SPFNA

TELCO CUSTOMER CHURN

CPE213 Data Modeling

Group Member

SPFNA

Chanon Khanijoh	3408
Tunwa Satianrapapong	3419
Napas Vinitnantharat	3422
Pechdanai Saepong	3434
Fasai Sae-Tae	3436

Objective of Telco Customer Churn Project

To understand customer behaviour
How they using company product.

To detecting which customers are likely to leave a service or to cancel a subscription to a service

Reduce company churn rate. which make company to higher profit margin

Dataset (1)

COLUMN NAME	CustomerID	Gender	SeniorCitizen	Partner	Dependents
DESCRIPTION	Customer ID	The customer’s gender: Male, Female	Indicates if the customer is 65 or older: Yes, No	Indicates if the customer is married: Yes, No	Indicates if the customer lives with any dependents: Yes, No. Dependents could be children, parents, grandparents, etc.

Dataset (2)

COLUMN NAME	Tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity
DESCRIPTION	Number of months the customer has stayed with the company	Indicates if the customer subscribes to home phone service with the company: Yes, No	Indicates if the customer subscribes to multiple telephone lines with the company: Yes, No	Indicates if the customer subscribes to Internet service with the company: No, DSL, Fiber Optic, Cable.	Indicates if the customer subscribes to an additional online security service provided by the company: Yes, No

Dataset (3)

COLUMN NAME	OnlineBackup	DeviceProtection	TechSupport	StreamingTV	StreamingMovies
DESCRIPTION	Indicates if the customer subscribes to an additional online backup service provided by the company: Yes, No	Indicates if the customer subscribes to an additional device protection plan for their Internet equipment provided by the company: Yes, No	Indicates if the customer subscribes to an additional technical support plan from the company with reduced wait times: Yes, No	Indicates if the customer uses their Internet service to stream television programing from a third party provider: Yes, No. The company does not charge an additional fee for this service.	Indicates if the customer uses their Internet service to stream movies from a third party provider: Yes, No. The company does not charge an additional fee for this service.

Dataset (4)

COLUMN NAME	Contract	PaperlessBilling	PaymentMethod	MonthlyCharges
DESCRIPTION	Indicates the customer's current contract type: Month-to-Month, One Year, Two Year.	Indicates if the customer has chosen paperless billing: Yes, No	Indicates how the customer pays their bill: Bank Withdrawal, Credit Card, Mailed Check	Indicates the customer's current total monthly charge for all their services from the company.

Dataset (4)

COLUMN NAME	TotalCharges	Churn (Target prediction) ★
DESCRIPTION	Indicates the customer's total charges, calculated to the end of the quarter specified above.	Yes = the customer left the company this quarter. No = the customer remained with the company. Directly related to Churn Value.

