Data Science Tips and Tricks

DSKC Meetup, February 2025

Micah Melling

Micah Melling

Chief Data Scientist at Americo Financial

Former Sr. Director of Data Science at Spring Venture Group

Adjunct Professor at Park University

UMKC Analytics Board Member

Masters in Data Science from Rockhurst

Goal: Share useful - but sometimes lesser-known - tips and tricks for effective data science projects.

How can this impact us?

Implementing a nifty “trick” can lift model performance with minimal added effort.

Incremental model improvements can be important differentiators.

Why Can’t We Just Let GenAI Do All This for Us?

1. We must know the correct and possible questions to ask.
2. We must have the ability to evaluate and debug answers.
3. GenAI models produce a rough average of their training data - I want to strive for deep excellence at specific problems.
4. We still need to be able to think critically and be wary of cognitive offloading.
5. The following tips are less about knowledge and more about leveraging context and implementing a bespoke strategic approach (humans have an edge here).

GenAI can potentially help us on the way (e.g., serve as a better Stack Overflow) as long as we heed the above points.

Only Do Over- and Under-Sampling on Training Data

The above point means we should only perform this action during the training folds in cross validation. We should never change the distribution on any test set, including those in cross validation (script).

The Wide World of Hyperparameter Optimization

Unless we are building a very simple model, we likely should not use grid search or random search.

A powerful alternative is Hyperopt (script, video).

In fact, a great variety of methodologies exist to optimize model hyperparameters.

How to Tune Feature Engineering / Data Cleaning

We know that tuning model hyperparameters is important. However, we can also tune our feature engineering and data cleaning!

For example, do we get the best cross validation loss when we impute missing values with the mean, median, or a fancy imputation strategy (e.g., MICE)?

We can accomplish this aim with scikit-learn mixins (script, video).

Custom Cross Validation

In cross validation, we can control more than the number of folds!

For example, out of the box, scikit-learn provides a time series CV splitter.

We can also create our own custom CV splits (script, video).

Custom Loss Functions

When training models, we often use a standard loss function (mean squared error, log loss, F1).

However, we might desire custom behavior based on our problem.

> In a regression problem, we might be OK if we over-predict but not if we under-predict.

We can embody such preferences in a custom loss function (script, video).

Model Calibration

Except for logistic regression, out-of-the-box class probabilities are often not well calibrated (i.e., they cannot be directly interpreted as confidence levels).

In many cases, we should care more about predicted probabilities than predicted classes (e.g., if someone is predicted to churn with 51% likelihood, shouldn’t they be treated differently than someone who is predicted at 95%?).

Model Calibration

ML models are simply tools to solve problems; the goal is not to naively optimize predictive power but rather to maximize usefulness.

We have many tool at our disposable to calibrate models and assess their calibration (script, video, textbook chapter).

Comparing Models to a Heuristic

In data science we are often asked, “how good is our model?”. To business stakeholders, the most truthful test set metrics might be unintuitive (e.g., log loss, F1).

Alternatively, we can build a heuristic model - and even optimize it! - to give a better status quo baseline (script, textbook chapter - scroll to the bottom).

Upcoming Presentations

“Building a Data Science App”
February 22 at Park University
(Plaster Center from 9:30 am - 12:00 pm)

“AI, Philosophy, and Faith”
March 1 at the downtown Central Library
(Durwood Film Vault Room from 2:00 pm - 4:00 pm)

Get Involved with PrepKC!

https://prepkc.org/dia

Data Science Tips and Tricks

DSKC Meetup, February 2025

Micah Melling