Machine Learning Capstone Project

BANK MARKETING CAMPAIGN

Presented by Mutiara Tambunan

Table of Contents

CONTEXT

- Term deposits will allow banks to invest to generate profits.
- More deposits, more investments and profits or Loan Deposit Ratio (LDR) is Ideal.

- Effective strategy to attract customers to subscribe deposit is through marketing campaigns.
- Phone-call-based campaign.

- Marketing campaign serves as a bridge between bank and customers.
- Promoting products and services, building trust and maintaining long-term relationships.

PROBLEM STATEMENT

Bank should be able to conduct a marketing campaign to the right customers who will subscribe a term deposit

Minimize the possibility of targeting customers who are not willing to make a deposit but also keep maintaining relationship with them

GOALS !!

Bank wants to have the ability to predict which customers are likely to deposit and conduct marketing campaign to those who genuinely want to subscribe a term deposit

Bank also wants to identify the most important features in subscribing deposits

Analytical Approcah & Metric Evaluation

TARGET

0 : <u>will not</u> subscribe a term deposit

1 : <u>will</u> subscribed a term deposit

<u>**Build a classification model to predict whether a customer will subscribe to a term deposit or not**</u>

CONFUSION MATRIX		Predicted	
		Negative (N) -	Positive (P) +
Actual	Negative -	True Negative (TN)	False Positive (FP) Type I Error
	Positive +	False Negative (FN) Type II Error	True Positive (TP)

- **FP**

Consequences : bank will waste time and money by conducting campaigns for the wrong customers.

- **FN**

Consequences : bank will create a negative impression, worsening the relationship and miss opportunity to promote products.

