Salma Ghoneim Follow

Apr 7, 2019 · 6 min read ★ · Listen

Save

5 Steps to correctly prepare your data for your machine learning model.

How to prepare your dataset in order to get the most out of it?

Photo by Brett Sayles from Pexels

previous e

> *"Data is the new oil." — **Clive Humb —** Chief Data Scientist and Executive Director of Starcount*

Data is at the core of nearly every business decision made.
Human resources directors are gathering data from online resources to determine the best people to recruit and confirm details about them. Marketing departments are lasering in on market segmentation data to find consumers who are ready to buy, speeding up the sale-closing process whenever possible.
Business executives must examine bigger trends in the market, such as changes in pricing of resources, shipping or manufacturing.

Your project is only as powerful as the data you bring.

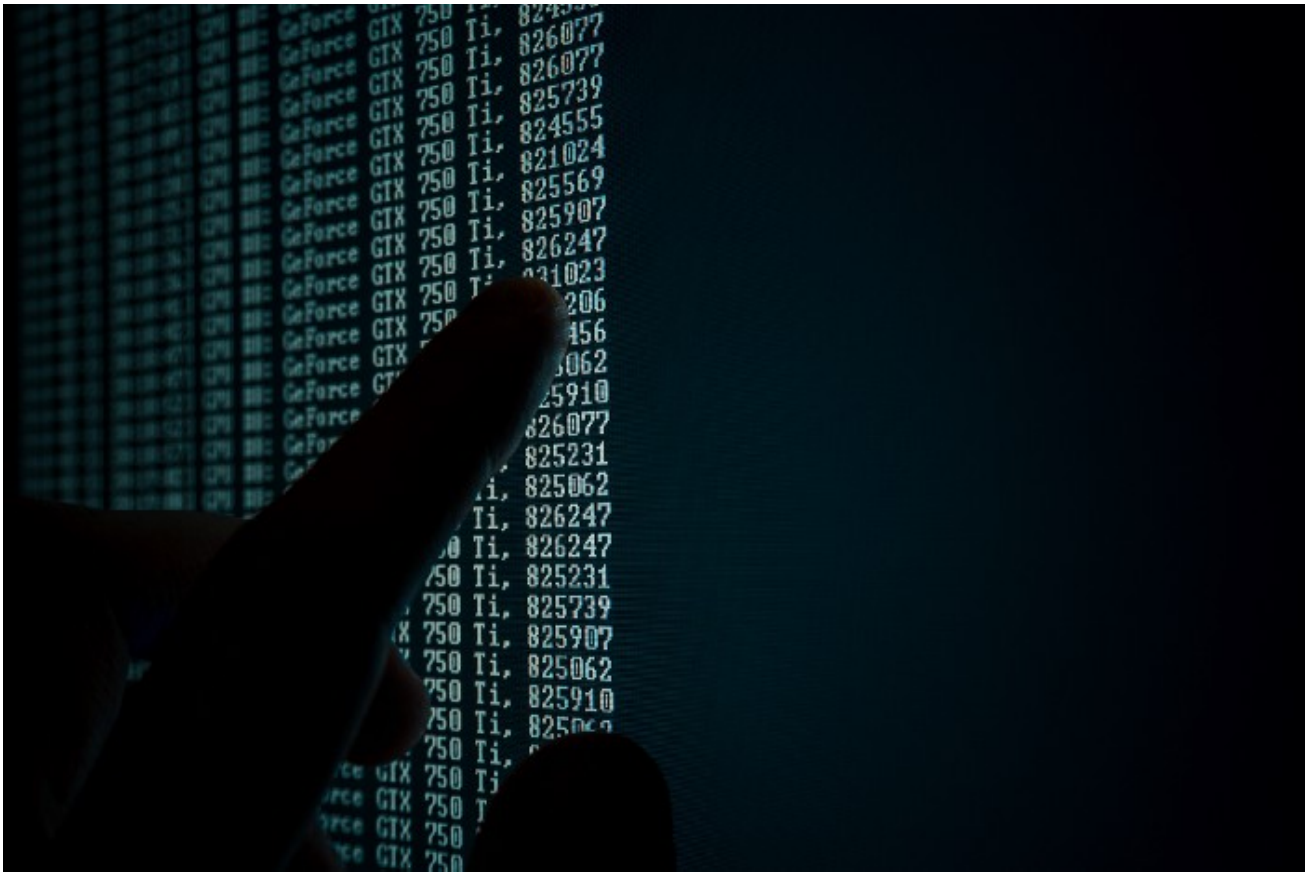
Step 1: Gathering the data

The choice of data entirely depends on the problem you're trying to solve.
Picking the right data must be your goal, luckily, almost every topic you can think of has several datasets which are public & free.
3 of my favorite free awesome website for dataset hunting are:

