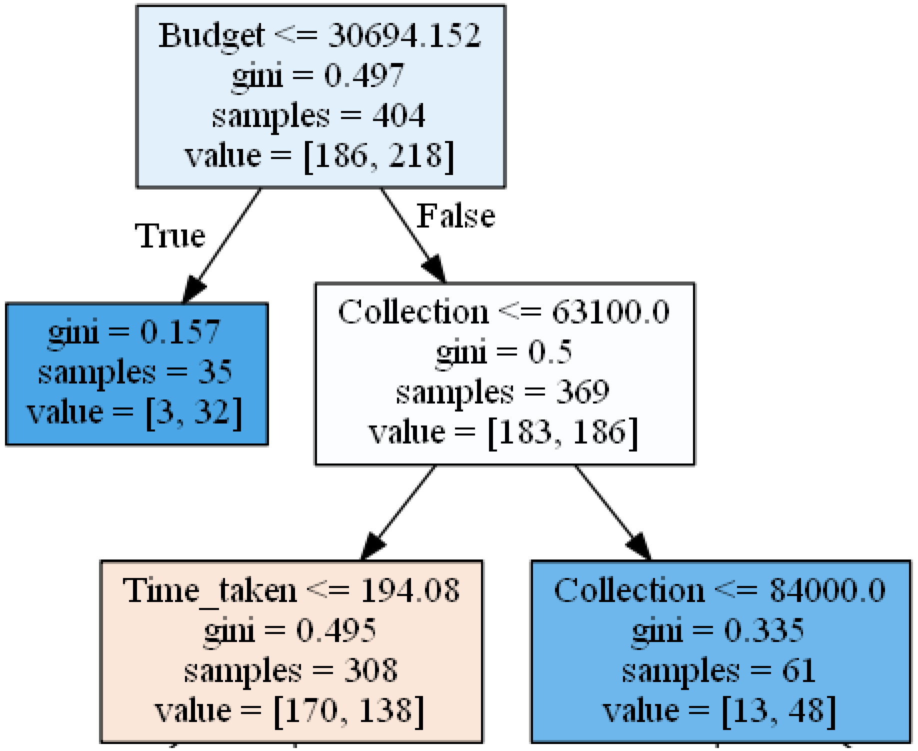
Decision Trees

Decisions Trees are the Most Popular technique of Machine learning

But why?

1. Simplicity



Decision Trees

Decisions Trees are the Most Popular technique of Machine learning

But why?

2. Accuracy

Decision Trees

Course Structure

- Crash Course
- Machine Learning Basics
- Decision Tree Basics
- Decision Tree Advanced

Decision Trees

Topics Covered

Simple Decision Trees

- Classification Trees
- Regression Trees
- Tree Pruning

Advanced Ensemble techniques

- Bagging
- Random Forest
- Gradient Boosting
- ADA Boost
- XG Boost

Basics

Regression vs Classification

Chart is a visual representation of numerical data. It can make your number more representable.

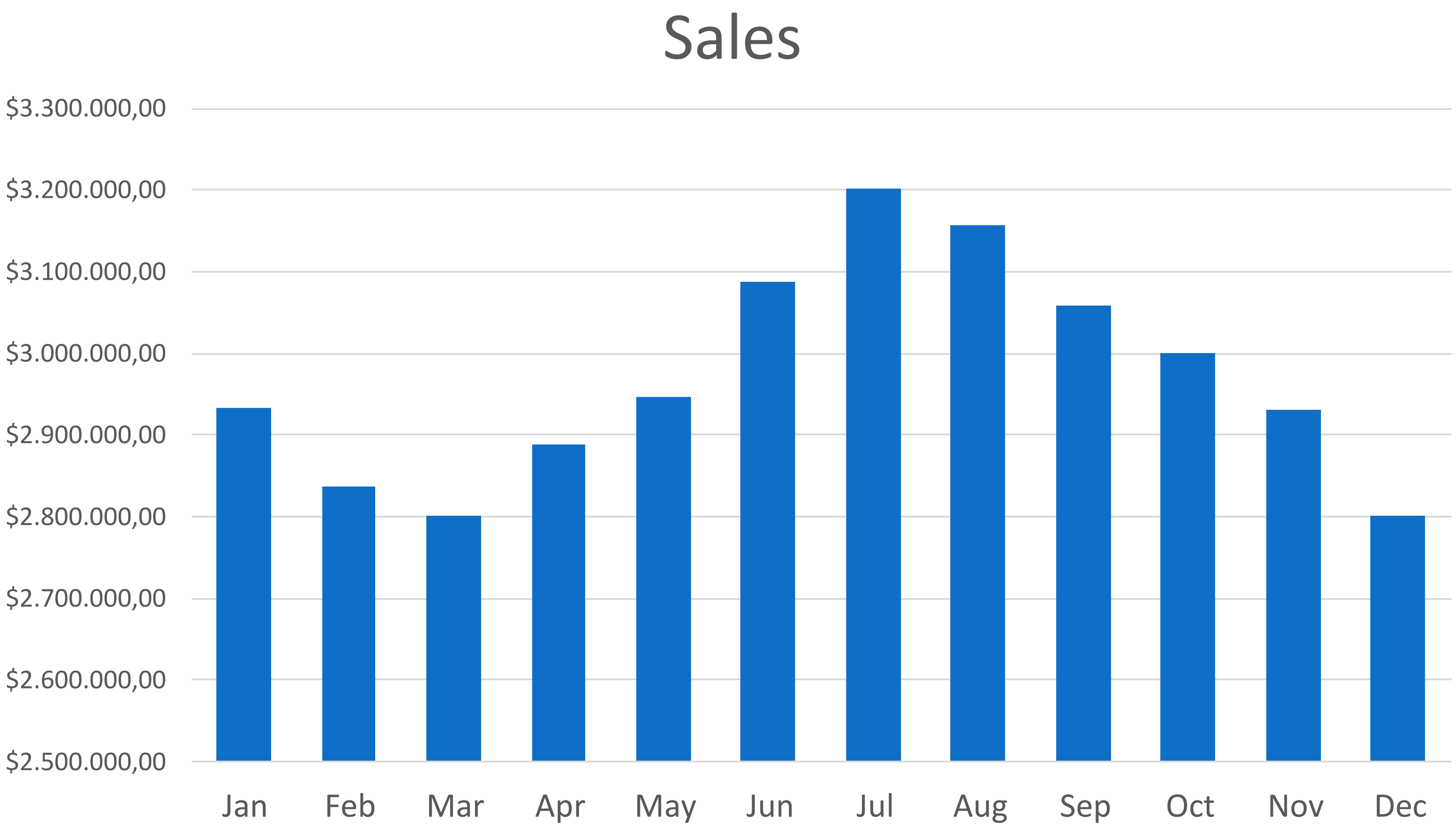
QUANTITATIVE DATA

Bias variance Tradeoff

Month	Sales
Jan	$ 2,933,743.00
Feb	$ 2,836,435.00
Mar	$ 2,799,982.00
Apr	$ 2,888,563.00
May	$ 2,945,629.00
Jun	$ 3,087,680.00
Jul	$ 3,202,347.00
Aug	$ 3,156,729.00
Sep	$ 3,057,932.00
Oct	$ 3,000,123.00
Nov	$ 2,930,987.00
Dec	$ 2,801,240.00

