Bank Telemarketing Project

Machine Learning

By: Aya Hamrouni and Safa Trabelsi

Outline

1. Introduction
2. Data Explanation
3. Data Exploration
4. Data Visualization
5. Data Pre-processing
6. Data Modeling
7. Recommendations

Introduction 01

Introduction

UNDERSTANDING THE PROBLEM

What makes a telemarketing compaign successul?

What are the factors that influence a customer’s decision to subscribe for a term depositN

Data Explanation

02

Our variables

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').

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)

Target Variable

21 - **y** - has the client subscribed a term deposit? (binary: 'yes', 'no')

Libraries

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import cross_val_score
from sklearn.ensemble import AdaBoostClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import confusion_matrix
from sklearn.metrics import plot_confusion_matrix
from sklearn.model_selection import GridSearchCV
from sklearn.tree import DecisionTreeClassifier
from sklearn.tree import plot_tree
```

Data Exploration

03

Dataset: (41188, 20)

	age	job	marital	education	default	housing	loan	contact	month	day_of_week	duration	campaign	pdays	previous	poutcome
0	56	housemaid	married	basic.4y	no	no	no	telephone	may	mon	261	1	999	0	nonexistent
1	57	services	married	high.school	unknown	no	no	telephone	may	mon	149	1	999	0	nonexistent
2	37	services	married	high.school	no	yes	no	telephone	may	mon	226	1	999	0	nonexistent
3	40	admin.	married	basic.6y	no	no	no	telephone	may	mon	151	1	999	0	nonexistent
4	56	services	married	high.school	no	no	yes	telephone	may	mon	307	1	999	0	nonexistent
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
41183	73	retired	married	professional.course	no	yes	no	cellular	nov	fri	334	1	999	0	nonexistent
41184	46	blue-collar	married	professional.course	no	no	no	cellular	nov	fri	383	1	999	0	nonexistent
41185	56	retired	married	university.degree	no	yes	no	cellular	nov	fri	189	2	999	0	nonexistent
41186	44	technician	married	professional.course	no	no	no	cellular	nov	fri	442	1	999	0	nonexistent
41187	74	retired	married	professional.course	no	yes	no	cellular	nov	fri	239	3	999	1	failure

41188 rows × 21 columns

Types of Variables

```
 #   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)
memory usage: 6.6+ MB
```

Missing Values?

```
df.isnull().sum()

age               0
job               0
marital           0
education         0
default           0
housing           0
loan              0
contact           0
month             0
day_of_week       0
duration          0
campaign          0
pdays             0
previous          0
poutcome          0
emp.var.rate      0
cons.price.idx    0
cons.conf.idx     0
euribor3m         0
nr.employed       0
y                 0
dtype: int64
```

Unique Values

```
age [56 57 37 40 45 59 41 24 25 29 35 54 46 50 39 30 55 49 34 52 58 32 38 44
 42 60 53 47 51 48 33 31 43 36 28 27 26 22 23 20 21 61 19 18 70 66 76 67
 73 88 95 77 68 75 63 80 62 65 72 82 64 71 69 78 85 79 83 81 74 17 87 91
 86 98 94 84 92 89]
job ['housemaid' 'services' 'admin.' 'blue-collar' 'technician' 'retired'
 'management' 'unemployed' 'self-employed' 'unknown' 'entrepreneur'
 'student']
marital ['married' 'single' 'divorced' 'unknown']
education ['basic.4y' 'high.school' 'basic.6y' 'basic.9y' 'professional.course'
 'unknown' 'university.degree' 'illiterate']
default ['no' 'unknown' 'yes']
housing ['no' 'yes' 'unknown']
loan ['no' 'yes' 'unknown']
contact ['telephone' 'cellular']
month ['may' 'jun' 'jul' 'aug' 'oct' 'nov' 'dec' 'mar' 'apr' 'sep']
day_of_week ['mon' 'tue' 'wed' 'thu' 'fri']
duration [ 261  149  226 ... 1246 1556 1868]
campaign [ 1  2  3  4  5  6  7  8  9 10 11 12 13 19 18 23 14 22 25 16 17 15 20 56
 39 35 42 28 26 27 32 21 24 29 31 30 41 37 40 33 34 43]
pdays [999   6   4   3   5   1   0  10   7   8   9  11   2  12  13  14  15  16
  21  17  18  22  25  26  19  27  20]
previous [0 1 2 3 4 5 6 7]
poutcome ['nonexistent' 'failure' 'success']
emp.var.rate [ 1.1  1.4 -0.1 -0.2 -1.8 -2.9 -3.4 -3.  -1.7 -1.1]
```

Continued...

