AMS 559: Smart Energy in the Information age

Machine Learning for Energy Price and Renewable Energy Prediction

Team members:
Pratik Mukund Velhal (112675099)
Palak Jain (112675008)
Navpreet Kaler (112689117)

Introduction:

As part of our project, we have implemented a machine learning based approach to analyze and predict the prices of energy at a given time based on historical price and weather data, along with analysis and prediction of solar energy generated from weather data. Both of these domains- price prediction and energy prediction are a vital part of the smart energy market. We have devised more efficient techniques to perform these predictions and to get accurate results. We will be discussing the implementation of these two mentioned topics of this project separately.

Background:

In deregulated power markets, the energy market clearing prices (MCPs) are quite volatile. Accurate energy price predictions would assist utilities and independent power producers to submit effective bids with low risks. However, the electricity market has its own complexities since on short time scales, most users of electricity are unaware of or indifferent to its price, and effective storage of electricity is also very difficult. These two facts enforce the extreme price volatility or even price spikes of the electricity market. Various Machine learning algorithms have been used in the past to predict energy prices based on demand and supply information.

Moreover, solar energy significantly reduces reliance on conventional energy sources. If predicted accurately, we can adjust total corresponding provision from other energy sources to meet the same demand.

This is why we have explored these two domains in particular, given the impact they have on the energy sector. In this project, we have leveraged machine learning algorithms to provide more accurate prediction results than existing techniques by making use of seasonal and periodic patterns along with supplementary information like weather data in addition to energy and load data.

Dataset:

We have used the energy price, demand, generation and weather data for Spain available on Kaggle. This is a rich dataset for over four years from 2015 to 2018 with a number of useful features that are used to perform in-depth analysis on the data and develop accurate prediction models. The data has around 34000 rows with each row as an hourly segment of information.

Motivation:

Today, everything including the energy sector is shifting towards a more decentralized approach requiring bidding of prices by the different vendors. In such a deregulated power market, price forecasting has become a very valuable tool and a secret to efficient and cost-effective trading of energy. This is our motivation to work on price prediction models since this would definitely be a huge advantage to the future of the energy sector.

Also as we know, the traditional energy resources are bound to become extinct some day, thus the research interest these days has been shifted to renewable sources of energy like solar energy. These sources can do wonders in the power and energy sector if utilized properly. In order to successfully harness these energy resources, accurate predictions of the generation of energy is required. This is the main reason many researchers today have inclined their research towards these demand and price predictions and for the same reason, we decided to delve in this domain as a supplementary topic.

Energy Price Prediction:

Literature Review:

Energy price prediction is a crucial area of interest in the electricity markets. Accurate price prediction is important for an organization to adjust it's own bid price and change production thresholds. In the past, significant research has been made to utilize Machine Learning approaches to predict energy prices for the future. [5] broadly mentions the prospective solutions to this problem using machine learning and artificial intelligence. [4] talks about using feed forward neural networks to capture patterns in historical energy prices to predict future prices. Similarly, [1] uses cascaded neural networks to predict energy MCPs. Both [6] and [2] mention using weighted nearest neighbour techniques along with time series forecasting to predict energy prices.

Most of the existing techniques to predict energy prices use neural networks and time series analysis algorithms, since the energy data follows a time series pattern. Moreover, neural networks can efficiently detect hidden patterns and features from data that can be used for prediction.

Proposed Idea:

Even though neural networks and time series analysis perform well with respect to detecting hidden features and seasonal patterns, they can get slow and inefficient on addition of external variables like socio-physical attributes, weather data, etc. This limits their applicability on a real-time system or environment

To overcome these shortcomings, we propose a solution that uses regressive machine learning algorithms to predict energy prices accurately. We are capturing seasonal and periodic elements by extracting seasonal features from the data(discussed below) along and providing them to the machine learning algorithm. In addition to this, we have also integrated external physical factors like weather elements to make accurate price prediction.

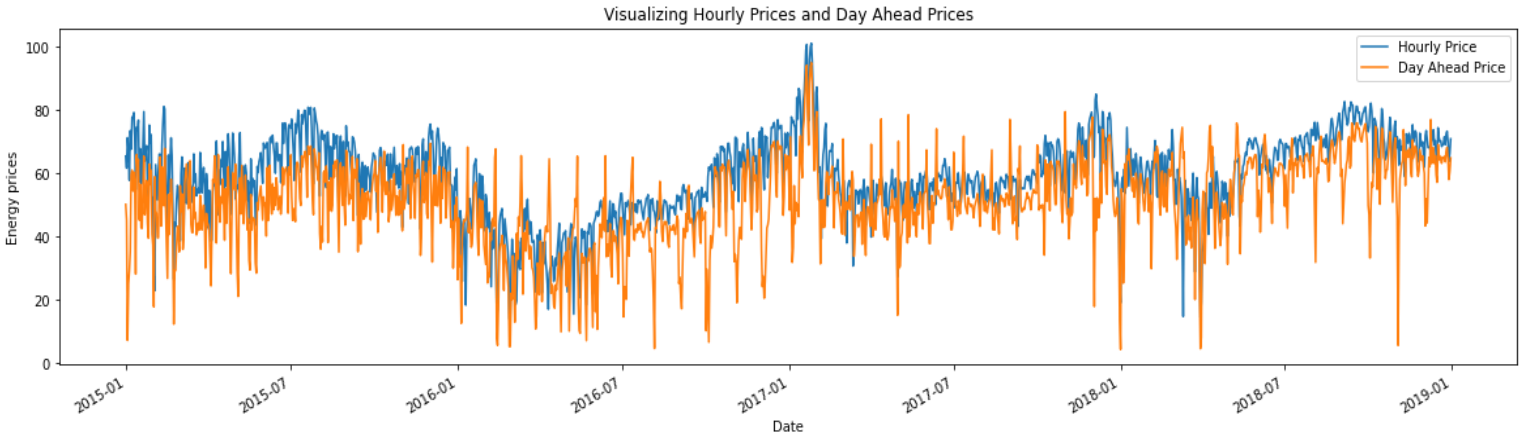
Implementation:

1. *Baseline Model:*

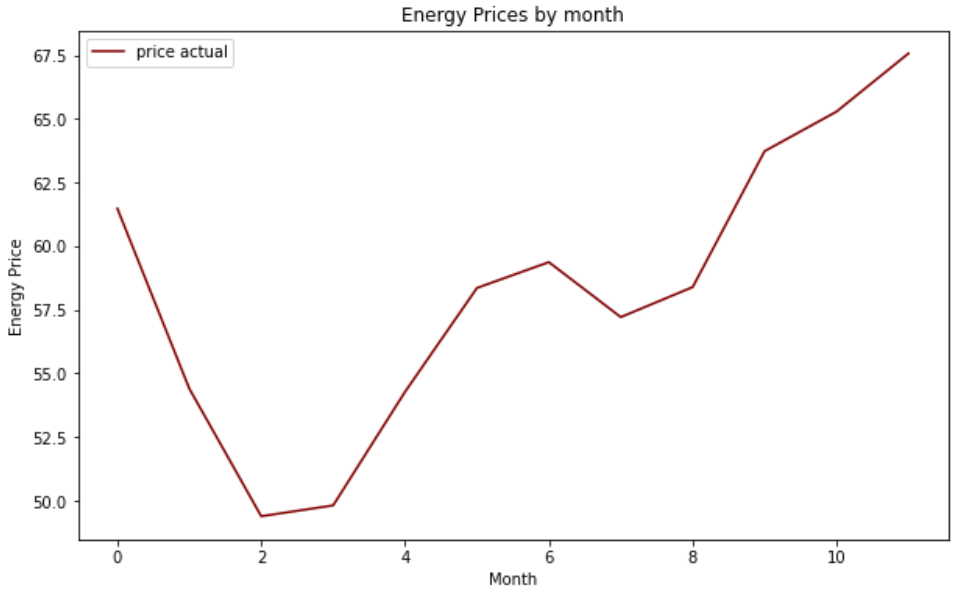
Since our main objective is to develop an approach better than existing neural networks and time series prediction methods, we have set our baseline model as ARIMA that has been used in [2] to give decent results for energy price prediction. As mentioned in the paper, this model gives a mean absolute percentage error (MAPE) of 10% on average for prediction of daily and hourly energy market clearing prices. We ran this model on our dataset to get MAPE of 20.63% and 35.07% respectively. We will be comparing accuracy of all our implemented models with this baseline model. Using our models, we are predicting both hourly energy prices and market clearing prices for the next day. To maintain consistency, we are using the mean absolute percentage error (MAPE) metric to determine how each prediction algorithm performs.

