DSA Case Study Report

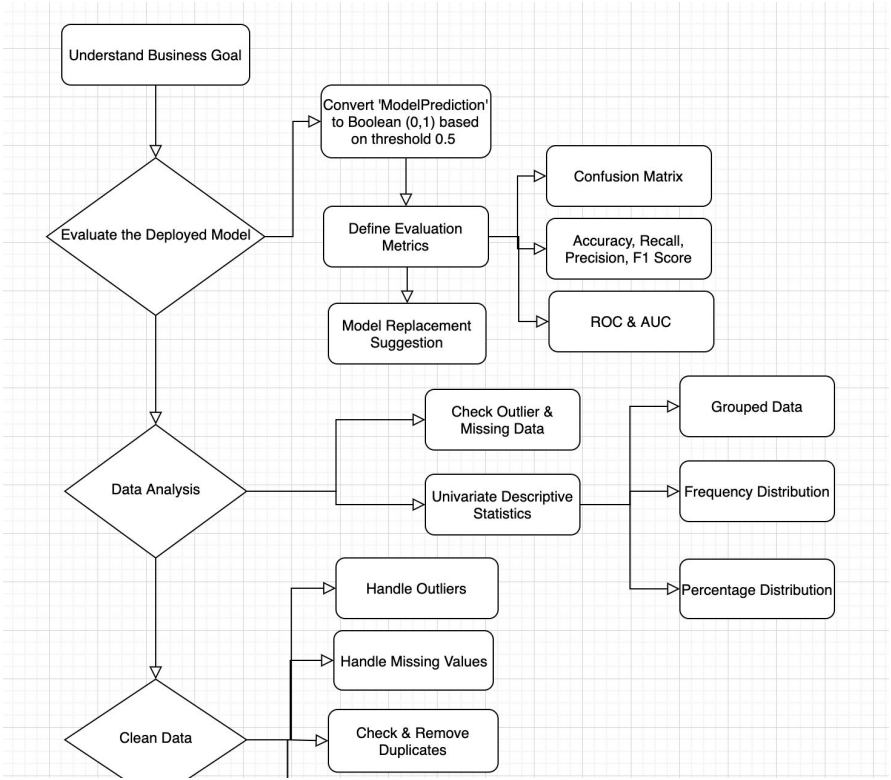
Created & Presented by Lin Sang
2021-08-17

Task Description

- Evaluate performance of the deployed model
- State the evaluation metrics used
- Suggest on whether replacing the deployed model with a new one
- Study and clean the data for further modeling

