SUCCESS OF BANK TELEMARKETING

A DATA DRIVEN APPROACH

Prepared by: Wone Eui Jung
Yiting Li
Yuchen Bi
Tianqing Huang
Sebastian Marx
Sohrab Gill

TABLE OF CONTENTS

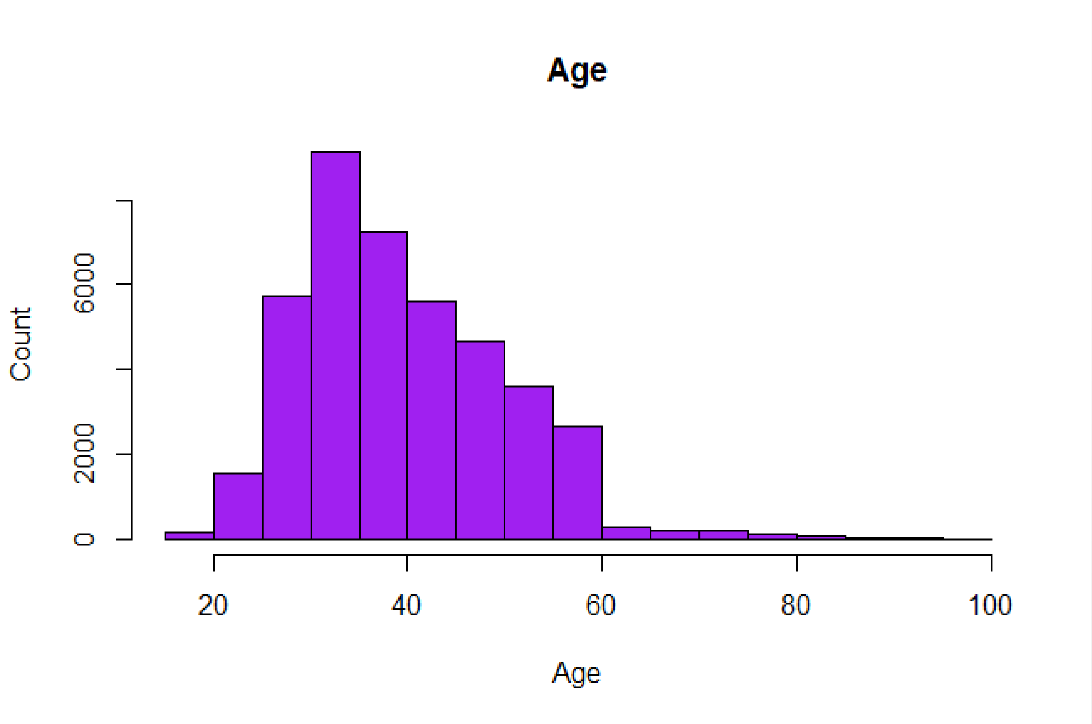
INTRODUCTION

- Our data contains several variables with descriptions, have understandable context about useful real-life marketing
- Portuguese retail bank aims to sell bank deposits to acquire capital post 2008 financial crisis
- Bank products, client profile and social economic important attributes
- We use various models to shortlist important variables

Executive Summary

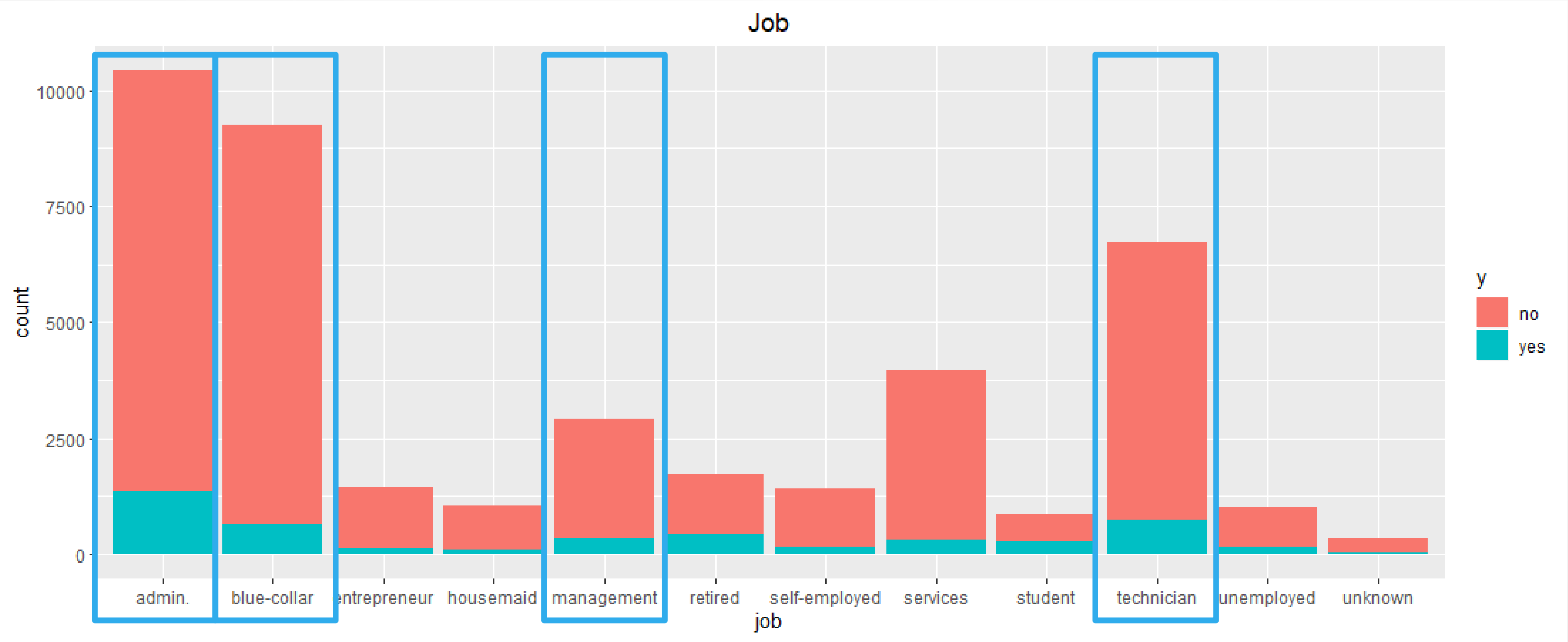
- Analyzed Portuguese Bank marketing dataset from UCI machine learning
- Data cleaning, data exploration, and data manipulation
- Under sampling imbalanced training data set to detect fewer y=1 samples.
- Classification using KNN, decision trees, logistic regression to analyze the impact of variables for the marketing results.
- Random forest and XGBoost to predict whether the client will subscribe to a term deposit (variable y) for telephone marketing.
- The results showed the call duration, and the Euribor3m are the most important factor.

