DATA SCIENCE BOOTCAMP BATCH 7

HOMEWORK DAY 27

Telco Customer Churn Predictive Learning

Author:
Indra Yanto Simanihuruk

Supervisor:
Louis Owen

January 2022

Contents

Contents

1 Introduction

1.1 Background

Telephone and Internet services are now considered essential in most aspects of human life. This leads to rapid and persistent growth in the telecom industries and makes customer loyalty one of the key aspects of surviving competitive market conditions. According to a study carried out by Bain & Company [1], the acquisition of new customers can cost the company 5 times more than what customer retaining will do. Moreover, it also states that increasing customer retention rates by 5% can increase profits by 25% at least.

Churn rate is one of the important metrics for customer retention study that indicates if customers stop doing business with the corresponding company. It surely has negative effect on company's profit and need to be minimized so the company can continue to run. The reason for this churn decision are usually a combination of some factors such as poor service quality, better competitor exists, high prices, and so on.

The valuable basis to analyze this churn metric is none other than the data of customer itself. In this work, the data of certain telecom company will be studied to understand spesific customer who is likely to churn and the reasons behind that decision. Thus, the company can learn from it and come with reactive plans in the future. The dataset has also been pre-modified so it won't harm any customer's privacy.

1.2 Data Source

The customer base dataset used in this work is made available by IBM and downloaded from Kaggle [2]. It is related to an anonymous telecom company and contains 7043 customers data with 21 attributes where each row represents a customer and each column contains customer's attributes. Overall, the dataset provides information about :

- The target variable is Churn, indicates customers who left within the last month.
- Services that each customer signed up for, consist of phone service, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies.
- The information about customer account consists of tenure, contract, payment method, paperless billing, monthly charges, and total charges.
- There are also gender, age range, partners, and dependents to give information about customer demographic.

The predictive analysis is carried out using Google Colab (Python environment). This work uses a few of libraries such as pandas, NumPy, Matplotlib, Seaborn, and Sklearn [3].

1.3 Business Objectives

The analysis is carried out to answer this problem :

1. Which machine learning model is the most accurate on predicting the churned customer and how is the performance? (Random Forest, KNN, and Logistic Regression are the only models that will be examined for now)

2 Data Profiling and Cleansing

2.1 Data Profiling

The dataset has 7043 rows and 21 columns which include informations about :

Table 1: Description of each column.

Column	Description
customerID	Represents Unique Id Number of each customer (do not give any useful informations)
gender	Represents the customer gender
SeniorCitizen	Whether the customer is a senior citizen or not (1, 0)
Partner	Categorical data, identifies whether the customer has a partner or not (Yes, No)
Dependents	Whether the customer has dependents or not (Yes, No)
tenure	Number of months the customer has stayed with the company
PhoneService	Whether the customer has a phone service or not (Yes, No)
MultipleLines	Whether the customer has multiple lines or not (Yes, No, No phone service)
InternetService	Customer's internet service provider (DSL, Fiber optic, No)
OnlineSecurity	Whether the customer has online security or not (Yes, No, No internet service)
OnlineBackup	Whether the customer has online backup or not (Yes, No, No internet service)
DeviceProtection	Whether the customer has device protection or not (Yes, No, No internet service)
TechSupport	Whether the customer has tech support or not (Yes, No, No internet service)
StreamingTV	Whether the customer has streaming TV or not (Yes, No, No internet service)
StreamingMovies	Whether the customer has streaming movies or not (Yes, No, No internet service)
Contract	The contract term of the customer (Month-to-month, One year, Two year)

PaperlessBilling	Whether the customer has paperless billing or not (Yes, No)
PaymentMethod	The customer's payment method (Electronic check, Mailed check, Bank transfer/automatic, Credit card/automatic)
MonthlyCharges	The amount charged to the customer monthly
TotalCharges	The total amount charged to the customer
Churn	Target variable, define whether the customer churns or not (Yes or No)

One can clearly see from Table 1 that most columns (17) are categorical variables except customerID, tenure, MontlyCharges, and TotalCharges. The customerID variable is feature that represents unique Id number of each customer (in total has 7043 unique values) hence it won't give any useful informations and will be dropped from analysis meanwhile tenure, MonthlyCharges and TotalCharges are considered as numerical variables.

2.2 Data Cleansing

It's important to eliminate all missing values and anomalies from the data before it will be analyzed further. In this telco customer dataset, TotalCharges is the only column where a few of missing values are detected.

	tenure	MonthlyCharges	TotalCharges	Churn
488	0	52.55		No
753	0	20.25		No
936	0	80.85		No
1082	0	25.75		No
1340	0	56.05		No
3331	0	19.85		No
3826	0	25.35		No
4380	0	20.00		No
5218	0	19.70		No
6670	0	73.35		No
6754	0	61.90		No

Figure 1: All missing values detected in TotalCharges match 0 tenure period.

There are in total 11 rows with missing value found in TotalCharges column and the reason behind these missing values is actually related to 0 tenure month (there are also 11 rows with 0 tenure). In addition, these rows with 0 period of tenure may emphasize the customer that has just signed up for the telecom service within the last month. Since they won't provide any useful information (the customers have

just signed up) and the amount of rows are only 11, i.e 0.1% of total rows, these rows will be eliminated and not included in the analysis.

Another thing one should be wary of before doing the analysis is the outlier. An outlier is a value that deviates extremely from the rest observations within the sample data. It may indicate bad and wrong observations or anomalies that need to be eliminated from the data. However, there are certain circumstances that outliers can reveal insights into special cases within the data that one may not otherwise notice. The best practice to detect the outlier is to use boxplot and it is shown in Figure 2 that there is no outlier in the dataset.

