Module 1

Video Transcripts

Video 1: Why Learn Data Science? (4:09)

Hi, and welcome to week one of Python for data science. In week one, we're going to give an, a general overview of data science. So, why would one want to learn data science? Data science is an incredibly exciting field that allows you to be a lifelong learner. So, just as our ability to implement algorithms that work on larger datasets has increased over time.

There's going to continue to be more tools and more software that allows us to work with larger datasets, help us make even cooler discoveries and insights, and we're going to, as a data scientist, have to keep up with the evolution of these tools as, the field evolves. It is absolutely the opportunity to be a detective. We always start a data science problem with a question that we're trying to solve, and our job is to find the relevant information to answer that question, leveraging data.

We can impact business strategy directly. Companies have been making a move towards making all of their decisions informed by data. And so, when a question arises, "What's the next move we should make?" It may be the job of the data scientist to figure out what would be the most likely profitable move for the business. Data science is relevant across so many different industries, whether we're using image data to diagnose cancer, whether we're fighting global warming.

The ads that you see that pop-up on your browser window are specifically targeted for you, and they are powered by data. I've seen fantastic examples of people using data around their populations to determine city planning, and the list really goes on. Another reason why we might want to study data science is that the current demand was over 10,000 data scientist positions listed on Indeed.com in the beginning of 2020. That's a lot of positions, and it's actually up over double from where it was a year ago.

As I already said, the idea of moving to a data driven or a data centric business is a really hot topic. People don't want to make decisions with their gut anymore. They want to make decisions based on data, and you are allowed to help make recommendations for those big business decisions. The amount of data that we collect is increasing at a rapid pace, and the tools are constantly being created and modified to help manage this.

And so, as we said, you get to be a lifelong learner, and you get to see the evolution of this field. And personally, myself, I've been a data scientist in industry, for over 10 years, and I've always felt that my opinions are valued and respected, I've found the work fulfilling, and I've been lucky to have reasonable working conditions and flexibility that have given me good work-life balance. So, in summary, if you consider yourself a lifelong learner, data science is a fantastic area to be with a lot of demand for work opportunities and fulfilling work.

****

Video 2: What is Data Science? (3:01)

Alright, let's continue our introduction to data science. So, what is data science? The term data scientist was only coined in 2008, although people have been doing predictive analytics for much longer than that. And the term continues to evolve what it represents as the role changes, keeping up with new technology. So, a model that you might have built eight years ago, that potentially took over the weekend to train, might train very quickly now, with the increases in computing power.

So, this is also changed what the day-to-day role of a data scientist actually looks like. My definition of data science is that it is the understanding and utilization of tools, data, and methodology that enable you to effectively solve problems utilizing data. And so, it's really an interdisciplinary field that involves coding to get at and manipulate your data, understanding principles of mathematics and statistics.

Because it is mathematics that's powering under the hood, these models, and the business domain expertise, so that the recommendations and the things that we're predicting, actually add value for the business.

So, you'll notice that in the intersection of the Venn diagram, we have the term unicorn, and that has sort of been, what the industry has called this person that we're looking for that's mystical, that is an expert programmer, an expert statistician, and has a wealth of business expertise. And so, over time, we've realized that that person doesn't exist, and a data science team is going to be made up of people who are going to build on each other's strengths.

So, you'll have somebody who is an incredibly strong coder, and they're the subject matter expert for coding on the team, and you will have somebody who has a deeper knowledge of statistics, and you'll have the people who are really well rounded and understand the business well, and as a team, you'll work together to arrive at the best results. So, in summary, data science is an interdisciplinary field where you utilize tools, data, and methodologies that enable you to solve problems utilizing data.

****

Video 3: Essential Data Science Tools (7:21)

In this video, we're going to be talking all about the tools the data scientist has in their toolbox. So, what are the data scientist’s tools? Typically, they leverage a combination of SQL, Python, Git and GitHub, and big data technologies to do their daily work, and we're going to go through all these different types of tools, and give you an intro into what these are. One important note about this is that the tools are going to change over time. Right now, Python is the most popular in the industry.

