FEATURE ENGINEERING & MODELLING

3.1 NOW IT'S TIME FOR FEATURE ENGINEERING

Well done for your analysis on the influence of price sensitivity relative to churn!

"I think that the difference between off-peak prices in December and January the preceding year could be a significant feature when predicting churn".

As the Data Scientist on the team, you need to investigate this question. So, in this task you'll be responsible for completing feature engineering for the dataset. Before we start on this task, let's explain what feature engineering is.

3.2 WHAT IS FEATURE ENGINEERING?

Feature engineering refers to Addition, Deletion, Combination, and Mutation of your data set to improve machine learning model training, leading to better performance and greater accuracy.

In context of this task, feature engineering refers to the engineering of the price and client data to create new columns that will help us to predict churn more accurately.

Effective feature engineering is based on sound knowledge of the business problem and the available data sources.

3.3 QUESTION

Which of the following techniques would not fall under feature engineering?

A. Merging
B. Multiplication
C. Aggregation
D. Inspection

3.4 CREATING THE NEW FEATURES

We'll show you some tips on the next step before you start your task, but make sure you have the relevant files downloaded before moving on!

- Download the new CSV data Jupyter notebook
- Run the cells in the Jupyter notebook to create the feature

- We'll continue working in this Jupyter notebook to create some more columns.

3.5 WHAT YOU NEED TO THINK ABOUT THE WORK

Your task is to create new features for your analysis and upload your complete python file. Below are some of the tips on how to get started.

As before, a good way to quickly learn how to effectively build a framework to follow . Below is an example of how you could attempt this task.

First – can we remove any of the columns in the dataset ?

- There will almost always be columns in a dataset that can be removed, perhaps because they are not relevant to the analysis, or they only have 1 unique value.

Second- Can we expand the datasets and use existing columns to create new features?

- For example, if you have "date" columns, in their raw form they are not so useful. But if you were to extract month, day of month, day of year and year into individual columns, these could be more useful.

Third- can we combine some columns together to create "better" columns?

- How do we define a "better" column and how do we know which columns to combine?
- We're tying to accurately predict churn- so a "better" column could be a column that improves the accuracy of the model.
- And which columns to combine? This can sometimes be a matter of experimenting until you find something useful, or you may notice that 2 columns share very similar information so you want to combine them.

Finally- can we combine these datasets and if so, how?

- To conbine datasets, you need a column that features in both dataset that share the same value to join them on.