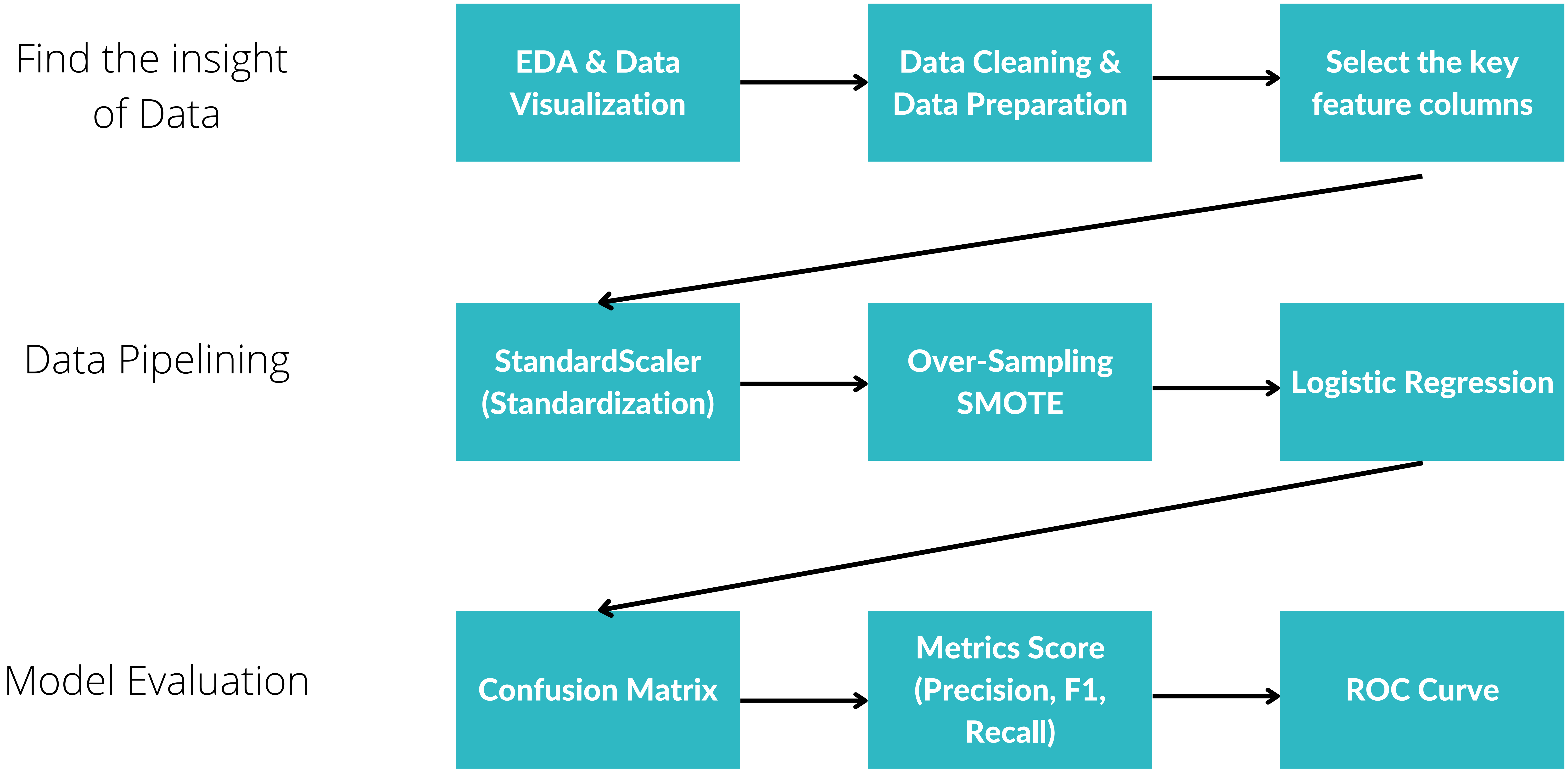
Tools and Techniques

- Data Management

- Data Visualization

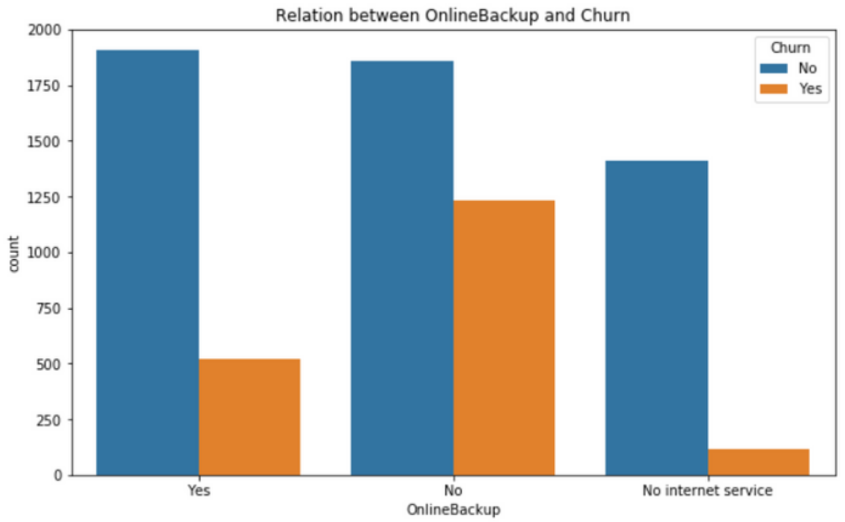
- Statistic model and ML library

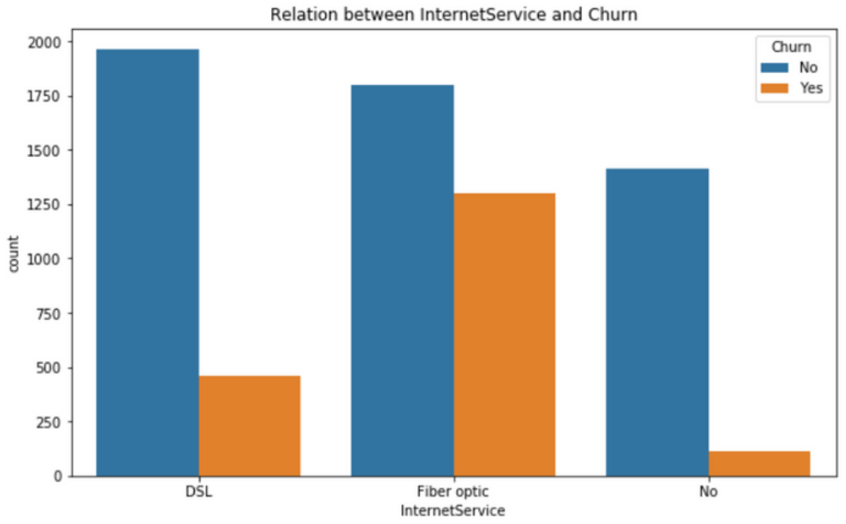
Flowchart of Processing

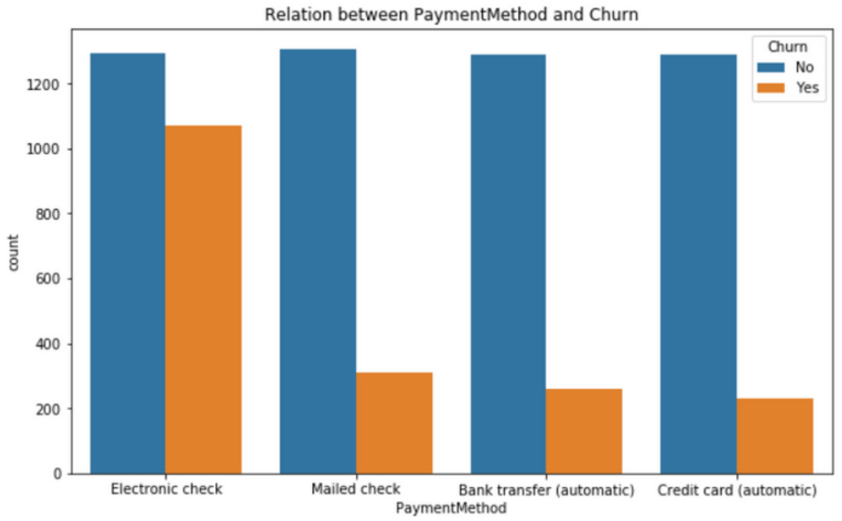


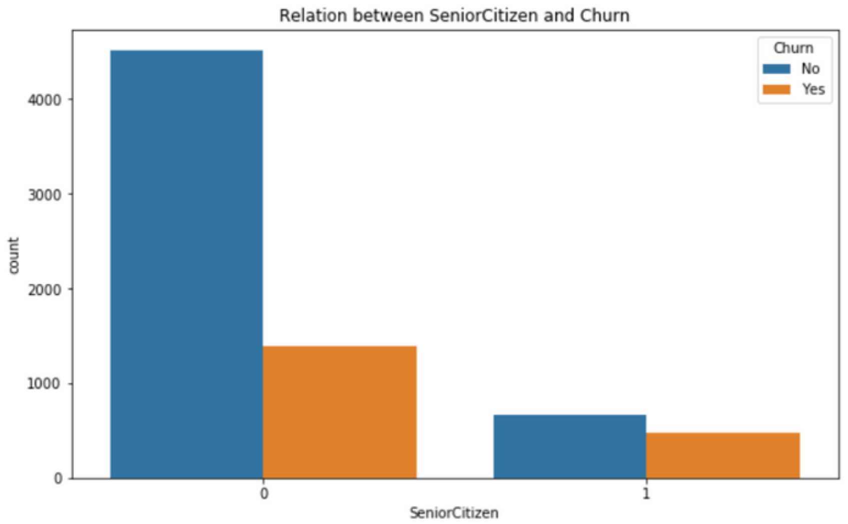
EDA

Exploratory data analysis
and
Data Visualization

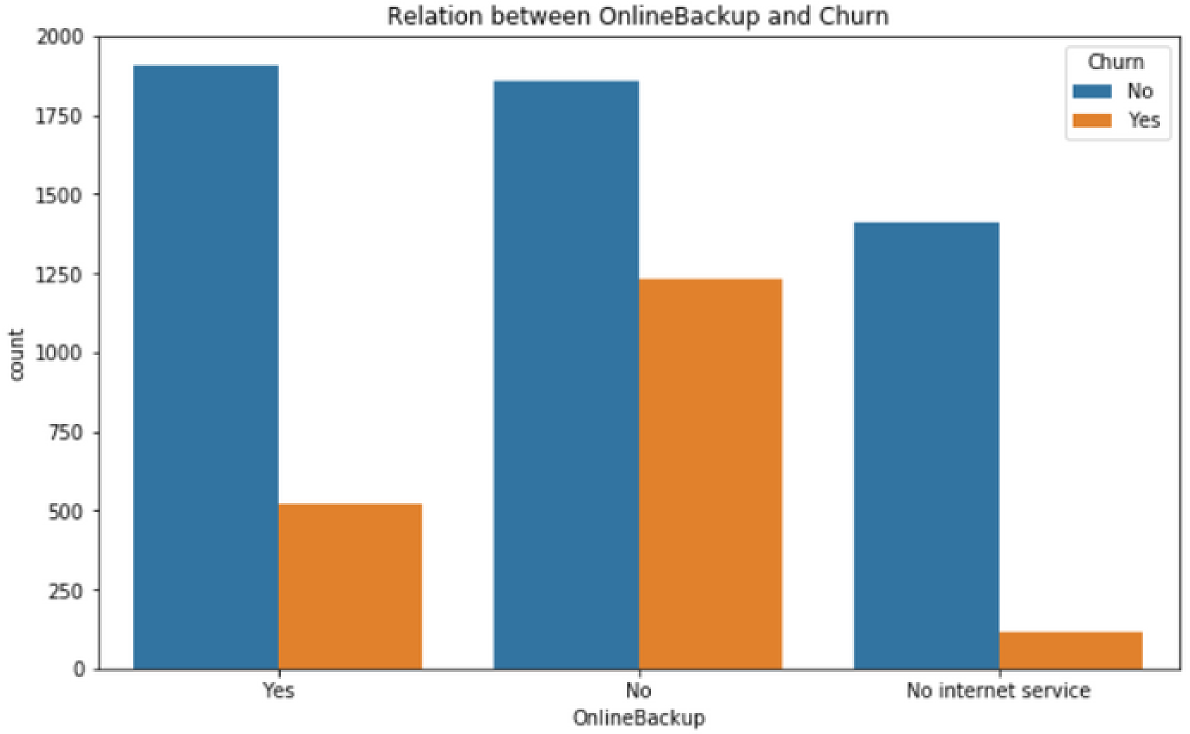


