TELECOM CUSTOMER CHURN ANALYSIS

Pragati Mishra, Ishan Jain, Vaibhav Shrivastava, Ashwin Kondapalli

May 3, 2019

Table of contents

Contents

1 Executive Summary

stuff to be added - by Ashwin

2 Project background

Customer attrition, also known as customer churn, customer turnover, or customer defection, is the loss of clients or customers. Telephone service companies, Internet service providers, pay TV companies and insurance firms often use customer attrition 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.

3 Data Description

Data being used: We will be working on the "Telco Customer Churn" data set taken from IBM Watson Analytics community https://community.watsonanalytics.com/resources/. IBM Watson Analytics team posted this dataset of a telecommunications company which is troubled by the number of customers leaving their landline business for other competitors. They need to understand who is leaving. This data set provides information to help us predict customer behavior in order to retain customers. We can analyze all relevant customer data and develop focused customer retention programs.

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. Some of the columns from dataset are defined below:

customerID: Customer ID

gender: Whether the customer is a male or a female

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	...	DeviceProtection	TechSupp
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 x 21 columns

Rows and columns

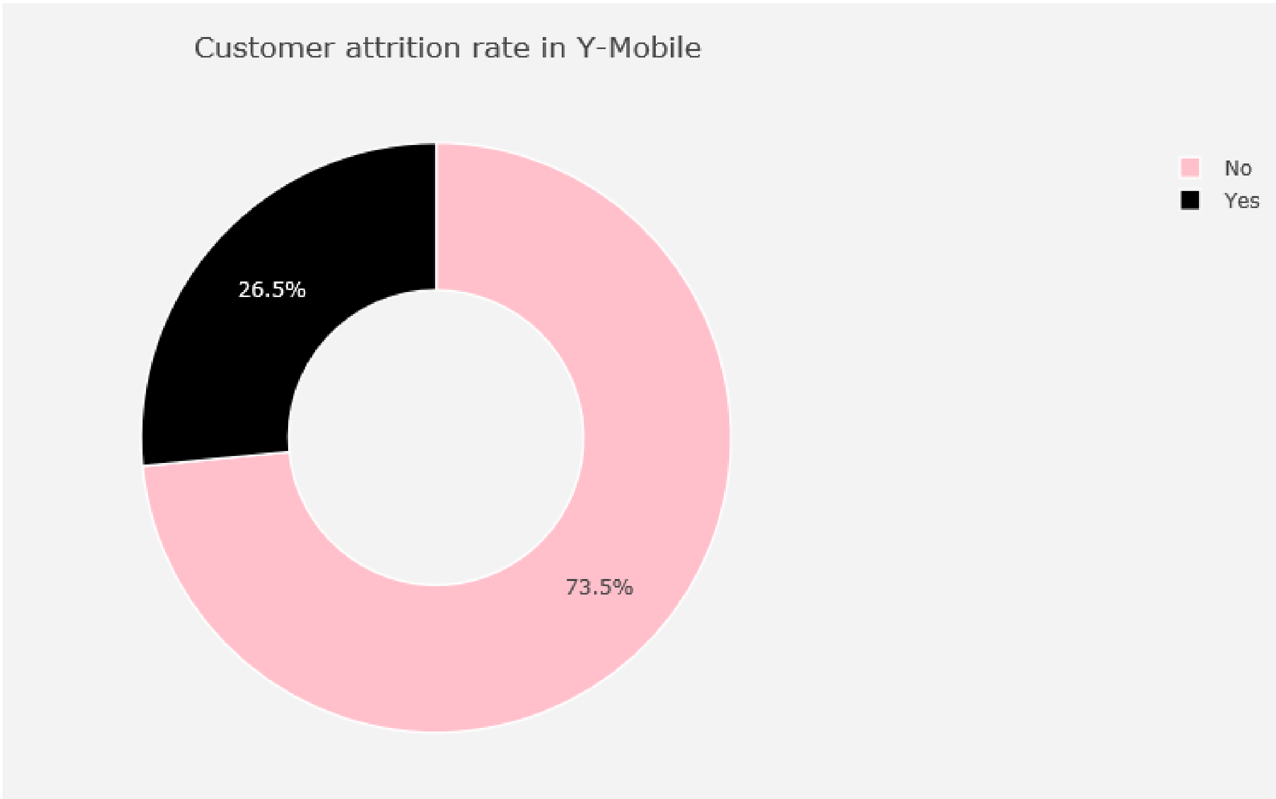
```
(7043, 21)
```

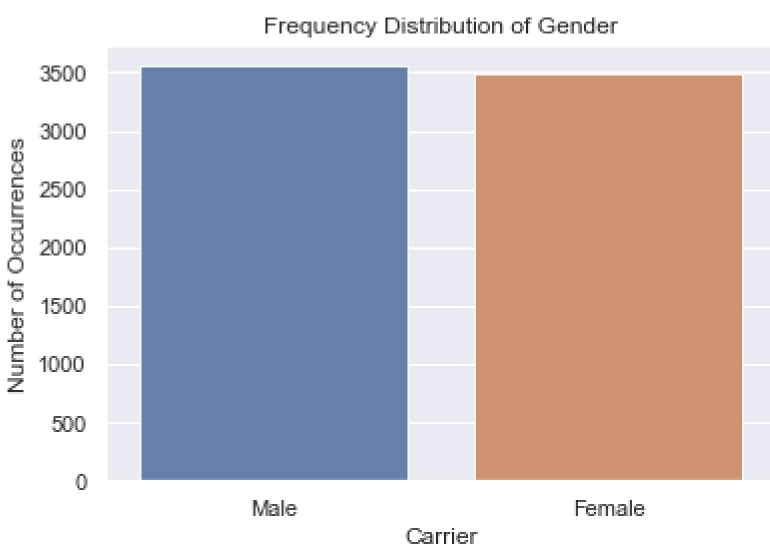
	SeniorCitizen	tenure	MonthlyCharges
count	7043.000000	7043.000000	7043.000000
mean	0.162147	32.371149	64.761692
std	0.368612	24.559481	30.090047
min	0.000000	0.000000	18.250000
25%	0.000000	9.000000	35.500000
50%	0.000000	29.000000	70.350000
75%	0.000000	55.000000	89.850000
max	1.000000	72.000000	118.750000

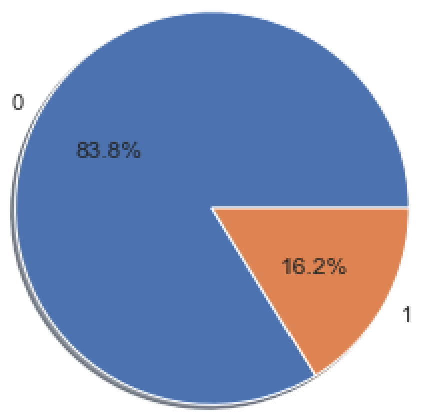
4 Exploratory Data Analysis

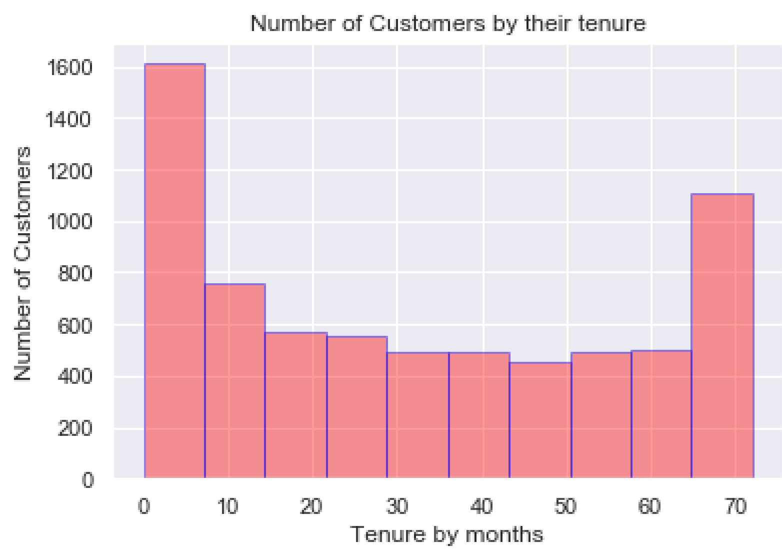
4.1 Frequency Distribution of Senior Citizen

