**31. Have you used AB testing in your project So far? If yes, Explain. If not, Tell me about AB testing.**

If yes:

Yes we have used a/b testing in our data science project for Analyzing the Results of two models.It is one of the most effective methods in making conclusions about any hypothesis one may have.We create the A and B version of our model and calculate the success rate of the that based on the comparison

**Tell me about AB testing**

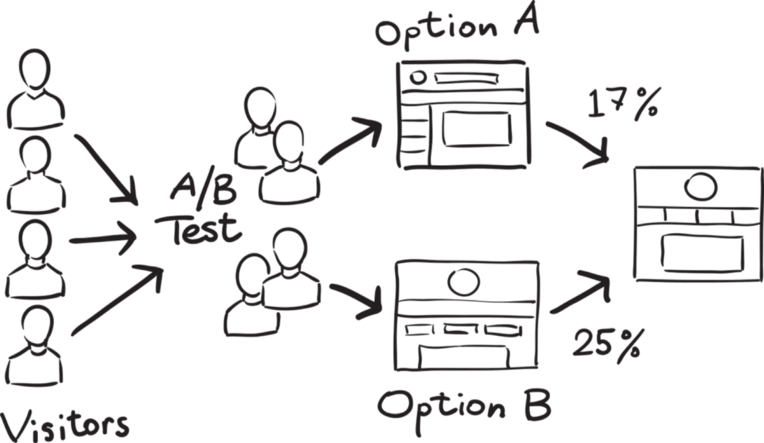
A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better.

Or

(definition from wikipedia)

A/B testing is a method of comparing two versions of a product or app against each other to determine which one performs better. A/B testing is essentially an experiment where two or more variants of a product are shown to users at random, and statistical analysis is used to determine which variation performs better.

For instance, let’s say you own a company and want to increase the sales of your product. Here, either you can use random experiments, or you can apply scientific and statistical methods. A/B testing is one of the most prominent and widely used statistical tools.



In the above scenario, you may divide the products into two parts – A and B. Here A will remain unchanged while you make significant changes in B’s packaging. Now, on the basis of the response from customer groups who used A and B respectively, you try to decide which is performing better.

**32. Can we use the Alternate hypothesis as a null Hypothesis?**

Hypothesis is a statement, assumption or claim about the value of the parameter (mean, variance, median etc.

Like, if we make a statement that “Dhoni is the best Indian Captain ever.” This is an assumption that we are making based on the average wins and loses team had under his captaincy. We can test this statement based on all the match data.

**Null Hypothesis**

The null hypothesis is the hypothesis to be tested for possible rejection under the assumption that it is true. The concept of the null is similar to innocent until proven guilty We assume innocence until we have enough evidence to prove that a suspect is guilty.

It is denoted by H0.

Alternate Hypothesis

The alternative hypothesis complements the Null hypothesis. It is opposite of the null hypothesis such that both Alternate and null hypothesis together cover all the possible values of the population parameter.

It is denoted by H1.

Let’s understand this with an example:

**A soap company claims that it’s product kills on an average 99% of the germs. To test the claim of this company we will formulate the null and alternate hypothesis.**

Null Hypothesis(H0): Average =99%

Alternate Hypothesis(H1): Average is not equal to 99%.

Note: The thumb rule is that a statement containing equality is the null hypothesis.

**Hypothesis Testing**

When we test a hypothesis, we assume the null hypothesis to be true until there is sufficient evidence in the sample to prove it false. In that case we reject the null hypothesis and support the alternate hypothesis.

If the sample fails to provide sufficient evidence for us to reject the null hypothesis, we cannot say that the null hypothesis is true because it is based on just the sample data. For saying the null hypothesis is true we will have to study the whole population data.

So the main question is: Can we use the Alternate hypothesis as a null Hypothesis?

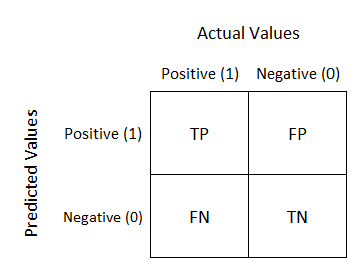
No, We can’t use it based on the above explanation.The **alternate hypothesis** is the opposite of the null hypothesis.

**33. Can you please explain the confusion matrix for more than 2 variables?**

Confusion matrix is a performance measurement for machine learning classification problems where output can be two or more classes.

