Machine Learning Project

Title - Energy Usage Prediction.

1. Introduction

In 2021 the operation of buildings accounted for 30% of global final energy consumption according to a statistics study. In this assignment, we are tasked with developing a machine learning model that predicts energy usage based on historical operational big data and observed weather readings. we began by conducting an exploratory analysis of the dataset, which allowed us to gain a better understanding of its characteristics and identify any potential issues that would need to be addressed in the pre-processing phase. Next, we pre-processed the data to prepare it for analysis. This included merging the datasets, optimizing memory usage, handled missing values and ensuring that all variables are in a suitable form. Lastly, we implemented various machine learning models, trained and tested them with the processed datasets.

2. Literature review/Related Work

There have been many distinct efforts made in the past to predict energy usage, some of which include the following:

M. Elementally et al [1] used artificial neural networks (ANNs) to predict the energy consumption of an office building with one-year data. They tested several ANN models like recurrent neural network and feedforward neural networks with one and two hidden layers. The later resulted to being the best performing model predicting energy consumption with a mean absolute percentage error (MAPE) of less than 3%. The models used were able to make accurate enough predictions, flexible in application use and handled huge data, however, they are also difficult to understand and complex computationally and do not generalize well to new set of data.

Another paper by A. I. Faruque et al [2] for an academic building using data collected and trained over a year's period showed linear regression model technique was able to accurately predict energy consumption with a MAPE of 4.8%. Although this technique is manual, it has a limited prediction accuracy and assumes a linear relationship between the input variables and the output variable, and a normal distributions or residuals, it is simple, easy to understand and apply, trains fast and flexible.

A. I. Akinola et al [3] presented a case study on the prediction of energy consumption in a public building using multiple linear regression (MLR). Results showed high level of accuracy in predicting energy consumption, with a MAPE of less than 10% and coefficient of determination (R-squared) value of 0.94. It also found the building's energy consumption was significantly affected by the number of occupants and time of day. MLR in this study showed transparency by showing clear understanding of how the predictors relates to energy consumption, accommodation of new variables and data sources and efficiency, also, an assumption or linearity between predictors and energy consumption.

T. Hong et al [4] wrote an article on a comparative performance of support vector regression (SVR) and artificial neural network (ANN) models in predicting the hourly energy consumption of the building using MAPE, root mean squared error (RMSE), and coefficient of determination (R2). SVR outperformed ANN with an MAPE of 7.24%, R2 score of 0.86 and RMSE of 32.10 kWh. Major takeaways from this article include both SVR and ANN models, benefited performance by using principal component analysis (PCA), feature selection data pre-processing techniques and selecting appropriate hyperparameters like kernel functions & regularization parameters, and also with building automation systems to provide real-time consumption prediction and the performance of the models varies depending on the specific building and energy consumption data used.

Research by R. Jain et al [5] investigated the use of SVR for forecasting energy consumption in multi-family residential buildings and evaluated the impact of temporal and spatial monitoring granularity on model performance accuracy. A performance of R2 values ranging from 0.81 to 0.93, RMSE values from 16.77 to 35.88, and MAE values from 10.86 to 22.28 were reported. They found the temporal granularity from hourly to daily or weekly and expanding the spatial granularity from building level to apartment level, resulted in accuracy of the model used and allows for a more detailed modelling and analysis of energy consumption patterns. Also, they implored a balance between granularity and model complexity as the use of low temporal or spatial resolution can result in an oversimplified models that may not capture the complex usage patterns of multi-family residential buildings.

Depending on what we have found from these literature review, we decided to implement linear regression, support vector machine, random forest and LSTM models for this project.

3. Description of the task

To predict the energy usage of various buildings, we use machine learning models that considers various relevant variables to the power consumption/energy usage which includes demographic information such as Building size, building type, weather. Also, since it was a time-series dataset, we believed we should also consider time factor when we implement the models.

Once we have identified and prepared the relevant variables, we trained the machine learning models using the given and processed dataset that includes information on historic readings. Then we used these trained models to make predictions in the future energy consumption of the corresponding buildings and evaluated their performance using different elevation metrices.

By using machine learning model to predict energy readings, we believed it can effectively reduce building energy consumption, as to manage cost and improve energy efficiency.

4. Exploratory Data Analysis

We have tried to explore datasets and their features individually to understand the data pattern in the data frames.

