PROJECT

STAN45: Data Mining and Visualisation

Predictions for a Portugese Banking Institution.

Meera Damaraju & Sebastian Martensson

Disposition:

1. Background
2. Objective
3. Methods and rationale
4. Inference Objective
5. Predictive Objective
6. Conclusion and Recommendation

Background:

-Direct marketing campaign for a Portugese banking Institution.

-Asked about the subscription of a bank term deposit?

-Personal characteristics were collected.

2 objectives: Inference & Predictive

Objective 1 - Inference Objective:

-What can the bank do? What gives us a better success rate in providing a loan to a client?

-Drive growth.

-Method used: Classification tree

Objective - 2 - Predictive Objective

-Can we properly identify people that would subscribe to a term deposit?

-Useful for validating the Inference objective.

-Method used: Random Forests.

Data Overview:

-In total 41,188 observations of 20 features and 1 outcome variable.

-Success rate of 11%

-7 variables - individual characteristics.

-6 variables - contemporary and previous campaigns.

-5 variables -social and economic context.

-2 variables -the date the contact was made

Data - 2

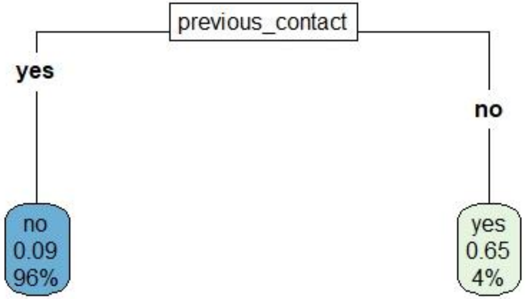
1) age: numeric
2) job: categorical; 12 categories
3) marital status: categorical; 3 categories
4) education: categorical; 8 categories
5) default: categorical: 'no','yes','unknown'
6) housing: categorical: 'no','yes','unknown'
7) loan: categorical: 'no','yes','unknown'
8) contact: categorical: 'cellular','telephone'
9) Month: Categorical
10) day_of_week: Categorical
11) duration: Numeric
12) campaign: Numeric
13) pdays: Numeric
14) previous: Numeric.
15) poutcome: Categorical: "failure","nonexistent","success"
16) emp.var.rate: numeric
17) cons.price.idx: numeric
18) cons.conf.idx: numeric
19) euribor3m: numeric
20) nr.employed: numeric
21) y: binary: "yes","no" - outcome variable

Classification trees

- Inference Objective
- Unweighted and Weighted data
- Pruned using cost complexity
- Variables - directly affect or perfectly predict
- Feature “Pdays” turned into factor

Inference Objective - Base

Pruned and unweighted



Should contact those who have not been contacted before

Result from Validation set

		Real	
		no	yes
Predicted	no	7196	751
	yes	113	177

Accuracy	0.86
False Negative Rate	0.81
False Positive Rate	0.02

Inference Objective - Base

Pruned and unweighted

Decrease false negative rate (FNR)

Result from Validation set

		Real	
		no	yes
Predicted	no	7196	751
	yes	113	177

Accuracy	0.86
False Negative Rate	0.81
False Positive Rate	0.02

Inference Objective - Weighting

Reweight so that classifying “Yes” correctly is more important

Weight of Yes / Weight of No= 7.88

One “Yes” is worth 7.88 “No”´s

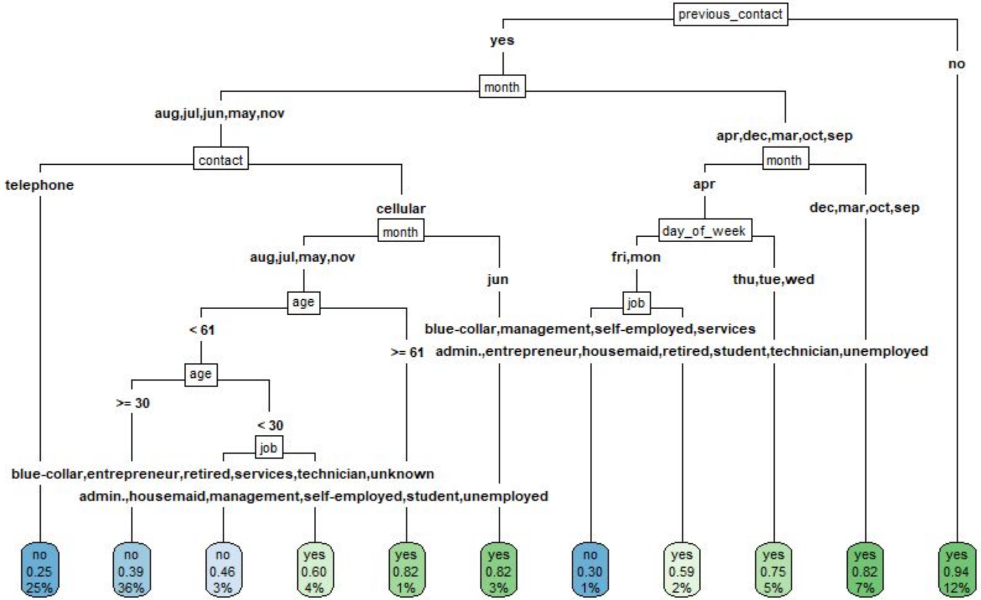
Inference Objective - Weighting

- Recommendation:

1. Contact those who haven't been contacted before
2. March, September, October and December
3. Above age 60 using a cell phone

Still not good enough FNR

Accuracy	0.84
False Negative Rate	0.47
False Positive Rate	0.13



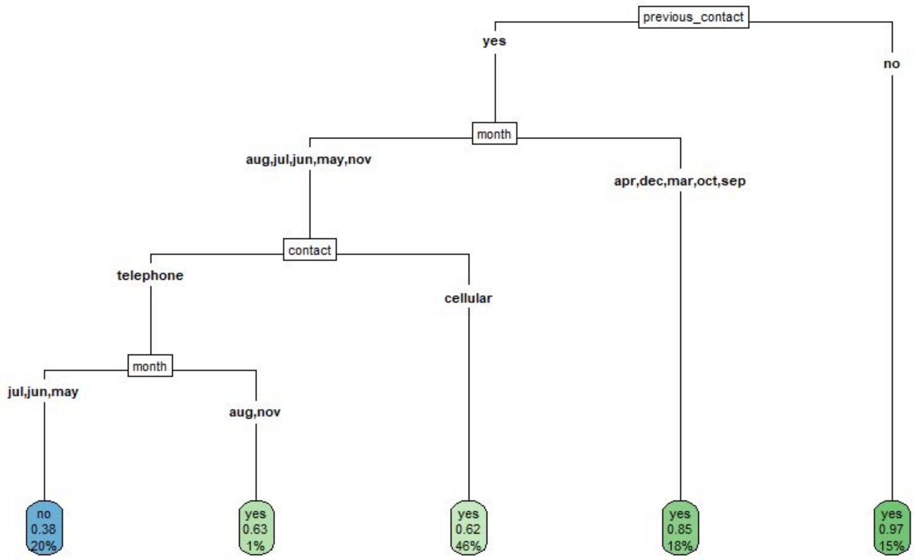
Inference Objective - Reweighting

Further doubling the weight of "Yes"

- Recommendations:

1. Contact those who haven't been contacted before
2. April, March, September, October and December
3. Cell phone

Accuracy	0.42
False Negative Rate	0.11
False Positive Rate	0.63



Random Forests: Overview

An overview as to what we need to determine effectiveness:

1) We need features with at least some predictive power- attested by previous section
2) The trees of the forest need to be uncorrelated to each other- fulfilled via the package.

Procedure:

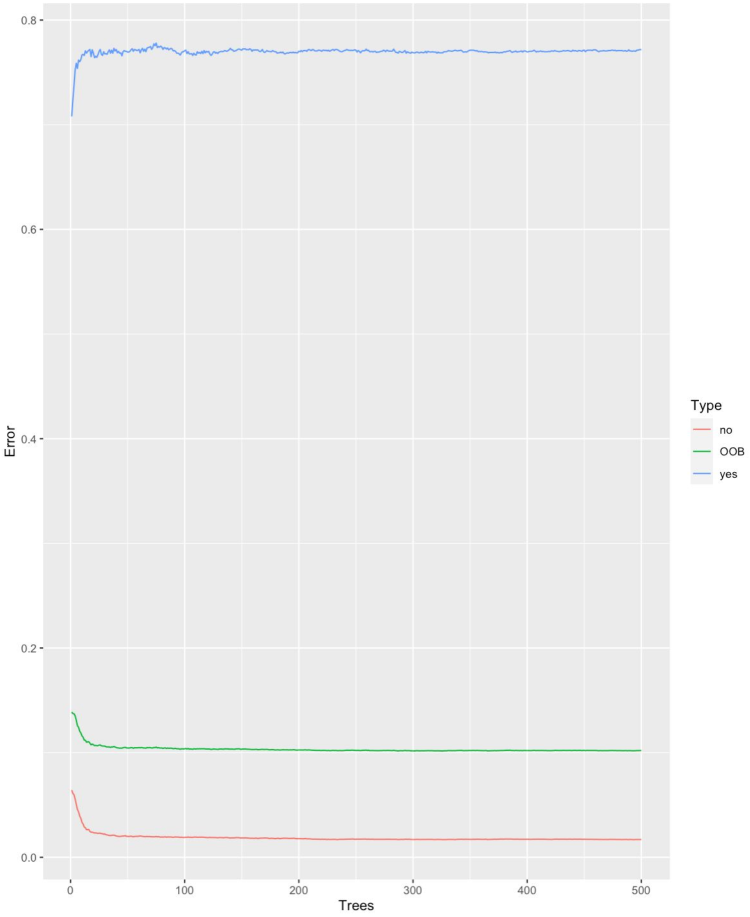
1) Division of the data set.
2) Generate of the random forest and calculation of the OOB estimates.
3) Fine-tuning the parameters.
4) Comparing accuracy for validation set.
5) 16 variables versus 21 variables : an overview and comparison

Out-of-Bag Error estimate - I

- For random forest set with 16 variables.
- Blue: Error rate of ‘yes’
- Green: Out of bag error rate
- Red: Error rate of ‘no’.
- Higher false positive than false negative.

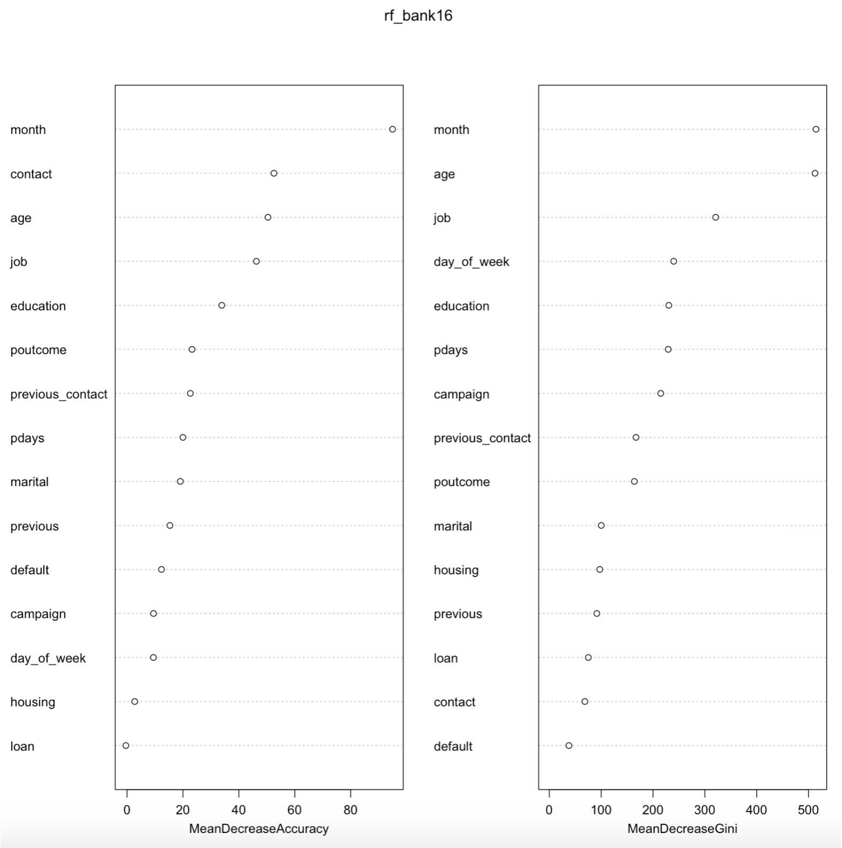
```
Confusion matrix:
        no yes class.error
no   21555 373  0.01701022
yes   2148 636  0.77155172
```

