ENERGY CONSUMPTION PREDICTION AND ANOMALY DETECTION

Submitted by

AYSHA SANA	**223020**
BABY SHERIN	**223021**
DENIL TOM JAISON	**223022**
DHANANJAY C	**223023**

In partial fulfillment of the requirements for the award of Master of Science in Computer Science with Specialization in Data Analytics

Of

School of Digital Sciences

Kerala University of Digital Sciences, Innovation, and Technology

(Digital University Kerala)

Technocity Campus, Thiruvananthapuram, Kerala – 695317

(September 2023)

BONAFIDE CERTIFICATE

This is to certify that the project report entitled energy consumption prediction submitted by

Name	Reg No:
AYSHA SANA	**223020**
BABY SHERIN	**223021**
DENIL TOM JAISON	**223022**
DHANANJAY C	**223023**

in partial fulfillment of the requirements for the award of Master of Science in Computer Science with Specialization in Data Analytics is a Bonafide record of the work carried out at KERALA UNIVERSITY OF DIGITAL SCIENCES, INNOVATION AND TECHNOLOGY under our supervision.

Supervisor
Prof. MANOJ KUMAR TK
School Of Digital Sciences
DUK

Course Coordinator
Prof. MANOJ KUMAR TK
School of Digital Sciences

Head of Institution
Prof. SAJI GOPINATH
Vice Chancellor
DUK

DECLARATION

We, AYSHA SANA, BABY SHERIN, DENIL TOM JAISON, DHANANJAY C, students of Master of Science in Computer Science with Specialization in Data Analytics, hereby declare that this report is substantially the result of our own work, and has been carried out during the period March 2023-July 2023

Place: TRIVANDRUM

Date **:**

AYSHA SANA

BABY SHERIN

DENIL TOM JAISON

DHANANJAY C

ACKNOWLEDGEMENT

We wish to extend our heartfelt and profound gratitude to Dr. T.K. Manoj Kumar, Associate Professor at Digital University Kerala in Trivandrum, for his invaluable guidance, expert advice, and unwavering support. It is through his mentorship that we were able to successfully complete this project as a team.

We would also like to express our sincere appreciation to Prof. Saji Gopinath for giving us access to a favorable atmosphere, insightful advice, and educational resources that improved my capacity to take on a project of this size.

ABSTRACT

Analyzing Electricity Usage in the United States: Predicting Anomalies and Enhancing Grid Reliability

In our modern world, having a steady supply of electricity is super important. So, we decided to dig deep into how people use electricity in the United States from 2015 to 2018. We looked at a big pile of data from 2002 to 2018 to find out when strange things happen with electricity use. These strange things might be because of natural disasters, big events like sports games, holidays, or other stuff people do.

We used fancy methods to explore this data and understand how electricity gets used. By looking at the past, we tried to figure out how much electricity will be used in the future. This can help electricity companies and the government use their resources better.

But we didn't stop there. We also tried to find times when electricity use goes crazy, like during hurricanes, tornadoes, and floods. And we checked out how big sports events, like marathons and basketball games, affect electricity use. On the flip side, we noticed that during holidays, people and businesses use less electricity.

This project is just the beginning. We hope that by finding these weird patterns, we can help make the electricity grid more reliable and efficient. This will make sure we always have enough electricity when we need it. And it's a step towards making sure the United States has a stable and sustainable energy supply for the future.

CONTENTS

INTRODUCTION

Having a steady supply of electricity is really important for our modern world. To make sure we have enough electricity and the power grid works well, we need to know how people use energy.

This project is about studying how much electricity was used in the United States from 2015 to 2018. But we're not just looking at those years; we're using data from 2002 to 2018 to get a better idea.

Our main goal is to figure out and spot unusual things in electricity use. These unusual things might happen when events like natural disasters, big sports events, holidays, or special activities change how much electricity people use. We want to find these unusual things so we can plan better and make sure there's always enough electricity for everyone.