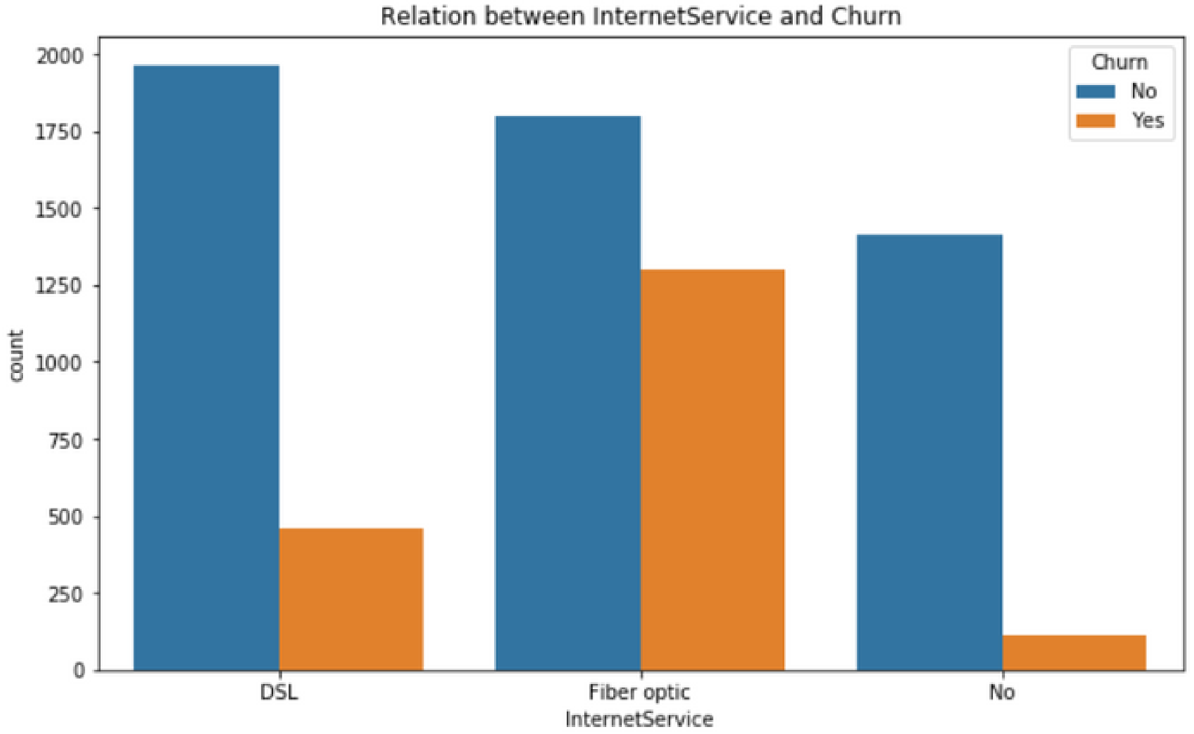


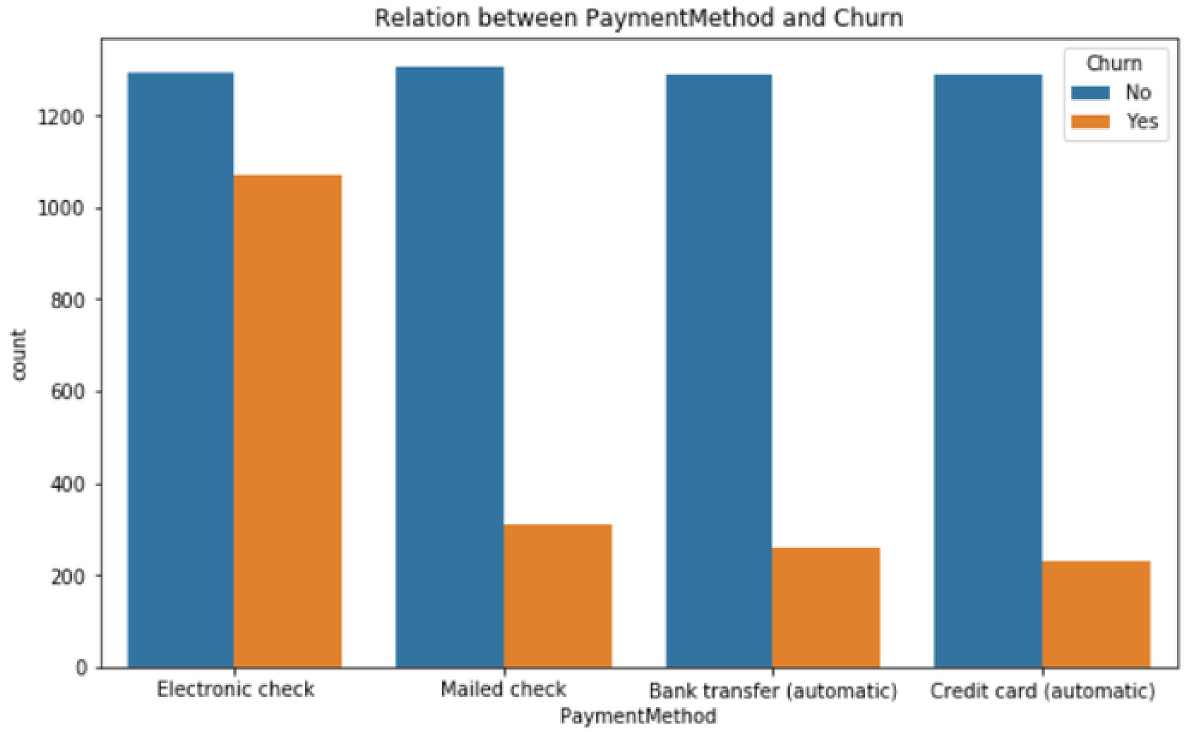


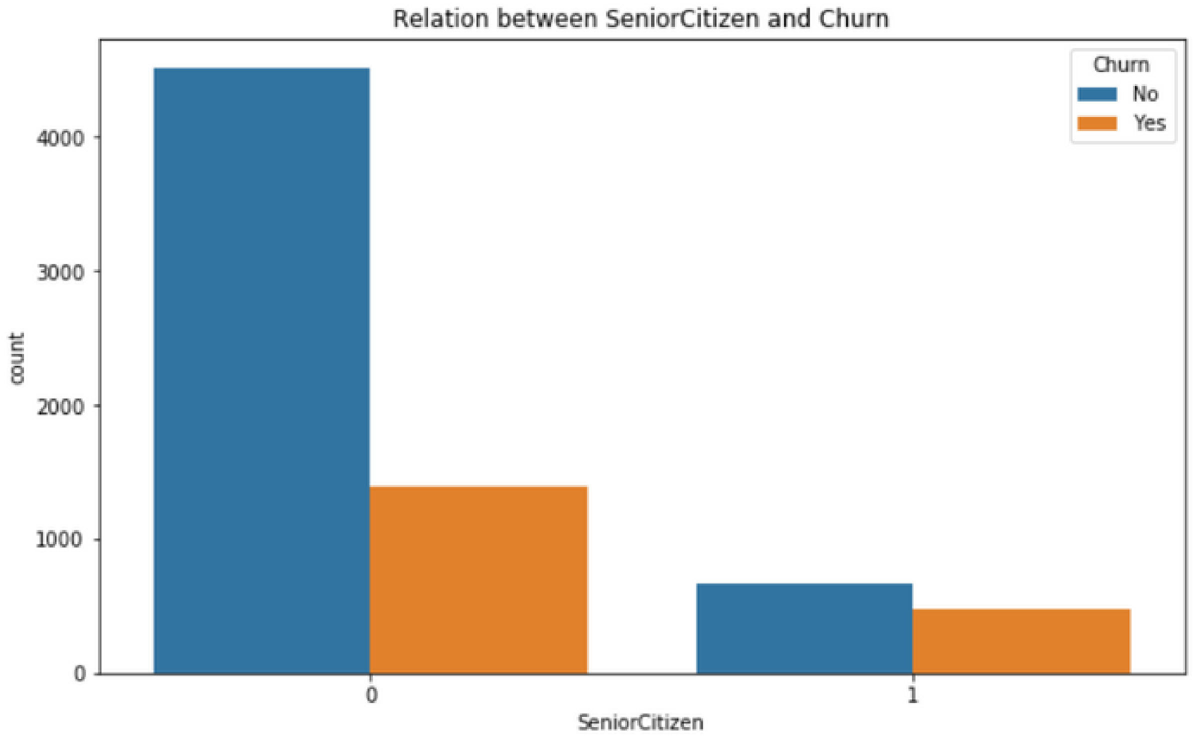


Graph of relation between dataset and churn

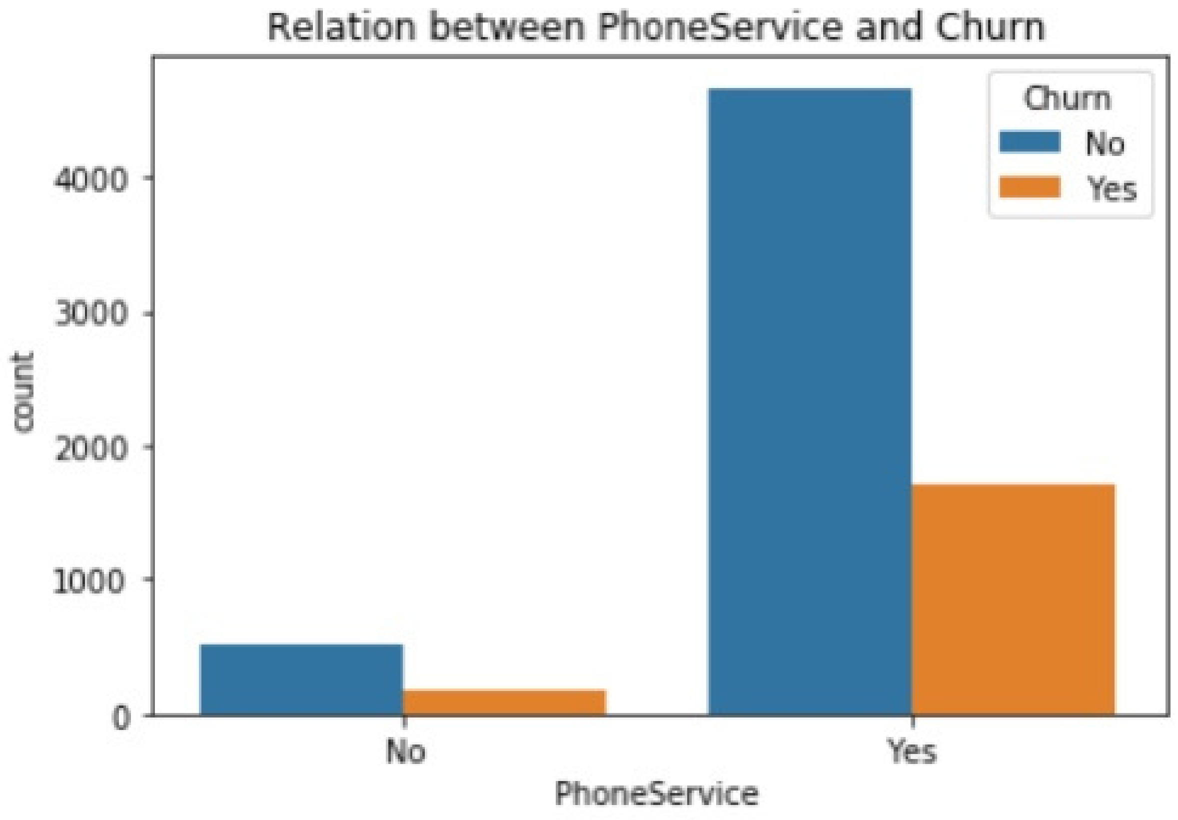


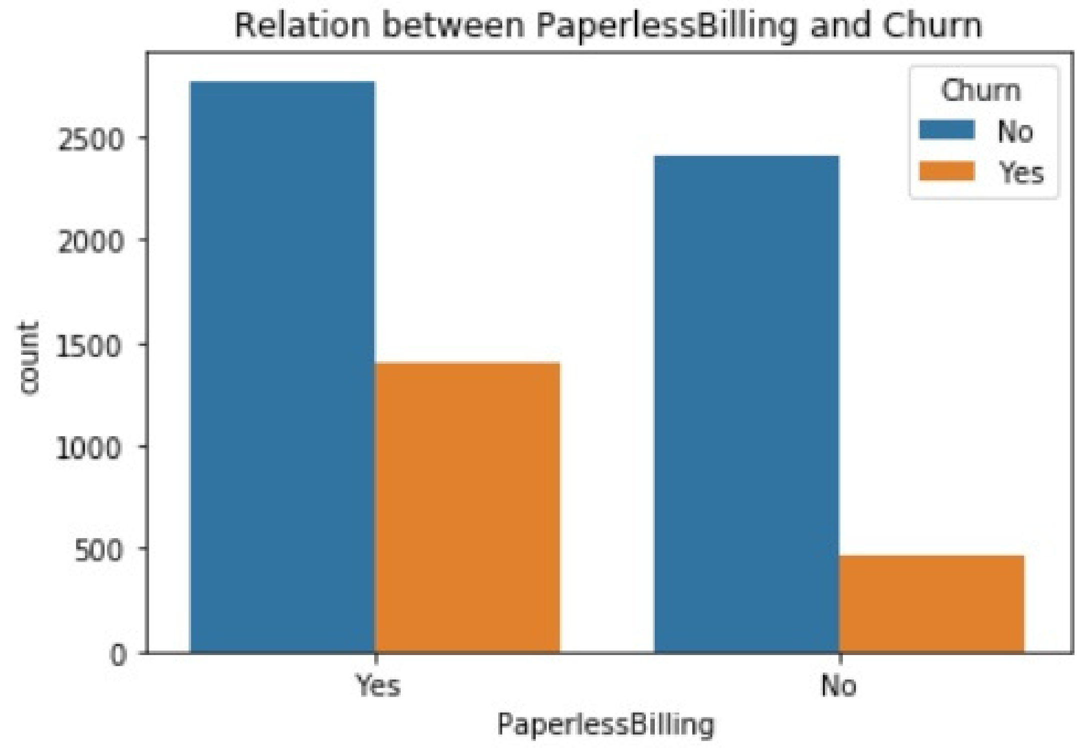


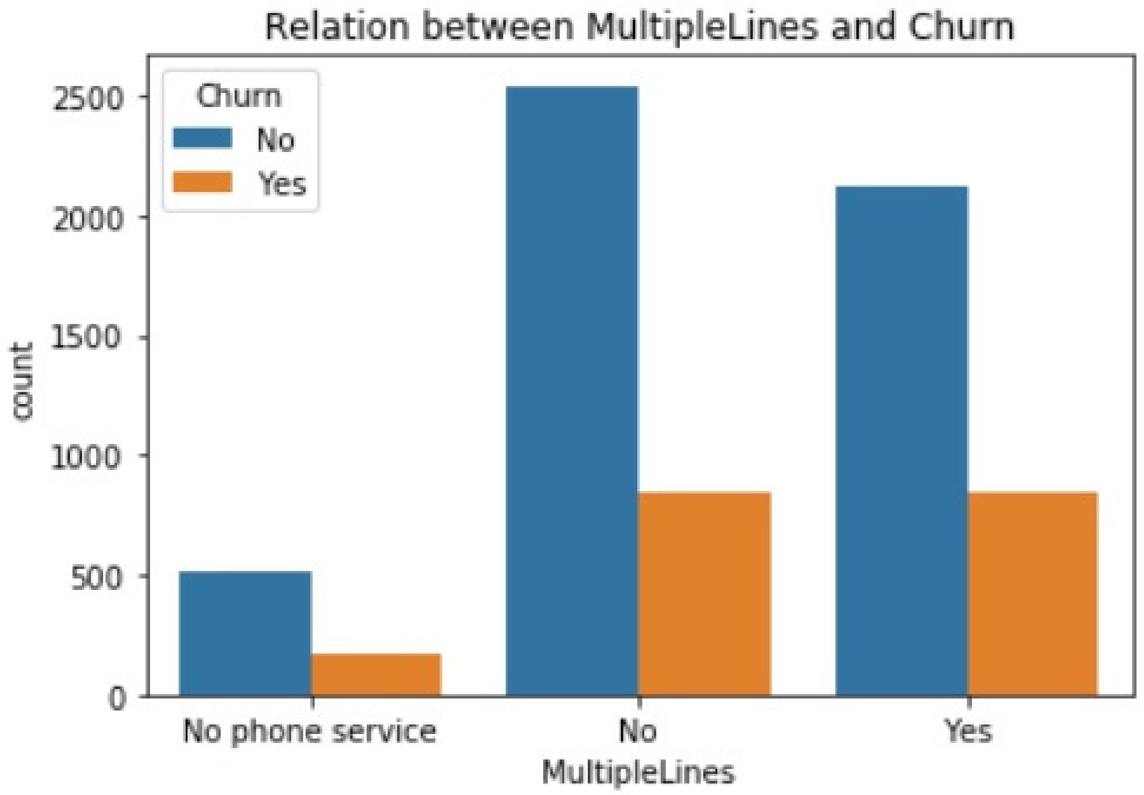


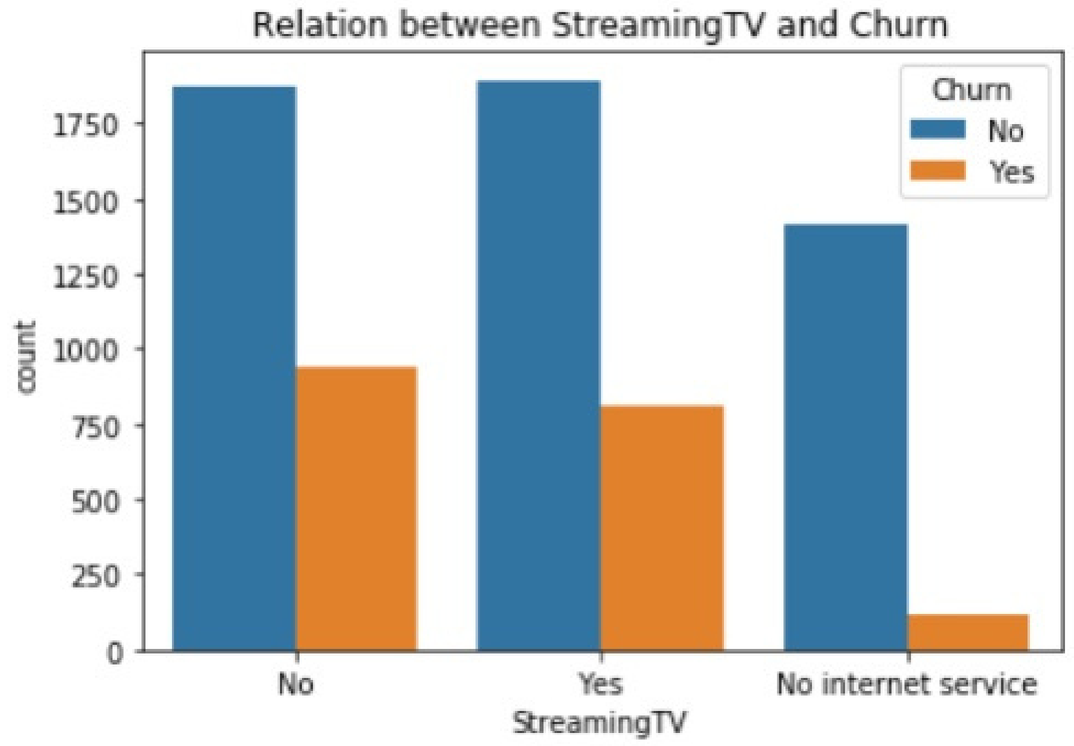


Graph of relation between dataset and churn

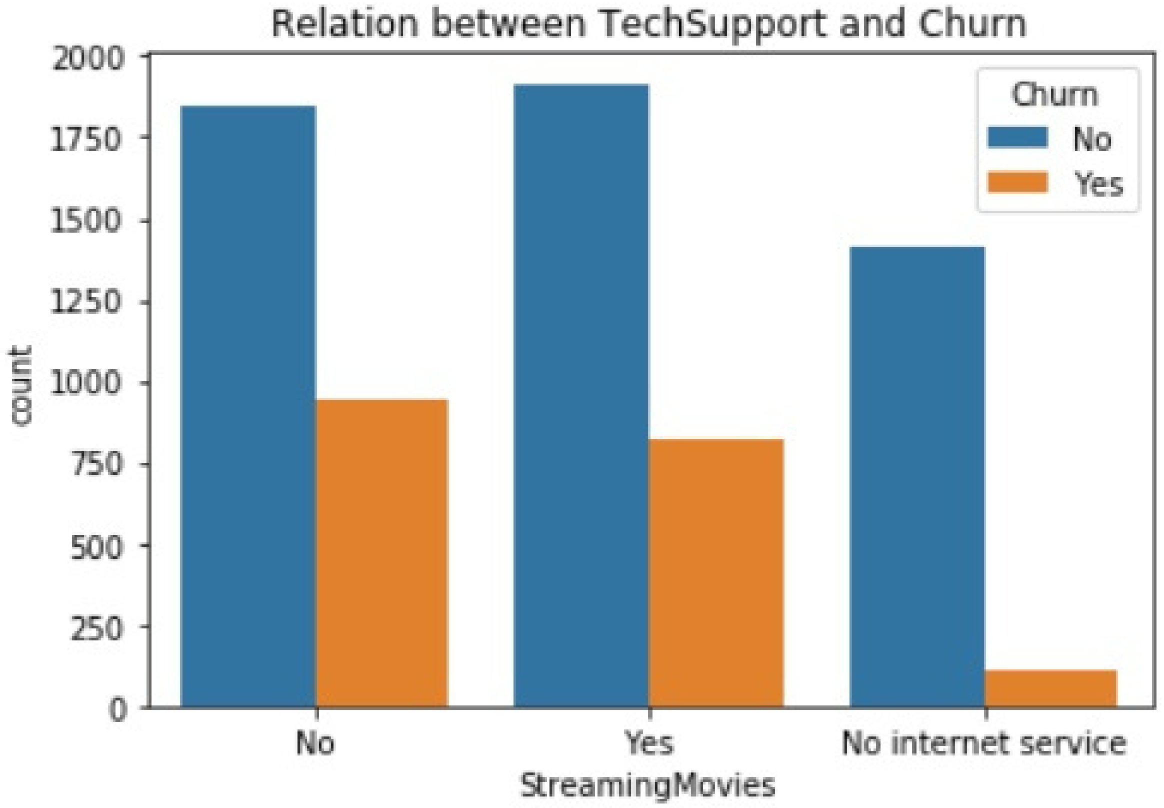


