Albert School | 20/11/2025

SUPERVISED MACHINE LEARNING

by Alexandre Waerniers & Vincent Lamy

NAME OF PROJECT:
Bank Marketing Campaign

PRESENTED BY:
Alexandre Waerniers
Vincent Lamy

PRESENTED TO:
Guillaume Desforges

Table of Contents

Context

Real Life

Portuguese **banking** institution.
Direct **marketing campaign** - phone calls.
Goal is to offer a banking contract - **long term deposit**.
Donated on 2/13/2012 [1]

Machine Learning

Predict if the client **will subscribe** a term deposit.
Classification - yes/no

[1] https://archive.ics.uci.edu/dataset/222/bank+marketing

Dataset

Provided by

UCI Machine Learning Repository

Features include

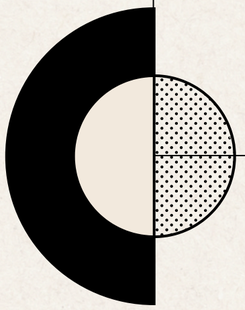
Client information

Current campaign

Previous campaign

Socioeconomic context

Content	Values
Observations	41 188
Input features	20
Target variable	yes/no
Temporal ordering	May 2008 - November 2010



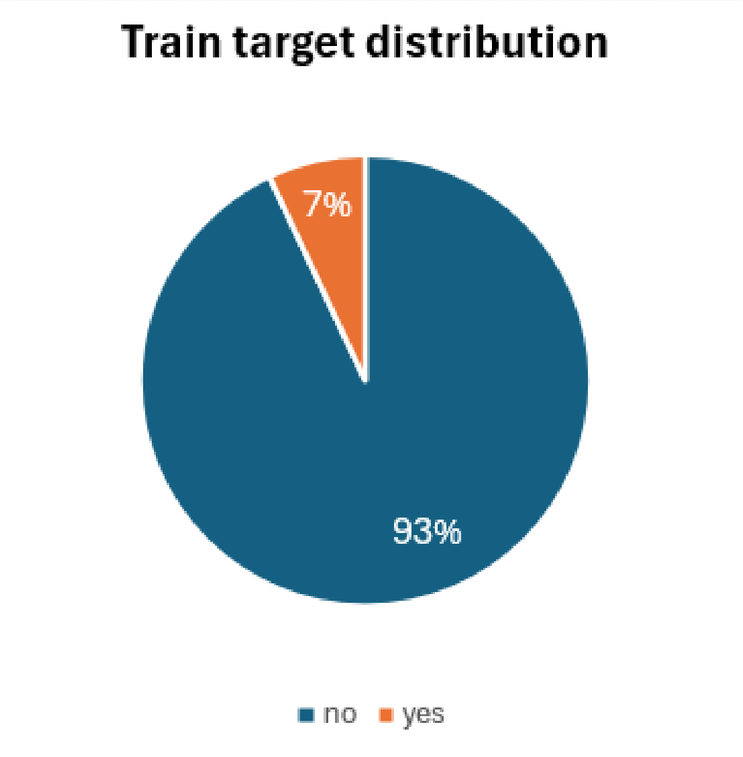
Dataset - Numerical Features

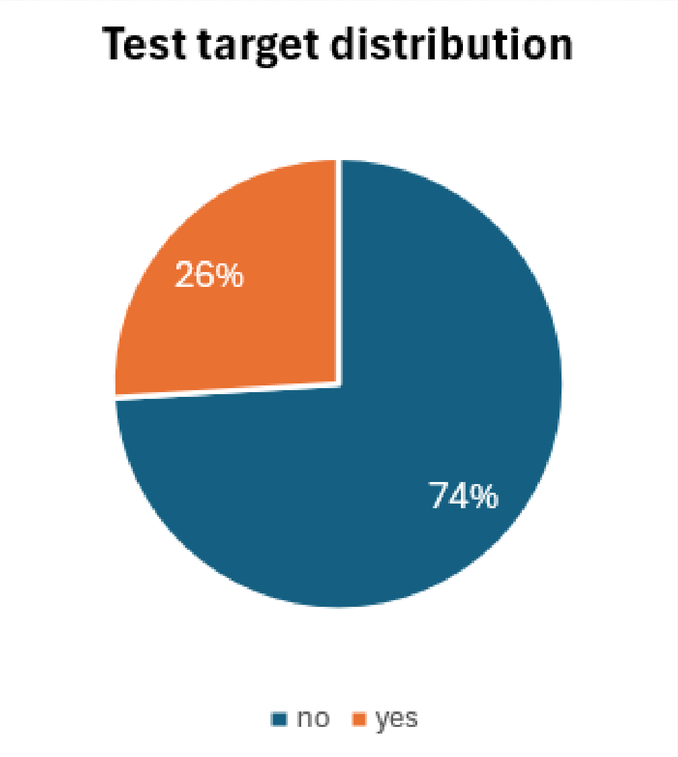
Features	Meaning
age	client age
duration	last phone call duration, in seconds
campaign	number of contacts performed during this campaign
previous	number of contacts performed before this campaign
pdays	number of days that passed by after the client was last contacted from a previous campaign
emp.var.rate	employment variation rate - quarterly indicator
cons.price.idx	consumer price index - monthly indicator
cons.conf.idx	consumer confidence index - monthly indicator
euribor3m	euribor 3 month rate - daily indicator
nr.employed	number of employees - quarterly indicator

Dataset - Categorical Features

Features	Meaning
job	client type of job
marital	marital status
education	educational status
default	has credit card default ?
housing	has housing loan ?
loan	has personal loan ?
contact	contact communication type
month	last contact month
day_of_week	last contact day of the week
poutcome	outcome of the previous marketing campaign

