WHITEPAPER

A Buyer’s Guide to Enterprise Data Science Platforms

RAPIDMINER

An ALTAIR Company

Today's enterprises are drowning in data. It's being collected faster, in greater quantities, and in more formats than ever before. In theory, more information should mean better decision-making. The reality isn't that simple. The speed and volume of data collection make it impossible for humans to process and make decisions based on it, and the variety of formats makes traditional analytics techniques less useful by the day.

Because you're reading this, we're guessing you already recognize that data science offers a way to combat this information overload. By automating the analysis of large volumes of data—and subsequent generation of insights—the right models can help you better understand your customers, protect against risk, and actually use the data you're gathering to remove the uncertainty from process decisions.

This guide isn't just designed to give you a checklist of features to look for in a data science platform. It takes a hard look at the challenges that prevent so many enterprises from delivering on the promise of data science, and shows how the right strategy paired with the right technology can help you overcome them.

Table of Contents

The "honeymoon" for enterprise data science is over. With almost 90% of models never making it into production, and even less having their desired impact, organizations are second-guessing the promise of data science—and are unsure how to deliver on it.

The challenge is that the underlying causes for failed data science projects usually don't have a quick fix, and reflect larger issues that span people, process, and technology (more on this later).

As a software provider that's been in this space for over a decade, we'd love to tell you that buying a no-code platform and training your teams on it is the solution to your problems. In reality, success in enterprise data science requires a more comprehensive approach. You'll need to establish a strategy that helps to break down organizational silos, prioritize high-value use cases, embed models where they can provide their most valuable predictions, and ensure long-term value creation.

This guide is broken down into three key sections that will help you understand:

1. **Why** you should still look to data science as a way to create long-term value for your organization
2. **How** an upskilling-focused strategy paired with the right data science platform can lead you to faster, more sustainable wins
3. **What** to look for in a platform to support your data science projects

Let's get started.

The Impact of Successful Data Science Programs

If you're involved with your company's data strategy, you're likely already familiar with the concept of digital transformation. In case you're not, the premise is straightforward: digital transformation is the process of embedding technology throughout your business in a way that helps you change how you operate and create more value for your customers. Many companies have already found success with digital transformation initiatives like modernizing applications and moving to cloud-first architectures, making previously hidden organizational data accessible.

That's where data science comes into play. Data science is a multi-disciplinary approach that allows computers to learn from data that represents real-world events rather than explicit programming. By helping companies use their existing data to make predictions about what's likely to happen in the future, data science can help companies change how they design processes, assess risk, and understand customer behavior. The benefits also aren't limited to any one industry—everyone from manufacturers to online retailers can use data science to make smarter decisions.

Here's the problem. Because of poor stakeholder management, shortcomings in approach, insufficient technology, or some combination of the three, most data science projects fail. For the companies who take them on, that failure represents wasted investments, misallocated resources, and mismanaged expectations.

But it doesn't have to be that way. The honeymoon for data science may be over, but that doesn't mean your efforts have been a waste—the evidence actually suggests that the sooner you align your people, processes and technology to facilitate data science projects, the greater your ROI will be. A recent Forrester Consulting study found that early adopters of data science who've successfully brought models into production are seeing the kind of results that most companies are hoping to see.

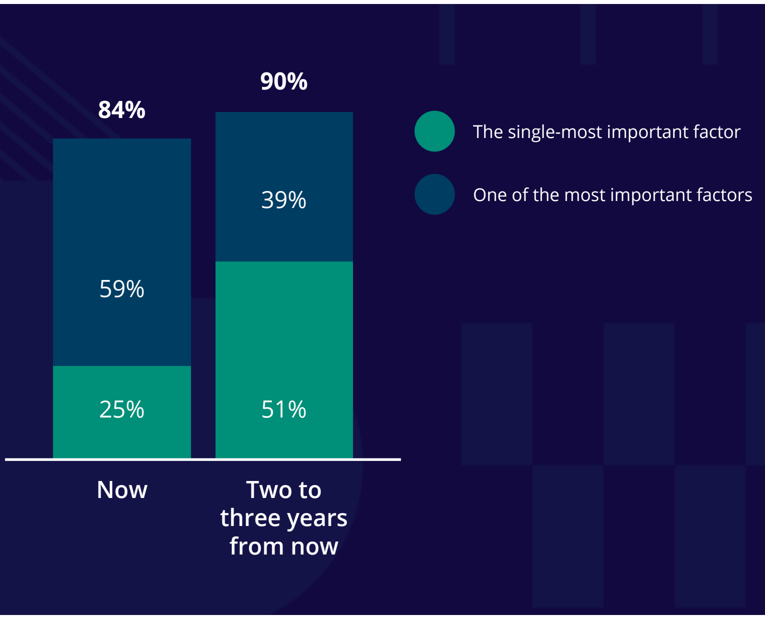
The data

While we'd encourage you to check out the full report, there are few key data points that illustrate how early adopters of data science are successfully transforming their processes and organizations.