1. Kaggle which is so organized. You'll love how detailed their datasets are, they give you info on the features, data types, number of records. You can use their kernel too and you won't have to download the dataset.
2. Reddit which is great for requesting the datasets you want.
3. Google Dataset Search which is still Beta, but it's amazing.
4. UCI Machine Learning Repository, this one maintains 468 data sets as a service to the machine learning community.

The good thing is that **data is means to an end,** in other words, the quantity of the

It is a libr[illegible] favorite parser to provide idiomatic ways of navigating, searching and modifying the parse tree. It commonly saves programmers hours or days of work.



pexels

Step 2: Handling missing data

This is one of the hardest step and the one that will probably take the longest unless you're lucky with a complete perfect dataset, which is rarely the case. **Handling missing data in the wrong way can cause disasters.**

Generally, there are many solutions such as:

- null value replacement
- mode/median/average value replacement
- deleting the whole record

- Forwa
- Multiple imputation

But you have to be smart.
For example, you're working in a traveling agency and you've been collecting data about travelers, for some reason, around 5% of your dataset is missing the traveler's nationality. Filling those missing values with null values is a terrible mistake. Cause a person's nationality greatly affects the traveling papers & procedure that they have to go through on traveling. In this case, nationality is a sensitive area. You're better off deleting the whole record.

Age, on the other hand, can be safely replaced by the average value. It isn't that sensitive in this case.

Step 3: Taking your data further with feature extraction

Feature extraction can be a turning point for you. It is what makes a dataset unique. Getting insight by making relations between features is an outstanding creative thing.

For example, you're still casually doing your job at the travel agency, **They asked you to create a model to prioritize your clients using clustering.**
You have 3 features in your dataset:

- **Ticket request date**: The date a person requests their ticket(s).
- **Departure date**: The date a person wants to travel on.
- **Return date**: The date of returning.

You know you can't feed your model non-numerical data, still, you want to include them.
Some of the insights the data scientist in you can observe:
1) **Stay duration**: Difference between departure day and return date. That's the duration they're staying.
2) **Request duration**: Difference between the arrival date and the request date.


they're ou

Applying this concept (closeness of departure date) on the other two can be faulty though. When was the last time you booked a flight a year in advance? **This person could be a regular customer** who travels yearly with the agency. That would make him a priority too. Maybe it's time to send them some free kilometers.

Step 4: Deciding which key factors are important

Now, this one is tricky. Originally this should be the model's job. You could simply dump the whole dataset you have and just let the artificial intelligence be intelligent. AI is able to decide which features truly affect the output and which doesn't. On the downside, The more data you give your model, it costs you money (computer power) & time. Both not always available. So, Giving your program a little help isn't always a bad idea. If you're sure that a certain feature is completely unrelated to the output, you should just disregard it altogether.

Step 5: Splitting the data into training & testing sets

The famous rule of splitting the data is 80–20 percent training & testing sets respectively. Sometimes, that 20% for the test set should be engineered in a way that they're not just randomly cut out of the dataset.

As an example, you're working in your travel agency and they want you to create a program to detect whether a person will book a ticket this year or not.

So you go on with your data collecting of all travelers since the agency ever started, processing your data. It's time to split your dataset now. You get the feeling that you want to cut the test set in a way that it spans all years. Is that really smart though? If you want to predict whether a person will travel or not in 2019 why care whether he(or similar people) traveled in 2008? A lot of global changes had happened since then. Maybe the country they're from is having problems with your country, maybe it's not going so well in his country. In this case, **a smart thing is to engineer the test set for it to be in most recent years only.**

Your training set may contain all data from across the years, cause the happier the merrier. **But it's a wise idea to give weights according to the year.** A client record from 2008 should have a lighter weight than a record from 2018.

each grou del on the other groups and evaluate your model on this mini-test set, retain the evaluation score and discard the model, then move on to the next group.

Bottom line is, In order to get most out of your data, you have to look further into it & into the business as a whole. Not just applying techniques & strategies you learned rather than whether these strategies go well with your case or not.
