Data Science 4-R

G.E Group |Binar Academy

FINAL PRESENTATION BINAR ACADEMY DATA SCIENCE

Muhammad Ghifari Kusuma
https://github.com/MasGhiff/

CHURN ESTIMATOR WITH MACHINE LEARNING APPROACH

- Data Understanding
- Data Preprocessing
- Data Cleansing
- Exploratory Data Analyst
- Model
- Result Evaluation

Dataset

No	Column Name	Type Data	Description
1	State	Object	The caller’s country
2	Account_length	Int	-
3	Area Code	Object	-
4	International Plan	Object	Option for international calling
5	Voice Mail Plan	Object	Option for Voicemail
6	Number Vmail Messages	Int	-
7	Total Day Minutes	Float	Duration of call in midday
8	Total Day Calls	Int	Number of call in midday
9	Total Day Charge	Float	Total charge per call in midday
10	Total Eve Minutes	Float	Duration of call in evening

No	Column Name	Type Data	Description
11	Total Eve Calls	Int	Number of calls in evening
12	Total Eve Charge	Float	Total charge per call in evening
13	Total Night Minutes	Float	Duration of night calling
14	Total Night Calls	Int	Number of calls in night
15	Total Night Charge	Float	Total charge per call in night
16	Total Intl Minutes	Gloat	Duration of international call
17	Total Intl Calls	Int	Number of international call
18	Total Intl Charge	Float	Total charge for international calling
19	Number Customer Service Calls	Int	Total of Customer Services
20	Churn	Object	Target Prediction

DATA PREPROCESSING

Data Cleansing

In cleansing data phase, we prepared the data, reduce the potential of mistake in training process. We're checked the potential of missing values, and duplicated data by using a command in Python.

```
[ ] df.isna().sum()
```

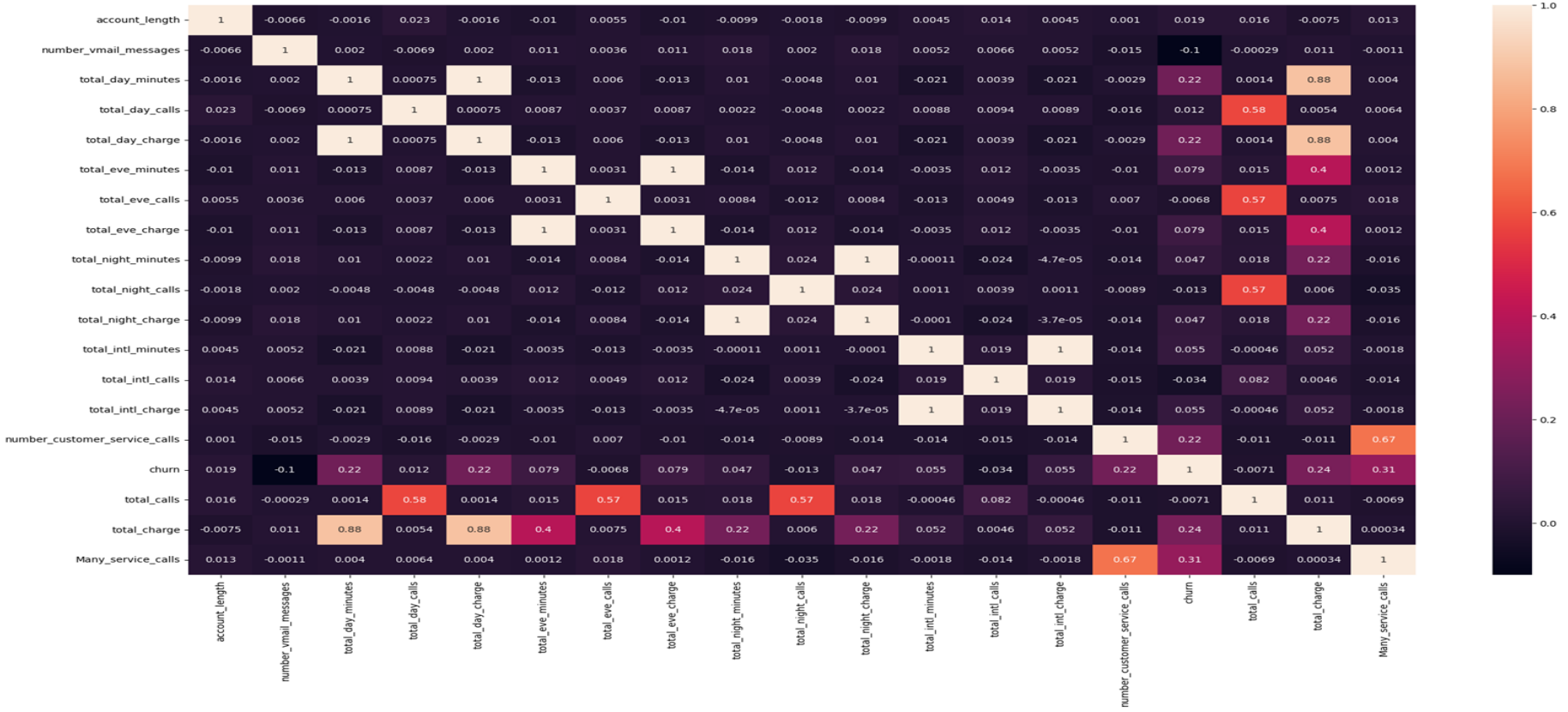
```
state                          0
account_length                 0
area_code                      0
international_plan             0
voice_mail_plan                0
number_vmail_messages          0
total_day_minutes              0
total_day_calls                0
total_day_charge               0
total_eve_minutes              0
total_eve_calls                0
total_eve_charge               0
total_night_minutes            0
total_night_calls              0
total_night_charge             0
total_intl_minutes             0
total_intl_calls               0
total_intl_charge              0
number_customer_service_calls  0
churn                          0
dtype: int64
```

```
[ ] df.duplicated().sum()
```

```
0
```

