Decision Tree

1. implement a decision tree and compute its accuracy for year 2

GME: 0.8846153846153846
SPY: 0.9423076923076923

2. compute the confusion matrix for year 2

GME:[[30, 1],[2, 19]]
SPY:[[30, 1],[2, 19]]

3. 3. what is true positive rate and true negative rate for year 2?

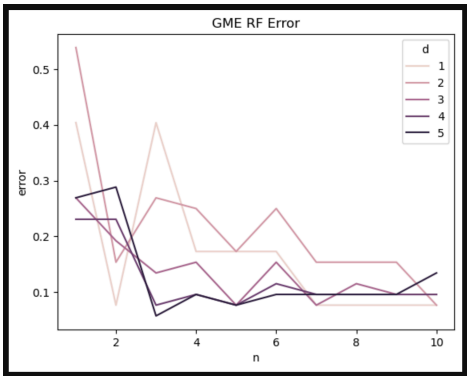
GME:(0.85, 0.94)
SPY: .9 and .97

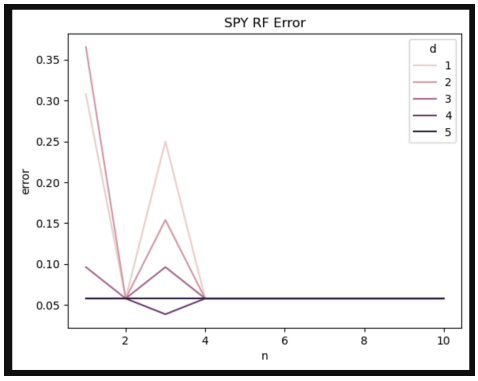
4. 4. implement a trading strategy based on your labels for year 2 and compare the performance with the ”buy-and-hold” strategy. Which strategy results in a larger amount at the end of the year?

GME: For gme, b&h yields 807.43 while trading based on decision tree labels yields 41867.41
SPY:164.84 & 127.54. trading labels is better

Random Forest

1. take N = 1, . . . , 10 and d = 1, 2, . . . , 5. For each value of N and d construct a random tree classifier (use ”entropy” as splitting criteria - this is the default) use your year 1 labels as training set and compute the error rate for year 2. Plot your error rates and find the best combination of N and d.





2. using the optimal values from year 1, compute the confusion matrix for year 2

GME: [[30, 1],[6, 15]]
SPY: [[30, 1], [1, 20]]

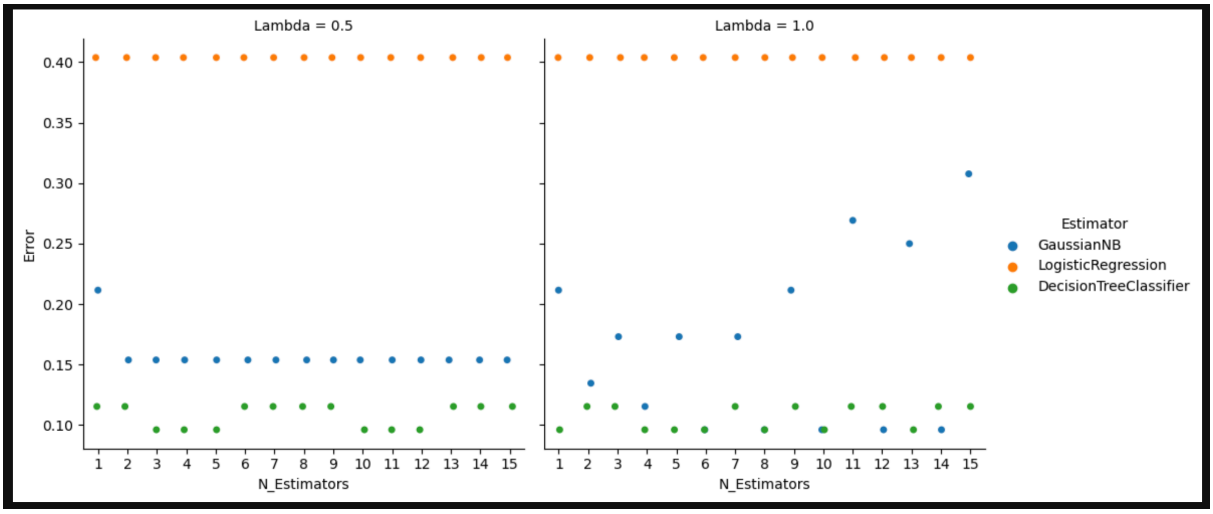
3. what is true positive rate and true negative rate for year 2?

GME: tpr: 0.71, tnr: 0.97
SPY: tpr: 0.95, tnr: 0.97

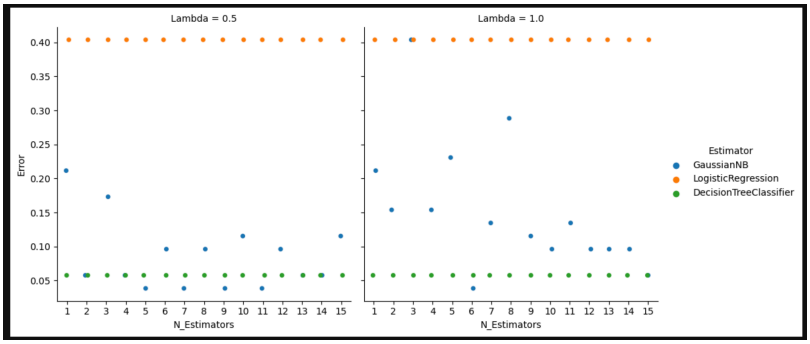
4. implement a trading strategy based on your labels for year 2 and compare the performance with the ”buy-and-hold” strategy. Which strategy results in a larger amount at the end of the year?
GME:labels:54072.1 b&h: 807.43
SPY:labs: 163.44, bh:127.54. labs better

