INNOMATICS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

TELECOM CHURN PREDICTION

About me

- Vijay sada
- B-tech in Electronics and Communications Engineering.
- Fresher
- LinkedIn - https://www.linkedin.com/in/vijay-sada/
- GitHub - https://github.com/vijaysada

INNOMATICS RESEARCH LABS

Introduction

The project aims to develop a predictive model for telecom churn prediction, leveraging data analytics techniques to anticipate and mitigate customer attrition in the telecom industry. By analyzing historical customer data, the objective is to identify patterns and factors influencing churn behavior, ultimately aiding telecom companies in implementing targeted retention strategies.

Telecom churn prediction is crucial for businesses as it directly impacts revenue and profitability. By proactively identifying customers at risk of churning, companies can take preemptive measures to retain them, such as offering personalized incentives, improving service quality, or implementing loyalty programs. This not only helps reduce customer churn rates but also enhances customer satisfaction and loyalty, leading to long-term business growth and sustainability in a competitive market landscape.

Data Overview

The dataset was provided to us by the client.

The dataset includes the following features:

Churn: Indicates whether a customer left within the last month.

Services: Details the services each customer has signed up for, such as phone, internet, and additional features like online security and streaming TV.

Customer Account Information: Includes tenure, contract type, payment method, billing preferences, and billing charges.

Demographic Information: Covers gender, age range, and whether the customer has partners and dependents.

S No	Feature Name	Data Type
1	Gender	Object
2	Senior Citizen	Int64
3	Partner	Object
4	Dependents	Object
5	Tenure	Int64
6	Phone Service	Object
s7	Multiple Lines	Object
8	Internet Service	Object
9	Online Security	Object
10	Online Backup	Object
11	Device Protection	Object
12	Tech Support	Object
13	Streaming TV	Object
14	Streaming Movies	Object
15	Contract	Object
16	Paperless Billing	Object
17	Payment Method	Object
18	Monthly Charges	float64
19	Total Charges	Object
20	Customer ID	Object
21	Churn	Object

