University of Essex

Department of Computer Science and Electronics Engineering

Part 1: Pilot Study

Subject:
CE802 Machine Learning

Registration number: 2201829

Supervisor: Dr. Vito De Feo

Date of submission (January 18 2023)

Word count: 698

Summary

The electricity bill increase is one of the most recent hot topics in the news headlines. The energy company AENERGY would like to predict whether the customers will encounter difficulty in paying electricity bills due to an increase in electricity bills, using machine learning. The company will give the required historical data of recent customers for prediction.

1. Solution

For predicting whether customers will suffer or not due to surge in electricity bills, the following four points were considered for the pilot study.

1.1 Type of Predictive Task

In this task, we need to find the type of machine learning method we will be using for the predictions. As we have the historical data of customers, supervised learning methods can be used. Classification is a type of supervised learning, that can be used when we need to categorise or predict discreet class labels using the new input data. Binary classification can be used if we have only two class labels. Regression is also a type of supervised learning, it is used to predict continuous values based on the new input variables. According to the nature of the problem, we have to classify the customers into two categories, whether they will suffer(True) or not suffer(False). So it can be considered to be a classification problem.

1.2 Informative Features

Choosing the most appropriate feature has a huge impact on getting an accurate prediction. Following are some of the important features that will be useful in predicting whether the customer will be affected by the increase in electricity bill.

a) **Age of Occupants** – People in different age groups need a different type of heating requirement. If there are infants or very old people there they may need more heating requirements.

b) Number of Occupants – This feature can be merged with other features to find the trend of usage of electricity.

c) **Electronic equipment usage** – Households with more electronic equipment like kettles, induction cooktops, microwaves etc will be more impacted when compared to those using gas stoves.

d) Heating Systems – If the heating system is based on electricity rather than gas.

e) Council Tax Band – This feature can be used to find the approximate size of the house

f) Recent Electricity usage(Monthly & Weekly) – Monthly and weekly electricity will have more visibility of usage of electricity. For specific months electricity usage will be more because of the weather or other circumstances.

g) **Billing Period** – Monthly billing or billing of 2 months or 3 months

1.3 Learning Procedures

The following binary classification techniques can be considered for the prediction task:

a) Decision Tree – A decision tree classifier has the capability to capture descriptive decision-making knowledge from the given input data. Decision Trees are easy to explain and interpret. While using a decision tree, we need not want to be bothered about whether the data is linearly separable or whether they have outliers or not[1].

b) XG Boost – XG Boost is a gradient-boosting decision tree. It uses sequentially built decision trees to provide high-accuracy scores and a highly scalable method (for training), also it is less prone to overfitting[2].

c) Extra Tree – Extra Tree Classifier is an ensemble learning technique in which, it takes the aggregated results of de-correlated decision trees in the forest and output it as the classification result.

d) SVM – The main characteristics of SVM is very high accuracy and performance. It provides us great flexibility to choose the kernel for even non-linearly separable data, also SVM is comparatively less prone to overfitting.

e) Logistic Regression –If the dependent variable is dichotomous (binary), logistic regression can be used for classification. In logistic regression, we can regularize our model easily without worrying about the correlation of features. Also, in addition to that it has a very good probabilistic interpretation, which helps us to update our model easily when predicting with a new set of inputs.

1.4 Evaluation

For the evaluation of the model confusion matrix and classification report can be used. From the classification report we will get the Accuracy score, Precision, Recall, f1 score etc and the confusion matrix will give an insight classification of data with the count of True Positive, True Negative, False Positive, and False Negative values.

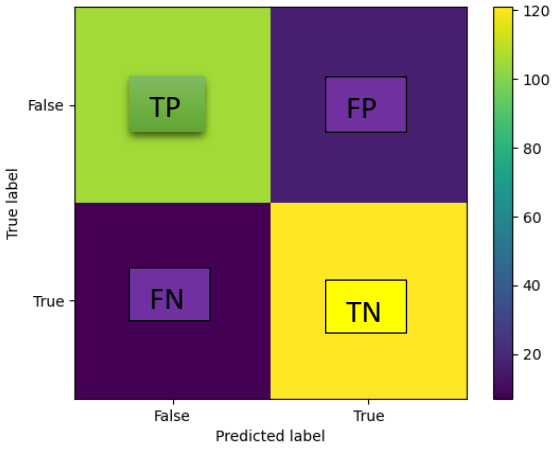

Figure 1. Confusion Matrix

References

1. Priyanka, and Dharmender Kumar. "Decision tree classifier: A detailed survey." *International Journal of Information and Decision Sciences* 12.3 (2020): 246-269.

2. https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/