ENERGY DEMAND PREDICTION

- Chaithanya Krishna Vadlamudi
- Robert Reiter
- Bharath Reddy Bora
- Adam Ronald Nygaard

Introduction :

In 2021, the state of Texas experienced long-term blackouts as a result of an unexpected blizzard attacking the deregulated energy economy's weak point. This blackout was due to a lack of preventative maintenance preceding this blizzard, but having a weather-based energy demand could also be used as a signal to perform this preventative maintenance before anything occurs. However, severe weather events are not the only factor that affects utility demand. Utility demand is almost always directly proportional to the difference in temperature from around 20℃. Of course, there are other factors that can impact this relationship, which we plan to research in depth.

On the flip side, overproduction raises costs for everyone since the energy is lost because there is really no mainstream infrastructure to support utility-scale grid energy storage. Unfortunately, this puts utility companies in a difficult situation because nobody wants to waste energy, but nobody wants to pay more than they absolutely need to. Additionally, blackouts cause unnecessary wear and tear on utility systems, of course requiring costly repairs right away to bring electricity back. Whether not enough or too much energy is produced, utilities operate with a certain level of excess, which is wasted. As mentioned previously, the end goal is to develop widespread, utility-scale energy storage, but until then we could always use Artificial Intelligence and Data Science to help increase utility efficiencies. Our proposed solution would be to generate a model that would accurately determine the expected utility demand at any given moment based on weather.

Energy Production Case Study:

The issue of energy demand that exceeds energy production has been on the rise over the past decade and is expected to worsen. The demand for energy is highly volatile and depends heavily on current seasons, temperatures, and weather conditions. Spikes in energy demand can cause rolling blackouts to occur, most frequently experienced in the summer months. According to a study done by the North American Electric Reliability Corporation, rolling blackouts will be a rising problem in the coming summers. Specifically, NAERC estimates that over two thirds of the country will experience blackouts in the near summers [1]. These regions are mostly concentrated in the West and Midwest, with the risk of blackouts being highest in the northern Midwest region. This can be visualized in figure 1 below, which is a map of the blackout warnings from the study.

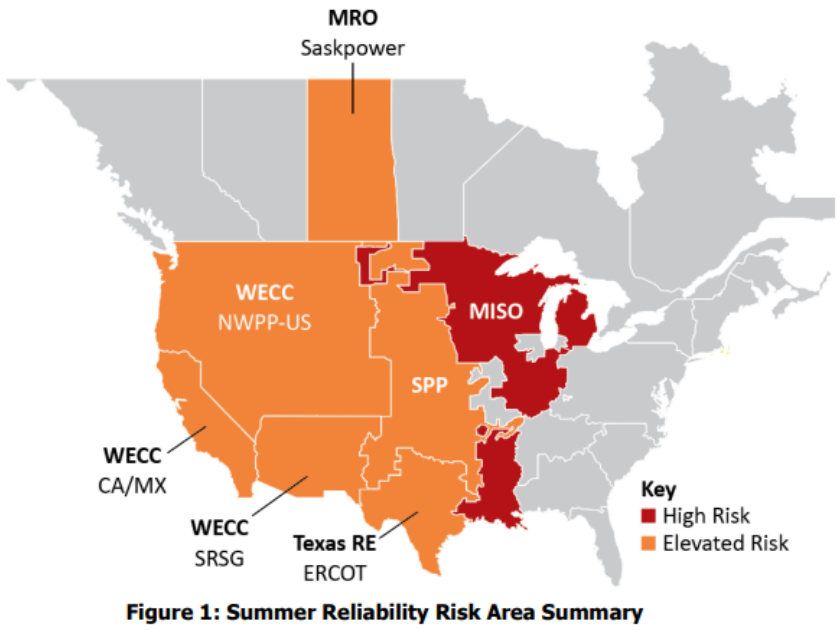

Figure 1: High and elevated risks of rolling blackouts in North America

This study is important to the estimation of power production for multiple reasons. First, this case study proves that the issue of energy supply and demand is relevant and needs addressing. Secondly, the production of energy needs to be precise due to the lack of energy storage technology. Without a means to store the produced energy, it is either used or wasted. This is also the issue when facing blackouts. Because there is no means to effectively store energy at a regional scale, spikes in energy demand during the warm summer months will be too much for the power plants to handle. The energy demand in these cases will be higher than the energy supply, and large swaths of residential homes will be cut from power.

