FeatureSelection

March 13, 2023

[4]:

```
%matplotlib inline

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

import warnings
warnings.filterwarnings('ignore')
pd.options.display.max_columns = None

from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import StratifiedKFold
from sklearn.feature_selection import RFECV

from numpy import mean
from numpy import std
from sklearn.datasets import make_classification
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import RepeatedKFold
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestRegressor
from sklearn.pipeline import Pipeline
from sklearn import preprocessing
from sklearn.preprocessing import MinMaxScaler
```

[78]:

```
#read dataset
data = pd.read_csv("/content/drive/MyDrive/CIND 820 Capstone Project/
 ↪merged_completedata.csv")
```

[79]:

```
#checking dimensions of data
data.head()
```

[79]:

```
   RecordID          X          Y  FID  BusinessID  \
0         1 -79.689829  43.644181    1        1055
1         2 -79.689419  43.644988    2        1057
2         3 -79.689419  43.644988    3        1058
3         4 -79.689419  43.644988    4        1060
```

```
4         5 -79.690664  43.645493    5        1061

                               Name            Address  StreetNo  \
0                  Golf Trends Inc.  300 Ambassador Dr       300   
1                Apex Graphics Inc.  320 Ambassador Dr       320   
2  Sands, John & Associates Limited  320 Ambassador Dr       320   
3       Printmedia-Tackaberry Times  320 Ambassador Dr       320   
4             S W R Industries Ltd.  321 Ambassador Dr       321   

      StreetName BldgNo UnitNo PostalCode           Location  Ward  NAICSCode  \
0  Ambassador Dr     No     No        L5T  Gateway EA (East)     5         41   
1  Ambassador Dr     No     No        L5T  Gateway EA (East)     5         32   
2  Ambassador Dr     No     No        L5T  Gateway EA (East)     5         32   
3  Ambassador Dr     No     No        L5T  Gateway EA (East)     5         32   
4  Ambassador Dr     No     No        L5T  Gateway EA (East)     5         41   

          NAICSCat                                         NAICSDescr  \
0  Wholesale Trade  Amusement and Sporting Goods Wholesaler-Distri…   
1    Manufacturing                    Support Activities for Printing   
2    Manufacturing                    Support Activities for Printing   
3    Manufacturing                                     Other Printing   
4  Wholesale Trade  Industrial Machinery, Equipment and Supplies W…   

          Phone           Fax TollFree EMail WebAddress  EmplRange  \
0  905-795-8900  905-795-8988      Yes   Yes        Yes          3   
1  905-795-9575  905-795-8775       No   Yes        Yes          4   
2  905-795-9519  905-795-8775       No    No         No          5   
3  905-564-8121  905-564-7395       No   Yes        Yes          1   
4  905-564-8080  905-564-5003       No   Yes        Yes          2   

        CENT_X        CENT_Y  Year isnew Closed  
0  605668.2538  4.833187e+06  2016    No     No  
1  605699.9370  4.833277e+06  2016    No     No  
2  605699.9370  4.833277e+06  2016    No     No  
3  605699.9370  4.833277e+06  2016    No     No  
4  605598.6442  4.833332e+06  2016    No     No  
```

[80]:

```
#decribe categorical data
data.describe(include='O')
```

[80]:

```
          Name             Address   StreetName BldgNo UnitNo PostalCode  \
count    78032               78032        78032  78032  78032      78032   
unique   22710                6618          669      2      2         37   
top     Subway  100 City Centre Dr  Dundas St E     No    Yes        L4W   
freq       212                 953         3202  73798  53665      12410   

                   Location      NAICSCat                     NAICSDescr  \
count                 78032        78032                           78032
unique                   56           19                            1039
top     Northeast EA (West)  Retail Trade  Limited-service eating places
freq                  21104        11071                            3647

        Phone    Fax TollFree  EMail WebAddress  isnew Closed
count   78032  78032    78032  78032      78032  78032  78032
unique  25064  15752        2      2          2      2      2
top                        No    Yes        Yes     No     No
freq     1457  29473    66596  47406      56765  71148  71617
```

