Problem we tried to solve

- Predict the level of solar radiation from the measurable meteorological information provided by NASA HI-SEAS (Hawaii's Space Exploration Analog and Simulation mission) sensors.
- This allows the HI-SEAS crew, or other explorers reliant on solar panels, to plan their daily energy consumption for all their human necessities and other planned activities based on expected energy output from solar panels.
- Dataset: The dataset is meteorological data from the HI-SEAS weather station from four months (September through December 2016) between Mission IV and Mission V.
- Dataset location: https://www.kaggle.com/dronio/SolarEnergy
- Features in the dataset: Date, Time of Day, Temperature, Pressure, Humidity,

  Wind Direction, Wind Speed, Time at Sunrise, Time at Sunset, Solar Radiation.
- Solar Radiation is the response variable. Since the response variable is continuous, this is a regression problem.

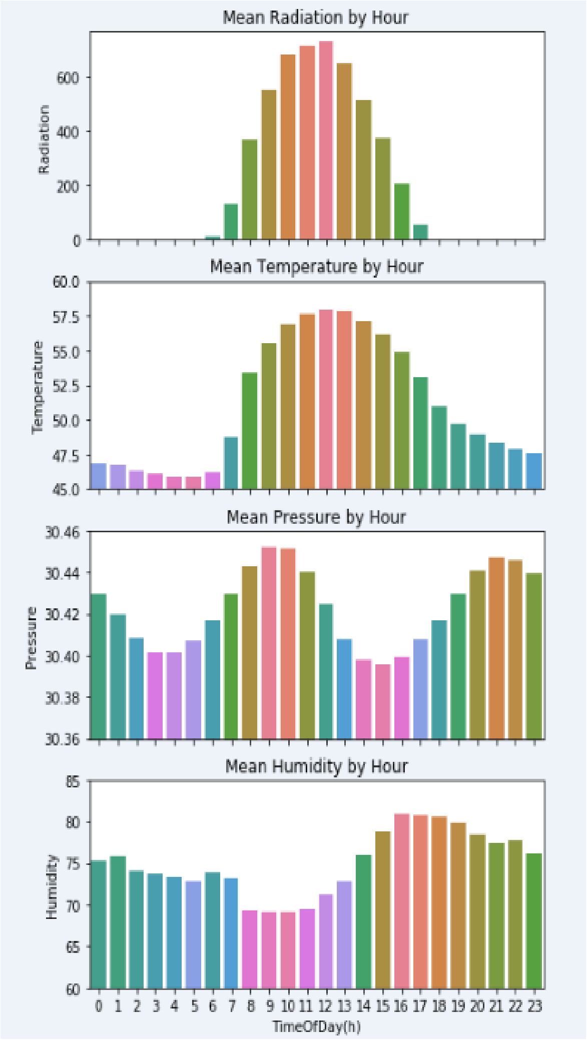
Project Stages

- 1) Data Collection and Data preprocessing
- 2) Exploratory Data Analysis
- 3) Feature Engineering
- 4) Dataset splitting
- 5) Modeling

  (i) Model Training

  (ii) Model evaluation and parameters tuning

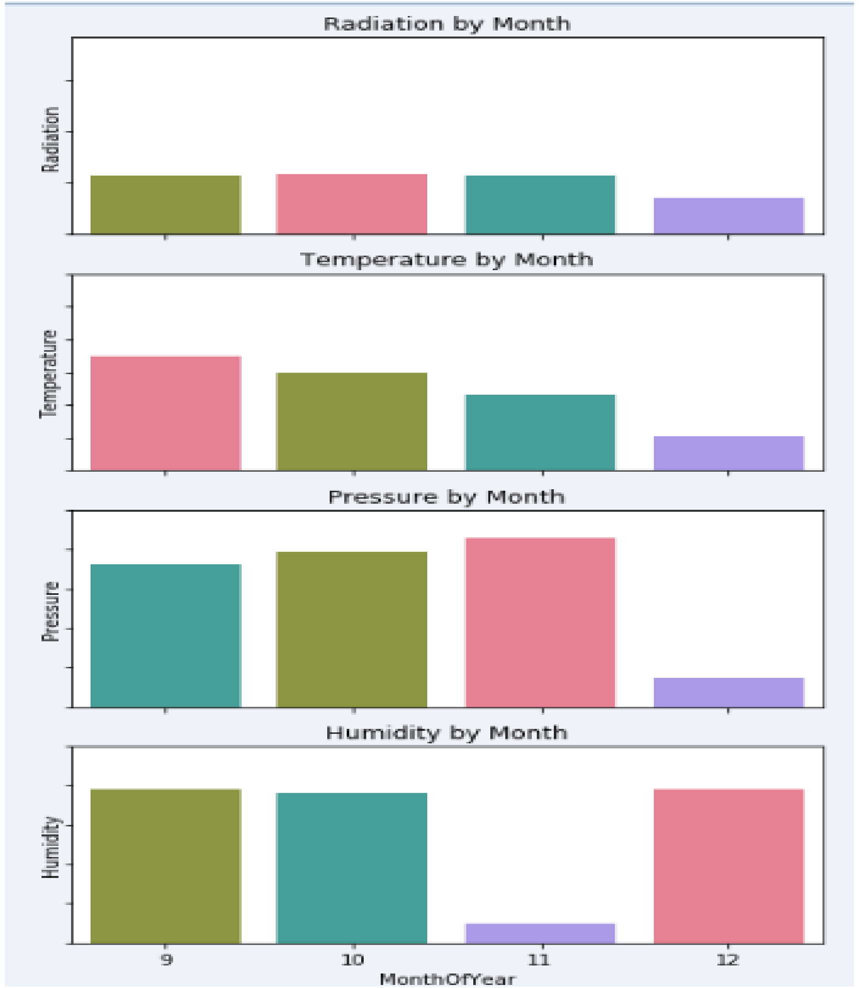
Exploratory Data Analysis



- From the above plots, its clear that temperature has strong correlation with solar radiation.

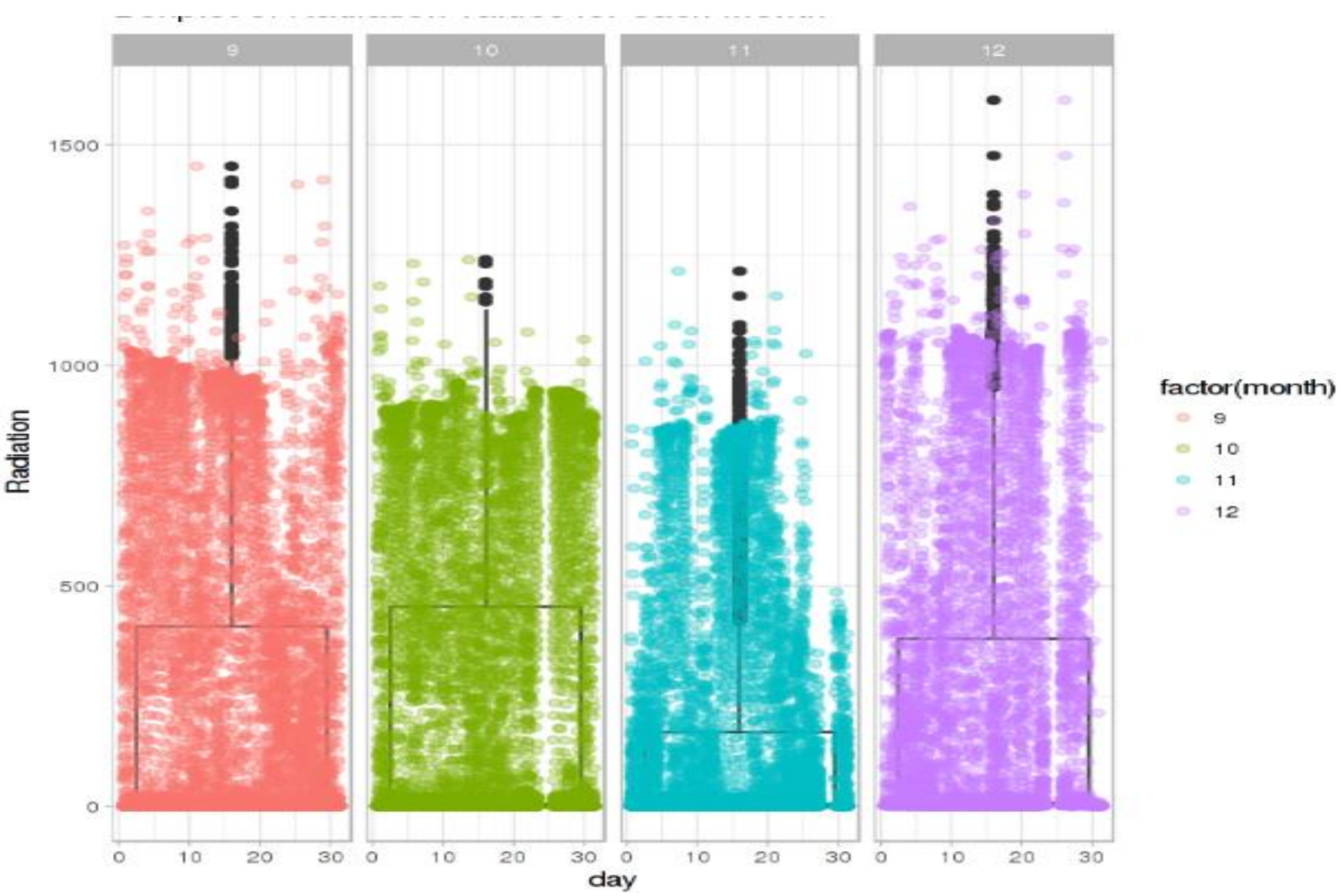
- Relationships between pressure/humidity and solar radiation is less clear but it does appear that humidity has a negative correlation with solar radiation, temperature and pressure.

