Customer Churn Prediction in Telecom using Machine Learning

DISSERTATION

Submitted in partial fulfillment of the requirements of

MTech Software Engineering Degree Programme

By

Samarth Malhotra

2018AP04535

Under the supervision of

Mayank Jain

Principal Software Developer

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE

PILANI (RAJASTHAN)

May, 2021

DSE CL ZG628T DISSERTATION

Customer Churn Prediction in Telecom using Machine Learning

Submitted in partial fulfillment of the requirements of the

M. Tech. Data Science and Engineering Degree Programme

By

Samarth Malhotra

2018AP04535

Under the supervision of

Mayank Jain

Principal Software Developer

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE

PILANI (RAJASTHAN)

May, 2021

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI

CERTIFICATE

This is to certify that the Dissertation entitled **"Customer Churn Prediction in Telecom using Machine Learning"** and submitted by Mr. **Samarth Malhotra** ID No. **2018AP04535** in partial fulfillment of the requirements of DSE CL ZG628T Dissertation, embodies the work done by him/her under my supervision.

Signature of the Supervisor

Name: Mayank Jain

Designation: Principal Software Developer

Place: _______________

Date: ________________

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI

I SEMESTER 2020-21

DSECLZG628T DISSERTATION

Dissertation Outline

BITS ID No. 2018AP04535 **Name of Student:** Samarth Malhotra

Name of Supervisor: Mayank Jain

Designation of Supervisor: Principal Software Developer

Qualification and Experience: B.E. Computer Science

E- mail ID of Supervisor: Mayank.nirmal@gmail.com

Topic of Dissertation: Customer Churn in Telecom using Machine Learning

Name of First Examiner: Prof. M. J. Shankar Raman

Designation of First Examiner: ____

Qualification and Experience: ____

E- mail ID of First Examiner: ____

Name of Second Examiner: ____

Designation of Second Examiner: ____

Qualification and Experience: ____

E- mail ID of Second Examiner: ____

(Signature of Student)
Date:

(Signature of Supervisor)
Date:

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI

Work Integrated Learning Programmes Division

I SEMESTER 2020-21

DSE CL ZG628T **DISSERTATION**

(EC-2 Mid-Semester Progress Evaluation Sheet)

Scheduled Month August:

NAME OF THE STUDENT : Samarth Malhotra

ID NO. : 2018AP04535

Email Address : malhotra.samarth@hotmail.com

NAME OF SUPERVISOR : Mayank Jain

PROJECT TITLE : Customer Churn in Telecom using Machine Learning

Evaluation Details

EC No.	**Component**	**Weightage**	**Comments** (Technical Quality, Originality, Approach, Progress, Business value)	**Marks Awarded**
1	Dissertation Outline	10%		
2.	Mid-Sem Progress Seminar Viva Work Progress	 10% 5% 15%		

	Supervisor	Additional Examiner
Name	Mayank Jain	
Qualification	BE Computer Science	
Designation & Address	Principal Software Developer E804, Palsh Society, Wakad, Pune, Maharashtra, 411057	
Email Address	Mayank.nirmal@gmail.com	
Signature		
Date		

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI

Work Integrated Learning Programmes Division

I SEMESTER 2020-21

Supervisor's Rating of the Technical Quality of this Dissertation Outline

EXCELLENT / GOOD / FAIR/ POOR (Please specify): _________________________

Supervisor's suggestions and remarks about the outline (if applicable).

Date: ____________________ (Signature of Supervisor)

Name of the supervisor: Mayank Jain

Email Id of Supervisor: Mayank.nirmal@gmail.com

Mob # of supervisor: +91 8805413880

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI

SECOND SEMESTER 2020-21

DSE CL ZG628T **DISSERTATION**

Dissertation Title : Customer Churn Prediction in Telecom using Machine Learning

Name of Supervisor : Mayank Jain

Name of Student : Samarth Malhotra

ID No. of Student : 2018AP04535

Abstract

Customer churn is the likelihood of a customer to leave a brand, stop using its services and switching over to other providers. It is a major challenge in businesses with subscription-based model and has direct impact on the revenue of the company, especially in the telecom field. The cost of churn includes both the loss of revenue and the marketing costs involved in replacing those customers with new ones, therefore, predicting and preventing customer churn has a potential revenue source therefore the telecom companies must make an effort to retain their customers.

In the face of stiff competition in the market, the customers have very wide choice and often they switch over from one product to another, there is always a search for better options.

There can be several factors responsible for customer churn, including:

- The availability of quality services
- Low-cost alternatives
- Better features and content
- Customer experience
- Availability of self-service options
- Easy access to the maintenance staff
- Network coverage

Customer churn prediction modelling aims to understand the customer's behavior and attributes (gender, age, dependents, financial status), also the likelihood to switching of the brand, possible reasons and the remedial measures to retain the customer.

With a better understanding and an insight into potential customers leaving the brand in the volatile market condition, the brand can take a suitable action after the analysis which will lead to most retention impact on the customer.

Contents

List of Symbols and Abbreviations

ML : Machine Learning
EDA : Exploratory Data Analysis
RFC : Random Forest Classifier
FE : Feature Engineering

List of Tables

List of Figures

Chapter 1

Introduction

Customer churn is the likelihood of a customer to leave a brand, stop using its services and switching over to other providers. It is a major challenge in businesses with subscription-based model and has direct impact on the revenue of the company, especially in the telecom field. The cost of churn includes both the loss of revenue and the marketing costs involved in replacing those customers with new ones, therefore, predicting and preventing customer churn has a potential revenue source therefore the telecom companies must make an effort to retain their customers.

In the face of stiff competition in the market, the customers have very wide choice and often they switch over from one product to another, there is always a search for better options.

There can be several factors responsible for customer churn, including:

- The availability of quality services
- Low-cost alternatives
- Better features and content
- Customer experience
- Availability of self-service options
- Easy access to the maintenance staff
- Network coverage

The above-mentioned list is not exhaustive, the inventory could vary depending upon service provider and would require domain knowledge.

