TELECOM CUSTOMER CHURN ANALYSIS

Prepared by: Susan Lum

November 2022

AGENDA

1. What is customer churn?
2. Why customer churn is important?
3. Objectives
4. About the Dataset
5. Pre-processing of Data
6. Exploratory Data Analysis (EDA)
7. Correlation of all variables with "Churn"
8. Model Building
9. Recommendations

Google Colab link:
https://colab.research.google.com/drive/1A5n-bFNv027iheLrrLUTiMH8Nq91yyla?usp=sharing

WHAT IS CUSTOMER CHURN ?

It is the phenomenon where customers of a business no longer purchase or interact with the business.

High churn means that a **higher** number of customers no longer want to purchase goods and services from the business.

WHY CUSTOMER CHURN IS IMPORTANT?

Churn leads to higher Customer Acquisition Cost (CAC) and reduce sales revenue.

Hence, it is important to analyze churn frequently and accurately for business sustainability.

Note: CAC = the total cost of sales and marketing required to acquire a customer.

OBJECTIVES:

1) To analyze and predict customer who churn
2) To Highlight the main variables/ factors influencing customer churn
3) To use various Machine Learning (ML) algorithms to build prediction models, evaluate accuracy & performance of these models.
4) To Find out the best model for business case & providing executive summary

ABOUT THE DATASET

Consists of 7,043 rows (customers) and 21 columns (variables)

No.	Variables	Description	Data Types
1	customerID	Customer ID	object
2	gender	Male / Female	object
3	**SeniorCitizen**	**0 = non-Senior Citizen; 1 = Senior Citizen**	**int64**
4	Partner	Yes = have partner; No = don’t have partner	object
5	Dependents	Yes = have dependents; No = don’t have dependents	object
6	**tenure**	**Number of months the customer has stayed with the company**	**int64**
7	PhoneService	Yes = have phone service; No = don’t have phone service	object
8	MultipleLines	Yes = have multiple lines; No = don’t have multiple lines	object
9	InternetService	Customer’s internet service provider (DSL, Fiber optic, No)	object
10	OnlineSecurity	Yes = have online security; No = don’t have online security; No internet service	object
11	OnlineBackup	Yes = have online backup; No = don’t have online backup; No internet service	object

ABOUT THE DATASET - continued

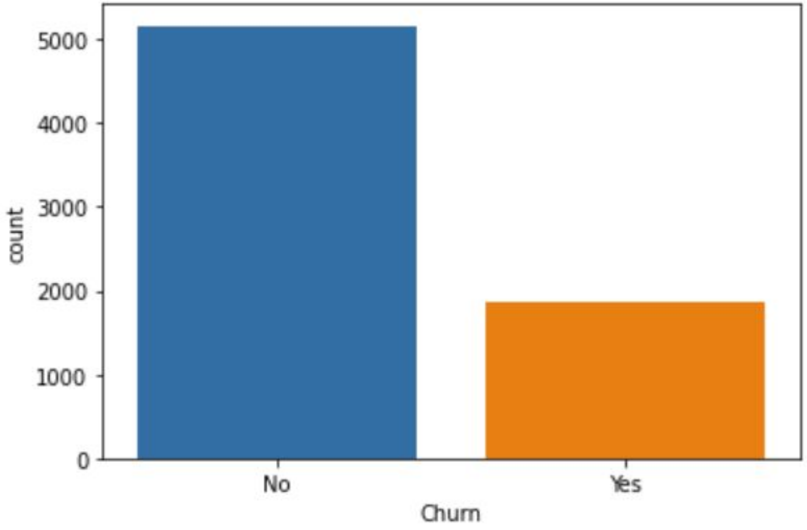
No.	Variables	Description	Data Types
12	DeviceProtection	Yes = have device protection; No = don't have device protection; No internet service	object
13	TechSupport	Yes = have tech support; No = don't have tech support; No internet service	object
14	StreamingTV	Yes = have streaming TV; No = don't have streaming TV; No internet service	object
15	StreamingMovies	Yes = have streaming TV; No = don't have streaming TV; No internet service	object
16	Contract	Contract term of the customer (month-to-month, one year, two year)	object
17	PaperlessBilling	Yes = have paperless billing; No = don't have paperless billing	object
18	PaymentMethod	customer's payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))	object
19	**MonthlyCharges**	**The amount charged to the customer monthly**	**float64**
20	TotalCharges	The total amount charged to the customer **(To change datatype)**	**object**
21	**Churn (Target Variable)**	**Yes = Churned customer; No = non-churned customer**	**object**

PRE-PROCESSING OF DATA

Items	Results/ FIndings/ Action
Check for null values	No missing values found
Check for duplicate data	No duplicate data found
Change data type of "TotalCharges" variable	Changed data type from Object to float
Check for null values (2nd time)	11 rows of NaN values found in "TotalCharges" column. Since the percentage of NaN values is very low (ie. 11/7,043 x 100 % = 0.15%), I decided to **remove** the rows.
Change "tenure" variable in bins of 12 months	Tenure is ranging from 0 to 72 months. For better visualization, I decided to group them in bins with new variable, "tenure_group" (ie. 1-12, 13-24, 25-36, 37-48, 49-60, 61-72)
Drop columns that do not have impact to Target variable	Dropped "customerID" and "tenure"
Convert target variable (Churn) in a binary numeric variable (ie. 0 & 1)	1 = Churn; 0 = Not Churn. This is for model building later.
Convert all categorical variables into numeric variables	This is for model building later.

