Telemarketing Campaign Predictive Analysis

Background

It is common for firms to spend massive amounts of money on marketing campaigns. Such expensive campaigns are required to maximize the return on investment (ROI). Understanding the customers along with their needs is crucial for achieving an optimized and effective marketing strategy. Through proper data analysis, the bank can understand the customers and obtain insights that can be used for intelligently targeted marketing campaigns.

Problem Identification

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

About the Data

The data provided by the Portuguese Bank came from a direct telemarketing marketing campaign. The data include 41,188 observations. Each observation represents an existing client that is reached via a phone call.

The dataset has 20 features and a binary target variable “y”, a total of 21 columns. This data is publicly available in the UCI Machine Learning Repository, which can be retrieved from: http://archive.ics.uci.edu/ml/datasets/Bank+Marketing#.
The features as decried by the data souce are:

- Bank client data:
  1. age (numeric)
  2. job : type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid', 'management', 'retired','self-employed','services','student','technician','unemployed','unknown')
  3. marital : marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)
  4. education (categorical: 'basic.4y' ,'basic.6y' ,'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree', 'unknown')
  5. default: has credit in default? (categorical: 'no','yes','unknown')
  6. housing: has housing loan? (categorical: 'no','yes','unknown')

7. loan: has personal loan? (categorical: 'no','yes','unknown')

- Related with the last contact of the current campaign:
  8. contact: contact communication type (categorical: 'cellular','telephone')
  9. month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
  10. day_of_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')
  11. duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.
- other attributes:
  12. campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
  13. pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
  14. previous: number of contacts performed before this campaign and for this client (numeric)
  15. poutcome: outcome of the previous marketing campaign (categorical: 'failure', 'nonexistent', 'success')
- Social and economic context attributes
  16. emp.var.rate: employment variation rate - quarterly indicator (numeric)
  17. cons.price.idx: consumer price index - monthly indicator (numeric)
  18. cons.conf.idx: consumer confidence index - monthly indicator (numeric)
  19. euribor3m: euribor 3 month rate - daily indicator (numeric)
  20. nr.employed: number of employees - quarterly indicator (numeric)

- Output variable (desired target):
  21. y - has the client subscribed a term deposit? (binary: 'yes','no')

The Analysis Method

The problem is a binary classification that predicts whether the client will subscribe to the term deposit service. Several classification machine learning algorithms were used to develop classification models. The one with the best performance is chosen. The performance measure is chosen based on the business objectives.

The objective is to maximize targeting customers who are likely to subscribe. The true positives rate is more important in this case, in particular in the subscribed class. Thus, more attention was made to the Recall score.

The tired models are:

- Logistic Regression
- Decision Tree
- Random Forest
- Light Gradient Boosting
- XGboost

The project went through all of the six steps of the data science method Problem identification, Data wrangling, Exploratory data analysis, Pre-processing & training data, Modeling, and Documentation.

- Oversampling using SMOTE is used to cope with the data imbalance.
- Bayesian hyperparameter optimization using hyperopt package is used to optimize the hyperparameters of each model.
- Various objective functions were applied to the various classification algorithms.

Data Wrangling

Below are the major issues encountered in the data:

1. There were 12 duplicates data rows. Duplicates are removed.
2. There were “unknown” values for some categorical variables, there are handled as follows:
    a. “Unknown” values for “housing”, “loan”, and “default” are replaced by the most frequent values since these features can be only “yes” or “no.”
3. “Unknown” values for “job”, “marital”, and “education” are replaced by the value “other.” These variables are not binary, the values can be other than what exists in the data (e.g. job can be a teacher and it does not exist in the data), and there is not enough information in the data to impute the missing values. Thus, it made sense to create a new categorical variable as “other.”
4. pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted). The value 999 is replaced by 0.
5. The duration feature is not known before the call is made. Thus, this feature is not practical for predictive modeling.

Exploratory Data Analysis

First, Each variable is inspected individually. Second, the correlation between the variables and the target variable is examined using different methods for numerical and categorical variables.

1. **Categorical Features**

The distribution of the variables across classes is calculated and visualized. Also, the correlation between categorical variables is calculated. The used method is Cramers V statistic for the categorical-categorical association, which uses correction from Bergsma and Wicher, Journal of the Korean Statistical Society 42 (2013): 323-328. The figures below show examples of the analysis and key findings.

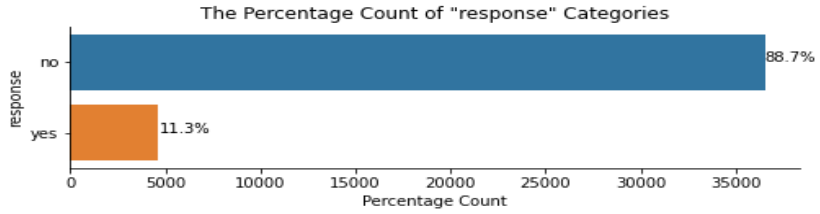
Cramér's V statistic is used as a measure of association between two categorical variables. Cramer's V is a normalized version of the chi-square test statistic, that determines the effect size. Its values are between [0 and 1] Interpreting Cramer's V, To interpret Cramer's V, the following approach is often used:

V in [0.1, 0.3]: weak association, in [0.4, 0.5]: medium association, V > 0.5: strong.

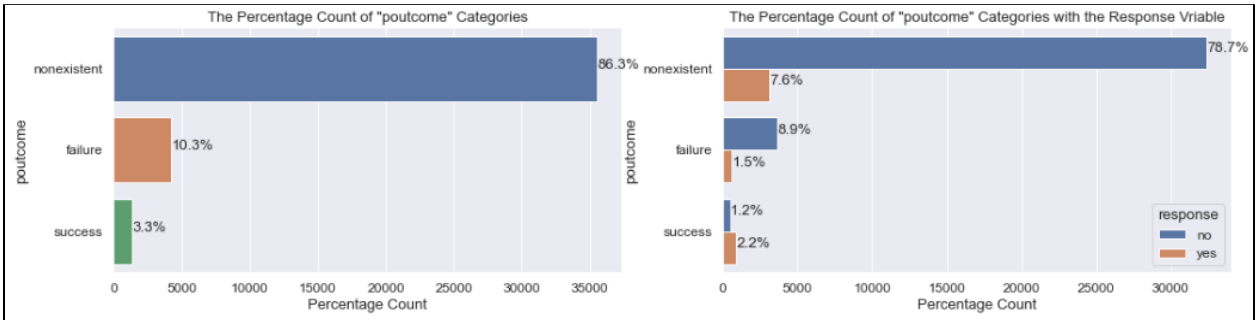
- The table below shows the categorical variables ranked by strength of the association with the response variable identified by Cramers V (aka Φ). This is a pair-wise association, does not account for variables interaction.

	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

- The response variable is imbalance. 88.7% none subscribers and 11.3% subscribers.



- **Poutcome** feature. The figure below shows that most of the subscribers were not previously contacted. The total of subscribers in this data set is 11.27%, 7.6% of the subscribers in this data set are new customers not previously contacted. Total subscribers are 4640, and 3141 of them are new customers. The percentages in the figure are based on the total number of customers (the data points) Approximately 68% of the subscribers are new customers. However, most of the contacted customers (86.3%) are new. As shown in the frequency table below, the subscription rate among the none previously contacted customers is the lowest with an 8.8% rate although they were the most contacted group.
- About 65% of the successfully subscribed customers through the previous campaign have subscribed as a result of this campaign. Previously subscribed customers showed more tendency to subscribe to other products.
- 14.2% of the 4252 customers who did not subscribe in the previous campaign have subscribed.



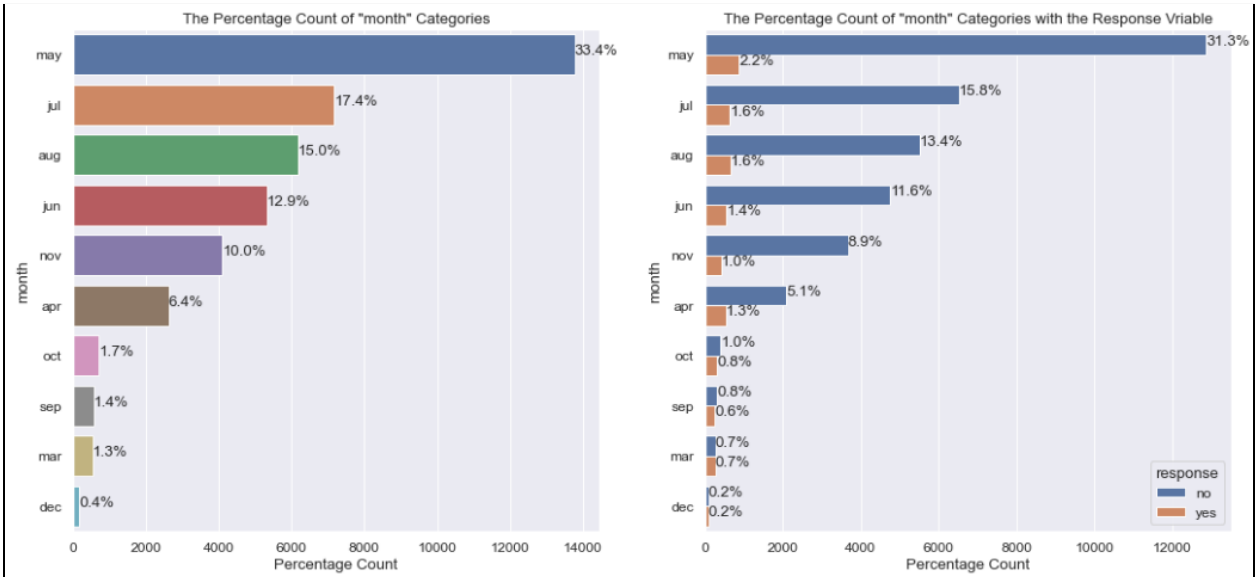

```
Cramér's V =  0.32044835703020846
```

```
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

- The **month** feature has the second-highest Cramer's V = 0.27. However, the association is weak. The campaign spanned 10 months.

- May month has the highest subscription rate; 2.2% of the total data. However, May also has received the most campaign calls with 33.4%. Thus, looking at the subscription rate and ignoring the call rate can be misleading.
- The table below shows the subscription rate per month in descending order. Although Mars and December have the fewest campaign calls, their subscription rates are the highest among other months, 50.5% and 50% accordingly.