2. *Preprocessing and Data Analysis:*

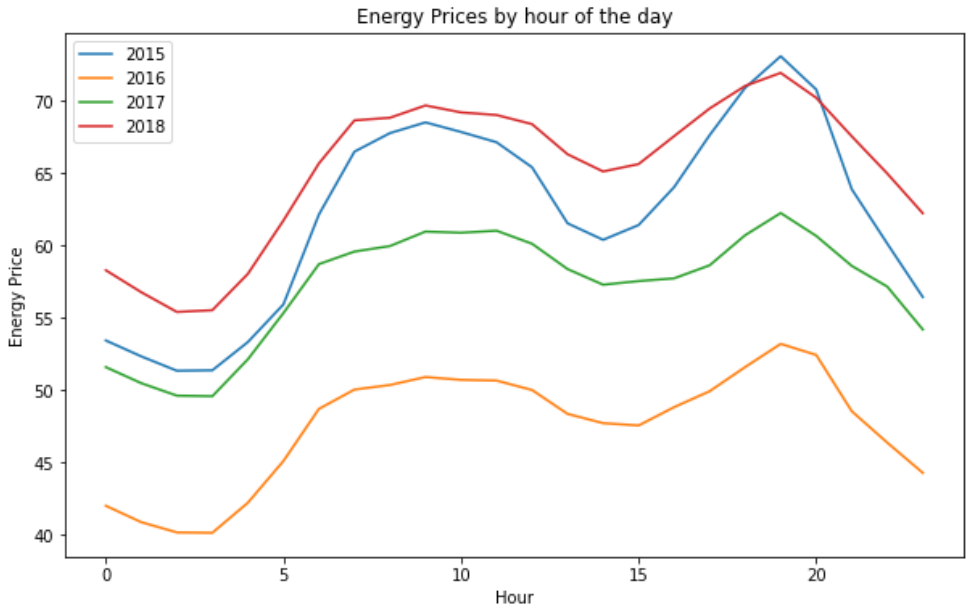
On loading the data, we have first removed useless columns to reduce the data size. We have also handled N.A values in the data. Since the data follows a time series pattern, we cannot remove rows that contain N.A values. Instead, we have replaced such N.A values using the forward fill method. In addition, we have performed encoding on categorical variables.



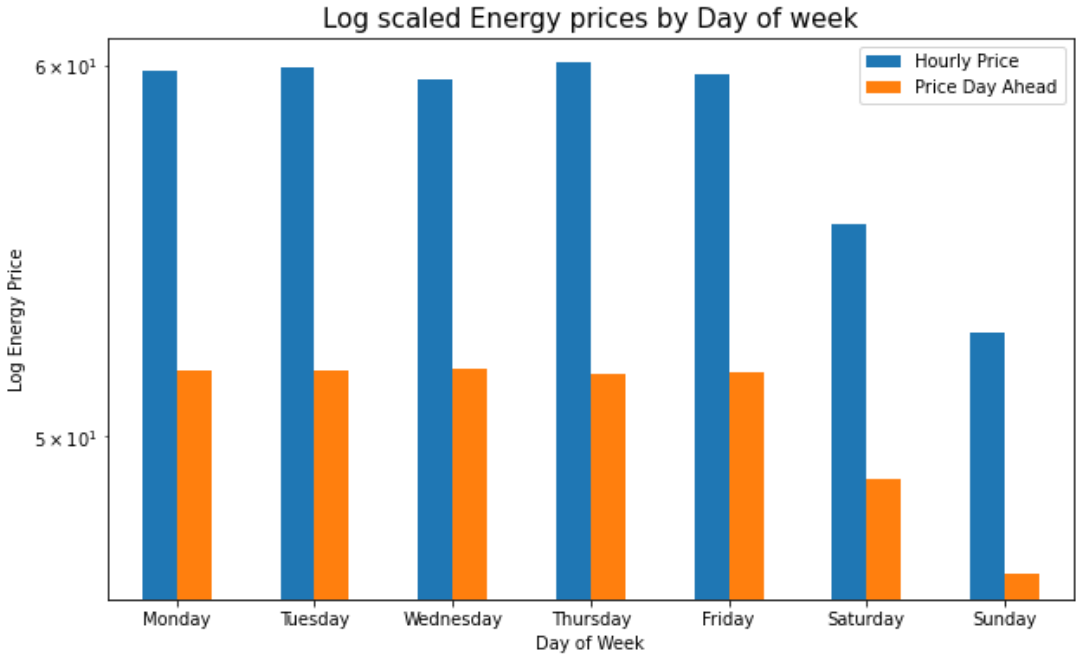
From the plot above, we can see that both hourly energy prices and the day ahead prices (market clearing prices) roughly follow the same time series pattern, with the day ahead prices being more volatile. Thus, we can assume that the seasonal patterns that we observe for hourly energy prices are also relevant to the day ahead prices.



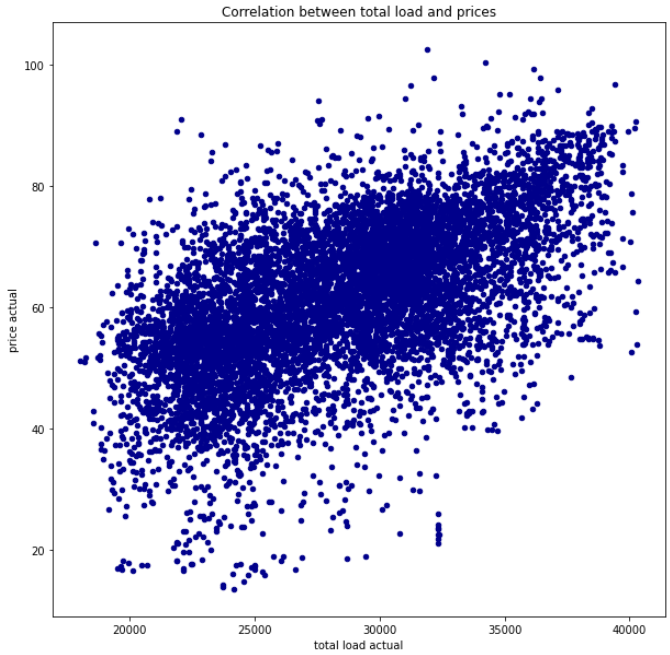
We can see that the energy price fluctuates with respect to different months. For example, the median daily energy price is high during the months of summer(May and June) and the months of winter(November, December and January). We are capturing this monthly seasonality of data by creating a new column in the data called “Month” and using it in our prediction models.



