Predictive modelling for wind power estimation

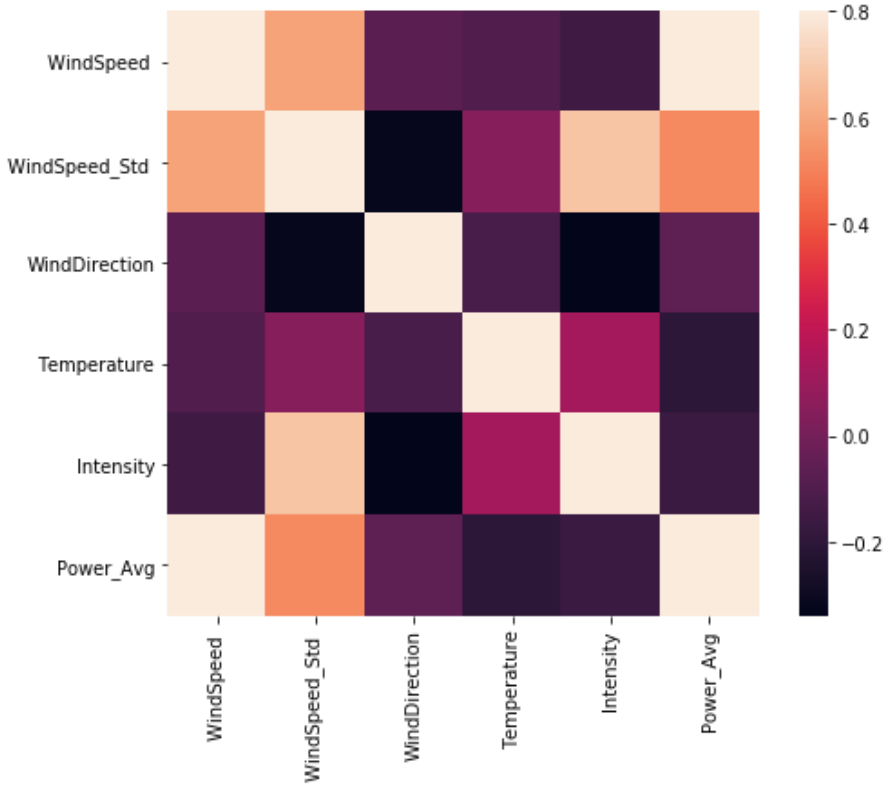
PRESENTED BY: VEDANT MEHTA

KANCHAN SATPUTE

Insight into the data

- The data consists of following variables:
  - Wind Speed
  - Wind Speed Stdev
  - Wind direction
  - Environment temperature
  - Turbulence Intensity (added manually after referring the paper[1])
- The output is Power generated
- The data has been collected over a period of one year so as to factor in the seasonal effect
- All the data recorded has been averaged over 10 minute time intervals

Cont.



- The figure shown on the left is a heat map
- This plot shows the correlation among the predictors and the output variable
- As seen from the plot, power and wind speed have the highest correlation
- Turbulence intensity also has correlation with wind speed and its std. deviation, because it is derived from those two values

Average value of predictors over each month

Month	wind speed	Wind Speed_Std	Wind Direction	Environment Temperature	TI
January	6.18882762	0.665538233	173.7761183	6.402845089	0.110353542
February	6.137140499	0.694159543	164.8385138	7.346446043	0.115861298
March	6.165356194	0.697589228	151.6910919	10.78608079	0.114760932
April	6.665079419	0.796706996	170.6862501	15.56413417	0.122128408
May	6.209380688	0.772476554	155.7210757	20.50823341	0.125958896
June	6.263534855	0.821642071	159.6317888	25.12843335	0.13337773
July	6.233259857	0.754149169	158.1749344	27.23974869	0.120259257
August	6.262362179	0.76313141	155.5356763	28.93998535	0.121109984
September	6.030573995	0.749104965	135.4471957	25.29694799	0.126558718
October	6.034052343	0.65773111	182.2127556	20.94305276	0.111203613
November	6.908429521	0.818605107	180.748135	13.33846974	0.120768332
December	6.372286851	0.650090311	236.2062578	8.428433218	0.104903643

As seen from the above figure, wind direction and temperature have a lot of variability., which may hamper prediction accuracy. Turbulence intensity already has std. dev. in it, so we can drop wind speed std. deviation as well