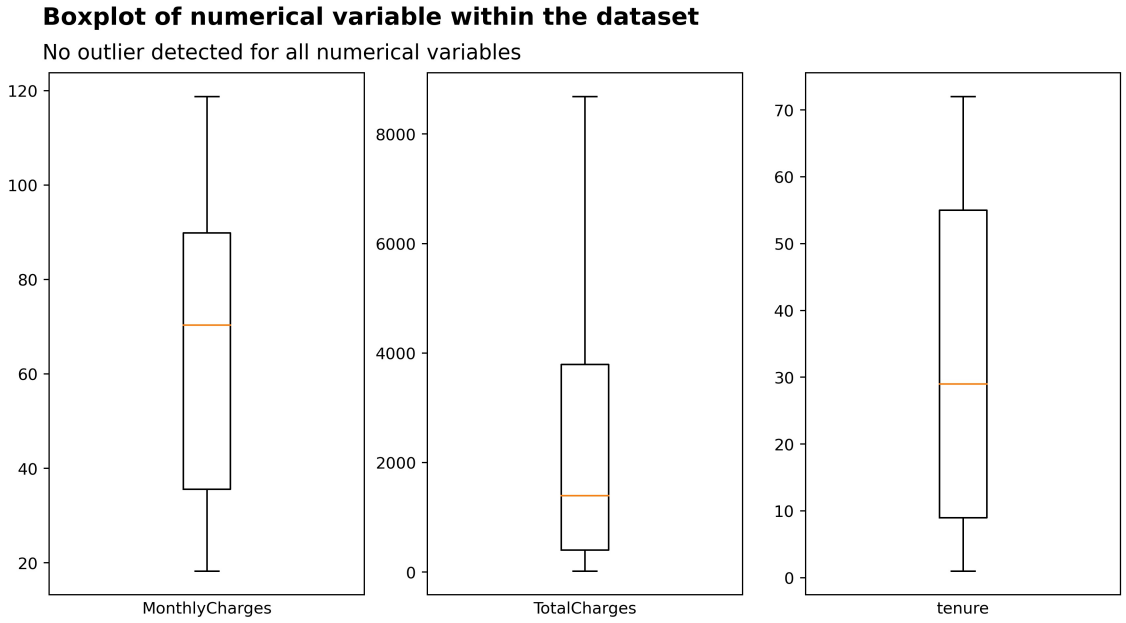

Figure 2: No outlier found in the boxplots of MonthlyCharges, TotalCharges, and tenure.

2.3 Feature Engineering

It can be seen that there is a huge imbalance for the target variable since it contains 5163 rows of No entries (73.42%) and 1869 rows of Yes entries (26.58%), indicates that the corresponding company has 26.57% churn rate within the last month. In the preceding section, one also can understand that there 3 numerical variables exist within the dataset such as tenure, MonthlyCharges and TotalCharges. The values of TotalCharges is very far huge compared to tenure and MonthlyCharges, as a result, these 3 features will be Standard Scaled.

Sklearn model can't process any data types aside from numerical variables. Hence, label encoding is a must. To avoid multicollinearity and increase effectivity, these steps have to be done beforehand :

1. No and No phone service in MultipleLines feature will be merged as one value No to avoid the multicollinearity with No from PhoneService feature.

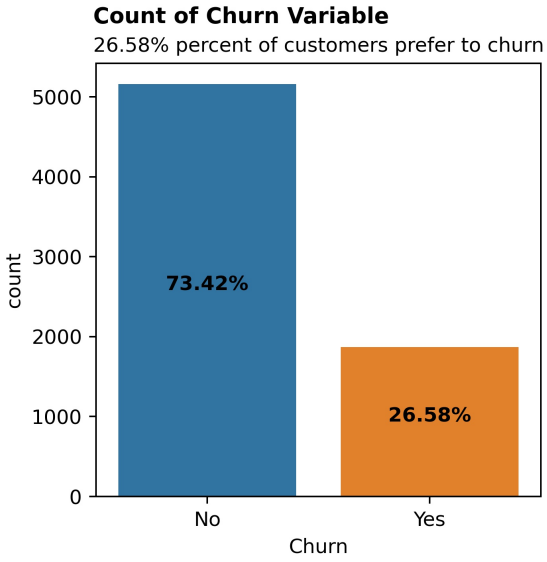

Figure 3: Univariate analysis of Churn variable.

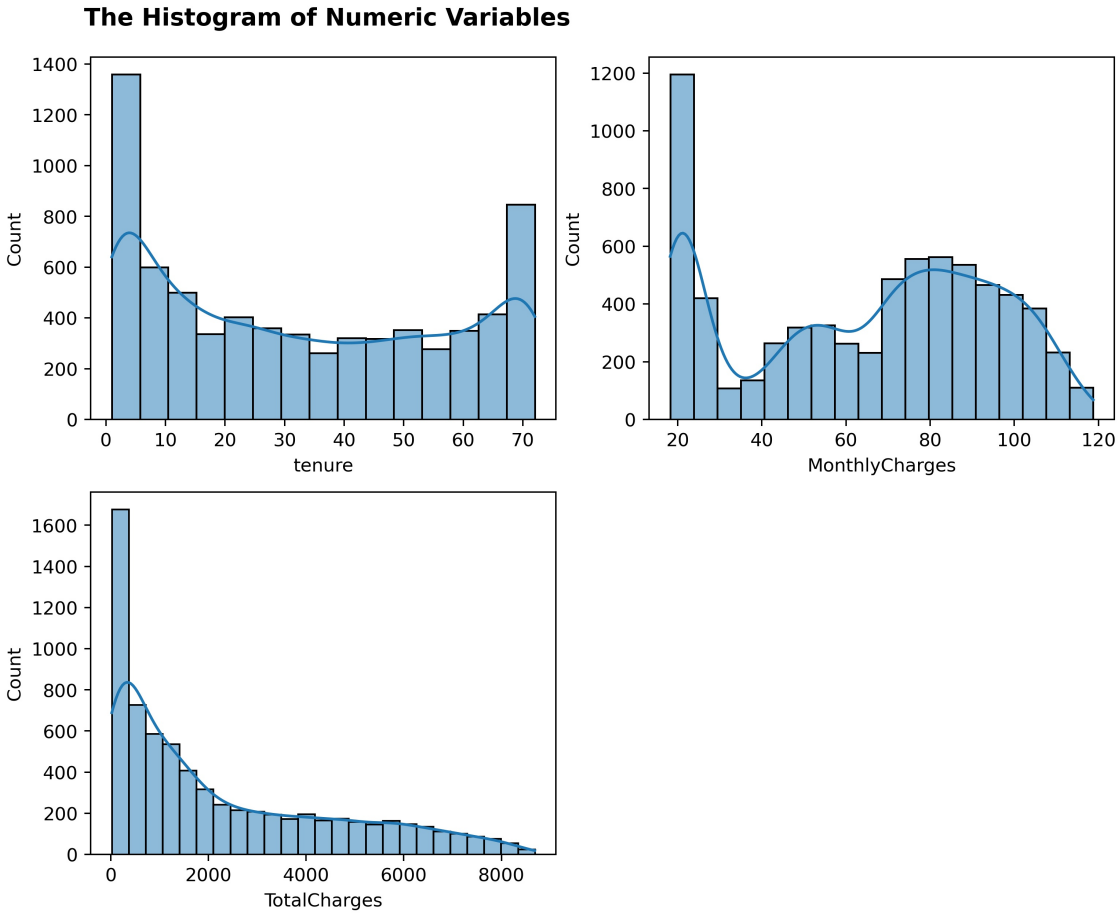

Figure 4: Univariate analysis of numeric variables.

2. No and No internet service in OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, and StreamingMovies will also be merged as one value No to avoid the multicollinearity with No from InternetService feature.

Combining these 2 steps with drop_first feature enabled will reduce the amount of encoded columns significantly. The correlation heatmap for the encoded dataset can be seen in Figure 6.

Special only for the regression model, different model of dataset will be created. First, The VIF analysis of each feature will be generated to prevent the multicollinearity. As a result, MonthlyCharges, PhoneService_Yes, and TotalCharges will be deleted since these columns have high score of VIF factor.

	VIF Factor	features
0	7.583795	tenure
1	17.269869	MonthlyCharges
2	10.809329	TotalCharges
3	2.020069	gender_Male
4	1.376281	SeniorCitizen_Yes
5	2.827071	Partner_Yes
6	1.969201	Dependents_Yes
7	10.299291	PhoneService_Yes
8	2.640337	MultipleLines_Yes

Figure 5: Vif analysis for the encoded dataset.

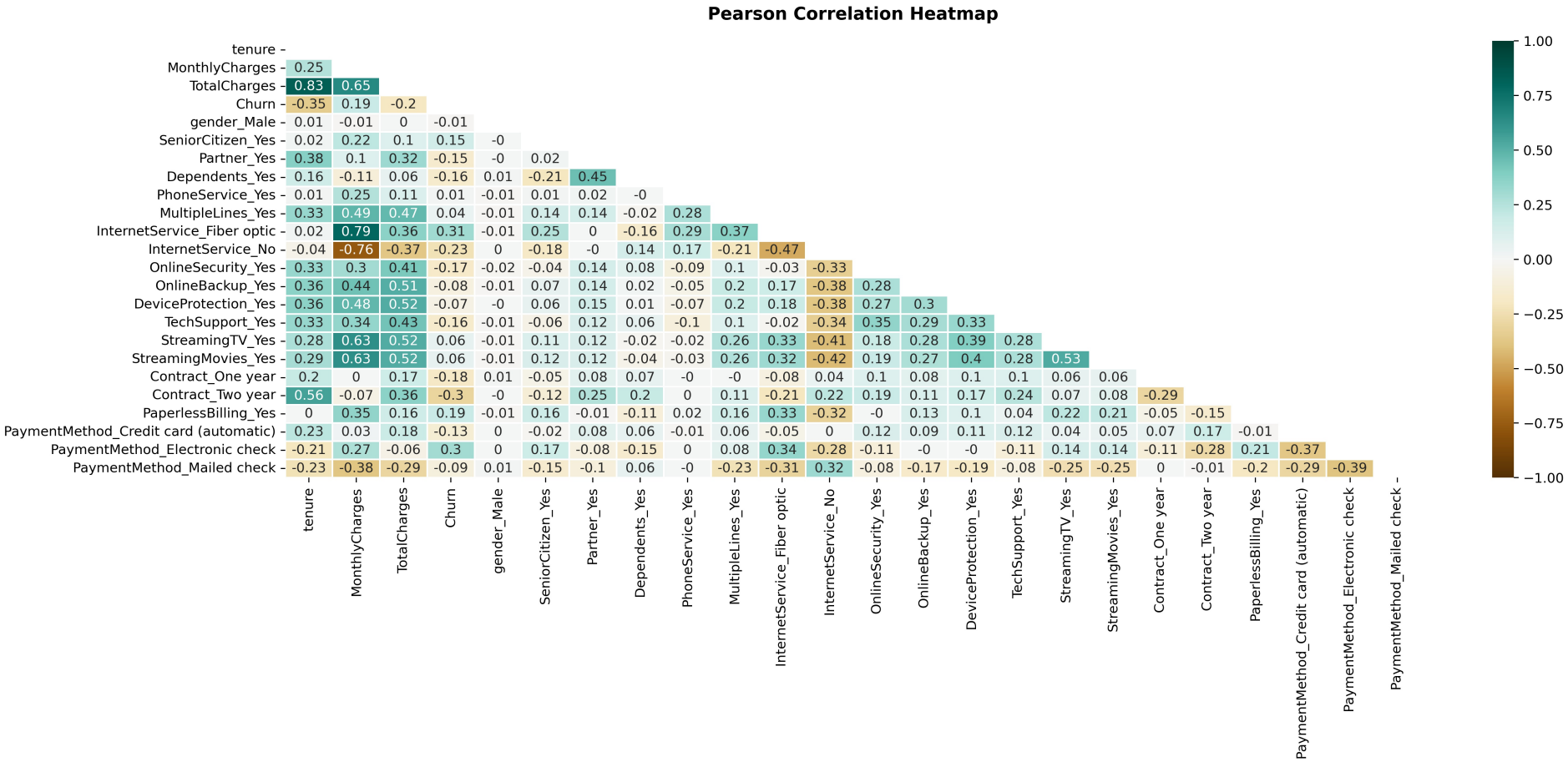

Figure 6: Pearson Correlation Heatmap for the encoded dataset.