[81]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 78032 entries, 0 to 78031
Data columns (total 28 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   RecordID    78032 non-null  int64
 1   X           78032 non-null  float64
 2   Y           78032 non-null  float64
 3   FID         78032 non-null  int64
 4   BusinessID  78032 non-null  int64
 5   Name        78032 non-null  object
 6   Address     78032 non-null  object
 7   StreetNo    78032 non-null  int64
 8   StreetName  78032 non-null  object
 9   BldgNo      78032 non-null  object
 10  UnitNo      78032 non-null  object
 11  PostalCode  78032 non-null  object
 12  Location    78032 non-null  object
 13  Ward        78032 non-null  int64
 14  NAICSCode   78032 non-null  int64
 15  NAICSCat    78032 non-null  object
 16  NAICSDescr  78032 non-null  object
 17  Phone       78032 non-null  object
 18  Fax         78032 non-null  object
 19  TollFree    78032 non-null  object
 20  EMail       78032 non-null  object
 21  WebAddress  78032 non-null  object
 22  EmplRange   78032 non-null  int64
 23  CENT_X      78032 non-null  float64
 24  CENT_Y      78032 non-null  float64
 25  Year        78032 non-null  int64
 26  isnew       78032 non-null  object
 27  Closed      78032 non-null  object
dtypes: float64(4), int64(8), object(16)
```

```
memory usage: 16.7+ MB
```

[82]:

```
#NAICSCode back to object
data['NAICSCode'] = data['NAICSCode'].astype(str)
```

[83]:

```
data.info()
```

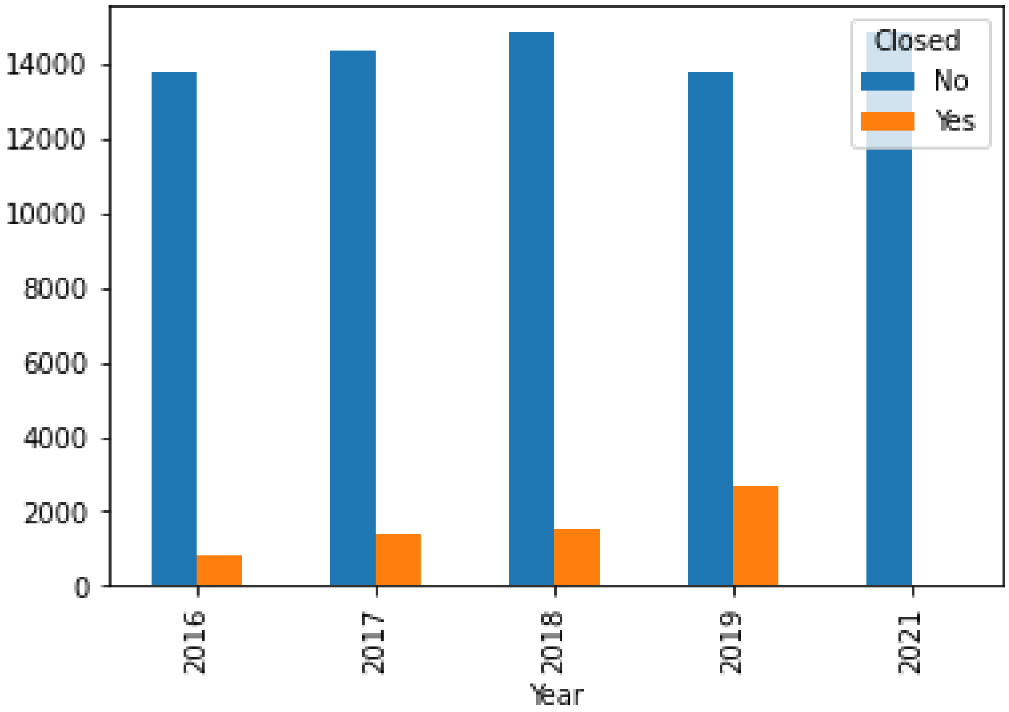
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 78032 entries, 0 to 78031
Data columns (total 28 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   RecordID     78032 non-null  int64
 1   X            78032 non-null  float64
 2   Y            78032 non-null  float64
 3   FID          78032 non-null  int64
 4   BusinessID   78032 non-null  int64
 5   Name         78032 non-null  object
 6   Address      78032 non-null  object
 7   StreetNo     78032 non-null  int64
 8   StreetName   78032 non-null  object
 9   BldgNo       78032 non-null  object
 10  UnitNo       78032 non-null  object
 11  PostalCode   78032 non-null  object
 12  Location     78032 non-null  object
 13  Ward         78032 non-null  int64
 14  NAICSCode    78032 non-null  object
 15  NAICSCat     78032 non-null  object
 16  NAICSDescr   78032 non-null  object
 17  Phone        78032 non-null  object
 18  Fax          78032 non-null  object
 19  TollFree     78032 non-null  object
 20  EMail        78032 non-null  object
 21  WebAddress   78032 non-null  object
 22  EmplRange    78032 non-null  int64
 23  CENT_X       78032 non-null  float64
 24  CENT_Y       78032 non-null  float64
 25  Year         78032 non-null  int64
 26  isnew        78032 non-null  object
 27  Closed       78032 non-null  object
dtypes: float64(4), int64(7), object(17)
memory usage: 16.7+ MB
```