Similarly, we can also observe seasonality in the energy prices with respect to hour of the day for the four years of 2015, 2016, 2017 and 2018. Prices for all the four years follow a similar pattern, where the hourly energy prices are lower early in the morning and increase till just before midnight (max at 7 pm). We are again utilizing this seasonal pattern by creating a new column called “Hour” of the day.



We have also found out another interesting insight, that both the hourly and daily energy prices are much lower on weekends as compared to weekdays. To consider this correlation for price prediction, we create a new column that stores the day of week.



We can also notice a positive correlation between energy prices and total load from the diagram above.

In addition, we have also merged aggregated weather data for the country with energy data to use it in our prediction models. Some weather related features include Temperature, Snow, Rainfall, Humidity, etc. We have found that these features that are otherwise neglected by traditional prediction models are highly correlated with energy prices.

A few variables that are highly correlated (both positively and negatively) are:

Attribute	Pearson's Correlation
Total Load Actual	0.435573
Month	0.732155
Rain	-0.418546
Temperature	0.82370
Generation fossil hard coal	0.465957

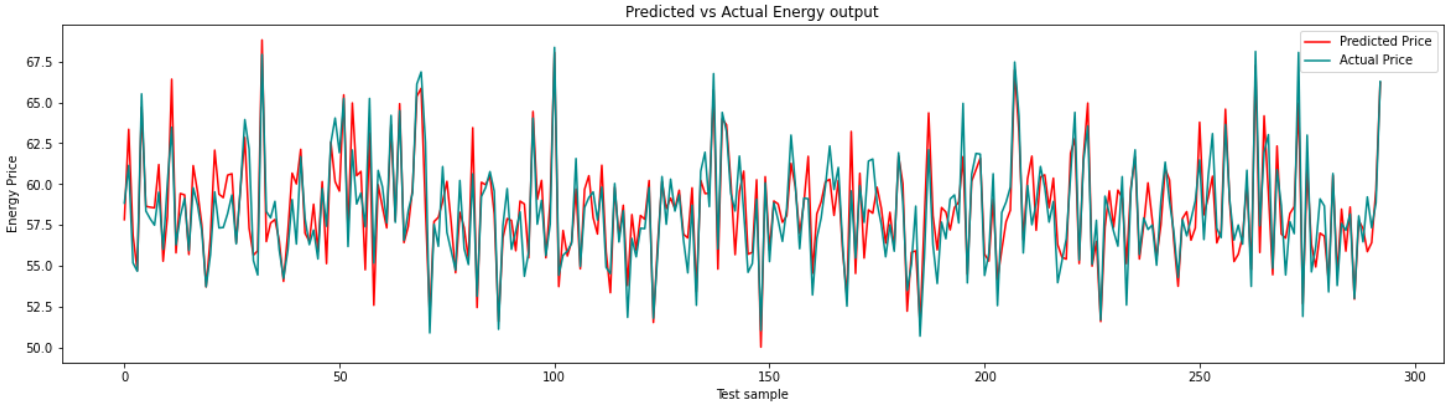
The data is now processed and ready to be used by prediction algorithms.

3. *Machine Learning Models and Results:*

Along with the baseline model (ARIMA) decided from the research paper, we have tried a few prediction algorithms using our own approach and evaluated their results to find the most optimal algorithm.

3.1. Linear Regression:

We are using linear regression over a train test split of 80%-20% to predict both hourly energy prices and next day energy prices respectively based on historical prices and other mentioned attributes.

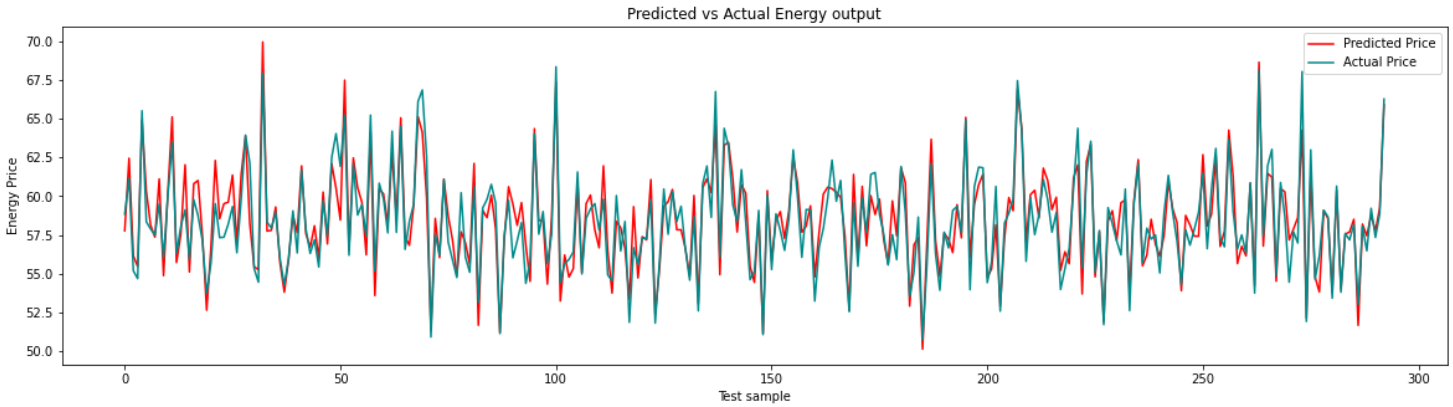


Since day ahead prices are more volatile, we are getting slightly higher MAPE for them as compared to hourly predicted values. We are getting the following values of Mean Absolute Percentage Error (MAPE):

Predictions (Linear Regression)	Mean Absolute Percentage Error (MAPE)
Hourly Energy Prices	4.598
Day Ahead Prices (MCP)	6.299

3.2. Random Forest:

Random forest builds multiple decision trees and merge their predictions together to get a more accurate and stable prediction rather than relying on individual decision trees. In this model, we are running Random forests with max depth of 8 and over a train test split of 80%-20% to predict both hourly energy prices and next day energy prices respectively.

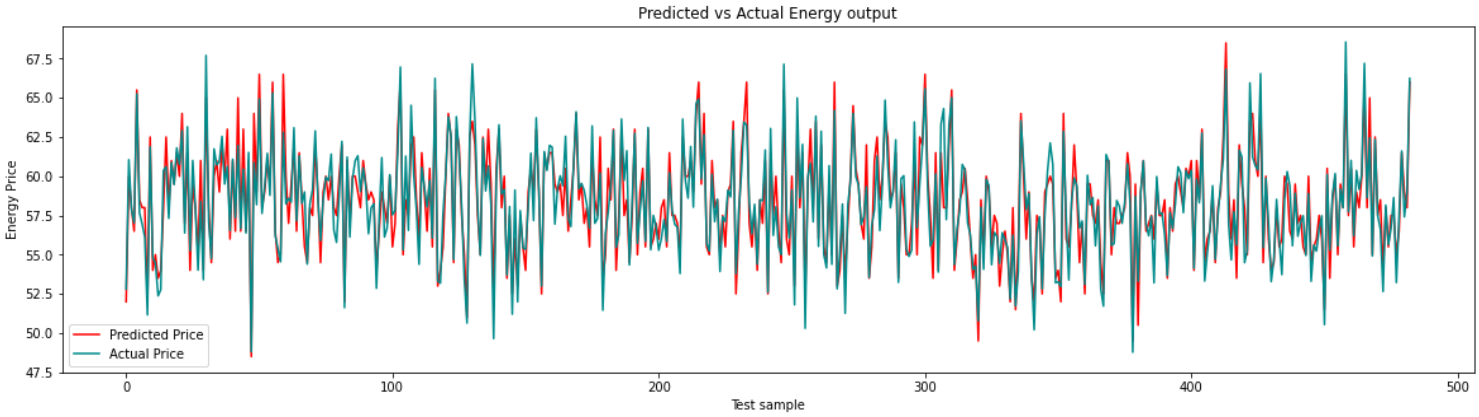


From the graph, we can observe that the predicted and true values are almost coincident except for a few spikes in the true value. We have the following values of Mean Absolute Percentage Error (MAPE):

Predictions (Random Forests)	Mean Absolute Percentage Error (MAPE)
Hourly Energy Prices	3.715
Day Ahead Prices (MCP)	5.306

3.3. XGBoost:

XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements machine learning algorithms under the Gradient Boosting framework. In this model, we are using a train-test split of 80%-20% to predict both hourly energy prices and next day energy prices respectively.

