Data Science Notes

Barclays Research's Data & Investment Sciences blog, where we dive into some of the frequently asked questions about data science and how it is reshaping investing.

BARCLAYS

26 June 2020

Alternative Data Can Break Traditional Ad-hoc Workflows

Data Science
Adam Kelleher

www.barclays.com

The world is becoming more complex, and investment professionals are increasingly expected to have to work with large data sets. Ad hoc workflows in Excel have their limits, either because of the size of the data or because of the complexity of models required to make sense of the data, or both. In this entry, we consider how best to optimise workflows, balancing the needs for larger datasets and more sophisticated models.

How to enable ad hoc analyses on large data

- Data science projects can be expensive, with costs for labour, development time (for infrastructure or analysis), input data, and hardware (purchased or rented), all of which add up quickly.
- The right tools, methods and infrastructure support workflows. Which are appropriate depends on the type of workflow: ad hoc (answering one-off questions) or production (developing repeated and automated processes).
- Our focus in this post is on managing ad hoc analysis with large or complex datasets.

Ad hoc versus production workflows

Ad hoc analysis in data science is more complex than in Excel

Using data to inform investment decisions can be straightforward, especially when you are doing ad hoc analysis of traditional datasets. You can often download an Excel sheet, do some aggregations to answer a question, and derive meaningful results. Moving the same ad hoc approach into a data science context can be complex and expensive, involving an open-ended commitment to data licenses, analyst and developer time, as well as hardware and software infrastructure.

As workflows and analyses grow more complex, they tend to become costlier. In general, there are two types of data science workflows:

- "Ad hoc" workflows are the most common, based on our experience. These include analysis to support one-off tasks such as investigating a trade idea, prototyping and optimisation of machine learning models, and exploratory analysis of new datasets for quality evaluations, feasibility studies, etc.
- "Production" workflows have tasks and processes that are repeated and (potentially) automated. Examples include software development of narrow data collection infrastructure, deploying machine learning models to run "live", automating analysis and reporting, and building dashboards. Broadly, these also include any other system one might build "in production" from an IT perspective.

Either of the above might have subtasks that could be outsourced to a data engineering team, especially where these require new infrastructure.

We focus in this post on ad hoc workflows

In this post, our focus is on managing complexity in ad hoc workflows. We draw our insights from our personal experience as well as the efforts of other technology-based companies. We recognize that every situation is unique. There is no one-size-fits-all solution.

A traditional analyst workflow might involve downloading data to Excel, doing some quality checks, and then building a model to answer a specific question. There might be tables and graphs, which may go into a report. The whole workflow can generally be contained to one analyst using one computer and a few externally supported programs (eg, a web browser, Excel, and a word processor). Those programs are developed and supported externally, so as long as things are working as expected, the entire workflow has no IT engineering requirements.

The traditional workflow of an analyst with Excel can break down due to size or complexity

There are a few ways this workflow can break down when analysts start working with larger, alternative datasets. These breakdowns can introduce new costs. Two of the more frequent reasons for failure are dataset size and model complexity.

Dataset size

The size of the dataset you want to work with gives us a progression of workflow breakdowns as datasets get larger.

Small Data – up to 500k rows

In our experience, datasets under 500k rows, or smaller than 15MB, can be used comfortably in spreadsheet tools like Excel. Many alternative data vendors provide aggregations that can be used in spreadsheets, or data exploration tools that can be used to filter their data down to Excel-friendly volumes.

Analysts can answer ad hoc questions the traditional way – by opening the data in a spreadsheet, and building some straightforward formulas. But as in the world of alternative data, many of the most interesting questions require working with larger sets than that.