DATA PREPROCESSING

Data Correlation

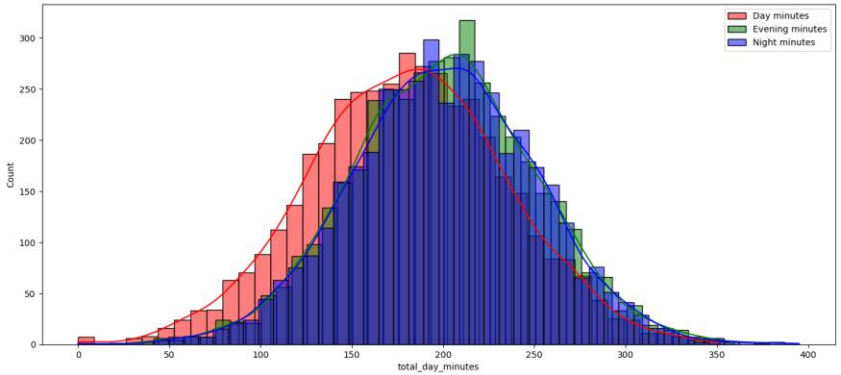


In this phase, we are used confusion matrix to found the correlation between the variable.

Data Preprocessing

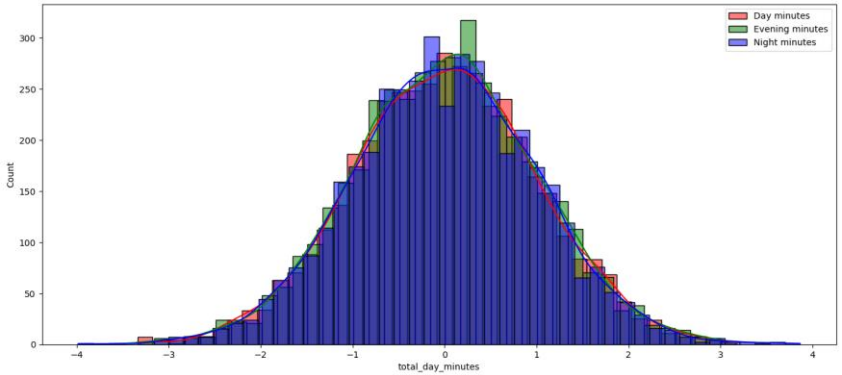
Normalization & Standardization

Before



	account_length	number_vmail_messages	total_day_minutes	total_day_calls
count	4250.000000	4250.000000	4250.000000	4250.000000
mean	100.236235	7.631765	180.259600	99.907294
std	39.698401	13.439882	54.012373	19.850817
min	1.000000	0.000000	0.000000	0.000000
25%	73.000000	0.000000	143.325000	87.000000
50%	100.000000	0.000000	180.450000	100.000000
75%	127.000000	16.000000	216.200000	113.000000
max	243.000000	52.000000	351.500000	165.000000

