Regression Trees

Seminar Paper

Modern Methods in Statistics: Modern Data Sources for Statistical Indicators

Dr. Charlotte Articus

submitted by:

Siva Nanda Reddy Pokala
Universitätsring 8c
54296 Trier
Student number: 1519035
M.Sc. Data Science

Chaya Maskeri Subraya
Universitätsring 8f
54296 Trier
Student number: 1510125
M.Sc. Data Science

March 30, 2021

Contents

List of Figures **II**

1 Introduction **1**

2 Regression Trees **2**
2.1 Building Regression Tree 2
2.1.1 Case 1: One exogenous variable 2
2.1.2 Case 2: Multiple exogenous variable 4
2.2 Advantages and disadvantages of Regression Trees 5
2.2.1 Advantages 5
2.2.2 Disadvantages 5

3 Pruning **6**
3.1 Intuition 6
3.2 Pruning algorithms 6

4 Ensemble Methods **6**
4.1 Intuition 6
4.2 Types of ensemble methods 7

5 Random Forest **7**
5.1 Intuition 7
5.2 Advantages and Disadvantages 8
5.2.1 Advantages 8
5.2.2 Disadvantages 8

6 Working Example **9**
6.1 Dataset description 9
6.2 Approach 9
6.3 Hyper parameter Tuning with cross validation 10
6.4 Parameters to be tuned in Regression tree or Random forest 10

7 Conclusion **13**

List of Figures

1 Regression Tree 2
2 Case 1: Building Regression Tree with one feature 3
3 Case 1: Building Regression Tree with one feature 4
4 Case 2: Building Regression Tree with multiple features 5
5 Ensemble method 6
6 Random Forest 8
7 Summary of the dataset 9
8 mtry tuning 10
9 ntree tuning 11
10 maxnodes tuning 12

1 Introduction

In the practical world, when the data is non-linear, it is hard to predict endogenous variables using exogenous variables by linear models. In general, every independent variable has a separate and strictly additive effect on dependent variables irrespective of what the remaining variables are doing. It is also possible to include interactions but the problem is handling a massive number of features in the model.

Let's consider linear regression, where a single predictive formula holds over the entire data set. When the data has many features which interact in complicated, non-linear ways, assembling a single generalised model can be difficult. So an alternative approach to non-linear regression is to divide the space into smaller partitions where the interactions are more manageable. We can partition these sub-partitions recursively, which is the fundamental idea behind regression trees.

There is a saying, two heads are better than one. More than one regression trees can ensemble together to produce better results in random forest. Final prediction is produced after aggregating the results over the ensemble. Random forest algorithms can handle missing values in high-dimensional data and can handle continuous, categorical and binary data.This method is preferable, when small changes in the data have a huge impact on the outcome.

In this paper, we will describe briefly the theoretical aspects of regression trees, ensemble methods and random forests. Regression tree and random forest algorithms on the Combined Cycle Power Plant dataset had been applied to predict electrical power output. Finally, we compared the performance of both algorithms. Using cross validation, we found the optimal hyperparameters.

2 Regression Trees

Regression trees (1) are decision trees, where each leaf node represents the numerical value. Tree branches are formed because of nested if-else conditions. In contrast to regression trees, classification trees have binary labels or discrete labels as leaf nodes. Both have tree structure with a root node, intermediate nodes and leaf nodes. Plotting begins from the root node, ends at leaf node and in-between there are intermediate nodes.

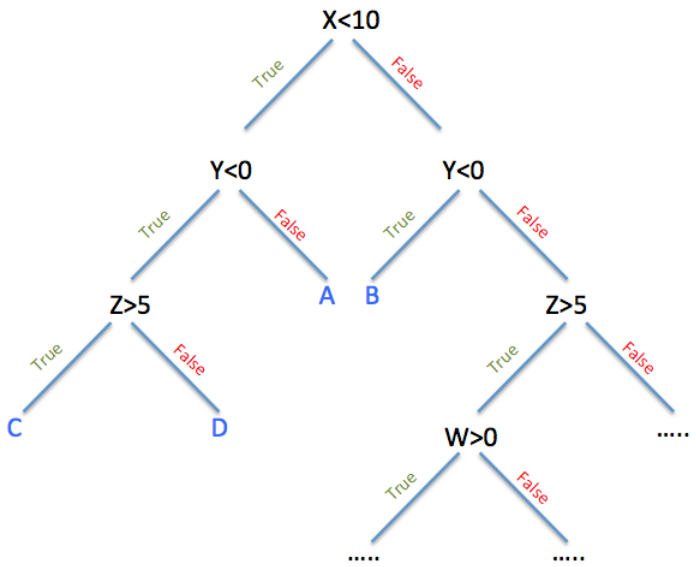

Figure 1: Regression Tree
Source: (5)

