Chapter 5. Predictive Analytics and Performance Optimization

In today's software-driven world, businesses generate vast amounts of data from their applications, users, and operations. This data holds valuable insights that can be used to make decisions, predict trends, and improve system performance. The ability to analyze and act on this data has become a critical skill for software engineers.

In this chapter, I tap into data analytics and business intelligence, and I'll test how state-of-the-art AI tools can help businesses understand their data and improve their results. Whether it's predicting future user behavior or optimizing resource usage, data analytics opens up new possibilities. Here are three key angles this chapter will cover:

Analyzing data

The first promise these tools make is that users can query large datasets by asking questions in natural language. This seems attractive, given how costly it is for companies to build visualization dashboards on top of their databases.

Distilling advanced insights

Companies often want to correlate data points and find patterns in order to understand user behavior or spot some malfunction in their processes, in ways that go well beyond simply querying the data to count and sum fields. Again, the costs of such projects prevent many companies from even stepping into this territory, so the promise of AI tools removing this barrier is a big one.

Predicting future behavior

The ultimate promise of data science and business intelligence is to pick up data about what happened in the past and use it to draw conclusions about what will happen in the near future. This forecasting can be a game-changer for businesses that do it well, and the

companies that develop and use it as a key part of their decision-making processes treat these projects and algorithms as sensitive intellectual property.

These angles show how software engineers and data analysts can turn raw data into actionable insights to help teams make smarter decisions. They also show how expensive and time-consuming these projects have been, historically. High costs and complexity have limited sophisticated projects in these fields to companies that have the funds and the technically capable teams to drive them. Most small to medium businesses, startups, and nontech businesses face high barriers to developing such tools, which hinders their ability to compete in this global market. The promise of AI tools in this sector is to democratize access to such tools among all market participants, regardless of their size, sector, or the technical abilities of their teams.

Before we look at the tools themselves, let's quickly review a few data analysis basics.

Data Collection and Sources

At the heart of data analytics is the data itself. In software engineering, this data comes from many sources, such as:

User activity

Information about how users interact with software, including page views, click paths, and session lengths

System logs

Detailed records of system and application performance, which help engineers monitor health and performance

Tracking tools

Automatically collected real-time data on application performance, such as response times and error rates

Customer feedback

Insights from user reviews, support tickets, and surveys, providing a qualitative perspective on user satisfaction

Market research

: Competitive analysis, news, market reports, and all relevant information that's published every day

There's a long tail of other possible sources of valuable data that businesses can use to shape their decisions about the future. These data sources form the foundation of data analysis. They also shape the quality of the data and determine how much data cleaning is required. For example, if a dataset contains lots of empty fields or inconsistent field types, it takes significant specialized work to clean the dataset for analysis, as well as advanced techniques like data normalization and clustering. I'll be exploring these techniques in the tools analyzed in this chapter.

Use Cases for Data Analytics

With valuable data in hand, data analysts can tackle a variety of challenges. We'll tap into some of these key use cases for data analysis and business intelligence in the tool evaluation section of this chapter.

Performance insights

: By analyzing system metrics and logs, engineers can identify performance bottlenecks and inefficiencies early on. This helps in optimizing resource usage, improving response times, and ensuring the application remains scalable as demand grows. For example, tracking CPU and memory usage over time can reveal patterns that signal when a system needs scaling or optimization.

User behavior prediction

: Data analytics can uncover patterns in user behavior, helping teams anticipate future user needs and preferences. By analyzing user activity data, such as click paths and session lengths, engineers can predict which features users will likely use more and tailor their product development efforts accordingly. This allows teams to focus on enhancements that will have the most impact on user satisfaction and engagement.

Capacity planning

Analyzing historical usage data can help teams predict future resource needs and scale infrastructure appropriately to meet demand. By understanding traffic patterns, engineers can forecast peak usage periods and prepare systems to handle higher loads without compromising performance.

Anomaly detection

Automated systems can analyze operational data to detect unusual patterns that could signal potential security breaches, system failures, or fraudulent activity. This proactive approach allows engineers to address issues before they escalate, minimizing downtime and protecting user data.

Business intelligence

Beyond performance and system optimization, data analytics can offer broader insights into business performance. This includes tracking product adoption, analyzing market trends, and evaluating key business metrics. These insights help guide strategic decisions, such as which new features to prioritize or how to allocate resources more effectively.