```
Cramér's V =  0.2741230962483959
```

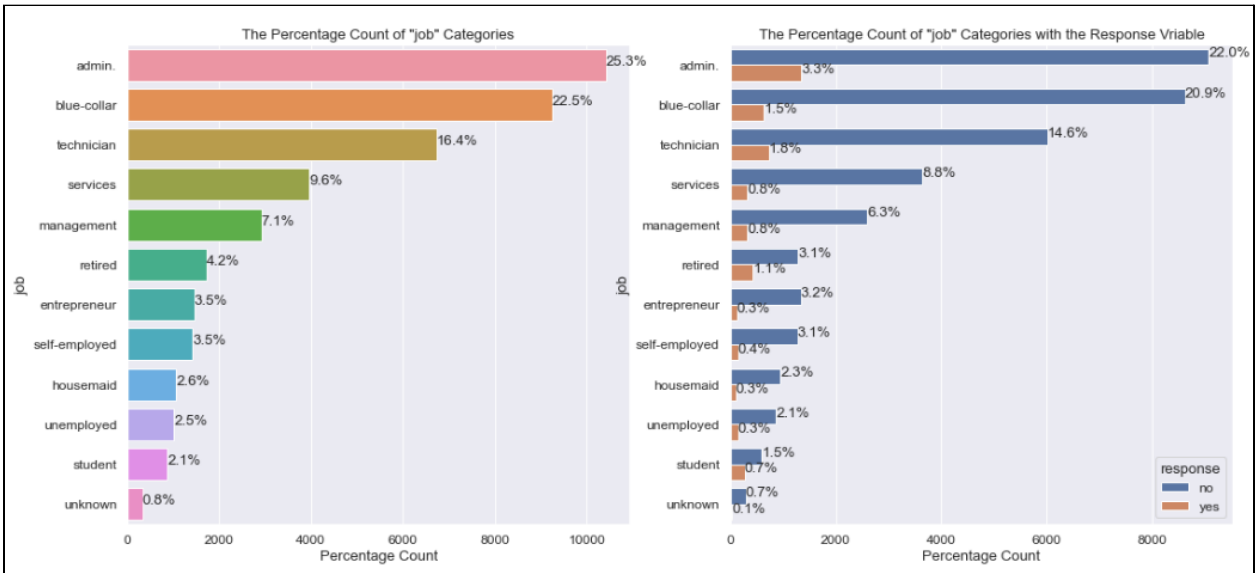
```
Notice: The "All" row and column from the calculated matrix is the sum of values in that row and column respectively.
It possible to be not located at the very end because of sorting by subscription rate.
```

response	no	yes	All	within class subscription per 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

- The **job** feature has the third-highest Cramer's V = 0.15. However, the association is also weak. The admin job has the highest subscription rate;

3.3% of the total customers. However, admins also have received the most campaign calls with 25.3%. The campaign probably focused on the admins because they are expected to have better income, leading to more chances of profitable subscription to a term deposit. However, out of 10419 admins contacted only 1351 have subscribed, about only 13% of the contacted admins were subscribed. Admins are ranked 4th among the within-class subscription rate as shown in the frequency table.

- Ironically, students are ranked first among the within-class subscription rate with 31.4%, followed by the retired with 25.3% although students are the least contacted class in this dataset.

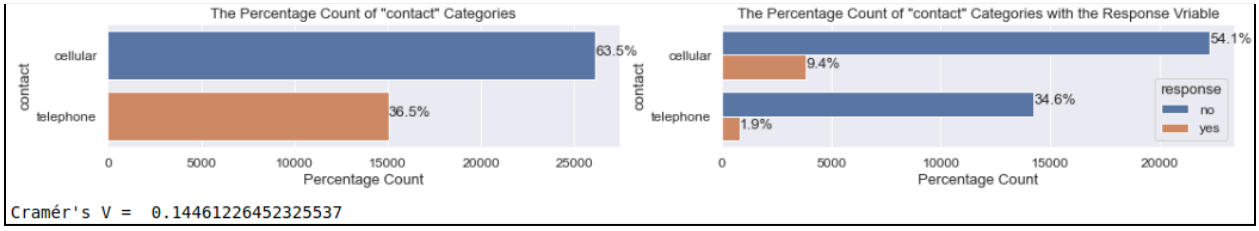



```
Cramér's V =  0.15195496226188143
```

```
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.
```

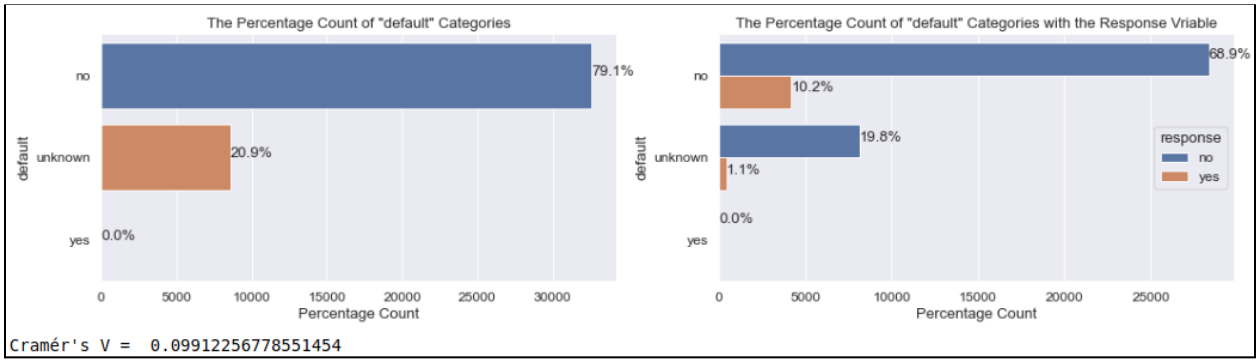
response	no	yes	All	within class subscription per job %
job				
student	600	275	875	31.428571
retired	1284	434	1718	25.261932
unemployed	870	144	1014	14.201183
admin.	9068	1351	10419	12.966695
All	36537	4639	41176	11.266272
management	2596	328	2924	11.217510
unknown	293	37	330	11.212121
technician	6009	730	6739	10.832468
self-employed	1272	149	1421	10.485574
housemaid	954	106	1060	10.000000
entrepreneur	1332	124	1456	8.516484
services	3644	323	3967	8.142173
blue-collar	8615	638	9253	6.895061

- The **contact** feature has the fourth-highest Cramer's V = 0.14. However, the association is also weak. The cellular has received the highest calls and also has the highest within class subscription rate with about 15% subscription of the called group of 26135 customers.

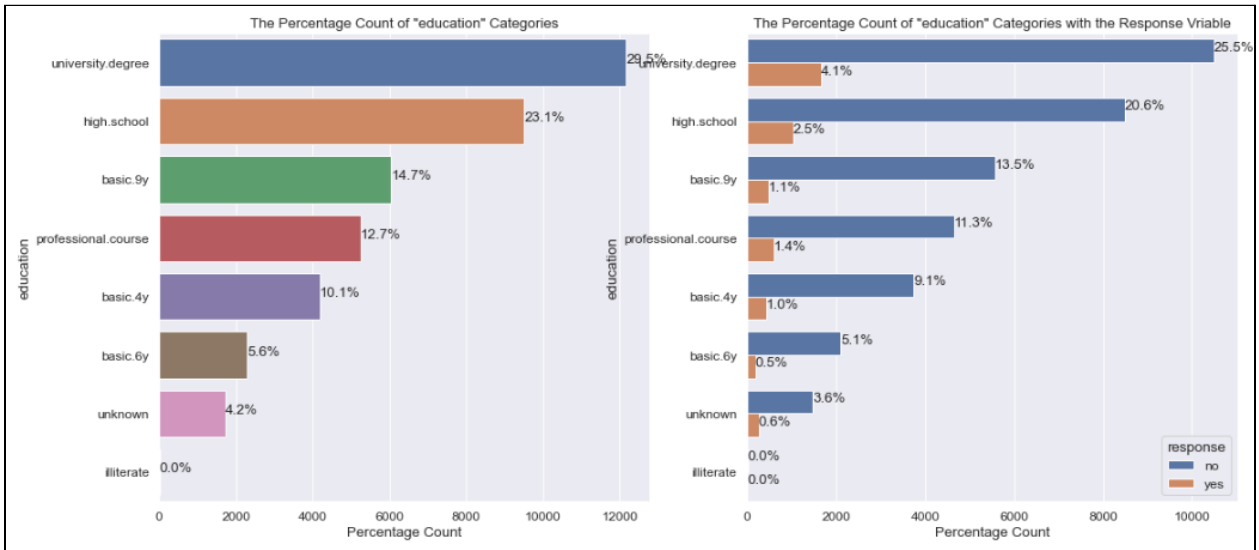


response	no	yes	All	within class subscription per contact %
contact				
cellular	22283	3852	26135	14.738856
All	36537	4639	41176	11.266272
telephone	14254	787	15041	5.232365

- The **default** feature has the fifth-highest Cramer's V = 0.1; indicates a weak association with the response variable.
- The known non-defaulted customers tend to have more subscriptions with a 12.9% within-class subscription rate. However, about 20.9% of default statuses are unknown. The known 3 defaulted customers did not subscribe. The unknown default status has a 5.1% within-class subscription rate. Customers of this class can be defaulted or not.



- The **education** feature has the sixth-highest Cramer's V = 0.07, which indicates a weak association with the response variable.
- Customers who hold a university degree are the most contacted, they represent 29.5% of the customers in this data set. However, this class is ranked third in the within-class subscription rate with 13.7%.
- Surprisingly, the illiterate class has the highest within-class subscription rate with 22.2%. However, only 18 of this class were contacted and 4 of them subscribed. Other than this notice, it seems like higher education has more chances of subscription.



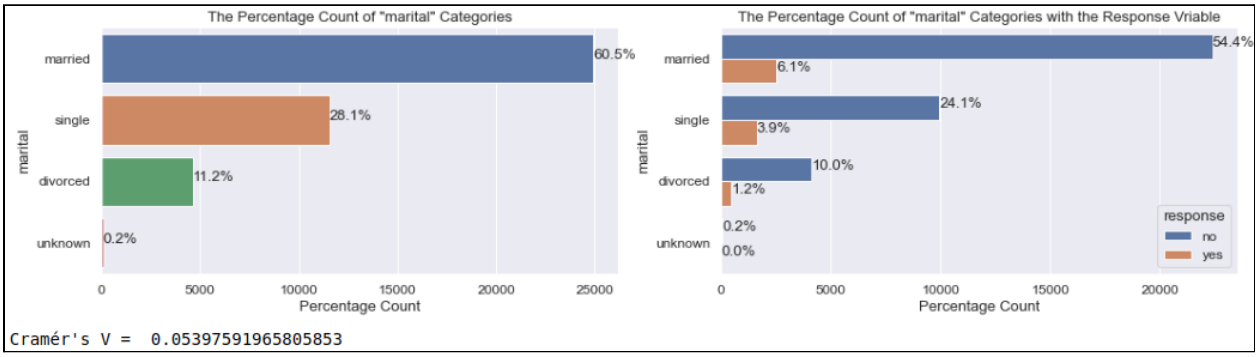