[84]:

```
df_gb_openclosed = data.groupby(['Year', 'Closed']).size().unstack(level=1)
df_gb_openclosed.plot(kind = 'bar')
```

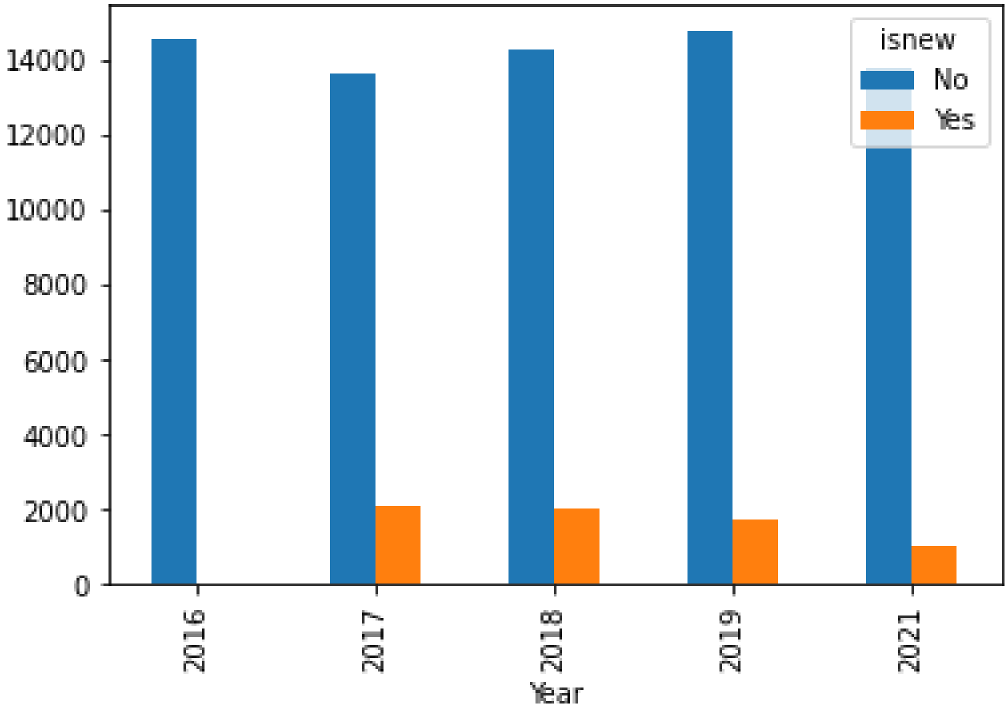
[84]:

```
<AxesSubplot:xlabel='Year'>
```



```
[11]: df_gb_openclosed = data.groupby(['Year', 'isnew']).size().unstack(level=1)
      df_gb_openclosed.plot(kind = 'bar')
```

```
[11]: <AxesSubplot:xlabel='Year'>
```



```
[85]: data.head()
```

```
[85]:    RecordID          X          Y  FID  BusinessID  \
      0         1 -79.689829  43.644181    1        1055   
      1         2 -79.689419  43.644988    2        1057   
      2         3 -79.689419  43.644988    3        1058   
      3         4 -79.689419  43.644988    4        1060   
      4         5 -79.690664  43.645493    5        1061   

                                     Name            Address  StreetNo  \
      0                 Golf Trends Inc.  300 Ambassador Dr       300   
      1               Apex Graphics Inc.  320 Ambassador Dr       320   
      2 Sands, John & Associates Limited  320 Ambassador Dr       320   
      3      Printmedia-Tackaberry Times  320 Ambassador Dr       320   
      4            S W R Industries Ltd.  321 Ambassador Dr       321   

            StreetName BldgNo UnitNo PostalCode           Location  Ward NAICSCode  \
      0  Ambassador Dr     No     No        L5T  Gateway EA (East)     5        41   
      1  Ambassador Dr     No     No        L5T  Gateway EA (East)     5        32   
      2  Ambassador Dr     No     No        L5T  Gateway EA (East)     5        32   
      3  Ambassador Dr     No     No        L5T  Gateway EA (East)     5        32   
      4  Ambassador Dr     No     No        L5T  Gateway EA (East)     5        41   
```

```
          NAICSCat                                         NAICSDescr  \
0  Wholesale Trade  Amusement and Sporting Goods Wholesaler-Distri…
1    Manufacturing                    Support Activities for Printing
2    Manufacturing                    Support Activities for Printing
3    Manufacturing                                     Other Printing
4  Wholesale Trade  Industrial Machinery, Equipment and Supplies W…

          Phone           Fax TollFree EMail WebAddress  EmplRange  \
0  905-795-8900  905-795-8988      Yes   Yes        Yes          3
1  905-795-9575  905-795-8775       No   Yes        Yes          4
2  905-795-9519  905-795-8775       No    No         No          5
3  905-564-8121  905-564-7395       No   Yes        Yes          1
4  905-564-8080  905-564-5003       No   Yes        Yes          2

        CENT_X        CENT_Y  Year isnew Closed
0  605668.2538  4.833187e+06  2016    No     No
1  605699.9370  4.833277e+06  2016    No     No
2  605699.9370  4.833277e+06  2016    No     No
3  605699.9370  4.833277e+06  2016    No     No
4  605598.6442  4.833332e+06  2016    No     No
```