By gaining an understanding of the effects that weather has on energy demand, utility companies can anticipate when to produce more electricity. Temperature and other weather factors play a large role when determining the amount of energy that will be consumed daily. With weather forecasting becoming increasingly accurate, utility companies can apply machine learning algorithms, such as the one developed in this project, to better predict the necessary energy production and thus mitigate any potential blackouts.

Methodology:

We considered implementing two approaches:

1. Regression analysis
2. Bayesian analysis.

Bayesian analysis is used to establish causality and to determine how changes in one feature affect other features. Additionally, rather than just giving us a single predicted number, it gives us the confidence interval.

EDA and feature analysis:

To check for null values, we performed data analysis after downloading the dataset from Kaggle. There are no null values in the data. Even though the data contains some outliers, we considered leaving them in place as they are true outliers.

Column	Non-Null Count	Dtype
Location	21045 non-null	object
Date	21045 non-null	int64
Time	21045 non-null	int64
Latitude	21045 non-null	float64
Longitude	21045 non-null	float64
Altitude	21045 non-null	int64
YRMODAHRMI	21045 non-null	float64
Month	21045 non-null	int64
Hour	21045 non-null	int64
Season	21045 non-null	object
Humidity	21045 non-null	float64
AmbientTemp	21045 non-null	float64
PolyPwr	21045 non-null	float64
Wind.Speed	21045 non-null	int64
Visibility	21045 non-null	float64
Pressure	21045 non-null	float64
Cloud.Ceiling	21045 non-null	int64

Figure 2: Data analysis for null values

Figure 2 shows that there are 17 columns with PolyPwr as the target variable.
The seasons and the location are both discrete variables. One hot encoding was employed by us for the analysis. Since time and date are cyclic variables in the data set. Time and month were converted into cyclic characteristics using cyclic encoding. We used the heat map depicted in figure 3 to identify the features that have a strong correlation to the target variable. Additionally, pair plots have been utilized to clarify the relationship between the variables. We initially intended to use normal linear regression to gain an understanding of the fundamental mean square error and r^2 score, which are our performance measures for our implementable models. To ensure that our validation was accurate, we intended to utilize a 0.3 test size.