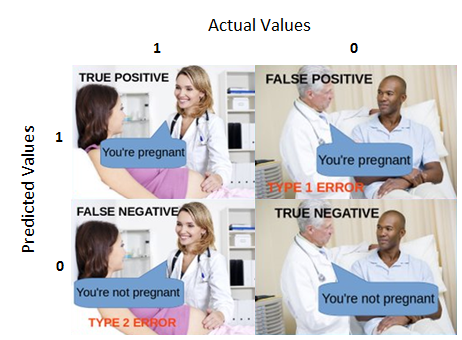
**For 2 variable:**

It is a table with 4 different combinations of predicted and actual values.



It is extremely useful for measuring Recall, Precision, Specificity, Accuracy, and most importantly AUC-ROC curves.

Let’s understand TP, FP, FN, TN in terms of pregnancy analogy.



**Confusion matrix for a 3 class classification:**

Let’s try to answer the above question with a popular dataset – IRIS DATASET.

The dataset has 3 flowers as outputs or classes, Versicolor, Virginia, Setosa.



Source: Google

With the help of petal length, petal width, sepal length, sepal width the model has to classify the given instance as Versicolor or Virginia or Setosa flower.

Let’s apply a classifier model here: decision Tree classifier is applied on the above dataset. The dataset has 3 classes hence we get a 3 X 3 confusion matrix.

But how to know TP, TN, FP, FN values !!!!!

In the multi-class classification problem, we won’t get TP, TN, FP, FN values directly as in the binary classification problem. We need to calculate for each class.

How to calculate FN, FP, TN, TP :

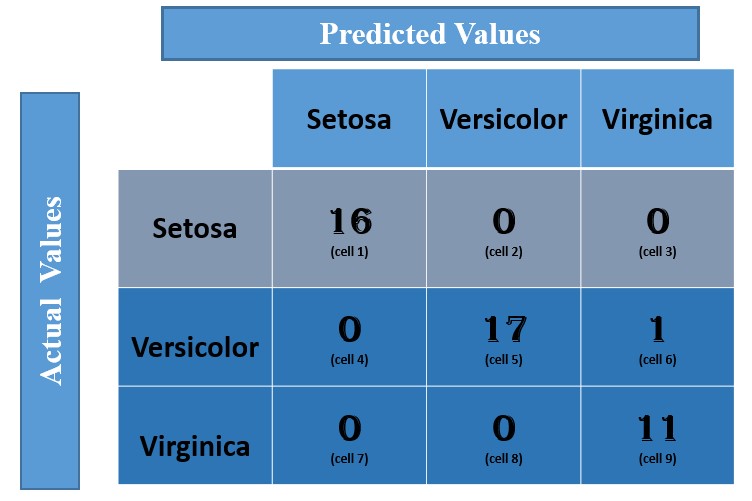
FN: The False-negative value for a class will be the sum of values of corresponding rows except for the TP value.

FP: The False-positive value for a class will be the sum of values of the corresponding column except for the TP value.

TN: The True Negative value for a class will be the sum of values of all columns and rows except the values of that class that we are calculating the values for.

TP: The True positive value is where the actual value and predicted value are the same.

The confusion matrix for the IRIS dataset is as below:



1.Let us calculate the TP, TN, FP, FN values for the class Setosa using the Above tricks:

TP: The actual value and predicted value should be the same. So concerning the Setosa class, the value of cell 1 is the TP value.

FN: The sum of values of corresponding rows except the TP value

FN = (cell 2 + cell3)

= (0 + 0)

= 0

FP : The sum of values of the corresponding column except the TP value.

FP = (cell 4 + cell 7)

= (0 + 0)

= 0

TN: The sum of values of all columns and rows except the values of that class that we are calculating the values for.

TN = (cell 5 + cell 6 + cell 8 + cell 9)

= 17 + 1 +0 + 11

= 29

Similarly, for Versicolor class the values/ metrics are calculated as below:

TP : 17 (cell 5)

FN : 0 + 1 = 1 (cell 4 +cell 6)

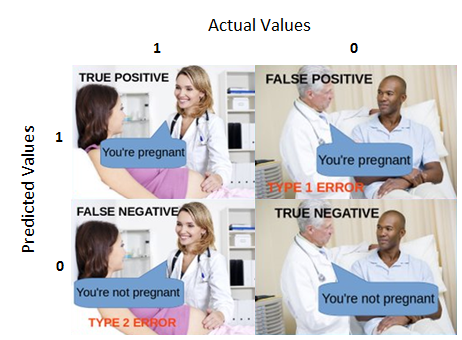
FP : 0 + 0 = 0 (cell 2 + cell 8)

TN : 16 +0 +0 + 11 =27 (cell 1 + cell 3 + cell 7 + cell 9).

