ECON 425 HW8

February 29, 2024

Due Thu, Mar 7, 6pm in Bruinlearn

Problem 1 Use the dataset 'marketing-campaign.csv' uploaded on Canvas (description can be found here: https://archive.ics.uci.edu/dataset/222/bank+marketing). Convert categorical variables of you choice (you need not use all of them) into dummies and allocate a third of your data to the testing sample.

(i) First, suppose each tree in a random forest picks a random subset of $m = \sqrt{p}$ features at each split, where p is the number of features in the data. On the training sample, fit the random forest to predict subscription to a term deposit. Vary the number of trees in the range $\{1, 2, 3, 4, 5, 10, 20, 50\}$. Plot accuracy, precision, recall, and F1 score on the testing sample against the number of trees.

(ii) Repeat (i) with *bagging*, i.e. $m = p$.

(iii) Pick the best-performing model and use the feature_importances_ attribute of RandomForestClassifier to evaluate importance of different features. Is there a clearly dominating feature? Explain.

(iv) Beyond sampling variation, is there any other explanation for the alternating pattern in some performance metrics arising when the number of trees is very small?

[1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.metrics import accuracy_score, precision_score, recall_score,␣
 ↪f1_score
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split
import warnings
warnings.filterwarnings('ignore')
```

[2]:
```
df=pd.read_csv('/Users/sandinatatu/Desktop/marketing_campaign.csv',sep=';')
catcols=['job','marital','education','default','housing','loan','contact','day','month','pouto

df=pd.get_dummies(df, columns=catcols)
```

[3]:
```
X=df.drop('y', axis=1).copy()
y=df[['y']]

X_train, X_test, y_train, y_test = train_test_split(X,y, test_size=0.33,␣
 ↪random_state=0)

n_trees=[1,2,3,4,5,10,20,50]
n_features=[int(np.sqrt(X.shape[1])), X.shape[1]]
```

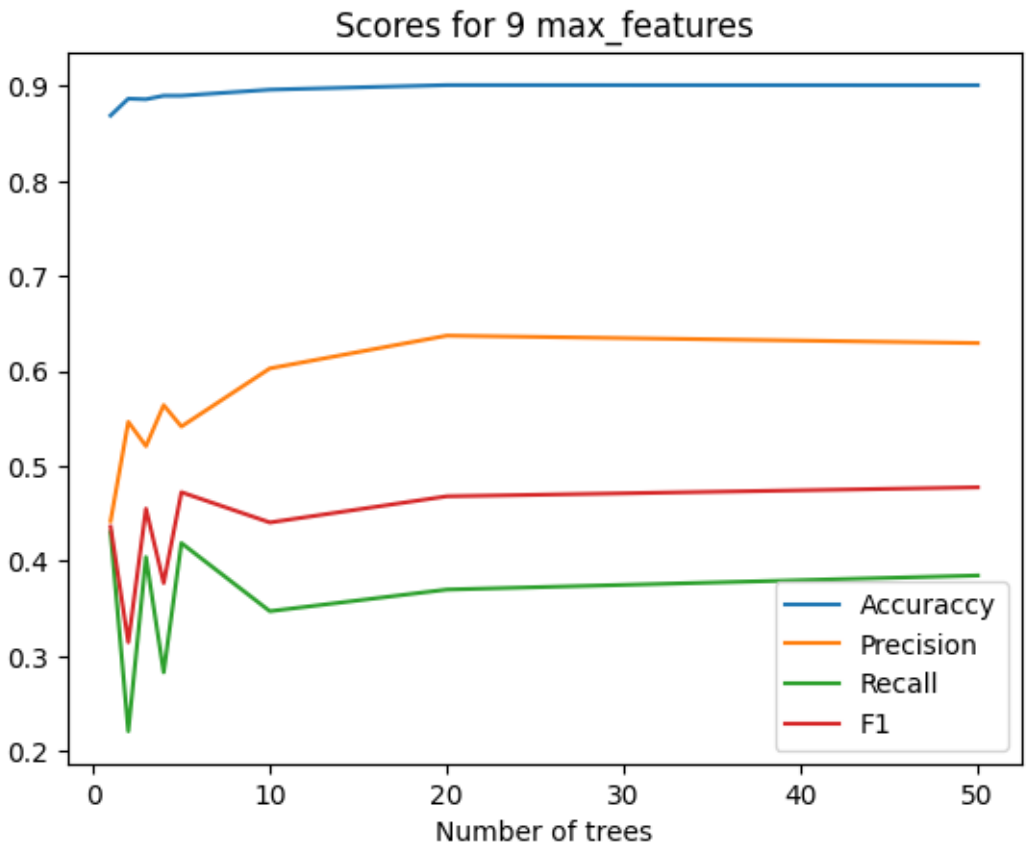
[4]:
```
for feature in n_features:
    print(f"We are at max_features {feature}")
    accuracies, precisions, recalls, f1s=[],[],[],[]

