Exercise for understanding Data Science Life Cycle

Business Problem

You are a data scientist working for an e-commerce company. The company is expanding aggressively and the leadership team has asked you to come up with a plan to **increase revenue by 25%** in the next quarter **without having to cut down** on any existing operations.

Now, use this statement and plug it into the overall data science lifecycle we covered in this module. We have provided the six stages here – you need to fill in each stage with your answers and thoughts.

1) Problem Definition

Convert the business problem into a data problem

- **Business Problem:** To increase revenue by 25% in the next quarter without having to cut down on any existing operations.
- **Data Problem:** To identify the factors that impact revenue and develop a strategy to optimize those factors.

2) Hypothesis Generation

Generate a set of hypotheses based on the problem definition

- Factors that impact revenue of an e-commerce company:
    - Website traffic volume and sources
    - User engagement metrics- Easy to navigate, Engaging content, Time on site, Page view per sessions, Bounce rate
    - Sales data- Number of orders, Order value, Revenue Generated,
    - Marketing campaign metrics-Impressions, Click, Conversions and Spend
    - Customer data- Demographics, Behaviour & Preferences
    - Supply chain data-Lead Time, Order Fulfiment Rate, Shipping Time
    - Product assortment and pricing-Wide Range of Products at competitive pricing
    - Customer service
- Hypothesis based on Problem Statement:
    - Increasing marketing spend will result in more sales and revenue
    - Improving the user experience on the website will result in increased sales and revenue
    - Offering promotions or discounts will incentivize customers to buy more and increase revenue
    - Analyzing customer behavior patterns to identify upsell and cross-sell opportunities
    - Improving supply chain efficiency and reducing shipping times will lead to increased customer satisfaction and repeat purchases, resulting in higher revenue

3) Data Collection/Extraction

What kind of data do you need based on the above hypotheses? Which variables do you require and how would you collect them?

Data may include website traffic, customer behavior data, sales data, marketing campaign data, supply chain data, customer reviews and feedback data. Variables required for analysis may include:

- Website traffic volume and sources
- User engagement metrics- Time on site, Page view per sessions, Bounce rate
- Sales data- Number of orders, Order value, Revenue Generated
- Marketing campaign metrics-Impressions, Click, Conversions and Spend
- Customer data- Demographics, Behaviour & Preferences
- Supply chain data-Lead Time, Order Fulfiment Rate, Shipping Time

These data can be collected through web analytics tools, transactional systems, marketing automation platforms, customer feedback surveys and other sources.

The metric for the problem of increasing revenue by 25% in the next quarter will depend on the specific strategy and models developed by the data team. However, some common target variable that could be used include:

1. **Revenue:** This is the most direct metric for evaluating the success of the strategy. The team can track the actual revenue generated in the next quarter and compare it to the target revenue.
2. **Conversion rate:** The conversion rate measures the percentage of website visitors who complete a desired action, such as making a purchase. Increasing the conversion rate can lead to increased revenue.
3. **Customer acquisition cost (CAC):** The CAC measures the cost of acquiring a new customer. Reducing the CAC can increase profitability and revenue.
4. **Return on advertising spend (ROAS):** The ROAS measures the revenue generated for every dollar spent on advertising. Improving ROAS can lead to increased revenue and profitability.
5. **Average order value (AOV):** The AOV measures the average amount spent by customers per order. Increasing the AOV can lead to increased revenue.

The choice of metric will depend on the specific strategy and models developed by the data team. It is important to choose a metric that is relevant to the goals of the project and can provide actionable insights for future optimization. Here, the most relevant metric for this problem is likely to be the actual revenue generated in the next quarter as this directly measures the success of the strategy in achieving the revenue target. While the other metrics mentioned such as conversion rate, CAC, ROAS, AOV can be useful in understanding the factors that contribute to revenue growth and identifying areas for optimization they are ultimately indirect measures of success. For example, a high conversion rate or low CAC may not necessarily result in the desired revenue increase if the average order value is low.

Therefore, the actual revenue generated in the next quarter should be the primary metric for this problem and the other metrics can be used to guide the development and optimization of the strategy.

4) Data Transformation and Exploration

a) What kind of visualization techniques will you use to explore the data?

- Visualization techniques can include scatterplots, heat maps and line charts to explore relationships between variables.

b) Do you need to transform any variables before proceeding with the analysis?

- Yes, Data may need to be transformed by normalizing variables, creating new features or removing outliers before analysis.

5) Model Building

a) What is the evaluation metric for your problem?

We can use Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R-squared (R2) etc. For example, MSE measures the average squared difference between the predicted revenue and the actual revenue. A lower MSE indicates better accuracy, but can be sensitive to outliers.

The choice of evaluation metric will depend on the specific models and techniques used by the data team. It is important to choose a metric that is relevant to the goals of the project and can provide actionable insights for future optimization.

b) What kind of models will you build?

Models can include regression analysis, decision trees and random forests to identify the factors that impact revenue.

c) What if your model validation strategy?

Model validation strategy can include cross-validation, hold-out validation and A/B testing.

6) Model Implementation

a) Which model, based on the ones you have built, is best suited to your business problem? Is there any trade-off between the accuracy and the interpretability?

The best model suited to the business problem may depend on the trade-off between accuracy and interpretability. A simple regression model may be more interpretable but less accurate while a complex machine learning model may be more accurate but less interpretable.

b) Any specific steps you'll follow for monitoring your model's performance?

Steps for monitoring model performance can include regular tracking of revenue and profit metrics, ongoing A/B testing of marketing campaigns and monitoring of customer feedback and satisfaction metrics.