Customer churn prediction modelling aims to understand the customer's behavior and attributes (gender, age, dependents, financial status), also the likelihood to switching of the brand, possible reasons and the remedial measures to retain the customer.

With a better understanding and an insight into potential customers leaving the brand in the volatile market condition, the brand can take a suitable action after the analysis which will lead to most retention impact on the customer.

Objective

The main objective of the present project is to design a churn prediction model that could help telecom operators to foresee the customer behavior and accurately predict the customers who are likely to churn.

In order to know the customer behavior, the relevant historical data will be used and as the current research in the field confirms machine learning could be efficiently applied to predict the customer churn and take the retention measures.

Principal objectives:

a) Create visualizations to showcase how each feature is affecting the target class
b) Create multiple machine learning models to predict the target variable and evaluate them with multiple metrics (AUC Score, precision, recall, f1 score)
c) Identify the features which are important for the chosen model

Uniqueness of the project

In many organizations the customer churn is reactive in the sense that when customer calls to end the subscription only then offers are rolled out to retain the customers.

In this project we are aiming to make this process proactive by actively predicting unhappy customers in advance and making necessary adjustments to retain them.

Benefit to the organization

Oracle provides end to end cloud solutions to telecommunication providers, it spans everything from capturing the network calling data to billing and processing payments to generating audit reports.

The project will directly benefit Oracle in providing up to date information about customer churn to its client i.e., telecom operators and operators in turn will ensure that the customer churn could be prevented in time by opting for retention strategies which will have direct impact on their revenue.

Chapter 2

Data Acquisition

Here we are using sample dataset provided by IBM community, the dataset contains information about a fictional telco company that provided home phone and internet services to 7043 customers in California in Q3. It indicates which customers have left, stayed, or signed up for their service. Multiple important demographics are included for each customer, as well as Customer Lifetime Value (CLTV) index.

The dataset includes information about:

- Customers who left within the last month – the column is called Churn
- Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device
- protection, tech support, and streaming TV and movies
- Customer account information – how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges
- Demographic info about customers – gender, and if they have partners and dependents

Data Preprocessing

Handling Missing data

The dataset has some missing values for features Total Charges and Churn Reason, for records where Tenure Months is 0 there is no value for Total Charges, also for feature Churn Reason for all the data points where Churn Label is No. i.e. The customers who have not left the company there will be no Churn Reason. Before imputing any value for Total Charges, we are changing the data type of the feature from object to float.

Table 1 : Features with missing values

Name	dtype	Missing	Unique
Total Charges	object	11	6531
Churn Reason	object	5174	20

We can impute the values for Churn Reason for the missing values as Not Available as the churn didn't happen also upon closer observation, we can observe that there is a strong correlation between the numerical features Tenure Months, Monthly Charges and Total Charges. If we calculate correlation coefficient for (Tenure Months x Monthly Charges) and Total Charges it is 0.9995605537972277 and therefore we can impute the missing values of the feature Total Charges with (Tenure Months x Monthly Charges).

Removing Unnecessary Features

We are removing unnecessary features such as latitude longitude, zip code, country, state and churn score as the data is only for United States of America also its for state of California, we have removed latitude, longitude and information as we will be using City to identify the location. We are removing the churn score as its not part of actual data but generated by IBM SPSS tool.

Data Exploration

We are firstly performing Univariate Analysis

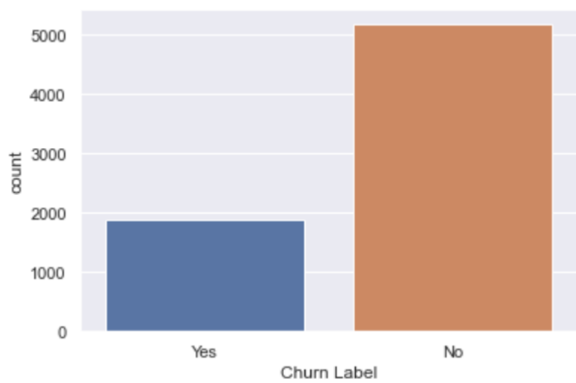

Figure 1: Distribution of Target Variable

Churn Label is the target feature and as we can observe from above plot the dataset is unbalanced, there are more data points for customers who have not churned while performing model creation and any analysis we have to keep this in mind.

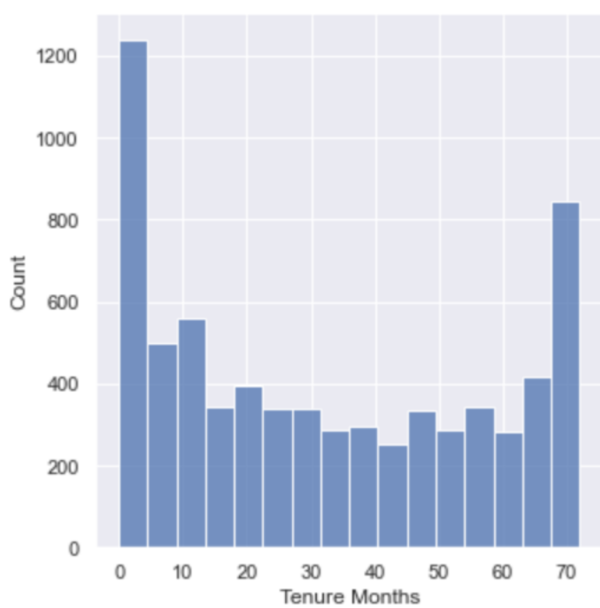

Figure 2: Distribution of Tenure in Months

From the above numerical feature Tenue Months, we can observe that it is a bimodal distribution, which means there are two different kinds among customers and we can find out what services are kept by those who stay more than 70 months.

Secondly, we will perform bivariate analysis of numerical features, we will see how different numerical features are distributed in terms of target variable.

