Data Mining Project

Telecom Customer Churn Prediction

Anastasia Karuzina 85536
Jiyoung Kim 110075

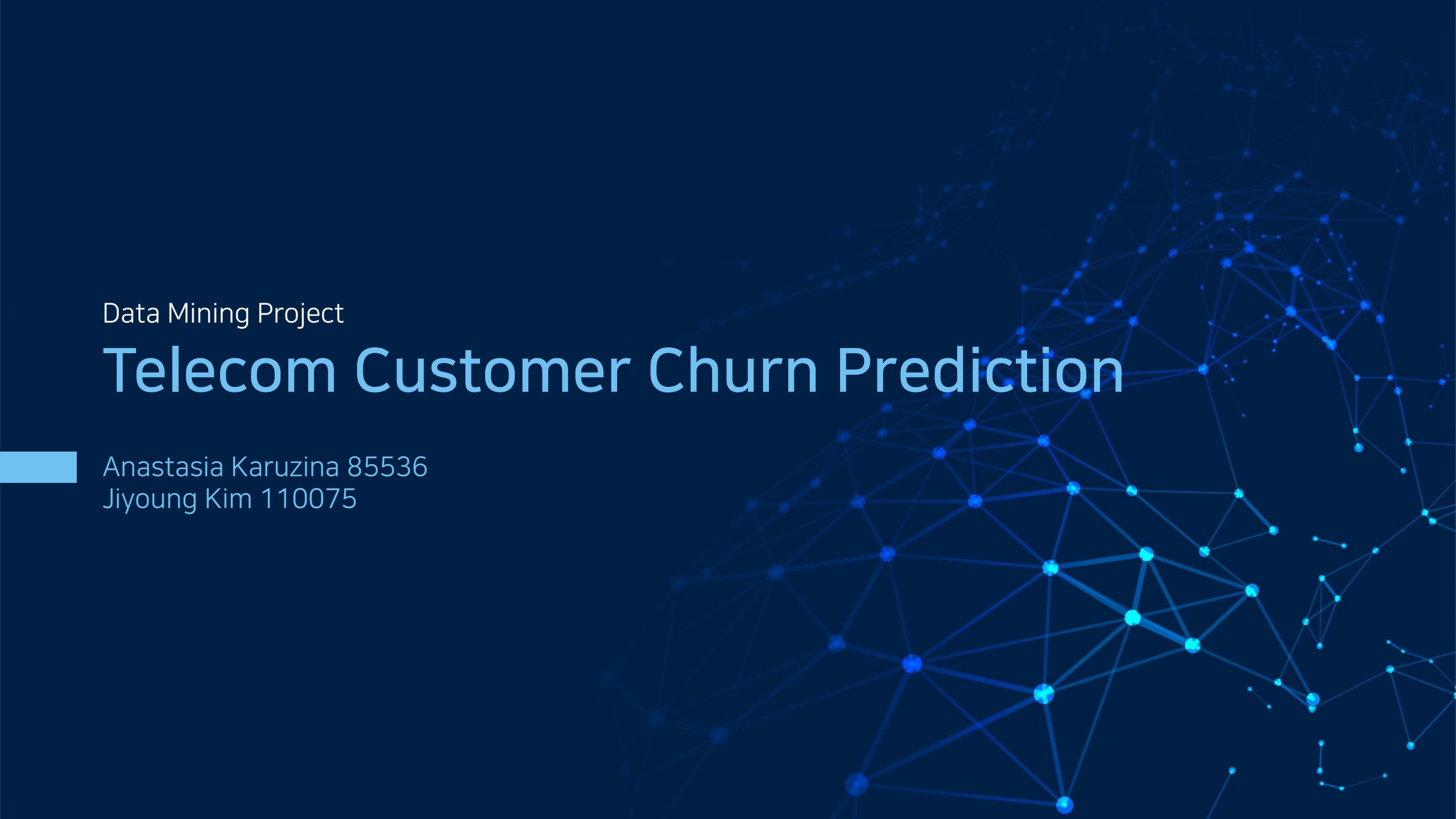


Table of Contents

Introduction

About Dataset

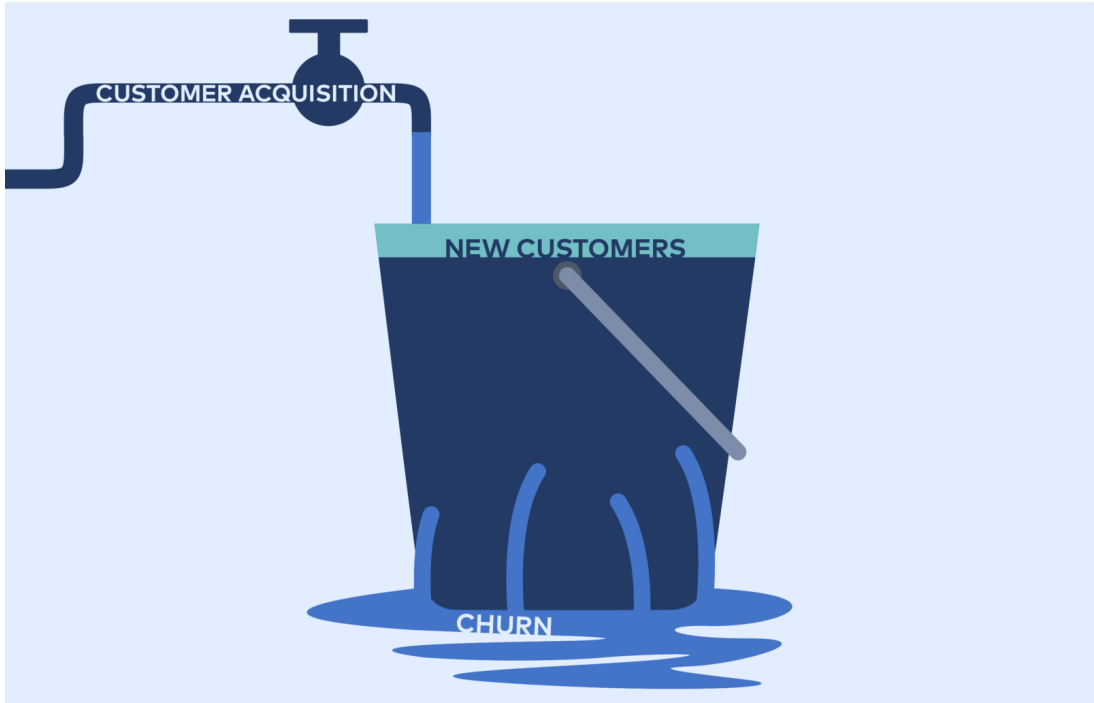
```
RangeIndex: 7043 entries, 0 to 7042
Data columns (total 21 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   customerID        7043 non-null   object
 1   gender            7043 non-null   object
 2   SeniorCitizen     7043 non-null   int64
 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      7032 non-null   float64
 20  Churn             7043 non-null   object
dtypes: float64(2), int64(2), object(17)
```

- Dataset from a telecommunications company - TELCO
- The data contains information about 7043 users , their demographic characteristics, the services they use, the duration of using the operator's se rvices, the method of payment, and the amount of payment.
- Any business wants to maximize the number of customers. The dataset includes churn of custo mers, meaning the company knows characterist ics of those users who left and those who are us ing Telco services.
- The company wants to know which customers will leave and which will stay
- Strategy is to investigate reasons of churn and i mplement marketing campaign that will increas e churn rate and improve company's revenues.

Project Purpose & Goal

Goal:

Identify the background reasons, characteristic, pattern of customers who are highly / less likely to Churn using Telecom company data



Our tasks:

- To derive useful insights from EDA
- Build models that predict Churn
- Assess, Compare those models and explain why one of them is better than the other
- Explain the reasons behind why customers stop using Telecom services (churn)

Hypothesis

Hypothesis 1: Customers who pay for the service more money (MonthlyCharges) than on average ha s higher chance to stop using the service.

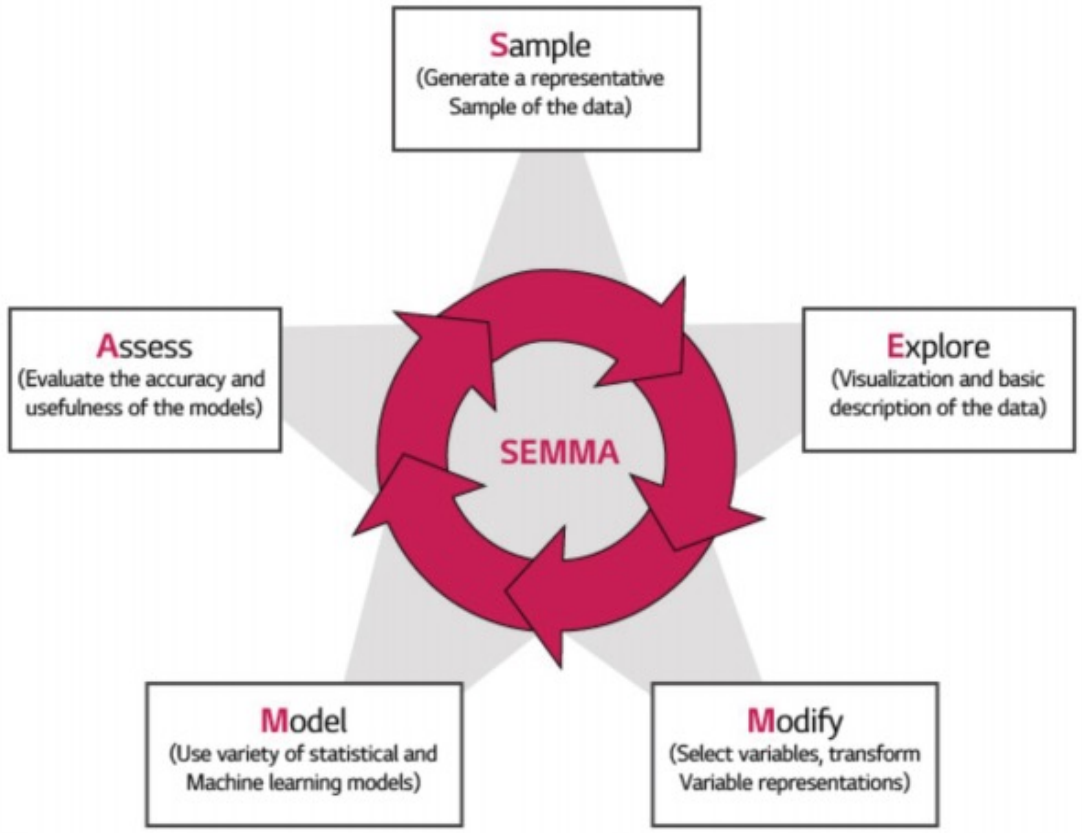
Hypothesis 2: Annual term contract customer has less chance to leave than monthly term contract c ustomer.

Hypothesis 3: If a customer uses additional services of Telecom, she will not stop using the service.

Hypothesis 4: Internet service can affect the churn rate.

SEMMA Methodology

- **Sample:** Split Train / Test set: 70:30
- **Explore:**
  - Correlation - Spearman & Cramer's V
  - Variable Selection - Univariate Statistical Tests (χ^2 Test of Independence) / RFE (Recursive Feature Elimination)
  - Variable Clustering: VarClusHi (Select Best variables based on $(1 - R^2)$ ratio)
- **Model:** Logistic Regression, Decision Tree, XGBoost
- **Assess:** Metrics, ROC Curve



Data Preparation (Data Preprocessing)

- Drop unnecessary variable: ‘customerID’
- Check Data type: changed ‘SeniorCitizen’

  from int64 → object (factor variable with 1/0)
- Check Missing value: Drop 11 missing in ‘TotalCharges’

  (since it’s only 0.15% out of whole dataset)

```
# missing value check
telco.isnull().sum() # 11 missing values in TotalCharges
```

```
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
```

```
gender              object
SeniorCitizen       object
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       float64
Churn               object
dtype: object
```

Conversion of Categorical Variables (non-numeric)

- Grouped level of variables & Created Dummy Columns

  1) Variables with 2 level Yes / No → 1/0

  2) Variables with 3 levels Yes / No / No Internet Service

     → 1/0 (In here, we regarded No Internet Service as No)

  3) Variables with more than 3 levels (which are not yes/no)

     → create dummy columns

```
telco.gender = [1 if each == "Male" else 0 for each in telco.gender]
telco.SeniorCitizen = [1 if each == "1" else 0 for each in telco.SeniorCitizen]

columns_to_change1 = ['OnlineSecurity', 'OnlineBackup', 'DeviceProtection',
                      'TechSupport', 'StreamingTV', 'StreamingMovies']

for column in columns_to_change1:
   telco[column] = telco[column].replace({'No internet service':'No'})

columns_to_change2 = ['Partner','Dependents','PhoneService','OnlineSecurity',
                      'OnlineBackup','DeviceProtection','TechSupport',
                      'StreamingTV','StreamingMovies','PaperlessBilling','Churn']

for column in columns_to_change2:
    telco[column] = telco[column].map({'Yes':1, 'No':0})

dummy_col = ['MultipleLines', 'InternetService', 'Contract', 'PaymentMethod']
telco_new = pd.get_dummies(data=telco, columns = dummy_col)
```

```
RangeIndex: 7032 entries, 0 to 7031
Data columns (total 29 columns):
 #   Column                                  Non-Null Count  Dtype
---  ------                                  --------------  -----
 0   gender                                  7032 non-null   int64
 1   SeniorCitizen                           7032 non-null   int64
 2   Partner                                 7032 non-null   int64
 3   Dependents                              7032 non-null   int64
 4   tenure                                  7032 non-null   int64
 5   PhoneService                            7032 non-null   int64
 6   OnlineSecurity                          7032 non-null   int64
 7   OnlineBackup                            7032 non-null   int64
 8   DeviceProtection                        7032 non-null   int64
 9   TechSupport                             7032 non-null   int64
 10  StreamingTV                             7032 non-null   int64
 11  StreamingMovies                         7032 non-null   int64
 12  PaperlessBilling                        7032 non-null   int64
 13  MonthlyCharges                          7032 non-null   float64
 14  TotalCharges                            7032 non-null   float64
 15  Churn                                   7032 non-null   int64
 16  MultipleLines_No                        7032 non-null   uint8
 17  MultipleLines_No phone service          7032 non-null   uint8
 18  MultipleLines_Yes                       7032 non-null   uint8
 19  InternetService_DSL                     7032 non-null   uint8
 20  InternetService_Fiber optic             7032 non-null   uint8
 21  InternetService_No                      7032 non-null   uint8
 22  Contract_Month-to-month                 7032 non-null   uint8
 23  Contract_One year                       7032 non-null   uint8
 24  Contract_Two year                       7032 non-null   uint8
 25  PaymentMethod_Bank transfer (automatic) 7032 non-null   uint8
 26  PaymentMethod_Credit card (automatic)   7032 non-null   uint8
 27  PaymentMethod_Electronic check          7032 non-null   uint8
 28  PaymentMethod_Mailed check              7032 non-null   uint8
dtypes: float64(2), int64(14), uint8(13)
```

Step 1. SAMPLE

Data Partitioning

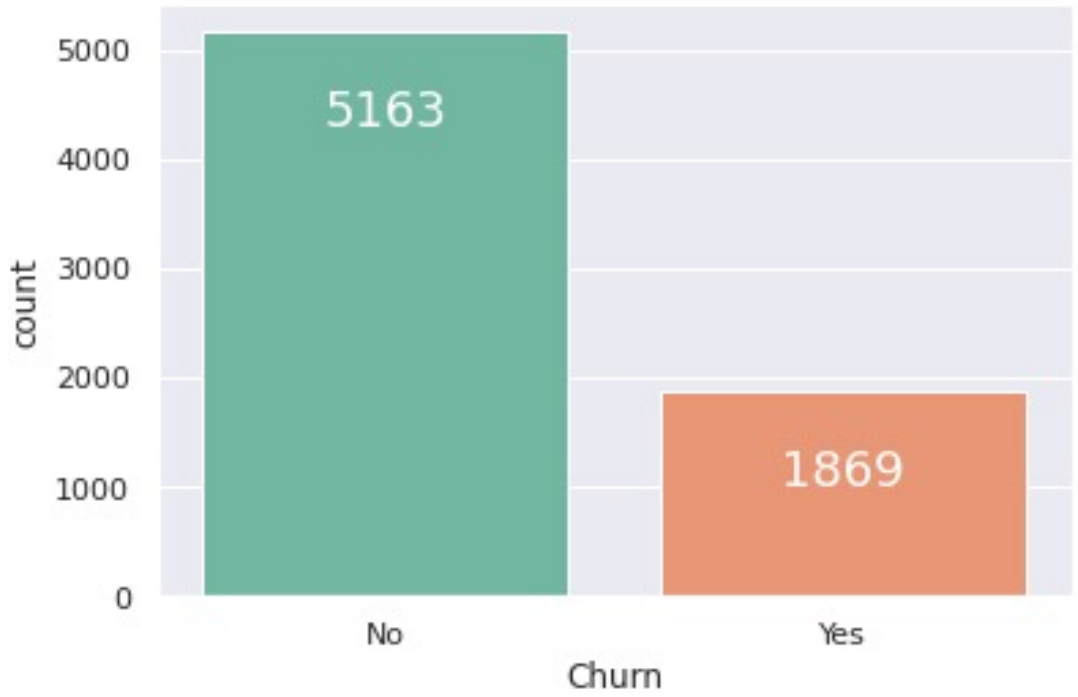
```
[20] # 1) Data Partitioning (70/30)

     from sklearn.model_selection import train_test_split

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

     # Train - Test Split
     X_train, X_test, y_train, y_test =
     train_test_split(X, y, test_size=0.3, stratify = y, random_state = 42)

     X_train_shape = X_train.shape
     X_test_shape = X_test.shape
     print("X_train shape = {}\nX_test shape = {}".format(X_train_shape, X_test_shape))
```