    for tree in n_trees:
        print(f"We are considering {tree} trees")

        rf=RandomForestClassifier(max_features=feature, n_estimators=tree,␣
 ↪random_state=0).fit(X_train, y_train)

        y_pred=rf.predict(X_test)

        accuracies.append(accuracy_score(y_test, y_pred))
```

```
        precisions.append(precision_score(y_test, y_pred, pos_label='yes'))
        recalls.append(recall_score(y_test, y_pred, pos_label='yes'))
        f1s.append(f1_score(y_test, y_pred, pos_label='yes'))

        frame=pd.DataFrame({'Importance':rf.feature_importances_}, index=list(X.
↪columns))
        print(frame.sort_values(by='Importance', ascending=False))

    plt.plot(n_trees, accuracies)
    plt.plot(n_trees, precisions)
    plt.plot(n_trees, recalls)
    plt.plot(n_trees, f1s)
    plt.xlabel("Number of trees")
    plt.legend(['Accuraccy','Precision','Recall','F1'])
    plt.title(f"Scores for {feature} max_features")
    plt.show()
```

```
We are at max_features 9
We are considering 1 trees
                  Importance
duration            0.264673
poutcome_success    0.095791
balance             0.072229
age                 0.069830
campaign            0.035303
…                        …
day_27              0.001931
default_no          0.001928
default_yes         0.000852
job_unknown         0.000729
day_31              0.000573

[81 rows x 1 columns]
We are considering 2 trees
                  Importance
duration            0.254467
balance             0.078974
age                 0.073560
poutcome_success    0.058036
campaign            0.037304
…                        …
day_24              0.001920
default_no          0.001252
default_yes         0.001054
day_31              0.000797
job_unknown         0.000694

[81 rows x 1 columns]
```

```
We are considering 3 trees
                  Importance
duration            0.251260
balance             0.081173
age                 0.076500
poutcome_success    0.069186
campaign            0.035508
…                        …
job_housemaid       0.002216
default_no          0.001243
default_yes         0.001120
day_31              0.000885
job_unknown         0.000855

[81 rows x 1 columns]
We are considering 4 trees
                  Importance
duration            0.255956
balance             0.079927
age                 0.074760
poutcome_success    0.058347
pdays               0.039236
…                        …
job_housemaid       0.002215
default_no          0.001323
default_yes         0.001158
day_31              0.000911
job_unknown         0.000805

[81 rows x 1 columns]
We are considering 5 trees
                  Importance
duration            0.256133
balance             0.080618
age                 0.074960
poutcome_success    0.050996
pdays               0.036250
…                        …
job_housemaid       0.002342
default_no          0.001380
default_yes         0.001059
job_unknown         0.000912
day_31              0.000880

[81 rows x 1 columns]
We are considering 10 trees
                  Importance
duration            0.250900
```

```
balance             0.081934
age                 0.079230
poutcome_success    0.044083
pdays               0.040527
…                      …
job_housemaid       0.002504
default_no          0.001468
default_yes         0.001119
day_31              0.001019
job_unknown         0.000887

[81 rows x 1 columns]
We are considering 20 trees
                  Importance
duration            0.252402
balance             0.082086
age                 0.080178
pdays               0.041009
poutcome_success    0.040894
…                      …
job_housemaid       0.002692
default_no          0.001223
default_yes         0.001156
day_31              0.001078
job_unknown         0.000916

[81 rows x 1 columns]
We are considering 50 trees
                  Importance
duration            0.250921
balance             0.082004
age                 0.081320
poutcome_success    0.041847
pdays               0.037019
…                      …
day_24              0.002518
day_31              0.001311
default_yes         0.001098
job_unknown         0.001067
default_no          0.001061

