(n_splits=6, random_state=10, shuffle=True)
    search = GridSearchCV(p, param_grid, n_jobs=7, cv=cv1, scoring=scorer)
    search.fit(X, y)
    print("SVC")
    print(f"Best Params = {search.best_params_}")
    print(f"Best score = {search.best_score_}")
    print('-' * 120)
    return search.best_estimator_

random.seed(22)
selector = []
svc = grid_search_svc(scaled_X.to_numpy(), y.to_numpy())
```

```
SVC
Best Params = {'sm__k_neighbors': 20, 'svc__class_weight': None, 'svc__gamm
a': 'auto', 'svc__kernel': 'rbf'}
Best score = 0.5958117676964181
------------------------------------------------------------------------
----------------------------------------
```

In [218]:
```
scaled_X.shape
```

Out[218]:
```
(10000, 13)
```

Test MLP with GridSearchCV

In [48]:
```
def grid_search_mlp(X, y):
    sm = SMOTE(random_state=1)
    mlp = MLPClassifier(random_state=1, n_iter_no_change=100, max_iter=10000)
    p = imblearn.pipeline.Pipeline([('sm', sm), ('mlp', mlp)])
    param_grid = {
        'mlp__hidden_layer_sizes': [(100,), (100, 50,),],
        'mlp__activation': ['relu', 'logistic'],
        'mlp__solver': ['adam'],
        'mlp__learning_rate': ['constant', 'adaptive'],
        #'mlp__alpha': [0.1, 0.01, 0.001, 0.0001]
    }
    scorer = make_scorer(metrics.f1_score)
    cv1 = StratifiedKFold(n_splits=6, random_state=10, shuffle=True)
    search = GridSearchCV(p, param_grid, n_jobs=7, cv=cv1, scoring=scorer)
    search.fit(X, y)
    print("MULTI LEVEL PERCEPTRON")
    print(f"Best Params = {search.best_params_}")
    print(f"Best score = {search.best_score_}")
    print('-' * 120)
    return search.best_estimator_

random.seed(22)
mlp = grid_search_mlp(scaled_X.to_numpy(), y.to_numpy())
```

```
MULTI LEVEL PERCEPTRON
Best Params = {'mlp__activation': 'relu', 'mlp__hidden_layer_sizes': (100,),
'mlp__learning_rate': 'constant', 'mlp__solver': 'adam'}
Best score = 0.5404128313643786
------------------------------------------------------------------------------
----------------------------------------
```

Final Baseline

GRADIENT BOOSTING Best Params = {'gbc**max_depth': 3, 'gbc**max_features': None, 'gbc**min_samples_split': 42, 'gbc**n_estimators': 100} Best score = 0.6311883955640584

RANDOM FOREST Best Params = {'rfc**class_weight': 'balanced_subsample', 'rfc**criterion': 'entropy', 'rfc**max_depth': None, 'rfc**min_samples_split': 100, 'rfc__n_estimators': 1000} Best score = 0.624375225286892

SVC Best Params = {'sm**k_neighbors': 20, 'svc**class_weight': None, 'svc**gamma': 'auto', 'svc**kernel': 'rbf'} Best score = 0.5958117676964181

MULTI LEVEL PERCEPTRON Best Params = {'mlp**activation': 'relu', 'mlp**hidden_layer_sizes': (100,), 'mlp**learning_rate': 'constant', 'mlp**solver': 'adam'} Best score = 0.5404128313643786

Basic Experimentation

Are there any outliers detected by Random Forest?

In [213]:
```
tdf = read_file()
tdf = engineer_features2(tdf)
clf = IsolationForest(contamination=1.0e-2000).fit(StandardScaler().fit_transf
orm(tdf))
outliers = clf.predict(tdf)
out = 0
inl = 0
for i in outliers:
    if 1 == i:
        inl += 1
    else:
        out += 1
print(f"Outliers = {out} in = {inl}")
```

```
Outliers = 20 in = 9980
```

In [245]:
```
# Check the number of items categorized as outliers as the random number is va
ried
for r in range(10):
    tdf = read_file()
    tdf = engineer_features2(tdf)
    clf = IsolationForest(contamination=0.01, random_state=r, n_estimators=100
0, n_jobs=-1).fit(StandardScaler().fit_transform(tdf))
    outliers = clf.predict(tdf)
    tdf = tdf[outliers == 1]
    print(r, tdf.shape)

print('-' * 80)
for r in range(10):
    tdf = read_file()
    tdf = engineer_features2(tdf)
    clf = IsolationForest(contamination=0.001, random_state=r, n_estimators=10
00, n_jobs=-1).fit(StandardScaler().fit_transform(tdf))
    outliers = clf.predict(tdf)
    tdf = tdf[outliers == 1]
    print(r, tdf.shape)
```

```
0 (155, 14)
1 (231, 14)
2 (386, 14)
3 (540, 14)
4 (323, 14)
5 (435, 14)
6 (147, 14)
7 (166, 14)
8 (320, 14)
9 (272, 14)
--------------------------------------------------------------------------------
---
0 (1658, 14)
1 (3905, 14)
2 (4221, 14)
3 (3715, 14)
4 (2616, 14)
5 (2760, 14)
6 (2492, 14)
7 (2296, 14)
8 (2865, 14)
9 (2856, 14)
```

We'll not use isolation forest because it is very sensitive to the random state, and with the same contamination it produces a vastly different output depending on the random value

In [272]:
```
def perform_grid_search(X, y, params, param_grid, description):
    p = imblearn.pipeline.Pipeline(params)
    cv1 = StratifiedKFold(n_splits=6, random_state=10, shuffle=True)
    scorer = make_scorer(metrics.f1_score)
    search = GridSearchCV(p, param_grid, n_jobs=7, cv=cv1, scoring=scorer)
    search.fit(X, y)
    print(description)
    print(f"Best Params = {search.best_params_}")
    print(f"Best score = {search.best_score_}")
    print('-' * 120)
    return search.best_estimator_
```

Feature Selection, PCA, SVC

In [288]:
```
def grid_search_svc_1(X, y):
    sm = SMOTE(random_state=1)
    svc = SVC(random_state=1)
    ms = MinMaxScaler()
    pca = PCA(random_state=0)
    rfecv = RFECV(estimator=RandomForestClassifier())
    params = [('sm', sm), ('pca', pca), ('rfe', rfecv), ('svc', svc)]
    param_grid = {
        'svc__kernel': ['rbf'],
        'svc__class_weight': [None],
        'svc__gamma': ['auto'],
        'sm__k_neighbors': [20],
        'pca__n_components': range(6, X.shape[1]+1)
    }
    return perform_grid_search(X, y, params, param_grid, description="PCA, RFE
CV, SVC")

random.seed(22)
selector = []
svc = grid_search_svc_1(scaled_X.to_numpy(), y.to_numpy())
```

```
PCA, RFECV, SVC
Best Params = {'pca__n_components': 10, 'sm__k_neighbors': 20, 'svc__class_we
ight': None, 'svc__gamma': 'auto', 'svc__kernel': 'rbf'}
Best score = 0.5963218633254905
------------------------------------------------------------------------------
----------------------------------------
```

In [289]:
```
print(svc.named_steps['rfe'].support_)
for i, j in zip(svc.named_steps['rfe'].support_, scaled_X.columns):
    print(j, i)
```

```
[ True  True  True  True  True  True  True  True  True  True]
CreditScore True
Age True
Tenure True
Balance True
NumOfProducts True
HasCrCard True
IsActiveMember True
EstimatedSalary True
Gender_Female True
Gender_Male True
```

Select the best parameters from above, and run them again, this time without feature selection

In [283]:
```
def grid_search_svc_1(X, y):
    sm = SMOTE(random_state=1)
    svc = SVC(random_state=1)
    ms = MinMaxScaler()
    pca = PCA(random_state=0)
    rfecv = RFECV(estimator=RandomForestClassifier())
    params = [('sm', sm), ('pca', pca), ('svc', svc)]
    param_grid = {
        'svc__kernel': ['rbf'],
        'svc__class_weight': [None],
        'svc__gamma': ['auto'],
        'sm__k_neighbors': [20],
        'pca__n_components': range(10, X.shape[1]+1)
    }
    return perform_grid_search(X, y, params, param_grid, description="PCA, SV
C")

random.seed(22)
selector = []
svc = grid_search_svc_1(scaled_X.to_numpy(), y.to_numpy())
```

```
PCA, RFECV, SVC
Best Params = {'pca__n_components': 10, 'sm__k_neighbors': 20, 'svc__class_we
ight': None, 'svc__gamma': 'auto', 'svc__kernel': 'rbf'}
Best score = 0.5963218633254905
--------------------------------------------------------------------------------
----------------------------------------
```

Do the same steps again, except use an SVC classifier for feature selection

In [286]:
```
def grid_search_svc_1(X, y):
    sm = SMOTE(random_state=1)
    svc = SVC(random_state=1)
    ms = MinMaxScaler()
    pca = PCA(random_state=0)
    rfecv = RFECV(estimator=DecisionTreeClassifier())
    params = [('sm', sm), ('pca', pca), ('rfe', rfecv), ('svc', svc)]
    param_grid = {
        'svc__kernel': ['rbf'],
        'svc__class_weight': [None],
        'svc__gamma': ['auto'],
        'sm__k_neighbors': [20],
        'pca__n_components': range(10, X.shape[1]+1)
    }
    return perform_grid_search(X, y, params, param_grid, description="PCA, RFE
CV, SVC")

random.seed(22)
selector = []
svc = grid_search_svc_1(scaled_X.to_numpy(), y.to_numpy())
```

```
PCA, RFECV, SVC
Best Params = {'pca__n_components': 13, 'sm__k_neighbors': 20, 'svc__class_we
ight': None, 'svc__gamma': 'auto', 'svc__kernel': 'rbf'}
Best score = 0.5967162189237114
--------------------------------------------------------------------------------
----------------------------------------
```

In [287]:
```
print(svc.named_steps['rfe'].support_)
for i, j in zip(svc.named_steps['rfe'].support_, scaled_X.columns):
    print(j, i)
```

```
[ True  True  True  True  True  True  True  True  True  True  True False
 False]
CreditScore True
Age True
Tenure True
Balance True
NumOfProducts True
HasCrCard True
IsActiveMember True
EstimatedSalary True
Gender_Female True
Gender_Male True
Geography_France True
Geography_Germany False
Geography_Spain False
```

Feature Selection, Random Forest

In [281]:
```
def grid_search_random_forest_fs(X, y):
    sm = SMOTE(random_state=1)
    rf = RandomForestClassifier(random_state=1, criterion='entropy', n_estimat
ors=1000, class_weight='balanced_subsample')
    rfecv = RFECV(estimator=RandomForestClassifier())
    param_grid = {'rf__min_samples_split': [100, 150]}
    params = [('sm', sm), ('rf', rf)]
    return perform_grid_search(X, y, params, param_grid, description="Random F
orest Classifier with Feature Selection")

random.seed(22)
rfcfs = grid_search_random_forest_fs(scaled_X.to_numpy(), y.to_numpy())
```

```
Random Forest Classifier with Feature Selection
Best Params = {'rf__min_samples_split': 100}
Best score = 0.624375225286892
--------------------------------------------------------------------------------
----------------------------------------
```

In [277]:
```
def grid_search_random_forest_fs(X, y):
    sm = SMOTE(random_state=1)
    rf = RandomForestClassifier(random_state=1, criterion='entropy', n_estimat
ors=1000, class_weight='balanced_subsample')
    rfecv = RFECV(estimator=RandomForestClassifier())
    param_grid = {'rf__min_samples_split': [100, 150]}
    params = [('sm', sm), ('rfecv', rfecv), ('rf', rf)]
    return perform_grid_search(X, y, params, param_grid, description="Random F
orest Classifier with Feature Selection")

random.seed(22)
rfcfs = grid_search_random_forest_fs(scaled_X.to_numpy(), y.to_numpy())
```

```
Random Forest Classifier with Feature Selection
Best Params = {'rf__min_samples_split': 100}
Best score = 0.6255208031561018
--------------------------------------------------------------------------------
----------------------------------------
```

In [280]:
```
print(svc.named_steps['rfe'].support_)
for i, j in zip(rfcfs.named_steps['rfecv'].support_, scaled_X.columns):
    print(j, i)
```

```
[ True  True  True  True  True  True  True  True  True  True]
CreditScore True
Age True
Tenure True
Balance True
NumOfProducts True
HasCrCard True
IsActiveMember True
EstimatedSalary True
Gender_Female True
Gender_Male True
Geography_France True
Geography_Germany True
Geography_Spain False
Index(['CreditScore', 'Age', 'Tenure', 'Balance', 'NumOfProducts', 'HasCrCar
d',
       'IsActiveMember', 'EstimatedSalary', 'Gender_Female', 'Gender_Male',
       'Geography_France', 'Geography_Germany', 'Geography_Spain'],
      dtype='object')
```

Compare Feature Selection Methods

In [164]:

```
def feature_select_score(X, y, selector, model_gen_fn, model_desc, selector_de
sc,\
                         min_max_scale=False, do_pca=False, n_pca=-1):
    p = []
    if min_max_scale:
        p.append(('ms', MinMaxScaler(),))
    p.append(('sm', SMOTE(),))
    if None != selector:
        p.append(('sel', selector(),))
    if do_pca:
        if -1 == n_pca:
            n_pca = 10
        p.append(('pca', PCA(n_components=n_pca),))
    p.append(('model', model_gen_fn(),))
    print(p)
    p = imblearn.pipeline.Pipeline(p)
    y_pred = cross_val_predict(p, X, y, \
                           cv=model_selection.StratifiedKFold(n_splits=6, shuf
fle=True, random_state=2),\
                           n_jobs=7)
    print(f"Selector: {selector_desc} Model: {model_desc}\n{metrics.classifica
tion_report(y, y_pred)}")
    return metrics.accuracy_score(y, y_pred), metrics.f1_score(y, y_pred),\
            metrics.classification_report(y, y_pred), metrics.confusion_matrix
(y, y_pred)

best_models_ = [
    ('GradientBoosting', lambda: GradientBoostingClassifier(random_state=0, ma
x_depth=3, min_samples_split=42, n_estimators=100)),
    ('RandomForest', lambda: RandomForestClassifier(class_weight='balanced_sub
sample', criterion='entropy', max_depth=None, min_samples_split=100, n_estimat
ors=1000, random_state=0)),
    ('SVC', lambda: SVC(class_weight=None, gamma='auto', kernel='rbf', random_
state=0)),
]


all_feature_selectors_ = [
    ('LinearSVC', lambda: SelectFromModel(LinearSVC())),
    ('LinearSVC L1', lambda: SelectFromModel(LinearSVC(penalty="l1"))),
    ('Variance Threshold 0.8', lambda: VarianceThreshold(threshold=(.8 * (1 -
.8)))),
    ('RFECV RandomForest', lambda: RFECV(estimator=RandomForestClassifier())),
    ('FWE chi2', lambda: SelectFwe(chi2, alpha=0.01)),
    ('FWE f_classif', lambda: SelectFwe(f_classif, alpha=0.01)),
    ('Tree Based Feature Selection', lambda: SelectFromModel(ExtraTreesClassif
ier(n_estimators=50))),
    ('NONE', None),
]

def compare_selectors_with_models(models, selectors,\
                                  min_max_scale=False, do_pca=False, n_pca=-1):
    """
    For each combination of model and feature selector, run cross validation a
nd aggregate
    the accuracy and f1 scores
```

```
    """
    all_models = {}
    desc_string = ""
    for model_desc, model_gen in models:
        accuracy_scores = []
        f1_scores = []
        selector_descs = []
        for selector_desc, selector in selectors:
            acc = 0
            f1 = 0
            try:
                acc, f1, report, conf_mat = feature_select_score(scaled_X.to_n
umpy(), y.to_numpy(),\
                                                 selector, model_gen, model_desc
, selector_desc,\
                                                  min_max_scale, do_pca, n_pca)
                accuracy_scores.append(acc)
                f1_scores.append(f1)
            except:
                print(f'====>>>> Failure in {model_desc}, {selector_desc}')
                accuracy_scores.append(acc)
                f1_scores.append(f1)
            selector_descs.append(selector_desc)
        sorted_indices = np.argsort(f1_scores)
        desc_string += "\nMODEL: %s\n----------------------\n" % model_desc
        for i in sorted_indices:
            desc_string += ("%40s %10.5f %10.5f\n" % (selector_descs[i], accur
acy_scores[i], f1_scores[i]))
        desc_string += '----------------------------------------------------
------------------------\n'
        print(desc_string)
        all_models[desc_string] = {'accuracy': accuracy_scores, 'f1': f1_score
s, 'selectors': selector_descs}
    return all_models, desc_string
```

In [124]:
```
all_models, desc_string = compare_selectors_with_models(best_models_, all_feat
ure_selectors_)
```

```
Selector: LinearSVC Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.88      0.86      0.87      7963
           1       0.50      0.56      0.53      2037

    accuracy                           0.80     10000
   macro avg       0.69      0.71      0.70     10000
weighted avg       0.81      0.80      0.80     10000

====>>>> Failure in GradientBoosting, LinearSVC L1
Selector: Variance Threshold 0.8 Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.91      0.89      0.90      7963
           1       0.61      0.65      0.63      2037

    accuracy                           0.84     10000
   macro avg       0.76      0.77      0.76     10000
weighted avg       0.85      0.84      0.84     10000

Selector: RFECV RandomForest Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.91      0.89      0.90      7963
           1       0.61      0.65      0.63      2037

    accuracy                           0.84     10000
   macro avg       0.76      0.77      0.76     10000
weighted avg       0.85      0.84      0.85     10000

====>>>> Failure in GradientBoosting, FWE chi2
Selector: FWE f_classif Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.91      0.87      0.89      7963
           1       0.57      0.68      0.62      2037

    accuracy                           0.83     10000
   macro avg       0.74      0.78      0.76     10000
weighted avg       0.85      0.83      0.84     10000

Selector: Tree Based Feature Selection Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.90      0.89      0.89      7963
           1       0.58      0.61      0.60      2037

    accuracy                           0.83     10000
   macro avg       0.74      0.75      0.74     10000
weighted avg       0.83      0.83      0.83     10000

Selector: NONE Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.91      0.89      0.90      7963
           1       0.61      0.65      0.63      2037
```

```
    accuracy                           0.84     10000
   macro avg       0.76      0.77      0.76     10000
weighted avg       0.85      0.84      0.85     10000


MODEL: GradientBoosting
----------------------
                           LinearSVC L1    0.00000    0.00000
                               FWE chi2    0.00000    0.00000
                              LinearSVC    0.79660    0.52938
            Tree Based Feature Selection    0.83160    0.59636
                           FWE f_classif    0.83210    0.62396
                                    NONE    0.84340    0.62750
                  Variance Threshold 0.8    0.84270    0.62822
                      RFECV RandomForest    0.84300    0.62831
--------------------------------------------------------------------------------
--

Selector: LinearSVC Model: RandomForest
              precision    recall  f1-score   support

           0       0.87      0.89      0.88      7963
           1       0.53      0.50      0.52      2037

    accuracy                           0.81     10000
   macro avg       0.70      0.69      0.70     10000
weighted avg       0.80      0.81      0.81     10000

====>>>> Failure in RandomForest, LinearSVC L1
Selector: Variance Threshold 0.8 Model: RandomForest
              precision    recall  f1-score   support

           0       0.91      0.88      0.89      7963
           1       0.59      0.66      0.62      2037

    accuracy                           0.84     10000
   macro avg       0.75      0.77      0.76     10000
weighted avg       0.84      0.84      0.84     10000

Selector: RFECV RandomForest Model: RandomForest
              precision    recall  f1-score   support

           0       0.91      0.88      0.90      7963
           1       0.59      0.66      0.62      2037

    accuracy                           0.84     10000
   macro avg       0.75      0.77      0.76     10000
weighted avg       0.84      0.84      0.84     10000

====>>>> Failure in RandomForest, FWE chi2
Selector: FWE f_classif Model: RandomForest
              precision    recall  f1-score   support

           0       0.91      0.87      0.89      7963
           1       0.58      0.68      0.63      2037
```

```
    accuracy                           0.84     10000
   macro avg       0.75      0.78      0.76     10000
weighted avg       0.85      0.84      0.84     10000

Selector: Tree Based Feature Selection Model: RandomForest
              precision    recall  f1-score   support

           0       0.90      0.87      0.89      7963
           1       0.56      0.63      0.59      2037

    accuracy                           0.82     10000
   macro avg       0.73      0.75      0.74     10000
weighted avg       0.83      0.82      0.83     10000

Selector: NONE Model: RandomForest
              precision    recall  f1-score   support

           0       0.91      0.88      0.89      7963
           1       0.59      0.66      0.62      2037

    accuracy                           0.84     10000
   macro avg       0.75      0.77      0.76     10000
weighted avg       0.84      0.84      0.84     10000


MODEL: GradientBoosting
----------------------
                           LinearSVC L1    0.00000    0.00000
                               FWE chi2    0.00000    0.00000
                              LinearSVC    0.79660    0.52938
            Tree Based Feature Selection    0.83160    0.59636
                           FWE f_classif    0.83210    0.62396
                                    NONE    0.84340    0.62750
                 Variance Threshold 0.8    0.84270    0.62822
                     RFECV RandomForest    0.84300    0.62831
------------------------------------------------------------------------------
--

MODEL: RandomForest
----------------------
                           LinearSVC L1    0.00000    0.00000
                               FWE chi2    0.00000    0.00000
                              LinearSVC    0.80910    0.51511
            Tree Based Feature Selection    0.82420    0.59230
                                    NONE    0.83520    0.61834
                 Variance Threshold 0.8    0.83520    0.61852
                     RFECV RandomForest    0.83620    0.62013
                           FWE f_classif    0.83510    0.62684
------------------------------------------------------------------------------
--

Selector: LinearSVC Model: SVC
              precision    recall  f1-score   support

           0       0.90      0.80      0.85      7963
           1       0.45      0.64      0.53      2037
```

```
    accuracy                           0.77     10000
   macro avg       0.67      0.72      0.69     10000
weighted avg       0.81      0.77      0.78     10000

====>>>> Failure in SVC, LinearSVC L1
Selector: Variance Threshold 0.8 Model: SVC
              precision    recall  f1-score   support

           0       0.92      0.82      0.87      7963
           1       0.50      0.72      0.59      2037

    accuracy                           0.80     10000
   macro avg       0.71      0.77      0.73     10000
weighted avg       0.84      0.80      0.81     10000

Selector: RFECV RandomForest Model: SVC
              precision    recall  f1-score   support

           0       0.92      0.82      0.87      7963
           1       0.50      0.72      0.59      2037

    accuracy                           0.80     10000
   macro avg       0.71      0.77      0.73     10000
weighted avg       0.83      0.80      0.81     10000

====>>>> Failure in SVC, FWE chi2
Selector: FWE f_classif Model: SVC
              precision    recall  f1-score   support

           0       0.92      0.82      0.87      7963
           1       0.51      0.73      0.60      2037

    accuracy                           0.80     10000
   macro avg       0.71      0.77      0.73     10000
weighted avg       0.84      0.80      0.81     10000

Selector: Tree Based Feature Selection Model: SVC
              precision    recall  f1-score   support

           0       0.91      0.80      0.85      7963
           1       0.47      0.71      0.57      2037

    accuracy                           0.78     10000
   macro avg       0.69      0.75      0.71     10000
weighted avg       0.82      0.78      0.79     10000

Selector: NONE Model: SVC
              precision    recall  f1-score   support

           0       0.92      0.82      0.87      7963
           1       0.50      0.72      0.59      2037

    accuracy                           0.80     10000
   macro avg       0.71      0.77      0.73     10000
weighted avg       0.83      0.80      0.81     10000
```

MODEL: GradientBoosting

LinearSVC L1	0.00000	0.00000
FWE chi2	0.00000	0.00000
LinearSVC	0.79660	0.52938
Tree Based Feature Selection	0.83160	0.59636
FWE f_classif	0.83210	0.62396
NONE	0.84340	0.62750
Variance Threshold 0.8	0.84270	0.62822
RFECV RandomForest	0.84300	0.62831

MODEL: RandomForest

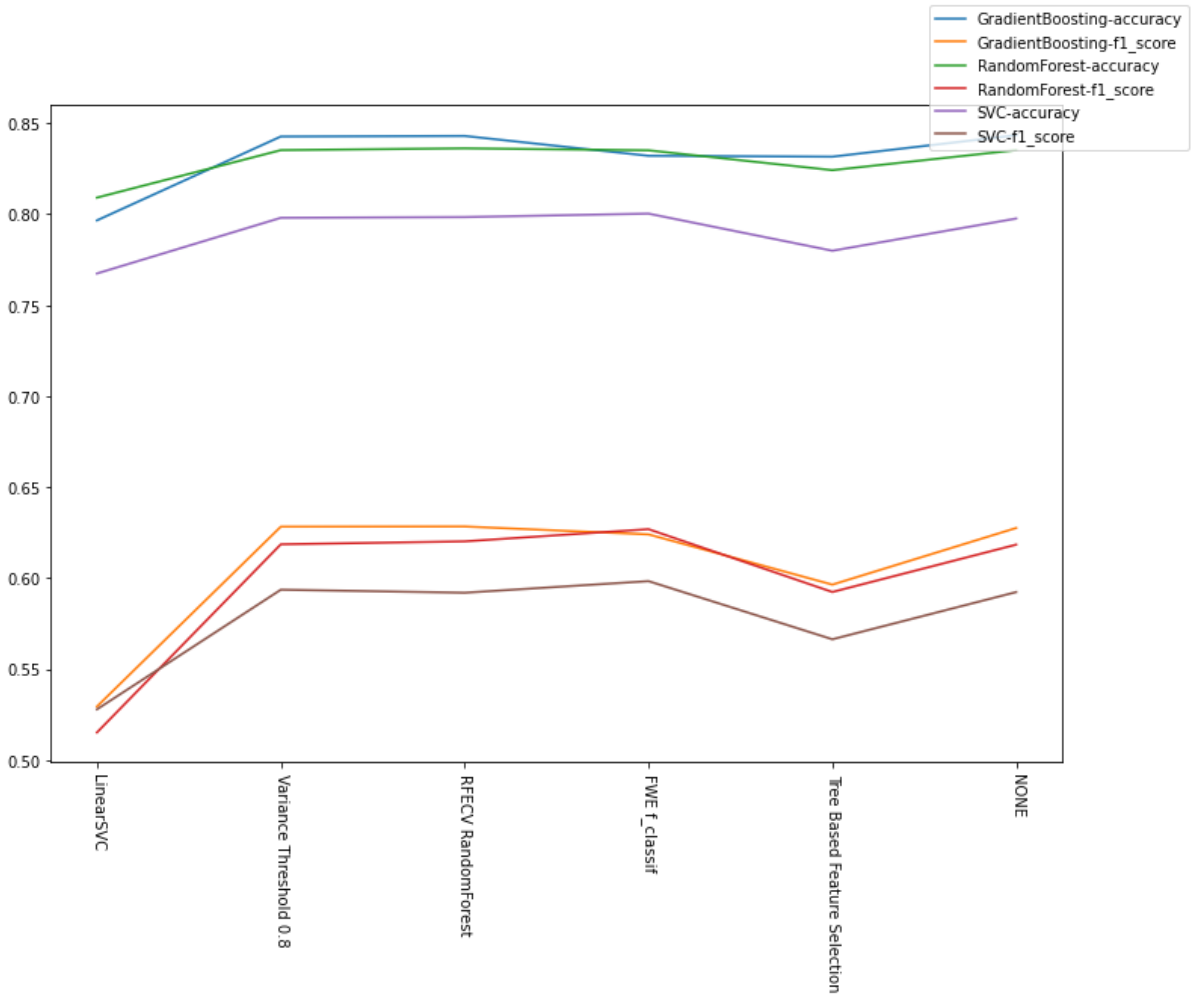
LinearSVC L1	0.00000	0.00000
FWE chi2	0.00000	0.00000
LinearSVC	0.80910	0.51511
Tree Based Feature Selection	0.82420	0.59230
NONE	0.83520	0.61834
Variance Threshold 0.8	0.83520	0.61852
RFECV RandomForest	0.83620	0.62013
FWE f_classif	0.83510	0.62684

MODEL: SVC

LinearSVC L1	0.00000	0.00000
FWE chi2	0.00000	0.00000
LinearSVC	0.76740	0.52781
Tree Based Feature Selection	0.77990	0.56631
RFECV RandomForest	0.79840	0.59190
NONE	0.79760	0.59226
Variance Threshold 0.8	0.79800	0.59356
FWE f_classif	0.80030	0.59827

In [161]:

```
def plot_graph_1(all_models):
    """
    Plot the accuracy and F1 scores from the data output by the previous step
    """
    am = ['GradientBoosting', 'RandomForest', 'SVC']

    i = 0
    accuracy_scores = None
    f1_scores = None
    selector_descs = None
    fig, ax = plt.subplots()

    for key, val in all_models.items():
        model = am[i]
        i+=1
        accuracy_scores = val['accuracy']
        f1_scores = val['f1']
        selector_descs = val['selectors']
        new_f1 = []
        new_acc = []
        new_descs = []
        for acc, f1, d in zip(accuracy_scores, f1_scores, selector_descs):
            if 0 != f1:
                new_f1.append(f1)
                new_acc.append(acc)
                new_descs.append(d)
        accuracy_scores = new_acc
        f1_scores = new_f1
        selector_descs = new_descs
        ax.plot(accuracy_scores, label="%s-accuracy" % model)
        ax.plot(f1_scores, label="%s-f1_score" % model)
    plt.xticks(np.arange(0, len(f1_scores), 1.0))
    ax.set_xticklabels(selector_descs, rotation=270)
    fig.legend()

plot_graph_1(all_models)
```


GradientBoosting-accuracy
GradientBoosting-f1_score
RandomForest-accuracy
RandomForest-f1_score
SVC-accuracy
SVC-f1_score
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
LinearSVC
Variance Threshold 0.8
RFECV RandomForest
FWE f_classif
Tree Based Feature Selection
NONE

In [126]:
```
print("Best models with Feature Selection")
print(desc_string)
```

```
Best models with Feature Selection

MODEL: GradientBoosting
----------------------
                            LinearSVC L1    0.00000    0.00000
                                FWE chi2    0.00000    0.00000
                              LinearSVC     0.79660    0.52938
            Tree Based Feature Selection    0.83160    0.59636
                           FWE f_classif    0.83210    0.62396
                                    NONE    0.84340    0.62750
                  Variance Threshold 0.8    0.84270    0.62822
                      RFECV RandomForest    0.84300    0.62831
-------------------------------------------------------------------------------
--

MODEL: RandomForest
----------------------
                            LinearSVC L1    0.00000    0.00000
                                FWE chi2    0.00000    0.00000
                              LinearSVC     0.80910    0.51511
            Tree Based Feature Selection    0.82420    0.59230
                                    NONE    0.83520    0.61834
                  Variance Threshold 0.8    0.83520    0.61852
                      RFECV RandomForest    0.83620    0.62013
                           FWE f_classif    0.83510    0.62684
-------------------------------------------------------------------------------
--

MODEL: SVC
----------------------
                            LinearSVC L1    0.00000    0.00000
                                FWE chi2    0.00000    0.00000
                              LinearSVC     0.76740    0.52781
            Tree Based Feature Selection    0.77990    0.56631
                      RFECV RandomForest    0.79840    0.59190
                                    NONE    0.79760    0.59226
                  Variance Threshold 0.8    0.79800    0.59356
                           FWE f_classif    0.80030    0.59827
-------------------------------------------------------------------------------
--
```

The best feature selectors are:

1. FWE f_classif
2. NONE
3. RFECV Random Forest
4. Variance Threshold 0.8

However the difference is not much

In [135]:
```
best_selectors_ = [
    ('Variance Threshold 0.8', lambda: VarianceThreshold(threshold=(.8 * (1 -
.8)))),
    ('RFECV RandomForest', lambda: RFECV(estimator=RandomForestClassifier())),
    ('FWE f_classif', lambda: SelectFwe(f_classif, alpha=0.01)),
    ('NONE', None),
]
```

Check if MinMaxScaling makes a difference

See if MinMax Scaling makes a difference. Standardization has already been performed earlier with unit SD

In [158]:
```
all_models_mms, desc_string_mms = compare_selectors_with_models(best_models_,
best_feature_selectors_,\
                                                                min_max_scale=True, do
_pca=False, n_pca=-1)
```

```
[('ms', MinMaxScaler()), ('sm', SMOTE()), ('sel', VarianceThreshold(threshold
=0.15999999999999998)), ('model', GradientBoostingClassifier(min_samples_spli
t=42, random_state=0))]
Selector: Variance Threshold 0.8 Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.86      0.69      0.77      7963
           1       0.32      0.57      0.41      2037

    accuracy                           0.66     10000
   macro avg       0.59      0.63      0.59     10000
weighted avg       0.75      0.66      0.69     10000

[('ms', MinMaxScaler()), ('sm', SMOTE()), ('sel', RFECV(estimator=RandomFores
tClassifier())), ('model', GradientBoostingClassifier(min_samples_split=42, r
andom_state=0))]
Selector: RFECV RandomForest Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.91      0.88      0.89      7963
           1       0.58      0.67      0.62      2037

    accuracy                           0.83     10000
   macro avg       0.75      0.77      0.76     10000
weighted avg       0.84      0.83      0.84     10000

[('ms', MinMaxScaler()), ('sm', SMOTE()), ('sel', SelectFwe(alpha=0.01)), ('m
odel', GradientBoostingClassifier(min_samples_split=42, random_state=0))]
Selector: FWE f_classif Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.92      0.86      0.89      7963
           1       0.56      0.70      0.62      2037

    accuracy                           0.83     10000
   macro avg       0.74      0.78      0.76     10000
weighted avg       0.85      0.83      0.83     10000

[('ms', MinMaxScaler()), ('sm', SMOTE()), ('model', GradientBoostingClassifie
r(min_samples_split=42, random_state=0))]
Selector: NONE Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.91      0.88      0.89      7963
           1       0.58      0.67      0.62      2037

    accuracy                           0.84     10000
   macro avg       0.75      0.77      0.76     10000
weighted avg       0.84      0.84      0.84     10000


MODEL: GradientBoosting
----------------------
                  Variance Threshold 0.8    0.66370    0.40782
                                    NONE    0.83510    0.62170
                       RFECV RandomForest    0.83440    0.62278
                            FWE f_classif    0.82730    0.62383
```

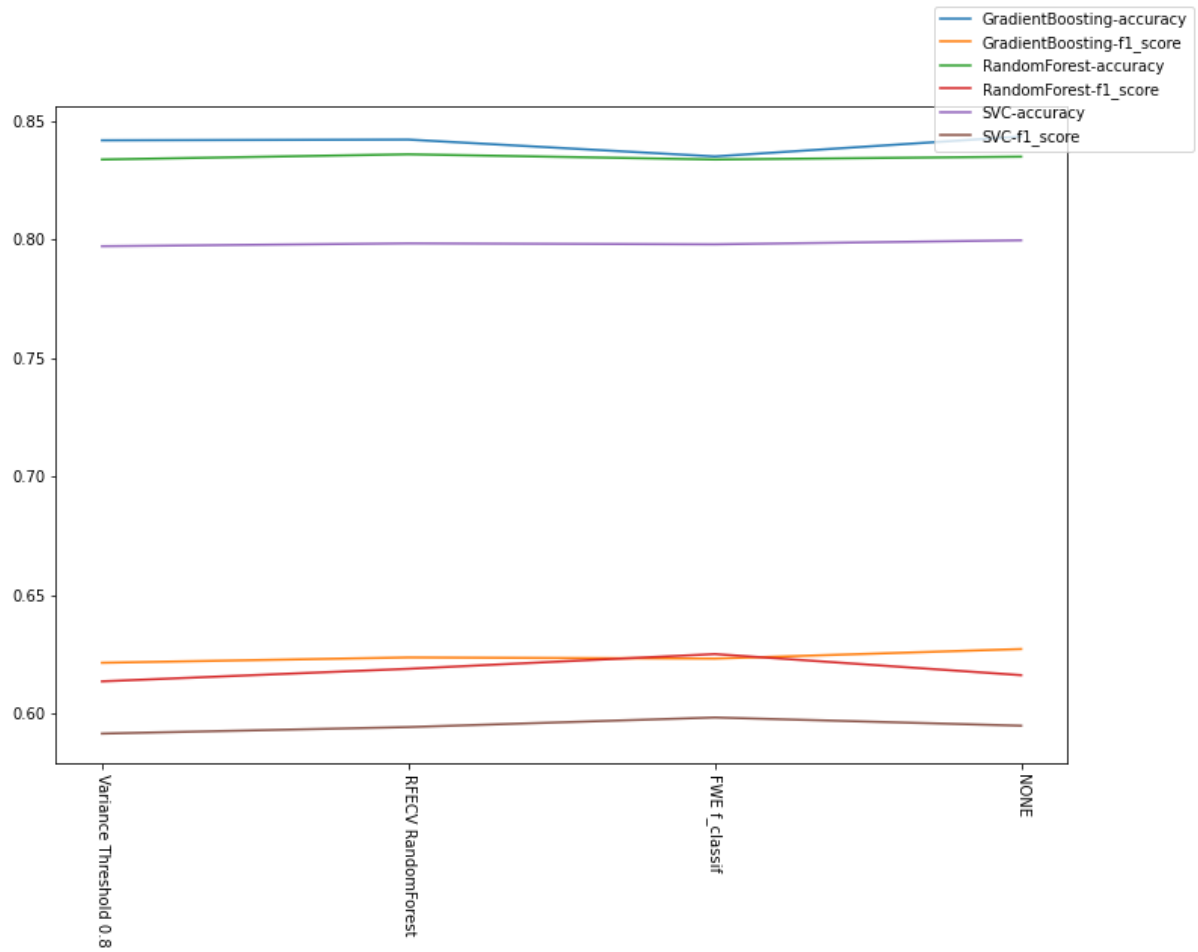
```
--------------------------------------------------------------------------------

[('ms', MinMaxScaler()), ('sm', SMOTE()), ('sel', VarianceThreshold(threshold
=0.15999999999999998)), ('model', RandomForestClassifier(class_weight='balanc
ed_subsample', criterion='entropy',
                       min_samples_split=100, n_estimators=1000,
                       random_state=0))]
Selector: Variance Threshold 0.8 Model: RandomForest
              precision    recall  f1-score   support

           0       0.86      0.68      0.76      7963
           1       0.32      0.58      0.41      2037

    accuracy                           0.66     10000
   macro avg       0.59      0.63      0.58     10000
weighted avg       0.75      0.66      0.69     10000

[('ms', MinMaxScaler()), ('sm', SMOTE()), ('sel', RFECV(estimator=RandomFores
tClassifier())), ('model', RandomForestClassifier(class_weight='balanced_subs
ample', criterion='entropy',
                       min_samples_split=100, n_estimators=1000,
                       random_state=0))]
Selector: RFECV RandomForest Model: RandomForest
              precision    recall  f1-score   support

           0       0.92      0.85      0.88      7963
           1       0.55      0.69      0.61      2037

    accuracy                           0.82     10000
   macro avg       0.73      0.77      0.75     10000
weighted avg       0.84      0.82      0.83     10000

[('ms', MinMaxScaler()), ('sm', SMOTE()), ('sel', SelectFwe(alpha=0.01)), ('m
odel', RandomForestClassifier(class_weight='balanced_subsample', criterion='e
ntropy',
                       min_samples_split=100, n_estimators=1000,
                       random_state=0))]
Selector: FWE f_classif Model: RandomForest
              precision    recall  f1-score   support