Experimental Design

- Link



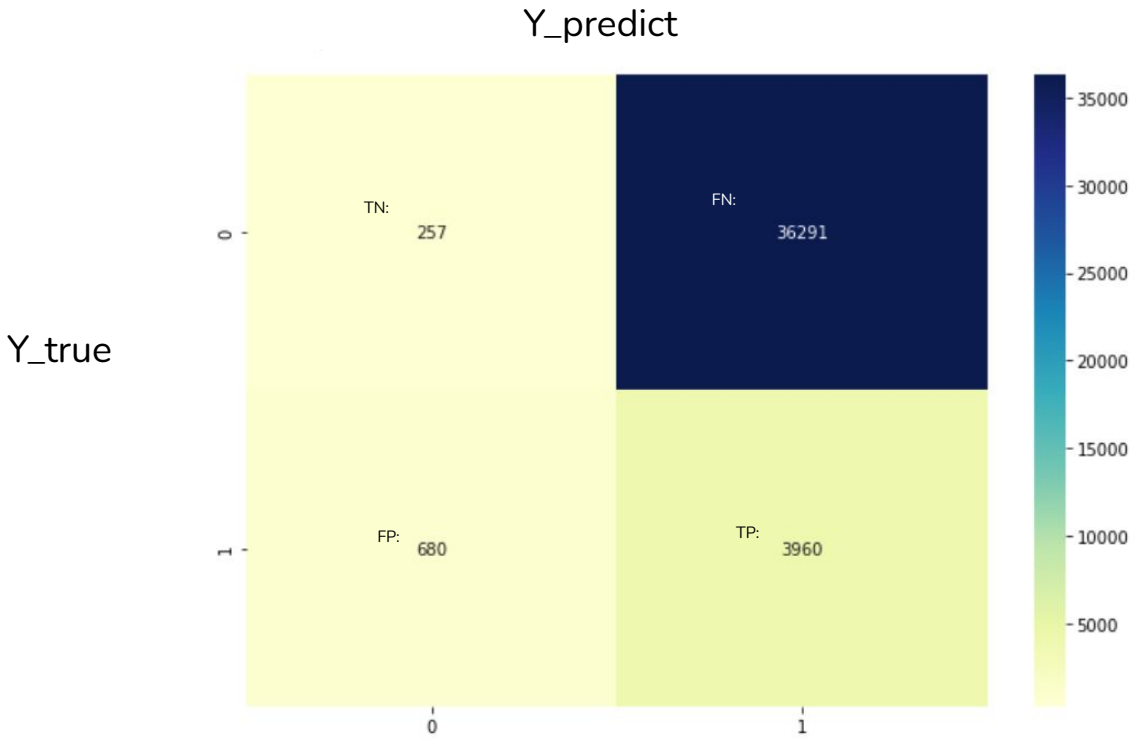
Our Business Goal

- Precisely identify real buyers & non buyers
- The importance of Type II Error
- Trade Offs

Evaluation Metrics for the Deployed Model

- Preprocessing
  - Convert 'ModelPrediction' to Boolean (0,1) based on threshold 0.5
- Evaluation Metrics
  - Confusion Matrix
  - Accuracy, Precision, Recall, F1 & F1 beta Score
  - AUC & ROC

