TELCO CUSTOMER CHURN

Data Analysis - Machine Learning

BY

ALFIAN (JCDSVL-005-013)

TABLE OF CONTENT

TELCO CUSTOMER CHURN

1. PURPOSE & INTRODUCTION

2. BUSINESS PROBLEM UNDERSTANDING

3. LIBRARY & DEFINED FUNCTION
- 3.1. PYTHON PACKAGE LIBRARY
- 3.2. DEFINED FUNCTION

4. DATA UNDERSTANDING
- 4.1. FEATURE CHECK
- 4.2. FEATURE CORRELATION
  - 4.2.1. NUMERICAL FEATURE WITH TARGET
  - 4.2.2. CATEGORICAL FEATURE WITH TARGET
  - 4.2.3. MULTIPLE FEATURE WITH TARGET

5. MACHINE LEARNING
- 5.1. DATA PREPARATION
- 5.2. BASE MODEL EVALUATION
- 5.3. BASE MODEL WITH OVERSAMPLING
- 5.4. BASE MODEL WITH UNDERSAMPLING
- 5.5. BASE MODEL OUTPUT RECAP

CONT.

5. MACHINE LEARNING

CONT.
- 5.6. HYPERPARAMETER TUNING
  - 5.6.1. HYPERPARAMETER TUNING - LOGISTIC REGRESSION
  - 5.6.2. HYPERPARAMETER TUNING - GRADIENT BOOST
  - 5.6.3. HYPERPARAMETER TUNING OUTPUT RECAP
- 5.7. FEATURE IMPORTANCE
  - 5.7.1. BEST MODEL WITH CURRENT FEATURES
  - 5.7.2. BEST MODEL WITH SELECTED FEATURES
  - 5.7.3. RECAP FOR FEATURE IMPORTANCE & FEATURE SELECTION

6. CONCLUSION

7. RECOMMENDATION

8. DEPLOYMENT

This is the table of content of the whole project. In this presentation I try to explain my project briefly. For more detail information please check

Repositories	: click here
Notebook	: click here

BUSINESS PROBLEM UNDERSTANDING

Repositories : click here
Notebook : click here

TELCO CUSTOMER CHURN

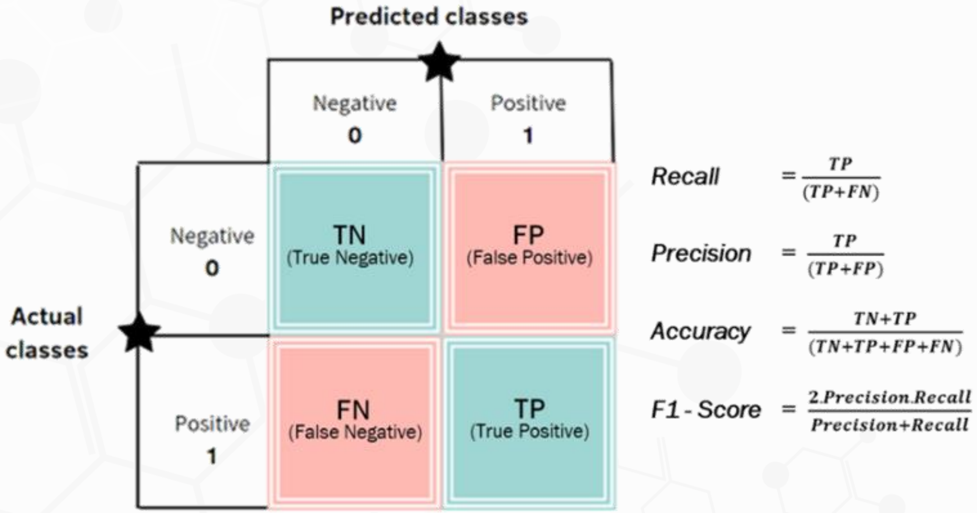
CHURN

Customer who left within the last month

OBJECTIVES

TELCO want to prevent & reduce customer churn, But first we need to understand why customer churn in the first place. From there TELCO can improve their services or create mitigation plan to prevent & reduce customer churn.

METRICS CHOOSEN



		Predicted classes	
		Negative 0	Positive 1
Actual classes	Negative 0	TN (True Negative)	FP (False Positive)
	Positive 1	FN (False Negative)	TP (True Positive)

$$Recall = \frac{TP}{(TP+FN)}$$

$$Precision = \frac{TP}{(TP+FP)}$$

$$Accuracy = \frac{TN+TP}{(TN+TP+FP+FN)}$$

$$F1\text{-}Score = \frac{2.Precision.Recall}{Precision+Recall}$$

After understanding the consequence of FP and FN. Metrics that I'll be use in this study is **<u>recall score .</u>**

Please check repositories or notebook for more in depth explanation

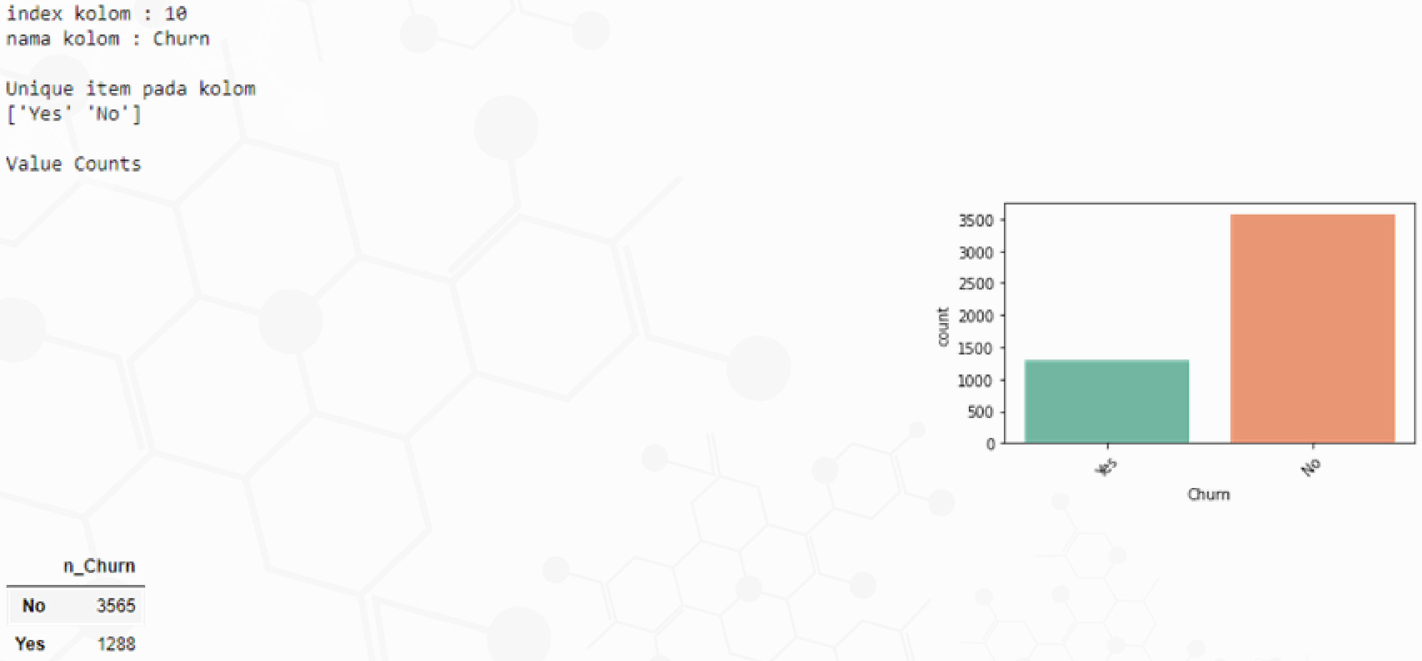
DATA ANALYSIS

TELCO CUSTOMER CHURN

Repositories : click here
Notebook : click here

Several things that I do in this section are as follows :

