National College of Ireland
Project Submission Sheet – 2019

Student Name:	Antonio Caruso		
Student ID:	19203608		
Programme:	Higher Diploma in Science in Data Analytics	**Year:**	2020
Module:	Data & Web Mining		
Lecturer:	John Kelly		
Submission Due Date:	24/07/2020		
Project Title:	CA PART 3: Decision Tree Model and Random forest applied on a marketing dataset		
Word Count:	1072		

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the references section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature: ……
……

Date: 24/07/2020

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. Projects should be submitted to your Programme Coordinator.

3. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. **Late submissions will incur penalties.**
5. All projects must be submitted and passed in order to successfully complete the year. **Any project/assignment not submitted will be marked as a fail.**

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Table of Contents

Table of Figures

1 Introduction

This document reports the application of a Decision Tree Classification model to the “Bank Marketing” dataset (Moro, et al., 2014).

The data refer to a direct phone marketing campaigns executed by a Portuguese Banking Institute. The reason for choosing this particular dataset lies in the author’s personal interest in marketing and advertising.

The scope of the application of the decision tree model was to predict whether a client will subscribe to a term deposit. The Random forest algorithm was applied and to improve the efficiency of the model and subsequently interpreted. Due to the nature of the dependent variable (categorical as expressed in yes or no preference), the Rpart algorithm used built a classification tree rather than a regression tree which applies to numerical variables or continuous variables (Lantz, 2015).

2 Dataset Description

Variable	Data Type	Description
age	integer	Age of the potential clients
job	character	Type of the jobs as following: "admin.","unknown","unemployed","management","housemaid","entrepreneur","student", "blue-collar"
marital	Character	"married","divorced","single"; note: "divorced" means divorced or widowed
education	character	As"unknown","secondary","primary","tertiary"
default	character	“Yes” or “No” referred to potential credit in default
balance	integer	Average year balance in Euro
housing	character	“Yes” or “No” referring to housing loan
loan	character	“Yes” or “No” referring to personal loan
contact	character	Communication type as “unknown”, “telephone”, “cellular”
day	integer	Last contact day of the month
month	character	Last contact month of the year
duration	integer	Last contact duration in seconds
campaign	integer	Number of contacts performed during the campaign for a specific client
pdays	integer	Number of days passed after a client was last contacted from a previous campaign (-1 means the client had never been contacted before)
previous	integer	Number of contacts performed before this campaign
poutcome	character	Outcome of the previous marketing campaign as “unknown”, “other, “failure”, “success”)
Y (Dependent Variable)	character	Has the client subscribed to a term deposit? Yes or No

Figure 1"The Bank Marketing Dataset", source: Moro et al,2011

The full dataset includes 45,211 observations and 17 variables described in the table above. In the presence of integer and characters data types, the dataset could be considered multivariate. Neither missing value nor blank were detected, thus the data pre-processing was accelerated.

3 Data Pre-Processing

For the data pre-processing, all the variables were considered for the analysis. In fact, the decision tree model would select the best variables and plot the trees accordingly (Bennette, 2011). After checking the structure of the data, the Y dependent variable corresponding to clients which subscribed or not to the term deposit, was converted into a factor to start building the decision tree. A quick summary view displayed 39,922 no and 5,289 yes to the subscription term.

4 The Classification Three

The first step to build the classification three was to split the dataset into train and test datasets, with the first calculating 70% of the values and the second the remaining 30%. The code applied resulted in 31,647 observation for the train dataset and 13,654 for the test dataset. The Rpart algorithm was then applied (Lantz, 2015) as it iteratively split the data into partitions forming a tree which nodes referred to the values of attributes trained by the algorithm, edges corresponding to the outcome of the test, and finally leaf nodes predicting the outcome. As minimum number of observations 20 was chosen as default value (Milborrow, 2019), The output showed a result of 11 leaf nodes with splitting variables of “duration” and “poutcome” respectively with importance of 61 and 38 as shown in the output and graphs below.

```
Classification tree:
rpart(formula = y ~ ., data = train, method = "class", control = rpart.control(minsplit = 20,
    minbucket = 7, maxdepth = 10, usesurrogate = 2, xval = 10))

Variables actually used in tree construction:
[1] duration poutcome

Root node error: 3753/31647 = 0.11859

n= 31647

        CP nsplit rel error  xerror     xstd
1 0.042988      0   1.00000 1.00000 0.015325
2 0.023448      3   0.87104 0.90354 0.014661
3 0.014655      4   0.84759 0.85638 0.014318
4 0.010000      5   0.83293 0.84253 0.014215
```

