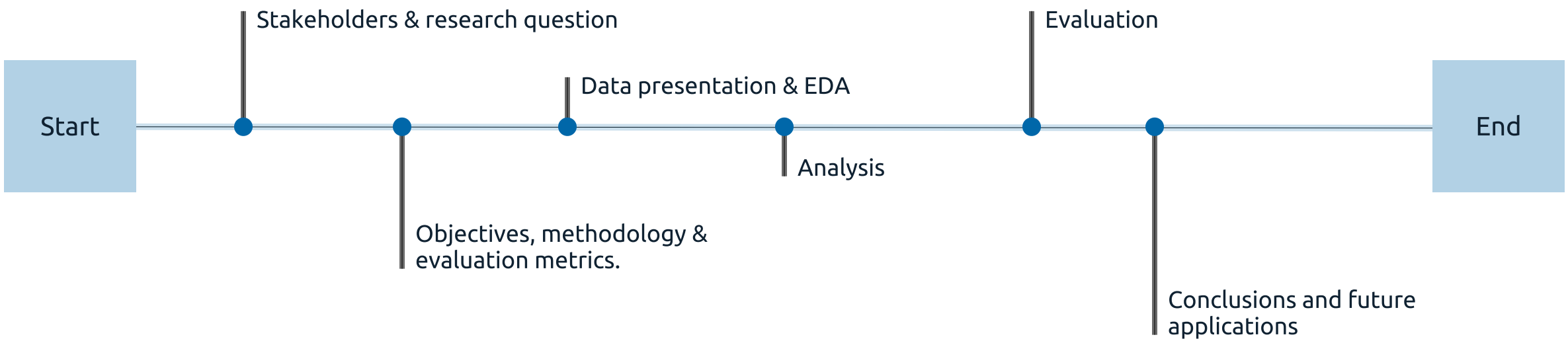
APPLIED MACHINE LEARNING FOR DIRECT MARKETING

An analysis over bank's clients

Statistical Learning Project

Andrea Corvi, Andrea Hüscher - AA 2019-2020 UCSC

Agenda


Start
Stakeholders & research question
Objectives, methodology & evaluation metrics.
Data presentation & EDA
Analysis
Evaluation
Conclusions and future applications
End

Section 1

Summary

Project Summary

- ## Stakeholder

  Caixa bank

- ## Objective

  Identify clients more likely to subscribe the long term deposit, in order to avoid waste of resources (human and monetary).

- ## Methodology

  Building four different binary classification models in order to predict the outcome for each client.

  - Logistic Regression
  - Support Vector Machines
  - Decision Trees
  - Gradient Boosting/Xgboost

- ## Evaluation

  Metric for evaluation is Area Under the Curve (AUC) since sensitivity is very important for this problem. Accuracy isn't enough

Data Set

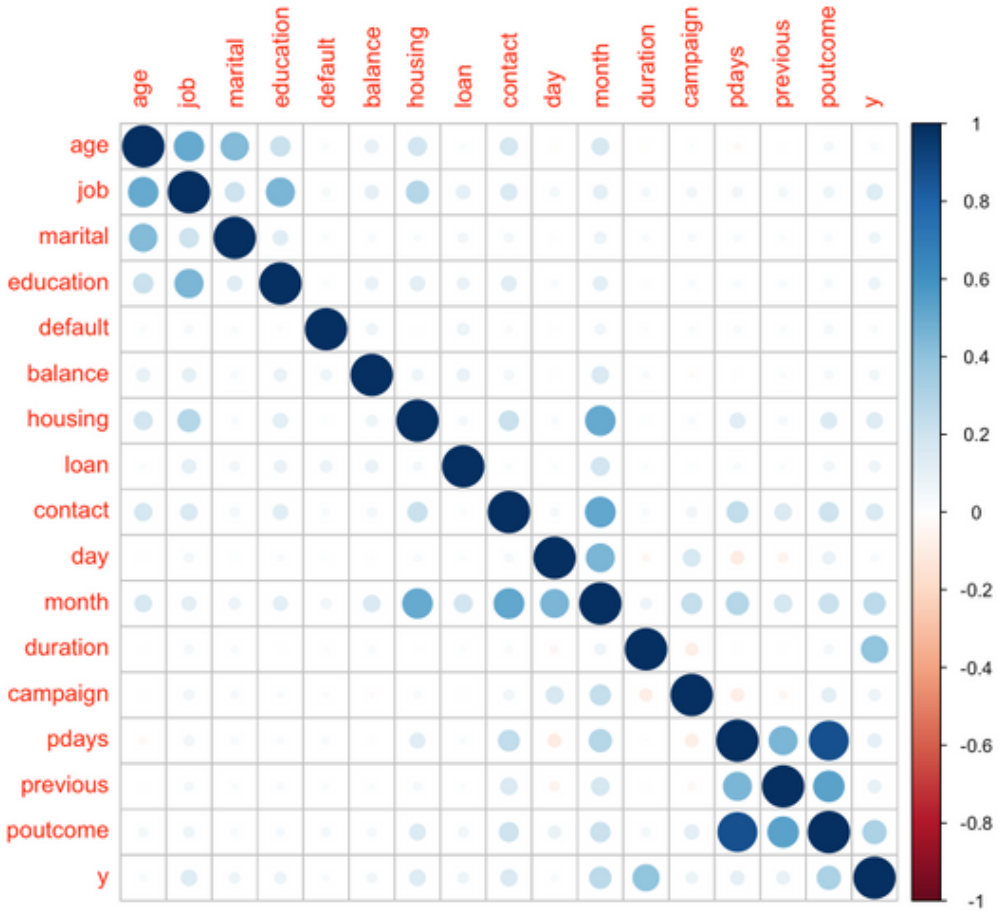
age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
30	unemployed	married	primary	no	1787	no	no	cellular	19	oct	79	1	-1	0	unknown	no
33	services	married	secondary	no	4789	yes	yes	cellular	11	may	220	1	339	4	failure	no
35	management	single	tertiary	no	1350	yes	no	cellular	16	apr	185	1	330	1	failure	no
30	management	married	tertiary	no	1476	yes	yes	unknown	3	jun	199	4	-1	0	unknown	no
59	blue-collar	married	secondary	no	0	yes	no	unknown	5	may	226	1	-1	0	unknown	no
35	management	single	tertiary	no	747	no	no	cellular	23	feb	141	2	176	3	failure	no
36	self-employed	married	tertiary	no	307	yes	no	cellular	14	may	341	1	330	2	other	no
39	technician	married	secondary	no	147	yes	no	cellular	6	may	151	2	-1	0	unknown	no
41	entrepreneur	married	tertiary	no	221	yes	no	unknown	14	may	57	2	-1	0	unknown	no
43	services	married	primary	no	-88	yes	yes	cellular	17	apr	313	1	147	2	failure	no
39	services	married	secondary	no	9374	yes	no	unknown	20	may	273	1	-1	0	unknown	no
43	admin.	married	secondary	no	264	yes	no	cellular	17	apr	113	2	-1	0	unknown	no
36	technician	married	tertiary	no	1109	no	no	cellular	13	aug	328	2	-1	0	unknown	no
20	student	single	secondary	no	502	no	no	cellular	30	apr	261	1	-1	0	unknown	yes
31	blue-collar	married	secondary	no	360	yes	yes	cellular	29	jan	89	1	241	1	failure	no
40	management	married	tertiary	no	194	no	yes	cellular	29	aug	189	2	-1	0	unknown	no
56	technician	married	secondary	no	4073	no	no	cellular	27	aug	239	5	-1	0	unknown	no
37	admin.	single	tertiary	no	2317	yes	no	cellular	20	apr	114	1	152	2	failure	no
25	blue-collar	single	primary	no	-221	yes	no	unknown	23	may	250	1	-1	0	unknown	no
31	services	married	secondary	no	132	no	no	cellular	7	jul	148	1	152	1	other	no
38	management	divorced	unknown	no	0	yes	no	cellular	18	nov	96	2	-1	0	unknown	no
42	management	divorced	tertiary	no	16	no	no	cellular	19	nov	140	3	-1	0	unknown	no
44	services	single	secondary	no	106	no	no	unknown	12	jun	109	2	-1	0	unknown	no
44	entrepreneur	married	secondary	no	93	no	no	cellular	7	jul	125	2	-1	0	unknown	no