3 Data Processing

Data Processing is the phase where the machine learning model is being applied. Before applying the models, the dataset will be splitted into train and test data with quantity of 80% and 20% respectively. The train data contains 4125 of No entries and 1500 of Yes entries.

Imbalanced dataset that have been prepared will be used to train all the models. Random Forest, KNN, and Logistic Regression are the only models that will be examined for now. Before training the model, the analysis of the confusion matrix will be carried out first to determine what metric explains our problem better :

- TP : Prediction is Churn and the actual data is also Churn (good)
- TN : Prediction is Not Churn and the actual data is also Not Churn (good)
- FP : Prediction is Churn, but the actual data is Not Churn (bad)
- FN : Prediction is Not Churn, but the actual data is Churn (worst, loss for the company)

For all the cases, it is evident that FN must be kept as low as possible to minimize the company losses. FN is considered as the worst case, that having not churn customer predicted as churn is slightly better than churn customer predicted as not churn. As a result, Recall metric is very important to be examined on this problem. Keep in mind that there is a huge imbalance in the actual train data, making the AUC and F1 score are also quite important to be analyzed.

3.1 Random Forest Classifier

First, the sklearn's model of RandomForestClassifier will be applied to learn the dataset with all hyperparameter are set to default. It is found that the ROC AUC score of this model is 82% meanwhile the accuracy is 79% (Figure 7).

Please note that the hyperparameter tuning method used in this work is the GridSearch method (GridSearchCV). To obtain better performance of this model, the parameters that will be tuned consist of :

- criterion : function to measure the quality of a split (gini, entropy).
- max_features : number of features to consider when looking for the best split ('sqrt',0.3,0.4,0.5).

Unlike the DecisionTree, RandomForest is quite robust to overfitting and the result is determined by using cumulative voting of all the trees. That's why min_samples_split and max_depth are not included in the parameter space. The n_estimators hyperparameter is also kept constant on its default value, i.e 100 to avoid very long training time. The optimum tuning will be determined by prioritizing the ROC AUC score.

The hyperparameter tuning results in 82% AUC score (train data) with tuned parameters criterion is entropy, max_features is 0.3, and n_estimators is 100. Applying the tuned model on test data gives such results in Figure 8.

Test Data Evaluation :

	precision	recall	f1-score	support
0	0.83	0.90	0.86	1038
1	0.63	0.49	0.55	369
accuracy			0.79	1407
macro avg	0.73	0.69	0.71	1407
weighted avg	0.78	0.79	0.78	1407

AUC score is 0.8227099226676275

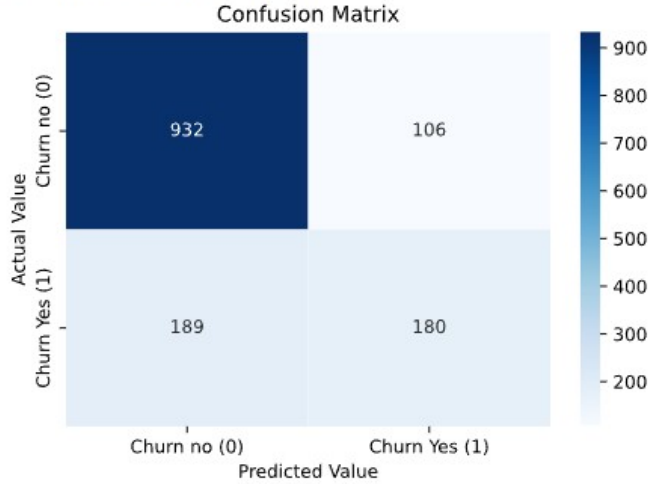

Figure 7: Test data evaluation of default RF model.

Test Data Evaluation :

	precision	recall	f1-score	support
0	0.83	0.90	0.87	1038
1	0.64	0.49	0.55	369
accuracy			0.79	1407
macro avg	0.73	0.69	0.71	1407
weighted avg	0.78	0.79	0.78	1407

AUC score is 0.8300893943428835

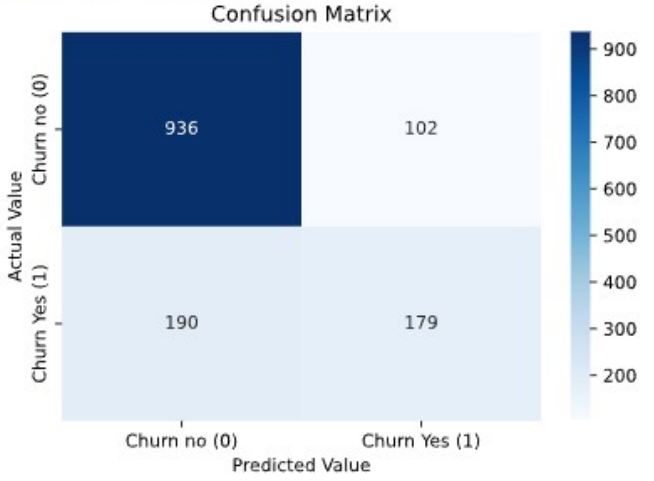

Figure 8: Test data evaluation of tuned RF model.

As can be seen from the results, the tuning process does not result in any significant improvement over the default model. Only in terms of ROC-AUC and Precision score does the tuned model perform better, with a score of 83 percent and 64 percent, respectively.

As explained before, the amount of FN is very important to be kept as low as possible to minimize the loss for the company. One way to check how far we can maximize the Recall score of our model is by plotting the Recall and Precision curve vs the Decision probability threshold (Figure 9). Based on the chart above, it is found that Precision and Recall have such inverse relationship that maximizing the score of Recall will respectively reduce the score of Precision. As a result, setting the recall score of 83% is not considered as good approach since the Precision score will also be significantly reduced to 45% , as well as the accuracy score (70%) and F1 score (59%). In other words, not knowing the allowed minimum limit of Precision score is the obstacles on determining the optimum Recall point.

