Customer Churn Prediction

ANALYSIS ON CUSTOMER CHURN RATE

Laiba Gohar | 11-07-2023

Contents

Data Collection

The Telco customer churn data contains information about a fictional telco company that provided home phone and Internet services to 7043 customers in California. It indicates which customers have left, stayed, or signed up for their service.

Customer churn is the number of customers that stopped using your company's product or service during a certain time frame.

Pre Processing

For the preprocessing part, when the features where looked into, the customer ID seemed not useful for data analysis so it was dropped.

- CustomerID column was dropped.
- Few rows were dropped (11 rows)

STANDARDIZATION

The dataset has been standardized (z-score normalization) to transform the data to have a mean of zero and a standard deviation of one.

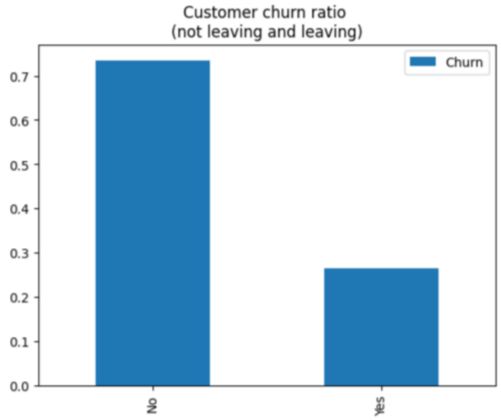
Exploratory Data Analysis

VARIABLE IDENTIFICATION

After dropping the customerID column there were 20 features left. Out of them two were of data type int64, two were of type float64 and rest were object type. The features which had object data types only InternetService, Contract and PaymentMethod had some strings rest had only yes or no.

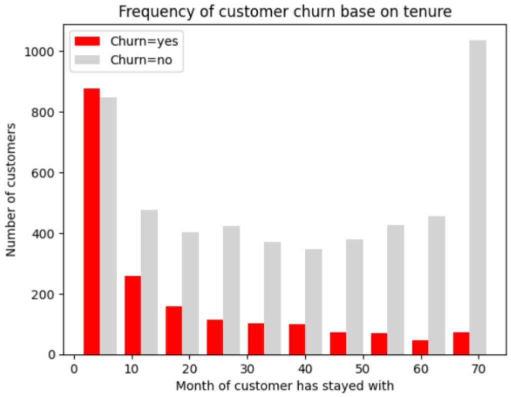
UNIVARIATE ANALYSIS

The churn ration was identified using univariate analysis. There ia a total of 7032 rows in dataset out of which 73.42% did not churned and 26.58% churned.

