techready PRO Professional Academy digitalent KOMINFO

Machine Learning with Tensorflow

Prediction Model for Direct Marketing Campaign Conversion on Term Deposit Product in Banking Sector

Group FIN7

INTRODUCTION

01

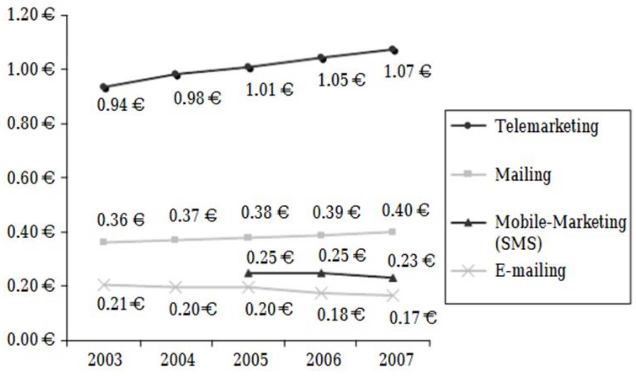
BACKGROUND SUMMARY

01

Problem Statement

Recent studies suggest that as of June 2020, companies and organizations within the banking and financial sector are spending an **average of 13% of their overall budget on marketing**. One of alternative in marketing is telemarketing or direct marketing by telephone. Telemarketing can be an effective method for reaching your customers. The telephone can be used to contact new customers, to maintain contact with current customers and to remind slow payers personally. One advantage to telemarketing is that you will get an instant response.

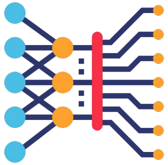
However, as seen at Figure 1-1, cost per contact for telemarketing has rising up from 2003 to 2007 and expected to be much higher now at 2022. Therefore, **to maximize profitability, we need to reduce these marketing cost** or increase the effectiveness of the cost with higher conversion.



Proposed Solution

To be cost effective, companies must target their marketing plan. It is more cost effective for companies to spend resource on customers which more likely to convert or subscribe to their products. We will develop a **model based on historical marketing campaign efforts and results to predict with high accuracy whether a potential customer will subscribe to our products upon contact**. Using this model, companies can save resources on marketing cost by focusing marketing efforts on customers which yield positive result.

Business Formulation



ML Objective

Predict the customer who will subscribe a term deposit upon contact

Actions

Call only potential customers, Give treatment to customers who are likely to reject the campaign

Values

Reduce marketing cost or increase the effectiveness of the cost with higher conversion, Maximize profitability

DATA UNDERSTANDING

02

Dataset

The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

The data folder contains two data sets:

- train.csv: 45,211 rows and 18 columns sorted by date (from May 2008 to November 2010)
- test.csv: 4521 rows and 18 columns with 10% examples (4521), randomly selected from train .csv

Dataset

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
0	58	management	married	tertiary	no	2143	yes	no	unknown	5	may	261	1	-1	0	unknown	no
1	44	technician	single	secondary	no	29	yes	no	unknown	5	may	151	1	-1	0	unknown	no
2	33	entrepreneur	married	secondary	no	2	yes	yes	unknown	5	may	76	1	-1	0	unknown	no
3	47	blue-collar	married	unknown	no	1506	yes	no	unknown	5	may	92	1	-1	0	unknown	no
4	33	unknown	single	unknown	no	1	no	no	unknown	5	may	198	1	-1	0	unknown	no
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
45206	51	technician	married	tertiary	no	825	no	no	cellular	17	nov	977	3	-1	0	unknown	yes
45207	71	retired	divorced	primary	no	1729	no	no	cellular	17	nov	456	2	-1	0	unknown	yes
45208	72	retired	married	secondary	no	5715	no	no	cellular	17	nov	1127	5	184	3	success	yes
45209	57	blue-collar	married	secondary	no	668	no	no	telephone	17	nov	508	4	-1	0	unknown	no
45210	37	entrepreneur	married	secondary	no	2971	no	no	cellular	17	nov	361	2	188	11	other	no

Attribute Information

Bank Client Data			
1	age	numeric	(in years)
2	job	categorical	“admin”,”unknown”, “unemployed”, ”management”, ”housemaid”, ”entrepreneur”, ”student”, ”blue-collar”, ”self-employed”, ”retired”, ”technician”, ”services”
3	marital	categorical	“married”,”divorced”,”single” (divorced means divorced or widowed)
4	education	categorical	“unknown”,”secondary”,”primary”,”tertiary”
5	default	binary	Has credit in default? (“yes” or “no”)
6	balance	numeric	Average yearly balance (in euros)
7	housing	binary	Has housing loan? (“yes” or “no”)
8	loan	binary	Has personal loan? (“yes” or “no”)

Attribute Information

Related with the last contact of the current campaign			
9	contact	categorical	Contact communication type (“unknown”, “telephone”, “cellular”)
10	day	numeric	Last contact day of the month (1 - 31)
11	month	categorical	Last contact month of the year (“jan”, “feb”, “mar”, … , “nov”, “dec”)
12	duration	numeric	Last contact duration (in seconds)
13	campaign	numerical	Number of contacts performed **during** this campaign for this client (include last contact)
14	pdays	numeric	Number of days passed by after last contact from previous campaign (in days) (-1 means client was not previously contacted)
15	previous	numeric	Number of contacts performed **before** this campaign for this client
16	poutcome	categorical	Outcome of previous marketing campaign (“unknown”, “other”, “failure”, “success”)

Attribute Information

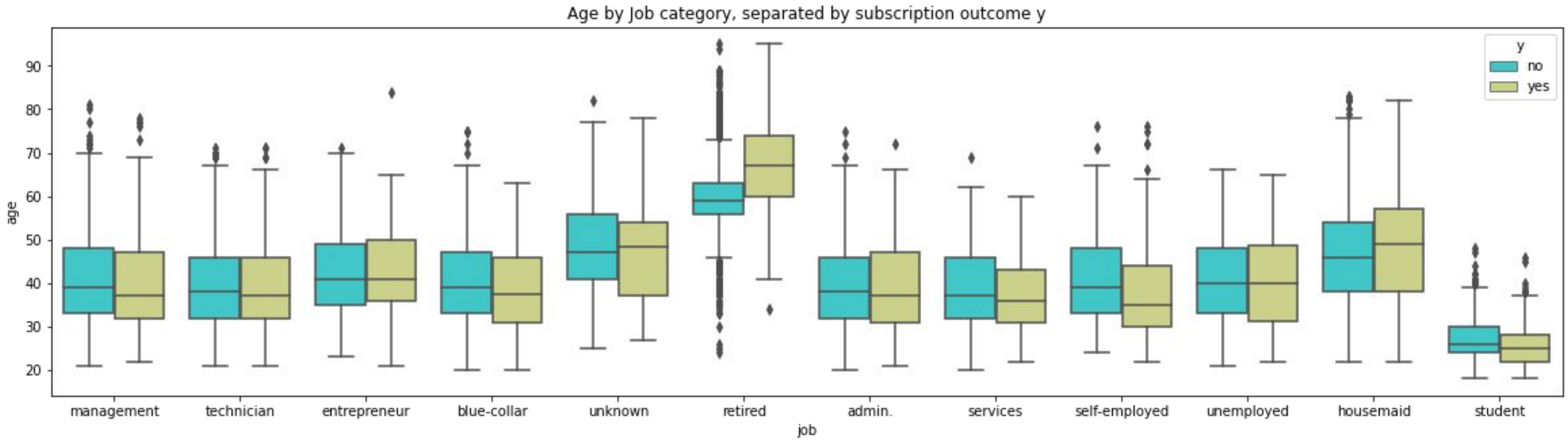
Output variable (desired target)			
17	y	binary	has the client subscribed a term deposit? ("yes" or "no")

