HI-SEAS Solar Irradiance Prediction

https://www.kaggle.com/dronio/SolarEnergy

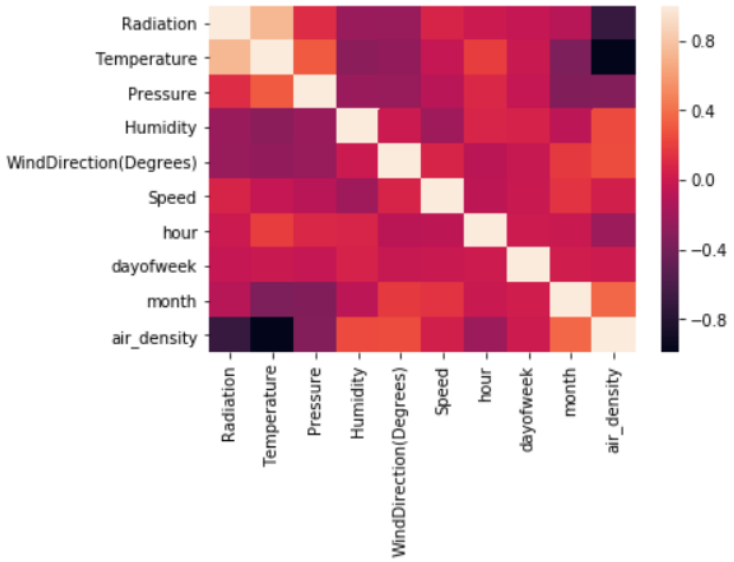
The following dataset can be found at the kaggle link above. Solar radiation data are collected at Moscow for over a span of a month. The data includes Date, Time, Radiation, Temperature, Pressure, Humidity, Wind Direction, Speed, Sun Rise, and Sun Set time.

	UNIXTime	Data	Time	Radiation	Temperature	Pressure	Humidity	WindDirection(Degrees)	Speed	TimeSunRise	TimeSunSet
0	1475229326	9/29/2016 12:00:00 AM	23:55:26	1.21	48	30.46	59	177.39	5.62	06:13:00	18:13:00
1	1475229023	9/29/2016 12:00:00 AM	23:50:23	1.21	48	30.46	58	176.78	3.37	06:13:00	18:13:00
2	1475228726	9/29/2016 12:00:00 AM	23:45:26	1.23	48	30.46	57	158.75	3.37	06:13:00	18:13:00
3	1475228421	9/29/2016 12:00:00 AM	23:40:21	1.21	48	30.46	60	137.71	3.37	06:13:00	18:13:00
4	1475228124	9/29/2016 12:00:00 AM	23:35:24	1.17	48	30.46	62	104.95	5.62	06:13:00	18:13:00

The premise of this report is to build a model and predict the radiation presumably taken at ground level. This report will explore the dataset, build additional features, and train a model to predict level of radiation at any given time in the city.

