EXERCISE 1

1. For each of parts (a) through (d), indicate whether we would generally expect the performance of a flexible statistical learning method to be better or worse than an inflexible method. Justify your answer.

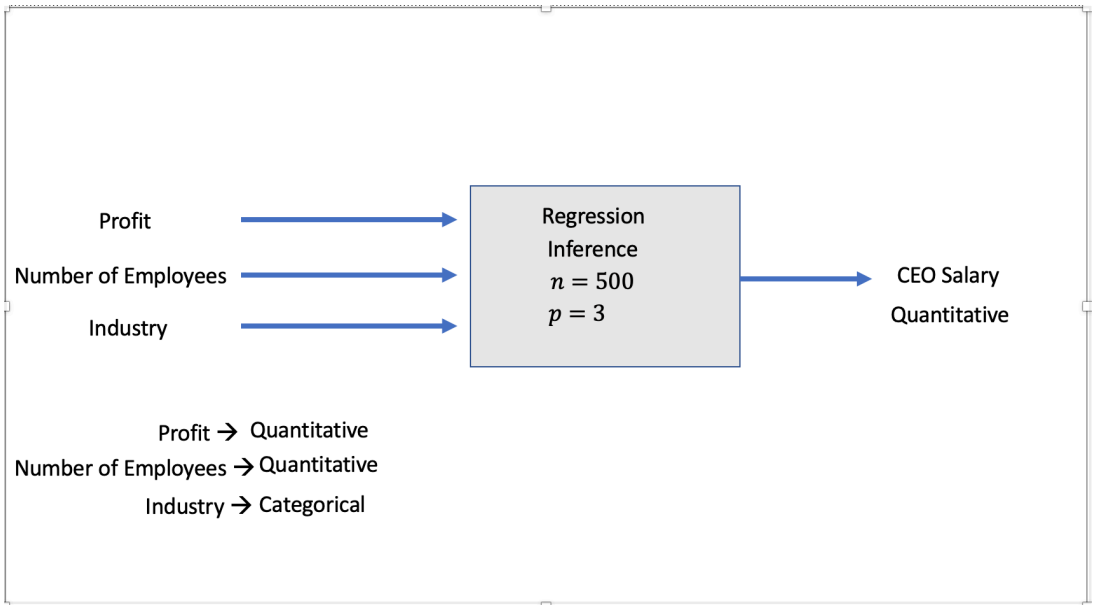
(a) The sample size n is extremely large, and the number of predictors p is small.
A. Permormance will improve.
Reason: As the sample size increases, the algorithm will get to learn various pattern in data. Hence the algorithm predicts well for unseen data . As the sample size is large, there is high probability that most of the pattern is captured / algorithm is trained. The data is better fit for flexible method with large sample size.

(b) The number of predictors p is extremely large, and the number of observations n is small.
B. Reason: The algorithm does not get enough data to get it trained or captured. Hence with larger predictors and small sample, most of the pattern of the data will be missing. The algorithm will get less pattern to get trained as there is small sample size. As as result when a new pattern is encountered in the unseen data, the performance will get worse.

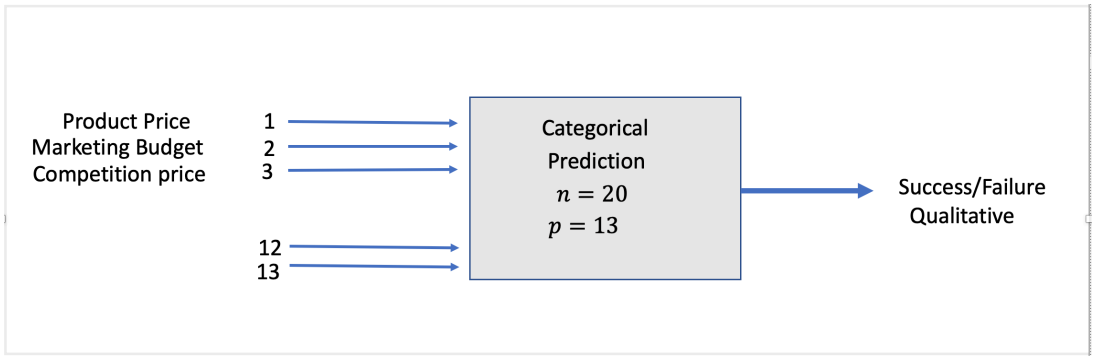
(c) The relationship between the predictors and response is highly non-linear.
C. Reason : Flexible method works well for non-linear data . The non linear data is captured with Flexible methods as they have more degree of freedom.

(d) The variance of the error terms, i.e. σ2 = Var(), is extremely high.
D) Reason : As the method is flexible, the algorithm captures most of the data. Along the data, the algorithm also pick the noise associated with data when there is high variance of error.