METRIC : F1 score

DATA CLEANING

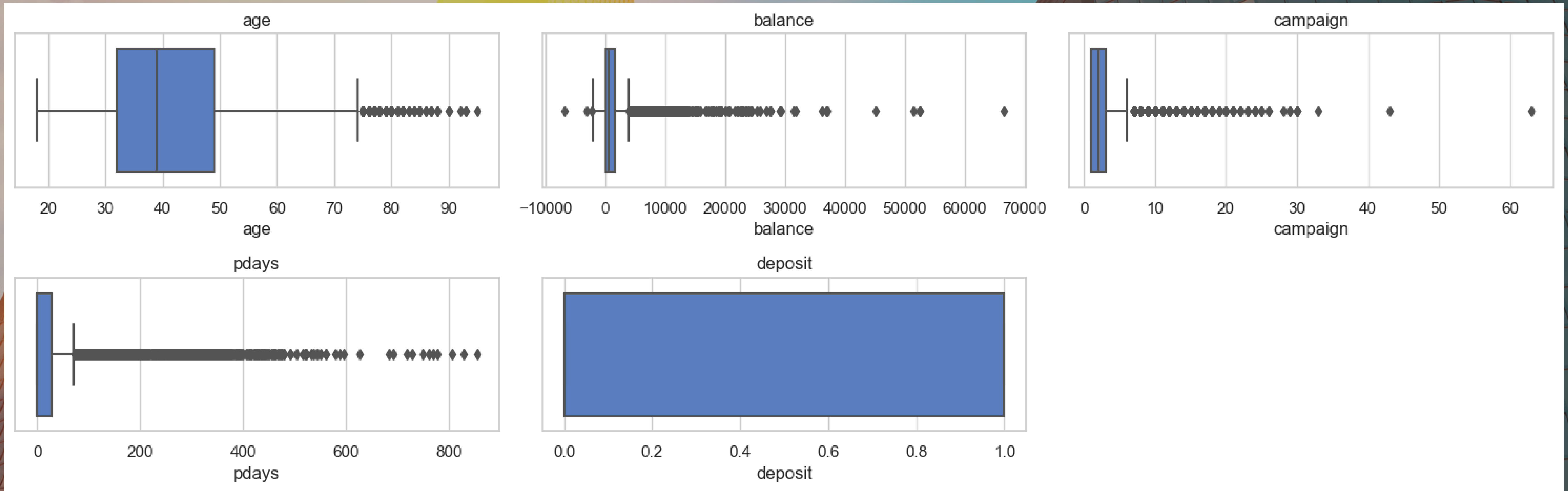
Features Information

Feature	Data Type	Description
Customer Profile		
age	numerical	Age of customer
job	categorical	Job of customer
balance	numerical	Customer's individual balance
housing	categorical	If costumer has housing loan
loan	categorical	If costumer Has Personal Loan
Marketing Data		
contact	categorical	Communication type
month	categorical	Last contact month of year
campaign	numerical	Number of contacts performed during this campaign and for this client
pdays	numerical	Number of days after the client was contacted from the previous campaign
poutcome	categorical	Outcome of the previous marketing campaign
deposit	categorical	Whether the customer deposits or not

	Column Name	Number of Unique	Unique Value
0	age	75	[55, 39, 51, 38, 36, 41, 37, 35, 57, 23, 33, 31, 53, 30, 46, 48, 25, 29, 28, 52, 49, 44, 42, 27, 47, 64, 26, 34, 56, 32, 58, 45, 54, 50, 79, 65, 40, 24, 60, 43, 61, 59, 62, 68, 82, 71, 73, 76, 69, 20, 72, 22, 67, 19, 70, 75, 63, 93, 77, 80, 66, 21, 87, 81, 92, 88, 84, 83, 78, 74, 18, 85, 95, 86, 90]
1	job	12	[admin., self-employed, services, housemaid, technician, management, student, blue-collar, entrepreneur, retired, unemployed, unknown]
2	balance	3153	[1662, -3058, 3025, -87, 205, -76, 4803, 911, 805, 0, 1234, 1107, 1170, 341, 4808, 88, 169, 863, 242, 2597, 4929, 277, 1438, 15, 3733, 204, 1684, 1025, 55, 19, 348, 785, 742, 511, 6651, 1612, 555, 54, 1185, 110, 950, 412, 228, 367, 3993, 2599, 3528, 32, 551, 3161, 533, 8725, 349, 514, 2688, -194, 154, 874, 2, 5953, 1269, -327, 235, 7, 2661, 1948, 20, 502, 193, 13658, 1716, 172, 1667, 157, 8, 951, 427, 241, 469, 2060, 7177, 655, -114, 588, -971, 4570, 250, 131, 93, 22, 15341, 356, 190, -124, 2228, -60, 376, 1567, 855, 4151, ...]
3	housing	2	[no, yes]
4	loan	2	[no, yes]
5	contact	3	[cellular, telephone, unknown]
6	month	12	[jun, apr, may, nov, jan, sep, feb, mar, aug, jul, oct, dec]
7	campaign	32	[2, 3, 1, 4, 5, 6, 7, 30, 8, 9, 11, 14, 10, 28, 63, 12, 24, 17, 15, 18, 19, 13, 21, 23, 22, 33, 16, 25, 26, 20, 29, 43]
8	pdays	422	[-1, 352, 21, 91, 186, 263, 96, 355, 294, 412, 89, 114, 276, 93, 175, 57, 323, 156, 86, 95, 271, 182, 289, 334, 269, 309, 144, 183, 417, 138, 254, 337, 171, 389, 87, 170, 165, 372, 247, 98, 196, 469, 272, 104, 63, 587, 336, 145, 130, 28, 202, 324, 147, 94, 328, 420, 179, 90, 81, 160, 298, 356, 357, 267, 430, 52, 181, 365, 237, 330, 103, 374, 75, 133, 321, 204, 782, 266, 197, 270, 318, 349, 187, 359, 490, 192, 227, 100, 168, 177, 251, 301, 350, 92, 184, 345, 290, 199, 333, 169, ...]
9	poutcome	4	[unknown, other, failure, success]
10	deposit	2	[1, 0]

- Dataset contains information of Bank Deposit Marketing Campaign. Phone calls based campaign.
- There is NO missing value and ordinal categorical.
- pdays starts with -1, indicating that the customer was not previously contacted.
- 'other' and 'unknown' in poutcome have same meaning.