```
Cramér's V =  0.0671834857776241
```

```
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 education %
education				
illiterate	14	4	18	22.222222
unknown	1479	251	1730	14.508671
university.degree	10495	1669	12164	13.720816
professional.course	4645	595	5240	11.354962
All	36537	4639	41176	11.266272
high.school	8481	1031	9512	10.838940
basic.4y	3748	428	4176	10.249042
basic.6y	2103	188	2291	8.206024
basic.9y	5572	473	6045	7.824648

- The **marital** feature has a weak association with the response variable, Cramer's V = 0.05.

- 60.5% of the called customers are married. However, the within-class subscription rate of married customers is the lowest with 10.6%. Single customers have a higher within-class subscription rate of 14%. 80 customers have unknown marital status.



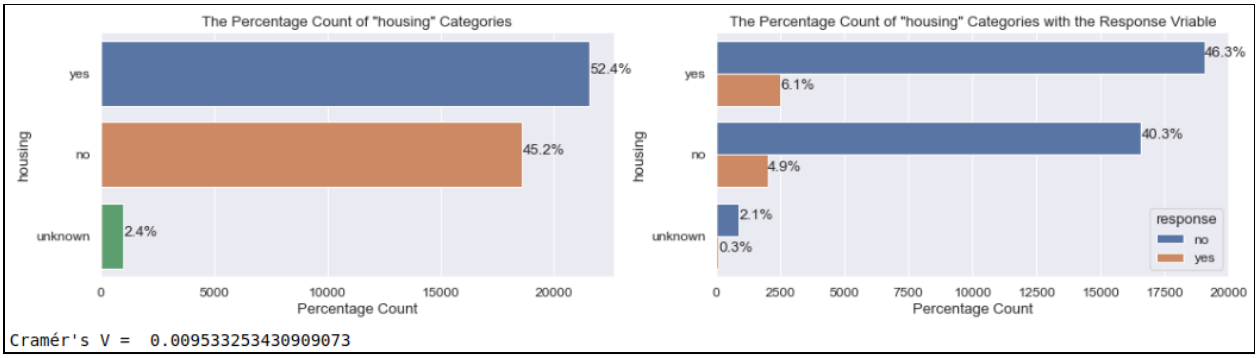

```
Cramér's V =  0.05397591965805853
```

```
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 marital %
marital				
unknown	68	12	80	15.000000
single	9944	1620	11564	14.008993
All	36537	4639	41176	11.266272
divorced	4135	476	4611	10.323140
married	22390	2531	24921	10.156093

- The **day_of_week** feature has a weak association with the response variable, Cramer's V = 0.02.
- The within-class subscription rate of day_of_week is mostly the same for all days, Thursday is the highest with 12.1% and Monday is the lowest with 10%.

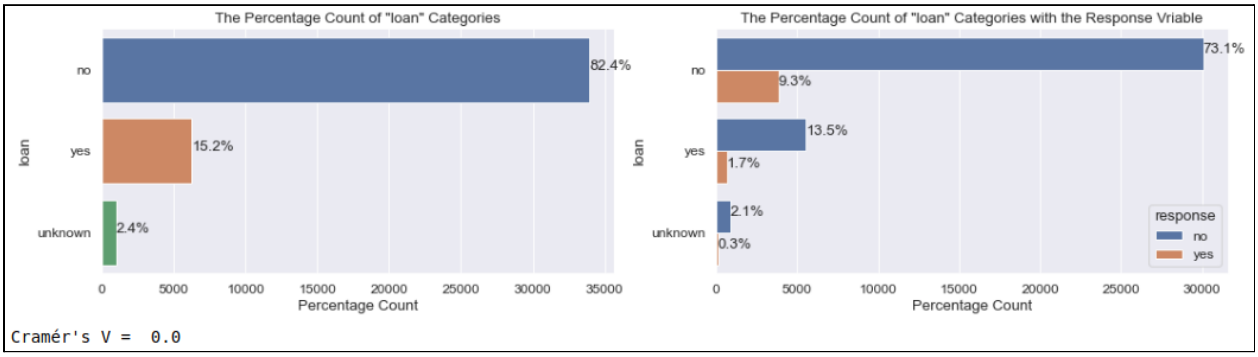
- The **housing** feature has a weak association with the response variable, Cramer's V= 0.0095.
- The within-class subscription rate is mostly the same for all classes, around 11%. There are 990 customers (2.4%) with unknown housing status.




```
Cramér's V =  0.009533253430909073
```

response	no	yes	All	within class subscription per housing %
housing				
yes	19064	2507	21571	11.622085
All	36537	4639	41176	11.266272
no	16590	2025	18615	10.878324
unknown	883	107	990	10.808081

- The housing feature has no association with the response variable, Cramer's V = 0.0.
- The within-class subscription rate is mostly the same for all loan classes including the 990 customers with unknown loan status.



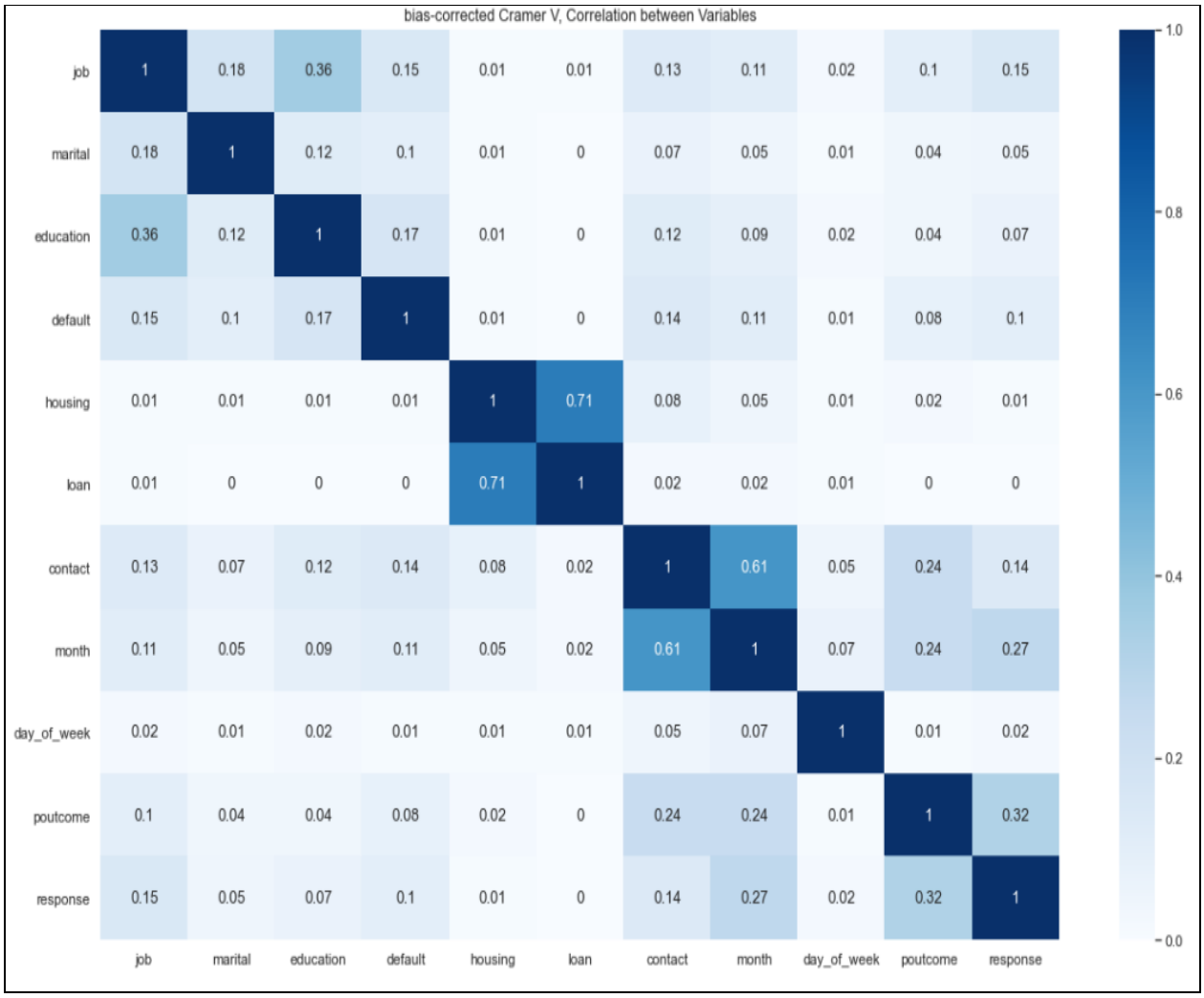

