

Machine Learning For Beginners

CAPTECH TRENDS | PODCAST | EPISODE 13

Proprietary and Confidential • © 2020 CapTech Ventures, Inc. All Rights Reserved.

Vinnie

Hello, and welcome to CapTech Trends, a place where we meet with thought leaders and subject matter experts to discuss emerging technology design and project methodology. I'm your host, Vinnie Schoenfelder, Principal and Chief Technology Officer at CapTech Consulting. Today, we're discussing machine learning for the non-expert. We'll go beyond the basics for sure, but we'll be taking the perspective of someone who may be entering into machine learning for the first time or wants to move beyond white papers and into experimentation. I have two guests with me today. First is Tom Estella. If you're a frequent listener, you'll remember Tom from an earlier podcast. Tom is a thought leader in advanced analytics and machine learning, and a Managing Director working out of our Atlanta office. Also with us today is Ramon Matos an Engagement Manager who recently delivered a machine learning initiative. Ramon is a Director and also works out of our Atlanta office. Tom, Ramon—Welcome.

Vinnie

So why don't we just get started at the very beginning. I don't necessarily want to take a rigid approach of a project life cycle through this, but maybe loosely. So, let's start at the beginning. You're new to machine learning. You want to dig deeper. What's the first thing you do?

Tom

Hey, this is Tom. The first thing we do is we try to come up with a business idea. Sometimes clients come to us and they say, "Hey, we have this concept. We have this idea of what we want to do." And sometimes they just say, "We've got a lot of data, a lot information. We're not sure exactly what we want to do, but we know we want to start getting into machine learning. We want to get into advanced analytics." So, the first thing to do is to really find out what that concept is that we're trying to work on. Now, if they don't have an idea, or we can't build on it, then we'll start with a workshop to start to understand what it is that they're trying to accomplish.

Vinnie

If they don't know what they can do with it, why do they feel a need to do it? Is it external

pressures that maybe they're being left behind, the competition's doing it? Where's that desire coming from if they don't know how to apply it?

Vinnie

Everyone's talking about it. You see it in the news. You see other clients and various colleagues talking about it, and it's just a general interest in what Machine Learning and Advanced Analytics can do and help them to solve things. I think that's part of the equation there.

Tom

Yeah. And when you hear about it, you think, it makes us more efficient, it reduces costs, it increases revenue—It does all these things. So, people are interested and want to understand what's the art of the possible. So how could we leverage this to make a more efficient environment?

Vinnie

So where I get wrapped around the axle on that from a starting point, is you have people in the business who know their industry really well. And they maybe are light on machine learning or haven't made the paradigm switch in their head in terms of how to think in those ways. So, they're struggling coming up with a business case. You can hire a data scientist but that person coming in may be really strong in math and statistics, but they don't know your business. And so how do you take those two extreme positions and find something meaningful to work on that will also be sort of a quick win so you're not spinning your wheels?

Ramon

Tom hit on this one a little bit in the previous part of the conversation around the workshops, just coming up in ideating to understand what are some areas that are ideal in the business environment to be able to address with Advanced Analytics. But I think another part of that is also needing to make sure you have really good subject matter expertise within the business, around the business processes, what they're trying to accomplish, and good understanding with some stakeholders that know and understand the data that's going to be used for the analytics.

Tom

Yeah. It's about asking the question of what do you do today that you want to do better. If you're marketing to people and you want to market to people that are more relevant or more targeted, whatever it is that you're trying to do, let's just talk about those business problems you have, and let's get 10 of them on a piece of paper and we can then start to narrow it down to which ones are best suited for Advanced Analytics.

Vinnie

Right. And if you're in a workshop situation, you have your data engineers in there to tell you whether or not they have the right data, or enough data to do that. And you have the Machine Learning people in there to know what kind of models may be available or how to build them. Is there anyone else who needs to be in that room or developers in that room? Are there any usability people in the room thinking about the front end or the customer experience?

Tom

There's actually a lot of people. In fact, I think you have to have a full team there. So, if you think about marketing, as I just said, and that's just one example, we have to understand if we could actually market to them? Do we have their contact information? Do we know what we would say to them, why we would say that to them, when we would say it to them, what we're trying to do—all of those different pieces. So, you may need UX people. You may need creative people. You may need strategists. You really need a full team of everyone together to kind of come up with the ideation of it.