- Data Cleaning (our data is great, the cleaning that I should do only to remove duplicate data)
- Exploring each features
- Determine target label and adjust it for machine learning.
- Understanding correlation between features & target.
- Link some features with target
- I'll explain several graph that affect machine learning, more graph on notebook more or less just an exploration on the data.
- Our data is imbalance, I try to handle the imbalance data with resampling method & adjust the threshold



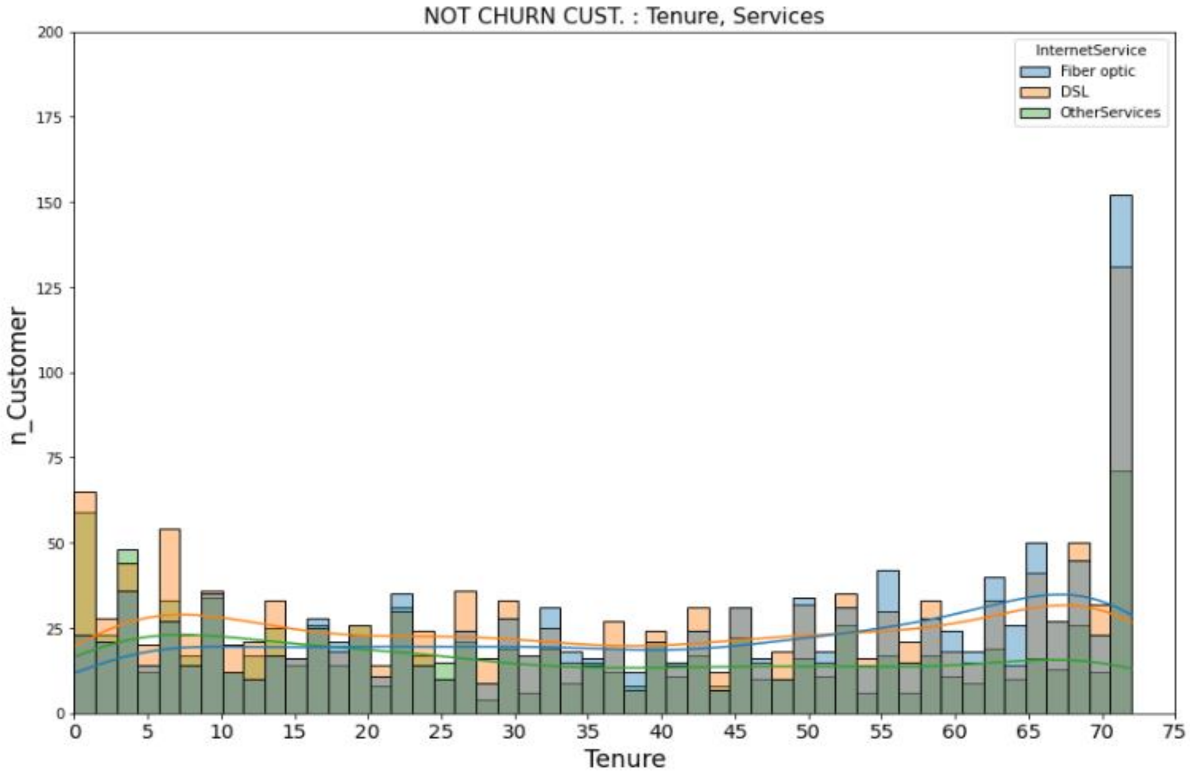
```
index kolom : 10
nama kolom : Churn

Unique item pada kolom
['Yes' 'No']

Value Counts
```

