Data Cleaning: The Most Important Step in Machine Learning

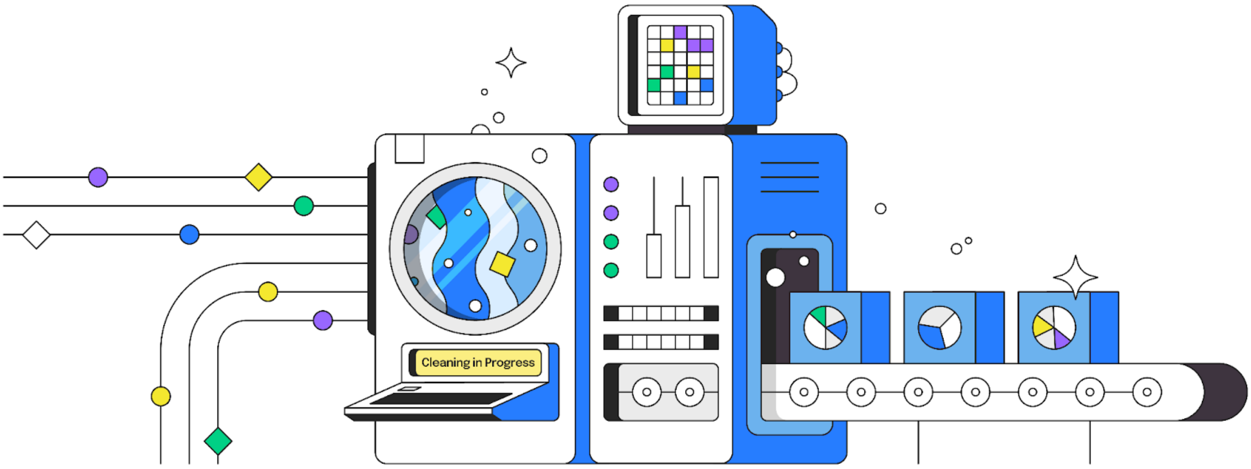
obviously.ai/post/data-cleaning-in-machine-learning

Data enrichment, data preparation, data cleaning, data scrubbing—these are all different names for the same thing: the process of fixing or removing incorrect, corrupt, or weirdly formatted data within a dataset.

But what does good, clean data look like? It's more than just reorganizing some rows and calling it a day. We asked our team of data scientists to tell us what exactly they mean when they tell you to prep, clean, or enrich your data. Here's what they had to say.

What is Data Cleaning?

First things first: let's define data cleaning.



Data cleaning is the process of preparing data for analysis by removing or modifying data that is incorrect, incomplete, irrelevant, duplicated, or improperly formatted.

But, as we mentioned above, it isn't as simple as organizing some rows or erasing information to make space for new data.

Data cleaning is a lot of muscle work. There's a reason data cleaning is the most important step if you want to create a data-culture, let alone make airtight predictions. It involves:

- Fixing spelling and syntax errors
- Standardizing data sets
- Correcting mistakes such as empty fields
- Identifying duplicate data points

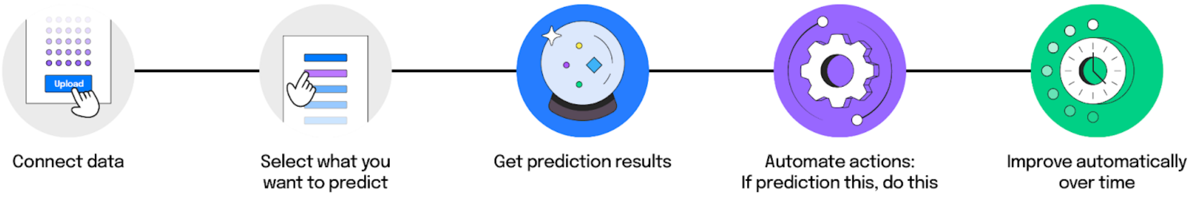
It’s said that the majority of a data scientist's time is spent on cleaning, rather than machine learning. In fact, 45% of data scientist's time is spent on preparing data.

And to us, that makes sense—if there’s data that doesn’t belong in your dataset, you aren’t going to get accurate results. And with so much data these days, usually combined from multiple sources, and so many critical business decisions to make, you want to be extra sure that your data is clean.

Why is Data Cleaning so Important?

Businesses have a plethora of data. But not all of it is accurate or organized. When it comes to machine learning, if data is not cleaned thoroughly, the accuracy of your model stands on shaky grounds.

We've talked about how no-code simplifies the traditional machine learning process. What is typically a 10-step process instantly becomes a much simpler route with platforms like Obviously AI.


Upload
Connect data
Select what you want to predict
Get prediction results
Automate actions: If prediction this, do this
Improve automatically over time

But the traditional and no-code process still require the same important first step: connecting to your data. And all that time you've saved by opting for a no-code tool won't matter if those large stacks of data aren't properly organized, formatted, or accurate.

Preparing your data helps you maintain quality and makes for more accurate analytics, which increases effective, intelligent decision-making.

These are the kinds of benefits you’ll see:

- Better decision making
- Boost in revenue
- Save time
- Increase productivity
- Streamline business practices

Related Reading: The Ultimate Guide to Machine Learning

I’ve Already Cleaned My Data Myself. Can I Start Predicting?

