Indian Institute of Technology, Jodhpur

Project Report
Machine Learning (CS7620)

Project: System for forecasting customer attrition

Submitted By:
Gaurav Choudhary (M22CS055)
Prashant Gautam (M22CS057)
Deepak Kumar (M22MA004)

Submitted To:
Deepak Mishra
Assistant Professor
CSE Department
IIT Jodhpur

Table of Contents

1.Abstract:

Customer attrition rate or churn rate is a critical measure of quality of service provided by service or subscription based businesses, marketing and financial institutions, Telecommunication industry etc.

Churn rate, sometimes called attrition rate, is the rate at which customers stop doing business with a company over a given period of time.

Businesses can learn about weak points in their products or pricing strategies, operational problems, consumer preferences, and expectations by identifying customers who are dissatisfied with the solutions they are receiving, thereby proactively reducing reasons for churn.

2.Introduction:

Churn may also apply to the number of subscribers who cancel or don't renew a subscription. It is an effective measure for subscription-based businesses.

Churn rate can be as percentage of number of subscribers lost during a time period divided by number of subscribers at the beginning of the time period.

This project is aimed at developing a customer churn prediction system which will forecast if an existing customer is likely to cancel or continue the subscription to a service. The ML Algorithm that will be used in the project is Logistic Regression. We shall also use other models like Decision Tree Classifier then compare performance metrics of each model and select the best one.

The dataset has been sourced from IBM Developer Platform which is a sample data of a non-existing telecommunication company. It has 20 features for 7043 customers. It includes a target label classifying whether or not a customer left within the previous month and some dependent variables like customer information, location, services they are subscribed to, etc.

3.Objective:

The primary objective is to develop a customer attrition prediction system using a suitable machine learning algorithm on telecom service data to help determine whether a customer will churn or not.

4.About the Dataset:

The dataset has been sourced from IBM Developer Platform which is a sample data of a non-existing telecommunication company. It has 21 features for 7043 customers. It includes target label classifying whether or not a customer left within the previous month and some dependent variables like customer information, location, services they are subscribed to, etc.

The feature set can be divided into 3 parts:

- Customer Account Information:
  - Tenure, Contract , PaperlessBilling , PaymentMethod , MonthlyCharges , TotalCharges
- Services that the customers has signed up for:
  - PhoneService , MultipleLines , InternetService , OnlineSecurity , OnlineBackup , DeviceProtection , TechSupport , StreamingTV , StreamingMovies,

- Customer Personal Information:
    - Gender , SeniorCitizen , Partner , Dependents

The features can also be defined as:

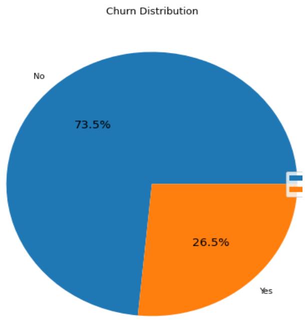
- Features with discrete values
    - CustomerID: Unique number for each Customer used for identification purposes.
    - Gender: Showing if a customer is a male or a female
    - SeniorCitizen: Shows that a customer is a senior citizen or not (1, 0)
    - Partner: Showing whether or not a customer has a partner. (Yes, No)
    - Dependent: Showing whether or not a customer has dependents. (Yes, No)
    - PhoneService: Showing whether or not a customer has a phone service. (Yes, No)
    - MultipeLines: Whether or not a customer has multiple lines. (Yes, No, No phone service)
    - InternetService: Showing the type of internet service a customer has opted. (DSL, Fiber optic, No)
    - OnlineSecurity: Showing whether or not a customer has online security. (Yes, No, No internet service)
    - OnlineBackup: Whether or not a customer has an online backup. (Yes, No, No internet service)
    - DeviceProtection: Whether or not a customer has device protection. (Yes, No, No internet service)
    - TechSupport: Whether or not a customer has tech support. (Yes, No, No internet service)
    - StreamingTV: Whether or not a customer has streaming TV service. (Yes, No, No internet service)
    - StreamingMovies: Whether or not a customer has streaming movies. (Yes, No, No internet service)
    - Contract: The contract term of the customer (Month-to-month, One year, Two years)
    - PaperlessBilling: The duration of the customer's contract. (Month-to-month, One year, Two years)
    - PaymentMethod: The payment method of the customer. (Electronic check, Mailed check, Bank transfer, Credit card)
- Features with continuous values (numeric values)
    - Tenure: Number of months the customer has stayed with the company.
    - MonthlyCharges: The monthly amount charged to the customer.
    - TotalCharges: The total amount charged to the customer.
- Target Feature
    - Churn: Whether or not a customer had churned. (Yes or No)

