

# PROJECT 2 FINAL REPORT

## Optimizing Energy Efficiency: A Comprehensive Analysis of Household Electricity Consumption for Cost Savings

*Harish Balasubramaniam (hb425)*

*Niyati Jain (nsj39)*

*Sarthak Dalal (spd115)*

### INTRODUCTION

In the contemporary era, where sustainable living and economic prudence are at the forefront of societal concerns, understanding and optimizing household electricity consumption have become paramount. The dataset at hand, spanning from January 2007 to June 2007, provides a minutely detailed account of electricity usage in a household located in Sceaux, Paris, France.

The increasing demand for energy necessitates a proactive approach towards optimizing electricity consumption. Leveraging advanced data analytics techniques like Time-Series Analysis presents an opportunity to gain actionable insights from historical consumption data. This project aims to investigate the efficacy of different modeling approaches in predicting electricity consumption, thereby empowering individuals to make informed decisions towards a more energy-efficient lifestyle.

### OBJECTIVES

The primary objective of this project is to delve into this wealth of data and unearth insights that empower individuals to make informed decisions, thereby curbing unnecessary expenditure and contributing to a more energy-efficient lifestyle.

We aim to achieve this through Machine Learning modeling of the data, specifically through Time-Series Analysis, which would try to closely predict the expected electricity consumption on a future date, based on the utilizations between the months of January 2007 and June 2007.

## PROBLEM STATEMENT

We have mentioned the problem statement in a few points that emphasize what we want to achieve in this project:

1. To develop robust predictive models that accurately forecast electricity consumption on future dates.
2. To compare the predictive performance of Model 1, which relies solely on date-time information, against Model 2, which incorporates sub meter readings along with date-time information.
3. To identify the most effective modeling approach that empowers individuals to make informed decisions, thereby curbing unnecessary expenditure and promoting energy efficiency.
4. To explore insights derived from the comparative analysis of the two models and assess their potential implications for practical implementation in promoting energy conservation.

## DATA COLLECTION

We acquired our dataset from Kaggle

(<https://www.kaggle.com/datasets/thedevastator/240000-household-electricity-consumption-records/data>)

The dataset encapsulates various facets of electricity consumption, ranging from the total active and reactive power to voltage levels and current intensity. Additionally, it provides a breakdown of specific appliance usage through sub-metering values for the kitchen, laundry room, and electric water heater/air conditioner. Through meticulous analysis, we aim to identify patterns, peak usage times, and areas of potential energy optimization.

*Columns:*

Column name	Description
Date	The date of the observation. (Date)
Time	The time of the observation. (Time)
Global_active_power	The total active power consumed by the household (kilowatts). (Numeric)

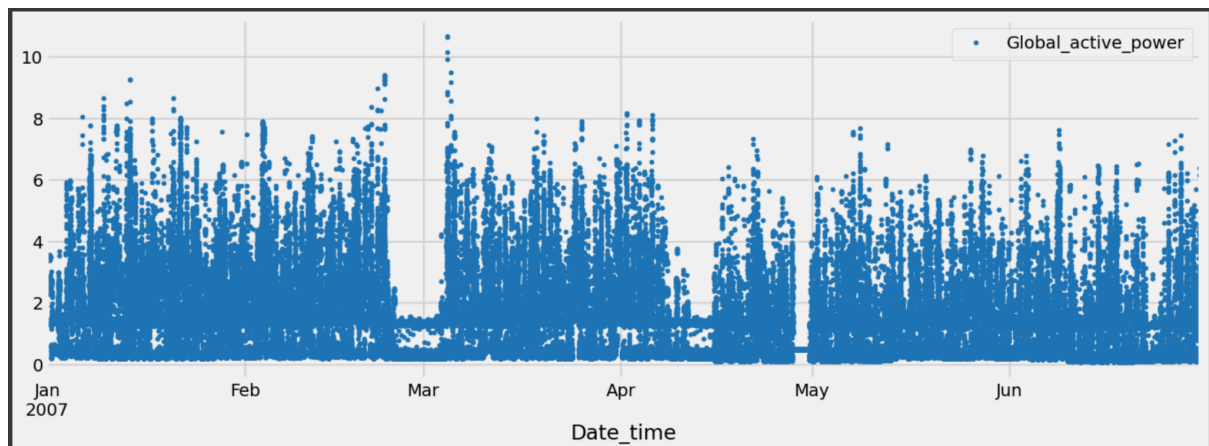
Global_reactive_power	The total reactive power consumed by the household (kilowatts). (Numeric)
Voltage	The voltage at which the electricity is delivered to the household (volts). (Numeric)
Global_intensity	The average current intensity delivered to the household (amps). (Numeric)
Sub_metering_1	The active power consumed by the kitchen (watts). (Numeric)
Sub_metering_2	The active power consumed by the laundry room (watts). (Numeric)
Sub_metering_3	The active power consumed by the electric water heater and air conditioner (watts). (Numeric)