[81 rows x 1 columns]
```

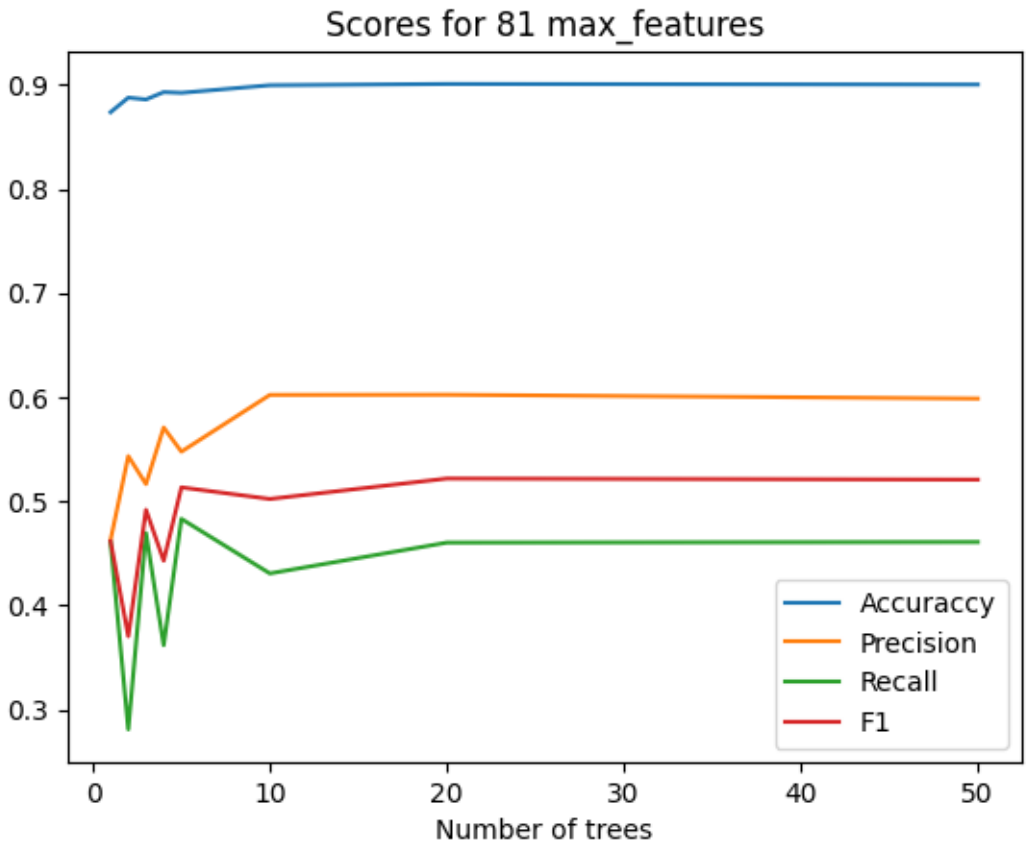



```
We are at max_features 81
We are considering 1 trees
                  Importance
duration            0.274813
balance             0.093819
poutcome_success    0.091346
age                 0.075659
pdays               0.031564
…                        …
default_yes         0.000734
day_31              0.000580
default_no          0.000554
job_unknown         0.000450
poutcome_unknown    0.000446

[81 rows x 1 columns]
We are considering 2 trees
                  Importance
duration            0.267694
poutcome_success    0.094011
```

```
balance             0.087792
age                 0.076685
pdays               0.036679
…                      …
poutcome_unknown    0.001317
job_unknown         0.001060
day_31              0.000763
default_yes         0.000676
default_no          0.000277

[81 rows x 1 columns]
We are considering 3 trees
                  Importance
duration            0.268442
poutcome_success    0.095686
balance             0.083181
age                 0.079971
pdays               0.037954
…                      …
poutcome_unknown    0.001195
default_yes         0.001184
job_unknown         0.000712
day_31              0.000696
default_no          0.000368

[81 rows x 1 columns]
We are considering 4 trees
                  Importance
duration            0.268039
poutcome_success    0.095112
balance             0.083130
age                 0.081234
pdays               0.036710
…                      …
poutcome_unknown    0.001614
default_yes         0.001163
job_unknown         0.000788
default_no          0.000607
day_31              0.000598

[81 rows x 1 columns]
We are considering 5 trees
                  Importance
duration            0.268928
poutcome_success    0.096801
age                 0.082590
balance             0.082309
pdays               0.038978
```

```
…                        …
poutcome_unknown    0.001722
default_yes         0.001162
default_no          0.000833
job_unknown         0.000790
day_31              0.000735

[81 rows x 1 columns]
We are considering 10 trees
                  Importance
duration            0.266941
poutcome_success    0.096939
balance             0.083348
age                 0.080388
pdays               0.038788
…                        …
poutcome_unknown    0.001859
default_yes         0.001322
day_31              0.001057
job_unknown         0.000735
default_no          0.000700

[81 rows x 1 columns]
We are considering 20 trees
                  Importance
duration            0.267467
poutcome_success    0.095287
balance             0.085369
age                 0.079069
pdays               0.038033
…                        …
poutcome_unknown    0.001868
day_31              0.001313
default_yes         0.001220
job_unknown         0.000928
default_no          0.000666

[81 rows x 1 columns]
We are considering 50 trees
                  Importance
duration            0.267728
poutcome_success    0.094582
balance             0.085523
age                 0.077207
pdays               0.038924
…                        …
poutcome_unknown    0.001971
day_31              0.001460
```

```
default_yes          0.000996
job_unknown          0.000884
default_no           0.000841

[81 rows x 1 columns]
```



[ ]: