Telco - Customer Churn

Annapoorani

Springboard

Capstone project – Milestone Report

Problem Statement

Predict behavior to retain customers. To can analyze all relevant customer data and develop focused customer retention programs.

Outcomes

Telco - a Telecommunication company will be beneficial by this project. This will help them in the following ways

- Improve user experience with their products
- Identifying important features that are key for customer retention
- To come up with new Marketing Strategies.

Data

Telco data collected from Kaggle.

https://www.kaggle.com/blastchar/telco-customer-churn

Approach

The Steps to solve the identified problems are as below

- Data Wrangling using Python Pandas, so that data is cleaned and readily available for analysis
- EDA to discover the underlying facts about the data and to study the different features
- Applying machine learning algorithms like Logistic regression, Decision Trees and Deep Learning.

Data Wrangling

Success of analysis depends upon how the data is cleaned up as usable for analysis with columns as separate features and each row as single observation. As it is highlighted always as Data Wrangling is time consuming, wrangling for this project also was very challenging.

As this data is downloaded from Kaggle it is already cleaned. Two main steps performed is

1. Converting the Total and Monthly charges to numeric
2. Checked for any NA values in the columns, identified very few rows with Total Charges as NA and those rows are dropped

Data Pre-Processing

Most features of the Telco – Customer Churn Dataset are categorical variables. One hot encoding is performed on those categories and redundant columns are removed.

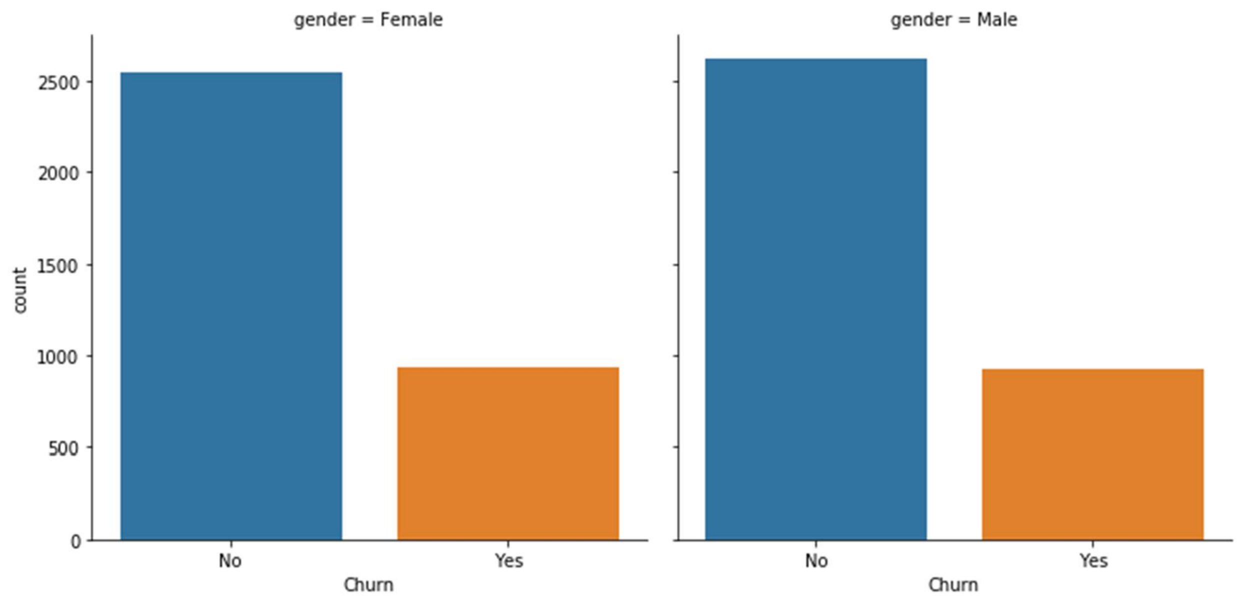
One major column of interest is the Internet Service, if the column is 'No' then obviously the concern customer won't have Streaming TV or Streaming Movies.

Henceforth "Internet_Service_No" column is retained by dropping Streaming_TV_No and Streaming_Movies_No columns.

Exploratory Data Analysis

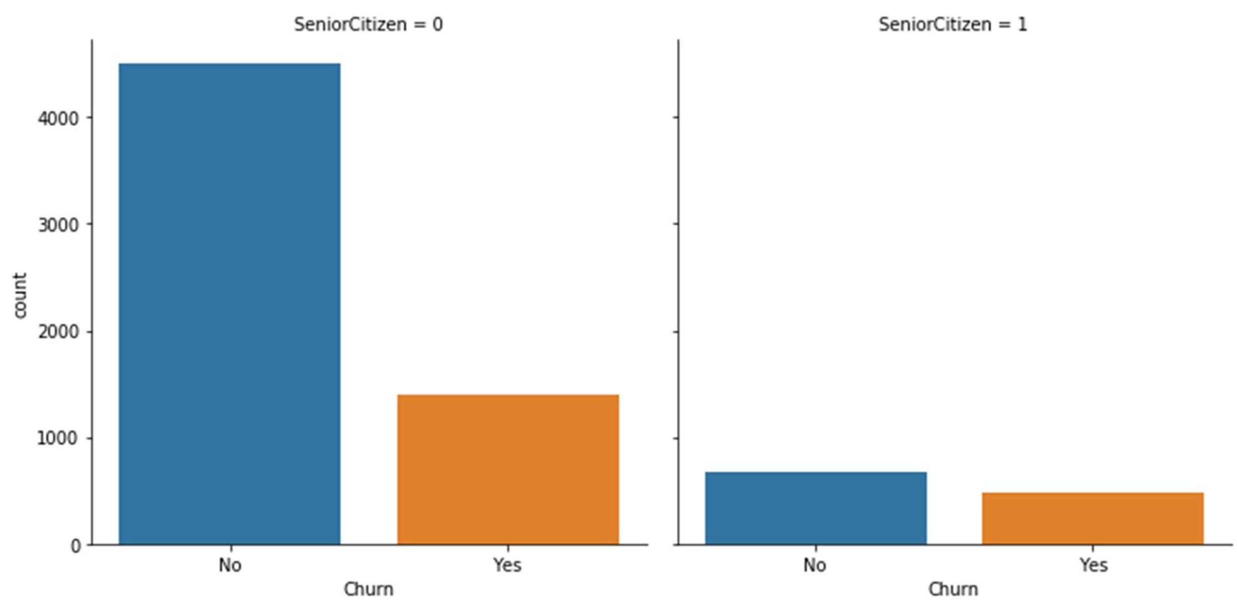
EDA helps in Visualizing underlying insights from the data. It helps in Feature Enginerring also paves way for identifying new scope of data. As churn is the key field of study,first part of the notebook deals with the picturization of churn data based on different features.

Churn Based on Gender



As from above chart churning rate looks similar for both male and female so there is no significant difference in churning rate based on gender.

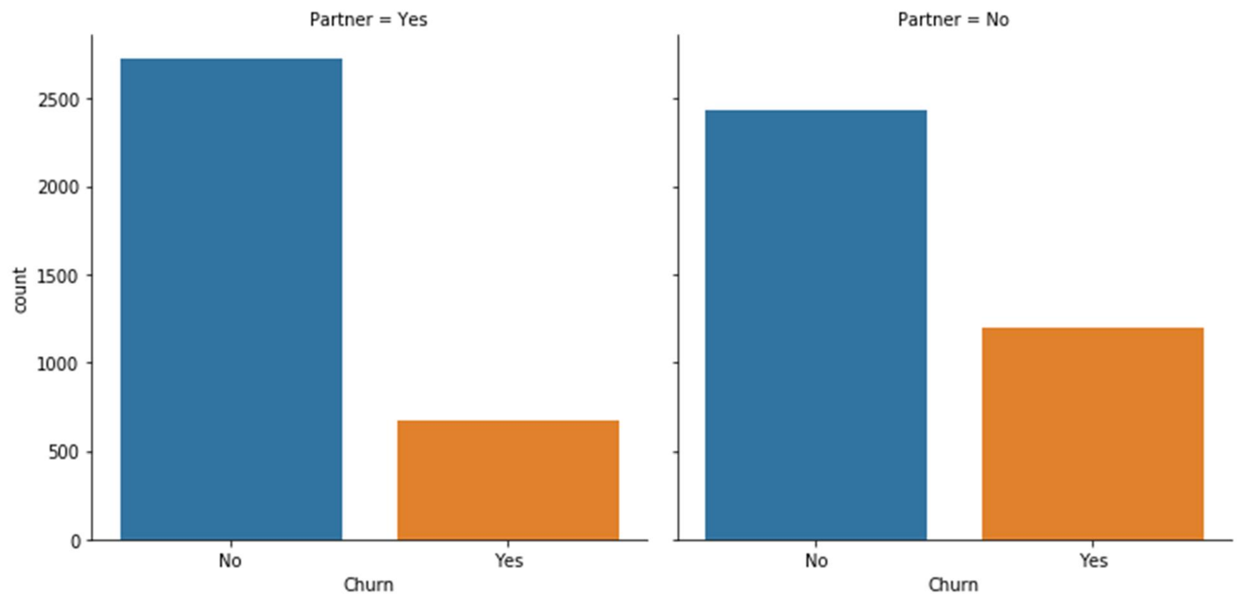
Churn Behavior with Senior Citizen



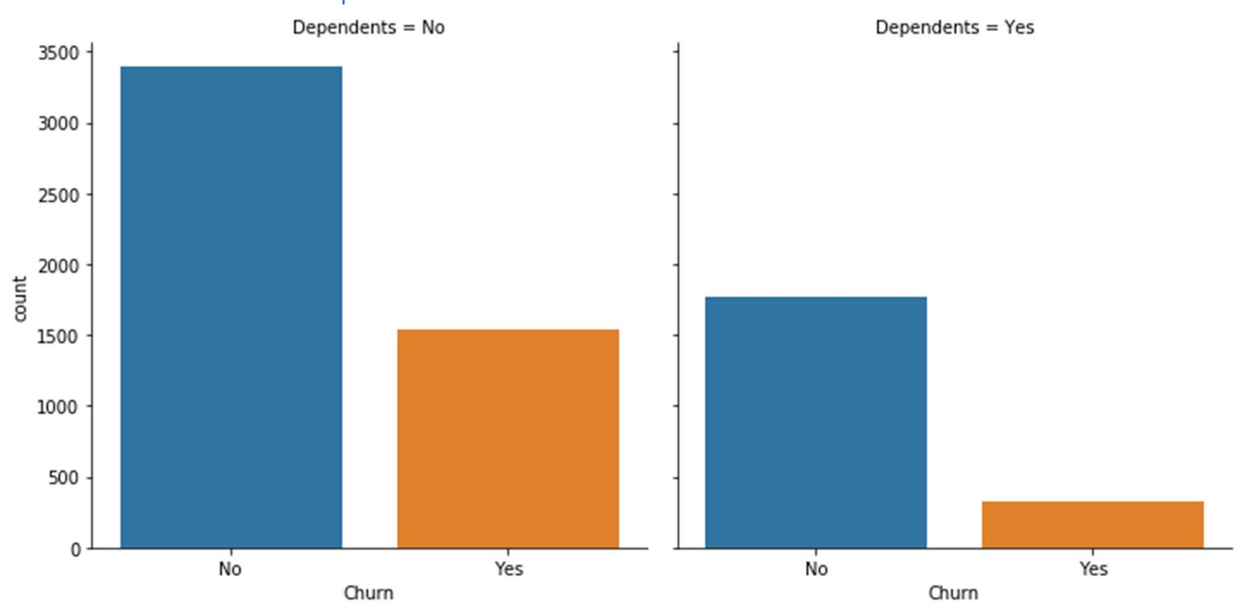
It seems Senior citizens churned and not churned numbers have very less difference, also the number of Total Senior Citizens is less compared to remaining population.

Partners Churn

If there is no partner then churn rate is little higher.

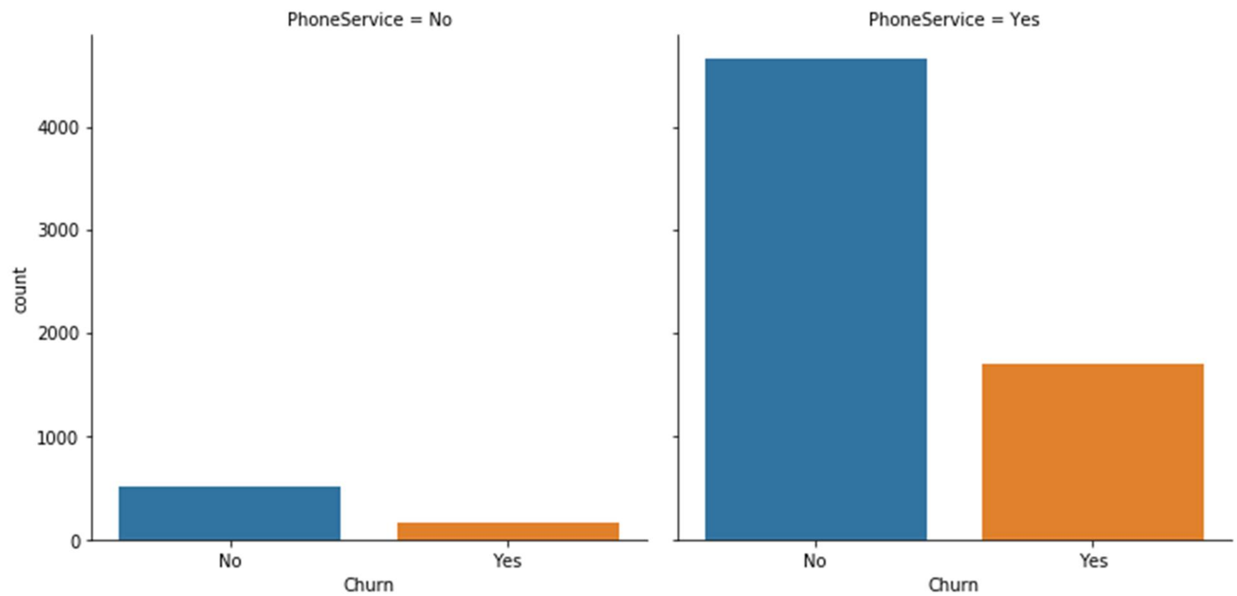


Churn Numbers for Dependents



If there are no Dependents then Churning numbers are higher

Customer Churn with Phone Service

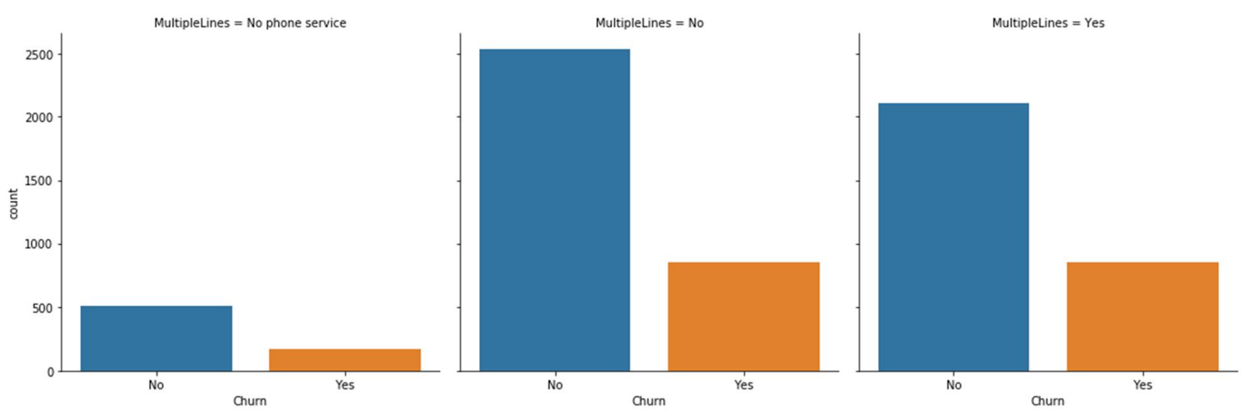



```
data.PhoneService.value_counts()
```

```
Yes    6352
No      680
Name: PhoneService, dtype: int64
```

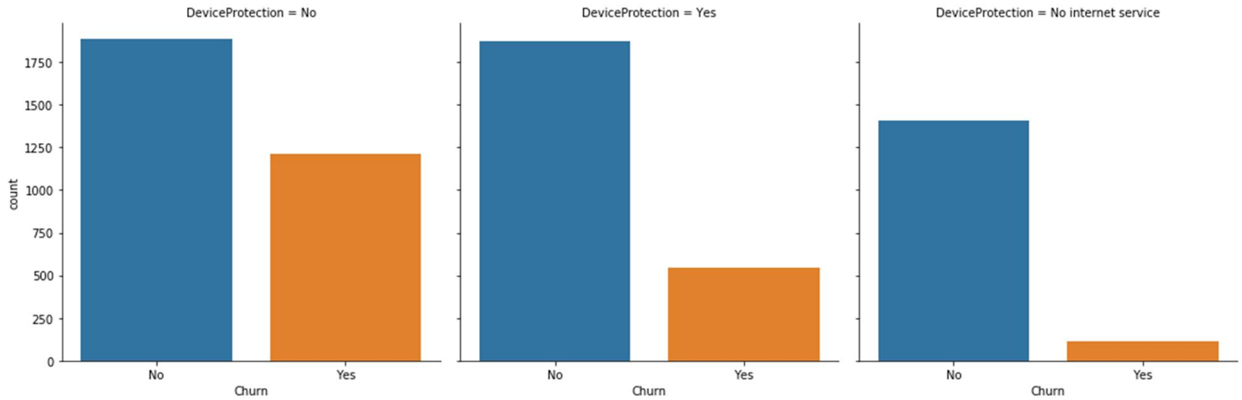
Most of the Telcom customers have Phone service too, Nearly 10% of them Phone Service customers are churned

