Exploratory Data Analysis

Bank Marketing Campaign

Name: Samuel Alejandro Cueva Lozano
Email: samuelcl7@gmail.com
Country: Peru
Specialization: Data Science

Agenda

Executive Summary

Problem Description

EDA

Feature Selection

Model Recommendations

Executive Summary

Client: ABC bank: Portuguese banking institution

Problem Description: ABC Bank wants to sell it's term deposit product to customers and before launching the product they want to know whether a particular customer will buy their product or not (based on customer's past interaction with bank or other Financial Institution).

Business goal: Shortlist which customers have more chances to subscribe to the term deposit.

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

Exploratory Data Analysis : Statistical data analysis will be performed on the Dataset

- Descriptive analysis (univariate analysis)
- Correlation analysis (bivariate analysis)
  - Qualitative analysis
  - Quantitative analysis
- Feature selection and engineering
  - Feature selection based on descriptive analysis
  - Feature selection based on correlation analysis
- Recommended models

Data

The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls.

There are three files:

- bank-additional-full.csv with all examples (41188) and 20 inputs, ordered by date (from May 2008 to November 2010).
- bank-additional.csv with 10% of the examples (4119), randomly selected from bank-additional-full.csv, and 20 inputs.
- bank-additional-names.txt with information about the attributes.

```
../Project_data/
└── bank-additional
    ├── bank-additional.csv
    ├── bank-additional-full.csv
    └── bank-additional-names.txt

1 directory, 3 files
```

There are two files with data but since bank-additional.csv is in bank-additional-full.csv, this file will be ignored.

Attributes or features

Input features

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).

Attributes or features

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')

Descriptive Analysis

Data types and missing values

Numerical attributes
- Quantile Statistics
- Descriptive Statistics
- Distribution histograms
- Outliers and tranformations
- Feature Selection

Categorical attributes
- Cardinality
- Unique counts
- Feature selection

Data Types and Missing values

	Pandas types	Python types	Number of records	Number of missing values	% of missing values
age	int64	int	41188	0	0.0
job	object	str	41188	0	0.0
marital	object	str	41188	0	0.0
education	object	str	41188	0	0.0
default	object	str	41188	0	0.0
housing	object	str	41188	0	0.0
loan	object	str	41188	0	0.0
contact	object	str	41188	0	0.0
month	object	str	41188	0	0.0
day_of_week	object	str	41188	0	0.0
duration	int64	int	41188	0	0.0
campaign	int64	int	41188	0	0.0
pdays	int64	int	41188	0	0.0
previous	int64	int	41188	0	0.0
poutcome	object	str	41188	0	0.0
emp.var.rate	float64	float	41188	0	0.0
cons.price.idx	float64	float	41188	0	0.0
cons.conf.idx	float64	float	41188	0	0.0
euribor3m	float64	float	41188	0	0.0
nr.employed	float64	float	41188	0	0.0
y	object	str	41188	0	0.0

No missing values found in the data set

Duplicate values

These records are duplicates and will be removed

	age	job	marital	education	default	housing
1266	39	blue-collar	married	basic.6y	no	no
12261	36	retired	married	unknown	no	no
14234	27	technician	single	professional.course	no	no
16956	47	technician	divorced	high.school	no	yes
18465	32	technician	single	professional.course	no	yes
20216	55	services	married	high.school	unknown	no
20534	41	technician	married	professional.course	no	yes
25217	39	admin.	married	university.degree	no	no
28477	24	services	single	high.school	no	yes
32516	35	admin.	married	university.degree	no	yes
36951	45	admin.	married	university.degree	no	no
38281	71	retired	single	university.degree	no	no

Note: only some attributes are displayed due to the table size.

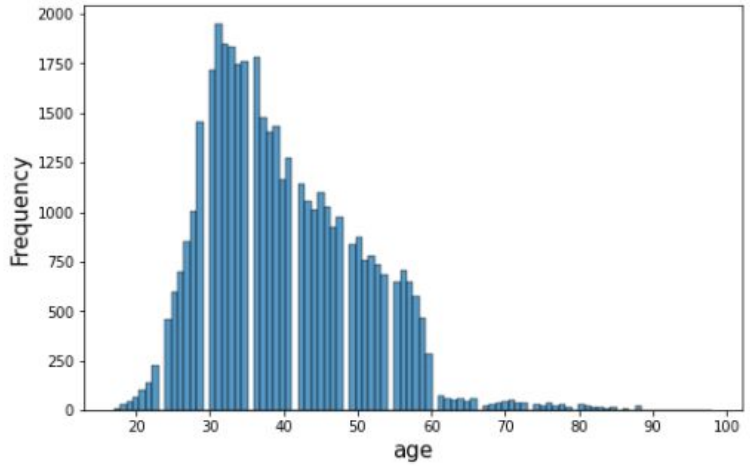
Descriptive Statistics

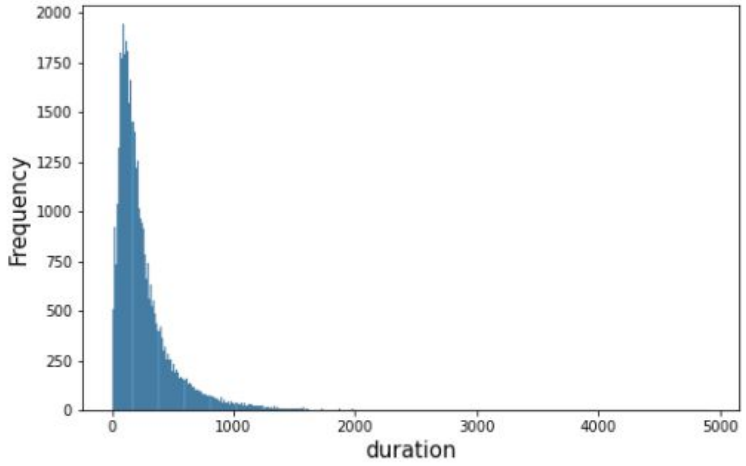
Low variance(std^2)

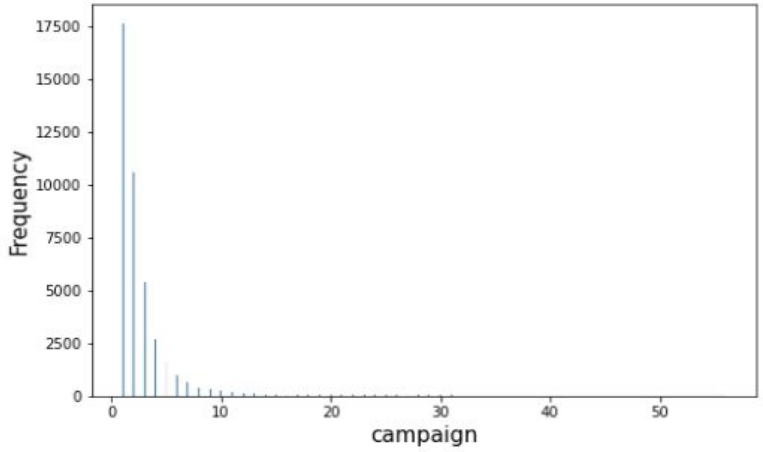
	age	duration	campaign	pdays	previous	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed
count	41176.00000	41176.000000	41176.000000	41176.000000	41176.000000	41176.000000	41176.000000	41176.000000	41176.000000	41176.000000
mean	40.02380	258.315815	2.567879	962.464810	0.173013	0.081922	93.575720	-40.502863	3.621293	5167.034870
std	10.42068	259.305321	2.770318	186.937102	0.494964	1.570883	0.578839	4.627860	1.734437	72.251364
min	17.00000	0.000000	1.000000	0.000000	0.000000	-3.400000	92.201000	-50.800000	0.634000	4963.600000
25%	32.00000	102.000000	1.000000	999.000000	0.000000	-1.800000	93.075000	-42.700000	1.344000	5099.100000
50%	38.00000	180.000000	2.000000	999.000000	0.000000	1.100000	93.749000	-41.800000	4.857000	5191.000000
75%	47.00000	319.000000	3.000000	999.000000	0.000000	1.400000	93.994000	-36.400000	4.961000	5228.100000
max	98.00000	4918.000000	56.000000	999.000000	7.000000	1.400000	94.767000	-26.900000	5.045000	5228.100000
IQR	15.00000	217.000000	2.000000	0.000000	0.000000	3.200000	0.919000	6.300000	3.617000	129.000000

