Analytics

Assignment 1

Tinotenda Mutsemi (MTSTIN007)

2024-03-05

Question 1

Question 1a

We split the data into training and test sets using an 80/20 split.

We fit a homegrown tree by setting the minimum size of the terminal nodes to 2 and the minimum deviance to 0.

This tree is overfitted, and we see that loacation features and count result in the most misclassification error reduction.

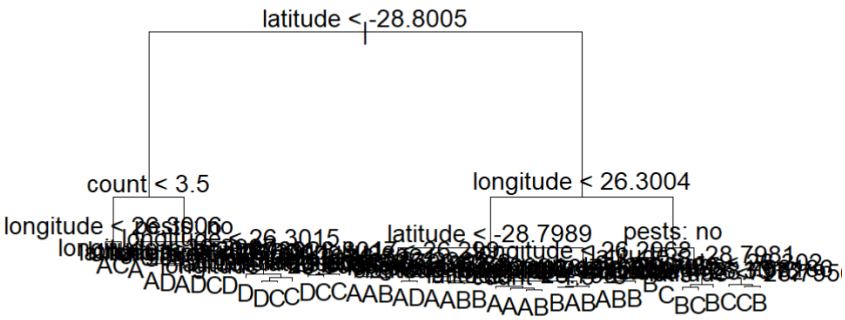

Figure 1: Overfitted Homogenous Mealies Tree

We then test the perfomance with the 20% we splitted from the data. Below is the confusion matrix of the test showing a misclassification rate of 8%.

	A Predicted	B Predicted	C Predicted	D Predicted
A Actual	40	3	0	1
B Actual	2	70	1	0
C Actual	0	1	31	5
D Actual	4	0	3	89

Question 1b

We use cross validation to prune the tree. We use 10 fold cross validation and the misclassification error as the cost function. We plot the cross validation error against the number of terminal nodes and the penality alpha.

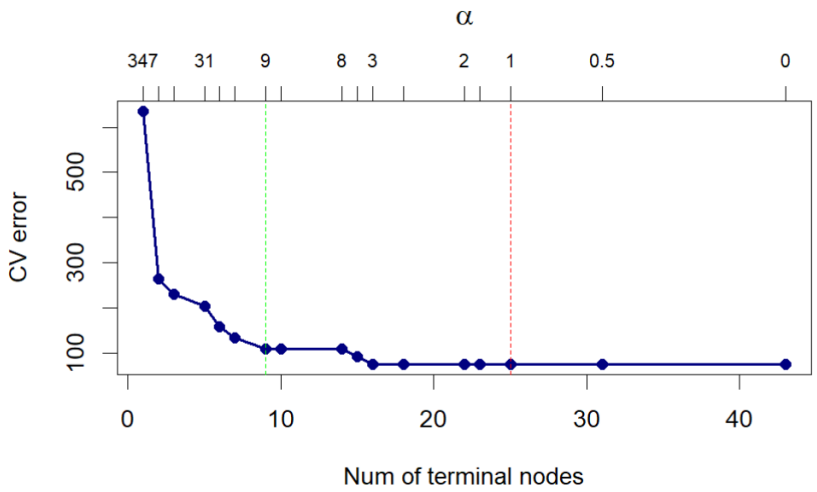

Figure 2: CV error for each tree size

We see the tree with the minimum cross validation error (red) line has over 20 nodes. We choose the tree with 9 nodes (green line) because it has good interpretability at a low error rate, compared to the lost of interpretability we get from the minimal error tree. The pruned tree has a misclassification rate of 13.2% and the confusion matrix is shown below.

These results are worse than from an overfitted tree. This tree has a higher cross validation error as shown by the fig above with CV error size for each tree. Therefore this is expected. We however gain interpretability at the cost of a sightly higher missclassification rate.

	A Predicted	B Predicted	C Predicted	D Predicted
A Actual	30	10	1	4
B Actual	1	71	1	0
C Actual	0	3	22	12
D Actual	0	2	0	94

Question 1c

We plot the location of the mealies on a scatter plot with the quality of the mealies represented by the color of the points. Tree methods partion the feature space othogonally and therefore we rotate the feature space to see if we can get a better partitioning of the feature space.

Now we apply 10 fold cross validation to the rotated data. We rotate the feature space over thetas. We then plot the misclassification error against the rotation angle. We see that the misclassification error is lowest at 0.97

The misclassification rate is 0.4%. This is expected because the tree method partions the feature space orthogonally and therefore rotating the feature space will find lower misclassification rates.