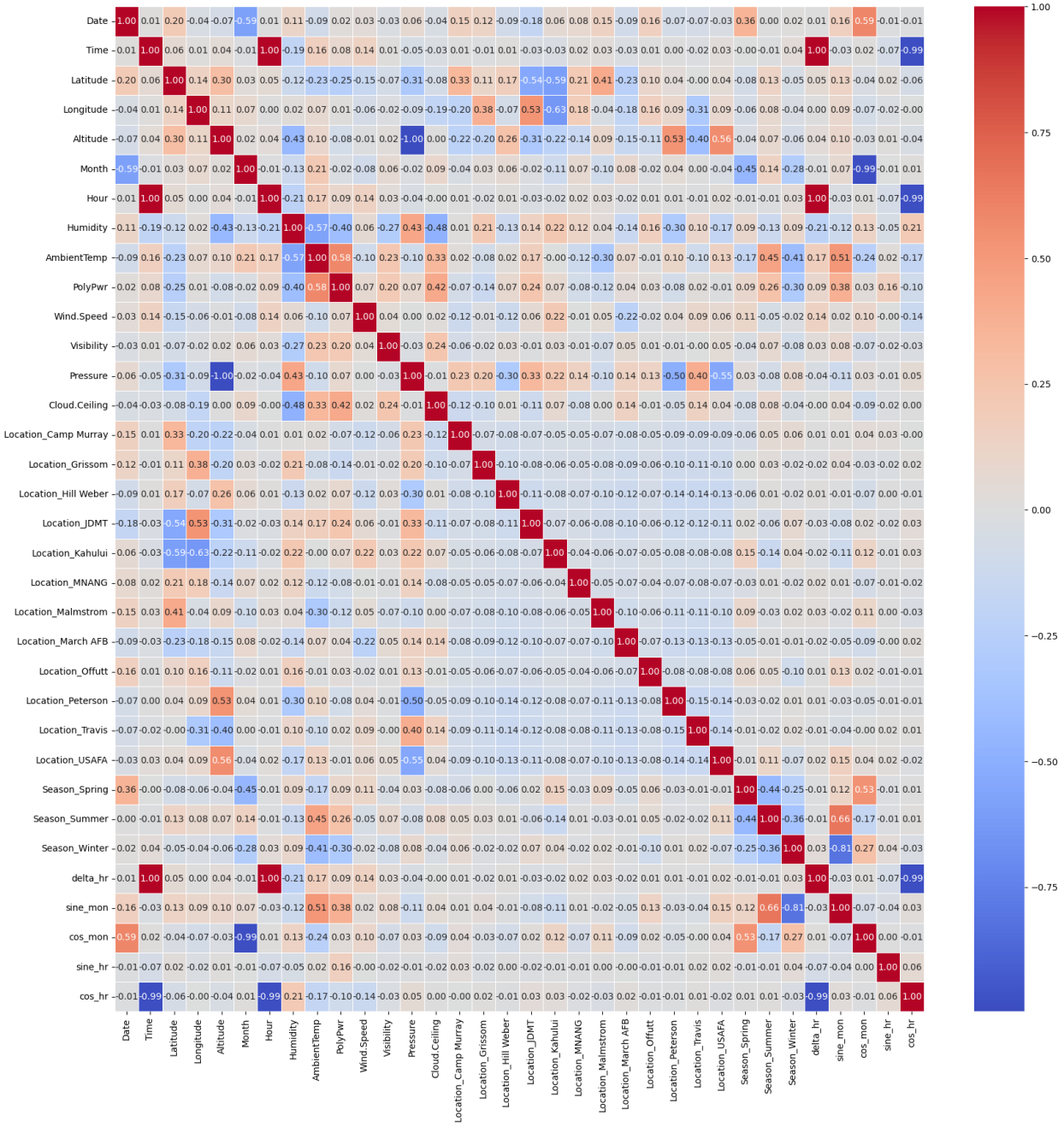

Figure 3: Heat map

Following a heat map analysis. Location, Cloud Ceiling, Latitude, Humidity, Ambient temperature, Season, and Visibility were considered as the necessary features for the additional analysis.

We considered ensemble techniques to raise the performance score. Three ensemble algorithms have been used.

1. Random Forest Regressor
2. LGBM regressor
3. XGBR regressor

Random Forest Regression analysis:

It is one of the ensemble algorithms that employs decision trees, executes each one of them in parallel, and calculates the mean of the results. For the random forest regression analysis, we employed Quantile transformer since we considered the outliers to be true outliers. Compared to linear regression, the mean square error has been reduced by 13%. The R^2 score increased from 0.568 to 0.6245.

LGBM regression analysis:

LGBM is another instance of an ensemble method that does leaf-wise processing. It is an excellent boosting algorithm that accelerates execution while using the least amount of processing power. Grid Search CV was used to fine-tune the parameters. With a cross validation of three and the subsequent search space, the hyper tuning is carried out.

```
search_space = {"num_leaves": [7,61,500],
"n_estimators”: [100,200,500],
"max_depth" : [3,6,9],
"learning_rate" :[0.0001,0.01,0.1,1],
"objective": [‘rmse','mae','mape'],
"feature_fraction":[0.5,1]}
```

With the help of the study discussed above, we were able to estimate the ideal parameters for the LGBM regression analysis, which are displayed below. Table 1, illustrated later, displays the mean square error and R^2 score.

```
GS2 = LGBMRegressor (
objective='rmse',
num_leaves=900,
n_estimators=1400,
max_depth=11,
learning_rate=0.008,
feature_fraction=0.6,
random_state=42)
```

XGBR regression analysis:

Another boosting approach is XGBR regression, which uses level-wise tree growth rather than leaf-wise tree growth. Due to the limited computing resources, we used the Grid Search CV with a cross validation of 2 to extract the best parameters. The following settings were used for the hyper tuning.

```
parameters = {
‘objective’:['reg:linear','reg:squarederror','reg:absoluteerror'],
‘learning_rate':[0.0001,0.01,0.1,1],
‘max_depth':[3,6,9,11,13],
‘min_child_weight':[4],
‘silent':[1],
‘subsample':[0.7],
‘colsample_bytree':[0.7],
'n_estimators':[500,1000,2000]}
```

By selecting the best parameters, shown below, and the evaluation metrics findings in Table 1 we have continued our regression study.

```
GSx2 = XGBRegressor(
colsample_bytree=0.7,
learning_rate=0.01,
max_depth=9,
n_estimators=1000,
objective='reg: linear')
```

Bayesian Analysis:

As ML regression models cannot provide causal analysis. We considered performing Bayesian analysis on the same data, using the same training and testing datasets, with the same features as we did for the regression analysis. We have used two discretization techniques:

1. EWD(Equal width discretization): Each column is separated into a certain number of intervals that are all the same size. The number of intervals in each column does not always have to be equal.
2. EFD(Equal frequency discretization): Each column has an established number of intervals of the same size and a fixed number of intervals per column.

Three Bayesian models were constructed with the above two discretization techniques:

Naïve Bayes model:

Assuming that every feature is conditionally independent of each other. We used Netica to implement our Bayesian models.

