Customers on Telco

CIND-820

Name: Jiahui Peng

Supervisor: Tamer Abdou, PhD

Date of submission: 2022-12-05

Abstract

With the advancement of the technology, internet is becoming widespread. People is used to using internet in their daily life. Internet is not only used for checking the text information from the website, but also internet can be used for calling, streaming and e-gaming. Specially during COVID-19, outdoor activities reducing, the demand of internet is keeping growing up. And now, people could not leave without internet. In such this high demand market, more and more telecom companies are found, the competitions are getting intense. Therefore, the main questions for the telecom companies is customer churn instead of attracting new customers, and how to get the retention customers.

Literature Review

John et al. [1] focused on comparing several data with LTV (Life Time Value) by using 80/20 rule of total charges. There are 20% of top average LTV of leaked customers brought 60% of the revenue. 81% of high LTVs tend to used lines, and 75%-90% of high LTVs used Fiber optic. Almost 80% of Low LTVs of current customer used streaming TV and streaming movies which were two top subsets of internet service use.

Shuheng et al. [2] used Simple Regression Model, Logistic Regression Model and binomial probit regression Model to calculate the P-values of all attributes with churn. Shuheng also used Random Forest Model to find out the significant variable. They were Tenure, Total Charges, Monthly Charges, Contract, Internet Service.

Dataset

Telco customer churn dataset is from Kaggle.com which stems from the IBM company. This dataset had 7043 customers basic service information which including not only phone line and internet service, but also has second subsets service such as multiple phone lines, streaming TV and so on. We were going to using the machine learning to get the prediction model which can help the telecom companies much easier to know who are going to churn in the short future and also building up the retention program.

```
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
```

Here are the first 10 records of the dataset.

Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	...	DeviceProtection	TechSupport	StreamingTV	StreamingMovies	Contract	PaperlessBilling	PaymentMethod	MonthlyCharges	TotalCharges	Churn
Yes	No	1	No	No phone service	DSL	No	...	No	No	No	No	Month-to-month	Yes	Electronic check	29.85	29.85	No
No	No	34	Yes	No	DSL	Yes	...	Yes	No	No	No	One year	No	Mailed check	56.95	1889.5	No
No	No	2	Yes	No	DSL	Yes	...	No	No	No	No	Month-to-month	Yes	Mailed check	53.85	108.15	Yes
No	No	45	No	No phone service	DSL	Yes	...	Yes	Yes	No	No	One year	No	Bank transfer (automatic)	42.30	1840.75	No
No	No	2	Yes	No	Fiber optic	No	...	No	No	No	No	Month-to-month	Yes	Electronic check	70.70	151.65	Yes
No	No	8	Yes	Yes	Fiber optic	No	...	Yes	No	Yes	Yes	Month-to-month	Yes	Electronic check	99.65	820.5	Yes
No	Yes	22	Yes	Yes	Fiber optic	No	...	No	No	Yes	No	Month-to-month	Yes	Credit card (automatic)	89.10	1949.4	No
No	No	10	No	No phone service	DSL	Yes	...	No	No	No	No	Month-to-month	No	Mailed check	29.75	301.9	No
Yes	No	28	Yes	Yes	Fiber optic	No	...	Yes	Yes	Yes	Yes	Month-to-month	Yes	Electronic check	104.80	3046.05	Yes
No	Yes	62	Yes	No	DSL	Yes	...	No	No	No	No	One year	No	Bank transfer (automatic)	56.15	3487.95	No