EDA - target distribution





Train

Target distribution strongly imbalanced

93% no ; 7% yes

Test

Target distribution less imbalanced

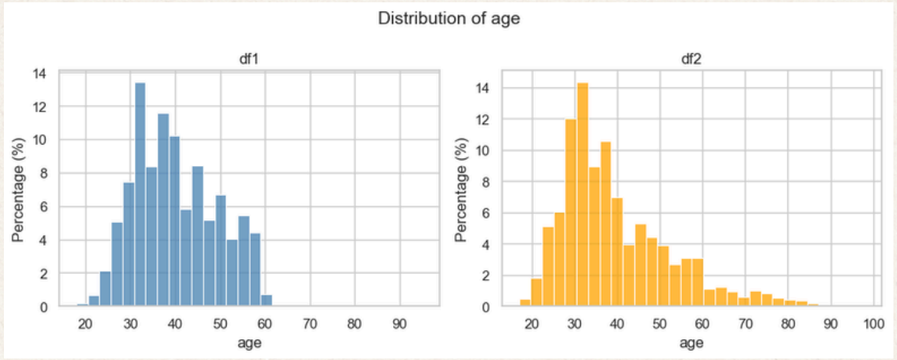
Split

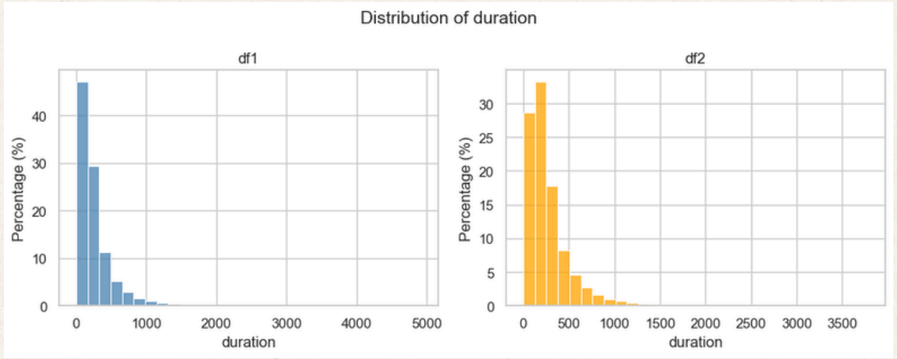
Train size 75%

Test szie 25%

Best compromise

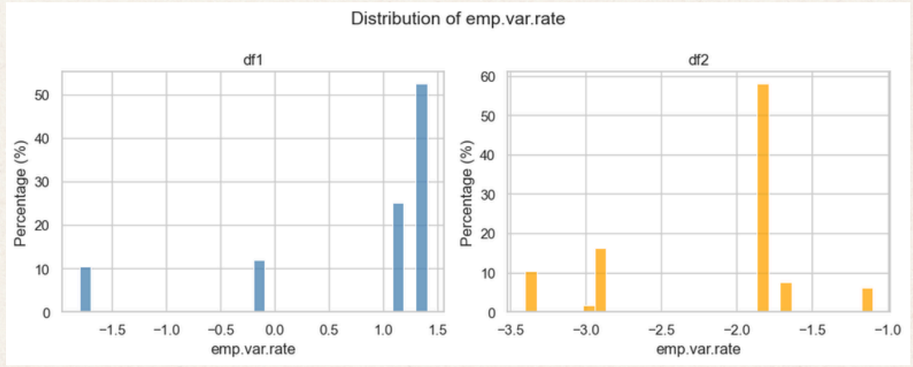
EDA - Numerical features distributions

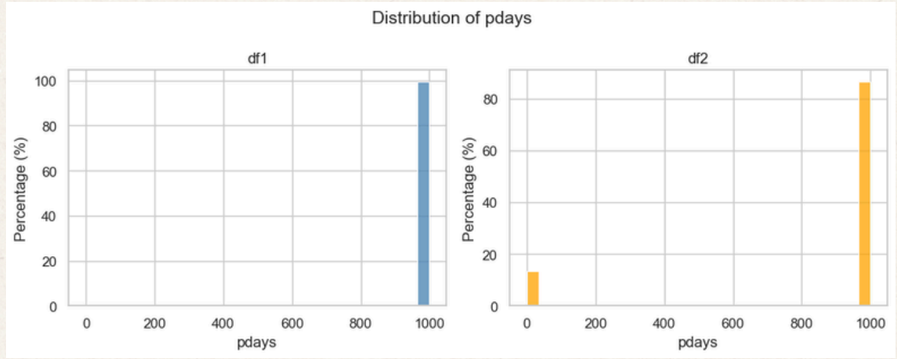




Observations

Similar distributions

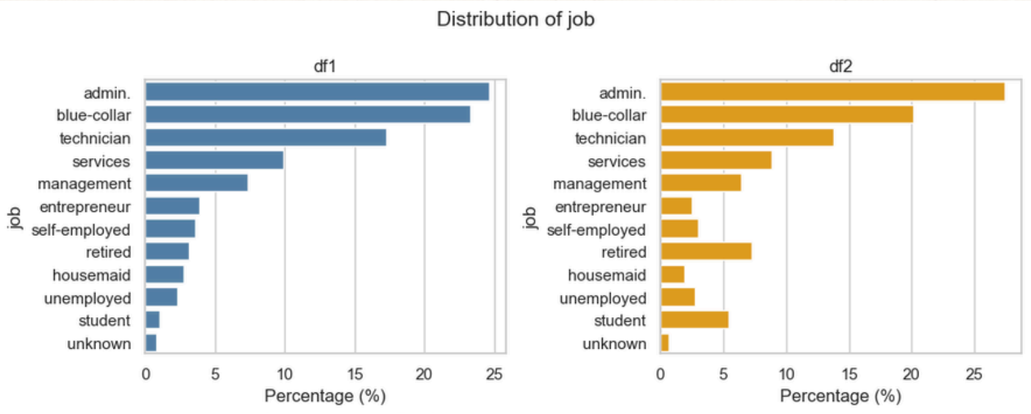




Observations

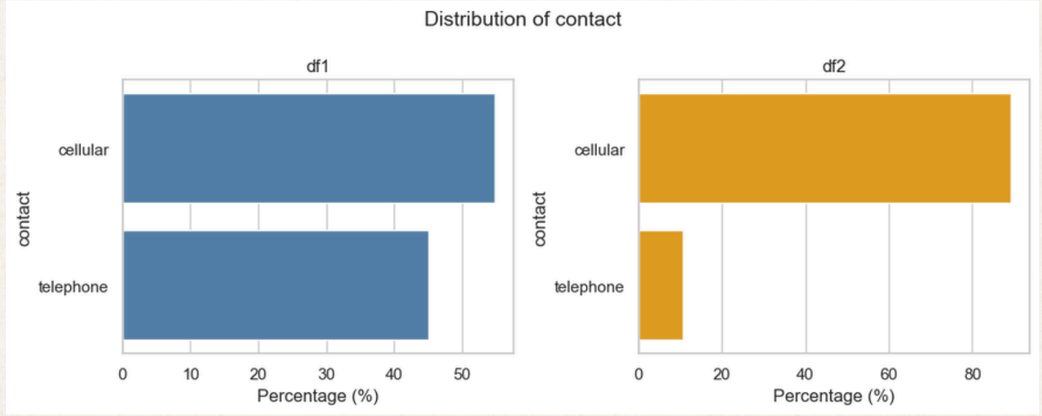
Different distributions

EDA - Categorical features distributions



Observations

Almost identical distributions