EXPLORATORY DATA ANALYSIS (EDA)

Target Variable, Churn



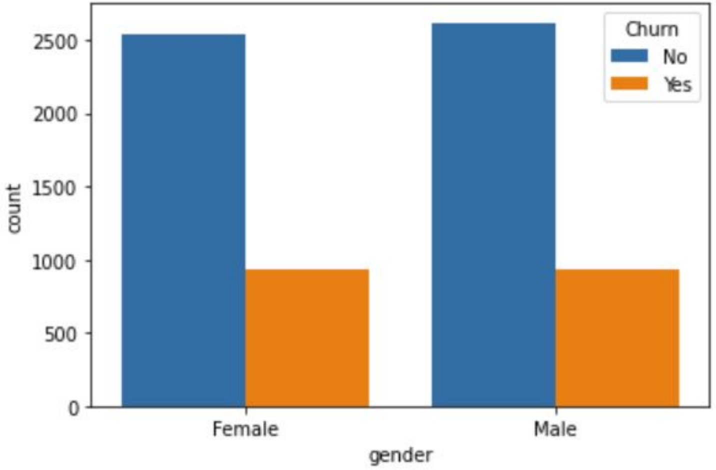
Insights:
Non churn customer contributed around 73% of total dataset.

The remaining 27% is contributed by customer who churn.

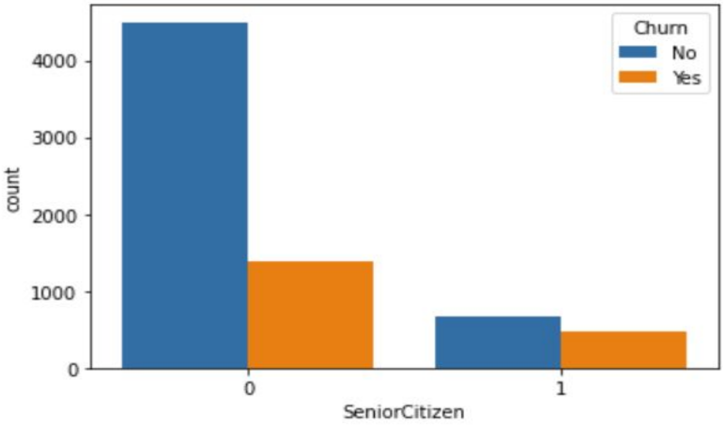
The dataset is highly imbalanced.

