Springboard—Data Science Career Track
Capstone Project 2
Predicting Customer Churn
By Sheema Murugesh Babu
September 2019

1 Introduction

Churn quantifies the number of customers who have unsubscribed or canceled their service contract. Customers turning their back to a service or product are no fun for any business. It is very expensive to win them back once lost. Keeping the right customers can be quite valuable for a company. Not only because customer acquisition is much more expensive, but as more and more business models are shifting towards subscription plans, a customer can be worth thousands of dollars in future. Reducing churn ultimately leads to a sustainable growing business.

1.1 Problem Statement

The challenge here is to build a model that identifies customers with the intention to leave a service in the near future. The data contains Demographic information like gender, age range, and whether they have partners and dependents, contains customer account information, services that each customer has signed up for and lastly customers who left within the last month – the column is called Churn.
My solution looks into building machine learning models to predict customer churn. Given the dataset, I could estimate whether a customer may or may not be unsubscribed to a service.

2 Approach

2.1 Data Acquisition and Wrangling

About the Dataset

The data was found from the “Telcom Customer Churn” dataset provided by Kaggle’s website. https://www.kaggle.com/blastchar/telco-customer-churn and was saved into the local as ‘**Telco-Customer-Churn.csv**’.

This is the current data they have available:

Variable	**Definition**
customerID	Customer ID
gender	Sex of User
SeniorCitizen	Senior Citizen
Partner	Partner
Dependents	Dependents
tenure	Tenure
PhoneService	Phone Service
MultipleLines	Multiple Lines

InternetService	Internet Service
OnlineSecurity	Online Security
OnlineBackup	Online Backup
DeviceProtection	Device Protection
TechSupport	Tech Support
StreamingTV	Streaming TV
StreamingMovies	Streaming Movies
Contract	Contract
PaperlessBilling	Paperless Billing
PaymentMethod	Payment Method
MonthlyCharges	Monthly Charges
TotalCharges	Total Charges
Churn	Churn

Steps on Data-Wrangling

First, I imported the required packages that I would need for Data Wrangling. Then, I used pandas read_csv module to import the csv file (i.e. 'Telco-Customer-Churn.csv') which was saved in my local repository. Upon executing the read_csv function using the info() method, the csv file has 7043 entries and 21 columns with different formats of data. Also, it seems that there are no missing values. Digging deeper into it using the describe() method along with the head(). Next step would be to find the distinct values in the dataset.

Second, I used nunique() method to find the count of distinct values in the dataset. All the columns except tenure, MonthlyCharges and TotalCharges have numerical values and all the other columns would be considered as categorical. Note: 'tenure' column is also categorical only its values are numerical.

Third step would be to perform some Data Manupulation like, changing the type 'TotalCharges' column from object type to integer type. When I tried to do that it gave me an error. There were actually some blank spaces in that column. I replaced the blank spaces with the mean. Also, changed the data in few other columns like 'OnlineSecurity', 'OnlineBackup', 'DeviceProtection', 'TechSupport', 'StreamingTV', 'StreamingMovies', 'MultipleLines' and 'SeniorCitizen'. Added a column called 'Tenure_group' to group in data from the tenure column. Dropped the irrelevant column, separated the churn, non-churn customers, categorical and numerical columns.

2.2 Storytelling and Inferential Statistics

After I wrangled and cleaned the dataset, I started to explore the data in detail. To put great visualizations, I used 'Plotly' library. Let's see some of the visualizations here.

