Report Assignment 5

Part (1) Decision Trees - Bagging - Random Forest

Data Pre-processing:

1) First column (serial number) from the data is dropped.
2) There are missing values in the column “pm2.5” .These missing values are replaced by mean of the values under this column.
3) The values under column “cbwd” are categorical therefore label encoding is performed to convert them into numeric values.
   {'NW': 1, 'cv': 3, 'NE': 0, 'SE': 2}

Assumptions:

1) While constructing decision/regression tree only binary split of the nodes is performed to maintain simplicity.
2) All the attributes except “cbwd” are considered as continuous.Therefore are managed in the same fashion.
3) Replacing the missing value of the feature “pm2.5” by mean is as assumption.The same can be done using median as well.But mean is giving better accuracy/MSE

Data Summerization

Data shape : (43824, 12)

	year	month	day	hour	pm2.5	DEWP	TEMP	PRES	cbwd	Iws	Is	Ir
0	2010	1	1	0	0.0	-21	-11.0	1021.0	NW	1.79	0	0
1	2010	1	1	1	0.0	-21	-12.0	1020.0	NW	4.92	0	0
2	2010	1	1	2	0.0	-21	-11.0	1019.0	NW	6.71	0	0
3	2010	1	1	3	0.0	-21	-14.0	1019.0	NW	9.84	0	0
4	2010	1	1	4	0.0	-20	-12.0	1018.0	NW	12.97	0	0
...	...	...	...	...	...	...	...	...	...	...	..	..
43819	2014	12	31	19	8.0	-23	-2.0	1034.0	NW	231.97	0	0
43820	2014	12	31	20	10.0	-22	-3.0	1034.0	NW	237.78	0	0
43821	2014	12	31	21	10.0	-22	-3.0	1034.0	NW	242.70	0	0
43822	2014	12	31	22	8.0	-22	-4.0	1034.0	NW	246.72	0	0
43823	2014	12	31	23	12.0	-21	-3.0	1034.0	NW	249.85	0	0

Training Data: Data points with year = 2010 and 2012 are taken for traing model.

Shape: (17520 X 12)

Testing Data: Data points with year = 2011 and 2014 are taken for testing

Shape: (17544 X 12)

Measure of Impurity

Tree Type	Impurity Measure	Formula /Description
Classification Tree	Gain Split (Information Gain)	$Gain_{split} = Entropy(p) - \sum_{i=1}^{k} \frac{n_i}{n} Entropy(i)$ Parent Node, p is split into k partitions (children) n_i is number of records in child node i
Regression Tree	Residue Squared Sum(RSS)	$\text{RSS} = \sum_{m=1}^{M} \sum_{i \in R_m} (y_i - \hat{y}_{R_m})^2$

Tree Algorithm (Classification/Regression)

Steps:

- Select the feature to be predicted { attribute: month (in case of classification) ,attribute : pm2.5 (in case of regression) }
- Recursively develop nodes of the tree. Split a node corresponding to a feature value based on the **Maximum Gain Split (Classification) / Minimum RSS (Regression)**.
- The anchor condition of the recursion would be :

  Classification :

  - When all the points in the node belong to the same class.
  - When number of points in a node is less than or equal to 29 (empirical value)

  Regression :

  - When number of points in a node is less than or equal to 19 (empirical value)

Results Classification Tree

Note : Please note that I have not performed any kind of pruning.

Height	Accuracy Reported
H=35	35.02 %
H=40	37.68%
H=57	43.085%

Results Bagging of Classification Tree

Note For the following results I have used classification tree of height=57

Bag Size	Accuracy Reported
Size=5	45.09%
Size=10	49.18%
Size= 20	54.036%

Results Random Forest of Classification Tree

Note For the following results I have used classification tree of height=57

Forest Size	Accuracy Reported
Size=5	44.13%
Size=10	51.34%
Size=19	58.29%

Results Regression Tree

Note : Please note that I have not performed any kind of pruning.