[86]:

```python
#decribe categorical data
data.describe(include='O')
```

[86]:

```
          Name             Address   StreetName BldgNo UnitNo PostalCode  \
count    78032               78032        78032  78032  78032      78032
unique   22710                6618          669      2      2         37
top     Subway  100 City Centre Dr  Dundas St E     No    Yes        L4W
freq       212                 953         3202  73798  53665      12410

                   Location NAICSCode      NAICSCat  \
count                 78032     78032         78032
unique                   56        24            19
top     Northeast EA (West)        81  Retail Trade
freq                  21104      9052         11071

                           NAICSDescr  Phone    Fax TollFree  EMail  \
count                           78032  78032  78032    78032  78032
unique                           1039  25064  15752        2      2
top     Limited-service eating places                      No    Yes
freq                             3647   1457  29473    66596  47406

       WebAddress  isnew Closed
count       78032  78032  78032
unique          2      2      2
top           Yes     No     No
freq        56765  71148  71617
```

[87]:
```
#drop columns that have unique values and categorical
data.drop(['FID','BusinessID','RecordID', 'Name','StreetNo','Address',␣
 ↪'NAICSCat',␣
 ↪'StreetName','Location','Phone','Fax','NAICSDescr','EMail','PostalCode','BldgNo','UnitNo','
 ↪'NAICSCode'], axis=1, inplace=True)
```

[88]:
```
# Save the new data set to a new file
data.to_csv("/content/drive/MyDrive/CIND 820 Capstone Project/
 ↪categoricaltonumericdata.csv", index=False)
```