Datasets may fit in memory, but be too large to fit in Excel – 500k - ~10mn rows

Excel's limits vary depending on your version, but once you are past them it has a tendency to become unstable. In this case, the most efficient solution is usually to transition to a scripting language such as Python or R (our Data and Investment Sciences teams are primarily Python users). In our experience, these work well with datasets up to about 10mn rows (the exact number depends on the encoding of the data and specs for given machines). There are a number of challenges here for investment analysts: the first is the learning curve of working with a coding-style interface, which can take analysts a long time, or intensive training, to become comfortable with. They are also limited in that some models which are well developed in excel (like the accounting-type "three statement" models) are much more cumbersome in a scripting language. On the other hand, python or R expand the range of possible analyses, so there are side benefits beyond just accommodating larger datasets. We have found that many open datasets fit in this range: examples include the US National Household Travel Survey or New York MTA Subway Ridership.

The ad hoc workflow for data in this scale is not too different from working with data in Excel: open a scripting interface (we prefer Jupyter notebooks), and ask questions of the data directly.

Switching to a scripting language can handle bigger files

Datasets may fit on a single computer, but be too large to fit in memory – ~10mn - ~1bn rows

But local machine memory is still a limit

At the smaller end of this range, you can still use a single computer, but now you have to work with it from disk. It can still be done with scripting languages plus some specialised skill, but in practice or if the dataset is too large to fit on a single hard drive (tens of GB on older computers to a few TB on newer computers), it needs to be stored externally. The typical solution is to store the data in a database and write queries (often in SQL) to perform basic operations and request subsets of the data. The cost for this solution includes the hardware for the database, the engineering time to set it up, and the extra analyst time to write the queries that perform the aggregations. Writing queries requires specialised skills, which takes either significant overhead to acquire (through training) for existing analysts or hiring costs to bring in analysts with the right skillset. You can outsource writing queries to IT or data science, but you often get better results when the person writing the query has the full context for what the query is for: there are many filtering steps and domain-specific choices of which engineers might not be aware. These are where the really fun alternative data really start; examples at this scale include some open datasets like commercial credit card transaction databases, customs data, and our ongoing collection of Twitter data with sentiment.

Here, ad hoc work requires some production like pre-work. Although databases can be set up on local machines, we usually start by having them set up properly by our IT engineering team. From there, ad hoc questions can be asked via structured query; it adds a layer of additional expertise in query writing, but once the database is set up, it's not so different.

Cloud or cluster solutions are complex and require specialist skills

Large or loosely structured datasets may also need a computing solution - >1bn rows or unstructured

Large and loosely structured data is where things get really different. At scales greater than about 1bn rows (details vary, of course), datasets enter the realm of cluster and cloud storage and computing.

At this scale, some real engineering choices have to be made before ad hoc questions can be asked. Approaches include using very large cloud instances, which act as one very large computer, or a collection of connected computers called a "cluster," which coordinates the work in a way that can be abstracted away from the analyst. These solutions are the most complex and require IT resources to set up and manage clusters and specialized skills (like working with PySpark, our usual choice) for doing analysis in that environment.

Even once those choices are made, asking ad hoc question tends to be more difficult in that scale – how queries are structured can affect runtime, or may cause the process to fail without delivering an answer, so more serious expertise is required for even simple tasks. But this is also where the complex and interesting investment questions can be structured. For us, these datasets have included ride-hailing and geolocation.

Model complexity

A second source of workflow breakdown is the need for more flexible and powerful models. You can do aggregations (eg, sums and averages) and regressions in Excel, but generally have to switch to scripting languages such as R or Python for machine learning models and working with novel data such as geolocation pings. It is not easy to attach a geolocation data point to a building in Excel.

Aside from the workflow breakdown of having to change to a scripting language, the models themselves can add complexity and cost to analysis workflows by, for example:

- **Parameter tuning.** Models might have many parameters, and their performance can be very sensitive to parameter choices. Manually adjusting the parameters to get good performance can take a very long time. Automatically adjusting parameters can be computationally very expensive. For example, if a model has six parameters that can each take ten values, you would have 1,000,000 choices to test.
- **Fragile pipelines.** The model prototyping workflow relies on repeatedly training a model and examining the results all the way through the analysis pipeline, where you are asking analysis questions. If you change your pre-processing, you have to re-train your model and re-run your analysis. This could involve finding the best parameters all over again. Changing anything changes everything.[1]
- **Specialized tools and skills.** While specialized computing tools introduced new costs due to training and hiring, the problem is amplified with advanced machine learning models. In our experience, it is very hard to hire machine learning specialists (but relatively easy to hire at the entry level).

