Olivier Maugain

Interpreting Data

Course Notes

365√DataScience

Table of Contents

ABSTRACT 3
1. Interpreting Data 4
1.1. Correlation analysis 5
1.1.1. Correlation coefficient 7
1.1.2. Correlation and causation 8
1.2. Simple linear regression 9
1.2.1. R-squared 11
1.3. Forecasting 13
1.3.1. Forecast errors 14
1.4. Statistical tests 17
1.4.1. Hypothesis testing 18
1.4.2. P-value 20
1.4.3. Statistical significance 21
1.5. Classification 22
1.5.1. Accuracy 24
1.5.2. Recall and precision 25

ABSTRACT

Now that you have built a solid theoretical foundation and understanding of data literacy qualities, real-life data applications, data terminology, storage methods and quality assessment techniques, in the final section of the data literacy course notes with, you will learn how to

The ability to carry out various analytical methods on a dataset is instrumental for communicating invaluable business insights that guide organizational decision-making. Data interpretation techniques are commonly used to evaluate the performance of any asset on the financial market, predict consumer behavior and forecast future sales.

Leaning on the more technical side, these course notes and their supplementary graphs, will familiarize you with:

- The different types of fundamental analyses
- The ability to draw meaningful conclusions from variable relationships
- The common causation and correlation fallacy
- The types of forecasting methods
- The process of performing statistical tests and hypothesis

Keywords: data interpretation, data literacy, statistical tests, data analysis, classification

1. Interpreting Data

Different types of analyses or method produce various forms of output. These can be statistics, coefficients, probabilities, errors, etc., which can provide different insights to the reader.

It is important to be able to interpret these results, to understand what they mean, to know what kind of conclusions can be drawn and how to apply them for further decision making.

Data interpretation requires domain expertise, but also curiosity to ask the right questions. When the results are not in line with one's expectations, these should be met with a sufficient level of doubt and examined in further depth. Such sense of mistrust and inquisitiveness can be trained.

Five types of data interpretation approaches:

1) Correlation

2) Linear regression

3) Forecasting

4) Statistical tests

5) Classification

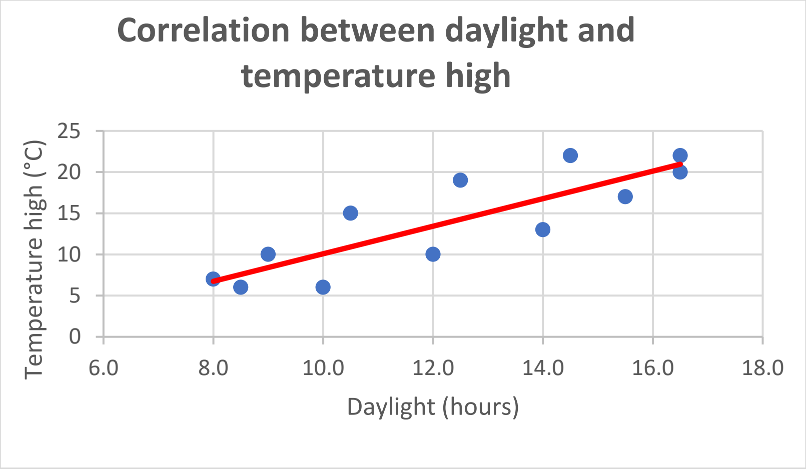
1.1. Correlation analysis

Correlation analysis is a *technique aimed at establishing whether two quantitative variables are related*

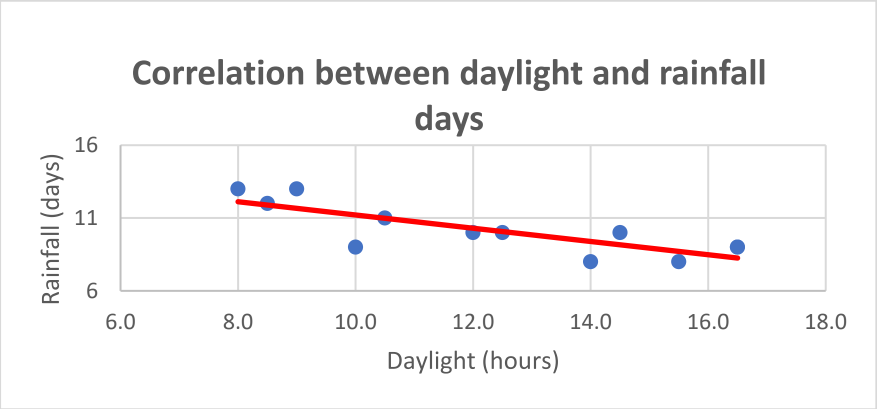
- It provides an indication of the direction and strength of the linear relationship between the two variables
- It is only applicable with two numerical (quantitative) variables.
- Correlations between two variables can be presented graphically in a scatter plot:

When a **correlation** exists, you should be able to draw a straight line (called "regression line") that fits te data well.

