MINISTRY OF SCIENCE AND HIGHER EDUCATION OF THE RUSSIAN FEDERATION
FEDERAL STATE AUTONOMOUS EDUCATIONAL INSTITUTION OF HIGHER EDUCATION
NATIONAL UNIVERSITY OF SCIENCE AND TECHNOLOGY
« M I S I S »

Informatics and Computer Technologies
Major code 09.04.01

Course work

In Modern IT systems in Economics and Industry

Telecom Customer Churn Analysis

Student: Asmerom Fessehaye
Group: МИВТ-19-7-12А
Checked by: Aslan Agabubaev

Moscow
2020

Contents

1. Introduction

The customer churn task deals with the analysis of Customer data of telecom organization facing sever customer attrition rate. As part of the analysis firstly the data is preprocessed to make it suitable for exploratory data analysis. Then further the dataset is explored by basic visualizations using Seaborn and Plotly visualization libraries, before moving to advanced predictive analytics of the dataset. Then exploratory data analysis is carried out to understand the influence of individual predictors on the target variable (**Customer churn).** Then feature Selection is done by analyzing how each independent variable will influence customer churn rate and thereby decide the top predictors or key drivers of Customer churn.

The latter portion of this task examines the impact different parameters have on Customer Churn using some of the major statistical learning. We have built predictive models for customer churn using some of the major statistical methods like logistic regression and Random Forest Classifier. Using the predictive models to identify customers who fit the churn profile some tips have suggested the telecom organization with few strategies so that we can proactively target them with marketing and retention programs.

2. Problem statement

Customer churn is known as loss of customer:
Service Company often uses customer attrition(churn) analysis and customer attrition rates as one of their key business metrics because the cost of retaining an existing customer is far less than acquiring a new one. Long term customers can be worth much more to a company than newly recruited clients.

The objective is to find relation between potential defectors and churn rate, quantify their relation. Following relations are important to understand churn rate:

- ✓ Effect of Gender with respect to churn rate;
- ✓ Senior Citizen with respect to churn rate;
- ✓ Effect of tenure period on churn rate;
- ✓ Relation of Customer contract type and churn rate;
- ✓ Relation of Fiber optics service with churn rate;
- ✓ Relation of phone service, online security and multiple line with churn rate;
- ✓ Relation of Online backup, device protection and tech support with churn rate; and
- ✓ Effect of monthly charges and total charges on churn rate

3. Data Description and Basic Checkup

Data being used: We have worked on the "Telco Customer Churn" data set taken from KAGGLE https://www.kaggle.com/blastchar/telco-customer-churn.

Each row represents a customer; each column contains customer's attributes described on the column Metadata. The raw data contains 7043 rows (customers) and 21 columns (features). The "Churn" column is our target. This data set contains 21 columns (features), but we will be selecting a subset of the most important columns for analysis purposes.

The data set includes information about:

- Customers who left within the last month – the column is called Churn
- Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
- Customer account information – how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges
- Demographic info about customers – gender, age range, and if they have partners and dependents

Variable name	Type of Variable	Explanation
Customer ID	Numerical	Unique ID of customers
Gender	Categorical	Male / Female
Senior Citizen	Categorical	Yes/No
Partner	Categorical	Yes/No
Dependents	Categorical	Yes/No
Tenure	Numerical	Number of months customer used service
Phone Service	Categorical	Yes/No
Multiple Lines	Categorical	Yes/No/No phone services
Internet Services	Categorical	DSL/Fiber optics/No
Online Security	Categorical	No/Yes/No internet service
Online Backup	Categorical	No/Yes/No internet service
Device protection	Categorical	No/Yes/No internet service
Tech Support	Categorical	No/Yes/No internet service
Streaming TV	Categorical	No/Yes/No internet service
Streaming Movies	Categorical	No/Yes/No internet service
Contract	Categorical	Month to Month/1Year/2Years
Monthly Charges	Numerical	Amount charged in USD
Total Charges	Numerical	Amount charged in USD

Importing Libraries:-

```
#Importing libraries
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)

import os
import matplotlib.pyplot as plt#visualization
from PIL import  Image
%matplotlib inline
import pandas as pd
import seaborn as sns#visualization
import itertools
import warnings
warnings.filterwarnings("ignore")

import io
import plotly.offline as py#visualization
py.init_notebook_mode(connected=True)#visualization
import plotly.graph_objs as go#visualization
import plotly.tools as tls#visualization
import plotly.figure_factory as ff#visualization

from sklearn.linear_model import LogisticRegression
from sklearn.metrics import confusion_matrix,accuracy_score,classification_report
from sklearn.model_selection import train_test_split, cross_val_score, GridSearchCV, cross_val_predict
from sklearn.metrics import roc_auc_score,roc_curve,scorer
from sklearn.metrics import f1_score
import statsmodels.api as sm
from sklearn.metrics import precision_score,recall_score
from yellowbrick.classifier import DiscriminationThreshold

from sklearn.ensemble import RandomForestClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.tree import export_graphviz
```

Load data and display top 5 rows

```
# Load raw data
raw_data = pd.read_csv(r"WA_Fn-UseC_-Telco-Customer-Churn.csv")

raw_data.head()
```

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	...	DeviceProtection	TechSup
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 rows × 21 columns

Display the shape of the data which is the total number of rows and columns (features)

```
#Dimension check

raw_data.shape
```

```
(7043, 21)
```

Check for missing value, the result indicates that there are no missing values in the dataset. Here as we can see there are no missing values in the data set.

```
#Check if there's a missing data at each column

raw_data.isnull().any()
#customer.isnull().sum().max()
#customer.isnull().sum()
```

```
customerID          False
gender              False
SeniorCitizen       False
Partner             False
Dependents          False
tenure              False
PhoneService        False
MultipleLines       False
InternetService     False
OnlineSecurity      False
OnlineBackup        False
DeviceProtection    False
TechSupport         False
StreamingTV         False
StreamingMovies     False
Contract            False
PaperlessBilling    False
PaymentMethod       False
MonthlyCharges      False
TotalCharges        False
Churn               False
dtype: bool
```

Check for Duplicate values: - False indicates no duplicate values

```
# Duplicate value check

raw_data.duplicated().any()
```

```
False
```

Check data formatting

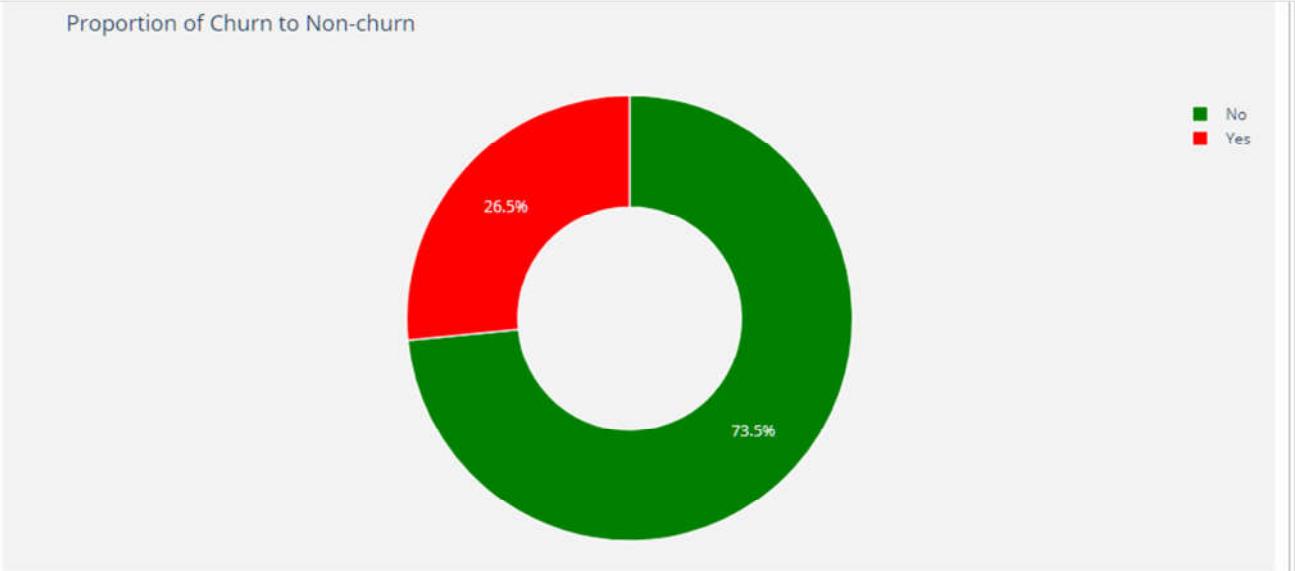
```
#check data formating / Exploring data types of each feature

#customer.dtypes
raw_data.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   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
dtypes: float64(1), int64(2), object(18)
memory usage: 1.1+ MB
```

[6]

4. Exploratory Data Analysis

Proportion of Churn to Non-churn

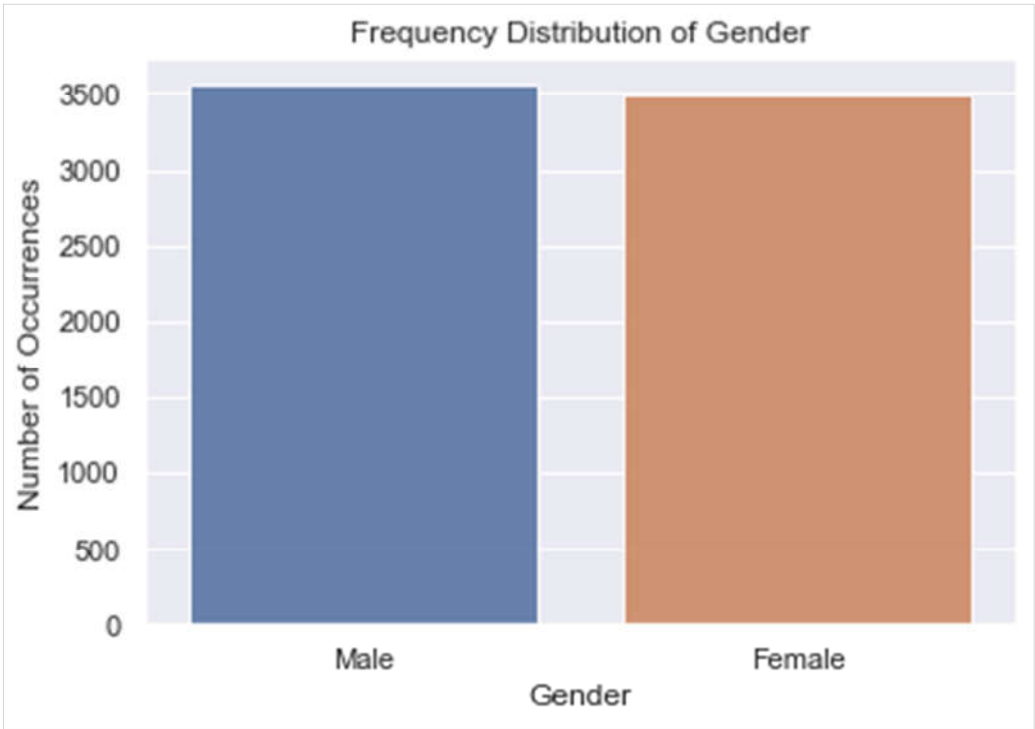


The churn rate is the percentage of subscribers to a service who discontinue their subscriptions to the service within a given time. So, current attrition rate is 26.5%. Objective of project is to determine bottle neck issue in the business to decrease churn rate.