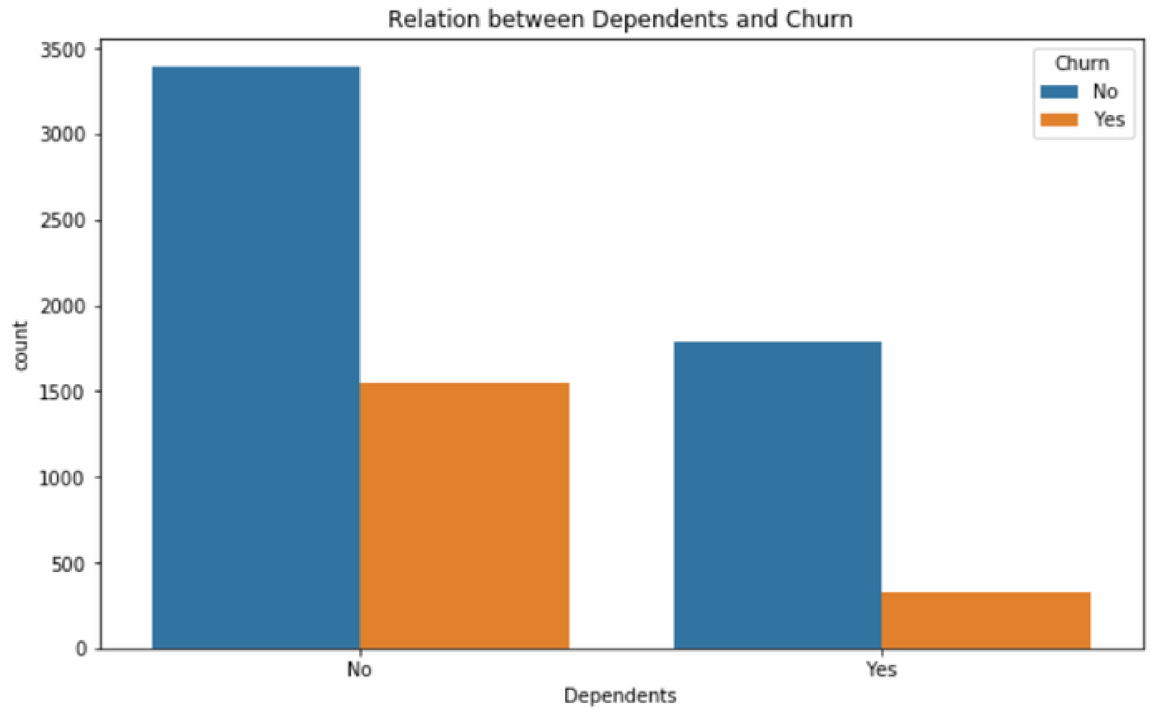


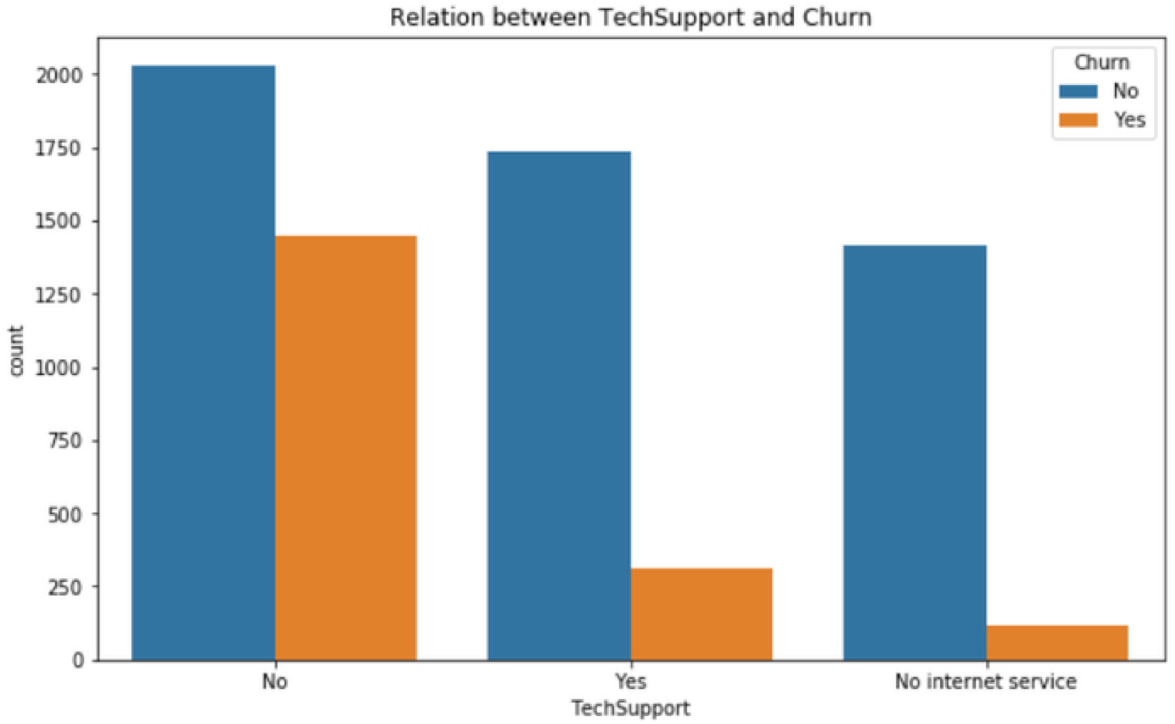


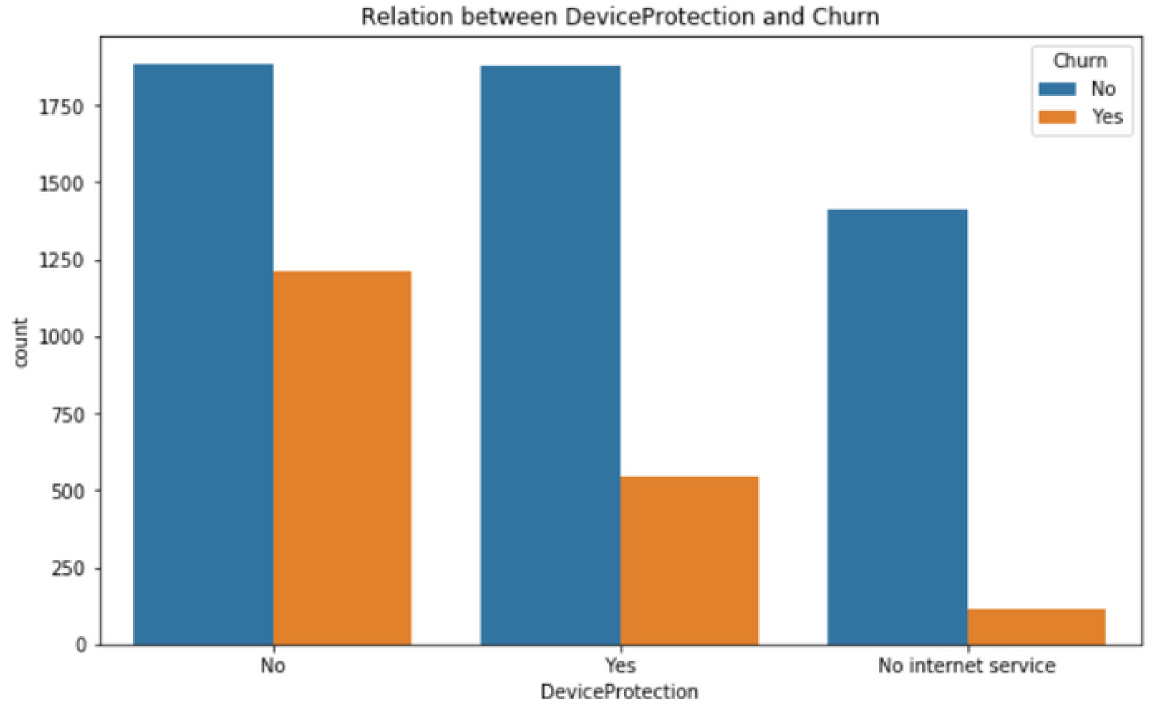


Graph of relation between dataset and churn

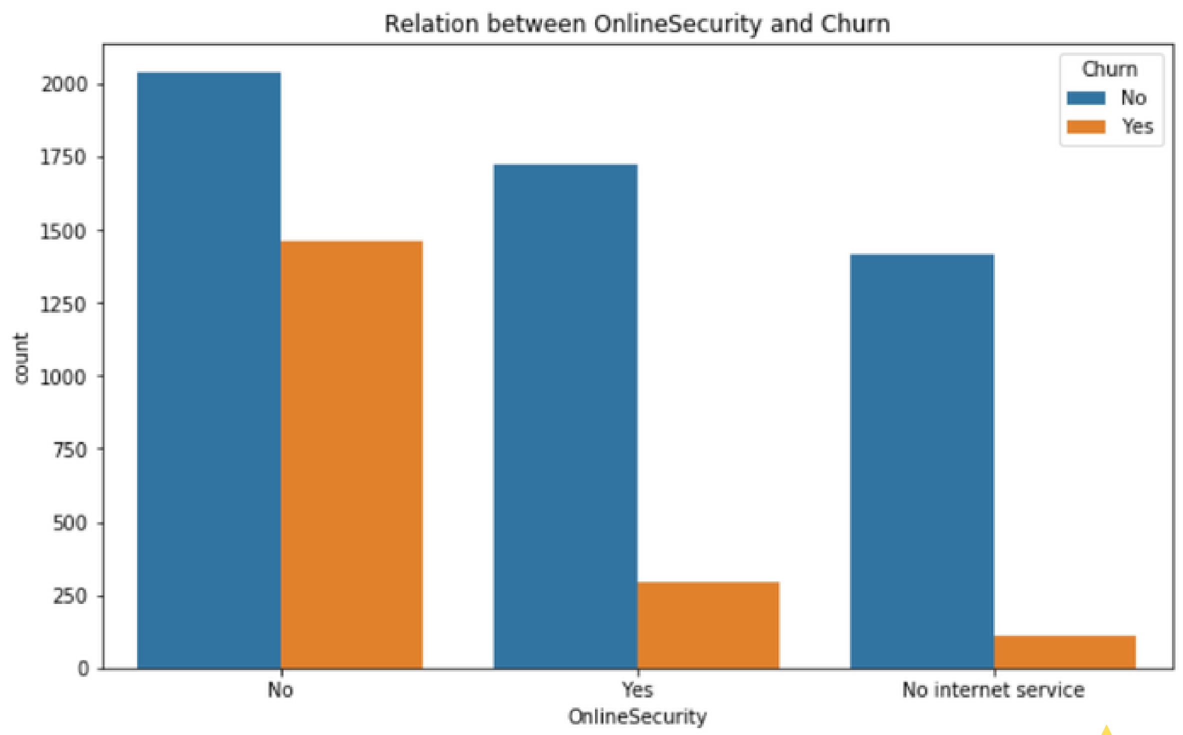


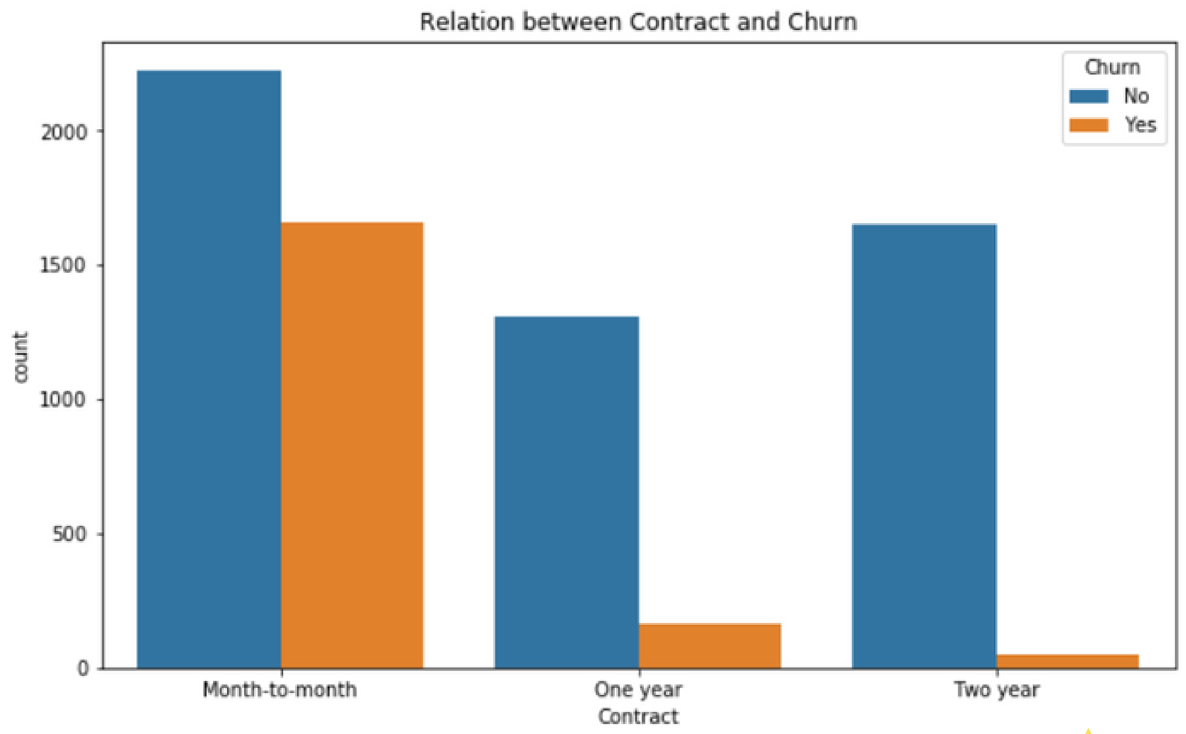




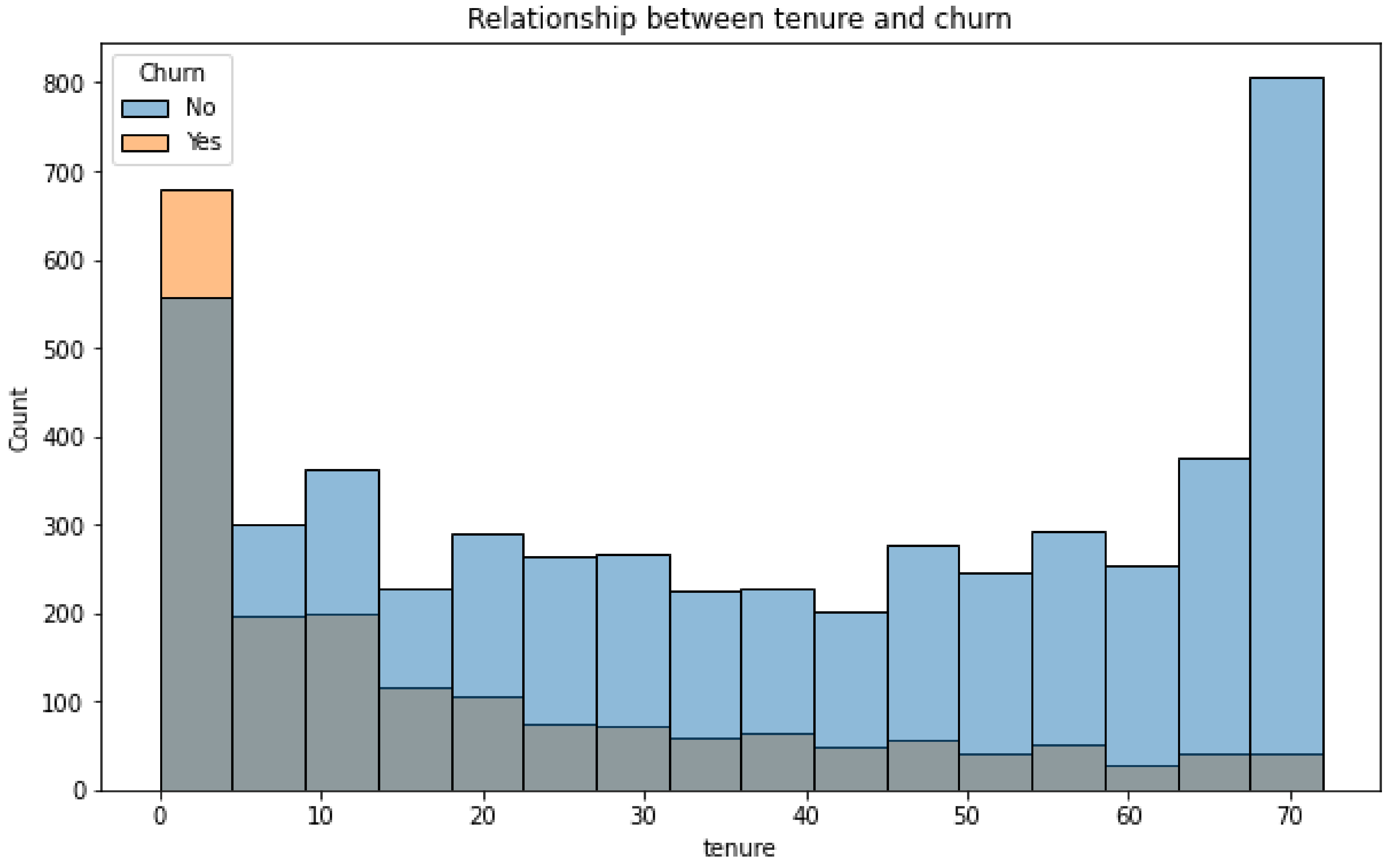


Graph of relation between dataset and churn



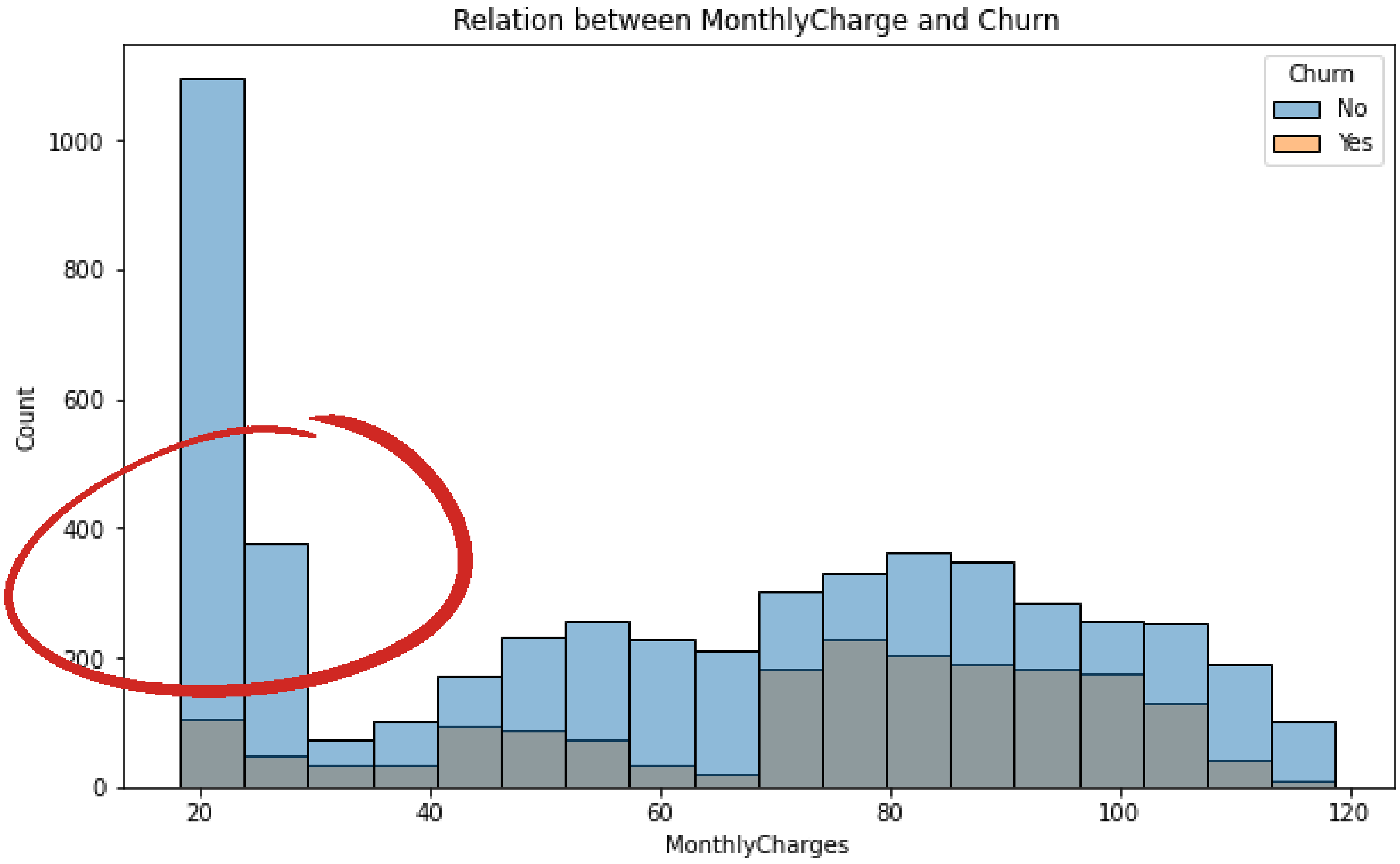


