Telco Customer Churn Prediction

Capstone Project

APRIL 29

Kalum De Alwis
Index Number: 265

Telco Customer Churn Prediction

Introduction

The telecom industry has traditionally experienced an average churn rate of 30% to 35%, while the cost per acquisition (CPA) typically ranges from five to 10 times higher than the cost of retaining an existing customer, according to the 2016 sector review Big Data Applications in the Telecommunications Industry (reference 1).

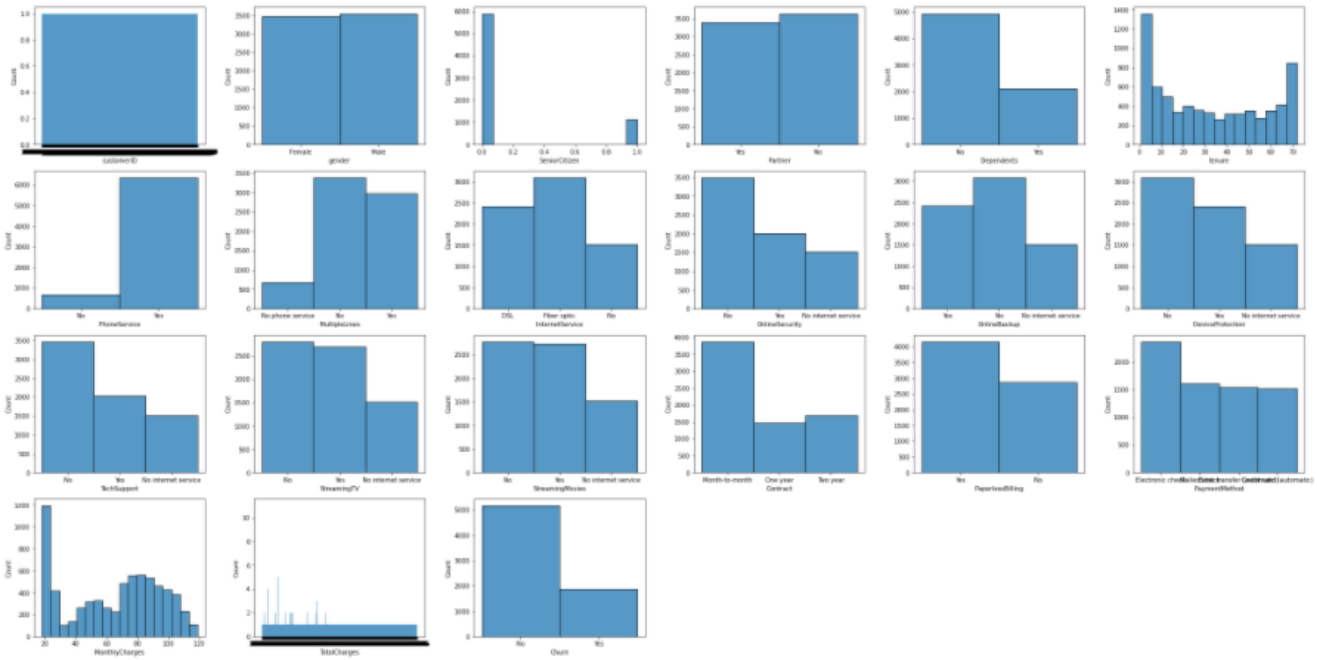
This Project tries to predict churn of customer in a telco environment using machine learning. Data set is WA_Fn-UseC_-Telco-Customer-Churn.csv which is sourced from https://www.kaggle.com/datasets/palashfendarkar/wa-fnusec-telcocustomerchurn.

Since this is a binary classification problem Random forest and logistic regression were targeted as possible ML algorithms.

