Customer Churn

22..08.2022

—

Roushan Raj
Oeson

Problem Statement

Customers' behavior is hard to predict these days, causing customers to move to some other beneficial organization. This can cause a huge loss and only retention is possible when the customer churns are evaluated and worked out at the right time.

Data Collection

This particular dataset, called ' Telco Customer Churn' is collected from Kaggle. This data is for 7043 customers having 21 different features.

Data Description

In this section, a little description and overview of data will be provided.

Fig: Overview of Data

Dataset statistics

Number of variables	22
Number of observations	7032
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	1.2 MiB
Average record size in memory	176.0 B

Variable types

Numeric	4
Categorical	13
Boolean	5

Here, we can see that stats of the dataset is provided which consists of num of variables, and their types.

Fig: Data Sample

Out[4]:

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	...	DeviceProtection
0	7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No	...	No
1	5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	...	Yes
2	3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	...	No
3	7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes	...	Yes
4	9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No	...	No
5	9305-CDSKC	Female	0	No	No	8	Yes	Yes	Fiber optic	No	...	Yes
6	1452-KIOVK	Male	0	No	Yes	22	Yes	Yes	Fiber optic	No	...	No

This is a sample of a dataset. Here we can see some of the attributes and columns it has.

Fig: Data Information

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7043 entries, 0 to 7042
Data columns (total 21 columns):
 #   Column            Non-Null Count  Dtype  
---  ------            --------------  -----  
 0   customerID        7043 non-null   object 
 1   gender            7043 non-null   object 
 2   SeniorCitizen     7043 non-null   int64  
 3   Partner           7043 non-null   object 
 4   Dependents        7043 non-null   object 
 5   tenure            7043 non-null   int64  
