University of Essex

Department of Computer Science and Electronics Engineering

Report: Comparative Study

Subject:
CE802 Machine Learning

Registration number: 2201829

Supervisor: Dr. Vito De Feo

Date of submission (January 18 2023)

Word count:1339

Introduction

The report consists of experiments and analyses done for the comparative study of the Classification and Regression problem. In the classification problem, customers were classified on the basis of whether they will be affected by the increase in electricity bill or not. In the Regression problem, we need to predict the variation in annual expenditure due to the increase in electricity bills.

1. Classification Problem

I. Experiment and Analysis

Binary classification is used to solve the problem, as we have to classify whether customers will suffer (True or False) from an increase in electricity bills. With provided historical data, different methods have been tried to get the most accurate result. Following are the analysis and experiments that were done to get the most accurate result.

a) **Target Data Distribution Analysis** - The dataset contains a total of 1000 records, out of which 506 data have the Target value of 'True' and 494 have the Target value of 'False'. Figure 1. depicts the data distribution of Target values.

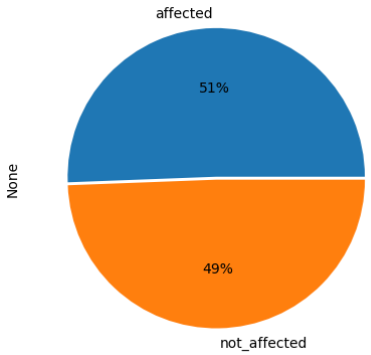

Figure 1. Target Value Distribution

b) **Data Distribution of Features** –The data set contains 21 features and all are numerical fields. Only one field contains null values and imputation techniques have been used to fill it. From figure 2., It is clear that input data is not normally distributed and scaling is required.

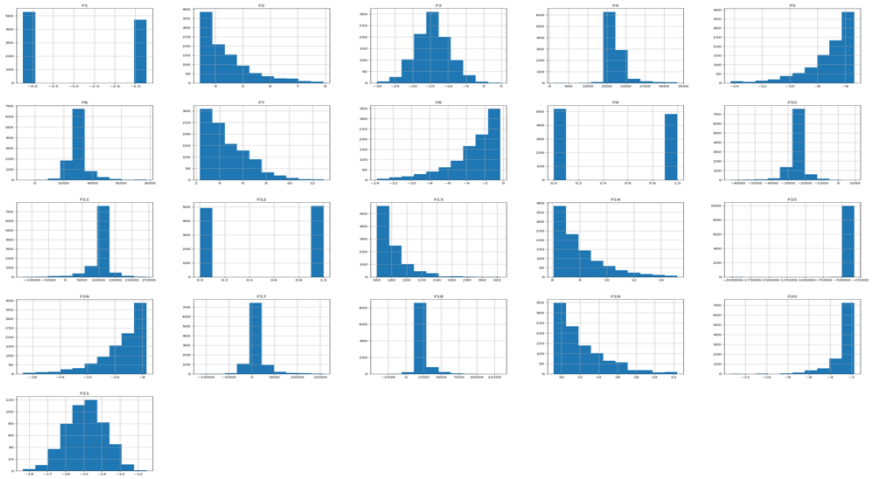

Figure 2. Feature Distribution

c) **Outlier Detection** – Outlier plays an important role in getting a more accurate result. Some models like XG Boost are very sensitive towards outliers. Figure 3 depicts the picture of outliers in the dataset. The first part of the picture shows the outliers in the input data and the second part shows the picture of the dataset after removing outliers. Outliers with values above the higher threshold replace with upper boundary value and values less than the lower boundary has been replaced with lower boundary value.