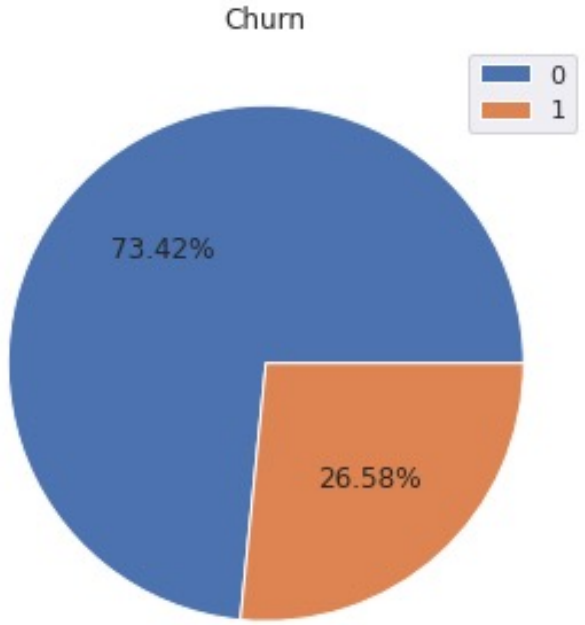
```
X_train shape = (4922, 28)
X_test shape = (2110, 28)
```

- Since our dataset is relatively small, we only split it without validation set
- Used ‘stratify = y’ to keep the same proportion of Y variable (target) in each sets

Step 2. Explore

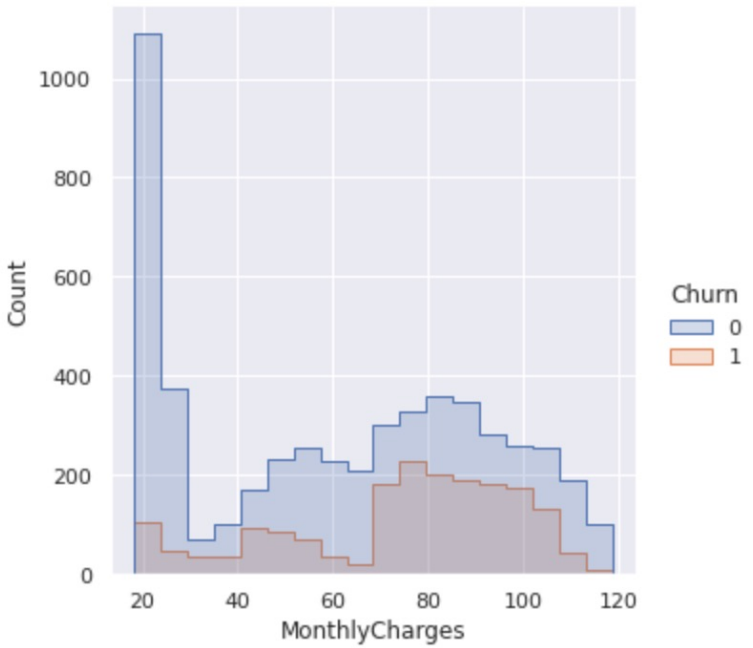
EDA – Target Variable ‘Churn’

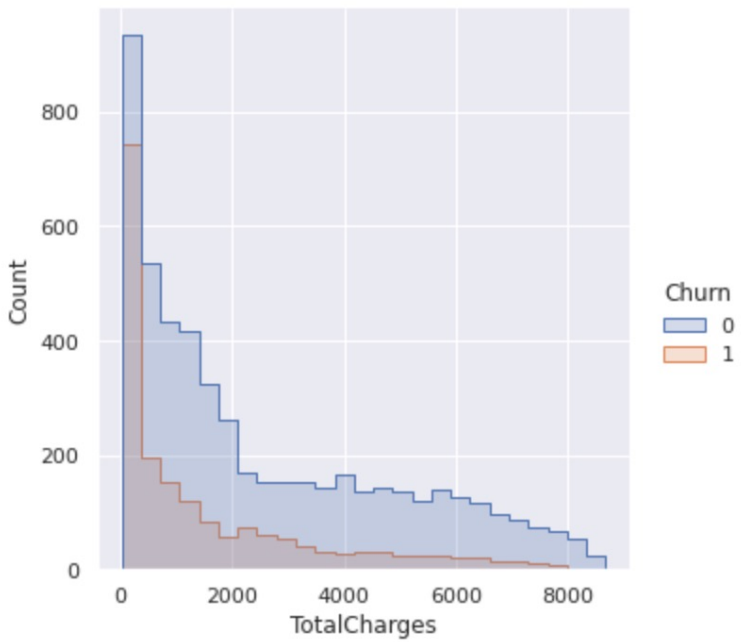


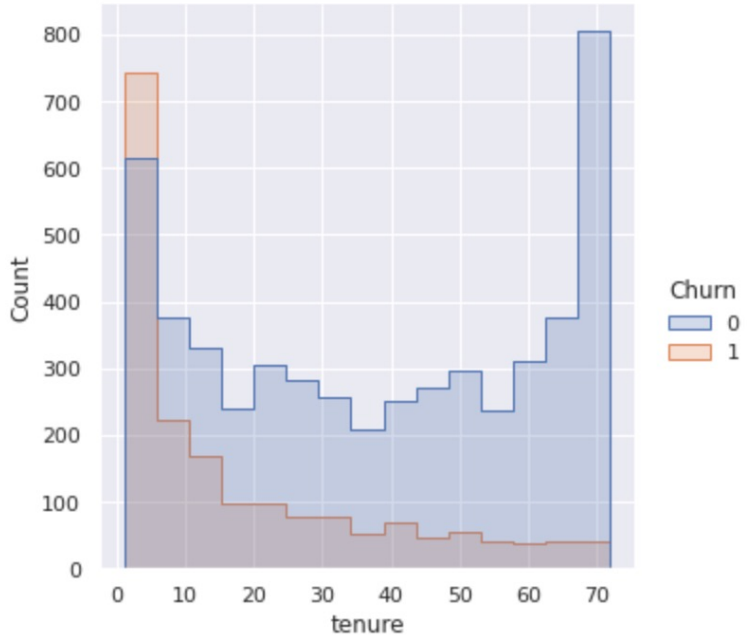


- In general, threshold for 'imbalanced data' is 80:20, so we can still embrace our target variables just the way it is even though it’s a bit imbalanced.

EDA - Numerical

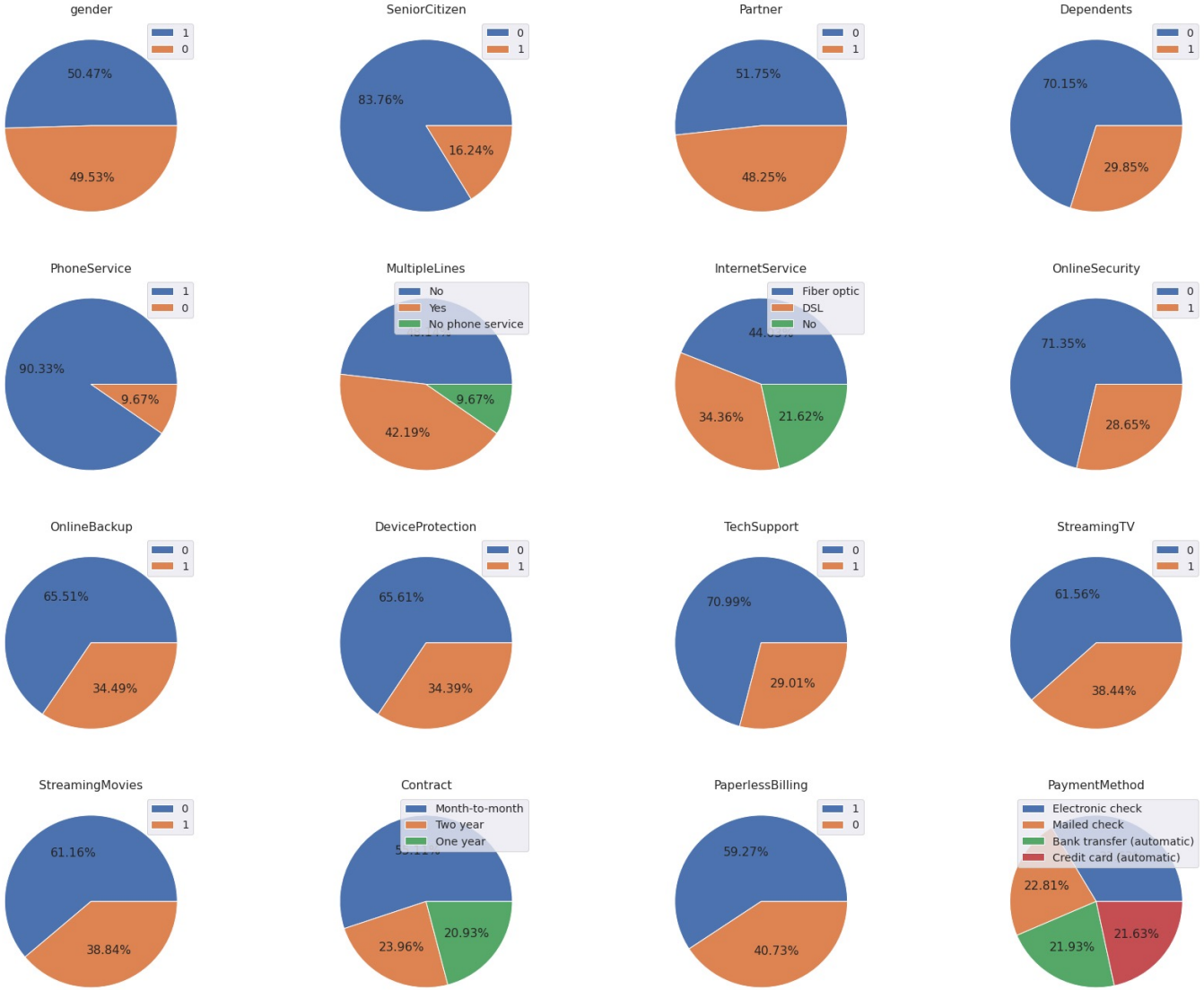




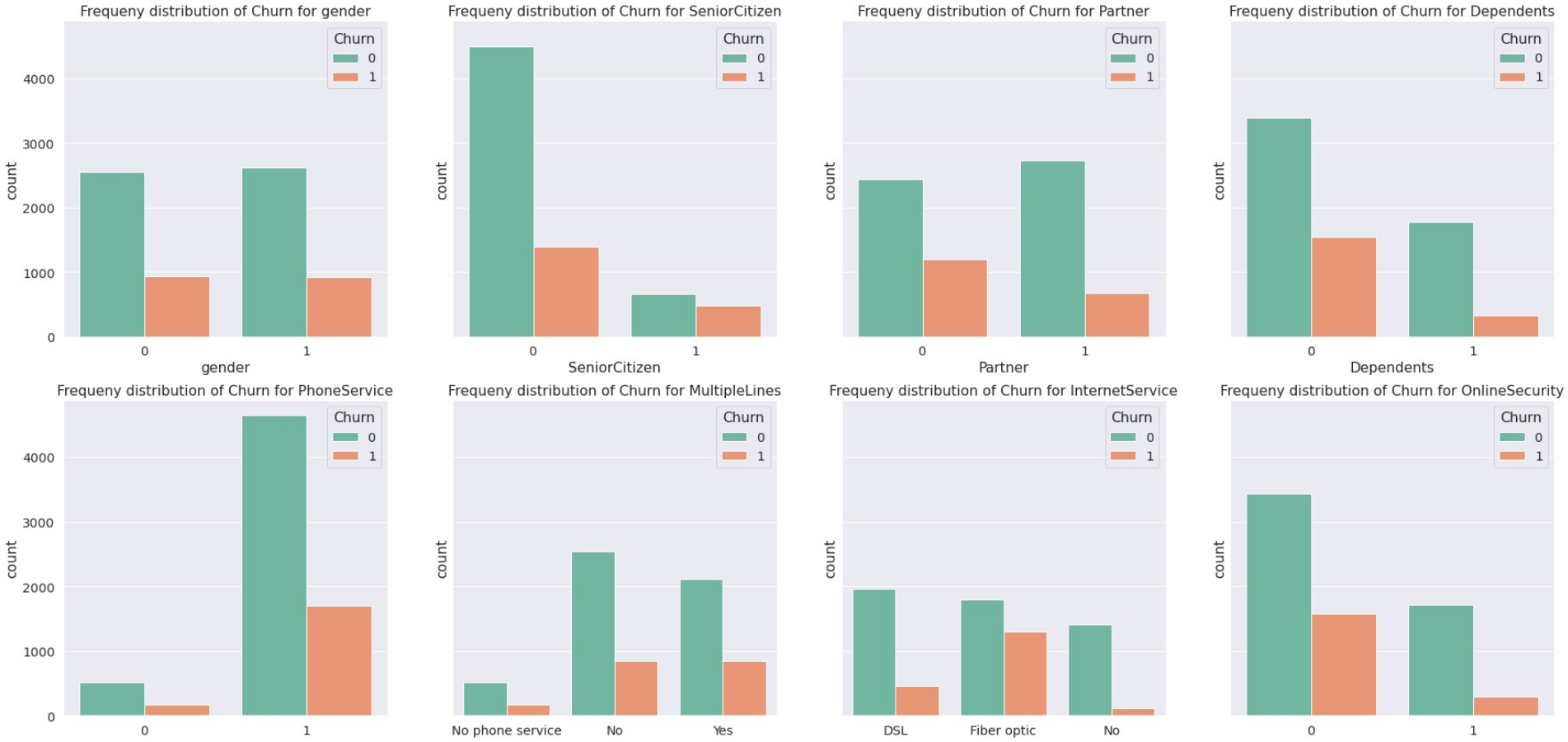


- The people who pays relatively less MonthlyCharges are less likely to churn (will stay)
- The longer the tenure is, the customers are less likely to churn (will stay)

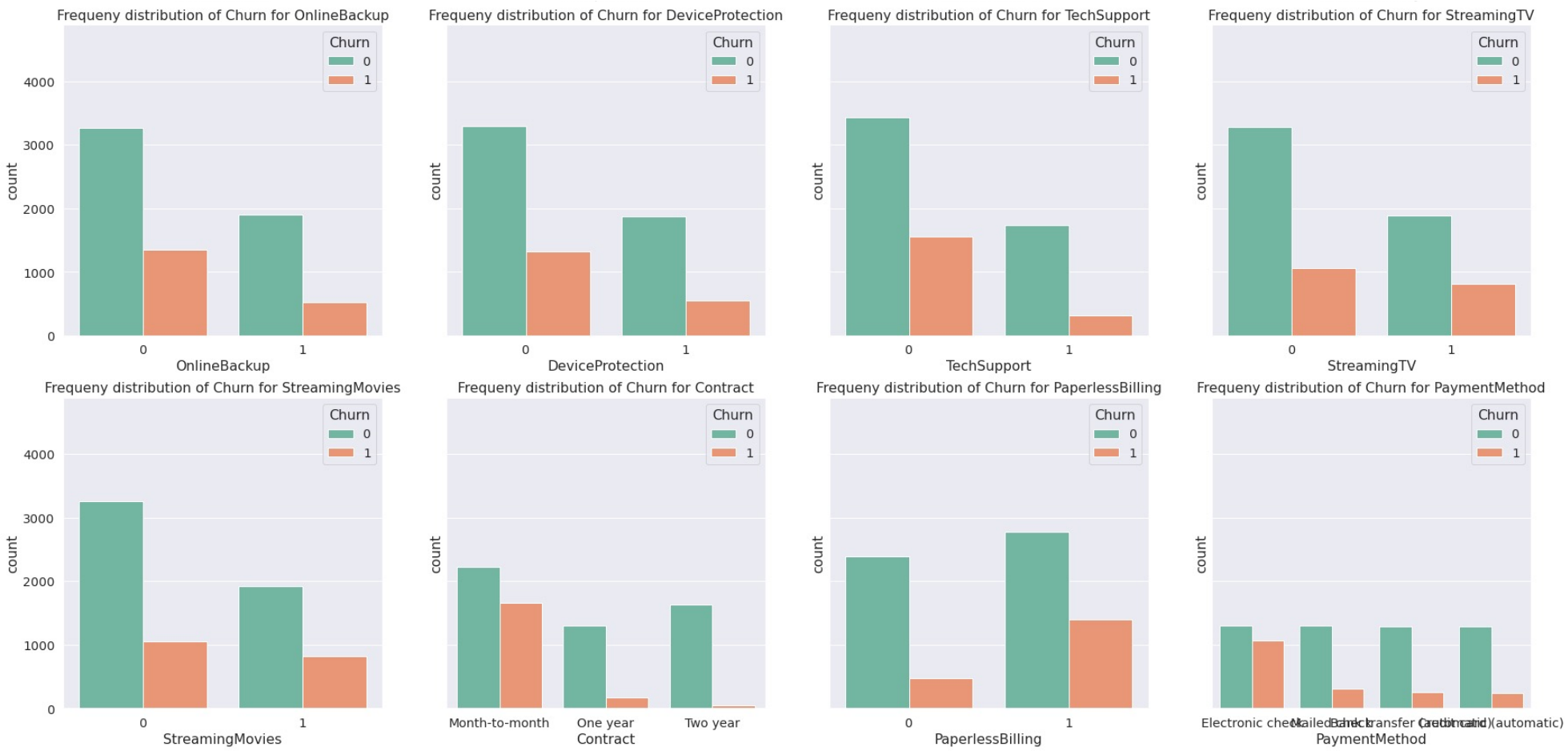
EDA - Categorical



EDA - Categorical (Multiplot in SAS Miner)