- Source: UCI Machine Learning
- Bank clients data
- No NAs
- 45k observations, 16 covariates
- 80% train data, 20% test

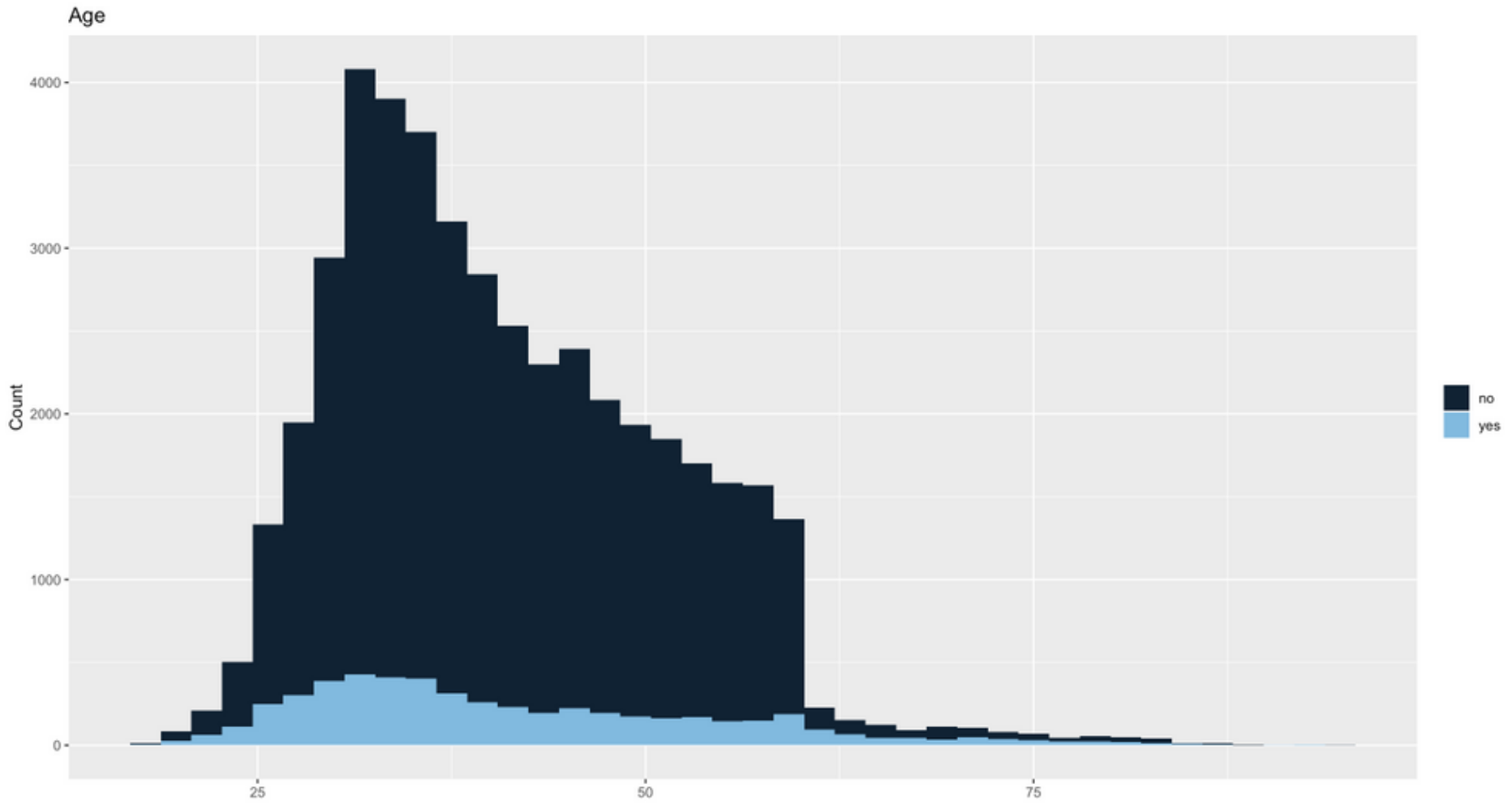
Section 2

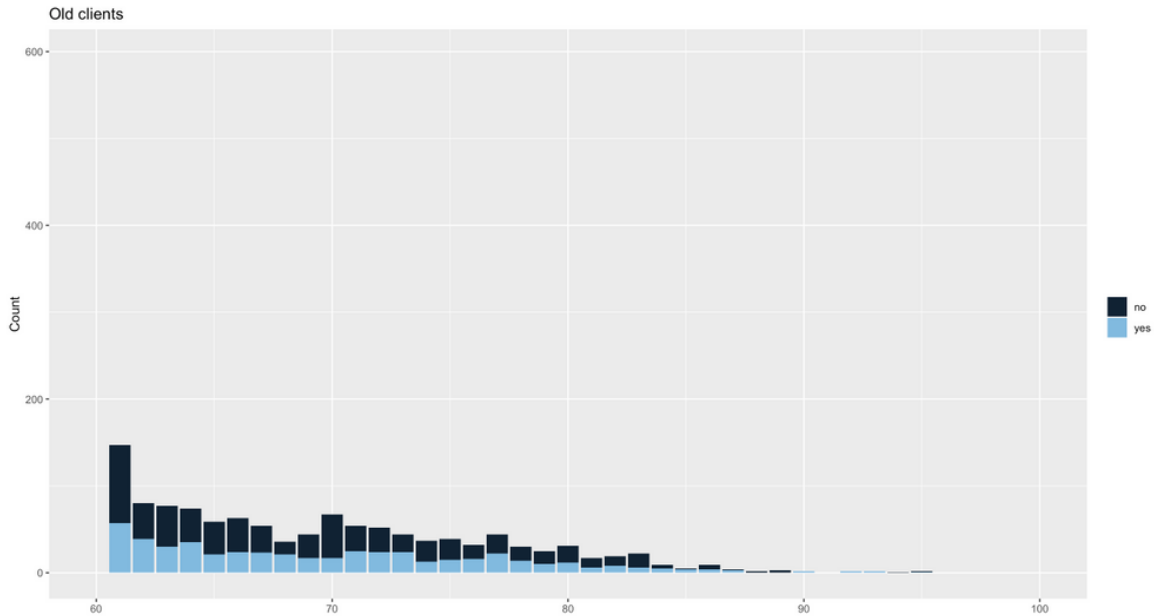
EDA

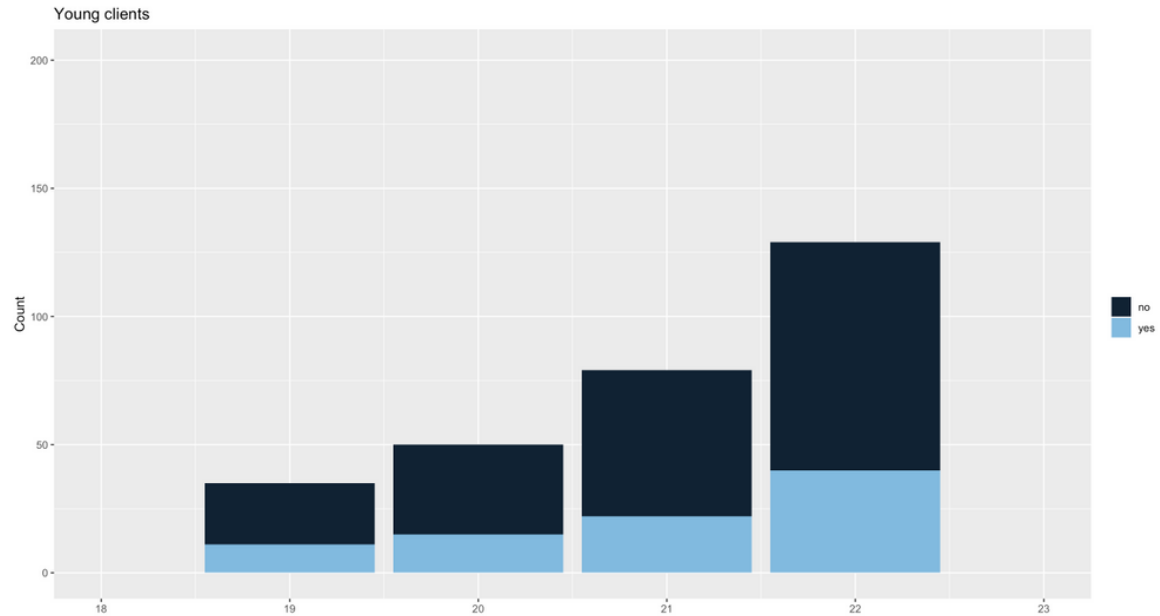
Correlation matrix



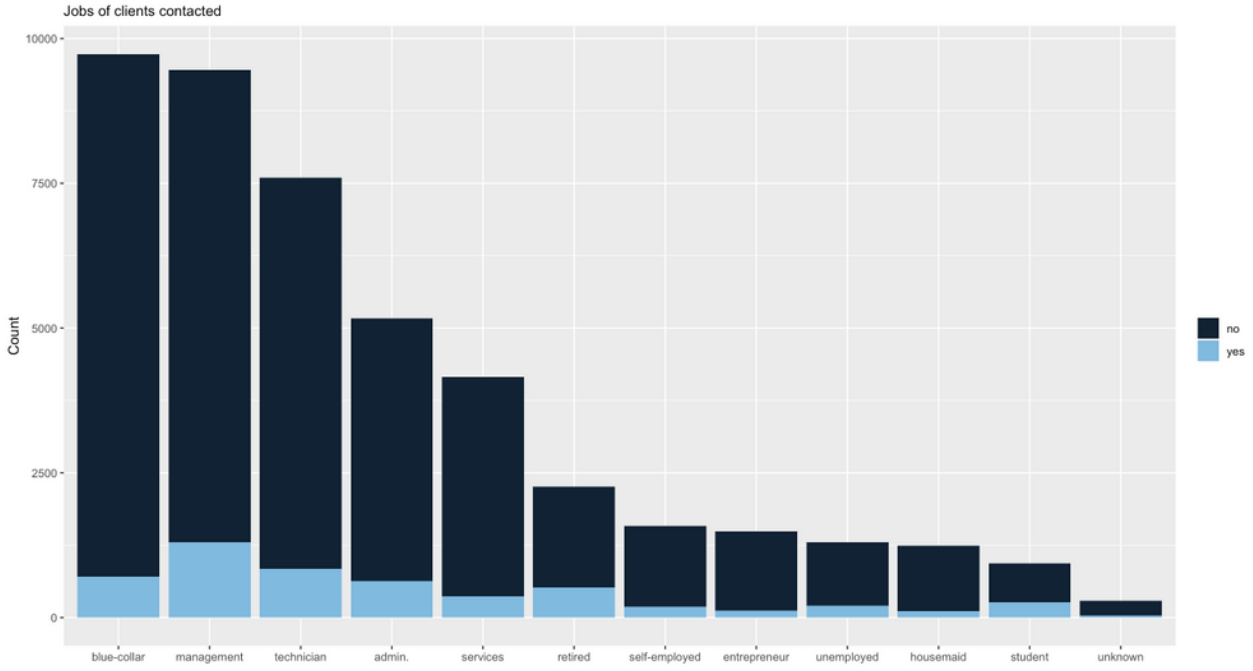
Age variable




Old clients
Count
