ART OF SOLVING BUSINESS PROBLEMS

THE DATA SCIENCE WAY

SHWETA DOSHI
DILEEP K.V.S

GREYATOM

We are sharing our obsession to answer business questions using Data Science for all stakeholders

FOREWORD

Data Science touches every aspect of our lives on a daily basis. Whether we visit the doctor, drive our cars, get on an airplane, or shop for services, Data Science is changing the way we interact with and explore our world.

Data Science is the art of turning data into actions. Performing Data Science requires the extraction of timely, actionable information from diverse data sources to drive decision-making and answer questions like:

- Which of my products should I advertise more heavily to increase profit?
- How can I improve my compliance program, while reducing costs?

There is much excitement about the field of Data Science, as there should be. It is an exciting, burgeoning field and has much to contribute to the industry as a whole. We all know how much data is generated on a second-by-second basis, and that data contains valuable insights which can be used to make important business decisions.

However, Data Science is a technical field, and new practitioners tend to get caught up in the technicalities of machine learning techniques, model-building and optimisation, and often lose track of the larger picture of what their work means to their company.

According to studies, **87% of Data Science projects fail**, and hence we wanted to share a systematic framework to take your Data Science project to success and not be a part of that statistic.

Data Science is a way of applying our curiosity and technical tradecraft to solve humanity's toughest challenges. Our own biases and assumptions can have profound outcomes in arenas as diverse as business and national security, to our daily lives. Hence a new class of practitioners and leaders are needed to navigate this new landscape.

This ebook has been built with the express purpose of explaining to Data Science aspirants what they will realistically have to do when they join the industry as Data Scientists. And how to harness their Data Science skills to bring about real value for their companies.

The story behind Art of Solving Business Problems - The Data Science Way

We created **Art of Solving Business Problems - The Data Science Way** because we wanted to help leaders and Data Science aspirants. There is unlimited content on Data Science techniques, algorithms describing what Data Science is and why we should care, but very little information on how to solve business problems with Data Science.

At GreyAtom, we have built our entire Data Science curriculum on this premise. The focus is always primarily on solving a problem first; and then learning the techniques needed to accomplish this goal along the way. Thousands of learners have gone through this journey with us and our goal in creating **Art of Solving Business Problems - The Data Science Way** was to capture what we have learned and to share it.

To bring the importance of business problem-solving into clearer focus, we have used a fictionalised case study to illustrate key concepts. The scenario of a food aggregator app is all too familiar to most of us, and encompasses both the B2C and B2B business models for maximum understanding.

We take you through the process of examining a business problem and converting it into a Data Science problem step-by-step, in order to highlight the key skills required at each stage.

You will gain an understanding of who you will be expected to deal with, and more importantly, how you should approach each interaction, so you can use that knowledge in your subsequent work.

Image credit: https://www.webalys.com/

TABLE OF CONTENTS

THE STORYTELLERS

Shweta Doshi

Data science and AI literacy isn't optional anymore. Whether you are a business owner, a leader, product owner, project manager, or software developer, knowing data, understanding it, and getting insights from therein, can fuel your entire business.

Data Science is about asking a business question and working your way backwards. Starting with technology first is the wrong approach.

KVS Dileep

Behind the buzz and jargon of data science is a valuable skill that is high in demand but hidden in the hype. Cut through the clutter, and the clarity can provide real value for business.

INTRODUCTION

One problem I face constantly when speaking to or counselling a lot of Data Scientist aspirants is their focus on machine learning techniques, rather than the business problems that Data science can solve.

After years of working in the industry, working on diverse use cases and solving problems , I realise that this is a common misconception, especially among those who are joining the industry for the first time.

Let me assure you that the CXO doesn't care about your accuracy score metric for the incredibly innovative algorithm you have dreamed up, or the cutting-edge library you have used. Don't get me wrong; these are admirable things to have! But the CXO doesn't care HOW you arrive at a solution; he or she cares only about the solution to the business problem at hand.

That's a very different kind of problem to be solved for the company! And you are going to do it with Data Science.

After seeing how new Data Scientists struggle to understand how to use their exceptional technical skills in a new environment, Dileep and I put together this ebook to help people think about business problem-solving with Data Science in a structured way!

I hope you find some real insights in this ebook.

Shweta Doshi
Co-Founder, GreyAtom

0

DATA SCIENCE: THE BASICS

DATA SCIENCE AND AI

It is not an exaggeration to say that the world we live in today is being redefined by Data Science. Data Science and AI touches every aspect of our lives. Whether we are shopping online, taking an Uber or driving our cars, working with our smartphone assistants, connecting on social networks — Data Science is transforming the way we interact with the world and the way world interacts with us.

We are generating more data now than ever before. With every video we upload, photo we tag, sensor data we generate, medical reports that we process, we make data a catalysing force. Data is the new currency.

Because of these profound implications, it is essential for leaders, educators, business owners, policymakers — indeed, everyone — to understand what AI and Data Science is, how it works, what it does, and what it means for them, their careers, families, friends and communities.

Everyone must be armed with AI literacy, job-relevant AI skills, and awareness of how to maximise its benefits and deal with its risks and pitfalls. A new class of practitioners and leaders are needed to navigate this new future.

THE BASICS

Data Science is the science of turning data into decisions using statistical methods and machine learning. Taking a line from **Paul Graham's article** on Donald Knuth, the father of computer programming:

> *Science is knowledge which we can understand so well that we can teach it to a computer. Everything else is art.*
>
> ***- Donald Knuth -***

Performing Data Science requires us to extract data from different sources, and use various techniques to answer questions like

- Which customer will subscribe to my marketing campaigns?
- Predict which customer will uninstall my app next

The scientific aspects of Data Science are well understood but when it comes to solving business problems, there is some art involved and that is what we wish to showcase.

A simpler definition of Data Science would be: "a discipline of making the data useful."

Usefulness is the keyword here. If you are not able to make useful decisions from data which impact your business metrics, you aren't doing 'true' Data Science.

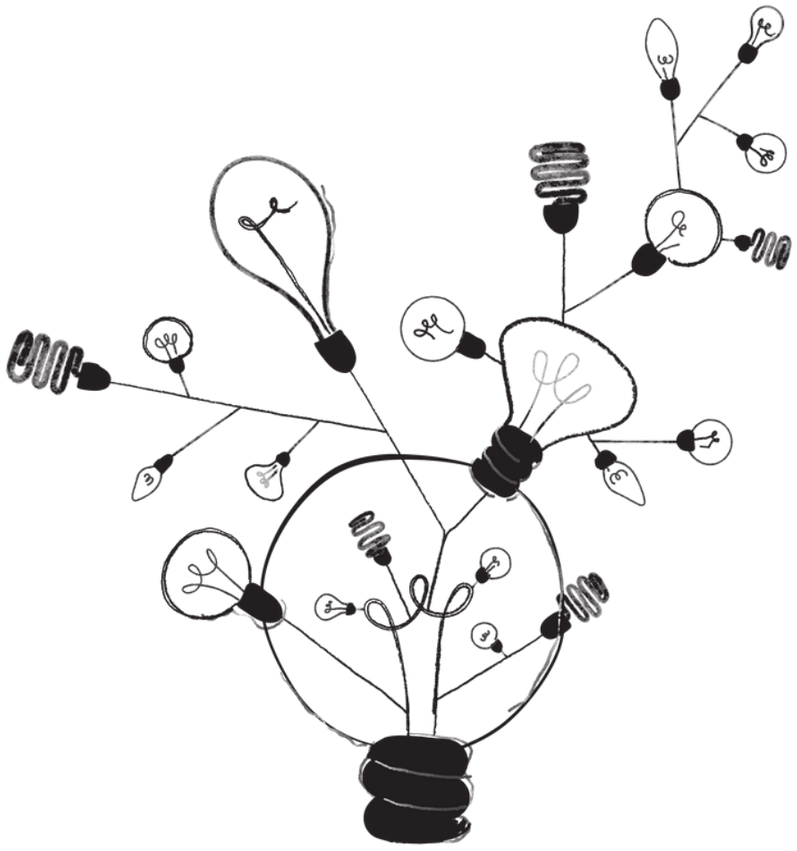


Image credit: https://absurd.design/

DECISIONS, DECISIONS, SO MANY DECISIONS

We love this image that describes what method to use for decision-making, which techniques to use, and when to use them.

