Task 2: Exploratory Data Analysis

Task Overview

What you'll learn

- How to investigate whether price sensitivity is the most influential factor for a customer churning
- How to use frameworks to conduct exploratory data analysis

What you'll do

- Use python to analyze client data
- Create data visualizations to help you interpret key trends

Your AD is giving you more responsibility!

Well done for your initial understanding of the case with PowerCo. After reviewing your project plan, the AD would like you lead on the Data Science deliverables for the rest of the project.

The AD would like you to investigate whether price sensitivity is the most influential factor for a customer churning, and if not, to what extent does price sensitivity influence churn.

Before we begin on this task, what exactly is price sensitivity?

What is price sensitivity?

Price sensitivity is **the degree to which demand changes when the cost of a product or service changes.**

In the context of PowerCo, the “demand” refers to the demand for energy consumption.

Price sensitivity is commonly measured using the price elasticity of demand, which states that some consumers won't pay more if a lower-priced option is available.

What is price elasticity of demand?

Price elasticity of demand is a measurement of the change in consumption of a product in relation to a change in its price

Complete the quick knowledge check and move onto your exploratory data analysis.

Exploratory data analysis

The client has sent over 2 datasets and it your responsibility to perform some exploratory data analysis.

What is exploratory data analysis?

Exploratory data analysis (EDA) is a technique used by a Data Scientist to gain a holistic understanding of the data that they are working with.

It is mainly based around using statistical techniques (such as descriptive statistics) and visualizations to gain a deeper understanding of the statistical properties that the data holds.

Complete the quick knowledge check on the next step and let's get started.

Let's get familiar with the data

As a Data Scientist at BCG, there will be occasions when you need to analyse data or investigate an issue and you are not provided strict instructions or guidance. You may be thinking, where do I start?

It is a highly valuable skill to begin to learn how to investigate a problem independently. A great way to learn this skill is to build a framework for analysis that works for you.

In this step, you'll need to analyse client data sets using Python and upload your work as a Jupyter notebook. We'll show you an example answer on the next step, but we encourage you to give it a go first!

The client has sent over 3 data sets (shown below):

1. Historical customer data: Customer data such as usage, sign up date, forecasted usage etc
2. Historical pricing data: variable and fixed pricing data etc
3. Churn indicator: whether each customer has churned or not

You need to analyze the following using Python:

- The data types of each column
- Descriptive statistics of the dataset
- Distributions of columns

Estelle has provided a starter Jupyter notebook has been provided for you to use as a template to complete your work.

Here are some tips to help you:

Let's take a look at the 3 data sets PowerCo. has sent over:

- A first good step is to review the data to make sense of the columns…
    - **Hint:** Look *at data types of column to gain a better understanding of what the columns mean. This is why data description documents are important - they describe exactly what the columns represent.*

- Once you understand the columns in the dataset. Now you want to look at how the values in the data vary…
    - ***Hint:*** *this is why reporting descriptive statistics is useful because it'll tell you some basic statistical properties of the columns in the data. It will also tell you how many values feature within a column, e.g. does a column only have 1 unique value or 100? This is useful to know because you can then start to*

build a picture of what this data represents.

- You now understand how the values vary and what the data represents - next up, it can be useful to visualize some of this…
    - ***Hint:*** *Not all visualizations are useful. Keep the visualizations simple and always keep in mind what you're trying to show. E.g. if you want to see how the distribution of a column looks and that column has 1000 unique values, using a pie chart would not be good because it would become too crowded! If the values are numeric, a distribution plot would be more appropriate.*
    - ***Hint:*** *make sure to use the starter Jupyter notebook provided, as this will show you some example visualizations and sample code to use!*

At this stage, you should now have a clearer understanding of what the data is and how it looks. This framework is not exhaustive, but it shows how you could start to build your own framework for analysing data.

When you're ready, upload your Python code notebook to complete this task.

Example Answer

Great work! Take a look at the example answer below to see how a professional would have attempted this task. Think about what you did well and how you can improve.

We've also provided an explanation for you to read through as you review the Jupyter Notebook.

Explanation

Getting set up - This task is focused on exploratory data analysis of the client and price data provided:

- The first thing you should do is download the provided Jupyter notebook and the CSV datasets.
- To run the notebook, you need to make sure that you provide the path for the CSV files so that you can load the data.
- By running the cells that exist within the notebook from Estelle, this will show you what the two datasets look like, it will provide you with code to produce descriptive statistics and it will also give some examples and sample code on how to visualize the data.

Analysis - Once you've run the cells provided, it was your job to build on this exploratory analysis:

- The visualization provided by Estelle shows how many companies churned vs. how many companies did not churn. We can see from this that the churn rate is approximately 10%. This is actually a very good churn rate, the closer the rate is to 0%, the better.
- The next series of visualizations were created in an attempt to try and dive deeper into how churn changes based on other factors (using other columns). This is useful for us to investigate because it may help us to understand factors that drive churn.
- In the notebook we visualize churn vs. sales channel, contract type, number of products, number of years and origin/contract offer.
- For example:

    - We see that for sales channel, there are some sales channels that yield customers churning but there are also other sales channels that have no customers churning.
    - For contract type, we see quite an even split for customers churning. This is interesting because this may suggest that contract type is not a driving factor towards churn rate.
- Additionally, for some columns their distributions with churn rate included. This is useful for us to understand because based on the distribution of a column, this could affect our feature engineering later.
- We look at the distribution of consumption, subscribed power and forecast in the notebook.
- For example:
    - We notice that the distribution of consumption is very skewed, this is called a positive skew since it is biased towards lower values on the x axis.
    - This is interesting because you may decide to treat this column to reduce the skewness later on during feature engineering. But also because we may want to visualize if there are any outliers within this column.
    - To investigate outliers, we use a boxplot. From the boxplot we can see that with the column as it is there are definitely some outliers. Once again this is interesting because we may choose to remove some of these outliers later.