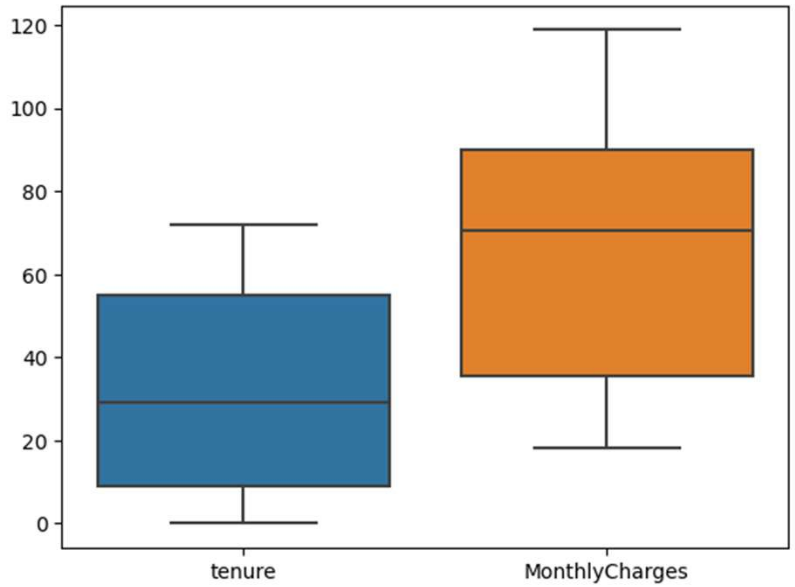
INNOMATICS RESEARCH LABS

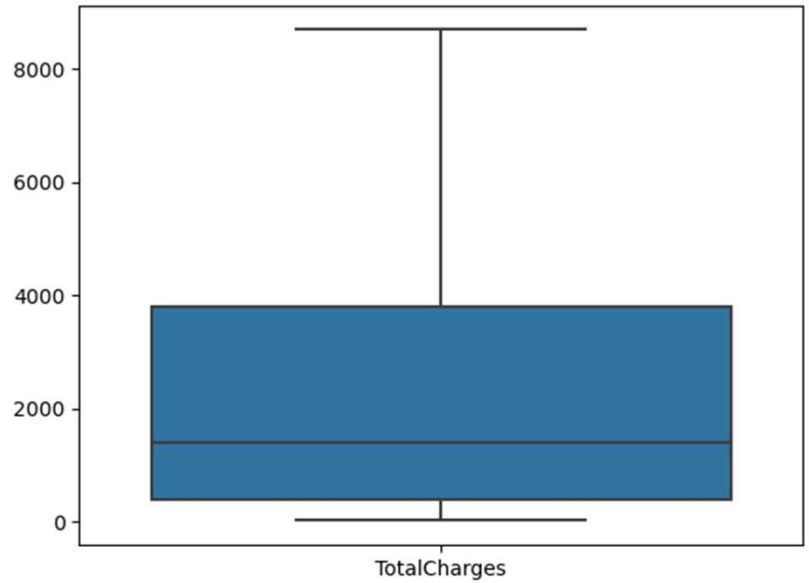
Pre-Processing

For the project, the following preprocessing steps were performed:

- **Data Cleaning:** Identified and handled inconsistencies like missing values, errors like wrong data type, and no outliers were found in the dataset.
- **Handling Missing Values:** Addressed missing values in the total charges column using imputation techniques like median and corrected the data type errors to ensure that the data is ready for accurate analysis.
- **Standardization:** Standardized the numerical columns to ensure they were on a similar scale, ready to be use for the model training.
- **Handling Categorical Variables:** Encoded categorical variables into numerical format using one-hot encoding, making them suitable for modeling.

Exploratory Data Analysis (EDA)

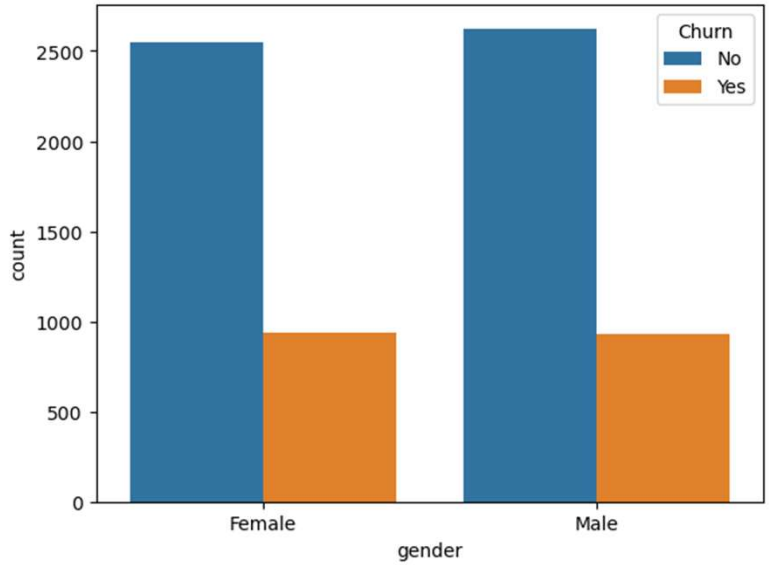




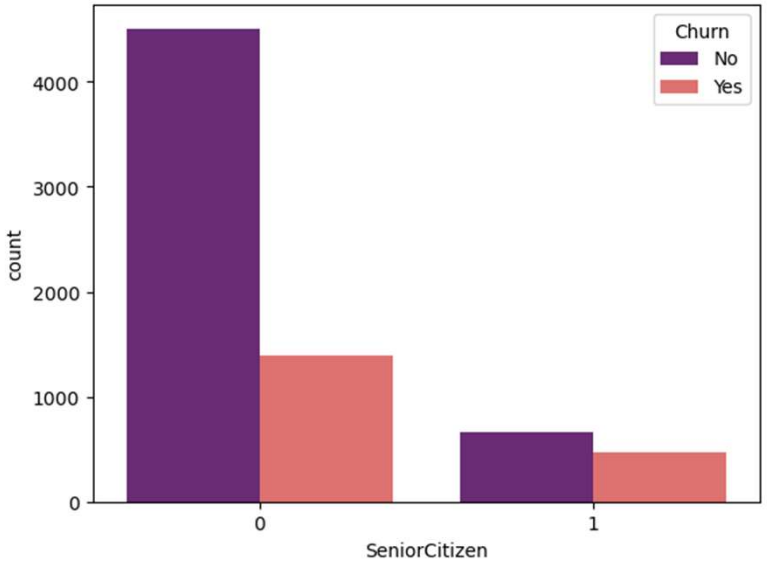
From the above boxplots we can observe that there aren't any outliers visible in the numerical columns tenure, monthly charges and total charges.

