



# Telemarketing Campaign Predictive Analysis for Portuguese Banking Institution

# The problem

## Company

Portuguese Banking  
Institution

## Context

- Firms spend massive amounts of money on marketing campaigns hoping to maximize the return on investment (ROI).
- Understanding the customers along is crucial for achieving an effective marketing strategy.
- Data analysis insights leads to an intelligent targeted marketing.

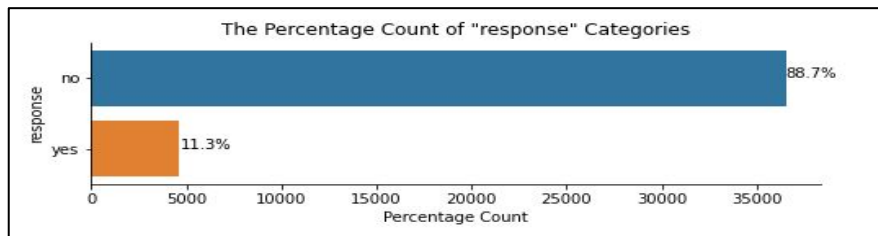
## Problem statement

- Not satisfying ROI

Through a telemarketing campaign, there is a need for a model that predicts whether a given customer will subscribe to a term deposit service offered by the bank.

# Exploratory Data Analysis (EDA)

- Contains 20 features and a binary target variable. Available in the UCI Machine Learning Repository. <http://archive.ics.uci.edu/ml/datasets/Bank+Marketing#>.
- 11 categorical variables including the response variable.
- 10 Numeric Features
- Imbalanced dataset
- 41188 Data points



Data columns (total 21 columns):				
#	Column	Non-Null Count		Dtype
0	age	41188	non-null	int64
1	job	41188	non-null	object
2	marital	41188	non-null	object
3	education	41188	non-null	object
4	default	41188	non-null	object
5	housing	41188	non-null	object
6	loan	41188	non-null	object
7	contact	41188	non-null	object
8	month	41188	non-null	object
9	day_of_week	41188	non-null	object
10	duration	41188	non-null	int64
11	campaign	41188	non-null	int64
12	pdays	41188	non-null	int64
13	previous	41188	non-null	int64
14	poutcome	41188	non-null	object
15	emp.var.rate	41188	non-null	float64
16	cons.price.idx	41188	non-null	float64
17	cons.conf.idx	41188	non-null	float64
18	euribor3m	41188	non-null	float64
19	nr.employed	41188	non-null	float64
20	y	41188	non-null	object
dtypes: float64(5), int64(5), object(11)				

# 1. Categorical Features

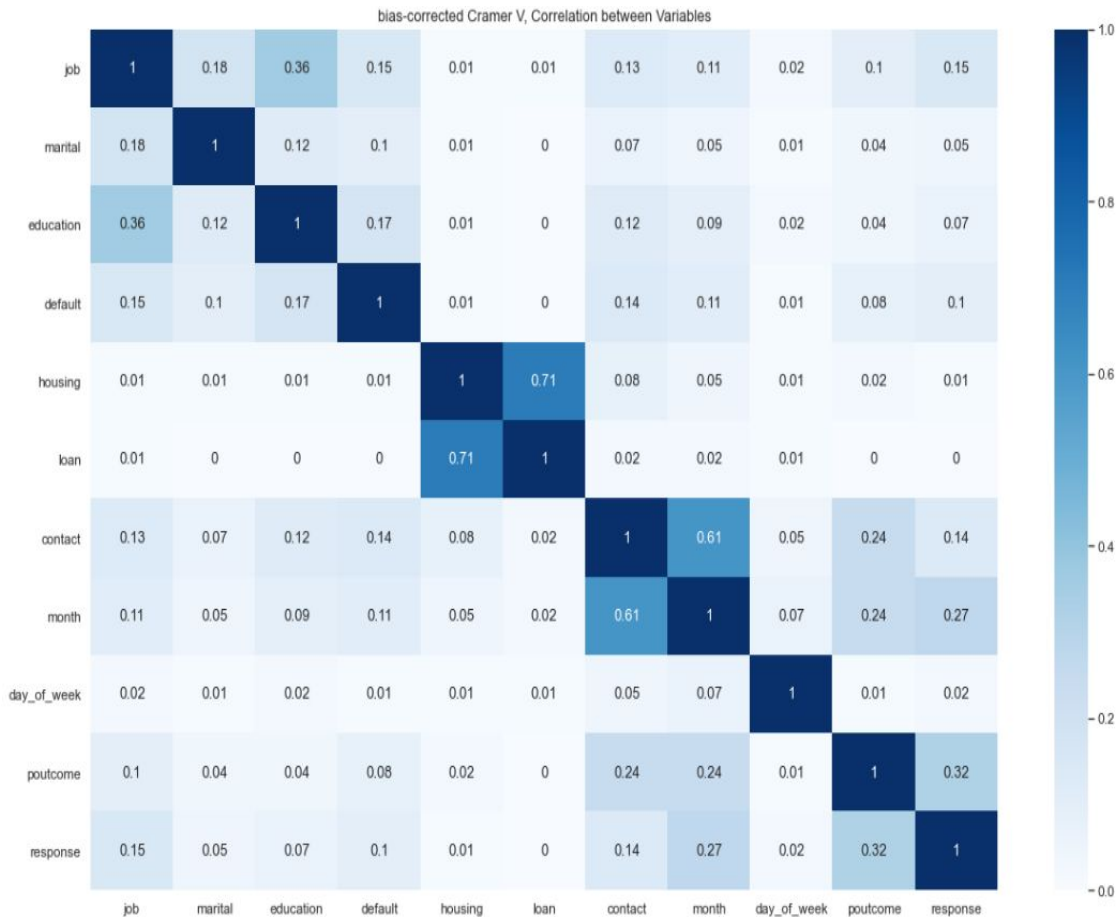
- Correlation with the response variable using Cramer's V
- Highest Correlation
  - poutcome: outcome of the previous marketing campaign
  - Month
  - Job