600
400
200
0
60
70
80
90
100
no
yes

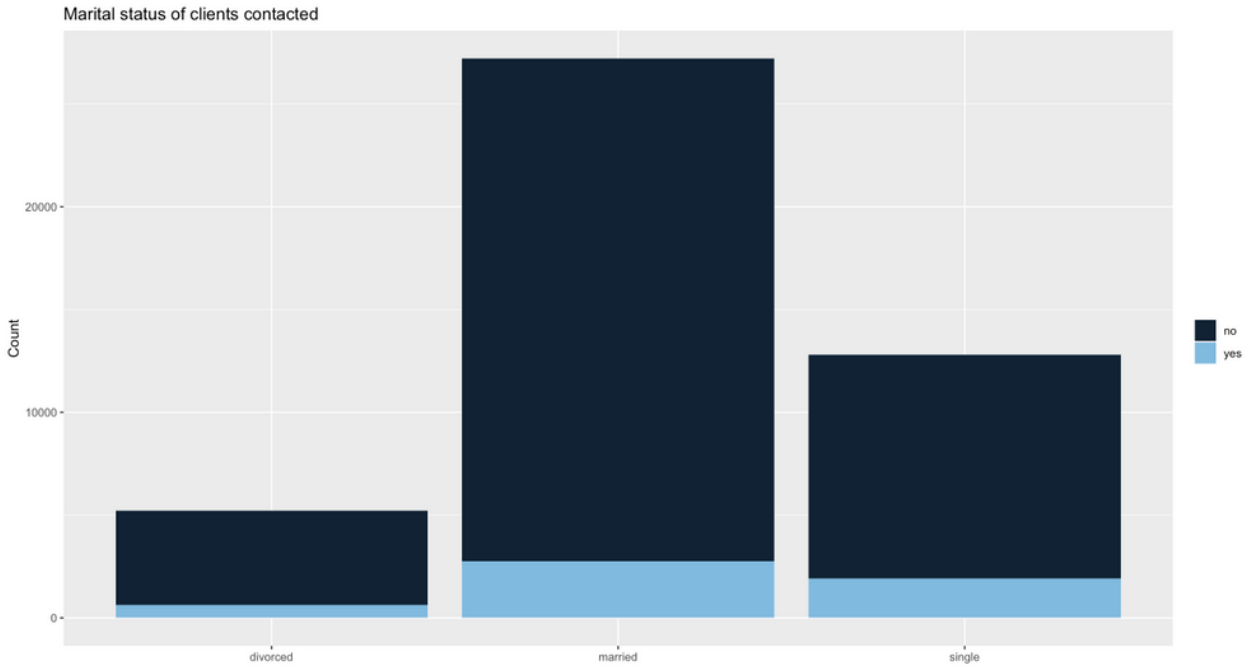

Young clients
Count
200
150
100
50
0
18
19
20
21
22
23
no
yes

Job variable



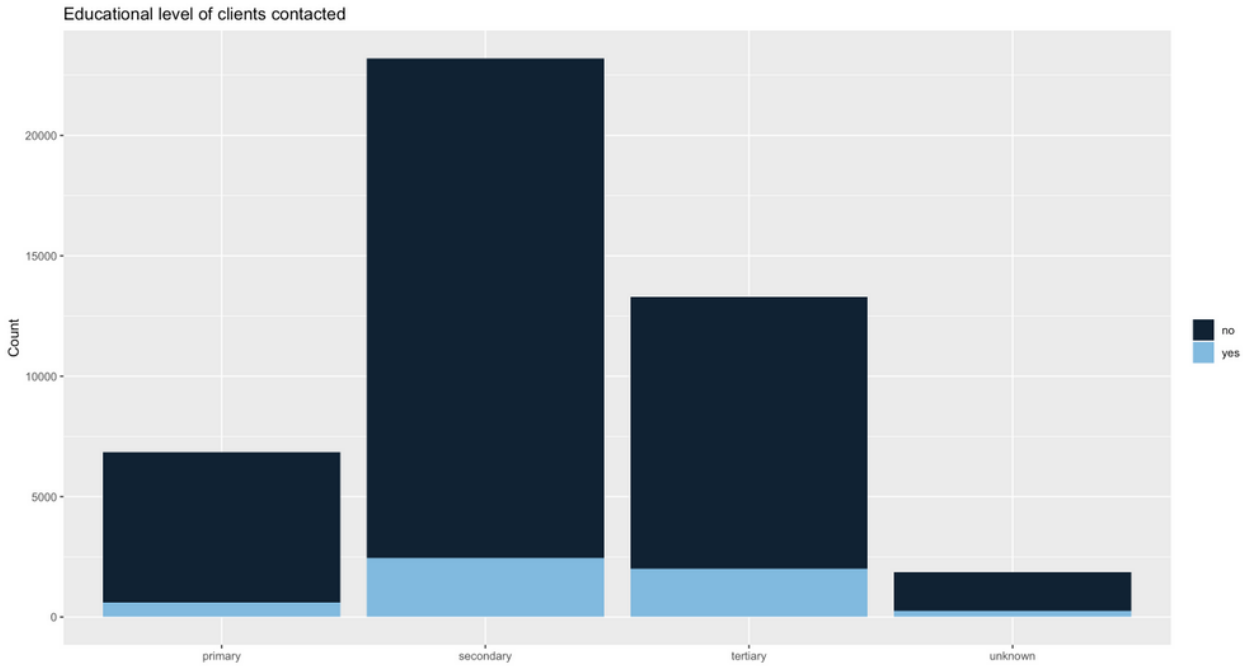
- Blue collar new deposit: 13%
- Management new deposit: 24%
- Technician new deposit: 15%
- Student new deposit: 5%. But with percentage of success 28%.