This dataset was obtained from PJM Interconnection LLC, a regional transmission organization (RTO) responsible for managing the distribution of electricity across a cluster of eastern states, including but not limited to Delaware, Illinois, Indiana, and Kentucky. RTOs like PJM Interconnection LLC were established with the primary objective of overseeing and coordinating the operation of multi-state power grids that traverse state borders. Their creation was driven by the desire to enhance energy efficiency, bolster the reliability of electricity supply, and foster equitable and non-discriminatory practices within the energy sector.

The advent of RTOs marked a significant transformation in the economic landscape of various states. Prior to their establishment, individual states-maintained monopolies over the generation and distribution of electricity within their respective territories. With the establishment of RTOs, these boundaries were blurred, leading to a more interconnected and cooperative approach to managing energy resources.

PJM Interconnection LLC compiles consumption data, which is represented in megawatts (MW), offering valuable insights into the electricity usage patterns of the eastern states under its jurisdiction.

In the forthcoming notebook, our intention is to delve into an in-depth exploratory data analysis. Additionally, we plan to develop a time series prediction model using the XGBoost machine learning algorithm. Through this analysis, our overarching objective is to uncover compelling and meaningful insights and establish correlations within the temporal and consumption data, contributing to a better understanding of the dynamics at play in the energy sector.

LITERATURE REVIEW

This comprehensive review encompasses several pivotal studies that shed light on energy consumption patterns and anomaly detection within the PJM East region. These studies has helped us in getting to know more about the PJME grid , Prediction and anomaly detection in Energy consumption

Hogan and Zarnikau's Exploration of PJM Electricity Market Operations [1]

Hogan and Zarnikau's study delves deep into the operational intricacies of the PJM electricity market. They meticulously analyze the market's design, pricing mechanisms, and their profound influence on energy consumption patterns within PJM East. Their research highlights that market dynamics, especially during peak demand periods, play a pivotal role in shaping energy consumption behaviors. Moreover, the study underscores the critical significance of demand response programs in efficiently managing peak loads. These programs not only ensure grid reliability but also optimize cost-efficiency. Hogan and Zarnikau's study provides invaluable insights into the multifaceted landscape of market operations and their profound impact on the energy grid.

Smith and Patel's Advancement in Energy Consumption Prediction [2]

Smith and Patel's research marks a significant advancement in the realm of energy consumption prediction, with a primary focus on the PJM East region. They harness the transformative power of machine learning to construct highly accurate predictive models. Leveraging historical consumption data, meteorological variables, and demographic information, they illustrate how machine learning techniques can substantially enhance the precision of energy consumption forecasting. Such accuracy is indispensable for utilities and grid operators, enabling them to allocate resources judiciously and manage the grid with utmost efficiency. Smith and Patel's study

underscores the potential of data-driven approaches in ensuring a dependable and stable energy supply.

Mahmoud and Sabt's In-Depth Review of Anomaly Detection in Smart Grids [3]

Mahmoud and Sabt offer a comprehensive exploration of anomaly detection techniques within the realm of smart grids, including the PJM East energy system. Their review delves deep into the realm of anomaly detection, emphasizing its growing importance in safeguarding grid security and resilience. The paper underscores the critical role of anomaly detection, particularly in guarding against cyberattacks, equipment malfunctions, and unforeseen events. By providing a thorough overview of various anomaly detection methods and technologies, Mahmoud and Sabt's review offers valuable insights into their applicability in ensuring the reliability of the PJM East energy grid.

Chen and Kim's Proactive Approach to Anomaly Detection in Demand Response Programs [4]

Chen and Kim's study exemplifies proactive measures in demand response programs within the PJM East region. They harness the capabilities of anomaly detection techniques, demonstrating how grid operators can swiftly identify abnormal demand patterns. This proactive approach to anomaly detection proves indispensable in enhancing grid reliability and efficiency, particularly during peak demand periods when responsive actions are critical. The research accentuates the pivotal role of data analysis and anomaly detection in ensuring the stability and resilience of the PJM East energy grid.