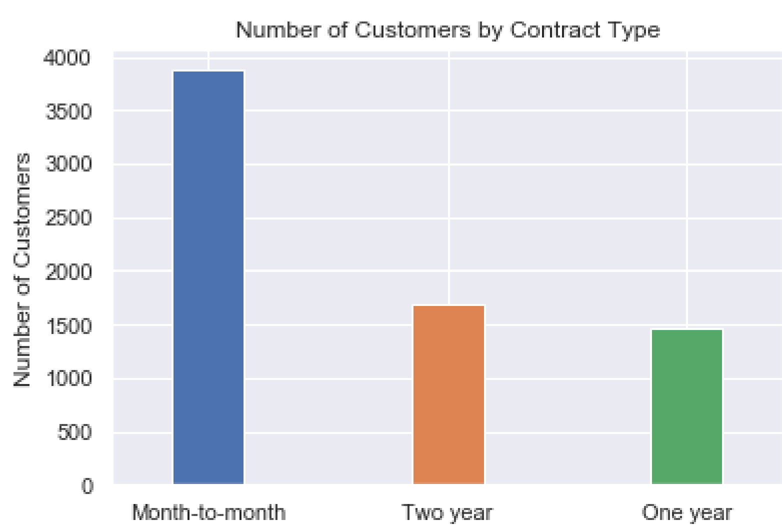
- xxxxxxx
- xxxxxx
- xxxxxxx

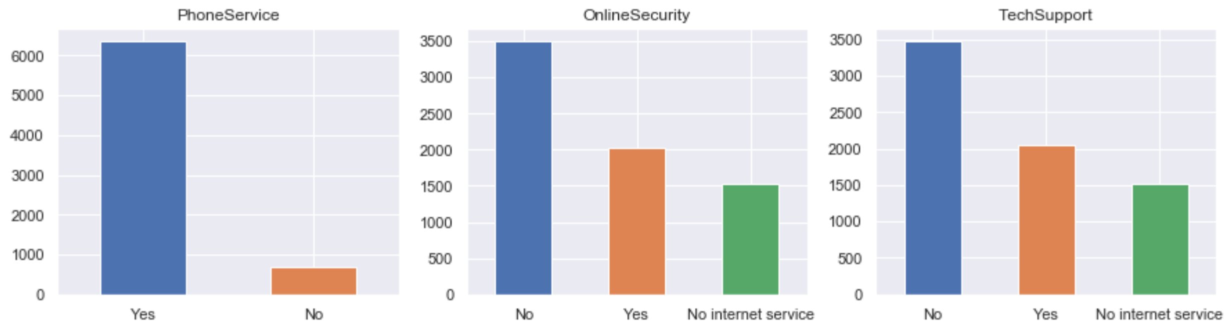

Customer attrition rate in Y-Mobile
26.5%
73.5%
No
Yes

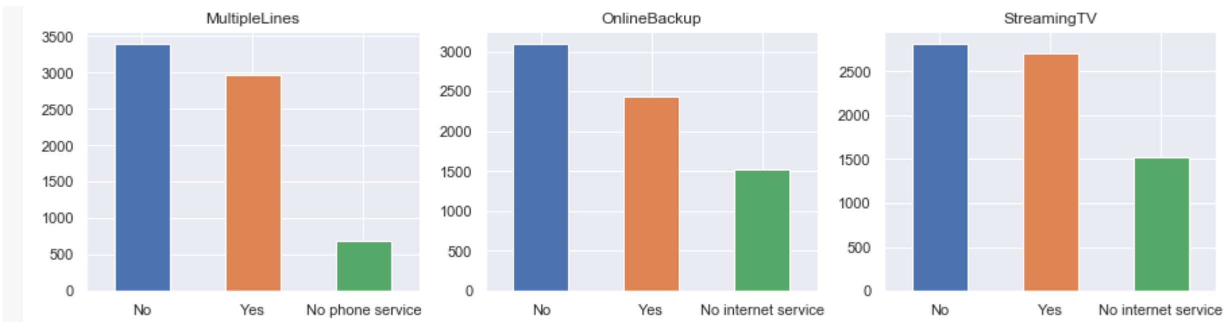

Frequency Distribution of Gender
Number of Occurrences
3500
3000
2500
2000
1500
1000
500
0
Male
Female
Carrier

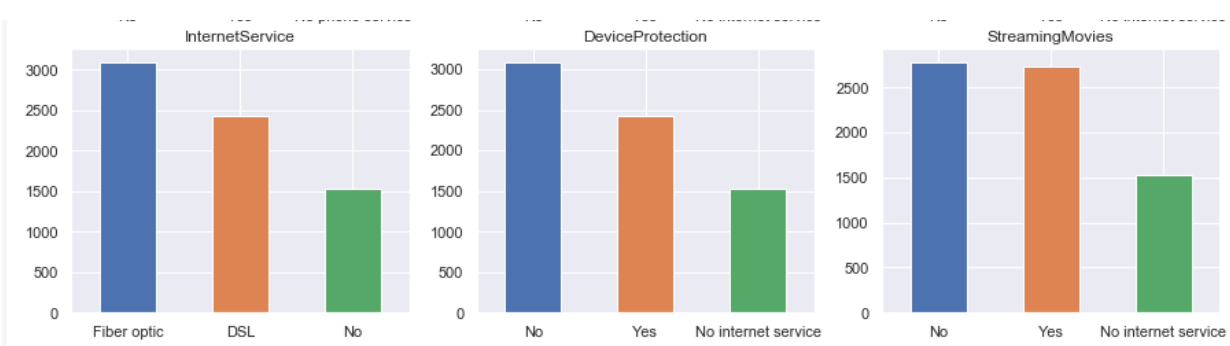

0
83.8%
16.2%
1


Number of Customers by their tenure
Number of Customers
Tenure by months
0
200
400
600
800
1000
1200
1400
1600
0
10
20
30
40
50
60
70


Number of Customers by Contract Type
Number of Customers
0
500
1000
1500
2000
2500
3000
3500
4000
Month-to-month
Two year
One year

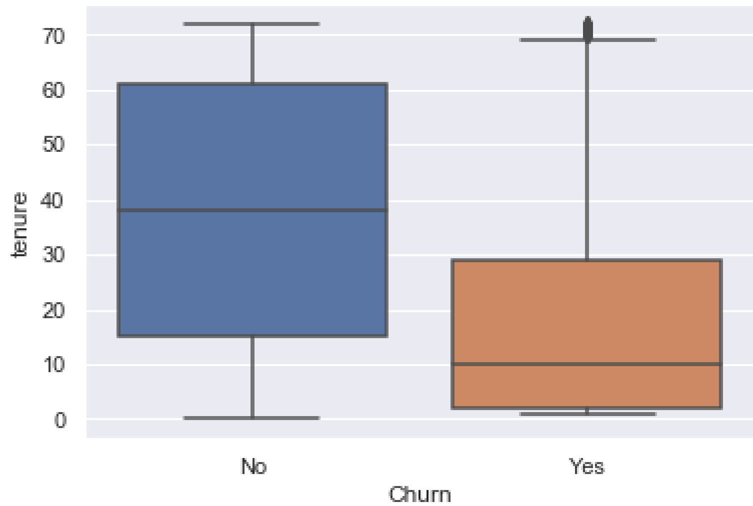

PhoneService
0
1000
2000
3000
4000
5000
6000
Yes
No
OnlineSecurity
0
500
1000
1500
2000
2500
3000
3500
No
Yes
No internet service
TechSupport
0
500
1000
1500
2000
2500
3000
3500
No
Yes
No internet service

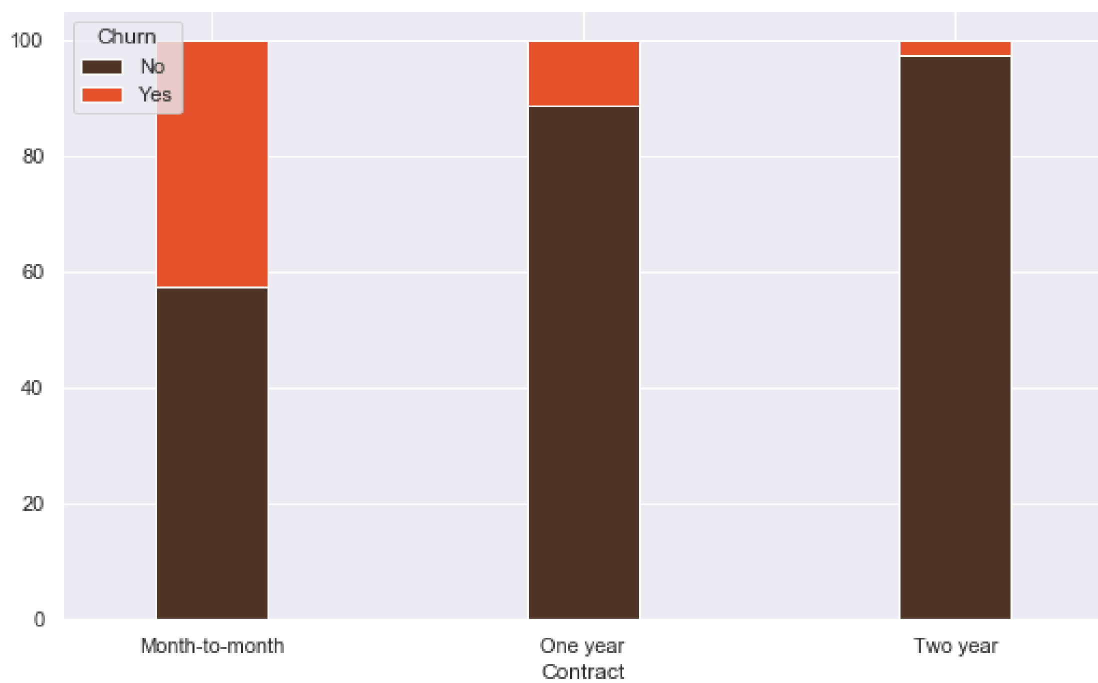

MultipleLines
0
500
1000
1500
2000
2500
3000
3500
No
Yes
No phone service
OnlineBackup
0
500
1000
1500
2000
2500
3000
No
Yes
No internet service
StreamingTV
0
500
1000
1500
2000
2500
No
Yes
No internet service

