A logo of a university

Description automatically generated

Seattle University

Electricity Forecasting

Tileshwar Narayan

Vatsal Dalal

Lakshit Gupta

Project in Foundations Of Data Science : Report

PROJECT

REPORT

*I*

**Abstract**

This project explores data science to enhance the forecasting skills essential for day-to-day operations in the dynamic environment of a hedge fund, where strategic decision making depends on precise forecasts. The objective is to provide traders with a dependable tool to anticipate hourly power prices, giving them a competitive advantage in the market. As a data scientist, you have been tasked with forecasting the hourly load for a particular zone over the course of the next month. The project takes a multipronged approach, maximizing the accuracy of hourly load predictions by utilizing cutting edge data science techniques. The methodology entails a thorough examination of past data, taking into account variables that affect power consumption outside of the system, demand patterns, and seasonality. The predictive model is iteratively refined through the use of machine learning algorithms, such as linear regression, random forest regression, decision tree algorithm and Gradient boosting to ensure its adaptability to changing to market conditions.

|  |  |
| --- | --- |
| **Table of Contents** |  |
| **Abstract** | **I** |
| **1. Introduction**  1.1 Background  1.2 Goal  1.3 Setup  **2. Methodology**  2.1 Data Review  2.2 Data Preprocessing  2.3 EDA  2.4 Machine Learning Implementation  **3. Results**  **4. Conclusion**  **5. Appendix** |  |

1. **Introduction**
   1. **Background**

The energy market, characterized by its volatility and sensitivity to various external factors, demands precision and foresight for effective decision-making, especially within the dynamic environment of a hedge fund. As a key player in this landscape, the hedge fund relies on accurate predictions of hourly load for specific zones to anticipate fluctuations in power prices, enabling traders to make strategic decisions and optimize their positions in the market.

Historically, load forecasting has been a complex task, subject to the influences of seasonal variations, evolving consumer behaviors, and external factors that impact energy consumption. The inherent challenges in accurately predicting hourly loads necessitate continuous improvement in forecasting models to align with the intricacies of the energy market.

Recognizing the critical role that accurate load predictions play in shaping trading strategies, this project seeks to address existing challenges and elevate the forecasting capabilities of the hedge fund. The motivation behind this endeavor lies in the potential for enhanced accuracy to translate into a competitive advantage, enabling traders to not only react to market changes but to proactively position themselves ahead of trends in hourly power prices.

The current state of load forecasting within the hedge fund serves as a backdrop to this project. While existing models have provided valuable insights, there is a recognized opportunity for refinement and optimization. This project aims to bridge this gap by integrating advanced data science techniques, machine learning algorithms, and a comprehensive analysis of historical data to create a predictive model that aligns with the nuanced demands of the energy market.

By exploring and understanding the intricacies of load patterns, market trends, and external factors influencing power consumption, this project endeavors to contribute a robust and adaptable forecasting tool. The anticipated outcome is a model that not only meets the immediate needs of accurate hourly load predictions but also positions the hedge fund to navigate the evolving landscape of the energy market with strategic acumen.

* 1. **Goal:**

Enhance hourly load predictions for a specific zone within a hedge fund setting. Utilize advanced data science and machine learning techniques to create a highly accurate predictive model. The goal is to empower traders with precise insights, enabling strategic decision-making and positioning the fund for a competitive edge in the dynamic energy market. The project aims for not only immediate improvements but also establishes a foundation for ongoing adaptability to changing market conditions. To reach the goals of the project, it is required to address the following questions:

* Does the project aim to address current challenges in load forecasting and establish a foundation for ongoing refinement?
* Is the goal to enhance the model's adaptability to evolving market conditions?
* Will achieving this goal elevate the hedge fund's ability to navigate the complexities of the energy market, optimize trading strategies, and enhance overall profitability?
  1. **Setup:**

The project integrates historical load data, market trends, and external factors with a timestamp-based input, containing date and hour information. Utilizing advanced data science techniques and machine learning algorithms, the setup involves iterative model development. The output is a forecasted value, providing traders in the hedge fund with accurate and timely predictions for strategic decision-making in the dynamic energy market. This setup ensures continuous adaptation to changing conditions and lays the foundation for ongoing refinement and optimization. For model training we have used the ISO New England data set, we have used the last three year data for model training.