```
cons.price.idx [93.994 94.465 93.918 93.444 93.798 93.2   92.756 92.843 93.075 92.893
 92.963 92.469 92.201 92.379 92.431 92.649 92.713 93.369 93.749 93.876
 94.055 94.215 94.027 94.199 94.601 94.767]
cons.conf.idx [-36.4 -41.8 -42.7 -36.1 -40.4 -42.  -45.9 -50.  -47.1 -46.2 -40.8 -33.6
 -31.4 -29.8 -26.9 -30.1 -33.  -34.8 -34.6 -40.  -39.8 -40.3 -38.3 -37.5
 -49.5 -50.8]
euribor3m [4.857 4.856 4.855 4.859 4.86  4.858 4.864 4.865 4.866 4.967 4.961 4.959
 4.958 4.96  4.962 4.955 4.947 4.956 4.966 4.963 4.957 4.968 4.97  4.965
 4.964 5.045 5.    4.936 4.921 4.918 4.912 4.827 4.794 4.76  4.733 4.7
 4.663 4.592 4.474 4.406 4.343 4.286 4.245 4.223 4.191 4.153 4.12  4.076
 4.021 3.901 3.879 3.853 3.816 3.743 3.669 3.563 3.488 3.428 3.329 3.282
 3.053 1.811 1.799 1.778 1.757 1.726 1.703 1.687 1.663 1.65  1.64  1.629
 1.614 1.602 1.584 1.574 1.56  1.556 1.548 1.538 1.531 1.52  1.51  1.498
 1.483 1.479 1.466 1.453 1.445 1.435 1.423 1.415 1.41  1.405 1.406 1.4
 1.392 1.384 1.372 1.365 1.354 1.344 1.334 1.327 1.313 1.299 1.291 1.281
 1.266 1.25  1.244 1.259 1.264 1.27  1.262 1.26  1.268 1.286 1.252 1.235
 1.224 1.215 1.206 1.099 1.085 1.072 1.059 1.048 1.044 1.029 1.018 1.007
 0.996 0.979 0.969 0.944 0.937 0.933 0.927 0.921 0.914 0.908 0.903 0.899
 0.884 0.883 0.881 0.879 0.873 0.869 0.861 0.859 0.854 0.851 0.849 0.843
