NAME: Hritik Ranjan

Roll: CSE/20047

Reg: 567

Assignment:

Predicting Customer Churn in a Telecommunications Company

Objective:

- The primary objective of this project is to develop a predictive model that can identify customers at risk of churning, enabling the company to take proactive measures to retain them.

Tasks:

- Data Collection and Preprocessing:
- You can use this dataset for the assignment: https://www.kaggle.com/datasets/blastchar/telco-customer-churn
- Preprocess the data to handle missing values, encode categorical variables, and prepare it for analysis.

Exploratory Data Analysis (EDA):

- Perform EDA on the dataset to understand customer behavior and factors influencing churn.
- Visualize key findings using appropriate graphs and charts.

Feature Engineering:

- Create relevant features that can help in predicting churn.

Building the Churn Prediction Model:

Choose and implement a machine learning algorithms for churn prediction. You can consider algorithms like logistic regression, random forests, gradient boosting, or any other suitable models. Train and fine-tune the models using the dataset.

Telco Customer Churn Analysis Report

Contents

Introduction

This project aims to build a smart tool that can predict which customers might leave as per the provided objectives. This is a business problem arises in the telecom companies and the idea is to help the company find these customers early on so they can take action to keep them happy and retain probable customers who are going to churn. By using smart computer methods, this project aims to provide the company a heads-up about customers who might leave soon. This way, the company can do things to make them stay. The main goal is to make sure the company keeps more customers and does well in the long run.

Analyzing the data for this project will uncover valuable insights for the business. Subscriber churn can be in different forms some of them are product churn(ex.-postpaid to prepaid),plan churn(ex. – 599Rs. plan to 349Rs. plan), subscriber churn (ex.- port from airtel to jio). And there can be various reasons for churn either conditionally or unconditionally. By understanding those patterns and trends in customer behavior, the company can gain useful information. For instance, the model might reveal specific factors that contribute to customers churn, allowing the business to address those issues proactively. Additionally, insights could highlight characteristics of customers who are more likely to stay, helping the company focus its efforts on retaining those groups. These findings can guide strategic decisions, resource allocation, and the development of targeted retention strategies, ultimately leading to a more informed and effective approach to customer management. Moving towards predictive building…..

Data Collection

Data for this project has been systematically gathered from the provided website link (https://www.kaggle.com/datasets/blastchar/telco-customer-churn). The collection process involved downloading the dataset. By leveraging data from the site, this project

aims to derive valuable insights and patterns that will contribute to the development of a robust predictive model for customer churn analysis.

Data Preprocessing and Feature Engineering

Steps performed across the project are the following:

- Getting insights about the data types and null values. Here in the provided dataset the datatype of some of the columns were not according with the values in column so it needed to change. And there were very few around 0.15% rows which had empty value in one of the column . So, those rows has been removed.

- Duplicate values: The given data has been checked for the duplicate rows. There were no duplicate rows present so no operation needed to perform.

- Data Transformation: In this step Standardization/Normalization i.e. Scaling numerical features to a standard range, making it easier to compare and analyze them. In given Dataset normalization was performed on the column named ‘Tenure’ which had many distinct values .By creating new column named ‘Tenure group’ with categorical format analyzation became easier.

- Encoding : In this project one-hot encoding method has been used which involves converting categorical variables into numerical format, to make them suitable for machine learning algorithms. In one-hot encoding each categorical value treats as a column and value will be 1/0 in the column.

- Data Reduction: Here in the project the columns such as ‘CustomerId‘ and others were dropped as the they were not a relevant feature.

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a robust technique for familiarizing with Data and extracting useful insights. Data Scientists sift through Unstructured Data to find patterns and infer relationships between Data elements. Data Scientists use Statistics and Visualization tools to summarize Central Measurements and variability to perform EDA.

Here in this project Univariate, Bivariate and Multivariate Analysis were performed.

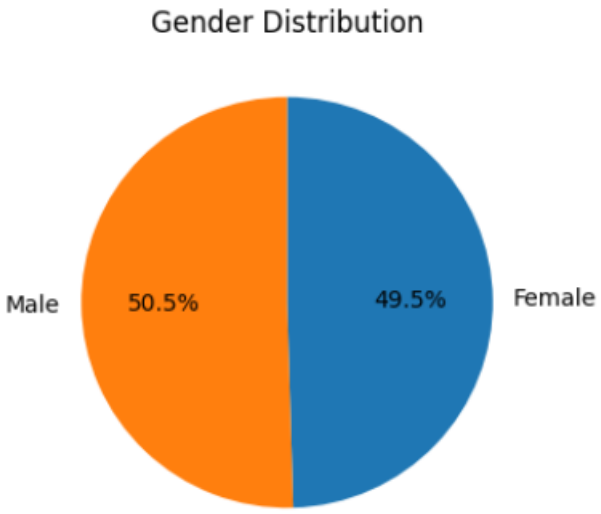
- Univariate Analysis: Univariate analysis focuses on examining and summarizing the characteristics of a single variable in isolation. It's all about understanding the distribution and patterns within that one variable.
- Bivariate Analysis: Bivariate analysis involves the study of two variables to determine if there is any relationship or correlation between them. It explores how changes in one variable might be associated with changes in another.
- Multivariate Analysis: Multivariate analysis extends the analysis to more than two variables. It explores how multiple variables interact with each other and contribute to a particular outcome or behavior.

Note: Most of the Eda analysis performed has been described in the markdown in code section. All the plots are present in code section.

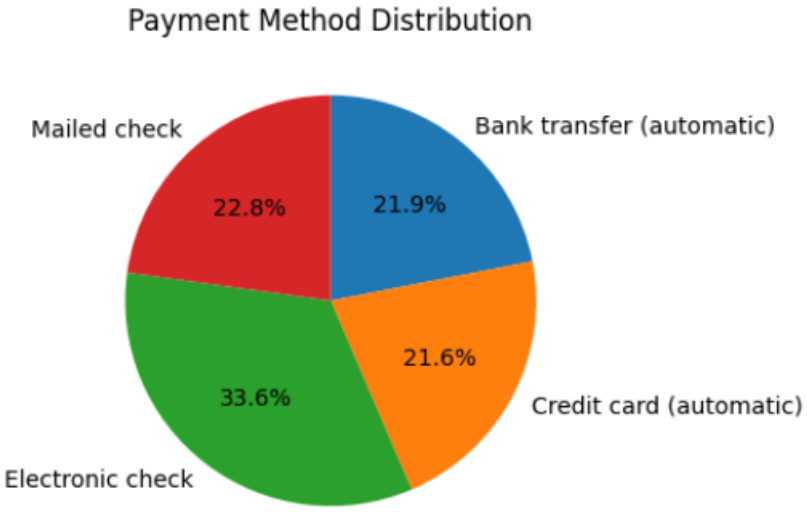
Univariate Analysis:

Here in this project each variable is plotted with pie graph and found the distribution insights of different category of the customers. Some of them are as follow.

