STAT 615 Final Project

Pirro Prifti

December 18, 2024

1 Introduction

Energy plays a crucial role in our modern lives. Without it, our world becomes dark, and we return to a more primitive way of life. That being said, the supply and demand of natural gas, energy, and other commodities represents a complex dynamic which is governed by multiple players. Whether you are a producer, consumer, seller, or buyer, or even a combination of these, being able to predict the movements of this market is essential if you seek to be an informed participant.

That begs the question, what exactly do we need to predict, and which variables and methods will help us in this endeavor? One key metric will be the demand of energy itself. If we can accurately predict the dynamics of this phenomena, then we can make better decisions on when to buy, sell, and at what price and quantity.

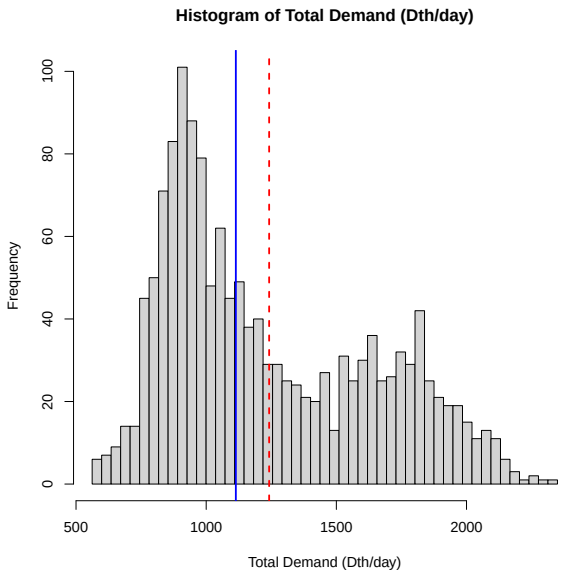
Given that demand is a continuous random variable (as opposed to a categorical one), it lends itself well to regression analysis, which will be the main tool used for both prediction and inference in this study. Energy markets, especially natural gas, are often strongly influenced by weather fundamentals. From an intuitive perspective, when temperatures rise, people need more energy to keep themselves cool. Similarly, when temperatures drop, people require more energy to stay warm. Therefore, weather data is likely to be a strong predictor of energy demand.

With a clear understanding of what we aim to uncover, the tools we'll use, and the data that will inform our analysis, lets explore a particular case study.

2 Data

Natural gas is delivered to customers in the Northeast (Connecticut, Massachusetts, Vermont, New Hampshire, Rhode Island, Maine) on the Algonquin (AGT) pipeline. As the largest city in the region, Boston's weather has a large impact on AGT gas demand each day.

This project aims to understand and model AGT natural gas demand. AGT publishes its daily natural gas demand in two categories: Residential and Commercial, and demand from natural gas fired power plants. Both categories are measured in dekatherms per day (Dth/day). For this study, the variable of interest is total demand, calculated as the sum of these two categories. The data ranges from the start of 2019 to the end of 2022. Below is a snap shot of what the total demand looks like through its histogram. The solid blue line represents its median with the dashed red line representing the mean. It exhibits a positive skew which might pose challenges for regression modelling down the road. As expected, due to the skew, its mean its larger than its median.



To explain the pipeline's demand, weather data from the National Oceanic and Atmospheric Administration (NOAA) is used. Specifically, the data is sourced from the Integrated Surface Database (ISD), a global repository of hourly weather observations collected from over 35,000 weather stations worldwide. This database is managed by the National Centers for Environmental Information (NCEI), a branch of the NOAA. The key features of this dataset include air temperature, dew point temperature, sea level pressure, wind direction, wind speed, total cloud cover, and one- and six-hour accumulated precipitation. These hourly observations span the same time period as the demand data, from 2019 to 2022, resulting in a total of 43,818 instances. According to the documentation, missing values are coded as -9999, temperatures are recorded in Celsius and scaled by a factor of ten, and wind speeds are also scaled by a factor of ten.

The preprocessing of these datasets involved several steps to ensure compatibility and readiness for analysis. First, the weather data's temperature values were converted from Celsius to Fahrenheit, and the wind speed variable was scaled to its original unit. After parsing through the different features, it was revealed that most variables had little to no missing values except for the six-hour accumulated precipitation variable, which was comprised of over 50 percent missing values. This variable was subsequently removed from the dataset, while the remaining missing values were interpolated

A critical step in the data preparation process was aligning the demand and weather data dates to ensure a one to one merge. AGT reports demand based on its "gas day" which runs from 10AM EPT (Eastern Prevailing Time) to 10AM EPT the following day. As an example, the reported gas demand for January 1, 2023 would be the demand from Jan 1, 2023 10:00 AM EPT to Jan 2, 2023 10:00 AM EPT. In contrast, the weather data follows a standard calendar day format. To account for this difference, the weather data was adjusted to the AGT's gas day. Additionally, the weather data was aggregated to daily values in order to match the granularity of the demand data.

