Telecom_Churn_Final

October 9, 2024

0.1 Telecom Churn Case Study

With 21 predictor variables we need to predict whether a particular customer will switch to another telecom provider or not. In telecom terminology, this is referred to as churning and not churning, respectively.

[205]:
```
# Suppressing Warnings
import warnings
warnings.filterwarnings('ignore')
```

0.1.1 Step 1: Importing and Merging Data

[206]:
```
# Importing Pandas and NumPy
import pandas as pd, numpy as np
```

[207]:
```
# Importing all datasets
churn_data = pd.read_csv("churn_data.csv")
churn_data.head()
```

[207]:

	customerID	tenure	PhoneService	Contract	PaperlessBilling	PaymentMethod	MonthlyCharges	TotalCharges	Churn
0	7590-VHVEG	1	No	Month-to-month	Yes	Electronic check	29.85	29.85	No
1	5575-GNVDE	34	Yes	One year	No	Mailed check	56.95	1889.5	No
2	3668-QPYBK	2	Yes	Month-to-month	Yes	Mailed check	53.85	108.15	Yes
3	7795-CFOCW	45	No	One year	No	Bank transfer (automatic)	42.30	1840.75	No
4	9237-HQITU	2	Yes	Month-to-month	Yes	Electronic check	70.70	151.65	Yes

[208]:
```
customer_data = pd.read_csv("customer_data.csv")
customer_data.head()
```

[208]:

	customerID	gender	SeniorCitizen	Partner	Dependents
0	7590-VHVEG	Female	0	Yes	No

```
1  5575-GNVDE    Male              0      No         No
2  3668-QPYBK    Male              0      No         No
3  7795-CFOCW    Male              0      No         No
4  9237-HQITU  Female              0      No         No
```

[209]:
```
internet_data = pd.read_csv("internet_data.csv")
internet_data.head()
```

[209]:
```
   customerID     MultipleLines InternetService OnlineSecurity OnlineBackup  \
0  7590-VHVEG  No phone service             DSL             No          Yes
1  5575-GNVDE                No             DSL            Yes           No
2  3668-QPYBK                No             DSL            Yes          Yes
3  7795-CFOCW  No phone service             DSL            Yes           No
4  9237-HQITU                No     Fiber optic             No           No

  DeviceProtection TechSupport StreamingTV StreamingMovies
0               No          No          No              No
1              Yes          No          No              No
2               No          No          No              No
3              Yes         Yes          No              No
4               No          No          No              No
```

Combining all data files into one consolidated dataframe

[210]:
```
# Merging on 'customerID'
df_1 = pd.merge(churn_data, customer_data, how='inner', on='customerID')
```

[211]:
```
# Final dataframe with all predictor variables
telecom = pd.merge(df_1, internet_data, how='inner', on='customerID')
```

0.1.2 Step 2: Inspecting the Dataframe

[212]:
```
# Let's see the head of our master dataset
telecom.head()
```

[212]:
```
   customerID  tenure PhoneService        Contract PaperlessBilling  \
0  7590-VHVEG       1           No  Month-to-month              Yes
1  5575-GNVDE      34          Yes        One year               No
2  3668-QPYBK       2          Yes  Month-to-month              Yes
3  7795-CFOCW      45           No        One year               No
4  9237-HQITU       2          Yes  Month-to-month              Yes

                PaymentMethod  MonthlyCharges TotalCharges Churn  gender  … \
0           Electronic check           29.85        29.85    No  Female  …
1               Mailed check           56.95       1889.5    No    Male  …
2               Mailed check           53.85       108.15   Yes    Male  …
3  Bank transfer (automatic)           42.30      1840.75    No    Male  …
```

```
4           Electronic check          70.70       151.65   Yes  Female  …

   Partner Dependents     MultipleLines InternetService OnlineSecurity  \
0      Yes         No  No phone service             DSL             No
1       No         No                No             DSL            Yes
2       No         No                No             DSL            Yes
3       No         No  No phone service             DSL            Yes
4       No         No                No     Fiber optic             No

  OnlineBackup DeviceProtection TechSupport StreamingTV StreamingMovies
0          Yes               No          No          No              No
1           No              Yes          No          No              No
2          Yes               No          No          No              No
3           No              Yes         Yes          No              No
4           No               No          No          No              No

[5 rows x 21 columns]
```