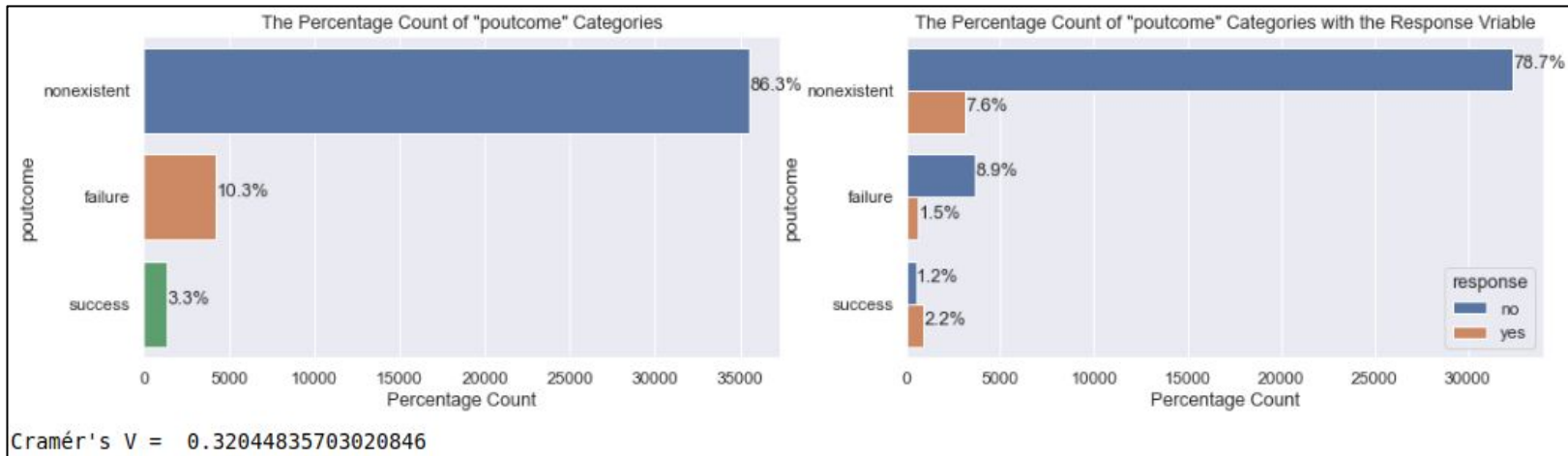
	Feature	Cramers_V
0	poutcome	0.320448
1	month	0.274123
2	job	0.151955
3	contact	0.144612
4	default	0.099123
5	education	0.067183
6	marital	0.053976
7	day_of_week	0.023143
8	housing	0.009533
9	loan	0.000000

# 1. Categorical Features

## Correlation Analysis using Cramer's V

- The heatmap below shows a strong association between loan and house variables with Cramer's  $V = 0.71$ , and a moderate association between education and job  $V = 0.36$ .
- Also, a strong association between month and contact with  $V = 0.61$ .
- poutcome, Month, and Job have weak correlation with the response variable.





Notice: The "All" row and column from the calculated matrix is the sum of values in that row and column respectively. It possible to be not located at the very end because of sorting by subscription rate.

response	no	yes	All	within class subscription per poutcome %
poutcome				
success	479	894	1373	65.112891
failure	3647	605	4252	14.228598
All	36537	4639	41176	11.266272
nonexistent	32411	3140	35551	8.832382

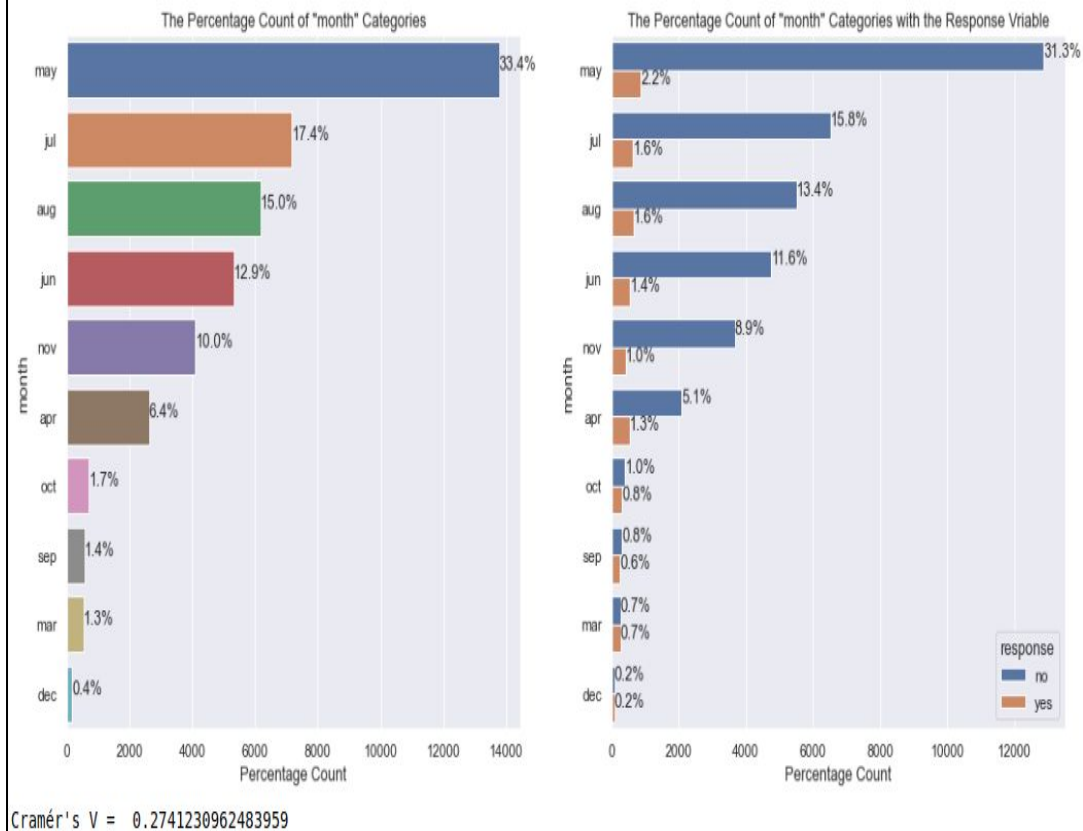
65% of cantaccted previously subscribed subscribed again



