Data Preparation

Tuesday, September 3, 2024 7:43 PM

Data Prep Introduction

Data Preparation: the process of cleaning and transforming raw data to prepare it for analysis. Proper data preparation can prevent errors before processing, improve data quality, produce better analytical results, lead to better business decisions, future proof models, improve scalability, facilitates feature engineering, among other benefits. Putting in a lot of effort during the data preparation stage can save time and effort down the line during the modeling and maintenance phases.

Common steps taken during the data preparation stage include:

1. Gathering the right data
2. Assess the data to understand what further steps need to be taken to make the data as useful as possible for your particular modeling goal
3. Clean the data:
    a. Handling extraneous data and outliers, potentially by removing them
    b. Filling in missing values
    c. Conforming data to standardized scale
    d. Masking private or sensitive data entries
    e. Identify potential issues with the data and fix what is fixable
4. Transforming and enriching the data: improving the dataset by adding new features or columns, increasing accuracy and reliability, and verifying the data
    a. Labeling the data
    b. Updating the format or value entries
    c. Enrich the data with additional context or information
5. Validate the data: ensure data accuracy, completeness, and consistency and align with the goals of your modeling problem
    a. Keep track of flaws or discrepancies
    b. Correct mistakes and maintain and audit record of modifications
    c. Automate data validation to ensure consistent maintenance of the model
    d. Choose a representative sample of the dataset for validation

Outliers

Outlier: a data point that is very different from the rest of the data. There are different kinds of outliers, including:

- Point outlier: a single data point that deviates significantly from the overall distribution of a dataset
- Collective outlier: a set of data points that collectively deviate significantly from the overall distribution of a dataset
- Contextual outlier: one or more datapoints that may not be outliers when considered in the entire dataset, but they exhibit unusual behavior within a specific context or subgroup

There are different methods for identifying outliers. Box-and-Whisker plots are effective for finding outliers in a single dimension. For multidimensional models, it is more difficult to identify outliers, but one method is to fit a model and then look at points that have very large error; these points may be outliers.

How to handle outliers within your dataset depends on the context of why the outlier is there. Sometimes outliers are the result of bad data (poor data collection, incorrect data input, contaminated experiment, etc.).

- In such cases, the data can safely be removed from the dataset or imputed.

But sometimes outliers are real data, real deviations in the dataset's distribution. Understanding why and how these outliers came to be is an important part of getting an overall understanding of your dataset. Investigating real outliers could include finding:

- Where the data came from
- How the data was compiled
- Unique situations that impacted the data

Removing real data outliers may result in too optimistic evaluation. In large datasets, it's basically guaranteed that outliers will exist. There are different methods for handling real outliers and the correct approach will depend on your modeling goal. Some common methods include:

- Reducing the weights of outliers
- Changing the values of outliers using imputation or other methods
- Using estimation techniques
- Removing outliers (in certain cases where keeping the outlier in would negatively impact the model)