Problem 1:

' ABC' is an online multiplayer real time cash game, a product of the company 'XYZ'. Thousands of users play daily and win real cash. Very small percentage of the amount is taken from the users as a charge for the platform which the company is providing to users. The amount taken from users is the revenue for the company. Users are segmented as 'High Value', 'Mid Value' and 'Low Value' based on the revenue generated by users in a month.

Among high value users, on an average 5000 users play on a daily basis. But, among 5000 users who played on a day, x1% users become inactive i.e. do not play for next day, x2% users become inactive for next consecutive 2 days, x3% users become inactive for next consecutive 3 days, ……
,xn% users become inactive for next consecutive n days and there are users are inactive till now. This happens on a daily basis and the percentage of users becoming inactive for next 1, 2, 3, …. ,n consecutive days almost remains the same.

The problem gets exacerbated by the event calendar pertaining to the game play. For example, if 'XYZ' is an online Fantasy platform, then the active days of the users would have to be based on the sporting event calendar, e.g., for cricket Fantasy, this would mean the days where users can make teams to enter a contest would be aligned with the days when real cricket matches are scheduled. As an extension of the example pertaining to cricket, this would mean during the IPL days for couple of months, this will be daily and on certain days multiple contests (because of multiple matches happening). However, for the other periods this may be far and few.

Like the user categories, in case of Fantasy sports, there would be certain distinction of the sporting events depending on the interest levels, which might also vary depending on the user preference. For example, in cricket, a high value users may be very interested in IPL and India matches, while another one might be interested in all the cricket matches across the year. Similarly, a low value user might be interested in all the matches or otherwise. So the definition of "consecutive" in the previous paragraph might be very tricky.

Another dimension of user affinity beyond the interest levels in the matches remain in the interest of various different types of contests offered by the platform. Some contests are low entry but high return with millions of users participating while only a few winning with skewed distribution of prizes (i.e., first position getting very high prize, second position getting a exponentially lower prize but still net positive, and so on. The liquidity and the resultant prize pool of such contests can be a function of the mass interest in a particular sporting event and hence the affinity of a user can be direct function of the prize pool. Some other contests can have a high entry but low return with higher chance of getting a winning position (e.g., 1:1 head-to-head contests). Users' affinity to contest might remain constant of change across events as well.

Now, the Marketing Team of 'XYZ' would like to identify various users on the day of their playing, whether they are going to remain inactive for n days. If the company can identify such users who are going to be inactive, then it can give some promotional offers on that day for the appropriate contest(s) with the expectation that the users would not become inactive after having the offer. But the Marketing Team is unable to find out answers to the following questions:

1. The Utopian strategy would be to identify users who become inactive for even one day for most popular contests. But this strategy would incur huge cost because:
    i. Percentage of such players would be very high
    ii. Among them a large percentage would anyways become active without any promos. So, giving promos to such users would not generate any business value

iii. And finally, for some contests the burn on the user by the platform can be significantly high depending on the prize pool offered

Similar to the case for “consecutive” 2, 3, or relatively less days of inactivity.

But if inactivity reaches beyond a certain number of “consecutive” days, then it would be a huge revenue loss for the company as these are high value players.

Now the challenge is to identify the optimal value of ‘n’.

2. How to identify users who are active on a given day but would remain inactive for next “consecutive” n days.

Assume that you are a ‘Data Scientist’ in ‘XYZ’ and the Marketing Team expects following solution from you:

1. Come up with data driven methodology to find out the value of ‘n’, the definition of “consecutive”, the potential action (e.g., in terms of contests) in a combined manner.
2. Assume that you come up with a value of ‘n’ and you find that only 1% players become inactive for “consecutive” n days along with the pertinent contests.
   Propose comprehensive data science approach to identify (predict) users who are active on a given day but would become inactive for next consecutive n days. It is to be noted that the approach should include all steps that are necessary in typical data science project lifecycle from **problem identification** to **model deployment**.

Also it would be preferable if you can provide some strategies to plan the project and interact with the marketing and product teams so as to meet their time-bound needs (often driven by the event calendar as mentioned above) and quality of the data science and AI solutioning.

Problem 2:

'XYZ' is a Multiplayer Online Gaming company. In the process of business operations, a humongous amount of data gets generated. The company leverages the power of data for business growth through various process optimization and automation, user journey personalisation, game dynamics modeling, etc. But now the company has realized limitations on the power of using data to drive business activities. The major identified limitations are as follows:

- No control over the external environment. Also, the company does not have any process in place to parameterize the external factors which can be fed into the existing models. Some of the categories of external factors could be: Sporting event calendar (e.g., Cricket schedule), Natural Disasters, Rarest of rare (pandemic), Cyclical, Seasonal, etc.
- The data gets generated as a result of micro level business activities which are very internal to the company. Hence, all the data driven processes are subjected to selection bias.
- Short term success does not always guarantee its sustainability in the longer term. E.g. An experiment can increase short term LTV of a user but it can negatively impact long term LTV. While experiments need to be concluded as early as possible, making inference becomes challenging.
- Temporal shifts in the data distribution which raise a few questions like, is the shift permanent or temporary? Should we keep making decisions based on current deployed models? etc.
- Enabling robust projections of the performance of the business under all the aforementioned volatility and limitations in an ever evolving environment of feature deployment and dynamic offers by the marketing and product teams

Now as a Data Science Leader, the company needs your thoughts on overcoming the above limitations. The expectation is to design a robust framework which can automatically take care of most of the limitations. The output of the framework then can be fed into various processes which can keep running without the need of manual intervention in order to account for the limitations.