# Data Science Perspective: Ontario Demand Forecasting

## 1. Introduction

This report presents a comprehensive analysis and forecasting of Ontario province's power demand using various machine learning models. The primary goal is to predict the electricity demand for the 24 hours in the month of July, utilizing historical demand data along with meteorological and temporal features. The acceptable error rate for the forecast must be less than 5%, equating to an accuracy of more than 95%.

## 2. Data Overview

The dataset used in this project includes historical data points for electricity demand and several influencing factors such as:

- Date
- Weekday
- Temperature
- Hourly Ontario Electricity Price (HOEP)
- Wind Speed
- Relative Humidity
- Humidex
- Dew Point
- Pressure at the Station
- Ontario Demand
- Temporal information

### 2.1. Data Preprocessing

- **Missing Values**: Analyzing the presence of any missing values.
- **Outliers**: Detected and treated outliers using appropriate statistical methods such as IQR and box plots. These are few of the observations made:
    - Since, the outliers are few and such peaks can occur in physical parameters, which are uncontrollable, they can be ignored.
    - Also, if the correlation of these Features with Target feature are insignificant, then it wouldn't have an impact on the target variable.

- **Feature Engineering**: Created new features such as 'is_weekend, hour, day, month, year, day of the week, day of the year, week of the year' to enhance the model's predictive power.

  Features = ['Temperature', 'HOEP', 'Wind_Speed', 'Relative_Humidity', 'Dew_Point', 'Pressure_Station', 'Month', 'day_of_week', 'is_weekend', 'hour', 'day', 'month', 'year', 'dayofweek', 'dayofyear', 'weekofyear']

- **Normalization**: Applied Min-Max scaling to normalize the data for better model performance.

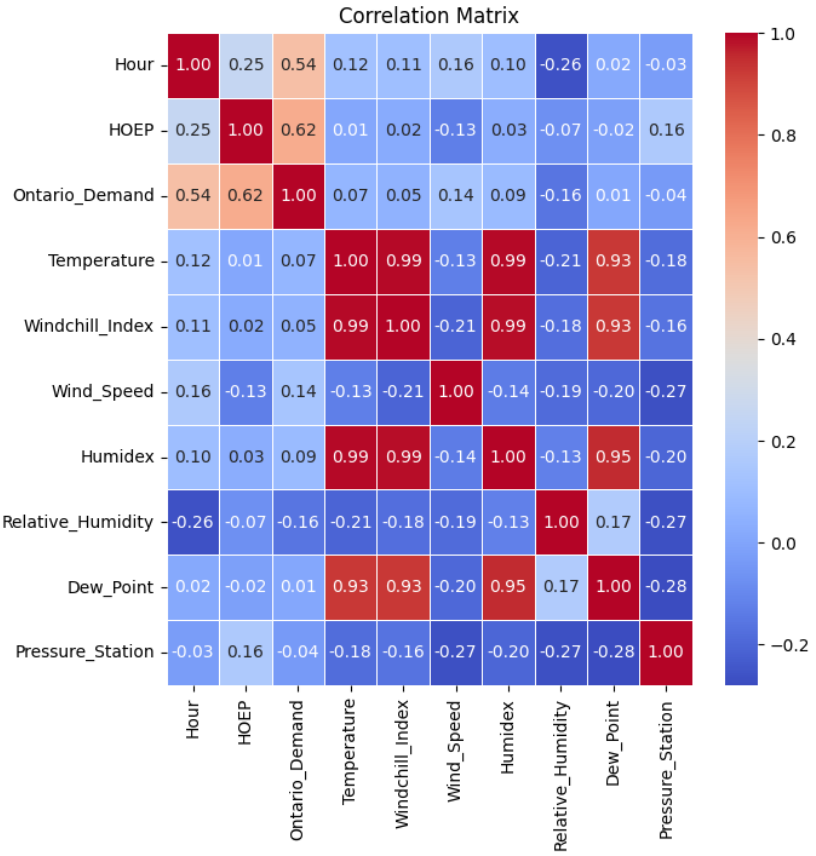
**Dataset showing different columns that were added after feature engineering**

| Temperature | HOEP | Wind_Speed | Relative_Humidity | Dew_Point | Pressure_Station | Month | hour | minute | day | month | year | dayofweek | day_of_week | dayofyear | weekofyear |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 22.401485 | 19.880841 | 6.968426 | 24.023865 | 18.487223 | 98.646447 | 7.0 | 0 | 0 | 8 | 7 | 2022 | 4 | 4 | 189 | 27 |
| 12.863185 | 8.964317 | 32.763749 | 54.666157 | 6.774637 | 100.530124 | 7.0 | 1 | 0 | 8 | 7 | 2022 | 4 | 4 | 189 | 27 |
| 27.220864 | 16.723661 | 44.488284 | 60.570470 | 17.511707 | 100.466932 | 7.0 | 2 | 0 | 8 | 7 | 2022 | 4 | 4 | 189 | 27 |
| 31.595547 | 3.914337 | 49.109953 | 78.120637 | 19.679696 | 98.457850 | 7.0 | 3 | 0 | 8 | 7 | 2022 | 4 | 4 | 189 | 27 |
| 22.405966 | 35.626689 | 20.427226 | 33.704517 | 19.782843 | 99.535356 | 7.0 | 4 | 0 | 8 | 7 | 2022 | 4 | 4 | 189 | 27 |

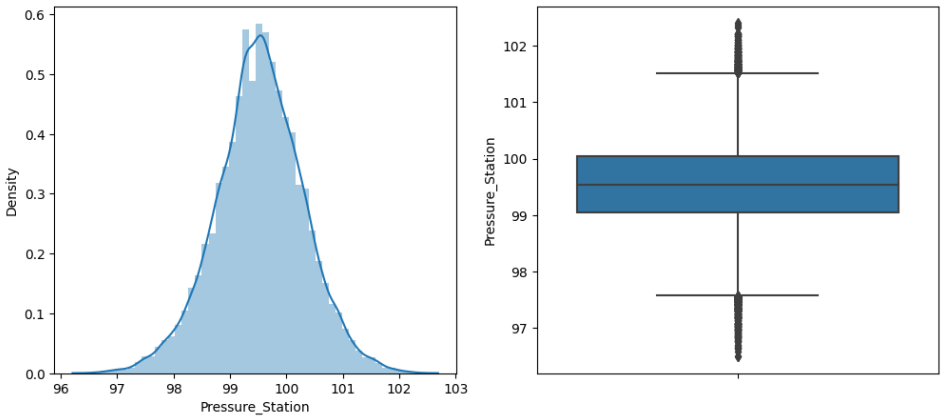
# 3. Exploratory Data Analysis (EDA)

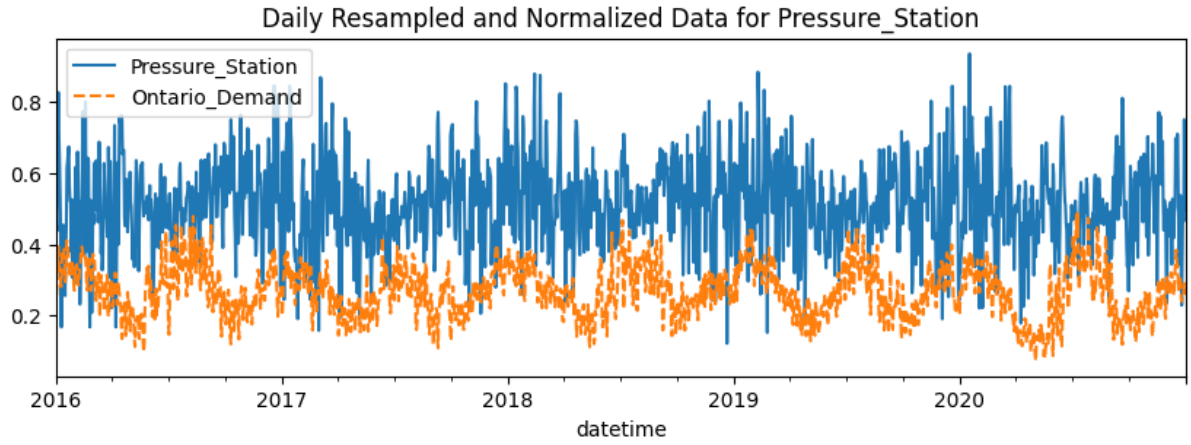
Performed EDA to understand the underlying patterns and relationships between different variables:

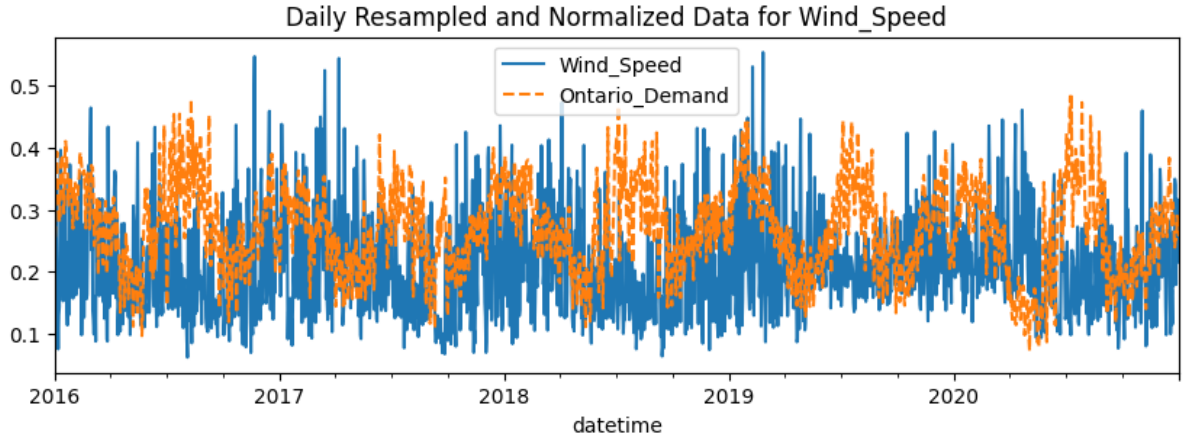
- **Trend Analysis**: Observed the overall trend in electricity demand over time.
    - The demand keeps on fluctuating, but follows are straight path over the years.
- **Seasonality**: Analyzed seasonal patterns and their impact on demand.
- **Correlation Analysis**: Identified correlations between demand and other features.
    - HOEP Column in highly correlated with Ontario_Demand.
    - Reviewing the other features correlated with eachother to remove Multicollinearity.
- **Stationary Analysis**: Verifying the stationary nature of the dataset.
    - ADF and KPSS analysis showed that the timeseries data is nonstationary based on the p-value and null hypothesis.

