PILOT STUDY PROPOSAL

678 words

In this pilot study, we will perform a binary classification task, where the goal is to predict whether a customer will likely encounter difficulties in paying the increasing electricity cost based on the provided features.
Possible informative features that could be provided by AENERGY include:

- Demographic information, such as age, gender, and income level
- Family composition, including the number of members and their ages
- Information about the heating system, including model and power rating
- Energy consumption habits, such as usage patterns and peak consumption times
- Information about the customer's history with AENERGY, including payment history and account balance

We will use a supervised machine learning algorithm such as logistic regression or decision trees for the learning procedure. Both of these algorithms are suitable for binary classification tasks and can handle a large number of features. We would choose logistic regression for this task because it is simple to implement and easy to interpret, it's a powerful tool for the classification problem, it can handle a large number of features, and it also performs well with categorical and numerical data. Logistic regression is also able to identify the relationship between the features and the target variable, and how much the features are likely to impact the target variable. Additionally, Logistic regression can handle outliers and missing values. There are some other types of the Classification algorithm, they are as follows:

Types of Classification: -

Decision Trees: -
It is a non-direct model that beats a couple of the disadvantages of straight calculations like Logistic relapse. It fabricates the characterization model as a tree structure that incorporates hubs and leaves. This calculation includes different if-else proclamations which help separate the design into more modest constructions and give the ultimate result in the long run. It tends to be utilized for relapse just as grouping issues.

Random Forest: -
It is an outfit learning technique that includes numerous choice trees to foresee the result of the objective variable. Every choice tree gives its result. On account of the grouping issue, it takes the larger part vote of these different choice trees to order the ultimate result. On account of the relapse issue, it takes the normal of the qualities anticipated by the choice trees.

Naïve Bayes: -

It is a calculation that depends on Bayes' hypothesis. It expects that a specific element is free of the incorporation of different elements. They are not connected. It, by and large, doesn't function admirably with complex information because of this presumption as in the vast majority of the informational collections there exists a connection between the highlights of some sort or another.

K-Nearest Neighbours (kNN): -

It utilizes distance measurements like Euclidean distance, Manhattan distance, and so on to work out the distance of one element from every other informative element. To characterize the result, it takes a larger part vote from k closest neighbours of every main item.

Support Vector Machine: -

It addresses the items in complex spaces. These information focuses are then isolated into classes with the assistance of hyperplanes. It plots an n-layered space for the n number of highlights in the dataset and afterwards attempts to make the hyperplanes to such an extent that it isolates the informative items with the greatest edge.

To evaluate the performance of the system before deployment, we will use a technique known as cross-validation. This will involve splitting the provided data into training and testing sets and using the training set to train the model, and then evaluating its performance on the testing set. In this way, we will be able to get an estimate of how well the model is likely to perform on unseen data. Additionally, we will also use metrics such as precision, recall, accuracy and F1 score, to measure the performance of our model.

In conclusion, A machine learning-based approach using logistic regression is suitable for predicting whether a customer is likely to encounter difficulties in paying the increasing electricity cost. The performance of the system will be evaluated using cross-validation and different evaluation metrics.