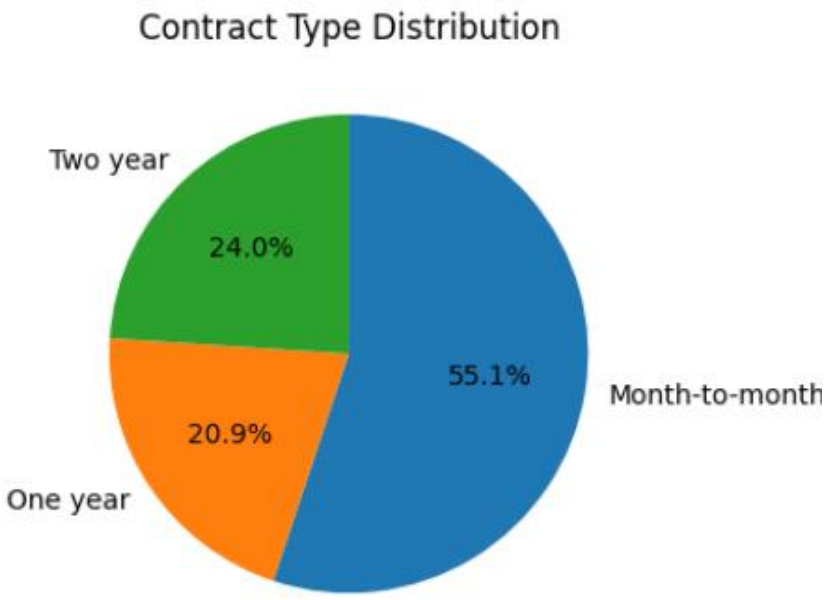
- Males are more in dataset than Females



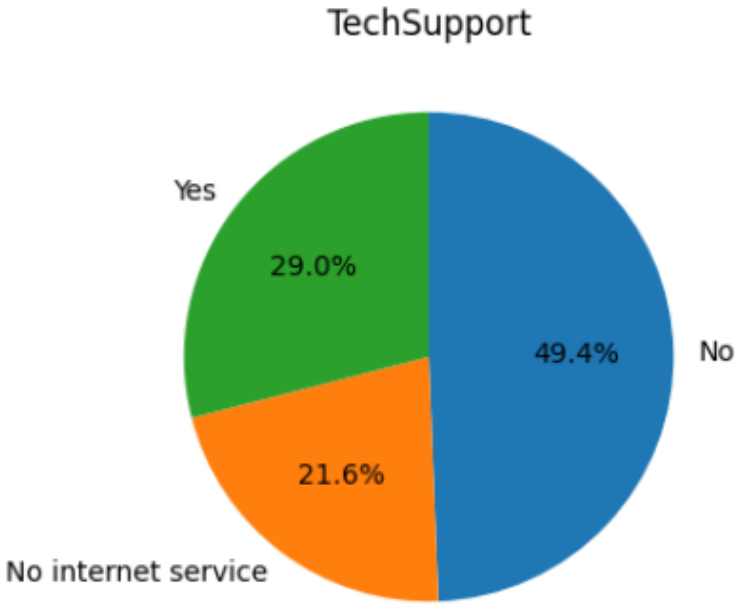
- More customers are having payment method Electronic check then Mailed likewise



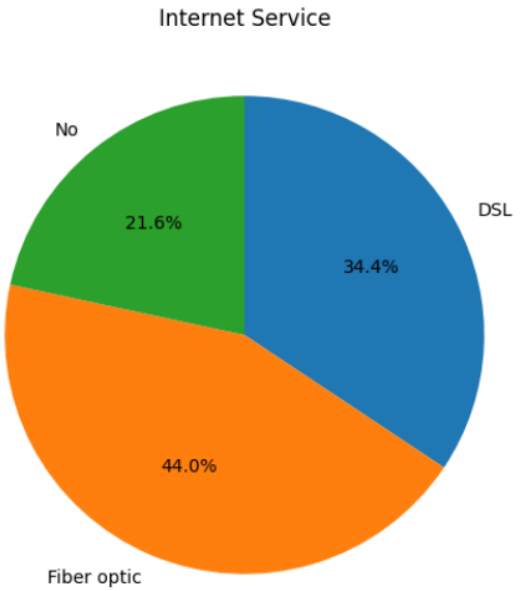
- Customer are more with month-to-month contract type than others.



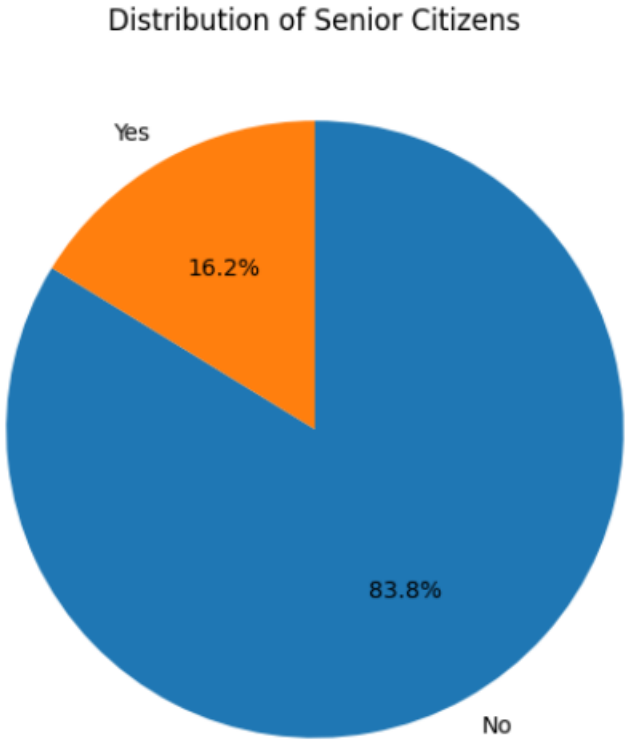
- Most of the customer do not have tech support