Churn based on Multiple Lines



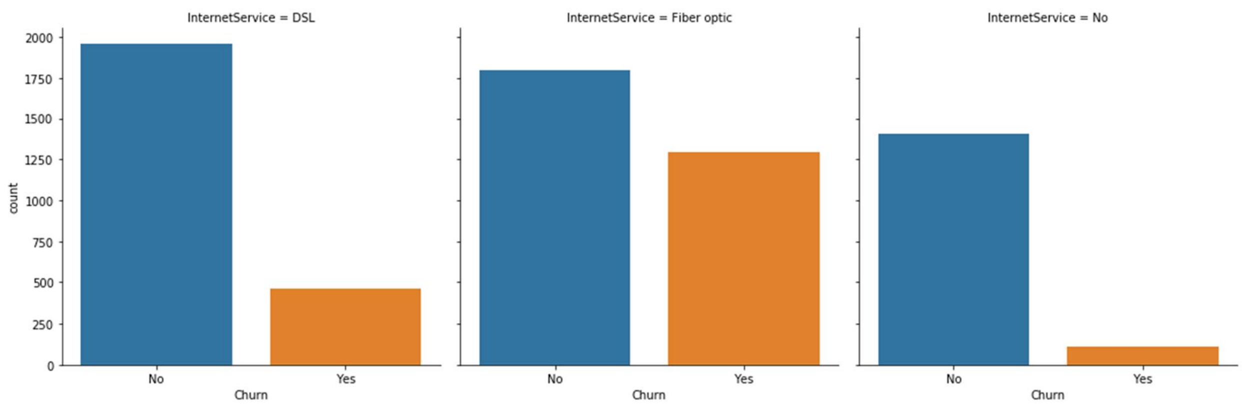
Irrespective of customer have multiple lines or not the Churning numbers are same.

Churn based on Device Protection



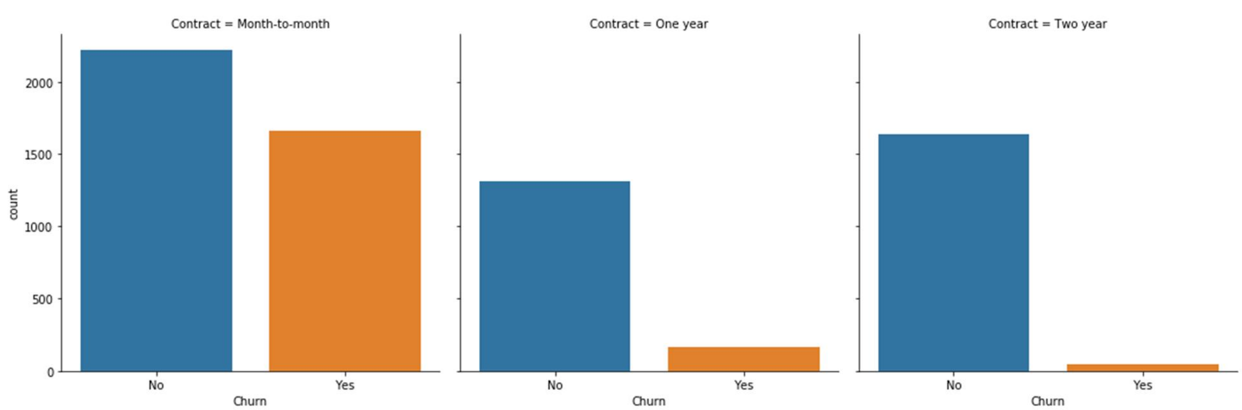
Customer without device protection churned higher

Churn based on Customer option on Internet service



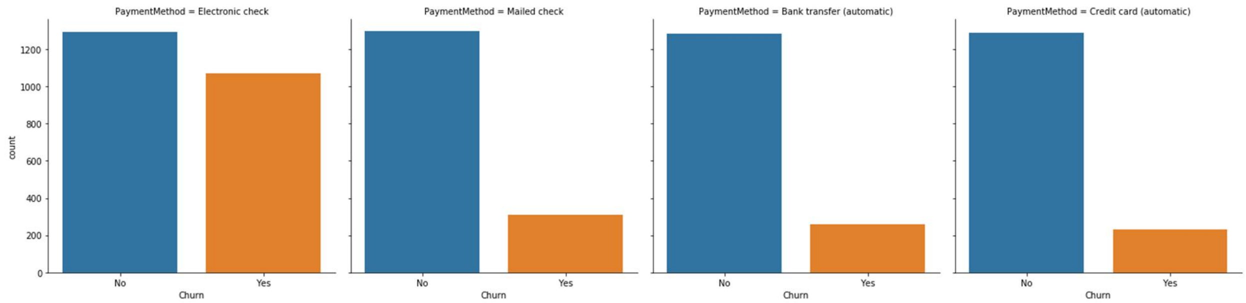
Customers getting Internet Service by Fiber optics Churned more, so definetly that may be a important feature.