Ramon

I would add in there the product management team. So, a lot of the things that we're looking at from an analytics standpoint potentially will be leveraged or used by a downstream application. So, having the product management team or representatives from the applications team that are going to be consuming. The output of the analytics being part of those conversations are important as well.

Vinnie

Gotcha. Maybe you guys can hit on this for me. One of the things that I've seen people struggle with is that paradigm shift I mentioned earlier—probabilistic models in place of deterministic SQL queries? I'm thinking about what's changed, how the problem solving changes, and how the workflows change. Do you see that being a problem that you guys have a mental model you have to work through early on in Machine Learning? Or is it really not that big a deal when people make that turn rather quickly?

Ramon

Hit on this one a little bit, Tom, because I think there are some other pieces here, like what type of analytics we want to use, that are going to be part of the deeper conversation around what we want to solve and why, and also just the general data. We've mentioned that a couple times, but I think there's a deeper sense. And I think the example that you give around the pieces like, let's just say that the Coke bottle in the ocean example that you like using, maybe hit on that. I think that that's going to help some of the conversation here.

Tom

Yeah. When we talk about deterministic versus probabilistic— if you can create a deterministic query to understand whether something's happening or not happening, then there's no need for Machine Learning. There's no need for Artificial Intelligence. There's no need for any of this. If I can tell you that... I'm trying to think of an example of that... if we're trying to predict fraud, and we look at something and say, without a doubt if they break into the vault of a bank it's fraud, they've stolen something. Well, if we see that they've broken into a vault, then we know they've stolen something, but it's the things that are not as clear. So when we try to predict whether or not something may have happened, that's when we talk about probabilistic modeling.

And so, Ramon, I think the example you were given is really to talk a little bit differently about what modeling technique we use instead of [asking] do we use modeling, or do we use queries? But when it comes to a query, they're, they're really useful. We would rather create queries and just say, "Hey, is this happening? Is that happening?" That's much simpler. And it gives you a more determined answer. Yes or no, what happened or didn't happen? Modeling comes in in the gray areas. So it's not black, it's not white. We're not sure exactly where it is. Well, where should we spend our time? So we try to get the transactions, the customers that are on the cusp. We're not sure whether it's going to go one way or the other, and those are the ones that we're able to kind of push in the direction that we need them to go.

Vinnie

Yeah. I think about workflows a lot too. Like if you take an example of college admissions and a bunch of applications are coming in and you run them through a Machine Learning model, there's an 80% chance that these people would not be good fits. [But you let them] fail at the first year or whatever else, [because] you can't just reject them. There's a 20% chance you're wrong about all those people. And you'd hate to have to try to explain that that's why somebody wasn't accepted. But you could flip it and say, you know, we can accept, with a much lighter review process, the ones that we predict will do very well and then the others go into a more human process. So I'm curious about those optimistic workflows versus pessimistic workflows and how you ensure that the models are being used in an ethical way.

Tom

That's a great point. So if you think of college admissions, anyone that received over 1550 on their SATs and had it over 'X' GPA can be almost an automatic approval. Or a two question, do you have this? Do you have that? And we prove that it's the other ones in the middle that you can say, "Which ones do we spend the most time on?" So there's a willingness to go deeper because you're right there on the cusp. And based on the statistical model of who was accepted last time, we could say their likelihood to be accepted. And that could give us a better understanding of what the key factors are who are at the top of the priority list.

Vinnie

Yeah. I was talking to a university, I guess probably a year ago pre-COVID. And they were saying that they were looking historically at who ended up doing well and looking at those models. And it turns out that a lot of the traditional things they looked at, like you said, a couple of key elements in many, many cases were wrong. Those students did do well. Yes. But the students that really did well perhaps didn't have the easy markers—they had 10 or 15 other subtle markers that would be very difficult to figure out. So it's kind of encouraging to think that it's not reduced to SATs and GPA and whether or not you were in one or two clubs or, or sporting teams, but that there are 15 to 20 more subtle things that Machine Learning can help us understand. So they were using those findings after people graduated to then build models for admissions. So what are some other key use cases? So our clients are coming in new to this, and it's going to be a lot of people on the call who are thinking through their first problems. Let's just talk through some common use cases since we're in the getting started section.