```
Cramér's V =  0.0
```

```
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 loan %
loan				
no	30089	3849	33938	11.341269
All	36537	4639	41176	11.266272
yes	5565	683	6248	10.931498
unknown	883	107	990	10.808081

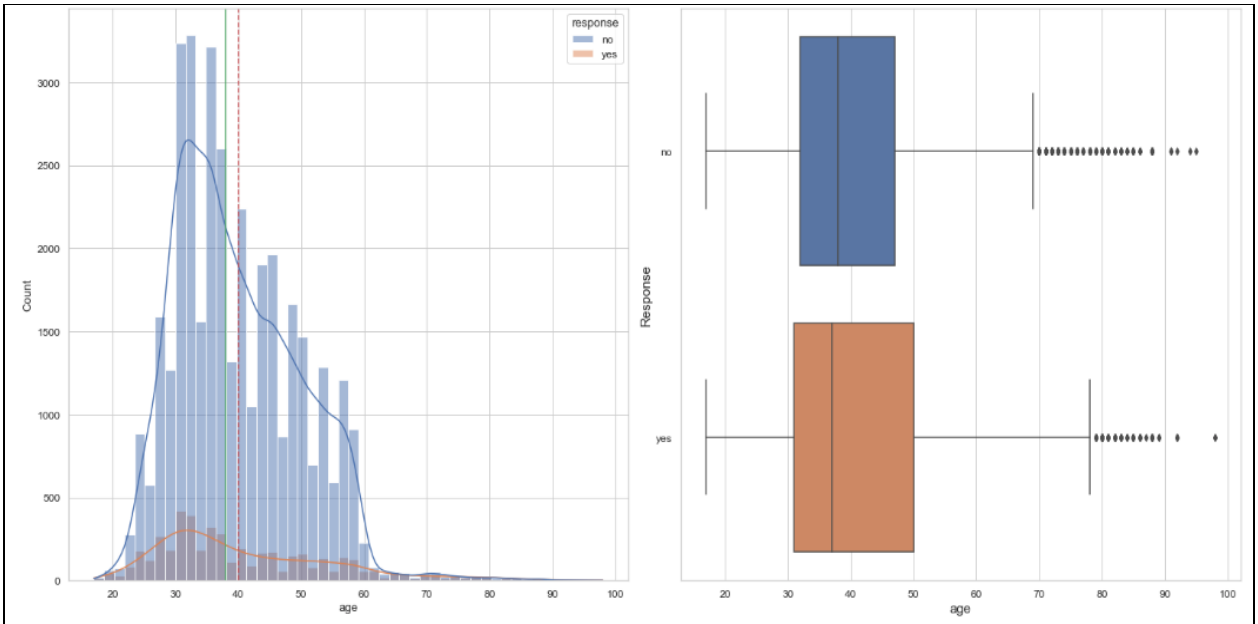
Correlation between all of the categorical variables

- 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.
- The association with the response variable is discussed above.



2. Numerical Features

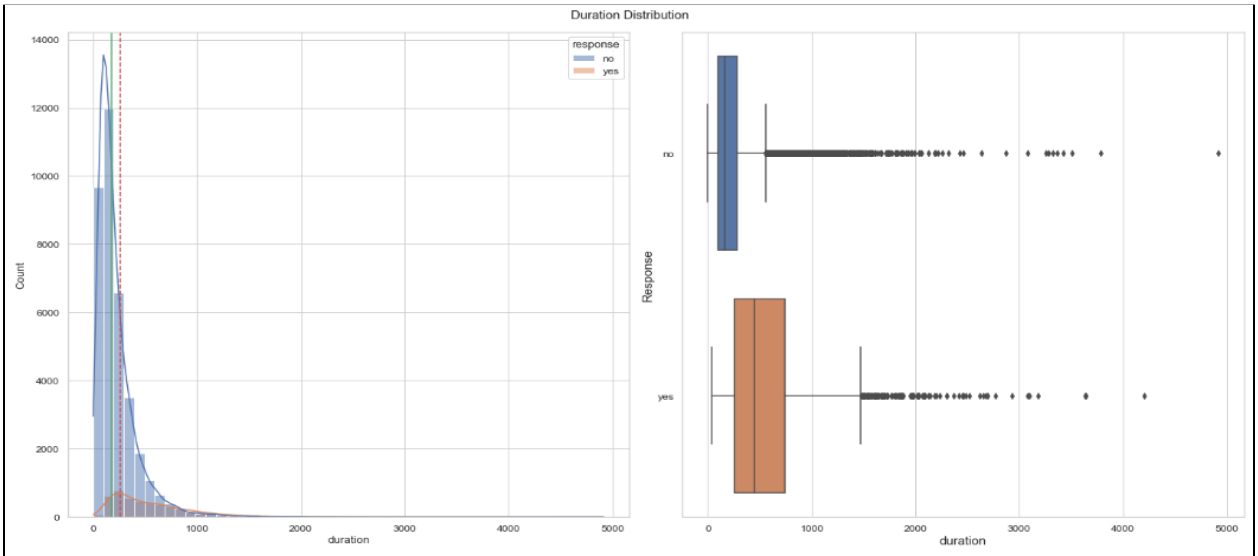
- The **age** distribution for both classes of the target variable is mostly overlapping. The median age for subscribers is 37 and none subscribers are 38.
- The maximum age is 98 and the minimum is 17.
- The age feature seems not to be a good indicator of the target variable.
- Outliers appear in both classes.



The age description for for both respnse classes =

response	age 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

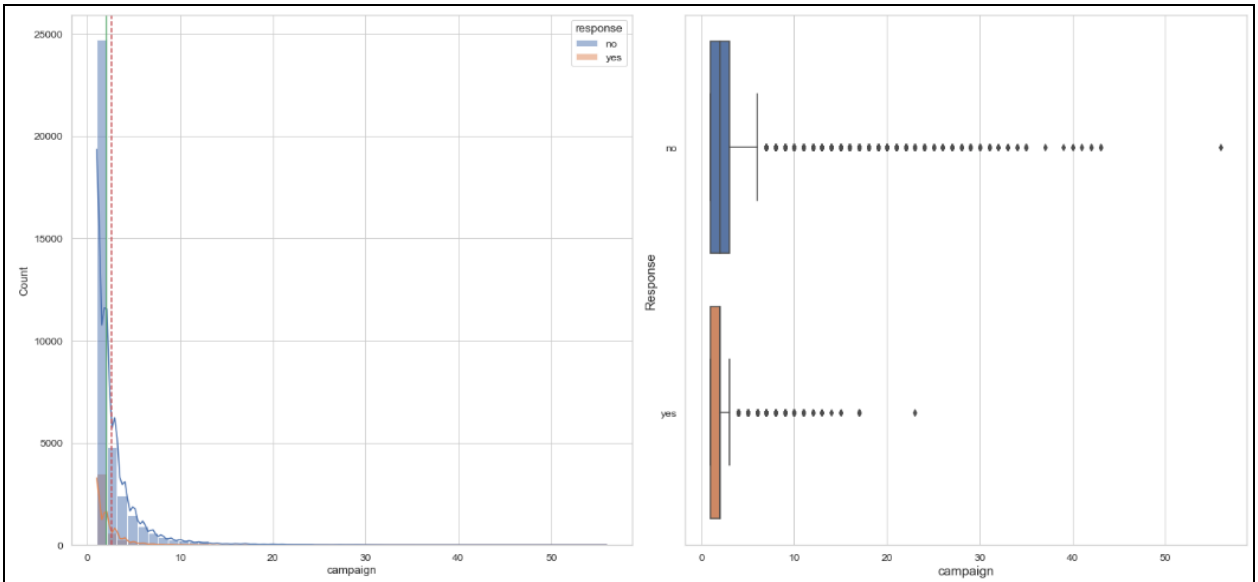
- **Duration.** The box plot below shows that subscribed customers had a longer call duration. The duration seems to be correlated with the response variable.
- However, the call duration is known after the call and the outcome of the call is likely to be known. Thus, this feature is not available as a predictor for prediction before the call is made.



