PATTERN EXTRACTION FROM AN INDIVIDUAL HOUSEHOLD POWER CONSUMPTION DATASET

MACHINE LEARNING FOR SIGNALS (EE 660)

ADIGUN OLAOLUWA
adigun@usc.edu
6th December 2015.

https://github.com/peruz/EE660_Project.git

ABSTRACT

The aim of this project is to apply various machine learning tools to analyze the dataset of an individual household electric power consumption and provide suggestions on how to reduce the consumption rate. The dataset gives minute-by-minute details of power consumption over the period of 4 years. The project focuses on identifying how the power consumption fluctuates with respect to season of the year, day of the week, and time of the day.

I applied various machine learning models because the project is broad. Some aspects of the tasks involved supervised learning models. I applied linear regression model to predict the total power consumption in the household per minute.

I also used logistic regression for the classification task involved in the analysis. I considered logistic regression to formulate a model for predicting whether the metered power or unmetered power accounts for majority of the household power consumption or not in general.

Based on the knowledge of feature importance obtained from the linear regression (using Lasso regression), I also carried out unsupervised learning using K-means clustering algorithm. I used this to gain more insight into the pattern of power consumption with respect to the key features identified from sparse linear regression model.

The results show that the unmetered power consumed in the house significantly affects the total power consumption. Also the power consumption due to the electric water heater and air-conditioner influences the meter power significantly. The results also show the pattern of power consumption over the periods of the day, days of the week, and seasons of the year. This pattern gives an insight into how to efficiently reduce power consumption in the house

PROBLEM STATEMENT AND GOALS

The aim of this project is to analyze the power consumption pattern of a household from the available dataset and suggest how to reduce power consumption. The dataset contains record of some features in addition to the actual power usage of the household over a period time. We

expect some of these features to influence the power consumption but we cannot rely on this hypothesis. Instead we would rather let the data speak for itself. Ultimately we want to have a deeper insight into how the power consumption fluctuates for this household.

Some of the insights I wanted to acquire include the following:

- Getting a model for predicting total power consumption in the household. This model will help understand how the input features affects the model. It will also explain the key features that can be used for some other models.
- I also want to know the significance of meter power as compared with the unmetered power used in the household. This will give an insight into which appliances are responsible for high power consumption.
- I considered the acquiring information about how features such as time of the day, day of the week and season of the year influences total power consumption of the household.
- Determine best predictor for total power consumed in the house out of the three sub-meters.

This set of information will help the household manage its power consumption effectively. The information can also help the household decide whether it is economical to consider another source of electric power or not. We can also predict the expected power usage over a period of time. A model of this type can be used for designing a mobile application that helps users manage power consumption in the household.

This project involved some difficulties discussed below.

- **Significant Amount of Preprocessing**: This dataset has approximately 2,000,000 data points that accounts for a data size of 130MB. All the software I tried could not open the dataset because of its size (MATLAB, R, and Microsoft Excel). I had to perform a line-by-line processing of the dataset into different files. Random sampling was not possible for me without access to the content of the data file.
- **Incomplete Data:** Dealing with the missing data was not a difficult thing to deal with because of the huge size of complete data compared to the missing samples. The

complete data accounts for approximately 98% of the entire dataset. So I relied on the complete samples in formulating the models.

- **High Dimensionality:** High dimensionality of the feature space will affect the quality of the model with respect to clustering. Even if the learning algorithm should converge to a solution it will be difficult to interpret the outcome of the clustering method. So I used the linear regression model (Lasso) to pick the most significant features before proceeding with the clustering.

LITERATURE REVIEW

There exist some works that are similar to this project. Amir et al [1] proposed a method to examine the determinants of residential electricity consumption. The model considered interaction between features and selected the subset that best explained the pattern of electricity consumption. The model uses weighted linear regression to explain the variation in power consumption with respect to the input features. For selection the best subset they considered the stepwise selection approach because it ranks variables based on their importance. To further gain insight into the power consumption pattern the work considered the peak and minimum load periods separately.

The input features were grouped into 4 categories namely:

- Weather and location (E.g. Ambient temperature and climate zone).
- Appliance and Electronics stock of the occupants (E.g. Number of refrigerators or computers).
- Occupancy and Occupants' behavior towards energy consumption (E.g. Purchasing energy-efficient appliances).
- Physical characteristics of the building (E.g. Level of insulation).

The work considered a dataset collected from 168 households over a period of 238 days. The result of the model shows that the daily minimum consumption is influenced the most by weather, location, and physical characteristics of the building. The energy intensive end usage influences the maximum daily consumption more than any other feature. For example the electric water heater.

One of the major differences between this work and my project is that the analysis in my case is specific to a single household with dataset coming from the household.

PRIOR WORK (NONE)

For this particular dataset no document prior work has been done on it.

PROJECT FORMULATION AND SETUP

The learning algorithms I considered for this project are described briefly below. These machine learning algorithms were used to extract information from the household power consumption dataset.

- **Linear Regression**

  This is a discriminative model of the form $p(y|\underline{x}, \theta)$ where $y \in \mathbb{R}$. The model represents a Gaussian pdf as show in equation (1) below.

$$p(y|\underline{x}, \theta) = \mathcal{N}(y \mid \underline{w}^T \underline{x}, \sigma^2) \qquad (1)$$

  The first algorithm I considered was the linear regression for predicting the power consumption in the household with respect to input features. I also used the linear regression to pick the most prominent features from the set of input features for clustering. This reduced the computational cost of clustering and simplified the result of the model. The result from running clustering algorithm on the important features is easier to interpret than considering all the input features.

I used Lasso method for the linear regression model. Lasso gives a sparse model for linear regression because of the condition attached to the optimization task. The Lasso regression seeks to maximize the posterior while assuming that the prior is has a Laplace pdf. The cost function is:

$$f(\underline{w}) = \sum_{i=1}^{N} -\frac{1}{2\sigma^2}\left(y_i - \left(w_0 + \underline{w}^T x_i\right)\right)^2 + \lambda\|\underline{w}\|_1 \qquad (2)$$

where λ the regularization is term and $\underline{w}$ is the weight vector. The dataset is $\{(\underline{x}_1, y_1), \ldots\ldots, (\underline{x}_N, y_N)\}$, w_0 is the intercept term, and σ^2 is the variance. The Lasso regression is a constrained optimization of equation (2) such that $\|\underline{w}\| < B$.

Algorithm: Coordinate Descent for Lasso [3].

Initialize $\underline{w} = (X^T X + \lambda I)^{-1} X^T Y$;

Repeat until convergence

For j = 1,2,. . . , d

$a_j = 2\sum_{i=1}^{N} {x_{ij}}^2$;

$c_j = 2\sum_{i=1}^{N} {x_{ij}}^2 \left(y_i - \underline{w}^T x_i + w_j x_{ij}\right)$;

$w_j = soft\left(\frac{c_j}{a_j}, \frac{\lambda}{a_j}\right)$

End

- **Logistic Regression**

The second technique I used was logistic regression. This is a discriminative model $p(y|\, x, \underline{w})$ for binary classification. I used this model to acquire information about the significance of metered power and unmetered power. This method find the parameter $\underline{\hat{w}}$ that maximizes the cross-entropy error function.

$$\underline{\widehat{w}} = \ argmax\left\{-\sum_{i=1}^{N} log\left(1 + \ e^{-y_i \ \underline{w}^T x_i}\right)\right\}$$

The cost function for Logistic regression is shown in equation (3) below:

$$f(\underline{w}) = \ -\sum_{i=1}^{N} log\left(1 + \ e^{-y_i \times \underline{w}^T x_i}\right)$$

The algorithm for $\underline{\widehat{w}}$ is shown below:

Algorithm: Logistic Regression with Newton's method.

Initialize w = 0_D*;*

$w_0 = log\left(\bar{y}/(1-\bar{y})\right);$

Repeat

$\eta_i = \ w_0 + \ \underline{w}^T x_i$

$\mu_i = sigm(\eta_i);$

$s_i = \ \mu_i(1-\mu_i);$

$z_i = \ \eta_i + \ \dfrac{y_i - \ \mu_i}{s_i}$

$S = diag(s_{1:N})$

$w = \ (X^T S X)^{-1} X^T S z;$

Until Convergence

- **K-Means Clustering**

The last algorithm I used was the K-means clustering algorithm for unsupervised learning. With the result of linear regression, I used the K-means clustering algorithm to learn the pattern that exist between the key input features (controllable by the consumer) and the power consumption. The K-means clustering is an algorithm for unsupervised learning.

The dataset for K-means clustering included unlabeled samples $\left\{\left(\underline{x}_1\right), \ldots \ldots, \left(\underline{x}_N\right)\right\}$. The aim of this algorithm is to minimize the cost function shown below:

$$z^* = \ arg\ min_k \left\|\underline{x}_i - \mu_k\right\|_2^{\ 2}$$

The algorithm for K-means clustering is shown below.

Algorithm: K-Means Algorithm [3].

Initialize the mean of the clusters: μ_k

Repeat Until Convergence.

Assign each data to the closest cluster:

Update the mean of the cluster:

$$\mu_k = \frac{1}{N_k} \sum_{i:\, z_i = k} \underline{x}_i$$

End;

METHODOLOGY

The aim of the project is to acquire information about the underlying pattern of power consumption in a household. The first I did was to reduce extract additional features needed for the project from the ones on the dataset. Then I divided the dataset into four parts (pre-training set, cross validation set, training set, and testing set). I took a pre-training set of samples from the dataset and plotted scatter plot of it to determine how to handle the features. My findings (as reported in the next section) informed my decision on the pre-processing approach I considered. Then I proceeded with feature space reduction by changing some features such as (time of the day, day of the week, and month of the year) into categorical features. This reduced the entire data set to 4,121 data points as shown below.

- Pre - Training set : 321 samples
- Training set: 3000 samples
- Cross – Validation set: 400 samples
- Testing set: 400 samples

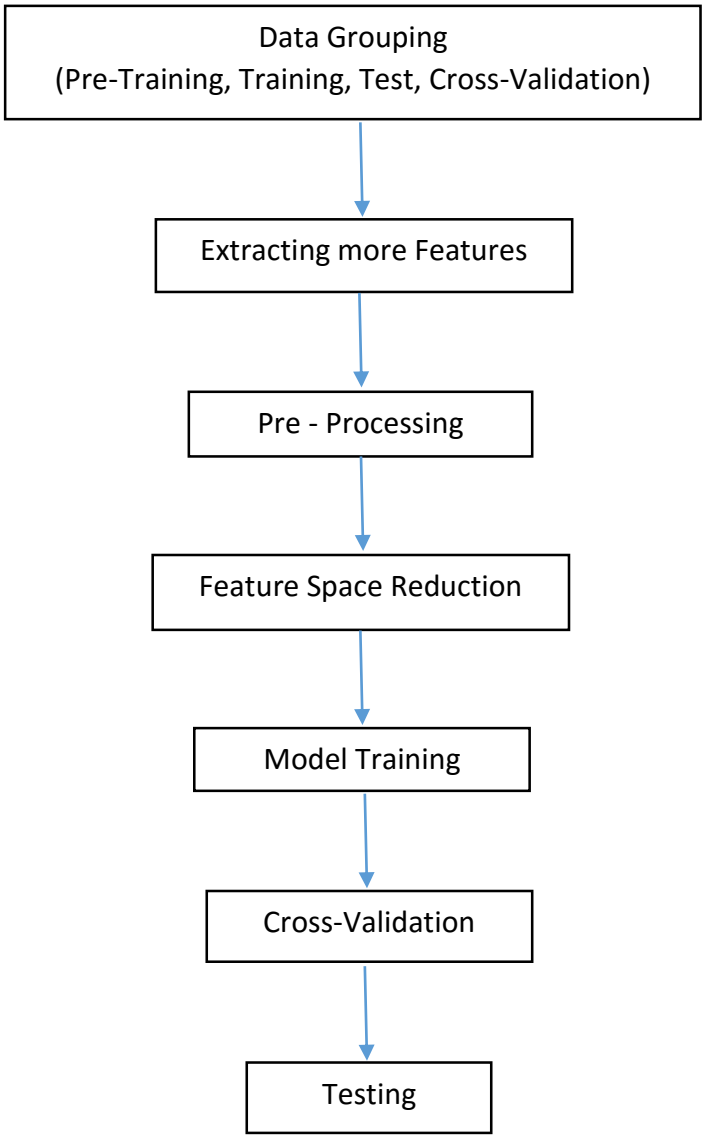
After this I moved on to the training of the Lasso model with the training set, and picking the best regularization constant with cross-validation set. The performance of the model was tested on the test data.

Then I proceeded to the part involving the logistic regression. I trained the data using the training set and I used the cross-validation dataset for picking the best regularization term for the model, and I used the testing set to test the performance of my model.

For the K-means algorithm, I relied on the result of the linear regression model to decide the features to consider for this method. Then in this case, there was no need using cross validation or testing data in this case but I only used the training dataset for the K-means clustering. I computed the Calinski and Harabasz index.

The pre-training set was only used to decide on how to proceed with the feature space reduction.

FLOW – CHART FOR THE LEARNING ALGORITHMS USED



IMPLEMENTATION AND RESULT

Feature Space

The dataset for this project contains the minute-by-minute power consumption of a single household over the period of two years. The features reported on the data set include:

- Date: This reports the corresponding date for each data point. The format for the date is dd/mm/yyyy.
- Time: This gives the corresponding time of the day for each data point in format hh:mm:ss.
- Global Active Power: The household global minute-averaged active power for the household.
- Global Reactive Power: This gives the household global minute-averaged reactive power for the household.
- Global Intensity: This is the averaged global current intensity over a minute.
- Voltage: This is the averaged global voltage over a minute.
- Sub-metering 1: This is the average power consumed by appliances connected to sub-meter 1. These include a dishwasher, an oven, and a microwave.
- Sub-metering 2: This reports the average power consumed by appliances in the laundry room. They include a washing-machine, a tumble-drier, a refrigerator, and a light.
- Sub-metering 3: This corresponds to averaged power consumed by electric water-heater and air-conditioner.

The sub-meters only record a fraction of the total power consumed by the household. The difference between the global active power and the sum of power consumption due to sub-meters 1, 2, and 3 accounts for the unmetered power consumed by the household. The dataset was collected over a period of 47-months and this corresponds to 2,000,004 samples. The complete data is approximately 98% of the entire dataset which is high enough to given a model that fits the entire dataset.

Pre-processing

The pre-processing tasks done are described below. I started with the conversion of the date and time of the day into categorical features.

- Time of the Day

  The power consumption reading spanned through every minutes of the day for a period of 47months. This represents about 1440 samples for a day. After picking the pre-training data from the dataset I plotted the power consumption per minute against time as shown in Figure (1) below. The plot prompted my decision to divide this feature into 3 classes as against using 1,440 distinct points to encode the time. I noticed that the trend in consumption rate can be classified into 3 major parts. So I divided the time of the day into *morning, afternoon,* and *evening*. Morning period spans through 12am till 8am, afternoon is 8am till 5pm, and evening covers 5pm till 12am.

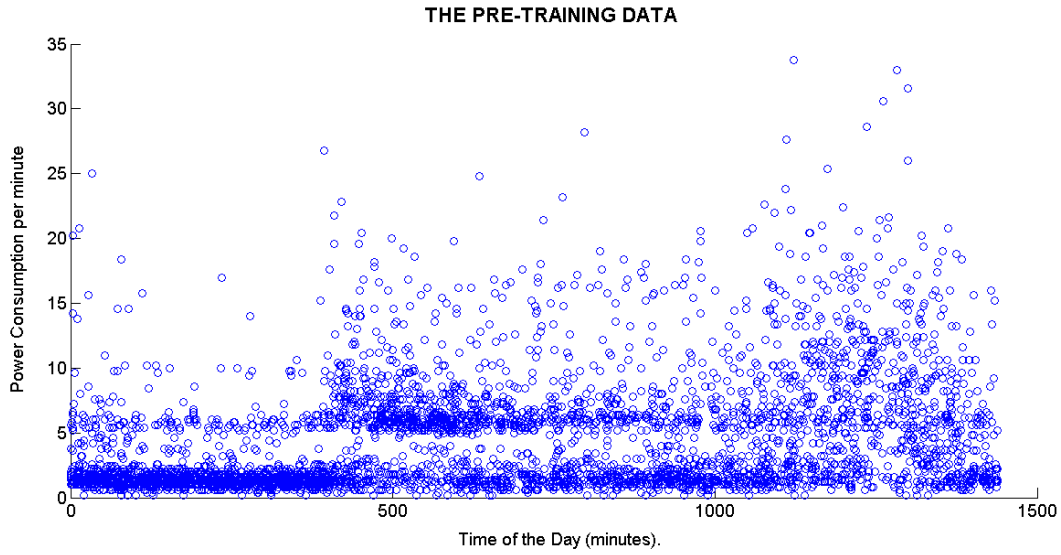


Figure 1: Pre – Training Data. The plot of Power consumption per minute plotted against the Time of the Day.