           0       0.92      0.85      0.88      7963
           1       0.54      0.71      0.61      2037

    accuracy                           0.82     10000
   macro avg       0.73      0.78      0.75     10000
weighted avg       0.84      0.82      0.83     10000

[('ms', MinMaxScaler()), ('sm', SMOTE()), ('model', RandomForestClassifier(cl
ass_weight='balanced_subsample', criterion='entropy',
                       min_samples_split=100, n_estimators=1000,
                       random_state=0))]
Selector: NONE Model: RandomForest
              precision    recall  f1-score   support

           0       0.92      0.85      0.88      7963
           1       0.55      0.69      0.61      2037
```

```
    accuracy                           0.82     10000
   macro avg       0.73      0.77      0.75     10000
weighted avg       0.84      0.82      0.83     10000


MODEL: GradientBoosting
----------------------
                 Variance Threshold 0.8    0.66370    0.40782
                                   NONE    0.83510    0.62170
                      RFECV RandomForest    0.83440    0.62278
                           FWE f_classif    0.82730    0.62383
--------------------------------------------------------------------------------
--

MODEL: RandomForest
----------------------
                 Variance Threshold 0.8    0.65870    0.40798
                           FWE f_classif    0.81750    0.61162
                                   NONE    0.82070    0.61182
                      RFECV RandomForest    0.82180    0.61278
--------------------------------------------------------------------------------
--

[('ms', MinMaxScaler()), ('sm', SMOTE()), ('sel', VarianceThreshold(threshold
=0.15999999999999998)), ('model', SVC(gamma='auto', random_state=0))]
Selector: Variance Threshold 0.8 Model: SVC
              precision    recall  f1-score   support

           0       0.87      0.65      0.74      7963
           1       0.31      0.61      0.41      2037

    accuracy                           0.64     10000
   macro avg       0.59      0.63      0.57     10000
weighted avg       0.75      0.64      0.67     10000

[('ms', MinMaxScaler()), ('sm', SMOTE()), ('sel', RFECV(estimator=RandomFores
tClassifier())), ('model', SVC(gamma='auto', random_state=0))]
Selector: RFECV RandomForest Model: SVC
              precision    recall  f1-score   support

           0       0.92      0.76      0.83      7963
           1       0.44      0.73      0.54      2037

    accuracy                           0.75     10000
   macro avg       0.68      0.74      0.69     10000
weighted avg       0.82      0.75      0.77     10000

[('ms', MinMaxScaler()), ('sm', SMOTE()), ('sel', SelectFwe(alpha=0.01)), ('m
odel', SVC(gamma='auto', random_state=0))]
Selector: FWE f_classif Model: SVC
              precision    recall  f1-score   support

           0       0.92      0.77      0.84      7963
           1       0.45      0.73      0.56      2037

    accuracy                           0.77     10000
```

```
   macro avg       0.69      0.75      0.70     10000
weighted avg       0.82      0.77      0.78     10000

[('ms', MinMaxScaler()), ('sm', SMOTE()), ('model', SVC(gamma='auto', random_
state=0))]
Selector: NONE Model: SVC
              precision    recall  f1-score   support

           0       0.92      0.76      0.83      7963
           1       0.43      0.73      0.54      2037

    accuracy                           0.75     10000
   macro avg       0.67      0.74      0.69     10000
weighted avg       0.82      0.75      0.77     10000


MODEL: GradientBoosting
----------------------
                 Variance Threshold 0.8    0.66370    0.40782
                                   NONE    0.83510    0.62170
                     RFECV RandomForest    0.83440    0.62278
                          FWE f_classif    0.82730    0.62383
--------------------------------------------------------------------------------
--

MODEL: RandomForest
----------------------
                 Variance Threshold 0.8    0.65870    0.40798
                          FWE f_classif    0.81750    0.61162
                                   NONE    0.82070    0.61182
                     RFECV RandomForest    0.82180    0.61278
--------------------------------------------------------------------------------
--

MODEL: SVC
----------------------
                 Variance Threshold 0.8    0.64010    0.40660
                                   NONE    0.75030    0.54259
                     RFECV RandomForest    0.75240    0.54401
                          FWE f_classif    0.76500    0.55976
--------------------------------------------------------------------------------
--
```

In [162]:
```
plot_graph_1(all_models_mms)
print(desc_string_mms)
```

```
MODEL: GradientBoosting
----------------------
                  Variance Threshold 0.8    0.84200    0.62147
                                FWE f_classif    0.83520    0.62323
                           RFECV RandomForest    0.84230    0.62372
                                        NONE    0.84360    0.62726
--------------------------------------------------------------------------------
--

MODEL: RandomForest
----------------------
                  Variance Threshold 0.8    0.83390    0.61363
                                        NONE    0.83510    0.61624
                           RFECV RandomForest    0.83610    0.61893
                                FWE f_classif    0.83390    0.62514
--------------------------------------------------------------------------------
--

MODEL: SVC
----------------------
                  Variance Threshold 0.8    0.79730    0.59158
                           RFECV RandomForest    0.79850    0.59432
                                        NONE    0.79980    0.59490
                                FWE f_classif    0.79810    0.59837
--------------------------------------------------------------------------------
--
```



Add PCA to the mix

In [167]:

```
all_models_mmspca_a = []
desc_strings_mmspca_a = []
ncap_a = []
for npca in range(5, 15):
    print('=' * 120)
    print(f'PCA {npca}')
    print('-' * 80)
    all_models_mmspca, desc_string_mmspca = compare_selectors_with_models(best
_models_, best_feature_selectors_,\
                                                                          min_max_scale=Fals
e, do_pca=True, n_pca=10)
    all_models_mmspca_a.append(all_models_mmspca)
    desc_strings_mmspca_a.append(desc_string_mmspca)
    ncap_a.append(npca)
```

```
================================================================================
========================================
PCA 5
--------------------------------------------------------------------------------
---
[('sm', SMOTE()), ('sel', VarianceThreshold(threshold=0.15999999999999998)),
('pca', PCA(n_components=10)), ('model', GradientBoostingClassifier(min_sampl
es_split=42, random_state=0))]
Selector: Variance Threshold 0.8 Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.92      0.81      0.86      7963
           1       0.49      0.72      0.59      2037

    accuracy                           0.79     10000
   macro avg       0.71      0.77      0.72     10000
weighted avg       0.83      0.79      0.81     10000

[('sm', SMOTE()), ('sel', RFECV(estimator=RandomForestClassifier())), ('pca',
PCA(n_components=10)), ('model', GradientBoostingClassifier(min_samples_split
=42, random_state=0))]
Selector: RFECV RandomForest Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.92      0.83      0.87      7963
           1       0.51      0.70      0.59      2037

    accuracy                           0.80     10000
   macro avg       0.71      0.77      0.73     10000
weighted avg       0.83      0.80      0.81     10000

[('sm', SMOTE()), ('sel', SelectFwe(alpha=0.01)), ('pca', PCA(n_components=1
0)), ('model', GradientBoostingClassifier(min_samples_split=42, random_state=
0))]
====>>>> Failure in GradientBoosting, FWE f_classif
[('sm', SMOTE()), ('pca', PCA(n_components=10)), ('model', GradientBoostingCl
assifier(min_samples_split=42, random_state=0))]
Selector: NONE Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.92      0.81      0.86      7963
           1       0.49      0.73      0.59      2037

    accuracy                           0.79     10000
   macro avg       0.71      0.77      0.72     10000
weighted avg       0.83      0.79      0.80     10000


MODEL: GradientBoosting
----------------------
                          FWE f_classif    0.00000    0.00000
                                   NONE    0.78990    0.58552
                Variance Threshold 0.8    0.79250    0.58690
                    RFECV RandomForest    0.80290    0.59285
--------------------------------------------------------------------------------
--
```

```
[('sm', SMOTE()), ('sel', VarianceThreshold(threshold=0.15999999999999998)),
('pca', PCA(n_components=10)), ('model', RandomForestClassifier(class_weight
='balanced_subsample', criterion='entropy',
                       min_samples_split=100, n_estimators=1000,
                       random_state=0))]
Selector: Variance Threshold 0.8 Model: RandomForest
              precision    recall  f1-score   support

           0       0.91      0.84      0.88      7963
           1       0.52      0.69      0.60      2037

    accuracy                           0.81     10000
   macro avg       0.72      0.76      0.74     10000
weighted avg       0.83      0.81      0.82     10000

[('sm', SMOTE()), ('sel', RFECV(estimator=RandomForestClassifier())), ('pca',
PCA(n_components=10)), ('model', RandomForestClassifier(class_weight='balance
d_subsample', criterion='entropy',
                       min_samples_split=100, n_estimators=1000,
                       random_state=0))]
Selector: RFECV RandomForest Model: RandomForest
              precision    recall  f1-score   support

           0       0.91      0.85      0.88      7963
           1       0.53      0.69      0.60      2037

    accuracy                           0.81     10000
   macro avg       0.72      0.77      0.74     10000
weighted avg       0.84      0.81      0.82     10000

[('sm', SMOTE()), ('sel', SelectFwe(alpha=0.01)), ('pca', PCA(n_components=1
0)), ('model', RandomForestClassifier(class_weight='balanced_subsample', crit
erion='entropy',
                       min_samples_split=100, n_estimators=1000,
                       random_state=0))]
====>>>> Failure in RandomForest, FWE f_classif
[('sm', SMOTE()), ('pca', PCA(n_components=10)), ('model', RandomForestClassi
fier(class_weight='balanced_subsample', criterion='entropy',
                       min_samples_split=100, n_estimators=1000,
                       random_state=0))]
Selector: NONE Model: RandomForest
              precision    recall  f1-score   support

           0       0.92      0.84      0.88      7963
           1       0.53      0.70      0.61      2037

    accuracy                           0.81     10000
   macro avg       0.72      0.77      0.74     10000
weighted avg       0.84      0.81      0.82     10000


MODEL: GradientBoosting
----------------------
                          FWE f_classif    0.00000    0.00000
                                   NONE    0.78990    0.58552
                  Variance Threshold 0.8    0.79250    0.58690
                      RFECV RandomForest    0.80290    0.59285
```

```
--------------------------------------------------------------------------------
--

MODEL: RandomForest
--------------------
                          FWE f_classif    0.00000    0.00000
                Variance Threshold 0.8    0.80930    0.59537
                    RFECV RandomForest    0.81420    0.60214
                                  NONE    0.81330    0.60604
--------------------------------------------------------------------------------
--

[('sm', SMOTE()), ('sel', VarianceThreshold(threshold=0.15999999999999998)),
('pca', PCA(n_components=10)), ('model', SVC(gamma='auto', random_state=0))]
Selector: Variance Threshold 0.8 Model: SVC
              precision    recall  f1-score   support

           0       0.92      0.81      0.86      7963
           1       0.50      0.73      0.59      2037

    accuracy                           0.80     10000
   macro avg       0.71      0.77      0.73     10000
weighted avg       0.83      0.80      0.81     10000

[('sm', SMOTE()), ('sel', RFECV(estimator=RandomForestClassifier())), ('pca',
PCA(n_components=10)), ('model', SVC(gamma='auto', random_state=0))]
Selector: RFECV RandomForest Model: SVC
              precision    recall  f1-score   support

           0       0.92      0.81      0.86      7963
           1       0.50      0.72      0.59      2037

    accuracy                           0.80     10000
   macro avg       0.71      0.77      0.73     10000
weighted avg       0.83      0.80      0.81     10000

[('sm', SMOTE()), ('sel', SelectFwe(alpha=0.01)), ('pca', PCA(n_components=1
0)), ('model', SVC(gamma='auto', random_state=0))]
====>>>> Failure in SVC, FWE f_classif
[('sm', SMOTE()), ('pca', PCA(n_components=10)), ('model', SVC(gamma='auto',
random_state=0))]
Selector: NONE Model: SVC
              precision    recall  f1-score   support

           0       0.92      0.82      0.87      7963
           1       0.50      0.73      0.59      2037

    accuracy                           0.80     10000
   macro avg       0.71      0.77      0.73     10000
weighted avg       0.84      0.80      0.81     10000


MODEL: GradientBoosting
--------------------
                          FWE f_classif    0.00000    0.00000
                                  NONE    0.78990    0.58552
                Variance Threshold 0.8    0.79250    0.58690
```

```
                     RFECV RandomForest    0.80290    0.59285
--------------------------------------------------------------------------------

MODEL: RandomForest
--------------------
                           FWE f_classif    0.00000    0.00000
                 Variance Threshold 0.8    0.80930    0.59537
                     RFECV RandomForest    0.81420    0.60214
                                   NONE    0.81330    0.60604
--------------------------------------------------------------------------------

MODEL: SVC
--------------------
                           FWE f_classif    0.00000    0.00000
                     RFECV RandomForest    0.79550    0.58960
                 Variance Threshold 0.8    0.79550    0.59125
                                   NONE    0.79820    0.59478
--------------------------------------------------------------------------------

================================================================================
==========================================
PCA 6
--------------------------------------------------------------------------------
---
[('sm', SMOTE()), ('sel', VarianceThreshold(threshold=0.15999999999999998)),
('pca', PCA(n_components=10)), ('model', GradientBoostingClassifier(min_sampl
es_split=42, random_state=0))]
Selector: Variance Threshold 0.8 Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.92      0.81      0.86      7963
           1       0.49      0.72      0.59      2037

    accuracy                           0.79     10000
   macro avg       0.71      0.77      0.72     10000
weighted avg       0.83      0.79      0.81     10000

[('sm', SMOTE()), ('sel', RFECV(estimator=RandomForestClassifier())), ('pca',
PCA(n_components=10)), ('model', GradientBoostingClassifier(min_samples_split
=42, random_state=0))]
Selector: RFECV RandomForest Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.92      0.81      0.86      7963
           1       0.49      0.72      0.58      2037

    accuracy                           0.79     10000
   macro avg       0.70      0.76      0.72     10000
weighted avg       0.83      0.79      0.80     10000

[('sm', SMOTE()), ('sel', SelectFwe(alpha=0.01)), ('pca', PCA(n_components=1
0)), ('model', GradientBoostingClassifier(min_samples_split=42, random_state=
0))]
====>>>> Failure in GradientBoosting, FWE f_classif
```

```
[('sm', SMOTE()), ('pca', PCA(n_components=10)), ('model', GradientBoostingCl
assifier(min_samples_split=42, random_state=0))]
Selector: NONE Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.92      0.81      0.86      7963
           1       0.49      0.72      0.58      2037

    accuracy                           0.79     10000
   macro avg       0.71      0.76      0.72     10000
weighted avg       0.83      0.79      0.80     10000


MODEL: GradientBoosting
----------------------
                          FWE f_classif    0.00000    0.00000
                     RFECV RandomForest    0.79120    0.58390
                                   NONE    0.79150    0.58458
                  Variance Threshold 0.8    0.79280    0.58626
--------------------------------------------------------------------------------

[('sm', SMOTE()), ('sel', VarianceThreshold(threshold=0.15999999999999998)),
('pca', PCA(n_components=10)), ('model', RandomForestClassifier(class_weight
='balanced_subsample', criterion='entropy',
                       min_samples_split=100, n_estimators=1000,
                       random_state=0))]
Selector: Variance Threshold 0.8 Model: RandomForest
              precision    recall  f1-score   support

           0       0.91      0.84      0.88      7963
           1       0.53      0.69      0.60      2037

    accuracy                           0.81     10000
   macro avg       0.72      0.77      0.74     10000
weighted avg       0.83      0.81      0.82     10000

[('sm', SMOTE()), ('sel', RFECV(estimator=RandomForestClassifier())), ('pca',
PCA(n_components=10)), ('model', RandomForestClassifier(class_weight='balance
d_subsample', criterion='entropy',
                       min_samples_split=100, n_estimators=1000,
                       random_state=0))]
Selector: RFECV RandomForest Model: RandomForest
              precision    recall  f1-score   support

           0       0.92      0.85      0.88      7963
           1       0.54      0.70      0.61      2037

    accuracy                           0.82     10000
   macro avg       0.73      0.77      0.74     10000
weighted avg       0.84      0.82      0.82     10000

[('sm', SMOTE()), ('sel', SelectFwe(alpha=0.01)), ('pca', PCA(n_components=1
0)), ('model', RandomForestClassifier(class_weight='balanced_subsample', crit
erion='entropy',
                       min_samples_split=100, n_estimators=1000,
                       random_state=0))]
```

```
====>>>> Failure in RandomForest, FWE f_classif
[('sm', SMOTE()), ('pca', PCA(n_components=10)), ('model', RandomForestClassi
fier(class_weight='balanced_subsample', criterion='entropy',
                       min_samples_split=100, n_estimators=1000,
                       random_state=0))]
Selector: NONE Model: RandomForest
              precision    recall  f1-score   support

           0       0.91      0.84      0.88      7963
           1       0.53      0.69      0.60      2037

    accuracy                           0.81     10000
   macro avg       0.72      0.77      0.74     10000
weighted avg       0.84      0.81      0.82     10000


MODEL: GradientBoosting
----------------------
                           FWE f_classif    0.00000    0.00000
                      RFECV RandomForest    0.79120    0.58390
                                    NONE    0.79150    0.58458
                  Variance Threshold 0.8    0.79280    0.58626
--------------------------------------------------------------------------------
--

MODEL: RandomForest
--------------------
                           FWE f_classif    0.00000    0.00000
                  Variance Threshold 0.8    0.80980    0.59669
                                    NONE    0.81250    0.59944
                      RFECV RandomForest    0.81510    0.60601
--------------------------------------------------------------------------------
--

[('sm', SMOTE()), ('sel', VarianceThreshold(threshold=0.15999999999999998)),
('pca', PCA(n_components=10)), ('model', SVC(gamma='auto', random_state=0))]
Selector: Variance Threshold 0.8 Model: SVC
              precision    recall  f1-score   support

           0       0.92      0.81      0.86      7963
           1       0.50      0.72      0.59      2037

    accuracy                           0.79     10000
   macro avg       0.71      0.77      0.73     10000
weighted avg       0.83      0.79      0.81     10000

[('sm', SMOTE()), ('sel', RFECV(estimator=RandomForestClassifier())), ('pca',
PCA(n_components=10)), ('model', SVC(gamma='auto', random_state=0))]
Selector: RFECV RandomForest Model: SVC
              precision    recall  f1-score   support

           0       0.92      0.82      0.86      7963
           1       0.50      0.72      0.59      2037

    accuracy                           0.80     10000
   macro avg       0.71      0.77      0.73     10000
weighted avg       0.83      0.80      0.81     10000
```

```
[('sm', SMOTE()), ('sel', SelectFwe(alpha=0.01)), ('pca', PCA(n_components=1
0)), ('model', SVC(gamma='auto', random_state=0))]
Selector: FWE f_classif Model: SVC
              precision    recall  f1-score   support

           0       0.92      0.82      0.87      7963
           1       0.51      0.74      0.60      2037

    accuracy                           0.80     10000
   macro avg       0.71      0.78      0.73     10000
weighted avg       0.84      0.80      0.81     10000

[('sm', SMOTE()), ('pca', PCA(n_components=10)), ('model', SVC(gamma='auto',
random_state=0))]
Selector: NONE Model: SVC
              precision    recall  f1-score   support

           0       0.92      0.81      0.86      7963
           1       0.50      0.72      0.59      2037

    accuracy                           0.80     10000
   macro avg       0.71      0.77      0.73     10000
weighted avg       0.83      0.80      0.81     10000


MODEL: GradientBoosting
----------------------
                          FWE f_classif    0.00000    0.00000
                     RFECV RandomForest    0.79120    0.58390
                                   NONE    0.79150    0.58458
                 Variance Threshold 0.8    0.79280    0.58626
--------------------------------------------------------------------------------

MODEL: RandomForest
----------------------
                          FWE f_classif    0.00000    0.00000
                 Variance Threshold 0.8    0.80980    0.59669
                                   NONE    0.81250    0.59944
                     RFECV RandomForest    0.81510    0.60601
--------------------------------------------------------------------------------

MODEL: SVC
----------------------
                 Variance Threshold 0.8    0.79450    0.58957
                                   NONE    0.79550    0.58960
                     RFECV RandomForest    0.79670    0.59037
                          FWE f_classif    0.79990    0.60004
--------------------------------------------------------------------------------

================================================================================
========================================
PCA 7
--------------------------------------------------------------------------------
```

```
---
[('sm', SMOTE()), ('sel', VarianceThreshold(threshold=0.15999999999999998)),
('pca', PCA(n_components=10)), ('model', GradientBoostingClassifier(min_sampl
es_split=42, random_state=0))]
Selector: Variance Threshold 0.8 Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.92      0.81      0.86      7963
           1       0.49      0.73      0.59      2037

    accuracy                           0.79     10000
   macro avg       0.71      0.77      0.72     10000
weighted avg       0.83      0.79      0.81     10000

[('sm', SMOTE()), ('sel', RFECV(estimator=RandomForestClassifier())), ('pca',
PCA(n_components=10)), ('model', GradientBoostingClassifier(min_samples_split
=42, random_state=0))]
Selector: RFECV RandomForest Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.91      0.83      0.87      7963
           1       0.51      0.70      0.59      2037

    accuracy                           0.80     10000
   macro avg       0.71      0.76      0.73     10000
weighted avg       0.83      0.80      0.81     10000

[('sm', SMOTE()), ('sel', SelectFwe(alpha=0.01)), ('pca', PCA(n_components=1
0)), ('model', GradientBoostingClassifier(min_samples_split=42, random_state=
0))]
Selector: FWE f_classif Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.92      0.82      0.87      7963
           1       0.51      0.73      0.60      2037

    accuracy                           0.80     10000
   macro avg       0.72      0.78      0.74     10000
weighted avg       0.84      0.80      0.81     10000

[('sm', SMOTE()), ('pca', PCA(n_components=10)), ('model', GradientBoostingCl
assifier(min_samples_split=42, random_state=0))]
Selector: NONE Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.92      0.81      0.86      7963
           1       0.49      0.72      0.58      2037

    accuracy                           0.79     10000
   macro avg       0.70      0.76      0.72     10000
weighted avg       0.83      0.79      0.80     10000


MODEL: GradientBoosting
----------------------
                                    NONE    0.78970    0.58299
                 Variance Threshold 0.8    0.79190    0.58784
```

```
                     RFECV RandomForest    0.80210    0.58984
                          FWE f_classif    0.80290    0.60142
--------------------------------------------------------------------------------

[('sm', SMOTE()), ('sel', VarianceThreshold(threshold=0.15999999999999998)),
('pca', PCA(n_components=10)), ('model', RandomForestClassifier(class_weight
='balanced_subsample', criterion='entropy',
                       min_samples_split=100, n_estimators=1000,
                       random_state=0))]
Selector: Variance Threshold 0.8 Model: RandomForest
              precision    recall  f1-score   support

           0       0.92      0.84      0.88      7963
           1       0.53      0.70      0.60      2037

    accuracy                           0.81     10000
   macro avg       0.72      0.77      0.74     10000
weighted avg       0.84      0.81      0.82     10000

[('sm', SMOTE()), ('sel', RFECV(estimator=RandomForestClassifier())), ('pca',
PCA(n_components=10)), ('model', RandomForestClassifier(class_weight='balance
d_subsample', criterion='entropy',
                       min_samples_split=100, n_estimators=1000,
                       random_state=0))]
Selector: RFECV RandomForest Model: RandomForest
              precision    recall  f1-score   support

           0       0.92      0.84      0.88      7963
           1       0.53      0.70      0.61      2037

    accuracy                           0.81     10000
   macro avg       0.73      0.77      0.74     10000
weighted avg       0.84      0.81      0.82     10000

[('sm', SMOTE()), ('sel', SelectFwe(alpha=0.01)), ('pca', PCA(n_components=1
0)), ('model', RandomForestClassifier(class_weight='balanced_subsample', crit
erion='entropy',
                       min_samples_split=100, n_estimators=1000,
                       random_state=0))]
====>>>> Failure in RandomForest, FWE f_classif
[('sm', SMOTE()), ('pca', PCA(n_components=10)), ('model', RandomForestClassi
fier(class_weight='balanced_subsample', criterion='entropy',
                       min_samples_split=100, n_estimators=1000,
                       random_state=0))]
Selector: NONE Model: RandomForest
              precision    recall  f1-score   support

           0       0.92      0.84      0.88      7963
           1       0.53      0.70      0.60      2037

    accuracy                           0.81     10000
   macro avg       0.72      0.77      0.74     10000
weighted avg       0.84      0.81      0.82     10000


MODEL: GradientBoosting
```

```
---------------------
                                  NONE    0.78970    0.58299
                 Variance Threshold 0.8    0.79190    0.58784
                     RFECV RandomForest    0.80210    0.58984
                          FWE f_classif    0.80290    0.60142
--------------------------------------------------------------------------------

MODEL: RandomForest
---------------------
                          FWE f_classif    0.00000    0.00000
                 Variance Threshold 0.8    0.81050    0.59979
                                  NONE    0.81160    0.60387
                     RFECV RandomForest    0.81470    0.60600
--------------------------------------------------------------------------------

[('sm', SMOTE()), ('sel', VarianceThreshold(threshold=0.15999999999999998)),
('pca', PCA(n_components=10)), ('model', SVC(gamma='auto', random_state=0))]
Selector: Variance Threshold 0.8 Model: SVC
              precision    recall  f1-score   support

           0       0.92      0.82      0.87      7963
           1       0.50      0.73      0.60      2037

    accuracy                           0.80     10000
   macro avg       0.71      0.77      0.73     10000
weighted avg       0.84      0.80      0.81     10000

[('sm', SMOTE()), ('sel', RFECV(estimator=RandomForestClassifier())), ('pca',
PCA(n_components=10)), ('model', SVC(gamma='auto', random_state=0))]
Selector: RFECV RandomForest Model: SVC
              precision    recall  f1-score   support

           0       0.92      0.81      0.86      7963
           1       0.50      0.73      0.59      2037

    accuracy                           0.80     10000
   macro avg       0.71      0.77      0.73     10000
weighted avg       0.84      0.80      0.81     10000

[('sm', SMOTE()), ('sel', SelectFwe(alpha=0.01)), ('pca', PCA(n_components=1
0)), ('model', SVC(gamma='auto', random_state=0))]
====>>>> Failure in SVC, FWE f_classif
[('sm', SMOTE()), ('pca', PCA(n_components=10)), ('model', SVC(gamma='auto',
random_state=0))]
Selector: NONE Model: SVC
              precision    recall  f1-score   support

           0       0.92      0.82      0.87      7963
           1       0.50      0.72      0.59      2037

    accuracy                           0.80     10000
   macro avg       0.71      0.77      0.73     10000
weighted avg       0.84      0.80      0.81     10000
```

```
MODEL: GradientBoosting
---------------------
                                    NONE    0.78970    0.58299
                  Variance Threshold 0.8    0.79190    0.58784
                      RFECV RandomForest    0.80210    0.58984
                           FWE f_classif    0.80290    0.60142
-------------------------------------------------------------------------------
--

MODEL: RandomForest
---------------------
                           FWE f_classif    0.00000    0.00000
                  Variance Threshold 0.8    0.81050    0.59979
                                    NONE    0.81160    0.60387
                      RFECV RandomForest    0.81470    0.60600
-------------------------------------------------------------------------------
--

MODEL: SVC
---------------------
                           FWE f_classif    0.00000    0.00000
                                    NONE    0.79810    0.59384
                      RFECV RandomForest    0.79690    0.59404
                  Variance Threshold 0.8    0.79770    0.59516
-------------------------------------------------------------------------------
--

================================================================================
========================================
PCA 8
--------------------------------------------------------------------------------
---
[('sm', SMOTE()), ('sel', VarianceThreshold(threshold=0.15999999999999998)),
('pca', PCA(n_components=10)), ('model', GradientBoostingClassifier(min_sampl
es_split=42, random_state=0))]
Selector: Variance Threshold 0.8 Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.92      0.82      0.86      7963
           1       0.50      0.73      0.59      2037

    accuracy                           0.80     10000
   macro avg       0.71      0.77      0.73     10000
weighted avg       0.84      0.80      0.81     10000

[('sm', SMOTE()), ('sel', RFECV(estimator=RandomForestClassifier())), ('pca',
PCA(n_components=10)), ('model', GradientBoostingClassifier(min_samples_split
=42, random_state=0))]
Selector: RFECV RandomForest Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.92      0.82      0.87      7963
           1       0.50      0.70      0.59      2037

    accuracy                           0.80     10000
   macro avg       0.71      0.76      0.73     10000
weighted avg       0.83      0.80      0.81     10000
```

```
[('sm', SMOTE()), ('sel', SelectFwe(alpha=0.01)), ('pca', PCA(n_components=1
0)), ('model', GradientBoostingClassifier(min_samples_split=42, random_state=
0))]
====>>>> Failure in GradientBoosting, FWE f_classif
[('sm', SMOTE()), ('pca', PCA(n_components=10)), ('model', GradientBoostingCl
assifier(min_samples_split=42, random_state=0))]
Selector: NONE Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.92      0.81      0.86      7963
           1       0.49      0.72      0.59      2037

    accuracy                           0.79     10000
   macro avg       0.71      0.77      0.72     10000
weighted avg       0.83      0.79      0.80     10000


MODEL: GradientBoosting
--------------------
                          FWE f_classif    0.00000    0.00000
                                   NONE    0.79170    0.58564
                     RFECV RandomForest    0.79900    0.58778
                Variance Threshold 0.8    0.79700    0.59302
--------------------------------------------------------------------------
--

[('sm', SMOTE()), ('sel', VarianceThreshold(threshold=0.15999999999999998)),
('pca', PCA(n_components=10)), ('model', RandomForestClassifier(class_weight
='balanced_subsample', criterion='entropy',
                       min_samples_split=100, n_estimators=1000,
                       random_state=0))]
Selector: Variance Threshold 0.8 Model: RandomForest
              precision    recall  f1-score   support

           0       0.92      0.84      0.88      7963
           1       0.53      0.70      0.61      2037

    accuracy                           0.81     10000
   macro avg       0.72      0.77      0.74     10000
weighted avg       0.84      0.81      0.82     10000

[('sm', SMOTE()), ('sel', RFECV(estimator=RandomForestClassifier())), ('pca',
PCA(n_components=10)), ('model', RandomForestClassifier(class_weight='balance
d_subsample', criterion='entropy',
                       min_samples_split=100, n_estimators=1000,
                       random_state=0))]
Selector: RFECV RandomForest Model: RandomForest
              precision    recall  f1-score   support

           0       0.92      0.84      0.88      7963
           1       0.53      0.70      0.60      2037

    accuracy                           0.81     10000
   macro avg       0.72      0.77      0.74     10000
weighted avg       0.84      0.81      0.82     10000
```

```
[('sm', SMOTE()), ('sel', SelectFwe(alpha=0.01)), ('pca', PCA(n_components=1
0)), ('model', RandomForestClassifier(class_weight='balanced_subsample', crit
erion='entropy',
                       min_samples_split=100, n_estimators=1000,
                       random_state=0))]
====>>>> Failure in RandomForest, FWE f_classif
[('sm', SMOTE()), ('pca', PCA(n_components=10)), ('model', RandomForestClassi
fier(class_weight='balanced_subsample', criterion='entropy',
                       min_samples_split=100, n_estimators=1000,
                       random_state=0))]
Selector: NONE Model: RandomForest
              precision    recall  f1-score   support

           0       0.92      0.84      0.88      7963
           1       0.53      0.70      0.60      2037

    accuracy                           0.81     10000
   macro avg       0.72      0.77      0.74     10000
weighted avg       0.84      0.81      0.82     10000


MODEL: GradientBoosting
----------------------
                          FWE f_classif    0.00000    0.00000
                                   NONE    0.79170    0.58564
                      RFECV RandomForest    0.79900    0.58778
                  Variance Threshold 0.8    0.79700    0.59302
-------------------------------------------------------------------------------
--

MODEL: RandomForest
----------------------
                          FWE f_classif    0.00000    0.00000
                      RFECV RandomForest    0.81230    0.60140
                                   NONE    0.81250    0.60435
                  Variance Threshold 0.8    0.81320    0.60574
-------------------------------------------------------------------------------
--

[('sm', SMOTE()), ('sel', VarianceThreshold(threshold=0.15999999999999998)),
('pca', PCA(n_components=10)), ('model', SVC(gamma='auto', random_state=0))]
Selector: Variance Threshold 0.8 Model: SVC
              precision    recall  f1-score   support

           0       0.92      0.82      0.87      7963
           1       0.50      0.73      0.59      2037

    accuracy                           0.80     10000
   macro avg       0.71      0.77      0.73     10000
weighted avg       0.84      0.80      0.81     10000

[('sm', SMOTE()), ('sel', RFECV(estimator=RandomForestClassifier())), ('pca',
PCA(n_components=10)), ('model', SVC(gamma='auto', random_state=0))]
Selector: RFECV RandomForest Model: SVC
              precision    recall  f1-score   support

           0       0.92      0.81      0.86      7963
```

```
           1       0.50      0.73      0.59      2037

    accuracy                           0.79     10000
   macro avg       0.71      0.77      0.73     10000
weighted avg       0.83      0.79      0.81     10000

[('sm', SMOTE()), ('sel', SelectFwe(alpha=0.01)), ('pca', PCA(n_components=1
0)), ('model', SVC(gamma='auto', random_state=0))]
====>>>> Failure in SVC, FWE f_classif
[('sm', SMOTE()), ('pca', PCA(n_components=10)), ('model', SVC(gamma='auto',
random_state=0))]
Selector: NONE Model: SVC
              precision    recall  f1-score   support

           0       0.92      0.82      0.87      7963
           1       0.50      0.73      0.59      2037

    accuracy                           0.80     10000
   macro avg       0.71      0.77      0.73     10000
weighted avg       0.84      0.80      0.81     10000


MODEL: GradientBoosting
----------------------
                          FWE f_classif    0.00000    0.00000
                                   NONE    0.79170    0.58564
                     RFECV RandomForest    0.79900    0.58778
                  Variance Threshold 0.8    0.79700    0.59302
--------------------------------------------------------------------------
--

MODEL: RandomForest
----------------------
                          FWE f_classif    0.00000    0.00000
                     RFECV RandomForest    0.81230    0.60140
                                   NONE    0.81250    0.60435
                  Variance Threshold 0.8    0.81320    0.60574
--------------------------------------------------------------------------
--

MODEL: SVC
----------------------
                          FWE f_classif    0.00000    0.00000
                     RFECV RandomForest    0.79430    0.59081
                                   NONE    0.79770    0.59369
                  Variance Threshold 0.8    0.79810    0.59482
--------------------------------------------------------------------------
--

==========================================================================
==========================================
PCA 9
--------------------------------------------------------------------------
---
[('sm', SMOTE()), ('sel', VarianceThreshold(threshold=0.15999999999999998)),
('pca', PCA(n_components=10)), ('model', GradientBoostingClassifier(min_sampl
es_split=42, random_state=0))]
```

```
Selector: Variance Threshold 0.8 Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.92      0.81      0.86      7963
           1       0.50      0.73      0.59      2037

    accuracy                           0.80     10000
   macro avg       0.71      0.77      0.73     10000
weighted avg       0.84      0.80      0.81     10000

[('sm', SMOTE()), ('sel', RFECV(estimator=RandomForestClassifier())), ('pca',
PCA(n_components=10)), ('model', GradientBoostingClassifier(min_samples_split
=42, random_state=0))]
Selector: RFECV RandomForest Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.92      0.81      0.86      7963
           1       0.49      0.71      0.58      2037

    accuracy                           0.79     10000
   macro avg       0.70      0.76      0.72     10000
weighted avg       0.83      0.79      0.80     10000

[('sm', SMOTE()), ('sel', SelectFwe(alpha=0.01)), ('pca', PCA(n_components=1
0)), ('model', GradientBoostingClassifier(min_samples_split=42, random_state=
0))]
Selector: FWE f_classif Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.92      0.82      0.87      7963
           1       0.51      0.73      0.60      2037

    accuracy                           0.80     10000
   macro avg       0.71      0.78      0.73     10000
weighted avg       0.84      0.80      0.81     10000

[('sm', SMOTE()), ('pca', PCA(n_components=10)), ('model', GradientBoostingCl
assifier(min_samples_split=42, random_state=0))]
Selector: NONE Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.92      0.80      0.86      7963
           1       0.48      0.72      0.58      2037

    accuracy                           0.79     10000
   macro avg       0.70      0.76      0.72     10000
weighted avg       0.83      0.79      0.80     10000


MODEL: GradientBoosting
----------------------
                                  NONE    0.78660    0.57809
                     RFECV RandomForest    0.79170    0.58181
                  Variance Threshold 0.8    0.79690    0.59388
                           FWE f_classif    0.80050    0.59932
--------------------------------------------------------------------------
--
```

```
[('sm', SMOTE()), ('sel', VarianceThreshold(threshold=0.15999999999999998)),
('pca', PCA(n_components=10)), ('model', RandomForestClassifier(class_weight
='balanced_subsample', criterion='entropy',
                       min_samples_split=100, n_estimators=1000,
                       random_state=0))]
Selector: Variance Threshold 0.8 Model: RandomForest
              precision    recall  f1-score   support

           0       0.92      0.84      0.88      7963
           1       0.53      0.70      0.60      2037

    accuracy                           0.81     10000
   macro avg       0.72      0.77      0.74     10000
weighted avg       0.84      0.81      0.82     10000

[('sm', SMOTE()), ('sel', RFECV(estimator=RandomForestClassifier())), ('pca',
PCA(n_components=10)), ('model', RandomForestClassifier(class_weight='balance
d_subsample', criterion='entropy',
                       min_samples_split=100, n_estimators=1000,
                       random_state=0))]
Selector: RFECV RandomForest Model: RandomForest
              precision    recall  f1-score   support

           0       0.91      0.85      0.88      7963
           1       0.53      0.69      0.60      2037

    accuracy                           0.81     10000
   macro avg       0.72      0.77      0.74     10000
weighted avg       0.84      0.81      0.82     10000

[('sm', SMOTE()), ('sel', SelectFwe(alpha=0.01)), ('pca', PCA(n_components=1
0)), ('model', RandomForestClassifier(class_weight='balanced_subsample', crit
erion='entropy',
                       min_samples_split=100, n_estimators=1000,
                       random_state=0))]
Selector: FWE f_classif Model: RandomForest
              precision    recall  f1-score   support

           0       0.92      0.84      0.88      7963
           1       0.54      0.72      0.62      2037

    accuracy                           0.82     10000
   macro avg       0.73      0.78      0.75     10000
weighted avg       0.84      0.82      0.83     10000

[('sm', SMOTE()), ('pca', PCA(n_components=10)), ('model', RandomForestClassi
fier(class_weight='balanced_subsample', criterion='entropy',
                       min_samples_split=100, n_estimators=1000,
                       random_state=0))]
Selector: NONE Model: RandomForest
              precision    recall  f1-score   support

           0       0.92      0.84      0.88      7963
           1       0.53      0.70      0.60      2037

    accuracy                           0.81     10000
```

```
   macro avg       0.72      0.77      0.74     10000
weighted avg       0.84      0.81      0.82     10000


MODEL: GradientBoosting
--------------------
                                    NONE    0.78660    0.57809
                       RFECV RandomForest    0.79170    0.58181
                  Variance Threshold 0.8    0.79690    0.59388
                           FWE f_classif    0.80050    0.59932
--------------------------------------------------------------------------------
--

MODEL: RandomForest
--------------------
                                    NONE    0.81140    0.60127
                       RFECV RandomForest    0.81440    0.60137
                  Variance Threshold 0.8    0.81280    0.60289
                           FWE f_classif    0.81890    0.61769
--------------------------------------------------------------------------------
--