Python and R are the programming tools you're going to hear the most about when you read articles on the internet. However, there are other programming languages such as Rust, and Julia that are also good for data science tasks. And so, as time goes on, just as we've seen in history, the tools that we use are going to change. So, prepare to be flexible. So, SQL is a query language for relational databases, and I highly suggest that you become familiar with it. Organizations, typically save their data in large databases, so in tables that are columns and rows. So, these are two-by-two tables of all their transactional data, the data that they have on their customers, maybe where people are clicking on their website or the actions that they're taking, which emails they've responded to. And so, this example that I'm giving is in e-commerce, but you can imagine that companies in general, are storing large quantities of data, and a lot of it is in two-by-two tables in this relational database format, and there's many different flavors of SQL.

So, there's Oracle, Postgres, MySQL, SQLite, Microsoft SQL Server, and they are all using a Structured Query Language, which is what SQL stands for. So, this is a standardized language. So, although there's going to be some slight differences between Postgres and Microsoft SQL Server. Once you understand how relational databases work, and how to query them, you really can learn any different language rather quickly. So, I always advise people that, let's say you learn PostgresSQL.

Don't let that keep you from applying to a job that is asking for somebody who is familiar with MySQL because really it's the same, and once you learn SQL and you understand how to read the documentation, you can easily read the documentation for what I refer to as a different flavor of SQL. Definitely don't let it limit you, and don't put too much thought into where you get your start with SQL, just start learning the language.

So, certainly don't try and learn all the different types of SQL. It's not really going to help you, but on your first day as a data scientist, your boss is going to say to you, "Welcome to the company. We house all our data in this data warehouse or in the database, and this is where you're going to query your data to answer these business questions." Okay, and then we have, of course, our programming languages. And like I said, Python is currently the most popular programming language in the field, and people typically ask, "Should I learn Python or R?" And honestly, my response is "Yes."

I had heard this from Kirk borne, and the thought behind it is that you do need to learn a programming language to be a, an effective data scientist, but the language that you learn shouldn't be too much of a concern as long as you can be effective. However, the most popular programming language right now on job descriptions is Python, and it is sort of taking over the market share. The concept of version control is incredibly important in data science. There's nothing worse than emailing your code to somebody else or having Excel files with version 45 at the end of them.

That leaves open plenty of area for people to make mistakes, to be leveraging old code, to end up with two analysts doing analysis and ending up with different results. And so, to avoid this, we keep all of our most recent versions of our code up on versioning control software. And so, Git and GitHub is the most popular software for doing this on the market right now. Nowadays, you can't talk about data science tools without talking about tools to access big data, and there is a wide variety of different software that companies are now using to manage their big data.

So, this includes Hadoop, Spark, Amazon Web Services, the list goes on, and I think my big takeaway here is that big data is not scary. So, if you were using Hive in Hadoop, you'd realize that Hive looks very much like SQL with a couple different capabilities for dealing with unstructured data. And Spark, if you're writing in PySpark, it's going to look a lot like Python. So, years ago when people were talking all these different buzzwords about big data, it looked really scary, but once you get in there, and you have the foundations of Python and understanding how data is typically stored.

All of a sudden, this doesn't seem like such a huge hurdle. So, in summary, there are a number of tools that data scientists use in the day-to-day to complete their work. And those include SQL, a programming language, in this case, we're talking about Python in this course. We use version control software, and the most popular in the industry is Git and GitHub, and we also leverage big data technologies to allow us to access our big data.

****

Video 4: The Data Science Lifecycle (11:41)

Hi, again. In this video, we're going to discuss the data science lifecycle or pipeline. Maybe you haven't heard before about the data science lifecycle, but all data science projects typically follow a similar pipeline. First, we need to scope the project, and this typically involves meeting with stakeholders, really trying to nail down the question that we're trying to solve, and what success looks like. So, this may sound simple, but success needs to be quantified in a certain way, and we need to be really clear about what it is we're trying to predict.

Next, we'll gather the data that we believe will help us to answer that question, and then we go into an iterative process of cleaning the data and doing exploratory data analysis. We're going to go over this a little bit more in the next couple of slides.

So, after we've done all our cleaning and all our exploratory data analysis, then we have the opportunity to build a model on the data, and once we've built that model, we're going to want to validate and test the, how the model is performing, and if it's not performing well, we might go back and gather more data, clean the data, look at the data, or even if we don't go all the way back to step two, we at least need to go back to step five, and maybe try a new type of model, or maybe we just need to adjust the parameters of our model, and then we'd go back to testing and validating again.