Pre-Processing

Data set consisted of 21 columns and 7043 rows. Data set had only 3 numerical features and 17 categorical features, it did not contain null values. Distribution of data was checked with histogram plots (Figure 1 Histograms)

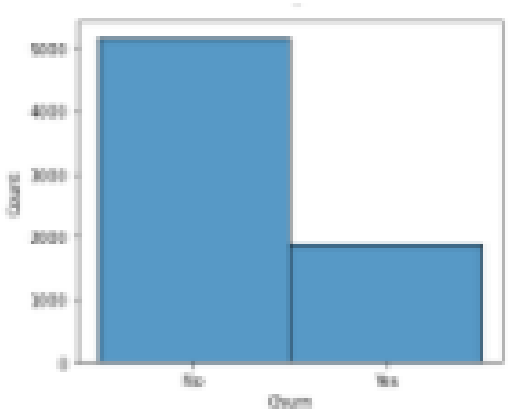
Figure 1 Histograms



Target variable which is ‘Churn’ had 1869 (26%) samples with ‘Yes” values and 5174 with ‘no’ values (Figure 2 Target Variable Bias).

Therefore, no special sampling was not done to the test data set considering the nature of telco churn.

Figure 2 Target Variable Bias



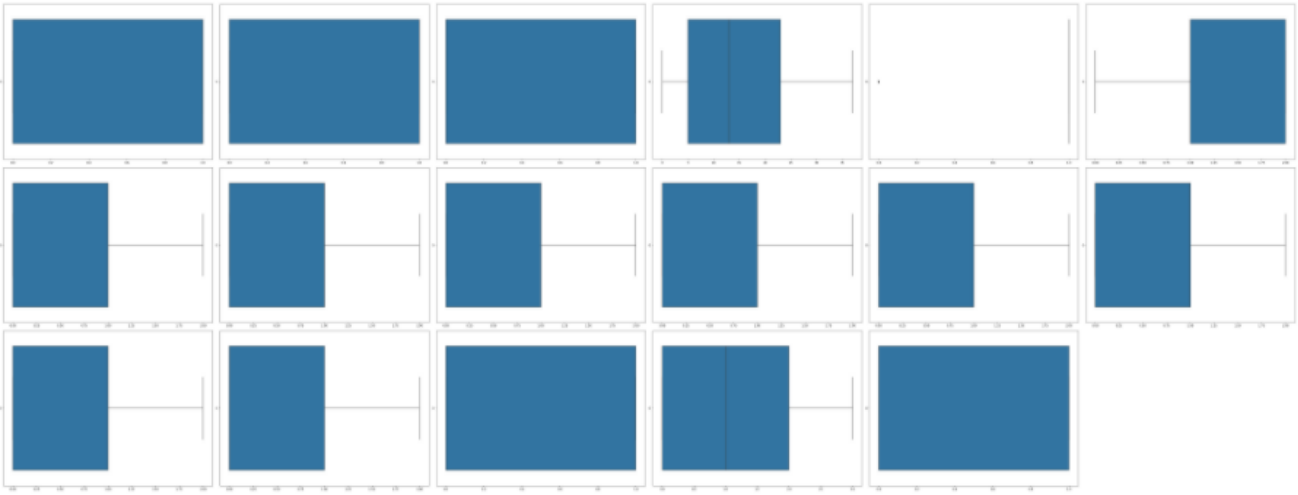
a

All categorical data were transformed to numerical data to support further processing.

The feature “TotalCharges” was store as object and needed to be converted to float. It also contained several samples with spaces and needed to be removed before converting to float.

Features ware checked for significant outliers and no outliers were present (Figure 3 Box Plots*).*

Figure 3 Box Plots



Machine Learning Model Building

Two Machine learning algorithms were evaluated for the `binary classification model.