DATA CLEANING



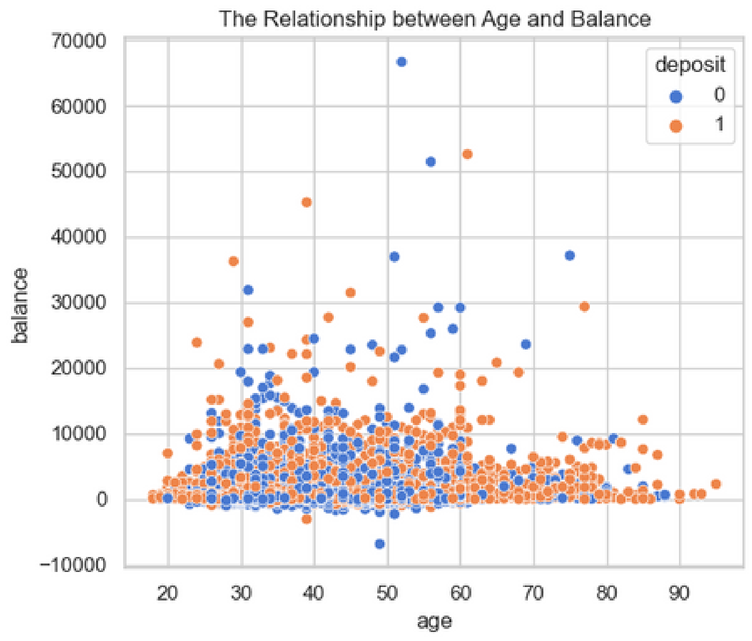
- Each feature outside deposit has numerous outliers, indicating that they are not normally distributed.

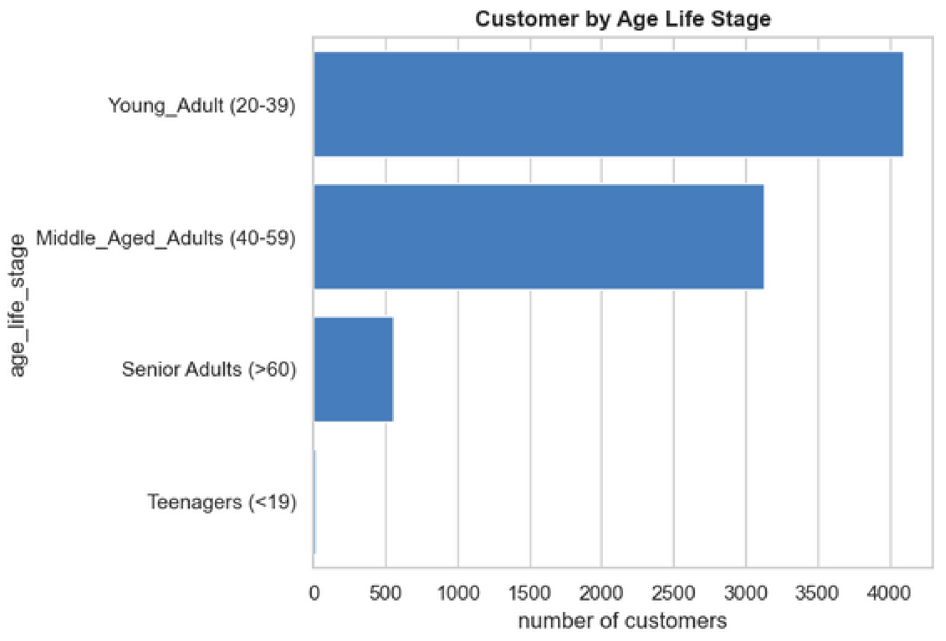
- Following the domain knowledge. Data has successfully cleaned to make it ready for analysis and modeling.