Below is the confusion matrix for the rotated feature space. This tree perform very well even on unseen data.

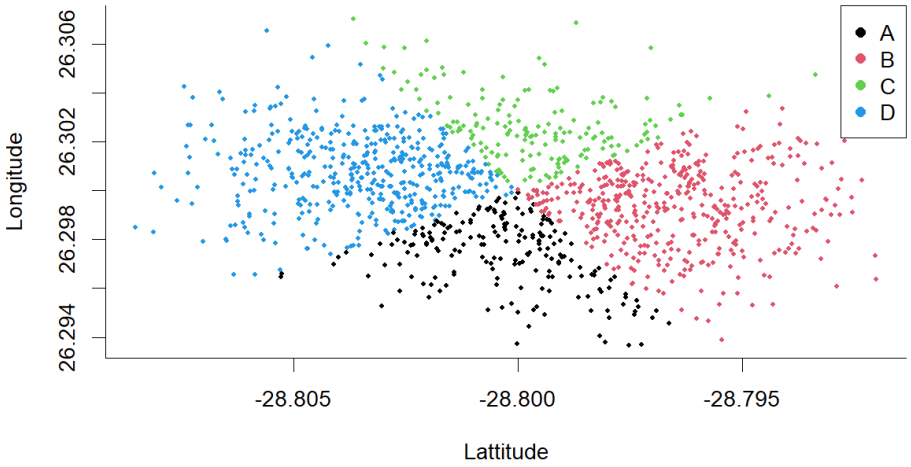

Figure 3: Mealies feature space

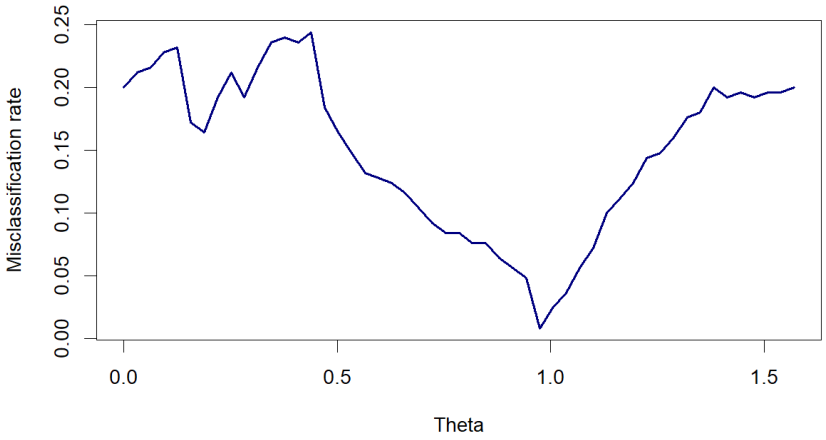

Figure 4: Misclassification rates for theta

	A Predicted	B Predicted	C Predicted	D Predicted
A Actual	44	0	0	0
B Actual	0	73	0	1
C Actual	0	0	37	0
D Actual	0	0	0	95

Question 2

We do some exploratory analysis to see the variables correlation.

Below is the scatter plot of First Inn Score and Bowl2 Strength. We can see that there is a positive correlation between the two variables.

We remove the match id from the regression, standardize the data and use dummy variables for the categorical variables.

We also remove defending team and chasing team from the regression. As they are not consistent across seasons.

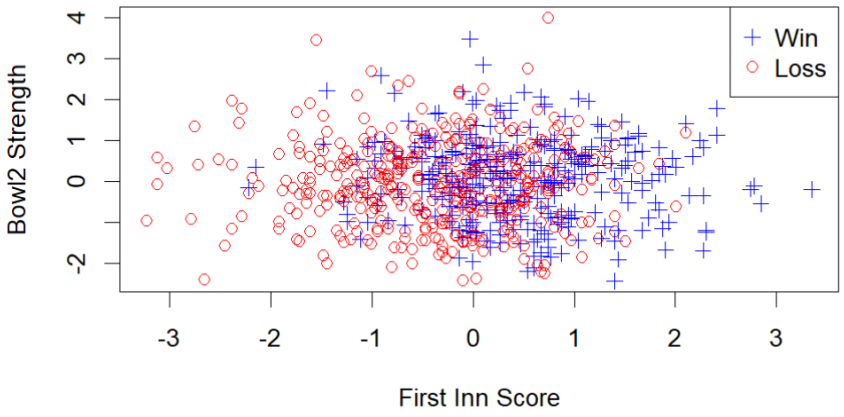

Figure 5: First Inn Score vs the team Bowling in the second round

Question 2a

We fit check the model after feature selection and standardization.