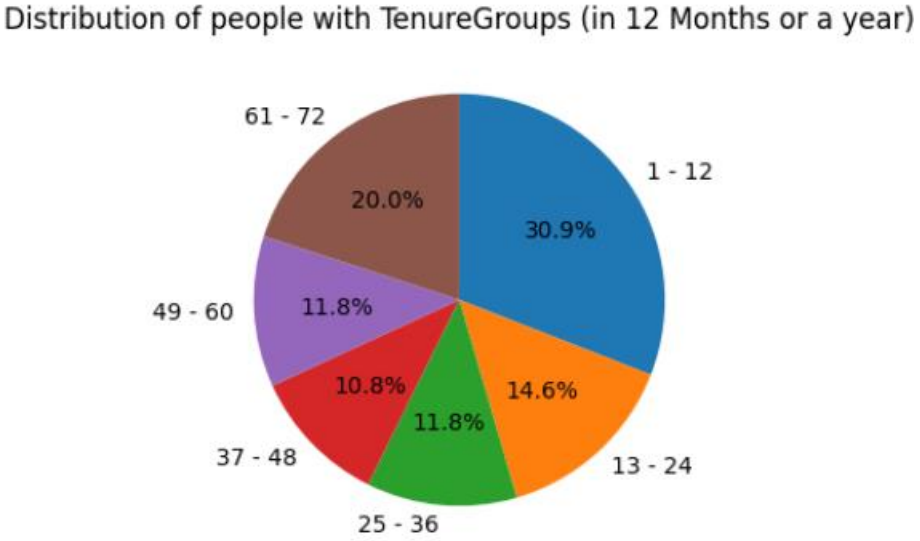
- Customers use fiber optic then DSL and likewise for Internet Service in priority



- Among customers Senior citizens are very less.



- Here as per dateset 30.9% of people have tenure ranges from 1 month to 12 months and 14.6% people have tenure ranges from 13 months to 24 months and likewise. (Tenure group was created in the process of data preprocessing)

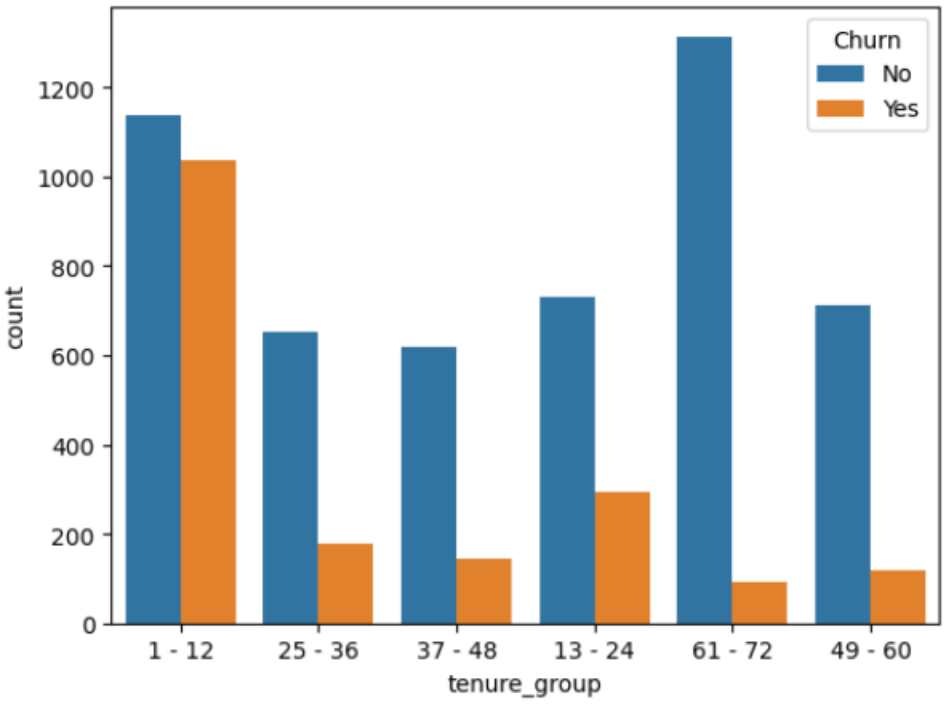


- Others….(mentioned in markdown)

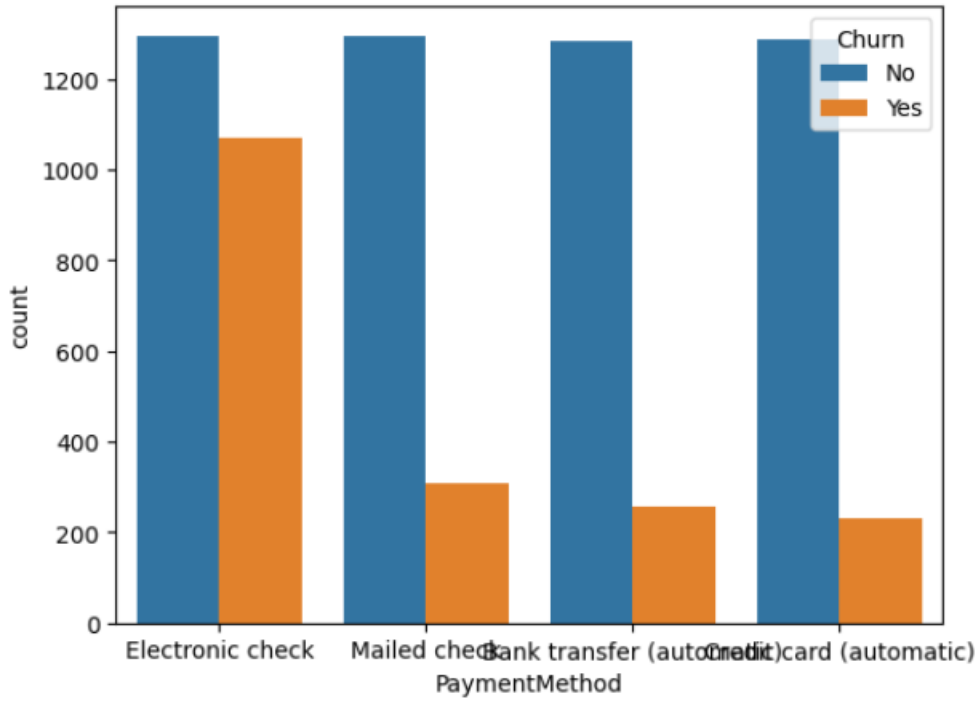
Bivariate Analysis

Here in this project bar graph are plotted with considering the churn and found the different insights of different category of the customers. Some of them are as follow.