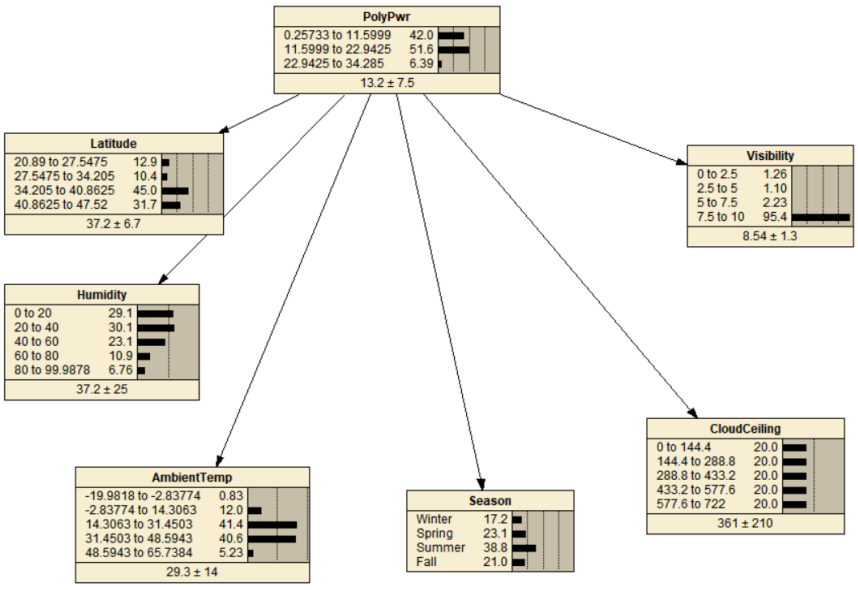

Figure 4: EWD Naïve Bayes

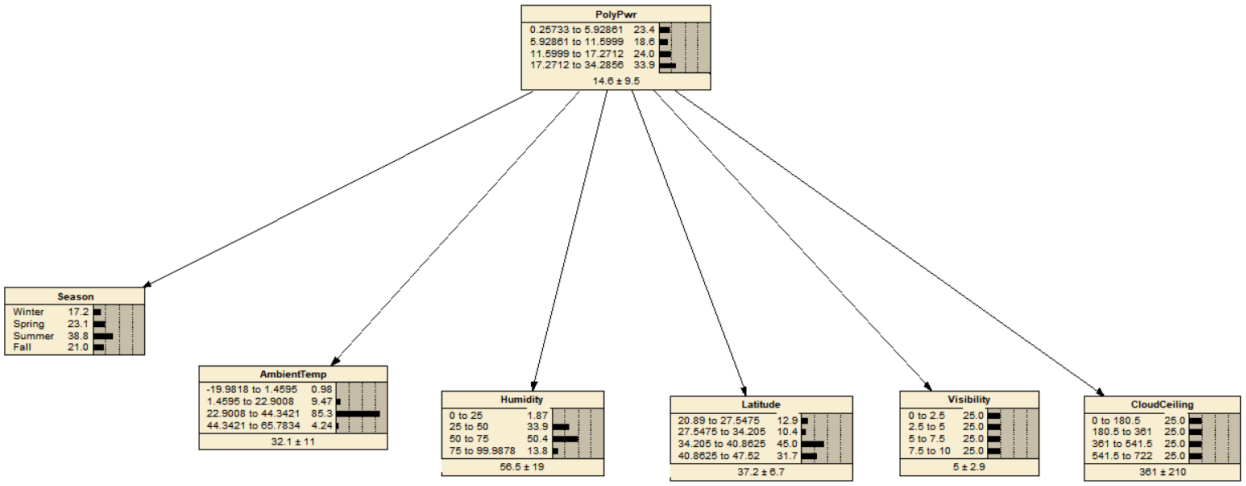

Figure 5: EFD Naïve Bayes

The bias of the model rises as the number of intervals increases. The interval size and number of intervals are implemented as shown in figure 8 in order to obtain the best training and testing accuracy.

TAN model:

Assuming that the features are not conditionally independent. We have constructed TAN model for both the EWD and EFD.

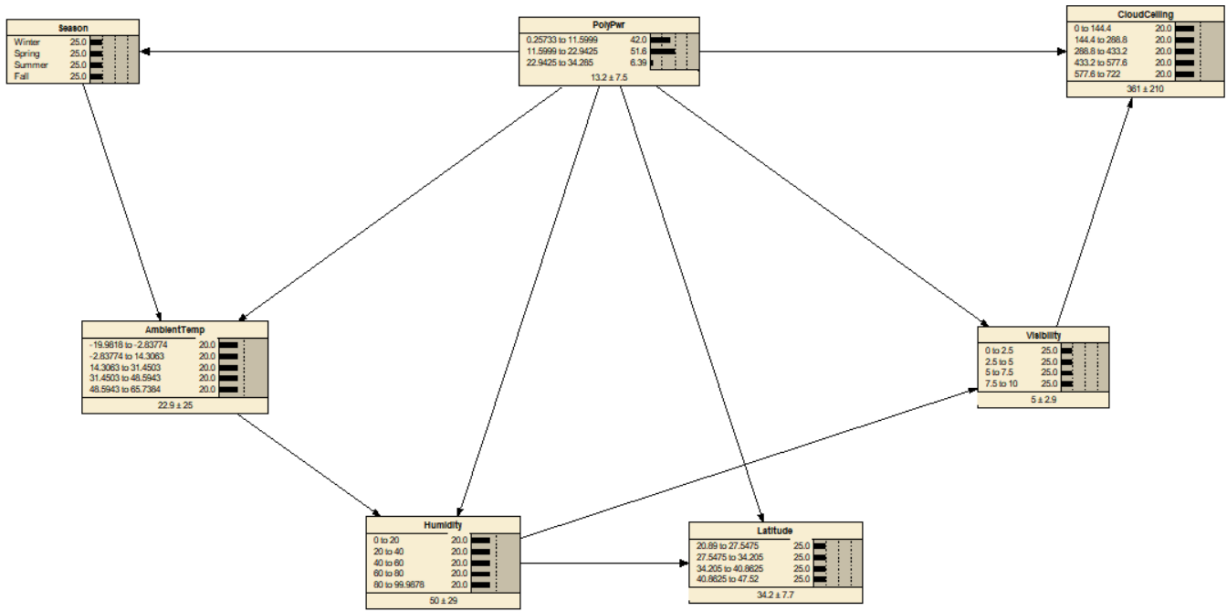

Figure 6: EWD TAN model

The same Network model was implemented for both EWD and EFD techniques and the results are obtained.

Net Bayesian:

Assuming the prior knowledge causal relationship between the features, we constructed net Bayesian. Both the procedures and the model were put into practice and examined. Figure 7 displays the net Bayesian's Bayesian analysis. The same Network model was implemented for both EWD and EFD techniques and the results are obtained.

