The Complete Guide to Data Acquisition For Machine Learning

EXPLORIUM

Glossary

Data Acquisition For Machine Learning

Finding a new data source that actually improves your model's accuracy and performance can feel like finding the tiniest needle in the largest haystack. While algorithm improvement has diminishing returns, finding better data and thus better features (the tiny needle) has the potential to be a breakthrough in model improvement.

When trying to improve a model's accuracy and performance there are two paths you can follow:

1. **Algorithm improvement:** Fine-tuning the algorithm a bit further (A.K.A. hyperparameter tuning). Although this may be more convenient (and cheap), at some point, it will only get you a fraction of a percentage of improvement before you hit a wall. You might turn to even darker magic (e.g. stacking, multi-level ensembles, etc.), but doing so would hurt your model's transparency and explainability.

2. **Data improvement:** Generating, testing, and integrating new features from various internal and/or external sources. This process is time-consuming, difficult, and more "artistic." But — and that's a big but — it could be a major discovery and move the needle much more.

For this whitepaper, we will focus on data improvement and walk you through different steps in the process of finding, testing, and acquiring external data sources that could potentially provide an uplift in your machine learning models. Those data sources should then be matched and joined to your current data to extract new features that will (if all goes to plan) give your model a wider context.

If you do find a valuable data source, it could mean a major breakthrough that provides more value and ROI than any other method you would use to improve a model.

The process of data acquisition can be broken down into six steps:

Step one:

Hypothesizing

At first, you’ll have to use your domain knowledge, creativity, and familiarity with the problem to try and scope the types of data that could be relevant to your model. This means thinking about what data you currently have that could be impactful. For this step, you should try and think about data that would be “orthogonal,” meaning that it would genuinely give you additional information.

For example, if you already have credit score data from a credit bureau, chances are that a credit score from another bureau will be obsolete and not add anything to your model. However, perhaps the academic background of a person would be relevant and impactful.

Obviously, this process could lead to hundreds of different ideas. You should try and focus on the one with the most potential.

Step two:

Generating a list of potential data providers

In this step, you’ll create a shortlist of sources (data partners, open data websites, commercial entities) that actually provide the type of data you hypothesized would be relevant. This step includes scouring the web and starting to engage with data partners through their website. Depending on the type of data you’re after, this can take a long time.

In a case where you’re dealing with commercial data providers (and not open data sources), always try to generate multiple options. This is crucial, especially if the project you’re working on will end up in production. Even if you end up choosing one provider, having multiple options will allow for fallback (e.g. in case a data provider disappears/in production downtime), competitive pricing, and avoid vendor lock-in.

Generally speaking, there are two types of sources out there we can acquire:

1. **Open and publicly available data:** These are datasets that are publicly available. Some of them are easy to download and are usually distributed in the form of a file (e.g. csv) and not an API. A good example of this is the U.S census. It may be more convenient to search for these sources in open data aggregators like Google BigQuery public data.

2. **Premium data providers:** These are data companies that provide all kinds of data. Some of it is highly restricted and regulated like credit scores, and some of it is collected, aggregated, and cleaned from open sources like social network data.

For data about individuals (enriching data about emails, for example), it's better to choose a data partner than to try to collect it yourself from an open source. Data providers can help you stay compliant. And, generally speaking, data about individuals is not as open and accessible as, say, geospatial data.

At the end of step two, you should have a list of potential data providers that might be relevant/impactful for the project you're working on.

Here is a table that can help you understand which type of data should you go after:

	Premium Data Provider	Open Data
Advantages ↓	• Cleaned • Updated • Unique and proprietary (sometimes)	Usually: • Free • One click to download and use • License: less limitation on what you can do with the data • Can hold data for eternity • No damage in case of breach • Complete dataset (and not samples) — good for aggregations

Disadvantages	• Could be expensive depending on the type of data and volume • Usually won't give you access to the complete dataset (only by API queries) • Some providers may require being a controller of your query data (save the query data to improve the service), which could be problematic for GDPR • Trust and integrate with an external API for runtime. Downtime of the provider's API could cause your prediction service damage • Vendor lock-in	• Not updated • Not proprietary or unique • Sometimes messy/uncleaned • No API to integrate with real-time systems
Integration type	• API • Batch	• Batch
Types of data it is best for	• Companies • People	• Aggregated statistics • Geospatial data

Step three:

Data provider due diligence

Due diligence is an absolute must. The list of parameters below will help you disqualify irrelevant data providers before you even get into the time-consuming and labor-intensive process of checking the actual data.

Trust us, you don't want to skip this step. The list below was generated through lessons learned the hard way.

1. **Historical snapshots:** While not true for every use case, historical snapshots are critical to training a machine learning model. For example, let's say you're trying to

build a model that would predict if a startup is going to succeed. To do this, you're feeding the model examples of startups that went through an IPO and others who went bankrupt. In order for the model to not "cheat" (A.K.A data leakage), you have to feed the model with data that used to be correct when the organization was a startup, and not a full-blown public company (e.g. the number of website visitors, etc).

2. **Transparent and compliant:** Unfortunately, a lot of data companies out there are acquiring their data through non-compliant measures. Make sure the data provider is transparent about how it sources the data. Then, make sure those sources can be used for the purpose you're trying to use them for. For some use cases, especially finance-related, specific types of data are forbidden to use. For example, the Fair Lending Law in the U.S. only allows specific types of data to be used.

3. **Processor versus controller:** Some data providers keep the data you send them through an API and use this data as part of their offering. Make sure you're aware of the way the provider collects data from you, not only from others.