Graph of relation between dataset and churn



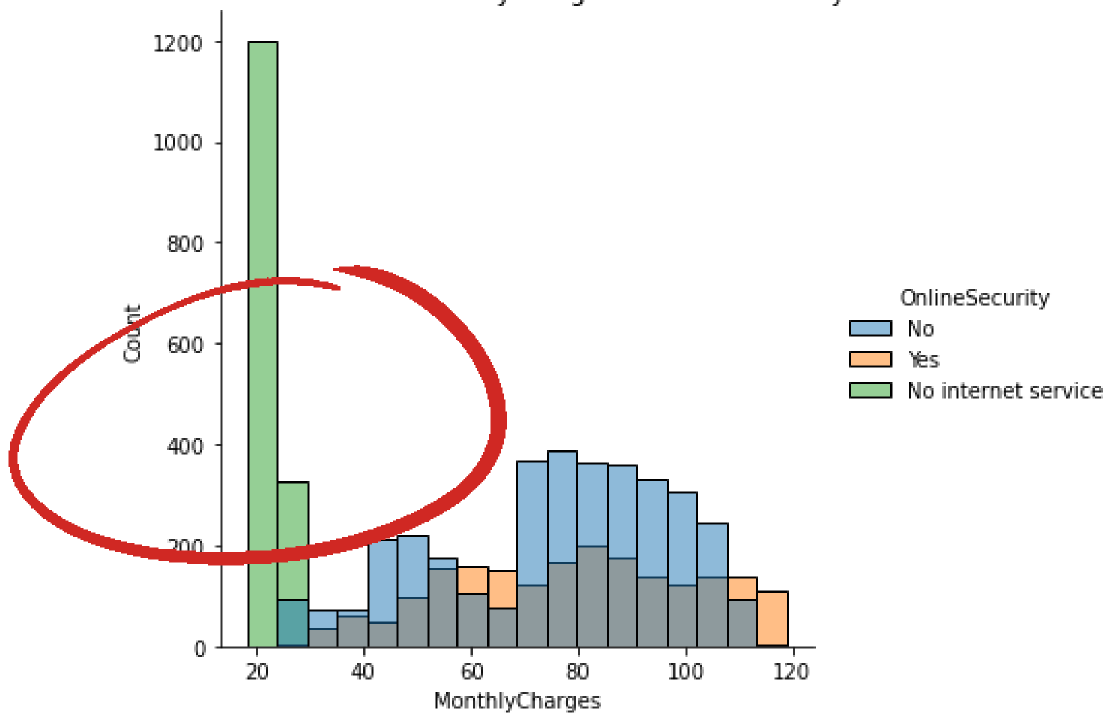
This graph show that as long as the customer sticks to the company product, the less likely the customer will churn.

Graph of relation between dataset and churn

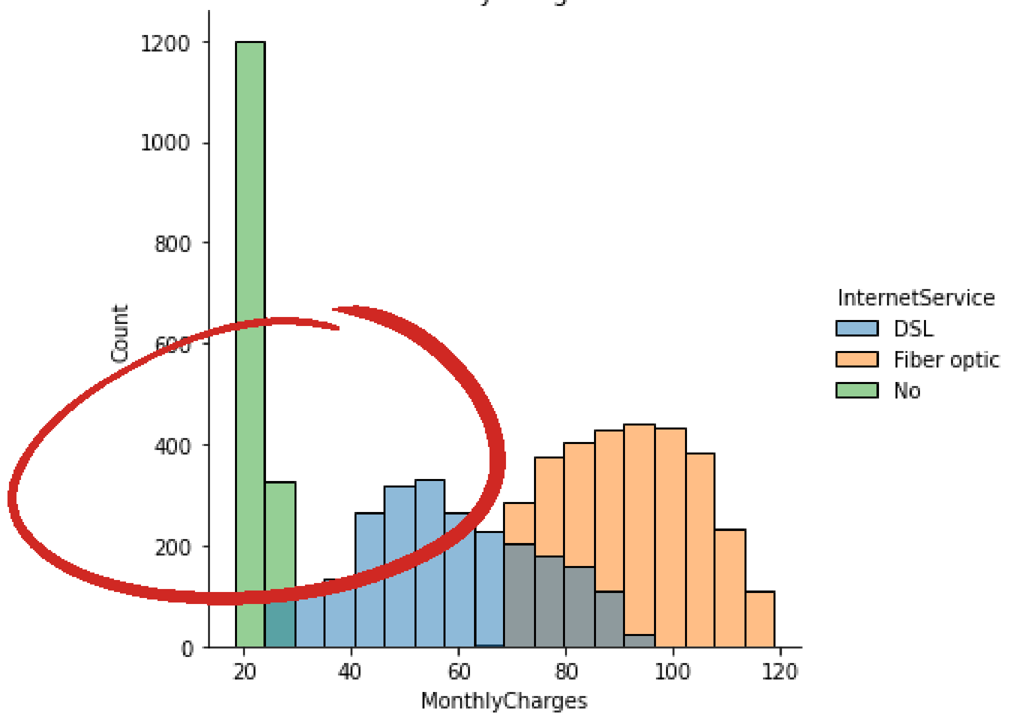


Graph of relation between dataset and churn

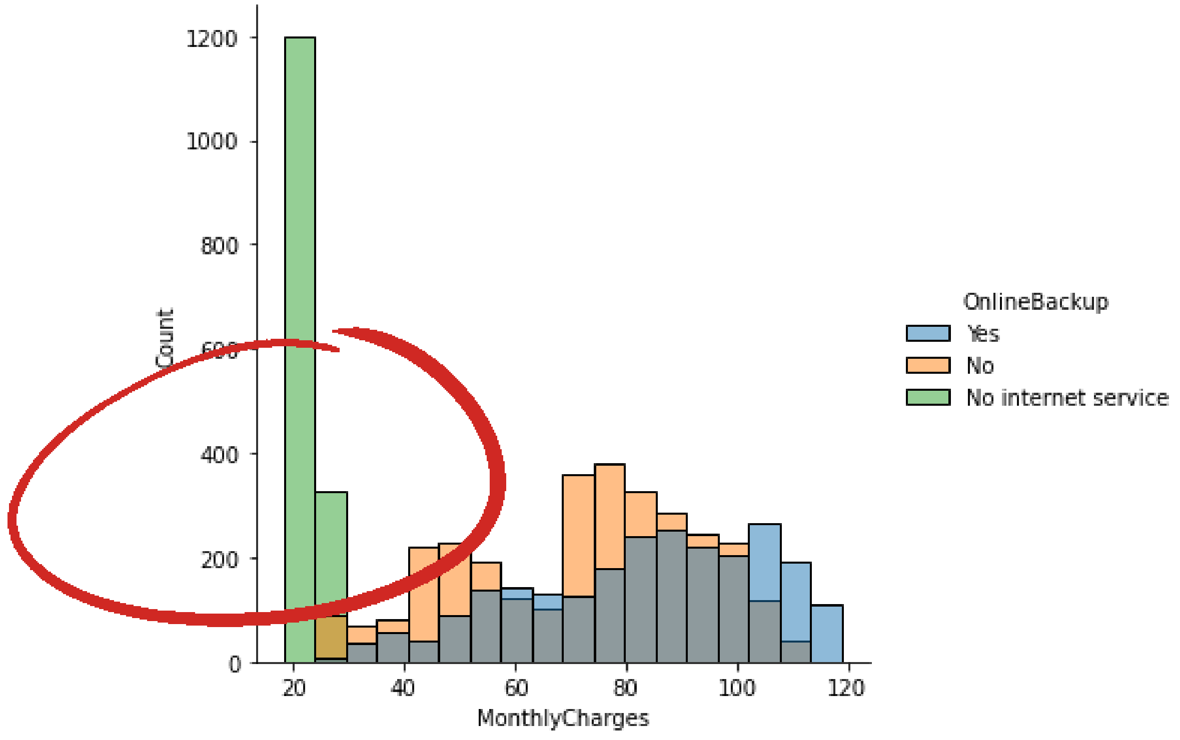
Relation between MonthlyCharge and OnlineSecurity