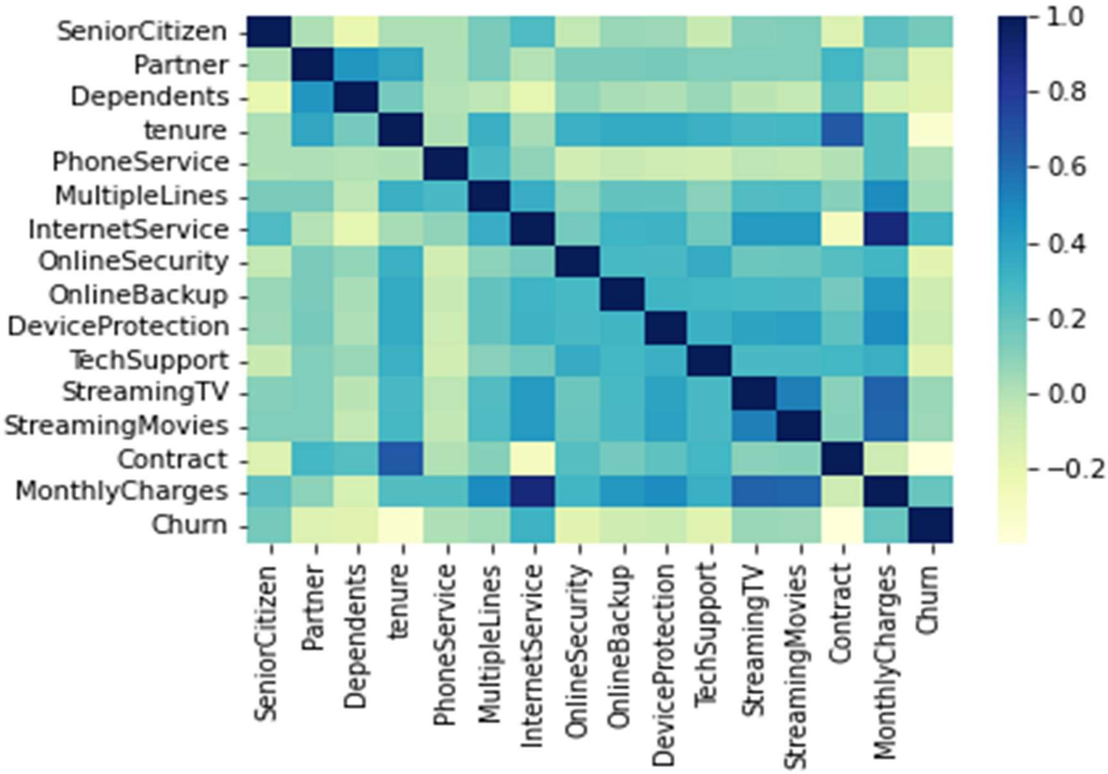
1. What is the main age area of customers? Are they the families' users or single users?

	SeniorCitizen	Partner	Dependents	tenure
count	7043.000000	7043.000000	7043.000000	7043.000000
mean	0.162147	0.483033	0.299588	32.371149
std	0.368612	0.499748	0.458110	24.559481
min	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	9.000000
50%	0.000000	0.000000	0.000000	29.000000
75%	0.000000	1.000000	1.000000	55.000000
max	1.000000	1.000000	1.000000	72.000000

(Senior Citizen:0=No, 1=Yes; Partner:0=No, 1=Yes; Dependents: 0=No, 1=Yes)

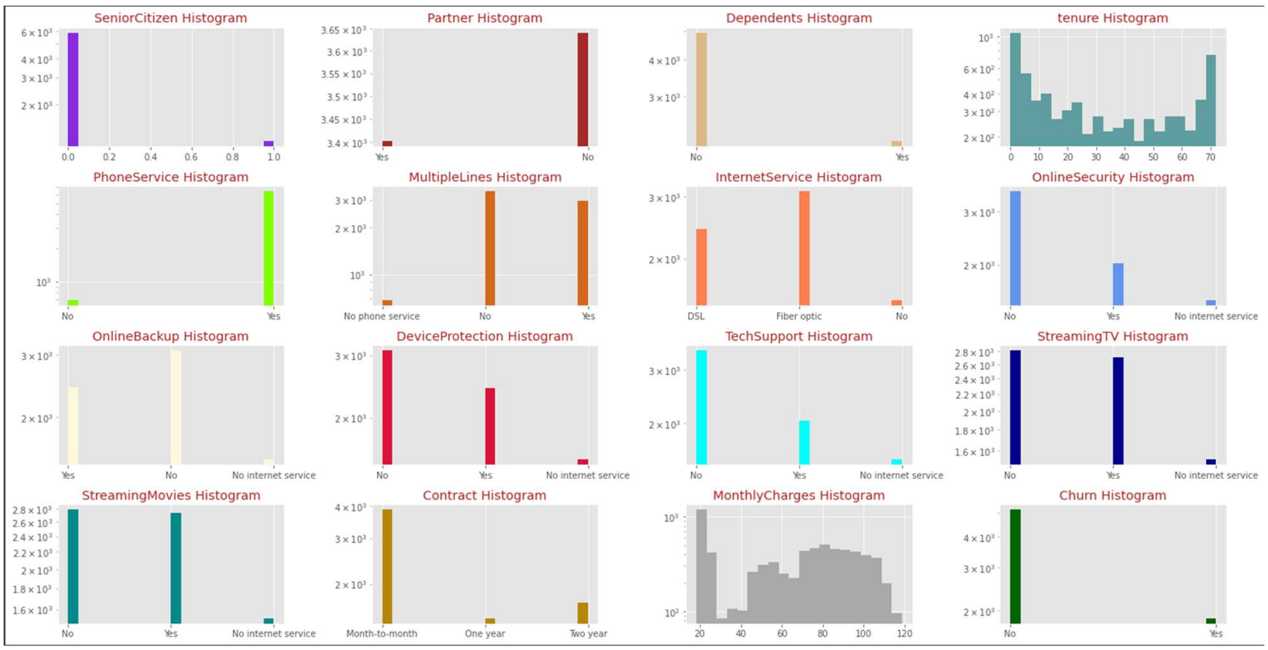
As we could see in the table, Senior customers were 16% of total customers, younger or mid-age customers were the main users, 84%. And 48% of them are the family's users and 29% of customers had dependents.