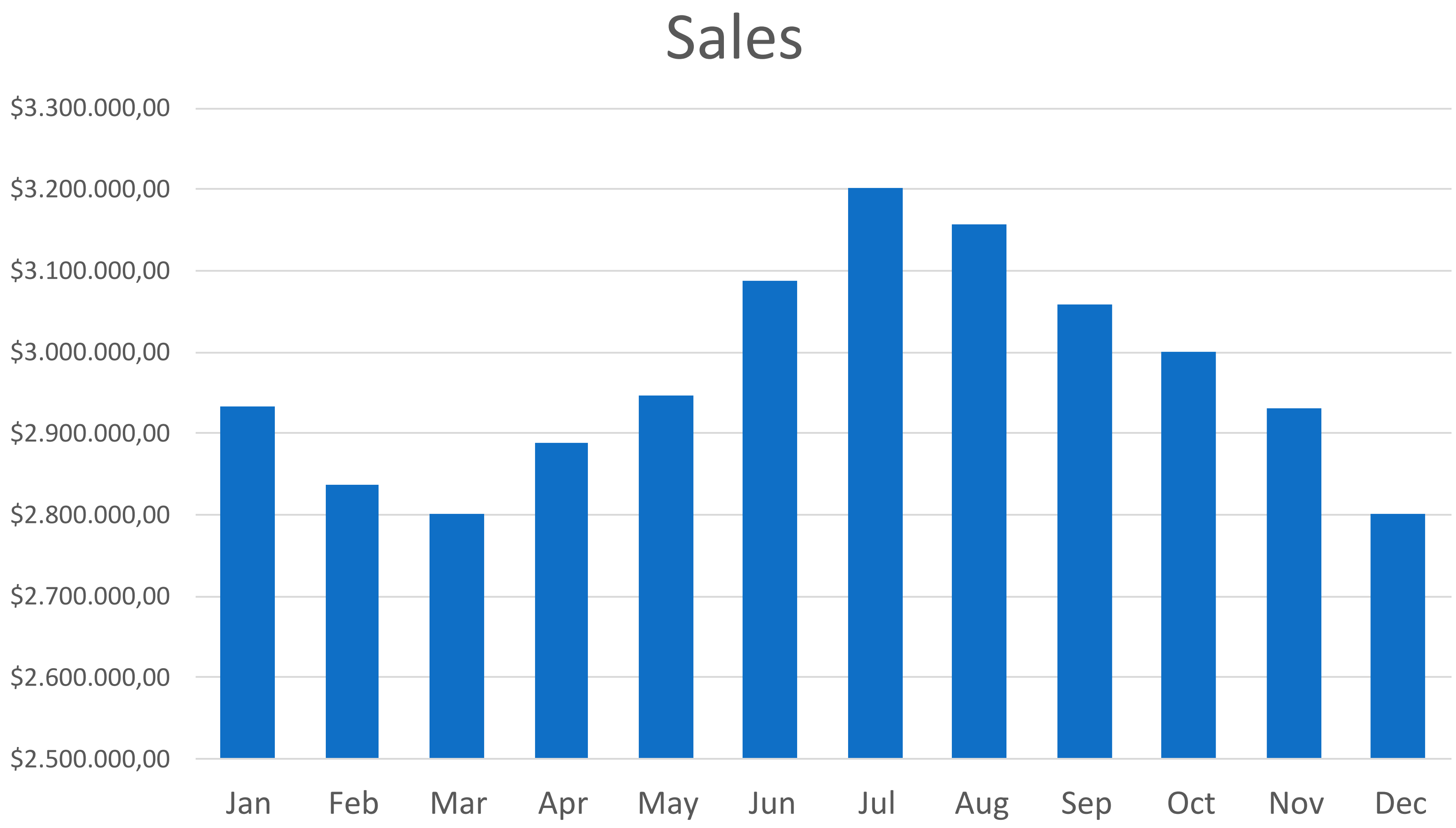
QUANTITATIVE DATA

OLS Method



QUANTITATIVE DATA

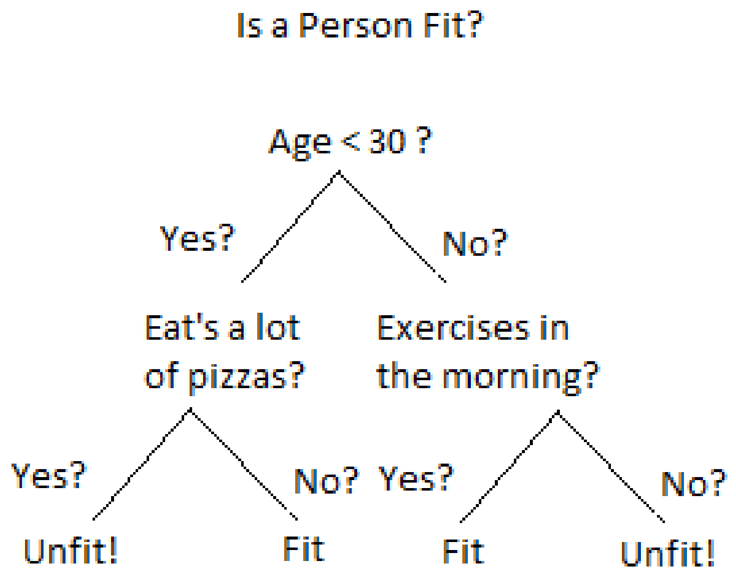
Overfitting



Decision Trees - Basics

Definition

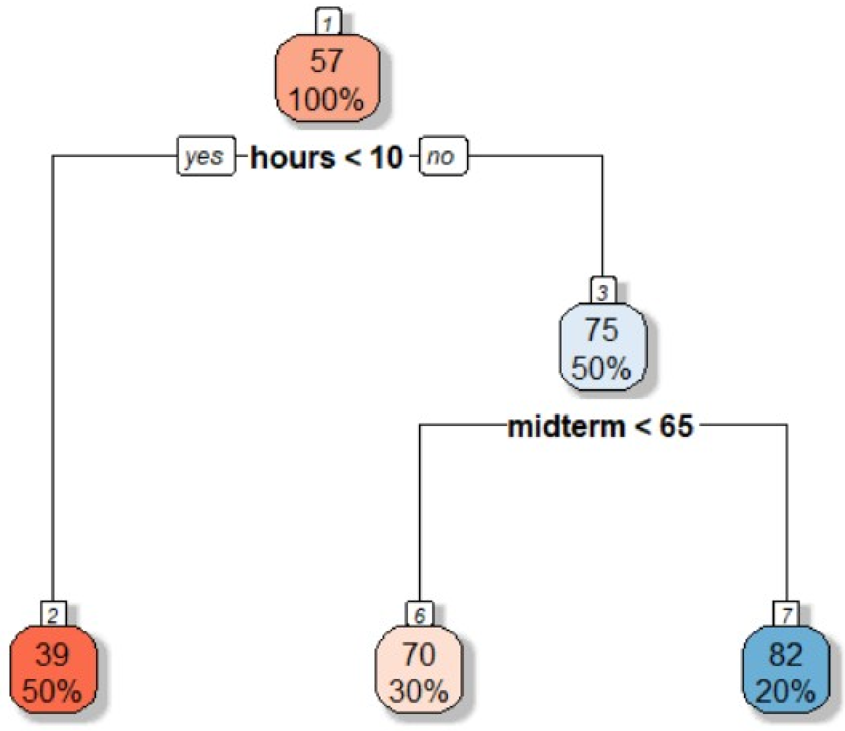
A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.