Four separate datasets were given and downloaded which included:

train.csv: included historical meter reading from different buildings, with a given building id, type of meter and the time recorded.

building_meta.csv: included the features from different buildings with given building id, like primary usage, floor area, floor count and year build. Also included site id to represent the location of the building.

weather_[train/test].csv: Weather data from a meteorological station as close as possible to the site which showed recorded temperature, cloud coverage, sea level, wind direction and wind speed with corresponding site id.

test.csv: raw data that included only building id, type of meter and corresponding timestamp. It was provided for us to test the developed models. However, it was not used as it didn't contain the real data for evaluation. We used the train-test split method on the training dataset instead.

We merged the **train.csv, building_meta.csv and weather_[train/test].csv** file into a single data frame using the primary and foreign keys. It was a huge dataset so to handle it we used some to reduce the memory usage. Also, since it was a time series data, we extracted the timestamp feature into hour, day, month and year by the python date and time feature for analyzing the data and their patterns periodically.

4.1. Meter types

We analyzed the first categorical feature- meter, which are four types of meters (0: electricity, 1: chilled water, 2: steam, 3: hot water). We can observe that the proportion of electricity data is much more than the other meters.

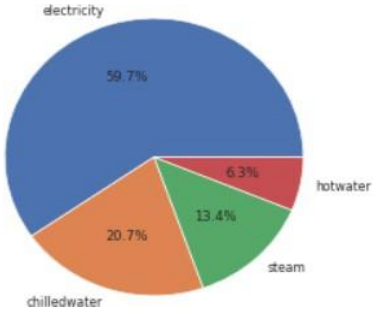


Figure 1. Proportion of different type of meters in the dataset

4.2. Meter reading

Then we analyzed the target variable - meter reading and checked the distribution pattern of it. We observed that it was heavily skewed and could not gather much information from it. Therefore, we did log transformation to it and the result as below:

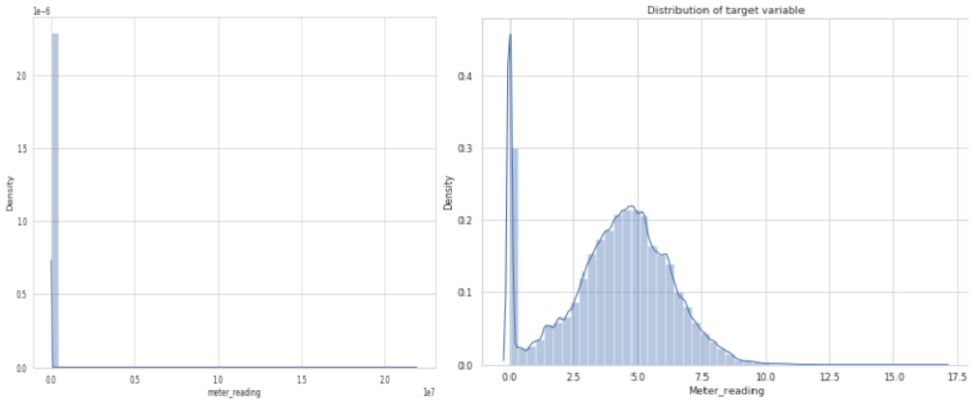


Figure 2. Distribution of meter reading

Then we checked the periodic patterns of meter reading with respect to the time in a day and the month in a year.

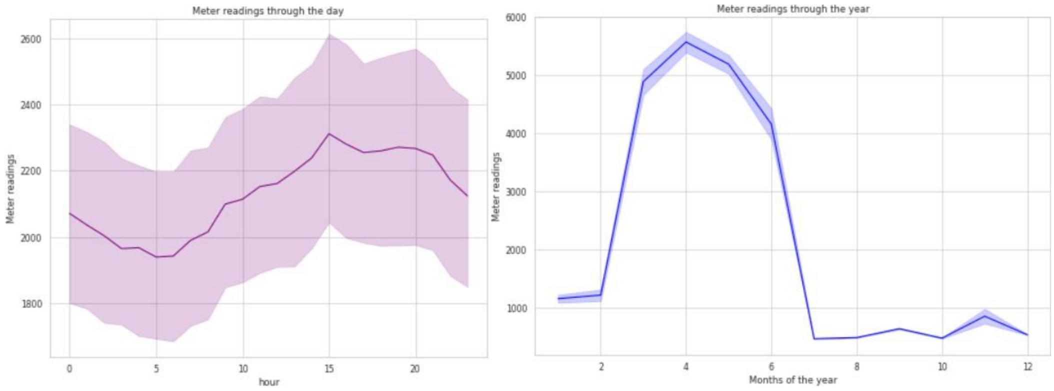


Figure 3. Periodic pattern of meter reading against time

It was observed that energy consumption was more in daytime than in nighttime. Also, in the second chart it was shown that energy consumption is higher in summer than winter, it was quite contradicting the problem statement, but we still went further.