Churn based on Contract Pattern



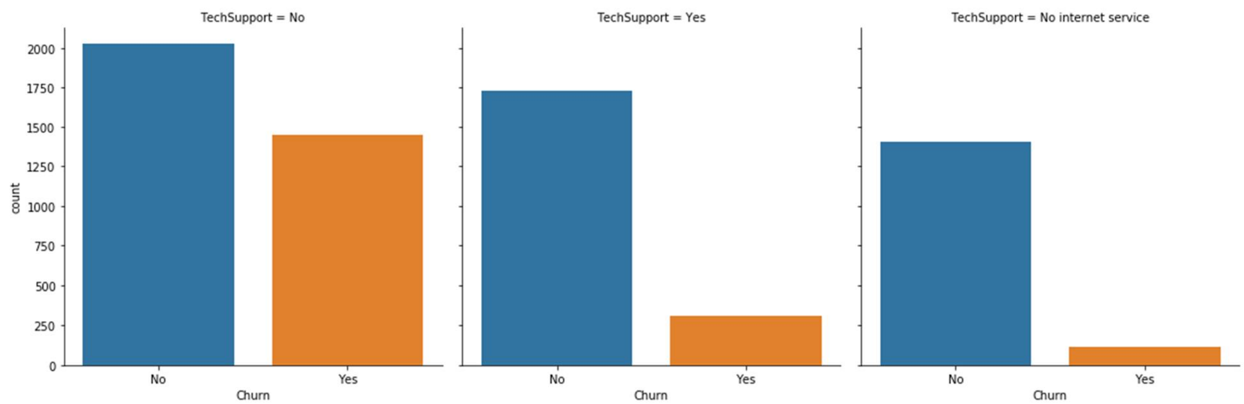
Customers chosen Month-to-month contract churned higher,this also may be a important feature

Payment Method – Churn Behavior



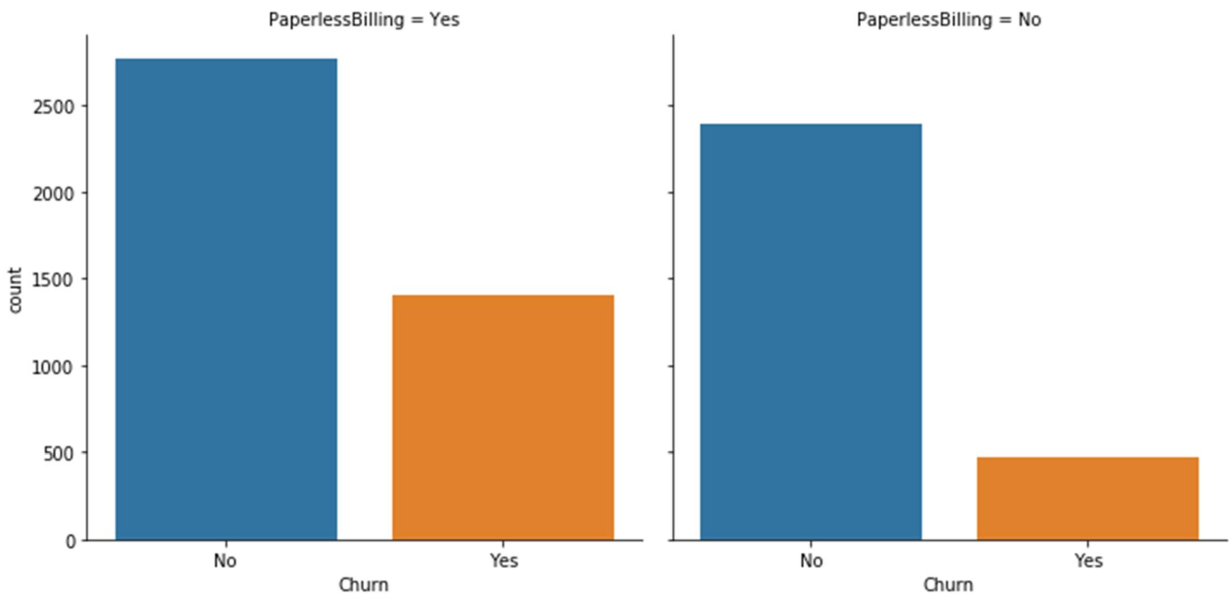
Customer paid by electronic check churned more