After



	account_length	number_vmail_messages	total_day_minutes	total_day_calls
count	4.250000e+03	4.250000e+03	4.250000e+03	4.250000e+03
mean	2.340611e-17	2.340611e-17	6.771054e-17	2.566313e-16
std	1.000118e+00	1.000118e+00	1.000118e+00	1.000118e+00
min	-3.716268e+00	-3.716268e+00	-3.337769e+00	-5.033498e+00
25%	-6.363089e-01	-6.363089e-01	-6.838979e-01	-6.502913e-01
50%	1.591774e-02	1.591774e-02	3.525533e-03	4.670679e-03
75%	6.319095e-01	6.319095e-01	6.654888e-01	6.596326e-01
max	3.530694e+00	3.530694e+00	3.170765e+00	3.279480e+00

Standard Scaler

$$\frac{X - \mu}{\sigma}$$

Min-Max Scaler

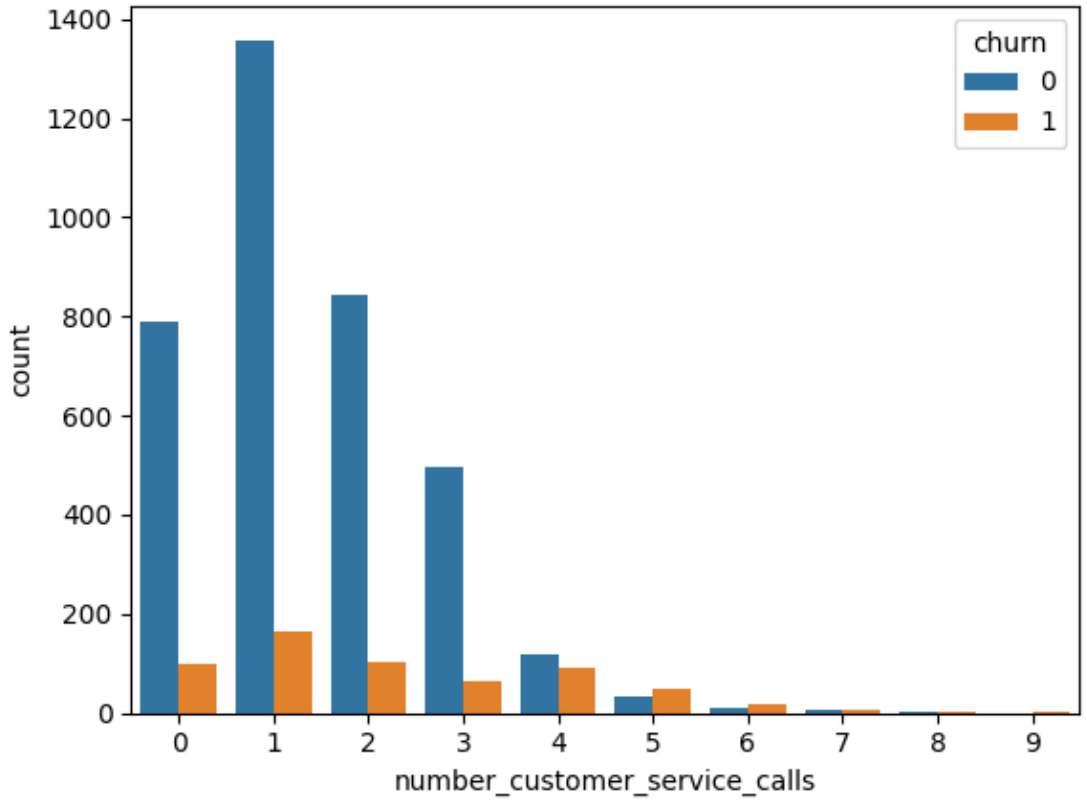
$$\frac{X - X_{min}}{X_{max} - X_{min}}$$

We're using standardization to rescale the range of dataset for centralization, and considering the high and low values we're using normalization to obtain a values with scale between 0-1.