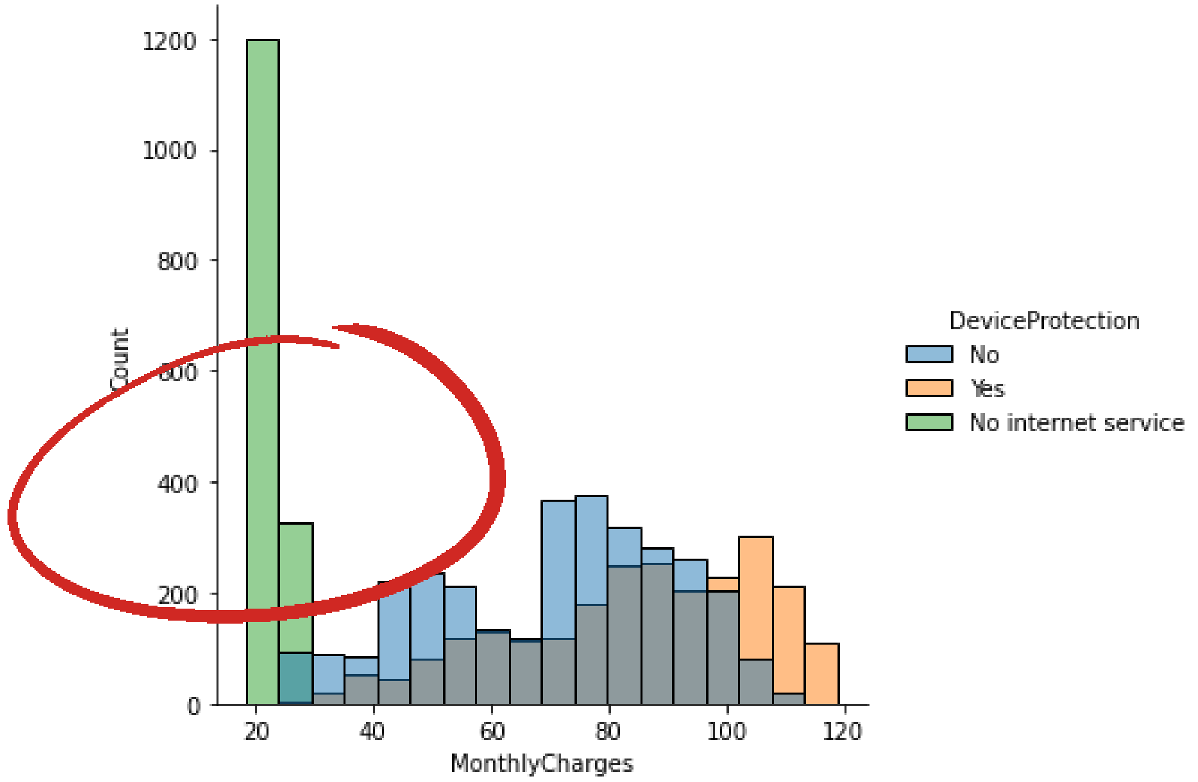
Relation between MonthlyCharge and InternetService



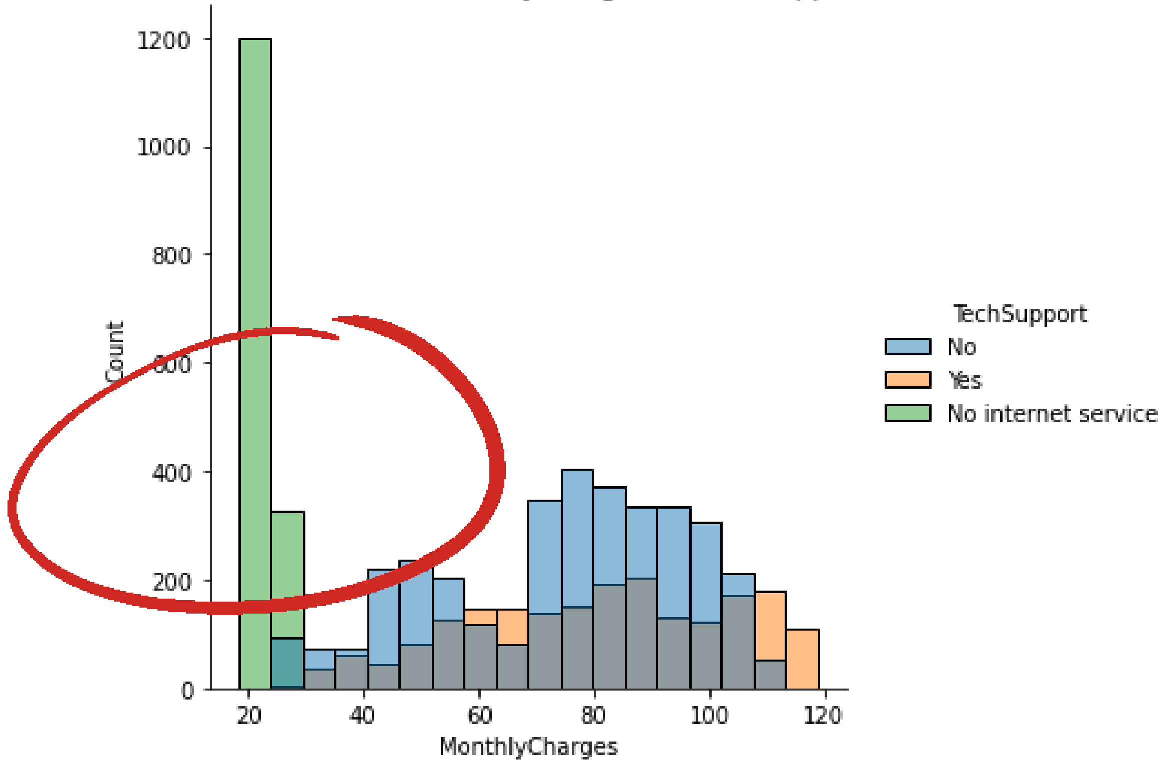
Relation between MonthlyCharge and OnlineBackup



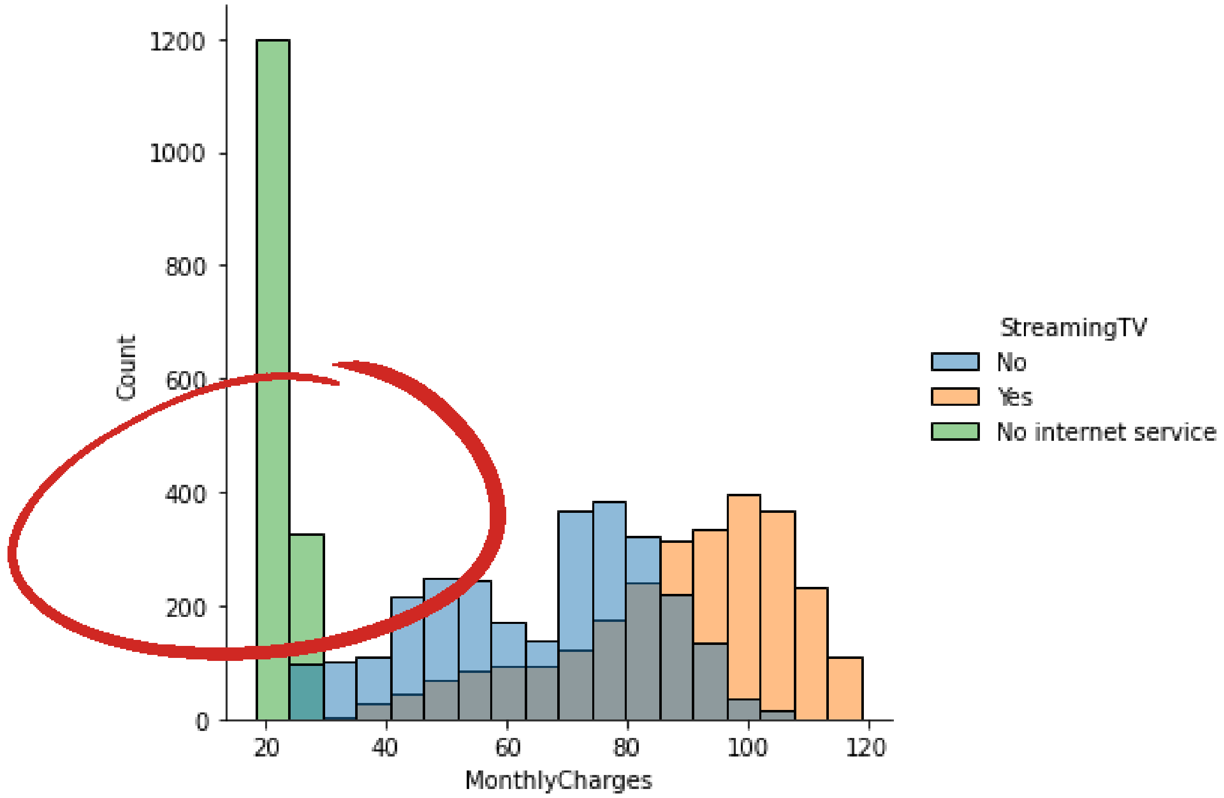
Relation between MonthlyCharge and DeviceProtection



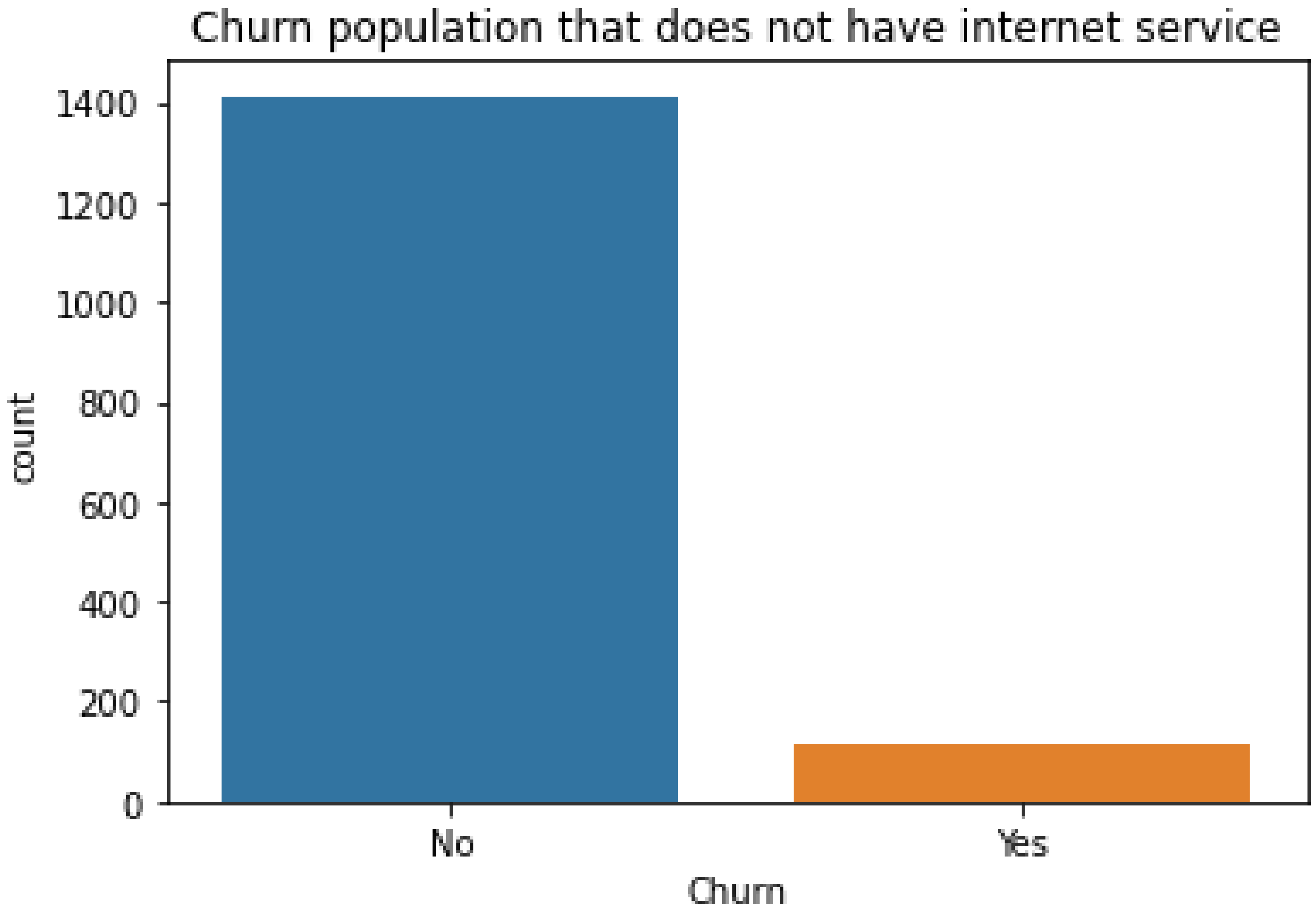
Relation between MonthlyCharge and TechSupport



