Detailed Report on Machine Learning Procedures Carried out for AENERGY

CE802: Machine Learning

Department of Computer Science and Electronic Engineering

University of Essex

Prepared By

Onyemelonu Nneoma Charity

[2207938]

Word Count:1370

January, 2023.

Part 1

The Utilization of Machine Learning Models in Forecasting Payment Challenges for Customers

Goal: The aim of this procedure is to predict if a customer is going to encounter difficulties in paying the increasing cost of electricity provided by AENERGY on the basis of few features of Customers' historical data.

Exploratory Analysis

The Data after imported were just having column titles of F1 – F22, of which they are not explainable, I had to rename these titles considering possible numerical factors that can influence the payment of electricity while still considering the kind of data filled in each columns and their measurement patterns. These were all based on assumptions as they may not be 100% correct, however a proper name to the column will better explain and add more insights to the pre-processing of the data.

Some were integers others were floats. Most floats had negative ranges and some others were positive.

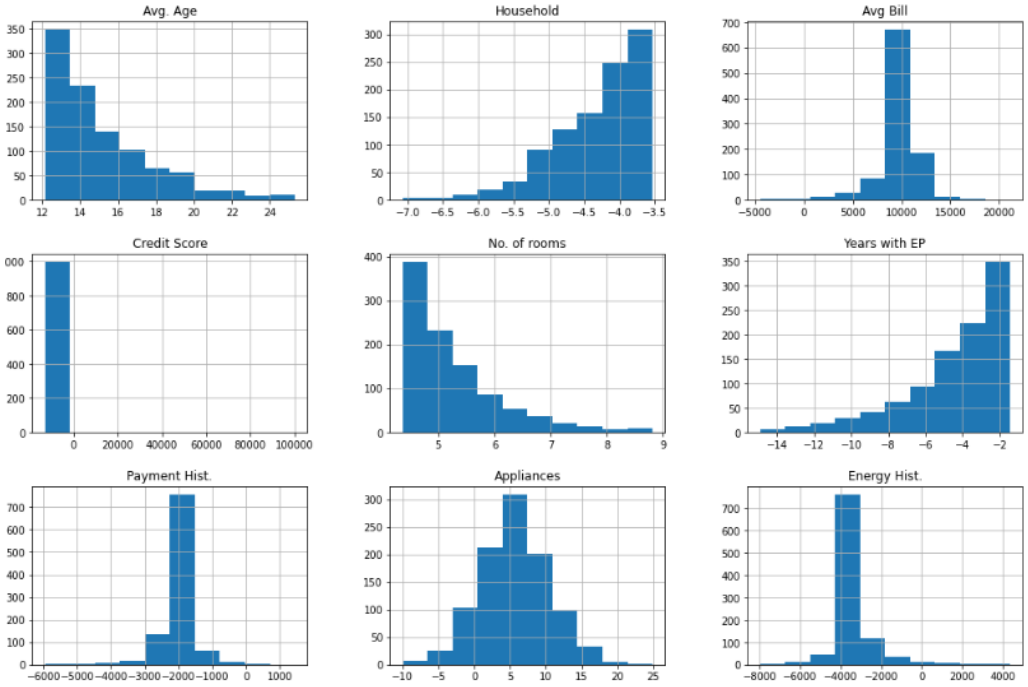
In line with the titles, my argument on this will be that since this is a historical data, the negatives, could be indicating a past information as a present information may not necessarily be the same. e.g. in the credit score column, a customer's credit score as of March 2022 will not be the same as in January 2023. Personally, this analysis helped made more meaning to why the data are the way, they are.