[('sm', SMOTE()), ('sel', VarianceThreshold(threshold=0.15999999999999998)),
('pca', PCA(n_components=10)), ('model', SVC(gamma='auto', random_state=0))]
Selector: Variance Threshold 0.8 Model: SVC
              precision    recall  f1-score   support

           0       0.92      0.81      0.86      7963
           1       0.50      0.72      0.59      2037

    accuracy                           0.79     10000
   macro avg       0.71      0.77      0.73     10000
weighted avg       0.83      0.79      0.81     10000

[('sm', SMOTE()), ('sel', RFECV(estimator=RandomForestClassifier())), ('pca',
PCA(n_components=10)), ('model', SVC(gamma='auto', random_state=0))]
Selector: RFECV RandomForest Model: SVC
              precision    recall  f1-score   support

           0       0.92      0.81      0.86      7963
           1       0.50      0.72      0.59      2037

    accuracy                           0.79     10000
   macro avg       0.71      0.77      0.73     10000
weighted avg       0.83      0.79      0.81     10000

[('sm', SMOTE()), ('sel', SelectFwe(alpha=0.01)), ('pca', PCA(n_components=1
0)), ('model', SVC(gamma='auto', random_state=0))]
====>>>> Failure in SVC, FWE f_classif
[('sm', SMOTE()), ('pca', PCA(n_components=10)), ('model', SVC(gamma='auto',
random_state=0))]
Selector: NONE Model: SVC
              precision    recall  f1-score   support

           0       0.92      0.82      0.87      7963
           1       0.50      0.72      0.59      2037
```

```
    accuracy                           0.80     10000
   macro avg       0.71      0.77      0.73     10000
weighted avg       0.84      0.80      0.81     10000


MODEL: GradientBoosting
----------------------
                                    NONE    0.78660    0.57809
                       RFECV RandomForest    0.79170    0.58181
                  Variance Threshold 0.8    0.79690    0.59388
                            FWE f_classif    0.80050    0.59932
--------------------------------------------------------------------------------
--

MODEL: RandomForest
----------------------
                                    NONE    0.81140    0.60127
                       RFECV RandomForest    0.81440    0.60137
                  Variance Threshold 0.8    0.81280    0.60289
                            FWE f_classif    0.81890    0.61769
--------------------------------------------------------------------------------
--

MODEL: SVC
----------------------
                            FWE f_classif    0.00000    0.00000
                  Variance Threshold 0.8    0.79380    0.58842
                       RFECV RandomForest    0.79490    0.58988
                                    NONE    0.79860    0.59346
--------------------------------------------------------------------------------
--

================================================================================
========================================
PCA 10
--------------------------------------------------------------------------------
---
[('sm', SMOTE()), ('sel', VarianceThreshold(threshold=0.15999999999999998)),
('pca', PCA(n_components=10)), ('model', GradientBoostingClassifier(min_sampl
es_split=42, random_state=0))]
Selector: Variance Threshold 0.8 Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.92      0.82      0.86      7963
           1       0.50      0.72      0.59      2037

    accuracy                           0.80     10000
   macro avg       0.71      0.77      0.73     10000
weighted avg       0.83      0.80      0.81     10000

[('sm', SMOTE()), ('sel', RFECV(estimator=RandomForestClassifier())), ('pca',
PCA(n_components=10)), ('model', GradientBoostingClassifier(min_samples_split
=42, random_state=0))]
Selector: RFECV RandomForest Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.92      0.83      0.87      7963
```

```
           1       0.51      0.70      0.59      2037

    accuracy                           0.80     10000
   macro avg       0.71      0.76      0.73     10000
weighted avg       0.83      0.80      0.81     10000

[('sm', SMOTE()), ('sel', SelectFwe(alpha=0.01)), ('pca', PCA(n_components=1
0)), ('model', GradientBoostingClassifier(min_samples_split=42, random_state=
0))]
====>>>> Failure in GradientBoosting, FWE f_classif
[('sm', SMOTE()), ('pca', PCA(n_components=10)), ('model', GradientBoostingCl
assifier(min_samples_split=42, random_state=0))]
Selector: NONE Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.92      0.81      0.86      7963
           1       0.50      0.72      0.59      2037

    accuracy                           0.80     10000
   macro avg       0.71      0.77      0.73     10000
weighted avg       0.83      0.80      0.81     10000


MODEL: GradientBoosting
----------------------
                          FWE f_classif    0.00000    0.00000
                     RFECV RandomForest    0.80030    0.58867
                                   NONE    0.79500    0.58918
                  Variance Threshold 0.8    0.79650    0.58947
--------------------------------------------------------------------------------

[('sm', SMOTE()), ('sel', VarianceThreshold(threshold=0.15999999999999998)),
('pca', PCA(n_components=10)), ('model', RandomForestClassifier(class_weight
='balanced_subsample', criterion='entropy',
                       min_samples_split=100, n_estimators=1000,
                       random_state=0))]
Selector: Variance Threshold 0.8 Model: RandomForest
              precision    recall  f1-score   support

           0       0.92      0.84      0.88      7963
           1       0.53      0.70      0.60      2037

    accuracy                           0.81     10000
   macro avg       0.72      0.77      0.74     10000
weighted avg       0.84      0.81      0.82     10000

[('sm', SMOTE()), ('sel', RFECV(estimator=RandomForestClassifier())), ('pca',
PCA(n_components=10)), ('model', RandomForestClassifier(class_weight='balance
d_subsample', criterion='entropy',
                       min_samples_split=100, n_estimators=1000,
                       random_state=0))]
Selector: RFECV RandomForest Model: RandomForest
              precision    recall  f1-score   support

           0       0.91      0.84      0.88      7963
           1       0.53      0.69      0.60      2037
```

```
    accuracy                           0.81     10000
   macro avg       0.72      0.77      0.74     10000
weighted avg       0.84      0.81      0.82     10000

[('sm', SMOTE()), ('sel', SelectFwe(alpha=0.01)), ('pca', PCA(n_components=1
0)), ('model', RandomForestClassifier(class_weight='balanced_subsample', crit
erion='entropy',
                       min_samples_split=100, n_estimators=1000,
                       random_state=0))]
Selector: FWE f_classif Model: RandomForest
              precision    recall  f1-score   support

           0       0.92      0.85      0.88      7963
           1       0.54      0.72      0.62      2037

    accuracy                           0.82     10000
   macro avg       0.73      0.78      0.75     10000
weighted avg       0.84      0.82      0.83     10000

[('sm', SMOTE()), ('pca', PCA(n_components=10)), ('model', RandomForestClassi
fier(class_weight='balanced_subsample', criterion='entropy',
                       min_samples_split=100, n_estimators=1000,
                       random_state=0))]
Selector: NONE Model: RandomForest
              precision    recall  f1-score   support

           0       0.92      0.84      0.88      7963
           1       0.53      0.70      0.61      2037

    accuracy                           0.81     10000
   macro avg       0.73      0.77      0.74     10000
weighted avg       0.84      0.81      0.82     10000


MODEL: GradientBoosting
----------------------
                          FWE f_classif    0.00000    0.00000
                     RFECV RandomForest    0.80030    0.58867
                                   NONE    0.79500    0.58918
                 Variance Threshold 0.8    0.79650    0.58947
------------------------------------------------------------------------------
--

MODEL: RandomForest
----------------------
                     RFECV RandomForest    0.81080    0.59881
                 Variance Threshold 0.8    0.81170    0.60165
                                   NONE    0.81460    0.60737
                          FWE f_classif    0.81890    0.61704
------------------------------------------------------------------------------
--

[('sm', SMOTE()), ('sel', VarianceThreshold(threshold=0.15999999999999998)),
('pca', PCA(n_components=10)), ('model', SVC(gamma='auto', random_state=0))]
Selector: Variance Threshold 0.8 Model: SVC
              precision    recall  f1-score   support
```

```
           0       0.92      0.82      0.86      7963
           1       0.50      0.72      0.59      2037

    accuracy                           0.80     10000
   macro avg       0.71      0.77      0.73     10000
weighted avg       0.83      0.80      0.81     10000

[('sm', SMOTE()), ('sel', RFECV(estimator=RandomForestClassifier())), ('pca',
PCA(n_components=10)), ('model', SVC(gamma='auto', random_state=0))]
Selector: RFECV RandomForest Model: SVC
              precision    recall  f1-score   support

           0       0.92      0.81      0.86      7963
           1       0.49      0.73      0.59      2037

    accuracy                           0.79     10000
   macro avg       0.71      0.77      0.72     10000
weighted avg       0.83      0.79      0.80     10000

[('sm', SMOTE()), ('sel', SelectFwe(alpha=0.01)), ('pca', PCA(n_components=1
0)), ('model', SVC(gamma='auto', random_state=0))]
Selector: FWE f_classif Model: SVC
              precision    recall  f1-score   support

           0       0.92      0.82      0.87      7963
           1       0.51      0.74      0.60      2037

    accuracy                           0.80     10000
   macro avg       0.72      0.78      0.74     10000
weighted avg       0.84      0.80      0.81     10000

[('sm', SMOTE()), ('pca', PCA(n_components=10)), ('model', SVC(gamma='auto',
random_state=0))]
Selector: NONE Model: SVC
              precision    recall  f1-score   support

           0       0.92      0.82      0.87      7963
           1       0.50      0.73      0.60      2037

    accuracy                           0.80     10000
   macro avg       0.71      0.77      0.73     10000
weighted avg       0.84      0.80      0.81     10000


MODEL: GradientBoosting
----------------------
                          FWE f_classif    0.00000    0.00000
                     RFECV RandomForest    0.80030    0.58867
                                   NONE    0.79500    0.58918
                 Variance Threshold 0.8    0.79650    0.58947
-------------------------------------------------------------------------
--

MODEL: RandomForest
----------------------
                     RFECV RandomForest    0.81080    0.59881
```

```
                Variance Threshold 0.8    0.81170    0.60165
                                  NONE    0.81460    0.60737
                           FWE f_classif    0.81890    0.61704
--------------------------------------------------------------------------------

MODEL: SVC
----------------------
                    RFECV RandomForest    0.79120    0.58653
                Variance Threshold 0.8    0.79710    0.59232
                                  NONE    0.79820    0.59510
                           FWE f_classif    0.80200    0.60337
--------------------------------------------------------------------------------

================================================================================
==========================================
PCA 11
--------------------------------------------------------------------------------
---
[('sm', SMOTE()), ('sel', VarianceThreshold(threshold=0.15999999999999998)),
('pca', PCA(n_components=10)), ('model', GradientBoostingClassifier(min_sampl
es_split=42, random_state=0))]
Selector: Variance Threshold 0.8 Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.92      0.81      0.86      7963
           1       0.50      0.72      0.59      2037

    accuracy                           0.80     10000
   macro avg       0.71      0.77      0.73     10000
weighted avg       0.83      0.80      0.81     10000

[('sm', SMOTE()), ('sel', RFECV(estimator=RandomForestClassifier())), ('pca',
PCA(n_components=10)), ('model', GradientBoostingClassifier(min_samples_split
=42, random_state=0))]
Selector: RFECV RandomForest Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.92      0.82      0.87      7963
           1       0.51      0.71      0.59      2037

    accuracy                           0.80     10000
   macro avg       0.71      0.77      0.73     10000
weighted avg       0.83      0.80      0.81     10000

[('sm', SMOTE()), ('sel', SelectFwe(alpha=0.01)), ('pca', PCA(n_components=1
0)), ('model', GradientBoostingClassifier(min_samples_split=42, random_state=
0))]
Selector: FWE f_classif Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.92      0.82      0.87      7963
           1       0.51      0.73      0.60      2037

    accuracy                           0.80     10000
   macro avg       0.72      0.77      0.74     10000
```

```
weighted avg       0.84      0.80      0.82     10000

[('sm', SMOTE()), ('pca', PCA(n_components=10)), ('model', GradientBoostingCl
assifier(min_samples_split=42, random_state=0))]
Selector: NONE Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.92      0.81      0.86      7963
           1       0.49      0.72      0.59      2037

    accuracy                           0.79     10000
   macro avg       0.71      0.77      0.72     10000
weighted avg       0.83      0.79      0.81     10000


MODEL: GradientBoosting
----------------------
                                    NONE    0.79300    0.58781
                 Variance Threshold 0.8    0.79520    0.58859
                     RFECV RandomForest    0.80020    0.59141
                          FWE f_classif    0.80360    0.60114
--------------------------------------------------------------------------------

[('sm', SMOTE()), ('sel', VarianceThreshold(threshold=0.15999999999999998)),
('pca', PCA(n_components=10)), ('model', RandomForestClassifier(class_weight
='balanced_subsample', criterion='entropy',
                       min_samples_split=100, n_estimators=1000,
                       random_state=0))]
Selector: Variance Threshold 0.8 Model: RandomForest
              precision    recall  f1-score   support

           0       0.92      0.84      0.88      7963
           1       0.53      0.70      0.60      2037

    accuracy                           0.81     10000
   macro avg       0.72      0.77      0.74     10000
weighted avg       0.84      0.81      0.82     10000

[('sm', SMOTE()), ('sel', RFECV(estimator=RandomForestClassifier())), ('pca',
PCA(n_components=10)), ('model', RandomForestClassifier(class_weight='balance
d_subsample', criterion='entropy',
                       min_samples_split=100, n_estimators=1000,
                       random_state=0))]
Selector: RFECV RandomForest Model: RandomForest
              precision    recall  f1-score   support

           0       0.91      0.85      0.88      7963
           1       0.53      0.67      0.59      2037

    accuracy                           0.81     10000
   macro avg       0.72      0.76      0.74     10000
weighted avg       0.83      0.81      0.82     10000

[('sm', SMOTE()), ('sel', SelectFwe(alpha=0.01)), ('pca', PCA(n_components=1
0)), ('model', RandomForestClassifier(class_weight='balanced_subsample', crit
erion='entropy',
```

```
                     min_samples_split=100, n_estimators=1000,
                     random_state=0))]
Selector: FWE f_classif Model: RandomForest
              precision    recall  f1-score   support

           0       0.92      0.84      0.88      7963
           1       0.54      0.71      0.61      2037

    accuracy                           0.82     10000
   macro avg       0.73      0.78      0.75     10000
weighted avg       0.84      0.82      0.83     10000

[('sm', SMOTE()), ('pca', PCA(n_components=10)), ('model', RandomForestClassi
fier(class_weight='balanced_subsample', criterion='entropy',
                     min_samples_split=100, n_estimators=1000,
                     random_state=0))]
Selector: NONE Model: RandomForest
              precision    recall  f1-score   support

           0       0.92      0.84      0.88      7963
           1       0.53      0.70      0.60      2037

    accuracy                           0.81     10000
   macro avg       0.72      0.77      0.74     10000
weighted avg       0.84      0.81      0.82     10000


MODEL: GradientBoosting
----------------------
                                  NONE    0.79300    0.58781
                 Variance Threshold 0.8    0.79520    0.58859
                     RFECV RandomForest    0.80020    0.59141
                          FWE f_classif    0.80360    0.60114
--------------------------------------------------------------------------
--

MODEL: RandomForest
----------------------
                     RFECV RandomForest    0.81240    0.59288
                 Variance Threshold 0.8    0.81210    0.60400
                                  NONE    0.81270    0.60460
                          FWE f_classif    0.81740    0.61379
--------------------------------------------------------------------------
--

[('sm', SMOTE()), ('sel', VarianceThreshold(threshold=0.15999999999999998)),
('pca', PCA(n_components=10)), ('model', SVC(gamma='auto', random_state=0))]
Selector: Variance Threshold 0.8 Model: SVC
              precision    recall  f1-score   support

           0       0.92      0.82      0.87      7963
           1       0.50      0.72      0.59      2037

    accuracy                           0.80     10000
   macro avg       0.71      0.77      0.73     10000
weighted avg       0.84      0.80      0.81     10000
```

```
[('sm', SMOTE()), ('sel', RFECV(estimator=RandomForestClassifier())), ('pca',
PCA(n_components=10)), ('model', SVC(gamma='auto', random_state=0))]
Selector: RFECV RandomForest Model: SVC
              precision    recall  f1-score   support

           0       0.92      0.81      0.86      7963
           1       0.50      0.73      0.59      2037

    accuracy                           0.80     10000
   macro avg       0.71      0.77      0.73     10000
weighted avg       0.83      0.80      0.81     10000

[('sm', SMOTE()), ('sel', SelectFwe(alpha=0.01)), ('pca', PCA(n_components=1
0)), ('model', SVC(gamma='auto', random_state=0))]
====>>>> Failure in SVC, FWE f_classif
[('sm', SMOTE()), ('pca', PCA(n_components=10)), ('model', SVC(gamma='auto',
random_state=0))]
Selector: NONE Model: SVC
              precision    recall  f1-score   support

           0       0.92      0.82      0.87      7963
           1       0.50      0.73      0.59      2037

    accuracy                           0.80     10000
   macro avg       0.71      0.77      0.73     10000
weighted avg       0.84      0.80      0.81     10000


MODEL: GradientBoosting
----------------------
                                  NONE    0.79300    0.58781
                 Variance Threshold 0.8    0.79520    0.58859
                     RFECV RandomForest    0.80020    0.59141
                          FWE f_classif    0.80360    0.60114
--------------------------------------------------------------------------
--

MODEL: RandomForest
----------------------
                     RFECV RandomForest    0.81240    0.59288
                 Variance Threshold 0.8    0.81210    0.60400
                                  NONE    0.81270    0.60460
                          FWE f_classif    0.81740    0.61379
--------------------------------------------------------------------------
--

MODEL: SVC
----------------------
                          FWE f_classif    0.00000    0.00000
                     RFECV RandomForest    0.79630    0.59187
                 Variance Threshold 0.8    0.79810    0.59335
                                  NONE    0.79780    0.59365
--------------------------------------------------------------------------
--

==========================================================================
========================================
```

PCA 12

[('sm', SMOTE()), ('sel', VarianceThreshold(threshold=0.15999999999999998)),
('pca', PCA(n_components=10)), ('model', GradientBoostingClassifier(min_sampl
es_split=42, random_state=0))]
Selector: Variance Threshold 0.8 Model: GradientBoosting
              precision   recall f1-score   support

           0   0.92   0.81   0.86   7963
           1   0.50   0.73   0.59   2037

    accuracy   0.79   10000
   macro avg   0.71   0.77   0.73   10000
weighted avg   0.83   0.79   0.81   10000

[('sm', SMOTE()), ('sel', RFECV(estimator=RandomForestClassifier())), ('pca',
PCA(n_components=10)), ('model', GradientBoostingClassifier(min_samples_split
=42, random_state=0))]
Selector: RFECV RandomForest Model: GradientBoosting
              precision   recall f1-score   support

           0   0.92   0.83   0.87   7963
           1   0.51   0.71   0.59   2037

    accuracy   0.80   10000
   macro avg   0.71   0.77   0.73   10000
weighted avg   0.83   0.80   0.81   10000

[('sm', SMOTE()), ('sel', SelectFwe(alpha=0.01)), ('pca', PCA(n_components=1
0)), ('model', GradientBoostingClassifier(min_samples_split=42, random_state=
0))]
====>>>> Failure in GradientBoosting, FWE f_classif
[('sm', SMOTE()), ('pca', PCA(n_components=10)), ('model', GradientBoostingCl
assifier(min_samples_split=42, random_state=0))]
Selector: NONE Model: GradientBoosting
              precision   recall f1-score   support

           0   0.92   0.81   0.86   7963
           1   0.50   0.73   0.59   2037

    accuracy   0.79   10000
   macro avg   0.71   0.77   0.73   10000
weighted avg   0.83   0.79   0.81   10000

MODEL: GradientBoosting

                          FWE f_classif   0.00000   0.00000
                Variance Threshold 0.8   0.79380   0.58941
                                  NONE   0.79400   0.58997
                     RFECV RandomForest   0.80150   0.59182

--

[('sm', SMOTE()), ('sel', VarianceThreshold(threshold=0.15999999999999998)),
('pca', PCA(n_components=10)), ('model', RandomForestClassifier(class_weight

```
='balanced_subsample', criterion='entropy',
                       min_samples_split=100, n_estimators=1000,
                       random_state=0))]
Selector: Variance Threshold 0.8 Model: RandomForest
              precision    recall  f1-score   support

           0       0.92      0.84      0.88      7963
           1       0.53      0.70      0.60      2037

    accuracy                           0.81     10000
   macro avg       0.72      0.77      0.74     10000
weighted avg       0.84      0.81      0.82     10000

[('sm', SMOTE()), ('sel', RFECV(estimator=RandomForestClassifier())), ('pca',
PCA(n_components=10)), ('model', RandomForestClassifier(class_weight='balance
d_subsample', criterion='entropy',
                       min_samples_split=100, n_estimators=1000,
                       random_state=0))]
Selector: RFECV RandomForest Model: RandomForest
              precision    recall  f1-score   support

           0       0.92      0.85      0.88      7963
           1       0.54      0.70      0.61      2037

    accuracy                           0.82     10000
   macro avg       0.73      0.77      0.74     10000
weighted avg       0.84      0.82      0.82     10000

[('sm', SMOTE()), ('sel', SelectFwe(alpha=0.01)), ('pca', PCA(n_components=1
0)), ('model', RandomForestClassifier(class_weight='balanced_subsample', crit
erion='entropy',
                       min_samples_split=100, n_estimators=1000,
                       random_state=0))]
Selector: FWE f_classif Model: RandomForest
              precision    recall  f1-score   support

           0       0.92      0.84      0.88      7963
           1       0.53      0.71      0.61      2037

    accuracy                           0.81     10000
   macro avg       0.73      0.77      0.74     10000
weighted avg       0.84      0.81      0.82     10000

[('sm', SMOTE()), ('pca', PCA(n_components=10)), ('model', RandomForestClassi
fier(class_weight='balanced_subsample', criterion='entropy',
                       min_samples_split=100, n_estimators=1000,
                       random_state=0))]
Selector: NONE Model: RandomForest
              precision    recall  f1-score   support

           0       0.92      0.84      0.87      7963
           1       0.52      0.70      0.60      2037

    accuracy                           0.81     10000
   macro avg       0.72      0.77      0.74     10000
weighted avg       0.84      0.81      0.82     10000
```

```
MODEL: GradientBoosting
----------------------
                          FWE f_classif    0.00000    0.00000
                Variance Threshold 0.8    0.79380    0.58941
                                  NONE    0.79400    0.58997
                    RFECV RandomForest    0.80150    0.59182
--------------------------------------------------------------------------------
--

MODEL: RandomForest
----------------------
                                  NONE    0.80930    0.59929
                Variance Threshold 0.8    0.81100    0.60059
                    RFECV RandomForest    0.81520    0.60597
                          FWE f_classif    0.81350    0.60778
--------------------------------------------------------------------------------
--

[('sm', SMOTE()), ('sel', VarianceThreshold(threshold=0.15999999999999998)),
 ('pca', PCA(n_components=10)), ('model', SVC(gamma='auto', random_state=0))]
Selector: Variance Threshold 0.8 Model: SVC
              precision    recall  f1-score   support

           0       0.92      0.82      0.87      7963
           1       0.50      0.72      0.59      2037

    accuracy                           0.80     10000
   macro avg       0.71      0.77      0.73     10000
weighted avg       0.84      0.80      0.81     10000

[('sm', SMOTE()), ('sel', RFECV(estimator=RandomForestClassifier())), ('pca',
PCA(n_components=10)), ('model', SVC(gamma='auto', random_state=0))]
Selector: RFECV RandomForest Model: SVC
              precision    recall  f1-score   support

           0       0.92      0.81      0.86      7963
           1       0.50      0.73      0.59      2037

    accuracy                           0.80     10000
   macro avg       0.71      0.77      0.73     10000
weighted avg       0.83      0.80      0.81     10000

[('sm', SMOTE()), ('sel', SelectFwe(alpha=0.01)), ('pca', PCA(n_components=1
0)), ('model', SVC(gamma='auto', random_state=0))]
====>>>> Failure in SVC, FWE f_classif
[('sm', SMOTE()), ('pca', PCA(n_components=10)), ('model', SVC(gamma='auto',
 random_state=0))]
Selector: NONE Model: SVC
              precision    recall  f1-score   support

           0       0.92      0.81      0.86      7963
           1       0.50      0.73      0.59      2037

    accuracy                           0.80     10000
   macro avg       0.71      0.77      0.73     10000
weighted avg       0.84      0.80      0.81     10000
```

```
MODEL: GradientBoosting
---------------------
                          FWE f_classif    0.00000    0.00000
                Variance Threshold 0.8    0.79380    0.58941
                                  NONE    0.79400    0.58997
                    RFECV RandomForest    0.80150    0.59182
-------------------------------------------------------------------------------

MODEL: RandomForest
---------------------
                                  NONE    0.80930    0.59929
                Variance Threshold 0.8    0.81100    0.60059
                    RFECV RandomForest    0.81520    0.60597
                          FWE f_classif    0.81350    0.60778
-------------------------------------------------------------------------------

MODEL: SVC
---------------------
                          FWE f_classif    0.00000    0.00000
                    RFECV RandomForest    0.79530    0.59068
                Variance Threshold 0.8    0.79730    0.59256
                                  NONE    0.79650    0.59259
-------------------------------------------------------------------------------

================================================================================
==========================================
PCA 13
--------------------------------------------------------------------------------
---
[('sm', SMOTE()), ('sel', VarianceThreshold(threshold=0.15999999999999998)),
 ('pca', PCA(n_components=10)), ('model', GradientBoostingClassifier(min_samp
les_split=42, random_state=0))]
Selector: Variance Threshold 0.8 Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.92      0.81      0.86      7963
           1       0.49      0.72      0.59      2037

    accuracy                           0.79     10000
   macro avg       0.71      0.77      0.72     10000
weighted avg       0.83      0.79      0.81     10000

[('sm', SMOTE()), ('sel', RFECV(estimator=RandomForestClassifier())), ('pca',
PCA(n_components=10)), ('model', GradientBoostingClassifier(min_samples_split
=42, random_state=0))]
Selector: RFECV RandomForest Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.92      0.81      0.86      7963
           1       0.49      0.72      0.59      2037

    accuracy                           0.79     10000
```

```
   macro avg       0.71      0.77      0.72     10000
weighted avg       0.83      0.79      0.81     10000

[('sm', SMOTE()), ('sel', SelectFwe(alpha=0.01)), ('pca', PCA(n_components=1
0)), ('model', GradientBoostingClassifier(min_samples_split=42, random_state=
0))]
====>>>> Failure in GradientBoosting, FWE f_classif
[('sm', SMOTE()), ('pca', PCA(n_components=10)), ('model', GradientBoostingCl
assifier(min_samples_split=42, random_state=0))]
Selector: NONE Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.92      0.81      0.86      7963
           1       0.50      0.73      0.59      2037

    accuracy                           0.79     10000
   macro avg       0.71      0.77      0.73     10000
weighted avg       0.83      0.79      0.81     10000


MODEL: GradientBoosting
----------------------
                          FWE f_classif    0.00000    0.00000
                     RFECV RandomForest    0.79200    0.58615
                 Variance Threshold 0.8    0.79260    0.58735
                                   NONE    0.79380    0.58908
--------------------------------------------------------------------------------

[('sm', SMOTE()), ('sel', VarianceThreshold(threshold=0.15999999999999998)),
 ('pca', PCA(n_components=10)), ('model', RandomForestClassifier(class_weight
='balanced_subsample', criterion='entropy',
                       min_samples_split=100, n_estimators=1000,
                       random_state=0))]
Selector: Variance Threshold 0.8 Model: RandomForest
              precision    recall  f1-score   support

           0       0.92      0.84      0.88      7963
           1       0.53      0.70      0.61      2037

    accuracy                           0.81     10000
   macro avg       0.73      0.77      0.74     10000
weighted avg       0.84      0.81      0.82     10000

[('sm', SMOTE()), ('sel', RFECV(estimator=RandomForestClassifier())), ('pca',
PCA(n_components=10)), ('model', RandomForestClassifier(class_weight='balance
d_subsample', criterion='entropy',
                       min_samples_split=100, n_estimators=1000,
                       random_state=0))]
Selector: RFECV RandomForest Model: RandomForest
              precision    recall  f1-score   support

           0       0.92      0.84      0.88      7963
           1       0.53      0.69      0.60      2037

    accuracy                           0.81     10000
   macro avg       0.72      0.77      0.74     10000
```

```
weighted avg       0.84      0.81      0.82     10000

[('sm', SMOTE()), ('sel', SelectFwe(alpha=0.01)), ('pca', PCA(n_components=1
0)), ('model', RandomForestClassifier(class_weight='balanced_subsample', crit
erion='entropy',
                       min_samples_split=100, n_estimators=1000,
                       random_state=0))]
Selector: FWE f_classif Model: RandomForest
              precision    recall  f1-score   support

           0       0.92      0.84      0.88      7963
           1       0.54      0.71      0.61      2037

    accuracy                           0.82     10000
   macro avg       0.73      0.78      0.75     10000
weighted avg       0.84      0.82      0.83     10000

[('sm', SMOTE()), ('pca', PCA(n_components=10)), ('model', RandomForestClassi
fier(class_weight='balanced_subsample', criterion='entropy',
                       min_samples_split=100, n_estimators=1000,
                       random_state=0))]
Selector: NONE Model: RandomForest
              precision    recall  f1-score   support

           0       0.91      0.84      0.87      7963
           1       0.52      0.69      0.60      2037

    accuracy                           0.81     10000
   macro avg       0.72      0.77      0.74     10000
weighted avg       0.84      0.81      0.82     10000


MODEL: GradientBoosting
----------------------
                           FWE f_classif    0.00000    0.00000
                      RFECV RandomForest    0.79200    0.58615
                  Variance Threshold 0.8    0.79260    0.58735
                                    NONE    0.79380    0.58908
-------------------------------------------------------------------------
--

MODEL: RandomForest
----------------------
                                    NONE    0.80920    0.59730
                      RFECV RandomForest    0.81220    0.60093
                  Variance Threshold 0.8    0.81490    0.60726
                           FWE f_classif    0.81710    0.61225
-------------------------------------------------------------------------
--

[('sm', SMOTE()), ('sel', VarianceThreshold(threshold=0.15999999999999998)),
 ('pca', PCA(n_components=10)), ('model', SVC(gamma='auto', random_state=0))]
Selector: Variance Threshold 0.8 Model: SVC
              precision    recall  f1-score   support

           0       0.92      0.82      0.87      7963
           1       0.50      0.73      0.59      2037
```

```
    accuracy                           0.80     10000
   macro avg       0.71      0.77      0.73     10000
weighted avg       0.84      0.80      0.81     10000

[('sm', SMOTE()), ('sel', RFECV(estimator=RandomForestClassifier())), ('pca',
PCA(n_components=10)), ('model', SVC(gamma='auto', random_state=0))]
Selector: RFECV RandomForest Model: SVC
              precision    recall  f1-score   support

           0       0.92      0.82      0.87      7963
           1       0.50      0.73      0.60      2037

    accuracy                           0.80     10000
   macro avg       0.71      0.77      0.73     10000
weighted avg       0.84      0.80      0.81     10000

[('sm', SMOTE()), ('sel', SelectFwe(alpha=0.01)), ('pca', PCA(n_components=1
0)), ('model', SVC(gamma='auto', random_state=0))]
====>>>> Failure in SVC, FWE f_classif
[('sm', SMOTE()), ('pca', PCA(n_components=10)), ('model', SVC(gamma='auto',
 random_state=0))]
Selector: NONE Model: SVC
              precision    recall  f1-score   support

           0       0.92      0.82      0.87      7963
           1       0.50      0.72      0.59      2037

    accuracy                           0.80     10000
   macro avg       0.71      0.77      0.73     10000
weighted avg       0.84      0.80      0.81     10000


MODEL: GradientBoosting
----------------------
                          FWE f_classif    0.00000    0.00000
                     RFECV RandomForest    0.79200    0.58615
                  Variance Threshold 0.8    0.79260    0.58735
                                   NONE    0.79380    0.58908
--------------------------------------------------------------------------
--

MODEL: RandomForest
----------------------
                                   NONE    0.80920    0.59730
                     RFECV RandomForest    0.81220    0.60093
                  Variance Threshold 0.8    0.81490    0.60726
                          FWE f_classif    0.81710    0.61225
--------------------------------------------------------------------------
--

MODEL: SVC
----------------------
                          FWE f_classif    0.00000    0.00000
                                   NONE    0.79810    0.59368
                  Variance Threshold 0.8    0.79810    0.59450
                     RFECV RandomForest    0.79870    0.59586
```

```
--------------------------------------------------------------------------------

================================================================================
PCA 14
--------------------------------------------------------------------------------
[('sm', SMOTE()), ('sel', VarianceThreshold(threshold=0.15999999999999998)),
 ('pca', PCA(n_components=10)), ('model', GradientBoostingClassifier(min_samp
les_split=42, random_state=0))]
Selector: Variance Threshold 0.8 Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.92      0.81      0.86      7963
           1       0.50      0.73      0.59      2037

    accuracy                           0.79     10000
   macro avg       0.71      0.77      0.73     10000
weighted avg       0.83      0.79      0.81     10000

[('sm', SMOTE()), ('sel', RFECV(estimator=RandomForestClassifier())), ('pca',
PCA(n_components=10)), ('model', GradientBoostingClassifier(min_samples_split
=42, random_state=0))]
Selector: RFECV RandomForest Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.92      0.82      0.86      7963
           1       0.50      0.71      0.59      2037

    accuracy                           0.80     10000
   macro avg       0.71      0.76      0.73     10000
weighted avg       0.83      0.80      0.81     10000

[('sm', SMOTE()), ('sel', SelectFwe(alpha=0.01)), ('pca', PCA(n_components=1
0)), ('model', GradientBoostingClassifier(min_samples_split=42, random_state=
0))]
Selector: FWE f_classif Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.92      0.82      0.87      7963
           1       0.51      0.73      0.60      2037

    accuracy                           0.80     10000
   macro avg       0.71      0.77      0.73     10000
weighted avg       0.84      0.80      0.81     10000

[('sm', SMOTE()), ('pca', PCA(n_components=10)), ('model', GradientBoostingCl
assifier(min_samples_split=42, random_state=0))]
Selector: NONE Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.92      0.81      0.86      7963
           1       0.49      0.72      0.58      2037

    accuracy                           0.79     10000
   macro avg       0.70      0.76      0.72     10000
```

```
weighted avg       0.83      0.79      0.80     10000


MODEL: GradientBoosting
---------------------
                                    NONE    0.79010    0.58328
                     RFECV RandomForest    0.79540    0.58650
                  Variance Threshold 0.8    0.79430    0.58967
                           FWE f_classif    0.79940    0.59686
--------------------------------------------------------------------------------

[('sm', SMOTE()), ('sel', VarianceThreshold(threshold=0.15999999999999998)),
 ('pca', PCA(n_components=10)), ('model', RandomForestClassifier(class_weight
='balanced_subsample', criterion='entropy',
                       min_samples_split=100, n_estimators=1000,
                       random_state=0))]
Selector: Variance Threshold 0.8 Model: RandomForest
              precision    recall  f1-score   support

           0       0.92      0.84      0.88      7963
           1       0.53      0.70      0.60      2037

    accuracy                           0.81     10000
   macro avg       0.72      0.77      0.74     10000
weighted avg       0.84      0.81      0.82     10000

[('sm', SMOTE()), ('sel', RFECV(estimator=RandomForestClassifier())), ('pca',
PCA(n_components=10)), ('model', RandomForestClassifier(class_weight='balance
d_subsample', criterion='entropy',
                       min_samples_split=100, n_estimators=1000,
                       random_state=0))]
Selector: RFECV RandomForest Model: RandomForest
              precision    recall  f1-score   support

           0       0.91      0.84      0.88      7963
           1       0.53      0.69      0.60      2037

    accuracy                           0.81     10000
   macro avg       0.72      0.77      0.74     10000
weighted avg       0.84      0.81      0.82     10000

[('sm', SMOTE()), ('sel', SelectFwe(alpha=0.01)), ('pca', PCA(n_components=1
0)), ('model', RandomForestClassifier(class_weight='balanced_subsample', crit
erion='entropy',
                       min_samples_split=100, n_estimators=1000,
                       random_state=0))]
====>>>> Failure in RandomForest, FWE f_classif
[('sm', SMOTE()), ('pca', PCA(n_components=10)), ('model', RandomForestClassi
fier(class_weight='balanced_subsample', criterion='entropy',
                       min_samples_split=100, n_estimators=1000,
                       random_state=0))]
Selector: NONE Model: RandomForest
              precision    recall  f1-score   support

           0       0.92      0.84      0.87      7963
           1       0.52      0.70      0.60      2037
```

```
    accuracy                           0.81     10000
   macro avg       0.72      0.77      0.74     10000
weighted avg       0.84      0.81      0.82     10000


MODEL: GradientBoosting
----------------------
                                  NONE    0.79010    0.58328
                     RFECV RandomForest    0.79540    0.58650
                  Variance Threshold 0.8    0.79430    0.58967
                           FWE f_classif    0.79940    0.59686
--------------------------------------------------------------------------------
--

MODEL: RandomForest
----------------------
                           FWE f_classif    0.00000    0.00000
                     RFECV RandomForest    0.81150    0.59902
                                  NONE    0.80950    0.60038
                  Variance Threshold 0.8    0.81070    0.60156
--------------------------------------------------------------------------------
--

[('sm', SMOTE()), ('sel', VarianceThreshold(threshold=0.15999999999999998)),
 ('pca', PCA(n_components=10)), ('model', SVC(gamma='auto', random_state=0))]
Selector: Variance Threshold 0.8 Model: SVC
              precision    recall  f1-score   support

           0       0.92      0.82      0.87      7963
           1       0.50      0.73      0.59      2037

    accuracy                           0.80     10000
   macro avg       0.71      0.77      0.73     10000
weighted avg       0.84      0.80      0.81     10000

[('sm', SMOTE()), ('sel', RFECV(estimator=RandomForestClassifier())), ('pca',
PCA(n_components=10)), ('model', SVC(gamma='auto', random_state=0))]
Selector: RFECV RandomForest Model: SVC
              precision    recall  f1-score   support

           0       0.92      0.81      0.86      7963
           1       0.49      0.72      0.59      2037

    accuracy                           0.79     10000
   macro avg       0.71      0.77      0.72     10000
weighted avg       0.83      0.79      0.80     10000

[('sm', SMOTE()), ('sel', SelectFwe(alpha=0.01)), ('pca', PCA(n_components=1
0)), ('model', SVC(gamma='auto', random_state=0))]
====>>>> Failure in SVC, FWE f_classif
[('sm', SMOTE()), ('pca', PCA(n_components=10)), ('model', SVC(gamma='auto',
 random_state=0))]
Selector: NONE Model: SVC
              precision    recall  f1-score   support

           0       0.92      0.81      0.86      7963
```

```
           1       0.50      0.73      0.59      2037

    accuracy                           0.80     10000
   macro avg       0.71      0.77      0.73     10000
weighted avg       0.83      0.80      0.81     10000


MODEL: GradientBoosting
----------------------
                                   NONE    0.79010    0.58328
                      RFECV RandomForest    0.79540    0.58650
                  Variance Threshold 0.8    0.79430    0.58967
                           FWE f_classif    0.79940    0.59686
-------------------------------------------------------------------------------
--

MODEL: RandomForest
----------------------
                           FWE f_classif    0.00000    0.00000
                      RFECV RandomForest    0.81150    0.59902
                                   NONE    0.80950    0.60038
                  Variance Threshold 0.8    0.81070    0.60156
-------------------------------------------------------------------------------
--

