PILOT STUDY PROPOSAL

University of Essex

Monzur, Rifat ID#: 2200367

Word Count: 517

Table of Contents

Task

AENERGY company would like to predict if a customer will be able to pay increasing electricity cost on the basis of few features. Company likes to know if prediction is possible from historic data that will be provided using machine learning algorithm and if possible what kind of machine learning procedure needs to follow.

Feasibility

Feasibility of the task vastly depend on the quality and amount of data provided by AENERGY. Since company assured that they will be able to provide relevant and quality historic data, it should be feasible to use machine learning to solve the task.

Type of Prediction

AENERGY wants us to predict if customer will be able to pay or not after increasing energy price. So, there is only two possible outcomes. In other words, there are two classes of outcomes either positive or negative. From all those observations, we can conclude that this is a binary classification.

Possible Features

As company already offers to provide customers historical data with several features like age, family composition, heating system model and they are currently able to pay bill or not. Other than those features, customers salary, household income, bank balance, credit history, previous history of late bill payment, type of job and house owner or renting house could come handy. All those features need to be labeled with if customer has been paying bill or not. As we are going to apply supervised machine learning algorithm.

Learning Procedures

Decision Tree: It is one of most common and efficient classification algorithms. Easy to understand and visualize. It can handle both categorical and numerical data. Relationship between features do not affect performance.

K Nearest Neighbor: Like decision tree, KNN is also one of the most common classification procedures. It is easy understand and implement. It has only one hyper parameter so it Is also use to tune

Logistic Regression: If features have linear relationship, logistic regression performs well. It is also easy to implement and very quick to predict. However, for non linear features, logistic regression do not perform well

Boosting Classifiers: Boosting is a procedure where various weak algorithms are used together to make strong predictions. Some boosting perform horizontal and some perfom vertical stacking to make better predictions. Boosting procedures generally perform well. So, we will try few boosting algorithms like ada boosting classifier and gradient boosting classifier.

Performance Evaluation

Firstly, we will split data into two parts. One for test and another for training. Since we are using supervised learning, we will use cross validation to validate the models. Since it is a classification problem, we would check what percentage of the prediction is accurate in both test and training data.

We also need to check confusion matrix. As if majority of the predictions are one class and model predict one class only, it will have a good accuracy percentage. However, it will not perform well in the real world. It will always perform poorly for other class. So, we need to check confusion matrix for recall and precision accuracy.