Part 1 (clustering)

Means of different variables are similar with the range of 1.49 to 2.60, this shows that employees are satisfied with both their mental health and job.

According to Chart 1, we can see that the first loading vectors places emphasis on employees' mental health while the second loading vectors focus on employees' working environment. Variables of workers' mental health are located close to each other hinting correlation, and workers' job variables also seem to be correlated.
We also observe that majority of the respondents are gathered at the bottom half of the chart with only a few outliers such as observation 3945, 5408 and 7635.

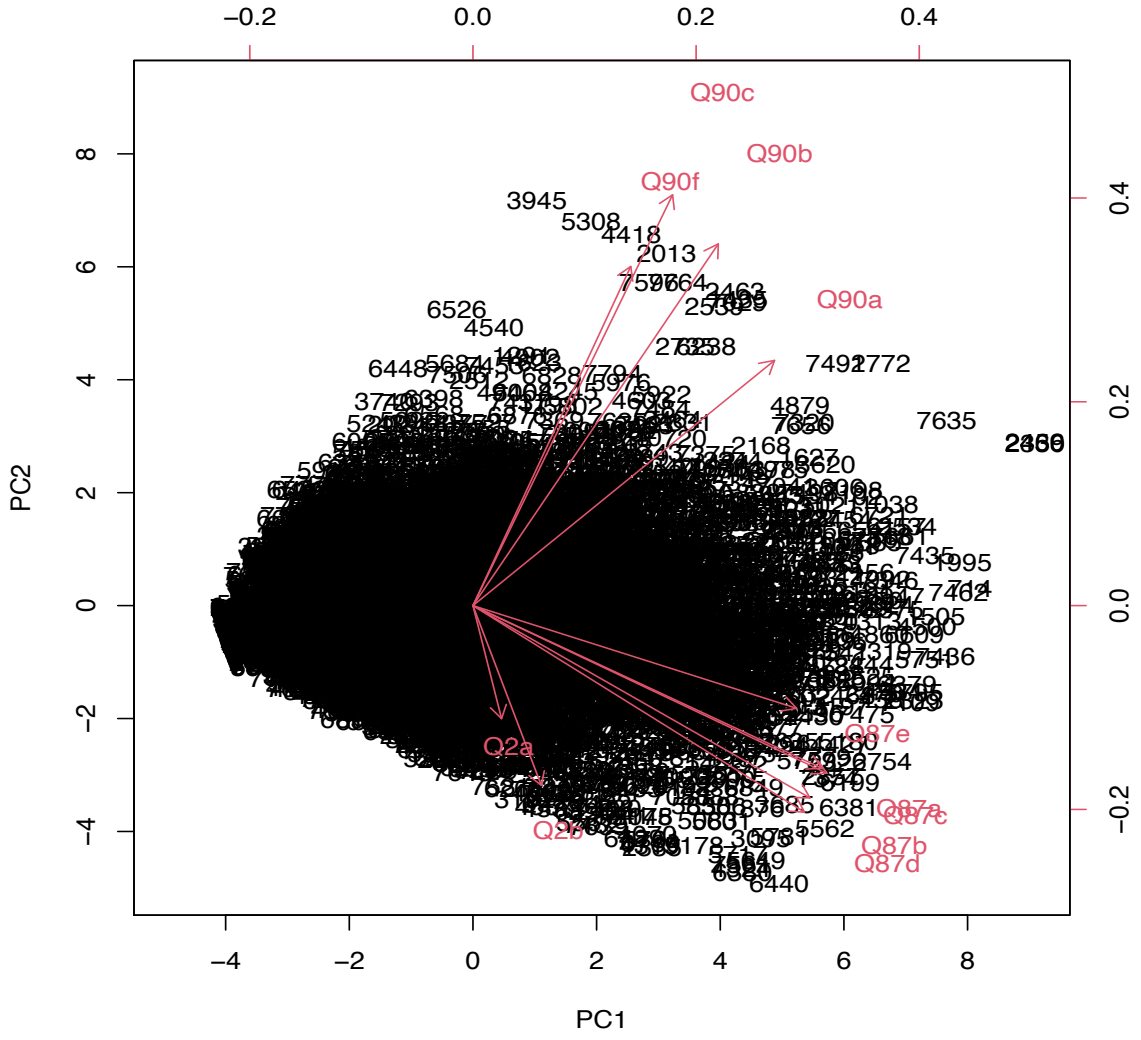

Chart 1

The first two Principal Components (PC) captured 52% of variance. There is a strong evidence that PC1 represents employees' mental health and PC 2 represents employees' working environment. Additionally, Q90b, Q90c and Q90f are relatively more important in PC 2 than in PC 1 which means these variables are important differentors.