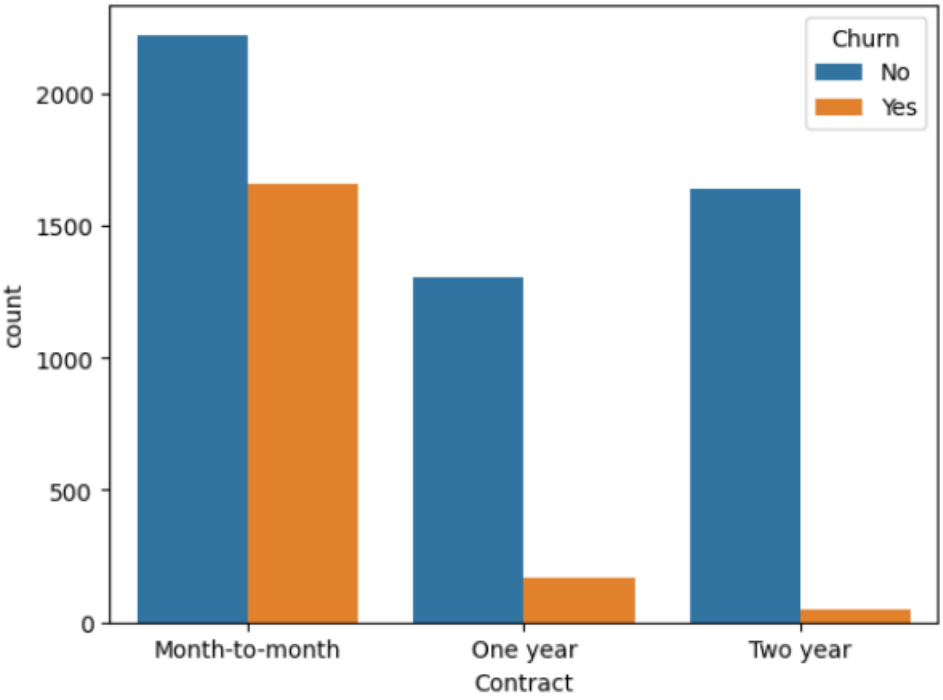
- As observed from plot that the customers of tenure group 1-12 churns more than other tenure groups



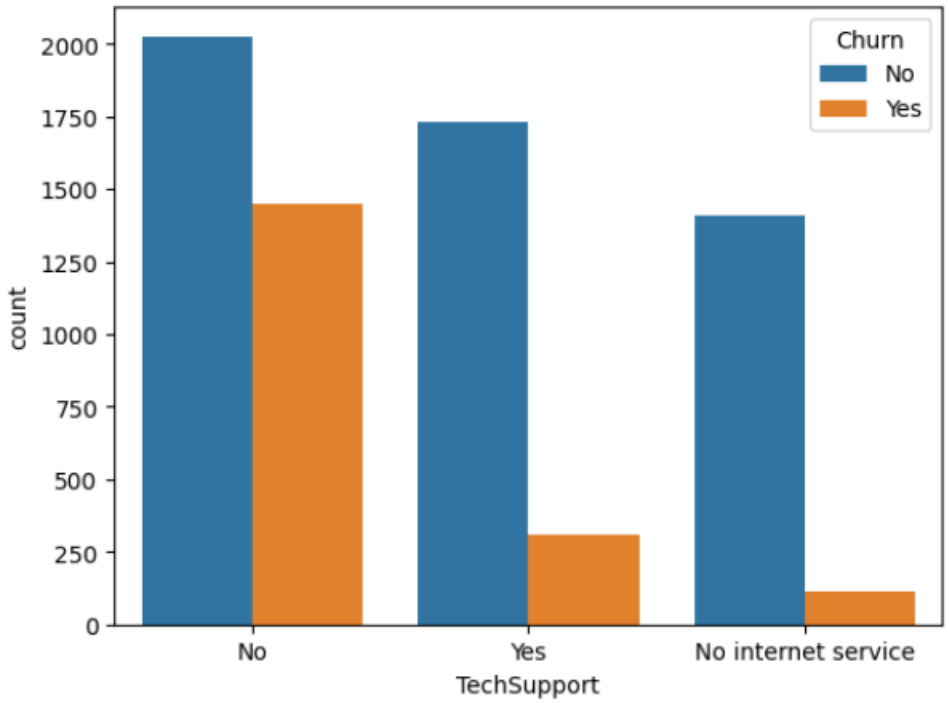
- As observed from plot that customers who uses Electronic Check Payment Method churns more than customers with other payment methods



- Observation from plot is that the customers with month-to-month contract churns more than other contract types



- It is seen through graph is that the customer with no tech support churns more than others

