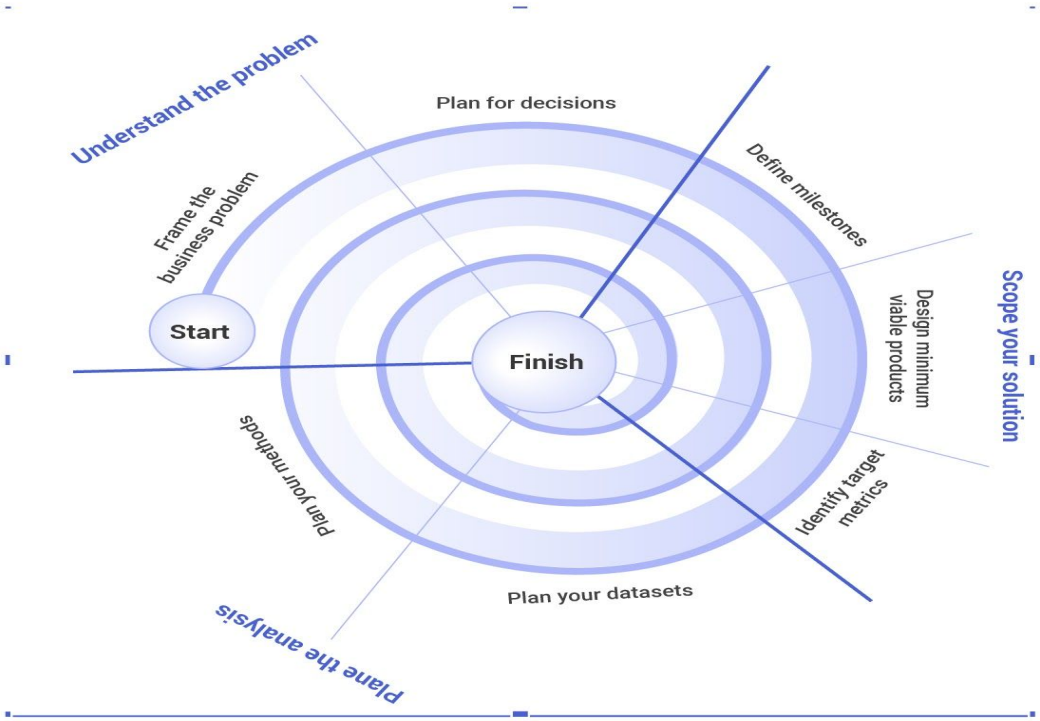
Problem Solving Pipeline

You know the basics of Machine Learning and understand the basic nuts and bolts of the 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, businesses look to data science teams to give insights and help solve problems. The journey from a business problem to a data science problem is not so straightforward, and hence in the next few topics, we will make an attempt to demystify the process. The process of constructing a data science solution to a business problem is often represented as the following path

While we will focus a lot on how to define the problem and setting the objectives in this concept, we will also briefly touch upon the other steps and show how to solve a business problem. The first step of the path - defining the problem - contains tasks such as understanding business needs, scoping a solution, and 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. 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, 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

At this point in your training, you will already know how to select and transform features, train and validate models, and obtain predictions. This module will emphasize the steps a data scientist needs to take to arrive at the point where it even makes sense to think about building a model. A large amount of a data scientist’s work takes place away from the computer: data scientists 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.

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 'to increase the visibility of a tweet and make it viral'

1. Define the problem
    - Predict the number of retweets, likes etc a new tweet would get.
2. Set the objectives
    - What is the ideal metric to chase - retweets/shares? Do we also assign a weighted score if a twitter influencer retweets the tweet?
3. Prepare the data
    - Trends corresponding to the hashtag accompanying the tweet would be a good signal. Identify more features that would influence the number of retweets.
4. Build and train the model
    - How much accurate is the model? How much of an error can the business tolerate?
5. Make predictions and fine-tune
    - The initial predictions were quite off. Think of further refinements. Can this problem be solved? What are the barriers to an effective solution - data? Computation?

A Business Case Study

An international restaurant aggregator company, YumEats! has decided to expand its operations and wished to enter the Indian sub-continent. YumEats! allows the users to select food from restaurants and order to their homes. Like every aggregator in the space, there are mainly 2 channels - B2C and B2B. The B2C revenue comes from a special rewards program or exclusive membership program that would allow the consumers to prescribe to a monthly or annual membership. The B2B revenue comes from collecting a portion of the fee per order collected as commission from the restaurants. Hence, the company has to acquire both consumers as well as other businesses.

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. This would be possible with both more Daily active users on the platform as well as lots of restaurants with a wide variety of cuisines that would further drive traffic to the app. They have already run a whirlwind campaign with deep discounts that are piling the losses and putting a lot of strain. You have been given a mandate - to inform where to cut the spends with minimal damage and also simultaneously find ways to increase the revenue stream with data-driven insights.