5.Methodology

We need to know about the dataset and the value distributions in order to work with it. So we'll need to critically explore and analyze the data to discover patterns and visualize how the features interact with each other and the target. We'll use libraries like plotly.express, matplotlib and seaborn for visualization. Analyzing the dataset will also help us find some missing values.

5.1.Exploratory Data Analysis

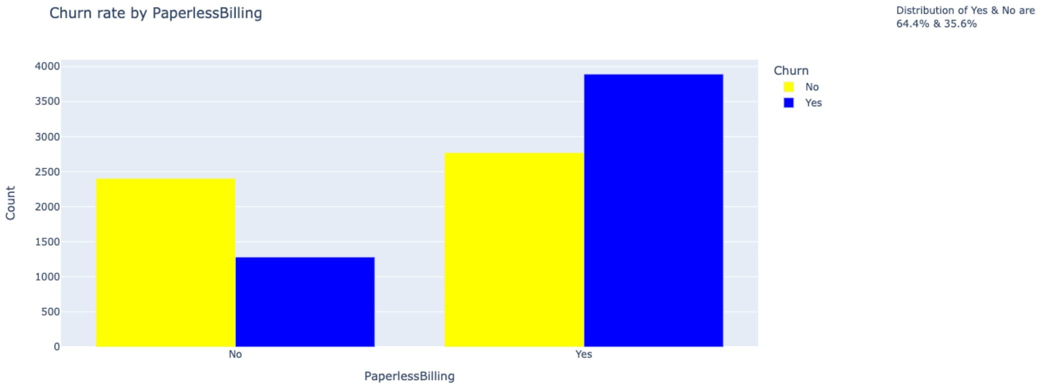
Exploring the target feature.



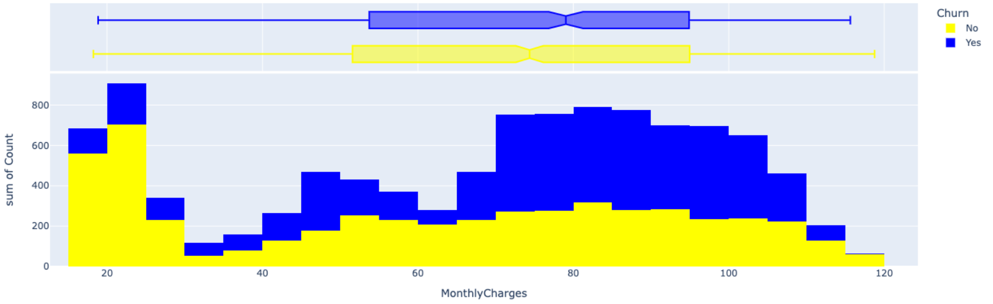
The target feature vector Churn has two values Yes/No which is unevenly distributed. We have balanced the distribution of 'Yes' and 'No' by downsampling the count of 'Yes', making it equal to the count of 'No'. The final distribution of 'Yes' and 'No' were 50% each.

Analyzing the features showing customer account information.