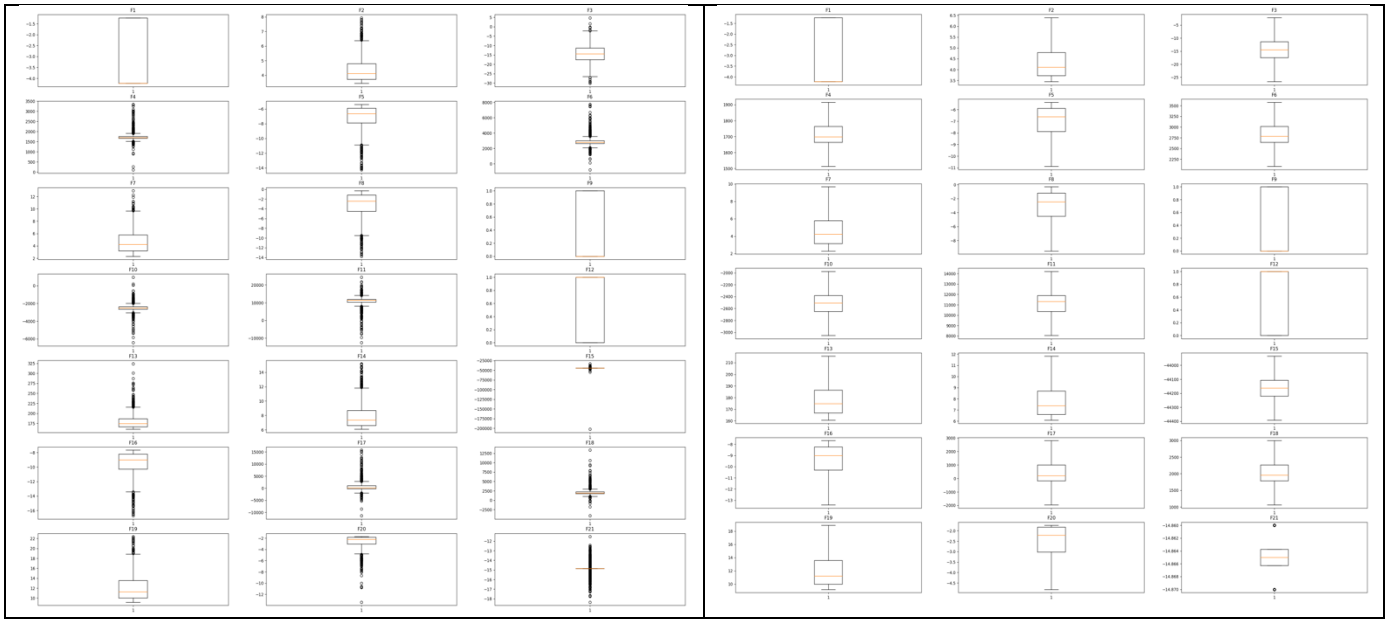

Figure 3. Before and after removal of Outliers.

d) Normalization - Normalization has been done to transform data to a similar scale in order to improve training stability and performance

e) Imputation – Input data contains null values in one of the columns, as a baseline approach, first droped the column and train the model and checked the accuracy, and later experiment with median, mean and mode values. Using the baseline approach

accuracy score was low compared to filling it with other imputation methods. Figure 4. Depicts the results of experiments done with different imputation methods

