MATH2319 Machine Learning Project Phase - 1

Identifying the factors affecting customer churn

Submitted by: Simarpreet Luthra (s3706588) & Pooja Dandge (s3698457)

Submitted to: Dr. Vural Aksakalli

Table of Contents

Introduction

The aim of this project is to investigate the reason behind discontinuing customers of a telecommunications company [1]. The dataset was obtained from the sample datasets provided in IBM Watson Analytics Community blog at:

https://community.watsonanalytics.com/wp-content/uploads/2015/03/WA_Fn-UseC_-Telco-Customer-Churn.csv.

The dataset provides the customer demographics, opted services and account preferences of a telecommunications company [2]. The project has two phases. Phase 1 focuses on data preprocessing and exploration, as covered in this report. Overall, the results indicate that customers who churn have noteworthy similarity in some features such as tenure, monthly charges, type of internet service and payment method preference. Further analysis will be conducted in Phase 2, using Machine Learning models on customer attributes to classify customers who are more likely to churn. This would help in developing focused customer retention programs.

Data Preprocessing

In [1]:
```
import pandas as pd
import numpy as np
```

In [2]:
```
churn = pd.read_csv("./Telco-Customer-Churn.csv", sep=',', decimal='.', header=0)
```

The dataset contains the specific attributes of 7043 customer and whether they churned within the last month or not. The dataset contains a CustomerID column, which is a unique identifier for each customer. The target feature is Churn, which refers to customers who left within the last month. The descriptive features are of 3 types:

- 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 information about customers – gender, age range, and if they have partners and dependents

In [3]:
```
churn.shape
```

Out[3]:
```
(7043, 21)
```

In [4]:
```
# Data type check

print(f"\nData type for each feature:")
churn.dtypes
```

```
Data type for each feature:
```

Out[4]:
```
customerID          object
gender              object
SeniorCitizen        int64
Partner             object
Dependents          object
tenure               int64
PhoneService        object
MultipleLines       object
InternetService     object
OnlineSecurity      object
OnlineBackup        object
DeviceProtection    object
```

```
TechSupport          object
StreamingTV          object
StreamingMovies      object
Contract             object
PaperlessBilling     object
PaymentMethod        object
MonthlyCharges      float64
TotalCharges         object
Churn                object
dtype: object
```

The data type **float64** and **int64** refer to numerical variables and **object** refers to categorical variables. Since, the data description states that Total charges is a numeric column, this implies there might be some extra string (usually for the missing value) which makes python read it as a categorical variable.

In [5]:
```
# Checking the extra string

(churn['TotalCharges'].value_counts())
```

Out[5]:
```
           11
20.2       11
19.75       9
19.65       8
19.9        8
20.05       8
19.55       7
45.3        7
19.45       6
20.25       6
20.15       6
20.3        5
20.45       5
19.3        4
69.95       4
19.85       4
49.9        4
...
Name: TotalCharges, Length: 6531, dtype: int64
```

Thus, by doing value counts, it can be seen that NA values were actually read as a whitespace, which converted numeric variable to object. Thus, the whitespace were replaced with with NAN and then converted to numeric.

In [6]:

```
# Replace blank with NA values and then then convert it to numerical variable

churn['TotalCharges'] = churn["TotalCharges"].replace(" ", np.nan)

churn['TotalCharges'] = pd.to_numeric(churn['TotalCharges'])

# Missing value check

print(f"\nNumber of missing value for each feature:")
churn.isna().sum()
```

```
Number of missing value for each feature:
```

Out[6]:

```
customerID           0
gender               0
SeniorCitizen        0
Partner              0
Dependents           0
tenure               0
PhoneService         0
MultipleLines        0
InternetService      0
OnlineSecurity       0
OnlineBackup         0
DeviceProtection     0
TechSupport          0
StreamingTV          0
StreamingMovies      0
Contract             0
PaperlessBilling     0
PaymentMethod        0
MonthlyCharges       0
TotalCharges        11
Churn                0
dtype: int64
```

The dataset was then checked for missing values. It was found that TotalCharges column was missing in eleven rows. These missing values were substituted by average of the column. The other features had no missing value.

In [7]:

```
# Replacing NA values with the mean of the column

churn['TotalCharges'].fillna(churn['TotalCharges'].mean(), inplace=True)
```

In [8]:
```
print(f"\n No. of NA values for TotalCharges column after replacing with mean
of the column:")
churn['TotalCharges'].isna().sum()
```

```
No. of NA values for TotalCharges column after replacing with mean of the col
umn:
```

Out[8]:
```
0
```

In [9]:
```
# Statistical analysis of all the numerical variables

churn.describe()
```

Out[9]:

	SeniorCitizen	tenure	MonthlyCharges	TotalCharges
count	7043.000000	7043.000000	7043.000000	7043.000000
mean	0.162147	32.371149	64.761692	2283.300441
std	0.368612	24.559481	30.090047	2265.000258
min	0.000000	0.000000	18.250000	18.800000
25%	0.000000	9.000000	35.500000	402.225000
50%	0.000000	29.000000	70.350000	1400.550000
75%	0.000000	55.000000	89.850000	3786.600000
max	1.000000	72.000000	118.750000	8684.800000

The best way to summarise numerical variables is using the summary statistics. Senior citizen, being a factor type column is coded as '0' being not a senior citizen and '1' being senior citizen. More than 75% values are '0', which is apt, describing that less than twenty five percent customers of the telco company hold senior citizenship status. This in turn, might entitle them to some special discounts and might have less churn rate or might find competition's offers appealing and thus being churned. However, this hypothesis would be catered to later.

Tenure value ranges from zero to seventy two months. Here zero is acceptable, as the customer might have been with the company for a few days only and not for the whole month. Since plans are billed monthly, it does not make any sense to count the number of days.

Monthly charges range between eighteen dollars and one hundred nineteen dollars with mean of almost sixty five dollars and median of seventy dollars. Thus, the distribution can be considered almost normal. Total charges refers to cumulative charges throughout the customer's tenure, lies between between about nineteen dollars and eight thousand six

hundred eighty five dollars and is highly right skewed with mean almost eight hundred dollars more than the median.

In [10]:
```
# Statistical analysis of all the categorical variables

for col in churn.columns[1:len(churn.columns)]:
    if (churn[col].dtype == 'object'):
        print(churn[col].value_counts(), '\n')
```

```
Male      3555
Female    3488
Name: gender, dtype: int64

No     3641
Yes    3402
Name: Partner, dtype: int64

No     4933
Yes    2110
Name: Dependents, dtype: int64

Yes    6361
No      682
Name: PhoneService, dtype: int64

No                  3390
Yes                 2971
No phone service     682
Name: MultipleLines, dtype: int64

Fiber optic    3096
DSL            2421
No             1526
Name: InternetService, dtype: int64

No                     3498
Yes                    2019
No internet service    1526
Name: OnlineSecurity, dtype: int64

No                     3088
Yes                    2429
No internet service    1526
Name: OnlineBackup, dtype: int64

No                     3095
Yes                    2422
No internet service    1526
```

```
Name: DeviceProtection, dtype: int64

No                     3473
Yes                    2044
No internet service    1526
Name: TechSupport, dtype: int64

No                     2810
Yes                    2707
No internet service    1526
Name: StreamingTV, dtype: int64

No                     2785
Yes                    2732
No internet service    1526
Name: StreamingMovies, dtype: int64

Month-to-month    3875
Two year          1695
One year          1473
Name: Contract, dtype: int64

Yes    4171
No     2872
Name: PaperlessBilling, dtype: int64

Electronic check             2365
Mailed check                 1612
Bank transfer (automatic)    1544
Credit card (automatic)      1522
Name: PaymentMethod, dtype: int64

No     5174
Yes    1869
Name: Churn, dtype: int64
```

The best way to summarise categorical column is by value counts. Here the sample is very representative of the population as possible as there are equal percentages of both the genders and equal percentages of single and married customers. The number of people with dependents is a third of the sample and thus, covering married couples without children in the case scenario.

Since this is a telecom company customer-base, the number of customers who have opted phone service are ten times as much as those who have not. Among the customers who have opted phone service, there are equal percentages of customers with single and multiple lines.

There are significant percentages of customers who have or haven't opted for the services and preferences described by other features such as online backup, payment method, paperless billing etcetera.

Finally, the target column Churn has three times more customers who stayed than churned. There's a need to explore why those twenty five percent customers left the company and whether they can be grouped using the descriptive features. This would help in understanding which customers are dissatisfied with the company and need urgent attention.

Data Exploration

Histogram and Box Plot of Numeric features

In [11]:
```
import matplotlib.pyplot as plt
import seaborn as sns
```

In [12]:
```
# plotting all numerical variables

# creating function for Boxplot and Histogram

def BoxHistogramPlot(x):
    f, (ax_box, ax_hist) = plt.subplots(2, sharex=True, gridspec_kw={"height_
ratios": (.15, .85)})
    sns.boxplot(x, ax=ax_box, color="Red")
    sns.distplot(x, ax=ax_hist, color="Red")
    ax_box.set(yticks=[])
    sns.despine(ax=ax_hist)
    sns.despine(ax=ax_box, left=True)
```

In [13]:
```
fig = [None]*40
i = 1
for col in churn[['tenure','MonthlyCharges', 'TotalCharges']]:
    BoxHistogramPlot(churn[col])
    fig[i] = "Figure " + str(i) + ": Histogram and Box Plot of " + col
    print(fig[i])
    i = i + 1
    plt.show()
    print("\n")
```