Now we fit an elastic net model to the data. We use cross validation to select the best alpha and lambda. We sequence alpha from 0 to 1 and use 10 fold cross validation to select the best lambda for each alpha.

The plot below shows the misclassificatioin rates for each alpha. The best alpha is 0.3

We fit the elastic net model with alpha = 0.3 and plot the coefficients. The best lambda is 0.28.

We see that at these hyperparameter we only have 1 feature in the model.

First Inn Score with a coefficient of 0.25. This means that a unit increase in the first innings score increases the log odds of winning by 0.25.

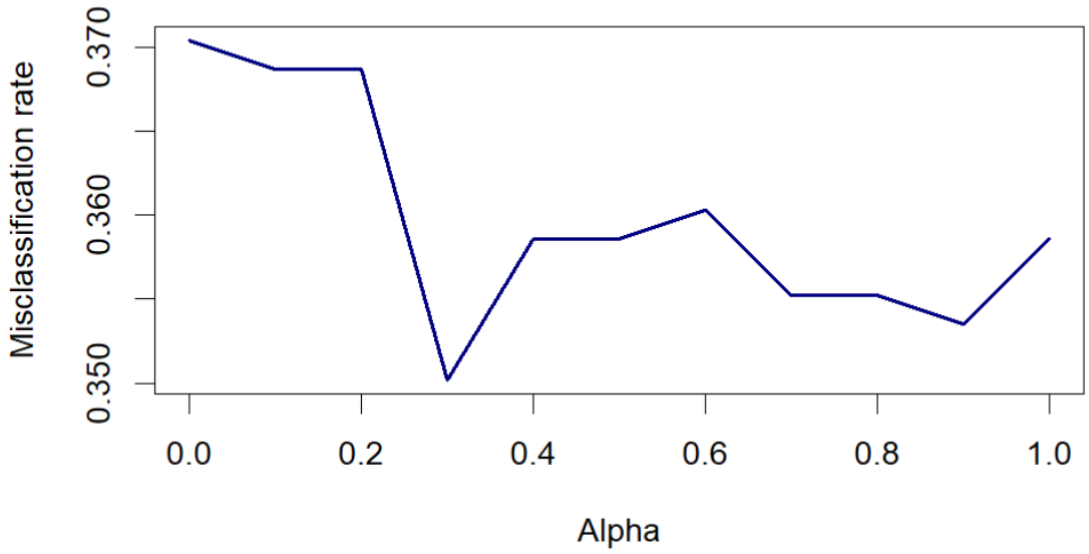

Figure 6: Misclassification Rates for each Alpha at the best Lambda

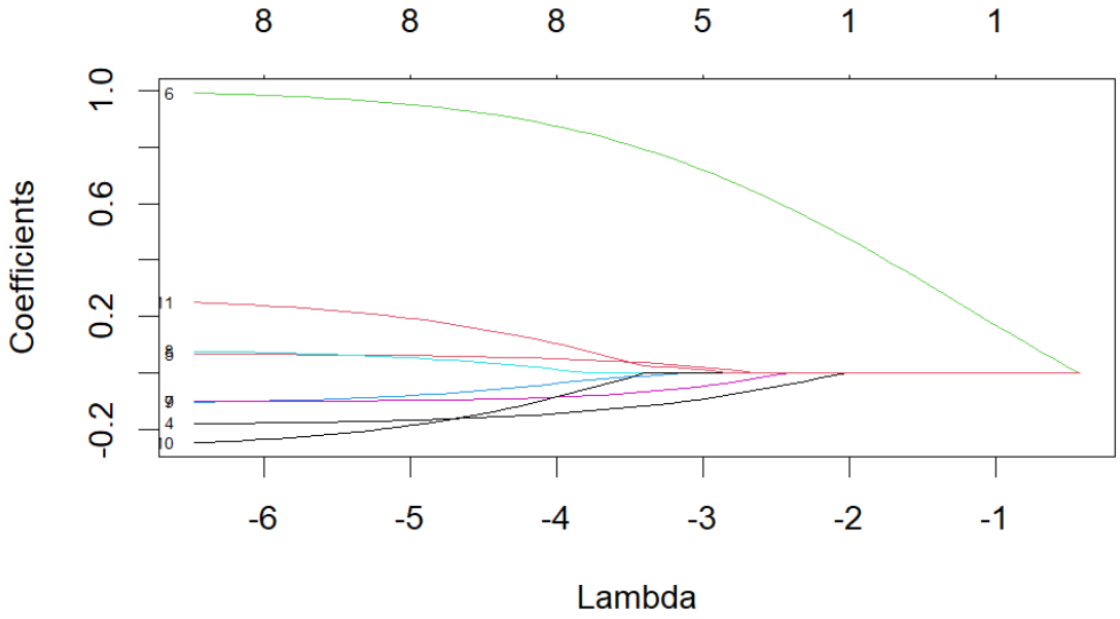

Figure 7: Coefficients at lambda 0.28 is First Inn Score only

Below is the confusion matrix from testing the model.

Table 4: Confusion Matrix

	Loss Predicted	Win Predicted
Loss Actual	80	10
Win Actual	40	19