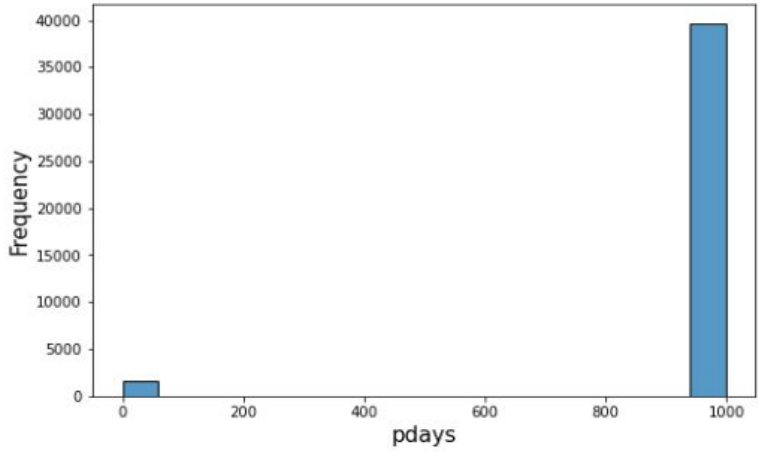
Potential outliers

Histograms

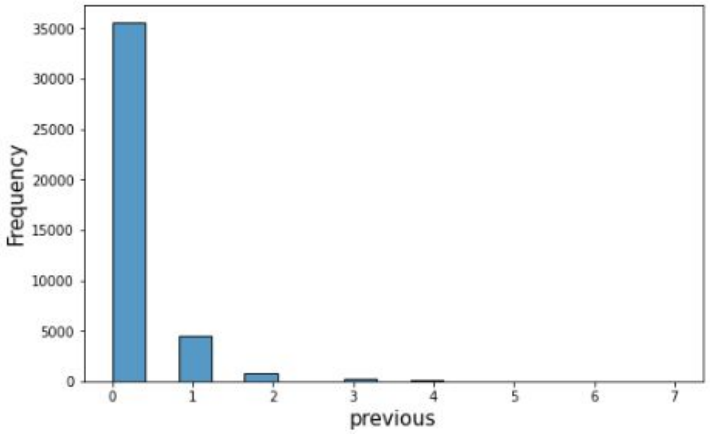


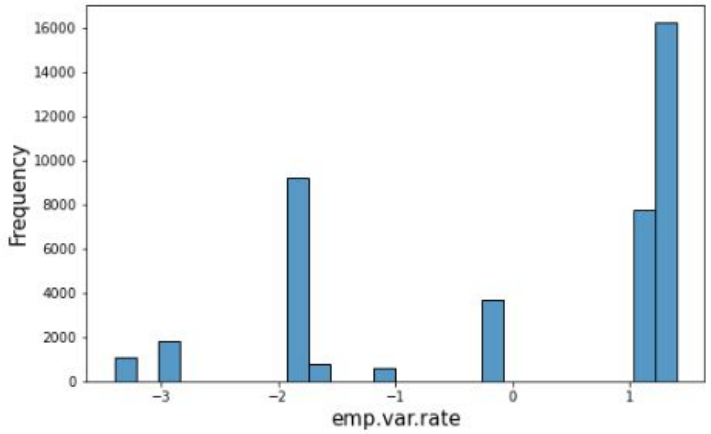


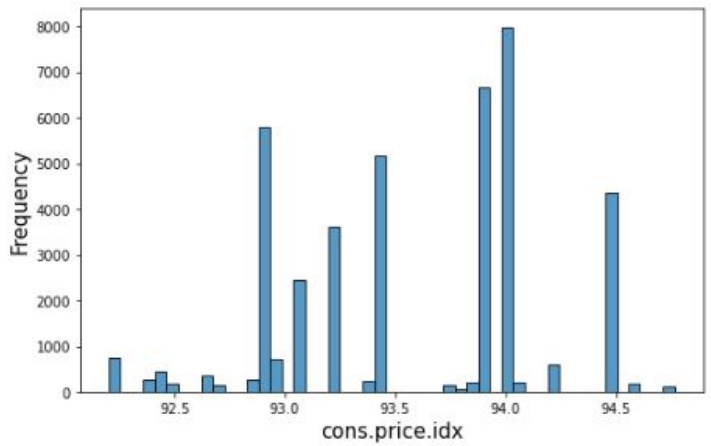


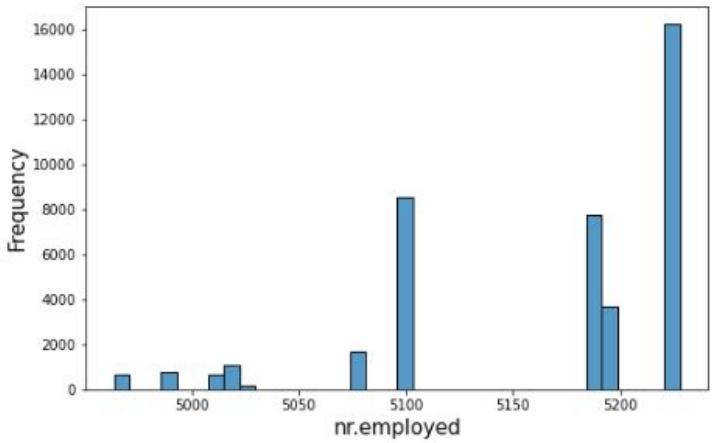


Histograms









In the graphics you can see the following:

- The **duration** and **campaign** attributes have a skewed distribution which may be due to outliers.
- The **age** attribute has a distribution that seems normal, but it also has a long tail shape. This may be because retired people are not contacted often.
- Most of the people weren't contacted before, that's why the **pdays** attribute has that distribution.
- The **previous** attribute has low variance, this means that most people were contacted between 1 and 4 times.
- The **emp.var.rate** and **nr.employed** attributes have few different values.

Outliers

Approaches to dealing with outliers

- Visualizations and descriptive statistics to detect potential outliers (done)
- Filtering by fixed threshold
- Clipping the attribute at a computed percentile (99%)
- log of every value
- IQR Score

The attributes with potential outliers are:

age, **duration** and **campaign**

Filtering by fixed threshold

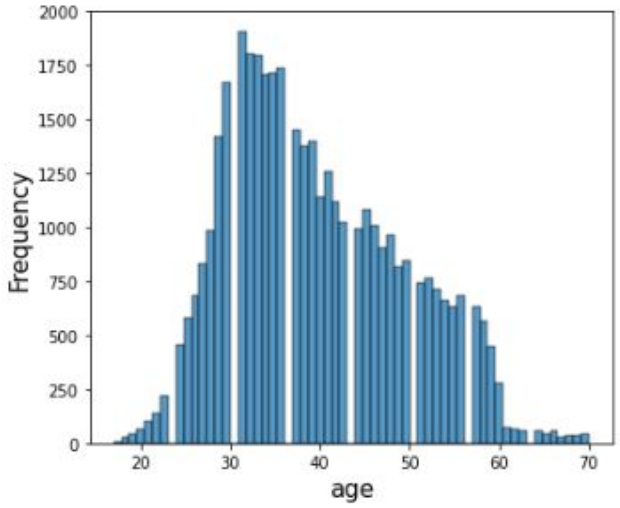
- There is not any attribute in which to apply this approach is reasonable.
- The **age** attribute has a maximum value of 98 and this value is correct.

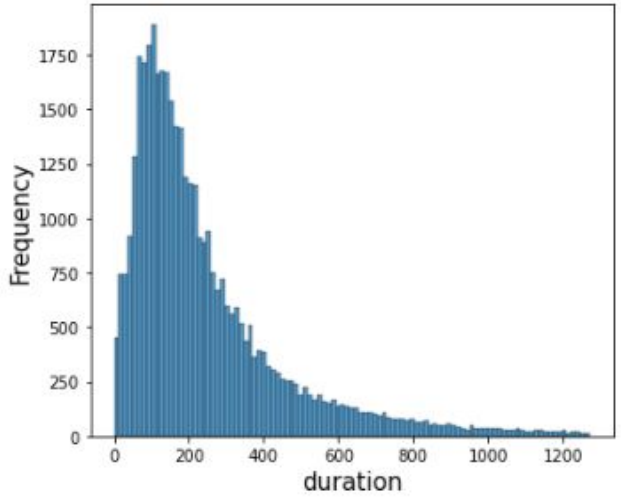
Clipping the attribute at a computed percentile(99%)

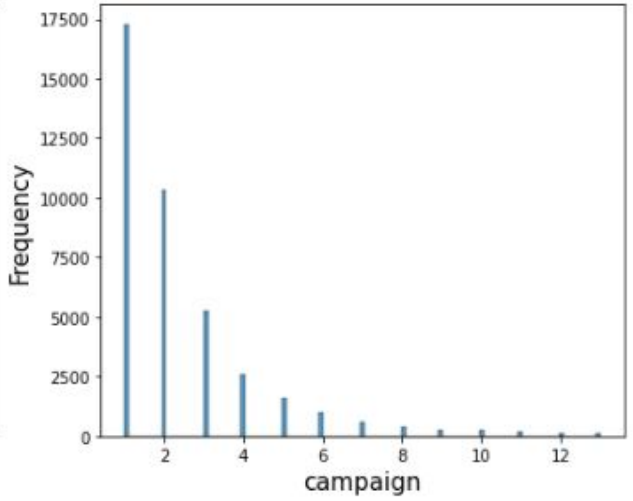
Percentile 99%		
age	duration	campaign
71	1271.25	14

Total values less than percetile 99%		
age	duration	campaign
40755	40764	40701

Histograms after clipping







Better, but there are still many outliers in **duration** and **campaign**.

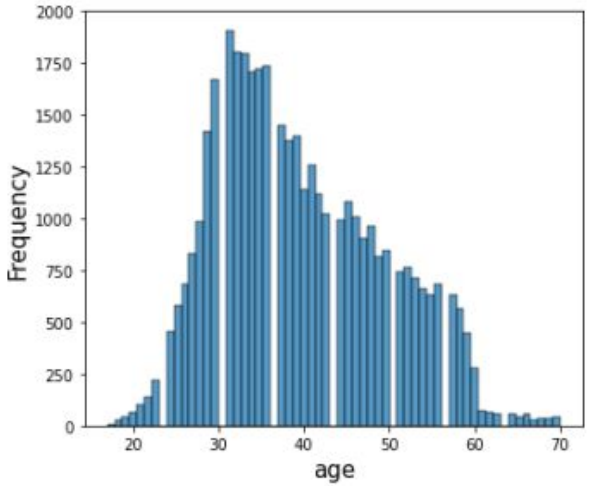
Taking the log of every value

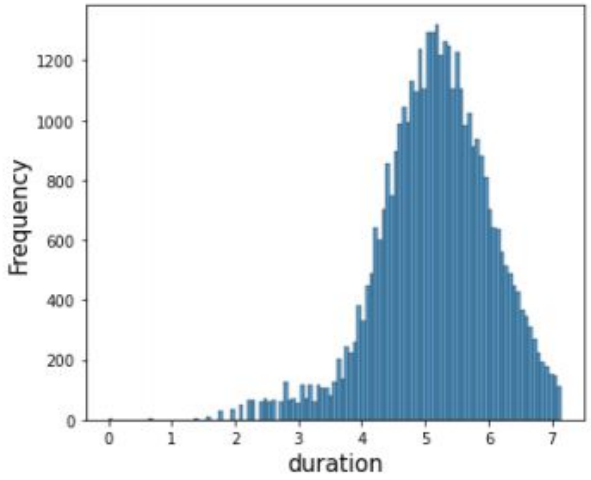
Campaign after log scaling

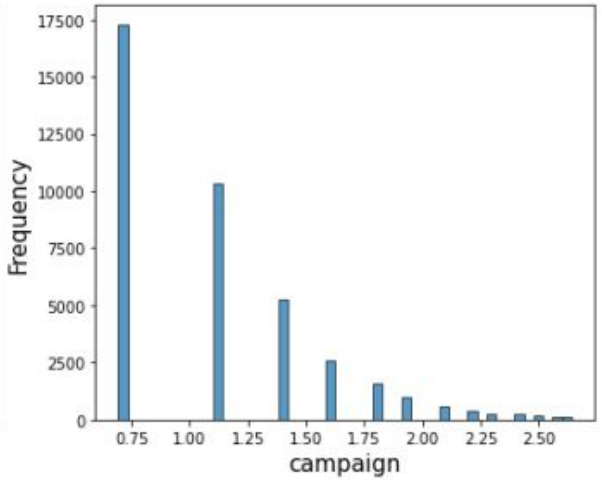
count	mean	Std	Min	25%	50%	75%	max
39872.00000	1.097673	0.449664	0.693147	0.693147	1.098612	1.386294	2.639057