4.2.1 Checking for outliers in meter reading

We also tried to locate any outliers in the meter reading of different meter types:

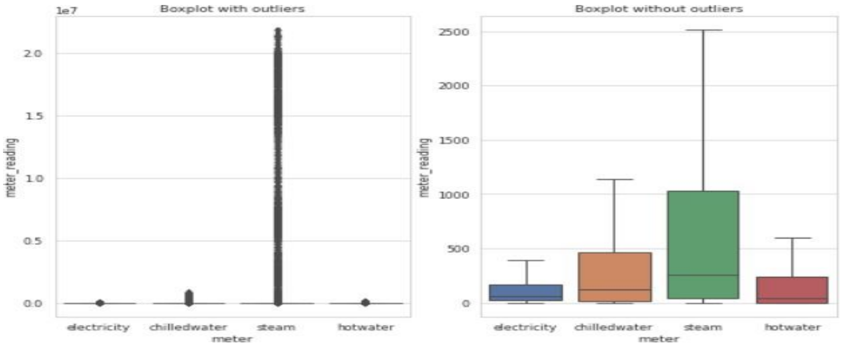


Figure 4. Boxplot of meter reading against types of meters.

We also tried to compare the average meter reading from different sites and buildings:

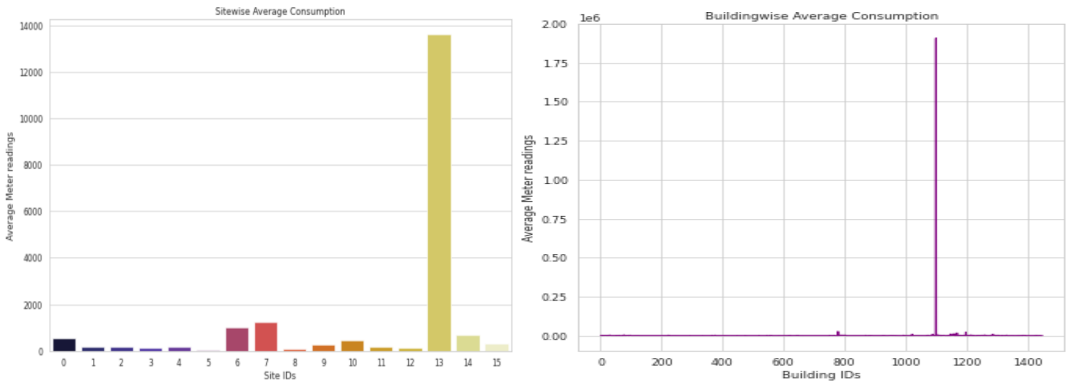
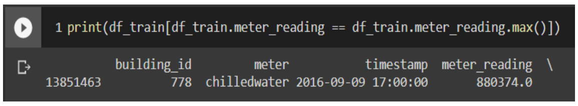


Figure 5. Average meter reading of different sites and building

It showed a similar pattern because of the outlier, we can see that site 13 has enormously higher energy consumption. The second chart also showed a huge spike of usage in one of the buildings which indicated the outliers in the data. Then we tried to find out which building it belonged to:



```
print(df_train[df_train.meter_reading == df_train.meter_reading.max()])
```

```
          building_id        meter            timestamp  meter_reading  \
13851463          778  chilledwater 2016-09-09 17:00:00       880374.0
```

Figure 6. Building ID with the highest meter reading

4.3. Primary usage

We also examined the distribution of different primary usage:

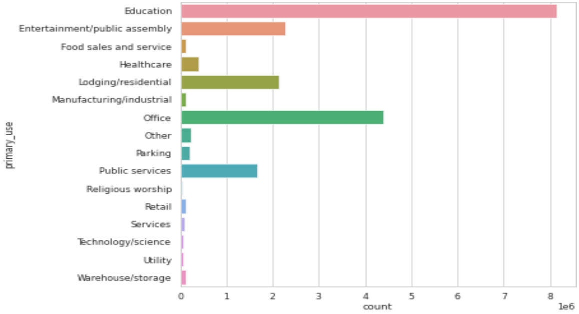


Figure 7. Distribution of buildings with different primary usage

It was found that most of the readings came from educational buildings, followed by the office space, entertainment and public assemblies respectively.