[213]:
```
# Let's check the dimensions of the dataframe
telecom.shape
```

[213]:
```
(7043, 21)
```

[214]:
```
# let's look at the statistical aspects of the dataframe
telecom.describe()
```

[214]:
```
             tenure  MonthlyCharges  SeniorCitizen
count  7043.000000     7043.000000    7043.000000
mean     32.371149       64.761692       0.162147
std      24.559481       30.090047       0.368612
min       0.000000       18.250000       0.000000
25%       9.000000       35.500000       0.000000
50%      29.000000       70.350000       0.000000
75%      55.000000       89.850000       0.000000
max      72.000000      118.750000       1.000000
```

[215]:
```
# Let's see the type of each column
telecom.info()
```

```
<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   tenure            7043 non-null   int64
 2   PhoneService      7043 non-null   object
```

```
 3   Contract          7043 non-null   object
 4   PaperlessBilling  7043 non-null   object
 5   PaymentMethod     7043 non-null   object
 6   MonthlyCharges    7043 non-null   float64
 7   TotalCharges      7043 non-null   object
 8   Churn             7043 non-null   object
 9   gender            7043 non-null   object
 10  SeniorCitizen     7043 non-null   int64
 11  Partner           7043 non-null   object
 12  Dependents        7043 non-null   object
 13  MultipleLines     7043 non-null   object
 14  InternetService   7043 non-null   object
 15  OnlineSecurity    7043 non-null   object
 16  OnlineBackup      7043 non-null   object
 17  DeviceProtection  7043 non-null   object
 18  TechSupport       7043 non-null   object
 19  StreamingTV       7043 non-null   object
 20  StreamingMovies   7043 non-null   object
dtypes: float64(1), int64(2), object(18)
memory usage: 1.1+ MB
```

0.1.3 Step 3: Data Preparation

Converting some binary variables (Yes/No) to 0/1

[216]:
```
# List of variables to map

varlist =  ['PhoneService', 'PaperlessBilling', 'Churn', 'Partner',␣
 ↪'Dependents']

# Defining the map function
def binary_map(x):
    return x.map({'Yes': 1, "No": 0})

# Applying the function to the telecom list
telecom[varlist] = telecom[varlist].apply(binary_map)
```

[217]:
```
telecom.head()
```

[217]:

	customerID	tenure	PhoneService	Contract	PaperlessBilling	\
0	7590-VHVEG	1	0	Month-to-month	1	
1	5575-GNVDE	34	1	One year	0	
2	3668-QPYBK	2	1	Month-to-month	1	
3	7795-CFOCW	45	0	One year	0	
4	9237-HQITU	2	1	Month-to-month	1	

	PaymentMethod	MonthlyCharges	TotalCharges	Churn	gender	…	\
0	Electronic check	29.85	29.85	0	Female	…	

```
1              Mailed check          56.95       1889.5      0    Male  …
2              Mailed check          53.85       108.15      1    Male  …
3  Bank transfer (automatic)          42.30      1840.75      0    Male  …
4           Electronic check          70.70       151.65      1  Female  …

   Partner  Dependents     MultipleLines InternetService OnlineSecurity  \
0        1           0  No phone service             DSL             No
1        0           0                No             DSL            Yes
2        0           0                No             DSL            Yes
3        0           0  No phone service             DSL            Yes
4        0           0                No     Fiber optic             No

  OnlineBackup DeviceProtection TechSupport StreamingTV StreamingMovies
0          Yes               No          No          No              No
1           No              Yes          No          No              No
2          Yes               No          No          No              No
3           No              Yes         Yes          No              No
4           No               No          No          No              No

[5 rows x 21 columns]
```

For categorical variables with multiple levels, create dummy features (one-hot encoded)

[218]:
```
# Creating a dummy variable for some of the categorical variables and dropping␣
 ↪the first one.
dummy1 = pd.get_dummies(telecom[['Contract', 'PaymentMethod', 'gender',␣
 ↪'InternetService']], drop_first=True)

# Adding the results to the master dataframe
telecom = pd.concat([telecom, dummy1], axis=1)
```