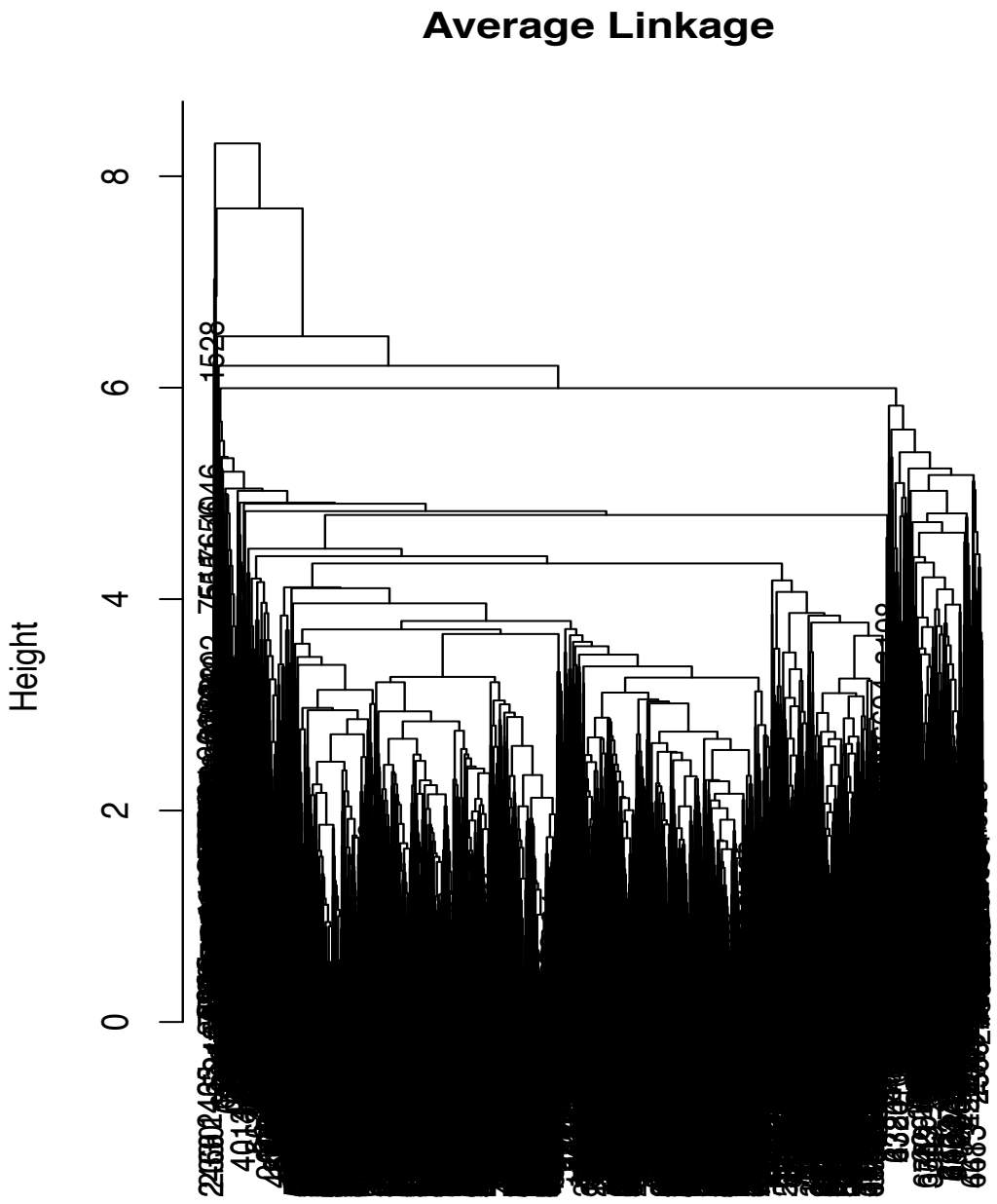

Chart 2

As seen in Chart 2, majority of the observations are lumped into 1 cluster, which means average linkage fails to provide sufficient sample for one cluster.

