MVA FINAL PROJECT PRESENTATION

Denaldo Lapi, Francesco Aristei, Samy Chouiti

Presentation Outline

Context

Exploratory Data Analysis

Algorithms
KNN, SVM, RF and NN

Conclusions

CONTEXT

What makes a client subscribe for a term deposit ?

Our problem

Dataset contents

- Marketing Campaigns of a Portuguese Bank
- 16 features: 8 categorical and 8 numerical
- A binary target variable (yes/no)
- 4521 observations (clients)

Name	Type	Description
Age	Num.	Age of the client
Job	Cat.	Job of the client
Martial Status	Cat.	Marital status of the client
Education	Cat.	Education type of the client
Default	Cat.	If the client has already been in default
Balance	Num.	Balance of the account
Housing	Cat.	Having a housing loan
Loan	Cat.	Having a personal loan
Contact	Cat.	Contacting method
Day	Num.	Last contact day of the week
Month	Num.	Last contact month of the year
Duration	Num.	Duration of last call
Campaign	Num.	Number of contacts (current campaign)
Pdays	Num.	Number of days since last contact (previous campaign)
Previous	Num.	Number of contacts performed (before current campaign)
Poutcome	Cat.	Outcome of the previous marketing campaign
y (Predictor)	**Bin.**	**Did the client subscribed for a term deposit ?**

EXPLORATORY DATA ANALYSIS

Basic Statistics

ID	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
Min. : 1	Min. :19.00	management :969	divorced: 528	primary : 678	no :4445	Min. :-3313	no :1962	no :3830	cellular :2896	Min. : 1.00	may :1398	Min. : 4	Min. : 1.000	Min. : -1.00	Min. : 0.0000	failure: 490	no :4000
1st Qu.:1131	1st Qu.:33.00	blue-collar:946	married :2797	secondary:2306	yes: 76	1st Qu.: 69	yes:2559	yes: 691	telephone: 301	1st Qu.: 9.00	jul : 706	1st Qu.: 104	1st Qu.: 1.000	1st Qu.: -1.00	1st Qu.: 0.0000	other : 197	yes: 521
Median :2261	Median :39.00	technician :768	single :1196	tertiary :1350	NA	Median : 444	NA	NA	unknown :1324	Median :16.00	aug : 633	Median : 185	Median : 2.000	Median : -1.00	Median : 0.0000	success: 129	NA
Mean :2261	Mean :41.17	admin. :478	NA	unknown : 187	NA	Mean : 1423	NA	NA	NA	Mean :15.92	jun : 531	Mean : 264	Mean : 2.794	Mean : 39.77	Mean : 0.5426	unknown:3705	NA
3rd Qu.:3391	3rd Qu.:49.00	services :417	NA	NA	NA	3rd Qu.: 1480	NA	NA	NA	3rd Qu.:21.00	nov : 389	3rd Qu.: 329	3rd Qu.: 3.000	3rd Qu.: -1.00	3rd Qu.: 0.0000	NA	NA
Max. :4521	Max. :87.00	retired :230	NA	NA	NA	Max. :71188	NA	NA	NA	Max. :31.00	apr : 293	Max. :3025	Max. :50.000	Max. :871.00	Max. :25.0000	NA	NA
NA	NA	(Other) :713	NA	NA	NA	NA	NA	NA	NA	NA	(Other): 571	NA	NA	NA	NA	NA	NA

Features Specificity

- A value of -1 in the 'pdays' variable represents clients who have never been contacted before for marketing campaign.

- Most of the values of the variable 'poutcome' are 'unknown', meaning that the bank doesn’t have any data regarding the outcome of previous campaigns for that client.

- No missing values

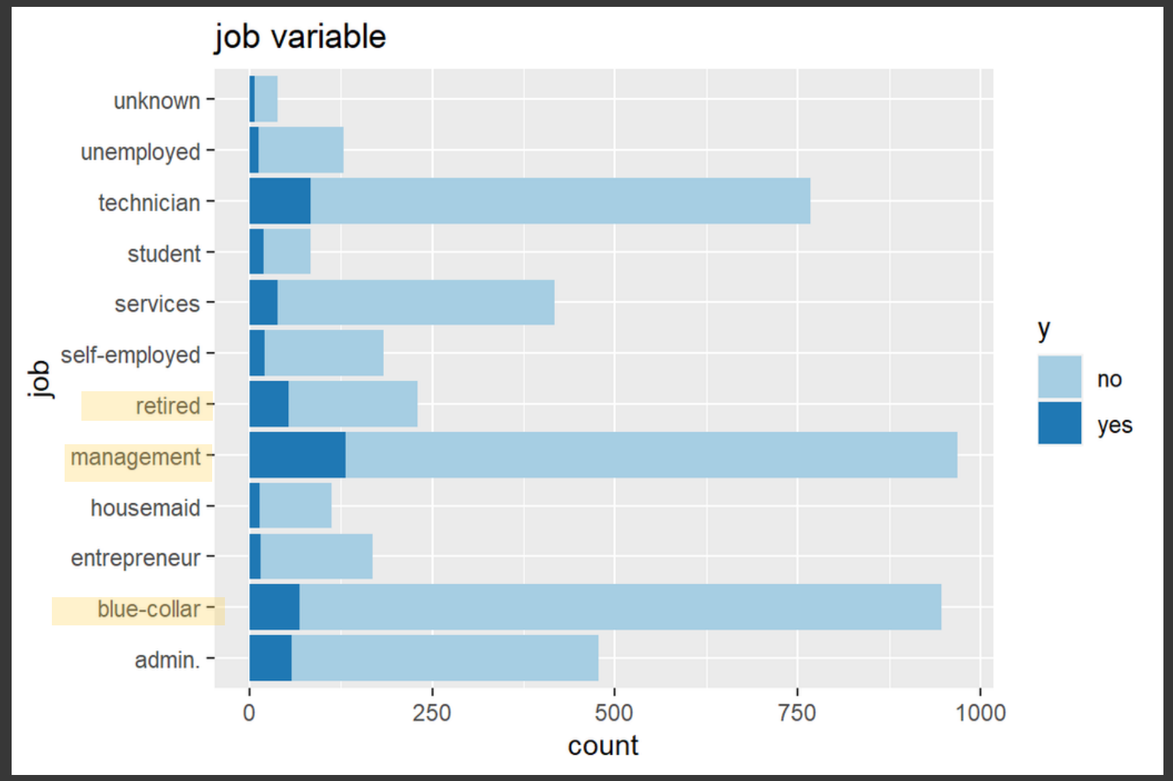
Distributions

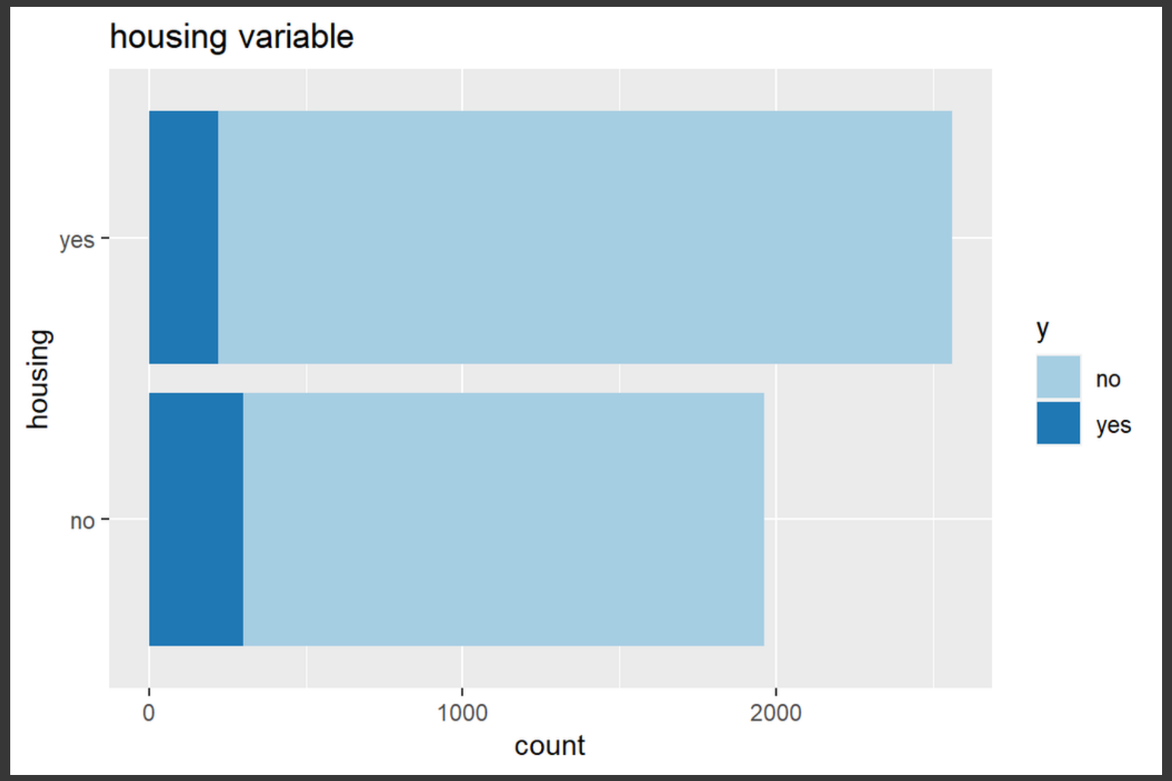
Categorical variables

Numerical variables

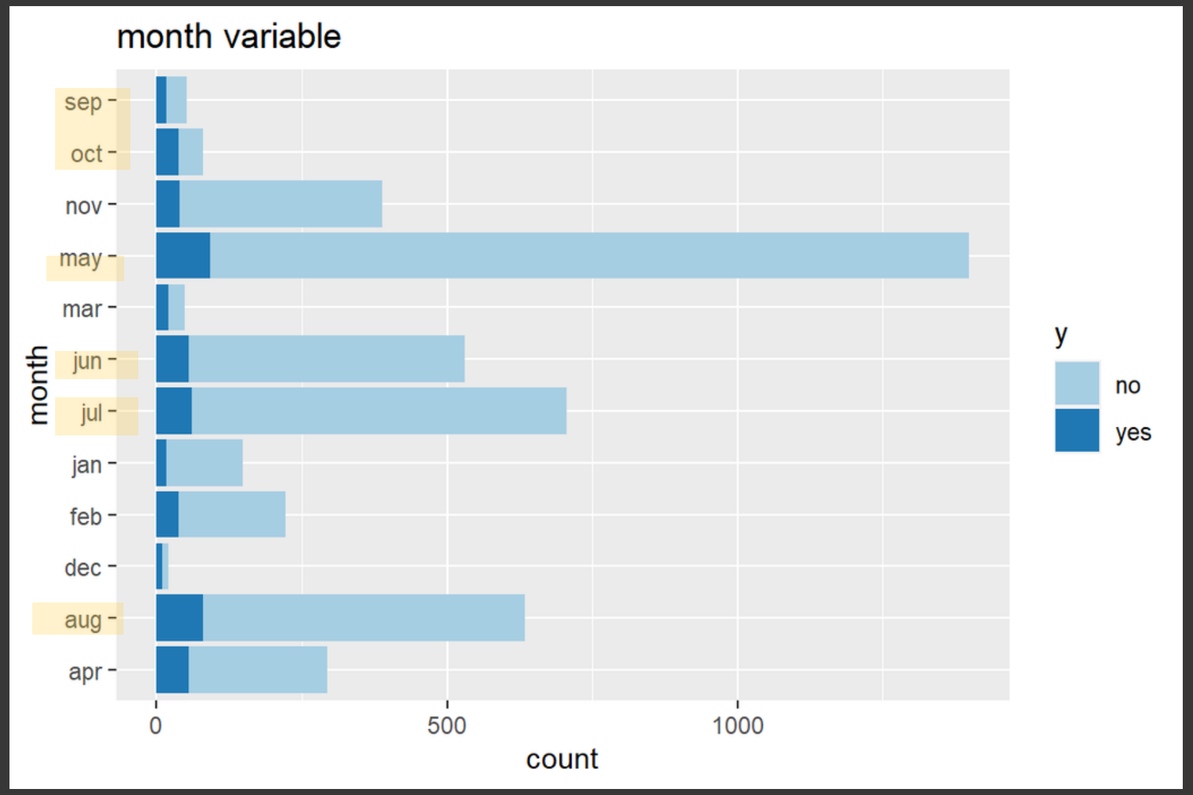
Subscription rates

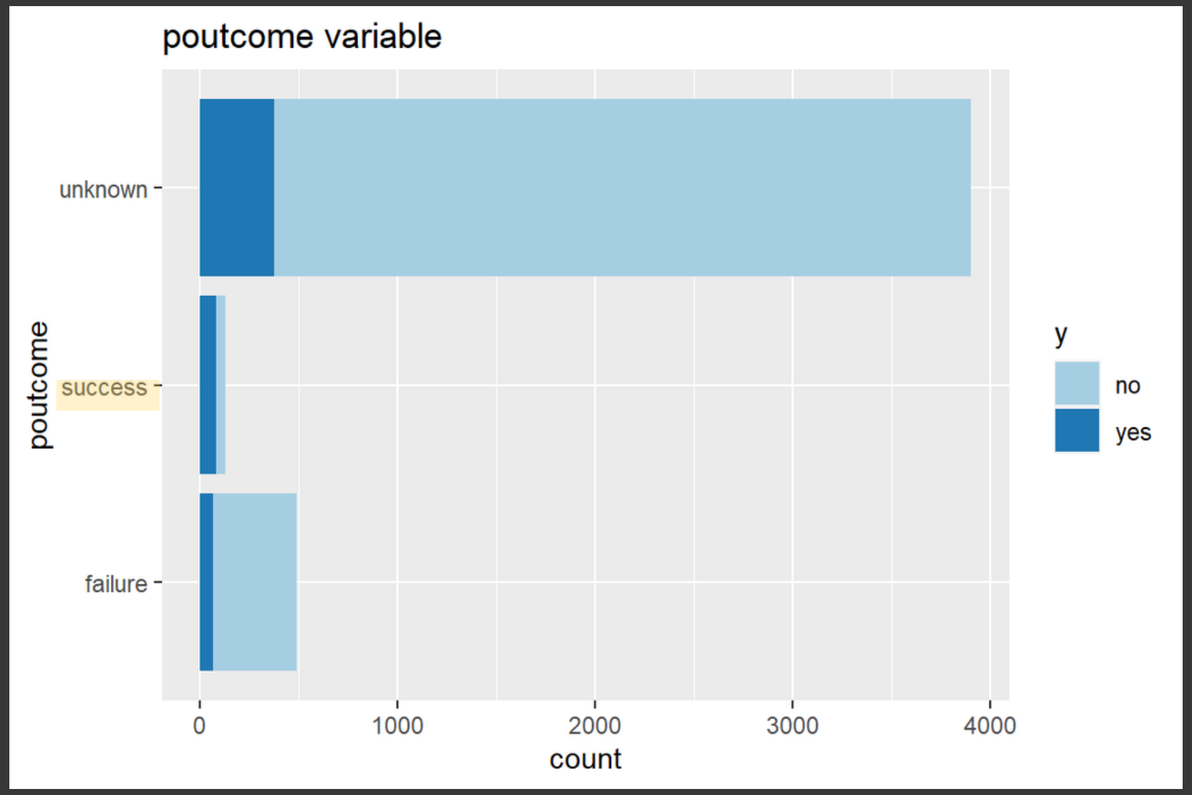
Categorical Variables





Categorical Variables



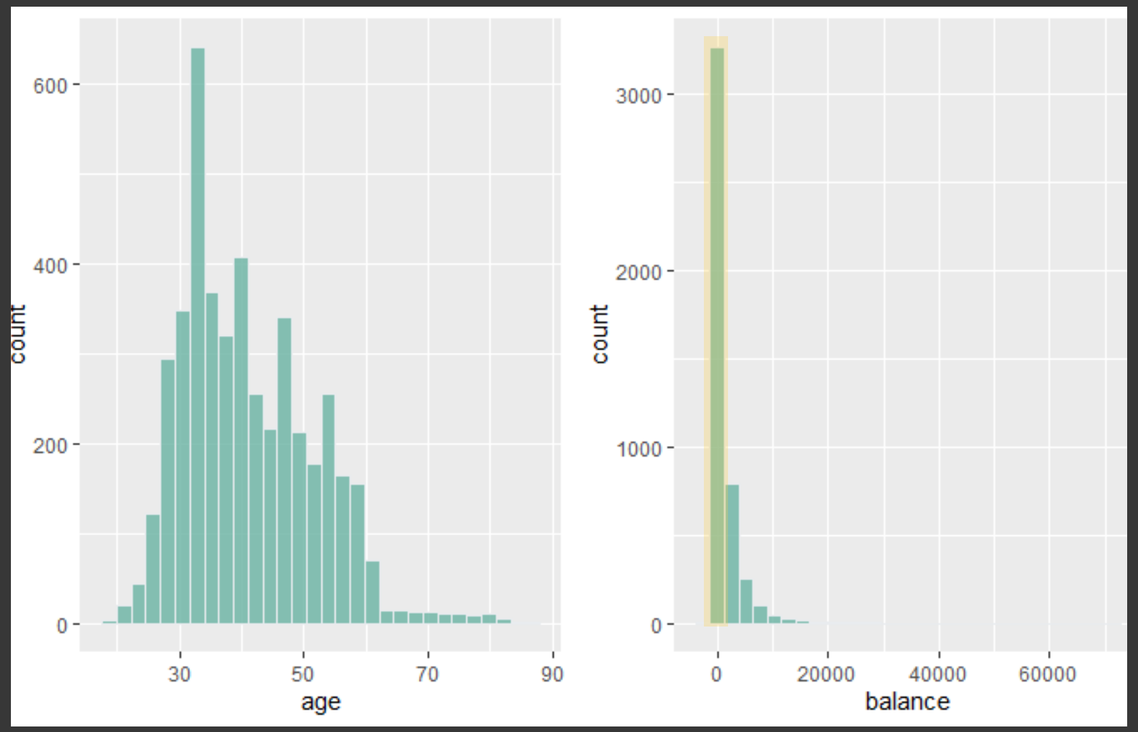


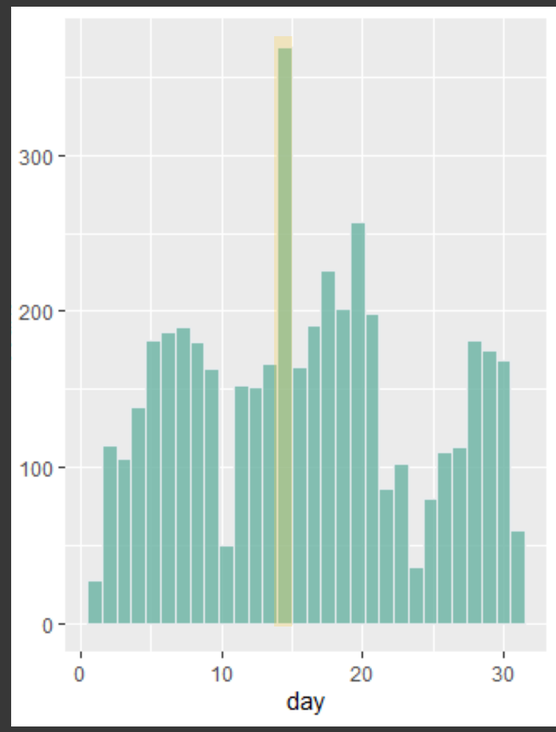
After merging 'other' and 'unknown'