- OOB estimate of error rate: 10.2%



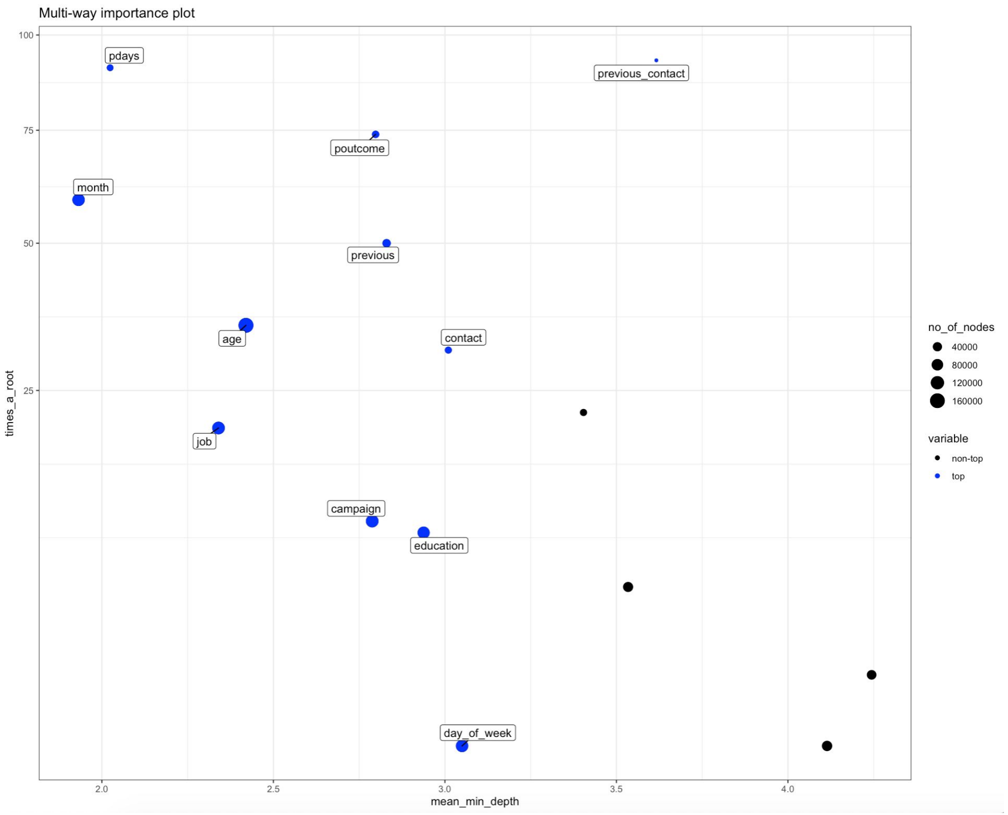
V.I.P of the Random Forest-I

- Predictions of training versus validation set: 93.62%. Validation: 1575 miss-classified.
- Month is the best.
- Others unclear.



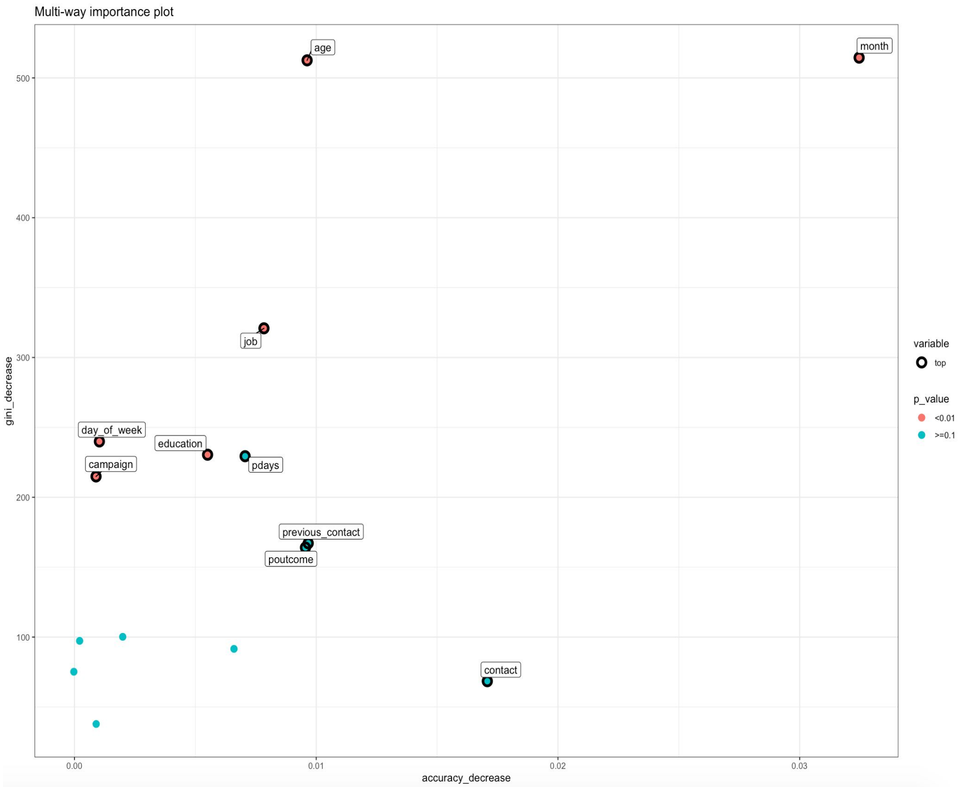
Multi-way Importance Plot- I

- Measures how important the variable is, based on the depth of the tree.
- Examines structure of the forest
- ‘Pdays’ > previous contact
- Month and age use by random forest - but not the top variable.