# Month

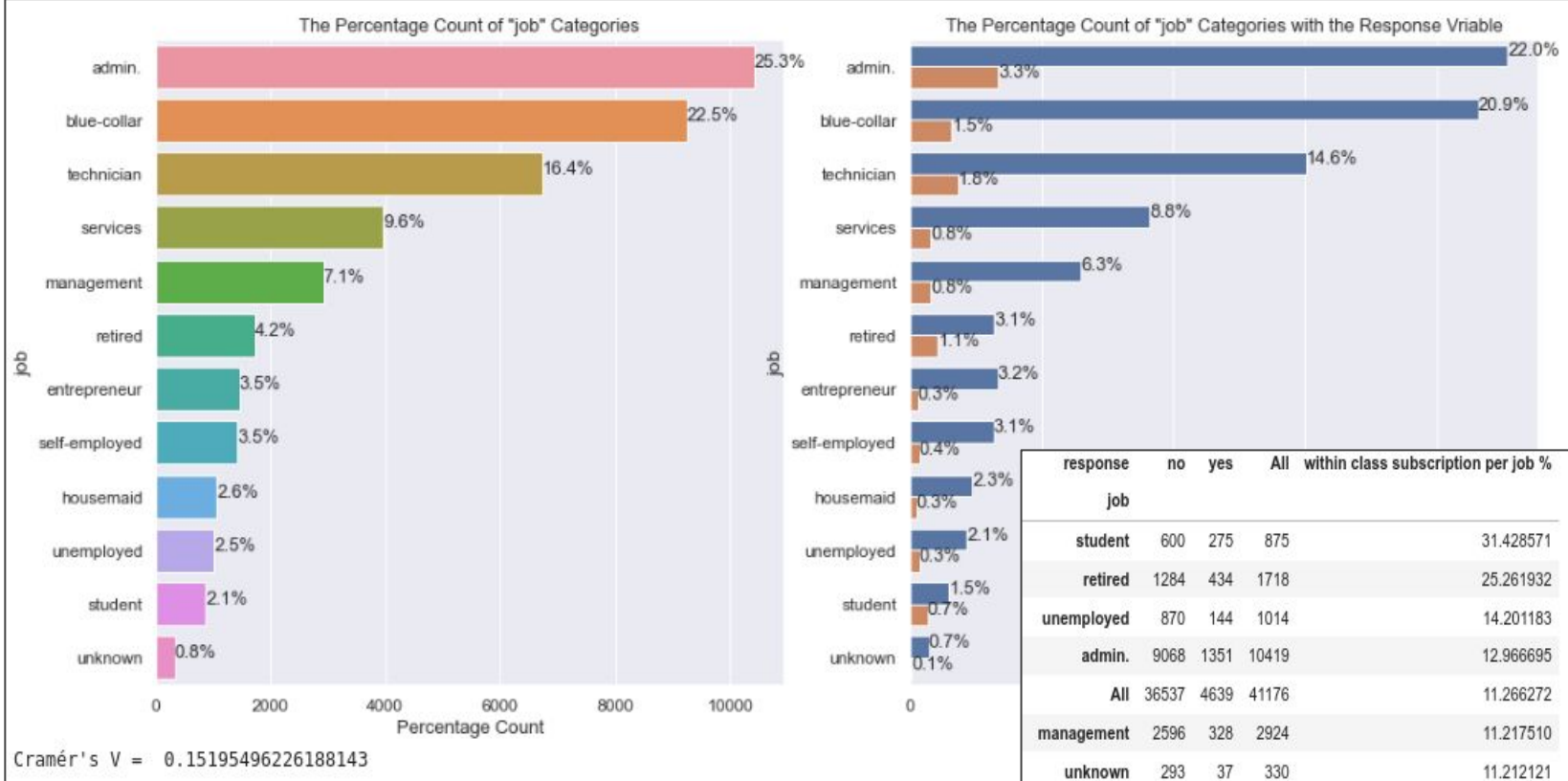
- May received most calls
- Mars has highest success rate

response	no	yes	All	within class subscription per month %
month				
mar	270	276	546	50.549451
dec	93	89	182	48.901099
sep	314	256	570	44.912281
oct	402	315	717	43.933054
apr	2092	539	2631	20.486507
All	36537	4639	41176	11.266272
aug	5521	655	6176	10.605570
jun	4759	559	5318	10.511470
nov	3684	416	4100	10.146341
jul	6521	648	7169	9.038918
may	12881	886	13767	6.435680