Figure 2 "Classification Tree output", own results

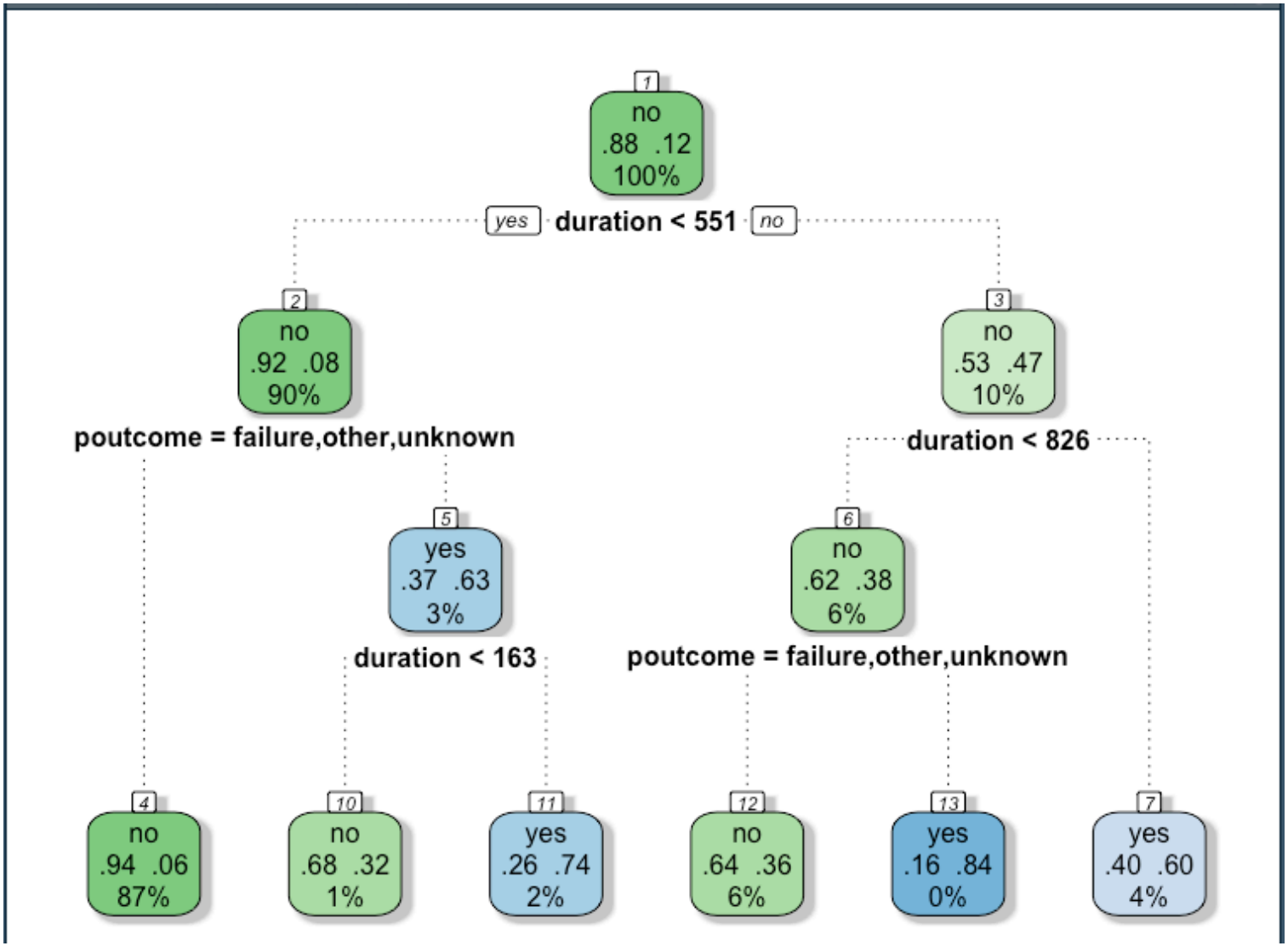


Figure 3 "Classification Tree", own result

The initial tree displayed 6 terminal leaf nodes. It is clearly visible how the first node begins the test with a duration of marketing call <551 splitting the prediction in yes and no, with the predicted value and the percentage of observation in the nodes. On the left side, 90% of observation classified as no, were tested on the succeeded outcome of a previous campaign, of which 3% is classified as yes and the remaining 87% as no following an outcome of failure, other or unknown from a past campaign. The final test on a duration of <163 of the marketing call, brought the result of 2% finally subscribing to the loan.

