Springboard Capstone Project Report: Wind Power Forecast

Are windmills an effective source of power? Do they generate enough power to offset the cost and loss of land required to install and maintain the windmills? How sustainable are the windmills themselves? In the controversial energy debates, renewable energy sources are highly sought after. As a starting point to assess the wind energy avenue, this project forecasts the Active Power generated by one windmill. My goal for this project was to forecast the active power one windmill would generate over the next 15 days.

As the project unfolded, I utilized the following python packages:

- numpy
- pandas
    - ProfileReport from pandas_profiling
- matplotlib.pyplot
- seaborn
- display from IPython.display
- os
- datetime
    - datetime
    - timedelta
- sklearn.preprocessing
    - StandardScaler
    - MinMaxScaler
- statsmodels.api
    - api
    - seasonal_decompose from tsa.seasonal
    - Adfuller from tsa.stattools
    - Graphics.tsaplots
        - Plot_acf
        - Plot_pacf
    - SARIMAX from statespace.sarimax
- Auto_arima from pmdarima
- Sklearn.metrics
    - Mean_absolute_error
    - Mean_squared_error
    - Mean_absolute_percentage_error
- Sqrt from math
- Itertools
- XGBRegressor from xgboost
- Sklearn.model_selection
    - GridSearchCV
    - Train_test_split
- Tensorflow.keras
    - Sequential from models
    - LSTM from layers
    - Dense from layers
    - Dropout from layers

Data Sourcing and Cleaning:

Data for this project was sourced from Kaggle. The data set is comprised of 19 features measured every 10 minutes for one windmill over two years. The featured measures included both external features such as time, wind speed, wind direction, and temperature, as well as features internal to the windmill such as rotor RPMs, blade positions, and turbine status. The data tracked both active and reactive power. This study focuses only on the active power. The original data set can be found here.

Initial inspection revealed that an unnamed column was the Timestamp data, and was renamed as such. Also, one feature was labeled as the name designating the windmill and did not contain actual values. This feature was dropped.

The description generated for the data set revealed that the Timestamp feature had no missing values, as to be expected. It also revealed that ControlBoxTemperature had a mean, min, and max value of 0. Most features, except for Timestamp, Active Power, Ambient Temperature, ReactivePower, and Windspeed were roughly 50% or more null values. Several of the first rows were null. Negative values were found in the three BladePitchAngle features as well as the ActivePower and ReactivePower features. The initial null percentages are displayed to the right.

Timestamp	0.0
ReactivePower	20.0
WindSpeed	20.0
ActivePower	20.0
AmbientTemperatue	21.0
NacellePosition	39.0
WindDirection	39.0
BearingShaftTemperature	47.0
TurbineStatus	47.0
RotorRPM	47.0
MainBoxTemperature	47.0
HubTemperature	47.0
GeneratorWinding2Temperature	47.0
GearboxOilTemperature	47.0
GearboxBearingTemperature	47.0
GeneratorWinding1Temperature	47.0
GeneratorRPM	47.0
Blade1PitchAngle	64.0
Blade3PitchAngle	65.0
Blade2PitchAngle	65.0
dtype: float64	

To deal with null values, non-essential features that were more than 50% null values were dropped. Numerous missing values were still evident. As I explored the correlation of missing values, it appeared that the windmill was recording feature measurements before the windmill was active. Instead of imputing missing values from feature averages or other imputation methods, the initial section of data comprised of missing values was deleted.

EDA:

A Profile Report is run to more extensively explore the data. The report gives 20 alerts regarding features with high cardinality or high correlation, data distribution, and unique values. The report also gives a summary of each feature, and visualizations, as well as notes features that are unique, common, rare, most frequent, and least frequent. This report indicates that the Timestamp feature is comprised of unique values. However, given the nature of time-based data, this is expected and acceptable. The report also indicated that Wind speed, generator RPM, rotor RPM, and generator winding temperature are highly correlated with the active power target feature. Any other area of interest or concern for this data set is explained by the nature of wind and the resulting windmill status and function.