Figure 1: Histogram and Box Plot of tenure

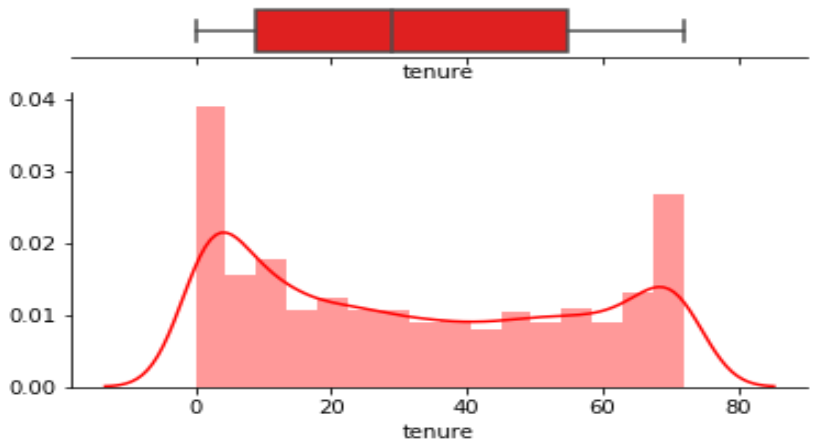


Figure 2: Histogram and Box Plot of MonthlyCharges

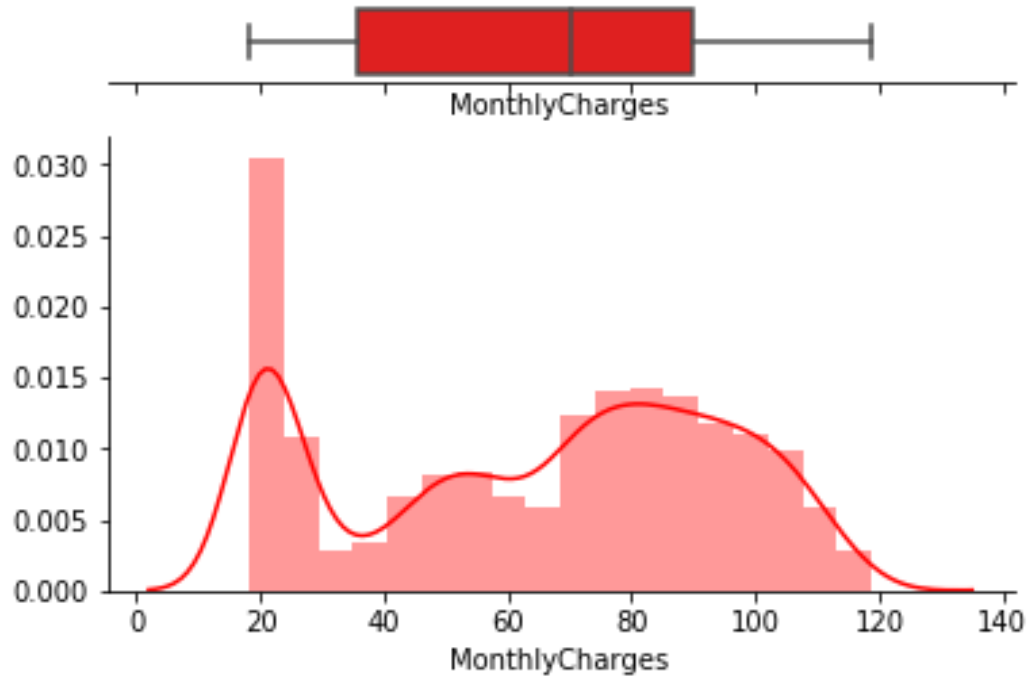
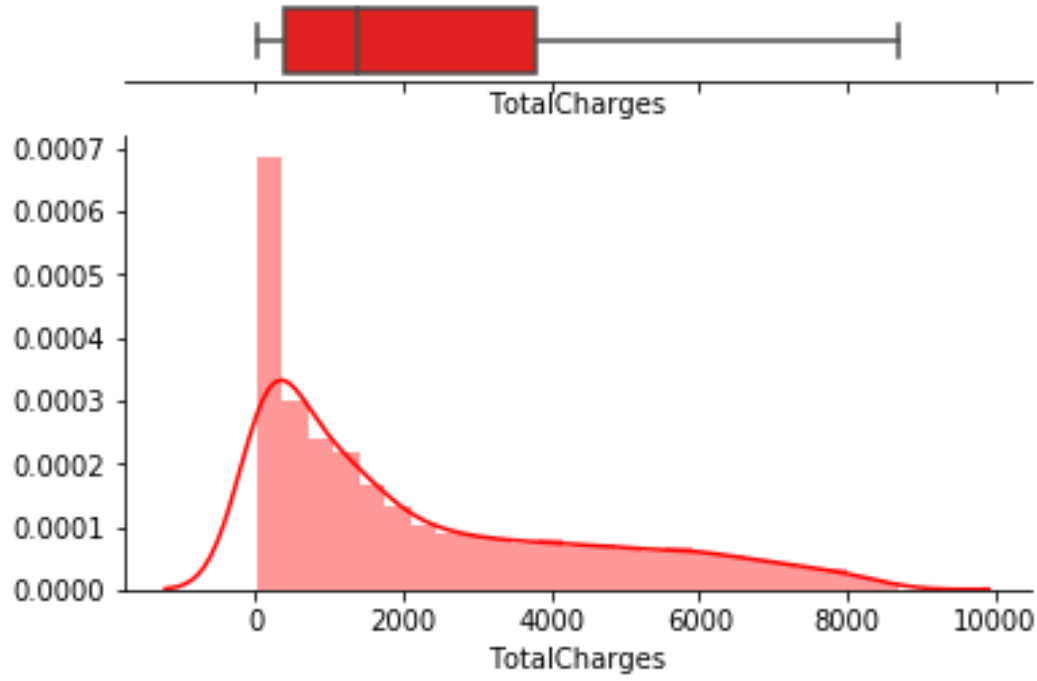


Figure 3: Histogram and Box Plot of TotalCharges



All of the numeric variables have high density around the starting value.

Tenure has high density around max value as well, which refers to the loyal customers and there is lower lower density for all the other values of tenures. This means that old customers are content and marketing strategies are in place, but they really need to reduce the churn rate.

Monthly charges can be divided into two top segments: Economical - The cheapest plans attract the major volume of customer. Premium - The plans above seventy dollars attract much more customers than mid range plans.

Total charges are very right skewed as expected of any cumulative frequency distribution.

Bar Plot of Categorical features

In [14]:
```
# plotting all categorical variables

# creating function for bar plot

sns.set(color_codes=True)
def BarPlot(x):
    total = float(len(churn))
    ax = churn[x].value_counts(normalize = True).plot(kind = "bar", alpha = 0
.5)
```

In [15]:
```
for col in ['gender', 'SeniorCitizen', 'Partner', 'Dependents', 'PhoneService
', 'MultipleLines', 'InternetService', 'OnlineSecurity', 'OnlineBackup', 'Dev
iceProtection', 'TechSupport', 'StreamingTV', 'StreamingMovies', 'Contract',
'PaperlessBilling', 'PaymentMethod', 'Churn']:
    plt.figure(figsize=(6,2))
    fig[i] = "Figure " + str(i) + ": Bar Chart of " + col
    plt.title(fig[i])
    BarPlot(col)
    plt.show()
    i = 1 + i
```

