Project Report: Predicting Customer Churn

Introduction:

The objective of this project is to develop a predictive model for customer churn in a telecommunications company. In this report, we outline the process of building a predictive model to identify potential churners and explore insights gained from Exploratory Data Analysis (EDA).

Dataset:

The dataset comprises customer information like ID, gender, Senior Citizen or not, Partner, Dependents, tenure, Phone service, Multiple Lines, type of internet service, Online Security, Device Protection, Tech Support, Streaming TV, Streaming Movies, Contract, Paperless Billing, Payment Method, Monthly Charges, Total Charges, Churn. The target variable is 'Churn,' which indicates whether a customer has churned (1) or not (0).

Methodology:

Form the data set they are so many categorical values such as

gender

Partner

Dependents

PhoneService

MultipleLines

InternetService

OnlineSecurity

OnlineBackup

DeviceProtection

TechSupport

StreamingTV

StreamingMovies

Contract

PaperlessBilling

PaymentMethod

Churn.

Convert them into numerical values because these algorithms use mathematical equations to find patterns and make predictions. Categorical data is not directly usable in these equations, so it needs to be converted to numerical form.

example :

if the categorical value is yes convert into 1 and 0 for no

dataset before converting

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	...	DeviceProtection
0	7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No	...	No
1	5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	...	Yes
2	3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	...	No
3	7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes	...	Yes
4	9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No	...	No

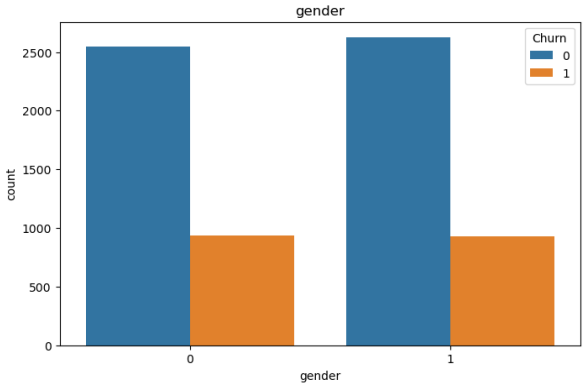
dataset after converting

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	...	DeviceProtection
0	7590-VHVEG	0	0	1	0	1	0	2	1	0	...	0
1	5575-GNVDE	1	0	0	0	34	1	0	1	1	...	1
2	3668-QPYBK	1	0	0	0	2	1	0	1	1	...	0
3	7795-CFOCW	1	0	0	0	45	0	2	1	1	...	1
4	9237-HQITU	0	0	0	0	2	1	0	2	0	...	0

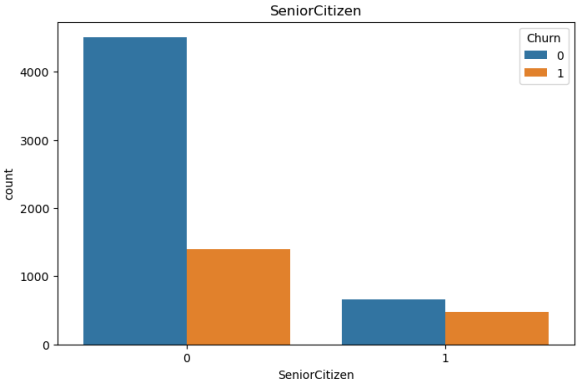
We have to find which attribute has more impact on target attribute churn by analysing the data and by generating correlation matrix.