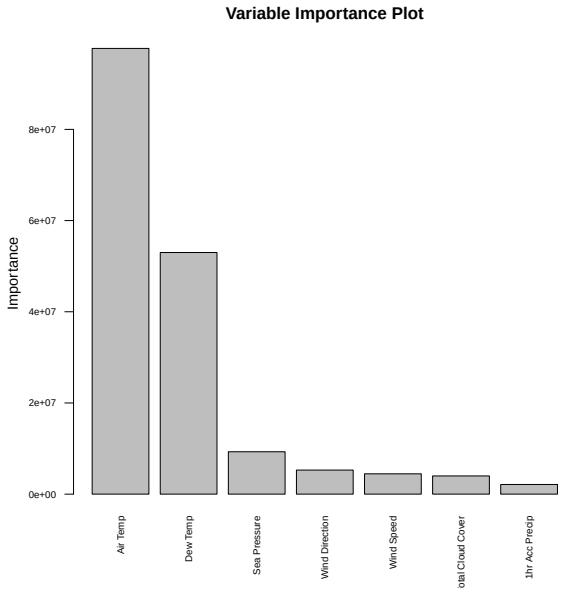
Once the datasets were full preprocessed and aligned, they were merged, and their observations were split into training and test sets. The training set consists of data from 2019 through the end of 2021, while the test set contains data from 2022.

3 Methods

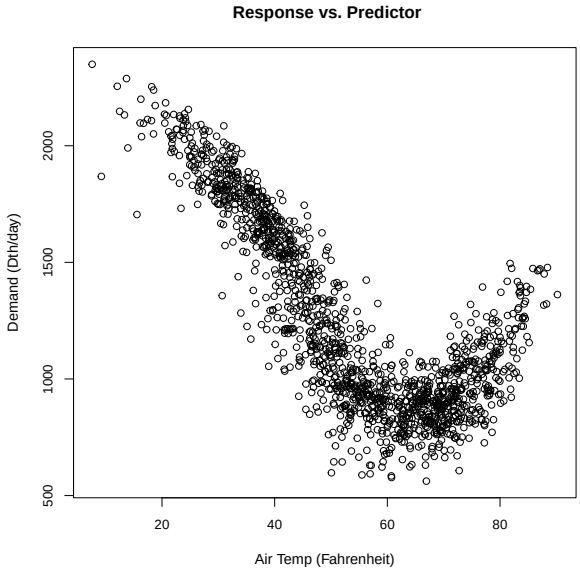
3.1 Polynomial Regression

Although the number of features in our dataset is relatively small, their importance likely varies significantly. To identify the most influential predictors, I began with variable selection using a random forests model. The model was run with default parameters for regression (500 trees, a minimum node size of 5, and an mtry equal to the number of features divided by three). From the resulting variable importance plot, Air

Temperature emerged as the most significant predictor by a wide margin, followed by Dew Point Temperature. The remaining variables contribute minimally to explaining the response. For simplicity and interpretability, we will proceed with Air Temperature as the primary variable for our analysis.



Now, lets examine the relationship between Air Temperature and our dependent variable, demand. A scatter plot reveals a clear non-linear relationship resembling that of a Nike swoosh.



To model this non-linear trend using linear regression, one technique we can employ is applying a polynomial transform to our independent variable to capture this non-linear trend. To balance model simplicity with predictive power and minimize the risk of overfitting, a third-degree polynomial transformation was applied to Air Temperature in the training set. The resulting model takes the form:

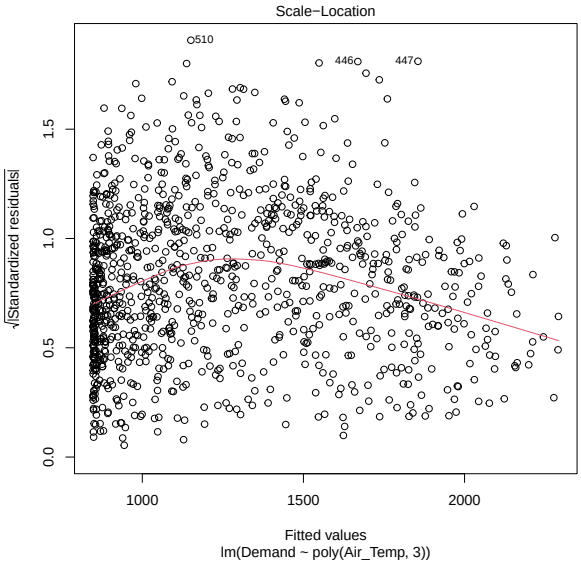
$$E[Y|X] = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$$

Next, we must check the assumptions of our model to ensure that we can draw reliable conclusions from it. The first assumption is linearity, which states that the relationship between the predictor and the response is linear. While this assumption was not initially met, we addressed it by applying a polynomial transformation to our predictor variable.