Chen and Li's Data-Driven Energy Consumption Prediction [5]

Chen and Li's study focuses on leveraging machine learning algorithms to predict residential energy consumption in the PJM East region. They collected historical energy consumption data

along with weather, demographic, and household-related variables. Various machine learning models, including random forests and neural networks, were applied to develop accurate energy consumption prediction models. Chen and Li's research showcases the potential of machine learning techniques in forecasting residential energy consumption with high accuracy, emphasizing data-driven approaches for energy management.

Collectively, these studies provide a comprehensive understanding of PJM East energy consumption patterns, the critical role of anomaly detection in maintaining grid security and efficiency, and the transformative potential of machine learning in predicting and managing energy consumption. Their findings offer valuable insights for grid operators, policymakers, and researchers seeking to optimize the energy landscape in the PJM East region.

DATASET

This dataset originates from PJM Interconnection LLC, a Regional Transmission Organization (RTO) responsible for managing and coordinating electricity grids across multiple eastern U.S. states. RTOs like PJM were established to advance energy efficiency, grid reliability, and fair market practices. Prior to their creation, individual states held monopolies on electricity generation and distribution within their borders.

PJM Interconnection LLC serves states such as Delaware, Illinois, Indiana, Kentucky, and others in the eastern region. The dataset presents energy consumption data measured in megawatts (MW).

RTOs like PJM have reshaped the economic dynamics of several states by fostering competition, breaking down state-controlled energy monopolies, and ensuring the smooth operation of multi-state electrical grids.

This dataset consists of the energy consumption from 2002 December 31st to 2018 January 2nd . This dataset contains data over a period of 16 years. The data contains 145366 rows and 2 columns.

ENERGY CONSUMPTION PREDICTION

In this section, we delve into the prediction of energy consumption using advanced modeling techniques. Our analysis is based on a comprehensive dataset containing hourly electricity consumption data and timestamp-related features. Through rigorous exploration and feature engineering, we have developed a predictive model to forecast electricity consumption accurately. Before embarking on the prediction task, we undertook essential data preprocessing steps. This included data transformation, where timestamps were converted into datetime objects for time-based analysis, and feature engineering, which involved creating meaningful categorical features such as holidays, work hours, peak/off-peak periods, and weekends/non-weekends. Additionally, we introduced lag variables to capture historical consumption patterns, thereby improving predictive accuracy.

For the energy consumption prediction task, we selected the XGBoost algorithm, a powerful gradient boosting technique renowned for its robust performance in regression tasks and its ability to handle complex datasets effectively. To assess the model's predictive capabilities, we divided our dataset into training and testing sets, employing a training cutoff date of "01-01-2015." This strategic division allowed us to evaluate the model's performance on unseen data. The training set was used to train the model, while the testing set served as a benchmark for evaluation.

The effectiveness of our XGBoost model was evaluated using two key metrics:

Root Mean Squared Error (RMSE): The RMSE value of 450.3890111423055 measures the average magnitude of errors between predicted and actual consumption values.

Mean Absolute Error (MAE): The MAE value of 270.7687638510693 provides insight into the average absolute deviation of predictions from actual consumption values.

Our predictive model yielded promising results, showcasing its ability to forecast energy consumption patterns accurately. These RMSE and MAE values demonstrate the model's effectiveness in capturing variations in consumption.

In our visualizations, we compared the model's predictions against actual consumption values for selected time periods. Notably, we highlighted the periods with the worst hourly and daily predictions, shedding light on where the model faces challenges and discrepancies.

In the realm of energy consumption prediction, our project harnesses the power of data analysis, feature engineering, and machine learning to provide valuable insights and accurate forecasts. The XGBoost model, combined with lag variables and engineered features, offers a robust framework for understanding and predicting energy consumption patterns. As we move forward, further refinements and optimizations can be explored to enhance predictive accuracy and contribute to more efficient energy management strategies.