EDA – continued

Categorical Independent Variable: Gender
Target Variable: Churn

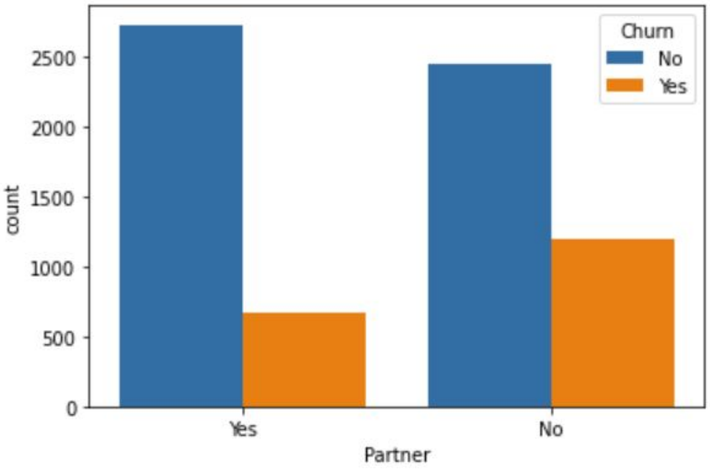


Categorical Independent Variable: SeniorCitizen
Target Variable: Churn

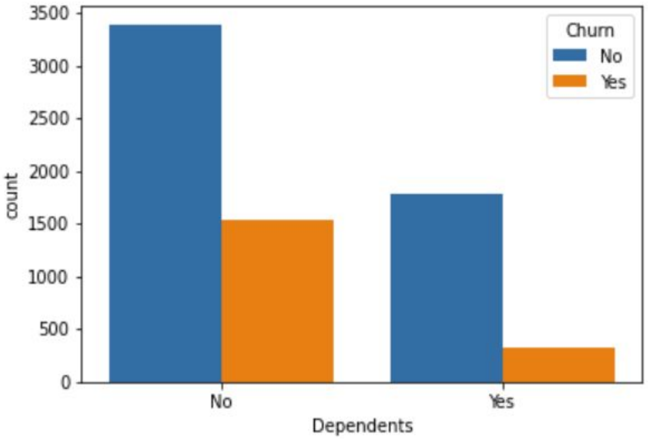


EDA – continued

Categorical Independent Variable: Partner
Target Variable: Churn

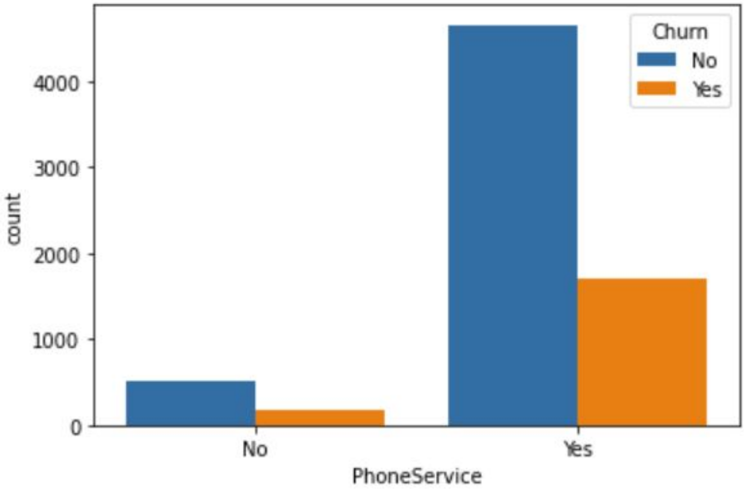


Categorical Independent Variable: Dependents
Target Variable: Churn

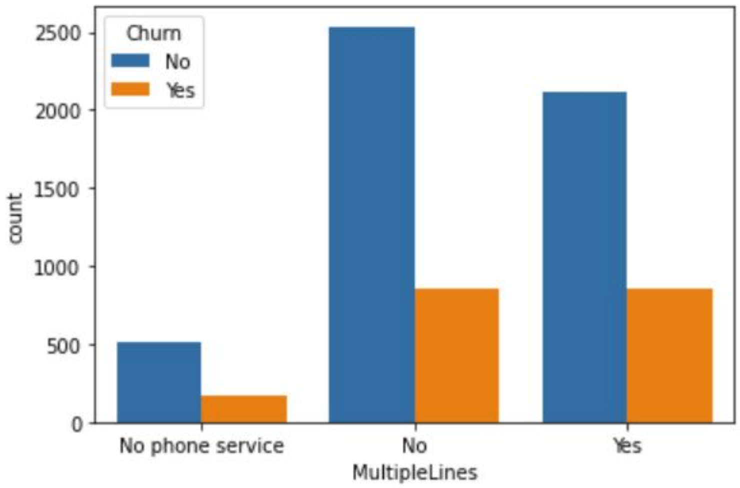


EDA - continued

Categorical Independent Variable: PhoneService