The Importance of Data Science for Competitiveness

25% of early adopters say that data science is the most important factor for their competitiveness today.

51% expect it to be the most important factor in 2-3 years.

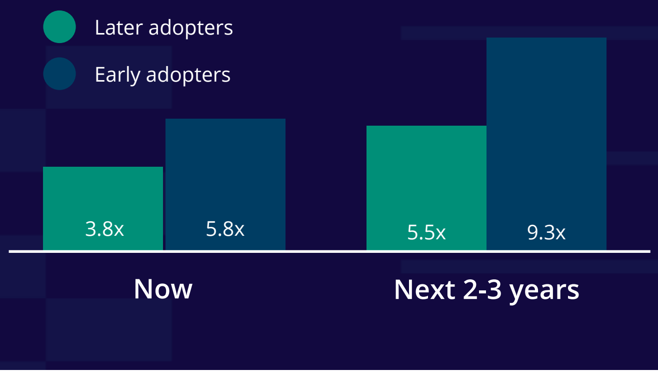


The Importance of Use Case Planning

Early adopters have more robust plans for how to use data science, including new customer acquisition, improvement of existing products or services, and the innovation of new products or services.

The ROI of Data Science

Early adopters are currently seeing a 5.8X return on their data science investments, compared to 3.8X for less mature organizations. In 2-3 years, the expected ROI for early adopters is 9.3X, compared to 5.5X for later adopters.



The difference in realized and expected ROI doesn't just show us that data science projects can deliver strong results in an enterprise setting, but also that success requires iteration. Companies who got started early have been able to bring more models to production, which in turn has allowed them to work on embedding data science into more areas of their business and see a higher return on investment.

While most organizations can agree on the potential of data science, every company has different people, expertise, and data they need to leverage to find success. But too often in enterprise data science, companies try one of two "one-size-fits-all" approaches: outsourcing projects to consultants, or hiring a team of data scientists.

Common Approaches to Data Science, Explained

There are 3 common ways that organizations approach data science: outsourcing, hiring, and upskilling. In this chapter, we'll provide an overview of these strategies and provide some pros and cons for each.

Relying primarily on outsourcing

Hiring consultants is a common approach, especially for companies who don't have in-house data science resources. On the surface, this makes sense—consultants exist to help fill their clients' capability gaps.

If your organization doesn't have widespread plans for data science usage as part of digital transformation initiatives, outsourcing projects on an as-needed basis may well make sense. However, if your goal is to improve data literacy throughout the enterprise and address a range of use cases across different areas of the business, relying exclusively on consultants raises a few common challenges. We'll explore those in more detail within this section.

Limited long-term value

The first issue that comes up when companies rely on outside consultants to handle their data science projects is that model maintenance isn't a priority.

For a data science solution to produce the greatest ROI, it needs to be tuned and adjusted over time. Even the strongest models become less accurate and make fundamentally flawed predictions if they're left alone for too long.

Again, if you've identified one or two high-value use cases that can quickly generate your desired return, this isn't much of an issue. But if your goal is to maximize long-term value, you should know that your models will have more staying power when your internal teams understand how to manage them. If they don't, your models won't stay in production for as long as they could, or worse—you risk making decisions based on inaccurate or otherwise problematic predictions.

Lack of true transformation

To reap the full value of data science, you need to improve the overall data literacy and analytics "know-how" within your company. Having a data-driven culture will yield greater long-term results than relying solely on data experts to handle your analytics work—whether they're outsourced or in-house.

As we've mentioned, outsourcing your projects to consultants can lead to wins for specific problems, but it does nothing to build data literacy within your company, which is a problem for any organization that doesn't want to stop at a single use-case. Outsourcing is expensive to begin with, and the idea of bringing in consultants for every data science project you're taking on will eventually be cost-prohibitive given that most organizations have hundreds of use cases to choose from.

> **"Often saddled with legacy data environments, business processes, skill sets, and traditional cultures that can be reluctant to change, mainstream companies appear to be confronting greater challenges as demands increase, data volumes grow, and companies seek to mature their data capabilities."**
>
> **- RANDY BEAN, HARVARD BUSINESS REVIEW**

Hiring & isolating an internal team

If you're just starting to think about how to create impactful data science projects at your company, hiring a team of data scientists may seem like the most obvious option. Just as you'd hire salespeople to hit a revenue goal or engineers to build products, you can start bringing in data scientists to create and deploy impactful models...right?

Well, not exactly. While having coding data scientists on your analytics team is largely a positive, far too many companies make the mistake of relying on hiring as their only strategy. Becoming a data-driven company requires wholesale change, and simply hiring and isolating a data science team from the rest of your business isn't just ineffective¬¬—it prevents analytics from becoming a core competency throughout your organization.

With that in mind, let's look at a few specific reasons why the "hire and isolate" approach doesn't typically work.