Decision Trees - Basics

Example

	score	hours	midterm
1	35	6	42
2	38	5	65
3	40	7	35
4	45	6	75
5	35	8	60
6	65	11	50
7	70	12	45
8	75	18	40
9	80	14	80
10	85	12	82



Decision Trees - Basics

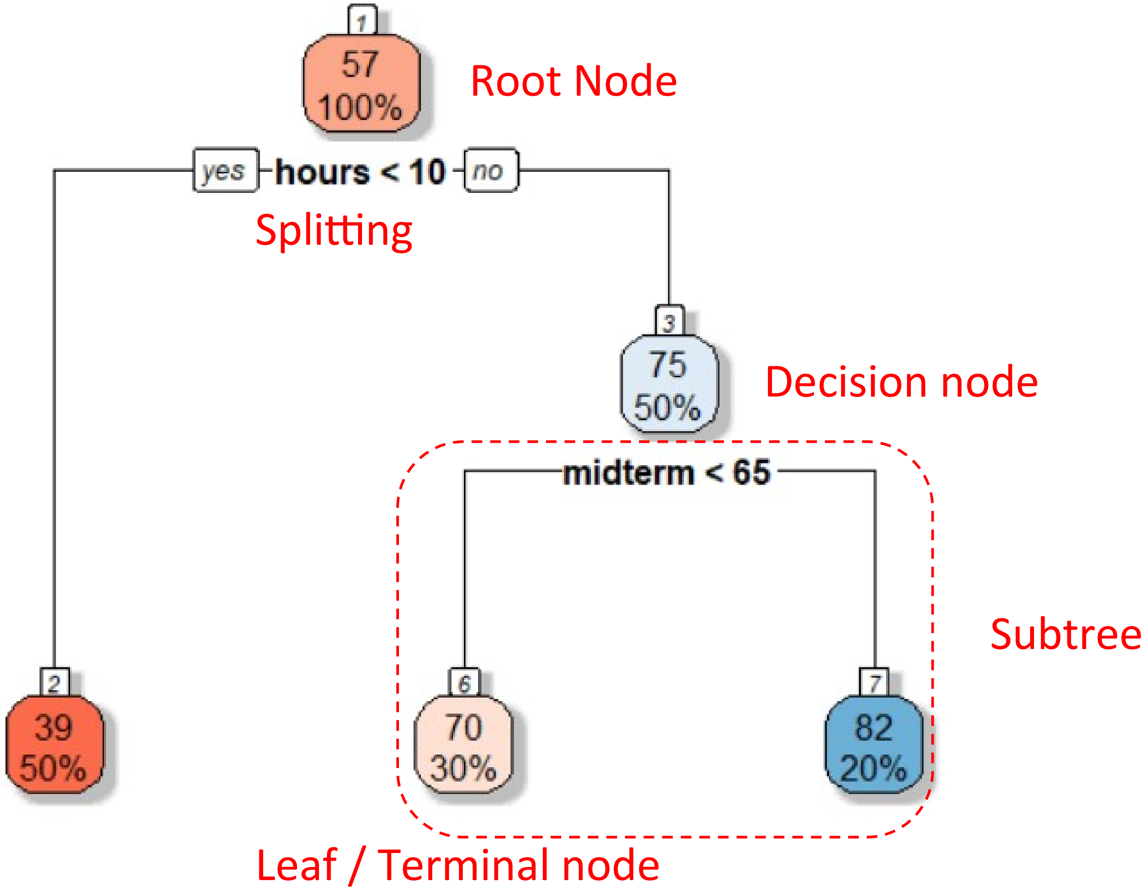
Types

1. Regression Tree
   For continuous quantitative target variable.
   Eg. Predicting rainfall, predicting revenue, predicting marks etc.

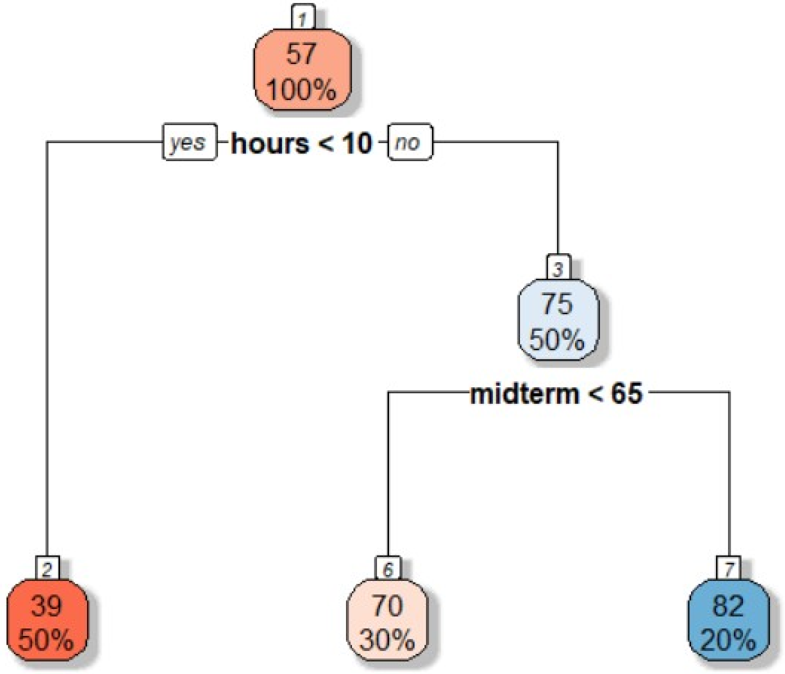
2. Classification Tree
   For discrete categorical target variables
   Eg. Predicting High or Low, Win or Loss, Healthy or Unhealthy etc