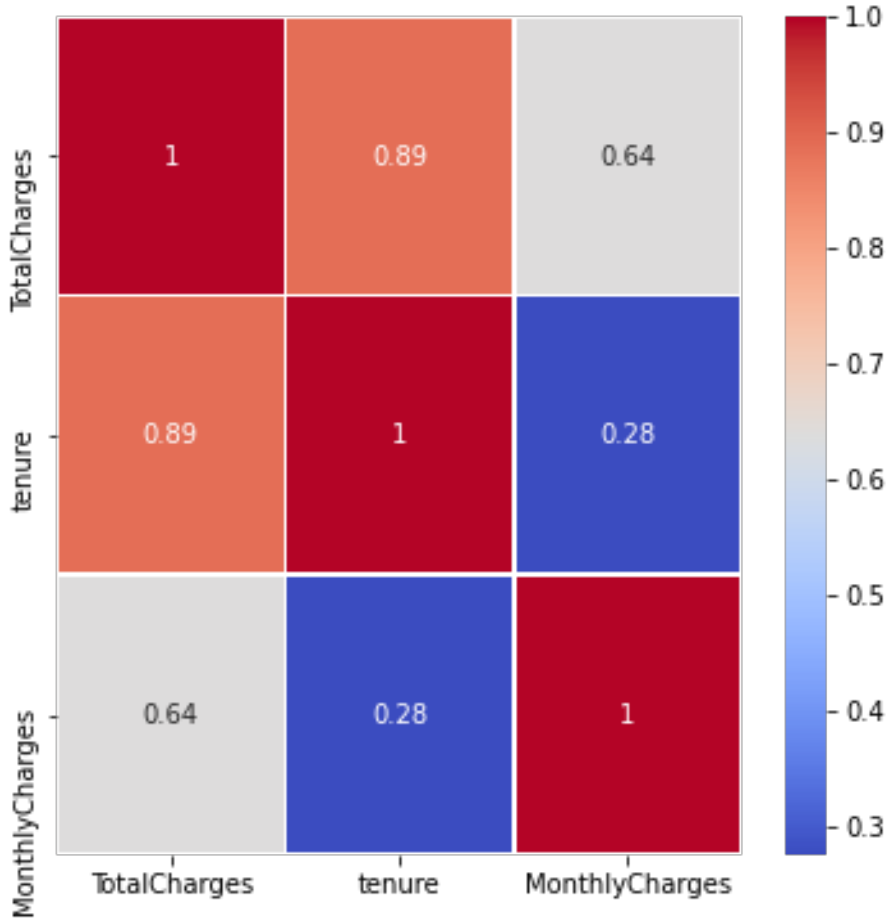


EDA - Categorical (Multiplot in SAS Miner)



Correlation - Numerical (StatExplore in SAS Miner)

Spearman correlation matrix for continuous variables:
Total Charges, Tenure, Monthly Charges

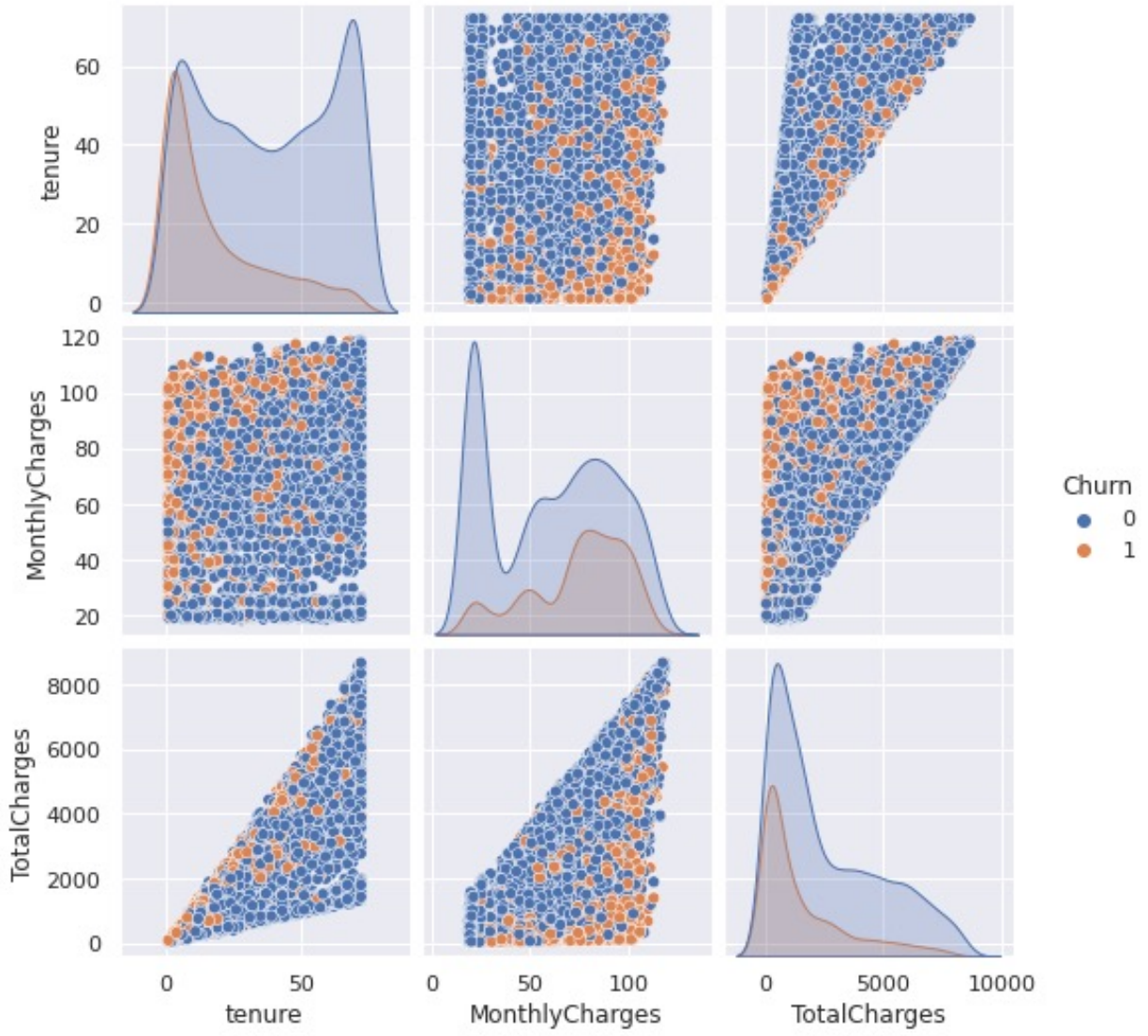


Total Charges and Tenure are highly correlated (0.89)

→ Considering that TotalCharges can be derived from tenure*Monthly Charges, we can just drop the Total Charges.

Correlation - Numerical (StatExplore in SAS Miner)

Spearman correlation matrix for continuous variables: Total Charges, Tenure, Monthly Charges

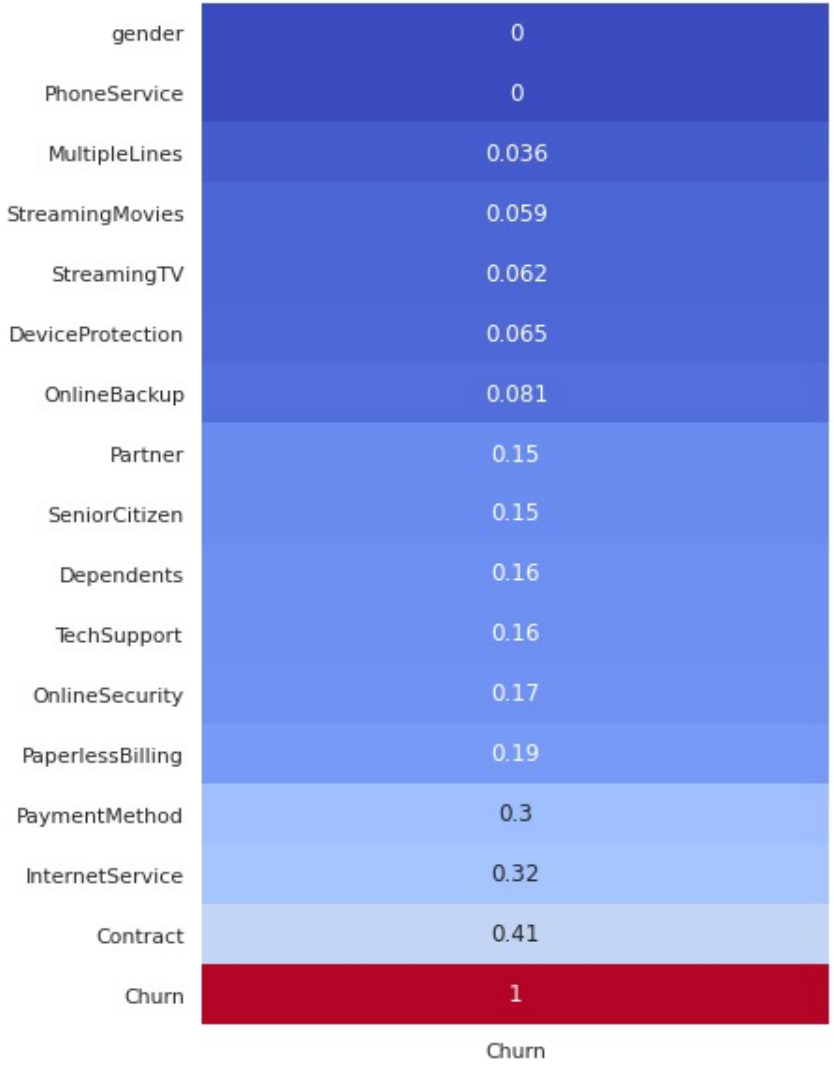


People having lower tenure and higher monthly charges are tend to churn more:

→ We need to closely observe these two variables and compare their distribution with churn.

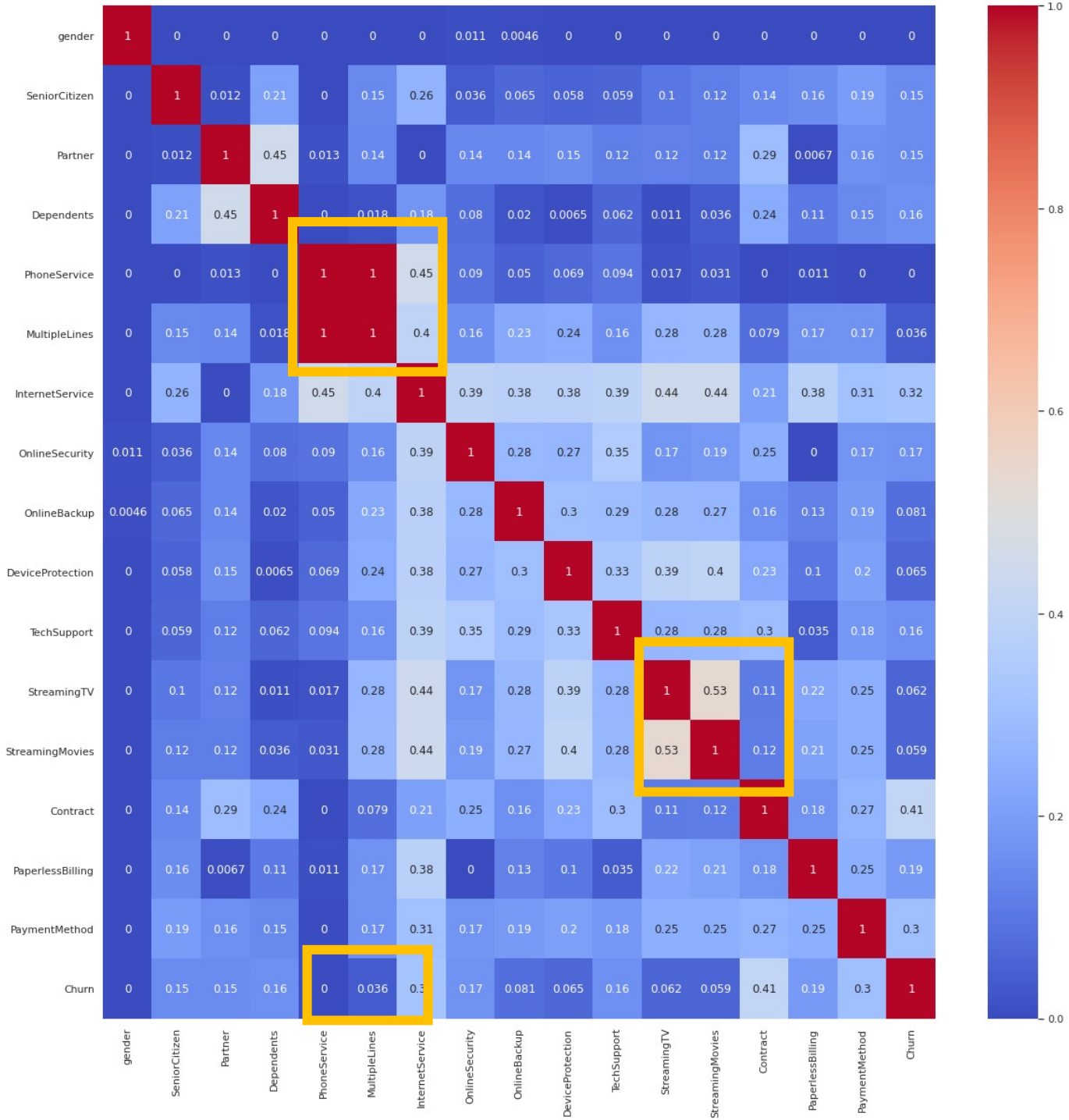
Correlation - Categorical (StatExplore in SAS Miner)

Cramer's V correlation matrix for categorical variables compared with our **target**



Contract is the most correlated with Churn among other categorical variables (correlation = 0.41)

→ We should take a closer investigation of Contract variable in our research and see if we can get insights from it.



Streaming TV, Streaming Movies services are quite correlated between each other (0.53)

→ Considering the EDA results, the distribution of values (1/0 and also churn distribution) of Streaming TV and Movies were almost same.

So, we decided to assemble those two and make it into one variable: **'Streaming Services'.**

PhoneService and **Multiple lines** are very correlated between each other (correlation = 1)

→ We will drop **PhoneService** (because it's less correlated with our target variable Churn than the Multiple lines).

Insights from Correlation

1) **Total Charges** and **Tenure** are highly correlated (0.89).
   → Considering that **TotalCharges** can be derived from **tenure*MonthlyCharges**, we can just drop the Total Charges.

2) **StreamingTV**, **StreamingMovies** services are quite correlated between each other (0.53)
   → Considering the EDA results, the distribution of values (1/0 and also churn distribution) of Stream ing TV and Movies were almost same. Therefore, we decided to assemble those two and make it in to one variable '**StreamingServices**'.

3) **PhoneService** and **MultipleLines** are very correlated between each other (correlation = 1)
   → We will drop **PhoneService** (Bcz it's less correlated with our varget Variable **Churn** than the Mul tiple lines).

4) **Contract** is 41% correlated with **Churn**. We should take it into consideration and investigate in later steps.