2.1 Building Regression Tree

To explain the process of building a regression tree (8), we will consider two cases. In the first case, let us assume, we have only one variable and in the second case, let us assume, we do have multiple variables.

2.1.1 Case 1: One exogenous variable

Basic idea is to divide the feature space into rectangular boxes. We will consider a scenario, where we need to predict the effectiveness of a drug based on dosage [1].

To decide the splitting condition, we will select the first observation of the data into one region, remaining all observations into another region. Later, calculate the sum of squared

[1] Example is considered from (12)

residuals(SSR) and note it.

$$\text{SSR} = \sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y_{R_j}}^2)$$

R is the number of regions in the space.
$\hat{y_{R_j}}$ is the mean of the particular region.
J is the number of observations in the region.

In the figure 2, the first observation is in one region, rests in another region and SSR is calculated.

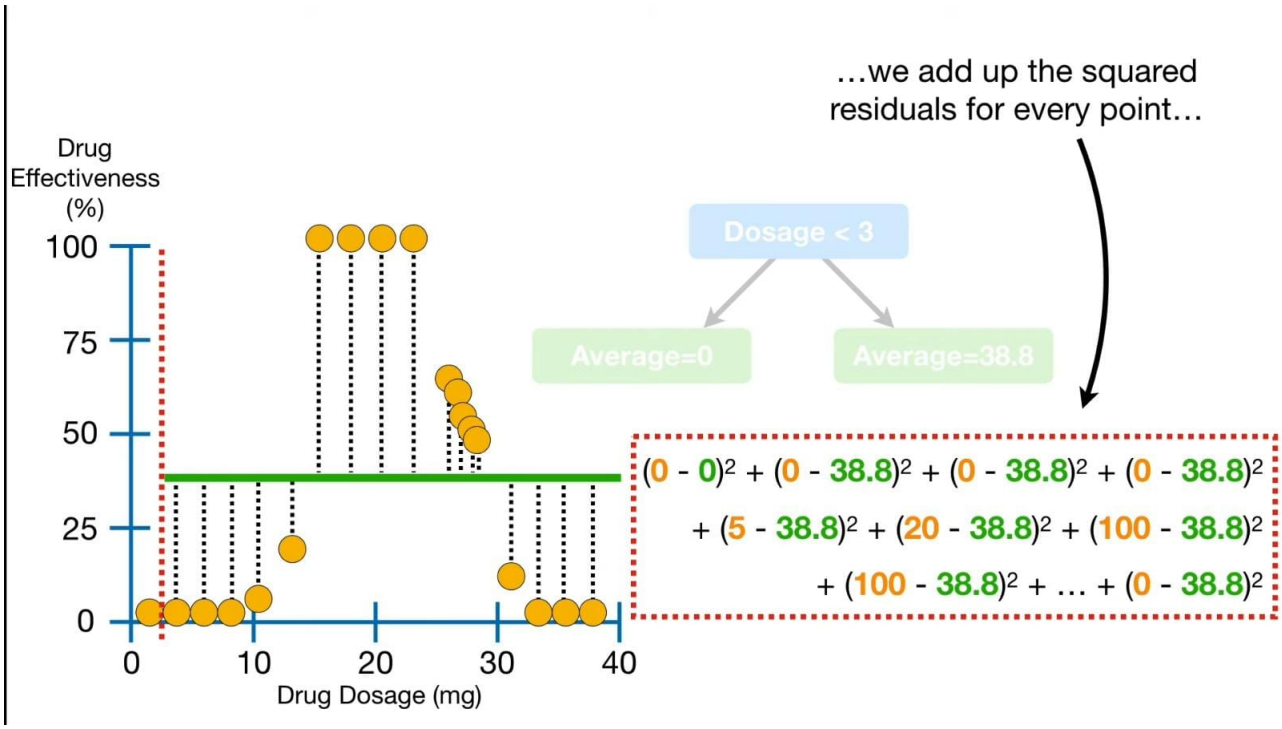

Figure 2: Case 1: Building Regression Tree with one feature
Source: (12)