2. Which attributes seem to be correlated? Which attributes seem to be most linked to the Churn attribute?



According to the table, monthly charges and Internet service would be the most correlated. Internet service were less correlated with Contract. Tenure and Contract seems to be the most linked to the Churn attribute.

3. Graph the frequency distribution of all attributes.



In the frequency distribution of all attributes, we could know 90% customers had the phone services. Most of customers had the internet service, and more than half of them would like to have high-speed internet (fiber optic). Around 38% of customers had internet service and streaming TV, and 38% of customers had internet service and streaming Movies service. Most of customers would like to pay month to month, less customers were willing to accept one year's or two years' contracts. At last, there was 26.5% customers churn rate.

Data Analyses

Today we are going to find out which attributes are the key attributes of the churn. We don’t need any prediction so far. Therefore, we chose Logistic Regression algorithm and KNN algorithm for the data model.

Logistic Regression algorithm is based in classification algorithm as a Sigmoid function. It used the giving dataset to fine a linear relationship and calculate the probability.

KNN algorithm is a kind of classification algorithm as well. KNN stands for K-Nearest Neighbors.

For the better calculation, we used “0” and “1” instead of “No” and “Yes”.

```
1 df.head(10)
```

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	...	DeviceProtection	TechSupport	StreamingTV	StreamingMovies	Contract
0	7590-VHVEG	Female	0	1	0	1	0	0	1	0	...	0	0	0	0	0
1	5575-GNVDE	Male	0	0	0	34	1	0	1	1	...	1	0	0	0	1
2	3668-QPYBK	Male	0	0	0	2	1	0	1	1	...	0	0	0	0	0
3	7795-CFOCW	Male	0	0	0	45	0	0	1	1	...	1	1	0	0	1
4	9237-HQITU	Female	0	0	0	2	1	0	2	0	...	0	0	0	0	0
5	9305-CDSKC	Female	0	0	0	8	1	1	2	0	...	1	0	1	1	0
6	1452-KIOVK	Male	0	0	1	22	1	1	2	0	...	0	0	1	0	0
7	6713-OKOMC	Female	0	0	0	10	0	0	1	1	...	0	0	0	0	0
8	7892-POOKP	Female	0	1	0	28	1	1	2	0	...	1	1	1	1	0
9	6388-TABGU	Male	0	0	1	62	1	0	1	1	...	0	0	0	0	1

10 rows × 21 columns

Used the Logistic Regression algorithm to predict the churn of customer using its attributes (70% of records as training data, 30% as testing data).

```
Optimization terminated successfully.
         Current function value: 0.432641
         Iterations 7
```