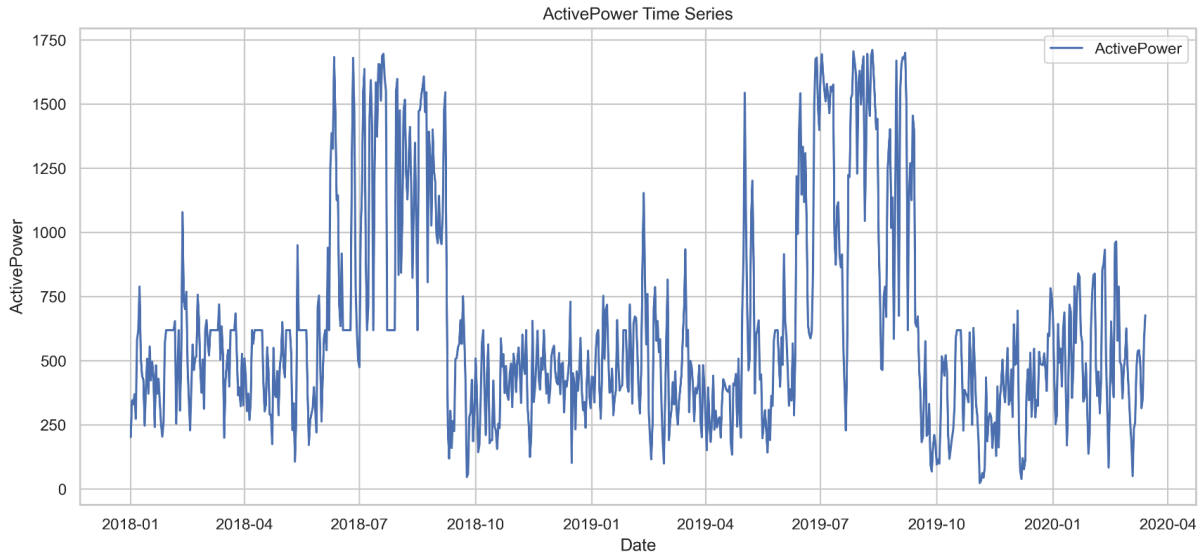
Later in the project, it became evident that due to computational requirements, the data set needed to be simplified. So, all features were dropped except for the following highly correlated features: Timestamp, ActivePower, WindSpeed, GeneratorRPM, RotorRPM, Generator1WindingTemperature.

Pre-Processing and Training:

Before building forecasting models, the Timestamp feature was converted to DateTime format and set as the index. The data was also aggregated to provide daily averages rather than exact measurements recorded every 10 minutes.

Other than the train/test splits that would be needed for building models, I wanted to reserve the last 15 days of the original model to compare to the data forecasted by the models. This was accomplished by sorting the Timestamp data, and splicing the last 15 days into its own data set. The last 15 days' data was saved as its own data set for comparison to the forecast generated at the end of the project.

The target feature ActivePower was visualized (displayed below) to gain a basic understanding of seasonality and trends.



An Augmented Dicky-Fuller test generated a p-value of 0.0011216483197323374, far less than the significance level of 0.05. It was therefore determined that the data was stationary. Considering that the data had both stationarity and seasonality, I decided to build a SARIMAX model. The ability to consider multivariate data was the deciding factor in choosing this model over the univariate SARIMA model.

Generating ACF and PACF plots suggested an autocorrelation lag, which may indicate a parameter p=1 benefiting the SARIMAX model. The seasonal decomposition appeared to indicate that the seasonal term was yearly. However, when I used auto_arrima, each iteration of attempting longer period parameters resulted in multi-hour computational time requirements, caused the program to crash, or yielded a lack of available memory error message. A period of 100 days was the closest I was reasonably able to utilize in the auto_arima parameter identification. Considering how many iterations this program searches, I chose to use the parameters it recommended rather than the potential parameters suggested by the PACF plot.

Modeling and Visualization:

Based on the nature of wind and windmill functions, as well as initial visualizations revealing seasonality in the data, SARIMAX, XGBoost with GridSearchCV parameter tuning, and LSTM modeling were selected. The models were evaluated by MAE, RMSE, and MAPE metrics.

For clarity purposes, note that unless otherwise stated, in the rest of this report, “useable data” will refer to the remaining original data after I spliced the final 15 days to compare to the forecasted active power.