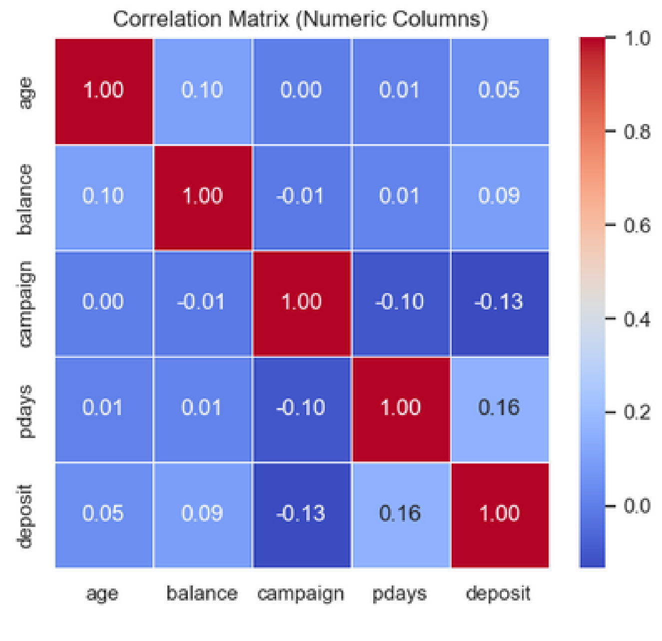
EXPLORATORY DATA ANALYSIS

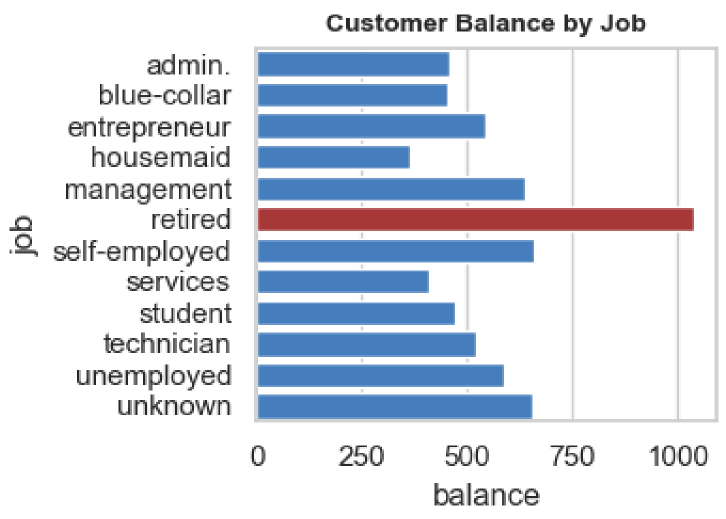
- Majority jobs in management with more deposits.
- Outcome of the previous marketing campaign has had very few successes.
- The dominant contact month for customers is in May, with mostly no deposits, and the preferred contact method is cellular.