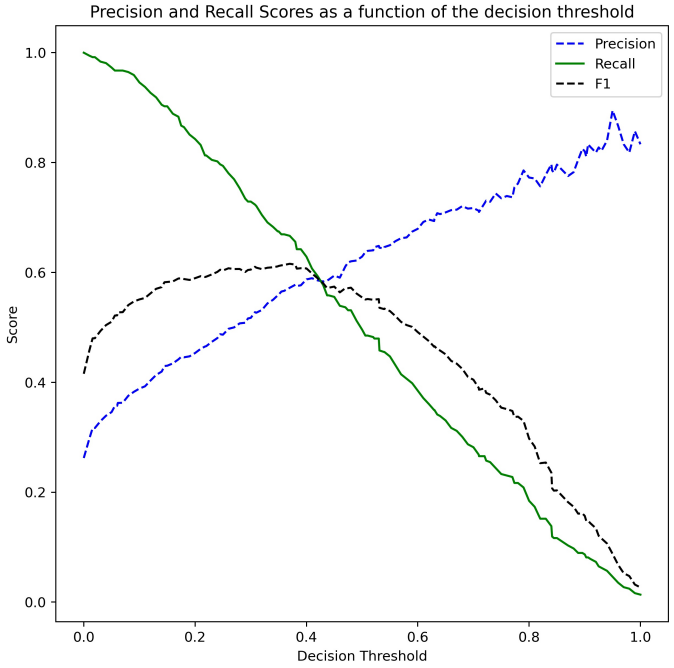

Figure 9: Precision and Recall scores as a function of the decision threshold for RF model.

To deal with this problem, the point where F1 reaches its maximum value will be set as the optimum point. In this case, the probability threshold where F1 is maximum is 0.37. Fortunately, adjusting the decision threshold based on the best F1 score improves the situation significantly. With an AUC score of 83% (+1%), the F1 and Recall scores are significantly higher, at 61% (+5%) and 66% (+17%), respectively, while the accuracy is 78%, only 1% lower compared to the default model. This result can be seen in Figure 10.

Test Data Evaluation :

	precision	recall	f1-score	support
0	0.87	0.83	0.85	1038
1	0.58	0.66	0.61	369
accuracy			0.78	1407
macro avg	0.72	0.74	0.73	1407
weighted avg	0.79	0.78	0.79	1407

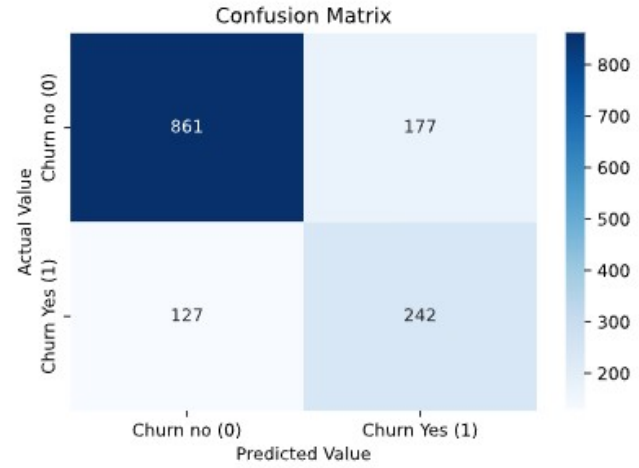

Figure 10: Test data evaluation of tuned RF model with adjusted decision threshold.

3.2 Logistic Regression Classifier

The sklearn's model of LogisticRegression will be applied to learn the dataset with all hyperparameter are set to default. Please note that the training and test for this one (only) are quite different since the MonthlyCharges, TotalCharges and PhoneService_Yes columns are dropped due to high VIF score. For the result, it is found that the ROC AUC score of this untuned model is 83% meanwhile the accuracy is 80% (Figure 11). For hyperparameter tuning, the parameters that will be tuned consist of :

- C : inverse of regularization strength, the smaller the stronger regularization will be (0.01, 0.1, 1, 10, 100).
- solver : optimization algorithm (newton-cg, lbfgs, saga, sag, liblinear).

The penalty hyperparameter will be specified as L2 (default). Similar as before, the optimum tuning will also be determined by prioritizing the ROC AUC score (method : GridSearch).

The hyperparameter tuning results in 84% AUC score (train data) with tuned parameters C is 0.1, and solver is lbfgs. As expected for Logistic Regression Model, the tuning process does not result in any significant improvement over the default model on test data (Figure 12). All the metrics for both models are almost identical with such values :

- Accuracy : 80%
- AUC : 83%

Test Data Evaluation:

	precision	recall	f1-score	support
0	0.84	0.90	0.87	1038
1	0.65	0.53	0.58	369
accuracy			0.80	1407
macro avg	0.74	0.71	0.72	1407
weighted avg	0.79	0.80	0.79	1407

AUC score: 0.8337393152351562

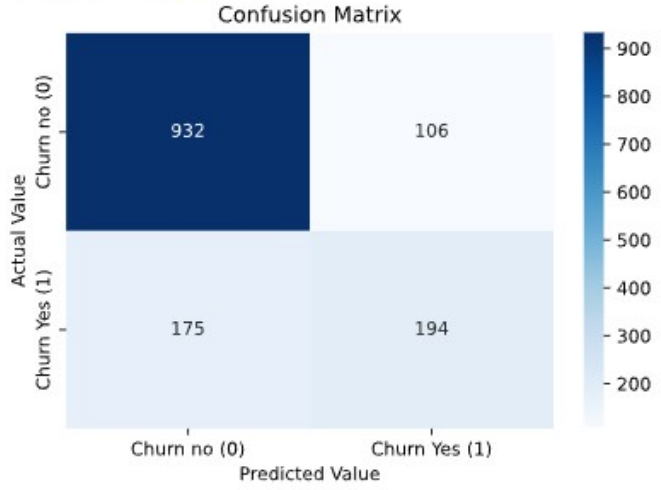

Figure 11: Test data evaluation of default LogisticRegression model.

