Title: A Comparative Study of Variable Selection Techniques for Predictive Analytics of Household Appliance Energy Usage

Abstract

This study "*A Comparative Study of Variable Selection Techniques for Predictive Analytics of Household Appliance Energy Usage*" aims to improve the management of household energy consumption by comparing various variable selection techniques and suggest the accurate feature selection technique for predictive analytics of appliance energy usage. The study uses data from the UCI Machine Learning Repository, conducts exploratory data analysis, and applies multiple feature selection methods such as dimensionality reduction, mutual information filtering, and wrapper methods like Recursive Feature Elimination with Cross-Validation, Boruta and stepwise forward regression. A linear regression model is then built using the selected variables, and its performance is evaluated using various metrics. The findings provide valuable insights for researchers in data analytics and serve as a case study for exploring different variable selection techniques. The study concludes among all used techniques that Boruta was the most accurate technique in predicting household appliance energy usage, which can help households accurately forecast their energy consumption and optimize appliance usage to reduce energy bills. The study suggest that future studies should explore other machine learning techniques and models to improve the accuracy of energy usage predictions.

I. Problem definition

Energy consumption is an ever-increasing concern for households, with the cost of energy continuing to rise. Low-energy homes have become increasingly popular, but households still struggle to manage their energy usage efficiently. One of the major contributors to energy consumption in homes is household appliances. However, determining the optimal time to use appliances is a challenge that households face, which often leads to higher energy bills.

To address this problem, the project entitled " ***A Comparative Study of Variable Selection Techniques for Predictive Analytics of Household Appliance Energy Usage*** " aims to compare different variable selection techniques to determine the most effective approach that determines most suitable variables that accurately helps in predicting household appliance energy usage. The accuracy of the predictive models generated using different variable selection techniques will be tested and compared. The study aims to provide insights into the effectiveness of different variable selection techniques and their impact on the accuracy of predictive models for household appliance energy usage.

Researchers in the field of data analytics can use the insights and techniques developed in this work to improve their own predictive modelling projects. The project can also serve as a valuable case study for exploring different variable selection techniques and their impact on the accuracy of predictive models.

II. Methodology

Data collection

The data used in this study was obtained from the UCI Machine Learning Repository, a compilation of databases, theories, and data generators utilized by the machine learning community [1]. It comprises temperature and humidity readings of a house that were monitored by a ZigBee wireless sensor network. The wireless data was collected in Realtime and then averaged at 10-minute intervals. Real-time energy usage was monitored using M-bus energy meters and recorded in 10-minute intervals. In addition, the dataset contains weather data obtained from a nearby airport weather station. The weather data was merged with the energy, temperature, and humidity data using the date and time column.

Data analysis

1. Exploratory data analysis

The purpose of conducting an exploratory analysis on the dataset was to understand its characteristics and gain insights into its structure, patterns, and relationships between variables. To achieve this, various aspects of the dataset were examined, including its length, attributes, and data types. Additionally, an evaluation was conducted to identify any data quality issues that may impact the analysis. A correlation matrix was computed to investigate patterns and relationships, and interactive visualizations were created to aid in the analysis.

2. Data pre-processing

The process of feature selection is a crucial step in the data pre-processing phase as it can greatly influence the performance of machine learning models [2]. To identify the most relevant variables that can improve the accuracy of energy usage predictions, several techniques were employed. These techniques include dimensionality reduction via principal component analysis, filtering through mutual information, and wrapper methods such as stepwise forward regression, Boruta, and recursive feature elimination with cross-validation. Through a comparative analysis of these techniques, the optimal method was identified and used to select the variables that provide the highest accuracy in energy usage predictions.

3. Model development and evaluation

Due to its simplicity, flexibility and interpretability, a linear regression model was built on several different scenario and it was used to predict the energy usage [3]. In fact, it was build using all variables given in the data sets, it was also built using variables from dimensionality reduction, using the variables from mutual information, as well as using variables from wrapper methods that were used. Standard steps for linear regression modelling were taken, including splitting the data into 70% training and 30% testing sets, model fitting, energy usage prediction, and model evaluation. Performance of the model was assessed in each scenario by measuring various metrics such as mean absolute error and R-squared on the test set.

III. Results:

The preliminary investigation was performed, and it revealed that the dataset contains 19735 observations and 29 attributes. Moreover, after calculating the correlation matrix as presented in Figure 1 below, it was discovered that there is a weak or low linear relationship between the independent variables and the dependent variable. Additionally, there were two columns that contained random variables which exhibited minimal association with the other feature variables in the data set.