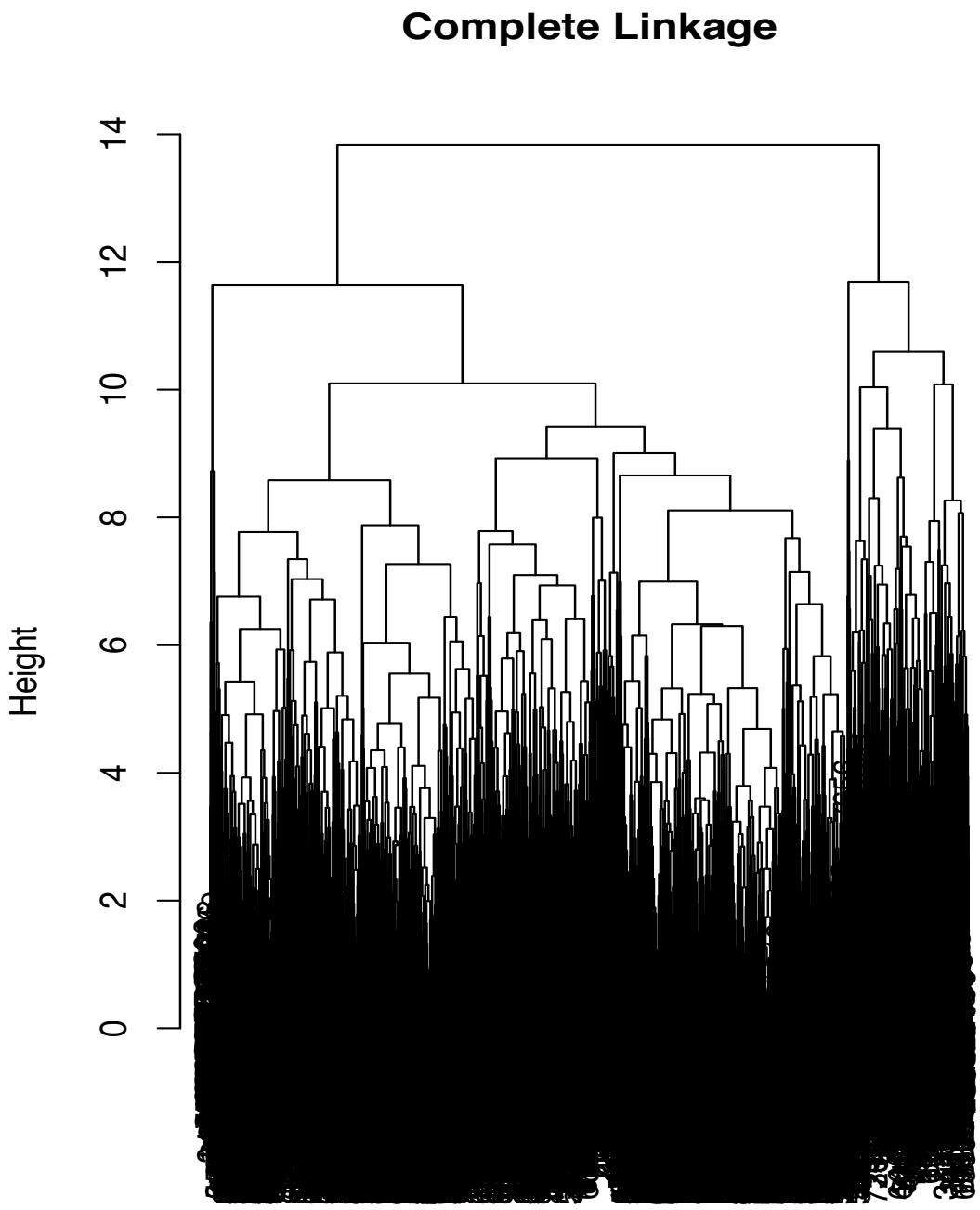

Chart 3

From Chart 3, there are obviously 2 clusters with sufficient samples for each cluster. By cutting the clusters into 2, 4 and 6, cluster 2 remains with the highest number of

observations, this shows that no matter the number of clusters, cluster 2 still remains with the majority of observation.

