2018

CS5014 – MACHINE LEARNING – P1 – PREDICTING ENERGY USE OF APPLIANCES

UNIVERSITY OF ST ANDREWS

SIDNEY OFOSU-OTTOPAHSTUDENT ID: 170027939

Table of Contents

Introduction

The report documents the set of procedures used undertaking the first Machine Learning practical where data, after being cleaned clean and processed, is used to develop a regression model to be used to predict an output i.e. he Power Consumption of an appliance, based on a set of inputs.

Aims and Objectives

The aim of the practical was to develop and reinforce knowledge and experience gained from working and using several machine learning skills on 'real' data while creating and developing a regression model.

The objectives were as follows:

- Loading and cleaning the data.
- Analysing and visualising the data.
- Preparing the inputs and choosing a suitable subset of features.
- Selecting and training a regression model.
- Evaluating the performance of the model.
- Critical discussion of the results and your approach.

Tasks Achieved

Data Description

The data used in this practical was that collected in (Candanedo, et al., 2017). The paper, a report, which described and documents an experiment, aimed at understanding "the relationships between appliances energy consumption and different predictors", explains that the report was conducted on a house which served as a data collection point from which several types of data regarding its (i.e. the house's) energy consumption across its rooms and outside of it were gathered.

The energy data, collected over five months from January 2016 through to mid-May 2016, (across Bedrooms, Laundry room, Garage, Dining room, Kitchen, Ironing room and Bathrooms), included the following attributes (Candanedo, et al., 2017):

- Temperature in degree Celsius
- Humidity as a percentage
- Date and Time on recording the (single) data set
- Energy Consumption in Kilowatt-hours
- Pressure in millimetres of mercury
- Wind speed in meters per second
- Visibility in Kilometres and
- Temperature-Dew Point in degree Celsius

A Table with a list of the data collected and used along with the definitions and the units has been attached as Appendix 2.

Data Loading and Cleaning

The data, made available as a CSV[1] file, was loaded into the python (script) by using the Pandas[2] Data Analysis Library and stored as a Pandas data-frame. This decision was made for flexibility and to allow for easier data manipulation.

Initially, not much cleaning was done, with the reason being that, an initial correlation matrix was generated (in MS[3] Excel) to find out how the several attributes affected (if at all), any other attributes. This also severed as a means of pre-determining the attributes would therefore be useful for training the model besides others that wouldn't be.

The names of series with the CSV file was also changed to easily recognisable ones to enable easier referencing within the python source code.

Instead, a glance through each row in the CSV was made to check for any irregularities, inconsistencies and/or null cells, which revealed the following:

- No empty cells were found.
- The date was concatenated with the time into one cell as a string
- There were unknown attributes: Random Variable 1 and 2

Some of the above issues were later addressed in the Data preparation and Pruning section of this report.

Particularly, it was found that the Random Variable 1 and Random Variable 2 had almost little to no correlation, relatively to the other series data recorded. This was taken into consideration and the two-data series were removed at the pruning stage. The resulting Correlation matrix from this data analysis has been attached as Appendix 1.

Data Analysis and Visualisation

For this part of the assignment, a pair(matrix) plot was generated using the Seaborn[4] python package. Other scatter diagrams were also generated with the Mat-Plot-Lib libraries[5].

Analysis 1

A pair plot matrix, showing the visual relationships between the Average Indoor Humidity, Average Indoor Temperature, Light Consumption and Appliance Consumption was plotted to initially characterise the data. The resulting chart has been attached as Figure 1 on page x.

All of the 'room' data were (for temperature and Humidity percentage) were averaged and plotted (as TempHouseAVG and HumHouseAVG). This was done because indoor temperature and humidity recordings, in the individual rooms, particularly the kitchen had some of the highest correlation with the Lights and Appliance Consumption data recorded.

[1] Comma-Separated Value
[2] Pandas - https://pandas.pydata.org
[3] Microsoft Excel
[4] Seaborn - https://seaborn.pydata.org/
[5] Mat-Plot-Lib - https://matplotlib.org

From the chart, the following was noticed:

- Humidity and Temperature averages increased with increases with their corresponding Lights and Appliance consumption
- The increase however was noticed to kind of tape cut off, i.e. was not strictly linear and was more of a scatter than following a single trend line.