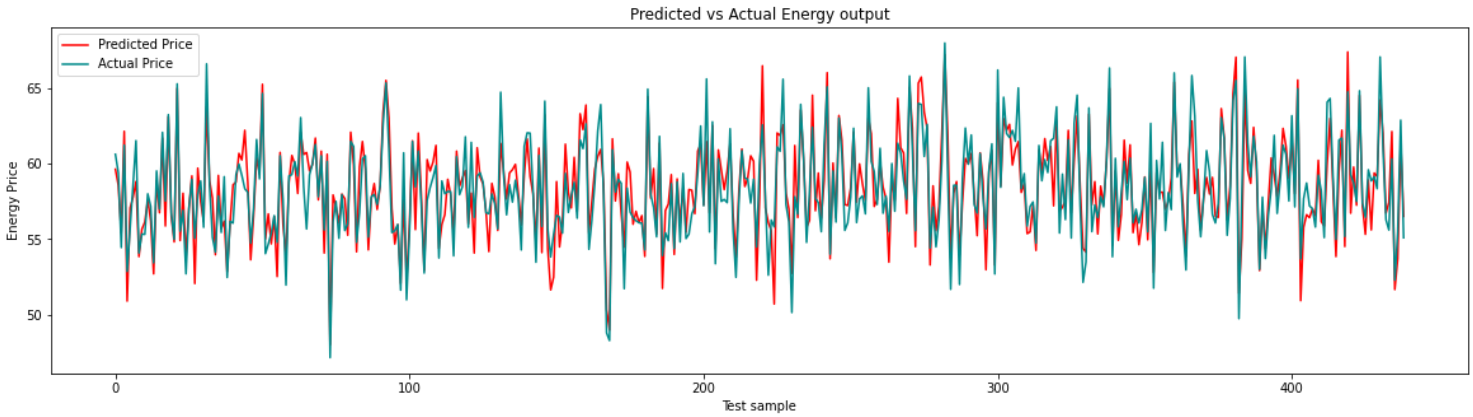


We have the following values of Mean Absolute Percentage Error (MAPE):

Predictions (XGBoost)	Mean Absolute Percentage Error (MAPE)
Hourly Energy Prices	3.463
Day Ahead Prices (MCP)	5.029

3.4. Lasso Regression:

Lasso regression is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces. Even in this model, we are using a train-test split of 80%-20% to predict both hourly energy prices and next day energy prices.



Using Lasso, we get the following values of Mean Absolute Percentage Error (MAPE):

Predictions (Lasso)	Mean Absolute Percentage Error (MAPE)
Hourly Energy Prices	4.569
Day Ahead Prices (MCP)	6.079

4. *Comparison and Results:*

Prediction Algorithm	Mean Absolute Percentage Error	
	Hourly Energy Prices	Day Ahead Prices (MCP)
ARIMA (Baseline Model from paper)	20.63%	35.07%
Linear Regression	4.598%	6.299%
Random Forests	3.715%	5.306%
XGBoost	3.463%	5.029%
Lasso Regression	4.569%	6.079%

We can observe that all the prediction models that we implemented using our approach performed better than the baseline model mentioned in the research paper(ARIMA). Among all the implemented algorithms, XGBoost gave the best results with a mean absolute percentage error of 3.46 and 5.02 for hourly and day ahead energy prices respectively.

We have further discussed the implications of these results in the conclusion at the end of the document.

Solar Energy Prediction:

Literature Review:

Today an increasing attention has been devoted to the energy production from renewable sources, because they represent a valid alternative to the traditional fossil fuel resources, whose future availability is uncertain and whose cost is constantly increasing [8]. Energy forecasting can be used to mitigate some of the challenges that arise from the uncertainty in such renewable resources. One of the main renewable energy sources available in nature is the sun, thus discussion on solar energy prediction is one of the current topics of research interest. This solar energy prediction is dependent on multiple factors such as weather conditions and site-specific conditions and is not easily predictable [7]. As using renewable resources is the future of our planet, finding accurate predictions is important so that we can find a way to use these resources wisely. These predictions can be validated by RMSE and a number of other metrics as mentioned in [9].

Since the solar energy is abundant, proper harnessing of the solar energy is required. Current usage limits its scope of benefits. So self-consumption and local energy communities as

described in [10] play a significant role in the energy transition and the development of solar energies.

Proposed Idea:

Most of the existing approaches take into account the solar irradiance to estimate generation of solar energy for a given time frame. However, this metric is not readily available across all the regions, making it bound to a given region. This might negatively affect the ability of the prediction algorithm to make accurate predictions over a large region.

In our approach, we have used Machine Learning algorithms to calculate the solar generation forecast for every hour based on climatic parameters such as humidity, temperature, pressure, wind speed, etc. These features either directly or indirectly have an effect on the solar forecast. Combining weather features along with energy data helps us provide more accurate results for our predictions. We have then performed recursive feature elimination(RFE) to find the best combination of features, which would thus give us the least error.

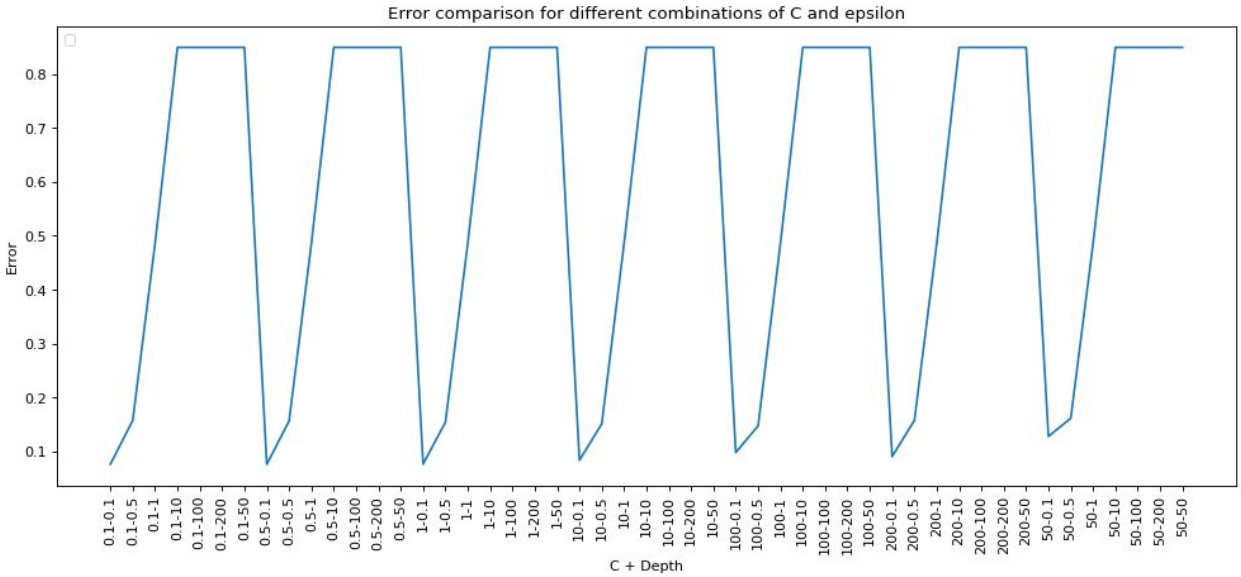
Apart from that, we have also varied each of the micro-parameters of every prediction model to find the perfect model that gives us the highest accuracy. Both regression and time series analysis were used to predict the hourly solar forecast. As weather is unpredictable and changes constantly, performing time series on such data could result in higher errors. To overcome these challenges, we have also performed various regression models. We eliminated outliers in the data which could be the result of extreme weather conditions, as this could affect the predictions in a negative manner. Taking the above measures will thus help us to obtain a higher accuracy.