Variables Selection #1 – Univariate Statistical Tests

- Variable Selection is required to filter those variables which are not strongly associated with the response variable (Churn)

- Used 'Chi-squared(χ^2) Test of Independence':

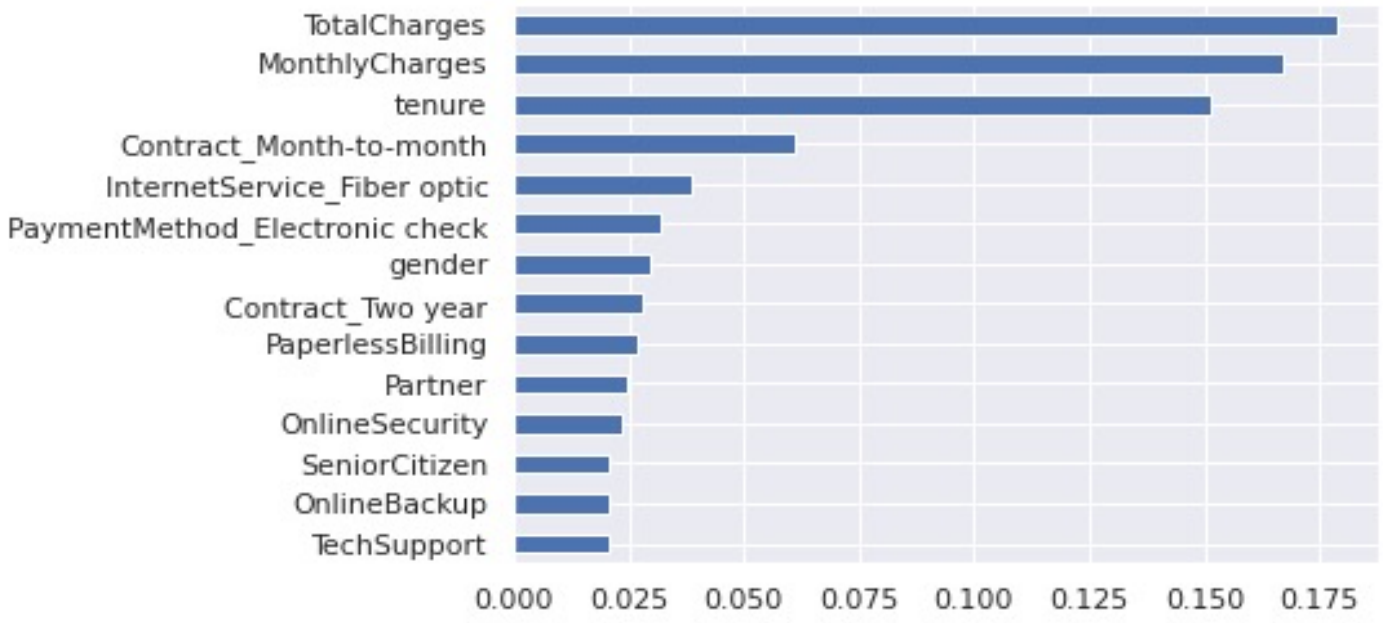
  $\boldsymbol{H_0}$ *: [Churn] is independent of [Variable]*

  $\boldsymbol{H_1}$ *: [Churn] is not independent of [Variable]*

- Selected best 14 variables (around half of current # of variables) based on low p-value

Variable	Score	p_value
TotalCharges	414772.526053529	0.0
tenure	11205.746655938865	0.0
MonthlyCharges	2769.3479131696445	0.0
Contract_Month-to-month	367.63954613108984	6.111424711641595e-82
Contract_Two year	341.1605521468674	3.565531205207441 5e-76
PaymentMethod_Electronic check	302.580581645805	9.027290802174378e-68
InternetService_Fiber optic	262.32426544131465	5.344293230167541e-59
InternetService_No	213.2737571639247	2.6524696751110032e-48
Contract_One year	125.02151596808329	5.0345834004237956e-29
Dependents	103.46115189395368	2.6555040711655292e-24
OnlineSecurity	102.22127279872525	4.9653824638368354e-24
SeniorCitizen	95.98093426282263	1.159951759409824e-22
TechSupport	93.21230688374366	4.697510774440431e-22
PaperlessBilling	74.9244837297093	4.8906725991694385e-18
PaymentMethod_Credit card (automatic)	69.8548458466187	6.383308928287899e-17
PaymentMethod_Bank transfer (automatic)	63.692020914461885	1.4547260734863301e-15
Partner	56.20618826509397	6.525582830964217e-14
InternetService_DSL	45.86628319654367	1.2660637411907248e-11
PaymentMethod_Mailed check	27.30222928355083	1.7401067575412844e-07
OnlineBackup	20.02065802614543	7.661004389707829e-06
StreamingTV	16.50335919701574	4.856395749520768e-05
StreamingMovies	14.464096692137682	0.0001428566572981041
DeviceProtection	11.31786332705957	0.0007676499499475058
MultipleLines_Yes	6.897034512815152	0.00863388478501551
MultipleLines_No	4.668931169662576	0.030713035862479574
MultipleLines_No phone service	0.4910751522881234	0.4834480952767507
PhoneService	0.052698053462639276	0.8184332523015435
gender	0.007289900596751923	0.9319585656060292

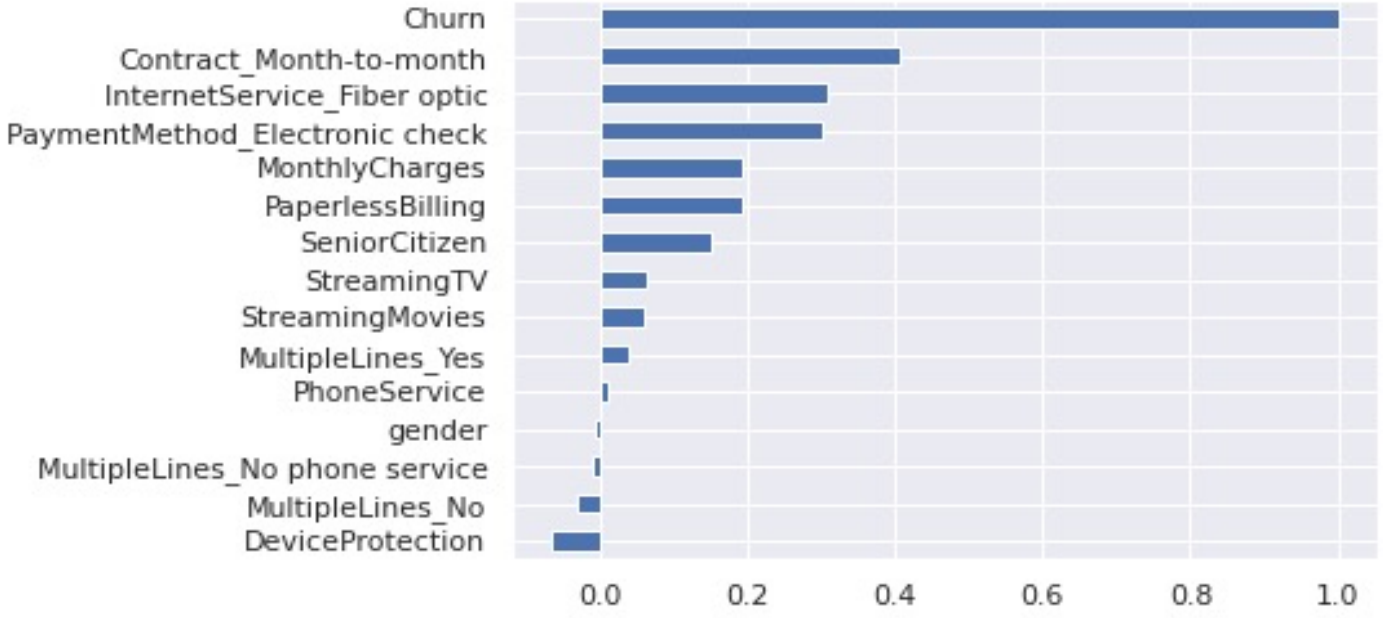
Variables Selection #2 – Feature Importance Method



TotalCharges	0.178761
MonthlyCharges	0.167127
tenure	0.151542
Contract_Month-to-month	0.061039
InternetService_Fiber optic	0.038941
PaymentMethod_Electronic check	0.031939
gender	0.029535
Contract_Two year	0.027853
PaperlessBilling	0.026686
Partner	0.024683
OnlineSecurity	0.023607
SeniorCitizen	0.021011
OnlineBackup	0.020934
TechSupport	0.020659

- Used RandomForestClassifier to check the Feature Importance
- We can see that most of the variables on top importance (except for 3 variables) are similar to variables selected from Univariate Statistical Tests

Variables Selection #3 – Feature Correlation Method



Churn	1.000000
Contract_Month-to-month	0.404565
InternetService_Fiber optic	0.307463
PaymentMethod_Electronic check	0.301455
MonthlyCharges	0.192858
PaperlessBilling	0.191454
SeniorCitizen	0.150541
StreamingTV	0.063254
StreamingMovies	0.060860
MultipleLines_Yes	0.040033
PhoneService	0.011691
gender	-0.008545
MultipleLines_No phone service	-0.011691
MultipleLines_No	-0.032654
DeviceProtection	-0.066193

- Default corr() function shows Correlation plot as above
- Selected 14 most correlated features with ‘Churn’
- Top 6 variables are also included in the variables selected from Univariate Statistical Tests

Feature Scaling - Standardization

```
#Feature Scaling for continuous(num) variables before Clustering

from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
sc.fit(X_train_selected[['tenure','MonthlyCharges']])
X_train_selected[['tenure','MonthlyCharges']] = sc.transform(
    X_train_selected[['tenure','MonthlyCharges']])

X_train_selected[['tenure','MonthlyCharges']].describe()
```

	tenure	MonthlyCharges
count	4.922000e+03	4.922000e+03
mean	1.001050e-16	-8.616522e-18
std	1.000102e+00	1.000102e+00
min	-1.280909e+00	-1.535103e+00
25%	-9.554981e-01	-9.794060e-01
50%	-1.419706e-01	1.885914e-01
75%	9.562914e-01	8.331178e-01
max	1.607113e+00	1.776770e+00

- Numeric variables in our current dataset has different scale
- In case of using Logistic Regression model, it uses ‘Regularization’ which makes the predictor dependent on the scale of the features, Feature Scaling is required
- **Standardization:** the values are centered around the mean with a unit standard deviation. (mean ≒ zero, unit standard deviation)

Variable Clustering

	Cluster	N_Vars	Eigval1	Eigval2	VarProp
0	0	3	2.231318	0.435094	0.743773
1	1	3	2.359791	0.530191	0.786597
2	2	1	1.000000	0.000000	1.000000
3	3	2	1.377281	0.622719	0.688641
4	4	2	1.346628	0.653372	0.673314
5	5	1	1.000000	0.000000	1.000000
6	6	1	1.000000	0.000000	1.000000
7	7	1	1.000000	0.000000	1.000000

	Cluster	Variable	RS_Own	RS_NC	RS_Ratio
2	0	Contract_Two year	0.715169	0.084011	0.310955
0	0	tenure	0.733741	0.149956	0.313230
1	0	Contract_Month-to-month	0.782408	0.315929	0.318083
3	1	MonthlyCharges	0.933246	0.159253	0.079399
4	1	InternetService_Fiber optic	0.725741	0.103221	0.305826
5	1	InternetService_No	0.700804	0.168584	0.359863
6	2	Contract_One year	1.000000	0.035789	0.000000
7	3	PaymentMethod_Credit card (automatic)	0.688641	0.058729	0.330786
8	3	PaymentMethod_Electronic check	0.688641	0.106583	0.348504
9	4	OnlineSecurity	0.673314	0.083676	0.356518
10	4	TechSupport	0.673314	0.101036	0.363403
11	5	Dependents	1.000000	0.050586	0.000000
12	6	SeniorCitizen	1.000000	0.059203	0.000000
13	7	PaperlessBilling	1.000000	0.136039	0.000000

- Used VarClusHi() function to perform variable clustering with keeping a hierarchical structure
- Selected ‘Best Variables’ based on $(1 - R^2)$ ratio $\text{1 - R**2 Ratio} = \frac{\text{1- R}^2 \text{ own cluster}}{\text{1- R}^2 \text{ next cluster}}$
- min $(1 - R^2)$ ratio means when a variable has maximum correlation with own cluster and minimum correlation with next cluster

Step 3. Modify

Modification based on the findings from Step 2

What's modified?

- Correlation result

→ Dropping **PhoneService** variable

→ Creating **StreamingServices variable**

- Variable Selection result

→ Selection of 14 most important features

- Feature Scaling result

→ Preparation of data for Logistic Regression modeling

- Variable Clustering result

→ Best variables selection

- Missing values

→ Dropping 11 missing values from **TotalCharges**

Step 4. Model

Logistic Regression Model

Logistic Regression Model

We will build 3 different Logistic Regression Model:

1) Model with **full range** of variables

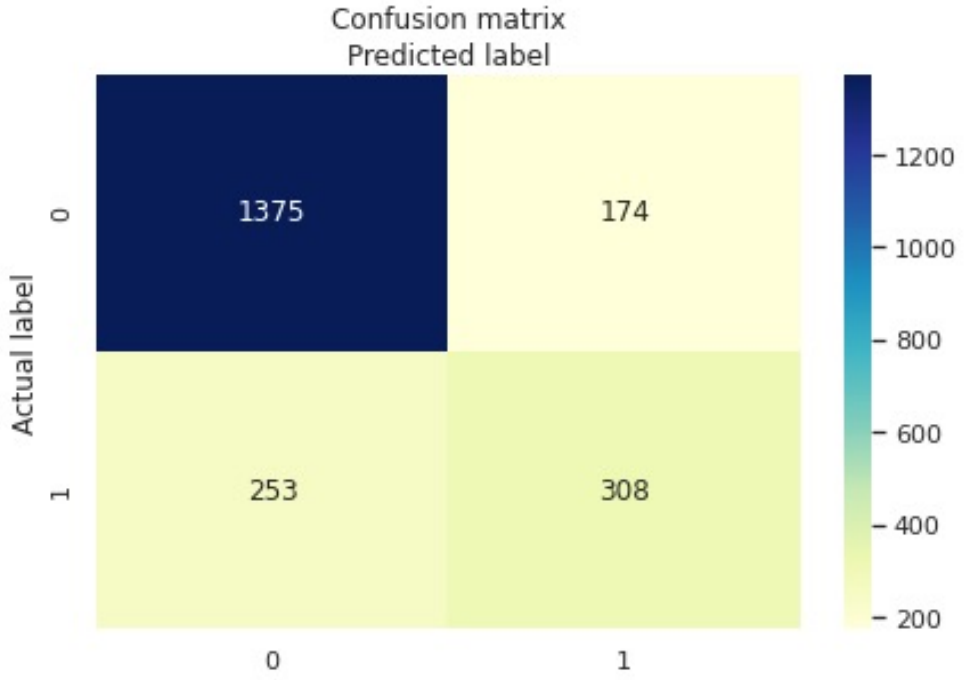
2) Model with Selected Variables

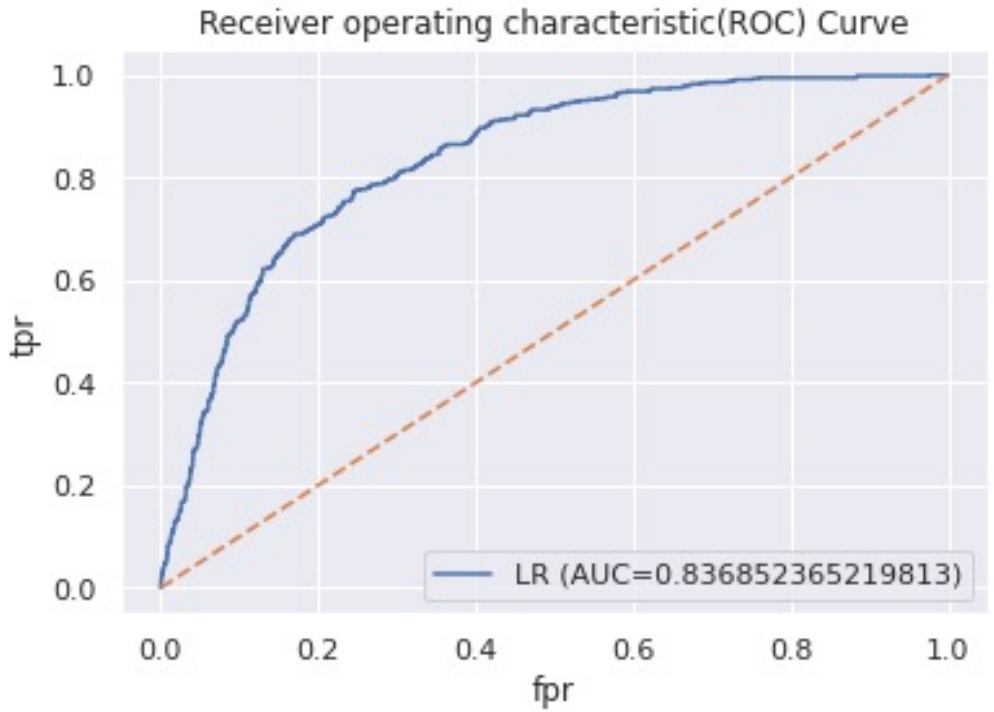
    2-1) Model with Selected variables using **RFE** (Recursive Feature Elimination)

    2-2) Model with Selected variables from Previous Explore Step

        (**Univariate selection with Chi-square test + Variable Clustering**)

LR Model #1 – Full variables





	Model	Test Accuracy	Train Accuracy	Precision	Recall	F1 Score
0	Logistic Regression (Base Model)	0.79763	0.804348	0.54902	0.639004	0.590604

LR Model #1 – Full variables - Interpretation

Generalized Linear Model Regression Results

Dep. Variable:	Churn	**No. Observations:**	4922
Model:	GLM	**Df Residuals:**	4900
Model Family:	Binomial	**Df Model:**	21
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2028.6
Date:	Mon, 31 May 2021	**Deviance:**	4057.2
Time:	01:10:29	**Pearson chi2:**	5.09e+03
No. Iterations:	8		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
gender	-0.0039	0.078	-0.050	0.960	-0.157	0.149
SeniorCitizen	0.1743	0.102	1.709	0.087	-0.026	0.374
Partner	0.0263	0.093	0.282	0.778	-0.156	0.209
Dependents	-0.2781	0.108	-2.585	0.010	-0.489	-0.067
tenure	-0.8185	0.070	-11.699	0.000	-0.956	-0.681
OnlineSecurity	-0.6201	0.114	-5.456	0.000	-0.843	-0.397
OnlineBackup	-0.3460	0.105	-3.309	0.001	-0.551	-0.141
DeviceProtection	-0.1659	0.114	-1.449	0.147	-0.390	0.058
TechSupport	-0.6147	0.117	-5.275	0.000	-0.843	-0.386
PaperlessBilling	0.3311	0.089	3.705	0.000	0.156	0.506
MonthlyCharges	1.2661	0.326	3.881	0.000	0.627	1.905
MultipleLines_No	-0.7703	0.069	-11.187	0.000	-0.905	-0.635
MultipleLines_No phone service	0.5489	0.255	2.152	0.031	0.049	1.049
MultipleLines_Yes	-0.5997	0.084	-7.099	0.000	-0.765	-0.434
InternetService_DSL	-0.2508	0.110	-2.288	0.022	-0.466	-0.036
InternetService_Fiber optic	-0.3936	0.221	-1.783	0.075	-0.826	0.039
InternetService_No	-0.1766	0.354	-0.499	0.618	-0.871	0.518
Contract_Month-to-month	0.3905	0.113	3.460	0.001	0.169	0.612
Contract_One year	-0.3034	0.117	-2.597	0.009	-0.532	-0.074
Contract_Two year	-0.9082	0.157	-5.773	0.000	-1.217	-0.600
PaymentMethod_Bank transfer (automatic)	-0.3399	0.102	-3.325	0.001	-0.540	-0.140
PaymentMethod_Credit card (automatic)	-0.3194	0.103	-3.105	0.002	-0.521	-0.118
PaymentMethod_Electronic check	0.0304	0.084	0.363	0.717	-0.134	0.195
PaymentMethod_Mailed check	-0.1923	0.099	-1.946	0.052	-0.386	0.001
Streaming	-0.2931	0.179	-1.634	0.102	-0.645	0.058