Figure 4: Bar Chart of gender

Figure 5: Bar Chart of SeniorCitizen

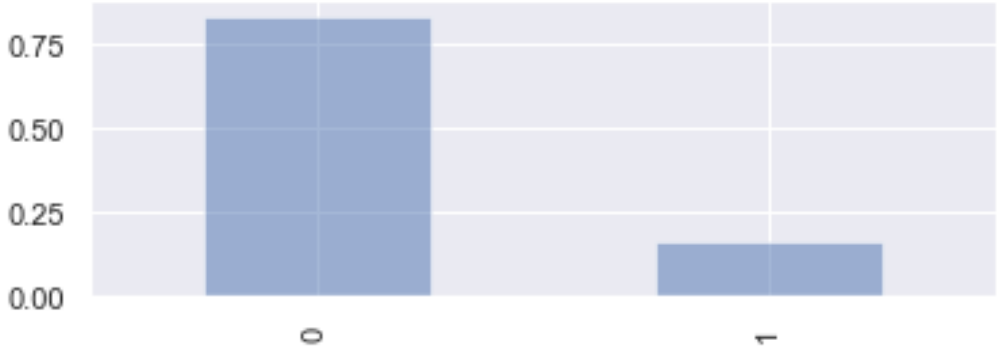


Figure 6: Bar Chart of Partner

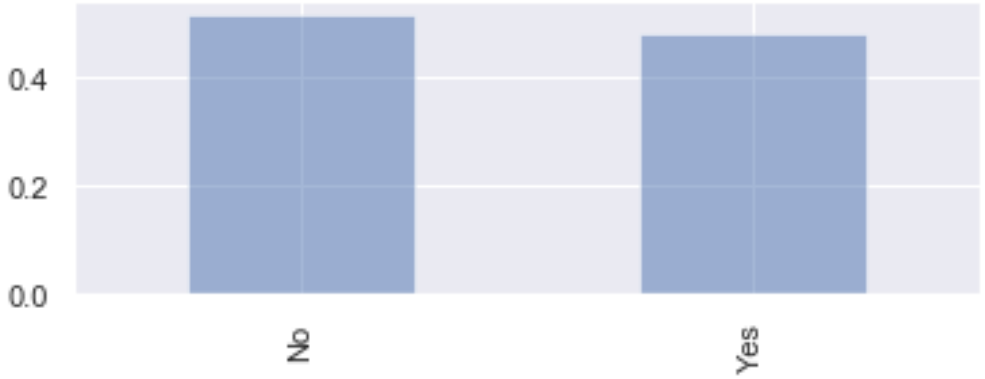


Figure 7: Bar Chart of Dependents

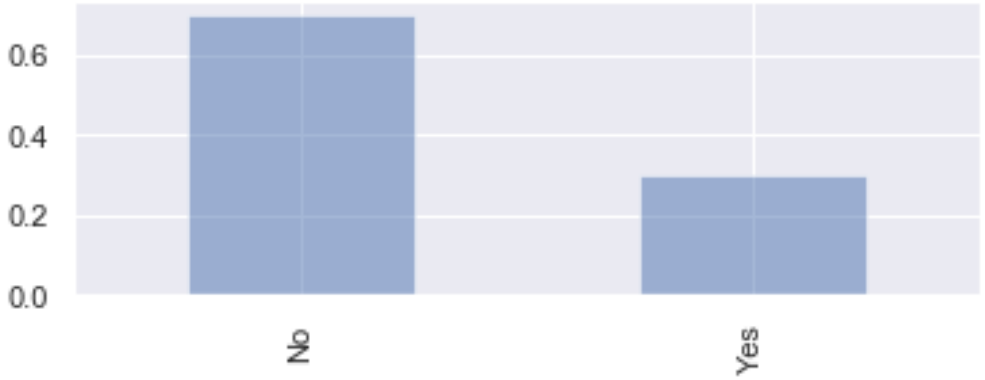


Figure 8: Bar Chart of PhoneService

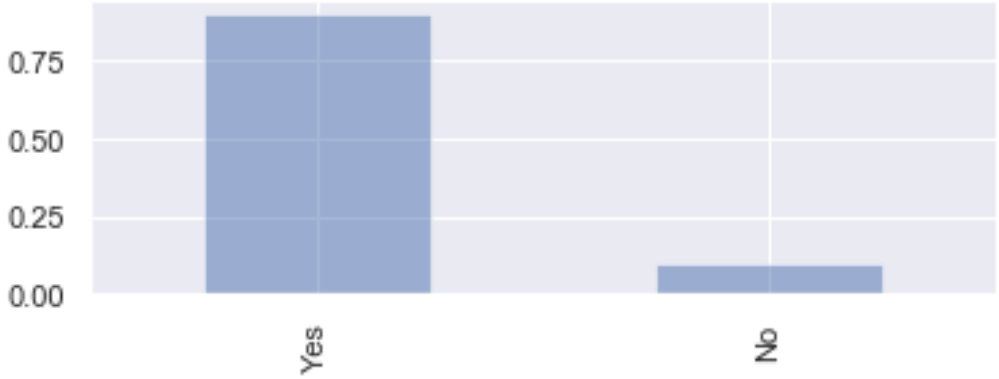


Figure 9: Bar Chart of MultipleLines

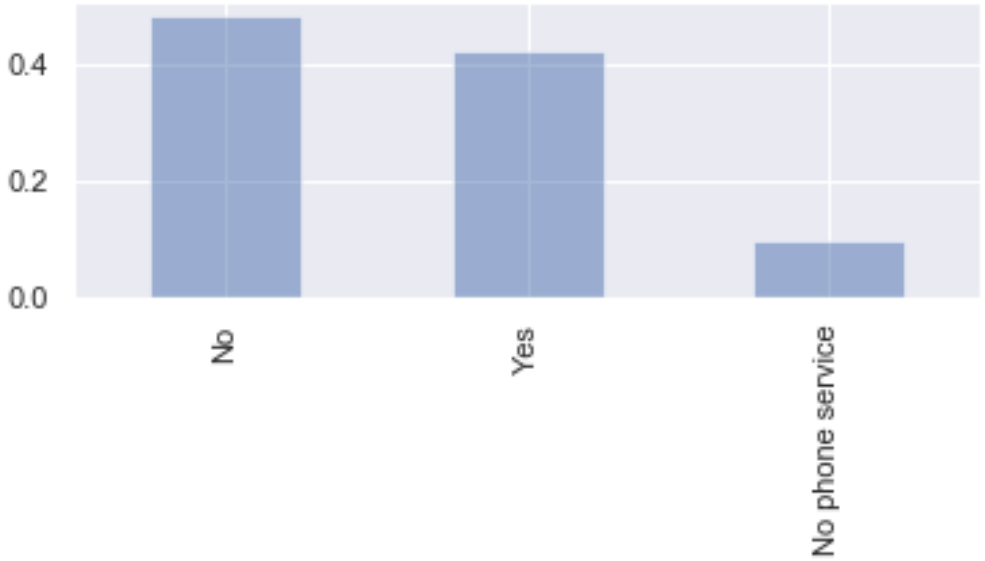


Figure 10: Bar Chart of InternetService

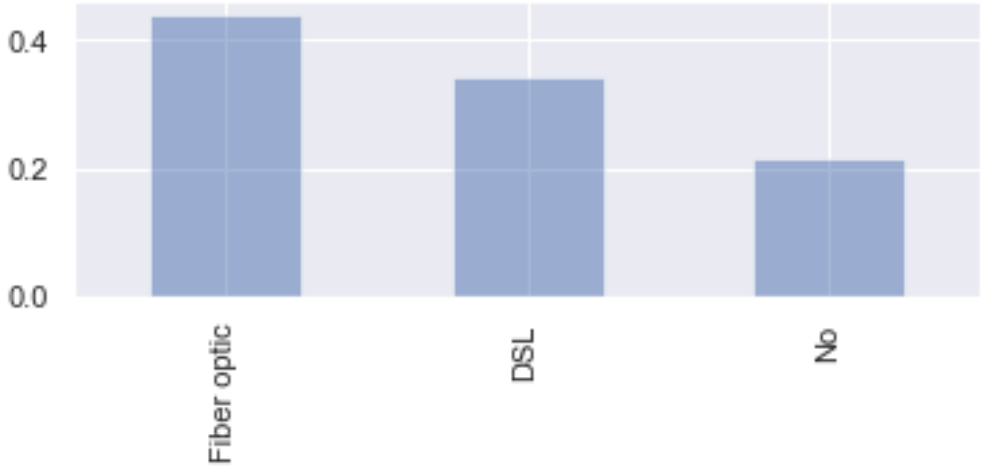


Figure 11: Bar Chart of OnlineSecurity

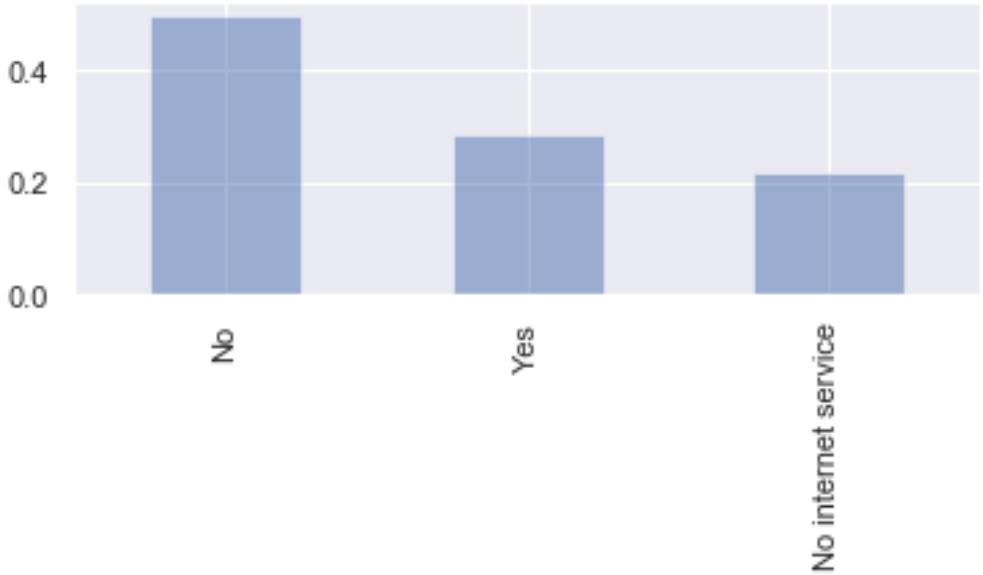


Figure 12: Bar Chart of OnlineBackup

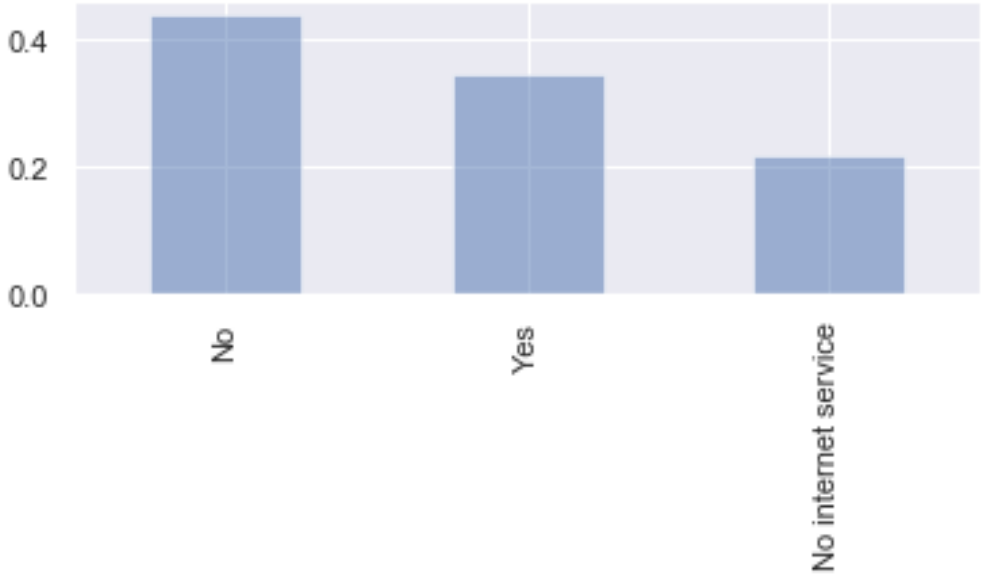


Figure 13: Bar Chart of DeviceProtection

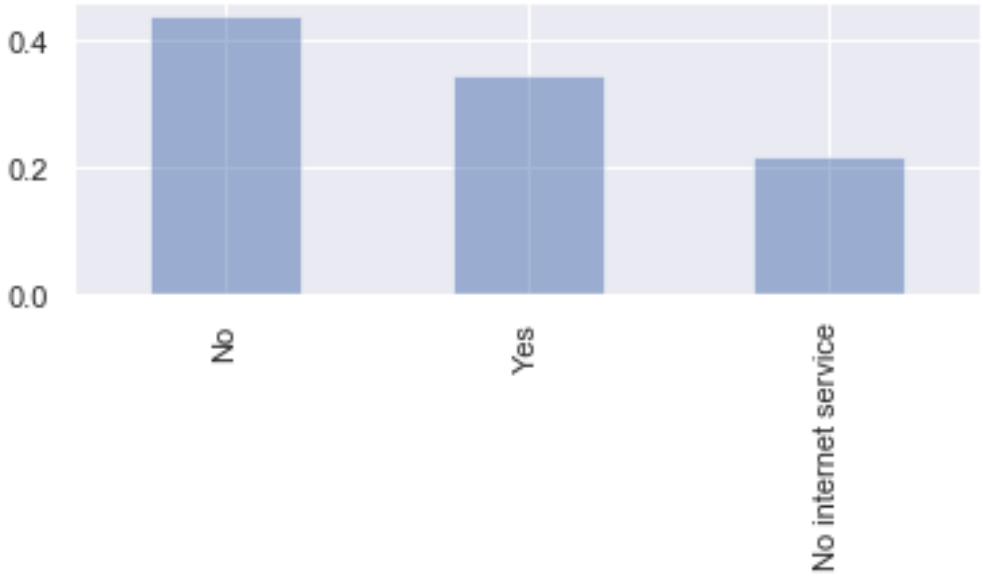


Figure 14: Bar Chart of TechSupport

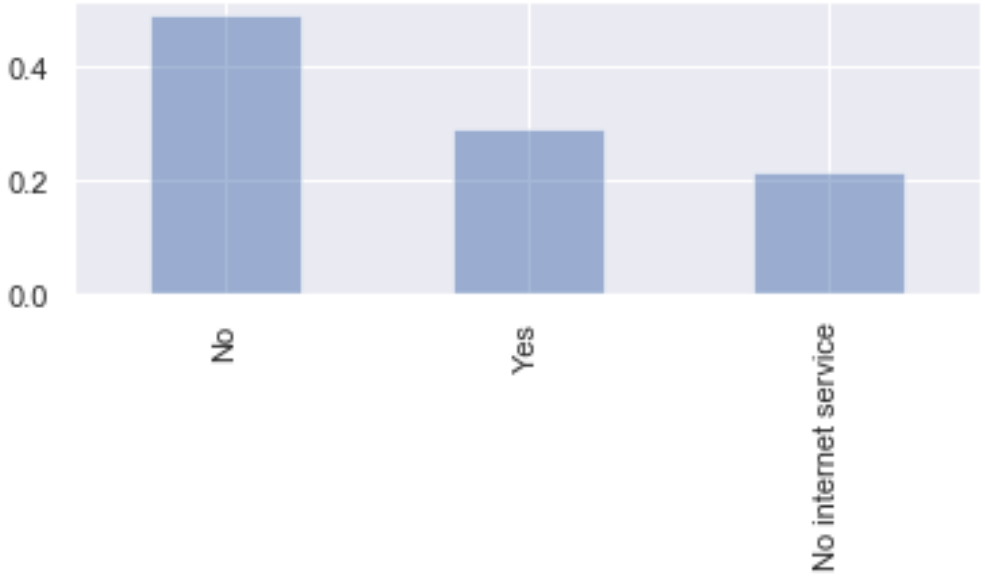