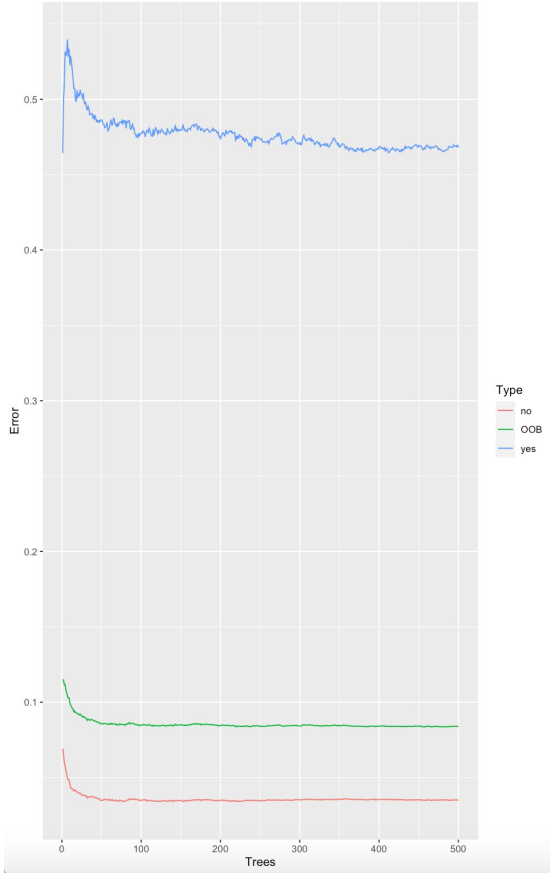
Multi-way Importance Plot-II

- Plot showing variable importance based on decrease in accuracy and decrease in gini.
- Structure versus prediction.
- Month is best.



The Out-of-Bag Error Estimate-II

- This is for the training set.
- Key takeaway: False positive >> false negative



Base Parameters of the random forest

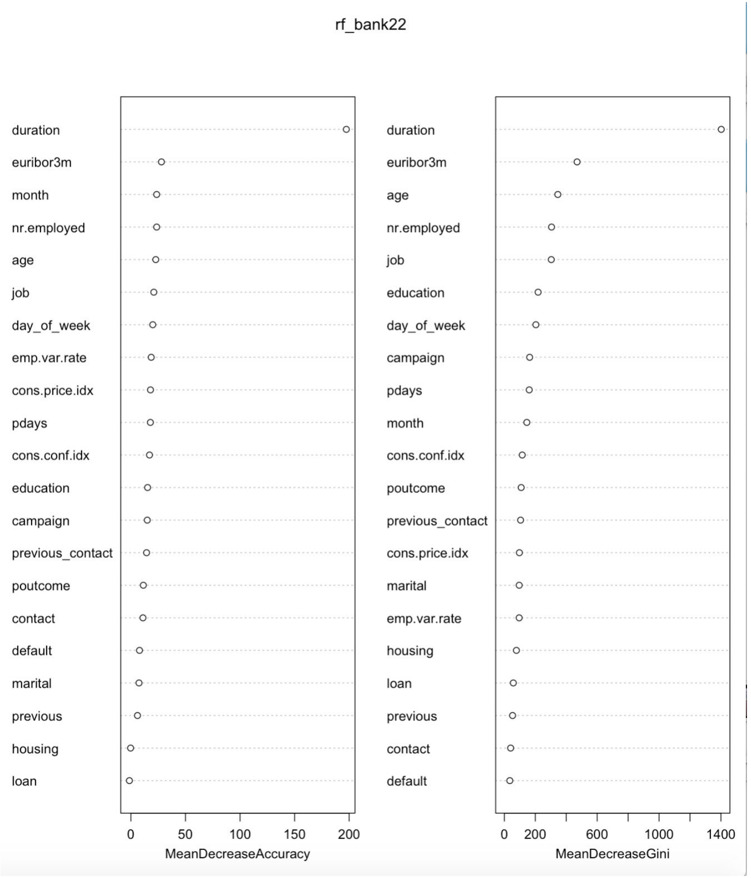
- Not possible to determine manually.
- Statistics:
- Number of trees: 500
- No. of variables tried at each split: 4
- OOB estimate of error rate: 8.39%

Confusion matrix:

	no	yes	class.error
no	21162	766	0.03493251
yes	1307	1477	0.46946839

V.I.P of the Random Forest-II

- Predictions of training versus validation set: 91.48%. Training: 92. Validation: 702 miss-classified.