We also tried to find out how they were distributed hourly by graphs below:

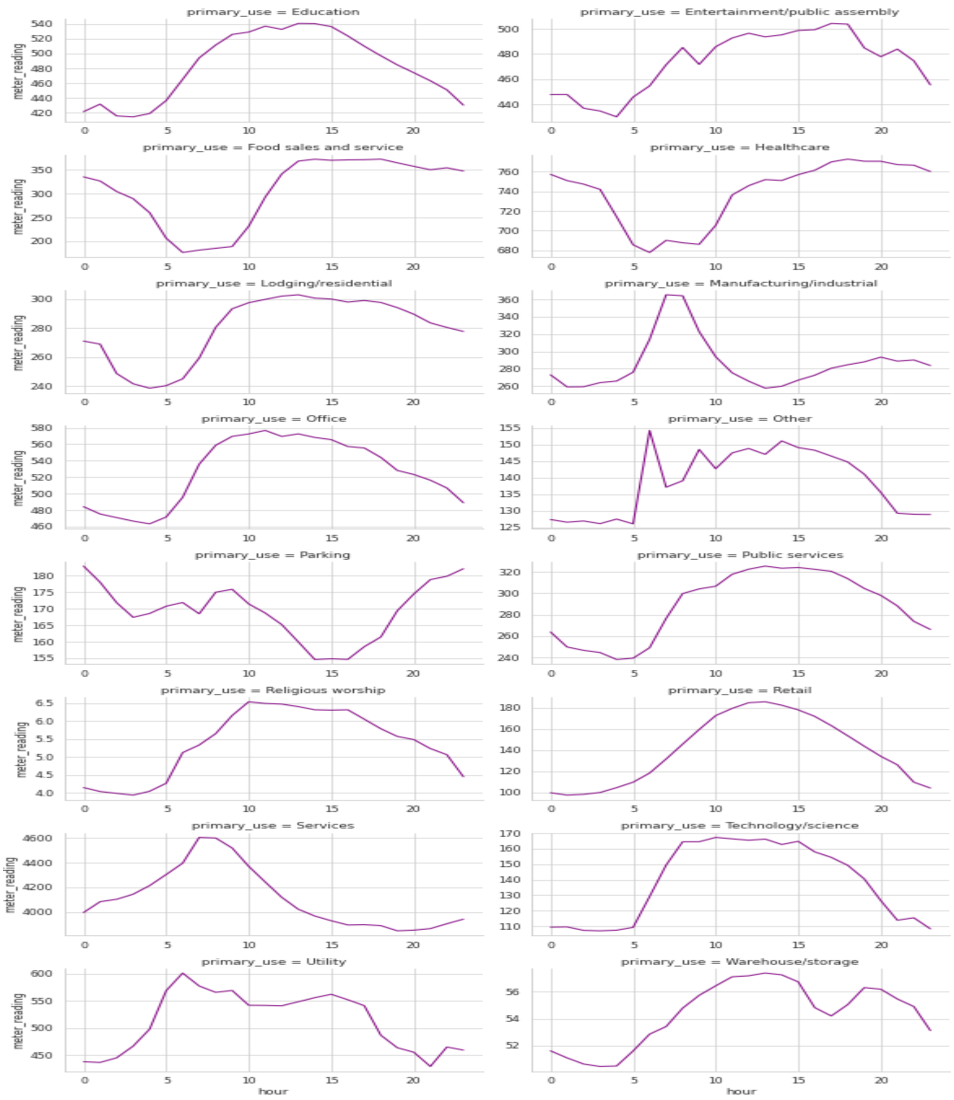


Figure 8. Energy consumption of different primary usage against time

4.4. Weather features

Then we looked at the features in weather data, air temperature was directly proportional to the energy usage for households and the public premises.

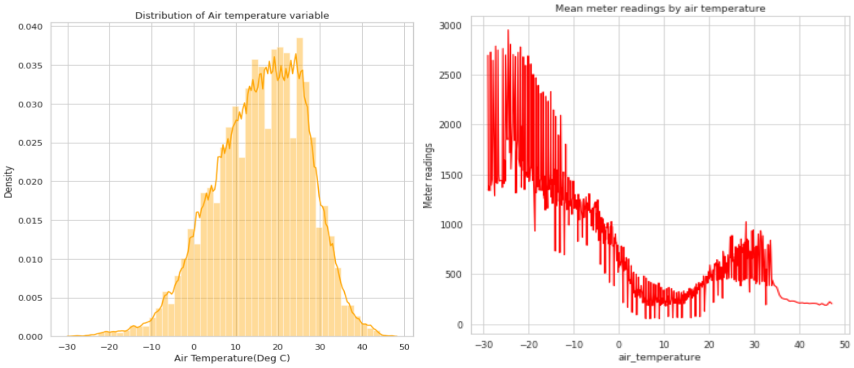


Figure 9. Distribution of air temperature and the meter reading against it

Air temperature followed a normal distribution which lies between 0-30 degree Celsius.