What the dataset looks like:

	index	Date	Time	Global_active_power	Global_reactive_power	Voltage	Global_intensity	Sub_metering_1	Sub_metering_2	Sub_metering_3
	0	1/1/07	0:00:00	2.58	0.136	241.97	10.6	0	0	0.0
	1	1/1/07	0:01:00	2.552	0.1	241.75	10.4	0	0	0.0
	2	1/1/07	0:02:00	2.55	0.1	241.64	10.4	0	0	0.0
	3	1/1/07	0:03:00	2.55	0.1	241.71	10.4	0	0	0.0
	4	1/1/07	0:04:00	2.554	0.1	241.98	10.4	0	0	0.0
...	...	...	...	...	...	...	...	...	...	...
	8995	7/1/07	5:55:00	0.25	0.056	244.79	1.2	0	0	0.0
	8996	7/1/07	5:56:00	0.25	0.056	244.9	1.2	0	0	0.0
	8997	7/1/07	5:57:00	0.252	0.06	245.6	1.2	0	0	0.0
	8998	7/1/07	5:58:00	0.254	0.062	246.33	1.2	0	0	0.0
	8999	7/1/07	5:59:00	0.256	0.064	246.73	1.2	0	0	0.0

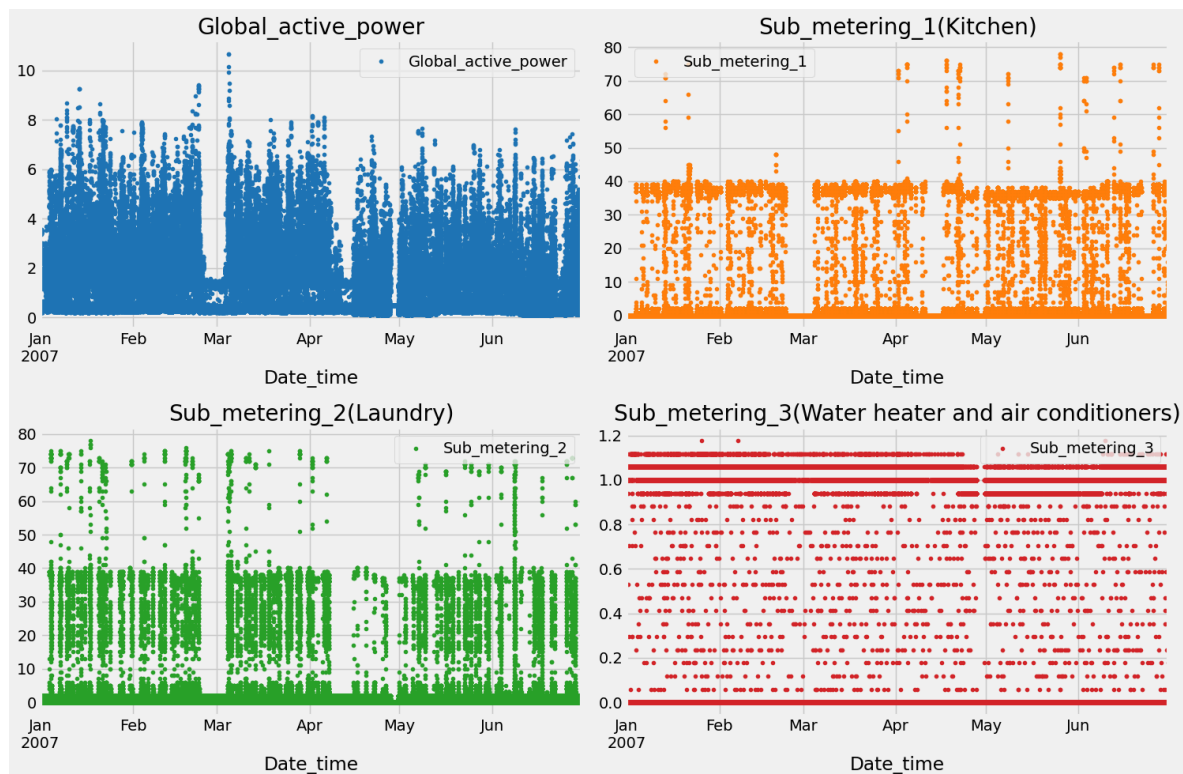
## ANALYSIS

Plot of Date-Time vs Global Active Power:

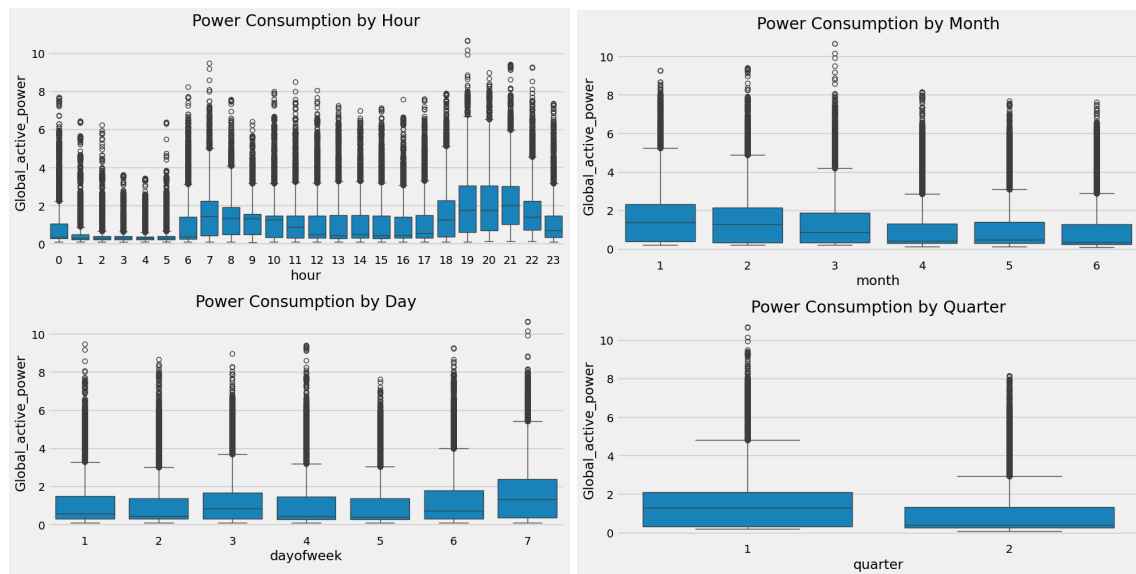


The plot shows us that the peak electricity consumption in the household is greater than 10 kW. There is a dip in electricity consumption for a few days in March, likely suggesting that the residents left the house for a spring break vacation. A similar dip can be observed at the middle of April but this is for a lesser timespan than the one in March.

*Insights acquired from the dataset:*



These graphs show us the plot of the three meter readings vs time. We can see that Meter 1 and Meter 2 have peaks around 80 W while most of the data points are concentrated towards 0 W. This is expected and it suggests that for the most part of the day, the kitchen and the laundry room are not in use but when they do see some use, the utilization is large. Furthermore, there's a noticeable pattern of intermittent high usage in meter reading 2, which may align with specific days of the week designated for laundry. Meter reading 3 on the other hand, shows consistent usage of the water heater and air conditioner, suggesting that these devices are at least on standby mode if not operating in full power mode.



These graphs offer a detailed breakdown of power consumption by hour, day of the week, month, and quarter, which provides valuable insights into usage patterns:

#### *Power Consumption by Hour:*

The graph shows that power consumption varies significantly throughout the day. There is a low usage period during the late night and early morning hours (from 0 to 5 AM), which increases as the day progresses.

A peak in consumption is visible in the evening (around 6 PM to 9 PM), which may coincide with the time when people are generally home from work and using appliances for cooking, heating, or entertainment.

The median values and the spread of the data (as shown by the interquartile range) are highest during the evening, indicating not only higher consumption but also more variability during these hours.

