Support Documentation for BeverageCo

Overview

Congrats! You have been chosen as our data science partners for our horizontal expansion journey. Please always keep in mind the scope of this project: the effective **identification**, **prioritization**, and **activation** of potential customers.

Business Case Steps

1. Identification

To horizontally scale our distribution, **we first need to understand our current customer base**. This involves analyzing the features behind our best-selling customers.

For example, pet-friendly hotels might not be our best customer, whereas a more crowded restaurant could be.

In this section you should go through the following steps:

- **Data Preprocessing**: ensure data quality and consistency.
- **Combining Data**: join datasets carefully, considering which join strategy best fits the problem.
- **EDA**: analyze and visualize all variables using appropriate plots. This step should help you understand the profile of our best-selling customer.
- **Feature Engineering**: transform your variables into more relevant ones.

2. Prioritization

Based on the customer profile obtained from the EDA, **prioritize the pool of potential customers that best align with our interests**. The objective is to identify the best customers for our expansion. In this section you should go through the following steps:

- **Linear Regression Modeling**: find the best linear regression model, check its assumptions, and interpret it. What is the impact of breaking multicollinearity?
- **Alternative model (optional)**: we encourage you to train more advanced models. Please prioritize explainability over performance.
- **Evaluation**: perform an inner evaluation of your best performing model and use it to prioritize potential customers.

3. Activate

We currently have 33 sales representatives covering our existing customers. Based on the sales potential of prospective customers, decide **how many additional sales representatives should we hire and where should we place them**. Our budget allows us to hire up to 50 additional sales representatives. In this section, you should go through the following steps:

- **Clustering**: generate localized customers groups, both current and potential, and rank them by total sales. Our n sales representatives should cover the n localized customer groups with the highest total sales. Please bear in mind that each sales representative can cover a maximum of 300 customers.
- **Potential curve**: based on the clustering, analyze the additional sales each additional sales representative would bring to the table. Can you find an ideal number of sales representatives?
- **Hiring Locations**: Based on the clustering and the ideal number of sales representatives, identify the optimal locations for hiring additional sales representatives. Keep in mind that sales representatives are best located when they are centered in their group of customers, and that current sales representatives cannot relocate.

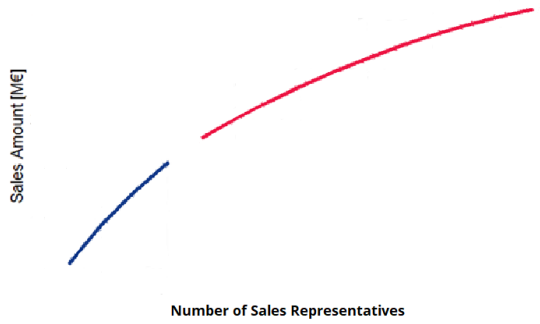

Fig.1: Example of potential curve without tickers. The blue line goes from 1 to the current number of sales representatives, and the red line goes from that number on. Hence, the red line represents the added sales by hiring additional sales representatives.

Data Dictionary

Datasets

Customer Info: General information of each customer, including contact details and location.

Variables

- **Distance**: distance is meters to the closest warehouse
- **City**: city the customer is located at.
- **Store Size**: size of the store in sqft.
- **Opening Hours**: opening hours for each customer.
- **Phone Number**: phone contact of each customer.
- **Free Wifi**: free wifi availability at the customers' location.
- **Parking**: parking availability at the customers' location.
- **Description**: verbose description of the customers' offering and activities.
- **Zip Code**: zip code of the customer.
- **Latitude**: positional latitude in degrees.
- **Longitude**: positional longitude in degrees.
- **Date Opening**: debut day of the customer.

- **Sales Representative Longitude**: positional longitude of the sales representative.
- **Sales Representative Latitude**: positional latitude of the sales representative.
- **Sales Representative ID**: identifier of the sales representative.

Customer Economics: Economic features of current customers, such as sales and number of reviews.

Variables

- **Rating**: average rating of the customer.
- **Number of revie**ws: total number of reviews of the customer.
- **Sales Amount**: total sales from the customer of our products.
- **Number of employees**: number of employees working for the customer.
- **ID**: identifier of the customer.

Sociodemographic: Additional sociodemographic features like population and unemployment rates.

Variables

- **City**: city the customer is located at.
- **Province**: province the customer is located at.
- **Population**: total population of the city.
- **Young Population**: average percentage of young population (16-25 years old).
- **GDP per capita**: average GDP per capita.
- **Unemployment rate**: average percentage of unemployment rate.

Potential Customers: Information about potential customers across the country.

Variables

- **Distance**
- **City**
- **Store Size**
- **Opening Hours**
- **Phone Number**
- **Free Wifi**
- **Parking**
- **Description**
- **Zip Code**
- **Latitude**
- **Longitude**
- **Date Opening**
- **Sales_representative_longitude**
- **Sales_representative_latitude**
- **Sales Representative ID**
- **Rating**
- **Number of reviews**
- **Sales Amount**
- **Number of employees**
- **ID**

Tools and Libraries

Here is the stack we usually leverage at the company; feel free to expand or use different libraries.

- General
  - Python
  - Jupyter Notebooks
- Preprocessing
  - Pandas
  - NumPy
  - Groq
- Exploratory Data Analysis (EDA)
  - Matplotlib
  - Plotly
  - Folium
- Modeling
  - Statsmodels
  - Scikit-learn
- GUI
  - Streamlit
  - Plotly Dash
- Version Control
  - GitHub

Frequently Asked Questions

- Can a current sales representative be assigned potential customers?
  - Yes, as long as the sales representative does not cover more than 300 customers. The sales representative can also stop visiting a current customer if it is preferable to visit a potential customer instead.
- Can I create the customer groups without using clustering?
  - Yes, you can use other methods to generate customer groups, for example linear optimization. Make sure to include the distance in your model so that the groups are localized.
- Is it necessary to consider the cost per sales representative when finding the optimal number of sales reps.?
  - No, you can use the sales per customer to get an idea of each cluster's potential.