Data-centric AI Recommendations

renumics

Getting relevant

Industrial AI Canvas

Think ML early by using the Industrial AI Canvas.

Perform EDA

Use EDA as a tool to understand your data and identify problems early on.

Renumics Spotlight

Check out Renumics Spotlight or a comparable tool which can help with data exploration immensely.

Feature Selection

Use automatic feature extraction/selection libraries.

Incorporate Domain Knowledge

To get a robust model and resolve problems, make sure you can use domain knowledge!

Productionizing

Integrate Early

Integrate your models into your system early and think of user interaction as a critical component.

Modeling for Insights

Model your problem early on to generate additional insights, then iterate.

Use ML Toolkits

Use (AutoML) Toolkits like FLAML for quick prototyping and combine them with interactive model evaluation to avoid garbage in garbage out.

Gather Feedback with Gradio

Gradio lets you deploy your models quickly and gather user feedback.

Find Outliers with PyOD

Use PyOD to identify edge cases and outliers.

Clean labels with Cleanlab

Identify inconsistent annotations with Cleanlab.

Treat Biases with Fairlearn

Fairlearn can help you detect biases in your data, especially if combined with interactive data visualizations.

Generate Insights with XAI

Use tools in the space of trustworthy AI such as SHAP to generate insights and understand problems quicker.

Staying relevant

Proactively revisit models

Proactively revisit your models using monitoring tools like Evidently.

Model serving

Use model serving like TorchServe.

Pipeline Management

Manage your pipelines e.g. with ZenML.

Model Validation

Use model validation tools such as Deepchecks.

Data Validation

Use data validation tools such as GreatExpectations.

Collboration & Engagement

Collaborate

Embrace collaboration between stakeholders like domain experts and data scientists to make your models robust.

EDA for Evaluation

Use the features of data exploration tools in showing data to the domain expert. It helps uncovering problems and patterns immensely.

Iterating Quickly

Iterate

Iterate on your data and models to become robust!

Track decisions

Make decisions traceable. This is really hard when just communicating in Emails or Verbally.

Automate and Review

Automate detection of data issues and review them subsequently in a smart way to allow scaling your process.

Use data versioning

Use data versioning tools such as DVC.

Version and track models

Version and track your models e.g. with MLFlow.

Best practices

Focus on data

Focus on improving your data at least as much as tuning your model!

Focus on label consistency

Label consistency significantly reduces the amount of data needed.

Labeling Instructions

Define labeling instructions to avoid inconsistencies. Adapt them if necessary.

Consensus Labeling

Set up consensus labeling to find errors by disagreement.

Use metadata

Metadata is super helpful and sometimes necessary to resolve data issues. Preserve it if possible!

Use Vector Representations

Use libraries for computing vector representations to uncover hidden patterns and problems like duplicates in your data.

Meaningful Metrics

Define meaningful metrics that really show how well your model performs on your use case.