 0.838 0.834 0.829 0.825 0.821 0.819 0.813 0.809 0.803 0.797 0.788 0.781
 0.778 0.773 0.771 0.77  0.768 0.766 0.762 0.755 0.749 0.743 0.741 0.739
 0.75  0.753 0.754 0.752 0.744 0.74  0.742 0.737 0.735 0.733 0.73  0.731
 0.728 0.724 0.722 0.72  0.719 0.716 0.715 0.714 0.718 0.721 0.717 0.712
 0.71  0.709 0.708 0.706 0.707 0.7   0.655 0.654 0.653 0.652 0.651 0.65
 0.649 0.646 0.644 0.643 0.639 0.637 0.635 0.636 0.634 0.638 0.64  0.642
 0.645 0.659 0.663 0.668 0.672 0.677 0.682 0.683 0.684 0.685 0.688 0.69
 0.692 0.695 0.697 0.699 0.701 0.702 0.704 0.711 0.713 0.723 0.727 0.729
```

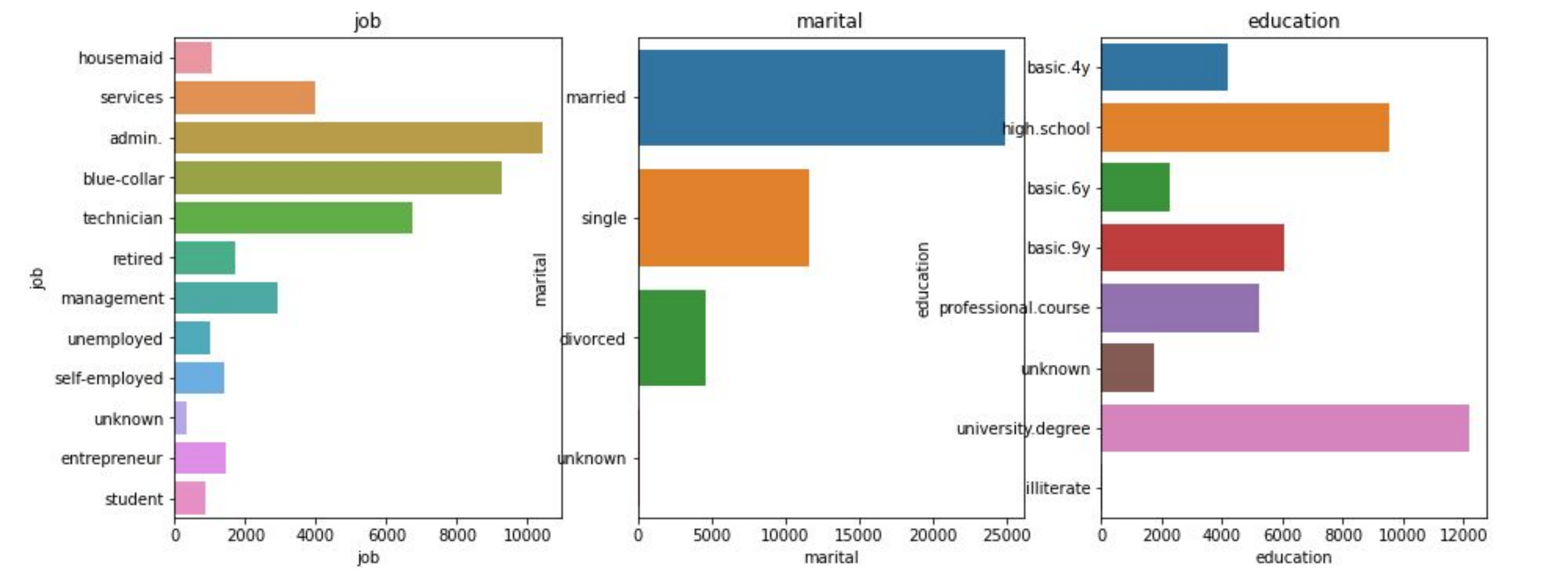
Descriptive Statistics

	age	duration	campaign	pdays	previous	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed
count	41188.00000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000
mean	40.02406	258.285010	2.567593	962.475454	0.172963	0.081886	93.575664	-40.502600	3.621291	5167.035911
std	10.42125	259.279249	2.770014	186.910907	0.494901	1.570960	0.578840	4.628198	1.734447	72.251528
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

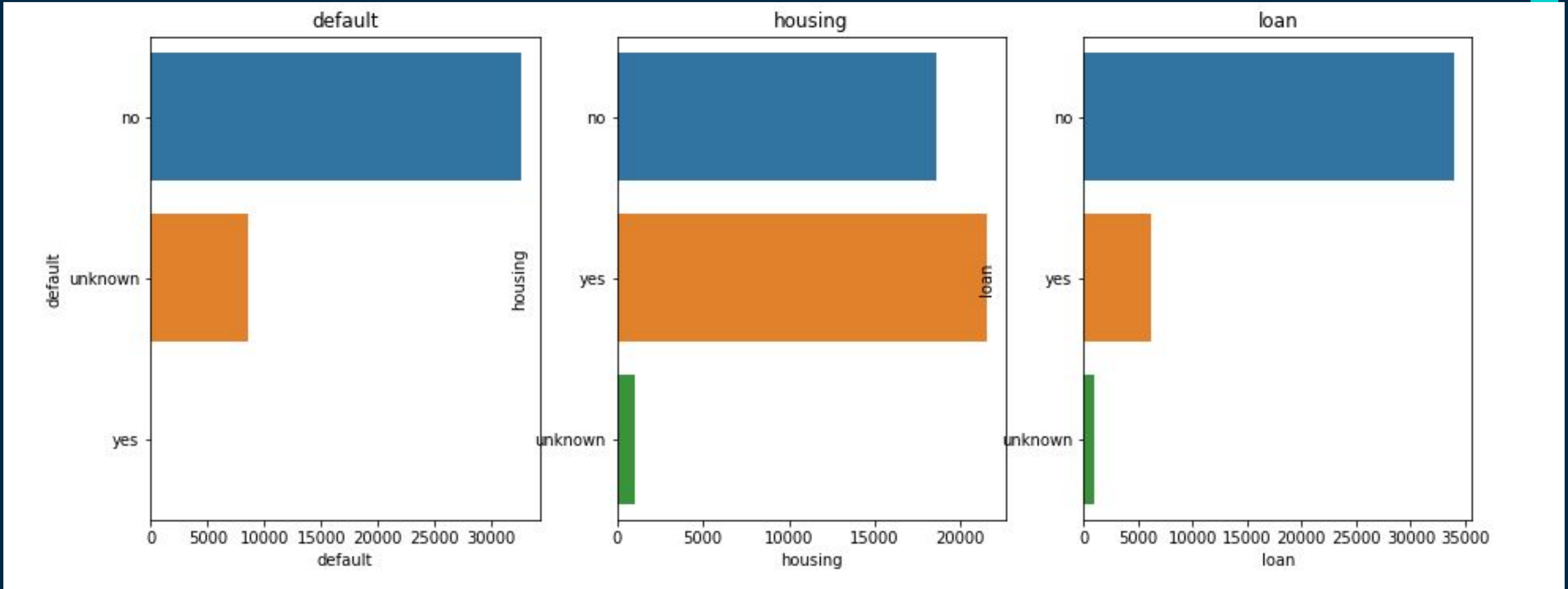
Data Visualization

04

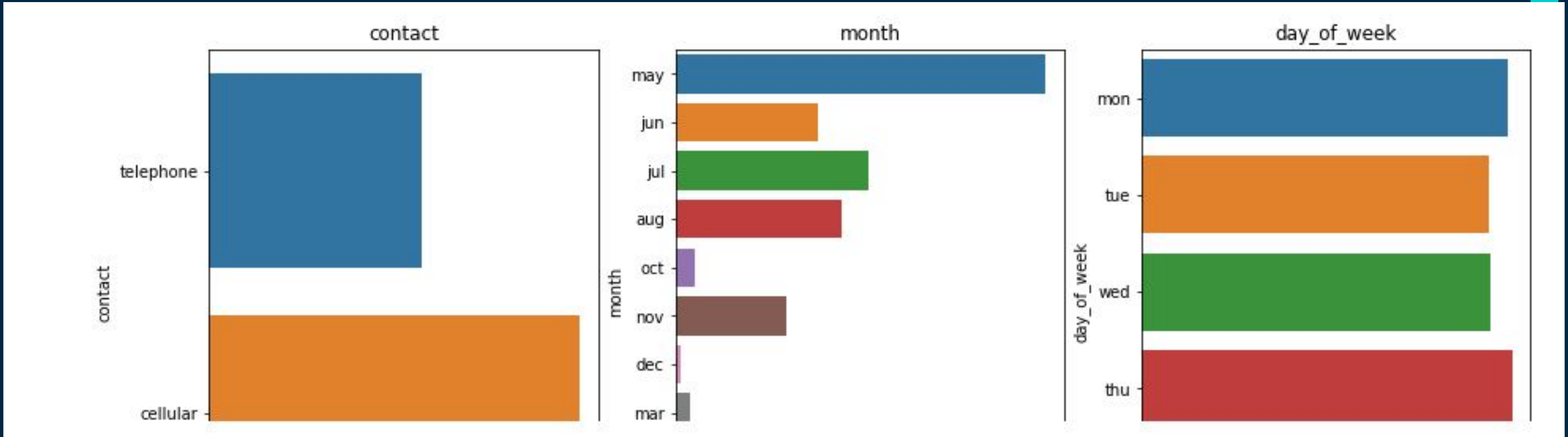
Categorical Data



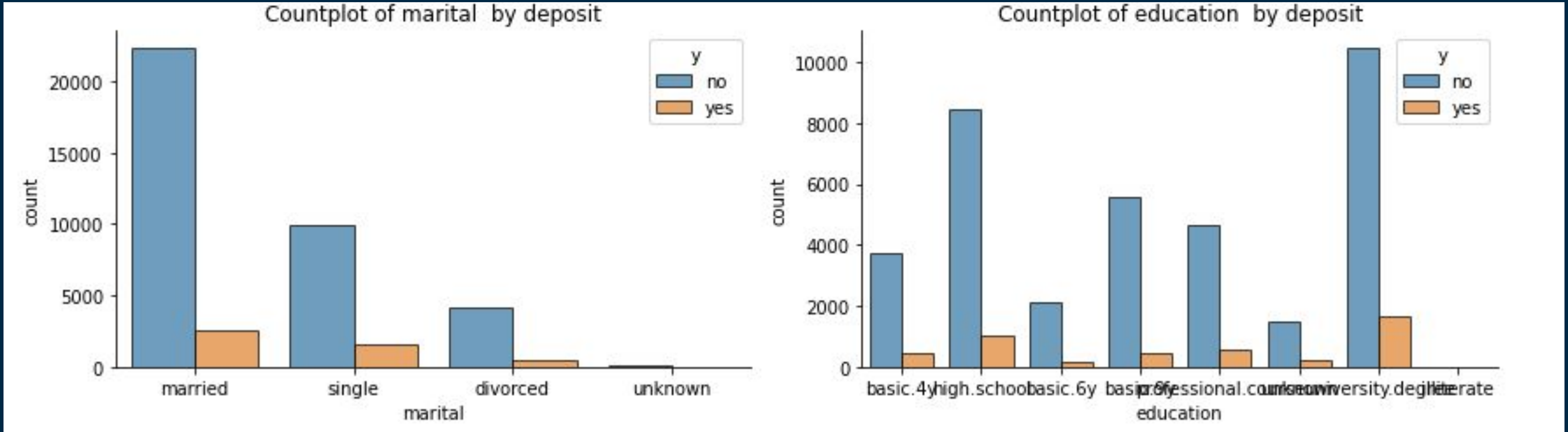
Continued (2/3)...



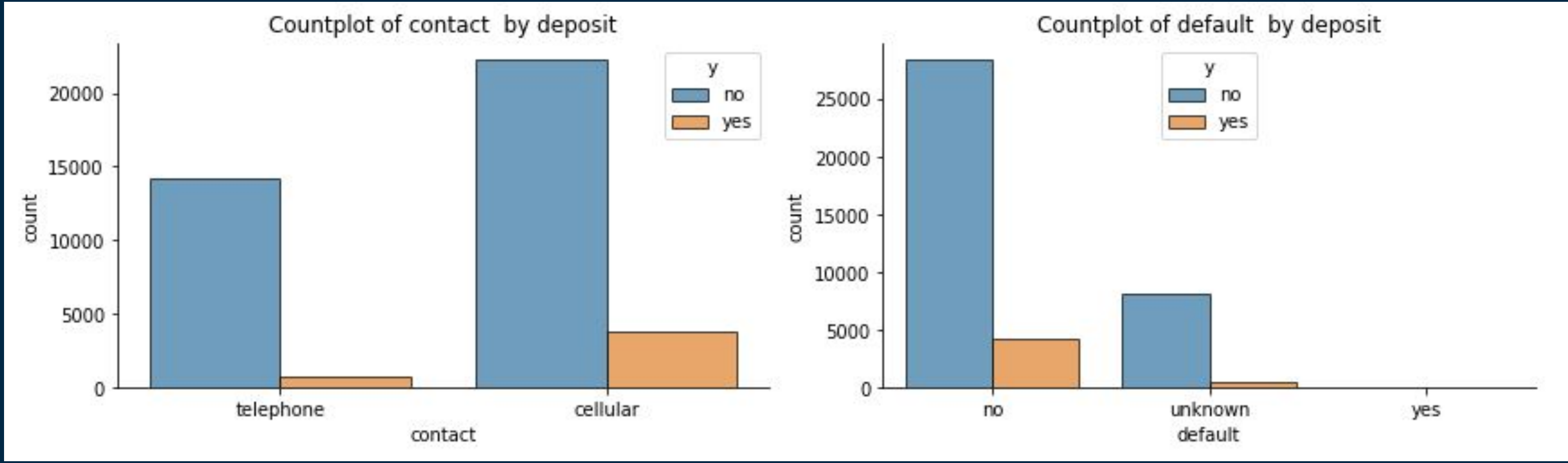