On the right side of the tree, the classified No for a campaign of <551, resulted in 10% tested No, of which 6% would respond yes to subscription on a call duration of <826 seconds. Of The remaining 6%, only dozens of people would finally subscribe to the term following success from a past campaign outcome, whereas almost the entire 6% would not subscribe to the loan following a failure or unknown result from a past campaign.

The lowest level of Complexity parameter of 0.1 (Basile, et al., 2017) was then selected for pruning the tree to a more compact tree with 4 nsplits as below;

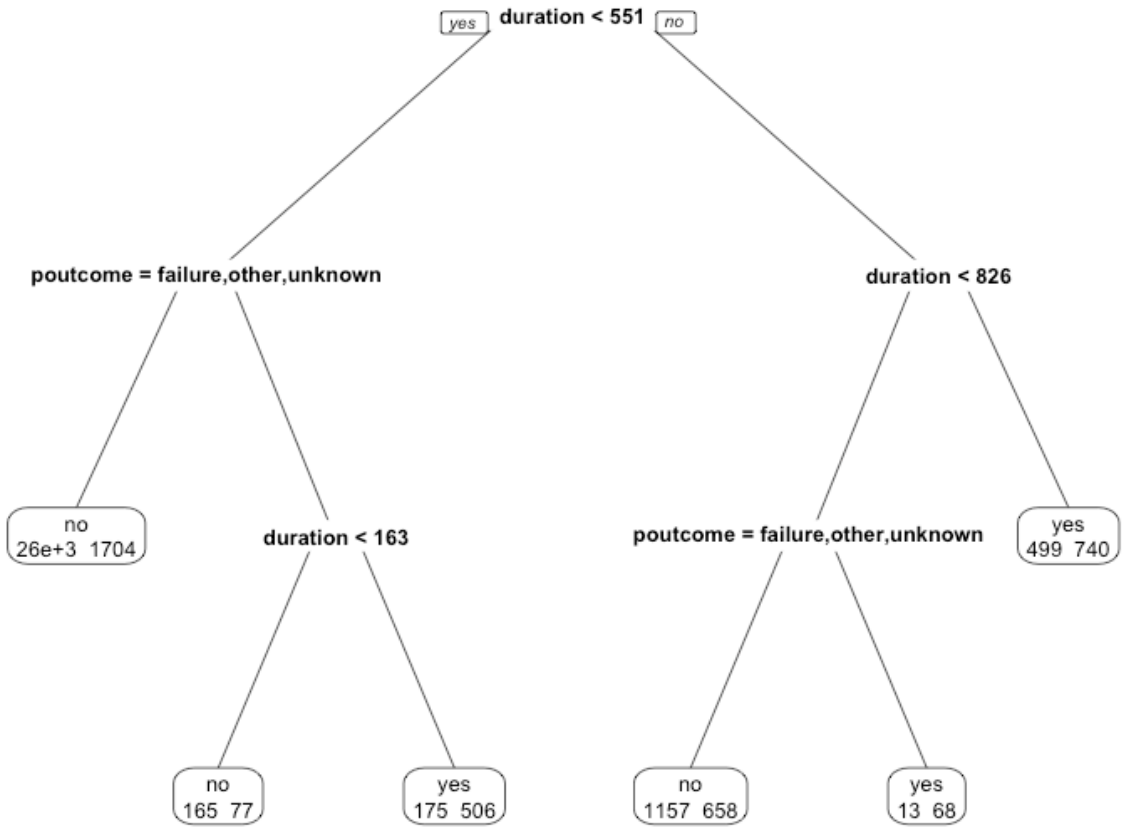
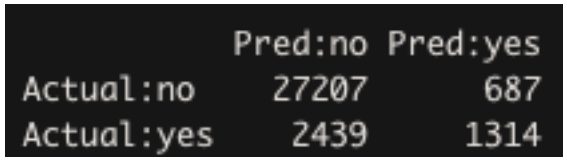