Implementation:

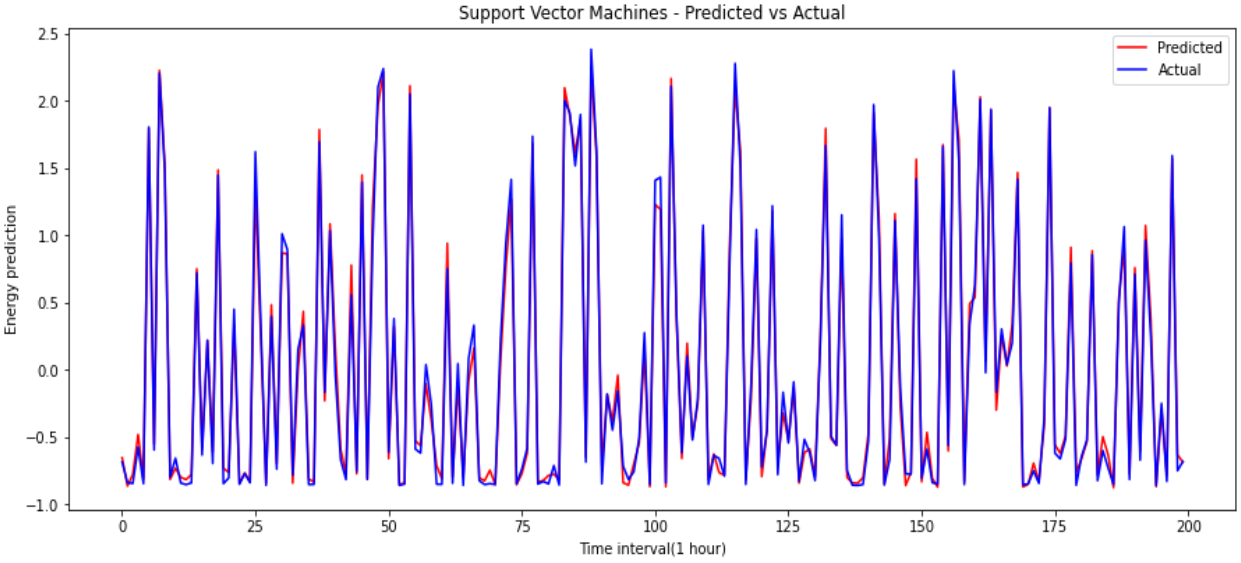
1. Baseline Model:

The baseline model chosen for this prediction is Support Vector Regression Model as described in [8]. The RMSE for this paper's prediction is 0.5275, while ours was 0.309 which means the accuracy for our model was better. Support Regression models are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. This model gives a Root Mean Square value (RMSE) of 0.309. We have then implemented other models using our own approach to predict energy and compared their performance with this baseline model. The data we are using is weather conditions and experimental solar forecast for 3 years with an interval of 1 hour.

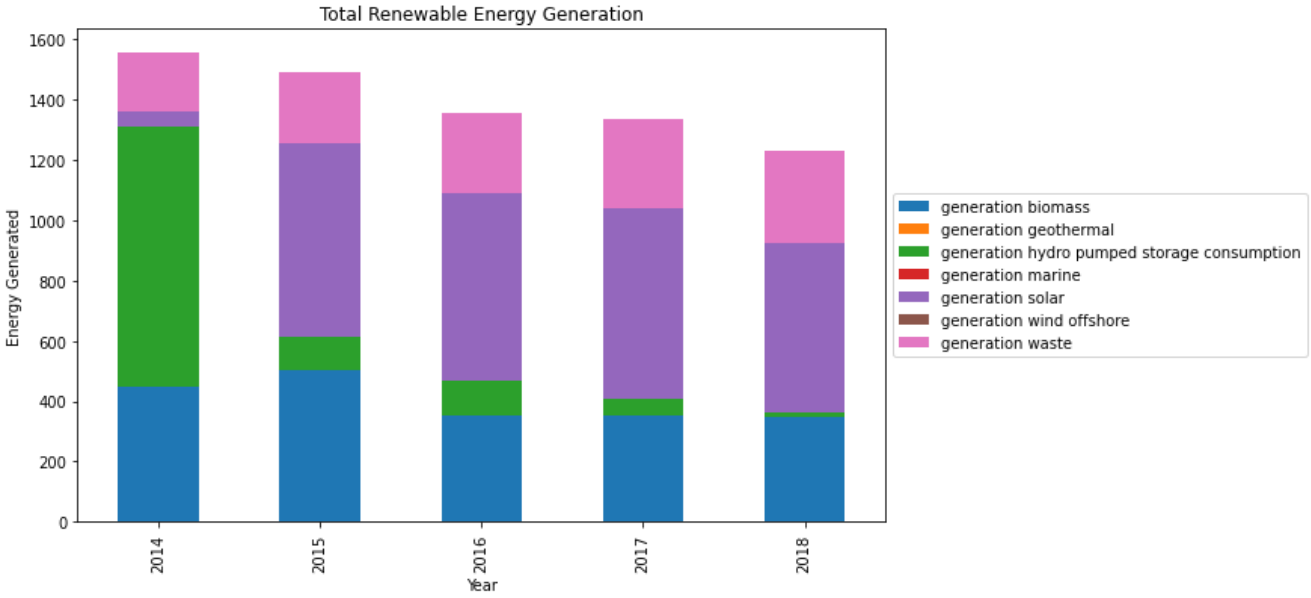
The following is the graph for the error comparison for the different values of C and epsilon in our baseline model, support vector machines.



