**ASSIGNMENT REPORT**

**PREDICTING CUSTOMER DIFFICULTY IN PAYING ELECTRICITY COSTS AND VARIATION IN THEIR ANNUAL EXPENDITURE USING MACHINE LEARNING**

**WORD COUNT WITHOUT REFERENCES: 1317**

**INTRODUCTION**

Rising energy prices are a growing concern for households, with 66% of adults in the United Kingdom reporting an increase in their monthly cost of living (Office for National Statistics, 2022). This study presents two machine learning systems, each with different goals. The first system aims to predict customer difficulty in paying electricity costs. It uses a decision tree classifier to analyze multiple features. Its performance will be compared to five other machine learning algorithms.

The second system aims to predict the variation in the annual expenditure of a customer using a linear regression model and evaluate against three alternative machine learning methods. The datasets used were consumer attributes. Plots were used to graphically represent the results. The results were also summarized in a panda’s data frame.

The goal of both systems is to determine which machine learning algorithms perform best in predicting how rising energy prices are affecting customers' expenditures and their ability to pay the costs.

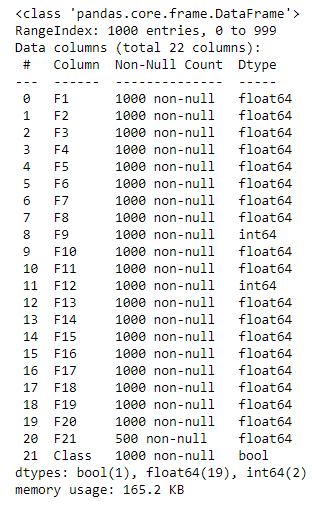
**METHODOLOGY**

**Data**

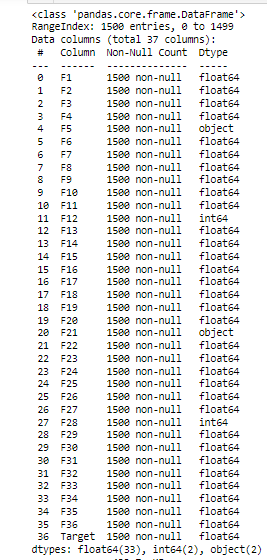
The data used to develop the model for predicting customer difficulty in paying electricity costs was provided by the energy firm AENERGY and included a variety of characteristics such as age, family composition, heating system model, habits, and so on. It also had markings that indicated whether the customer was having difficulty paying their energy bills. The dataset also included 1000 samples with 20 characteristics and one label. One of the features, "F21," was missing half of its data and was represented by NaN values.

The data used to develop the model for predicting variation in the annual expenditure of a customer was also provided by the energy firm AENERGY. It had a column that contained positive and negative values that indicated the changes in customers' yearly spending due to the rising energy prices. The dataset also included 1500 samples with 35 features and one label.

**Fig 1 - Dataset information for the first system**



**Fig 2 - Dataset information for the second system**



**Metrics**

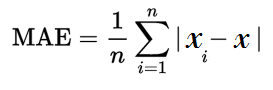
1. **Accuracy:**

This is calculated by dividing the number of correct predictions by the total number of predictions (Sweeney et al., 2022).



1. **Mean Absolute Error:**

The Mean Absolute Error (MAE) is the mean of the absolute difference between actual and predicted value. It is used as an evaluation metric in regression tasks and the formula is represented by:



**Preprocessing**

Data preprocessing is an essential step in the machine learning process, as it helps to improve the quality and structure of the data used to train a model (Gupta et al., 2022).

In the study of the first system, data preprocessing was performed on the dataset to address issues such as missing values, incorrect data formats, errors in data entry, and noise. Specifically, the missing values in the F21 column were addressed using four different strategies: filling the missing values with the mean of the column, filling the missing values with the mode of the column, filling the missing values with the median of the column, and removing the F21 column from the dataset altogether.

The performance of different strategies for filling in missing values in the F21 column was evaluated using a decision tree classifier model. The model was trained on a training set consisting of 90% of the data, and the filling strategy that resulted in the highest accuracy on the test set, which comprised 10% of the data, was selected.

In addition to the decision tree classifier, five other machine learning algorithms were trained on the dataset: logistic regression, gradient boosting classifier, random forest classifier, AdaBoost classifier, and SVC. The training was conducted using two different sample sizes for the training data (80% and 90%). The training results were compiled in a data frame to facilitate model comparison and selection of the best-performing model.

In the study of the second system, data preprocessing was performed on the dataset to address issues such as incorrect data formats. The dataset includes categorical variables in two columns, F5 and F21, which machine learning algorithms find difficult to process. These columns were manually label-encoded to facilitate their use.