Graphs of various attributes vs target attribute (churn)

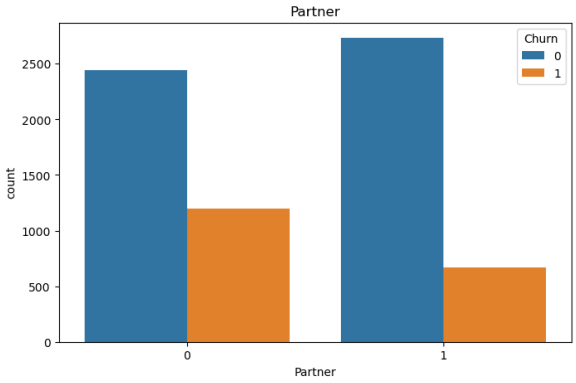
Gender vs Churn



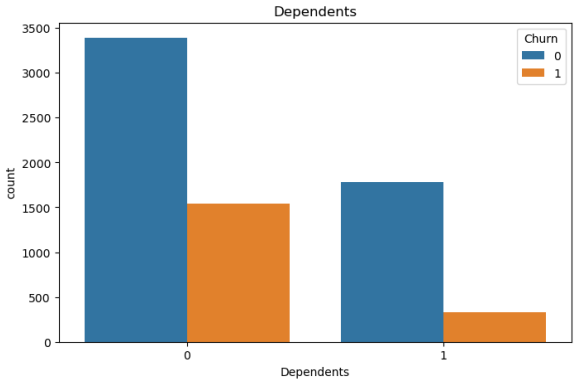
Seniorcitizen vs Churn



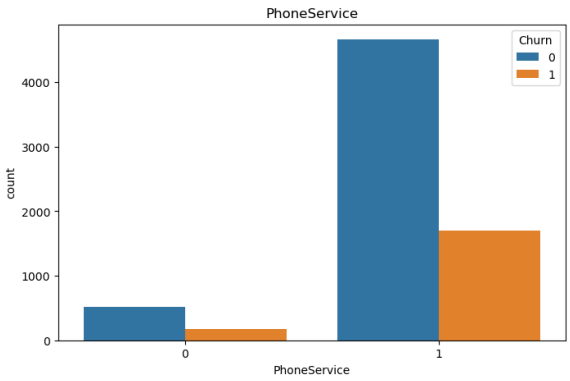
Partner vs Churn



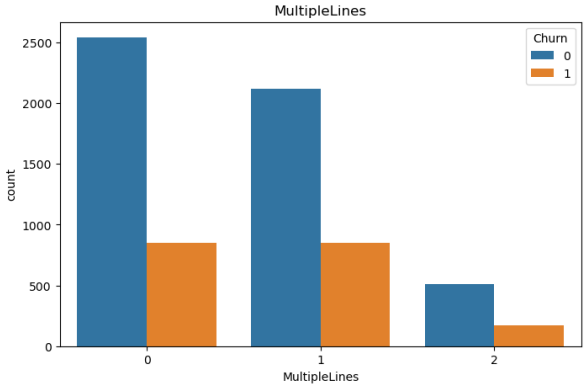
Dependents vs Churn



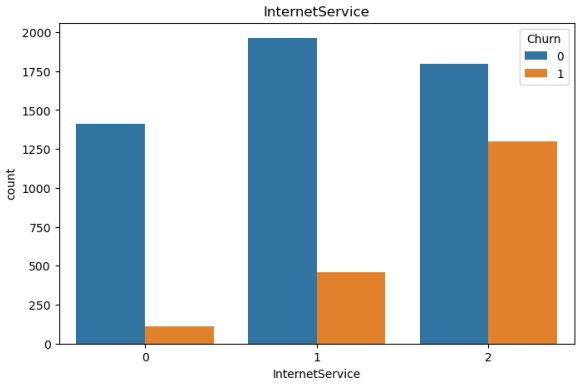
Phone Services vs Churn



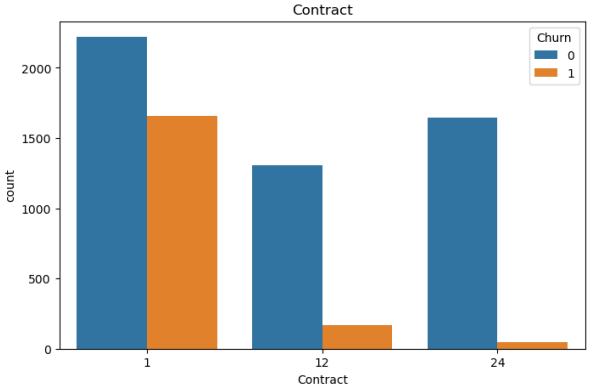
Multiple Lines vs Churn



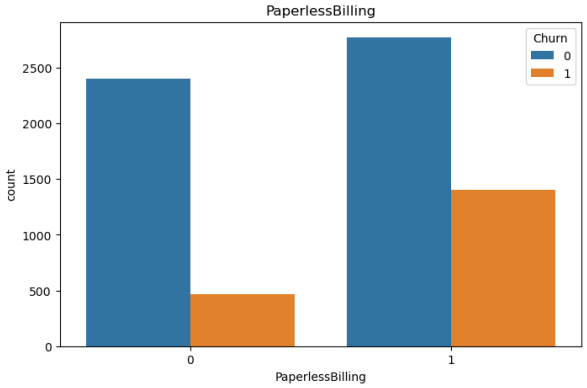
Internet Services vs Churn



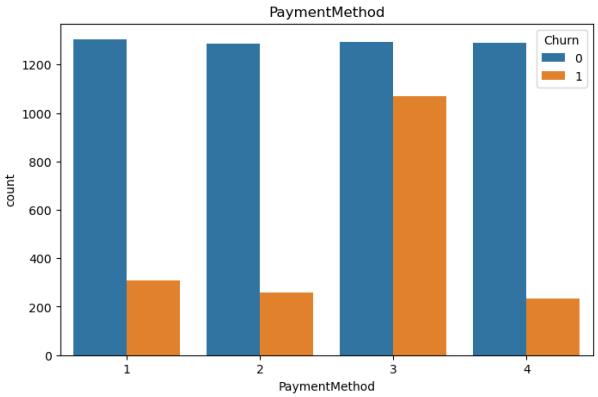
Contract vs churn



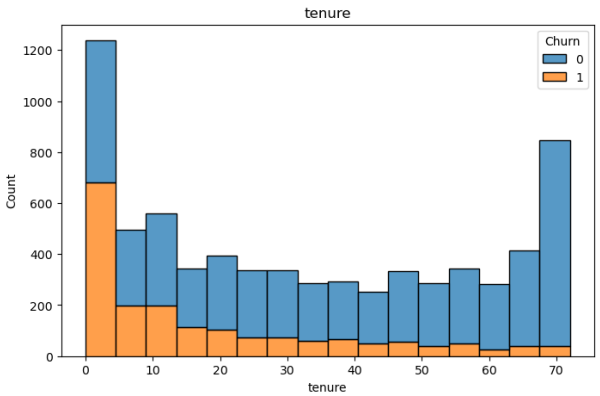
Paperless billing vs Churn



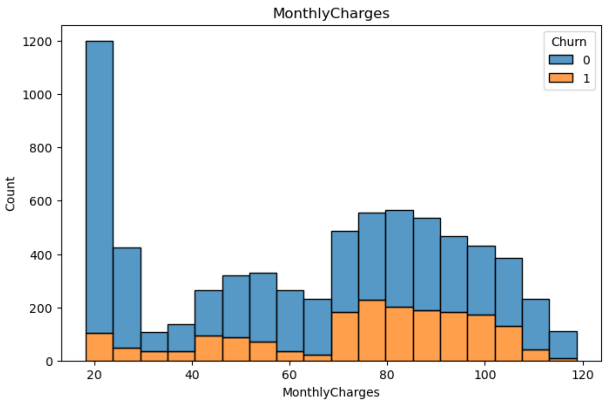
Payment method vs Churn



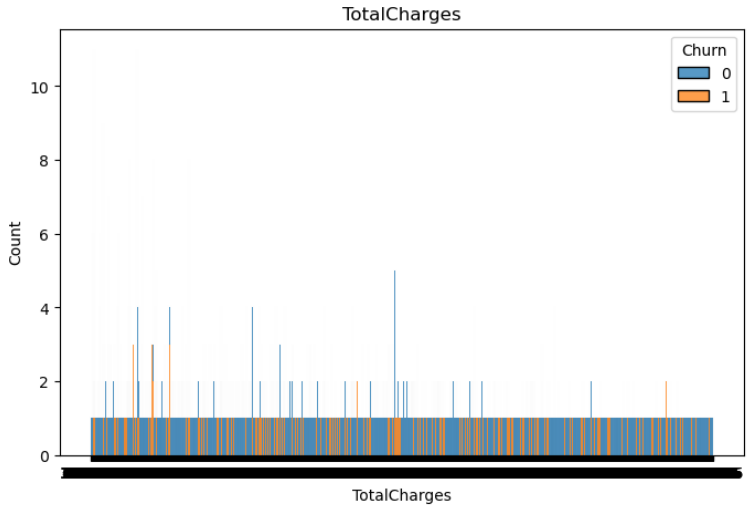
Tenure vs Churn



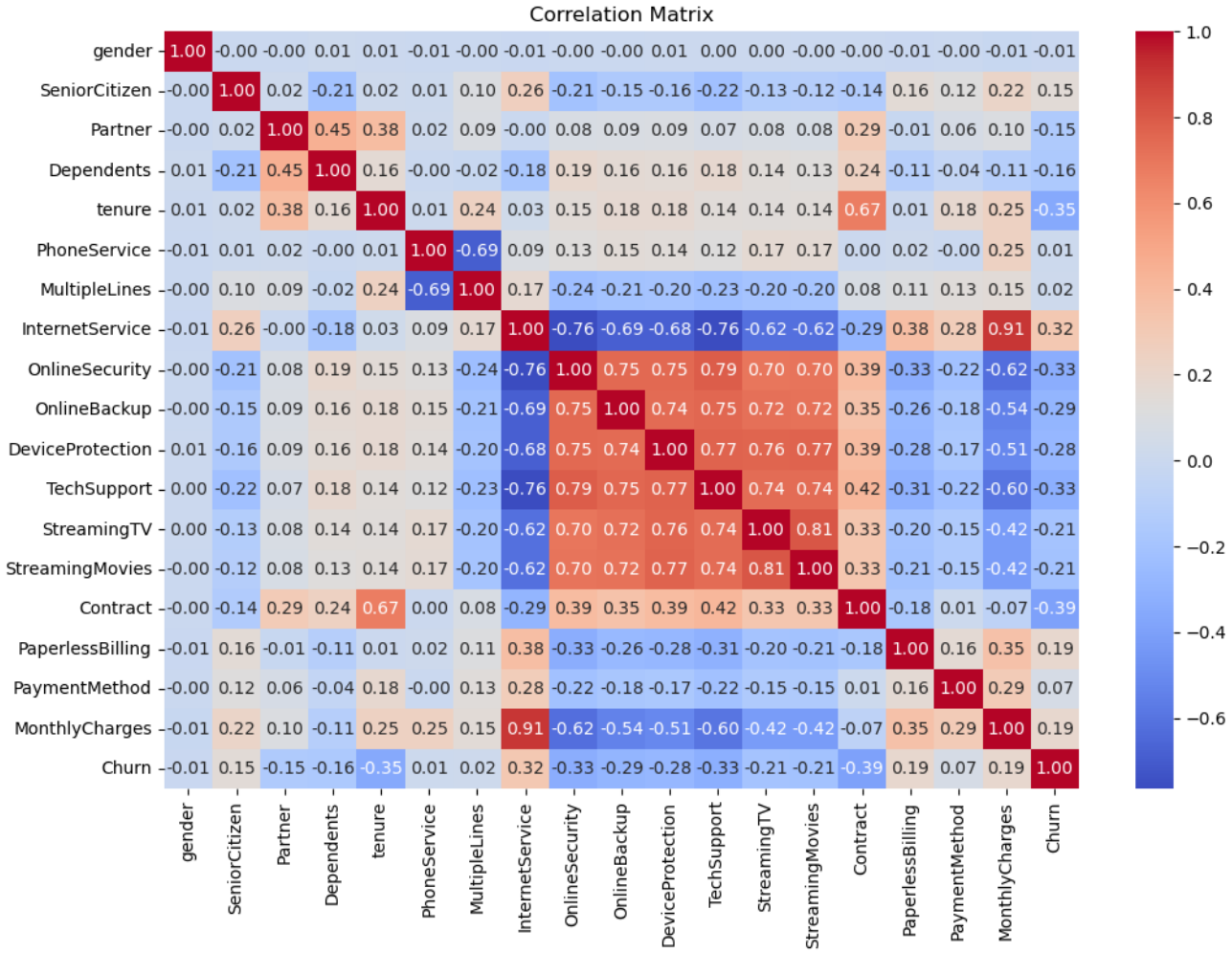
Monthly Charges vs Churn



Total Charges vs Churn



Correlation matrix



Monthly Charges, Paperless Billing, Internet Service, Senior Citizen has more impact on churn because the correlation value is more for churn