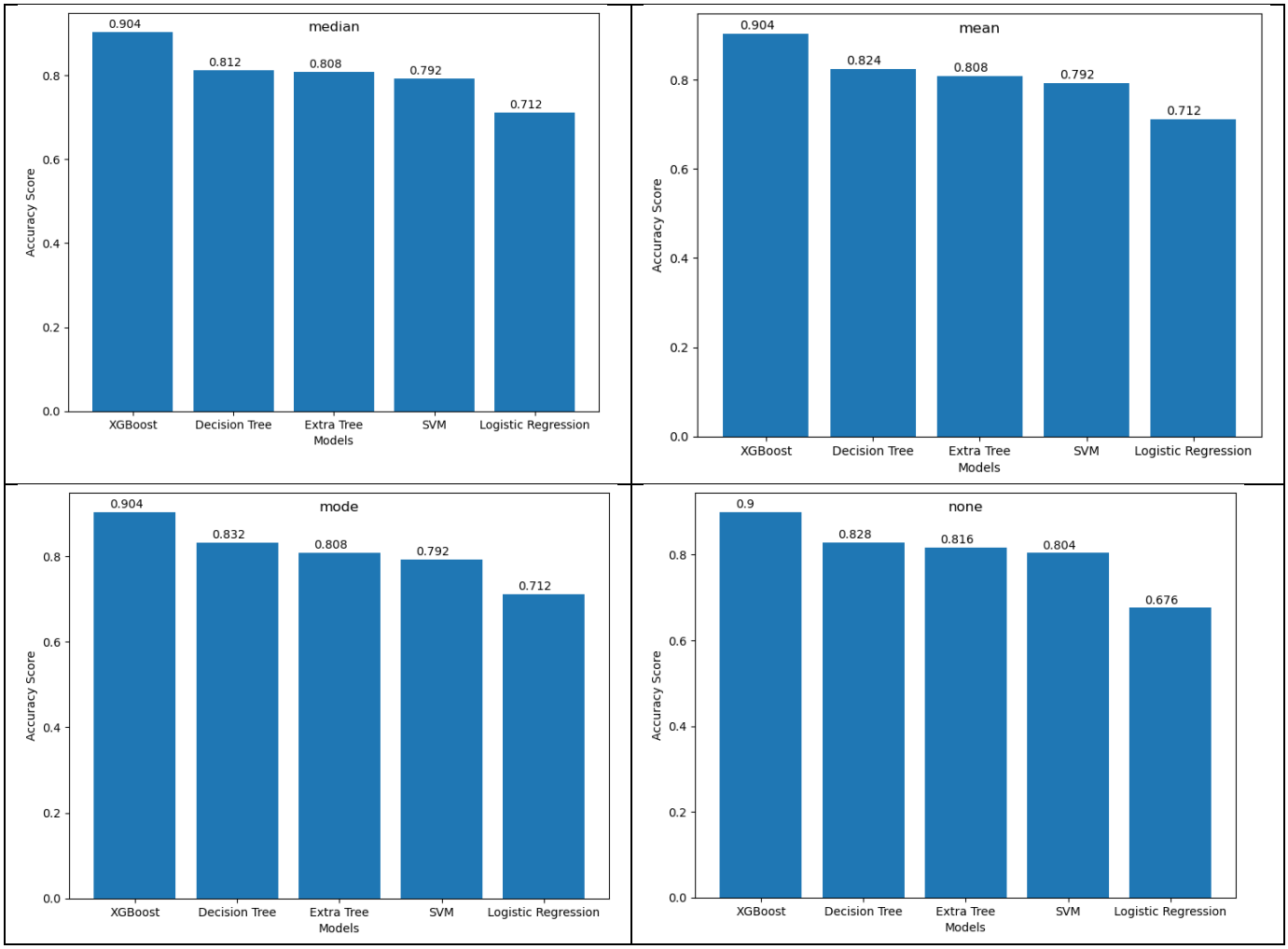

Figure 4. Accuracy Score with different imputation methods and dropping the column

f) **Hyperparameter Tuning** - Grid search has been used for finding the best parameter and best model.

g) **Cross Validation** – Cross Validation has been used to check whether the result is overfitted or not and get a more generalized result.

h) **Models** – After pre-processing, the models were trained using the pre-processed data. Following are the 5 supervised learning methods used to classify customers.

- XG Boost
- Decision Tree
- Extra Tree
- SVM
- Logistic Regression

i) **Model Evaluation -** Training data has been divided into 75% training and 25% validation dataset. The Models trained with the training set are evaluated using the prediction in the validation set. Accuracy score, Precision, Recall, and f1 score were used to check the accuracy score and also classification report and confusion matrix gave better visibility of classification.