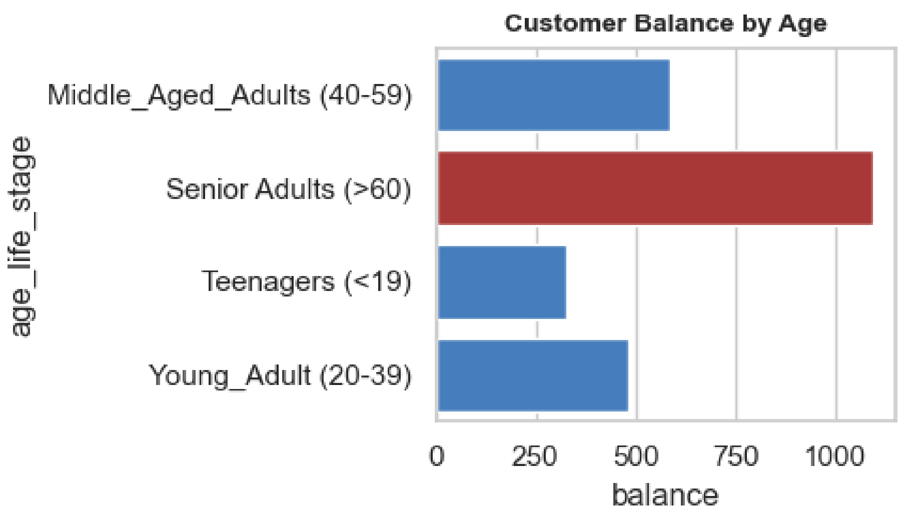
EXPLORATORY DATA ANALYSIS











- NO strong correlations among features.
- NO positive relationship between age and balance.
- Retired job have the highest median balance, and aged over 60 years old.

DATA UNDERSTANDING

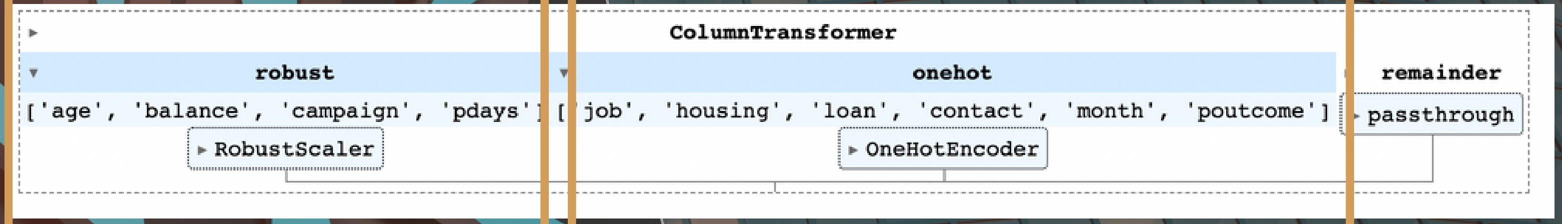
MODELING

CHECKING TARGET DEPOSIT VALUES (TARGET IS BALANCED)

DEFINE X AND Y & SPLIT DATA

- 75% for Train
- 25% for Test

PREPROCESSING (TRANSFORM DATA)



MODEL BENCHMARKING (TRAIN AND TEST)

