Activity

Problem Formulation: What this pipeline phase entails and why it is important

Introduction

The problem formulation phase of the ML Pipeline is critical, and it is where everything begins. Typically, this phase is kicked off with a question of some kind. Examples of these questions include: What additional product should we offer someone as they checkout? How much storage will clients need from a data center at a given time?

The problem formulation phase starts by seeing a problem and thinking, "what question if I could answer it, would provide the most value to my business?" If I knew the next product a customer was going to buy, is that most valuable? If I knew what was going to be popular over the holidays, is that most valuable? If I better understood who my customers are, is that most valuable?

However, some problems are not so obvious. When sales drop, new competitors emerge, or there is a big change to a company/team/org, it can be easy to say, "I see the problem!" However, sometimes the problem is not so clear.

Part of the problem formulation phase includes seeing where there are opportunities to use machine learning.

Business Scenarios

In the following practice examples, you are presented with four different business scenarios. For each scenario, consider the following questions:

1. Is machine learning appropriate for this problem, and why or why not?
2. What is the ML problem, if there is one, and what would a success metric look like?
3. What kind of ML problem is this?
4. Is the data appropriate?

The first scenario has been completed for you. Remember that there are two ways to start an ML problem. The first is by addressing an obvious problem, the second is by seeing an opportunity. Lastly, be sure to consider whether this is even an ML problem at all. Take a look at scenarios 2 – 4 (three remaining scearios) below and see if you can answer the questions above.

Scenario 1

1) XYZ eCommerce recently began advertising to its customers when they visit the company website. The Director in charge of the initiative wants the advertisements to be as tailored to the customer as possible. You will have access to all the data from the retail webpage, as well as all the customer data.

    1) ML is appropriate because of the scale, variety, and speed required. There are potentially thousands of ads and millions of customers that need to be served customized ads immediately as they arrive at the site.
    2) The problem is that ads that are not useful to customers are a wasted opportunity and a nuisance to customers, yet not serving ads at all is a wasted opportunity. So how does XYZ eCommerce serve the most relevant advertisements to its retail customers?
        i. Success would be the purchase of a product that was advertised.
    3) This is a supervised learning problem because we have a labeled data point, our success metric, which is the purchase of a product.
    4) This data is appropriate because it is both the retail webpage data as well as the customer data.

Scenario 2

2) You're a Senior Business Analyst at a social media company that focuses on streaming. Streamers* use a combination of hashtags and predefined categories to be discoverable by your platform's consumers. You ran an analysis on unique streamer counts by hashtags and categories over the last month and found that out of tens of thousands of streamers, almost all use only 40 hashtags and 10 categories despite innumerable hashtags and hundreds of categories. You presume the predefined categories do not represent all the possibilities very well and that streamers are simply picking the closest fit. You figure there are likely many categories and groupings of streamers that are not accounted for. So you collect a dataset that consists of all streamer profile descriptions (all text), all the historical chat information for each streamer, and all their videos that have been streamed.
    1)
    2)
    3)
    4)

*[**Additional notes**: An online streamer or live streamer is a person who broadcasts themselves online through a live stream or prerecorded video. Genres include playing video games, tutorials, or solo chats. Streamers make money live streaming on platforms like Twitch, TikTok, YouTube, and Instagram. These social media platforms make money through ads.*

A hashtag is a word or phrase preceded by the pound symbol. On social media, it serves as an indication (for users and algorithms) that a piece of content relates to a specific topic or belongs to a category. Hashtags help make content discoverable in on-platform searches and, effectively, reach more people.]

Scenario 3

3) You are a headphone manufacturer who sells directly to big and small electronic stores. As an attempt to increase competitive pricing, Store 1 and Store 2 decided to put together the pricing details for all headphone manufacturers and their products (about 350 products) and conduct daily releases of the data. You will have all the specs from each manufacturer and their product's pricing. Your sales have recently been dropping, so your first concern is whether there are competing products that are priced lower than your flagship product.
    1)
    2)
    3)
    4)

Scenario 4

4) You're a Senior Product Manager at a leading ridesharing company. You did some market research, collected customer feedback, and discovered that both customers and drivers are not happy with an app feature. This feature allows customers to place a pin exactly where they want to be picked up. The customers say drivers rarely stop at the pin location. Drivers say customers most often put the pin in a place they cannot stop. Your company has a relationship with the most used maps app for the driver's navigation, so you leverage this existing relationship to get direct, backend access to their data. This includes latitude and longitude, visual photos of each lat/long, traffic delay details, and regulation data if available (ie, No Parking zones, bus-lanes operation zones, fire hydrants, etc.).
    1)
    2)
    3)
    4)

--- End of Activity ---