Predictive modelling

- Predictive models built, with their RMSE values:

Model	Training RMSE	Validation RMSE
Linear regression	0.0885	0.0895
Ridge regression	0.081	0.078
Lasso regression	0.0831	0.0845
SVM	0.0474	0.0470
CART	0.0017	0.0552
Tree with AdaBoost	0.0495	0.0494
Random Forest	0.0356	0.039

- We will discuss only two methods, decision tree and random forest
- And also explain why we have chosen them
- We have split the data randomly in the ratio of 7:3 for training and validation

Decision Tree

- Decision Tree is a much simpler algorithm as compared to other algorithms
- Decision tree gives better results as compared to linear, ridge, lasso and SVM
- This is because when you compare this method with the industry standard IEC binning method, they both are kind of analogous
- In IEC binning method, you make bins of predefined bin width and then put the data points in each of that bins and for prediction we average the corresponding output values in each bin

Cont.

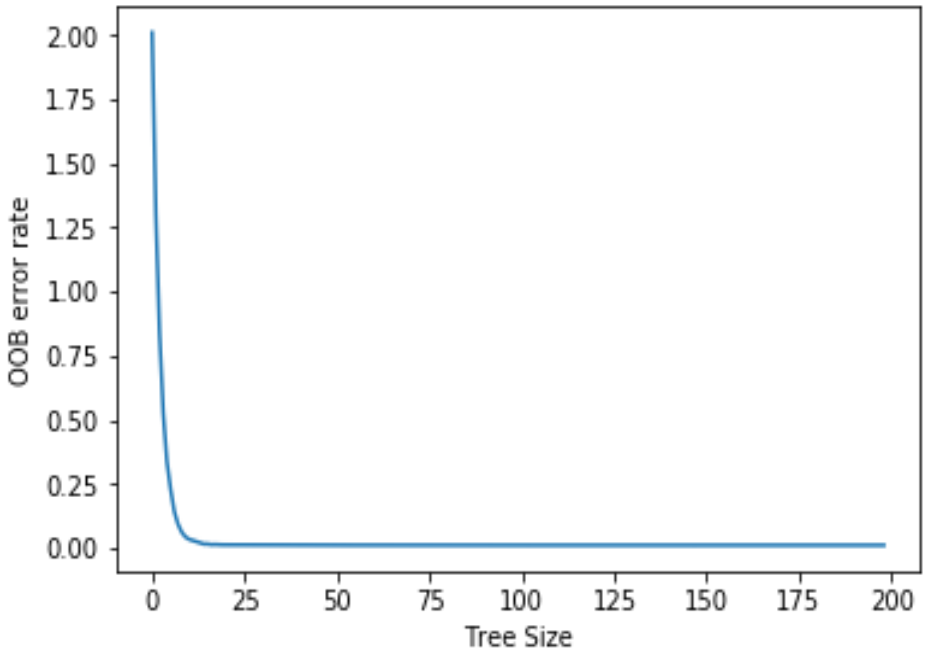
- In decision trees, we do a similar kind of procedure, except we don't specify any kind of "node size" other than the min. node size
- Also, we do the prediction by taking the average at each node

- RMSE values for CART:
  - Training RMSE = 0.0017
  - Validation RMSE = 0.0552
- Above RMSE values show that decision tree has overfit the data, thus giving a high validation RMSE

Random Forest

- Random Forest is an ensemble of trees, so our analogy with IEC binning method still holds
- But, Random Forest builds trees on bootstrap samples and bags them together
- Thus, reducing the variance as compared to decision trees
- So, Random Forest is a low bias low variance model as compared to decision trees

Fine tuning Random Forest



- We gave an array of tree sizes as an input and used 10-fold cross validation to see which tree size gave the least cross validation error
- The best parameter (tree size = 20) model gave the following RMSE values:
  - Training RMSE = 0.0356
  - Validation RMSE = 0.039
- We cross-checked this using a plot of OOB error rate vs tree size

Concluding remarks

We have chosen random forest as the best model due to following reasons:

- We get very low training RMSE and validation RMSE
- Random Forest has low variance as compared to decision trees
- SVM is computationally complex, so Random Forest is preferred for huge amount of data
- Also, Random Forest provides better RMSE values as compared to all other methods that we applied