

ISI HYD

RKMVERI

DECODING DEPARTURES: A CLASSIFICATION

JOURNEY INTO CUSTOMER CHURN

A MACHINE LEARNING PERSPECTIVE

SOHAM PAUL – ANIRBAN DEY



CONTENT

01

Introduction

02

SOP

03

Data Prep & Cleaning

04

EDA

05

Feature Engineering

06

Modelling

07

Evaluation

08

Challenges

09

Future Work

10

Conclusion

INTRODUCTION



Customer churn is a critical issue for telecom companies, as it can lead to significant revenue loss and damage to the company's reputation. ABC Voice, a telecom company that has been in operation since 1999, is no exception to this challenge.

Customer churn can be defined as the rate at which customers stop using a company's services. For telecom companies, this typically means the number of customers who cancel their subscriptions or switch to another provider. There are many reasons like:

- Dissatisfaction with the service
- Better offers from competitors
- Life changes



PROBLEM STATEMENT

ABC Voice is facing a growing challenge in retaining its telecom customers. With increasing competition and changing customer preferences, the company is experiencing a high rate of customer churn. The objective of this project is to develop a predictive model that can accurately identify customers at risk of churning. By doing so, ABC Voice aims to proactively implement customer retention strategies and reduce churn rates, ultimately improving customer satisfaction and the company's profitability. This project will leverage machine learning techniques to analyze historical customer data and build a predictive model capable of identifying potential churners, enabling ABC Voice to take timely actions to retain these customers.

