Dear Mr. Smith,
In relation to the Project commissioned by PowerCo, after a careful framing of the problem, I have defined what I believe to be the best way to solve it.
In the following Framework, I illustrate how I intend to proceed by breaking the problem down into steps to be solved one at a time.

	Problem Statement	**Choice of algorithm type**	**Data cleaning and collection**	**Exploratory Data Analysis and Data Processing**	**Model building**	**Results evaluation**
Key Tasks	How likely is the customer to abandon us?	Classification algorithm	• Choice of data sources • Cleaning the dataset	• Understand features importance • Categorical variable management • Features selection	• Splitting the dataset • Parameter setting • Prediction running	• Accuracy evaluation • Confusion Metrix • AUC metrics and ROC Curve

Next, I have described each step in a more appropriate manner.

Step 1 – Problem Statement

Based on the client's business problem, I defined a simple phrase as "problem statement" that will help us fully understand the type of problem and the input and output of the problem to be solved: How likely is the customer to abandon us?

Step 2 – Choice of algorithm type

The choice of algorithm to be used depends on the specific problem to be solved. The PowerCo goal is:

- Prevent customer churn by preemptively identify customers at risk
- Design suitable interventions to improve retention

In this scenario, I would proceed by modelling the categorical variable "churn" as a binary variable, using a binary classification algorithm where P(y=1) will represent the probability of output to the class "yes" and P(y=0) will represent the probability of output to the class "no" in reference to customer churn. For this problem, I would use the logistic regression algorithm to determine the classification of the input values.

Step 3 – Data cleaning and collection

For this step, we will require from the client company the historical data collected on customers through sources such as CRM software, web analysis, sentiment analysis, customer service and more others. Ideally, I would expect to get a dataset composed of a range of information about customers who have left the company in the last month, the services each customer has subscribed to, customer account information (such as contract, payment method, how long customer has been a customer, monthly expenses, and total expenses), customer demographic information (such as gender, age, residence location). Ideally, the dataset we should obtain at the end of the collection and cleaning process should look like the following:

Churn is our target variable and indicates whether the customer churned or not (Yes, No)

Customer ID	**SME**	**Residential Customer**	**how long Customer**	**Tenure**	**Monthly Charges**	**Total Charges**	**Gender**	**Age**	**Living City**	**Gas Service**	**Electricity Service**	**Contract Type**	**Payment Method**	**Churn**
….	….	….	….	….	….	….	….	….	….	….	….	….	….	YES
….	….	….	….	….	….	….	….	….	….	….	….	….	….	NO
….	….	….	….	….	….	….	….	….	….	….	….	….	….	YES
….	….	….	….	….	….	….	….	….	….	….	….	….	….	NO

Step 4 – EDA and Data Processing

We must first do a deeper dive into the process of exploratory data analysis (EDA) to gain a better understanding of our data.
During the exploration I would check a couple of different aspects of each feature (e.g., with the help of visualization I would look at the distribution of the sample for the output classes, such as its distribution for Tenure groups, the probability distributions of the variables related to the numerical features and its distribution for the categorical ones and so on). After looking at the key features and how they interact with each other, we can proceed with the pre-processing phase to accommodate the categorical features we have in our data and target our target variable, Churn so that our model is able to interpret it meaningfully. At this stage we are going to perform features selection because using all values of categorical variables as features would raise the problem of multicollinearity and break the model. To finish the process, we will standardize the data values and split the dataset into train and test.

Step 5 -Model Building

This step should be relatively simple and quick as it involves choosing a few parameters before building the model and, after it is built, predicting future values.

Step 6 - Results Evaluation

To assess the reliability of the model, we check its overall performance with the accuracy metric. To give us a holistic view of how well our classification model is performing and what kinds of errors it is making, I would proceed by plotting the confusion matrix. The matrix will return the observations divided by the 4 metrics: True Positive, True Negative, False Negative, False Positive. In this way, we will have the necessary data to calculate more descriptive metrics of the model's performance: Precision, Recall, F1 Score. As an alternative

to the confusion matrix to evaluate classifier performance, we could use the AUC metric and an ROC-curve graph.
This visual graph will illustrate the true positive rate against the false positive rate of our classifier. The AUC will give us a singular numeric to quantify the overall accuracy of our classifier model.

In conclusion, the model will be able to predict which customers will be at risk of dropping out each month, but it is also able to offer insights into the reasons behind the drop-out. In this way, by integrating the model with their existing software, our client will have several insights that will allow the marketing and customers service departments to choose the best approach in relation to the specific customer. If the motivation behind the abandonment is related to e.g., the perceived quality of the service maybe, a discount of 20% may not be the most effective tactics to prevent churn.
Thank you for your time.

Best regards,
Simone Travaglione