**The Art of Solving Any Data Science Problem — Simple Tips For A Better Outcome**

Unspoken Advice To Avoid Trouble and Increase Your Chances of Success

Most people interested in data science learn about tools and technology to solve data science problems. They are absolutely necessary to build a solution. But, remember, it is just not enough. To come up with an efficient solution, one needs to learn the art of problem-solving. There are many courses to teach about the tools and technology in data science. There aren’t many courses on how to solve a data science problem.

In this article, my aim is to use real-world use cases to help you understand the key aspects of solving a data science problem. We will also see how these would assist in identifying and solving the core business problem. Also, how they assist in avoiding common pitfalls that lead to issues.

To make sure the concepts discussed here are easier to understand, I will use customer churn in telecommunications as an example.

**Problem Conceptualization**

The first and foremost objective should be to clearly define the problem. Problems such as customer churn can be very tricky. Generally, when a customer stops using a product or a service they are considered churned. In the case of telecommunications, we could simply settle for customers who have moved to a competitor or discontinued as customers churned. There could be scenarios where it might not be straightforward to define customer churn, like, for example, in e-commerce. Different e-commerce customers could use the platform at different frequencies. So, in this case, extensive data analysis is required to accurately define customer churn. Another example is, dating sites, how to differentiate people who found a date on the platform from someone who quit after a bad experience.

Now, coming back to the example of customer churn in telecommunications. Time should be spent to better understand customer behaviour. Having a good understanding of customers will help in better problem conceptualization and, hence, better solution development. Below are some questions that could help in better understanding customer churn.

* What is the average tenure of a customer?
* What is the average revenue per customer?
* What is the average lifetime value of a customer?
* What is the cost associated with acquiring a new customer?
* How many new customers are acquired every month?
* What are the top 10 complaints reported to customer support?
* What events are likely to result in customer churn?
* Where does a churned customer go/is likely to go?
* What are the factors on which the competitions are rated better?
* What are the reasons for customer churn as per the internal teams? Usually, the reasons wouldn't be aligned.

These questions would help to clearly conceptualize the problem and focus on a solution that helps achieve the goal. Later in this article, I will talk about mental models, that can be used to better understand the problem.

* Should we focus on events leading to churn and hence prevent those from happening?
* Should we predict the customer's likely to churn and prevent them from churning?
* Should we just focus on acquiring more new customers?

While being data-driven in problem conceptualization is good. It is more important to be purpose-driven. Being data-driven we make use of the data in key decision making but by being purpose-driven, the intended goals and the purpose is also considered. Being purpose-driven ensures that the solutions are aligned with the overall objective of the organization.

**Good Understanding Of The Data Landscape and Being Flexible**

In real-life scenarios, data science teams are not provided with the data they need to solve a problem. One of the primary duties of a data scientist is to collect all data required to solve a problem. At most organizations, data is spread across multiple sources that come from different platforms.

While solving a data science problem, a data scientist would generally come up with several hypotheses. In the case of customer churn, some of the relevant hypotheses are,

* Are customers leaving due to issues not being handled well by the customer support team?
* Are customers leaving due to better deals by the competition?
* Are customers leaving due to poor service?
* Are customers leaving due to technology issues?
* Are young professionals likely to churn after their first unsolved complaint?
* Are long-time customers paying higher prices likely to churn?

Based on a few of the above hypotheses, some of the relevant datasets that need to be considered for the analysis are,

* Customer profile data
* Customer usage and billing data
* Complaints data and other customer interaction data
* Data on technology issues
* Data about the competition and their deals

A better understanding of the data landscape would help in,

* Easily identifying all relevant datasets to validate the hypothesis
* To integrate datasets across different sources and also to understand the limitations
* To be aware of the data quality issues and hence be better prepared to handle them
* To ensure the personally identifiable information is handled in a secure and ethical manner as per compliance
* Having a good understanding of the dataset helps a lot in data exploration.

**Avoiding Bias in Analysis**

It is absolutely normal for us to have an opinion on a problem we are solving. But care should be taken to ensure that it doesn’t impact the solution we are building or the analysis we are performing. Bias could have a very large impact not just on the project but also on the organization and its reputation. Below are key reasons why bias should be well handled,

* It results in an inaccurate result that leads to bad decisions and hence poor outcomes for business
* It could result in discriminatory outcomes that could put a segment of people at a disadvantage
* It could result in a lack of trust in the organization

To better explain this with an example, let us consider the field of medical science. If the models used for diagnosis and treatment decision is biased toward a group or a race because of insufficient data while training the model. This could lead to inaccurate diagnoses and incorrect treatment decisions. This is a classic example of selection bias.

Similarly, let's say a government agency or an organization makes use of historic crime data to identify people who are likely to reoffend. The bias in the historical data could result in discrimination and injustice to individuals from certain groups. Hence, while solving data science problems, it is always important to consider bias.

