State the problem, take the problem a step further

Think critically, ask relevant questions, explain your thought process

People pay and then drop out

How to solve this, reduce customer churn

Look at the data, clean? NAs? How to repair that? Remove or fill

Make sure you don’t lose quantity or any special features

Frequency of distribution in histogram; identify outliers or sequential patterns;

check for skewness, for a normal distribution, skewness is zero

very high skewness can violate logistic or linear regression

What is the target variable -- Churn

Modelling approach? -- more in class

Check for correlation, connect to the main business problem

Check for gender, payment type, frequency of use– who’s dropping off more

If credit card is on file and being used, customer might forget to remove the card; cash members might have more problem

Distribution by age, enrollment date

Data observations: age 0-99: look at the distribution, use logic; do customers not share their age correctly

Additional data: seasonal, promotional, occupation, income group, address, personal trainer?

Data collection: how is the data collected? Is there a cap in the system for the data collection? Medium through which data is collected? Additional historical data? Anything that could affect clarity or quality?

State how this data would benefit the analysis.

Check for correlation between down payment and annual fees

**What do you consider the specific business problem and the matching goal of analysis?**

Business problem: Customers sign up for the fitness center throughout the year, more in the first part of the year. But people drop out at some point or another. This also affects the business financially, losing people frequently, try to find the financial impact.

Try to find why? What could be the factors causing this turnover. Look for some potential patterns or segments where this happens more. This will help leadership come up with a strategy on how to target their marketing and other efforts.

The data comes from different fitness centers so data collection methodologies may vary.

From the data given, we observe that the number customers dropping is increasing year on year.   
A graph of blue bars

Description automatically generated

Looking at the chart above which shows the customers who dropped out and their respective enrollment dates, we observe that the trend has been on the rise. The greatest number of drops in the year 2019 were from the customers who enrolled in December 2019, same for 2020. One of the explanations could be that the customers sign up for the fitness center as a new year resolution and then eventually drop out. Although the customer drop could be a direct impact of a customer’s health habits and discipline, there are some other explanations that could be explored.

If we start looking at the other variables or features of a customer, we see that the mode of payment that the customer uses plays a major role. The customers who pay using cash are more likely to drop out, considering that their payments don’t get debit automatically and then they can choose to cancel the membership by just not paying the cash on time. Similar trend goes for cheque since it works in a similar way.  
 A graph of a bar chart

Description automatically generated

Here, it is clearly visible that the percent of people who drop out are majorly using cash or cheque payments.

Additionally, we also notice that the customers who have done a lower down payment are more likely to drop out. This is natural since the initial investment for them has been low.

A graph of a down payment

Description automatically generated

Additionally, the usage rate of customers is also impactful here. The lesser a customer uses the fitness center, more likely he/she is to drop out. We can clearly see a pattern where the red stack which represents the churned customers is higher when usage is low.A graph of a number of blue and red bars

Description automatically generated

**What modelling approach do you suggest and what did you notice in the data that suggests it will be helpful here?**

Remove rows with ages less than 16 and more than 80

Convert enrolment date to date-time data type.

Convert categorical variables using one-hot encoding or dummy coding

Check for nulls or missing data

Create new features necessary that could improve model performance

Remove highly correlated features (if correlation more than 90%)

Since our data has more data about non-churned customers (skewed data), our models could be biased towards that as well

Logistic regression: Useful for binary outcome, split the data into train and test, fit the model parameters on train set and predict the results on the test set.

Use confusion matrix and classification reports to understand model accuracy

**What questions, if any, do you have about the client organization, or its broader industry?**

As a fitness centre, what is unique about your fitness centre, what is that defining quality that sets you in this industry?

Are trainers available for the customers who need it? If yes, are they charged extra for that?

Is an exit survey taken from the customers who drop out? If yes, what are some of the main reasons they site for dropping out?

Are there any promotional schemes ongoing that could impact the enrolment and churn numbers? Is there a possibility that some customers could have enrolled solely for the scheme and then drop out when the scheme ends?

There is a definitive peak in enrolment in the middle of the year (July-August) in additional to the seasonality during the first month of the year. What are the seasonal factors that drive this general behaviour?

All of these questions will help us understand the general trend, seasonal trend, customer behaviours better and help us develop a more targeted and contextual strategy with regards to the business problem.

**What questions, if any, do you have about the data?**

The ages of customers range from 0-99. The customers below age of 16 and above 80 do not seem legitimate. Can these be corrected in some way or should those be removed altogether? If the data needs to be removed, we might end up losing some key information or key components of the dataset. Another approach could be replacing these fields with mean/median of the ages we have so that the other data is retained.

It is mentioned in the initial client meeting that every fitness centre might use a slightly different method of data collection. What are these different methodologies and how do they affect the data quality? Also, do these data collection methodologies impact the amount of customer data that can be stored in the systems?

Can we get access to additional historical data before 2019 which will help us understand the general trend and seasonality aspect better? Additionally, any geographical data relating to the locations of the fitness centres, the economic, ethnic and social label of the surrounding population could also be helpful. This could help us in understanding customer behaviours and if the fitness centres are catering to the requirements of those specific areas.