 6   PhoneService      7043 non-null   object 
 7   MultipleLines     7043 non-null   object 
 8   InternetService   7043 non-null   object 
 9   OnlineSecurity    7043 non-null   object 
 10  OnlineBackup      7043 non-null   object 
 11  DeviceProtection  7043 non-null   object 
 12  TechSupport       7043 non-null   object 
 13  StreamingTV       7043 non-null   object 
 14  StreamingMovies   7043 non-null   object 
 15  Contract          7043 non-null   object 
 16  PaperlessBilling  7043 non-null   object 
 17  PaymentMethod     7043 non-null   object 
 18  MonthlyCharges    7043 non-null   float64
 19  TotalCharges      7043 non-null   object 
 20  Churn             7043 non-null   object 
```

This figure shows the complete information of the dataset. Here, none of the columns or data is null. The count of data is also available.

Fig: Describing the data

Out[27]:

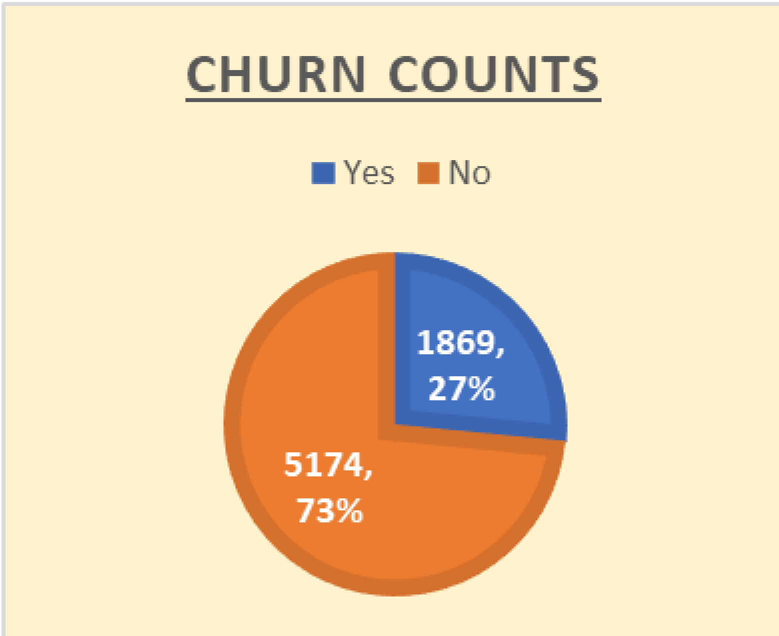
	SeniorCitizen	tenure	MonthlyCharges
count	7043.000000	7043.000000	7043.000000
mean	0.162147	32.371149	64.761692
std	0.368612	24.559481	30.090047
min	0.000000	0.000000	18.250000
25%	0.000000	9.000000	35.500000
50%	0.000000	29.000000	70.350000
75%	0.000000	55.000000	89.850000
max	1.000000	72.000000	118.750000

Here is a short description of the dataset where ‘senior citizen’, ‘tenure’ & ‘monthly charges’ are being taken under consideration. The count, mean, max etc are provided.

EDA (Exploratory Data Analysis)

Here, we will go through a series of visuals to get a brief idea about the data.

Fig: Counts of the Churn



This pie chart shows the number of the churns, where it is found that a total of 1869 customers moved out(churn, Yes) while 5174 were loyal (churn, No).

Fig: Correlation

Out[7]:

	SeniorCitizen	tenure	MonthlyCharges
SeniorCitizen	1.000000	0.016567	0.220173
tenure	0.016567	1.000000	0.247900
MonthlyCharges	0.220173	0.247900	1.000000

This is a correlation between the three variables.

Fig: F1 and Accuracy Score of Algorithms

Out[31]:

	Accuracy Score	f1 Score	AUC Score	ML Models