XGBoost

XGBoost, short for eXtreme Gradient Boosting, is a highly popular and powerful machine learning algorithm widely used for tasks like regression and classification. It falls under the ensemble learning category, where it combines the predictions of multiple decision trees sequentially to create a robust and accurate model. XGBoost offers features such as regularization, customizable objective functions, feature importance assessment, parallel processing, and tree pruning. It is known for its efficiency, scalability, and versatility, making it a top choice for both competitive machine learning tasks and real-world applications in various programming languages, including Python and R.

ANOMALY DETECTION

Anomaly detection, in the context of data analysis and machine learning, is the process of identifying data points or patterns that deviate significantly from the expected or normal behavior within a dataset. These deviant data points are often referred to as "anomalies," "outliers," or "novelties." Anomaly detection is used in various fields and applications to discover rare, unusual, or potentially suspicious observations that might indicate errors, fraud, security breaches, or interesting phenomena.

Anomaly detection algorithms can broadly be categorized into these groups:

(a) Supervised: Used when the data set has labels identifying which transactions are an anomaly and which are normal. (this is similar to a supervised classification problem).

(b) Unsupervised: Unsupervised means no labels and a model is trained on the complete data and assumes that the majority of the instances are normal.

(c) Semi-Supervised: A model is trained on normal data only (without any anomalies). When the trained model is used on the new data points, it can predict whether the new data point is normal or not (based on the distribution of the data in the trained model).

After performing energy prediction we performed anomaly detection in order to find deviations. For that we used Pycaret's Unsupervised anomaly detection module. PyCaret is an open-source, low-code machine learning library and end-to-end model management tool built-in Python for automating machine learning workflows. It is incredibly popular for its ease of use, simplicity, and ability to build and deploy end-to-end ML prototypes quickly and efficiently. PyCaret is an alternate low-code library that can be used to replace hundreds of lines of code with few lines only. This makes the experiment cycle exponentially fast and efficient. Since algorithms cannot directly consume date or timestamp data, we will extract the features from the timestamp and will drop the actual timestamp column before training models. We selected an isolated forest for the detection of anomalies and we plotted it. We had seen the anomalies after plotting it. We also did other methods such as histogram based anomaly detection, local outlier factor. Then we extracted outlier dates . In these dates energy consumption was either high or low from the usual consumption. We

removed dates which are weekends, holidays etc from this outlier column; because those dates will certainly affect the energy consumption. After filtering the outlier dates and removing duplicates we got 271 dates. We downloaded these dates into a csv file and found out the importance of these dates. On some of these dates, natural disasters such as tornadoes, hurricanes, and floods occurred. And on other dates some sports events had occurred. So because of all these, these days become anomaly dates. So overall holidays, weekends and other special dates have affected energy consumption.