Height	MSE	Mean Absolute Error	Std Deviation
H=25	17978.56801	102.066	87.29
H=30	11197.23566	95.288	81.33
H=35	977.244289	91.875	75.269

Results Bagging of Regression Tree

Tree Height =35

Bag Size	MSE	Mean Absolute Error	Std Deviation
Size=4	970.25689	91.02	75.09
Size=7	885.23657	90.78	71.46
Size=10	877.24428	89.75	69.99

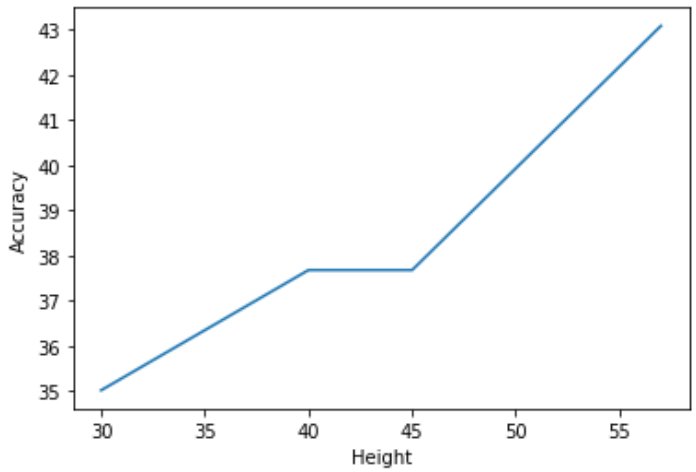
Results Random Forest of Regression Tree

Tree Height=35

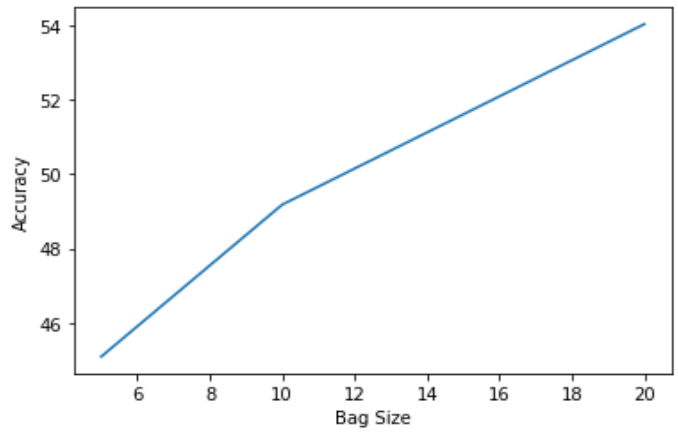
Bag Size	MSE	Mean Absolute Error	Std Deviation
Size=4	960.8829	90.149	74.89
Size=7	881.8697	88.88	70.77
Size-10	790.1257	79.875	65.267

Inferences

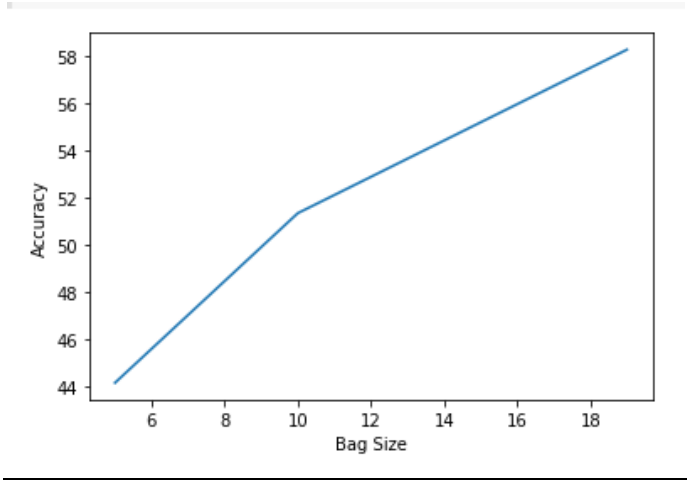
Classification Tree (Height vs Accuracy)



