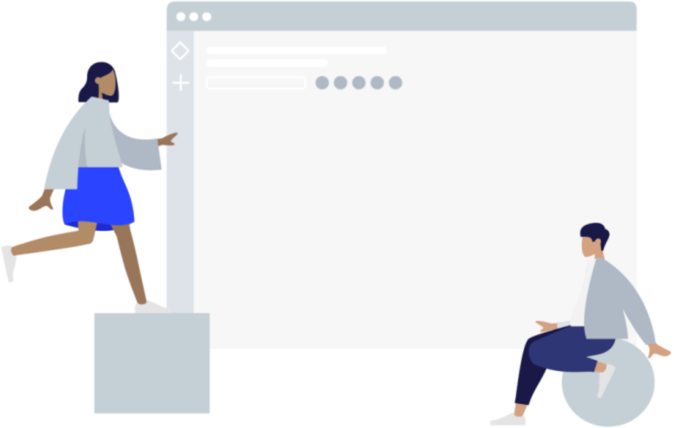
ALGORITHMIA

OCTOBER 24, 2019

Harnessing Machine Learning for Data Science Projects



We hear more and more everyday that businesses are sitting on troves of valuable data. It's compared to precious metals, unrefined oil, or cash in a vault. But those items aren't valuable simply because they exist. Their value comes from what is created out of them. The same holds true for data. Rows full of numbers and text only become useful when you can tell stories and draw insights from them.

For those less familiar with data-driven business initiatives, the path from raw data, to extracting insights, to making decisions based on those insights may seem like a black hole. But like any process of turning a raw material into a valuable product, there is a system to follow and a way to avoid the black hole. In the case of data, it comes in the form of data science projects.

The intent of this article is to guide you through the process of creating and executing a data science project, including selecting machine learning models most appropriate for your goals. While this is written in the business context, this process is relevant to those working on personal projects as well.

What is a data science project?

A data science project is the structured process of using data to answer a question or solve a business problem. Conducting data science projects is becoming more common as more companies become more proactive about finding value in the data they have been storing. Common goals for these projects include:

- Developing more targeted and effective marketing campaigns
- Increasing internal operational efficiency
- Revenue forecasting
- Predicting likelihood of default (banking/financial services)

Prompting a data science project

There are two common scenarios in which a data science project might start. The first begins at the top of an organization with directives from senior management. They may have outlined specific problems to be explored and are looking for employees to find opportunities for improvement through the use of data. It's common for organizations like this to have data scientists or senior analysts embedded in divisions of the organization. This helps them obtain the relevant business knowledge, in addition to their technical skills, to draw out relevant insights.

Data science projects can also begin at the individual level. It's not uncommon for an employee to notice a problem or inefficiency and want to fix it. If they have access to the company's data warehouse and analytics tools, they may begin their investigation alone before bringing others in on the project.

An example of a data science project

A data scientist at a brick-and-mortar retailer may be tasked with developing a predictive model to judge the likely success of new locations of the store. The business goal of this project is for real estate and facilities division team members of this company to understand the success of other established locations and use this knowledge to guide decision making in future transactions.

Note that we will use this retail location example and variations on it for the entirety of this piece to further emphasize points.

How does machine learning fit into a data science project?

Before getting too far into this discussion, we need to define a few terms. There is often some confusion between machine learning and data science, with some individuals believing that one is “better” than the other, or that they are somehow mutually exclusive.

Data science is an encompassing term that refers to a discipline whose main pillars are:

- Mathematics, specifically statistics
- Computer science
- Business acumen and domain knowledge

Machine learning is a subfield of artificial intelligence. It is the process of using algorithms to learn and understand large amounts of data and then make predictions based on specific questions asked. Machine learning regression modeling is where math and computer science intersect, as it takes compute power and a knowledge of programming to develop and build on these statistical models.

From these definitions, it should be clear that machine learning is a vital component of data science. It is the bridge between raw data and solving business problems. You will need to build models and validate them before drawing any conclusions or providing recommendations.

The data science workflow and project process

When beginning your data science project, it’s useful to frame it as a series of questions that we will discuss in detail.