Ramon

Yes. And actually, I wanted to give you guys another example, on Machine Learning. So, where we are today in the world with the pandemic—one of the things that we worked on with our client was a COVID-19 analytic project around fraud, waste and abuse for trying to understand how COVID was impacting health care and the needs of their patients. So, this was actually an interesting one. Tom, if you recall, we worked on this one very early in the Spring timeframe and as a team we did some analysis on the COVID 19 data. Come to find out they ended up identifying areas where behavioral health was really impacted as a result of the shelter-in-place orders that happened at the beginning of the year. And, as they were looking through that, they further dissected and looked through the data and identified that this was actually an area that was getting a lot of fraudulent claims coming in as a result of telehealth kind of picking up with patients needing to have those visits on a day-to-day basis. They ended up having those more often as a result of the shelter-in-place. So that was another area, kind of like the university example that you gave, where it helped to identify and provide some predictive things to look at as a result of the analytics that were positive by identifying who needed to help—not necessarily that it was fraud, but just at some certain areas where certain patients needed more help than others as a result of the shelter-in-place. And the impacts of that.

Tom

Yeah. In that use case for the healthcare industry, we often predict what normal behavior is for providers. How they interact with their customers, how they interact with their patients. We were able to very quickly see the changes in how they were interacting. So how often people come in,

how long they stay on a phone with the person, all of those different pieces, we were able to predict that, and we were able to see and create a new profile for that interaction. That allowed them to understand what was different from that profile and who was acting abnormally compared to their peers. So we were able to help find fraud, waste, and abuse within that sector.

Vinnie

So I hear a lot about fraud detection, you guys have mentioned that a couple of times, just like rapid fire, give our audience a couple other common use cases.

Tom:

Another one of them is a financial services. Understanding on a general ledger (GL file), if you were doing an audit, which transactions are the highest risk or which ones look the most abnormal. That's a great one to use.

Ramon

Workflow. So, looking at the outputs in analytics to help end users prioritize the work that is the riskiest, or most important, or highest value, and helping them to identify those quickly is another good use case.

Vinnie

And more generally, what I'm picking up from talking to you guys is common behavior versus uncommon behavior. So one of the ones I like to think of from an Internet of Things healthcare perspective is wearables. I know that for the elderly, falling and breaking a hip is a very serious thing. And they're able to now study the movement and behavior of people four or five days prior to that fall. They know how they're walking, how the gate has changed, the speed, whether or not their patterns are different, and they can predict that a fall is coming and then do preventative care. And again, from a workflow perspective, there's no harm there in reaching out to somebody and checking in with them, it doesn't create a negative experience.

Tom:

Yeah. I actually had a previous life where I worked on that exact project. We were predicting

activities of daily living. So is mom getting up, making three meals a day, using the restroom, and then bathing the correct amount of time—all of those things just based on sensors within the home, and then able to say, mom's not acting the way she did last week. So maybe you should just check in as well as having red flags that say, something looks extremely wrong today. It gave people just a little feedback as to how the elderly are doing.

Vinnie:

Gotcha. Great. Let's move more into the implementation of the project. Is this just like a regular software development life cycle? Is it—we got agile, we got coaches, we do stand ups, we do a two week sprint, we're good to go?

Ramon:

It's actually much harder and more complicated than that. Because there is no right answer to begin with. And even project set up and working through understanding the problem, and as you're doing the development work, that iteration back and forth and discussion has to happen between the data scientist, the business, SMEs to understand the outputs of that iterative process. You want to try to fail fast and try to understand what adjustments you need to make, but it's a lot harder from that perspective where you can't time box it in certain phases and really work through from initiation to delivery.

Tom:

Yeah, there's one other thing to remember here—when we talk about the implementation or the actual creation of this thing, it's a solution, right? So it's not going to be just the prototype of a model like statistical model. One aspect of this is connecting to data sources and storing that data, then there's the creation of the model. That's going to predict X, Y, or Z, whatever it is you're doing. Then that has to be the output, put into another system to give to a frontline person that can then make a decision and do something based on that. All of these different components need to touch each other and work together. So it's not just the data scientists there. It's really a team of people that work on a complete solution and architecture, not just this one little piece in the middle.