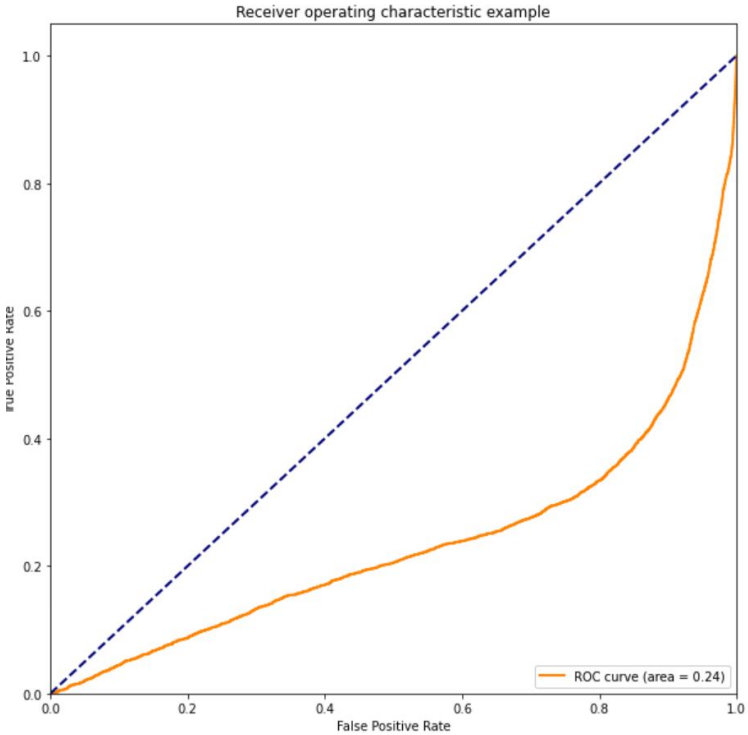
Confusion Matrix



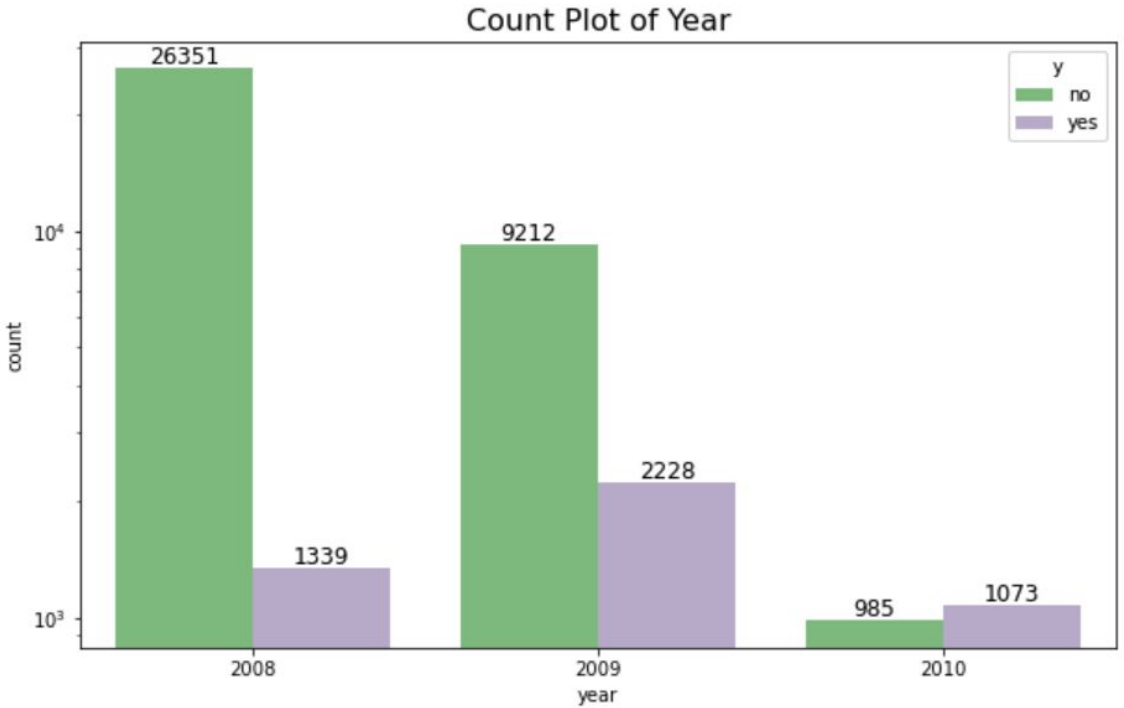
Recall & F-1 beta scores

```
Accuracy Score is:  10.24  %
Recall Score is:  85.34  %
Precision Score is:  9.84  %
F1 Score is:  17.64  %
F1 Score with Beta = 2 is: 33.67  %
Classification Report is:
              precision    recall  f1-score   support

           0       0.27      0.01      0.01     36548
           1       0.10      0.85      0.18      4640

    accuracy                           0.10     41188
   macro avg       0.19      0.43      0.10     41188
weighted avg       0.25      0.10      0.03     41188
```

AUC & ROC


Receiver operating characteristic example
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.24)

Evaluation Performance in Each Year



Evaluation Results in Each Year

In Year 2008:

Accuracy Score is: 4.84 %
Recall Score is: 100.0 %
Precision Score is: 4.84 %
F1 Score is: 9.229999999999999 %
F1 Score with Beta = 2 is: 20.26 %
Classification Report is:

	precision	recall	f1-score	support
0	0.00	0.00	0.00	26351
1	0.05	1.00	0.09	1339
accuracy			0.05	27690
macro avg	0.02	0.50	0.05	27690
weighted avg	0.00	0.05	0.00	27690

In Year 2009:

Accuracy Score is: 18.72 %
Recall Score is: 91.38 %
Precision Score is: 18.27 %
F1 Score is: 30.45 %
F1 Score with Beta = 2 is: 50.760000000000005 %
Classification Report is:

	precision	recall	f1-score	support
0	0.35	0.01	0.02	9212
1	0.18	0.91	0.30	2228
accuracy			0.19	11440
macro avg	0.27	0.46	0.16	11440
weighted avg	0.32	0.19	0.08	11440

In Year 2010:

Accuracy Score is: 35.809999999999995 %
Recall Score is: 54.52 %
Precision Score is: 41.260000000000005 %
F1 Score is: 46.97 %
F1 Score with Beta = 2 is: 51.23 %
Classification Report is:

	precision	recall	f1-score	support
0	0.24	0.15	0.19	985
1	0.41	0.55	0.47	1073
accuracy			0.36	2058
macro avg	0.33	0.35	0.33	2058
weighted avg	0.33	0.36	0.33	2058

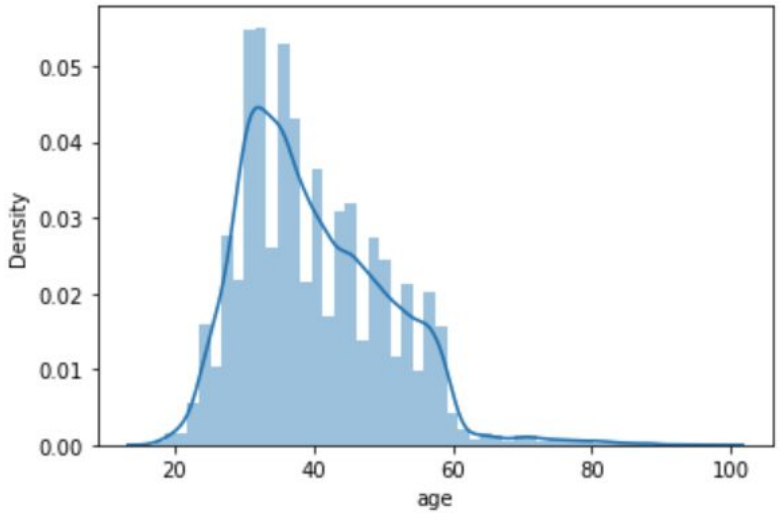
Data Analysis

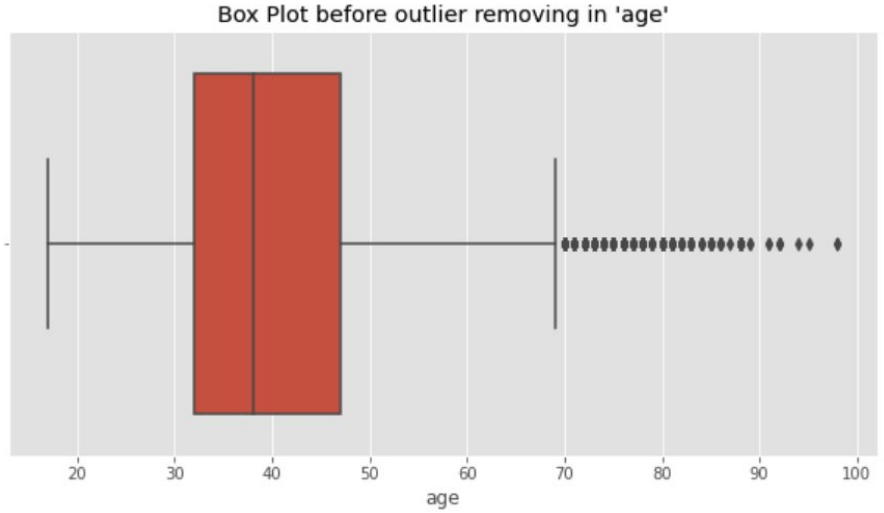
- Check uniqueness, missing values & outliers
- Univariate summaries