It was found that meter readings were higher when temperature was negative and above 15 degrees Celsius. It was probably because more energy was consumed by heaters and air conditioners.

The distributions of other weather variables are shown below:

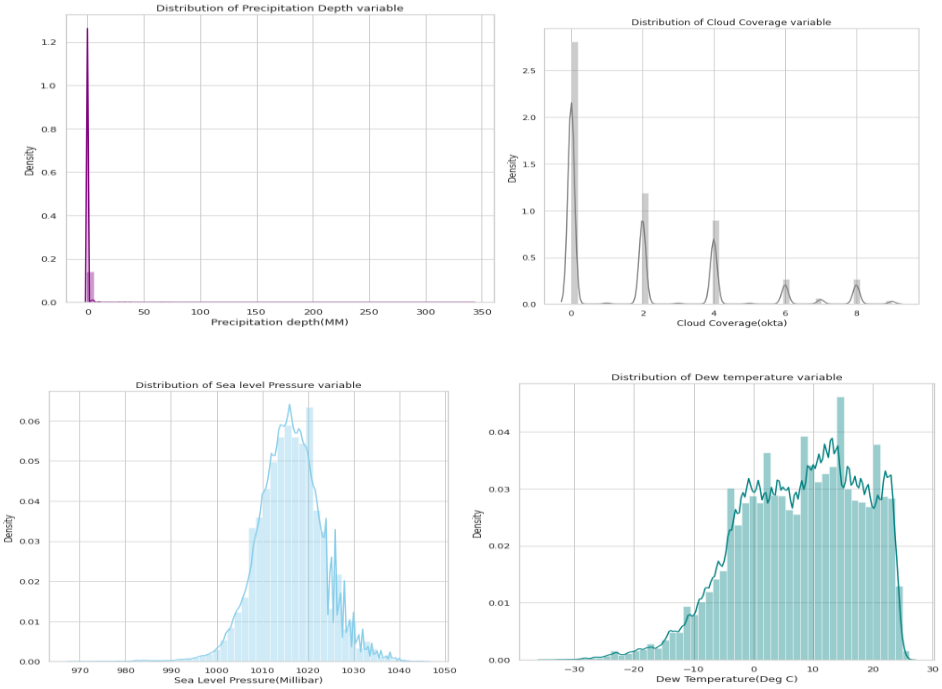


Figure 10. Distribution of other weather variable

4.5. Floor count

Also checked the meter reading against floor count:

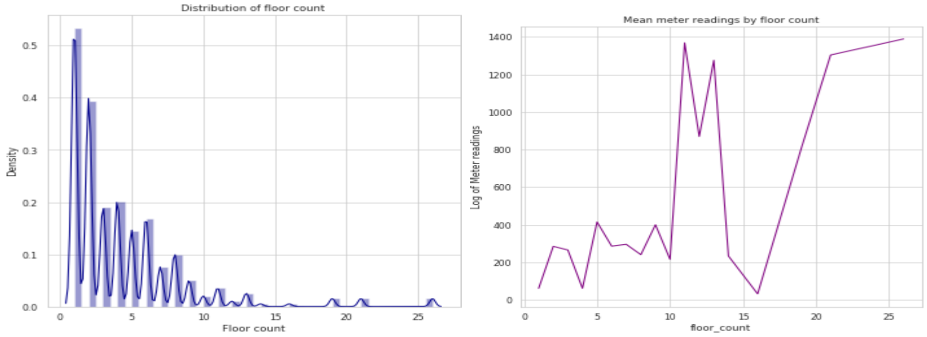


Figure 11. Distribution of floor count and meter reading against it

Generally, buildings with more floors consume more energy.

4.6. Year-built

Then we moved to the year-built feature.