Exploratory Data Analysis

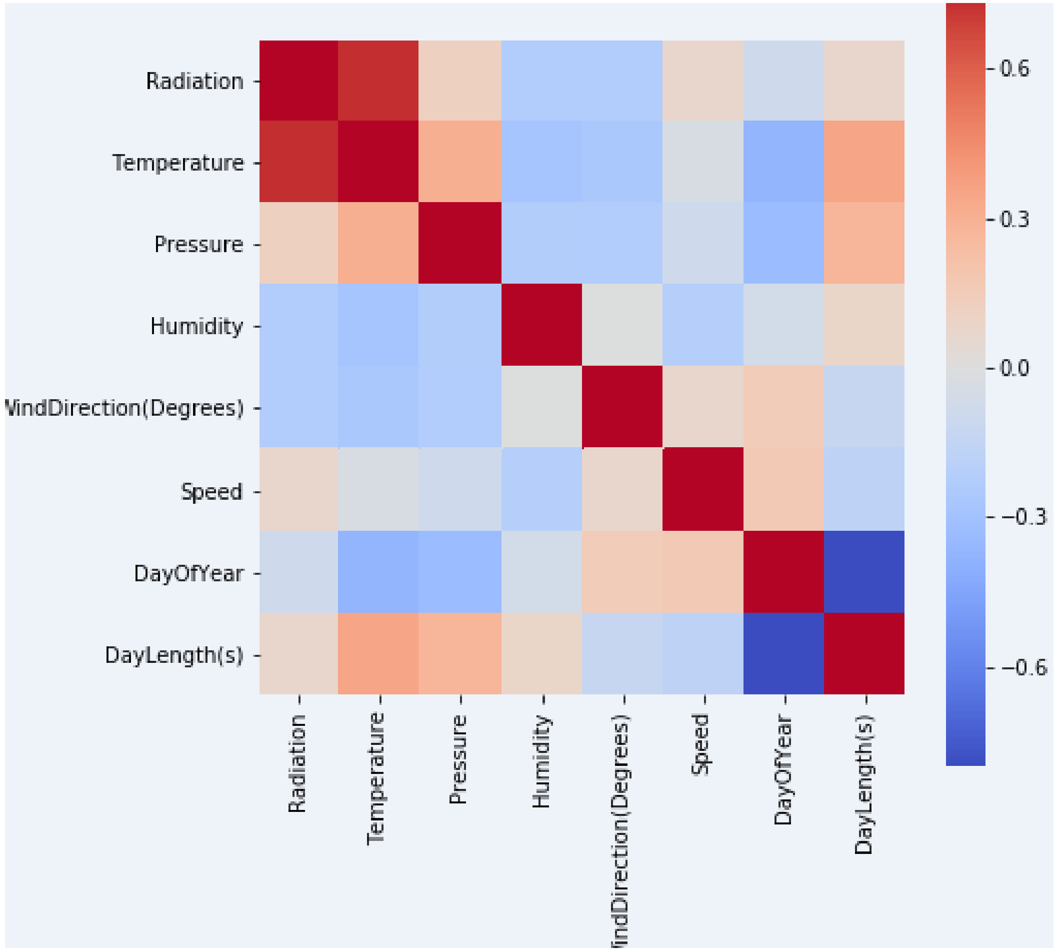


- As winter is approaching the radiation level relatively reduces in December month.
- Monthly temperature appear to decrease which can be observed.
- Humidity was relatively low in November month whereas pressure increased and then fall down in December.

Box Plot Radiation for each month



Correlation Plot

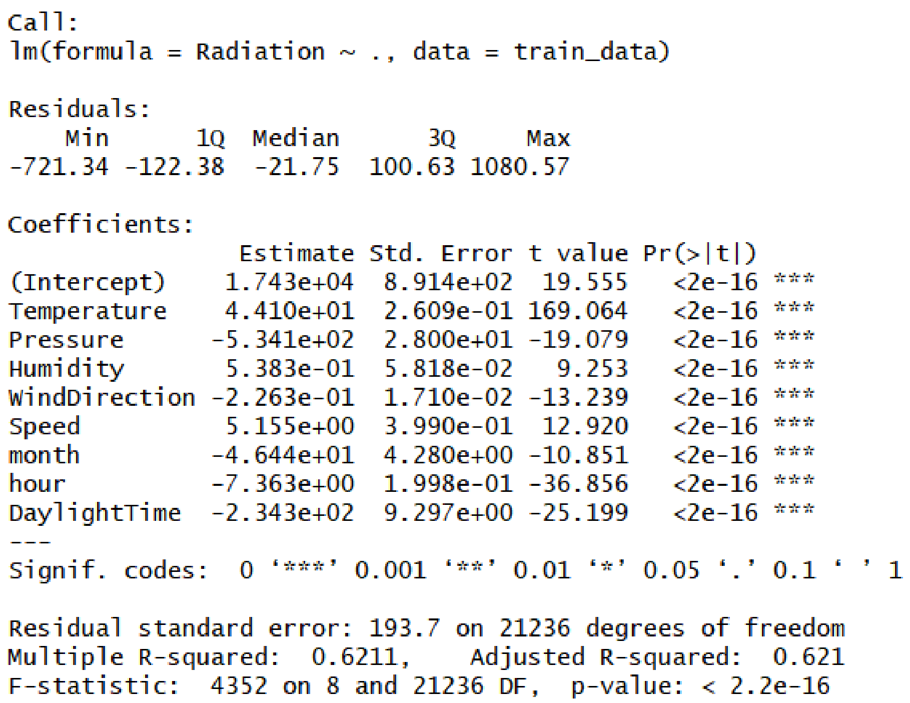


- As we can observe, Radiation and Temperature are highly correlated
- Radiation is inversely correlated to humidity and to the direction of the wind

Feature Engineering

- Feature engineering is the process of using domain knowledge of the data to create features that better represent the underlying problem to the predictive models. This results in improved model accuracy on unseen data.
- We have dates, timestamps of sunrise and sunset columns for the meteorological readings. It doesn`t make sense to use these columns as features directly. Instead, we generated new features from these columns like hour of the day, month, total Daylight time (sunset time – sunrise time).
- Finally, the unnecessary columns (UNIXTime, Date, Time, TimeSunRise, TimeSunSet) were dropped in the dataset. Now the data is ready for applying data mining models to make predictions for the response variable “Radiation”.
- Data Splitting: Before applying the models, We have split the data into train(65%) (21245 obs) and test sets(35%) (11441 obs).

Applying Linear Regression (OLS)



```
Call:
lm(formula = Radiation ~ ., data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-721.34 -122.38  -21.75  100.63 1080.57 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)    1.743e+04  8.914e+02  19.555   <2e-16 ***
Temperature    4.410e+01  2.609e-01 169.064   <2e-16 ***
Pressure      -5.341e+02  2.800e+01 -19.079   <2e-16 ***
Humidity       5.383e-01  5.818e-02   9.253   <2e-16 ***
WindDirection -2.263e-01  1.710e-02 -13.239   <2e-16 ***
Speed          5.155e+00  3.990e-01  12.920   <2e-16 ***
month         -4.644e+01  4.280e+00 -10.851   <2e-16 ***
hour          -7.363e+00  1.998e-01 -36.856   <2e-16 ***
DaylightTime  -2.343e+02  9.297e+00 -25.199   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 193.7 on 21236 degrees of freedom
Multiple R-squared:  0.6211,	Adjusted R-squared:  0.621 
F-statistic:  4352 on 8 and 21236 DF,  p-value: < 2.2e-16
```

- We can see that there is a relationship between the predictors and response as the F-Statistic is far from 1 (with a small p-value) indicating evidence against the null hypothesis.

- Looking at the p-values associated with each predictor's t-statistic, we see that all the predictors have statistically significant relationship with the response variable "Radiation".

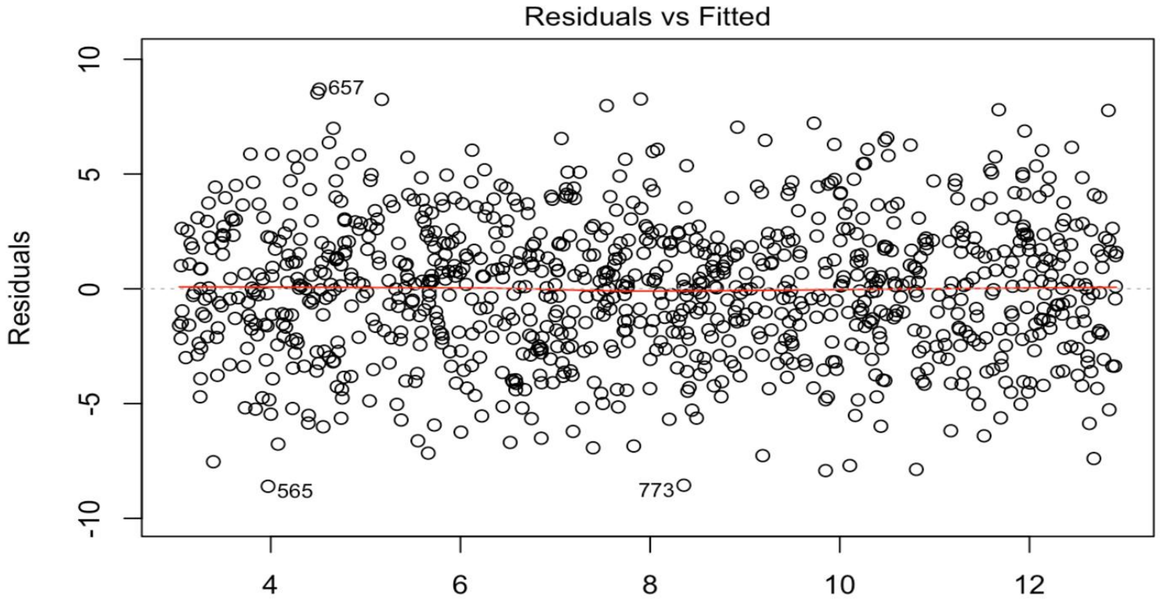
Applying Linear Regression (OLS)

- Couple of assumptions for Linear Regression (OLS estimates) are :

  (i) There should be linear relationship between independent variables and response variable.