CHOOSE THE BEST MODEL

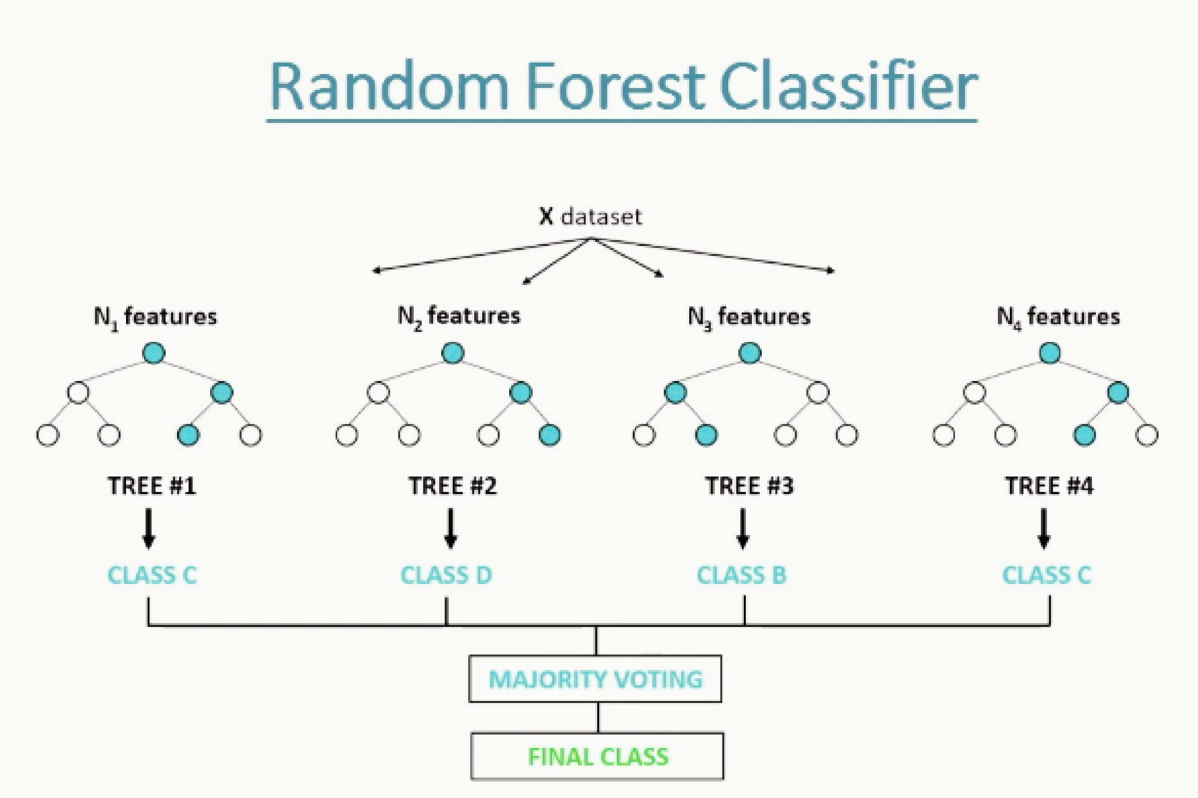
HYPERPARAMETER TUNING

UP TO SECOND TUNING

CLASSIFICATION MODEL

	algorithm model	accuracy	precision	recall	f1	f2
8	Random Forest	0.689667	0.701623	0.620430	0.658428	0.635070
9	XGBoost	0.688616	0.702496	0.615054	0.655862	0.630750
7	Stacking	0.695063	0.724274	0.597849	0.654906	0.619416
2	Logistic Regression Penalized	0.686244	0.701482	0.609319	0.651756	0.625531
6	Voting Classifier - Hard	0.682311	0.706883	0.586738	0.641172	0.607353
1	Logistic Regression	0.690099	0.732353	0.570251	0.640941	0.596522
5	Voting Classifier - Soft	0.652620	0.644705	0.614695	0.629236	0.620406
0	KNN	0.661409	0.665889	0.594982	0.628345	0.607873
3	Decision Tree	0.619155	0.603211	0.597491	0.600176	0.598524
4	Decision Tree Penalized	0.618620	0.604561	0.588889	0.596499	0.591882

BEST MODEL !!!

