Programmers Force

Analysis of Churn Rate

Exploratory Data Analysis and Model Training using Artificial Intelligence

Muhammad Arham
Data Scientist Trainee

CONTENTS

DATA COLLECTION

Data named Telco Customer Churn is the dataset of customers that have churned or not. Churn means if they left in the last month or not. Each row represents a customer, each column contains customer attributes.

The data set 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, age range, and if they have partners and dependents.

The data contains the following features.

Features:

customerID: Unique identifier for each customer.

gender: Customer's gender (Male or Female).

SeniorCitizen: Indicates if the customer is a senior citizen (1) or not (0).

Partner: Whether the customer has a partner or not (Yes or No).

Dependents: Whether the customer has dependents or not (Yes or No).

tenure: Number of months the customer has stayed with the company.

PhoneService: Whether the customer has a phone service or not (Yes or No).

MultipleLines: Whether the customer has multiple lines or not (Yes, No, or No phone service).

InternetService: Customer's internet service provider (DSL, Fiber optic, or No internet service).

OnlineSecurity: Whether the customer has online security or not (Yes, No, or No Internet service).

OnlineBackup: Whether the customer has online backup or not (Yes, No, or No Internet service).

DeviceProtection: Whether the customer has device protection or not (Yes, No, or No Internet service).

TechSupport: Whether the customer has tech support or not (Yes, No, or No internet service).

StreamingTV: Whether the customer has streaming TV or not (Yes, No, or No internet service).

StreamingMovies: Whether the customer has streaming movies or not (Yes, No, or No internet service).

Contract: The term of the customer's contract (Month-to-month, One year, or Two years).

PaperlessBilling: Whether the customer has paperless billing or not (Yes or No).

PaymentMethod: Customer's payment method.

MonthlyCharges: The amount charged to the customer monthly.

TotalCharges: The total amount charged to the customer.

Churn: Whether the customer churned or not (Yes or No).

Churn is the target variable. The data contains 7043 customer information with demographic and company information about the customer. The data consists of 20 features with dtypes: float64(1), int64(2), object(18).

PREPROCESSING

In the preprocessing part, the given steps are followed.

DROPPING FEATURES

CustomerID does not have any valuable information that can be used to predict a “Churn”. So it is dropped.

CHECKING FOR NULLS

No immediate nulls are found In the dataset. In total Charges, there are empty spaces saved, and numerical features are saved strings. So We fill the empty spaces by taking the mean of the feature values and converting them to Float type.

CHECKING FOR OUTLIERS

Outliers were checked by checking the Z-score of all features. There were no features that had a z score of above 2.9.

LABEL ENCODING

Three label encoding techniques were tried.

- Custom labeling
- Label encoder
- One hot encoder

By testing, custom labels gave the best results.

The labels were accounted for by checking the hierarchy of the features and making sure the labels are consistent across all features.

Features	**Labels**
Gender	Male: 1, Female: 0
Partner, Dependents PhoneService PaperlessBilling Churn	No: 0, Yes: 1
InternetService	Fiber optic: 0, DSL: 1, No: -1
OnlineSecurity OnlineBackup DeviceProtection TechSupport StreamingTV StreamingMovies	No: 0, Yes: 1, No internet service: -1
Contract	Month-to-month: 0, One year: 1, Two years: 2
Payment	Electronic check: 0, Mailed check: 1, Bank transfer (automatic): 2, Credit card (automatic): 3

SCALING/NORMALIZING

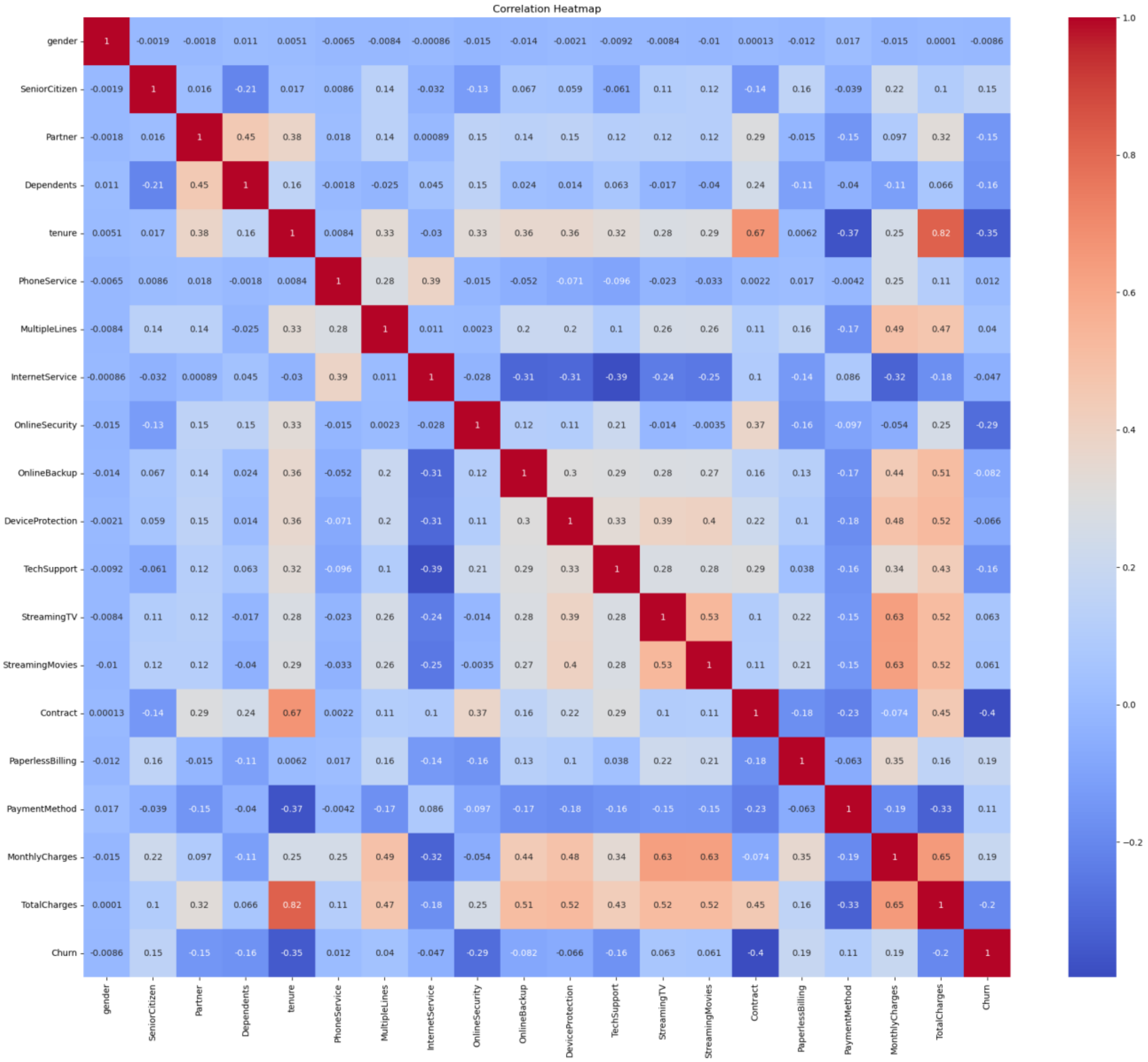
Min Max Scaler and Standard Scaler were tested while normalizing the dataset. The MinMax gave better and simpler labels after normalizing.