Target Variable: Churn

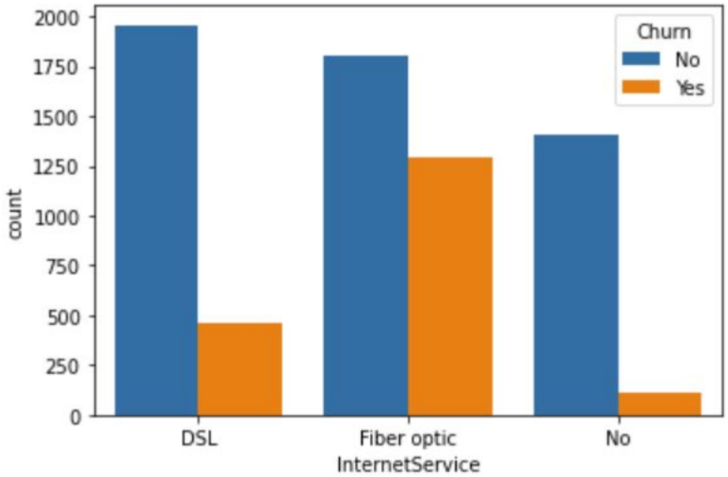


Categorical Independent Variable: MultipleLines
Target Variable: Churn

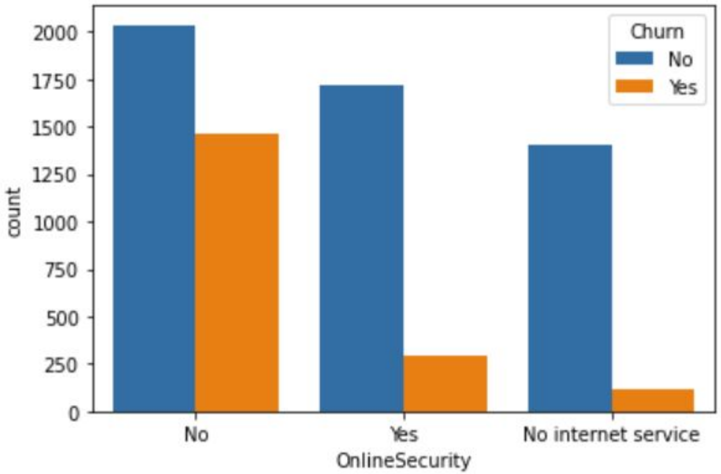


EDA - continued

Categorical Independent Variable: InternetService
Target Variable: Churn

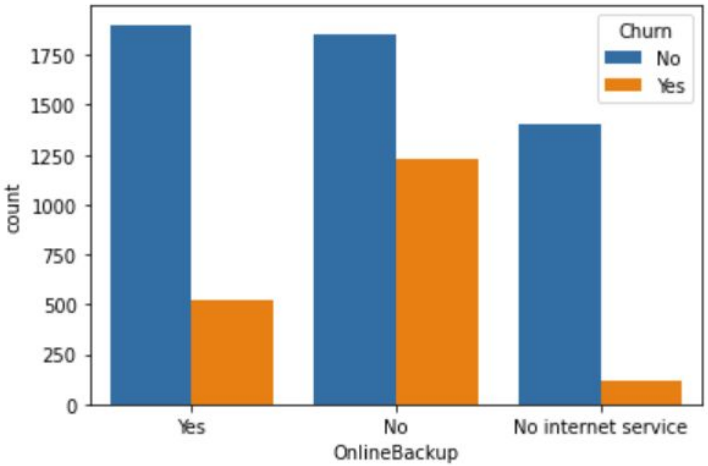


Categorical Independent Variable: OnlineSecurity
Target Variable: Churn

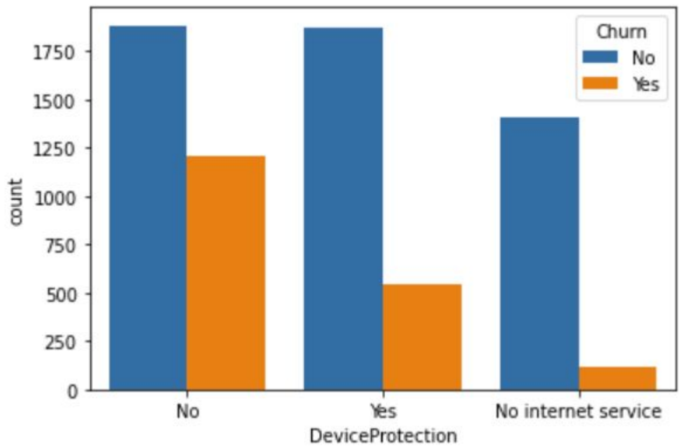


EDA – continued

Categorical Independent Variable: OnlineBackup
Target Variable: Churn