The above problem is a business problem that is presented in the rawest form - which is to somehow improve revenue and cut losses by analysing the data. 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.

For example, in the case study just presented, because of the direct revenue impacts and cuts on spending needed, the stakeholder could be a leadership level person on the Marketing Team or Sales Team (CMO or Director, Sales). Instead, if it was a product manager, then the person is more likely to be focused on customer-facing features. A stakeholder who is an engineering manager is more likely to be concerned with the maintainability of a product, and 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 more 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.

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. Here are some of the things to consider:

1. The stakeholders have presented a very vague problem statement. Improving revenue and cutting costs is too wide a problem that is presented. There is a need to narrow this problem.
2. The stakeholders are the ones who will be consuming the insights you present and drive 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.

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 on. For YumEats!, the CxOs 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.

Always remember

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.

Exercise - Identify the stakeholder

As part of the concept, solve the exercises also to get a clear understanding of what you have learnt. This is another business problem that you will solve by yourself based on the principles that are highlighted.

You were recently hired as a data scientist for a travel firm that books flights, hotels, and car rentals (for example: TripAdvisor, Booking.com, Yatra, MakeMyTrip). It became clear on your very first day that the company is very concerned about retention: it saves the company money and effort to have previous customers return for more business than it does to go out and convince new people to try the service, but in a highly-competitive market consumers are used to shopping around and picking whatever deal most appeals to them, regardless of who offers that deal. The problem posed to you is to help in solving the problem of customer retention.

The problem mentioned here is that of customer retention. Before going ahead in refining the problem. First, identify the stakeholders who you will need to talk to. Think which department would be benefited in solving the problem. Then get to understand who is the right stakeholder. Who do you think is the right stakeholder?

Frame the Business Problem

For the broad questions that were mentioned previously, there could be a lot of sub-problems that could be solved. Talking to the stakeholders and brainstorming with them has led to these possible problems that can be solved.

- Increase in Revenue
    - Conduct campaigns to acquire new users, understand user profiles to deliver customized campaigns.
    - Predict when the users are likely to order next from the different trends and nudge them to order.
    - Recommend restaurants to the users to order from next, based on the previous orders of the users.
    - Perform existing customer segmentation and run a tiered campaign - budget restaurant recommendations for low paying users, premium restaurants from high paying users.
    - Identify good restaurants that are not yet onboarded on the app to increase the roster.
    - Provide recommendations to new onboarded restaurants on cuisine preferences of the people in the locality and give on-the-fly menu suggestions based on trends.
    - Automatically check the image qualities of restaurant and food photos and enhance the low-quality photos to drive traffic.
- Reduce costs
    - Understand which customers are likely to leave the platform and only provide discounts to retain them. (Only provide discounts if someone is at risk to leave the platform)
    - Automatically identify orders from nearby localities and assign it to the same delivery person.
    - Automatically identify fraudulent orders beforehand, and trigger orders only upon verification.

These problems have further narrowed the scope of the different kinds of the problems that can be solved here. Here is one thing to avoid while talking to a stakeholder. As a data scientist, you use a specialized 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.

Similarly, people who work on marketing, product management, or an executive team 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. In the case study, some good follow-up questions that could lead to the above sub-problems could be:

- “You’ve talked about increasing revenue. What were some of the initiatives for increasing revenue? What are the major pain-points that you experienced that led to the loss of revenue?
- “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 that have left the platform. What are the 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.

Here are some guidelines for asking 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 user wants”, 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. For example, you might ask, “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 the pain points and if the problem you are thinking can alleviate the pain points. Prioritize the pain points to in turn prioritize the data science problems that might ease out these pain points.
- Look for opposites. If we 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. An important thing that you need to be aware of is also to identify possible biases - human bias usually gets reflected into algorithms and might be harmful.
- 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. For example, the stakeholder pointed out that the users spend quite a bit of time on 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 soon?”

Asking clarifying questions serves several purposes:

1. It demonstrates that what is important to your stakeholders is important to you, too. It builds trust and establishes rapport, which are two things you will need when it comes time to share the results of your work.
2. If 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.
3. 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.

Exercise - Refining the problem

From the previous exercise, you would have identified the stakeholder for your problem. As next steps, 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.