Following relations have been studied to understand which variables are more related to churn rate:

1. Frequency distribution of **Gender** with respect to churn rate.
2. Frequency Distribution of **Senior Citizen** with respect to churn rate.
3. Frequency Distribution of **Dependents** with respect to churn rate.
4. Frequency Distribution of **Partner** with respect to churn rate.
5. Frequency Distribution of **Payment Method** with respect to churn rate.
6. Frequency Distribution of **Paperless Billing** with respect to churn rate.
7. Relation of **Monthly Salary** and Churn rate
8. Understanding the effect of **tenure** period on churn
9. Relation of Customer **contract type** and churn rate
10. Relation of **all services** with churn rate

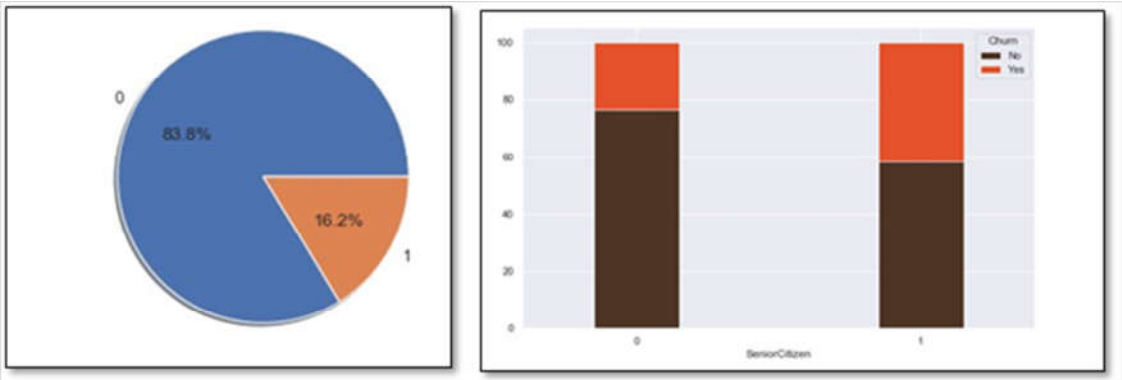
Frequency distribution of Gender with respect to churn rate:



Number of customers in dataset has almost **50:50** distribution of male and female. So, there no effect of gender biasedness

Relative frequency distribution of customer as per they are senior citizen or not

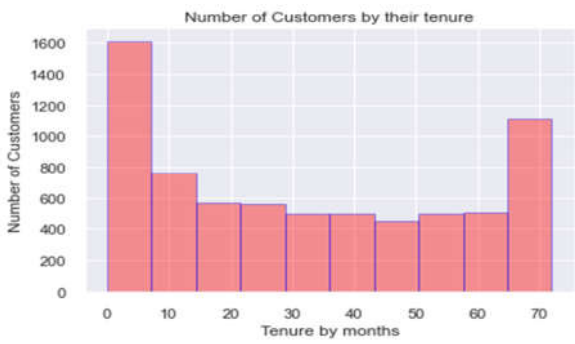
Distribution of senior citizen - 0 denotes young citizen and 1 denotes senior citizen



Senior citizen pie chart Young Citizen Senior Citizen Bar chart of churn rate in

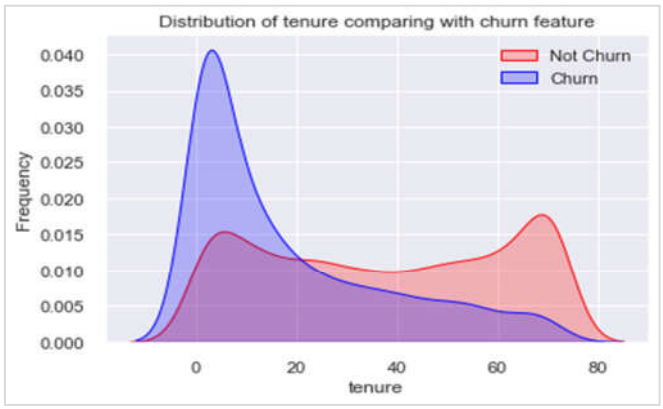
Out of total customers, senior citizens are 16.2%. In senior citizen, churn rate is higher than in non-senior citizen group.

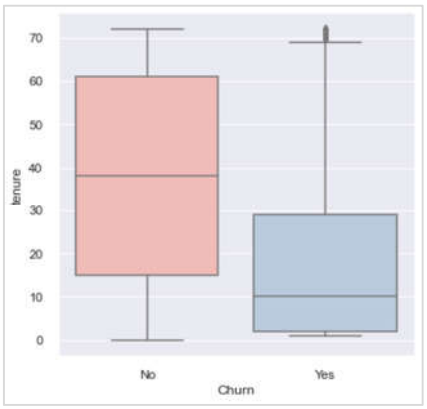
Effect of tenure period on churn rate



Tenure distribution

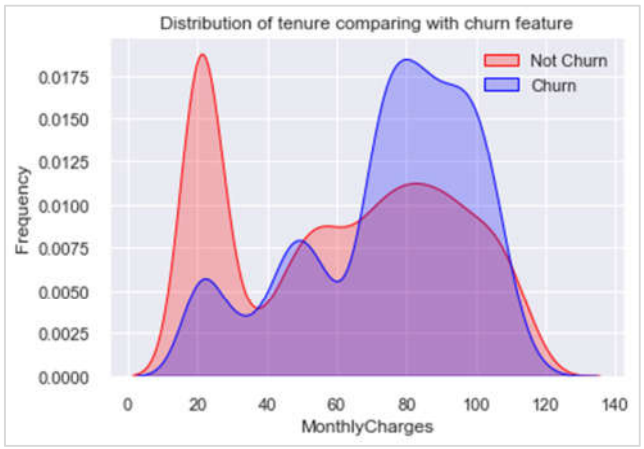
As customer is in initial period of tenure, chances of leave service are higher.

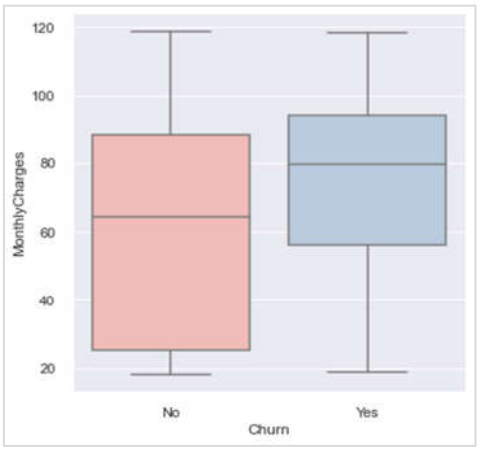




Relation of Monthly Salary and Churn rate

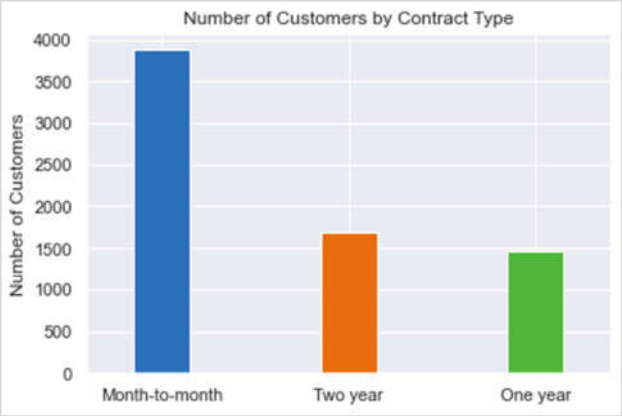
Using distribution and Box plot



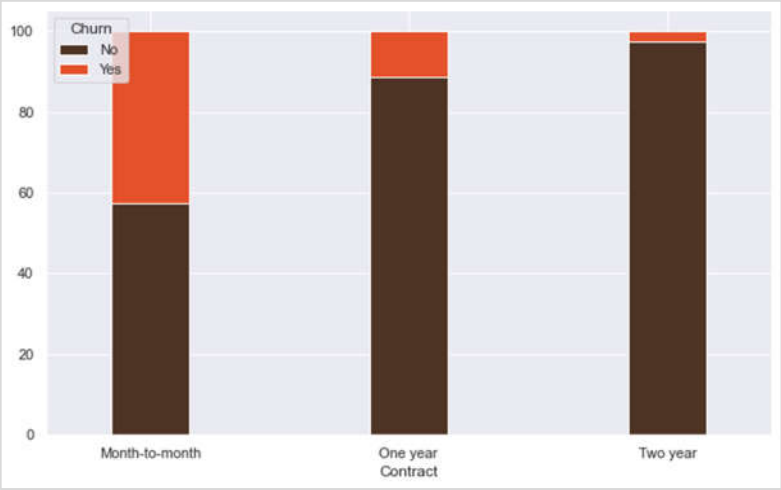


As we can see from above distribution, customer will more likely leave his/her monthly charges more than $60

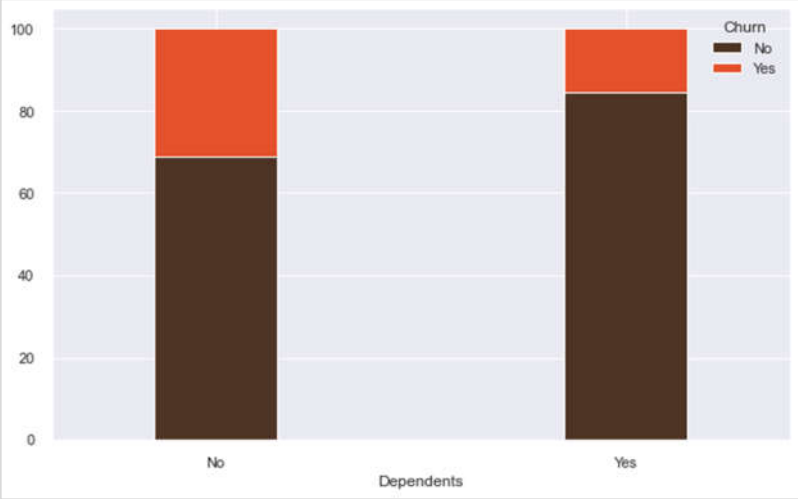
Relation of Customer contract type and churn rate:



Distribution of Customer by contract type
Month-to-month contract type is having highest number of customers following by two years and one year.
Stacked plot of contract type w.r.t to churn rate: - Month to month is having highest churn