Observations

Different distributions

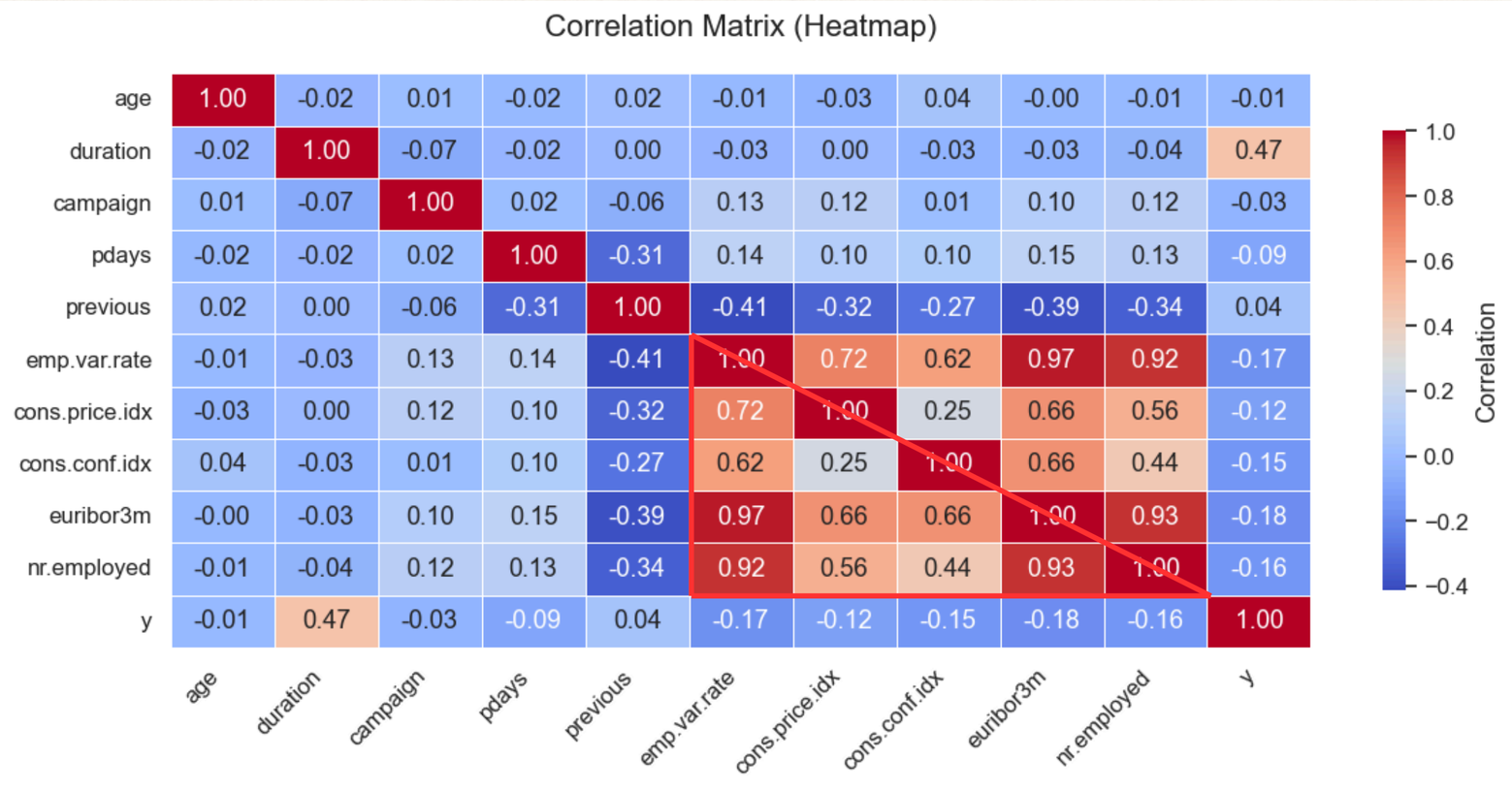
Train

55% cellular ; 45% telephone

Test

90% cellular : 10% telephone

EDA - Correlation Heatmap



Parameter

Pearson - linear relationship

Observations

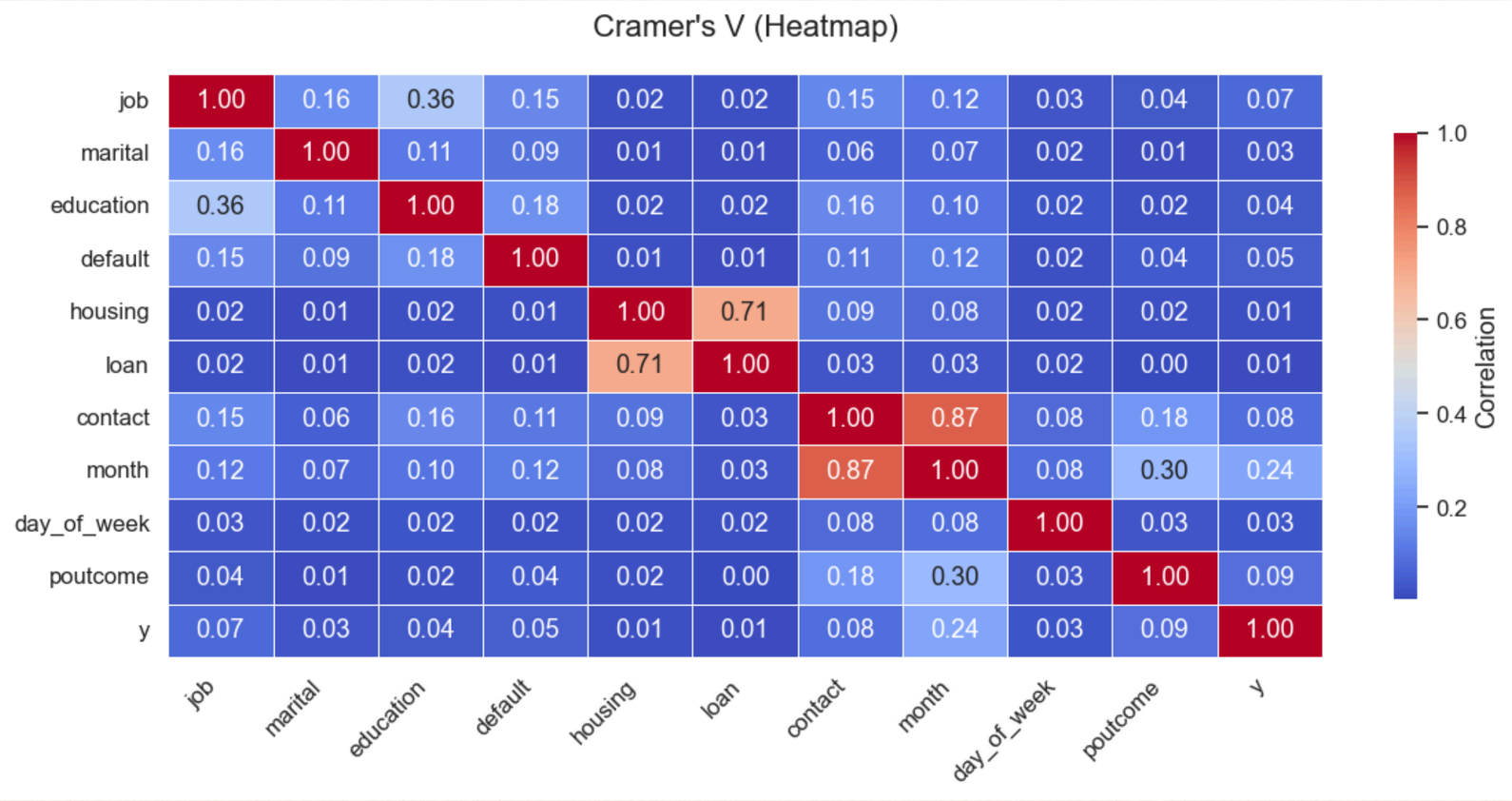
Socioeconomic cluster growth/recession

Correlation to target

duration but data leakage

Nothing else

EDA - Cramer's V



Observations

education/job

loan/housing

month/contact

month/poutcome

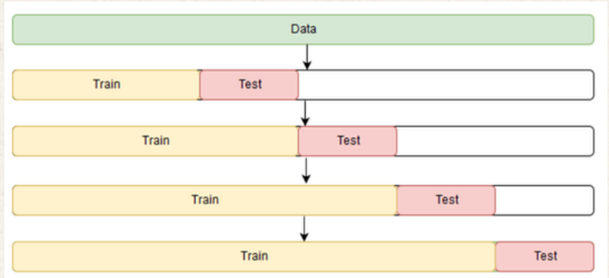
Correlation to target

month → contact

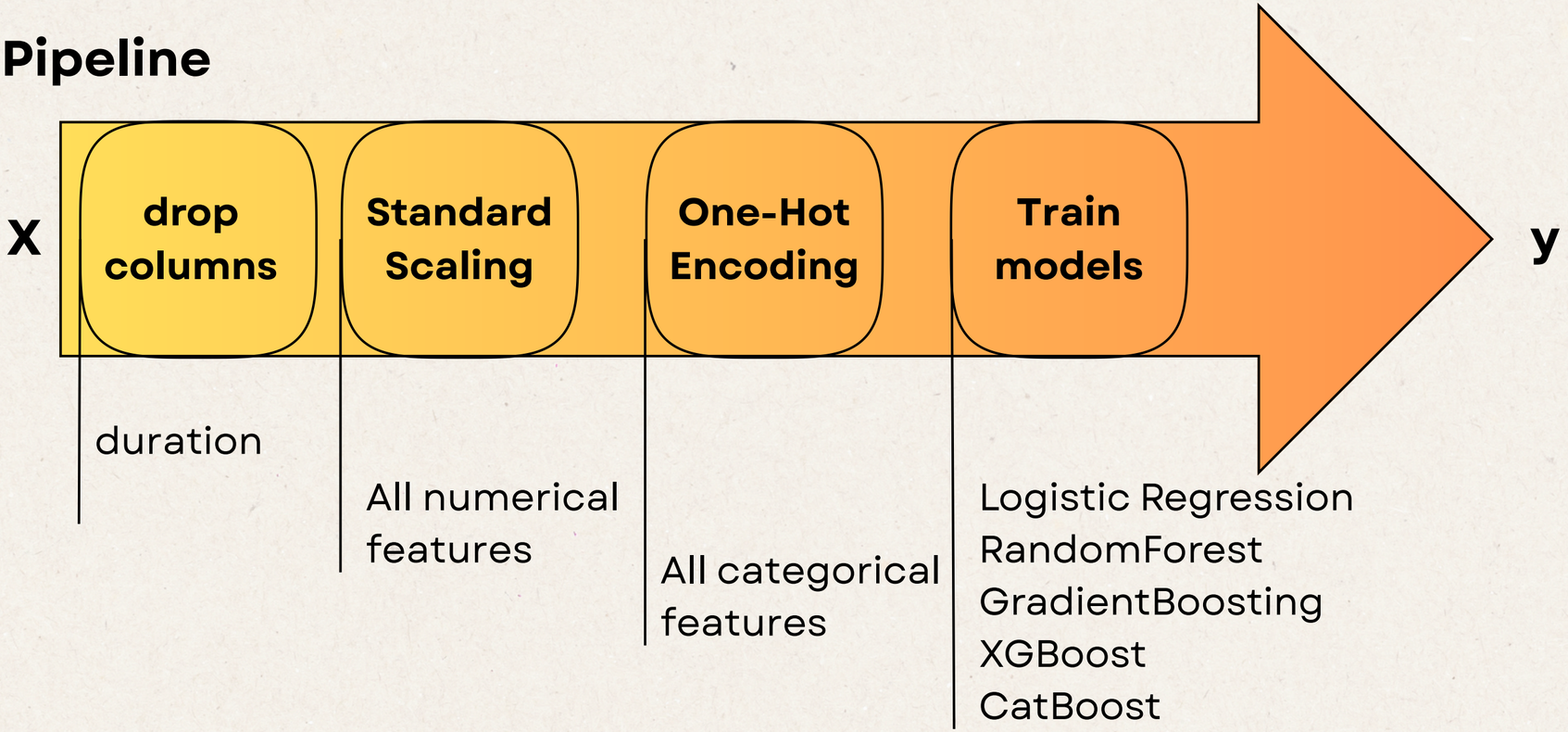
Train - Baseline

Validation