The duration description for for both respnse classes =

response	duration count	mean	std	min	25%	50%	75%	max
no	36537.0	220.868079	207.116178	0.0	95.0	164.0	279.0	4918.0
yes	4639.0	553.256090	401.190736	37.0	253.5	449.0	741.5	4199.0

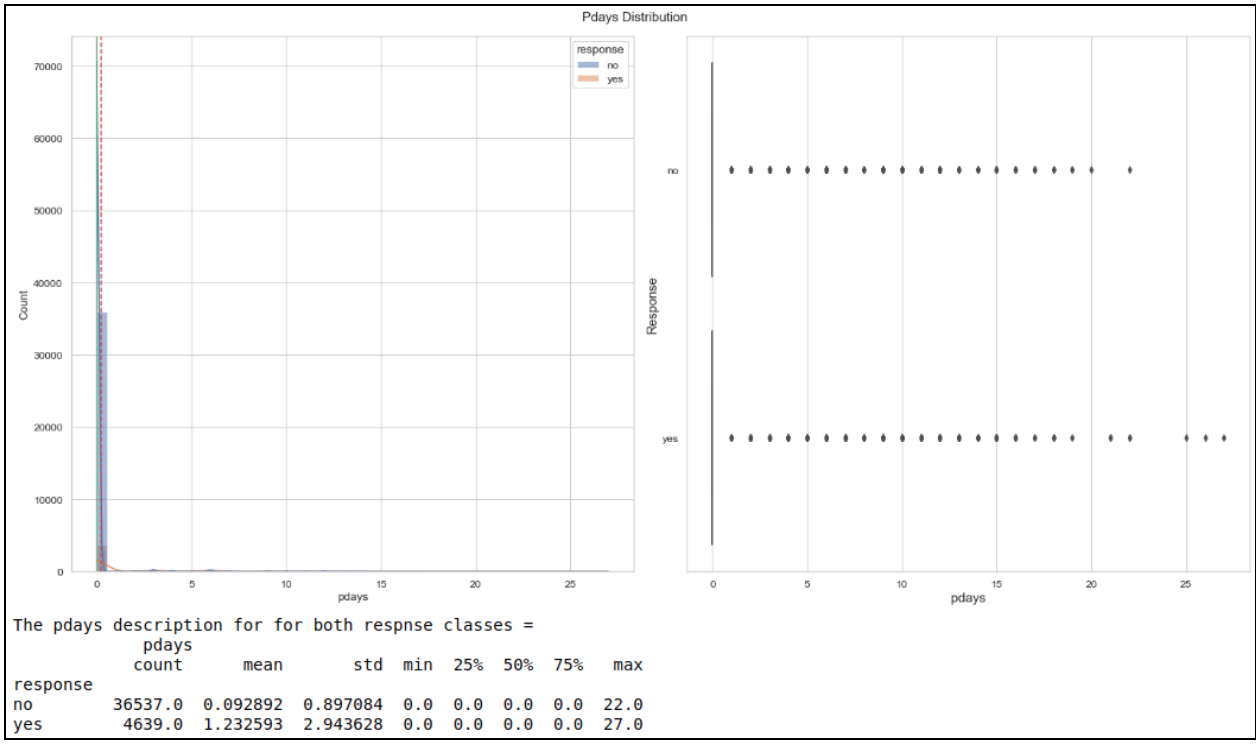
- The **campaign** feature refers to the number of contacts performed during this campaign and for this client (includes the last contact).
- More campaign calls for a particular customer did not lead to the service subscription. The plots show that most of the subscribed customers received one or two calls.
- Also, most of the none subscribed customers has received a fewer call.
- This feature does not seem a strong predictor.



The campaign description for for both respnse classes =

	campaign							
	count	mean	std	min	25%	50%	75%	max
response								
no	36537.0	2.633385	2.873768	1.0	1.0	2.0	3.0	56.0
yes	4639.0	2.051951	1.666353	1.0	1.0	2.0	2.0	23.0

- The **pdays** feature refers to the number of days that passed by after the client was last contacted from a previous campaign.
- Most of the clients were not contacted before.
- This feature by itself can not be a predictor of the response variable.

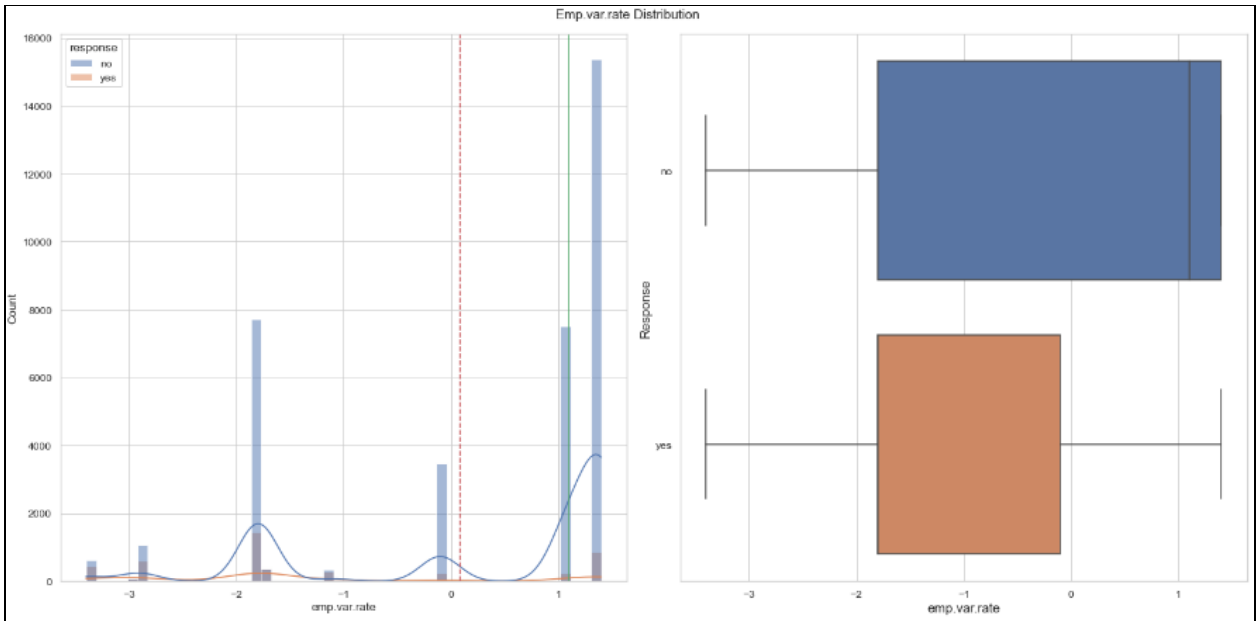



```
The pdays description for for both respnse classes =
            pdays
            count      mean       std  min  25%  50%  75%   max
response
no        36537.0  0.092892  0.897084  0.0  0.0  0.0  0.0  22.0
yes        4639.0  1.232593  2.943628  0.0  0.0  0.0  0.0  27.0
```

- The **previous** feature refers to the number of contacts performed before this campaign and for this client.
- Only 13.7% of the clients were previously contacted. The highest number of contacts is 7 and it only happened one time.
- As the table shows, increasing the number of contacts did not lead to more subscriptions.

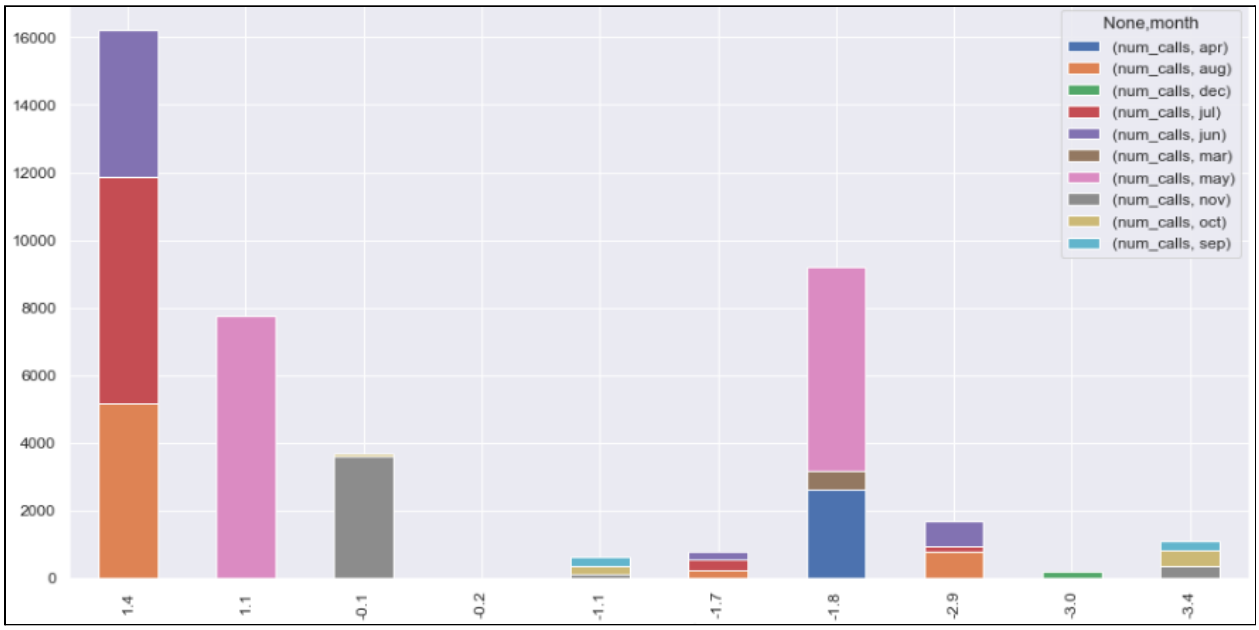
previous	0	1	2	3	4	5	6	7
response								
no	32411.0	3594.0	404.0	88.0	32.0	5.0	2.0	1.0
yes	3140.0	967.0	350.0	128.0	38.0	13.0	3.0	NaN

- The **emp.var.rate** feature refers to employment variation rate — quarterly indicator.
- 10 distinct rates were reported during the campaign. It seems like most of the campaign time happened when the employment variation rate is negative; in 4 months out of 10, this rate was positive.
- From the graph, the campaign calls focused more on the time with a positive rate. However, a higher subscription rate per call was reported when the employment variation rate -1.7 = 52% and -3.0 = 51%. Only 5% of the called customers subscribed when emp.var.rate = 1.4 and 3% when it was (1.1).
- The higher rate did not lead to more subscriptions.