Part 2 (Regression)

There are 1044 observations with 33 variables in this dataset, with age, absences, G1, G2 and G3 as integers and the rest as factor variables. We remove variables G1 and G2 to get a better predictive performance of G3. From the summary of model, which is a linear regression with G3 as dependent variable, there are many dummy variables created for the categorical independent variables. However, only some variables such are significant, while majority of the variables are insignificant. Adjusted R-squared is 0.2465 which means the predictor variables are not good in predicting G3. As such, we use backward elimination method to select a model with the most significant variables to predict G3. According to backward elimination, the model selected (bk_model) is **G3 = address + famsize + Mjob + Fjob + internet + studytime + failures + schoolsup + paid + higher + romantic + freetime + goout + health** , with lowest AIC = 2557.11. Adjusted R-squared is 0.2477.

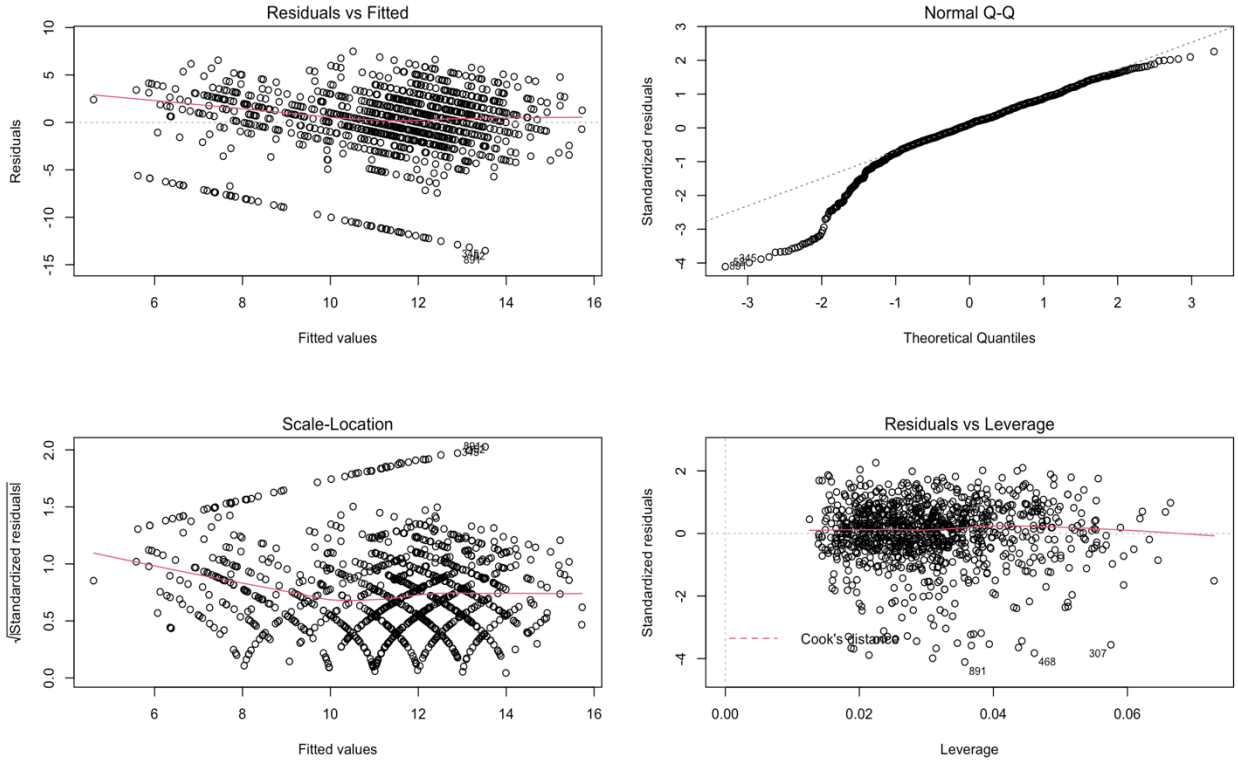

Chart 4

Chart 4 shows 4 diagnostic plots of bk_model. Residuals vs Fitted plot display equally spread residuals around the horizontal line with no distinct pattern. This indicates the model is linearly related. Normal Q-Q plot shows that majority of residuals are lined on the straight dashed line, which means majority of the residuals are normally distributed. From Scale-Location plot, we see horizontal line with equally spread residuals, which checks homoscedasticity. Residuals vs Leverage plot does not display Cook's distance, this means there are no influential outliers.