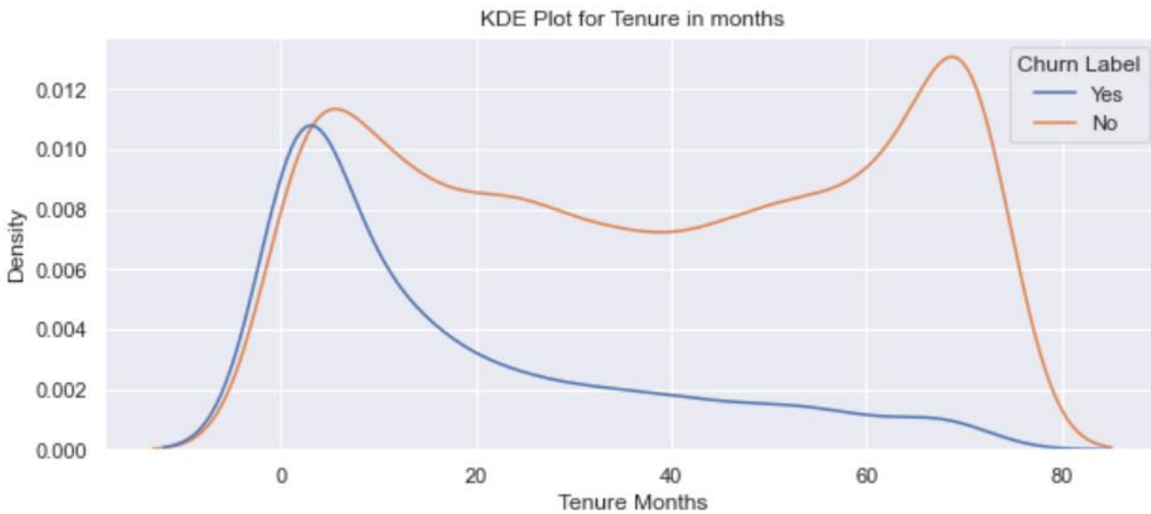

Figure 3: Kernel Density Estimation of Tenure Months

From above kernel density estimation plot we can see the probability density function of numerical feature Tenure Months; it can be observed that customers who have recently joined are more likely to churn.

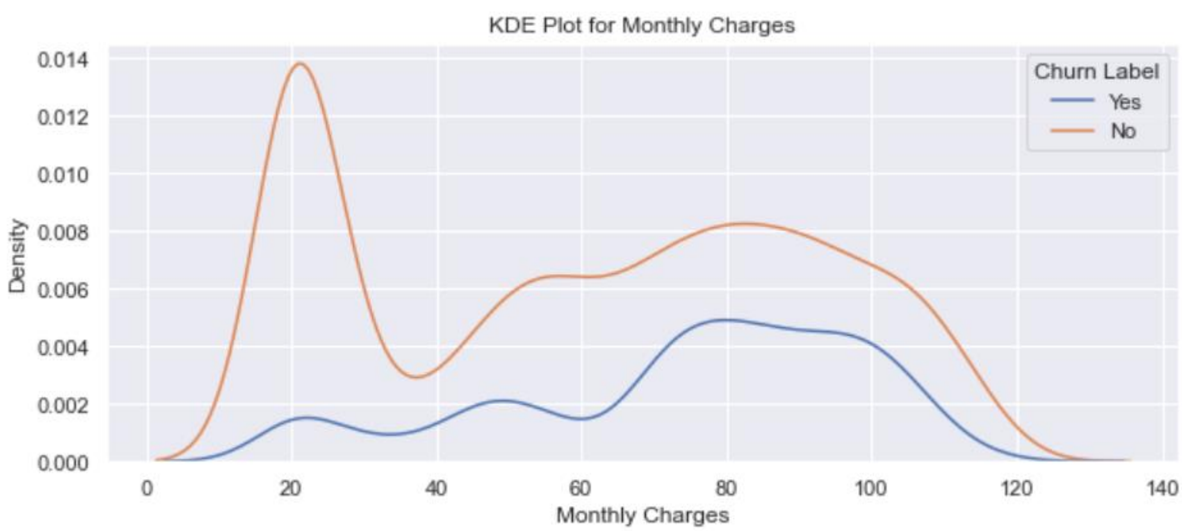

Figure 4: Kernel Density Estimation of Monthly Charges

From above kernel density estimation of numerical feature Monthly charges, we can observe that customers with higher monthly charges are more likely to churn than those with lower monthly charges.

Now, we will see how categorical features are distributed in terms of target feature

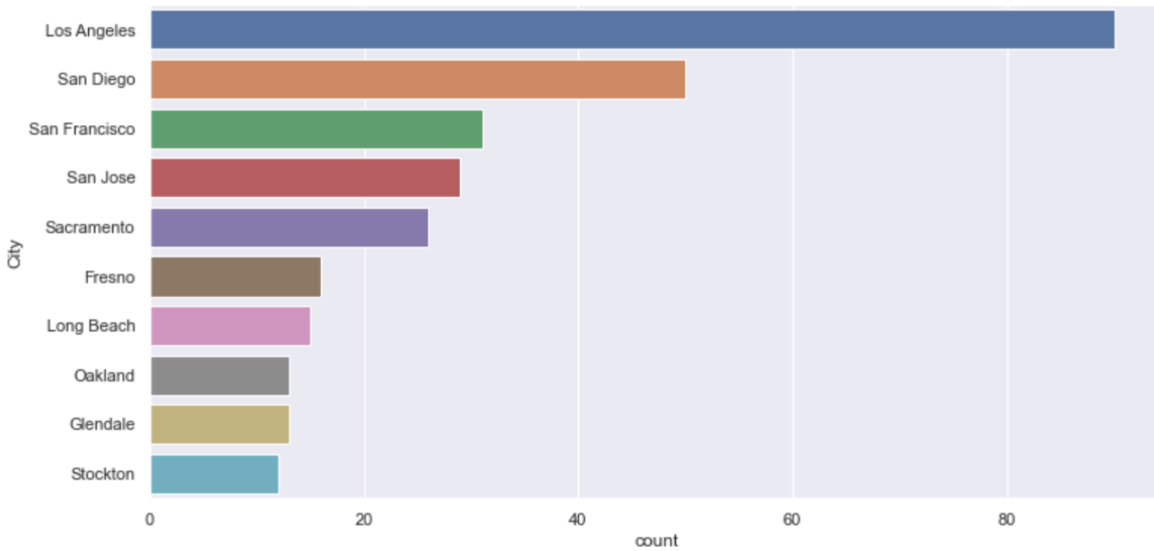

Figure 5: Distribution of customer who churned for different cities

From the above plot it is easily visible that city Los Angeles, San Diego, San Francisco, San Jose, Sacramento accounts for most churn customers and therefore we need to investigate further why so many customers are leaving from these particular locations.