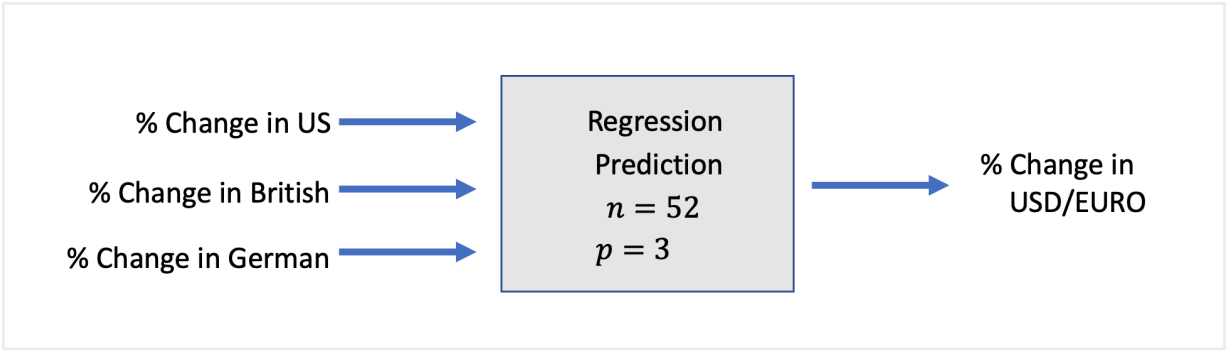
Explain whether each scenario is a classification or regression problem and indicate whether we are most interested in inference or prediction. Finally, provide n and p.
(a) We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the
CEO salary. We are interested in understanding which factors affect CEO salary.



(b) We are considering launching a new product and wish to know whether it will be a *success* or a *failure*. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.

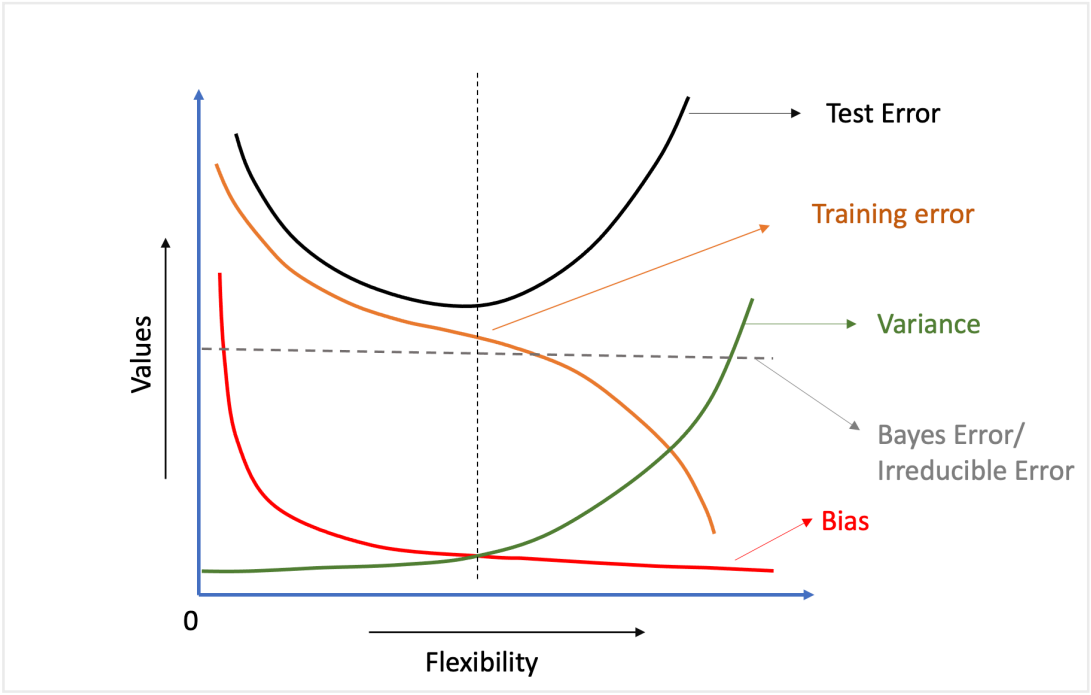


(c) We are interested in predicting the % change in the USD/Euro exchange rate in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record the % change in the USD/Euro, the % change in the US market, the % change in the British market, and the % change in the German market.



N is Weekly data the year - 52 weeks in a year(2012)

3. We now revisit the bias-variance decomposition.
(a) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, and the y-axis should represent the values for each curve. There should be five curves.Make sure to label each one.



b) Explain why each of the five curves has the shape displayed in part (a).

Bayes Error : This is part of dataset and remains constant.It is not altered by modeling Features.

Training Error : This is calculated using training data.This decreases monotonically as flexibility incresases. At higher flexibility, the data overfits and there is much decrease in training error.
Test Error : The test error decreases monotonically to certain point and as the data overfits, the Test Error increases.
Bias : As the flexibility increases, the bias reduces monotonically to certain point and remains constant .
Variance **:** The variance is low at low flexibility and increases monotonically after certain point.