- Based on the results, we rejected (dropped) some variables with p-value > 0.05 which are not significant variable

Generalized Linear Model Regression Results

Dep. Variable:	Churn	**No. Observations:**	4922
Model:	GLM	**Df Residuals:**	4907
Model Family:	Binomial	**Df Model:**	14
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2035.8
Date:	Mon, 31 May 2021	**Deviance:**	4071.7
Time:	01:12:11	**Pearson chi2:**	5.16e+03
No. Iterations:	11		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Dependents	-0.3089	0.097	-3.196	0.001	-0.498	-0.119
tenure	-0.8218	0.067	-12.292	0.000	-0.953	-0.691
OnlineSecurity	-0.5860	0.102	-5.746	0.000	-0.786	-0.386
OnlineBackup	-0.3036	0.093	-3.273	0.001	-0.485	-0.122
TechSupport	-0.6201	0.103	-6.007	0.000	-0.822	-0.418
PaperlessBilling	0.3581	0.089	4.037	0.000	0.184	0.532
MonthlyCharges	1.0504	0.069	15.320	0.000	0.916	1.185
MultipleLines_No	-1.0368	0.076	-13.671	0.000	-1.185	-0.888
MultipleLines_No phone service	0.1508	0.142	1.061	0.289	-0.128	0.430
MultipleLines_Yes	-0.8120	0.083	-9.806	0.000	-0.974	-0.650
InternetService_DSL	-0.0295	0.110	-0.269	0.788	-0.244	0.185
Contract_Month-to-month	0.1550	0.081	1.918	0.055	-0.003	0.313
Contract_One year	-0.6129	0.107	-5.740	0.000	-0.822	-0.404
Contract_Two year	-1.2401	0.162	-7.671	0.000	-1.557	-0.923
PaymentMethod_Bank transfer (automatic)	-0.3166	0.110	-2.868	0.004	-0.533	-0.100
PaymentMethod_Credit card (automatic)	-0.2913	0.112	-2.605	0.009	-0.511	-0.072

LR Model #1 – Full variables (filtered) - Interpretation

	coef	Odds Ratio
Dependents	-0.3089	0.734261
tenure	-0.8218	0.439619
OnlineSecurity	-0.5860	0.556533
OnlineBackup	-0.3036	0.738135
TechSupport	-0.6201	0.537910
PaperlessBilling	0.3581	1.430568
MonthlyCharges	1.0504	2.858770
MultipleLines_No	-1.0368	0.354577
MultipleLines_No phone service	0.1508	1.162792
MultipleLines_Yes	-0.8120	0.443986
InternetService_DSL	-0.0295	0.970951
Contract_Month-to-month	0.1550	1.167639
Contract_One year	-0.6129	0.541781
Contract_Two year	-1.2401	0.289368
PaymentMethod_Bank transfer (automatic)	-0.3166	0.728596
PaymentMethod_Credit card (automatic)	-0.2913	0.747256

- Odds ratio Interpretation: Customers who have...
  - Dependents are 27% less likely to churn
  - Longer tenure are 56% less likely to churn
  - Online Security are 44% less likely to churn
  - Online Backup are 26% less likely to churn
  - TechSupport are 46% less likely to churn
  - No Multiplelines are 65% / Multiplelines are 56% less likely to churn
  - InternetService_DSL are 3% less likely to churn
  - One year contract are 46% / Two year contract are 71% less likely to churn
  - Payment with bank transfer(automatic) are 27% less likely to churn
  - Payment with Credit card(automatic) are 25% less likely to churn
  - Paperless Billing are 1.4 times (40%) more likely to churn
  - Higher MonthlyCharges are 2.9 times more likely to churn

    (However, since MonthlyCharges 95% Confidence interval [0.916, 1.185] include '1', this result is not statistically significant)
  - No phone service (no multiplelines) are 1.2 times (20%) more likely to churn
  - Month to month contract are 1.2 times (20%) more likely to churn

LR Model #1 – Full variables (filtered) - Insights

Summary

- **Customers with shorter contract (like month-to-month) are more likely to churn than the others with longer contract form / Customers with longer tenure are less likely to churn**

  → Think of the method to make customers stick to your company for long-term

  → Consider a marketing campaign that gives more promotion when the short-term contract / relatively new customers switch to the longer contract

- **Customers who uses automatic payment method with card or transfer are less likely to churn**

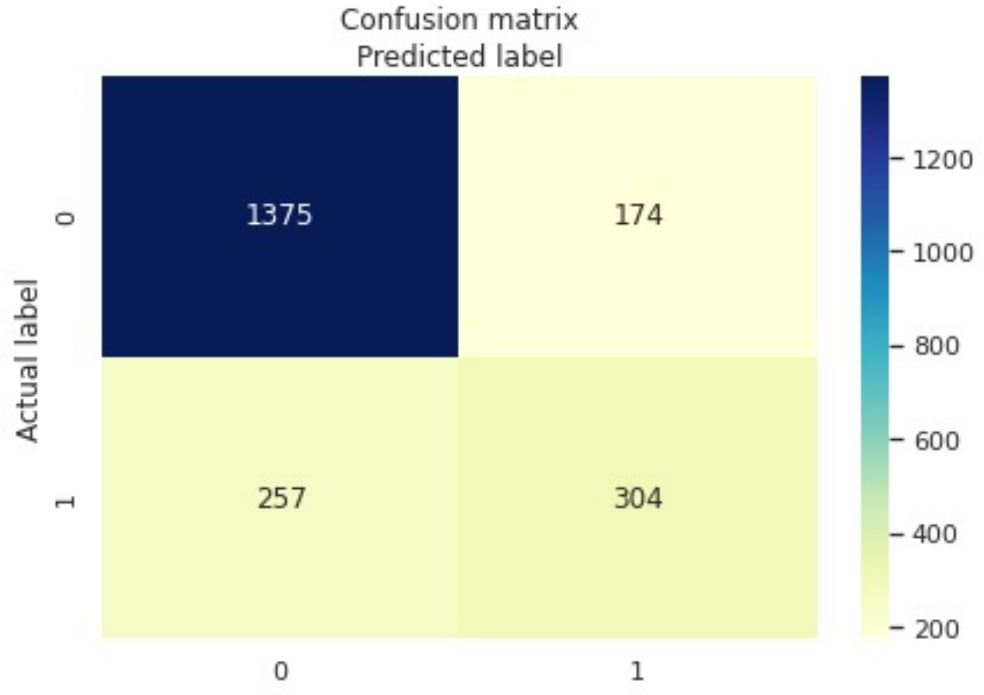
  → Targeting the customers who are still not using automatic payment to enroll for it

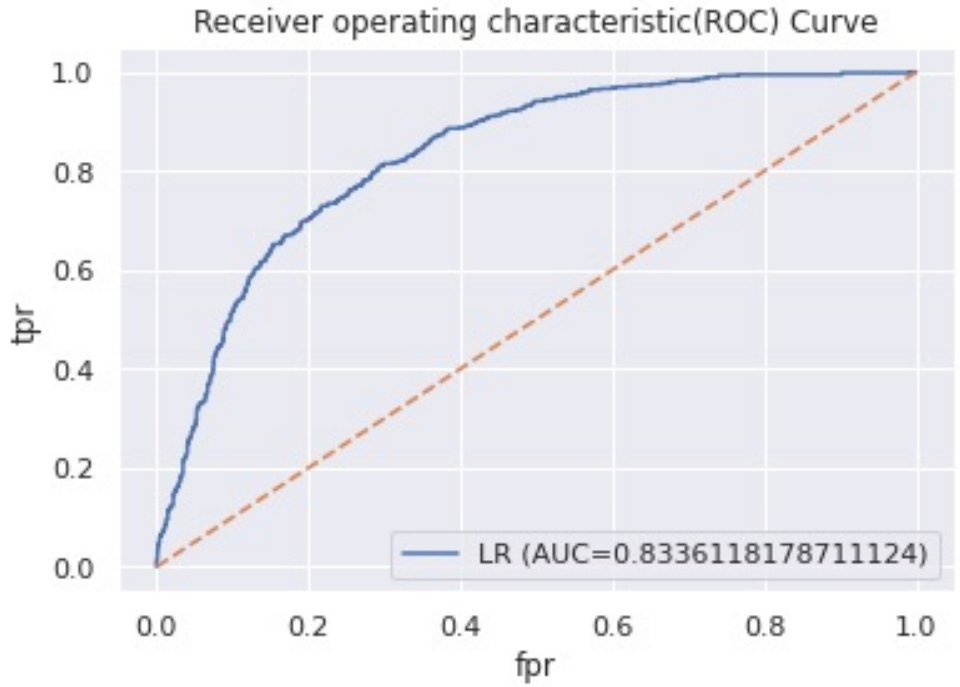
- **Customers who uses Paperless Billing are more likely to churn**

  → It's hard to guess what's the exact cause in here, but maybe we need to check how the company's paperless billing service is working and what we can improve.

  (ex. Does payment amount error occurs often?, Are there any complaints regarding this billing service?)

LR Model #1 – Full variables filtered (p-value)





	Model	Test Accuracy	Train Accuracy	Precision	Recall	F1 Score
0	Logistic Regression (var select with p-value)	0.795735	0.802519	0.541889	0.635983	0.585178

LR Model #2 – Selected variables using RFE

What is RFE (Recursive Feature Elimination)?

- Similar to Backward selection, but still a bit more advanced
- This approach does the whole cycle of eliminations and then chooses the best subset of it while ordinary Backward method stops at the point where the score starts decreasing
- RFE is based on the idea to repeatedly construct a model (ex. SVM / Regression) and choose either the best or worst performing feature (ex. based on coefficients), 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.
- Features are then ranked according to when they were eliminated.
- As such, it is a **greedy optimization**

LR Model #2 – Selected variables using RFE

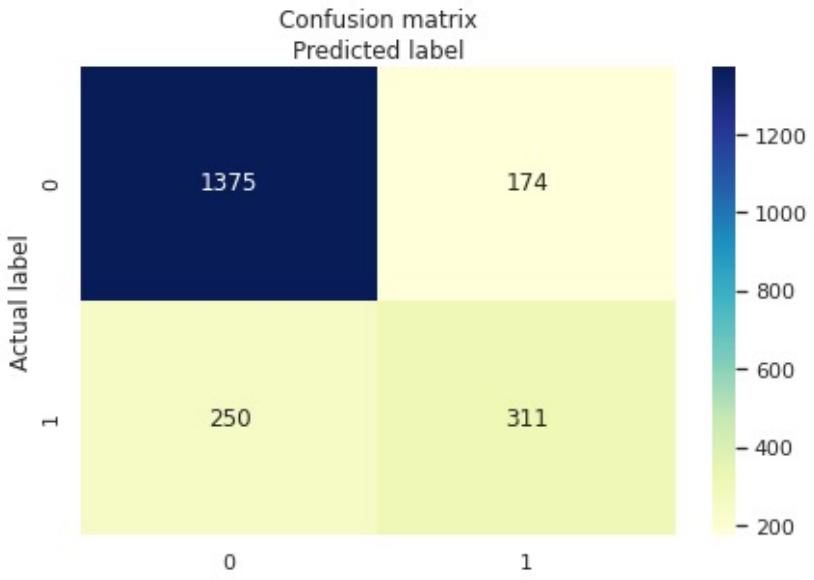
```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import confusion_matrix
from sklearn import preprocessing
from sklearn.feature_selection import RFE

# Feature extraction
lr2_model = LogisticRegression(solver='lbfgs', max_iter=15000)
rfe = RFE(lr2_model, n_features_to_select=None) #None means just half of the features are selected
fit = rfe.fit(X_train, y_train)
print("Num Features: %s" % (fit.n_features_))
print("Selected Features: %s" % (fit.support_))
print("Feature Ranking: %s" % (fit.ranking_))
```

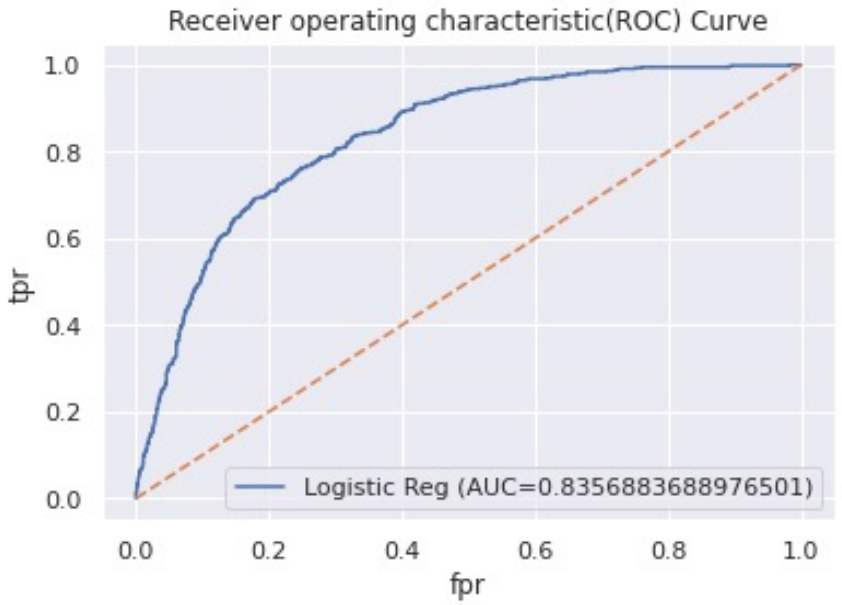
```
Num Features: 12
Selected Features: [False False False  True  True  True  True False  True  True  True  True
  True False False False False  True False  True False False  True False
 False]
Feature Ranking: [14  4 10  1  1  1  1  5  1  1  1  1  1  2 12 11  8  1  9  1  6  7  1 13
  3]
```

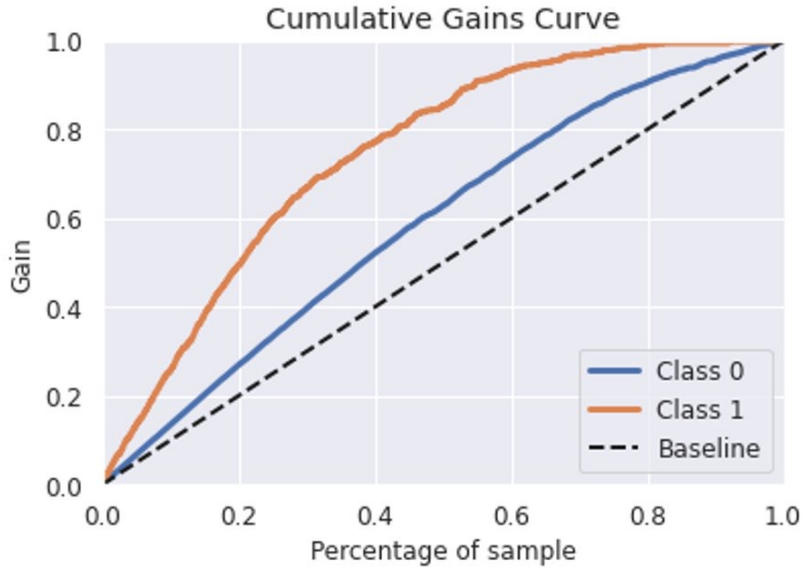
- RFE Selected Variables (12)

```
['Dependents',
 'tenure',
 'OnlineSecurity',
 'OnlineBackup',
 'TechSupport',
 'PaperlessBilling',
 'MonthlyCharges',
 'MultipleLines_No',
 'MultipleLines_No phone service',
 'Contract_Month-to-month',
 'Contract_Two year',
 'PaymentMethod_Electronic check']
```

LR Model #2 – Selected variables using RFE







	Model	Test Accuracy	Train Accuracy	Precision	Recall	F1 Score
0	Logistic Regression (selected var with RFE)	0.799052	0.802113	0.554367	0.641237	0.594646

LR Model #2 – Selected variables using RFE - Interpretation

Generalized Linear Model Regression Results

Dep. Variable:	Churn	**No. Observations:**	4922
Model:	GLM	**Df Residuals:**	4910
Model Family:	Binomial	**Df Model:**	11
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2105.8
Date:	Mon, 31 May 2021	**Deviance:**	4211.7
Time:	01:13:32	**Pearson chi2:**	5.09e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Dependents	-0.5606	0.092	-6.062	0.000	-0.742	-0.379
tenure	-0.9348	0.065	-14.446	0.000	-1.062	-0.808
OnlineSecurity	-0.8350	0.097	-8.651	0.000	-1.024	-0.646
OnlineBackup	-0.5420	0.089	-6.100	0.000	-0.716	-0.368
TechSupport	-0.8529	0.098	-8.737	0.000	-1.044	-0.662
PaperlessBilling	-0.0661	0.080	-0.826	0.409	-0.223	0.091
MonthlyCharges	1.0401	0.063	16.432	0.000	0.916	1.164
MultipleLines_No	-0.6603	0.088	-7.496	0.000	-0.833	-0.488
MultipleLines_No phone service	0.5482	0.159	3.451	0.001	0.237	0.860
Contract_Month-to-month	-0.1447	0.092	-1.578	0.115	-0.325	0.035
Contract_Two year	-1.3701	0.185	-7.392	0.000	-1.733	-1.007
PaymentMethod_Electronic check	0.1459	0.082	1.786	0.074	-0.014	0.306