Min Max scaler is used for normalizing.

EDA

HEATMAP

Correlation in features analysis through a heatmap.



Here we can say the following features are heavily correlated.

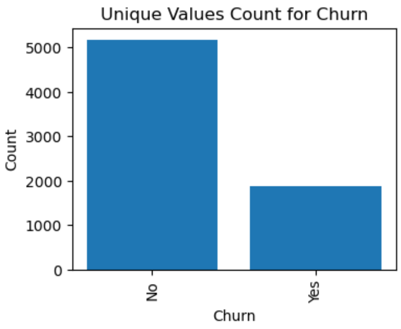
Tenure and Total charges correlate.

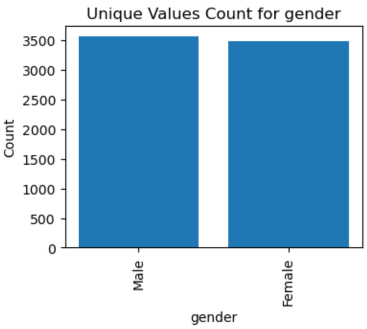
Monthly Charges and Total Charges correlate.

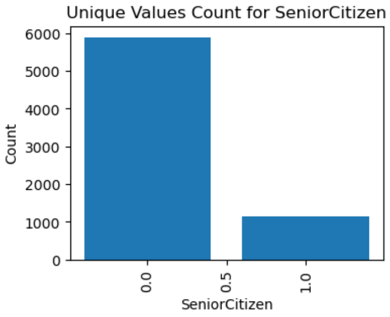
Tenure and contract a correlate.

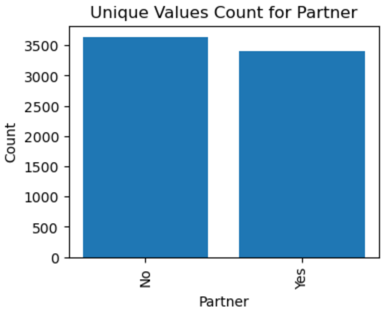
UNIVARIATE ANALYSIS.

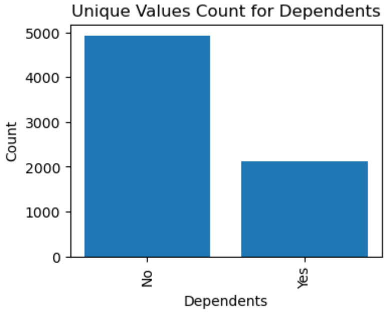
Here We will look at each feature individually.