INNOMATICS RESEARCH LABS

EDA (Cont.):



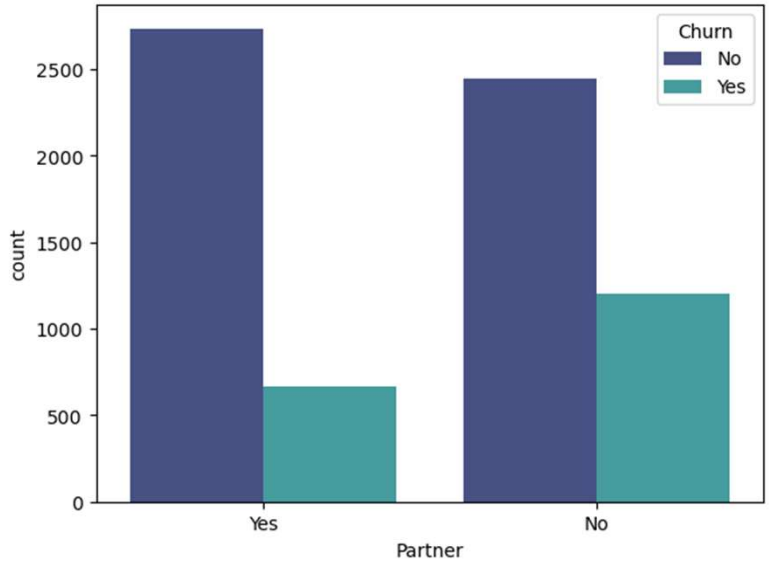
From the above plot on the churn distribution by gender we can observe that females tend to churn when compared to men.



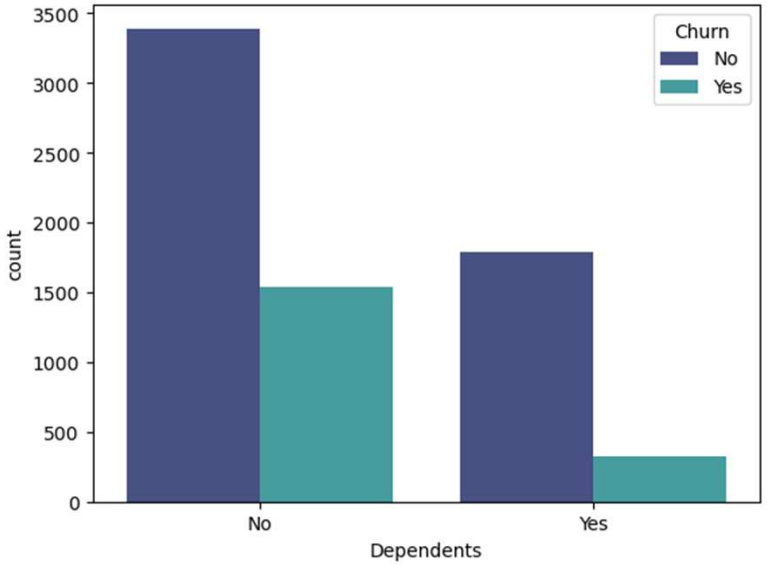
From the above we can say that the old people are less likely to churn in comparison to younger people

INNOMATICS RESEARCH LABS

EDA (Cont.):