1. What business problem am I trying to solve?

While your personal projects don’t necessarily require a specific focus, businesses are looking to reach certain targets like increasing revenue, cutting costs, operating more efficiently, decreasing customer churn, etc.

With that in mind, consider how the answer to your project question would influence the business. Ideally, it would give the company the information it needs to develop a plan of action.

Let’s illustrate this using our retail store example. Instead of asking “Which store brought in the most revenue during Q2?” frame it as “Why did store 123 bring in the most revenue in Q2?” The first question gives you a simple answer that probably can’t be acted upon without further research. The second question suggests that recommendations can easily be extracted out of the answer.

If you are unsure of the question you want to ask, it's helpful to first engage in exploratory analysis—making visualizations and small manipulations of the raw data, especially in your area of the business. If anything jumps out, or looks like an opportunity for further research, you can begin your question there.

2. Do I have all of the data I need to answer this question?

To develop a predictive model about retail store success, you probably need some the following information:

- Store address
- Type of location (In a mall? Standalone building?)
- Revenue by period
- Square footage
- Daily traffic
- Number of employees per location

Your company likely has all of this information, but it's probably stored within various SaaS applications and databases. In addition, you may need some information from publicly available data sources like demographics, population, and weather trends, to round out your picture of the location.

3. How will I put everything together in a manageable form?

Combining data sources into a form that you can analyze usually involves the ETL (Extract, Transform, Load) process through the use of one or multiple tools.

Here's an overview of ETL:

- **Extraction** – The process of pulling data from various sources (relational databases, SaaS applications, etc.).
- **Transformation** – Data undergoes a series of changes based on rules that meet the requirements needed for analysis. This step includes data cleaning and normalization (putting numerical values in standard units).
- **Load** – Extracted and transformed data is sent to the end system, usually a data warehouse where it can be linked to an analytics tool.

4. How will I approach the analysis?

Before deciding on the machine learning model you will use (we'll get into some actual use cases in the next section), think about how you would frame the answer to your question. Maybe you're going to make a prediction or possibly uncover segments. What you choose to do will depend on the type of data available to you and your business goals.

5. How will I communicate my results to a broader audience?

In other words, what do you plan on doing with the results of your data science project? For example, will you create a dashboard, send a report to interested parties once a month? Or only discuss when asked about it—remember, you are trying to provide value to the business. This is particularly important point to keep in mind for self-directed projects.

Which algorithms are used for machine learning?

Machine learning algorithms can be broken down broadly into two methods: supervised learning and unsupervised learning. A supervised method requires there to be a defined target with data to compare it to. An unsupervised method does not have any specific target.

Let's illustrate this difference with two questions related to retail stores in our hypothetical example.

- **Unsupervised**: Do our retail stores fall into natural groupings?
- **Supervised**: How can we identify stores with a high likelihood of converting customers into store credit card holders?

The supervised question has an explicit target: we want to find stores that share a business-specific characteristic. The unsupervised grouping isn't looking for anything in particular.

It's important to note that neither of these methods is "better" or more useful than the other. Their value depends completely on business goals. An unsupervised method is particularly useful when trying to uncover segments that don't appear obvious by just looking at data laid out in spreadsheets.

In our retail store example, once placing stores in natural groupings, business teams might be able to use their domain knowledge and intuition to infer something about these stores that is not explicitly laid out in the data. The supervised example is useful for a company that has a goal in mind, and wants to bring all stores up to the level of the successful ones.