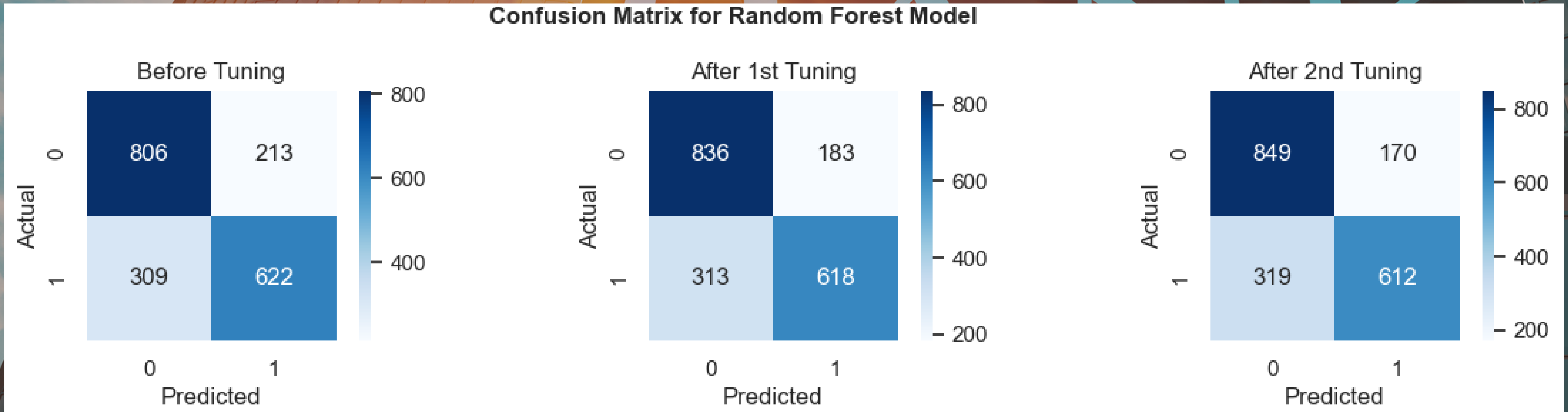


- Random Forest Classifier is an ensemble ML algorithm.
- Versatile and powerful technique that combines multiple decision trees to make predictions.
- Bootstrap Sampling >> Random Feature Selection >> Decision Tree Building >> Voting

CLASSIFICATION MODEL

	Accuracy	Precision	Recall	F1	F2
Random Forest before Tuning	0.732308	0.744910	0.668099	0.704417	0.682167
Random Forest after Tuning	0.745641	0.771536	0.663802	0.713626	0.682873
Random Forest after 2nd Tuning	0.749231	0.782609	0.657358	0.714536	0.679095

- After applying the second round of hyperparameter tuning, F1 score improved very slightly.



- Number of FP has significantly decreased from 213 to 170 customers.
- BUT FN increased from 309 to 319.
- FN can be profitable for the bank, but also can worsening the relationship.