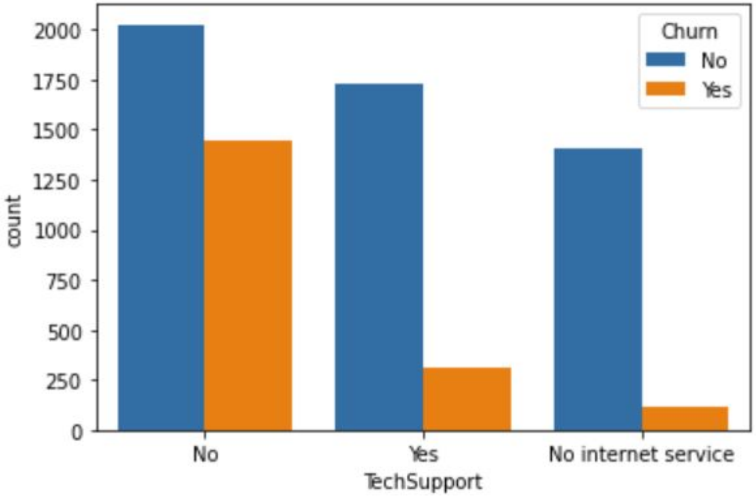


Categorical Independent Variable: DeviceProtection
Target Variable: Churn

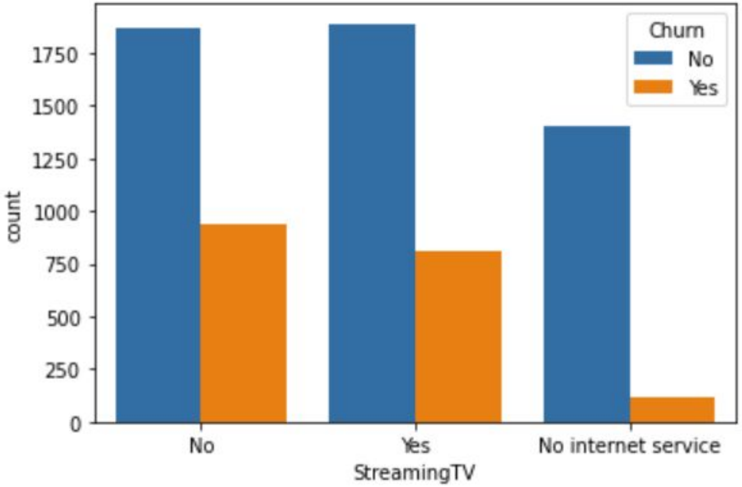


EDA – continued

Categorical Independent Variable: TechSupport
Target Variable: Churn

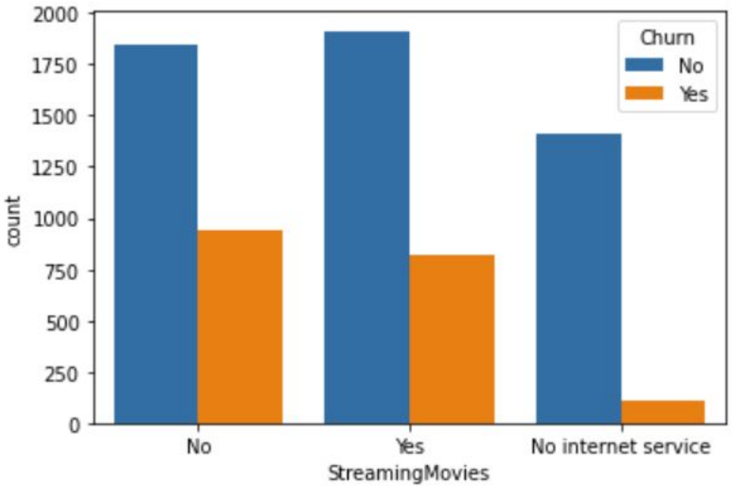


Categorical Independent Variable: StreamingTV
Target Variable: Churn

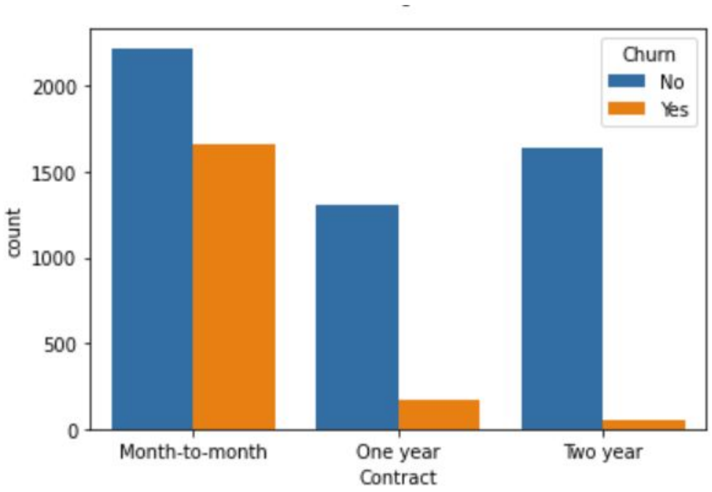


EDA – continued

Categorical Independent Variable: StreamingMovies
Target Variable: Churn

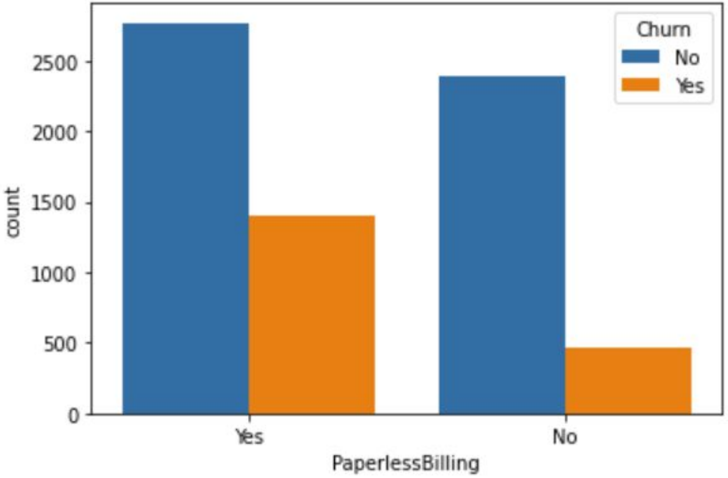