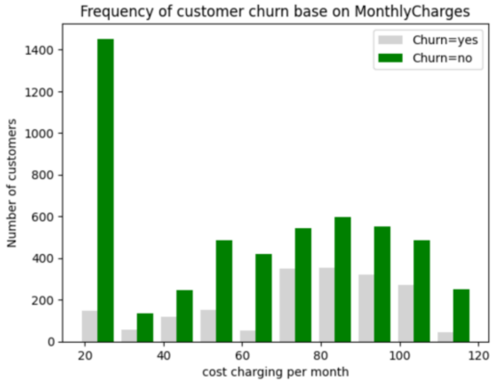


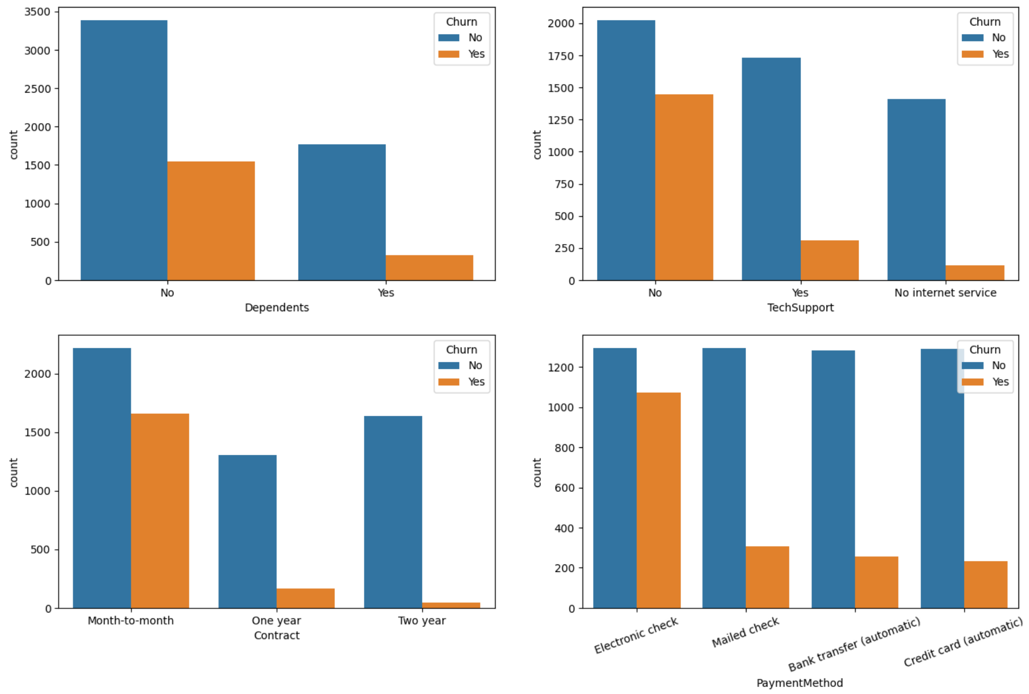
BIVARIATE ANALYSIS

Here we can see that the people are leaving more in the start of the tenure rather than as time passes.

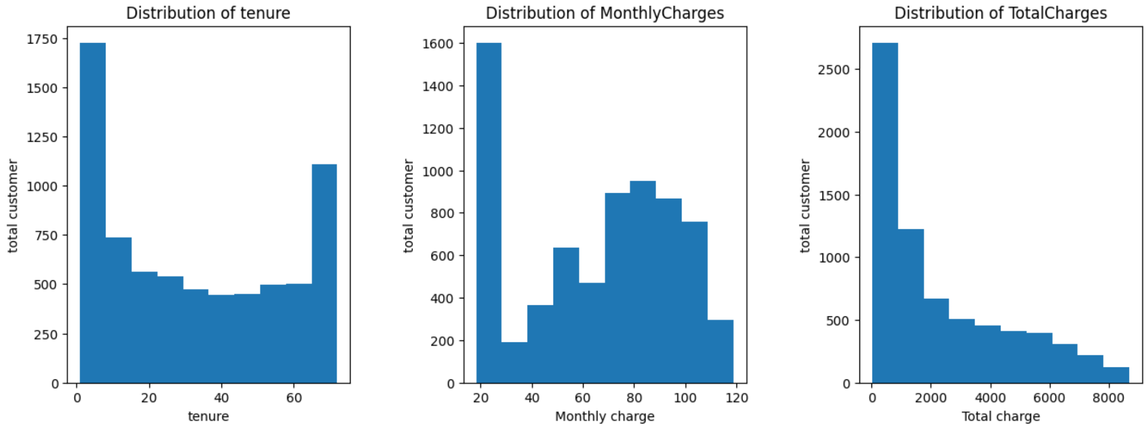


Now for this analysis people who have been charged <30 per month are staying more.