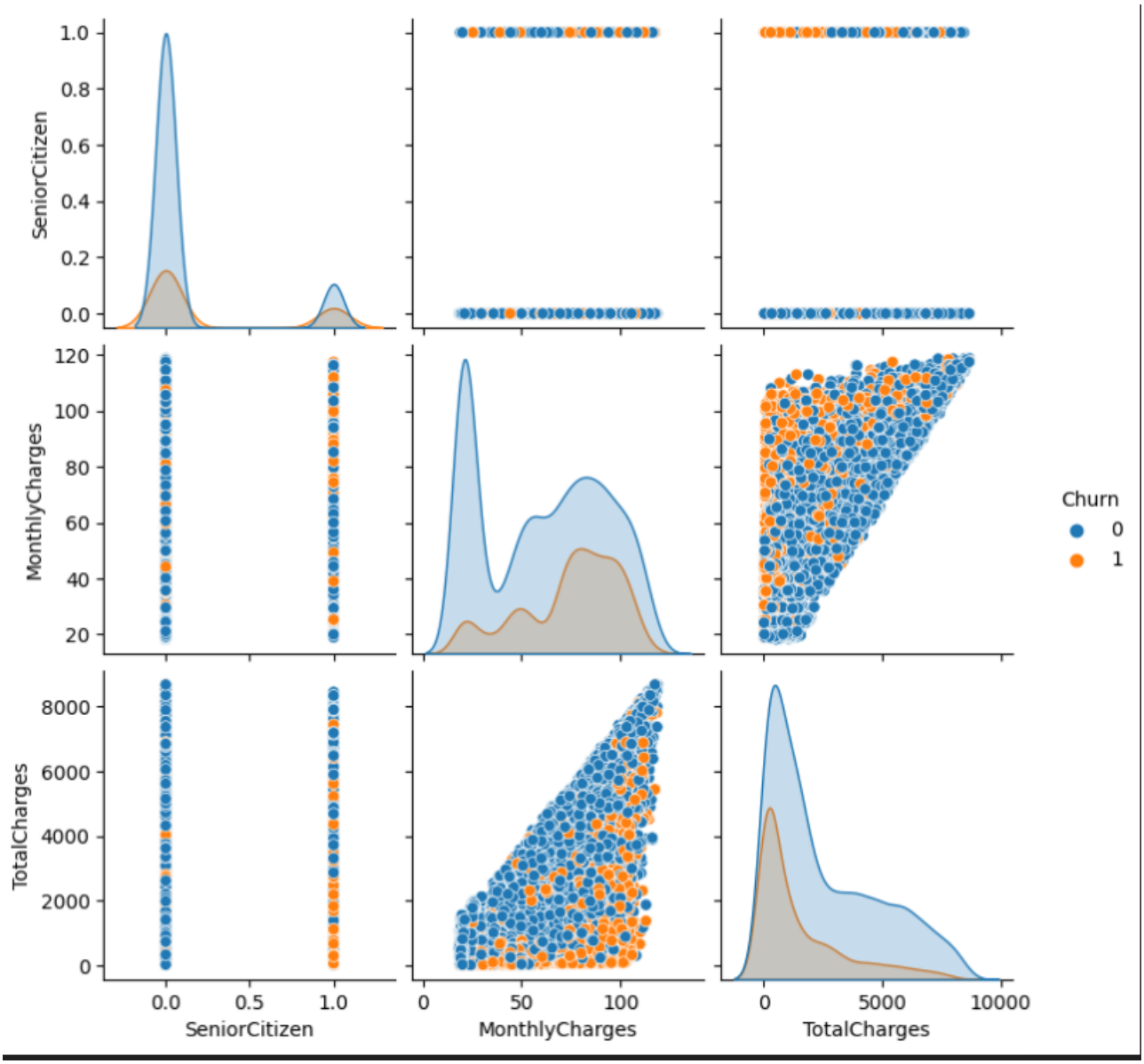


- Others….(mentioned in markdown)

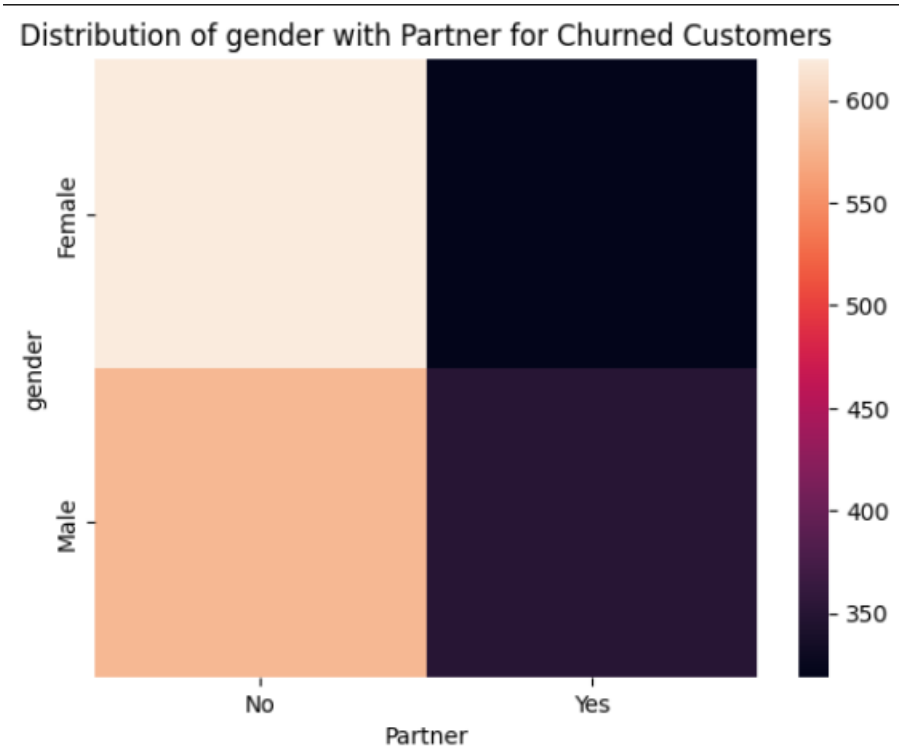
Multivariate Analysis

Here in this project correlations are calculated and heatmap are plotted with considering the churn and found the different insights of different category of the customers. Some of them are as follow.

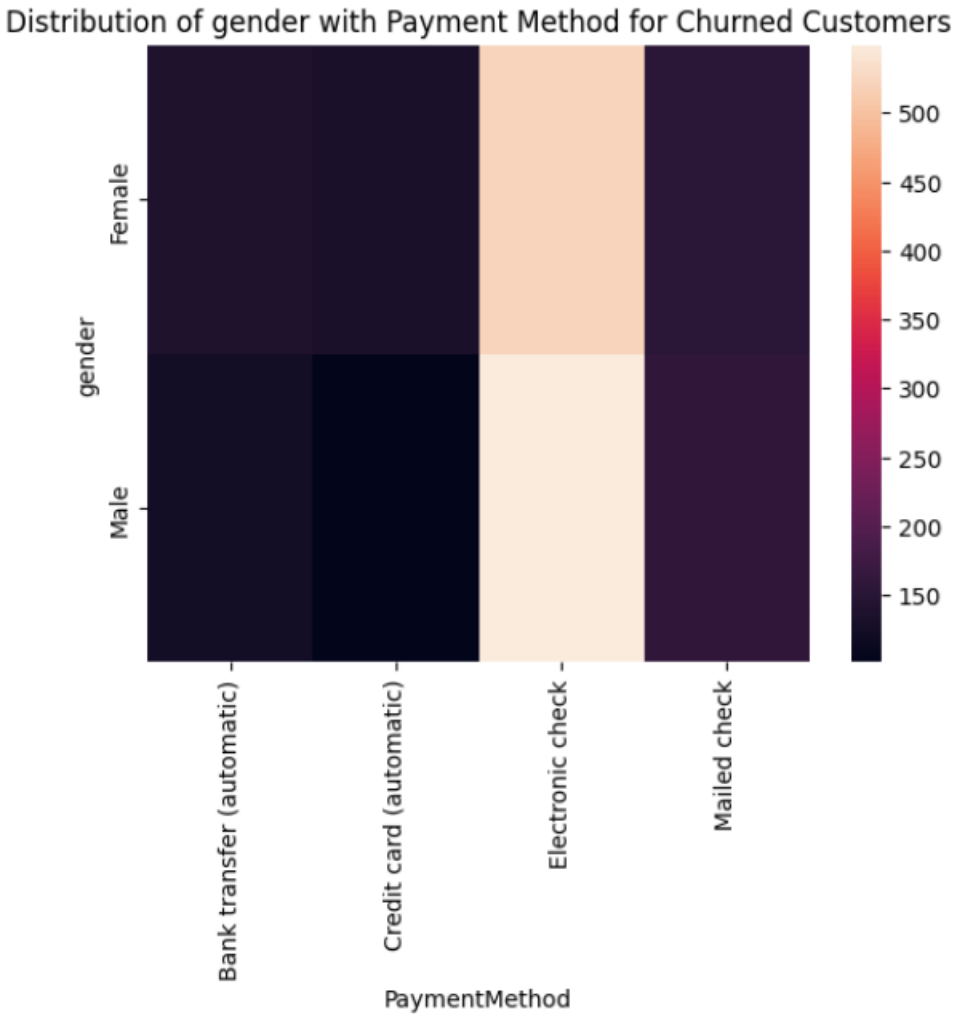
- From plot(in code section) it can be observed that the customer churns more when the monthly charges are high. And total charges grow linearly with monthly charges