Positive correlation (upward sloping regression line) - both variables move in the same direction. When one increases/decreases, the other increases/decreases as well.

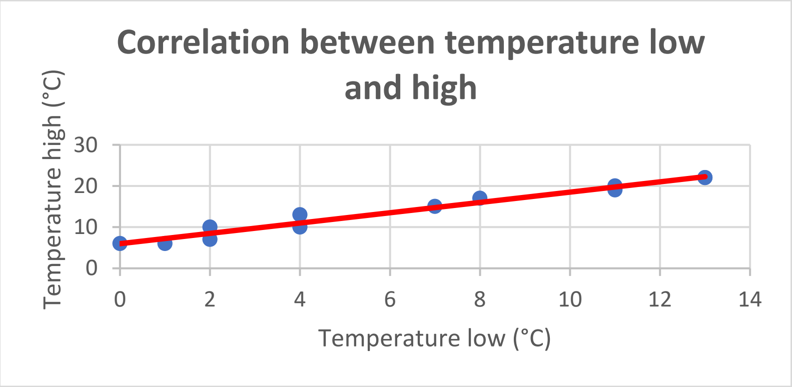


Negative correlation (downward sloping regression line) - the variables move in opposite directions. When one increases/decreases, the other decreases/increases.

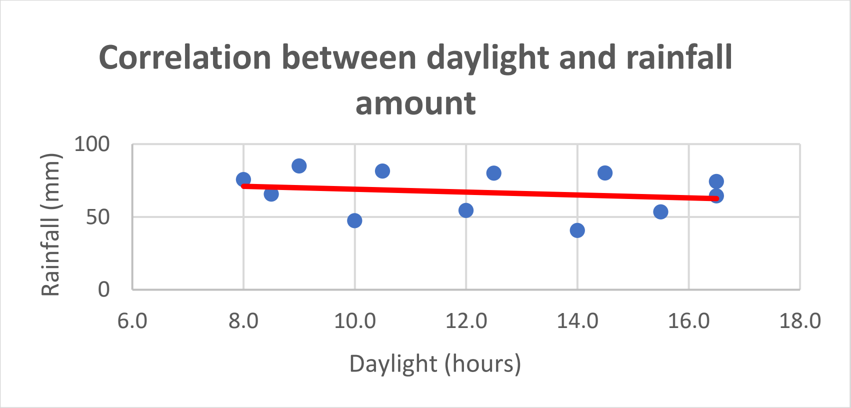


The more tightly the plot forms a line rising from left to right, the stronger the correlation.

"Perfect" correlation - all the points would lie on the regression line itself



Lack of correlation - If a line does not fit, i.e. if the dots are located far away from the line, it means that the variables are not correlated. In that case, the line is relatively flat.



1.1.1. Correlation coefficient

Although the visual examination of a scatter plot can provide some initial clues about the correlation between two variables, relying solely on them may lead to an interpretation that is too subjective.

It is also possible to determine the correlation of variables numerically, thus offering more precise evidence.

Correlation coefficient - a statistic that summarizes in a single number the relationship that can be seen in a scatter plot.

Basics about correlation coefficients:

The possible values of a correlation coefficient range from -1 to +1

Correlation coefficient = -1 => perfect negative correlation

Correlation coefficient = +1 => perfect negative correlation

When the value of one variable increases/decreases, the value of the other one moves in perfect sync.

The higher the absolute value of the correlation coefficient, the stronger the correlation

Strong positive (respectively negative) correlation, when the value of one variable increases/decreases, the value of the other variable increases (respectively decreases) in a similar manner

Correlation coefficient = 0 => there is no linear relationship between the two variables.

However, this does not necessarily mean the variables are not related at all. They may have some other form of relationship (e.g., an exponential or logarithmic relationship), but not a linear one.