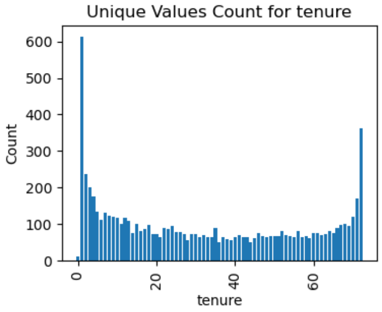


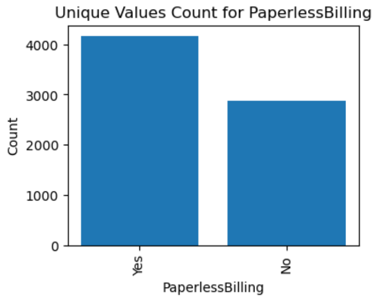


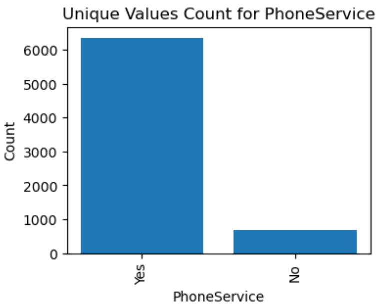


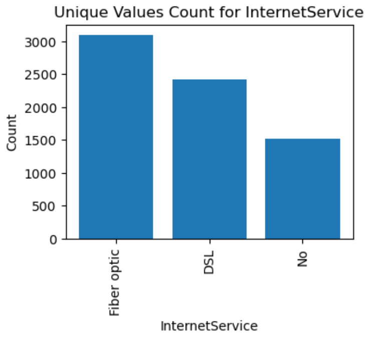


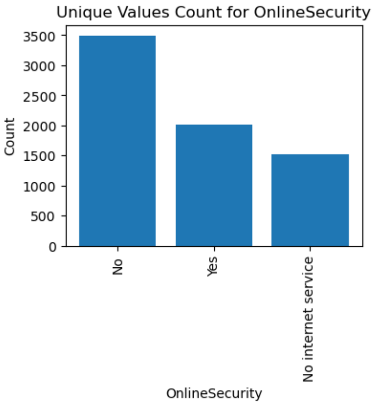


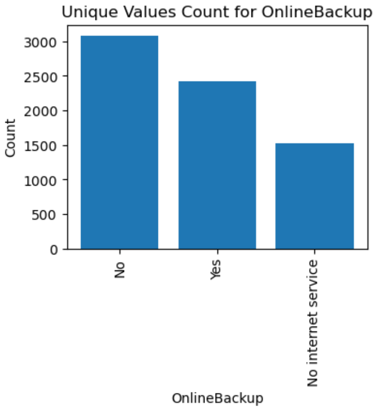


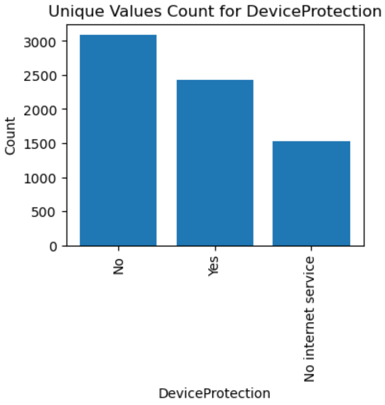


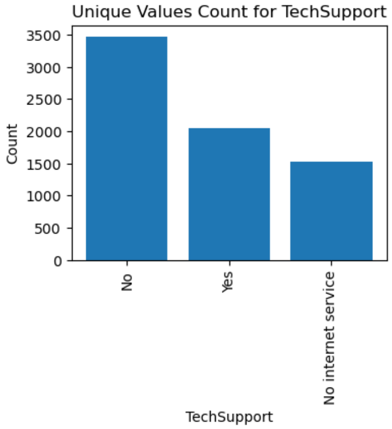


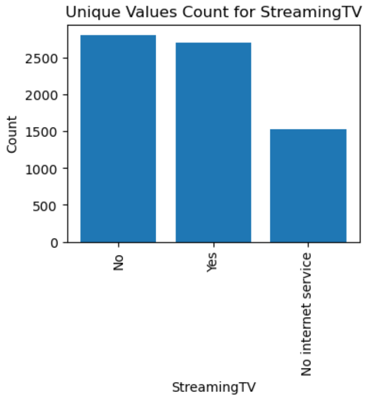


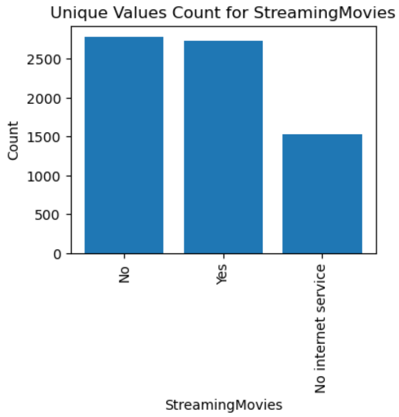


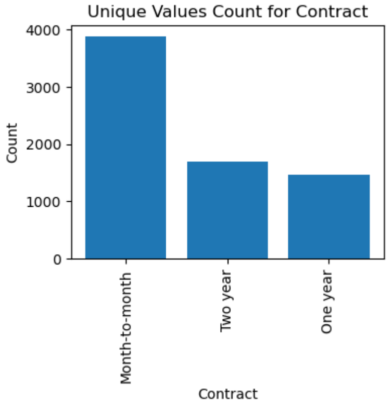


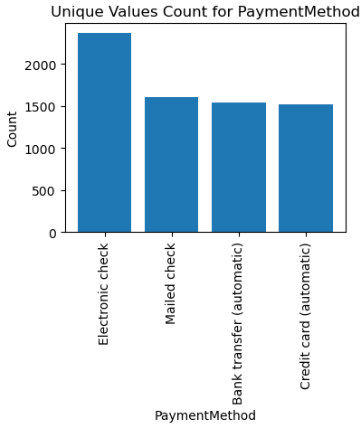


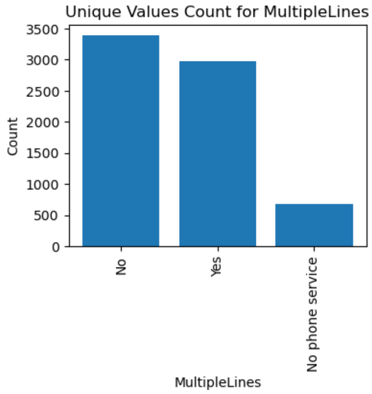


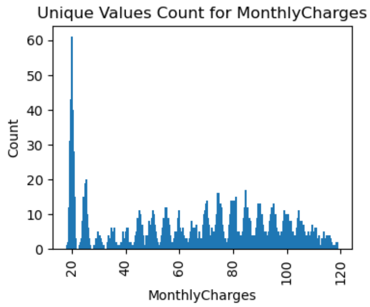






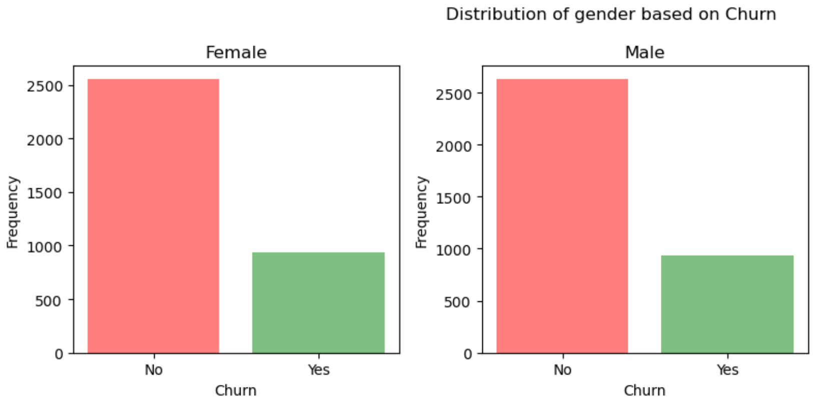




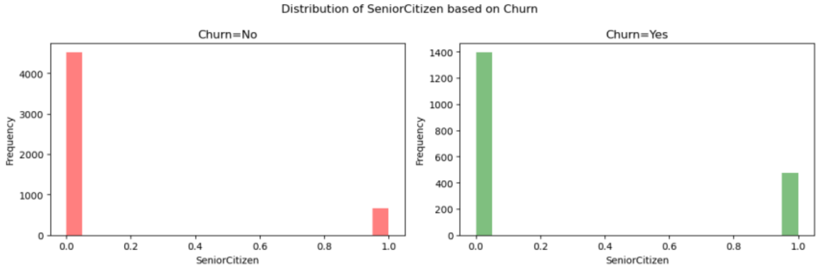