Multicollinearity do not exist in bk_model as GVIF of the predictor variables are all within 1.
Root mean square error (RMSE) of original data is 3.3.
We set seed at 123 before splitting train and test sets into 70% and 30% respectively.
Trainset (Model_train) have adjusted R-squared of 0.2461 and RMSE of 3.29.
We predict linear model using the test dataset which gives 3.38 RMSE of testset.

We then proceed to use train and test set using Lasso and Ridge regression. Lasso gives 3.38 RMSE of training set and 3.80 RMSE of test set. Ridge provide 3.71 RMSE of trainset and 3.73 RMSE of testset.

As such, dataset did not improve in prediction by using trainset and testset and we conclude that simple linear regression model is the best model with lowest RMSE among the 3 linear regression models.

Part 3 (Classification)

The bank's client data represents 4521 observations with 17 variables. Our predictive variable y is a categorical binary variable (yes/no). We remove 'duration' variable for a realistic predictive assessment From summary, there are 4000 'no' and 521 'yes', as such majority of the bank client did not subscribe to the product.

Maximal Tree in bank.csv

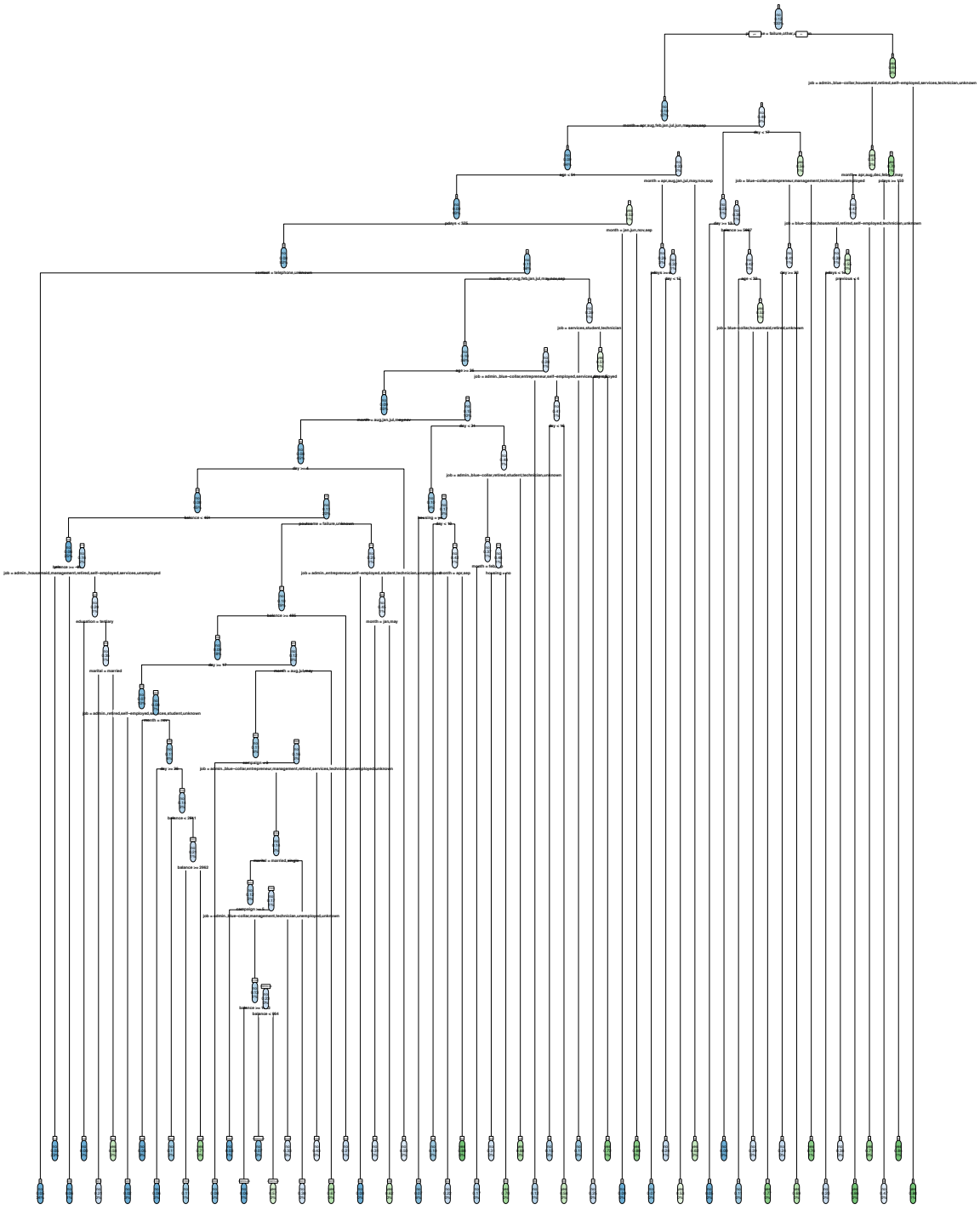

Chart 5