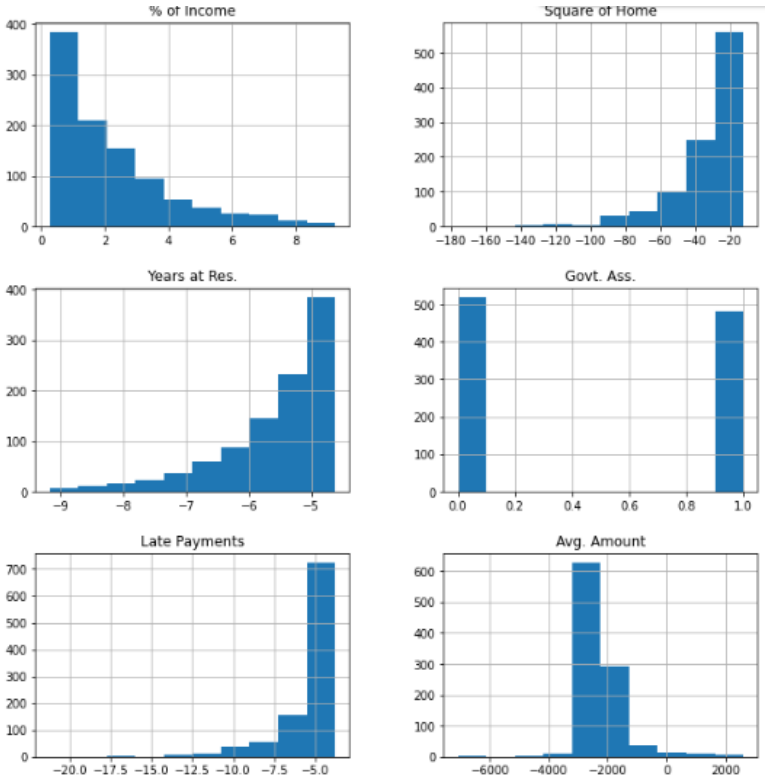
Hence, resulting to the following renaming of F1- F22 respectively. In the code, the columns were titled according to the attached abbreviations.

	FEATURE	MEASUREMENT	COLUMN TITLE
F1	Average Age of Household	Scale of 10 -50	Avg. Age
F2	Number of Household	Scale of -6 to 0	Household
F3	Average monthly electricity bill	Scale of 1,000-20,000	Avg. Bill
F4	Percentage of income spent on energy bills	Scale of 1 – 20%	% of Income
F5	Total square footage of home	Scale of -100 to -10	Square of Home
F6	Credit score	Scale of -10000 to – 1000	Credit Score
F7	Number of rooms in home	Scale of 1-10	No of rooms
F8	Number of years with current energy provider	Scale of -10 to -1	Years with EP
F9	Number of years at current residence	Scale of -10 to -1	Years at Res
F10	Access to government assistance	Scale of 0 or 1	Govt. Ass
F11	Previous payment history	Scale of -10000 to -500 in currency	Payment Hist.
F12	Number of Electrical appliances	Scale of 0 - 20	Appliances
F13	Energy consumption history	Scale of -5000 to 0	Energy Hist.
F14	Number of late payments in the past year	Scale of -10 to 0	Late Payments
F15	Average Monthly payment made	Scale of -10000 to 0 in currency	Avg. Amount
F16	Home ownership status	Scale of 0 or 1	Ownership
F17	Number of complaints to the energy provider	Scale of 0 to 20	Complaints
F18	Average Daily energy consumption	Scale of -1000 to 5000	Daily Consumption
F19	Number of energy-efficient upgrades made	Scale of 0 - 5	Upgrades
F20	Monthly income	Scale of 1000-2000 Currency	Income
F21	Home energy efficiency rating	Scale of 0 - 20	Rating
F22	Class	True or False	Prediction

Table 1.1 ***Identification and measurement of Data Columns***

A histogram distribution of the features as renamed above were presented as shown below using plot from matplotlib and histogram from seaborn.