Categorical Independent Variable: Contract
Target Variable: Churn

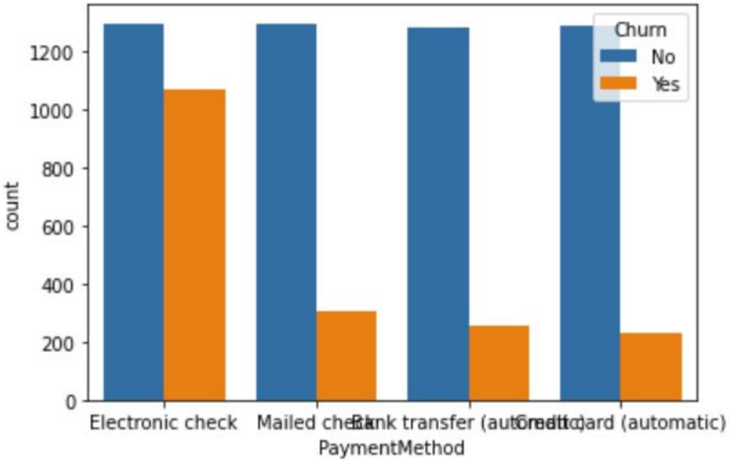


EDA – continued

Categorical Independent Variable: PaperlessBilling
Target Variable: Churn

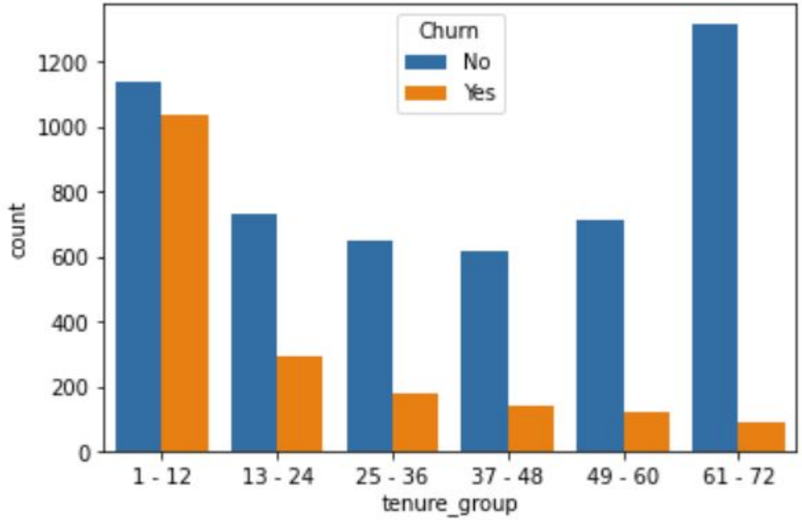


Categorical Independent Variable: PaymentMethod
Target Variable: Churn

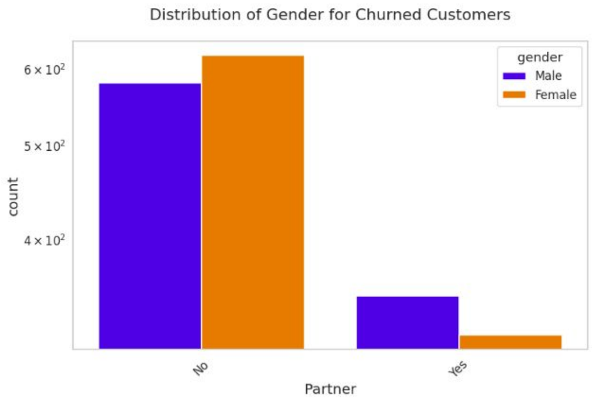


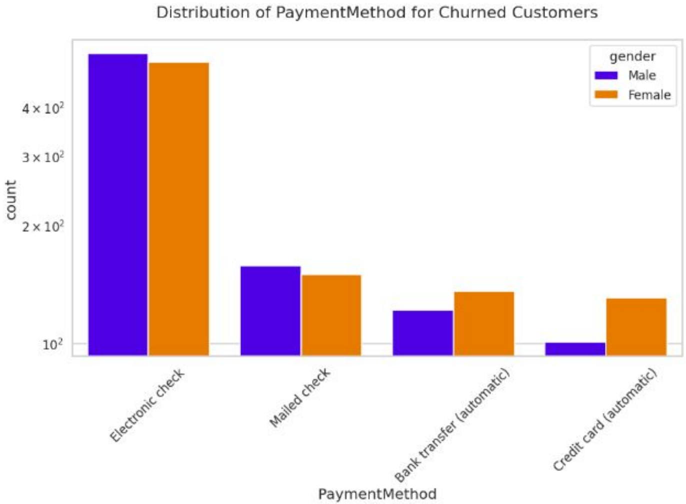
EDA – continued

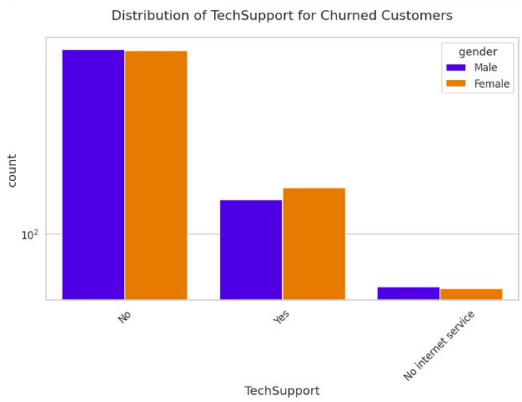
Categorical Independent Variable: tenure_group
Target Variable: Churn