I converted the time of the day into a categorical feature and I encoded it as (Morning = 1, Afternoon = 2, and Evening = 3).

- Day of the Week.

  I plotted the power consumption against day of the week as show in Figure (2). I noticed that using all the days will increase the computational cost of the model without any substantive benefit of doing so. I decided to encode the day of the week as a categorical feature into *weekday* and *weekend.* From the pre-training plot in figure 2, one can see that there seems to be more consumption during the weekend (Saturday and Sunday) than weekdays (Monday – Friday).

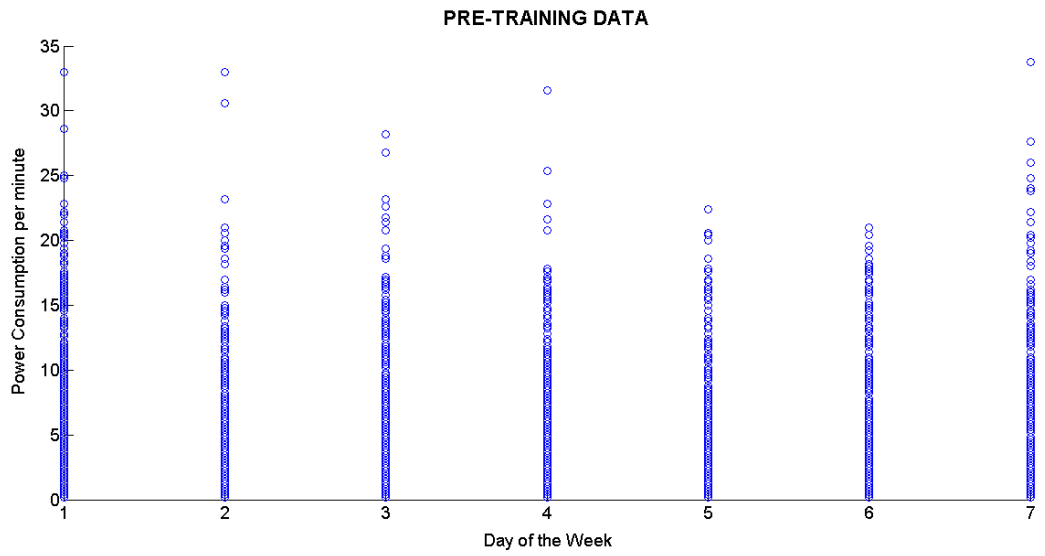


Figure 2: Pre – Training Data. The plot of Power consumption against the Day of the Week. The labels are listed here: (Sunday = 1, Monday = 2, Tuesday = 3, Wednesday = 4, Thursday = 5, Friday = 6, Saturday = 7).

I encoded the day of the week as (Weekday = 1, and Weekend = 2).

- Season of the Year.

  The month of the year does not provide so much individual information as much as classifying the month into two categories namely: *hot season* and *cold season.*

Figure 3: Pre – Training Data. The plot of Power consumption against the Month of the Year. The labels are listed here: (January = 1, February = 2, March = 3, April = 4, May = 5, June = 6, July = 7, August = 8, September = 9, October = 10, November = 11, December = 12).

Figure 3 above shows the plot of power consumption against the month of the year for each data. The plot shows that we have more power consumption during the cold season (October – March) than the hot season (April – September).

Due to the result of the plot I decided to convert the months of the year into a categorical data encoded as (Hot season = 1, and Cold season = 2).

After the pre-processing of the feature space of date and time of the day, I had to re-compute the corresponding features for data points that correspond to the date and time as specified by the new space for these two features. For instance, I now have just 3 possible values for time of the day, 2 for day of the week, and 2 for month of the year. So I took the average of all the sample points with the same time within a day.

- Global Active Power: This is a continuous feature that gives the global average active power per minute for the household corresponding to a period of a day (morning, afternoon, or evening). The average was taken over the samples within the same class of the day. This feature space is the real line.
- Voltage: This is the average voltage per minute taken over the samples that belong to the same class of time of the day. This feature is continuous over the real line.

- Global Intensity: Global intensity gives the averaged-minute global intensity of all the samples collected over a specific class of time of a day (morning, afternoon, or evening). The feature is also continuous over the real line.
- Sub-metering 1: This is the average energy consumed by appliances connected to sub-meter 1. These include the dishwasher, an oven, and a microwave. I took the average of this feature over the samples that fall within the same period of a day (morning, afternoon, and evening). The feature is continuous over the real line.
- Sub-metering 2: This reports the average power consumed by appliances in the laundry room. They include a washing-machine, a tumble-drier, a refrigerator, and a light. I took the average over the samples within the same class of time of the day. The feature is continuous over the real line.
- Sub-metering 3: This corresponds to power consumption due to the usage of electric water-heater and air-conditioner. I took the average of the power consumption due to these appliances over a give class of time the day.

The pre-processing reduced the entire dataset to 4,121 data points from approximately 2,000,000 samples.

Feature Extraction

The available sub-meters only accounts for a portion of the entire power consumption in the household. Knowing the unmetered power consumed in the house will give an insight into how to control power consumption in the house. This will give us an idea of which appliances are contributing significantly to the total power consumption in the house.

- Unmetered Power Consumed:

  It was computed using the formula below, where sub-metering 1 is SM1, sub-metering 2 is SM, and sub-metering 3 is SM3.

$$Unmetered\ Power = \left(Global\ Active\ Power\ \times \frac{1000}{60}\right) - \quad SM1 - SM2 - SM3$$

- Meter vs Unmetered :

  This is binary feature that shows when the meter power consumed accounts for more consumption than the unmetered power consumed. This feature is a binary feature.