Data scientists are difficult to hire

The gap between the demand for data scientists and the available talent pool is getting wider—so much so that it's now a common challenge for companies to hire for the role, much less retain employees over the long-term. In fact, a Burtch Works study from 2019 showed that the average tenure for data scientists is just 2.6 years, with 17.6% of respondents having changed jobs the year before.

All the signs tell us that it's only going to get more difficult to hire data scientists at all, let alone bring in the ones who have an aptitude for your unique business. As more companies look to transform their business through data science, the hiring shortage will only become more pronounced.

Data scientists need business context

In many cases, data scientists are hired to work in industries that they're new to. This means that bringing them into projects isn't as simple as giving them access to relevant data and letting them take the wheel—in order for them to be successful, they need to develop a thorough understanding of what that data actually means. If you're trying to optimize the yield from a manufacturing process and plan to primarily rely on a data scientist who has no idea what condenser temperature is, you likely won't get very far.

While a data scientist could theoretically spend time learning the ins and outs your unique business, this type of approach extends time-to-value and is very difficult to scale. On the other hand, finding ways to more closely involve business stakeholders from the beginning of your data science projects ensures that they have all the context they need.

This brings us to the next common challenge created by a "hire and isolate" approach.

Business stakeholders need to clearly understand results

For a domain expert, more traditional analyst, or executive to make decisions based on a model's predictions, they need to understand the work that went into creating it and easily be able to interpret those predictions. This means being able to understand every step of a given data pipeline and go beyond model accuracy to determine a decision's financial impact.

When models are custom-coded using Python or R, this becomes impossible at worst and overly time-consuming at best. It's difficult for data scientists to deliver results in an intuitive way without business stakeholder involvement, not to mention unreasonable for them to manually document every step of their pipelines from data access to testing and validation.

Again, we're not advocating against hiring data scientists—they're in-demand for a reason, and their skills can help you create truly differentiated solutions. But if hiring is your only strategy, it'll be difficult for your data scientists to truly understand business problems and get on the same page with non-coding decision makers.

Here, "domain experts" refers to anyone responsible for a functional area of the business. This could be a plant manager, marketer, customer support leader, etc.

As we mentioned earlier, you'll need to take a more comprehensive approach to find success with data science. That involves maximizing the value of your existing people, expertise, and data.

Benefits of Bringing Data Science Closer to Your Business

As we've previously alluded to, the most proven way to generate value from your data science initiatives is to closely tie them to key business goals and processes—as well as the people who are responsible for them.

On a "people" level, bringing data science closer to the business means two things: making it accessible for employees without coding skills, and upskilling your teams so they have a shared understanding of the both business goal and data science concepts that can help to achieve it.

The right technology can be tremendously helpful here. It's unlikely that someone who's spent their career on a production floor or marketing team will learn how to create models using Python, so investing in a code-optional data science platform can remove what's arguably the largest obstacle that's preventing them from getting involved in projects.

Upskilling is the practice of teaching your employees new skills—both to advance their careers and help them create more value for your organization. In this context, it means helping non-data scientists build skills around data preparation, modeling, interpretation of results, and general data science best practices.

Effective upskilling is a prerequisite for successful data science. Don't take our word for it—a McKinsey study found that 70% of companies who've invested in upskilling report business results that exceed those investments.

Enabling non-coders to contribute to projects while learning data science concepts is a great way to bridge the data skills gap and create value for your company right away. In this section, we'll talk through the benefits of bringing data science closer to your business by making it accessible across the enterprise and prioritizing an upskilling-first approach.

Shared project ownership

Sharing of knowledge and expertise

As we mentioned earlier, the problems that companies are trying to solve with data science are complex, and it can take a while for data scientists to develop a full understanding of the data they're working with. By involving business experts in the process, you can accelerate that learning and provide crucial context for data scientists.

And, while business experts have a strong understanding of the challenges they face and what potential solutions look like, they don't have the data science knowledge required to build those solutions. By working with data scientists, they can learn concepts that are relevant to the models they're building, allowing them to create additional value on new projects down the road. We should mention that while many DSML vendors pitch collaboration as a benefit, the only way to deliver on that promise is to enable non-technical stakeholders to do things that normally rely on code.

Joint understanding of what "success" looks like

To effectively create and operationalize a model, you need to go into the process with a shared definition of success among all stakeholders. Having joint buy-in on what accuracy threshold a model should meet and well-defined goals for its financial impact will ensure that everyone is aligned when it comes time to put a project into production.

Sharing of data assets in a governed environment

Many data science projects fail to gain traction due to a lack of relevant data assets. The average enterprise uses hundreds of applications, which means that critical process data is stored across disparate systems. By centralizing relevant data and project work, you not only ensure that it can be used to build a solution a given problem, but also maximize its value in the long-term since it can be used to solve multiple challenges.

Simplifying data access doesn't have to come at the cost of security. The right data science platform will also allow admins (e.g. IT & Security Professionals) to control that access down to the employee level while meeting ever-increasing standards for enterprise data security.