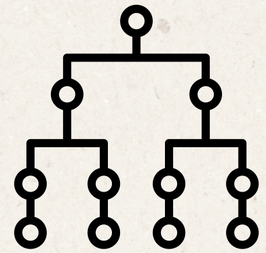
Time Series
Cross Validation



Pipeline



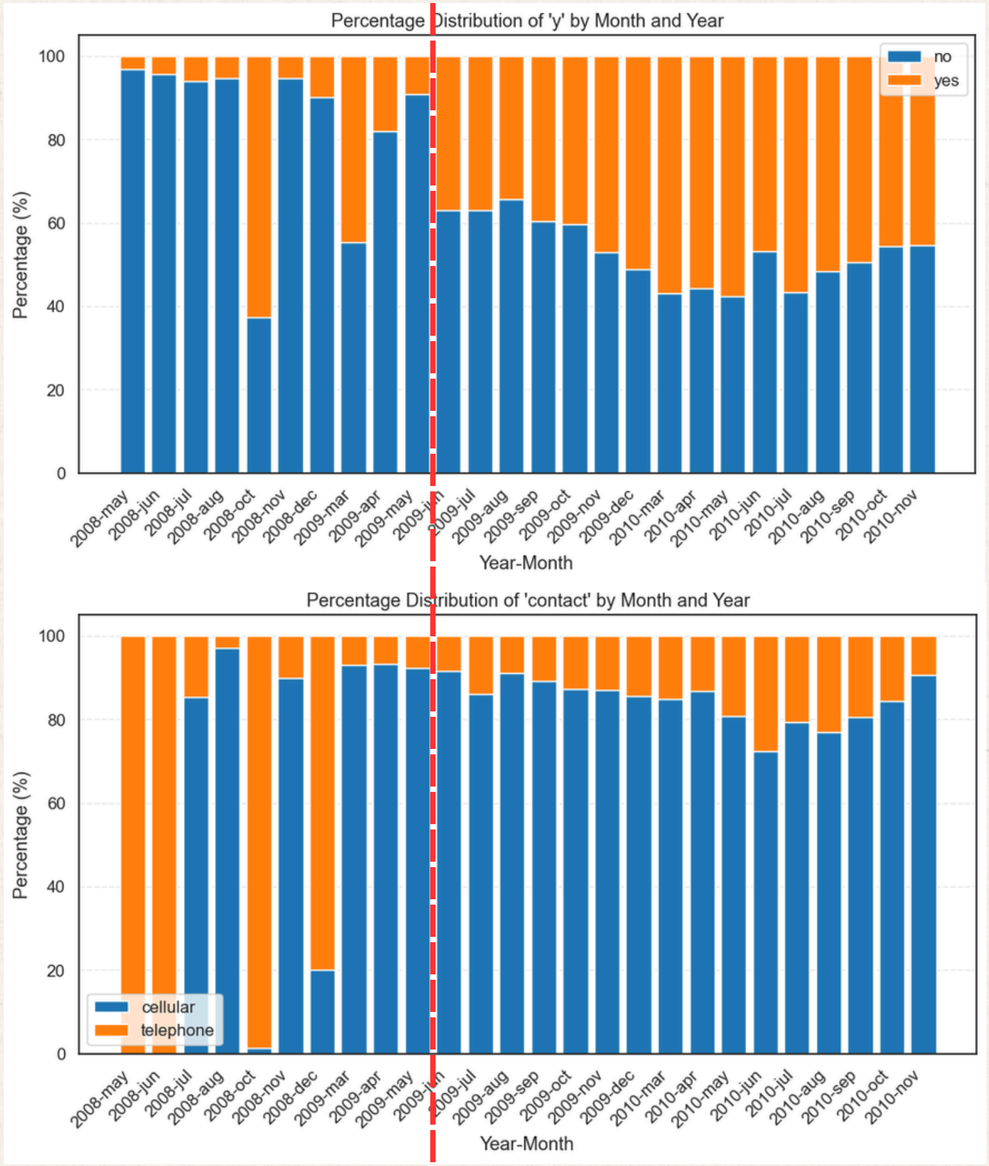
Best model:
RandomForest



Metrics

Accuracy = 0.55
Precision = 0.57
Recall = 0.60
F1-score= 0.58

EDA - Understand the data



Context

Post subprime crisis
Financial uncertainty

Smartphone boom
Between 2008 and 2009
People switched to cellular phones

New Split
Keep data from June 2008 to November 2010

Train (80%)
target distribution : 55% / 45%

Test (20%)
target distribution : 52% / 48%

EDA - Population Stability Index

$$PSI = \sum_{i=1}^{n} \left((P_i - Q_i) * \ln\left(\frac{P_i}{Q_i}\right) \right)$$

where:

Pi = Proportion in bin i for the expected (training) population

Qi = Proportion in bin i for the actual (testing) population

	Feature	Psi	Psi_cut
0	age	0.352378	0.057033
1	duration	0.022452	0.083758
2	campaign	0.081058	0.010943
3	pdays	0.461901	0.096170
4	previous	0.940028	0.274494
5	emp.var.rate	8.727344	7.411689
6	cons.price.idx	10.530461	7.417164
7	cons.conf.idx	10.678185	10.634205
8	euribor3m	7.361996	7.892721
9	nr.employed	7.978768	7.411689
10	job	0.136014	0.030247
11	marital	0.060851	0.012201
12	education	0.009382	0.019768
13	default	0.164107	0.003178
14	housing	0.008237	0.001421
15	loan	0.000008	0.004277
16	contact	0.663215	0.021500
17	month	1.305064	3.589962
18	day_of_week	0.001828	0.021608
19	poutcome	0.956514	0.115783
20	y	0.315080	0.018680

Pros

Train and test dataset are much more similar

Better suited for generalization

Still relevant of the current situation

Cons

examples down to **4964**

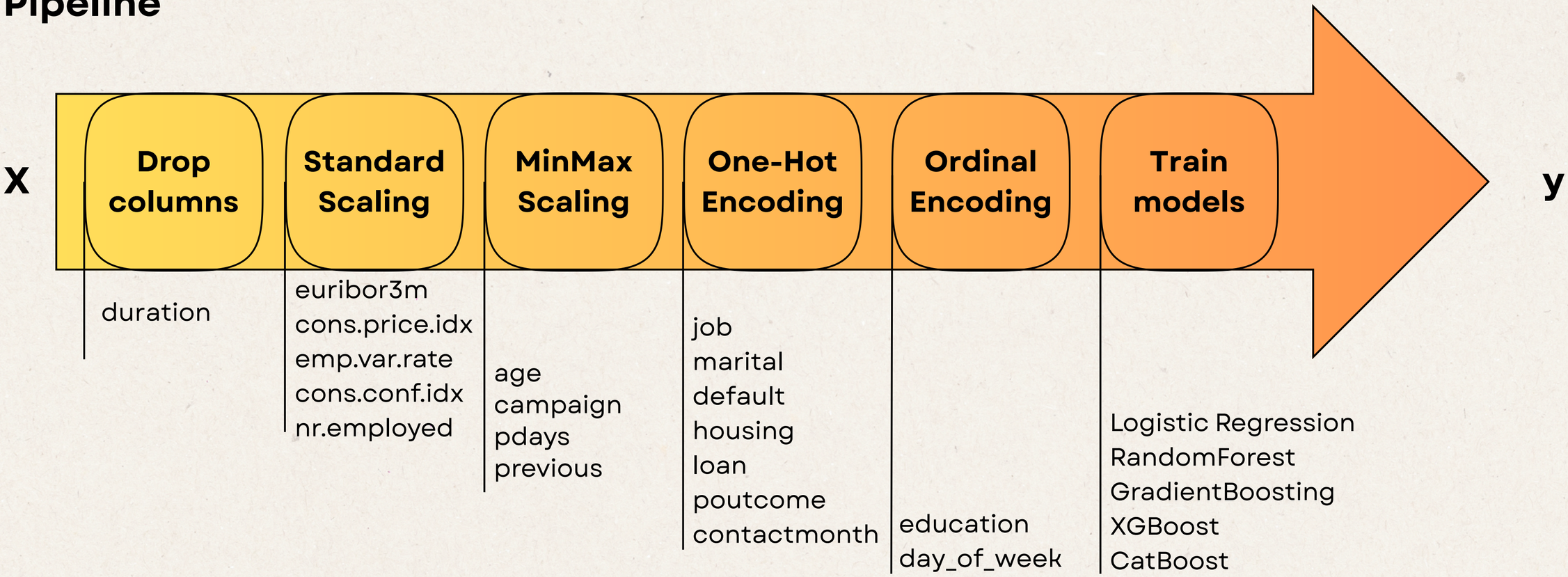