Attribute Information

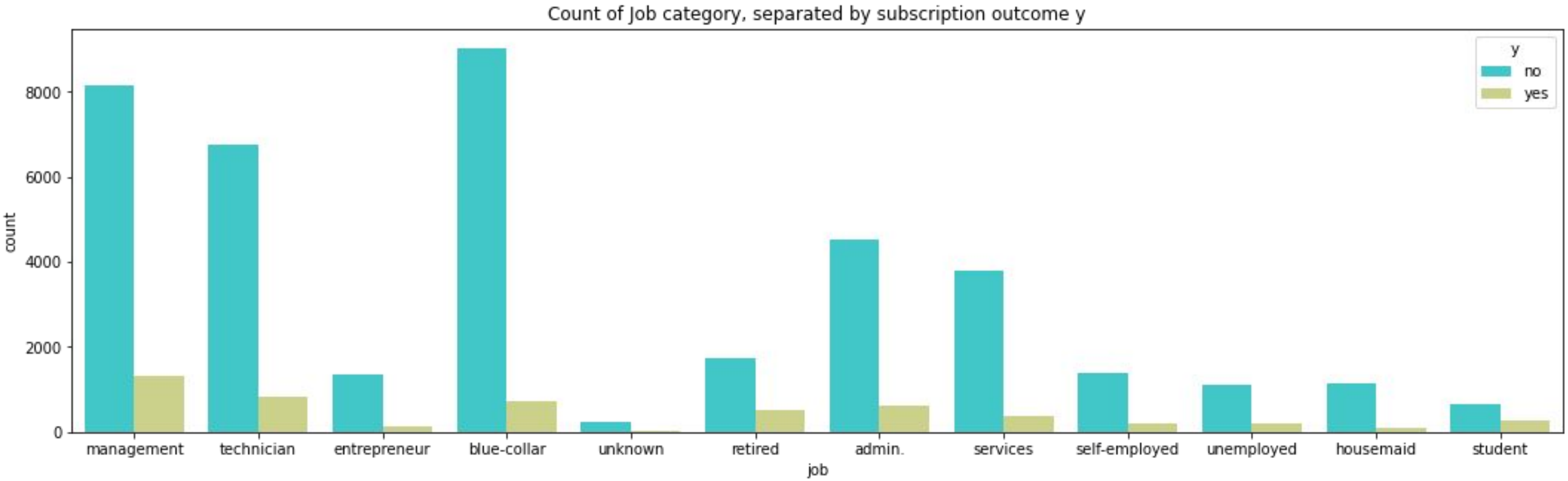
df_train.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 45211 entries, 0 to 45210
Data columns (total 17 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   age        45211 non-null  int64
 1   job        45211 non-null  object
 2   marital    45211 non-null  object
 3   education  45211 non-null  object
 4   default    45211 non-null  object
 5   balance    45211 non-null  int64
 6   housing    45211 non-null  object
 7   loan       45211 non-null  object
 8   contact    45211 non-null  object
 9   day        45211 non-null  int64
 10  month      45211 non-null  object
 11  duration   45211 non-null  int64
 12  campaign   45211 non-null  int64
 13  pdays      45211 non-null  int64
 14  previous   45211 non-null  int64
 15  poutcome   45211 non-null  object
 16  y          45211 non-null  object
dtypes: int64(7), object(10)
memory usage: 5.9+ MB
```

Insights

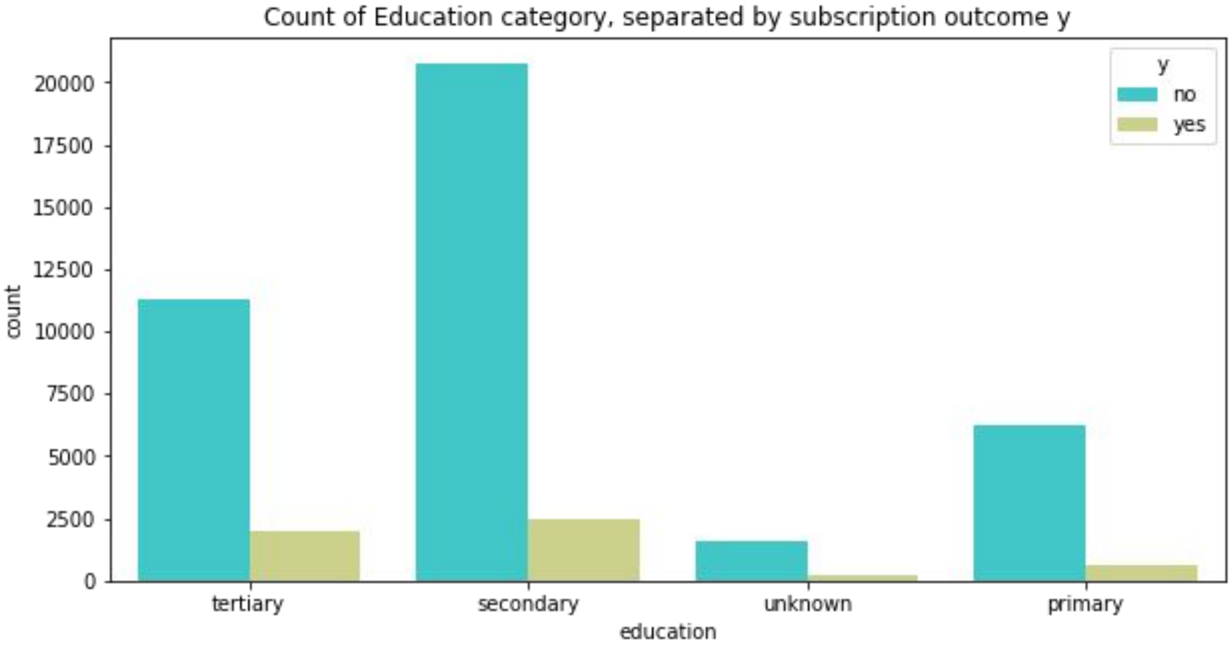

Age by Job category, separated by subscription outcome y
y
no
yes
age
90
80
70
60
50
40
30
20
management
technician
entrepreneur
blue-collar
unknown
retired
admin.
services
self-employed
unemployed
housemaid
student
job

Insights



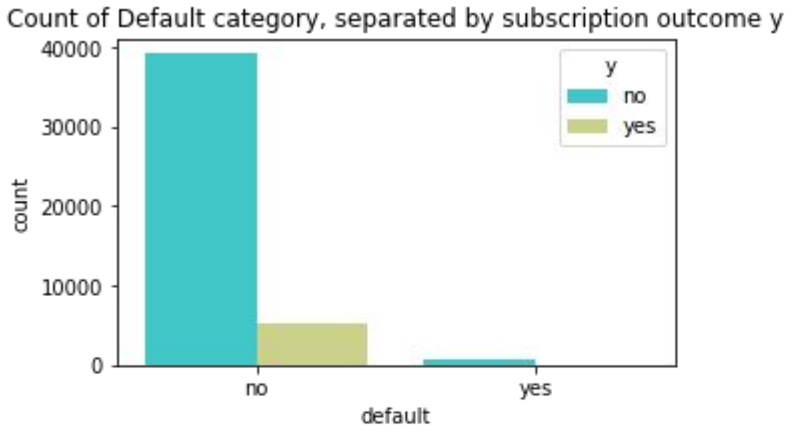
Count of Job category, separated by subscription outcome y