Monthly Charges - 0.19

Paperless Billing - 0.19

Internet Service - 0.32

Senior Citizen – 0.15

Now, we need to determine the most suitable algorithm for our dataset. We're choosing from options like Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, and K-Nearest Neighbors. The goal is to select an algorithm that works well with our data and fits the specific things we want to find out.

Because the dataset is relatively small, consisting of 7,043 entries, and we're looking for a method that is efficient and performs well, I've decided to use logistic regression.

Explanation:

Small Dataset : The dataset contains a limited number of entries (7,043).

Efficiency and Performance : The objective is to find an approach that is both efficient and performs well in making accurate predictions.

Choice of Logistic Regression : Logistic regression is chosen as the modeling technique. Logistic regression is a statistical method used for binary classification problems, where the outcome is either 0 or 1. It's favored for its simplicity, efficiency, and interpretability, making it suitable for small datasets while still delivering good performance in many cases.

Steps in Logistic Regression

1. Data Pre-processing step
2. Fitting Logistic Regression to the Training set
3. Predicting the test result
4. Test accuracy of the result (Creation of Confusion matrix)
5. Visualizing the test set result.

Step 1: We've already prepared the data.

Step 2: Now, we're dividing the data into two parts - 80% for training and 20% for testing. This helps us teach the model and check how well it's doing.

Step 3: Time to teach the model! We're using a tool (Logistic Regression) from a library called sklearn. We're showing the model our training data, and it learns from it using a function called "fit."

Step 4: We let the trained model make predictions using the testing data. We use a function called "predict" for this.

Step 5: Finally, we want to know how good our model is. We check its accuracy using a tool from the sklearn library called "accuracy_score."

Accuracy of the model :

```
Accuracy: 0.76
              precision    recall  f1-score   support

           0       0.78      0.93      0.85      1036
           1       0.61      0.29      0.39       373

    accuracy                           0.76      1409
   macro avg       0.70      0.61      0.62      1409
weighted avg       0.74      0.76      0.73      1409
```

Challenges faced :

Handling Categorical Variables:

Challenge: Logistic regression requires numerical input, and dealing with categorical variables (like 'gender' or 'internet service type') can be challenging.

Solution: Encode categorical variables using label encoding.

Feature Selection:

Challenge: Identifying the most relevant features for predicting churn can be tricky, especially if the dataset has many variables.

Solution: Use techniques like feature importance from correlation analysis to select the most informative features.