This process will be repeated by adding one succeeding observation to the first region. In the figure 3, the second observation is added to the first region and rests in another region. Again note down the SSR. Repeat this process until we have only one observation in the second region and rests in the first region. Minimum SSR, decides the splitting condition of the space. Now, our single space is divided into two subspaces.

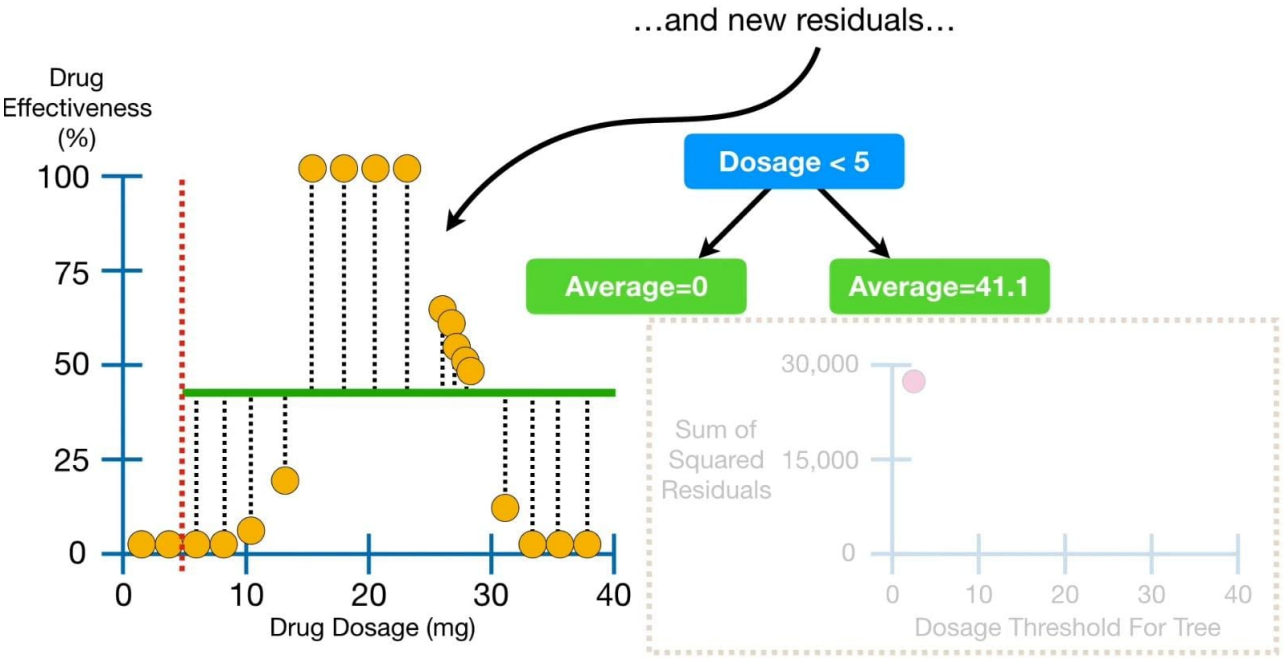

Figure 3: Case 1: Building Regression Tree with one feature
Source: (12)

To divide the subspaces into smaller units, above explained process should be repeated. To avoid overfitting, select optimal value for number of samples in a region. Split the region, if number of observations are greater than the pre-selected optimal value in the region.

2.1.2 Case 2: Multiple exogenous variable

Let us consider, we need to predict the effectiveness of a drug not only based on Dosage also on Age and Sex. For each individual variable, we need to follow the Case 1 process. Whichever feature has minimum SSR, will become the root node. Figure 4 shows, among all three features, Age has the lowest SSR. So we should consider Age as first, Dosage as second and Sex as the last priority.