Duration after log scaling

count	mean	Std	Min	25%	50%	75%	max
39872.00000	5.161927	0.881968	0.00	4.644391	5.192957	5.752573	7.148346



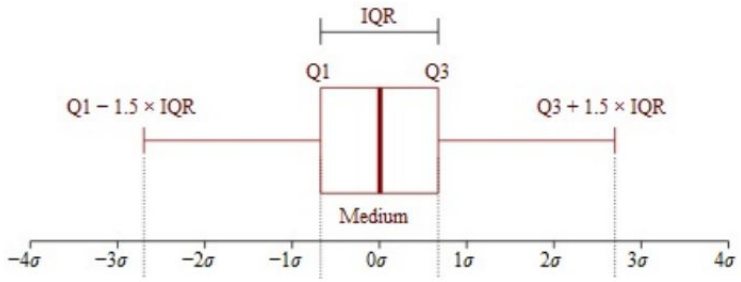


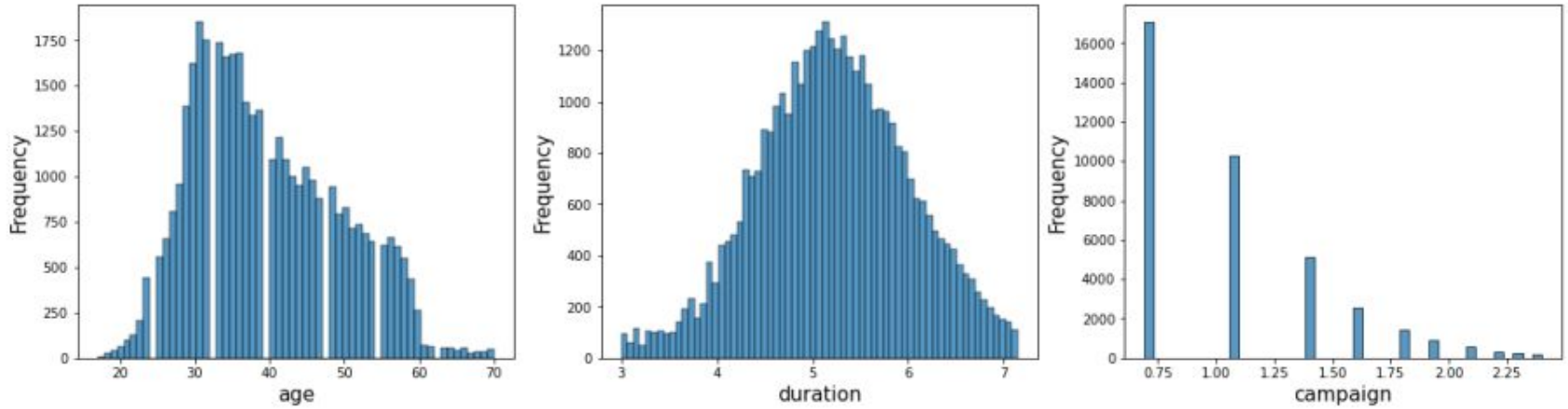


Now, **duration** looks more normal and **campaign** has less tail

IQR Score

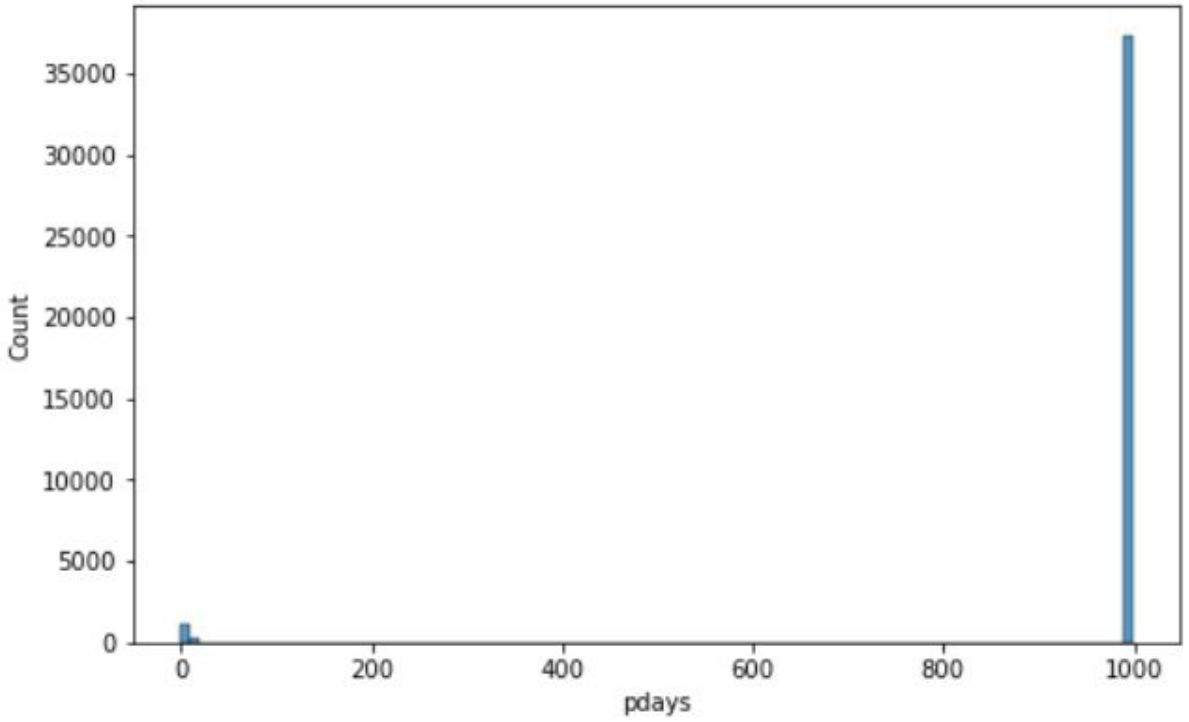
The IQR can be used to identify outliers by defining limits on the sample values that are a factor k of the IQR below the 25th percentile or above the 75th percentile. The common value for the factor k is the value 1.5. A factor k of 3 or more can be used to identify values that are extreme outliers or “far outs” when described in the context of box and whisker plots.





Now, our data is more useful than the original data

Transformation



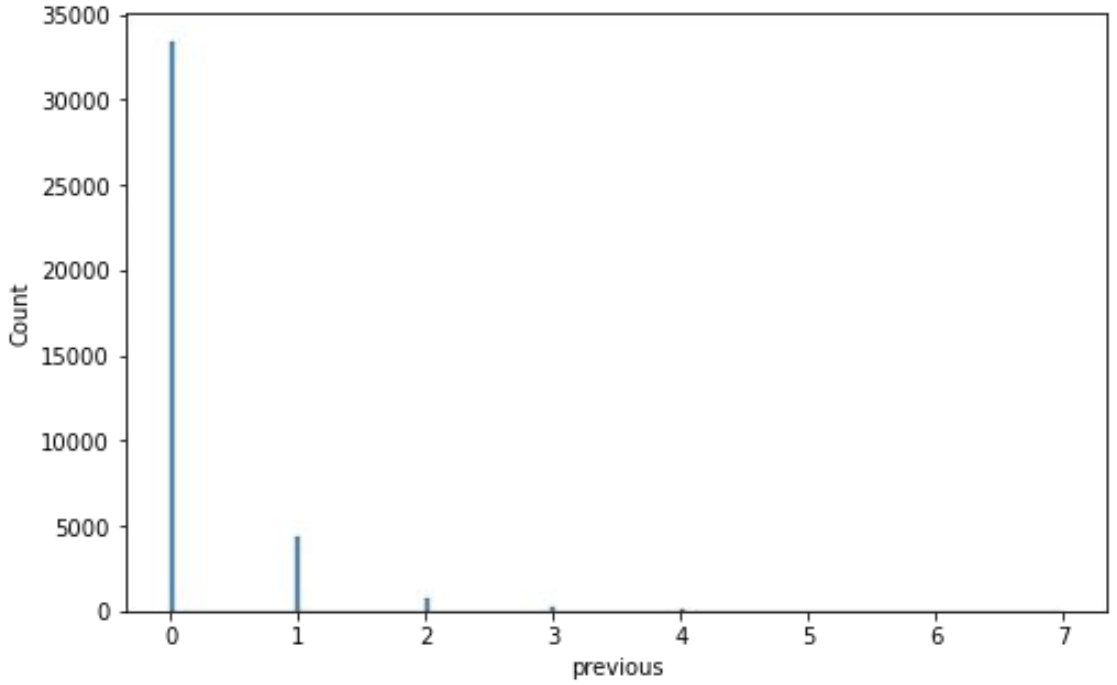
discretizing **pdays**	
bad	0<= value <=3
good	3< value <=7
excellent	7< value <=30
fair	30<value

unique counts **pdays**	
bad	37314
good	588
excellent	501
fair	315

Most of the people weren't contacted before, that's why most values are .