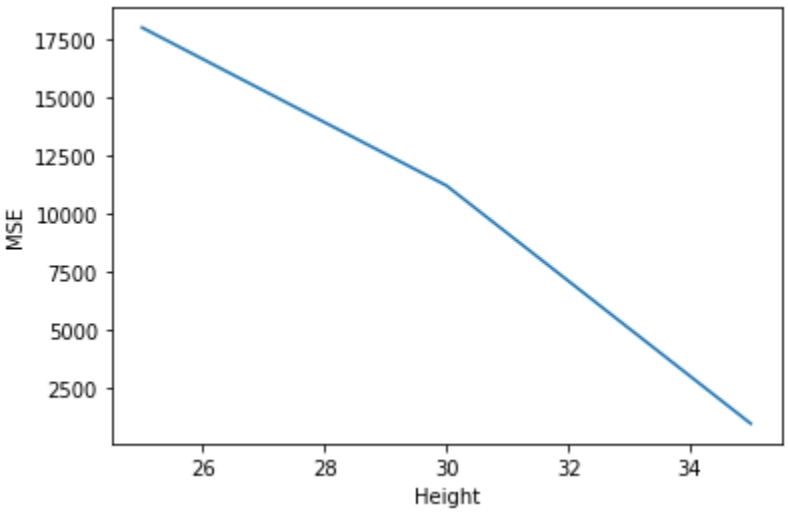
Bag Size vs Accuracy



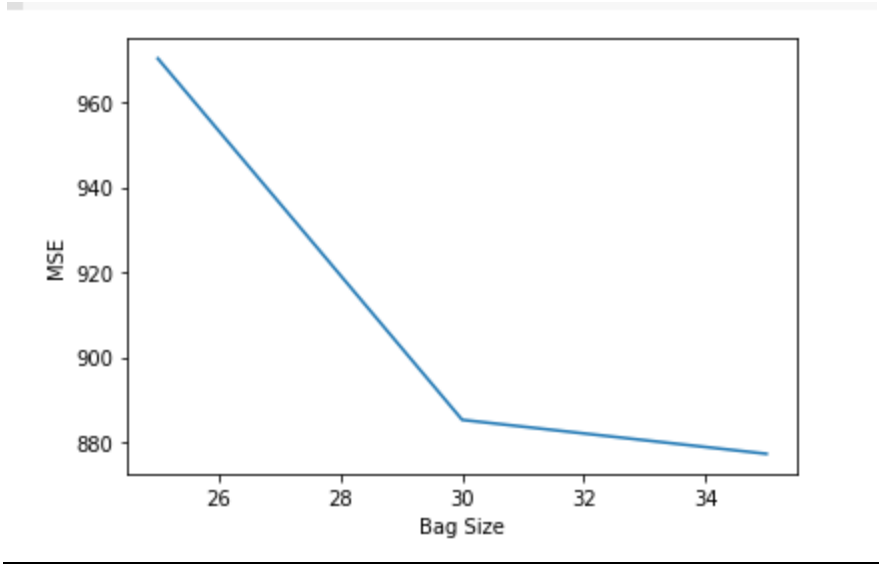
Forest Size vs Accuracy



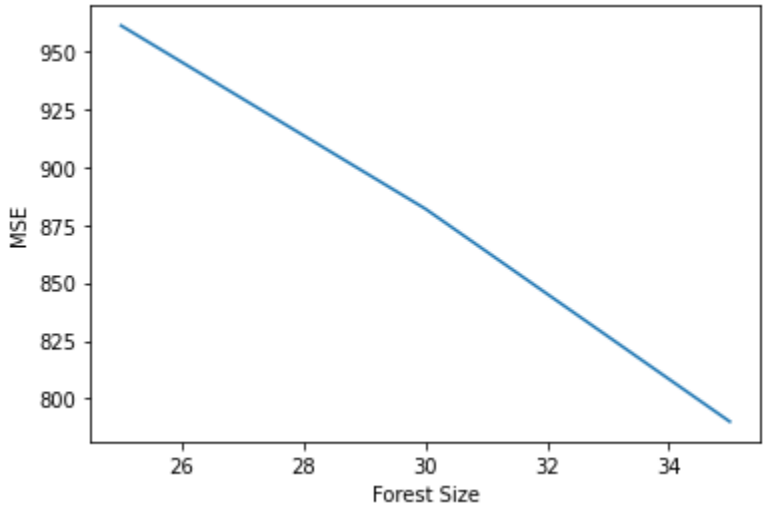
Regression Tree (Height vs MSE)



Regression Tree(Bag size vs MSE)



Regression Tree (Forest Size vs MSE)