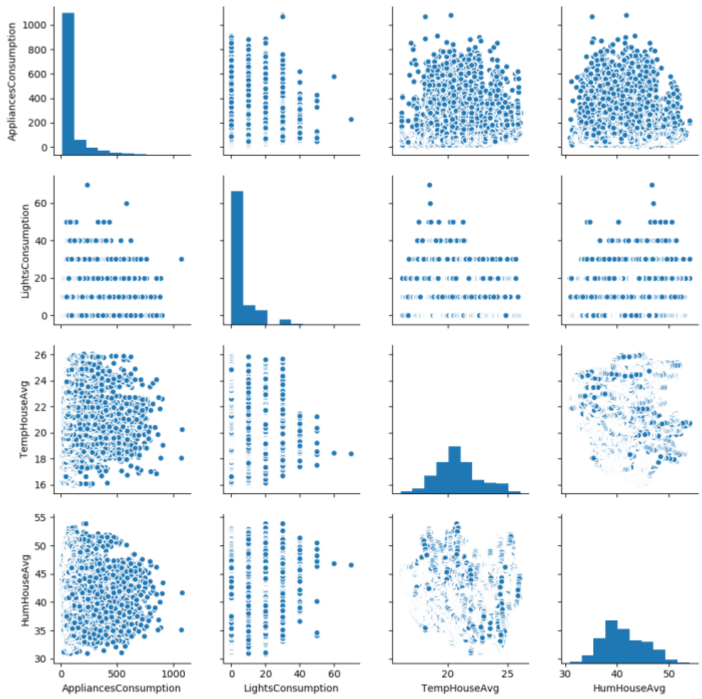

Figure 1 – Pair Plot [Light and Appliance Consumption V Average Temperature and Humidity]

Analysis 2

A closer inspection of the Temperature average and its relationship with the Appliance and lights consumption was made. The following four scatter charts were plotted to determine this relationship: Temperature V Lights, Temperature V Appliances, Humidity V Lights and Humidity V Appliances.
These charts have been attached as Figure 2.

- The trend lines (i.e. the red-coloured line) for each, stayed almost consistently level indicating the even though they influenced the results of Lights and Appliance Consumption, the changes were fairly small.
- Temperature V Appliance trend line had a similar gradient to that of Humidity V Lights, with the trend lines having a similarly positive gradient.
- Temperature had a constant trend line with Light consumption, suggesting that irrespective of how hot or cold it was, the need for switching on more lights wouldn't necessarily change in such a household.
- Similarly, as above with Humidity V Appliance, the trend line, even though has a slightly lower (negative) gradient compared to the above, suggests that Appliance consumption reduces as humidity percentages in the house increases.

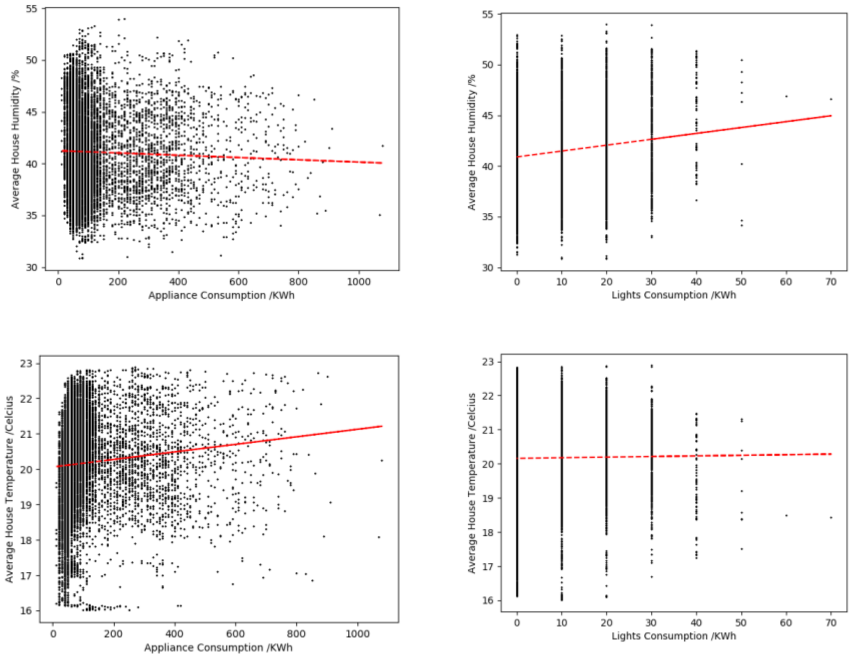

Figure 2 – Scatter Charts [Light and Appliance Consumption V Average Temperature and Humidity]

Data Preparation and Pruning

Pruning

The series data which were deemed somewhat not as impactful on the results of the experiments or otherwise wouldn't serve much purpose to the experiments and the purpose of this practical, were removed. The series removed were Random Variable 1 and Random Variable 2.