Due to these costs, from our experience, it is better as a first pass to start with the simplest possible model and use that to develop your analysis workflow. From there, you can assess whether your model needs refinement by asking "what is the value of an improvement to this model?"

There are some notable exceptions.

- **Frameworks.** Modern machine learning frameworks such as tensorflow and pytorch have greatly reduced the development time for neural network models. Simple models such as deep feedforward networks are cheap to implement.

[1] D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, "Machine Learning: The High-Interest Credit Card of Technical Debt" (2014)

- **Pre-trained models**. You can often find open-source pre-trained models that can be useful for specific tasks. There are a few pre-trained image recognition networks in keras, for example, and there are many text embedding models such as GLOVE, FastText, and Word2Vec, which can be useful.
- **Transfer learning**. This can be a significant way to reduce training time, where you start with a pre-trained model, but then continue training with a new specific use case. You might, for example, build on a pre-trained image recognition model to identify a new object of interest; you might build on a pre-trained text embedding model to train a text-based regression or classification model.
- **Third-party APIs**. There are several "solved problems" in neural networks, where computers have reached human-level performance on specific tasks. These problems have been adapted to sit behind APIs and services that are provided by third parties. Google offers several such products for image recognition, video tagging, sentiment analysis, language translation, and others. Amazon offers several of these services as well, and there are very competitive smaller companies with similar product offerings.

Often, you can get more improvement by using better data instead of better models

With these solutions in mind, you might find a balance by building simple models yourself and licensing the use of more complex models as your use cases warrant. Often, you can get more improvement by using better data instead of better models. Better models can be justified when you are working with small datasets that require careful treatment of statistics. They can also be fun projects for your team to work on, so a nice change of pace to mix up the day-to-day work.

Finding a balance

Resist the urge to use the latest tools; instead, ask your staff where workflows are breaking down

Following agile principles, the optimal approach might be to start simple and shift workflows carefully and intentionally as needed to maintain a steady pace of development. Resist the urge to use the latest tools, but instead ask your staff where workflows are breaking down and what you can do to improve them. In addition to technological advancement, you will likely find other ways to boost productivity.

Consider whether you need large-scale datasets or whether you can get away with more aggregated (and, thus, smaller) datasets. You can avoid changing your workflows too much by looking at aggregated datasets as a stop-gap and can often find aggregations sold by third parties.

Also consider whether you need more advanced models or whether you should focus on more datasets. If you are refining a problem you have solved before, try upgrading your model sophistication a little and see if the performance improvement is comparable to the last time you added a new data source. That can help guide your decisions on where you should invest to improve more.

Disclaimer:

BARCLAYS

This communication has been prepared by Barclays.
"Barclays" means any entity within the Barclays Group of companies, where "Barclays Group" means Barclays Bank PLC, Barclays PLC and any of their subsidiaries, affiliates, ultimate holding company and any subsidiaries or affiliates of such holding company.

CONFLICTS OF INTEREST

BARCLAYS IS A FULL SERVICE INVESTMENT BANK. In the normal course of offering investment banking products and services to clients, Barclays may act in several capacities (including issuer, market maker and/or liquidity provider, underwriter, distributor, index sponsor, swap counterparty and calculation agent) simultaneously with respect to a product, giving rise to potential conflicts of interest which may impact the performance of a product.

NOT RESEARCH

The information provided does not constitute 'investment research' or a 'research report' and should not be relied on as such, although it may contain references to views or information published by the Barclays Research department. Investment decisions should not be based upon the information provided.

BARCLAYS POSITIONS

Barclays may at any time acquire, hold or dispose of long or short positions (including hedging and trading positions) and trade or otherwise effect transactions for their own account or the account of their customers in the products referred to herein which may impact the performance of a product.