Marital variable



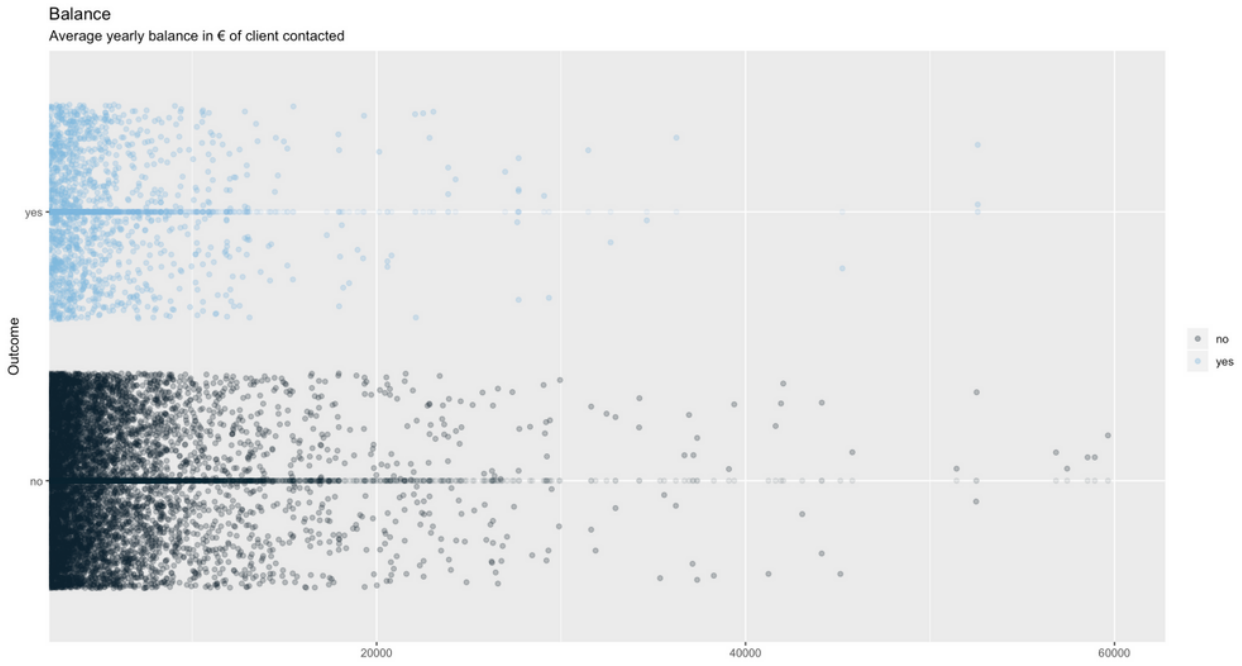
- About 52% of clients that opened a new deposit is married, 12% results divorced and 36% single.
- Percentage of success from single is 15%, for divorced is 12% and for married is 10%.

Education variable



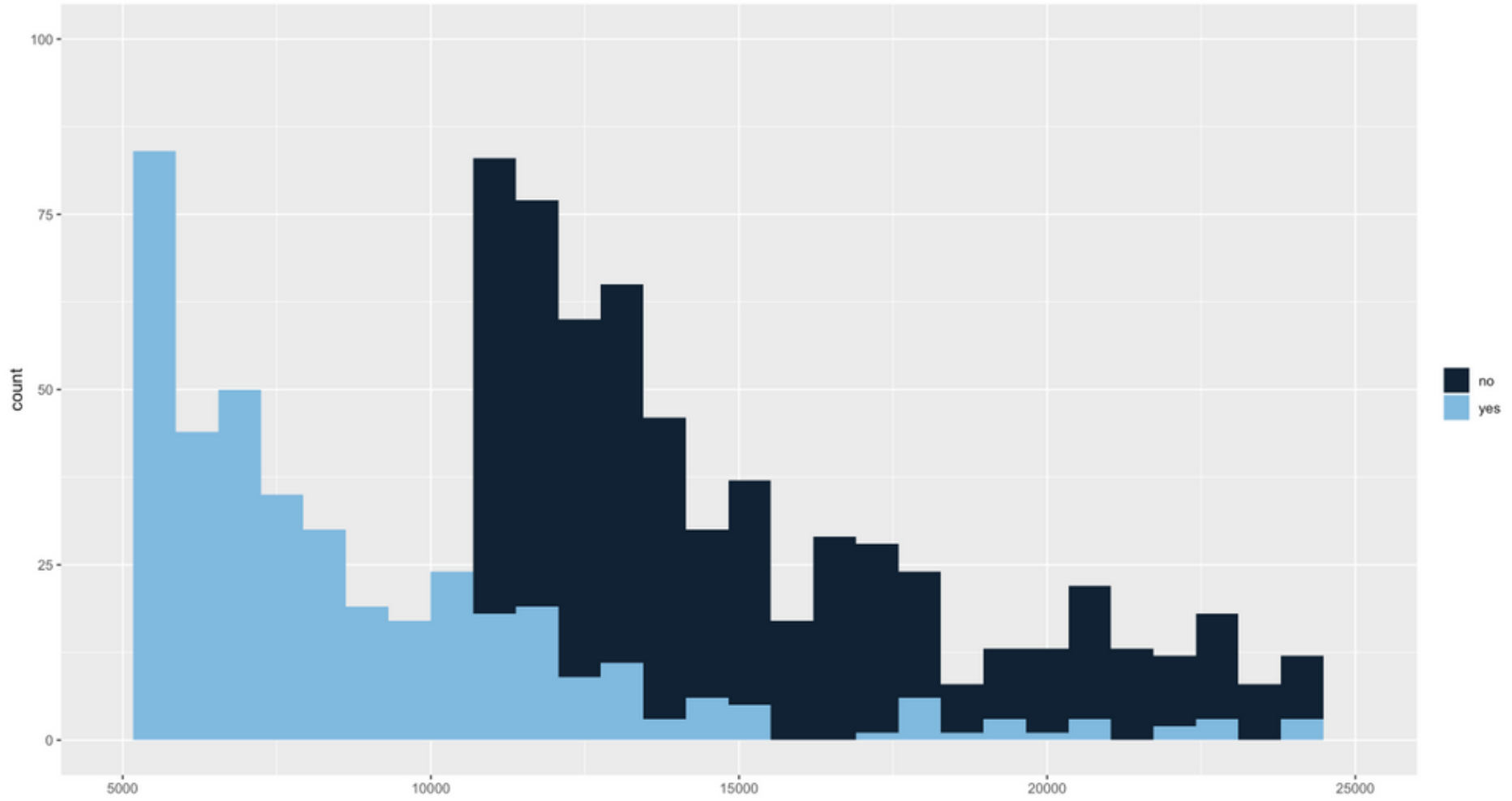
- New deposit from secondary: 44%
- New deposit from tertiary: 38%
- New deposit from primary: 11%
- New deposit from unknown: 5%
- The highest percentage of success is from tertiary and unknown respectively 18% and 15%