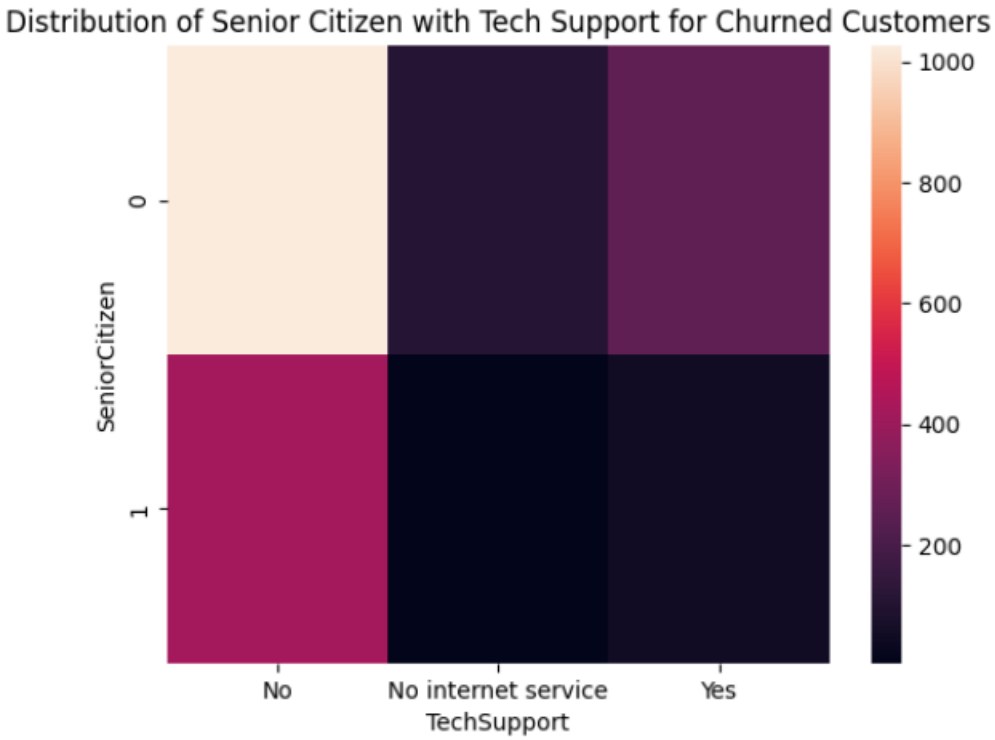
- From plot it is analyzed that female are more likely to churn when there is no partner and when there is a partner male churns more comparatively



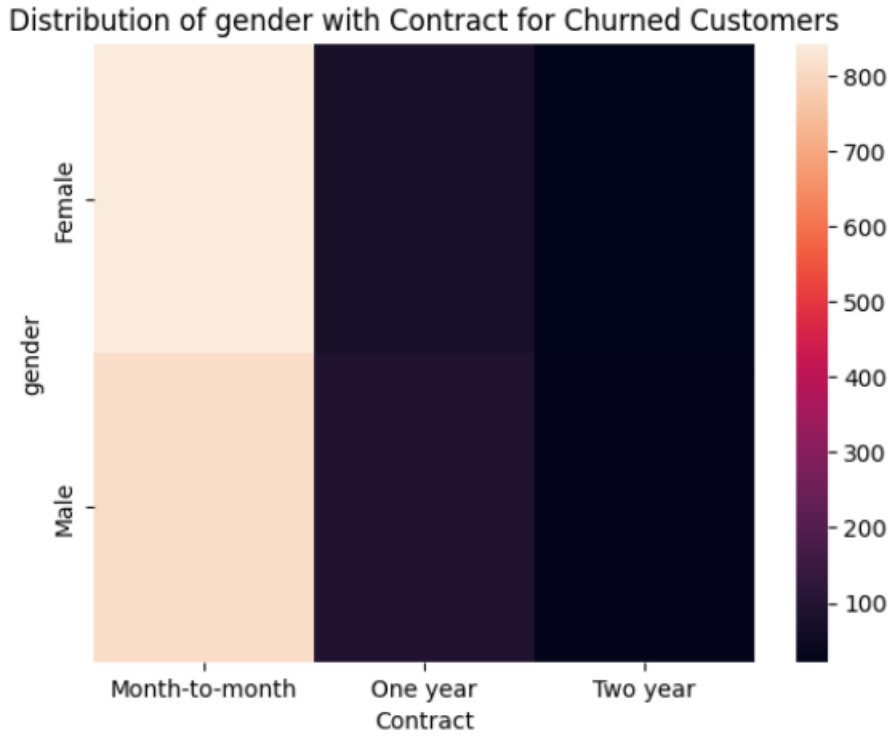
- Hera as analyzed from plot that churners are more for electronic check payment method among them male are more likely to churn.



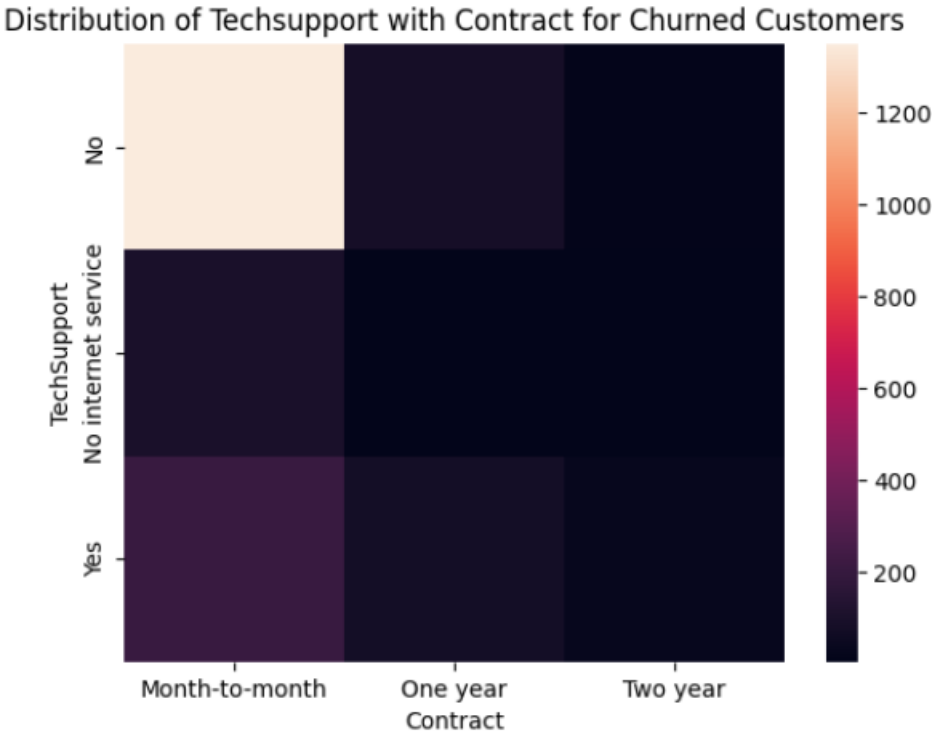
- not senior citizen i.e. younger citizens churns more when there is no tech support