The second assumption is that the residuals of the model are normally distributed. We tested this using the Shapiro-Wilks test, which returned a significant p-value close to zero. This signifies that our residuals deviate from normality.

The third assumption is independence, stating that our residuals are not auto-correlated or related with one another. To check this, we conducted the Durbin-Watson test, which also returned a significant p-value of zero suggesting auto-correlation is present.

The fourth assumption is homoscedasticity, which requires that the variance of the residuals remain constant across all levels of the predictor. We initially assessed this visually using the scale-location plot. As shown below, the fitted line is not fully horizontal and flat and there seems to be some pattern in the residuals as well. This suggests that the homoscedasticity assumption has also been violated.



Finally, because we introduced additional features by transforming our independent variable, it is prudent upon us to also check for multicollinearity. Multicollinearity occurs when predictors are highly correlated with one another. Fortunately, R's `poly` function generates orthogonal polynomials, ensuring the transformed variables are uncorrelated. To confirm this, we computed the correlation between the polynomial terms, which were indeed zero. Thus, we can confidently conclude that multicollinearity is not a concern.

Our initial model reveals significant issues, as it fails to meet the assumptions of normality, independence, and homoscedasticity. These violations highlight a need for a different or refined appraoch to address these issues.

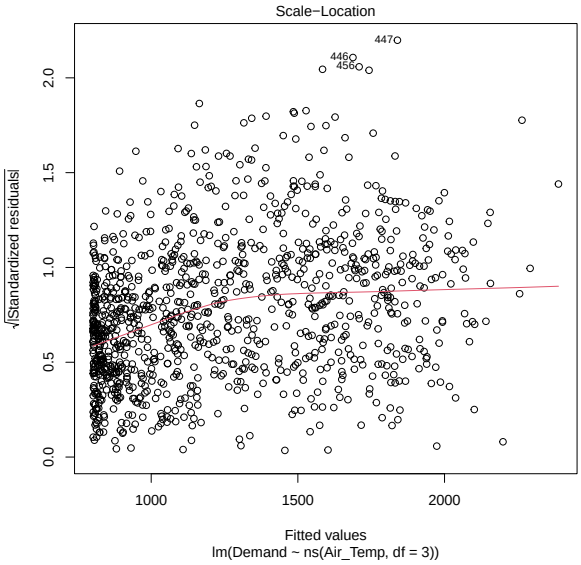
3.2 Natural Cubic Splines

Considering that we still need to model a non-linear trend, but our previous model violated the normality assumption, exploring non-parametric methods is a logical next step. Natural cubic splines provide the flexibility to capture non-linear relationships while offering user-defined customization, such as selecting the number of knots, and ensuring a smooth fit. For simplicity and to minimize the risk of overfitting, we chose a spline with three degrees of freedom, consistent with the approach used in polynomial regression. While the spline model addresses the issue of non-linearity and normality, the problem of heteroscedasticity remains unresolved.

To tackle heteroscedasticity, we assigned weights to the model by taking the inverse of the square of the predictor values, Air Temperature. The intuition behind this approach is that if the variance increases as the predictor grows, assigning these weights reduces the influence of predictor values where variance is higher. Where the functions scaled by each coefficient represent the spline basis functions, the resulting weighted natural cubic spline model takes the form:

$$\hat{Y}_i = \beta_0 + \beta_1 N_1(X_i) + \beta_2 N_2(X_i) + \beta_3 N_3(X_i)$$

Next, we validated the assumptions of this model. Since normality is no longer a concern, our focus shifts to homoscedasticity and independence. Examining the scale-location plot for the weighted spline model, we observe significant improvements, with the fitted line appearing flatter and the pattern largely eliminated. To confirm this improvement, we performed the Breusch-Pagan hypothesis test, which yielded a p-value of approximately 0.99. This statistically insignificant result indicates that the homoscedasticity assumption is satisfied.