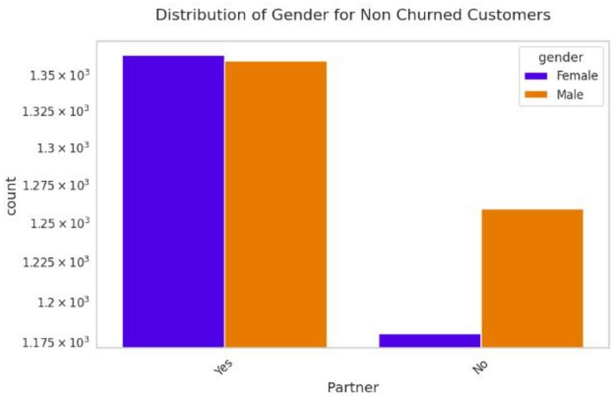


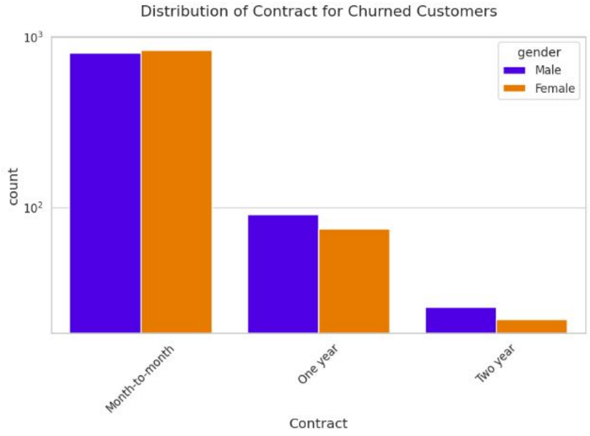
EDA - continued

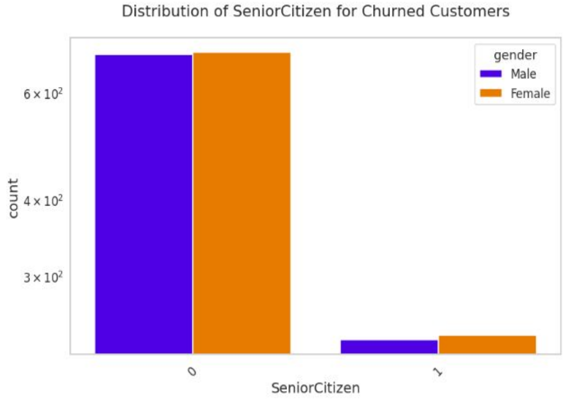








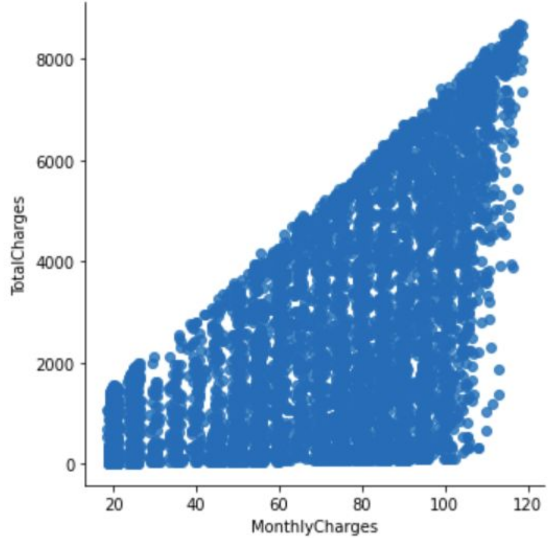




EDA - Insights

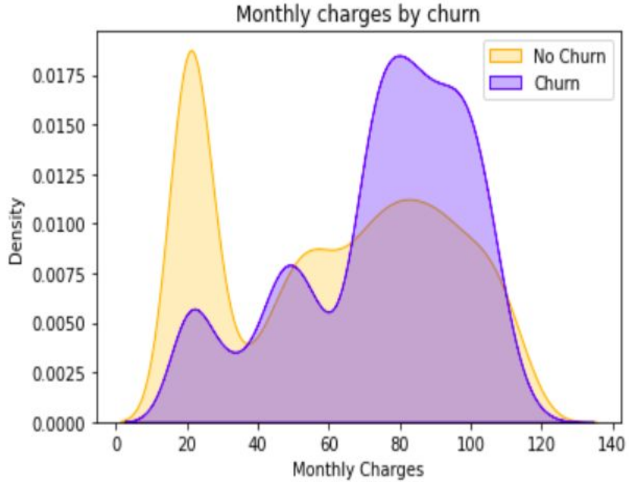
- Electronic check medium are the highest churners
- Contract Type - Monthly customers are more likely to churn because of no contract terms, as they are free to go customers.
- No Online security, No Tech Support category are high churners
- Non senior Citizens are high churners