I hope the concept is clear and you can try for the Virginia class.

**34. Give me an example of False Negative From this interview?**

A false negative error, or false negative, is a test result which wrongly indicates that a condition does not hold. For example, when **a pregnancy test indicates a woman is not pregnant**, but she is, or when a person guilty of a crime is acquitted, these are false negatives.



**35. What do you understand by Precision, Recall and F1 Score with example?**

**Confusion Matrix**

A typical confusion matrix looks like the figure shown.

Where the terms have the meaning:

\_\_True Positive(TP):\_\_ A result that was predicted as positive by the classification model and also is positive

\_\_True Negative(TN):\_\_ A result that was predicted as negative by the classification model and also is negative

\_\_False Positive(FP):\_\_ A result that was predicted as positive by the classification model but actually is negative

\_\_False Negative(FN):\_\_ A result that was predicted as negative by the classification model but actually is positive.

The Credibility of the model is based on how many correct predictions the model did.

## What is the accuracy of the machine learning model for this classification task?

Accuracy represents the number of correctly classified data instances over the total number of data instances.

In this example, Accuracy = (55 + 30)/(55 + 5 + 30 + 10 ) = 0.85 and in percentage the accuracy will be 85%.

Is accuracy the best measure?

Accuracy may not be a good measure if the dataset is not balanced (both negative and positive classes have different numbers of data instances). We will explain this with an example.

Consider the following scenario: There are 90 people who are healthy (negative) and 10 people who have some disease (positive). Now let’s say our machine learning model perfectly classified the 90 people as healthy but it also classified the unhealthy people as healthy. What will happen in this scenario? Let us see the confusion matrix and find out the accuracy?

In this example, TN = 90, FP = 0, FN = 10 and TP = 0. The confusion matrix is as follows.

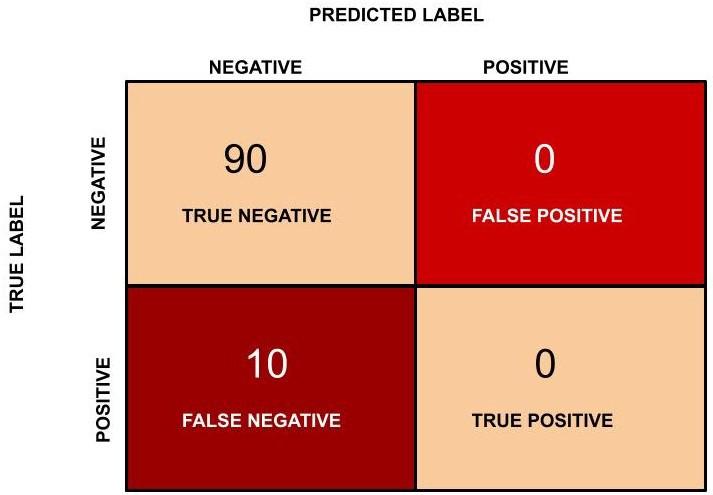


Figure 7: Confusion matrix for healthy vs unhealthy people classification task.

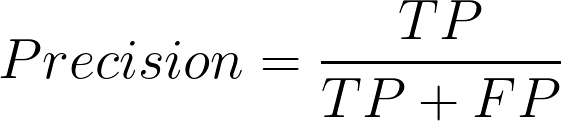
Accuracy in this case will be (90 + 0)/(100) = 0.9 and in percentage the accuracy is 90 %.

Is there anything fishy?

The accuracy, in this case, is 90 % but this model is very poor because all the 10 people who are unhealthy are classified as healthy. By this example what we are trying to say is that accuracy is not a good metric when the data set is unbalanced. Using accuracy in such scenarios can result in misleading interpretation of results.

So now we move further to find out another metric for classification. Again we go back to the pregnancy classification example.

Now we will find the precision (positive predictive value) in classifying the data instances. Precision is defined as follows:

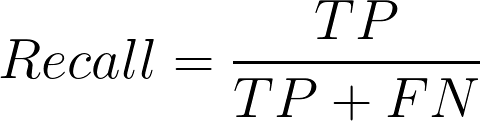


What does precision mean?

Precision should ideally be 1 (high) for a good classifier. Precision becomes 1 only when the numerator and denominator are equal i.e TP = TP +FP, this also means FP is zero. As FP increases the value of the denominator becomes greater than the numerator and precision value decreases (which we don’t want).