BIVARIATE ANALYSIS WITH CHURN

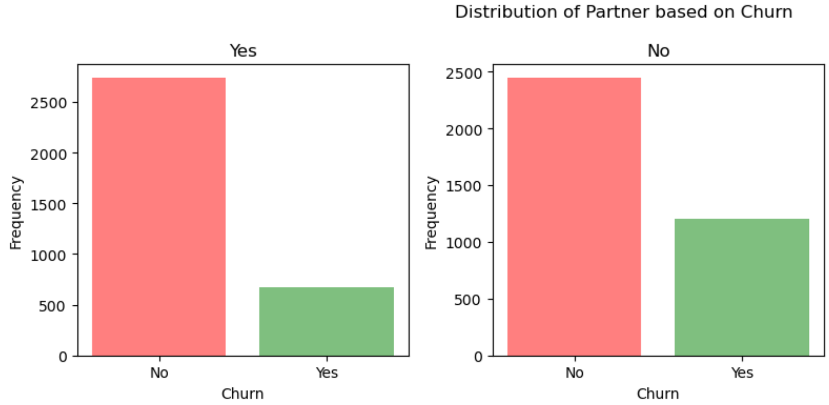
generates histograms and bar plots to visualize the distribution of numerical and categorical features, respectively, based on the 'Churn' label. This will give insights into how each feature behaves according to Churn.



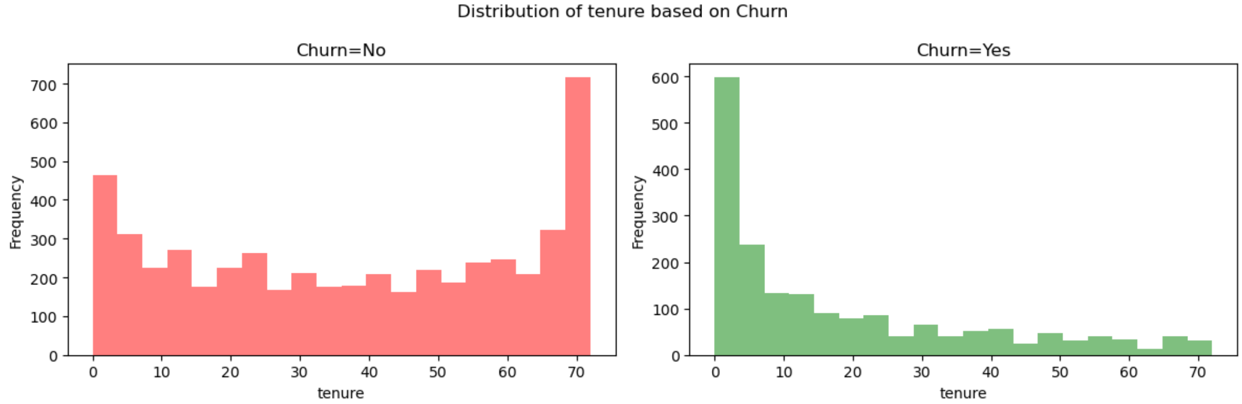
Gender labels do not matter according to churn rate.



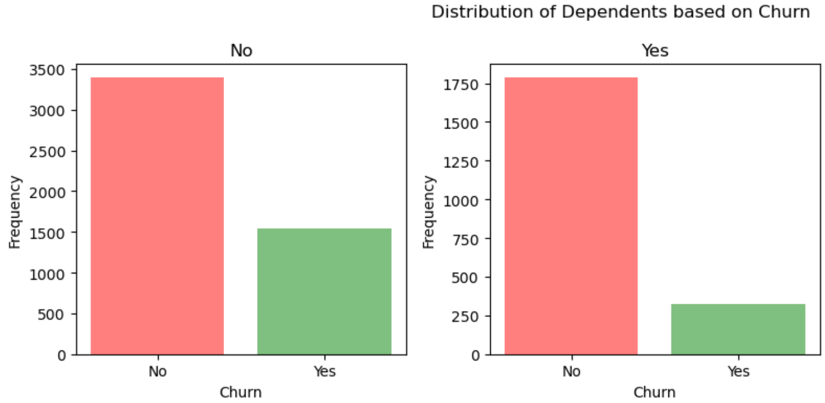
In the churned population bracket, more people churned who are senior citizens rather than seniors in the non-churn bracket.



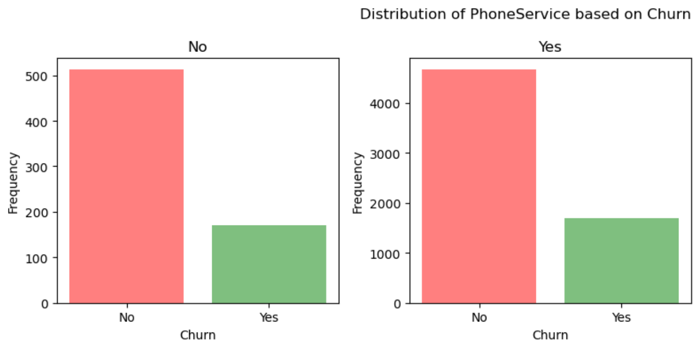
Acquainted population has a lower churn rate rather than a single population.



