Project : Does being a plus member increase customer lifetime value? (Causal Inference)

Problem Statement

Q1.1. Background

Briefly describe the **need**, **problem** or **opportunity** - from a business perspective, not a technical one.

Our online store offers a subscription-based loyalty program, *Plus*. Customers can keep on using the platform without one. Yet, subscriptions ensure a recurrent revenue stream and may increase customer lifetime value (CLV). Using a regression analysis, Analytics team has already found that Plus members, on average, has higher customer lifetime value. However, this is simply an association and may stem from a spurious correlation. Product and marketing teams are considering channeling more resources into Plus membership, but **does being a plus member ACTUALLY increase customer lifetime value?**

Q1.2. Stakeholders and impact

Which **teams**, **business functions** or **people** are likely to be **impacted** or **benefit** from the project?

Marketing team needs to make a decision whether to promote brand loyalty programs or Plus membership, their content calendar is already packed. Website developers are interested in this question as they determine how much real estate to devote to the Plus membership. Logistics team is interested in this question as more Plus members translate into more personnel demand due to prime delivery.

Q1.3. Key domain knowledge

Briefly name & describe key **concepts**, **processes**, **risks** and **constraints** which are relevant to the project and might not be understood by all readers of this document.

Randomized Control Trials (RCTs) are the gold-standard for causal inference. Our team has successfully leveraged RCTs via A/B testing for recommendation engines promoting products via notifications, targeted marketing campaigns, checkout flows and more. However, an A/B test (or an RCT) is not feasible for testing Plus membership. **We cannot randomly assign customers to treatment (Plus) and control (non-Plus) conditions.** In other words, Plus members are a self-selected subsample and may differ from the control group on many levels, alongside being Plus members. For example, they may have higher disposable income and do not mind paying for Plus features or buying more products. Causal inference is the process of determining whether a cause-and-effect relationship exists between variables. It involves using statistical methods to infer causality from observational or experimental data, rather than mere correlation. Randomized control trials (RCTs) are experimental studies where participants are randomly assigned to either the treatment group or the control group. This randomization helps eliminate bias and establish

causality between an intervention and its outcomes. A/B testing is a method of comparing two versions of a variable (A and B) to determine which one performs better. It is commonly used in marketing and web development to test changes to a webpage or product against the current version to identify improvements. Subscription-based loyalty programs are services where customers pay a recurring fee to receive special benefits, such as free or expedited shipping, exclusive access to products, and premium customer service. These programs aim to enhance customer loyalty and increase lifetime value by offering consistent value to subscribers.

Value Proposition

Q2.1. How will the project help the business?

Explain **what you want to achieve**, and **how this provides value** to the business.

We are trying to achieve higher customer lifetime value by increasing order frequency and reducing churn. This will increase revenue.

Q2.2. Address the Problem Statement

Explain **how** the project will tackle the issues or opportunities described in the Problem Statement.

To project will uncover and quantify the causal relationship between Plus membership and customer lifetime value (CLV), if any. If Plus membership increases CLV, then Plus membership will be promoted to make sure more and more customers are becoming Plus members. If there is no such relationship, Plus membership will be further tested whether or not it provides any benefits, such as customer loyalty, or satisfaction.

Q2.3. Measuring value

Describe how the value provided by the project can be demonstrated and measured objectively. **Try to quantify this value**.

1. Model will estimate a coefficient for Plus membership, that is the extra CLV generated by being a Plus member. If positive, 2p. Plus membership will be promoted. 3p. New members on top of baseline member acquisition will be calculated 4p. New members will be tracked in 6 months or multiplied with the model coefficient for estimation. If negative or non-existent: 2n. Plus membership will be subject to further tests to see if it provides value.
CLV=Average Purchase Value (APV)×Purchase Frequency (PF)×Customer Lifespan (CL).
CLV=α+β(Plus) · Plus Membership+ϵ ΔN = N_posttreatment −N_baseline
Estimated Revenue Increase=ΔN × β_Plus

Q2.4. Viability

What level of impact, performance or gain is **necessary** for the project to deliver a **financial benefit**?