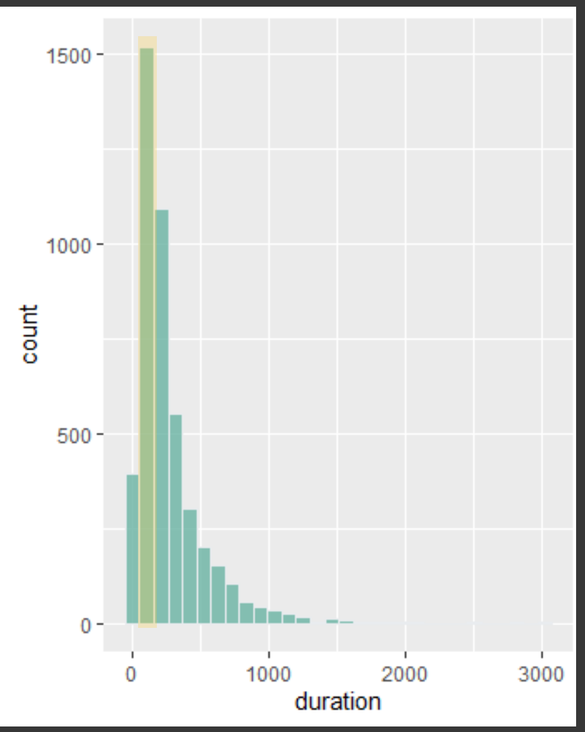
Numerical variables

- Distribution
- Distribution by target
- Box plots
- Correlations

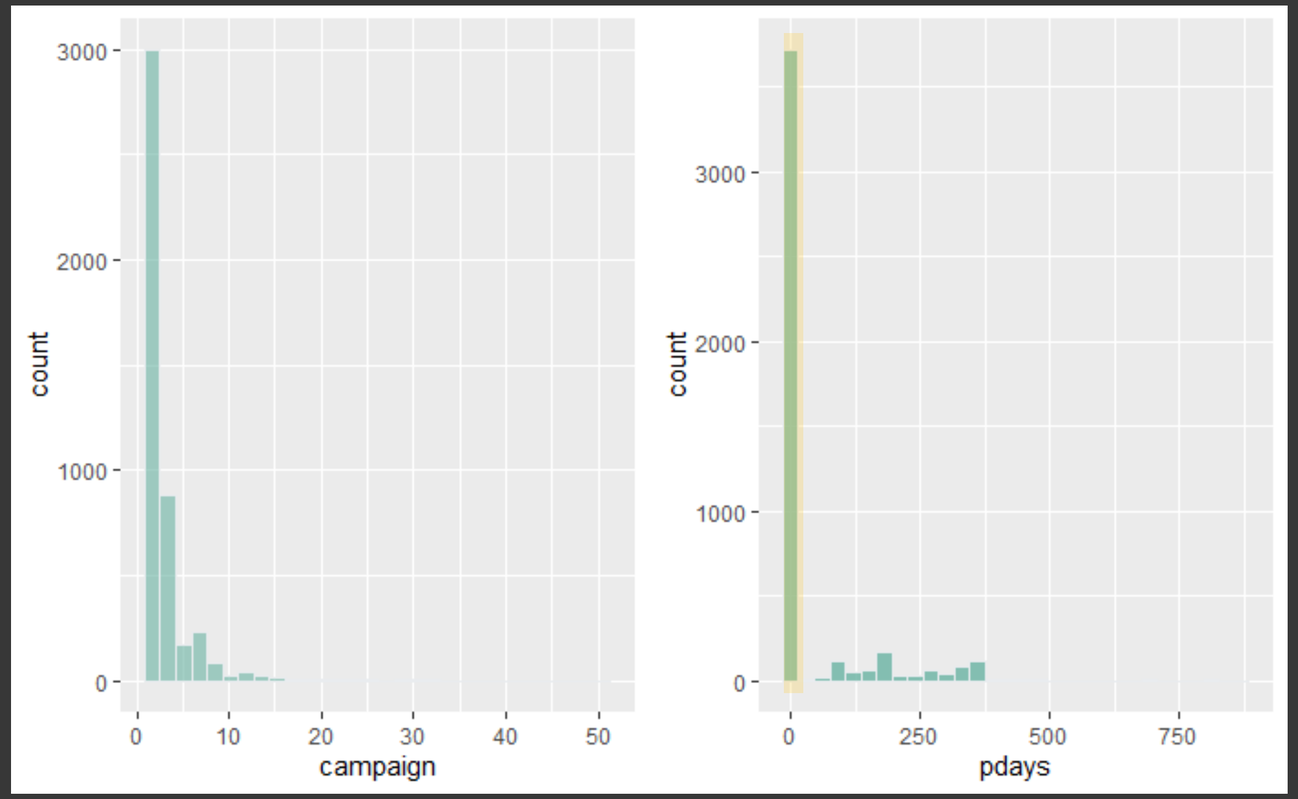
Distribution



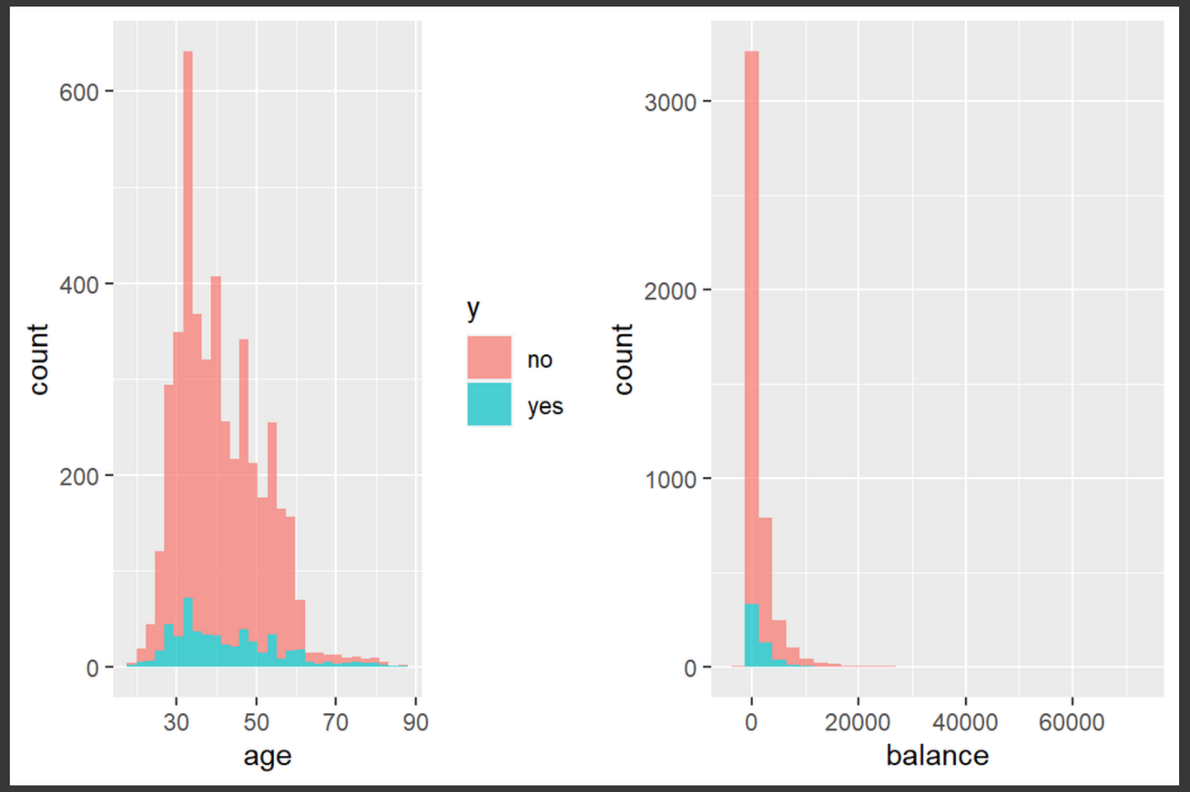


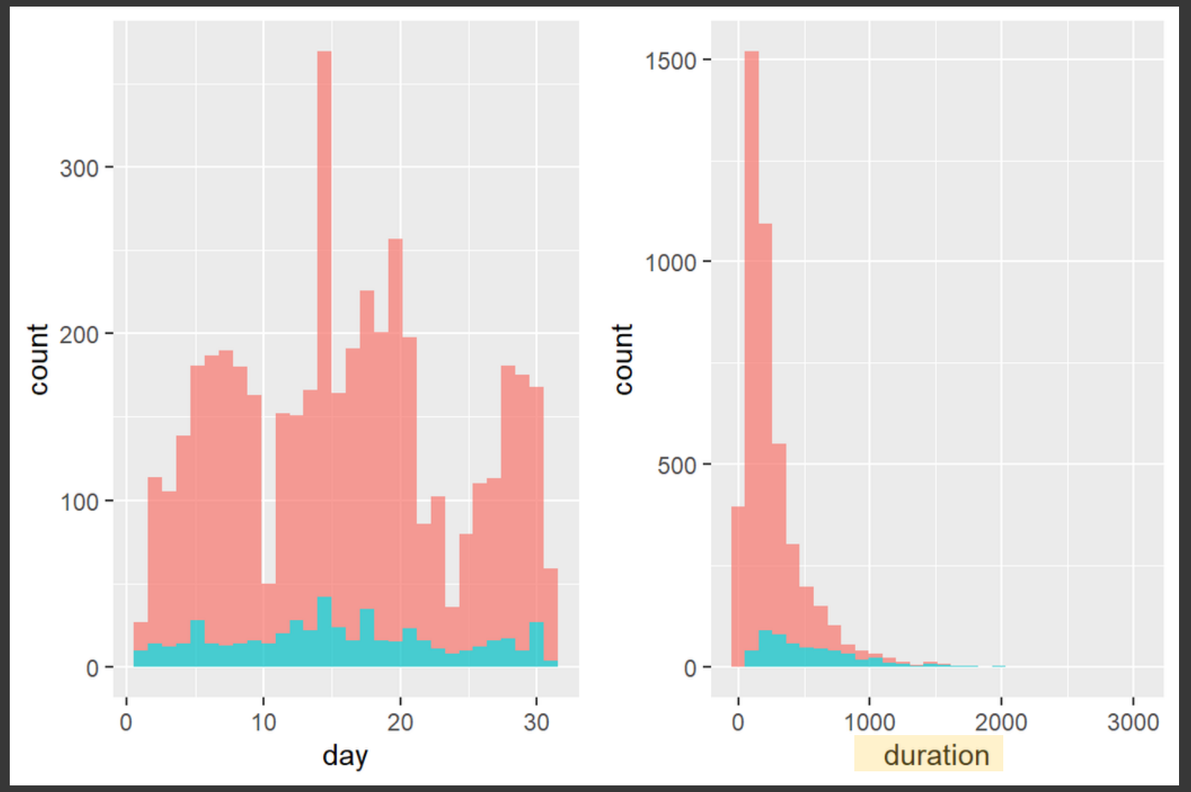


Distribution

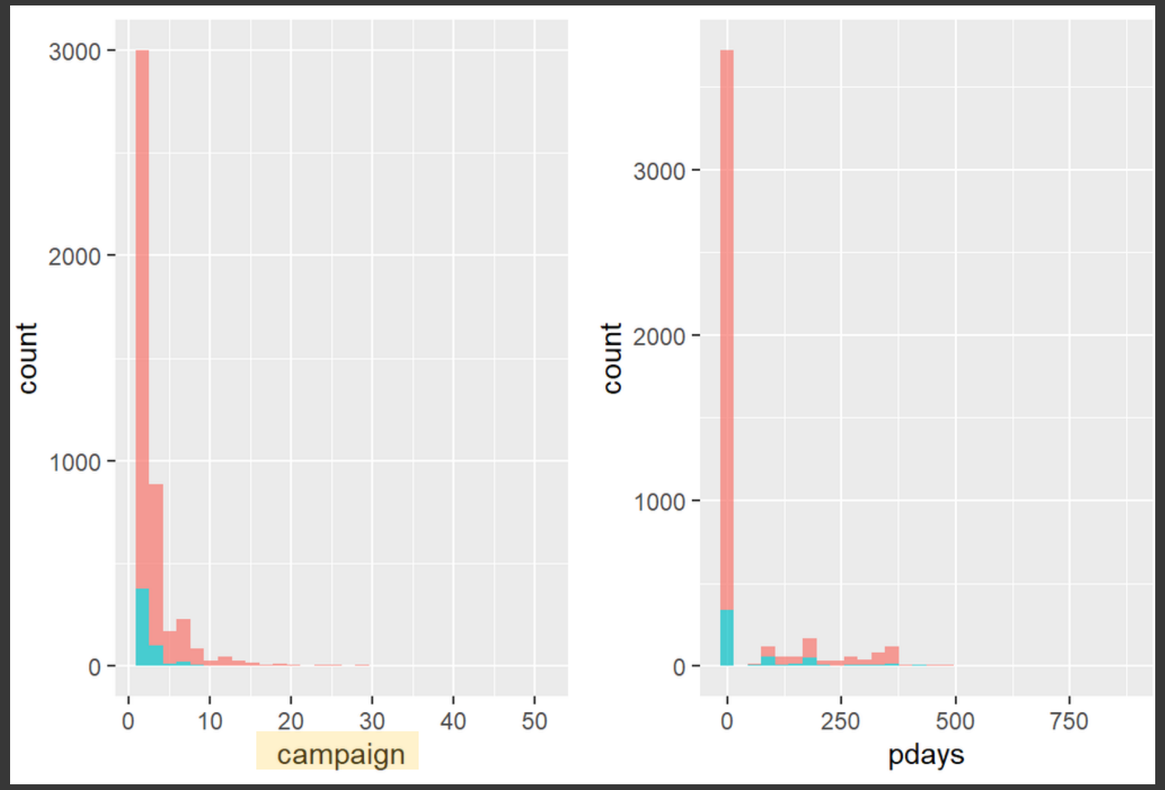


Distribution by target

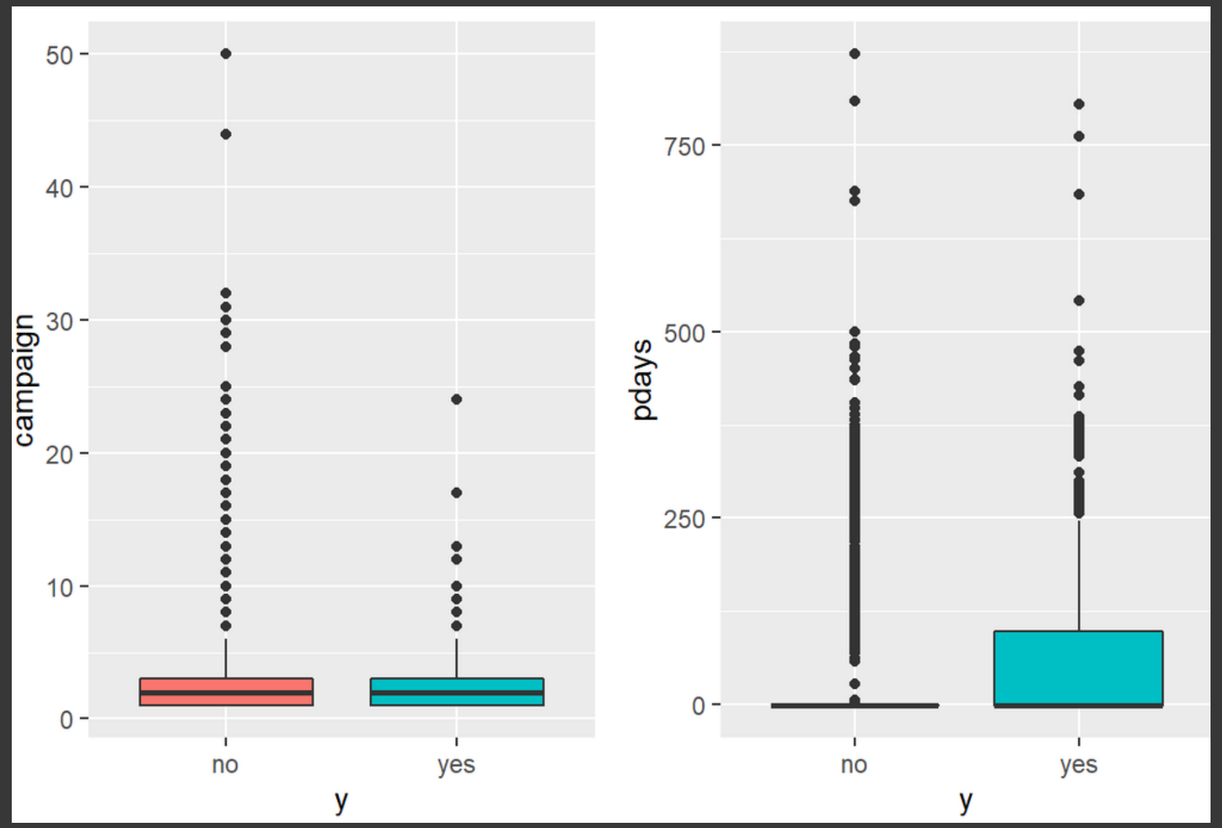


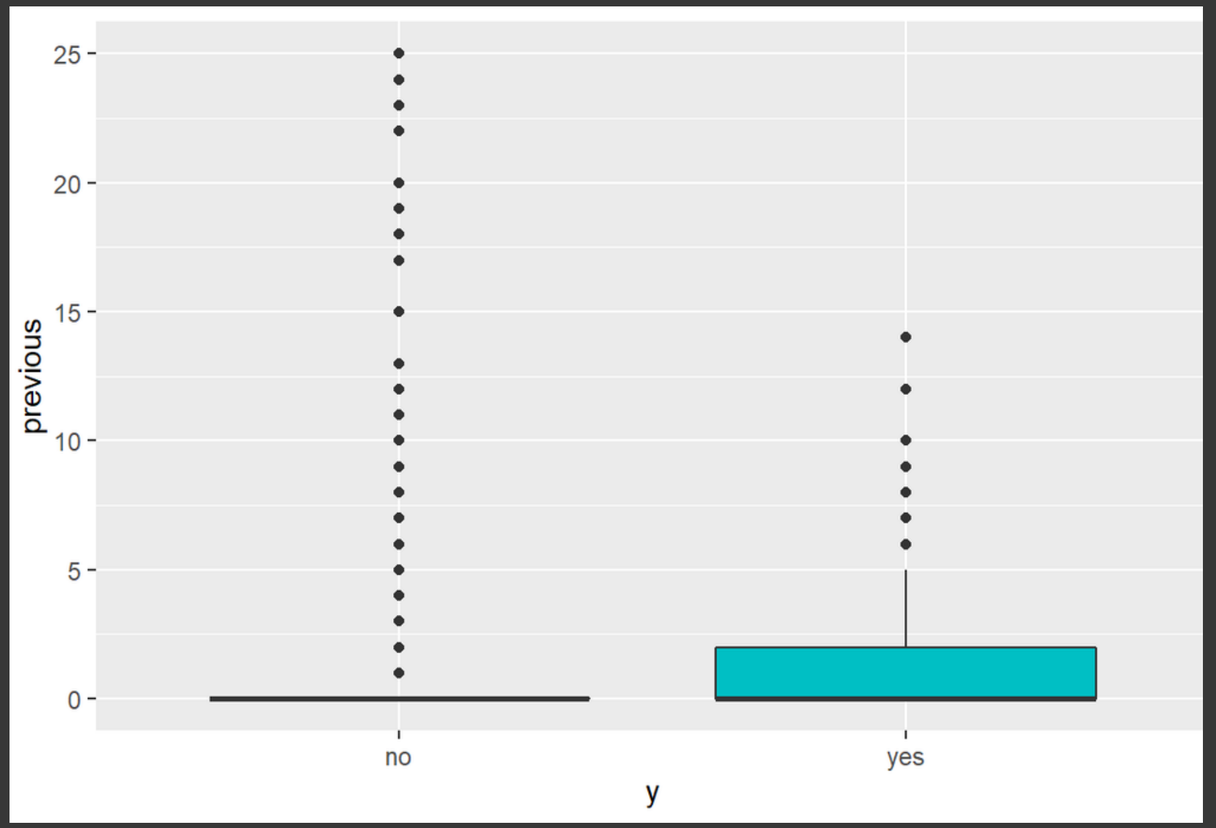


Distribution by target



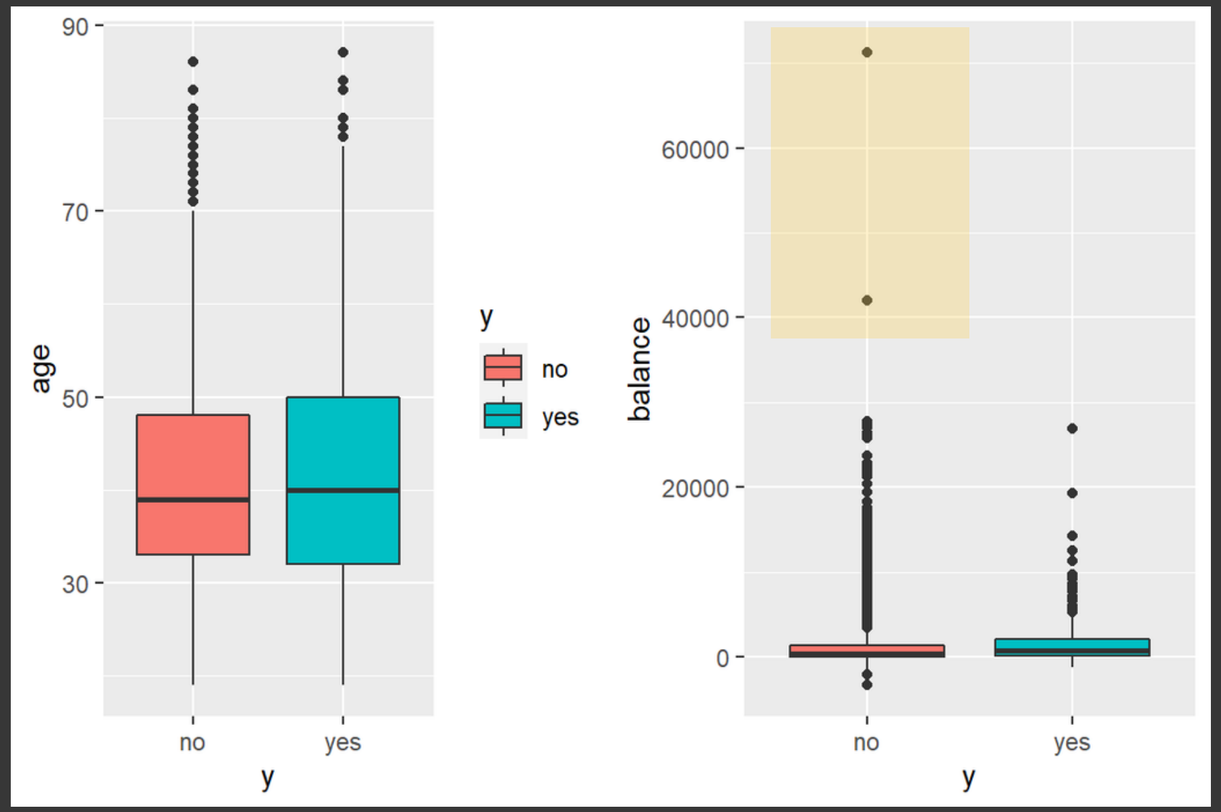
Box Plots

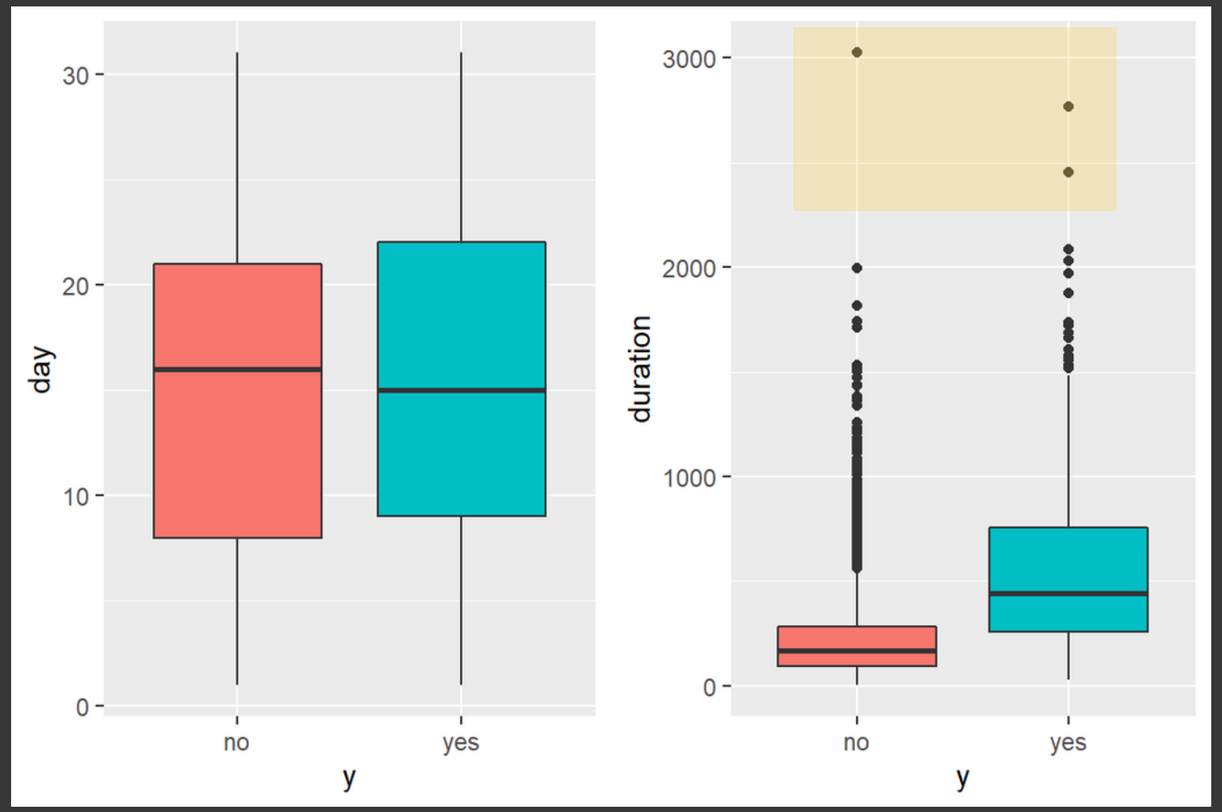