[219]:
```
telecom.head()
```

[219]:
```
    customerID  tenure  PhoneService        Contract  PaperlessBilling  \
0  7590-VHVEG       1             0  Month-to-month                 1
1  5575-GNVDE      34             1        One year                 0
2  3668-QPYBK       2             1  Month-to-month                 1
3  7795-CFOCW      45             0        One year                 0
4  9237-HQITU       2             1  Month-to-month                 1

                PaymentMethod  MonthlyCharges TotalCharges  Churn  gender  …  \
0           Electronic check           29.85        29.85      0  Female  …
1               Mailed check           56.95       1889.5      0    Male  …
2               Mailed check           53.85       108.15      1    Male  …
3  Bank transfer (automatic)           42.30      1840.75      0    Male  …
4           Electronic check           70.70       151.65      1  Female  …
```

```
   StreamingTV  StreamingMovies  Contract_One year Contract_Two year  \
0           No               No              False             False
1           No               No               True             False
2           No               No              False             False
3           No               No               True             False
4           No               No              False             False

  PaymentMethod_Credit card (automatic) PaymentMethod_Electronic check  \
0                                 False                           True
1                                 False                          False
2                                 False                          False
3                                 False                          False
4                                 False                           True

  PaymentMethod_Mailed check gender_Male InternetService_Fiber optic  \
0                      False       False                        False
1                       True        True                        False
2                       True        True                        False
3                      False        True                        False
4                      False       False                         True

  InternetService_No
0              False
1              False
2              False
3              False
4              False

[5 rows x 29 columns]
```

[220]:
```
telecom.MultipleLines.value_counts()
```

[220]:
```
MultipleLines
No                  3390
Yes                 2971
No phone service     682
Name: count, dtype: int64
```

1 Creating dummy variables for the remaining categorical variables and dropping the level with big names.

[221]:
```
# Creating dummy variables for the variable 'MultipleLines'
ml = pd.get_dummies(telecom['MultipleLines'], prefix='MultipleLines')
# Dropping MultipleLines_No phone service column
ml1 = ml.drop(['MultipleLines_No phone service'], axis=1)
```

```
#Adding the results to the master dataframe
telecom = pd.concat([telecom,ml1], axis=1)

# Creating dummy variables for the variable 'OnlineSecurity'.
os = pd.get_dummies(telecom['OnlineSecurity'], prefix='OnlineSecurity')
os1 = os.drop(['OnlineSecurity_No internet service'], axis = 1)
# Adding the results to the master dataframe
telecom = pd.concat([telecom,os1], axis=1)

# Creating dummy variables for the variable 'OnlineBackup'.
ob = pd.get_dummies(telecom['OnlineBackup'], prefix='OnlineBackup')
ob1 = ob.drop(['OnlineBackup_No internet service'], axis = 1)
# Adding the results to the master dataframe
telecom = pd.concat([telecom,ob1], axis=1)

# Creating dummy variables for the variable 'DeviceProtection'.
dp = pd.get_dummies(telecom['DeviceProtection'], prefix='DeviceProtection')
dp1 = dp.drop(['DeviceProtection_No internet service'], axis = 1)
# Adding the results to the master dataframe
telecom = pd.concat([telecom,dp1], axis=1)

# Creating dummy variables for the variable 'TechSupport'.
ts = pd.get_dummies(telecom['TechSupport'], prefix='TechSupport')
ts1 = ts.drop(['TechSupport_No internet service'], axis =1)
# Adding the results to the master dataframe
telecom = pd.concat([telecom,ts1], axis=1)

# Creating dummy variables for the variable 'StreamingTV'.
st =pd.get_dummies(telecom['StreamingTV'], prefix='StreamingTV')
st1 = st.drop(['StreamingTV_No internet service'], axis =1)
# Adding the results to the master dataframe
telecom = pd.concat([telecom,st1], axis=1)

# Creating dummy variables for the variable 'StreamingMovies'.
sm = pd.get_dummies(telecom['StreamingMovies'], prefix='StreamingMovies')
sm1 = sm.drop(['StreamingMovies_No internet service'], axis =1)
# Adding the results to the master dataframe
telecom = pd.concat([telecom,sm1], axis=1)
```