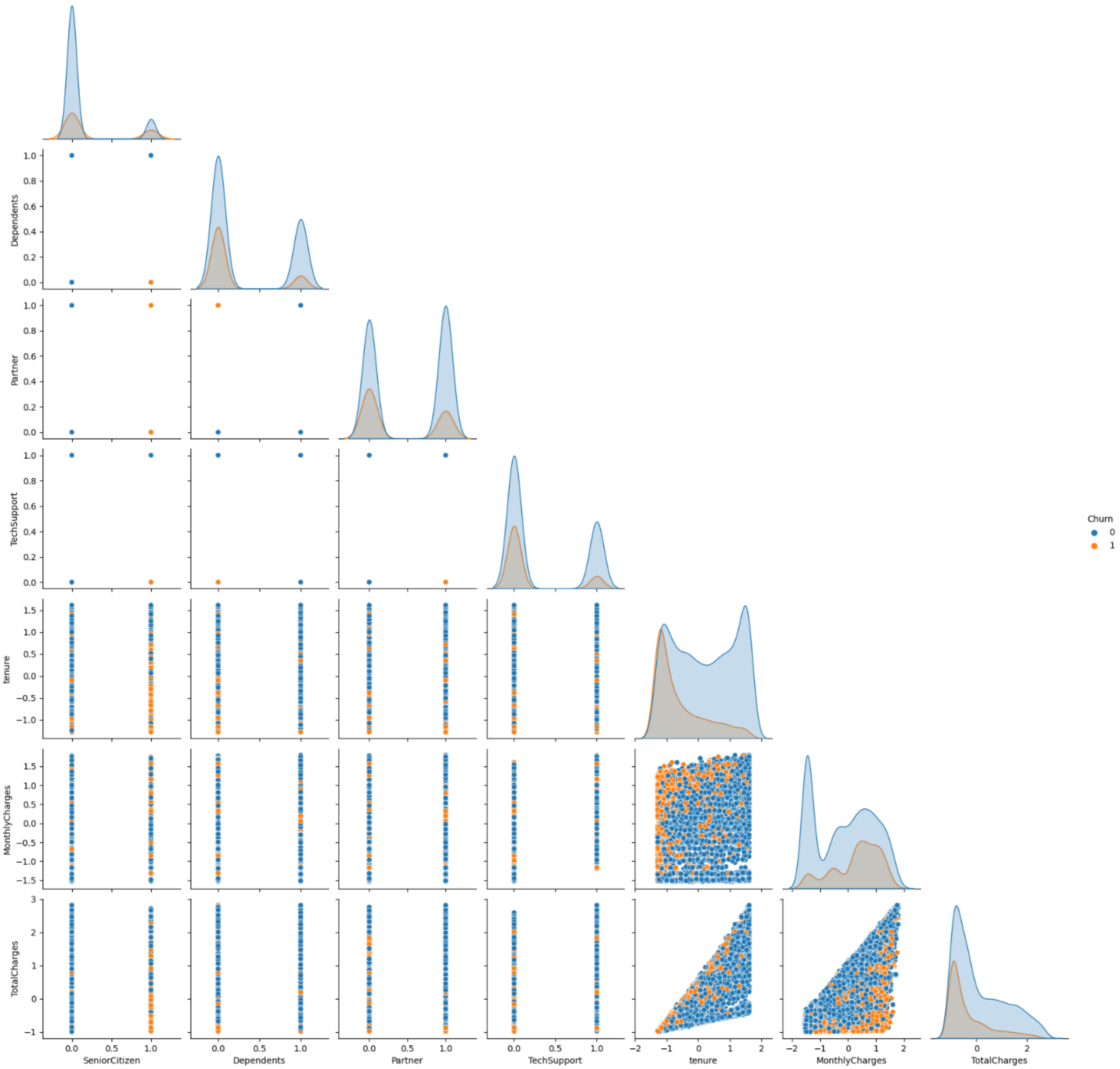




- Dependents: People who have no dependents are not churning out.
- TechSupport: People with no technical support are tend to leave more than that who have any sort of tech support.
- Contract: Month-to-Month contracts are apparently making people easy to decide to leave.
- PaymentMethod: the people with payment method of electronic checks are tend to leave more than any other payment method.



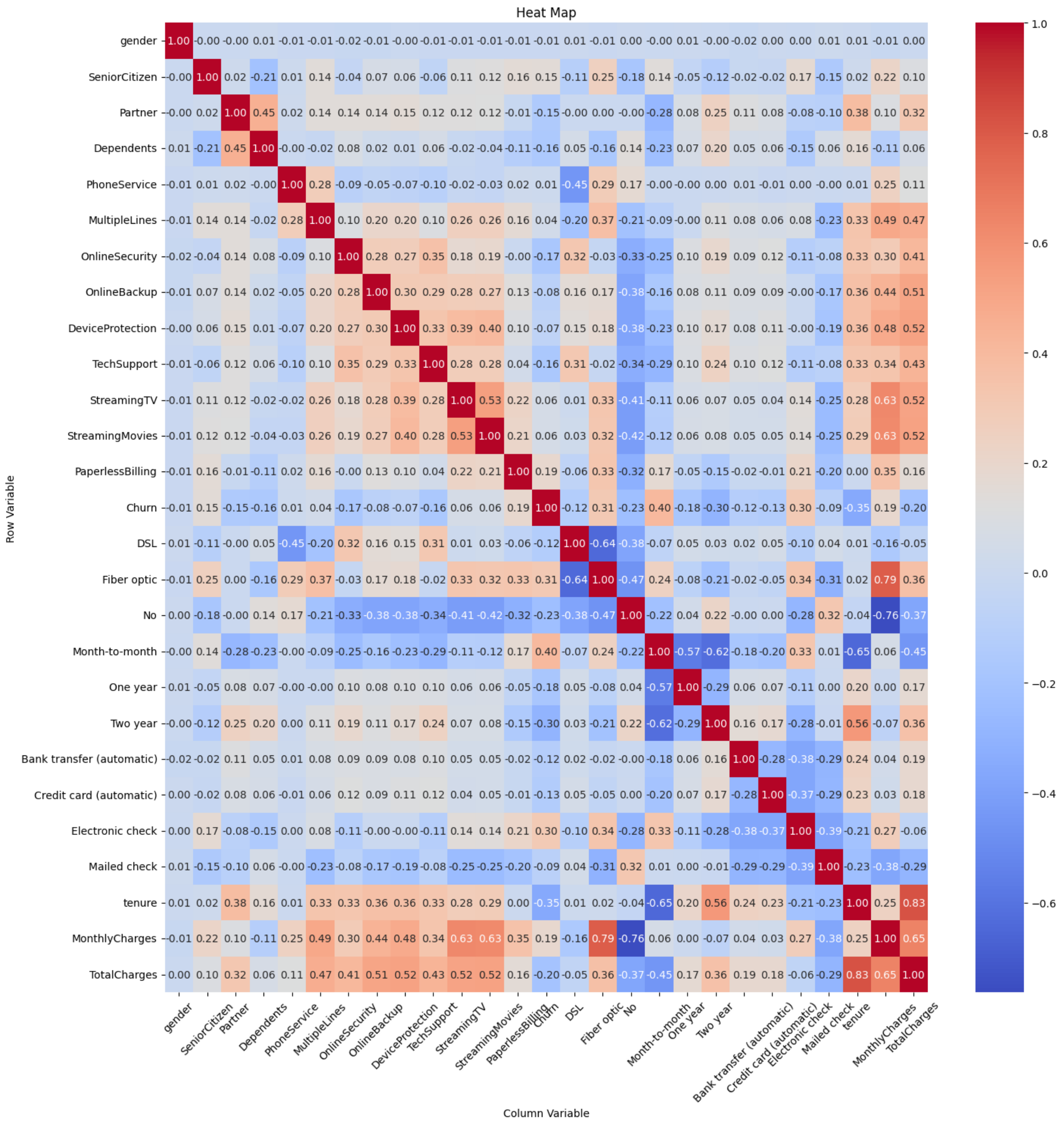
Now we can clearly see people are choosing the options with less monthly and total charges. But if we have a look at the tenure to the total customer graph, we can see people are leaving earlier in their tenure with a major dip in the 1st month. There must be something that is causing them to leave this early.