Figure 4 "Pruned Classification Tree", own result

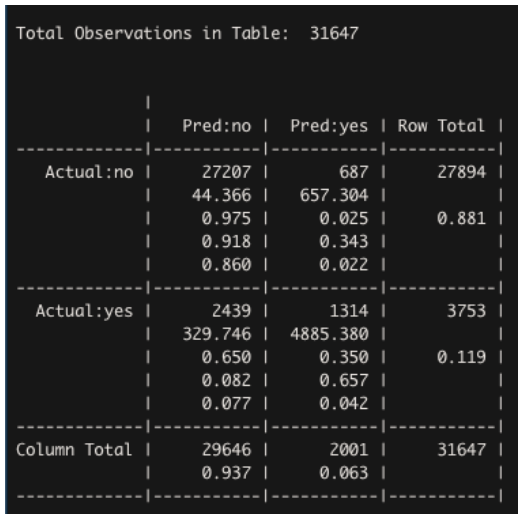
The result showed improved prediction, and a reversed confusion matrix was built to describe the performance of the model and generated the following result (Riou, et al., 2015).



	Pred:no	Pred:yes
Actual:no	27207	687
Actual:yes	2439	1314

Figure 5 "Confusion Matrix output", own results

The output shows that of the 27894 that actually were classified as no, the model predicted 27,207 as responding no to the subscription loan (true negative) and 687 as responding yes (false positive). Equally, of the 3753 that responded actually yes, the model classified incorrectly 2,439 no (false negative) and 1,314 (true positive) correctly predicted. To understand the validity of the model, a cross-table was analysed (Vuk, 2006).



Total Observations in Table: 31647

	Pred:no	Pred:yes	Row Total
Actual:no	27207	687	27894
	44.366	657.304	
	0.975	0.025	0.881
	0.918	0.343	
	0.860	0.022	
Actual:yes	2439	1314	3753
	329.746	4885.380	
	0.650	0.350	0.119
	0.082	0.657	
	0.077	0.042	
Column Total	29646	2001	31647
	0.937	0.063	

Figure 6 "Cross Table" own results

The cross table shows the accuracy of the model. The true negative rate resulted in 9% and the true positive is 6.3%, therefore very low values for the model.