So in the pregnancy example, precision = 30/(30+ 5) = 0.857

Now we will introduce another important metric called recall. Recall is also known as sensitivity or true positive rate and is defined as follows:

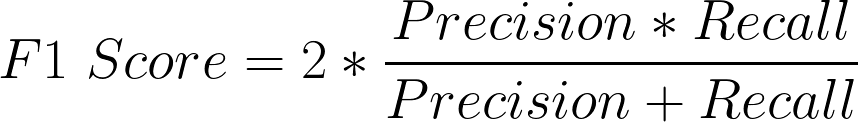


Recall should ideally be 1 (high) for a good classifier. Recall becomes 1 only when the numerator and denominator are equal i.e TP = TP +FN, this also means FN is zero. As FN increases the value of the denominator becomes greater than the numerator and recall value decreases (which we don’t want).

So in the pregnancy example let us see what the recall will be.

Recall = 30/(30+ 10) = 0.75

So ideally in a good classifier, we want both precision and recall to be one which also means FP and FN are zero. Therefore we need a metric that takes into account both precision and recall. F1-score is a metric which takes into account both precision and recall and is defined as follows:



F1 Score becomes 1 only when precision and recall are both 1. F1 score becomes high only when both precision and recall are high. F1 score is the harmonic mean of precision and recall and is a better measure than accuracy.

In the pregnancy example, F1 Score = 2\* ( 0.857 \* 0.75)/(0.857 + 0.75) = 0.799.

36. What kind of questions do you ask your client if they give you a dataset?

* How was the data compiled? Was it aggregated from multiple sources? ...
* Is the data accurate? ...
* Is the data clean? ...
* How much data should you have? ...
* Remember why: what problem do you want to tackle?
* Dimension of the dataset
* Type of the attributes in the dataset
* For predictive analytics, target attribute
* Missing values in the data set
* How to fill missing values?

**37. Have you ever done an F test on your dataset, if yes, give an example. If No, then explain F distribution?**

F Distribution

F-Test (variance ratio test)

When we run a regression analysis, we get f value to find out the means between two populations. It's similar to a T statistic from a T-Test. A T-test will tell you if a single variable is related statistically, and an F test will tell you if a group of variables is jointly significant.

* F-test is used to test the two independent estimations of population variances(S1^2 & S2^2).
* F-test is used by comparing the ratio of the two variances S1^2 & S2^2.
* The samples must be independent.
* F-test is a small sample test.
  + F = (Larger estimate of population variance) / (Smaller estimate Of population variance)
* The variance ratio = S1^2 & S2^2
* F-test never is -ve because the upper value is greater than lower.
* Degree of freedom for larger population[vS1] variance is V1[vS2] and smaller V2
* The null hypothesis of two population variance are equal, i.e.,
  + HO: S1^2 = S2^2

Determining the Values of F

F Distribution using Python

#impolrt scipy, numpy and matplotlib

x=np.linspace(-10, 10, 100)

dfn = 29

dfd = 18

mean, var, skew, kurt = scipy.stats.f.stats(dfn, dfd, moments='mvsk')

print('mean: {:.2f}, skewness: {:.2f}, kurtosis: {:.2f}'.format(mean, var, skew, kurt))

plt.plot(x, scipy.stats.f.pdf(x,dfn, dfd))

plt.show()

mean: 1.12, skewness: 0.28, kurtosis: 1.81

Note:

* The Student ‘t’ distribution is robust, which means that if the population is non-normal, the results of the t-test and confidence interval estimate are still valid provided that the population is not extremely non-normal.
* To check this requirement, draw a histogram of the data and see how bell-shaped the resulting figure is. If a histogram is extremely skewed (say in that case of an exponential distribution), that could be considered “extremely non-normal,” and hence, t-statistics would not be valid in this case.

Example

Question: From a population of women, suppose you randomly select 7 women, and from the population of men, 12 men are selected.

| Population | Population standard deviation | Sample standard deviation |
| --- | --- | --- |
| Women | 30 | 35 |
| Men | 50 | 45 |

To calculate f statistics.

Answer: The f statistic can be calculated from the sample standard deviations and population, using the following equation: f = [ s12/σ12 ] / [ s22/σ22 ]

where Standard deviation of the sample drawn from population 1 is s1 and s2 in the denominator is the standard deviation of the sample drawn from population 2, σ1 is the standard deviation of population 1, Population 2’s standard deviation is σ2.

As we can see from the equation, there are two ways to compute an f statistic from these data. If the data of women appears in the numerator, we can compute f statistic as follows:

f = ( 552 / 202 ) / ( 452 / 502 )

f = (3025 / 400) / (2025 / 2500).

f = 1.361 / 0.81 = 1.68

For calculations, the numerator degrees of freedom v1 are 7 - 1 or 6; and the degrees of freedom for denominator v2 are 12 - 1 or 11.

On the other hand, if the men's data appears in the numerator, we can calculate the f statistic as follows:

f = ( 452 / 502 ) / ( 552 / 202 )

f = (2025 / 2500) / (3025 / 400)

f = 0.812 / 1.3610 = 0.5955

For this calculation, the denominator degrees of freedom v2 is 7 - 1 or 6 and the numerator degrees of freedom v1 is 12 - 1 or 11

When we are trying to find the cumulative probability associated with an f statistic, you need to know v1 and v2.

Question: Find the cumulative probability related to each of the f statistics from the above example:

Answer: First, we need to find the degrees of freedom for each sample. Then, probabilities can be found.

* The sample of women’s degrees of freedom is equal to n - 1 = 7 - 1 = 6.
* The sample of men’s degrees of freedom is equal to n - 1 = 12 - 1 = 11.

Therefore, when data of women appear in the numerator, then v1 is equal to 6; and then v2 is equal to 11. And, the f statistic is equal to 1.68. So, 0.78 is the cumulative probability.

When data of men appear in the numerator, then v1 is equal to 11; and then v2 is equal to 6. And, the f statistic is equal to 0.595. Thus the cumulative probability is 0.22.

**38. What is AUC & ROC Curve? Explain with uses.**

We know that the classification algorithms work on the concept of probability of occurrence of the possible outcomes. A probability value lies between 0 and 1. Zero means that there is no probability of occurrence and one means that the occurrence is certain.

But while working with real-time data, it has been observed that we seldom get a perfect 0 or 1 value. Instead of that, we get different decimal values lying between 0 and 1. Now the question is if we are not getting binary probability values how are we actually determining the class in our classification problem?

There comes the concept of Threshold. A threshold is set, any probability value below the threshold is a negative outcome, and anything more than the threshold is a favourable or the positive outcome. For Example, if the threshold is 0.5, any probability value below 0.5 means a negative or an unfavourable outcome and any value above 0.5 indicates a positive or favourable outcome.

<image>

Now, the question is, what should be an ideal threshold?

The horizontal lines represent the various values of thresholds ranging from 0 to 1.

\* Let’s suppose our classification problem was to identify the obese people from the given data.

\* The green markers represent obese people and the red markers represent the non-obese people.

\* Our confusion matrix will depend on the value of the threshold chosen by us.

\* For Example, if 0.25 is the threshold then

TP(actually obese)=3

TN(Not obese)=2

FP(Not obese but predicted obese)=2(the two red squares above the 0.25 line)

FN(Obese but predicted as not obese )=1(Green circle below 0.25line )

A typical ROC curve looks like the following figure.

<img src="ROC.PNG" width="300">

\* Mathematically, it represents the various confusion matrices for various thresholds. Each black dot is one confusion matrix.

\* The green dotted line represents the scenario when the true positive rate equals the false positive rate.

\* As evident from the curve, as we move from the rightmost dot towards left, after a certain threshold, the false positive rate decreases.

\* After some time, the false positive rate becomes zero.

\* The point encircled in green is the best point as it predicts all the values correctly and keeps the False positive as a minimum.

\* But that is not a rule of thumb. Based on the requirement, we need to select the point of a threshold.

\* The ROC curve answers our question of which threshold to choose.

### But we are confused!!

Let’s suppose that we used different classification algorithms, and different ROCs for the corresponding algorithms have been plotted.

The question is: which algorithm to choose now?

The answer is to calculate the area under each ROC curve.

#### AUC(Area Under Curve)

<img src="AUC.PNG" width="300">

\* It helps us to choose the best model amongst the models for which we have plotted the ROC curves

\* The best model is the one which encompasses the maximum area under it.