Continued (3/3)...



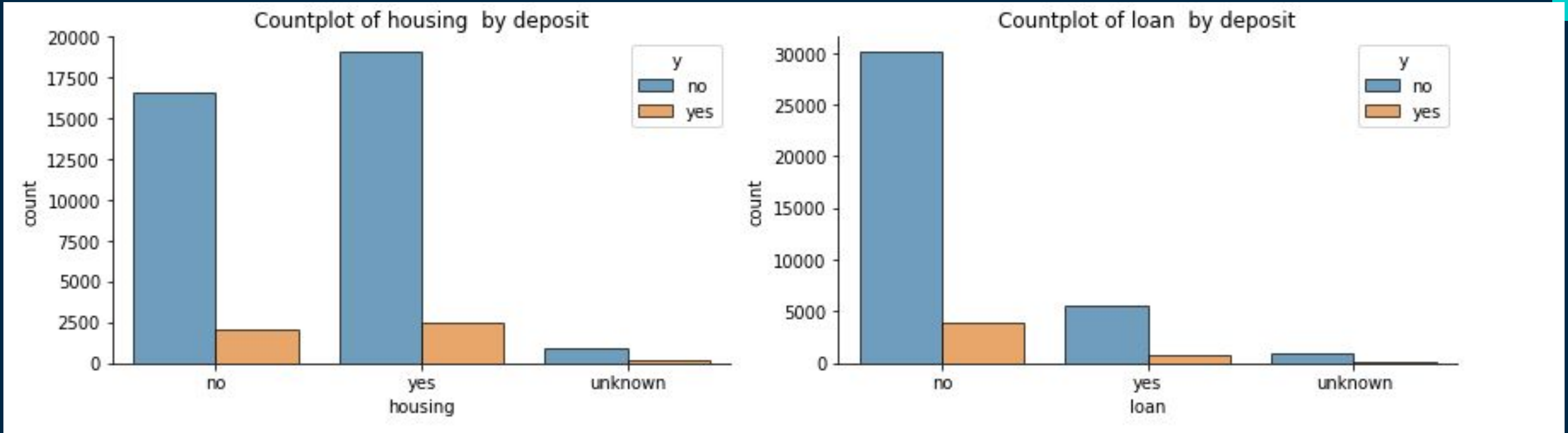
Categorical features (based on deposit counts)



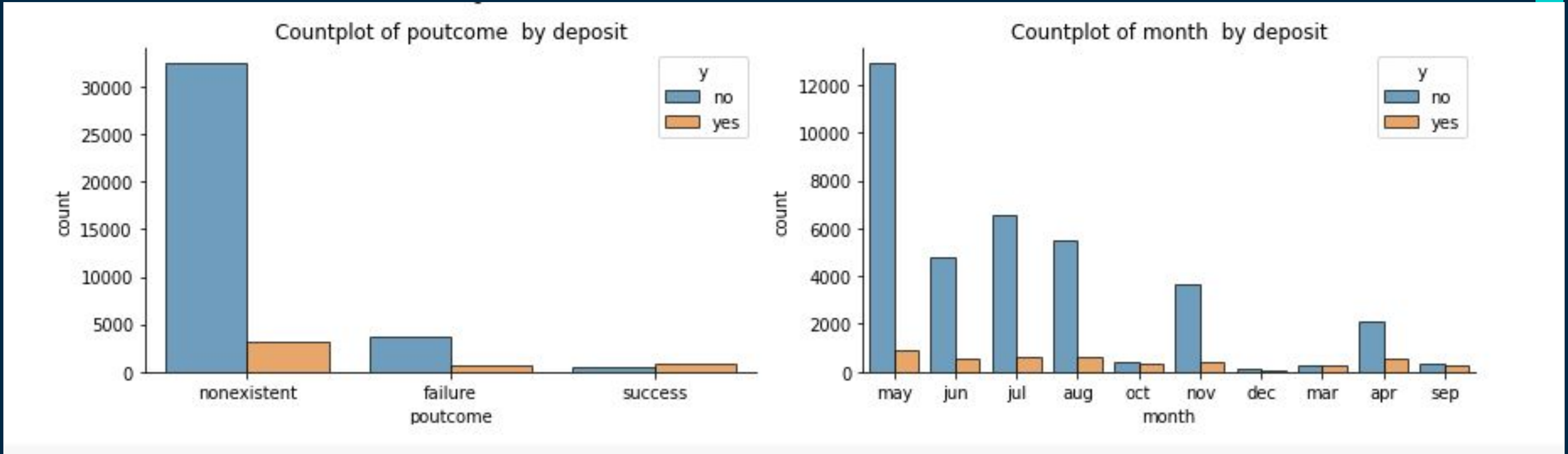
Continued (2/4)...



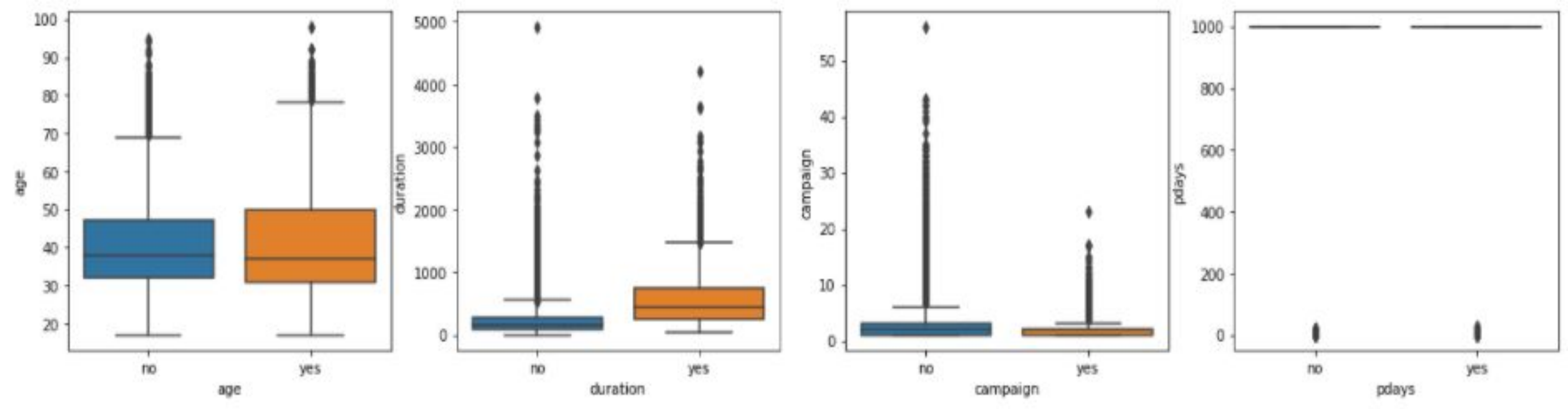
Continued (3/4)...

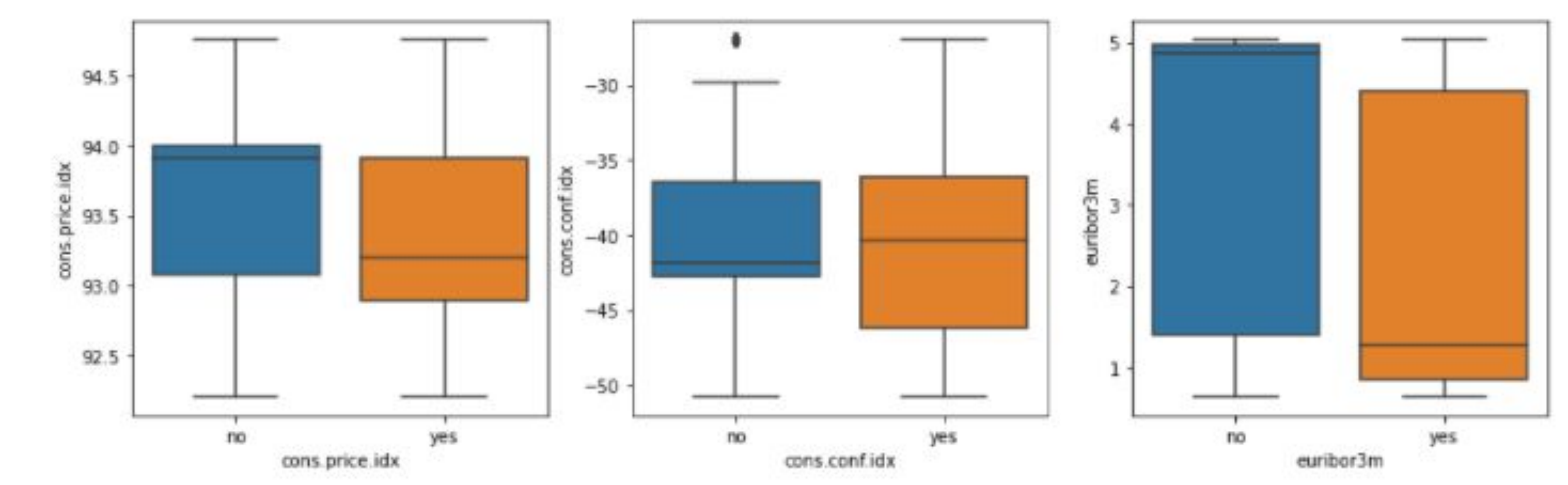


Continued (4/4)...



Numerical features (yes/no term deposit counts)




94.5
94.0
93.5
93.0
92.5
cons.price.idx
no
yes
cons.price.idx
−30
−35
−40
−45
−50
cons.conf.idx
no
yes
cons.conf.idx
5
4
3
2
1
euribor3m
no
yes
euribor3m

Correlation Matrix

Data Preprocessing

05

5.1 Duplicated Data

	age	job	marital	education	default	housing	loan	contact	month	day_of_week	duration	campaign	pdays	previous	poutcome	emp.var.rate
1265	39	blue-collar	married	basic.6y	no	no	no	telephone	may	thu	124	1	999	0	nonexistent	1.1
12260	36	retired	married	unknown	no	no	[illegible]	telephone	[illegible]	thu	88	1	999	0	nonexistent	1.4
14155	27	technician	single	professional.course	no	no	[illegible]	cellular	[illegible]	mon	331	2	999	0	nonexistent	1.4
16819	47	technician	divorced	high.school	no	yes	no	[illegible]	jul	thu	43	3	999	0	nonexistent	1.4
18464	32	technician	single	professional.course	no	yes	no	[illegible]	jul	thu	128	1	999	0	nonexistent	1.4
20072	55	services	married	high.school	unknown	no	no	[illegible]	aug	mon	33	1	999	0	nonexistent	1.4
