Assignment

In many data science applications, you want to identify patterns, labels or classes based on available data. In this assignment we will focus on discovering patterns in your past stock behavior.

To each trading day i you will assign a "trading" label "+" or "−", depending whether the corresponding daily return for that day $r_i \geq 0$ or $r_i < 0$. We will call these "true" labels and we compute these for all days in all 5 years.

We will use years 1,2 ans 3 as training years and we will use years 4 and 5 as testing years. For each day in years 4 and 5 we will predict a label based on some patterns that we observe in training years. We will call these "predicted" labels. We know the "true" labels for years 4 and 5 and we compute "predicted" labels for years 4 and 5. Therefore, we can analyze how good are our predictions for all labels, "+" labels only and "-" labels only in years 4 and 5.

Question 1: You have a csv table of daily returns for your stosk and for S&P-500 ("spy" ticker).

1. For each file, read them into a pandas frame and add a

column "  In that column, for each day (row) i with d≥ 0 you assign a "$+$" label ("up day"). F with daily return $r_i < 0$ you assign "$-$" (" You do this for every day for all 5 years bo.

For example, if your initial dataframe were

Date	...	Return
1/2/2015	...	0.015
1/3/2015	...	-0.01
1/6/2015	...	0.02
...	...	...
...	...	...
12/30/2019	...	0
12/31/2019	...	-0.03

Table 1: Initial data

you will add an additional column "True Label" and have data as shown in Table 2.

Your daily "true labels" sequence is $+, -, +, \cdots +, -$.

2. take years 1,2 and 3. Let L be the number of trading days. Assuming 250 trading days per year, L will contain about 750 days. Let L^- be all trading days with $-$ labels and let L^+ be all trading days with $+$ labels. Assuming that

	...	Return	True Label
	...	0.015	+
	...	-0.01	−
	...	0.02	+
	...	...	...
	...	...	...
12/30/2019	...	0	+
12/31/2019	...	-0.03	−

Table 2: Adding True Labels

all days are independent of each other and that the ratio of "up" and "down" days remains the same in the future, compute the default probability p^* that the next day is a "up" day.

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 $k = 1, 2, 3$.

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 $k = 1, 2, 3$.

Predicting We will now describe a procedure to predict labels in years 4 and 5 from "true" labels in training years.

For each day in years 4 and 5, we look at the pattern of last W true labels (including this day d). By looking at the frequency of this pattern and true label for the next day in the training set, we will predict label for day $d+1$. Here W is the **hyperparameter** that we will choose based on our prediction accuracy.

Suppose $W = 3$. You look at a particular day d and suppose that the sequence of last W labels is $s =$ "$-, +, -$". We want to predict the label for next day $d+1$. To do this, we count the number of sequences of length $W+1$ in the training set where the first W labels coincide with s. In other words, we count the number $N^-(s)$ of sequences "$s, -$" and the number of sequences $N^+(s)$ of sequences "$s, +$". If $N^+(s) \geq N^-(s)$ then the next day is assigned "$+$". If $N^+(s) < N^-(s)$ then the next day is assigned "$-$". In the unlikely event that $N^+(s) = N^-(s) = 0$ we will assign a label based on default probability p^* that we computed in the previous question.

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 [redacted] mpute the accuracy - what percentage of t[redacted]h positive and negative) have you predicte[redacted]he last two years.

3. which W [redacted] you the highest accuracy for your stock and and which W^* valuegave you the highest accuracy for S&P-500?

Question 3. One of the most powerful methods to (potentially) improve predictions is to combine predictions by some "averaging". This is called *ensemble learning*. Let us consider the following procedure: for every day d, you have 3 predicted labels: for $W = 2$, $W = 3$ and $W = 4$. Let us compute an "ensemble" label for day d by taking the majority of your labels for that day. For example if your predicted labels were "−","−" and "+", then we would take "−" as ensemble label for day d (the majority of three labels is "−"). If, on the other hand, your predicted labels were "−", "+" and "+" then we would take "+" as ensemble label for day d (the majority of predicted labels is "+"). Compute such ensemble labels and answer the following:

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?

3. did you improve accuracy on predicting "−" labels by using ensembles compared to $W = 2, 3, 4$?
4. did you improve accuracy on predicting "+" labels by using ensembles compared to $W = 2, 3, 4$?

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. summari[illegible]s in the table as shown below:

W	t[illegible]	[illegible]	FP	TN	FN	accuracy	TPR	TNR
2	S[illegible]							
3	S[illegible]							
4	S[illegible]							
ensemble	S&P-500							
2	your stock							
3	your stock							
4	your stock							
ensemble	your stock							

Table 3: Prediction Results for $W = 1, 2, 3$ and ensemble

8. discuss your findings

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)