Capstone Project
Wind Power Forecast

By Lindsey Caughlin, Special Thanks to Mentor: Kenneth Gil-Pasquel

Problem: How much wind will one windmill generate over 15 days?

- 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?
- Often, conversations around renewable energy are heated and seep into strongly held beliefs and opinions.
- This project creates a tool to make data available to inform discussions and decisions.

Information Benefactors:

- Renewable energy enthusiasts
- Policy makers and lobbyist
- Energy companies
- Land owners
- Business owners
- Home owners

Data Sourcing

Data for this project was sourced from a Kaggle dataset recording measurements every 10 minutes over 2 years for a windmill.

Null data presented a significant issue in this data set, as evidenced by the null percentage. Features that were more than 50% null were dropped from the data set.

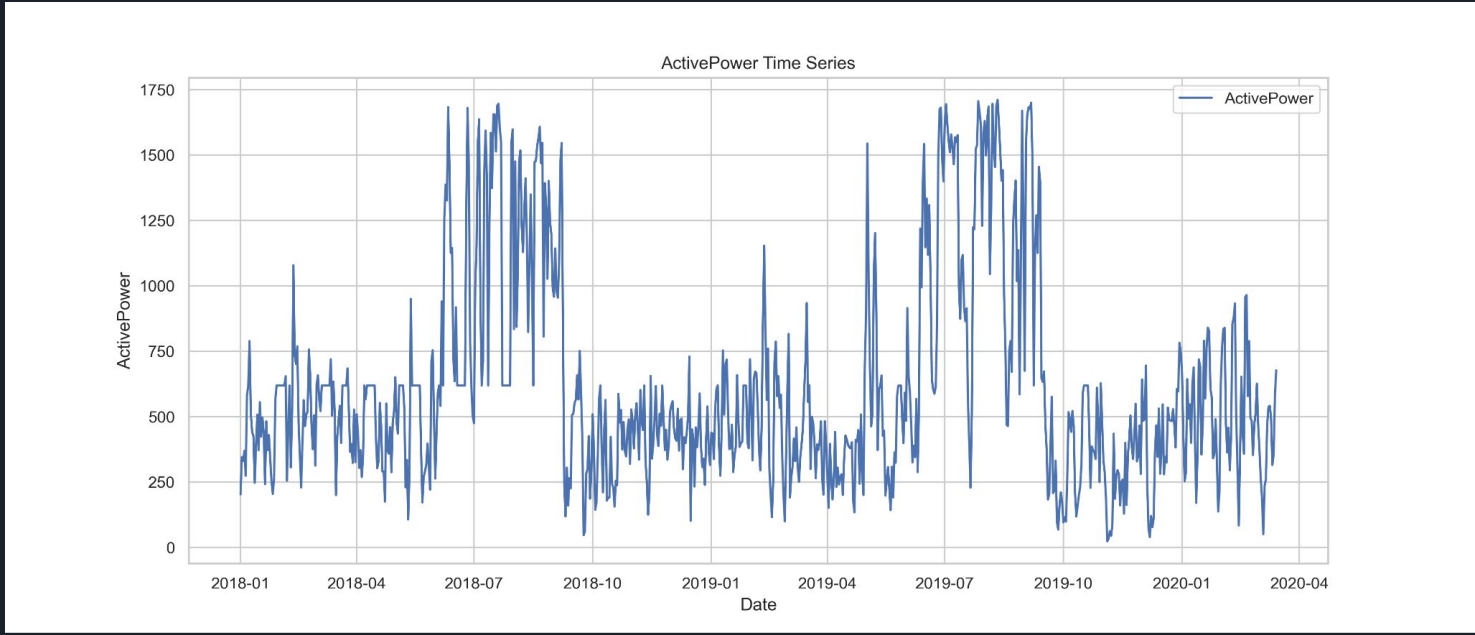
Apparently, the recording device was active for one day prior to the windmill being activated. Dropping the corresponding 144 initial entries resolved this problem, leaving no null values in the data set.

Null Value Percentages

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

Data Exploration



Given the seasonality of the target feature and the importance of the supporting multivariate data, a SARIMAX model was the first model built for forecasting future Active Power quantities.

XGBoost and LSTM models were also built as options to compare and select the best performing model.

Data Pre-Processing

It became evident that the data complexity would need to be reduced in order to accomodate computational requirements.

All features except for the following highly correlated features were dropped:

- Timestamp
- Active Power
- Wind Speed
- Generator RPM
- Rotor RPM
- Generator Winding 1 Temperature

Data was resampled to report daily averages rather than 10 minute measurements.

SARIMAX, XGBoost, and LSTM Modeling

The final 15 days of data were separated from the original data set to compare against the best performing model's forecast.