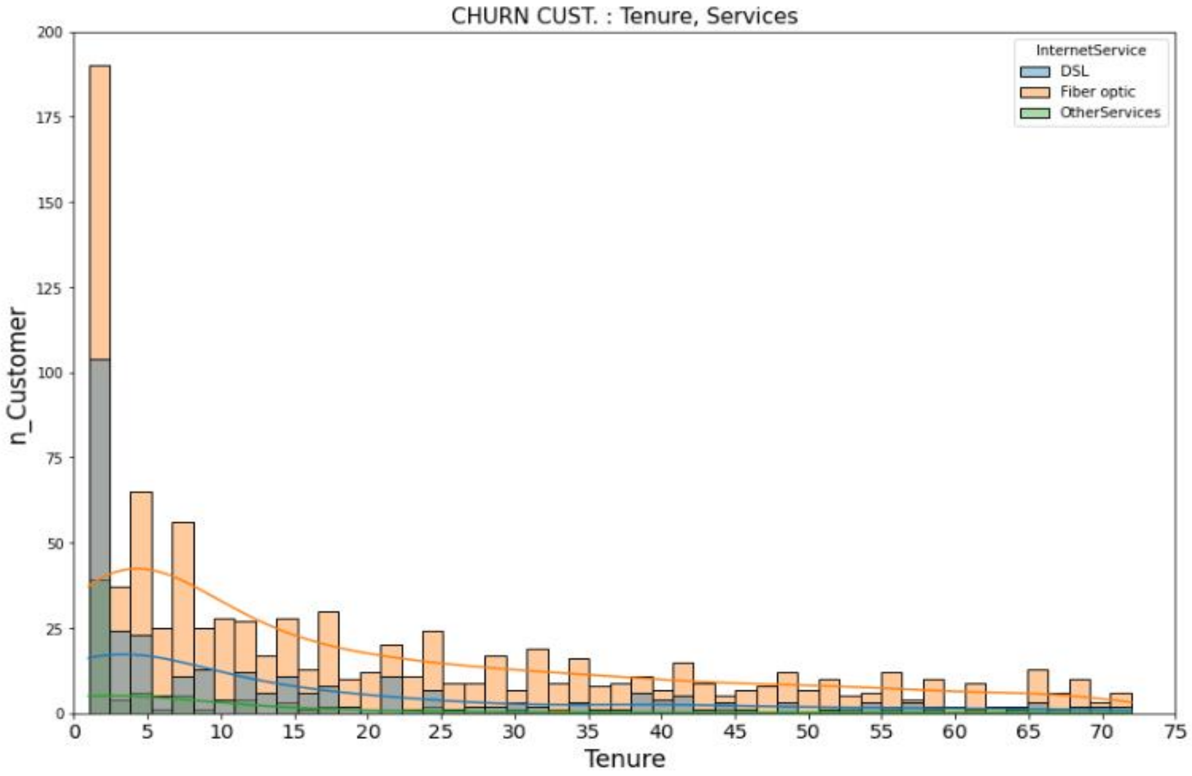
	n_Churn
No	3565
Yes	1288

DATA ANALYSIS

TELCO CUSTOMER CHURN

Repositories : click here
Notebook : click here


NOT CHURN CUST. : Tenure, Services


CHURN CUST. : Tenure, Services

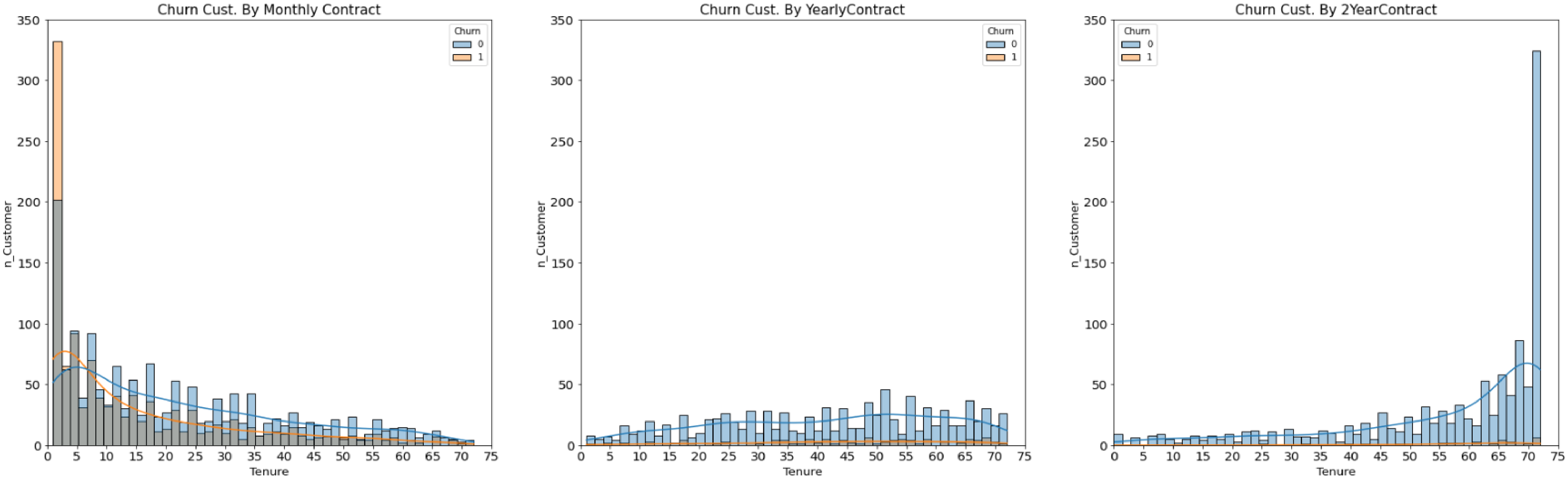
What can we see on the graph :

- Much fiber optic customer will churn compared to other service categories
- customer will churn in the first several month (0 - 5) especially fiber optic services (churn rate for fiber optic category is consistent along tenure)
- customer who not churn is well diverse.
- We can see that TELCO have loyal customer that have subscribe for >70 months

DATA ANALYSIS

TELCO CUSTOMER CHURN

Repositories : click here
Notebook : click here



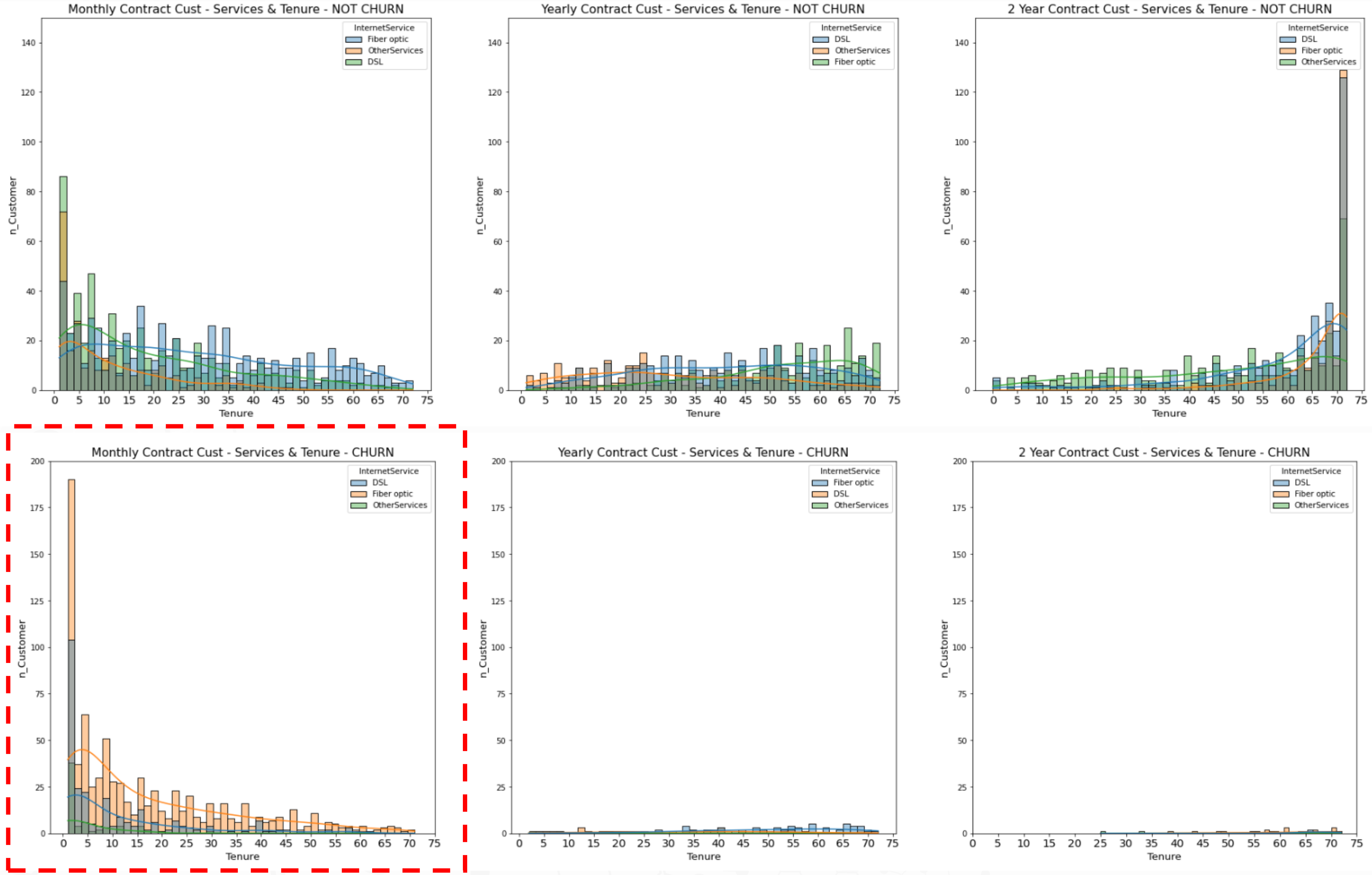
What can we see on the graph :

- Many customer that will churn is from Month to Month category (especially first 0 - 25 month).
- But we can see that the longer the customer subscribe, the less customer will churn. I'll assume that customer transitioning from Month to month to Yearlycontract to 2Year contract.
- Small percentage of customer who use yearly contract will churn, I assume that this small percentage customer doesn't need TELCO services anymore.