#### *Power Consumption by Day of the Week:*

The power usage does not show significant variation across the days of the week, although the median values on the weekends (days 6 and 7) are slightly higher than on weekdays. This is expected as the residents are at home on weekends and consume more electricity.

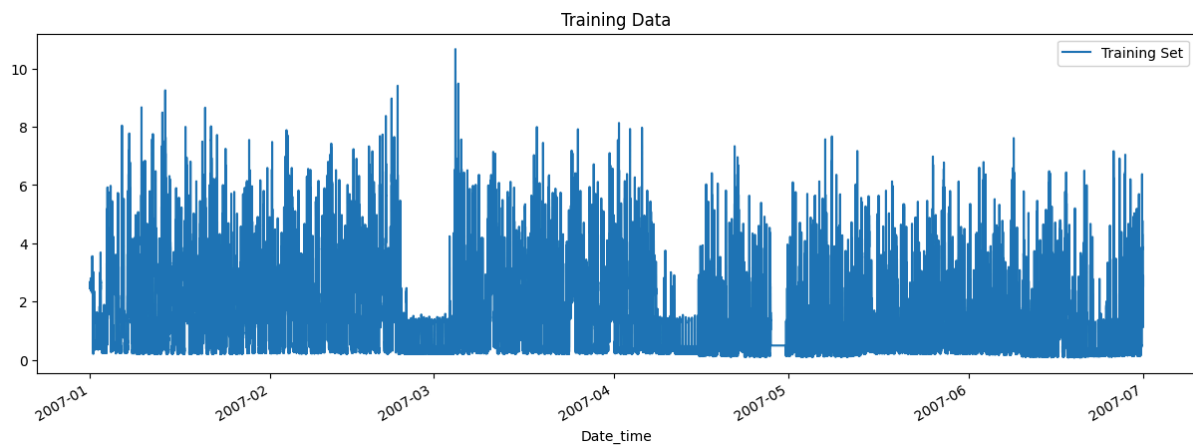
The overall range and outliers are similar across all days, suggesting a consistent pattern of consumption with occasional spikes.

### *Power Consumption by Month:*

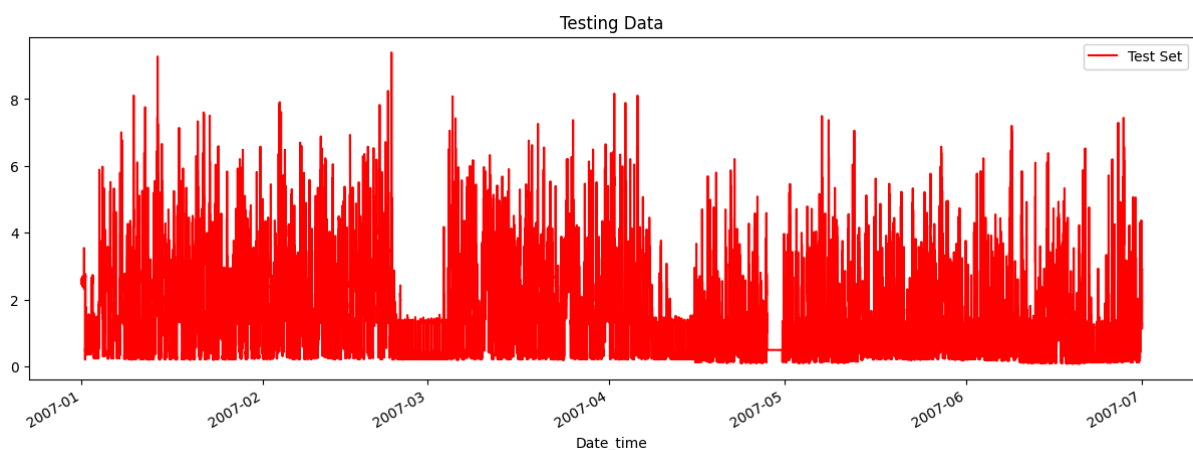
The monthly breakdown shows a noticeable peak in month 3, suggesting a seasonal influence where certain months may have higher energy needs due to heating or cooling demands.

The spread and outliers are particularly pronounced in months 3 and 6, which might correlate with extreme weather conditions necessitating more intensive use of heating or cooling appliances.