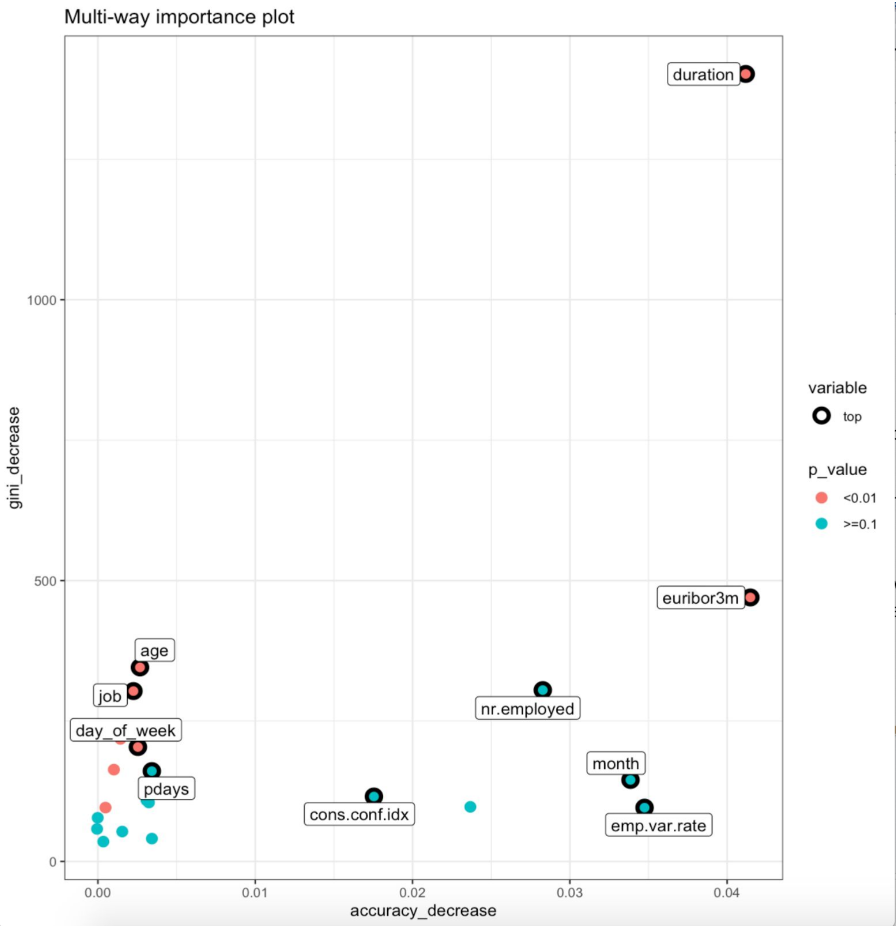


Multi-way Importance Plot-III

- Clear negative relationship between y and x.

Multi-way Importance Plot IV

- Duration consistently drives the forest- contrast with earlier study.



Comparison in accuracy:

- Model (1) that includes socio-economic variables has higher predictive accuracy even though the other model (2) initially suggested a better predictive power.

	Model	Accuracy
1	Random Forests	0.9150210
2	Random Forests	0.8981062

- Predictive power of 96.14% on test set for model with 22 attributes.

predTest22	no	yes
no	7211	220
yes	98	708

Limitations:

Classification Trees:

- Did not use Boosting because of time constraints.

Random Forests:

- Software limitations, unable to generate proximity matrix.

Conclusion

-Duration - month - age are significant variables.

-No previous contact, cellphone, month.

-Key stand out variables: Duration for 20 attributes, Month for 16

-The Trade-off between practicality versus interpretability.

	Accuracy	False Negative Rate	False Positive Rate
Classification tree (first reweight)	83.77%	46%	12%
Random Forest	96.14%	23.71%	1.34%

Appendix A: Packages

Packages used

- Rpart
- Rpart.plot
- Groupdata2
- RandomForest
- Readr
- Cowplot
- randomForestExplainer
- tidymodels

Appendix B: Detailed info on data

1 - age (numeric)

2 - job : type of job (categorical: "admin.","blue-collar","entrepreneur","housemaid","management","retired","self-employed","services","student","technician","unemployed","unknown")

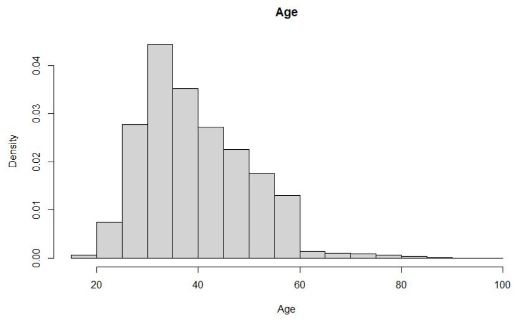
3 - marital : marital status (categorical: "divorced","married","single","unknown"; note: "divorced" means divorced or widowed)

4 - education (categorical: "basic.4y","basic.6y","basic.9y","high.school","illiterate","professional.course","university.degree","unknown")