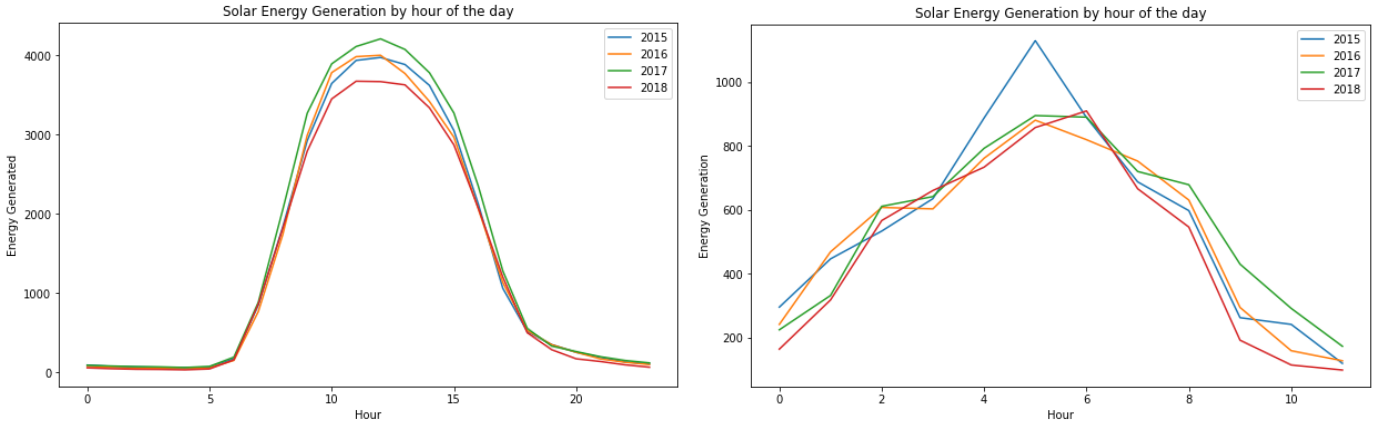
The following is the graph for the predicted values vs. the actual values.



2. Preprocessing and Analysis:



From the plot above, we can see that solar energy has a major contribution among energy generated from renewable sources. Hence, it makes sense to analyse and predict generation of solar energy.

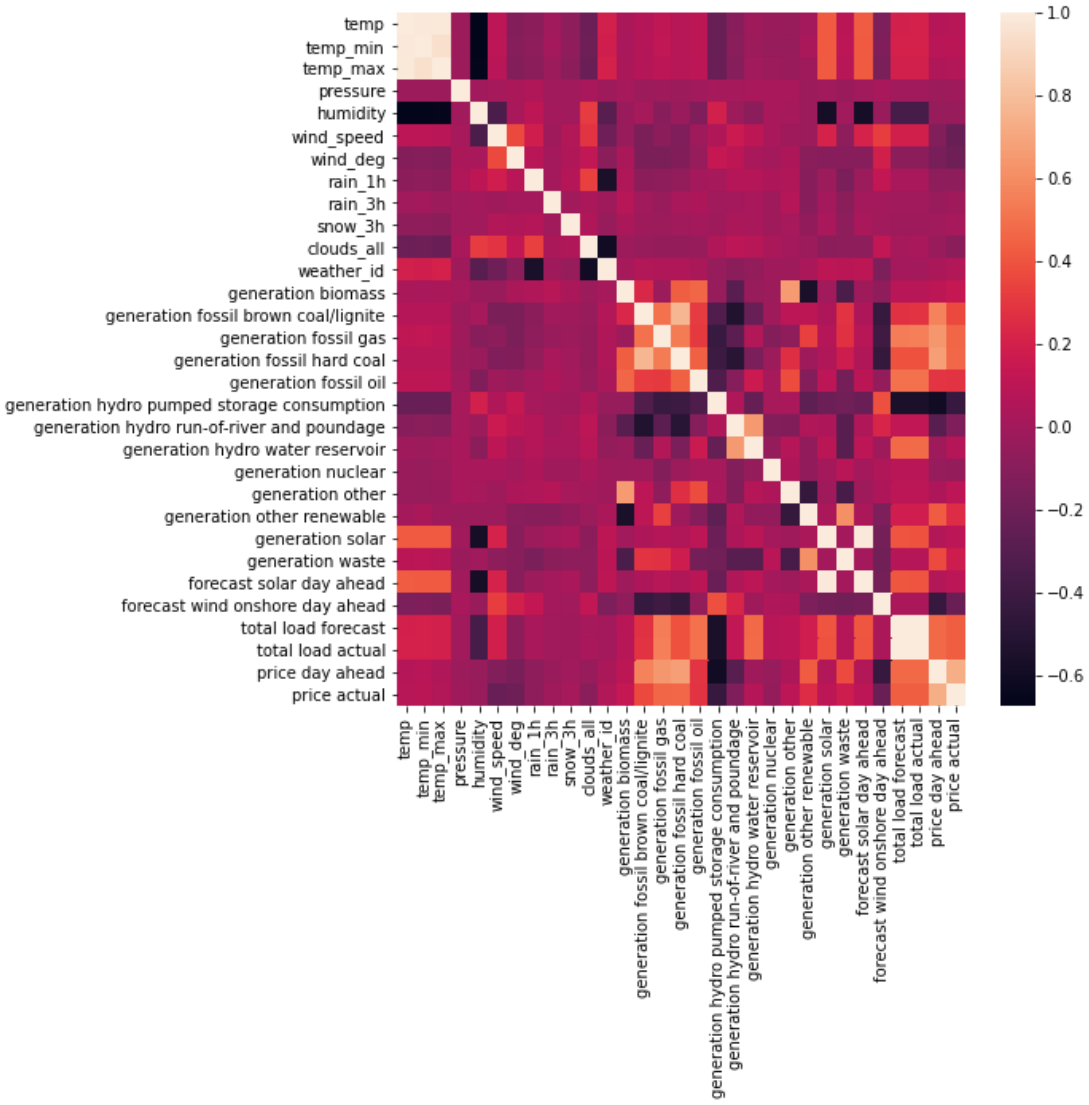


In the graphs above, we have plotted solar energy generation for four years from 2015 to 2018. We can see that solar energy generation follows a normal distribution with respect to hours of the day. We can also observe that the generation is high during months of summer and gradually goes down during the months of winter. We are using this seasonal behaviour to improve accuracy of our prediction models.

In order to correctly see the overall solar forecast results, we have merged these 2 data sets on the common timestamp value. We have analysed all the features present in the dataset by

finding the Pearson correlation between each and every feature and selected the ones that correlate the most with the output label.

According to the heat map of Pearson correlation between features as shown in the below figure, we have selected the 17 best features to calculate the prediction.



After finalizing the features, we split the entire data into a testing set and training set. The training set contains a known output and the model learns on this data in order to be generalized to other data later on. We have the test dataset in order to test our model's prediction on this subset. The ratio for splitting used is 80-20. We have also removed outliers from our data by filtering the data using z-scores. The data above and below three standard deviations from the mean was removed. These outliers could be results of extreme weather conditions and thus, would negatively affect our predictions.

We also tried to separate the data as day and night data to see if that would provide us with better results. This, however, performed poorly and we used the entire dataset to train and test the model.