Customer Churn based on support

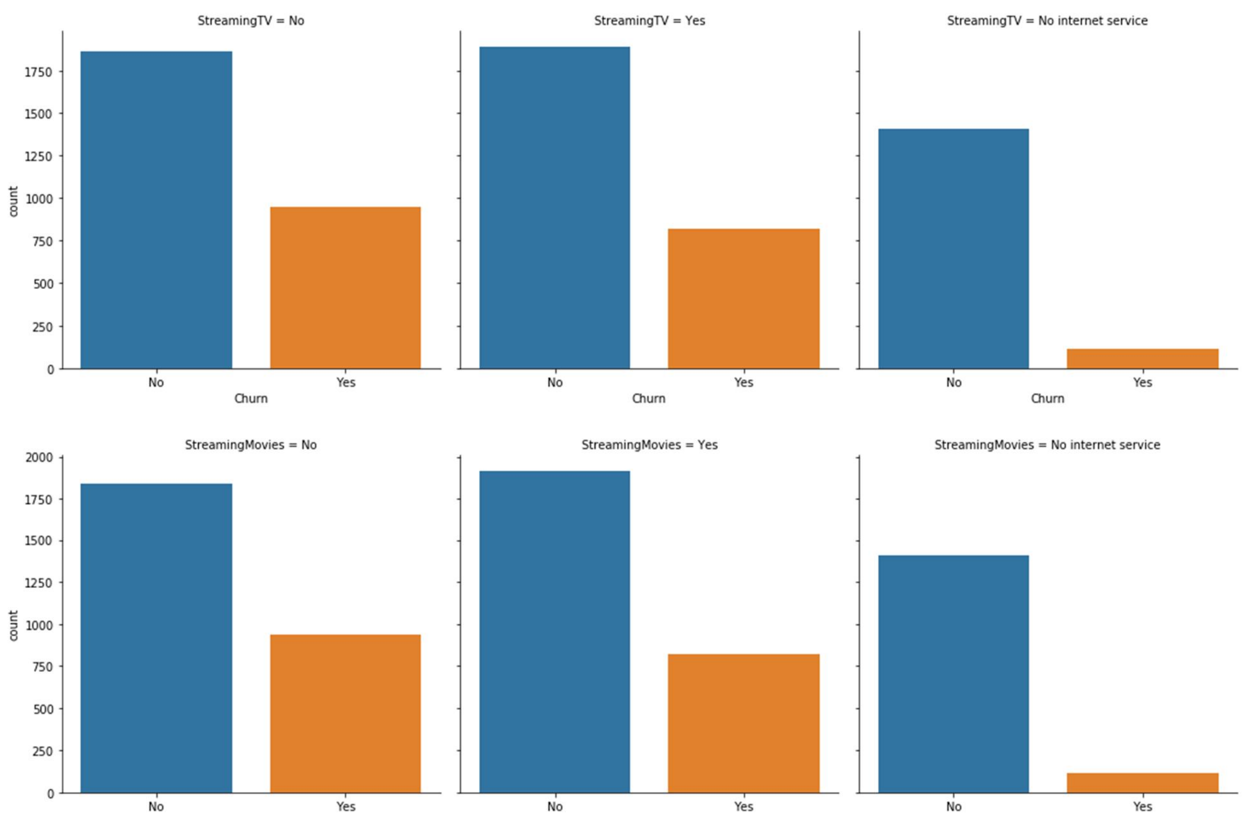


Customer without tech support churned higher than customer with tech support

Churn based on Billing Method



Churn by Streaming Services



Both Streaming Tv and Streaming Movies have greater churn rate

Feature Engineering

From EDA Internet Service, Contract, Payment Method, Streaming Tv and Streaming Movies, Tech Support are the Primary features and Paperless billing and Senior Citizen can be of secondary importance.

Second way is by applying statsmodel Logit function

```
from statsmodels.formula.api import logit
m = logit('Churn_Yes ~ gender_Female + tenure + MonthlyCharges + TotalCharges + SeniorCitizen_1 + Partner_Yes + Dependents_Yes +
print(m.summary())
```

```
Optimization terminated successfully.
         Current function value: 0.414311
         Iterations 9
                           Logit Regression Results
==============================================================================
Dep. Variable:              Churn_Yes   No. Observations:                 7032
Model:                          Logit   Df Residuals:                     7009
Method:                           MLE   Df Model:                           22
Date:                Wed, 03 Oct 2018   Pseudo R-squ.:                  0.2845
Time:                        12:51:43   Log-Likelihood:                -2913.4
converged:                       True   LL-Null:                       -4071.7
                                        LLR p-value:                     0.000
```