A new attribute will be created from pdays

Feature Selection

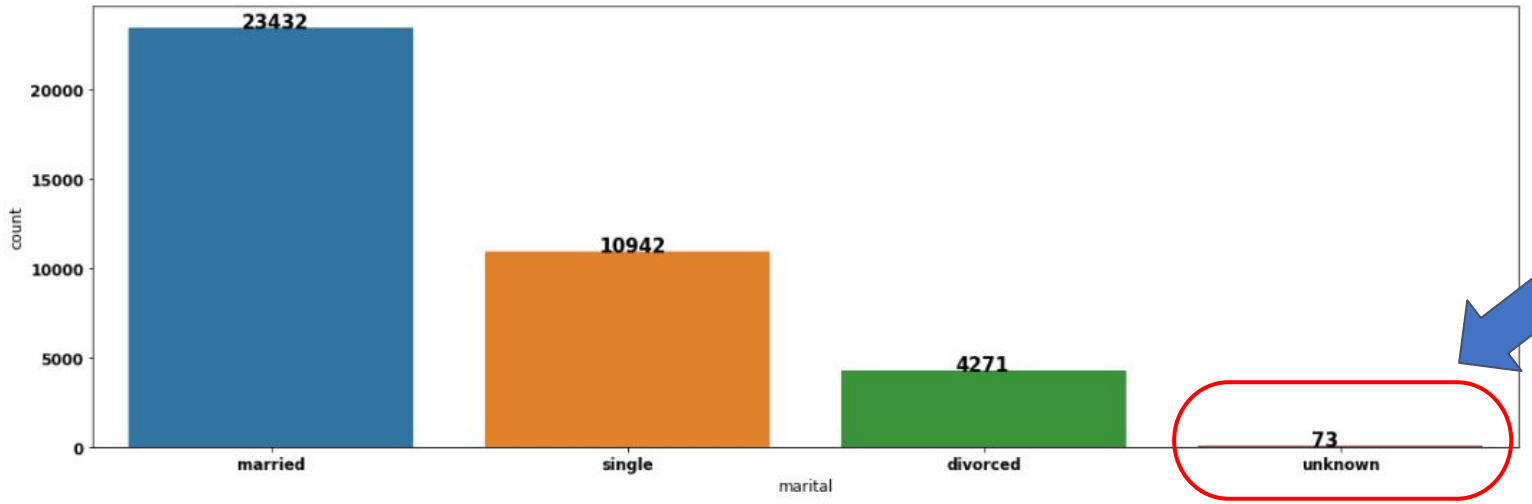


Before **previous** attribute has very low variance, it will be removed since this will not harm the performance of the model, and it can reduce the complexity of the model.

After dealing missing values, duplicate values, outliers and transformations of numerical attributes, all numerical attributes will be preserved except **previous.**

