Topic: Data Driven prediction of model of energy use of appliances

Research Paper 1

#Data Source:

1. Data from home-
2. sensor data from home - Temp, Humidity
3. components of energy consumption: Appliances, Light
4. Data from weather: recorded from nearest Airport

Data format:

1. Time series data collected over a period of 6 months
2. sensor data collected every 10 minutes

Target Parameter:

1. Energy consumption

EDA:

1. Average Electricity consumption and frequency of variety of power consumption
2. Correlation diagram between electricity consumption of appliances, Room Temperature, Relative Humidity are measured.
3. Relationship between indoor temperature and weather conditions outside {Pressure, Temperature, Windspeed}
4. positive correlation energy consumption vs lights, Appliances & T2, Appliances & Outdoor Temp, NSM & Appliances
5. Negative correlation appliances & outdoor humidity

Data Filtering

1. Parameter selection based on which features are improving the accuracy of model.

Boruta Algorithm:

1. Comparison of random parameter (as base) with other parameters
2. Ranks the variables NSM: high ranked, week status- weak

Correlation with RMSE:

1. RFE algorithm applied and dummy variables are introduced for week of day and week status. Number of optimal parameters are 34.

Performance of regression model: {LM, GBM, SVM, Random Forest are used}

1. 10 cross validation are applied for making the model robust.
2. Since the RMSE for linear model are not normally distributed hence is not accurate to use.

Four Parameter used for performance evaluation of model:

1. RMSE

2. R2

3. MAE

4. MAPE

SVR radial kernel - {sigma and cost function are tuning parameters apart from 12 parameters}

Random Forest model - Tree based model

1. Tree based on random set of predictors for removing the correlation between trees & improve the prediction.

2. predicts optimal number of trees (300) - {RMSE stops improving after that}

3. Number of optimal variables for the model calculated as 18.

GBM model:

1. Improve prediction of information from first trees.

2. Require selection of optimal parameter for number of trees(10900) & tree depth (5)

Model Selection:

1. 5 Model from 10 cross validation & 3 repeats.

2. RFE & GBM- similar performance on RMSE & R2.

3. SVR is better than LM.

Variable importance measured by residual sum of squares

GBM is best.

Evaluating further GBM & variable importance:

1. subsets of variables as model variables - RMSE parameter is computed

Research Paper 2:

1.Engineering and hybrid approaches use thermodynamic equations to estimate energy use, the AI-based approach uses historical data to predict future energy use under constraints.

2.Ease of use and adaptability to seek optimal solutions in a rapid manner, the AI-based approach has gained popularity in recent years.

3.Approaches for building energy use prediction, conducts an in-depth review of single AI-based methods such as multiple linear regression, artificial neural networks, and support vector regression, and ensemble prediction method

4.Combining multiple single AI-based prediction models improves the prediction accuracy manifold.

This paper elaborates the principles, applications, advantages and limitations of these AI-based prediction methods and concludes with a discussion on the future directions of the research on AI-based methods for building energy use prediction.

Research Paper3:

Prediction of Appliance use in smart homes:

1. The paper predicts the energy consumption on the next day. Two basic predictors and one stochastic predictor is proposed.

2. Predictors gives better performance than other approaches.

3. Two processing are proposed to improve the prediction, segmentation, aggregation of data.

4. Data was collected from European countries including Central and Eastern European Countries (REMODECE database)

5. Hourly data was analysed for appliances over a year.

6. The performance of the predictor is evaluated using e(h), which is 1 if the appliance is consuming the electricity, else 0. P(h) be the prediction provided by the predictor a , which is equal to 1 if the appliance is actually consuming the electricity else 0. Then with the formula the precision of the predictor is calculated for any time.

Proposed algorithm for assessing a predictor a involves the following steps:

1. Set the time window dimension to n hours within the period for which the historical data was registered where n goes from 24 to 364 * 24;

2. Compute the predictions for the data corresponding to the historical sliding time window;

3. Compute the predictor precision pa(h) based on the “next day” data for all possible hour’s h and compute an average precision for the predictor.