Data Exploration

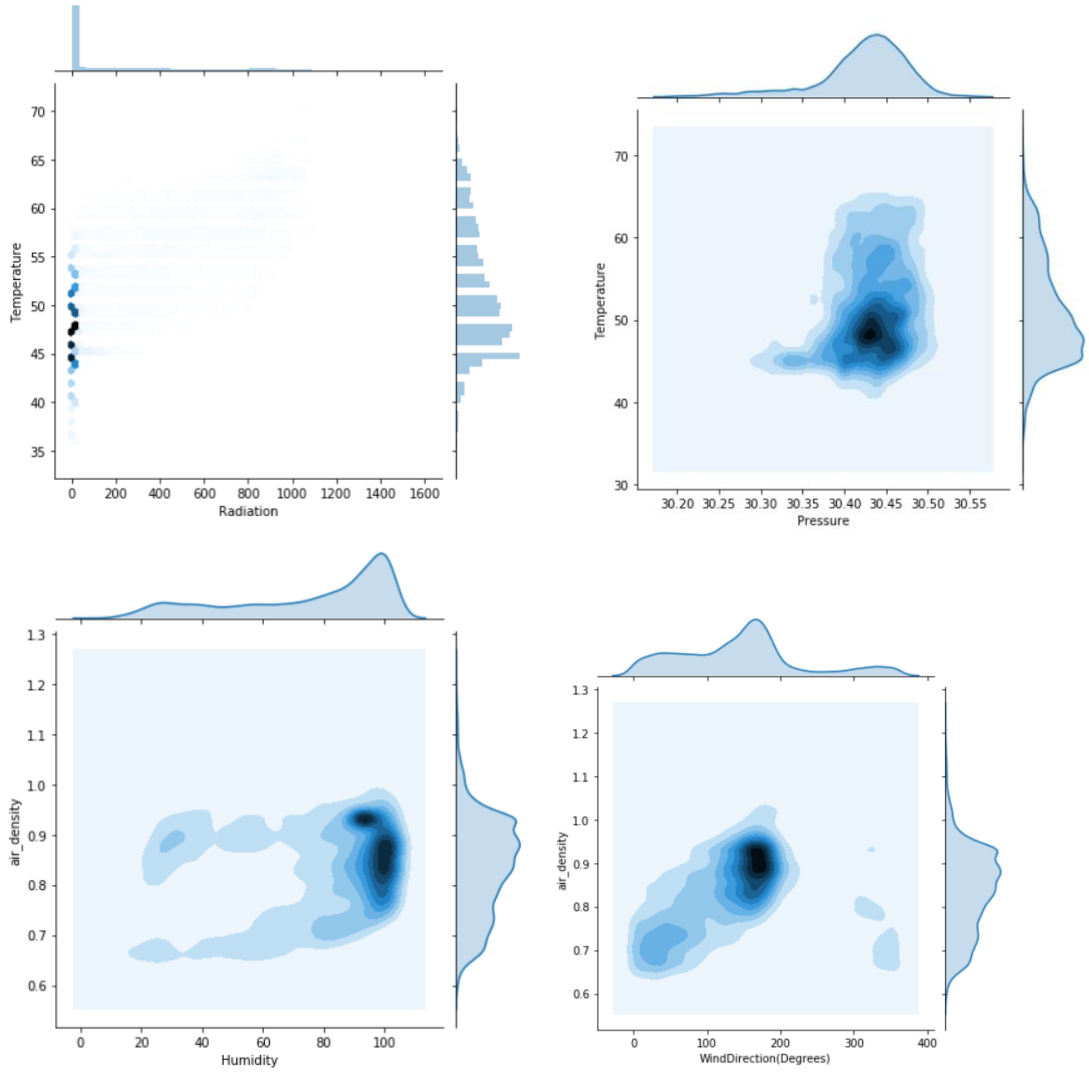
Before data exploration an additional feature must be added and that is air density. This can be calculated through PVNRT equation. This is important due to the radiation's attenuation of matter and now dense it is. When travelling several miles it is important to note even if it is though air. Thus air density is added in as a feature. Next the correlation of the data is shown below.

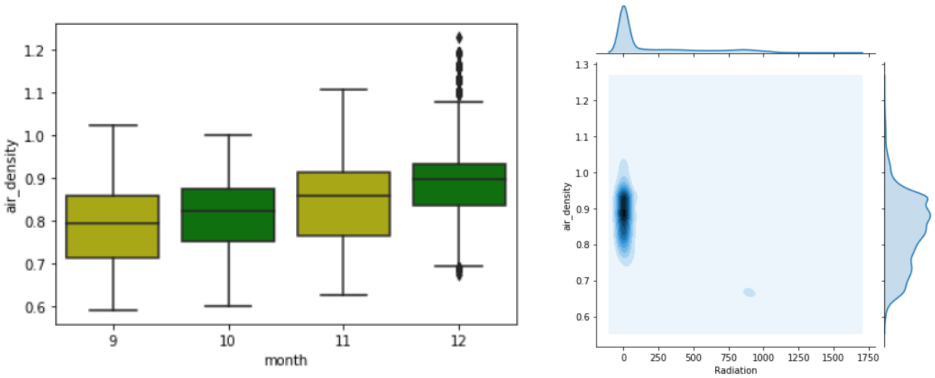


	Radiation	Temperature	Pressure	Humidity	WindDirection(Degrees)	Speed	hour	dayofweek	month	air_density
Radiation	1.000000	0.734955	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Temperature	0.734955	1.000000	0.311173	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Pressure	NaN	0.311173	1.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Humidity	NaN	NaN	NaN	1.00000	NaN	NaN	NaN	NaN	NaN	0.246860
WindDirection(Degrees)	NaN	NaN	NaN	NaN	1.000000	NaN	NaN	NaN	NaN	0.256429
Speed	NaN	NaN	NaN	NaN	NaN	1.0	NaN	NaN	NaN	NaN
hour	NaN	NaN	NaN	NaN	NaN	NaN	1.0	NaN	NaN	NaN
dayofweek	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.0	NaN	NaN
month	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.000000	0.368221
air_density	NaN	NaN	NaN	0.24686	0.256429	NaN	NaN	NaN	0.368221	1.000000

For correlations bigger than 0.2.

