Title : **Pilot-Study Proposal**

Word Count: 654 (excluding table of content & References)

Table of Contents

Abstract:

A pilot study proposal to investigate the machine learning procedures that can be used successfully to solve the problem. The purpose of this assignment is to predict whether a customer is going to face the difficulty in paying the increasing electricity bills or not considering the following features like heating system power, age, habits etc. This assignment is given to study and investigate different machine learning procedures and find the best fitted model for the problem.

Type of predictive task:

According to the dataset provided and the description given it is a classification task because the output is binary variable as it indicates whether the customer will struggle or not. As the classification model tries to get the conclusion from the input values given the train the model

Possible informative features:

According to me some possible informative features that can be useful for this study are:

- Customer's income
- Customer's education level
- Family size
- Previous bill history
- Size of customer's residence & family size

The learning procedure/ procedures:

There are different classification procedures that can be used to solve this problem to name few of them:

a) Random forest classifier
b) Gradient boosting classifier
c) Decision Tree
d) Logical Regression
e) K- Nearest neighbours

We use the above given models to get the best accuracy of the model. The above-mentioned procedures are very efficient for such types of investigation. Random forest classifier uses the average to improve the accuracy, which is predicted by the model and is one of the best feature as it controls overfitting of the model. Gradient boosting classifier is good for large dataset, as it gives good accuracy with large data.

Decision Tree is very easy to understand and the best part is that it can handle both numerical and categorical data really well. Logical Regression is considered one of the best algorithms in the classification procedures as it explains the several input variables on a single output variable.

K- Nearest neighbours is simple to implement and efficient if training data is large. It simply stores the instances of the training data as it does not constructs a general internal model.[1]

To get the best model we need to train the models and identify the one, which gives the best accuracy without overfitting the model.

Evaluate the performance of system before deploying it:

There are many ways by which systems performance can be evaluated before deploying it. One method is to split the data into training and test variables. Basically, dataset is spilt into 80-20 ratio. So to train the model 80 percent of data will be used and the remaining 20 percent will be used to test the model if it is predicting accurately or not. Then, cross validation can be used.

There are different scaling features also which can help in improving the accuracy like Standard scaler, Robust scaler, MinMax scaler. We can also improve the accuracy by removing the skewness of dataset. One more point should be considered while training a model that there should be no categorical data. If there is categorical data that should be changed because it will not let the model train efficiently. To get the best accuracy results one have to train the models and evaluate, which model will give the best accuracy.

Conclusion: To conclude, there are many procedures which can give results(accuracy). To find the one with best(highest) accuracy, the models should be train on them and the various features should be tested and the one which gives the best result should be selected. There should be always comparative approach while solving the problem, because when we try, test and compare with other models it gives you the best results and as explained above there are different features, which can be used to improve the accuracy.

So for the above investigation I will train the models individually on above procedures and evaluate, which model will give the best accuracy result.

References:

https://analyticsindiamag.com/ [1]

Lecture notes on machine learning

Scikit-learn online documentation and tutorials : https://scikit-learn.org