The F score is 0.43 showing that the model is not good because the F1 score is less than 0.5.

Question 2b

Question 3

Question 3a

We do some exploratory analysis to see the variables correlation.

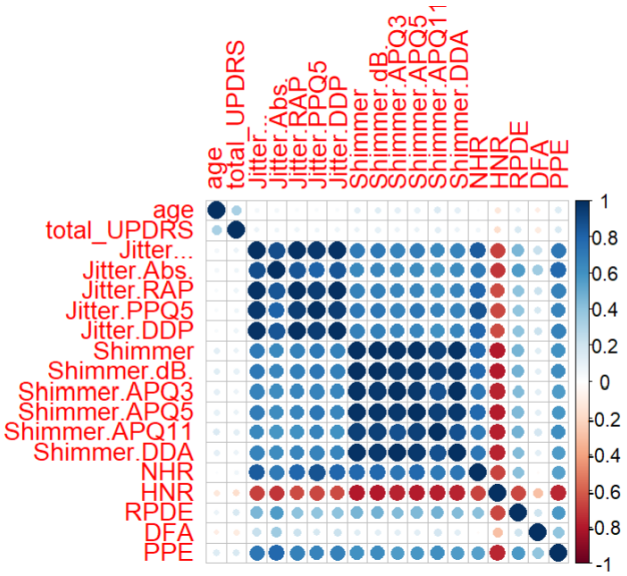

Figure 8: Variable correlations for Parkinsons data set

The table below shows the covariance, p value, and the correlation coefficient for the variables in the Parkinsons data set.

Question 3a i

We fit an elastic net with alpha = 1 for lasso regularization.

Our final model (min lambda.1se) has test MSE of 95, and variables age, sex, DFA with high significance.

The plot shows lambda and MSE for the elastic net model.

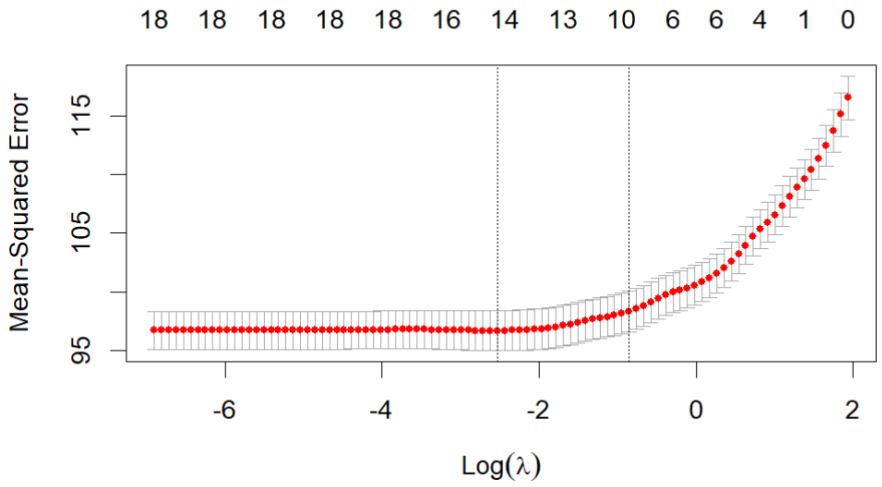

Figure 9: MSE Elastic net over lambda

Question 3a ii

We range k from 1 to 15 and set up the control for model training using 10-fold cross-validation, repeated 10 times to provide a robust estimate of model performance.

The choose k = 4 because it has the lowest RMSE.