DATA PREPROCESSING
Data Exploration



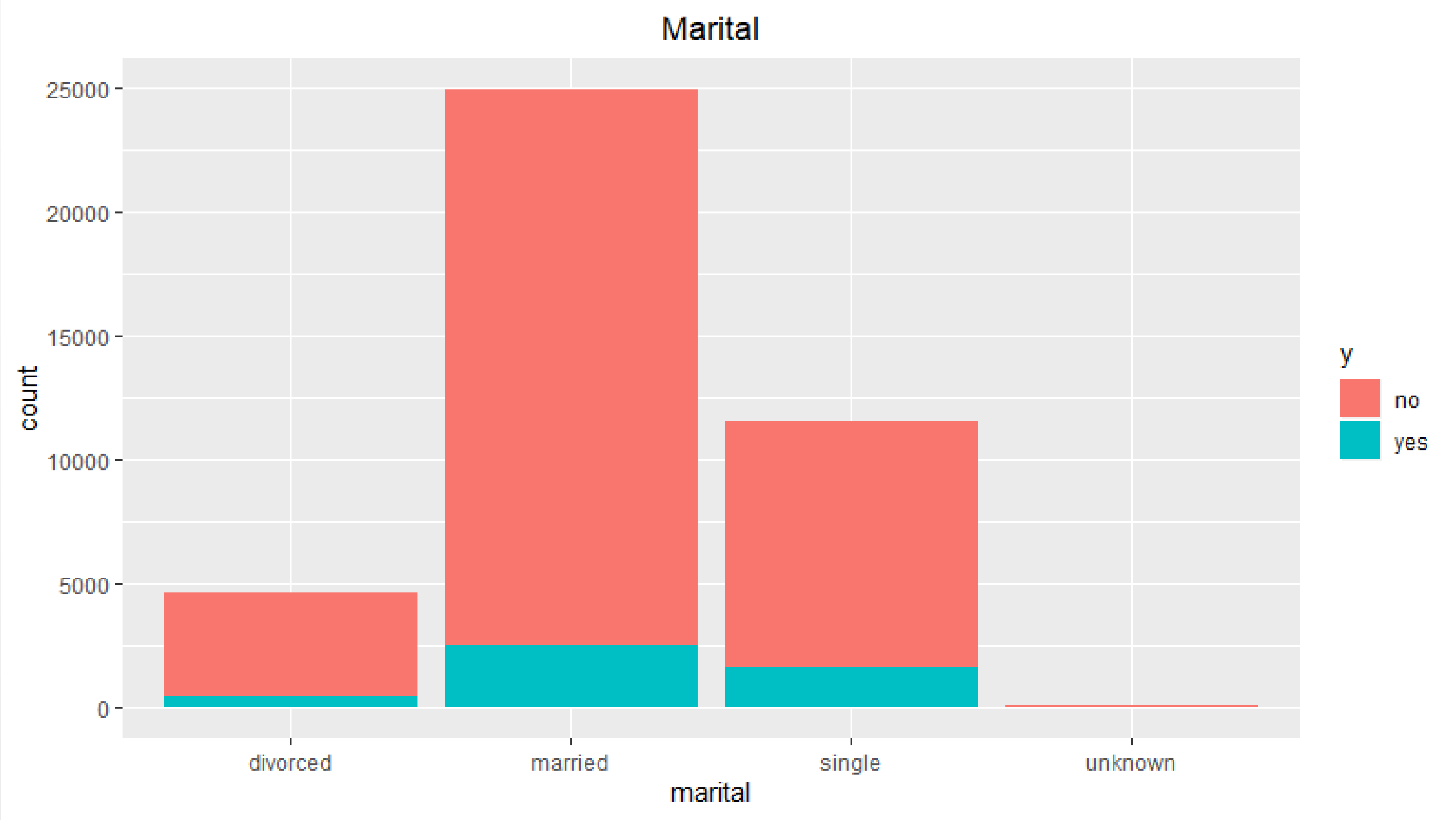
50% of the data is concentrated between age 32 to 47

DATA PREPROCESSING
Data Exploration



The top four jobs (admin, blue-collar, technician, management) takes part 80% of data.

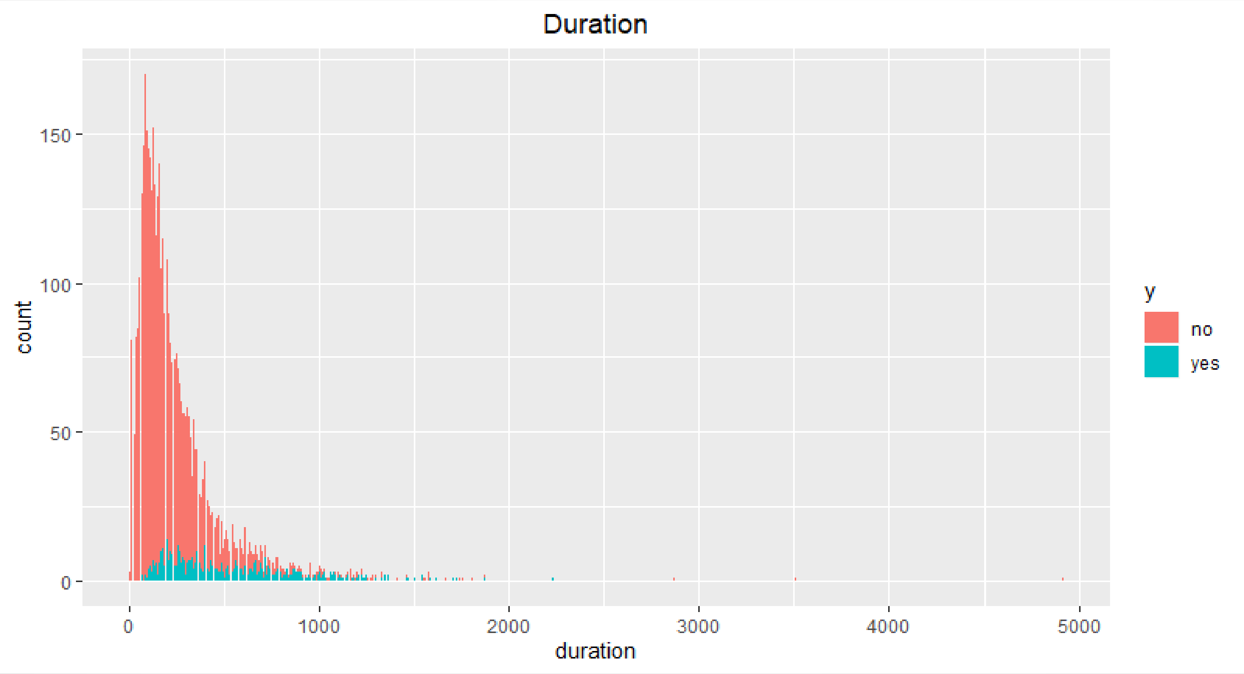
DATA PREPROCESSING
Data Exploration



Single and unknown group tends to have higher rate for subscribing the term deposit than other groups. For the marketing, bank can put more effort for those groups