Each of these use cases highlights how data analytics allows teams to make informed decisions, optimize processes, and improve both software performance and the overall business strategy. The sections that follow look at how to approach these use cases with the right tools, models, and techniques. We'll also explore how AI and machine learning can further enhance data analytics, helping engineers automate processes and uncover insights faster.

Types of AI Tools for Data Analysis

AI tools have been emerging in data analysis, as they have in many other industries and verticals. Just using many enterprise tools requires complex sales and onboarding processes; I've left those out of the scope of this book, with the goal of steering you toward the most accessible options.

I've also found some tools that offer infrastructure-level support for data analysis. While many of them are valuable, this chapter's use case is about a

business owner who wants to extract business-worthy insights from a dataset, and such tools are overkill for such cases.

I ended up with tools that offer self-service onboarding and that have a free tier that allows readers to test the software. Almost all of these tools contain a chatbot UX that lets you upload a data file and ask analytical questions about the data. This seems to be the winning UX pattern for data analysis use cases.

Evaluation Process

I evaluated more than 20 AI tools in the data analysis and business intelligence space in order to shortlist the ones I highlight in this chapter. Every tool covered here meets the following criteria:

- It is a professional project with a competent team behind it.
- It generates high-quality results.
- It offers some level of functionality for free or on a trial basis.
- It has a high level of adoption at the time of writing (mid-2025).

For this test I'm using a public dataset of one year of online retail transactions from the Machine Learning Repository at the University of California, Irvine. It contains over 500,000 transactions, with eight data columns for each transaction:

- InvoiceNo
- StockCode
- Description
- Quantity
- InvoiceDate
- UnitPrice
- CustomerID
- Country

You can see a sample in Figure 5-1.

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/10 8:26	2.55	17850	United Kingdom
536365	71053	WHITE METAL LANTERN	6	12/1/10 8:26	3.39	17850	United Kingdom
536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/10 8:26	2.75	17850	United Kingdom
536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/10 8:26	3.39	17850	United Kingdom
536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/10 8:26	3.39	17850	United Kingdom
536365	22752	SET 7 BABUSHKA NESTING BOXES	2	12/1/10 8:26	7.65	17850	United Kingdom
536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	12/1/10 8:26	4.25	17850	United Kingdom
536366	22633	HAND WARMER UNION JACK	6	12/1/10 8:28	1.85	17850	United Kingdom
536366	22632	HAND WARMER RED POLKA DOT	6	12/1/10 8:28	1.85	17850	United Kingdom

Figure 5-1. Sample of the online retail transactions dataset

For this comparison test, I'll act as if I'm the owner of an online retailer and attempt to use the AI tools to draw business-worthy insights from this data. I'll ask questions like:

- What are my top-selling products?
- Which of my customer segments have the highest lifetime value?
- What is my sales forecast for the following year?

In this test, I'm importing the raw dataset into each tool, working through the UX to get insights, and taking note of the results I get, how easy it is to get them, and how the tools compare to each other.

Julius

Julius AI automates complex data analysis processes and provides interpretations, visualizations, and predictive analytics. It uses a combination of OpenAI's GPT-4 and Anthropic's Claude as its underlying data-processing models.

Julius's instant-messaging UI (Figure 5-2) resembles those of ChatGPT and other popular AI tools. I used it to upload my dataset, then asked my first question, in natural language:

```
What are the top selling products, both by number of ur
revenue generated?
```

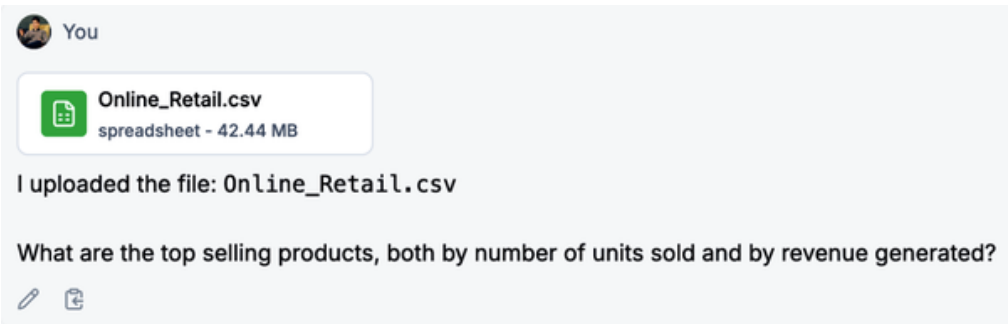

Figure 5-2. My file upload and first prompt to Julius

Julius took just over a minute to run several Python scripts, then returned the tables shown in Figures 5-3 and 5-4, which list the 10 top-selling products by number of units sold and by revenue generated.