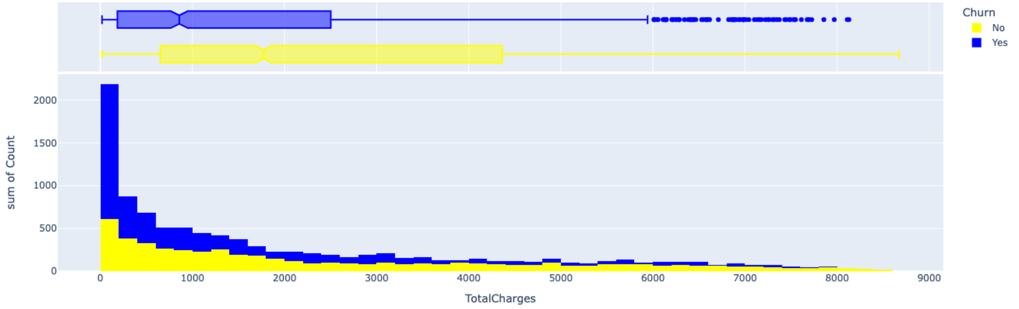
We've use bar graphs to show distribution in features with discrete values and histograms to show distributions in features with continuous values.



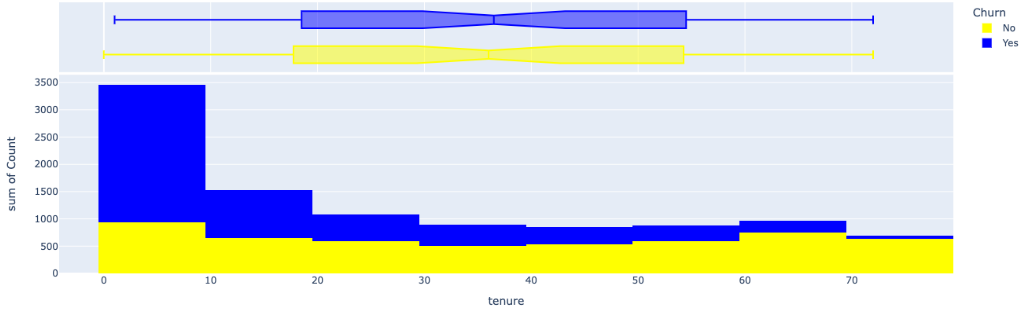
Churn rate frequency to MonthlyCharges distribution



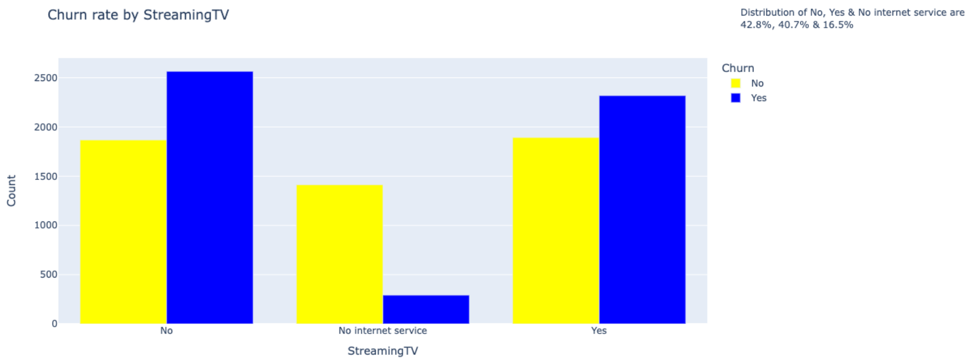
Churn rate frequency to TotalCharges distribution