Test Data Evaluation:

	precision	recall	f1-score	support
0	0.84	0.90	0.87	1038
1	0.66	0.52	0.58	369
accuracy			0.80	1407
macro avg	0.75	0.71	0.73	1407
weighted avg	0.79	0.80	0.80	1407

AUC score is 0.8330370057072439

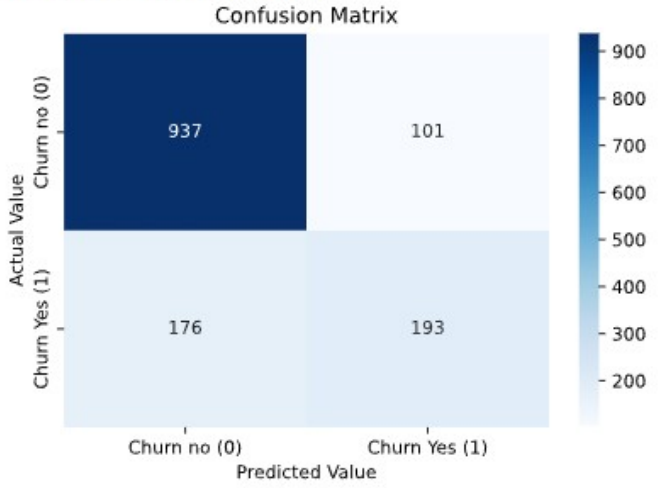

Figure 12: Test data evaluation of tuned LogisticRegression model.

- Precision : 66%
- Recall : 52%
- F1 : 58

Similar as before, the decision/probability threshold will be adjusted to maximize the F1 score and obtain better recall score. From Figure 13 , it is found that the best decision threshold for maximum F1 score is 0.3.

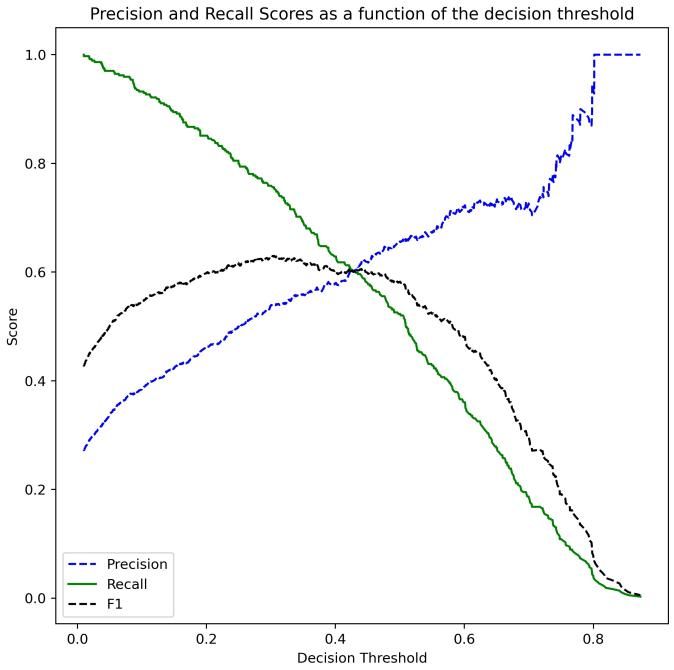

Figure 13: Precision and Recall scores as a function of the decision threshold for LogisticRegression model.

After adjusting the decision threshold, the improvement becomes very clear as the Recall and F1 score are significantly increased to 76% (+23%) and 63% (+5%) respectively (Figure 14), compared to the default model. With an AUC score of 83% (+0%), those metrics are successfully increased without a lot of reduction in accuracy, only 3% lower from the default model with the score of 77% .

Test Data Evaluation:

	precision	recall	f1-score	support
0	0.90	0.77	0.83	1038
1	0.54	0.76	0.63	369
accuracy			0.77	1407
macro avg	0.72	0.76	0.73	1407
weighted avg	0.80	0.77	0.78	1407

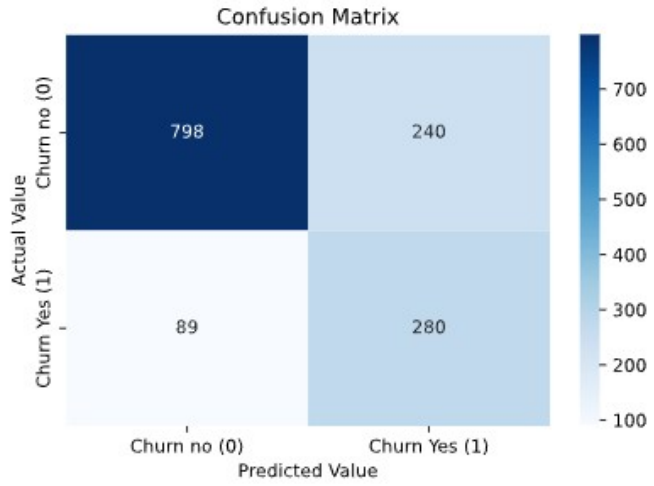

Figure 14: Test data evaluation of tuned LogisticRegression model with adjusted decision threshold.

3.3 KNN Classifier

In this section, The sklearn's model of KNeighborsClassifier will be applied to learn the dataset with all hyperparameter are set to default. Remember that the default value of neighbors in this model is 5. That's why the AUC and accuracy for this model are relatively lower than other models with score of 78% and 77% respectively (Figure 15) .

For this model, The parameters that will be tuned is n_neighbors or number of neighbors with value range of 3-81. Remember that the performance result of KNN model heavily relies on n_neighbors parameter or number of neighbors. The space of n_neighbors is set from 3 to 81 since the best value usually lies around the root of number of samples in training data (75). Best number of neighbors will be determined based on maximum ROC AUC score. Other hyperparameters will be set to default.