From the correlation plot there is high correlation between temperature and radiation. Note that is only for linear correlation, and this is expected since radiation does heat up the surrounding environment. For the rest of the features there are a certain level of correlation but not as high as the former. Next a closer look at the correlation features.





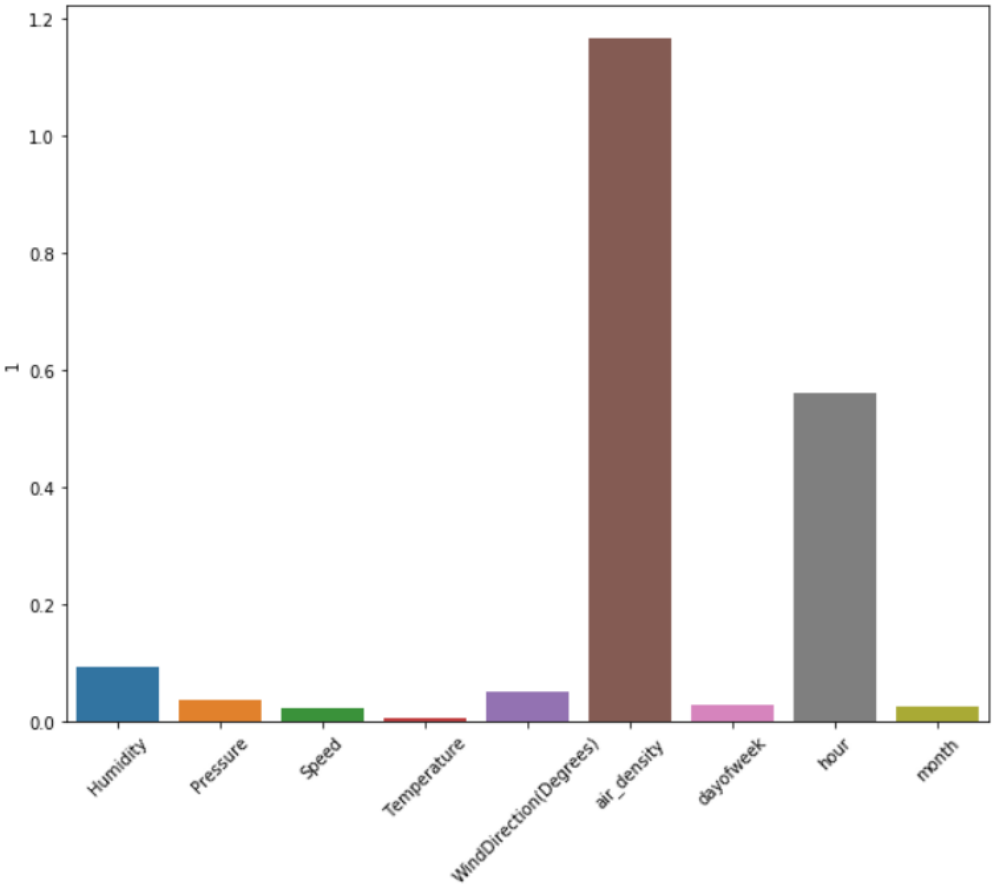
The plot above shows the correlation between Temperature and Radiation. There doesn't seem to be that much correlation between the two from the plot and most of the radiation level is at the lower end regardless of the temperature. The plot is also similar to air density and radiation. Next, is Pressure and Temperature, there appears to be a concentrated area where pressure and temperature coincide. The same goes for Humidity and Wind Direction with air density. Lastly there appears to be a relation between density and the month as expected since air gets denser as it gets colder. From these plots there indicates that temperature and radiation does not have any real relation and its linearity is very steep and should not be taken at face value for how much radiation is absorbed.

Model Training and Prediction

The data is already cleaned which leaves selecting the right features for modeling. The follow features are use in the dataset: 'Radiation', 'Temperature', 'Pressure', 'Humidity', 'WindDirection(Degrees)', 'Speed', 'hour', 'dayofweek', 'month', 'air_density'. Sun set and rise time is taken out since it follows the radiation pattern and can be counted as redundant data. With these data 3 models are used, these are XGB, Random Forest, and Linear Regression.

From initial fit Linear regression gave a R2 score of 0.638 and MAE of 134, while XGB got a R2 score of 0.91 and MAE of 8892. For Random Forest the R2 score is 0.9286 and MAE of 7271. From these scores RF is definitely the model to use and be applied for feature selection.

Before fitting the data into a model for prediction feature selection is performed. For the current case Random Forest is used for feature selection. For feature selection a 8 split Kfold is used and fitted to the model to find the feature importance score. That score is tallied up from each fold. The plot below shows the tallied score.