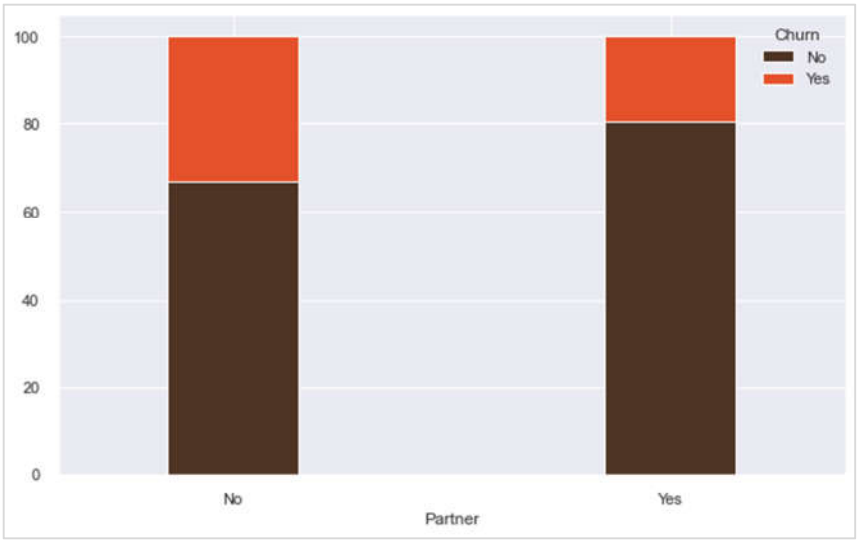


Effect of Dependents on Churn rate



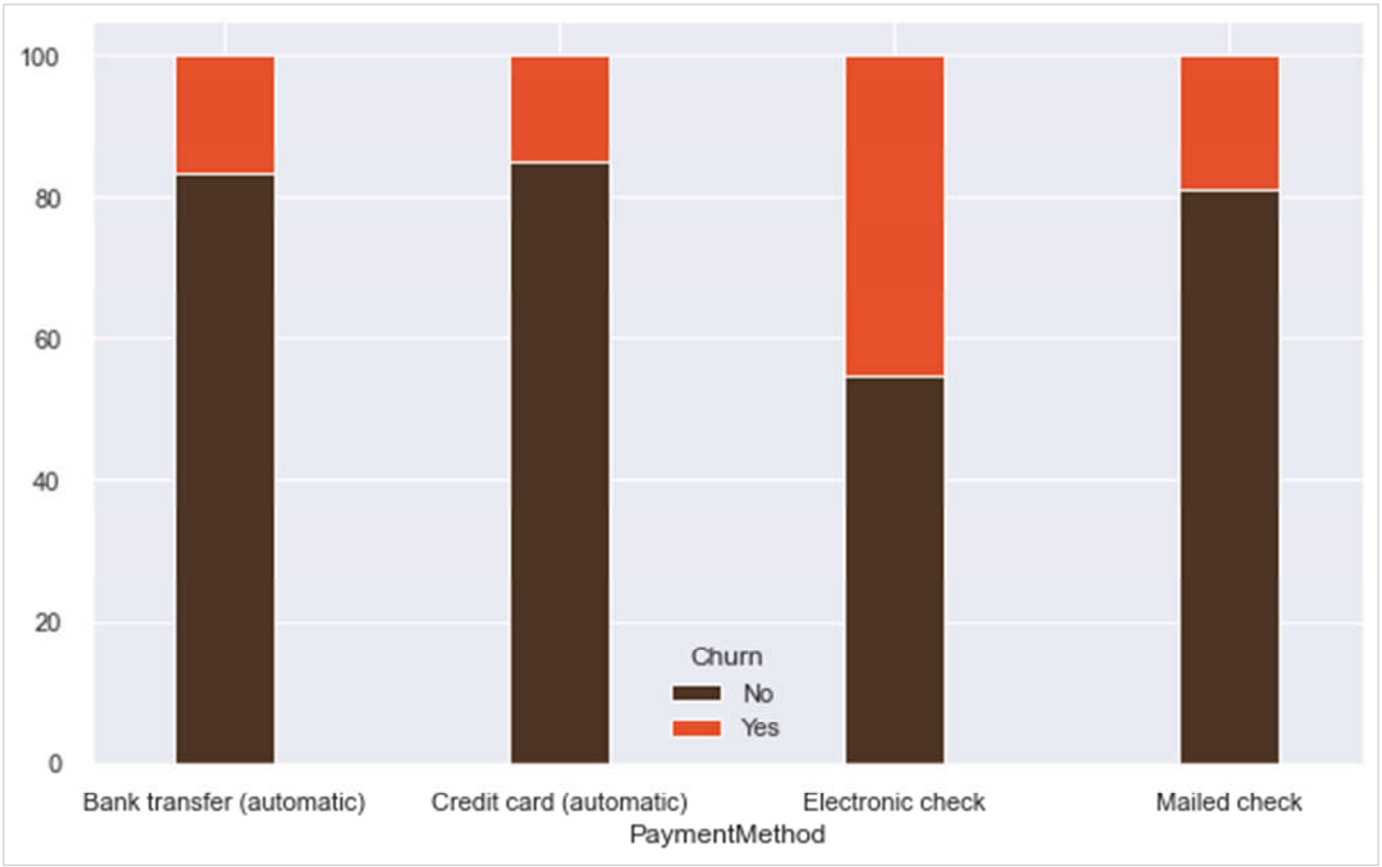
As we can see the customer with no dependents have high probability to leave the company.

Effect of Partner on Churn rate



As we can see from the above the customer with no Partner has high probability to leave the company.

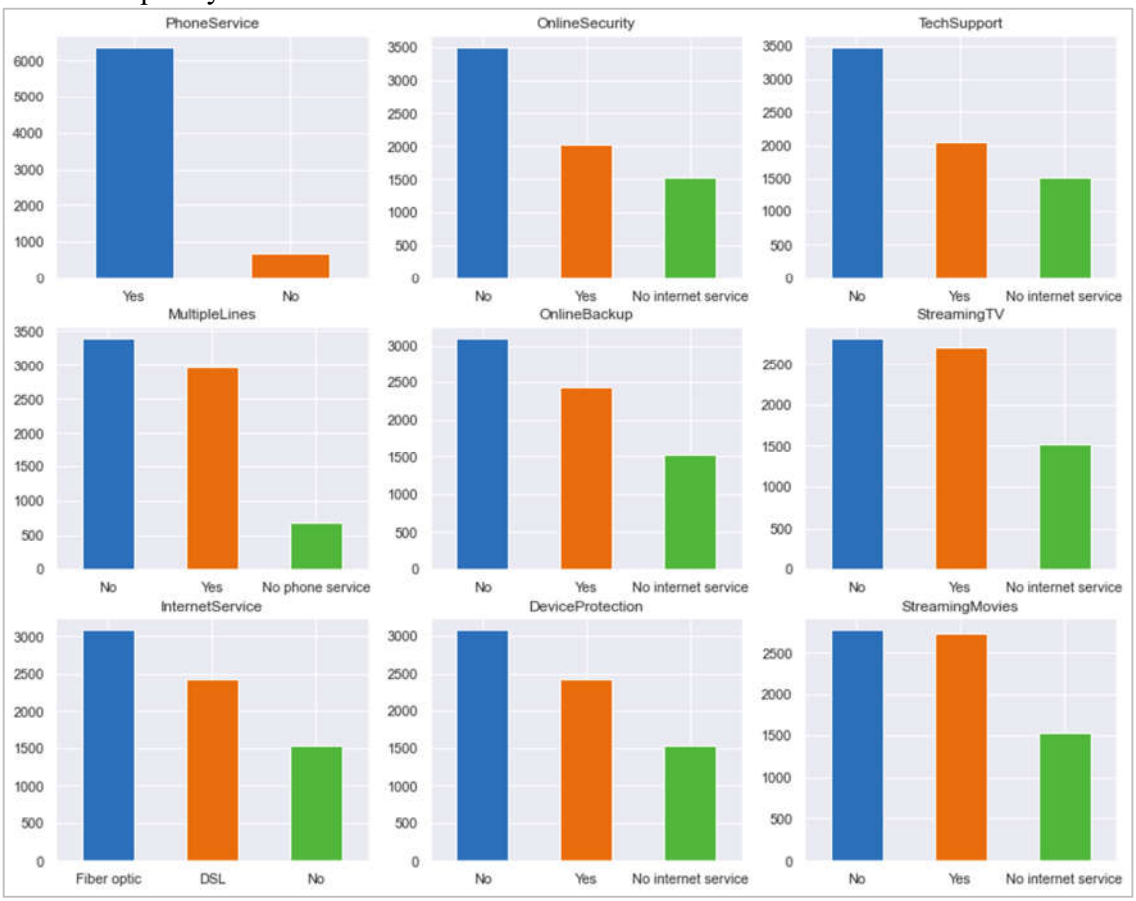
Effect of Payment Method on Churn rate



Those Customers who pay with electronic check have high probability of leaving (Churn).

Study of effect of services on churn rate

Here is frequency distribution of all services.



Relation of Fiber optics service with churn rate:

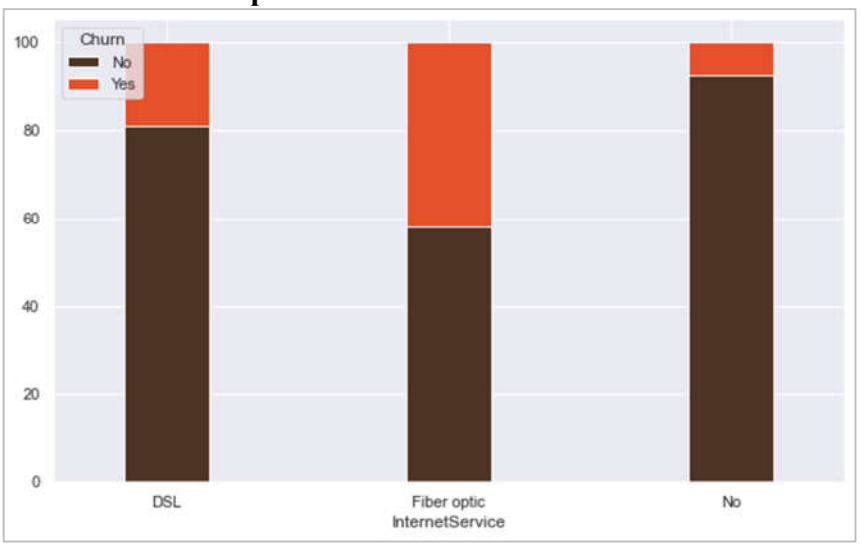


Fig:- Stacked plot of internet service

Observation: Customers using Fiber optics churn more than other group customers.

Relation of phone service with churn rate:

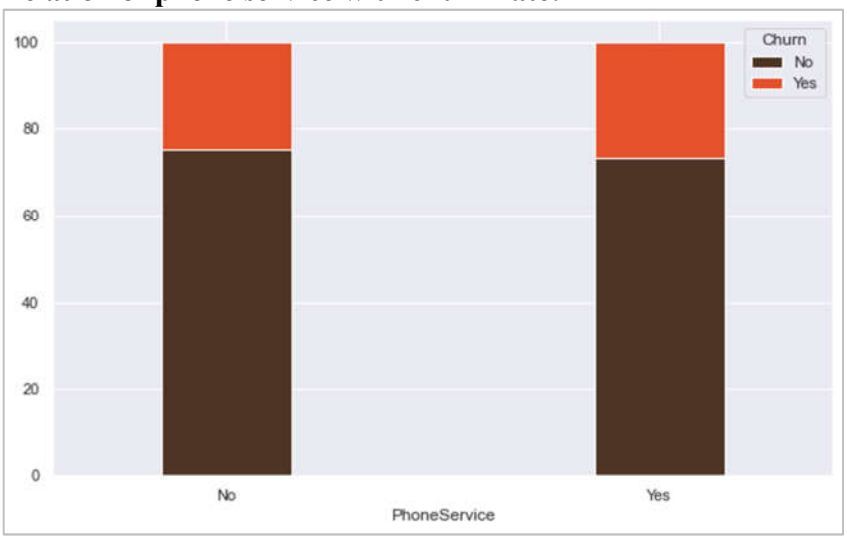


Fig: - Stacked plot of phone service

Churn rate is approximately same for customer using phone service vs. those not using.

Relation of internet service subscription with churn rate

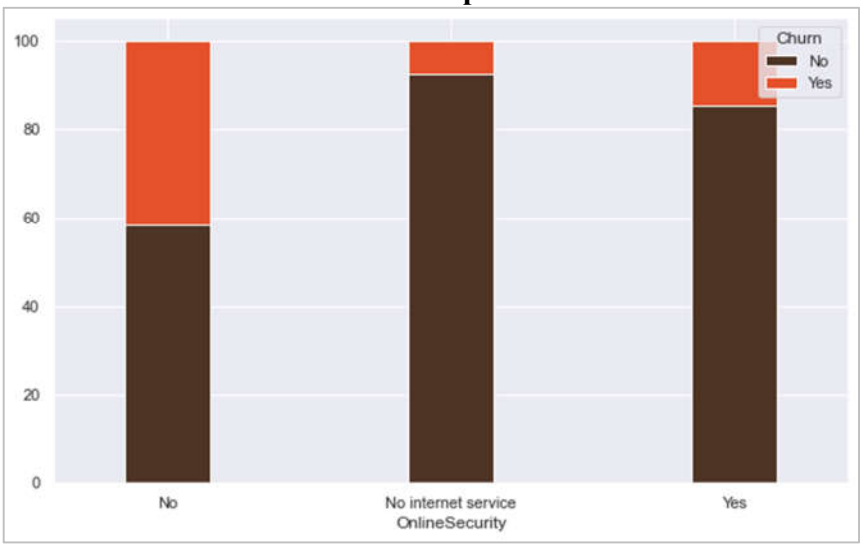


Fig:- Stacked plot of online security

Customer who doesn't use internet services churn more than other group customers.

Relation of multiple line service with churn rate:

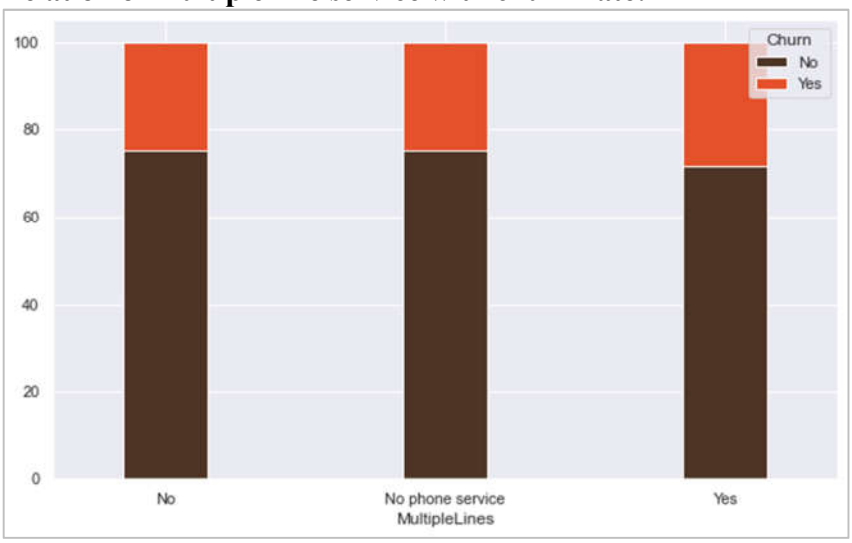


Fig:- Stacked plot of multiple lines
Multiple line service doesn't have any relation with churn rate.