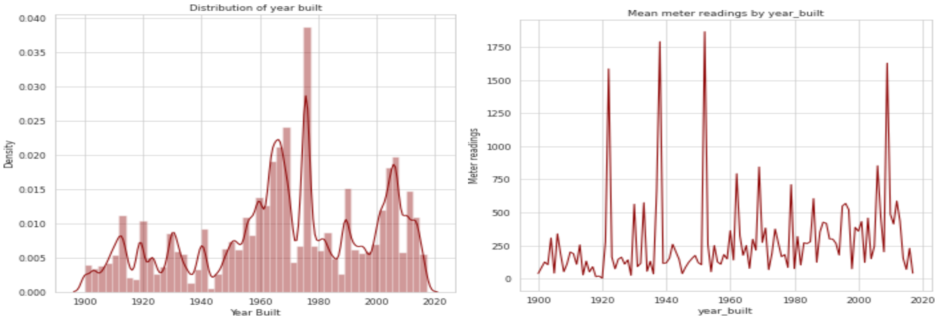


Figure 12. Distribution of year build and meter reading against it

It showed data over 120 years and generally more energy were consumed for the buildings built before 1960.

4.7. Correlation matrix

The correlation matrix below summarizes the relation of features in the dataset.

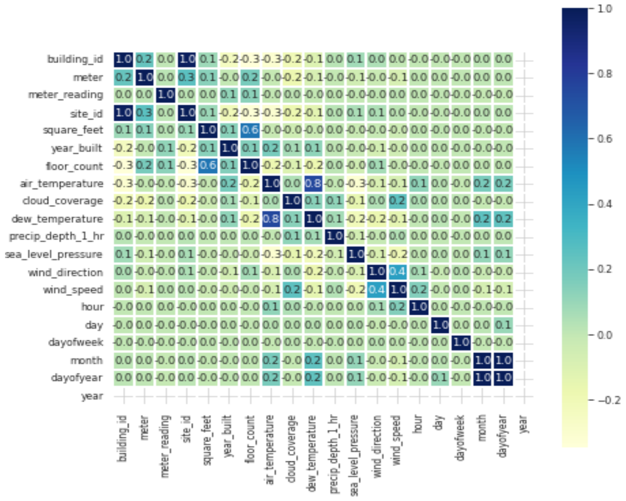
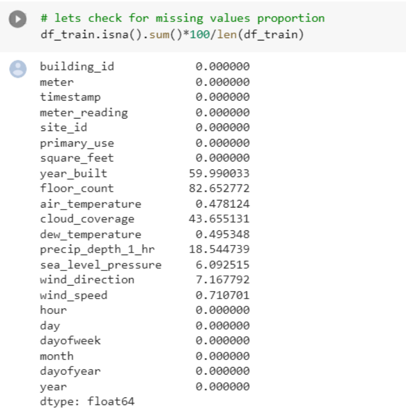


Figure 13. Correlation matrix of features

4.8. Missing values

Also, the number of missing values in different features was found and would be further handled in the following section.



```
# lets check for missing values proportion
df_train.isna().sum()*100/len(df_train)

building_id            0.000000
meter                  0.000000
timestamp              0.000000
meter_reading          0.000000
site_id                0.000000
primary_use            0.000000
square_feet            0.000000
year_built            59.990033
floor_count           82.652772
air_temperature        0.478124
cloud_coverage        43.655131
dew_temperature        0.495348
precip_depth_1_hr     18.544739
sea_level_pressure     6.092515
wind_direction         7.167792
wind_speed             0.710701
hour                   0.000000
day                    0.000000
dayofweek              0.000000
month                  0.000000
dayofyear              0.000000
year                   0.000000
dtype: float64
```

Figure 14. Proportion of missing value in different features

5. Data Preprocessing

5.1. Dealing with missing values.

It was found that there were many missing values in the dataset. For example, more than 50% of the data in floor count and year-built were missing. We simply dropped these features.

For the other features with less than 50% value. We filled them with median values or dropped the entire rows.

5.2. Outlier Treatment.

The identified outliers in the previous section were dropped to prevent the training being affected by extreme data.

5.2. Normalization

One hot encoder was used to categorical features in the dataset which included ‘primary use’, ‘meter’, ’site_id’, ‘hour’, ‘dayofweek’ and ‘month’. While min max scaler was used to handle the remaining features which were all numerical. The encoder and scaler were combined in a column transformer which is then applied to the input training data while only min max scaler was applied to the output data ‘meter_reading’.