MODEL: SVC
----------------------
                           FWE f_classif    0.00000    0.00000
                      RFECV RandomForest    0.79180    0.58559
                                   NONE    0.79650    0.59210
                  Variance Threshold 0.8    0.79800    0.59405
-------------------------------------------------------------------------------
--
```

In [168]:
```
npca_a = ncap_a
```

In [169]:
```
for i, d, m in zip(ncap_a, desc_strings_mmspca_a, all_models_mmspca_a):
    print(f'PCA {i}')
    print('-' * 120)
    print(d)
    plot_graph_1(m)
```

PCA 5

MODEL: GradientBoosting

FWE f_classif	0.00000	0.00000
NONE	0.78990	0.58552
Variance Threshold 0.8	0.79250	0.58690
RFECV RandomForest	0.80290	0.59285

MODEL: RandomForest

FWE f_classif	0.00000	0.00000
Variance Threshold 0.8	0.80930	0.59537
RFECV RandomForest	0.81420	0.60214
NONE	0.81330	0.60604

MODEL: SVC

FWE f_classif	0.00000	0.00000
RFECV RandomForest	0.79550	0.58960
Variance Threshold 0.8	0.79550	0.59125
NONE	0.79820	0.59478

PCA 6

MODEL: GradientBoosting

FWE f_classif	0.00000	0.00000
RFECV RandomForest	0.79120	0.58390
NONE	0.79150	0.58458
Variance Threshold 0.8	0.79280	0.58626

MODEL: RandomForest

FWE f_classif	0.00000	0.00000
Variance Threshold 0.8	0.80980	0.59669
NONE	0.81250	0.59944
RFECV RandomForest	0.81510	0.60601

MODEL: SVC

Variance Threshold 0.8	0.79450	0.58957
NONE	0.79550	0.58960

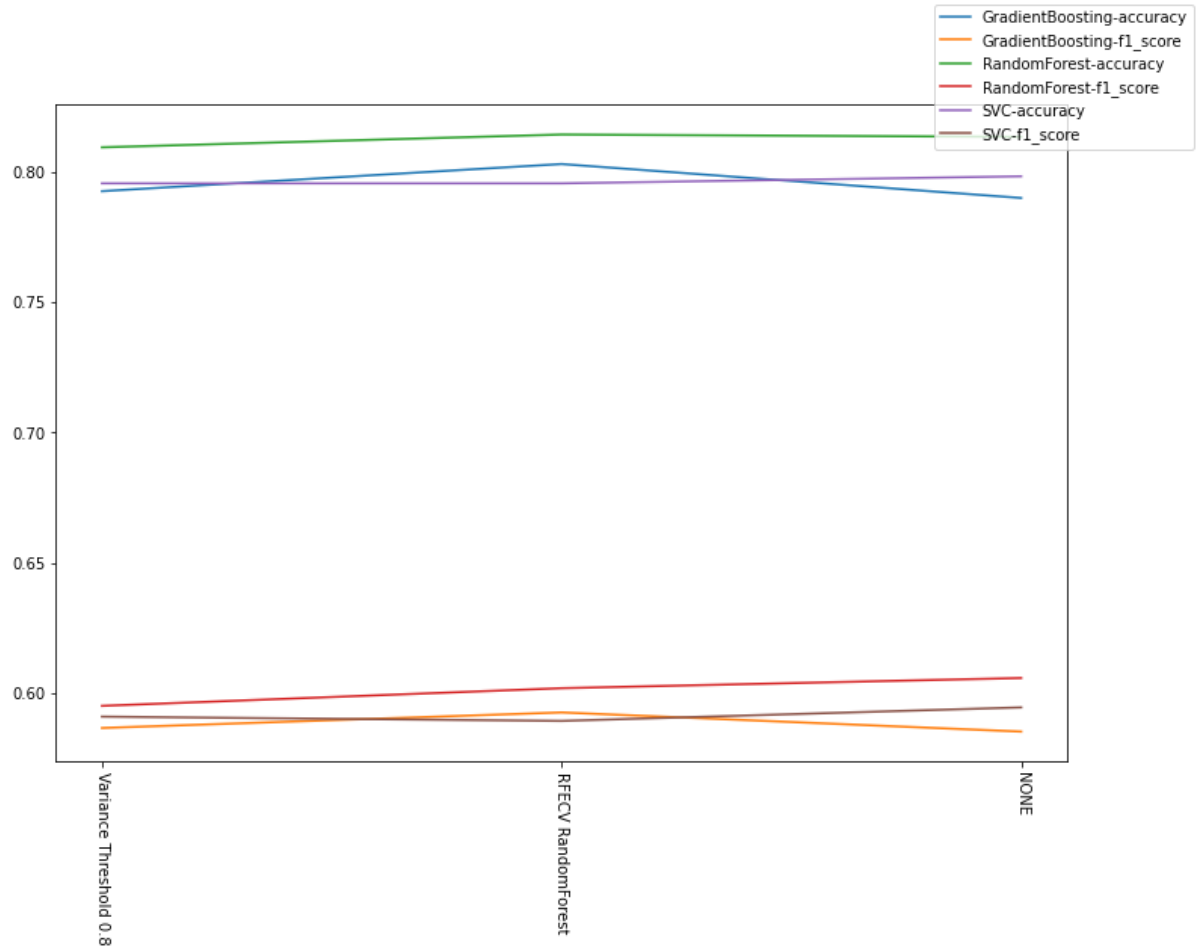
```
                     RFECV RandomForest    0.79670    0.59037
                          FWE f_classif    0.79990    0.60004
--------------------------------------------------------------------------------

PCA 7
--------------------------------------------------------------------------------
----------------------------------------

MODEL: GradientBoosting
---------------------
                                  NONE    0.78970    0.58299
                 Variance Threshold 0.8    0.79190    0.58784
                     RFECV RandomForest    0.80210    0.58984
                          FWE f_classif    0.80290    0.60142
--------------------------------------------------------------------------------

MODEL: RandomForest
---------------------
                          FWE f_classif    0.00000    0.00000
                 Variance Threshold 0.8    0.81050    0.59979
                                  NONE    0.81160    0.60387
                     RFECV RandomForest    0.81470    0.60600
--------------------------------------------------------------------------------

MODEL: SVC
---------------------
                          FWE f_classif    0.00000    0.00000
                                  NONE    0.79810    0.59384
                     RFECV RandomForest    0.79690    0.59404
                 Variance Threshold 0.8    0.79770    0.59516
--------------------------------------------------------------------------------

PCA 8
--------------------------------------------------------------------------------
----------------------------------------

MODEL: GradientBoosting
---------------------
                          FWE f_classif    0.00000    0.00000
                                  NONE    0.79170    0.58564
                     RFECV RandomForest    0.79900    0.58778
                 Variance Threshold 0.8    0.79700    0.59302
--------------------------------------------------------------------------------

MODEL: RandomForest
---------------------
                          FWE f_classif    0.00000    0.00000
                     RFECV RandomForest    0.81230    0.60140
                                  NONE    0.81250    0.60435
                 Variance Threshold 0.8    0.81320    0.60574
--------------------------------------------------------------------------------
```

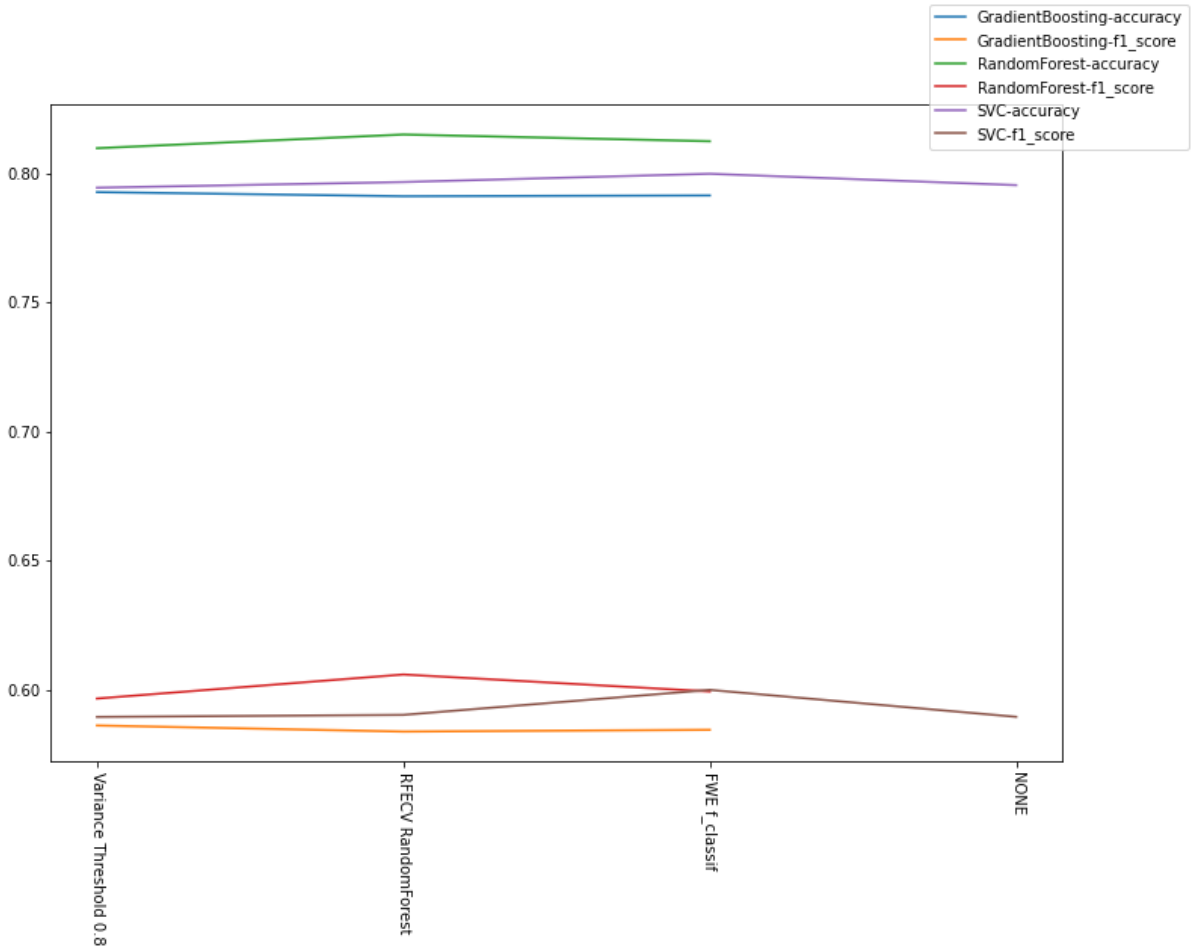
```
MODEL: SVC
---------------------
                          FWE f_classif    0.00000    0.00000
                    RFECV RandomForest    0.79430    0.59081
                                  NONE    0.79770    0.59369
                Variance Threshold 0.8    0.79810    0.59482
-------------------------------------------------------------------------
--

PCA 9
-------------------------------------------------------------------------
-----------------------------------------

MODEL: GradientBoosting
---------------------
                                  NONE    0.78660    0.57809
                    RFECV RandomForest    0.79170    0.58181
                Variance Threshold 0.8    0.79690    0.59388
                          FWE f_classif    0.80050    0.59932
-------------------------------------------------------------------------
--

MODEL: RandomForest
---------------------
                                  NONE    0.81140    0.60127
                    RFECV RandomForest    0.81440    0.60137
                Variance Threshold 0.8    0.81280    0.60289
                          FWE f_classif    0.81890    0.61769
-------------------------------------------------------------------------
--

MODEL: SVC
---------------------
                          FWE f_classif    0.00000    0.00000
                Variance Threshold 0.8    0.79380    0.58842
                    RFECV RandomForest    0.79490    0.58988
                                  NONE    0.79860    0.59346
-------------------------------------------------------------------------
--

PCA 10
-------------------------------------------------------------------------
-----------------------------------------

MODEL: GradientBoosting
---------------------
                          FWE f_classif    0.00000    0.00000
                    RFECV RandomForest    0.80030    0.58867
                                  NONE    0.79500    0.58918
                Variance Threshold 0.8    0.79650    0.58947
-------------------------------------------------------------------------
--

MODEL: RandomForest
---------------------
                    RFECV RandomForest    0.81080    0.59881
```

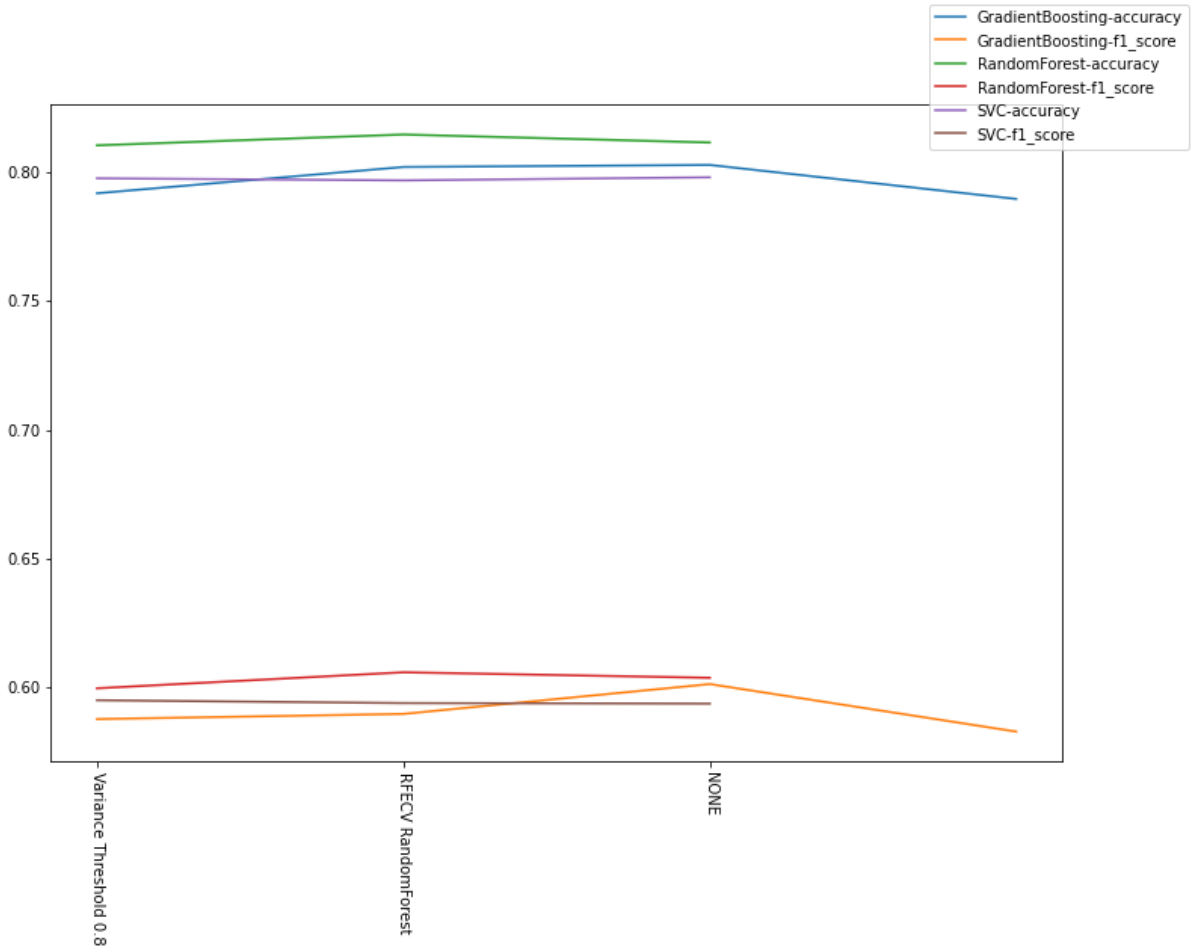
```
                  Variance Threshold 0.8    0.81170    0.60165
                                    NONE    0.81460    0.60737
                             FWE f_classif    0.81890    0.61704
-------------------------------------------------------------------------
--

MODEL: SVC
----------------------
                      RFECV RandomForest    0.79120    0.58653
                  Variance Threshold 0.8    0.79710    0.59232
                                    NONE    0.79820    0.59510
                             FWE f_classif    0.80200    0.60337
-------------------------------------------------------------------------
--

PCA 11
-------------------------------------------------------------------------
-----------------------------------------

MODEL: GradientBoosting
----------------------
                                    NONE    0.79300    0.58781
                  Variance Threshold 0.8    0.79520    0.58859
                      RFECV RandomForest    0.80020    0.59141
                             FWE f_classif    0.80360    0.60114
-------------------------------------------------------------------------
--

MODEL: RandomForest
----------------------
                      RFECV RandomForest    0.81240    0.59288
                  Variance Threshold 0.8    0.81210    0.60400
                                    NONE    0.81270    0.60460
                             FWE f_classif    0.81740    0.61379
-------------------------------------------------------------------------
--

MODEL: SVC
----------------------
                             FWE f_classif    0.00000    0.00000
                      RFECV RandomForest    0.79630    0.59187
                  Variance Threshold 0.8    0.79810    0.59335
                                    NONE    0.79780    0.59365
-------------------------------------------------------------------------
--

PCA 12
-------------------------------------------------------------------------
-----------------------------------------

MODEL: GradientBoosting
----------------------
                             FWE f_classif    0.00000    0.00000
                  Variance Threshold 0.8    0.79380    0.58941
                                    NONE    0.79400    0.58997
                      RFECV RandomForest    0.80150    0.59182
-------------------------------------------------------------------------
```

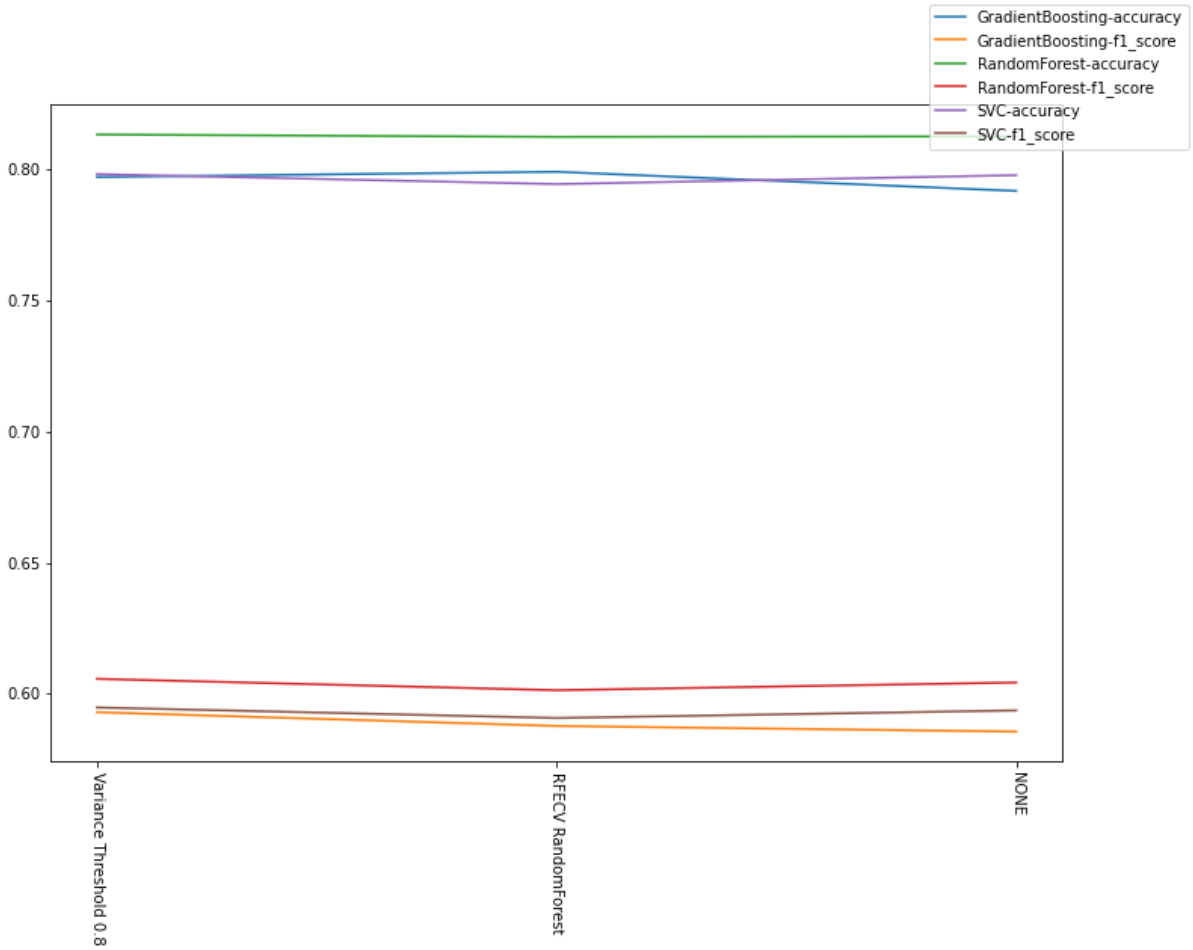
```
--

MODEL: RandomForest
---------------------
                                  NONE    0.80930    0.59929
                 Variance Threshold 0.8    0.81100    0.60059
                     RFECV RandomForest    0.81520    0.60597
                          FWE f_classif    0.81350    0.60778
-------------------------------------------------------------------------------
--

MODEL: SVC
---------------------
                          FWE f_classif    0.00000    0.00000
                     RFECV RandomForest    0.79530    0.59068
                 Variance Threshold 0.8    0.79730    0.59256
                                  NONE    0.79650    0.59259
-------------------------------------------------------------------------------
--

PCA 13
-------------------------------------------------------------------------------
-----------------------------------------

MODEL: GradientBoosting
---------------------
                          FWE f_classif    0.00000    0.00000
                     RFECV RandomForest    0.79200    0.58615
                 Variance Threshold 0.8    0.79260    0.58735
                                  NONE    0.79380    0.58908
-------------------------------------------------------------------------------
--

MODEL: RandomForest
---------------------
                                  NONE    0.80920    0.59730
                     RFECV RandomForest    0.81220    0.60093
                 Variance Threshold 0.8    0.81490    0.60726
                          FWE f_classif    0.81710    0.61225
-------------------------------------------------------------------------------
--

MODEL: SVC
---------------------
                          FWE f_classif    0.00000    0.00000
                                  NONE    0.79810    0.59368
                 Variance Threshold 0.8    0.79810    0.59450
                     RFECV RandomForest    0.79870    0.59586
-------------------------------------------------------------------------------
--

PCA 14
-------------------------------------------------------------------------------
-----------------------------------------

MODEL: GradientBoosting
---------------------
```

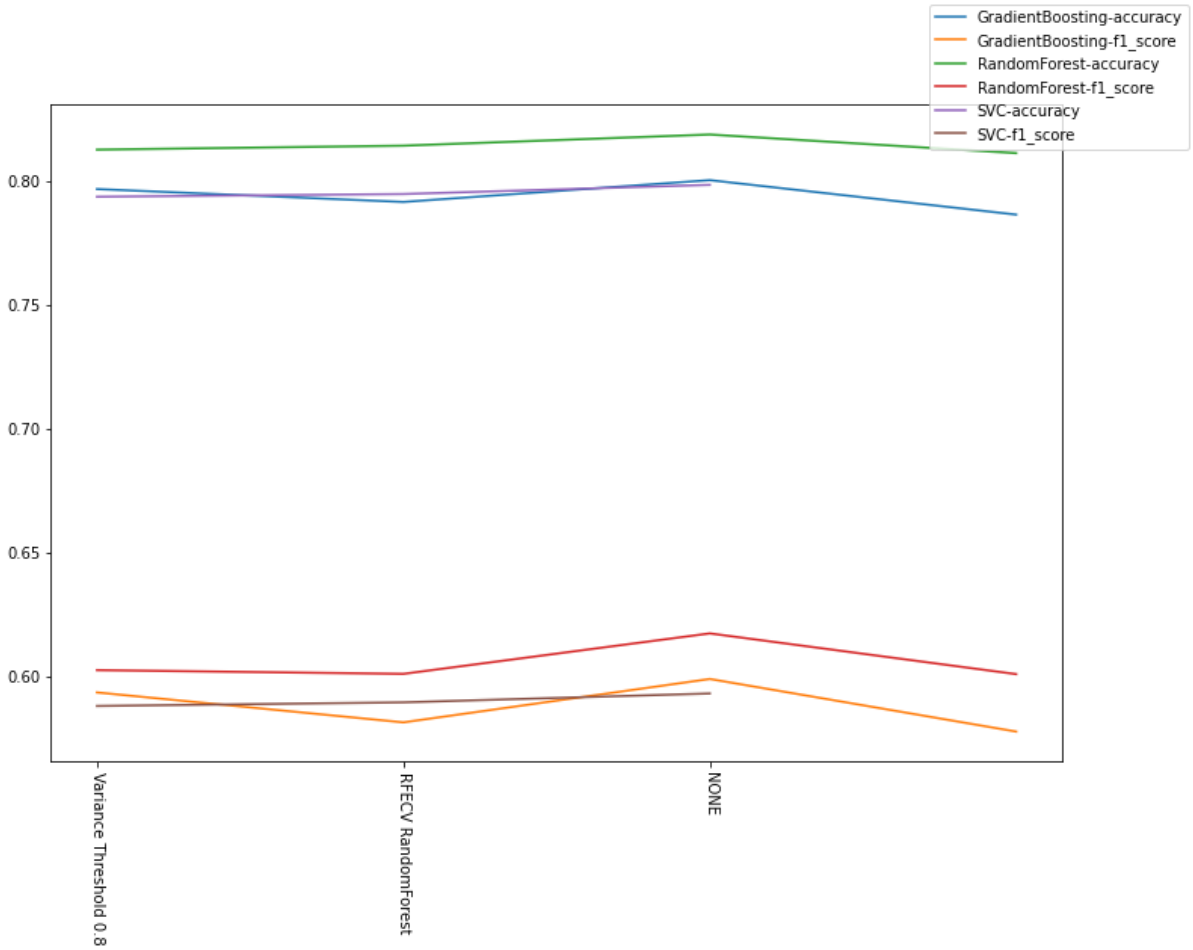
```
                                NONE    0.79010    0.58328
                    RFECV RandomForest    0.79540    0.58650
                Variance Threshold 0.8    0.79430    0.58967
                        FWE f_classif    0.79940    0.59686
--------------------------------------------------------------------------------
--

MODEL: RandomForest
----------------------
                        FWE f_classif    0.00000    0.00000
                    RFECV RandomForest    0.81150    0.59902
                                NONE    0.80950    0.60038
                Variance Threshold 0.8    0.81070    0.60156
--------------------------------------------------------------------------------
--