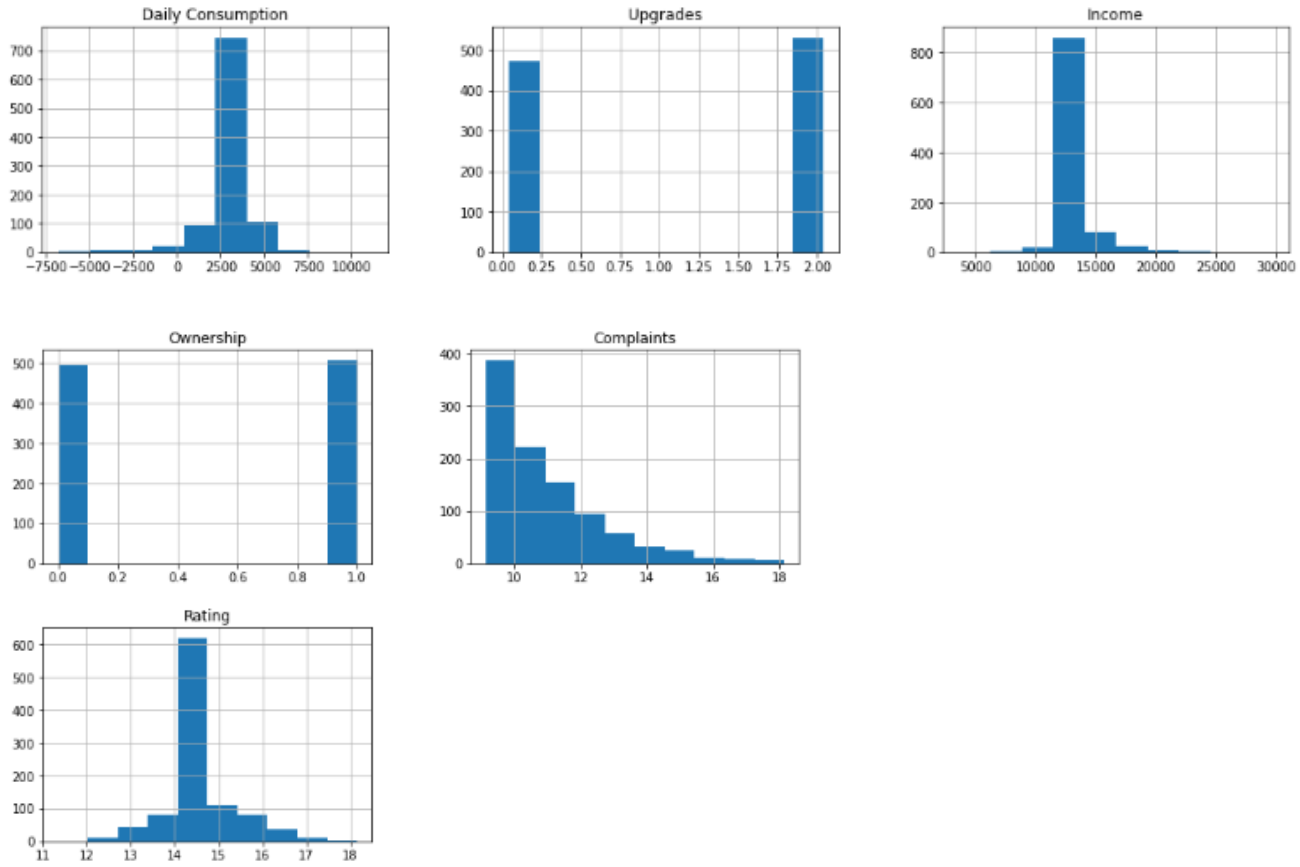


Fig 1.1 ***Chart representation of each input features***

Also, the correlation of each column to another, was demonstrated using a heat map as shown below:

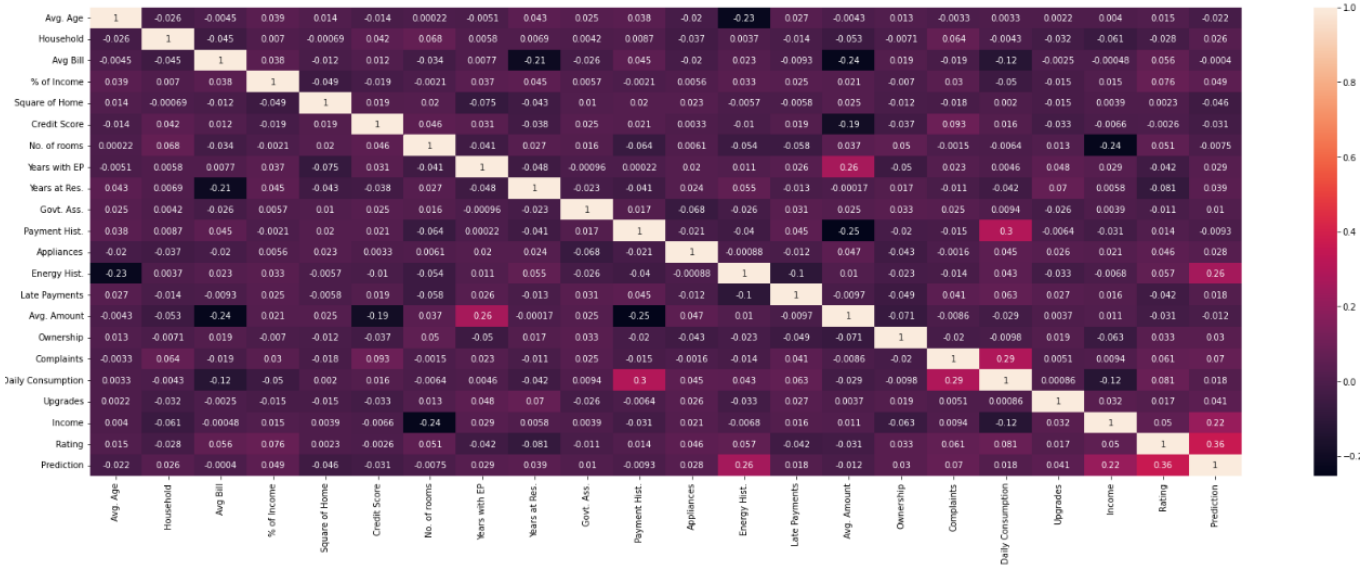


Fig 1.2 ***Heat map representation of Features' correlation***

Data Pre-Processing

The information on the data frame indicated that there were 500 missing values out of 1000 rows of the ‘Rating’ column. To handle this, three experiments were carried out; where experiment 1 was dropping the column entirely, experiment 2 was filling the missing values with the mean and experiment 3 filling with the median. Individually these experiments were tried on the 3 listed models used (DTC, KNN, and RFM) and their accuracy score helped determine which of the imputation method was to be used. The median gave the highest accuracy score on all models, hence was adopted.

The outliers present were tackled by normalizing the data within the range of 0 to 1 using the min max scaler formula.

The ‘prediction’ column data were then replaced from True to 1 and False to zero using a Label Encoder. For experimentation and model confirmation purposes, the data was split to two using the train test split function resulting to a mini-test data of 20% of the original train data.

Model Implementation

Three models were employed to explore the accuracy of prediction of the input data; Decision Tree Classifier, K –Nearest Model and Random Forest Model. They were all fitted to the train data and predicted on the mini test data, which led to 84%, 64%, and 84% accuracies respectively, alongside a classification report indicating the f1, recall and precision scores.

The best model was then selected to be Decision Tree Classifier because it can easily learn a series of decision rules based on the available features in order to predict the class label. So even if given an unknown data, prediction will be close to 100%.

Evaluating Performance

Importing the attached original test data, the model was then tested on the data after the columns must have been renamed to match the train data and the missing values of the Rating Colum of the test data, filled with the median number. The Predicted values were then attached to a newly generated csv file.

Conclusion

In conclusion, the utilization of decision tree model in forecasting payment challenges for customers has proven to be an effective approach and one of the best model for predicting if customers of AENERGY were going to encounter difficulties in paying the increasing electricity cost based on their historical data. The model achieved high accuracy, precision and recall scores. The results of this analysis can help AENERGY in taking necessary actions to mitigate the risk by providing support to these customers. In future research, other models and incorporating more customer data could be considered to improve the prediction.

PART 2

Predicting the variation in the annual expenditure of customers for future years

Goal: The purpose of developing this machine model is to be able to predict the variation in the annual expenditure that a customer is going to have due to the increase of the energy cost in pounds per year.