Supervised machine learning methods

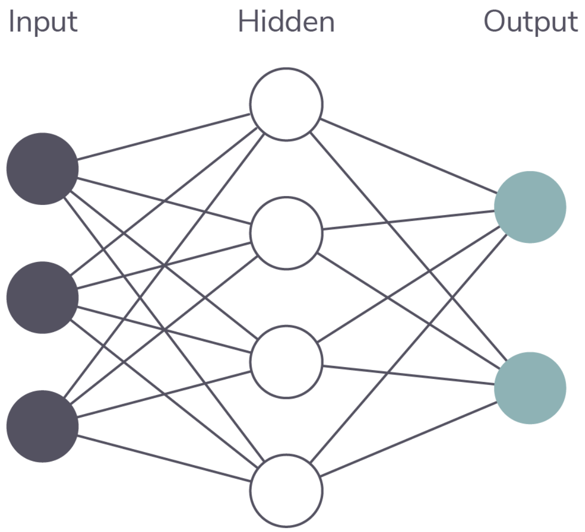
- **Regression** – This is a predictive data science algorithm that explores the relationship between a dependent variable and one or more independent variables. The output is always a numeric value. Continuing with our example, you could use a linear regression to predict a new loctaion's potential revenue, given a set of numeric variables.
- **Classification** – This is a predictive method used to determine which category a new observation belongs to. The target output is two or more categories, often framed simply as "yes" or "no." Example: Given the data we have about other store locations, and our definition of success, should we open a new store in this location? Yes/No.
- **Class probability estimation** – A binary classification is not always useful in every situation. Even our retail store example requires more nuance than a simple yes or no. This is the advantage of class probability estimation, which predicts the likelihood that a new observation belongs to a specific class. Example: Given the data we have about other stores, and our definition of success, what is the likelihood this new store will be successful? The output is a numeric estimate between 0 and 1.

Unsupervised machine learning methods

- **Clustering** – The unsupervised question examples earlier would probably lead a data scientist to develop a clustering model. Clustering means grouping observations based on similarities. It's also a form of exploratory data analysis. When interpreting clusters, you will need to look at the underlying components of each group, conduct summary statistics, and compare this information to other groups. It's important to determine if these clusters have any significant meaning based on your knowledge of the business.
- **Dimension reduction** – When attempting to analyze multiple large data sets, you can run into the problem of having too many variables that are intercorrelated. Dimension reduction is the process of eliminating redundant variables in a data set. This is a reduction of the number of variables in a data set. Breaking down data into vital components can be analysis in and of itself, or it can be a first step in refining linear regression models. A commonly used dimension reduction data science algorithm is principal components analysis (PCA).

Neural networks and how they fit into data science algorithms

Neural networks have come in and out of fashion in the computer science and cognitive computing communities for the past seven decades. They have seen a resurgence recently because of an increase in compute power and more practical applications of the technology. Neural networks are also the underlying architecture of deep learning AI.



While neural networks are really their own discipline, we'll discuss them briefly here. Neural networks have three parts: the input layer, output layer, and hidden layer. The input and output layers are part of almost any algorithm—you provide data, and the computer returns some information. The hidden layer is the interesting part. You can think of it as a stack of algorithms (supervised or unsupervised), that build on each other until it reaches a final output.

Neural networks are often referred to as "black boxes," meaning you don't really have an understanding of the "thought" process. In some situations it may be fine not to know, but in other business contexts like financial services and credit scoring, this lack of transparency can be problematic. Keep this in mind if you are considering incorporating neural networks into your data science project.

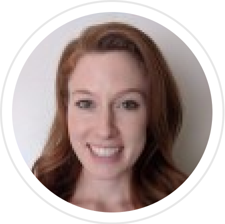
An additional risk of neural networks is that they can fit training data too well, and become irrelevant when trying to analyze general population data.

The importance of data structures and algorithms in data science

As we mentioned earlier, the technical component of data science skills is where math and computer science meet. Having a foundation in statistical methods is essential to data science, as is having an understanding of not just programming, but computer science itself.

Data structures and algorithms are the foundation of computer science. A data structure is an organized way of storing data and using it efficiently. And as discussed, an algorithm is an unambiguous, finite, step-by-step procedure to reach a desired output.

So why is this important to a data scientist? For one, developing algorithms for data science projects is not a one-time task. You will be constantly refining the model with new variables and rows of data. With more data comes more demands on processors and records that take longer to access. Large-scale data science projects cannot be efficiently modified or replicated without the base understanding of how data is organized and processed in a computer. Data scientists should not be reinventing the wheel every time they develop an algorithm. Instead, they should be thinking about how an algorithm can be easily scaled and reproduced.



Whitney Nistrian

More Posts

Search

Enter your query here...

A.I. Topic Guides

Algorithm Spotlight

Blog Posts

Content Hub

Demos

Developer Spotlight

Emergent Future

Events

Integrations

Newsletter

Algorithmia

AI in every application.