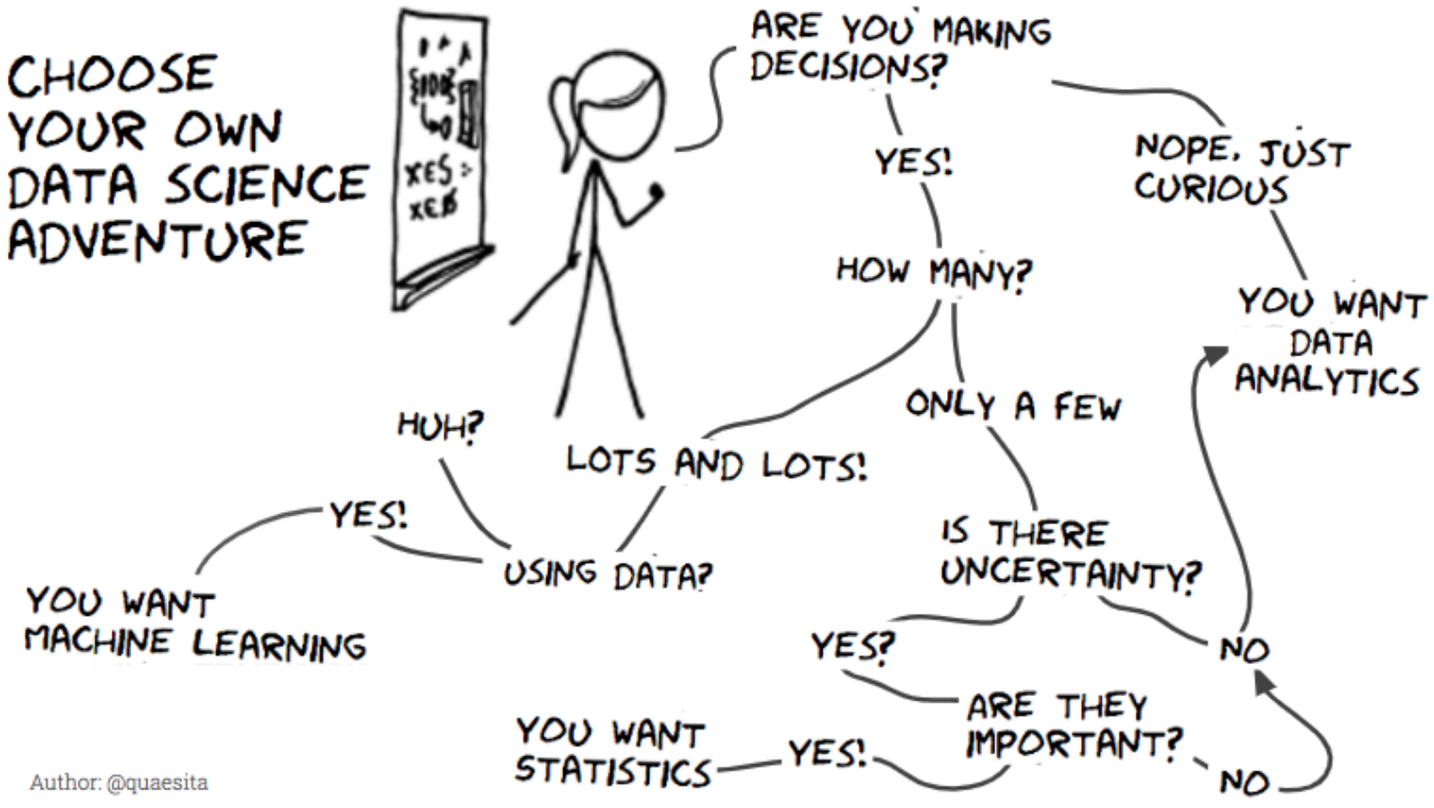


Image credit: Cassie Kozyrkov

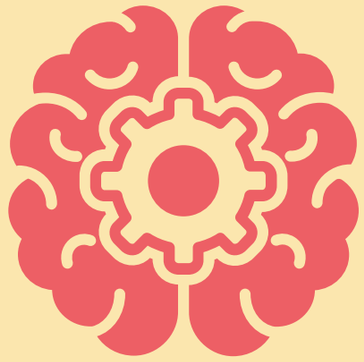
BREAKING IT DOWN A LITTLE MORE

How do we positively impact businesses through Data Science? This is accomplished through the creation of data products, which provide actionable information without exposing decision-makers to the underlying data or analytics (e.g. buy/sell strategies for financial instruments, a set of actions to improve product yield, or steps to improve product marketing).

Here's a map of Data Science for you, perfectly faithful to the Wikipedia definition:

Data Mining
a.k.a. Analytics

Statistical
Inference



Machine
Learning

Analytics

If you don't know what decisions you want to make yet, the best you can do is look at all the data you have collected and be inspired to make decisions based on what you see. This is plain data analytics, or exploratory analysis.

For example, trying to establish a pattern between car brand owned and neighbourhoods, and then deciding what kind of marketing campaigns to run where.

Exploratory Data Analysis is a very important step and a lot of value could be delivered to the business through a detailed analysis.

Let's take the example above and expand it a bit, to see what kind of questions we can answer:

- *Which car brands are sold most in each of the neighbourhoods?*
- *What is the age bracket of the car owners in each of the neighbourhoods?*
- *What is the average price of the car in each neighbourhood?*

These answers would give the company a better idea of the customer persona and purchasing behaviour in those neighbourhoods.

At this stage, we are conducting a simple 'exploration' of the data, which is typically done through data visualisation. At this stage, no complicated inferences are done, and yet a lot of value is generated for business.

Statistical inference

We love the way Cassie Kozyrkov describes this:

> *Statistics is the science of changing your mind under uncertainty.*
>
> ***- Cassie Kozyrkov -***

If you intend to make high-quality, risk-controlled, important decisions that rely on conclusions about the world, beyond the data available to you, you're going to have to bring statistical skills into your team.

For example, inferring car holding patterns in similar neighbourhoods where you don't have full data. Collecting data about all the cars in all the neighbourhoods along with the owners is practically impossible. All you can do is collect the data of few houses in each neighbourhood and then make inferences about interesting business questions.

An example of a business question that can be solved through statistical inference is: Does the age of the buyer affect the cost of the car he purchases?

Posing interesting questions and then inferring answers based on the incomplete data is statistical inference.

Machine learning

Machine learning is a new programming paradigm, a new way of communicating your wishes to a computer.

Machine learning algorithms **find patterns** in your data, and turn them into instructions you couldn't write yourself. To elaborate further, it tries to find patterns in data and tries to figure out the answers on data where the information is not present. Let's look at an example.

> *Predicting if a customer will purchase a new car, based on the marketing campaign.*
>
> *Machine learning works on historical data of which campaigns resulted or did not result in a car purchase. It figures out the patterns and based on what it learnt, for new campaigns it gives a prediction on whether the customer will purchase the car or not.*

Don't worry about these definitions; let's keep our focus on solving problems, because that's what Data Science is for!

DATA SCIENCE IN THE WILD

Let's look at some general use cases of Data Science in various organizations.

Amazon Go

Walk in and walk out of a store without stopping near a cashier.

The cashier-less checkout store at the Amazon headquarters has a lot of Data Science going on in the background. It uses different techniques to figure out which items a customer is taking on and off a shelf. Once the customer leaves the store, the final bill is calculated and charged against the customer's Amazon account.

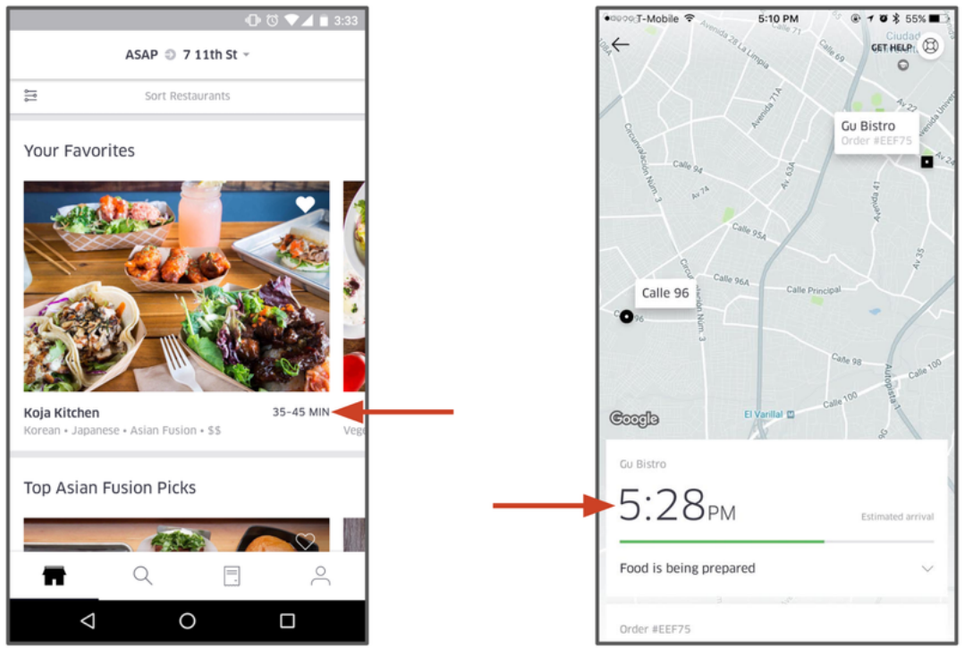
Image credit: Shara Tibken/CNET

Uber

An interesting use case of Data Science at Uber is the restaurant delivery estimate for UberEats.

The Data Science models predict each step of the delivery process, right from how much time it will take to prepare a meal to its delivery time, all before the order is issued. Food delivery is a multi-step process, starting from a restaurant acknowledging an order, moving on to food preparation, to the assignment of a delivery person, and finally to delivery at the user's doorstep.

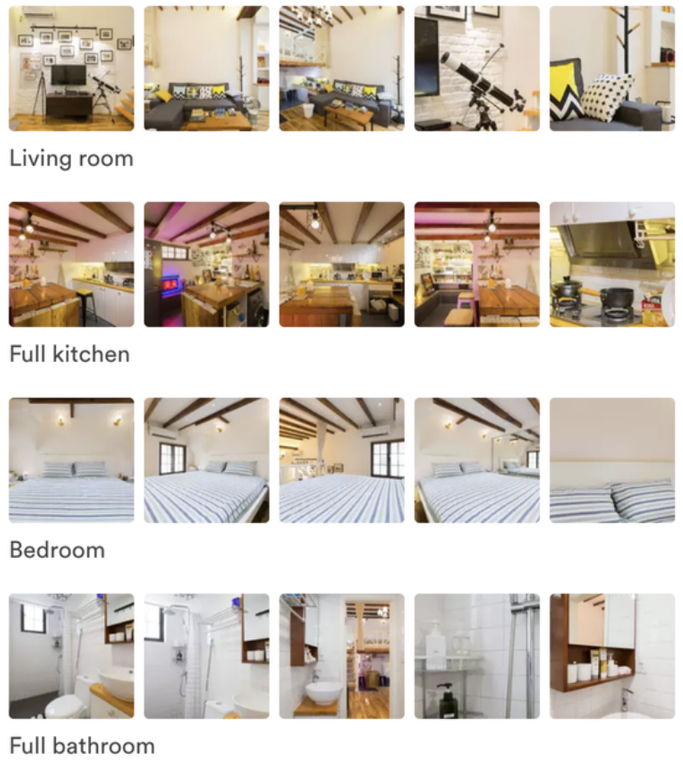
To make a decent prediction end-to-end, and readjust these predictions at each step of the process, lots of Data Science processes take place in the background.



Airbnb

Airbnb is a rental marketplace which has pictures of houses and rooms that are available for users to rent.

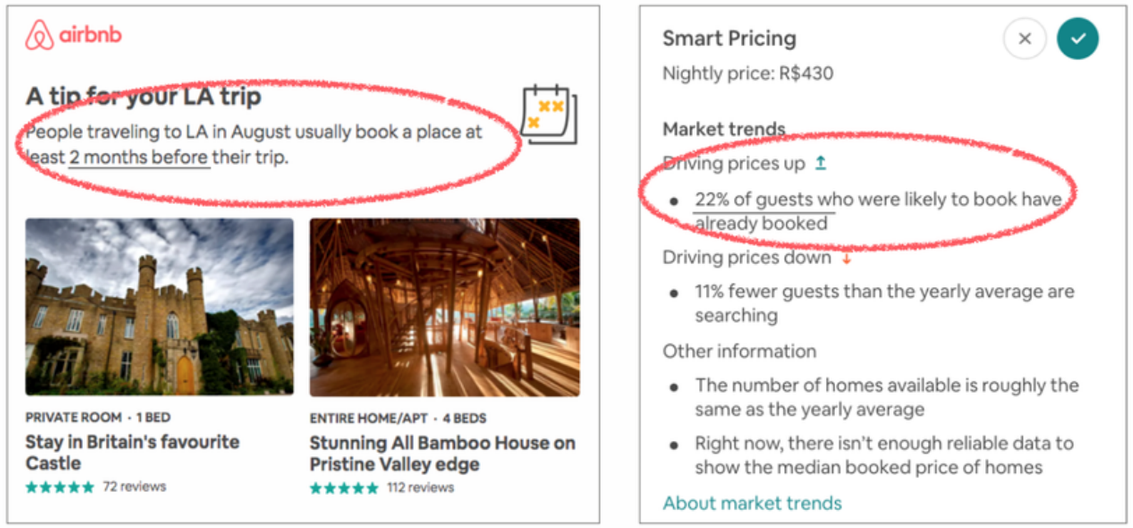
As these pictures are user-uploaded, they often do not have the correct tagging information, which part of the house they are in, for example. The Data Science team at Airbnb **built models to automatically classify images** of the different parts of home, bedroom, living room, etc, to better communicate the listing to users, and increase the chances of a house being booked.