Correlation in the real world rarely return coefficients of exactly +1.0, –1.0, or 0. Most of the time, they fall somewhere in between.

A rule of thumb for the interpretation of the absolute value of correlation coefficients:

- Below 0.25: No correlation
- Between 0.25 and 0.50: Weak correlation
- Between 0.50 and 0.75: Moderate correlation
- Above 0.75: Strong correlation

When applying this rule of thumb, always bear in mind that the definition of a weak, moderate, or strong correlation may vary from one domain to the next.

1.1.2. Correlation and causation

Correlation ≠ causal relationship.

Correlation only establishes a statistical relationship. No matter how strong or significant the correlation is, however, it never provides sufficient evidence to claim a causal link between variables. This applies to both the *existence* or the *direction* of a cause-and-effect relationship, for which no conclusion can made whatsoever.

The strong correlation between two variables A and B can be due to various scenarios:

- Direct causality: A causes B
- Reverse causality: B causes A
- Bidirectional causality: A causes B and B causes A
- A and B are both caused by C
- Pure coincidence, i.e., there is no connection between A and B

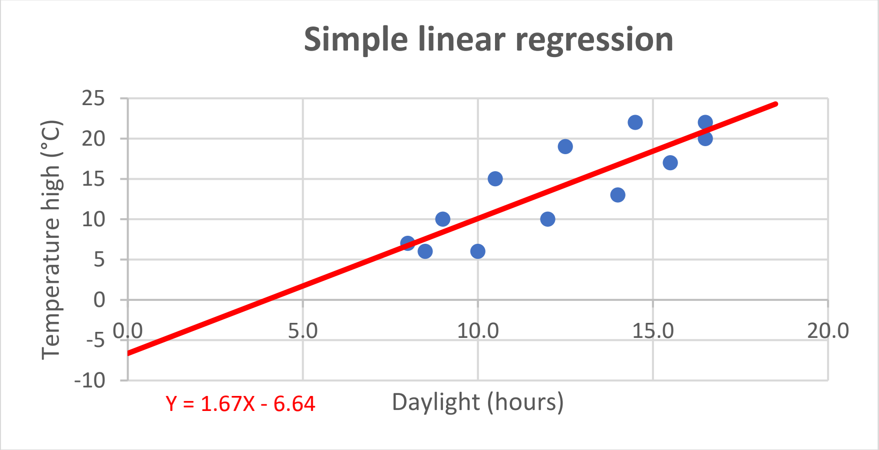
Determining an actual cause-and-effect relationship between the variables requires further investigation.

The presence of correlation, combined with sound business judgment, is sometimes enough to make quality decisions. A good analyst or a wise decision maker will prefer to strive for an explanation and always bear in mind the old statistical adage:

Correlation does not imply causation.

1.2. Simple linear regression

With simple linear regression, it is possible to predict the value of the **dependent variable** based on the value of one single independent variable.



Regression equation: Y = a + bX

Y (dependent variable) - the one to be explained or to be predicted

X (independent variable, predictor) – the one explaining or predicting the value of Y

a (intercept) - a constant, corresponding to the point at which the line crosses the vertical axis, i.e., when X is equal to zero

b (slope) - the coefficient of X, quantifying how much Y changes for each incremental (one-unit) change in X.

N.B. The higher the absolute value of b, the steeper the regression curve

The sign of b indicates the direction of the relationship between the Y and X

b > 0 - the regression line shows an upward slope. An increase of X results in an increase of Y

b < 0 - the regression line shows a downward slope. An increase of X results in a decrease of Y

Notice that a prediction using simple linear regression does not prove any causality. The coefficient b, no matter its absolute value, says nothing about a causal relationship between the dependent and the independent variables.

Summary: The objective of the regression is to plot the one line that best characterizes the cloud of dots.

1.2.1. R-squared

After running a linear regression analysis, you need to examine how well the model fits the data., i.e., determine if the regression equation does a good job explaining changes in the dependent variable.

The regression model fits the data well when the differences between the observations and the predicted values are relatively small. If these differences are too large, or if the model is biased, you cannot trust the results.