- NOTE:
  - Removed observations with highest 'campaign' (top 2)

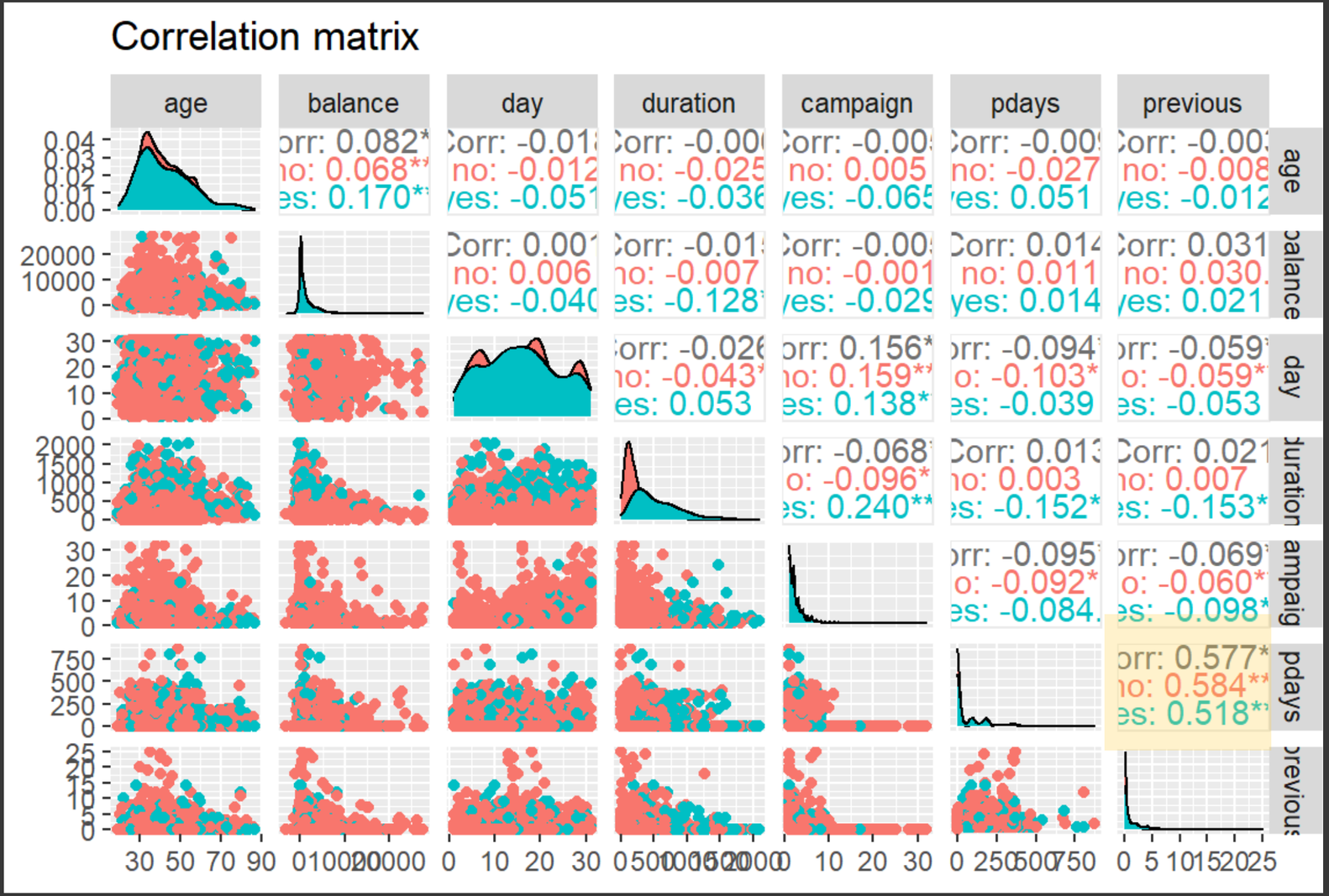
Box Plots





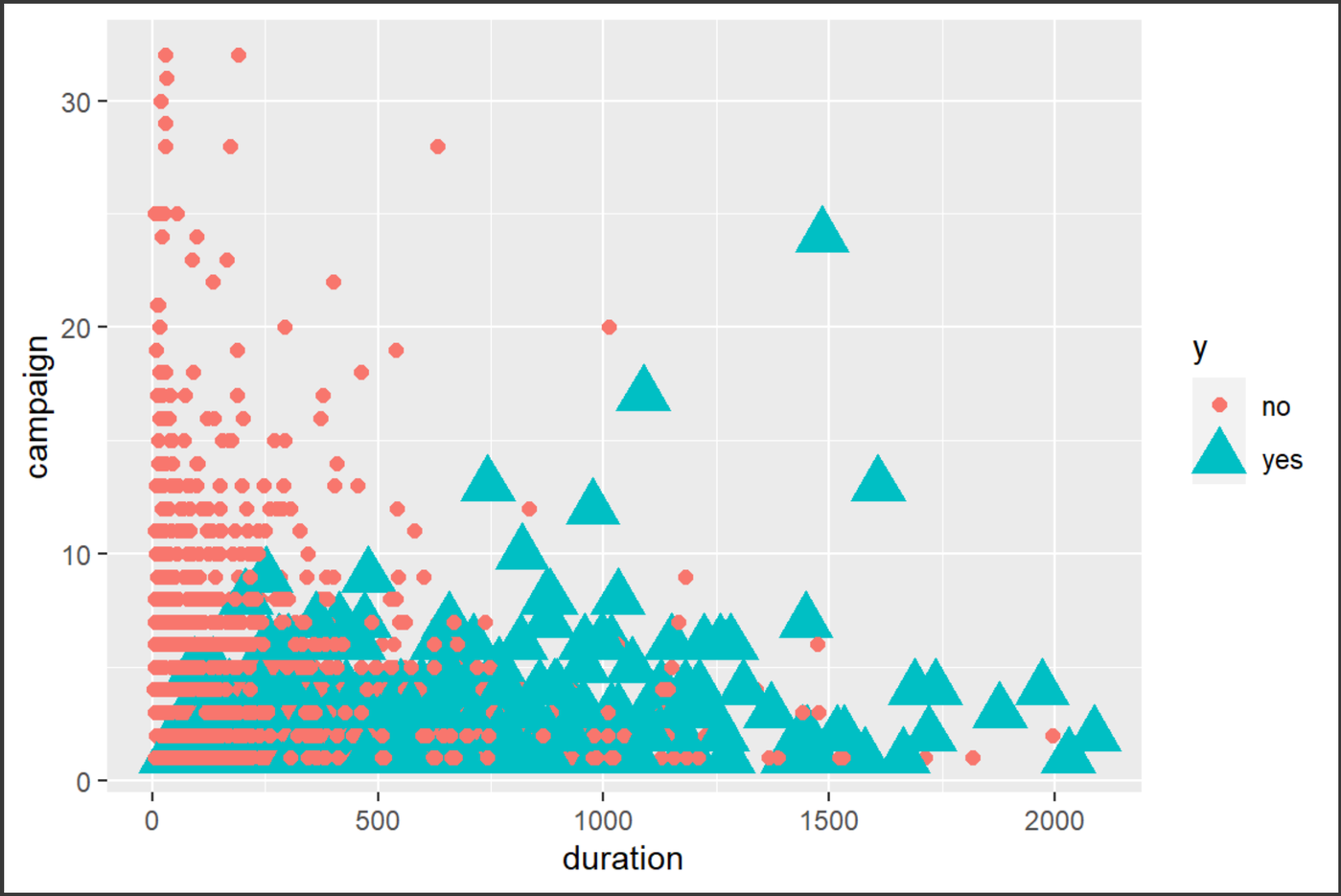
- NOTE:
  - Removed observations with highest 'balance' (top 2) and 'duration' (top 3)

Correlation Matrix



- Very low correlation values
- 'pdays' and 'previous' positively correlated

Scatter plot



- Two clusters of customers:
  - 'Yes' clients: low value of 'campaign' and high 'duration'
  - 'No' clients: higher values on 'campaign' and low 'duration' values