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	0.3202	3.87e+06	8.27e-08	1.000	-7.59e+06	7.59e+06
gender_Female	0.0221	0.065	0.341	0.733	-0.105	0.149
tenure	-0.0606	0.006	-9.713	0.000	-0.073	-0.048
MonthlyCharges	-0.0404	0.032	-1.271	0.204	-0.103	0.022
TotalCharges	0.0003	7.06e-05	4.657	0.000	0.000	0.000
SeniorCitizen_1	0.2168	0.085	2.564	0.010	0.051	0.382
Partner_Yes	0.0013	0.078	0.017	0.986	-0.151	0.154
Dependents_Yes	-0.1491	0.090	-1.662	0.097	-0.325	0.027
PhoneService_Yes	0.1743	0.649	0.269	0.788	-1.097	1.446
MultipleLines_Yes	0.4485	0.177	2.530	0.011	0.101	0.796
InternetService_Fiber_optic	1.7490	0.798	2.191	0.028	0.185	3.314
InternetService_No	-1.7879	0.807	-2.214	0.027	-3.370	-0.205
OnlineSecurity_Yes	-0.2049	0.179	-1.146	0.252	-0.555	0.145
OnlineBackup_Yes	0.0259	0.175	0.148	0.883	-0.318	0.370
DeviceProtection_Yes	0.1470	0.176	0.834	0.405	-0.199	0.493
TechSupport_Yes	-0.1810	0.181	-1.002	0.316	-0.535	0.173
StreamingTV_Yes	0.5916	0.326	1.813	0.070	-0.048	1.231
StreamingMovies_Yes	0.5998	0.327	1.836	0.066	-0.041	1.240
Contract_Month_to_month	0.7798	3.87e+06	2.01e-07	1.000	-7.59e+06	7.59e+06
Contract_One_year	0.1182	3.87e+06	3.05e-08	1.000	-7.59e+06	7.59e+06
Contract_Two_year	-0.5777	3.87e+06	-1.49e-07	1.000	-7.59e+06	7.59e+06
PaperlessBilling_Yes	0.3411	0.074	4.580	0.000	0.195	0.487
PaymentMethod_Electronic_check	0.3469	0.077	4.503	0.000	0.196	0.498
PaymentMethod_Mailed_check	-0.0146	0.101	-0.145	0.885	-0.212	0.183

The third approach is to apply Random Forest Classifier to identify the important approach

```python
from sklearn.ensemble import RandomForestClassifier
from mlxtend.feature_selection import SequentialFeatureSelector as sfs
clf = RandomForestClassifier(n_estimators=100, n_jobs=-1)

sfs1 = sfs(clf,
           k_features=10,
           forward=True,
           floating=False,
           verbose=2,
           scoring='accuracy',
           cv=5)

# Perform SFFS
sfs1 = sfs1.fit(X_train, y_train)

feat_cols = list(sfs1.k_feature_idx_)
print(feat_cols)
```

```
[0, 6, 9, 10, 11, 14, 15, 18, 19, 21]
```

```
[Parallel(n_jobs=1)]: Done  15 out of  15 | elapsed:   36.4s finished

[2018-10-03 12:59:23] Features: 10/10 -- score: 0.7750940036583287
```

```
new_data.columns
```

```
Index(['tenure', 'MonthlyCharges', 'TotalCharges', 'gender_Female',
       'SeniorCitizen_1', 'Partner_Yes', 'Dependents_Yes', 'PhoneService_Yes',
       'MultipleLines_Yes', 'InternetService_Fiber_optic',
       'InternetService_No', 'OnlineSecurity_Yes', 'OnlineBackup_Yes',
       'DeviceProtection_Yes', 'TechSupport_Yes', 'StreamingTV_Yes',
       'StreamingMovies_Yes', 'Contract_Month_to_month', 'Contract_One_year',
       'Contract_Two_year', 'PaperlessBilling_Yes',
       'PaymentMethod_Bank_transfer_automatic',
       'PaymentMethod_Electronic_check', 'PaymentMethod_Mailed_check',
       'Churn_Yes'],
      dtype='object')
```

Tenure,MonthlyCharges,SeniorCitizen_1,PhoneService_Yes,InternetService_Fiber_optic,InternetService_No,Contract_Month_to_month,Contract_One_year, Contract_Two_year,PaymentMethod_Electronic_check

From all the above three methods the concluded features are Tenure, MonthlyCharges, SeniorCitizen_1,PhoneService_Yes,InternetService_Fiber_optic,InternetService_No,Contract_Month_to_month,Contract_One_year, Contract_Two_year, PaymentMethod_Electronic_check