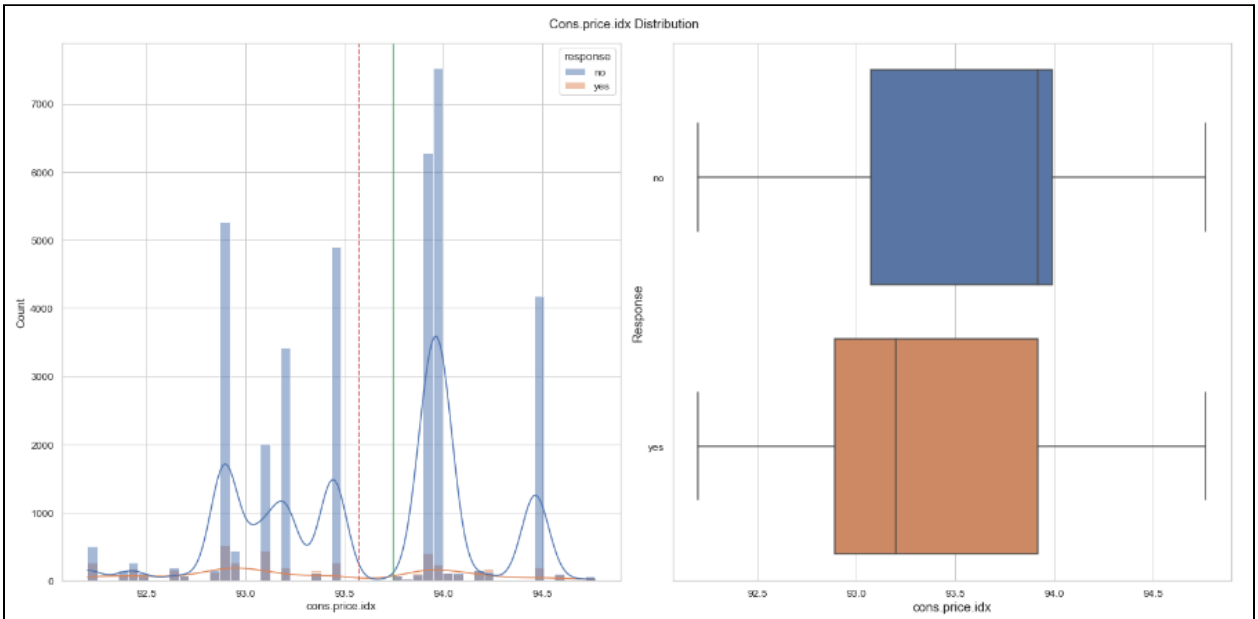
The emp.var.rate description for for both respnse classes =

	emp.var.rate							
	count	mean	std	min	25%	50%	75%	max
response								
no	36537.0	0.248885	1.482873	-3.4	-1.8	1.1	1.4	1.4
yes	4639.0	-1.233089	1.623616	-3.4	-1.8	-1.8	-0.1	1.4



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

- The **cons.price.idx** feature is consumer price index — monthly indicator.
- The box plot shows no relationship between this feature and the response variable. There is an overlap of the two response classes across most values of cons.price.idx



The cons.price.idx description for for both respnse classes =

	cons.price.idx						
	count	mean	std	min	25%	50%	75%
response							
no	36537.0	93.603798	0.559006	92.201	93.075	93.918	93.994
yes	4639.0	93.354577	0.676592	92.201	92.893	93.200	93.918

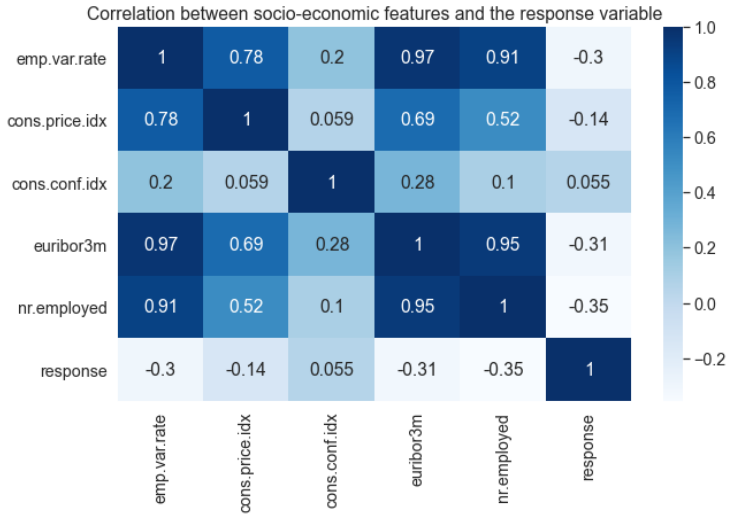
- Similarly, '**cons.conf.idx**', '**euribor3m**', '**nr.employed**', have no relationship with the response variable. There is an overlap of the two response classes across most values of the socio-economic features.

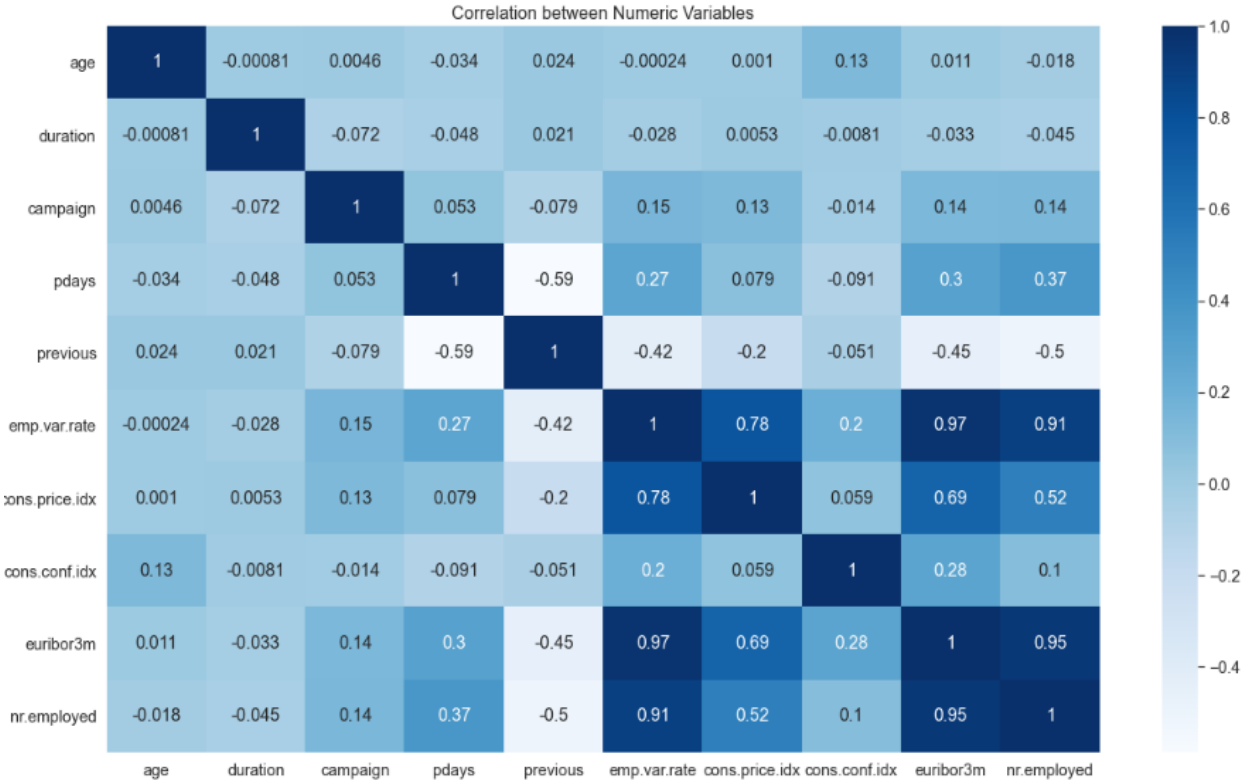
- The correlation matrix of Pearson correlation below shows that all socio-economic features are correlated with each other except the cons.conf.idx.

- As we see from the correlation analysis below, when changing the target variable to numeric of (1 and 0), we found that none of the socio-economic features correlates with the response variable.

Correlation between all of the categorical variables (Pearson Correlation)

- We use here Pearson (r): standard correlation coefficient to examine the linear relationship between the numerical variables.