Subscription Rates

Subscription rate by age groups with target

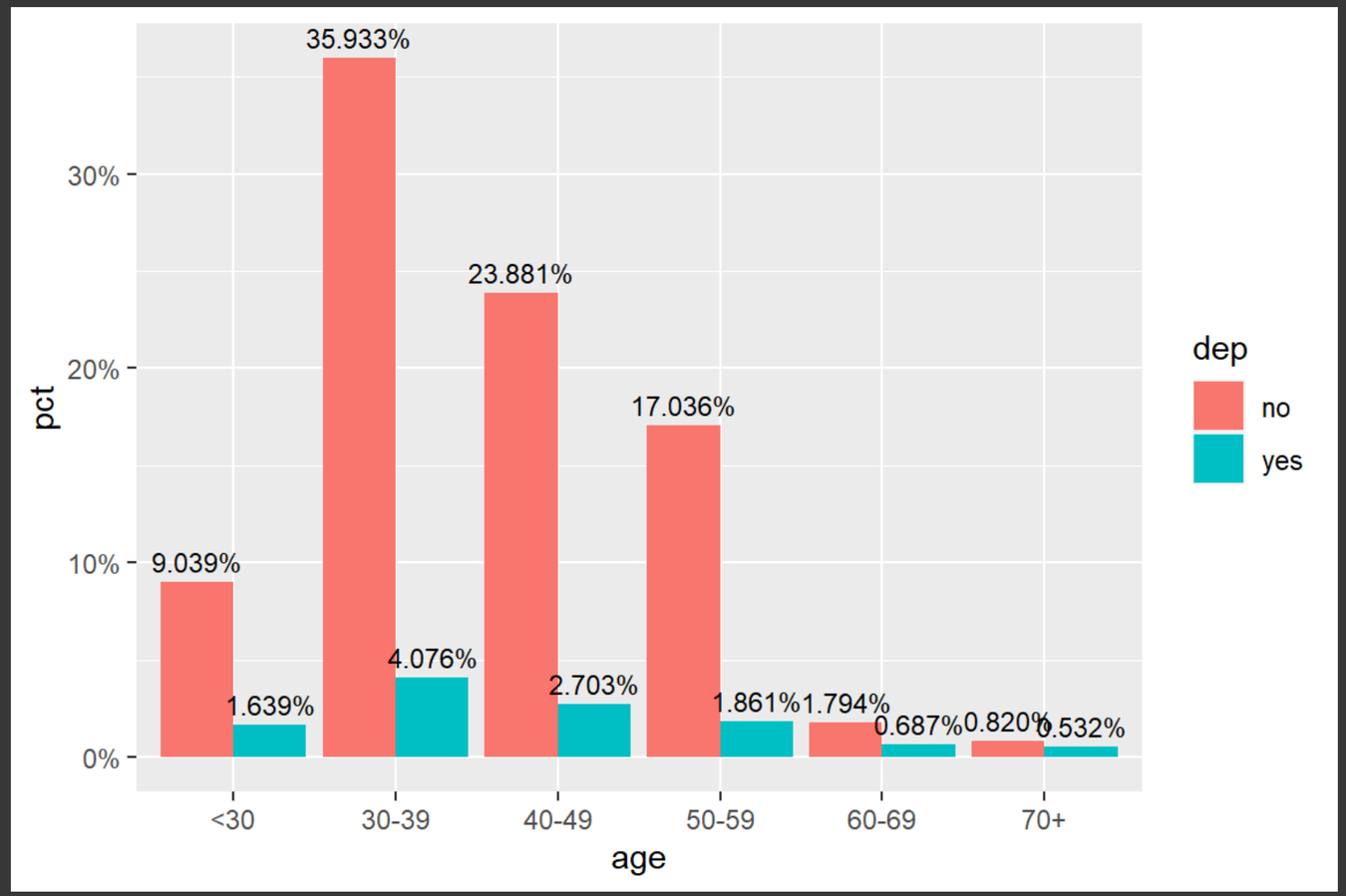
Subscription rate by age groups

Subscription rate by job

Subscription rate by balance

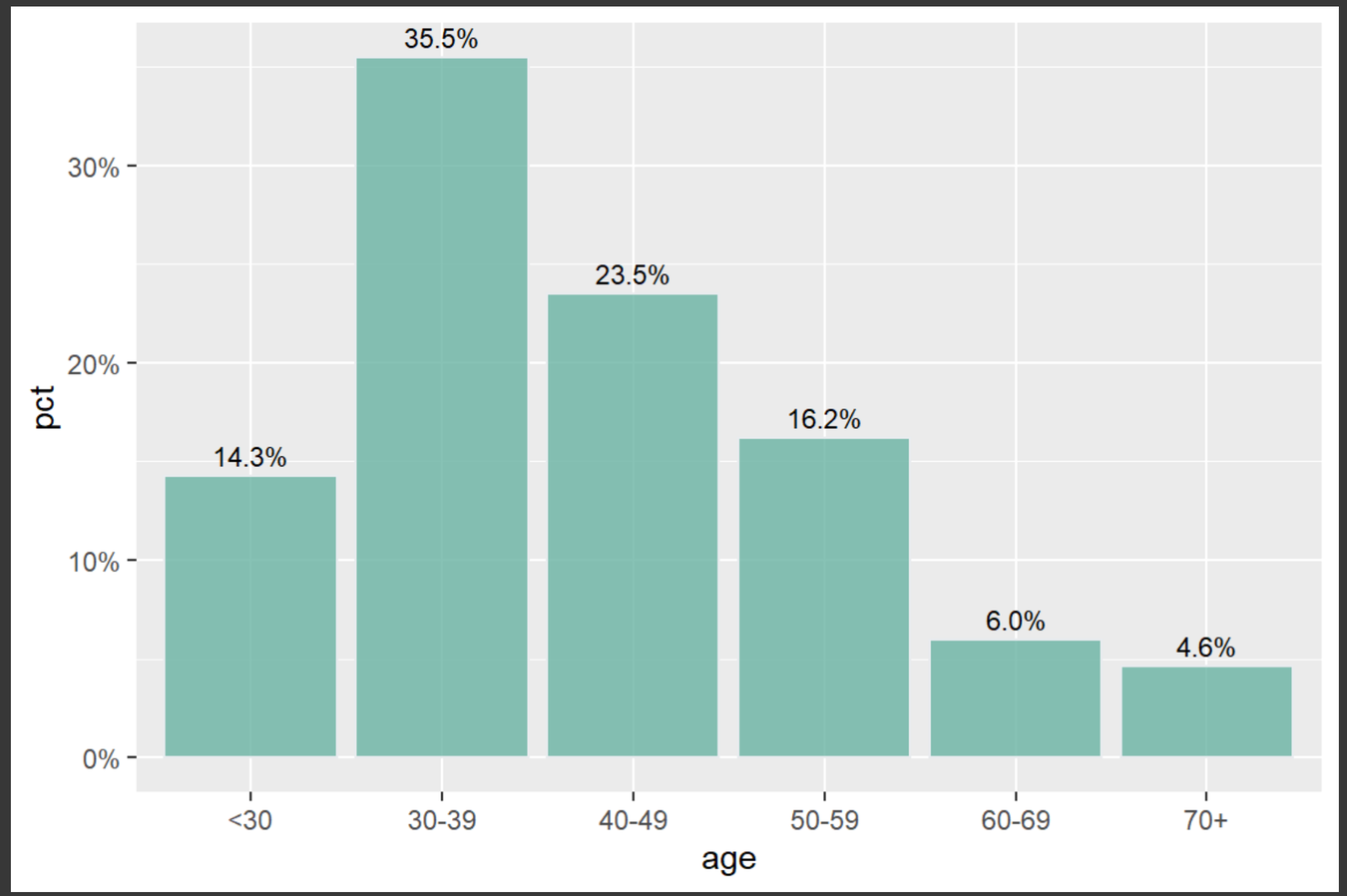
Subscription rate by age and balance

Subscription rate by Age Groups with target



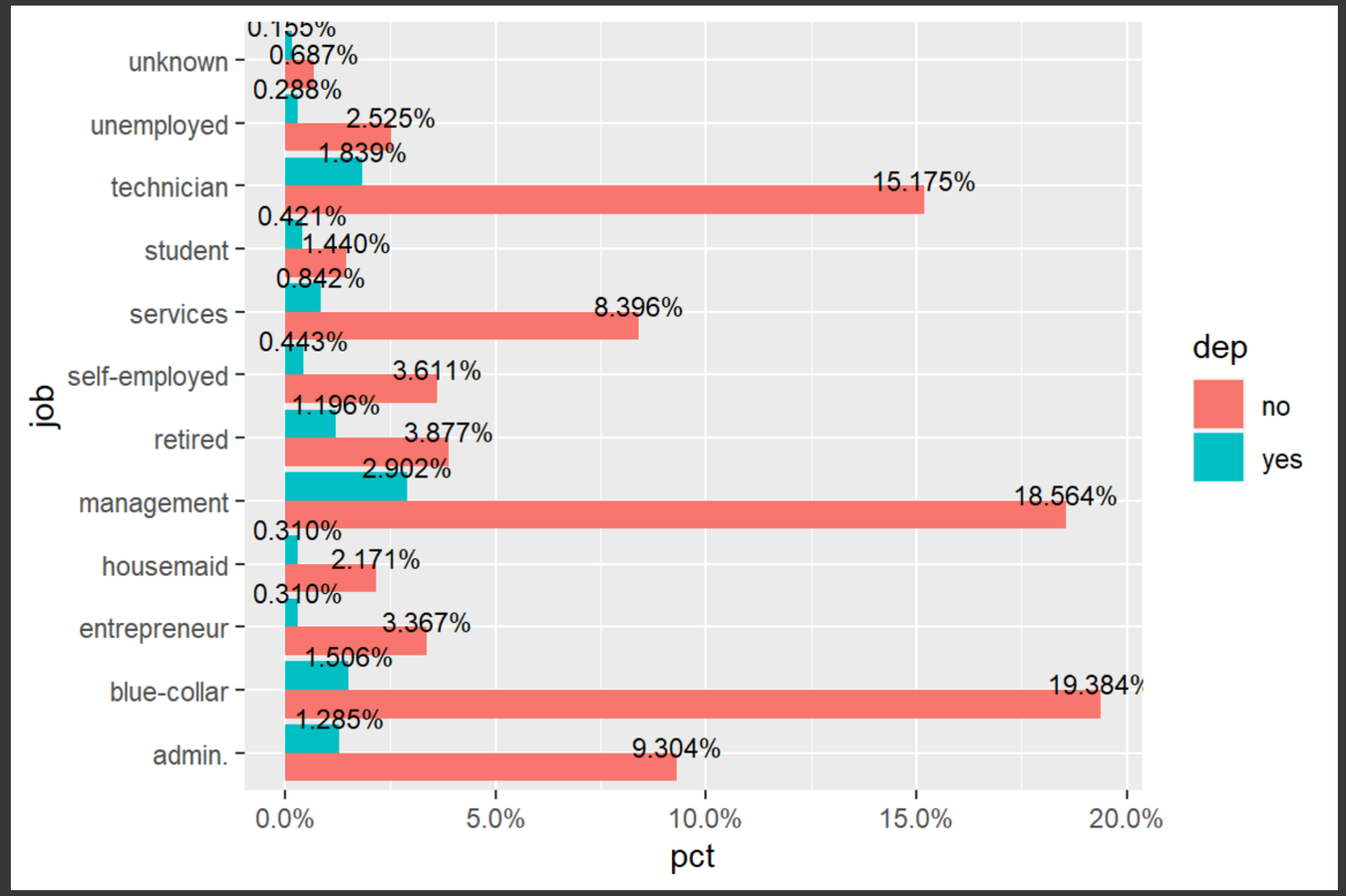
- Clients with an age above 60 have the highest subscription rate
- Young people have high subscription rate

Subscription rate by Age Groups



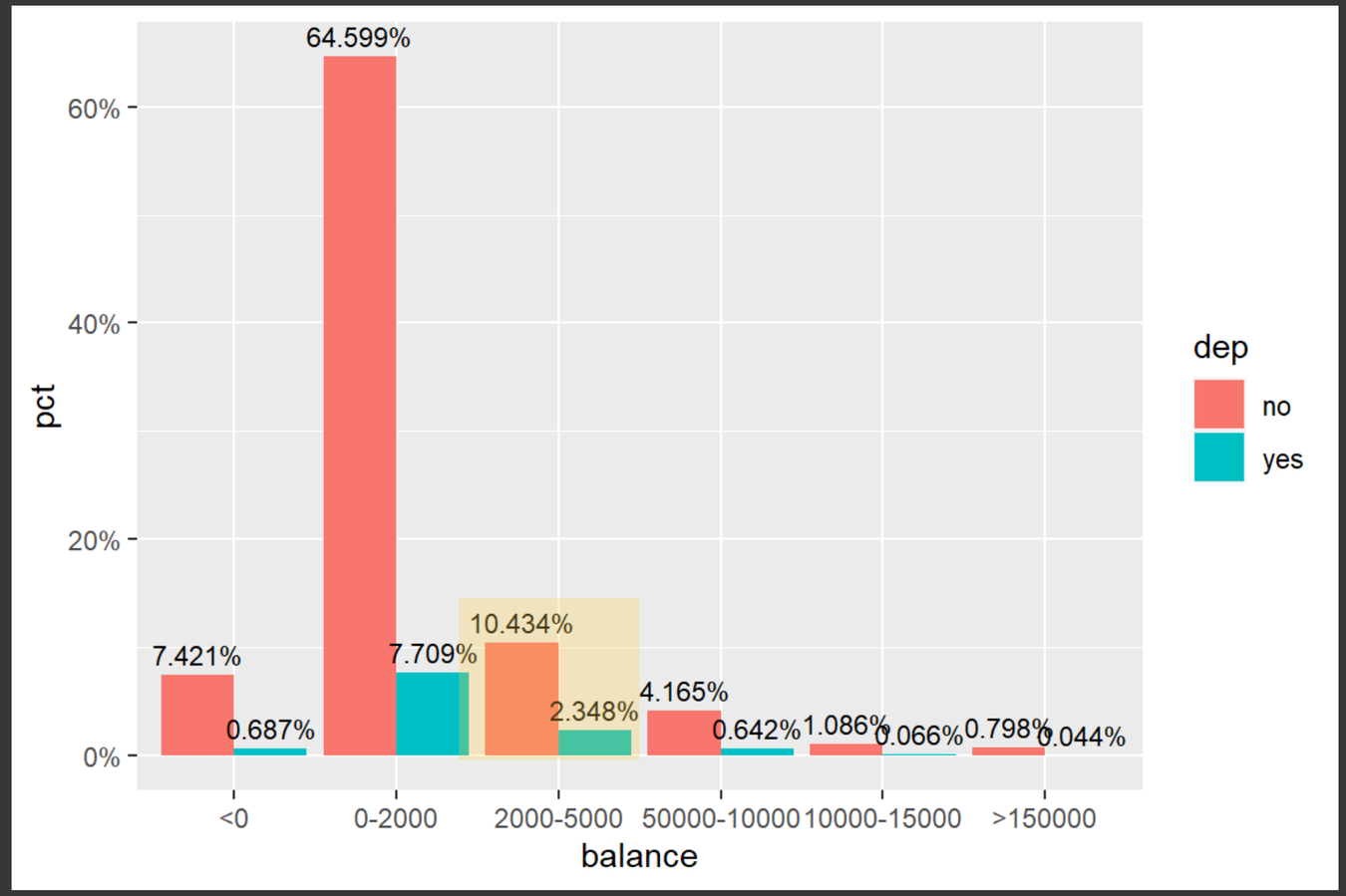
- Age range of contacted people between 30 and 60
  - Explains the high percentage of subscriptions

Subscription rate by Job



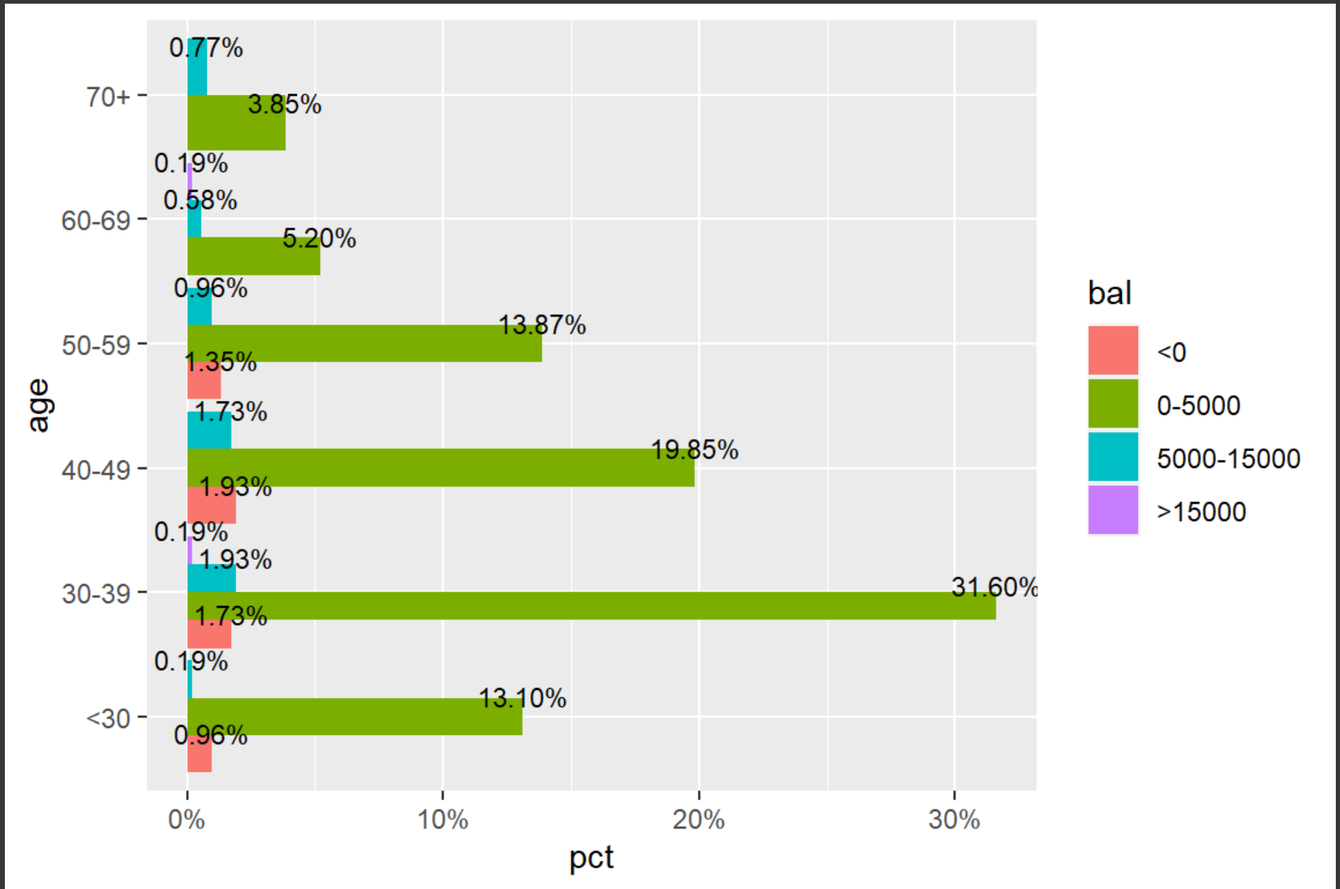
- Percentage of 'yes' very high for 'retired' and 'student' categories
- Confirms previous plot on age

Subscription rate by Balance



- Clients with a balance in the range of 2000 to 5000 are the ones with a highest subscription rate

Subscription rate by Age with Balance



- Highest number of 'yes' in the age range 30-39 for clients with a balance between 0-5000, since they represent the most called people by the bank

Preparing the dataset

- Feature selection
- Feature engineering
- Multivariate outliers on continuous features
- Splitting

Feature Selection

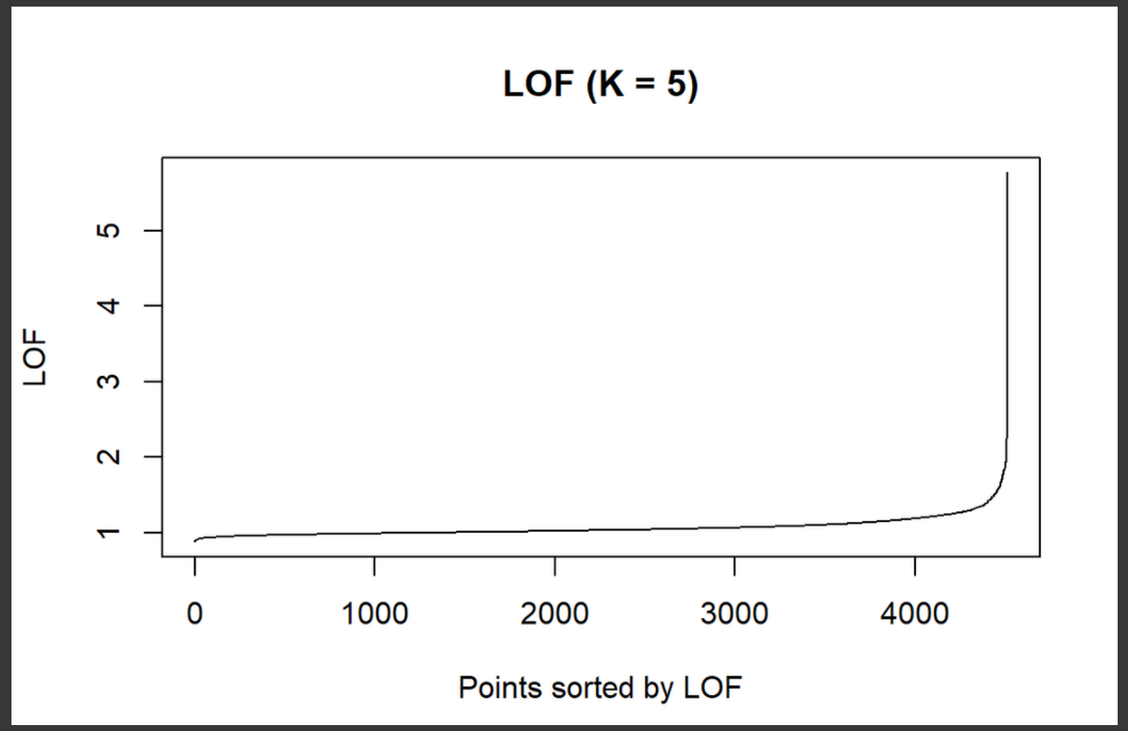
- 'Contact':
  - No interesting relationship spotted
- 'Day':
  - Density pretty uniform along the days of the month

Features Engineering

- 'pdays' :
  - Very spread in the past and has a lot of '-1' values: grouped values into 7 bins

- Convert categorical variables:
  - One-hot encoding through **dummy variables**

Multivariate Outliers



- LOF (Local outlier factor):
  - Non-parametric
  - Density-based local outliers
  - Outlier score for each sample
  - Outliers start around a LOF value of 2.0