DATA ANALYSIS

TELCO CUSTOMER CHURN

Repositories : click here
Notebook : click here



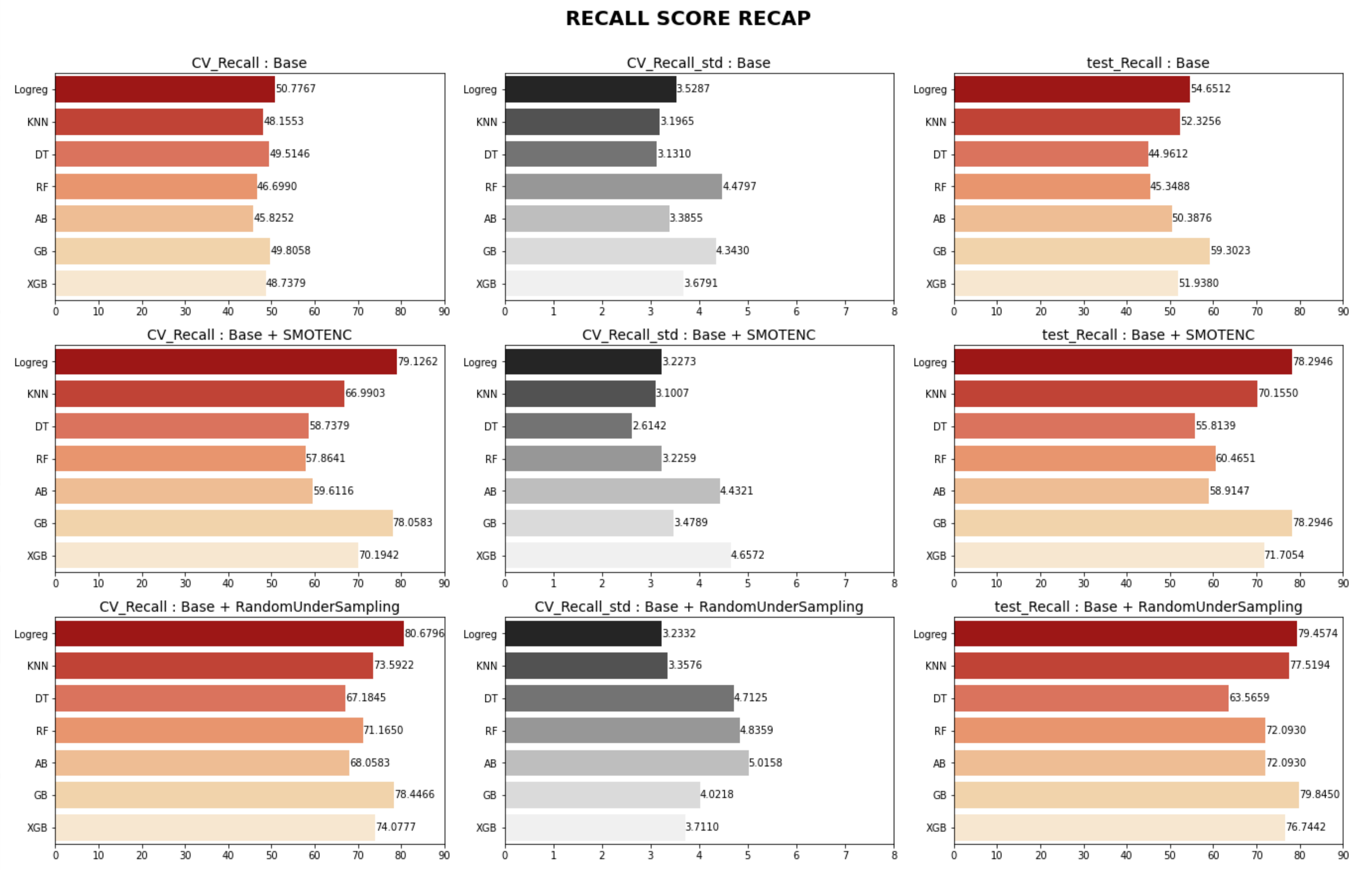
What can we see on the graph :

- We now can see clearly that customer who use fiber optic, month to month contract & who has tenure within range 0 – 5 will churn

MACHINE LEARNING – MODEL SELECTION

Repositories : click here
Notebook : click here

TELCO CUSTOMER CHURN



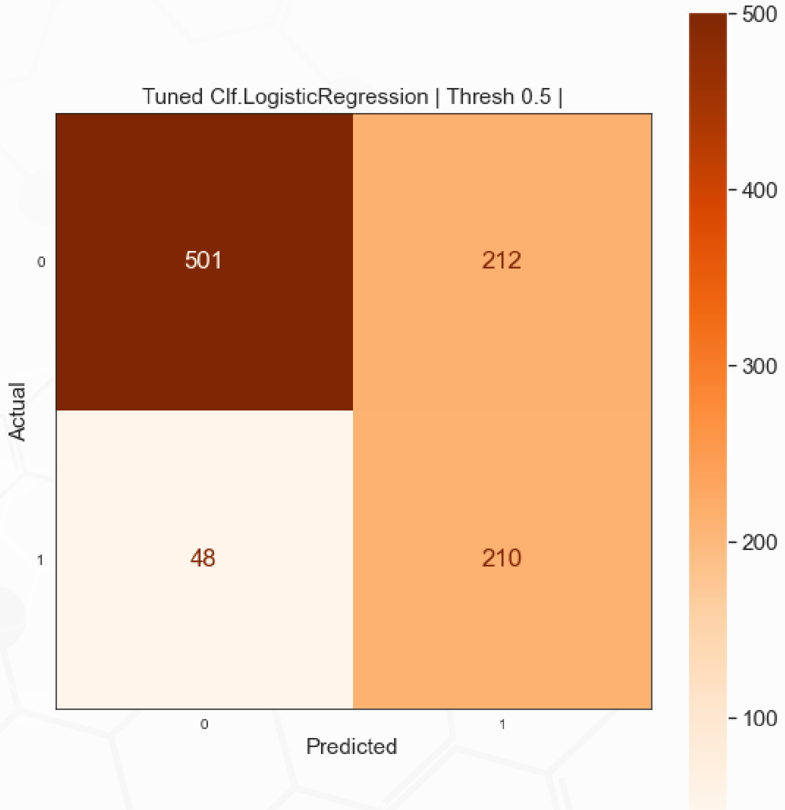
What can we see on the graph :

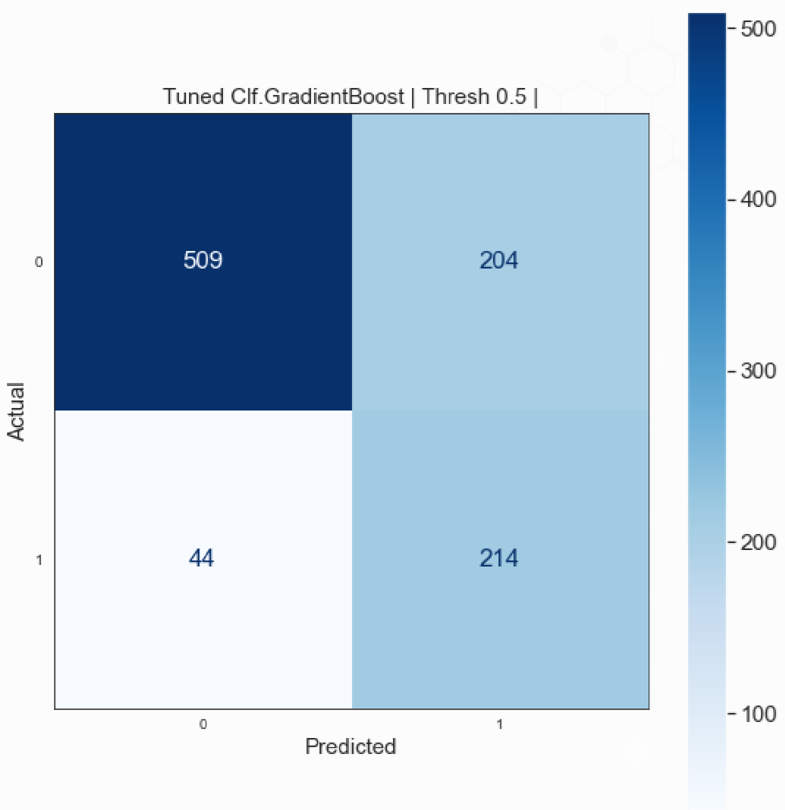
- Because there's no perfect algorithm that can be used for all type of case, I need to find that one algorithm that suitable for TELCO customer churn.
- This graph consist of CV score & test score for 7 algorithm + resampling method
- I find 2 candidates : LogisticRegression & GradientBoost that I try to tuned further to see which will yield the best result
- I use SMOTENC to resample our imbalance data