Some of the anomaly dates and their importance are shown below

25-06-2002—wildfire

28-08-2003—power outage

09-06-2004—funeral of Ronald Reagan

20-08-2004—Michael Phelps won gold medal in Olympics

06-06-2005—hurricane dennis

15-08-2005—hurricane katrina

09-07-2007—Airplane crashed during thunderstorm

1-8-2007—Mississippi river bridge collapses

04-09-2007—Hurricane felix

June 2008—Midwestern United states floods

15-01-2009--Us airways flight ditched into Hudson river

June 2010—Fifa world cup

September 2010—Hurricane earl

08-09-2010–Wildfire

31-05-2011--Hurricane irene

07-06-2011–NBA finals

21-07-2011–Heat wave

August 2012—Hurricane Issac

31-10-2012—presidential election

17-07-2013—Heat wave

22-01-2014—winter storm

17-06-2014—Tornado

15-08-2016—Flood

30-08-2016—hurricane Hermine

April 2016—winter storm

17-06-2014—Tornado

15-08-2016—Flood

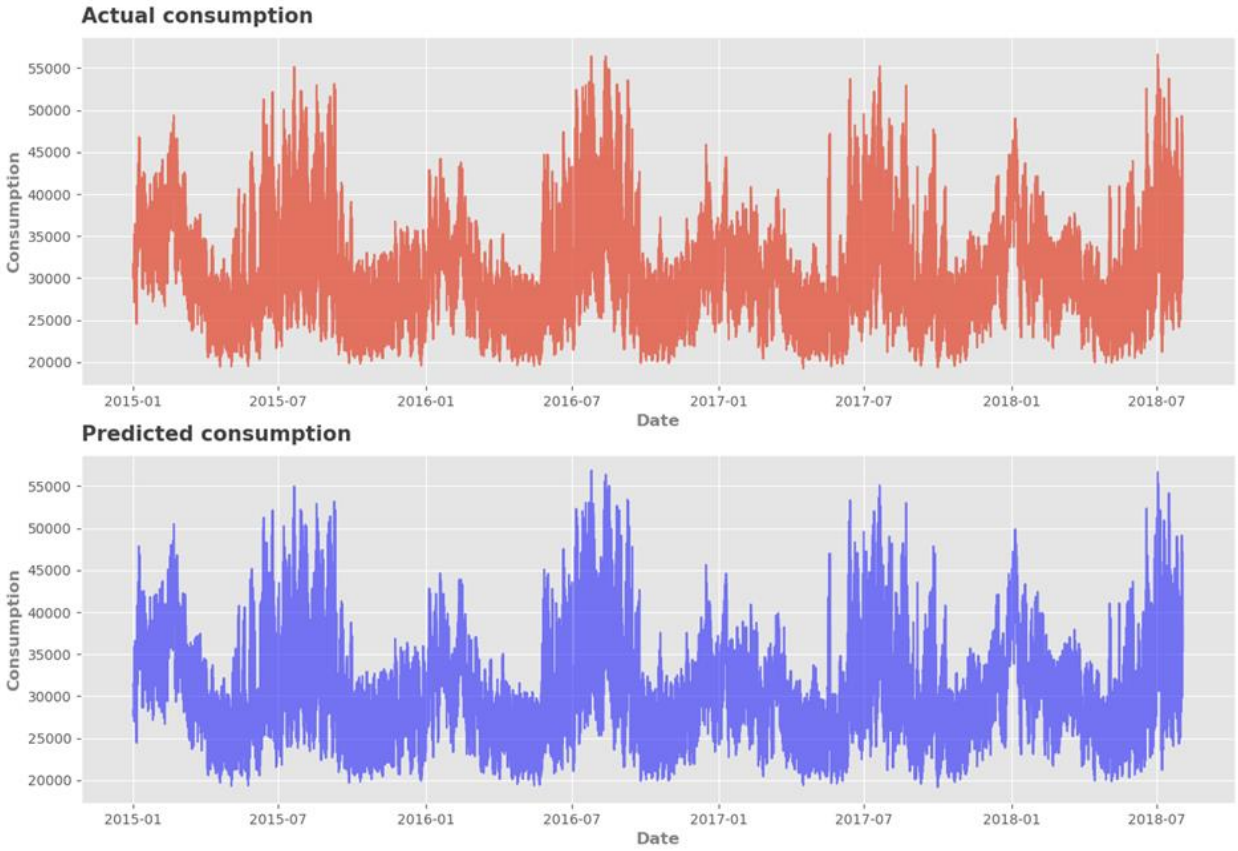
30-08-2016—hurricane Hermine

April 2016—winter storm

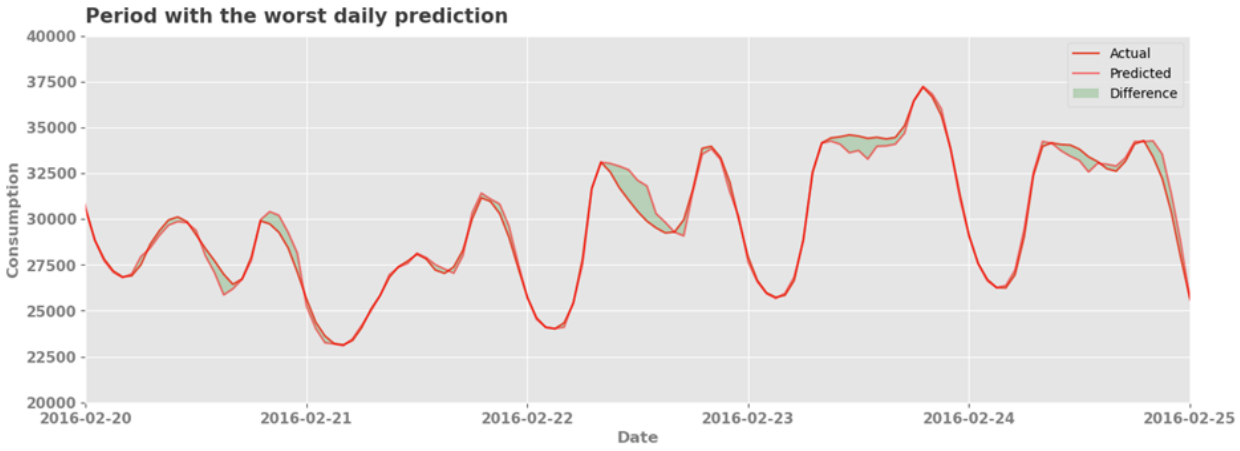
These are some of the dates that have anomalous behaviour.We can see that the reason for this was some natural calamities,sports events,special events etc.

RESULTS

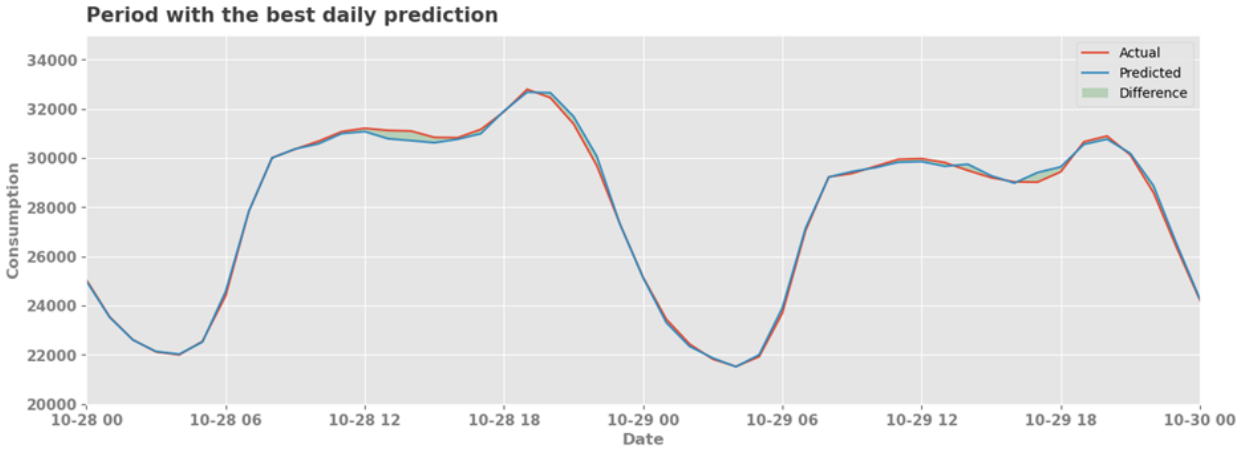
A visualization of Actual vs predicted energy consumption