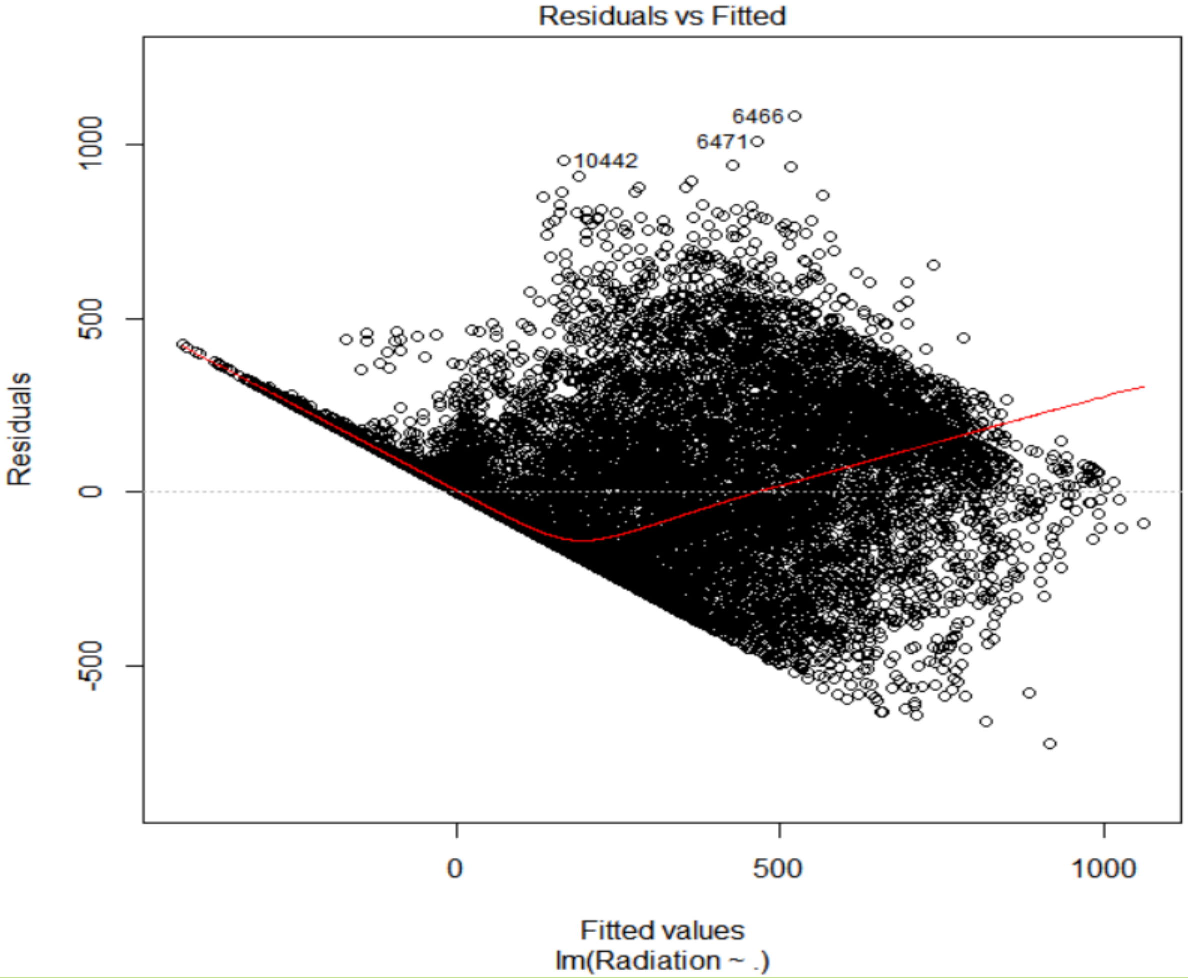
  (ii) The error terms should have constant variance.

- Residuals vs Fitted Plot can be used to test whether the first two assumptions hold true.



- This is an ideal Residuals vs Fitted graph.
- The residuals are randomly spread(without any pattern) around the 0 line. This implies that there is no non-linearity in the data.
- The residuals spread around the line is constant for all fitted values. Hence the variance in Residuals is constant.

Applying Linear Regression (OLS)

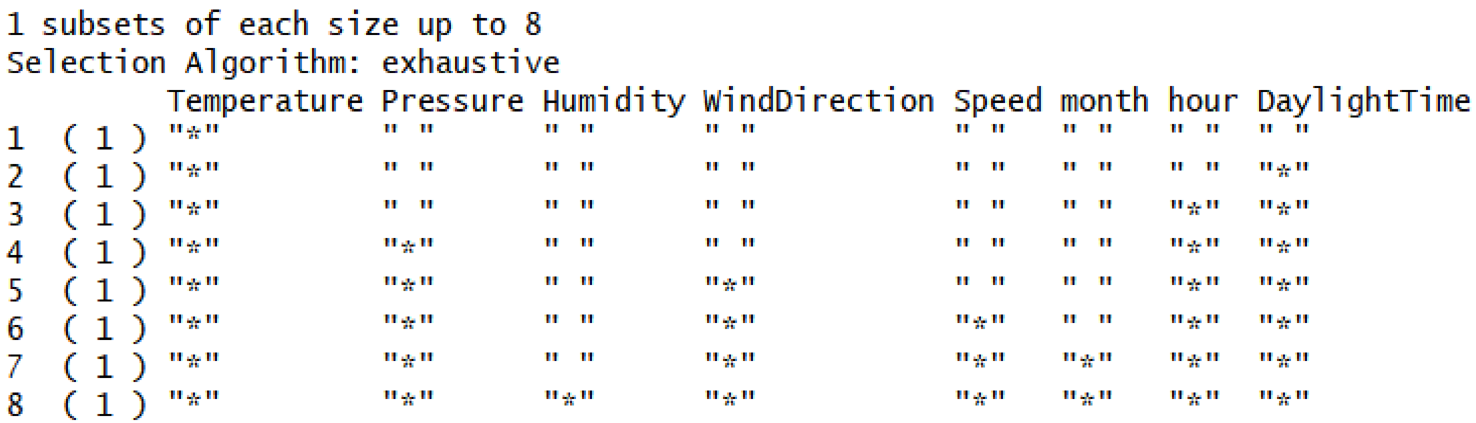


Residuals vs Fitted plot for our data

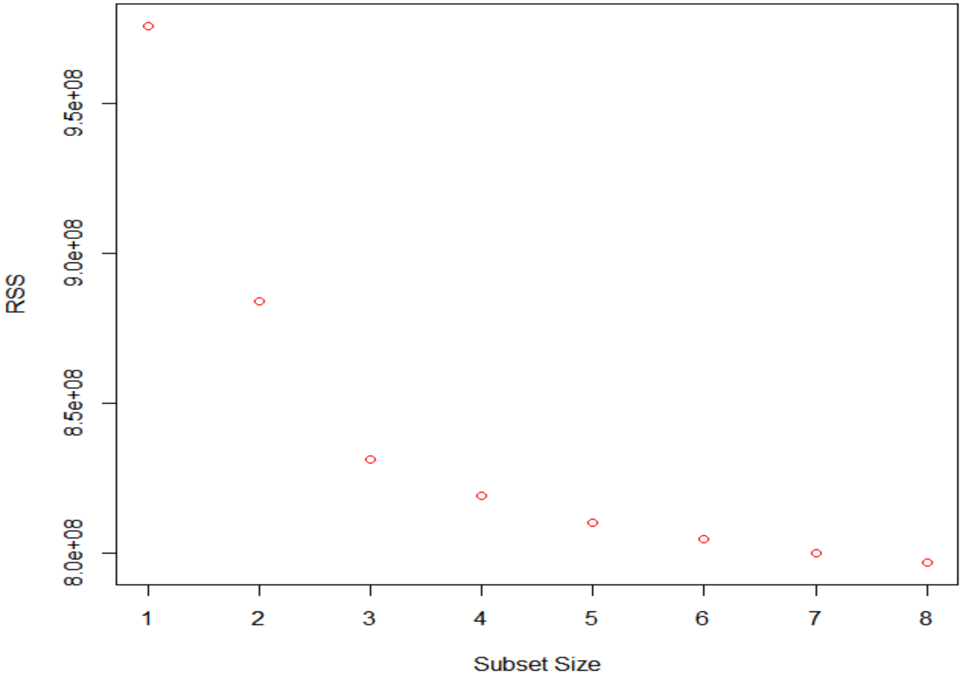
- There is a pattern in the graph implying some non-linearity in the data.
- Also the residuals spread varies with fitted values. So the residuals have non-constant variance (Heteroscedasticity).
- Hence the assumptions of OLS-linear regression didn`t hold true here and so the inferences - such as calculation of standard errors, confidence intervals and p-values which rely on these assumptions might no longer hold true.
- Hence, a non-linear model might perform well on this data.