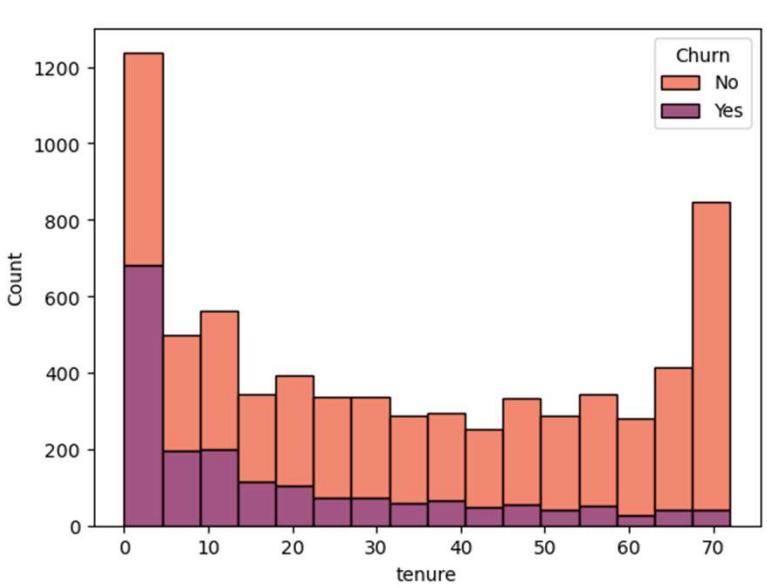
From the above we can say that people without a partner tend to churn more than people with a partner.

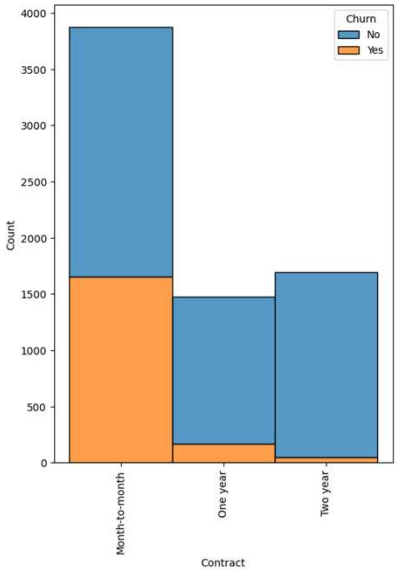


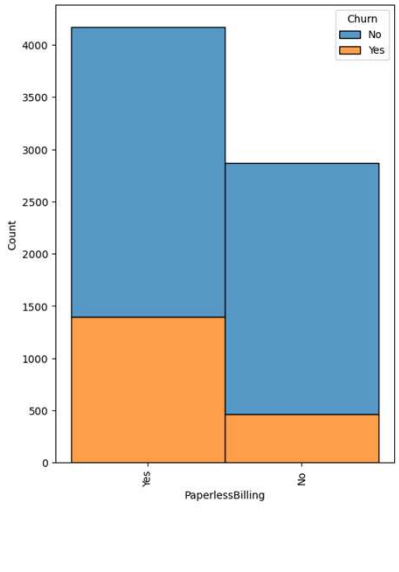
From the above we can say that people with no dependents are most likely to churn compared to people with dependents.

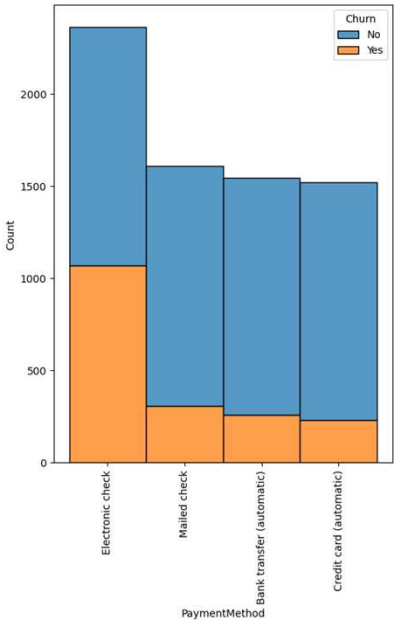
INNOMATICS RESEARCH LABS

EDA (Cont.):