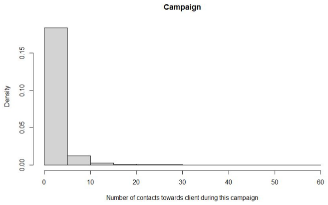
5 - default: has credit in default? (categorical: "no","yes","unknown")

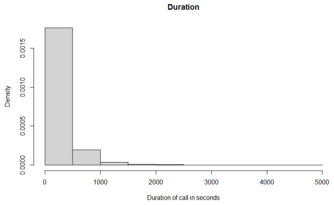
6 - housing: has housing loan? (categorical: "no","yes","unknown")

7 - loan: has personal loan? (categorical: "no","yes","unknown")



Appendix B: Detailed info on data-2

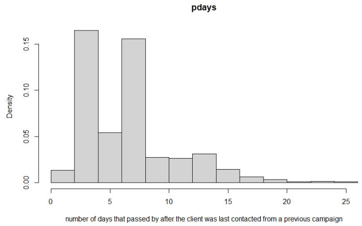




8 - contact: contact communication type (categorical: "cellular","telephone")

9 - month: last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")

10 - day_of_week: last contact day of the week (categorical: "mon","tue","wed","thu","fri")

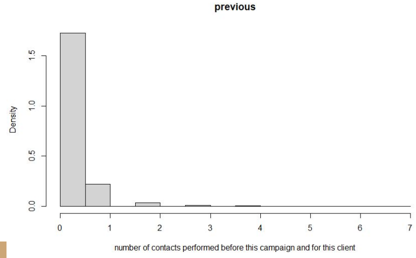


11 - duration: last contact duration, in seconds (numeric).

other attributes:

12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)

13 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)



14 - previous: number of contacts performed before this campaign and for this client (numeric)

15 - poutcome: outcome of the previous marketing campaign (categorical: "failure","nonexistent","success")

Appendix - detailed info on data

social and economic context attributes

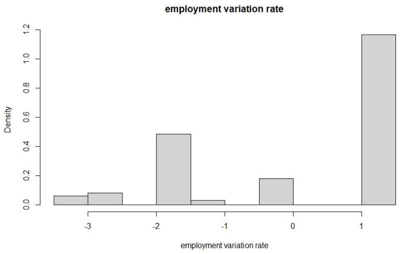
16 - emp.var.rate: employment variation rate - quarterly indicator (numeric)

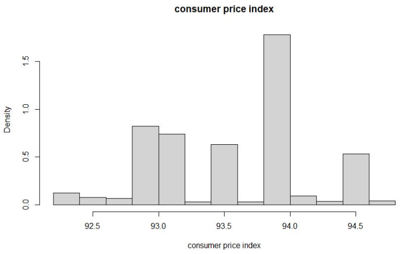
17 - cons.price.idx: consumer price index - monthly indicator (numeric)

18 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)

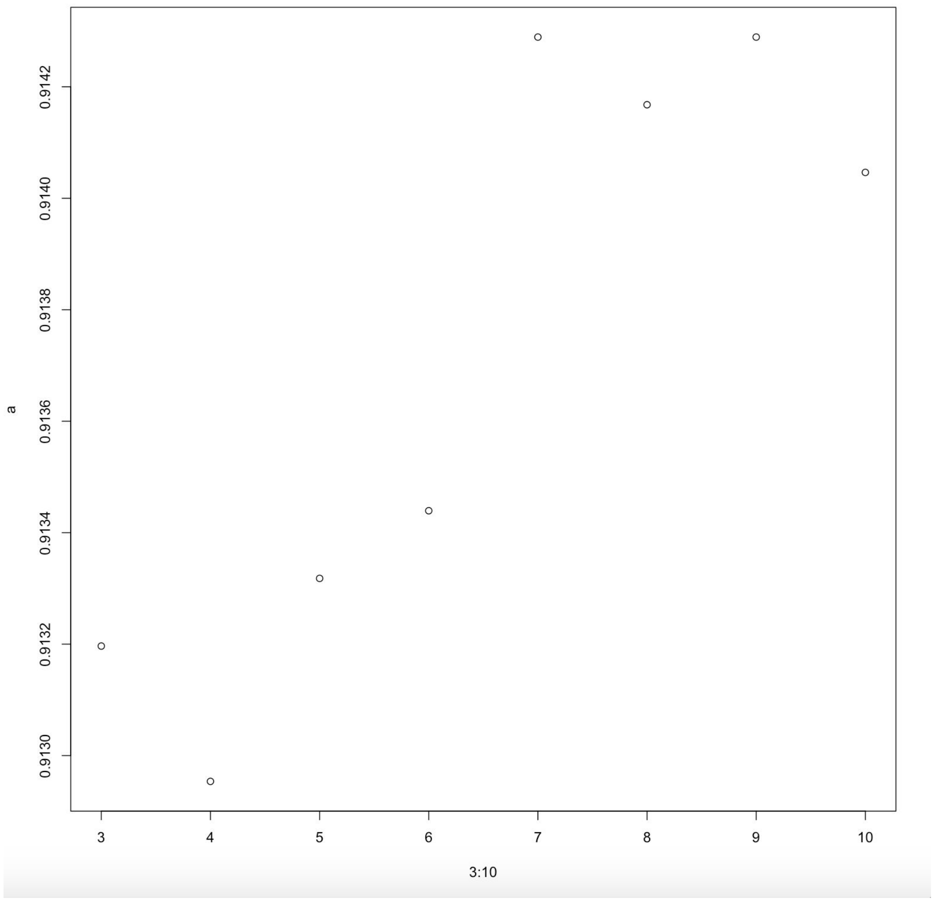
19 - euribor3m: euribor 3 month rate - daily indicator (numeric)