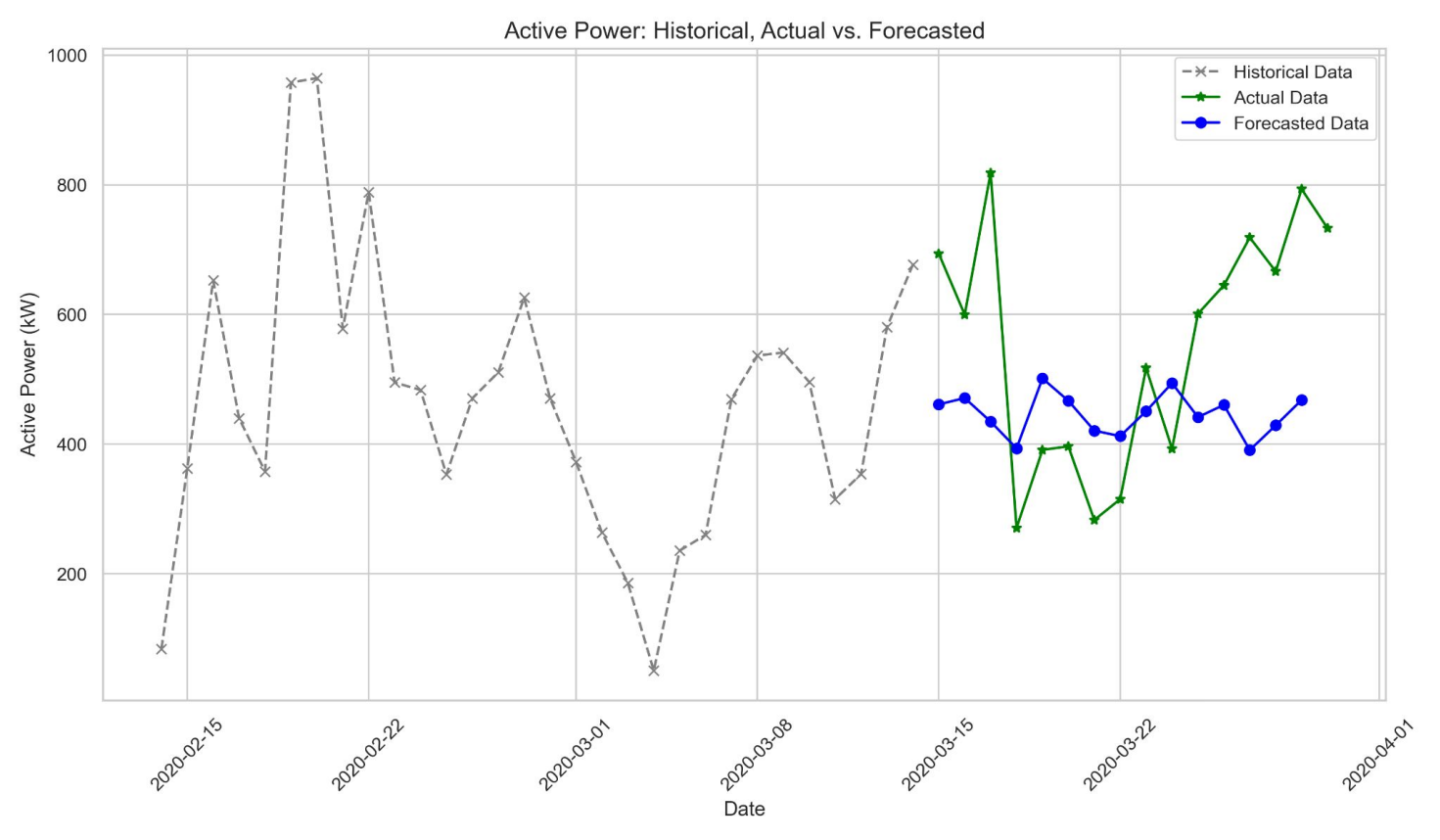
Models were trained on the majority of the remaining data, reserving 15 days (SARIMAX) or 30 days (XGBoost and LSTM) for model testing.

Models were assessed based on their MAE, RMSE, and MAPE scores. XGBoost was selected for final forecasting because it had the best 2 out of 3 metrics.

Model Metrics

	Model	MAE	RMSE	MAPE
0	SARIMAX	31.023569	43.030384	0.12028836230760867%
1	XGBoosting	31.249542	40.671548	0.09430550877546205%
2	LSTM	166.127010	197.897075	48.517090848142445%

Active Power Forecast (15 Days)



Total Forecasted Active Power: 6696.6279296875 units

Total Actual Active Power: 8836.047383400117 units

Next Steps
{Research & Optimization}

- Improve model parameters
  - Test SARIMAX on 30 days instead of 15 days
  - Test alternate parameters for XGBoost and LSTM
- Forecast based of weather forecasts instead of generated data
- Research average power usage
- Compare Active Power against Reactive Power

Next Steps
{Future Projects}

- Create user interface for various users to input their custom data
- Market to program to energy producing companies
- Market to educational programs teaching about renewable energy sources