Clean the Data

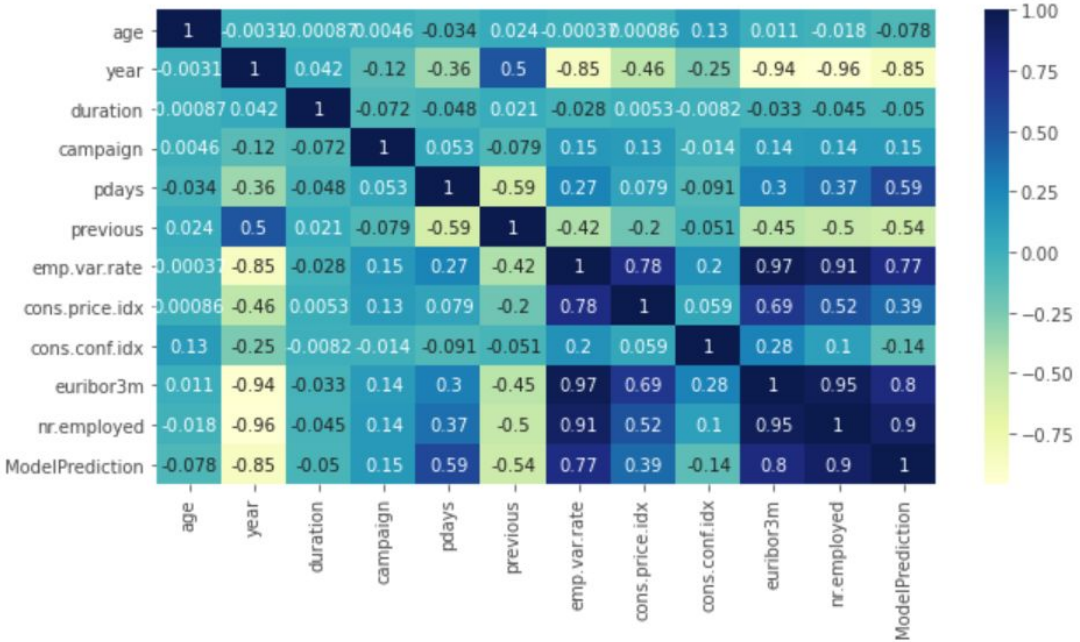
- Handle outliers
- Handle missing values
- Check & remove duplicates ✓
- Check & remove highly correlated inputs
- Implement Label Encoding

Check Outliers





Correlation between Inputs



Actions Taken

- Drop irrelevant inputs
  - ‘duration’
  - ‘ModelPrediction’
- Drop time-related inputs
  - ‘month’
  - ‘day_of_week’
- Drop the input with too many missing values
  - 'pdays'
- Drop columns with high redundancy cores
  - 'emp.var.rate'
  - 'euribor3m'
  - 'nr.employed'

Label Encoding

- Ordinal Encoding
  - ‘education’
  - ‘poutcome’
  - ‘y’
- One-Hot Encoding
  - for all categorical variable that do not have a natural rank

Preview the cleaned dataset

	age	job_housemaid	job_services	job_admin.	job_blue-collar	job_technician	job_retired	job_management	job_unemployed	job_self-employed	...	contac
0	56	1	0	0	0	0	0	0	0	0	...	
1	57	0	1	0	0	0	0	0	0	0	...	
2	37	0	1	0	0	0	0	0	0	0	...	
3	40	0	0	1	0	0	0	0	0	0	...	
4	56	0	1	0	0	0	0	0	0	0	...	
...	...	...	...	...	...	...	...	...	...	...	...	
41183	73	0	0	0	0	0	1	0	0	0	...	
41184	46	0	0	0	1	0	0	0	0	0	...	
41185	56	0	0	0	0	0	1	0	0	0	...	
41186	44	0	0	0	0	1	0	0	0	0	...	
41187	74	0	0	0	0	0	1	0	0	0	...	

41188 rows × 48 columns


Thank
you!