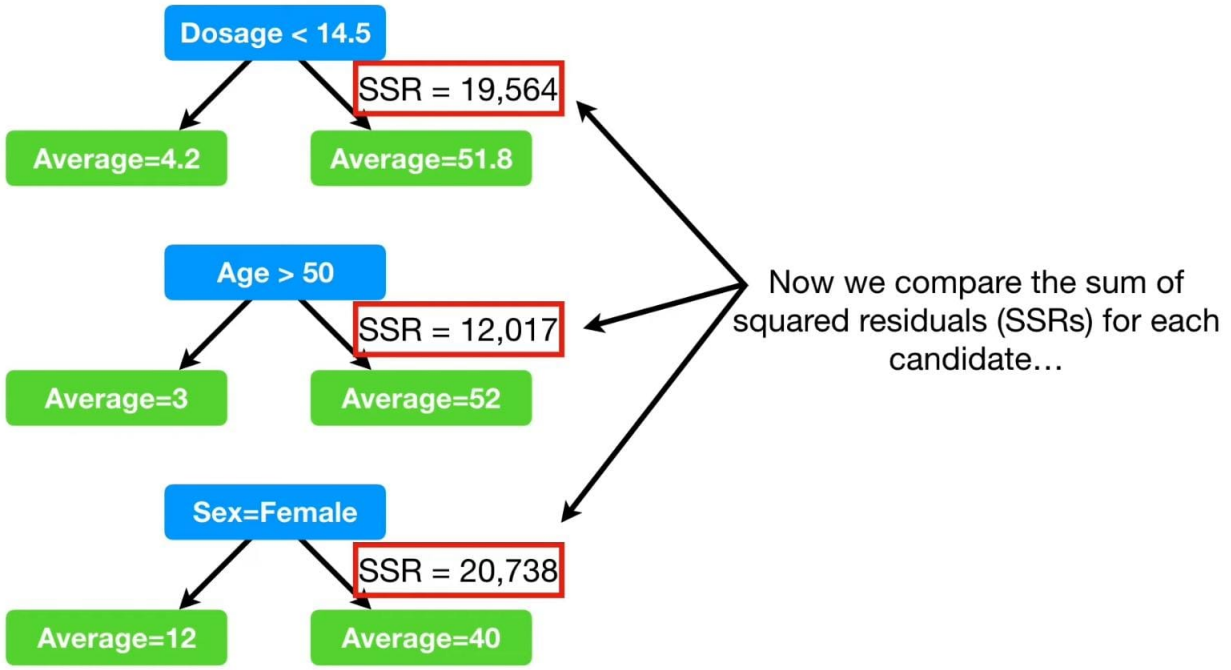

Figure 4: Case 2: Building Regression Tree with multiple features
Source: (12)

2.2 Advantages and disadvantages of Regression Trees

2.2.1 Advantages

1. Interpretation of the results are easy: It allows the rapid prediction of the new observation.
2. Non-linear problems can be addressed effortlessly.
3. Feature selection is automatic in regression trees. We don't need to perform it explicitly.

2.2.2 Disadvantages

1. Overfitting: We may face the problem of overfitting as regression trees consider noise that exists in the data for prediction. This leads to incorrect results. We will shortly describe, how we can avoid this problem in the next section.
2. High variance: If there is small change in the data, we might face high variance in the predicted values.

3 Pruning

3.1 Intuition

Pruning is a technique which is used to avoid overfitting by removing certain parts of the tree that are non-critical and redundant.

3.2 Pruning algorithms

There are different types (10) of pruning. We will shortly define below algorithms (7).

1. Pre-pruning: In pre-pruning, before constructing the tree, a stopping criterion is set. So the tree will grow only until the stopping condition is met. Examples of stopping criterion are maximum depth of the tree, information gain of a particular attribute $<$ threshold.
2. Post-pruning: In this case, we build a complete tree and later remove the branches which are not necessary. This method is computationally expensive.
3. Bottom-up pruning: Here, pruning begins at the last node and follows upward direction. If a node is not relevant, it will be replaced by a leaf node or simply dropped.
4. Top-down pruning: Unlike bottom-up pruning, this method begins from the root node. There is a chance of dropping the entire subtree, if the intermediate node is not relevant.

