Calgary Solar Power and Energy Output Estimator

Problem Statement

The goal of this project is to generate a data-based solar power and energy estimating web app. By using time, date, and weather data, I utilized machine learning techniques to build a regression model that can predict the amount of solar radiance being in Calgary. The model can also predict the long term (monthly) solar radiation received in Calgary by learning the historical data using time series models. Potential and current residential solar energy customers could benefit from this project by gaining more unbiased information about what they can expect by installing solar panels.

Background

Located in the province of Alberta, Calgary is the sunniest city in the country in all respects. It receives 2396 hours of bright sunshine on average every year and experiences about 333 sunny days annually. 52% of the daylight hours in the city are sunny (Osborn, n.d.). This makes Calgary a big potential solar market. The goal of my project is to create a data-based solar energy and power predictor/estimator for the Calgary area using time and weather. The current data I can obtain is based on a system called TMY. TMY is Typical Meteorological Year data. To determine TMY data, various meteorological measurements are made at hourly intervals over several years to build up a picture of the local climate. A simple average of the yearly data underestimates the amount of variability, so the month that is most representative of the location is selected (peveducation.org, n.d.). However, with the climate change we are experiencing these years, I found it hard to determine what “typical” is.

Data Source

Data was requested from the National Solar Radiation Database (National Solar Radiation Database, n.d.). It provides serially complete collections of meteorological and solar irradiance data sets for the United States and a growing list of international locations. The data is publicly available at no cost to the user. Many academic research papers have been published based on the data provided by them. In this dataset, the Global Horizontal Irradiance (GHI) is our target for solar power and energy estimation.

Data Processing

The data set was clean, with no missing data or missing data points. One minor problem that occurred during the initial EDA is that since the data was recorded yearly over 23 years, some

data was recorded in a different format from different years. To keep the uniform formality of the data, all the features are transformed into the format most of the data was recorded in.

Exploratory Data Analysis and Data Transform

This dataset only contains power data. To make an energy output prediction, time needed to be incorporated with the power data to generate energy data.
Figure 1 shows how the monthly mean differs from the annual mean, as expected. If we look at the monthly data over the years in figure 2, however, we see dips in radiation in June for many years. After cross-checking with some weather data, this is most likely caused by the heavier rainfall in June for some years (after cross-checking with historical weather data in Calgary).