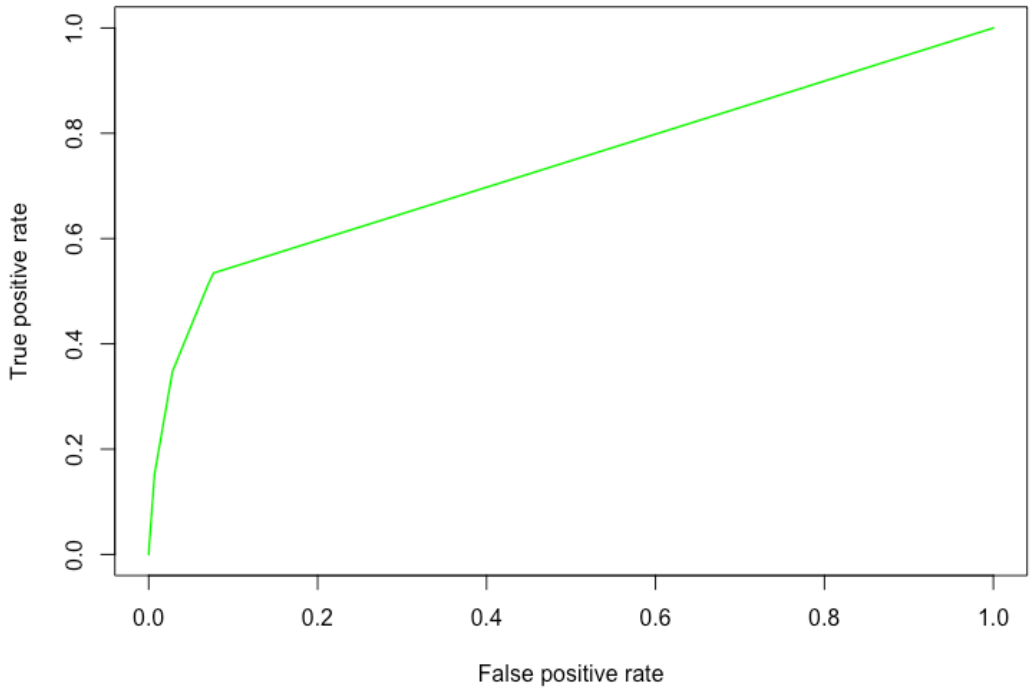


Figure 7 "ROC Curve", own results

To analyse the validity of the classification model performance, the Rock curve was plotted (Vuk,2006) . It displays the specificity, true negative rate, against the sensitivity, true positive rate, at various threshold settings. The main measure returned was the Area Under the Curve (AUC) that is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. In the model applied, the curve fell under 0.60, determining that the test was not good enough as 1 would equal to 100% correctness. The KS statistics as measure of degree between the positive and negative distribution was calculated at 0.49 (Utgo & Clouse, 1996), and since it should be in the top 10 deciles with a higher value, it confirms that the model should then be improved with the application of random forest.

5 Random Forest Model

The Random forest was applied to improve the performance of the classification tree, obtained using randomForest library (Cutler, et al., 2007), which already estimates the cross-validation.

```
Call:
 randomForest(formula = y ~ ., data = train, importance = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 4

        OOB estimate of  error rate: 9.42%
Confusion matrix:
       no  yes class.error
no  27018  876   0.0314046
yes  2104 1649   0.5606182
```

Figure 8 "Random Forest Output", own results

The output reported a number of 500 trees, 4 variables at each split and an OOB estimate error of 9.42%, therefore the combination of trees produced a . The reverse confusion matrix predicted 38,716 as responding no to the campaign (true negative) And 1206 as responding positive to the campaign (false positive). Then, for 3025 the model classified incorrectly (false negative) and 2264 are the (true positive) correctly predicted.

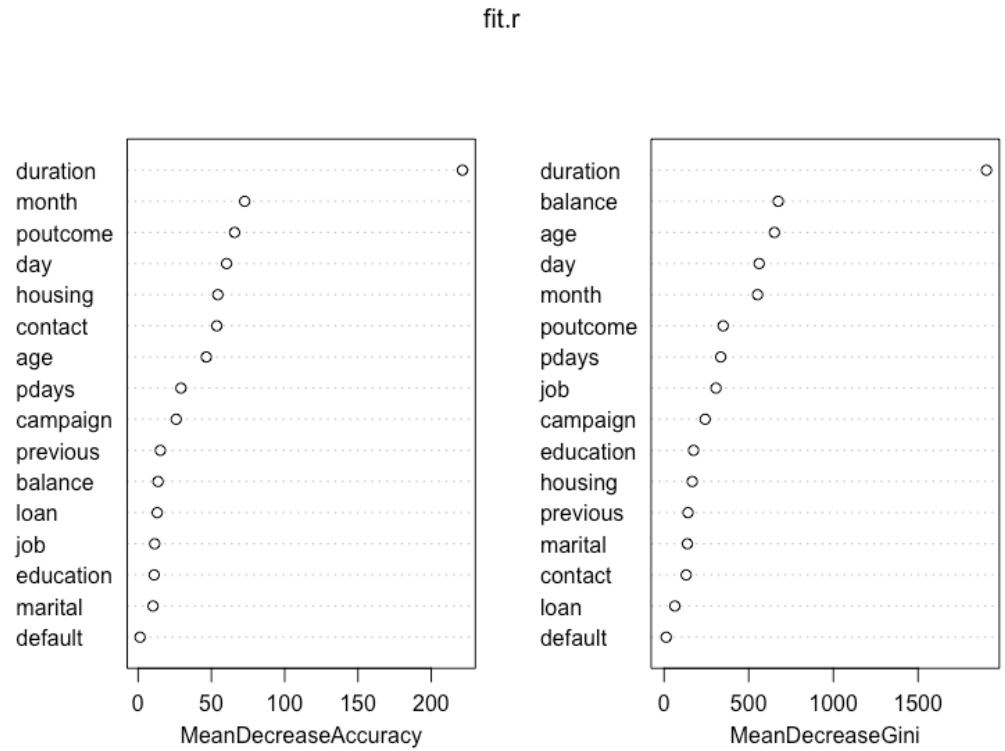


Figure 9 "Variable Importance", own results

The Variable importance plot above shows the Mean Decrease Accuracy values on the left and mean decrease Gini on the right (Cutler, et al., 2007) about the importance given by the model to each variable. Variable with a large mean accuracy are the most important (duration of the marketing call with a 250 accuracy), followed by almost the same level for month and outcome.The mean Decrease Gini measured how each variable contributed to the homogeineity of the nodes, thus it can be considered more important for the evaluation. The graph shows the highest value for “Duration”,followed by account “balance” and customer’s “age”.