# Job



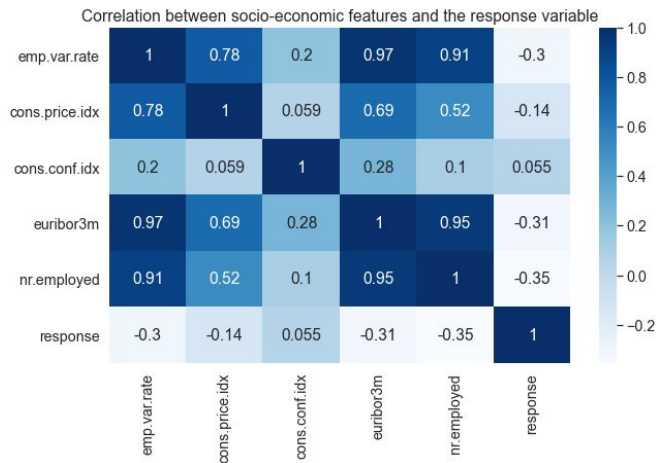
- admins received most calls
- Students and retired has highest success rate



# 1. Numerical Features

## Correlation Analysis using Pearson

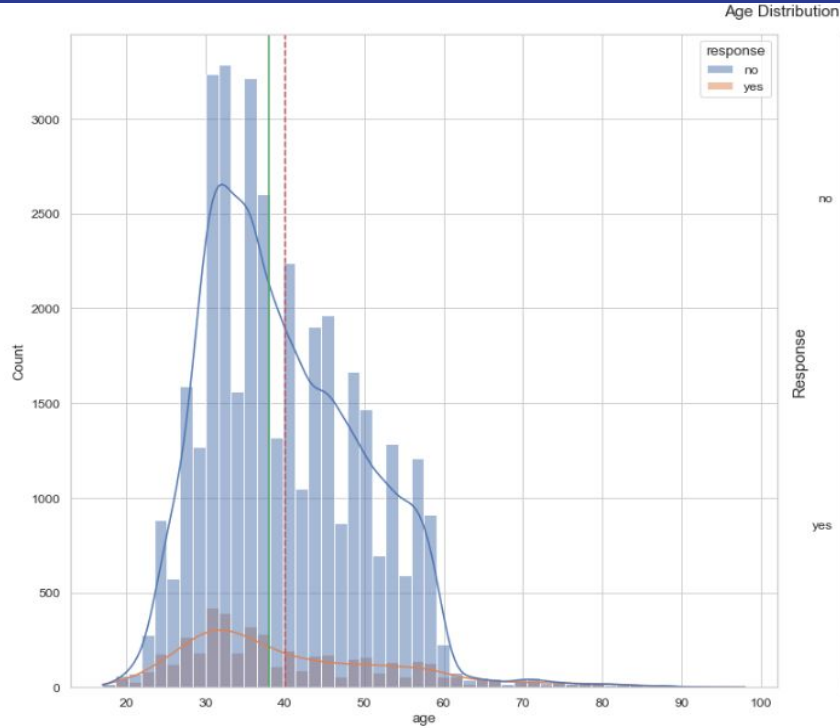
Positive correlation between most of the social and economic context attributes.  
None of these attributes correlate with the response variable



# 1. Numerical Features



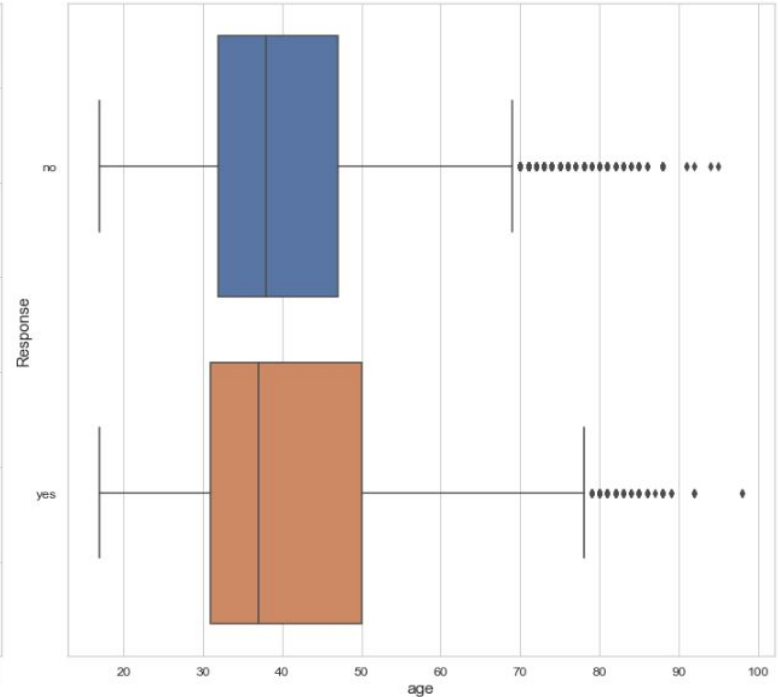
age



The age description for for both response classes =

	count	mean	std	min	25%	50%	75%	max
no	36537.0	39.910994	9.897176	17.0	32.0	38.0	47.0	95.0
yes	4639.0	40.912266	13.838838	17.0	31.0	37.0	50.0	98.0

```
bank_data.groupby('response')['age'].describe()
```