DATA PREPROCESSING

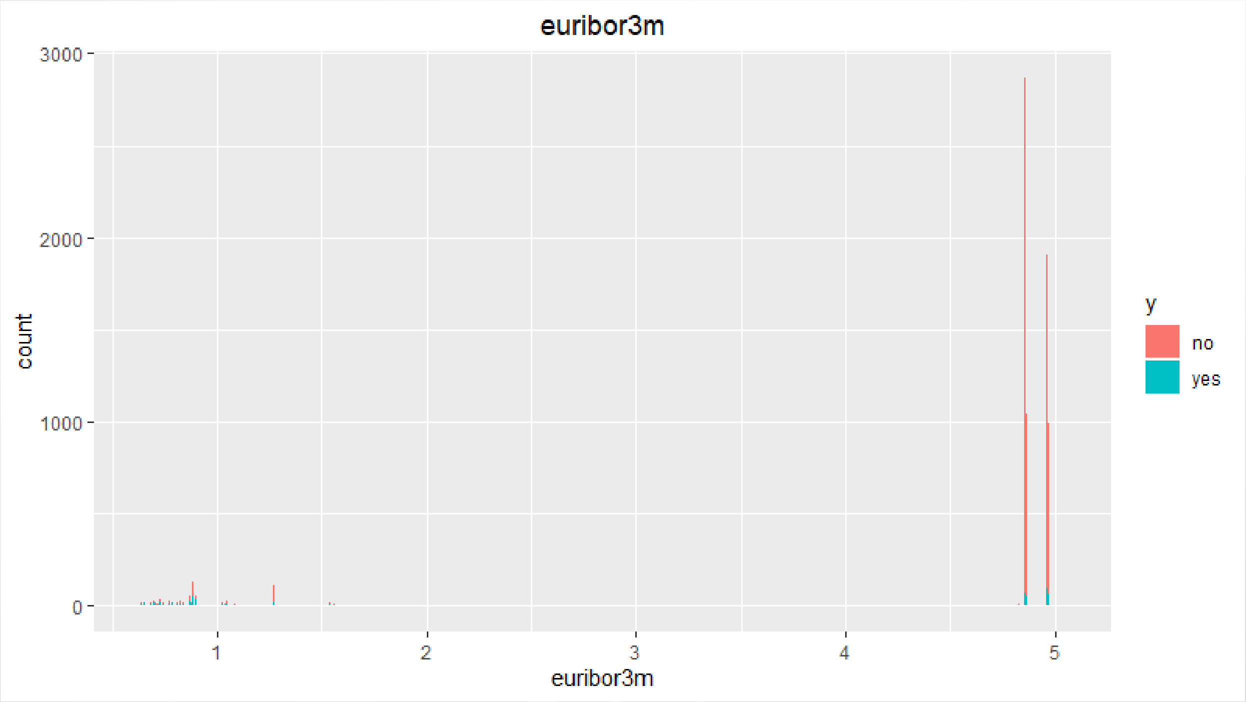
Data Exploration



Duration of phone calls chart shows shorter duration calls have maximum counts but most of them resulting in no sale.

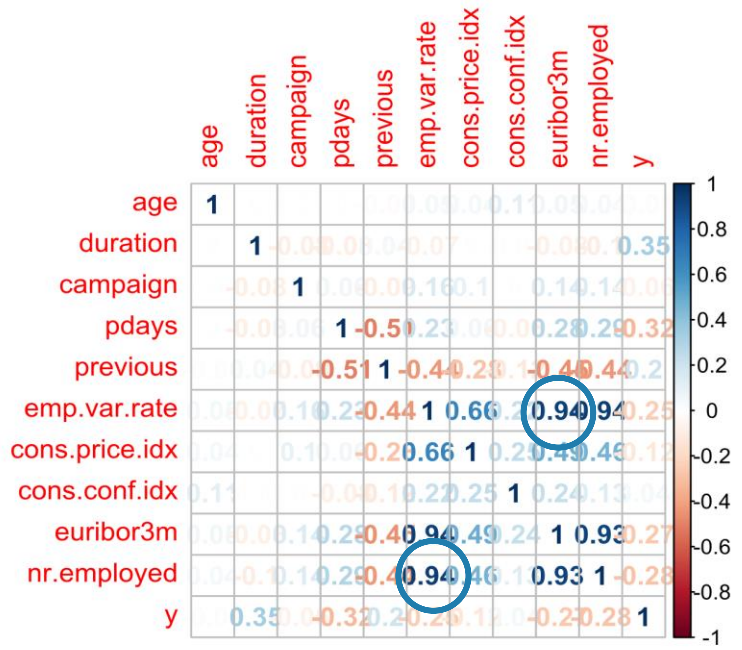
DATA PREPROCESSING
Data Exploration

Bank Loan Interest Rate



DATA PREPROCESSING
Data Cleaning

- Our dataset did not have any missing values
- Unbalanced dataset – no outliers deleted
- Numerical variables Euribor3m, employment variation rate and nr.employed rate (number of employees)



DATA PREPROCESSING
Data Partition

- Original Dataset – 41,188 records, 21 variables
- Sample had 11.3% Yes ; We used 50% for model building in training dataset
- Remaining 4.66% Yes and 95.34% No for Validation Data

Target(y)	raw data	Training data	Validation data
yes	4640	3000	1640
no	36548	3000	33548
total	41180	6000	35188

CLASSIFICATION
Logistic Regression

Variable Selection

- Forward/Stepwise
- Backward

CLASSIFICATION
Logistic Regression

- Forward/Stepwise – Coefficients

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.044e+01	7.911e+00	-2.584	0.00978	**
duration	6.692e-03	1.990e-04	33.633	< 2e-16	***
euribor3m	-7.778e-01	3.613e-02	-21.528	< 2e-16	***
month_may	-1.274e+00	1.089e-01	-11.705	< 2e-16	***
poutcome_success	1.209e+00	4.646e-01	2.603	0.00924	**
age_E	1.006e+00	2.072e-01	4.856	1.20e-06	***
education_university.degree	3.830e-01	8.615e-02	4.445	8.78e-06	***
poutcome_nonexistent	4.986e-01	1.227e-01	4.065	4.81e-05	***
cons.conf.idx	3.368e-02	7.777e-03	4.330	1.49e-05	***
cons.price.idx	2.317e-01	8.534e-02	2.715	0.00663	**
default_no	3.334e-01	1.229e-01	2.712	0.00670	**
age_A	1.363e+00	5.332e-01	2.555	0.01061	*
age_B	2.724e-01	9.692e-02	2.811	0.00494	**
marital_unknown	-1.765e+00	7.816e-01	-2.258	0.02393	*
campaign	-4.580e-02	2.151e-02	-2.129	0.03325	*
month_nov	-2.464e-01	1.460e-01	-1.688	0.09138	.
day_of_week_wed	1.929e-01	9.955e-02	1.938	0.05265	.
pdays	-7.321e-04	4.495e-04	-1.629	0.10337	
job_technician	1.791e-01	1.083e-01	1.654	0.09813	.
month_Other_Month	1.917e-01	1.203e-01	1.594	0.11095	

CLASSIFICATION
Logistic Regression

- Forward / Stepwise – OR, SE, p-value

	OR	SE	95% CI, lower	95% CI, upper	p value
duration	1.0067145	0.0002003089	0.0070820213	0.0070820213	5.539511e-248
euribor3m	0.4594111	0.0165983896	-0.7069969025	-0.7069969025	8.465018e-103
month_may	0.2796825	0.0304442648	-1.0607522645	-1.0607522645	1.204651e-31
poutcome_success	3.3513169	1.5570767318	2.1199847476	2.1199847476	9.243694e-03
age_E	2.7355804	0.5668702835	1.4124897436	1.4124897436	1.195549e-06
education_university.degree	1.4666238	0.1263508627	0.5518155345	0.5518155345	8.778499e-06
poutcome_nonexistent	1.6464654	0.2019875163	0.7390781821	0.7390781821	4.813537e-05
cons.conf.idx	1.0342510	0.0080437749	0.0489209267	0.0489209267	1.489861e-05
cons.price.idx	1.2606951	0.1075820178	0.3989177031	0.3989177031	6.632873e-03
default_no	1.3956774	0.1715922960	0.5743486962	0.5743486962	6.695804e-03
age_A	3.9062533	2.0829074339	2.4076782265	2.4076782265	1.060782e-02
age_B	1.3131235	0.1272679603	0.4623684156	0.4623684156	4.944090e-03
marital_unknown	0.1711651	0.1337843058	-0.2331998261	-0.2331998261	2.392557e-02
campaign	0.9552307	0.0205497999	-0.0036378846	-0.0036378846	3.324877e-02
month_nov	0.7815818	0.1140946160	0.0396783733	0.0396783733	9.138140e-02
day_of_week_wed	1.2127588	0.1207244048	0.3880029414	0.3880029414	5.264888e-02
pdays	0.9992681	0.0004491921	0.0001489019	0.0001489019	1.033726e-01
job_technician	1.1961029	0.1294955492	0.3912633083	0.3912633083	9.812921e-02
month_Other_Month	1.2113268	0.1456976607	0.4274595901	0.4274595901	1.109529e-01