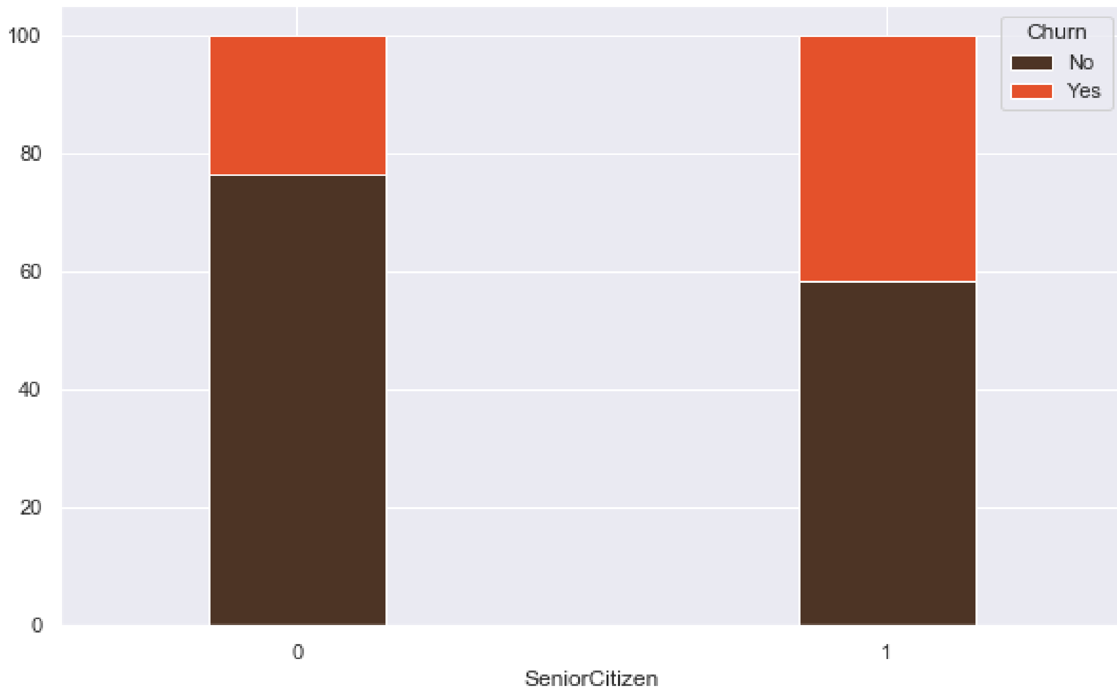

InternetService
0
500
1000
1500
2000
2500
3000
Fiber optic
DSL
No
DeviceProtection
0
500
1000
1500
2000
2500
3000
No
Yes
No internet service
StreamingMovies
0
500
1000
1500
2000
2500
No
Yes
No internet service

```
#attrition rate with tenure
sns.boxplot(x = customer.Churn, y = customer.tenure)
```

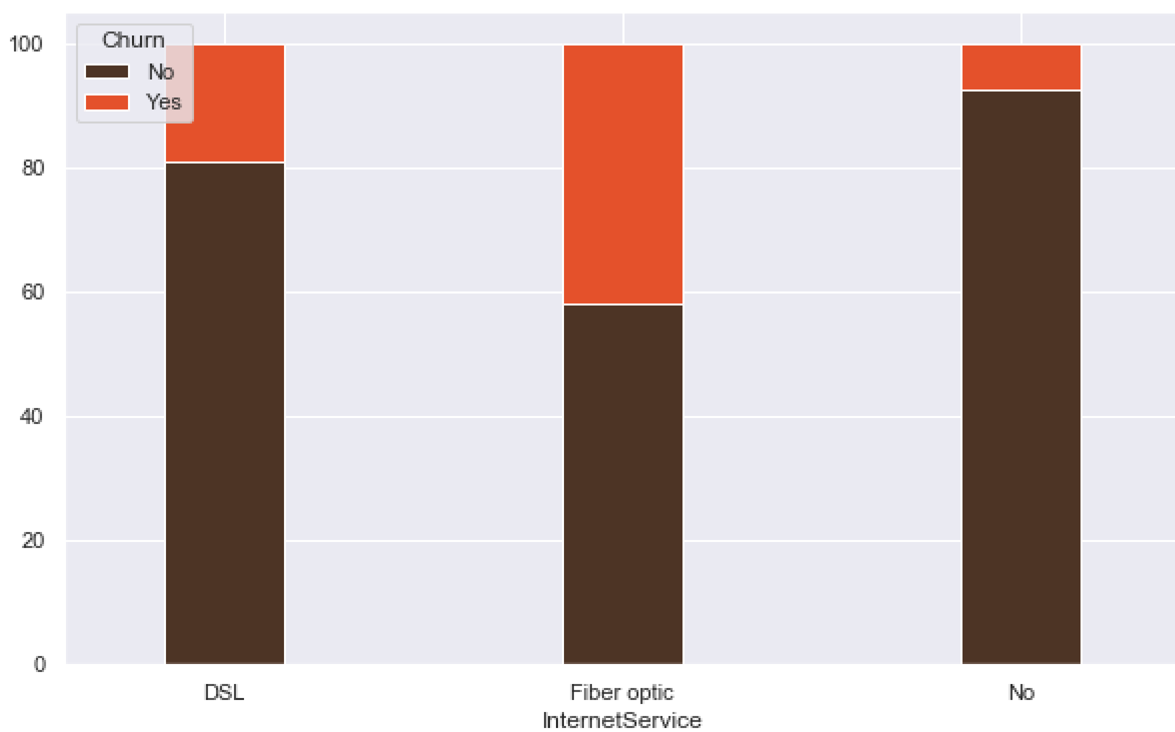
```
<matplotlib.axes._subplots.AxesSubplot at 0x2a12568cb38>
```