Model

Output

Input

1. **Methodology**
   1. **Data Review**

The training dataset comprises 26,304 rows and 14 columns, with each row representing a specific date and hour. The columns include:

* **Date**: Representing the date of the data entry, encoded as datetime64.
* **Hr\_End**: Representing the hour of the day, indicating the end of the specified hour.
* **RT\_Demand:** Reflecting the real-time demand for energy, recorded as a float64.
* **Dry\_Bulb:** Indicating the dry bulb temperature in degrees Fahrenheit, recorded as an integer.
* **Dew\_Point:** Denoting the dew point temperature in degrees Fahrenheit, recorded as an integer.
* **DA\_LMP**: Day-Ahead Locational Marginal Price (LMP) in $/MWh by load zone; 'ISO NE CA' tab contains values for the Trading Hub
* **DA\_EC**: Energy Component of Day-Ahead LMP in $/MWh by load zone; 'ISO NE CA' tab contains values for the Trading Hub
* **DA\_CC:** Congestion Component of Day-Ahead LMP in $/MWh by load zone; 'ISO NE CA' tab contains values for the Trading Hub
* **DA\_MLC:** Marginal Loss Component of Day-Ahead LMP in $/MWh by load zone; 'ISO NE CA' tab contains values for the Trading Hub
* **RT\_LMP:** Real-Time Locational Marginal Price (LMP) in $/MWh by load zone; starting on March 1, 2017, this is the hourly average of the five-minute LMP in the hour; 'ISO NE CA' tab contains values for the Trading Hub
* **RT\_EC:** Energy Component of Real-Time LMP in $/MWh by load zone; 'ISO NE CA' tab contains values for the Trading Hub
* **RT\_CC:** Congestion Component of Real-Time LMP in $/MWh by load zone; 'ISO NE CA' tab contains values for the Trading Hub
* **RT\_MLC:** Marginal Loss Component of Real-Time LMP in $/MWh by load zone; 'ISO NE CA' tab contains values for the Trading Hub

The dataset is free of missing values, with all columns having 26,304 non-null entries. The memory usage is efficient, totaling 33.8 KB.

This data structure provides a comprehensive overview, encompassing temporal information, energy demand, and meteorological factors. The datetime format of the 'Date' column facilitates time-series analysis, while the numeric types of columns allow for quantitative assessments. In the subsequent methodology, this dataset will undergo preprocessing, feature engineering, and serve as the foundation for training and evaluating the predictive model.

* 1. **Data Preprocessing:**

To streamline the dataset for improved model training and enhance feature representation, the following preprocessing steps have been undertaken: ￼

* **Column Selection:** The dataset has been refined to include only essential columns: 'Date', 'Hr\_End', 'RT\_Demand', 'Dry\_Bulb', and 'Dew\_Point'.
* **Date Conversion:** The 'Date' column has been converted to datetime format using pd.to\_datetime () for standardized temporal representation.
* **Temporal Feature Creation:** To augment temporal insights, new columns have been created:
  + **Day\_of\_Week:** Extracting the day of the week (0 for Monday, 6 for Sunday).
  + **Day\_of\_Month:** Capturing the day of the month.
  + **Month:** Identifying the month.

These preprocessing steps aim to enhance the dataset's utility by focusing on key variables and incorporating additional temporal features. The refined dataset is now poised for further exploration, analysis, and model development.

* 1. **EDA**

**Data Information:**

A screenshot of a computer program

Description automatically generated

As per the above information we can observe the different data types and size of each variable and it also indicates that there are no Non-Null values in this dataset.