Figure 15: Bar Chart of StreamingTV

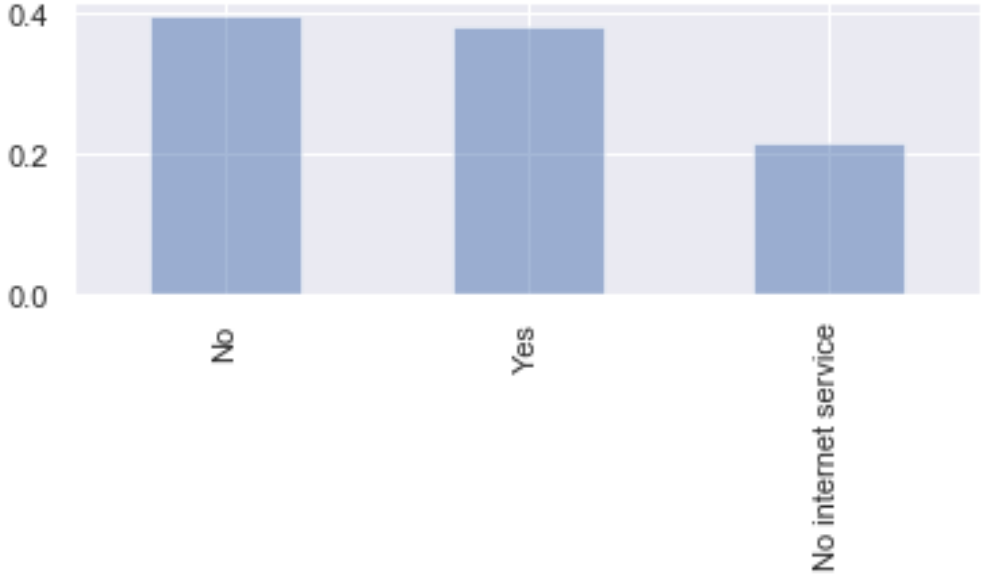


Figure 16: Bar Chart of StreamingMovies

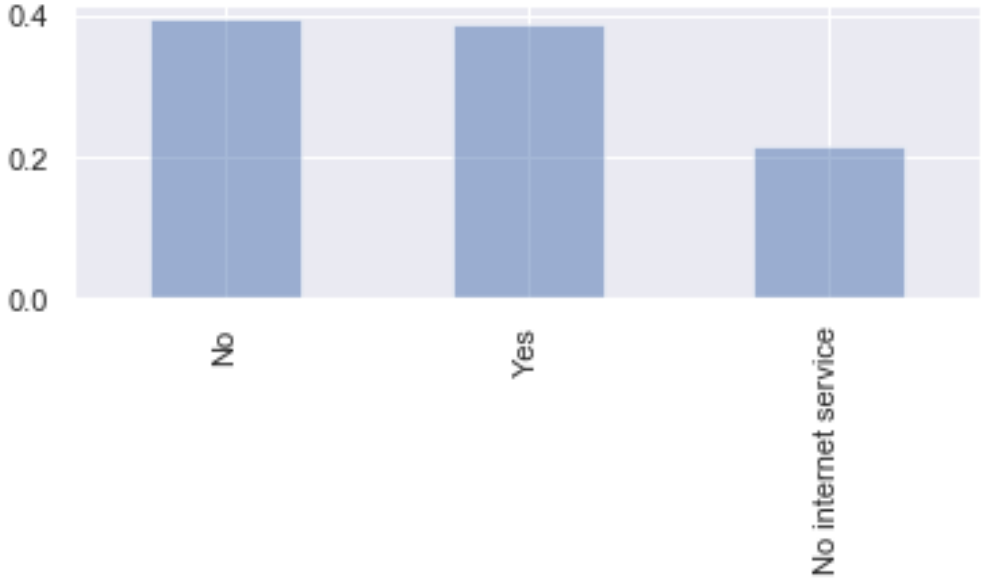


Figure 17: Bar Chart of Contract

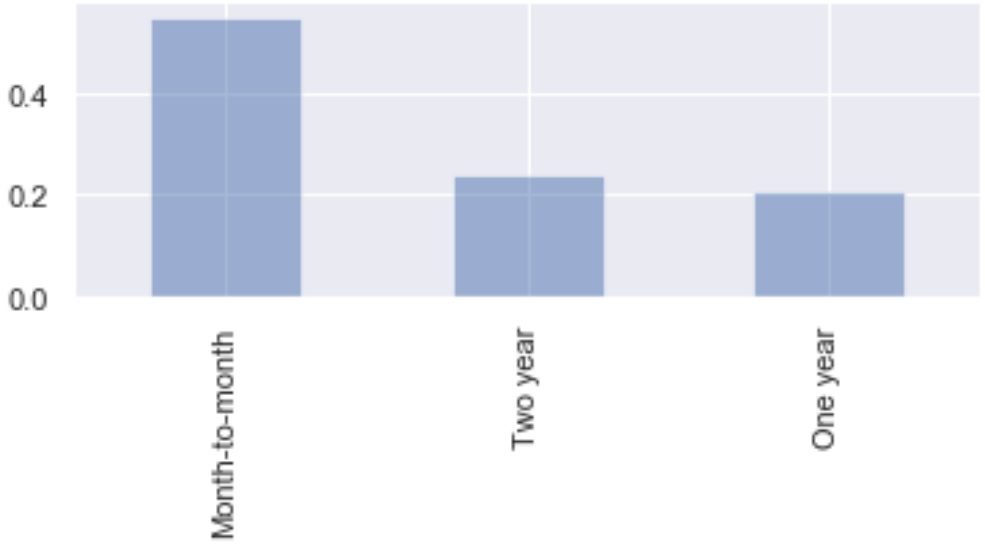


Figure 18: Bar Chart of PaperlessBilling

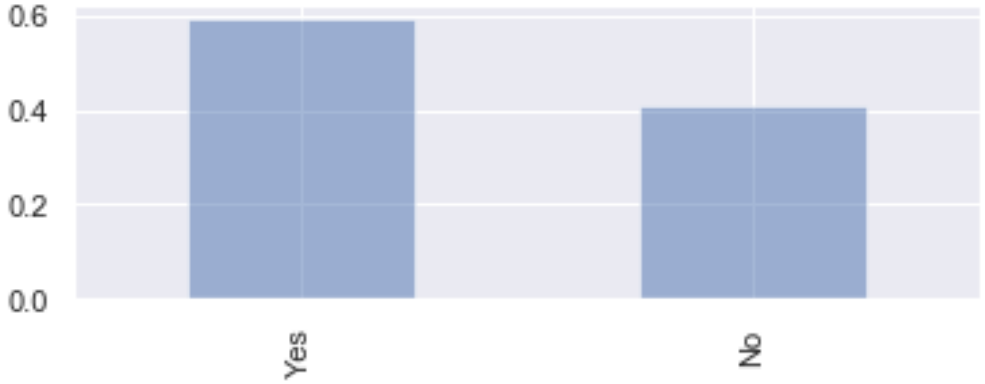


Figure 19: Bar Chart of PaymentMethod

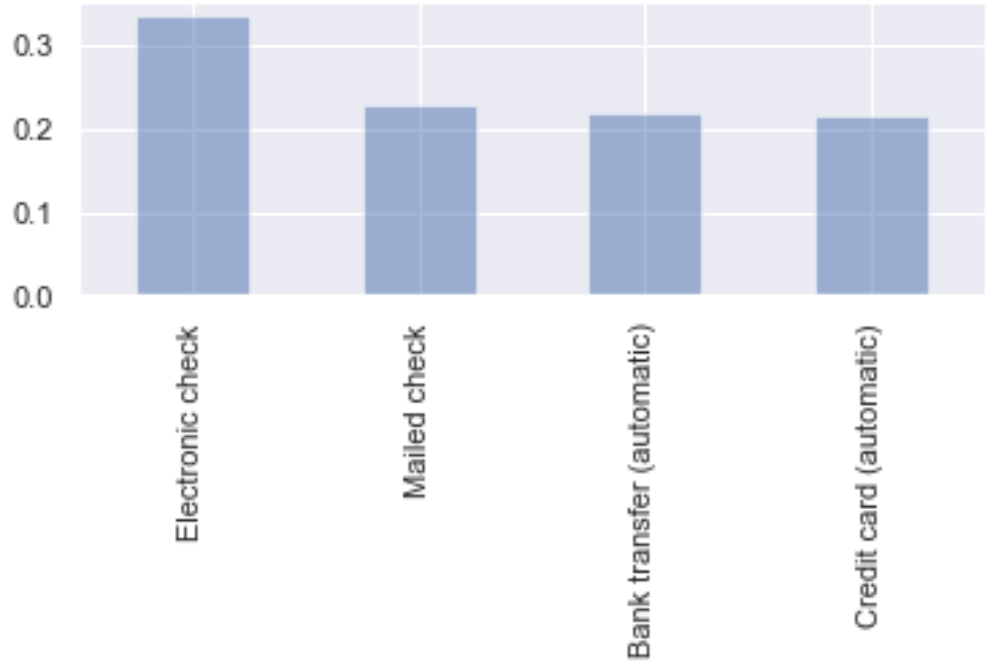
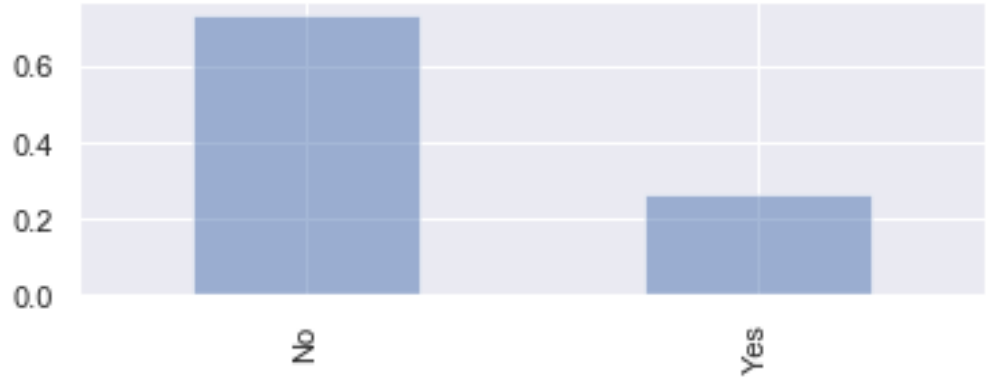


Figure 20: Bar Chart of Churn



Significant proportions can be seen for each of the feature levels, which should ideally represent all types of customers and give a good view of the customer-base on a whole.

Multivariate Visualisation

Now, The relationship of all variables will be check with the Churn to identify and segregate the important variables for the analysis.