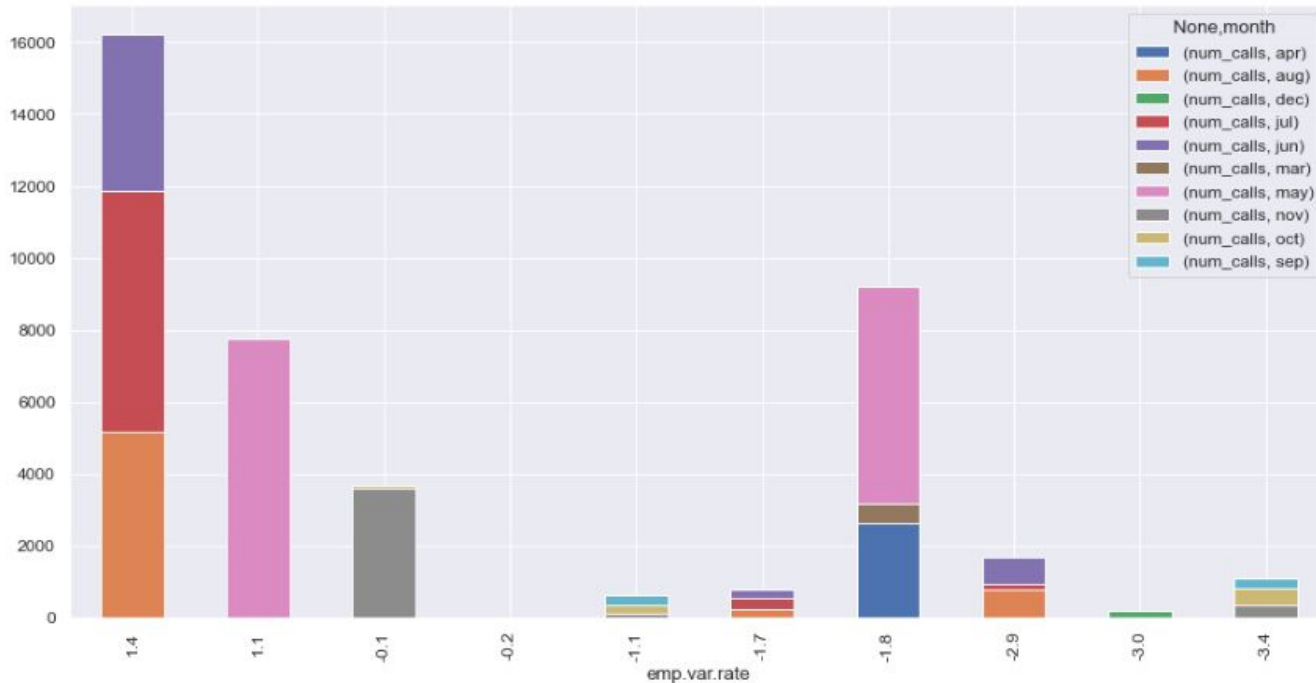


- Age overlaps across both classes
- Outliers in both classes



## emp.var.rate

- **emp.var.rate** feature refers to employment variation rate — quarterly indicator
- 10 distinct values
- Most calls in positive var rate
- Subscription rate highest when emp.var.rate = -1.7



emp.var.rate	-3.4	-3.0	-2.9	-1.8	-1.7	-1.1	-0.2	-0.1	1.1	1.4	All
response											
no	616.0	84.0	1069.0	7721.0	370.0	334.0	9.0	3450.0	7522.0	15362.0	36537.0
yes	454.0	88.0	593.0	1461.0	403.0	301.0	1.0	232.0	240.0	866.0	4639.0
All	1070.0	172.0	1662.0	9182.0	773.0	635.0	10.0	3682.0	7762.0	16228.0	41176.0
success_rate %	42.0	51.0	36.0	16.0	52.0	47.0	10.0	6.0	3.0	5.0	11.0

# Correlation Analysis

## Cramer's V

- Pairwise Correlation between categorical variables
- Assumes symmetrical relation

## Pearson Correlation ( $r$ )

- Pairwise Correlation between numerical variables
- Assumes symmetrical relation

## Predictive power Score (PPS)

- Pairwise Correlation between numerical and Categorical variables
- Asymmetrical relation

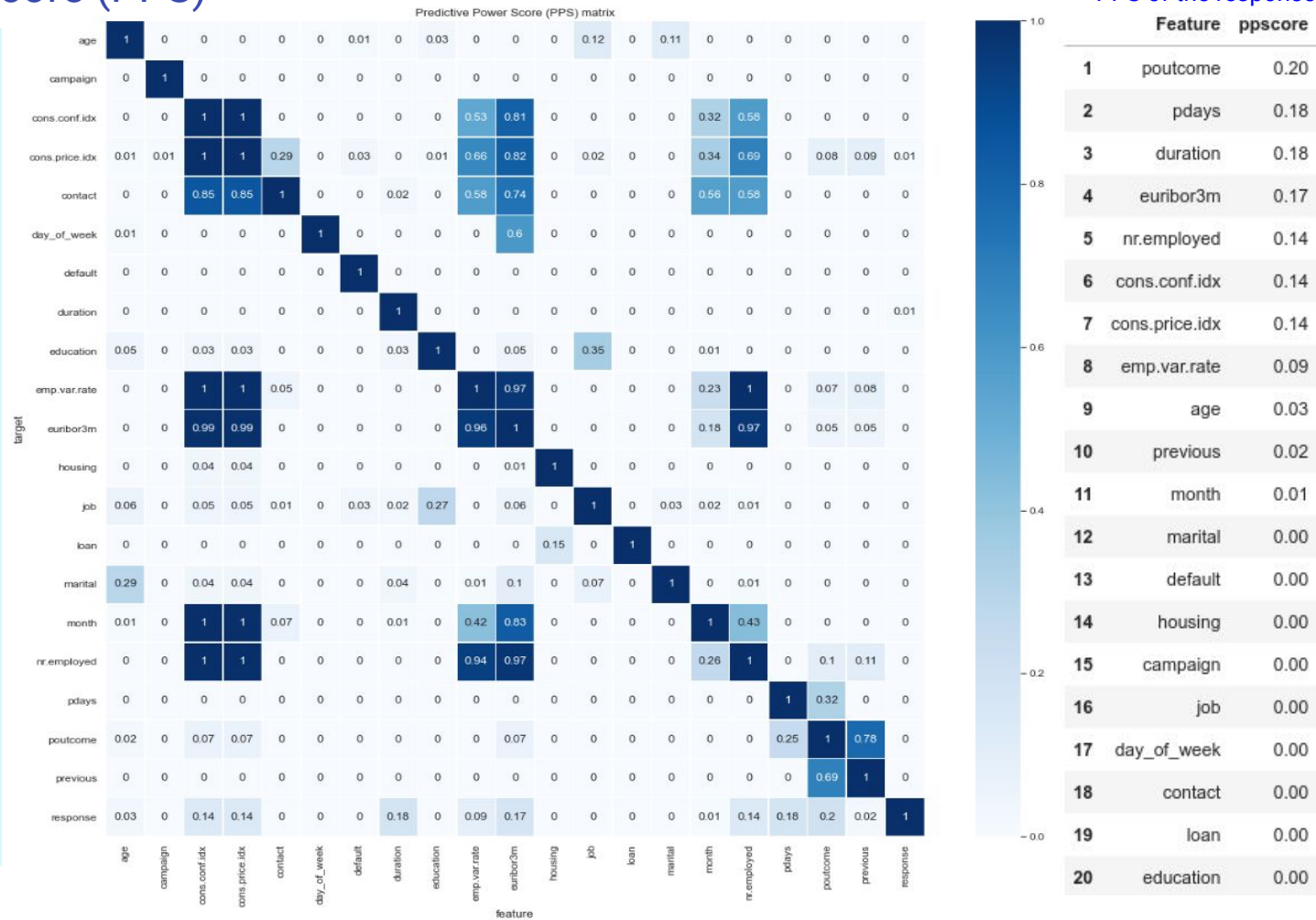
# Predictive power Score (PPS)