Churn rate frequency to tenure distribution

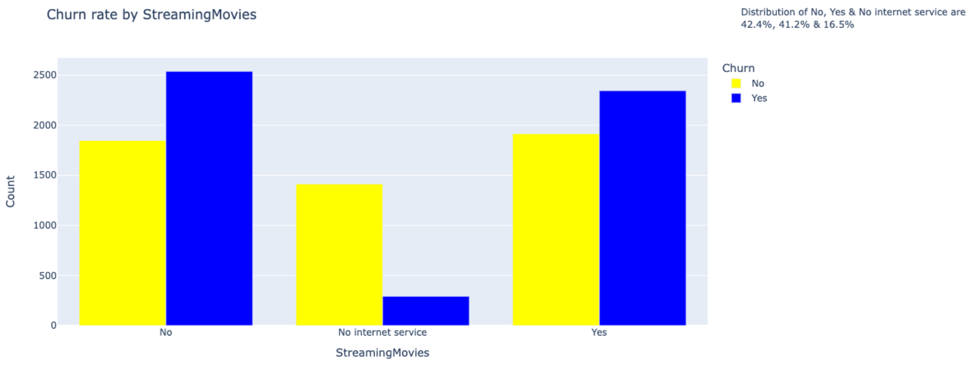


Exploring the features showing the services a customer has signed up for.



Churn rate by StreamingTV

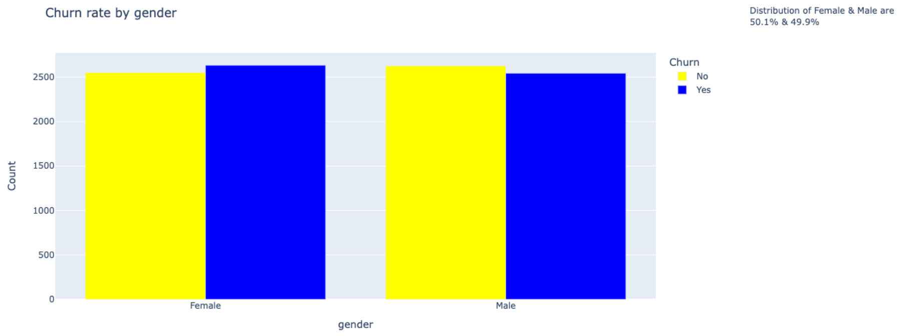
Distribution of No, Yes & No internet service are 42.8%, 40.7% & 16.5%