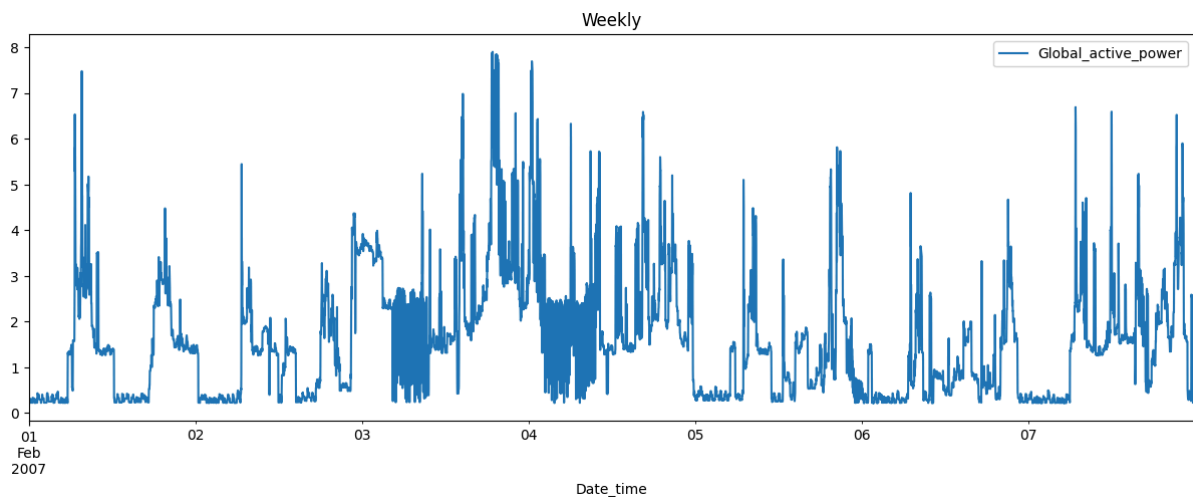
### *Visualizing training data on a plot:*



### *Visualizing testing data on a plot:*



*Plotting the fluctuations of the Global Active Power from February 1, 2007 to February 8, 2007:*



## METHODOLOGIES

1. We combined the separate columns "Date" and "Time" into a single column called "Date\_Time"
2. We dropped the columns "Global\_reactive\_power" and "Global\_intensity" as they do not affect our predictions model
3. We checked for missing values in the dataset, and if found, we replaced them using the "ffill" method, which basically fills the missing values with the nearest non-missing value
4. We split the "Date\_Time" into "hour", "day", "dayofweek", "month", "minute", and "quarter" to facilitate effective Time-Series modelling
5. We used the "Date\_Time" column as our index
6. We split the data into training and testing sets
7. We then create two separate machine learning python notebooks, for Models 1 and 2 respectively, which are mentioned in the Problem Statement
8. This is followed by the fitting of our data into multiple machine learning models capable of performing a Time-Series analysis such as Polynomial Regression, XGBoost, Random Forest and CatBoost
9. On finding the model that gives us the minimum Root Mean Squared Error, we perform GridSearchCV to identify the best parameters for that model, with respect to our data
10. We check the feature importance to gain insight on which feature most affects the predictions of our models
11. Finally, we fit the model onto our test data set to acquire insights on the electricity consumption predictions and compare them with the actual electricity consumptions

## MODELING ANALYSIS

*Dataset after performing the aforementioned preprocessing steps:*

	Global_active_power	Date_time	Sub_metering_1	Sub_metering_2	Sub_metering_3	hour	day	dayofweek	quarter	month	minute
Date_time											
2007-01-01 00:00:00	2.580	2007-01-01 00:00:00	0.0	0.0	0.0	0	1	1	1	1	0
2007-01-01 00:01:00	2.552	2007-01-01 00:01:00	0.0	0.0	0.0	0	1	1	1	1	1
2007-01-01 00:02:00	2.550	2007-01-01 00:02:00	0.0	0.0	0.0	0	1	1	1	1	2
2007-01-01 00:03:00	2.550	2007-01-01 00:03:00	0.0	0.0	0.0	0	1	1	1	1	3
2007-01-01 00:04:00	2.554	2007-01-01 00:04:00	0.0	0.0	0.0	0	1	1	1	1	4

### Modeling Analysis for Model 1:

*Root Mean Square Errors of the aforementioned models on our dataset:*

	Model	RMSE	MSE	MAE
0	Polynomial Regression	1.072555	1.150373	0.798539
1	XGBoost	0.943807	0.890771	0.666697
2	Random Forest	0.334261	0.111731	0.140007
3	CatBoost	0.692457	0.479497	0.454405