Represents 12% of original dataset

Train - Upgraded

Validation

Time Series Cross Validation

Pipeline



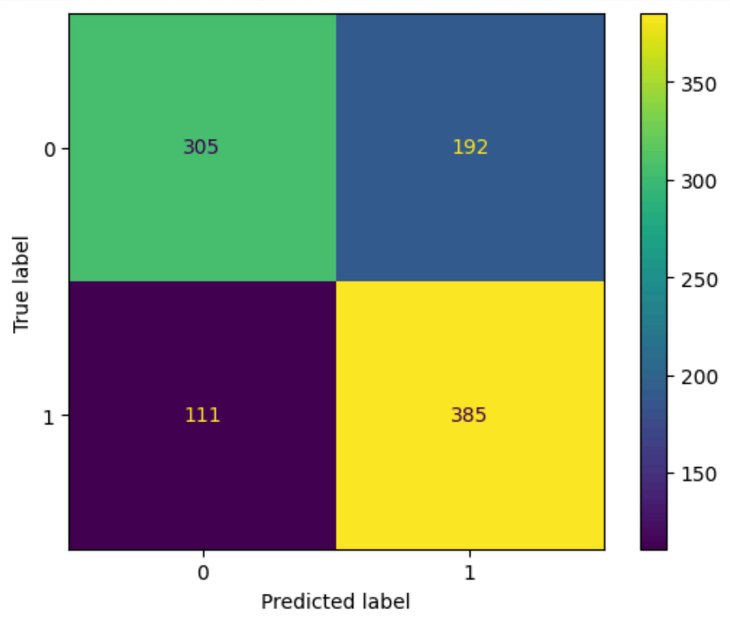
Test

Best Model

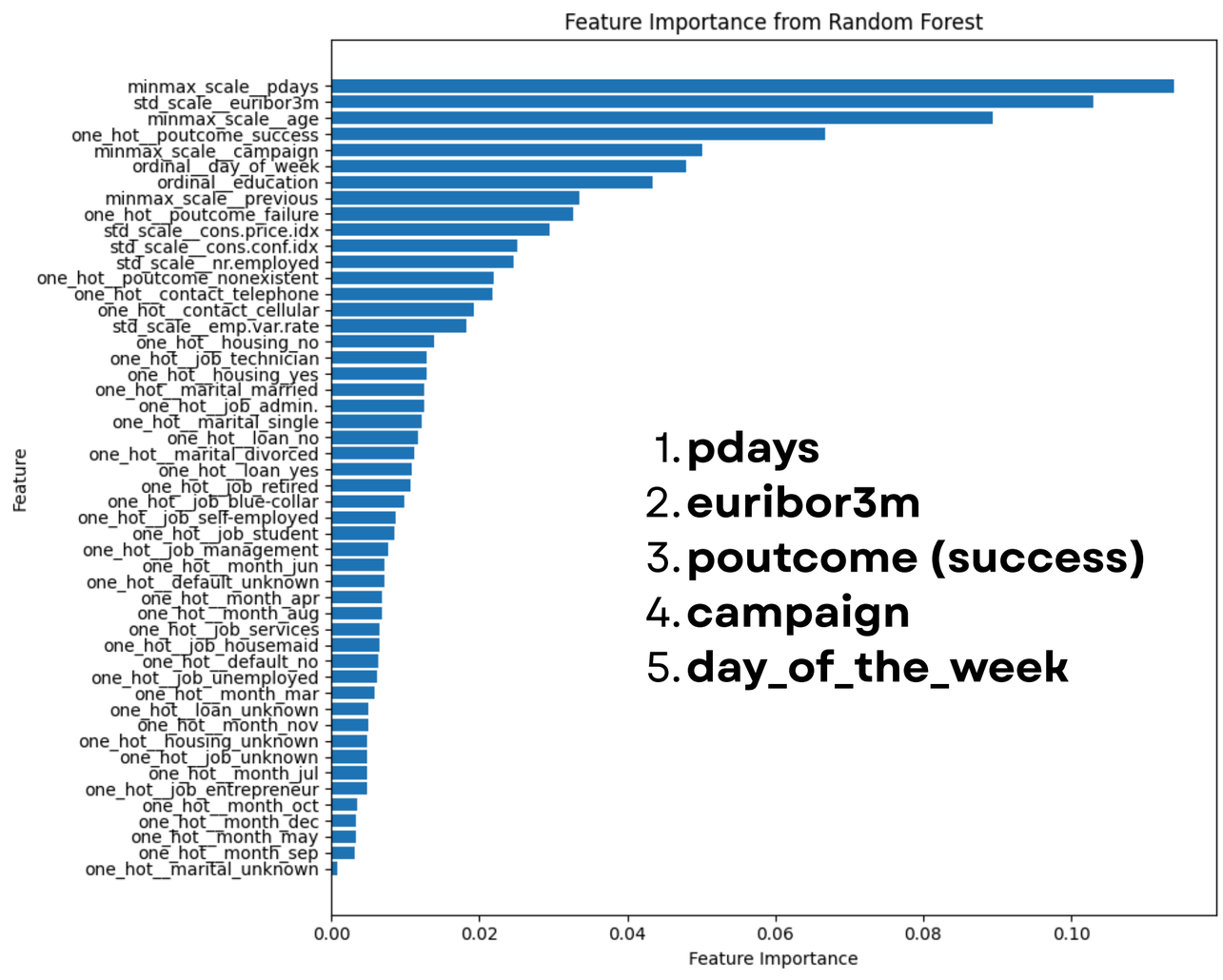
Random Forest

Results

Accuracy : 0.694
Precision: 0.667
Recall : 0.776
F1 score : 0.717



Confusion Matrix

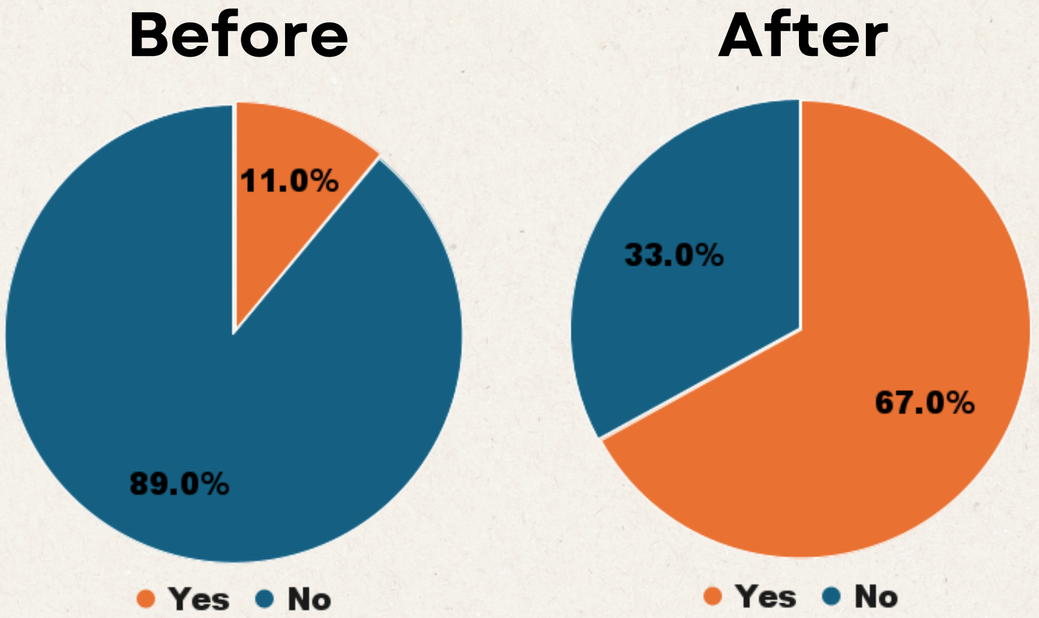


Business impact !

What does that mean?

67% Precision means:
When the model says “YES”
→ It’s right 67 times out of 100

77% Recall means:
Out of all who would have said yes
→ The model catches 77% of them



Convertion rate from **11%** to **67%**
By targeting only the positive predictions out of a lead dataset

Thank you

Alexandre Waemiers

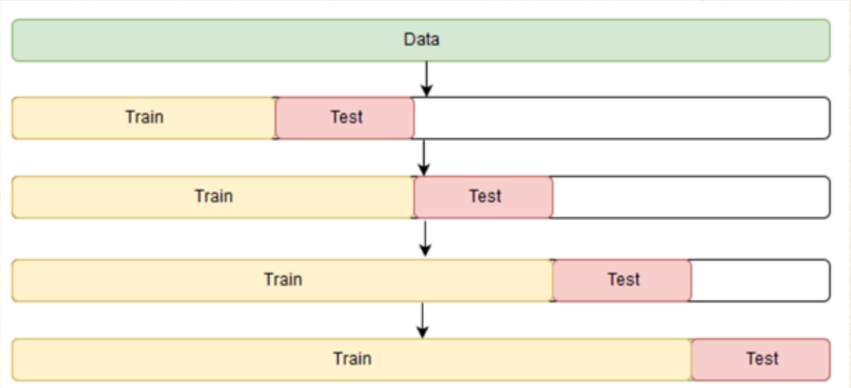
Vincent Lamy

Train

Feature Engineering	Preprocessing	Training
None	**Drop** duration **StandardScaler** all numerical One-Hot Encoder all categorical	**BaseLine** Logistic Regression **Tree-based** RandomForest GradientBoosting XGBoost CatBoost