Prediction with different predictors:

1. "Will Always consume" and "Will never consume predictors"

It refers to the probability of the appliance will consume and vice versa

ARMA predictor:

1. In the algorithm the current value of a time variable is made a function of its past values and is expresses as sum of weights.

2. This ARMA model was used in order to predict the next day energy consumption.

Proposed Predictor:

1. Proposed predictor specifies the probability of the appliance to consume on an hourly base

Improving the Prediction precision:

1. temporal segmentation, that considers each day of the week as a partition was done.

2. the hourly predictions are made considering the proposed predictor. A k-Means clustering algorithm is applied in order to group the similar consumption days.

3. Each cluster is defined by its cluster centre and clustering proceeds by assigning each of the input data to the cluster with the closest centre.

Prediction precision after clustering:

1. After applying the iterative k-Means algorithm, two clusters are obtained. In the presented case, cluster C1 groups weekdays data and cluster C2 gathers Saturday and Sunday data

Part 2: EDA:

Exploratory analysis was done on the data set to find out the relationship of the variables with the appliances consumption and key insights were evaluated.

1. The plots 1 & 2 shows marked peak in energy consumption during day.
2. The Plot 2 shows the average energy consumption is below the 50 Wh.
3. The plot 2 shows the average energy consumption is less than ¼ the peak energy consumption during the day.
4. The plot 2 shows high average consumption on Sunday marked by date 14th Feb., as approximated by the fact most the members of the family stays at home.
5. The Fig. 3 shows the frequency of binned energy consumption over a period of six months.
6. The plot is skewed towards left and shows the frequency of energy consumption is high for 50 Wh which validates the findings in the plot 1 and plot2.
7. The Box plot shows the energy over the period of six months.
8. The maximum energy consumption is little less then 200Wh and minimum is more than zero.
9. Average consumption occurs approximately at 50 Wh.
10. The outliers show the energy consumption has peaked many times in a day.
11. The correlation diag. shows the relationship of the energy consumption of various components of the house recorded over a period of six months.
12. The plot shows the positive correlation between energy consumption of lights and Appliances.
13. Positive correlation is high i.e. 0.19 between indoor Temperature (T2) and Appliances energy consumption.
14. Positive correlation is expected between T1 and T3.
15. Plot shows highest correlation between outdoor temperature and Appliances energy consumption.
16. Negative correlation exists between Appliances and outdoor humidity RH6.
17. positive correlations between the consumption of appliances and T7, T8 and T9 being 0.03, 0.05 and 0.02 respectively.
18. Highest correlation between the energy consumption of appliances and NSM with a value of 0.22.

Part 3: Feature engineering

Data Clean:

The data is clean and need no cleaning as shown by the below information.

```
Data columns (total 28 columns):
Appliances      19735 non-null int64
lights          19735 non-null int64
T1              19735 non-null float64
RH_1            19735 non-null float64
T2              19735 non-null float64
RH_2            19735 non-null float64
T3              19735 non-null float64
RH_3            19735 non-null float64
T4              19735 non-null float64
RH_4            19735 non-null float64
T5              19735 non-null float64
RH_5            19735 non-null float64
T6              19735 non-null float64
RH_6            19735 non-null float64
T7              19735 non-null float64
RH_7            19735 non-null float64
T8              19735 non-null float64
RH_8            19735 non-null float64
T9              19735 non-null float64
RH_9            19735 non-null float64
T_out           19735 non-null float64
Press_mm_hg     19735 non-null float64
RH_out          19735 non-null float64
Windspeed       19735 non-null float64
Visibility      19735 non-null float64
Tdewpoint       19735 non-null float64
rv1             19735 non-null float64
rv2             19735 non-null float64
dtypes: float64(26), int64(2)
```

Data Transformation

Add the columns many of columns such as week of day - Sunday, Monday, Tuesday, Wednesday, Thursday, Friday, Saturday to denote the day of the week status - with attributes such as weekends, weekdays to signify the day is weekends or weekdays. NSM - to denote the number of seconds from the midnight of a day

Features significance are measured by Correlation of Appliances with other variables:

1. Plot shows highest correlation between outdoor temperature and Appliances energy consumption.

2. Negative correlation exists between Appliances and outdoor humidity RH6.

3. Positive correlations between the consumption of appliances and T7, T8 and T9 being 0.03, 0.05 and 0.02 respectively.

4. Highest correlation between the energy consumption of appliances and NSM with a value of 0.22.

5. The correlation diag. shows the relationship of the energy consumption of various components of the house recorded over a period of six months.