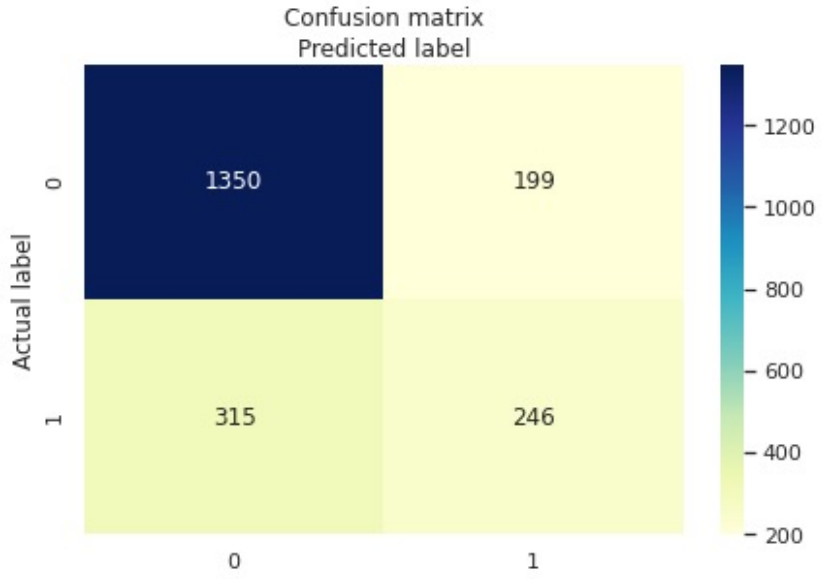
LR Model #2 – Selected variables using RFE - Interpretation

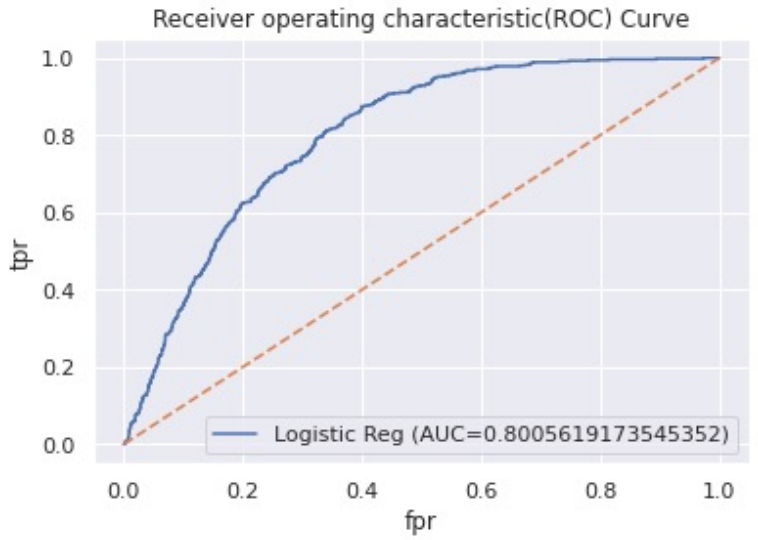
	coef	Odds Ratio
Dependents	-0.5606	0.570861
tenure	-0.9348	0.392652
OnlineSecurity	-0.8350	0.433876
OnlineBackup	-0.5420	0.581599
TechSupport	-0.8529	0.426159
PaperlessBilling	-0.0661	0.936048
MonthlyCharges	1.0401	2.829472
MultipleLines_No	-0.6603	0.516676
MultipleLines_No phone service	0.5482	1.730216
Contract_Month-to-month	-0.1447	0.865252
Contract_Two year	-1.3701	0.254073
PaymentMethod_Electronic check	0.1459	1.157077

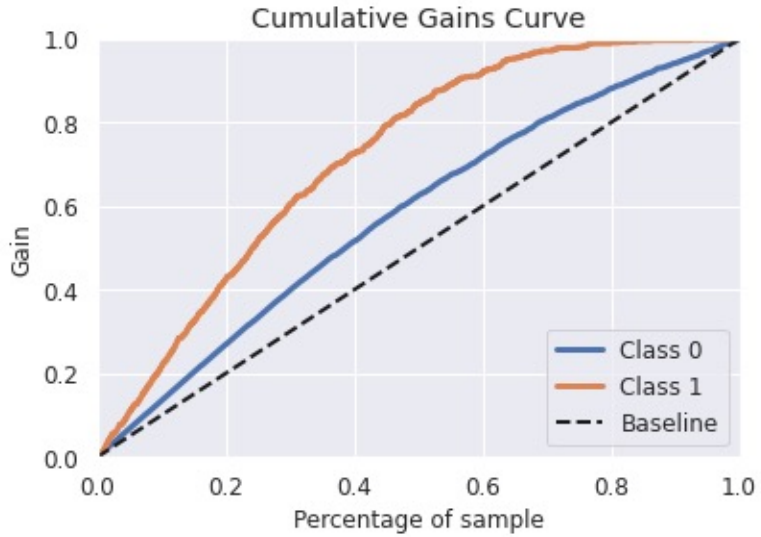
- Odds ratio Interpretation: Customers who have...
  - Dependents are 43% less likely to churn
  - Longer tenure are 61% less likely to churn
  - Online Security are 57% less likely to churn
  - Online Backup are 42% less likely to churn
  - TechSupport are 58% less likely to churn
  - Paperless Billing are 7% less likely to churn
  - No Multiplelines are 49% less likely to churn
  - Month to month contract are 13% less likely to churn
  - InternetService_DSL are 3% less likely to churn
  - Two year contract are 75% less likely to churn
  - Higher MonthlyCharges are 2.8 times more likely to churn

    (However, since MonthlyCharges 95% Confidence interval [0.916, 1.164] include '1', this result is not statistically significant)
  - No phone service (no multiplelines) are 1.7 times (70%) more likely to churn
  - Payment method with electronic check are 1.15 times (15%) more likely to churn

LR Model #3 – Selected variables (from Chi-square test + Variable Clustering)







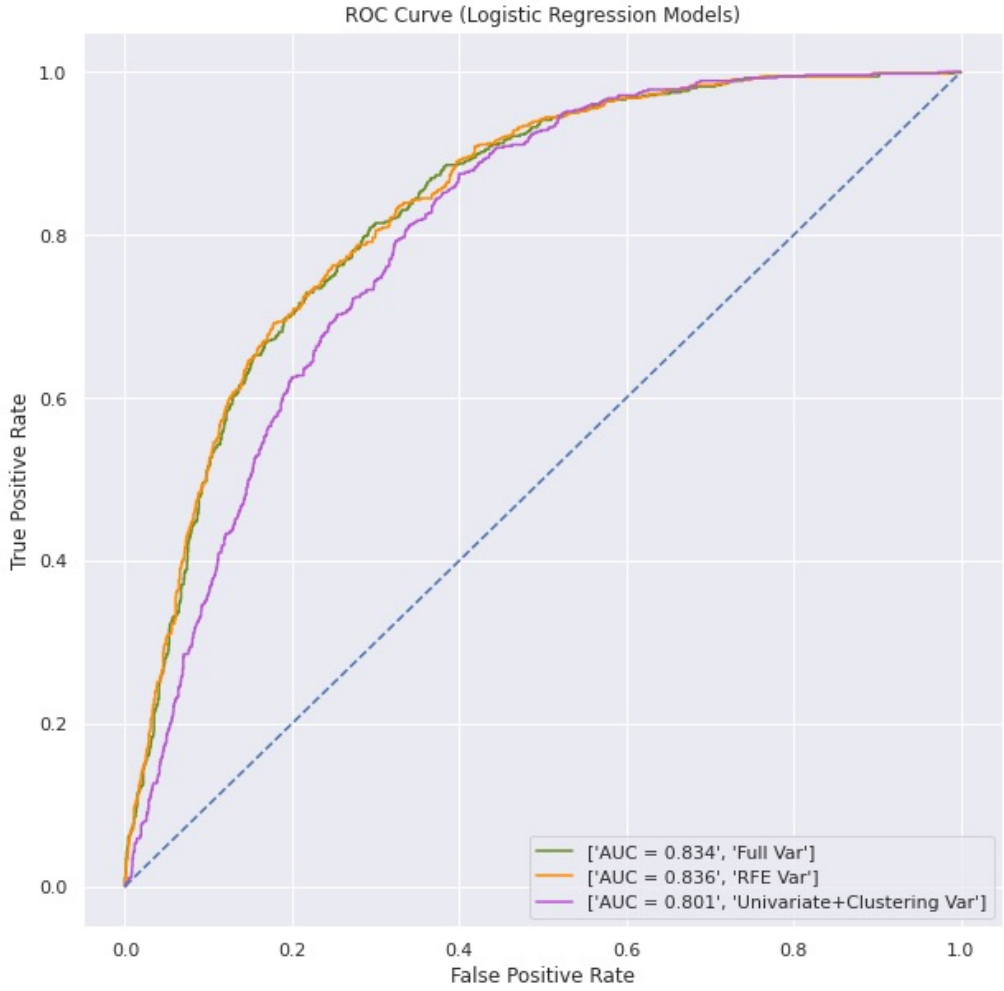
	Model	Test Accuracy	Train Accuracy	Precision	Recall	F1 Score
0	Logistic Regression (Univariate+Clustering)	0.756398	0.776107	0.438503	0.552809	0.489066

LR Model Summary

Comparing 4 types of Logistic Regression models with different feature selection methods,

- Model using selected variables from RFE method has the highest accuracy
- Don't have train, test overfitting
- Also has the best score for Precision, Recall, F1 Score

	Model	Test Accuracy	Train Accuracy	Precision	Recall	F1 Score
0	Logistic Regression (Base Model)	0.797630	0.804348	0.549020	0.639004	0.590604
1	Logistic Regression (var select with p-value)	0.795735	0.802519	0.541889	0.635983	0.585178
2	Logistic Regression (selected var with RFE)	0.799052	0.802113	0.554367	0.641237	0.594646
3	Logistic Regression (Univariate+Clustering)	0.756398	0.776107	0.438503	0.552809	0.489066



Decision Tree Model

Decision Tree Model

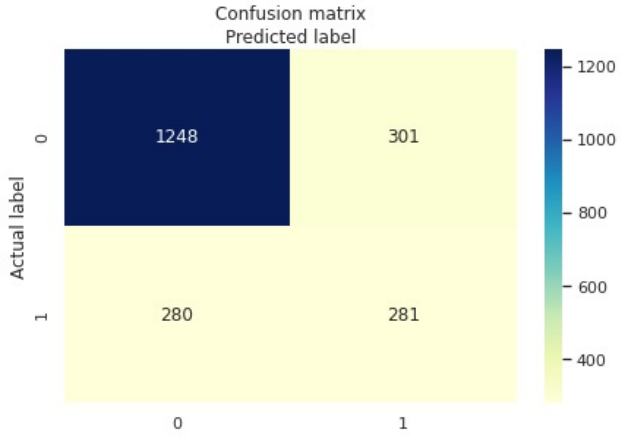
We will build 3 different DECISION TREES and later compare their accuracy:

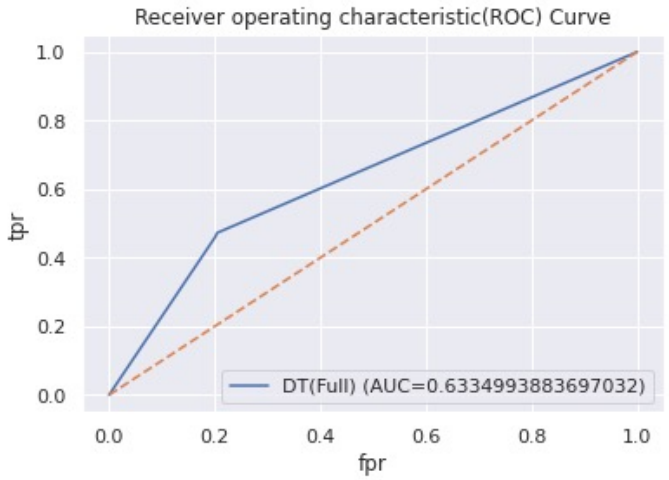
1) Model with full range of variables

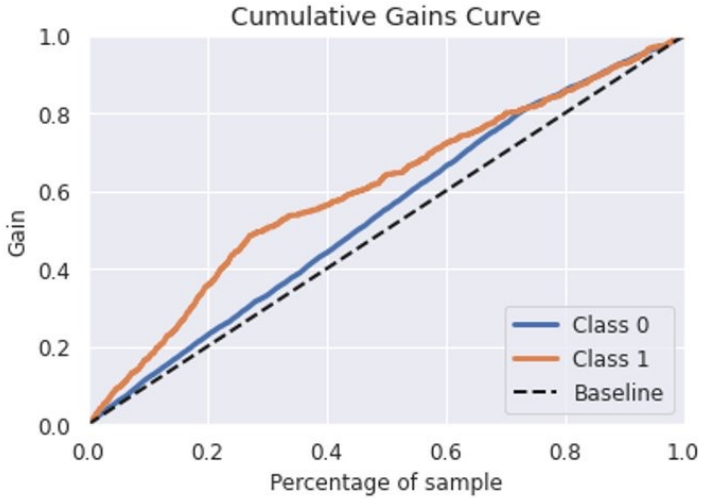
2) Prunned model

2) Model with Selected variables from Previous Explore Step (Univariate selection with Chi-square test + Variable Clustering)

DT Model #1 – Full variables



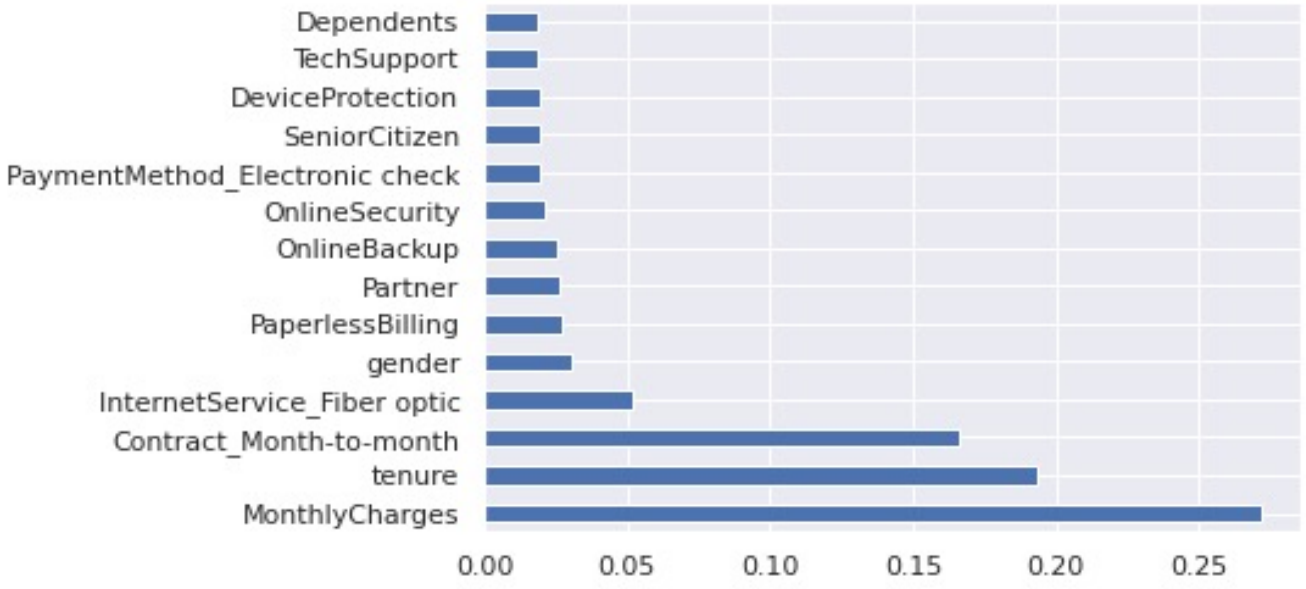




AUC score	
Test data	0.6334993883697032
Train data	0.9999956633140912

	precision	recall	F1-score	support
0	0.82	0.81	0.81	1549
1	0.48	0.50	0.49	561
accuracy			0.72	561

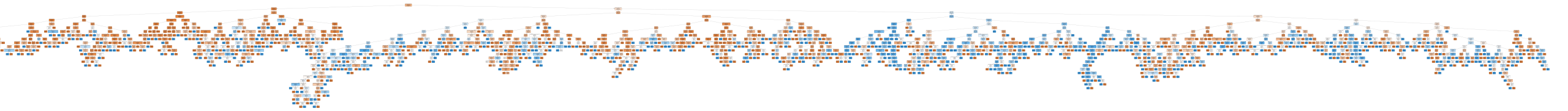
DT Model #1 – Full variables



Importance of variables based on Decision Tree Classifier:

Monthly Charges
Tenure
Contract Month-to-Month

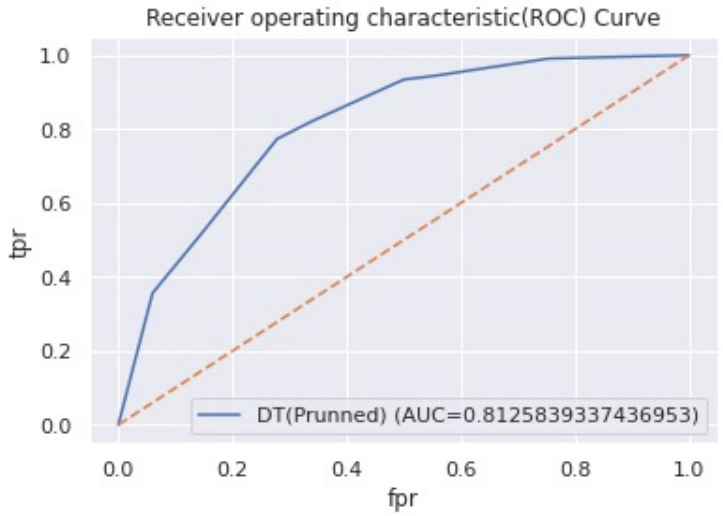
These variables influence churn target variable the most comparing to other variables.

