**Interview with Sandeep Pawar**

Data Analyst at Cree Lighting (Analyzing data using Machine Learning and Power BI)

Interview Conducted on Jan 12, 2021

Q: In relation to what you have been doing in the past years, what has your observation on the learnings that can be derived at each stage of a project: Problem Formulation, Data Collection and Cleaning, Analysis and Presentation?

My observation is that the most time consuming and important aspect is the ETL and data collection. Suppose you are working with a manager, they want to solve a problem.

When you work on a project on say Kaggle, all the data is presented very neatly for you, columns and so on. If you are working on marketing, you have customers and what they have purchased and all that information is with you.

In real life, you have to access tens of database and table, aggregate that data, you have hundreds of features that you have to boil down to 10-15 that are important. It is a very iterative process and time consuming process. There is no way out of it. We had to go through it and keep in mind that it has to be repeatable when you are creating your model. When you create a model with these hundred features that you have boiled down to a few features, be aware that the data exploration you are doing comes from those hundred features and it has to be done in a way that is very adaptable.

Q. In your department, who comes up with the questions that makes you chase certain data.

It is the usually department that I am working with. For instance if you are working with marketing data, it doesn’t always work that way. For example in my company Data Analysis and BI is performed a lot more than ML. And ML is based on certain business challenges. Say I am working on an analytics project with Bi and someone comes up and says, “This is good” We can look at the past data and forecast. Tell us what has happened in the past. But then the manager also be interested in “how do I prepare, How do I allocate my resources. What should I expect. Those are the questions. What can I expect. And those are the questions I also ask when I start a project. What exactly are they trying to do. What kind of decisions are they hoping to make from the dashboard. Most of the time they don’t know and we have to ask them.

And so the trick that I have learnt is asking them what decisions they would like to make? They might then say, after looking at data for a month or so, should I hire five more people. That gives you an indication that what they are looking for process bottlenecks, and your goal is to help them make those decisions.

So they might be looking at a non-concrete thing and you have to tease out the requirement out of that. And most time it is not easy in the sense and you can ask questions but you have to work with the domain since initially you do not know how they work or what decisions they make.

So it start with the process, what kind of data they have, where does the data come from? What decisions have they made in the past by looking at the data. That gives you an insight into all the available things and then you try to find the missing pieces.   
They might not be aware of all the data that is really available. They might only be ware of certain excel sheets that is circulated or shared with them. But there might be some other places where you can find data.

Q. What are the questions you typically ask to elicit the information you are looking for?

I usually ask first is what data exactly do they have are the familiar with and then What I usually ask is what are top 3 or 4 or 5 things are they trying to get. It is also critical to understand or keep the overall strategic goal in mind. If the department as a whole has the goal of say reducing inventory – or some such goal. SO say I am working in something and I am also aware that the department wants to reduce inventory or improve something, then while they are working on their job. You have to ask them how your work impacts what they are trying to do.

And that might help them think how the system they are hoping to change might affect their overall goal. So this is an iterative process. So 80% of the time your goal is to guide them through the process. Because if you create something that is not insightful or actionable, then no matter what machine learning model to BI dashboard you create is not going to be useful at all.

Q. You mentioned earlier that your company uses analytics and BI more than ML. What are these analytics?

It mainly involves plotting things. So if you create a dashboard, there should not be more than 3-5 things on that. KPIS for example. When a manager opens it, your goal is to ensure there are 3-5 things that capture the same thing right away. For example, are there any anomalies, are there any trends we should be worried about? Are there top 3 items I need to be aware of. Because he has very little time to understand all of this as he is working on a lot of stuff.

So the analysis could be as simple as showing a KPI or as complicated as maybe doing clustering or it might be doing anomaly detection or it could be TPA or more complicated methods like regression. So it depends and varies from project to project and what we are trying to do. But the goal is first if you can, keep it simple first in the dashboard and then see where it goes and if they are not satisfied, go with more complicated analysis.

NOTE: This is the [Occam’s razor](https://simple.wikipedia.org/wiki/Occam%27s_razor) test. If two models have comparable performance, then you should usually pick the simpler one.

Q. Since the initial stages such as formulation and data collection and cleaning as well as analysis take quite long, what is the expectation on presentation cadence?

There are varied views on that and some say you should work with the manager or stakeholder and present early so you can get feedback. I am on the other end of the spectrum.

When someone comes to me, I make them write it out and maybe even draw a picture of the dashboard they had in mind. So that helps me but more importantly it helps them. Or instance, if they say, I need a moving average, then to have Sanjeev create a moving average, he needs this and this information. 99% of tie what they have in mind is not realistic or the real dashboard they need is completely different from what they have in mind. And so the picture helps because I can get back with, instead of doing this, can I do it this way because if you do it this way, it will be more insightful. In a nutshell, the process is very transactional. It is never that ta manager gives me something and I give something back to them. But you have to absolutely make sure it captures the minimum requirements they have in mind. If they are not excited the first time, then they will not come back. It has happened in the past that I build something and they are excited with the visual and things but if it is not useful the they are not excited enough to use it again.

Q. With respect to reproducibility, how much of the data collection and manipulation is manual and how much is automated.

With respect to enterprise data, when it comes to the Oracle DB and SAP then we do not have to do much for the extraction and cleaning of the data. But if it is in excel sheets, then instead of doing it in Power BI, then I ask them to do it in the source level. Because it is in their own interest as well. If they already have an excel sheet they are working on, then I show them how they can improve the data quality by showing them how to use dropdowns and lists. If that cannot be doe then I handle it in Power Query. But unfortunately there is always some data cleaning.

Q. What led you to Data Science from a Mechanical engineering background and what led you to Power BI and ML

My background in research in mechanical engineering and in building numerical models. I was always involved in statistics and it was always part of what I was doing. I was in conversation with one of my colleagues who mentioned they were having issues with a client. I looked into it and performed statistical analysis on it and told them what the supplier is giving them is incorrect. The colleague accepted the results but told him that there was not way for the company to make it reproducible.

And that is where Power BI came in and then we talked about value and that is how I got pulled into more projects and that is what led to data analysis. I also did Masters in Business Analytics and when I was doing that I picked up projects to use those skills.

In Mechanical engineering, we have to do test that involve statistical analysis. But I am not a statistical expert. I know where to look and know what to find keep myself updated.

Q. Do the Power BI visuals do all the talking or is there anyway he uses to communicate the results.

Presentation – I keep it as simple as possible with Line charts and bar charts and comparative analysis. n and it is an iterative process. I ask them would they use the dashboard and ask them to talk about it as they are using it. This lets me know what their thinking process is and if they are using it correctly and if something does not make sense as they are looking at it.

I had further discussion regarding Power BI specifically.

Once you create your model in Power BI and feed it to Azure ML and get it back into Power BI you have all you need. It is not cheap but with

You do not need Premium licensing you can have Dataflows.

it.