6. Models implementation

6.1. Linear Regression, SVM, and Random Forest Models

For these three models, only *building_id*, and some of the time variables (*day*, *dayofyear*, and *year*) were excluded from the data, mostly to reduce the number of columns when encoding and because there were other columns to cover the time without these. Among the dataset, 70% and 30% of the data was split into training and testing set respectively.

6.1.1. Linear Regression

As a basis to compare from, the first model we used was a linear regression since it is an easy to fit model. Indeed, it was fast to fit on the entire dataset but a big problem with the linear regression is that some of the predictions are negative (see figure.15) which is not possible.

6.1.2. SVM

The other two models, support vector machines and random forests, were mentioned a few times in other literature so we also decided to try those. Starting with a SVM, which also didn't take long to fit on the whole dataset. However, it performed extremely poorly so we decided not to pursue this model.

6.1.3. Random Forest Models

The random forest model was by far the most promising, however we estimated that training it on the whole dataset would take roughly 9 hours, which is far too long before considering hyperparameter optimisation. As a result, we decided to take a small sample of 20,000 rows of the data to train and optimise the model on. We performed hyperparameter optimisation on the maximum depth of the trees, the minimum number of samples required to be at a leaf node, the minimum number of samples required to split an internal node, and the number of trees in the forest. We used grid search CV because it would still take a long time even with predefined values, and it did take almost 5 hours. Interestingly, the base model without hyperparameter optimisation performed better than the one after hyperparameter optimisation. With more time it may have made more sense to use grid search to get an initial idea of the parameter values, and then used random search around those values to fine tune.

6.1.3. Model evaluation

Below are some plots showing the predictions of each model on a random building in the validation dataset:

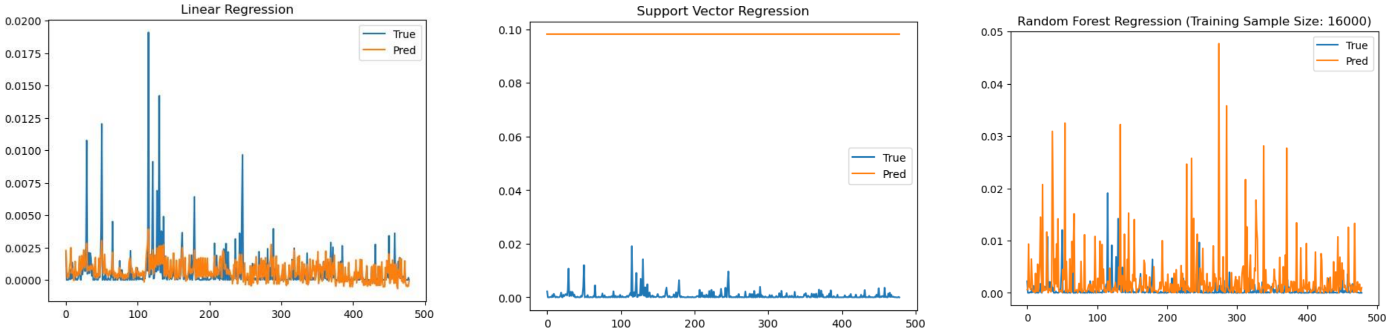


Figure 15. Performance of Linear regression, SVM and random forest on validation dataset by comparing prediction and actual values

6.2. Long short-term memory (LSTM) networks

LSTM networks provide 'memory' for older information and has specific ability to deal with time series data. It was used as one of the models to predict energy consumption in this project.

6.2.1. Train-test split

Due to the special feature of LSTM model, we picked a specific building to implement the LSTM model instead of the whole dataset. In this project, building 46 was used as an example to train this model. The data was sorted by time series. Therefore, different from the previous approach, we took out the time-related features from the input set and the first 80% of the data were used as the training set while the other 20% of the data were used as the testing set.

6.2.2. Convert series to supervised learning by Jason Brownlee

A pre-designed function that takes a univariate or multivariate time series and reframes it as a supervised learning dataset. Time step was set to 2 in this example. After splitting into input and out set and before training the model, the input dataset had to be reshaped into a 3D structure.

6.2.3. Design of LSTM network

This LSTM network consisted of 3 layers, including an input LSTM layer with 50 unit, a hidden layer with 100 unit and ReLU activation function and an output layer with a single unit. Between each layer, dropout was inserted for better performance of the network. Besides, early stopping was applied to prevent overfitting.

MAE was used as the loss function and adaptive moment estimation (adam), a keras pre-defined optimizer, was used as the optimizer during training for parameter optimization of the network. Epochs and batch size were set to 70 and 72 respectively.