Table 1 below shows the features at the end of preprocessing and feature extraction.

Table 1: Features after Pre-Processing and Extraction.

Features	Type	Classes
Time of the Day	Categorical	Morning, Afternoon, or Evening
Day of the Week	Categorical	Weekend or Weekday
Season of the Year	Categorical	Hot season or Cold season
Global Active Power	Real	
Voltage	Real	
Global Intensity	Real	
Sub –Metering 1	Real	
Sub-Metering 2	Real	
Sub-Metering 3	Real	
Unmetered Power	Real	
Meter vs. Unmetered	Binary	0 or 1

For the categorical features, I encoded them as binary features when working with regression models such that each class of the categorical feature was defined as a binary feature with possible values 0 or 1.

Training Process

Linear Regression:

I compared the result of Lasso model and ridge regression but I settled for the Lasso model because it encourages sparsity. For the categorical datasets I converted them into binary features for each class and I standardized each feature on the training set.

The features I considered include:

- Time of the day
- Day of the Year
- Season of the Year
- Voltage
- Global Intensity
- Sub-Metering 1
- Sub – Metering 2
- Sub –Metering 3

The training set has 3,000 samples and 8 input features. The complexity of the hypothesis set affects the outcome of a model. A model with high complexity results in likely to overfit with high out of sample error. Although the size of training data is slightly high compared to the number of features, I decided to include the regularization term to control the overfitting problem. I used the cross-validation discussed in the next section to pick the best regularization term for the model. I specified a range for the regularization and picked the best value based on the performance with respect to the cross-validation set. The cross-validation was done in 5 folds.

Logistic Regression

I trained the model with training set. I encoded the categorical features as binary feature for each class and then standardized the input features on the dataset. The features I considered were:

- Time of the day
- Day of the week
- Season of the year

- Sub - Metering 1
- Sub – Metering 2
- Sub –Metering 3

The training set has 3,000 samples and 4 features. The complexity of hypothesis set shouldn't affect the model that much in this case because the feature space is relatively small. To avoid any possibility of overfitting I used a regularization term in the model. For the training I considered a set of possible terms and the best term as iced based on the performance on the cross validation set. I used the zero-one loss function to compute the performance of the model.

K –Means Clustering:

I considered the features that are most significant as suggested by the sparse linear regression model. I also restricted myself to features under the user's control. The features I considered include:

- Sub-Metering 1
- Sub – Metering 2
- Sub –Metering 3
- Unmetered Power

I clustered these features together with the total power consumed per minute. I decided to cluster the data into 3 parts because I am interested in how these features partitions the power consumption into regions of high, medium, and low power consumption.

For the training I initialized the mean of the clusters using 3 data points chosen randomly from the training set. Then I computed the nearest cluster for each sample and re-adjusted over several loops until convergence. To ensure an efficient result I repeated the process a couple of times with different initial mean for the clusters.

Testing, Validation and Model Selection

Linear Regression:

I considered ridge regression and Lasso regression. I used the cross validation set to pick the best regularization term for the linear regression models. The cross-validation was done in 5 folds over the cross validation set. I compared the result of ridge regression and Lasso regression. The best regularization terms are 6.9 and 10 for lasso regression and ridge regression respectively. I also considered the sparse nature of lasso model to decide the import features in the model. The next section contains the full result of this model.

Logistic Regression:

I used the training set to try different values of regularization terms. Then I used the cross-validation set to pick the best regularization term. It was necessary to include a regularization term to control overfitting. The best regularization term for this model is 0.01. I considered only the features under the user's control for the logistic regression model.

K-Means Clustering:

I did not use cross validation in this case. I used a prior knowledge of K = 3 because my aim is to see how the appliances in the household (sub-metering 1, sub-metering 2, sub-metering 3, and unmetered power) separates the power consumption space into low, medium, and high consumption. The Calinski and Harabasz index for the model is 3602.9107.

RESULTS

Linear Regression

All Features:

The models is:

$$f(\underline{x}) = w_0 + \underline{w}^T \underline{x}$$

where $f(\underline{x})$ is the predicted power consumed by the household and $\underline{w}$ is the weight matrix for the features.

The result of Lasso regression and ridge regression are shown in table 2 below:

Table 2: Weights for Lasso Model.

Feature	Lasso Regression	Ridge Regression
Intercept	18.5219	18.5130
Morning	0.0000	-0.0112
Afternoon	0.0000	-0.0268
Evening	0.0194	0.0379
Weekday	-0.0136	-0.0063
Weekend	0.0136	0.0063
Hot Season	-0.0563	-0.0786
Cold Season	0.0563	0.0786
Voltage	0.1075	0.1221
Global Intensity (Current)	10.8792	10.6774
Sub – Metering 1	-0.1141	-0.0226
Sub – Metering 2	-0.0215	0.0019
Sub – Metering 3	0.1488	0.2699

Table 3 shows the training error, testing error, and the best regularization term obtained from the cross-validation process.

Table 3: The Results for Lasso Regression and Ridge Regression.

Parameters	Lasso Regression	Ridge Regression
Training Error	0.0625	0.0666
Testing Error	0.0860	0.0843
Regularization Constant	6.9	10

From the result above we can see that the Lasso regression encourages sparsity.

Best Two Features:

I picked the best two features that the household can control (sub-metering 1 and sub-metering 3) and use it to find the model for predicting the power consumed in the household.

Table 4: Weights for Lasso model with two best features.

Feature	Weight
Intercept	18.5142
Sub – Metering 1	4.4494
Sub – Metering 3	6.2008

Table 5: Results for the Model with best two features under household control.

Parameters	Value
Regularization Term	1.1
Training Error	44.1503
Testing Error	50.2905

Figure 4 below compares the prediction of the model and the actual value on the testing set.

