Yubin Ye U53651145

Question 1

1. For each file, read them into a Pandas frame and add a column ‘True Label’.
2. compute the default probability $p*$ that the next day is a ‘up’ day.
SPY: P= 0.535099
Y: P= 0.525828

3. take years 1, 2 and 3 What is the probability that after seeing k consecutive ‘down days’, the next day is an ‘up day’? For example, if k = 3, what is the probability of seeing ‘−, −, −, +’ as opposed to seeing ‘−, −, −, −’. Compute this for k = 1,2,3.

SPY:
K=1, P= 0.591429 K=2, P= 0.643357 K=3, P= 0.627451
Y:
K= 1, P= 0.560224 K=2, P= 0.573248 K=3, P= 0.537313

4. take years 1, 2 and 3. What is the probability that after seeing k consecutive ‘up days’, the next day is still an ‘up day’? For example, if k = 3, what is the probability of seeing ‘+, +, +, +’ as opposed to seeing ‘+, +, +, −’? Compute this for k = 1, 2, 3.

SPY:
K= 1, P= 0.485149 K= 2, P= 0.469388 K= 3, P= 0.467391
Y:
K= 1, P= 0.493703 K=2, P= 0.545918 K=3, P= 0.542056

Question 2

1. for W = 2, 3, 4, compute predicted labels for each day in year 4 and 5 based on true labels in years 1,2 and 3 only. Perform this for your ticker and for ‘spy’.
2. for each W = 2, 3, 4, compute the accuracy - what percent- age of true labels (both positive and negative) have you predicted correctly for the last two years.

SPY:
W=2 'all': 0.500998; '+': 0.549549; '-': 0.404762
W=3 'all': 0.466, '+': 0.527473; '-': 0.392070
W=4 'all': 0.470942; '+': 0.537118; '-': 0.414815

Y:
W=2 'all': 0.512974; '+': 0.531335; '-': 0.462686
W=3 'all': 0.514, '+': 0.531335; '-': 0.466165
W=4 'all': 0.513026; '+': 0.538983; '-': 0.475490

3. which W ∗ value gave you the highest accuracy for your stock and which W∗ value gave you the highest accuracy for S&P-500?
For SPY, when W=2, the accuracy is highest
For Y, when W=3, the accuracy is highest

Question 3

1. compute ensemble labels for year 4 and 5 for both your stock and S&P-500.
2. for both S&P-500 and your ticker, what percentage of labels in year 4 and 5 do you compute correctly by using ensemble?
SPY: W=ensemble 'all': 0.466934; '+': 0.527473; '-': 0.393805
Y: W=ensemble 'all': 0.515030; '+': 0.532786; '-': 0.466165

3. did you improve your accuracy on predicting '−' labels by using ensemble compared to W = 2, 3, 4?
For SPY, using ensemble not improved compare to using w=2 or 4 but improved a little compare to using w=3.
W=2 '-': 0.404762
W=3 '-': 0.392070
W=4 '-': 0.414815
ensemble '-': 0.393805

For Y, using ensemble not improved compare to using w=2,3 or 4
W=2 '-': 0.462686
W=3 '-': 0.466165
W=4 '-': 0.475490
ensemble '-': 0.466165

4. did you improve your accuracy on predicting '+' labels by using ensemble compared to W = 2, 3, 4?
For SPY, using ensemble not improved compare to using w=2,3 or 4
W=2 '+': 0.549549
W=3 '+': 0.527473
W=4 '+': 0.537118
W=eb. '+': 0.527473

For Y, using ensemble not improved compare to using w= 4 but improved a little compare to using w=2 or 3.
W=2 '+': 0.531335
W=3 '+': 0.531335
W=4 '+': 0.538983
W=eb '+': 0.532786

Question 4

For W = 2, 3, 4 and ensemble, compute the following (both for your ticker and ‘spy’) statistics based on years 4 and 5:
1. TP - true positives (your predicted label is + and true label is +
2. FP - false positives (your predicted label is + but true label is −
3. TN - true negatives (your predicted label is − and true label is −
4. FN - false negatives (your predicted label is − but true label is +
5. TPR = TP/(TP + FN) - true positive rate. This is the fraction of positive labels that your predicted correctly. This is also called sensitivity, recall or hit rate.
6. TNR = TN/(TN + FP) - true negative rate. This is the fraction of negative labels that your predicted correctly. This is also called specificity or selectivity.