Trust in results

Easier to get buy-in for deployment

When working on complex data science problems, it's natural that someone may not think to document every step they're taking and all the ways in which data has been transformed. However, when it comes time for deployment, it's also understandable that executives wouldn't feel comfortable making key business decisions based on models they don't fully understand.

By adopting the right data science platform, you can both ensure that every project step will be automatically documented for full transparency and that results will be displayed in an intuitive way.

Bias detection & mitigation

As enterprise data science adoption increases, the emphasis on creating models that make ethical and sound decisions does too. While models won't make prejudiced decisions or misinterpret outcomes due to personal viewpoints, they will make decisions based on the data they're being fed.

By leveraging a data science platform, your team can gain a full understanding of the data that was used to create & train a model, see how its been transformed over time, and avoid various forms of bias as a result.

Best practice sharing & reuse

Often, organizations with multiple data science use-cases find that there's an overlap between projects—things like the data that's being used, steps to prepare it for machine learning, and how to operationalize tested models.

It's not just important to be able to reference documented workflows if you decide to take on a similar projects—being able to bring new users up to speed quickly means that your models will be more resilient when you're faced with employee turnover, which is a growing challenge for most of today's enterprises.

Future-proof value

Built-in fit with existing systems

For a platform to help deliver on the promise of data science, it needs to work with the existing systems and architecture that your organization has in place. As mentioned in previous sections, projects are often held up due to logistical issues like a lack of relevant data access. In some cases, this can significantly delay deployments and eat into a project's ROI—given that promising projects often have $100K+ of business impact, it's important to make sure that data accessibility doesn't get in their way.

The right data science platforms can help address those challenges right away, while providing enough flexibility to adapt to infrastructure changes in the future. Enterprises should seek out platforms that can access relevant business data from where it's stored and also integrate with other tools within their analytics technology stack (e.g.. Business intelligence tools for easy visualization of results).

Smoother project handoff

The data science to ops handoff is a stage where many projects stall, especially when those responsible for implementing projects don't understand how models work and what the desired business outcomes are. To put a model into production, your team needs to understand all of its' dependencies.

By leveraging a data science platform, you provide those responsible for operationalizing models with full visibility into the end-to-end data pipeline, as well as project notes & logs containing all the information they need.

More long-term value

Effective long-term monitoring and iteration help to maximize the value of an analytics model. The problem is, maintenance work isn't particularly interesting to data scientists, who are more concerned with applying cutting-edge data science techniques to new problems. For organizations without a plan, this means that models will ultimately lose their value and fail to provide a sustainable competitive advantage.

By making data science accessible to anyone through a platform, companies can enable non-coding business experts to intuitively and effectively monitor models and guard against threats like deterioration and concept drift.

Key Capabilities to Look For in a Data Science Platform

When handled properly, an investment in a data science platform can propel your digital transformation initiatives and have a positive impact on everything from traditional BI and reporting efforts to the identification of new markets and revenue opportunities. Needless to say, that investment should be handled with care.

Now that you've seen how the right data science platform can help you create and operationalize models more effectively, let's take a more detailed look the type of functionality that'll help you achieve the benefits described above.

Complete experiences for every type of user

Regardless of their initial skill level, every user within your organization should be able to work productively within your data science platform.

The market is filled with tools that empower specific user types—usually to the detriment of others. For example, platforms that cater to coding data scientists fail to bring non-coding domain experts along, even though they're the ones with the most knowledge of business processes and the data collected from them.

In some cases, data science vendors will tack on capabilities to make it seem like they cater to users with diverse backgrounds, but offer no ability for those users to work interchangeably.

The right platforms offer in-depth experiences for everyone while giving different user types the ability to work together. In this section of the guide, we'll go through the different authoring types you'll need to cater to employees with varying skill levels and ensure that they can positively contribute to data science projects.

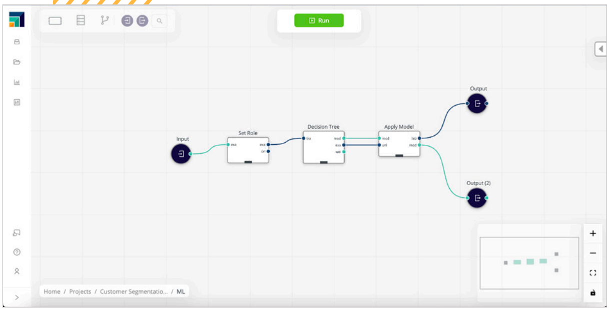
In-platform coding notebooks

It's no secret that data scientists typically prefer to use code, especially when taking on complex projects. Python and R both have relatively simple syntaxes and a wide selection of libraries, which have made them extremely popular in data science circles. Having a platform that supports the use of those languages is essential.