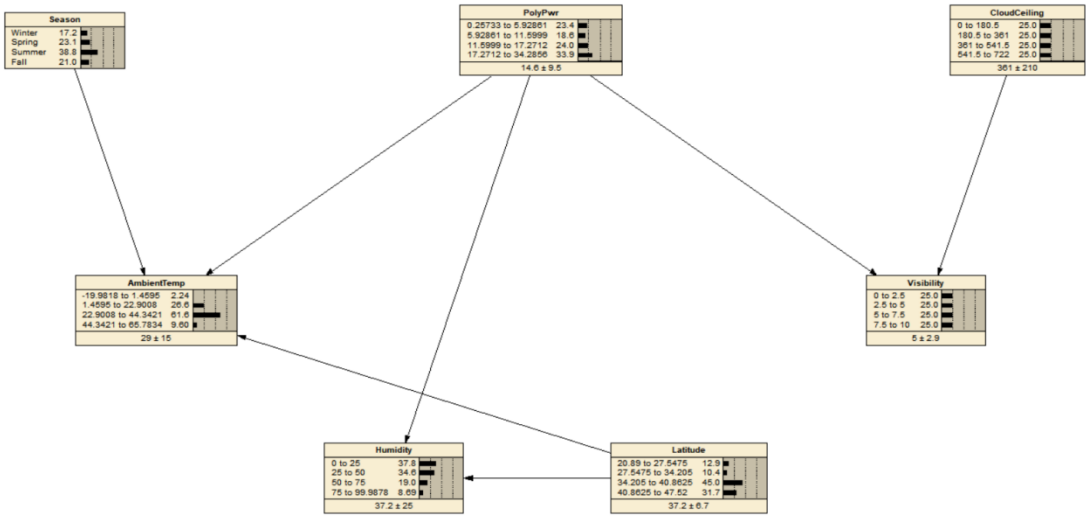

Figure 7: EWD Net Bayesian model

Expected output:

The following results were obtained for the Bayesian models implemented:

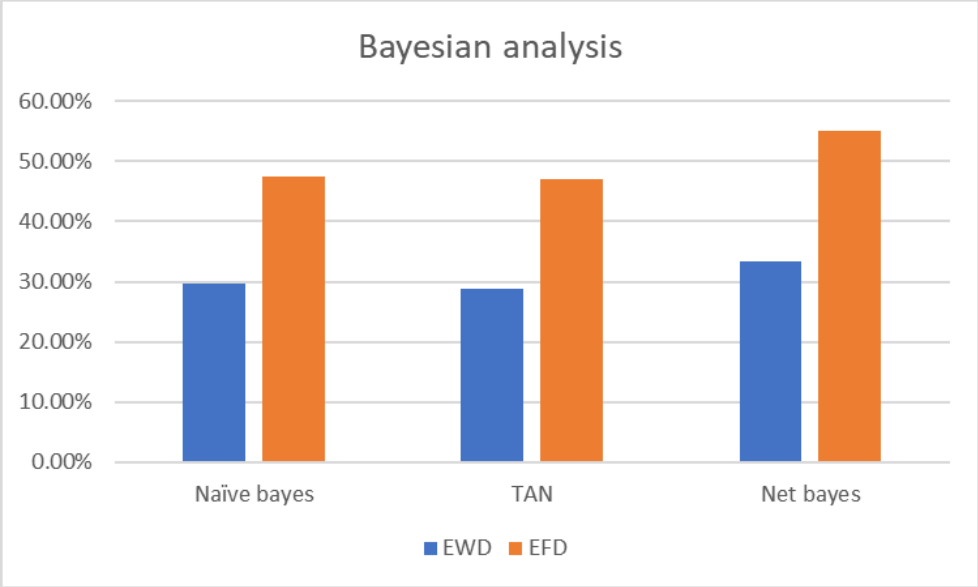

Figure 8: Bayesian Analysis results

From figure 8, it is observed that the TAN model with EWD discretization technique has the lowest error rate when compared to the remaining models.

Model	**Mean Squared Error**	**R2 Score**
Linear	21.51	0.569
Random Forest	18.6347	0.6272743
LGBM regressor	17.30431	0.6272743
XGB Regressor	17.3166	0.6536

Table 1: Regression Analysis results

Comparing the XGBR regressor to the other regression models, the table demonstrates that it is the most suitable regression model. Since we are using confidence intervals, our evaluation metrics are Mean squared error and R^2 score.