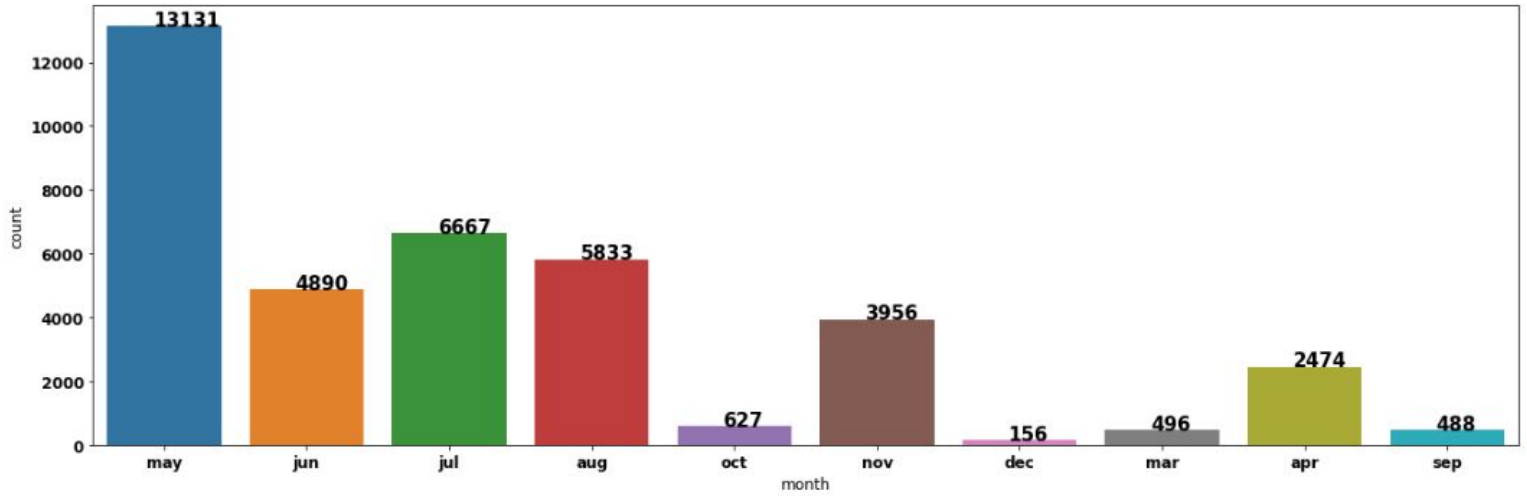
Cardinality and Unique counts

marital

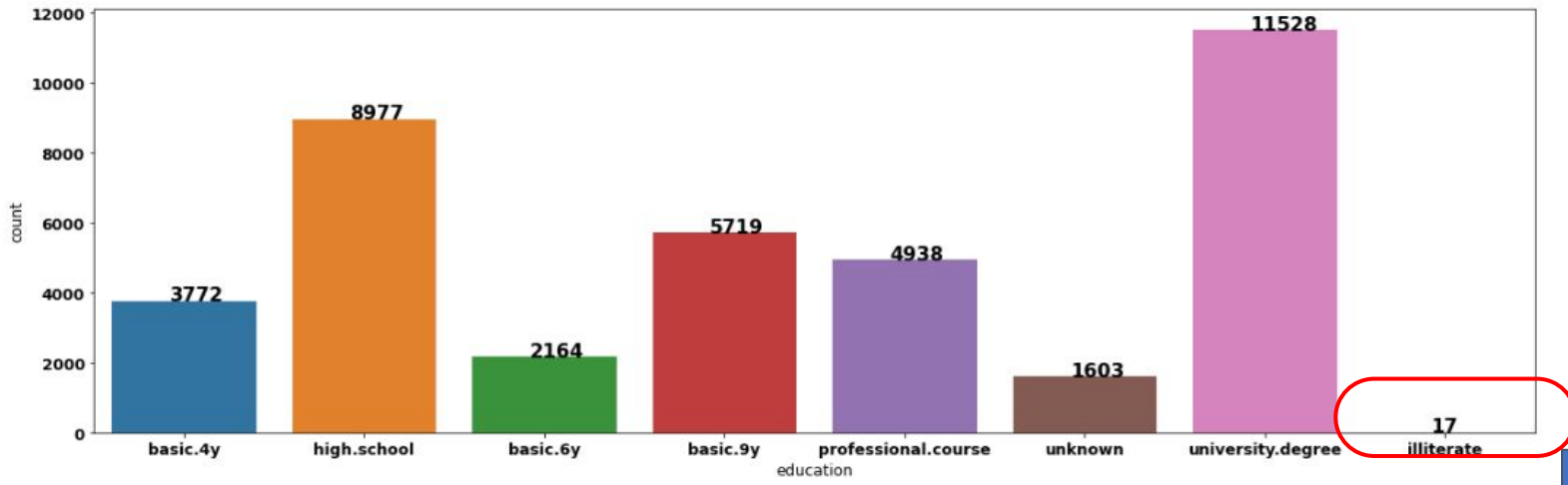


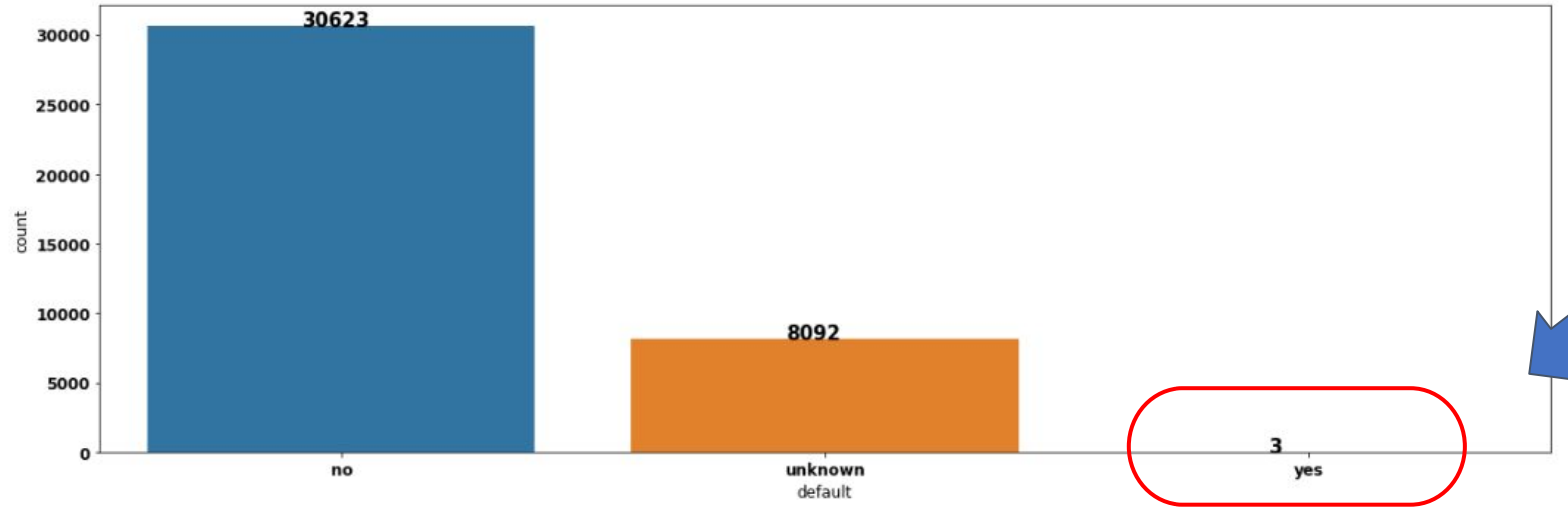
very few values

month



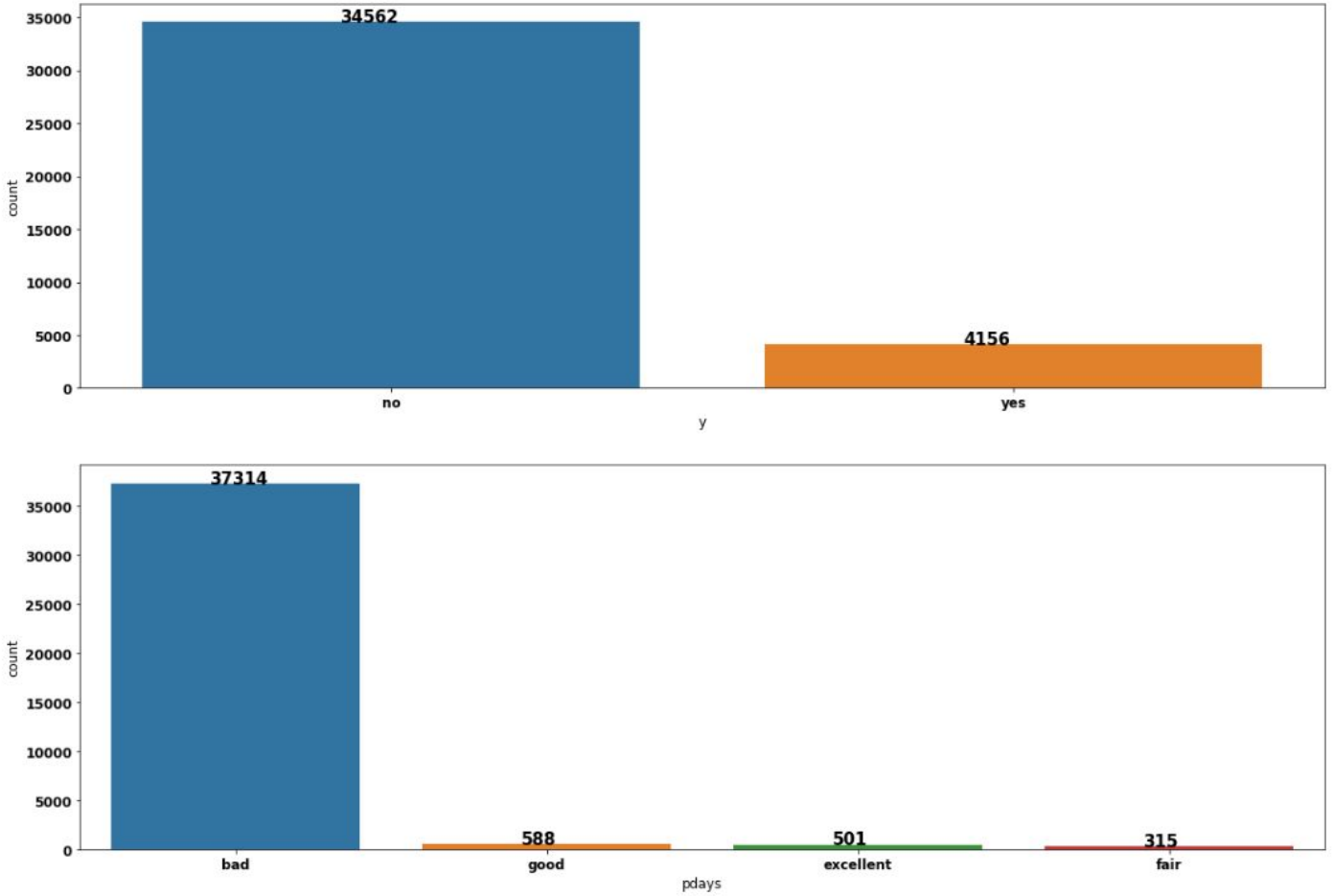
Cardinality and Unique counts





very few values

Feature Selection



- **pdays** attribute will be removed because it has a very small entropy, which means that most of the records are in **bad** category.
- The categories **illiterate,yes** and **unknown** will be removed from **education, default** and **marital** respectively.
- The Target **y** has class imbalance and will be treated to mitigate the imbalance.