This needs to be discussed with a more finance-savvy teammate. As a placeholder, if Plus membership increases the CLV by 10%, the efforts are justified.

Q2.5. Potential risks

Briefly identify key financial and reputational **risks**, business **disruption**, and **change management** issues that

Briefly identify key financial and reputational **risks**, business **disruption**, and **change management issues** that could arise from adoption of the solution.

1. Customers may think a Plus membership is necessary for using our platform. 2. Plus membership may distract customers and reduce checkout conversion. 3. Discontinuing Plus membership may damage reputation.

Team

Q3.1. Key stakeholders

List **teams**, **systems** and **individuals** whose responsibilities will be affected by deployment of the solution.

Q3.2. Project Champion[s]

Champions are crucial to maintaining project **momentum** and have a **clear vision of the solution**. Who is your Champion? Note: The champion is often also a stakeholder.

Q3.3. Subject matter experts (SMEs)

SMEs know key **data**, **systems**, or business-**processes** better than almost anyone else. List SMEs whose knowledge and experience will be important to your project.

Q3.4. Team Composition

Estimate the **size** and **composition** of the project team, including:

- Project champion (from above)
- Project management role
- Selected hands-on Subject-Matter Experts (SMEs)
- Business analyst (BA) role
- Data science / AI / ML technical specialist role
- Software development / engineering role
- DevOps role
- Operations and technical support roles

Note that in smaller projects, one person might fill multiple roles. Answering this question may require help from AI/ML and software development experts. That's OK - feel free to guess and refine the team later.

- Someone from customer analytics team for reporting Plus membership and its impact so far. - Someone from financials team responsible for subscription revenue stream, Plus membership. - Someone from customer success team responsible for Plus membership - A data scientist for implementing the causal inference solution.

Data

Q4.1. Key sources

List data which is necessary or highly desirable for the function or evaluation of the solution. In each case, make a note:

- Who owns, controls or maintains the data?
- **How much** data is available (e.g. date ranges, number of samples)?
- Which version or source of the data is most definitive?

- Which version or source of the data is most definitive?

N/A

Q4.2. Data structure

Explore with your team the **structure** of the data sources, and how they can be **linked** together, e.g. via unique identifiers or date ranges. List any potential issues with linking the data together, such as cardinality changes (e.g. many to 1 relations), gaps or missing data. Make notes here, and consider making an Entity-Relationship Diagram.

We need the following tables: - customers - orders - plus_memberships - support_tickets

Q4.3. Origins and sources

Explore and note the **process** by which new data would **continually** be obtained, including **cadence**, **latency**, and any **manual** processes which might be difficult to automate. Who is responsible for this and how will continuity be guaranteed?

N/A

Q4.4. Data quality

Note any known issues, concerns or risks due to data quality. Consider obtaining Exploratory Data Analysis (EDA), which should identify potential issues such as:

- Missing or sparse data
- Very uneven value distribution, or many rare values in categorical data
- Inconsistent data types or encoding / recording
- Need for dimensionality reduction

N/A

Q4.5. Recognise the value of your data

Integrated, trustworthy and coherent datasets are very valuable, even without use in an AI/ML solution. Consider how the dataset you are producing for the project can be used in other business functions, perhaps replacing less well maintained or more inconsistent / outdated data sources. This may be a key value proposition of your project. **Where can you find uses for your data?**

Our data can be used by the customer success team to better understand demographics and purchasing behavior of Plus vs non-Plus members.

Solution design

Q5.1. Identify key entities

Take a minute to think about key entities and concepts involved in your solution. Make notes here. Try to identify:

- **Classes** - key types of entity
- **Samples** - many independent instances of classes, to which inference or optimisation is applied
- **Features** - properties or attributes of each sample, such as measurements
- **Targets** - labels, known correct outputs, evaluation function, etc.

is_member, member_start_date, age, income, n_orders, avg_order_amount, registration_date and some more variables will be integrated into the causal graph. First, a discussion with SMEs is vital

some more variables will be integrated into the causal graph. First, a discussion with SMEs is vital for finding out these factors.