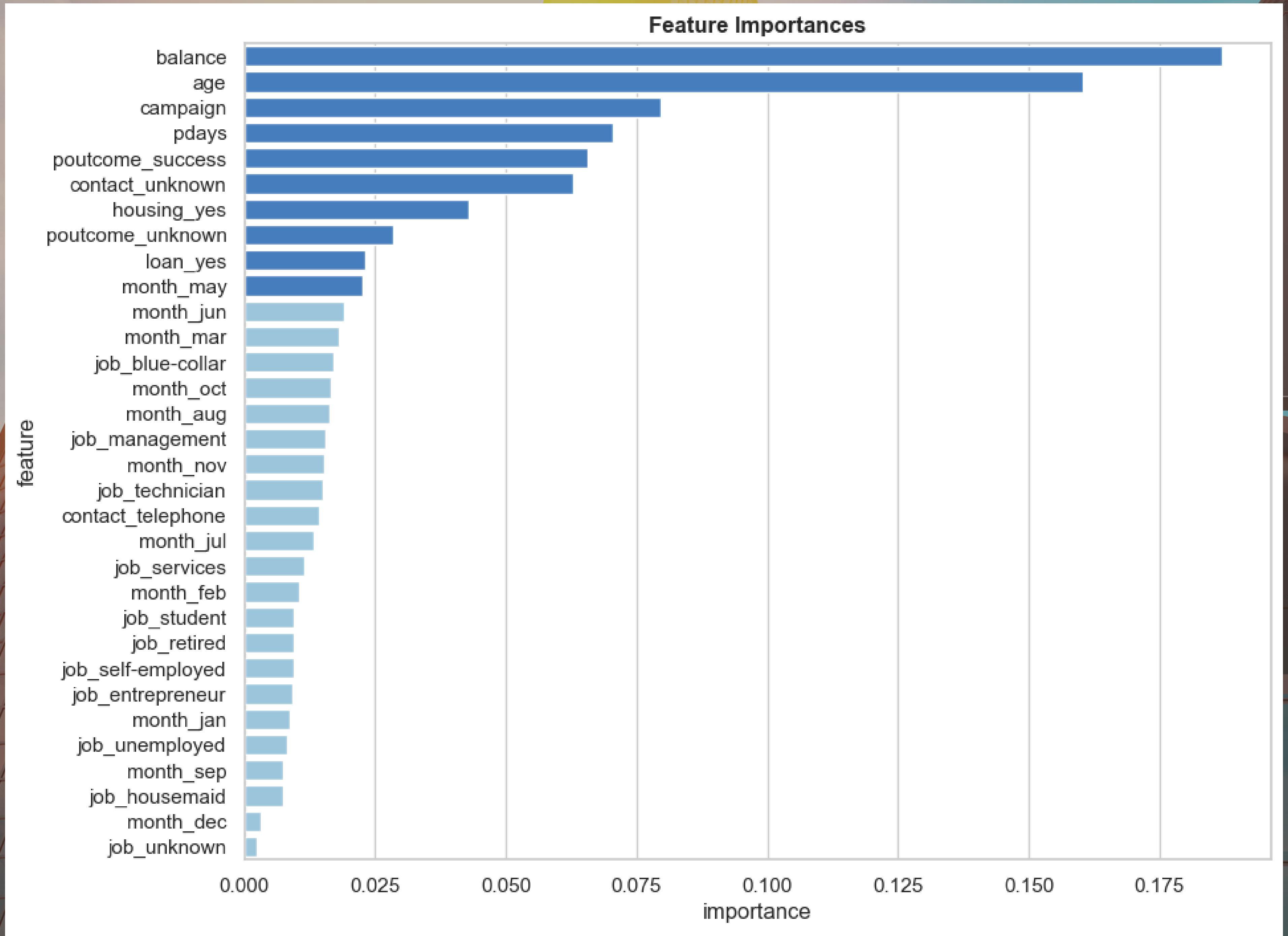
CLASSIFICATION MODEL

Classification Report Random Forest After 2nd Tuning :

	precision	recall	f1-score	support
0	0.73	0.83	0.78	1019
1	0.78	0.66	0.71	931
...				
accuracy			0.75	1950
macro avg	0.75	0.75	0.75	1950
weighted avg	0.75	0.75	0.75	1950

- **From Precision, out of all customers that predicted to subscribe deposit, 22% of them were incorrect predictions.**
- **From Recall, model managed to identify correctly 83% of customers who will not subscribe to a term deposit.**
- **An F1 score of 0.7 or higher is often considered good. Model have a good balance between precision and recall.**

CLASSIFICATION MODEL



- The most significant influence on the prediction result is '<u>balance</u>' followed by '<u>age</u>' and '<u>campaign</u>'.

CONCLUSION

- The best model for predicting which customers will subscribe to a term deposit is Random Forest.
- What is the impact of applying this model for marketing campaigns?

Without Model : all customer will be contacted

- Total time consumed : 1.950 x 4 = 7.800 minutes ~ 2.2 hours
- Total cost : 1.950 x 0.5 = $975

With Model : ONLY customer who predicted will be contacted

- Total time consumed : 782 x 4 = 3.128 minutes ~ 0,9 hour
- Total cost : 782 x 0.5 = $391

- Time Saved : 7.800 - 3.128 = 4.672 minutes ~ 1,3 hours
- Cost Saved : 975 - 391 = $584

RECOMMENDATION

- Explore alternative campaign methods such as utilizing WhatsApp or email blasts.
- Consider offering higher interest rates for longer-term and higher deposits.
- Introducing limited-time offers or promotions.
- Ensuring online and offline application process for opening a term deposit account are user-friendly and efficient.

- Adding a primary key column to ensure there are no duplicate data.
- Adding new features that could be related to the bank's marketing campaign such as: ‘duration’, ‘previous’ etc

- Continually updating with new machine learning model
- Try another machine learning algorithm such as clustering method.


Machine Learning Capstone Project
THANK YOU!
Mutiara Tambunan
Batam, September 2023