Balance variable

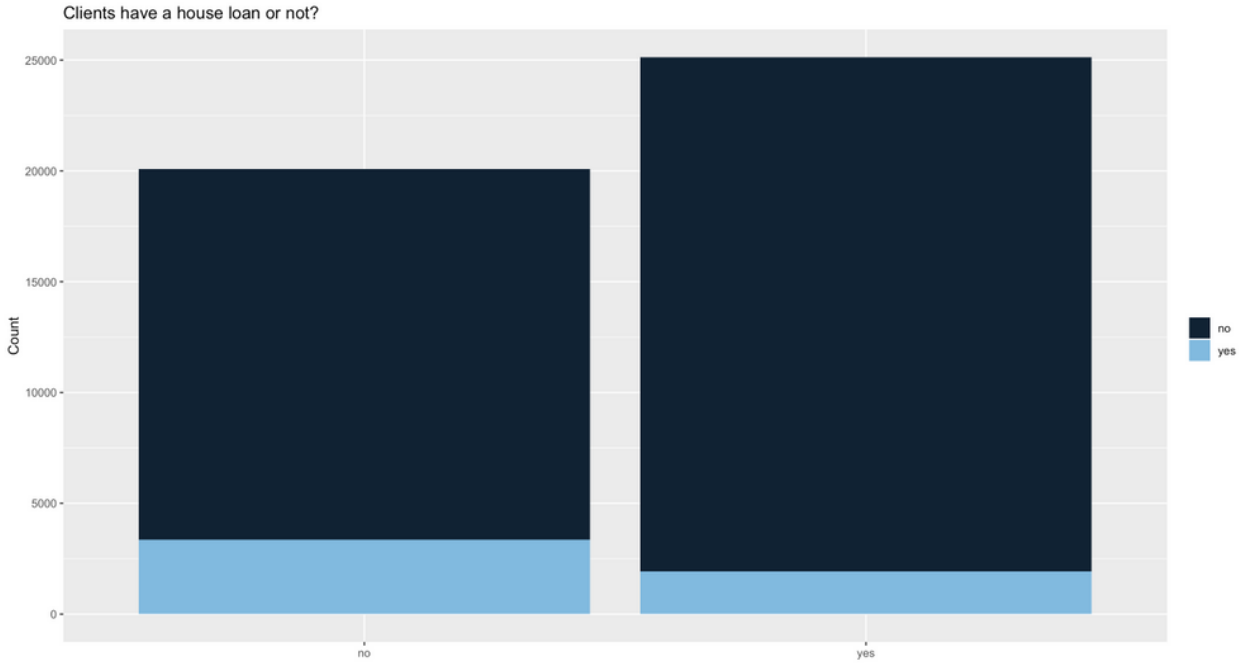


- Most of the time the balance of people who opened a new long term deposit results lower than the one of who didn't

Low balance clients

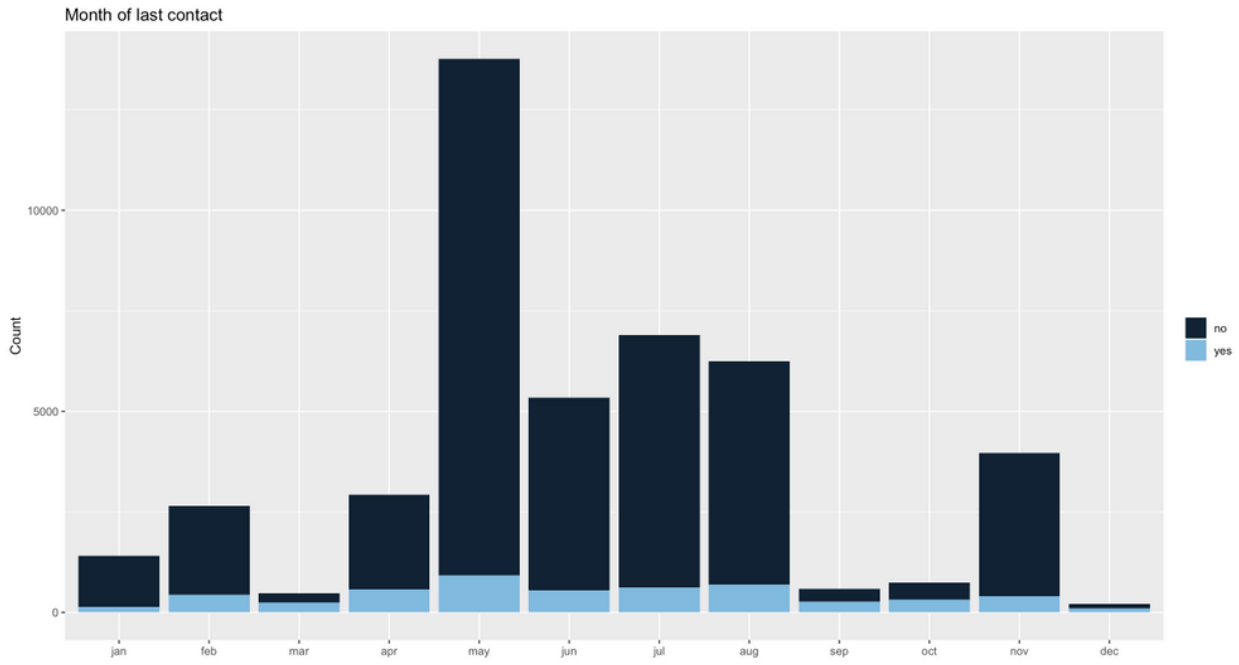


Housing variable



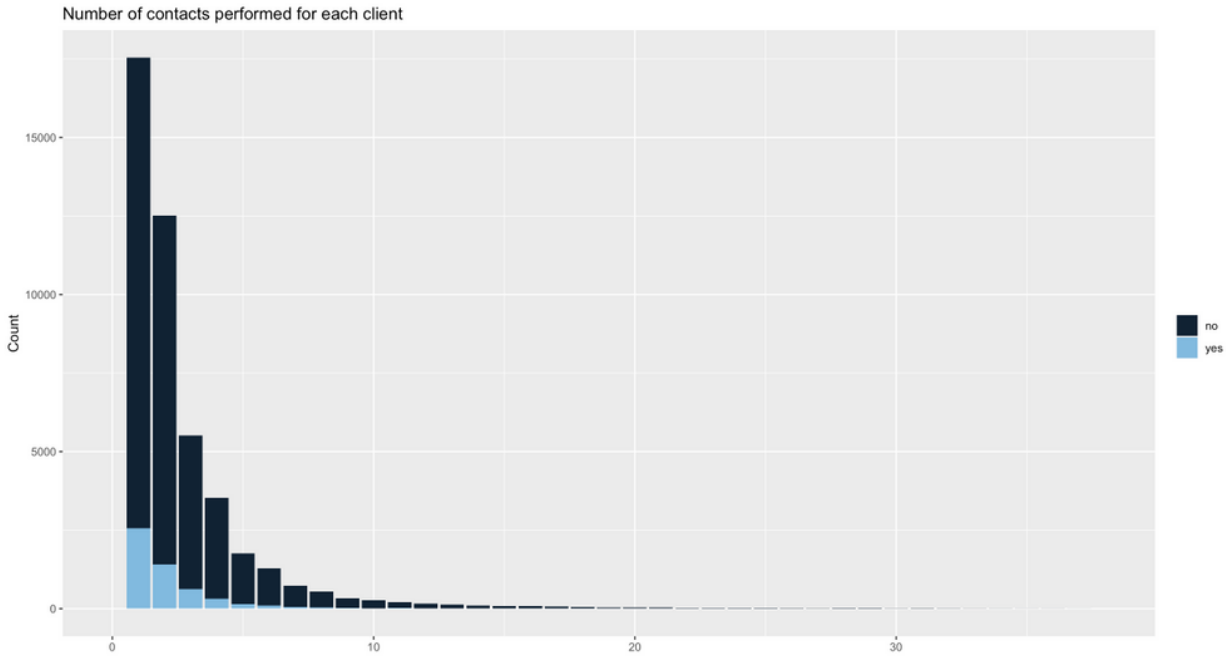
- New deposit from people without house loan: 63%
- New deposit from people with house loan: 37%