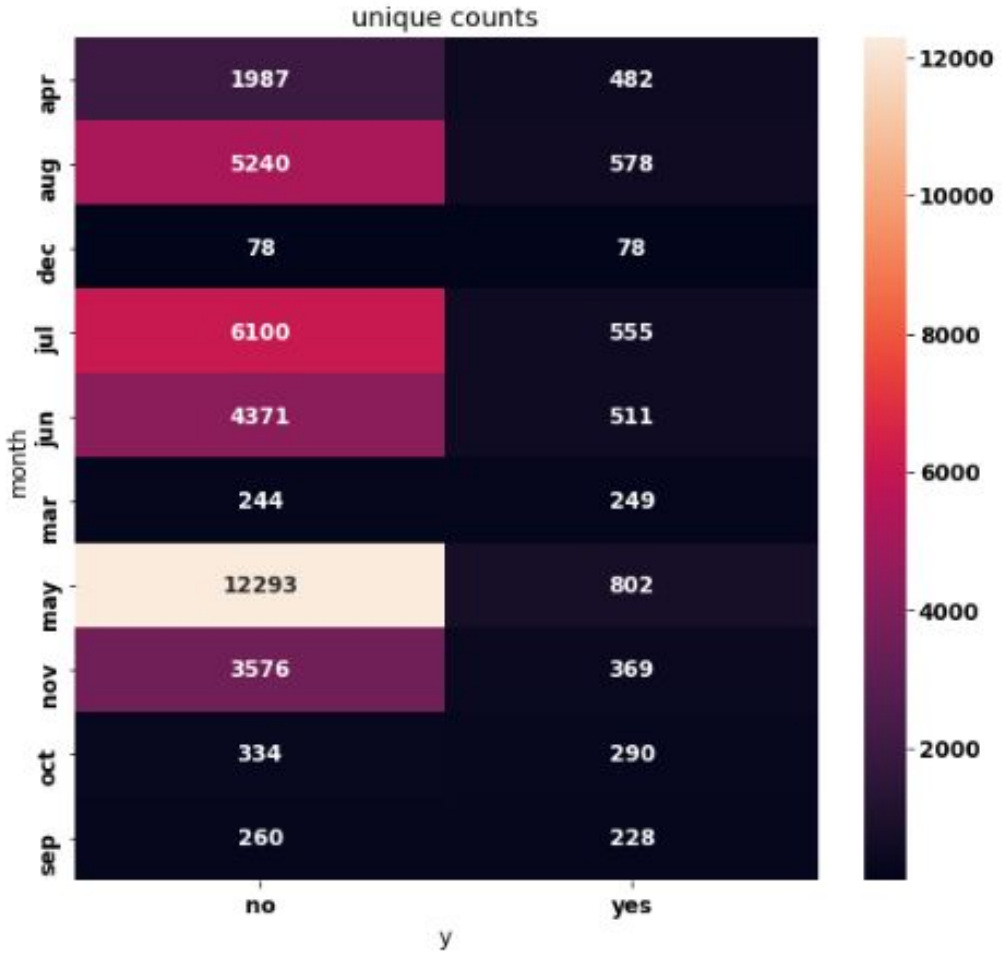
Correlation Analysis

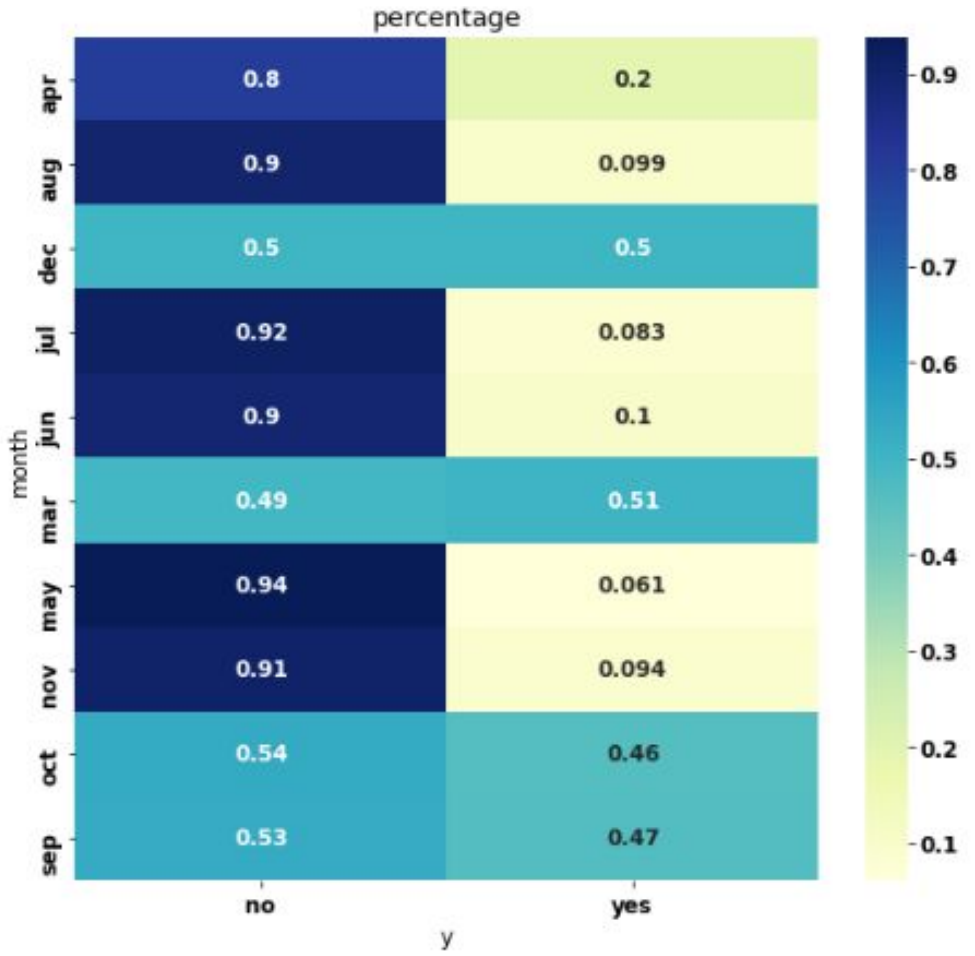
Categorical vs Categorical

Numerical vs Numerical

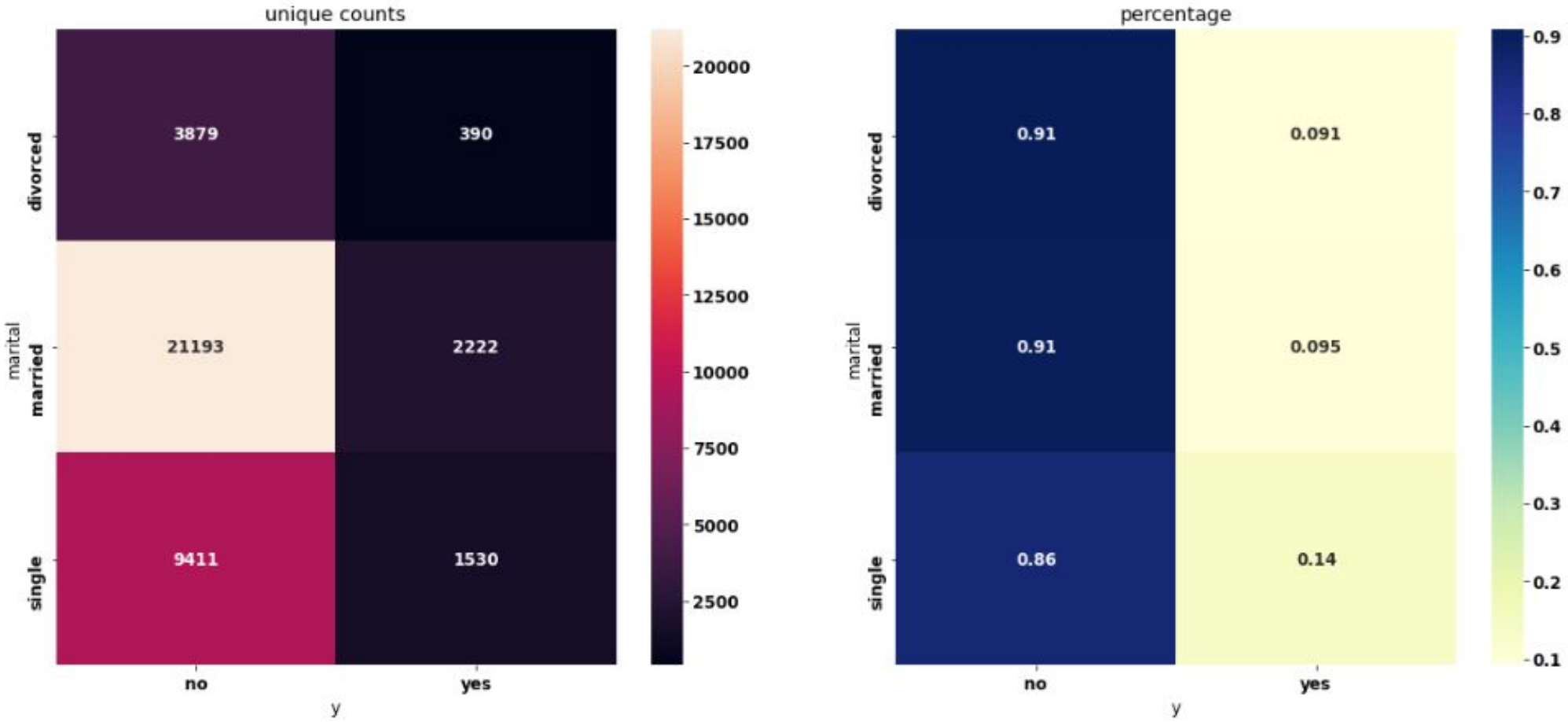
Numerical vs Categorical

Categorical vs Categorical - month vs target

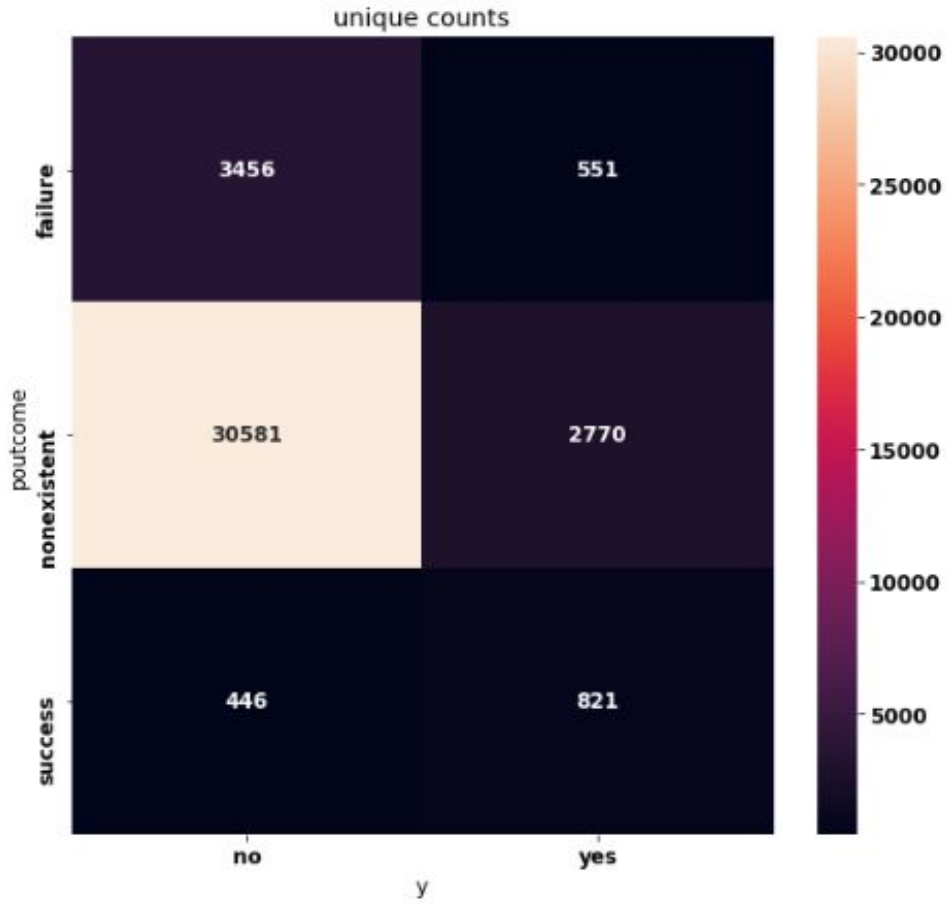


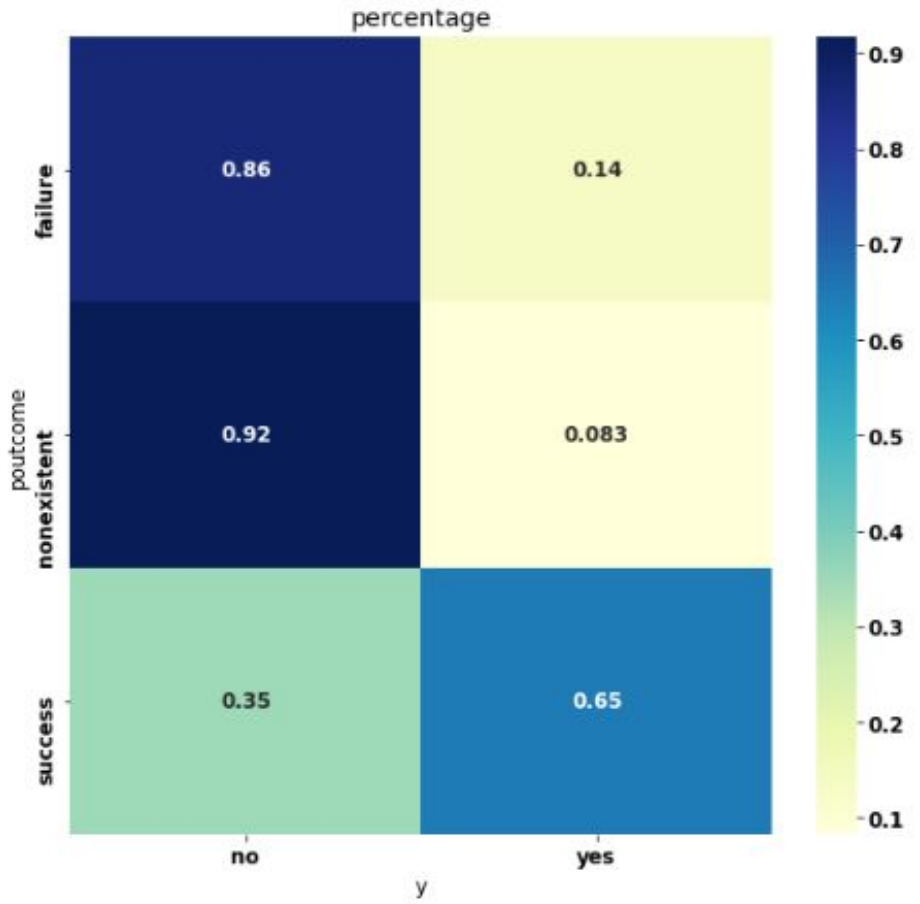


Categorical vs Categorical - marital vs target

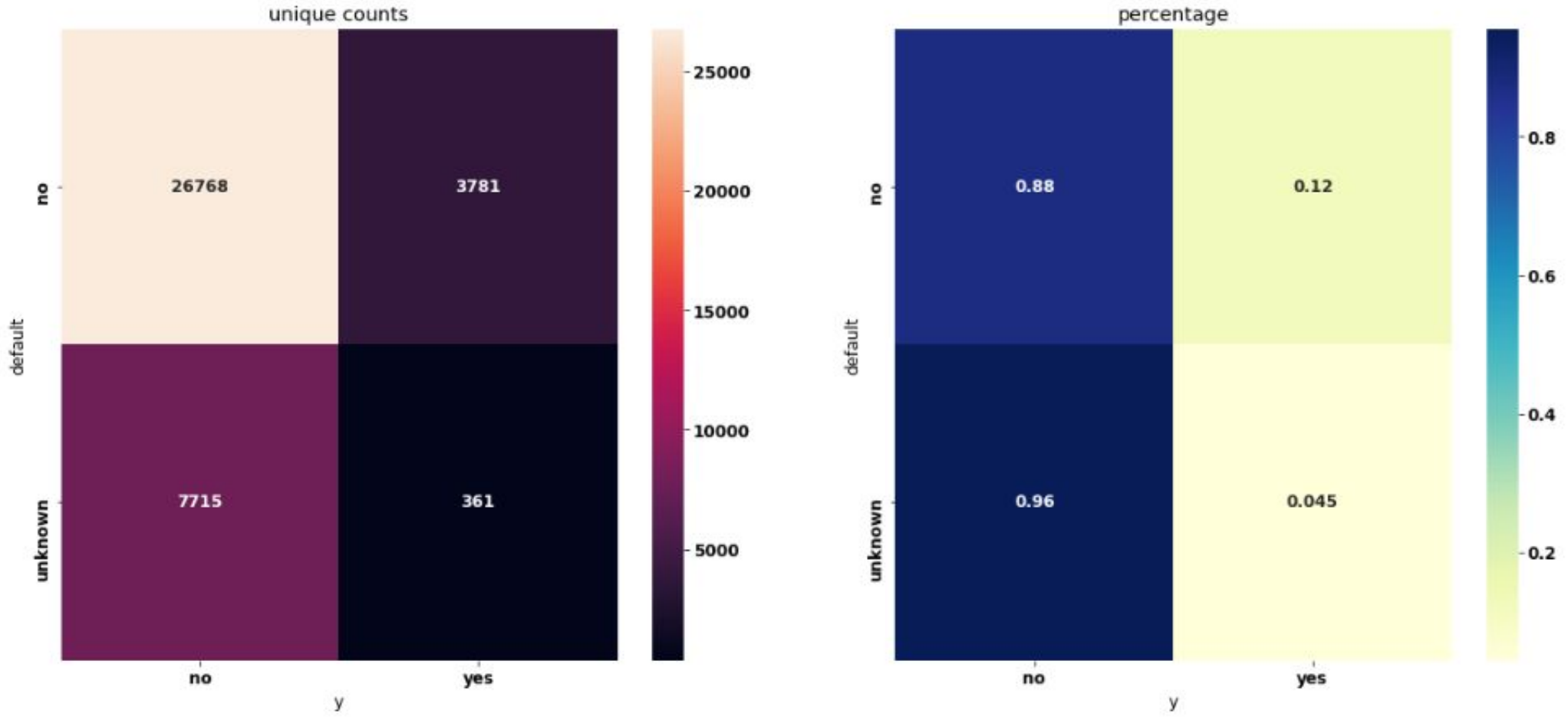


Categorical vs Categorical - poutcome vs target

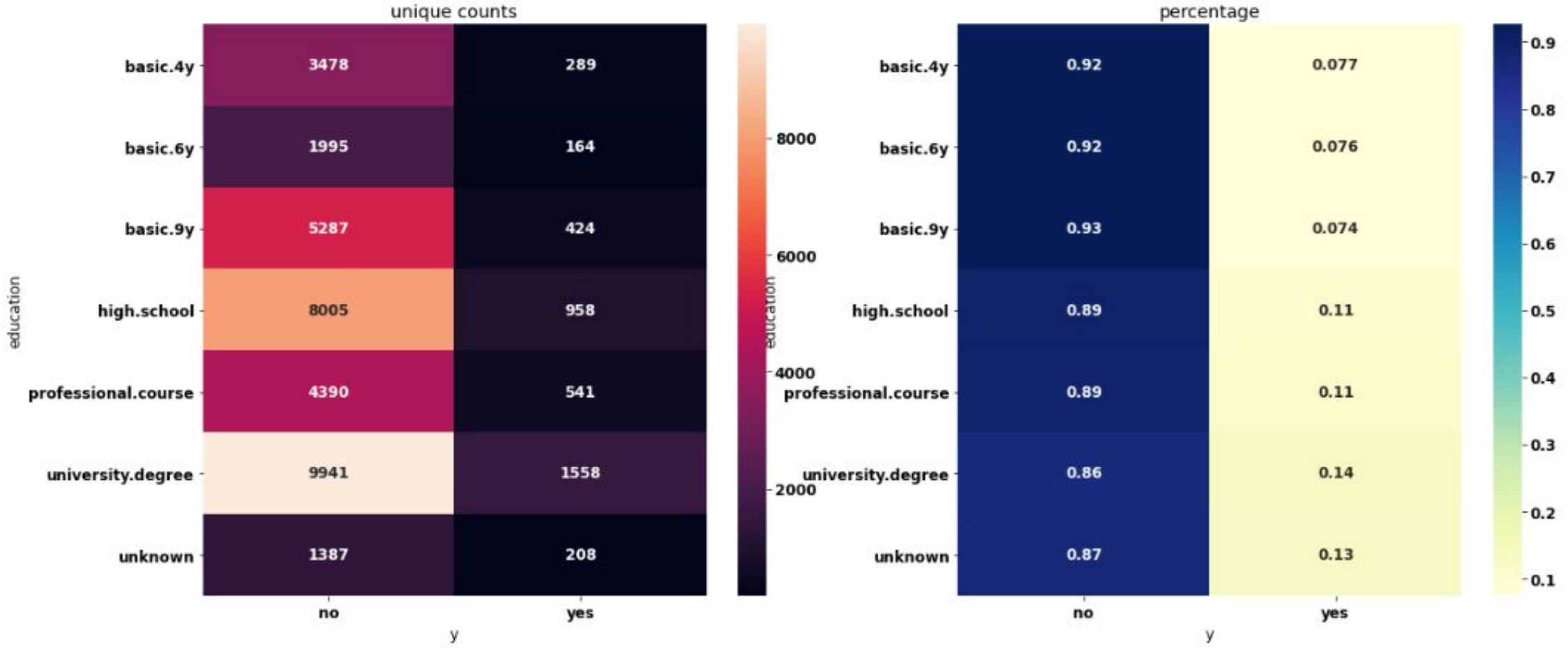




Categorical vs Categorical- default vs target

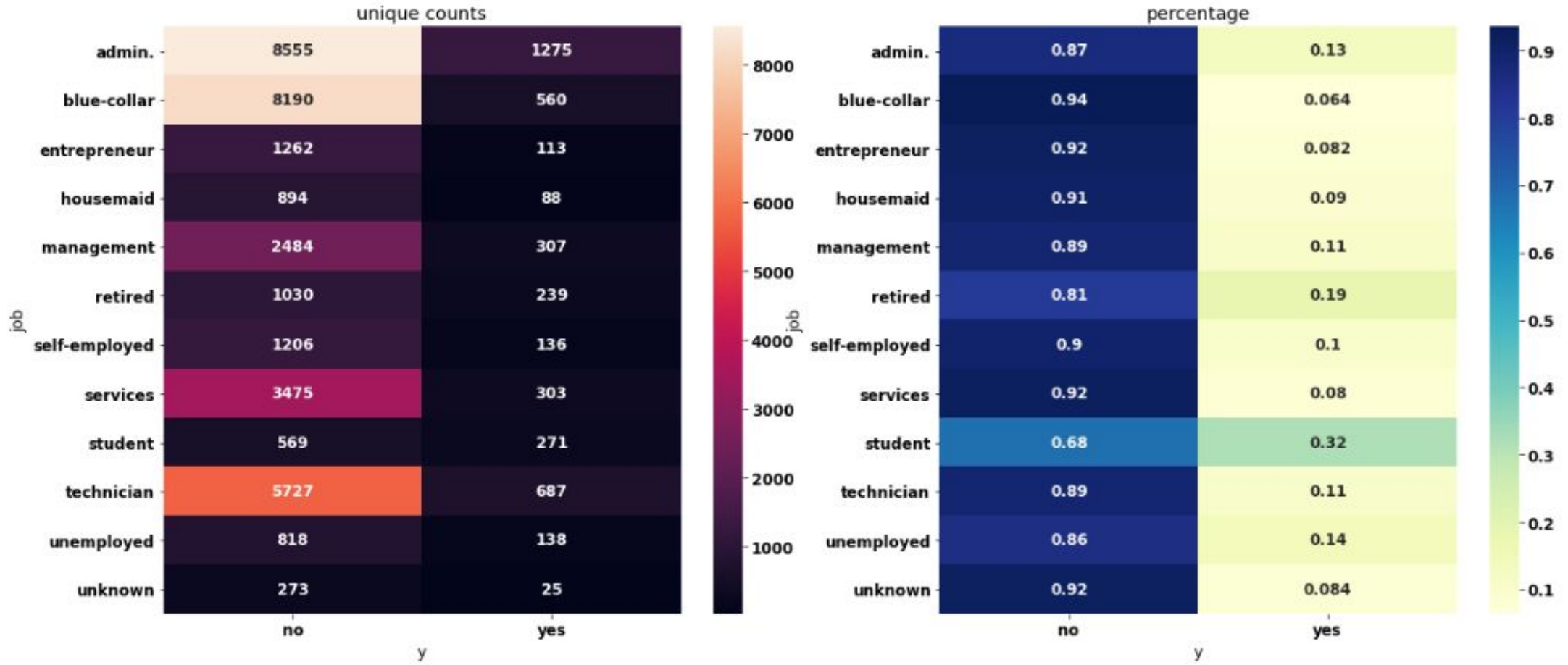


Categorical vs Categorical - education vs target

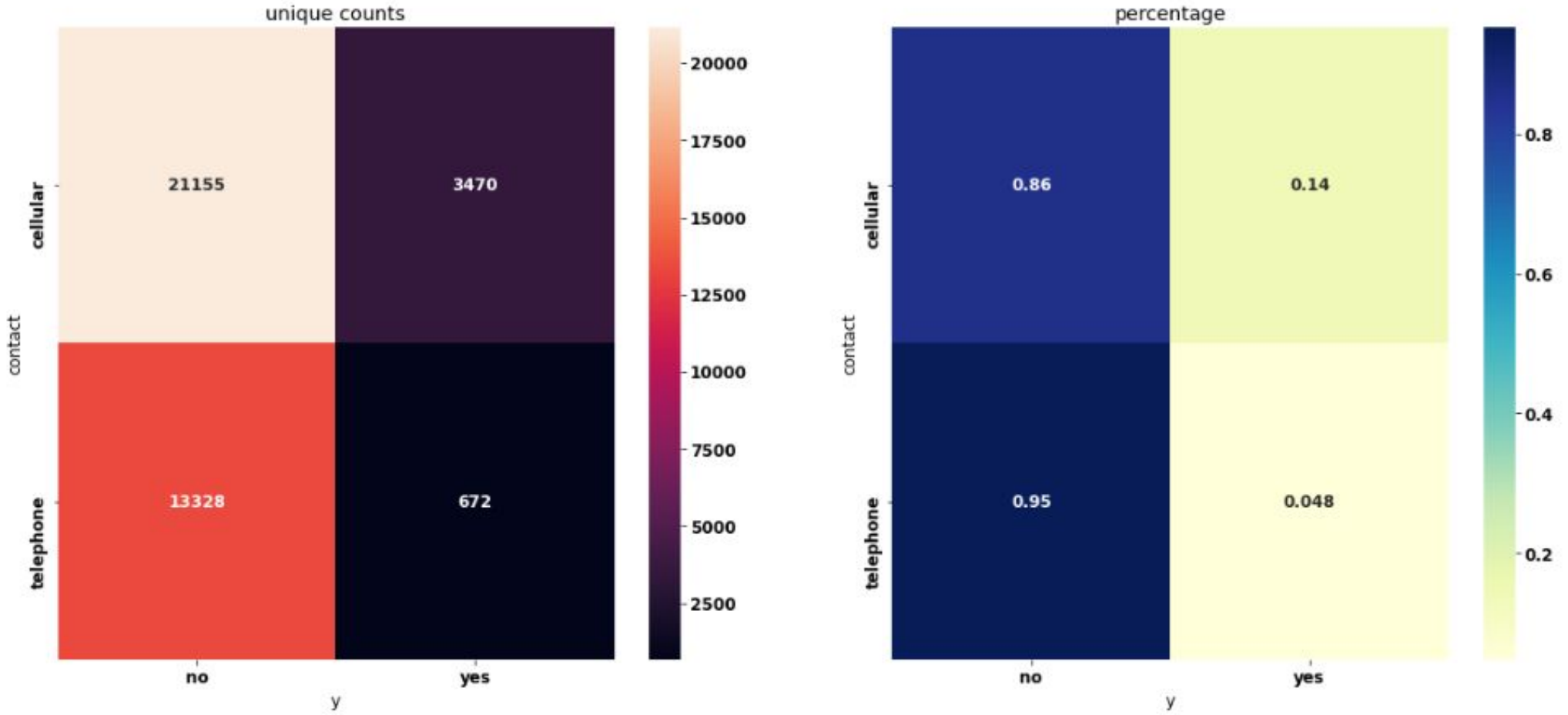


Data Glacier
Your Deep Learning Partner

Categorical vs Categorical - job vs target



Categorical vs Categorical- contact vs target



Qualitative analysis - Insights

Based on the contingency tables (heat maps):

- There is an apparent slight relationship between the **marital** attribute and the target **y**, as single people are more likely to subscribe to the product.
- There is an apparent relationship between the **default**, **job**, **contact**, **education**, **month**, **poutcome** attributes and the target **y.**
  - The relationship between **job**, **education** and the **target** is intuitive.
  - **poutcome** is the output of the last campaign, so this attribute is the most predictive.

    These relationships will be confirmed with a quantitative analysis.

- The **loan** and **housing** attributes have no relationship with the target **y**, even though the 'unknown' category is removed in both attributes, the relationship is null. These attributes will be removed.
- The **day_of_week** attribute has no relationship with the target **y**, it will be removed.

Quantitative analysis - Chi-squared Test

```
----------------------- marital vs target -------------------------
Interpret test-statistic
probability=0.950, critical=5.991, stat=169.976
Dependent (reject H0)


Interpret p-value
significance=0.050, p=0.000
Dependent (reject H0)


----------------------- default vs target -------------------------
Interpret test-statistic
probability=0.950, critical=3.841, stat=416.282