PPS confirmed the correlation between most of the social and economic context attributes. Correlated features:

- 'euribor3m',
- 'cons.price.idx',
- 'cons.price.idx'
- 'emp.var.rate',
- 'nr.employed'

- Job is predictive of education

- From the table, only few features are weak predictor to the response



# Preprocessing

## Preprocessing

- Label Encoder
- One Hot Encoder
- Train Test Split

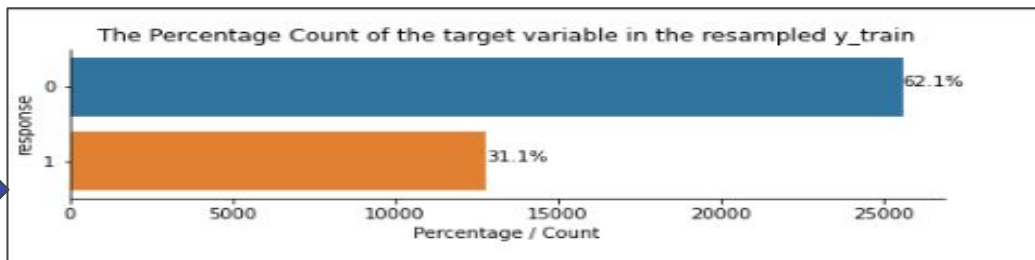
- Oversampling using SMOTE,  
Synthetic Minority Oversampling  
Technique,

## Results

The new features are 61:

```
['age', 'job_housemaid', 'job_services', 'job_admin.', 'job_blue-collar', 'job_technician', 'job_retired', 'job_management', 'job_unemployed', 'job_self-employed', 'job_other', 'job_entrepreneur', 'job_student', 'marital_married', 'marital_single', 'marital_divorced', 'marital_other', 'education_basic.4y', 'education_high.school', 'education_basic.6y', 'education_basic.9y', 'education_professional.course', 'education_other', 'education_university.degree', 'education_illiterate', 'default_no', 'default_yes', 'housing_no', 'housing_yes', 'loan_no', 'loan_yes', 'contact_telephone', 'contact_cellular', 'month_may', 'month_jun', 'month_jul', 'month_aug', 'month_oct', 'month_nov', 'month_dec', 'month_mar', 'month_apr', 'month_sep', 'day_of_week_mon', 'day_of_week_tue', 'day_of_week_wed', 'day_of_week_thu', 'day_of_week_fri', 'duration', 'campaign', 'pdays', 'previous', 'poutcome_nonexistent', 'poutcome_failure', 'poutcome_success', 'emp_var_rate', 'cons_price_idx', 'cons_conf_idx', 'euribor3m', 'nr_employed', 'response']
```

60 variables including the target after dropping duration.



# Modeling results and analysis

# Modeling Method

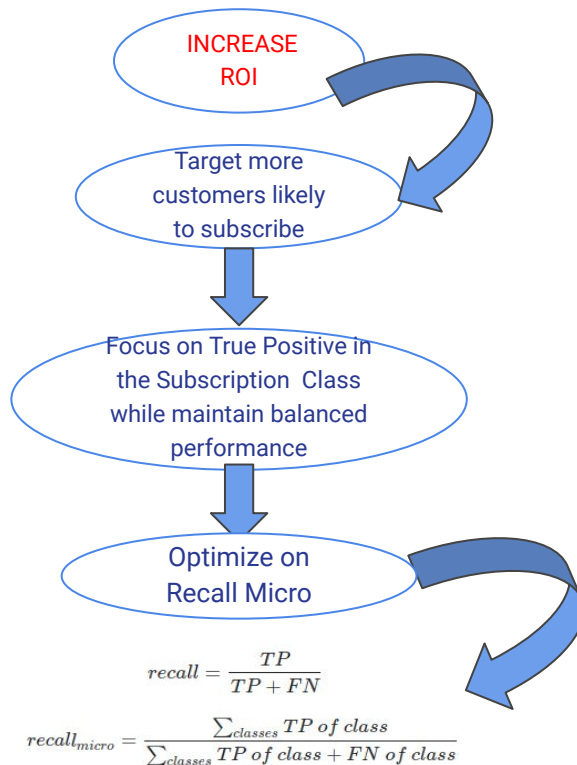
## Models

- Tested classification algorithms:
  - Logistic Regression
  - Decision Tree
  - Random Forest
  - Light Gradient Boosting
  - XGboost

## Hyperparameter Tuning

- Bayesian hyperparameter optimization, using Tree-structured Parzen Estimator Approach (TPE).
- Various objective functions applied.
- Recall\_micro gave more balanced.