After we have a model that we're happy with; we move to the important step of interpreting the results, so that we're able to communicate with our stakeholders exactly what's going on and the types of recommendations that we might make around this model. And so, we interpret the results, we create a presentation, and then we go, and we present that presentation to stakeholders. And once we have buy-in from the business that this is a model that we want to move forward with. We'll put it into production, and that is the step where we automate.

So, often in blogs online, they don't focus as much in the pipeline on scoping the project and presenting recommendations, but those are very important steps that you will absolutely run into in your job as a data scientist. And so, just as you saw that there might be some back and forth, I don't want to undersell the amount of back and forth that there is, and we're going to see that in the next couple slides.

Of course, scoping and presenting the results requires us to talk to stakeholders, but the techniques that allow us to get through this pipeline are things like the Data Cleaning, the Feature Engineering, which is, part of both exploratory data analysis and data cleaning. We have Analysis, Statistics, Modeling, and Prediction.

So, data cleaning is taking the data that we've aggregated, and making sure that it's accurate, are their duplicate rows? Are there missing data in our rows? And if there's missing data, is it something that we can re-code as a zero or do we need to impute that? And imputing means that we find a logical value to put in place of the missing value based on the data. We move on to feature engineering. Feature engineering is most often creating variables out of a combination of other variables that we already have.

So, the example that I give here is that if we're in an e-commerce business and you have a customer that's purchased 400 times, and another customer that's purchased 20 times; it might look like the person who's purchased 400 times has, you know, an incredible usage rate. However, maybe they've been a customer for 15 years, and maybe the person who purchased 20 times has only been a customer for a week, right? So, if we didn't divide that by the length of time that someone has been a customer.

We'd get a really unrealistic picture of what was going on. One hot-encoding, which we typically care about in data science is when you take a categorical variable, and for each category, you instead give it its own variable that's binary. So, if we had a categories that were cat, dog, and horse, we then have a variable cat, and it would be 1 when it is a cat and 0 when it was either a dog or a horse, and we can continue to create the three variables for that one variable and use that instead.

Analysis: so this is the "A" in Exploratory Data Analysis, and this is the step where we look and determine what the distribution of our data looks like. And we'll be talking about distributions when we get to the statistical portion of this course. But basically, it's a series of tables, charts, and graphs that allow us to get a full understanding of our data, the outliers, and how correlated certain variables are with other variables so that we have a full picture of our data before we try and build a model. You'd never want to build a model without first understanding the structure of your data.

Okay, so data scientists can often also be the subject matter experts on designing, scoping, analyzing hypothesis test for the organization. In addition, when we see model output, a lot of times there is stats involved. And even in linear regression, which is the first machine learning algorithm that people typically learn, you're assuming normality of the data, right? So, there is statistics embedded in all of these tools, and understanding of statistics also helps you to understand when a methodology is appropriate, and more importantly, when it is not appropriate.

Modelling is the act of taking your data and building a statistical model or a machine-learning model for the purpose of understanding the data with multiple variables, right? So, when we do our analysis, we're typically looking at a single variable, and how it's distributed, but what are the effects on that variable controlling for maybe 20 variables, right? So, this allows us to get at the effects of each variable when we are controlling for confounding factors. So, simple models are always preferred over complex.

If you can do something quickly that delivers the same amount of accuracy or close to the same amount of accuracy as doing something more complex or computationally intensive. We would always prefer to use the more simple model, especially more interpretable models that make it easier for us to explain what's going on to our stakeholders, and then our stakeholders have more confidence in the work that we're doing. And once we have our model like we were discussing, we go ahead and make a prediction. So, we've already talked about Descriptive analytics, right?

We didn't talk about the term, but there's four types of analytics, and Descriptive is when we're looking at historical data and talking about what happened, which is what we do in analysis. Diagnostic is finding root causes. So, in the data cleaning process, if there was a duplicate row in our data, we'd want to dig into that and find out, is this a true duplicate?

Did this person purchase twice? Why is this data coming through this way? And we want to, not only ensure the integrity of our data, but dive in and find out what caused any errors or problems, so that we can correct it, and hopefully see it less than the future.