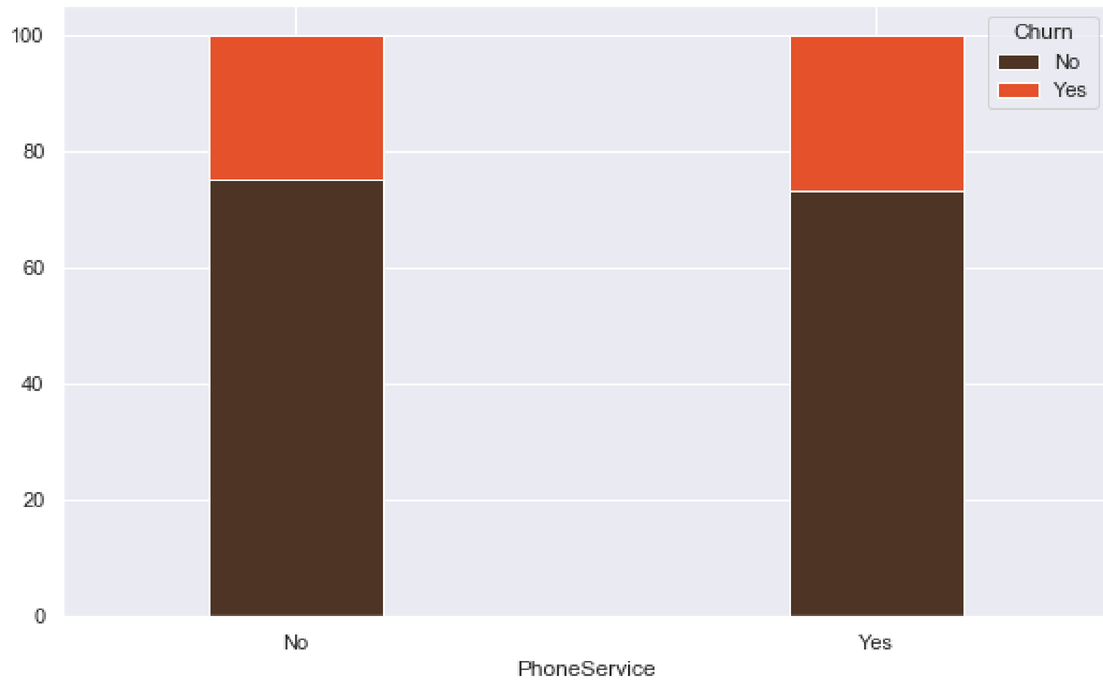


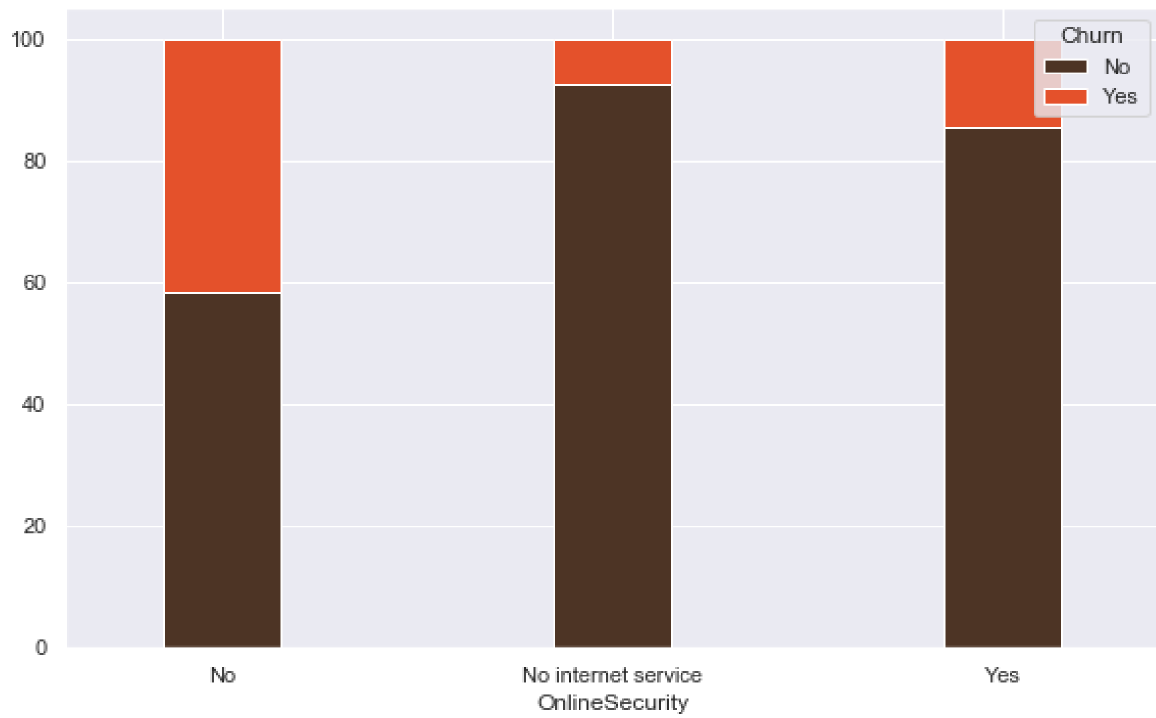


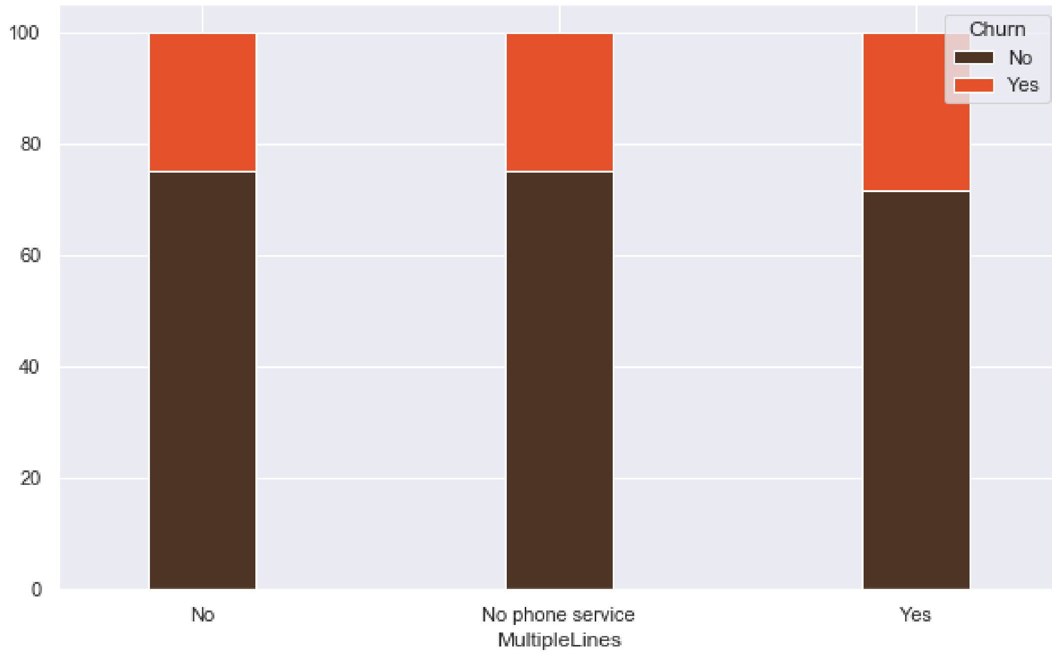
Churn
No
Yes
100
80
60
40
20
0
0
1
SeniorCitizen



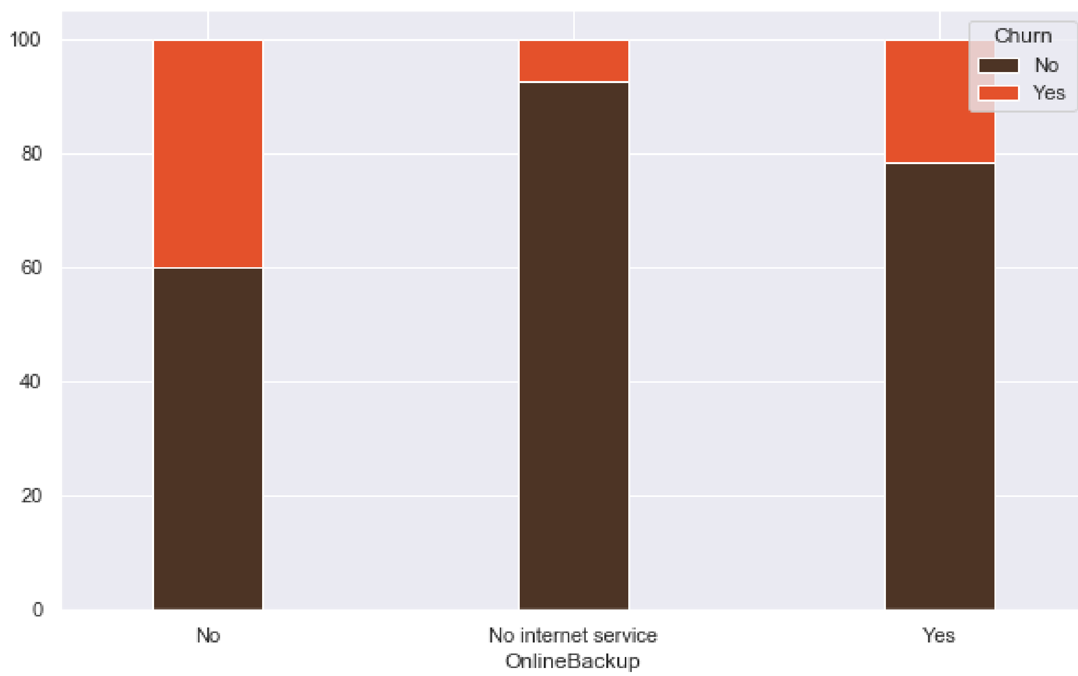
Churn
No
Yes
100
80
60
40
20
0
DSL
Fiber optic
No
InternetService


