Dear name[LDS],

After researching PowerCo problem context (our client), I come up with the following hypothesis.

Customer’s churn is driven by the price sensitivity, which means that the higher the price, the lower the quantity of product customers will buy. If the customer is offered 20% price discount, it is very likely that the churn rate would be lower, which means that the customers are less likely to quit consuming PowerCo’s utility product.

The dataset needed to conduct the test of the hypothesis should have the following features:

1. **Customer data** - which should include characteristics of each client, for example, industry, historical electricity consumption, date joined as customer etc

2. **Churn data** - which should indicate if customer has churned

3. **Historical price data** – which should indicate the prices the client charges to each customer for both electricity and gas at granular time intervals

The following steps are laid out to test the hypothesis:

**Step 1: Perform exploratory analysis on the dataset given.**

*Checking out how our data looks like and visualize how it interacts with our label (churned or not?)*

**Step 2: Feature engineering.**

*2.1. Group the numerical columns by using clustering techniques*

*2.2. Apply****Label Encoder****to categorical features which are binary*

*2.3. Apply****get\_dummies()****to categorical features which have multiple values.*

**Step 3: Investigating how the features affect Retention by using Logistic Regression.**

*Prdicting churn is a binary classification problem. Customers either churn or retain in a given period.*

**Step 4: Binary classification model with XGBoost.**

*We should prepare features (X) and label (y) sets and do train & test split to come up with an optimal model.*

I look forward to hearing your feedback on this finding.

Best regards,

Tram