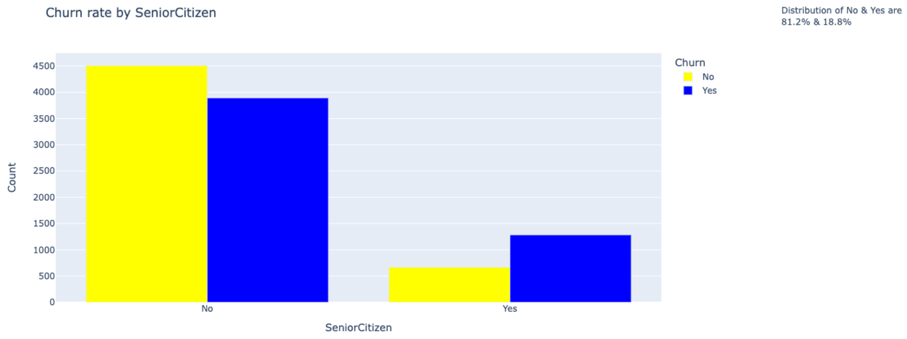


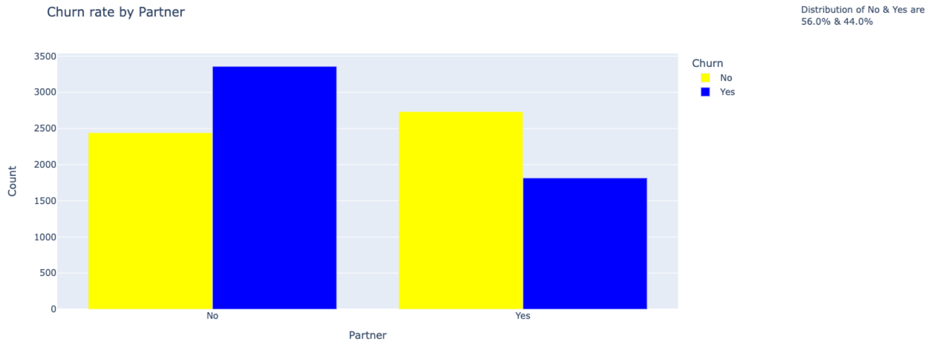
Churn rate by StreamingMovies

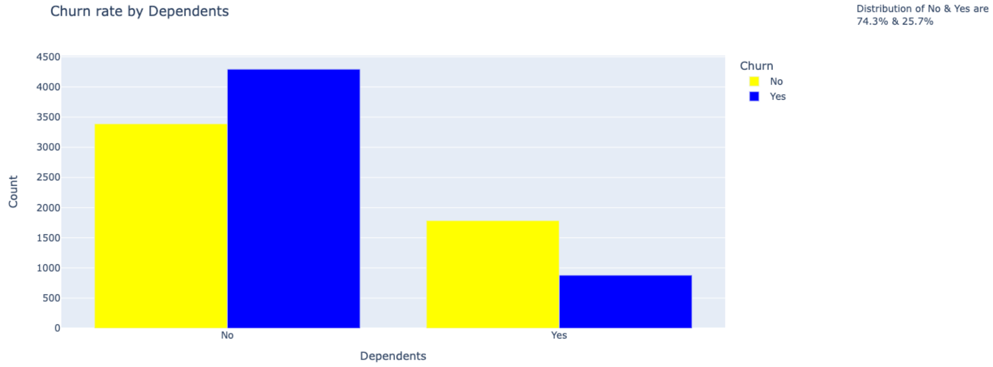
Distribution of No, Yes & No internet service are 42.4%, 41.2% & 16.5%

Exploring the features showing customer's personal information.









5.2.Data Preprocessing:

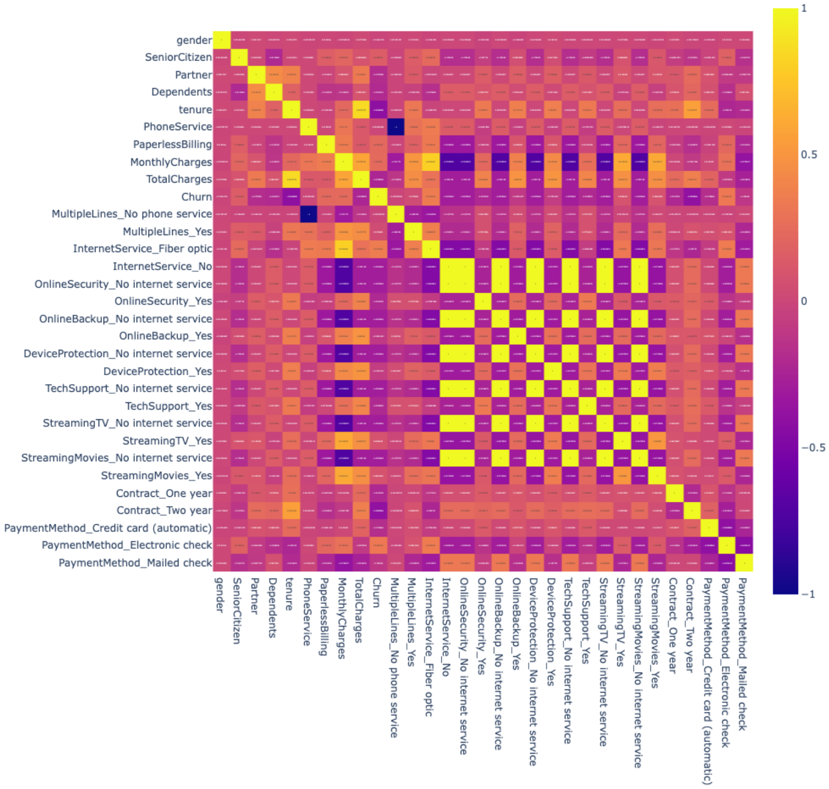
We can make data easier to use and analyze by preprocessing it. The accuracy of a model is improved by removing data inconsistencies or duplicates. Preprocessing the data makes sure there aren't any incorrect or missing values brought on by bugs or human error.