20531	41	technician	married	professional.course	no	yes	[illegible]	[illegible]	aug	tue	127	1	999	0	nonexistent	1.4
25183	39	admin.	married	university.degree	no	no	[illegible]	cellular	[illegible]	tue	123	2	999	0	nonexistent	-0.1
28476	24	services	single	high.school	no	[illegible]	no	cellular	[illegible]	tue	114	1	999	0	nonexistent	-1.8
32505	35	admin.	married	university.degree	no	yes	no	cellular	may	fri	348	4	999	0	nonexistent	-1.8

5.2 Rescaling Numerical Data

campaign	pdays	previous
-0.565963	0.195443	-0.349551
-0.565963	0.195443	-0.349551
-0.565963	0.195443	-0.349551
-0.565963	0.195443	-0.349551
-0.565963	0.195443	-0.349551
...	...	...
-0.565963	0.195443	-0.349551
-0.565963	0.195443	-0.349551
-0.204990	0.195443	-0.349551
-0.565963	0.195443	-0.349551
0.155984	0.195443	1.670821

emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed
0.648101	0.722628	0.886568	0.712463	0.331695
0.648101	0.722628	0.886568	0.712463	0.331695
0.648101	0.722628	0.886568	0.712463	0.331695
0.648101	0.722628	0.886568	0.712463	0.331695
0.648101	0.722628	0.886568	0.712463	0.331695
...	...	...	...	...
-0.752402	2.058076	-2.225059	-1.495197	-2.815689
-0.752402	2.058076	-2.225059	-1.495197	-2.815689
-0.752402	2.058076	-2.225059	-1.495197	-2.815689
-0.752402	2.058076	-2.225059	-1.495197	-2.815689