Relation of online backup service with churn rate:

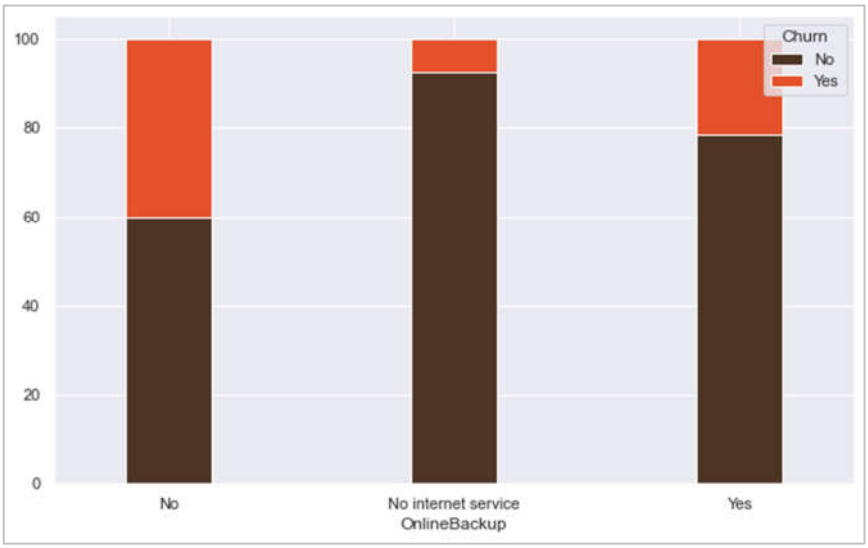


Fig: - Stacked plot of online backup

Customers who doesn't use online backup option, churn more than other group of customers

Relation of device protection service with churn rate:

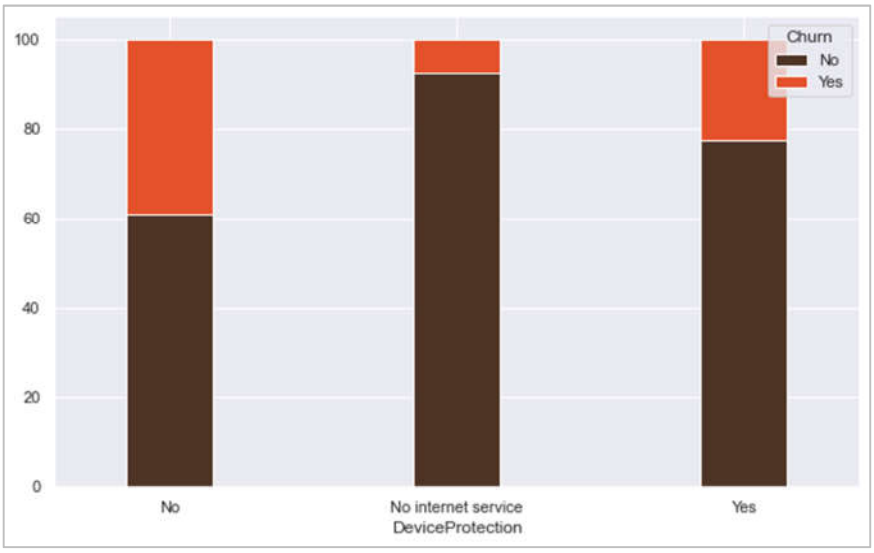


Fig: - Stacked plot of device protection

Customer with internet services doesn't use device protection, churn more than other group of customers

Relation of tech support service with churn rate:

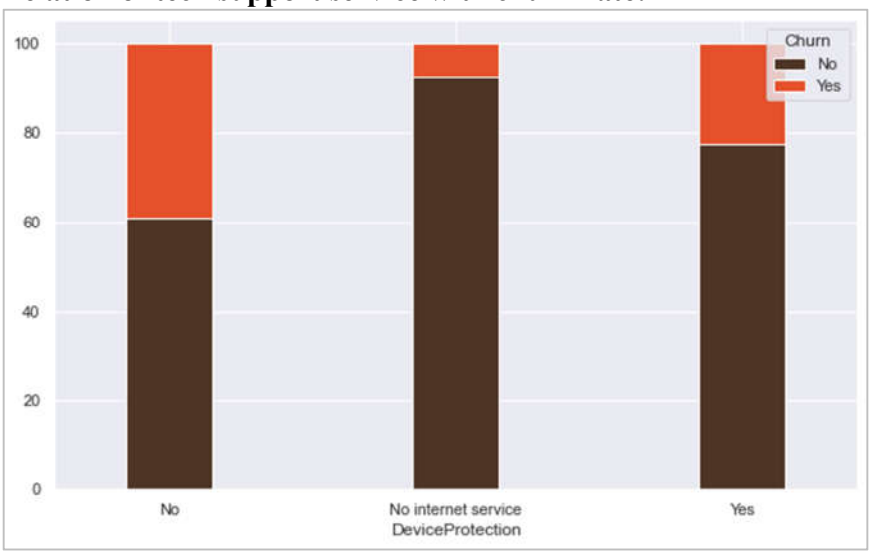


Fig: - Stacked plot of tech support

Customer with internet services doesn't use tech support, churn more than other group of customers

Relation of streaming TV with churn rate:

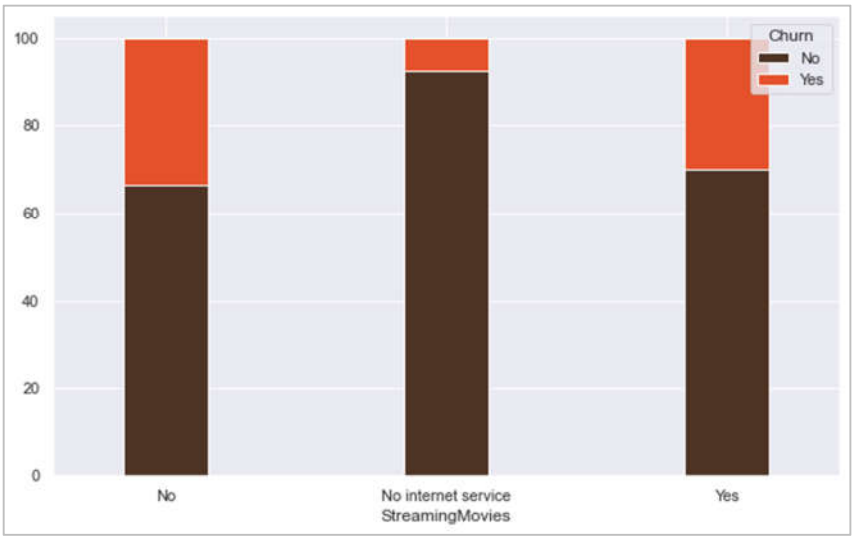


Fig: - Stacked plot of streaming TV

Customer with internet services who use or doesn't use streaming TV, churn more than other group of customers.

Conclusion of data exploration:

- ❖ Senior citizen has higher churn rate.
- ❖ Customers in initial period of tenure have higher churn rate.
- ❖ Customers with Monthly charges more that $60 have higher churn rates
- ❖ Churn rate is higher in initial period of tenure.
- ❖ Month-to month contract has highest churn rate.
- ❖ Fiber optics customer churns more frequently than other group customer.
- ❖ Customers with no partners and Dependents has highest churn rate
- ❖ Customers who don't have online security service, online backup service, device protection service, Tech support service have high churn rate
- ❖ Customers who have streaming TV and streaming Movies services have highest churn rate
- ❖ Customers with monthly based contract have higher churn rate
- ❖ Customers with paperless billing service have highest churn rate
- ❖ Customers who have electronic check payment method have higher churn rate

5. Data Preprocessing and Transformation

In this step the main tasks are:-

- Replacing spaces with null values in total charges column
- Dropping null values from total charges column which contain .15% missing data
- Replace 'No internet service' to No for the following columns
- Label encoding Binary columns (Yes=1 and No=0)
- Scaling Numerical columns using **StandardScaler**

Here is the result of after performing the above operation

```
customer = Data_transformation(raw_data)
customer.head()
```

	CustomerID	Gender	SeniorCitizen	Partner	Dependents	Tenure	PhoneService	MultipleLines	InternetServiceType	OnlineSecurity	...	StreamingTV	Stream
0	7590-VHVEG	Female	0	1	0	1	0	0	DSL	0	...	0	
1	5575-GNVDE	Male	0	0	0	34	1	0	DSL	1	...	0	
2	3668-QPYBK	Male	0	0	0	2	1	0	DSL	1	...	0	
3	7795-CFOCW	Male	0	0	0	45	0	0	DSL	1	...	0	
4	9237-HQITU	Female	0	0	0	2	1	0	Fiber optic	0	...	0	

5 rows × 23 columns

Descriptive statistics

```
customer[cat_cols].describe(include='all')
```

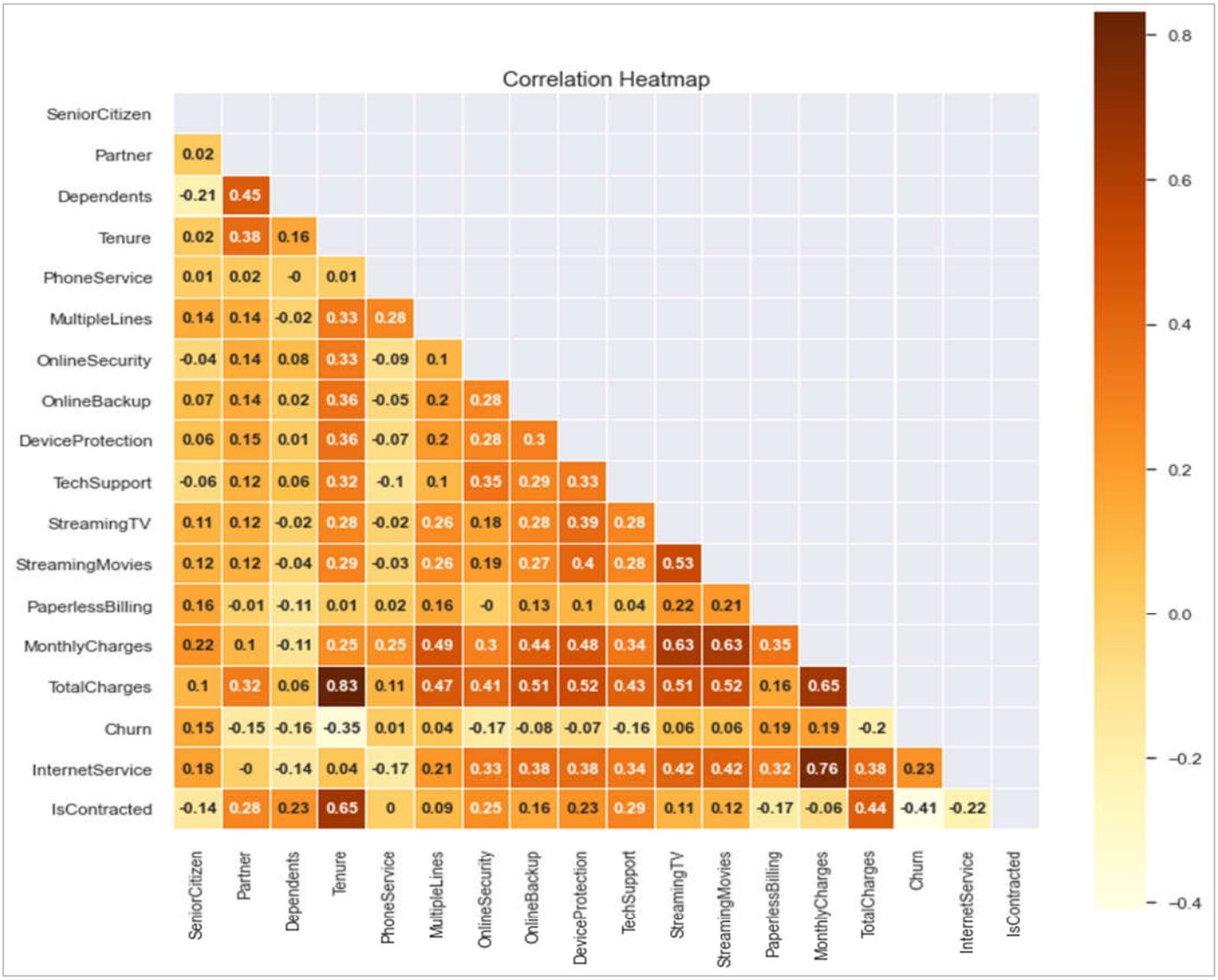
	Gender	Partner	Dependents	SeniorCitizen	PhoneService	MultipleLines	InternetServiceType	OnlineSecurity	OnlineBackup	DeviceProtection	Te
count	7043	7043.000000	7043.000000	7043.000000	7043.000000	7043.000000	7043	7043.000000	7043.000000	7043.000000	70
unique	2	NaN	NaN	NaN	NaN	NaN	3	NaN	NaN	NaN	
top	Male	NaN	NaN	NaN	NaN	NaN	Fiber optic	NaN	NaN	NaN	
freq	3555	NaN	NaN	NaN	NaN	NaN	3096	NaN	NaN	NaN	
mean	NaN	0.483033	0.299588	0.162147	0.903166	0.421837	NaN	0.286668	0.344881	0.343888	
std	NaN	0.499748	0.458110	0.368612	0.295752	0.493888	NaN	0.452237	0.475363	0.475038	
min	NaN	0.000000	0.000000	0.000000	0.000000	0.000000	NaN	0.000000	0.000000	0.000000	
25%	NaN	0.000000	0.000000	0.000000	1.000000	0.000000	NaN	0.000000	0.000000	0.000000	
50%	NaN	0.000000	0.000000	0.000000	1.000000	0.000000	NaN	0.000000	0.000000	0.000000	
75%	NaN	1.000000	1.000000	0.000000	1.000000	1.000000	NaN	1.000000	1.000000	1.000000	
max	NaN	1.000000	1.000000	1.000000	1.000000	1.000000	NaN	1.000000	1.000000	1.000000	

```
customer[num_cols].describe()
```