Chart 5 displays the maximal tree using the dataset provided, excluding duration variable.For node number 32, which is a terminal node, decision rule is pday<374.5 and contact=telephone or unknown. Total cases 1556 in the node of which 75 cases will be misclassified if predict by majority (no). Node number 32 has an intense dark blue colour which means the node is pure.

From printcp, the tree can be pruned 14 times and in total 60 splits.

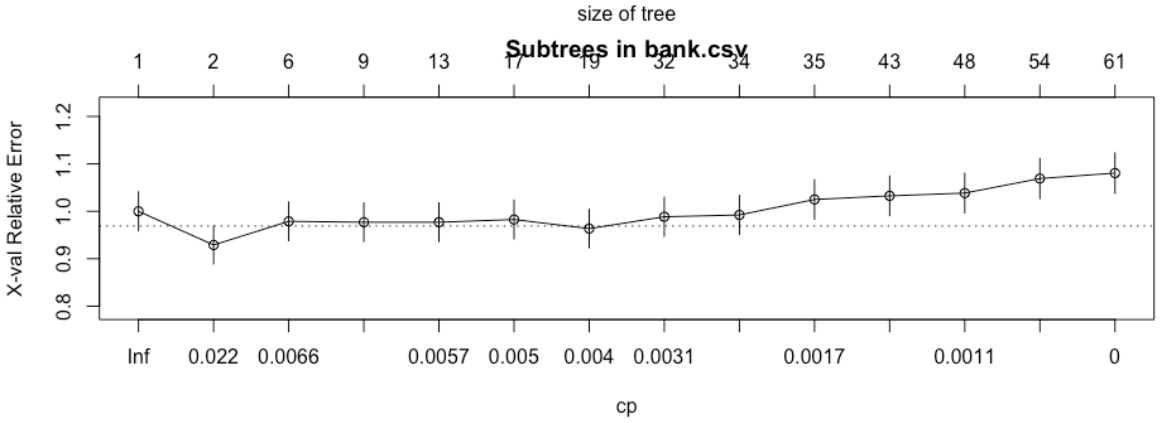

Chart 6

Chart 6 shows the pruning sequence and 10-fold cross validation error.

Minimum cross validation error tree is 2. Since there is only 1 tree below the dotted line which is the cross validation error cap, the minimum cross validation of tree 2 is our optimal tree with cp 0.022.

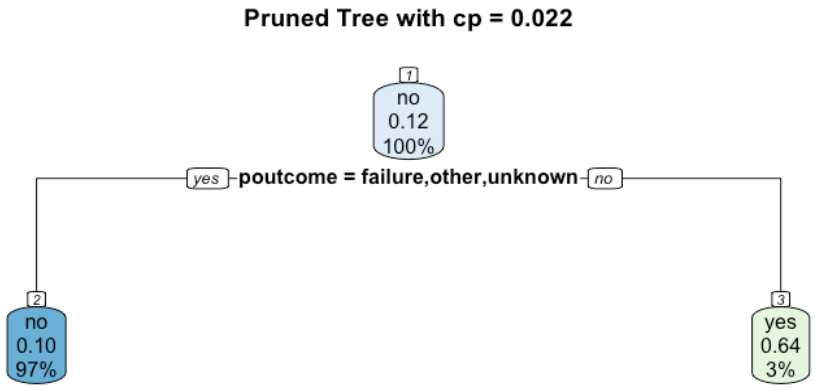

Chart 7

Mod2 is our specific optimal subtree, thus we prune the tree when CP is 0.022. Chart 7 shows the plot of pruned tree. Majority (97%) of the clients with outcome of the

previous marketing campaign = failure, other, unknown, did not subscribe to the marketing product. Only 3% subscribed to the product with poutcome = success

Using m2 prediction, we test cart model specifically with poutcome = success. The result of predicted y variable is ‘yes’ as expected. As such, we conclude that clients’ data with poutcome = failure, other, unknown are more likely to not subscribe to the marketing bank product.