4 Ensemble Methods

4.1 Intuition

Ensemble methods are the technique, which combines a group of base models to make more accurate predictions than a single base model.

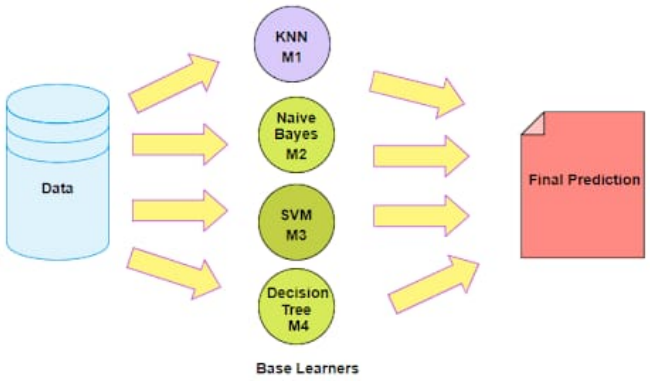

Figure 5: Ensemble method
Source: (6)

4.2 Types of ensemble methods

1. Bagging(Bootstrap Aggregation): When we have high variance and low bias models, we can make use of the bagging technique (2) to reduce the variance.

   Generalization error = $bias^2$+variance + irreducible error

   Bagging results in a low bias and low variance model. In turn the generalization error of the model will be reduced. In Bagging, each model will take a sample with replacement from the original data(bootstrap sampling method). Final prediction in regression will be the average prediction of all models and in classification it will be the majority vote. Example of Bagging is random forest.

2. Boosting: In order to reduce the bias in a model, we can make use of boosting. Here, each individual model is dependent on the previous model and the execution order is sequential (11). Examples of Boosting are XGBoost, Ada-Boost.

3. Stacking: It is technique, where multiple classification models of different types are joined together to form a meta classifier (14). If we are interested to increase the prediction accuracy of a model, we can make use of stacking.

4. Cascading: In cascading a sequence of models are built to make sure about the prediction accuracy. Normally, the complexity of the model is high. Mainly used in scenarios, where even small mistakes in prediction are not acceptable.[2]

5 Random Forest

5.1 Intuition

Random forest (4) consists of a group of decision trees, which are independent and run in parallel. This technique can be used for both classification and regression problems. Using random sampling with replacement each training example is drawn from the original dataset. Later, a tree is grown without pruning on the sampled data using random feature selection. By this method, the depth of each tree is reduced, which leads to less variance.

As we consider subset of features in random forest, trees grown from different bootstrap sample will be different. Even if we consider two trees from same sample, they will likely differ. So the correlation in prediction is low compared to bagging.

Bagging and random feature selection are resistant to the noise in the data. Even if the outliers or missing values exist in the data, random forest shows small changes (4).

[2] This paper has information on Cascaded Ensemble of Convolutional Neural Networks in medical field (15)