Exploratory Analysis

The train data were all explored using different plots:

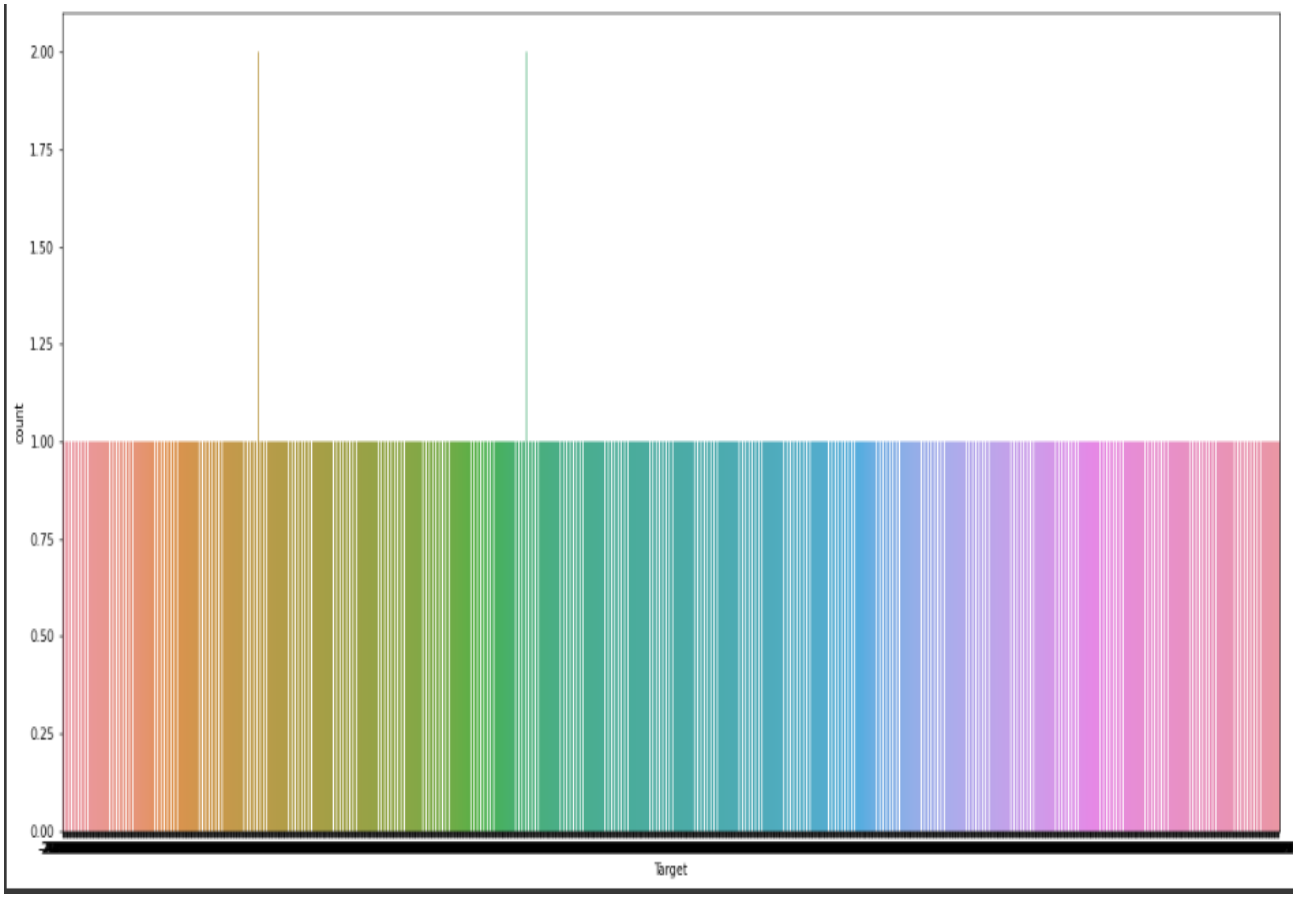


Fig 2.1 ***Plot of the Values found in the Target Data***

Because the first plot was not explanatory enough so a histogram was built as shown below:

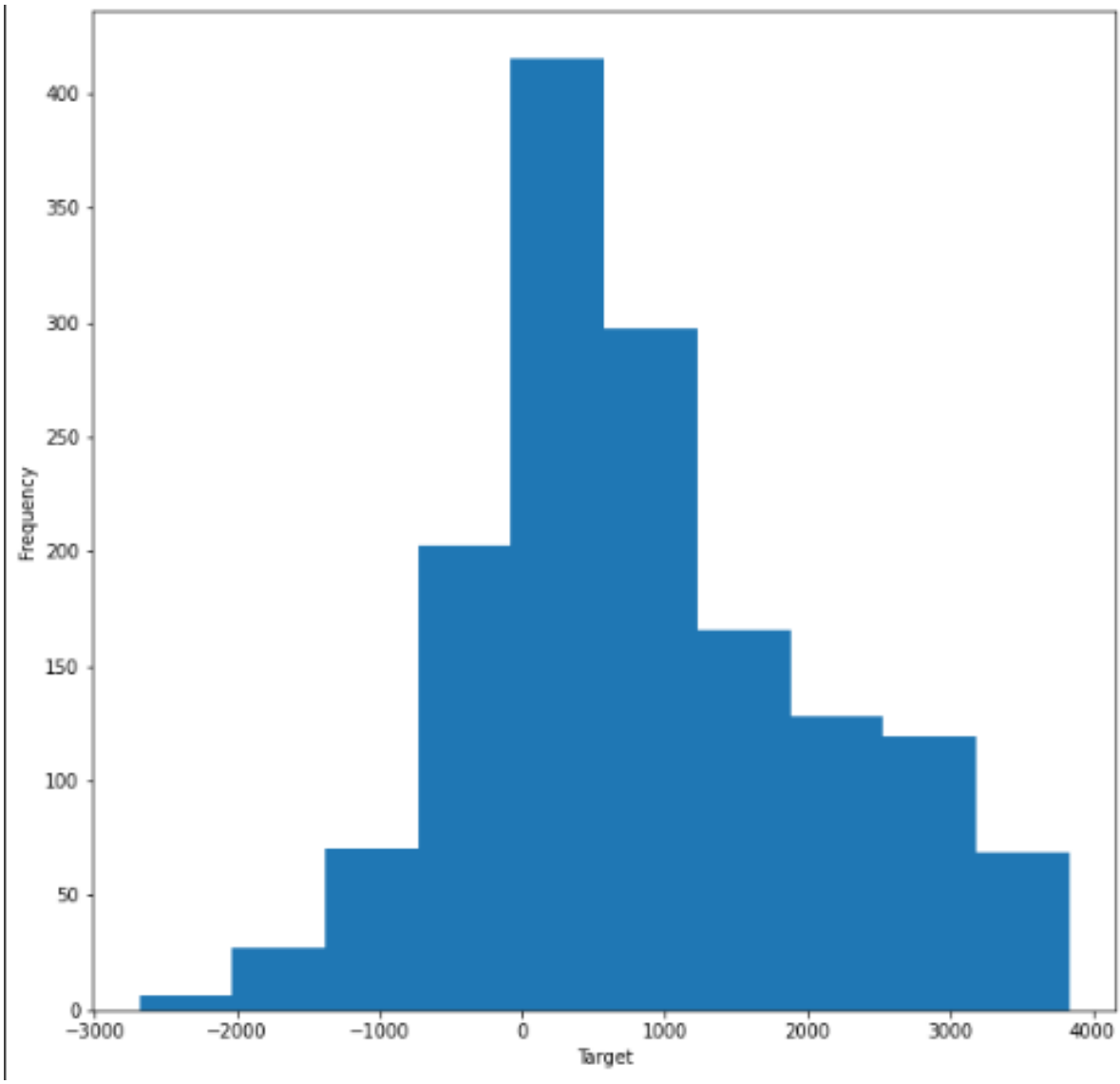


Fig 2.2 ***Histogram of the Values found in the Target Data***

The outliers showed that some columns had data out of range, this was reduced using normalisation to enable proper fitting into the model.