CLASSIFICATION
Logistic Regression

- Forward / Stepwise – Confusion Matrix

		Actual Class	
Predicted Class		0	1
	0	28723	203
	1	4825	1437

Cutoff = 0.5

- Sensitivity = 0.8762
- Specificity = 0.8561

		Actual Class	
Predicted Class		0	1
	0	25912	71
	1	7636	1569

Cutoff = 0.3

- Sensitivity = 0.9567
- Specificity = 0.7723

CLASSIFICATION
Logistic Regression

- Backward – Coefficients

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.108e+01	7.766e+00	-2.714	0.006654	**
duration	6.692e-03	1.990e-04	33.629	< 2e-16	***
campaign	-4.493e-02	2.149e-02	-2.091	0.036535	*
pdays	-1.336e-03	1.799e-04	-7.427	1.11e-13	***
cons.price.idx	2.459e-01	8.373e-02	2.937	0.003312	**
cons.conf.idx	3.329e-02	7.762e-03	4.289	1.80e-05	***
euribor3m	-7.768e-01	3.599e-02	-21.583	< 2e-16	***
age_B	-7.636e-01	2.146e-01	-3.558	0.000373	***
age_C	-1.023e+00	2.002e-01	-5.109	3.24e-07	***
age_D	-9.498e-01	2.054e-01	-4.625	3.75e-06	***
marital_married	1.712e+00	7.642e-01	2.240	0.025070	*
marital_single	1.790e+00	7.648e-01	2.341	0.019252	*
marital_divorced	1.678e+00	7.721e-01	2.173	0.029797	*
education_Basic_education	-4.176e-01	1.062e-01	-3.931	8.47e-05	***
education_high.school	-4.036e-01	1.041e-01	-3.876	0.000106	***
education_professional.course	-2.193e-01	1.263e-01	-1.737	0.082392	.
default_no	3.420e-01	1.242e-01	2.754	0.005890	**
month_may	-1.210e+00	1.045e-01	-11.587	< 2e-16	***
month_Other_Month	2.492e-01	1.158e-01	2.152	0.031415	*
day_of_week_wed	1.940e-01	9.952e-02	1.949	0.051299	.
poutcome_failure	-5.634e-01	1.191e-01	-4.730	2.24e-06	***

CLASSIFICATION
Logistic Regression

- Backward – OR, SE, p-value

	OR	SE	95% CI, lower	95% CI, upper	p value
duration	1.0067140	0.0002003190	0.0070815749	0.0070815749	6.350292e-248
campaign	0.9560630	0.0205447067	-0.0028141135	-0.0028141135	3.653516e-02
pdays	0.9986649	0.0001796377	-0.0009834193	-0.0009834193	1.109931e-13
cons.price.idx	1.2787948	0.1070690362	0.4100190225	0.4100190225	3.312347e-03
cons.conf.idx	1.0338475	0.0080242220	0.0484995683	0.0484995683	1.796725e-05
euribor3m	0.4598690	0.0165516920	-0.7062701807	-0.7062701807	2.603822e-103
age_B	0.4659822	0.0999999270	-0.3429988919	-0.3429988919	3.732925e-04
age_C	0.3595987	0.0719874275	-0.6304049353	-0.6304049353	3.238262e-07
age_D	0.3868327	0.0794405606	-0.5472617965	-0.5472617965	3.748991e-06
marital_married	5.5402672	4.2338473421	3.2098384177	3.2098384177	2.507014e-02
marital_single	5.9900128	4.5811575948	3.2890726350	3.2890726350	1.925249e-02
marital_divorced	5.3528284	4.1329750850	3.1909338973	3.1909338973	2.979700e-02
education_Basic_education	0.6586551	0.0699682836	-0.2093501706	-0.2093501706	8.469615e-05
education_high.school	0.6679135	0.0695482921	-0.1995100696	-0.1995100696	1.061975e-04
education_professional.course	0.8030458	0.1014079553	0.0281590452	0.0281590452	8.239184e-02
default_no	1.4077866	0.1748436669	0.5854414220	0.5854414220	5.890282e-03
month_may	0.2980666	0.0311373663	-1.0056916778	-1.0056916778	4.791755e-31
month_Other_Month	1.2829725	0.1485706467	0.4761471728	0.4761471728	3.141508e-02
day_of_week_wed	1.2140540	0.1208247727	0.3890242417	0.3890242417	5.129862e-02
poutcome_failure	0.5692807	0.0678019995	-0.3299476898	-0.3299476898	2.242124e-06