MODEL: SVC
----------------------
                        FWE f_classif    0.00000    0.00000
                    RFECV RandomForest    0.79180    0.58559
                                NONE    0.79650    0.59210
                Variance Threshold 0.8    0.79800    0.59405
--------------------------------------------------------------------------------
--
```

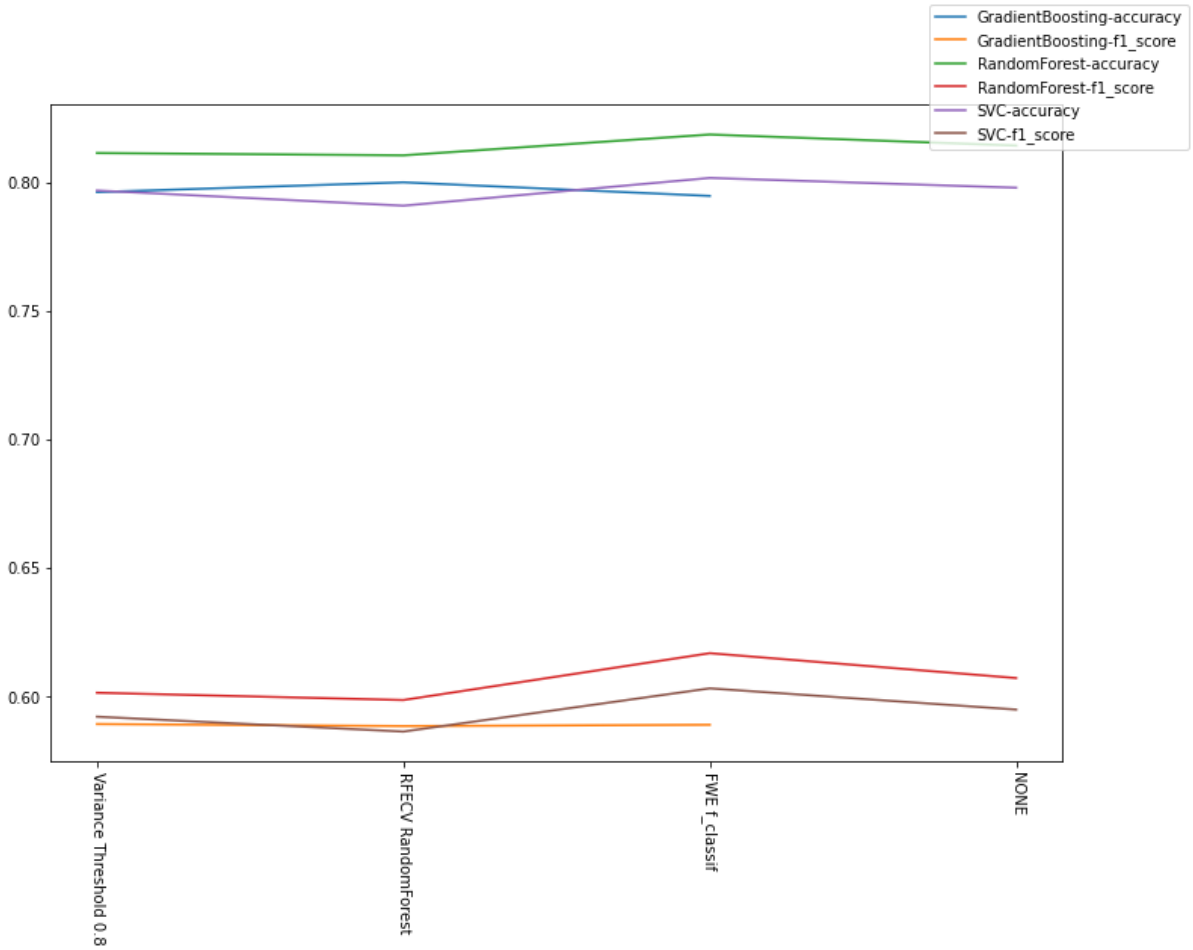


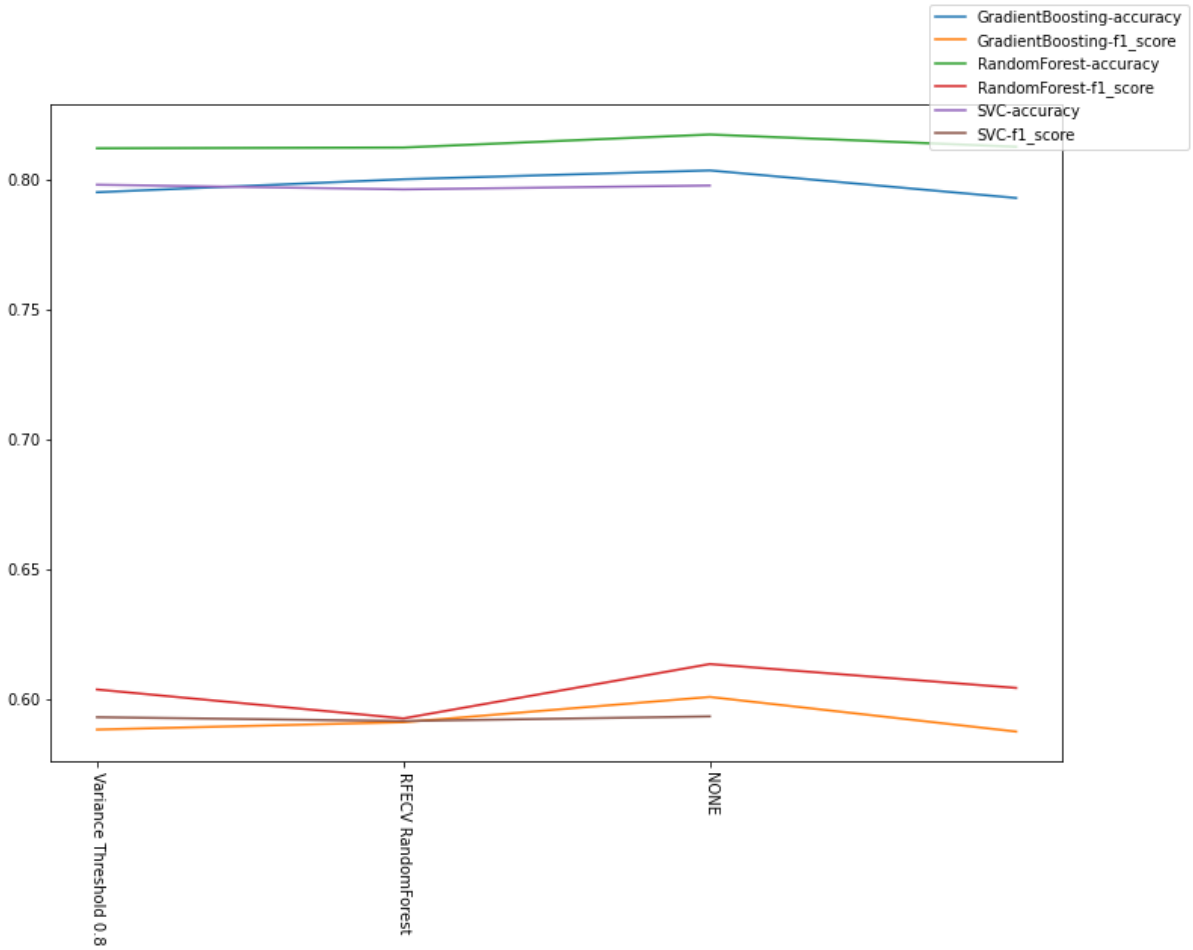

GradientBoosting-accuracy
GradientBoosting-f1_score
RandomForest-accuracy
RandomForest-f1_score
SVC-accuracy
SVC-f1_score
0.80
0.75
0.70
0.65
0.60
Variance Threshold 0.8
RFECV RandomForest
FWE f_classif
NONE

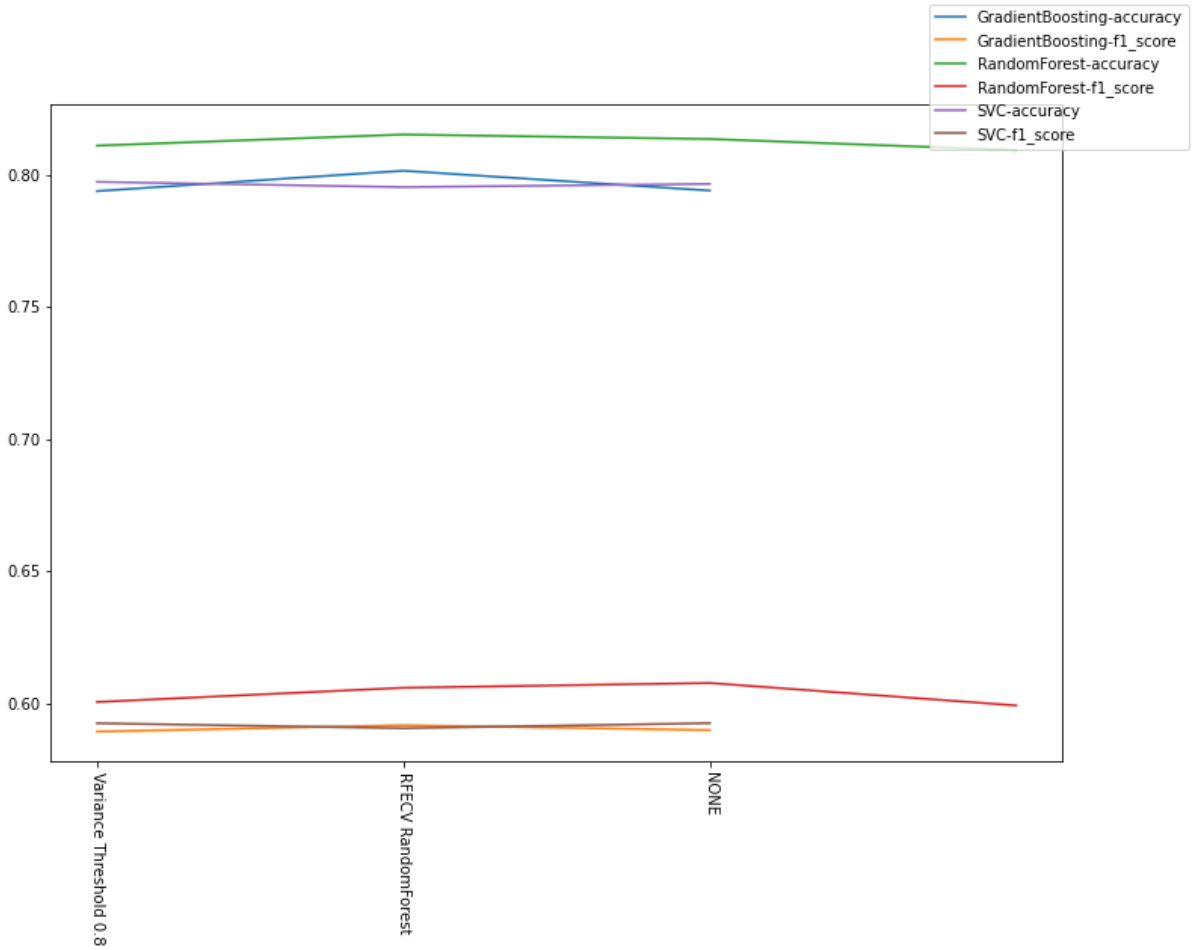

GradientBoosting-accuracy
GradientBoosting-f1_score
RandomForest-accuracy
RandomForest-f1_score
SVC-accuracy
SVC-f1_score
0.80
0.75
0.70
0.65
0.60
Variance Threshold 0.8
RFECV RandomForest
NONE

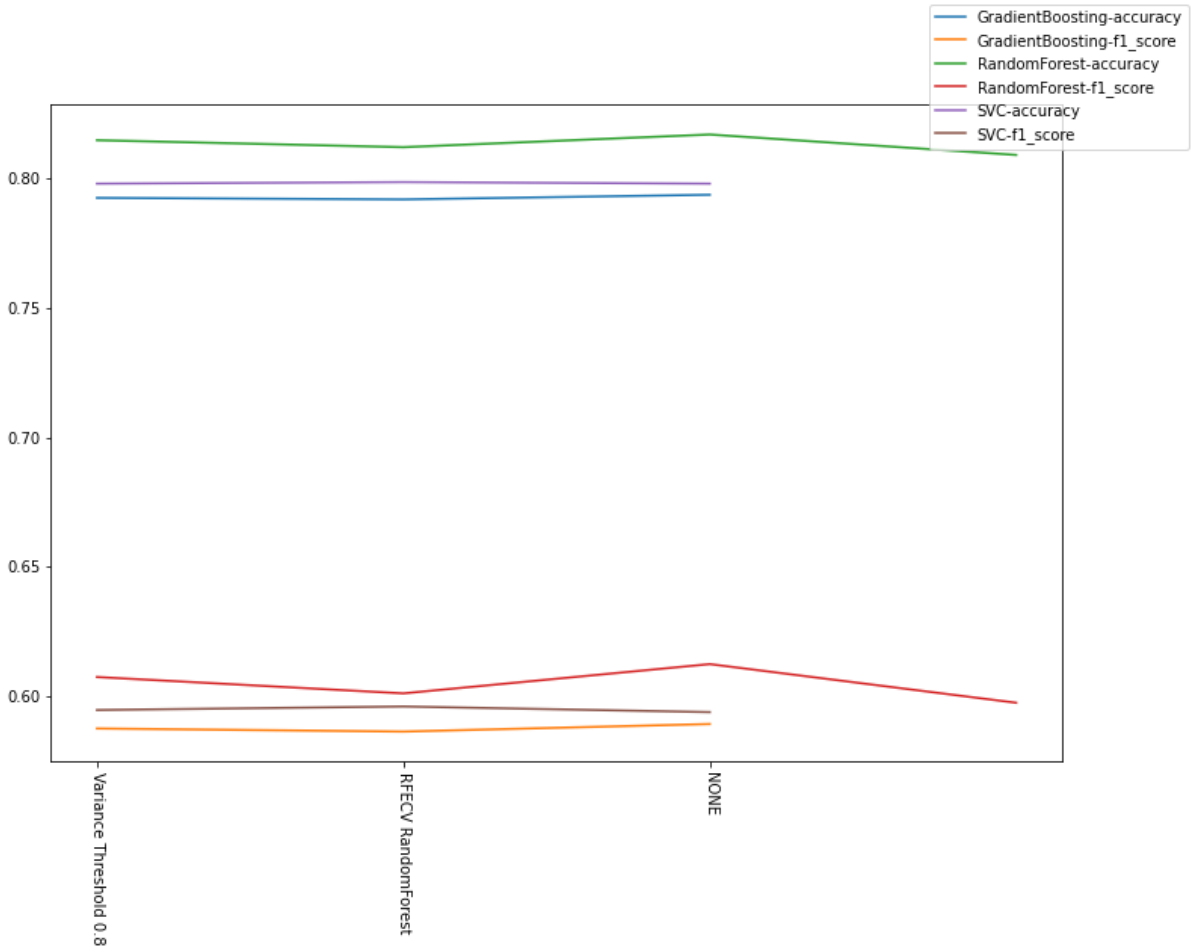

GradientBoosting-accuracy
GradientBoosting-f1_score
RandomForest-accuracy
RandomForest-f1_score
SVC-accuracy
SVC-f1_score
0.80
0.75
0.70
0.65
0.60
Variance Threshold 0.8
RFECV RandomForest
NONE

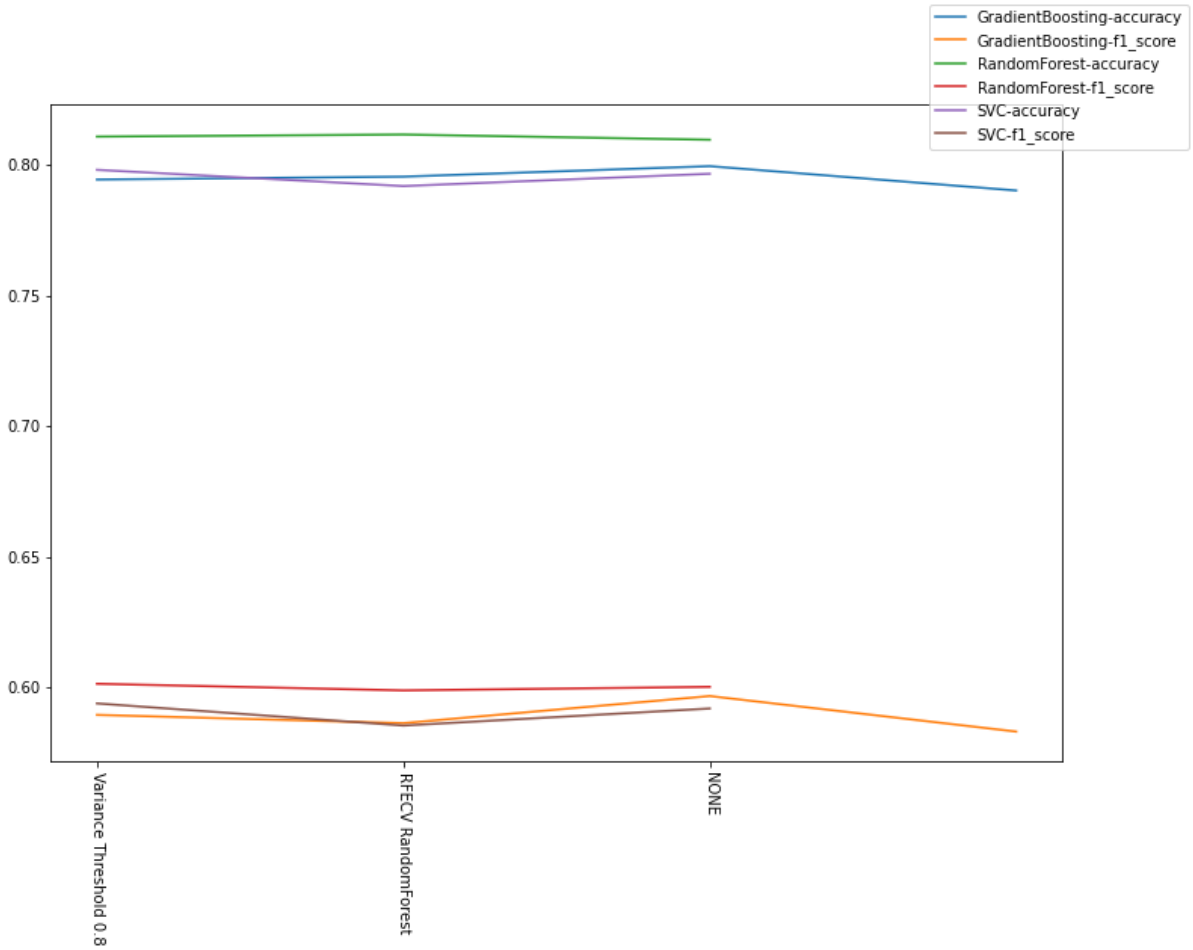

GradientBoosting-accuracy
GradientBoosting-f1_score
RandomForest-accuracy
RandomForest-f1_score
SVC-accuracy
SVC-f1_score
0.80
0.75
0.70
0.65
0.60
Variance Threshold 0.8
RFECV RandomForest
NONE


GradientBoosting-accuracy
GradientBoosting-f1_score
RandomForest-accuracy
RandomForest-f1_score
SVC-accuracy
SVC-f1_score
0.80
0.75
0.70
0.65
0.60
Variance Threshold 0.8
RFECV RandomForest
FWE f_classif
NONE


GradientBoosting-accuracy
GradientBoosting-f1_score
RandomForest-accuracy
RandomForest-f1_score
SVC-accuracy
SVC-f1_score
0.80
0.75
0.70
0.65
0.60
Variance Threshold 0.8
RFECV RandomForest
NONE


GradientBoosting-accuracy
GradientBoosting-f1_score
RandomForest-accuracy
RandomForest-f1_score
SVC-accuracy
SVC-f1_score
0.80
0.75
0.70
0.65
0.60
Variance Threshold 0.8
RFECV RandomForest
NONE


GradientBoosting-accuracy
GradientBoosting-f1_score
RandomForest-accuracy
RandomForest-f1_score
SVC-accuracy
SVC-f1_score
0.80
0.75
0.70
0.65
0.60
Variance Threshold 0.8
RFECV RandomForest
NONE

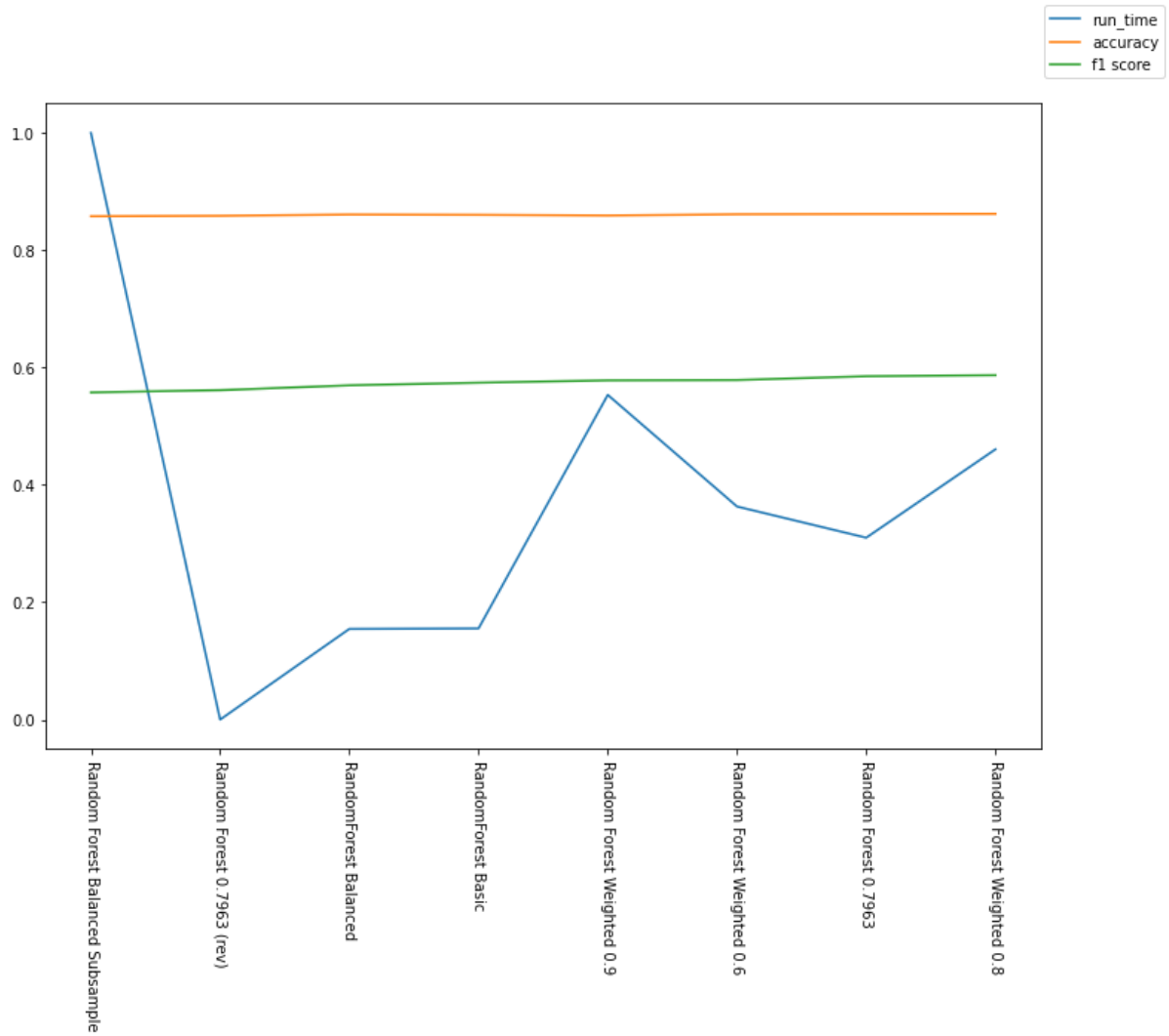


Research

Experiment with weighted features in Random Forest

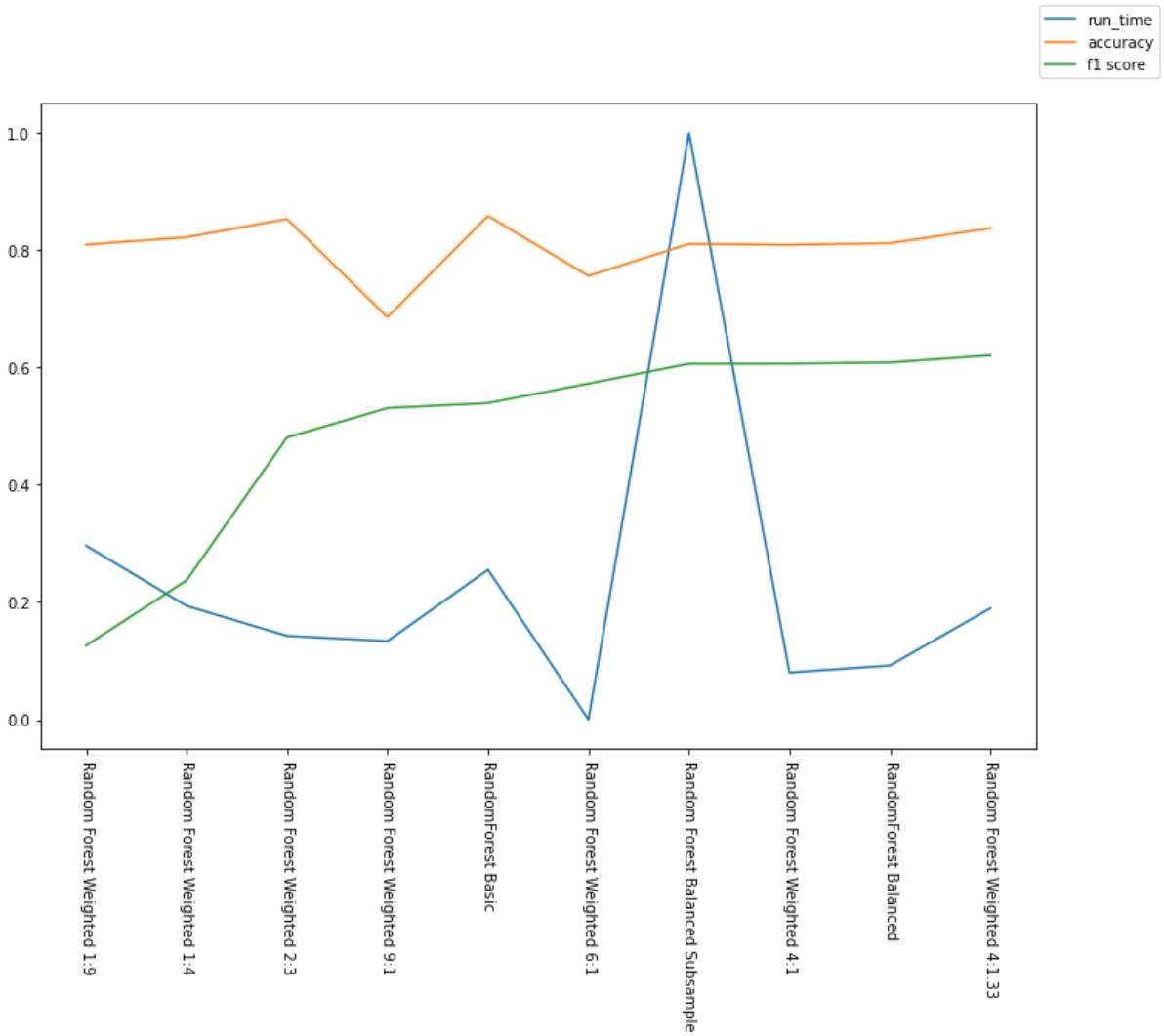
In [182]:
```
# First check with default parameters, with just this parameter changed
basic_models = [
    ('RandomForest Basic', lambda: RandomForestClassifier(random_state=1)),
    ('RandomForest Balanced', lambda: RandomForestClassifier(random_state=1, c
lass_weight='balanced')),
    ('Random Forest Balanced Subsample', lambda: RandomForestClassifier(random
_state=1, class_weight='balanced_subsample')),
    ('Random Forest Weighted 0.8', lambda: RandomForestClassifier(random_state
=1, class_weight={1: 0.2, 0: 0.8})),
    ('Random Forest Weighted 0.6', lambda: RandomForestClassifier(random_state
=1, class_weight={1: 0.4, 0: 0.6})),
    ('Random Forest Weighted 0.9', lambda: RandomForestClassifier(random_state
=1, class_weight={1: 0.1, 0: 0.9})),
    ('Random Forest 0.7963', lambda: RandomForestClassifier(random_state=1, cl
ass_weight={1: 2037, 0: 7963})),
    ('Random Forest 0.7963 (rev)', lambda: RandomForestClassifier(random_state
=1, class_weight={0: 2037, 1: 7963})),
]
compare_basic_models(run_basic_model, basic_models)
```

```
RandomForest Basic <function <lambda> at 0x000001EC43C1E5E0>
RandomForest Balanced <function <lambda> at 0x000001EC43C1E550>
Random Forest Balanced Subsample <function <lambda> at 0x000001EC43FA2D30>
Random Forest Weighted 0.8 <function <lambda> at 0x000001EC43FA23A0>
Random Forest Weighted 0.6 <function <lambda> at 0x000001EC43FA2550>
Random Forest Weighted 0.9 <function <lambda> at 0x000001EC43FA2700>
Random Forest 0.7963 <function <lambda> at 0x000001EC43FA28B0>
Random Forest 0.7963 (rev) <function <lambda> at 0x000001EC43FA2A60>

[0.8603999488217603, 0.8609005089738355, 0.8579003286977658, 0.86200028901782
67, 0.8613003689778251, 0.8592000087377483, 0.8617004690298435, 0.85870076884
98315]

[0.5742653346426003, 0.5697548162058486, 0.5577101012466912, 0.58704804321528
09, 0.578779895715622, 0.5782437561433658, 0.5852288885179198, 0.561476369894
2369]

-------------------------------------------------------------------------------
-----------------------------------------
                                  Model  ----     Accuracy   F1 Score   Run
Time
-------------------------------------------------------------------------------
-----------------------------------------
        Random Forest Balanced Subsample  ----        0.858      0.558
1.000
              Random Forest 0.7963 (rev)  ----        0.859      0.561
0.000
                   RandomForest Balanced  ----        0.861      0.570
0.155
                      RandomForest Basic  ----        0.860      0.574
0.155
              Random Forest Weighted 0.9  ----        0.859      0.578
0.553
              Random Forest Weighted 0.6  ----        0.861      0.579
0.363
                    Random Forest 0.7963  ----        0.862      0.585
0.310
              Random Forest Weighted 0.8  ----        0.862      0.587
0.461
-------------------------------------------------------------------------------
-----------------------------------------
```


run_time
accuracy
f1 score
1.0
0.8
0.6
0.4
0.2
0.0
Random Forest Balanced Subsample
Random Forest 0.7963 (rev)
RandomForest Balanced
RandomForest Basic
Random Forest Weighted 0.9
Random Forest Weighted 0.6
Random Forest 0.7963
Random Forest Weighted 0.8

In [245]:
```
# Now check with the parameters obtained in the baseline
basic_models = [
    ('RandomForest Basic', lambda: RandomForestClassifier(random_state=1, crit
erion='entropy', min_samples_split=100, n_estimators=1000)),
    ('RandomForest Balanced', lambda: RandomForestClassifier(random_state=1, c
lass_weight='balanced', criterion='entropy', min_samples_split=100, n_estimato
rs=1000)),
    ('Random Forest Balanced Subsample', lambda: RandomForestClassifier(random
_state=1, class_weight='balanced_subsample', criterion='entropy', min_samples_
split=100, n_estimators=1000)),
    ('Random Forest Weighted 1:4', lambda: RandomForestClassifier(random_state
=1, class_weight={1: 0.2, 0: 0.8}, criterion='entropy', min_samples_split=100,
n_estimators=1000)),
    ('Random Forest Weighted 2:3', lambda: RandomForestClassifier(random_state
=1, class_weight={1: 0.4, 0: 0.6}, criterion='entropy', min_samples_split=100,
n_estimators=1000)),
    ('Random Forest Weighted 4:1', lambda: RandomForestClassifier(random_state
=1, class_weight={1: 4, 0: 1}, criterion='entropy', min_samples_split=100, n_e
stimators=1000)),
    ('Random Forest Weighted 4:1.33', lambda: RandomForestClassifier(random_st
ate=1, class_weight={1: 4, 0: 1.33}, criterion='entropy', min_samples_split=10
0, n_estimators=1000)),
    ('Random Forest Weighted 6:1', lambda: RandomForestClassifier(random_state
=1, class_weight={1: 6, 0: 1}, criterion='entropy', min_samples_split=100, n_e
stimators=1000)),
    ('Random Forest Weighted 1:9', lambda: RandomForestClassifier(random_state
=1, class_weight={1: 0.1, 0: 0.9}, criterion='entropy', min_samples_split=100,
n_estimators=1000)),
    ('Random Forest Weighted 9:1', lambda: RandomForestClassifier(random_state
=1, class_weight={1: 0.9, 0: 0.1}, criterion='entropy', min_samples_split=100,
n_estimators=1000)),

]
compare_basic_models(run_basic_model, basic_models)
```

```
RandomForest Basic <function <lambda> at 0x000001D924FA4EE0>
RandomForest Balanced <function <lambda> at 0x000001D924FA4940>
Random Forest Balanced Subsample <function <lambda> at 0x000001D924FA4160>
Random Forest Weighted 1:4 <function <lambda> at 0x000001D924FA40D0>
Random Forest Weighted 2:3 <function <lambda> at 0x000001D924FA41F0>
Random Forest Weighted 4:1 <function <lambda> at 0x000001D924FA4D30>
Random Forest Weighted 4:1.33 <function <lambda> at 0x000001D924FA4DC0>
Random Forest Weighted 6:1 <function <lambda> at 0x000001D924FA49D0>
Random Forest Weighted 1:9 <function <lambda> at 0x000001D924FA4700>
Random Forest Weighted 9:1 <function <lambda> at 0x000001D924FA4280>

[0.8584003487417762, 0.8120002050010166, 0.8108001448929902, 0.82210010578916
64, 0.8532003683296954, 0.8092998447129301, 0.8374993068613168, 0.75629928036
00144, 0.8096005648810214, 0.6858987146148201]

[0.5394276378866715, 0.6085874497161349, 0.6065484980293258, 0.23713837863635
384, 0.48082540319439254, 0.6066130501763358, 0.6208418323813478, 0.572675352
1214187, 0.12621721702162433, 0.5309915187663873]

------------------------------------------------------------------------------
-----------------------------------------
                                  Model  ----     Accuracy   F1 Score   Run
Time
------------------------------------------------------------------------------
-----------------------------------------
             Random Forest Weighted 1:9  ----        0.810      0.126
0.296
             Random Forest Weighted 1:4  ----        0.822      0.237
0.194
             Random Forest Weighted 2:3  ----        0.853      0.481
0.143
             Random Forest Weighted 9:1  ----        0.686      0.531
0.134
                     RandomForest Basic  ----        0.858      0.539
0.255
             Random Forest Weighted 6:1  ----        0.756      0.573
0.000
       Random Forest Balanced Subsample  ----        0.811      0.607
1.000
             Random Forest Weighted 4:1  ----        0.809      0.607
0.080
                  RandomForest Balanced  ----        0.812      0.609
0.092
          Random Forest Weighted 4:1.33  ----        0.837      0.621
0.190
------------------------------------------------------------------------------
-----------------------------------------
```

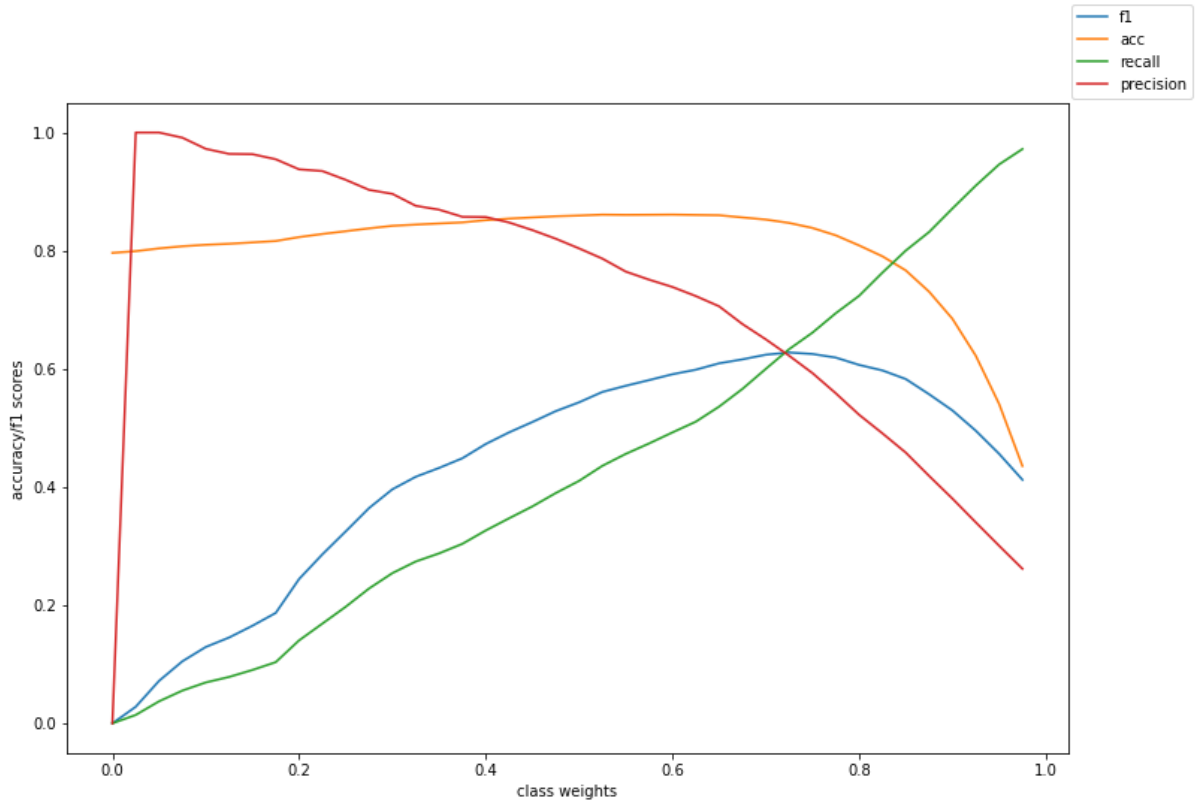

run_time
accuracy
f1 score
1.0
0.8
0.6
0.4
0.2
0.0
Random Forest Weighted 1:9
Random Forest Weighted 1:4
Random Forest Weighted 2:3
Random Forest Weighted 9:1
RandomForest Basic
Random Forest Weighted 6:1
Random Forest Balanced Subsample
Random Forest Weighted 4:1
RandomForest Balanced
Random Forest Weighted 4:1.33

In [282]:
```
# Now plot with by varying the params
import copy
def compare_weighted_rfcs(X, y):
    """
    Compare the effect of alternate class weights on RFC
    """
    weights = []
    f1_scores = []
    acc_scores = []
    recall_scores = []
    precision_scores = []
    for i in np.arange(0, 1.1, 0.025):
        desc = "Random Forest: Weight = %f" % i
        if (1-i) > 0:
            cw = {1: i, 0: 1-i}
            rfc = RandomForestClassifier(random_state=1, class_weight=cw, crit
erion='entropy',\
                                                    min_samples_split=100, n_est
imators=1000)
            rfc.fit(X, y)
            y_pred = cross_val_predict(rfc, X, y,\
                                       cv=model_selection.StratifiedKFold(n_sp
lits=6, shuffle=True, random_state=2),\
                                       n_jobs=7)
            f1 = metrics.f1_score(y, y_pred)
            weights.append(i)
            f1_scores.append(f1)
            acc_scores.append(metrics.accuracy_score(y, y_pred))
            recall_scores.append(metrics.recall_score(y, y_pred))
            precision_scores.append(metrics.precision_score(y, y_pred))
            print("%8.4f %8.4f" % (i, f1, ))
    return weights, f1_scores, acc_scores, recall_scores, precision_scores

weights, f1_scores, acc_scores, rec_scores, prec_scores = compare_weighted_rfc
s(scaled_X.to_numpy(), y.to_numpy())
```

0.0000	0.0000
0.0250	0.0281
0.0500	0.0719
0.0750	0.1051
0.1000	0.1292
0.1250	0.1453
0.1500	0.1652
0.1750	0.1869
0.2000	0.2442
0.2250	0.2861
0.2500	0.3250
0.2750	0.3644
0.3000	0.3962
0.3250	0.4174
0.3500	0.4323
0.3750	0.4487
0.4000	0.4728
0.4250	0.4925
0.4500	0.5101
0.4750	0.5285
0.5000	0.5434
0.5250	0.5610
0.5500	0.5713
0.5750	0.5810
0.6000	0.5909
0.6250	0.5986
0.6500	0.6094
0.6750	0.6161
0.7000	0.6239
0.7250	0.6277
0.7500	0.6252
0.7750	0.6191
0.8000	0.6067
0.8250	0.5973
0.8500	0.5828
0.8750	0.5571
0.9000	0.5296
0.9250	0.4953
0.9500	0.4563
0.9750	0.4123

In [284]:
```
fig, ax = plt.subplots()
ax.plot(weights, f1_scores, label='f1')
ax.plot(weights, acc_scores, label='acc')
ax.plot(weights, rec_scores, label="recall")
ax.plot(weights, prec_scores, label="precision")
ax.set_xlabel('class weights')
ax.set_ylabel('accuracy/f1 scores')
fig.legend()
```

Out[284]: <matplotlib.legend.Legend at 0x1d85a47c850>



In [288]:
```
for i, j, k, l, m in zip(weights, f1_scores, acc_scores, rec_scores, prec_scor
es):
    print("weight = %4.2f | f1 = %5.4f | accuracy = %5.4f | recall = %5.4f | p
recision = %5.4f" % (i, j, k, l, m))
```

weight = 0.00 | f1 = 0.0000 | accuracy = 0.7963 | recall = 0.0000 | precision = 0.0000
weight = 0.03 | f1 = 0.0281 | accuracy = 0.7992 | recall = 0.0142 | precision = 1.0000
weight = 0.05 | f1 = 0.0719 | accuracy = 0.8039 | recall = 0.0373 | precision = 1.0000
weight = 0.08 | f1 = 0.1051 | accuracy = 0.8075 | recall = 0.0555 | precision = 0.9912
weight = 0.10 | f1 = 0.1292 | accuracy = 0.8100 | recall = 0.0692 | precision = 0.9724
weight = 0.12 | f1 = 0.1453 | accuracy = 0.8117 | recall = 0.0785 | precision = 0.9639
weight = 0.15 | f1 = 0.1652 | accuracy = 0.8140 | recall = 0.0903 | precision = 0.9634
weight = 0.18 | f1 = 0.1869 | accuracy = 0.8164 | recall = 0.1036 | precision = 0.9548
weight = 0.20 | f1 = 0.2442 | accuracy = 0.8230 | recall = 0.1404 | precision = 0.9377
weight = 0.23 | f1 = 0.2861 | accuracy = 0.8283 | recall = 0.1689 | precision = 0.9348
weight = 0.25 | f1 = 0.3250 | accuracy = 0.8330 | recall = 0.1973 | precision = 0.9199
weight = 0.28 | f1 = 0.3644 | accuracy = 0.8378 | recall = 0.2283 | precision = 0.9029
weight = 0.30 | f1 = 0.3962 | accuracy = 0.8421 | recall = 0.2543 | precision = 0.8962
weight = 0.33 | f1 = 0.4174 | accuracy = 0.8442 | recall = 0.2739 | precision = 0.8760
weight = 0.35 | f1 = 0.4323 | accuracy = 0.8461 | recall = 0.2877 | precision = 0.8694
weight = 0.38 | f1 = 0.4487 | accuracy = 0.8479 | recall = 0.3039 | precision = 0.8573
weight = 0.40 | f1 = 0.4728 | accuracy = 0.8517 | recall = 0.3265 | precision = 0.8570
weight = 0.43 | f1 = 0.4925 | accuracy = 0.8543 | recall = 0.3471 | precision = 0.8477
weight = 0.45 | f1 = 0.5101 | accuracy = 0.8563 | recall = 0.3672 | precision = 0.8348
weight = 0.48 | f1 = 0.5285 | accuracy = 0.8583 | recall = 0.3898 | precision = 0.8202
weight = 0.50 | f1 = 0.5434 | accuracy = 0.8595 | recall = 0.4104 | precision = 0.8038
weight = 0.53 | f1 = 0.5610 | accuracy = 0.8610 | recall = 0.4359 | precision = 0.7865
weight = 0.55 | f1 = 0.5713 | accuracy = 0.8606 | recall = 0.4561 | precision = 0.7646
weight = 0.58 | f1 = 0.5810 | accuracy = 0.8608 | recall = 0.4737 | precision = 0.7510
weight = 0.60 | f1 = 0.5909 | accuracy = 0.8611 | recall = 0.4924 | precision = 0.7386
weight = 0.62 | f1 = 0.5986 | accuracy = 0.8605 | recall = 0.5106 | precision = 0.7232
weight = 0.65 | f1 = 0.6094 | accuracy = 0.8600 | recall = 0.5361 | precision = 0.7059
weight = 0.68 | f1 = 0.6161 | accuracy = 0.8563 | recall = 0.5660 | precision = 0.6758
weight = 0.70 | f1 = 0.6239 | accuracy = 0.8527 | recall = 0.5999 | precision

```
= 0.6500
weight = 0.73 | f1 = 0.6277 | accuracy = 0.8470 | recall = 0.6333 | precision
= 0.6223
weight = 0.75 | f1 = 0.6252 | accuracy = 0.8385 | recall = 0.6613 | precision
= 0.5929
weight = 0.78 | f1 = 0.6191 | accuracy = 0.8260 | recall = 0.6942 | precision
= 0.5587
weight = 0.80 | f1 = 0.6067 | accuracy = 0.8089 | recall = 0.7236 | precision
= 0.5223
weight = 0.83 | f1 = 0.5973 | accuracy = 0.7905 | recall = 0.7629 | precision
= 0.4908
weight = 0.85 | f1 = 0.5828 | accuracy = 0.7666 | recall = 0.8002 | precision
= 0.4583
weight = 0.88 | f1 = 0.5571 | accuracy = 0.7306 | recall = 0.8316 | precision
= 0.4188
weight = 0.90 | f1 = 0.5296 | accuracy = 0.6847 | recall = 0.8714 | precision
= 0.3804
weight = 0.93 | f1 = 0.4953 | accuracy = 0.6222 | recall = 0.9102 | precision
= 0.3402
weight = 0.95 | f1 = 0.4563 | accuracy = 0.5407 | recall = 0.9460 | precision
= 0.3006
weight = 0.98 | f1 = 0.4123 | accuracy = 0.4355 | recall = 0.9720 | precision
= 0.2616
```

Grid Search different weights (no SMOTE)

In [243]:
```
# Try a grid wearch with different weights, and with some other parameters
def grid_search_random_forest_weighted(X, y):
    rfc = RandomForestClassifier(random_state=1)
    p = imblearn.pipeline.Pipeline([('rfc', rfc)])

    class_weights = [{0: i, 1: 1-i} for i in np.arange(0, 1, 0.05)]
    class_weights.append({1: 4, 0: 1.33})
    class_weights.append({1: 4, 0: 1})
    class_weights.append({1: 5, 0: 1})
    param_grid = {
        'rfc__n_estimators': [10, 100, 1000],
        'rfc__criterion': ['gini', 'entropy'],
        'rfc__min_samples_split': [i for i in range(20, 200, 40)],
        'rfc__class_weight': class_weights
    }
    #scoring = {'F1': make_scorer(metrics.f1_score), 'ACC': make_scorer(metric
s.accuracy_score)}
    scorer = make_scorer(metrics.f1_score)
    cv1 = StratifiedKFold(n_splits=6, random_state=10, shuffle=True)
    search = GridSearchCV(p, param_grid, n_jobs=7, cv=cv1, scoring=scorer)
    search.fit(X, y)
    print("RANDOM FOREST")
    print(f"Best Params = {search.best_params_}")
    print(f"Best score = {search.best_score_}")
    print('-' * 120)
    return search.best_estimator_

random.seed(22)
rfcw  = grid_search_random_forest_weighted(scaled_X.to_numpy(), y.to_numpy())
```

```
RANDOM FOREST
Best Params = {'rfc__class_weight': {0: 0.2, 1: 0.8}, 'rfc__criterion': 'gin
i', 'rfc__min_samples_split': 20, 'rfc__n_estimators': 100}
Best score = 0.6323982569024459
------------------------------------------------------------------------------
----------------------------------------
```

GridSearch BalancedRandomForestClassifier from Imblearn

In [276]:
```
def grid_search_balanced_random_forest_weighted(X, y):
    rfc = BalancedRandomForestClassifier(random_state=1)
    p = imblearn.pipeline.Pipeline([('rfc', rfc)])

    param_grid = {
        'rfc__n_estimators': [10, 100, 1000],
        'rfc__criterion': ['gini', 'entropy'],
        'rfc__min_samples_split': [i for i in range(20, 200, 40)],
        'rfc__class_weight': [{0: i, 1: 1-i} for i in np.arange(0, 1, 0.1)],
        'rfc__bootstrap': [True, False],
        'rfc__oob_score': [True, False],
    }
    #scoring = {'F1': make_scorer(metrics.f1_score), 'ACC': make_scorer(metric
s.accuracy_score)}
    scorer = make_scorer(metrics.f1_score)
    cv1 = StratifiedKFold(n_splits=6, random_state=10, shuffle=True)
    search = GridSearchCV(p, param_grid, n_jobs=7, cv=cv1, scoring=scorer)
    search.fit(X, y)
    print("RANDOM FOREST")
    print(f"Best Params = {search.best_params_}")
    print(f"Best score = {search.best_score_}")
    print('-' * 120)
    return search.best_estimator_


random.seed(22)
bwrfc  = grid_search_balanced_random_forest_weighted(scaled_X.to_numpy(), y.to
_numpy())
```