6.2.4. Model evaluation

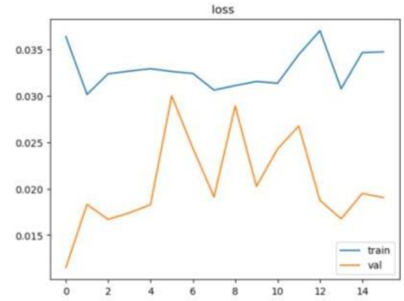


Figure 16. Training history of LSTM model

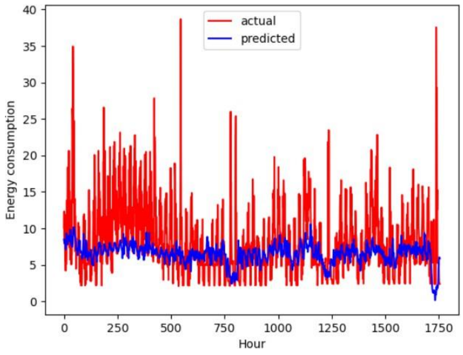


Figure 17. Performance of LSTM model on validation dataset by comparing prediction and actual values

7. Results

7.1. Model evaluation

	Linear Regression (LR)	Support Vector Machine (SVM)	Random Forest (RF)	LSTM network (LSTM)
R-squared score	0.015	-187.247	0.393	0.153
Mean absolute error	0.001	0.098	0.002	0.093
Root means square error	0.007	0.098	0.014	0.134

Table 1. Evaluation matrices on different models

Based on the table above, one could conclude that SVM has the worst fit. LR had a much better fit than SVM however not as good a fit as RF and LSTM due to the limitation stated before. Both random forest and LSTM showed good performance. However, in terms of efficiency, since random forest took a much longer time to train which would be quite infeasible in real life, it was better to use LSTM model for time-series prediction.

7.2. Further analysis on the project

If we compare the MAE and RMSE we could see that there was large variation between the errors that the LSTM model produces a lot of variation between errors since RMSE_LSTM – MAE_LSTM = 0.041 which was a largest difference. In statistical terms, a higher variance in the individual errors can make it more difficult to estimate the true population parameters from the sample data, as there is more random fluctuation in the data. It can also increase the margin of error and reduce the accuracy of statistical inferences based on the sample. This further adds to the previous comments mentioned above that MAE of LSTM is the least accurate of all the models.

Also, apart from the error that the models could make as mentioned before, some other error might be unintentionally included. For example, due to the limitation of memory, we had to use a subset of the dataset instead of the whole dataset and might affect the model performance. Besides, there could be other factors affecting the energy consumption except those we were given such as human factors. For example, a lower income family would tend to save energy consumption.

8. Conclusion and future work

It was found in this project that LSTM model and random forest could both perform well in energy consumption prediction. However, in term of efficiency, we would still prefer LSTM rather than random forest due to the huge difference in training time. Apart from these, there are still future works could be done. For example, there were still more types of models that could possibly work for this issue such as a convolutional neural network. Furthermore, an advanced platform with better memory usage could possibly help our models perform better so we can train them with a whole dataset instead of a subset. Finally, apart from the factors that we were given, we should also consider some other factors that may possibly affect energy consumption to get a better picture.

References

[1] Mohamed Elmitwally, Ahmed Al-Masri, Essam Eldin Khalil, and Ahmed Fahmy, “Energy Consumption Prediction Using Artificial Neural Networks: A Case Study of an Office Building” published in Energies in 2018.

[2] A. I. Faruque, A. V. Babu, M. T. Iqbal, and M. F. Akbar, “A Data-Driven Approach to Energy Use Prediction in an Academic Building Using Linear Regression" published in Energies in 2020.

[3] A. I. Akinola and A. M. Hassan, “Energy Consumption Prediction of Public Buildings using Multiple Linear Regression: A Case Study" in International Journal of Innovative Technology and Exploring Engineering in 2019.

[4] T. Hong, H. Sun, Y. Chen, Q. Yan, L. Lu and W. Tong, "Predictive Models for Energy Consumption of Commercial Buildings using Support Vector Regression and Artificial Neural Networks" published in Energy and Buildings in 2018.

[5] Rishee K. Jain ,Kevin M. Smith ,Patricia J. Culligan and John E. Taylor, Forecasting energy consumption of multi-family residential buildings using support vector regression: Investigating the impact of temporal and spatial monitoring granularity on performance accuracy, Volume 8 in 2012.