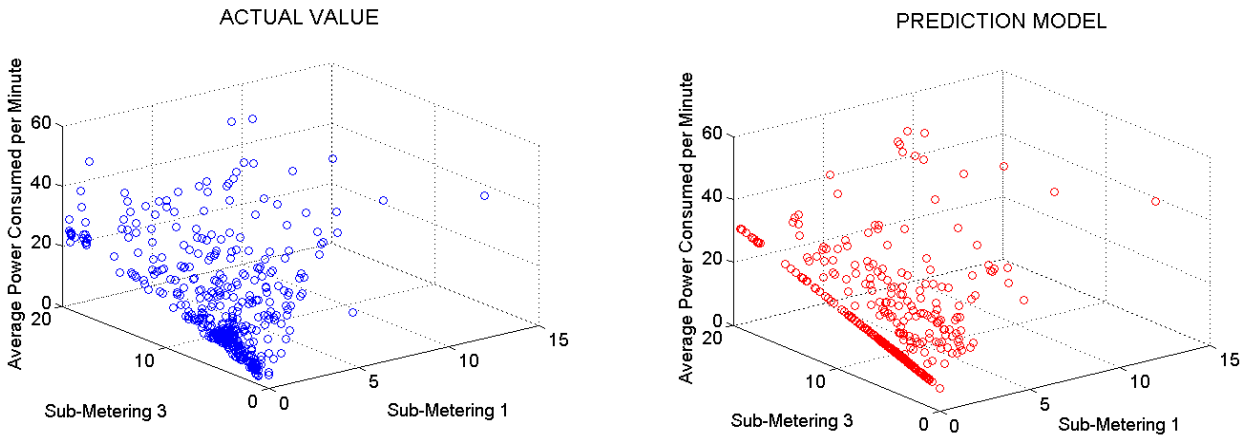


Figure 4: Plots for the predictions from the model and the actual value using best two features under the household control.

Logistic Regression

All Features:

The model is:

$$f(\underline{x}_i) = \frac{1}{1 + e^{-(w_0 + \underline{w}^T \underline{x}_i)}}$$

$$\hat{y}_i = \begin{cases} 1, & if\ f(\underline{x}_i) > 0.5 \\ -1, & if\ f(\underline{x}_i) < 0.5 \end{cases}$$

Table 6 below shows the weights associated to the features according to the model. Table 7 shows the result obtained from the testing, training, and cross validating the model. It also gives the best regularization term.

Table 6: Weights of the Logistic Regression Model.

Feature	Logistic Regression
Intercept	-0.6601
Morning	-0.0388
Afternoon	0.4365
Night	-0.3982
Weekday	-0.0604
Weekend	0.0604
Hot Season	0.4893
Cold Season	0.4893
Sub – Metering 1	0.2969
Sub – Metering 2	0.5531
Sub – Metering 3	1.3165

Table 7: Other Results.

Parameters	Value
Regularization Term	0.01
Training Error	0.1868
Testing Error	0.1700

Best Two Features:

From the result of the logistic regression model, the best two features are sub-metering 2 and sub-metering 3. With these two features I came up with a Logistic regression model shown below. Tables 8 and 9 below shows the result of the model using the best two features. The

regularization term was picked based on the performance with respect to the cross validation set.

Table 8: Weights for Logistic regression model with the best two features.

Feature	Weight
Intercept	-0.5284
Sub – Metering 2	0.4141
Sub – Metering 3	1.1933

Table 9: Results for the Model with best two features under household control.

Parameters	Value
Regularization Term	3.5313
Training Error	0.2350
Testing Error	0.2494

Figure 5: Logistic regression model.

K-Means Clustering

I clustered the data into 3 regions using 5 features earlier mentioned and I plotted the result in 2D for each feature against the global active power. Figures 6 – 9 show the result of the clustering.

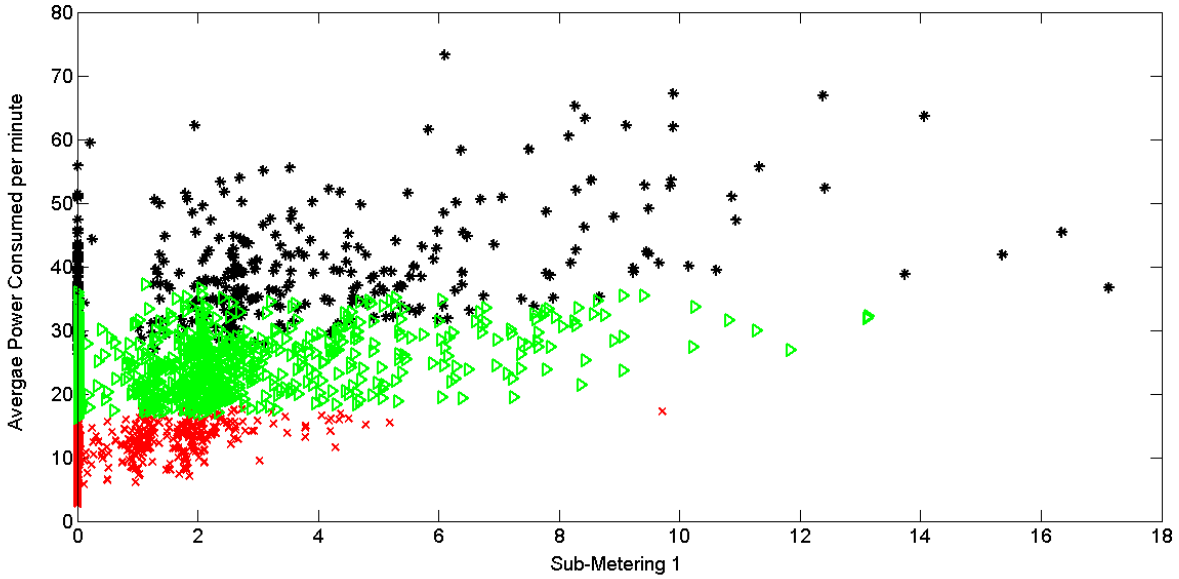


Figure 6: K-Means Clustering (Power Consumption against Sub-Metering 1)