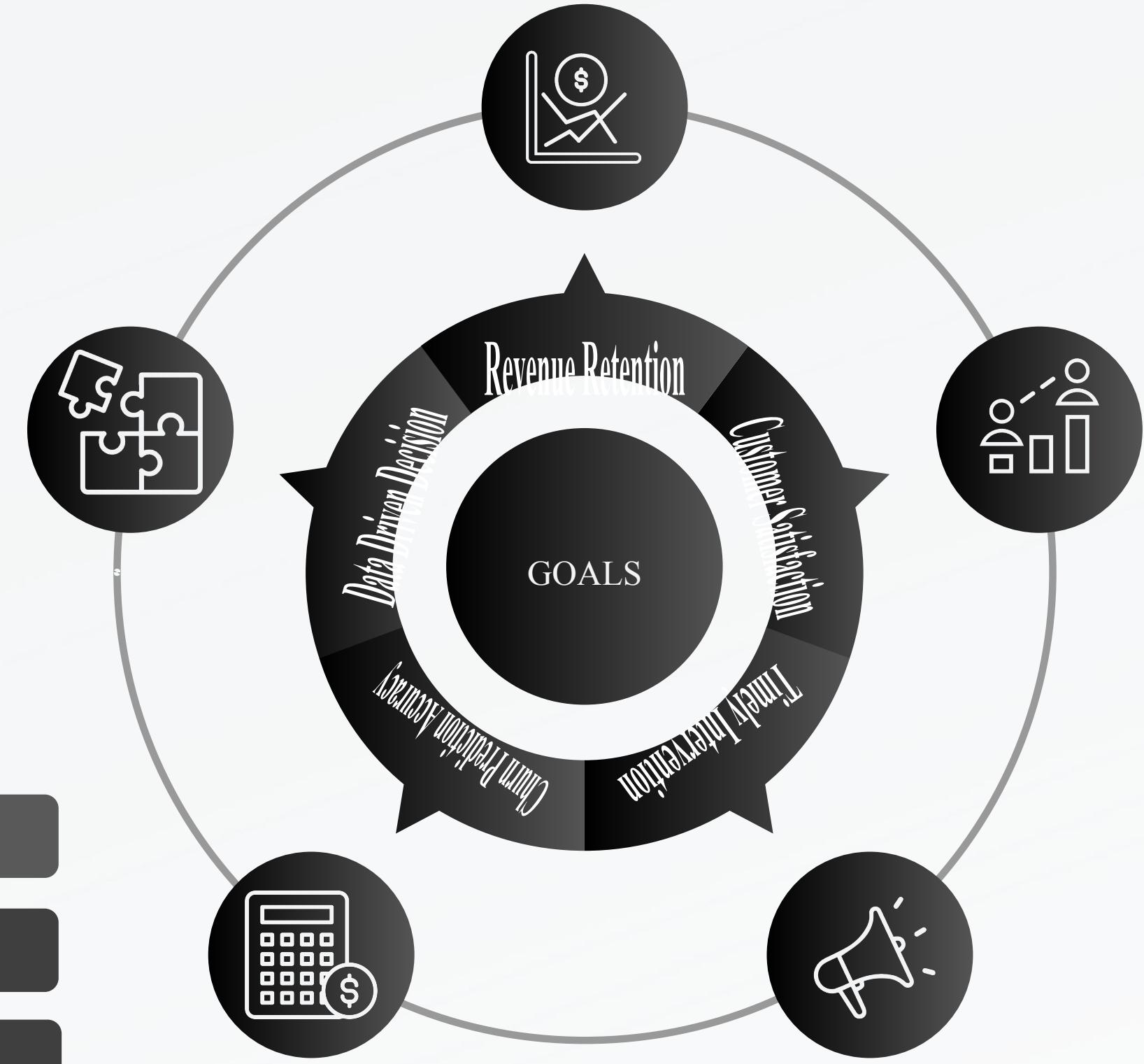
BENEFITS

ENHANCED PROFITABILITY

PROACTIVE CHURN MANAGEMENT

COMPETITIVE ADVANTAGE

ENHANCED SERVICE QUALITY



DATA PREP & CLEAN



Data Collection



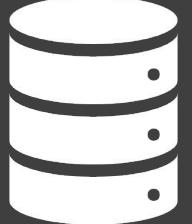
Collected Customer Churn Data containing info on all 7,043 customers from a Telecom company (say, ABC Voice) in California in Quarter 2 of 2022



Data Reduction



Dropped irrelevant columns from the dataset like Customer ID, Total Refunds, Zip Code, Latitude, Longitude, Churn Category & Churn Reason



Data Cleaning



Eliminated rows having Null Values for several columns. E.g., in 'Avg Monthly GB Download' column, 21.6% of the data instances contain NULL values