Box Plots of Numeric Features Segregated by Churn

In [16]:

```
for col in ['tenure', 'MonthlyCharges', 'TotalCharges']:
    fig[i] = "Figure " + str(i) + ": Box Plot of " + col + " segregated by ch
urn"
```

```
    plt.suptitle(fig[i])
    ax = sns.boxplot(x=col, y="Churn", data=churn)
    plt.show()
    i=i+1
```

Figure 21: Box Plot of tenure segregated by churn

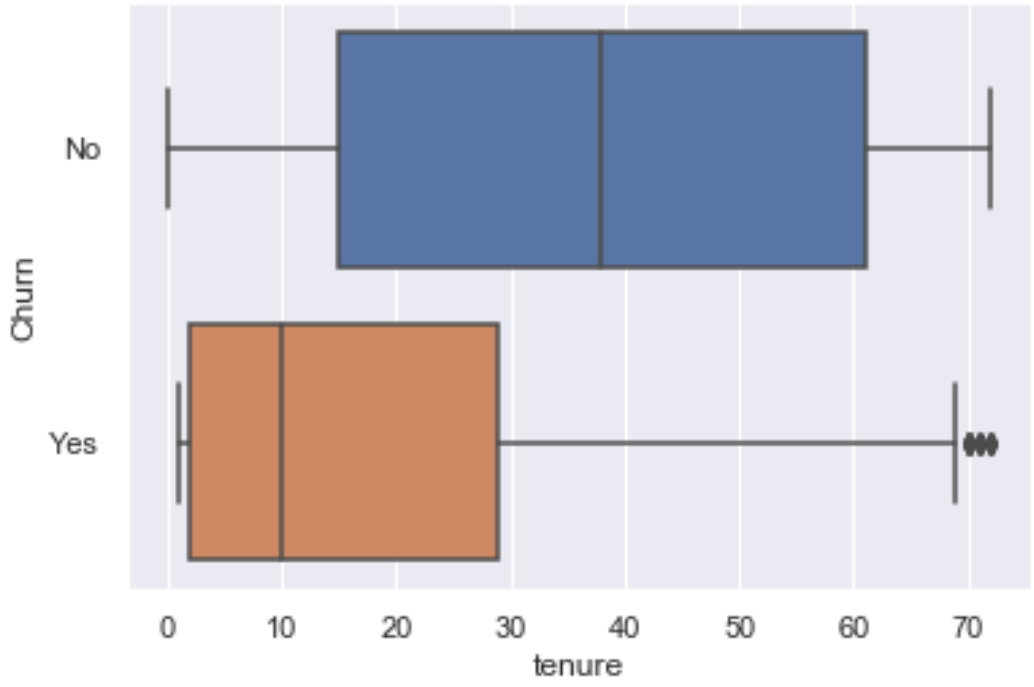


Figure 22: Box Plot of MonthlyCharges segregated by churn

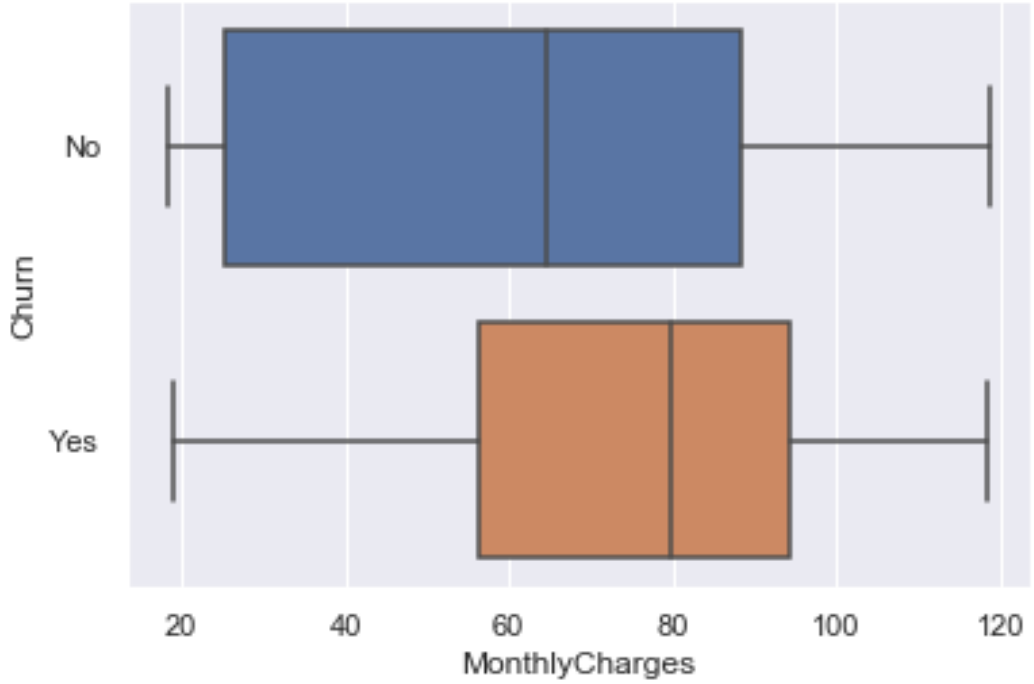


Figure 23: Box Plot of TotalCharges segregated by churn

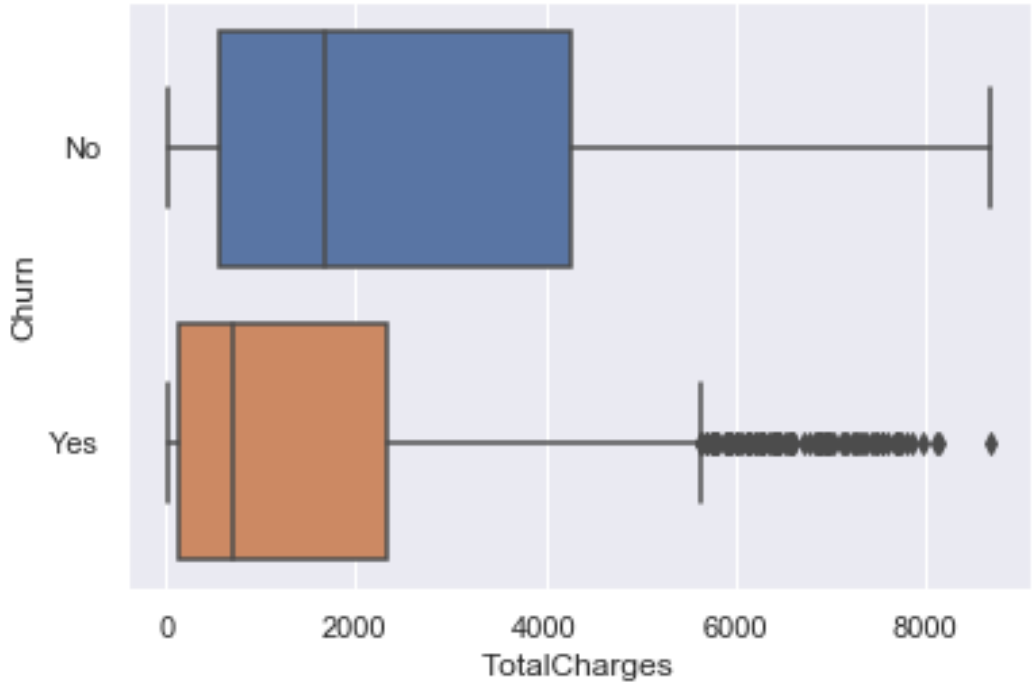


Figure 21 clearly shows that tenure has significant relationship with churn as most of the churned customers have much less tenure as compared to their counterparts.

Monthly charges also have a significant relationship with churn as half the churned customer-base has monthly charges between sixty and ninety dollars.

Since, the very old customers tend to stay with the company, mostly the customers with lower total charges get churned.

Heatmaps for Categorical Features Segregated by Churn

In [17]:

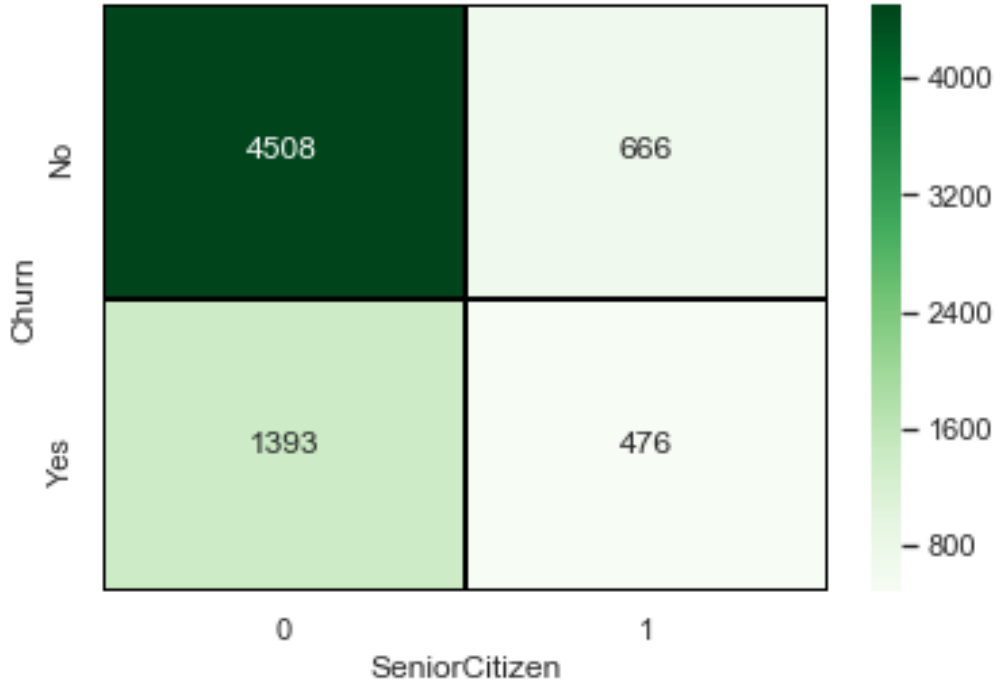
```
for col in ['gender', 'SeniorCitizen', 'Partner', 'Dependents', 'PhoneService
', 'MultipleLines', 'InternetService', 'OnlineSecurity', 'OnlineBackup', 'Dev
iceProtection', 'TechSupport', 'StreamingTV', 'StreamingMovies', 'Contract',
'PaperlessBilling', 'PaymentMethod']:
    fig[i] = "Figure " + str(i) + ": Heatmap of " + col + " and churn"
    plt.suptitle(fig[i])
    a = pd.crosstab(churn['Churn'], churn[col])
    ax = sns.heatmap(a, annot=True, fmt='g', cmap="Greens", linewidths=1, lin
ecolor="black")
    plt.show()
    if col == 'gender':
        print("Gender does not have significant effect on churn as the propor
tions are equal in Fig 24.\n")
    if col == 'SeniorCitizen':
        print("Senior Citizens usually do not like changes and tend to stay w
ith the company.\n")
    if col == 'Partner':
        print("Not having a partner makes you explore more.\nThus, the churn
rate is higher for scustomers who do not have partner as displayed above.\n")
    if col == 'Dependents':
        print("Dependents have similar impact as Partner and the churn rate i
s observed to be lower for customers with dependents.\n")
    if col == 'PhoneService':
        print("Phone Service seemingly does not have much impact on Churn sin
ce the proportions are the same.\n")
    if col == 'MultipleLines':
        print("Having multiple lines also do not have seemingly important dif
ference in the churn rate.\n")
    if col == 'InternetService':
        print("Internet Service has a very high impact on churn rate. The chu
rn rate is especially high for Fiber optic customers.\nThey might be expectin
g better service.\n")
    if col == 'OnlineSecurity':
        print("The people who do not opt for online security churn must faste
r as compared to those who do opt for it.\n")
    if col == 'OnlineBackup':
        print("Online Backup also has a very high impact on churn rate.\nThe
customers who opt for online backup have much lower churn rate compared to th
eir counterparts.\n")
    if col == 'DeviceProtection':
        print("Device Protection has similar impact on churn rate as Online B
```

```
ackup. Those without it, have a very high churn rate.\n")
    if col == 'TechSupport':
        print("Tech Support has similar impact on churn rate as Online Securi
ty. Those without it, have a very high churn rate.\nThey might be expecting b
etter service.\n")
    if col == 'StreamingTV':
        print("Streaming TV service does not have much correlation with churn
as the proportions of yes and no are similar.\n")
    if col == 'StreamingMovies':
        print("Streaming Movies service does not have much correlation with c
hurn as the proportions of yes and no are similar.\n")
    if col == 'Contract':
        print("Customers with month-to-month contract have a higher chance of
churning than yearly or bi-yearly contracts due to flexibility.\n")
    if col == 'PaperlessBilling':
        print("Customers with paperless billing churn more than those without
.\nThis is similar to the Senior Citizens case as these people do not like to
change")
    if col == 'PaymentMethod':
        print("People who pay using Electronic check iare very highly likely
to churn than those that pay using other methods.\n")
    i = i + 1
```