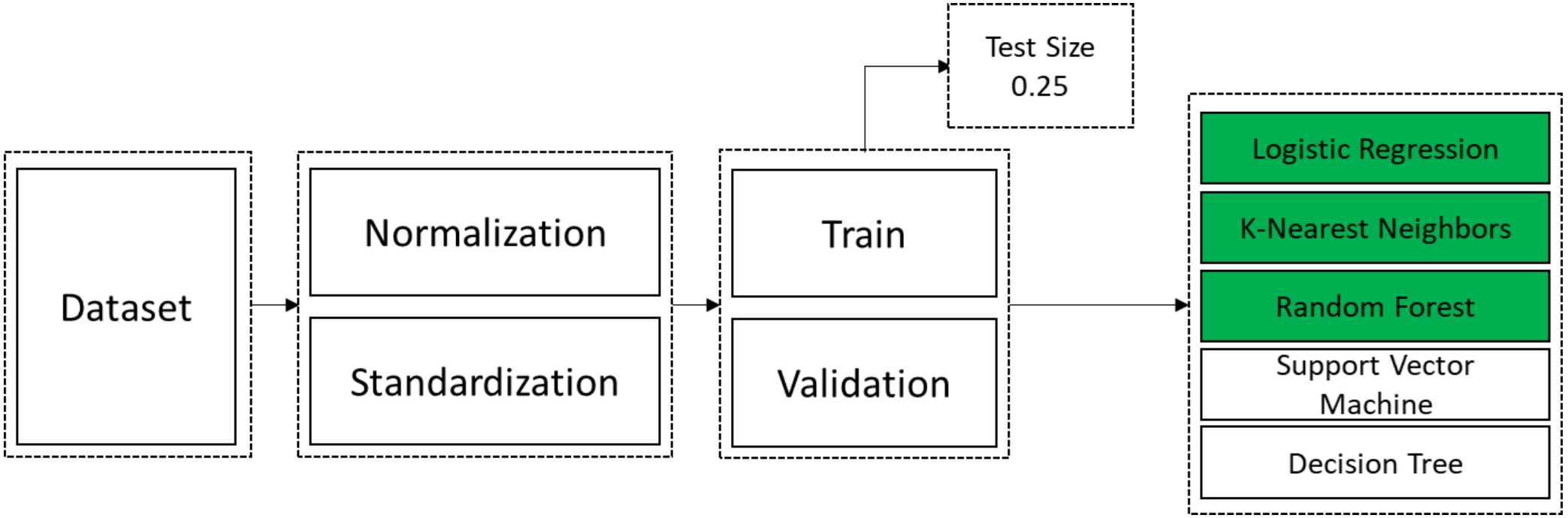
EXPLORATORY DATA ANALYST

Number Customer Service Call With Churn

The image shown the comparison of customer service to churning behavior of customers. Based on the data acquired, churn behavior relatively to low compared to non-churn behavior. The highest number of customer services shown the decreasing of churn and non churn behavior.



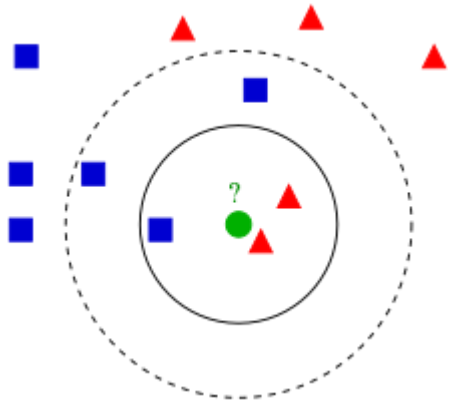
MODELING



In modelling phase, we are using a couple of method to compare the result acquired. The method consist of Support Vector Machine, Logistic Regression, KNN, Decision Tree, and Random Forest. Although many methods are used, we've concern of result in Logistic Regression, KNN, and Random Forest with expected good evaluation result. The result of each the method will evaluate in accuracy, precision, recall, and F*1*-score.

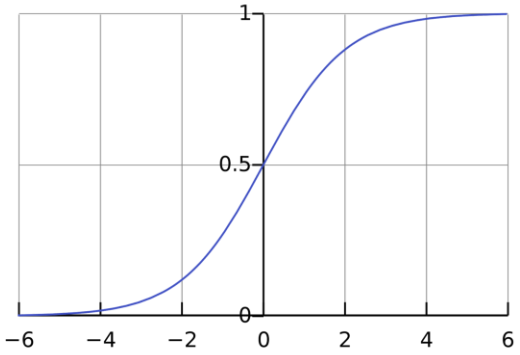
The Reason?

K-Nearest Neighbors



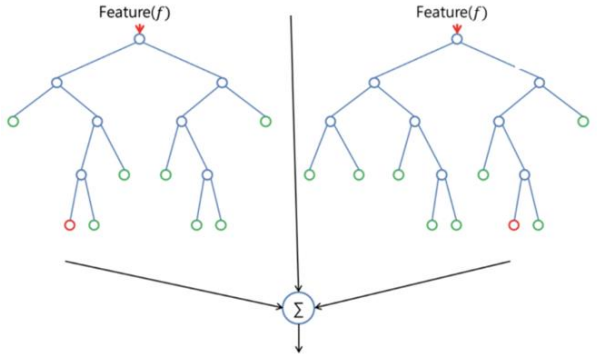
KNN is an approach for object classification considering the closer neighbors label. KNN is one of the best most popular method for binary classification