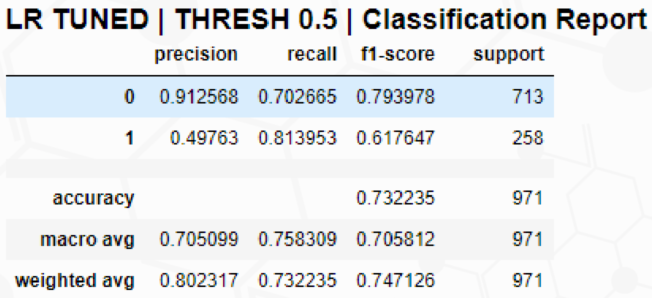
MACHINE LEARNING – TUNING

TELCO CUSTOMER CHURN

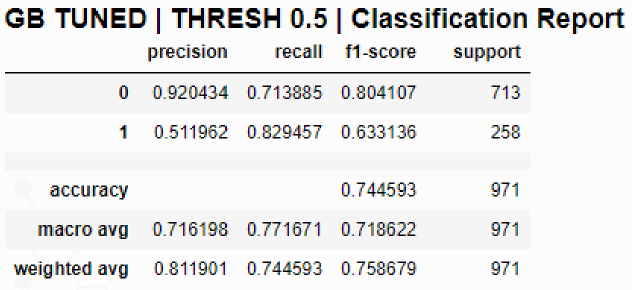
Repositories : click here
Notebook : click here


Tuned Clf.LogisticRegression | Thresh 0.5 |


Tuned Clf.GradientBoost | Thresh 0.5 |


LR TUNED | THRESH 0.5 | Classification Report

	precision	recall	f1-score	support
0	0.912568	0.702665	0.793978	713
1	0.49763	0.813953	0.617647	258
accuracy			0.732235	971
macro avg	0.705099	0.758309	0.705812	971
weighted avg	0.802317	0.732235	0.747126	971


GB TUNED | THRESH 0.5 | Classification Report

	precision	recall	f1-score	support
0	0.920434	0.713885	0.804107	713
1	0.511962	0.829457	0.633136	258
accuracy			0.744593	971
macro avg	0.716198	0.771671	0.718622	971
weighted avg	0.811901	0.744593	0.758679	971

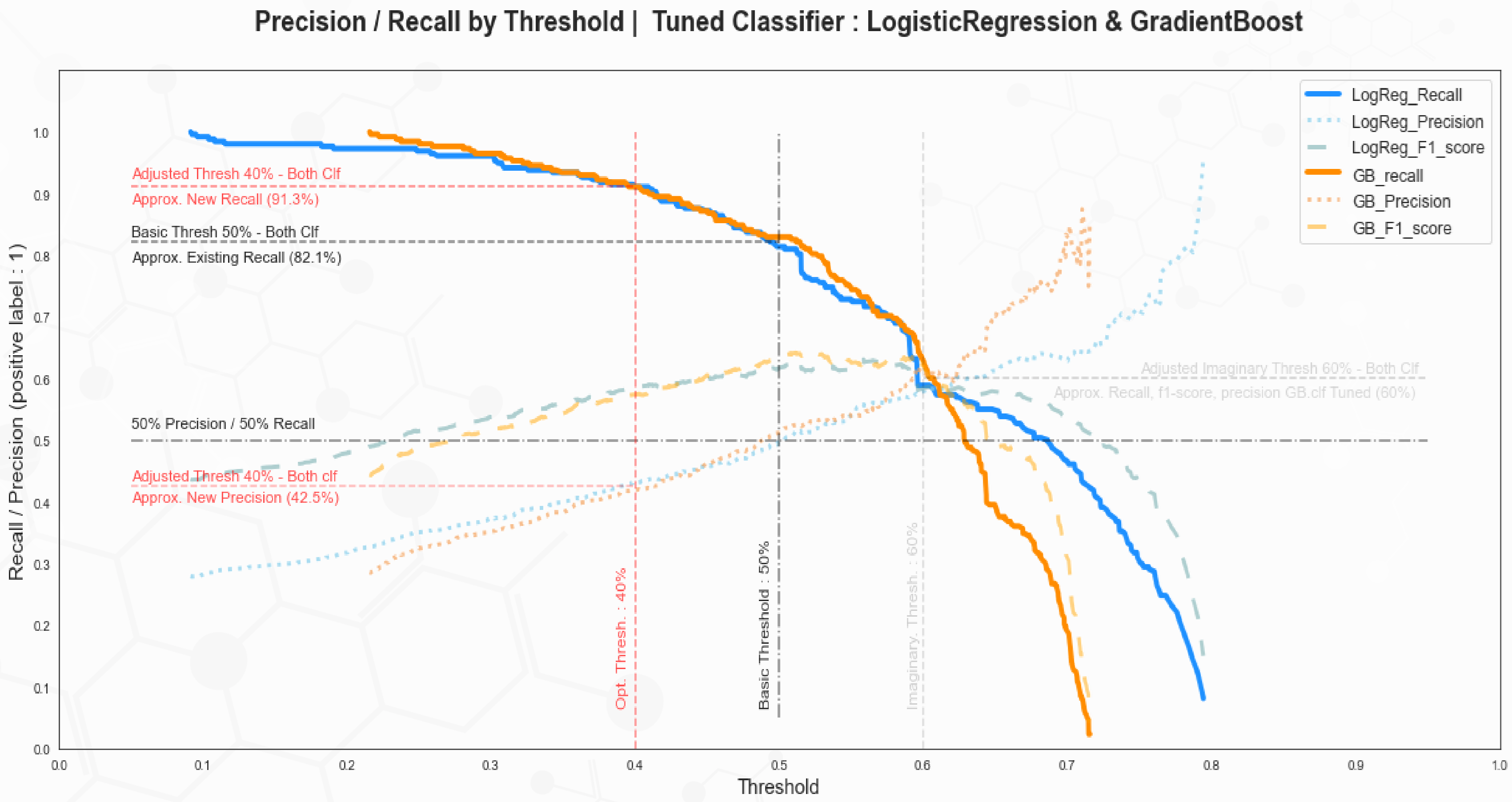
What can we see on the graph :

- Remember that metrics that I use in TELCO customer churn case is recall score.
- From confusion matrix & classification report, with 50% threshold, GradientBoostClassifier will yield the best result with recall score of 82.94%

MACHINE LEARNING – ADJUST THRESHOLD

Repositories : click here
Notebook : click here

TELCO CUSTOMER CHURN



NOTES

- Please understand that there's precision recall tradeoff. If we try to increase recall, the precision will decrease. Vice versa. For more detail explanation please check my repo or notebook

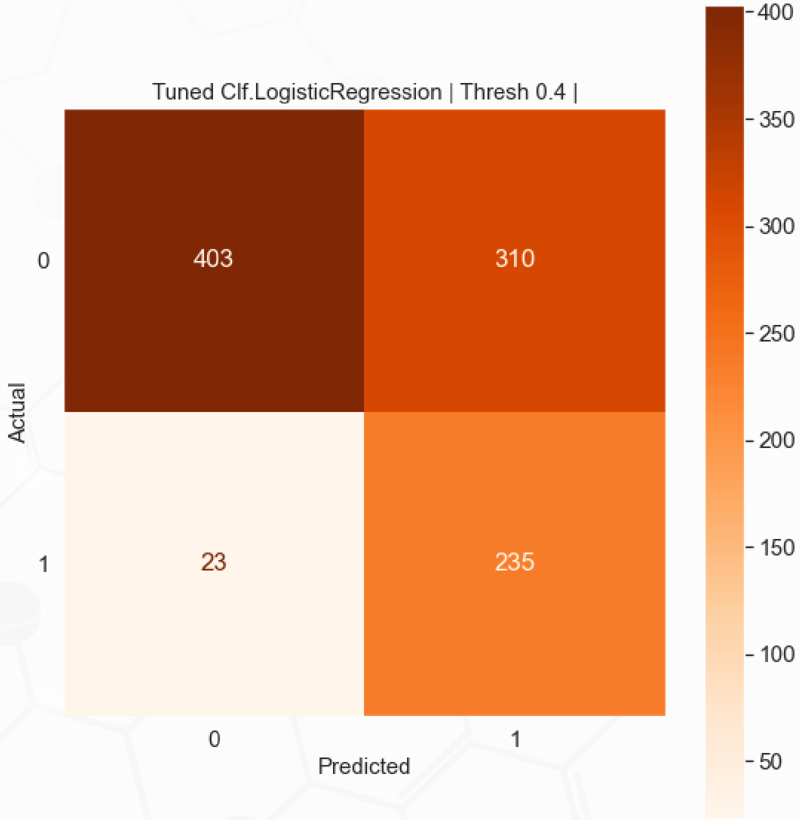
What can we see on the graph :