```
RANDOM FOREST
Best Params = {'rfc__bootstrap': True, 'rfc__class_weight': {0: 0.70000000000
00001, 1: 0.29999999999999993}, 'rfc__criterion': 'entropy', 'rfc__min_sample
s_split': 20, 'rfc__n_estimators': 100, 'rfc__oob_score': True}
Best score = 0.6329106408569257
------------------------------------------------------------------------------
----------------------------------------
```

Random Forest with Adjusted Cutoff

In [290]:

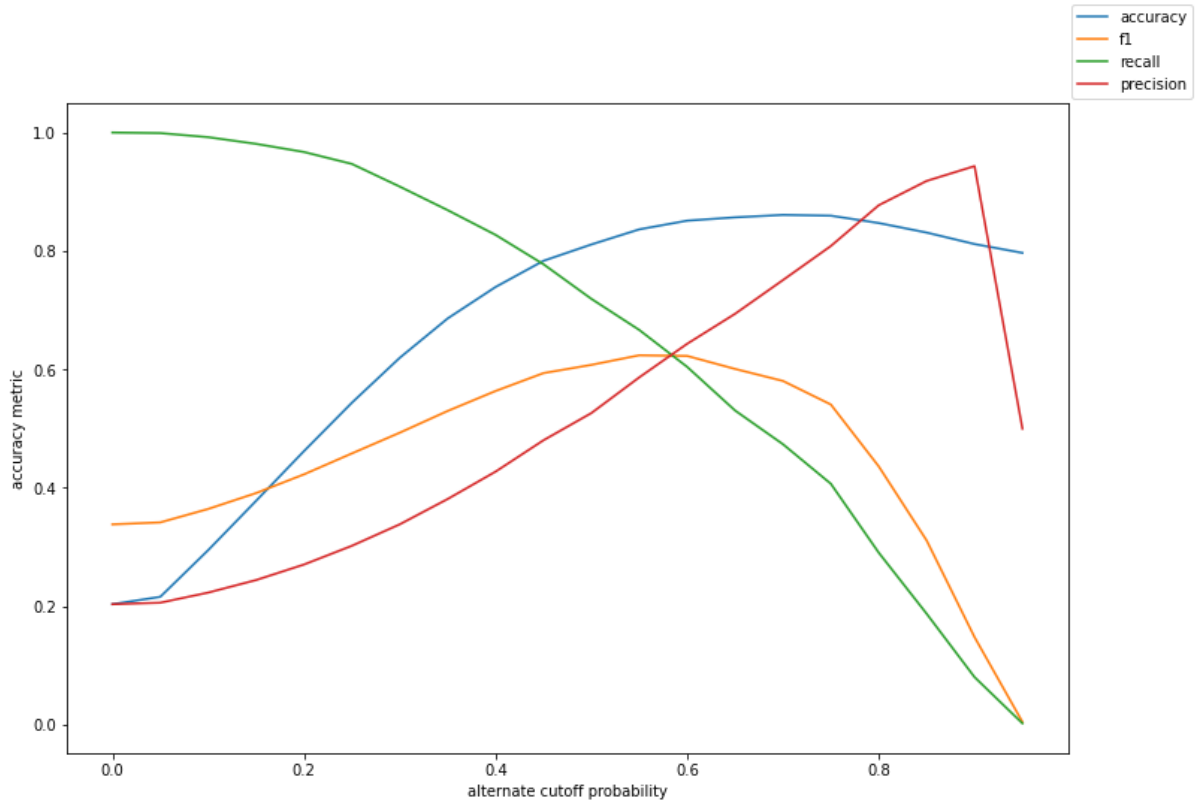
```
def random_forest_alternate_cutoff(X, y, cutoff, use_smote=False):
    """
    Using Alternate cut-off is different from assigning class weights. Assigni
ng class weights
    actually changes the mode, but this doens't change the model.
    The only thing that changes is that we predict a probability instead of a
 definite value,
    and then set a cut-off on that probability.
    """
    f1_scores = []
    accuracy_scores = []
    recall_scores = []
    precision_scores = []
    kf = model_selection.StratifiedKFold(n_splits=6, shuffle=True, random_stat
e=2)
    for train_index, test_index in kf.split(X, y):
        rfc = RandomForestClassifier(random_state=0, class_weight='balanced_su
bsample', criterion='entropy', min_samples_split=100, n_estimators=1000, n_job
s=7)
        X_train, y_train = X[train_index, :], y[train_index]
        X_test, y_test = X[test_index, :], y[test_index]
        if use_smote:
            sm = SMOTE()
            X_train, y_train = sm.fit_resample(X_train, y_train)
        rfc.fit(X_train, y_train)
        y_pred_proba = rfc.predict_proba(X_test)
        y_pred = (y_pred_proba[:, 1] >= cutoff).astype('int')
        f1s = metrics.f1_score(y_test, y_pred)
        accs = metrics.accuracy_score(y_test, y_pred)
        f1_scores.append(f1s)
        recall_scores.append(metrics.recall_score(y_test, y_pred))
        precision_scores.append(metrics.precision_score(y_test, y_pred))
        accuracy_scores.append(accs)
    return mean(accuracy_scores), mean(f1_scores), mean(recall_scores), mean(p
recision_scores)
```

In [295]:

```
acc_scores = []
f1_scores = []
rec_scores = []
prec_scores = []
cutoffs = []
for i in np.arange(0, 1, 0.05):
    accs, f1s, rec, prec = random_forest_alternate_cutoff(scaled_X.to_numpy(),
y.to_numpy(), i)
    acc_scores.append(accs)
    f1_scores.append(f1s)
    rec_scores.append(rec)
    prec_scores.append(prec)
    cutoffs.append(i)

for i, j, k, l, m in zip(cutoffs, acc_scores, f1_scores, prec_scores, rec_scor
es):
    print("%10.5f %20.10f %20.10f %20.10f %20.10f" % (i, j, k, l, m))
fig, ax = plt.subplots()
ax.plot(cutoffs, acc_scores, label="accuracy")
ax.plot(cutoffs, f1_scores, label="f1")
ax.plot(cutoffs, rec_scores, label="recall")
ax.plot(cutoffs, prec_scores, label="precision")
ax.set_xlabel('alternate cutoff probability')
ax.set_ylabel('accuracy metric')
fig.legend()
```

```
   0.00000         0.2036999563         0.3384562867         0.2036999563
1.0000000000
   0.05000         0.2160994572         0.3417630524         0.2061432254
0.9990196078
   0.10000         0.2948994235         0.3644404072         0.2232327069
0.9921481867
   0.15000         0.3781006104         0.3912567866         0.2443955141
0.9808534328
   0.20000         0.4621009172         0.4228523914         0.2706232485
0.9671149286
   0.25000         0.5438002436         0.4581976529         0.3022398960
0.9469879692
   0.30000         0.6196992894         0.4932123362         0.3385166326
0.9086977269
   0.35000         0.6860987346         0.5298961211         0.3812581525
0.8689325583
   0.40000         0.7391995791         0.5636000164         0.4274662791
0.8271863613
   0.45000         0.7834007028         0.5938122218         0.4804228219
0.7776100411
   0.50000         0.8109996848         0.6077237432         0.5263355725
0.7191957314
   0.55000         0.8364991467         0.6239149352         0.5866065161
0.6666724507
   0.60000         0.8511993880         0.6228026748         0.6433087395
0.6043250043
   0.65000         0.8567996485         0.6008714103         0.6939108514
0.5306900341
   0.70000         0.8608993087         0.5805817165         0.7508777502
0.4737463127
   0.75000         0.8595988485         0.5407757560         0.8083735609
0.4069625195
   0.80000         0.8471991076         0.4362901545         0.8772331919
0.2906327723
   0.85000         0.8310996864         0.3112439357         0.9182463509
0.1875368732
   0.90000         0.8116997249         0.1480510594         0.9432839380
0.0805092834
   0.95000         0.7967999437         0.0048790276         0.5000000000
0.0024524264
```

Out[295]: <matplotlib.legend.Legend at 0x1d8c551d4c0>



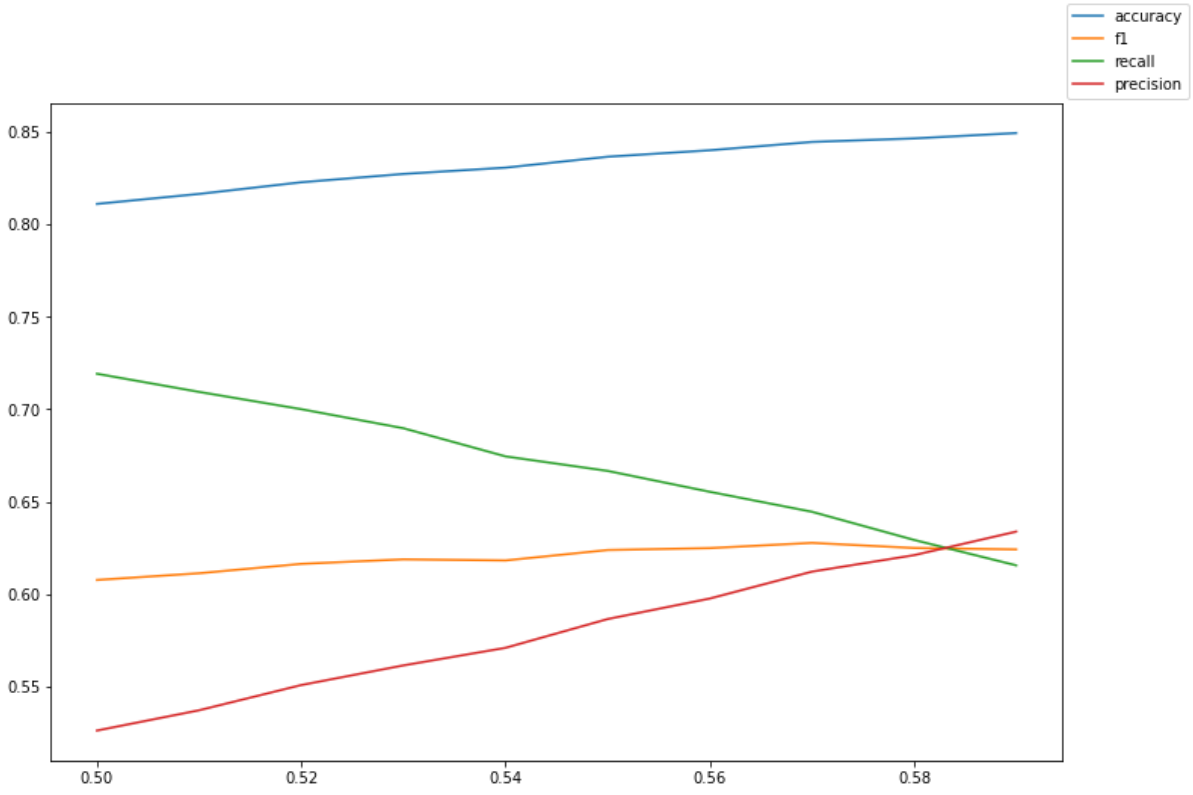
Do a more fine grained one with a smalle rinterval

In [297]:

```
acc_scores = []
f1_scores = []
recall_scores = []
precision_scores = []
cutoffs = []
for i in np.arange(0.50, 0.60, 0.01):
    accs, f1s, rec, prec = random_forest_alternate_cutoff(scaled_X.to_numpy(),
y.to_numpy(), i)
    acc_scores.append(accs)
    f1_scores.append(f1s)
    recall_scores.append(rec)
    precision_scores.append(prec)
    cutoffs.append(i)

for i, j, k, l, m in zip(cutoffs, acc_scores, f1_scores, prec_scores, rec_scor
es):
    print("%10.5f %20.10f %20.10f %20.10f %20.10f" % (i, j, k, l, m))
fig, ax = plt.subplots()
ax.plot(cutoffs, acc_scores, label="accuracy")
ax.plot(cutoffs, f1_scores, label="f1")
ax.plot(cutoffs, recall_scores, label="recall")
ax.plot(cutoffs, precision_scores, label="precision")
fig.legend()
```

```
   0.50000         0.8109996848         0.6077237432         0.2036999563
1.0000000000
   0.51000         0.8163995052         0.6113544138         0.2061432254
0.9990196078
   0.52000         0.8226993257         0.6164340182         0.2232327069
0.9921481867
   0.53000         0.8271995661         0.6188813431         0.2443955141
0.9808534328
   0.54000         0.8305992463         0.6182805793         0.2706232485
0.9671149286
   0.55000         0.8364991467         0.6239149352         0.3022398960
0.9469879692
   0.56000         0.8399995871         0.6249520083         0.3385166326
0.9086977269
   0.57000         0.8444997675         0.6277714194         0.3812581525
0.8689325583
   0.58000         0.8463995676         0.6250138357         0.4274662791
0.8271863613
   0.59000         0.8492999479         0.6242401350         0.4804228219
0.7776100411
```

Out[297]: `<matplotlib.legend.Legend at 0x1d92481f160>`

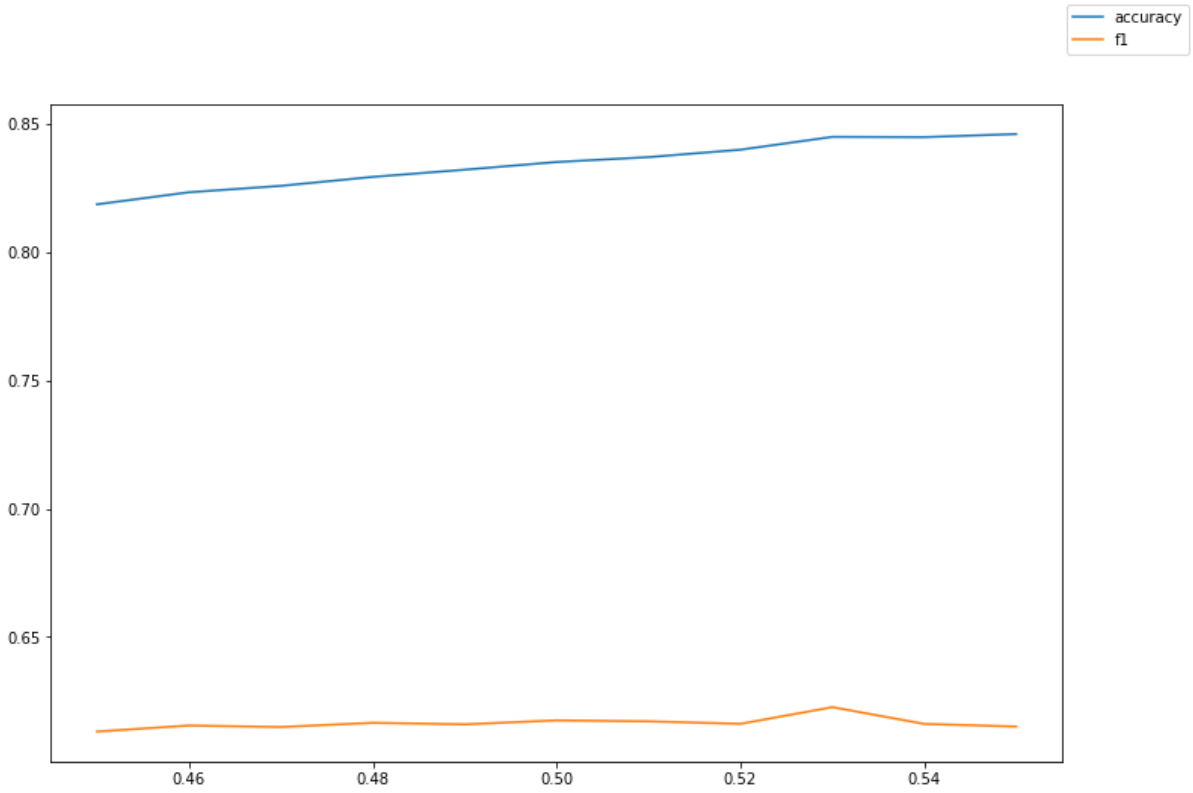


Do a more fine-grained for a smaller interval, use SMOTE this time

In [343]:
```
acc_scores = []
f1_scores = []
cutoffs = []
for i in np.arange(0.45, 0.55, 0.01):
    accs, f1s = random_forest_alternate_cutoff(scaled_X.to_numpy(), y.to_numpy
(), i, use_smote=True)
    acc_scores.append(accs)
    f1_scores.append(f1s)
    cutoffs.append(i)

for i, j, k in zip(cutoffs, acc_scores, f1_scores):
    print("%3.2f %10.5f %10.5f" % (i, j, k))
fig, ax = plt.subplots()
ax.plot(cutoffs, acc_scores, label="accuracy")
ax.plot(cutoffs, f1_scores, label="f1")
fig.legend()
```

```
0.45    0.81860    0.61310
0.46    0.82330    0.61544
0.47    0.82580    0.61486
0.48    0.82930    0.61650
0.49    0.83210    0.61591
0.50    0.83510    0.61746
0.51    0.83700    0.61707
0.52    0.83990    0.61611
0.53    0.84490    0.62261
0.54    0.84480    0.61607
0.55    0.84600    0.61504
```

Out[343]: <matplotlib.legend.Legend at 0x1ec41eb7ee0>



Classification with Other Samplers (over-samplers and combine-samplers)

In [84]:
```
def all_models_compare_samplers2(X, y, sampler, description, model_gen_fn, mod
el_description):
    if None != sampler:
        p = imblearn.pipeline.Pipeline([('sampler', sampler), ('model', model_
gen_fn())])
    else:
        p = model_gen_fn()
    y_pred = cross_val_predict(p, X, y, \
                               cv=model_selection.StratifiedKFold(n_splits=6,
shuffle=True, random_state=2),\
                               n_jobs=7)
    print(f"Description: {description} Model: {model_description}\n{metrics.cl
assification_report(y, y_pred)}")
    print('-' * 80)
    return metrics.accuracy_score(y, y_pred), metrics.f1_score(y, y_pred)
```

In [117]:

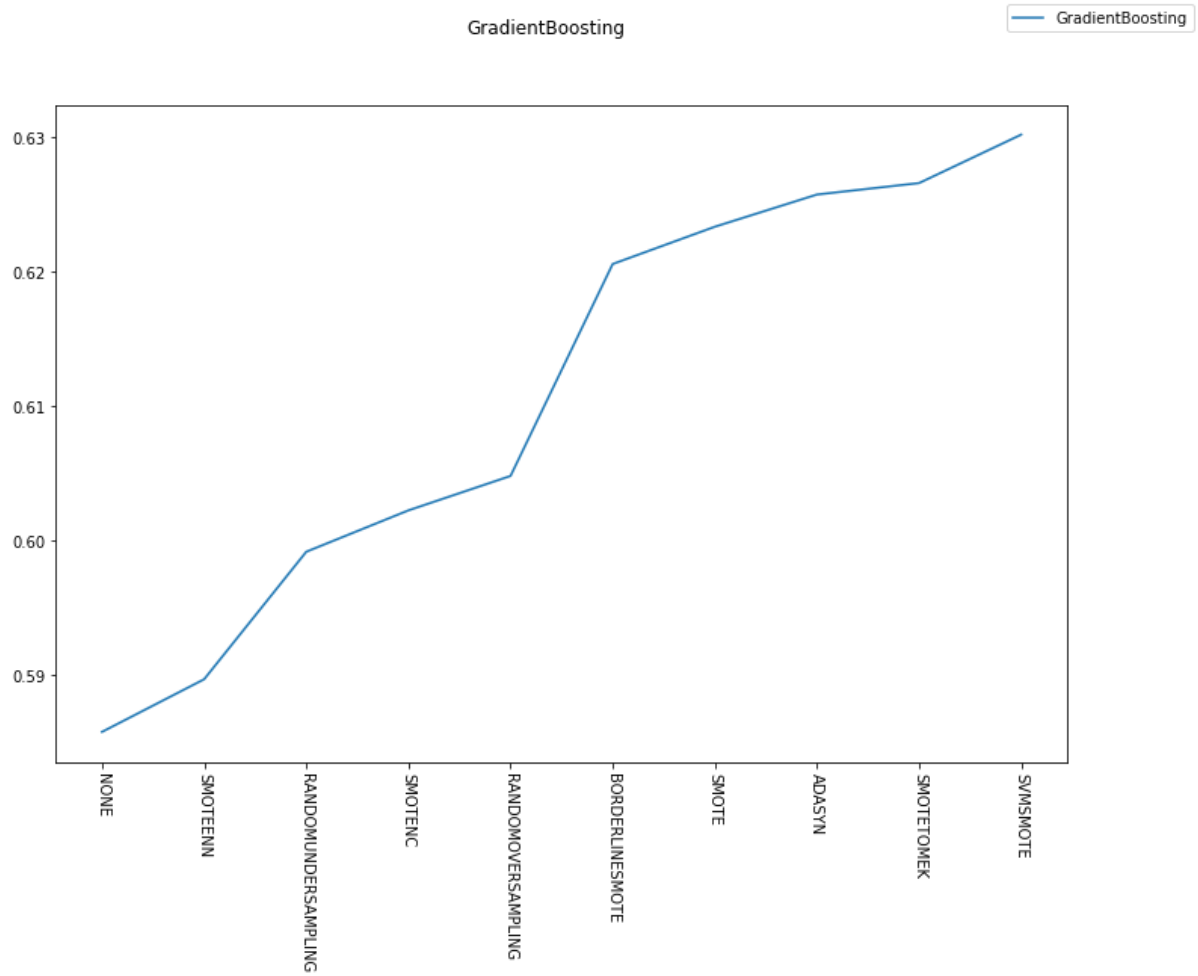
```
samplers = [
    ('NONE', None),
    ('RandomOversampling', RandomOverSampler(sampling_strategy='minority')),
    ('RandomUnderSampling', RandomUnderSampler(sampling_strategy='majority')),
    ('smote', SMOTE(random_state=0)),
    ('adasyn', ADASYN(random_state=0)),
    ('smotenc', SMOTENC([False, False, True, True, True, True, False, True, Tr
ue, True, True, True], random_state=0)),
    ('borderlinesmote', BorderlineSMOTE(random_state=0)),
    ('svmsmote', SVMSMOTE(random_state=0)),
    ('smoteenn', SMOTEENN(random_state=0)),
    ('smotetomek', SMOTETomek(random_state=0))
]

models = [
    ('GradientBoosting', lambda: GradientBoostingClassifier(random_state=0, ma
x_depth=3, min_samples_split=42, n_estimators=100)),
    ('RandomForest', lambda: RandomForestClassifier(class_weight='balanced_sub
sample', criterion='entropy', max_depth=None, min_samples_split=100, n_estimat
ors=1000, random_state=0)),
    ('SVC', lambda: SVC(class_weight=None, gamma='auto', kernel='rbf', random_
state=0)),
]
descriptions = []
sorted_indices = None
indices_calculated = False
comparison_string = ""

plot_data = []
for a, b in models:
    fig1, ax1 = plt.subplots()
    accuracy_scores = []
    f1_scores = []
    descriptions = []
    print('-' * 80)
    print(f"ITERATING FOR {a.upper()}\n")
    for i, j in samplers:
        accs, f1s = all_models_compare_samplers2(scaled_X.to_numpy(), y.to_num
py(), j, i, b, a)
        accuracy_scores.append(accs)
        f1_scores.append(f1s)
        descriptions.append(i)
    if False == indices_calculated:
        sorted_indices = np.argsort(f1_scores)
        indices_calculated = True
    plot_data.append(([f1_scores[i] for i in sorted_indices], f"{a}-f1",) )
    plot_data.append(([accuracy_scores[i] for i in sorted_indices], f'{a}-accu
racy',))
    ax1.plot([f1_scores[i] for i in np.argsort(f1_scores)], label=a)
    plt.xticks(np.arange(0, len(f1_scores), 1.0))
    ax1.set_xticklabels([descriptions[i].upper() for i in np.argsort(f1_scores
)], rotation=270)
    fig1.legend()
    fig1.suptitle(a)
    plt.show()
    sorted_order = np.argsort(f1_scores)
```

```
    comparison_string += "EVALUATING MODEL %s\n" % a
    comparison_string += '------\n'
    for i in sorted_order:
        comparison_string += "%20s %10.5f %10.5f" % (descriptions[i], accuracy
_scores[i], f1_scores[i])
        comparison_string += "\n"
    comparison_string += "----------------------------------------\n"
    print(comparison_string)
fig, ax = plt.subplots()
plt.xticks(np.arange(0, len(f1_scores), 1.0))
for array, lab in plot_data:
    ax.plot(array, label=lab)
ax.set_xticklabels([descriptions[i].upper() for i in sorted_indices], rotation
=270)
ax.legend()
fig.legend()

print(plot_data)
```

```
------------------------------------------------------------------------
---
ITERATING FOR GRADIENTBOOSTING

Description: NONE Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.88      0.97      0.92      7963
           1       0.78      0.47      0.59      2037

    accuracy                           0.86     10000
   macro avg       0.83      0.72      0.75     10000
weighted avg       0.86      0.86      0.85     10000

------------------------------------------------------------------------
---
Description: RandomOversampling Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.93      0.82      0.87      7963
           1       0.51      0.74      0.60      2037

    accuracy                           0.80     10000
   macro avg       0.72      0.78      0.74     10000
weighted avg       0.84      0.80      0.81     10000

------------------------------------------------------------------------
---
Description: RandomUnderSampling Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.93      0.80      0.86      7963
           1       0.49      0.76      0.60      2037

    accuracy                           0.79     10000
   macro avg       0.71      0.78      0.73     10000
weighted avg       0.84      0.79      0.81     10000

------------------------------------------------------------------------
---
Description: smote Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.91      0.89      0.90      7963
           1       0.60      0.65      0.62      2037

    accuracy                           0.84     10000
   macro avg       0.76      0.77      0.76     10000
weighted avg       0.85      0.84      0.84     10000

------------------------------------------------------------------------
---
Description: adasyn Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.91      0.89      0.90      7963
           1       0.60      0.65      0.63      2037
```

```
    accuracy                           0.84     10000
   macro avg       0.76      0.77      0.76     10000
weighted avg       0.85      0.84      0.84     10000

--------------------------------------------------------------------------------
Description: smotenc Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.92      0.83      0.87      7963
           1       0.52      0.72      0.60      2037

    accuracy                           0.81     10000
   macro avg       0.72      0.77      0.74     10000
weighted avg       0.84      0.81      0.82     10000

--------------------------------------------------------------------------------
Description: borderlinesmote Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.91      0.87      0.89      7963
           1       0.57      0.68      0.62      2037

    accuracy                           0.83     10000
   macro avg       0.74      0.77      0.76     10000
weighted avg       0.84      0.83      0.84     10000

--------------------------------------------------------------------------------
Description: svmsmote Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.91      0.88      0.90      7963
           1       0.59      0.67      0.63      2037

    accuracy                           0.84     10000
   macro avg       0.75      0.78      0.76     10000
weighted avg       0.85      0.84      0.84     10000

--------------------------------------------------------------------------------
Description: smoteenn Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.93      0.79      0.85      7963
           1       0.48      0.77      0.59      2037

    accuracy                           0.78     10000
   macro avg       0.70      0.78      0.72     10000
weighted avg       0.84      0.78      0.80     10000

--------------------------------------------------------------------------------
Description: smotetomek Model: GradientBoosting
              precision    recall  f1-score   support
```

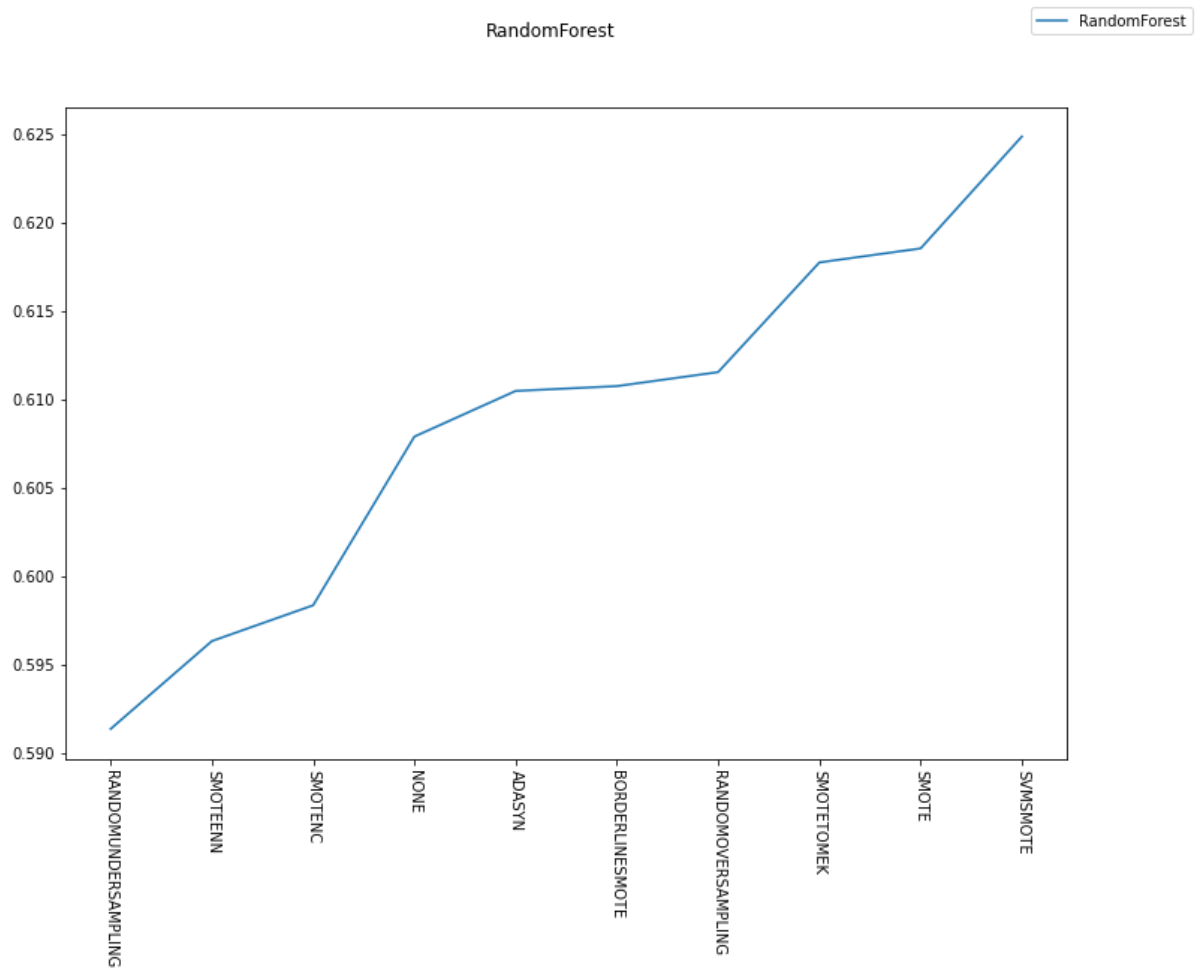
```
              0       0.91      0.89      0.90      7963
              1       0.60      0.65      0.63      2037

       accuracy                           0.84     10000
      macro avg       0.76      0.77      0.76     10000
   weighted avg       0.85      0.84      0.84     10000

------------------------------------------------------------------------------
---
```



```
EVALUATING MODEL GradientBoosting
------
                NONE    0.86480    0.58578
            smoteenn    0.78250    0.58970
 RandomUnderSampling    0.79290    0.59919
             smotenc    0.80690    0.60227
  RandomOversampling    0.80190    0.60483
     borderlinesmote    0.83090    0.62060
              smote    0.84110    0.62337
             adasyn    0.84140    0.62577
          smotetomek    0.84150    0.62662
            svmsmote    0.83900    0.63023
----------------------------------------

--------------------------------------------------------------------------------
---
ITERATING FOR RANDOMFOREST

Description: NONE Model: RandomForest
              precision    recall  f1-score   support

           0       0.92      0.83      0.88      7963
           1       0.53      0.72      0.61      2037

    accuracy                           0.81     10000
   macro avg       0.72      0.78      0.74     10000
weighted avg       0.84      0.81      0.82     10000

--------------------------------------------------------------------------------
---
Description: RandomOversampling Model: RandomForest
              precision    recall  f1-score   support

           0       0.92      0.84      0.88      7963
           1       0.54      0.71      0.61      2037

    accuracy                           0.82     10000
   macro avg       0.73      0.78      0.75     10000
weighted avg       0.84      0.82      0.82     10000

--------------------------------------------------------------------------------
---
Description: RandomUnderSampling Model: RandomForest
              precision    recall  f1-score   support

           0       0.93      0.80      0.86      7963
           1       0.49      0.75      0.59      2037

    accuracy                           0.79     10000
   macro avg       0.71      0.77      0.72     10000
weighted avg       0.84      0.79      0.80     10000

--------------------------------------------------------------------------------
---
Description: smote Model: RandomForest
              precision    recall  f1-score   support
```

```
              0       0.91      0.88      0.90      7963
              1       0.59      0.65      0.62      2037

       accuracy                           0.84     10000
      macro avg       0.75      0.77      0.76     10000
   weighted avg       0.84      0.84      0.84     10000

------------------------------------------------------------------------
---
Description: adasyn Model: RandomForest
                 precision    recall  f1-score   support

              0       0.91      0.86      0.88      7963
              1       0.55      0.68      0.61      2037

       accuracy                           0.82     10000
      macro avg       0.73      0.77      0.75     10000
   weighted avg       0.84      0.82      0.83     10000

------------------------------------------------------------------------
---
Description: smotenc Model: RandomForest
                 precision    recall  f1-score   support

              0       0.91      0.84      0.88      7963
              1       0.53      0.69      0.60      2037

       accuracy                           0.81     10000
      macro avg       0.72      0.77      0.74     10000
   weighted avg       0.84      0.81      0.82     10000

------------------------------------------------------------------------
---
Description: borderlinesmote Model: RandomForest
                 precision    recall  f1-score   support

              0       0.92      0.85      0.88      7963
              1       0.54      0.70      0.61      2037

       accuracy                           0.82     10000
      macro avg       0.73      0.77      0.75     10000
   weighted avg       0.84      0.82      0.83     10000

------------------------------------------------------------------------
---
Description: svmsmote Model: RandomForest
                 precision    recall  f1-score   support

              0       0.91      0.87      0.89      7963
              1       0.58      0.68      0.62      2037

       accuracy                           0.83     10000
      macro avg       0.75      0.78      0.76     10000
   weighted avg       0.85      0.83      0.84     10000

------------------------------------------------------------------------
---
```

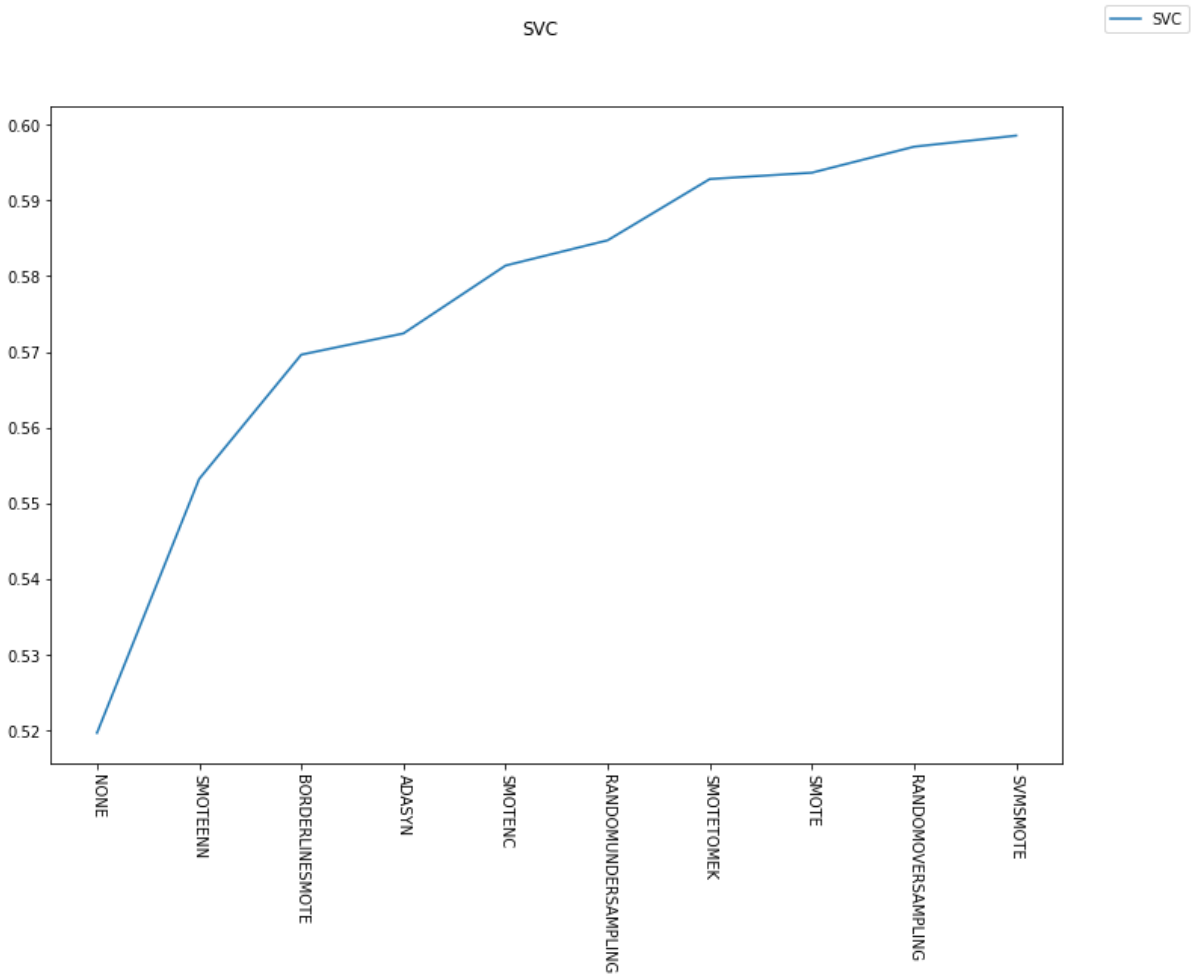
```
Description: smoteenn Model: RandomForest
              precision    recall  f1-score   support

           0       0.92      0.81      0.86      7963
           1       0.50      0.74      0.60      2037

    accuracy                           0.80     10000
   macro avg       0.71      0.77      0.73     10000
weighted avg       0.84      0.80      0.81     10000

--------------------------------------------------------------------------------
---
Description: smotetomek Model: RandomForest
              precision    recall  f1-score   support

           0       0.91      0.88      0.89      7963
           1       0.58      0.65      0.62      2037

    accuracy                           0.83     10000
   macro avg       0.75      0.77      0.76     10000
weighted avg       0.84      0.83      0.84     10000

--------------------------------------------------------------------------------
---
```



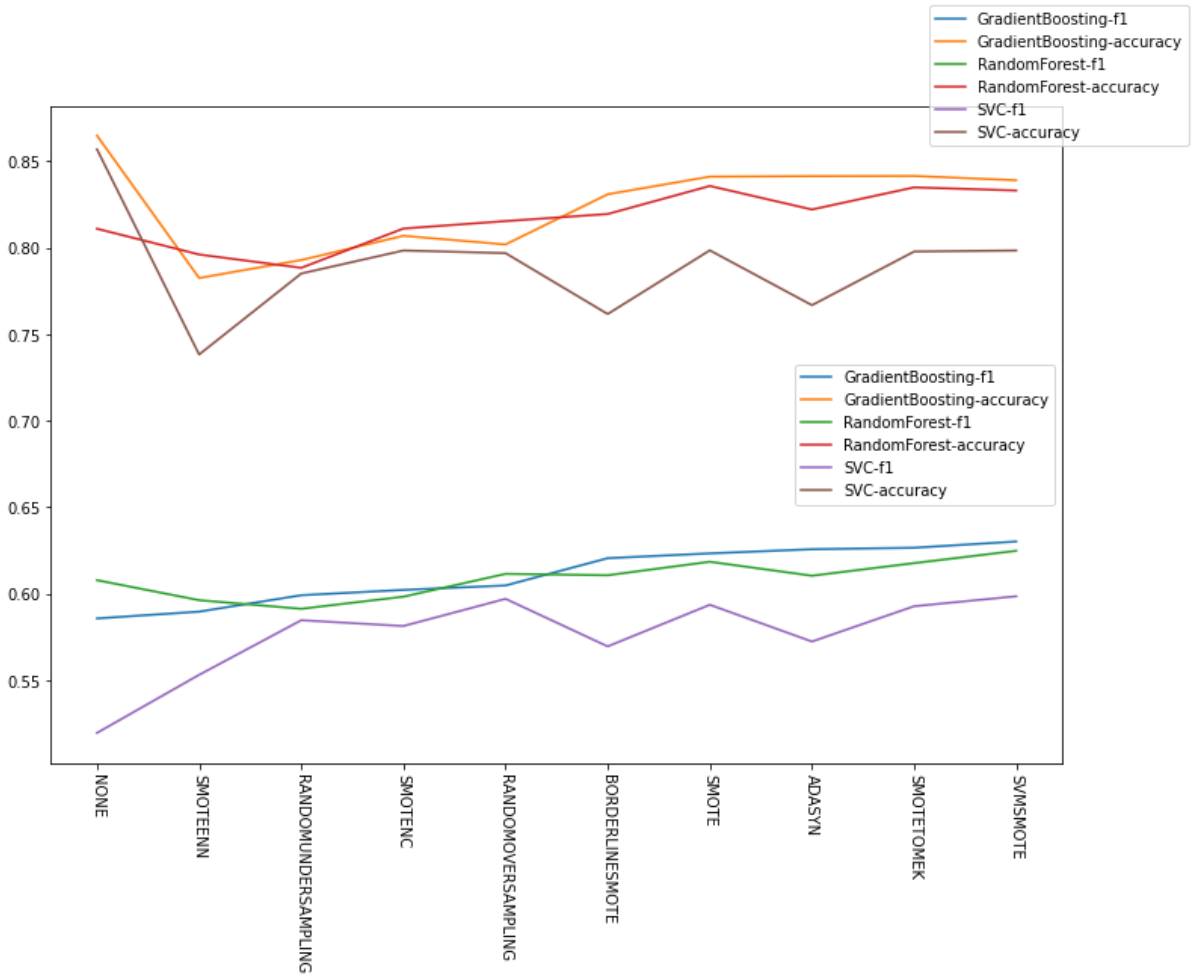
```
EVALUATING MODEL GradientBoosting
------
                NONE    0.86480    0.58578
            smoteenn    0.78250    0.58970
 RandomUnderSampling    0.79290    0.59919
             smotenc    0.80690    0.60227
  RandomOversampling    0.80190    0.60483
     borderlinesmote    0.83090    0.62060
               smote    0.84110    0.62337
              adasyn    0.84140    0.62577
          smotetomek    0.84150    0.62662
            svmsmote    0.83900    0.63023
---------------------------------------
EVALUATING MODEL RandomForest
------
 RandomUnderSampling    0.78840    0.59135
            smoteenn    0.79610    0.59632
             smotenc    0.81110    0.59834
                NONE    0.81100    0.60788
              adasyn    0.82210    0.61047
     borderlinesmote    0.81950    0.61074
  RandomOversampling    0.81540    0.61153
          smotetomek    0.83490    0.61774
               smote    0.83570    0.61853
            svmsmote    0.83310    0.62486
---------------------------------------