Adaboost:

1) GME:



SPY:



2)
GME:

lambda_pt_5

	Lambda	Estimator	N_Estimators	Accuracy	Error
32	0.5	DecisionTreeClassifier	3	0.903846	0.096154
1	0.5	GaussianNB	2	0.846154	0.153846
15	0.5	LogisticRegression	1	0.596154	0.403846

SPY:

	Lambda	Estimator	N_Estimators	Accuracy	Error
30	0.5	DecisionTreeClassifier	1	0.942308	0.057692
4	0.5	GaussianNB	5	0.961538	0.038462
15	0.5	LogisticRegression	1	0.596154	0.403846

3) GME: .904. SPY: .962

4) GME: DecisionTree. SPY: GaussianNB

5) GME labels: 45250.34 bh:807.43. ; labs better.
SPY: labs 165.06, b&h: 127.54; labs better

Tips:

1. what is the average tip (as a percentage of meal cost) for lunch and for dinner?

```
time
Dinner    0.159518
Lunch     0.164128
```

2. what is average tip for each day of the week (as a percentage of meal cost)?

```
day
Fri     0.169913
Sat     0.153152
Sun     0.166897
Thur    0.161276
```

3. when are tips highest (which day and time)?

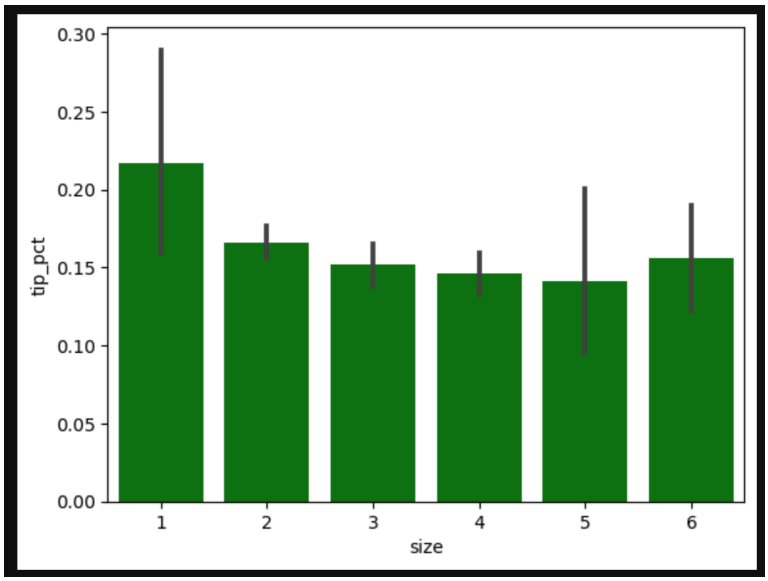
```
day   time
Fri   Dinner    0.158916
      Lunch     0.188765
Sat   Dinner    0.153152
Sun   Dinner    0.166897
Thur  Dinner    0.159744
      Lunch     0.161301
```

Friday lunch , tip pct 18.88%

4. compute the correlation between meal prices and tips

-.339

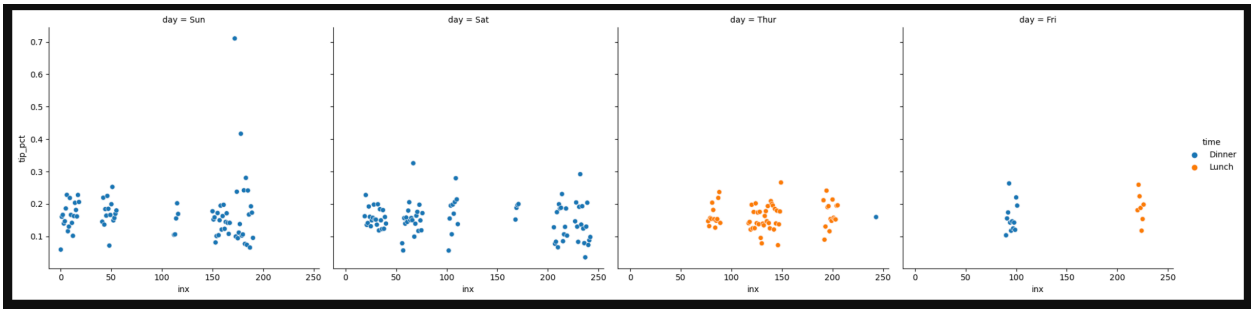
5. is there any relationship between tips and size of the group?



Group size of 1 seems to have larger tips.

6. what percentage of people are smoking?
38.11% of people are smokers

7. assume that rows in the tips.csv file are arranged in time. Are tips increasing with time in each day?



After sorting by day, I don't see an association between index and tip pct

8. is there any difference in correlation between tip amounts from smokers and non-smokers?
Non-smoker

total_bill	-0.199772
tip	0.349519
size	-0.121354
tip_pct	1.000000

smoker

total_bill	-0.457351
tip	0.377667
size	-0.191993
tip_pct	1.000000

Smokers have a more negative correlation between total_bill and tip_pct.