Relation between MonthlyCharge and StreamingTV



observation in population who doesn't have internet

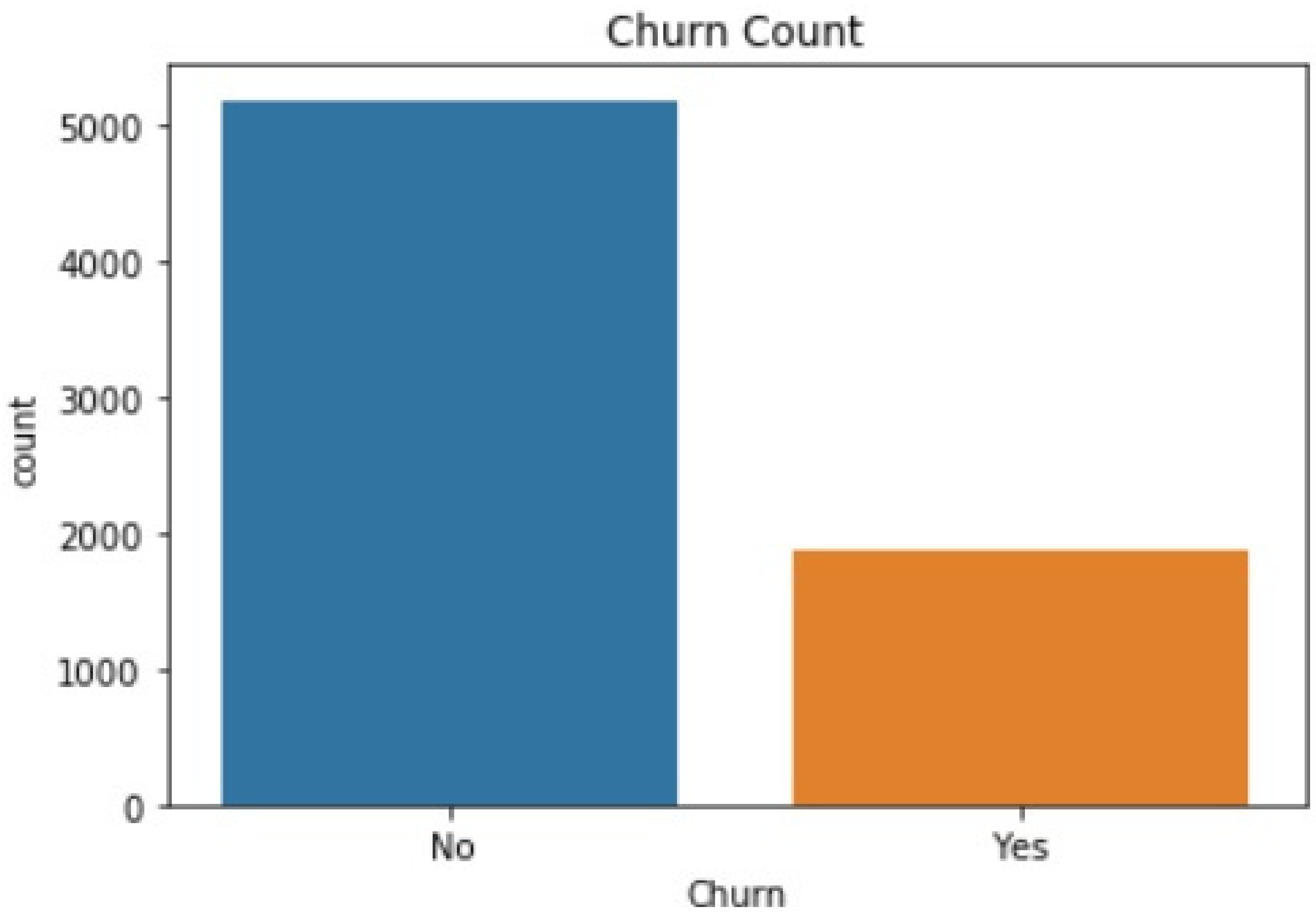


percentage of churn customer who doesn't have internet service is 7.40

percentage of not churn customer who doesn't have internet service is 92.59

seems to be a good predictor of the outcome variable

Graph of relation between dataset and churn



percentage of churn customer
is 73.46%

percentage of not churn customer
is 26.53%

This plot we can summarize that
the target is imbalance dataset

Data preparation

1. Select feature base on data exploration

2. One Hot encoded catergorical feature (Creating dummies)

3. Standardise features by removing the mean and scaling to unit variance

4. Over-sampling using SMOTE

Select feature base on data exploration

```
df_model = df[['tenure','Contract','OnlineSecurity','InternetService','PaymentMethod', 'Churn']]
df_model
```

	tenure	Contract	OnlineSecurity	InternetService	PaymentMethod	Churn
0	1	Month-to-month	No	DSL	Electronic check	No
1	34	One year	Yes	DSL	Mailed check	No
2	2	Month-to-month	Yes	DSL	Mailed check	Yes
3	45	One year	Yes	DSL	Bank transfer (automatic)	No
4	2	Month-to-month	No	Fiber optic	Electronic check	Yes
...	...	...	...	...	...	...
7038	24	One year	Yes	DSL	Mailed check	No
7039	72	One year	No	Fiber optic	Credit card (automatic)	No
7040	11	Month-to-month	Yes	DSL	Electronic check	No
7041	4	Month-to-month	No	Fiber optic	Mailed check	Yes
7042	66	Two year	Yes	Fiber optic	Bank transfer (automatic)	No

7043 rows × 6 columns

One Hot encoded catergorical feature (Creating dummies)

```
contract = pd.get_dummies(df_model['Contract'],prefix='Contract')
onlinesecurity = pd.get_dummies(df_model['OnlineSecurity'],prefix='OnlineSecurity')
payment = pd.get_dummies(df_model['PaymentMethod'],prefix='PaymentMethod')
internet = pd.get_dummies(df_model['InternetService'],prefix='InternetService')
```

```
df_model = pd.concat([df_model, contract, onlinesecurity, payment,internet], axis=1)
```

```
df_model.drop(['Contract','OnlineSecurity','PaymentMethod', 'InternetService'], axis=1, inplace=True)
```

```
#df_model.drop(['OnlineSecurity_No internet service'], axis=1, inplace=True)
df_model.drop(['InternetService_No'], axis=1, inplace=True)
```

```
df_model['Churn'] = df_model['Churn'].map({'Yes':1, 'No':0})
df_model
```

	tenure	Churn	Contract_Month-to-month	Contract_One year	Contract_Two year	OnlineSecurity_No	OnlineSecurity_No internet service	OnlineSecurity_Yes	PaymentMethod_Bank transfer (automatic)	PaymentMethod_Credit card (automatic)
0	1	0	1	0	0	1	0	0	0	0
1	34	0	0	1	0	0	0	1	0	0
2	2	1	1	0	0	0	0	1	0	0
3	45	0	0	1	0	0	0	1	1	0
4	2	1	1	0	0	1	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...
7038	24	0	0	1	0	0	0	1	0	0
7039	72	0	0	1	0	1	0	0	0	1
7040	11	0	1	0	0	0	0	1	0	0
7041	4	1	1	0	0	1	0	0	0	0
7042	66	0	0	0	1	0	0	1	1	0

7043 rows × 14 columns

Standardise features

```
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
df_model[['tenure']] = scaler.fit_transform(df_model[['tenure']])
df_model
```

	tenure	Churn	Contract_Month-to-month	Contract_One year	Contract_Two year	OnlineSecurity_No	OnlineSecurity_No internet service	OnlineSecurity_Yes	PaymentMethod_Bank transfer (automatic)	PaymentMethod_Credi card (automatic
0	-1.277445	0	1	0	0	1	0	0	0	[illegible]
1	0.066327	0	0	1	0	0	0	1	0	[illegible]
2	-1.236724	1	1	0	0	0	0	1	0	[illegible]
3	0.514251	0	0	1	0	0	0	1	1	[illegible]
4	-1.236724	1	1	0	0	1	0	0	0	[illegible]
...	...	...	...	...	...	...	...	...	...	...
7038	-0.340876	0	0	1	0	0	0	1	0	[illegible]
7039	1.613701	0	0	1	0	1	0	0	0	[illegible]
7040	-0.870241	0	1	0	0	0	0	1	0	[illegible]
7041	-1.155283	1	1	0	0	1	0	0	0	[illegible]
7042	1.369379	[illegible]	0	0	1	0	0	1	1	[illegible]

7043 rows × 14 columns

Over-sampling using SMOTE

```
X = df_model.drop('Churn', axis = 1)
y = df_model['Churn']
```

```
os = SMOTE(random_state=0)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=0)
columns = X_train.columns
```

```
os_data_X,os_data_y = os.fit_resample(X_train, y_train)
os_data_X = pd.DataFrame(data=os_data_X,columns=columns )
os_data_y= pd.DataFrame(data=os_data_y,columns=['Churn'])

os_data_X_test,os_data_y_test = os.fit_resample(X_test, y_test)
```

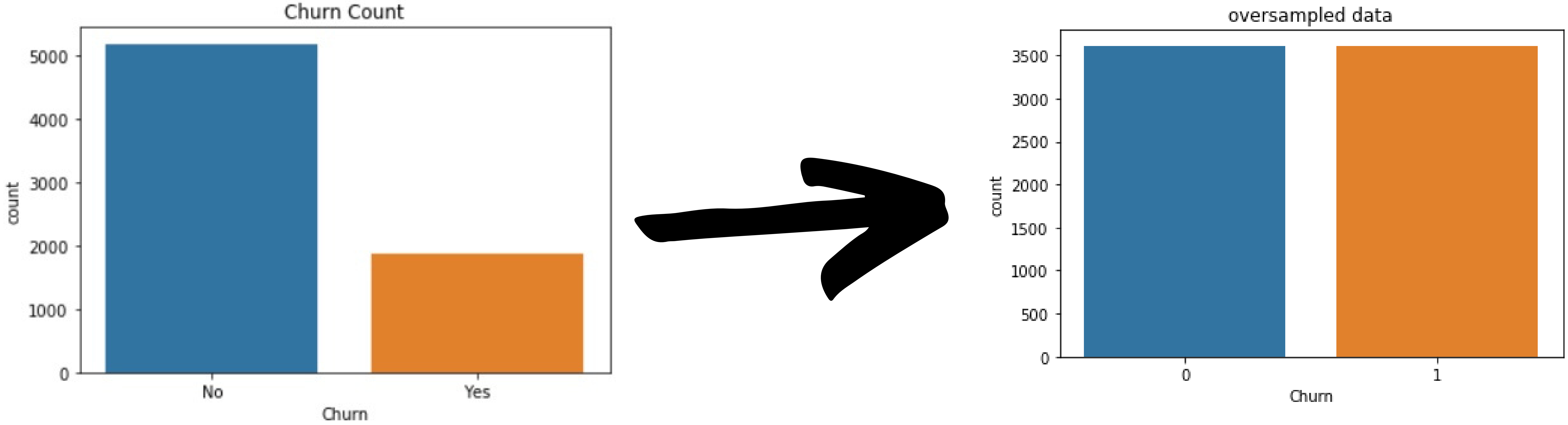
```
print("length of oversampled data is ",len(os_data_X))
print("Number of not churn in oversampled data",len(os_data_y[os_data_y['Churn']==0]))
print("Number of churn customer in oversampled data",len(os_data_y[os_data_y['Churn']==1]))
print("Proportion of churn data in oversampled data is ",len(os_data_y[os_data_y['Churn']==1])/len(os_data_X))
```

```
length of oversampled data is  7228
Number of not churn in oversampled data 3614
Number of churn customer in oversampled data 3614
Proportion of churn data in oversampled data is  0.5
```