After calculating the correlation of features, few features were chosen who have high impact on the churn out rate rather than who have less impact.

This made it a little better view into looking at the features who are impacting the churn rate.

MULTIVARIATE ANALYSIS



Here the most correlated features are tenure with total charges, the 2nd most correlated are monthlyCharges with the people who are using fiber optics as connection.

Model Selection

DATA SPLITTING

The data was firstly split into training and testing data using sklearn library.

SUPPORT VECTOR REGRESSION (SVR)

Implementing the SVR model yielded an accuracy of

Validation RMSE: 0.399526209114001

LINEAR REGRESSION

The integration of the linear regression led to a noticeable enhancement in accuracy rather than SVR

Validation RMSE: 0.4636832023350158

GRADIENT BOOST FEATURE

Upon implementing the gradient boost feature model, the resulting accuracy showed notable improvement.

Validation Accuracy: 0.7732764747690121

LOGISTIC REGRESSION

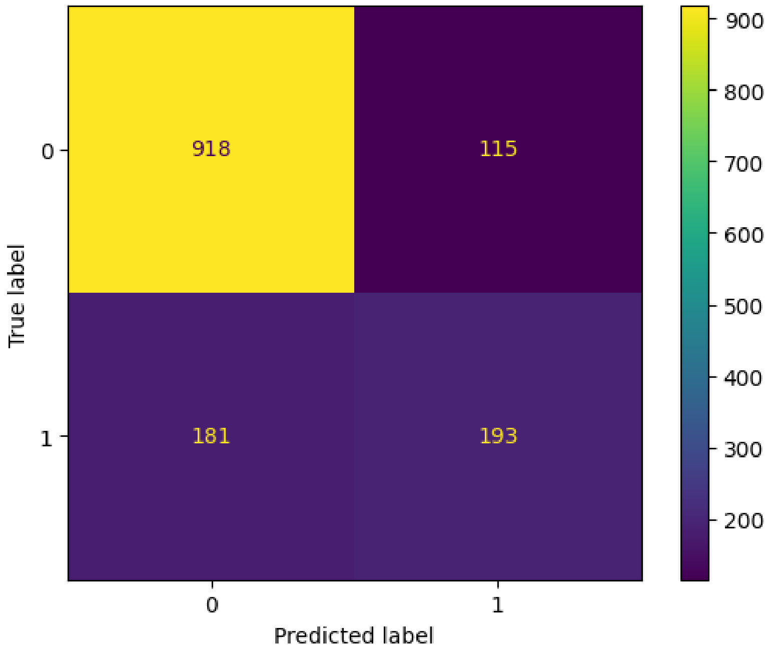
The accuracy of the model significantly increased following the implementation of the logistic regression technique.

Validation Accuracy: 0.7896233120113717

Here the best accuracy was being given by the logistic regression technique. The attempt to improve it using both SMOTE and Pipelining resulted in the decrease of the accuracy so its is not recommended.

Model Performance

CONFUSION MATRIX



CLASSIFICATION REPORT

	Precision	*Recall*	*F1-score*	*Support*
0	0.84	0.89	0.86	1033
1	0.63	0.52	0.57	374
Accuracy			0.79	1407
Macro Avg	0.73	0.70	0.71	1407
Weighted Avg	0.78	0.79	0.78	1407