6 Conclusion

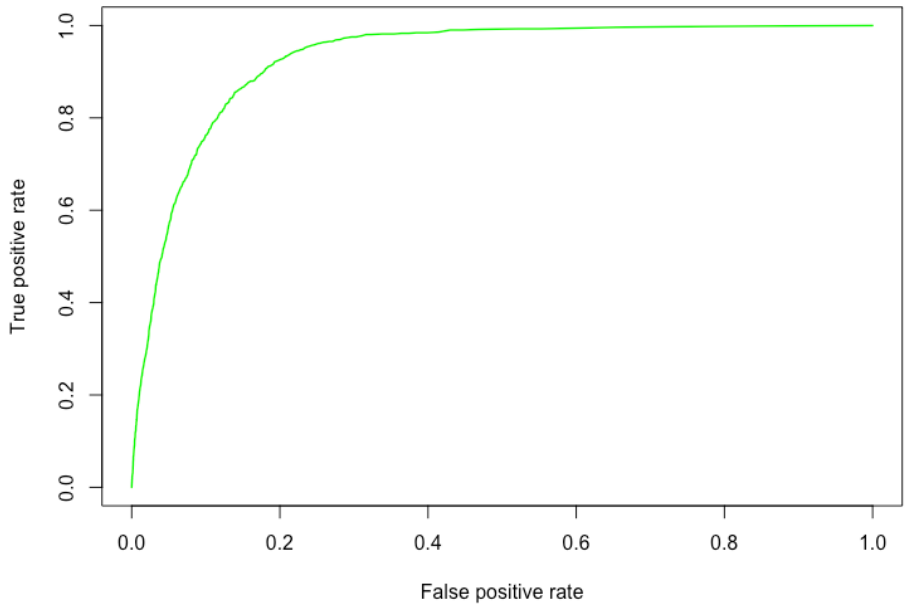


Figure 10 " Comparison of Lift Curve and ROC Curve of Random Forest Model", own results

With a 0.90 area under prediction, as 1 would equal to 100% correctness (Vuk,2006), It can be stated that based on the data the random forest more accurately improved the prediction initially set by the classification tree with a strong model.

7 References

Basile, A., Dwyer, G. & Medvedeva, M., 2017. . . *N-GrAM: New Groningen Author-profiling Model..* [Online] Available at: https://www.researchgate.net/figure/Decision-Tree-output_fig1_318392575 [Accessed 21 July 2020].

Bennette, W. D., 2011. *Instance selection for simplified decision treesthrough the generation and selection of instancecandidate subsets.* Iowa: Iowa State University Capstones.

Cutler, D. et al., 2007. Random Forests for Classification in Ecology.. *Ecology.,* Volume 88, pp. 2783-92.

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Lantz, B., 2015. *Machine Learning with R.* Second ed. Birmingham: Packt. Milborrow, S., 2019. *Plot 'rpart' Models: An Enhanced Version of 'plot.rpart'.* [Online] Available at: https://cran.r-project.org/web/packages/rpart.plot/rpart.plot.pdf [Accessed 18 July 2020].

Moro, S., Cortez, P. & Rita, P., 2014. *A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems.* [Online] Available at: https://archive.ics.uci.edu/ml/datasets/Bank+Marketing [Accessed 20 July 2020].

Riou, M.-E., Rioux, F., Lamothe, G. & Doucet , É., 2015. *Validation and Reliability of a Classification Method to Measure the Time Spent Performing Different Activities.* [Online]
Available at: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0128299 [Accessed 20 July 2020].

Utgo, P. E. & Clouse, J. A., 1996. *A Kolmogorov-Smirnoff Metric for Decision Tree Induction,* Amherst: Department of Computer Science, University of Massachusetts.

Vuk, M., 2006. *ROC Curve, Lift Chart and Calibration Plot.* [Online] Available at: https://www.semanticscholar.org/paper/ROC-Curve%2C-Lift-Chart-and-Calibration-Plot-Vuk/ebe9a6b158bb20275e78a6bf35371de6b0523344 [Accessed 20 July 2020].