100
80
60
40
20
0
Churn
No
Yes
No
Yes
PhoneService

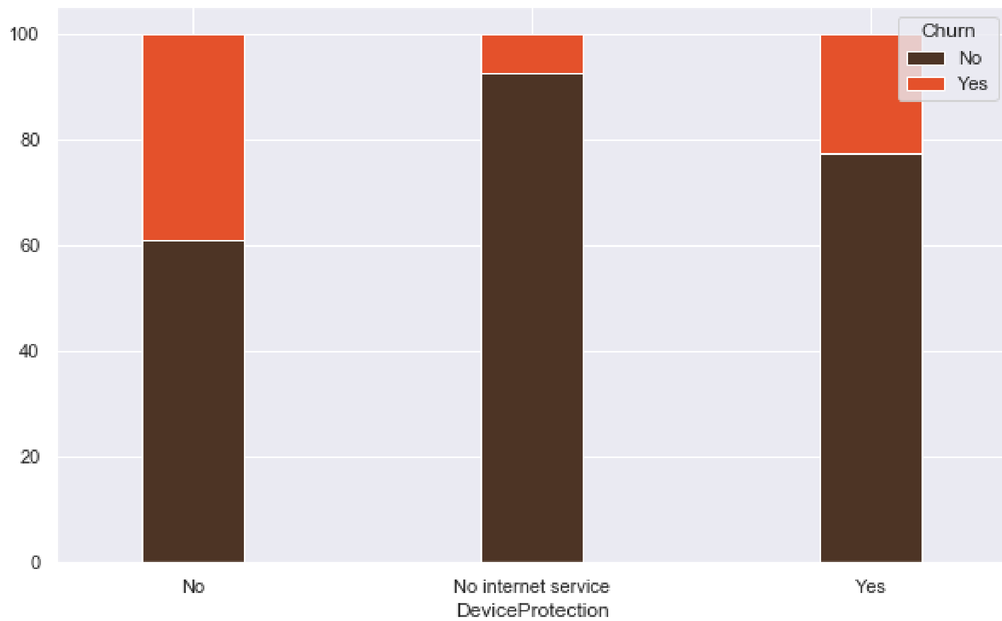

100
80
60
40
20
0
Churn
No
Yes
No
No internet service
Yes
OnlineSecurity

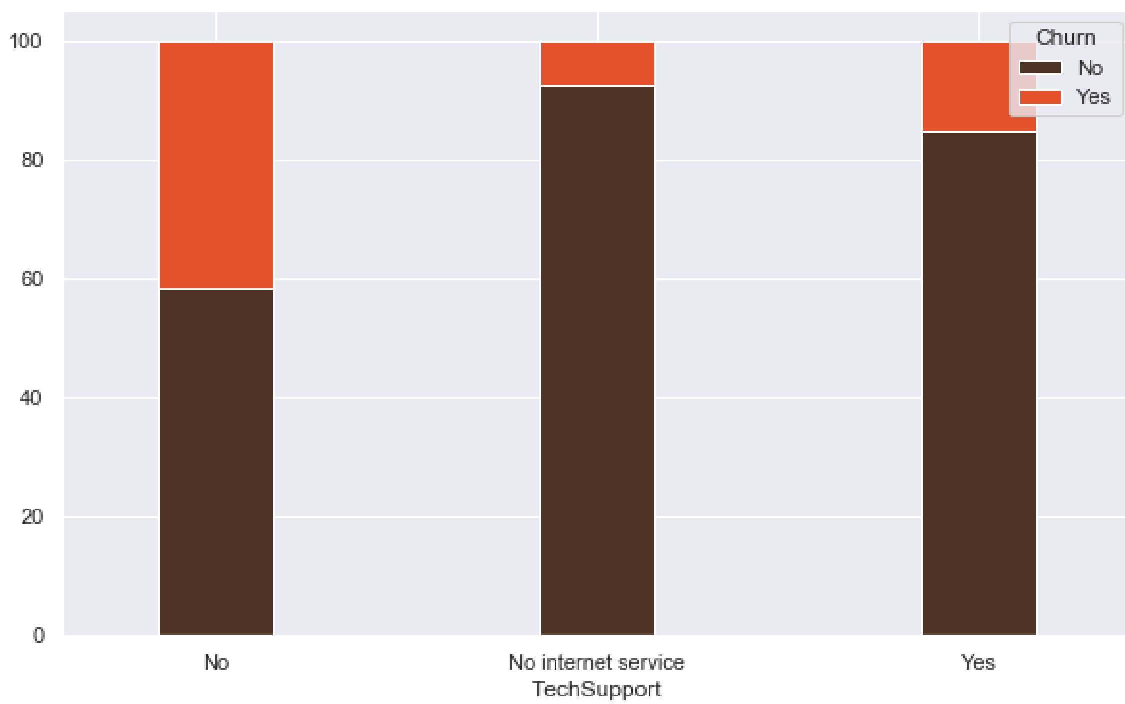


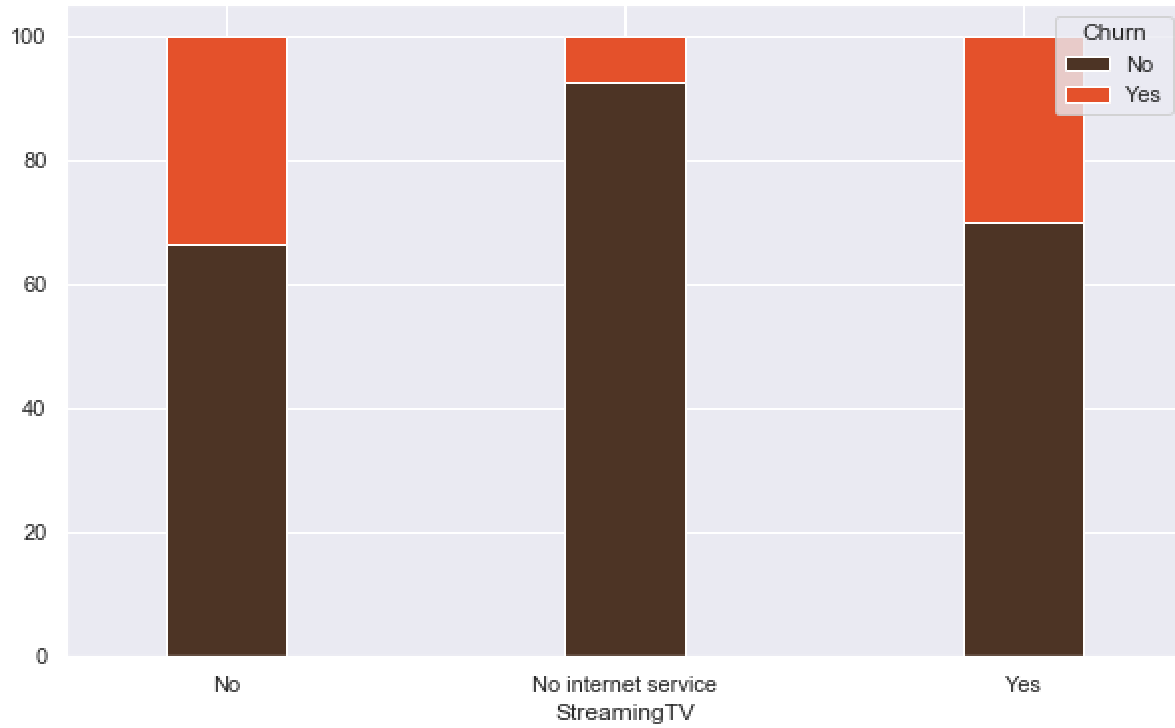
100
80
60
40
20
0
Churn
No
Yes
No
No phone service
Yes
MultipleLines



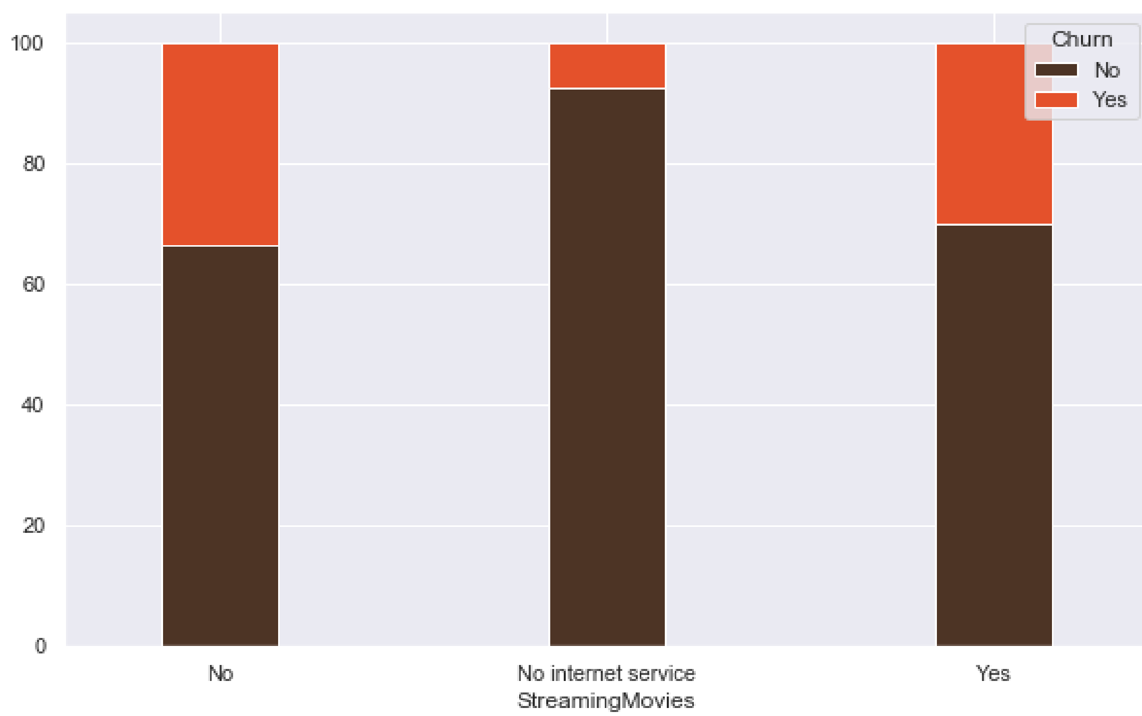
100
80
60
40
20
0
Churn
No
Yes
No
No internet service
Yes
OnlineBackup