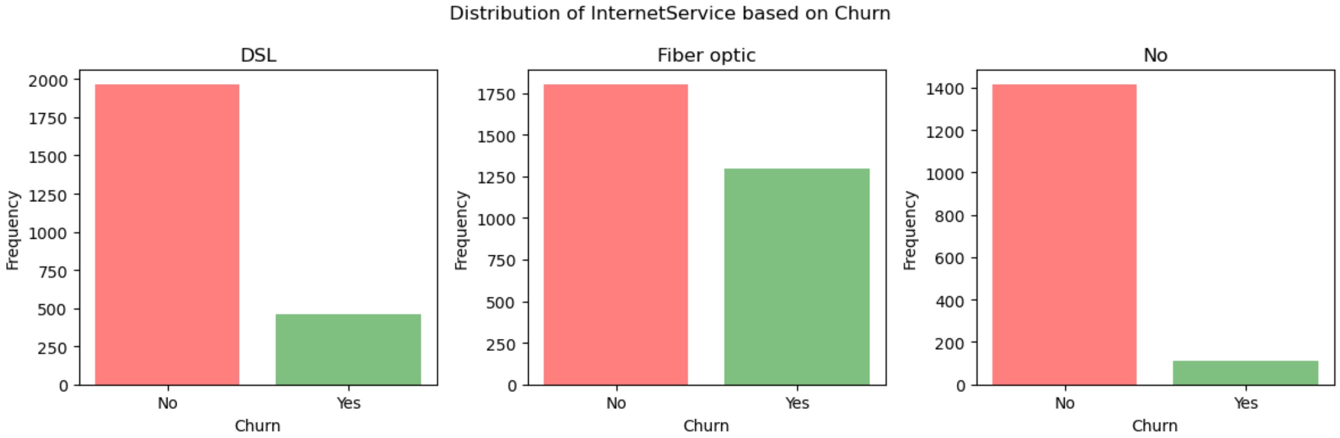
This graph indicated early customers are churning at a very high rate and customers who are staying are not churning, which is indicated by the non-churn graph giving a spike in the farther end.



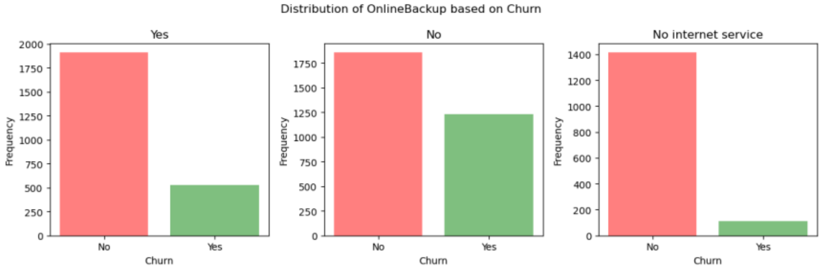
The distribution is not evenly distributed, but probability applies that non-dependent people have a higher churn rate than non-dependent people.



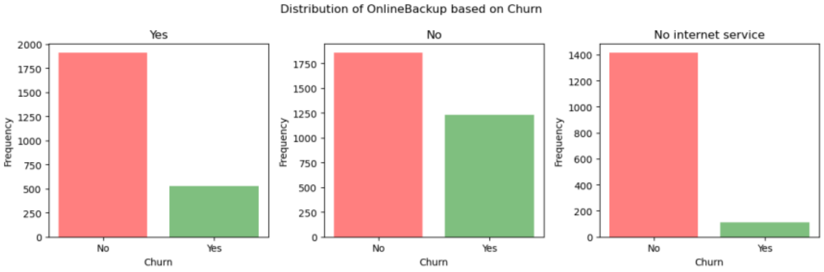
There is no clear correlation between phone service and churn rate.



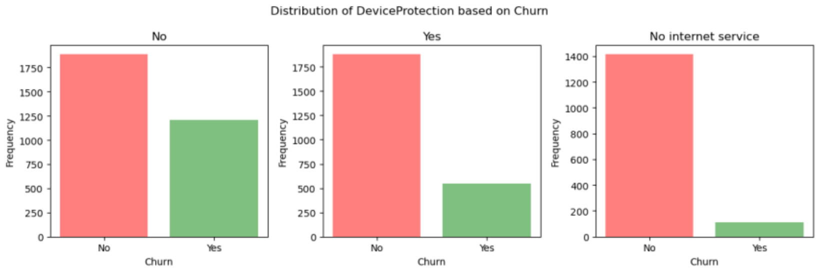
This graph shows Fiber Optic customers are churning at a very high rate then followed by DSL and customers with no internet service are not churning.



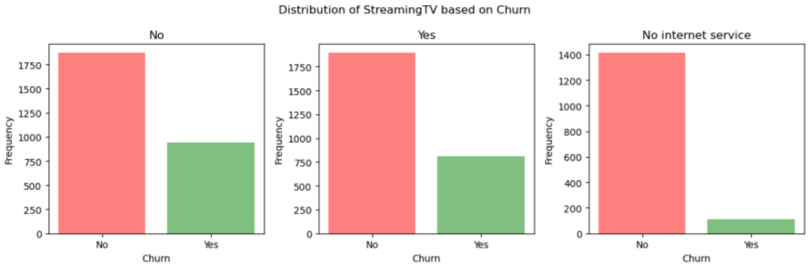
Customers with no Online backup are churning at a higher rate.



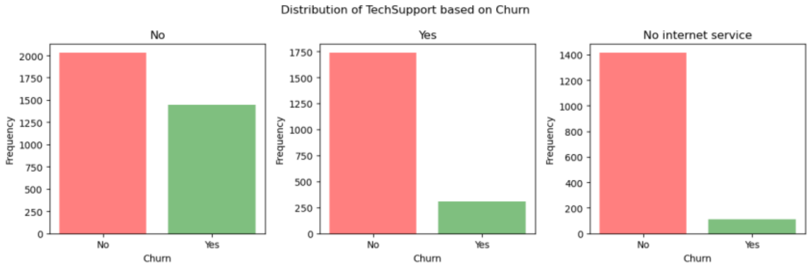
Note: No internet service churn rate is consistent in all these features.