- By adjusting the threshold to <50% we can increase the recall score, by adjusting the threshold to >50% we can increase the precision.
- Since I use recall score for this case, I adjust the threshold to 40%.
- The projected recall is increase from default thresh (50%) by 9.2% (from 82.1% to 91.3%)
- There may be a slight difference, but the difference will not be significant

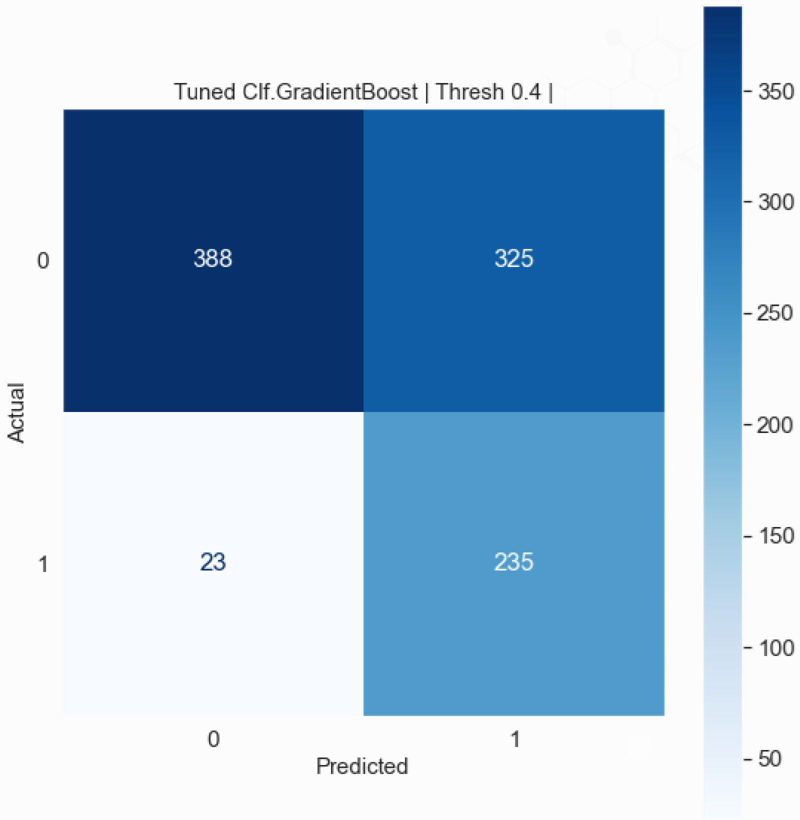
MACHINE LEARNING – TUNED + 40% THRESH

TELCO CUSTOMER CHURN

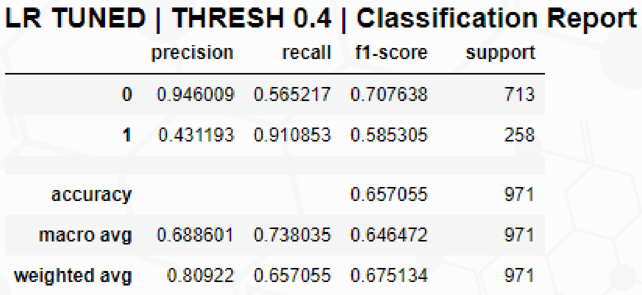
Repositories : [click here]
Notebook : [click here]



Tuned Clf.LogisticRegression | Thresh 0.4 |

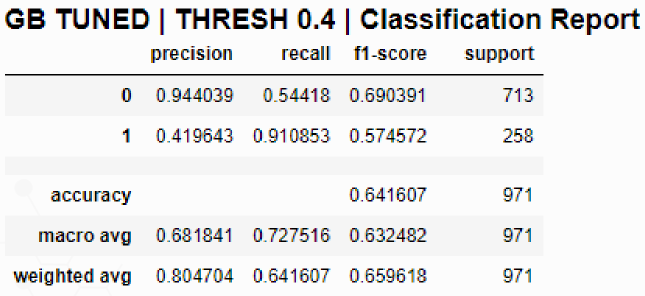


Tuned Clf.GradientBoost | Thresh 0.4 |



LR TUNED | THRESH 0.4 | Classification Report

	precision	recall	f1-score	support
0	0.946009	0.565217	0.707638	713
1	0.431193	0.910853	0.585305	258
accuracy			0.657055	971
macro avg	0.688601	0.738035	0.646472	971
weighted avg	0.80922	0.657055	0.675134	971



GB TUNED | THRESH 0.4 | Classification Report

	precision	recall	f1-score	support
0	0.944039	0.54418	0.690391	713
1	0.419643	0.910853	0.574572	258
accuracy			0.641607	971
macro avg	0.681841	0.727516	0.632482	971
weighted avg	0.804704	0.641607	0.659618	971

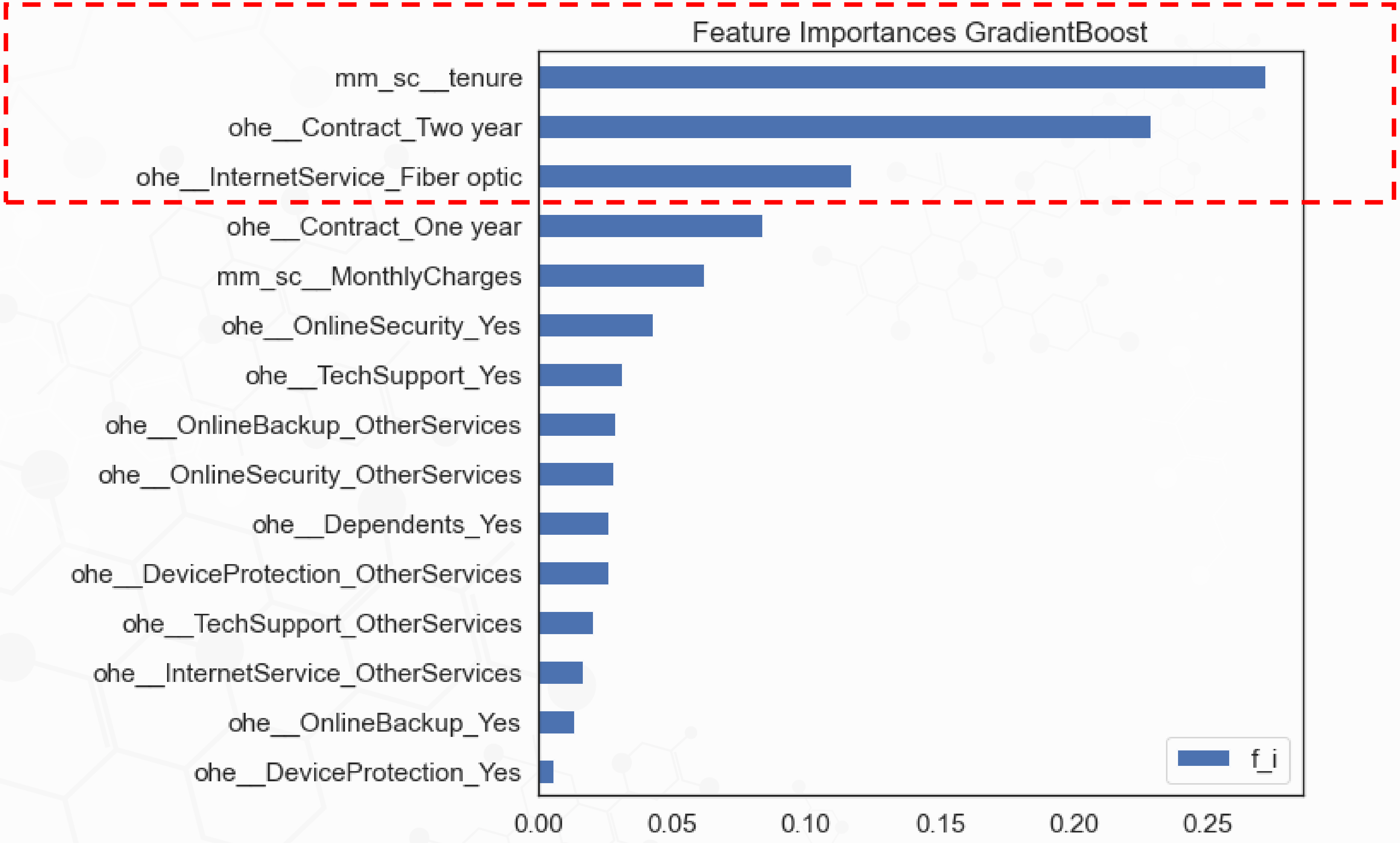
What can we see on the graph :