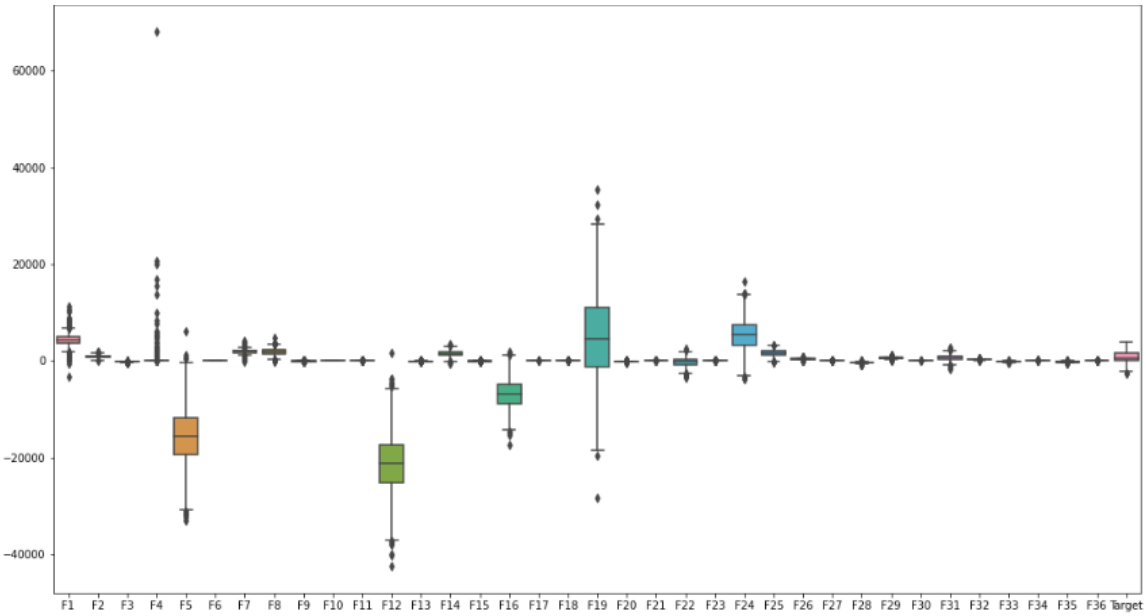


Fig 2.2 ***Boxplot of the Outliers found in the data***

Data Pre-processing

There weren't any missing values found in the data, however column F6 and F10, were having strings i.e. object type. To handle F6, the strings were replaced with ascending numbers since they were all in increasing and decreasing orders, while F10 were just replaced with real numbers beginning from 1 to identify each country. The outliers were brought into range using normalization of 0 to 1.A train test split was also carried out within the range of 80 and 20, to fit and test the model before experimenting on the original test data.

Model Implementation

Being a regression problem, the models used were all regressors namely; Linear Regression Model, Support Vector Regression Model, and Random Forest Regression Model. They were all measured using the Mean Square error metrics which generated outputs of 0.0104, 0.0100 and 0.0105 respectively. The support Vector regressor proved to be the best model having the lesser value of error. It was then used on the test data as provided.

Conclusion

In conclusion, the results of the analysis showed that the support vector regressor was able to make almost accurate predictions of the variation in the annual expenditure of customers. The findings of this analysis have important implications for AENERGY as they can use these predictions to plan for future expenses, and also to develop strategies to mitigate the financial impact of increased energy costs on their customers.

Appendices

[All generated from the attached code files]

Table 1.1 Identification and measurement of Data Columns

Fig. 1.1 Chart representation of each input features

Figure 1.2 Heat map representation of Features' Relationship

Figure 2.1 Plot of the Values found in the Target Data

Figure 2.2 Histogram of the Values found in the Target Data

Figure 2.3 Boxplot of the Outliers found in the data