[89]:
```
data = data[data['Year'] == 2019]
```

[90]:
```
data.head()
```

[90]:
```
               X          Y  Ward  EmplRange       CENT_X        CENT_Y  Year  \
46689 -79.665386  43.684736     5          1  607567.2334  4.837723e+06  2019
46690 -79.642760  43.593515     4          2  609556.5032  4.827621e+06  2019
46691 -79.667311  43.682752     5          3  607415.6044  4.837500e+06  2019
46692 -79.629235  43.698932     4          2  610454.8654  4.839347e+06  2019
46693 -79.629235  43.698932     4          4  610454.8654  4.839347e+06  2019

      Closed
46689     No
46690     No
46691     No
46692     No
46693     No
```

[ ]:
```
#data = data[data['Closed'] == 0]
#use this for when taking 2021 and is new!!!
```

[63]:
```
data.head()
```

[63]:
```
               X          Y  StreetNo  Ward  EmplRange       CENT_X  \
46689 -79.665386  43.684736      7050     5          1  607567.2334
46690 -79.642760  43.593515       100     4          2  609556.5032
46691 -79.667311  43.682752      7015     5          3  607415.6044
46692 -79.629235  43.698932       100     4          2  610454.8654
46693 -79.629235  43.698932       100     4          4  610454.8654

             CENT_Y  Year Closed
46689  4.837723e+06  2019     No
46690  4.827621e+06  2019     No
46691  4.837500e+06  2019     No
46692  4.839347e+06  2019     No
46693  4.839347e+06  2019     No
```

[91]:

```
df2 = data.mean(axis=0)
print(df2)
```

```
X            -7.965769e+01
Y             4.360136e+01
Ward          5.372927e+00
EmplRange     2.183981e+00
CENT_X        6.088039e+05
CENT_Y        4.828662e+06
Year          2.019000e+03
dtype: float64
```

[92]:

```
correlated_features= set()
correlation_matrix = data.drop('Closed', axis=1).corr()

for i in range(len(correlation_matrix.columns)):
  for j in range(i):
    if abs(correlation_matrix.iloc[i,j]) > 0.8:
      colname = correlation_matrix.columns[i]
      correlated_features.add(colname)
```

[93]:

```
correlated_features
```

[93]:

```
{'CENT_X', 'CENT_Y'}
```

[94]:

```
data.drop(['CENT_X', 'CENT_Y'], axis=1, inplace=True)
#use for closed analysys
```

[95]:

```
#checking dimensions of data
data.head()
```

[95]:

```
               X          Y  Ward  EmplRange  Year Closed
46689 -79.665386  43.684736     5          1  2019     No
46690 -79.642760  43.593515     4          2  2019     No
46691 -79.667311  43.682752     5          3  2019     No
46692 -79.629235  43.698932     4          2  2019     No
46693 -79.629235  43.698932     4          4  2019     No
```

[96]:

```
data.dtypes
```

[96]:

```
X             float64
Y             float64
Ward            int64
EmplRange       int64
Year            int64
Closed         object
dtype: object
```

[97]:
```
data['EmplRange'].isnull().values.any()
```

[97]:
```
False
```

[98]:
```
data.isnull().sum().sum()
```

[98]:
```
0
```

[116]:
```
X = data.drop('Closed', axis=1)
target = data['Closed']

rfc = RandomForestClassifier(random_state=101)
rfecv = RFECV(estimator=rfc, step=1, cv=StratifiedKFold(10), scoring='accuracy')
rfecv.fit(X, target)
#Recursive feature elimination
#Takes around 2-3 minutes to run. Not as effecient for feature selection.
```

[116]:
```
RFECV(cv=StratifiedKFold(n_splits=10, random_state=None, shuffle=False),
      estimator=RandomForestClassifier(random_state=101), scoring='accuracy')
```

[100]:
```
print('Optimal number of features: {}'.format(rfecv.n_features_))
```

```
Optimal number of features: 3
```

[101]:
```
print(np.where(rfecv.support_ == False)[0])

X.drop(X.columns[np.where(rfecv.support_ == False)[0]], axis=1, inplace=True)
```

```
[2 4]
```