It's also important to note that for coding data scientists, flexibility doesn't have to come at the expense of productivity. Having a platform that allows them to use more augmented functionality for tasks like model prototyping and evaluation frees them up to focus more of their time creating impactful solutions.

Visual ML & drag-and-drop designers

This is one of the most common interfaces that you'll find when shopping around for data science platforms. While the terminology that vendors use may vary, what they're describing is a user interface that takes complex data science algorithms and functions and makes them available in drag-and-drop building blocks. This allows non-technical users to create processes without writing code by simplifying everything from data access to deployment.

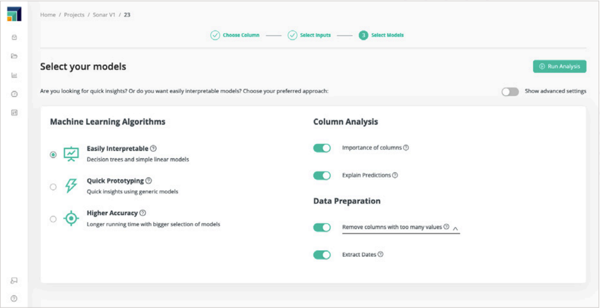


While making data science accessible to non-coders will help you bring more projects into production, bear in mind that some tools only support drag-and-drop for basic tasks and don't offer the same level of control as code. These platforms present many of their functions as being "visual," but actually require code to work in production environments. To check for this, see if a platform's drag-and-drop functions go beyond data science algorithms and help to incorporate true coding elements. A couple of examples of this would be **looping** (repeating an instruction until a specified condition is met) and **branching** (creating a central code base that multiple people can base their independent work on.)

We should also mention that the building blocks you use within a visual interface shouldn't be black boxes. Instead, you should be able to open, inspect, and edit their characteristics and behavior depending on your use case.

Automated machine learning

Automated machine learning (or AutoML) guides users through data prep, model selection, and deployment and monitoring. It's primarily used by data science novices who require more structured help, but can also be used by coding data scientists who want to eliminate busywork (e.g. preprocessing).



Select your models

Are you looking for quick insights? Or do you want easily interpretable models? Choose your preferred approach:

Show advanced settings

Machine Learning Algorithms

● Easily Interpretable
Decision trees and simple linear models

○ Quick Prototyping
Quick insights using generic models

○ Higher Accuracy
Longer running time with bigger selection of models

Column Analysis

Importance of columns

Explain Predictions

Data Preparation

Remove columns with too many values

Extract Dates

By this point, most vendors have automated capabilities for common tasks like model selection and validation. We'd encourage you to look for a tool that can further empower non-data scientists by automating the more nuanced aspects of the data science lifecycle—things like feature engineering, use case analysis, and selection of results to display in end-user apps.

It's also worth mentioning that, similar to how hiring and isolating data scientists won't yield results, solely relying on software to do the work for you isn't always realistic in data science—especially for more advanced use-cases. No matter how well an AutoML solution simplifies model creation, real-world data science projects require some tuning to ensure that models align with a given use-case and dataset. Put simply, AutoML solutions rarely deliver the type of ROI enterprises are looking for without some additional work.

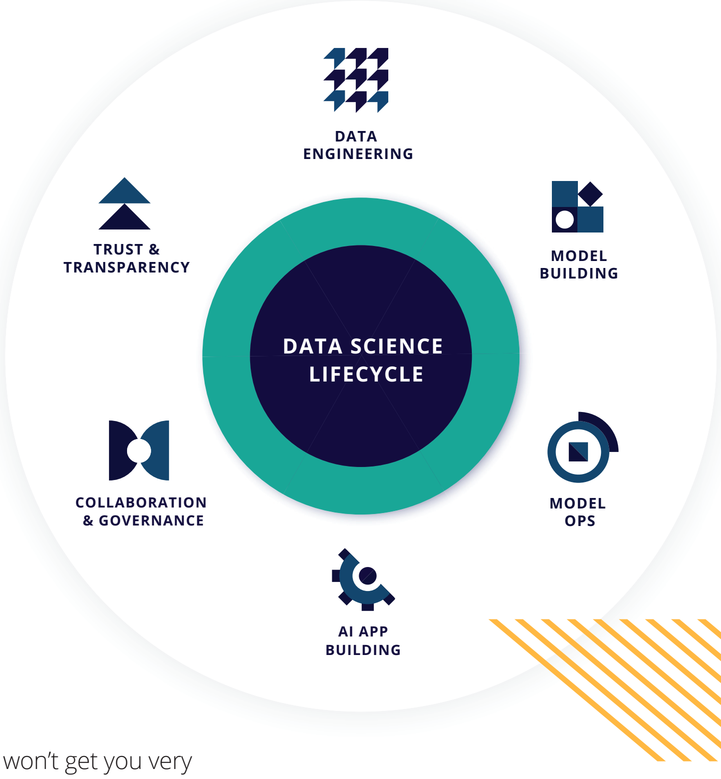
Checklist: Complete Experiences for All

	RapidMiner	Vendor B	Vendor C	Vendor D
In-platform coding notebooks	✓			
Visual ML & drag-and-drop designers	✓			
Automated machine learning	✓			
Total Score: Complete Experiences for All	●			

End-to-end lifecycle coverage

Data science projects are made up of a series of interdependent steps, and each one needs to be handled with care. Models can't make useful predictions on data that isn't prepared properly, and also won't generate the desired ROI without effective long-term maintenance.