R-squared (R^2, coefficient of determination) - estimates the scatter of the data points around the fitted regression line

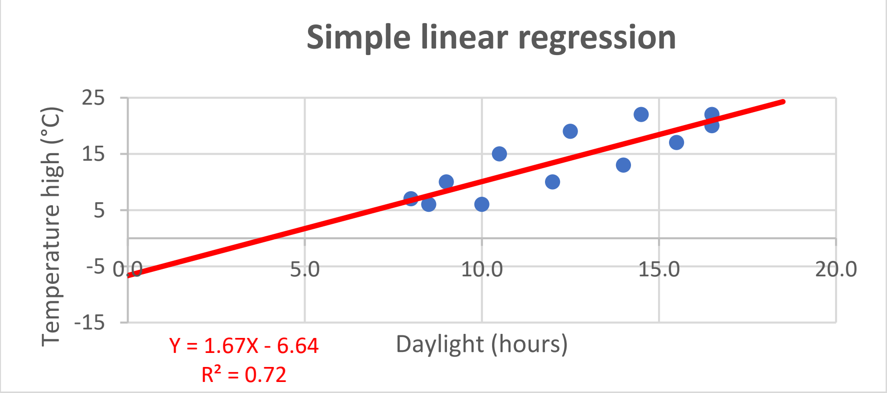
It represents the proportion of the variance in a dependent variable that can be explained by the independent variable.

The remaining variance can be attributed to additional, unknown, variables or inherent variability.

In simple linear regression models, the R-squared statistic is always a number between 0 and 1, respectively between 0% and 100%

- **R-squared = 0** => the model does not explain any of the variance in the dependent variable. The model is useless and should not be used to make predictions
- **R-squared = 1** => the model explains the whole variance in the dependent variable. Its predictions perfectly fit the data, as all the observations fall exactly on the regression line.

Example:



R-squared = 0.72

=> the number of daylight accounts for about 72% of the variance of the temperature

=> the number of daylight hours is a good predictor of the temperature.

What is a good R-squared value? At what level can you trust a model?

- Some people or textbooks claim that 0.70 is such a threshold, i.e., that if a model returns an R-squared of 0.70, it fits well enough to make predictions based on it. Inversely, a value of R-squared below 0.70 indicates that the model does not fit well.
- Feel free to use this rule of thumb. At the same time, you should be aware of its caveats.
- Indeed, the properties of R-squared are not as clear-cut as one may think

A higher R-squared indicates a better fit for the model, producing more accurate predictions

**A model that explains 70% of the variance is likely to be much better than one that explains 30% of the variance. However, such a conclusion is not necessarily correct.*

R-squared depends on various factors, such as:

- The sample size: The larger the number of observations, lower the R-squared typically gets
- The granularity of the data: Models based on case-level data have lower R-squared statistics than those based on aggregated data (e.g., city or country data)
- The type of data employed in the model: When the variables are categorical or counts, the R-squared will typically be lower than with continuous data
- The field of research: studies that aim at explaining human behavior tend to have lower R-squared values than those dealing with natural phenomena. This is simply because people are harder to predict than stars, molecules, cells, viruses, etc.

1.3. Forecasting

Forecast accuracy – the closeness of the forecasted value to the actual value.

For a business decision maker, the key question is how to determine the accuracy of forecasts:

"Can you trust the forecast enough to make a decision based on it?

As the actual value cannot be measured at the time the forecast, the accuracy can only be determined retrospectively.

Accuracy of forecasts is tested by calculating different statistical indicators, also known as errors. These errors can be assessed without knowing anything about a forecast except its past values.

The three most common forecast errors:

1) the Mean Absolute Error (MAE)

2) the Mean Absolute Percent Error (MAPE)

3) the Root Mean Square Error (RMSE)

**The MAE, MAPE and RMSE only measure typical errors. They cannot anticipate black swan events like financial crises, global pandemics, terrorist attacks, or the Brexit.*

As the errors associated with these events are not covered by the time series data, they cannot be modeled. Accordingly, it is impossible to determine in advance how big the error will be.

1.3.1. Forecast errors

Mean Absolute Error (MAE) - the absolute value of the difference between the forecasted values and the actual value.

- With the MAE, we can get an idea about how large the error from the forecast is expected to be on average.
- The main problem with it is that it can be difficult to anticipate the relative size of the error. How can we tell a big error from a small error?

What does a MAE of 100 mean? Is it a good or bad forecast?

That depends on the underlying quantities and their units. For example, if your monthly average sales volume is 10’000 units and the MAE of your forecast for the next month is 100, then this is an amazing forecasting accuracy. However, if sales volume is only 10 units on average, then we are talking of a rather poor accuracy.

