Model Objective Validation

Avoiding risks associated with Design Objectives and Bias in Data

Microsoft

Objectives

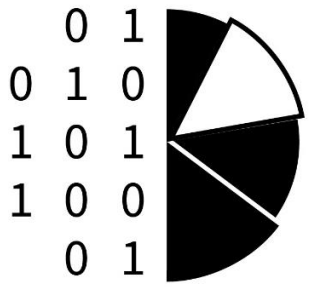
- How to determine the right set of metrics according to your business problem.
- How to build a baseline by using the defined metrics.
- How to identify weaknesses on the dataset. Evaluating fairness and representation on dataset.
- How to determine quality samples vs. target. i.e. miss-labeled samples.
- How to determine your Dev/Test strategies to ensure robustness of evaluation.
- Build baseline models.
- How to identify cases where you need to enrich dataset.

Agenda

- **Data Composition**
- **Data preprocessing or pre-formatting. (Before Feature Engineering)**
- **Data Use**
- **Model Baseline**
- **Questionnaire**

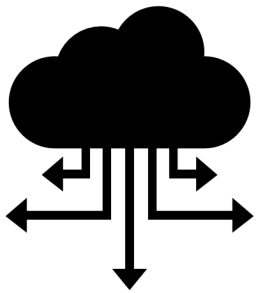
Data Ingestion| Composition

Composition



Data Processing

Use & Distribution

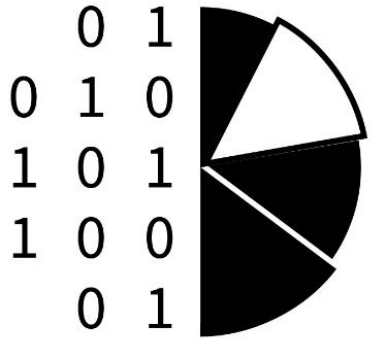


Model Baseline

Before Feature Engineering

Data Ingestion| Composition

Composition



Questions

- **Does the dataset contain data that might be considered sensitive in any way?**: (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data)
- **Does the dataset identify any subpopulations? If yes, please identify those variables**: This question can help in the identification of possible biases or under-representation of certain subpopulations. *Age, Sex, Gender and Race* are examples.
- **Is the dataset self-contained, or does it link to or otherwise rely on external resource?** The answer here can be *additional websites, tweets or other datasets*.

Data Ingestion| Collection Process

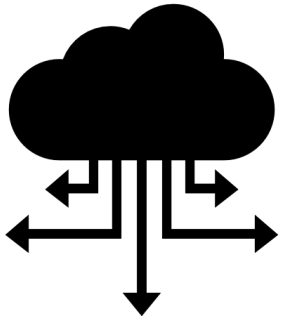
Data Processing

Questions

- **Do you have the documentation preprocessing/cleaning/labeling of the data? If so, provide link to code**: This question is based on the pre-processing of the data before storage. This is independend from the data preprocessing during *Feature Engineering*.
- **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data?** (e.g., to support unanticipated future uses)? This only applies to changes in data to be stored.
- **Are there tasks for which the dataset should not be used?**: Please specify whether this data cannot be used in other processes. This can be because several reasons: regulations, data not being adequate for other processes, etc...

Data Ingestion| Storage

Use & Distribution



Questions

- **Did you identify the set of metrics that will translate the business problem into Machine Learning?**: Specify whether the data is currently being used by other process or model. This is important, as any change that you do to this dataset might affect other models.
- **What is the structure of data?**: What is the structure of the dataset? *Tabular?, Non-Tabular? Graph**. More information on data types here

Data Ingestion| Collection Process

Model Baseline

Questions

- **Did you identify the set of metrics that will translate the business problem into Machine Learning?**: This is very important; we need to have a way of measuring the improvement of the Machine Learning models.
- **Do you have a business explanation why you chose those metrics?**: It is important to have performance metrics that translated your business problems into an optimization problems, so that each improvement in those metrics can be translated to actual improvements in your processes.
- **How are you evaluating the overall consistency labels (target) vs. input data? How are you identifying wrong labels or outputs?**: it is important to identify possible outliers, you can do this by using an **anomaly detection model**.

Data Science Questionnaire for Data Ingestion

Link:

Microsoft