When shopping for a platform, it's important to make sure that every step from initial data intake through drift prevention is accounted for to prevent project bottlenecks. Here's what to look for within each stage of the data science lifecycle.



Data access

Platforms that can't connect to all your relevant data won't get you very far. The safest bet is to find a platform that offers flexibility—something that can fit with your architecture today and adapt as your data management strategy changes.

It can be helpful to think about this in the form of **structured** and **unstructured** data. Structured data refers to data that can be easily categorized and fits nicely into spreadsheets and software programs (think standard customer profiles—things like names, addresses, and contact information). Unstructured data is data that doesn't fit traditional structures and can't be natively stored in relational databases (this can be text on a web page, social media posts, and images).

Having a platform that can access all of this information is critical, because you usually need a combination of structured and unstructured data to create impactful models. Let's say you want to predict how likely your customers are to churn. A model that makes predictions based on past transaction history & lifetime value would undoubtedly be useful, but a model that could also take a customer's email correspondence and social media engagement into account is a game-changer.

Data preparation

It's been widely cited that until recently, close to 80% of a data scientist's time was spent on data prep. Having clean and properly formatted data will lead to more desirable outcomes for your data science projects, but spending too much time on data prep takes away from the time your team spends on creating and testing models. That's why it's important to find a data science platform that can eliminate much of the busywork that's associated with data prep.

To start, the right platform will cover the basics like joining, appending, and removing duplicates from your data. From there, you'll need the ability to partition data for different purposes—things like separating training, validation, and test datasets.

Finally, it's crucial to find a platform that truly supports "last-mile data prep," which are the steps that are specific to data science projects. The most important of these is **feature engineering**. Platforms that can suggest useful features based on the use-cases you're working on will go a long way towards making this process more efficient.

Data exploration

Your data science platform should help your teams study a dataset to understand it's characteristics, usually through the use of charts and other visualizations. There's not necessarily an "end goal" in mind at this stage—really, your teams will be trying to get a sense for any patterns, trends, or other key areas of interest.

The right data science platform will a.) provide a variety of visualization options to help uncover areas of interest b.) generate some statistics that help users to understand what they're working with.

For visuals, look for a combination of charts (bar & area charts for example), graphs (line graphs, scatter plots), and 3-D capabilities. As far as statistics that can help users make sense of their data, it's helpful to have numerical attributes like mean, median, and standard deviation, as well as categorical attributes like number of categories and missing values.

By providing the right tools to initially explore data, you can accelerate the process of model creation by giving your teams a sense for where to aim.

Model creation & validation

Once your team has accessed the right data, made sure it's ready for analysis, and begun to identify areas that are worth investigating further, it's time to create a model.

There's no "one-size-fits-all" approach to model creation. There are usually a number of algorithms that could help with any given use-case—the trick is to find the simplest one that can yield an impactful model, then train it.

The platform you adopt should support a wide variety of algorithms and machine learning techniques, both supervised and unsupervised. It's also worth looking into vendors' ability to help with advanced data science techniques like deep learning.

Once your team creates a model that could work, they'll also need the ability to test it before putting anything into production. Look for platforms that support a variety of model validation techniques (cross & split validation, for example).

Model operationalization

Now that your team has built a model to address a given business problem, the hard part should be over—unfortunately, operationalization is where many data science projects get held up.

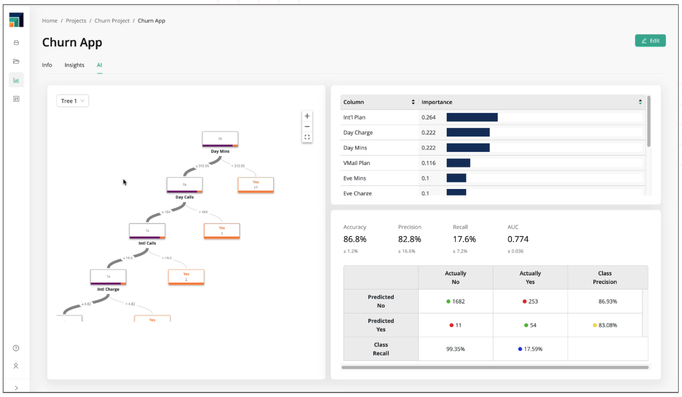
When it comes time for deployment, there are two key areas to consider: **containerization** and flexible deployment.