Month variable



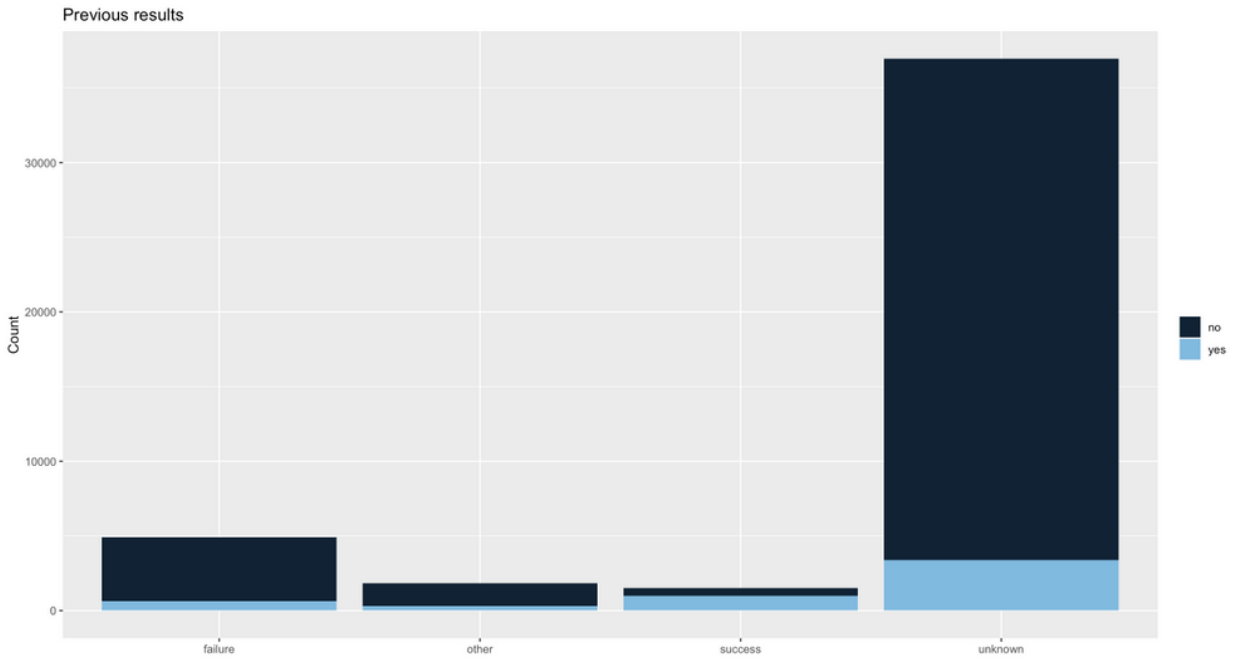
- In the last month of each (almost) trimester the percentage of success is larger
- Percentage of success in March: 52%
- Percentage of success in September: 46%
- Percentage of success in December: 46%

Campaign variable



- It show the number of contacts performed during this campaign for a specific client

Poutcome variable



- It indicates the outcome of previous marketing campaing

Other variables

- **Loan**

  Categorical variable which tells if a client has a personal loan or not

- **Day**

  Numerical variable indicating the last contact day of the month

- **Duration**

  Numerical variable that returns the last contact duration in seconds

- **Pdays**

  Numerical variable which indicate the number of days that passed by after the client was last contacted from a previous campaign

- **Previous**

  Numerical variable that refers to the number of contacts performed before this campaign and for each client