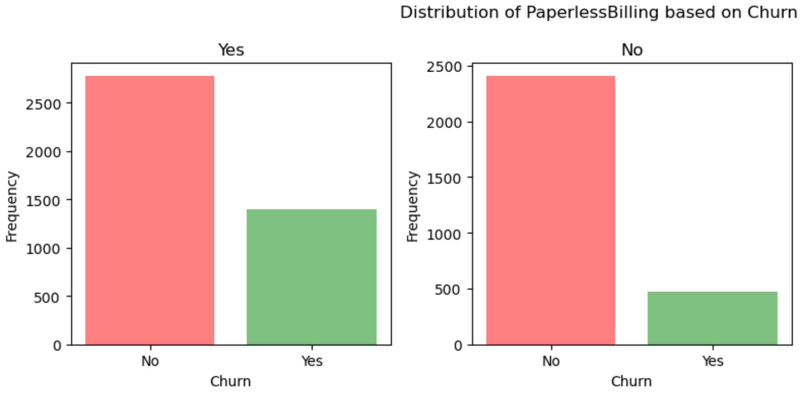
Device protection and tech support with no is churning at a higher rate.



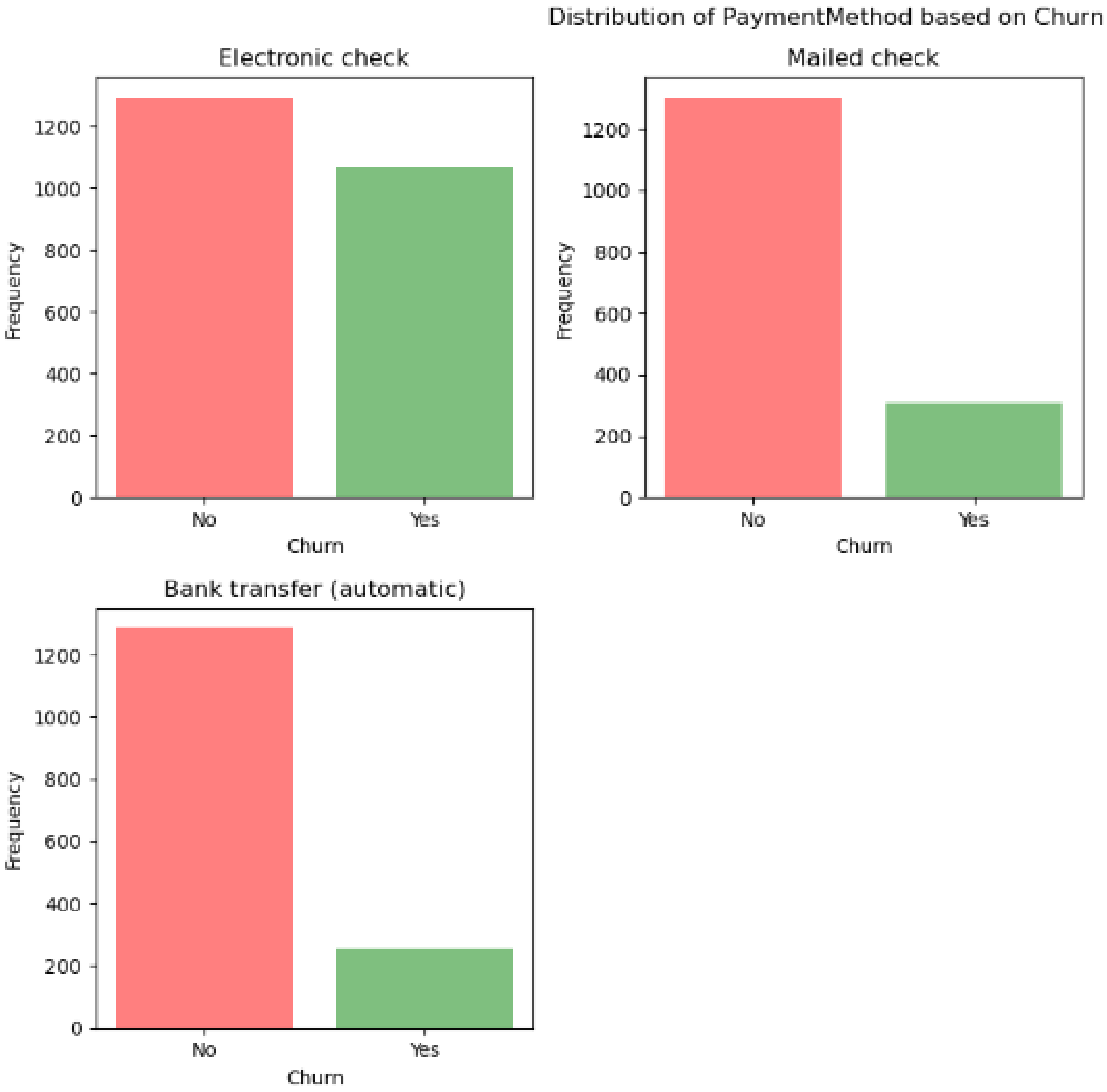
Streaming movies and Streaming TV have no clear correlation with the Churn rate according to this figure.