--------------------------------------------------------------------------------
---
ITERATING FOR SVC

Description: NONE Model: SVC
              precision    recall  f1-score   support

           0       0.86      0.98      0.92      7963
           1       0.82      0.38      0.52      2037

    accuracy                           0.86     10000
   macro avg       0.84      0.68      0.72     10000
weighted avg       0.85      0.86      0.84     10000

--------------------------------------------------------------------------------
---
Description: RandomOversampling Model: SVC
              precision    recall  f1-score   support

           0       0.92      0.81      0.86      7963
           1       0.50      0.74      0.60      2037

    accuracy                           0.80     10000
   macro avg       0.71      0.78      0.73     10000
weighted avg       0.84      0.80      0.81     10000

--------------------------------------------------------------------------------
---
Description: RandomUnderSampling Model: SVC
              precision    recall  f1-score   support
```

```
           0       0.92      0.80      0.86      7963
           1       0.48      0.74      0.58      2037

    accuracy                           0.79     10000
   macro avg       0.70      0.77      0.72     10000
weighted avg       0.83      0.79      0.80     10000

--------------------------------------------------------------------------------
---
Description: smote Model: SVC
              precision    recall  f1-score   support

           0       0.92      0.82      0.87      7963
           1       0.50      0.72      0.59      2037

    accuracy                           0.80     10000
   macro avg       0.71      0.77      0.73     10000
weighted avg       0.84      0.80      0.81     10000

--------------------------------------------------------------------------------
---
Description: adasyn Model: SVC
              precision    recall  f1-score   support

           0       0.93      0.77      0.84      7963
           1       0.46      0.77      0.57      2037

    accuracy                           0.77     10000
   macro avg       0.69      0.77      0.71     10000
weighted avg       0.83      0.77      0.79     10000

--------------------------------------------------------------------------------
---
Description: smotenc Model: SVC
              precision    recall  f1-score   support

           0       0.91      0.83      0.87      7963
           1       0.50      0.69      0.58      2037

    accuracy                           0.80     10000
   macro avg       0.71      0.76      0.72     10000
weighted avg       0.83      0.80      0.81     10000

--------------------------------------------------------------------------------
---
Description: borderlinesmote Model: SVC
              precision    recall  f1-score   support

           0       0.93      0.76      0.84      7963
           1       0.45      0.77      0.57      2037

    accuracy                           0.76     10000
   macro avg       0.69      0.77      0.70     10000
weighted avg       0.83      0.76      0.78     10000

--------------------------------------------------------------------------------
```

```
---
Description: svmsmote Model: SVC
              precision    recall  f1-score   support

           0       0.92      0.81      0.87      7963
           1       0.50      0.74      0.60      2037

    accuracy                           0.80     10000
   macro avg       0.71      0.78      0.73     10000
weighted avg       0.84      0.80      0.81     10000

--------------------------------------------------------------------------------
---
Description: smoteenn Model: SVC
              precision    recall  f1-score   support

           0       0.93      0.72      0.81      7963
           1       0.42      0.80      0.55      2037

    accuracy                           0.74     10000
   macro avg       0.68      0.76      0.68     10000
weighted avg       0.83      0.74      0.76     10000

--------------------------------------------------------------------------------
---
Description: smotetomek Model: SVC
              precision    recall  f1-score   support

           0       0.92      0.82      0.87      7963
           1       0.50      0.72      0.59      2037

    accuracy                           0.80     10000
   macro avg       0.71      0.77      0.73     10000
weighted avg       0.84      0.80      0.81     10000

--------------------------------------------------------------------------------
---
```


SVC
SVC
0.60
0.59
0.58
0.57
0.56
0.55
0.54
0.53
0.52
NONE
SMOTEENN
BORDERLINESMOTE
ADASYN
SMOTENC
RANDOMUNDERSAMPLING
SMOTETOMEK
SMOTE
RANDOMOVERSAMPLING
SVMSMOTE

```
EVALUATING MODEL GradientBoosting
------
                NONE    0.86480    0.58578
            smoteenn    0.78250    0.58970
 RandomUnderSampling    0.79290    0.59919
             smotenc    0.80690    0.60227
  RandomOversampling    0.80190    0.60483
     borderlinesmote    0.83090    0.62060
               smote    0.84110    0.62337
              adasyn    0.84140    0.62577
          smotetomek    0.84150    0.62662
            svmsmote    0.83900    0.63023
---------------------------------------
EVALUATING MODEL RandomForest
------
 RandomUnderSampling    0.78840    0.59135
            smoteenn    0.79610    0.59632
             smotenc    0.81110    0.59834
                NONE    0.81100    0.60788
              adasyn    0.82210    0.61047
     borderlinesmote    0.81950    0.61074
  RandomOversampling    0.81540    0.61153
          smotetomek    0.83490    0.61774
               smote    0.83570    0.61853
            svmsmote    0.83310    0.62486
---------------------------------------
EVALUATING MODEL SVC
------
                NONE    0.85690    0.51964
            smoteenn    0.73830    0.55318
     borderlinesmote    0.76170    0.56962
              adasyn    0.76680    0.57242
             smotenc    0.79840    0.58140
 RandomUnderSampling    0.78510    0.58473
          smotetomek    0.79780    0.59283
               smote    0.79850    0.59367
  RandomOversampling    0.79690    0.59710
            svmsmote    0.79840    0.59857
---------------------------------------

[([0.5857843137254902, 0.5897000565930957, 0.5991871492161795, 0.602265705458
2904, 0.6048274486335529, 0.6205968140004489, 0.6233704669352926, 0.625766871
1656442, 0.6266195524146054, 0.630225080385852], 'GradientBoosting-f1'), ([0.
8648, 0.7825, 0.7929, 0.8069, 0.8019, 0.8309, 0.8411, 0.8414, 0.8415, 0.839],
'GradientBoosting-accuracy'), ([0.6078838174273858, 0.5963175608790339, 0.591
3480108149866, 0.5983414841590474, 0.6115319865319865, 0.6107397023937892, 0.
6185279777107034, 0.6104663893146486, 0.6177355869414217, 0.624859518993032],
'RandomForest-f1'), ([0.811, 0.7961, 0.7884, 0.8111, 0.8154, 0.8195, 0.8357,
0.8221, 0.8349, 0.8331], 'RandomForest-accuracy'), ([0.5196374622356495, 0.55
31842240054635, 0.5847342995169083, 0.5813953488372092, 0.5971037492560999,
0.5696225392811992, 0.593668078241581, 0.5724239090575723, 0.592831252517116
4, 0.5985663082437276], 'SVC-f1'), ([0.8569, 0.7383, 0.7851, 0.7984, 0.7969,
0.7617, 0.7985, 0.7668, 0.7978, 0.7984], 'SVC-accuracy')]
```



Classifiers Samplers and Feature-Selectors compared

In [100]:
```
def model_compare_3(X, y,\
                    sampler, description,\
                    model_gen_fn, model_description,\
                    feature_selector_fn, feature_selector_description):
    p = []
    if None != sampler:
        p.append(('sampler', sampler,))
    if None != feature_selector_fn:
        p.append(('fs', feature_selector_fn(),))
    p.append(('model', model_gen_fn(),))
    p = imblearn.pipeline.Pipeline(p)
    y_pred = cross_val_predict(p, X, y, \
                               cv=model_selection.StratifiedKFold(n_splits=6,
shuffle=True, random_state=2),\
                               n_jobs=7)
    print(f"Description: {description} Model: {model_description}\n{metrics.cl
assification_report(y, y_pred)}")
    print('-' * 80)
    return metrics.accuracy_score(y, y_pred), metrics.f1_score(y, y_pred)
```

In [244]:

```
selectors = [
    ('Variance Threshold 0.8', lambda: VarianceThreshold(threshold=(.8 * (1 -
.8)))),
    ('RFECV RandomForest', lambda: RFECV(estimator=RandomForestClassifier())),
    ('FWE f_classif', lambda: SelectFwe(f_classif, alpha=0.01)),
    ('NONE', None),
]


samplers = [
    ('smote', SMOTE(random_state=0)),
    ('adasyn', ADASYN(random_state=0)),
    ('svmsmote', SVMSMOTE(random_state=0)),
    ('smotetomek', SMOTETomek(random_state=0)),
    ('RandomOversampling', RandomOverSampler(sampling_strategy='minority')),
    ('NONE', None)
]

models = [
    ('GradientBoosting', lambda: GradientBoostingClassifier(random_state=0, ma
x_depth=3, min_samples_split=42, n_estimators=100)),
    ('RandomForest', lambda: RandomForestClassifier(class_weight='balanced_sub
sample', criterion='entropy', max_depth=None, min_samples_split=100, n_estimat
ors=1000, random_state=0)),
    ('SVC', lambda: SVC(class_weight=None, gamma='auto', kernel='rbf', random_
state=0)),
]

all_descs = ""
all_results = []
for model_desc, model_fn in models:
    for sampler_desc, sampler in samplers:
        for selector_desc, selector in selectors:
            print('-'*80)
            print("EVALUATING: ", model_desc, sampler_desc, selector_desc)
            acc, f1 = model_compare_3(scaled_X.to_numpy(), y.to_numpy(),\
                                      sampler, sampler_desc,\
                                      model_fn, model_desc,\
                                      selector, selector_desc)
            print("%20s %30s %30s %20.10f %20.10f" % (model_desc, sampler_desc
, selector_desc, acc, f1))
            all_descs += "%20s %30s %30s %20.10f %20.10f" % (model_desc, sampl
er_desc, selector_desc, acc, f1)
            all_descs += '\n'
            all_results.append((model_desc, sampler_desc, selector_desc, acc,
f1,))
print(all_descs)
```

```
------------------------------------------------------------------------
---
EVALUATING:  GradientBoosting smote Variance Threshold 0.8
Description: smote Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.91      0.89      0.90      7963
           1       0.60      0.65      0.62      2037

    accuracy                           0.84     10000
   macro avg       0.76      0.77      0.76     10000
weighted avg       0.85      0.84      0.84     10000

------------------------------------------------------------------------
---
    GradientBoosting                          smote         Variance Threshol
d 0.8         0.8411000000         0.6233704669
------------------------------------------------------------------------
---
EVALUATING:  GradientBoosting smote RFECV RandomForest
Description: smote Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.91      0.89      0.90      7963
           1       0.61      0.65      0.63      2037

    accuracy                           0.84     10000
   macro avg       0.76      0.77      0.76     10000
weighted avg       0.85      0.84      0.85     10000

------------------------------------------------------------------------
---
    GradientBoosting                          smote             RFECV RandomF
orest         0.8433000000         0.6268159085
------------------------------------------------------------------------
---
EVALUATING:  GradientBoosting smote FWE f_classif
Description: smote Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.91      0.88      0.89      7963
           1       0.58      0.67      0.62      2037

    accuracy                           0.83     10000
   macro avg       0.75      0.77      0.76     10000
weighted avg       0.84      0.83      0.84     10000

------------------------------------------------------------------------
---
    GradientBoosting                          smote                  FWE f_cl
assif         0.8340000000         0.6227272727
------------------------------------------------------------------------
---
EVALUATING:  GradientBoosting smote NONE
Description: smote Model: GradientBoosting
              precision    recall  f1-score   support
```

```
           0       0.91      0.89      0.90      7963
           1       0.60      0.65      0.62      2037

    accuracy                           0.84     10000
   macro avg       0.76      0.77      0.76     10000
weighted avg       0.85      0.84      0.84     10000

--------------------------------------------------------------------------------
---
    GradientBoosting                          smote
NONE         0.8411000000         0.6233704669
--------------------------------------------------------------------------------
---
EVALUATING:  GradientBoosting adasyn Variance Threshold 0.8
Description: adasyn Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.91      0.89      0.90      7963
           1       0.60      0.65      0.63      2037

    accuracy                           0.84     10000
   macro avg       0.76      0.77      0.76     10000
weighted avg       0.85      0.84      0.84     10000

--------------------------------------------------------------------------------
---
    GradientBoosting                          adasyn         Variance Threshol
d 0.8         0.8414000000         0.6257668712
--------------------------------------------------------------------------------
---
EVALUATING:  GradientBoosting adasyn RFECV RandomForest
Description: adasyn Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.91      0.89      0.90      7963
           1       0.60      0.65      0.63      2037

    accuracy                           0.84     10000
   macro avg       0.76      0.77      0.76     10000
weighted avg       0.85      0.84      0.84     10000

--------------------------------------------------------------------------------
---
    GradientBoosting                          adasyn             RFECV RandomF
orest         0.8417000000         0.6263865943
--------------------------------------------------------------------------------
---
EVALUATING:  GradientBoosting adasyn FWE f_classif
Description: adasyn Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.92      0.86      0.89      7963
           1       0.55      0.70      0.62      2037

    accuracy                           0.82     10000
   macro avg       0.74      0.78      0.75     10000
weighted avg       0.84      0.82      0.83     10000
```

```
--------------------------------------------------------------------------------
    GradientBoosting                          adasyn                  FWE f_cl
assif         0.8239000000         0.6169240809
--------------------------------------------------------------------------------
EVALUATING:  GradientBoosting adasyn NONE
Description: adasyn Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.91      0.89      0.90      7963
           1       0.60      0.65      0.63      2037

    accuracy                           0.84     10000
   macro avg       0.76      0.77      0.76     10000
weighted avg       0.85      0.84      0.84     10000

--------------------------------------------------------------------------------
    GradientBoosting                          adasyn
NONE          0.8414000000         0.6257668712
--------------------------------------------------------------------------------
EVALUATING:  GradientBoosting svmsmote Variance Threshold 0.8
Description: svmsmote Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.91      0.88      0.90      7963
           1       0.59      0.67      0.63      2037

    accuracy                           0.84     10000
   macro avg       0.75      0.78      0.76     10000
weighted avg       0.85      0.84      0.84     10000

--------------------------------------------------------------------------------
    GradientBoosting                         svmsmote         Variance Threshol
d 0.8         0.8390000000         0.6302250804
--------------------------------------------------------------------------------
EVALUATING:  GradientBoosting svmsmote RFECV RandomForest
Description: svmsmote Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.91      0.88      0.90      7963
           1       0.59      0.67      0.63      2037

    accuracy                           0.84     10000
   macro avg       0.75      0.78      0.76     10000
weighted avg       0.85      0.84      0.84     10000

--------------------------------------------------------------------------------
    GradientBoosting                         svmsmote             RFECV RandomF
orest         0.8383000000         0.6275051831
--------------------------------------------------------------------------------
```

```
---
EVALUATING:  GradientBoosting svmsmote FWE f_classif
Description: svmsmote Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.92      0.87      0.89      7963
           1       0.57      0.70      0.63      2037

    accuracy                           0.83     10000
   macro avg       0.75      0.78      0.76     10000
weighted avg       0.85      0.83      0.84     10000

------------------------------------------------------------------------------
---
    GradientBoosting                      svmsmote                  FWE f_cl
assif         0.8324000000         0.6290393980
------------------------------------------------------------------------------
---
EVALUATING:  GradientBoosting svmsmote NONE
Description: svmsmote Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.91      0.88      0.90      7963
           1       0.59      0.67      0.63      2037

    accuracy                           0.84     10000
   macro avg       0.75      0.78      0.76     10000
weighted avg       0.85      0.84      0.84     10000

------------------------------------------------------------------------------
---
    GradientBoosting                      svmsmote
NONE         0.8390000000         0.6302250804
------------------------------------------------------------------------------
---
EVALUATING:  GradientBoosting smotetomek Variance Threshold 0.8
Description: smotetomek Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.91      0.89      0.90      7963
           1       0.60      0.65      0.63      2037

    accuracy                           0.84     10000
   macro avg       0.76      0.77      0.76     10000
weighted avg       0.85      0.84      0.84     10000

------------------------------------------------------------------------------
---
    GradientBoosting                    smotetomek         Variance Threshol
d 0.8         0.8415000000         0.6266195524
------------------------------------------------------------------------------
---
EVALUATING:  GradientBoosting smotetomek RFECV RandomForest
Description: smotetomek Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.91      0.89      0.90      7963
```

```
           1       0.61      0.65      0.63      2037

    accuracy                           0.84     10000
   macro avg       0.76      0.77      0.76     10000
weighted avg       0.85      0.84      0.85     10000

--------------------------------------------------------------------------------
---
    GradientBoosting                    smotetomek             RFECV RandomF
orest         0.8432000000         0.6277302944
--------------------------------------------------------------------------------
---
EVALUATING:  GradientBoosting smotetomek FWE f_classif
Description: smotetomek Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.91      0.88      0.89      7963
           1       0.58      0.67      0.63      2037

    accuracy                           0.84     10000
   macro avg       0.75      0.77      0.76     10000
weighted avg       0.85      0.84      0.84     10000

--------------------------------------------------------------------------------
---
    GradientBoosting                    smotetomek                   FWE f_cl
assif         0.8358000000         0.6251141553
--------------------------------------------------------------------------------
---
EVALUATING:  GradientBoosting smotetomek NONE
Description: smotetomek Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.91      0.89      0.90      7963
           1       0.60      0.65      0.63      2037

    accuracy                           0.84     10000
   macro avg       0.76      0.77      0.76     10000
weighted avg       0.85      0.84      0.84     10000

--------------------------------------------------------------------------------
---
    GradientBoosting                    smotetomek
NONE          0.8415000000         0.6266195524
--------------------------------------------------------------------------------
---
EVALUATING:  GradientBoosting RandomOversampling Variance Threshold 0.8
Description: RandomOversampling Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.93      0.82      0.87      7963
           1       0.51      0.75      0.61      2037

    accuracy                           0.80     10000
   macro avg       0.72      0.78      0.74     10000
weighted avg       0.84      0.80      0.81     10000
```

```
------------------------------------------------------------------------
---
    GradientBoosting             RandomOversampling         Variance Threshol
d 0.8         0.8023000000         0.6065671642
------------------------------------------------------------------------
---
EVALUATING:  GradientBoosting RandomOversampling RFECV RandomForest
Description: RandomOversampling Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.93      0.82      0.87      7963
           1       0.51      0.74      0.61      2037

    accuracy                           0.80     10000
   macro avg       0.72      0.78      0.74     10000
weighted avg       0.84      0.80      0.82     10000

------------------------------------------------------------------------
---
    GradientBoosting             RandomOversampling             RFECV RandomF
orest         0.8043000000         0.6077370214
------------------------------------------------------------------------
---
EVALUATING:  GradientBoosting RandomOversampling FWE f_classif
Description: RandomOversampling Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.93      0.81      0.87      7963
           1       0.51      0.75      0.61      2037

    accuracy                           0.80     10000
   macro avg       0.72      0.78      0.74     10000
weighted avg       0.84      0.80      0.81     10000

------------------------------------------------------------------------
---
    GradientBoosting             RandomOversampling                  FWE f_cl
assif         0.8011000000         0.6063724520
------------------------------------------------------------------------
---
EVALUATING:  GradientBoosting RandomOversampling NONE
Description: RandomOversampling Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.93      0.82      0.87      7963
           1       0.51      0.75      0.61      2037

    accuracy                           0.80     10000
   macro avg       0.72      0.78      0.74     10000
weighted avg       0.84      0.80      0.81     10000

------------------------------------------------------------------------
---
    GradientBoosting             RandomOversampling
NONE          0.8023000000          0.6062537343
------------------------------------------------------------------------
---
```

```
EVALUATING:  GradientBoosting NONE Variance Threshold 0.8
Description: NONE Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.88      0.97      0.92      7963
           1       0.78      0.47      0.59      2037

    accuracy                           0.86     10000
   macro avg       0.83      0.72      0.75     10000
weighted avg       0.86      0.86      0.85     10000

--------------------------------------------------------------------------------
    GradientBoosting                           NONE         Variance Threshol
d 0.8         0.8648000000         0.5857843137
--------------------------------------------------------------------------------
EVALUATING:  GradientBoosting NONE RFECV RandomForest
Description: NONE Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.88      0.97      0.92      7963
           1       0.78      0.47      0.58      2037

    accuracy                           0.86     10000
   macro avg       0.83      0.72      0.75     10000
weighted avg       0.86      0.86      0.85     10000

--------------------------------------------------------------------------------
    GradientBoosting                           NONE             RFECV RandomF
orest         0.8639000000         0.5823872353
--------------------------------------------------------------------------------
EVALUATING:  GradientBoosting NONE FWE f_classif
Description: NONE Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.88      0.96      0.92      7963
           1       0.77      0.47      0.59      2037

    accuracy                           0.86     10000
   macro avg       0.82      0.72      0.75     10000
weighted avg       0.86      0.86      0.85     10000

--------------------------------------------------------------------------------
    GradientBoosting                           NONE                  FWE f_cl
assif         0.8639000000         0.5851874429
--------------------------------------------------------------------------------
EVALUATING:  GradientBoosting NONE NONE
Description: NONE Model: GradientBoosting
              precision    recall  f1-score   support

           0       0.88      0.97      0.92      7963
           1       0.78      0.47      0.59      2037
```

```
    accuracy                           0.86     10000
   macro avg       0.83      0.72      0.75     10000
weighted avg       0.86      0.86      0.85     10000

--------------------------------------------------------------------------------
    GradientBoosting                           NONE
NONE         0.8648000000         0.5857843137
--------------------------------------------------------------------------------
EVALUATING:  RandomForest smote Variance Threshold 0.8
Description: smote Model: RandomForest
              precision    recall  f1-score   support

           0       0.91      0.88      0.90      7963
           1       0.59      0.65      0.62      2037

    accuracy                           0.84     10000
   macro avg       0.75      0.77      0.76     10000
weighted avg       0.84      0.84      0.84     10000

--------------------------------------------------------------------------------
        RandomForest                           smote         Variance Threshol
d 0.8         0.8357000000         0.6185279777
--------------------------------------------------------------------------------
EVALUATING:  RandomForest smote RFECV RandomForest
Description: smote Model: RandomForest
              precision    recall  f1-score   support

           0       0.91      0.88      0.90      7963
           1       0.59      0.66      0.62      2037

    accuracy                           0.84     10000
   macro avg       0.75      0.77      0.76     10000
weighted avg       0.84      0.84      0.84     10000

--------------------------------------------------------------------------------
        RandomForest                           smote             RFECV RandomF
orest         0.8372000000         0.6226240148
--------------------------------------------------------------------------------
EVALUATING:  RandomForest smote FWE f_classif
Description: smote Model: RandomForest
              precision    recall  f1-score   support

           0       0.91      0.88      0.89      7963
           1       0.58      0.68      0.63      2037

    accuracy                           0.84     10000
   macro avg       0.75      0.78      0.76     10000
weighted avg       0.85      0.84      0.84     10000

--------------------------------------------------------------------------------
```

```
---
        RandomForest                            smote                   FWE f_cl
assif         0.8352000000         0.6261343013
--------------------------------------------------------------------------------
---
EVALUATING:  RandomForest smote NONE
Description: smote Model: RandomForest
              precision    recall  f1-score   support

           0       0.91      0.88      0.90      7963
           1       0.59      0.65      0.62      2037

    accuracy                           0.84     10000
   macro avg       0.75      0.77      0.76     10000
weighted avg       0.84      0.84      0.84     10000

--------------------------------------------------------------------------------
---
        RandomForest                            smote
NONE          0.8357000000         0.6185279777
--------------------------------------------------------------------------------
---
EVALUATING:  RandomForest adasyn Variance Threshold 0.8
Description: adasyn Model: RandomForest
              precision    recall  f1-score   support

           0       0.91      0.86      0.88      7963
           1       0.55      0.68      0.61      2037

    accuracy                           0.82     10000
   macro avg       0.73      0.77      0.75     10000
weighted avg       0.84      0.82      0.83     10000

--------------------------------------------------------------------------------
---
        RandomForest                          adasyn         Variance Threshol
d 0.8         0.8221000000         0.6104663893
--------------------------------------------------------------------------------
---
EVALUATING:  RandomForest adasyn RFECV RandomForest
Description: adasyn Model: RandomForest
              precision    recall  f1-score   support

           0       0.91      0.86      0.88      7963
           1       0.55      0.69      0.61      2037

    accuracy                           0.82     10000
   macro avg       0.73      0.77      0.75     10000
weighted avg       0.84      0.82      0.83     10000

--------------------------------------------------------------------------------
---
        RandomForest                            adasyn             RFECV RandomF
orest         0.8205000000         0.6086766950
--------------------------------------------------------------------------------
---
EVALUATING:  RandomForest adasyn FWE f_classif
```

```
Description: adasyn Model: RandomForest
              precision    recall  f1-score   support

           0       0.92      0.84      0.88      7963
           1       0.54      0.71      0.61      2037

    accuracy                           0.82     10000
   macro avg       0.73      0.78      0.75     10000
weighted avg       0.84      0.82      0.83     10000

--------------------------------------------------------------------------------
        RandomForest                         adasyn                  FWE f_cl
assif         0.8170000000         0.6129441624
--------------------------------------------------------------------------------
EVALUATING:  RandomForest adasyn NONE
Description: adasyn Model: RandomForest
              precision    recall  f1-score   support

           0       0.91      0.86      0.88      7963
           1       0.55      0.68      0.61      2037

    accuracy                           0.82     10000
   macro avg       0.73      0.77      0.75     10000
weighted avg       0.84      0.82      0.83     10000

--------------------------------------------------------------------------------
        RandomForest                         adasyn
NONE          0.8221000000         0.6104663893
--------------------------------------------------------------------------------
EVALUATING:  RandomForest svmsmote Variance Threshold 0.8
Description: svmsmote Model: RandomForest
              precision    recall  f1-score   support

           0       0.91      0.87      0.89      7963
           1       0.58      0.68      0.62      2037

    accuracy                           0.83     10000
   macro avg       0.75      0.78      0.76     10000
weighted avg       0.85      0.83      0.84     10000

--------------------------------------------------------------------------------
        RandomForest                        svmsmote         Variance Threshol
d 0.8         0.8331000000         0.6248595190
--------------------------------------------------------------------------------
EVALUATING:  RandomForest svmsmote RFECV RandomForest
Description: svmsmote Model: RandomForest
              precision    recall  f1-score   support

           0       0.92      0.87      0.89      7963
           1       0.58      0.68      0.62      2037
```

```
    accuracy                           0.83     10000
   macro avg       0.75      0.78      0.76     10000
weighted avg       0.85      0.83      0.84     10000

------------------------------------------------------------------------------
---
        RandomForest                        svmsmote             RFECV RandomF
orest         0.8327000000         0.6248037677
------------------------------------------------------------------------------
---
EVALUATING:  RandomForest svmsmote FWE f_classif
Description: svmsmote Model: RandomForest
              precision    recall  f1-score   support

           0       0.92      0.87      0.89      7963
           1       0.57      0.70      0.63      2037

    accuracy                           0.83     10000
   macro avg       0.74      0.78      0.76     10000
weighted avg       0.85      0.83      0.84     10000

------------------------------------------------------------------------------
---
        RandomForest                        svmsmote                   FWE f_cl
assif         0.8321000000         0.6282931149
------------------------------------------------------------------------------
---
EVALUATING:  RandomForest svmsmote NONE
Description: svmsmote Model: RandomForest
              precision    recall  f1-score   support

           0       0.91      0.87      0.89      7963
           1       0.58      0.68      0.62      2037

    accuracy                           0.83     10000
   macro avg       0.75      0.78      0.76     10000
weighted avg       0.85      0.83      0.84     10000

------------------------------------------------------------------------------
---
        RandomForest                        svmsmote
NONE          0.8331000000         0.6248595190
------------------------------------------------------------------------------
---
EVALUATING:  RandomForest smotetomek Variance Threshold 0.8
Description: smotetomek Model: RandomForest
              precision    recall  f1-score   support

           0       0.91      0.88      0.89      7963
           1       0.58      0.65      0.62      2037

    accuracy                           0.83     10000
   macro avg       0.75      0.77      0.76     10000
weighted avg       0.84      0.83      0.84     10000

------------------------------------------------------------------------------
---
```

```
        RandomForest                    smotetomek         Variance Threshol
d 0.8         0.8349000000         0.6177355869
------------------------------------------------------------------------------
---
EVALUATING:  RandomForest smotetomek RFECV RandomForest
Description: smotetomek Model: RandomForest
              precision    recall  f1-score   support

           0       0.91      0.88      0.89      7963
           1       0.59      0.66      0.62      2037

    accuracy                           0.84     10000
   macro avg       0.75      0.77      0.76     10000
weighted avg       0.84      0.84      0.84     10000

------------------------------------------------------------------------------
---
        RandomForest                    smotetomek             RFECV RandomF
orest         0.8352000000         0.6185185185
------------------------------------------------------------------------------
---
EVALUATING:  RandomForest smotetomek FWE f_classif
Description: smotetomek Model: RandomForest
              precision    recall  f1-score   support

           0       0.91      0.88      0.90      7963
           1       0.59      0.68      0.63      2037

    accuracy                           0.84     10000
   macro avg       0.75      0.78      0.76     10000
weighted avg       0.85      0.84      0.84     10000

------------------------------------------------------------------------------
---
        RandomForest                    smotetomek                  FWE f_cl
assif         0.8365000000         0.6283246192
------------------------------------------------------------------------------
---
EVALUATING:  RandomForest smotetomek NONE
Description: smotetomek Model: RandomForest
              precision    recall  f1-score   support

           0       0.91      0.88      0.89      7963
           1       0.58      0.65      0.62      2037

    accuracy                           0.83     10000
   macro avg       0.75      0.77      0.76     10000
weighted avg       0.84      0.83      0.84     10000

------------------------------------------------------------------------------
---
        RandomForest                    smotetomek
NONE         0.8349000000         0.6177355869
------------------------------------------------------------------------------
---
EVALUATING:  RandomForest RandomOversampling Variance Threshold 0.8
Description: RandomOversampling Model: RandomForest
```

```
              precision    recall  f1-score   support

           0       0.92      0.84      0.88      7963
           1       0.54      0.72      0.62      2037

    accuracy                           0.82     10000
   macro avg       0.73      0.78      0.75     10000
weighted avg       0.84      0.82      0.83     10000

--------------------------------------------------------------------------------
        RandomForest             RandomOversampling         Variance Threshol
d 0.8         0.8175000000         0.6160319798
--------------------------------------------------------------------------------
EVALUATING:  RandomForest RandomOversampling RFECV RandomForest
Description: RandomOversampling Model: RandomForest
              precision    recall  f1-score   support

           0       0.92      0.84      0.88      7963
           1       0.54      0.71      0.61      2037

    accuracy                           0.82     10000
   macro avg       0.73      0.78      0.75     10000
weighted avg       0.84      0.82      0.83     10000

--------------------------------------------------------------------------------
        RandomForest             RandomOversampling             RFECV RandomF
orest         0.8172000000         0.6140202703
--------------------------------------------------------------------------------
EVALUATING:  RandomForest RandomOversampling FWE f_classif
Description: RandomOversampling Model: RandomForest
              precision    recall  f1-score   support

           0       0.92      0.84      0.88      7963
           1       0.54      0.72      0.62      2037

    accuracy                           0.82     10000
   macro avg       0.73      0.78      0.75     10000
weighted avg       0.84      0.82      0.83     10000

--------------------------------------------------------------------------------
        RandomForest             RandomOversampling                  FWE f_cl
assif         0.8161000000         0.6158345519
--------------------------------------------------------------------------------
EVALUATING:  RandomForest RandomOversampling NONE
Description: RandomOversampling Model: RandomForest
              precision    recall  f1-score   support

           0       0.92      0.84      0.88      7963
           1       0.54      0.72      0.61      2037

    accuracy                           0.82     10000
```

```
   macro avg       0.73      0.78      0.75     10000
weighted avg       0.84      0.82      0.83     10000

--------------------------------------------------------------------------------
        RandomForest             RandomOversampling
NONE         0.8163000000         0.6138322472
--------------------------------------------------------------------------------
EVALUATING:  RandomForest NONE Variance Threshold 0.8
Description: NONE Model: RandomForest
              precision    recall  f1-score   support

           0       0.92      0.83      0.88      7963
           1       0.53      0.72      0.61      2037

    accuracy                           0.81     10000
   macro avg       0.72      0.78      0.74     10000
weighted avg       0.84      0.81      0.82     10000

--------------------------------------------------------------------------------
        RandomForest                             NONE         Variance Threshol
d 0.8         0.8110000000         0.6078838174
--------------------------------------------------------------------------------
EVALUATING:  RandomForest NONE RFECV RandomForest
Description: NONE Model: RandomForest
              precision    recall  f1-score   support

           0       0.92      0.83      0.88      7963
           1       0.53      0.72      0.61      2037

    accuracy                           0.81     10000
   macro avg       0.72      0.78      0.74     10000
weighted avg       0.84      0.81      0.82     10000

--------------------------------------------------------------------------------
        RandomForest                             NONE             RFECV RandomF
orest         0.8119000000         0.6103169671
--------------------------------------------------------------------------------
EVALUATING:  RandomForest NONE FWE f_classif
Description: NONE Model: RandomForest
              precision    recall  f1-score   support

           0       0.93      0.82      0.87      7963
           1       0.52      0.74      0.61      2037

    accuracy                           0.81     10000
   macro avg       0.72      0.78      0.74     10000
weighted avg       0.84      0.81      0.82     10000

--------------------------------------------------------------------------------
        RandomForest                             NONE                  FWE f_cl
```

```
assif         0.8062000000         0.6088009689
------------------------------------------------------------------------------
---
EVALUATING:  RandomForest NONE NONE
Description: NONE Model: RandomForest
              precision    recall  f1-score   support

           0       0.92      0.83      0.88      7963
           1       0.53      0.72      0.61      2037

    accuracy                           0.81     10000
   macro avg       0.72      0.78      0.74     10000
weighted avg       0.84      0.81      0.82     10000

------------------------------------------------------------------------------
---
        RandomForest                            NONE
NONE         0.8110000000         0.6078838174
------------------------------------------------------------------------------
---
EVALUATING:  SVC smote Variance Threshold 0.8
Description: smote Model: SVC
              precision    recall  f1-score   support

           0       0.92      0.82      0.87      7963
           1       0.50      0.72      0.59      2037

    accuracy                           0.80     10000
   macro avg       0.71      0.77      0.73     10000
weighted avg       0.84      0.80      0.81     10000

------------------------------------------------------------------------------
---
                SVC                            smote         Variance Threshol
d 0.8         0.7985000000         0.5936680782
------------------------------------------------------------------------------
---
EVALUATING:  SVC smote RFECV RandomForest
Description: smote Model: SVC
              precision    recall  f1-score   support

           0       0.92      0.82      0.87      7963
           1       0.50      0.72      0.59      2037

    accuracy                           0.80     10000
   macro avg       0.71      0.77      0.73     10000
weighted avg       0.84      0.80      0.81     10000

------------------------------------------------------------------------------
---
                SVC                            smote             RFECV RandomF
orest         0.7986000000         0.5939516129
------------------------------------------------------------------------------
---
EVALUATING:  SVC smote FWE f_classif
Description: smote Model: SVC
              precision    recall  f1-score   support
```

```
           0       0.92      0.82      0.87      7963
           1       0.51      0.73      0.60      2037

    accuracy                           0.80     10000
   macro avg       0.71      0.77      0.73     10000
weighted avg       0.84      0.80      0.81     10000

------------------------------------------------------------------------------
---
                SVC                           smote                  FWE f_cl
assif         0.8006000000         0.5991154001
------------------------------------------------------------------------------
---
EVALUATING:  SVC smote NONE
Description: smote Model: SVC
              precision    recall  f1-score   support

           0       0.92      0.82      0.87      7963
           1       0.50      0.72      0.59      2037

    accuracy                           0.80     10000
   macro avg       0.71      0.77      0.73     10000
weighted avg       0.84      0.80      0.81     10000

------------------------------------------------------------------------------
---
                SVC                           smote
NONE          0.7985000000         0.5936680782
------------------------------------------------------------------------------
---
EVALUATING:  SVC adasyn Variance Threshold 0.8
Description: adasyn Model: SVC
              precision    recall  f1-score   support

           0       0.93      0.77      0.84      7963
           1       0.46      0.77      0.57      2037

    accuracy                           0.77     10000
   macro avg       0.69      0.77      0.71     10000
weighted avg       0.83      0.77      0.79     10000

------------------------------------------------------------------------------
---
                SVC                         adasyn         Variance Threshol
d 0.8         0.7668000000         0.5724239091
------------------------------------------------------------------------------
---
EVALUATING:  SVC adasyn RFECV RandomForest
Description: adasyn Model: SVC
              precision    recall  f1-score   support

           0       0.93      0.77      0.84      7963
           1       0.46      0.77      0.57      2037

    accuracy                           0.77     10000
   macro avg       0.69      0.77      0.71     10000
```

```
weighted avg       0.83      0.77      0.79     10000

------------------------------------------------------------------------------
---
                SVC                          adasyn             RFECV RandomF
orest         0.7668000000         0.5724239091
------------------------------------------------------------------------------
---
EVALUATING:  SVC adasyn FWE f_classif
Description: adasyn Model: SVC
              precision    recall  f1-score   support

           0       0.93      0.76      0.84      7963
           1       0.46      0.78      0.58      2037

    accuracy                           0.77     10000
   macro avg       0.70      0.77      0.71     10000
weighted avg       0.84      0.77      0.79     10000

------------------------------------------------------------------------------
---
                SVC                          adasyn                   FWE f_cl
assif         0.7672000000         0.5784136183
------------------------------------------------------------------------------
---
EVALUATING:  SVC adasyn NONE
Description: adasyn Model: SVC
              precision    recall  f1-score   support

           0       0.93      0.77      0.84      7963
           1       0.46      0.77      0.57      2037

    accuracy                           0.77     10000
   macro avg       0.69      0.77      0.71     10000
weighted avg       0.83      0.77      0.79     10000

------------------------------------------------------------------------------
---
                SVC                          adasyn
NONE          0.7668000000         0.5724239091
------------------------------------------------------------------------------
---
EVALUATING:  SVC svmsmote Variance Threshold 0.8
Description: svmsmote Model: SVC
              precision    recall  f1-score   support

           0       0.92      0.81      0.87      7963
           1       0.50      0.74      0.60      2037

    accuracy                           0.80     10000
   macro avg       0.71      0.78      0.73     10000
weighted avg       0.84      0.80      0.81     10000

------------------------------------------------------------------------------
---
                SVC                        svmsmote         Variance Threshol
d 0.8         0.7984000000         0.5985663082
```

```
------------------------------------------------------------------------
---
EVALUATING:  SVC svmsmote RFECV RandomForest
Description: svmsmote Model: SVC
              precision    recall  f1-score   support