	Tenure	MonthlyCharges	TotalCharges
count	7043.000000	7043.000000	7043.000000
mean	32.371149	64.761692	2279.734304
std	24.559481	30.090047	2266.794470
min	0.000000	18.250000	0.000000
25%	9.000000	35.500000	398.550000
50%	29.000000	70.350000	1394.550000
75%	55.000000	89.850000	3786.600000
max	72.000000	118.750000	8684.800000

Based on numerical columns descriptive analysis, average tenure, average monthly charges and total charges of subscriber is 32 months, 64 dollars and 2279 dollars, respectively.

Correlation Matrix



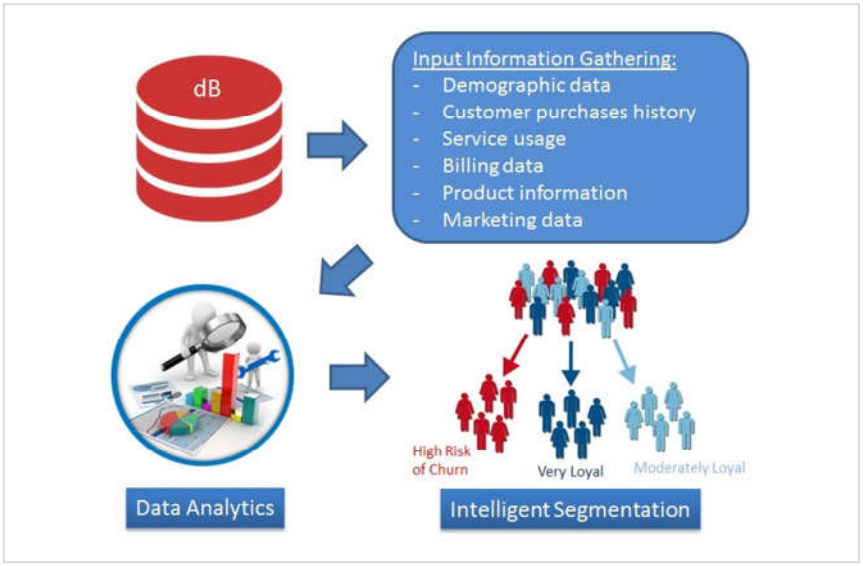
Features like Tenure, Monthly charges and Total charges are highly correlated with services like MultipleLines of phone services and Internet services like OnlineSecurity, OnlineBackup,DeviceProtection, TechSupport, StreamingTV and Streaming Movies services.

6. Customer Segmentation using cluster analysis

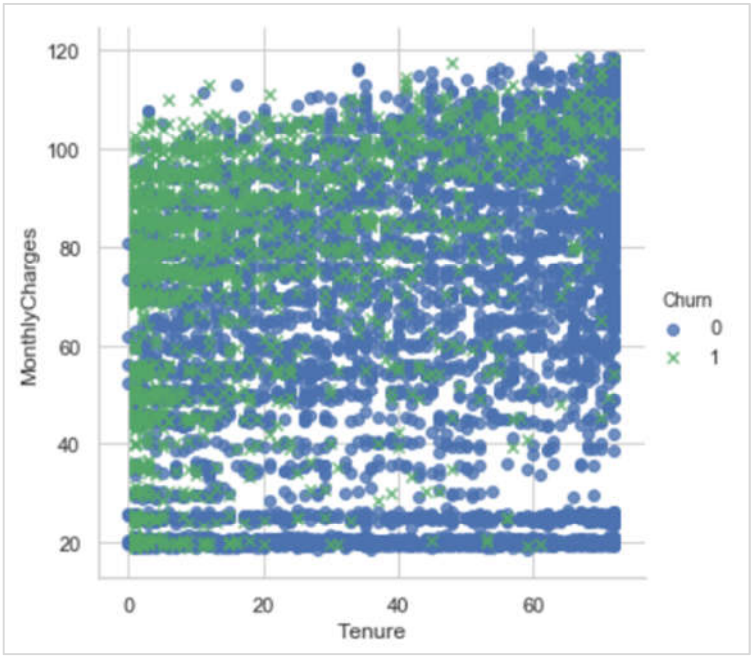
In this part, we'll use KMeans to build a Customer Segmentation model, and then we can compare the different distributions of each cluster.

Customer segmentation is the practice of dividing customers into groups of individuals that are similar in specific ways relevant to marketing, such as age, months in service, services usage and services spending.

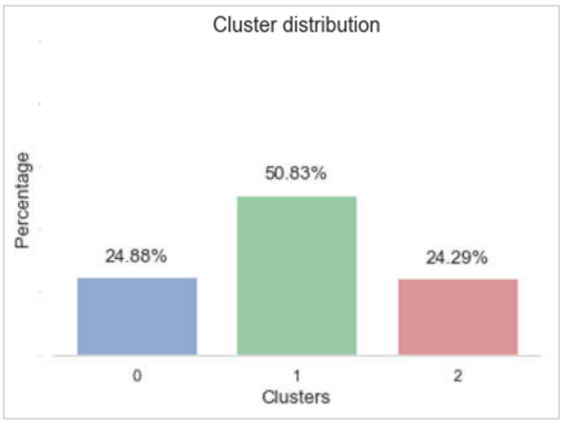
Based on the customer segmentation, the telecommunications company can customize their marketing efforts for different cluster, and gain a deeper understanding of their customers' preferences in order to more accurately tailor marketing materials.



First let's check if there is any relationship between Tenure and Monthly Charges.



From the analysis, there are some clusters based on Tenure and Monthly Charges. Let's apply K-means cluster algorithm to see clusters. Before passing data to K-means algorithm, need to normalize Tenure and Monthly Charges.



Based on K-means cluster graph, we can see that there are three groups.

1. Low Tenure and High Monthly Charges (Green cluster)
2. High Tenure and High Monthly Charges (Red cluster)
3. Low Tenure and Low Monthly Charges (Blue cluster)

And From distribution graph shows that around 50% of the customers belong to cluster Low Tenure and High Monthly Charges

Let's check there average Tenure and Monthly Charges by cluster.

Cluster	Tenure	MonthlyCharges
0	7.808602	38.329140
1	8.805263	83.795000
2	47.592511	91.855837

Based on demographic/usage/account related information, we need to explore characteristics of churn customer by each cluster.

Cluster 1 - Low Tenure and Low Monthly Charges

	Category	Label	Percentage	Cluster	Avg_Tenure	Avg_MonthlyCharges	Represent_in_graph	Label_in_graph
0	Gender	Male	0.535484	0	7.73	38.82	1	Male
3	SeniorCitizen	Not have a SeniorCitizen	0.862366	0	7.72	38.33	0	SeniorCitizen
6	Partner	Not have a Partner	0.752688	0	6.31	38.52	0	Partner
9	Dependents	Have a Dependents	0.197849	0	9.61	38.21	1	Dependents
12	PhoneService	Not have a PhoneService	0.318280	0	10.19	35.52	0	PhoneService
15	MultipleLines	Not have a MultipleLines	0.909677	0	7.31	37.50	0	MultipleLines
18	InternetServiceType	DSL	0.756989	0	7.67	44.10	1	DSL
19	InternetServiceType	No	0.243011	0	8.24	20.37	0	No
22	OnlineSecurity	Not have a OnlineSecurity	0.864516	0	7.32	36.76	0	OnlineSecurity
25	OnlineBackup	Not have a OnlineBackup	0.858065	0	7.16	37.12	0	OnlineBackup
28	DeviceProtection	Not have a DeviceProtection	0.864516	0	7.15	36.93	0	DeviceProtection
31	TechSupport	Not have a TechSupport	0.873118	0	7.36	36.85	0	TechSupport
34	StreamingTV	Not have a StreamingTV	0.870968	0	7.27	36.94	0	StreamingTV
37	StreamingMovies	Not have a StreamingMovies	0.862366	0	7.57	37.02	0	StreamingMovies
40	ContractType	Month-to-month	0.939785	0	6.59	38.62	1	Month-to-month
44	PaperlessBilling	Not have a PaperlessBilling	0.434409	0	7.83	35.76	0	PaperlessBilling
47	PaymentMethod	Mailed check	0.380645	0	5.02	35.08	1	Mailed check
51	InternetService	Not have a InternetService	0.243011	0	8.24	20.37	0	InternetService

Cluster 2 - Low Tenure and High Monthly Charges

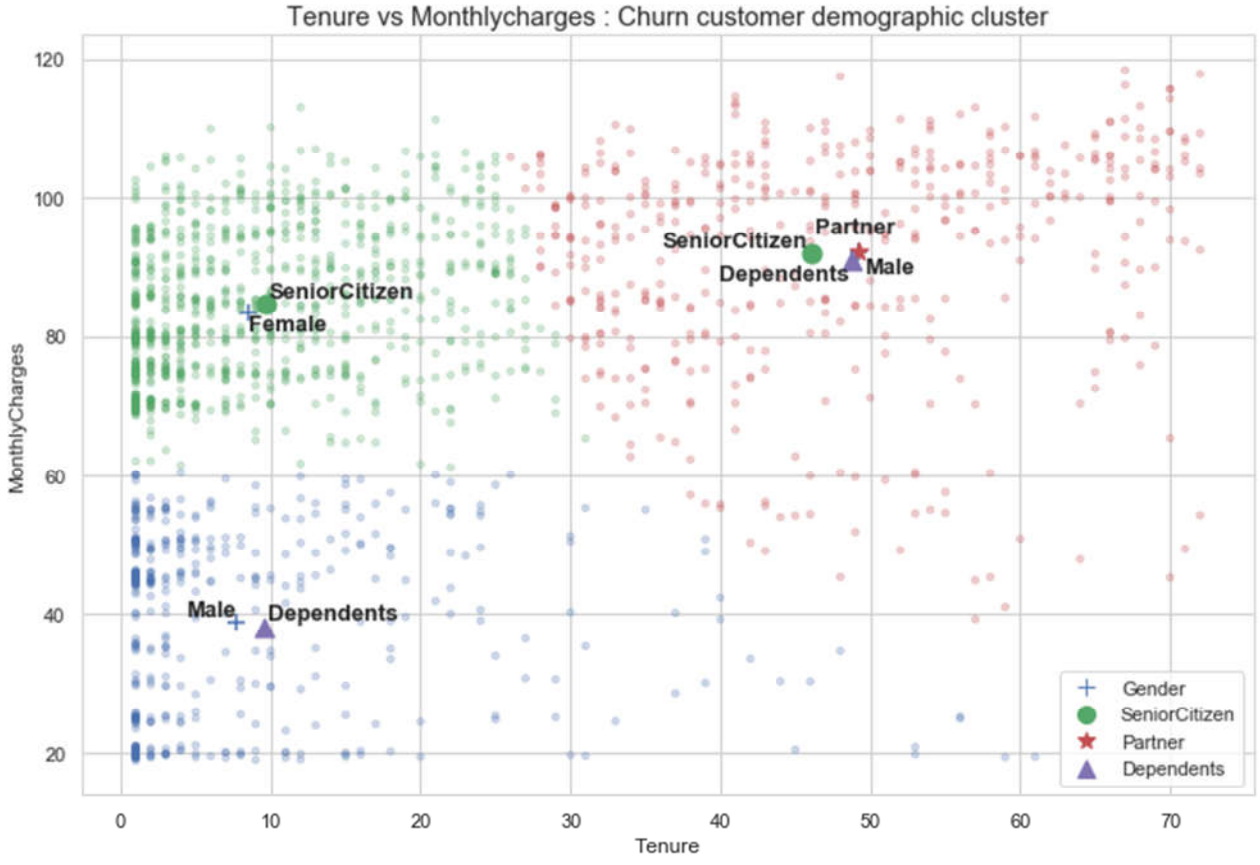
	Category	Label	Percentage	Cluster	Avg_Tenure	Avg_MonthlyCharges	Represent_in_graph	Label_in_graph
1	Gender	Female	0.531579	1	8.47	83.47	1	Female
4	SeniorCitizen	Have a SeniorCitizen	0.274737	1	9.67	84.56	1	SeniorCitizen
7	Partner	Not have a Partner	0.701053	1	7.94	83.28	0	Partner
10	Dependents	Not have a Dependents	0.860000	1	8.56	83.91	0	Dependents
13	PhoneService	Have a PhoneService	1.000000	1	8.81	83.80	1	PhoneService
16	MultipleLines	Have a MultipleLines	0.492632	1	10.57	87.30	1	MultipleLines
20	InternetServiceType	Fiber optic	0.956842	1	8.69	84.47	1	Fiber optic
23	OnlineSecurity	Not have a OnlineSecurity	0.892632	1	8.30	83.39	0	OnlineSecurity
26	OnlineBackup	Not have a OnlineBackup	0.776842	1	7.94	82.29	0	OnlineBackup
29	DeviceProtection	Not have a DeviceProtection	0.749474	1	8.07	81.48	0	DeviceProtection
32	TechSupport	Not have a TechSupport	0.883158	1	8.49	83.12	0	TechSupport
35	StreamingTV	Have a StreamingTV	0.451579	1	10.25	91.51	1	StreamingTV
38	StreamingMovies	Have a StreamingMovies	0.445263	1	10.50	91.67	1	StreamingMovies
41	ContractType	Month-to-month	0.983158	1	8.66	83.76	1	Month-to-month
45	PaperlessBilling	Have a PaperlessBilling	0.812632	1	9.03	84.26	1	PaperlessBilling
48	PaymentMethod	Electronic check	0.677895	1	8.46	84.32	1	Electronic check
52	InternetService	Have a InternetService	1.000000	1	8.81	83.80	1	InternetService