Figure 5-3. Part of Julius's reply to the first prompt, showing the top-selling products by quantity

Top 10 Products by Revenue:

Optional: Click on columns or rows to focus on specific data — Export

	TotalRevenue
DOTCOM POSTAGE	206245.48
REGENCY CAKESTAND 3 TIER	164762.19
WHITE HANGING HEART T-LIGHT HOLDER	99668.47
PARTY BUNTING	98302.98
JUMBO BAG RED RETROSPOT	92356.03
RABBIT NIGHT LIGHT	66756.59
POSTAGE	66230.64
PAPER CHAIN KIT 50'S CHRISTMAS	63791.94
ASSORTED COLOUR BIRD ORNAMENT	58959.73
CHILLI LIGHTS	53768.06

Figure 5-4. Part of Julius's reply to the first prompt, showing the top-selling products by revenue

Along with the hard data, Julius offered some contextual explanation about high-value products (Figure 5-5), highlighting some interesting aspects of the data.

Figure 5-5. Some of Julius's observations about the data in response to the first prompt

Next, I asked Julius to identify which customer segments were generating the highest value for this online retail store:

```
From the same dataset, do customer segmentation and cal
value of each client segment. I want to identify the hi
customer segments.
```

Julius decided to follow the Recency, Frequency, Monetary (RFM) method and ran some Python scripts again for almost two minutes, then presented the customer clusters and their lifetime value (Figure 5-6).

Figure 5-6. Julius's reply to the second prompt, with its analysis of high-value customers

Julius figured that clients that make an average 5,914 purchases and generate £64k per year are likely to be resellers, and gave me pointers on how to use

that information for my marketing efforts. This is a unique and very good insight; however, key information is still missing, such as how large this cluster is and how representative these figures are. Is it an outlier case of an extremely high-value client, or does it include hundreds or even thousands of clients that I can market toward to grow the business?

Also, Julius clearly hallucinated on its lifetime value (LTV) calculation: by no means would any client be worth £13 billion. It's not clear what went wrong, but my hunch is that Julius used the wrong field in this calculation.

My third question about this dataset asked Julius to do some forecasting:

```
Assume I do a marketing investment of £500k to grow thi
segment that you've identified as high-LTV. Forecast my
volume for the next 12 months, both in units sold and i
want to provision inventory based on your forecast. Ger
for the 20 top selling products in the forecast, with p
and quantity to provision for next year.
```

In this forecast (Figure 5-7), Julius first made the fair assumption that these marketing efforts would generate a 20% sales increase. However, note that its total revenue forecast is off: the current year's revenue is around £9.7 million, so this £565k forecast would be a massive decrease.

Figure 5-7. Part of Julius's reply to the third prompt, with sales forecast

Second, the table it created with the number of items to provision for next year, shown in Figure 5-8, is also off (probably a cascade from the wrong revenue estimation). The quantities listed are way below the current year's sales volume for those items. So, while Julius's forecast indicates a 20% sales increase, its stock-provisioning figures suggest a *decrease* of 60% or more.

Figure 5-8. Part of Julius's reply to the third prompt, with its stock-provisioning forecast

Julius did well in the objective data analysis in the first prompt and showed promise in customer segmentation, where it offered an insightful analysis despite the error in its calculation. However, its forecasting was far off the mark. While this could perhaps be mitigated with some prompt engineering, I believe my question was specific enough that I could reasonably expect a better answer.

As such, I'm rating Julius a 7/10 in this test.

Akkio

Akkio provides AI-driven data analysis and predictive modeling aimed at digital-marketing and ad-targeting clients. Akkio uses its own model, called AD LLM, which it claims to have trained on advertising-specific data to understand data structures, business requirements, and other context specific to ad targeting.

Akkio's polished UI starts with a file upload, prompting users to select from a number of file formats. Once I uploaded my file, Akkio took some two minutes to fully ingest it and make the product functionality available to me. While the file renders as a spreadsheet table in the UI's Prepare tab, the product offers several different features (Figure 5-9): Prepare, Explore, Predict, Deployments, and Reports.