Validation

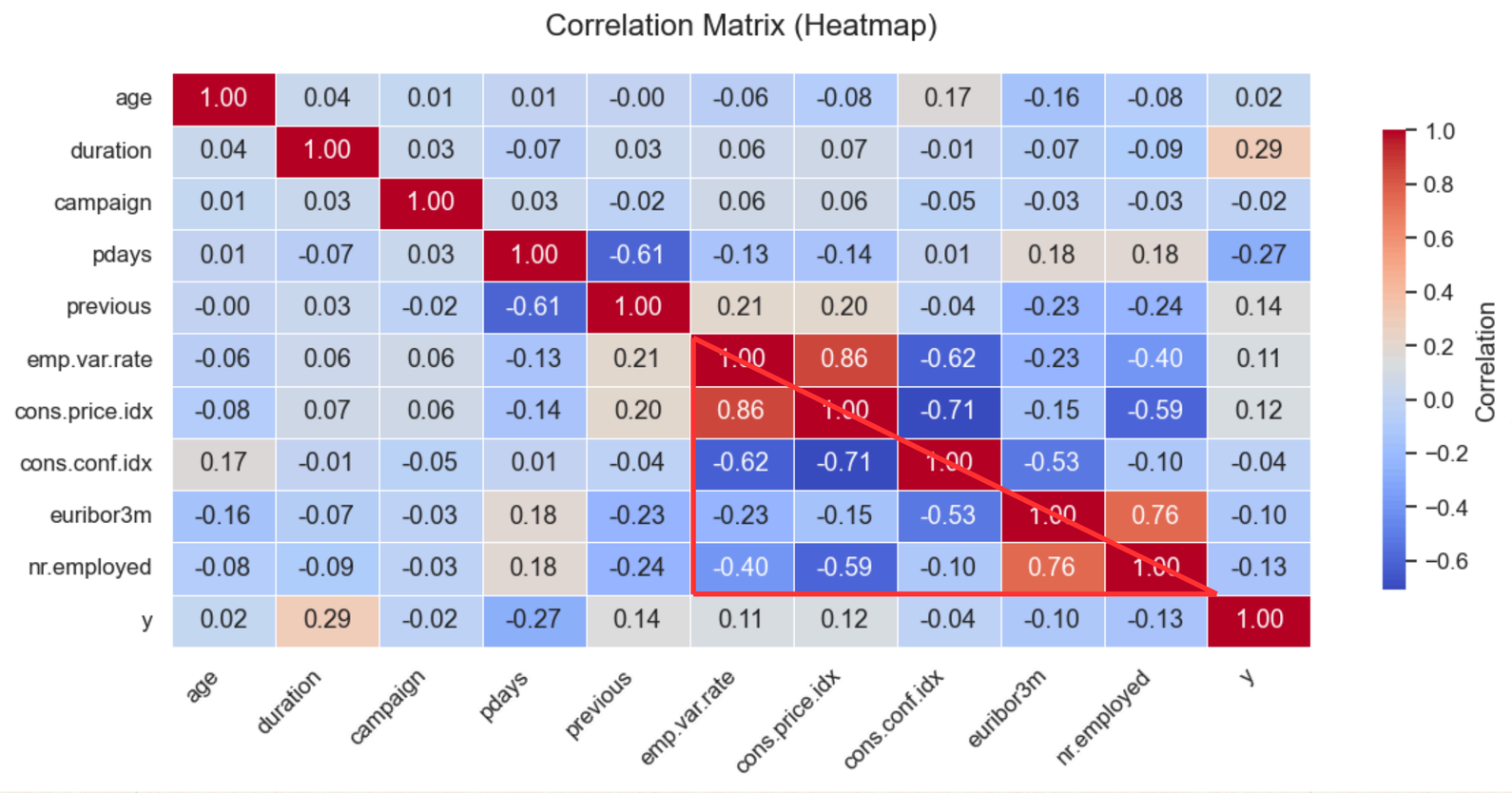
Time Series Cross Validation



Metrics

Accuracy : avg = 0.55
Precision : avg = 0.57
Recall : avg = 0.60
F1-score : avg = 0.58