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. 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:

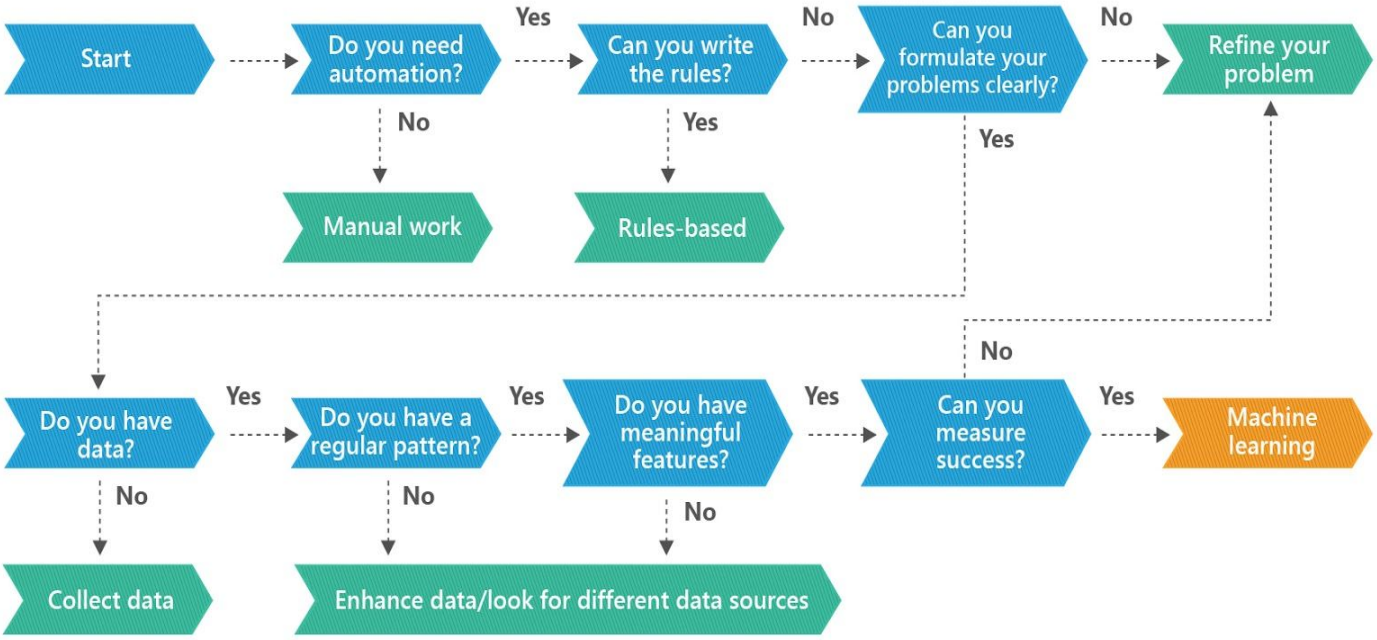
- “Which users must be given the discounts to stay back on the app and when to trigger them?”
- “To a new user who has just landed on the app, what is the right campaign to show?”
- “For a new restaurant onboarded on the app, which dishes must be highlighted on the menu?”

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 the 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.

Here are some guidelines for mapping out the relevant decisions, which your stakeholder has just more alluded to 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. For example, you could ask, “When should a particular campaign be shown for a user to ensure maximum conversion?”
- Understand expectations. Set the expectations clear 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. For example, discounts might make the user happy but the restaurant partner might be worried about the impact on the offline business. The respective B2B sales director must be aware of the impact.
- Understand when the business problem isn’t a data science problem. The most important thing to realize 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. You can use the checklist to decide:



Most importantly analyze among all the possible data science problems to solve what would be the ideal sweet spot which would lead to

- Quantifiable impact for users - increase in daily active users, increase in the daily orders
- Quantifiable impact for stakeholders - increase in revenue with lesser costs

By now it must have become 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 which would incur further cost. After a lot of deliberation with the stakeholders, you now get to the actual business problem that you like to solve

Identify the good restaurants to be targeted for onboarding onto the app.

With this problem, you are trying to improve the choices of good restaurants to 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.

Exercise - Identify ONE problem to solve.

From the different subproblems that you are 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 Introduction to Data Science…). Next, think of what decision can the stakeholders take on solving the shortlisted subproblems. (Remember that we are focused more on decisions and less on findings). Further, analyze what could be the impact of solving each of the shortlisted subproblems. Discuss and arrive on ONE problem statement that you would like to tackle.

Identifying the milestones

Let us revisit the case study in terms of the problem we just framed in the previous chapter.

In consultation with the management of YumEats!, we had identified the business problem to be solved - identify good restaurants to be onboarded for the app. The stakeholders for this business problem would be the Director of Sales, B2B. The impact would be

- *More active restaurant partners with high rating would improve the brand, drive more active users and increase daily orders.*
- *More restaurants onboarded brings greater choice and more users to the app.*
- *If famous yet not so visible restaurants are on the app, it would lead to more value proposition to the users.*

The business problem is now more or less clearly defined and framed. This is the first step to go towards the data science problem. 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 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.

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 re-frame "identify the right restaurants to onboard on the app" 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 buckets of restaurants - excellent, decent, bad?
- How do we estimate the popularity of the restaurants?
- What are the features that differentiate the good and 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.