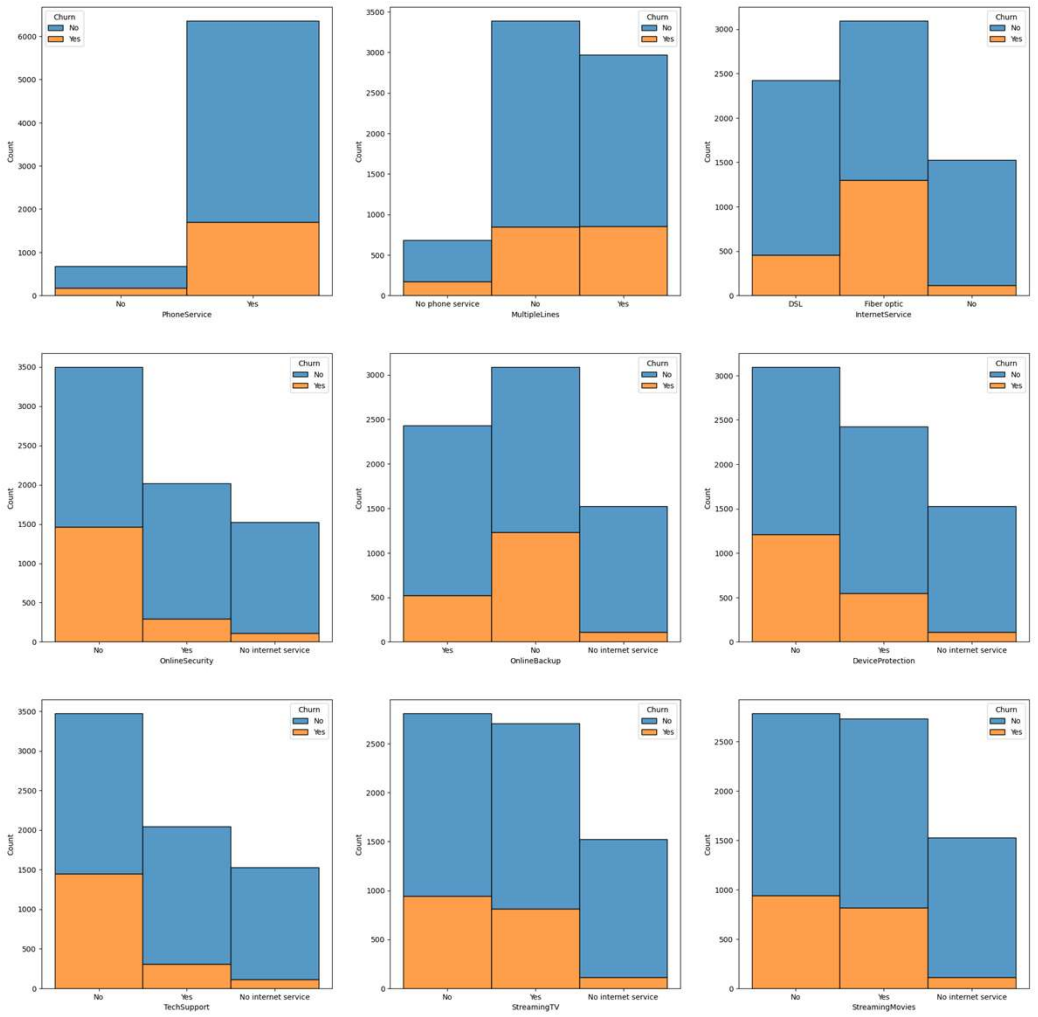
- **From the above we can say that people tend to churn less the longer they use the service.**
- **people with a month to month contract are most likely churn compared to those with a year or more contract**
- **Customers with paperless billing have higher churn rates.**
- **Customers who use Electronic check have a higher churn rates in comparison with other payment methods**

INNOMATICS RESEARCH LABS

EDA (Cont.):