II. Results

a) **Best imputation method** - The Median gave the best result to fill the null values in the 'F21' feature compared to other imputation techniques' mean and mode. Also tried a baseline approach by dropping the 'F21' column. The result obtained after the grid search confirmed that the median was the best method to replace the null values.

b) **Best Model** - XG Boost, Decision Tree, Extra Tree, SVM and Logistic Regression were the classification methods used for training the model.

- SVM and Logistic Regression ended with low accuracy, precision, recall and f1 score compared to other classification models used in the experiment. Logistic Regression has the lowest accuracy score of 71.2 % and SVM has an accuracy score of 79.2%.
- The Extra Tree classifier performed comparatively well with an accuracy score of 80.8 %.

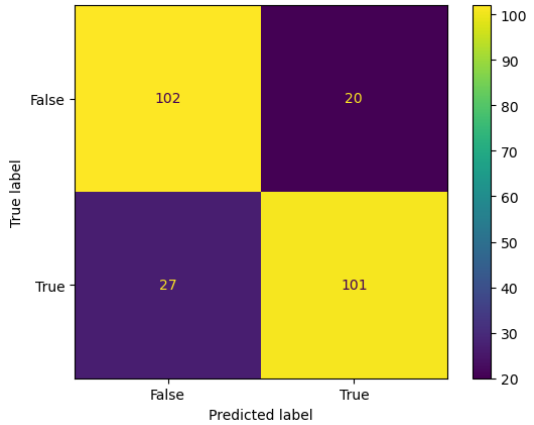

Figure 8. Confusion Matrix of Extra Tree Classifier

- XG Boost and Decision Tree got better scores compared to other models. Among all models, XG Boost outperforms Decision Tree

and other models with the highest accuracy score of 90.4 % and Decision Tree has an accuracy score of 81.2 %.(Figure 11.)

Decision Tree XG Boost

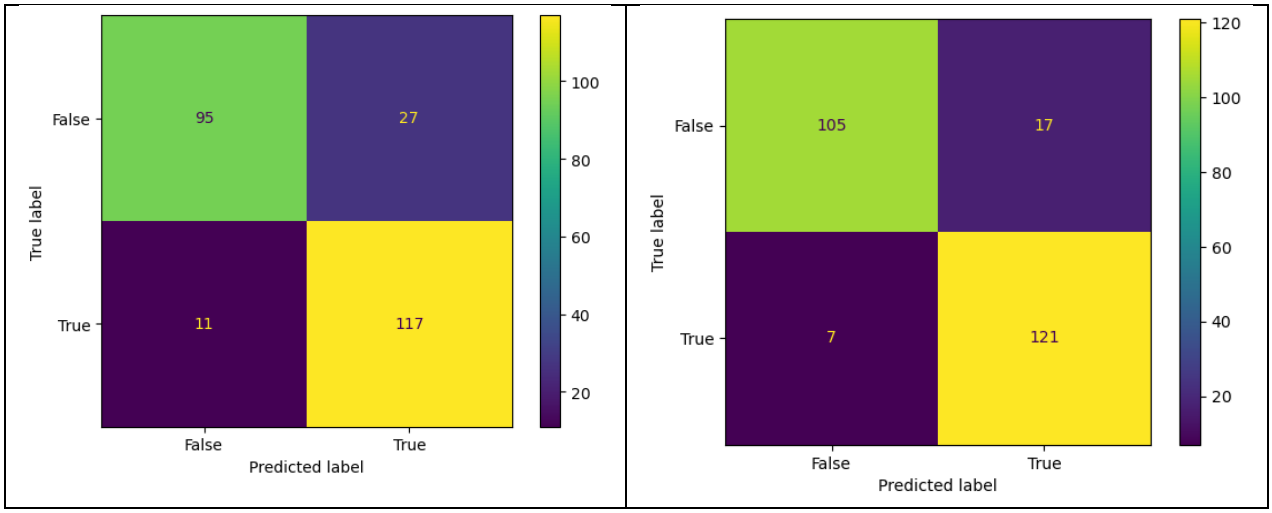

Figure 9. Confusion matrix obtained from Decision Tree and XG Boost Classifier

The following figure shows the final result obtained after training the training set with the best parameters obtained from the grid search.