- **Contact**

  Categorical variable referring to the type of contact given by the client

Section 3

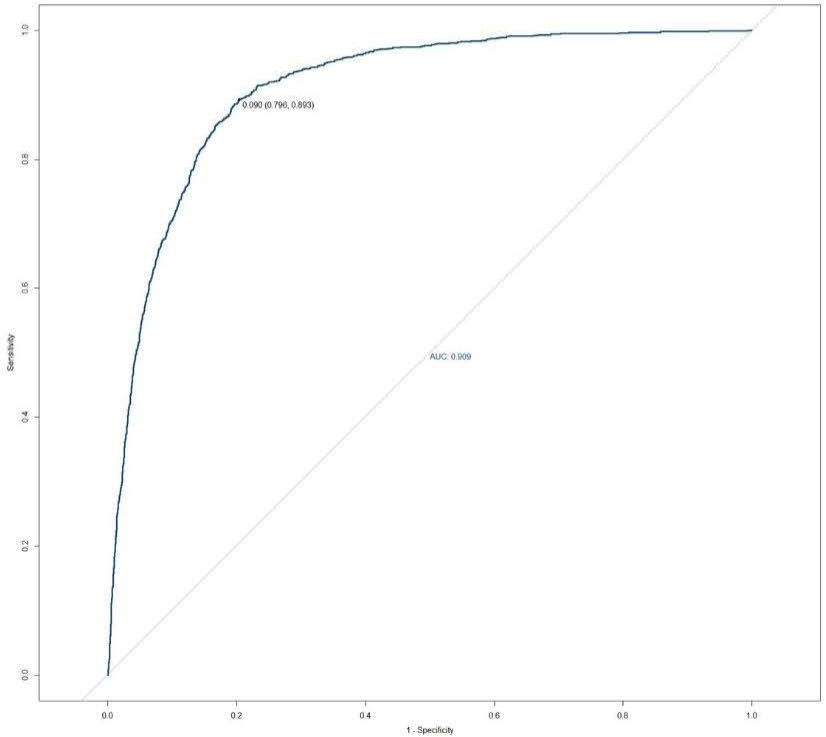
Data Analysis

DATA ANALYSIS

1. Logistic Regression
2. Decision Trees
3. Support Vector Machines
4. Gradient Boosting

Logistic Regression

Logistic Regression



Model Fitting

Optimal threshold: 0.09

Specificity: 0.79

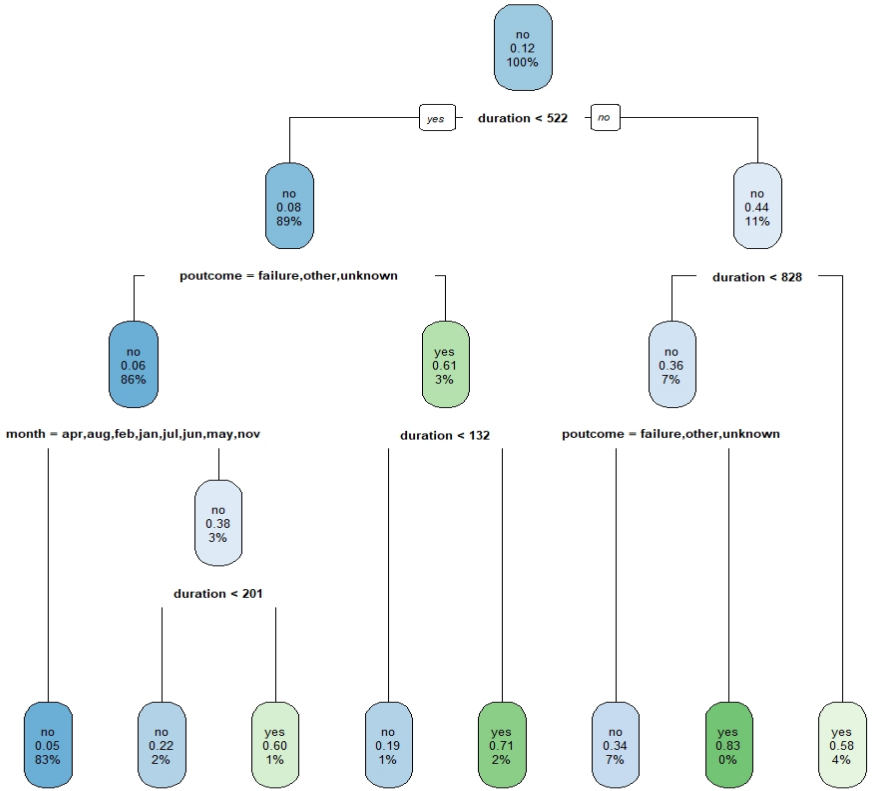
Sensitivity: 0.89

Area under the curve: 0.909

	Actual Values	
LOGISTIC	NO	YES
NO	6372	111
YES	1632	928

Decision Trees

Decision Trees



Model Fitting

Optimal threshold: 0.08

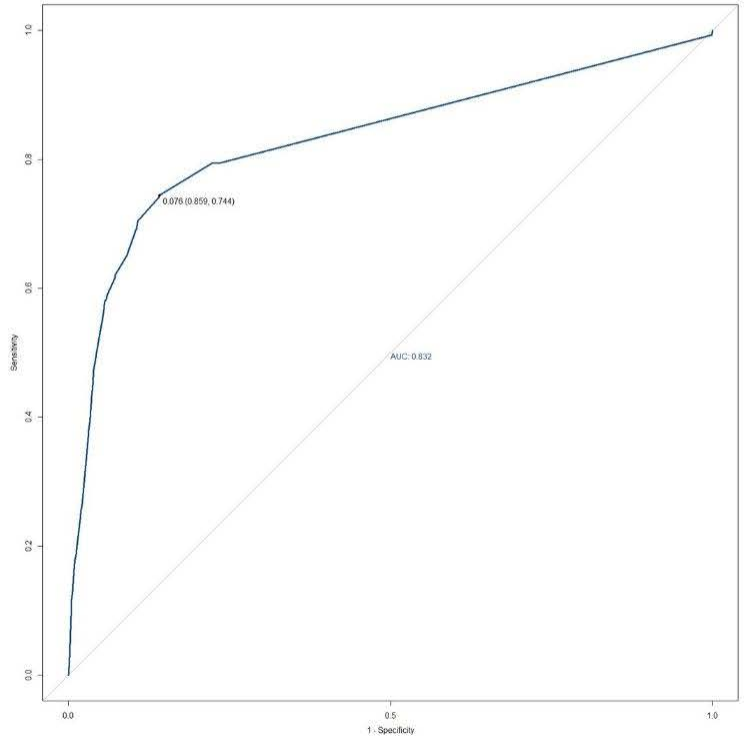
Specificity: 0.86

Sensitivity: 0.74

Area under the curve: 0.832

DT	Actual Values NO	YES
NO	6873	266
YES	1131	773

Decision Trees



Model Fitting

Optimal threshold: 0.08

Specificity: 0.86

Sensitivity: 0.74

Area under the curve: 0.832

		Actual Values
DT	NO	YES
NO	6873	266
YES	1131	773

Support Vector Machines

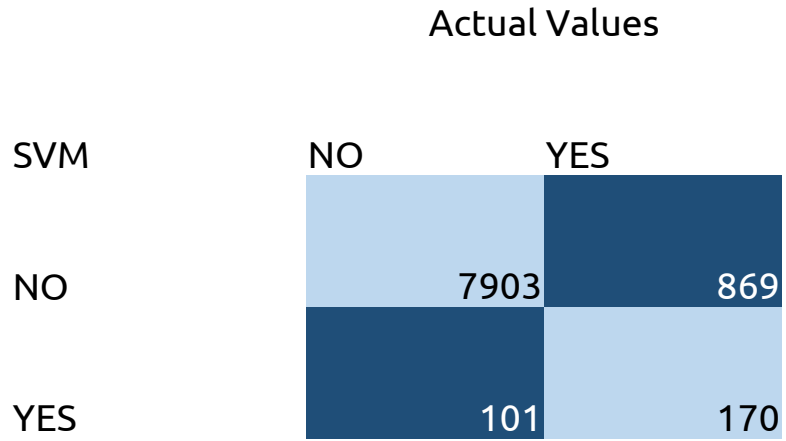
Support Vector Machines

Model Fitting

Specificity: 0.98

Sensitivity: 0.16

Area under the curve: 0.58



Gradient Boosting

Tree-based algorithms over time

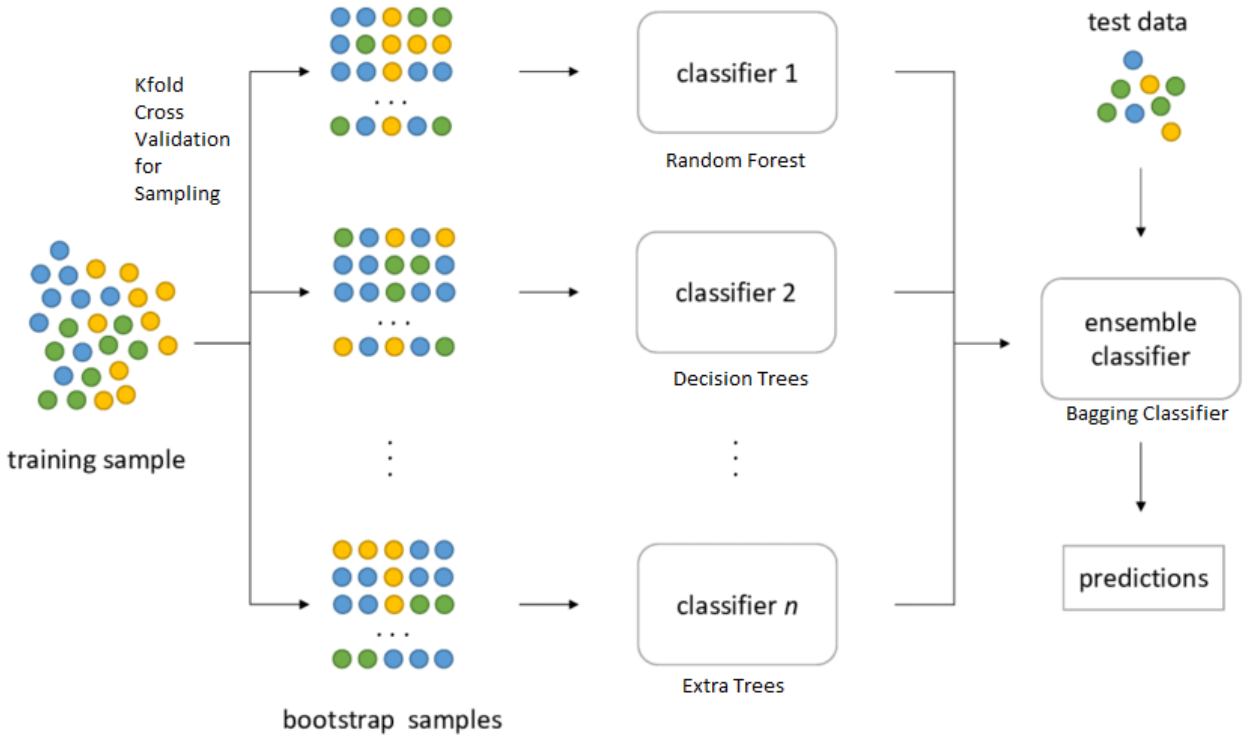
Decision Trees

Supervised learning algorithms based on stratifying or segmenting the predictors space into homogeneous subgroups.

Bagging

Bootstrap Aggregation, is an ensemble algorithm combining predictions from multiple decision tress.

Bagging



Bagging Classifier Process Flow

Bagging is used when the goal is to reduce variance. The idea is to create several subsets of data from training samples chosen randomly. Each collection of subset data is used to train the decision trees. As a result, we end up with an ensemble of different models.

Tree-based algorithms over time

Decision Trees

Supervised learning algorithms based on stratifying or segmenting the predictors space into homogeneous subgroups.

Bagging

Bootstrap Aggregation, is an ensemble algorithm combining predictions from multiple decision tress.

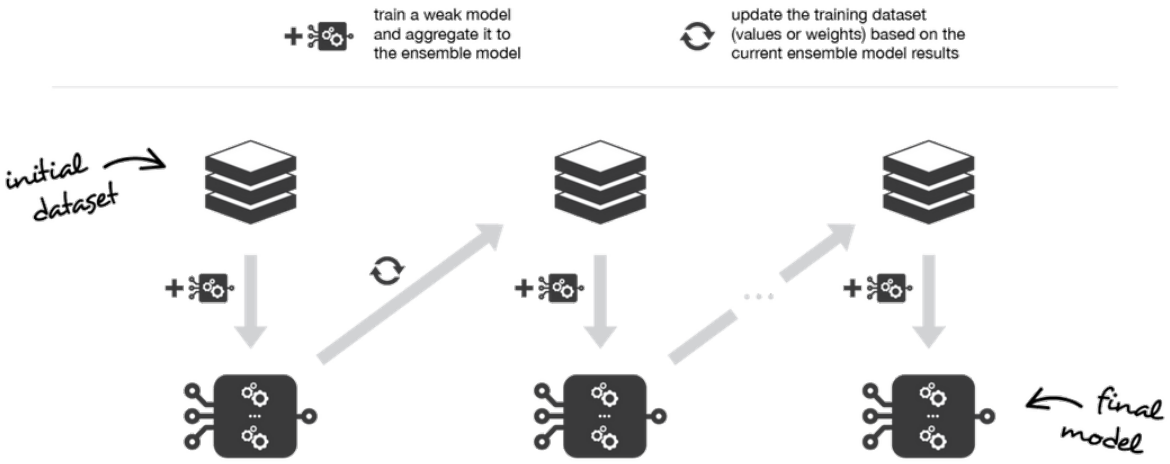
Random Forest

Bagging based algorithm where only a subset of features are selected at random to build a collection of decision trees.

Boosting

Models are built sequentially by minimizing the errors from previous models while increasing influence of high performing models

Boosting



Boosting is another ensemble technique to create a collection of models. In this technique, models are learned sequentially with early models fitting simple models to the data and then analyzing the data for errors. Recall that bagging had each model run independently and then aggregate the outputs at the end without preference to any model. In other words, with boosting, we fit consecutive trees and at every step. The goal is to solve for the net error from the prior tree.

Tree-based algorithms over time

Decision Trees

Supervised learning algorithms based on stratifying or segmenting the predictors space into homogeneous subgroups.

Bagging

Bootstrap Aggregation, is an ensemble algorithm combining predictions from multiple decision tress.