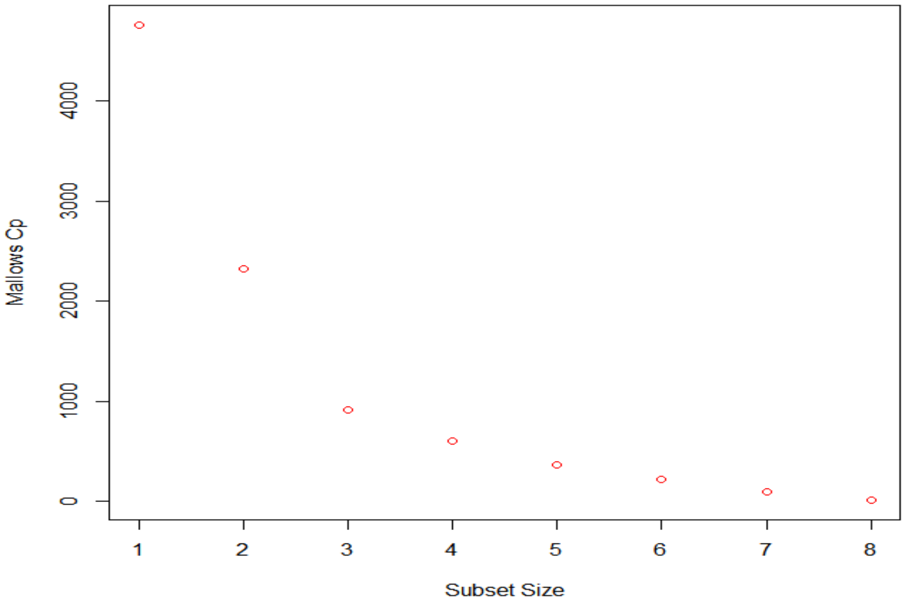
Mean Squared Error on test data: 38230.

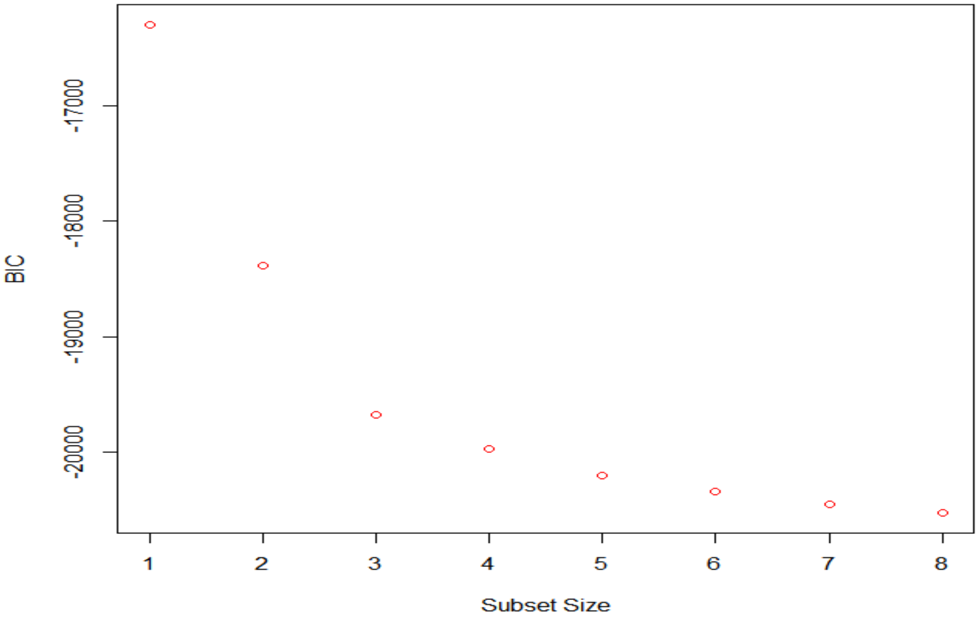
Best Subset selection (Exhaustive)



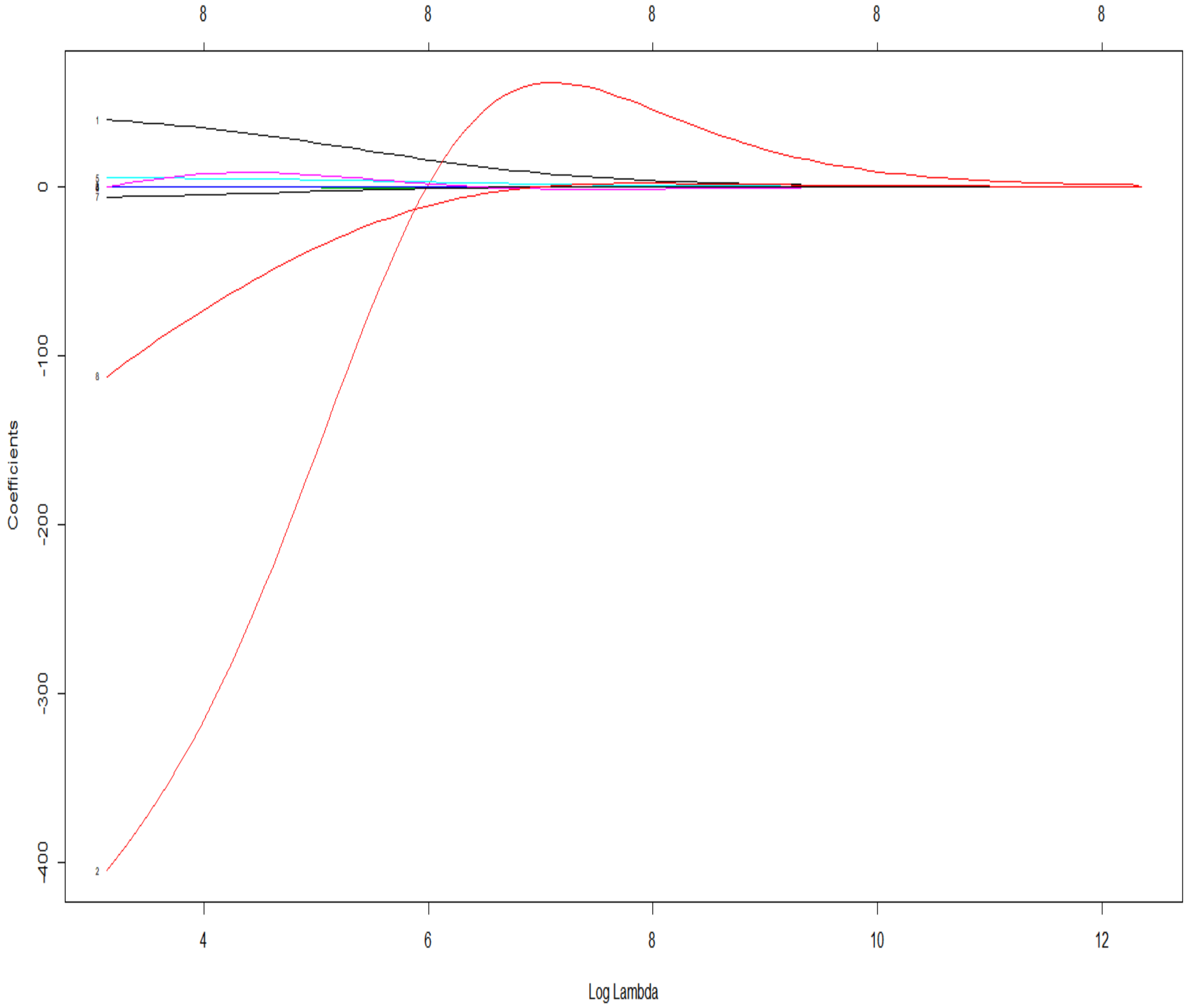
```
1 subsets of each size up to 8
Selection Algorithm: exhaustive
         Temperature Pressure Humidity WindDirection Speed month hour DaylightTime
1  ( 1 ) "*"         " "      " "      " "           " "   " "   " "  " "
2  ( 1 ) "*"         " "      " "      " "           " "   " "   " "  "*"
3  ( 1 ) "*"         " "      " "      " "           " "   " "   "*"  "*"
4  ( 1 ) "*"         "*"      " "      " "           " "   " "   "*"  "*"
5  ( 1 ) "*"         "*"      " "      "*"           " "   " "   "*"  "*"
6  ( 1 ) "*"         "*"      " "      "*"           "*"   " "   "*"  "*"
7  ( 1 ) "*"         "*"      " "      "*"           "*"   "*"   "*"  "*"
8  ( 1 ) "*"         "*"      "*"      "*"           "*"   "*"   "*"  "*"
```