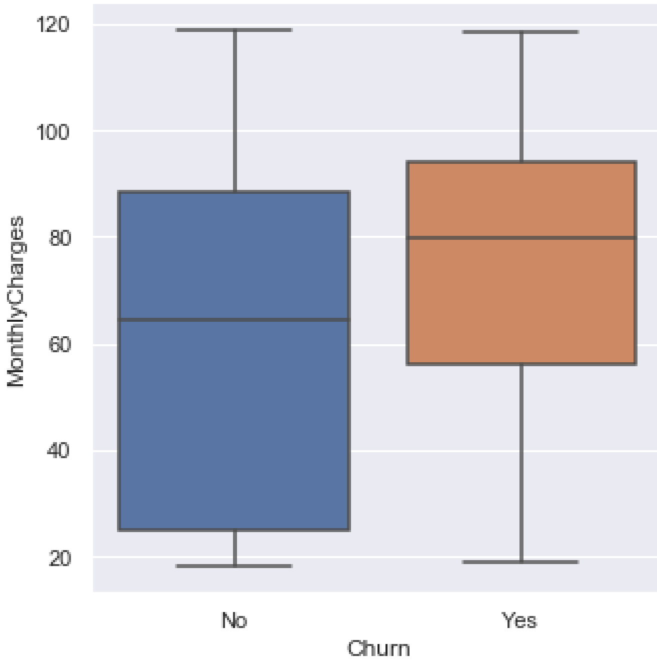

100
80
60
40
20
0
Churn
No
Yes
No
No internet service
Yes
DeviceProtection

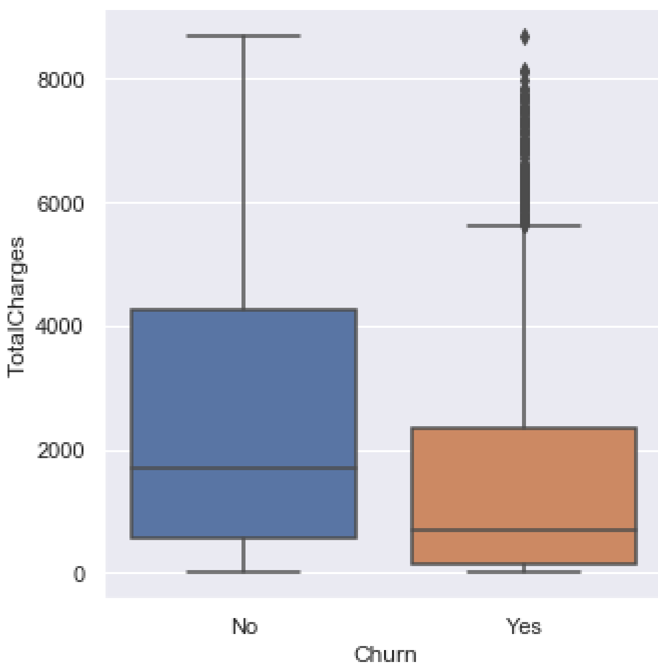

100
80
60
40
20
0
Churn
No
Yes
No
No internet service
Yes
TechSupport

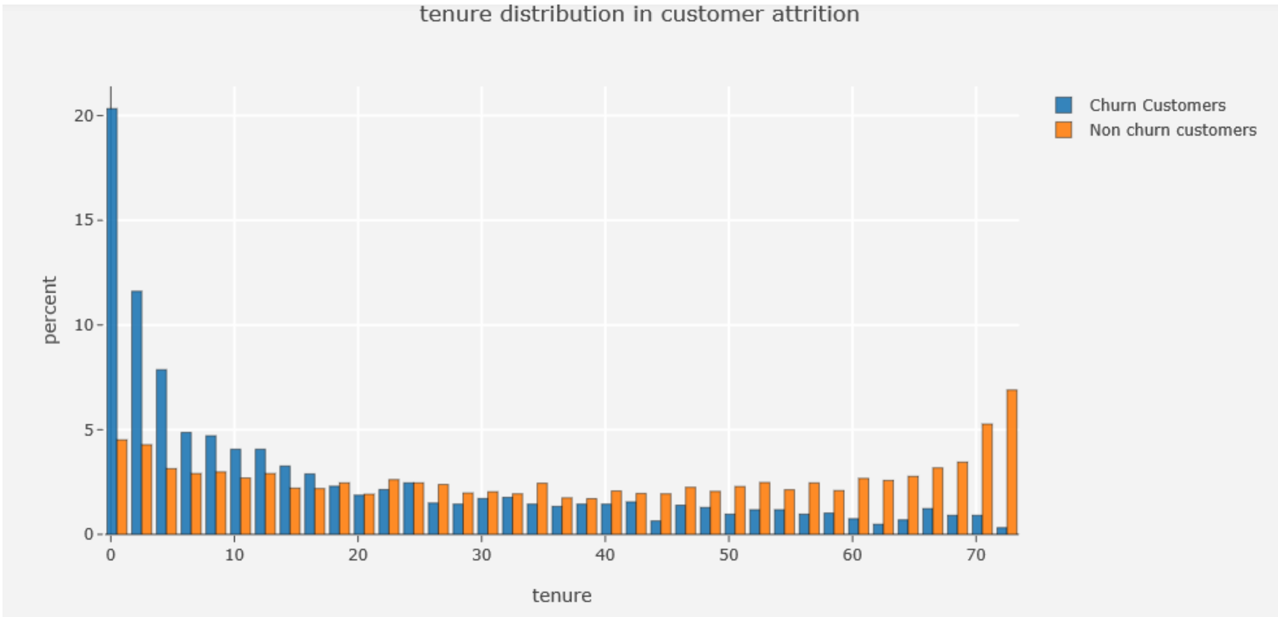

100
80
60
40
20
0
Churn
No
Yes
No
No internet service
Yes
StreamingTV

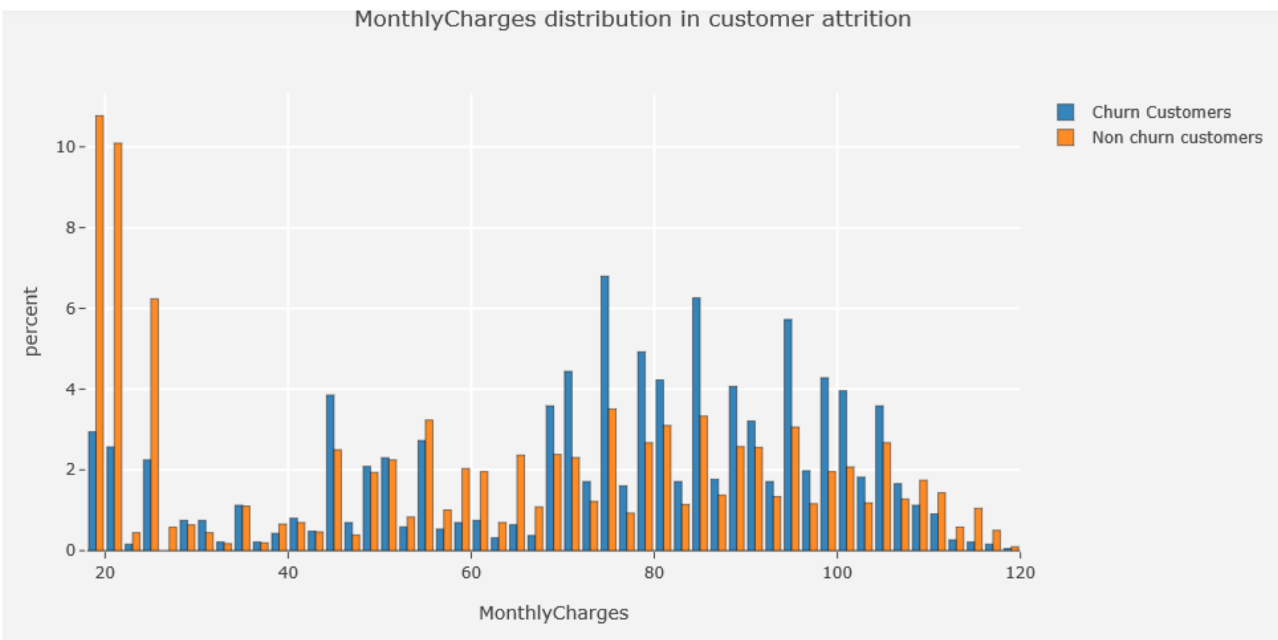
,

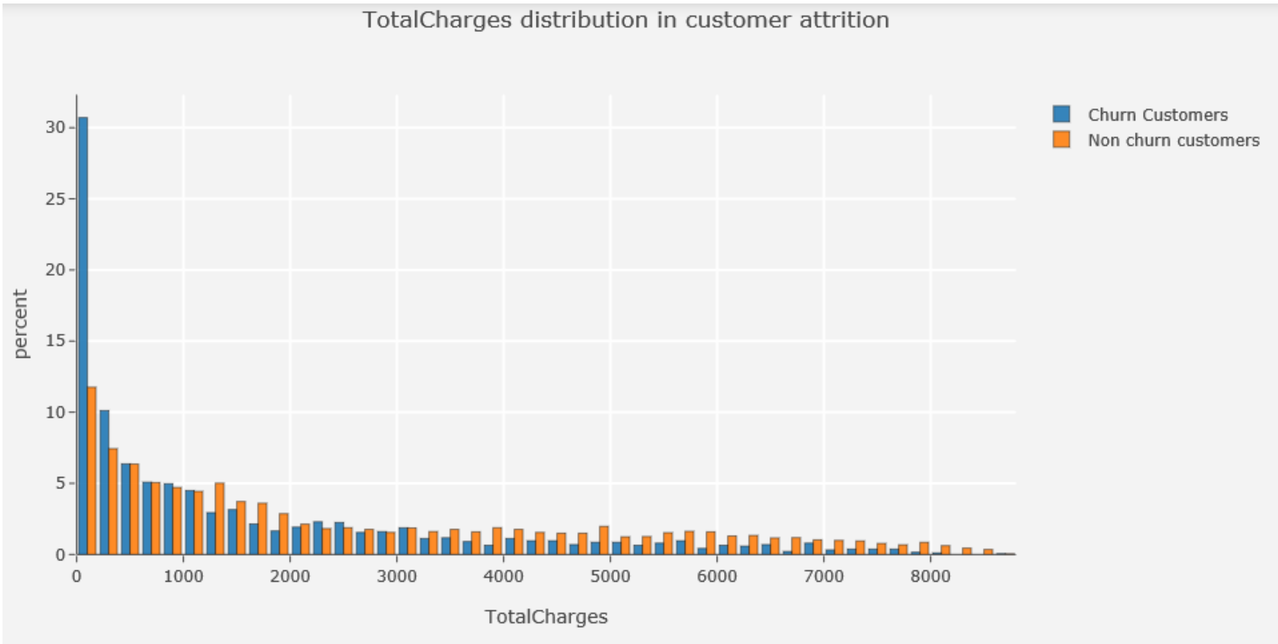

100
80
60
40
20
0
Churn
No
Yes
No
No internet service
Yes
StreamingMovies


120
100
80
60
40
20
MonthlyCharges
No
Yes
Churn


8000
6000
4000
2000
0
TotalCharges
No
Yes
Churn


tenure distribution in customer attrition
Churn Customers
Non churn customers
percent
tenure
0
5
10
15
20
0
10
20
30
40
50
60
70


MonthlyCharges distribution in customer attrition
Churn Customers
Non churn customers
percent
MonthlyCharges
0
2
4
6
8
10
20
40
60
80
100
120



5 Models and Analysis

5.1 Logistic Regression

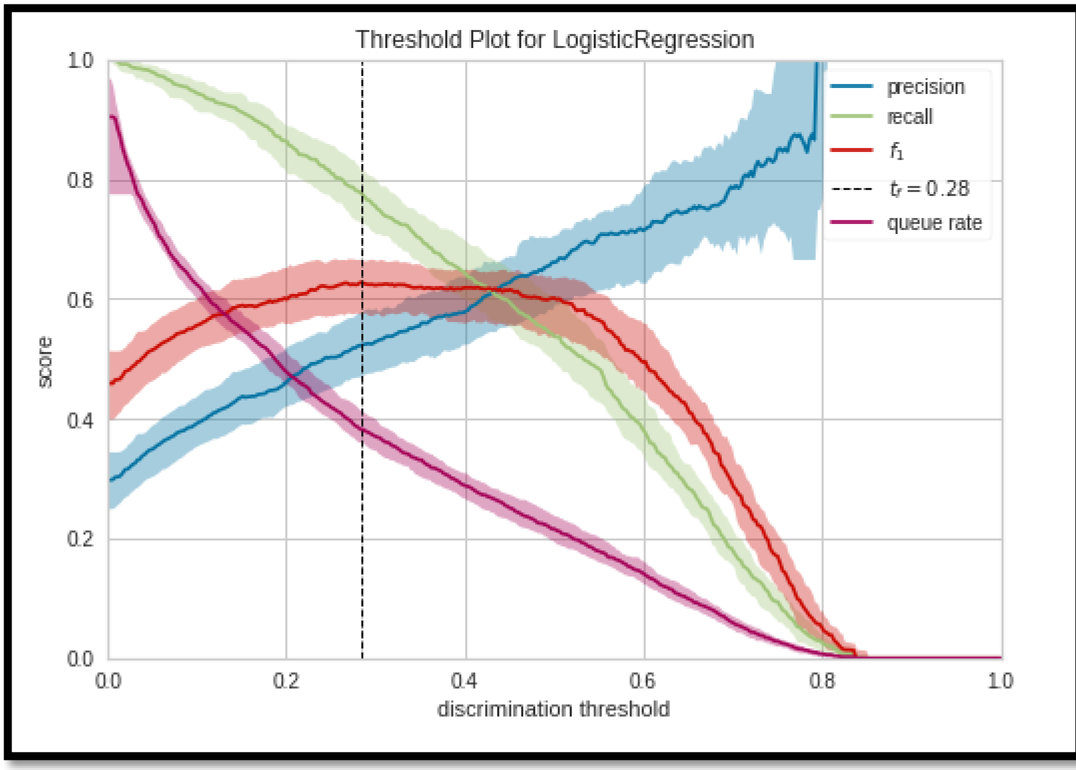
