BUSINESS UNDERSTANDING & HYPOTHESIS FRAMING

1.1 BEFORE WE BEGIN

You will need to have an understanding of Python before starting this program. If you do not know Python, you are welcome to give it a try, but it might be a little challenging for you.

1.2 WHAT IS A DATA SCIENTIST?

What exactly is a Data Scientist?

A Data scientist works with data to try and deliver value to a business. But how is that achieved? Data Science can sometimes overlap with domains such as Data Analysts, Data Engineers and DevOps. How much the Data Scientists role overlaps with these other roles really depends on the company, but in essence a Data Scientist is generally focused on modeling data to be able to accurately predict an outcome, for example, predicting how likely customers are to leave. 4 core skills are used as a Data Scientist:

- Statistics
- Mathematics
- Programming
- Communication

1.3 QUESTION

What are the 4 core skills of Data Scientist?

A. Statistics, mathematics, software architecture, programming

B. [✓ Statistics, mathematics, programming, communication]

C. Mathematics, linguistics, blockchain, security

1.4 KEY ROLES AND RESPONSIBILITIES OF A DATA SCIENTIST AT BCG X

BCG X is transforming businesses using data science to help companies generate competitive advantage. To do this, we typically follow a 5-step methodology:

1. **Business understanding & Problem framing:** What is the context of this problem and why are they trying to solve it?.
2. **Exploratory data analysis & data cleaning:** What data are we working with, what does it look like and how can we make it better?
3. **Feature Engineering:** Can we enrich this dataset using our own expertise or third party information?
4. **Modeling and evaluation:** Can we use this dataset to accurately make predictions? If so, are they reliable?
5. **Insights & Recommendations:** How we can communicate the value of these predictions by explaining them in a way that matters to the business?

1.5 THE BRIEF FROM POWERCO

Your client is **PowerCo** – a major gas and electricity utility that supplies to small and medium sized enterprises. The energy market has had a lot of change in recent years and there are more options than ever for customers to choose form. PowerCo are concerned about their customers leaving for better offers from other energy providers. When a customer leaves to use another service provider, this is called **churn.** This is becoming a big issue for PowerCo and they have engaged BCG to help diagnose the reason why their customers are churning.

1.6 WE NEED TO UNDERSTAND POWERCO'S PROBLEM IN DETAILS

First things first, someone need to understand the problem that PowerCo is facing at a deeper level and plan how you'll tackle it. If you recall the 5 steps in the Data science methodology, this is called **"Business Understanding & Problem Framing".** To formulate PowerCo's issue as a problem using the 5 step data science process and lay out the major steps needed to test it.

1. What do you think are the key reasons for a customer deciding to stay with or switch energy providers? For example: price, is it clean energy, customer service, location etc.
2. What data do you think would be useful in order to investigate these key reasons? E.g Customer purchasing trends over five years, location of business etc
3. If you were to get this data, how could you analyze or visualize it to test whether these reasons may have an impact on churn?

1.7 ANSWER

In order to test the hypothesis of whether churn is driven by the customers' price sensitivity, we would need to model churn probabilities of customers, and derive the effect of prices on churn rates. We would need the following data to be able to build the models.

1. Customer Data: Which should include characteristics of each client, for example, industry, historical electricity consumption, data joined as customer etc.

2. Churn data: Which should indicate if customer has churned.

3. Historical price data: Which should indicate the prices the client charges to each customer for both electricity and gas at granular time intervals.

Once we have the data, the work plan would be:

1. We need to define what price sensitivity is and calculate it.

2. We need to prepare the data and engineer features.

3. Then, we can test our hypothesis using a binary classification model (e.g Logistic Regression, Random Forest, Gradient Boosting Classifier).

4. We would choose a model from one of the tested algorithms based on the model complexity, the explainability, and the accuracy of the model.

5. With the trained model, we would be able to extrapolate the extent to which price sensitivity influence churn.