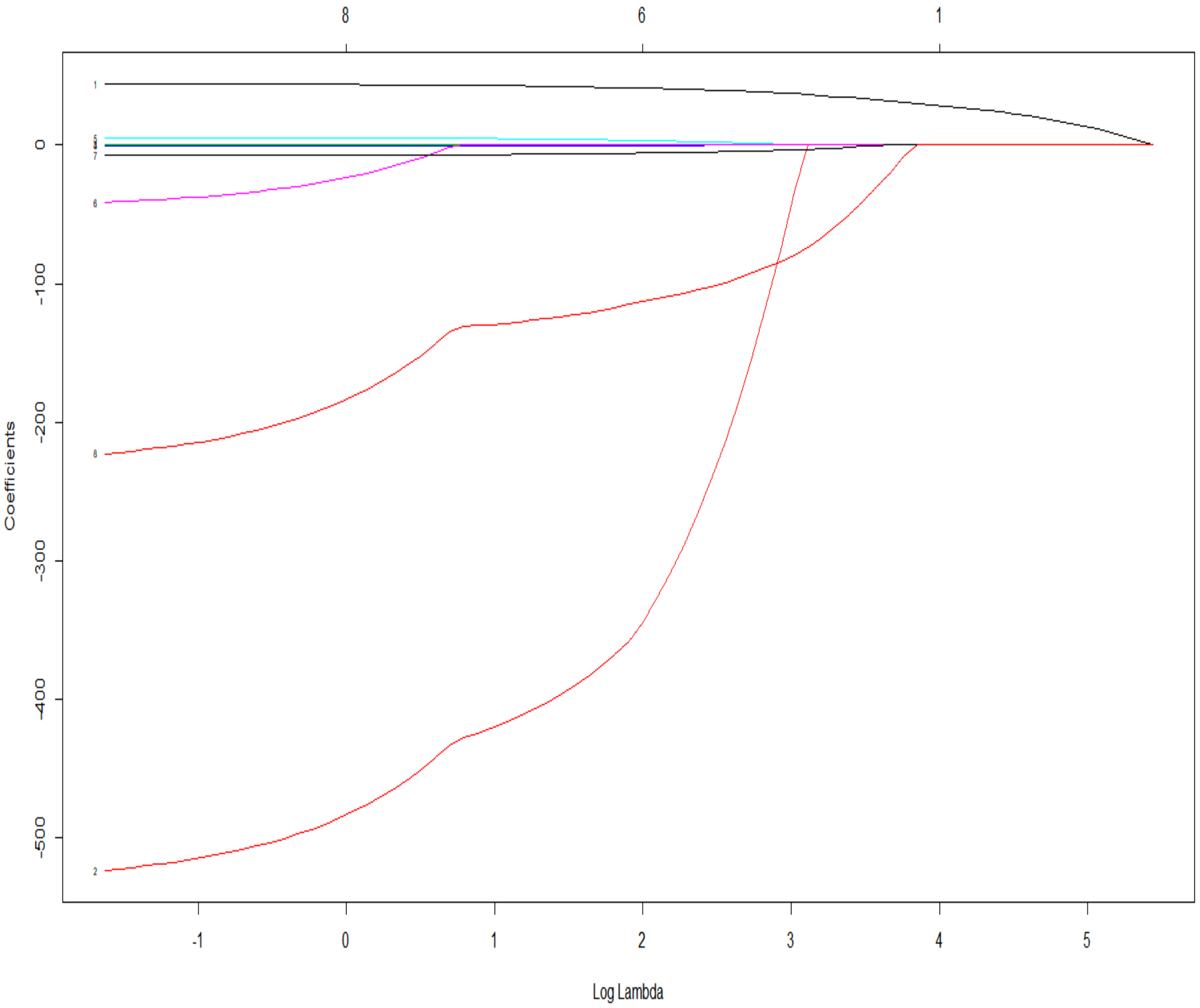


Ridge Regression



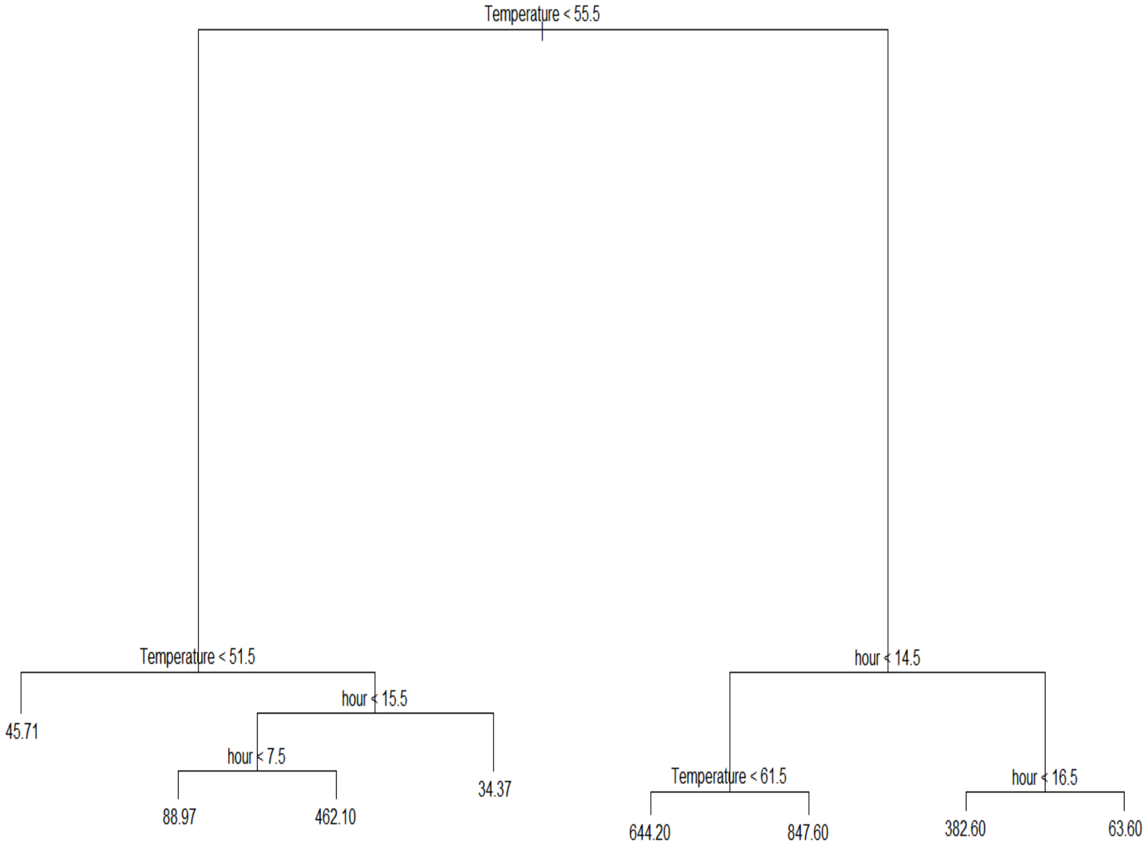
- For the best lambda (obtained from cross validation), the significant predictors are Temperature, Pressure, Daylight time.
- Mean Squared Error obtained on test data = 39159.

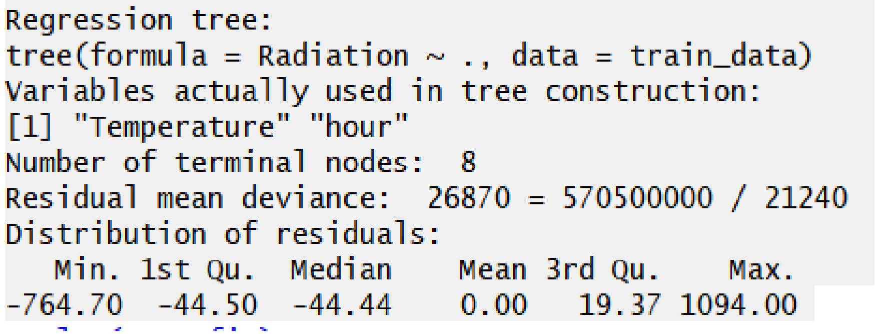
Lasso Regression



- For the best lambda (obtained from cross validation), the significant predictors are Temperature, Pressure, Daylight time.
- Mean Squared Error obtained on test data = 38236.
- Regularized models are typically used to overcome the problem of overfitting in OLS.
- Since overfitting was not a problem for our data, there was no improvement in using regularized models.