y
no
yes
count
8000
6000
4000
2000
0
management
technician
entrepreneur
blue-collar
unknown
retired
admin.
services
self-employed
unemployed
housemaid
student
job

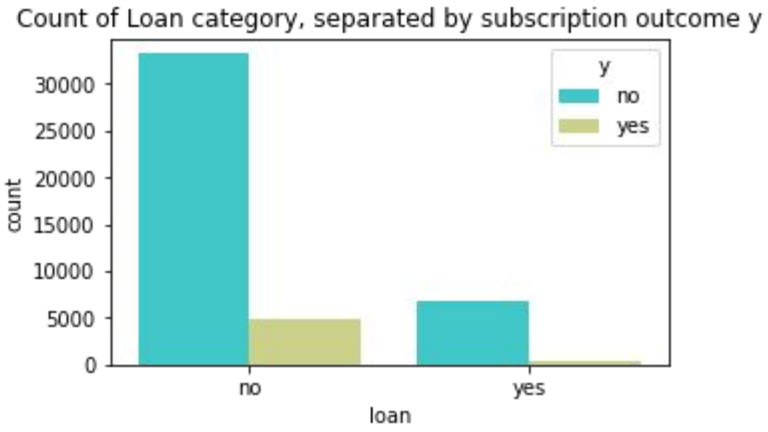
Insights

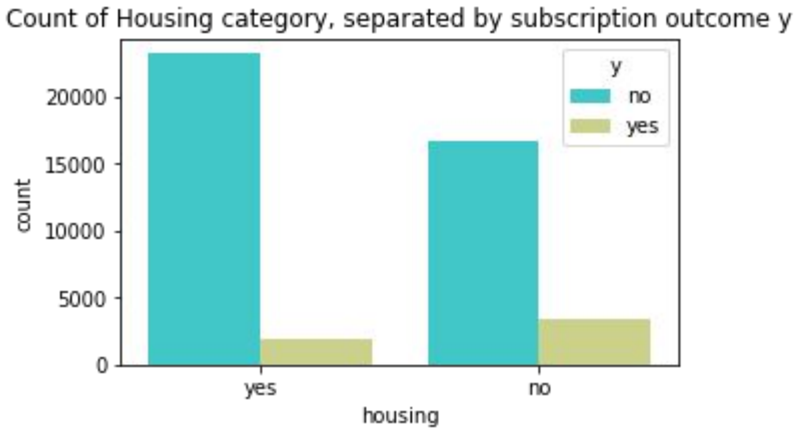


Count of Education category, separated by subscription outcome y
y
no
yes
20000
17500
15000
12500
10000
7500
5000
2500
0
count
tertiary
secondary
unknown
primary
education

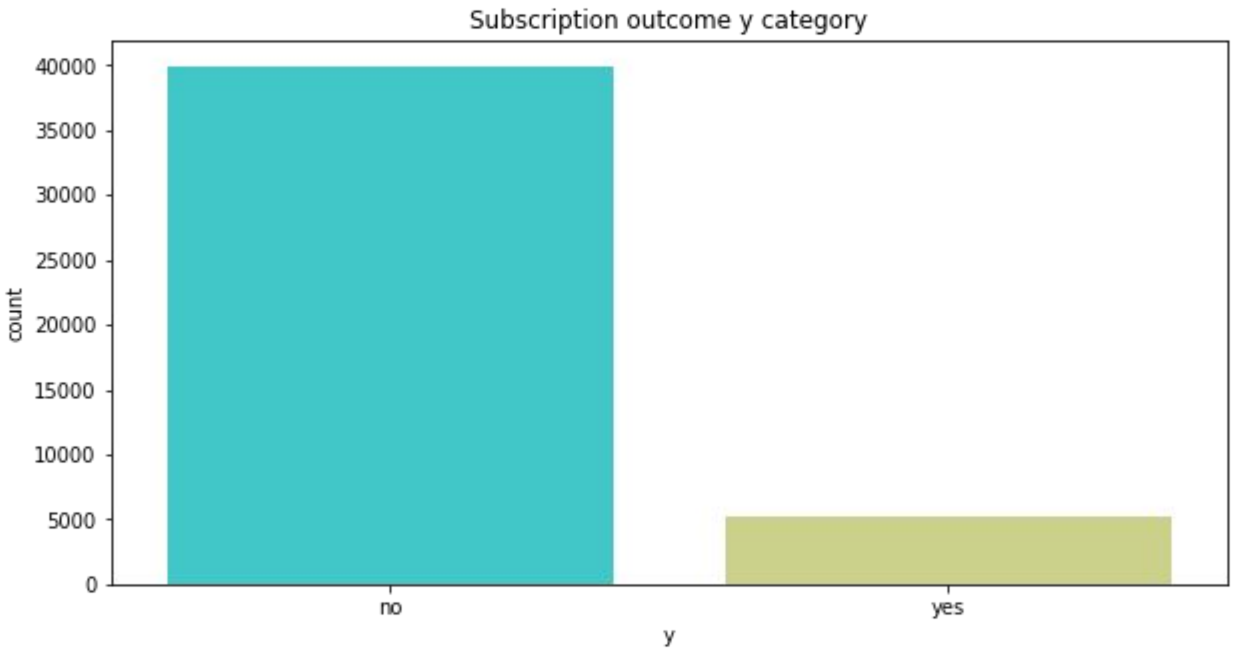
Insights


Count of Default category, separated by subscription outcome y
40000
30000
20000
10000
0
count
y
no
yes
no
yes
default


Count of Loan category, separated by subscription outcome y
30000
25000
20000
15000
10000
5000
0
count
y
no
yes
no
yes
loan


Count of Housing category, separated by subscription outcome y
20000
15000
10000
5000
0
count
y
no
yes
yes
no
housing

Insights


Subscription outcome y category
count
40000
35000
30000
25000
20000
15000
10000
5000
0
no
yes
y

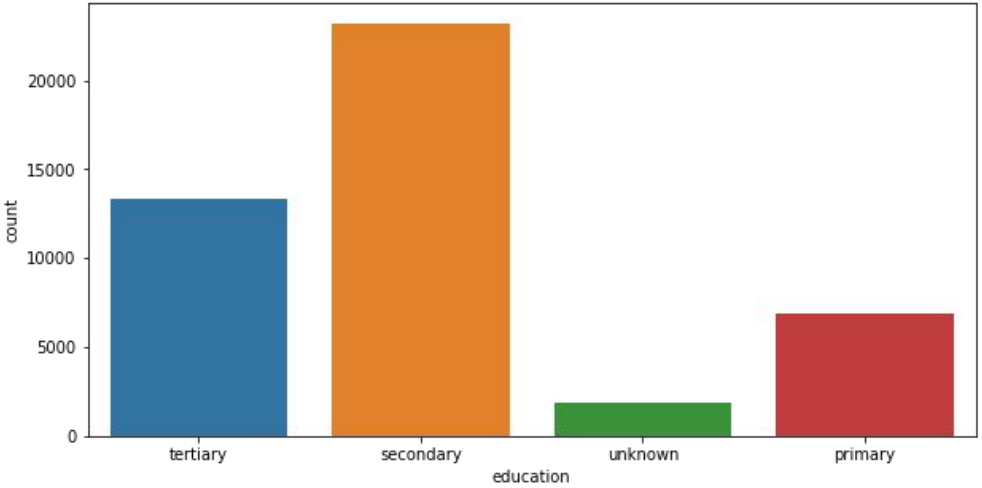
DATA MODELING

03

Preprocessing

- Replace unknown data
  - Education, Contact, and Poutcome columns.
- Object type to Integer type
  - Month and Job columns.
- Integer type to Categorial type
  - Age, Campaign, Pdays, and Previous columns.
- Standard Scaler
  - Balance column.