- There's an improment for recall score by using 40% threshold (compared to model with 50% threshold).
- We'll continue next process with 40% threshold

MACHINE LEARNING – FEATURE SELECTION

Repositories : click here
Notebook : click here

TELCO CUSTOMER CHURN



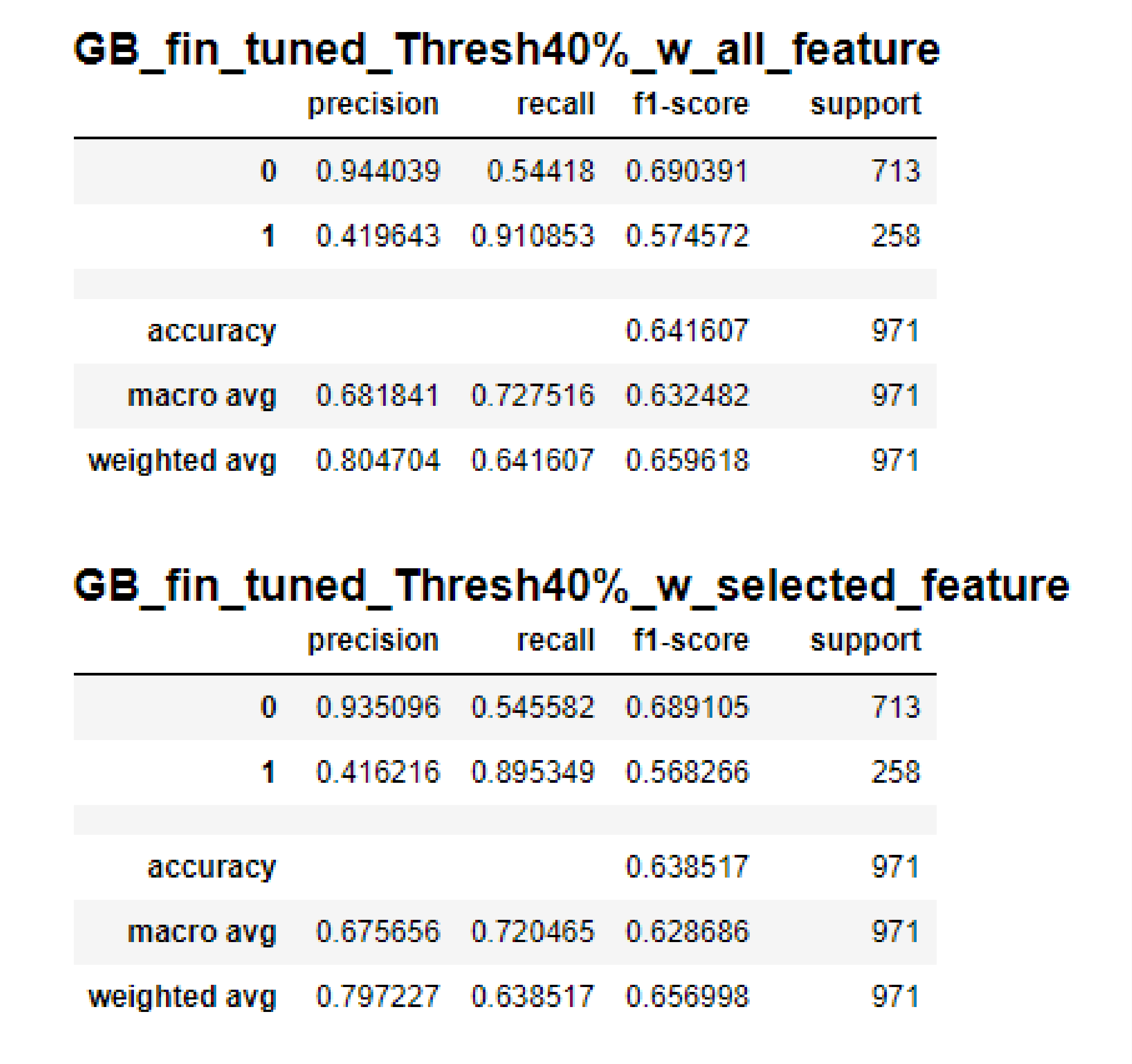
What can we see on the graph :

- The most important features area as follows Tenure, Contract, InternetService

MACHINE LEARNING – MODEL WITH FS

TELCO CUSTOMER CHURN

Repositories : click here
Notebook : click here



GB_fin_tuned_Thresh40%_w_all_feature

	precision	recall	f1-score	support
0	0.944039	0.54418	0.690391	713
1	0.419643	0.910853	0.574572	258
accuracy			0.641607	971
macro avg	0.681841	0.727516	0.632482	971
weighted avg	0.804704	0.641607	0.659618	971

GB_fin_tuned_Thresh40%_w_selected_feature

	precision	recall	f1-score	support
0	0.935096	0.545582	0.689105	713
1	0.416216	0.895349	0.568266	258
accuracy			0.638517	971
macro avg	0.675656	0.720465	0.628686	971
weighted avg	0.797227	0.638517	0.656998	971