Mean Absolute Percentage Error (MAPE) - the sum of the individual absolute errors divided by the underlying value.

- It corresponds to the average of the percentage errors
- The MAPE allows us to compare forecasts of different series in different scales.
- Hence, we could, for example, compare the accuracy of a forecast of sales value and volume in US$, EUR, units, liters or gallons - even though these numbers are in different units.
- Thanks to this property, the MAPE is one of the most used indicators to measure forecast accuracy
- One key problem with the MAPE is that it may understate the influence of big, but rare, errors. Consequently, the error is smaller than it should be, which could mislead the decision maker.

Root Mean Square Error (RMSE) - the square root of the average squared error.

RMSE has the key advantage of giving more importance to the most significant errors. Accordingly, one big error is enough to lead to a higher RMSE. The decision maker is not as easily pointed in the wrong direction as with the MAE or MAPE.

- Best practice is to compare the MAE and RMSE to determine whether the forecast contains large errors. The smaller the difference between RMSE and MAE, the more consistent the error size, and the more reliable the value.

- As with all “errors”, the objective is always to avoid them.

The smaller the MAE, MAPE or RMSE, the better!

What is a “good” or “acceptable” value of the MAE, MAPE or RSME depends on:

- The environment: In stable environment, in which demand or prices do not vary so much over time (e.g., electricity or water distribution), demand or sales volumes are likely to be rather steady and predictable. A forecasting model may therefore yield a very low MAPE, possibly under 5%.
- The industry: In volatile industries (e.g., machine building, oil & gas, chemicals) or if the company is exposed to hyper competition and constantly has to run advertising campaigns or price promotions (like in the FMCG or travel industries), sales volumes vary significantly over time and are much more difficult to be forecasted accurately. Accordingly, the MAPE of a model could be much higher than 5%, and yet be useful for decision makers in the sales, finance or supply chain departments.
- The type of company: Forecasts for larger geographic areas (e.g., continental or national level) are generally more accurate than for smaller areas (e.g., regional or local).
- The time frame: Longer period (e.g., monthly) forecasts usually yield higher accuracies than shorter period (e.g., daily or hourly) forecasts.

With three well-established indicators available, one cannot conclude that one is better than the other. Each indicator can help you avoid some shortcomings but will *be prone to others. Only experimentation with all three indicators can tell you which one is best, depending on the phenomenon to be forecasted.*

1.4. Statistical tests

Hypothesis testing is a key tool in inferential statistics, and used in various domains - social sciences, medicine, and market research. The purpose of hypothesis testing is to establish whether there is enough statistical evidence in favor of a certain idea or assumption, i.e., the hypothesis.

The process involves testing an assumption regarding a population by measuring and analyzing a random sample taken from that population

Population *- the entire group that is being examined.*

If you want to compare the satisfaction levels of male and female employees in your company, the population is made of the entire workforce (25,421 people)

Sample *- the specific group that data are collected from. Its size is always smaller than that of the population.*

If you randomly select 189 men and 193 women among these employees to carry out a survey, these 382 employees constitute your sample.

Hypothesis testing becomes particularly relevant when census data cannot be collected, for example because the process would be too lengthy or too expensive. In these cases, researchers need to develop specific experiment designs, and rely on survey samples to collect the necessary data.

Modern statistical software is there to calculate various relevant statistics, test values, probabilities, etc. All you need to do is to learn interpret the most important ones:

- null hypothesis
- the p-value
- statistical significance.

1.4.1. Hypothesis testing

A hypothesis resembles a theory in science. But it is "less" than that, because it first needs to go through extensive testing before it can be deemed a proper theory.

Hypotheses are formulated as statements, not as questions.

A hypothesis constitutes a starting point for further analytical investigation.

Step 1 in any statistical test is to define a hypothesis, which is a statement that helps communicate an understanding of the question or issue at stake. It is common to propose two opposite, mutually exclusive, hypotheses so that only one can be right: *The null hypothesis (Ho) and the alternative hypothesis (H1).*

The null hypothesis (Ho)

The null hypothesis represents the commonly accepted fact

It is usually expressed as a hypothesis of "no difference" in a set of given observations, for example:

- *A blue conversion button on the website results in the same CTR as a red button*

In statistics terms, the null hypothesis is therefore usually stated as the equality between population parameters. For example:

- *The mean CTRs of the red and blue conversion buttons are the same.*

OR: the difference of the mean CTRs of the red and the blue conversion buttons is equal to zero

It is called "null" hypothesis, because it is usually the hypothesis that we want to nullify or to disprove.