[222]:
```
telecom.head()
```

[222]:

	customerID	tenure	PhoneService	Contract	PaperlessBilling	\
0	7590-VHVEG	1	0	Month-to-month	1	
1	5575-GNVDE	34	1	One year	0	
2	3668-QPYBK	2	1	Month-to-month	1	
3	7795-CFOCW	45	0	One year	0	
4	9237-HQITU	2	1	Month-to-month	1	

```
              PaymentMethod  MonthlyCharges TotalCharges  Churn  gender  …  \
0          Electronic check           29.85        29.85      0  Female  …
1              Mailed check           56.95       1889.5      0    Male  …
2              Mailed check           53.85       108.15      1    Male  …
3  Bank transfer (automatic)           42.30      1840.75      0    Male  …
4          Electronic check           70.70       151.65      1  Female  …

   OnlineBackup_No  OnlineBackup_Yes  DeviceProtection_No  \
0            False              True                 True
1             True             False                False
2            False              True                 True
3             True             False                False
4             True             False                 True

  DeviceProtection_Yes TechSupport_No TechSupport_Yes StreamingTV_No  \
0                False           True           False           True
1                 True           True           False           True
2                False           True           False           True
3                 True          False            True           True
4                False           True           False           True

  StreamingTV_Yes StreamingMovies_No StreamingMovies_Yes
0           False               True               False
1           False               True               False
2           False               True               False
3           False               True               False
4           False               True               False

[5 rows x 43 columns]
```

Dropping the repeated variables

[223]:
```
# We have created dummies for the below variables, so we can drop them
telecom = telecom.
 ↪drop(['Contract','PaymentMethod','gender','MultipleLines','InternetService',␣
 ↪'OnlineSecurity', 'OnlineBackup', 'DeviceProtection',
       'TechSupport', 'StreamingTV', 'StreamingMovies'], axis = 1)
```

[224]:
```
#The varaible was imported as a string we need to convert it to float
telecom['TotalCharges'] = pd.to_numeric(telecom['TotalCharges'],␣
 ↪errors='coerce')
#telecom['TotalCharges'] = telecom['TotalCharges'].
 ↪convert_objects(convert_numeric=True)
```

[225]:
```
telecom.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7043 entries, 0 to 7042
Data columns (total 32 columns):
 #   Column                                   Non-Null Count  Dtype
---  ------                                   --------------  -----
 0   customerID                               7043 non-null   object
 1   tenure                                   7043 non-null   int64
 2   PhoneService                             7043 non-null   int64
 3   PaperlessBilling                         7043 non-null   int64
 4   MonthlyCharges                           7043 non-null   float64
 5   TotalCharges                             7032 non-null   float64
 6   Churn                                    7043 non-null   int64
 7   SeniorCitizen                            7043 non-null   int64
 8   Partner                                  7043 non-null   int64
 9   Dependents                               7043 non-null   int64
 10  Contract_One year                        7043 non-null   bool
 11  Contract_Two year                        7043 non-null   bool
 12  PaymentMethod_Credit card (automatic)    7043 non-null   bool
 13  PaymentMethod_Electronic check           7043 non-null   bool
 14  PaymentMethod_Mailed check               7043 non-null   bool
 15  gender_Male                              7043 non-null   bool
 16  InternetService_Fiber optic              7043 non-null   bool
 17  InternetService_No                       7043 non-null   bool
 18  MultipleLines_No                         7043 non-null   bool
 19  MultipleLines_Yes                        7043 non-null   bool
 20  OnlineSecurity_No                        7043 non-null   bool
 21  OnlineSecurity_Yes                       7043 non-null   bool
 22  OnlineBackup_No                          7043 non-null   bool
 23  OnlineBackup_Yes                         7043 non-null   bool
 24  DeviceProtection_No                      7043 non-null   bool
 25  DeviceProtection_Yes                     7043 non-null   bool
 26  TechSupport_No                           7043 non-null   bool
 27  TechSupport_Yes                          7043 non-null   bool
 28  StreamingTV_No                           7043 non-null   bool
 29  StreamingTV_Yes                          7043 non-null   bool
 30  StreamingMovies_No                       7043 non-null   bool
 31  StreamingMovies_Yes                      7043 non-null   bool
dtypes: bool(22), float64(2), int64(7), object(1)
memory usage: 701.7+ KB
```

Checking for Outliers

[226]:
```
# Checking for outliers in the continuous variables
num_telecom =␣
  ↪telecom[['tenure','MonthlyCharges','SeniorCitizen','TotalCharges']]
```

[227]:
```
# Checking outliers at 25%, 50%, 75%, 90%, 95% and 99%
num_telecom.describe(percentiles=[.25, .5, .75, .90, .95, .99])
```