## Business Objective Reflected

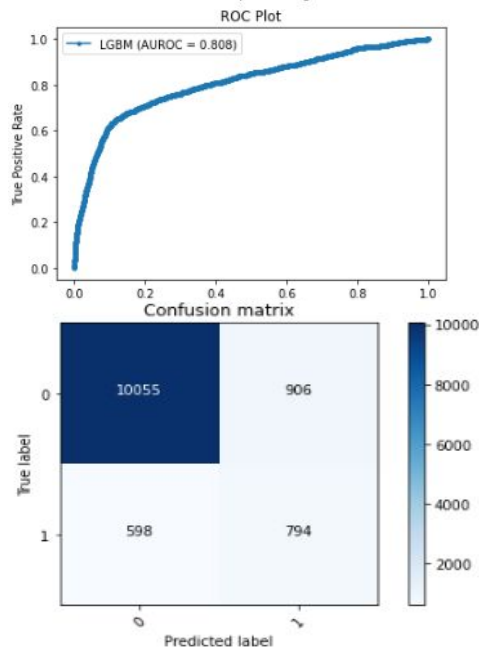




# Models Comparison

## Winning Model

- Light Gradient Boosting Model
  - ROC AUC = 0.81
  - Recall Weighted avg = 0.88

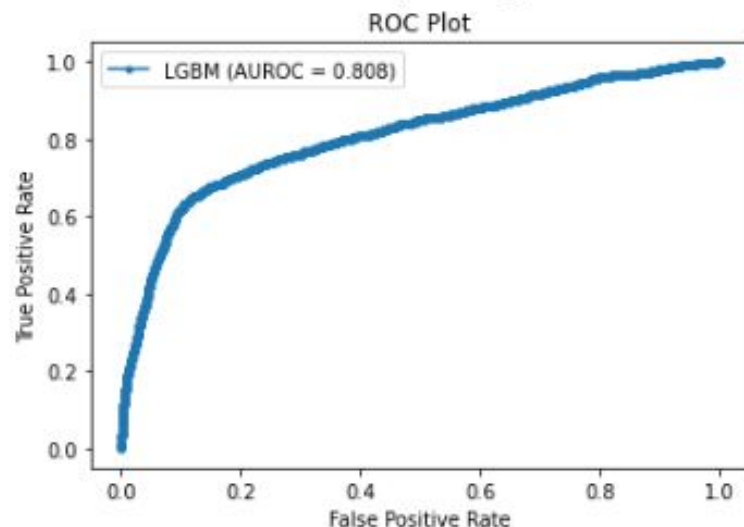
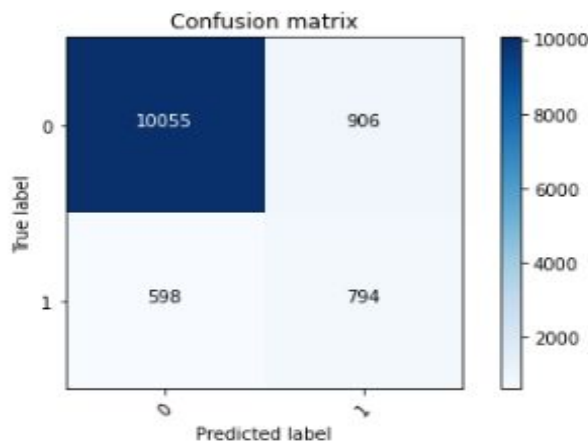


		precision	recall	f1-score	support
model / objective function	Logistic Regression/ F1				
	Not Subscribed 0	0.95	0.85	0.9	10961
	Subscribed 1	0.35	0.64	0.46	1392
	accuracy				0.83
	macro avg	0.65	0.75	0.68	12353
	weighted avg	0.88	0.83	0.85	12353
	ROC AUC	0.8	0.8	0.8	0.8
lgbm / recall_micro	Not Subscribed 0	0.94	0.92	0.93	10961
	Subscribed 1	0.47	0.57	0.51	1392
	accuracy				0.88
	macro avg	0.71	0.74	0.72	12353
	weighted avg	0.89	0.88	0.88	12353
	ROC AUC	0.81	0.81	0.81	0.81
xgb / recall_micro	Not Subscribed 0	0.93	0.95	0.94	10961
	Subscribed 1	0.5	0.4	0.45	1392
	accuracy				0.89
	macro avg	0.72	0.68	0.69	12353
	weighted avg	0.88	0.89	0.88	12353
	ROC AUC	0.79	0.79	0.79	0.79
Decision Tree / recall_micro	Not Subscribed 0	0.93	0.93	0.93	10961
	Subscribed 1	0.46	0.48	0.47	1392
	accuracy				0.88
	macro avg	0.7	0.7	0.7	12353
	weighted avg	0.88	0.88	0.88	12353
	ROC AUC	0.78	0.78	0.78	0.78
Random Forest / Recall_Mic.	Not Subscribed 0	0.93	0.95	0.94	10961
	Subscribed 1	0.5	0.4	0.45	1392
	accuracy				0.89
	macro avg	0.72	0.68	0.69	12353
	weighted avg	0.88	0.89	0.88	12353
	ROC AUC	0.8	0.8	0.8	0.8