We can see that RandomForest gives us the minimum RMSE of 0.334261

Hence, we choose **RandomForest** as our **Model 1**

*We perform GridSearchCV on RandomForest to find the best parameters:*

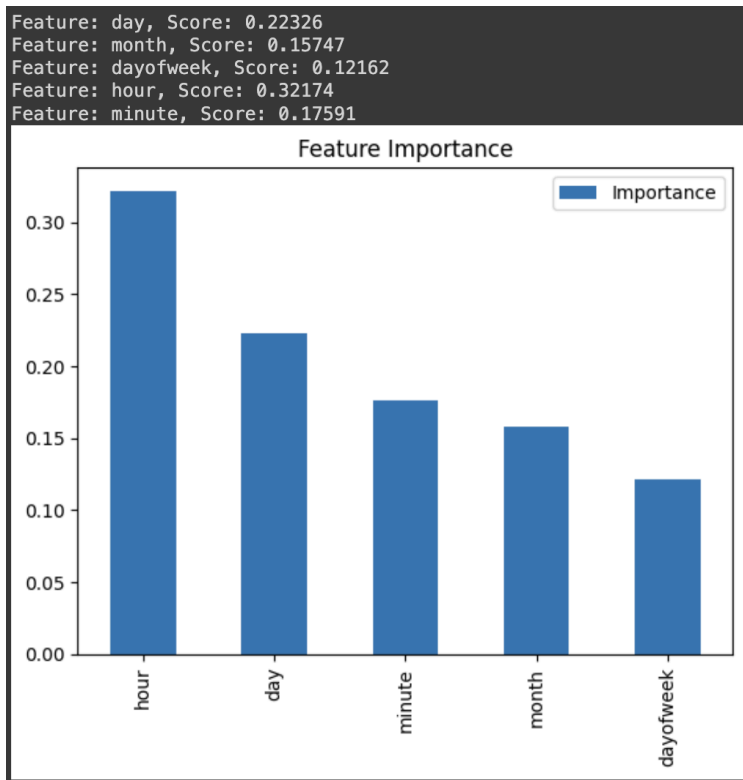
```
Best parameters found: {'max_features': 'auto', 'n_estimators': 300}
```

▼ RandomForestRegressor

```
RandomForestRegressor(max_features='auto', n_estimators=300)
```

*Feature importance in case of Model 1:*





## Modeling Analysis for Model 2:

Root Mean Square Errors of the aforementioned models on our dataset:

	Model	RMSE	MSE	MAE
0	Polynomial Regression	0.602507	0.363015	0.394001
1	XGBoost	0.549705	0.302175	0.350701
2	Random Forest	0.267473	0.071542	0.105481
3	CatBoost	0.384864	0.148120	0.224732

We can see that RandomForest gives us the minimum RMSE of 0.267473

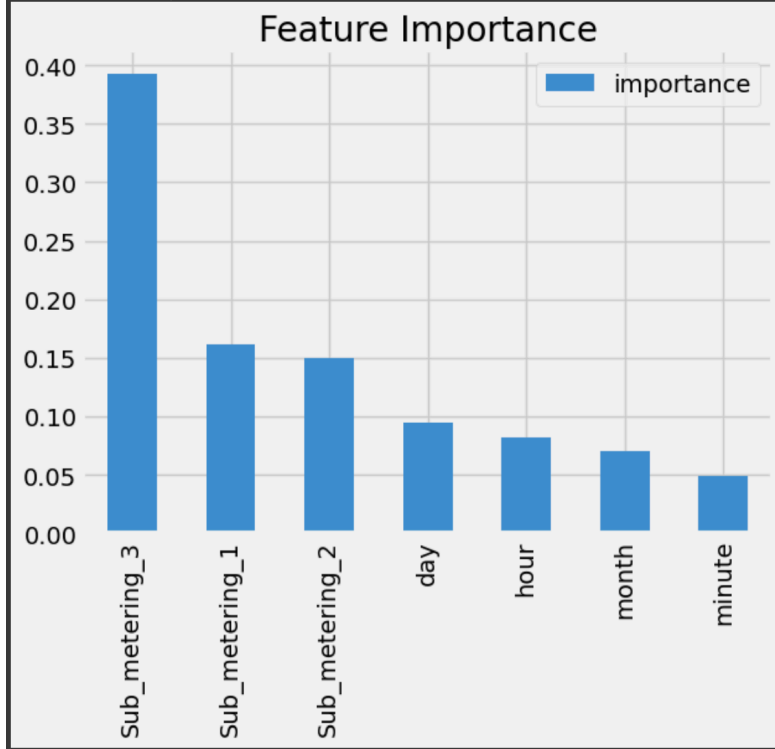
Hence, we choose **RandomForest** as our **Model 2**