-0.752402	2.058076	-2.225059	-1.495197	-2.815689

5.3 Other Pre-Processing Steps

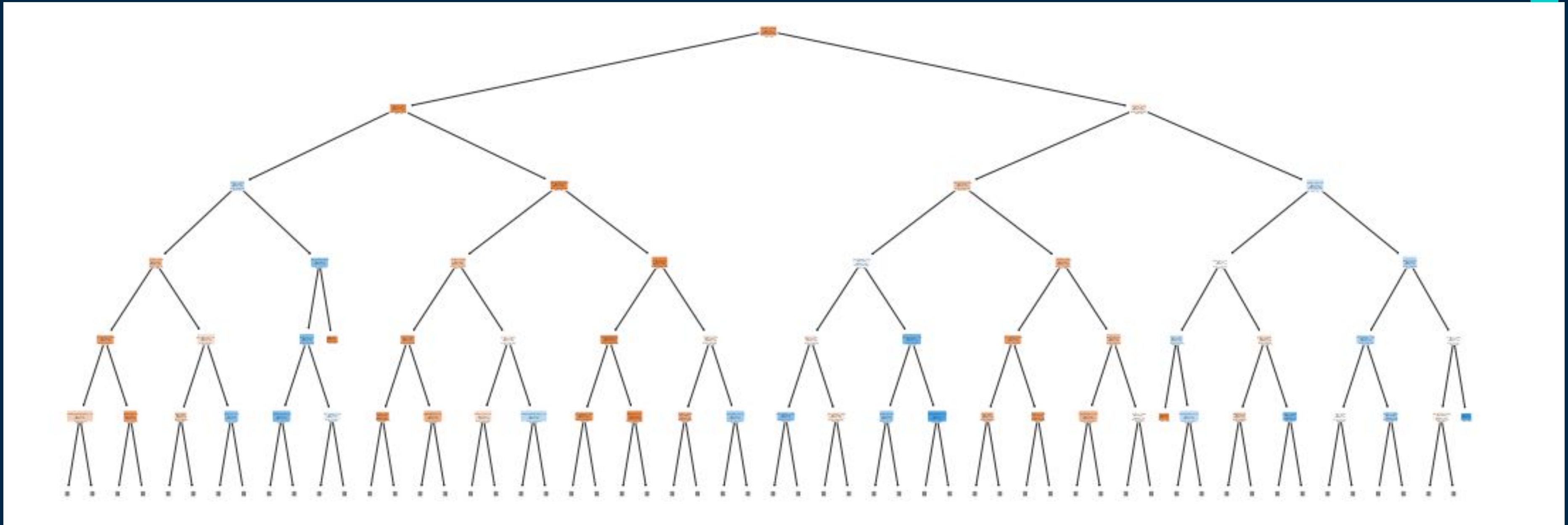
- Turning y into an integer with Yes= 1 and No=0
- Encoding categorical variables using pd.get_dummies
- Splitting the data using the random seed 43
- Balancing data : Random Oversampling

Data Modeling

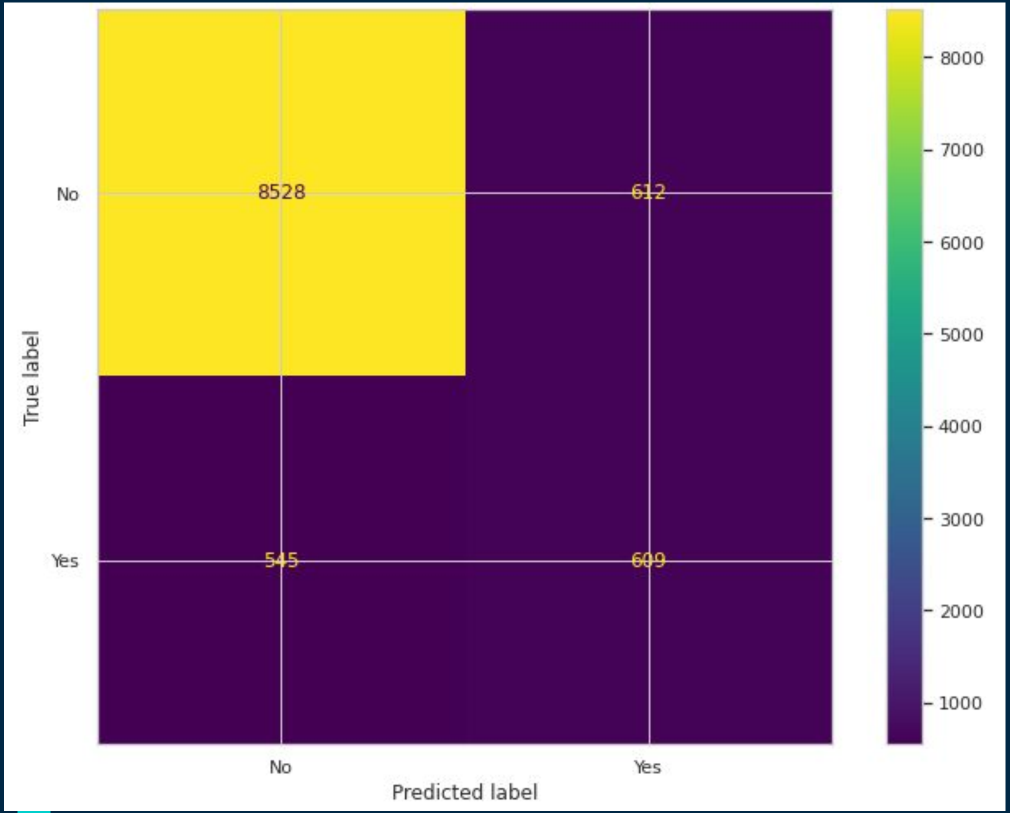
06

Decision Tree

Decision Tree (imbalanced data)



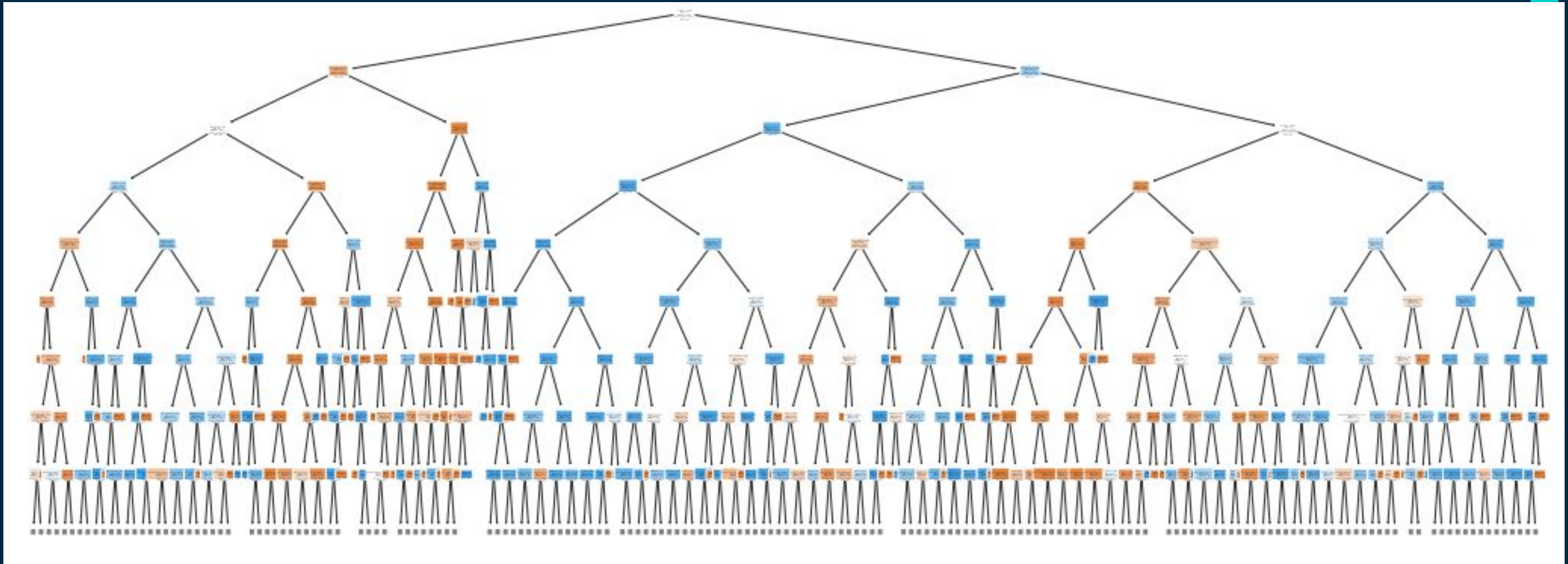
Confusion Matrix of the Testing Set



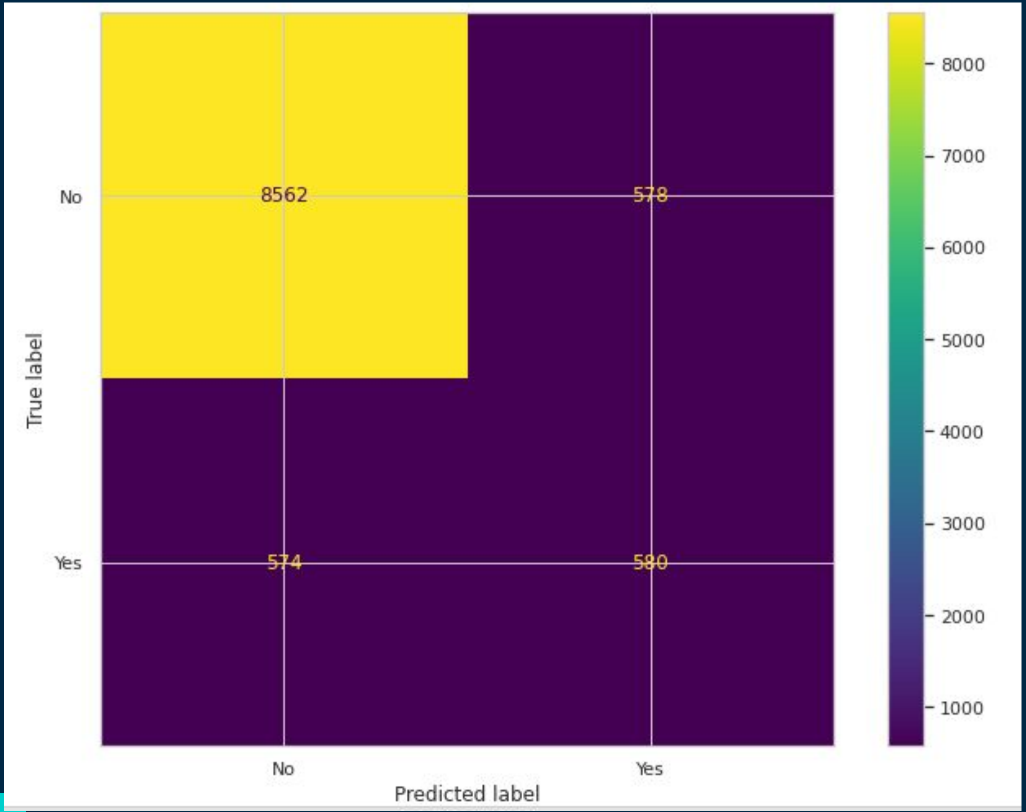
Accuracy =
0.888

Precision =
0.49877

Decision Tree (Balanced Data)



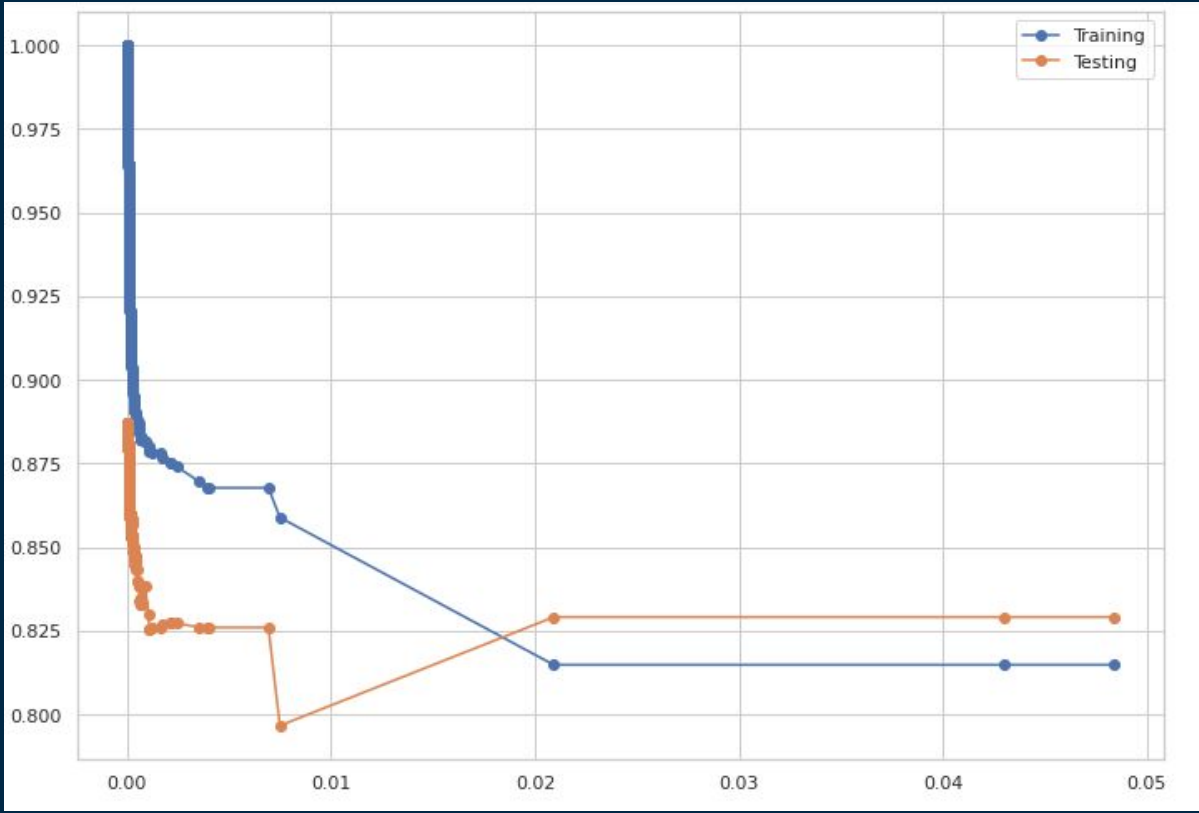
6.2.1 Confusion Matrix of the Testing Set



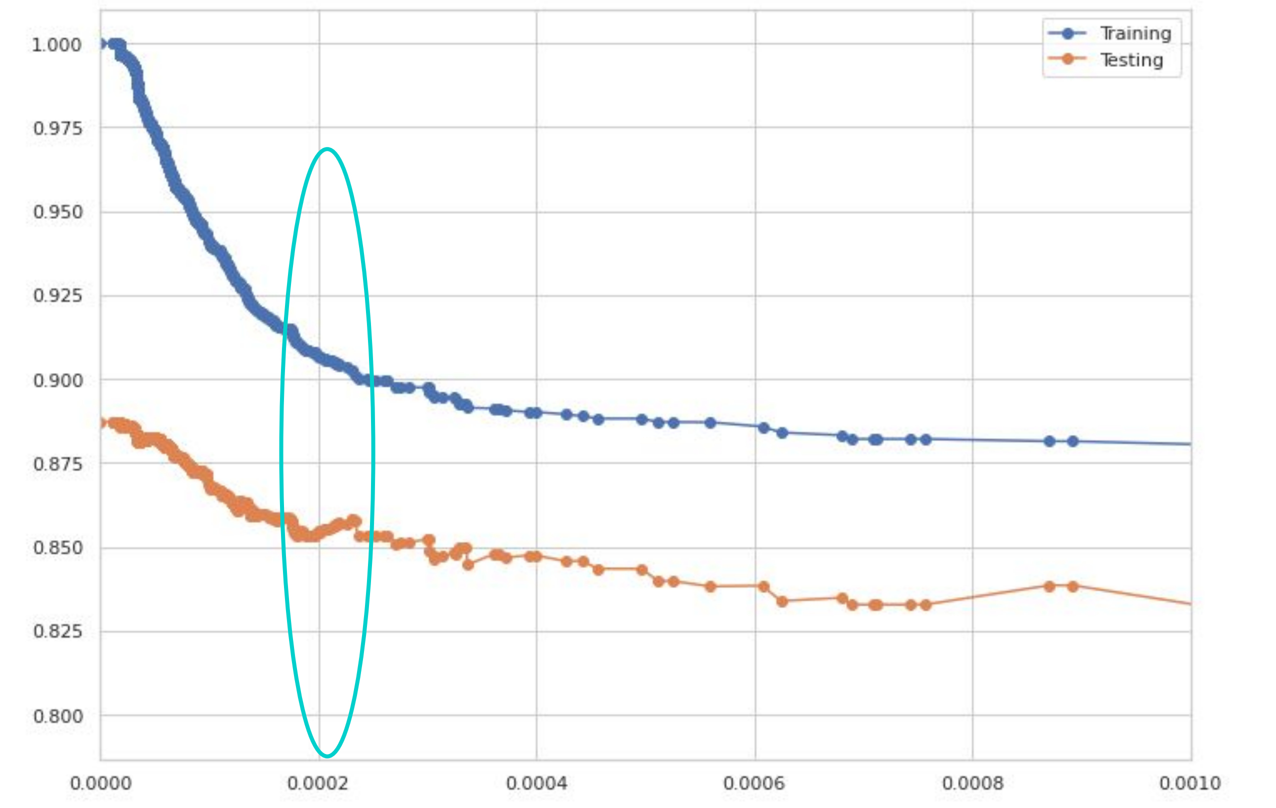
Accuracy:
0.88809

Precision:
0.5008

6.2.2 Cost Complexity Pruning



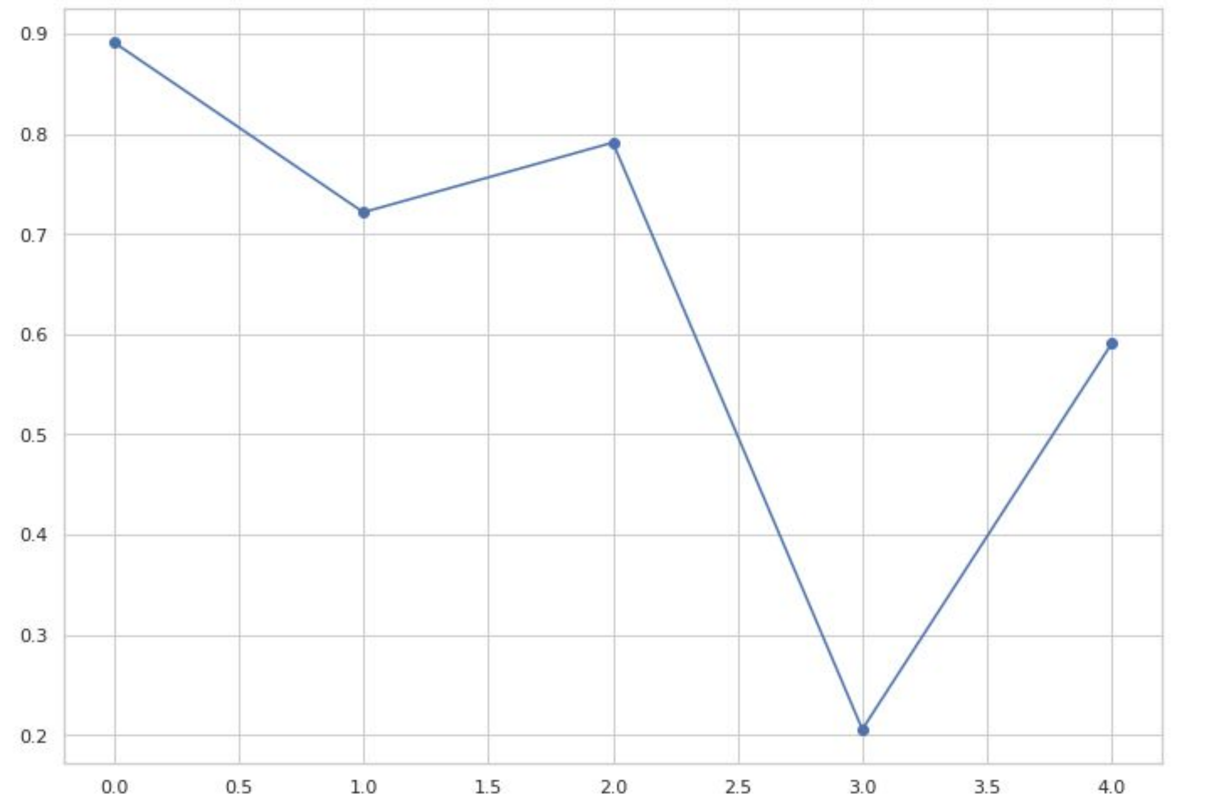
ZOOMED-IN GRAPH BETWEEN 0 AND 0.001



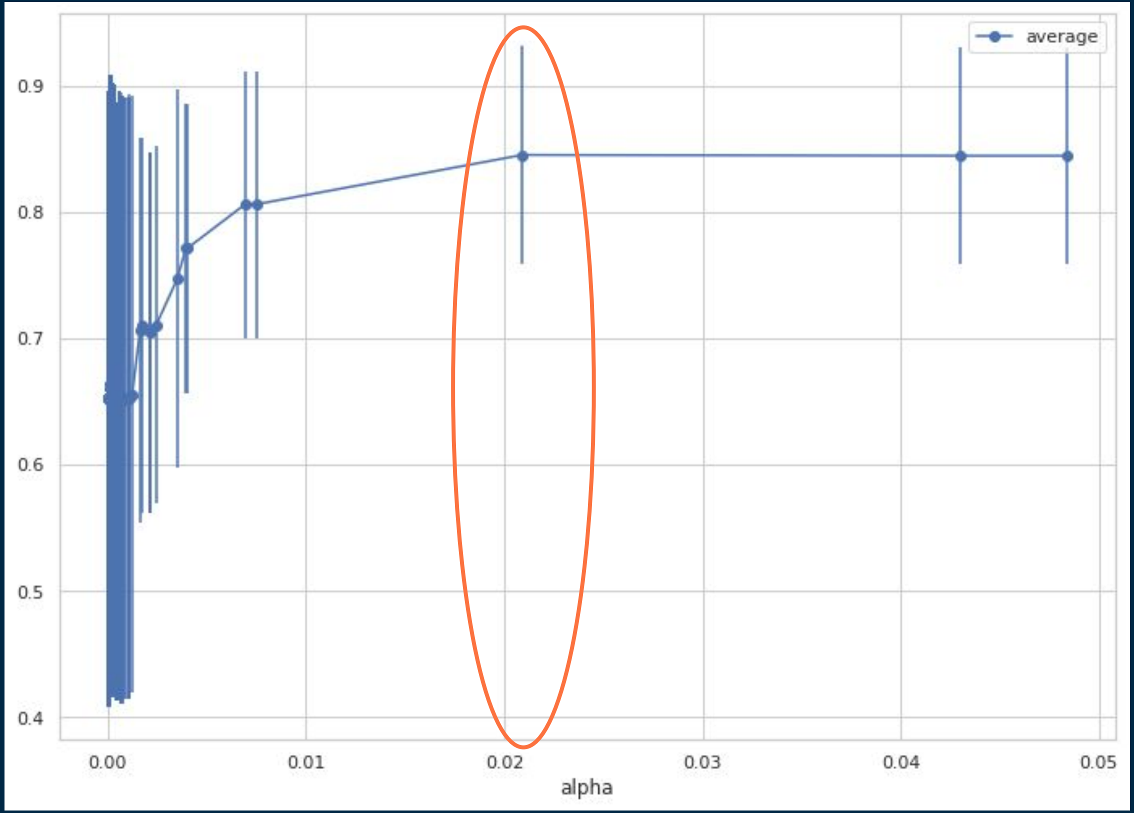
alpha =

0.000206207852
253893

Cross Validation for the Best Alpha

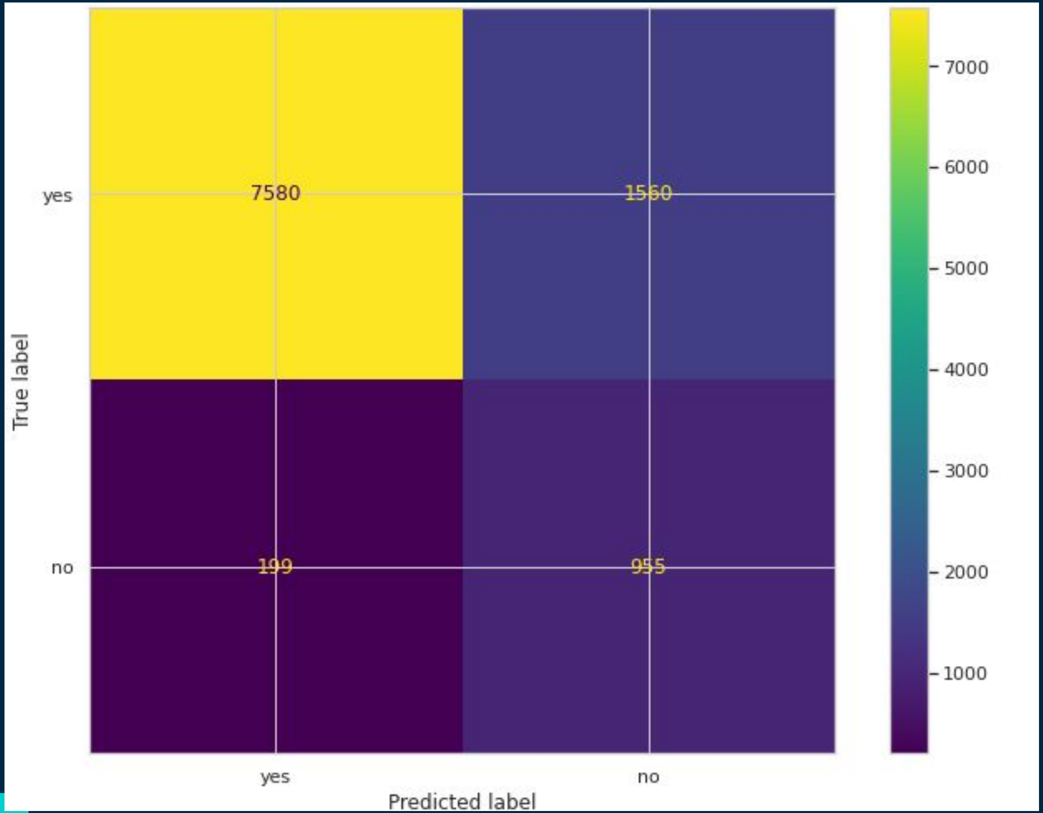


Plot Overall Scores



alpha_ideal = 0.021

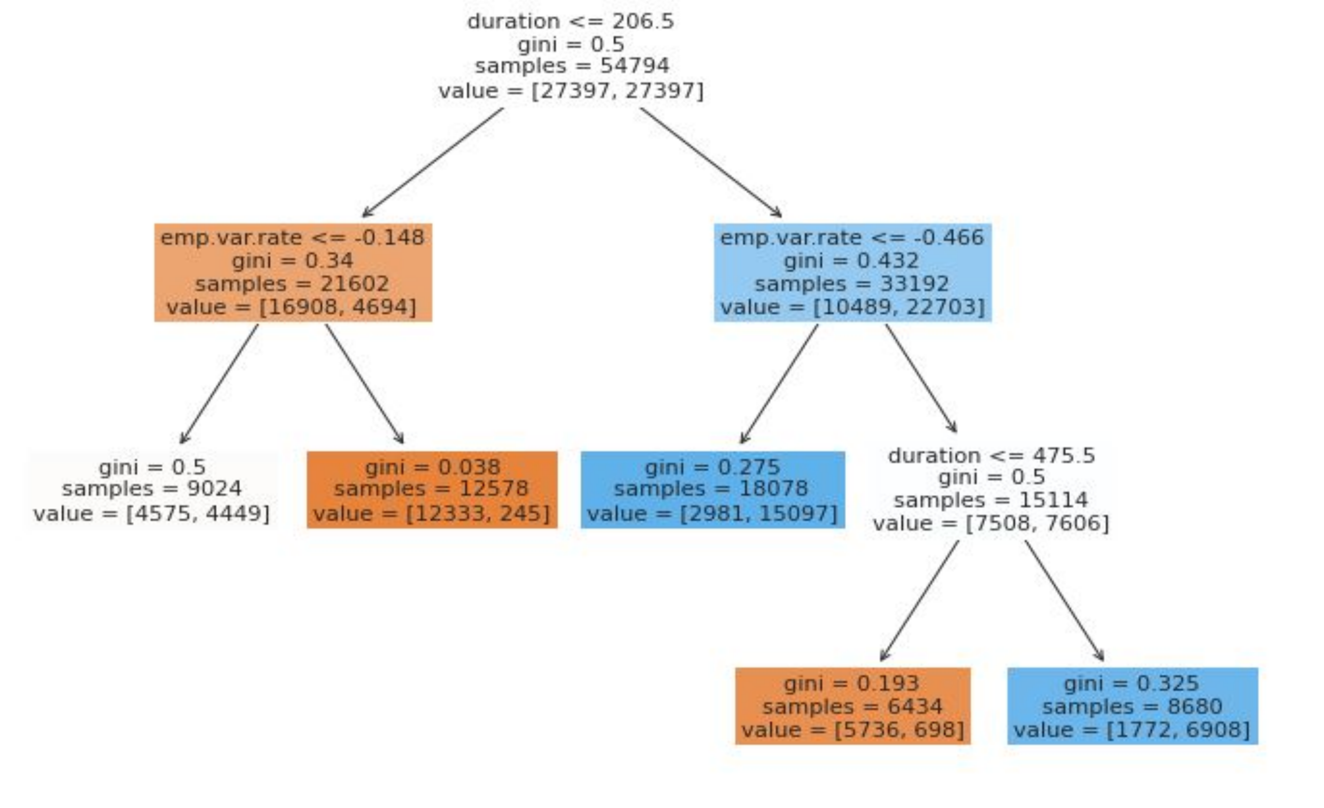
Confusion Matrix of the Testing Set after finding the Ideal Alpha



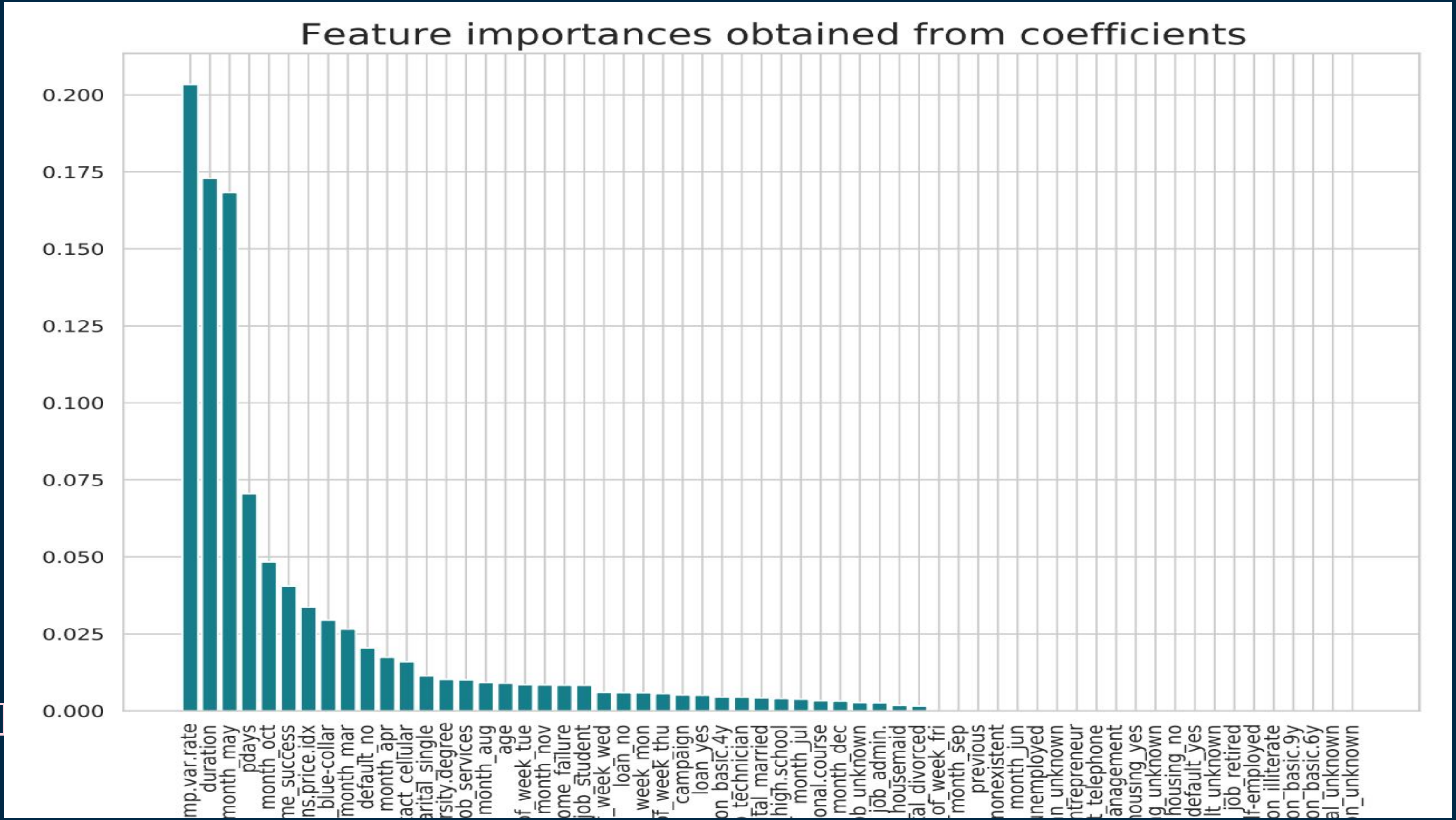
Accuracy :
0.84387

Precision:
0.9744

Ideal Decision Tree