- Most of the social and economic context attributes are highly correlated.
- euribor3m is highly correlated with emp.var.rate (r = 0.79), and with cons.price.idx (r = 0.69).
- emp.var.rate is highly correlated with euribor3m and with nr.employed (r = 0.91). Also, it has a strong relationship with cons.price.idx (r = 0.78)

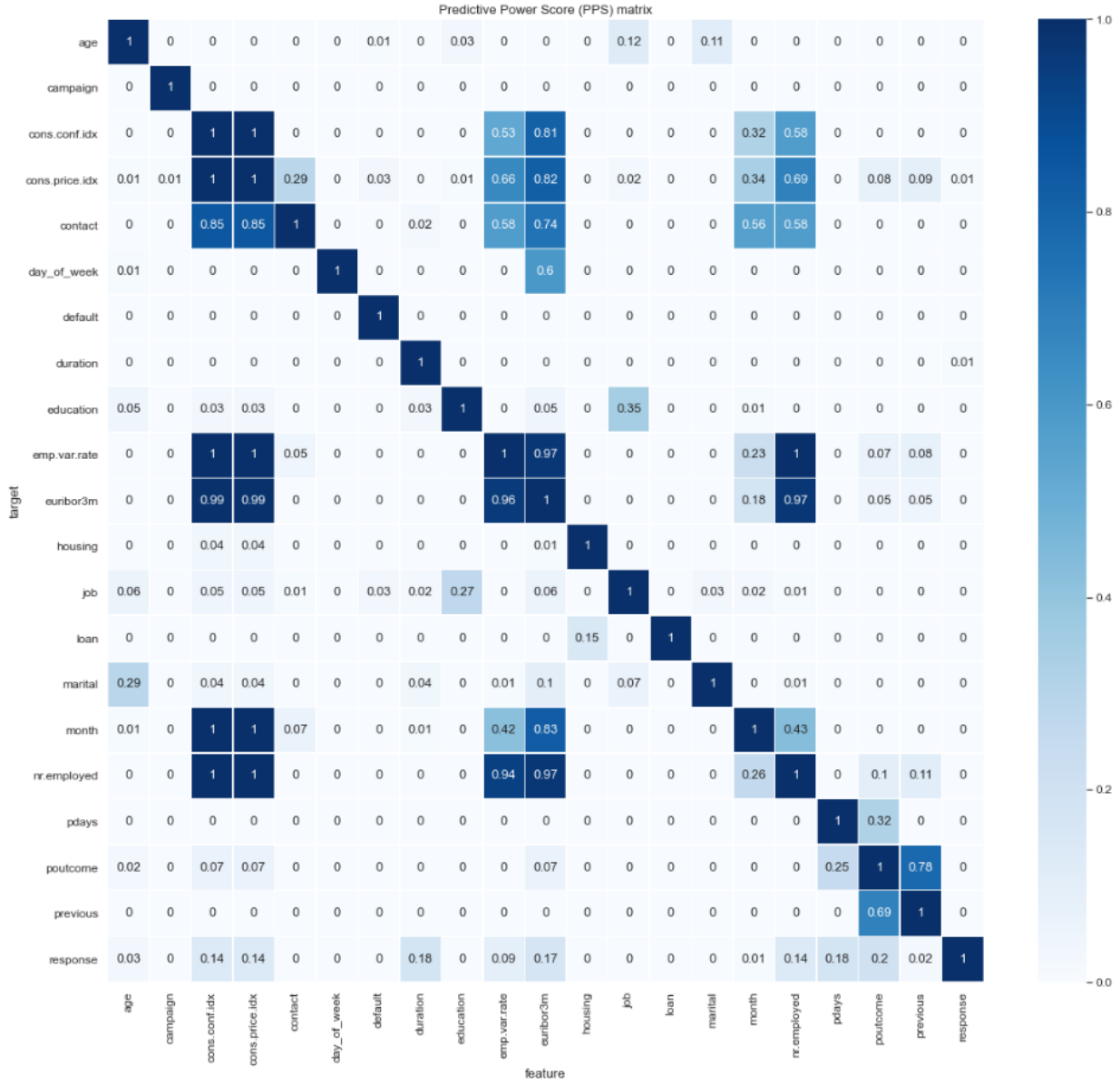
- nr.employed is highly correlated with emp.var.rate (r = .91) and with euribor3m (r = 0.95). Also, it has a relationship with cons.price,idx (r = 0.52).

The Predictive Power Score (PPS)

- Another approach to finding the correlation through calculating the Predictive Power Score (PPS). Among the issues of the previous approaches is that Cramers V works with categorical data and Pearson correlation works with numeric data. Both approaches indicate symmetry of the relationship, which is not always the case.
- The PPS is an asymmetric, data-type-agnostic score that can detect linear or non-linear relationships between two columns. The score ranges from 0 (no predictive power) to 1 (perfect predictive power). It can be used as an alternative to the correlation (matrix).
- The PPS approach is a way to find the correlation or the predictive power between the numerical and categorical variable which is not possible via Pearson Correlation. The table below shows the predictive power of all variables with the response variable using PPS.

	Feature	ppscore	y
1	poutcome	0.20	response
2	pdays	0.18	response
3	duration	0.18	response
4	euribor3m	0.17	response
5	nr.employed	0.14	response
6	cons.conf.idx	0.14	response
7	cons.price.idx	0.14	response
8	emp.var.rate	0.09	response
9	age	0.03	response
10	previous	0.02	response
11	month	0.01	response
12	marital	0.00	response
13	default	0.00	response
14	housing	0.00	response
15	campaign	0.00	response
16	job	0.00	response
17	day_of_week	0.00	response
18	contact	0.00	response
19	loan	0.00	response
20	education	0.00	response

- Below is the PPS matrix that shows the asynchronous pairwise predictive power between all variables (categorical and numerical).



- The output of PPS shows that there are features have sort of predictive and ranked from the highest to the lowest as follows:
    - 'poutcome',
    - 'pdays',
    - 'duration',
    - 'euribor3m',
    - 'cons.price.idx',
    - 'cons.conf.idx',

    - 'nr.employed',
    - 'emp.var.rate',
    - 'age',
    - 'previous',
    - 'month'.
- PPS agrees with Cramers V that **poutcome** has the highest predictive power, PPS = 0.2 & V = 0.32. poutcome refers to the outcome of the previous marketing campaign
- PPS revealed the predictive power (the correlation) between numeric and categorical variables, which cannot be found by any of the other methods.
- PPS confirmed the correlation between most of the social and economic context attributes ('euribor3m', 'cons.price.idx', 'cons.conf.idx', 'nr.employed'). Thus, it makes sense that they have similar PPS with the response variable around 14.

- However, none of the features by itself indicate a strong predictive power. None of these methods account for the interaction between the variables. The interaction can be verified by Modeling.

- The subscribers and non-subscriber classes overlap across most of the variables. This is a challenging classification problem.

Data Preprocessing

- This part involved handling the categorical variable for modeling :
    - The response variable transformed to a binary variable using Label Encoder.
    - One Hot Encoding is used to Transform the Remaining Categorical Variable.

```
print('The new features are {}:\n\n {}'.format(len(df_encoded.columns),  encoder.get_feature_names()))
```

```
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', 'ma
rital_married', 'marital_single', 'marital_divorced', 'marital_other', 'education_basic.4y', 'education_hig
h.school', 'education_basic.6y', 'education_basic.9y', 'education_professional.course', 'education_other',
'education_university.degree', 'education_illiterate', 'default_no', 'default_yes', 'housing_no', 'housing_y
es', '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_mo
n', 'day_of_week_tue', 'day_of_week_wed', 'day_of_week_thu', 'day_of_week_fri', 'duration', 'campaign', 'pda
ys', 'previous', 'poutcome_nonexistent', 'poutcome_failure', 'poutcome_success', 'emp_var_rate', 'cons_price
_idx', 'cons_conf_idx', 'euribor3m', 'nr_employed', 'response']
```

    - We ended with 61 features including the response variable. By dropping the **duration** feature, the total is 60 columns.
- The resulted data is splitted into Training (70%) and Test (30%) sets.

```
y = df_encoded.response
X = df_encoded.drop('response', axis=1)

from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42, stratify=y)

data_sets =[X_train, X_test, y_train, y_test]

for d in data_sets:
    print(f'The shape of {get_df_name(d)}: {d.shape} {round(d.shape[0]/df_encoded.shape[0] *100, 0)}% \n')
```

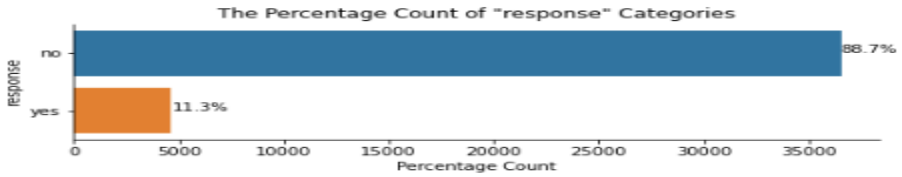
```
The shape of X_train: (28823, 59) 70.0%

The shape of X_test: (12353, 59) 30.0%

The shape of y_train: (28823,) 70.0%

The shape of y_test: (12353,) 30.0%
```

- Over-Sampling Using SMOTE
  - There is a class imbalance in the original dataset.