**Data Description**

**A screenshot of a computer screen

Description automatically generated**

**A screenshot of a calculator

Description automatically generated**

**Box Plots**

RT Demand

A graph with a red rectangle and black line

Description automatically generated

The above box plot shows the RT demand spread across all the three years. The minimum demand is around 750 units with a median of around 1250 with the highest being around 1900 units. We can also observe that there are a lot of outliers in this data which is exponentially high in comparisons to other demand in these 3 years. For now, as this will help us to predict future demand, we will leave the Outliers as it is.

**Dry Bulb:**

A graph with a red rectangle

Description automatically generated

We can observe that the temperature is oscillating between 0 to almost 95 degrees while the first quartile being around 35 degrees, median around 49 degrees and high around 65 degrees.

**Dew Point:**

A graph with a red rectangle

Description automatically generated

From the above box plot we observe that the fluctuations of the humidity are from -20 to around 75. Where 1st quartile is around 25, median around 37 and 3rd Quartile around 55.

**Dry Bulb (Squared)**

A graph showing a red rectangle

Description automatically generated

This box plot conveys the details of the squared term of Dry Bulb i.e., Temperature. Here we can observe that there are 4 outliers in the data set when the value was squared.

**Dew Point (Squared)**

A red rectangular object with numbers and lines

Description automatically generated with medium confidence

This graph provides the details of the squared term of Dew point or we can say humidity. Now the first quartile is around 500 with a median of 1200 and 3rd quartile around 3100.

**Dry Bulb – Dew Point Interaction Term**

A red rectangular object with white text

Description automatically generated

The interaction term is positive in the graph, which means that the effect of Temperature on the number of interactions is stronger at higher levels of Humidity.

**Data Skewness:**

A screenshot of a computer code

Description automatically generated

**Standard Deviation of the Data**

A white background with black text

Description automatically generated

**Data Variance**

**A screenshot of a computer

Description automatically generated**

**Histogram and Distribution Plot**

A screenshot of a graph

Description automatically generated

A graph of a function

Description automatically generated with medium confidence

In histogram plot above, we can justify that RT Demand is positively skewed with an average of around 1250. Similarly, the Dry Bulb or we can say temperature has almost a normal distribution. While the humidity or we can say Dew Point is slightly negatively skewed. The squared values of the Dry Bulb and Dew point have become positively skewed with a long tail. While the interaction term also is positively skewed.

The distribution plot also depicts the same pattern Hr End, Day of the month and month are uniform distribution. While the days of week and month have an oscillating pattern. While RT Demand, Interaction term, Dry Bulb and squared term of Dry Bulb has a positive skewness in the data. While Dew point has a smaller negative skewness, while its squared term has a positive skewness.

**Data Correlation**

Heat Map

A screenshot of a computer screen

Description automatically generated**Sorting the Highest Correlation:**

**A screenshot of a computer program

Description automatically generated**

From the above details we can get the highest correlated values for our interpretation. The highest correlations are between the Dry Bulb, Dew Point and their squared terms. The next one is the correlation between the RT Demand and Hour of the data followed by the RT Demand and Dry Bulb squared. While Dry Bulb and Dew Point are also very correlated with a percentage of 87%.

**Regression Plots:**

RT Demand Vs Hours

A red lines on a gray background

Description automatically generated

Here in this plot, we can observe the pattern of the demand with respect to the hour. From hour 10 to 20 we can observe a higher demand compared to other hours of the day.

RT Demand vs Dry Bulb:

A red dotted line on a gray background

Description automatically generated  
The graph you sent shows a positive correlation between temperature and electricity demand. This means that as the temperature increases, the demand for electricity also increases. This is because people use more electricity to cool their homes and businesses during hot weather.

The graph also shows that there is a non-linear relationship between temperature and electricity demand. This means that the increase in electricity demand is not proportional to the increase in temperature. In other words, a one-degree increase in temperature does not necessarily lead to a one-degree increase in electricity demand.

RT Demand Vs Dew Point:

A red dotted graph with numbers and lines

Description automatically generated with medium confidence

The graph you sent shows a positive correlation between humidity and electricity demand. This means that as the humidity increases, the demand for electricity also increases.