3. Machine Learning Models and Results:

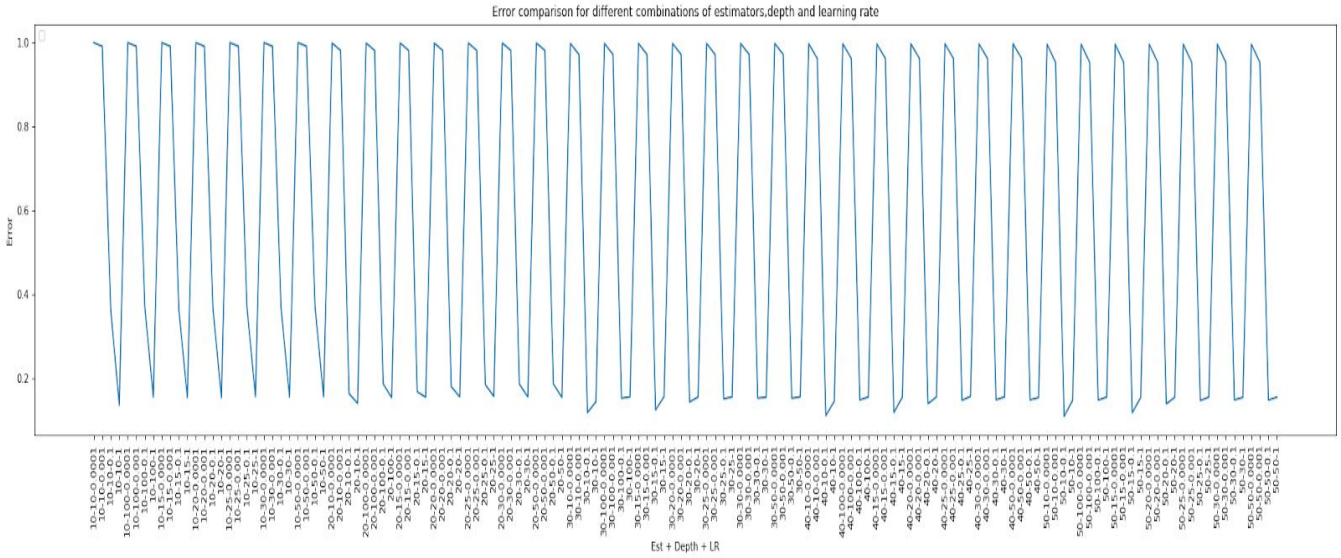
3.1 Gradient Boosting Regression:

Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

Here size of tree or the depth plays an important role. This parameter can be adjusted for a data set at hand. It controls the maximum allowed level of interaction between variables in the model. For example, with J = 2(decision stumps), no interaction between variables is allowed, with J = 3 the model may include effects of the interaction between up to two variables, and so on.

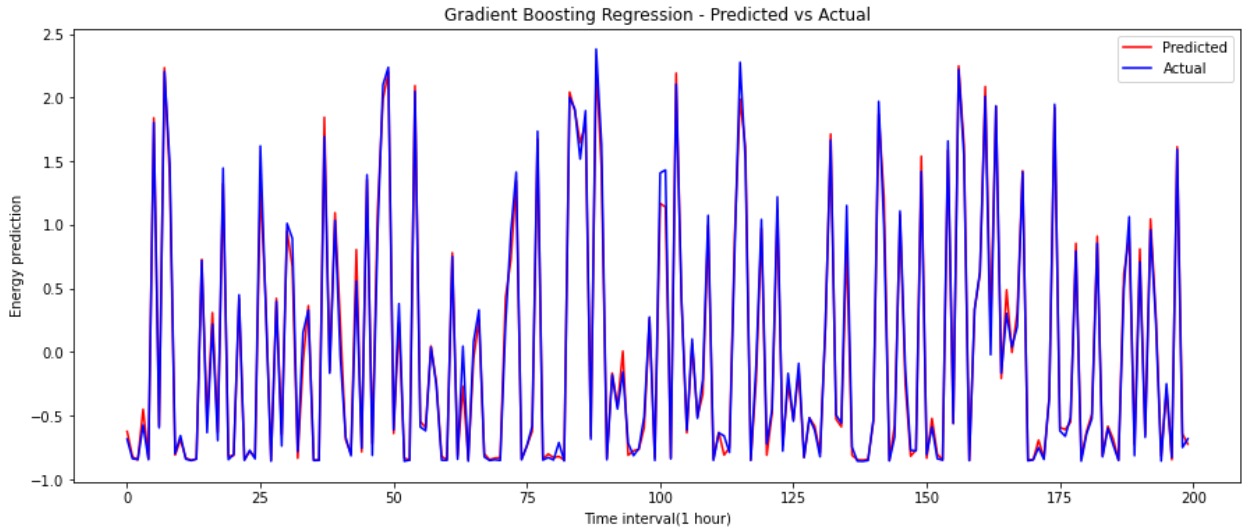
Also, the number of estimators and learning rate plays an important role in the efficiency of this algorithm.

So, we have tried different values of estimator, depth(or J) and learning rate and checked the best combination of both these values to produce the least error. The figure below represents error values for all these combinations.



After implementing the model for different estimators depth and model, the best combination we got is the error value of 0.1106 for estimators = 50, depth = 10 and learning rate = 0.1. This error value is a great improvement over the original baseline model's RMSE value.

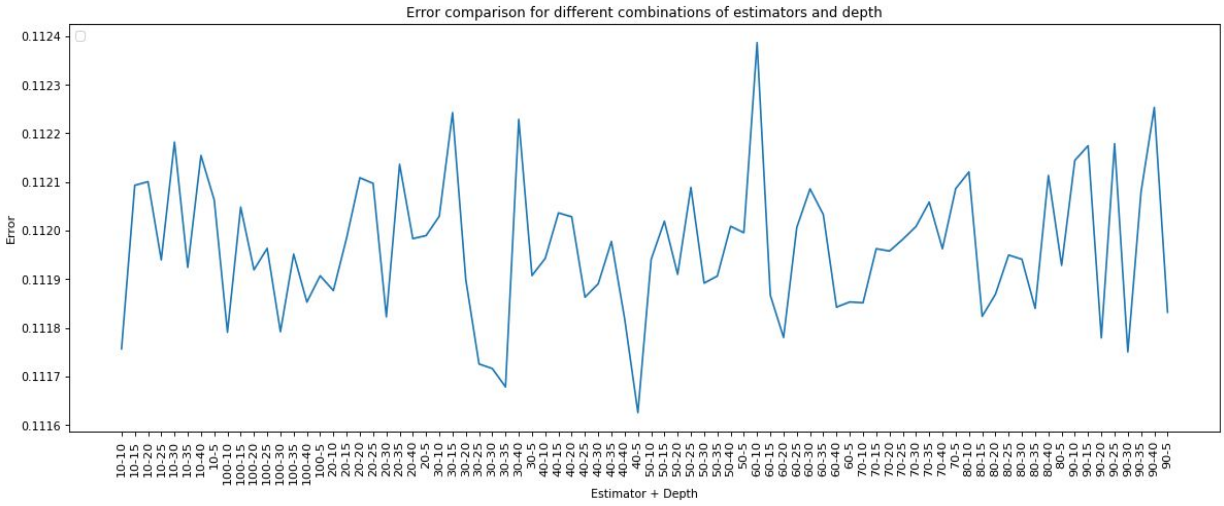
The following is the graph for the predicted values vs. the actual values.



3.2 Random Forest Regression:

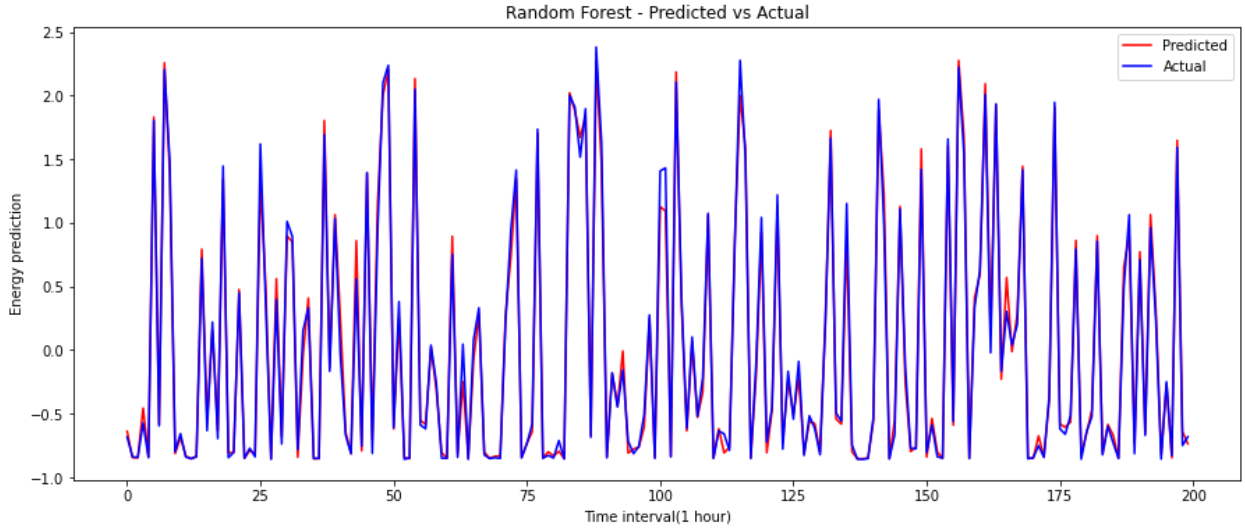
Random forest builds multiple decision trees and merge their predictions together to get a more accurate and stable prediction rather than relying on individual decision trees. We tried random forests with different estimators and depth in order to figure out the best combination minimizing the error to the greatest extent.