Many other models that have been developed so far have used a test size of 0.2. For improved validation, we used a test size of 0.3. Additionally, we have increased the R^2 score using the XGB regressor, which is superior to the previous regression models used up till now.

The causal relationship between other features was not considered by the other models that have been used up to this point. Therefore, we have put these Bayesian models into practice, which will give the confidence interval of the output and analysis of how other features are impacted by changes in observed feature.

We have also built a simple User Interface model using Gradio. This model can provide us the predictions according to real time data.

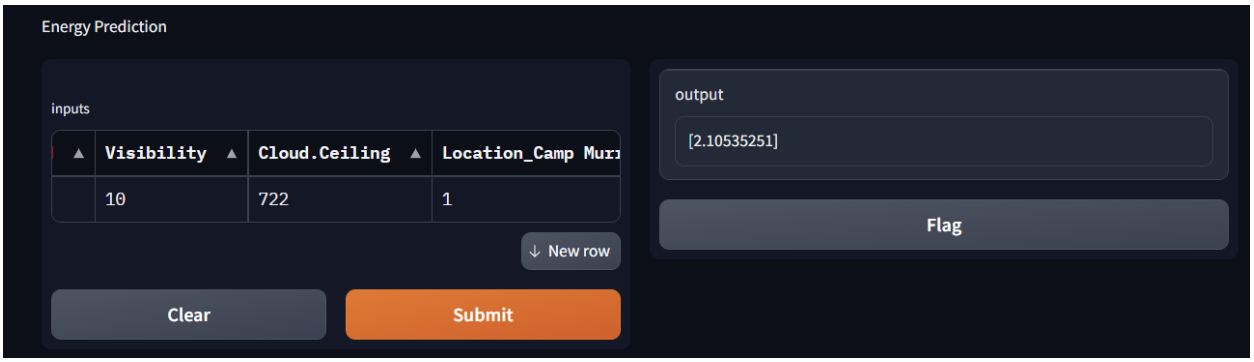

Figure 9: Gradio Interface model

Issues Faced:

When working on this project, the main issue that was faced was the error rate of the Bayesian network. Variables were chosen that had the highest correlation with the target variable PolyPwr, but these variables were unable to produce a Bayesian network with low error rates. Different discretization methods were used in order to lower the error rate, but this still proved to be only a partial improvement.

The first discretization method that was implemented was the equal frequency distribution. This method divides the datasets for each variable into groups with equal occurrences. The error rate for this method of discretization was found to be at or near 50% for all three Bayesian network models, which is too large of an error rate to be an accurate model. The Bayesian network was reconstructed to try to lower the error rate, but this had little effect on the error rate. Often, interconnecting the nodes of the Bayesian network made the error rate increase.

The second discretization method proved to lower the error rate, but not enough to be considered an accurate network. This method is known as the equal width method, which takes the lowest value of a dataset and the highest value of a dataset and then splits the data into classes with equal ranges. For example, splitting a data set of the integers 1-9 into three classes using equal width distribution would result in classes for 1-3, 4-6, and 7-9. Each class has the same range, although the number of instances within each class can vary. This method proved to lower the error rate to 29% at best when used with the tree-augmented naive Bayesian network. The other networks experienced higher error rates at 30% and 33%, but this is still lower than the equal frequency discretization.

The most likely way to decrease the error rate would be to experiment with different discretization methods and classes. In this project, two different methods were used, but many more discretization techniques exist beyond equal frequency and equal width. By implementing varying methods of discretization and changing the class sizes, this would most likely drive the error rate down to a percentage that would be acceptable when considering the Bayesian network as accurate.

Conclusion:

Significance and potential impact of the project:

The project's significance lies in its potential to address the growing issue of energy supply and demand, paving the way for more sustainable and efficient energy production and consumption practices. Energy demand is very unpredictable and significantly influenced by the time of year, temperature, and weather. Rolling blackouts are more frequently observed throughout the summer when energy consumption spikes. Machine learning methods developed for the research can be used to forecast the amount of energy needed to avoid blackouts. The foundation for utilities to take a proactive rather than a reactive approach is laid by their capacity to supply several metrics and produce accurate energy consumption predictions.

