Machine Learning

|Main Challenges – Bad Data

Prof. Dr. Mostafa Elhosseini

Professor of Smart Systems Engineering

https://youtube.com/drmelhosseini

Main Challenges

- Since your main task is to select a model and train it on some data, the two things that can go wrong are:
  - bad model and
  - bad data

Bad Data

Poor-Quality Data:

- Data with errors, inconsistencies, or noise can mislead the training process. Clean, well-preprocessed data is essential for effective model training. For example, typos in customer records or wrong measurements in scientific data.
- **Incomplete Data**: **Missing values** or incomplete records can significantly affect the quality of data. Incomplete data can lead to biased analyses or models if the missingness is not random.
  - 5% of your customers did not specify their age
  - Ignore this attribute altogether, ignore these instances, fill in the missing values (e.g., with the median age), or train one model with the feature and one model without it

Bad Data

Poor-Quality Data:

- Inconsistent Data: This occurs when there are discrepancies in the data, often due to different data sources or data entry standards. For instance, varying formats for dates (DD/MM/YYYY vs. MM/DD/YYYY) or inconsistent naming conventions can create inconsistencies.
- **Outdated Data**: Data that is not current or updated can lead to incorrect conclusions or predictions, particularly in rapidly changing domains like market trends or technology.
- It is often well worth the effort to spend time cleaning up your training data

Bad Data

Irrelevant Features:

- **Irrelevant features** in data refer to attributes or variables in a dataset that do not contribute useful information for a specific analysis, prediction, or decision-making process.
  - features that don't have a meaningful relationship with the target variable or outcome of interest
- **Feature engineering** is the process of using domain knowledge to create or transform variables in a dataset to make them more suitable and informative for use in predictive modeling. It involves **selecting**, **modifying**, and **creating** features to improve the performance and accuracy of machine learning algorithms

Bad Data

Irrelevant Features:

Feature engineering

- **Feature Creation**: Developing new features from the raw data. This might involve combining two or more variables, extracting parts of a date or time stamp, or converting categorical variables into numerical variables through encoding.
- **Feature Transformation**: Modifying features to enhance their relationship with the target variable. Examples include normalization (scaling all numeric variables to a particular range) and transformation (like taking the log of variables to handle skewed data).

Bad Data

Irrelevant Features:

Feature engineering

- **Feature Selection**: Identifying the most relevant features for the model. This involves removing irrelevant, redundant, or highly correlated features that might not contribute to or might degrade the model's performance.
- **Feature Extraction**: In cases like text and image data, extracting features involves converting raw data into a set of representative features. For example, converting text into vectors using techniques like TF-IDF or word embeddings.

Bad Data

Irrelevant Features:

Importance of Feature Engineering

- Improves Model Performance: Well-engineered features increase the predictive power of the model.
- Reduces Model Complexity: By selecting only relevant features, the complexity of the model can be reduced, making it faster and less resource-intensive.
- Enhances Model Interpretability: Models built with well-thought-out features are often easier to interpret and understand.
- Adapts Data to Model Requirements: Different models have different requirements regarding input data. Feature engineering tailors the data to these requirements.