Hello,

After going through the problem statement related to PowerCo and understanding the hypothesis, it is understood that we need a structured dataset to build models and test our hypothesis.

The data should comprise of following:

- Customer Details – It should account for relevant information in form of variables like Concerned Industry, Units Consumed, Whether Discount offered, Spend on power bill, and other independent factors in the form of variables.
- The Churning detail should be recorded and included in the dataset. It should be done in the form of categorical variable, where binary encoding will depict whether customer leaves or stays, 0 – No Churn, 1- Churn happens.
- Moreover, the data should be collaborated in a way so that it represents historical data, thereby helping in identifying patterns and generating a robust model.

Following the formation of the dataset, we need to identify the type of model which best aligns with the business requirement and the data paradigm.

Since our dependent variable, Churn, is categorical, it best suits to design a classification model to identify the pattern and significant variables impacting the customer decisions.

Logistic Regression algorithm is the primitive Classification technique to initiate the analysis. After this, advanced algorithms like Naïve Bayes, SVM, Random Forest, and other Boosting techniques could be used to enhance the accuracy and find out which model best fits our dataset.