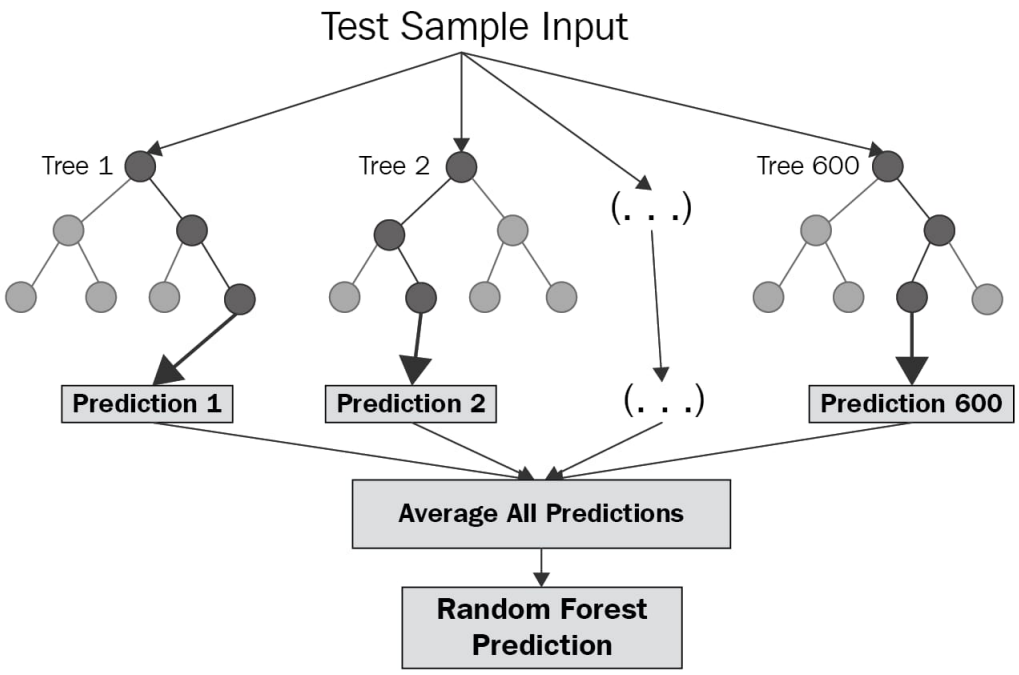

Figure 6: Random Forest
Source: (6)

5.2 Advantages and Disadvantages

5.2.1 Advantages

1. Overfitting in the decision tree can be reduced and prediction can be improved.
2. Flexible both for classification and regression problems.
3. Normalisation of data is not necessary as it uses rule based approach.

5.2.2 Disadvantages

1. It requires much computational power because it constructs multiple number of trees and averages their output.
2. More time is required for training the dataset.

6 Working Example

6.1 Dataset description

We are interested in predicting full load electrical power output based on hourly average of Ambient pressure(AP), Relative Humidity(RH), Exhaust Vacuum (EV) and Ambient Temperature(AT). The data set has 9568 observations recorded from a Combined Cycle Power Plant over 6 years(2006 - 2011) [3]. The dataset is not having any missing values. So, imputation is not necessary.

```
> summary(data)
       AT               V               AP               RH              PE
 Min.   : 1.81   Min.   :25.36   Min.   : 992.9   Min.   : 25.56   Min.   :420.3
 1st Qu.:13.51   1st Qu.:41.74   1st Qu.:1009.1   1st Qu.: 63.33   1st Qu.:439.8
 Median :20.34   Median :52.08   Median :1012.9   Median : 74.97   Median :451.6
 Mean   :19.65   Mean   :54.31   Mean   :1013.3   Mean   : 73.31   Mean   :454.4
 3rd Qu.:25.72   3rd Qu.:66.54   3rd Qu.:1017.3   3rd Qu.: 84.83   3rd Qu.:468.4
 Max.   :37.11   Max.   :81.56   Max.   :1033.3   Max.   :100.16   Max.   :495.8
```

Figure 7: Summary of the dataset

6.2 Approach

Dataset is split into two parts: train and test. Training split has 80% of the observation and the testing split has 20% of the observation. Feature importance is in the below order.

1. Ambient Temperature
2. Exhaust Vacuum
3. Ambient pressure
4. Relative Humidity

We have applied the decision tree regressor and random forest on the training dataset separately and predicted the testing dataset. RMSE and Rsquare are noted.
We cross checked prediction error of training and test data separately and concluded that there is no overfitting problem in this scenario as the prediction error in training and testing don't have much difference.

Algorithm	RMSE_Train	RMSE_Test	Rsquare_Train	Rsquare_Test
Decision Tree	4.47	4.67	0.93	0.92
Random Forest	3.34	3.43	0.96	0.95

[3] Dataset is considered from https://archive.ics.uci.edu/ml/datasets/combined+cycle+power+plant

6.3 Hyper parameter Tuning with cross validation

Hyper parameter decides whether we encounter an overfitting or underfitting problem. So, we need to find the parameters which can predict unseen data well.

To optimize the performance of the model, we make use of hyper parameter tuning. The best hyperparameter is determined by trial and error method. To evaluate the performance of the model, different combinations can be tried and best will be selected. Using cross validation, we can achieve hyperparameter tuning. In cross validation, each data point participates in training and testing, but not at the same time. If we are not choosing cross validation for hyperparameter tuning, the optimal parameter might be suitable only for training, and can't be generalized to testing.