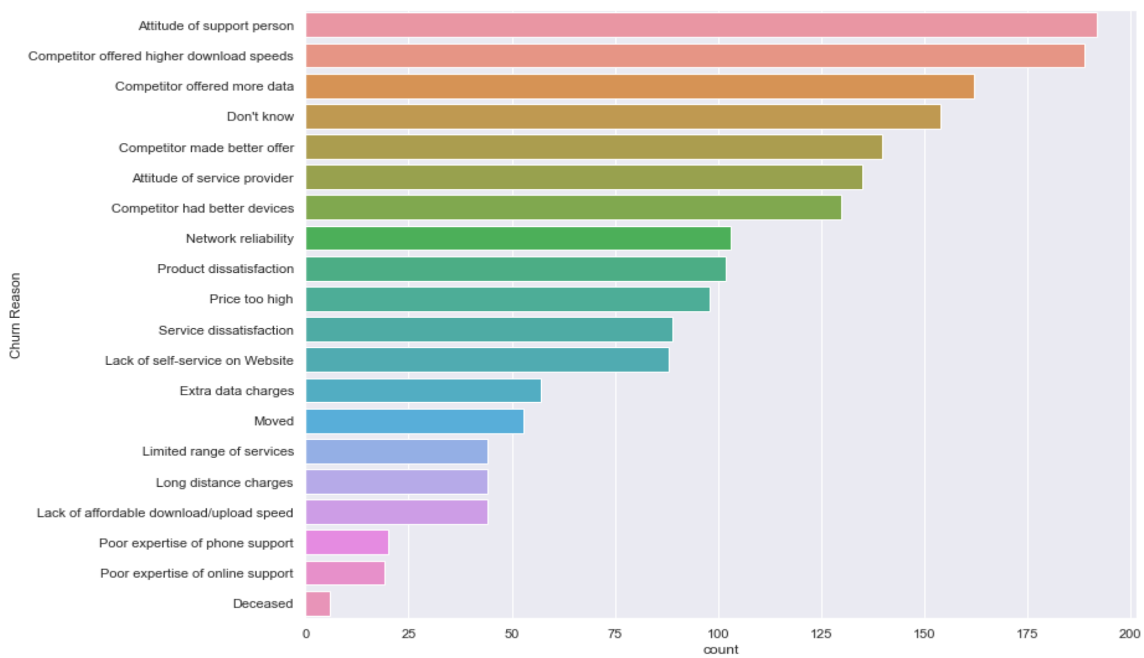

Figure 6: Distribution of customer who churned for different reasons

From the above distribution plot of Churn Reason, it can be observed that the reason of highest churn among customers is dissatisfaction from support services and internet data and speed and therefore remedial actions can be taken to improve support service and internet speed also we can come up with better internet plans.

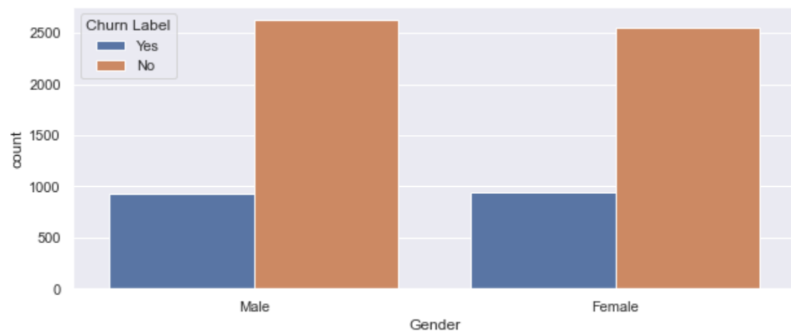

Figure 7: Distribution of customer for gender

From the above figure we can see that feature gender has no influence on customer churn

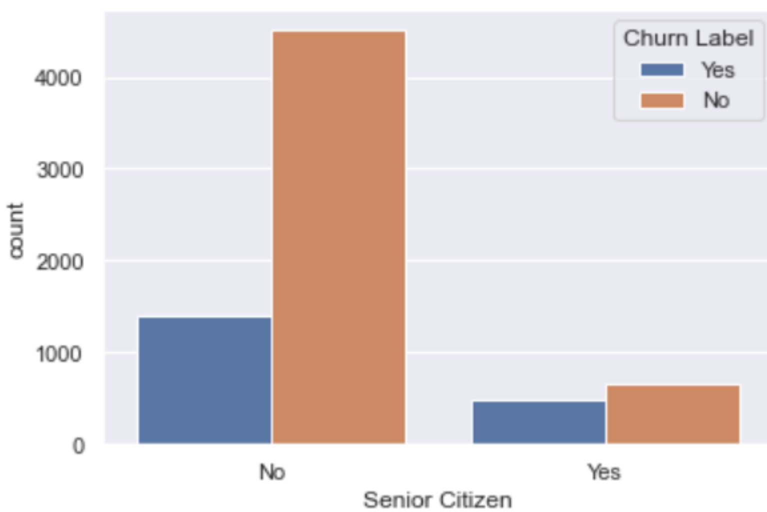

Figure 8: Distribution of customers for senior citizens

From the above distribution plot, we can observe that, even though there are only 16 % senior citizen among total customers but the churn rate among senior citizens in 41.6 % compared to 23.6 % in younger customers.