# Winning Model Details

```
XGBClassifier(base_score=0.5, booster='gbtree', colsample_bylevel=1,
              colsample_bynode=1, colsample_bytree='0.675', eval_metric='aucpr',
              gamma='0.073', gpu_id=-1, importance_type='gain',
              interaction_constraints='', learning_rate=0.300000012,
              max_delta_step=0, max_depth=10, min_child_weight=1, missing=nan,
              monotone_constraints='()', n_estimators=20, n_jobs=4,
              num_parallel_tree=1, random_state=0, reg_alpha=0, reg_lambda=1,
              scale_pos_weight=2, subsample=1, tree_method='exact',
              validate_parameters=1, verbosity=None)
```

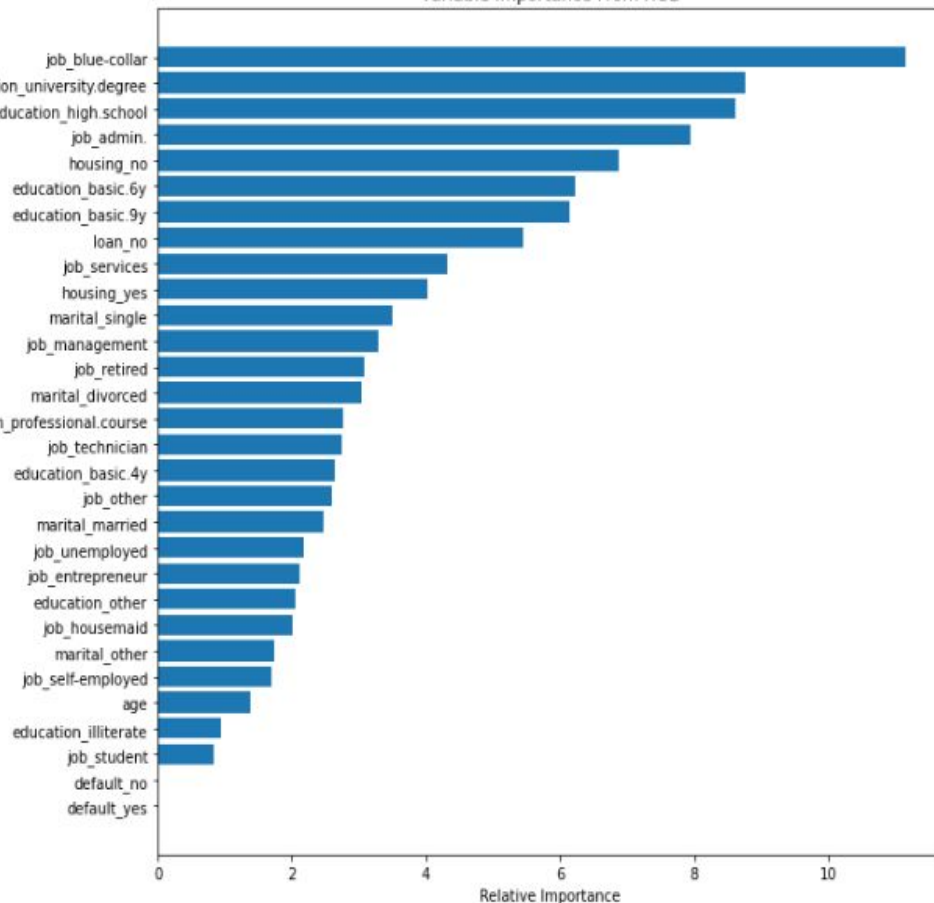
	precision	recall	f1-score	support
0	0.94	0.92	0.93	10961
1	0.47	0.57	0.51	1392
accuracy			0.88	12353
macro avg	0.71	0.74	0.72	12353
weighted avg	0.89	0.88	0.88	12353



# Look as Interesting Business Features

## Model Feature Importance

Variable Importance From XGB



## Permutation Importance

Weight	Feature
0.0027 ± 0.0007	contact_telephone
0.0023 ± 0.0009	contact_cellular
0.0021 ± 0.0005	month_apr
0.0011 ± 0.0005	poutcome_failure
0.0009 ± 0.0005	cons_price_idx
0.0006 ± 0.0003	poutcome_nonexistent
0.0004 ± 0.0004	pdays
0.0004 ± 0.0002	month_oct
0.0004 ± 0.0005	poutcome_success
0.0003 ± 0.0007	campaign
0.0003 ± 0.0001	month_mar
0.0003 ± 0.0005	day_of_week_mon
0.0002 ± 0.0007	day_of_week_fri
0.0002 ± 0.0004	job_technician
0.0001 ± 0.0001	marital_divorced
0.0001 ± 0.0001	job_management
0.0000 ± 0.0001	previous
0.0000 ± 0.0001	job_housemaid
0.0000 ± 0.0001	education_basic.6y
0.0000 ± 0.0002	education_other
... 39 more ...	

# Summary and conclusion

## Business Perspective

- Without the model, the bank has 11.3% chance of randomly targeting those who are likely subscribe
- With Recall 0.57 on the subscribers class, Bank has 57% chance of targeting subscribers. ROI can be increased by 45.7% using the model
- The bank needs to invest more on collecting customers predictive features for more profitable predictive modeling

## Technical Perspective

- The Light Gradient Boosting algorithm optimized for Recall Micro gave the best performance. ROC AUC = 0.81 and F1= 0.51 one minority class, (waited avg F1= 0.88).
- Optimizing the hyperparameters on the Recall Micro gave a better performance of most of the models applied on this imbalanced dataset.
- The SMOTE oversampling seemed to be helpful with this imbalanced dataset. Scores on the 30% split test set indicates that the model can be generalizable.