- The probability of the response taking a particular value is modeled based on a combination of values taken by the predictors.
- The advantage of Logistic Regression model is that it gives the confidence of prediction as a probability.
- The disadvantage is that it assumes that the classes are linearly separable in feature space.

Table 1: Confusion Matrix

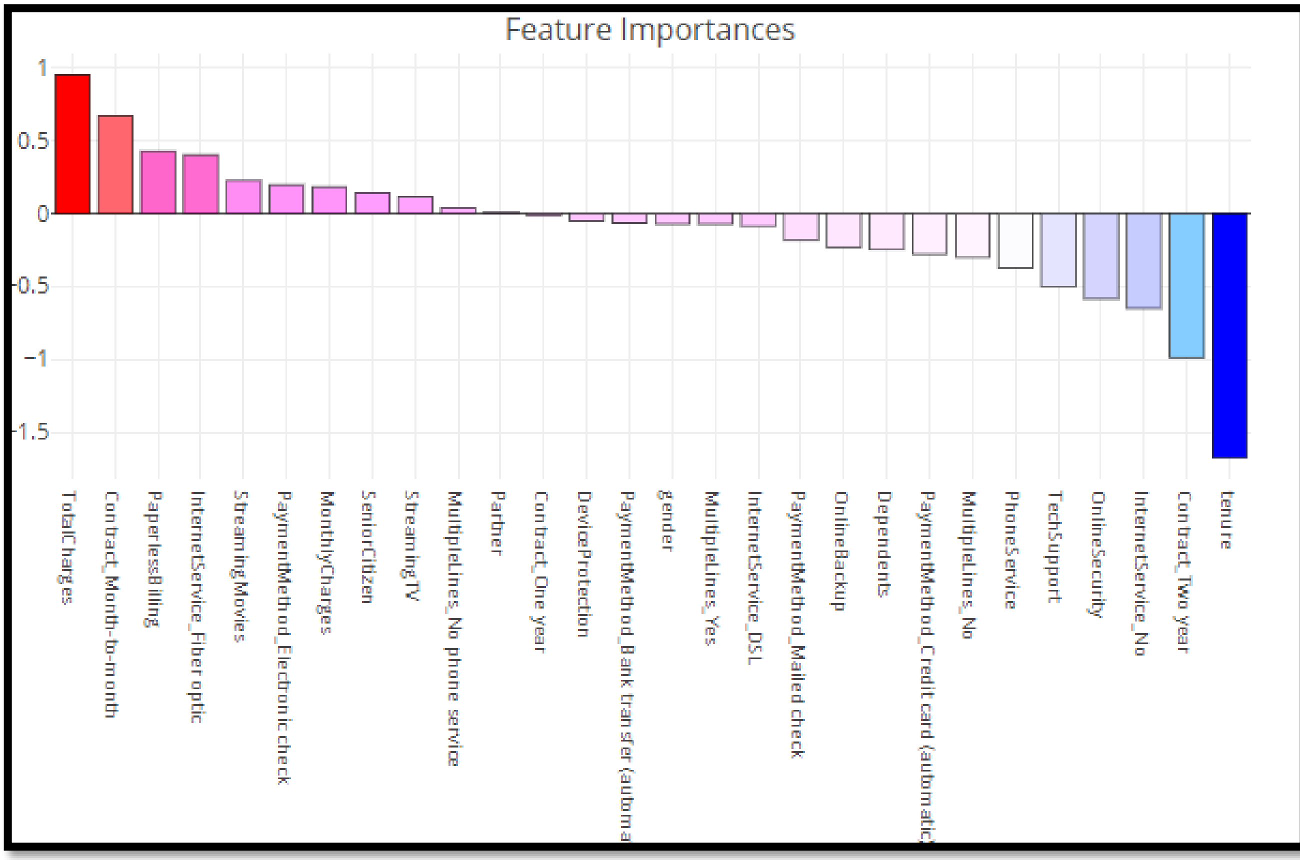
	Not Churn	Churn
Churn	227	12
Not Churn	17702	291

```
 Classification report :
              precision    recall  f1-score   support

           0       0.83      0.90      0.87      1268
           1       0.68      0.54      0.60       490

avg / total       0.79      0.80      0.79      1758

Accuracy   Score :  0.8003412969283277
Area under curve :  0.7194714478851477
```

The model has a very good accuracy i.e. 80% and AUC of 71.9%



We see from the plot that the threshold is 0.28 which tells that customers below the threshold are less likely to churn whereas the customers above the threshold are more likely to churn