Decision Trees - Basics

Terminologies



Steps



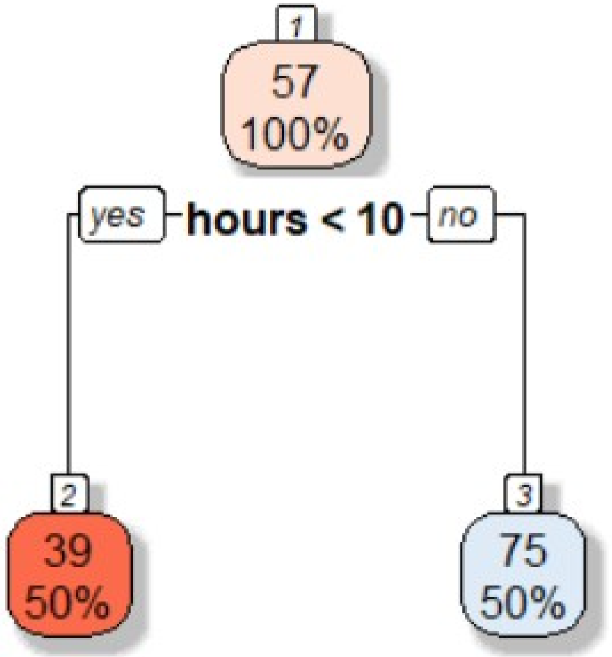
1. We divide the predictor space—that is, the set of possible values for X1,X2, . . .,Xp—into J distinct and non-overlapping regions, R1,R2, . . . , RJ .

2. For every observation that falls into the region Rj, we make the same prediction, which is simply the mean of the response values for the training observations in Rj .

Goal is to minimize RSS

$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2$$

Building tree



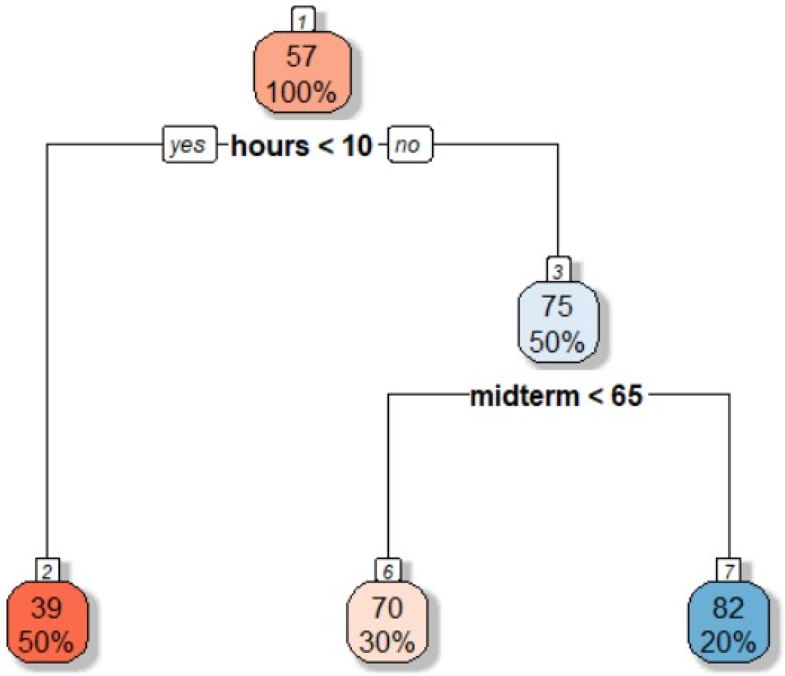
	score	hours	midterm
1	35	6	42
2	38	5	65
3	40	7	35
4	45	6	75
5	35	8	60
6	65	11	50
7	70	12	45
8	75	18	40
9	80	14	80
10	85	12	82

Mean score 39

$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2$$

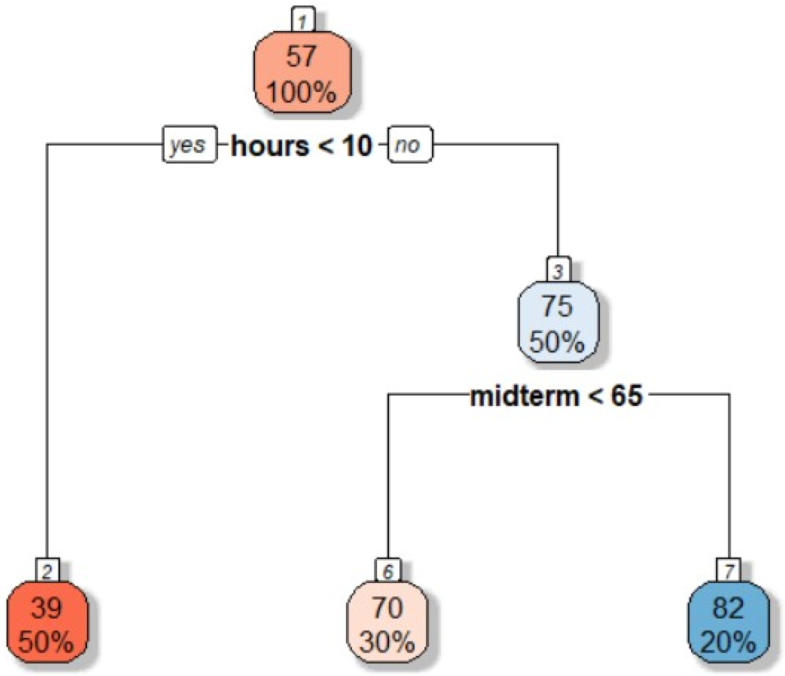
Mean score 75

Approach



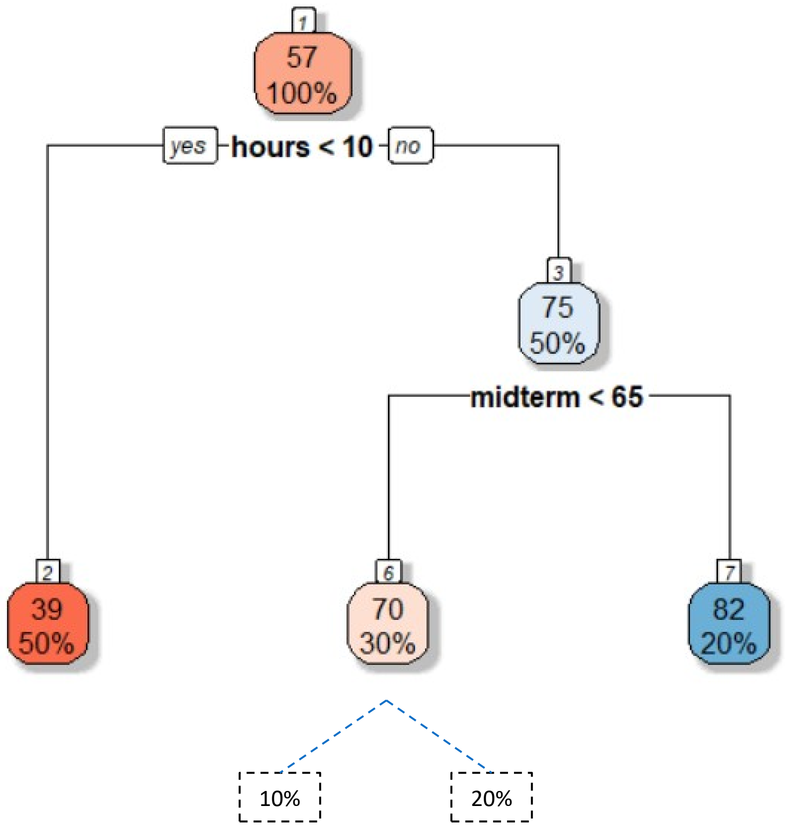
- Top-down, greedy approach that is known as recursive binary splitting.
- Top-down because it begins at the top of the tree and then successively splits the predictor space
- Each split is indicated via two new branches further down on the tree.
- It is greedy because at each step of the tree-building process, the best split is made at that particular step, rather than looking ahead and picking a split that will lead to a better tree in some future step.