The alternative hypothesis (H1) is the one that you want to investigate, because you think that it can help explain a phenomenon

It represents what you believe to be true or hope to prove true.

- In the example above, the alternative hypothesis could be formulated as follows:
  - *A blue conversion button on the website will lead to a different CTR than the one with a red button*

In this case, the objective is to determine whether the population parameter is generally distinct or differs in either direction from the hypothesized value. It is called a two-sided (or non-directional) alternative hypothesis.

Sometimes, it can be useful to determine whether the population parameter differs from the hypothesized value in a specific direction, i.e. is smaller or greater than the value. This is known as a one-sided (or directional) alternative hypothesis

- *Example: The difference of the mean CTRs of the blue and the red conversion button is positive.*

- *Here we only care about the blue button yielding a higher CTR than the red button*

We can also be even more aggressive in our statement and quantify that difference, for example:

- The difference of the mean CTRs of the red and the blue conversion button is higher than 2 percentage points
- That would be equivalent to stating that the mean CTR of the blue button is 2 percentage points higher than mean CTR of the red button

N.B. You do not have to specify the alternative hypothesis. Given that the two hypotheses are opposites and mutually exclusive, only one can, and will, be true. For the purpose of statistical testing, it is enough to reject the null hypothesis. It is therefore very important to work out a clear null hypothesis.

1.4.2. P-value

P-value is a measure of the probability that an observed result could have occurred just by random chance, as opposed to a certain pattern.

The greater the dissimilarity between these patterns, the less likely it is that the difference occurred by chance.

Examples:

If p-value = 0.0326 => there is a 0.0326 (or 3.26%) chance that the results happened randomly.

If p-value = 0.9429 => the results have a 94.29% chance of being random

The smaller the p-value, the stronger the evidence that you should reject the null hypothesis

When you see a report with the results of statistical tests, look out for the p-value. Normally, the closer to 0.000, the better – depending, of course, on the hypotheses stated in that report.

1.4.3. Statistical significance

The **significance level** (alpha, α) is a number stated in advance to determine how small the p-value must be to reject the null hypothesis.

- The researcher sets it arbitrarily, before running the statistical test or experiment
- If the p-value falls below the significance level, the result of the test is statistically significant.
- Unlike the p-value, the alpha does not depend on the underlying hypotheses, nor is it derived from any observational data.
- The alpha will often depend on the scientific domain the research is being carried out in

If p-value < alpha => you can reject the null hypothesis at the level alpha.

If the P-value is lower than the significance level alpha, which should be set in advance, then we can conclude that the results are strong enough to reject the old notion (the null hypothesis) in favor of a new one (the alternative hypothesis).

Example:

If p-value = 0.0321, alpha = 0.05 => "Based on the results, we can reject the null hypothesis at the level of significance α = 0.05 (as p = 0.0321 < 0.05).

If the p-value = 0.1474, alpha = 0.05 => "Based on the results, we accept the null hypothesis at the level of significance α = 0.05 (as p = 0.1474 >= 0.05).

N.B. Statistical significance ≠ practical (theoretical) significance.

A result that is statistically significant is not necessarily "meaningful" or "important". That will depend on the real-world relevance of that result, which the researcher must determine.

1.5. Classification

A classification model can only achieve two results: Either the prediction is correct (i.e., the observation was placed in the right category), or it is incorrect.

This characteristic makes it rather straightforward to estimate the quality of a classification model, especially when there are only two available categories or labels.

Out-of-sample validation - withholding some of the sample data used for the training of the model. Once the model is ready, it is validated with the data initially set aside for this very purpose.

Example:

Imagine that we trained a model for a direct marketing campaign. We used the data available to predict each recipient's response to the marketing offer:

- "yes"- favorable response

- “no” – negative response

We set aside 100 customer records, which constitute our validation data.

- For these 100 customers, we use the model to predict their responses. These constitute the predicted classes.
- As these customers also receive the marketing offer, we also get to know who responded favorably, and who did not. These responses constitute the actual classes.
- You can compare the predicted with the actual classes, and find out which predictions were correct.