Similarly,

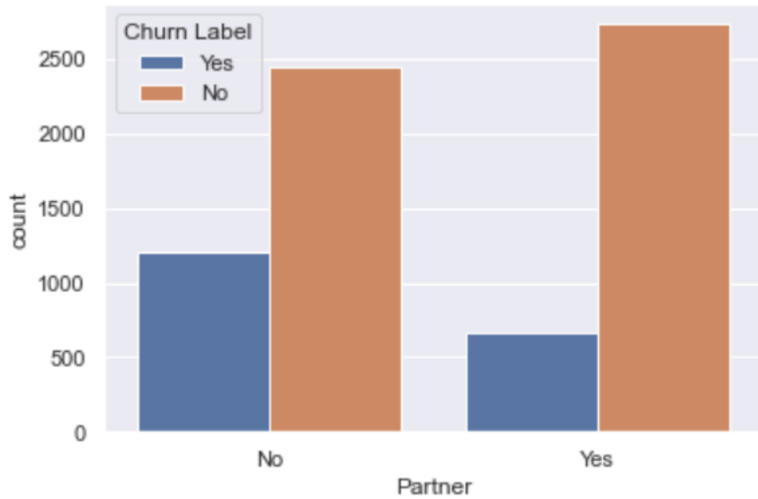

Figure 9: Distribution of customers for customer with partners

From above distribution plot it is evident that customer without partners is more likely to churn in comparison to customers with partners.

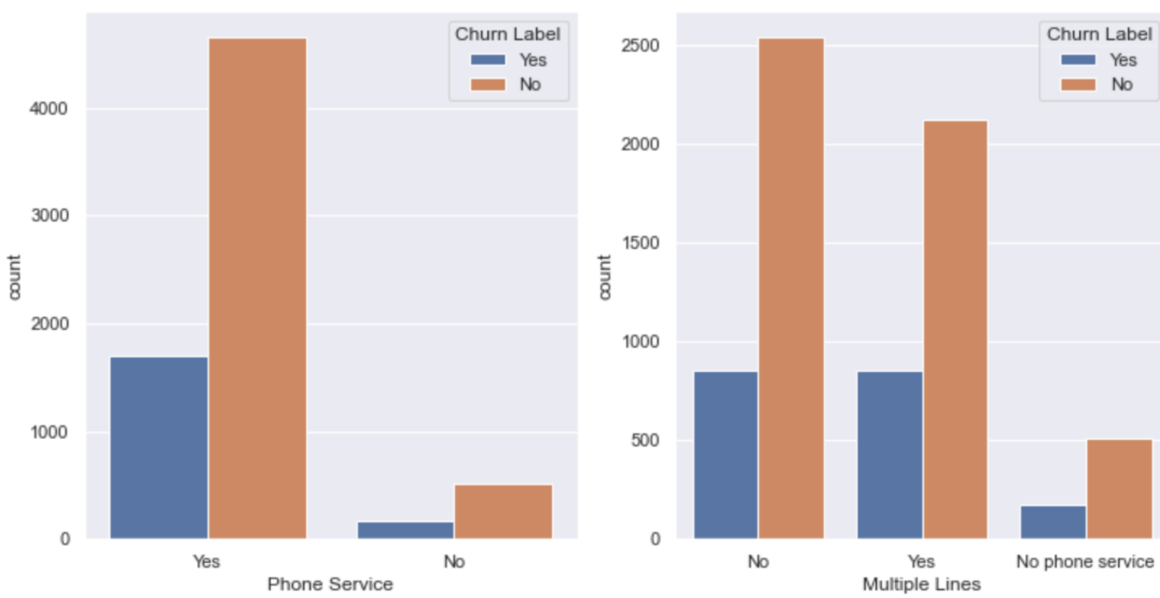

Figure 10: Distribution of customers for Phone Service and Multiple Lines

From the above distribution plots, it is evident that customer with no phone service is less and customers with multiple lines have slightly higher churn.

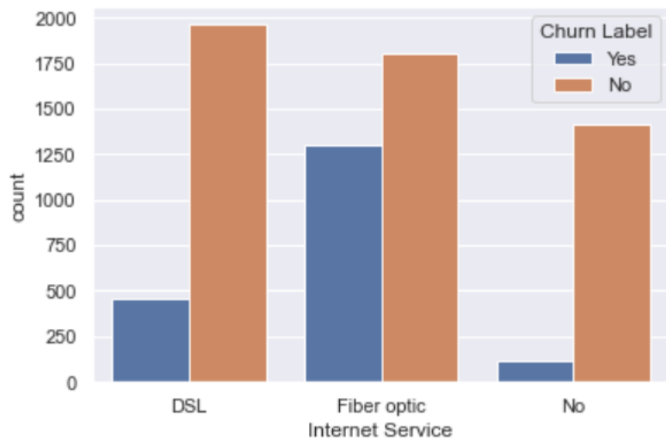

Figure 11: Distribution of customers for fiber optic service

It can be observed from above plot that customers without internet have very low churn, also customers with fiber optic cable are more likely to churn than customers with DSL connection.