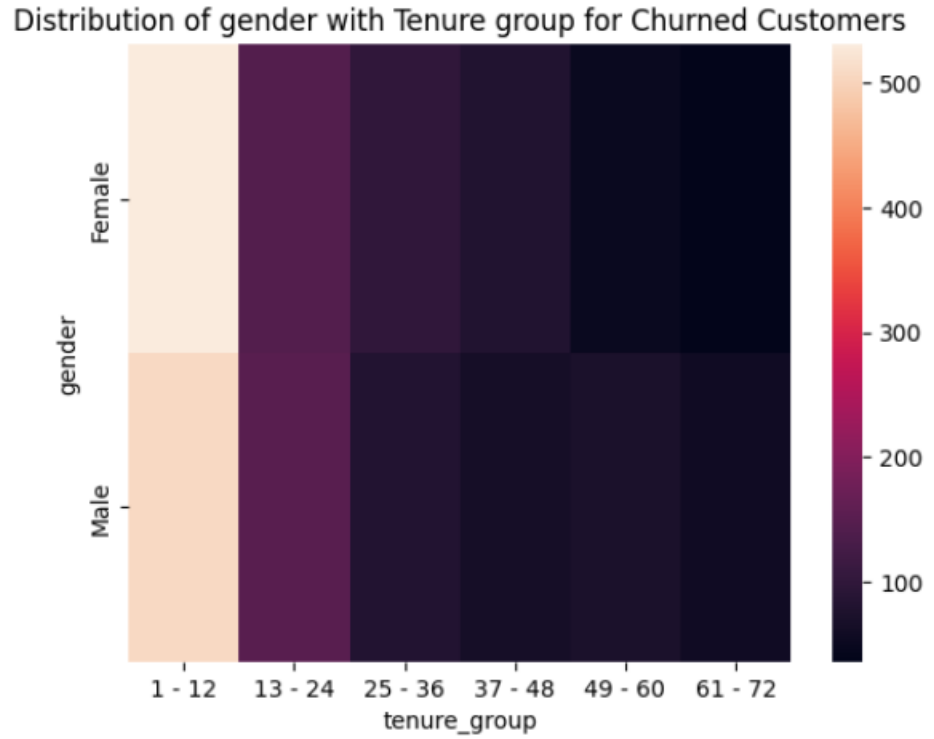
- Among the churners female churns more than male for month-to-month contract comparatively



- Frequency of churners are substantially higher when there is no tech support and the month-to-month contract



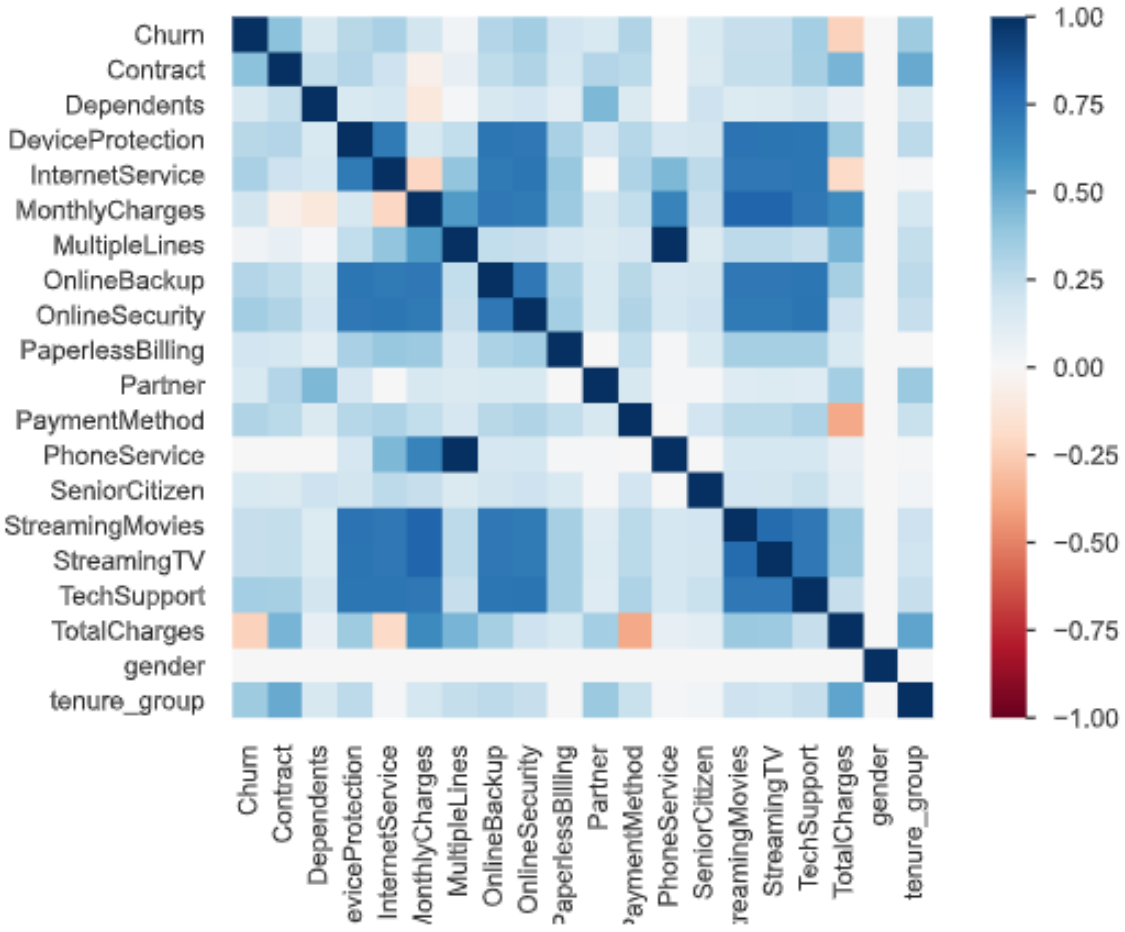
- Customers of tenure ranges from 1 to 12 likes to churn more , among them Female churns more