20 - nr.employed: number of employees - quarterly indicator (numeric)





Appendix B: Mtry values, and manual execution.



Appendix D: Importance breakdown (20 attributes data set)

	variable	mean_min_depth	no_of_nodes	accuracy_decrease	gini_decrease	no_of_trees	times_a_root	p_value
1	age	3.556000	97689	2.669945e-03	345.48735	500	4	0.000000e+00
2	campaign	4.154000	60020	1.013195e-03	163.48891	500	2	0.000000e+00
3	cons.conf.idx	3.202000	19807	1.754288e-02	115.59953	500	23	1.000000e+00
4	cons.price.idx	3.090000	19971	2.367212e-02	97.19286	500	16	1.000000e+00
5	contact	4.110000	11786	3.427760e-03	40.65314	500	4	1.000000e+00
6	day_of_week	3.674000	60946	2.537495e-03	203.70115	500	0	0.000000e+00
7	default	5.560000	16133	3.408643e-04	35.28442	500	0	1.000000e+00
8	duration	1.740000	119550	4.117944e-02	1401.93947	500	64	0.000000e+00
9	education	3.660000	62557	1.427262e-03	218.20077	500	0	0.000000e+00
10	emp.var.rate	3.434000	10483	3.474827e-02	95.76910	500	23	1.000000e+00
11	euribor3m	2.132000	79174	4.147375e-02	469.98694	500	74	0.000000e+00
12	housing	5.142000	35228	-1.556710e-05	77.74819	500	0	1.000000e+00
13	job	3.230000	72659	2.254049e-03	303.30309	500	1	0.000000e+00
14	loan	5.188000	25293	-5.656314e-05	57.89062	500	0	1.000000e+00
15	marital	4.678000	38943	4.774257e-04	95.90936	500	0	6.490547e-05
16	month	2.720000	20681	3.385426e-02	145.09951	500	23	1.000000e+00
17	nr.employed	2.168000	10974	2.827810e-02	305.05958	500	108	1.000000e+00
18	pdays	2.484000	8503	3.425183e-03	160.83568	500	57	1.000000e+00
19	poutcome	3.576000	12121	3.091133e-03	108.68637	500	39	1.000000e+00
20	previous	4.356000	17942	1.542784e-03	53.20811	500	10	1.000000e+00
21	previous_contact	4.680632	1959	3.237750e-03	104.86973	483	52	1.000000e+00

Appendix D: Importance breakdown (16 attributes data set)

	variable	mean_min_depth	no_of_nodes	accuracy_decrease	gini_decrease	no_of_trees	times_a_root	p_value
1	age	2.4200	163856	9.623157e-03	512.57677	500	35	0
2	campaign	2.7880	103888	8.920558e-04	214.90455	500	10	0
3	contact	3.0100	15382	1.707729e-02	68.44714	500	31	1
4	day_of_week	3.0500	99132	1.030435e-03	239.95076	500	0	0
5	default	3.4040	17937	8.982097e-04	37.82384	500	22	1
6	education	2.9380	94786	5.506494e-03	230.45403	500	9	0
7	housing	4.1140	60450	2.136203e-04	97.30178	500	0	1
8	job	2.3400	105015	7.837808e-03	320.87488	500	20	0
9	loan	4.2440	48582	-2.549082e-05	75.20270	500	1	1
10	marital	3.5340	55729	1.996090e-03	100.18684	500	5	1
11	month	1.9320	99555	3.245766e-02	514.49697	500	59	0
12	pdays	2.0240	12839	7.058746e-03	229.38470	500	91	1
13	poutcome	2.7980	18545	9.554907e-03	164.03909	500	74	1
14	previous	2.8300	29639	6.597808e-03	91.56607	500	50	1
15	previous_contact	3.6164	1384	9.669425e-03	167.14595	485	93	1