We perform GridSearchCV on RandomForest to find the best parameters:

```
Best parameters found: {'max_features': 'auto', 'n_estimators': 200}
▼ RandomForestRegressor
RandomForestRegressor(max_features='auto', n_estimators=200)
```

### Feature importance in case of Model 2:

```
Feature: Sub_metering_1, Score: 0.16193  
Feature: Sub_metering_2, Score: 0.14941  
Feature: Sub_metering_3, Score: 0.39285  
Feature: day, Score: 0.09468  
Feature: month, Score: 0.07036  
Feature: hour, Score: 0.08163  
Feature: minute, Score: 0.04913
```



## RESULTS

Final results when using Model 1 for our predictions:

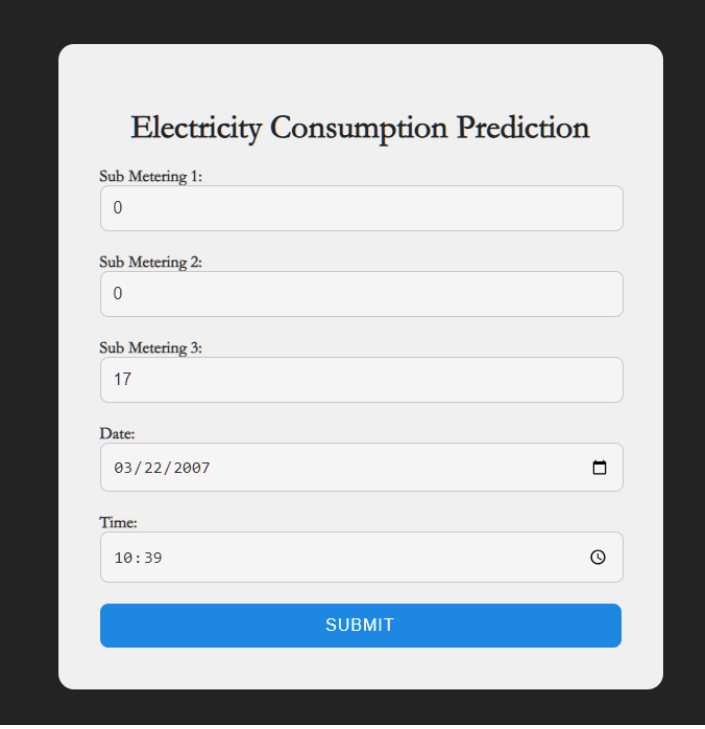
### Electricity Consumption Prediction

Date:

Time:

**Your Predicted Electricity Consumption: 1.392 kW.**

Final results when using Model 2 for our predictions:



The screenshot shows a web form titled "Electricity Consumption Prediction". It contains five input fields: "Sub Metering 1:" with value "0", "Sub Metering 2:" with value "0", "Sub Metering 3:" with value "17", "Date:" with value "03 / 22 / 2007" and a calendar icon, and "Time:" with value "10 : 39" and a clock icon. A blue "SUBMIT" button is at the bottom.

**Your Predicted Electricity Consumption: 1.357 kW.**

## CONCLUSION

In this project, we journeyed through the landscape of household electricity consumption, leveraging machine learning algorithms like Polynomial Regression, XGBoost, Random Forest, and CatBoost to uncover predictive insights from historical data spanning January 2007 to June 2007. Through rigorous experimentation, we found that Random Forest yielded the most accurate predictions, minimizing Root Mean Square Error (RMSE).

Translating these insights into actionable outcomes, we integrated our findings into a user-friendly interface using the Django framework. This fusion of data science and web development empowers users to make informed decisions about their energy usage.

In conclusion, our project underscores the efficacy of ensemble learning techniques like Random Forest in modeling electricity consumption.