           0       0.92      0.81      0.87      7963
           1       0.50      0.74      0.60      2037

    accuracy                           0.80     10000
   macro avg       0.71      0.78      0.73     10000
weighted avg       0.84      0.80      0.81     10000

------------------------------------------------------------------------
---
                SVC                       svmsmote             RFECV RandomF
orest         0.7991000000         0.5997210600
------------------------------------------------------------------------
---
EVALUATING:  SVC svmsmote FWE f_classif
Description: svmsmote Model: SVC
              precision    recall  f1-score   support

           0       0.93      0.81      0.86      7963
           1       0.50      0.74      0.60      2037

    accuracy                           0.80     10000
   macro avg       0.71      0.78      0.73     10000
weighted avg       0.84      0.80      0.81     10000

------------------------------------------------------------------------
---
                SVC                       svmsmote                  FWE f_cl
assif         0.7979000000         0.6001978239
------------------------------------------------------------------------
---
EVALUATING:  SVC svmsmote NONE
Description: svmsmote Model: SVC
              precision    recall  f1-score   support

           0       0.92      0.81      0.87      7963
           1       0.50      0.74      0.60      2037

    accuracy                           0.80     10000
   macro avg       0.71      0.78      0.73     10000
weighted avg       0.84      0.80      0.81     10000

------------------------------------------------------------------------
---
                SVC                       svmsmote
NONE          0.7984000000         0.5985663082
------------------------------------------------------------------------
---
EVALUATING:  SVC smotetomek Variance Threshold 0.8
Description: smotetomek Model: SVC
              precision    recall  f1-score   support
```

```
              precision    recall  f1-score   support

           0       0.92      0.82      0.87      7963
           1       0.50      0.72      0.59      2037

    accuracy                           0.80     10000
   macro avg       0.71      0.77      0.73     10000
weighted avg       0.84      0.80      0.81     10000

--------------------------------------------------------------------------------
                SVC                    smotetomek         Variance Threshol
d 0.8         0.7978000000         0.5928312525
--------------------------------------------------------------------------------
EVALUATING:  SVC smotetomek RFECV RandomForest
Description: smotetomek Model: SVC
              precision    recall  f1-score   support

           0       0.92      0.82      0.87      7963
           1       0.50      0.72      0.59      2037

    accuracy                           0.80     10000
   macro avg       0.71      0.77      0.73     10000
weighted avg       0.83      0.80      0.81     10000

--------------------------------------------------------------------------------
                SVC                    smotetomek             RFECV RandomF
orest         0.7973000000         0.5915776748
--------------------------------------------------------------------------------
EVALUATING:  SVC smotetomek FWE f_classif
Description: smotetomek Model: SVC
              precision    recall  f1-score   support

           0       0.92      0.82      0.87      7963
           1       0.50      0.73      0.60      2037

    accuracy                           0.80     10000
   macro avg       0.71      0.77      0.73     10000
weighted avg       0.84      0.80      0.81     10000

--------------------------------------------------------------------------------
                SVC                    smotetomek                  FWE f_cl
assif         0.7986000000         0.5957446809
--------------------------------------------------------------------------------
EVALUATING:  SVC smotetomek NONE
Description: smotetomek Model: SVC
              precision    recall  f1-score   support

           0       0.92      0.82      0.87      7963
           1       0.50      0.72      0.59      2037

    accuracy                           0.80     10000
   macro avg       0.71      0.77      0.73     10000
weighted avg       0.84      0.80      0.81     10000
```

```
--------------------------------------------------------------------------------
---
                SVC                    smotetomek
NONE         0.7978000000         0.5928312525
--------------------------------------------------------------------------------
---
EVALUATING:  SVC RandomOversampling Variance Threshold 0.8
Description: RandomOversampling Model: SVC
              precision    recall  f1-score   support

           0       0.93      0.81      0.86      7963
           1       0.50      0.75      0.60      2037

    accuracy                           0.79     10000
   macro avg       0.71      0.78      0.73     10000
weighted avg       0.84      0.79      0.81     10000

--------------------------------------------------------------------------------
---
                SVC             RandomOversampling         Variance Threshol
d 0.8         0.7942000000         0.5961538462
--------------------------------------------------------------------------------
---
EVALUATING:  SVC RandomOversampling RFECV RandomForest
Description: RandomOversampling Model: SVC
              precision    recall  f1-score   support

           0       0.93      0.81      0.86      7963
           1       0.50      0.75      0.60      2037

    accuracy                           0.79     10000
   macro avg       0.71      0.78      0.73     10000
weighted avg       0.84      0.79      0.81     10000

--------------------------------------------------------------------------------
---
                SVC             RandomOversampling             RFECV RandomF
orest         0.7934000000         0.5950607605
--------------------------------------------------------------------------------
---
EVALUATING:  SVC RandomOversampling FWE f_classif
Description: RandomOversampling Model: SVC
              precision    recall  f1-score   support

           0       0.93      0.81      0.86      7963
           1       0.50      0.75      0.60      2037

    accuracy                           0.80     10000
   macro avg       0.71      0.78      0.73     10000
weighted avg       0.84      0.80      0.81     10000

--------------------------------------------------------------------------------
---
                SVC             RandomOversampling                  FWE f_cl
assif         0.7950000000         0.5997657165
--------------------------------------------------------------------------------
```

```
---
EVALUATING:  SVC RandomOversampling NONE
Description: RandomOversampling Model: SVC
              precision    recall  f1-score   support

           0       0.93      0.81      0.86      7963
           1       0.50      0.75      0.60      2037

    accuracy                           0.80     10000
   macro avg       0.71      0.78      0.73     10000
weighted avg       0.84      0.80      0.81     10000

------------------------------------------------------------------------
---
                SVC             RandomOversampling
NONE         0.7973000000         0.5999605289
------------------------------------------------------------------------
---
EVALUATING:  SVC NONE Variance Threshold 0.8
Description: NONE Model: SVC
              precision    recall  f1-score   support

           0       0.86      0.98      0.92      7963
           1       0.82      0.38      0.52      2037

    accuracy                           0.86     10000
   macro avg       0.84      0.68      0.72     10000
weighted avg       0.85      0.86      0.84     10000

------------------------------------------------------------------------
---
                SVC                           NONE         Variance Threshol
d 0.8         0.8569000000         0.5196374622
------------------------------------------------------------------------
---
EVALUATING:  SVC NONE RFECV RandomForest
Description: NONE Model: SVC
              precision    recall  f1-score   support

           0       0.86      0.98      0.92      7963
           1       0.82      0.38      0.52      2037

    accuracy                           0.86     10000
   macro avg       0.84      0.68      0.72     10000
weighted avg       0.85      0.86      0.84     10000

------------------------------------------------------------------------
---
                SVC                           NONE             RFECV RandomF
orest         0.8575000000         0.5235707121
------------------------------------------------------------------------
---
EVALUATING:  SVC NONE FWE f_classif
Description: NONE Model: SVC
              precision    recall  f1-score   support

           0       0.86      0.98      0.92      7963
```

```
           1       0.82      0.39      0.53      2037

    accuracy                           0.86     10000
   macro avg       0.84      0.68      0.72     10000
weighted avg       0.85      0.86      0.84     10000

--------------------------------------------------------------------------------
---
                SVC                           NONE                  FWE f_cl
assif         0.8581000000         0.5274725275
--------------------------------------------------------------------------------
---
EVALUATING:  SVC NONE NONE
Description: NONE Model: SVC
              precision    recall  f1-score   support

           0       0.86      0.98      0.92      7963
           1       0.82      0.38      0.52      2037

    accuracy                           0.86     10000
   macro avg       0.84      0.68      0.72     10000
weighted avg       0.85      0.86      0.84     10000

--------------------------------------------------------------------------------
---
                SVC                           NONE
NONE          0.8569000000         0.5196374622
    GradientBoosting                         smote         Variance Threshol
d 0.8         0.8411000000         0.6233704669
    GradientBoosting                         smote             RFECV RandomF
orest         0.8433000000         0.6268159085
    GradientBoosting                         smote                  FWE f_cl
assif         0.8340000000         0.6227272727
    GradientBoosting                         smote
NONE          0.8411000000         0.6233704669
    GradientBoosting                        adasyn         Variance Threshol
d 0.8         0.8414000000         0.6257668712
    GradientBoosting                        adasyn             RFECV RandomF
orest         0.8417000000         0.6263865943
    GradientBoosting                        adasyn                  FWE f_cl
assif         0.8239000000         0.6169240809
    GradientBoosting                        adasyn
NONE          0.8414000000         0.6257668712
    GradientBoosting                       svmsmote         Variance Threshol
d 0.8         0.8390000000         0.6302250804
    GradientBoosting                       svmsmote             RFECV RandomF
orest         0.8383000000         0.6275051831
    GradientBoosting                       svmsmote                  FWE f_cl
assif         0.8324000000         0.6290393980
    GradientBoosting                       svmsmote
NONE          0.8390000000         0.6302250804
    GradientBoosting                     smotetomek         Variance Threshol
d 0.8         0.8415000000         0.6266195524
    GradientBoosting                     smotetomek             RFECV RandomF
orest         0.8432000000         0.6277302944
    GradientBoosting                     smotetomek                  FWE f_cl
assif         0.8358000000         0.6251141553
```

```
    GradientBoosting                 smotetomek
NONE         0.8415000000        0.6266195524
    GradientBoosting            RandomOversampling         Variance Threshol
d 0.8         0.8023000000          0.6065671642
    GradientBoosting            RandomOversampling             RFECV RandomF
orest         0.8043000000          0.6077370214
    GradientBoosting            RandomOversampling                 FWE f_cl
assif         0.8011000000          0.6063724520
    GradientBoosting            RandomOversampling
NONE         0.8023000000        0.6062537343
    GradientBoosting                       NONE         Variance Threshol
d 0.8         0.8648000000          0.5857843137
    GradientBoosting                       NONE             RFECV RandomF
orest         0.8639000000          0.5823872353
    GradientBoosting                       NONE                 FWE f_cl
assif         0.8639000000          0.5851874429
    GradientBoosting                       NONE
NONE         0.8648000000        0.5857843137
        RandomForest                      smote         Variance Threshol
d 0.8         0.8357000000          0.6185279777
        RandomForest                      smote             RFECV RandomF
orest         0.8372000000          0.6226240148
        RandomForest                      smote                 FWE f_cl
assif         0.8352000000          0.6261343013
        RandomForest                      smote
NONE         0.8357000000        0.6185279777
        RandomForest                     adasyn         Variance Threshol
d 0.8         0.8221000000          0.6104663893
        RandomForest                     adasyn             RFECV RandomF
orest         0.8205000000          0.6086766950
        RandomForest                     adasyn                 FWE f_cl
assif         0.8170000000          0.6129441624
        RandomForest                     adasyn
NONE         0.8221000000        0.6104663893
        RandomForest                    svmsmote         Variance Threshol
d 0.8         0.8331000000          0.6248595190
        RandomForest                    svmsmote             RFECV RandomF
orest         0.8327000000          0.6248037677
        RandomForest                    svmsmote                 FWE f_cl
assif         0.8321000000          0.6282931149
        RandomForest                    svmsmote
NONE         0.8331000000        0.6248595190
        RandomForest                   smotetomek         Variance Threshol
d 0.8         0.8349000000          0.6177355869
        RandomForest                   smotetomek             RFECV RandomF
orest         0.8352000000          0.6185185185
        RandomForest                   smotetomek                 FWE f_cl
assif         0.8365000000          0.6283246192
        RandomForest                   smotetomek
NONE         0.8349000000        0.6177355869
        RandomForest            RandomOversampling         Variance Threshol
d 0.8         0.8175000000          0.6160319798
        RandomForest            RandomOversampling             RFECV RandomF
orest         0.8172000000          0.6140202703
        RandomForest            RandomOversampling                 FWE f_cl
assif         0.8161000000          0.6158345519
        RandomForest            RandomOversampling
```

```
NONE         0.8163000000        0.6138322472
       RandomForest                        NONE         Variance Threshol
d 0.8        0.8110000000         0.6078838174
       RandomForest                        NONE             RFECV RandomF
orest        0.8119000000         0.6103169671
       RandomForest                        NONE                 FWE f_cl
assif        0.8062000000         0.6088009689
       RandomForest                        NONE
NONE         0.8110000000        0.6078838174
                  SVC                       smote         Variance Threshol
d 0.8         0.7985000000         0.5936680782
                  SVC                       smote             RFECV RandomF
orest         0.7986000000         0.5939516129
                  SVC                       smote                 FWE f_cl
assif         0.8006000000         0.5991154001
                  SVC                       smote
NONE         0.7985000000        0.5936680782
                  SVC                      adasyn         Variance Threshol
d 0.8         0.7668000000         0.5724239091
                  SVC                      adasyn             RFECV RandomF
orest         0.7668000000         0.5724239091
                  SVC                      adasyn                 FWE f_cl
assif         0.7672000000         0.5784136183
                  SVC                      adasyn
NONE         0.7668000000        0.5724239091
                  SVC                    svmsmote         Variance Threshol
d 0.8         0.7984000000         0.5985663082
                  SVC                    svmsmote             RFECV RandomF
orest         0.7991000000         0.5997210600
                  SVC                    svmsmote                 FWE f_cl
assif         0.7979000000         0.6001978239
                  SVC                    svmsmote
NONE         0.7984000000        0.5985663082
                  SVC                  smotetomek         Variance Threshol
d 0.8         0.7978000000         0.5928312525
                  SVC                  smotetomek             RFECV RandomF
orest         0.7973000000         0.5915776748
                  SVC                  smotetomek                 FWE f_cl
assif         0.7986000000         0.5957446809
                  SVC                  smotetomek
NONE         0.7978000000         0.5928312525
                  SVC           RandomOversampling         Variance Threshol
d 0.8         0.7942000000          0.5961538462
                  SVC           RandomOversampling             RFECV RandomF
orest         0.7934000000          0.5950607605
                  SVC           RandomOversampling                 FWE f_cl
assif         0.7950000000          0.5997657165
                  SVC           RandomOversampling
NONE         0.7973000000         0.5999605289
                  SVC                       NONE         Variance Threshol
d 0.8         0.8569000000          0.5196374622
                  SVC                       NONE             RFECV RandomF
orest         0.8575000000         0.5235707121
                  SVC                       NONE                 FWE f_cl
assif         0.8581000000         0.5274725275
                  SVC                       NONE
```

NONE	0.8569000000	0.5196374622

In [90]:
```
all_f1s = []
for model_desc, sampler_desc, selector_desc, acc, f1 in all_results:
    all_f1s.append(f1)
sorted_indices = np.argsort(all_f1s)[::-1]
print("%20s %22s %25s %10s %10s" % ('ALGORITHM', 'SAMPLING', 'FEATURE-SELECTIO
N', 'ACC', 'F1'))
print("%20s %22s %25s %10s %10s" % ('-'*20, '-'*22,  '-'*25, '-'*10, '-'*10))
for i in sorted_indices:
    model_desc, sampler_desc, selector_desc, acc, f1 = all_results[i]
    print("%20s %22s %25s %10.5f %10.5f" % (model_desc, sampler_desc, selector
_desc, acc, f1))
```

ALGORITHM	SAMPLING	FEATURE-SELECTION	ACC	F1
GradientBoosting	svmsmote	Variance Threshold 0.8	0.83900	0.63023
GradientBoosting	svmsmote	NONE	0.83900	0.63023
GradientBoosting	svmsmote	RFECV RandomForest	0.83850	0.62933
GradientBoosting	svmsmote	FWE f_classif	0.83240	0.62904
RandomForest	smotetomek	FWE f_classif	0.83650	0.62832
RandomForest	svmsmote	FWE f_classif	0.83210	0.62829
GradientBoosting	smotetomek	Variance Threshold 0.8	0.84150	0.62662
GradientBoosting	smotetomek	NONE	0.84150	0.62662
GradientBoosting	smotetomek	RFECV RandomForest	0.84140	0.62630
RandomForest	smote	FWE f_classif	0.83520	0.62613
GradientBoosting	adasyn	RFECV RandomForest	0.84160	0.62589
GradientBoosting	adasyn	NONE	0.84140	0.62577
GradientBoosting	adasyn	Variance Threshold 0.8	0.84140	0.62577
GradientBoosting	smotetomek	FWE f_classif	0.83580	0.62511
RandomForest	svmsmote	Variance Threshold 0.8	0.83310	0.62486
RandomForest	svmsmote	NONE	0.83310	0.62486
GradientBoosting	smote	RFECV RandomForest	0.84190	0.62348
GradientBoosting	smote	NONE	0.84110	0.62337
GradientBoosting	smote	Variance Threshold 0.8	0.84110	0.62337
RandomForest	svmsmote	RFECV RandomForest	0.83200	0.62281
GradientBoosting	smote	FWE f_classif	0.83400	0.62273
RandomForest	smotetomek	RFECV RandomForest	0.83640	0.62059
RandomForest	smote	RFECV RandomForest	0.83630	0.62027
RandomForest	RandomOversampling	RFECV RandomForest	0.81940	0.61883
RandomForest	smote	NONE	0.83570	0.61853
RandomForest	smote	Variance Threshold 0.8	0.83570	0.61853
RandomForest	smotetomek	NONE	0.83	

```
490    0.61774
       RandomForest             smotetomek    Variance Threshold 0.8    0.83
490    0.61774
    GradientBoosting                adasyn             FWE f_classif    0.82
390    0.61692
       RandomForest     RandomOversampling             FWE f_classif    0.81
400    0.61395
       RandomForest                adasyn        RFECV RandomForest    0.82
360    0.61299
       RandomForest                adasyn             FWE f_classif    0.81
700    0.61294
       RandomForest     RandomOversampling    Variance Threshold 0.8    0.81
520    0.61193
       RandomForest     RandomOversampling                     NONE    0.81
560    0.61113
       RandomForest                adasyn                     NONE    0.82
210    0.61047
       RandomForest                adasyn    Variance Threshold 0.8    0.82
210    0.61047
       RandomForest                  NONE        RFECV RandomForest    0.81
260    0.60926
       RandomForest                  NONE             FWE f_classif    0.80
620    0.60880
    GradientBoosting     RandomOversampling                     NONE    0.80
410    0.60875
    GradientBoosting     RandomOversampling        RFECV RandomForest    0.80
490    0.60815
       RandomForest                  NONE    Variance Threshold 0.8    0.81
100    0.60788
       RandomForest                  NONE                     NONE    0.81
100    0.60788
    GradientBoosting     RandomOversampling             FWE f_classif    0.80
280    0.60686
    GradientBoosting     RandomOversampling    Variance Threshold 0.8    0.80
220    0.60487
                SVC     RandomOversampling                     NONE    0.79
760    0.60063
                SVC               svmsmote             FWE f_classif    0.79
790    0.60020
                SVC                  smote             FWE f_classif    0.80
060    0.59912
                SVC               svmsmote                     NONE    0.79
840    0.59857
                SVC               svmsmote    Variance Threshold 0.8    0.79
840    0.59857
                SVC     RandomOversampling             FWE f_classif    0.79
440    0.59797
                SVC               svmsmote        RFECV RandomForest    0.79
790    0.59717
                SVC     RandomOversampling        RFECV RandomForest    0.79
570    0.59696
                SVC             smotetomek             FWE f_classif    0.79
860    0.59574
                SVC     RandomOversampling    Variance Threshold 0.8    0.79
290    0.59559
                SVC                  smote                     NONE    0.79
850    0.59367
```

```
                SVC                  smote    Variance Threshold 0.8    0.79
850    0.59367
                SVC             smotetomek    Variance Threshold 0.8    0.79
780    0.59283
                SVC             smotetomek                      NONE    0.79
780    0.59283
                SVC             smotetomek        RFECV RandomForest    0.79
760    0.59243
                SVC                  smote        RFECV RandomForest    0.79
720    0.59179
    GradientBoosting                   NONE    Variance Threshold 0.8    0.86
480    0.58578
    GradientBoosting                   NONE                      NONE    0.86
480    0.58578
    GradientBoosting                   NONE             FWE f_classif    0.86
390    0.58519
    GradientBoosting                   NONE        RFECV RandomForest    0.86
400    0.58154
                SVC                 adasyn             FWE f_classif    0.76
720    0.57841
                SVC                 adasyn    Variance Threshold 0.8    0.76
680    0.57242
                SVC                 adasyn                      NONE    0.76
680    0.57242
                SVC                 adasyn        RFECV RandomForest    0.76
600    0.57127
                SVC                   NONE             FWE f_classif    0.85
810    0.52747
                SVC                   NONE        RFECV RandomForest    0.85
720    0.52336
                SVC                   NONE    Variance Threshold 0.8    0.85
690    0.51964
                SVC                   NONE                      NONE    0.85
690    0.51964
```

Here we can see that applying variance-threshold feature selection and no feature selection gave the exact same results. This is because no feature was eliminated.

In [99]:

```
selectors = [
    ('NONE', None),
]

for i in np.arange(0, 1.01, 0.05):
    selectors.append(('Variance Threshold %f' % i, lambda: VarianceThreshold(i
* (1-i)),))




model = lambda: GradientBoostingClassifier(random_state=0, max_depth=3, min_sa
mples_split=42, n_estimators=100)

all_descs2 = ""
all_results2 = []
for selector_desc, selector in selectors:
    model_desc = "GradientBoosting"
    sampler_desc = "svmsmote"
    print('-'*80)
    print("EVALUATING: ", model_desc, sampler_desc, selector_desc)
    acc, f1 = model_compare_3(scaled_X.to_numpy(), y.to_numpy(),\
                              SVMSMOTE(random_state=0), 'svmsmote',\
                              model, 'gradient-boosting',\
                              selector, selector_desc)
    print("%20s %30s %30s %20.10f %20.10f" % (model_desc, sampler_desc, select
or_desc, acc, f1))
    all_descs2 += "%20s %30s %30s %20.10f %20.10f" % (model_desc, sampler_desc
, selector_desc, acc, f1)
    all_descs2 += '\n'
    all_results2.append((model_desc, sampler_desc, selector_desc, acc, f1,))


    print(all_descs2)
```

```
--------------------------------------------------------------------------------
---
EVALUATING:  GradientBoosting svmsmote NONE
Description: svmsmote Model: gradient-boosting
              precision    recall  f1-score   support

           0       0.91      0.88      0.90      7963
           1       0.59      0.67      0.63      2037

    accuracy                           0.84     10000
   macro avg       0.75      0.78      0.76     10000
weighted avg       0.85      0.84      0.84     10000

--------------------------------------------------------------------------------
---
    GradientBoosting                      svmsmote
NONE         0.8390000000         0.6302250804
--------------------------------------------------------------------------------
---
EVALUATING:  GradientBoosting svmsmote Variance Threshold 0.000000
Description: svmsmote Model: gradient-boosting
              precision    recall  f1-score   support

           0       0.91      0.88      0.90      7963
           1       0.59      0.67      0.63      2037

    accuracy                           0.84     10000
   macro avg       0.75      0.78      0.76     10000
weighted avg       0.85      0.84      0.84     10000

--------------------------------------------------------------------------------
---
    GradientBoosting                      svmsmote    Variance Threshold 0.0
00000         0.8390000000         0.6302250804
--------------------------------------------------------------------------------
---
EVALUATING:  GradientBoosting svmsmote Variance Threshold 0.050000
Description: svmsmote Model: gradient-boosting
              precision    recall  f1-score   support

           0       0.91      0.88      0.90      7963
           1       0.59      0.67      0.63      2037

    accuracy                           0.84     10000
   macro avg       0.75      0.78      0.76     10000
weighted avg       0.85      0.84      0.84     10000

--------------------------------------------------------------------------------
---
    GradientBoosting                      svmsmote    Variance Threshold 0.0
50000         0.8390000000         0.6302250804
--------------------------------------------------------------------------------
---
EVALUATING:  GradientBoosting svmsmote Variance Threshold 0.100000
Description: svmsmote Model: gradient-boosting
              precision    recall  f1-score   support
```

```
           0       0.91      0.88      0.90      7963
           1       0.59      0.67      0.63      2037

    accuracy                           0.84     10000
   macro avg       0.75      0.78      0.76     10000
weighted avg       0.85      0.84      0.84     10000

--------------------------------------------------------------------------------
---
    GradientBoosting                      svmsmote    Variance Threshold 0.1
00000         0.8390000000         0.6302250804
--------------------------------------------------------------------------------
---
EVALUATING:  GradientBoosting svmsmote Variance Threshold 0.150000
Description: svmsmote Model: gradient-boosting
              precision    recall  f1-score   support

           0       0.91      0.88      0.90      7963
           1       0.59      0.67      0.63      2037

    accuracy                           0.84     10000
   macro avg       0.75      0.78      0.76     10000
weighted avg       0.85      0.84      0.84     10000

--------------------------------------------------------------------------------
---
    GradientBoosting                      svmsmote    Variance Threshold 0.1
50000         0.8390000000         0.6302250804
--------------------------------------------------------------------------------
---
EVALUATING:  GradientBoosting svmsmote Variance Threshold 0.200000
Description: svmsmote Model: gradient-boosting
              precision    recall  f1-score   support

           0       0.91      0.88      0.90      7963
           1       0.59      0.67      0.63      2037

    accuracy                           0.84     10000
   macro avg       0.75      0.78      0.76     10000
weighted avg       0.85      0.84      0.84     10000

--------------------------------------------------------------------------------
---
    GradientBoosting                      svmsmote    Variance Threshold 0.2
00000         0.8390000000         0.6302250804
--------------------------------------------------------------------------------
---
EVALUATING:  GradientBoosting svmsmote Variance Threshold 0.250000
Description: svmsmote Model: gradient-boosting
              precision    recall  f1-score   support

           0       0.91      0.88      0.90      7963
           1       0.59      0.67      0.63      2037

    accuracy                           0.84     10000
   macro avg       0.75      0.78      0.76     10000
weighted avg       0.85      0.84      0.84     10000
```

```
--------------------------------------------------------------------------------
---
    GradientBoosting                      svmsmote    Variance Threshold 0.2
50000         0.8390000000         0.6302250804
--------------------------------------------------------------------------------
---
EVALUATING:  GradientBoosting svmsmote Variance Threshold 0.300000
Description: svmsmote Model: gradient-boosting
              precision    recall  f1-score   support

           0       0.91      0.88      0.90      7963
           1       0.59      0.67      0.63      2037

    accuracy                           0.84     10000
   macro avg       0.75      0.78      0.76     10000
weighted avg       0.85      0.84      0.84     10000

--------------------------------------------------------------------------------
---
    GradientBoosting                      svmsmote    Variance Threshold 0.3
00000         0.8390000000         0.6302250804
--------------------------------------------------------------------------------
---
EVALUATING:  GradientBoosting svmsmote Variance Threshold 0.350000
Description: svmsmote Model: gradient-boosting
              precision    recall  f1-score   support

           0       0.91      0.88      0.90      7963
           1       0.59      0.67      0.63      2037

    accuracy                           0.84     10000
   macro avg       0.75      0.78      0.76     10000
weighted avg       0.85      0.84      0.84     10000

--------------------------------------------------------------------------------
---
    GradientBoosting                      svmsmote    Variance Threshold 0.3
50000         0.8390000000         0.6302250804
--------------------------------------------------------------------------------
---
EVALUATING:  GradientBoosting svmsmote Variance Threshold 0.400000
Description: svmsmote Model: gradient-boosting
              precision    recall  f1-score   support

           0       0.91      0.88      0.90      7963
           1       0.59      0.67      0.63      2037

    accuracy                           0.84     10000
   macro avg       0.75      0.78      0.76     10000
weighted avg       0.85      0.84      0.84     10000

--------------------------------------------------------------------------------
---
    GradientBoosting                      svmsmote    Variance Threshold 0.4
00000         0.8390000000         0.6302250804
--------------------------------------------------------------------------------
```

```
---
EVALUATING:  GradientBoosting svmsmote Variance Threshold 0.450000
Description: svmsmote Model: gradient-boosting
              precision    recall  f1-score   support

           0       0.91      0.88      0.90      7963
           1       0.59      0.67      0.63      2037

    accuracy                           0.84     10000
   macro avg       0.75      0.78      0.76     10000
weighted avg       0.85      0.84      0.84     10000

--------------------------------------------------------------------------------
---
    GradientBoosting                       svmsmote    Variance Threshold 0.4
50000         0.8390000000         0.6302250804
--------------------------------------------------------------------------------
---
EVALUATING:  GradientBoosting svmsmote Variance Threshold 0.500000
Description: svmsmote Model: gradient-boosting
              precision    recall  f1-score   support

           0       0.91      0.88      0.90      7963
           1       0.59      0.67      0.63      2037

    accuracy                           0.84     10000
   macro avg       0.75      0.78      0.76     10000
weighted avg       0.85      0.84      0.84     10000

--------------------------------------------------------------------------------
---
    GradientBoosting                       svmsmote    Variance Threshold 0.5
00000         0.8390000000         0.6302250804
--------------------------------------------------------------------------------
---
EVALUATING:  GradientBoosting svmsmote Variance Threshold 0.550000
Description: svmsmote Model: gradient-boosting
              precision    recall  f1-score   support

           0       0.91      0.88      0.90      7963
           1       0.59      0.67      0.63      2037

    accuracy                           0.84     10000
   macro avg       0.75      0.78      0.76     10000
weighted avg       0.85      0.84      0.84     10000

--------------------------------------------------------------------------------
---
    GradientBoosting                       svmsmote    Variance Threshold 0.5
50000         0.8390000000         0.6302250804
--------------------------------------------------------------------------------
---
EVALUATING:  GradientBoosting svmsmote Variance Threshold 0.600000
Description: svmsmote Model: gradient-boosting
              precision    recall  f1-score   support

           0       0.91      0.88      0.90      7963
```

```
           1       0.59      0.67      0.63      2037

    accuracy                           0.84     10000
   macro avg       0.75      0.78      0.76     10000
weighted avg       0.85      0.84      0.84     10000

--------------------------------------------------------------------------------
---
    GradientBoosting                      svmsmote    Variance Threshold 0.6
00000         0.8390000000         0.6302250804
--------------------------------------------------------------------------------
---
EVALUATING:  GradientBoosting svmsmote Variance Threshold 0.650000
Description: svmsmote Model: gradient-boosting
              precision    recall  f1-score   support

           0       0.91      0.88      0.90      7963
           1       0.59      0.67      0.63      2037

    accuracy                           0.84     10000
   macro avg       0.75      0.78      0.76     10000
weighted avg       0.85      0.84      0.84     10000

--------------------------------------------------------------------------------
---
    GradientBoosting                      svmsmote    Variance Threshold 0.6
50000         0.8390000000         0.6302250804
--------------------------------------------------------------------------------
---
EVALUATING:  GradientBoosting svmsmote Variance Threshold 0.700000
Description: svmsmote Model: gradient-boosting
              precision    recall  f1-score   support

           0       0.91      0.88      0.90      7963
           1       0.59      0.67      0.63      2037

    accuracy                           0.84     10000
   macro avg       0.75      0.78      0.76     10000
weighted avg       0.85      0.84      0.84     10000

--------------------------------------------------------------------------------
---
    GradientBoosting                      svmsmote    Variance Threshold 0.7
00000         0.8390000000         0.6302250804
--------------------------------------------------------------------------------
---
EVALUATING:  GradientBoosting svmsmote Variance Threshold 0.750000
Description: svmsmote Model: gradient-boosting
              precision    recall  f1-score   support

           0       0.91      0.88      0.90      7963
           1       0.59      0.67      0.63      2037

    accuracy                           0.84     10000
   macro avg       0.75      0.78      0.76     10000
weighted avg       0.85      0.84      0.84     10000
```

```
------------------------------------------------------------------------
---
    GradientBoosting                      svmsmote    Variance Threshold 0.7
50000         0.8390000000         0.6302250804
------------------------------------------------------------------------
---
EVALUATING:  GradientBoosting svmsmote Variance Threshold 0.800000
Description: svmsmote Model: gradient-boosting
              precision    recall  f1-score   support

           0       0.91      0.88      0.90      7963
           1       0.59      0.67      0.63      2037

    accuracy                           0.84     10000
   macro avg       0.75      0.78      0.76     10000
weighted avg       0.85      0.84      0.84     10000

------------------------------------------------------------------------
---
    GradientBoosting                      svmsmote    Variance Threshold 0.8
00000         0.8390000000         0.6302250804
------------------------------------------------------------------------
---
EVALUATING:  GradientBoosting svmsmote Variance Threshold 0.850000
Description: svmsmote Model: gradient-boosting
              precision    recall  f1-score   support

           0       0.91      0.88      0.90      7963
           1       0.59      0.67      0.63      2037

    accuracy                           0.84     10000
   macro avg       0.75      0.78      0.76     10000
weighted avg       0.85      0.84      0.84     10000

------------------------------------------------------------------------
---
    GradientBoosting                      svmsmote    Variance Threshold 0.8
50000         0.8390000000         0.6302250804
------------------------------------------------------------------------
---
EVALUATING:  GradientBoosting svmsmote Variance Threshold 0.900000
Description: svmsmote Model: gradient-boosting
              precision    recall  f1-score   support

           0       0.91      0.88      0.90      7963
           1       0.59      0.67      0.63      2037

    accuracy                           0.84     10000
   macro avg       0.75      0.78      0.76     10000
weighted avg       0.85      0.84      0.84     10000

------------------------------------------------------------------------
---
    GradientBoosting                      svmsmote    Variance Threshold 0.9
00000         0.8390000000         0.6302250804
------------------------------------------------------------------------
---
```

```
EVALUATING:  GradientBoosting svmsmote Variance Threshold 0.950000
Description: svmsmote Model: gradient-boosting
              precision    recall  f1-score   support

           0       0.91      0.88      0.90      7963
           1       0.59      0.67      0.63      2037

    accuracy                           0.84     10000
   macro avg       0.75      0.78      0.76     10000
weighted avg       0.85      0.84      0.84     10000

------------------------------------------------------------------------------
---
    GradientBoosting                       svmsmote    Variance Threshold 0.9
50000         0.8390000000         0.6302250804
------------------------------------------------------------------------------
---
EVALUATING:  GradientBoosting svmsmote Variance Threshold 1.000000
Description: svmsmote Model: gradient-boosting
              precision    recall  f1-score   support

           0       0.91      0.88      0.90      7963
           1       0.59      0.67      0.63      2037

    accuracy                           0.84     10000
   macro avg       0.75      0.78      0.76     10000
weighted avg       0.85      0.84      0.84     10000

------------------------------------------------------------------------------
---
    GradientBoosting                       svmsmote    Variance Threshold 1.0
00000         0.8390000000         0.6302250804
    GradientBoosting                       svmsmote
NONE         0.8390000000         0.6302250804
    GradientBoosting                       svmsmote    Variance Threshold 0.0
00000         0.8390000000         0.6302250804
    GradientBoosting                       svmsmote    Variance Threshold 0.0
50000         0.8390000000         0.6302250804
    GradientBoosting                       svmsmote    Variance Threshold 0.1
00000         0.8390000000         0.6302250804
    GradientBoosting                       svmsmote    Variance Threshold 0.1
50000         0.8390000000         0.6302250804
    GradientBoosting                       svmsmote    Variance Threshold 0.2
00000         0.8390000000         0.6302250804
    GradientBoosting                       svmsmote    Variance Threshold 0.2
50000         0.8390000000         0.6302250804
    GradientBoosting                       svmsmote    Variance Threshold 0.3
00000         0.8390000000         0.6302250804
    GradientBoosting                       svmsmote    Variance Threshold 0.3
50000         0.8390000000         0.6302250804
    GradientBoosting                       svmsmote    Variance Threshold 0.4
00000         0.8390000000         0.6302250804
    GradientBoosting                       svmsmote    Variance Threshold 0.4
50000         0.8390000000         0.6302250804
    GradientBoosting                       svmsmote    Variance Threshold 0.5
00000         0.8390000000         0.6302250804
    GradientBoosting                       svmsmote    Variance Threshold 0.5
```

```
50000         0.8390000000         0.6302250804
    GradientBoosting                         svmsmote    Variance Threshold 0.6
00000         0.8390000000         0.6302250804
    GradientBoosting                         svmsmote    Variance Threshold 0.6
50000         0.8390000000         0.6302250804
    GradientBoosting                         svmsmote    Variance Threshold 0.7
00000         0.8390000000         0.6302250804
    GradientBoosting                         svmsmote    Variance Threshold 0.7
50000         0.8390000000         0.6302250804
    GradientBoosting                         svmsmote    Variance Threshold 0.8
00000         0.8390000000         0.6302250804
    GradientBoosting                         svmsmote    Variance Threshold 0.8
50000         0.8390000000         0.6302250804
    GradientBoosting                         svmsmote    Variance Threshold 0.9
00000         0.8390000000         0.6302250804
    GradientBoosting                         svmsmote    Variance Threshold 0.9
50000         0.8390000000         0.6302250804
    GradientBoosting                         svmsmote    Variance Threshold 1.0
00000         0.8390000000         0.6302250804
```

In [180]:

```python
from imblearn.ensemble import RUSBoostClassifier

def search_rus_boost(X, y):
    rusbc = RUSBoostClassifier(random_state=0)
    p = [('rusbc', rusbc,)]
    p = imblearn.pipeline.Pipeline(p)
    param_grid = {
        'rusbc__n_estimators': [1000, 10000],
        'rusbc__learning_rate': [1, 0.1, 0.01, 0.001],
        'rusbc__algorithm': ['SAMME', 'SAMME.R'],
    }
    scorer = make_scorer(metrics.f1_score)
    cv1 = StratifiedKFold(n_splits=6, random_state=10, shuffle=True)
    search = GridSearchCV(p, param_grid, n_jobs=7, cv=cv1, scoring=scorer)
    search.fit(X, y)
    print("GRADIENT BOOSTING")
    print(f"Best Params = {search.best_params_}")
    print(f"Best score = {search.best_score_}")
    print('-' * 120)
    return search.best_estimator_

rusboost = search_rus_boost(scaled_X.to_numpy(), y.to_numpy())
```

```
GRADIENT BOOSTING
Best Params = {'rusbc__algorithm': 'SAMME.R', 'rusbc__learning_rate': 0.001,
'rusbc__n_estimators': 10000}
Best score = 0.5837437094271175
------------------------------------------------------------------------------
------------------------------------------
```

RUSBOOST (Boost + UnderSampling at each stage)

In [181]:
```
from imblearn.ensemble import RUSBoostClassifier

def search_rus_boost(X, y):
    rusbc = RUSBoostClassifier(random_state=0)
    p = [('rusbc', rusbc,)]
    p = imblearn.pipeline.Pipeline(p)
    param_grid = {
        'rusbc__n_estimators': [10000, 100000],
        'rusbc__learning_rate': [0.001, 0.0001],
        'rusbc__algorithm': ['SAMME.R'],
    }
    scorer = make_scorer(metrics.f1_score)
    cv1 = StratifiedKFold(n_splits=6, random_state=10, shuffle=True)
    search = GridSearchCV(p, param_grid, n_jobs=7, cv=cv1, scoring=scorer)
    search.fit(X, y)
    print("GRADIENT BOOSTING")
    print(f"Best Params = {search.best_params_}")
    print(f"Best score = {search.best_score_}")
    print('-' * 120)
    return search.best_estimator_

rusboost = search_rus_boost(scaled_X.to_numpy(), y.to_numpy())
```

```
GRADIENT BOOSTING
Best Params = {'rusbc__algorithm': 'SAMME.R', 'rusbc__learning_rate': 0.001,
'rusbc__n_estimators': 100000}
Best score = 0.5847328429438647
------------------------------------------------------------------------------
----------------------------------------
```

In [ ]: