17.4.1

Assess Accuracy, Precision, and Sensitivity

It's not enough to use a machine learning model to create predictions. The model must answer an important question: how well does it perform? You have seen that accuracy score is one way of assessing a classification model's performance. That is, what percentage of predictions does it get right?

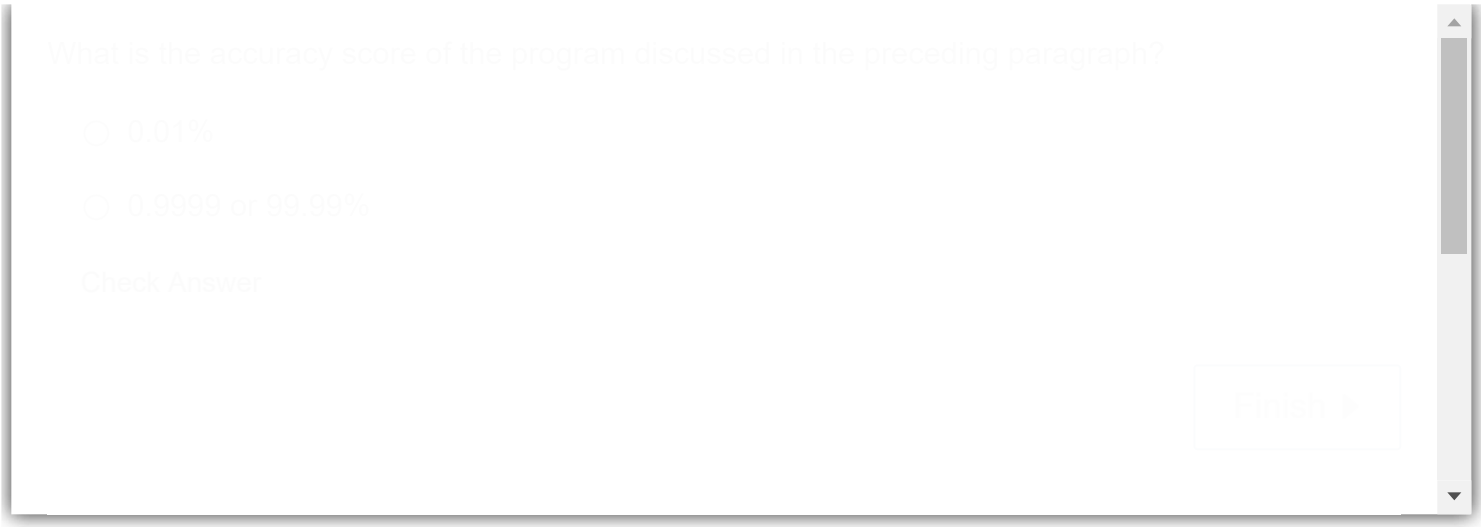
Jill explains that there are other ways to validate a classification model, and asks you to look into them. This is where the statistical rubber meets the road!

Accuracy

In an earlier module, you learned that the performance of a linear regression model is measured based on the difference between its predicted values and actual values.

However, this method cannot be used when the target values are not continuous. Different methods must be used to assess a model with discrete outcomes. We have already seen one way of validating such a model's performance: its **accuracy score.** An accuracy score is not always an appropriate or a meaningful performance metric, however.

Imagine the following scenario, in which a credit card company wishes to detect fraudulent transactions in real time. Historically, out of 100,000 transactions, 10 have been fraudulent. An analyst writes a program to detect fraudulent transactions, but due to an uncaught bug, it flags every transaction as not fraudulent. Out of 100,000 transactions, it correctly classifies the 99,990 transactions that are not fraudulent, and it erroneously classifies all 10 transactions that are fraudulent.



The program's accuracy score appears to be impressive at 99.99%. However, it fails spectacularly at its job, detecting 0 out of 10 fraudulent transactions, a success rate of 0%.

Predictions in such a scenario with binary outcomes can be categorized according to the table below:

	Predicted True	Predicted False
Actually True	TRUE POSITIVE	FALSE NEGATIVE
Actually False	FALSE POSITIVE	TRUE NEGATIVE

Any given prediction falls under one of two categories: true or false. In the context of fraud detection, a true prediction would mean that the model categorizes the transaction as fraudulent. A false prediction means that the model categorizes the transaction as not fraudulent.

If a transaction is predicted to be fraudulent and is really a fraudulent transaction, it is a true positive (TP).

If a transaction is predicted to be fraudulent but is not fraudulent, it is a false positive (FP). It falsely categorized an innocent transaction as fraudulent.

Similarly, if a transaction is predicted to be non-fraudulent but is actually fraudulent, it is a false negative (FN).

And when a transaction is predicted to be non-fraudulent and is in reality non-fraudulent, it is a true negative (TN).

A patient has a streptococcal infection, and a clinical test for the infection came back negative. What is the classification?

- ○ True positive
- ○ True negative
- ○ False negative
- ○ False positive

Check Answer

A 41-year-old woman takes a mammogram, which comes back positive for breast cancer. Subsequent examination of her breast tissue by a pathologist reveals that her tissue is noncancerous. What is the classification?

- ○ True positive
- ○ True negative
- ○ False negative
- ○ False positive

Check Answer

Finish ▶

Precision

Imagine that a man is experiencing weight loss and chills. He consults an online test that uses machine learning algorithms to see whether he might have cancer, which informs him that he indeed has cancer. However, the online test is not perfect. When the test was previously evaluated in a study, the results were collated into the following table, called a **confusion matrix**.

	Predicted True	Predicted False
Actually True	30	10

Actually False	20	40

How many people, in total, were assessed in this study?

- ○ 20
- ○ 40
- ○ 60
- ○ 100

Check Answer

How many people actually had cancer?

Check Answer

How many people were diagnosed with cancer?

Check Answer

Finish ▶

The man in this scenario has a positive diagnosis for cancer. He wants to know how likely it is that he actually has cancer. **Precision,** also known as positive predictive value (PPV), is a measure of this. Precision is obtained by dividing the number of true positives (TP) by the number of all positives (i.e., the sum of true positives and false positives, or TP + FP).

Precision = TP/(TP + FP)

How many true positives are there in the study?

- ○ 20
- ○ 30
- ○ 40

Check Answer

How many false positives are there in the study?

- ○ 20
- ○ 30
- ○ 40

Check Answer

What is the precision of this test?

- ○ 0.4
- ○ 0.5
- ○ 0.6

Check Answer

Finish ▶

In this study, a total of 50 people were predicted to have cancer. Of the 50, 30 people actually had cancer. The precision is therefore 30/50, or 0.6.

NOTE

The terms precision and positive predictive value (PPV) are interchangeable.

To summarize, in machine learning, precision is a measure of how reliable a positive classification is. The following formulation may help you in remembering precision: "I know that the test for cancer came back positive. How likely is it that I have cancer?"

Sensitivity

Another way to assess a model's performance is with sensitivity, also called recall. While the term **recall** is more commonly used in machine learning, the two terms are synonymous and will be used interchangeably from this point.

The following formulation may help you understand sensitivity: "I know that I have cancer. How likely is it that the test will diagnose it?" Here is the formula for sensitivity:

Sensitivity = TP/(TP + FN)

In this context, all who have cancer means true positives (those who have cancer and were correctly diagnosed) and false negatives (those who have cancer and were incorrectly diagnosed as not having cancer). **Sensitivity** is a measure of how many people who actually have cancer were correctly diagnosed.

NOTE

The terms sensitivity and recall are used interchangeably.

What is the sensitivity of this test for those who actually had cancer?

- ○ 0.25
- ○ 0.5
- ○ 0.75

Check Answer

Which is more important in a screening test to detect cancer: precision or sensitivity?

- ○ Precision
- ○ Sensitivity

Check Answer

Finish ▶

Tradeoff Between Precision and Sensitivity

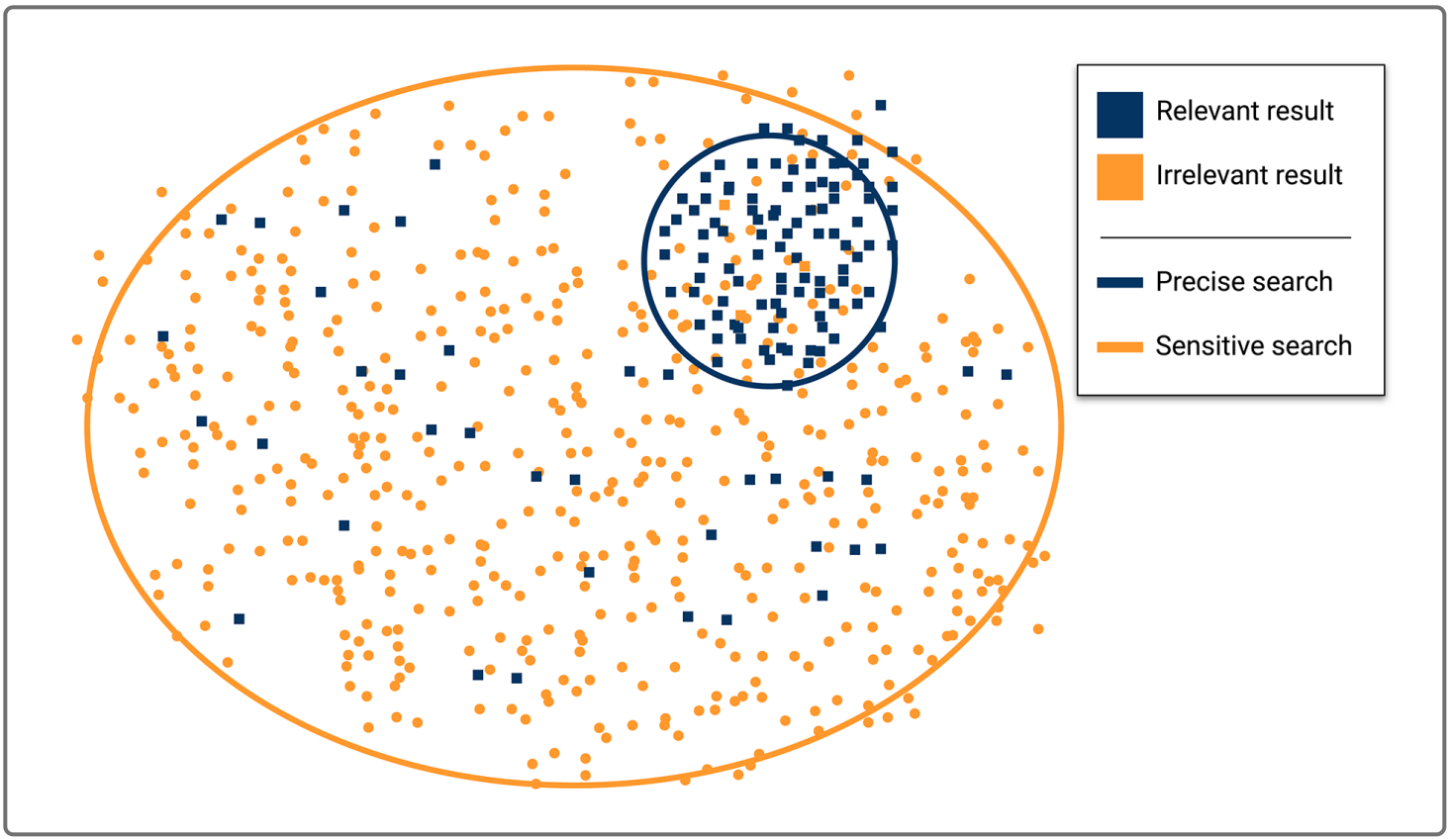
There are situations in which high sensitivity is more important. With cancer screening, for example, a high sensitivity is more important than high precision. Remember that high sensitivity means that among people who actually have cancer, most of them will be diagnosed correctly. High precision, on the other hand, means that if the test comes back positive, there's a high likelihood that the patient has cancer.

It may appear at first that the two terms refer to the same thing, but they do not. As an extreme example, let's say that among 100 people, 50 have cancer and 50 do not. A screening algorithm is extremely aggressive and labels everyone to have cancer. Since everyone who actually has cancer is detected, the sensitivity is 1.0, or 100%. However, the precision is low: being diagnosed with cancer in this case only means a 50% likelihood of actually having cancer. In other words, there are many false positives.

To return to a previous point: Why is high sensitivity more important than precision for a cancer screening test? It's better to detect everyone who might have cancer, even if it means a certain number of false positives, than to miss people who do have cancer. After all, those with a positive result for cancer can undergo further testing to confirm or rule out cancer. The false positives in a highly sensitive test are accepted as a cost of doing business.

In contrast, there are situations in which precision is more important than sensitivity. Imagine that the criminal justice system depended on a machine learning algorithm to judge the innocence or guilt of a person on trial. Is high sensitivity or high precision preferable? Perfect sensitivity would mean that everyone who committed a crime is

declared guilty. However, a potential consequence of such an aggressive algorithm is that people who didn't commit a crime are also declared guilty. Perfect precision would mean that someone who's been declared guilty actually is guilty. But it also means that there may be people who committed a crime who aren't found guilty. If the justice system values sparing innocent people false imprisonments more than punishing all guilty criminals, precision trumps sensitivity.



When using a machine learning algorithm to detect fraudulent credit card transactions, which is more important between sensitivity and precision?

- ○ Precision
- ○ Sensitivity

Check Answer

A political campaign has a database of potential donors. The role of the phone bank is to call potential donors for contributions. Due to limited time and staffing, however, not everyone on the database can be telephoned, and a machine learning algorithm is used to sort the list into likely and unlikely donors. Is sensitivity or precision more important in this case?

- ○ Precision
- ○ Sensitivity

Check Answer

In an algorithm for spam email detection, which is more important: precision or sensitivity?

- ○ Precision
- ○ Sensitivity

Check Answer

Finish ▶

In summary, there's a fundamental tension between precision and sensitivity. Highly sensitive tests and algorithms tend to be aggressive, as they do a good job of detecting the intended targets, but also risk resulting in a number of false positives. High precision, on the other hand, is usually the result of a conservative process, so that predicted positives are likely true positives; but a number of other true positives may not be predicted. In practice, there is a trade-off between sensitivity and precision that requires a balancing act between the two.

F1 Score

The F1 score, also called the harmonic mean, can be characterized as a single summary statistic of precision and sensitivity. The formula for the F1 score is the following:

2(Precision * Sensitivity)/(Precision + Sensitivity)

NOTE

The terms F1 score and harmonic mean are interchangeable.

To illustrate the F1 score, let's return to the scenario of a faulty algorithm for detecting fraudulent credit card transactions. Say that 100 transactions out of 100,000 are fraudulent.

If a faulty algorithm labels every transaction as fraudulent, what is the sensitivity?

- ○ 0
- ○ 1

Check Answer

Using the same scenario above, what is the precision?

- ○ 0.001
- ○ 0.00001
- ○ 0.1

Check Answer

Finish ▶

In such a scenario, the sensitivity is very high, while the precision is very low. Clearly, this is not a useful algorithm. Nor does averaging the sensitivity and precision yield a useful figure. Let's try calculating the F1 score.

Using the same scenario above, what is the F1 score in this case?

- ○ 0.001
- ○ 0.002
- ○ 0.003

Check Answer

Finish ▶

The F1 score is 0.002. We noted previously that there's usually a trade-off between sensitivity and precision, and that a balance must be struck between the two. A useful way to think about the F1 score is that a pronounced imbalance between sensitivity and precision will yield a low F1 score.