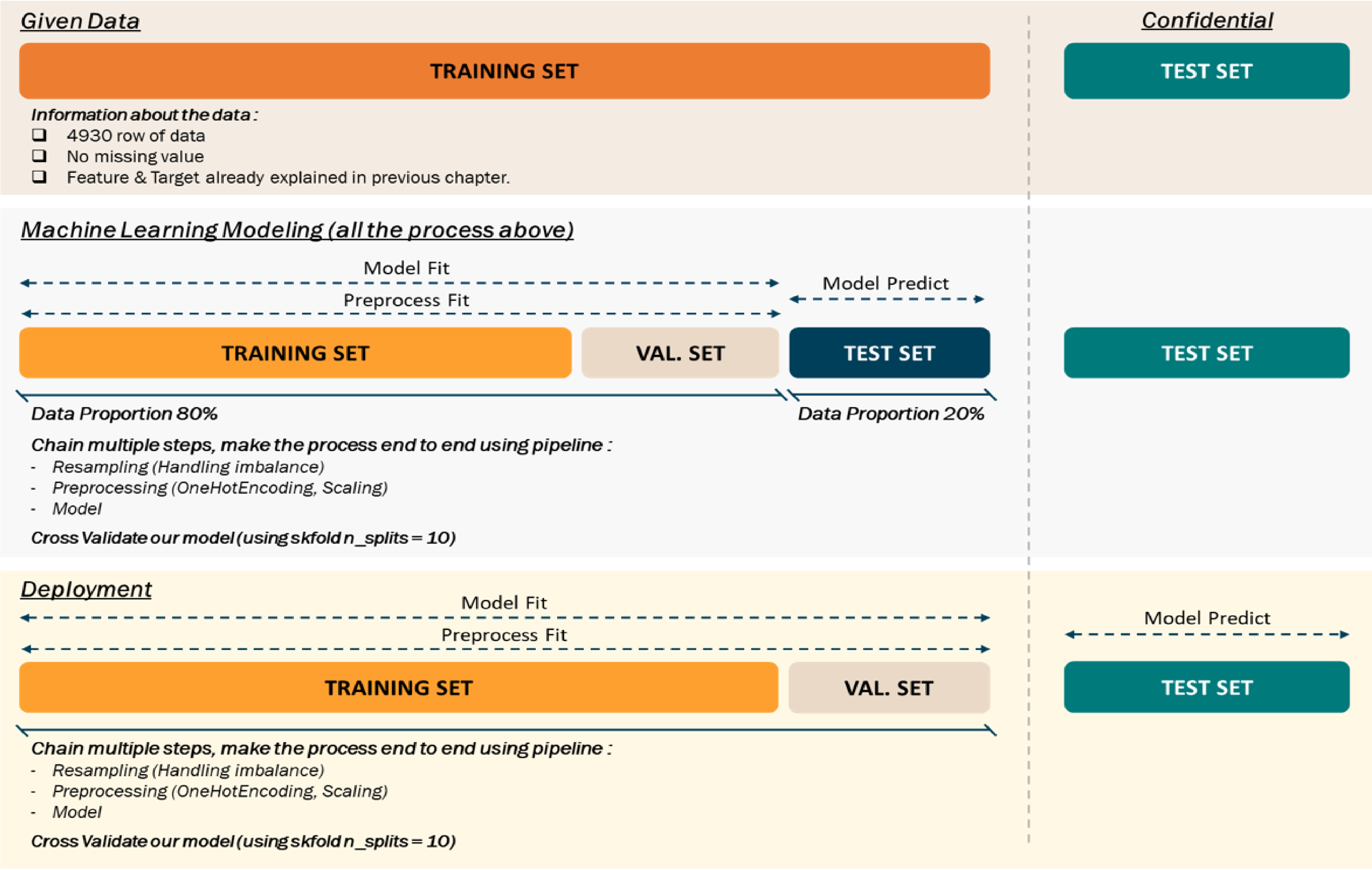
What can we see on the graph :

- There is no significant difference between models that use all features and models that use only selected features. (the recall decrease around 1.5%)
- Because of that I'll continue with model with selected feature only
- By doing this, we reduce redundant feature which does not affect the prediction result.

DEPLOYMENT

TELCO CUSTOMER CHURN

Repositories : click here
Notebook : click here



Given Data

TRAINING SET

Information about the data :

- ❑ 4930 row of data
- ❑ No missing value
- ❑ Feature & Target already explained in previous chapter.

Confidential

TEST SET

Machine Learning Modeling (all the process above)

Data Proportion 80%

Data Proportion 20%

Chain multiple steps, make the process end to end using pipeline :

- *Resampling (Handling imbalance)*
- *Preprocessing (OneHotEncoding, Scaling)*
- *Model*

Cross Validate our model (using skfold n_splits = 10)

Deployment

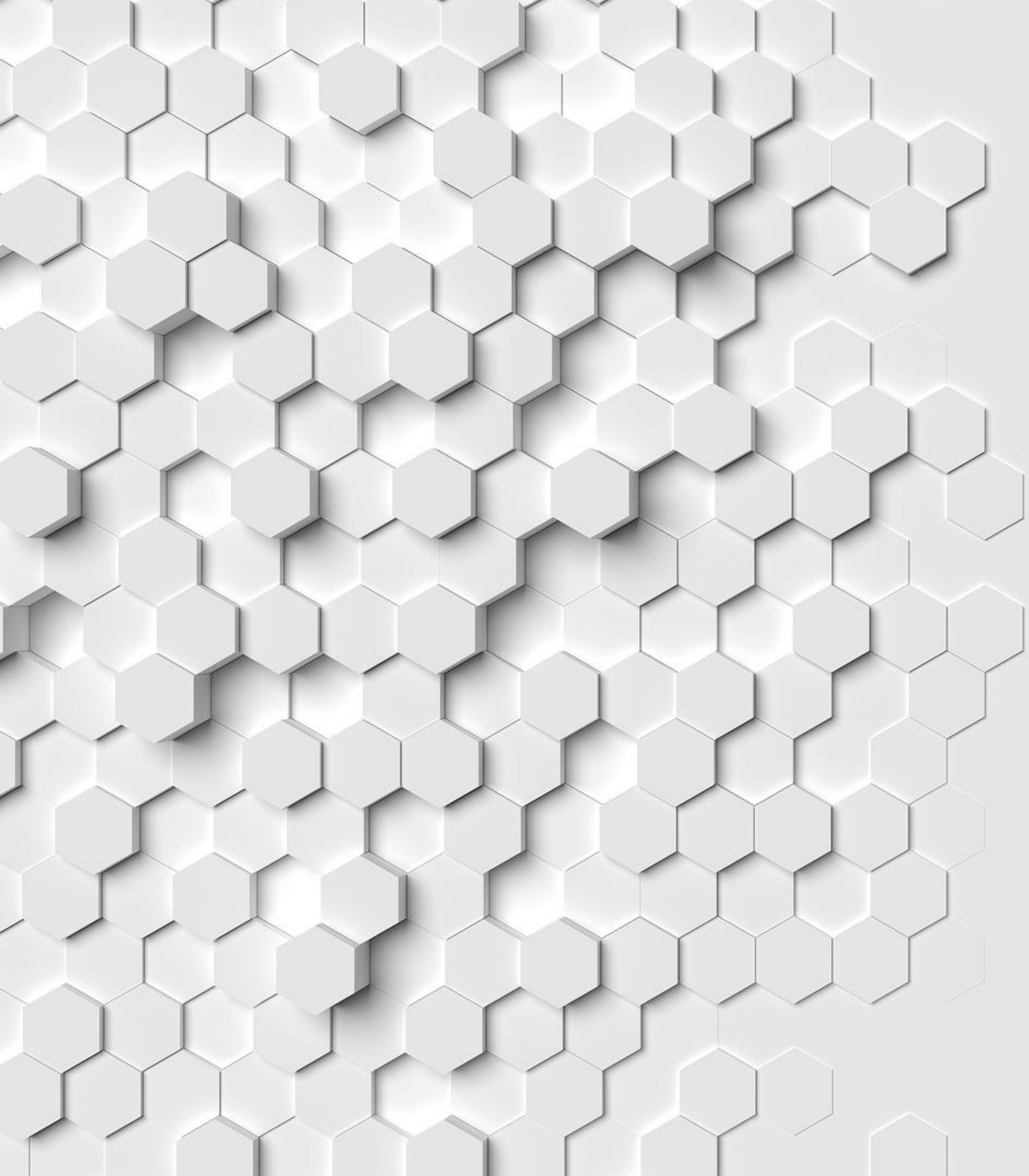
Chain multiple steps, make the process end to end using pipeline :

- *Resampling (Handling imbalance)*
- *Preprocessing (OneHotEncoding, Scaling)*
- *Model*

Cross Validate our model (using skfold n_splits = 10)

What can we see on the graph :

- From the given data, I split the data 80% for training and validation & 20% for test set.
- For deployment, I revert back the given data to 100% training and validation set.
- In the deployment phase, our model learn 20% more data (previously used as test set).
- By doing that I expect that the model will yield better result.



THANK YOU !

TELCO CUSTOMER CHURN

Data Analysis - Machine Learning

ALFIAN (JCDSVL-005-013)