EDA - Correlation Matrix



Parameter

Pearson - linear relationship

Observations

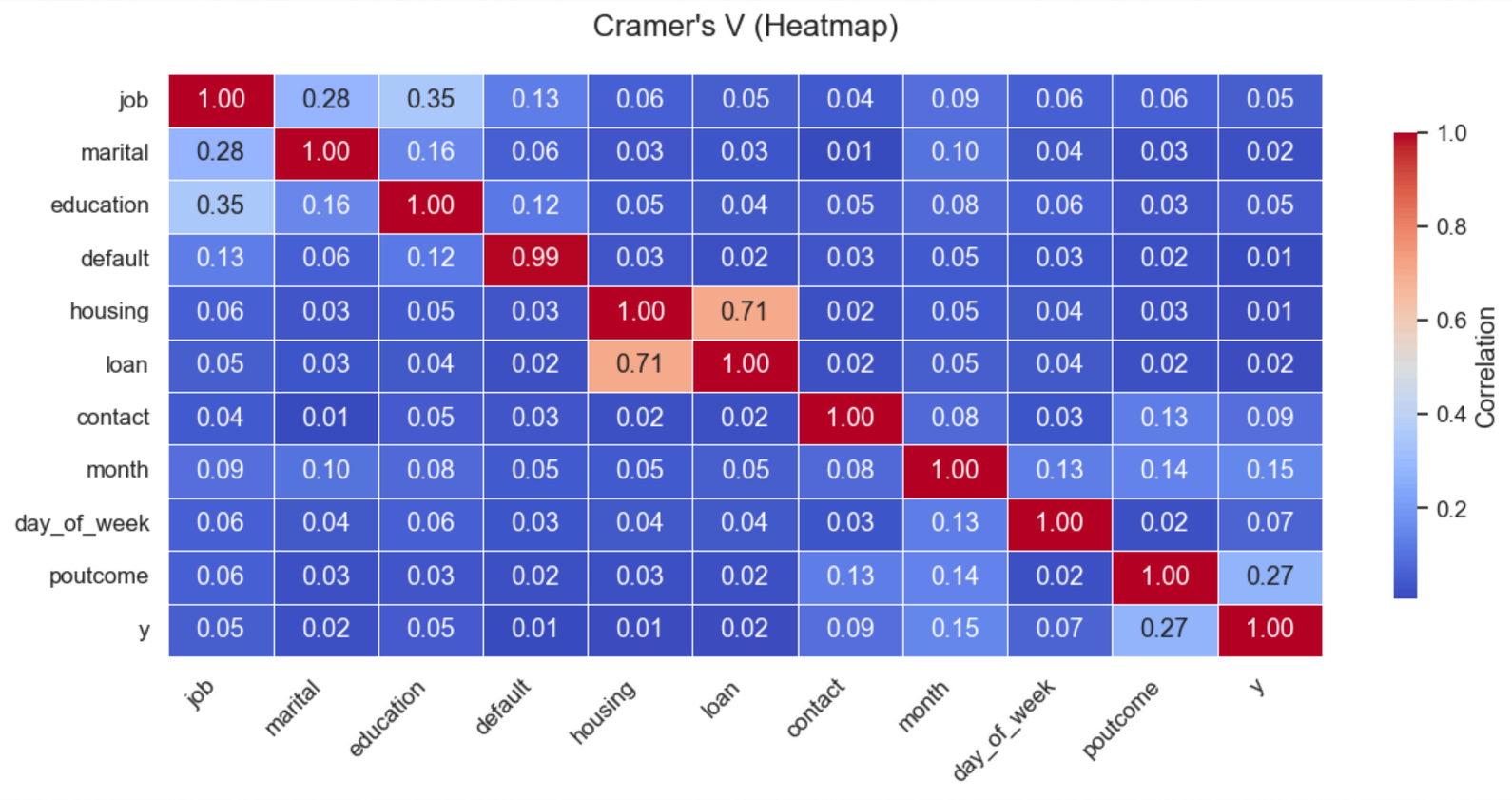
Socioeconomic cluster less impressive

Correlation to target

duration but data leakage

pdays → previous

EDA - Cramer's V



Observations

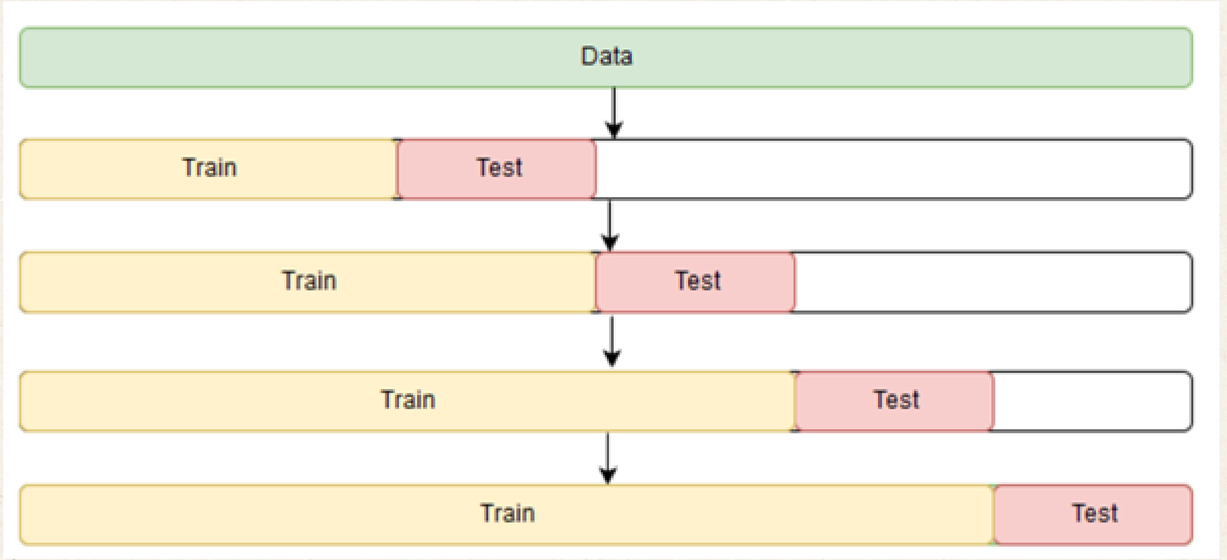
education/job/marital

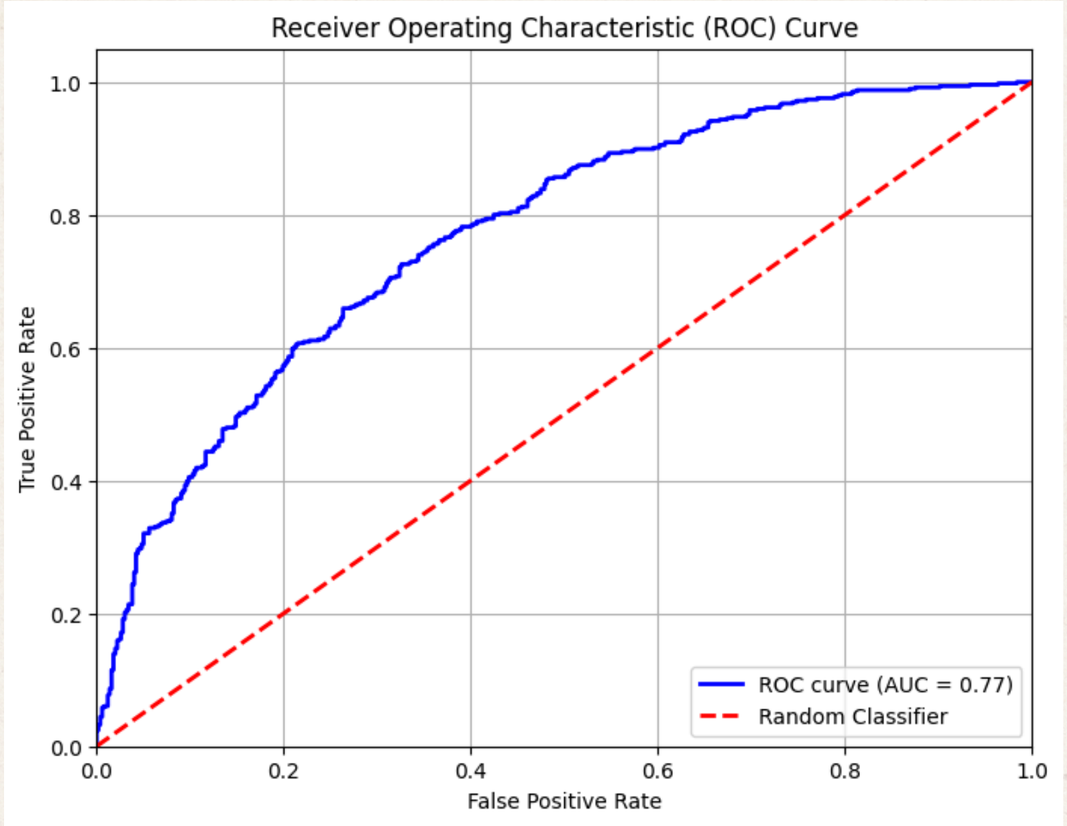
loan/housing

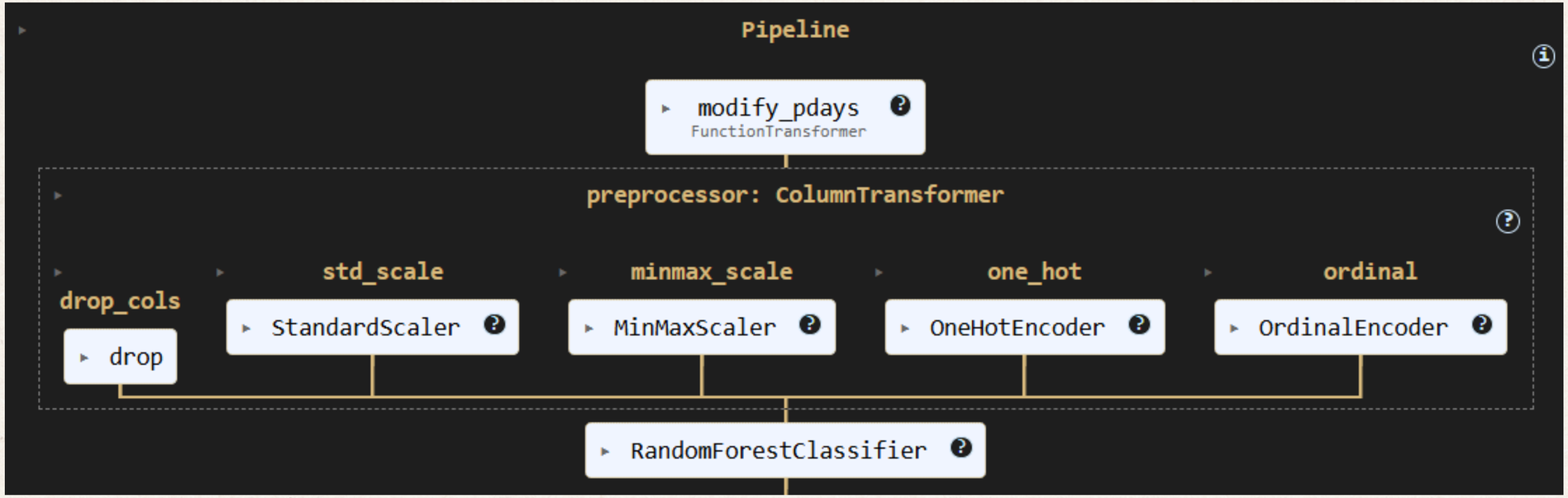
Correlation to target

poutcome

Time Series Cross Validation




Receiver Operating Characteristic (ROC) Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (AUC = 0.77)
Random Classifier

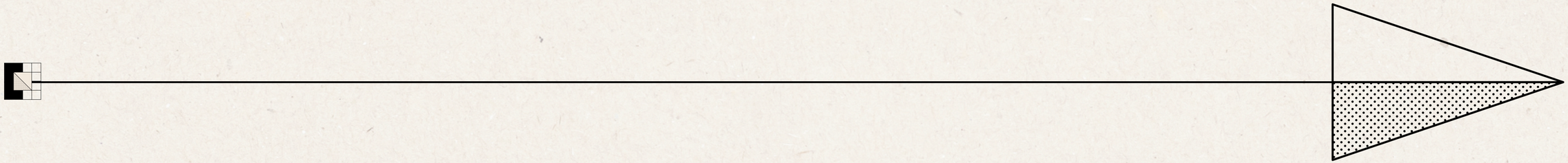

Pipeline
modify_pdays
FunctionTransformer
preprocessor: ColumnTransformer
drop_cols
drop
std_scale
StandardScaler
minmax_scale
MinMaxScaler
one_hot
OneHotEncoder
ordinal
OrdinalEncoder
RandomForestClassifier

What's Next ?

Data

Use more relevant macro economics

More client information : location, financial status, etc...



Models

Explore Deep Learning algorithms : FFNN, RNN