Implementing this model, we obtained different error values for multiple values of estimator and depth as shown in the figure below.



We got the highest optimization with an RMSE value of 0.112 for 80 estimators and a depth of 35. This is better than the baseline SVR model.

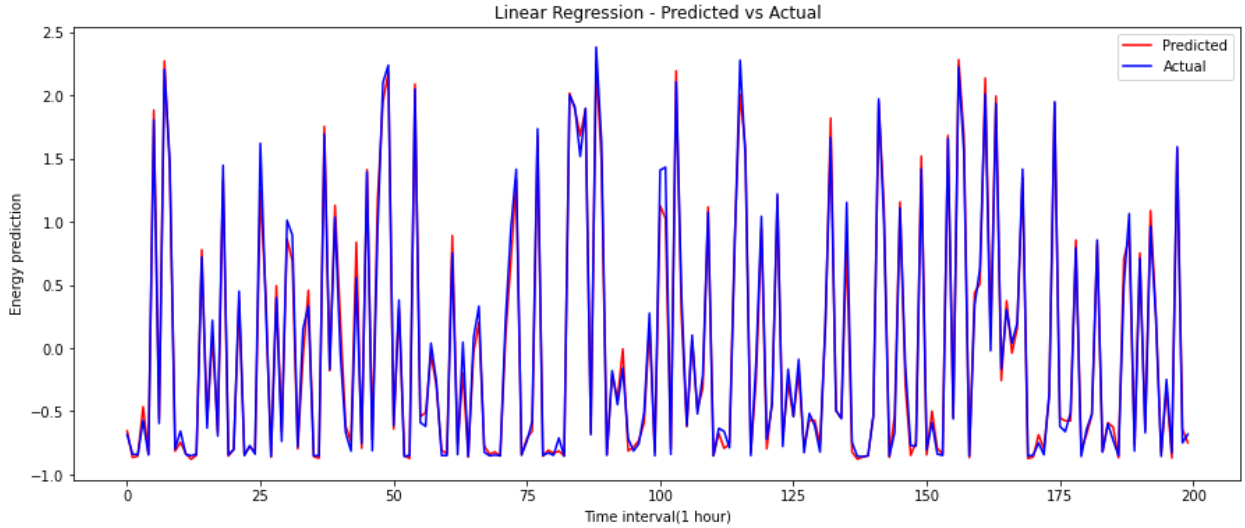
The following is the graph for the predicted values vs. the actual values.



3.3 Linear Regression:

In statistics, linear regression is a linear approach to modeling the relationship between a scalar response(or dependent variable) and one or more explanatory variables(or independent variables). Using this model, we obtained an RMSE value of 0.114.

The following is the graph for the predicted values vs. the actual values.



3.4 SARIMA:

Seasonal Autoregressive Integrated Moving Average, SARIMA is an extension of ARIMA that explicitly supports univariate time series data with a seasonal component. It adds three new hyperparameters to specify the autoregression (AR), differencing (I) and moving average (MA) for the seasonal component of the series, as well as an additional parameter for the period of the seasonality.

The RMSE value for this model is 0.116 which is again better than the baseline SVR model.

4. *Comparison and Results:*

Prediction Algorithm	Root Mean Squared Error
Support Vector Machines(Baseline Model from paper)	0.309590
Linear Regression	0.114997
Random Forest Regression	0.112024
Gradient Boosting Regression	0.110674
SARIMA	0.116086

From the above results, we can see that our baseline model which was Support Vector Machines, performed the worst with an RMSE value of 0.309590. The model which gave us the best results was Gradient Boosting Regression with an RMSE value of 0.110674.

Conclusion and Key Contributions:

In this project, we have touched upon the two key areas in the energy domain- price prediction and energy prediction.

As mentioned above, most of the existing techniques rely heavily on historical price data and load estimates for energy price prediction. Most of these techniques make use of time series analysis or neural networks. With our new approach, we are utilizing external factors like weather, seasonal patterns, daily and monthly trends, etc. to improve accuracy of energy price prediction using regression techniques. We have achieved a very significant improvement in accuracy of predictions as compared to the baseline model that we set from an existing research paper(3.5% vs 10% Mean absolute percentage error respectively).

Moreover, existing methods to predict solar energy generation rely heavily on solar irradiation and do not consider other external factors. With our approach, we predict solar energy

generation using weather based features along with seasonal patterns. This approach has given us a significant improvement in prediction accuracy as compared to the baseline model discussed in the research paper(0.11% vs 0.30% Mean absolute percentage error respectively).

Improvements and Future Work:

We can further improve the accuracy of our predictions by integrating data on more socio-economic features like festivals(during which demand of energy is high), national sport events, etc.

Also, these predictions will be more useful if the market structure is improved. In short, the entire energy market needs to be shifted towards a more decentralized approach. This decentralized approach can be implemented by blockchain technology using smart meters to calculate energy measurements and smart contracts used as a proof of traded energy as well as payment.

Such a market structure will remove the need for intermediaries and eventually reduce the energy price volatility, thus resulting in an efficient market. This will make our predictions even more significant.

Team Members and Contribution:

The team members are:
1. Pratik Mukund Velhal (112675099)
2. Palak Jain (112675008)
3. Navpreet Kaler (112689117)

Each of us contributed equally to the project, and worked on all the parts of this project together right from literature review to actual implementation.

Supplementary materials:

We have maintained two separate notebooks for energy price prediction and solar energy prediction respectively for ease of access.

Our dataset includes two files:

1. energy_dataset.csv
2. weather_features.csv

References:

[1]. Energy Clearing Price Prediction and Confidence Interval Estimation With Cascaded Neural Networks.

[2]. Day-Ahead Price Forecasting of Electricity Markets by a New Fuzzy Neural Network.

[3]. Optimal Residential Load Control With Price Prediction in Real-Time Electricity Pricing Environments

[4]. Machine Learning Applications for Load, Price and Wind Power Prediction in Power Systems

[5]. Energy price forecasting - problems and proposals for such predictions

[6]. Electricity Market Price Forecasting Based on Weighted Nearest Neighbors Techniques

[7]. Predicting Solar Generation from Weather Forecasts Using Machine Learning

[8]. Photovoltaic energy production forecast using support vector regression

[9]. Proposed Metric for Evaluation of Solar Forecasting Models

[10]. Distributed solar self-consumption and blockchain

[11]. Forecasting of total daily solar energy generation using ARIMA: A case study