For example, in overall churn customer, percentage of female is 50.83%. And in cluster 2 (Low tenure and high monthly charges), percentage of female is 53.15%. Meaning female are more likely to leave company due to high monthly charges and lower tenure.

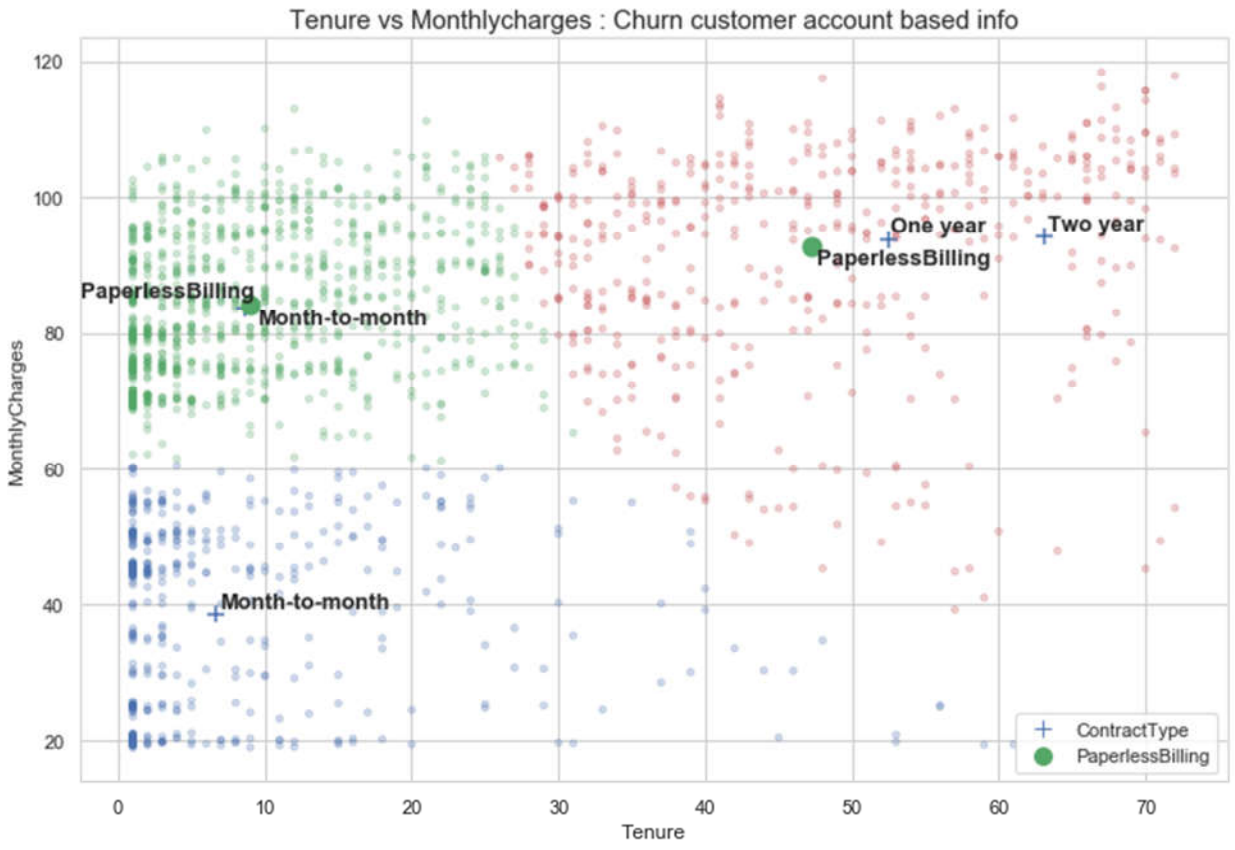
Cluster 3 - High Tenure and High Monthly Charges

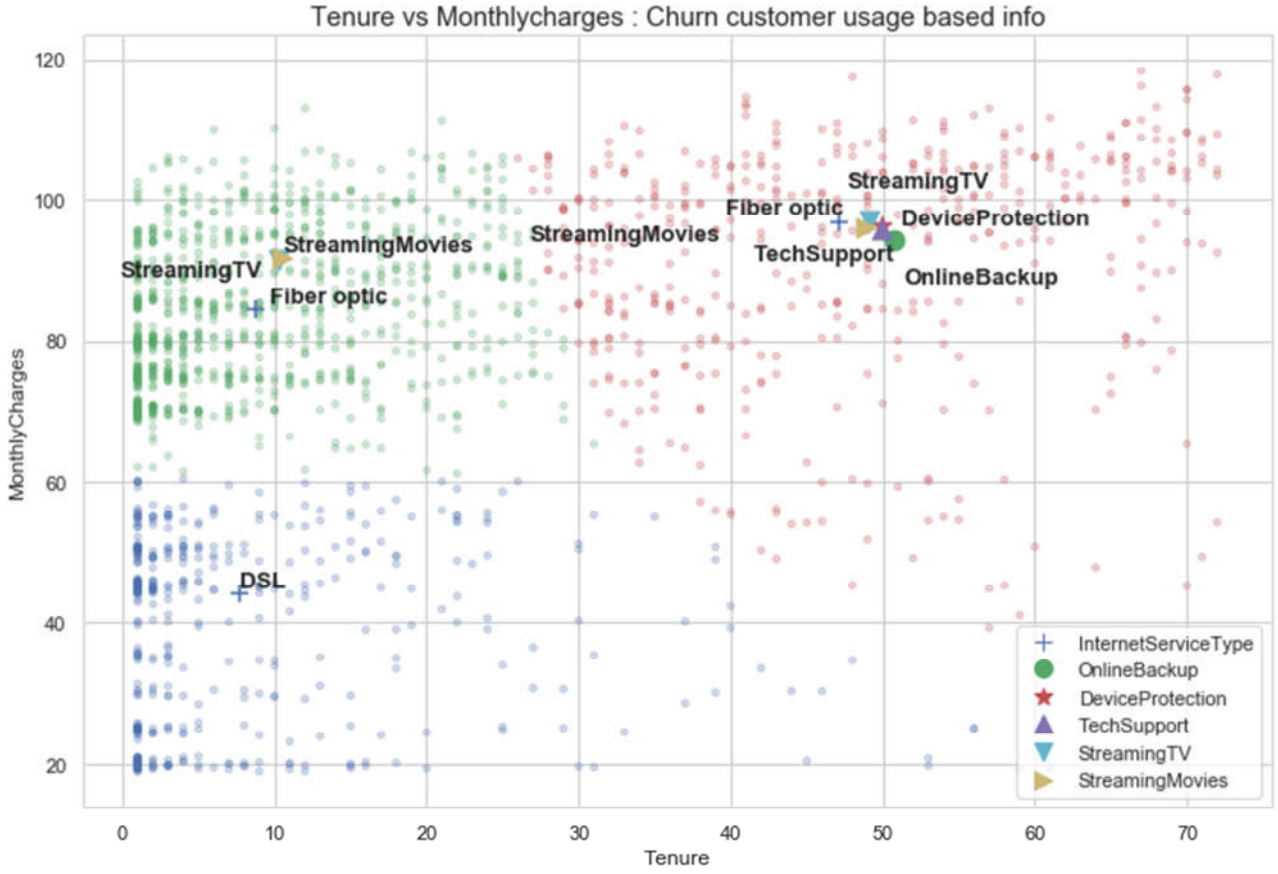
	Category	Label	Percentage	Cluster	Avg_Tenure	Avg_MonthlyCharges	Represent_in_graph	Label_in_graph
2	Gender	Male	0.519824	2	49.25	92.21	1	Male
5	SeniorCitizen	Have a SeniorCitizen	0.332599	2	46.03	92.03	1	SeniorCitizen
8	Partner	Have a Partner	0.594714	2	49.23	92.24	1	Partner
11	Dependents	Have a Dependents	0.222467	2	48.75	90.94	1	Dependents
14	PhoneService	Have a PhoneService	0.951542	2	47.28	93.93	1	PhoneService
17	MultipleLines	Have a MultipleLines	0.748899	2	48.33	96.74	1	MultipleLines
21	InternetServiceType	Fiber optic	0.854626	2	47.16	96.70	1	Fiber optic
24	OnlineSecurity	Have a OnlineSecurity	0.286344	2	50.43	93.32	1	OnlineSecurity
27	OnlineBackup	Have a OnlineBackup	0.539648	2	50.72	94.15	1	OnlineBackup
30	DeviceProtection	Have a DeviceProtection	0.537445	2	49.98	96.36	1	DeviceProtection
33	TechSupport	Have a TechSupport	0.308370	2	49.96	95.65	1	TechSupport
36	StreamingTV	Have a StreamingTV	0.715859	2	49.08	97.06	1	StreamingTV
39	StreamingMovies	Have a StreamingMovies	0.729075	2	48.89	96.10	1	StreamingMovies
42	ContractType	One year	0.279736	2	52.48	93.77	1	One year
43	ContractType	Two year	0.094714	2	63.14	94.23	1	Two year
46	PaperlessBilling	Have a PaperlessBilling	0.803965	2	47.25	92.77	1	PaperlessBilling
49	PaymentMethod	Bank transfer (automatic)	0.211454	2	50.73	93.62	1	Bank transfer (automatic)
50	PaymentMethod	Credit card (automatic)	0.215859	2	49.47	90.27	1	Credit card (automatic)
53	InternetService	Have a InternetService	1.000000	2	47.59	91.86	1	InternetService

Graphical representation of clusters



[22]


Tenure vs Monthlycharges : Churn customer account based info
MonthlyCharges
Tenure
One year
Two year
PaperlessBilling
PaperlessBilling
Month-to-month
Month-to-month
ContractType
PaperlessBilling


Tenure vs Monthlycharges : Churn customer usage based info
MonthlyCharges
Tenure
StreamingTV
Fiber optic
DeviceProtection
StreamingMovies
TechSupport
OnlineBackup
StreamingMovies
StreamingTV
Fiber optic
DSL
InternetServiceType
OnlineBackup
DeviceProtection
TechSupport
StreamingTV
StreamingMovies

7. Building Classification Models

Finally, our data is prepared and ready for modeling. The type of models we choose should be based on the type of problem which we are seeking to solve. In this case, our problem is a regression and classification one. We seek to establish a relationship between an output features (Churn), and identify the variables which impact this (eg. type of contract, customer tenure).