CLASSIFICATION
Logistic Regression

- Backward – Confusion Matrix

Predicted Class \ Actual Class	0	1
0	28741	203
1	$\boxed{4807}$	$\boxed{1437}$

Cutoff = 0.5

- Sensitivity = 0.8762
- Specificity = 0.8567

Predicted Class \ Actual Class	0	1
0	25950	72
1	$\boxed{7598}$	$\boxed{1568}$

Cutoff = 0.3

- Sensitivity = 0.9560
- Specificity = 0.7735

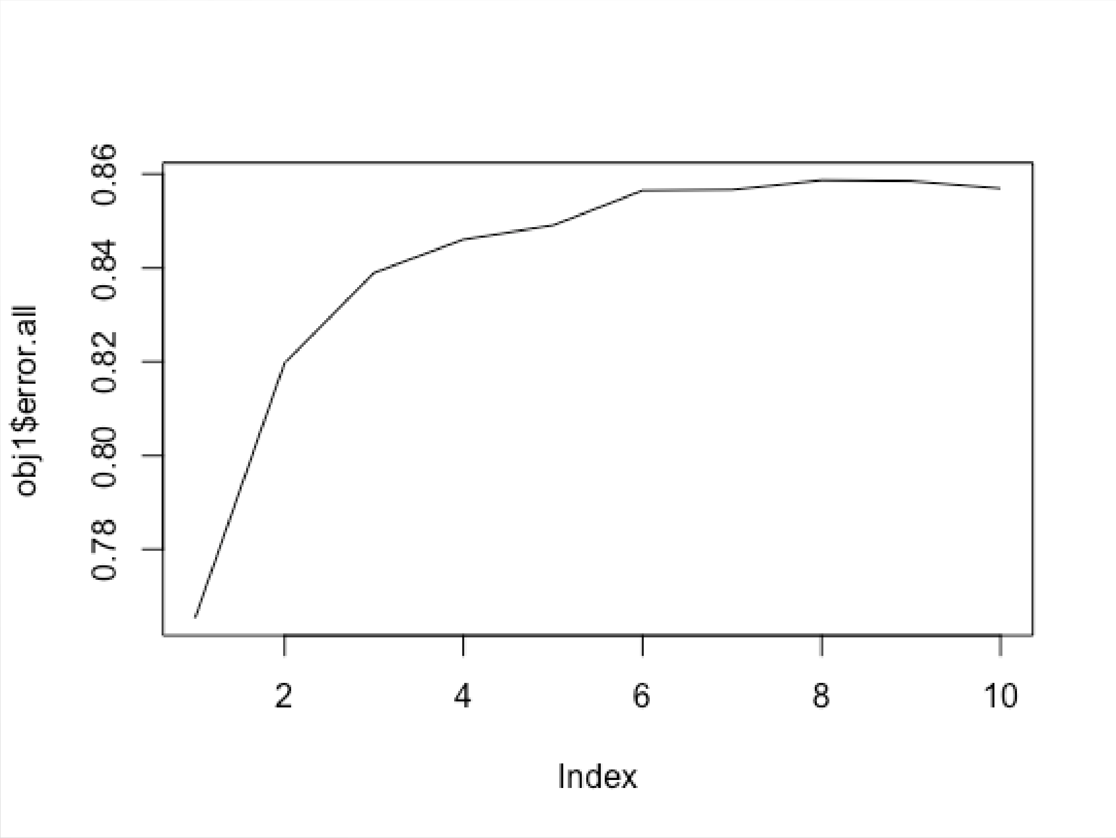
CLASSIFICATION
Logistic Regression

- Error Rate, Sensitivity, Specificity at Cutoff 0.5/0.3

Models		Logistic regression		
Methods		Forward	Backward	Stepwise
cutoff=.5	Valid ER	0.1428896	0.1423781	0.1428896
	Sensitivity	0.8762195	0.8762195	0.8762195
	Specificity	0.8561762	0.8567128	0.8561762
cutoff=.3	Valid ER	0.2190235	0.217972	0.2190235
	Sensitivity	0.9567073	0.9560976	0.9567073
	Specificity	0.7723858	0.7735185	0.7723858