Another interesting use case is the prediction of an optimal rental pricing for a listed property.

For tourist destinations, booking a stay close to a special occasion, like New Year's eve, will result in premium prices. In fact, as the user moves closer to the occasion, the higher the price will go.

Data science is used to predict important parameters for optimal pricing to better inform guests of the ideal time to book a home at a particular destination. Even the hosts are given information about the right pricing strategy, with insights into trends pushing the rent higher or lower.



MORE THAN MEETS THE EYE

These use cases are just a small sample of a large world. These are general use cases for tech companies. What about other industries? Data Science is ubiquitous in the true sense of the word, and is making its way to core industries too.

Data Science is required to maintain competitiveness in an increasingly data-rich environment. Much like the application of simple statistics, organizations that embrace Data Science will be rewarded, while those that do not will be challenged to keep pace.

We will look at how businesses define problems and then solve them using various Data Science techniques.

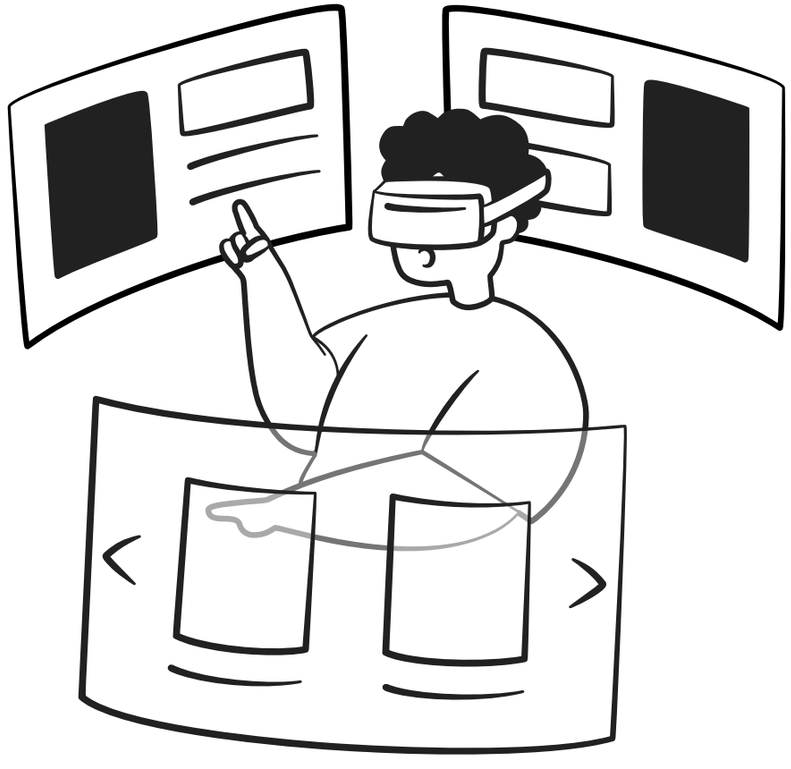


Image credit: https://www.webalys.com/

1

SETTING THE STAGE FOR OUR BUSINESS: A RESTAURANT AGGREGATOR

INTRODUCING YUMEATS!

An international restaurant aggregator company, YumEats! has decided to expand their operations to enter the Indian subcontinent. YumEats! allows users to select food from restaurants and get the orders home delivered.

The business model

Like every aggregator in the space, they have 2 business channels, B2C and B2B, that can be leveraged.

- B2C: The B2C revenue comes from a special rewards program or an exclusive membership program that allows consumers to subscribe to a monthly or annual membership.
- B2B: The B2B revenue comes from collecting a percentage of the order as commission from the restaurants.

Hence, the company has to acquire both consumers as well as other businesses.

Where do you fit in all of this

YumEats! has set up a Data Science wing in their company and hired you as a Data Scientist on the team. They want you to analyse data and help the company expand in terms of both growth and revenue.

Growth and revenue expansion would be possible with, both, more daily active users (DAUs) on the platform, as well as lots of restaurants with a wide variety of cuisines that would drive more traffic to the app. They have already run a whirlwind campaign with deep discounts that are increasing the losses and adding strain.

You have been given a mandate: figure out where to cut spends with minimal damage, and simultaneously find ways to increase revenue streams with data-driven insights.

THE PATH TO FINDING A SOLUTION

You know the basics of machine learning and understand the basic nuts and bolts of algorithms. Great! Awesome! But before you apply the algorithm, you need data. And before you collect data, you need to be clear about the actual business problem you are solving.

At the end of the day, businesses look to Data Science teams to generate insights and help solve problems. The journey from a business problem to a Data Science problem is not straightforward, hence we will make an attempt to demystify the process.

We will transform the field of Data Science into a set of simplified activities as shown in the figure. Data Science purists will likely disagree with this approach, but then again, they probably don't need a field guide, sitting as they do in their ivory towers! In the real world, we need clear and simple operating models to help drive us forward.

The process of constructing a Data Science solution to a business problem is often takes the following path:

Define the problem

This is one of the most underrated yet the most important phases of the Data Science lifecycle. In this phase, you choose the business problem you will try to solve and work with the stakeholder who is going to implement the actionable insights of the solution.

Construct the dataset

Once the problem is fixed, the next step is decide what data is needed for solving the problem and collect it. The data needs to be collected and cleaned before it can be used for analysis.

Select and transform features

Next, we need to select the right features that can aid in our analysis and help detect the patterns. Also new and novel features can be generated from existing features. These features are then given as input for ML algorithms.

Train and tune models

The features are fed as input to ML algorithms to get the ML model as output. This model is trained on historical data and tuned to get the best possible result.

Use the models to make predictions

The models are then applied on new data to make predictions for the business to action and make better decisions. If a core business pain point is not solved, then the Data Science project is a failure.

SUMMARY

While this ebook will focus mostly on how to define the business problem and setting up objectives for the case study, we will also briefly touch upon the other steps and ultimately demonstrate how to solve the business problem.

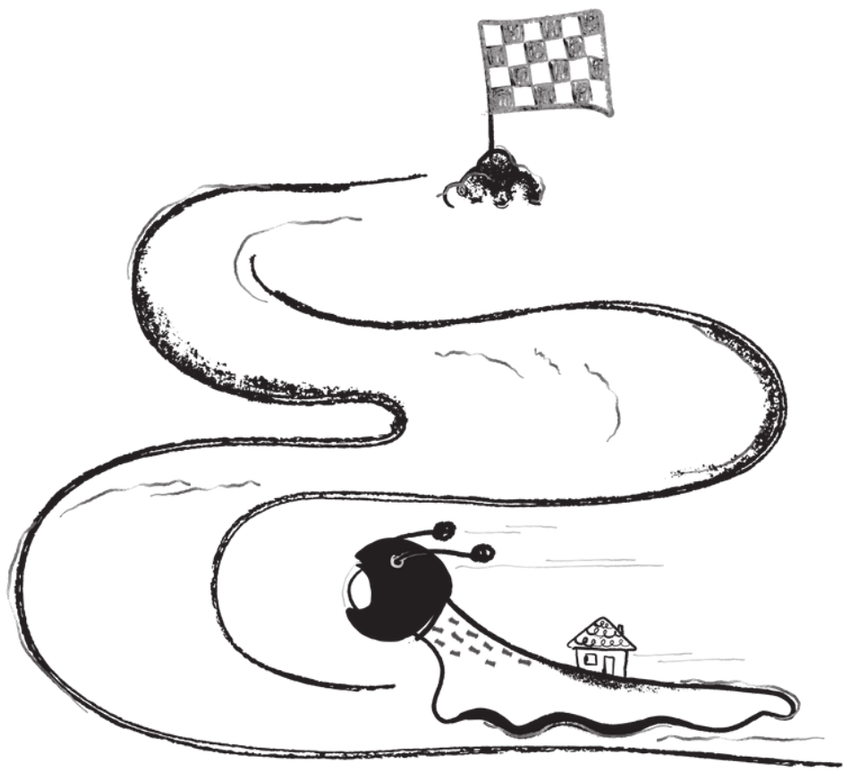
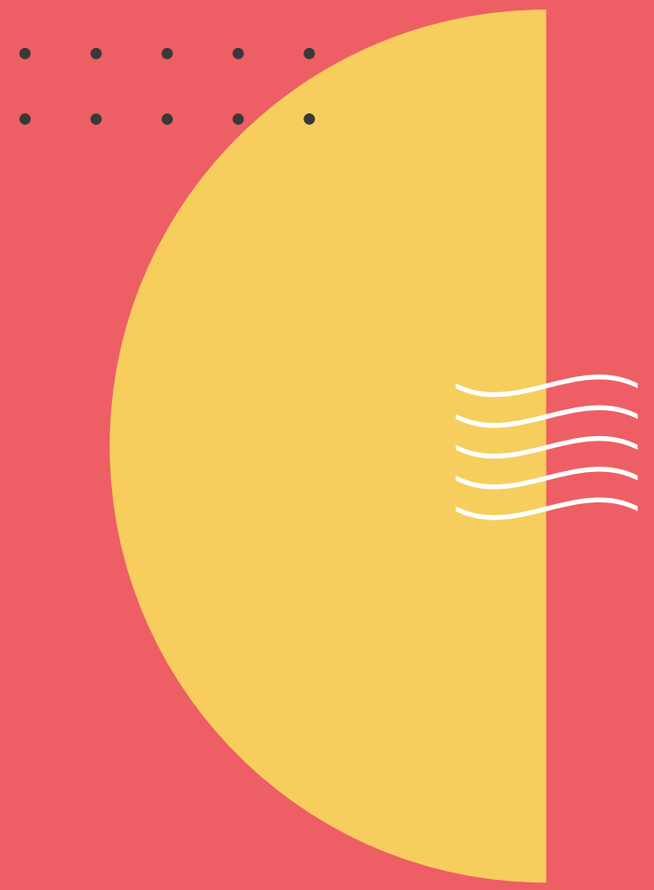


Image credit: https://absurd.design/

2

NARROWING DOWN THE PROBLEM STATEMENT



IT’S NOT AS SIMPLE AS IT LOOKS!