Steps



1. Considers all predictors and all possible cut point values
2. Calculates RSS for each possibility
3. Selects the one with least RSS
4. Continues till stopping criteria is reached

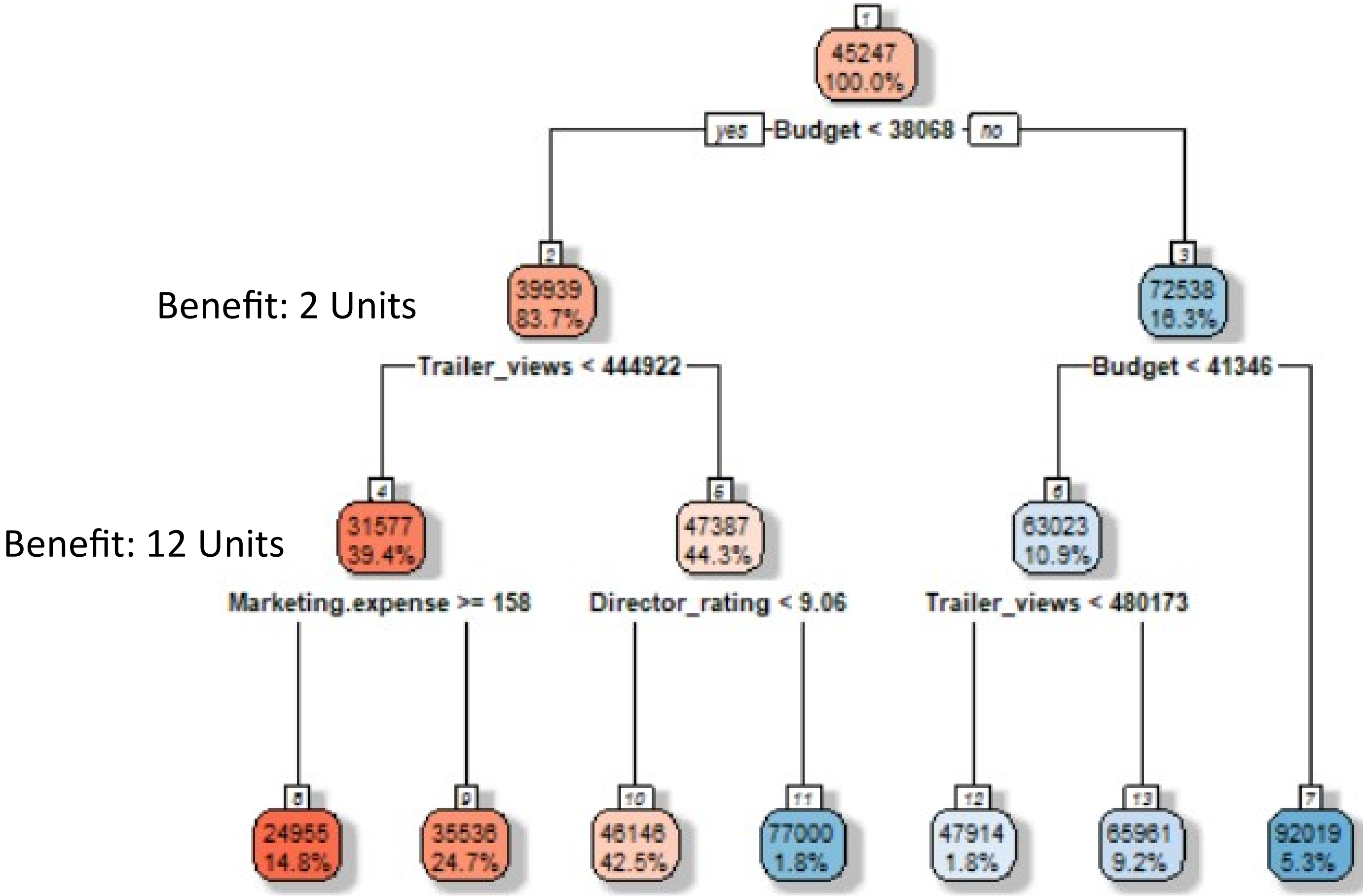
Stopping Criteria



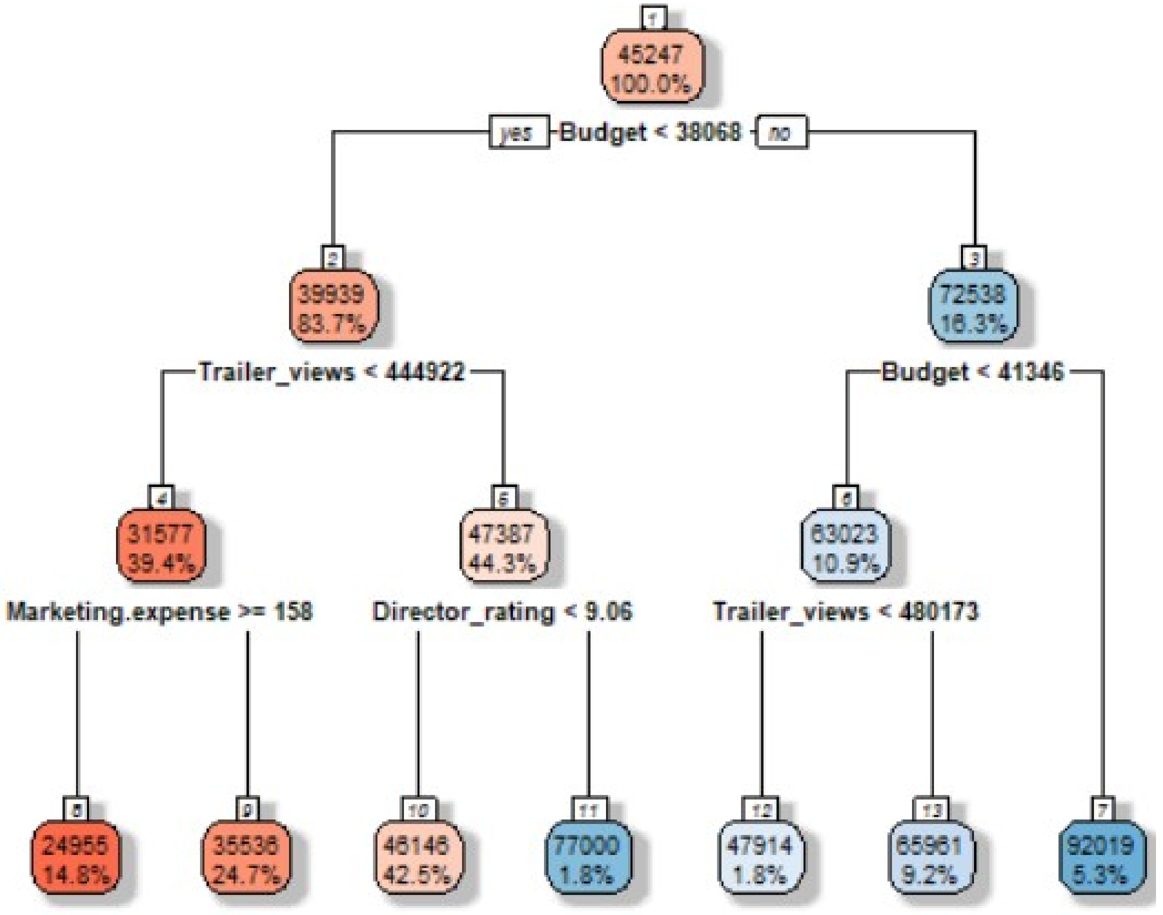
1. Minimum Observations at internal node
   Minimum numbers of observations required for further split
2. Minimum Observations at leaf node
   Minimum number of observation needed at each node after splitting
3. Maximum depth
   Maximum layers of tree possible