CLASSIFICATION
kNN

- Best K with all variables: K=1
- Prob>0.5



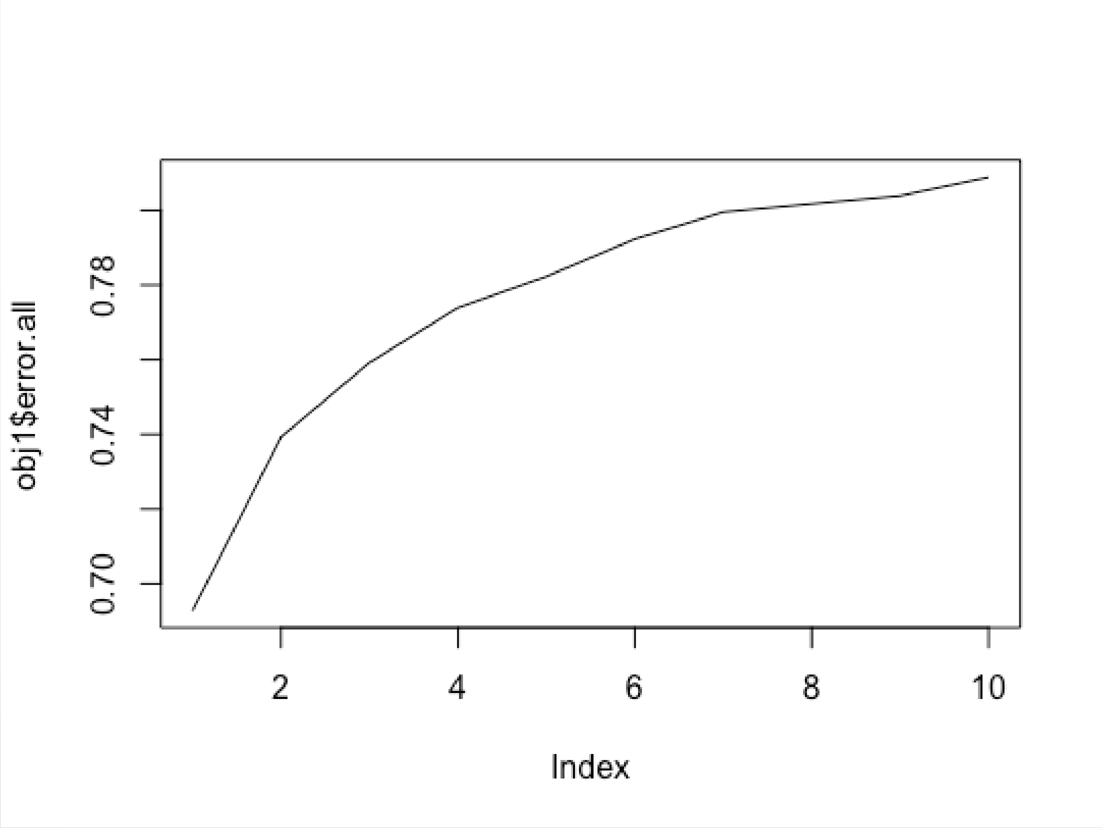
CLASSIFICATION
kNN

- Best K with all variables: K=1
- Prob>0.5
- Sensitivity=0.62

	Y Test	
Y Predict	0	1
0	22738	626
1	10810	1014

CLASSIFICATION
kNN

- Best K with all variables: K=3
- Prob>0.3

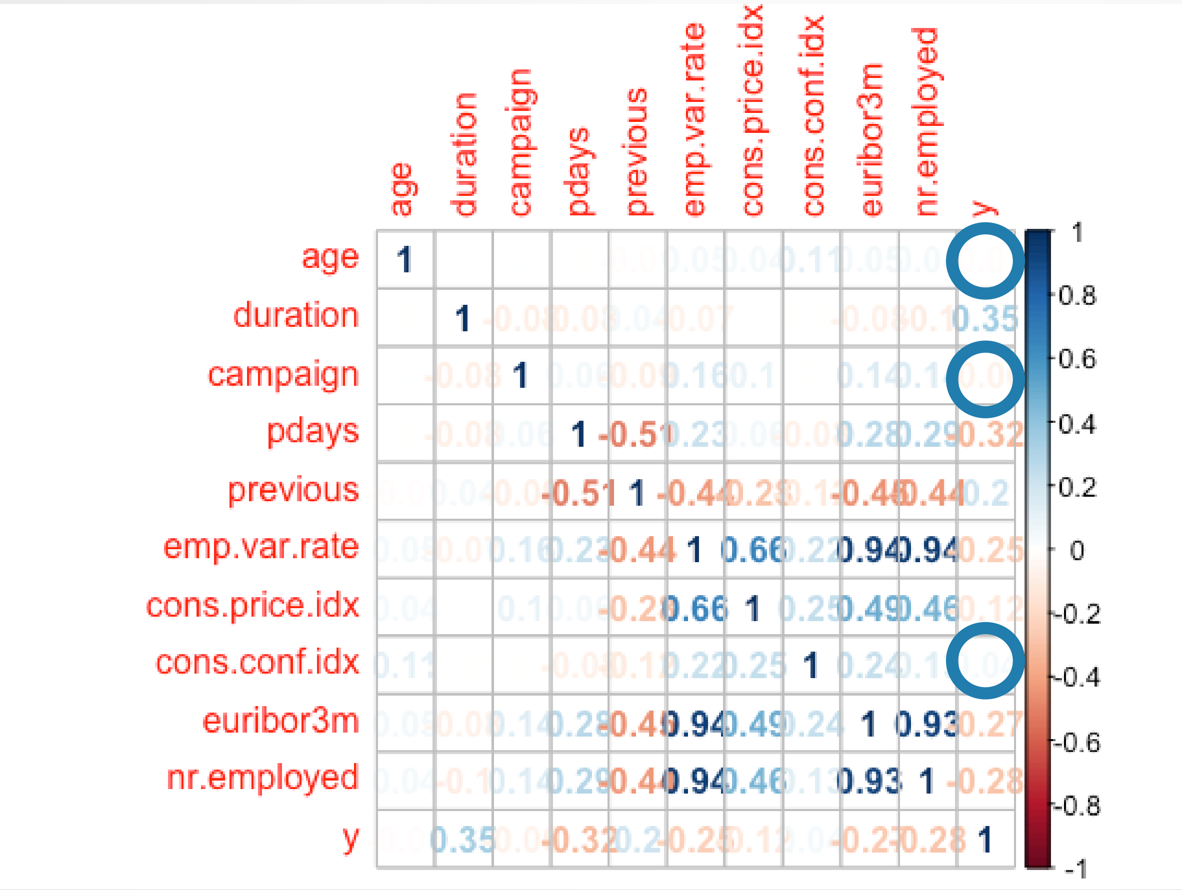


CLASSIFICATION kNN

- Best K with all variables: K=3
- Prob>0.3
- Sensitivity=0.87

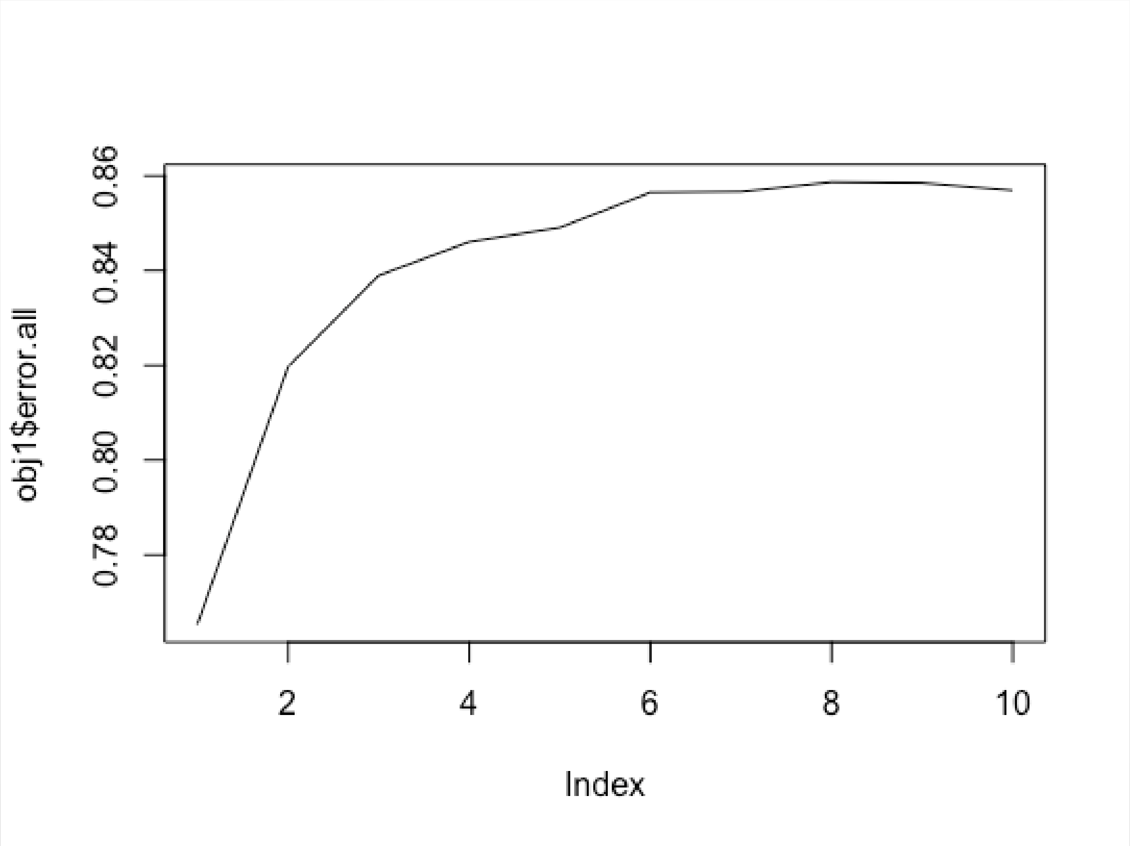
	Y Test	
Y Predict	0	1
0	12129	207
1	21419	1433

CLASSIFICATION
kNN



CLASSIFICATION kNN

- Best K with age, campaign, cons.conf.idx removed: K=1
- Prob>0.5



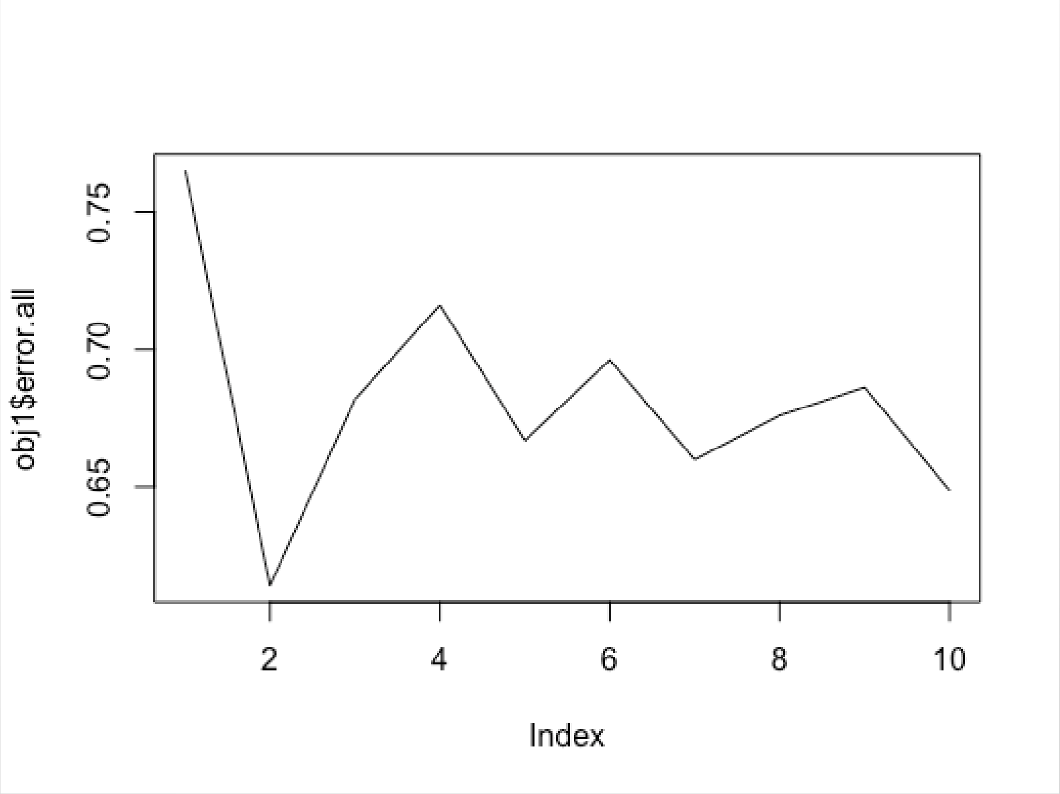
CLASSIFICATION
kNN

- Best K with age, campaign, cons.conf.idx removed: K=1
- Prob>0.5
- Sensitivity=0.70

