Abstract

Initially, I explored the data plotting various histograms, distplot, heatmap to understand the distribution, correlation and univariate importance of all the given variables. Found that temp and feel temp were highly correlated which can be anticipated.

Searched for missing values but data was complete without any values missed. I plotted boxplot and found some outliers in Temperature and Humidity data so I removed them because it didn't indicated possibility of occurrence of any special event.
Removed date column didn't indicate any importance to me.
Converted continuous data into factors so that model works smoothly and cost in terms of time is not compromised.

Build a function to calculate the error given in the data set. Now, I used split my data sets into ratio 80:20 for training and testing respectively. I removed PCA and opted not to go for it because model was pretty fast and was giving satisfactory results.

On my train dataset I built my regressors. I explored various regressors such as RandomForest, SVM, GradientBoostig and I found randomforest generating the best results. Algorithms were performing quite well in terms of time so rather than the data thai I classified I decided to go for the raw data as me test data. I performed parameter tuning for decision trees and Random Forest but gradientboosting was producing better results with default parameters. SVM was ruled out earlier because of slow performance.

Model here is highly simple and easy to understand and fast. I already cross checked for various parameters and tried methods like feature scaling but it already performing better on the raw features.