Dependent (reject H0)


Interpret p-value
significance=0.050, p=0.000
Dependent (reject H0)
```

Interpretation

marital vs target: p-value< significance, both attributes are dependent, this means that there is a relationship between them.

default vs target: p-value< significance, both attributes are dependent, this means that there is a relationship between them.

Quantitative analysis - Chi-squared Test

```
------------------------ job vs target -------------------------
Interpret test-statistic
probability=0.950, critical=19.675, stat=774.108
Dependent (reject H0)


Interpret p-value
significance=0.050, p=0.000
Dependent (reject H0)


------------------------ contact vs target -------------------------
Interpret test-statistic
probability=0.950, critical=3.841, stat=803.884
Dependent (reject H0)


Interpret p-value
significance=0.050, p=0.000
Dependent (reject H0)
```

Interpretation

job vs target: p-value< significance, both attributes are dependent, this means that there is a relationship between them.

contact vs target: p-value< significance, both attributes are dependent, this means that there is a relationship between them.

Quantitative analysis - Chi-squared Test

```
----------------------- education vs target -------------------------
Interpret test-statistic
probability=0.950, critical=12.592, stat=228.795
Dependent (reject H0)


Interpret p-value
significance=0.050, p=0.000
Dependent (reject H0)


----------------------- month vs target -------------------------
Interpret test-statistic
probability=0.950, critical=16.919, stat=3100.314
Dependent (reject H0)


Interpret p-value
significance=0.050, p=0.000
Dependent (reject H0)


----------------------- poutcome vs target -------------------------
Interpret test-statistic
probability=0.950, critical=5.991, stat=4111.883
Dependent (reject H0)


Interpret p-value
significance=0.050, p=0.000
Dependent (reject H0)
```

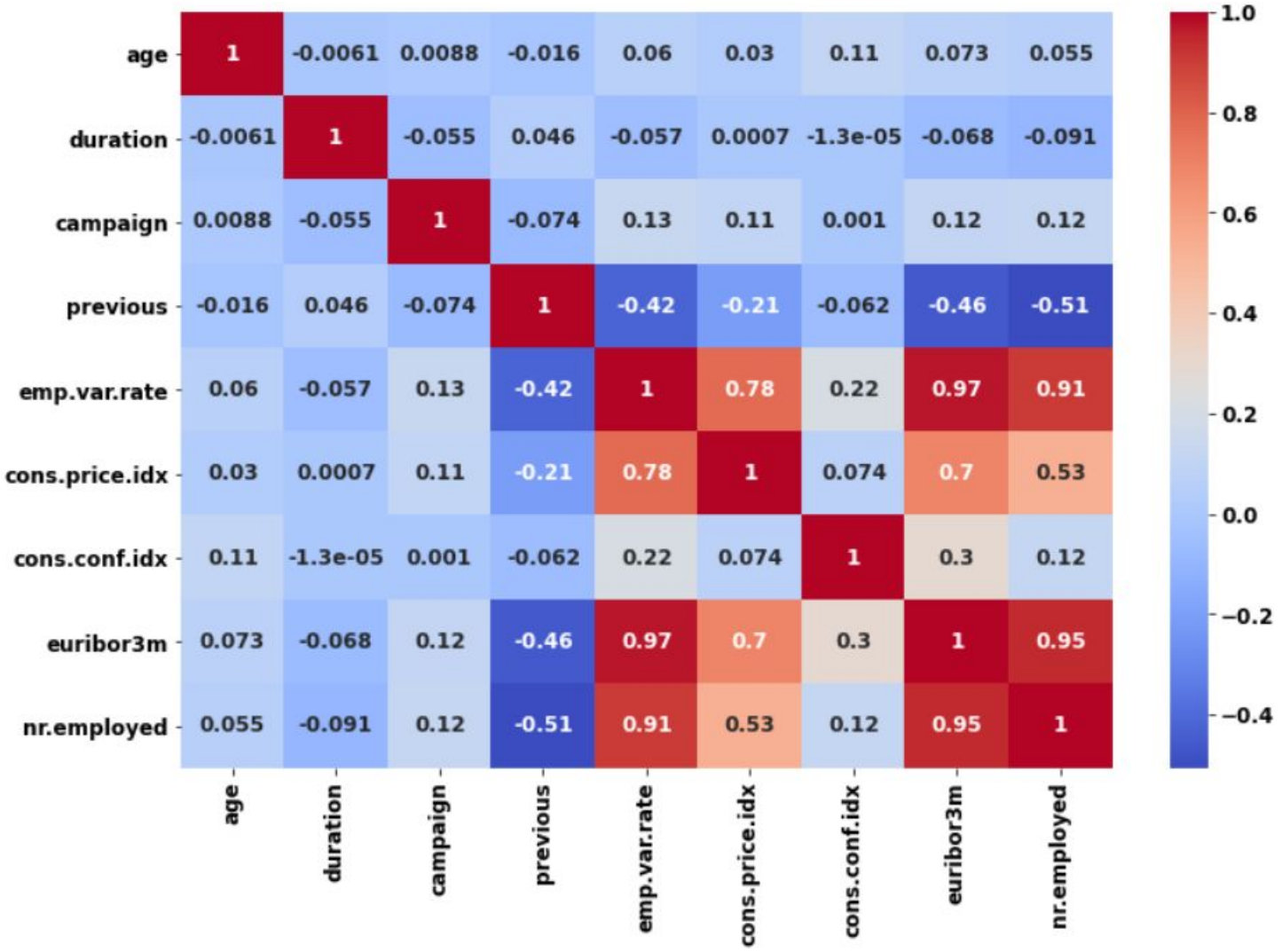
Interpretation

education vs target: p-value< significance, both attributes are dependent, this means that there is a relationship between them.

month vs target: p-value< significance, both attributes are dependent, this means that there is a relationship between them.