Random Forest

Bagging based algorithm where only a subset of features are selected at random to build a collection of decision trees.

Boosting

Models are built sequentially by minimizing the errors from previous models while increasing influence of high performing models

Gradient Boosting

Gradient boosting employs gradient descent algorithm to minimize errors in sequential models.

XGBoost

Optimized Gradient Boosting algorithm.

XGBoost

Definition

XGBoost is a decision-tree based ensemble machine learning algorithm that uses a gradient boosting framework.

How does it work?

As other GBM it involves three main steps:

- The first step that is required is that a loss function has to be optimized.

- The second step is the use of a weak learner, in our case a decision tree. Often, weak learners can be constrained using a maximum number of layers, nodes, splits or leaf nodes.

- The third step is combing many weak learners in an additive fashion. Decision trees are added one at a time. A gradient descent procedure is used to minimize the loss when adding trees. That's the gradient part of gradient boosters.

What is different?

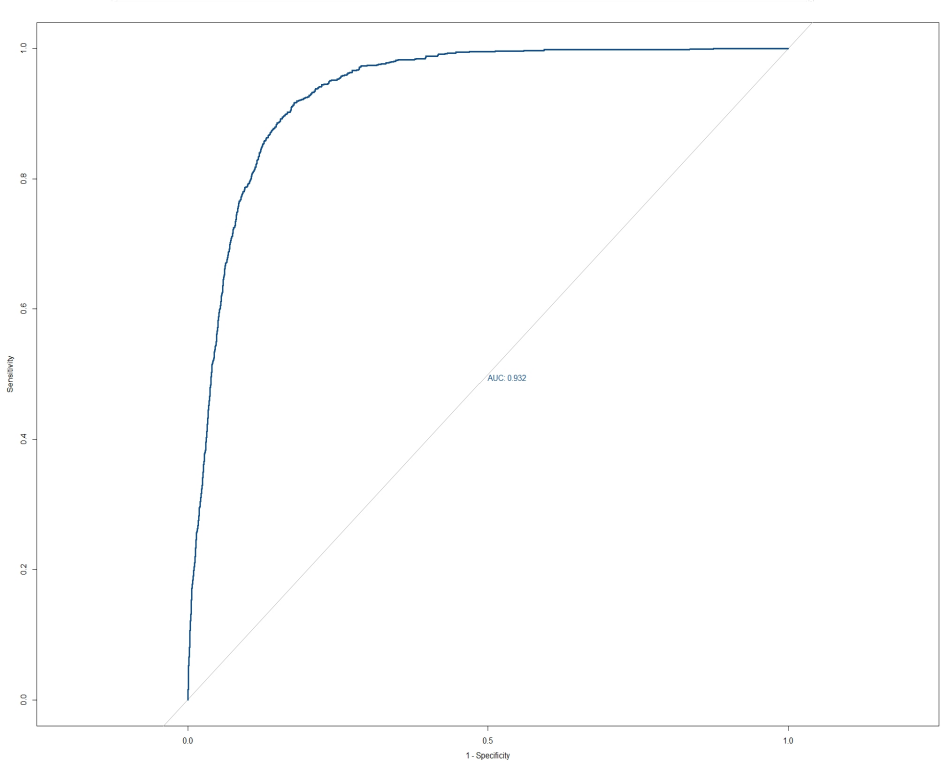
- A regularization term is added to the loss function to penalize complexity of the model.The additional regularization term helps to smooth the final learnt weights to avoid over-fitting.

- Shrinkage scales newly added weights by a factor η after each step of tree boosting. Similar to a learning rate in stochastic optimization, shrinkage reduces the influence of each individual tree and leaves space for future trees to improve the model.

- Others technical features for speeding up computation.

$$\mathrm{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f)_k$$

XGBoost



Model Fitting

Optimal threshold: 0.125

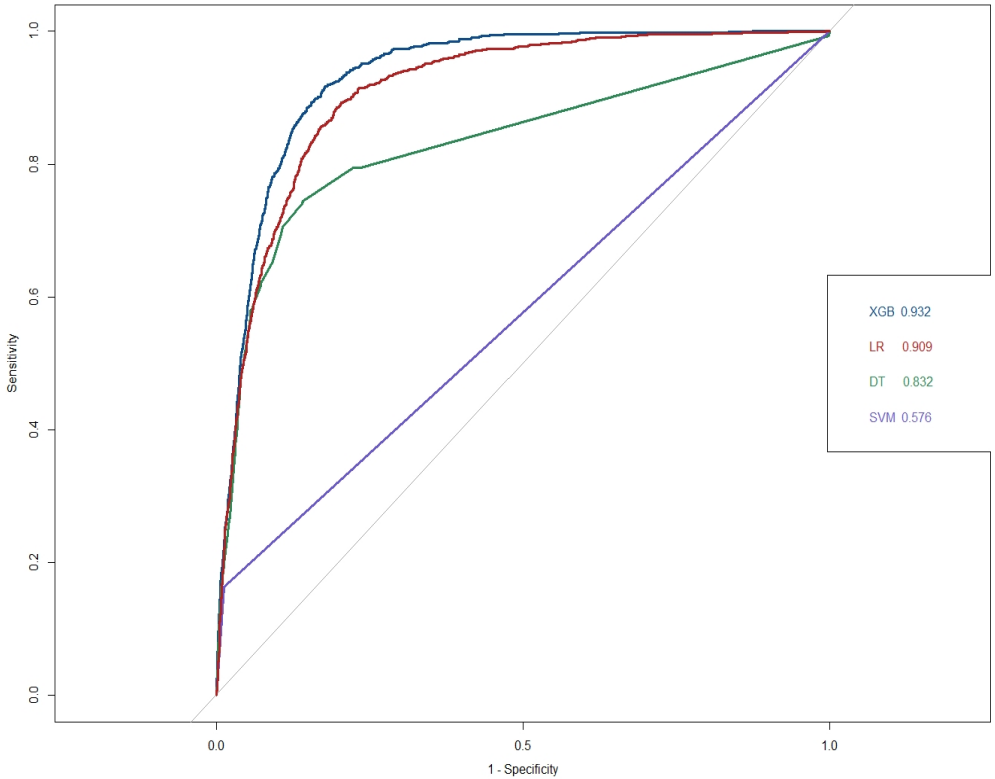
Specificity: 0.85

Sensitivity: 0.89

Area under the curve: 0.932

		Actual Values
XGBOOST	NO	YES
NO	6873	126
YES	1167	913

Model Comparison



Evaluation

As expected the Xgboost model is the one that performs better. It is followed from the logistic regression which despite being a much simpler model, it has a good performance and leaves room for interpretation.

Conclusions and future applications

- Based on AUC, Xgboost is the best model, but if we consider the trade-off between complexity and interpretability of the models the choice would be the logistic regression. Indeed, the AUC are marginally different but the second allows also interpretation and inference on the variables.

- We advise Caixa Bank to continue collecting this data since it has been proven that they're effective in terms of both prediction and inference. The dataset could be expanded by adding macroeconomics data; such as inflation, employment variation rate and three-month Euribor.

- Since this kind of client's data is quite standard in banks, we expect to be able to extend the use of this model easily also to other financial institutions.

Sources

- L. Breiman, J.H. Friedman, R.A. Olshen, , and C.J Stone. Classification and Regression Trees. Wadsworth, Belmont, Ca, 1983.
- G. James, D. Witten, T. Hastie and R. Tibshirani, An Introduction to Statistical Learning, Springer, London, 2017.
- Medium

  https://medium.com/ml-research-lab/bagging-ensemble-meta-algorithm-for-reducing-variance-c98fffa5489f

  https://medium.com/@gabrieltseng/gradient-boosting-and-xgboost-c306c1bcfaf5
- Quora

  https://www.quora.com/What-is-an-intuitive-explanation-of-Gradient-Boosting
- S. Moro, R. Laureano, P. Cortez, Using data mining for bank direct marketing:an application of the crisp-dm methodology, Lisbon, 2011
- Tianqi Chen, Carlos Guestrin, XGBoost: A Scalable Tree Boosting System, University of Washington, 2016
- Towards Data Science

  https://towardsdatascience.com/boosting-algorithm-gbm-97737c63daa3

  https://towardsdatascience.com/decision-trees-and-random-forests-df0c3123f991
- Uci Machine Learning repository

  https://archive.ics.uci.edu/ml/datasets/Bank+Marketing