Figure 24: Heatmap of gender and churn

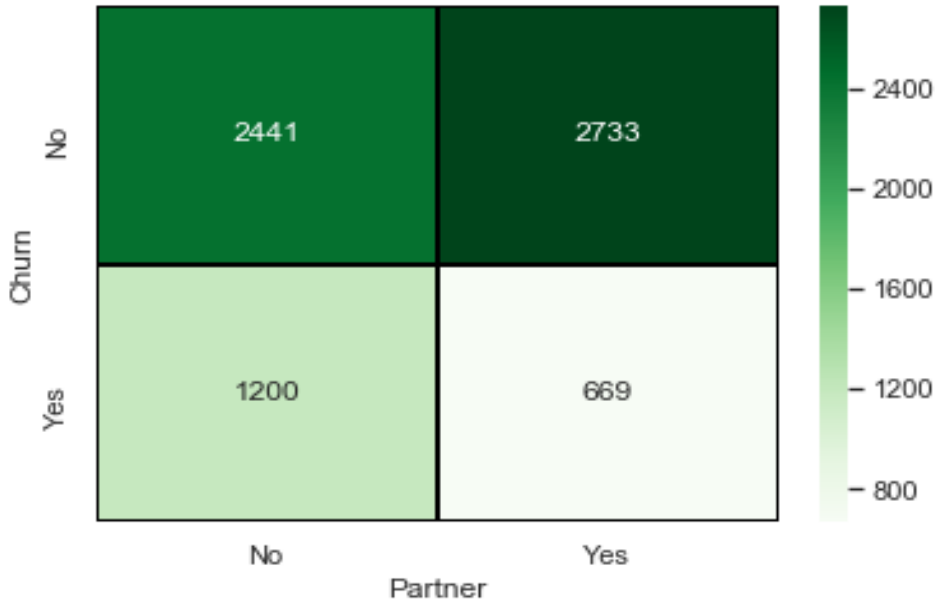
Gender does not have significant effect on churn as the proportions are equal in Fig 24.

Figure 25: Heatmap of SeniorCitizen and churn



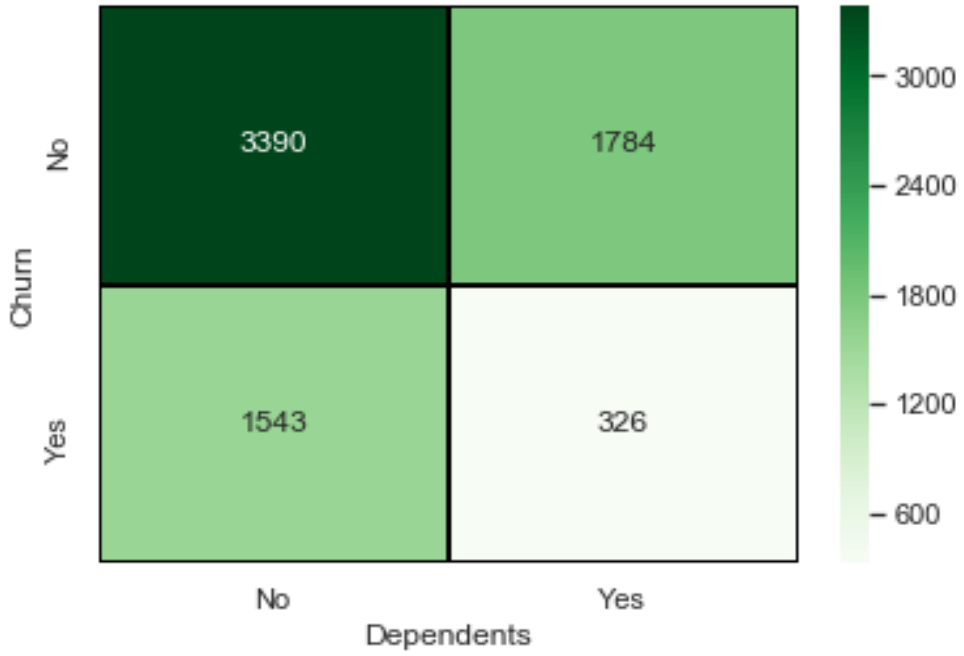
Senior Citizens usually do not like changes and tend to stay with the company .

Figure 26: Heatmap of Partner and churn



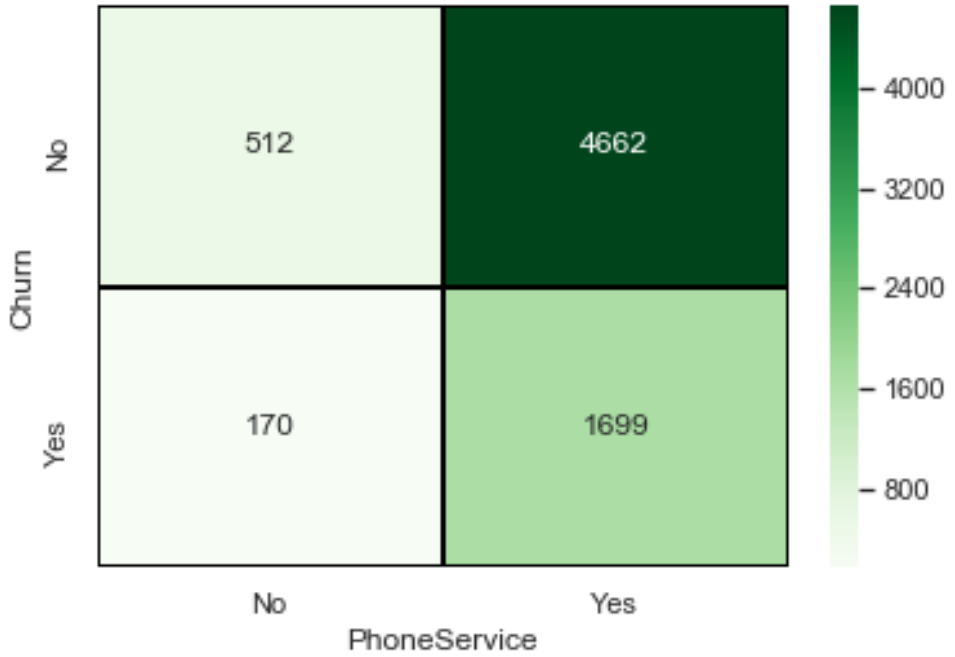
Not having a partner makes you explore more.
Thus, the churn rate is higher for scustomers who do not have partner as displayed above.

Figure 27: Heatmap of Dependents and churn



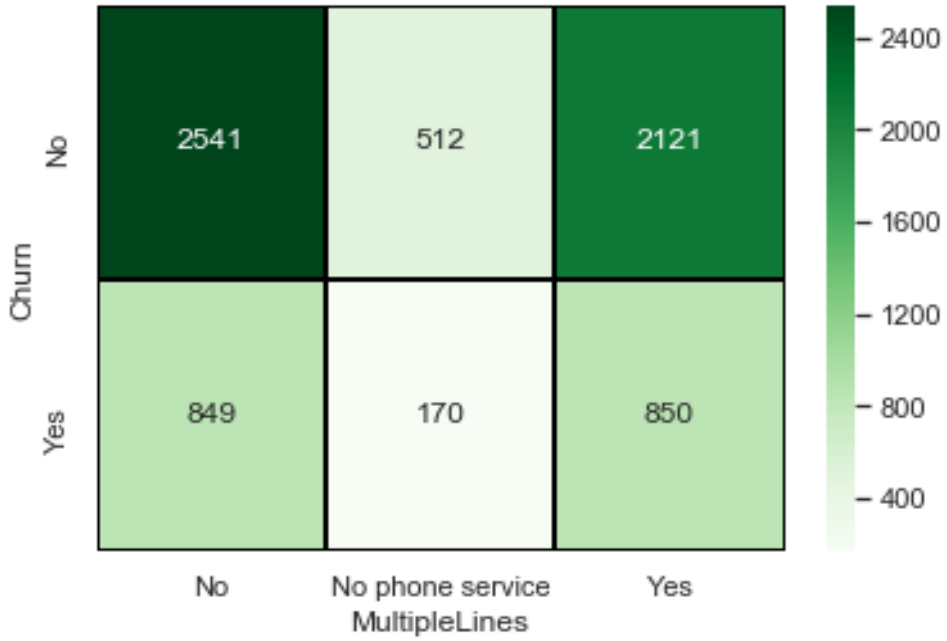
Dependents have similar impact as Partner and the churn rate is observed to b e lower for customers with dependents.

Figure 28: Heatmap of PhoneService and churn



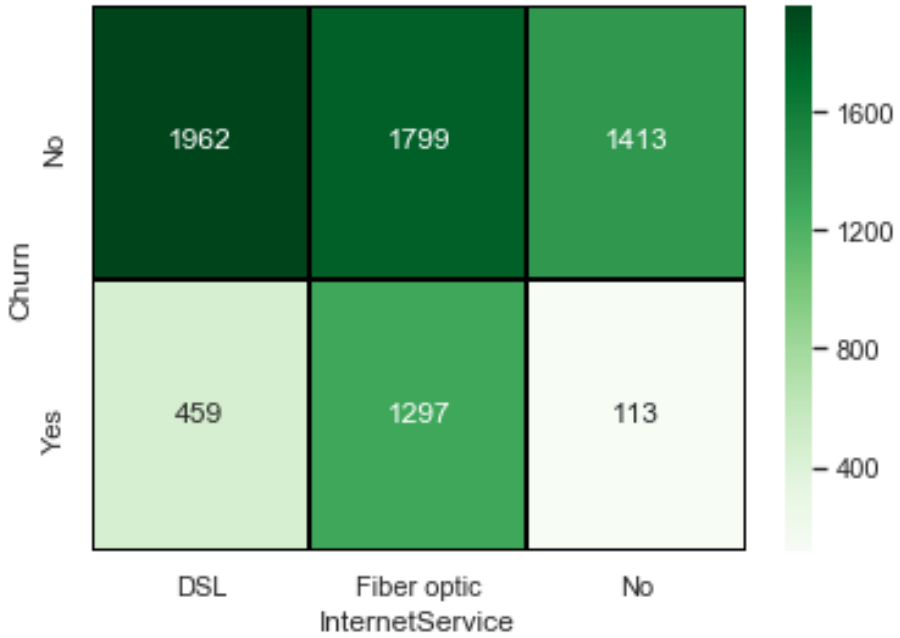
Phone Service seemingly does not have much impact on Churn since the proporti ons are the same.