Question 3a iii

Now we train a random forest model using ranger and caret. Number of variables randomly sampled as candidates at each split in the decision tree is varied from 2 to 20, and the minimum node size is 1, 5, 10, 15, 20. We use 10-fold cross-validation, repeated 10 times to provide a robust estimate of model performance.

The est MSE of 8.75.

Question 3a iv

We now fit the model using the xgboost algorithm. We search for the best hyperparameters using a grid search set to 100 boosting rounds and 10-fold cross validation repeated 10 times. We then fit the model using the best hyperparameters and predict the test data.

Below is the variable importance plot for the xgboost model. We notice that after the 8th variable the importance is 0. The most important features are age, DFA and sex.

Question 3b

The table below shows the MSE of the 4 different models. XGBoost produces the least MSE, and we shall choose it as the best model and use it to test the data performance with Q3testing.csv.

Model	MSE
Elastic net	95.59

Model	MSE
KNN	14.90
Random forest	19.70
XGBoost	9.00

Below is the plot showing the RMSE of the XGBoost model. The selected model has a depth of 8. There is a low RMSE however its weakness is lack of interpretability. For lesser tree depths, where we have good interpretability, the RMSE has not yet converged.

Question 3c

The predictions for Q3testing are in MTSTIN007.csv.

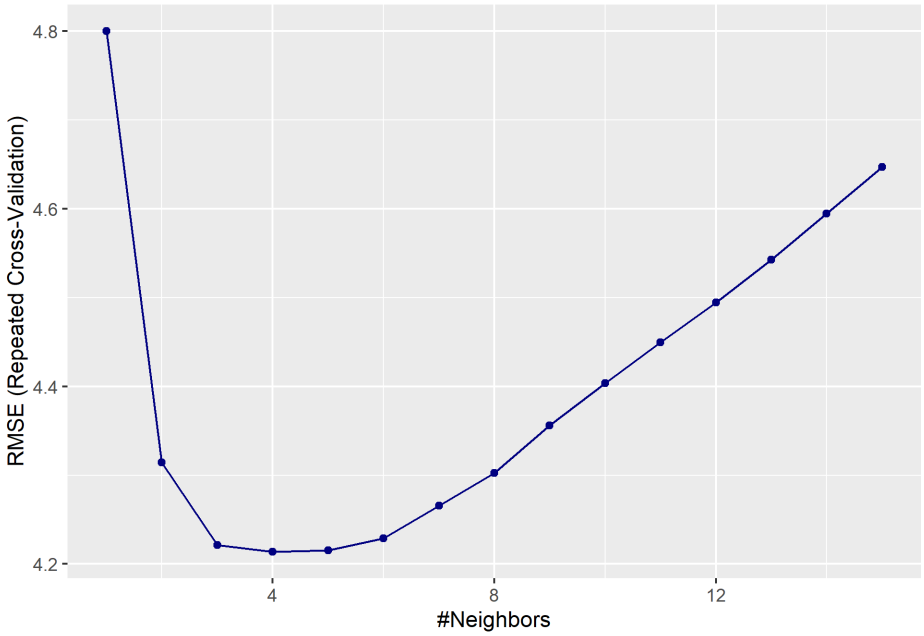

Figure 10: RMSE of KNN from k = 1 to 15

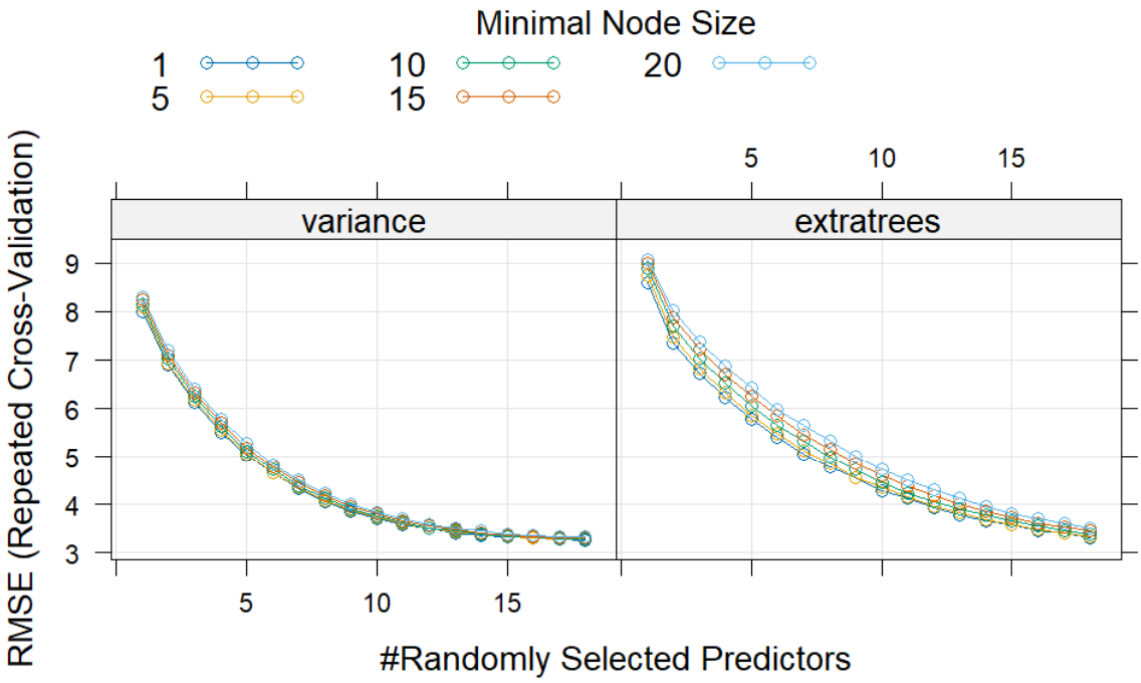

Figure 11: Random forests grid RMSE

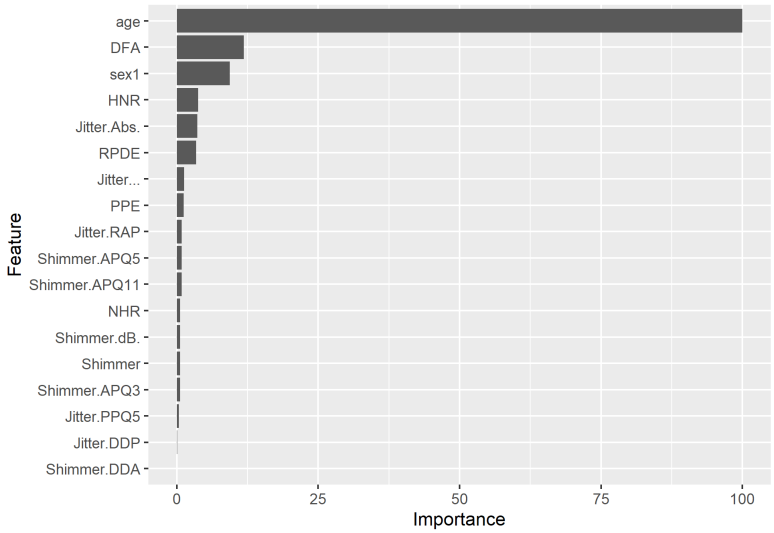

Figure 12: XGBoost Variable Importance

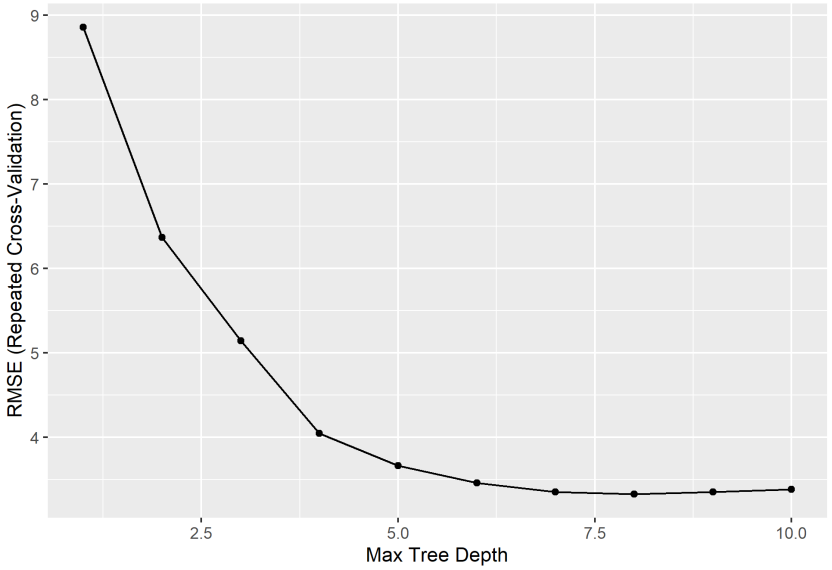

Figure 13: RMSE vs Maximum tree depth.