For example, given the case study above, you have a reasonably clear analytic goal: identify good restaurants that are likely to improve the brand. Here are some guidelines for creating good analytic milestones for the project:

- Eliminate possibilities. It’s easy to jump into using all available data to try to solve a problem, but it is often wiser to think of important data points - for the analysis. Further, you can bucket them as Critical, Good to Have, etc to prioritize which features need to be collected first. This kind of feature and data selection does not require any particular method - it relies upon domain knowledge, which is something you can get from your stakeholders. For example, a critical feature for the restaurant is location, year of establishment. Good to Have is average no of daily visitors to the restaurant.
- 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. For example, 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.
- Group milestone activities by 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. For example, 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 them both.

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 next week, that 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.

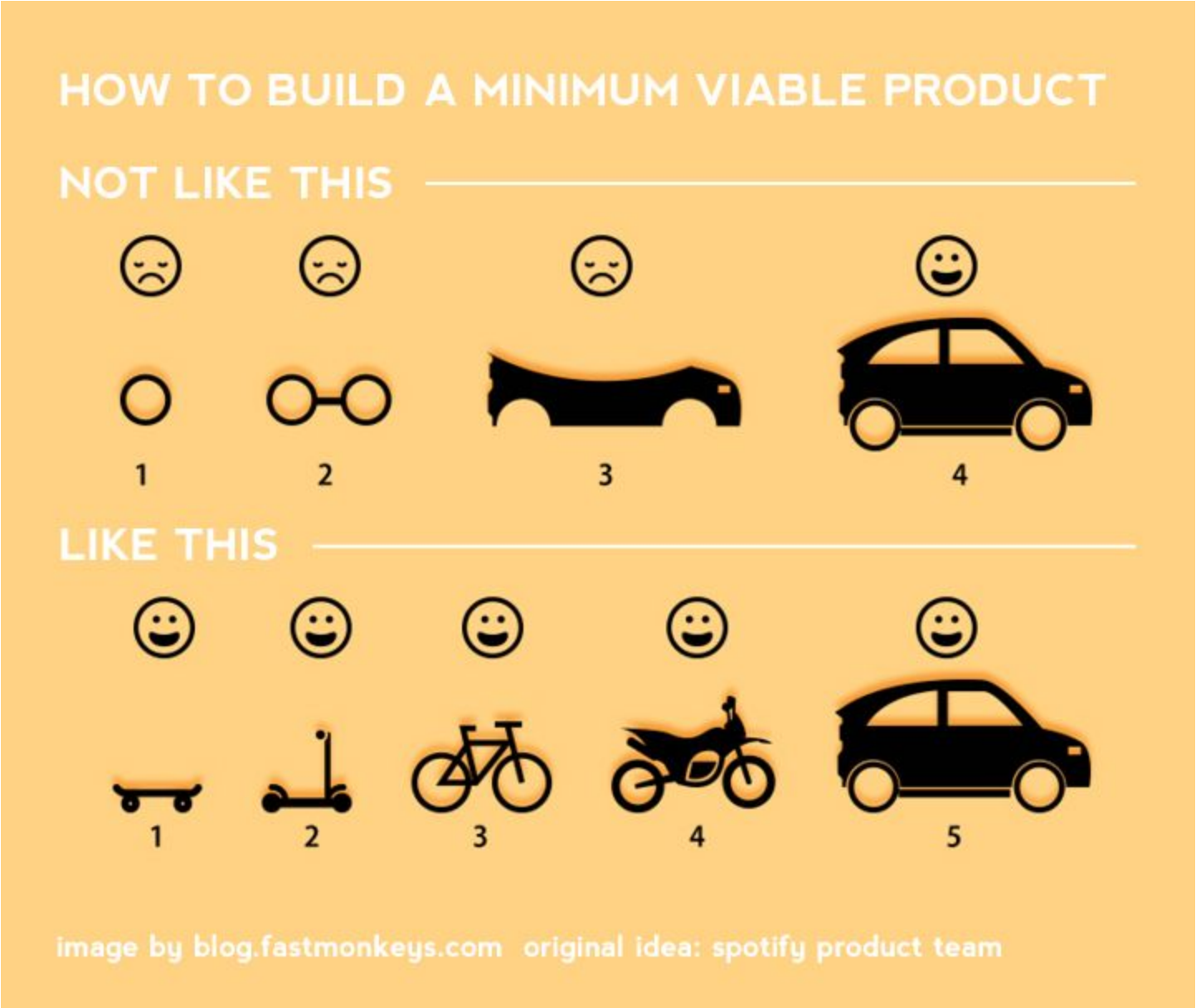
Now with the milestones identified, the next step is to build out a minimum viable product.

Exercise - Milestones for the problem

By now, you must have identified the problem you wish to solve which helps in achieving customer retention on the travel management app. If you haven't yet identified the problem, go back and formulate the problem you wish to solve using the steps in the previous chapter. Now with the problem identified, 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.

Design minimum viable products



To use the imagery from the above graphic, 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 results by? Do they need all the results at once, or do some results have a more pressing deadline than others?
- What is the simplest way to meet a benchmark, regardless of whether you consider it the "best" way?