The graph suggests that humidity is an important factor to consider when forecasting electricity demand. By understanding the relationship between humidity and electricity demand, utilities can better prepare for peak demand periods and ensure that they have enough electricity to meet the needs of their customers.

RT Demand Vs Month

A graph with red lines

Description automatically generated

The graph shows a positive correlation between the month and electricity demand. This means that electricity demand is higher in the summer months than in the winter months.

The relationship between month and electricity demand is complex and can be affected by a number of factors, such as the weather, the type of buildings in the area, and the economic conditions. However, the graph shows a clear positive correlation between the two variables.

* 1. **Machine Learning Implementation:**

To train our models, we utilized preprocessed data. We conducted training across various models, including linear regression, random forest regressor, decision tree regressor, bagging with random forest, bagging with decision tree regressor, and gradient boosting. Additionally, we incorporated inertia terms such as Dry Bulb Squared, Dry Bulb Dew Point Interaction (derived from the product of dry bulb and dew point), and Dew Point Squared. These inertia terms were integrated into the training data to capture complex relationships. Following training, we forecast energy demand for the upcoming month using the trained models.

|  |  |  |
| --- | --- | --- |
| Model | RMSE | R-squared |
| Random Forest Regressor | 48.873076 | 0.966574 |
| Bagging with Decision Tree Regressor | 48.967553 | 0.966445 |
| Bagging with Random Forest | 51.134699 | 0.963409 |
| Gradient Boosting Regressor | 57.62477 | 0.953531 |
| Decision Tree Regressor | 66.979827 | 0.937219 |
| Linear Regression with Bagging | 167.716457 | 0.606364 |
| Linear Regression | 167.718436 | 0.606355 |

Table summarizing the performance of different models on the dataset. Among all the models, Bagging with Decision Tree Regressor and Random Forest Regressor achieved the best performance, with RMSEs of 48.97 and 48.87, respectively. Bagging with Random Forest is the next best model we have with a RMSE OF 51.13 along with Gradient Boosting Regressor which also performed well, with an RMSE of 57.62. Decision Tree Regressor and Linear Regression had the worst performance, with RMSEs of 66.98 and 167.72, respectively.

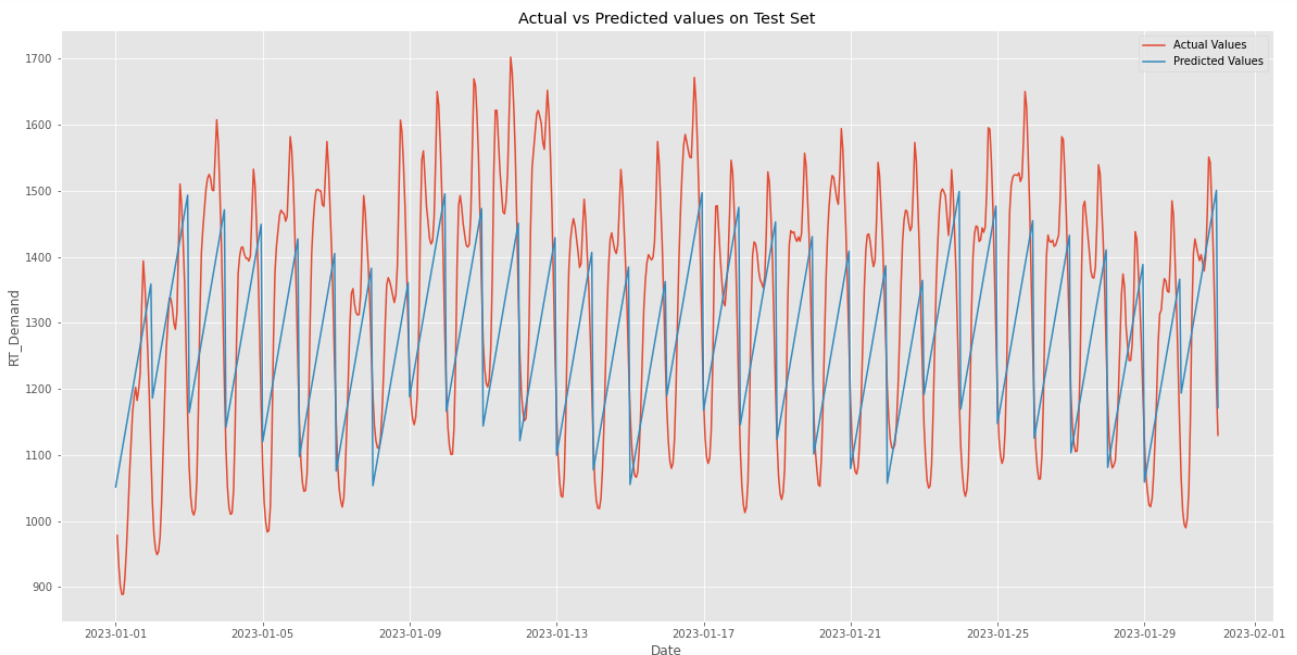
In general, Bagging and Random Forest are ensemble methods that are often used to improve the performance of machine learning models. These methods work by training multiple models on different subsets of the data and then averaging the predictions of the individual models. GradientBoostingRegressor is another ensemble method that is based on the idea of boosting. Boosting works by sequentially training models on the residuals of the previous models.