Here is the preprocessed data

```
#Data preprocessing

df_model = Data_transformation(raw_data)
df_model.head()
```

	CustomerID	Gender	SeniorCitizen	Partner	Dependents	Tenure	PhoneService	MultipleLines	InternetServiceType	OnlineSecurity	...	StreamingTV	Stream
0	7590-VHVEG	Female	0	1	0	1	0	0	DSL	0	...	0	
1	5575-GNVDE	Male	0	0	0	34	1	0	DSL	1	...	0	
2	3668-QPYBK	Male	0	0	0	2	1	0	DSL	1	...	0	
3	7795-CFOCW	Male	0	0	0	45	0	0	DSL	1	...	0	
4	9237-HQITU	Female	0	0	0	2	1	0	Fiber optic	0	...	0	

Normalized data is shown below

```
print(df_model_target)
df_model_feature.head()
```

```
[1 0 0 ... 1 0 0]
```

	Partner	Dependents	SeniorCitizen	PhoneService	MultipleLines	OnlineSecurity	OnlineBackup	DeviceProtection	TechSupport	StreamingTV	...	InternetSer
0	0	0	0	1	1	0	0	1	0	1	...	
1	0	1	0	0	0	0	1	0	1	1	...	
2	0	0	0	1	1	1	1	1	0	0	...	
3	0	0	0	1	0	0	0	0	0	0	...	
4	1	0	0	1	1	0	0	1	1	1	...	

5 rows × 28 columns

```
----------------------------
Original features shape,  (7043, 28)
Original target shape,  (7043,)
x train shape,  (5634, 28)
y train shape,  (5634,)
x test shape,  (1409, 28)
y test shape,  (1409,)
----------------------------
```

Then the normalized data is split into training and testing sets, as shown in the above picture.

Up sampling to balance data

Original Data

After up sampling train data

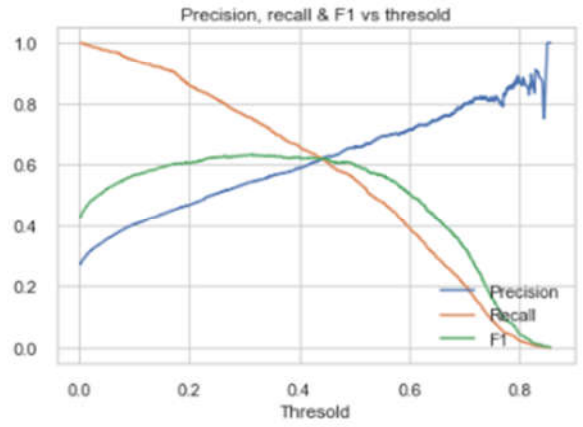
Here from the above distribution we have unbalanced data which is not good for creating effective models. To avoid this problem we will balance the data using **Up Sampling.** So after up sampling we can see the data is balanced.

We will be using the following models:

1. Logistic Regression
2. Random Forest

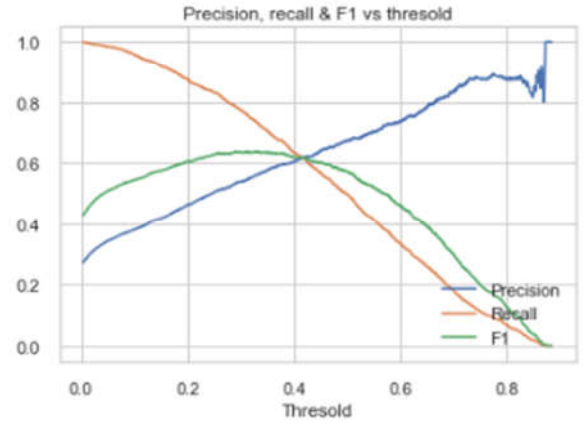
Building models using original data

```
================================================
Logistic Regression
************************************************
Original data
accuracy  score: 0.8029802947513085
precision  score: 0.6554033430233874
recall  score: 0.5525204872646733
f1  score: 0.5994931890474751
neg_log_loss  score: -0.42167012577392793
```




```
When F1 score is max
      Precisions  Recalls       F1  Threshold
3454    0.543969  0.75266  0.63152   0.312124
```

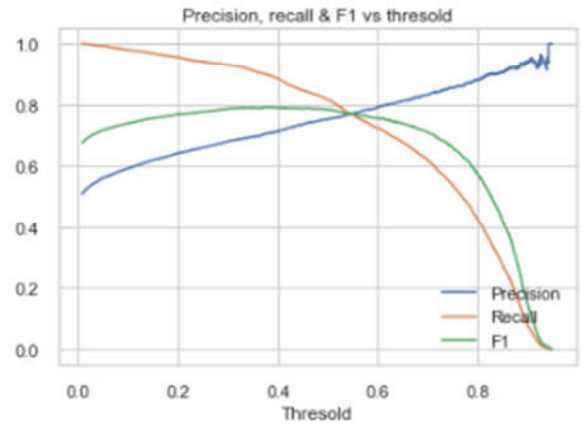
```
================================================
Random Forest
************************************************
Original data
accuracy  score: 0.8022702879900898
precision  score: 0.676539191623021
recall  score: 0.49868438538205984
f1  score: 0.5737256877307021
neg_log_loss  score: -0.417344488752713
```




```
When F1 score is max
      Precisions   Recalls       F1  Threshold
3544    0.565039  0.730718  0.637286   0.333171
```

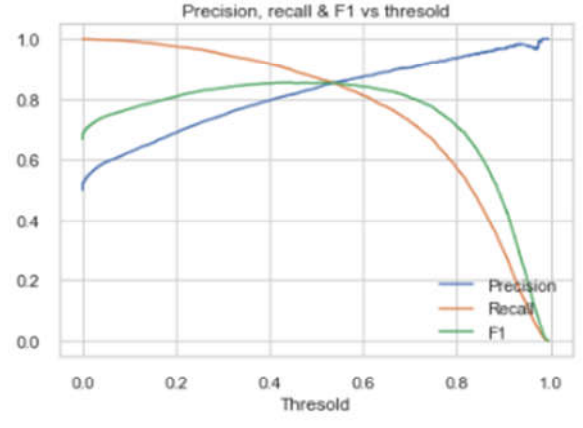
Building models using up sampled data

```
================================================
Logistic Regression
************************************************
Upsampled data
accuracy  score: 0.7734866828087167
precision  score: 0.7521956404051224
recall  score: 0.8157384987893461
f1  score: 0.7825554099098079
neg_log_loss  score: -0.4739315040436261
```




```
When F1 score is max
      Precisions   Recalls       F1  Threshold
2848     0.70669  0.897821  0.790871    0.37869
```

```
================================================
Random Forest
************************************************
Upsampled data
accuracy  score: 0.8487893462469733
precision  score: 0.8344879723826903
recall  score: 0.8685230024213075
f1  score: 0.8465652202136267
neg_log_loss  score: -0.3600175223998228
```



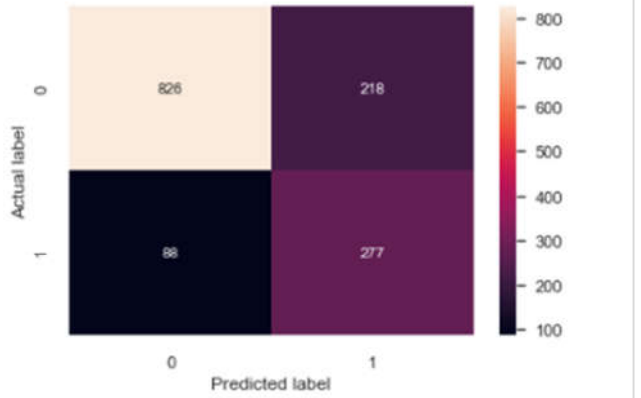

```
When F1 score is max
      Precisions   Recalls       F1  Threshold
2579    0.809835  0.901211  0.853083   0.435021
```

Results of confusion matrix on original data

********** Logistic Regression **********

	precision	recall	f1-score	support
0	0.90	0.79	0.84	1044
1	0.56	0.76	0.64	365
micro avg	0.78	0.78	0.78	1409
macro avg	0.73	0.78	0.74	1409
weighted avg	0.81	0.78	0.79	1409

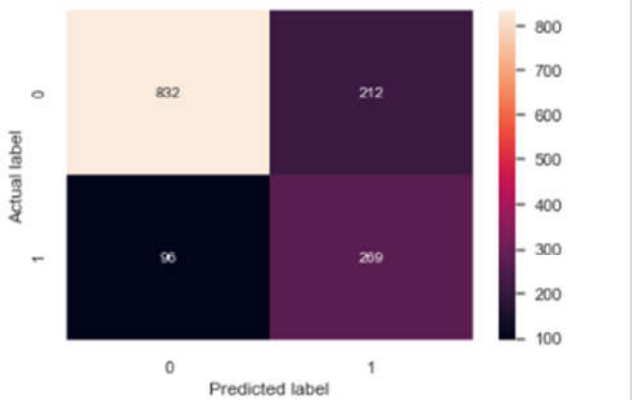
****
Log loss score 0.4
****
Confusion matrix



********** Random Forest **********

	precision	recall	f1-score	support
0	0.90	0.80	0.84	1044
1	0.56	0.74	0.64	365
micro avg	0.78	0.78	0.78	1409
macro avg	0.73	0.77	0.74	1409
weighted avg	0.81	0.78	0.79	1409

****
Log loss score 0.4
****
Confusion matrix

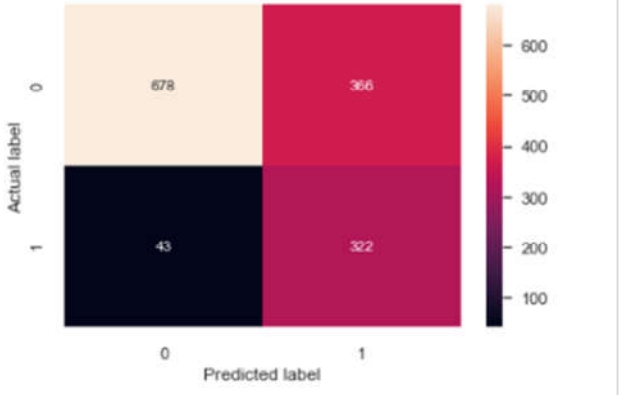


Results of confusion matrix on up sampled data

********** Logistic Regression **********

	precision	recall	f1-score	support
0	0.94	0.65	0.77	1044
1	0.47	0.88	0.61	365
micro avg	0.71	0.71	0.71	1409
macro avg	0.70	0.77	0.69	1409
weighted avg	0.82	0.71	0.73	1409

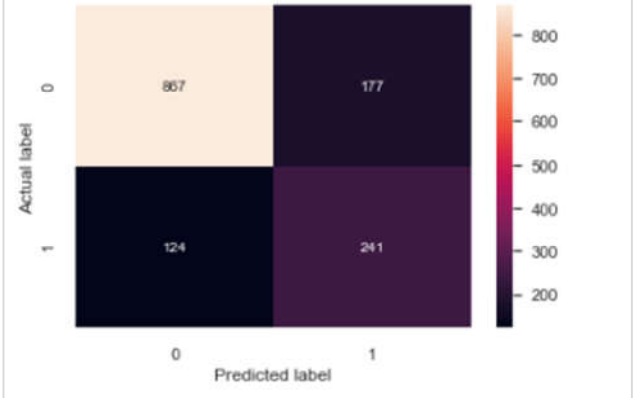
****
Log loss score 0.46
****
Confusion matrix



********** Random Forest **********

	precision	recall	f1-score	support
0	0.87	0.83	0.85	1044
1	0.58	0.66	0.62	365
micro avg	0.79	0.79	0.79	1409
macro avg	0.73	0.75	0.73	1409
weighted avg	0.80	0.79	0.79	1409

****
Log loss score 0.46
****
Confusion matrix



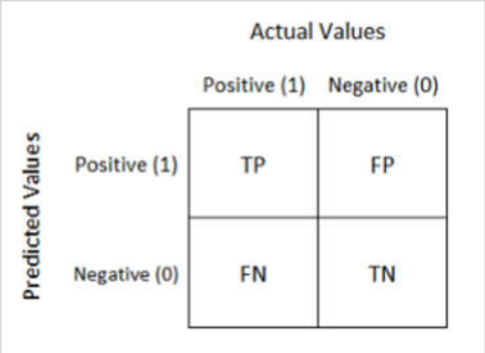
8. Evaluation and Comparison of Models

Evaluation Metrics

We will compare the two models by analyzing the following metrics:

- Classification Accuracy
- Logarithmic Loss
- Confusion Matrix
- Precision
- Recall
- F1 Score

Confusion Matrix *forms the basis for the other types of metrics.* It is extremely useful for measuring Recall, Precision, and Accuracy.