Regression Tree

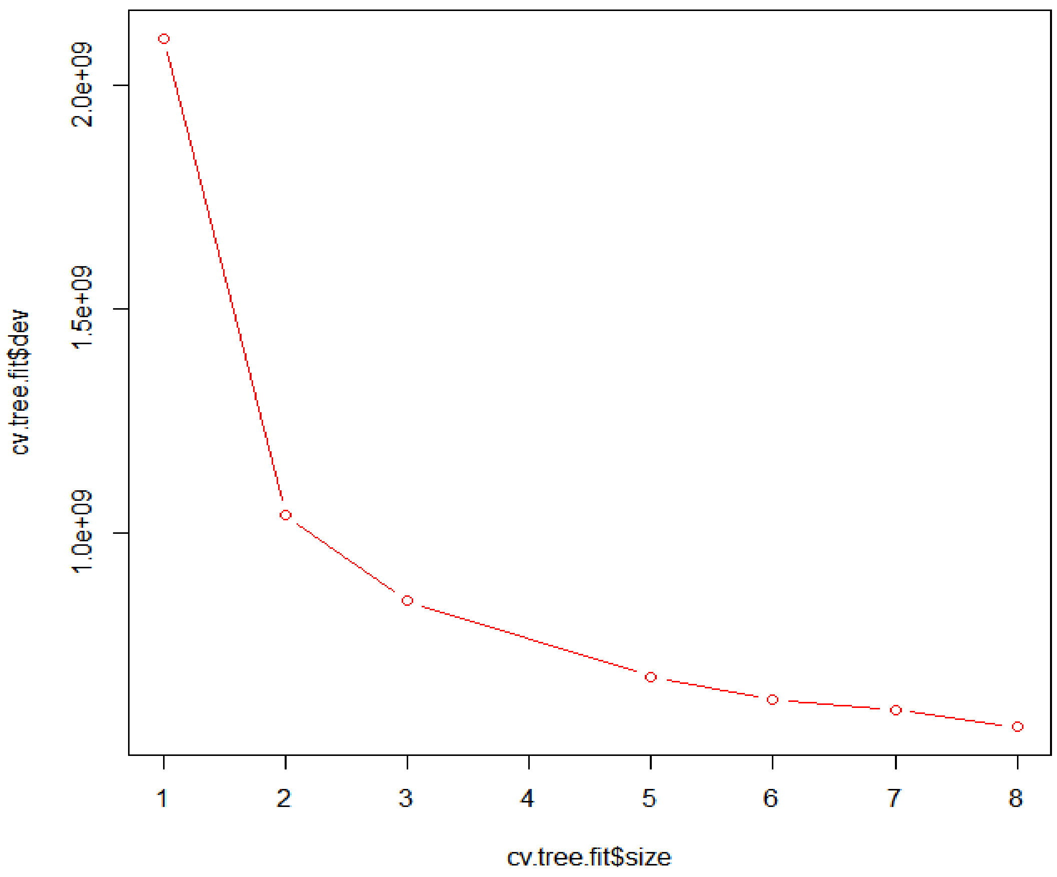




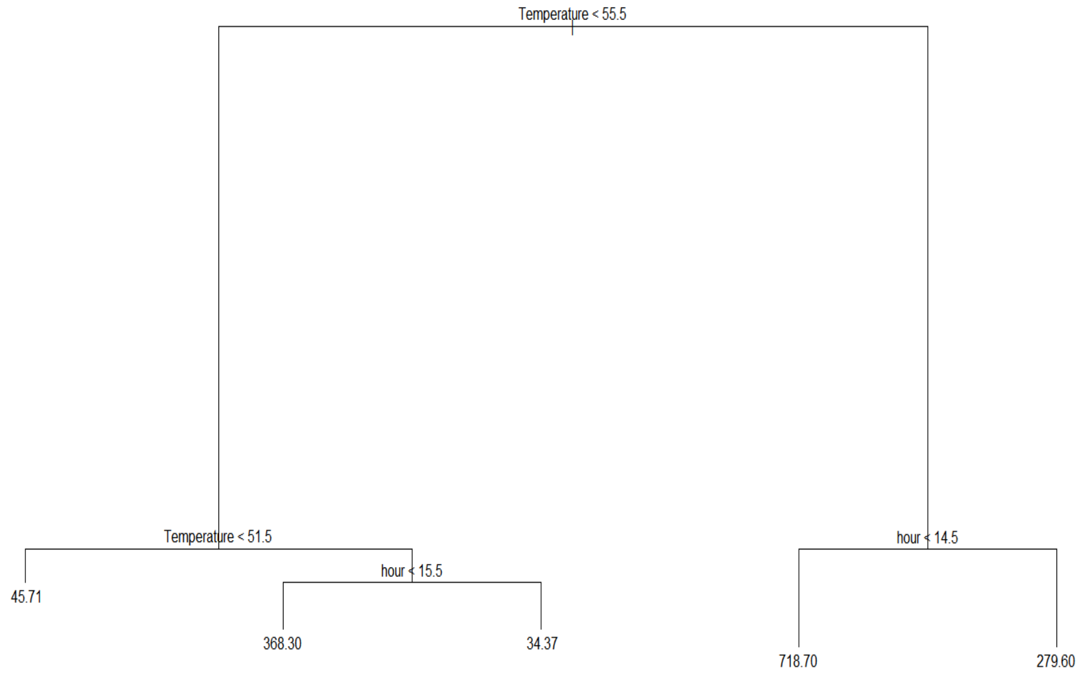
```
Regression tree:
tree(formula = Radiation ~ ., data = train_data)
Variables actually used in tree construction:
[1] "Temperature" "hour"
Number of terminal nodes:  8
Residual mean deviance:  26870 = 570500000 / 21240
Distribution of residuals:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-764.70  -44.50  -44.44    0.00   19.37 1094.00
```

- From the summary, we can see that only two predictors “Temperature” and “hour” were used in constructing the tree.
- Deviance (Train MSE for regression tree) = 26870.
- Test MSE obtained = 27740

Regression Tree

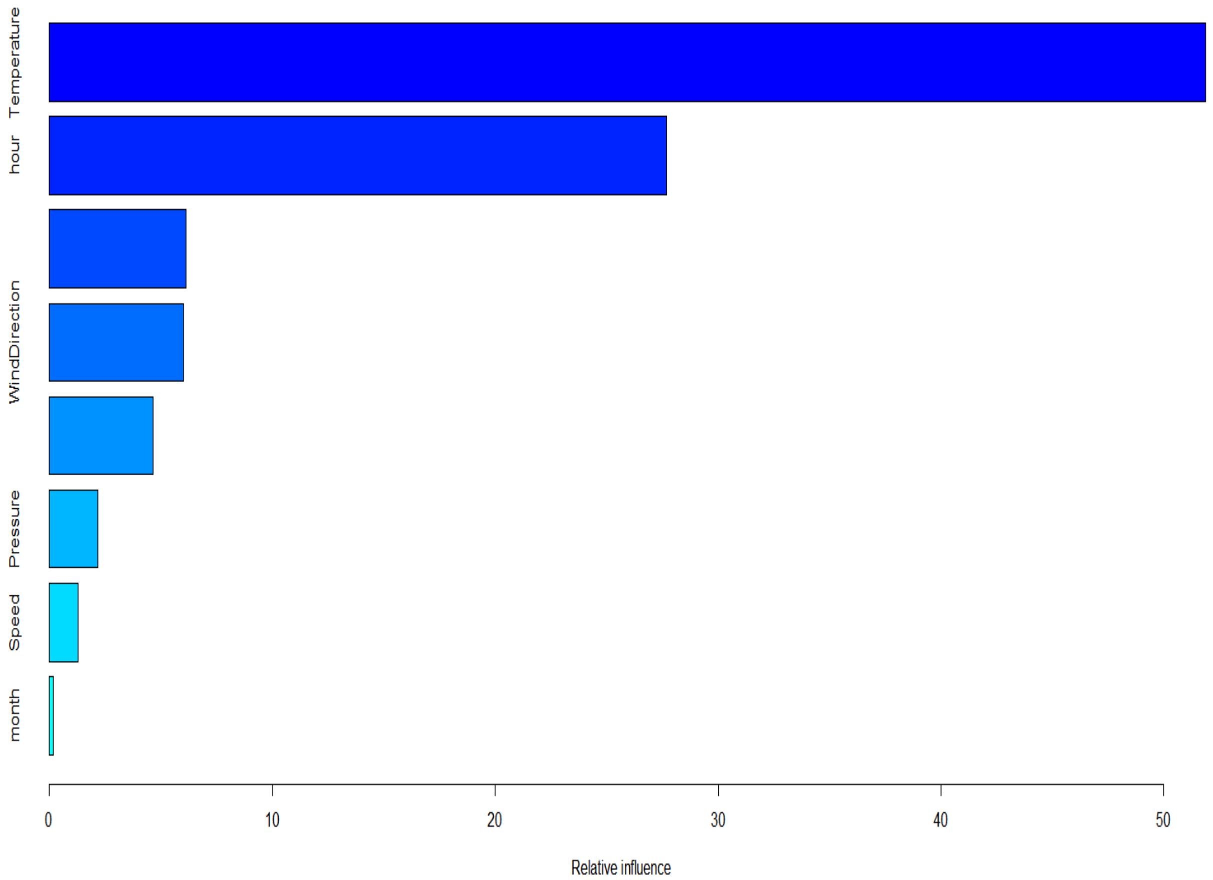


Pruned Tree with size 5



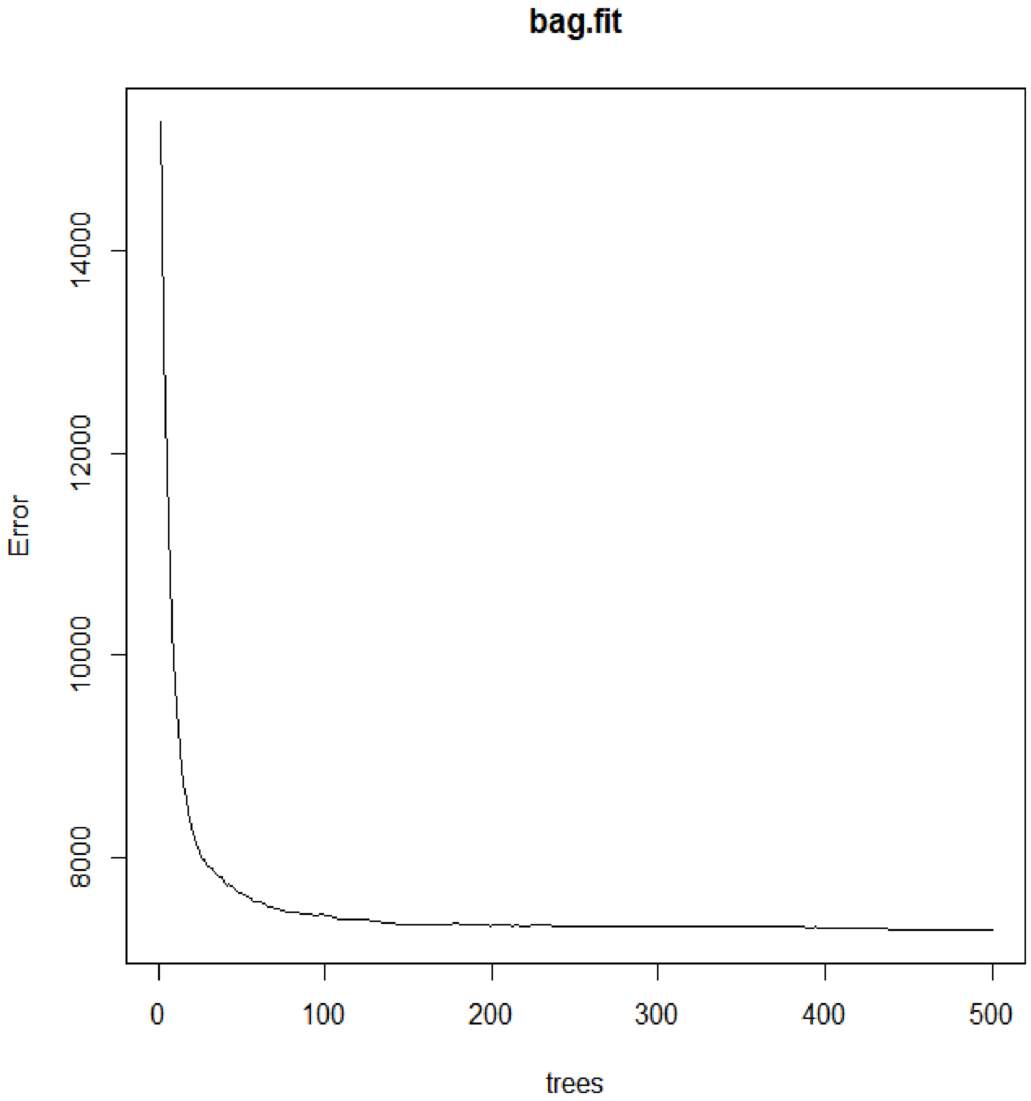
- Deviance (Train MSE for pruned regression tree) = 32160.
- Test MSE = 32849