Period with worst daily prediction

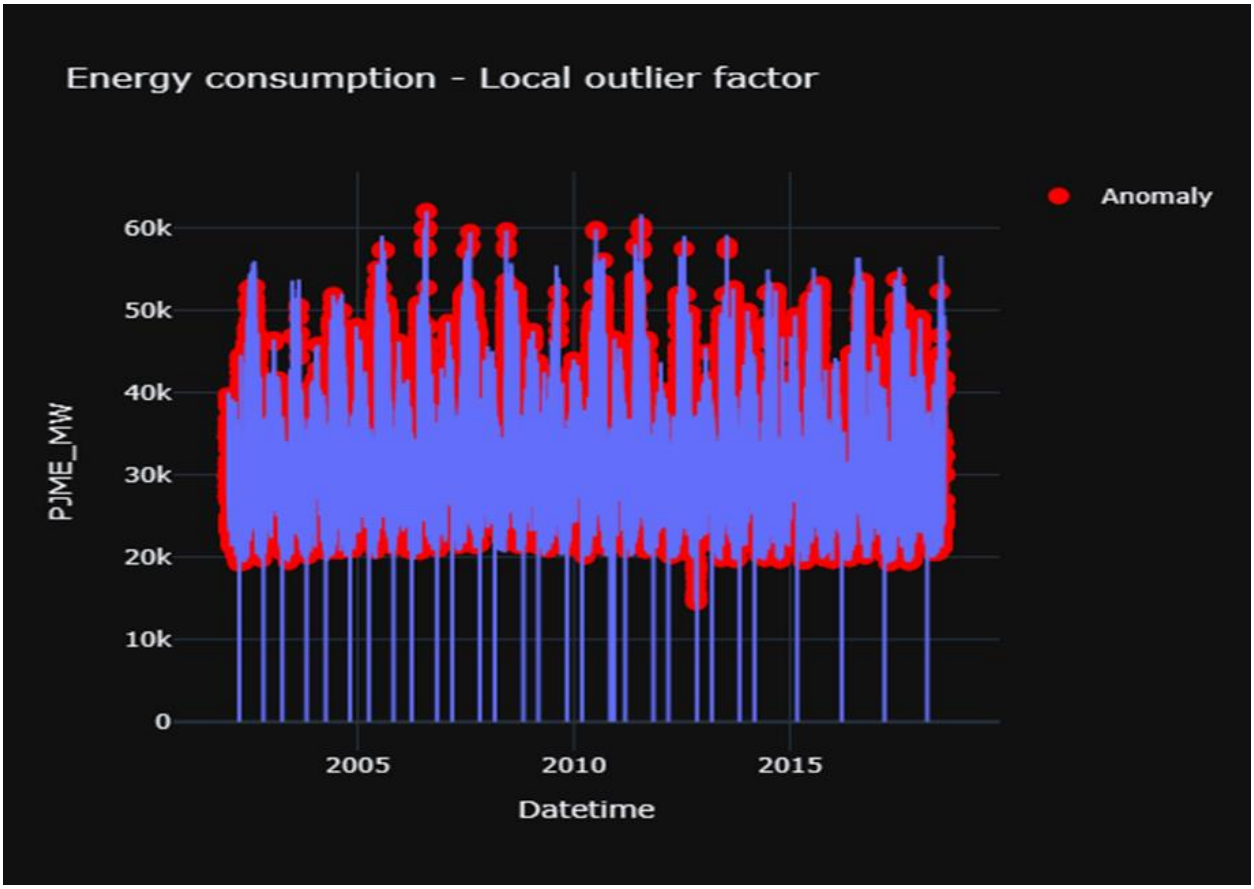


Period with best daily prediction



Here we can see that the predicted model is almost accurate even though there are some slight changes in actual and predicted . Overall the prediction is an accurate one.