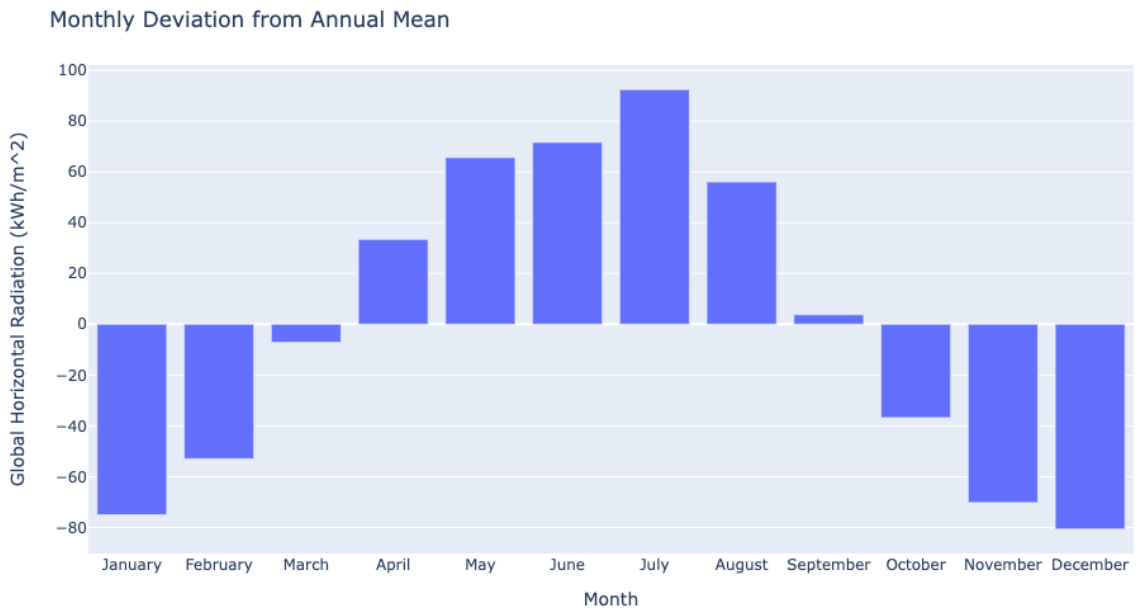

Figure 1. Monthly deviation from annual mean

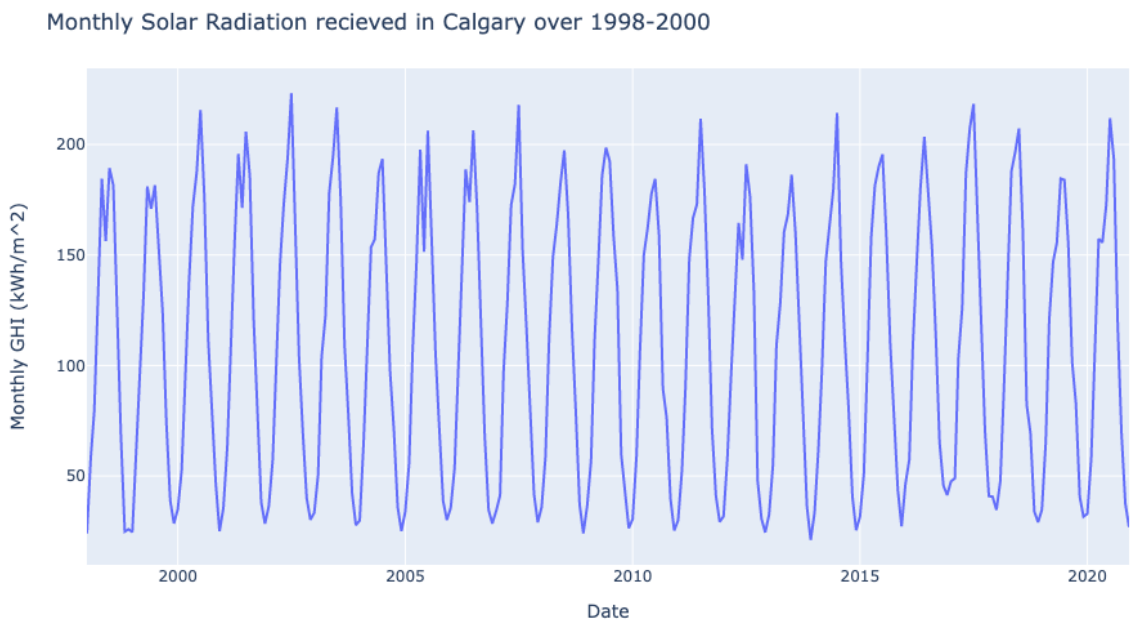

Figure 2. Monthly solar radiation received in Calgary

I also did a basic correlation check between all the features and GHI. Figure 3 shows that solar zenith angle has the strongest relationship with GHI, followed by relative humidity and temperature.

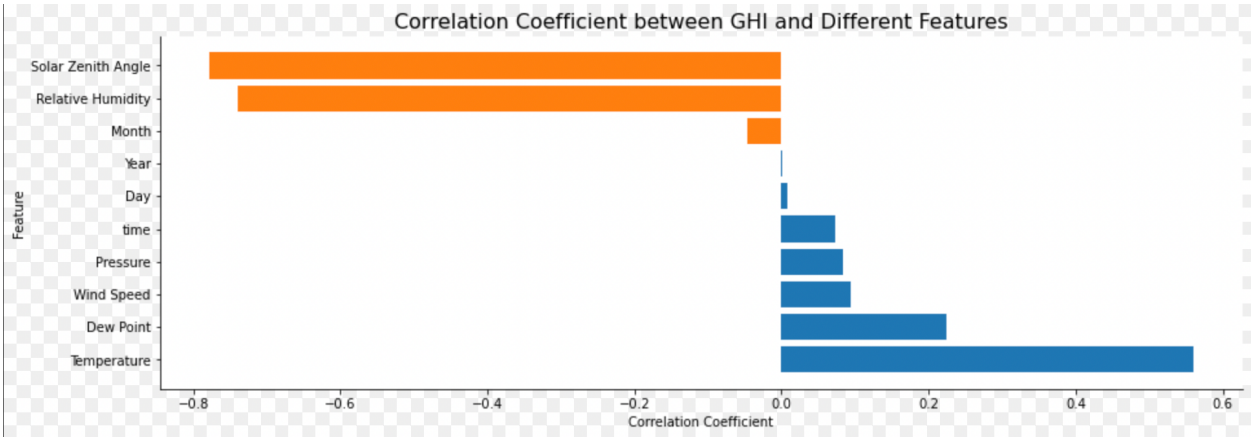

Figure 3. Correlation coefficient between GHI and different features

Time Series and Regression Modeling

Time series modeling was performed for energy estimation, RMSE in combination with R^2 value were compared to determine the best model. Table 1 shows us the comparison of each model.