Preprocessing Education column

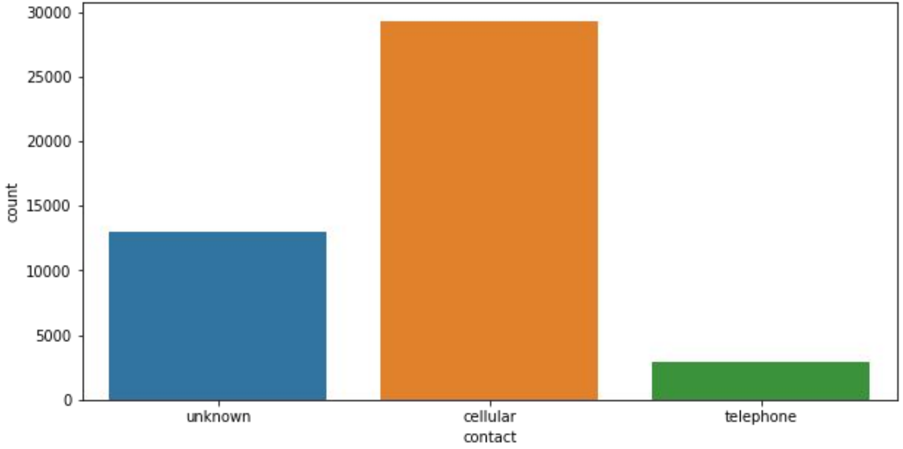


unknown → secondary

education
tertiary
secondary
unknown
primary

education	edu_imput
tertiary	0
secondary	0
secondary	1
primary	0

Preprocessing Cellular column

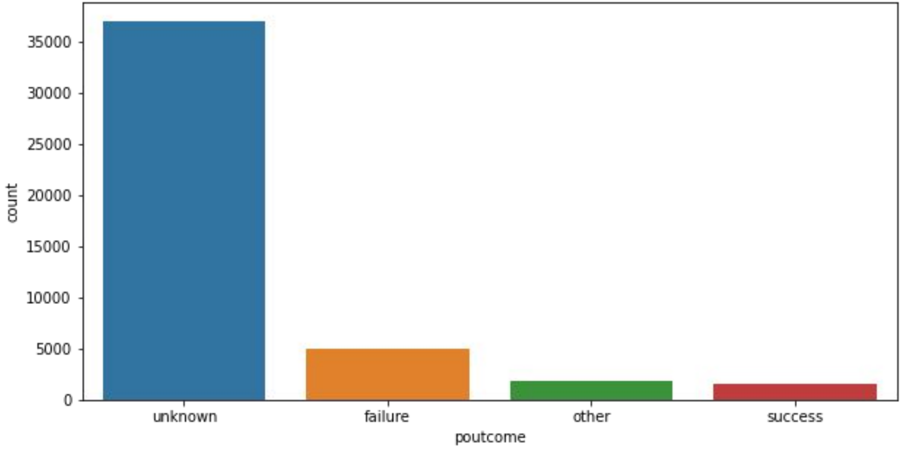


unknown → cellular

cellular
unknown
cellular
unknown
telephone

cellular	cell_imput
cellular	1
cellular	0
cellular	1
telephone	0

Preprocessing Poutcome column



unknown
other

→ no-data

poutcome
unknown
failure
other
succes
other
unknown

→

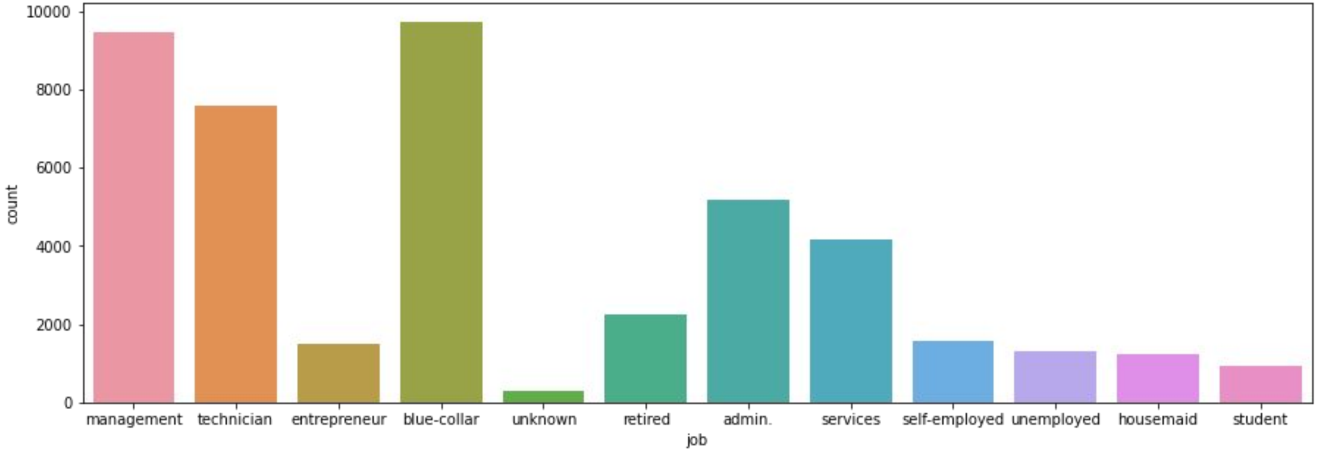
poutcome
no-data
failure
no-data
succes
no-data
no-data

Preprocessing Month and Job columns

- Month column

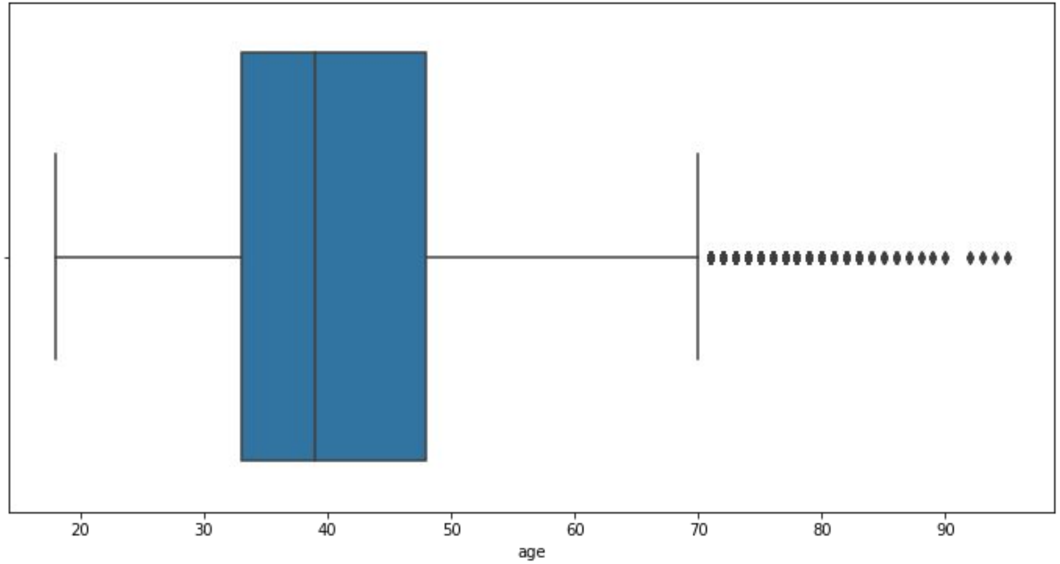
  ["jan", "feb", "mar", … , "nov", "dec"] → [1, 2, 3, …, 11, 12]

- Job column



→ [1, 2, 3, …, 11, 12]

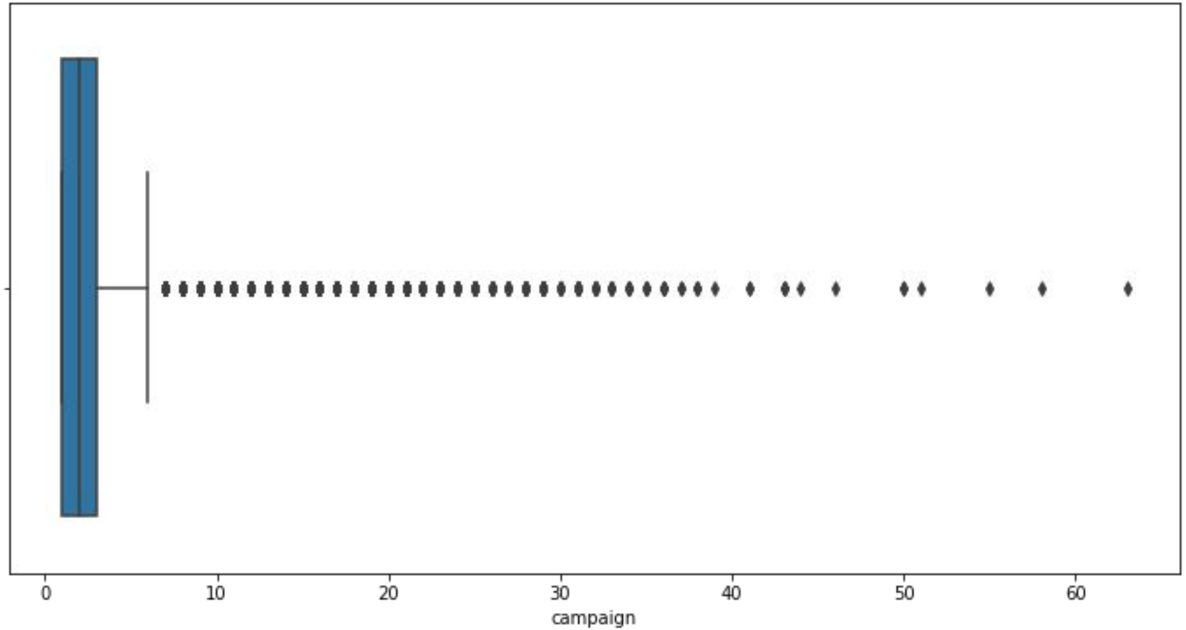
Preprocessing Age column



Age categories :

- Lower than equal 32
- Between 32 and 39
- Between 40 and 48
- Between 49 and 70
- Larger than 70

Preprocessing Campaign column

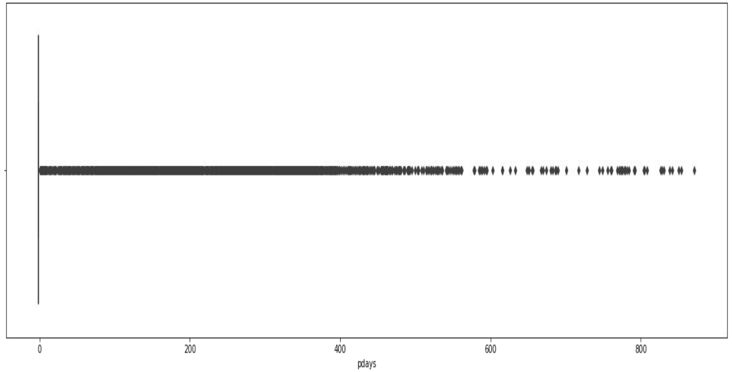


Campaign categories :

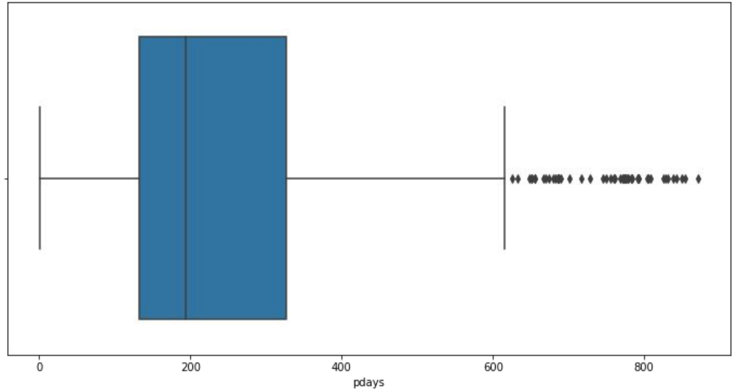
- Exactly 1
- Between 2 and 3
- Between 4 and 6
- Larger than 7

Preprocessing Pdays column

- Pdays



- Pdays without -1