Pruning

Constraint
A split should have a benefit of 10 Units



Weakest Link Pruning

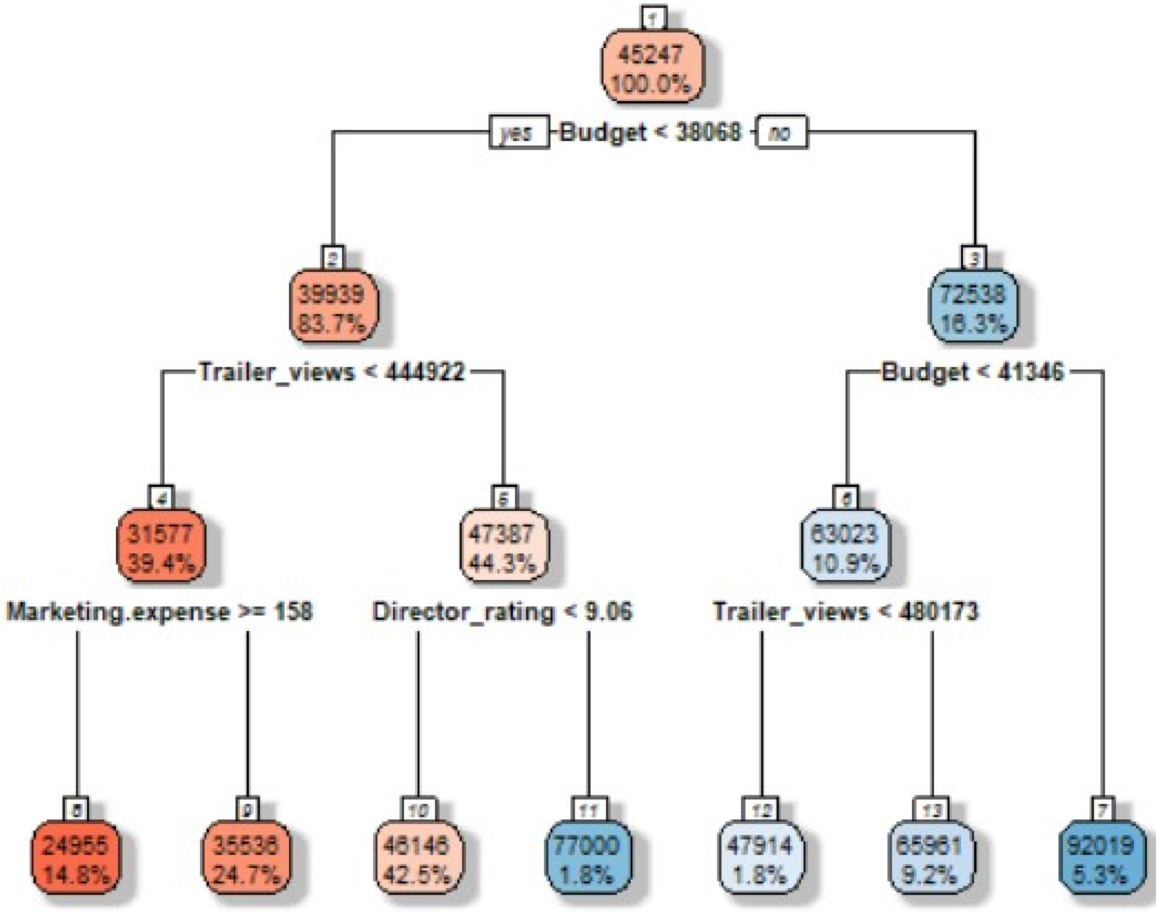


RSS

$$\sum_{m=1}^{|T|} \sum_{i:\, x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha|T|$$

a =Tuning Parameter
T =Number of Leaf nodes

Pruning



Steps

1. Grow a very large tree
2. Cut it back to get an optimal tree

Decision Trees

Types

1. Regression Tree
   For continuous quantitative target variable.
   Eg. Predicting rainfall, predicting revenue, predicting marks etc.

2. Classification Tree
   For discrete categorical target variables
   Eg. Predicting High or Low, Win or Loss, Healthy or Unhealthy etc

Classification Trees

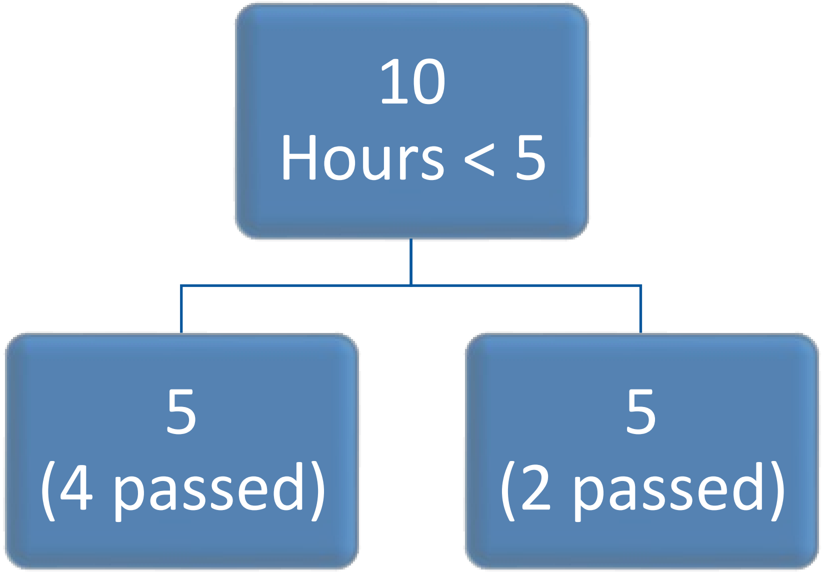
Prediction method

Regression

Mean of response variable became prediction for that class

Classification

We use mode (most frequent category in that region will be the prediction)