1. Plot for 'Percentage of churn in the dataset':

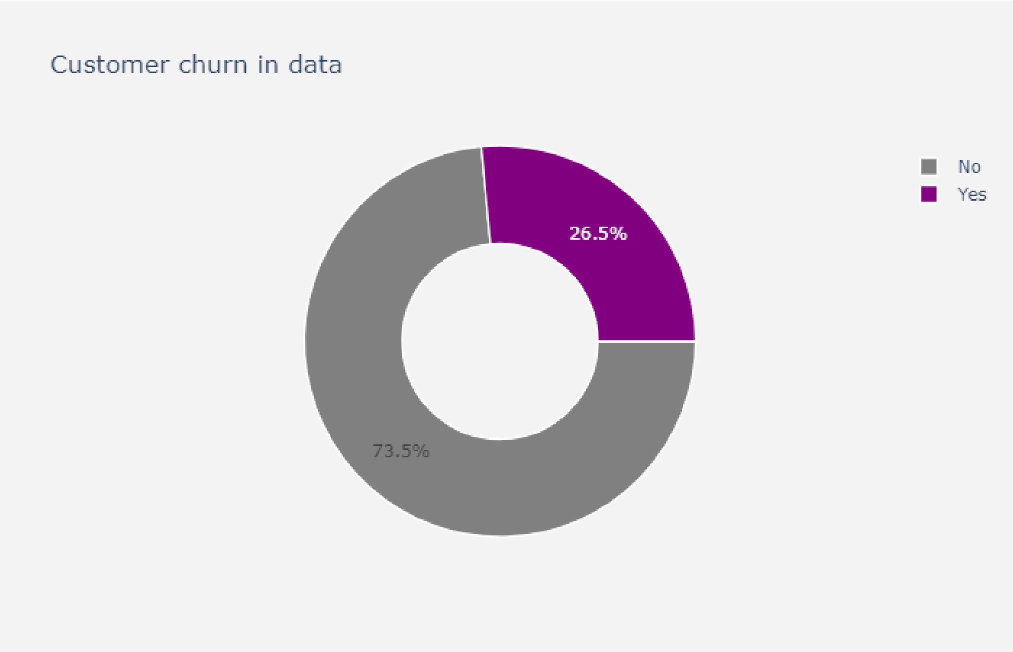

Figure: Customer Churn in data

From the above pie chart, we can see that only 26.6% of the data represents churn customers and the majority are the non-churn customers. This also shows that we might be dealing with a class imbalance problem as there are more non-churned customers than the churned ones.