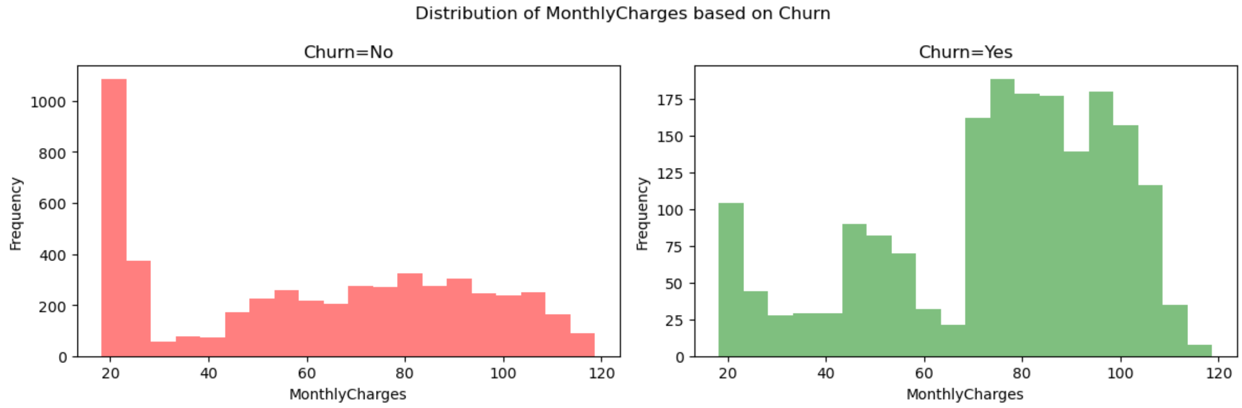
Tech support with No label has a higher Churn rate.



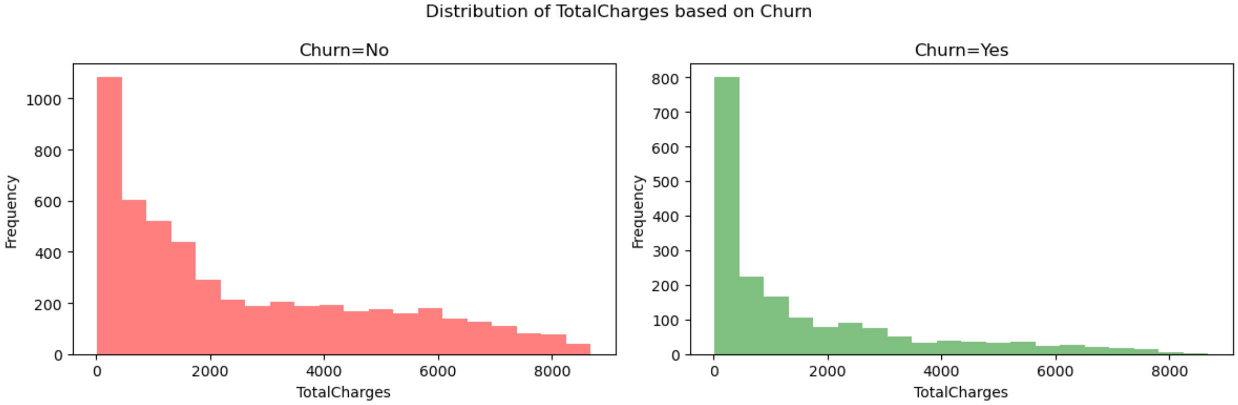
The electronic check has a much higher churn rate rather other ways for payments.



Paperless billing with the label yes has a higher churn rate rather than paper billing.

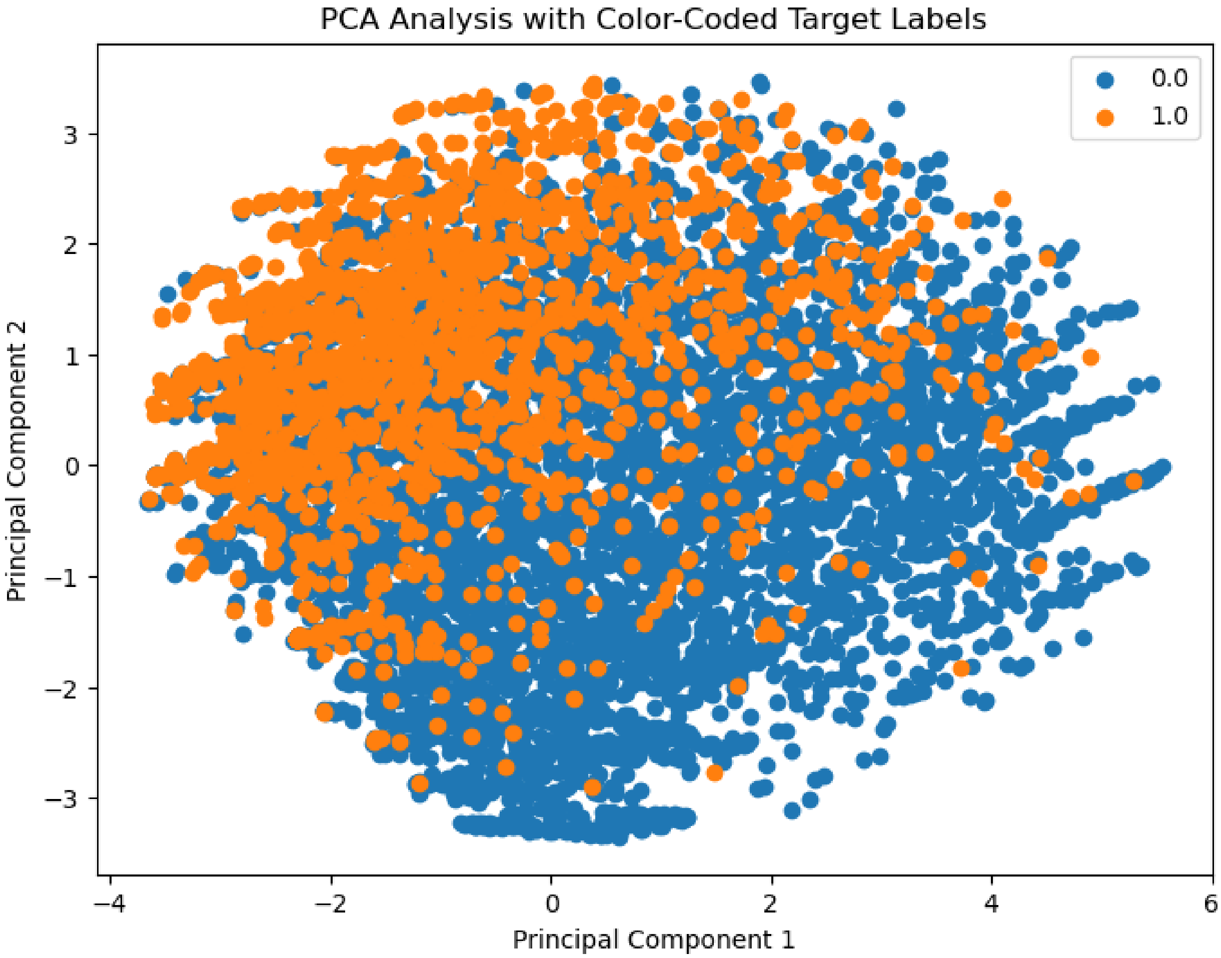


Customers with low monthly charges of around 20$ have a much lower churn rate rather than customers with a 100$ monthly charge indicated by the graph.