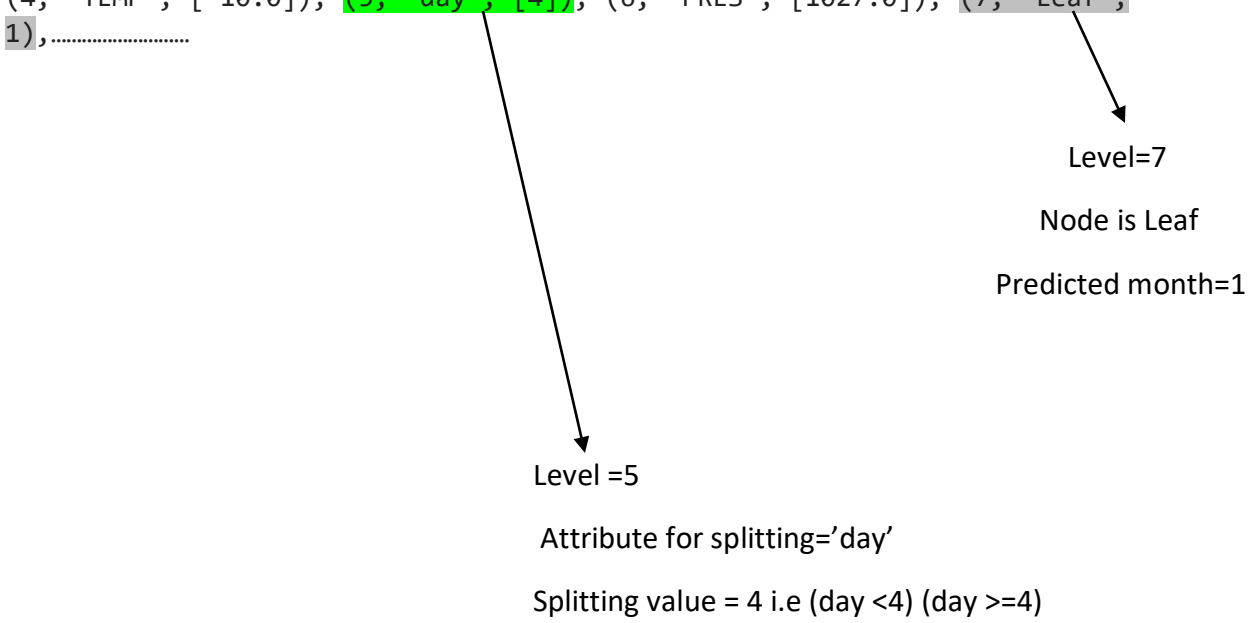
Decision Tree

Please note that the Decision Tree produced is Binary Search Tree. And a BST has **unique pre-order traversal** . My code is using this pre-order sequence to predict the target label/value.

##In code this pre order sequence is named as Decision_Tuple ####

Sample Pre-order sequence:

```
[(0, 'TEMP', [11.0]), (1, 'TEMP', [2.0]), (2, 'DEWP', [-10]), (3, 'day', [22]),
(4, 'TEMP', [-10.0]), (5, 'day', [4]), (6, 'PRES', [1027.0]), (7, 'Leaf',
1),……
```



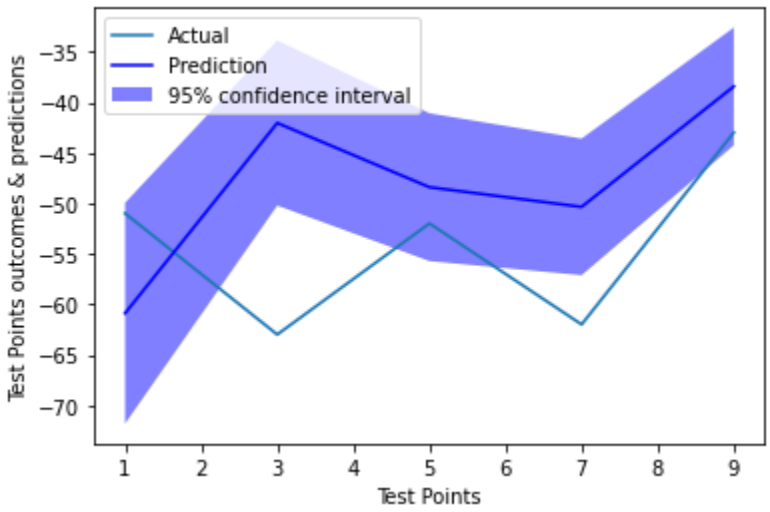
Part (2) Gaussian Processes

```
#######Special Credits for developing the code : AV Sir's lecture material
 ###########
```

Result :

Distance	Signal Strength(Actual)	Predicted Mean of Signal Strength	Variance of Signal Strength
1	-51	-60.8911	5.577
3	-63	-42.049	4.169
5	-52	-48.408	3.733
7	-62	-50.359	3.449
9	-43	-38.433	2.972

Graph



Kernel Used : RBF and constant kernel [kernel = C(1.0, (1e-10, 1e10))* RBF(2, (1e-4, 1e4))]

Inferences Drawn :

Predicted mean and variance values are pretty convincing at each test point except at x=3 and x=7