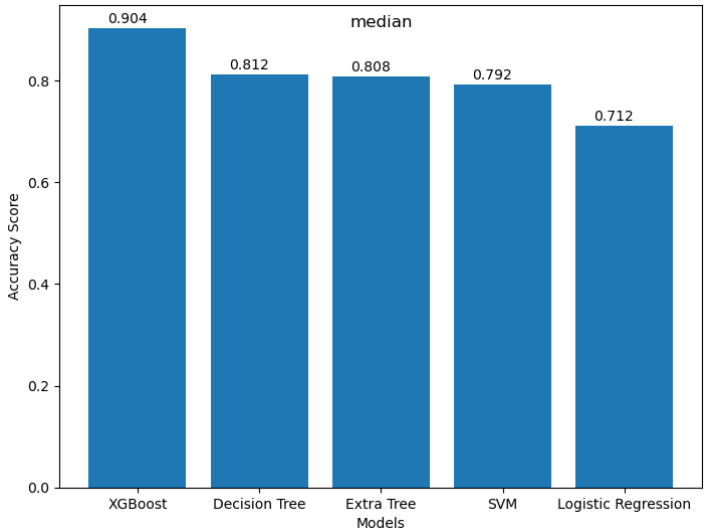

Figure 10. Accuracy score of each model with the median a the imputation method

Figure 9. depicts the confusion matrix obtained from the validation set evaluation using Decision Tree and XG Boost Classifier. From the confusion matrix, it is pretty clear how many values are correctly and wrongly classified with True Positive, True Negative, False Negative and False Positive values.

Figure 11. is the classification reports of XG Boost classifier, we got precision, recall, and f1 score of 91%, 90.5% and 90.5 % respectively.

	precision	recall	f1-score	support
False	0.94	0.86	0.90	122
True	0.88	0.95	0.91	128
accuracy			0.90	250
macro avg	0.91	0.90	0.90	250
weighted avg	0.91	0.90	0.90	250

Figure 11. Classification Report of XG Boost Classifier

2. Regression Problem

I. Experiment and Analysis

In the additional comparative study, the problem has been resolved using regression techniques, as the target values are continuous. The aim of this experiment is to find the variation in the annual expenditure of each customer due to the raise in electricity bills. Following analysis and experiment and analysis have been done to get the most accurate result.

a) **Feature Data Analysis –** A total of 36 features were there in the dataset. Datasets contain both numerical and categorical data. Categorical data has been converted to numerical data for training the model. Figure 12. Shows the distribution of features.

Figure 12. Feature Distribution

b) **Outliers** – Outliers from all features have been identified and removed to get the best result. Figure 13. Depicts the identified outliers and data set after the removal of outliers.