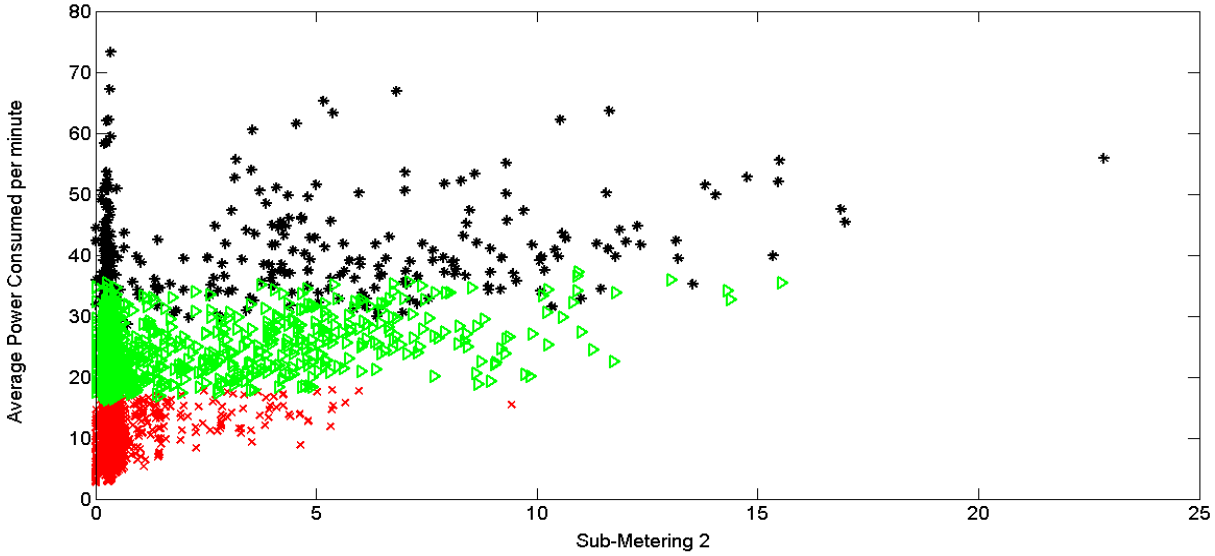


Figure 8: K-Means Clustering (Power Consumption against Sub-Metering 2)

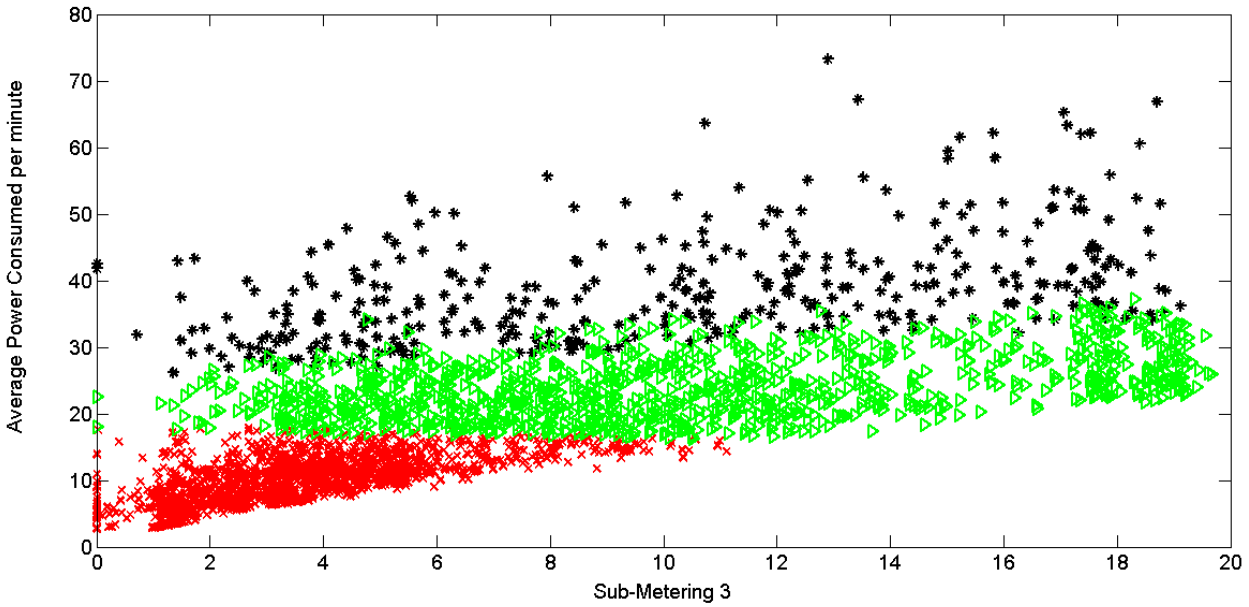


Figure 9: K-Means Clustering (Power Consumption against Sub-Metering 3)