EDA – continued

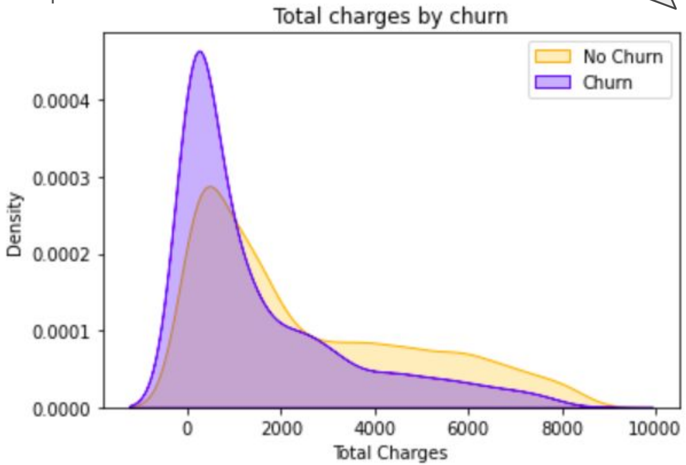


Monthly Charges vs Total Charges

Insight:
Total Charges increase as Monthly Charges increase

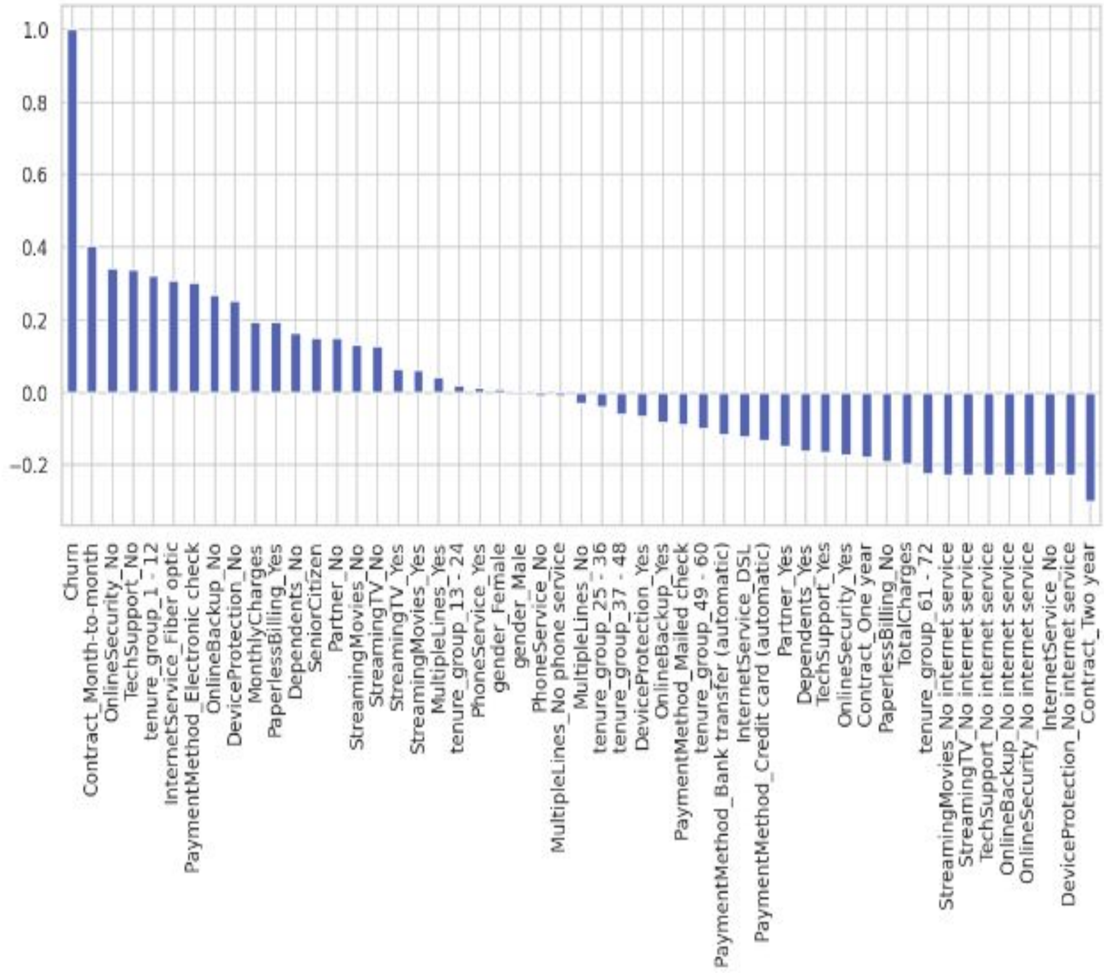


Insight:
Lower total charges led to higher churn.



Higher Monthly Charge, **Lower** tenure and **Lower** Total Charge are linked to **High** Churn.

Insight: Churn is high when monthly charges around USD70 to USD100.



CORRELATION OF ALL VARIABLES WITH "CHURN"

Insights:

HIGH Churn seen in case of Month to month contracts, No online security, No Tech support, First year of subscription and Fibre Optics Internet

LOW Churn is seens in case of Long term contracts, Subscriptions without internet service and The customers engaged for 5+ years

Factors like Gender, Availability of PhoneService and number of multiple lines have alomost **NO impact on Churn**

MODEL BUILDING

<table>
<tr><th>Items</th><th>Results/ FIndings/ Action</th></tr>
<tr><td>Drop “Churn” column from x</td><td>“Churn” is the target variable which is excluded for model building</td></tr>
<tr><td>Perform train-test-split with 80% as train data and 20% as test data</td><td><code>x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.2,random_state=0)</code></td></tr>
<tr><td>Perform Decision Tree Classifier</td><td>Classification Report

<table>
<tr><th></th><th>Precision</th><th>Recall</th><th>F1-score</th></tr>
<tr><td>0</td><td>0.83</td><td>0.78</td><td>0.81</td></tr>
<tr><td>1</td><td>0.47</td><td>0.55</td><td>0.51</td></tr>
<tr><td>Accuracy</td><td></td><td></td><td>0.72</td></tr>
</table>
As this is an imbalance dataset, we shouldn’t consider Accuracy as our metrics to measure model.

To check recall, precision & F1-score for the minority class. It shown that these scores are too low for Class 1 (ie. churned customers).</td></tr>
</table>

MODEL BUILDING - continued

Items	Results/ FIndings/ Action
Perform **SMOTEENN** (UpSampling + ENN) : combination of SMOTE and Edited Nearest Neighbor (ENN)	Use this method on **train data only** to solve imbalance dataset. SMOTEENN combine oversampling and undersampling techniques into a hybrid strategy. 1. (**Start of SMOTE**) Choose random data from the minority class. 2. Calculate the distance between the random data and its k nearest neighbors. 3. Multiply the difference with a random number between 0 and 1, then add the result to the minority class as a synthetic sample. 4. Repeat step number 2–3 until the desired proportion of minority class is met. (**End of SMOTE**) 5. (**Start of ENN**) Determine K, as the number of nearest neighbors. If not determined, then K=3. 6. Find the K-nearest neighbor of the observation among the other observations in the dataset, then return the majority class from the K-nearest neighbor. 7. If the class of the observation and the majority class from the observation's K-nearest neighbor is different, then the observation and its K-nearest neighbor are deleted from the dataset. 8. Repeat step 2 and 3 until the desired proportion of each class is fulfilled. (**End of ENN**)

MODEL BUILDING – continued

Decision Tree Classifier (before SMOTENN)

	Precision	Recall	F1-score
0	0.83	0.78	0.81
1	0.47	0.55	0.51
Accuracy			0.72

Decision Tree Classifier (after SMOTENN)

	Precision	Recall	F1-score
0	0.86	0.76	0.81
1	0.49	0.65	0.56
Accuracy			0.73

Random Forest Classifier

	Precision	Recall	F1-score
0	0.91	0.71	0.8
1	0.5	0.81	0.62
Accuracy			0.74

Bagging Classifier

	Precision	Recall	F1-score
0	0.91	0.71	0.8
1	0.5	0.81	0.62
Accuracy			0.74

Logistic Regression

	Precision	Recall	F1-score
0	0.88	0.8	0.84
1	0.55	0.68	0.61
Accuracy			**0.77**

- After we resolved imbalanced dataset, we can refer to Accuracy.
- Logistic Regression show the highest accuracy at 77%.

RECOMMENDATIONS

- To assess and identify customers who are about to churn and to further strengthen customers’ loyalty
- To strategize new marketing initiatives from predictive model experience
- To drive analytics led campaigns
- To study on competitors’ SWOT analysis and find any gap that company can fill up / achieve better than competitors

Thank You!