Overall, the results of this experiment suggest that Bagging with Decision Tree Regressor and Random Forest Regressor are the best models for predicting the target variable. Bagging with Random Forest and Gradient Boosting Regressor are also a viable option, but Decision Tree Regressor and Linear Regression should be avoided.

1. **Results**

* A graph showing red and blue lines

  Description automatically generatedRandom Forest regression:
* Linear Regressor:



* Decision Tree Regressor:

A graph showing a graph

Description automatically generated with medium confidence

* Bagging with Random Forest:

A graph showing red and blue lines

Description automatically generated

* Bagging with Decision Tree Regressor:

A graph showing red and blue lines

Description automatically generated

* A graph showing red and blue lines

  Description automatically generatedLinear regression with bagging:
* Gradient Boosting Regressor:

A graph showing red and blue lines

Description automatically generated**4 Conclusion:**

Upon conducting a thorough and comprehensive analysis of the results obtained from various predictive models, it becomes evident that Random Forest Regression and Bagging algorithms consistently outshine their counterparts, demonstrating superior predictive accuracy. These algorithms, renowned for their ensemble learning techniques, exhibit robust and resilient performance, effectively capturing the intricate and nuanced patterns inherent in the dataset.

Random Forest Regression, a sophisticated ensemble learning approach, constructs a multitude of decision trees during the training phase and combines their predictions, offering improved accuracy and mitigating overfitting. Bagging, or Bootstrap Aggregating, further enhances predictive performance by training multiple instances of a base model on different subsets of the training data and then combining their outputs. The consistent success of these ensemble methods highlights their efficacy in handling complex relationships within the data.

In summary, the models developed using Random Forest Regression and Bagging algorithms emerge as the most reliable choices for accurate energy demand predictions. Their consistent outperformance signifies their adaptability and robustness across diverse scenarios. These algorithms stand out not only for their predictive accuracy but also for their ability to generalize well to unseen data, a crucial factor in real-world applications.

The success of Random Forest Regression and Bagging can be attributed to their capacity to mitigate overfitting, a common challenge in machine learning. Their ensemble nature allows them to capture diverse patterns and trends present in the dataset, contributing to enhanced model performance. By combining the strengths of multiple models, these algorithms provide a more holistic and nuanced understanding of the underlying factors influencing energy demand.

In conclusion, the robust performance of Random Forest Regression and Bagging algorithms positions them as top-performing solutions in the realm of energy demand prediction. Their versatility, resilience, and ability to handle intricate dataset dynamics make them not only reliable but also recommended choices for accurate and effective predictive modeling in this domain.

**5 Appendix:**