The XGB Regressor model's ability to handle large datasets, balance accuracy, and speed up the training process is significant for energy production and consumption prediction. Making knowledgeable choices about energy supply and demand also depends on the TAN Bayesian network's capacity to recognize the most important aspects that influence energy production and consumption and offer insight into their relationships.

It is important for this project to have a user interface (UI) since it gives potential users a simple way to access the model. This UI makes it simpler to utilize and apply the model in real-world circumstances by allowing users to interact with it, enter data, and receive predictions and recommendations.

Overall, the project will have a major effect on utilities' capacity to predict energy demand, avoid blackouts, and minimize overproduction. Utility firms can better forecast the required energy production and prevent any potential blackouts by understanding how weather affects energy demand. The project offers a comprehensive response to the expanding supply and demand for energy.

Future directions for the project:

There are several future directions that can be considered to make the energy production project more effective and improve its overall performance. These include:

Expand the dataset to larger regions for utility company applicability: The project presently focuses on a particular location, however expanding the dataset to include data from other regions would make the project more applicable to utility businesses operating in other places. This would increase the dataset's usability for utility firms. In order to make decisions and allocate resources, this may also help identify regional trends in energy demand and production. It may also be possible to find new avenues for the delivery and production of renewable energy by incorporating data from other places.

1. Modify target variables for different energy sectors (solar): The project might be altered to concentrate on particular energy sectors, such solar energy, by changing the target variables. In order to take into consideration the particular elements that have a bearing on solar energy production and consumption, the target variables would have to be modified. The project can offer more in-depth insights into the variables that affect energy production and consumption for a certain sector by focusing on those industries specifically. Identifying possible areas for improvement and directing policy choices pertaining to the usage of renewable energy sources may be helped by this.
2. Further develop Bayesian Network and ML models for increased accuracy: The Bayesian Network and Machine Learning models used to forecast energy consumption and output could be improved upon for the project. This can entail experimenting with various methods, improving the models with more information, or applying novel strategies to boost accuracy. improved energy demand and production forecasting would be possible with increased precision, which might result in more effective resource allocation and improved planning for energy infrastructure.
3. Test different discretization methods to better analyze the dataset: The project could benefit from evaluating various discretization techniques to enhance the precision of the Bayesian Network models, as mentioned in the "Issues Faced" section. This can entail testing many other widely used discretization techniques or creating original techniques just for this project. The project may determine the best discretization technique for the dataset by testing various approaches, which would increase the precision of the Bayesian Network models' predictions.

Overall, implementing these future objectives into action could contribute to making the energy production project more reliable, precise, and adaptable to various geographies and energy industries. The study could offer useful insights for decision-makers in the energy sector and aid in guiding policy decisions on renewable energy by increasing the accuracy of the models used to anticipate energy demand and supply.

Resources:

[1] “2022 Summer Reliability Assessment,” 2022 SRA Draft, https://www.nerc.com/pa/RAPA/ra/Reliability%20Assessments%20DL/NERC_SRA_2020.pdf (accessed May 8, 2023).
[2]. “Random Forest Regression by Chaya” https://levelup.gitconnected.com/random-forest-regression-209c0f354c84
[3]” LGBM regressor” https://wiki.datrics.ai/lgbm-regressor.
[4] “Dataset from Kaggle” https://www.kaggle.com/datasets/saurabhshahane/northern-hemisphere-horizontal-photovoltaic
[5] “A comparative study of discretization methods for Naïve bayes classifiers” http://i.giwebb.com/wp-content/papercite-data/pdf/yangwebb02a.pdf
[6] “ Analysis of Solar Power using machine learning techniques” https://towardsdatascience.com/predicting-solar-power-output-using-machine-learning-techniques-56e7959acb1f
[7] “ Bayesian Network for predicting energy consumption in schools Florianópolis -Brazil” https://www.researchgate.net/publication/335988231_Bayesian_Network_for_Predicting_Energy_Consumption_in_Schools_in_Florianopolis_-Brazil -