poutcome vs target: p-value< significance, both attributes are dependent, this means that there is a relationship between them.

Numerical vs Numerical



Using the Heat map above, the following can be observed:

- The **previous** attribute has high negative correlation with **nr.employed**,so the latter will be removed.
- The **emp.var.rate**, **euribor3m** and **cons.price.idx** attributes have a high positive correlation between them, so **euribor3m** and **emp.var.rate** will be removed.

Categorical versus numerical

Student T-test

Categorical	Value1	Value2	Numerical	p-value	t-statistic
target 'y'	yes	no	age	0.002	-3.073700
target 'y'	yes	no	duration	0.000	72.474800
target 'y'	yes	no	campaign	0.000	-11.465300
target 'y'	yes	no	previous	0.000	47.233300
target 'y'	yes	no	cons.price.idx	0.000	-27.633000
target 'y'	yes	no	cons.conf.idx	0.000	8.957300

Taking a significance level of 0.05, this table shows that there is a strong correlation between the target y and all numerical variables.

Final Recommendations

After Exploratory Data Analysis and Feature Selection and Engineering, the features that should be fed to the model are:

Numerical:

- age
- duration
- campaign
- previous
- cons.price.idx
- cons.conf.idx

Categorical:

- marital
- default
- job
- contact
- education
- month
- poutcome

Target:

y : Imbalance of categorical target, This problem will be addressed when building the model using different methods:

- Oversampling minority class
- Alternative metric and/or loss function

Model recommendation

algorithms	advantages	disadvantages
Logistic Regression	• easy to train • easy to implement	• difficult to fit nonlinear data • does not easily grasp complex relationships
Decision Tree	• easy to interpret • models complex relationships	• bounded exploration of the variable space • very simple
Neural Networks	• models complex relationships • high performance with large amounts of data	• difficult to interpret
Random Forest	• good exploration of variable space • parallel work	• more complicated to interpret than a tree
SVM	• models complex relationships • robust model against noise	• difficult to interpret • need processing power
XGBoost	• high predictive power • power other algorithms	• black box

Thank You