Ramon:

And that's a great point, Tom. It cannot be a single point solution because of all of the downstream and upstream items that need to connect together to be able to get the data to the analytic, to the consumption end, to be able to leverage the outputs.

Vinnie:

You said, you have to fail fast and, and keep trying to get it right. How do you know you're right? We're making predictions. So let's say you're deciding whether or not to make a bank loan or write an insurance policy. You're not going to know for potentially a year or two that someone's going to default or fail. That was a bad idea. So as you're creating these models and looking at the results. How do you validate some of these long running threads as to whether or not you're in the box?

Ramon:

Hey, Tom, the way that I was thinking about that when I said it earlier was more so about analysis paralysis, right? So I think if you can hit more on, on what that really means from an analytics perspective, I think that we'll go deeper into those details.

Tom:

I'm going to go to Vinny's question. I'm going to write back to you for mine, but Vinnie to your point, when we build models, we look at past behaviors and past data. We build models and say, if we scored last year and implemented this, how many defaults would we have in the mortgage industry? From that we can say, "Yes, we are staying at the 1.5% default rate that we expect," or whatever that number is. So we actually take past results we hold them out from our modeling sample, so that we're now building the model on that data. And then we score it and say, how did we perform? That gives us an understanding or an idea as to where those cutoffs are, those score cutoffs to say, if we were to accept mortgages at this rate, here's where we would have this default rate.

Vinnie

Well, that makes the assumption that people have that historical data available. Right? So I know as a prerequisite to doing Machine Learning, it's getting your big data right. And getting valid data, enough data. So I guess the takeaway is to make sure your data engineering is in place as early as possible, because otherwise you're not going to be able to do that sort of historical trending to validate your models.

Tom:

And now for Ramon's question, and that's what I like to say knowing good enough is good enough. So when we're building models, it's almost more of an art than it is a science. There are scientific factors and, and guardrails on the process, but it truly is an art of how we create that statistical model. Just like any other piece of art, you could always make tweaks and changes. And at some point you have to say, this is good enough, let's move to the next step in the process. So understanding when to finish your masterpiece is a really difficult part and just takes years of experience doing it.

Vinnie:

Yeah. I wanted to talk about bias too, and I didn't know if we should introduce bias in this section or the previous section. And what I mean by bias is kind of two parts. One is, knowing that some of the data that's collected may be biased just because the nature of what's being collected, and then the second part is when you do show the results of the models bias in the analysis of that data, right? So do you talk about bias and think about that in the first section, when you're trying to find what problem to solve, or are you free of that and just brainstorming? And then as you get into the implementation, you start worrying about those constraints. Where does that come in?

Tom:

That's the data scientist. So that's really after you've gotten through the ideas and the concepts, you've gotten the data set there, there are ways that you can adjust your data set to make sure that you're not being biased. There it's actually one, Vinnie we've talked about this before, it's one of the reasons that these point and click solutions for advanced analytics I think are dangerous. Yes, you can create the model, but if you don't really understand the data and the mathematical process behind it, there's a good chance that you will have some bias in there, or

you will have something incorrect about your data infrastructure. So it really is important to have a data scientist, someone who's done this before to understand what that data is that goes into those models. Now, those tools, they make my life easier as well. Because once that data is prepped properly, I click a button and the model's created right. You gave that example yesterday.

Vinnie:

Yeah, I always think about chainsaws when I think about automated tools. So just because you can put an easy start button and a nice heavy strap on a chainsaw, you shouldn't hand them out to a bunch of five-year-olds and send them out into the woods. Someone's going to get hurt. So I joke about that with other technologies as well, but I think it's relevant here, because especially when you think about the biases and the predictions you're making, those have real impacts on people's lives. And so to that point, going further, we have bias. We also have the fact that models deteriorate over time. So they may be very accurate today, but then the nature of the data you're collecting, and the fact that you're acting on that data can change the data that you're collecting. Cause you're treating your customers differently, perhaps. So how do you A) ensure weeks and months into this, that a bias is not being introduced, and B) that your models haven't deteriorated to the point where they're no longer accurate in who does it, and how does that work? What's the process.