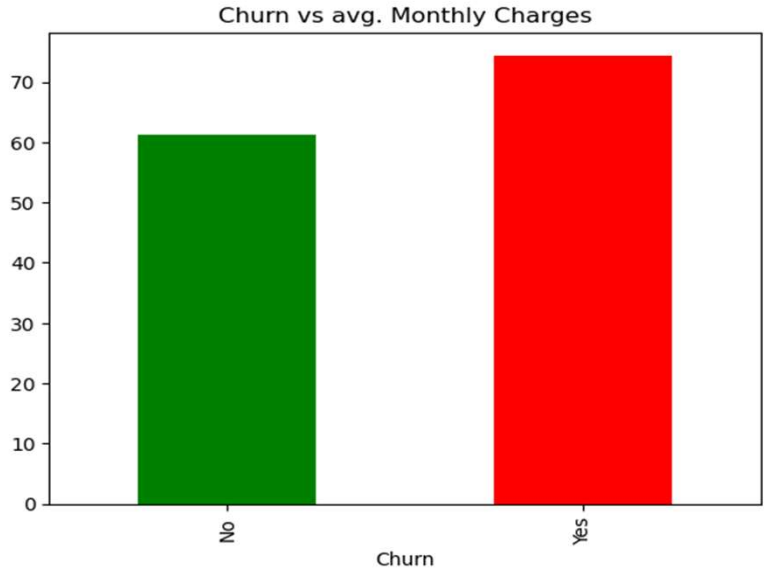
From the multiple plots we can say that

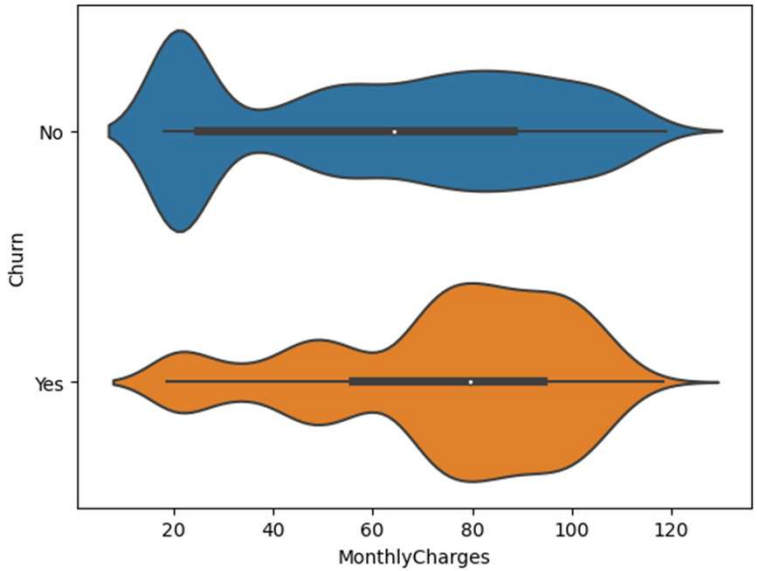
- **There is a higher churn rate for customers who have the phone service.**
- **Customers with internet service fiber optic have a higher churn rate compared with DSL and No. as internet helps with most of the activities.**
- **Customers who don't have access to tech support appear to churn more than those who have the service available.**
- **Customers without services like online security, online backup and device protection have a higher churn rate.¶**



INNOMATICS RESEARCH LABS

EDA (Cont.):