Q5.2. Approach

Will your solution be:

- **Automation**: The solution will perform a task without human intervention, although perhaps with human **review**.
- **Decision support**: The solution will help people to complete a process, perhaps with **recommendations**.
- Generate **insights** or data for people to use

Describe how your solution will do one of these things using specific terminology from your problem statement and explaining how it fulfills part of the value proposition.

Q5.3. Problem representation

This question may require some AI/ML expertise, but have a go anyway. You can always change the answer later. Popular AI/ML problem representations are listed below. **Which one will you use? How will it be fitted to your problem?**

1. **Optimization** (you must be able to generate and evaluate all possible solutions; AI can search through them efficiently)
2. **Unsupervised Learning** (discover patterns in data)
3. **Supervised Learning** - requires a large dataset of samples with "correct" answers. Will learn to generate "correct" answers for other samples. There are two main types:
    1. **Classification**: The answers are categorical labels, such as Case/Control or 0/1.
    2. **Regression**: Approximating a function; the answers are real numbers such as 5.18.
4. **Reinforcement Learning**. You must create a function which defines the quality (reward) of any action or output of the solution. Used when there's no "correct" answer, but the *quality* of answers can be evaluated.

Regression. We will approximate the effect of becoming a member.

Q5.4. Data transformation

AI/ML solutions rarely use structured (relational) data; instead, relational data is usually transformed to tabular format. Even if your data is images or video, it will usually still have the same structure - many **samples**, each with the same **features**. How will you transform your various data sources into a single tabular format? Pay particular attention to **links** between data and **cardinality changes**.

Q5.5. Outline the Pipeline

Sketch out the steps involved in obtaining and producing data for your solution, continuously.

If you must provide "correct answers" for your chosen approach, how will those answers be produced? How will you conduct human or automatic feedback to continuously measure solution performance?

How will users or operators interact with the solution?

How are its outputs integrated into other systems?

Evaluation

Q6.1. Qualitative

How will you evaluate the solution's performance in terms that are meaningful to stakeholders? For example, examination of system behaviour under specific conditions or results for known examples.

Q6.2. Quantitative

What numerical performance **metrics** can you use to evaluate your solution? Which **variables** or **outputs** will be measured using each metric? How good is "good-enough", or what minimum performance is **necessary**? Do you have any existing systems or human performance which can act as a **baseline**?

Plan to measure in ways which will reflect real-world utility and establish the viability of the identified use-case.

Q6.3. Fairness and generalization

How will you evaluate your solution appropriately and fairly, minimizing **bias**?

How will you ensure your solution **generalizes** from your existing data, to real-world conditions? How can you ensure your data is representative of the variability of future, real-world data?

Adoption

Q7.1. Ownership

Who will own, operate and maintain the solution? (Teams or individuals). Can more be done to achieve and sustain adoption of the solution? Who will provide technical support?

Q7.2. Users

Who are the users of the solution? **How many** users are expected? How can you **measure** uptake and use?

Q7.3. Consumers

Who - or **what** - will consume or rely on outputs of the solution?

Q7.4. Integration strategy

How will the solution be integrated into existing business proccesses and systems? What input and output **dependencies** will exist?

Q7.5. Deployment strategy

Consider how to deploy the solution to **minimise technical and change risks**. Potential strategies include:

- **Duplication and parallel operation** of existing process for verification
- **Canary** (gradual introduction across userbase, see who screams)
- **Blue/green** (enables seamless rollback after changes)
- **A/B testing** (helps to verify benefits of modified process)

Note: Ensure **validation** is always part of any new model deployment!

How would you rollback and defer deployment if problems are encountered? Plan for **regular deployments** over the lifetime of a production solution.

Q7.6. Documentation and training

How will you document the **design**, **use** and other **technical details** of the solution - **where** will the documentation live? How will you provide training to users and maintainers? In addition, note the legal status of

any **Intellectual Property (IP)** generated by the project.

Q7.7. Socialisation and awareness

Make a plan to **socialise** the benefits of the solution and raise awareness of your successes. How will you **build momentum** and **interest** in your project, both now and through to delivery and even after adoption?

This document was created with the Causal Wizard AI/ML project designer tool.