2. Plots for some of the categorical columns:

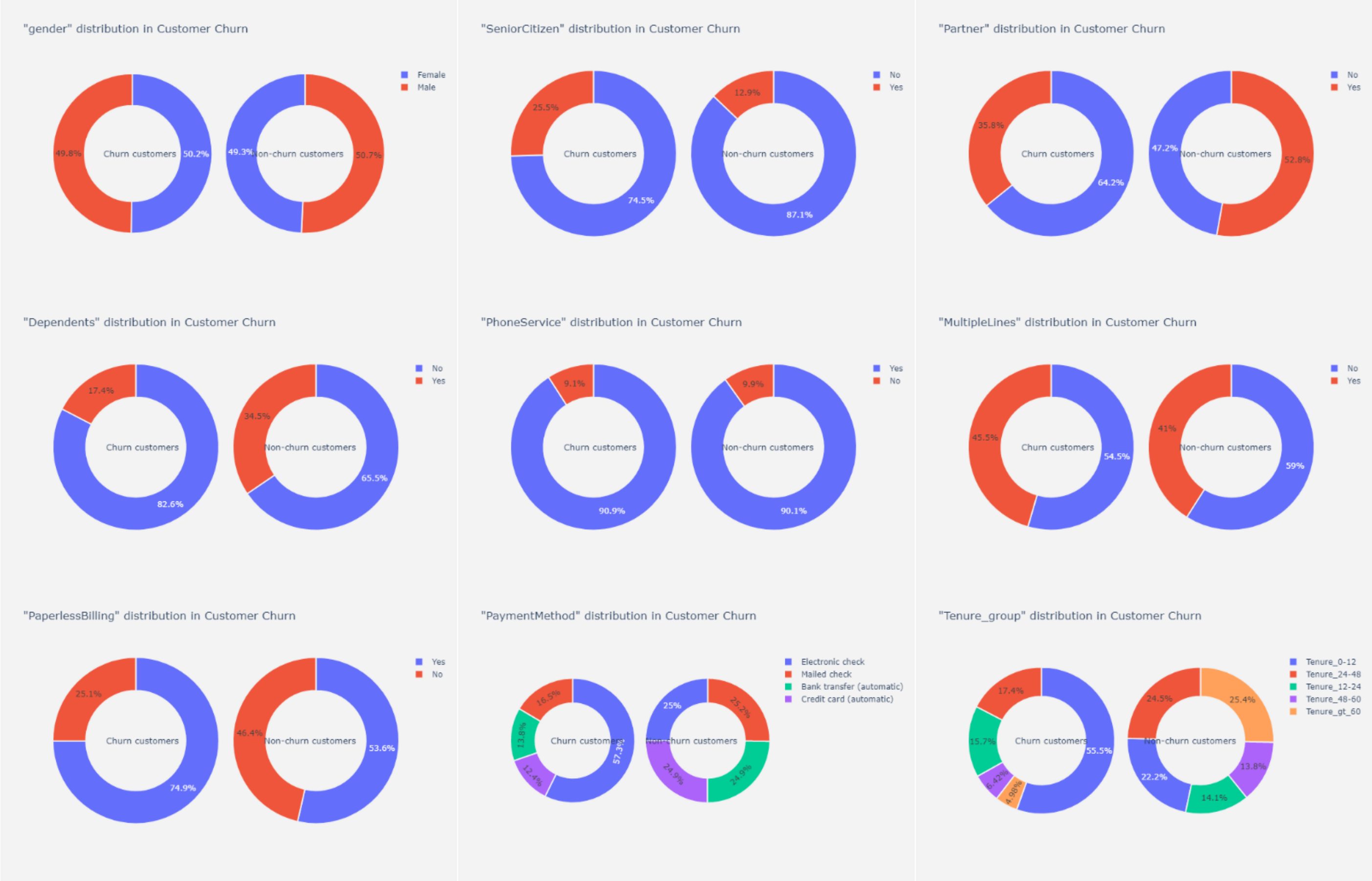

Figure: Pie Plots for few columns.

3. Histograms for all the numerical columns:

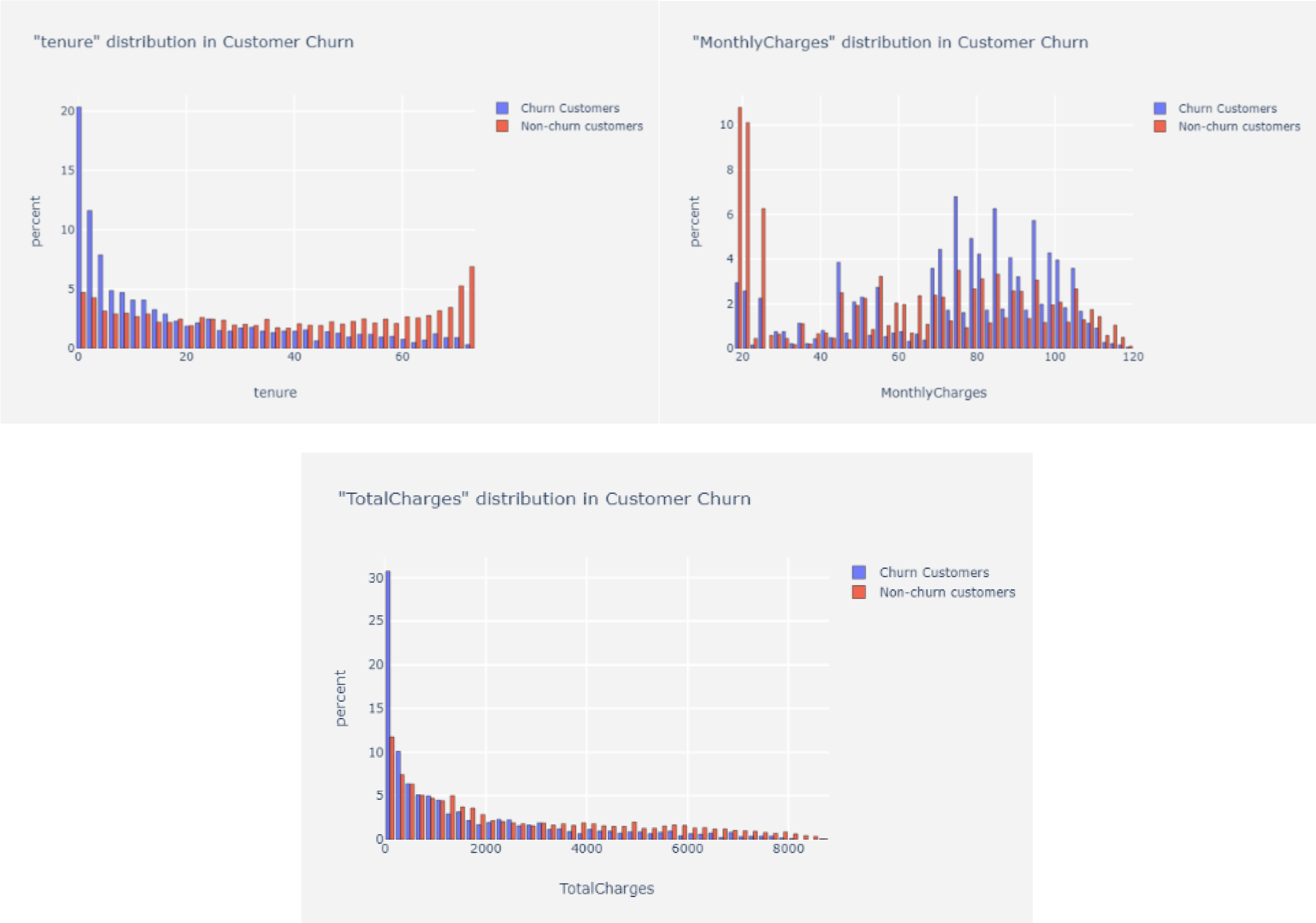

Figure: Histogram plot for the numerical columns.