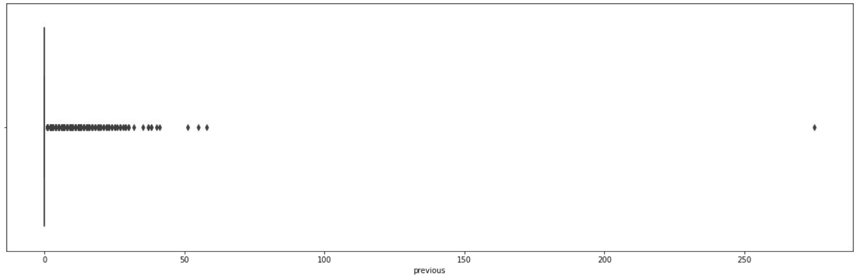


Pdays categories :

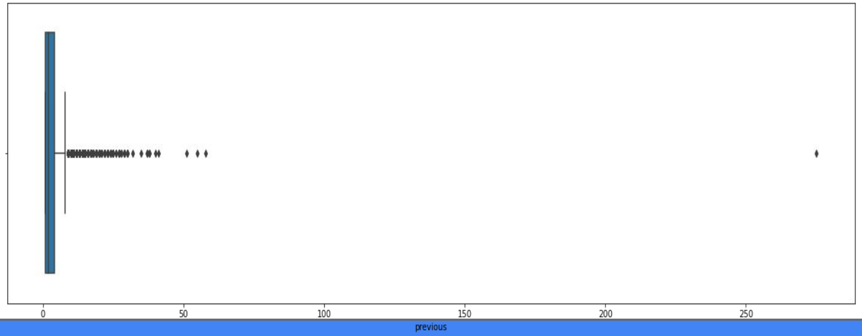
- Exactly -1
- Between 0 and 133
- Between 134 and 194
- Between 195 and 327
- Larger than 328

Preprocessing Previous column

- Previous



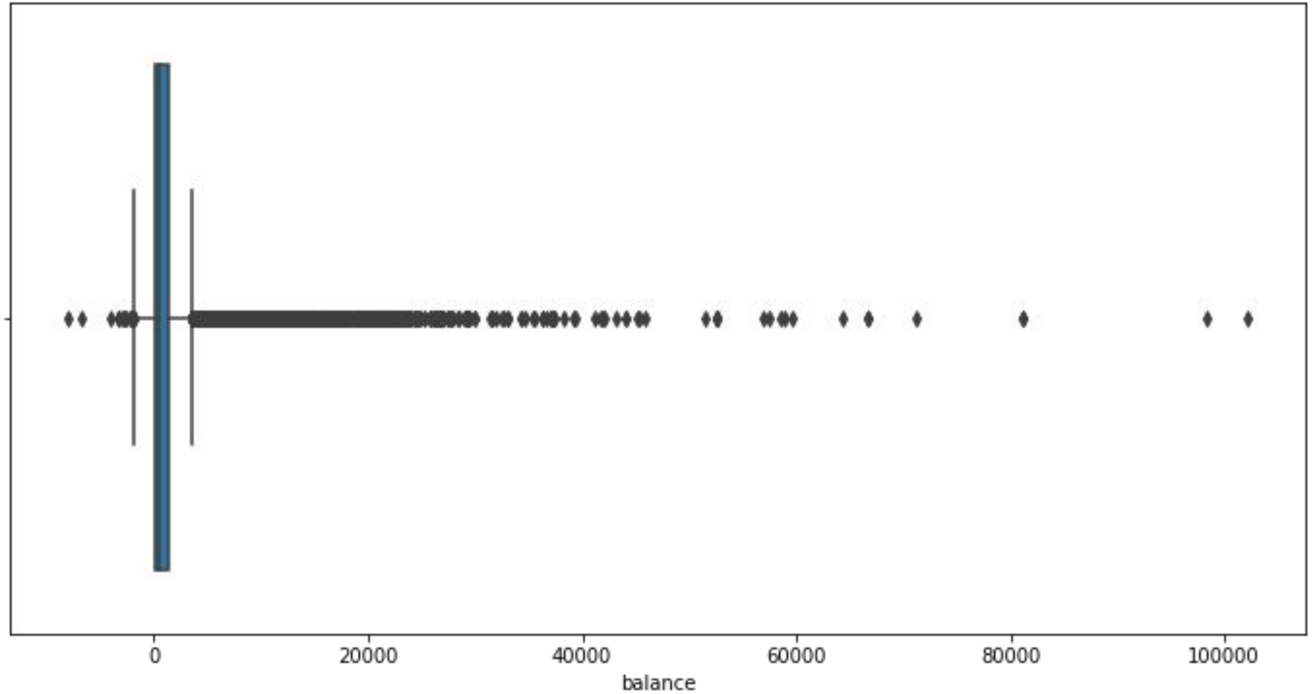
- Previous without 0



Previous categories :

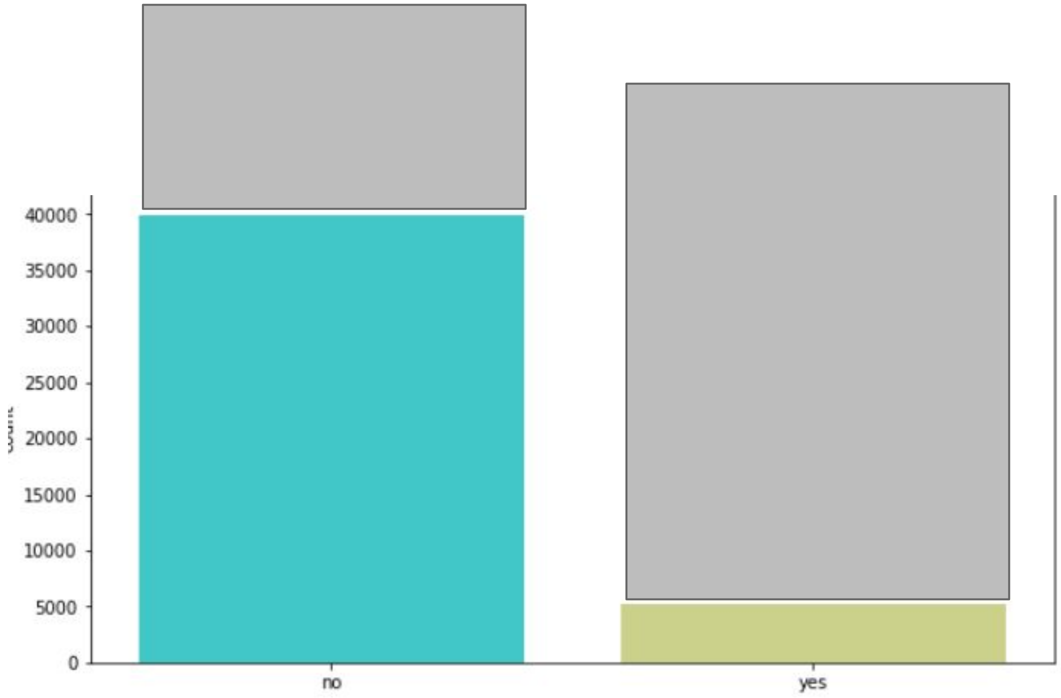
- Exactly 0
- Exactly 1
- Exactly 2
- Between 3 and 4
- Larger than 4

Preprocessing Balance column



$$z = \frac{x - \mu}{\sigma}$$

Oversampling Train Data



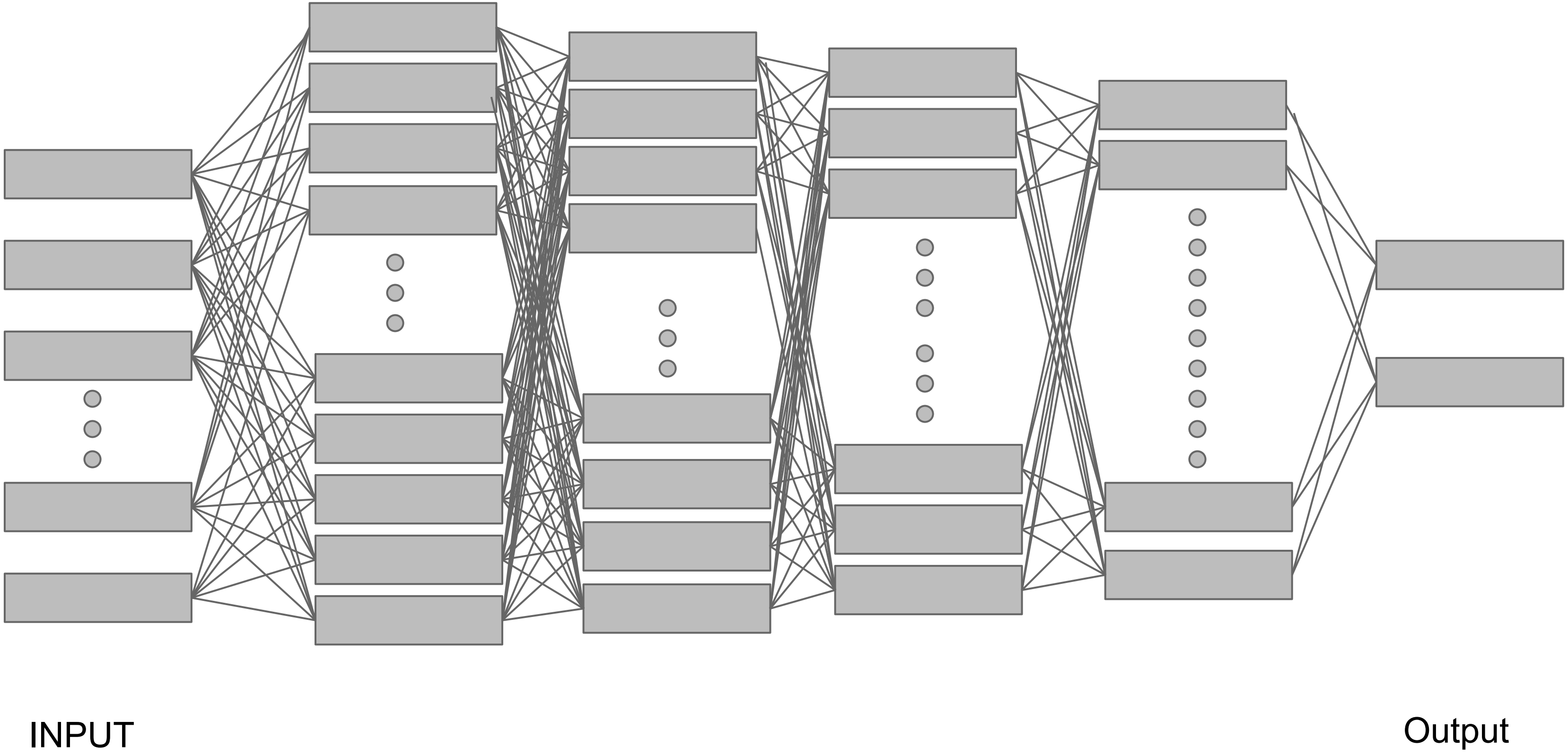
One Hot Encoding for Categorical Data

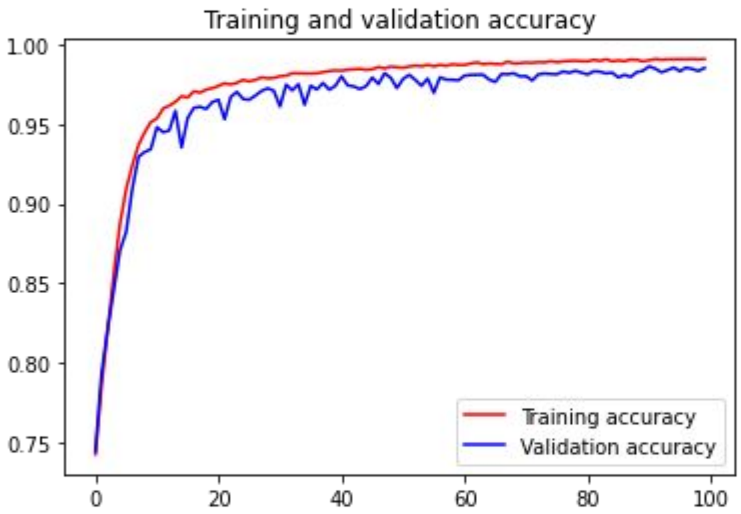
Cat
A
B
C
B
C
C

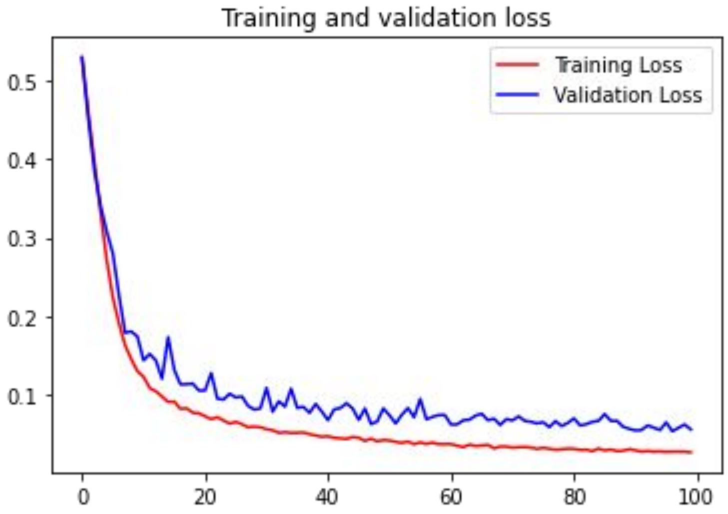
Cat_A	Cat_B	Cat_C
1	0	0
0	1	0
0	0	1
0	1	0
0	0	1
0	0	1

Cat_A	Cat_B
1	0
0	1
0	0
0	1
0	0
0	0

Modeling with ANN







Callback : ModelCheckPoint

Results

Confusion Matrix

ACTUAL \ PREDICT	no	yes
no	3952	48
yes	2	519

Classification Report

	Precision	Recall	F1-Score
no	1.00	0.99	0.99
yes	0.92	1	0.95

Accuracy = 0.99

BUSINESS IMPLICATIONS

05

Confusion Matrix Interpretation

Customers who predicted to subscribe and actually subscribe (**TP**): 519
Customers who predicted to NOT subscribe and actually NOT subscribe (**TN**): 3952
Customers who predicted to subscribe but actually NOT subscribe (**FN**): 48