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. The typical journey of a data science product is

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 some 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. For example: 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. For example: 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 (though might not be accurate).

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 the 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 your work will be outdated before it is even deployed.

Exercise - Identify the minimum viable product you will build

In the previous exercise, you have 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 data of apps. On a similar line, 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?

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. One obvious success metric is the actual business outcome stakeholders want to achieve: if they use your data science solution and onboard the suggested restaurants 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 in the prediction of the restaurants? Can they go blindly with the suggestion or some other checks are 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 suggest 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, 'goodness' of the restaurant is subjective, and to define the metric it must be objective. An objective measure of the goodness 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 in terms of a data science problem, it can be - predict the ratings of a restaurant to decide if they can be onboarded on the app.

Now let’s skip ahead in our thinking to consider what proof of the value we want or need to be able to deliver to the stakeholders in our case study. Here are some guidelines for selecting good metrics:

- 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. For example: Out of a rating of 5, consider ones with rating 4 and 5 as good restaurants and rest as bad restaurants. Which is more harmful - identifying good restaurant as bad, or identifying bad restaurant as good. The stakeholders are more 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: restaurant categories, cuisine, regions, etc. 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.
- 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. For example: assess the comfort level of stakeholders as it regards technical metrics. Consider re-framing technical concepts such as “false-positive rate”, and “false-negative rate” as “wrongly identified restaurants” and “missed opportunities”.

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 the 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 the rating. 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.

Exercise - 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 if the problem is going to be a regression problem or a classification problem).

Prepare the Data

To recap,

Business Problem Statement: Find good restaurants in the specified locality so that they can be onboarded on the app.

Business Impact: More good restaurants on the app, improve the brand value and drives traffic to the app further leading to increase in Daily Active Users, the daily value of orders and impacting revenue.

Data Science Problem statement: Predict the rating bucket of a given restaurant (good or bad) to decide if the restaurant can be onboarded to the app.

Data Science Metric: Precision

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.

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 your 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 the 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, 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. Do 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 actually 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. Data is the key for success or failure for the data science project.

The first step of the analysis is to collect the data. When it comes to datasets, trust no one unless they themselves created the dataset - and even then, be careful.

For our case study, our data sources could be:

- Scrape the data from other restaurant search and discovery sites like Zomato etc where a lot of information about the restaurants are present. (These would have the rating information also.)
- Explore trending venues in particular neighbourhoods especially restaurants - (using FourSquared API)

In addition, we can even scrape twitter to obtain the conversations or tweets regarding the restaurants of interest. These can be an additional input for the data.

Now the data that comes from Zomato would be a table that could contain the restaurant name, cuisine, location, est delivery time, min order value, a few photos, reviews of the restaurant, recent streak of the reviews etc. 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. And there could be even additional and extra information. For example, there could be a delivery facility for a particular restaurant by Swiggy but the facility is not available on Zomato. Now imagine, getting these disparate disjoint into a single table on which further analysis could be done. And if you plan to do this at scale, you need to get the engineering team also involved in designing a complete data pipeline that can store the huge 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 got the data you needed. Now since the data is collected through a third-party source or API, there could be lot of inconsistencies in the data and cleaning and transforming this data becomes really crucial. For example, estimated delivery time is returned as a string. But any valuable insights or analysis could only be possible if it is stored in a time format. So that semantic parsing must also 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, more such challenges would arise that need to be overcome.

Let's say you have persisted and got the data to the format you wanted to. Now, let us talk about the target variable. The target variable for the 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 say 1 to 10. 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 the various data quality issues. Any data corruption that can be identified? Incorrect data values? For example, if say delivery time is shown in negative values, then you know that the data is incorrect for time cannot be negative.
- Resolve missing entries, inconsistencies in the data and any semantic errors like wrongly labelled columns.
- Extract new features from existing features or identify new ones.
- Encode the data i.e convert non-numeric to numeric data.

Stakeholders often ask for things they can’t have. It is especially common for stakeholders to want answers to questions at the same time they lack the data needed to provide those answers. For example, 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. Sometimes you won’t even realize that a crucial data point is missing until you are in the thick of your analysis.

Now 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 in the direction of regression, it 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, know for sure that you have done aggregation of data sources. Now a basic dataset is ready and 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. For example, you could say: “At the very least, I need the location data, cost for 2, cuisine and reviews to start with.”
- 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. For example, 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 any ways it was filtered, transformed, or otherwise processed before it got to the place 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? For example, you could ask: “How is our location data stored? Do we have a separate record for every time a user gives a new review of the 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

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.

Exercise - 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 w.r.t data just that you are not coding in the jupyter notebook. Identify the bare minimum data that you can collect along with the sources of how you would collect. Identify the challenges you can anticipate. Look at the 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.

Build the model

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 halfway 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 a thorough EDA is essential before applying the methods. This also gives an opportunity to present to the stakeholders a few insights that might be useful for them.

