Mohsen Kazemian @mokazemian

How should outliers be handled in machine learning?

Swipe →

DATA MAVERICKS

An outlier is like the unusual friend in a group—it's a data point that stands out as very different from the rest.

These outliers can be unusually high or low values and are often caused by errors or unique events.

We need to spot and handle them because they can mess up our data analysis and model predictions.

Swipe →

WHY IS HANDLING OUTLIERS ESSENTIAL?

- **Accuracy Improvement**: Handling outliers enhances model accuracy and performance.

- **Data Quality Enhancement**: Outliers often result from data errors, and addressing them improves data quality.

- **Generalization Improvement**: Dealing with outliers helps models generalize better to new data.

Swipe →

WHY IS HANDLING OUTLIERS ESSENTIAL?

- **Distribution Normalization:** Outliers can distort data distributions; addressing them helps achieve normality.

- **Sensitivity Reduction:** Some models (like linear regression) are sensitive to outliers; treatment makes them more robust.

- **Interpretability:** Outliers can affect model interpretability; addressing them enhances it.

Swipe →

WHY IS HANDLING OUTLIERS ESSENTIAL?

- **Consistent Performance:** Treating outliers ensures model performance consistency over time.

- **Feature Engineering Insights:** Handling outliers provides insights for meaningful feature engineering.

- **Scaling and Normalization:** Outliers can affect data scaling and normalization processes.

Swipe →

BUT NEVER FORGET THAT

In some cases, outliers are relevant to the problem and should not be removed.

Handling outliers, which may represent critical information, allows the model to learn from them without being unduly influenced.

Swipe →

SO, WHAT DO YOU NEED?

- Understand your dataset.
- Possess domain knowledge or consult an expert.
- Be familiar with your model's idiosyncrasies.

Now, let's explore how to deal with outliers >>

Swipe →

FIRST STEP

The initial step is to spot and get a good look at those outliers.

You can run some number crunching using techniques like Z-Scores or the IQR (Interquartile Range).

Alternatively, you could tap into your domain know-how to sniff out these misfits.

If you're a more visual learner, you can always throw your data into a histogram, scatterplot, or a trusty old box plot to make things clearer.

Swipe →

SECOND STEP

The next step is to put your domain Knowledge to use, figuring out whether you want to **keep** those outliers, **kick** them to the curb, or **make** some adjustments to make them mesh better with your data.

It's all about making a **judgment call** based on what works best for your specific situation.

Swipe →

THIRD STEP

Now that you've got your data treatment plan locked in, it's time to put it into action.

In this section, we'll dig into eight different methods to see how we can make your step two decision, work.

So, let's dive right into these techniques! >>

Swipe →

REMOVE OUTLIERS

1

One common approach is to remove outliers from the dataset.

However, this should be done carefully, as removing too many outliers can lead to a **loss of information**.

You can use techniques like the Z-Score, IQR, or domain-specific knowledge to decide which data points to remove.

Swipe →

TRANSFORM DATA

2

Transforming data using mathematical functions can help in making outliers less influential.

For example, taking the logarithm or square root of data can help normalize skewed distributions.

Swipe

WINSORIZATION 3

In winsorization, you replace extreme values with the nearest values that are not outliers.

For instance, you could cap outliers at a certain percentile, such as the 95th percentile.

Swipe →

BINNING 4

Group data into bins or categories, which can help reduce the impact of outliers.

For instance, if you have a continuous variable, you can bin it into discrete intervals.

Swipe

FEATURE ENGINEERING 5

Create new features based on existing ones, which can help in reducing the impact of outliers.

For example, you could create a binary feature that flags whether a data point is an outlier or not.

Swipe →

ROBUST ALGORITHMS 6

Some machine learning algorithms are inherently robust to outliers.

Algorithms like Random Forests and Gradient Boosting Trees are less sensitive to outliers compared to linear models.

Swipe →

MODEL-BASED APPROACHES 7

You can also use models that are specifically designed to handle outliers, such as robust regression models or support vector machines with robust kernels.

Swipe →

IMPUTATION 8

In some cases, rather than removing outliers, you may choose to impute their values with a more reasonable estimate, such as the median or mean of the non-outliers.

Swipe →

It's essential to choose the appropriate method for handling outliers based on the **characteristics of your data** and the **goals of your machine learning project.**

Additionally, always carefully assess the impact of outlier treatment on model performance through **cross-validation** or other evaluation methods.

Swipe →

WAS THIS HELPFUL?

dont forget to save this post