Another data series which was also removed was the Date series. This was however not because it had no relevance, but instead because it was in an undesirable format that, was difficult to integrate into the learning algorithm, due to format conversion complications. The dates were made available and separated but weren't used in the end.

The remaining data was then split into their several series manually which meant that, I had maximum flexibility in making subsets of the data set for training and testing the algorithm

Splitting into Training, Validation and Test sets

The Training Data set for all series was chosen to start from the index 1 to 16000 (actually 0-15999 in python). The Testing data was the remaining data from the 1600 index to the end (i.e. 19735).
The validation data was chosen to be a subset of the Training data, from the 14000 indices to the 16000 indices.

The Validation data is only used to check and reassure the learning process, whereas the testing data is used to actually evaluate the algorithm.

Choosing Subsets

Keeping the above Training, Validation and Test Splits, the following subsets containing their respectively indicated sub series, were made and used to develop the algorithm.

Subset 1(Full):
Lights Consumption, Humidity and Temperatures of all rooms (i.e. Kitchen, Living Room, Parents Room, Parent's room, Bathroom, Laundry, Teenage room, Ironing rooms), Outside Temperature, Dew point, Visibility, Pressure, Outside Humidity, Humidity from Weather Station, Wind Speed and Temperature from the Weather station

Subset 2(Averages):
Lights Consumption, Average Temperature, Average Humidity, Outside Temperature, Dew point, Visibility, Pressure, Outside Humidity, Humidity from Weather Station, Wind Speed and Temperature from the Weather station

Subset 3(Indoor Factors only with Light Consumption):
Lights Consumption, Humidity and Temperatures of all rooms (i.e. Kitchen, Living Room, Parents Room, Parent's room, Bathroom, Laundry, Teenage room, Ironing rooms),

Subset 4(Outdoor Factors Only Without Light Consumption):
Outside Temperature, Dew point, Visibility, Pressure, Outside Humidity, Humidity from Weather Station, Wind Speed and Temperature from the Weather station

Model Training: Linear Regression Model

A linear Regression model was used to train the data on the four subsets; Subset 1, Subset 2, Subset 3 and Subset 4. The results were then printed out after the training.

Results

The following results were obtained:

	Mean Squared Error	**R-SQUARED**	**MEAN ABSOLUTE ERROR**	**MEAN ABSOLUTE ERROR %**
Subset 1	3.116	1	3.202	3.748
Subset 2	8.703	1	1.608	1.77
Subset 3	3.194	1	3.2869	3.905
Subset 4	N/A	N/A	N/A	N/A

Figure 3 – Table of Results from Linear Regression Model Training and Testing

Observations

- It was noticed that the R-Squared was not actually changing
- The subset 4 could not train because there was a non-number or an infinite number which threw up and error in compilation
- The lowest Mean absolute error was from training with Subset 2, though it had the highest Means Squared error.
- Subset 1 and 3 were close in their errors, resulting in 3.75% and 3.91% approximately respectively.

Evaluation of Model

- From the Observations above, the best data set in terms of error is suggested to be Subset 2, in which the averages of room temperatures were used instead on the individual room data readings. This could mean that the data was too much to actually fit, but could also mean

- Subset 2's model predicted a respectable 98.23% accurately according to the absolute mean error figures.

- The failure in computing Subset 4 proves that there were some over sights while cleaning where the invalid variable has found its way through.

- The values of r-squared would help shed more light on the levels of correlation between the two sets of data and how close to the model's predicted trend line the test data was, for each situation but unfortunately, the deadline was.

Conclusion

The data set proved to be very interesting in that it provided realistic 'everyday' data set and gave the experience of training a linear regression algorithm.

Unfortunately, with limited time, further investigation was made into the situation. A solution would be to have a method iterate through the cells in each series and check for the values, null, or invalid or otherwise not welcome in the dataset.

Bibliography

Candanedo, L. M., Feldheim, V. & Deramaix, D., 2017. Data driven prediction models of energy use of appliances in a low-energy house. *Energy and Buildings,* Volume 140, pp. 81-87.