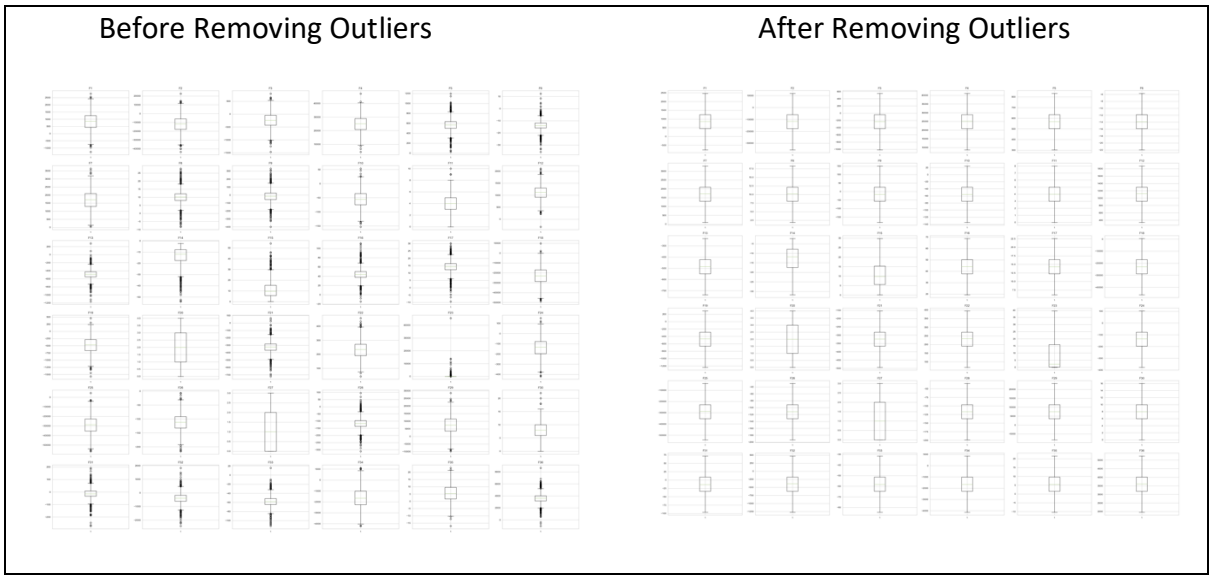

Figure 13. Before and after the removal of outliers

c) **Normalization** - Normalization has been done to transform data to a similar scale in order to improve training stability and performance

d) **Hyperparameter Tuning** - Grid search has been used for finding the best parameter and best model.

e) **Cross Validation** – Cross Validation has been used to check whether the result is overfitted or not and get a more generalized result.

f) **Model**s - The following regression methods have been used for the process.

- XG Boost Regressor
- Ada booster Regressor
- Linear Regression
- Gradient Boosting Regressor
- Extra Tree Regressor

g) **Evaluation Methods** - Root means square, Mean Square Error and Mean Absolute Error has been used for calculating the error and r2_score have been used for calculating the model score. Cross-validation helped to identify whether the model is overfitted or not and gave a better-generalized result.

II. Result

Best Model – Linear Regression, XG Boost Regressor, Ada Booster Regressor, and Gradient Boosting regressor are the regression methods used in the experiment.

- Extra Tree Regressor has the least accuracy compared to other models with an r2_score of 73.07.
- Ada Booster Regressor underperformed compared to Extra Tree Regressor with r2_score of 74.67 %.
- Linear Regression also performed well compared to above two models with a root mean square error score of 575.01 and r2_score of 80.09%.
- Both XG Boost and Gradient Booster performed well compared to all other models with high r2-score and low root mean square error values of 84.04 % and 82.09 %, and 514.85 and 545.45 respectively.

Table 1, shows the overall result obtained after the grid search, Figure 8 is a graphical representation of the result with respect to r2_score. When compared to all models XG Boost Regressor was the best model to get the best result according to the experiment.

Table 1. Error Values

model	rmse	mse	mae
XGBoost	514.854980	265075.650178	409.780997
Adabooster	648.644296	420739.423334	526.034139
Linear Regression	575.016847	330644.374630	474.696600
Gradient Boosting	545.459798	297526.391729	428.329098
Extra Tree Regressor	668.742782	447216.908216	515.568637

Figure 14. shows the graphical representation of r2_score values against each model.

