STATISTICS · DATA SCIENCE AND MACHINE LEARNING

# Predictive Modelling: Training, Testing and Validation

Why a predictive model must be judged on data it has not seen, how training, validation and test sets and cross-validation work, and the mistakes that make accuracy look better than it is.

ModQu research article · 718 words · 5 references (APA 7th edition)

A predictive model is built to make good guesses about cases it has never seen: next quarter's customers, the respondents in the next survey wave, the visitors arriving tomorrow. That goal changes how a model must be judged. Explanatory modelling asks whether coefficients make sense; predictive modelling asks how wrong the predictions will be on new data. The only honest way to answer that question is to test the model on data that played no part in building it (Hastie et al., 2009).

## Why training performance misleads

A model fitted to a data set learns two things at once: the real patterns that will recur, and the noise peculiar to that sample. Scored on the same data, the model gets credit for both, so its accuracy on the training data is almost always optimistic. A model can reproduce its training data perfectly and still predict poorly, a problem known as overfitting.

## Training, validation and test sets

The standard remedy is to divide the available data into three parts with distinct jobs.

- The **training set** is used to estimate the model's parameters.
- The **validation set** is used to compare candidate models and tune settings such as the number of predictors or the strength of a penalty.
- The **test set** is locked away and used once, at the end, to estimate how the chosen model will perform in practice.

The separation matters. Each time an analyst adjusts a model after looking at a validation score, that score becomes a little more optimistic; only an untouched test set stays unbiased.

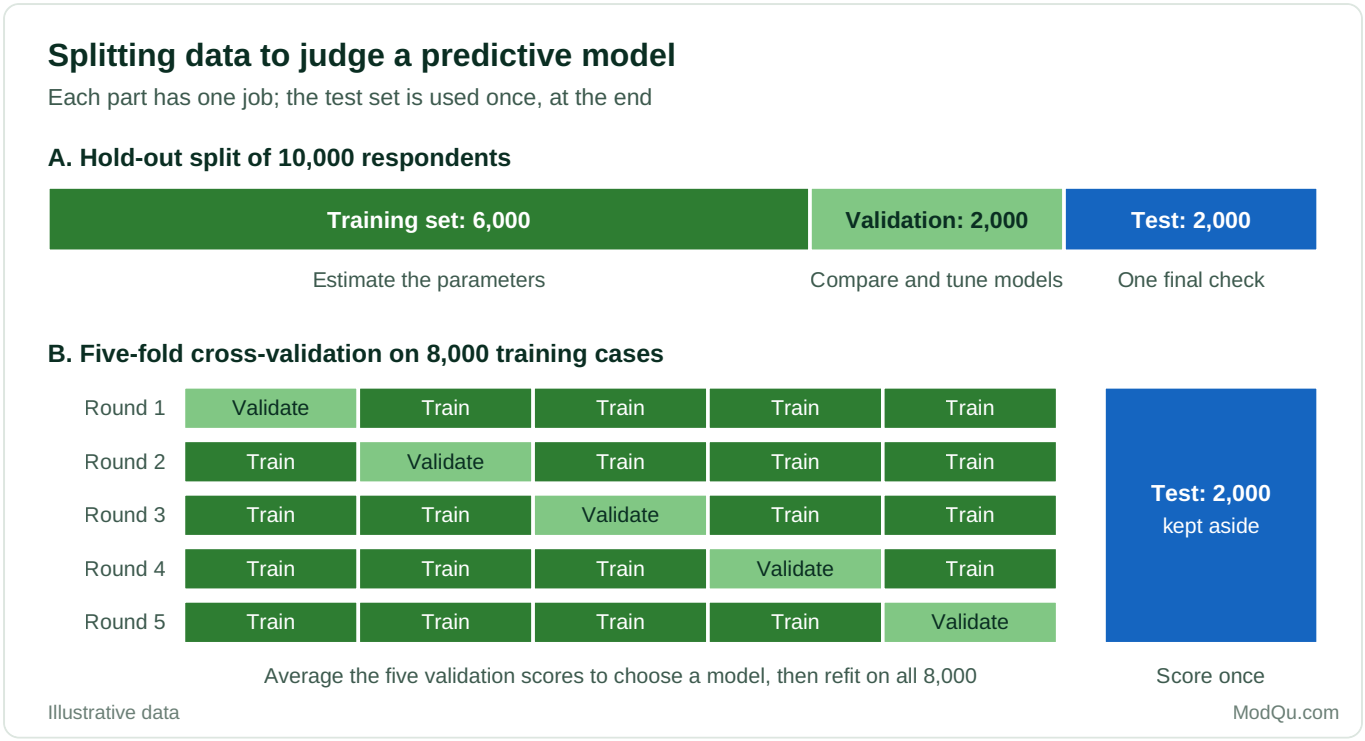



Figure 1. A hold-out split into training, validation and test sets (top) and five-fold cross-validation on the training data with the test set kept aside (bottom). Illustrative data; chart by ModQu.

## Cross-validation

When data are scarce, cross-validation reuses the data more efficiently (Stone, 1974). In *k*-fold cross-validation the training data are split into *k* equal folds, usually five or ten. The model is fitted *k* times, each time leaving out one fold and scoring the predictions on it, and the *k* scores are averaged. If cross-validation is used both to choose the model and to report its accuracy, the winning score is biased upwards; nested cross-validation, with an inner loop for tuning and an outer loop for assessment, corrects this (Varma & Simon, 2006).

## A worked example

Suppose, hypothetically, a subscription firm links 10,000 customer survey responses to whether each customer cancelled within six months. The analyst holds back 2,000 cases as a test set and runs five-fold cross-validation on the remaining 8,000. A large decision tree misclassifies 6% of training cases but 19% in cross-validation, a clear sign of overfitting. A logistic regression with eight predictors misclassifies 15% of training cases and 16% in cross-validation. The analyst chooses the logistic model, refits it on all 8,000 cases and scores it once on the test set: 16.4%. That figure, not 6%, is what to expect.

## Practical guidance for survey and business data

- **Split the way the model will be used.** If the model will predict future customers, train on earlier periods and test on later ones. Random splits of time-ordered, clustered or repeated-measures data let near-duplicates fall on both sides and inflate scores; blocked splits by period, region or respondent are safer (Roberts et al., 2017).
- **Prevent leakage.** Leakage occurs when information that would not be available at prediction time enters training, such as a variable recorded after the outcome (Kaufman et al., 2012). Do all preparation, including imputation, scaling and variable selection, inside each training fold.
- **Report uncertainty.** A test score from 2,000 cases still has sampling error; give an interval.

## Common mistakes

- Tuning repeatedly against the test set, which turns it into a second validation set.
- Selecting predictors on the full data and then cross-validating only the final fit.

## Key takeaways

A predictive model is only as good as its performance on unseen data. Use a training set to fit, validation or cross-validation to choose, and an untouched test set to report. Split the data the way the model will be used and keep every preparation step inside the training data.

## References

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7

Kaufman, S., Rosset, S., Perlich, C., & Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data*, *6*(4), Article 15. https://doi.org/10.1145/2382577.2382579

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., & Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography, 40*(8), 913–929. https://doi.org/10.1111/ecog.02881

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society Series B: Statistical Methodology, 36*(2), 111–133. https://doi.org/10.1111/j.2517-6161.1974.tb00994.x

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics, 7*, Article 91. https://doi.org/10.1186/1471-2105-7-91