Figure 5-9. Akkio navigation UI

The Explore tab displays an instant-messaging UI similar to the one in Julius. As such, I began with the same prompt, asking about the top-selling products:

```
What are the top selling products, both by number of ur
generated?
```

Akkio took just a few seconds to reply to my prompts. However, the output was quite raw and lacked context, often consisting of just a data table, with no accompanying text to provide context. For instance, in response to my first question, it simply returned charts and tables (Figures 5-10 and 5-11) with the top-selling products by quantity and by revenue generated, respectively.

Figure 5-10. Part of Akkio's reply to my first prompt, with top-selling products by quantity

I used the AI Interpretation widget above each chart, but it didn't help much, since the explanation was very technical. It simply described in natural language the technicalities of the query performed against the data; there was no functional context about the analysis being made or what we are seeing in the rendered visualization.

Figure 5-11. Part of Akkio's reply to my first prompt, with top-selling products by revenue

Akkio's response to the first question was correct, so I proceeded to the second question:

```
From the same dataset, do customer segmentation and cal
value of each client segment. I want to identify the hi
customer segments.
```

Akkio's reply is good (see Figure 5-12) since it tells me that the highest-value customer segment is composed of 2,539 clients, who generated an average of about £3,000 each. However, that cluster is too large. I'd expect such a cluster to be perhaps 10% to 20% of the total number of customers, but this one encompasses more than *half* of the 4,372 unique customers in the whole dataset.

Figure 5-12. Part of Akkio's reply to my second prompt, with customer segmentation and value calculations

Working with such a broad cluster of customers produces generic recommendations that aren't very actionable, as seen in Figure 5-13. These recommendations would apply to any business; none of them are specific to high-value customers in a way that I can use for marketing purposes.

Figure 5-13. Part of Akkio's reply to my second prompt, with an explanation about the clients with the highest lifetime value

I proceeded to the third question:

```
Assume I do a marketing investment of £500k to grow thi
you've identified as high-LTV. Forecast my total sales
months, both in units sold and in revenue. I want to pr
```

```
on your forecast. Generate a table for the 20 top selli
forecast, with product name and quantity to provision f
```

Akkio's reply is so devoid of context that it's hard to understand the rationale behind the numbers. The revenue figure of £9.8 million (Figure 5-14) is a slight increase over the current year's £9.7 million in revenue, which seems too low given the marketing investment I mentioned in my prompt—but, again, no context is provided for that forecast.

Figure 5-14. Part of Akkio's reply to my third prompt, with sales forecast

Also, there seems to be some hallucination in the stock-provisioning forecast (Figure 5-15). These numbers are way below the current year's sales for those products, by an order of magnitude. For example, Akkio forecasts 1,918 sales for the first item on the list, even though the current year's sales for that item total 53,847. This, too, suggests some confusion in the calculations, but without any visibility into the scripts or functional explanations of the context, it's hard to understand the process that led to those predictions.

Figure 5-15. Part of Akkio's reply to my third prompt, with product provisioning forecast

Here, too, using the AI Interpretation tab (Figure 5-16) doesn't help much. It provides a technical description of the query used to feed the chart, but it doesn't explain the broader thought process that led to that query.

Figure 5-16. AI interpretation of the results

Akkio did well on the objective data analysis with the first prompt, and it did some decent customer segmentation work with the second prompt. However, its forecasting was off. I think the most underwhelming aspect of using Akkio is how little context it provides for its replies. It comes across as a tool best used for simple dataset queries and charts that don't require much analysis or contextualization.

As such, I'm rating Akkio a 5/10.

ChatGPT

Since the top products that emerged in this category offer a chatbot-type experience, I decided to add ChatGPT for comparison, since it is most people's go-to AI tool. I'll quickly review its replies for each of the same three questions.

As seen in Figure 5-17, in reply to my first question, ChatGPT correctly counted the product sales and summed the revenue.

Figure 5-17. ChatGPT's reply to my first prompt

The second question is more nuanced. ChatGPT found a customer segment of 498 generating an LTV of £403,000. However, that value is higher than the revenue generated by *any* client in the current year, which seems to be a

mistake. ChatGPT could be assuming that clients will be purchasing for many years into the future, which would inflate the LTV calculations.

When I asked ChatGPT for more details about that highest LTV customer segment, its reply (Figure 5-18) was underwhelming and generic enough to apply to any business, rather than being specific to this dataset and the underlying business.

Figure 5-18. Part of ChatGPT's reply to my second prompt, with its explanation for the customer segmentation

ChatGPT's reply to my third question, asking it to forecast revenue and stock provisioning based on a significant marketing investment, started off quite well. ChatGPT estimated that the marketing investment would produce a sales increase of 20% to 40% in this customer segment.