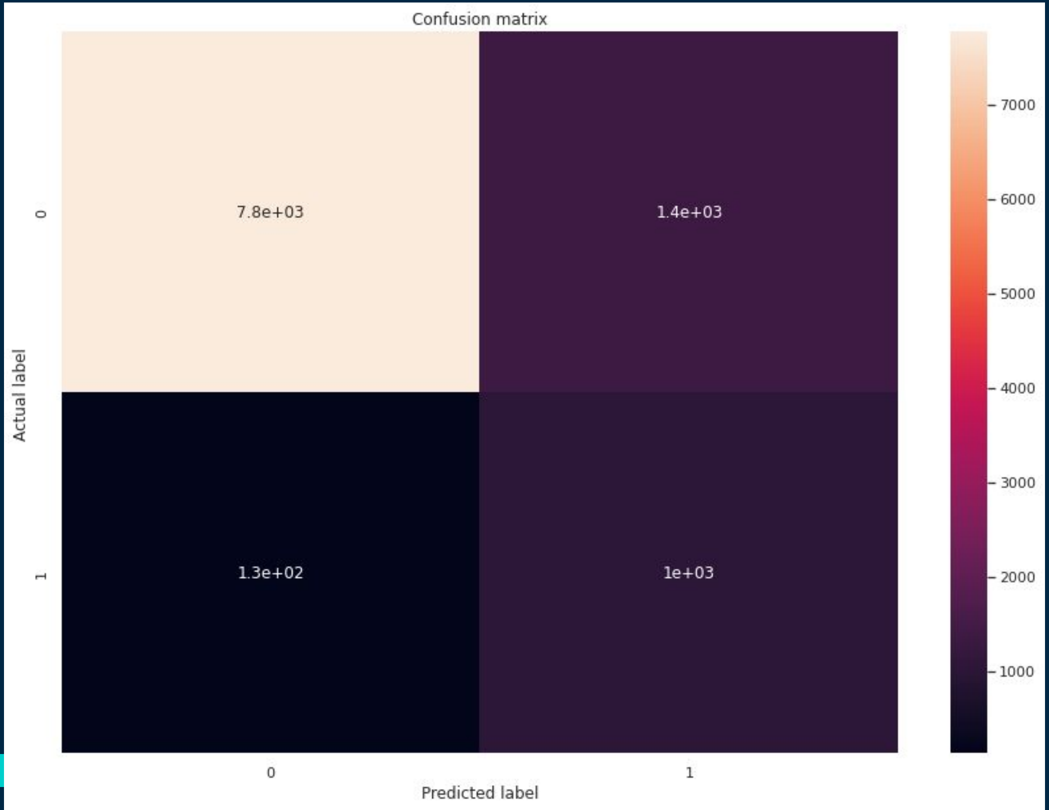


Feature Importance



Logistic Regression

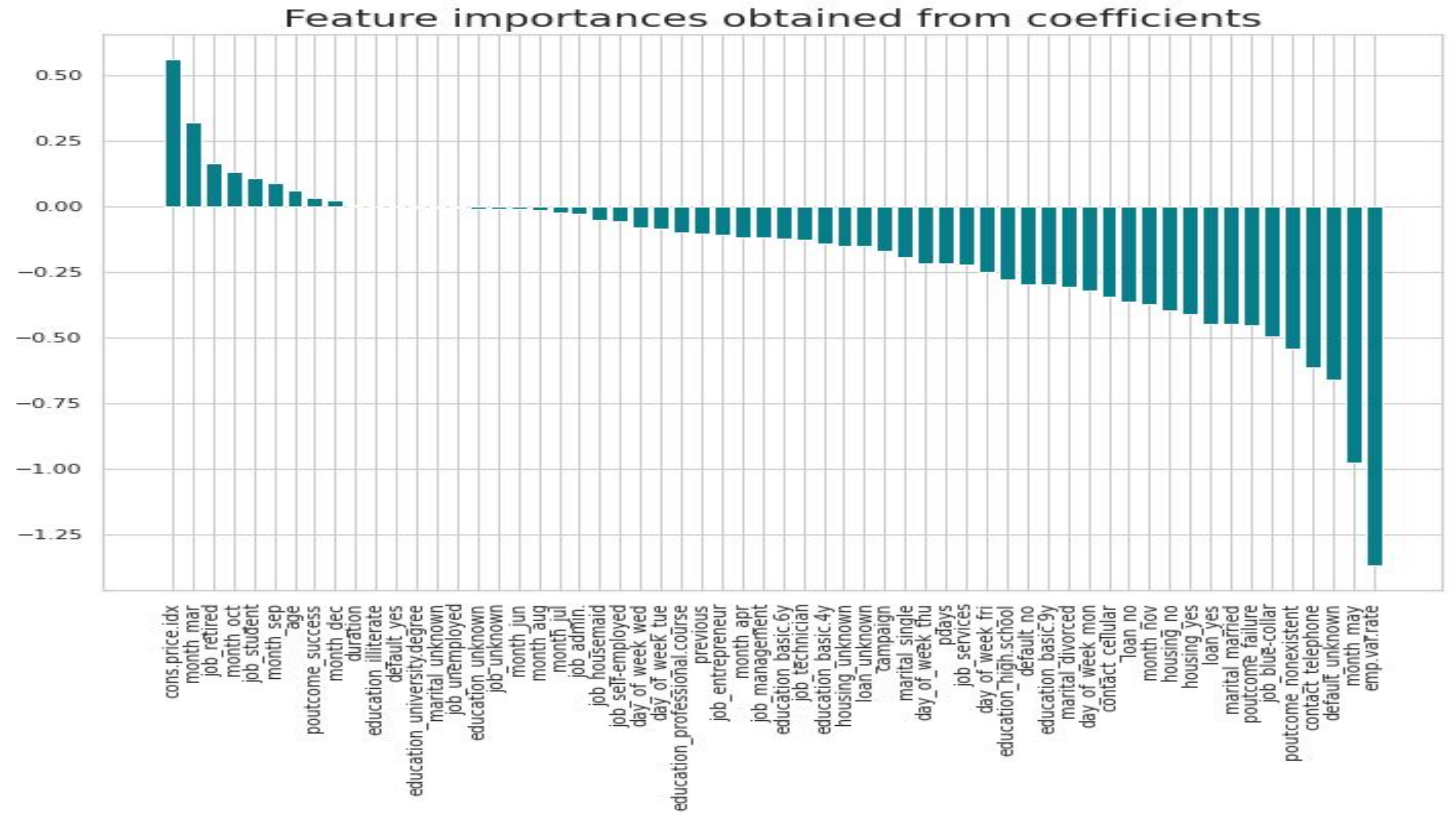
Logistic Regression Confusion Matrix



The accuracy:
0.708014

Precision:
0.4274395329441201

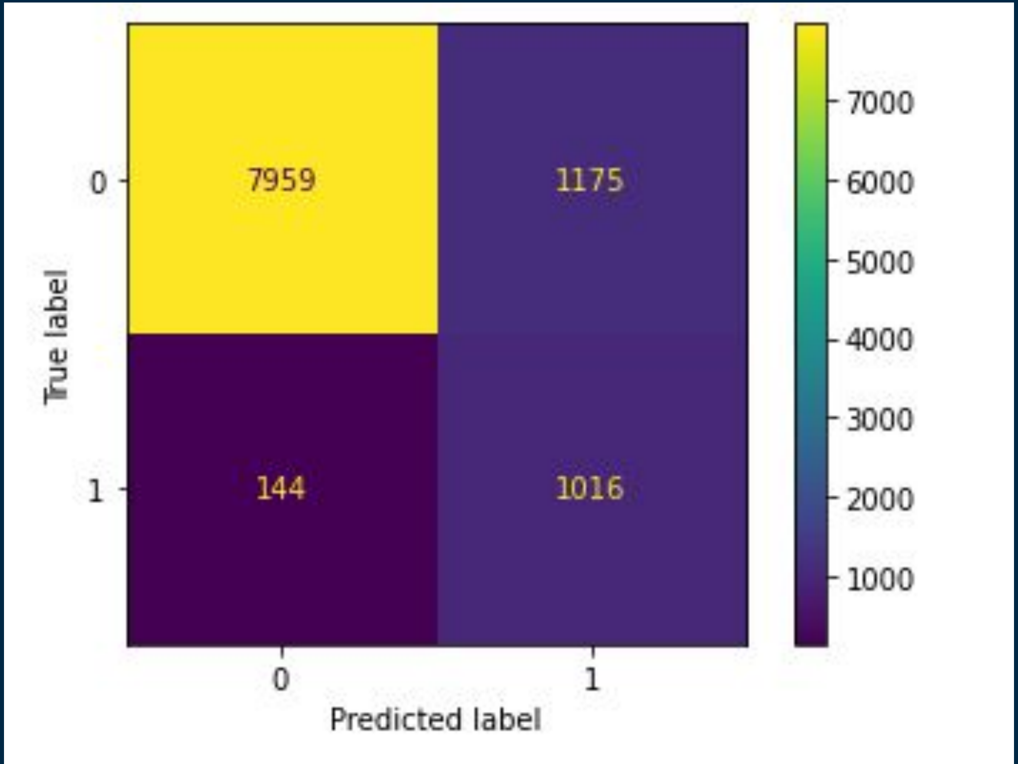
Recall:
0.8882149046793761



Feature importances obtained from coefficients
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−1.25
cons.price.idx
month_mar
job_retired
month_oct
job_student
month_sep
age
poutcome_success
month_dec
duration
education_illiterate
default_yes
education_university.degree
marital_unknown
job_unemployed
education_unknown
job_unknown
month_jun
month_aug
month_jul
job_admin.
job_housemaid
job_self-employed
day_of_week_wed
day_of_week_tue
education_professional.course
previous
job_entrepreneur
month_apr
job_management
education_basic.6y
job_technician
education_basic.4y
housing_unknown
loan_unknown
campaign
marital_single
day_of_week_thu
pdays
job_services
day_of_week_fri
education_high.school
default_no
education_basic.9y
marital_divorced
day_of_week_mon
contact_cellular
loan_no
month_nov
housing_no
housing_yes
loan_yes
marital_married
poutcome_failure
job_blue-collar
poutcome_nonexistent
contact_telephone
default_unknown
month_may
emp.var.rate

ADA BOOST

ADA BOOST Confusion Matrix after Finding the best Model