Finding Correlated Features:

Correlation measures linear dependency between any two features. Features with high correlation are more dependent and dropping any one of them will do no harm to the model learning and inference.

Having correlated features does not bring any supplementary information. It will increase the learning time complexity and also the risk of errors.

We’ve shown the correlation between every feature using a heat map.



From the heatmap, we conclude that we can drop the features like MultipleLines, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, and StreamingMovies.

Features scaling:

The range of all features need to be normalized so that each feature contributes approximately proportionately to the final value. We'll use MinMaxScaler from sklearn.preprocessing to normalize features like Tenure, MonthlyCharge and TotalCharge.

Train-Test Split:

We've split the dataset into a training set and test set. The training set is 70% of the original dataset sampled randomly and the test set is the rest 30%.

5.4.Model training and performance evaluation:

Model Selection:

As this is a binary classification problem, We've selected classification models that learn on labeled training dataset. This type of learning is called supervised learning. The models used were Logistic Regression, Naive Bayes, SVM, Decision Trees and Random Forest. We've evaluated the performance of these models based on metrics like accuracy scores, precision, f1 scores.

6. Results:

Insights gained on Analysis of the data:

Features showing customer personal information:

Females have somewhat higher churn. Churn is higher among younger customers, customers without partners, and customers without dependents. Senior citizens without partners or dependents are most likely to churn.

Features showing the services that the customers have signed up for:

A customer cannot have several lines if they do not have phone service. 90.3% of consumers have phone service and have a higher turnover rate. Customers that use fiber optic internet are more prone to churn. Customers may leave because fiber optic service is significantly more expensive than DSL. Customers who have access to OnlineSecurity, OnlineBackup, DeviceProtection, and TechSupport are less likely to leave. Because it is fairly allocated across yes and no options, streaming service is not predictive of churn.

Features showing the customer account information:

The higher the churn rate, the shorter the contract. Those with longer-term intentions encounter extra challenges when canceling early. This clearly illustrates why firms want to maintain long-term connections with their customers.

Customers that have chosen paperless billing have a greater churn rate. Paperless billing is used by approximately 59.2% of clients. Customers that pay using electronic checks are more prone to churn and this payment method is more common than others.

Majority of clients have only been using the telecom provider for a short period of time (0-9 months). Additionally, the first few months have the highest turnover rates (0-9 months). Within the first 30 months, 75% of clients who decide to leave the telco firm do so. Customers with higher monthly rates have a higher rate of churn. This implies that offers such as discounts and specials may persuade customers to stay.

Evaluation metrics obtained for all the models used on the test dataset.

	Accuracy	Precision	Recall	F1 score
Logistic Regression	0.7665056360708534	0.7418590882178804	0.8125810635538262	0.7660563149633365
Naive Bayes	0.7536231884057971	0.7254788160185722	0.8106355382619974	0.752881143986151
SVM	0.7838969404186795	0.7596899224806202	0.8261997405966277	0.7835515087538022
Decision Tree	0.8608695652173913	0.8153409090909091	0.930609597924773	0.860247303059138
Random Forest	0.8898550724637682	0.8500583430571762	0.9448767833981842	0.8895576491496541

Random Forest gives the best metrics. Its accuracy is about 89%. We’ve selected the Random Forest classifier as our final model.

7.Conclusion:

Churn rate is an important metric for subscription-based businesses. Identifying unhappy consumers can assist managers in identifying product or pricing plan flaws, operational challenges, and customer preferences and expectations. Knowing all of this makes it easier to implement proactive churn-reduction strategies.

Businesses can learn about weak points in their products or pricing strategies, operational problems, consumer preferences, and expectations by identifying customers who are dissatisfied with the solutions they are receiving, thereby proactively reducing reasons for churn.

8. Future Scope:

The predictions of these models can be improved by tuning their hyperparameters, finding appropriate train-test splits by using KFold cross validation etc.

The trained model obtained can be deployed as a system to instantaneously generate predictions for a set of observations.