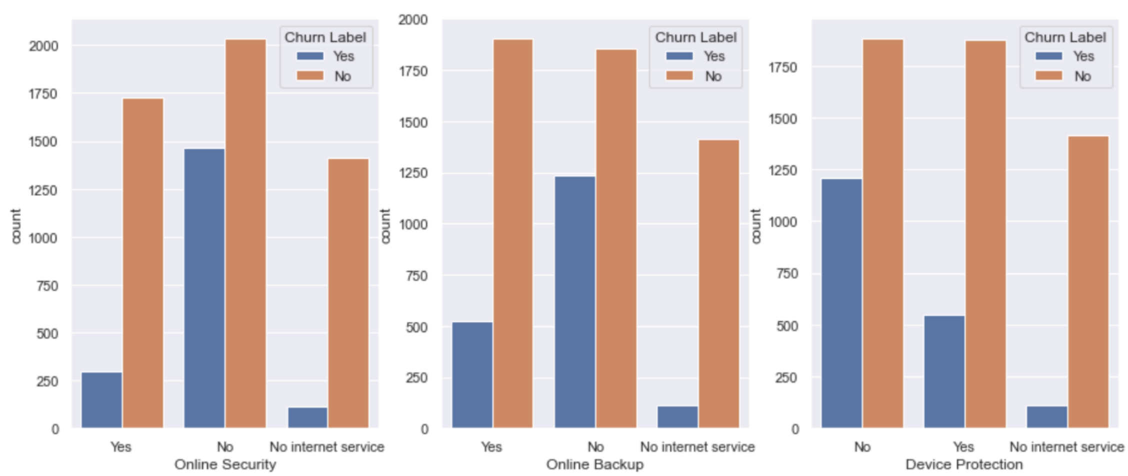

Figure 12: Distribution of customers for online security, online backup and device protection

From above plot it is evident that

- customers without internet are less likely to churn
- Customer with online security is less likely to churn
- Customers with online backup are less likely to churn
- Also, customers with device protection are less likely to churn

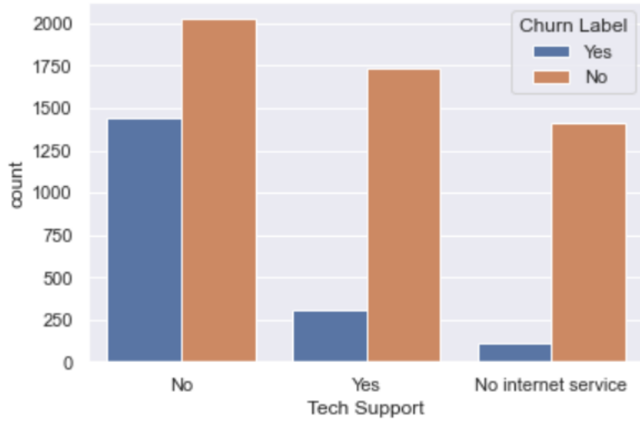

Figure 13: Distribution of customers for tech support

From the above plot we can see that customers with tech support are less likely to churn

Similarly,

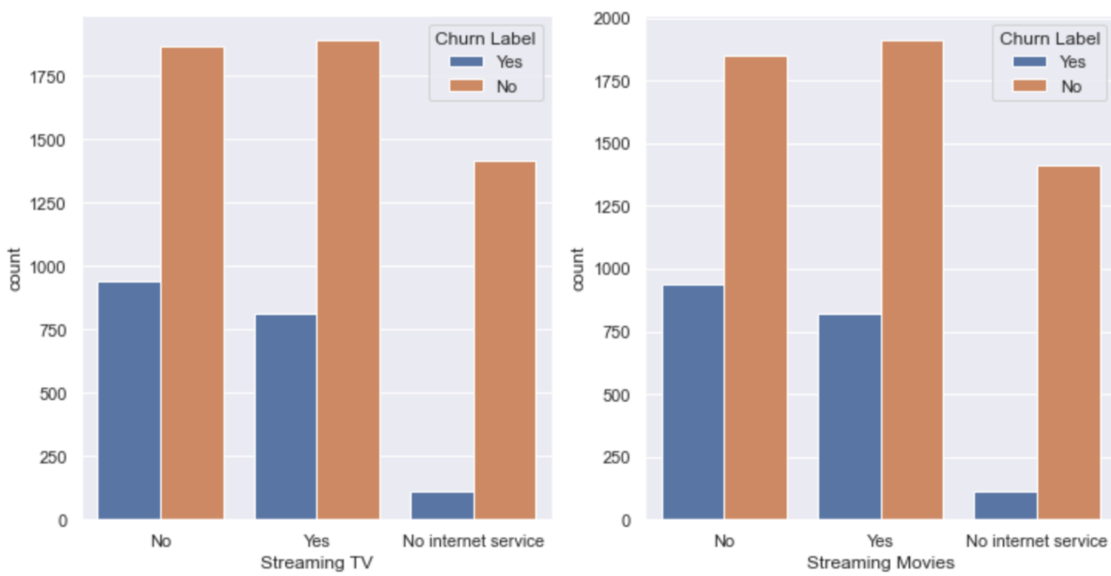

Figure 14: Distribution of customers for steaming tv and streaming movies

From above plot it is very evident that customers with Streaming TV and Streaming Movies are less likely to churn.

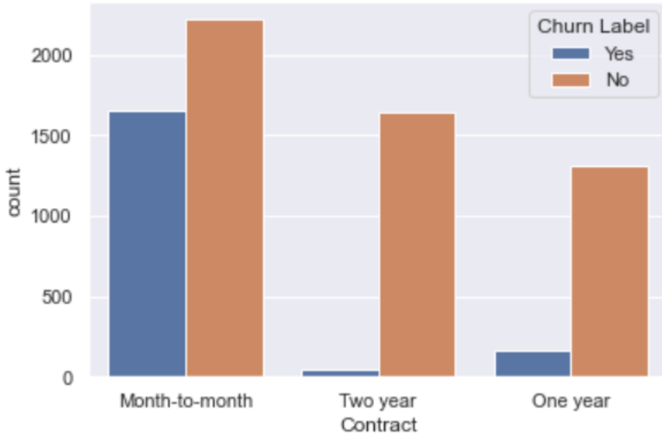

Figure 15: Distribution of customers for contract

From the above graph we can observe that customers one-year and two-year contracts are less likely to churn.