After successfully encoding the categorical columns, a linear regression model was trained on a training set consisting of 90% of the data. A plot was developed to show the relationship between the actual and predicted values of the model.

In addition to using linear regression as a machine learning technique, Bayesian Ridge, Gradient Boosting Regressor, and AdaBoost Regressor were also trained on the dataset. 90% of the data was used for training and the results were represented by a data frame to enable a comparison of the performance of various models. Plots were generated for each algorithm’s predicted values against the actual values.

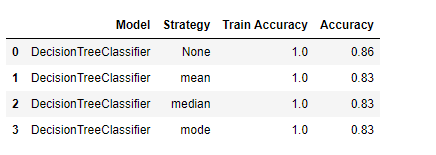
Finally, the model with the best results after training was fitted to the training data and used to predict labels for a holdout test dataset.

**RESULTS**

**First system:**

Fig 3. represents the results of the decision tree classifier trained on the dataset using different fill strategies to handle the missing values in the F21 column.

**Fig 3.**



The "None" strategy signifies the total elimination of the column with blank values. On the test dataset, the decision tree classifier had an accuracy of 86% in its predictions. The accuracy of the other methods, which included replacing the missing data column with its mean, median, or mode, was 83%. Thus, removing the column with missing values improved the decision tree classifier's performance.

Five more machine learning algorithms—Logistic Regression, Gradient Boosting Classifier, Random Forest Classifier, AdaBoost Classifier, and Support Vector Classifier—as well as the decision tree classifier, were put through additional tests. The column with missing values was completely removed, and the models were trained with two different splitting strategies (90% of train data and 80% of train data).

Fig 4. displays a data frame with the results of various machine learning methods with an 80% sample size for training. Random Forest performed the best with 85% accuracy, while Support Vector Classifier performed the least with 68% accuracy.

**Fig 4.**

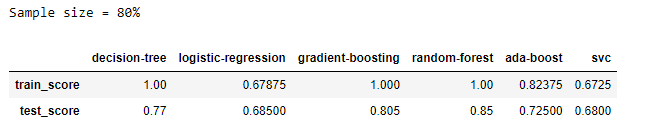
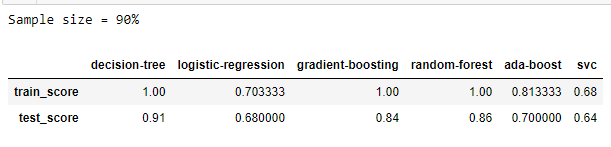


Fig 5. depicts the same data frame as Figure 3, but with a 90% sample size. It was revealed that by training on more data, the decision tree classifier outperformed all other algorithms, with an accuracy of 91%. The performance of the other machine learning methods changed only slightly.

**Fig 5.**



In conclusion, the decision tree classifier was discovered to be the best model for this task as it had an accuracy of 91%, and it was deduced that machine learning techniques can also be used to forecast if a client will struggle to pay their energy bills despite the price increase.

**Second system:**

Fig 6. represents the performance of the linear regression model after the manual label encoding was performed on the dataset. It had a train MAE of 593 and a test MAE of 653.

**Fig 6.**

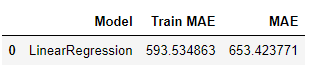


Fig 7 is a plot that shows the similarity of the actual values versus predicted values of the linear regression model.

**Fig 7.**

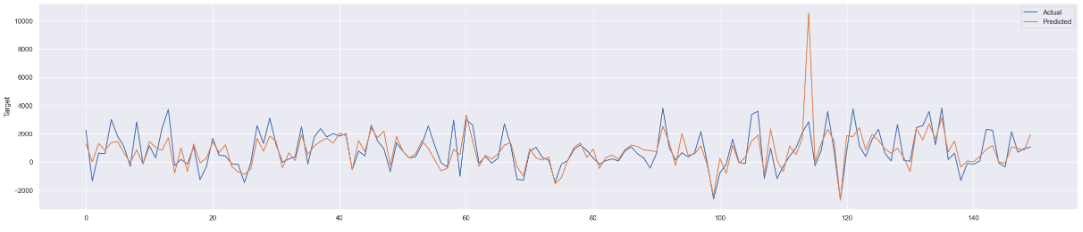
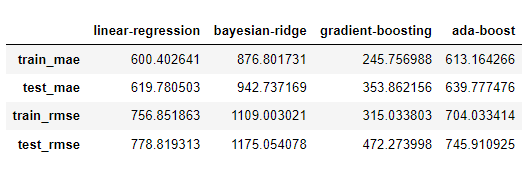