4. **Latency:** In the case of real-time use cases, make sure the provider has sufficient capabilities from a latency point of view. Ask the provider to commit to a certain QPS (queries per minute).

5. **Updating data:** Data, just like the world around us, changes all the time. Make sure to dig deeper into the cadence in which the data provider updates its data.

6. **Downtime handling:** In production, everything is certain to break at some point. Even Facebook and Whatsapp have production downtime once in a while. Ask the data provider about historical incidents and how they'll inform you about production downtime.

7. **Overall partner due diligence:** A data provider is just like any other business partner. Make sure you choose a company you trust and believe will provide ongoing value and good service.

8. **Global coverage:** Make sure the provider's global coverage matches your model's aspirations. For example, if you're looking to predict behavior of consumers around the globe, data providers focused only on the U.S. are not relevant.

Unfortunately, a lot of data companies out there are acquiring their data through non-compliant measures. Make sure the data provider is transparent about how it sources the data.

When it comes to open data, things are more straightforward. The main things to look at, besides the data itself, include:

1. **License:** Make sure you can use the data for whatever purpose you want to.

2. **Format:** In some cases, you'll need to spend a lot of time parsing the data (XML, nested JSON files, etc.) instead of just buying it from a commercial data provider. A good example of this is IRS data.

3. **Global coverage:** Make sure the provider's global coverage matches your model's aspirations. For example, if you're looking to predict behavior of consumers around the globe, data providers focused only on the U.S. are not relevant.

Step four:

Data provider tests

Now that you've narrowed down the list of potential data providers and done your due diligence, it's time to dig into the actual data. In order to do this, you should set up a test with each provider that will allow you to measure the data in an objective way.

Our goal here is to identify an uplift in the model's accuracy.

Data POCs are time-consuming so make sure you choose the providers you test carefully. Be sure to go through NDAs, commercial terms, and budget allocation that will enable you to enrich your data with the provider's data.

The next step is to test the data against the model you're building. **There would be no value in the data source if it does not improve your model.** The process here is quite technical and labor-intensive.

1. **Match your data with the provider's data:** This can take a while to do right, especially around names (e.g. company names, product names). This can be very frustrating because it means a lot of iterations on matching logics and digging into specific samples. Nonetheless, this step is crucial. Eventually, one of the ways you will measure the data provider is by the coverage metric — meaning the number of records that matched with the provider's data.

2. **Clean and extract meaningful features:** It’s time to engineer features that might be meaningful and relevant, which will then be fed into your model. Be creative, make sure you extract as many potentially relevant signals possible so you don’t miss any value you’d get from the data. Here are a couple of examples:

    a. **Fraud detection:** A lot of good features for fraud detection models are generated by analyzing interactions between a data provider’s data and the first-party data (e.g. if the address of a customer is the same address for both datasets).

    b. **Retail sales forecasting:** If you bought points of interest data for a retail sales forecasting model, you would probably want to calculate features like the number of malls around the store.

    c. **Reviews data:** If you bought reviews data, you can calculate the slope of reviews throughout time, the number of distinct reviewers, or the number of zero-star reviews.

3. **Measure model accuracy:** Time for the long-awaited answer. Feed the new features into your model and see if it improves. Obviously, not all models are the same, and some tweaking of the algorithm might be needed here to make sure the model adjusts to the new parameters. Additional features can make the model overfit the data and so you, therefore, need to make sure you adjust the model accordingly.

You might want to repeat these steps a couple of times to make sure you're not throwing away data that might be valuable.

Step five:

Calculate ROI

The good thing about this process is that once you have a quantified number for the model’s improvement, ROI can be calculated very easily. For example, if you're building a fraud model and found a data source that improved your model, you can calculate the extra number of fraud events you'll detect for every additional percentage of accuracy.

Some initiatives will end with negative ROI for all providers, but that’s the nature of this process — you never know beforehand what data will actually improve the model.

Once you calculate ROI for every provider on your list, it's relatively easy to choose a data provider. Some initiatives will end with negative ROI for all providers, but that's the nature of this process — you never know beforehand what data will actually improve the model.

Step six:

Integration and production

The last step in acquiring a new data source for your model is to actually integrate the data provider into your production pipeline. If the use case is a real-time use case you'll have to build the right infrastructure to consume data in real-time, extract features, and feed them into your model. If your use case is an offline use case, you'll need to build a batch job that will enrich and process a chunk of data every certain period.

When implementing the pipeline, make sure to handle exceptions and build fallbacks. A lot of things can happen in production (e.g. the data provider changed API schema, over-flooded with requests, authentication problems, etc.).

New data source: Acquired!

The process of searching and acquiring data for a predictive model can take a long time, a lot of resources, and end up providing no value to your model. It's a risk investing so much time and effort with no promise of results.

However, if you do find a valuable data source, it could mean a major breakthrough that provides more value and ROI than any other method you would use to improve a model. Just like the human mind, better information can help the model make better decisions and better predictions.

Don't have the time or resources for labor-intensive data acquisition? Explorium is here to change that.

About Explorium

Explorium is driving a new paradigm in the world of data science — one where companies can build models on the data they need, not the data they have. Our platform automatically extracts the most relevant features and integrates them to boost your AUC and power superior models.

Explorium is the leader and category builder of automated data discovery. We're turning the so-called "art" of data science models on its head by enabling any data scientist to automatically generate thousands of new features and immediately distill the top performers.

No need to waste time on searching for new data sets or trying to figure out how to test and integrate them. Explorium does it for you.

Ready to see how we do it?

Check out a demo

www.explorium.ai