FOR INFORMATION ONLY

This information has been prepared by the Research Department within the Investment Bank of Barclays. The information, analytic tools, and/or models referenced herein (and any reports or results derived from their use) are intended for informational purposes only. Barclays has no obligation to update this information and may cease provision of this information at any time and without notice.

NO OFFER

Barclays is not offering to sell or seeking offers to buy any product or enter into any transaction. Any offer or entry into any transaction requires Barclays' subsequent formal agreement which will be subject to internal approvals and execution of binding transaction documents. The products mentioned may not be eligible for sale in some states or countries, nor suitable for all types of investors.

NO LIABILITY

Neither Barclays nor any of its directors, officers, employees, representatives or agents, accepts any liability whatsoever for any direct, indirect or consequential losses (in contract, tort or otherwise) arising from the use of this communication or its contents or reliance on the information contained herein, except to the extent this would be prohibited by law or regulation.

NO ADVICE

Barclays is not acting as a fiduciary. Barclays does not provide, and has not provided, any investment advice or personal recommendation to you in relation to any transaction and/or any related securities described herein and is not responsible for providing or arranging for the provision of any general financial, strategic or specialist advice, including legal, regulatory, accounting, model auditing or taxation advice or services or any other services in relation to the transaction and/or any related securities described herein. Accordingly Barclays is under no obligation to, and shall not, determine the suitability for you of the transaction described herein. You must determine, on your own behalf or through independent professional advice, the merits, terms, conditions and risks of any transaction described herein.

NOT A BENCHMARK

The information provided does not constitute a financial benchmark and should not be used as a submission or contribution of input data for the purposes of determining a financial benchmark.

INFORMATION PROVIDED MAY NOT BE ACCURATE OR COMPLETE AND MAY BE SOURCED FROM THIRD PARTIES

All information is provided "as is" without warranty of any kind. Because of the possibility of human and mechanical errors as well as other factors, Barclays is not responsible for any errors or omissions in the information contained herein. Barclays is not responsible for information stated to be obtained or derived from third party sources or statistical services. Barclays makes no representation and disclaims all express, implied, and statutory warranties including warranties of accuracy, completeness, reliability, fitness for a particular purpose or merchantability of the information contained herein.

PAST & SIMULATED PAST PERFORMANCE

Any past or simulated past performance including back-testing, modelling or scenario analysis contained herein is no indication as to future performance. Past performance is not necessarily indicative of future results. No representation is made as to the accuracy of the assumptions made within, or completeness of, any modelling, scenario analysis or back-testing and no representation is made that any returns will be achieved.

OPINIONS SUBJECT TO CHANGE

All opinions and estimates are given as of the date hereof and are subject to change. The value of any investment may also fluctuate as a result of market changes. Barclays is not obliged to inform the recipients of this communication of any change to such opinions or estimates.

NOT FOR RETAIL

This document is being directed at persons who are professional investors and is not intended for retail customer use.
Not For Further Distribution or Distribution To Retail Investors.
For Discussion Purposes Only.

IMPORTANT DISCLOSURES

For important regional disclosures you must read, visit the link relevant to your region. Please contact your Barclays representative if you are unable to access.
EMEA
https://www.home.barclays/disclosures/important-emea-disclosures.html.
APAC
https://www.home.barclays/disclosures/important-apac-disclosures.html.
U.S.
https://www.home.barclays/disclosures/important-us-disclosures.html.

CONFIDENTIAL

This communication is confidential and is for the benefit and internal use of the recipient for the purpose of considering the securities/transaction described herein, and no part of it may be reproduced, distributed or transmitted without the prior written permission of Barclays.

ABOUT BARCLAYS

Barclays Bank PLC is authorised by the Prudential Regulation Authority and regulated by the Financial Conduct Authority and the Prudential Regulation Authority and is a member of the London Stock Exchange. Barclays Bank PLC is registered in England No. 1026167 with its registered office at 1 Churchill Place, London E14 5HP.

COPYRIGHT