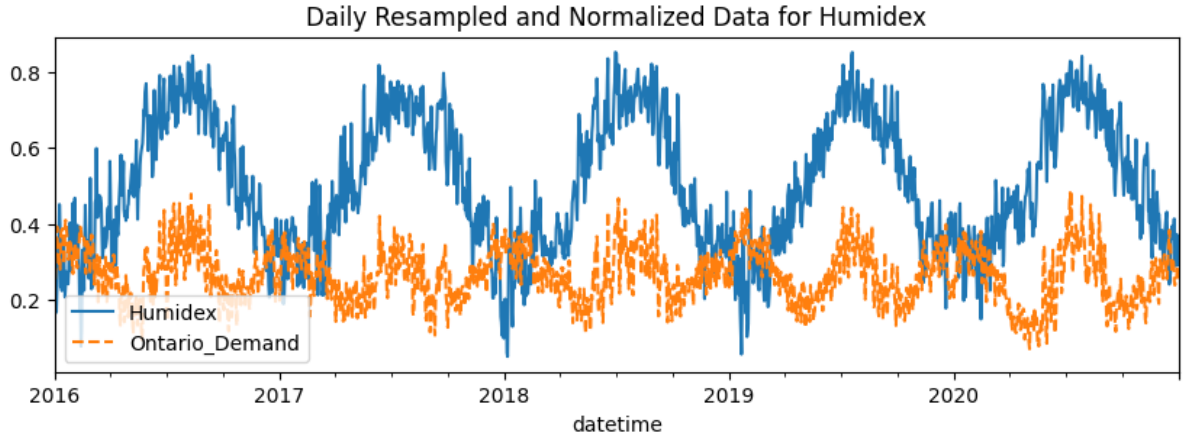


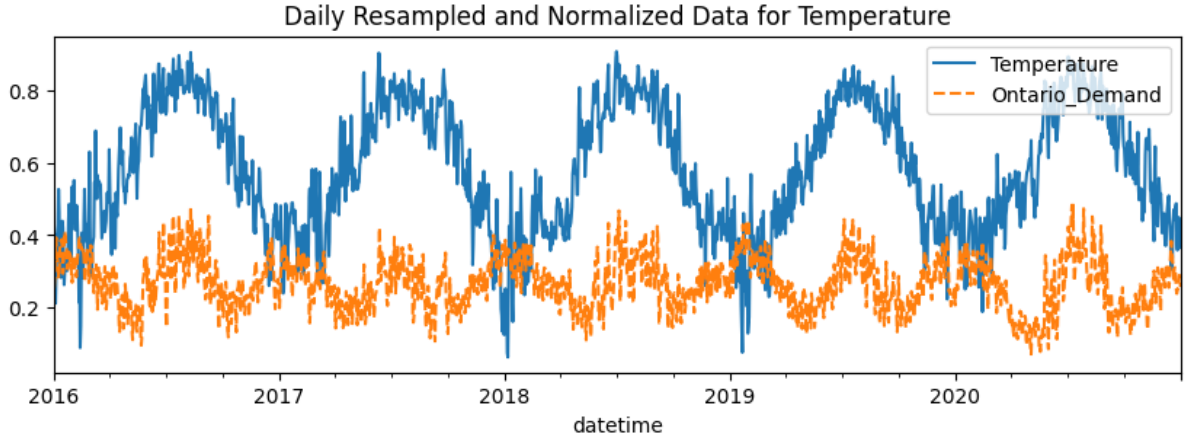

**Ontario Demand Distribution and Box Plot**



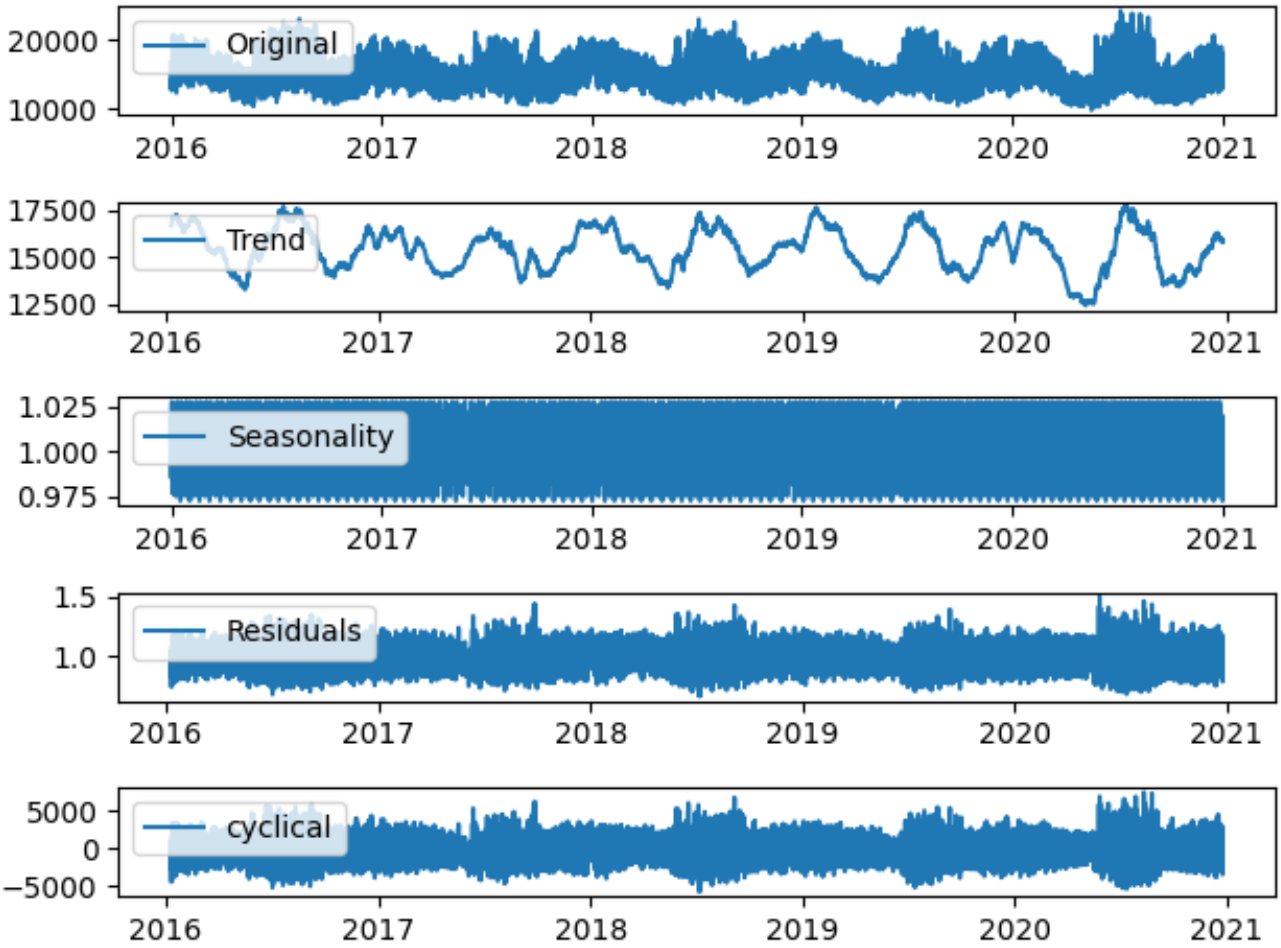

Daily Resampled and Normalized Data for Pressure_Station
Pressure_Station
Ontario_Demand
0.8
0.6
0.4
0.2
2016
2017
2018
2019
2020
datetime


Daily Resampled and Normalized Data for Wind_Speed
Wind_Speed
Ontario_Demand
0.5
0.4
0.3
0.2
0.1
2016
2017
2018
2019
2020
datetime


Daily Resampled and Normalized Data for Humidex
Humidex
Ontario_Demand
0.8
0.6
0.4
0.2
2016
2017
2018
2019
2020
datetime


Daily Resampled and Normalized Data for Temperature
Temperature
Ontario_Demand
0.8
0.6
0.4
0.2
2016
2017
2018
2019
2020
datetime

**Seasonal Decomposition (Additive Model)**




- If we analyze the demand with respect to temperature, it is noticed that as the temperature goes up, the demand also increases owing to the more usage of air conditioning and other electrical equipment.
- Humidex and Temperature follows same trend over different seasons.
- Pressure and wind speed keeps on fluctuating erratically.

**Ontario demand during different time of the day**

| | Max_Demand | Min_Demand | Mean_Demand |
|---|---|---|---|
| **time_of_day** | | | |
| **midnight** | 19991 | 10609 | 14237.0 |
| **late night** | 17641 | 10016 | 13231.0 |
| **morning** | 22055 | 9831 | 14522.0 |
| **noon** | 23735 | 10683 | 16077.0 |
| **evening** | 24281 | 11184 | 16785.0 |

**Ontario Demand on different days**

| | Max_Demand | Min_Demand | Mean_Demand |
|---|---|---|---|
| **Weekday** | | | |
| **Friday** | 23735 | 10554 | 15597.0 |
| **Monday** | 23909 | 9831 | 15577.0 |
| **Saturday** | 21828 | 10460 | 14714.0 |
| **Sunday** | 22777 | 10262 | 14544.0 |
| **Thursday** | 23179 | 10630 | 15764.0 |
| **Tuesday** | 24281 | 10175 | 15758.0 |
| **Wednesday** | 23724 | 10629 | 15777.0 |