\* In the adjacent diagram, amongst the two curves, the model that resulted in the red one should be chosen as it clearly covers more area than the blue one

**39. Who decided in your last project, what will be the accuracy of your model & what was the criterion to make the decision.**

Whether i was doing a classification problem so i have chosen parameter for classification model evaluation

For classification model evaluation we have different different parameter like performance matrix,pr curve, roc-auc curve in performance matrix also we have a different different evaluation parameter like accuracy, error rate , precision, recall so based on our class distribution we can choose any of them, first i have checked accuracy of the model and also i have gone through through with roc-auc curve inside that i have checked auc score of the given model.

In the auc score I had the criterion 0.5 or 50% based on that i have filtered the model then i have compared auc score between the models as well so whatever auc score i have found greater that i have model i have chosen finally.

**40. What do you understand by 1 tail test & 2 tail test? Give an example.**

If the alternate hypothesis gives the alternate in both directions (less than and greater than) of the value of the parameter specified in the null hypothesis, it is called a Two **tailed test**.

If the alternate hypothesis gives the alternate in only one direction (either less than or greater than) of the value of the parameter specified in the null hypothesis, it is called **One tailed test**.

e.g. if H0: mean= 100 H1: mean not equal to 100

Here according to H1, the mean can be greater than or less than 100. This is an example of Two tailed test

Similarly, if H0: mean>=100 then H1: mean< 100

Here, the mean is less than 100, it is called One tailed test.

**41. What do you understand by the power of a test?**

The statistical power of a binary hypothesis test is the probability that the test correctly rejects the null hypothesis H0 when a specific alternative hypothesis H1 is true. It is commonly denoted by 1-beta , and represents the chances of a "true positive" detection conditional on the actual existence of an effect to detect. Statistical power ranges from 0 to 1, and as the power of a test increases,

the probability beta of making a type II error by wrongly failing to reject the null hypothesis decreases.

**42. How do you set the level of significance for your dataset?**

In normal English, "significant" means important, while in Statistics "significant" means probably true (not due to chance). A research finding may be true without being important. When statisticians say a result is "highly significant" they mean it is very probably true. They do not (necessarily) mean it is highly important.

determining the significance level(alpha). This refers to the likelihood of rejecting the null hypothesis even when it's true. A common alpha is 0.05 or 5 percent.(We can choose 1% or 10% as well)

**43. Have you ever used a T table in any of your projects so far? If No, then why is statistics important for data scientists? If yes, explain the scenario.**

It is the science of conducting studies to collect, organize, summarize, analyze, and draw a conclusion out of data.It deals with collective informative data, interpreting those data, and drawing a conclusion from that data.It is used in many disciplines like marketing, business, healthcare, telecom, etc.

In any data science project, data helps us to analyze the initial level of insight.

Building models using popular statistical methods such as Regression, Classification, Time Series Analysis and Hypothesis Testing which is core of data science. Data Scientists run suitable experiments and interpret the results with the help of these statistical methods.

So we have a data and based on that data we are going to create a statistical model which will be able to learn from data itself above some of the statics techniques has been given based on this scenario you can understand statistics is important with respect to datascience.

**44. Can we productionise statistical model?**

**What is productionize**

It means testing and deploying an application to production such that it uses real data on a frequent basis to produce output for use by the business. When data scientists build and test models, it is often a very manual process.

Optimizing data science across the entire enterprise requires more than just cool tools for wrangling and analyzing data. Obviously, we can simply hardcode a data science **model** or rent a pre-trained predictive model in the cloud, embed it into an application in-house and we are done.yes so we can productionise our statistical model.

**45. How frequently do you build the model and test it?**

If a model’s predictive performance has fallen due to changes in the environment, the solution is to retrain the model on a new training set, which reflects the current reality. How often should you retrain your model? And how do you determine your new training set? The answer is that *it depends*. But what does it depend on?

**Sometimes the problem setting itself will suggest when to retrain your model. For instance, suppose you’re working for a university admissions department and are tasked with building a student attrition model that predicts whether a student will return the following semester. This model will be used to generate predictions on the current cohort of students directly after midterms. Students identified as being at risk of churning will automatically be enrolled in tutoring or some other such intervention.**

**Let’s think about the time horizon of such a model. Since we’re generating predictions in batches once a semester, it doesn’t make sense to retrain the model any more often than this because we won’t have access to any new training data. Therefore we might choose to retrain our model at the start of each semester after we’ve observed which students from the previous semester dropped out.**

This is an example of a **periodic** retraining schedule. It’s often a good idea to start with this simple strategy but you’ll need to determine based on your business problem exactly how frequently you’ll need to retrain.

Quickly changing training sets might require you to train as often as daily or weekly. Slower varying distributions might require monthly or annual retraining.