6.4 Parameters to be tuned in Regression tree or Random forest

Performance of the tree models depends on below parameters. Tuning of these parameters results in better performance of the model.

1. mtry: Number of features randomly selected : It is considered only for random forest. In our example the best mtry value is 2 as it is having the smallest MAE value.

```
mtry  RMSE      Rsquared   MAE
1     3.606522  0.9560939  2.661440
2     3.425840  0.9601062  2.479050
3     3.483789  0.9587180  2.517438
4     3.528282  0.9576806  2.538983

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was mtry = 2.
```

Figure 8: mtry tuning

2. ntree: Number of trees: Stable models are produced by using a large number of trees, but it requires large amount of memory and time. When we tuned ntree parameter by considering mtry as 2, it gave a minimum MAE at 80.

```
Models: 60, 70, 80, 90, 100
Number of resamples: 5

MAE
        Min.  1st Qu.   Median     Mean  3rd Qu.     Max. NA's
60  3.133787 3.170632 3.212613 3.224821 3.223044 3.384028    0
70  3.127239 3.164521 3.209388 3.223781 3.230905 3.386855    0
80  3.120988 3.170730 3.209733 3.221997 3.228182 3.380353    0
90  3.121176 3.174870 3.209337 3.222298 3.221086 3.385018    0
100 3.123690 3.181348 3.206966 3.222676 3.213768 3.387610    0

RMSE
        Min.  1st Qu.   Median     Mean  3rd Qu.     Max. NA's
60  3.980058 4.091996 4.117902 4.200850 4.174968 4.639327    0
70  3.972903 4.088146 4.114077 4.200506 4.178700 4.648705    0
80  3.979217 4.092134 4.108487 4.199032 4.175557 4.639764    0
90  3.988174 4.088948 4.108721 4.199502 4.168372 4.643297    0
100 3.994043 4.089026 4.109165 4.199091 4.160475 4.642745    0

Rsquared
         Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
60  0.9280629 0.9400510 0.9428223 0.9405270 0.9429382 0.9487606    0
70  0.9277883 0.9399305 0.9429107 0.9405242 0.9430197 0.9489718    0
80  0.9281004 0.9400407 0.9427900 0.9406007 0.9432014 0.9488711    0
90  0.9279837 0.9402462 0.9428947 0.9406037 0.9432100 0.9486839    0
100 0.9280099 0.9404622 0.9429214 0.9406366 0.9432070 0.9485822    0
```

Figure 9: ntree tuning

3. maxnodes: If this parameter is not given, trees will be grown to the fullest. Otherwise, it specifies, maximum number of terminal node in the tree. In our example, as we increased the maxnodes parameter, we saw better results and selected 1500 as the final value.

```
Models: 1500, 1501, 1502, 1503, 1504, 1505
Number of resamples: 5

MAE
         Min.  1st Qu.   Median     Mean  3rd Qu.    Max. NA's
1500 2.364152 2.386955 2.467974 2.473859 2.553694 2.59652    0
1501 2.364152 2.386955 2.467974 2.473859 2.553694 2.59652    0
1502 2.364152 2.386955 2.467974 2.473859 2.553694 2.59652    0
1503 2.364152 2.386955 2.467974 2.473859 2.553694 2.59652    0
1504 2.364152 2.386955 2.467974 2.473859 2.553694 2.59652    0
1505 2.364152 2.386955 2.467974 2.473859 2.553694 2.59652    0

RMSE
         Min. 1st Qu.   Median     Mean  3rd Qu.     Max. NA's
1500 3.231473 3.28826 3.303986 3.418287 3.578353 3.689364    0
1501 3.231473 3.28826 3.303986 3.418287 3.578353 3.689364    0
1502 3.231473 3.28826 3.303986 3.418287 3.578353 3.689364    0
1503 3.231473 3.28826 3.303986 3.418287 3.578353 3.689364    0
1504 3.231473 3.28826 3.303986 3.418287 3.578353 3.689364    0
1505 3.231473 3.28826 3.303986 3.418287 3.578353 3.689364    0

Rsquared
          Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
1500 0.9547579 0.9568032 0.9621401 0.9602847 0.9634038 0.9643184    0
1501 0.9547579 0.9568032 0.9621401 0.9602847 0.9634038 0.9643184    0
1502 0.9547579 0.9568032 0.9621401 0.9602847 0.9634038 0.9643184    0
1503 0.9547579 0.9568032 0.9621401 0.9602847 0.9634038 0.9643184    0
1504 0.9547579 0.9568032 0.9621401 0.9602847 0.9634038 0.9643184    0
1505 0.9547579 0.9568032 0.9621401 0.9602847 0.9634038 0.9643184    0
```

Figure 10: maxnodes tuning

After hyperparameter tuning, below optimal parameters are considered as part of final model.