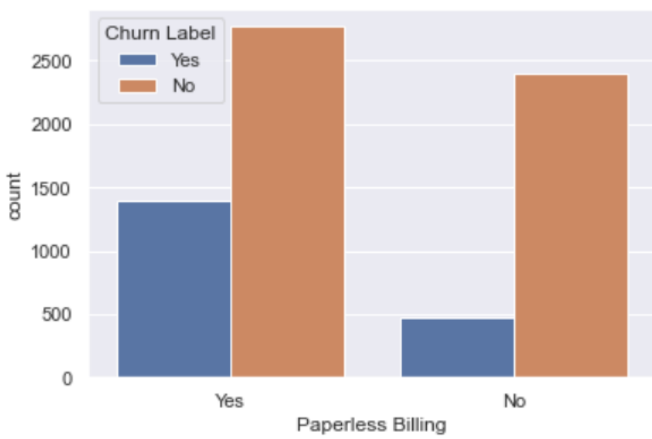

Figure 16: Distribution of customers for paperless billing

From above plot we can observe that customers with paperless billing are more likely to churn as compared to other customers.

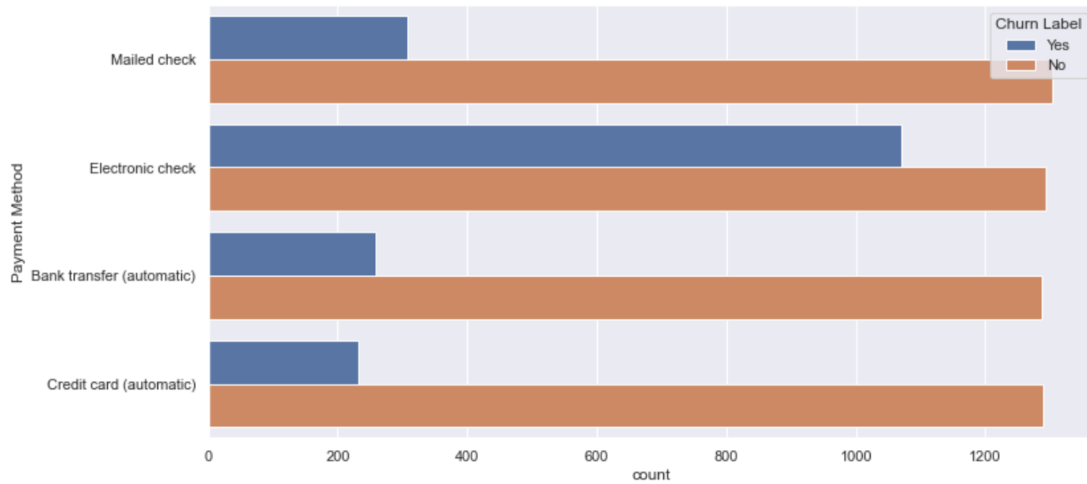

Figure 17: Distribution of customers for payment method

From the above plot we can observe that customers with electronic check are more likely to churn compared to other payment methods.
Now we will check for correlation between features, checking correlations is an important part of the exploratory data analysis process. This analysis is one of the methods used to decide which features affect the target variable the most, and in turn, get used in predicting this target variable. In other words, it's a commonly-used method for feature selection in machine learning. Here we have divided the dataset into two sets one for numerical features and one for categorical features. For numerical features Tenure Months, Monthly Charges, Total Charges and CLTV we have created correlation matric and plotted heatmap.

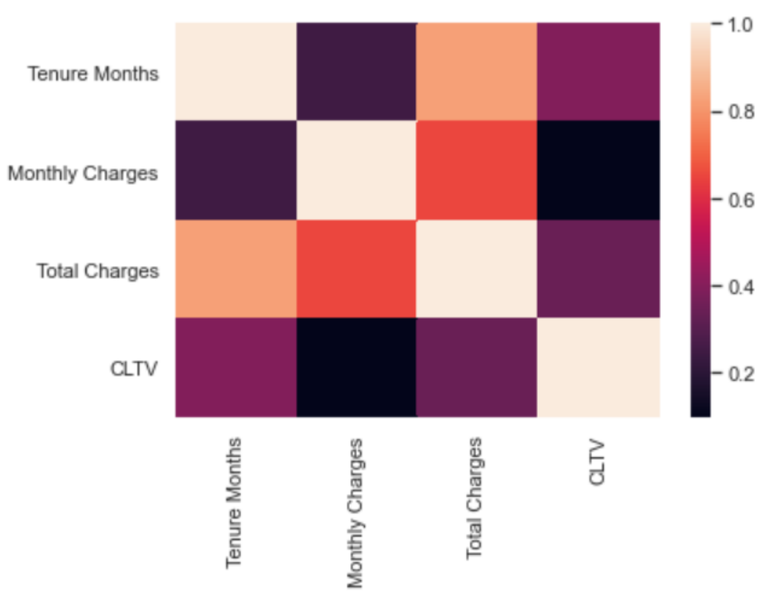

Figure 18: Correlation between numeric features

From the above heatmap it is very clear that Tenure Months and Total charges are highly correlated, similarly for categorical features we have created a separate dataset and created correlation matrix.