Splitting

	”no”	”yes”
Training	3185	412
Test	797	104

- Train:
  - 80 % of the original data will be used for training
- Test:
  - 20% will be used to evaluate the final performances
- **Stratified Splitting**:
  - Keep the same 'yes'/'no' partitions

Reducing the dataset

PCA

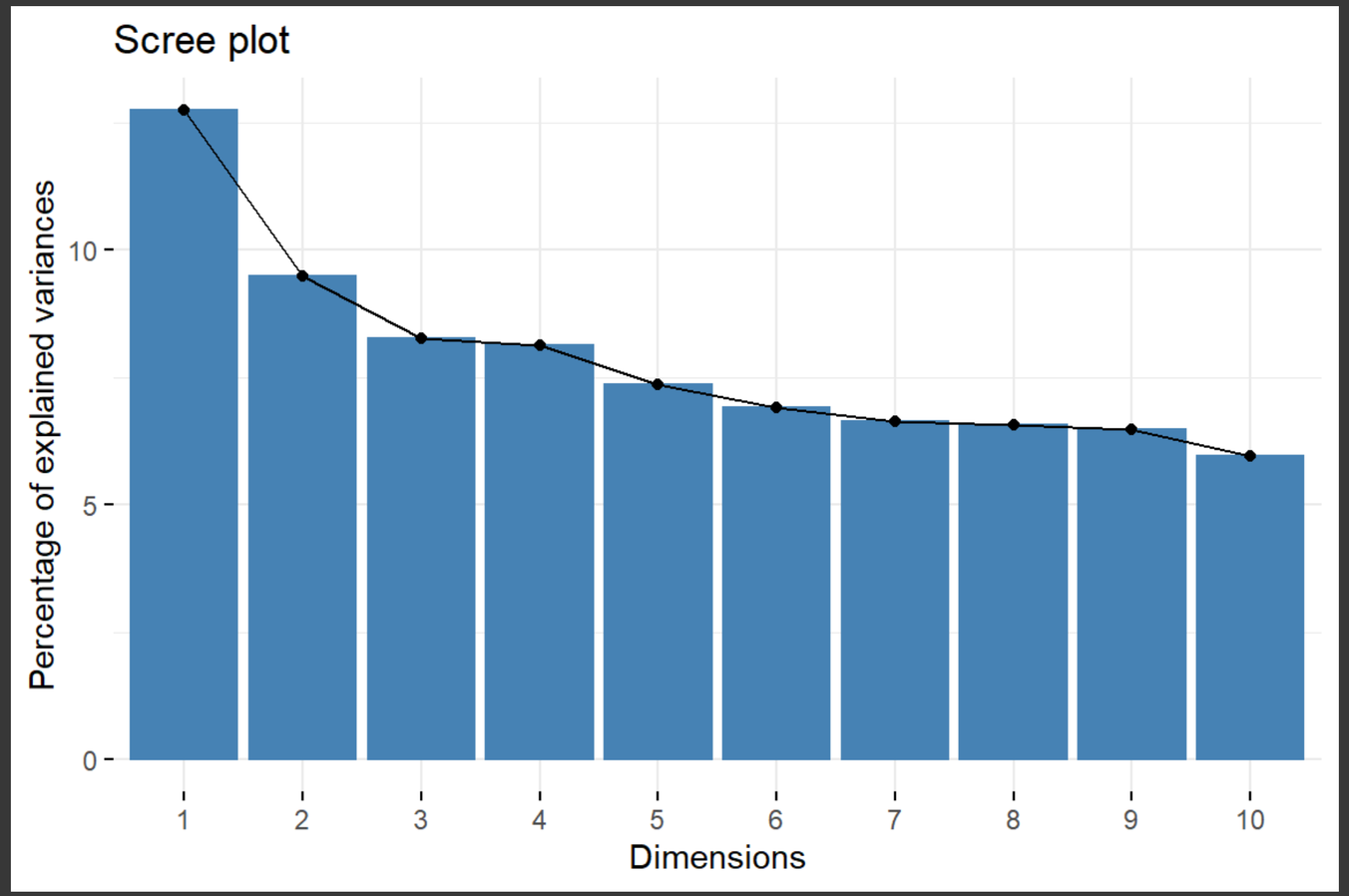
LDA

PCA

- Find useful **relationships** between variables and individuals

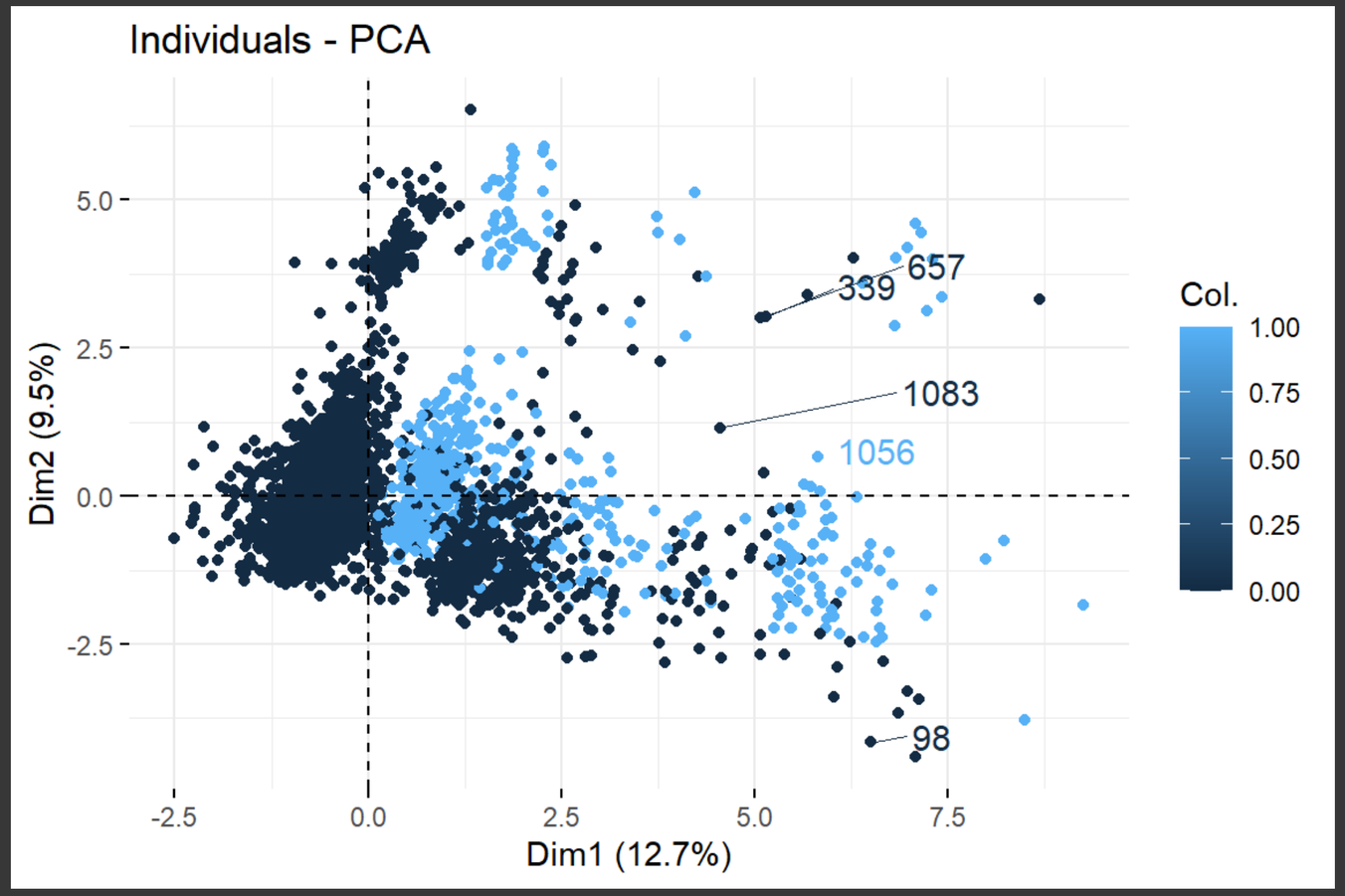
- Dimensionality reduction: **visualize the final dataset**

Scree Plot



- **Variance**:
  - Uniformly spread along more dimensions

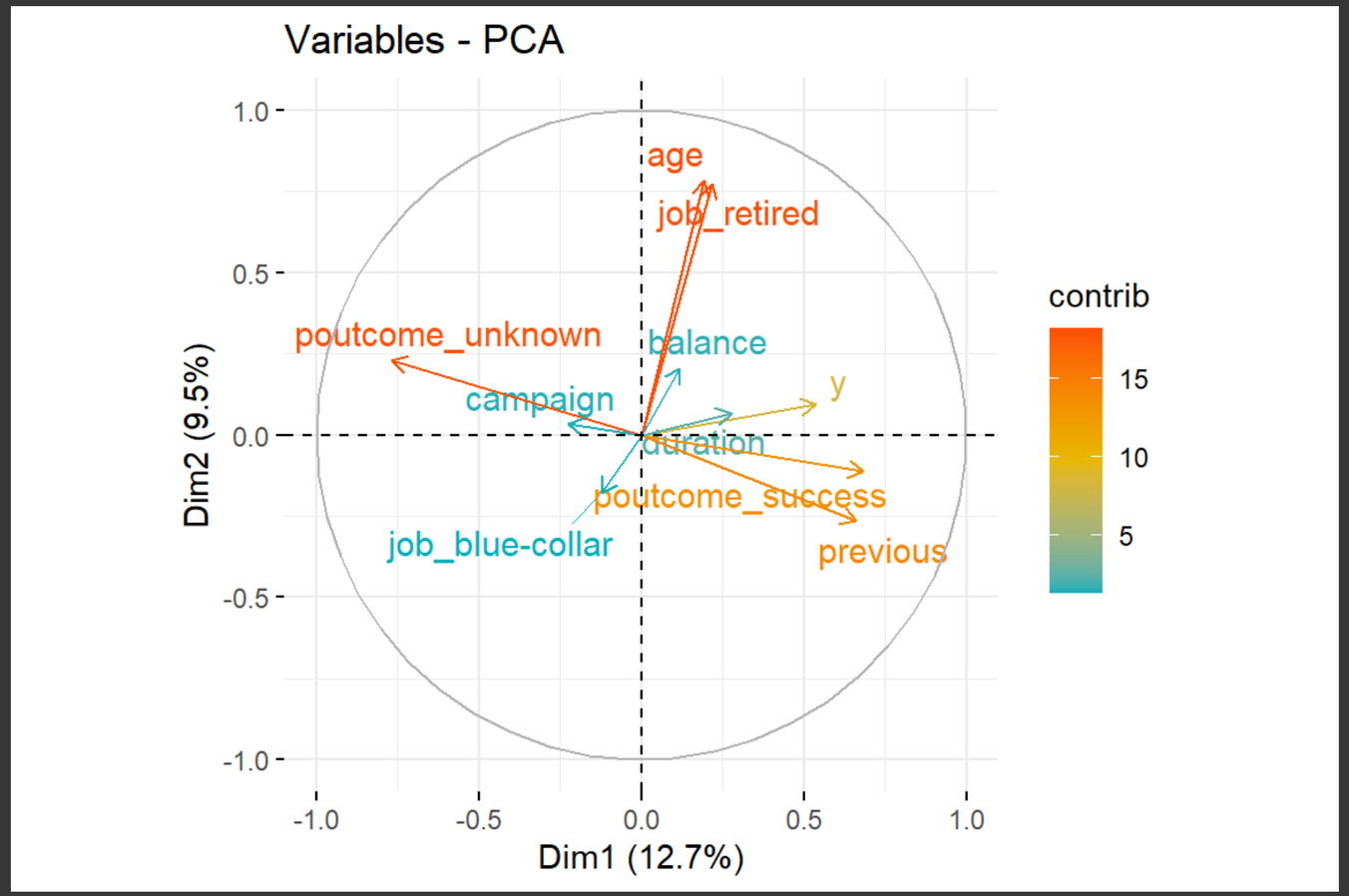
Individuals Plot by Target



- Negative samples in dark blue and positive ones in light blue
- **Interpretation**:
  - first factorial plane creates clusters of individuals

Variables Plot

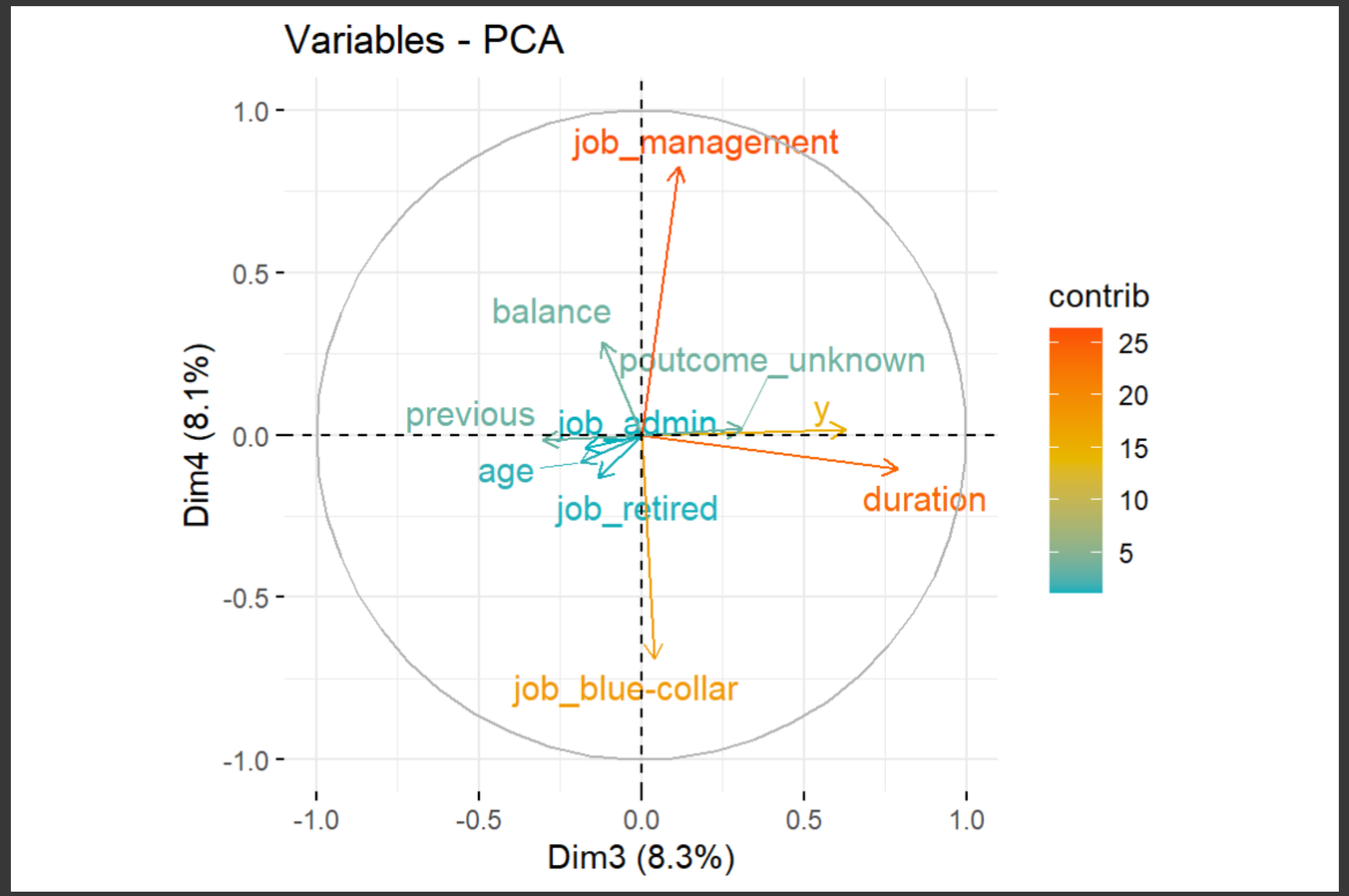
1st and 2nd axes



- **Observations**:
  - High correlation for the variables 'previous' and 'poutcome_success'
  - High correlation between 'job_retired' and 'age'
  - Negative correlation between 'poutcome_success' and 'poutcome_unknown'

Variables Plot

2nd and 3rd axes

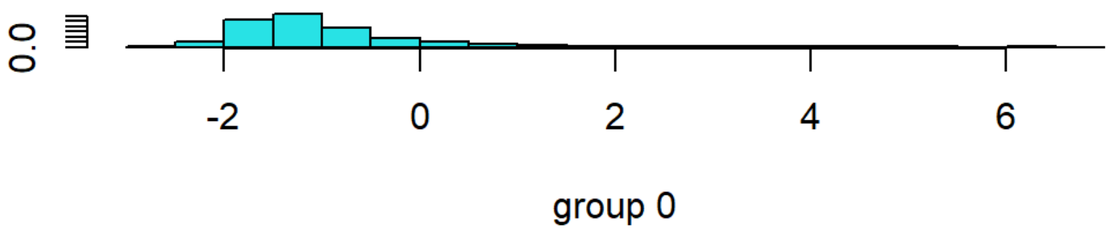


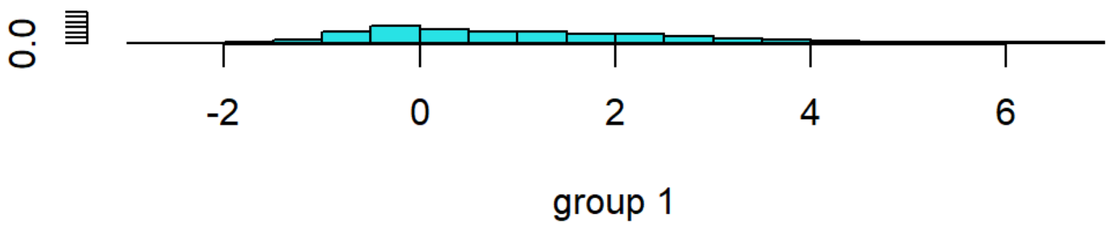
- **Observations:**
  - Same correlation among 'age' and 'job retired'.
  - Positive correlation between the 'y' and 'duration'

LDA

- Dimensionality reduction: **visualize the final dataset**
- Classification point of view: can the 2 classes be linearly separated?

LDA





- **Interpretation**:
  - New dimension created by LDA not able to separate the points into 2 clusters, a lot of overlapping around the value of 0 of the axes

Choosing the right performance metric

- Among all performance metrics available, which suits the most our problem ?
- F1 score for taking into account Recall and Precision
- Balanced Accuracy to weight class repartition

Predictions quality metrics

PART 1

True Positive Rate (TPR) $= \frac{TP}{TP + FN}$

or *Sensitivity*
or *Recall*
or *Hit Rate*

True Negative Rate (TNR) $= \frac{TN}{TN + FP}$

or *Specificity*
or *Selectivity*

Accuracy Score $= \frac{TP + TN}{TP + TN + FP + FN}$

ONLY for balanced data!