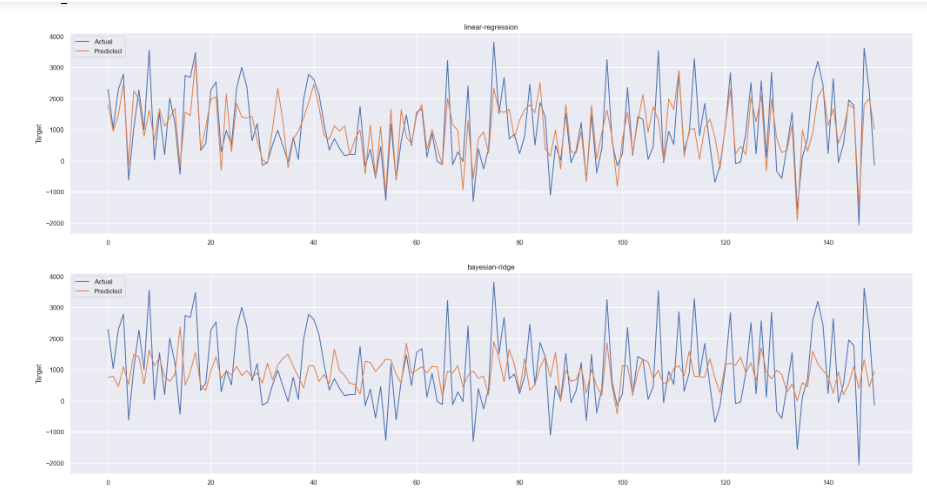
Fig 8 depicts the performance comparison of all four machine learning algorithms trained on the dataset. The gradient boosting regressor outperformed other algorithms with a MAE of 353 compared to linear regression with a MAE of 619, Bayesian ridge with a MAE of 942 and AdaBoost with a MAE of 639.

**Fig 8.**

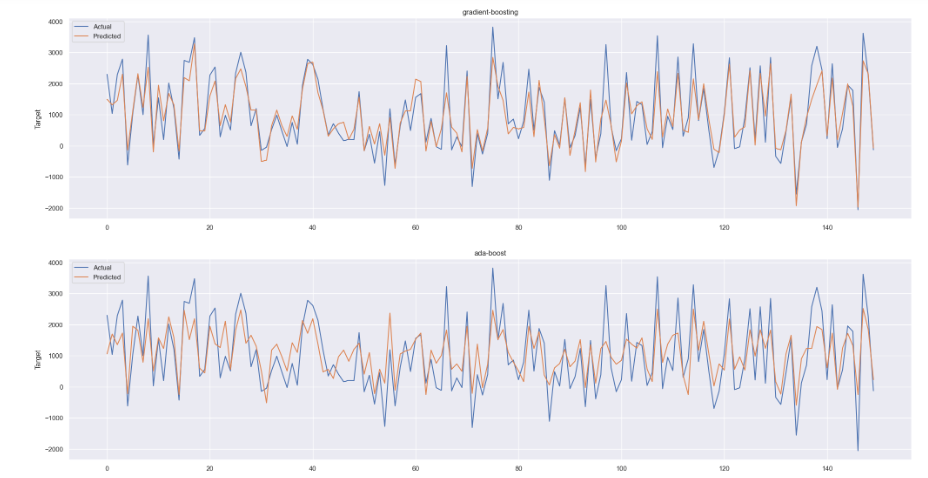


The relationship between the actual and predicted values of all 4 models is visualized in Fig 9 and Fig 10.

**Fig 9 - Linear Regression and Bayesian Ridge**



**Fig 10 - Gradient Boosting Regressor and AdaBoost Regressor**



In conclusion, Figure 8 shows that the gradient boosting regressor algorithm has the lowest mean absolute error (MAE) of 353, outperforming other algorithms, linear regression, Bayesian Ridge, and AdaBoost which had MAEs of 619, 942, and 639, respectively. This suggests that the gradient boosting regressor is the most suitable algorithm for this dataset and problem.

**REFERENCES**

*Energy prices and their effect on households.* (2022, February 1). Office for National Statistics. <https://www.ons.gov.uk/economy/inflationandpriceindices/articles/energypricesandtheireffectonhouseholds/2022-02-01>

Gupta, S., Saluja, K., Goyal, A., Vajpayee, A., & Tiwari, V. (2022). *Comparing the performance of machine learning algorithms using estimated accuracy*. Measurement: Sensors, *24*, 100432.<https://doi.org/10.1016/j.measen.2022.100432>

Sweeney, C., Ennis, E., Mulvenna, M., Bond, R., & O’Neill, S. (2022, May 23). *How machine learning classification accuracy changes in a happiness dataset with different demographic groups*. MDPI. Retrieved January 8, 2023, from<https://www.mdpi.com/2073-431X/11/5/83>