Boosting

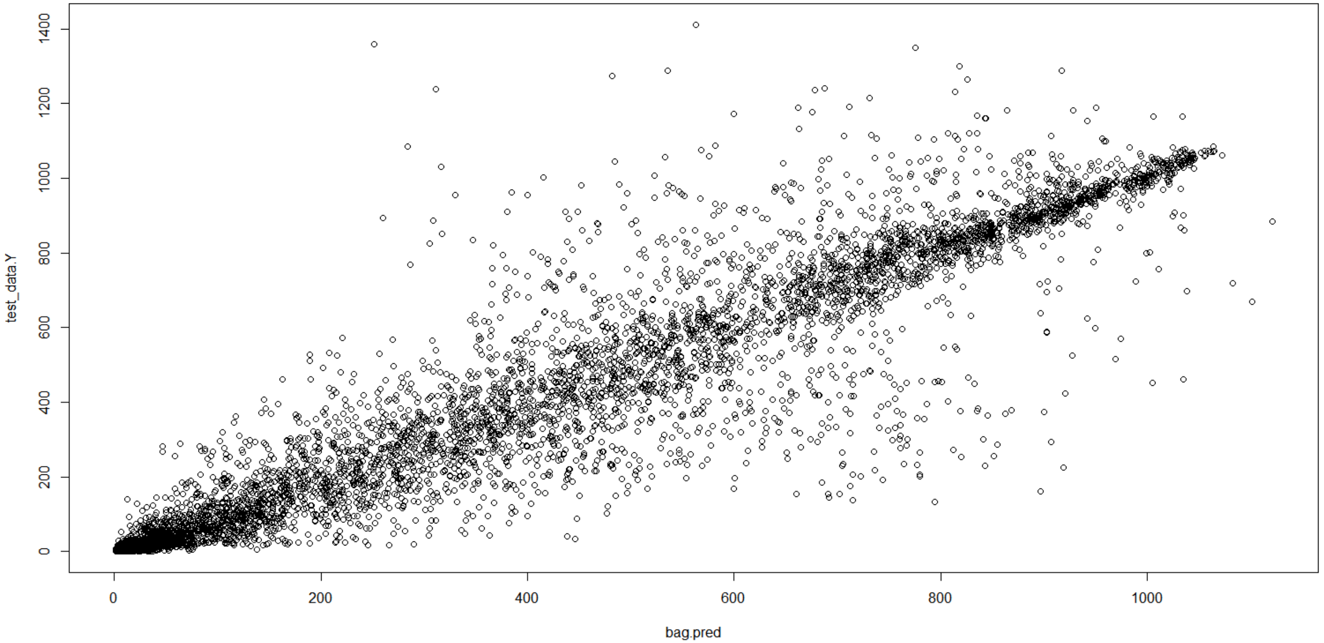


- Setting the number of trees to 3000, limiting the depth to each tree to 8 and shrinkage = 0.001, we achieved a MSE of 8673 on test data.
- Temperature and Hour are most important features.

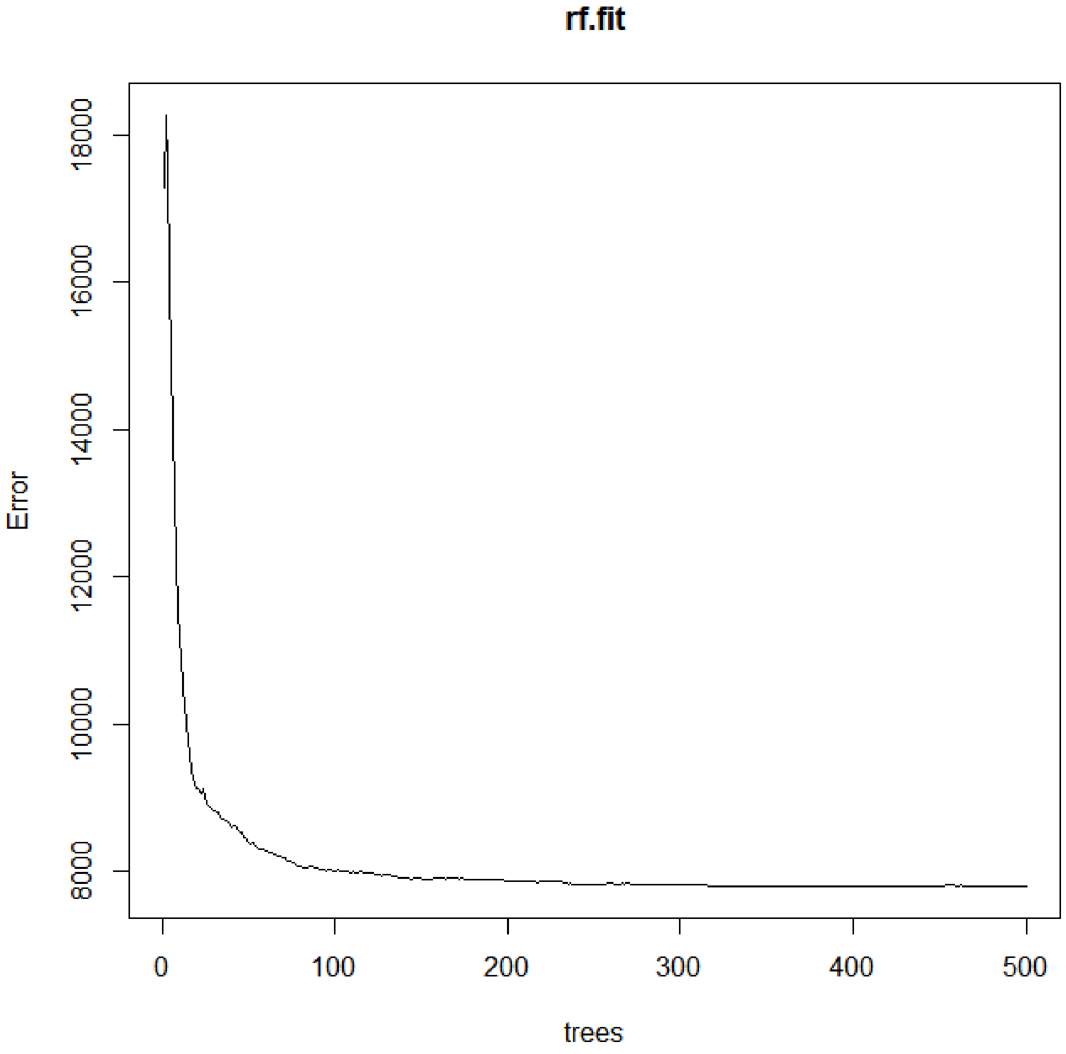
Bagging



- Bagging is simply a special case of a random forest where all predictors are considered for each split in the tree.
- There isn`t much improvement in performance as the number of trees is increased from 300 to 500
- MSE obtained on test data = 7752