And then we have Predictive analytics. So, we can make a prediction based on our historical data, but even more importantly, right? So, if I say that, "Bob is going to churn with probability 0.98." Well, that actually doesn't help us; what we really want to get to is Prescriptive analytics. What could we do to keep Bob knowing that he's likely to leave, and we know the variables that are driving that probability?

Which of those variables might give us some insight into actions that we could take to potentially change that behavior? Okay, so in summary, we talked about the techniques that we use to get us through this data pipeline. And so, we start with scoping the project; we then gathered data, we then clean the data and do analysis, and this is often an iterative process because I will try and clean the data, then I'll go to look at it, and I'll find something else that needs to be cleaned.

So, I'll clean that up, and then I'll look at the data again. Once my data is cleaned, and I understand the structure of the data, I will then model the data, I will validate my model, I will make a prediction, then I interpret the results. I present recommendations based on my findings to my stakeholders, and then we put the model into production if it's found that it's useful.

****

Video 5: Adopting a Data Scientist's Mindset (4:55)

Alright, welcome back. In this video, we're going to discuss how to adopt a data science mindset. So, what are data scientists thinking? So, this really relates to the data science soft skills that are required for you to be effective at your job. You want to make sure that you're asking questions, that you communicate effectively, and that you're able to balance the constant struggle of getting really accurate results in your model with the ROI of spending that extra time.

So, like I said, data scientists need to ask questions, and really there are questions at every single step of the process, and it's how you communicate with your stakeholders and get these questions answered, and dive deep into making sure you're completely aligned on what it is that you're looking for that is going to help the project be a success. So, of course, we want to scope the problem. So, we're going to talk about what the problem is, how do we define success?

But then we want to refine that, and really make sure that we're truly getting at answering the stakeholders' question, and you may think it's obvious, but it is really subtle sometimes. So, we're asking ourselves, who has the business domain knowledge or the institutional knowledge that's going to get us the answers that we need to be most effective in gathering data, building this model, making sure that the output make sense. What data is available? What is the deliverable? Is it an email with some answers? Is it a dashboard? Is it a PowerPoint presentation? Is it a model in production?

And so, that's going to help you determine the level of effort that you might want to put into the analysis, right? When is the deadline? How extensive do you expect this to be? What types of factors do you want me to look at? What are the caveats? What's the population that we're including? You know, what are the areas for improvement? Where are we going after this? What are the next steps? What decision is going to be made based on this analysis, right?

And that will all help us to determine who we should be speaking to, how much time we should be spending on the project, what the final deliverable should look like, and making sure that we deliver what was in the other person mind, right? This person has an expectation, our stakeholders have an expectation of what it is we're looking, they're looking for, and unless we ask enough questions to fully get at, what is their expectation, we're not going to be able to deliver to their expectations.

Okay, so communicating effectively. Yes, as a data scientist, you'll have the opportunity to build models, but so much of the job is really communication, right? When we're scoping a model, or even if you come up with an idea for an analysis or a predictive model that can have a big impact on the business. You have to go around and sell it a bit, and try and get buy-in to get the time to work on that project that you want to work on.

And because not all data scientists are good at communication, if this is your strong suit, it can really help to differentiate you. And as we already talked about, in asking questions, you really want to understand whether or not an analysis should require three weeks or should it require 30 minutes, and the results just be dropped into an email.

So, although we're answering these questions with data, and we need to follow the techniques and methodologies, and know about coding, and know about stats, and have the ability to implement those things. Thinking like a data scientist really involves asking a lot of questions, communicating effectively, and being able to determine how long you should be spending on projects.

****

Video 6: Core Principles: Collaboration, Reproducibility, and Ethics (9:10)

Welcome back to our introduction to data science. Here we're going to talk about collaboration, reproducibility, and ethics in data science. Data science is an incredibly collaborative field, for a number of reasons. So, first, you're typically on a team, and that team is going to have a set of different skill sets. Where people are really strong in one area, and they leverage the team to help them uplevel their skills in the other areas and they work together.

So, you will have somebody who is typically a little more experienced in terms of coding, somebody who is a little bit better at the stats and modeling aspect of data science, and someone who is more fluent in the business acumen side. So, that we can all work together to solve problems effectively leveraging data. So, even though you collaborate within the team itself, you're also collaborating with stakeholders. So, data science is typically a support function.