6. The plot show the positive correlation between energy consumption of lights and Appliances.

7. Positive correlation is high i.e. 0.19 between indoor Temperature (T2) and Appliances energy consumption.

8. Positive correlation is expected between T1 and T3.

9. Plot shows highest correlation between outdoor temperature and Appliances energy consumption.

10. Negative correlation exists between Appliances and outdoor humidity RH6.

11. positive correlations between the consumption of appliances and T7, T8 and T9 being 0.03, 0.05 and 0.02 respectively.

12. Highest correlation between the energy consumption of appliances and NSM with a value of 0.22.

Pre-processing is required as the data is measured with different scales

Part 4 Prediction algorithms

After we see four different prediction model, we compute RMS, MAPE,R2 and MAE for each of them. We find out the best model is using the random forest technique.

Part 5 Feature Selection

compare and use all feature selection method

1. removing features with low variance
    a. VarianceThreshold is a simple baseline approach to feature selection. It removes all features whose variance doesn't meet some threshold.

    b. use VarianceThreshold method to remove low variance features, as all features have high variance, this method not fits this model
2. univariate feature selection
    a. Univariate feature selection works by selecting the best features based on univariate statistical tests.

    b. only f_regression and mutual_info_regression score functions are available, not quite fit the random tree model
3. recursive feature elimination
    a. select features by recursively considering smaller and smaller sets of features. First, the estimator is trained on the initial set of features and the importance of each feature is obtained either through a coef_ attribute or through a feature_importances_ attribute. Then, the least important features are pruned from current set of features.That procedure is recursively repeated on the pruned set until the desired number of features to select is eventually reached.

    b. a good choice

4. Feature selection using SelectFromModel
    a. SelectFromModel is a meta-transformer that can be used along with any estimator that has a coef_ or feature_importances_ attribute after fitting. The features are considered unimportant and removed, if the corresponding coef_ or feature_importances_ values are below the provided threshold parameter.

    b. a good choice

5. Exhaustive search, Forward search and Backward search

    a. forward search and backward search are unstable since some features have similar importance scores, so they are easy to select different variables, especially from the start or end in the feature list

    b. exhaustive is a good choice but has long run time

tpot

- tpot automl select the features for you and try out different method to build up model. However, it takes a really long time to build the model. We got a Best pipeline: ExtraTreesRegressor(input_matrix, bootstrap=False, max_features=0.7000000000000001, min_samples_leaf=1, min_samples_split=7, n_estimators=100)
- -4660.639446960434
-

Boruta

- boruta is a feature selection tool and it picks the best features for you, it's not as time consuming as tpot, but the outcome is not good as tpot's. It choose two top features

tsfresh

- tsfresh is also a feature selection tool and it takes a long time to run as well

Part 6 Model Validation and Selection

validate random forest regression model from these contexts:

1. cross validation techniques
    a. do a sequence of fits where each subset of the data is used both as a training set and as a validation set.

    b. useful validation approach and always combine with other approaches
2. Bias-variance tradeoff
    a. Scikit-Learn offers a convenient utility for computing such learning curves

    b. As dataset is too large and has many 'outliers', the training error and validation error curves don't fit well

3. Regularization

    a. instead of L1, L2, and elastic net these regularization parameters, we choose oob_score to analyze the accuracy
    b. a good approach, find relationship easily
4. Grid search options
    a. use grid search with cross validation, choose GridSearchCV method, give parameter grid, to choose best parameters automatically
    b. very good choice, but need to set a good parameters grid to fit the model and dataset

Part 7 Final pipeline

In conclusion, provide full pipeline to select the best model for this regression problem. From feature engineering, to prediction algorithms, to feature selection and model validation, to the final model select. Create whole workflow process of machine learning.