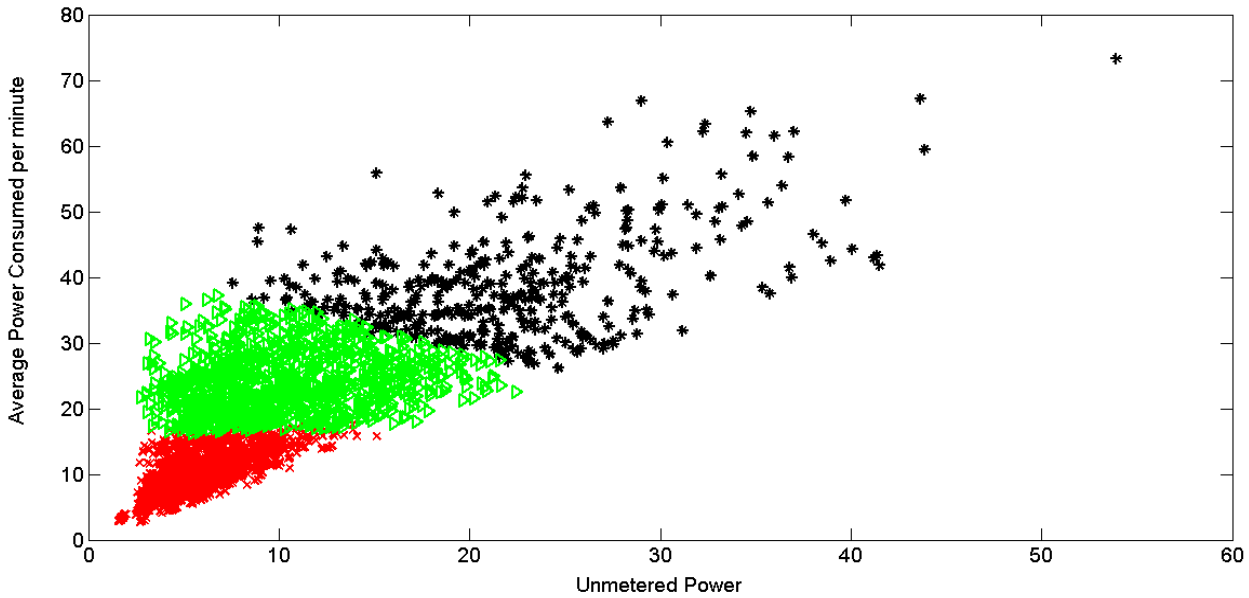


Figure 10: K-Means Clustering (Power Consumption against Unmetered Power)

The Calinski and Harabasz index for the model is 3602.9107.

INTERPRETATION

Linear Regression

For the linear regression model the features were standardize to mean 0 and standard deviation 1 therefore we can use the weights of the model to determine the influence of each feature on the power consumption rate.

During the evening period the household experiences more power consumption than either the morning or afternoon period. For the day of the week, there is more consumption during the weekend than the weekdays. This is not surprising because one would expect more activities during the weekend. Similarly the season of year also affected the power consumption as expected. There is more consumption during the cold season as compared to the hot season. The voltage and global intensity (current) intensity also affects the power consumption in the household. There is more consumption when any of this two increases. In fact the global intensity (current) is the feature with the greatest influence on the model.

For the sub-metering systems, the one that accounts for electric water-heater and air-conditioner influences the rate of power consumption in the house the most. The other two sub-metering systems does not influence the power consumption in the house as much as sub-metering 3.

Logistic Regression:

From the dataset we know that the sub-metering system only accounts for a fraction of the entire power consumption in the house. The logistic regression model provided an insight into the significance of unmetered power consumed in the house. A sense of this can help determine what influences the unmetered power in the house.

For a logistic regression model, a feature with positive weight causes increase in odds of the output when the feature is increased. On the other hand the odds reduces when the features with negative weights increases.

The logistic regression model shows that power consumption in the house is significantly determined by unmetered power during the night and morning periods. This is not surprising because these periods coincide with when the occupants are in and we expect more activities in the house due to other appliances that are not on the sub-meters. Also the unmetered power seems to be more significant than the metered power during the weekdays. The other features favor metered power to be more significant that unmetered power.

K-Means Clustering:

The K-means clustering gives an insight into how each group of appliances affects the average power consumption in the house. The results shows that unmetered power is the best indicator for the level of power consumption in the house. According to figure (10), the power consumption in the house is low, medium, and high when the unmetered power is low, medium, and high respectively. This shows that unmetered power influences the power consumption in the house significantly. Also sub-metering 3 influences the power consumption but not as much as unmetered power. The other two sub-metering systems does not influence the power consumption significantly.

We can predict effectively the range of power consumption (low, medium, or high) from the knowledge of unmetered power and sub-metering 3 only.

SUMMARY AND CONCLUSION

This project gives an insight into the power consumption pattern of the household. The results obtained show that some features are more important than the others. We can make inferences from this models on how to reduce power consumption in the house significantly.

During the weekend the household should focus of regulating the usage of appliances connected to the sub-meters because it influences the power consumption a lot.

In general the reduction in unmetered power can also reduce the total consumption in the house significantly. In cases where they can get an economical substitute this should be considered. The electric water heater and air-conditioner usage can be regulated to reduce the power consumption rate significantly.

Further analysis can be done to identify the exact variation of power consumption for each period of the day separately. This will provide more information on how to effectively control power consumption for each period exclusively. Also a more robust model of this type can be implemented on a software platform as a mobile application.

REFERENCES

[1] Kavousian A., Rajagopal R., Fischer M., " Determinants of residential electricity consumption: using smart meter data to examine the effect of climate, building characteristics, appliance stock, and occupants' behavior," Energy 2013 ;55:184e94.URL: http:// *www.amirkavousian .com/wp content/uploads/2015/04/Kavousian etal 2013 DataDrivenEnergyEfficiency.pdf*

[2] G. Hebrai , "Individual household electric power consumption Data Set," as of 2012. *http://archive.ics.uci.edu/ml/datasets/Individual+household+electric+power+consumption #*

[3] K. Murphy, Machine Learning: A Probabilistic Perspective. MIT Press, 2012.