Tom:

There should always be a measurement plan at the backend of your model so that you understand how it's performing, and you can watch that over time. So you'll see that degradation if the models not performing at the same level as you go forward, if your model is more of a machine learning model where it's actually recreating its coefficients on a regular basis, it should be keeping up with the trends within the data. But if it's not that type of model, if it's more of a deep learning model, then you, you will have to kind of recreate that, recreate that process on a more regular basis. What I fear most more than any of that is the data sources that you're getting your data from change. So before we were building models on this data, but now that data is not in that system anymore, it's coming in a different way. And maybe it's being captured in a different light. And that could change the actual predictiveness and change the accuracy of your model. So you do have to continue to monitor that model over time and see where those blips are happening. And when there is a significant change in the results, go in there and understand why it is.

Vinnie

Right. So, what are some common pitfalls then I've tried to throw a couple out there around like bias and model deterioration. But from your guys' point of view, if you were to give the top three reasons these initiatives fail, even if it's restating something we already said, what would you say are the most common?

Ramon:

Tom, I'll start us off here. I think, and I know we've said this one a couple of times, but it is so important to understand the problem that's trying to be solved or to the decisions that you're trying to help make. With the analytics, there is also making sure that you have strong stakeholders that are bought into the work and strong executive sponsorship, driving the initiatives so that when the decision-making needs occur, they're being made by the right stakeholders that are going to be impacted or going to be driving those decisions based on those analytics.

Vinnie:

Would you add anything to that, Tom?

Tom:

You know, I have a great example of that. One of what Ramon just said. So we created a process that was finding fraud, waste and abuse within financial services. So for audit purposes. When we built this tool, it actually saved the client like 75% of their time and effort of all of the data manipulation, all of the statistics, statistics that they were doing in Excel at the time, all of this was automated done online, quick, cheap, easy. This was a reseller of this and the, the executive sponsor at the end said, now that we have this, we don't have to sell it. So forget it. We're not going to use it, even though it was a better solution, it was created, it worked better, it was

faster. It gave better results. The executive sponsorship, wasn't behind the fact that we'd have to change the entire business model. And so it sat on the shelf. So you really do have to understand the entire situation, how this is going to work completely and have your executives behind you?

Ramon:

Take that point just a little bit further. There also, Tom, are a lot of times the analytics could potentially change a business process. And if you don't have the business stakeholders bought into what you're trying to solve for and understanding that there could be downstream changes to the business processes, you might build something that's not going to get used. It's not going to get adopted because people are not going to be bought into why they are doing this.

Tom:

I have another pitfall, and this happens a lot. So you really have to focus. I remember actually being a pretty junior person. This is probably 20 years ago, and I was asked to build a statistical model for the company I was working for. I built it, but they expected me to do all of the data engineering at the front end, which I did. I wasn't a data engineer. I was a data scientist, but I did that work. It wasn't the most efficient way of doing it. And then do all the system, the data visualization on the backend so that it could, it could be presented to the, to the end user and then do the system integration. So that, that it worked in the system. There was probably five or six different roles that I had to play on that, on that team. And I did them all. Okay. I did the data, scientists were great, but the other four roles, I probably did a good job on them, not a great job. And the solution was slow. It didn't have the UX aspect of it, of the user experience on the backend. All of those different pieces were done on an average rate, not on a great rate. Really you need to look at those different pieces and have the right person build each of them. I think that's an important part to think about.

Vinnie:

Yeah, we spoke about this. I probably in the last podcast that we did together, where, just because a data scientist has the word data in the title, it doesn't mean they're a data engineer. And so these, these are very different roles and you probably need what one data scientist for every six or seven data engineers you have, because it's just gonna take that much more work to get the data clean, the right data, the right volume, the right velocity, all your five V's of data. So, yeah, it's not a one-to-one and it's a different skillset, right?

Ramon:

Yeah, that's right. The other point there, Vinny is the prototype doesn't necessarily mean it's a production ready analytics, right there. There's engineering best practices that have to be accounted for in hardening of the code to make it production ready. And that takes time to go along with the overall effort.