The tuning result can be seen in Figure 16 and 17. Based on the tuning process (Figure 16), 78 is considered as the best number of neighbors with AUC score of 84% (training data). This number of neighbors will be set as the n_neighbors of the tuned model. From Figure 17, one can observe that the improvement of the tuned model is very significant compared to default model. The AUC score increases by 5% to 83%, as well as the accuracy, i.e 79% (+2%) and F1 score, 59% (+4%).

Let's check if the Recall and F1 score still can be improved by adjusting the decision threshold. From Figure 18, the threshold where F1 is maximum lies around 0.36. After adjusting the tuned KNN model to that value, the F1 score increases to

```
Test Data Evaluation :
              precision    recall  f1-score   support

           0       0.84      0.85      0.85      1038
           1       0.56      0.54      0.55       369

    accuracy                           0.77      1407
   macro avg       0.70      0.70      0.70      1407
weighted avg       0.77      0.77      0.77      1407

AUC score is  0.7816300369169396
```

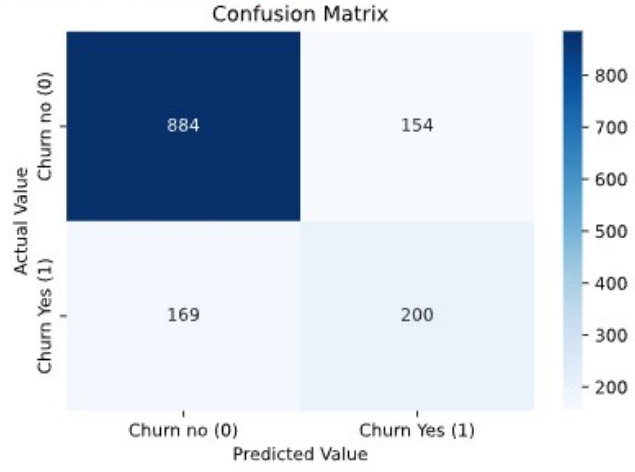

Figure 15: Test data evaluation of default KNN model.

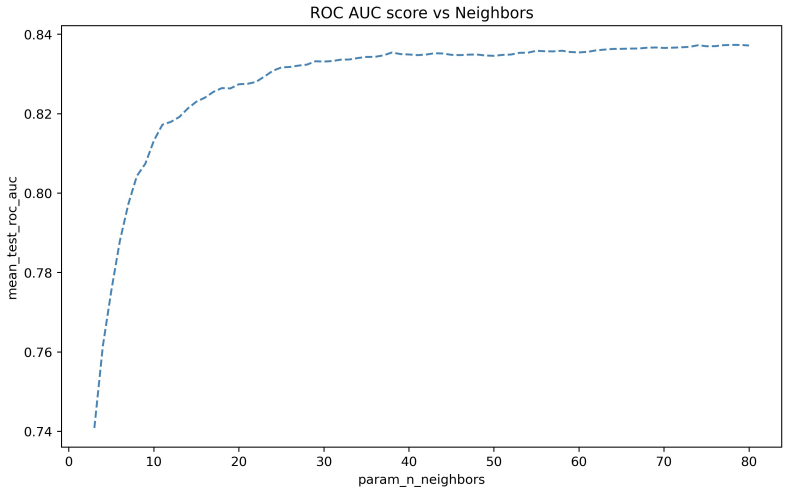

Figure 16: Line plot of ROC AUC score and Neighbors.

Test Data Evaluation :

	precision	recall	f1-score	support
0	0.85	0.88	0.86	1038
1	0.62	0.56	0.59	369
accuracy			0.79	1407
macro avg	0.73	0.72	0.72	1407
weighted avg	0.79	0.79	0.79	1407

AUC score is 0.8328973270464881

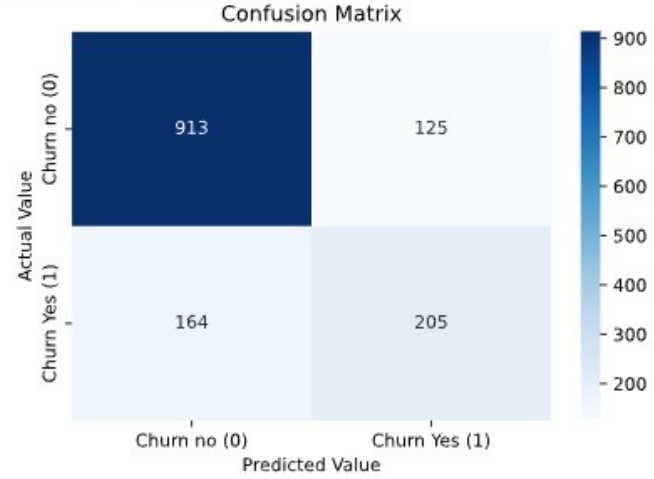

Figure 17: Test data evaluation of tuned KNN model.

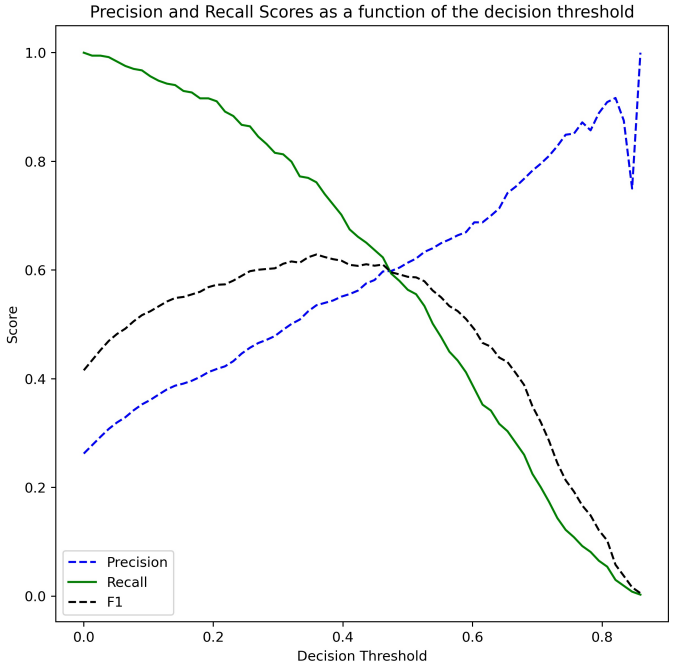

Figure 18: Precision and Recall scores as a function of the decision threshold for KNN model.