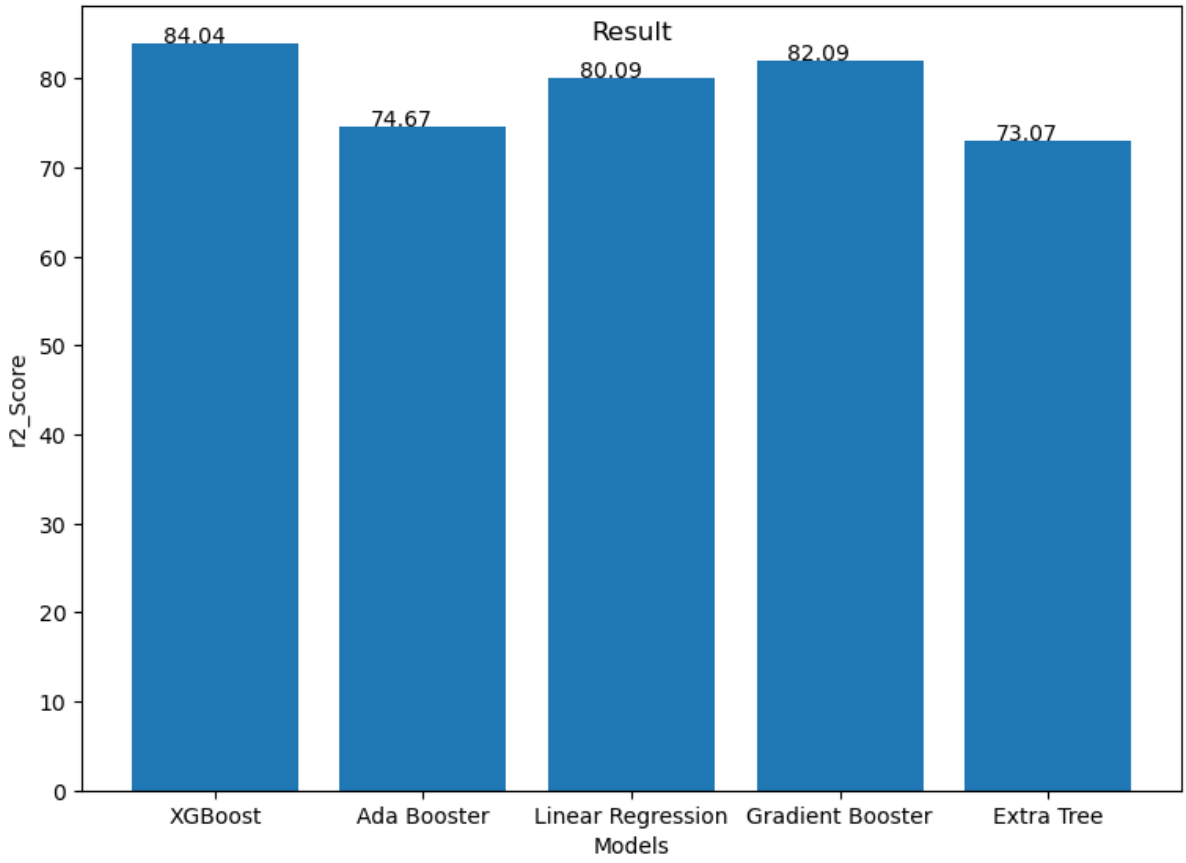

Figure 14. Result of the grid search on each model

Conclusion

Two experiments have been done for finding whether the customers will be affected by the rise in electricity bills or not and to predict the variation in the annual expenditure of customers after the increase in electricity bills respectively. For the former different classification methods have been tried and XG Boost Classifier gave the best accuracy score compared to the other models. And the second was a regression problem and the XG Boost regressor gave the least root mean square value and high r2_score value compared to other models.

References

1. https://www.edupristine.com/blog/decision-trees-development-and-scoring#:~:text=Ap art%20from%20overfitting%2C%20Decision%20Trees,are%20prone%20to%20sampling% 20errors.

2. https://www.analyticsvidhya.com/blog/2016/03/complete-guide-parameter-tuning-xgboost-with-codes-python/