The first step of the path — defining the problem — contains tasks such as:

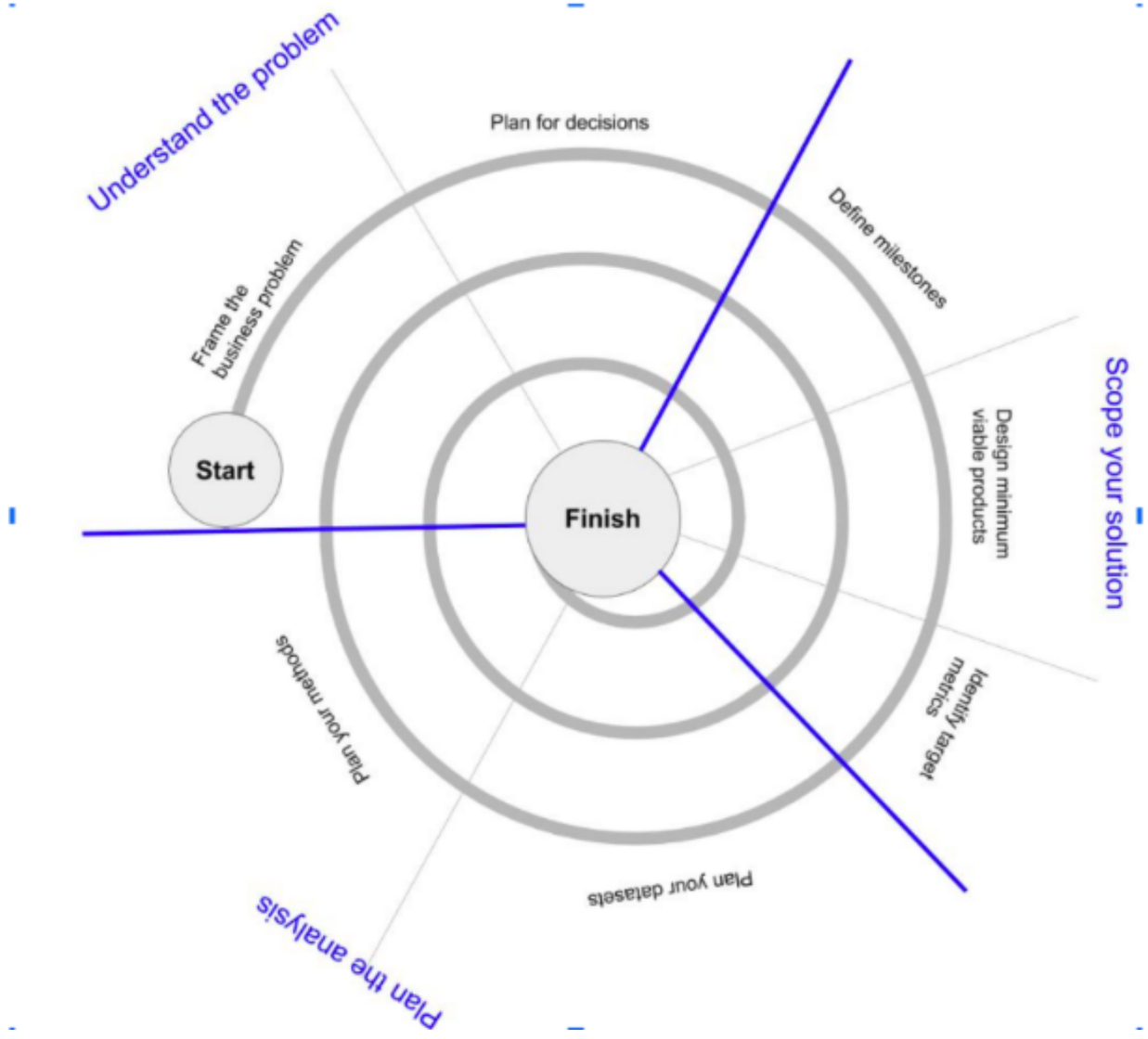
- Understanding business needs
- Scoping a solution
- Planning the analysis

However, while translating a business problem into a Data Science model is a process, it is not linear. Each step in the process usually needs to be revisited multiple times in order to arrive at an analytically sound, maintainable, and scalable solution.

Circular process not linear

The “define the problem” box in the simple linear diagram above can actually be exploded into a much more nuanced process:

It is very common to initially define a business need, but then, as you proceed to more fully scope the problem, you realize that an entirely different need is more pressing. Likewise, it is common to scope a solution, only to realize later that data access limitations or engineering constraints require a change in that scope.



Often, these changes in plans will even occur after you think you have left the “define the problem” stage of the process. For example, it is common for model tuning issues to raise scalability concerns which may require a substantial re-evaluation of what problems you are trying to solve and how you plan to solve them.

More than just algorithms

A large amount of a Data Scientist’s work takes place away from the computer, as Data Dcientists must work with non-technical co-workers to define the goals and scope of their projects. A well-understood business problem and a well-designed plan of action will lead to better results, less wasted effort, and happier stakeholders.

Image credit: https://www.webalys.com/

TEST DRIVE THE PROCESS WITH ANOTHER EXAMPLE

Let us take a simple example to look at each of the steps that we mentioned in our diagram. The problem we are looking at is to increase the number of retweets a tweet can get.

Steps

- ***Define the problem:*** *Predict the number of retweets a new tweet would get.*
- ***Set the objectives:*** *Identify the features that would influence the number of retweets.*
- ***Prepare the data:*** *Trends corresponding to the hashtag accompanying the tweet would be a good signal.*
- ***Build and train the model:*** *How accurate is the model? How much of an error can the business tolerate?*
- ***Make predictions and fine-tune:*** *The initial predictions were quite off. Think of further refinements. Can this problem be solved? What are the barriers for an effective solution; data? computation?*

We want to arrive at the answers to these questions in a systematic way. Our case study is going to help us to do just that.

Let's solve the YumEats! problem

The YumEats! problem is a business problem that is presented in its most raw form: improve revenue and cut losses using Data Science.

DON’T BREAK OUT THE ALGORITHMS JUST YET

Before any form of analysis can be performed, a thorough understanding of the problem is required. To understand the problem in its entirety, you first need to talk to the respective stakeholders.

Data scientists always work with stakeholders. Stakeholders are people who have a say in how the business operates and in what goals the business needs to prioritize. Stakeholders could be managers and executives, but they could also be individual contributors who have responsibilities over specific aspects of marketing, engineering, sales, finance, operations, or any of the facets of a business enterprise. Different stakeholders have different requirements

Types of stakeholders you may have to work with include:

- A product manager will likely be more focused on customer-facing features.
- An engineering manager is likely to be concerned with the maintainability of a product, and would aim to minimize the extent to which changes will create unanticipated work for his or her team.
- An executive stakeholder (like the one we are working with) is likely to be focused on the bottom line — he or she won’t care too much about the product’s maintainability or about specific features as long as he or she can be assured that revenue will increase, or a client’s business will be retained, or expenses can be cut.

Identify the YumEats! Stakeholders and their POVs

Because of the direct revenue impact and cuts on spending needed, the stakeholder could be a leadership level person on the marketing or sales teams, possibly a CMO or Director, Sales.

When faced with a scenario like the one given above, the first instinct of most Data Scientists is to consider different methods they could use to achieve the desired result. That is almost always the wrong reaction to this kind of situation.

Things to consider

- The stakeholders have presented a very vague problem statement. Improving revenue and cutting costs is too wide a problem. There is a need to narrow down this problem.
- The stakeholders are the ones who will be consuming the insights you present and driving the required change. Understand the constraints (if any) that you need to operate under. Align with their agenda and make your objective as close as possible to theirs.

Why you should take a moment to consider

Most of the time, however, people who ask for Data Science help do not know how to ask questions in a way that are Data Science-ready.

While most Data Scientists are used to thinking about analysis in terms of method, features and variables, and data transformations, most stakeholders are used to thinking about analysis in terms of spreadsheets or other tools they are familiar with. When they confront a business problem, they think: “How would I solve this if I had to do it myself?”

They are asking the best question they know how to ask, but that question needs to be translated and shaped into something a Data Science can act upon.

The CxOs of YumEats! are asking the right questions

- How to increase revenue for the company?
- How to cut spending and reduce the losses and move towards profitability?

As a Data Scientist, you must translate these questions into actual Data Science problems that need to be solved. When asked to use “Data Science” to solve a problem, your first task is to think of ways you can ensure that you understand the problem. You need to meet their problem on their terms, not yours.

SUMMARY

- Always remember, the right direction is more important than speed. Huge progress in a completely wrong direction makes no sense.
- Spend enough time to understand the problem and validate if it is the right direction to pursue.

Image credit: https://iconscout.com/

3

GETTING DOWN TO THE BRASSTACKS: FRAME THE BUSINESS PROBLEM

BREAK IT ALL DOWN INTO DIGESTIBLE CHUNKS

In order to further understand broad questions, you need to identify the subproblems that could be solved. Talking to stakeholders and brainstorming with them will lead to these possible problems being identified.

Increase revenues

- Conduct marketing campaigns to acquire new users; understand user profiles to deliver customized campaigns.
- Predict when the users are likely to order next from different trends and nudge them to order.
- Recommend restaurants to users to order from next, based on their previous orders.
- Perform existing customer segmentation and run a tiered campaign: budget restaurant recommendations for low-paying users, premium restaurants for high-paying users.
- Identify good restaurants that are not yet onboarded on the platform to increase the roster.
- Provide recommendations to new onboarded restaurants on cuisine preferences of the people in their localities and give on-the-fly menu suggestions based on trends.
- Automatically check the image quality of restaurant and food photos, and enhance low quality photos to drive traffic.

Reduce costs

- Understand which customers are likely to leave the platform and only provide discounts to retain them.
- Automatically identify orders from nearby localities and assign it to a single delivery person.
- Automatically identify fraudulent orders beforehand, and trigger orders only upon verification.

Identifying subproblems has narrowed the scope on the different kinds of the problems that can be solved.

SPEAK THE RIGHT LINGO

As a Data Scientist, you use a specialised vocabulary to describe your work and the results of that work.

Some people, such as engineers, will understand some of your vocabulary. But most people (especially the stakeholders) will understand practically none of it. If you start talking about gradient boosting trees or k-means clustering, you will at best get blank stares and sometimes you might even encounter hostility.

Develop your own dictionary

Similarly, people who work on marketing, product management, or executive teams will speak about their work in terms that are not quite clear to you. To understand business stakeholders on their own terms, learn to ask good follow-up questions.

A good follow-up question encourages stakeholders to illustrate what they mean without realizing that that is what they are doing.

Good follow-up questions for YumEats! CxOs

- “You’ve talked about increasing revenue. What were some of the past initiatives for increasing revenue? What are the major pain points that you experienced that led to loss of revenue?”Predict when the users are likely to order next from different trends and nudge them to order.
- “In your rich experience of market research, what is it that the user is looking for? How can we improve the user experience on the app?”
- “You might have seen users leave the platform. What are their main reasons for leaving the app?”