[102]:
```
rfecv.estimator_.feature_importances_
```

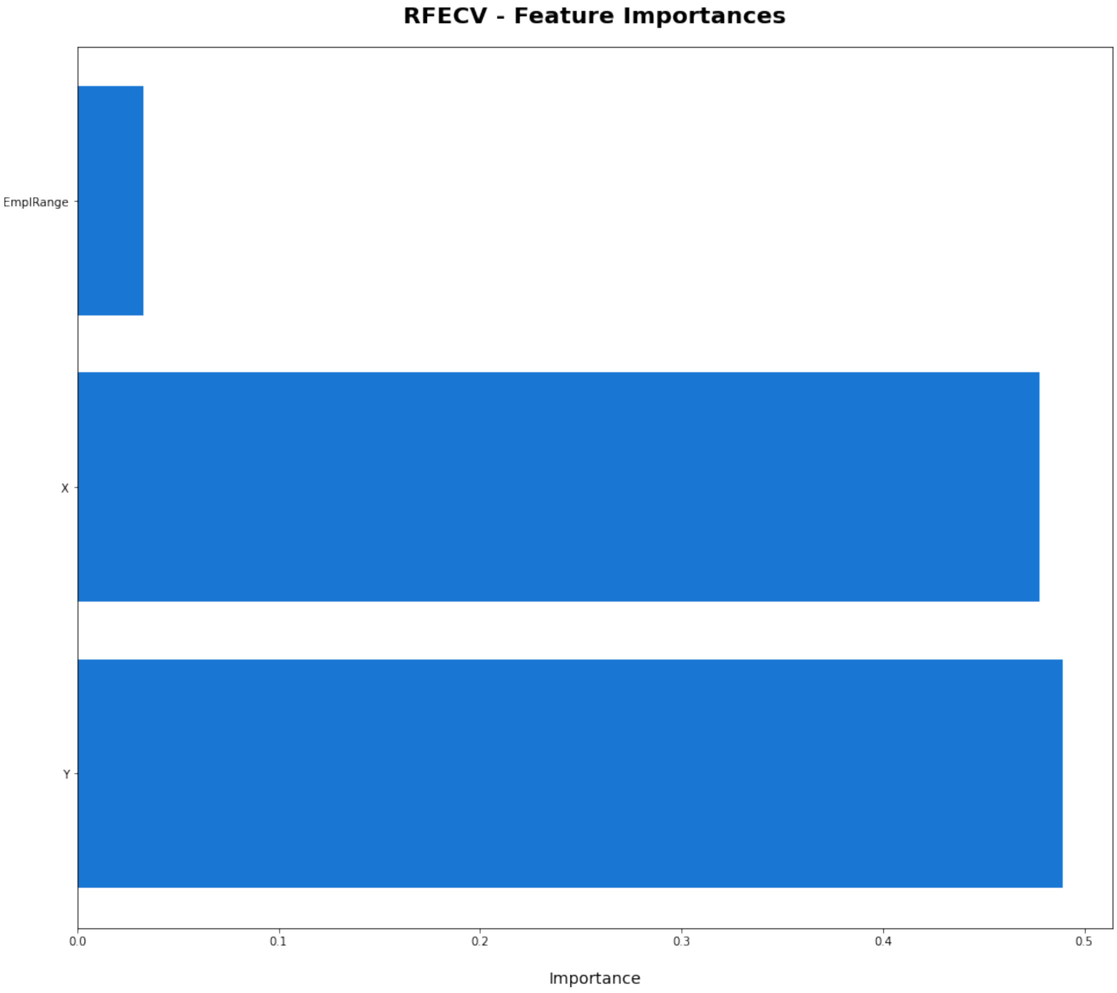
[102]:
```
array([0.47756572, 0.48930314, 0.03313114])
```