- Others….(mentioned in markdown)

Other observations

The below are the observations from the correlations, heatmap , report generated in code and written as per dataset :



- 'Phone Service' column has high correlation with monthly charges column and multiple lines column.
- Tech support column has high correlation with monthly charges column and Internet service column and other columns.
- Contract column is highly correlated with Tenure group column.
- Monthly charges and total charges are highly correlated with each other. They grow linearly with each other. With growing monthly charges and total charges customer is more likely to churn. (as per pair plot)

Model Building

This is one of the most crucial processes in Data Science Modelling as the Machine Learning Algorithm aids in creating a usable Data Model. The choice of the model depends on various factors related to the dataset and the goals of the analysis. Some models are mentioned below.

- Logistic Regression: Why: Simple and interpretable; suitable for linear relationships.
- Decision Trees: Why: Captures non-linear patterns; easy to interpret.
- Random Forest: Why: Ensemble method for higher accuracy; robust to overfitting.
- Support Vector Machines (SVM): Why: Effective in high-dimensional data; captures complex relationships.

All models are imported using libraries.

In the project ,Data splitting happened the 80% data is used for training and 20% for testing. Test data separation is done using library only. 'Accuracy', 'Precision', 'Recall', 'F1 Score', 'ROC AUC' are the measures used for the model evaluation. All models used in the project is mentioned : 'Logistic Regression'

- 'Decision Tree'
- 'Random Forest'
- 'Gradient Boosting'
- 'Support Vector Machine'
- 'K-Nearest Neighbors'
- 'Naive Bayes'

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
Logistic Regression	0.781805	0.621818	0.457219	0.526965	0.678271
Decision Tree	0.763326	0.559420	0.516043	0.536857	0.684449
Random Forest	0.786780	0.633094	0.470588	0.539877	0.685923
Gradient Boosting	0.792466	0.643357	0.491979	0.557576	0.696619
Support Vector Machine	0.785359	0.647541	0.422460	0.511327	0.669604
K-Nearest Neighbors	0.745558	0.522599	0.494652	0.508242	0.665526
Naive Bayes	0.692964	0.457101	0.826203	0.588571	0.735464
Neural Network	0.744847	0.523077	0.454545	0.486409	0.652249

As seen through the model training that the result of the parameters were not very good and reason for that was the imbalance in the data as the the count of churners were very low compared to the not churners.

Solution for this is sampling. Here in this project up-sampling is performed over dataset. After training the score is as below

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
Logistic Regression	94.251902	95.276873	93.750000	94.507270	94.281082
Decision Tree	93.744717	93.929712	94.230769	94.080000	93.716458
Random Forest	94.843618	94.896332	95.352564	95.123901	94.814028
Gradient Boosting	95.350803	95.230525	95.993590	95.610535	95.313432
Support Vector Machine	94.759087	95.469256	94.551282	95.008052	94.771169
K-Nearest Neighbors	93.068470	90.813253	96.634615	93.633540	92.861136
Naive Bayes	87.827557	86.809816	90.705128	88.714734	87.660256
Neural Network	93.829248	93.799682	94.551282	94.173982	93.787269

The results after sampling were quite well but more tuning to the model would be better that’s why some others steps were also performed.

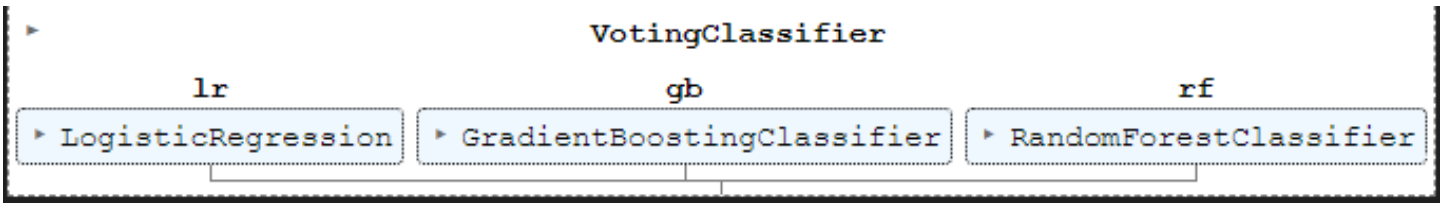
Among them PCA was used at first as PCA helps in dimensionality reduction collinearity handling, Noise Reduction and improves the model performance . As measures of the Gradient boosting were highest so PCA was performed over it. Scores of it is .. overall Accuracy-93.153% and measures for 0/1 value of churn are-

0.9315300084530854

	precision	recall	f1-score	support
0	0.93	0.92	0.93	559
1	0.93	0.94	0.94	624
accuracy			0.93	1183
macro avg	0.93	0.93	0.93	1183
weighted avg	0.93	0.93	0.93	1183

Later performed Voting classifier. A Voting Classifier is an ensemble learning technique that combines the predictions of multiple individual classifiers to improve overall predictive accuracy and robustness. It works by aggregating the predictions of each classifier and selecting the most predicted class (for classification tasks) or averaging the predicted probabilities (for probability estimation tasks). For voting classifier the best three performing model is used which are

- LogisticRegression
- GradientBoostingClassifier
- RandomForestClassifier



And scores of measures were like this.

Accuracy 0.9509721048182587
Precision 0.9477848101265823

After that Stacking classifier was also tried.

A Stacking Classifier is another form of ensemble learning, like the Voting Classifier, but with a more complex architecture. Instead of simply averaging or taking a majority vote, a Stacking Classifier leverages the predictions of multiple base classifiers to train a meta-classifier, which then makes the final predictions. It introduces a higher level of abstraction by incorporating a second layer of models to learn how to best combine the predictions of the base classifiers.

estimators=[logistic regression, gradient boosting , Random forest]

final_estimator= Support Vector Machine

StackingClassifier(estimators=estimators, final_estimator=final_estimator)

And the results of this was better from all.

```
Accuracy 0.9518174133558749
Precision 0.9536
```

...	precision	recall	f1-score	support
0	0.95	0.95	0.95	559
1	0.95	0.96	0.95	624
accuracy			0.95	1183
macro avg	0.95	0.95	0.95	1183
weighted avg	0.95	0.95	0.95	1183

This model performed best among all. So this has been further used.