Rank	Model	Train R^2	Test R^2	Train RMSE	Test RMSE
1	Sarima (0, 0, 0) (2, 1, 2, 12) n	0.968	0.961	10.97	11.7
2	Sarima (0, 0, 0) (0, 1, 1, 12)	0.967	0.959	11.16	11.99
3	Facebook Prophet	0.974	0.959	9.82	12.02

Table 1. Performance of different time series models

Ridge Regression, Lasso Regression, Random Forest Regression, and XGBoost Regression were performed on the power estimation. Grid search, random search, and cross-validation were used to select the best combination of features and model hyperparameters. I also used MAE and RMSE to evaluate each model, combined with R^2 value. Table 2 shows us the comparison of the performance from different types of models.

Rank	Model	MAE	RMSE	R^2
1	Random Forest	22.31	52.73	0.9479
2	XGBoot	22.31	53.11	0.9472
3	Ridge	83.61	12008.49	0.7751
4	Lasso	84.27	12202.2	0.7715

Table 2. Performance of different regression models

The choice of the energy estimation model to be implemented in the app was easy to make, I simply picked the best performing one. However, I can not use the same approach to choose the power prediction model because of the model size. The size of the Random Forest model is over 16GB, which will crash the app, therefore, the second-best model (by a margin) was chosen to be implemented in the app. One thing to note here file size for XGBoost is only 314MB.

Findings

The figure below shows the predicted energy output in comparison to the actual energy output for 2019 and 2020. We can see that our model did capture the extreme cases that can happen in June and balanced it between the years with heavy rainfall and without.

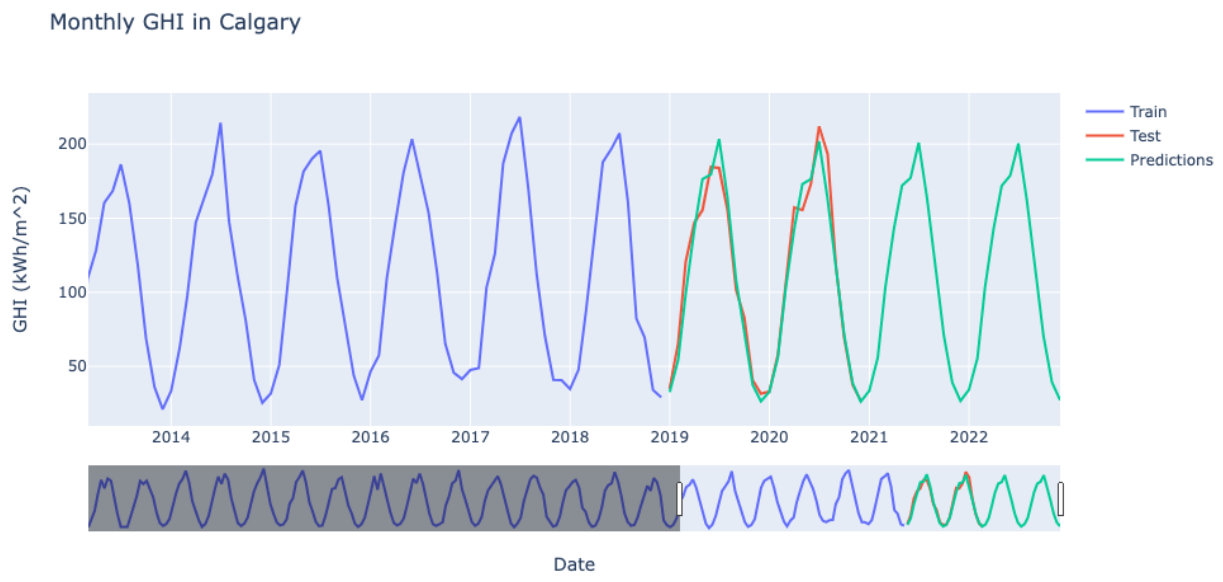

Figure 4. Comparison of train, test, and predicted data

For Power estimation, there is no “best” way to find out how important a feature is to predict the target for a complicated model like Random Forest or XGboost. Therefore 2 methods were used to give a better understanding of the model. They calculated the importance of each feature in different ways, and each has its advantages and disadvantages. If we look at the plot in combination with figure 3, which shows us how strong of a relationship each feature has with GHI, I suggest that the solar zenith angle is the most important feature for the XGBoost model.

