1. Introduction

Nowadays, where global demand for energy increases and prices increase, there is also a strong need to reduce waste when possible. Thus, a prediction of energy usage is an important step that can allow humans to use energy more efficiently, and consequently more economically. This is crucial to reduce the negative consequences of global warming. To meet international regulations regarding environmental issues, such as the 2DS pathway for global climate change mitigation designed by International Energy Agency (IEA) [1]. The pathway aims to keep an increase in global temperature lesser than two Celsius degrees by 2100. To meet the purpose it is highly required to maintain annual growth in energy consumption at the rate of 1.2%. However, the current rate is 2.9%. Thus, it is desirable to create a system that can forecast energy use so that IEA's establishments can be met. Also, a successful prediction of energy usage would reduce wastage, which is highly advantageous for energetic companies, governments bodies, and human beings.

Energy usage prediction is a great task for Machine Learning (ML), which can be seen as nonlinear time series with numerous of complex factors/variables. These include a variety of building stocks such as public services, residential and industrial, weather changes, as well as periods of a year such as holidays, Christmas, etc. Also, there might be sudden changes in energy consumption due to unexpected events like equipment failure or blackout. Thus, it is extremely important to develop models which will have a wide exposure to various factors and will bring accurate results. Precise predictions will help us in achieving the following goals:

* Creating a reliable ML model that can effectively predict the use of energy in various categories of buildings in the following areas: hilled water, electric, hot water, and steam meters, and different weather conditions
* Gaining valuable insights into factors affecting a building’s energy demand, which allows managers to improve energy efficiency
* Enable managers to identify anomalously high/low energy consumption and alert them to problems with buildings

1. Description of the task and dataset

Dataset comes from Kaggle’s ASHRAE competition. It consists of data/readings from over 1,000 buildings over a three-year timeframe. The dataset consists of five CSV files. The building data file consists of 6 variables that provide information on buildings’ primary use, covered area, built year and floor count with a number of values ranging from 0 to 1448. Next, there are training and test data files for weather readings. They consist of 9 variables that provide information on air temperature, cloud coverage, dew temperature, precipitation depth, sea pressure, wind direction and wind speed. Lastly, there are two files, test and training both of which provide details on buildings meter readings. Overall there are thirty-four columns that have four data types: decimal, integer, date and string.

To measure a quality of developed models the following evaluation metrics: Root Mean Squared Logarithmic Error (RMLE).

The RMLE is calculated as

Where:

– RMLE value(score)

n- number of observations in the public/private dataset

– prediction of target

– actual target for i

log(x) – natural logarithm of x

To ease the analysis of the dataset, merging techniques were used. A merge was performed on three train data files and two test files to obtain a single test and training dataset. It is crucial to observe that nine out of sixteen variables have large missing values that need to be adjusted at a later stage. In order to explore timely variations, we broke the timestamp variable into six new columns: hour, day, dayOfWeek, dayOfYear, month and year in order to explore timely variations in data.

2.1 Exploratory Data Analysis

Once the data was combined we started exploring the target variable of meter reading. Firstly, a log transformation of the variable was taken to adjust for high skewness and then plot a density graph that shows a good variation in values along with a high number of 0-meter reading values. Then an exploration of any seasonality changes was made by plotting meter readings against time.

Chart, line chart

Description automatically generated

Figure 1 shows energy usage throughout the day. The energy usage is reasonably very low during the early morning hours and fairly high during the evening as it is the peak working time of all operating sites such as educational institutions and industrial areas.

Chart, line chart

Description automatically generated

Figure 2 shows the energy usage per annum. It can be observed that energy consumption is fairly low during the starting months of the year, rises sharply in the spring season and fluctuates during the summer season. June-September summer months report the highest level of energy consumption which may be due to the high AC usage in offices and institutions

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

Both bar plots show the distribution of meters at the building (fig.a) and energy consumption from each of these meter types(fig. b). Electricity meters are the most commonly used by chilled water and steam meter types. Also, steam and electricity meter types consume the greatest energy followed by chilled water type. It might be useful to replace hot water meter type with electricity type as it can greatly save energy.

Graphical user interface

Description automatically generated

Graphical user interface, chart, line chart

Description automatically generated

Since data assumes energy consumption in various types of buildings, it is worth exploring meter reading distribution based on primary usage in different areas. It is seen that educational institutions, offices and retail sites have the most energy consumption during the morning and evening time of the day and the least consumption during the night. This result is quite expected as these sites mostly have fixed operational timings. For entertainment and public assembly sites, there is low energy usage during night times and greater energy usage during evening times. For residential areas, there is a sharp decline after midnight and then meter reading keeps on increasing and reaches a high level and remains stable until midnight. This result is directly related to the higher level of activities being performed throughout the day in a house that utilizes various appliances. Also, an overall analysis of the graphs shows that utility, industrial, healthcare and food sales sites report higher levels of energy consumption whereas worship areas and retail sites consume a lower level of energy.

2.1.1 Analysis of weather data

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

A density plot graph shows that air temperature variable follows a normal distribution. The mean value of air temperature is between 14 and 15 degree Celsius and most values lie between the range of 0-30 degrees. Cloud\_coverage is measured between a 0 to 9 scale where 0 means it is a clear sky and 9 means it is rainy. It can be observed from the density plot that most of the cloud coverage is zero. Sea level pressure follows a normal distribution with most values in the range of 1000-1025. Dew temperature has a skewed distribution with most values between 0-25 degrees

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Chart, radar chart

Description automatically generatedThe windrose diagram shows that for most of the sites, the wind blows from north direction(90 degrees) most of the time, followed by south direction(270 degrees). It also shows that wind blows the least from NE direction(45 degrees). When the wind blows from the north, maximum of the times the speed is between 0 to 3.8 m/s.

2.2.2 Analysis of building data

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Chart, line chart, histogram

Description automatically generated

From the above analysis, it is clear that most of the buildings were built around 1975. The growing trend is maintained before 1980. A second growing trend is observed after 2000 with a peak around the 2010 year. Average meter readings greatly fluctuate with the year built variable with no clear trend. Such fluctuations were difficult describable and there is no valid rationale why they happened. Moreover, most of the buildings have 0-15000 covered square feet area and they have two floors. For a floor count greater than 10, meter reading has a drastic increase, then around 14 floors have a sharp decrease and then again for floor count greater than 15 there is a drastic increase.