Classification Trees

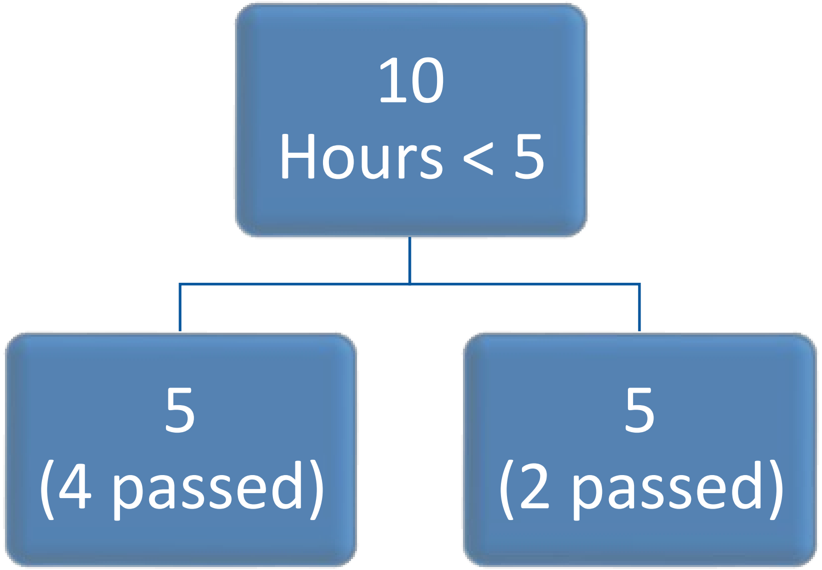
Methods

Both Regression and classification use recursive binary splitting

In Regression RSS is used to decide the split

In Classification we can use

1. Classification error rate
2. Gini Index
3. Cross Entropy



Classification Trees

Methods

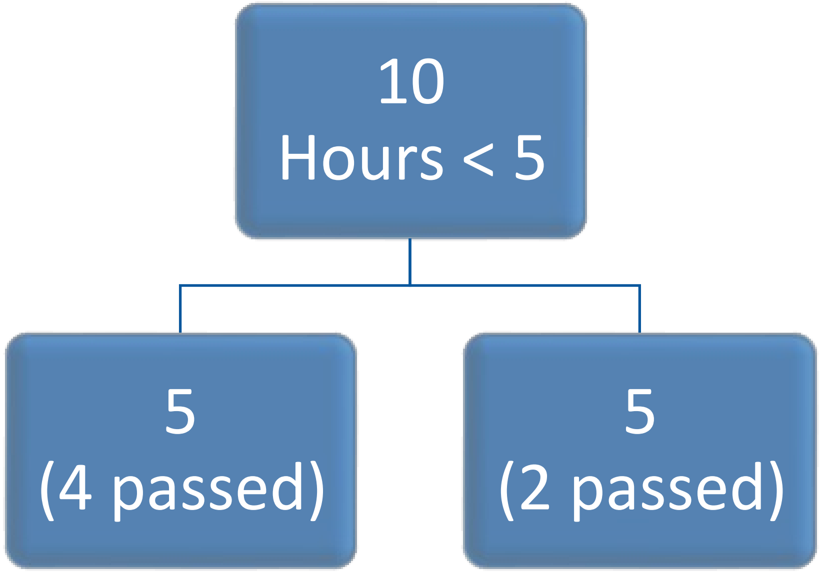
In Classification we can use

1. Classification error rate
2. Gini Index
3. Cross Entropy

Gini index and cross entropy signifies node purity

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

$$D = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}$$



Ensemble Methods

Types

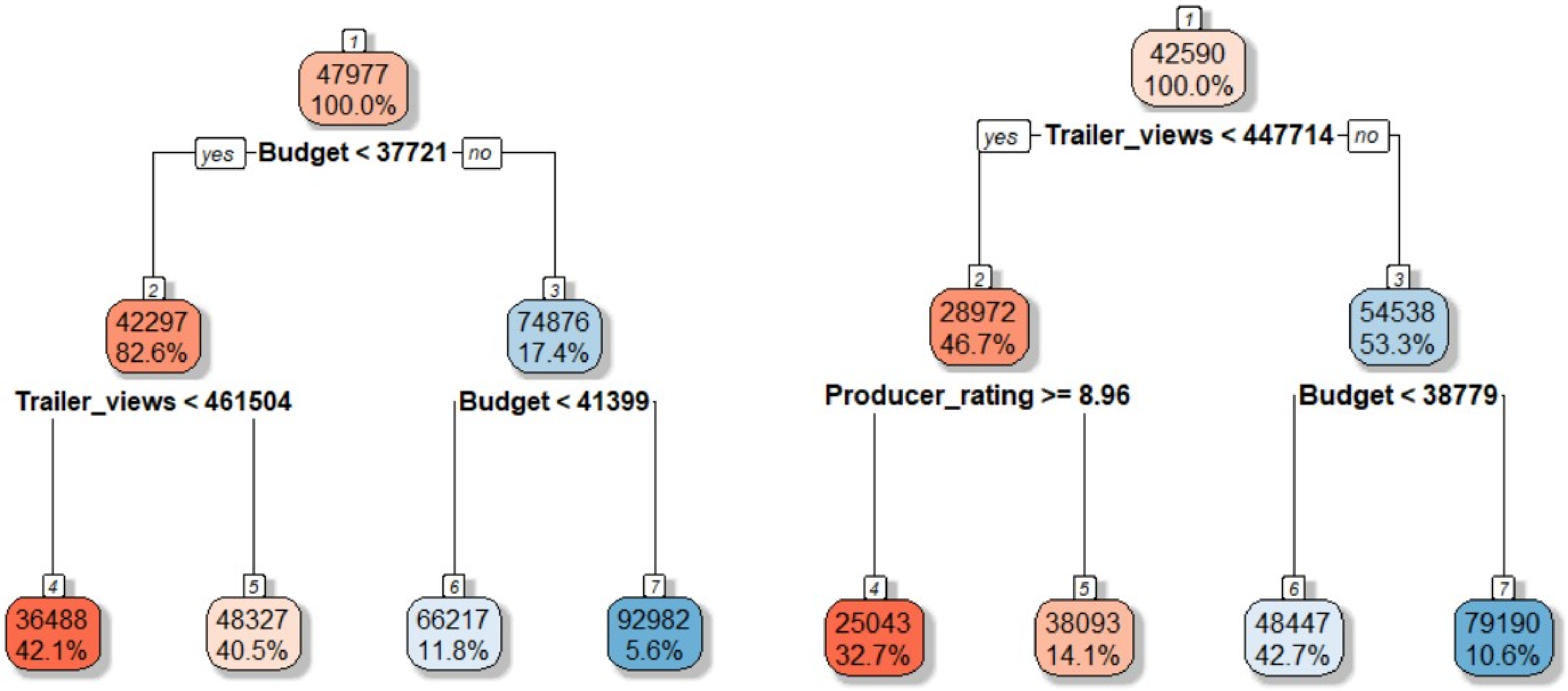
1. Bagging
2. Random Forest
3. Boosting

Problem with normal decision tree

- High Variance

Ensemble Methods

Example

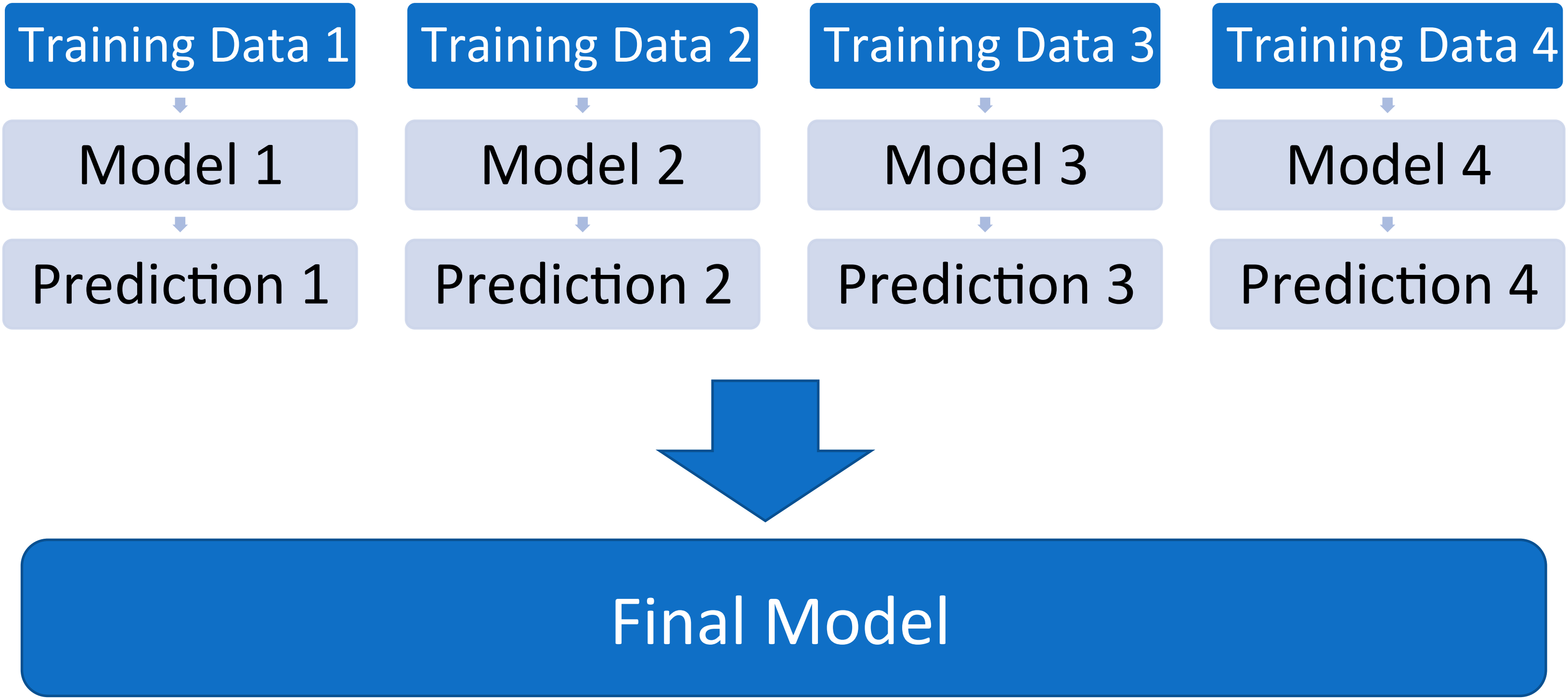


Bagging

Methods

Concept:-

If N observations have variance sigma sq. (s^2), then variance of mean of these observations is (s^2)/N



Bagging

Bootstrapping

	7	9	5	4	3
Sample 1 -	9	5	4	3	4
Sample 2 -	7	9	5	4	7
Sample 3 -	7	9	9	4	3

Bagging

Methods

1. While bagging pruning is not done, Full length trees are grown
2. Individual trees have high variance and low bias, averaging reduces the variance
3. In regression, we take the average of predicted values
4. In Classification, we take majority vote i.e. most predicted class will be taken as the final prediction

Random Forest

Shortcomings Of Bagging

Problem:-
Bagging creates correlated trees

Created models are very similar

Final Model

Random Forest

Shortcomings

Concept:-

We use subset of predictor variables so that we get different splits in each model

Training Data 1	Training Data 2	Training Data 3	Training Data 4
Random m pred	Random m pred	Random m pred	Random m pred
Model 1	Model 2	Model 3	Model 4
Prediction 1	Prediction 2	Prediction 3	Prediction 4