Customers who predicted to NOT subscribe but actually subscribe (**FP**): 2

		PREDICT	
		no	**yes**
ACTUAL	**no**	3952	48
	yes	2	519

Number of People to Contact

From the 4,521 customers, we only have contact 567 (519+48) customers.

Supposed there are 100 total listed customers. Rather than contacting all those 100 customers, **we only have to contact 13 customers** (based on confusion matrix). This means our model will **decrease number of people to contact up to 87%.** This means we can save energy and resources which used to contact the customers.

From
100
listed customers

Only
13
are contacted

The ML model will help to decrease the number of people contact up to
87%

Reduce Marketing Cost

Supposed we spend 5 euros for each contact. Before applying ML we spend:
4,521 x 5 = 22,605 €

After applying ML we spend only:
567 x 5 euros = 2,835 €

We save 19,770 € or 87% less than before applying ML model

Spend

22,605

€ of marketing cost **before** applying ML

Spend only

2,835

€ of marketing cost **after** applying ML

The ML model will help to decrease the marketing cost up to

87%

Increase Profit

Supposed we spend 5 euros for each contact and every customers who decide to subscribe will lead to 50 euros revenue.

	Actual	After Applying ML
Spending	4,521 x 5 = 22,605 €	567 x 5 euros = 2,835 €
Revenue	521 x 50 = 26,050 €	519 x 50 = 25,950 €
Profit	26,050 - 22,605 = 3,445 €	25,950 - 2,835 = 23,115 €

Note that we also lose 2 customers by applying ML. Let's do a calculation.
= Profit after applying ML - Lost profit 23,115 - (2x50) = **23,015**
See… at the end we still earn more when applying ML, even earning almost **7 times** than before applying ML model

Comparison

With oversampling vs Without oversampling

Confusion Matrix

PREDICT

ACTUAL	no	yes
no	3952	48
yes	2	519

Classification Report

	Precision	Recall	F1-Score
no	1.00	0.99	0.99
yes	0.92	1	0.95

Accuracy = 0.99

Confusion Matrix

PREDICT

ACTUAL	no	yes
no	3946	54
yes	432	89

Classification Report

	Precision	Recall	F1-Score
no	0.90	0.99	0.94
yes	0.62	0.17	0.27

Accuracy = 0.89

Comparison (Vs ML based model: KNN)

Confusion Matrix

PREDICT

ACTUAL	no	yes
no	3952	48
yes	2	519

Classification Report

	Precision	Recall	F1-Score
no	1.00	0.99	0.99
yes	0.92	1	0.95

Accuracy = 0.99

Confusion Matrix

PREDICT

ACTUAL	no	yes
no	3162	838
yes	135	386

Classification Report

	Precision	Recall	F1-Score
no	0.96	0.79	0.87
yes	0.32	0.74	0.44

Accuracy = 0.78

Comparison (Vs ML based model: Logistic Regression)

Confusion Matrix

ACTUAL \ PREDICT	no	yes
no	3952	48
yes	2	519