Appendix 1

CORRELATION MATRIX OF ENERGY DATA

	Date	ancesConsum	htsConsumpt	TempKitchen	HumKitchen	mpLivingRoo	umLivingRoo	TempLaundry	HumLaundry	TempOffice	HumOffice	empBathroo	lumBathroom	npOusideNo	mOutsideNo	empIronRoon	lumIronRoom	pTeenageRcn	TeenageRom	pParentRoo	mParentRoo	mpFromStati	sureFromSta	idityFromSt	speedFromSt	ilityFromSta	Tdewpoint	RV1	RV2
Date	1																												
AppliancesC(	-0.00963	1																											
LightsConsur	-0.1898349	0.19727756	1																										
TempKitcher	0.69697401	0.05544747	-0.0235281	1																									
HumKitchen	-0.1027598	0.08603143	0.10696811	0.16400579	1																								
TempLivingR	0.51532097	0.12007328	-0.0056217	0.83683399	0.26983924	1																							
HumLivingR	-0.0929451	-0.0604648	0.05098502	-0.0025088	0.79753543	-0.16561	1																						
TempLaundr	0.79293795	0.0850599	-0.0973931	0.89240229	0.25322995	0.73524547	0.13731871	1																					
HumLaundry	-0.4319389	0.03629172	0.13116053	-0.0285502	0.84467692	0.12149667	0.67832638	-0.0112339	1																				
TempOffice	0.78195356	0.04028098	-0.0088587	0.87700106	0.10617954	0.76206605	-0.0473041	0.85277773	-0.1404572	1																			
HumOffice	-0.263074	0.01696543	0.11493571	0.09786124	0.88035854	0.23156287	0.72143511	0.12273738	0.89897829	-0.0486504	1																		
TempBathro	0.77730881	0.01975986	-0.0787446	0.88524687	0.20579699	0.72055019	0.11040933	0.88816891	-0.0500625	0.87181348	0.09181192	1																	
HumBathroo	-0.2378516	0.00695454	0.14123332	-0.0147819	0.30325825	0.0295952	0.25027068	-0.0663552	0.37542217	-0.0764893	0.35259097	0.03278556	1																
TempOuside	0.60140191	0.11763769	-0.0790288	0.65476873	0.31614129	0.80118605	-0.0096703	0.68688248	0.07683302	0.6523497	0.25904724	0.62916061	-0.0783372	1															
HumOutside	-0.8242293	-0.0831776	0.15375647	-0.6150449	0.24512617	-0.5803718	0.38993302	-0.6476722	0.51491208	-0.7031486	0.39217771	-0.6323838	0.26379653	-0.672137	1														
TempIronRo	0.83497383	0.02580104	-0.1353474	0.83870503	0.02139703	0.66366009	-0.0514218	0.84737415	-0.2500905	0.87776288	-0.1312036	0.8706242	-0.1428233	0.61908486	-0.753973	1													
HumIronRoo	-0.1707288	-0.0556424	0.03506932	0.13518236	0.80112213	0.2292119	0.69058432	0.17262397	0.83268499	0.04352678	0.89430125	0.14890541	0.32580758	0.25607652	0.35722161	-0.0339478	1												
TempTeenag	0.78364572	0.03957198	-0.0714577	0.82541331	-0.030053	0.57819084	-0.0410233	0.79528349	-0.283228	0.79625585	-0.1670663	0.8249806	-0.0869906	0.48283551	-0.665869	0.88212323	-0.1241023	1											
HumTeenag	-0.2815496	-0.0940386	0.01291533	-0.0064413	0.7361959	0.06853418	0.67977721	0.04442678	0.8288219	-0.0951922	0.84725904	0.01638752	0.35983964	0.07372138	0.48958032	-0.2099608	0.88398396	-0.2095324	1										
TempParent	0.88854621	0.01001028	-0.1575924	0.84477661	0.11526335	0.67553459	0.05454443	0.90132359	-0.1952697	0.88943915	-0.0445176	0.91105512	-0.138509	0.66717699	-0.7389402	0.94477642	0.02805475	0.86933772	-0.1130135	1									
HumParentR	-0.2446473	-0.051462	-0.0087659	0.07175561	0.76400124	0.15734631	0.6764668	0.13460215	0.83353797	-0.025549	0.85659108	0.0723079	0.2721971	0.1844243	0.39194343	-0.0776898	0.85868637	-0.15682	0.85581186	-0.0086834	1								
TempFromSt	0.58885209	0.09915468	-0.0744226	0.68285431	0.34078795	0.79224739	0.03370292	0.69941993	0.11823719	0.6634787	0.29330709	0.65133473	-0.0530904	0.97477846	-0.6415726	0.63129954	0.2942225	0.50284973	0.11716423	0.66821838	0.22328156	1							
PressureFror	-0.0267705	-0.0348848	-0.0105764	-0.1505745	-0.2939569	-0.1330283	-0.255646	-0.1899744	-0.2332743	-0.0752919	-0.2507475	-0.1709989	-0.1193274	-0.1416428	-0.0658142	-0.0977239	-0.2658497	-0.1628546	-0.2255912	-0.1568277	-0.1837393	-0.1432581	1						
HumidityFro	-0.3363621	-0.1522824	0.06854282	-0.3454815	0.27412601	-0.5052908	0.5849106	-0.2817177	0.35619197	-0.3886024	0.33681262	-0.2739528	0.18594103	-0.569119	0.7185867	-0.4112132	0.37851908	-0.3004375	0.48735452	-0.3188484	0.35937703	-0.5741864	-0.0920165	1					
WindspeedF	-0.257415	0.08712211	0.06028103	-0.0876535	0.20493249	0.05249513	0.06919046	-0.1007756	0.26318773	-0.1857465	0.30019234	-0.1450112	0.08204351	0.1691851	0.09831666	-0.185892	0.21297871	-0.218876	0.20237993	-0.1777564	0.23865456	0.19293995	-0.2350323	-0.1764583	1				
VisibilityFror	-0.0962136	0.00023036	0.02003774	-0.0762099	-0.0210573	-0.0697214	-0.0053676	-0.1023098	0.01704109	-0.1047684	0.00263635	-0.0841645	-0.0131333	-0.0813127	0.10870545	-0.1100885	-0.0071845	-0.0600981	0.04617803	-0.1039152	0.00866674	-0.0773675	0.04031489	0.08312504	-0.0075158	1			
Tdewpoint	0.46875139	0.01535664	-0.0363305	0.57133297	0.6391245	0.58263243	0.4991542	0.64590853	0.41439796	0.51950395	0.61650755	0.58839271	0.07816399	0.76424176	-0.2604265	0.46666431	0.64185703	0.39183926	0.50178146	0.58151652	0.54033738	0.79066425	-0.244103	0.03650932	0.12592619	-0.0421943	1		
RV1	0.00027678	-0.0111449	0.00052111	-0.006203	-0.0006993	-0.0110874	0.00627472	-0.0051944	-0.0004775	-0.0018147	-0.0017865	-0.0054902	-0.0110562	-0.0150862	0.01167136	-0.0038997	0.00181971	-0.0032103	0.00447966	-0.0012272	-0.0029545	-0.0152588	0.00069946	0.02044068	-0.0113418	-0.0058886	-0.0039503	1	
RV2	0.00027678	-0.0111449	0.00052111	-0.006203	-0.0006993	-0.0110874	0.00627472	-0.0051944	-0.0004775	-0.0018147	-0.0017865	-0.0054902	-0.0110562	-0.0150862	0.01167136	-0.0038997	0.00181971	-0.0032103	0.00447966	-0.0012272	-0.0029545	-0.0152588	0.00069946	0.02044068	-0.0113418	-0.0058886	-0.0039503	1	1