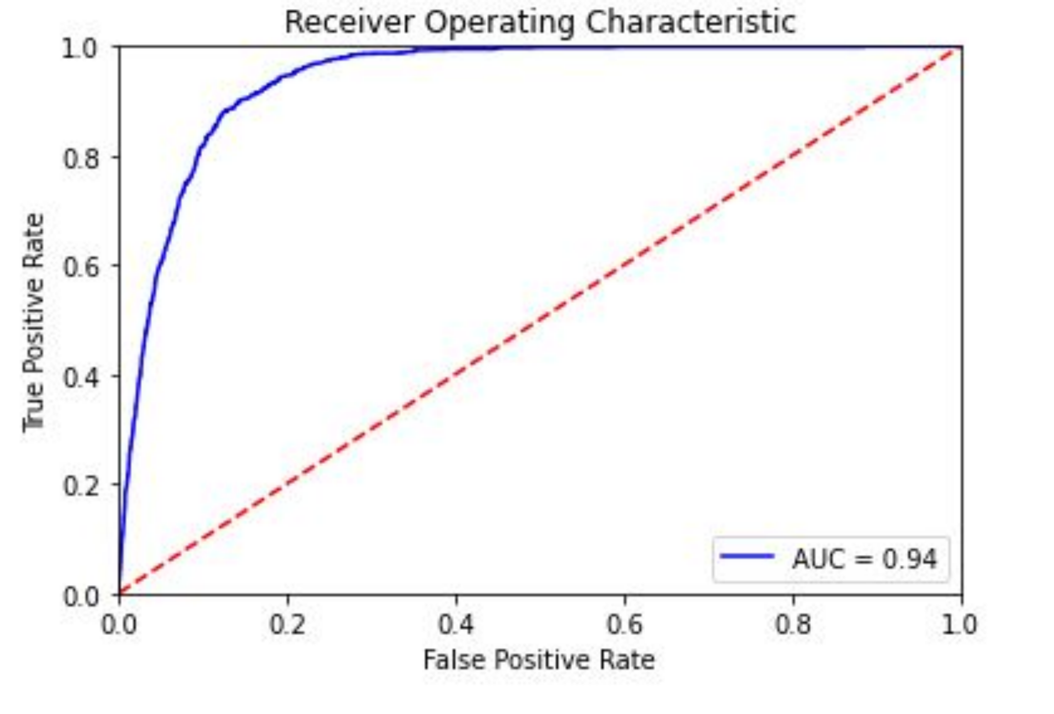


Accuracy: 0.87

Precision: 0.46

Learning rate of 0.2
Number of estimators as 200

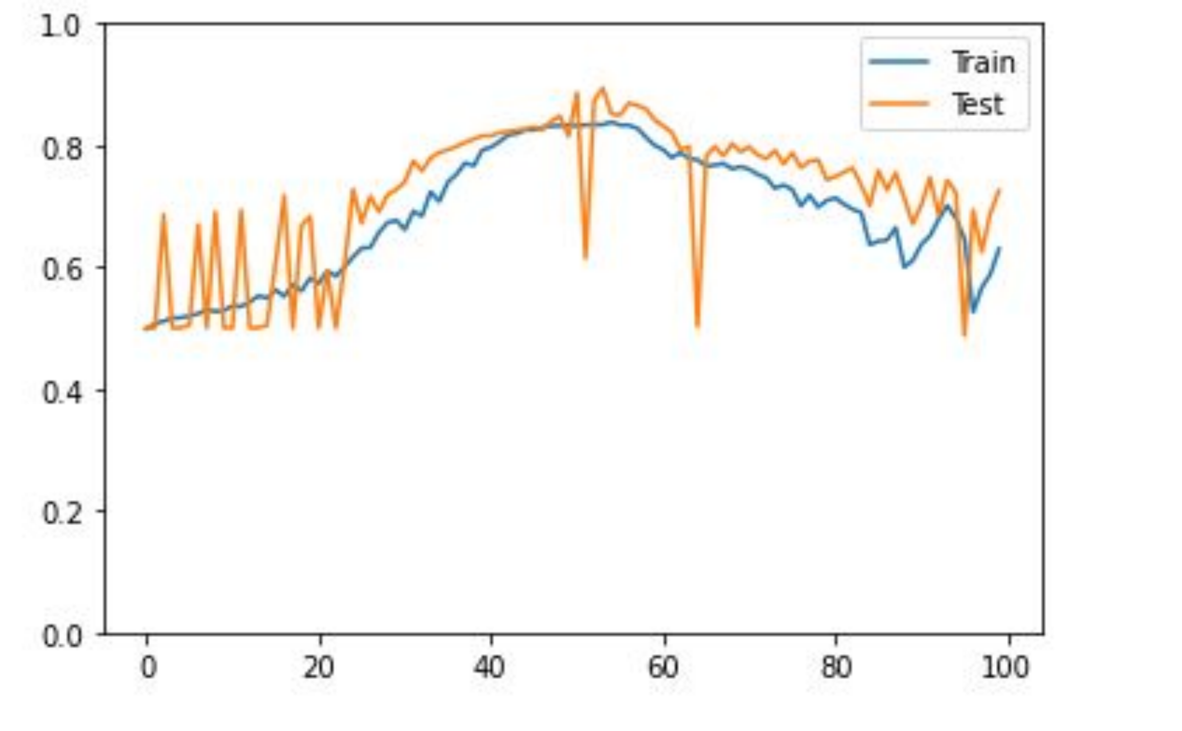
ADA BOOST AUC Curve



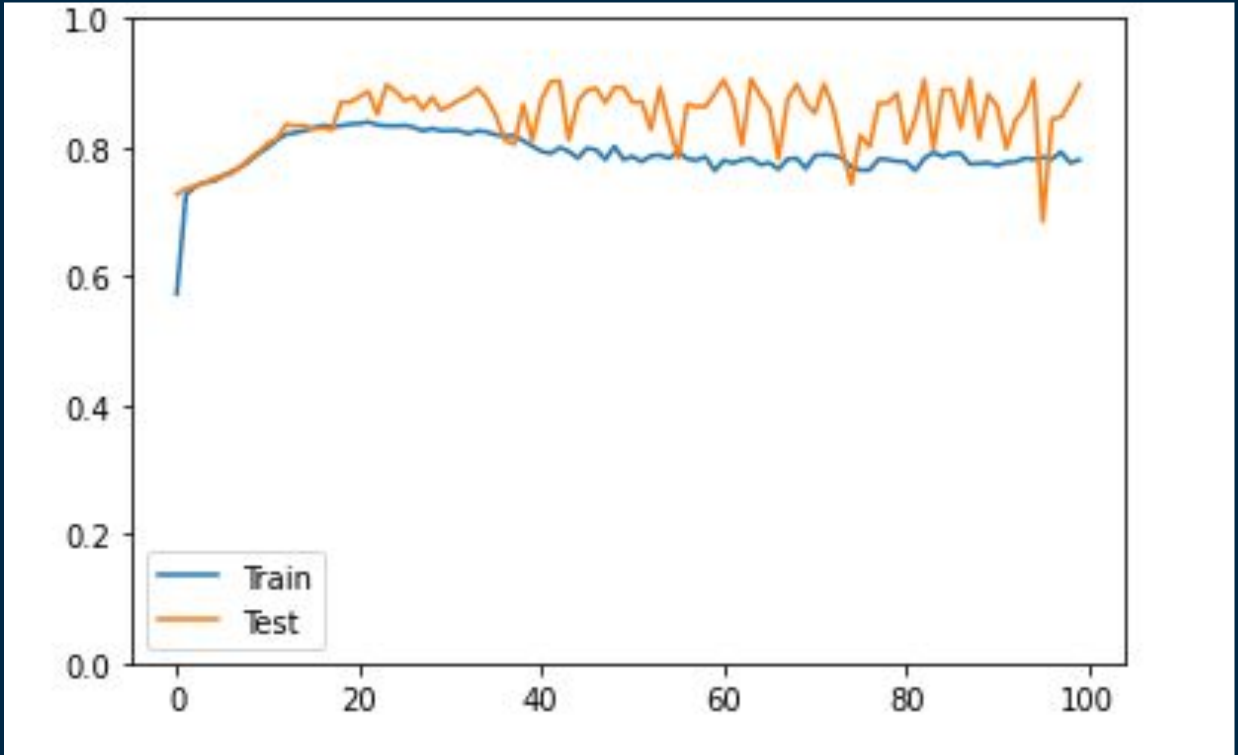
AUC score is = 0.94

Neural Network

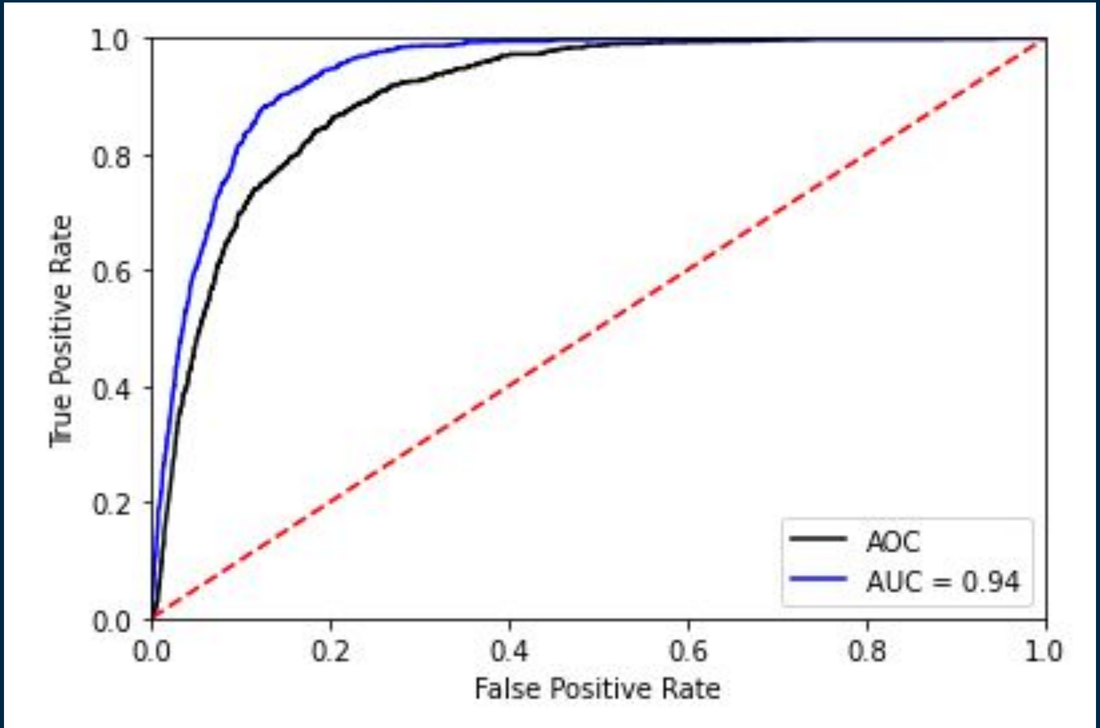
Model Accuracy (auc_1)



Model Accuracy (AUC_7)



6.5.3.Neural Network AUC Curve



AUC score is = 0.94

Recommendations 07

- All models produced an AUC score of around 0.94(all are similar).

- From the feature importance bar chart, the most important features are : employment variation rate, month, and pdays (number of days that passed by after the client was last contacted from a previous campaign).

- Other Features: jobs, marital status and p-outcome.

Jobs: Blue-Collars

Most people who subscribed are blue-collars.

Month: May, Oct,Mar

Most important months (peaks, in order):

- May
- October
- March

Least Important (troughs):

- December

Marital Status: Single

Most of subscribers are single.

Poutcome: Success

If they have been previously contacted and subscribed, then they are likely to subscribe again

Do you have any questions?

Thank you for your attention

CREDITS: This presentation template was created by Slidesgo, including icons by Flaticon, and infographics & images by Freepik