1. Random forest

Target variable= “Churn”

Feature list= 'gender', 'Partner', 'Dependents','tenure', 'PhoneService', 'MultipleLines','InternetService', 'OnlineSecurity','OnlineBackup', 'DeviceProtection', 'TechSupport','StreamingTV','StreamingMovies','Contract','PaperlessBilling','PaymentMethod','MonthlyCharges','SeniorCitizen','TotalCharges'

Confusion Metrix

Y_Predicted	0	1	All
Y_Actual			
0	1549	6	1555
1	1	554	555
All	1550	560	2110

2. Logistic Regression

Target variable= “Churn”

Feature list ='gender', 'Partner', 'Dependents','tenure', 'PhoneService', 'MultipleLines', 'InternetService', 'OnlineSecurity', 'OnlineBackup', 'DeviceProtection', 'TechSupport','StreamingTV', 'StreamingMovies', 'Contract', 'PaperlessBilling','PaymentMethod','MonthlyCharges','SeniorCitizen','TotalCharges'
Confusion Metrix

Y_Predicted	0	1	All
Y_Actual			
0	1384	171	1555
1	273	282	555
All	1657	453	2110

Hyper parameter change

Logistic regression failed to converge with default hyper [parameters and had to set “max_iter=300” to solve the issue.

Machine Learning Model

Machine Learning Model evaluation

	model_name	model	accuracy	precision	f1_score	roc_auc	recall_val
0	RFModel1	(DecisionTreeClassifier(max_depth=100, max_fea...	0.792417	0.633257	0.784037	0.824879	0.464865
1	RFModel2	(DecisionTreeClassifier(max_depth=200, max_fea...	0.784834	0.619952	0.774507	0.821455	0.464865
2	RFModel3	(DecisionTreeClassifier(max_depth=50, max_feat...	0.782938	0.617433	0.771757	0.823859	0.464865
3	LogModel1	LogisticRegression(max_iter=300)	0.788152	0.615880	0.781865	0.832325	0.464865

Random forest algorithm-based model RFModel1 gave much better F1, recall, precision scores and they were close to .75.

Therefore, Random forest algorithm based RFModel1 was selected to deploy.

Feature importance was also identified to consider for user input data.

	Feture Name	Featureimportance
18	TotalCharges	0.204934
16	MonthlyCharges	0.169308
3	tenure	0.126523
13	Contract	0.087023
15	PaymentMethod	0.062009
7	OnlineSecurity	0.045677
10	TechSupport	0.042145
0	gender	0.028672
14	PaperlessBilling	0.027526
8	OnlineBackup	0.025776
5	MultipleLines	0.025115
6	InternetService	0.024132
1	Partner	0.022592
9	DeviceProtection	0.022515
17	SeniorCitizen	0.021340
2	Dependents	0.020371
11	StreamingTV	0.019944
12	StreamingMovies	0.019528
4	PhoneService	0.004870

Post Processing

Model was saved using joblib and functions were created to load the model and provide churn prediction for input data sets by user.

Conclusion

Random forest algorithm-based model RFModel1 gave much better F1, recall, precision scores around 0.75 and was selected as the best model to deploy.

Discussion

Random forest algorithm gave better estimation in this scenario and importance of hyperparameter tuning was highlighted when logistic regression failed to converge initially.

Also, must pay attention when numeral is stored as text in data sets. During data type conversion, cells with blank spaces can create errors as space does not get dropped with general null removal functions such as dropna. Must use query and remove before converting,

References

1. https://books.google.es/books?id=vzfVDQAAQBAJ&pg=PA92&lpg=PA92&dq=telecom+it+costs+5+to+10+more+times+to+acquire+a+customer+than+to+retain+it&source=bl&ots=U3nat9l0FU&sig=ACfU3U1ObhjBWUU-qLlY-UC7hujXHpf45Q&hl=en&sa=X&ved=2ahUKEwjxoOOOjbvqAhWtzYUKHZoACkwQ6AEwCnoECAgQAQ#v=onepage&q=telecom%20it%20costs%205%20to%2010%20more%20times%20to%20acquire%20a%20customer%20than%20to%20retain%20it&f=false