Figure 29: Heatmap of MultipleLines and churn



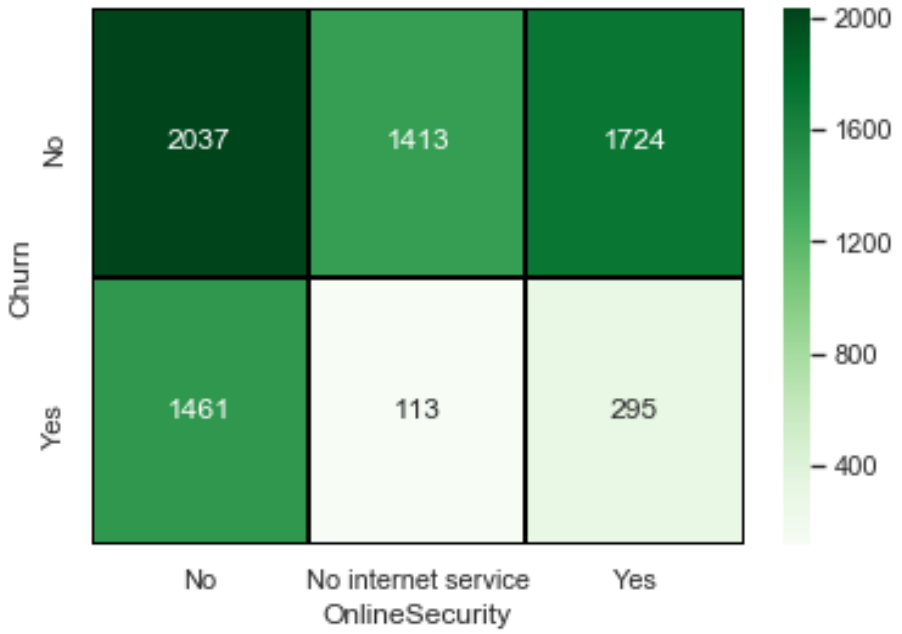
Having multiple lines also do not have seemingly important difference in the churn rate.

Figure 30: Heatmap of InternetService and churn



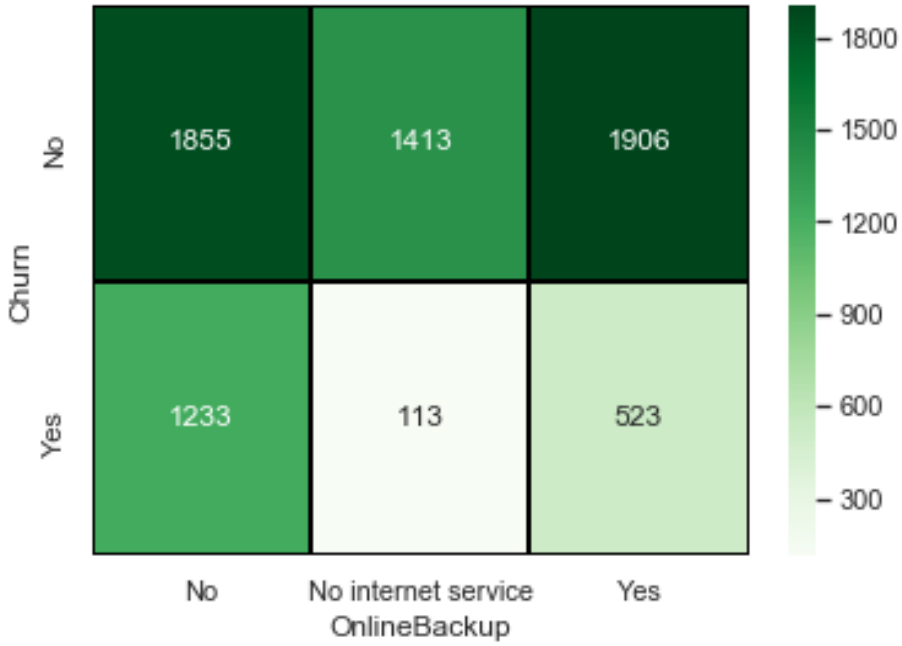
Internet Service has a very high impact on churn rate. The churn rate is espe cially high for Fiber optic customers.
They might be expecting better service.

Figure 31: Heatmap of OnlineSecurity and churn



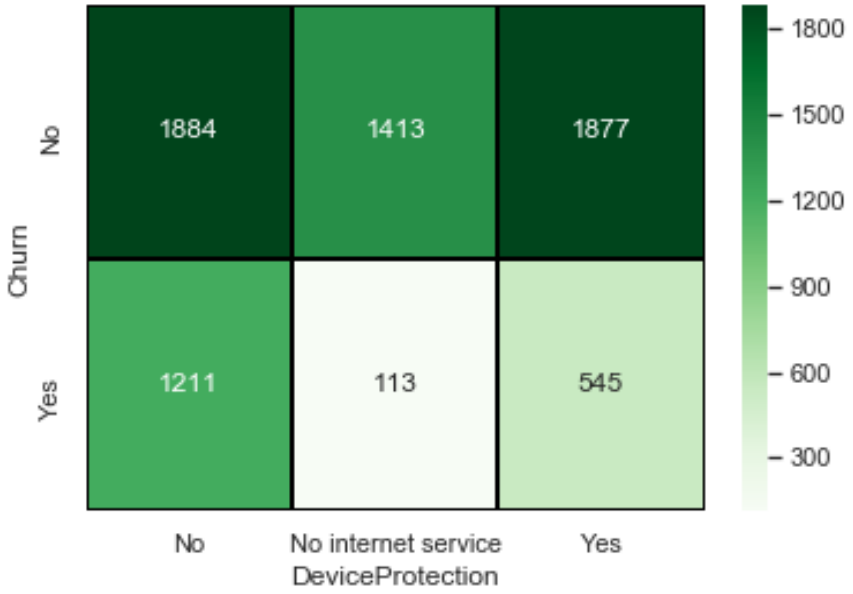
The people who do not opt for online security churn must faster as compared t
o those who do opt for it.

Figure 32: Heatmap of OnlineBackup and churn



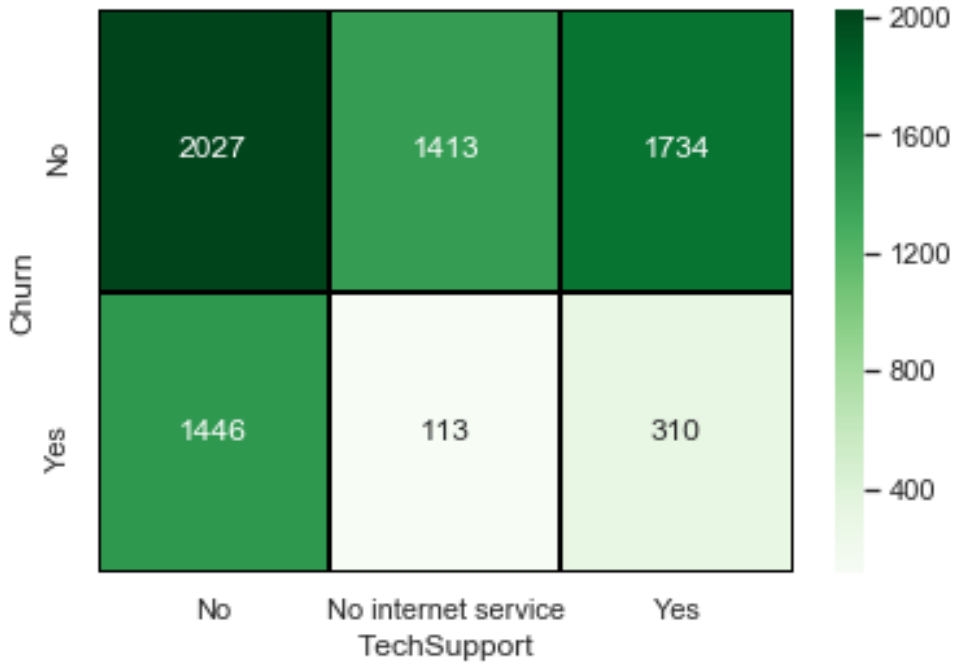
Online Backup also has a very high impact on churn rate.
The customers who opt for online backup have much lower churn rate compared t
o their counterparts.

Figure 33: Heatmap of DeviceProtection and churn



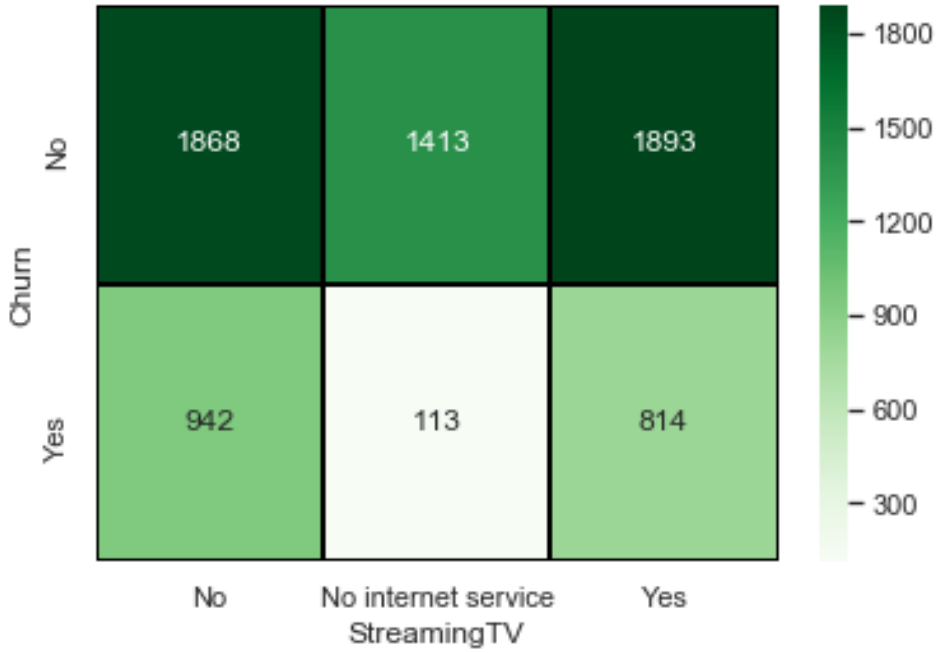
Device Protection has similar impact on churn rate as Online Backup. Those wi
thout it, have a very high churn rate.