	Y test	
Y Predict	0	1
0	25291	496
1	8257	1144

CLASSIFICATION
kNN

- Best K with age, campaign, cons.conf.idx removed: K=3
- Prob>0.3



CLASSIFICATION
kNN

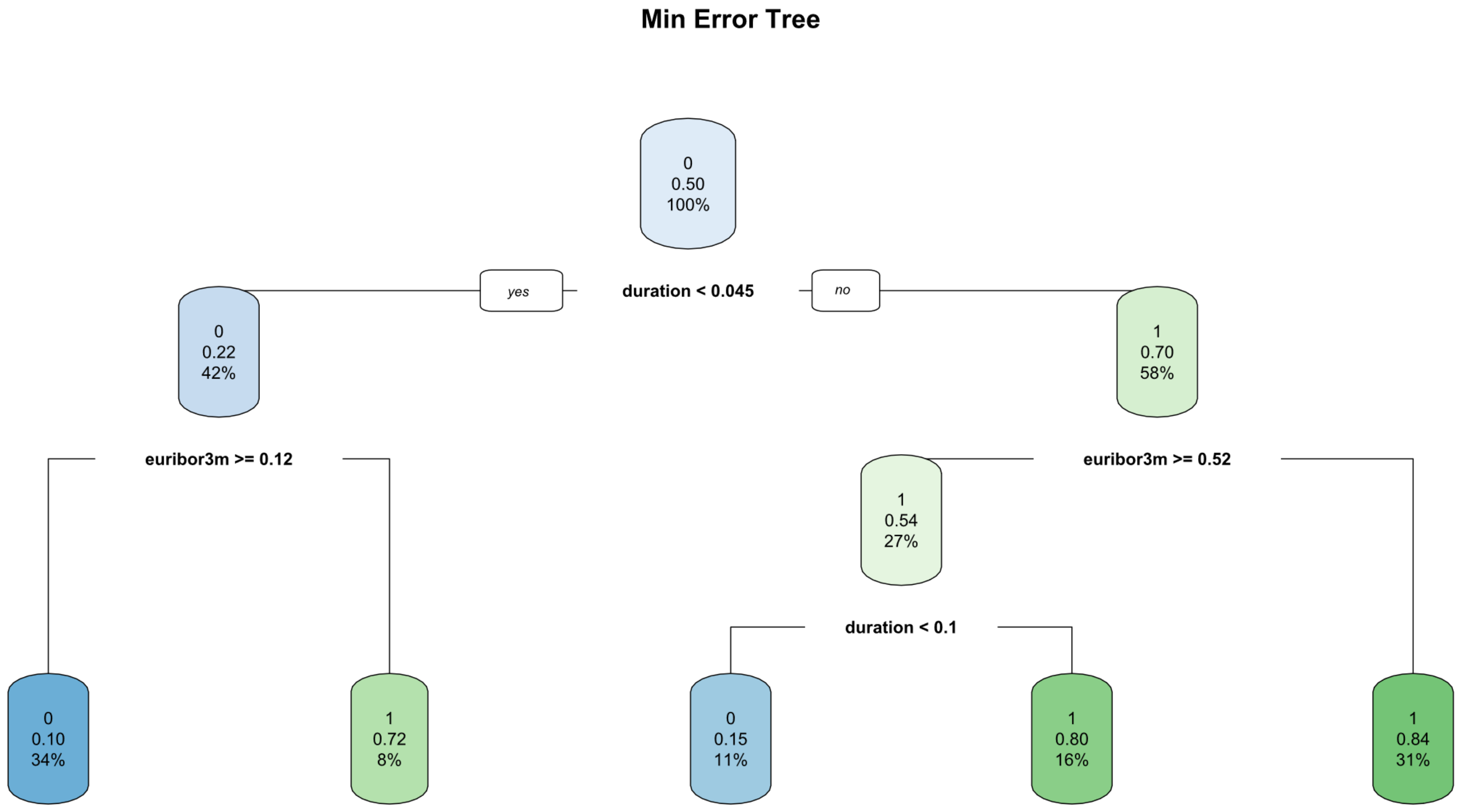
- Best K with age, campaign, cons.conf.idx removed: K=3
- Prob>0.3
- Sensitivity=0.70

	Y test	
Y Predict	0	1
0	25286	495
1	8262	1145

CLASSIFICATION
kNN

Models		W/ all	Removed 3 var
cutoff=.5	Valid ER	0.3249972	0.2487496
	Sensitivity	0.6182927	0.697561
	Specificity	0.6777751	0.753875
cutoff=.3	Valid ER	0.6145845	0.2488632
	Sensitivity	0.8737805	0.6981707
	Specificity	0.3615417	0.753726

CLASSIFICATION
Classification Tree



CLASSIFICATION
Classification Tree

Confusion Matrix

		Actual Class	
Predicted Class		0	1
	0	26889	180
	1	6659	1460

CLASSIFICATION
Classification Tree

- Validation ER = 0.194
- Accuracy = 0.8056
- Specificity = 0.8015
- Sensitivity = 0.8902

CLASSIFICATION
Classification Tree

Cutoff = 0.5 & Cutoff = 0.3

- Specificity = 0.8015
- Sensitivity = 0.8902

		Actual Class	
Predicted Class		0	1
	0	26889	180
	1	6659	1460

Cutoff	Y=1
10.38%	20669
14.91%	6400
71.86%	1664
79.98%	2312
84.42%	4143

CLASSIFICATION
Random Forest

Cutoff = 0.5

- Sensitivity = 0.912
- Specificity =0.841

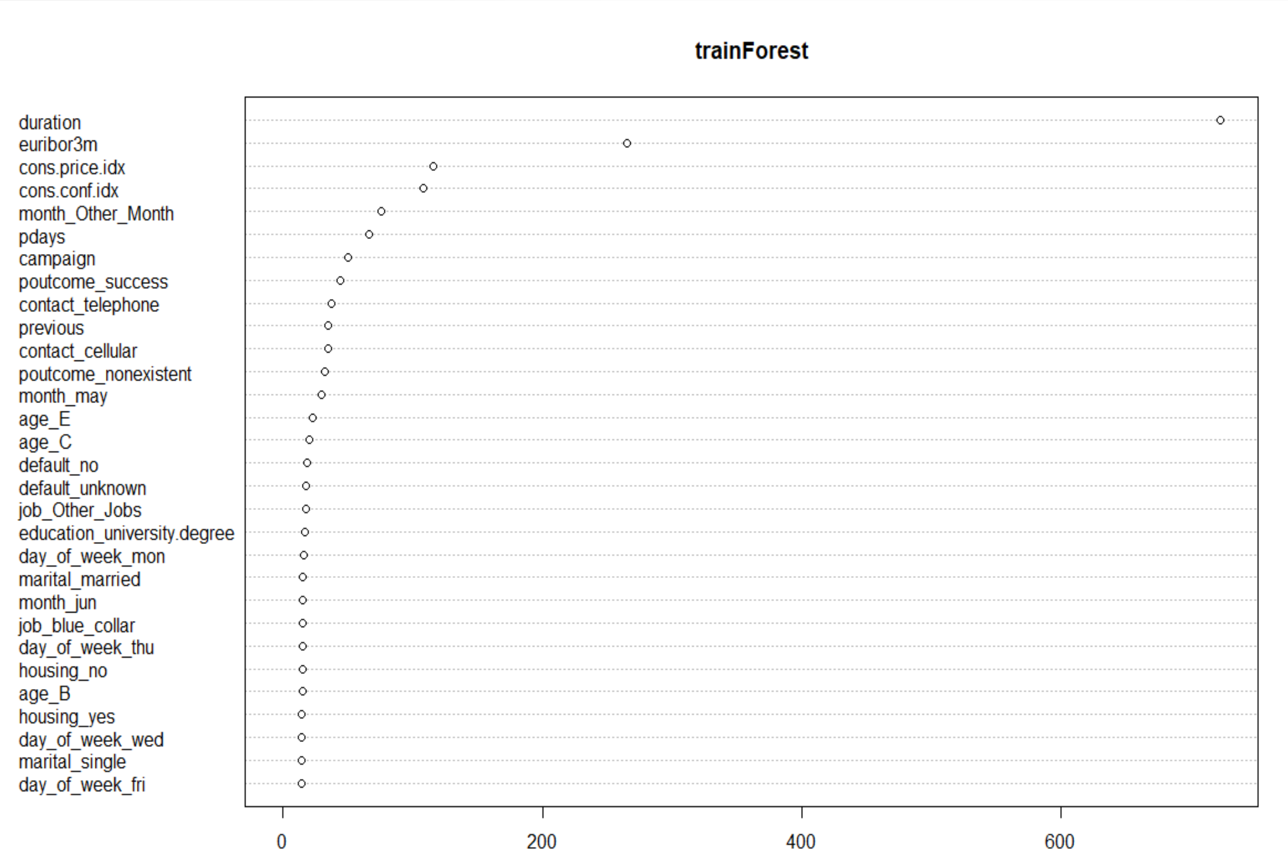
		Actual Class	
Predicted Class		0	1
	0	28201	144
	1	5347	1496