Containerization helps to package your models and their dependencies into an environment (container) that can run on any infrastructure. By adopting a platform that helps to create containers, you can ensure that you're accounting for more than just the code behind a solution by packaging any frameworks, libraries or dependencies that you need to in order to get a model to run where it can be most useful. Consider platforms that both provide you with the ability to create containers using technologies like Docker and also help you manage and "orchestrate" those containers using technologies like Kubernetes.

When we reference **flexible deployment,** we're mostly concerned with models' ability to fit into your data architecture today while adapting to any changes tomorrow. For example, while you may want to deploy solutions on-premise today, your organization may decide to move to a cloud-first infrastructure down the road. Because your strategy is likely to change, your data science platform should allow you to deploy models, well, just about anywhere. By giving your teams the ability to run solutions on premise, in your chosen cloud, or even at the edge in real-time scenarios, you can help ensure that the models they create won't go to waste.

Results evaluation

Once a model is successfully deployed, it's time to start analyzing predictions and using them to make more informed decisions. The right platforms will go beyond model accuracy measurements to help you quantify the estimated business impact of certain decisions. This can come in the form of projected revenue improvements, cost reductions, and overall profitability.



Column	Importance
Int'l Plan	0.264
Day Charge	0.222
Day Mins	0.222
VMail Plan	0.116
Eve Mins	0.1
Eve Charge	0.1

Accuracy	Precision	Recall	AUC
86.8%	82.8%	17.6%	0.774
± 1.2%	± 16.6%	± 7.2%	± 0.036

	Actually No	Actually Yes	Class Precision
Predicted No	● 1682	● 253	86.93%
Predicted Yes	● 11	● 54	● 83.08%
Class Recall	99.35%	● 17.59%	

Those results should also be displayed intuitively so that the people relying on them can review them with ease and play out different scenarios to determine the best course of action. That's why your data science platform should have some form of interactive dashboards, whether it's native functionality or an integration with widely-used visualization tools like Tableau or PowerBI.

Long-term maintenance

All too often, companies will operationalize and rely on a model to make decisions, but fail to properly maintain it over time. Models that aren't monitored will likely experience degradation or drift, making their predictions far less useful.

One example of concept drift would be creating a model that analyzes purchasing behavior without considering the strength of the overall economy. If a major macroeconomic event changes spending habits on a large scale, your model's predictions would become much less useful.

You'll need to make sure that the models you have in production are still the best ones for the job. One way to accomplish this is to test them against other models in deployment environments. When a new model outperforms the one you're actively using, it's time to revisit (and potentially adjust) your strategy.

It's also important to consider whether the underlying assumptions that your model is making still hold true. When they don't, you'll experience concept drift, which can have lasting consequences if it's not detected and addressed in a timely fashion.

Having a data science platform that can help you compare a model's latest performance with expected error rates, test multiple models together, and monitor for drift will help you ensure that you're getting the most value you can from your work.

Checklist: End-to-End Coverage

	RapidMiner	Vendor B	Vendor C	Vendor D
Data access	✓			
Data preparation	✓			
Data exploration	✓			
Model creation & validation	✓			
Model operationalization	✓			
Results evaluation	✓			
Long-term maintenance	✓			
Total Score: End-to-End Coverage	●			

Full transparency & explainability

As we mentioned in the previous section, the models whose steps are fully explained and documented are the ones that get deployed. It's not easy to change decision-making processes to begin with, but you're especially likely to be met with pushback if you ask a room full of people to make decisions based on predictions they don't fully understand.

That's why explainability is such an important consideration when evaluating different data science platforms. Broadly, explainability can be broken down into two major categories: **model-wide methods** and **outcome-specific methods.** Let's take a look at a few ways that vendors should be able to help you show your work.

Model-wide methods

First, it's crucial that your data science platform allows you to get a full birds-eye view of your entire data pipeline. While the model that you create is obviously an important part of that, there are many steps that take place before (intake, exploration, prep) and after (accuracy monitoring, drift prevention) model creation. For example, if the data you trained a model on is significantly different than the data it needs to analyze in production, that would impact its' performance—the trouble is, you may not be able to isolate the root cause without a full understanding of your data pipeline.

Visualizations are also extremely helpful here, because they can help you discern how a model is working on the whole and understand the logic that's used at each step in the process. Look for platforms that support visualizations such as Decision Trees and Random Forests.

Lastly, we should revisit the importance of features. We've already mentioned that selecting the right inputs is a prerequisite to creating an impactful machine learning model, but it's equally important to understand the roles those inputs are playing in generating predictions. The right data science platform will help you evaluate how much weight a model is giving to each individual input through **global feature weights.**

Outcome-specific methods

Like global feature weights, **partial dependencies** play a key role in understanding the relationship between inputs and predicted values. They do this by showing how a model responds to changes in a single input value. An example of this would be seeing how heavily a model weighs the length of a customer's relationship with you, as opposed to say, the overall amount they've spent. Look for a platform that can help you map and understand these relationships.