Figure 34: Heatmap of TechSupport and churn



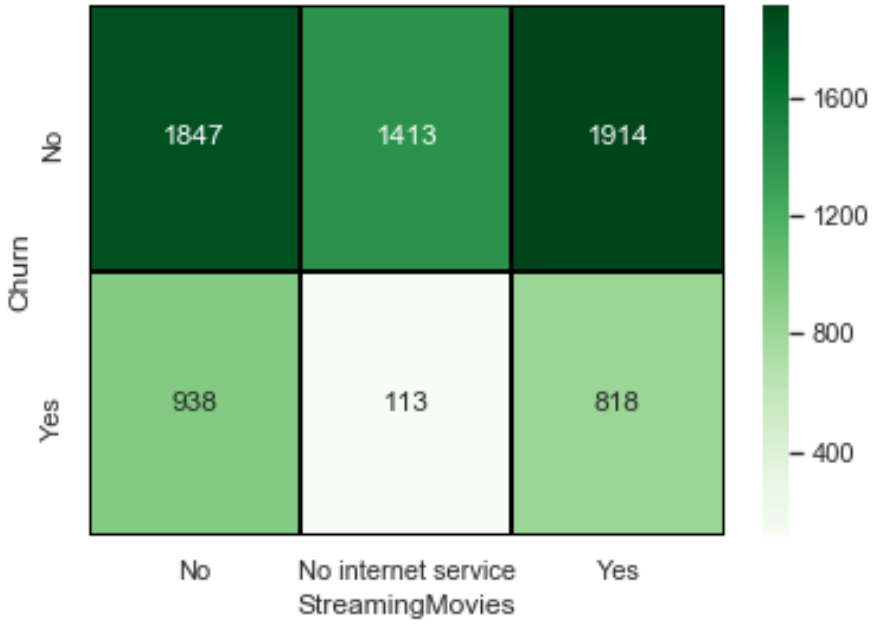
Tech Support has similar impact on churn rate as Online Security. Those witho
ut it, have a very high churn rate.
They might be expecting better service.

Figure 35: Heatmap of StreamingTV and churn



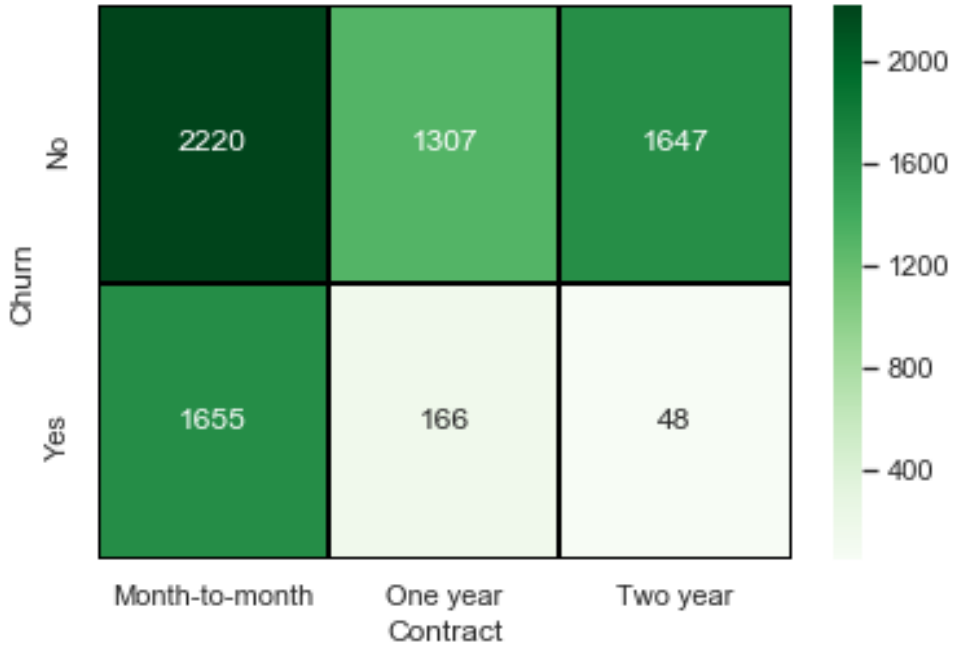
Streaming TV service does not have much correlation with churn as the proport ions of yes and no are similar.

Figure 36: Heatmap of StreamingMovies and churn



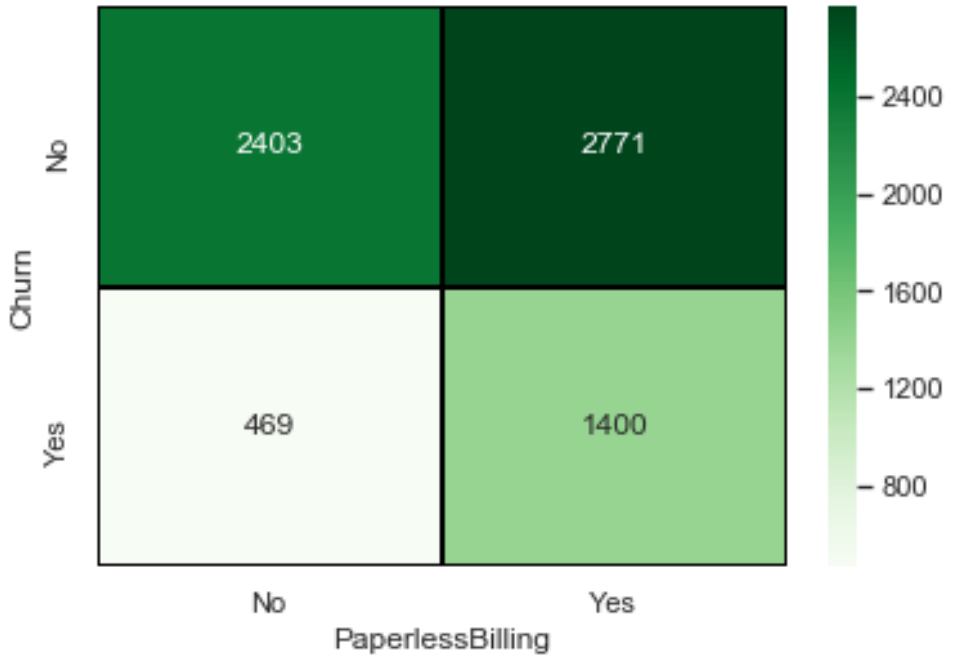
Streaming Movies service does not have much correlation with churn as the pro portions of yes and no are similar.

Figure 37: Heatmap of Contract and churn



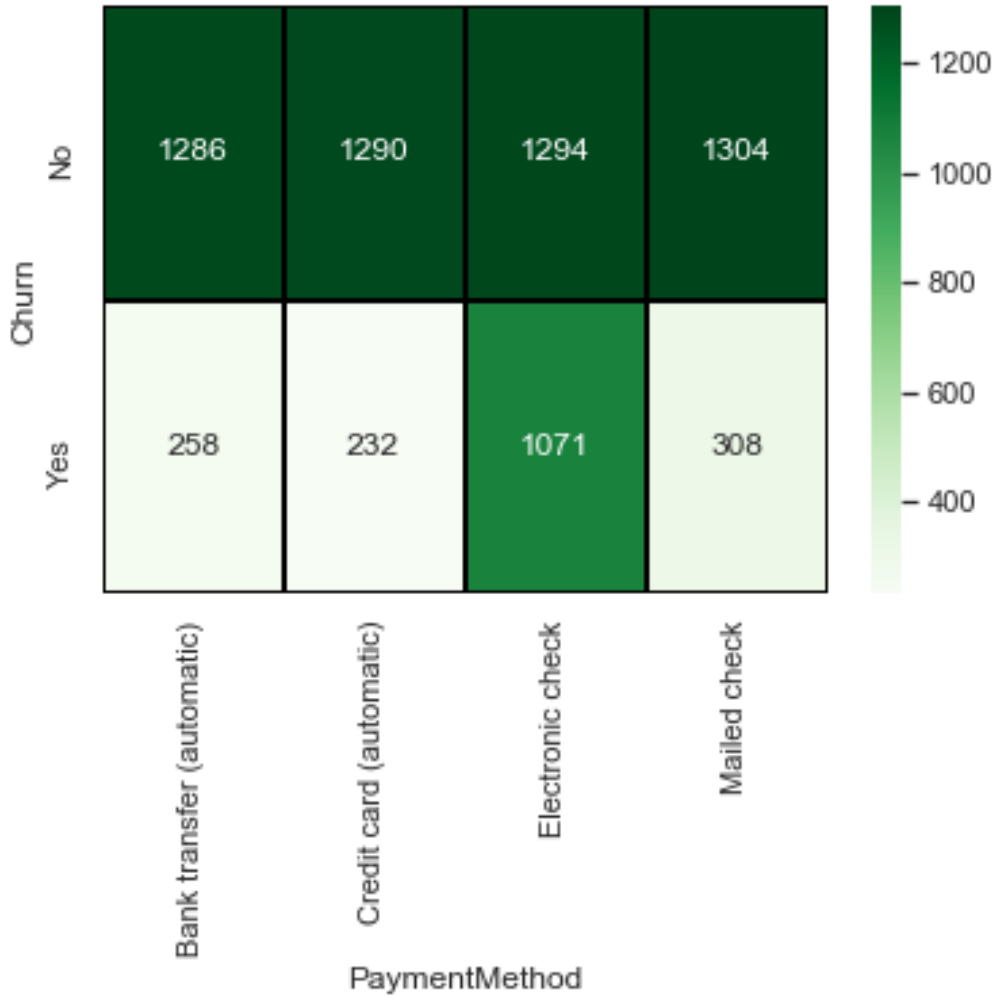
Customers with month-to-month contract have a higher chance of churning than yearly or bi-yearly contracts due to flexibility.

Figure 38: Heatmap of PaperlessBilling and churn



Customers with paperless billing churn more than those without.
This is similar to the Senior Citizens case as these people do not like to ch
ange

Figure 39: Heatmap of PaymentMethod and churn



People who pay using Electronic check iare very highly likely to churn than t hose that pay using other methods.

Conclusions

The conclusion of the analysis is that all the numeric variables, that is tenure, MonthlyCharges and TotalCharges are significant to identify churning customers. Further, categorical columns such as SeniorCitizen, Partner, Dependents, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, Contract, PaperlessBilling and PaymentMethod have significant effect on churn rate while the other features do not have a seemingly high impact on churn rate.

Recommendations

Services such as Online Security, Online Backup, Device Protection, Tech Support have high impact on churn rate. Thus, customers should be given lucrative offers to opt-in for these services. Feedback must be taken from the fiber optic internet users and Electronic check payment users on their expectations and problems faced from the service/preference to retain them better. New plans must be introduced for customers without partners or dependents as the customers with partners or dependents are most likely satisfied with their bundle plans.

References

[1] "Churn Rate", Investopedia, 2019. [Online]. Available: https://www.investopedia.com/terms/c/churnrate.asp.

[2] "Using Customer Behavior Data to Improve Customer Retention", IBM Analytics Communities, 2019. [Online]. Available: https://www.ibm.com/communities/analytics/watson-analytics-blog/predictive-insights-in-the-telco-customer-churn-data-set/.