Logit Regression Results

Dep. Variable:	Churn	**No. Observations:**	4930
Model:	Logit	**Df Residuals:**	4915
Method:	MLE	**Df Model:**	14
Date:	Tue, 27 Sep 2022	**Pseudo R-squ.:**	0.2608
Time:	01:51:47	**Log-Likelihood:**	-2132.9
converged:	True	**LL-Null:**	-2885.3
Covariance Type:	nonrobust	**LLR p-value:**	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
SeniorCitizen	0.2584	0.100	2.583	0.010	0.062	0.454
Partner	0.0067	0.091	0.073	0.942	-0.172	0.185
Dependents	-0.1748	0.104	-1.681	0.093	-0.379	0.029
tenure	-0.0370	0.003	-13.861	0.000	-0.042	-0.032
PhoneService	-0.0679	0.775	-0.088	0.930	-1.587	1.451
MultipleLines	0.7532	0.213	3.537	0.000	0.336	1.171
InternetService	2.0405	0.964	2.117	0.034	0.151	3.930
OnlineSecurity	-0.1898	0.215	-0.883	0.377	-0.611	0.231
OnlineBackup	0.1328	0.213	0.624	0.533	-0.284	0.550
DeviceProtection	0.1502	0.213	0.706	0.480	-0.267	0.567
TechSupport	-0.1794	0.215	-0.833	0.405	-0.601	0.243
StreamingTV	0.7887	0.394	2.004	0.045	0.017	1.560
StreamingMovies	0.8525	0.396	2.152	0.031	0.076	1.629
Contract	-0.7160	0.087	-8.216	0.000	-0.887	-0.545
MonthlyCharges	-0.0494	0.038	-1.284	0.199	-0.125	0.026

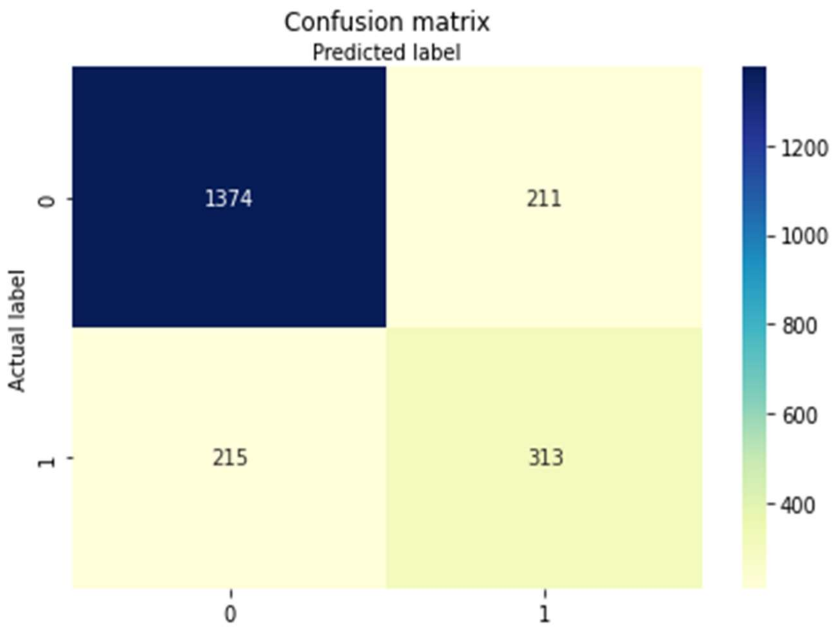
From the result, we could see the p-value of the attributes: SeniorCitizen, tenure, MultipleLines, InternetService, StreamingTV and StreamingMovies were smaller than 0.05. It means they were significant or less significant to churn.

Then we could get the simple calculation as:

Churn=0.2584* SeniorCitizen-0.037*tenure+0.7532* MultipleLines+2.0405* InternetService+0.7887* StreamingTV+0.8525* StreamingMovies-0.7160*Contract

Therefore, if the customer was a senior, had multiple lines, had internet service, had streaming TV and Movies was more probability to churn. If the customer was not a senior, had tenure, no internet service, had a contract was more chances to stay

Evaluate the model performance by computing Accuracy, Sensitivity, and Specificity.