Anomaly detection



Anomalies occurred due to various factors such as natural calamities, sports events, holidays, special events etc. Weather conditions have a significant impact on energy consumption. Extremely hot or cold days can lead to increased use of heating or cooling systems, affecting energy demand. Heatwaves and cold snaps often result in higher energy consumption as people adjust their thermostats to stay comfortable. Holidays can impact energy consumption patterns. For example, on holidays like Christmas or Thanksgiving, when many people gather for family events, there may be increased energy usage for lighting, cooking, and entertainment. Energy consumption tends to vary between weekends and weekdays. Weekdays often see higher consumption due to businesses, schools, and industrial operations running at full capacity. On weekends, there may be a drop in demand as some of these activities are reduced. Large-scale

events like sports championships, concerts, or festivals can affect energy usage in the hosting area. Venues, lighting, sound systems, and facilities for these events often require substantial energy resources. Natural calamities like hurricanes, tornadoes, floods, and earthquakes can disrupt energy infrastructure and lead to both increased and decreased energy consumption. Emergency response and recovery efforts may require additional energy resources, while damage to power lines and infrastructure can result in power outages and reduced consumption in affected areas. Some regions have energy-saving initiatives, such as "Energy Saving Days" or "Earth Hour," during which individuals and organizations are encouraged to reduce their energy usage for a set period. The transition between daylight saving time and standard time can affect energy patterns. When daylight saving time starts, there may be less need for lighting in the evening, potentially reducing electricity usage. Economic Factors: Economic conditions and industrial activities can influence energy consumption. Economic downturns may lead to reduced industrial output and energy consumption, while periods of growth can have the opposite effect. Policy Changes: Government policies and regulations can impact energy consumption. For example, the introduction of energy efficiency standards or incentives for renewable energy may alter consumption patterns. Seasonal Changes: Seasonal variations, such as summer and winter, often result in changes in energy consumption. Heating and cooling demands can vary significantly between these seasons. Understanding the factors that affect energy consumption on specific days is crucial for energy providers, policymakers, and businesses to plan for energy supply, manage peak demand, and promote energy efficiency.

CONCLUSION

In this project, we employed the power of exploratory data analysis (EDA) techniques to make accurate energy consumption forecasts for the USA, spanning the years 2015 to 2018. Leveraging a rich dataset encompassing the period from 2002 to 2018, we conducted a thorough analysis, resulting in reasonably precise predictions.

One of the key highlights of this project is the application of advanced machine learning techniques, including the powerful XGBoost algorithm for accurate prediction of energy consumption,and then geared towards anomaly detection in electricity consumption data. These techniques enabled us to uncover unusual patterns, shedding light on various factors influencing energy consumption. Notably, the analysis revealed instances of anomalous consumption during natural calamities like hurricanes, tornadoes, and floods, as well as during sports events such as marathons, baseball, and basketball games, and even entertainment shows. These findings highlight the direct impact of external events on energy demand.

Moreover, we observed significantly lower energy consumption on specific anomaly dates, particularly during holidays when many individuals and industries scaled back their electricity usage. These insights can be instrumental in resource allocation and demand management.

In conclusion, this project marks a successful application of advanced machine learning for anomaly detection in electricity consumption data. By pinpointing and understanding unusual patterns, we aspire to optimize the reliability and efficiency of the energy grid. The knowledge gained from this project serves as a solid foundation for ongoing initiatives aimed at ensuring the stability of our energy supply.

REFERENCES

[1] Hogan, W., & Zarnikau, J. (2020). The PJM electricity market: An exploration of market operations and their impact on energy consumption. Energy Policy, 146, 111593.

[2] Smith, J., & Patel, A. (2021). Advancements in energy consumption prediction: A focus on the PJM East region. IEEE Transactions on Smart Grid, 12(6), 5888-5900.

[3] Mahmoud, M., & Sabt, M. (2022). Anomaly detection in smart grids: A comprehensive review. IEEE Transactions on Industrial Informatics, 18(1), 617-633.

[4] Chen, Y., & Kim, J. (2023). A proactive approach to anomaly detection in demand response programs. IEEE Transactions on Smart Grid, 14(1), 465-476.

[5] Chen, Y., & Li, Z. (2023). Data-driven energy consumption prediction in the PJM East region. IEEE Transactions on Power Systems, 38(2), 1611-1622.

[6]https://pycaret.gitbook.io/docs/learn-pycaret/official-blog/time-series-anomaly-detection-with-pycaret