[227]:

	tenure	MonthlyCharges	SeniorCitizen	TotalCharges
count	7043.000000	7043.000000	7043.000000	7032.000000
mean	32.371149	64.761692	0.162147	2283.300441
std	24.559481	30.090047	0.368612	2266.771362
min	0.000000	18.250000	0.000000	18.800000
25%	9.000000	35.500000	0.000000	401.450000
50%	29.000000	70.350000	0.000000	1397.475000
75%	55.000000	89.850000	0.000000	3794.737500
90%	69.000000	102.600000	1.000000	5976.640000
95%	72.000000	107.400000	1.000000	6923.590000
99%	72.000000	114.729000	1.000000	8039.883000
max	72.000000	118.750000	1.000000	8684.800000

Checking for Missing Values and Inputing Them

[228]:
```
# Adding up the missing values (column-wise)
telecom.isnull().sum()
```

[228]:

customerID	0
tenure	0
PhoneService	0
PaperlessBilling	0
MonthlyCharges	0
TotalCharges	11
Churn	0
SeniorCitizen	0
Partner	0
Dependents	0
Contract_One year	0
Contract_Two year	0
PaymentMethod_Credit card (automatic)	0
PaymentMethod_Electronic check	0
PaymentMethod_Mailed check	0
gender_Male	0
InternetService_Fiber optic	0
InternetService_No	0
MultipleLines_No	0
MultipleLines_Yes	0
OnlineSecurity_No	0
OnlineSecurity_Yes	0
OnlineBackup_No	0
OnlineBackup_Yes	0
DeviceProtection_No	0
DeviceProtection_Yes	0
TechSupport_No	0
TechSupport_Yes	0
StreamingTV_No	0

```
StreamingTV_Yes                          0
StreamingMovies_No                       0
StreamingMovies_Yes                      0
dtype: int64
```

It means that $11/7043 = 0.001561834$ i.e 0.1%, best is to remove these observations from the analysis

[229]:
```
# Checking the percentage of missing values
round(100*(telecom.isnull().sum()/len(telecom.index)), 2)
```

[229]:
```
customerID                            0.00
tenure                                0.00
PhoneService                          0.00
PaperlessBilling                      0.00
MonthlyCharges                        0.00
TotalCharges                          0.16
Churn                                 0.00
SeniorCitizen                         0.00
Partner                               0.00
Dependents                            0.00
Contract_One year                     0.00
Contract_Two year                     0.00
PaymentMethod_Credit card (automatic) 0.00
PaymentMethod_Electronic check        0.00
PaymentMethod_Mailed check            0.00
gender_Male                           0.00
InternetService_Fiber optic           0.00
InternetService_No                    0.00
MultipleLines_No                      0.00
MultipleLines_Yes                     0.00
OnlineSecurity_No                     0.00
OnlineSecurity_Yes                    0.00
OnlineBackup_No                       0.00
OnlineBackup_Yes                      0.00
DeviceProtection_No                   0.00
DeviceProtection_Yes                  0.00
TechSupport_No                        0.00
TechSupport_Yes                       0.00
StreamingTV_No                        0.00
StreamingTV_Yes                       0.00
StreamingMovies_No                    0.00
StreamingMovies_Yes                   0.00
dtype: float64
```

[230]:
```
# Removing NaN TotalCharges rows
telecom = telecom[~np.isnan(telecom['TotalCharges'])]
```

[231]:
```
# Checking percentage of missing values after removing the missing values
round(100*(telecom.isnull().sum()/len(telecom.index)), 2)
```

[231]:
```
customerID                                0.0
tenure                                    0.0
PhoneService                              0.0
PaperlessBilling                          0.0
MonthlyCharges                            0.0
TotalCharges                              0.0
Churn                                     0.0
SeniorCitizen                             0.0
Partner                                   0.0
Dependents                                0.0
Contract_One year                         0.0
Contract_Two year                         0.0
PaymentMethod_Credit card (automatic)     0.0
PaymentMethod_Electronic check            0.0
PaymentMethod_Mailed check                0.0
gender_Male                               0.0
InternetService_Fiber optic               0.0
InternetService_No                        0.0
MultipleLines_No                          0.0
MultipleLines_Yes                         0.0
OnlineSecurity_No                         0.0
OnlineSecurity_Yes                        0.0
OnlineBackup_No                           0.0
OnlineBackup_Yes                          0.0
DeviceProtection_No                       0.0
DeviceProtection_Yes                      0.0
TechSupport_No                            0.0
TechSupport_Yes                           0.0
StreamingTV_No                            0.0
StreamingTV_Yes                           0.0
StreamingMovies_No                        0.0
StreamingMovies_Yes                       0.0
dtype: float64
```