BTW for binary classificator Balanced Accuracy is equal to AUC ROC

Balanced Accuracy $= \frac{TPR + TNR}{2}$

ok for imbalanced dataframe, cares about detecting positives and negatives

Precision $= \frac{TP}{TP + FP}$

should be in balance with Recall

F1 $= 2 * \frac{Precision * Recall}{Precision + Recall}$

ok for imbalanced dataframe, cares about detecting positives

Source : https://towardsdatascience.com/performance-metrics-for-binary-classifier-in-simple-words-be958535db49

ALGORITHMS: KNN

Quick intro to KNN

- Supervised learning classifier
- Based on distances:
  - standardized numerical variables
- Parameters:
  - number of neighbours (K): to be fine tuned
  - distance measure: euclidean
- Prediction:
  - majority vote

KNN: Optimization and Evaluation

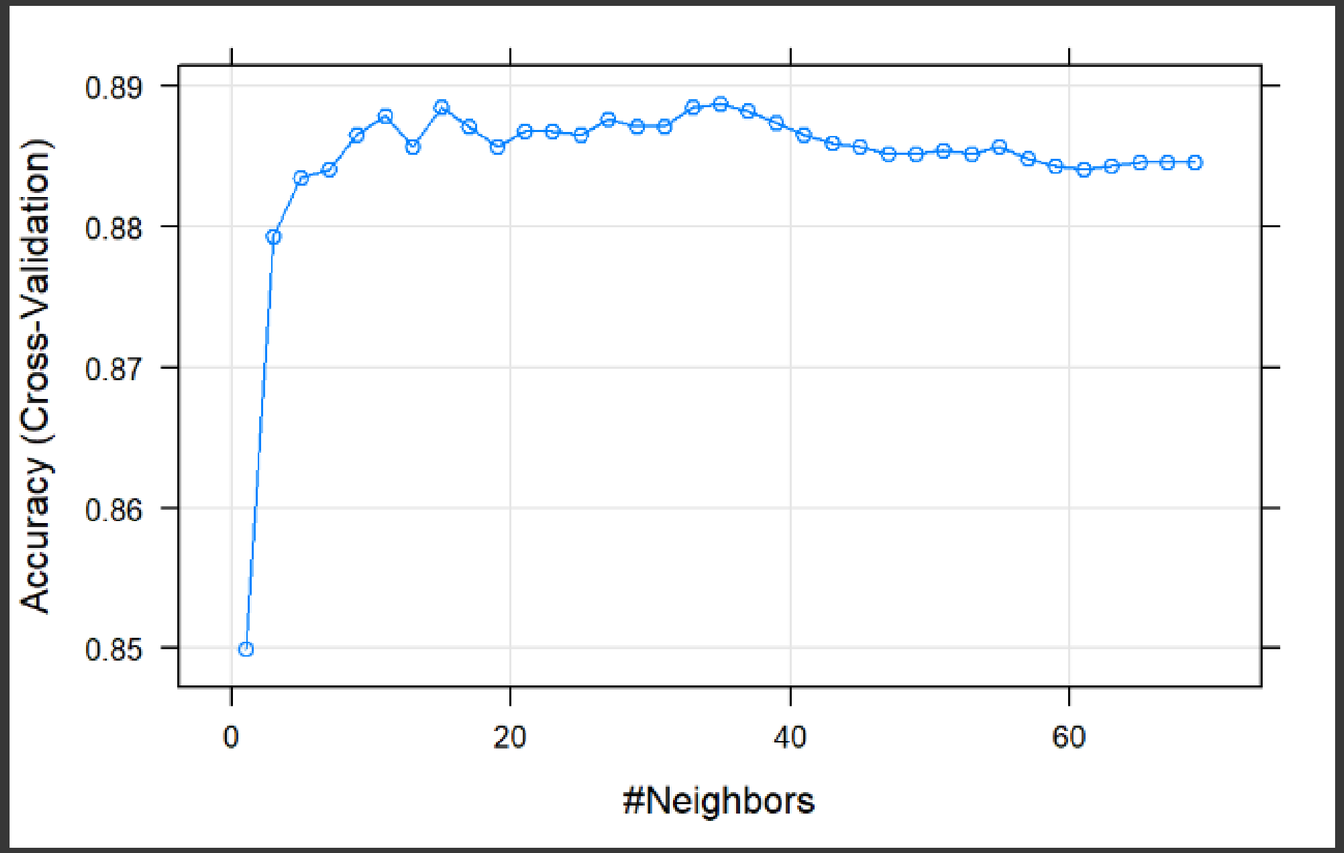
- **Optimization strategy**:
  - 10-fold cross-validation on the training dataset
- **Performance evaluation on test dataset**:
  - Confusion matrix
  - F1-score
  - Balanced accuracy

KNN:
Models

- **Baseline model**:
  - find K by optimizing Accuracy
- **Optimized model**:
  - find K by optimizing F1-score

Baseline model

Optimize K by using Accuracy

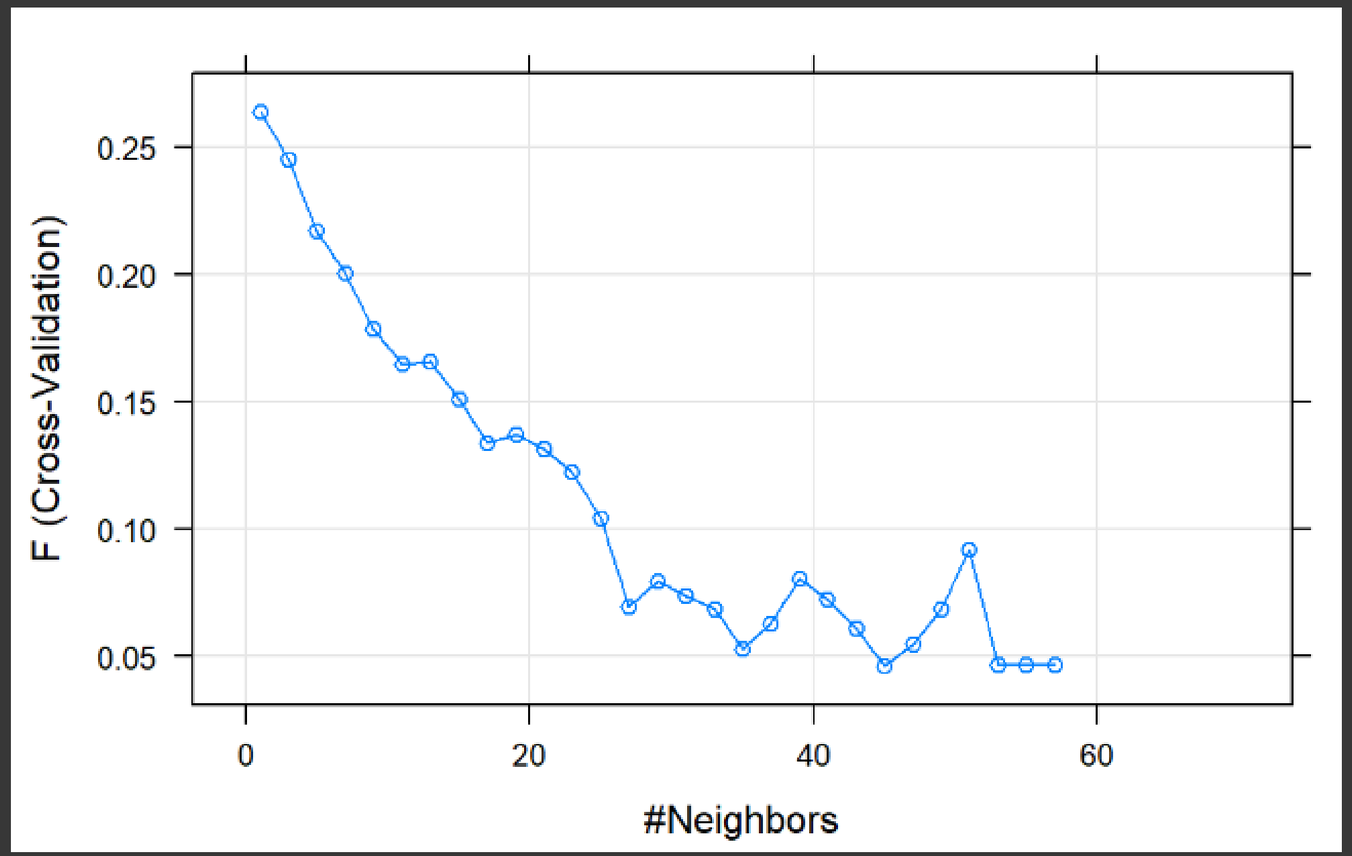


Optimal K: 35

```
Confusion Matrix and Statistics

          Reference
Prediction yes  no
       yes   1   0
       no  103 798

               Accuracy : 0.8858
                 95% CI : (0.8632, 0.9058)
    No Information Rate : 0.8847
    P-Value [Acc > NIR] : 0.4845

                  Kappa : 0.0169

 Mcnemar's Test P-Value : <2e-16

              Precision : 1.000000
                 Recall : 0.009615
                     F1 : 0.019048
             Prevalence : 0.115299
         Detection Rate : 0.001109
   Detection Prevalence : 0.001109
      Balanced Accuracy : 0.504808
```

Optimized model

Optimize K by using F1-score



Optimal K: 1

```
Confusion Matrix and Statistics

          Reference
Prediction yes  no
       yes  35  45
       no   69 753

               Accuracy : 0.8736
                 95% CI : (0.8501, 0.8946)
    No Information Rate : 0.8847
    P-Value [Acc > NIR] : 0.86267

                  Kappa : 0.3114

 Mcnemar's Test P-Value : 0.03123

              Precision : 0.43750
                 Recall : 0.33654
                     F1 : 0.38043
             Prevalence : 0.11530
         Detection Rate : 0.03880
   Detection Prevalence : 0.08869
      Balanced Accuracy : 0.64007
```

Optimized model: variable importance

Area under the ROC curve for each predictor:

A series of cutoffs is applied to the predictor data to predict the class. The sensitivity and specificity are computed for each cutoff and the ROC curve is computed.

ROC curve variable importance

only 20 most important variables shown

	Importance
duration	100.000
previous	30.411
pdays_Never contacted	29.592
poutcome_unknown	24.796
housing_yes	24.094
housing_no	24.094
poutcome_success	22.122
balance	21.956
month_may	21.670
campaign	17.407
marital_married	14.896
education_tertiary	14.351
job_blue-collar	13.634
pdays_Under 6months	12.908
loan_yes	12.196
loan_no	12.196
month_oct	11.035
job_retired	10.079
marital_single	9.757
education_secondary	8.946

ALGORITHMS: SVM

Quick intro to SVM classifier

- Supervised learning classifier
- Kernel-based method (instanced-based):
  - kernel function representing similarity between samples
- Parameters:
  - Kernel function
  - Cost (C)
- Training:
  - find hyperplane maximizing margin
  - scaled data
- Prediction:
  - step function applied to weighted sum of kernel similarity with support vectors

SVM: Optimization and Evaluation

- **Optimization strategy**:
  - 10-fold cross-validation on the training dataset
  - Optimize F1-score
- **Performance evaluation on test dataset**:
  - Confusion matrix
  - F1-score
  - Balanced accuracy

SVM: Models

- Baseline model
  - Linear Kernel
- Radial Kernel
- Polynomial Kernel
- Radial Kernel with down-sampling
- Radial Kernel with SMOTE sampling
- Optimized model:
  - down-sampling with more tuning

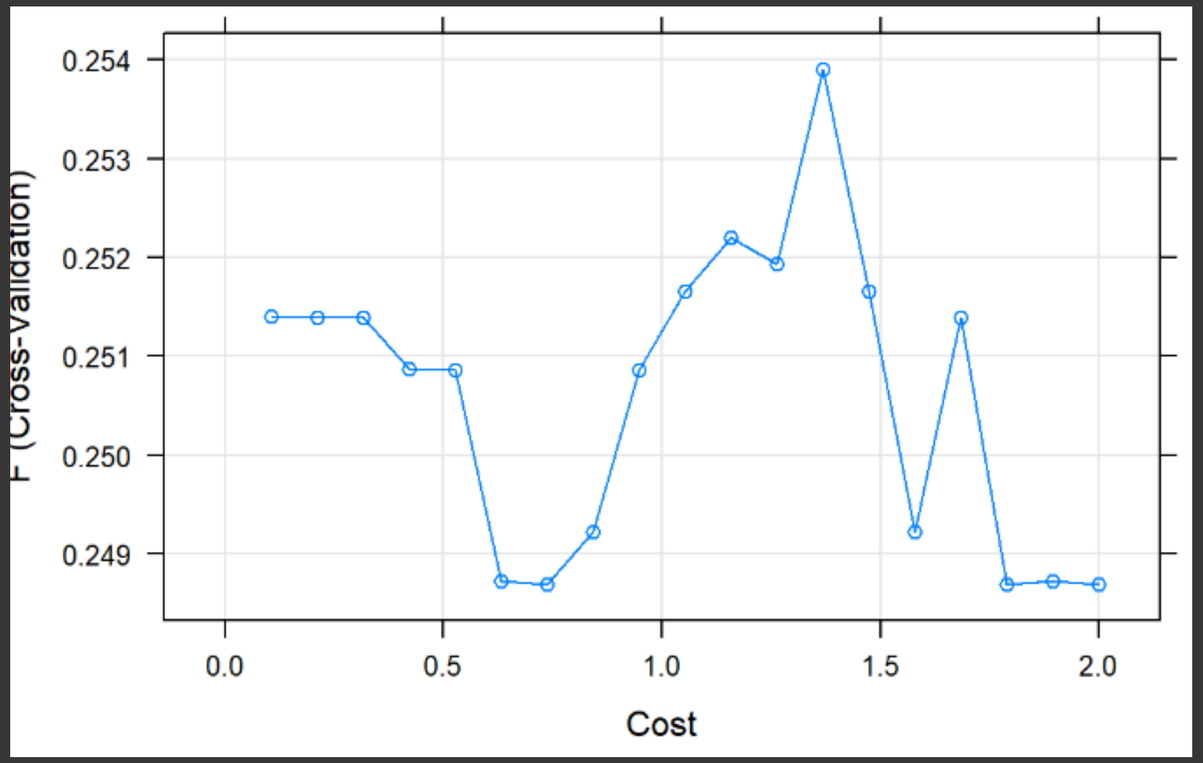
Baseline model

Linear Kernel, no tuning on C

```
          Reference
Prediction yes  no
       yes  19   5
       no   85 793

               Accuracy : 0.9002
                 95% CI : (0.8788, 0.919)
    No Information Rate : 0.8847
    P-Value [Acc > NIR] : 0.07755

                  Kappa : 0.2651

 Mcnemar's Test P-Value : < 2e-16

              Precision : 0.79167
                 Recall : 0.18269
                     F1 : 0.29688
             Prevalence : 0.11530
         Detection Rate : 0.02106
   Detection Prevalence : 0.02661
      Balanced Accuracy : 0.58821
```

Linear Kernel

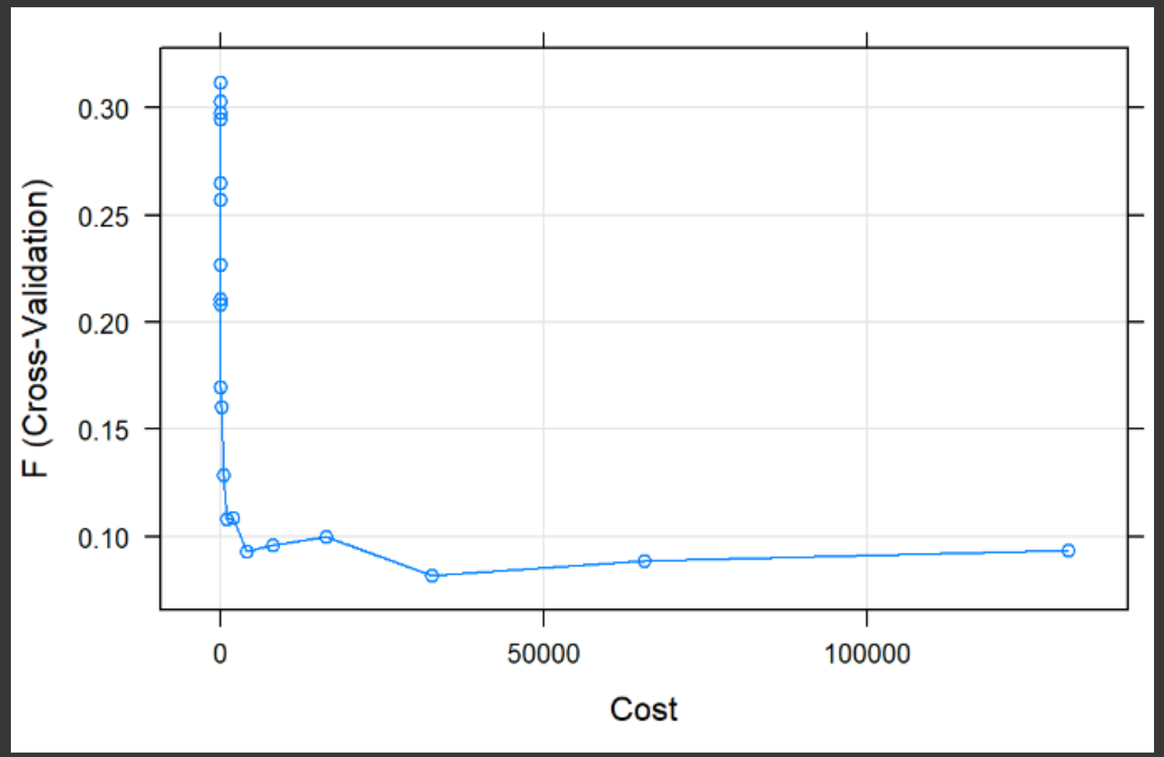
Tuning C parameter



- **Optimal C**: 1.4
- **Same performances** as C = 1
- Parameter C doesn't change the performances in the Linear Kernel