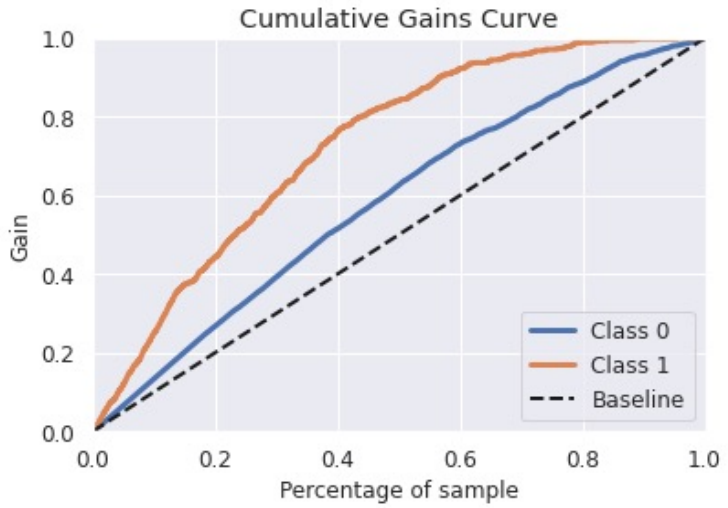


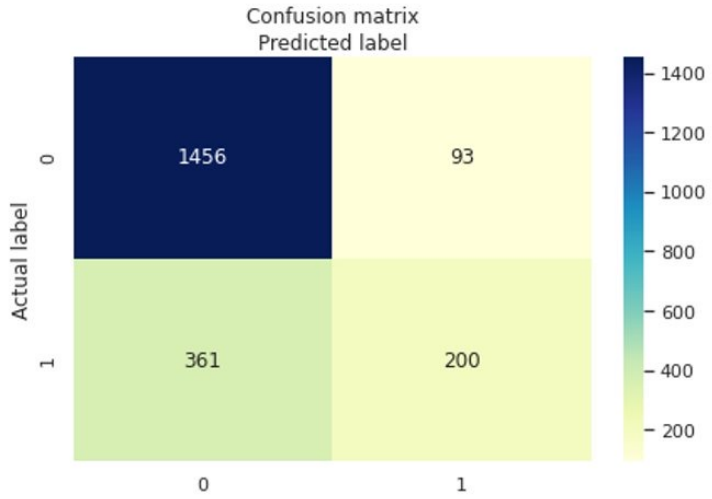
DT Model #2 – Prunned model

We prunned our model using Gini criterion. It is criterium that helps to find an optimal split for the decision tree. The optimum split is chosen by the features with less Gini index.

	precision	recall	F1-score	support
0	0.80	0.94	0.87	1549
1	0.68	0.36	0.47	561
accuracy			0.78	561



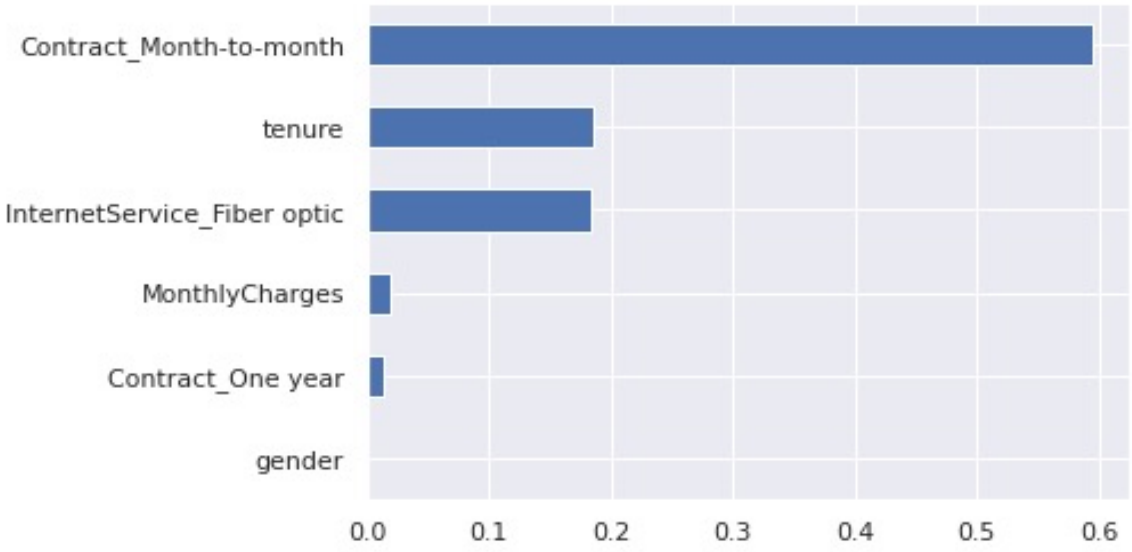




AUC score	
Test data	0.8129119010712448
Train data	0.8264480511568162

DT Model #2 – Prunned model

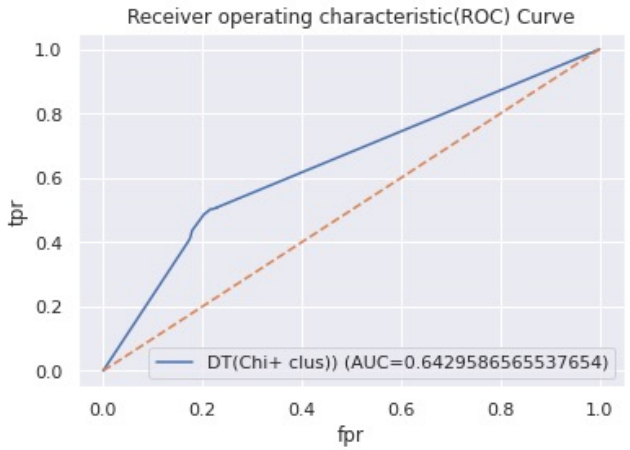
Importance of prunned decision tree variables

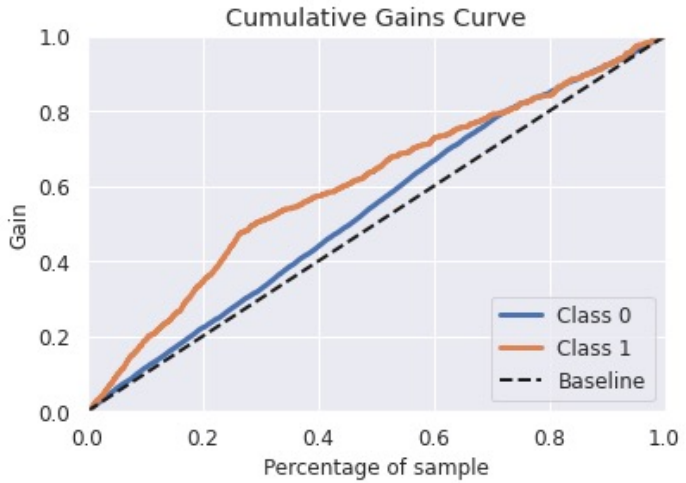


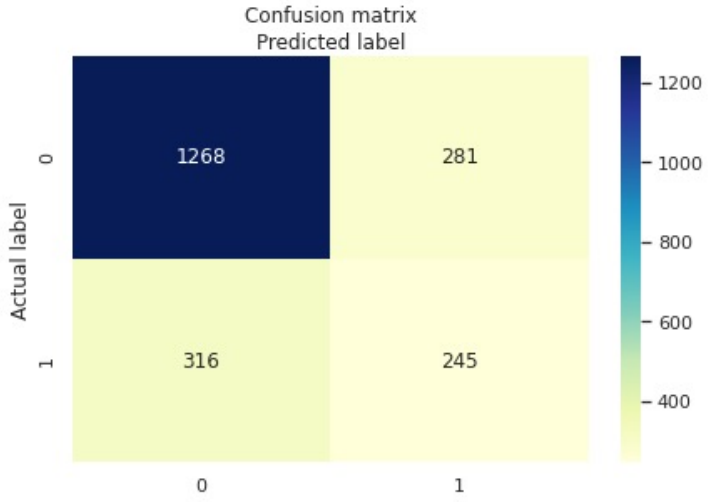
These variables influence churn target variable the most comparing to other Variables in prunned decision tree with Gini criterion:

Contract_Month-to-month
Tenure
InternetService_Fiber optic

DT Model #3 – Model with Selected variables (Univariate with Chi-square test + Variable Clustering)







AUC score	
Test data	0.6429586565537654
Train data	0.9955885327024196

	precision	recall	F1-score	support
0	0.80	0.82	0.81	1549
1	0.47	0.44	0.45	561
accuracy			0.72	561

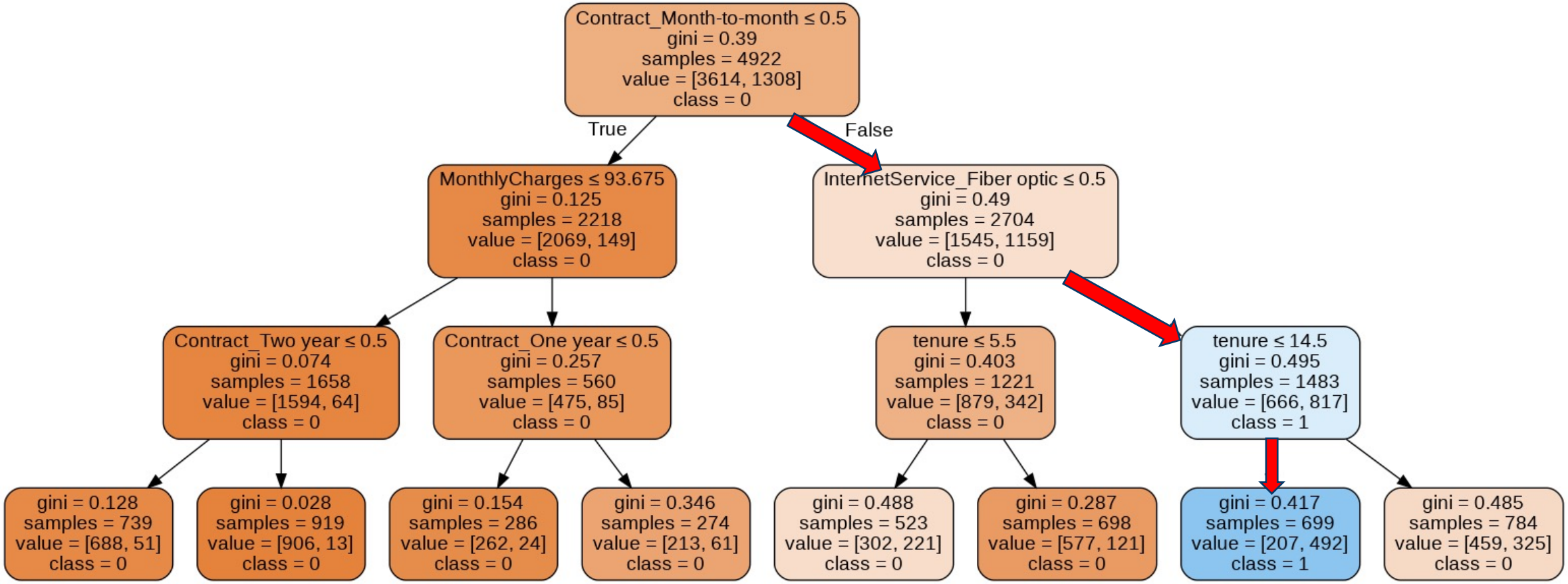
DT Model Summary

	Model	Test Accuracy	Train Accuracy	Precision	Recall	F1 Score
0	Decision Tree (Base Model: full set of variables)	0.728436	0.998375	0.495544	0.489437	0.492471
1	Decision Tree (Prunned model using Gini)	0.784834	0.792158	0.356506	0.682594	0.468384
2	Decision Tree (Chi+Clustered variables)	0.718483	0.959976	0.436720	0.468451	0.452030

Comparing 4 types of decision tree with different feature selection methods we can conclude that prunned decision tree using Gini criterion gives us the most accuracy among other decision trees.

In addition, it is not overfitted comparing to other decision trees.

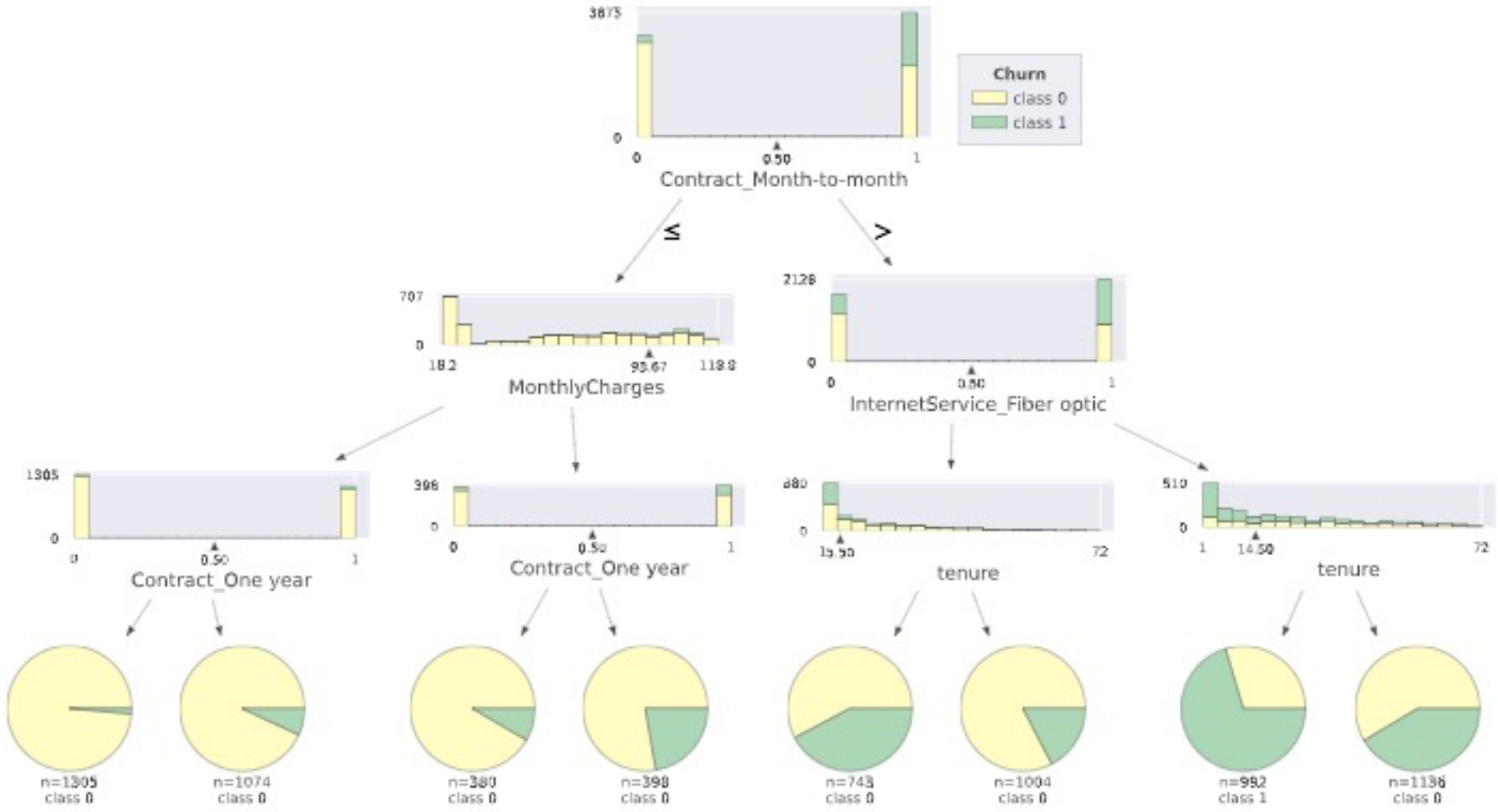
DT Prunned Model Interpretation



Insights from prunned decision tree:

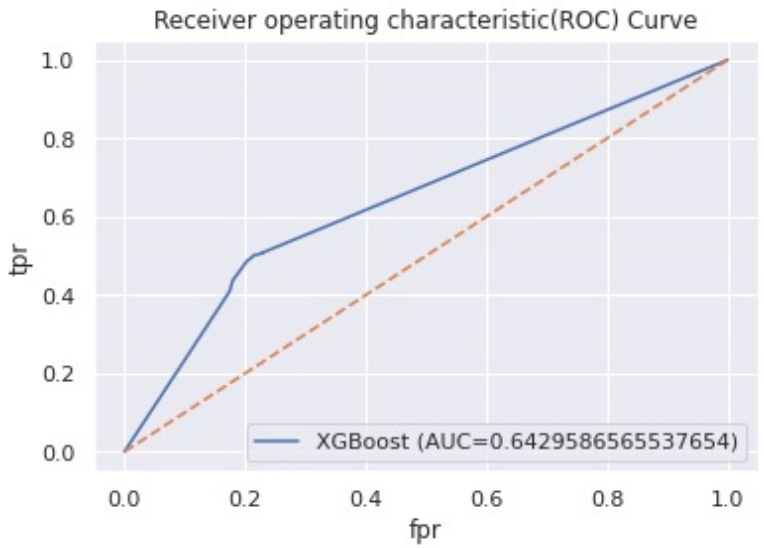
1. Having a month-to-month (best churn predictor) contract increases chance of a customer to leave
2. Having Fiber optic as the Internet service increases chance of a customer to leave
3. Having tenure less than 14.5 months increases the chance of a customer to leave