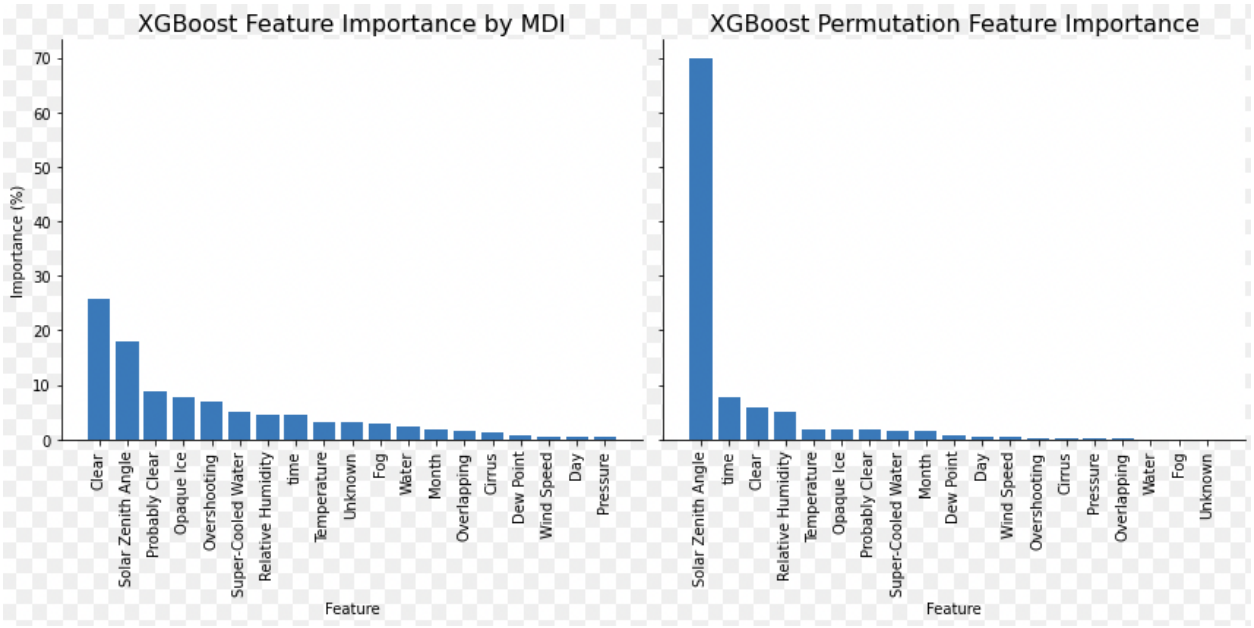

Figure 5. Feature importance comparison of XGBoost in different ways of calculation

Conclusion

In this study, I trained different models to estimate the solar energy and power output for Calgary and pick the most suitable model for the web app. I have also suggested the feature that has the biggest influence on GHI. A web app is built based on the best performing timeseries model (energy estimation) and the XGBoost model (power estimation). It can help the customers on the solar market to navigate their decision making on whether to install solar panels.

Future work

As a next step, I would like to gather more related data to combine with what I currently have to build a more accurate model. I also want to explore what neutral network can bring for time series and regression modeling in this case. I can easily expand our project horizontally to predict many locations. In the current app, the panel output using GHI is based on a simple calculation, I would like to find more sophisticated ways to relate them, which could lead to another interesting machine learning project.

Works Cited

(n.d.). Retrieved from peveducation.org: https://www.pveducation.org/pvcdrom/properties-of-sunlight/typical-meteorological-year-data-tmy

(n.d.). Retrieved from National Solar Radiation Database: https://nsrdb.nrel.gov/

(n.d.).

Osborn, L. (n.d.). Retrieved from Current Result: https://www.currentresults.com/Weather-Extremes/Canada/sunniest-cities.php