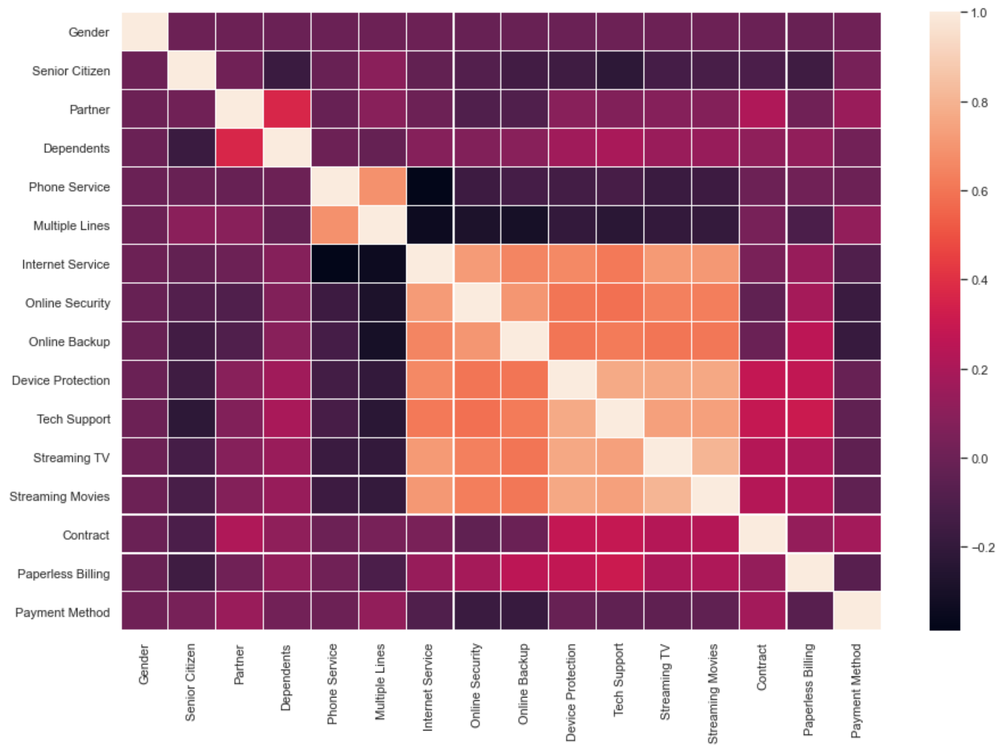

Figure 19: Correlation between categorical features

Form the above heatmap we can observe that Phone Service and Multiple Lines are correlated, similarly, internet Service, Online Security, Device Protection, Tech Support, Streaming TV and Streaming Movies are correlated.

Now we try to find the feature importance using **Random Forest Classifier here**, we have one-hot encoded the categorical features and dropped the features which are not required for analysis and also performed hyper parameter tuning by applying model in various settings.

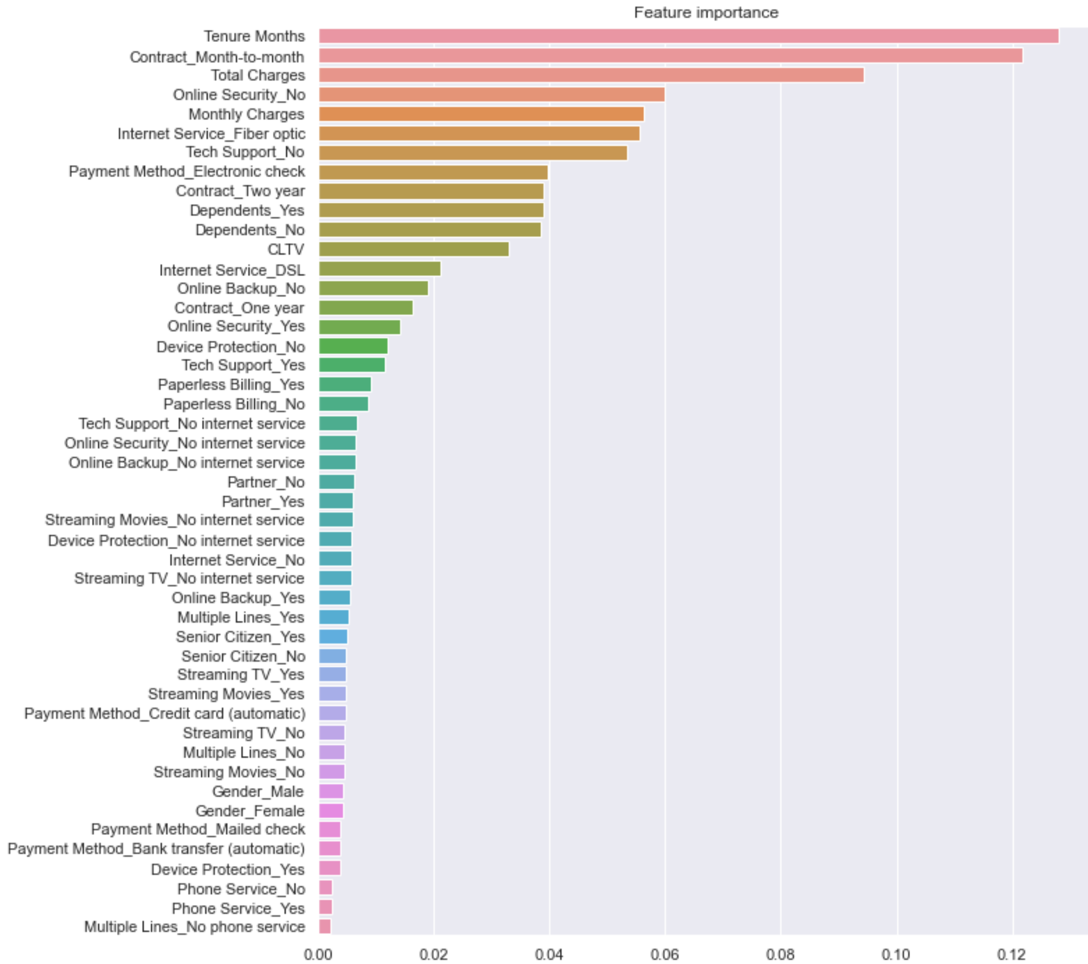

Figure 20: Feature Importance using Random Forest Classifier

As we have also observed in exploratory data analysis numerical features Tenure Months, Monthly Charges, and Total Charges are very important features to predict customer churn also categorical features Contract Mouth-to-mouth which are highly likely to churn. These results obtained are in line with the results obtained in exploratory data analysis.

Chapter 3

Scope of work

- Acquisition of data
- Preprocessing of data
- Exploration of data
- Implementation of model to predict churn using python
- Visualizations
- Report creation

Chapter 4

Resources needed for the project

- Telecom customer data
- Visualization libraries
- Windows machine
- Python libraries