[115]:
```
dset = pd.DataFrame()
dset['attr'] = X.columns
dset['importance'] = rfecv.estimator_.feature_importances_

dset = dset.sort_values(by='importance', ascending=False)


plt.figure(figsize=(16, 14))
plt.barh(y=dset['attr'], width=dset['importance'], color='#1976D2')
plt.title('RFECV - Feature Importances', fontsize=20, fontweight='bold', pad=20)
plt.xlabel('Importance', fontsize=14, labelpad=20)
#plt.show()
```

```
[115]: Text(0.5, 0, 'Importance')
```



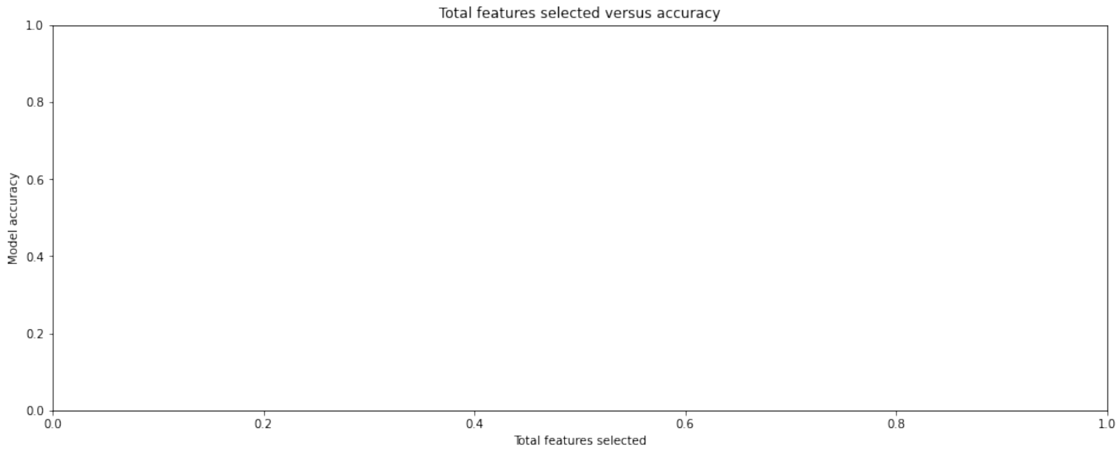
```
[ ]: plt.figure(figsize=(16, 9))
     plt.title('Recursive Feature Elimination with Cross-Validation', fontsize=18,␣
      ↪fontweight='bold', pad=20)
     plt.xlabel('Number of features selected', fontsize=14, labelpad=20)
     plt.ylabel('% Correct Classification', fontsize=14, labelpad=20)
     plt.plot(range(1, len(rfecv.grid_scores_) + 1), rfecv.grid_scores_,␣
      ↪color='#303F9F', linewidth=3)

     plt.show()
```

```
[110]: plt.figure( figsize=(16, 6))
       plt.title('Total features selected versus accuracy')
       plt.xlabel('Total features selected')
       plt.ylabel('Model accuracy')
       plt.plot(range(1, len(rfecv.grid_scores_) + 1), rfecv.grid_scores_)
       plt.show()
```

```
---------------------------------------------------------------------------
AttributeError                            Traceback (most recent call last)
<ipython-input-110-31a34b31d8b7> in <module>
      3 plt.xlabel('Total features selected')
      4 plt.ylabel('Model accuracy')
----> 5 plt.plot(range(1, len(rfecv.grid_scores_) + 1), rfecv.grid_scores_)
      6 plt.show()

AttributeError: 'RFECV' object has no attribute 'grid_scores_'
```



```
[111]: df_features = pd.DataFrame(columns = ['feature', 'support', 'ranking'])

       for i in range(X.shape[1]):
           row = {'feature': i, 'support': rfecv.support_[i], 'ranking': rfecv.
        ↪ranking_[i]}
           df_features = df_features.append(row, ignore_index=True)

       df_features.sort_values(by='ranking').head(10)
```

```
[111]:   feature support ranking
       0       0    True       1
       1       1    True       1
       2       2   False       2
```

```
[112]: df_features[df_features['support']==True]
```

```
[112]:   feature support ranking
       0       0    True       1
       1       1    True       1
```

```
[114]: selected_features = rfecv.get_support(1)
       X = data[data.columns[selected_features]]
```