You are supporting the marketing team, the supply chain team, operations, or depending on the industry, you may be supporting some other area of the business, or multiple areas of the business. This means that you're often going to have meetings outside of the data science team, and it's your job to collaborate with them. And they may not have the data skills, and the ability to access, and aggregate, and size opportunities leveraging data, but they're able to tell you the problem that they need solved, and you'll work together to refine the business needs in a collaborative fashion.

And then as a team in data science, you'll also be working with your coworkers in a collaborative way to make sure that you are delivering the best analyses possible. And part of delivering the best analyses possible is that your work needs to be reproducible. So, this saves tons of time and efficiency if somebody else is able to pick up the work that you've done and recreate it. A lot of times, after you deliver an analysis, a stakeholder is going to have follow-up question.

So, your work needs to be done in a way that allows for easily picking up where you left off, and continuing to dig in and this is why Python is so important. You probably can't fathom at this point how much quicker Python is going to make your workflow, but imagine this. If you've done work in Excel before, and you had to manually change the variables so that they are relevant for analysis, and then later you find out that you actually need to bring more data into the analysis.

Now you've got to go back into the database, you've got to join on some primary key, probably using a VLOOKUP in Excel, and all of this is manual and you may even need to start from the beginning and remember all of those manual steps that you did. When you're writing it in Python, it's literally click, click, click, click through your notebook. You know, you'd go back, you'd add the data, and then it would take you less than a minute to get back to having the data in that cleaned, transformed state that you had got it in. This also helps to build trust.

If somebody else was to reproduce your analysis and get different results, it's going to cause problems in terms of people questioning your analysis and why the results were not the same. And so having a common version of the truth, and being able to reproduce your results, both if you make a mistake, and you need to go back and add more data or change something, and being able to reproduce that quickly, which is much different than Excel, right?

Where you may have, you know, done certain steps one way the first time, and if you had to go back and recreate that, you might do it slightly different the second time. And it also improves your ability to work as a team because if your code is there available in the most recent version, and you are able to share it with a coworker, who can easily run your code, and get the same results, then if priorities shift, and you're required to work on something else, you are able to help onboard one of your teammates to continue that analysis.

It also helps in hiring when, you know, somebody more junior comes in after you, after you've been working a couple years, and you're able to teach them, you know, your process and how you arrived at things. And they're allowed to continue on with your analysis, and the integrity of that analysis is still intact. Okay, so let's talk about data integrity. So, there's two types of integrity going on here. One, data always contains some element of bias, and we want to be aware of those biases and that's why statistics is so important in data science.

It's that understanding where that bias is introduced in terms of how the data was collected, who's not included there. You know, really what does this population represent? We're able to then correctly list the caveats and communicate those with our stakeholders that we're collaborating with. There's another type of integrity that's also the data integrity. We want to make sure that we're preserving the validity of the data through checking, validation, and following up on errors that we see in the data.

So, there's often logic in the database that is calculating certain fields, and you may find an error and so it's often our job to get to the root cause of where that error is. Or bring it up with the BI team, or the data engineering team that is going to handle that data massaging, and data, and making the data available. So, it is possible to find duplicates, missing rows that shouldn't be missing.

And it's our job to follow up and make sure that our data is of the highest integrity, because if it is not, it will affect us when we're going to build machine learning models. Junk in leads to junk out, right? So, we need to make sure that the data that we're working on is correct if we're going to be reporting out on it. Or at least, you know, minimize bias as much as we can and when there is bias, we call it out.

In addition to data validity and building more on biases, we want to make sure that the biases that are in our data are not related to something like racial, gender, or socioeconomic biases that could be hurting people with our models. A fantastic book on the subject is "Weapons of Math Destruction" by Cathy O'Neil, and she goes over all sorts of interesting case studies, including you know, putting law enforcement in certain areas, and predatory education.

And so if you are interested in the topic of ethics in data science, I highly suggest you read her book. So, in summary, collaboration, reproducibility, and ethics are a huge part in data science, and something that we need to be constantly aware of to make sure that the models are of the highest quality that we're not harming anyone, and that our workflows are efficient.

****

---**END**---