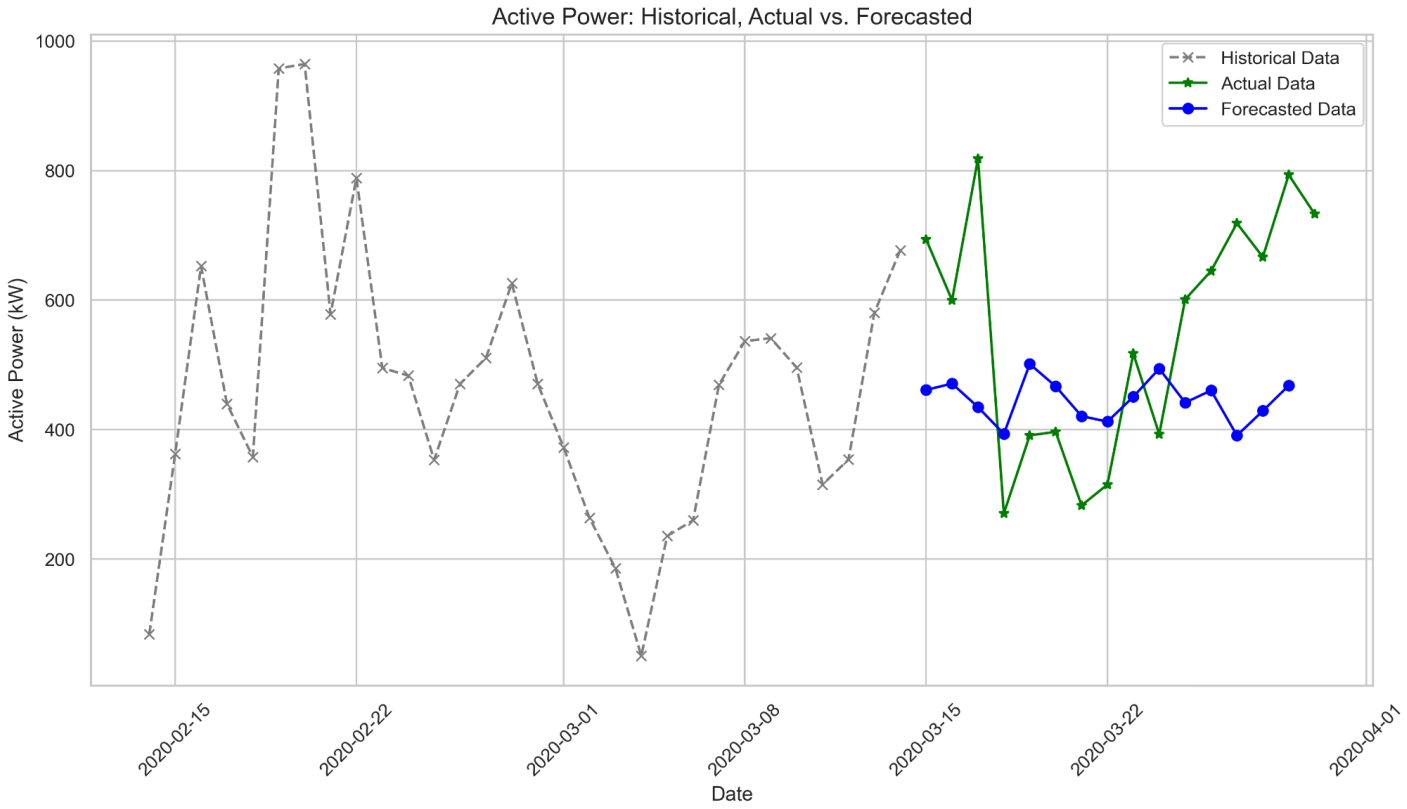
The train/test split I used for the SARIMAX model mirrored the original splicing of the last 15 days. Auto_arrima was used on a training data set comprised of all but the last 15 days of the useable data. The SARIMAX was trained using the resulting (2,0,2)(0,0,0)[100] parameters from auto_arrima. The SARIMAX model was fit on the remaining final 15 days of the useable data. This model was evaluated on its MAE score of 31.023569376895356, RMSE score of 43.030384497261345, and MAPE score of 0.12028836230760867%.

The XGBoost model used a train/test split that reserved the last 30 days of usable data for the test set. GridSearchCV was used for parameter selection. The recommended best parameters were 'colsample_bytree': 1.0, 'learning_rate': 0.1, 'max_depth': 5, 'subsample': 1.0. The XGBoost model was evaluated on its MAE score of 31.24954161423522, RMSE score of 40.67154788747206, and MAPE score of 0.09430550877546205%.

For the LSTM model, a function was defined to create sequences using a train/test split similar to the XGBoost model. The LSTM model was fit with the parameters epochs=50, batch_size=32, validation_split=0.1, verbose=1, and was evaluated on its MAE score of 166.12701042975024, RMSE score of 197.89707493603598, and MAPE score of 48.517090848142445%. This model may have run the fastest of all three, but it had the worst performance metrics. If continuing this project, I would learn more about LSTM parameter tuning.

The XGBoost model with GridSearchCV() proved to be the best model with the lowest RMSE and MAPE values, and the second-lowest MAE of all three models. The SARIMAX model produced the next best metrics and may have outperformed the XGBoost model with additional parameter tuning or increased computational abilities available.

After comparing model performance metrics and selecting XGBoost, the mean and standard deviation of supporting features during the last 60 days were utilized to generate 15 days' worth of new supporting feature data for model training. This data in the XGBoost model forecasted 15 data points for the active power feature. As a visual comparison, I created a line graph displaying the last 30 days of historical data (before the separated final 15 days), the predicted data, and the actual data reserved in the pre-processing stage. The line graph can be viewed on the next page. Adding the respective data points revealed that the XGBoost model forecasted a total of 6696.6279296875 units of active power, while the actual data totaled units of data in the final 15 days.



Next Steps:

Advocates for wind energy programs claim that implementing wind power for your home or business will reduce your electricity bills by large percentages. If customers requested similar data logs from the windmill company and used the XGBoost model developed in this program, they could then project how much power the suggested windmill would generate in the future. Similarly, the model could be used on historic customer energy usage data to predict how much energy the customers will need over the same period. This, along with analysis of windmill installation and maintenance cost and sustainability will better equip potential wind energy customers in their energy sourcing decisions.