Notice that none of the previous questions had a “straightforward” answer. Asking these questions helps stakeholders clarify their thinking, gives you additional concrete examples of what the problem looks like, and attempts to elicit information that might help scope the problem.

GUIDELINES TO ASK QUESTIONS

The goal is to ask clarifying questions that can get your stakeholder to give you more useful details:

- **Get concrete as fast as you can:** If a stakeholder talks about “what users want”, ask them to tell you about one specific user or a segment of users. People are natural storytellers: let them explain a problem in story form. You will get less distorted (though not necessarily less biased) information.

> “Can you walk me through the behaviours of some of the high frequency users of the YumEats app?”

- **Focus on pain points:** Find out what users are trying to do with the product. Find out their pain points and if the problem you are thinking of solving can alleviate those pain points. Prioritise the pain points to, in turn, prioritise the Data Science problems that might ease out these pain points.
- **Look for opposites:** If we have received inputs from stakeholders on users who were unhappy with the app, try to gauge information from the other group too — users who were happy with the app. This will help build a balanced perspective. Look for behaviours that must be rewarded and the ones that need to be changed and identify biases.
- **Find hidden problems:** The problems someone asks you to solve may not always be the most pressing problem. Look for problems that stakeholders mention incidentally as they tell about what they think is their main problem.

> The stakeholder pointed out that the users spend quite a bit of time on the platform without ordering. It may be worth it to ask: “Can we improve the ranking of recommendations to align with the user’s interests and push them to order sooner?”

Why the right questions are important

Asking clarifying questions serves several purposes:

- It demonstrates that what is important to your stakeholders is important to you, too. It builds trust and establishes a rapport, which are two things you will need when time comes to share the results of your work.
- It fleshes out your understanding of the problem. It is easy to assume that you understand what people want. It is much better to take a little extra time to reduce the possibility that you misunderstand.
- It forces stakeholders to confront some of the complexities of the problem they are asking you to solve. It is easy for a stakeholder to assume that a problem is easy to address because it is easy for them to talk about. By asking clarifying questions, you force them to consider contradictions and nuances in their story.

Refining the YumEats! problem

From the previous exercise, you would have identified the stakeholder for your problem. Do these next steps to prepare for the next stage:

- Think of customer retention as a large problem and then break it down into different subproblems.
- Think of different ways to retain customers.
- Take cues from any of such apps you have been using or talk to friends who regularly use such apps.
- List down the different ways or strategies to retain customers just like we have done above.
- After listing down the different subproblems, next list down the different clarifying questions you would ask the stakeholders.
- Ask your friend to assume the role of a stakeholder and pose your clarifying questions and jot down the answers.

SUMMARY

- Break a bigger problem statement into smaller chunks.
- Speak to stakeholders in their language to understand exactly how they view the problem.
- Ask clarifying questions to get a deeper understanding of the problem at hand, and also to get stakeholders to communicate the complexities of what they are asking for.

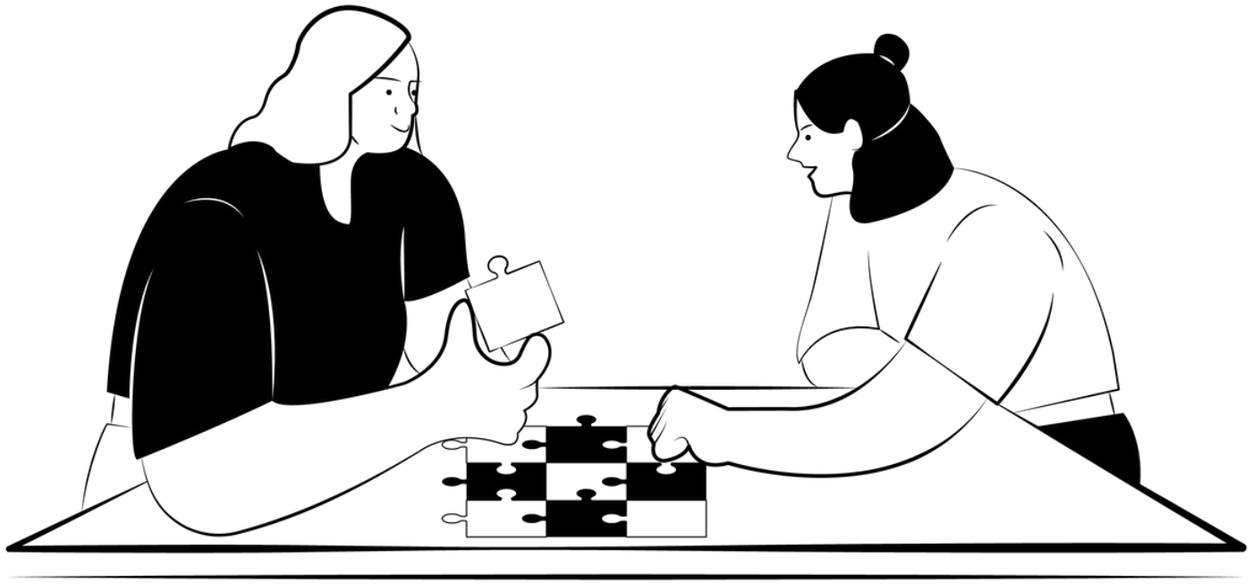


Image credit: https://iconscout.com/

4

WALKING THE TALK: PLAN FOR DECISIONS, NOT FINDINGS

As Data Scientists, we often think about the results we produce in terms of findings: we conduct an analysis, validate certain results, and those results say something useful about how the business is doing or what the business should do next.

For non-technical stakeholders, however, findings are almost always irrelevant.

DECISIONS, DECISIONS

Stakeholders need to make decisions, and you should never assume that you fully understand what those decisions are, and you should definitely never assume that stakeholders will be able to naturally map your findings to their decisions.

Consider the following questions related to the case study

“Which users must be given the discounts to stay on the app and when to trigger them?”

“To a new user who has just landed on the app, what is the right campaign to show?”

“For a new restaurant onboarded on the platform, which dishes must be highlighted on the menu?”

All of the above questions are designed to elicit information about decisions. A stakeholder needs to make a decision. Any findings you produce should help them make those decisions.

Notice that these questions point to the basic “who”, “what”, “where”, “when” and “why” of how the app uses data-driven insights. Asking these questions helps you create a map of decisions and outcomes that will need to be considered when implementing the solution you eventually develop.

INTERVIEW STRATEGY

Here are some guidelines for mapping out the relevant decisions, which your stakeholder has alluded to rather than defined, in a more formal and explicit way:

- **Understand timing:** People have to make decisions at certain times, within certain timeframes, and on certain schedules.

> You could ask, “When should a particular campaign be shown to a user to ensure maximum conversion?”

- **Understand expectations:** Set the expectations clearly, upfront. Clarify the timelines for the Data Science solutions to start showing results.
- **Understand downstream effects:** Even though one stakeholder might ask you for a solution, they might not be the only person impacted by what you deliver.

> Discounts might make users happy, but restaurant partners might get worried about the impact on offline business. The concerned B2B sales director must be made aware of the potential impact.

- **Understand when the business problem isn’t a Data Science problem:** The most important thing to realise is that not all business problems can or should be addressed through Data Science. Make sure you feel confident that the problem is solvable in principle, and that using Data Science to solve it is the most cost-effective way to go.

SOLVE FOR MAXIMUM IMPACT

Do spend time in analyzing all the possible data science problems, in order to determine which one to solve. The ideal sweet spot would be solutions which lead to:

- **Quantifiable impact for users:** increase in daily active users, increase in daily orders.
- **Quantifiable impact for stakeholders:** increase in revenue with reduced costs.

1 problem statement or 2 problem statements?

By now, it must have become quite clear that increasing revenue while cutting costs is a tricky proposition for they seem to contradict each other. To increase revenue, you need more DAUs which would be acquired by various campaigns that would incur further cost.

Another approach for improving revenue would be to get more restaurants onto the app. More restaurants implies more choice, which in turn drives more DAUs and more volume. Further, it gives scope for a membership program or exclusivity agreement to drive more users to the app. If more good restaurants can be onboarded then there could be an increase in revenue. Improving restaurant quality could be a long term growth lever.

You present your rationale to the stakeholders and they make the projections of cost vs value of different initiatives. After a lot of deliberation with the stakeholders, you now agree upon the actual business problem that you would like to solve: **Identify good restaurants to be targeted for onboarding onto the platform.**

With this problem, you are trying to improve the choices of good restaurants for a user and improve the quality of restaurant partners on the app.

In the next chapter, we will talk about scoping the solution to this problem.

Identify ONE problem to solve

From the different subproblems that you have listed, shortlist the problems where you feel Data Science can make an impact or solve the problem.

Go back to the workflow for applying Machine Learning covered in the beginning. Next, think about what decisions the stakeholders can take to solve the shortlisted subproblems.Remember that we are focused more on decisions and less on findings.

Further, analyze the potential impact of solving each of the shortlisted subproblems. Discuss and arrive at ONE problem statement that you would like to tackle.

SUMMARY

- Solve problems that will lead to concrete decisions, not just analysis.
- Try and anticipate the impact of the potential solutions that you will be presenting to stakeholders.

Image credit: https://www.webalys.com/

5

FIX THE GOALPOSTS: IDENTIFY PROJECT MILESTONES

Let's recap where we are with our case study

In consultation with the management of YumEats!, we identified the business problem to be solved: **Identify good restaurants to be onboarded for the app.**

The primary stakeholder for this business problem would be the Director of Sales, B2B.

The impact would be that increased active restaurant partners with high ratings would improve the brand, drive more active users and increase daily orders. More restaurants onboard bring greater choice and more users to the app.

If famous, yet not very visible, restaurants are on the app, it would lead to a greater value proposition for the users.

The business problem is now more or less clearly defined and framed. This is the first step to move towards solving the Data Science problem.

THINK IN TERMS OF RESULTS

The business stakeholders have helped you define what the business needs. It's now your job to translate that business need into analytic needs. That means there are a lot of questions that still need to be answered:

- What analytic goals do you need to accomplish in order to justifiably claim that you've found a solution to the business problem?
- What are options for reaching those goals, and which options are most cost-effective, in terms of both time and resources?
- How will you be able to measure the extent to which your proposed solution addresses the business problem?

In most cases, you cannot and should not wait for other people to tell you what steps you need to take to solve a problem. Part of your job as a Data Scientist is to define the path to a solution, not just take the path others have laid down.

Do not think about methods and algorithms yet. Your task right now is to plan out what a viable solution will look like. Later, you will consider how to turn that plan into a reality. For the very first step, we would need to define the milestones.