> **An example of this would be seeing how heavily a model weighs the length of a customer's relationship with you, as opposed to say, the overall amount they've spent.**

True explainability also requires an understanding of **local feature weights**, or the weighting of features for a single or small set of data points. This helps to troubleshoot unexpected predictions on an individual level. Your data science platform should support the ability to calculate local feature weights through industry-standard methods like **LIME** and **SHAP.**

Lastly, platforms that allow you to explore these relationships interactively help to answer any "what if" questions you have about the relationship between your inputs and predictions. By seeing what happens to predictions when you intentionally manipulate different model inputs, you can not only see how a model may behave in production, but also check that behavior against well-established domain knowledge and expertise.

Checklist: Full Transparency & Explainability

	RapidMiner	Vendor B	Vendor C	Vendor D
Model-wide methods	✓			
Outcome-specific methods	✓			
Total Score: **Full Transparency & Explainability**	●			

Centralized data & project assets

The organizations who are finding the most success with data science are taking steps to make sure that their efforts compound. Rather than engaging in one-off projects and moving on, they're looking for ways to share their work across different teams and functional areas so that others can benefit from what they've already done.

Your data science platform should help with this—not only by helping you foster a collaborative environment for data science, but also by centrally storing your projects and their related assets. By creating and providing access to a repository of past work, you can ensure that any successes and learnings scale across your organization.

Data & project repository

In many cases, the data used for a specific use-case will also apply to others, which is why having a central data repository is essential. A well-organized repository allows others within organization to benefit from your previous work by saving time on things like data connections and preparation steps.

A project repository provides shared access to project components such as models, processes, and results. Even if those components aren't explicitly reused, your teams can reference them when they're developing solutions that are similar to those they've worked on in the past.

Security, auditing & governance

Ensuring that your data and project assets can be properly secured and governed is a prerequisite for any data science project. In recent years, high profile data breaches have put customers' information at risk and had long-term brand implications for the organizations who've suffered them—here's what to look for in a data science platform that'll help you avoid that.

Authentication & authorization

Like any other business application, your data science platform should give admins the ability to effectively authorize users while controlling what they can access. Strong authentication and authorization protocols have become a standard aspect of enterprise security, especially as more employees seek to access work apps from multiple devices.

Your platform should support Single sign-on (SSO), two-factor authentication (2FA), and authorization protocols such as OAuth 2.0. By helping you verify users' identities and giving you fine-grained control over what they can access, your data science platform should help you ensure that sensitive information can never be accessed by unauthorized parties.

Data encryption

Encryption is one of the most crucial steps you can take to protect your organization's data—when handled correctly, it can impact several areas of security like authentication and data integrity.

Look for platforms that support widely adopted protocols such as Transport Layer Security (TLS) to make sure they support enterprise standards for privacy and encryption.

Auditing & data lineage

To deploy a model, decision makers within your organization need to trust the data its predictions are based on. This includes where it originated and how its been transformed over time.

Your data science platform should give administrators and project leaders the ability to audit processes and models over time. Automatic logging and versioning capabilities are table stakes—functionality that allows you to quickly view commit history and modifications to a process will help trace each transformation back to a member of your team.

Checklist: Security, Auditing & Governance

	RapidMiner	Vendor B	Vendor C	Vendor D
Authentication & authorization	✓			
Data encryption	✓			
Auditing & data lineage	✓			
Total Score: **Security, Auditing & Governance**	●			

The DSML Platform Checklist

The data science market is filled with tools that are designed for specific personas or purposes. Here's a look at how some of the common categories stack up.

	Complete Experiences for All	End-to-End Coverage	Transparent & Explainable	Centralized Data & Assets	Security, Auditing & Governance
RapidMiner	●	●	●	●	●
AutoML	◕	◑	◕	◑	◕
Open-Source Only	◔	●	◔	◔	◔
Drag & Drop Tools	●	◕	◑	◕	◕
Cloud Service Providers	◔	●	◔	●	◕

Your DSML Platform Checklist

Use this template to rate the vendors you're considering for your digital transformation. Add up the total scores from the previous, more detailed checklists and input them here.

	Complete Experiences for All	End-to-End Coverage	Transparent & Explainable	Centralized Data & Assets	Security, Auditing & Governance
RapidMiner	●	●	●	●	●
Vendor B					
Vendor C					
Vendor D					

To Wrap-Up

To find success with enterprise data science, you need to find ways to bring it closer to your actual business problems. Too often, organizations rely almost exclusively on coding data scientists—in-house or outsourced—to handle their projects. By prioritizing an upskilling-based approach, you can empower anyone who understands a business problem to generate data-driven insights that are easy to share, consume, and use to transform processes.

Data science software can help drive such an approach by providing one platform for every project and stakeholder, supporting the entire AI lifecycle, and making results easy to explain and understand without compromising governance or security.

By pairing the right strategy with the right tools, you can begin to deliver on the promise of enterprise data science and transform the way you do business as a result.