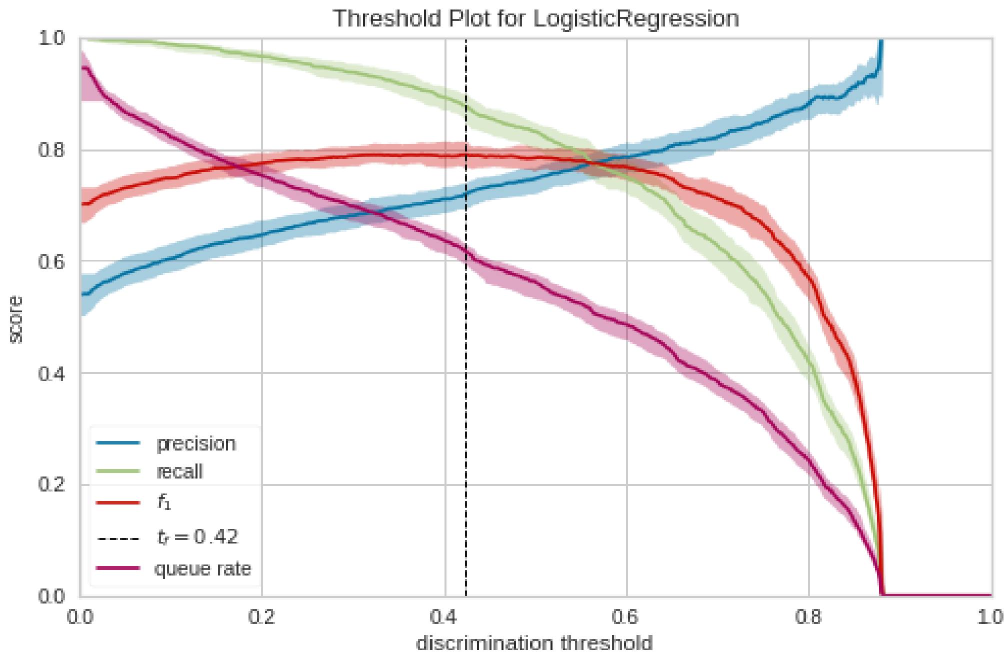


5.2 Logistic Regression (RFE)

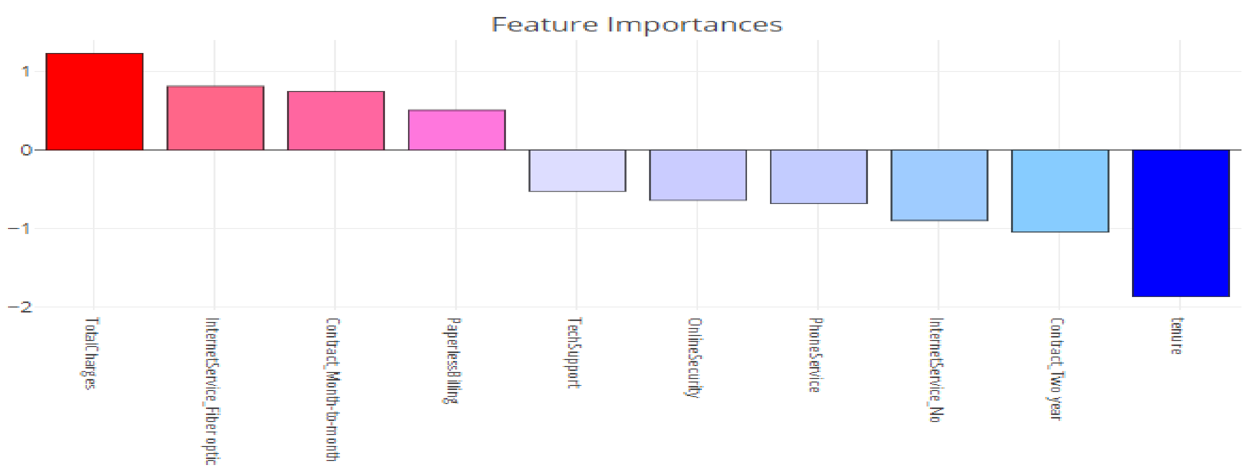
Recursive Feature Elimination (RFE) is based on the idea to repeatedly construct a model and choose either the best or worst performing feature, setting the feature aside and then repeating the process with the rest of the features. This process is applied until all features in the dataset are exhausted. The goal of RFE is to select features by recursively considering smaller and smaller sets of features.

Table 3: Confusion Matrix

	Not Churn	Churn
Churn	227	12
Not Churn	17702	291



Threshold= 0.42



The most important predictors in Customer attrition after applying RFE in Logistic Regresssion

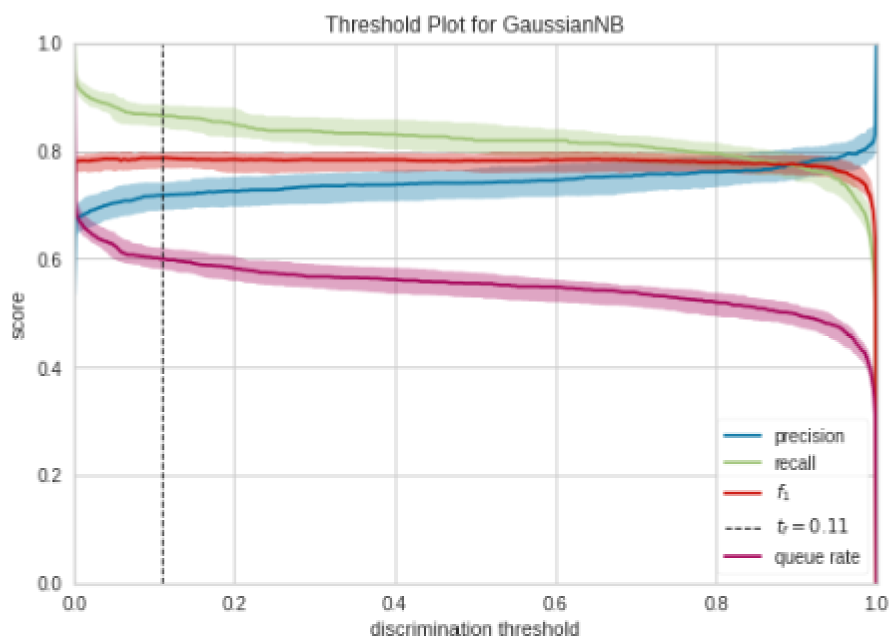
5.3 Naïve Bayes Classifier

Table 4: Confusion Matrix

	Not Churn	Churn
Churn	227	12
Not Churn	17702	291

```
Classification report :
              precision    recall  f1-score   support

          0       0.90      0.74      0.81      1268
          1       0.54      0.79      0.64       490

avg / total       0.80      0.75      0.77      1758

Accuracy Score   :  0.7542662116040956
Area under curve :  0.7657921843816391
```

The model has 75.4 % accuracy and AUC of 76.5 %



Threshold= 0.11

5.4 KNN Classifier

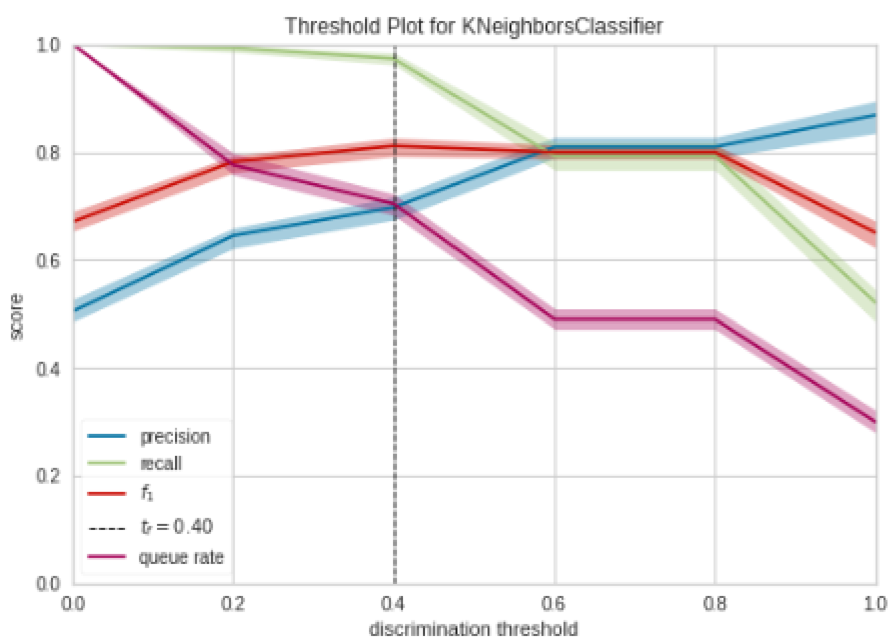
Applying knn algorithm to smote oversampled data.

Table 4: Confusion Matrix

	Not Churn	Churn
Churn	227	12
Not Churn	17702	291

```
 Classification report :
              precision    recall  f1-score   support

          0       0.86      0.69      0.76      1268
          1       0.47      0.71      0.56       490

avg / total       0.75      0.69      0.71      1758

Accuracy Score    :  0.6939704209328783
Area under curve :  0.6989506212579669
```

The model has 69.3% accuracy and AUC of 69.8% which is quite low compared to Logistic Regression and KNN



Threshold= 0.40

6 Findings and Managerial implications

To be added – by Ashwin

7 Conclusions

The initial part of our project mainly dealt with exploratory data analysis of the telecom data.