62% or 7% higher compared to the default model. The Recall score also significantly increases by 20% without any reduction on the Accuracy score, i.e 77%. These significant improvement comes with an AUC score of 83% (+5%) (Figure 19).

Test Data Evaluation :

	precision	recall	f1-score	support
0	0.89	0.78	0.83	1038
1	0.54	0.74	0.62	369
accuracy			0.77	1407
macro avg	0.72	0.76	0.73	1407
weighted avg	0.80	0.77	0.78	1407

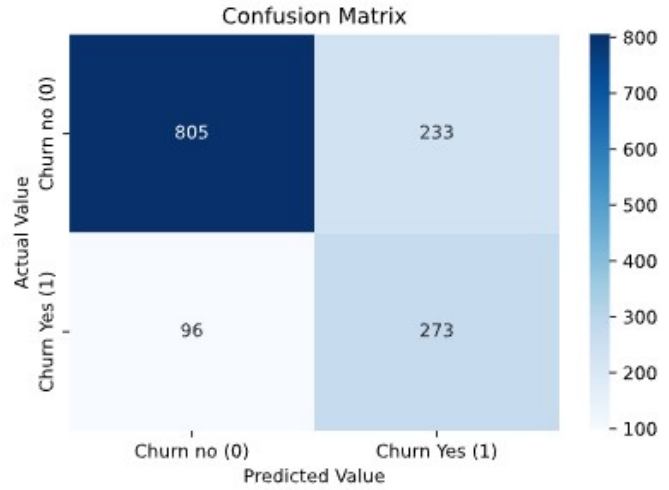

Figure 19: Test data evaluation of tuned KNN model with adjusted decision threshold.

4 Conclusion

From all the results above it can be concluded that :

1. Overall, one can see that the hyperparameter tuning does not always result in good improvement. Among the three models, KNN Classifier is the only model that significantly improves after the hyperparameter tuning was applied. This phenomenon is actually normal since KNN's performance heavily relies on the number of neighbours. Compared to the untuned model, ROC AUC, accuracy and F1 scores are increased by 5%, 2%, and 4% respectively to 83%, 79%, and 59%. Other than hyperparameter tuning, adjusting the decision threshold is able to effectively increase the Recall and F1 score for all models. Since the dominant category in this dataset is 0 (Not Churn), the maximum value of F1 score is inclined towards the Recall score, thus helps us to determine the best threshold (in this case, Recall is valued higher than Precision).

2. Logistic Regression Classifier is considered as the best model due to its performance. After being tuned with GridSearchCV method and adjusted to 0.3 decision/probability threshold, the improvement becomes significant compared

to the default model as the Recall and F1 score are increased to 76% (+23%) and 63% (+5%) respectively (Figure 14). With an AUC score of 83% (+0%), those metrics are successfully increased without a lot of reduction in accuracy, only 3% lower from the default model with the score of 77% . Even though the AUC score of KNN, RF, and Logistic Regression are almost identical, the combination of Recall, F1, and Accuracy owned by LogisticRegression Model is relatively better than KNN (Figure 19) and Random Forest (Figure 10).

3. Feature importance of the Logistic regression model can be seen below. Notice that the coefficients are both positive and negative. It can be elaborated as the predictor of Class 1 (Churn Yes) has positive coefficient whereas the predictor of Class 0 (Churn No) has negative coefficient. Overall, it is evident that the graph (Figure 20) is already in accordance to the result of EDA project carried out before on similar dataset [4]. Contract, tenure, InternetService, PaymentMethod and some additional internet services such as OnlineSecurity and Streaming are considered as the key features on which the business strategists should focus to improve the level of satisfaction and retent the customer.

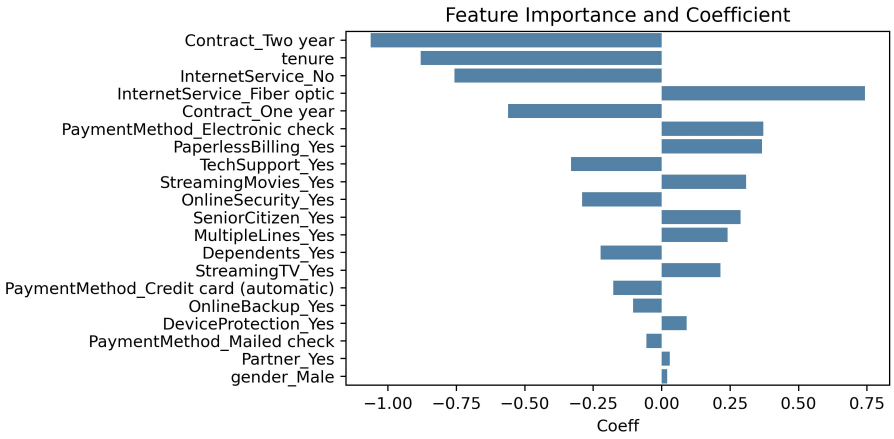

Figure 20: Coefficient of the LogisticRegression model.

References

[1] Ferd Reichheld. Prescription of cutting costs. http://www2.bain.com/Images/BB_Prescription_cutting_costs.pdf. (Accessed: 01-Dec-2021).

[2] IBM. Telco customer churn. https://www.kaggle.com/blastchar/telco-customer-churn. (Accessed: 01-Dec-2021).

[3] Indra Yanto. Telco customer churn predictive analysis. https://colab.research.google.com/drive/1_DIwM4A7kMZOEInNVh2GWwKaJ5IhoBFT?usp=sharing, . (Accessed: 22-Jan-2022).

[4] Indra Yanto. Telco customer churn exploratory data analysis. https://colab.research.google.com/drive/1iLpyBlI6tkHTbQYsEeiLglVIdbUq2onV?usp=sharing, . (Accessed: 01-Dec-2021).