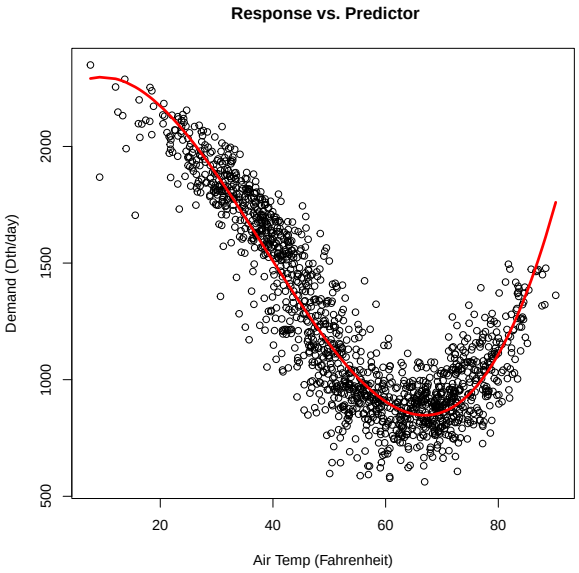


Finally, we evaluated the independence assumption using the Durbin-Watson test. After calculating the weighted residuals by multiplying the residuals of the model by the square root of the weights, the test still reveals the presence of autocorrelation. This result was expected, as the methods employed in this analysis do not explicitly address autocorrelation. Addressing this issue will require further refinement in future iterations.

4 Results

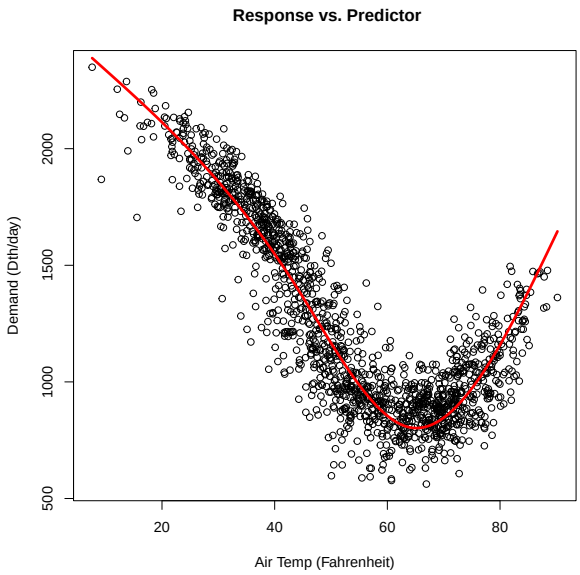
4.1 Polynomial Regression

A summary of our polynomial regression model indicated that all coefficients are statistically significant with p-values near zero. The model also achieves an outstanding in-sample adjusted R^2 of 85.75 percent. To evaluate its performance on unseen data, we applied the model to the test set, which yielded an out-of-sample R^2 of 88.88 percent. Interestingly, the out-of-sample performance is roughly 3 percent higher than its in-sample performance. This suggests the model generalizes well to new data and is not overfitting. Below, we observe how the fitted curve aligns with the scatterplot of the data. It does a great job of modelling the non-linear trend, as further evidenced by the strong R^2 metrics.



4.2 Natural Cubic Splines

A summary of our natural cubic spline regression model showed that all coefficients are statistically significant as well with p-values near zero. The model also achieves an in-sample adjusted R^2 of 90.69 percent which is better than that of our polynomial regression by 5 percent. In an effort to evaluate its performance on unseen data, we applied the model to the test set, which yielded an out-of-sample R^2 of 89.15 percent. This value is within 2 percent of its in-sample counterpart. This likely suggests that the model generalizes even better to new data and is not overfitting than our previous model. Below, we observe how the fitted curve aligns with the scatterplot of the data. Like the polynomial regression, it does a great job of modelling the non-linear trend, as further evidenced by the very strong R^2 metrics.



5 Conclusion

Our analysis has revealed several key takeaways. Despite our efforts to meet each of the model's assumptions, we were unable to resolve the issue of independence. This limitation is likely due to the time-series nature of the data, where autocorrelation is inherently present. On the other hand, we successfully addressed heteroscedasticity by introducing weights into the spline model, demonstrating the effectiveness of weighted regression in handling variance-related issues.

That said, while the inability to fully satisfy all assumptions limits either model's suitability for inference, it does not preclude its use for prediction. Given their high R^2 metrics and the close alignment between in-sample and out-of-sample performance, I posit that either model is still a strong tool for forecasting demand using Air Temperature.

Looking ahead, it would be prudent to explore time-series models, such as ARIMA, which are specifically designed to handle autocorrelation. Additionally, incorporating other relevant predictors, such as Dew Point Temperature, which showed relatively high importance, or lagged demand features, might further improve modelling performance and address issues with autocorrelation.