Cutoff = 0.3

- Sensitivity =0.977
- Specificity =0.756

		Actual Class	
Predicted Class		0	1
	0	25379	37
	1	8169	1603

CLASSIFICATION
Random Forest



CLASSIFICATION
XGBoost

Cutoff = 0.5

- Sensitivity = 0.938
- Specificity=0.838

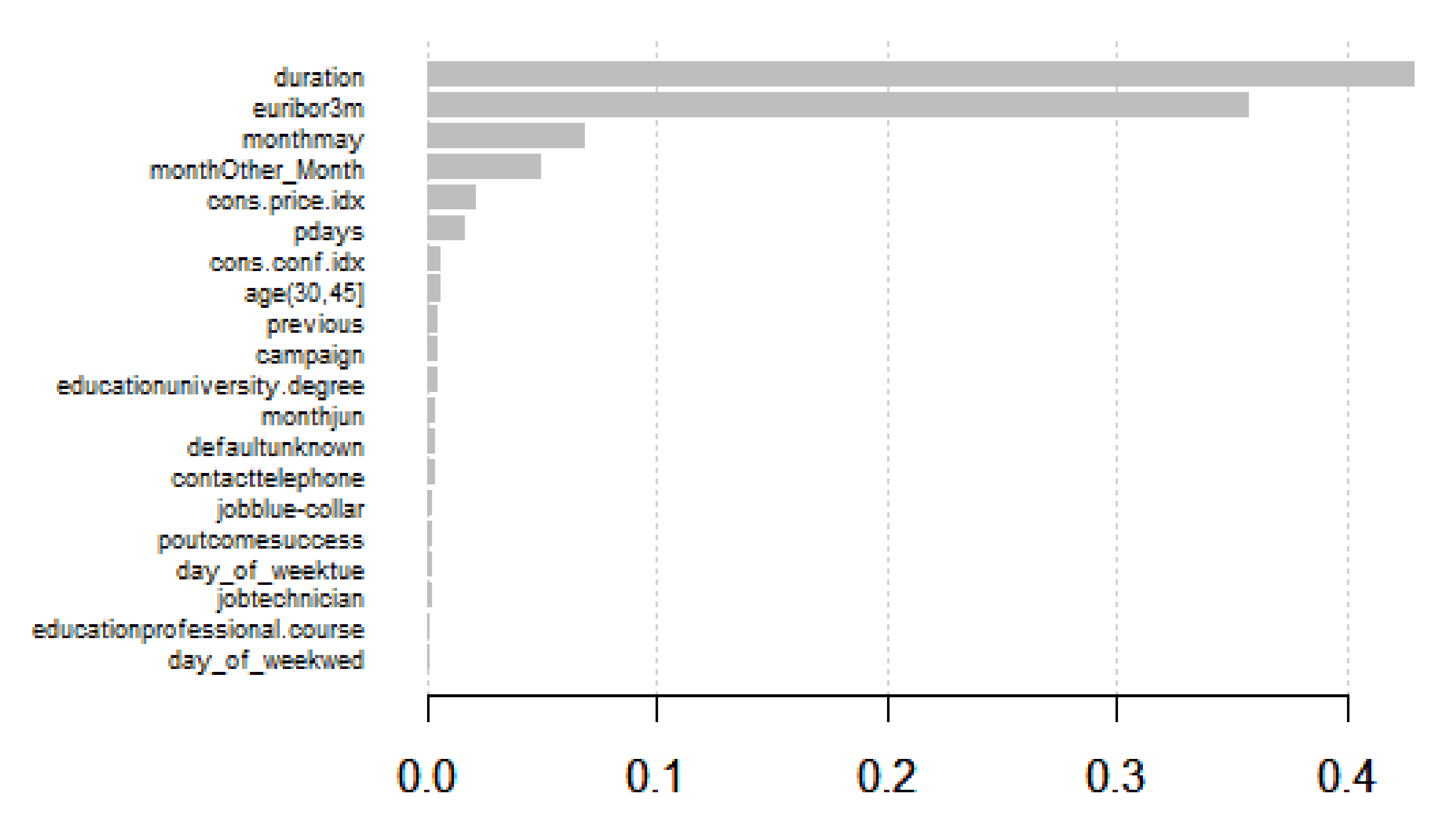
		Actual Class	
Predicted Class		0	1
	0	28100	101
	1	5448	1539

Cutoff = 0.3

- Sensitivity= 0.975
- Specificity = 0.786

		Actual Class	
Predicted Class		0	1
	0	26359	41
	1	7189	1599

CLASSIFICATION
XGBoost



CONCLUSION

Five Data Mining models

- Logistic Regression
- KNN
- Decision trees
- XGBoost
- Random Forest

Three Metrics

- Error Rate
- Sensitivity
- Specificity

RESULTS

Models		Logistic regression			kNN		Classification tree		Random Forest	XGBoost
Methods		Forward	Backward	Stepwise	w/ all	Removed 3	Best pruned	Min Error		
cutoff=.5	Valid ER	0.143	0.142	0.143	0.325	0.249	0.194	0.194	0.156	0.158
	Sensitivity	0.876	0.876	0.876	0.618	0.698	0.890	0.890	0.912	0.938
	Specificity	0.856	0.857	0.856	0.678	0.754	0.802	0.802	0.841	0.838
cutoff=.3	Valid ER	0.219	0.218	0.219	0.615	0.249	0.194	0.194	0.233	0.206
	Sensitivity	0.957	0.956	0.957	0.874	0.698	0.890	0.890	0.977	0.975
	Specificity	0.772	0.774	0.772	0.362	0.754	0.802	0.802	0.756	0.786

A data-driven approach to predict the success of bank telemarketing

This article is analyzing the same dataset than we did in our project

- Four DM models were compared:
  - logistic regression (LR)
  - decision trees (DTs)
  - neural networks (NNs)
  - support vector machines (SVMs)
- These models were compared using two metrics,
  - The best results were obtained by the NN, which resulted in an AUC of 0.80 and ALIFT of 0.67 during the rolling window evaluation.

REFERENCES

- Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, *62*, 22–31. https://doi.org/10.1016/j.dss.2014.03.001.
- Wang, Z., Lai, C., Chen, X., Yang, B., Zhao, S., & Bai, X. (2015). Flood hazard risk assessment model based on random forest. *Journal of Hydrology*, *527*(C), 1130-1141.
- Hearty, J; Sjardin, B; Prateek. Python: Real World Machine Learning by Luca Massaron; Alberto Boschetti; Joshi Published by Packt Publishing, 2016.
- Shmueli.G , Peter C. Bruce, Yahav. I, Nitin R. Patel, and Kenneth, Data mining for business analytics concepts techniques and applications in r answers, page 229-232.

SUCCESS OF BANK TELEMARKETING

A DATA DRIVEN APPROACH

Prepared by: Wone Eui Jung
Yiting Li
Yuchen Bi
Tianqing Huang
Sebastian Marx
Sohrab Gill