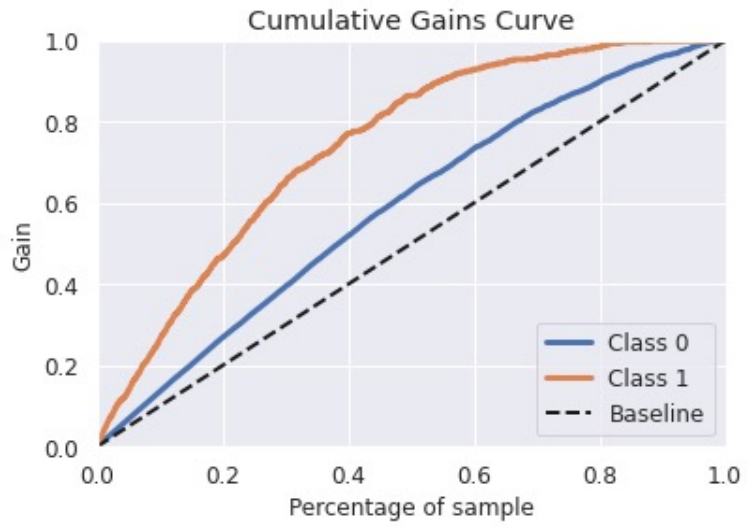
DT Prunned Model Interpretation

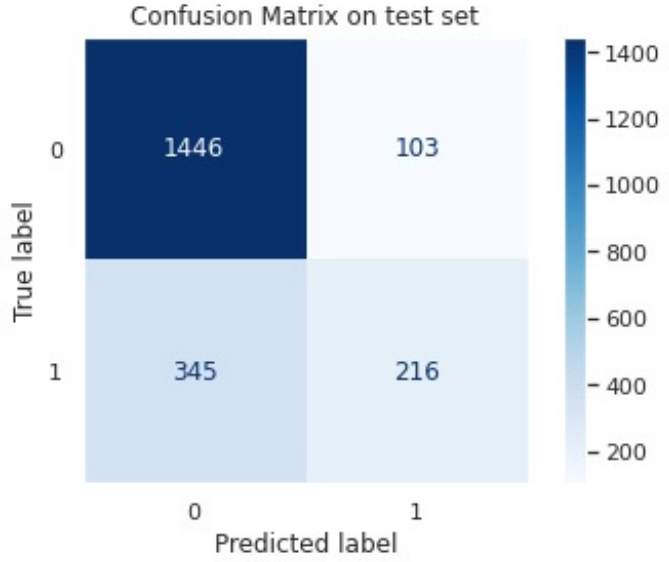


XGBoost Model

XGBoost Model



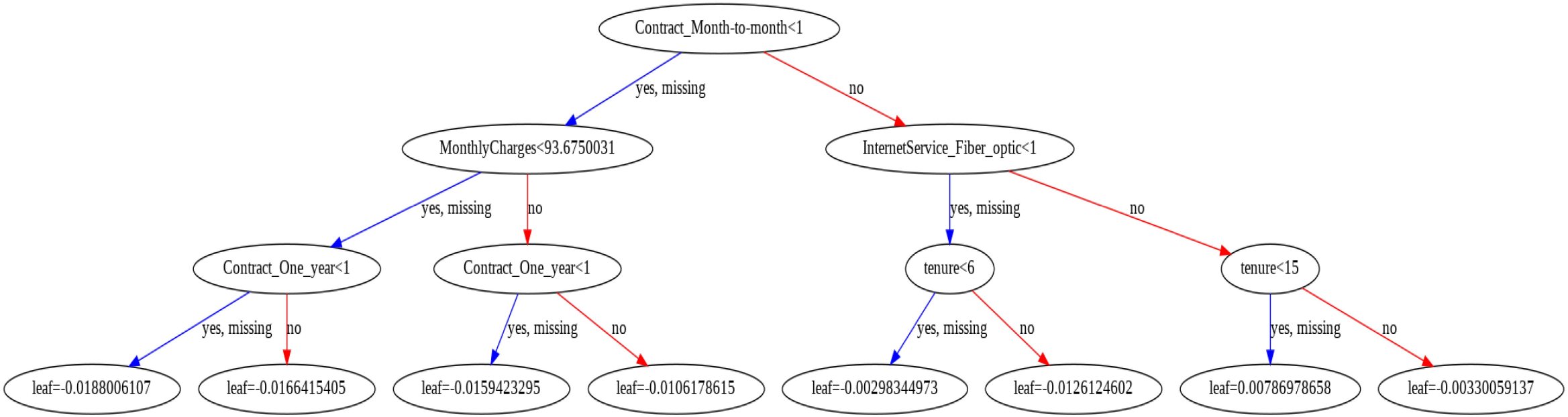




AUC score	
Test data	0.8281784924780407
Train data	0.8445341468533006

	precision	recall	F1-score	support
0	0.81	0.93	0.87	1549
1	0.68	0.39	0.49	561
accuracy			0.79	561

XGBoost Model



XGBoost model tree looks similar to Prunned decision tree. It leads us to the same conclusions about customer churn.

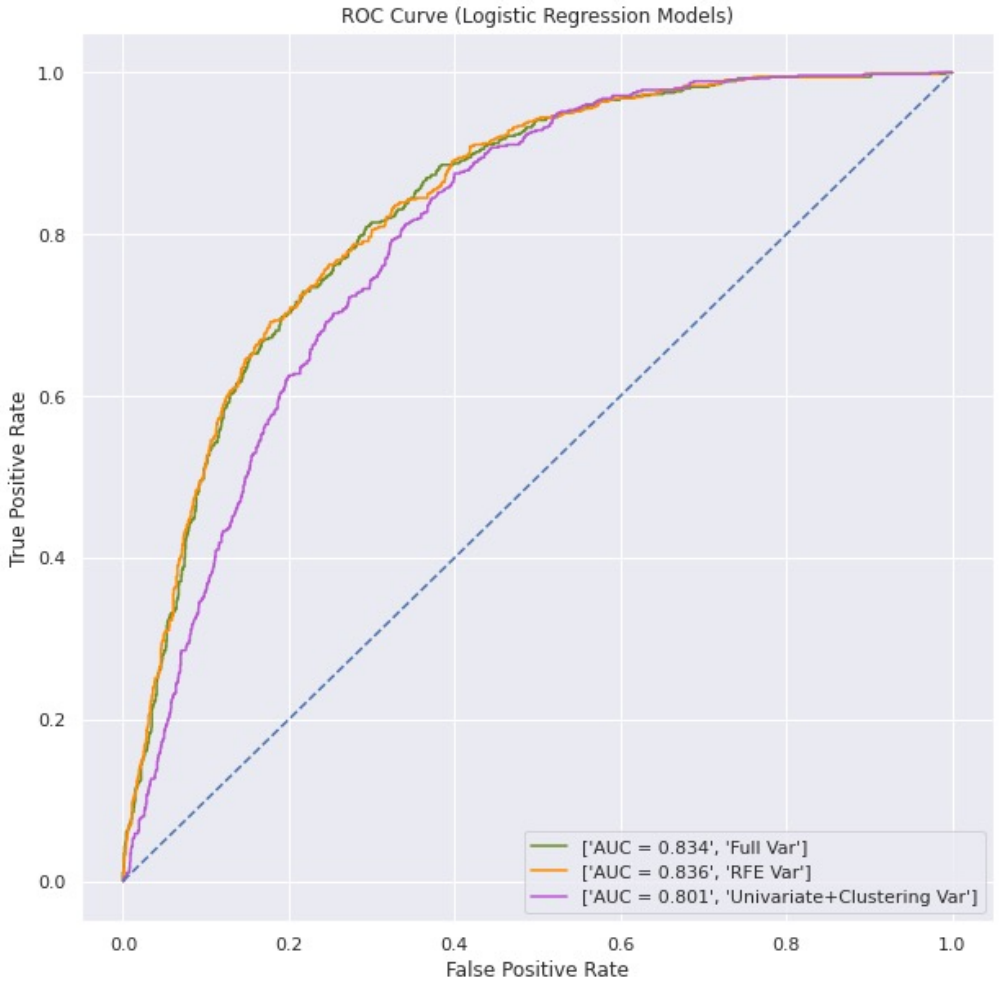
Step 5. Assess

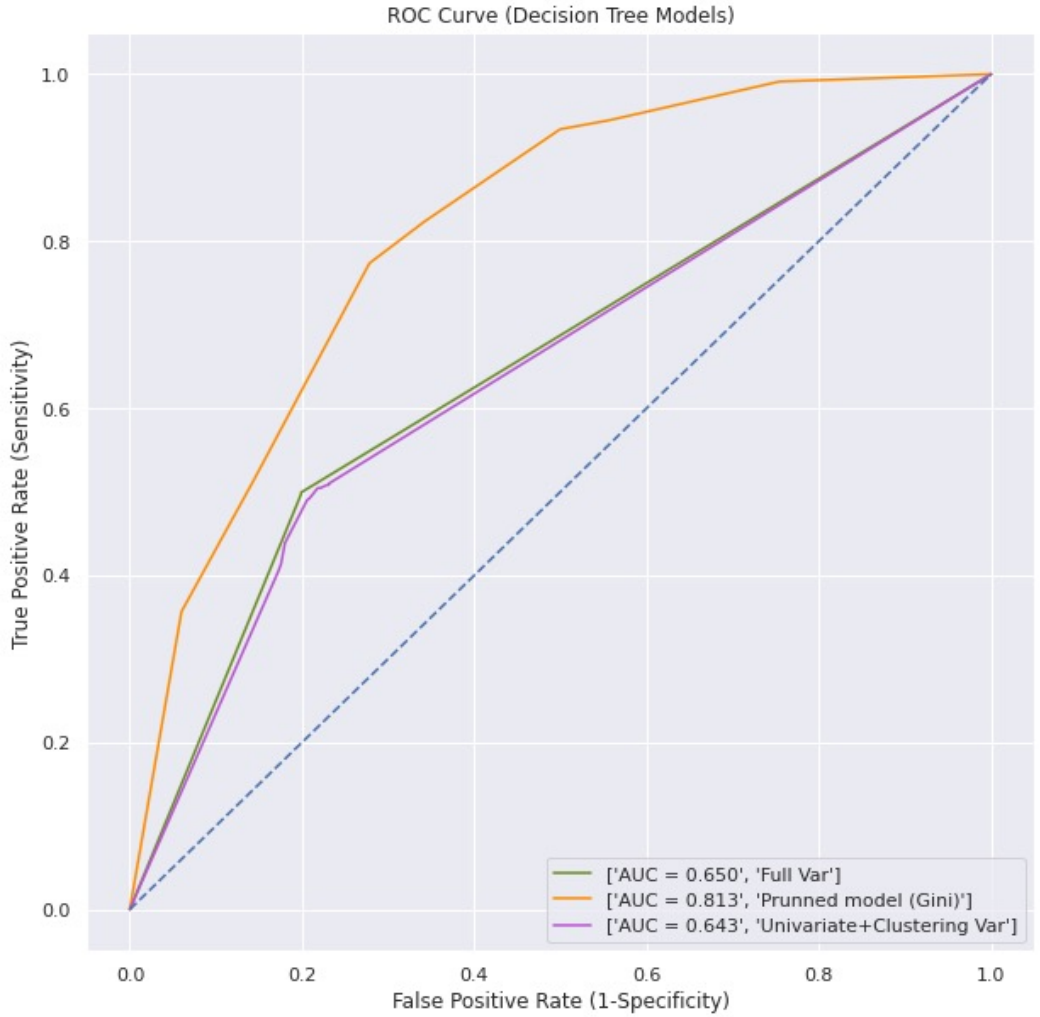
Model Comparison (metrics)

	Model	Test Accuracy	Train Accuracy	Precision	Recall	F1 Score
3	Logistic Regression (selected var with RFE)	0.799052	0.802113	0.554367	0.641237	0.594646
1	Logistic Regression (Base Model)	0.797630	0.804348	0.549020	0.639004	0.590604
2	Logistic Regression (var select with p-value)	0.795735	0.802519	0.541889	0.635983	0.585178
0	XGBoost	0.787678	0.795815	0.385027	0.677116	0.490909
6	Decision Tree (Prunned model using Gini criter...	0.784834	0.792158	0.356506	0.682594	0.468384
4	Logistic Regression (Univariate+Clustering)	0.756398	0.776107	0.438503	0.552809	0.489066
5	Decision Tree (Base Model: full set of variables)	0.720379	0.998375	0.493761	0.475129	0.484266
7	Decision Tree (Chi+Clustered variables)	0.718483	0.959976	0.438503	0.468571	0.453039

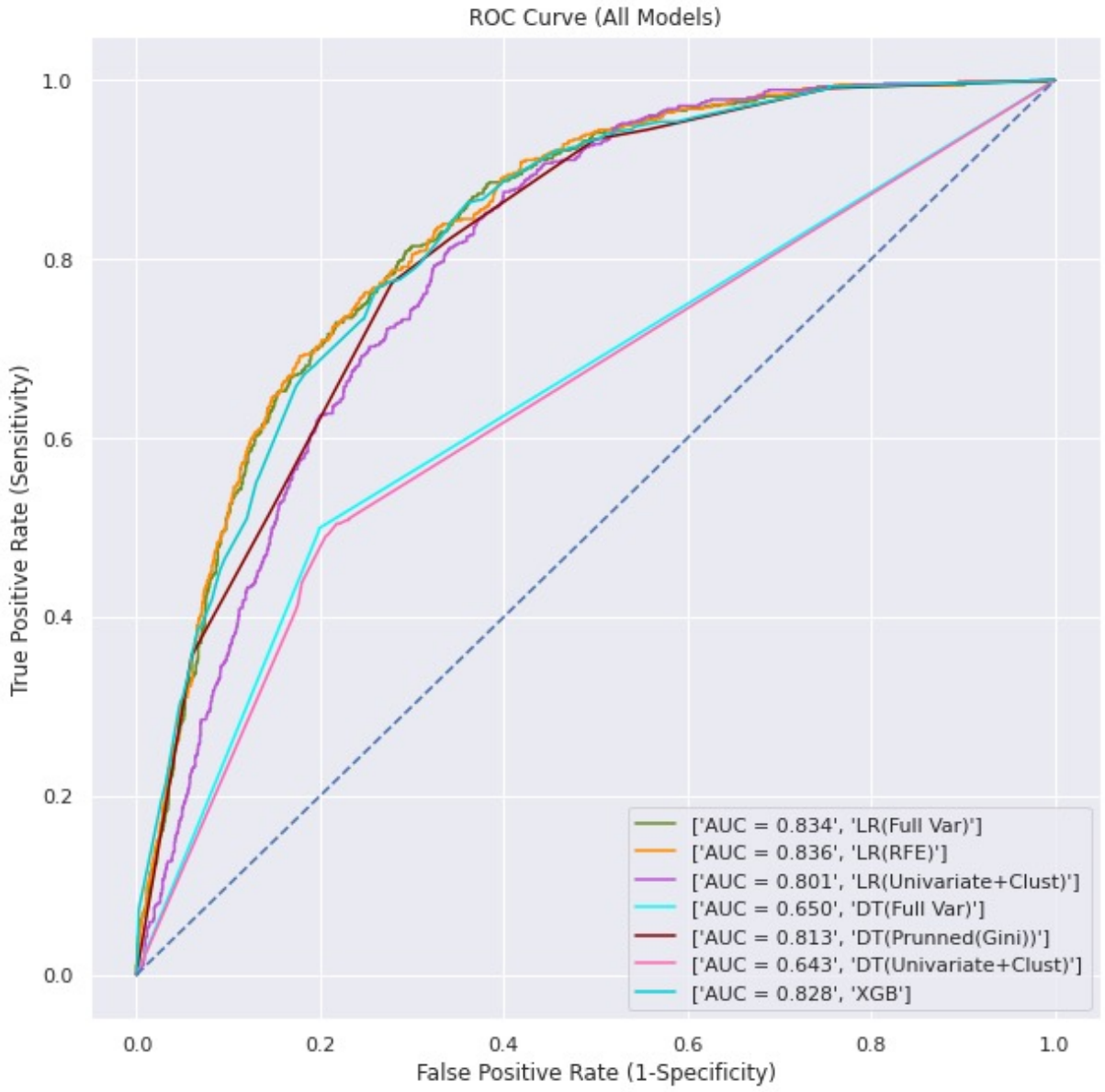
Conclusion: Logistic Regression model with selected RFE variables gives us the best performance among other models

Model Comparison (ROC Curve)





Model Comparison (ROC Curve)



Area under the curve for LR model with RFE variables is the biggest (AUC = 836) => Best performed model that gives us the best predictions.

Conclusion

Hypothesis Check

TRUE **Hypothesis 1:** Customers who pay for the service (MonthlyCharges) more money than on average has higher chance to stop using the service.

TRUE **Hypothesis 2:** Annual term contract customer has less chance to leave than monthly term contract customer.

TRUE **Hypothesis 3:** If a customer uses additional services of Telecom, she will not stop using the service.

TRUE **Hypothesis 4:** Internet service can affect the churn rate.

Main insights

- If we want to predict the churn rate of the customers on the future data, as Telco's business analysts we should use Logistic Regression model with RFE.

- Duration of the contract indeed influences on the churn rate.

- If the company wants to decrease churn rate, it should implement marketing campaign targeting clients who have more tendency to churn and offer them attractive deals.

- Main target would be to have clients signed up for 1-year or 2-year contracts promoting offers such as Online Security and Online Back up as well as Tech support.

Challenges

Challenges

- May have better model performance if we dealt with quite imbalanced target variable (Churn (1) 2 6.58%) using methods such as SMOTE
- Limited volume of dataset (7032 obs)
- Due to limited time, we couldn't check various results using different variable Selection and cluster ing method (ex. RFE CV method, PCA clustering, etc...)

Thank you