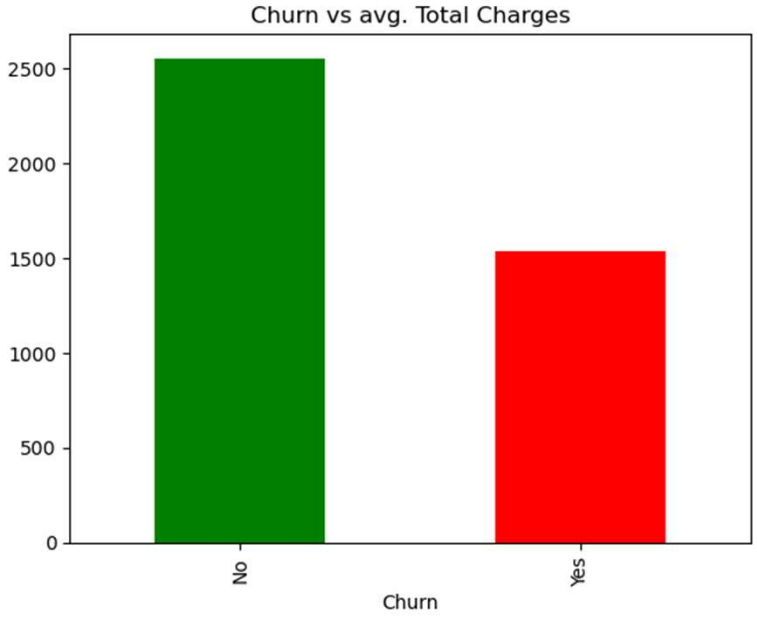


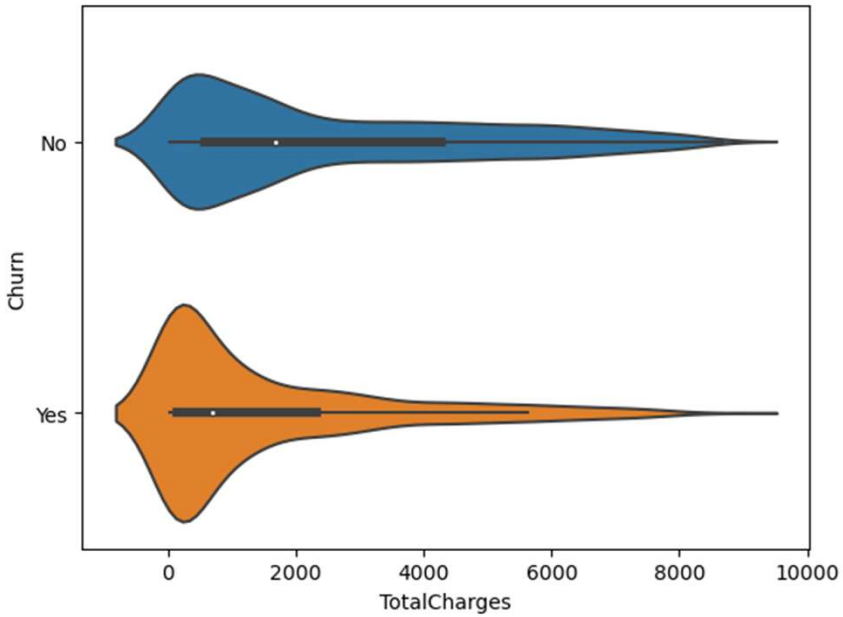


We can observe that majority of the people who churn are paying higher monthly charges whereas those who pay less are unlikely to churn in comparison. From the above we can say that people who pay more in monthly charges are likely to churn more.

INNOMATICS RESEARCH LABS

EDA (Cont.):





We can observe that most of the people who pay more of total charges on an average are not churn. From the above we can say that people who pay more in Total charges are likely to churn less as they might have subscribed to multiple services.

INNOMATICS RESEARCH LABS

Model Building

For the classification task of churn prediction, the below machine learning models were used:

- **K-Nearest Neighbors (KNN) Classifier:** A non-parametric method that classifies data points based on the majority class among their k-nearest neighbors, which can be effective for capturing local patterns in the data.
- **Logistic Regression:** A classic binary classification algorithm that models the probability of a customer churning based on input features. It's interpretable and efficient for linearly separable data.
- **Decision Trees:** These models partition the feature space into regions and make predictions based on majority voting within each region. They're intuitive and can capture non-linear relationships between features and the target variable.
- **Random Forest:** An ensemble method that builds multiple decision trees and combines their predictions to improve accuracy and robustness. It's effective for handling complex datasets and mitigating overfitting.
- **Support Vector Machines (SVM):** SVMs aim to find the hyperplane that best separates the data points into different classes. They're effective in high-dimensional spaces and suitable for datasets with clear margins of separation.

Model Training

The model training process involved:

- Splitting the dataset into training and testing sets (75:25 Ratio) to train the models on a subset of the data and evaluate their performance on unseen data.
- Training each model using the training set, where the model learns patterns and relationships between features and the target variable (churn).

S No	Type of Problem	Algorithm Name
1	Classification	KNeighbors Classifier
2	Classification	Logistic Regression
3	Classification	SVC
4	Classification	Random Forest Classifier
5	Classification	Decision Tree Classifier

Model Evaluation

Evaluation metrics employed to assess model effectiveness was accuracy:

- **Accuracy:** The accuracy metric measures the proportion of correctly classified instances out of total instances, providing a straightforward evaluation of overall model performance. It is commonly used in classification tasks to assess the model's ability to correctly predict the class labels of the dataset.

S No	Algorithm Name	Metric Score
1	KNN	0.768881
2	Logistic Regression	0.816014
3	SVC	0.809199
4	Random Forest Classifier	0.797842
5	Decision Tree Classifier	0.733674

Challenges Faced

While identifying the missing values inside the total charges column, we found that it wasn't an empty string but a single space character " " that was inside the data which made the column into object datatype.

Conclusion

From the above Accuracy results we can observe that the Logistic Regression model has the highest accuracy when compared with the other models. We can see that the Logistic Regression model has the accuracy of 0.816. Therefore, we can say that the Logistic Regression appears to be the best model for classification task on our dataset based on the evaluation metrics.

THANK YOU

INNOMATICS
RESEARCH LABS