Consider the following questions:

- For a particular city give a locality wise breakup of the number of restaurants in each locality.
- For each locality, break up the number of restaurants based on their cuisine i.e Indian, Continental, Chinese.
- Visualize the density of restaurants in the localities, whether the restaurants are very close or spread far apart.
- For every locality, give the average cost of restaurants in that locality.
- For every locality, what are the percentages of the restaurants that deliver food?
- Break up the data into good restaurants and bad restaurants and explore further
    - Avg cost of good and bad restaurants
    - Locality wise breakup
    - 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 answers the questions. You'll need to bring all of that data together, which means you need a plan.

If you are able to answer most of the questions in the EDA phase and identify the right insights to the stakeholders that is itself a huge value add. You could even do some statistical analysis like finding the answers to questions like:

Is there any relationship between the rating and the cost of the restaurant?

The above 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.

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 to 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. Given the problem at hand, here are some ways you could go about planning what methods you will investigate:

- Identify un-suitable 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 a GPU, 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 much 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 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 what 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:

1. 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.

2. It keeps your mind open to all opportunities - even the less appealing ones. It's often not 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.
3. 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. Consider all these, the logistic regression or random forests might be the right choices for modelling.

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 the data analysis is done. The approach and the questions might differ from case to case, but overall these guidelines would get the job done.

Exercise - 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 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.

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.

Define the Problem

- Get concrete as fast as you can
- Focus on the consequences
- Look for opposites
- Look for hidden problems
- Understand timing
- Understand expectations
- Understand downstream effects
- Understand when the business problem isn't a data science problem

Set the objectives

- Eliminate possibilities
- Think about dependencies
- Group milestone activities by entity
- Include housekeeping items
- Think modularly
- Get external advice
- Prioritize pain points
- Think explicitly about trade-offs
- Figure out the business's "value" units
- Subset all metrics
- Keep it as explainable as possible

Prepare the Data

- Identify all dataset needs ahead of time
- Differentiate between necessity and sufficiency.
- Understand the data-generating process
- Know the data refresh cycles

Build and Train the model

- Identify un-suitable methods first
- Keep constraints in mind
- Choose boring technology
- Be willing to walk away

The final step: Make predictions and fine-tune is about looking 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 to a data science problem is extremely important. Data science is not just about getting a 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. At the end of the day, data science is more about problem-solving and structured thinking that just chasing a metric.

How to tackle a Case Study?

A case study interview tests your:

- Technical Skills
- Decision-making skills
- Problem Solving Approach
- Thinking Process

How you use these techniques to solve any business problems?

A framework to tackle case studies is ASPER. It is a wonderful framework technique used to breakdown almost any case study.

Ask - Ask questions to uncover details that were kept hidden by the interviewer. Specifically, you want to answer the following questions: “what are the product requirements and evaluation metrics?”, “what data do I have access to?”, ”how much time and computational resources do I have to run experiments?” etc.

Suppose - Make justified assumptions to simplify the problem. Examples of assumptions are: “we are in small data regime”, “events are independent”, “the statistical significance level is 5%”, “the data distribution won’t change over time”, “we have three weeks”, etc.

Plan - Break down the problem into tasks. A common task sequence in the data science case study interview is: (i) data engineering, (ii) modeling, and (iii) business analysis.

Execution - Announce your plan, and tackle the tasks one by one. In this step, the interviewer might ask you to write code or explain the maths behind your proposed method. (OPTIONAL)

Recap - At the end of the interview, summarize your answer and mention the tools and frameworks you would use to perform the work. It is also a good time to express your ideas on how the problem can be extended.

A Few tips

- Articulate your thoughts in a compelling narrative.
- Tie your task to the business logic
- Brush up your data science foundations before the interview.
- Study topics relevant to the company.
- Listen to the hints given by your interviewer.
- When you are not sure of your answer, be honest and say so.
- When out of ideas or stuck, think out loud rather than staying silent.

Success Rate of LinkedIn Story Feature

Suppose you're working as a data scientist at Linkedin. How would you measure the success rate of the Linkedin story features?

