Data Scientist & ML Engineer English Essentials

Must-know terms and phrases for data scientists and ML engineers — models, evaluation, pipelines.

https://coderslingo.com/resources/cheatsheets/data-scientist-essentials/

ML basics

model — A function learned from data that makes predictions on new inputs.
feature — An input variable the model uses to make a prediction.
label / target — The correct answer the model is trying to predict.
training — The process of fitting a model to data by adjusting its parameters.
inference — Using a trained model to make predictions in production.
supervised learning — Learning from labelled examples; unsupervised has no labels.
overfitting — When a model memorises the training data and fails on new data.
underfitting — When a model is too simple to capture the patterns in the data.
hyperparameter — A setting you choose before training (learning rate, tree depth), not learned from data.
epoch — One full pass over the training dataset.
gradient descent — The optimisation method that nudges parameters to reduce error.
loss function — The number measuring how wrong the model's predictions are.
regularization — Techniques that penalise complexity to reduce overfitting.
embedding — A dense numeric vector representing something like a word or user.
fine-tuning — Adapting a pre-trained model to a specific task with extra training.

Evaluation

accuracy — The fraction of predictions that are correct — misleading on imbalanced data.
precision — Of the items flagged positive, how many really were positive.
recall — Of the truly positive items, how many the model caught.
F1 score — The harmonic mean of precision and recall, balancing the two.
AUC / ROC — A threshold-independent measure of how well a classifier ranks positives above negatives.
confusion matrix — A table of true/false positives and negatives.
baseline — A simple reference model your model must beat to be worth deploying.
cross-validation — Splitting data several ways to estimate performance more reliably.
train/test split — Holding out data to test on examples the model never saw.
bias–variance trade-off — Balancing systematic error against sensitivity to the training data.
A/B test — Comparing two versions on live traffic to measure real impact.
RMSE / MAE — Common error metrics for regression — how far predictions are from actuals.
leakage — When information from the future or the target sneaks into training, inflating scores.

Data & pipelines

pipeline — An automated sequence that ingests, transforms and serves data or predictions.
ETL — Extract, Transform, Load — moving and reshaping data into a warehouse.
feature store — A central system for storing and serving features consistently to models.
data drift — When the input data distribution shifts away from what the model trained on.
concept drift — When the relationship between inputs and the target changes over time.
data warehouse — A central store optimised for analytical queries over large datasets.
feature engineering — Creating useful input variables from raw data.
imputation — Filling in missing values in the data.
normalization / scaling — Putting features on a comparable numeric range.
labeling — Annotating data with the correct answers for supervised training.
batch vs streaming — Processing data in scheduled chunks vs continuously as it arrives.
reproducibility — Being able to rerun an experiment and get the same result.

Presenting findings

statistical significance — Confidence that a result is unlikely to be due to chance.
p-value — The probability of seeing the result if there were truly no effect.
confidence interval — A range that likely contains the true value, with a stated confidence.
correlation vs causation — Two things moving together doesn't mean one causes the other.
effect size — How big a difference is, not just whether it's statistically detectable.
sample size — How many observations the conclusion is based on.
distribution — How values are spread across a range.
outlier — A data point far from the rest that can distort results.
hypothesis — A testable statement about what the data will show.
interpretability — How easily a human can understand why a model made a prediction.

Key phrases used at work

"The data suggests a clear correlation, but I'd caution against reading causation into it."

"The model beats the baseline on F1, though precision dipped slightly."

"This result is statistically significant at $p < 0.05$, with a 95% confidence interval of 2–4%."

"We're seeing data drift — the input distribution has shifted since we trained."

"Recall matters more than precision here, because missing a fraud case is costly."

"I suspect leakage — the score looks too good, so let me audit the features."

"In plain terms, the model is most confident when the customer has a long history."

"Let me caveat this: the sample size is small, so treat the numbers as directional."

"We optimised for recall at the cost of more false positives — that's the trade-off the business asked for."

"The headline number is encouraging, but the confidence interval is wide."

"To translate that into impact: this should reduce churn by roughly two percentage points."

"The model is overfitting — training accuracy is high but it falls apart on the holdout set."

"I'd recommend we A/B test this before rolling it out to everyone."

"These are early results from a single experiment, not a final recommendation."

"The feature importance tells us location is the strongest signal by far."

"Happy to go deeper on the methodology, but the takeaway is the lift is real and repeatable."

"We retrained after detecting drift, and performance recovered to the previous level."

"Let me walk you through the assumptions before we get to the conclusion."

"Yesterday: finished the feature engineering. Today: tuning hyperparameters. No blockers."

"Quick question before I proceed: do we care more about accuracy or interpretability for this use case?"