The short answer is no. Unless you are trained in data science, or already know how to prepare your dataset, we typically advise against jumping right in.

The best way to explain this is with an analogy.

Most people know how to drive a car. But not everyone knows how to drive a race car. Driving a regular car and driving a race car are two very different things. But because they have the same fundamentals (press gas, brake, turn wheel), there are those who think that they'll be able to successfully drive a race car at the racetrack.

Race cars, however, put out raw power and it is largely up to the driver to filter that force as the car is driven. So, while you could absolutely go out on that racetrack and drive that race car, chances are, you won't be able to drive it well or get the most out of your experience. You might even crash.

The same thing can be said about data. Everyone knows how to operate an excel spreadsheet. But oftentimes, the dataset in that spreadsheet isn't set up for building machine learning models.

Let's say you're trying to predict housing prices. You have a lot of data on sellers: their demographics, the amount they sold their house for, etc. You might also have data that appears to be irrelevant to what you want to predict. But that outlier may be crucial to your predictions. And machine learning will catch that.

This is why we always advise meeting with our team first before getting started on your first predictions and what we mean when we say that data cleaning is more than just formatting spreadsheets. And typically, we find that most people that go through onboarding have the most success when they see how data needs to be prepped.

How to Clean Your Data

Once you know what to look out for, it's easier to know what to look for when prepping your data. While the techniques used for data cleaning may vary depending on the type of data you're working with, the steps to prepare your data are fairly consistent.

Here are some steps you can take to properly prepare your data.

1. Remove duplicate observations

Duplicate data most often occurs during the data collection process. This typically happens when you combine data from multiple places, or receive data from clients or multiple departments. You want to remove any instances where duplicate data exists.

You also want to remove any irrelevant observations from your dataset. This is where you data doesn't fit into the specific problem you're trying to analyze. This will help you make your analysis more efficient.

2. Filter unwanted outliers

Outliers are unusual values in your dataset. They're significantly different from other data point and can distort your analysis and violate assumptions. Removing them is a subjective practice and depends on what you're trying to analyze. Generally speaking, removing unwanted outliers will help improve the performance of the data you're working with.

Remove an outlier if:

- **You know that it's wrong.** For example, if you have a really good sense of what range the data should fall in, like people's ages, you can safely drop values that are outside of that range.
- **You have a lot of data.** Your sample won't be hurt by dropping a questionable outlier.
- **You can go back and recollect.** Or, you can verify the questionable data point.

Remember: just because an outlier exists, doesn't mean it is incorrect. Sometimes an outlier will help, for instance, prove a theory you're working on. If that's the case, keep the outlier.

3. Fix structural errors

Structural errors are things like strange naming conventions, typos, or incorrect capitalization. Anything that is inconsistent will create mislabeled categories.

A good example of this is when you have both "N/A" and "Not Applicable." Both are going to appear in separate categories, but they should both be analyzed as the same category.

4. Fix missing data

Make sure that any data that's missing is filled in.

A lot of algorithms won't accept missing values. You may either drop the observations that have missing values, or you may input the missing value based on other observations.

5. Validate your data

Once you've thoroughly prepped your data, you should be able to answer these questions to validate it:

- Does your data make complete sense now?
- Does the data follow the relevant rules for its category or class?
- Does it prove/disprove your working theory?

Data Prep Checklist: The Basics

Obviously AI requires a structured dataset to get meaningful prediction outcomes.

We made a quick DIY checklist to ensure your data is well structured and machine learning ready. It was prepared by the data science team at Obviously AI, so you know it's comprehensive.

- Dataset must have at least 1,000 rows
- Dataset must have at least 5 columns
- The first column must be an identifier column, such as a name, customer_id, etc.
- The first row should be column names
- The data should be aggregated in a single file or table
- The data must have as less missing values as possible
- No personally identifiable information is required, such as phone numbers, addresses, etc.
- No long text phrases—only use discrete values for text columns

What columns should I bring in my dataset?

A training dataset that's machine learning ready typically contains several types of columns (features), while you don't need them all, having as many as possible can help make better predictions.

Here's a list of most common column types:

- **Identifier column**: Anything we use to distinguish a customer from another. Only *one* is required. (e.g. User ID, Name, Customer ID, etc.)
- **Demographic columns**: Any columns with demographic data that relates to the user OR the line item in the row. (e.g. Age, Location, Income, etc.)
- **Product/Usage columns**: Any columns that record activity done by the customer on your product OR details of their account. (e.g. Number of sessions, Account type, etc.)
- **Transactional columns**: Any columns with details on transactions done by the customer. (e.g. Monthly charges, Payment method, Contract length, etc.)
- **Prediction column**: Data of historical activity that you would like to predict. (e.g. Churn, Lead status, Sales, Revenue, etc.)

To learn more about the type of columns, check out the following links:

Summary

Data cleaning is an extremely vital step for any business that is data-centric. Businesses that take proper care of their datasets are rewarded with high-quality predictions and are able to make leaps ahead of their competition. With clean and organized data, you can predict anything—from customer churn to hospital stay to employee attrition.

Obviously AI has a team of data scientists that become an extension of your team helping you make your datasets machine learning-ready. Book a demo with us today to learn more about how our dedicated data scientists team can help you get your data machine-learning ready.