- Points of Approach
    - Stories are temporary. Active for 24 hrs and then gone.
    - If you don't know how LinkedIn stories work, then go for the closest product you know like Instagram stories.
    - Different ways I interact with the stories
        - Click and view the post.
        - Click the link in the post.
        - Click the profile after viewing the post.
        - Reply to the post.
        - Unfollow… Report content
        - Forwarded or Reshared
        - Closed the story
    - Ask
        - A question to LinkedIn - what is more important - the CTA or just interaction with the post. 2 side engagement - created the story… or interacted with the story. Should I treat these 2 metrics separately?
        - What is the goal with this feature - Is it to make users spend more time on the LinkedIn app or website? Or is it to drive engagement between the users? Or is it another way to serve ads eventually?
    - Suppose
        - Assumption - LinkedIn is more interested in engagement.
        - Assumption - CTAs or generate additional interest - either click the link or click the profile of the person. Forwarded is a strong signal. Closed the story is a strong negative signal. Creation is also a good signal.
        - Assumption - out of 1000 people, ~10 people like, ~4 people comment, ~2 people forward, ~5 people close the story, ~1 person reports the story
    - Plan breakdown
        - Total # of linked in stories created - baseline.
        - Total # of people to whom the LinkedIn story is shown
        - Total # of people who like the story - Positive weights
        - Total # of people who commented on the story - Slightly higher positive weight
        - Total # of people who forwarded the story - higher intent, so higher weight.
        - Total # of people who reported the story - negative intent, so negative weight.
        - Take a weighted sum of all the metrics to define an engagement metric.
    - Execute and Recap
        - A simple dashboard to show a weekly tracking of various metrics can be set up using pandas and plotly.
        - Eventually, an ML model to predict future engagement

The idea here is not to be right/wrong here but to wear a thinking hat and analytically approach the problem.

What are the different ways you can reduce fraud and malicious claims leveraging data?

As a data scientist with a big insurance firm, what are the different ways you can reduce fraud and malicious claims leveraging data?

- Insurance companies are striving for a technologically advanced system that helps keep all their employees synchronized. These employees vary from agents, brokers, claim investigators to market and support team.
- Fraud detection is a set of activities undertaken to prevent money or property from being obtained through false pretenses. Fraud detection is applied to many industries such as banking or insurance. In banking, fraud may include forging checks or using stolen credit cards.
- ASK, SUPPOSE steps were used to ask questions regarding the problem statement helps us understand the key data. Knowing the nature of the fraud is more important as well as the time period of the data.
- Location, time period, the behavior of consumers, and other factors are important for the plan since they help us to find the correlations between the attributes.
- Planning the data and data points is an important step.
- Identifying the data and extracting patterns from the data will help us detect the fraud. All this goes into the plan with a lot of research and analysis.
- Then comes the final recap part where the narrative approach was used to tie the business statement to our solution. Explaining with the clear institution and field knowledge is key here.

Pharmaceutical Use Case

Business Problem Statement

The pharmaceutical industry plays a very crucial role in the healthcare sector. Their revenue is generated through the drugs sold to patients (at a pharmacy) or to the hospital (through a distributor). It is through marketing that these companies influence the healthcare providers/patients to prescribe/want their drugs. The bottom line is to get new products and services to the market through apt research and development.

Our client Pharma Art is one of the established names in the industry and provided almost 10% of the generic drugs consumed in the last decade. The total volume of medicines consumed globally will increase by about 3% annually through 2021. On the contrary, reports indicate that the market share of Pharma Art will reduce due to the increasing flow of foreign counterfeit drugs in the market, expansion in the self-medication sector, competition, and several other factors. As a result, the overall annual growth of Pharma Art is predicted to slow to single digits, between 6% and 9%, through 2021, which is already down from a 12% growth in 2015. Looking at these figures, the company wants to utilize all its resources in order to maximize the sale of their drugs.

Since the cost of developing a new drug is too high, Pharma Art wants to optimize all the marketing campaigns to target the correct audience in order to maximize the sale of the existing drugs. Now the main challenge here is defining the scope of the customer. Of course, the ultimate customer here is a patient, but there are numerous indirect entities through which the product reaches the patient. These entities, in turn, become the primary customer, like a healthcare professional or a physician prescribing the drug. Thus the business problem posed to the Data Science team at Pharma Art was - Determine which consumers and physicians are most likely to utilize any drug and create more targeted on-the-ground marketing efforts.

Since the marketing team would be most affected by this problem, the Head of Marketing would be the stakeholder. The stakeholders have dedicated Medical Representatives for defined territories, and they form the entire backbone of the marketing effort. They form the link between the primary customers and the company.

Framing the Data Science Problem

The Reps need to be prepared to present impactful content and make the best use of every opportunity in less time. Since now we know that we have to influence the providers and distributors to prescribe our drug, our first step would be to analyze the current sales rates and market conditions. This would also be our first delivery to the stakeholder.

Sales statistics within territories were determined and presented to the stakeholder. Using these statistics, one single objective was to identify doctors/distributors more interested in making time for reps who would have more treatments to discuss. This would help Pharma Art convert new sales opportunities from existing customers.

The reps need to identify prospective business opportunities for the companies and persuade them to purchase the company medication. Now the next question is, what is going to make more impact - reaching out to consumers who are more likely to utilize our drug in less time or spend more time with the consumer and try hard to sell maximum products? Since the reps have restricted time with distributors/physicians, reaching out to more interested consumers and physicians would help them maximize their effectiveness in every customer meeting.