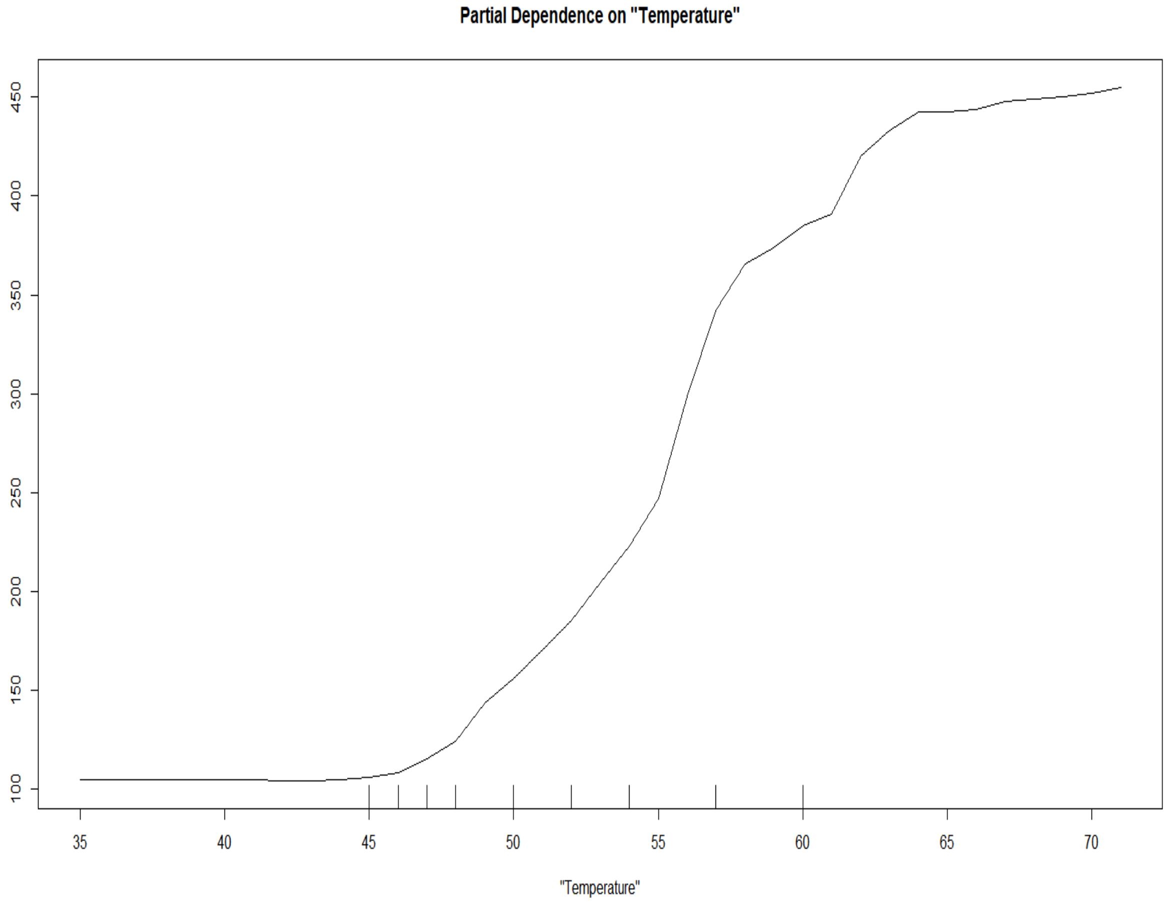


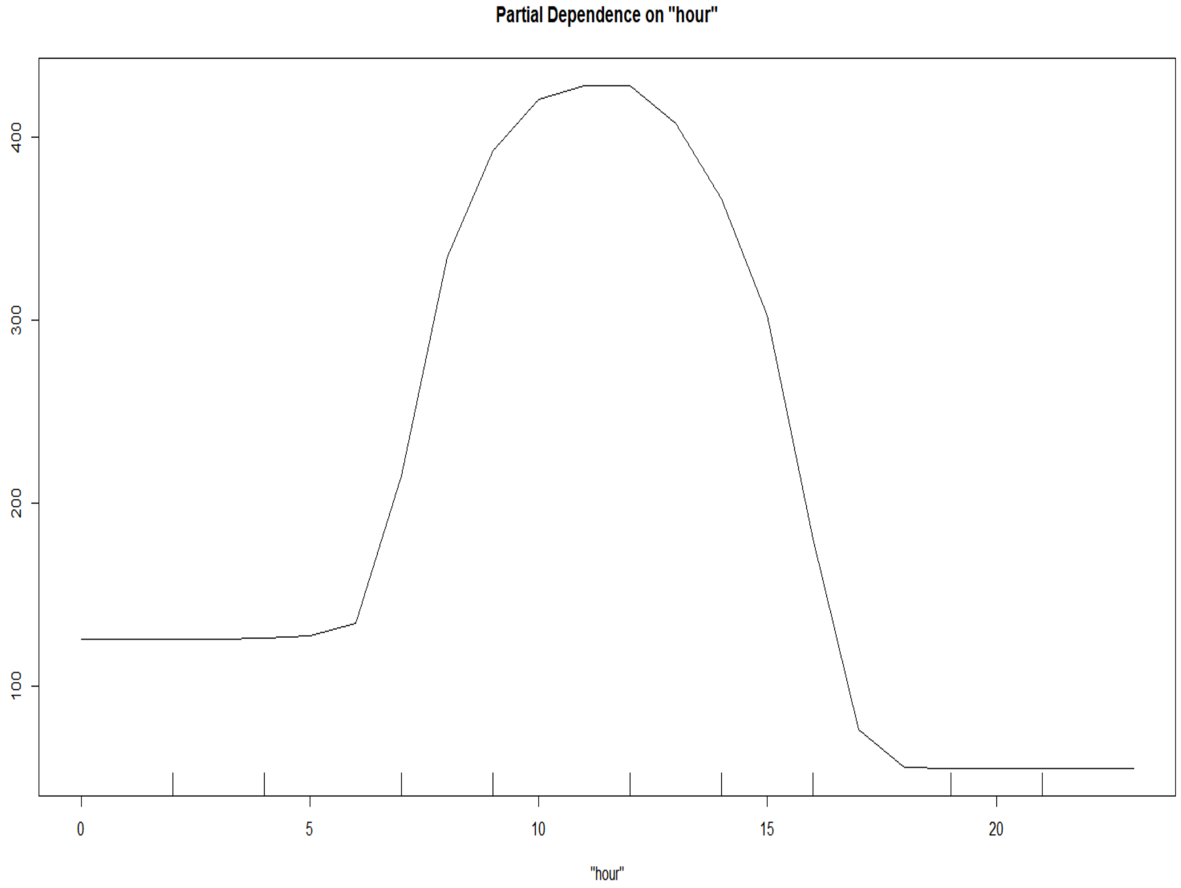
Random Forests



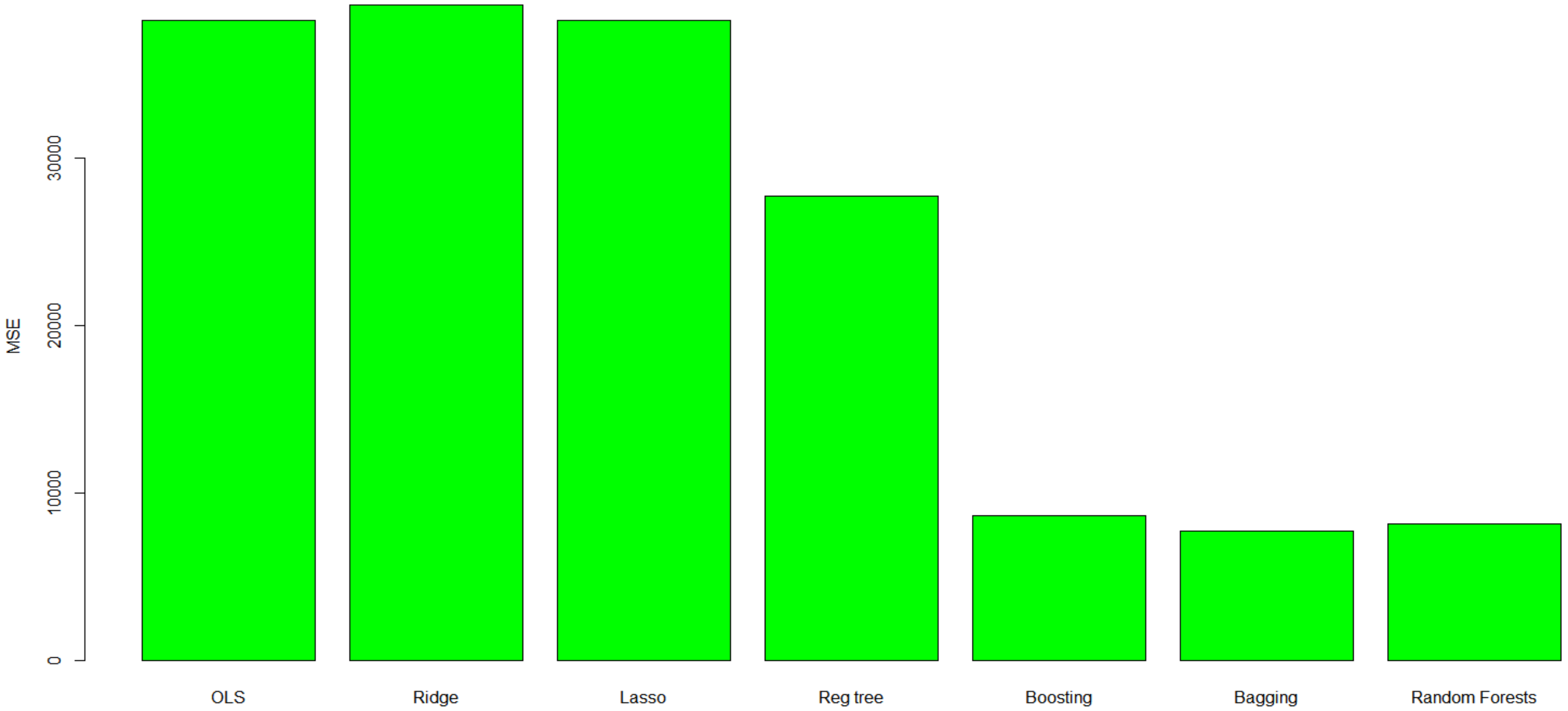
- There isn`t much improvement in performance as the number of trees is increased from 300 to 500
- MSE obtained on test data = 8180

Partial Dependence plots





Models Test MSE Comparison



Conclusion

- Taking MSE as a performance metric, ensemble methods like Bagging random forests and boosting performed well for our data when compared to non-ensemble models like linear regression, ridge and lasso.
- Including more features that correlate with solar radiation can improve the model performance.
- Including meteorological information from other months can generalize the model well for future data.