Logistic Regression



Logistic Regression is an approach for binary classification. Our dataset just has two labels or binary dataset. Considering the work principle, LR has an ability to obtain good evaluation for binary classification

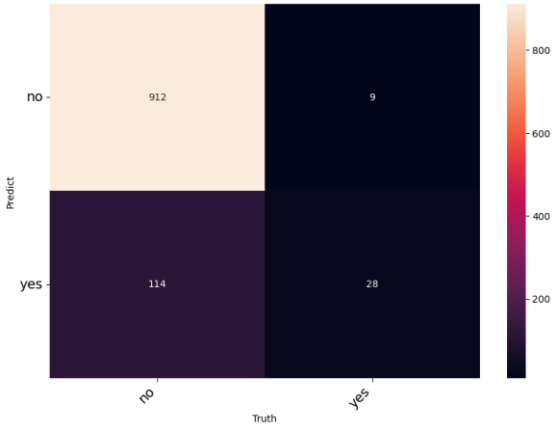
Random Forest



Random Forest is an approach for supervised learning. This method similar to decision tree but considering the principle of Random Forest, it has a potential for overfit reduction much better than decision tree for binary classification

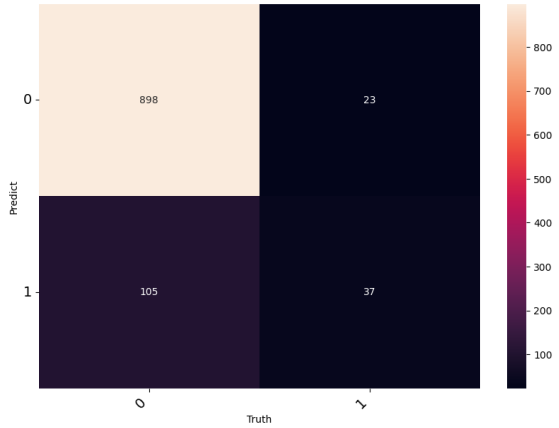
Confusion Result

K-Nearest Neighbors



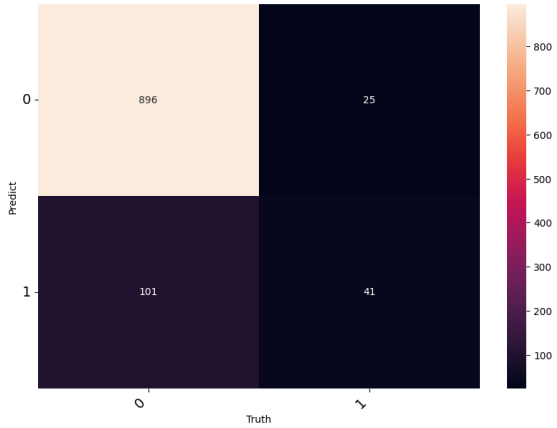
KNN Shown the evaluation in confusion matrix with “yes” churn prediction up to 28.

Logistic Regression



Logistic Regression Shown the evaluation in confusion matrix with “yes” churn prediction up to 37.

Random Forest



Random Forest Shown the evaluation in confusion matrix with “yes” churn prediction up to 41.

Result Comparison

Methods	Labels	Precision	Recall	F1-Score	Accuracy
Support Vector Machine	Yes	0.00	0.00	0.00	0.87
	No	0.87	1.00	0.93	
Logistic Regression	Yes	0.62	0.26	0.37	0.88
	No	0.90	0.98	0.93	
K-Nearest Neighbors	Yes	0.76	0.20	0.31	0.88
	No	0.89	0.99	0.94	
Decision Tree	Yes	0.28	0.30	0.29	0.80
	No	0.89	0.88	0.89	
Random Forest	Yes	0.62	0.29	0.39	0.88
	No	0.90	0.97	0.93	

As shown in the table, several of method acquired the accuracy up to 88%. The result still need to be improve by adding more dataset to balance the model.

Conclusion

- The increasing of number of customer service call shown the potential of churn positive behavior. It affected for social approach potential to customer from the provider.
- According the result from various of method, we have a concern in the highest of result from 3 methods. The 3 methods consist of KNN, Logistic Regression and Random Forest. These methods, shown the accuracy up to 88% with using 75% of training set, and 25% of validation set.
- Although the methods shown equal for accuracy obtained, Random Forest shown the highest of “yes” churn prediction up to 41. In addition, KNN shown the lowest result of “yes” churn prediction only 28.