Customers who have lower churn rates are evenly distributed in total charges with a linear decline after 2000$.

PRINCIPAL COMPONENT ANALYSIS



Red is non-Churned, and blue is churned customers.

Given the graph, the data will be prone to overfitting and a simpler decision line will produce a better-performing model in production.

FEATURE IMPORTANCE

FEATURE INFORMATION GAIN

Features	**Importance**
Contract	0.105487
tenure	0.082689
OnlineSecurity	0.066758
InternetService	0.055260
TechSupport	0.052365
OnlineBackup	0.047532
MonthlyCharges	0.045400
TotalCharges	0.042493
Payment method	0.042052
DeviceProtection	0.039667
StreamingTV	0.037522
StreamingMovies	0.034750
Partner	0.019030
Dependents	0.009624
SeniorCitizen	0.009358
PaperlessBilling	0.009187
gender	0.004814
PhoneService	0.002932
MultipleLines	0.000000

RANDOM FORREST FEATURE IMPORTANCE

Feature	**Importance**
0 TotalCharges	0.19193094849046255
1 MonthlyCharges	0.1775559148231411
2 tenure	0.15529367962437726
3 Contract	0.0841807405364959
4 PaymentMethod	0.07121601357937742
5 StreamingMovies	0.03219133452582449
6 DeviceProtection	0.030621719728959173
7 gender	0.02747851847460875
8 OnlineSecurity	0.026799019560124836
9 PaperlessBilling	0.02615196334170451
10 Partner	0.023474629993295235
11 MultipleLines	0.023155918923605114
12 OnlineBackup	0.022431895324713683
13 InternetService	0.02224223023325941
14 SeniorCitizen	0.022142845179906646
15 TechSupport	0.019696933640211085
16 Dependents	0.019310331570229423
17 StreamingTV	0.019135805975117954
18 PhoneService	0.004989556474585499

GRADIENT BOOST FEATURE IMPORTANCE

Feature	**Importance**
Contract	0.432800
tenure	0.167341
MonthlyCharges	0.156156
TotalCharges	0.062728
PaymentMethod	0.058495
OnlineSecurity	0.024872
TechSupport	0.019092
PaperlessBilling	0.018992
StreamingTV	0.013783
OnlineBackup	0.009929
StreamingMovies	0.007247
MultipleLines	0.006227
SeniorCitizen	0.005668
DeviceProtection	0.005454
PhoneService	0.004949
InternetService	0.004738
Partner	0.000771
Dependents	0.000569
gender	0.000188

Given three different feature importance graphs Contract is the highest performing feature with the highest information relating to churn rate. Followed by Charges and Payment Methods.

MODEL TRAINING

4 Machine learning models were tried.

- Random Forrest Classifier
- Gradient Boost Classifier
- Logistic Regression Classifier
- Support Vector Machines Classifier

RANDOM FOREST:

With hyperparameter tuning and grid search, the Random Forest model achieved an accuracy of 80.55%. The confusion matrix shows the following results:

Confusion matrix:

Predicted	**No**	**Yes**	**All**
True			
No	941	95	1036
Yes	179	194	373
All	1120	289	1409

- Accuracy: 80.55%
- Precision: 79.55%
- Recall: 80.55%
- F1 Score: 79.70%

GRADIENT BOOST CLASSIFIER:

With hyperparameter tuning, the Gradient Boost Classifier achieved an accuracy of 81.26%. The confusion matrix shows the following results:

Confusion matrix:

Predicted	**No**	**Yes**	**All**
True			
No	940	96	1036
Yes	168	205	373
All	1108	301	1409

- Accuracy: 81.26%
- Precision: 80.41%
- Recall: 81.26%
- F1 Score: 80.58%

LOGISTIC REGRESSION:

The Logistic Regression model achieved an accuracy of 81.97%. The confusion matrix shows the following results:

Predicted	**No**	**Yes**	**All**
True			
No	939	97	1036
Yes	157	216	373
All	1096	313	1409

- Accuracy: 81.97%
- Precision: 81.26%
- Recall: 81.97%
- F1 Score: 81.44%

SUPPORT VECTOR MACHINES:

The Support Vector Machines model achieved an accuracy of 82.19%. The confusion matrix shows the following results:

Predicted	**No**	**Yes**	**All**
True			
No	943	93	1036
Yes	158	308	373
All	1101	313	1409

- Accuracy: 82.19%
- Precision: 81.46%
- Recall: 82.19%
- F1 Score: 81.61%

Overall, the accuracy of each model on the given dataset is as follows:

- Random Forest: 80.55%
- Gradient Boost Classifier: 81.26%
- Logistic Regression: 81.97%
- Support Vector Machines: 82.19%

Support Vector Machines Is the best performer for this dataset.

CONCLUSION

Analysis of the churn rate at a telecom company using exploratory data analysis (EDA) and model training with artificial intelligence. The dataset consisted of 7,043 customer records with 20 features.

EDA revealed several interesting findings: tenure and total charges were correlated, fiber optic customers had higher churn rates, and the electronic check payment method had a higher churn rate compared to other methods.

Three machine learning models were trained and evaluated: Random Forest, Gradient Boosting, Logistic Regression, and Support Vector Machines (SVM).

Based on the results, Support Vector Machines performed the best in predicting churn rate. These findings provide valuable insights for the telecom company to develop effective strategies to reduce customer churn.