Radial Kernel

Random tuning

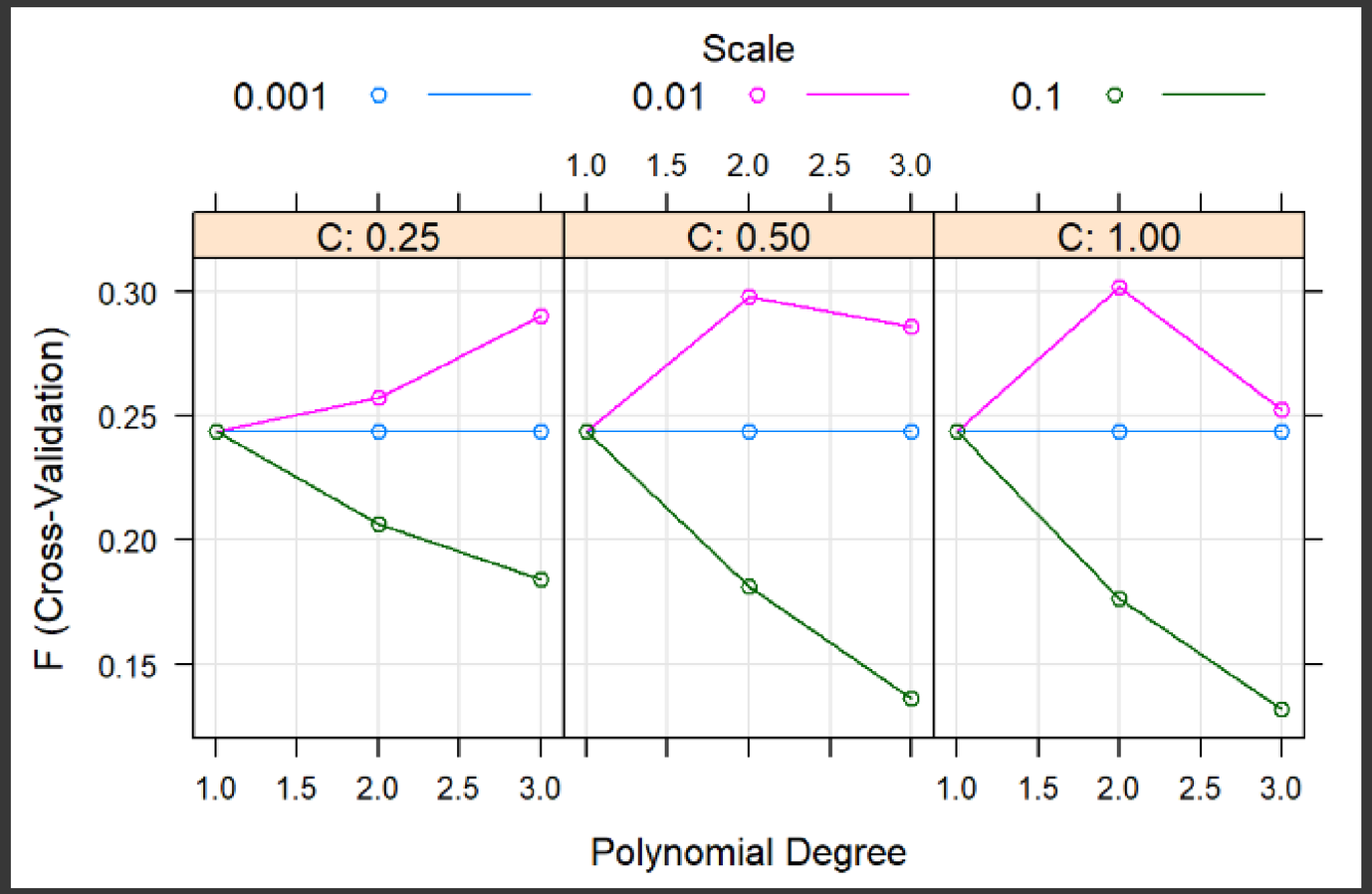


- Optimal C: 5
- sigma (fixed): 0.014

```
Confusion Matrix and Statistics

          Reference
Prediction yes  no
       yes  23   7
       no   81 791

               Accuracy : 0.9024
                 95% CI : (0.8812, 0.921)
    No Information Rate : 0.8847
    P-Value [Acc > NIR] : 0.05056

                  Kappa : 0.3075

 Mcnemar's Test P-Value : 7.149e-15

              Precision : 0.76667
                 Recall : 0.22115
                     F1 : 0.34328
             Prevalence : 0.11530
         Detection Rate : 0.02550
   Detection Prevalence : 0.03326
      Balanced Accuracy : 0.60619
```

Polynomial Kernel

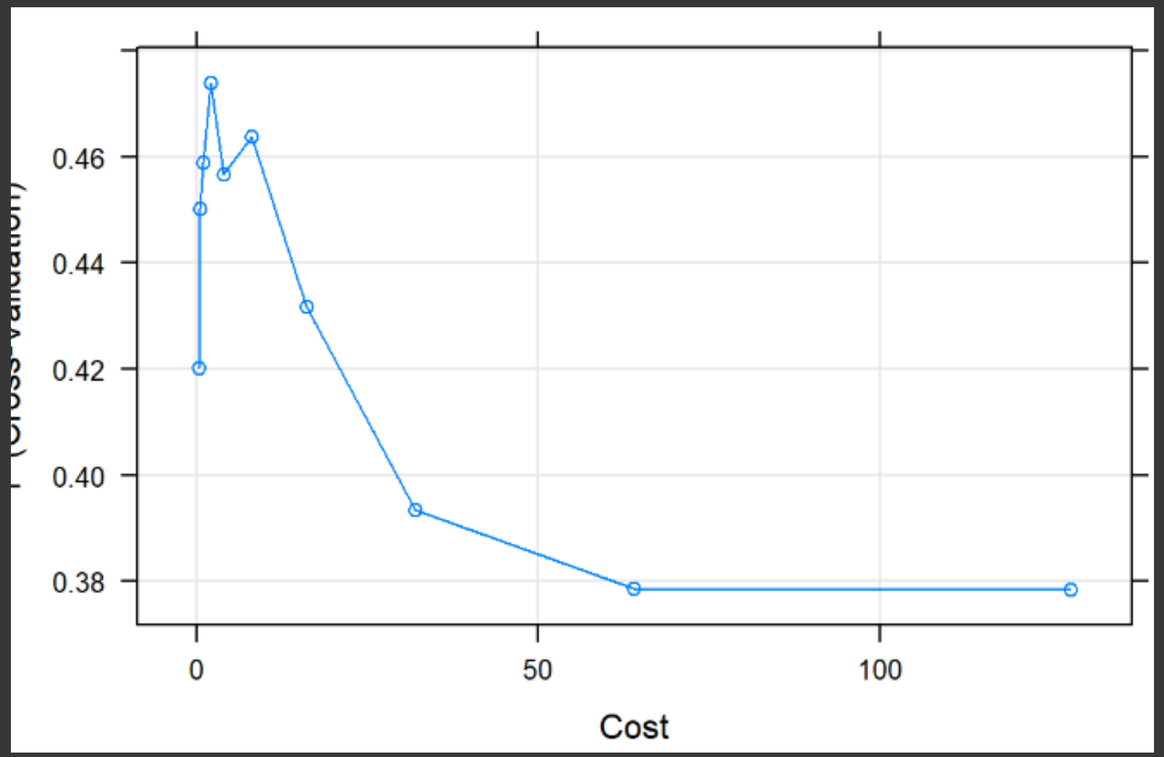
Random tuning




```
Confusion Matrix and Statistics

          Reference
Prediction yes  no
       yes  25   9
       no   79 789

               Accuracy : 0.9024
                 95% CI : (0.8812, 0.921)
    No Information Rate : 0.8847
    P-Value [Acc > NIR] : 0.05056

                  Kappa : 0.3239

 Mcnemar's Test P-Value : 1.903e-13

              Precision : 0.73529
                 Recall : 0.24038
                     F1 : 0.36232
             Prevalence : 0.11530
         Detection Rate : 0.02772
   Detection Prevalence : 0.03769
      Balanced Accuracy : 0.61455
```

Radial Kernel

Random tuning + Down-Sampling

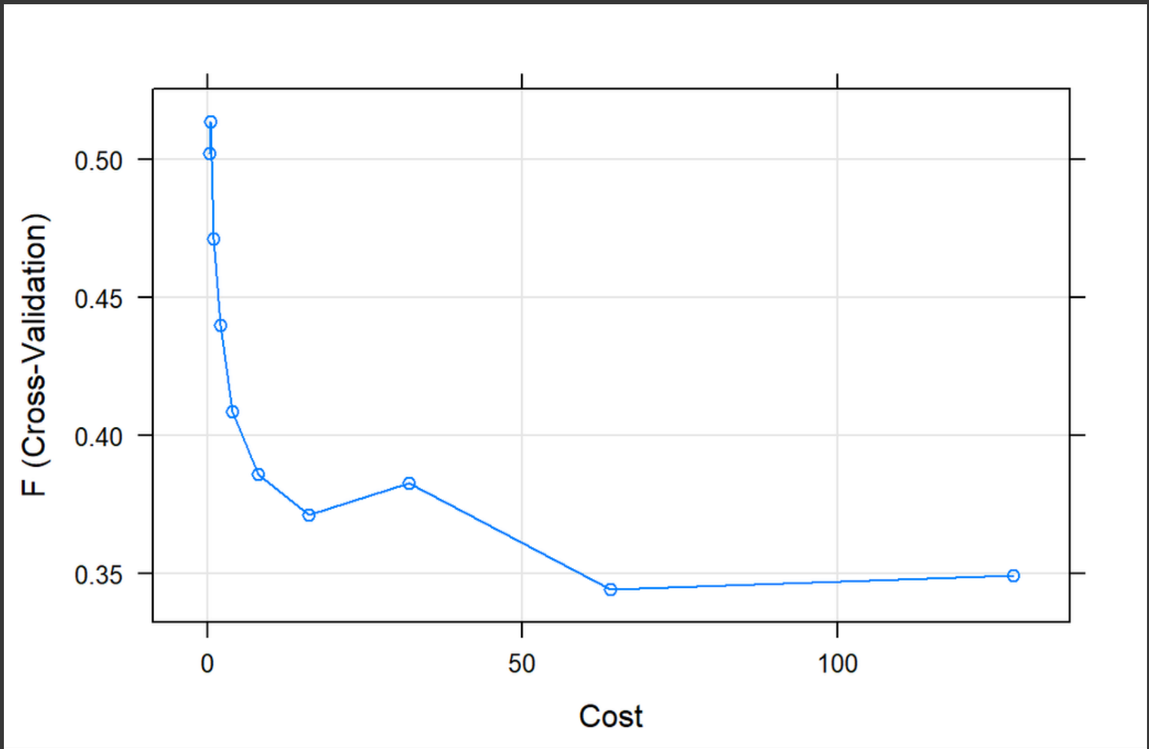


- Optimal C: 2
- Sigma (fixed): 0.014

```
Confusion Matrix and Statistics

          Reference
Prediction yes  no
       yes  86 150
       no   18 648

               Accuracy : 0.8137
                 95% CI : (0.7868, 0.8386)
    No Information Rate : 0.8847
    P-Value [Acc > NIR] : 1

                  Kappa : 0.4117

 Mcnemar's Test P-Value : <2e-16

              Precision : 0.36441
                 Recall : 0.82692
                     F1 : 0.50588
             Prevalence : 0.11530
         Detection Rate : 0.09534
   Detection Prevalence : 0.26164
      Balanced Accuracy : 0.81948
```

Radial Kernel

Random tuning + SMOTE Sampling



- Optimal C: 0.5
- Sigma (fixed): 0.014

```
Confusion Matrix and Statistics

          Reference
Prediction yes  no
       yes  59  58
       no   45 740

               Accuracy : 0.8858
                 95% CI : (0.8632, 0.9058)
    No Information Rate : 0.8847
    P-Value [Acc > NIR] : 0.4845

                  Kappa : 0.4691

 Mcnemar's Test P-Value : 0.2370

              Precision : 0.50427
                 Recall : 0.56731
                     F1 : 0.53394
             Prevalence : 0.11530
         Detection Rate : 0.06541
   Detection Prevalence : 0.12971
      Balanced Accuracy : 0.74731
```

Radial Kernel

Grid Tuning+ Down-Sampling

```
Confusion Matrix and Statistics

          Reference
Prediction yes  no
       yes  83 130
       no   21 668

               Accuracy : 0.8326
                 95% CI : (0.8066, 0.8564)
    No Information Rate : 0.8847
    P-Value [Acc > NIR] : 1

                  Kappa : 0.4363

 Mcnemar's Test P-Value : <2e-16

              Precision : 0.38967
                 Recall : 0.79808
                     F1 : 0.52366
             Prevalence : 0.11530
         Detection Rate : 0.09202
   Detection Prevalence : 0.23614
      Balanced Accuracy : 0.81758
```

- Optimal C: 2
- Sigma (**tuned**): 0.01

Final Radial SVM: variable importance

Area under the ROC curve for each predictor:

A series of cutoffs is applied to the predictor data to predict the class. The sensitivity and specificity are computed for each cutoff and the ROC curve is computed.

- NOTE: same as KNN

ROC curve variable importance

only 20 most important variables shown

	Importance
duration	100.000
previous	30.411
pdays_Never contacted	29.592
poutcome_unknown	24.796
housing_yes	24.094
housing_no	24.094
poutcome_success	22.122
balance	21.956
month_may	21.670
campaign	17.407
marital_married	14.896
education_tertiary	14.351
job_blue-collar	13.634
pdays_Under 6months	12.908
loan_no	12.196
loan_yes	12.196
month_oct	11.035
job_retired	10.079
marital_single	9.757
education_secondary	8.946

ALGORITHMS: RANDOM FORESTS

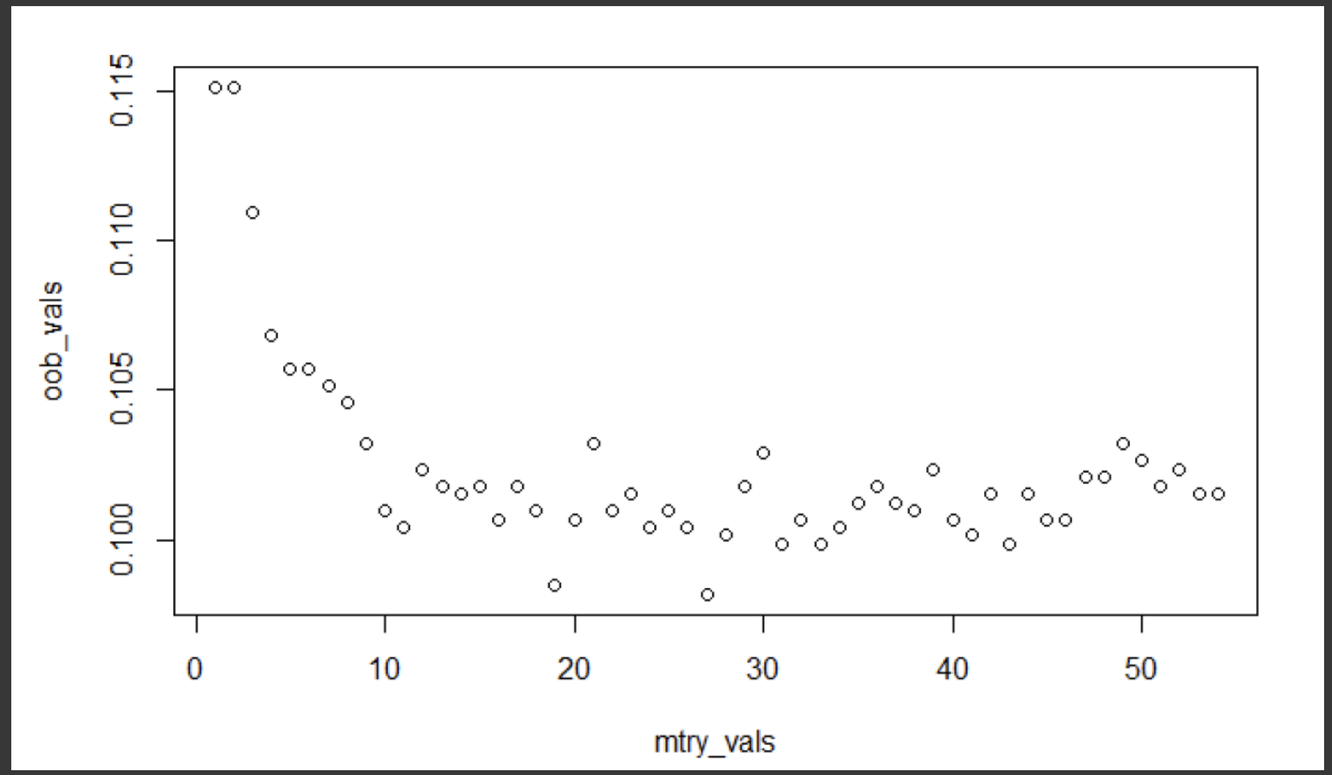
Quick intro to RF's

- Classifications trees trained on **bags of samples**
- Two parameters:
  - Number of trees (ntree)
  - Number of variables used (mtry)
- Evaluated through the **Out-Of-Bag error**

Default run & Optimal mtry

- Default Run (mtry=7 and ntree=500)
  - Balanced Accuracy 62.24 %
  - F1-Score 94.35 %
- Optimal number of variables (mtry=27):
  - Balanced Accuracy 67 (+5 %)
  - F1-Score 94.29 (~ unchanged)

Default run & Optimal mtry

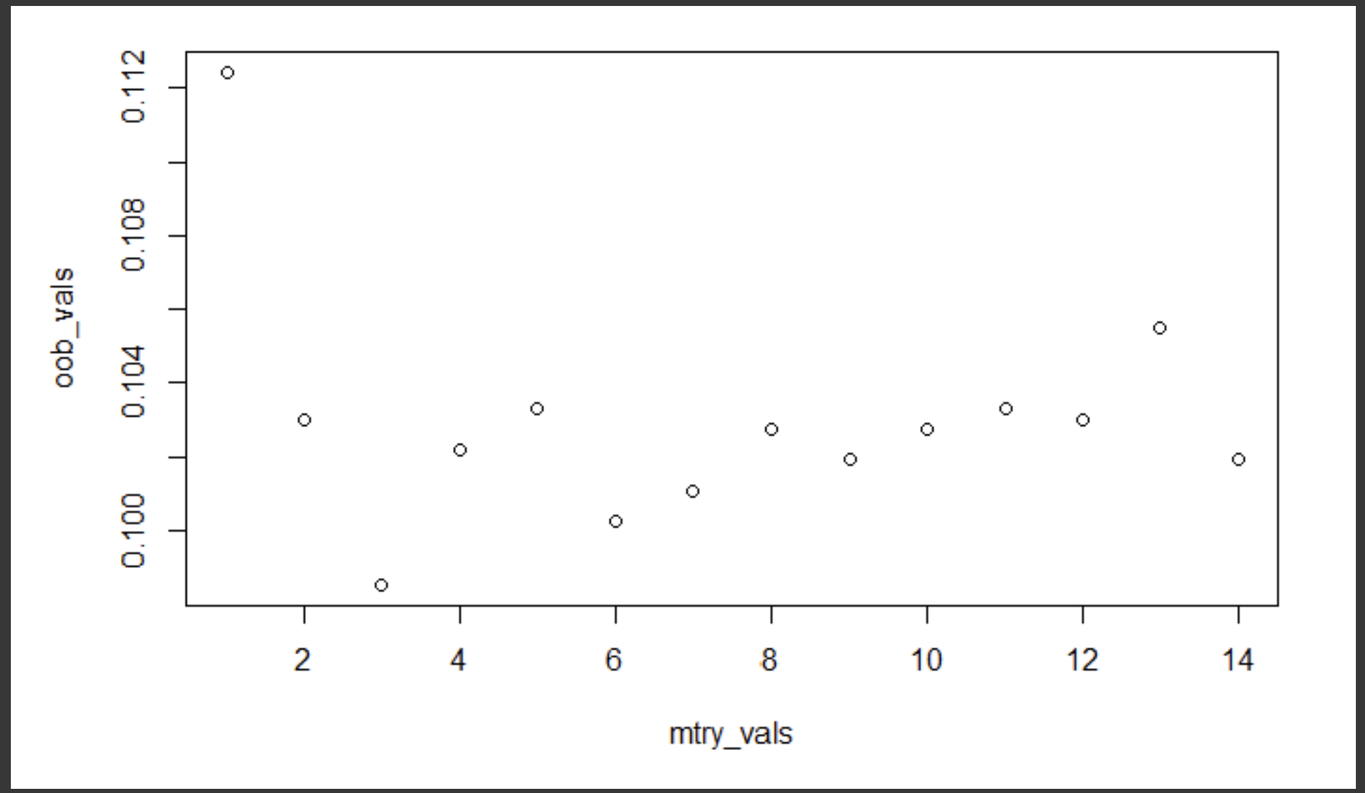


Selected mtry = 27

One-hot to categorical

- One-hot encoding of categorical variables (e.g. "job") introduced **many variables**
- Unecessary splits for the trees
- **Default run**, with **categorical**:
  - BA: 67 % (+5 % than with One-hot)
  - F1: 94 % (~ untouched)