0	82.302772	61.633282	86.381038	GradientBoosting
1	82.302772	61.275272	85.926958	CatBoost
2	80.952381	59.880240	84.070938	XGBoost
3	81.307747	59.724349	85.174585	LightGBM
4	79.246624	53.943218	82.708832	RandomForest
5	73.276475	49.732620	65.929982	DecisionTree
6	75.124378	48.529412	66.276392	AdaBoost

These are some of the ML algorithms implemented which are showing the F1, Accuracy and AUC score. Here we can see that Accuracy of Gradient boosting is the most.

Fig: Confusion Matrix

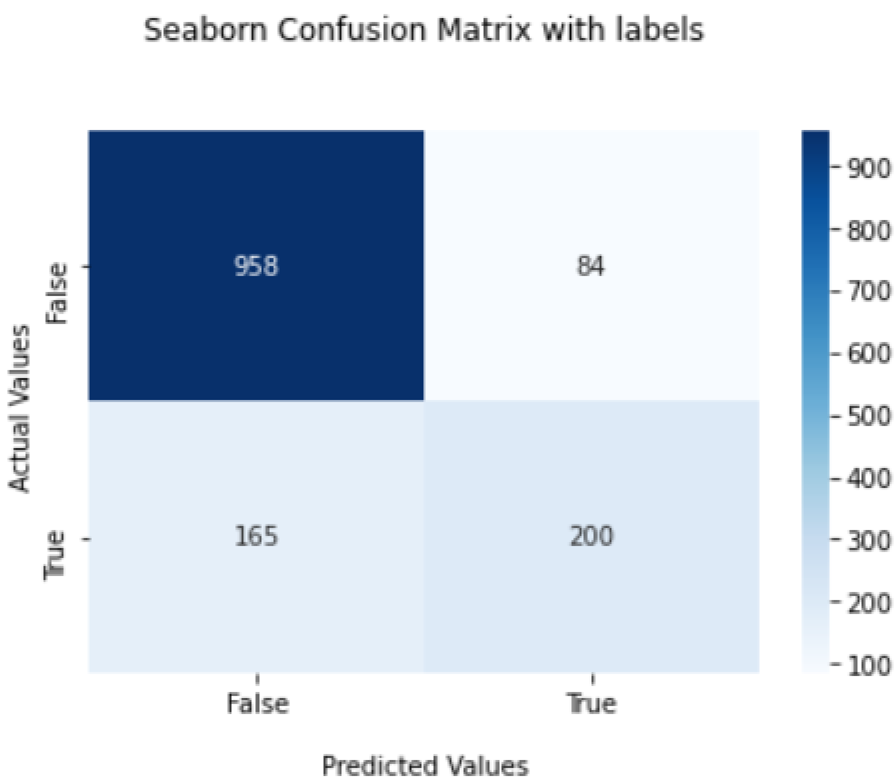
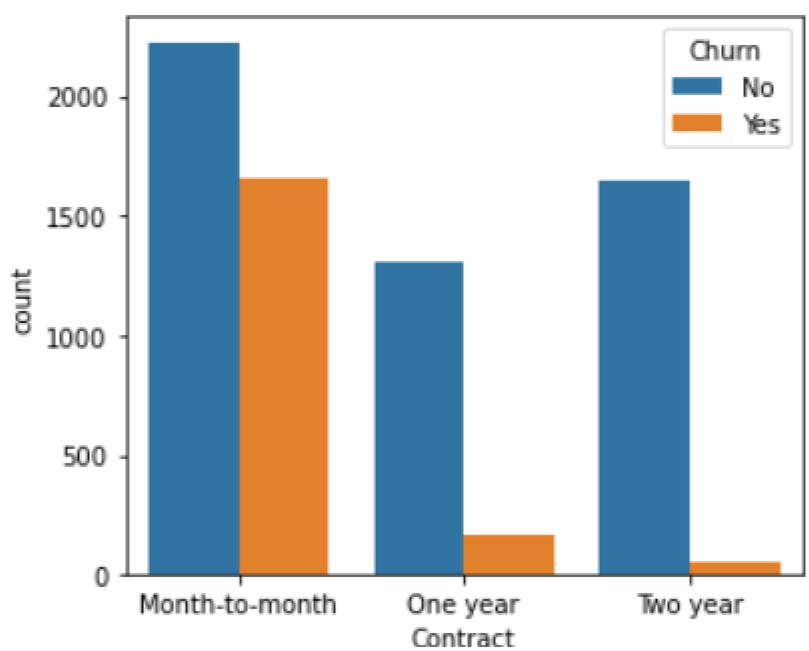


Fig: Churn vs Contract



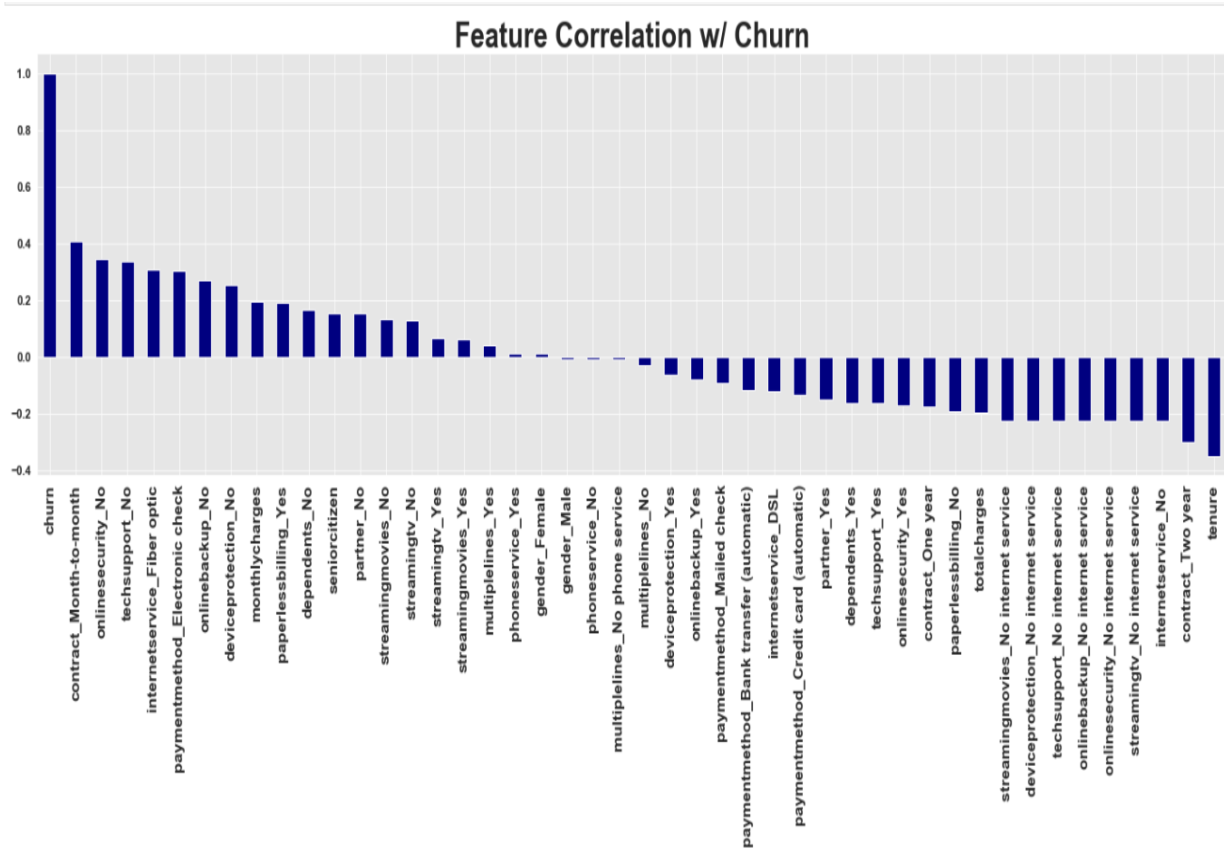
Here, it can be predicted that customers who have month-to-month contracts are more likely to churn. Churn probably decreases when contract length increases. Also it is seen that customers with online security churn the least.

Fig: Classification Report

	precision	recall	f1-score	support
0	0.85	0.92	0.88	1042
1	0.70	0.55	0.62	365
accuracy			0.82	1407
macro avg	0.78	0.73	0.75	1407
weighted avg	0.81	0.82	0.82	1407

This classification report gives a short idea for accuracy, ie. 82%. Her the F1 score is about 62%.

Fig: Feature Correlation



This is the correlation plots between all the variables of churn.

Fig: Descriptive Stat

Monthly charge		*Total charge*	
Mean	64.7666501	Mean	2286.909308
Standard Error	0.358561929	Standard Error	27.02065976
Median	70.35	Median	1400.425
Mode	20.05	Mode	20.2
Standard Deviation	30.08930702	Standard Deviation	2267.48258
Sample Variance	905.3663969	Sample Variance	5141477.252
Kurtosis	-1.256952501	Kurtosis	-0.239506028
Skewness	-0.220844567	Skewness	0.958124395
Range	100.5	Range	8666
Minimum	18.25	Minimum	18.8
Maximum	118.75	Maximum	8684.8
Sum	456086.75	Sum	16104415.35
Count	7042	Count	7042
Confidence Level(95.0%)	0.702889295	Confidence Level(95.0%)	52.96862538

Conclusion

- Customers with monthly contracts churn 20% more than those who have long term contracts.
- Tech-support and online security plays an important role for customer retention.
- When the monthly charge is more, customers churn more.