YumEats! milestones

The goal of **identify good restaurants to be onboarded** is clear enough from a business perspective, but in terms of running an actual analysis we need to further break it down into smaller milestones.

It often helps to re-frame the business goal as a question. In the case of the scenario above, we might reframe the goal to “how can we identify which restaurants are the right restaurants to onboard on the app?”

That question is still largely unanswerable — it still too vague — but it paves the way for a few smaller questions, such as:

- How do we define the various categories of restaurants — excellent, decent, bad?
- How do we estimate the popularity of the restaurants?
- What are the features that differentiate good from bad restaurants?

If you answer all of your milestone questions, you answer your large business question as a matter of course. If you answer your large business question, you’ve addressed your business problem.

The three milestone questions are not necessarily the “right” questions to solve the business problem. It’s less about finding the right milestones and more about making sure you have milestones to work with.

Now, you have a reasonably clear analytic goal: **identify good restaurants that are likely to improve the brand.**

LEARN TO PRIORITISE

Here are some guidelines for creating good analytic milestones for the project:

- **Eliminate possibilities:** It's natural to want to use all available data to try solving a problem, but it is often wiser to think of important data points for the analysis.

  Further, you can categorise them as Critical, Good-to-have, etc to prioritise which features need to be collected first. This kind of feature and data selection does not require any particular method; it relies on domain knowledge, which is something you can get from your stakeholders.

A Critical feature for the restaurant is its location, and its year of establishment. The average number of daily visitors to the restaurant is a good-to-have.

Here are some guidelines for creating good analytic milestones for the project:

- **Think about dependencies:** If you can identify one thing that you think you need to accomplish, ask yourself: "Is there anything I need to get done before I can do this?" and "Once I do this, what will I then be able to do?"

  You don't have to plan all milestones in order: identify just one and then work backwards and forwards from that point to identify the rest.

First, we need to identify how we can identify a restaurant as good or bad. Next, we can think of the metric that can help satisfy the objective. Then, we can start thinking of features and how to collect data.

ORGANISE YOUR MILESTONES

Group milestone activities by the entity. When an analysis involves multiple entities, it often makes sense to create at least one milestone per entity, and then at least one milestone to tie the entities together in a way that solves the business problem.

> Business stakeholders have already mentioned that they want good restaurants to be identified. So part of the analysis should include a comparative analysis of good and bad restaurants and find what separates both.

Importance of milestones

Defining milestones serve several purposes:

1. It helps you **anticipate difficulties** you confront as you develop your analysis. If you can see these difficulties before they actually occur, you can prepare for them or sometimes avoid them entirely.
2. It helps **communicate your work** to other stakeholders and **provide visibility**. If a project has five milestones and two are completed and one is set for completion the following week, this clarity helps stakeholders plan around the Data Science work and therefore support it better.
3. It **imposes order**. In large projects especially, it is easy to get lost in all the details of data cleaning and exploratory analysis. Relatively inexperienced Data Scientists will often see their original timelines balloon as they discover new aspects of the data which then require additional investigations. By setting out milestones ahead of time, it is easier to stay on track.

Now with the milestones identified, the next step is to build out a minimum viable product.

Identify milestones for the problem

Now that you have identified the problem you wish to solve; which helps in achieving customer retention on the app. Drill down into various milestones of the one problem you have set out to solve on the app. Go through the steps mentioned about eliminating possibilities and identifying dependencies. The output of this exercise is to identify milestones for your chosen problem.

6

DESIGN MINIMUM VIABLE PRODUCTS

To use the analogy illustrated in the graphic above, data scientists are often asked to deliver cars.

Inexperienced Data Scientists will then try to figure out how to build the specific car they were asked for.

Experienced Data Scientists will try to figure out how to build a skateboard, and then figure out how to turn that skateboard into a scooter, and then turn the scooter into a bicycle, and so on, until they finally have built a car. Even if they never build the car, they've still delivered enough skateboards and bicycles and other means of helping their customers do what they want to do.

Consider the following questions:

- What is the smallest benefit stakeholders could get from the analysis and still consider it valuable?
- When do stakeholders need the results by? Do they need all the results atonce, or do some results have a more pressing deadline than others?
- What is the simplest way to meet a benchmark, regardless of whether youconsider it the "best" way?

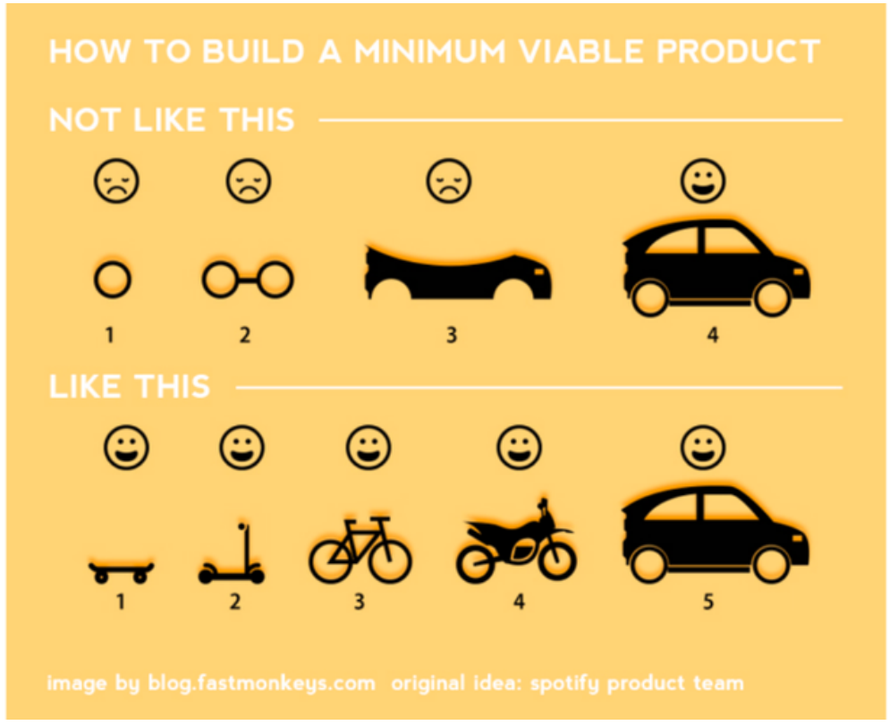


Photo credit: Fast Monkeys Official Blog

A minimum viable product (MVP) allows you to provide value to your stakeholders in smaller increments, which makes them happy, and reduces the risk of having to throw away months of work because of misunderstood or miscommunicated requirements, which makes you happy.

In our case, a minimum viable product could be just an analytic dashboard showing a visualization of the restaurants already on the app.

Group the good and bad restaurants separately and visualize the features in comparison with each other. Have a naive rule-based approach to identify good or bad restaurants to set up the benchmarks.

JOURNEY OF A DATA SCIENCE PRODUCT

Typically, a Data Science product will have a certain path:

1. **Analytic solution:** look at existing data to analyze patterns
2. **Diagnostic solution:** look at data to explain the past
3. **Prescriptive solution:** look at data to provide insights and decisions

Instead of simply trying to order the analytic work, here are some ways you could deliver minimum viable products over the course of producing your full results:

- **Plan in sprints:** Set an arbitrary amount of time — typically 2 or 3 weeks — and ask yourself: “what would I deliver if I had to deliver a solution by the end of that period?” Your answer to that question is probably a good “skateboard” approach.

What is the most you could hope to accomplish in two weeks? Maybe you feel it would be realistic to just show the general trends in good and bad restaurants (analytic solution).

- **Think modularly:** Once you have a general idea of something you want to deliver, pause and ask yourself if there is a way to split that deliverable into smaller deliverables that are useful all by themselves.
- **Get feedback:** At every step of the product, get feedback from stakeholders.

Make a simple dashboard out of your restaurant analysis. Showcase the rule-based simple implementation (older restaurants might be good) and check if the results are in the right form (even though they might not be accurate).

WHY MVP IS THE WAY FORWARD

Creating minimum viable products serves several purposes:

1. Data scientists usually find it easy to think about analytic details and relatively difficult to think about value delivered to the business. Building minimum viable products forces you, as the Data Scientist, to think more about the considerations that are easier for you to overlook.
2. It helps interested stakeholders make a case for ongoing support of your work. They will be more patient since you regularly show incremental value.
3. Business needs change constantly. If you take six months to finish a project, chances are half of the needs that motivated the project in the first place will no longer exist, and the other half will have substantially changed. By building incrementally, you minimize the chance that you work will be outdated before it is even deployed.

Image credit: https://www.webalys.com/

Identify the minimum viable product you will build

In the previous exercise, you identified the milestones for the app. Now, from these milestones, come up with a minimum viable first deliverable that will satisfy your stakeholder.

In the example above, it was a simple dashboard built upon the existing app data. On similar lines, can you think of what the minimum viable product would be?

What would you build in 1 week and 1 month from now? What could be the skateboard version of the problem you have identified for solving?

7

IDENTIFY TARGET METRICS

As we plan the roadmap for our Data Science project, one thing we must keep in mind is how we will measure the success of the project. This can take many forms; for instance, one obvious success metric is the actual business outcome stakeholders want to achieve.

In the YumEats! case study, the actual business outcome achieved would be, if they use your Data Science solution, and onboard the suggested restaurants, and attract higher DAUs in a shorter time compared to the manual checks and onboarding.

Also think in terms of:

- Why should anyone trust the results of this analysis?
- What is the confidence level of the prediction of the restaurants? Can they go ahead blindly with the suggestion, or are some other checks needed?
- Where does the bulk of the value come from? Are there parts of the analysis that are more valuable than others?
- Along with the suggested restaurants, can you solve other problems, like suggesting dishes to be highlighted for new restaurants?

You may have extremely high confidence in the quality of your analysis, and yet the results of the analysis might not be cost-effective for the business to implement.

Coming back to the case study, we need to identify the target metric that we would use to measure the success of the problem. The problem statement is to identify good restaurants to be onboarded on the app.

Now, though a well-defined business problem, it is still subjective. For the right metric to be applied a slight reframing of the problem is needed.

> Currently, the ‘goodness’ of the restaurant is subjective, and to define the metric it must be objective. An objective measure of the quality of a restaurant is the rating of the restaurant given by users. So the problem changes to identifying restaurants with a high rating.

A metric must be measurable, so you can use it in Data Science.

> Suppose you decide to predict the ratings of a restaurant to decide if they can be onboarded on the platform. Now let’s skip ahead in our thinking to consider what proof of value we want or need to be able to deliver to the stakeholders in our case study.

HOW TO SELECT GOOD METRICS

There are guidelines to establish good metrics. Here are some of them:

- **Think explicitly about trade-offs:** Almost any metric will involve a trade-off. For example, in a classification problem, “precision” focuses on minimizing false positives, while “recall” focuses on minimizing false negatives. False positives might be more important to the business than false negatives, or the reverse could be true. It helps interested stakeholders make a case for ongoing support of your work. They will be more patient since you regularly show incremental value.