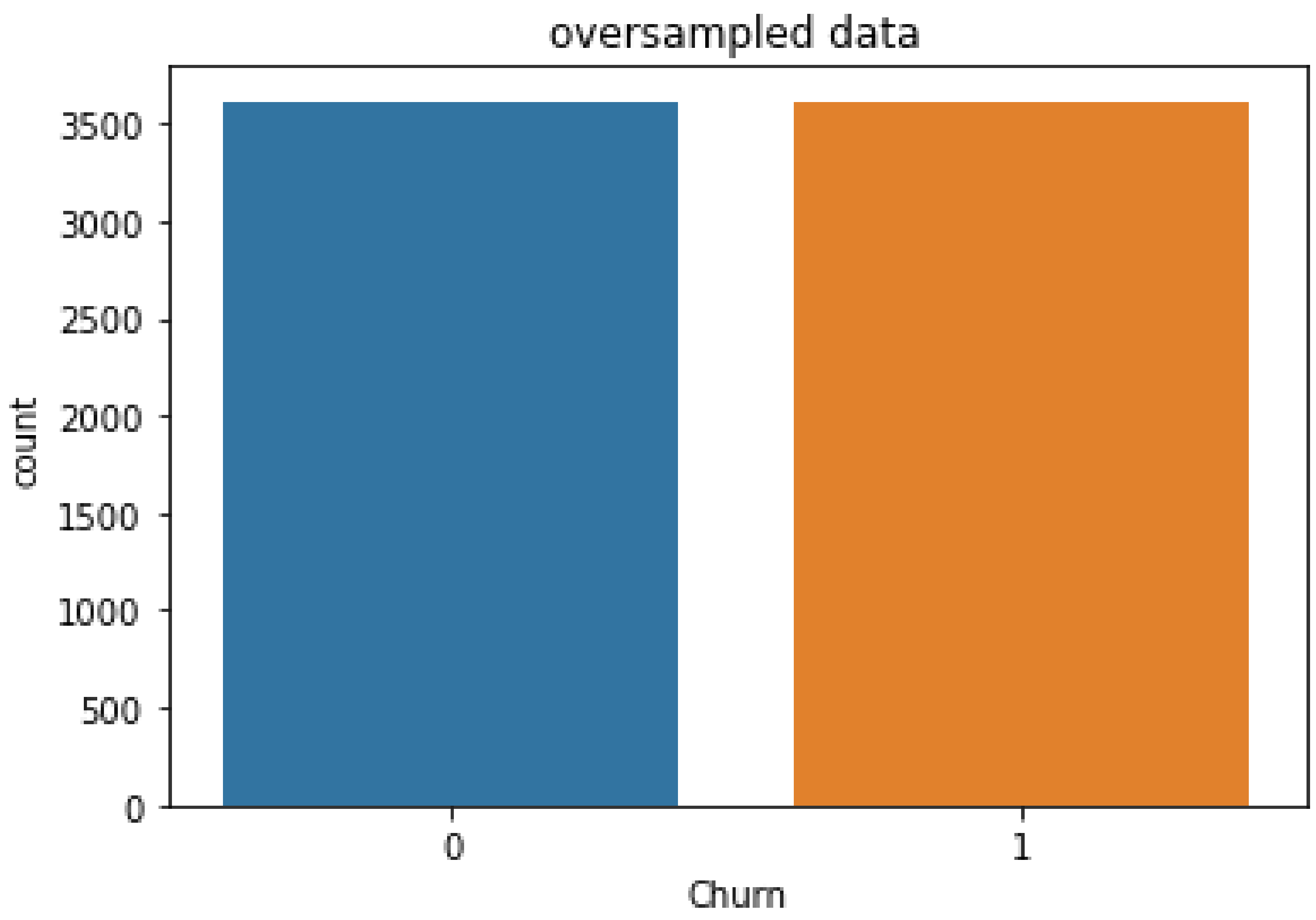
```
sns.countplot(data=os_data_y, x='Churn').set(title='oversampled data')
```

```
[Text(0.5, 1.0, 'oversampled data')]
```

Over-sampling using SMOTE



Over-sampling using SMOTE



length of oversampled data is 7228

Number of not churn customers in oversampled data 3614

Number of churn customers in oversampled data 3614

Proportion of churn data in oversampled data is 0.5

Implementing the model

Logistics Regression

Summarize model

```
Optimization terminated successfully.
         Current function value: 0.466417
         Iterations 7
                                 Results: Logit
=================================================================================
Model:                    Logit                    Pseudo R-squared:      0.327
Dependent Variable:       Churn                    AIC:                   6768.5246
Date:                     2022-05-25 18:04         BIC:                   6858.0390
No. Observations:         7228                     Log-Likelihood:        -3371.3
Df Model:                 12                       LL-Null:               -5010.1
Df Residuals:             7215                     LLR p-value:           0.0000
Converged:                1.0000                   Scale:                 1.0000
No. Iterations:           7.0000
---------------------------------------------------------------------------------
                                         Coef.  Std.Err.    z     P>|z|   [0.025  0.975]
---------------------------------------------------------------------------------
tenure                                  -0.7314   0.0461 -15.8625 0.0000 -0.8218 -0.6411
Contract_Month-to-month                 -3.2033   0.6286  -5.0961 0.0000 -4.4353 -1.9713
Contract_One year                       -4.1059   0.6287  -6.5309 0.0000 -5.3381 -2.8737
Contract_Two year                       -5.1351   0.6393  -8.0329 0.0000 -6.3880 -3.8822
OnlineSecurity_No                       14.0853   1.3936  10.1072 0.0000 11.3539 16.8167
OnlineSecurity_No internet service       6.9197   0.9034   7.6596 0.0000  5.1491  8.6903
OnlineSecurity_Yes                      13.5705   1.3918   9.7506 0.0000 10.8427 16.2983
PaymentMethod_Bank transfer (automatic) -5.1169   0.6499  -7.8735 0.0000 -6.3906 -3.8431
PaymentMethod_Credit card (automatic)   -5.1969   0.6500  -7.9954 0.0000 -6.4709 -3.9230
PaymentMethod_Electronic check          -4.6237   0.6496  -7.1180 0.0000 -5.8969 -3.3506
PaymentMethod_Mailed check              -5.1486   0.6507  -7.9119 0.0000 -6.4240 -3.8732
InternetService_DSL                     -6.1781   1.0478  -5.8962 0.0000 -8.2318 -4.1244
InternetService_Fiber optic             -5.0063   1.0472  -4.7807 0.0000 -7.0587 -2.9538
=================================================================================
```

Evaluatation model

```
clf = LogisticRegression(random_state=0)
clf.fit(os_data_X, os_data_y.values.ravel())

y_pred = clf.predict(os_data_X_test)
```

```
y_pred
```

```
array([0, 0, 1, ..., 1, 1, 1])
```

```
from sklearn.metrics import confusion_matrix
from sklearn.metrics import classification_report
```

```
cm = confusion_matrix(os_data_y_test, y_pred)
print(cm)
```

```
[[1149  411]
 [ 271 1289]]
```

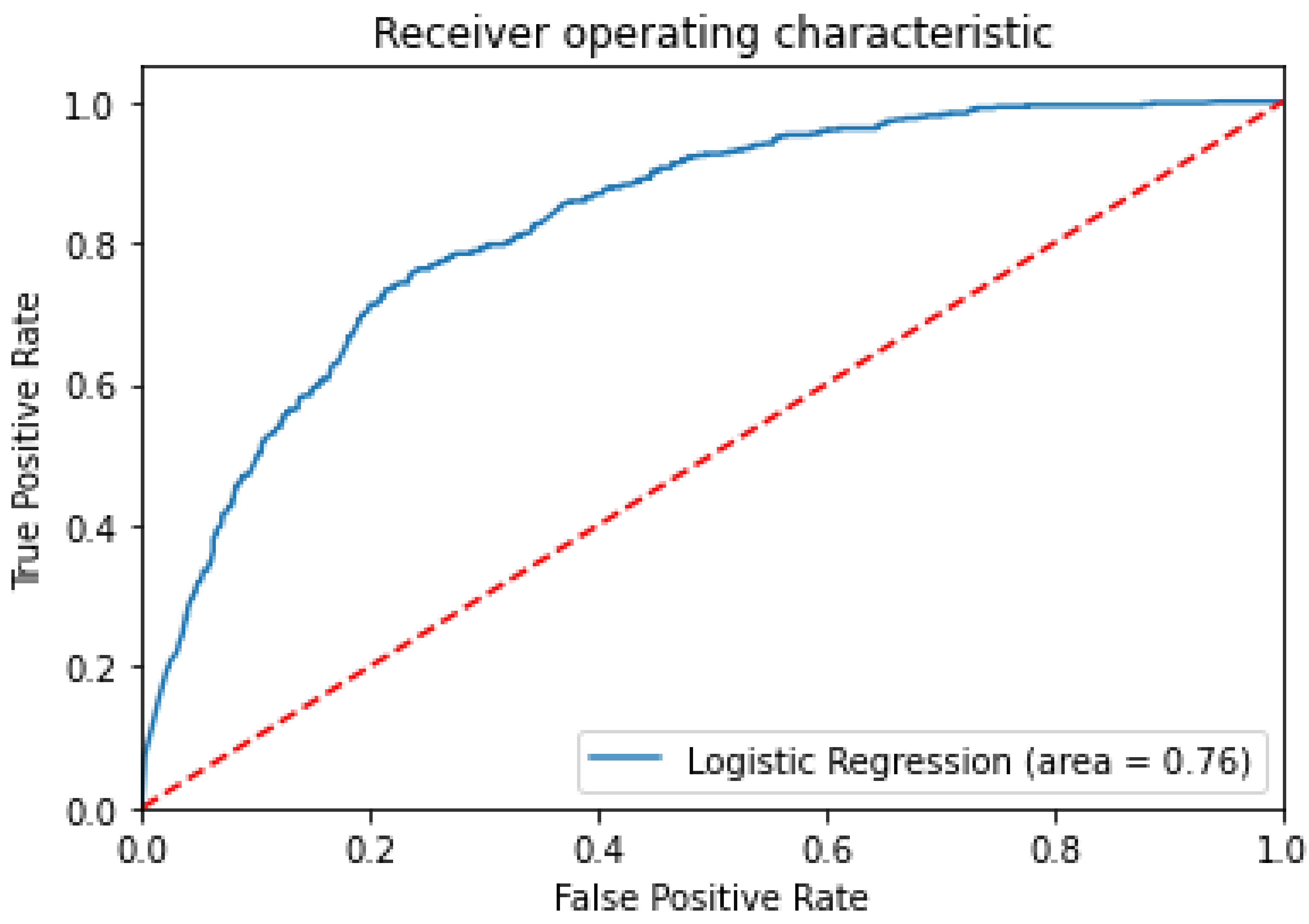
```
print('Accuracy = ', accuracy_score(os_data_y_test,y_pred))
print('F1-Score = ', f1_score(os_data_y_test,y_pred))
print('Precision = ', precision_score(os_data_y_test,y_pred))
print('Recall = ', recall_score(os_data_y_test,y_pred))
```

```
Accuracy =  0.7814102564102564
F1-Score =  0.7907975460122699
Precision =  0.758235294117647
Recall =  0.8262820512820512
```

TN	1149	411	FP
FN	271	1289	TP

Accuracy = 0.7814102564102564
F1-Score = 0.7907975460122699
Precision = 0.758235294117647
Recall = 0.8262820512820512

ROC Curve



AUC of Logistic Regression is equal to 0.76

In the summary, our model has good test quality with AUC values and good metric score. At least we can use this model to predict the churn customer.

Conclusions

From EDA we found that customer who doesn't use internet service is likely to not churn

From EDA we found that as long as the customer sticks to the company product, the less likely the customer will churn.

After we evaluate the Logistic model, we got the good result from all metric score
by using the following columns tenure, contract, online security, internet service and payment method

More information

IMPORT DATA

```
df = pd.read_csv('Telco-Customer-Churn.csv')
df
```

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	...	DeviceProtection	TechSupport	Streami
0	7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No	...	No	No	
1	5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	...	Yes	No	
2	3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	...	No	No	
3	7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes	...	Yes	Yes	
4	9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No	...	No	No	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
7038	6840-RESVB	Male	0	Yes	Yes	24	Yes	Yes	DSL	Yes	...	Yes	Yes	
7039	2234-XADUH	Female	0	Yes	Yes	72	Yes	Yes	Fiber optic	No	...	Yes	No	
7040	4801-JZAZL	Female	0	Yes	Yes	11	No	No phone service	DSL	Yes	...	No	No	
7041	8361-LTMKD	Male	1	Yes	No	4	Yes	Yes	Fiber optic	No	...	No	No	
7042	3186-AJIEK	Male	0	No	No	66	Yes	No	Fiber optic	Yes	...	Yes	Yes	

7043 rows × 21 columns

https://github.com/Nas-virat/Telco-Customer-Churn

Thank You