Thus the problem is to identify consumers and physicians who are likely to utilize a drug so that the reps can then tailor their agenda to suit their requirements. Currently, the identification of the consumers/physicians is still a subjective parameter. So the problem was reframed to develop a model with the outcome variable being a quantitative value related to Rx records (prescription records) of different drugs. The final Data Science problem, thus, was to predict whether a consumer/physician would prescribe to a specific drug or not.

Clearly, our Data Science problem is to identify maximum consumers/physicians who will prescribe to our drug so that the reps can convert most of their customer meetings. The metric of interest would be Recall. Now since we have a clear Data Science problem along with the evaluation metric, let us explore how the problem was solved using Machine Learning.

Under the hood ML

To begin with, Pharma Art already has some unstructured and semi-structured data that was abstracted on a structured pro forma for all patients who visited the several physicians/healthcare institutes. The first step would be to get the data in a structured format and do the analysis. Since this data alone would not be sufficient, more data was collected from the Point-of-Sale system of several pharmacies.

Aggregating these data, daily sales of drugs in different categories over a period of 5 years was consolidated. Again remember that most of the data science project cycle is spent in this stage. The final data used in doctor's medicine prescription pattern learning had the following features:

- Details of the Physician (Gender, Location, Credentials, Speciality)
- Details of the Patient (Gender, Education, Marriage, Occupation)
- Physical Examination (Height, Weight, BMI, Blood Pressure)
- Exercise and fitness level (Fitbit data, fruits, and veggie consumption, etc.)
- Habits (hours of TV watching, soft drink consumption, smoking, e-cigarette, binge drinking, etc.)
- Use of digital devices (computer, internet, etc.)
- Pollution (SO4, SO2, NO3, HNO3, NH4, Mg, Na, Ca, K, Cl, etc.)
- Climate (UB Exposure, Precipitation)
- Insurance coverage (e.g., Medicare Enrollment)
- Road Traffic and Commuter Stress index
- Laboratory Tests Results - Continuous/Categorical
- Category of Drug
- Recent Prescriptions - Boolean (as per the category of the drug understudy)
- Disease History - Boolean (as per the category of the drug understudy)
- Leading indicators of disease

Target Variable: Prescribed - Yes or No.

- While exploring the data,
    - prescriptions of a particular medicine were marked as increasing or decreasing on a monthly basis. This helped the stakeholders find previously not known patterns (like 75% of Physicians who accepted Pharma Art and were located in a metropolitan city wrote 25% more Rx of Medicine A this month compared to the prior month).
    - statistical analysis was conducted to evaluate the importance of each disease in disease history to the prediction. The importance is assessed for each disease in the prescription of each drug, were having this disease or not is counted towards having this drug prescribed or not.
    - drug dependencies were identified. For example, for diabetic patient prescriptions, around 13% cases have more than one drug prescribed. Many drug pairs show dependency, suggesting correlation between the drugs.
    - region wise tendency of the physicians to prescribe drugs was identified which showed where the demand for the product is highest and where it is lowest and then made a plan to tackle it.
    - segmentation of territories based on physicians' therapeutic tastes, geographic trends and peak prescription rates. These helped in changing marketing tactics for each physician.

These were the initial insights from the data. To interpret these results to our stakeholders, we want to identify the probability that the consumer and physician are likely to utilize the drug. Since we are trying to forecast a future outcome (whether the consumer would utilize our drug) which is binary in nature, along with a probability associated with it, the need was to create a binary classification model. Assuming that the features are independent Decision Trees seems to be a good option here. When evaluating the model, k-fold cross-validation was also implemented. Another main reason to choose this algorithm was that it will give a visual and explicit representation of the decision to our reps, and they can comprehend it easily. Accordingly they can tailor their agenda to suit the consumer's requirement and maximize the sale of the drugs.

Results and Impact

Impact to the Organization - The greatest impact this project had on the organization was to slash down the time consumed by the reps to sell the drugs. They were able to train the Reps in a better fashion so that they can effectively prepare the pre-call before any client visit.The drug reps were provided with mobile devices and real-time analytics on their prospects, which would help them reach out to the correct lot and approach them according to their needs. In a span of three months, the average time spent per call was reduced to 60%.

Impact to the Stakeholders - Machine Learning model specifically identified Prescribers that have high potential and high propensity to write materially more drug prescriptions than they do presently. For these prescribers, the stakeholders decided to distribute free drug samples depending on their therapeutic tastes and prescription pattern. This would help the company to convert new sales from their existing customers.

Better growth in domestic sales would also depend on the ability of companies to align their product portfolio towards chronic therapies for diseases such as cardiovascular, anti-diabetes, antidepressants and anti-cancers that are on the rise. This area can be explored next.
References

- Workera Data Science Case study interview: https://workera.ai/resources/data-science-case-study-interview/
- Prepare for Business Case Study Interview : https://towardsdatascience.com/how-to-prepare-for-business-case-interview-as-an-analyst-6e9d68ce2fd8
- Acing the Data Science Case Study Interview : https://www.interviewquery.com/blog-data-science-case-study-interview