Different set of m predictors out of p

Final Model

Random Forest

Thumb Rule for value of M

1. For Regression
   P/3

1. For Classification
   sqrt(P)

2. Don’t forget to use your business knowledge
   If the variables are highly correlated try a smaller value of M

Boosting

Boosting

Process of turning a weak learner into a strong learner

1. Gradient Boost
2. Ada Boost
3. XG Boost

Boosting

Gradient Boosting

1st Tree
- Calculate residuals

2nd tree
- Tree after adjusting for residuals

Nth tree
- Final tree after adjusting residuals multiple times

Boosting

Ada Boosting

1st Tree

- Assign more weightage to misclassified observations

2nd tree

- Retrain the tree after accounting for weightages

Nth tree

- Final tree after accounting for weightages N-1 Times

Boosting

XG Boost

- Almost similar to Gradient Boost
- XG-boost used a more regularized model formalization to control over-fitting, which gives it better performance.
- For model, it might be more suitable to be called as regularized gradient boosting.

Regularization

The cost function we are trying to optimize (MSE in regression etc) also contains a penalty term for number of variables. In a way, we want to minimize the number of variables in final model along with the MSE or accuracy. This helps in avoiding overfitting

XG-Boost contains regularization terms in the cost function.

Decision Trees

Advantages

1. Trees are very easy to explain to people
2. decision trees more closely mirror human decision-making than other regression and classification approaches
3. XGTrees can be displayed graphically, and are easily interpreted even by a non-expert
4. Trees can easily handle qualitative predictors without the need to create dummy variables.

Decision Trees

Disadvantages

1. Trees generally do not have the same level of predictive accuracy as some of the other regression and classification approaches