4. You will now think of some real-life applications for statistical learning

(a) Describe three real-life applications in which classification might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

Spam Detection: It is a two class classification. The incoming data is analyzed for the presence of keywords usually linked with the spam.
Here the application is Prediction.

5. **What are the advantages and disadvantages of a very flexible (versus a less flexible) approach for regression or classification? Under what circumstances might a more flexible approach be preferred to a less flexible approach? When might a less flexible approach be preferred ?**

Advantage Flexible approach : Yields less Bias. Suited for non-linear data and represents more complex system.
Disadvantage : Variance increases and data can overfit.

Obs.	X_1	X_2	X_3	Y
1	0	3	0	Red
2	2	0	0	Red
3	0	1	3	Red
4	0	1	2	Green
5	−1	0	1	Green
6	1	1	1	Red

6. Describe the differences between a parametric and a non-parametric statistical learning approach. What are the advantages of a parametric approach to regression or classification (as opposed to a non-parametric approach)? What are its disadvantages?

Less flexible

- (+) gives better results with few observations
- (+) simpler inference: the effect of each feature can be more easily understood
- (+) fewer parameters, faster optimisation
- (-) performs poorly if observations contain highly non-linear relationships

More flexible

- (+) gives better fit if observations contain non-linear relationships
- (-) can overfit the data providing poor predictions for new observations

7. The table below provides a training data set containing six observations, three predictors, and one qualitative response variable.

Obs.	X_1	X_2	X_3	Y
1	0	3	0	Red
2	2	0	0	Red
3	0	1	3	Red
4	0	1	2	Green
5	−1	0	1	Green
6	1	1	1	Red

Suppose we wish to use this data set to make a prediction for Y when $X_1 = X_2 = X_3 = 0$ using K-nearest neighbors.

(a) Compute the Euclidean distance between each observation and the test point, $X_1 = X_2 = X_3 = 0$.

In [7]:
```
import pandas as pd
import numpy as np
```

In [8]:
```
df = pd.DataFrame({'Obs': [1, 2, 3, 4, 5, 6],
                   'X1': [0, 2, 0, 0, -1, 1],
                   'X2': [3, 0, 1, 1, 0, -1],
                   'X3': [0, 0, 3, 2, 1, 1],
                   'Y': ['Red', 'Red', 'Red', 'Green', 'Green', 'Red']})
```

In [9]:
```
df
```

Out[9]:

	Obs	X1	X2	X3	Y
0	1	0	3	0	Red
1	2	2	0	0	Red
2	3	0	1	3	Red
3	4	0	1	2	Green
4	5	-1	0	1	Green
5	6	1	-1	1	Red

a) Compute the Euclidean distance between each observation and test point, X1 = X2 = X3 = 0

In [10]:
```
def euclidian_dist(x):

    return (np.sum(x**2, axis=1))**0.5

eucli_dist = pd.DataFrame({'Euclidean_Dist': euclidian_dist(df[['X1', 'X2', 'X3']])})
euclidian_dist = pd.concat([df, eucli_dist], axis=1)
euclidian_dist
```

Out[10]:

	Obs	X1	X2	X3	Y	Euclidean_Dist
0	1	0	3	0	Red	3.000000
1	2	2	0	0	Red	2.000000
2	3	0	1	3	Red	3.162278
3	4	0	1	2	Green	2.236068
4	5	-1	0	1	Green	1.414214
5	6	1	-1	1	Red	1.732051

b) What is the prediction with K = 1? Why ?

Soln: With K = 1, the prediction is Green .The nearest neighbour for Z -> (0,0,0) is Green.

	Obs	X1	X2	X3	Y	Euclidean_Dist
0	1	0	3	0	Red	3.000000
1	2	2	0	0	Red	2.000000
2	3	0	1	3	Red	3.162278
3	4	0	1	2	Green	2.236068
4	5	-1	0	1	Green	1.414214
5	6	1	-1	1	Red	1.732051

b) What is our prediction with K = 3 ? Why ?

Soln : With K = 3,the prediction is Red. The nearest 3 neighbours for Z -> (0,0,0) has majority Red.

	Obs	X1	X2	X3	Y	Euclidean_Dist
0	1	0	3	0	Red	3.000000
1	2	2	0	0	Red	2.000000
2	3	0	1	3	Red	3.162278
3	4	0	1	2	Green	2.236068
4	5	-1	0	1	Green	1.414214
5	6	1	-1	1	Red	1.732051

(d) If the Bayes decision boundary in this problem is highly non-linear, then would we expect the *best* value for K to be large or small? Why?

Soln : Small : If Bayes decision boundary in this problem is highly non-linear, then the best value for K should be small. As small values of K are more flexible for non-linear boundary. Large value of K smoothens the boundary.