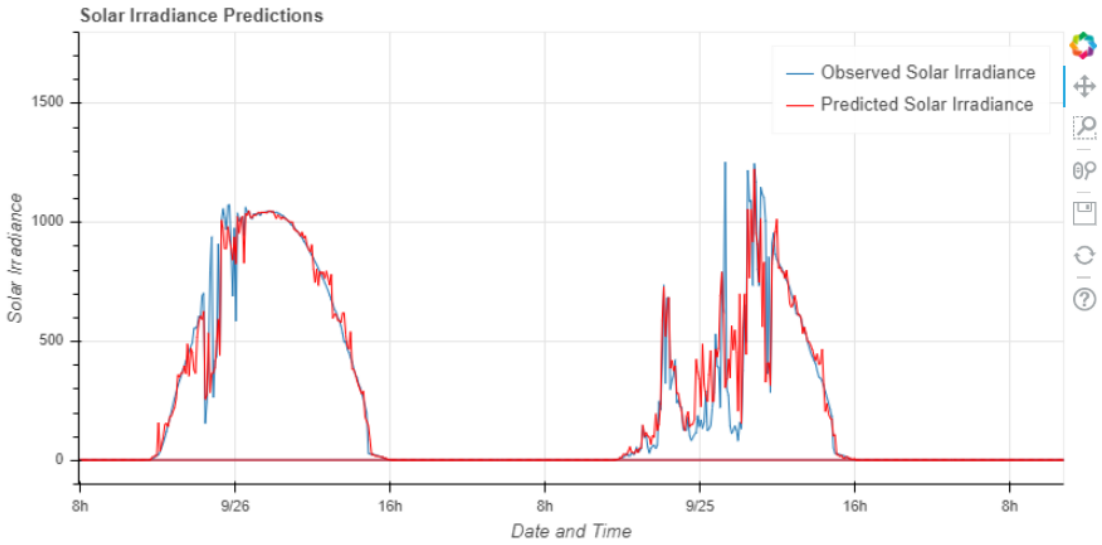
	0	1
0	Humidity	0.093209
1	Pressure	0.037482
2	Speed	0.023945
3	Temperature	0.007026
4	WindDirection(Degrees)	0.052149
5	air_density	1.166699
6	dayofweek	0.030350
7	hour	0.561442
8	month	0.027697

From the plot air density is the highest corresponding feature in the model while temperature is the lowest. Note that this is different than linear correlation as shown before where this is importance score is based on the model related to its prediction.

Based on the score the best features can be determined. In order to find the best value the weakest features are weeded out one by one in the model and based on the R2 score or the degree of fitting. The calculation of the R2 score is done through a 3 fold cross validation for an accurate fit.

	Features	r2 Score
0	Pressure, Humidity, WindDirection(Degrees), Speed, hour, dayofweek, month, air_density	0.917679
1	Pressure, Humidity, WindDirection(Degrees), hour, dayofweek, month, air_density	0.918822
2	Pressure, Humidity, WindDirection(Degrees), hour, dayofweek, air_density	0.912246
3	Pressure, Humidity, WindDirection(Degrees), hour, air_density	0.901912
4	Humidity, WindDirection(Degrees), hour, air_density	0.893810
5	Humidity, hour, air_density	0.892326

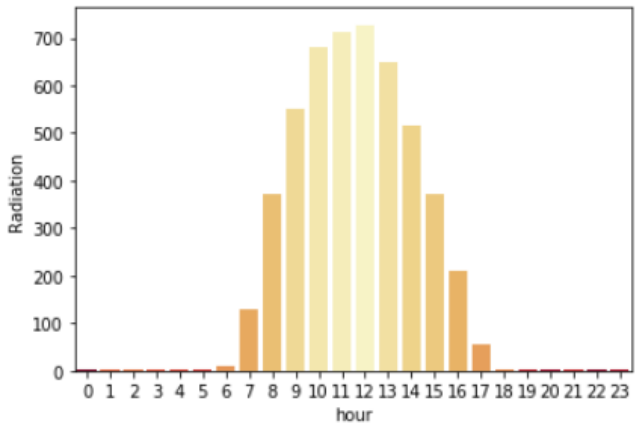
As seen from the table the R2 score is the highest for the features in index 1, and those are the features that will be used in the final model with RF. After training and testing the model the R2 score is 0.917 with MAE at 37.29. By performing feature selection the mean absolute error decreased by a lot making the model much more reliable.



Plotted prediction of the data with the trained RF model.

Conclusion

Overall for the model the temperature is not the main contributor but air density as expected since it all it comes down to matter attenuation. After performing feature selection the MAE is lowered by a lot with cross validation. Can gladly say this model is ready for deployment and prediction.



Average radiation vs hour