Number of variables optimization w/ categorical

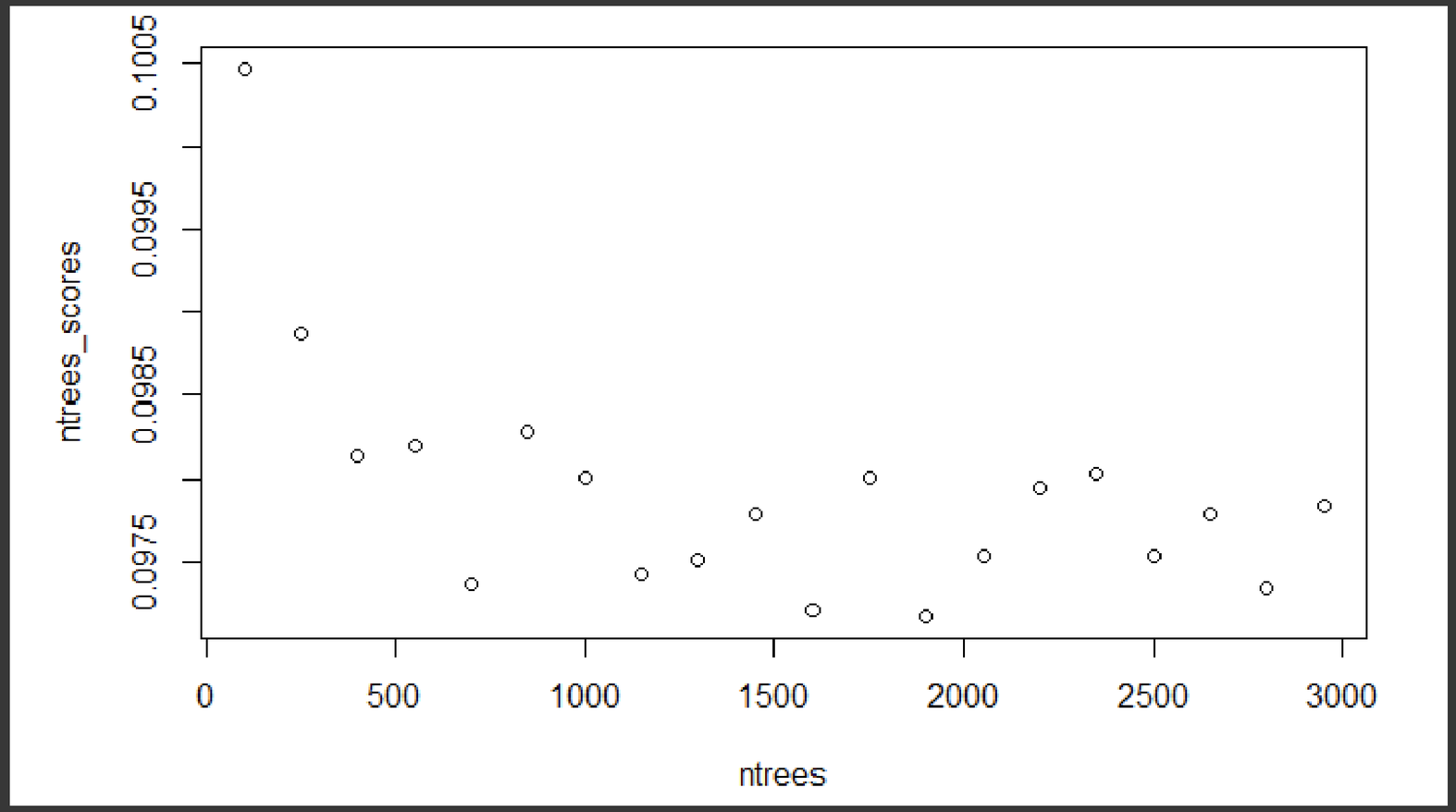


With mtry=3 (minimum) and ntree=500

BA: 69.48% (+2 %) and F1: 94% (untouched)

Number of trees
w/categorical

10 iterations of RF for each ntree, averaged OOB (mtry = 3)



BA: 67.83 % (-1.5%) and F1: 94.29%

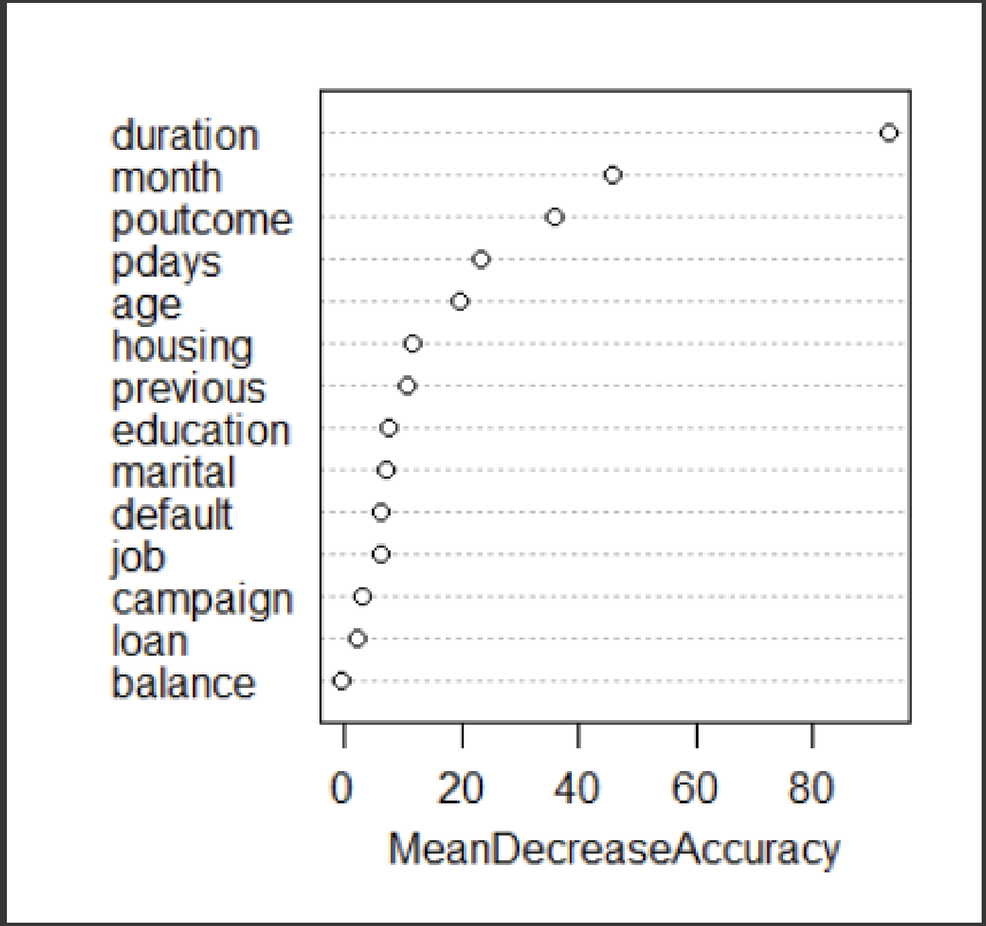
Method	OOB error	Balanced Accuracy	F1-Score
Default run	10.43	62.24	94.35
Opt. mtry	10.07	67	94.29
Default run, w/ cat.	9.94	67.35	94.24
Opt. mtry, w/ cat.	9.94	69.48	94.06
Opt. ntree, w/ cat.	9.94	67.83	94.29

Random Forest Performances

A word about variable importance

Mean Decrease Accuracy:

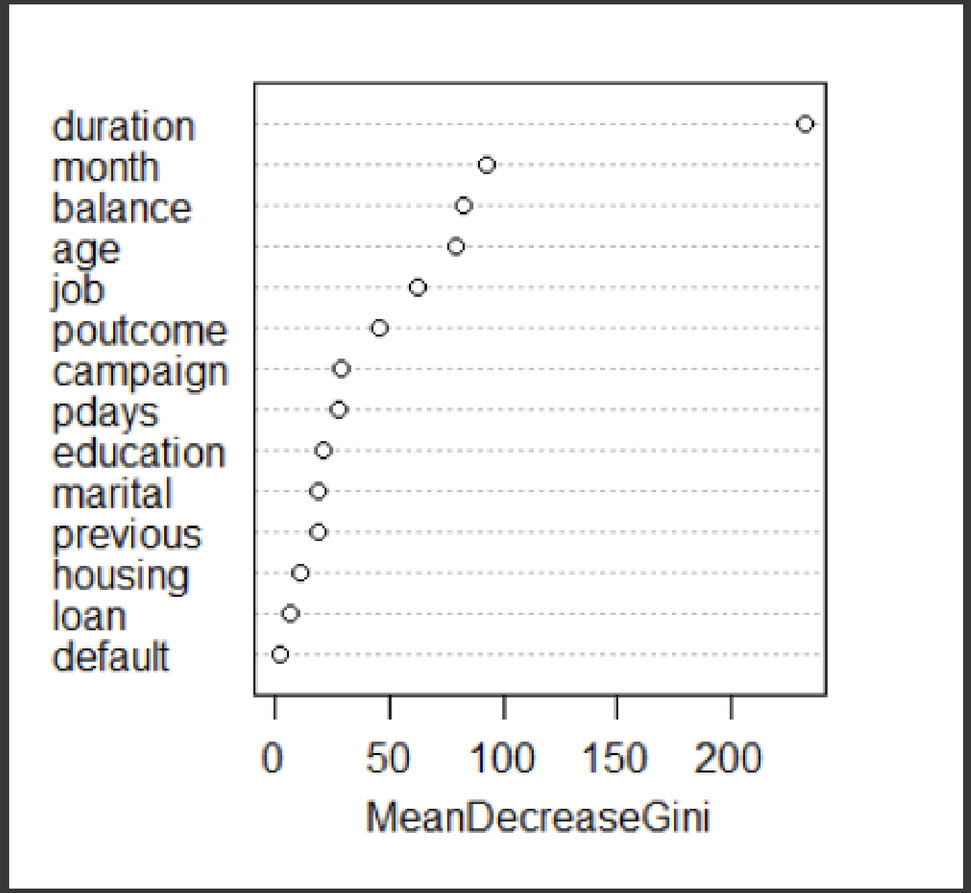
How much the accuracy decreases when removing a variable from the bag



A word about variable importance

Mean Decrease Gini Impurity:

How much each variable improve splits on average Gini purity ?



ALGORITHMS: NEURAL NETWORK

Quick intro to NN classifier

- Supervised learning classifier
- Structure of the NN:
  - Input layer
  - 1 hidden layer
  - Output layer
- Parameters:
  - neurons in the hidden layer
  - weight decay rate
- Prediction:
  - output neuron produces probability of the positive class

NN: Optimization and Evaluation

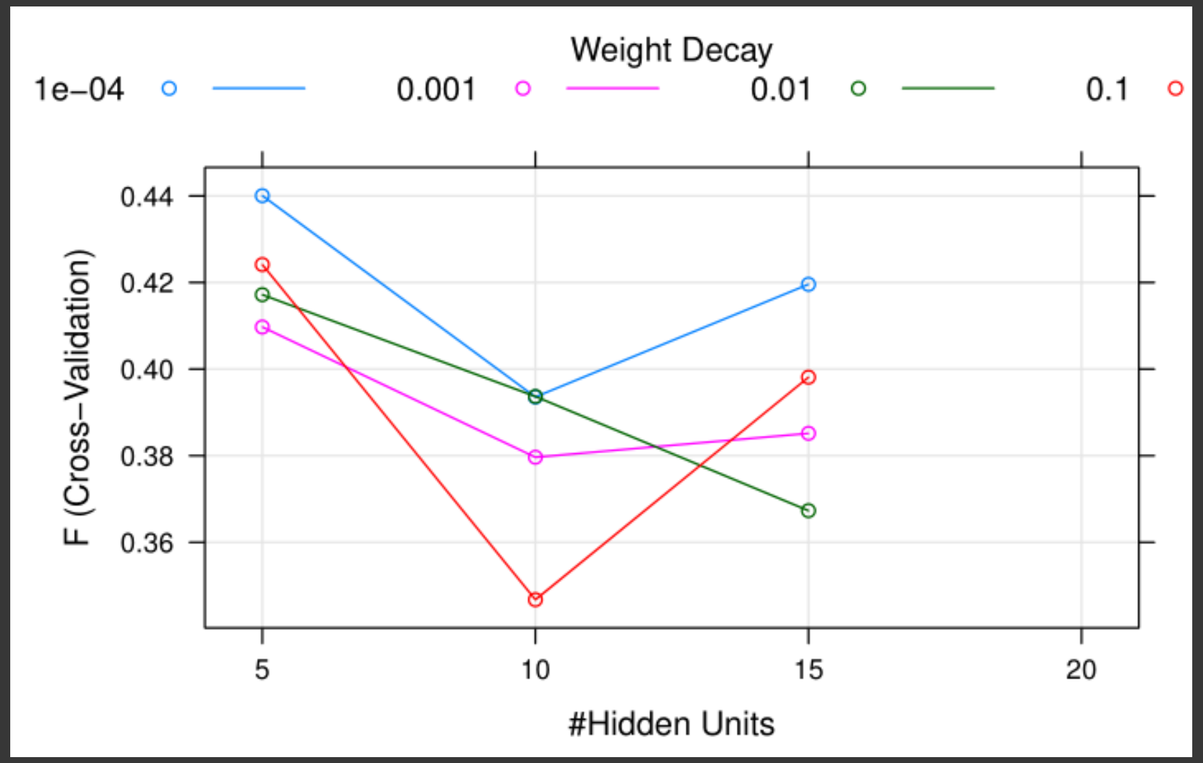
- **Optimization strategy**:
  - 10-fold cross-validation on the training dataset
  - Optimize F1-score
- **Performance evaluation on test dataset**:
  - Confusion matrix
  - F1-score
  - Balanced accuracy

NN: Models

- **Baseline model**:
  - grid search on hyperparameters using F1-score
- **Optimized model**:
  - applying sample weighting

Baseline model

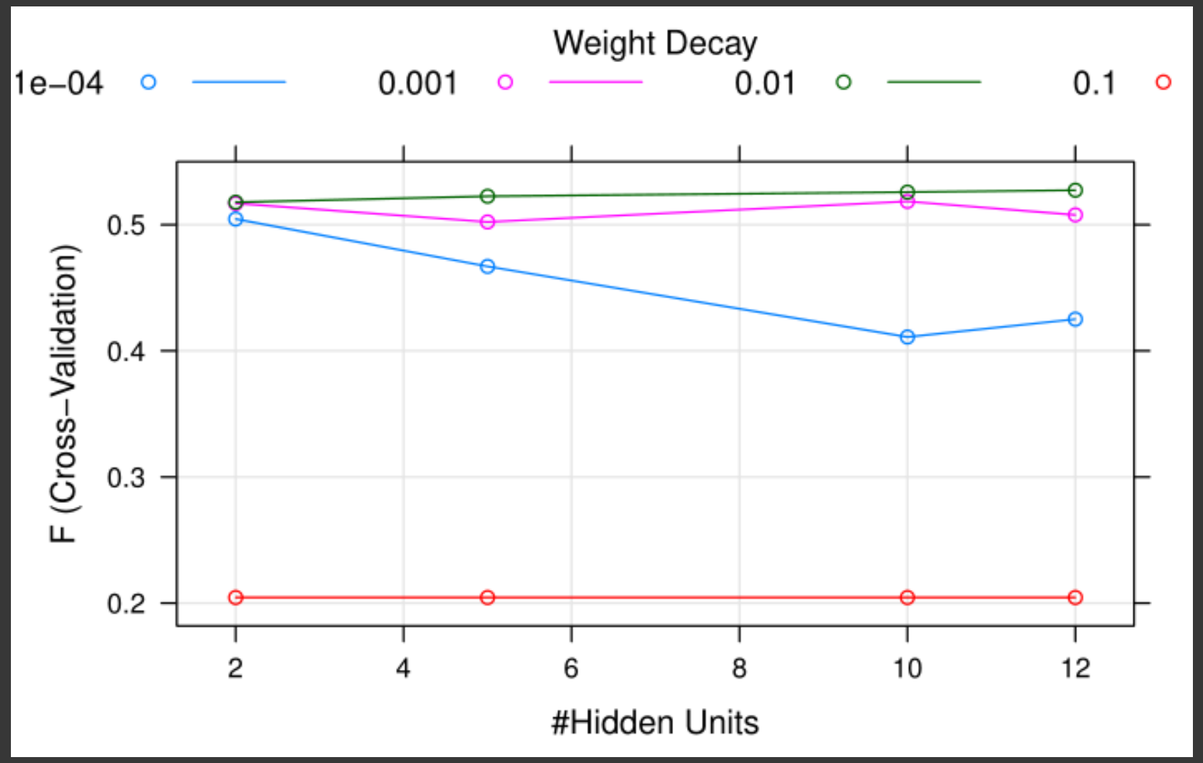
Not dealing with class unbalance




```
Confusion Matrix and Statistics

          Reference
Prediction yes  no
       yes  37  28
       no   67 770

               Accuracy : 0.8947
                 95% CI : (0.8728, 0.914)
    No Information Rate : 0.8847
    P-Value [Acc > NIR] : 0.1885

                  Kappa : 0.3832

 Mcnemar's Test P-Value : 9.67e-05

              Precision : 0.56923
                 Recall : 0.35577
                     F1 : 0.43787
             Prevalence : 0.11530
         Detection Rate : 0.04102
   Detection Prevalence : 0.07206
      Balanced Accuracy : 0.66034
```

Optimized model

Sample weighting




```
Confusion Matrix and Statistics

          Reference
Prediction yes  no
       yes  79 120
       no   25 678

               Accuracy : 0.8392
                 95% CI : (0.8136, 0.8626)
    No Information Rate : 0.8847
    P-Value [Acc > NIR] : 1

                  Kappa : 0.436

 Mcnemar's Test P-Value : 5.89e-15

              Precision : 0.39698
                 Recall : 0.75962
                     F1 : 0.52145
             Prevalence : 0.11530
         Detection Rate : 0.08758
   Detection Prevalence : 0.22062
      Balanced Accuracy : 0.80462
```

Optimized model: variable importance

Garson's algorithm:

Idenitfy all weights connecting the specific input node that pass through the hidden layer to the output neuron.

only 20 most important variables shown (out of 53)

	Overall
duration	100.000
poutcome_success	27.846
month_oct	24.890
month_mar	18.248
month_may	15.619
campaign	15.063
job_retired	14.175
‘pdays_Under 1.5yrs‘	13.051
‘pdays_Never contacted‘	12.527
month_sep	11.500
month_apr	10.957
loan_no	9.792
loan_yes	9.791
‘pdays_Under 6months‘	8.697
‘job_blue-collar‘	8.346
housing_yes	8.137
housing_no	8.137
‘pdays_Under 3months‘	7.959
poutcome_failure	7.823
‘pdays_Over 1.5yrs‘	7.496

CONCLUSIONS

Conclusions

Method	Balanced Accuracy	F1-Score
K-NN	64.01%	38.04%
SVM	81.76%	52.37%
Random Forest	67.35%	94.24%
Neural Network	80.46%	52.15%

Thanks for your attention

Preguntas ?