Vinnie:

Yeah. One more, question on that, because I remember, gosh, four years or so ago when it was a little more immature from a tooling perspective, you would see, Machine Learning happening on someone's laptop or on a server under the desk kind of thing, as opposed to the mature application development that the, that the systems integration teams will be working on. I know that for a lot of that, that that's fine because you're not creating a production element that needs to be used in production. You're using a model that can then be exported and used in production. Right. So is there a more form a lifecycle, I mean, maybe not a full SDLC, but something that's more formal now in terms of creating these in, in more robust environments.

Ramon:

And I would say, so that's actually one of the things that we've worked on our current client, is about standing up in analytics, and development process to help with the ideation help with the intake of the analytic help with the prototyping and then working our way through review and analysis with the business stakeholders and then working towards moving into production as production ready code. And so, so by helping put that structure in place and helping them understand the different phases of the life cycle, that it takes to get the models into production or the analytics into production, it's, it's helped tremendously with business stakeholders to understand, you know, I'm putting an ask now I understand what I'm getting and why, and I'm able to participate and be a voice of within the process to understand the analytics, and to validate the analytics is going to help me.

Vinnie:

Great. Thanks. So closing thoughts, I kind of want to get a little bit closer to the metal on this one and say, okay, if you guys were advising me, I've got a technical background, used to be developer, done architecture work, and I want to get into this. Where do you point me? Do I go to AWS? Do I use Azure? Are there, is there a machine learning in a box, you know, data science in a box or as a service? What's a good place for me to go maybe a good website, or what, how do I get started?

Tom:

My personal opinion is Azure and AWS are almost interchangeable. They both have tools that have the ability to do all to do everything that we need to do here. So whichever you're more comfortable in you've worked in, in the past is probably the right one for you. In terms of how to learn, how to, how to do this again, you're going to start with the business problem. So understood understanding what it is you're trying to accomplish and how you flow through the solution is going to be the most important part. So work with someone to understand the solution architecture, work with them to understand the different components, and then start to focus on each of those components of the solution and learning how to do each of those so that you, you grasp the concept of the solution.

Vinnie:

It sounds to me like we've said this a couple of times without coming out and saying it takes a team. So maybe the best way to get started, if you're looking to get started, is to find a couple of colleagues that are in those other disciplines and forming like a small study group/working team or something that can do a proof of concept. Get someone who knows the data, get someone who knows the data science, get someone who knows some of the cloud tooling and such, and just treat it that way I guess I would also suggest this is probably a common pitfall as well, not biting off more than you can chew. Doing something that has a small turnaround time and still has an obvious impact to the business so that you can get some excitement. And of course, that's just the best practice with technology. Anyway, get small frequent wins, get people excited. You get that momentum behind you, as opposed to taking months and months.

Ramon:

And set your goals. We're going to do X. We're going to be able to predict why we're going to, it should be better than Z. Have those goals and either hit them or don't hit them. But if you just build it and say, let's see what happens. You never know if you're successful.

Vinnie:

Or not. Yeah. State of vision.

Ramon:

Good. And just from an expectation setting standpoint, it's not a magic bullet, right? This is not an easy button to get things done. It takes time. It takes effort. It takes definitely subject matter expertise across all the disciplines that are going to be doing the work and have realistic expectations on the outcomes and how long it's going to take to achieve those goals.

Vinnie:

Great. Well that wraps up this podcast. Ramon, Tom, thanks so much for joining us. Great input. Really appreciate it for those who are listening. If you haven't had a chance to yet, please subscribe. So you can be notified when new podcasts come out. Thank you, guys. Thank you.

Vinnie

The entire contents in designing this podcast are the property of CapTech or used by CapTech with permission and are protected under U.S. and International copyright and trademark laws. Users of this podcast may save and use information contained in it only for personal or other non-commercial educational purposes. No other uses of this podcast may be made without CapTech's prior written permission. CapTech makes no warranty, guarantee, or representation as to the accuracy or sufficiency of the information featured in this podcast. The information opinions and recommendations presented in it are for general information only. And any reliance on the information provided in it is done at your own risk. CapTech. makes no warranty that this podcast or the server that makes it available is free of viruses, worms, or other elements or codes that manifest contaminating or destructive properties. CapTech expressly disclaims any and all liability



or responsibility for any direct, indirect, incidental, or any other damages arising out of any use of, or reference to, reliance on, or inability to use this podcast or the information presented in it.