However, its product stock-provisioning forecasts are off (Figure 5-19): again, the sales numbers it predicts are lower than the current year's. To me, this suggests that ChatGPT calculated the number of items to be sold *only* to this high-value customer segment and "forgot" to include the items sold to all other clients.

Figure 5-19. Part of ChatGPT's reply to my third prompt, with product stock-provisioning forecast

ChatGPT did well in the objective data analysis with the first prompt, and it offered a good amount of context and reasoning in response to the other questions, despite some obvious issues with the calculations. I believe some of these issues might be mitigated with prompt engineering. It's also worth noting that ChatGPT isn't really a native data analysis tool: unlike the other tools analyzed here, it renders clunky tables and has no ability to render charts.

As such, I'm rating ChatGPT a 6/10.

Tool Comparison

My first challenge with this comparison is that all three of the tools I analyzed were subject to a black-box effect. I input a large volume of data, and within seconds these tools output good-looking tables, charts, and write-ups with conclusions and insights that appear to make sense. It would be easy to assume that the information provided by these tools is correct, given their impressive speed and output.

However, I double-checked the results by running a script on my local machine against the reference dataset (Figure 5-20; this script is available in the book's GitHub repository).

Comparing the tools' results against my local tests, I first observed that all tools *missed* the product with most units sold ("Small Popcorn Holder"). I dug

a bit into this quirk, but I couldn’t figure out why. I can speculate that, since this item has a very low unit price, perhaps a rounding-to-zero type of error could have caused it to be missed.

Besides that, all tools performed quite similarly, both in terms of the value they provided and their pitfalls. From a UX perspective, Akkio stands out from the other tools. It sets a higher expectation by offering what seems like a very robust process with multiple steps and tools. However, it ends up standing out negatively, because the level of contextualization it provides for each interaction is way below what the other tools offer.

Julius and ChatGPT are simpler chatbot experiences that take longer to reply, but offer insights into what’s happening and how my data is being processed. Both of those tools include text in their replies alongside the tables and charts, to provide context and reasoning for their operations and to show users how to read the data and interpret the results.

Figure 5-20. Console log of my local tests to double-check the tools' calculations and reference figures for items sold and revenue generated

If I were to choose one of these tools, I'd select Julius. While its UX is very similar to ChatGPT's, and even the underlying model is in part the same (GPT-4, as I write this in mid-2024), its data analysis capabilities, such as rendering charts in the chat conversation, are not available in ChatGPT.

I rated all three tools between 5 and 7 (Table 5-1), given these shortcomings. I expect these tools to evolve a lot in the coming years, but in my opinion, they are not yet reliable enough that you can simply give them a large volume of data, ask questions, and trust the results. If you use them, I recommend running scripts locally to double-check the numbers. (It's OK if your scripts are generated by AI tools, since you can review and modify the code and have full visibility and control over the data analysis, as you saw in Chapters 2 and 3.)

Table 5-1. AI data analysis tools overview

Tool	UX	Test performance
Julius	Chatbot	7/10
Akkio	Chatbot	5/10
ChatGPT	Chatbot	6/10

Conclusion

After more than 15 years working with software development and data science teams, I can confidently say that AI tools have the potential to become game-changers in how we handle data analysis and business intelligence. Their ability to clean and analyze massive datasets in seconds, rather than days, will transform what's possible for businesses of all sizes.

Furthermore, from my experience working with a wide range of business stakeholders, from early-stage startup founders to business teams at Fortune 500 companies, I can easily imagine these AI tools empowering nontechnical stakeholders to extract insights from their data. The effects of that empowerment could be immense. In some cases, it might mean skipping costly data engineering projects; in other cases, it just makes those projects faster and less expensive.

With that, here's my word of caution: the tools are not there yet. While the results can be very impressive on the surface, they come with significant flaws, calculation errors, and generic explanations. A distracted user might be easily fooled by the instant reward of good-looking charts and insights, but overlooking such shortcomings can result in serious negative consequences. Business stakeholders could make decisions that reduce the value of their

business; data analysts who delegate their work to these tools might end up performing poorly in their jobs.

These tools are already powerful and useful. But they have limitations, and the "black box" effect can make it very hard to identify those limitations. Always be specific in your prompts, and always double-check the results by doing manual analysis or running local scripts. I always tell my teams to treat AI-generated insights like advice from a colleague: while it's valuable input, always validate it and do your own critical thinking before making any big decisions.