Correlation Matrix of Initial Data Set

Appendix 2

Table 2
Data variables and description.

Data variables	Units	Number of features
Appliances energy consumption	Wh	1
Light energy consumption	Wh	2
T1, Temperature in kitchen area	°C	3
RH1, Humidity in kitchen area	%	4
T2, Temperature in living room area	°C	5
RH2, Humidity in living room area	%	6
T3, Temperature in laundry room area	°C	7
RH3, Humidity in laundry room area	%	8
T4, Temperature in office room	°C	9
RH4, Humidity in office room	%	10
T5, Temperature in bathroom	°C	11
RH5, Humidity in bathroom	%	12
T6, Temperature outside the building (north side)	°C	13
RH6, Humidity outside the building (north side)	%	14
T7, Temperature in ironing room	°C	15
RH7, Humidity in ironing room	%	16
T8, Temperature in teenager room 2	°C	17
RH8, Humidity in teenager room 2	%	18
T9, Temperature in parents room	°C	19
RH9, Humidity in parents room	%	20
To, Temperature outside (from Chièvres weather station)	°C	21
Pressure (from Chièvres weather station)	mm Hg	22
RHo, Humidity outside (from Chièvres weather station)	%	23
Windspeed (from Chièvres weather station)	m/s	24
Visibility (from Chièvres weather station)	km	25
Tdewpoint (from Chièvres weather station)	°C	26
Random Variable 1 (RV_1)	Non dimensional	27
Random Variable 2 (RV_2)	Non dimensional	28
Number of seconds from midnight (NSM)	s	29
Week status (weekend (0) or a weekday (1))	Factor/categorical	30
Day of week (Monday, Tuesday... Sunday)	Factor/categorical	31
Date time stamp	year-month-day hour:min:s	–

Data Key and units of measurements