Confusion matrix - shows the actual and predicted classes of a classification problem (correct and incorrect matches). The rows represent the occurrences in the actual class, while the columns represent the occurrences in the predicted class.

n = 100		**Predicted class**		
		Yes	**No**	
Actual class	**Yes**	10	5	**15**
	No	15	70	**85**
		25	**75**	**100**

There are two possible responses to a marketing offer:

- "yes" - the customers accept it
- "no" - they ignore or reject it.

Out of the 100 customer who received the offer, the model predicted that 25 customers would accept it (i.e., 25 times "yes") and that 75 customers would reject it (i.e., 75 times "No"

After running the campaign, it turned out that 15 customers responded favorably ("Yes"), while 85 customers ignored it ("No").

Based on the confusion matrix, one can estimate the quality of a classification model by calculating its:

- **Accuracy**
- **Recall**
- **Precision**

1.5.1. Accuracy

Accuracy is the proportion of the total number of correct predictions.

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Number\ of\ total\ observations}$$

The model correctly predicted 10 "Yes" cases and 70 "No" cases =>

$$Accuracy = \frac{10 + 70}{100} = 80\%$$

- Accuracy is the most fundamental metric used to assess the performance of classification models.
- It is intuitive and widely used.
- It comes with a major flaw, which becomes apparent when the classes are imbalanced.
- Experienced analysts are familiar with this issue, and have at their disposal various techniques to handle imbalanced data sets

1.5.2. Recall and precision

Recall (also known as sensitivity) is the ability of a classification model to identify *all* relevant instances.

$$Recall = \frac{Total\ correct\ predictions}{Actual\ positive\ cases}$$

The model correctly predicted 10 out of 15 positive responses =>

$$Recall = \frac{10}{15} = 66.67\%$$

- This means that only two-thirds of the positives were identified as positive, which is not a very good score.

Precision is the ability of a classification model to return *only* relevant instances (to be correct when predicting "yes").

$Precision = \frac{Total\ correct\ predictions}{Predicted\ positive\ cases}$The model predicted 25 positive responses, out of which 10 were correctly predicted =>

The model predicted 25 positive responses, out of which 10 were correctly predicted =>

$$Precision\ = \frac{10}{25} = 40\%$$

There are two types of incorrect predictions: false positives and false negative

A false positive - when a case is labeled as positive although it is actually negative

- The model predicts "yes" where the case is actually "no"
- When the objective is to minimize false positives, it is advisable to assess classification models using precision

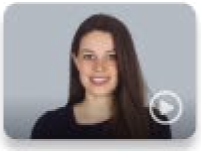
A false negative - when a case is labeled as negative although it is actually positive

- The model predicts "no" where the case is actually "yes"
- When the objective is to minimize false negatives, it is advisable to assess classification models using recall.

Learn **DATA SCIENCE**
anytime, anywhere, at your own pace.

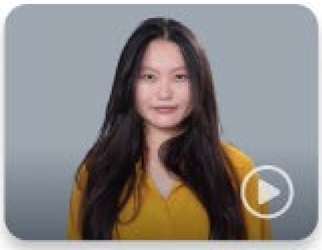
If you found this resource useful, check out our **e-learning program**. We have everything you need to succeed in data science.

Learn the most sought-after data science skills from the **best experts in the field**! Earn a **verifiable certificate** of achievement trusted by employers worldwide and future proof your car



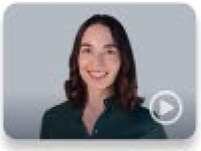
Danielle Thé
Esade Ramon Llull University

Bernard Marr
Cambridge University



Tina Huang
University of Pennsylvania

Ken Jee
DePaul University



Anastasia Kuznetsova
Université Côte d'Azur

Comprehensive training, exams, certificates.

- ✓ 160+ hours of video
- ✓ 599+ Exercises
- ✓ Downloadables
- ✓ Exams & Certification
- ✓ Personalized support
- ✓ Resume Builder & Feedback
- ✓ Portfolio advice
- ✓ New content
- ✓ Career tracks

Join a global community of 1.8 M successful students with an annual subscription at 60% OFF with coupon code 365RESOURCES.

~~$432~~ $172.80/year

Start at 60% Off

VAT may be applied

365√DataScience

Olivier Maugain

Email: team@365datascience.com

365√DataScience