There are 4 important terms:

- True Positives (**TP**): The cases which are predicted YES and the actual output was also YES.
- True Negatives (**TN**): The cases which are predicted NO and the actual output was NO.
- False Positives (**FP**): The cases which are predicted YES and the actual output was NO.
- False Negatives (**FN**): The cases which are predicted NO and the actual output was YES.

Classification accuracy is what we usually mean, when we use the term accuracy. It is the ratio of number of correct predictions to the total number of input samples.

		Predicted/Classified	
		Negative	Positive
Actual	Negative	998	0
	Positive	1	1

It works well only if there are equal number of samples belonging to each class. For example, consider that there are 98% samples of class A and 2% samples of class B in our training set. Then our model can easily get **98% training accuracy** by simply predicting every training sample belonging to class A.

$$Accuracy = \frac{TruePositive + TrueNegative}{TotalSample}$$

Precision: It is implied as the measure of the correctly identified positive cases from all the predicted positive cases. Thus, it is useful when the cost of False Positives is high.

$$Precision = \frac{TruePositive}{(TruePositive\ + FalsePositive)}$$

Recall: It is the measure of the correctly identified positive cases from all the actual positive cases. It is important when the cost of False Negatives is high.

$$Recall = \frac{TruePositive}{(TruePositive\ + FalseNegative)}$$

F1 Score:-This is the harmonic mean of Precision and Recall and gives a better measure of the incorrectly classified cases than the Accuracy Metric.

$$F1 - Score = 2 * \frac{(Precision\ *\ Recall)}{(Precision + Recall)}$$

F1 Score is needed when you want to seek a balance between Precision and Recall. We have previously seen that accuracy can be largely contributed by a large number of True Negatives which in most business circumstances, we do not focus on much whereas False Negative and False Positive usually has business costs (tangible & intangible) thus F1 Score might be a better measure to use if we need to seek a balance between Precision and Recall and there is an uneven class distribution (large number of Actual Negatives).

Logarithmic Loss: - is indicative of how close the prediction probability is to the corresponding actual/true value (0 or 1 in case of binary classification). The more the predicted probability diverges from the actual value, the higher is the log-loss value.

$$Logloss_i = -[y_i \, ln \, p_i + (1 - y_i) \, ln(1 - p_i)]$$

Where *i* is the given observation/record, *y* is the actual/true value, *p* is the prediction probability, and *ln* refers to the natural logarithm (logarithmic value using base of *e*) of a number.

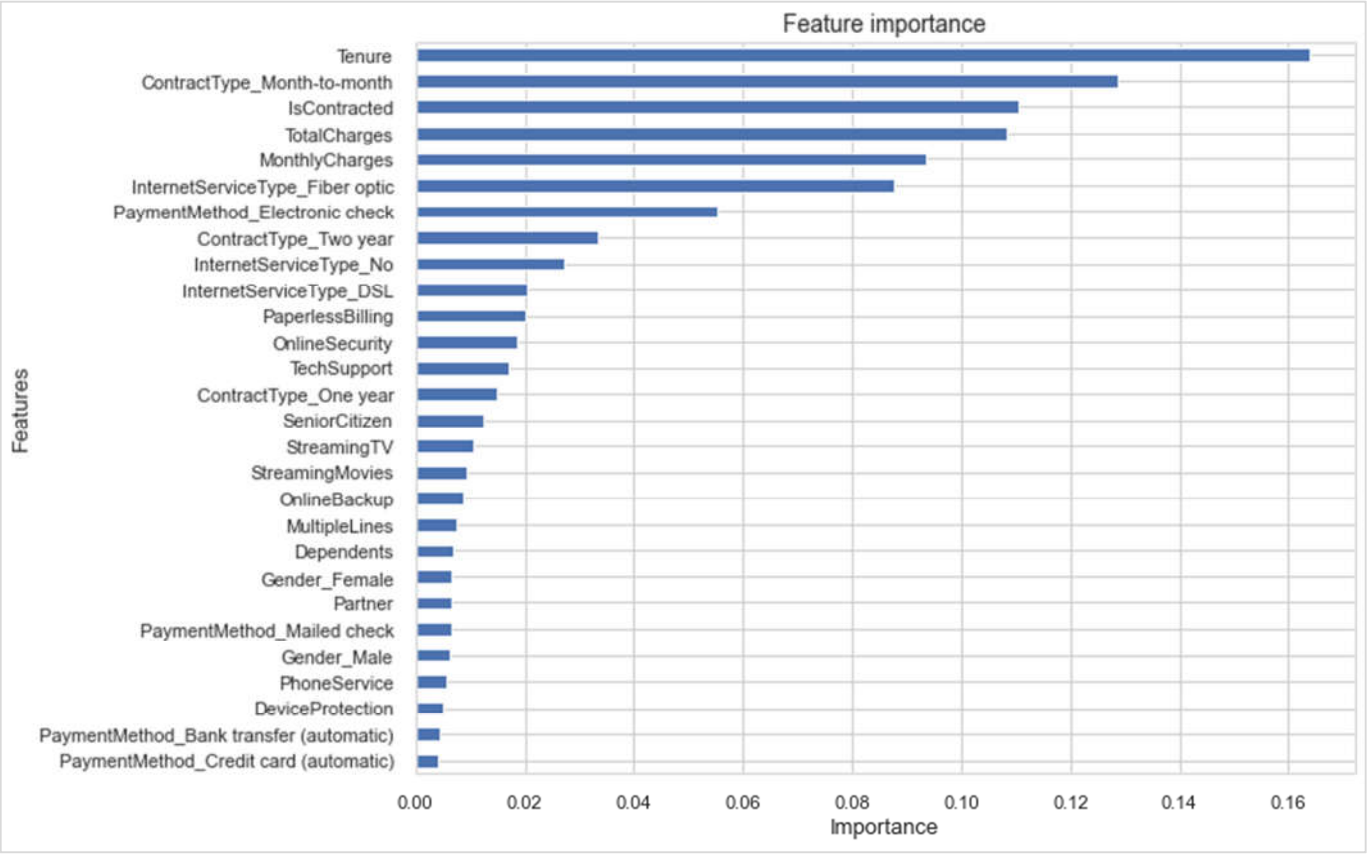
Conclusion on Model selection
From the analysis results, we have

Original Data			**Up Sampled Data**	
Metrics	Logistic Regression	Random Forest	Logistic Regression	Random Forest
Accuracy	0.803	0.802	0.773	0.849
Precision	0.655	0.676	0.752	0.834
Recall	0.553	0.498	0.816	0.869
F1-Score	0.599	0.574	0.783	0.847
Log loss	0.4	0.4	0.46	0.46

As we can see from the above summary table, using the original data which is unbalanced the two models have similar results in accuracy, F1-Score and log loss. After up sampling which means using a balanced data the accuracy of regression Model decreased and other metrics increased, but for Random Forest Model accuracy increased and other metrics also increased.

So we can conclude that the Random Forest model performed well on the original and up sampled data. We will consider this Model as our base model for making predictions and analyzing the feature importance.

Feature importance



Based on feature importance graph, it tells that Totalcharges, Tenure, Monthly charges, Contract type, Payment method, Internet sevice type, PaperlessBilling are some important features to predict churn customers.

Making Predictions

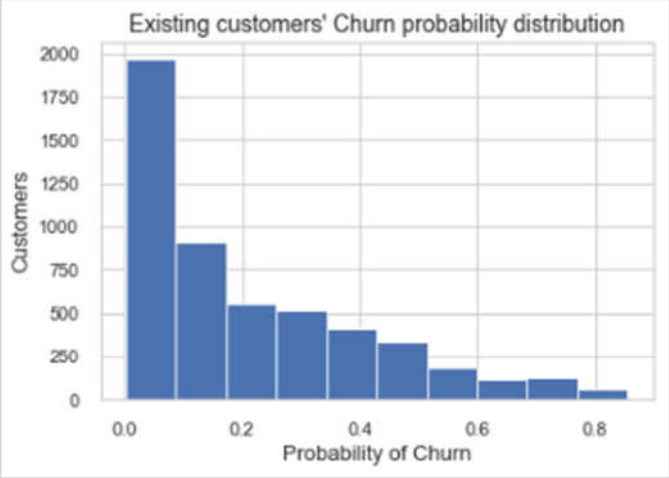
Here is the result of the prediction for test data

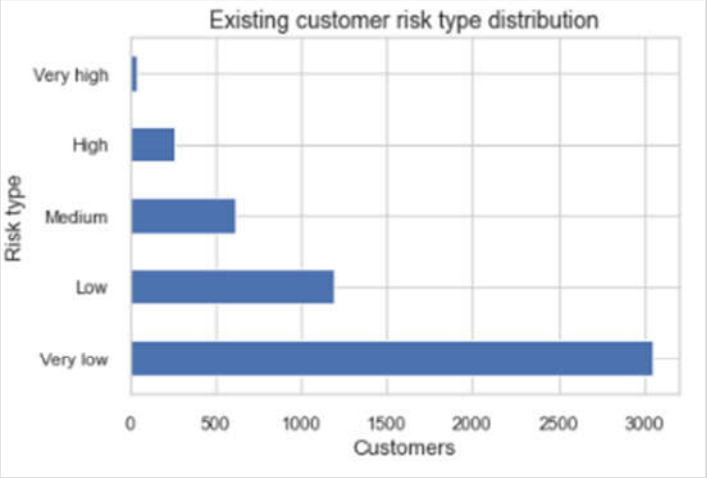
```
churn_customer_prob = pd.DataFrame({'Predicted_proba': clsRF.predict_proba(x_test)[:][:,1]
                                    , 'predicted': clsRF.predict(x_test)
                                    , 'Actual Values':y_test})
print(churn_customer_prob)
```

```
      Predicted_proba  predicted  Actual Values
0            0.041143          0              0
1            0.110440          0              0
2            0.042984          0              0
3            0.391750          0              0
4            0.066827          0              0
...               ...        ...            ...
1404         0.827567          1              0
1405         0.533259          1              1
1406         0.547218          1              1
1407         0.437020          0              1
1408         0.038043          0              0

[1409 rows x 3 columns]
```

Retention Plans

Retention strategies are policies and **plans** that organizations follow to reduce employee turnover and attrition and ensure employees are engaged and productive long-term.





We can provide retention plans to high risk and very high risk type customers.

Conclusions

Based on our analysis we can offer the final marketing strategies to the telecom company as follows:

1. Target new customers with 1 year or 2-year contract with special incentives, since getting new customers to sign a long-term contract will make them less likely to leave the company.

2. Set special bundle package offers to new customers in order to minimize future chance of churn since we have found out that customers with Fiber Optic internet are leaving the company at a higher rate and hence it would be beneficial if we offer them a bundled internet package.

3. Minimize total charges levied on customers with less tenure, since we found that very high costs levied on customers is a major factor leading customers to leave the company.

4. Offer high discount packages to customers most likely to churn as of today in order to stop the immediate attrition of customers and retain its existing customer base.

Link to Code in Google Colab

https://drive.google.com/file/d/1In5hYiKt13hUMx9mszhDeHR2ZJ9PhC4n/view?usp=sharing

References

- https://towardsdatascience.com/machine-learning-algorithms-part-9-k-means-example-in-python-f2ad05ed5203
- https://www.ibm.com/communities/analytics/watson-analytics-blog/predictive-insights- in-the-telco-customer-churn-data-set/
- Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. An Introduction to Statistical Learning : with Applications in R. New York :Springer, 2013. Print.
- VanderPlas, J. (2016). Python Data Science Handbook. O'Reilly Media, Inc.
- Beazley, D. and B.K. Jones (2013). Python cookbook, 3rd Edition. O'Reilly Media, Inc.
- https://www.kaggle.com/pavanraj159/telecom-customer-churn-prediction
- https://www.kaggle.com/bandiatindra/telecom-churn-prediction