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

Accuracy is the percentage of correct prediction from the model calculation.

$$Sensitivity = \frac{TP}{TP + FN}$$

Sensitivity is the percentage of correct positive prediction from all positive values

$$Specificity = \frac{TN}{TN + FP}$$

Specificity is the percentage of correct negative prediction from all negative values

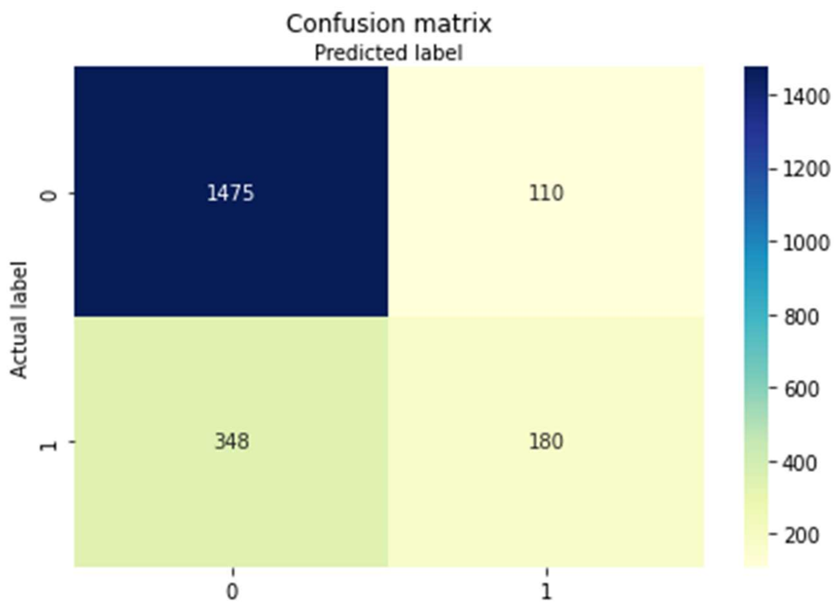
```
The Accuracy is 0.7983909133932797
The Sensitivity is 0.8646947765890497
The Specificity is 0.5973282442748091
```

Use the KNN algorithm to predict the churn of customer using its attributes;

evaluate the model performance by computing Accuracy, Sensitivity, and Specificity.

```
from sklearn.neighbors import KNeighborsClassifier
knn = KNeighborsClassifier(n_neighbors=2)
knn.fit(X_train, y_train)
knn.predict(X_test)
```

```
array(['0', '0', '0', ..., '0', '0', '0'], dtype=object)
```




```
The Accuracy is 0.7832465688594415
The Sensitivity is 0.8091058694459682
The Specificity is 0.6206896551724138
```

Conclusion and Recommendations

In conclusion, we could know about younger or mid-age customers were the main group users, half of them were not a single. We could assume around 4 devices connected to internet in the same time by each customer. Therefore, half of their internet service would more than 4 devices in the same time, and also they would use internet for streaming TV or movies. It means they highly required the stable and high-speed internet. Mean while we know Internet service would be the most correlated with monthly charges. And the customers churn was correlated with contract. In order to reducing the customers churn rate, company should attract more customers who paid month to month into contract customers.

Therefore, the telco company could provide a promotion as a new faster high-speed internet package with the 1 year’s or 2 years’ contract, and the price would be higher than the current package. Also, the company could provide the same current speed internet service with the 1 year’s or 2 years’ contract as a lower price.

Data From:

https://www.kaggle.com/datasets/blastchar/telco-customer-churn?select=WA_Fn-UseC_-Telco-Customer-Churn.csv

Base information from:

https://community.ibm.com/community/user/businessanalytics/blogs/steven-macko/2019/07/11/telco-customer-churn-1113

GitHub account:

GARYPENGSS/project

https://github.com/GARYPENGSS/project

Reference

CHEN, JOHN YUEH-HAN. Towards Data Science, 2020, *Data Analysis Project — Telco Customer Churn*, https://towardsdatascience.com/data-analysis-project-telco-customer-churn-fe5c0144e708. Accessed 2022 Oct. 3n.d..

Ma, Shuheng. Towards Data Science, 2021, *Telco Customer ChurnRate Analysis*, https://towardsdatascience.com/telco-customer-churnrate-analysis-d412f208cbbf. Accessed 30 Oct. 2022.