1. mtry = 2
2. ntree = 80
3. maxnodes = 1500

Algorithm	RMSE_Train	RMSE_Test	Rsquare_Train	Rsquare_Test
Random forest(parameters tuned)	2.07	3.42	0.99	0.96

7 Conclusion

Regression trees and random forest are preferable tools for classification and regression tasks. In our analysis of a regression tree and random forest, we have experienced approximately the same result as the dataset is small and is not having many features. In general, when we have a huge number of features and large dataset, random forest performs well, because it uses random features with boosting [4]. After Hyper parameter tuning, we have not experienced much improvement in the result. We conclude that hyper parameter tuning was not necessary for our dataset. Nevertheless, tuning and obtaining optimal parameters are recommended for large datasets.

[4] L. BREIMAN has made a point of this in the conclusion part of (4)

References

[1] Jehad Ali, Rehanullah Khan, Nasir Ahmad, and Imran Maqsood. Random forests and decision trees. *International Journal of Computer Science Issues (IJCSI)*, 9(5):272, 2012.

[2] Naomi Altman and Martin Krzywinski. Points of significance: Ensemble methods: bagging and random forests, 2017.

[3] Richard A. Berk. An introduction to ensemble methods for data analysis. *Sociological Methods & Research*, 34(3):263–295, 2006.

[4] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[5] Michele Cavaioni. "machine learning: Decision tree classifier". `https://medium.com/machine-learning-bites/machine-learning-decision-tree-classifier-9eb67cad263e`, 2017. [Online; accessed 19-February-2021].

[6] Afroz Chakure. "random forest regression". `https://medium.com/swlh/random-forest-and-its-implementation-71824ced454f`, 2019. [Online; accessed 21-February-2021].

[7] Johannes Fürnkranz. Pruning algorithms for rule learning. *Machine learning*, 27(2):139–172, 1997.

[8] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112, pages 310–314. Springer, 2013.

[9] Zardad Khan, Asma Gul, Aris Perperoglou, Miftahuddin Miftahuddin, Osama Mahmoud, Werner Adler, and Berthold Lausen. Ensemble of optimal trees, random forest and random projection ensemble classification. *Advances in Data Analysis and Classification*, 14(1):97–116, 2020.

[10] John Mingers. An empirical comparison of pruning methods for decision tree induction. *Machine learning*, 4(2):227–243, 1989.

[11] David Opitz and Richard Maclin. Popular ensemble methods: An empirical study. *Journal of artificial intelligence research*, 11:169–198, 1999.

[12] Josh Starmer. "regression trees, clearly explained!!!". `https://www.youtube.com/watch?v=g9c66TUylZ4&ab_channel=StatQuestwithJoshStarmer`, 2019. [Online; accessed 08-March-2021].

[13] Daniell Toth and John L. Eltinge. Building consistent regression trees from complex sample data. *Journal of the American Statistical Association*, 106(496):1626–1636, 2011.

[14] Anurag Kumar Verma and Saurabh Pal. Prediction of skin disease with three different feature selection techniques using stacking ensemble method. *Applied biochemistry and biotechnology*, pages 1–20, 2019.

[15] Haibo Wang, Angel Cruz-Roa, Ajay Basavanhally, Hannah Gilmore, Natalie Shih, Mike Feldman, John Tomaszewski, Fabio Gonzalez, and Anant Madabhushi. Cascaded ensemble of convolutional neural networks and handcrafted features for mitosis detection. In *Medical Imaging 2014: Digital Pathology*, volume 9041, page 90410B. International Society for Optics and Photonics, 2014.
(9), (3), (13), (12), (5), (6), (2) (1) (4) (8) (11) (14)