7. summarize your findings in the table as shown below:

SPY	TP	FP	TN	FN	accuracy	TPR	TNR
w=2	183	150	68	100	0.500998	0.646643	0.311927
w=3	144	129	89	138	0.466	0.510638	0.408257
w=4	123	106	112	158	0.470942	0.437722	0.513761
ensemble	144	129	89	137	0.466934	0.512456	0.408257

Y	TP	FP	TN	FN	accuracy	TPR	TNR
w=2	195	172	62	72	0.512974	0.730337	0.264957
w=3	195	172	62	71	0.514	0.733083	0.264957
w=4	159	136	97	107	0.513026	0.597744	0.416309
ensemble	195	171	62	71	0.51503	0.733083	0.266094

8. discuss your findings

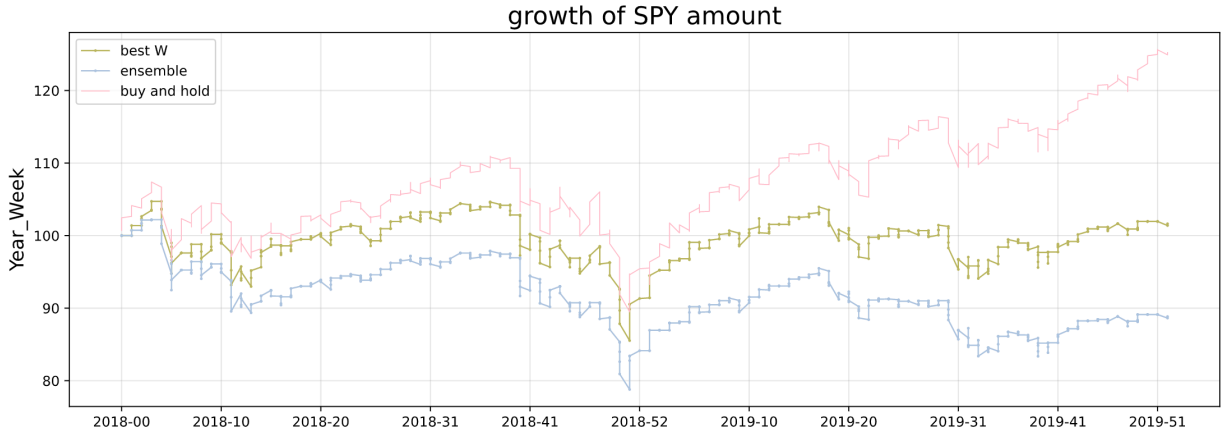
For SPY, as w increase from 2 to 4, the accuracy decrease, as well as TPR, but TNR increase. The ensemble prediction cannot exactly increase accuracy.

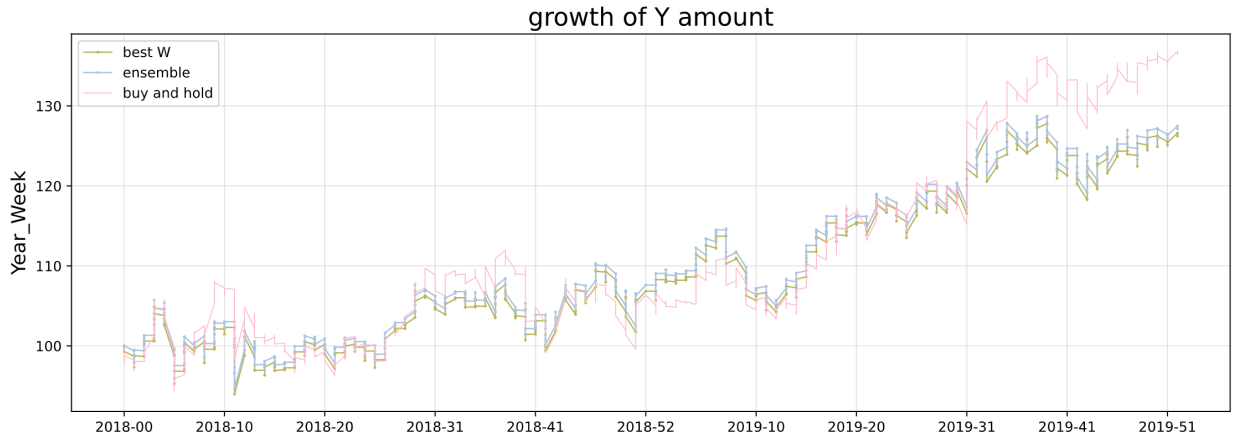
For Y, as w increase from 2 to 4, the accuracy increase, but TPR and TNR are not exactly increase.
The ensemble prediction can increase accuracy, but it cannot increase TNR.

Question 5

At the beginning of year 4 you start with $100 dollars and trade for 2 years based on predicted labels.

1. take your stock. Plot the growth of your amount for 2 years if you trade based on best W∗ and on ensemble. On the same graph, plot the growth of your portfolio for ‘buy-and- hold’ strategy





2. examine your chart. Any patterns? (e.g any differences in year 4 and year 5)

For SPY, the three lines have similar tendency, they drop down at the end of 2018 and rise since the beginning of 2019. The amount of ensemble is smaller than best w, and the buy and hold strategy have highest amount.

For Y, best w and ensemble have similar tendency and amount. While the buy and hold in this graph is different, sometimes it is higher than best w and ensemble, or sometimes it is lower than the two line.

If look at the two graph together, it is obviously that best w and ensemble, the predicted strategy, is more stable than the buy and hold strategy, and at least in the two graph, buy and hold earns more.