Out of a rating of 5, consider restaurants with ratings of 4 and 5 as good, and the rest as bad. Which is more harmful, identifying good restaurants as bad, or identifying bad restaurants as good? The stakeholders tend to be conscious of brand image and don’t want to onboard a bad restaurant. Hence the metric to optimize could reduce the false positives or ‘precision’.

- **Figure out the business’s “value” units:** Business stakeholders practically never think about value in terms of root mean squared error or precision. Maybe they think about customers served, or revenue generated, or hours saved. Find out what unit of value your stakeholders think in, and estimate the value of your analysis using that unit.

For example, stakeholders have said that they want to get good restaurants, but upon further investigation, you might find that what they really want is increased orders, which in turn impacts revenue and brand image.

- **Subset all metrics:** An analysis should almost never have only one set of metrics. All metrics used for the analysis as a whole should be repeated for any relevant subsets. An analysis may perform very well on average but abjectly fail for certain subsets. That is relevant information that your stakeholders should have when making decisions.

Some subsets could be restaurant categories, cuisine, regions, etc.

- **Keep it as explainable as possible:** A good metric does not always have to be easy for non-technical stakeholders to understand, but non-technical stakeholders do need to be able to understand whatever metrics you use.

  If you choose a metric that is hard to explain, then you will need to make the extra effort to help stakeholders understand it. If you can find an easily-explainable metric that is still appropriate, you can focus your time on other things.

Assess the comfort level of stakeholders towards technical metrics. Consider re-framing technical concepts such as “false-positive rate”, and “false-negative rate” as “wrongly identified restaurants” and “missed opportunities”.

Image credit: https://www.webalys.com/

GOOD REASONS TO IDENTIFY METRICS

Identifying target metrics serves several purposes:

1. It makes you clarify your and your stakeholders' thinking about what value the analysis is really meant to achieve.
2. It can keep you from pursuing interesting analytic questions that don't ultimately lead to value for the business. If a question won't help you produce one of your target metrics, it is probably out of scope of the project.
3. It keeps you focused on explaining and justifying your work, which helps those around you support you better. If other people understand what you are doing and understand what value you are providing, they can help get you the attention and resources you need to continue your work.

We can frame the Data Science problem as a regression problem, where we predict the rating of the restaurant; or a classification problem, where we bucket restaurants into good or bad, based on users' ratings.

For the first attempt at solving the business problem, let's go with the simple problem of classification i.e. identifying if it is a good restaurant or a bad restaurant.

Target metrics

If you have solved the exercises till now, great work! You are really close to a properly defined Data Science problem that is distilled out of the identified business problem.

Now for the problem and the minimum viable product that you have identified, figure out the target metric for the Data Science model.

Remember to think through if the target metric is in line with the business metric you are setting out to achieve. Also you will need to drill down to figure out if the problem is going to be a regression problem or a classification problem.

8

EXECUTING THE DATA SCIENCE PROJECT

So you've already answered a lot of important questions about your analysis of the business problem. You know how and why it is important to the business. You're focused on a specific decision stakeholders need to make. You've identified what metrics you will need to make your case to the company's stakeholders.

FINDING THE RIGHT DATA SOURCES

You have a few more questions to answer before you should begin your actual analysis.

1. What data is available to answer your questions, and is that data sufficient for you to give an answer you feel good about?
2. How difficult it is to obtain the data that you are looking for? Is the data in public domain or does it incur costs to obtain the data that you need?
3. What is the form factor for the data you need? Is it in a neatly labelled format? If not available in the required format, how much effort does it take to label the data?
4. Which data can be acquired easily, and which data needs additional effort to acquire? Align your milestones to make a minimum viable product with easily acquired data first, and then add more and more data.
5. Does all the data you need exist in datasets that can be easily joined together? Or will you have to spend time figuring out how to link records across datasets?
6. How many pieces of data that you want can be missing or inaccessible before you decide that the analysis is simply not feasible?

Always remember that the key to solving the problem is obtaining, cleaning and wrangling with the data. An estimated 80% of effort is spent in this stage, so you need to be patient and try to question the data at every stage.

The first step of the analysis is to collect the data. Data is the key for success or failure for the Data Science project.

Our data sources could be

Scrape the data from other restaurant search and discovery sites, like Zomato, where a lot of information about restaurants is already present. These sites would have the ratings information as well.
Explore trending venues in particular neighbourhoods especially restaurants, using the FourSquare API, for example.
In addition, we can even scrape twitter to obtain conversations or tweets about restaurants of interest. Those can be an additional input of data.

Now, the data that comes from Zomato would be a table that could contain the restaurant name, cuisine, location, best delivery time, minimum order value, a few photos, reviews of the restaurant, recent streak of the reviews, and so on.

Suppose you collect data from another competitor like Swiggy or UberEats, while the data points would be more or less the same, the form of data could be different, the column names could be different, for instance. And there could be even additional information. For example, Swiggy might provide delivery for a particular restaurant, but the same facility is not available on Zomato.

GATHERING ALL YOUR DATA TOGETHER

Now imagine, getting these disparate disjointed datasets into a single table, on which further analysis needs to be done.

If you plan to do this at scale, you also need to get the engineering team involved in designing a complete data pipeline that can store the huge amounts of incoming data, and then provide data in a form that you as a Data Scientist can then work upon.

For now, let us assume, you have painstakingly standardized the different columns and gotten the data you need. Since the data is collected through a third-party source or API, there could be several inconsistencies, and so cleaning and transforming this data becomes really crucial.

> Consider the estimated delivery time of a restaurant. Let's say it is returned as a string. But any valuable insights or analysis would only be possible, if it is stored in a timestamp format. Therefore semantic parsing must be done.

What is presented here is just the tip of the iceberg. As you go deeper into data wrangling, analysis and aligning the data to the problem, you will find more challenges like these will arise and will need to be overcome.

NEXT UP, THE TARGET VARIABLE

Assuming the data is the format you need it to be in, the next thing you need to think about is the target variable.

The target variable for analysis is the rating buckets

The first step is to convert the ratings into a standard format. Some sites might rate a restaurant from 1 to 5, others might use a 1 to 10 scale. We need to transform these ratings into the right buckets of good and bad restaurants.

Many times, such conversions, where you are deriving values from existing values, might not be straightforward.

Other data operations can be:

- **Identify various data quality issues:** Any data corruption that can be identified? Incorrect data values?

If delivery times are shown in negative values, then you know that the data is incorrect, because time cannot be negative.

- **Resolve inconsistencies in the data:** Like missing entries and semantic errors, like incorrectly labeled columns.
- **Extract new features:** From existing features or identify new ones.
- **Encode the data:** Convert non-numeric to numeric data.

WHAT TO DO ABOUT MISSING DATA

Stakeholders often ask for things they can't have. It is especially common for stakeholders to want answers to questions, even though they lack the data needed to provide those answers.

> Suppose they want to predict the rating of a brand new restaurant (whose data is not available).

It's your job as a Data Scientist to identify data problems before you conduct your analysis, and to only spend your time trying methods that are appropriate to the situation. In spite of this, sometimes you won't realise that a crucial data point is missing until you are in the thick of your analysis.

With all the operations that you have done and questions you have answered, you have a rough dataset to begin with.

Let's assume the final form of the dataset to be as follows:

Target variable: Rating Bucket — Good or Bad (If you decided to go with regression, the target variable would be the rating itself.)

Features:

- Url of the restaurant page from where data was scraped
- Name
- Location of the restaurant
- Cuisine
- Cost for 2
- Delivery available [Yes or No]
- City
- Reviews
- Dishes typically liked by people
- Conversations around the restaurant from Twitter

Till we reach this point, be absolutely certain that you have aggregated the data sources. Now that a basic dataset is ready, your stakeholders are eager for you to start your analysis.

Here are some guidelines for planning your datasets:

- **Identify all dataset needs ahead of time:** Make sure you have all the pieces to the data puzzle available.

> Suppose they want to predict the rating of a brand new restaurant (whose data is not available).

- **Differentiate between necessity and sufficiency:** It is relatively easy to identify when the lack of certain information will make your analysis hard to do. That is a focus on necessity — the things you need in order to proceed with your work.

  It is harder to focus on sufficiency: even if you have everything you need, that doesn't mean you'll still be able to complete the analysis as planned. If data from different datasets don't have a common key on which to join the information, or you can't get access to some datasets even though they exist, or some of the data have so many missing values that they cannot support your use case, then your analysis will disappoint both you and your stakeholders.

> You could say: "We've identified where all the data is. Do all the data stores have a common column like a restaurant ID or name that can tie the datasets together?"

- **Understand the data-generating process:** Even if the data technically exists somewhere in a database, take the time to figure out how it got there. Understand if it was filtered, transformed, or otherwise processed before it got to the place from where you will receive it.

  Also, focus on data refresh cycles: how old is the data? When does it get updated? How is it updated? What/who decides when it is updated?

> You could ask: “How is our location data stored? Do we have a separate record for every time a user reviews a restaurant?”

- **Know when additional data collection is necessary:** Sometimes the only way to complete an analysis is to collect more data. If additional data collection isn’t possible, then the scope and goals of the analysis need to be renegotiated with stakeholders.

Image credit: https://www.webalys.com/

PLAN YOUR DATASETS IN ADVANCE

Planning your datasets serves several purposes:

1. It minimizes surprises: It is always easier to plan for contingencies before you begin your analysis than it is to try to adapt in the middle of your work as deadlines approach.
2. It ensures you have all of the support you need: If stakeholders are made aware of problems in the data from the start, they will be more patient and sympathetic when you face delays or unexpected obstacles.
3. It generates good ideas for new datasets: With the dataset in our hand, it is time to do the exploratory data analysis and other explorations.

Get the dataset

Now you have the complete Data Science problem in hand. All you need to do now is run it through the usual ML pipeline.

Just as discussed above, plan out the data sources from where you would collect data. Some of the data that you require would be available, and some would be difficult. Think of doing everything with respect to the data, except that you are not coding in a Jupyter notebook.

- Identify the bare minimum of data that you can collect, along with the sources of how you would collect.
- Identify challenges you can anticipate.
- Look at apps in the same domain of the problem that you are solving and look at the different datapoints that are present.

At the end of this exercise, you need to have all the features of the dataset on paper along with the target metric.

PLAN YOUR METHODS

People create data sets for specific purposes — purposes that people will often have forgotten by the time you come around and want to use the dataset for an analysis.

It's easy to look at a column name and assume the dataset has what you need. Because of that, it's very common for Data Scientists to find out — at least half way into their analysis — that the data they have isn't really the data they need.

Some of those problems manifest themselves only through careful exploratory data analysis. Hence, thorough EDA is essential before applying the methods. This also gives you an opportunity to present a few insights to the stakeholders that might be useful for them.