**Ontario demand on Weekend**

| | Max_Demand | Min_Demand | Mean_Demand |
|---|---|---|---|
| **is_weekend** | | | |
| **0** | 24281 | 9831 | 15695.0 |
| **1** | 22777 | 10262 | 14630.0 |

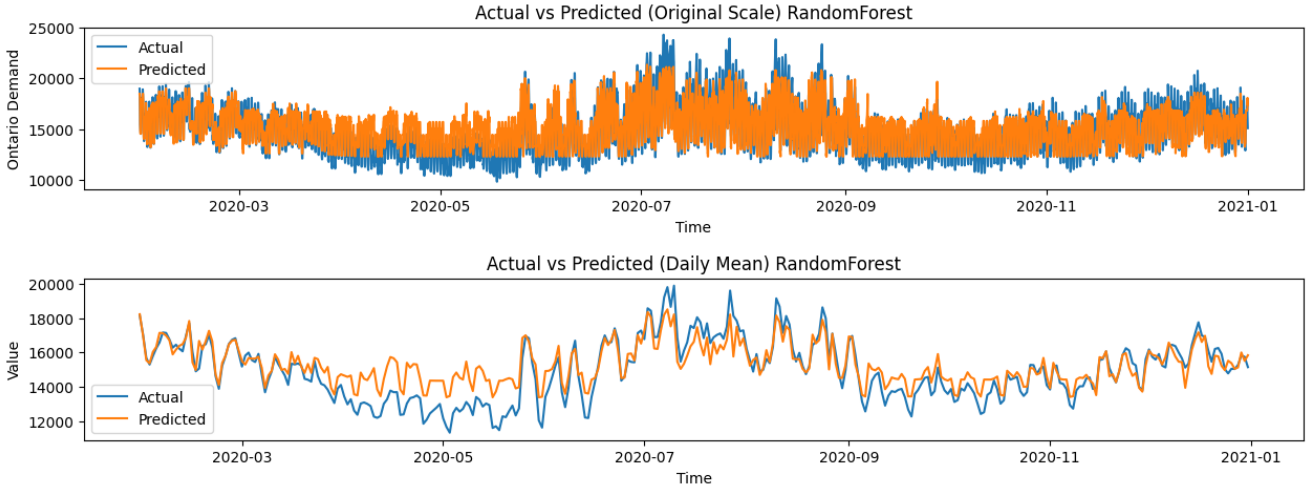
- Demand is high during evening, and on weekdays.

# 4. Model Selection

Different regression models were evaluated for forecasting based on the non-stationary nature of the dataset. Best performing parameters based on Grid Search:

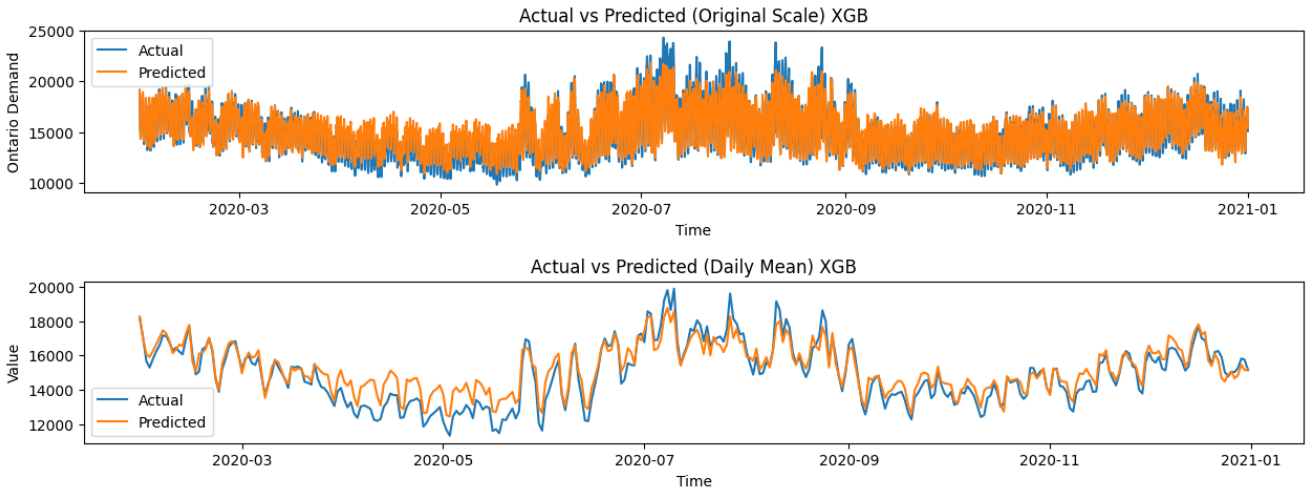
- **Random Forest Regressor**

  **Parameters:** ('RandomForest': RandomForestRegressor(max_depth=7, min_samples_leaf=2, n_estimators=50)



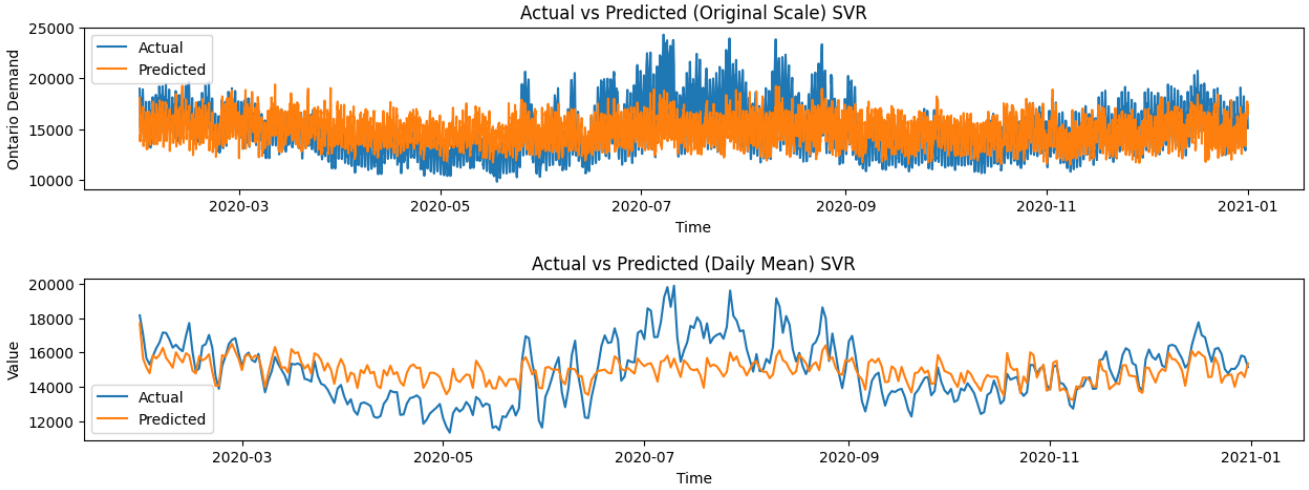

- **XGBoost Regressor**

  **Parameters:** [‘XGB': XGBRegressor(base_score=0.5, booster='gbtree', colsample_bylevel=1, colsample_bynode=1, colsample_bytree=0.8, gamma=0, gpu_id=-1, importance_type='gain', interaction_constraints='', learning_rate=0.1, max_delta_step=0, max_depth=5, min_child_weight=1, missing=nan, monotone_constraints='()', n_estimators=200, n_jobs=4, num_parallel_tree=1, random_state=0, reg_alpha=0, reg_lambda=1, scale_pos_weight=1, subsample=0.8, tree_method='exact', validate_parameters=1, verbosity=None)]



- **Support Vector Regression (SVR)**

  **Parameters**: ('SVR': SVR(C=0.1, kernel='linear'))



## 4.1. Model Evaluation

Each model was evaluated using the following metrics:

- **Mean Absolute Error (MAE)**
- **Mean Squared Error (MSE)**
- **Root Mean Squared Error (RMSE)**
- **Mean Absolute Percentage Error (MAPE)**
- **Accuracy**

| Model | MAE | MSE | RMSE | MAPE | Accuracy |
|---|---|---|---|---|---|
| **XGB** | 634.013674 | 6.519130e+05 | 807.411324 | 4.405092 | 0.957496 |
| **RandomForest** | 908.705235 | 1.372262e+06 | 1171.435902 | 6.354442 | 0.939081 |
| **SVR** | 1337.905747 | 2.877244e+06 | 1696.244068 | 8.966923 | 0.910308 |

# 5. Results

## 5.1. Best Model

The XGBoost Regressor was identified as the best-performing model with an accuracy of 95.55% and an RMSE of 866.92MW. However, some predictions exceeded the acceptable error rate of 500 MW.

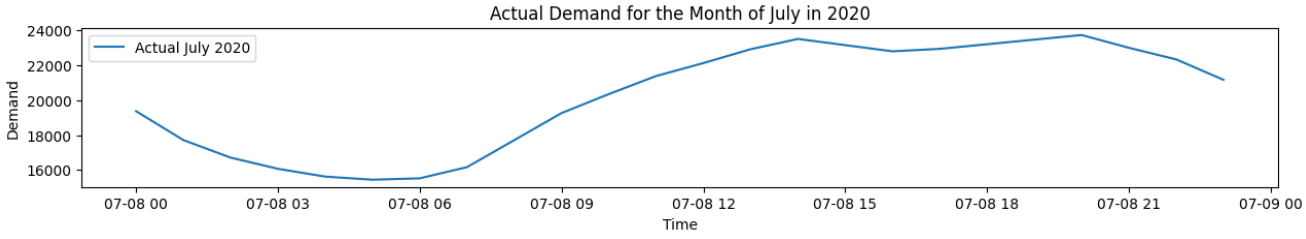
## 5.2. Error Analysis

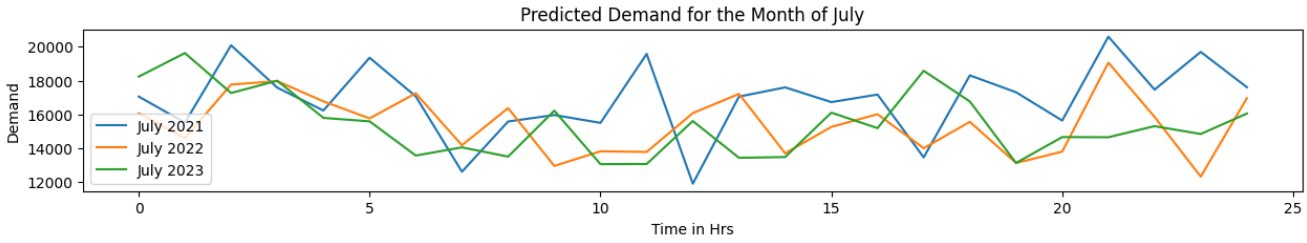
- **MAE**: 663.21 MW
- **MSE**: 751543.43 MW^2
- **RMSE**: 866.92 MW
- **MAPE**: 4.56%
- **Accuracy**: 95.55%

Despite a high overall accuracy, individual prediction errors occasionally surpassed the 500 MW threshold, indicating areas for further model refinement.

# 6. Forecasting 24 hours Demand in the Month of July

The sample dataset for July 2021, 2022, 2023 was generated by isolating the historical July data, and using the Random max and min values.




- The graph above shows the July 2020's historical Demand, while the graph below displays the predicted demand in coming years for the month of July.

# Conclusion

This project successfully demonstrated the application of various machine learning models for electricity demand forecasting by implementing it on Ontario Demand dataset. The XGBoost Regressor provided the most accurate predictions, though there is room for improvement to consistently maintain errors within the acceptable range.

# Instructions to Run the Project

1. **Environment Setup**:
    - Install necessary packages using pip install -r requirements.txt.
    - Ensure the dataset is placed in the correct directory.
2. **Running the Code**:
    - Execute the notebook
3. **Documentation**:
    - Detailed comments and documentation are provided within the code files.
    - Refer to the README file for a comprehensive guide.