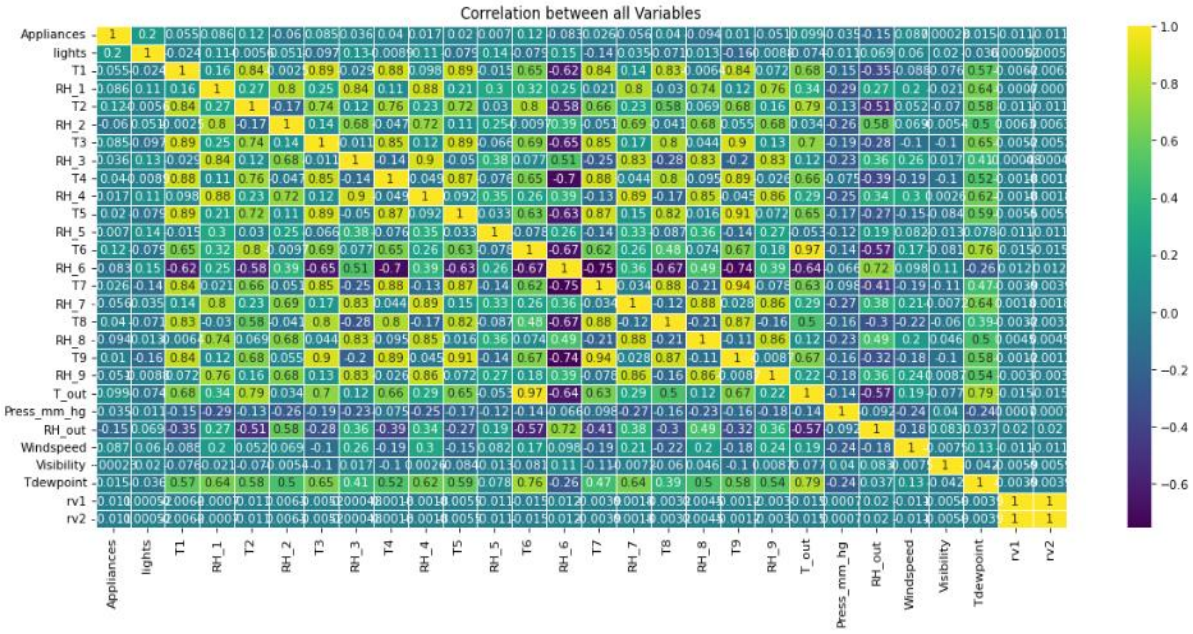

Figure 1: The correlation matrix of all variables in a data set

Table 1 below summarizes the metrics performance of the linear regression model built using feature variables resulted from different data pre-processing techniques. The results revealed that Boruta, a machine learning approach for feature selection, was the most effective technique in improving the model's performance. In contrast, the dimensionality reduction technique implemented, specifically PCA, resulted in poorer performance compared to other approaches.

Table 1:The metrics performance of a linear regression model using feature variables obtained from different feature selection techniques

	MAE	**R-squared**
All variables	53.7692	0.1605
Dimensionality reduction technique		
PCA	56.1549	0.0922
Feature selection techniques		
Stepwise forward regression	53.7985	0.1609
Boruta	53.6746	0.1612
RFECV	53.9653	0.1570
Mutual information	53.8806	0.1587

IV. Discussion and reflection:

The study aimed to identify the best variable selection technique to predict household appliance energy usage. After employing multiple techniques and evaluating the results in as shown in Table 1, It was found that Boruta was the most accurate technique. Using Boruta can help households in selecting variables that accurately predict their energy usage and determine the most suitable time to use their appliances, ultimately reducing their energy bills.

Boruta identifies the most relevant variables for prediction by comparing the importance of each variable to a set of random variables with the same distribution [4, 5]. It is known for its ability to handle complex, high-dimensional datasets, and it is less prone to overfitting compared to other methods which is the reason why I think it was the best technique [4].

One limitation of this study is that it only focused on one type of machine learning model (linear regression). Other models, such as decision trees, random forests, and neural networks, could have been used to compare the effectiveness of different variable selection techniques.

V. Conclusion

This work aimed to identify the most effective variable selection technique for predicting household appliance energy usage to help households manage their energy consumption efficiently. The study used multiple variable selection techniques and found that Boruta was the most accurate technique in predicting energy usage.

The study's results can be useful to data analytics researchers and serve as a case study for exploring different variable selection techniques and their impact on the accuracy of predictive models. Using the insights and techniques developed in this study, households can identify the most relevant variables to predict their energy usage, thus reducing their energy bills. Further research can explore other machine learning techniques and models to improve the accuracy of energy usage predictions.

References

[1] L. M. Candanedo and L. Candanedoibarra, "Appliances energy prediction Data Set," UCI Machine Learning Repository, [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Appliances+energy+prediction. [Accessed 01 April 2023].

[2] A. Subasi, "Data preprocessing," in *Practical \Machine Learning for Data Analysis Using Python*, Jeddah, Academic Press, 2020, pp. 27-89.

[3] Y. Chen, M. Guo, Z. Chen, Z. Chen and Y. ji, "Physical energy and data-driven models in building energy prediction: A review," *Energy Reports,* vol. 8, pp. 2656-2671, 2022.

[4] D. Bhalla, "FEATURE SELECTION : SELECT IMPORTANT VARIABLES WITH BORUTA PACKAGE," Listen Data, [Online]. Available: https://www.listendata.com/2017/05/feature-selection-boruta-package.html. [Accessed 01 April 2023].

[5] L. M. Candanedo, V. Feldheim and D. Deramaix, "Data driven prediction models of energy use of appliances in a low-energy house," *Energy and Buildings,* vol. 140, pp. 81-97, 2017.