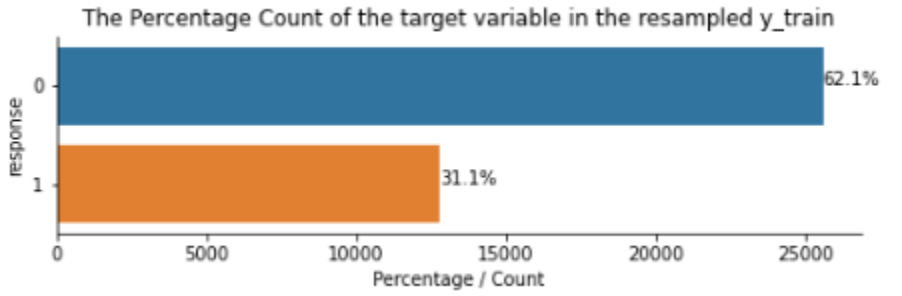
  - Synthetic Minority Oversampling Technique, or SMOTE for short is a method for oversampling the minority class. SMOTE allows augmenting new examples synthesized from the existing examples. This is a type of data augmentation for the minority class. Only the training sets were oversampled.
  - Oversampling made the classes ratio 50% to limit augmenting training artificial data. Many classifiers can handle such reduced imbalance.

```
from imblearn.over_sampling import SMOTE

# Adding data points to the minority calss to be increased to the half of the majority class.
oversample = SMOTE(sampling_strategy=0.5, n_jobs=-1, random_state=42)
X_train, y_train = oversample.fit_resample(X_train, y_train)
```

```
print('The resampled training data new classes ratio is:{}%'.format( y_train.value_counts()[1] / y_train.value_counts()[0] * 100))
```

```
The resampled training data new classes ratio is:50.0%
```



Modeling

- The following classification algorithms are used:
    - Logistic Regression
    - Decision Tree
    - Random Forest
    - Light Gradient Boosting
    - XGboost

- Bayesian hyperparameter optimization, using Tree-structured Parzen Estimator Approach (TPE), from the hyperopt package is used to optimize the hyperparameters of each model.
- Various objective functions were applied to the various classification algorithms. Recall_micro gave more balanced performance for several models giving the business objective.
- Models Performance:
    - **The objective** is to maximize targeting customers who are likely to subscribe. The true positive rate is more important in this case, in particular in the subscribed class. Thus, more attention was made to the Recall score.

$$Precision = \frac{TP}{TP+FP} \;\&\; Recall = \frac{TP}{TP+FN}$$

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

    - Precision implied as the measure of the correctly identified positive cases from all the predicted positive cases. Thus, it is useful when the cost of False Positives is high.
    - Recall the measure of the correctly identified positive cases from all the actual positive cases. It is important when the cost of False Negatives is high.
    - The tale below shows a comparison of the models' performance by testing the models on the test dataset.

		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_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	0.8	0.8	0.8	0.8

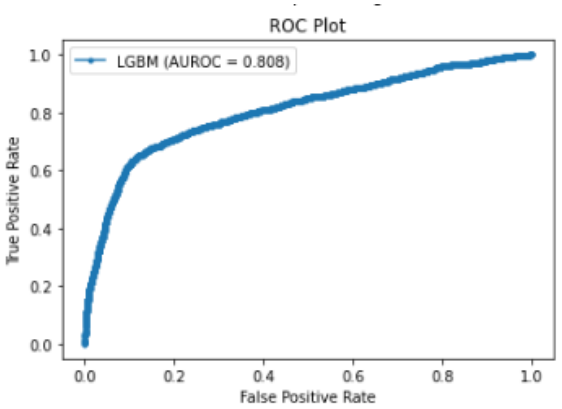
Conclusion

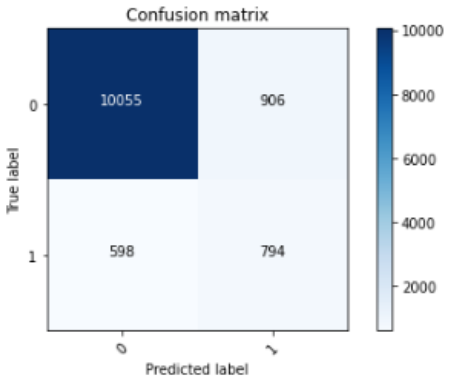
- The Light Gradient Boosting algorithm optimized for Recall Micro gave the best performance. ROC AUC = 0.81 and F1= 0.51 one minority class, (waited avg F1= 0.88).
- The LGB model gave ROC AUC = 0.81 on the test dataset which is slightly more than the 0.8 reported by the published paper using Neural Networks Model (S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014).
- This model showed a higher chance of predicting the minority class (subscribers) with Recall of 0.51 on this class.
- 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.

The Light Gradient Boosting with the following parameters is the winner .

```
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)
```

Model performance:

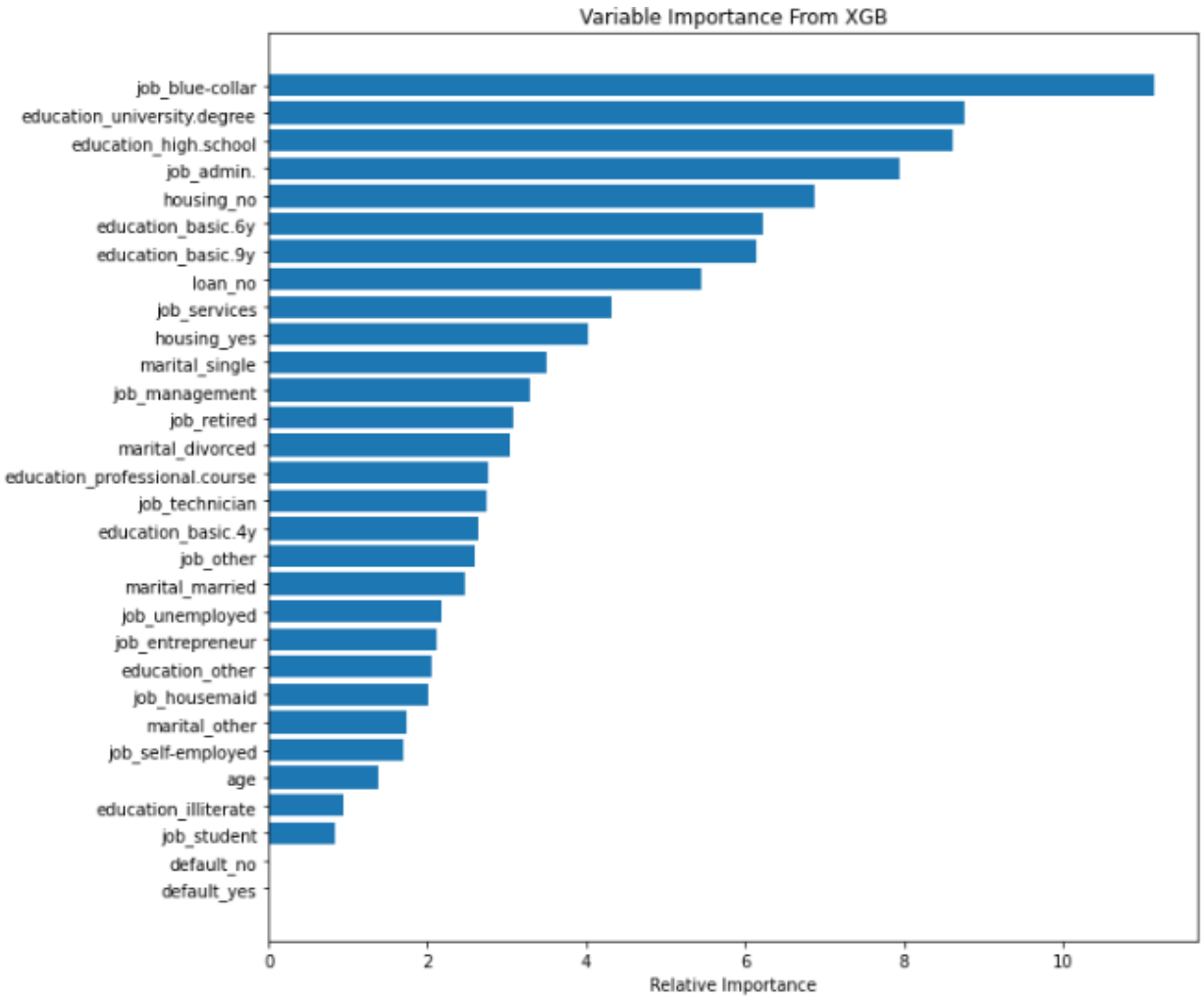




- Features Importance

  The default feature importance with LGBM package is as the following is shown in the figure below. Literature and many blogs talked about an issue of bias towards high cardinality variables with the tree based models. This issue seems to be reflected in this case.

  Another method of measuring the feature importance is through **Permutation Importance**. eli5 package website states: "eli5 provides a way to compute feature importances for any black-box estimator by measuring how score decreases when a feature is not available; the method is also known as “permutation importance” or “Mean Decrease Accuracy (MDA)”. One way to do permutation importance is shuffling values for a feature and assessing the model performance.



- Permutation Importance

The methods gave different outcomes. One can lean more toward the permutation importance in this case. The outcome of permutation importance is shown below.

```
import eli5
from eli5.sklearn import PermutationImportance

perm = PermutationImportance(lgbm_model, random_state=1).fit(X_test, y_test)
eli5.show_weights(perm, feature_names = X_test.columns.tolist())
```

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 …	

Future work

- Invest more time on Hyperparameter tuning using more computational resources.
- Investigate more and compare feature importance for various techniques. SHAP (SHapley Additive exPlanations) package is worth checking (https://github.com/slundberg/shap).
- One can compare applying SMOTE oversampling with under sampling and without sampling.
- Deploy the selected model on a web API.