Classification Report

	Precision	Recall	F1-Score
no	1.00	0.99	0.99
yes	0.92	1.00	0.95

Accuracy = 0.99

Confusion Matrix

ACTUAL \ PREDICT	no	yes
no	3056	944
yes	246	275

Classification Report

	Precision	Recall	F1-Score
no	0.96	0.79	0.87
yes	0.23	0.53	0.32

Accuracy = 0.74

Comparison (Vs ML based model: Decision Tree)

Confusion Matrix

ACTUAL \ PREDICT	no	yes
no	3952	48
yes	2	519

Classification Report

	Precision	Recall	F1-Score
no	1.00	0.99	0.99
yes	0.92	1.00	0.95

Accuracy = 0.99

Confusion Matrix

ACTUAL \ PREDICT	no	yes
no	3974	26
yes	0	521

Classification Report

	Precision	Recall	F1-Score
no	1.00	0.99	1.00
yes	0.95	1.00	0.98

Accuracy = 0.99

Comparison (Vs TensorFlow Decision Forest)

Confusion Matrix

ACTUAL \ PREDICT	no	yes
no	3952	48
yes	2	519

Classification Report

	Precision	Recall	F1-Score
no	1.00	0.99	0.99
yes	0.92	1.00	0.95

Accuracy = 0.99

Confusion Matrix

ACTUAL \ PREDICT	no	yes
no	3640	360
yes	14	507

Classification Report

	Precision	Recall	F1-Score
no	1.00	0.91	0.95
yes	0.58	0.97	0.73

Accuracy = 0.92

CONCLUSIONS

06

Conclusions

- NN Model managed to predict the customer who will subscribe with the accuracy of 99% (f1-score 95%, precision 92%, recall 100%)
- The application of this model is potential to reduce marketing costs by up to 87%

THANK YOU!