4. Customer Churn in Tenure groups:

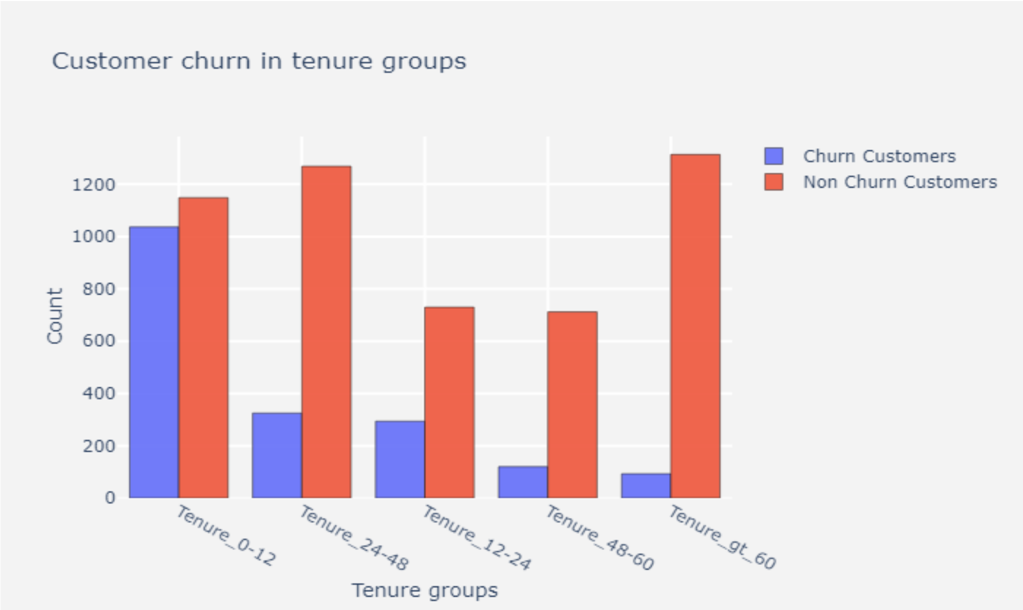

Figure: Customer Churn in Tenure groups

5. Monthly Charges and Total Charges by Tenure group and Churn group:

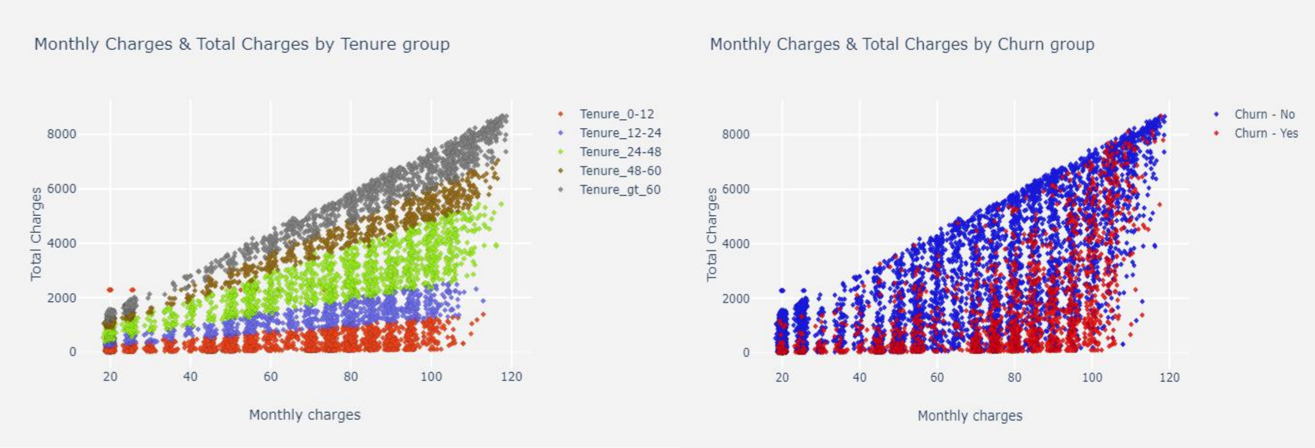

Figure: Monthly Charges and Total Charges by Tenure group and Churn group

6. Average charges by Tenure Groups:

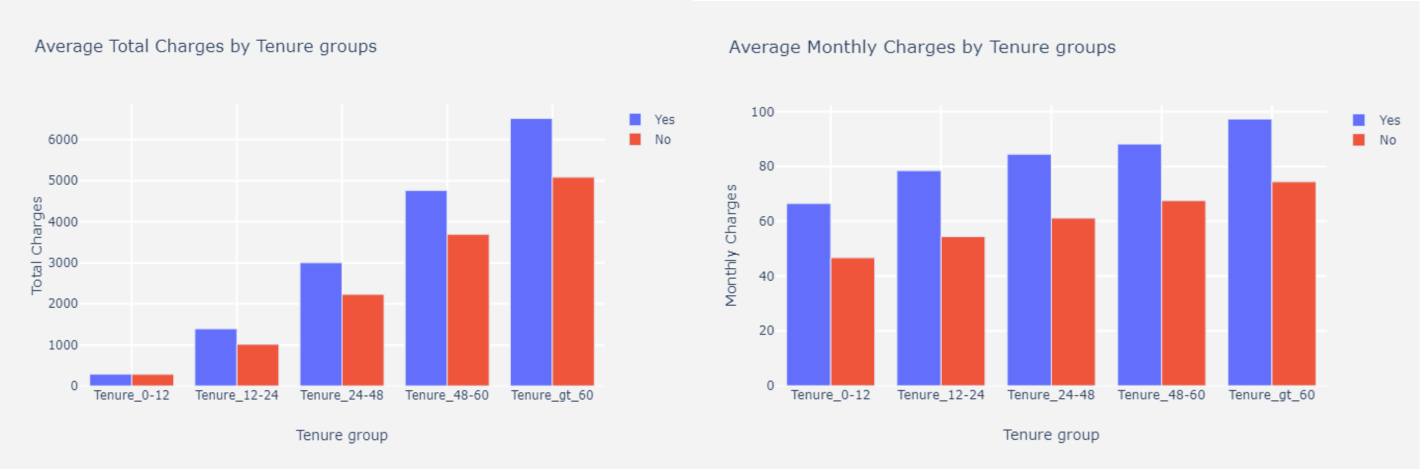

Figure: Average charges by Tenure Groups

The next step would be to provide steps on Inferential Statistics. I have performed Hypothesis testing using the Chi – Square test.

Hypothesis Test:

Chi-Square Test

The Pearson’s Chi-Squared test, or just Chi-Squared test is a statistical hypothesis test that assumes (the null hypothesis) that the observed frequencies for a categorical variable match the expected frequencies for the categorical variable. The test calculates a statistic that has a chi-squared distribution, named for the Greek capital letter Chi (X) pronounced “ki” as in kite.

The Chi-Squared test uses something called a contingency table, by first calculating the expected frequencies for the groups, then determining whether the division of the groups, called the observed frequencies, matches the expected frequencies. The result of the test is a test statistic that has a chi-squared distribution and can be interpreted to reject or fail to reject the assumption or null hypothesis that the observed and expected frequencies are the same.

Imported the ‘chi2’ and the ‘chi2_contingency’ from the scipy.stats library and wrote a function for Contingency table for all the categorical columns and the numerical columns. The function returned a Contingency table for each categorical/numerical column against the target variable i.e. Churn, degrees of freedom, expected values, the test statistics such as Probability, Critical values, Chi-square statistic, significance and the p-value. If the p-value < 0.05, it would mean there is a relationship between the 2 categorical variables.

Categorical Columns		**Numerical Columns**	
Column Name	**Significance**	**Column Name**	**Significance**
SeniorCitizen	YES	tenure	YES
Partner	YES	MonthlyCharges	YES
Dependents	YES	TotalCharges	NO
MultipleLines	YES		
InternetService	YES		
OnlineSecurity	YES		
OnlineBackup	YES		
DeviceProtection	YES		
TechSupport	YES		
StreamingTV	YES		
StreamingMovies	YES		
Contract	YES		
PaperlessBilling	YES		
PaymentMethod	YES		
Tenure_group	YES		
Gender	NO		
PhoneService	NO		

Table shows the significance of all the Categorical and Numerical Columns.