Consider the following questions:

- For a particular city, give an area-wise breakup of the number of restaurants in each locality.
- For each locality, break up the number of restaurants based on their cuisine, like Indian, Continental, Chinese.
- Visualize the density of restaurants in the localities; whether the restaurants are very close to each other or spread far apart.
- For every locality, give the average spend at restaurants in that locality.
- For every locality, what percentage of restaurants deliver food?
- Break up the data into good restaurants and bad restaurants, and explore further.
- Average cost of good and bad restaurants.
- Are all the good restaurants present in the same locality, or are they spread across different localities?
- Cuisine-wise breakup of good and bad restaurants. Is there a particular cuisine in a city or locality that is not doing so well?

It's unlikely that all of the information needed is stored in one place, already formatted in a way that makes it ready for your investigation and to answer questions. You'll need to bring all the data together, which means you need a plan.

If you are able to answer most of the questions in the EDA phase, and signal the right insights to the stakeholders, that in itself is a huge value-add.

> You could even do some statistical analysis like find the answers to questions like: "Is there a relationship between the rating and the cost of the restaurant?" The question can be answered by a chi-squared statistical test.

You have your business needs. You have your milestones. You have your data. It might seem like there is (finally) nothing left to do but conduct the analysis. But there is still one more step.

Image credit: https://www.webalys.com/

ONE LAST THING BEFORE ANALYSIS

Consider the following questions:

- Which methods are inappropriate for your analysis?
- Of those methods that are appropriate, what are the costs and benefits of using each one?
- If you find a number of methods that are appropriate and have roughly the same costs and benefits (and you probably will), how do you decide how to proceed?

This is the core competency of a Data Scientist: choosing and using analytic techniques to derive value from data.

Here are some ways you could go about planning what methods you will investigate:

- **Identify unsuitable methods first:** Judge whether a black box solution would suffice for the business needs or the model we apply needs to be interpretable to explain the results to the stakeholders.
- **Keep constraints in mind:** If your preferred method requires a GPU, but you don't have easy access to one, then it shouldn't be your preferred method, even if you think it is analytically superior to its alternatives.
  Similarly, some methods simply do not work well for large numbers of features, or only work if you know beforehand how many clusters you want. Save time by thinking about the constraints each method places on your work — because every method carries constraints of some kind.
- **Choose boring technology:** Analytic approaches like deep learning and reinforcement learning are exciting. As a general rule, the more exciting the technology is, the less you should use it.
  Technologies are exciting when they are relatively new, and when technologies are new, they are less stable and harder to support and maintain. A "boring" technology contains far fewer surprises. Look for surprises in your data, not in your technology, and you will tend to build tools that last longer and work better.
- **Be willing to walk away:** Even after you have eliminated unsuitable methods and further narrowed down your list to accommodate your project's constraints, you will still likely have more than one method that could plausibly work for you.

There is no way to know beforehand which of these methods is better — you will have to try as many of them as possible, and try each with as many initializing parameters as possible, to know which performs best.

You will probably run out of time before you run out of models and configurations to try. Don't fall into the trap of thinking you need to ask for more time in order to test everything — set yourself a time limit and go with the best you have at the end of that time.

Planning your methods serves several purposes:

It keeps you from wasting your time on methods that will not ultimately suit your purpose: If a method works beautifully but does not work at scale, and you need it to work at scale, then it is not a good method to choose. If a method can't handle a high number of variables without overfitting, and you have a high number of variables, it is not a good method to choose.
It keeps your mind open to all opportunities: Even the less appealing ones. It is not always particularly fun to implement a simple heuristic or use a model that has been around for decades, but that is often the most appropriate choice for a business.
It keeps your work compatible with the rest of the business: Be a good colleague and think about how your work is going to impact others. Your work shouldn't just accomplish your own commitments to stakeholders. It should make it as easy as possible for others, such as engineers, to accomplish their commitments. Build things in a way that others can use them as easily as possible.

Keeping the above thoughts in mind, and coming back to the case study. We need the solution to be interpretable so that the results can be explained to the stakeholders properly. There are constraints on the deployment costs so the use of GPU must be avoided.

Since the stakeholder, Director of Sales B2B, needs to hit the target of new restaurants, he needs the solution soon. So you require an easy to understand or interpretable method that also consumes less resources. Here, you are going with speed and ease as opposed to the power of the ML algorithm. Of the available choices of methods - **Logistic Regression** and **Random Forests** are decent choices. (A complete and thorough discussion on the intricacies of ML Algorithms are beyond the scope of the book.)

At every step of the analysis, as explained in the beginning, there could be roadblocks that cause us to revisit our assumptions and go back to the beginning. It is a spiral model of development of the solution.

What you have just seen is an illustration of how a business problem is converted into a data science problem and how data analysis is done. The approach and the questions might differ from case to case, but overall these guidelines would get the job done.

Models and analysis

With the dataset identified, you now have the final step — analysis and modelling.

- After arriving at the dataset, think of what data explorations would you do and what questions you will answer.
- Get the list of questions that you will definitely ask during the EDA.
- Looking at the data, think deeply on what ML algorithms are best suited for the analysis. Think of the pros and cons of each algorithm and which algorithm would you like to proceed with for your analysis.

Bonus: Think of the presentation that you would like to make for your stakeholders and what you would communicate with the stakeholders. Think on the lines of what is interesting to them, what would benefit the business objective etc.

The final step: make predictions and fine-tune

Look at the results and fine-tune the model, as well as the business problem if needed, and iteratively go through the steps. As mentioned in the beginning, the complete process represents an iterative spiral instead of a linear path.

To summarize, the entire process of converting a business problem into a data science problem is extremely important. Data science is not just about getting readymade data and applying EDA along with ML algorithms. No company will have data ready for analysis. Identifying the problem, distilling it into a data science solvable version is equally important.

SUMMARY

Let’s do a quick recap of what we have done till now, mapping the complete steps to the entire pipeline that we began with.

Here is a checklist that summarizes the various guidelines and principles for the journey from a business problem to a Data Science problem.

1. Frame the problem

2. Get concrete as fast as you can

3. Focus on consequences

4. Look for opposites

5. Look for hidden problems

6. Understand timing

7. Understand expectations

8. Understand downstream effects

9. Understand when the business problem isn’t a Data Science problem

10. Set the objectives

11. Eliminate possibilities

12. Think about dependencies

13. Group milestone activities by entity

14. Include housekeeping items

15. Think modularly

16. Get external advice

17. Prioritize pain points

18. Think explicitly about trade-offs

19. Figure out the business’s “value” units

20. Subset all metrics

21. Keep it as explainable as possible

22. Prepare the Data

23. Identify all dataset needs ahead of time

24. Differentiate between necessity and sufficiency

25. Understand the data-generating process

26. Know the data refresh cycles

27. Build and Train the model

28. Identify un-suitable methods first

29. Keep constraints in mind

30. Choose boring technology

31. Be willing to walk away

Together, we have now seen the complete journey - right from defining a business problem, to converting it into a Data Science problem and solving it. We began by talking about the art of Data Science and how many aspects of business problem-solving using Data Science are not showcased elsewhere.

You have a step-by-step framework of identifying the business problem and converting it into a Data Science problem. You are armed with this framework to apply these principles in your organisation and execute Data Science projects flawlessly.

A small caveat here: Every business is unique and hence the needs and intricacies of these businesses are unique. As you apply each of the steps in the framework to your business problem, remember to set the right expectations with your stakeholder and check the obtained result at each stage with your expectations. In case of a mismatch, identify the issue (whether it is with the problem definition or the expectations) and fix it before moving to the next step.

The best way to make this a second nature in your work is to pick up different apps and think about the possible problem statements and then apply each step to the problem and convert it into a Data Science problem statement. Validate your approach with anyone who is working in the same domain as the problem you have picked. Remember that companies look for problem solvers and critical thinkers - and this is the first step to become one.

PARTING THOUGHTS

Now that you've reached the end of this book, you have firmly begun your journey towards solving business problems with Data Science. You now know that a Data Science project involves much more than machine learning algorithms and crunching data.

Data Science can just be a few people trying to solve business problems on their own, or it can be a concerted organisational effort, spearheaded by the top brass. It doesn't matter how the effort begins, you will often find that people management and politics will cause greater challenges than technicalities.

You can head off some of these challenges by demonstrating the high value of a Data Science team. Do that by generating momentum, and solid proofs of concept. Here are 4 ways to build a prototype that sells the value of Data Science to the stakeholders:

- **Commitment from the team:** Any project requires dedication to reach its goals. There needs to be considerable investment of time and people in any project to make sure it emerges successfully. Getting feedback from middle management is a good way to assess the general sentiments of the larger team.
- **Concrete value**: There needs to be demonstrable value in any project to justify the resources that are pumped into it. The organisation needs to benefit in clear and measurable ways, by setting up goals and assigning them quantifiable parameters.
- **Be certain of the data**: Before going through the trouble of setting up a prototype, and getting approvals to move ahead, it is important to have the data requirements firmly established. If data needs to be extracted, how many resources will be required, and so on. Just like any project requirements, data must be considered beforehand, so as not to create hurdles midway.
- **Aim for balance**: It is tempting to pick an extremely complex problem to solve at the outset, to demonstrate the value most emphatically. However this is not the way to go. On the other end of the spectrum, picking an easy problem to solve will not showcase value at all. Therefore, it is best to pick a project that hits the right note somewhere in the middle. Look for problems that could be solved using large or disparate datasets. The technical wizardry required here trumps the analytics, but will often yield great insights for the organisation, with lower complexity overall.

As we said in the beginning, Data Science touches every aspect of our daily lives, whether we are cognizant of it or not. From shopping for groceries online, to driving to visit a friend, the way we interact with the world is changing.

We wanted to illustrate the potential of Data ScienceData Science in the chapters above, and how to activate your thinking capabilities by asking insightful questions. We hope you are excited about the possibilities of Data Science after this. Thank you for joining us on this journey. Please do share your thoughts on email at growth@greyatom.com, or reach out to us on Linkedin.

ABOUT GREYATOM

At GreyAtom, our mandate has always been to create a bridge between learners and the industry. In spite of the talent pool that exists in the world today, many jobs lie vacant because learners have gone through theoretical training, and lack the ability to use their knowledge in practical circumstances.

Over the course of **3 years**, over **70,000** learners have taken our programs to achieve the career transitions of their dreams. By creating a curriculum in collaboration with academia and the industry, we have taken the best of both worlds and developed a methodology that creates skilled individuals who the industry wants and needs.

In our endeavor to do so, our **175 hiring** partners have played an immense role. We will include a list of all of them in our acknowledgements, as their insight was invaluable and has contributed to our success in delivering outcomes in a significant way.

For more information, please visit: **www.greyatom.com**