Data Leakage in Machine Learning

Abstract

Data leakage happens when a machine learning model accidentally uses information it shouldn't have during training. This makes the model appear too good on test data but fails in real-life applications. This paper explains what data leakage is, its types, effects, and how to avoid it in simple terms.[1][2][3]

Introduction

In machine learning, we train models using data to predict outcomes such as customer churn or house prices. Data leakage is a common mistake where the training data includes hints about the answers (targets) that won't exist in real use. It deceives the model, giving a false impression of high accuracy.[2][4][5]

For example, if predicting tomorrow's weather but training includes tomorrow's data, the model "knows" the answer already.[6]

Types of Data Leakage

There are main types of data leakage.

- **Target Leakage**: Features directly tell the target. Like using "customer left = yes" as a feature to predict leaving.[7][1]
- **Temporal Leakage**: Future data mixes into past training. Common in time series, like stock prices, use the next day's info.[8][2]
- **Preprocessing Leakage**: Fixing data (like scaling) on the full dataset before splitting train/test. Test info leaks into training.[1][6]

Other cases include duplicate data in the train and test sets.[2]

Type	Example	Why It Happens
Target	Target in features [4]	Wrong data prep
Temporal	Future data in train [8]	No time split
Preprocessing	Scale before split [1]	Full dataset use

Effects on Models

Leakage makes models seem perfect (99% accuracy) but flop in production. It wastes time, money, and trust. Biased predictions harm decisions, like wrong medical diagnoses.[3][7][1]

Prevention Methods

Prevent leakage with good habits.

- Split data into train/test first, then preprocess each separately.[6][2]
- Use time-based splits for sequences: train on past, test on future.[8]
- Check features: remove any that reveals targets.[4]
- Validate pipelines end-to-end with holdout sets.[1]

Tools like cross-validation help spot issues early.[5]

Conclusion

Data leakage is easy to miss but ruins ML projects. By splitting data right and checking features, students and pros can build reliable models. Always test as if in the real world.[3][7]

⁂

1. https://airbyte.com/data-engineering-resources/what-is-data-leakage
2. https://ibarrond.github.io/data-leakage
3. https://builtin.com/machine-learning/data-leakage
4. https://towardsdatascience.com/data-leakage-in-machine-learning-6161c167e8ba/
5. https://www.machinelearningmastery.com/data-leakage-machine-learning/
6. https://www.sailpoint.com/identity-library/data-leakage
7. https://megaladata.com/blog/data-leakage-machine-learning
8. https://www.geeksforgeeks.org/machine-learning/what-is-data-leakage/
9. https://www.kaggle.com/code/alexisbcook/data-leakage
10. https://www.innovatiana.com/en/glossary/data-leakage