Let’s return to our customer churn example. If the data on which we have trained our model doesn’t represent a certain group of people then it could result in a skewed outcome. Here are some tips to avoid bias in our analysis,

* Using a sample that accurately represents the populations for the data analysis
* Ensure all the relevant data are considered for the data analysis
* Test model on independent datasets to measure real performance
* Having a diverse team and engaging them in key project-related decisions
* Always look for alternative explanations while interpreting outcomes.
* Explore the patterns in the dataset that might show signs of potential bias in the data
* Clearly document the process, results, and interpretation so that they could be reviewed by third parties for bias

**Brainstorming**

Brainstorming not only helps in avoiding bias, but it is also very helpful in the success of a data science project. Brainstorming helps in coming up with a wide range of ideas, which is critical to coming up with an innovative solution. Below are some of the benefits of brainstorming,

* It helps to come up with an ingenious approach to a problem
* It helps promote better collaboration among the team members
* It brings in different points of view to come up with a well-rounded solution
* It is one of the key reasons behind many innovative solutions

Brainstorming sessions are needed for more than just internal team discussions. But for better outcomes, these should be cross-functional. When working on the customer churn problem, all the parties that are directly or indirectly impacted should be included in relevant sessions. For example, problem conceptualization should include brainstorming with the following teams to start with.

* **Customer Support Team**: They do know some of the key issues faced by the customers
* **Technology Team**: They do know about newly released features and the issues across different platforms.
* **Marketing Team**: They have a better idea of the campaigns and offers at the competitions.

**Use Mental Models To Facilitate Structured Thinking**

Mental models are a set of tools that provide guidance on understanding and making sense of complex information. In data science, mental models are very helpful in better understanding the problem and simplifying the problem-solving process.

Let’s see how the mental models help in solving our customer churn problem. Below is an article that explains how the first principle of thinking is used to solve customer churn.

**[How to use First Principle Thinking to solve Data Science Problems?](https://towardsdatascience.com/how-to-use-first-principle-thinking-to-solve-data-science-problems-db94bc5af21" \t "_blank)**

[towardsdatascience.com](https://towardsdatascience.com/how-to-use-first-principle-thinking-to-solve-data-science-problems-db94bc5af21" \t "_blank)

Below is how the mental models are useful across below stages of a data science project.

* **Problem Definition**: The very first step in solving a data science problem is understanding the problem. A framework like First-Principle Thinking and Feynman’s Technique helps in better understanding the problem we are trying to solve.
* **Data Exploration**: Exploratory data analysis is all about asking relevant questions and making sure all the possibilities are considered while solving a problem. Some of the mental models that could be adopted here are First-Principle Thinking, Second-Ordering Thinking, Bottom-up & Top-down Approaches, and Probabilistic Thinking.
* **Feature Engineering**: Mental models help in better understanding the domain and identifying important features. Some of the mental models found useful here are Inverse Thinking, Multiple Causation, Root Cause & Proximate Use.

Mental models are frameworks that help with clear thinking. I strongly believe that mental models can be used to solve any type of data science problem. I always try to make use of a thinking framework to approach a problem. If you also find this interesting and want to explore more, here is an article about using mental models to enhance your career in data science.

**[5 Mental Models to Help Boost Your Data Science Career](https://towardsdatascience.com/5-mental-models-to-help-boost-your-data-science-career-2a40fd9b7d8f" \t "_blank)**

[Simple frameworks to stimulate super thinking](https://towardsdatascience.com/5-mental-models-to-help-boost-your-data-science-career-2a40fd9b7d8f" \t "_blank)

[towardsdatascience.com](https://towardsdatascience.com/5-mental-models-to-help-boost-your-data-science-career-2a40fd9b7d8f" \t "_blank)

**Things to Remember While Solving Any Data Science Problem**

Below are important factors to keep in mind while you work on a problem. These could be relevant to anything that you work on.

**No Job is Small**

All jobs are of great importance and should never be taken lightly. Many data science projects involve a lot of data analysis but no model building. People who are new to data science jobs generally have a notion that problems that do not require model building are not so important. It is very important to understand that business doesn’t care about the models being used or the sophisticated algorithms implemented. The single most important factor is whether it solves the problem. Never consider a job lightly just because it doesn’t involve algorithms!

This is very important for a successful career in data science. People will always be remembered by their peers for their effort. A strong work ethic and willingness to put in the effort are highly valued by employers, and they always lead to better career opportunities.

**Never Settle**

The other important thing to keep in mind is to have a “Never Settle” attitude. One should never settle for a simple solution. You should always keep looking for opportunities to enhance your solution or your approach. Though this is relevant to any career, I find it more relevant in a data science career. There will always be pressure to deliver a solution. But one should never use them as a reason to settle for a simpler solution. There will be cases where some options would require more effort, or sometimes they might demand that you learn a new skill. Keeping an open mind and having a never settle attitude help a lot with long-term career success.