Data Categorization



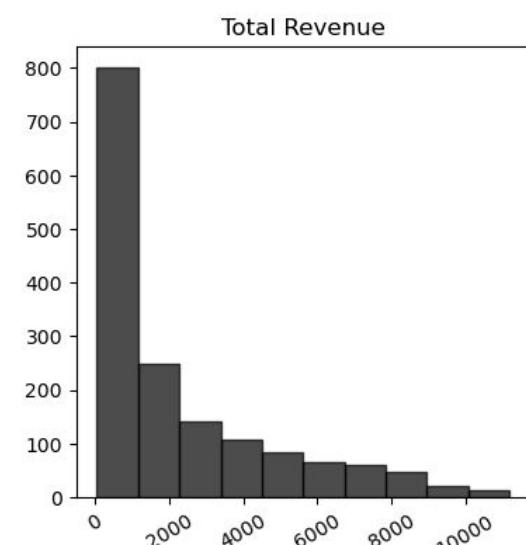
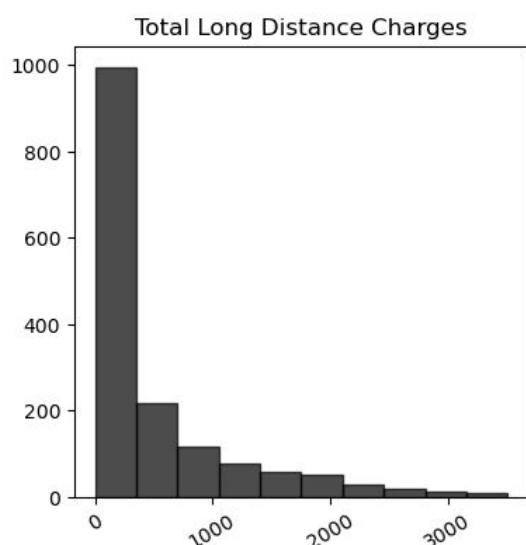
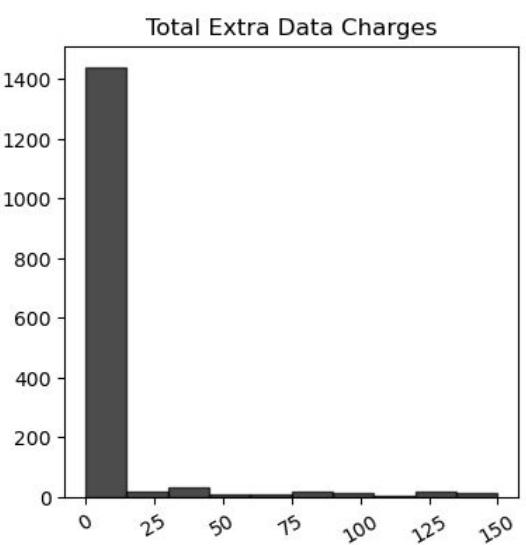
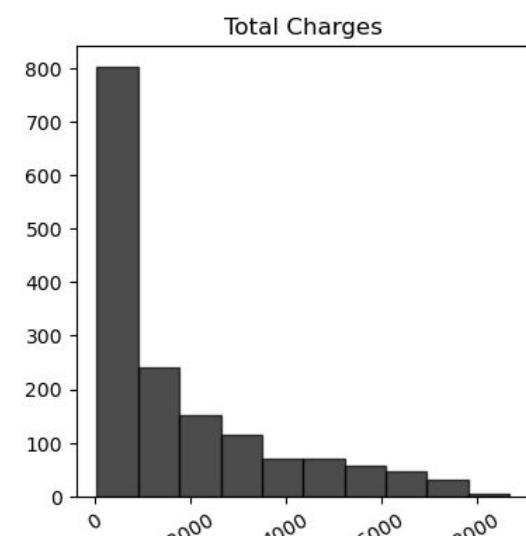
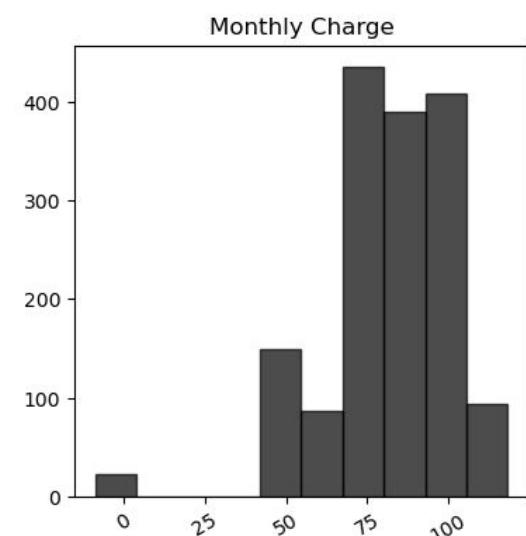
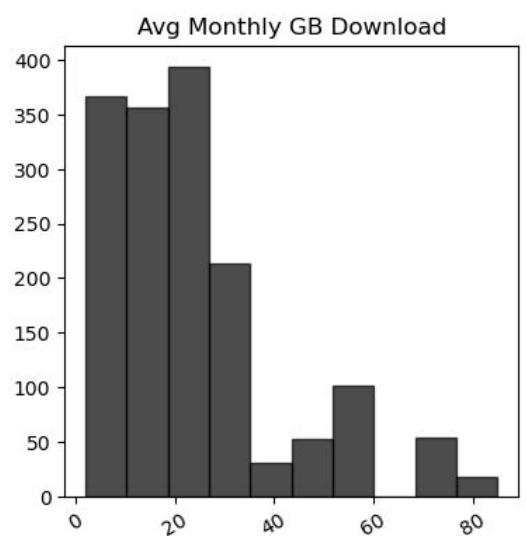
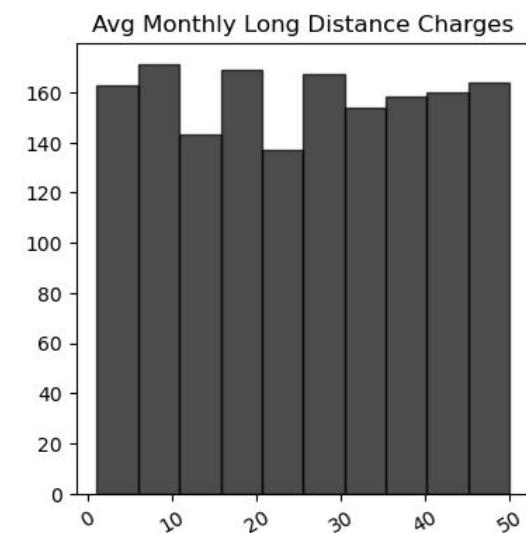
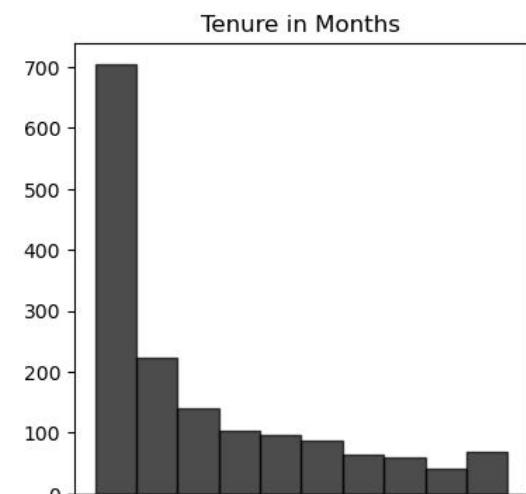
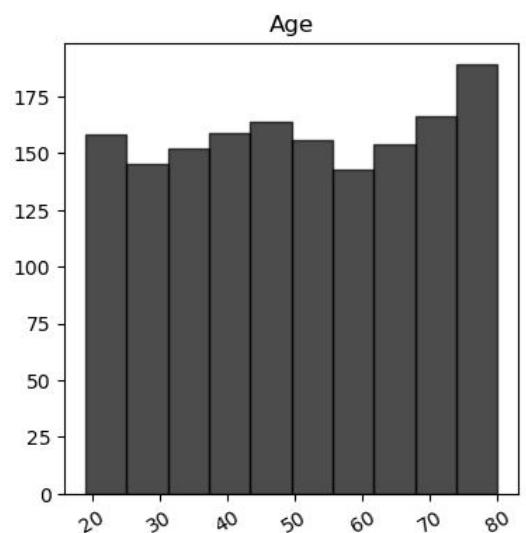
Categorized columns in Attributes, Discrete Variables & Continuous Variables for Analysis purpose



EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) and data visualization provide a comprehensive understanding of the dataset by revealing hidden patterns and relationships. They aid in identifying factors contributing to customer churn and guide feature selection. Histograms and box plots for continuous variables highlight data differences between churned and non-churned customers. Bar plots for categorical variables offer insights into factors influencing churn, enabling targeted retention strategies. Additionally, data visualization simplifies complex information, facilitating data-driven decisions and improving customer satisfaction.





HISTOGRAM OF DISTRIBUTION OF CHURNED CUSTOMERS



Customers in their late 80s & customers with shorter tenure have high churn; But different levels of avg monthly long distance charges has no effect on churn.

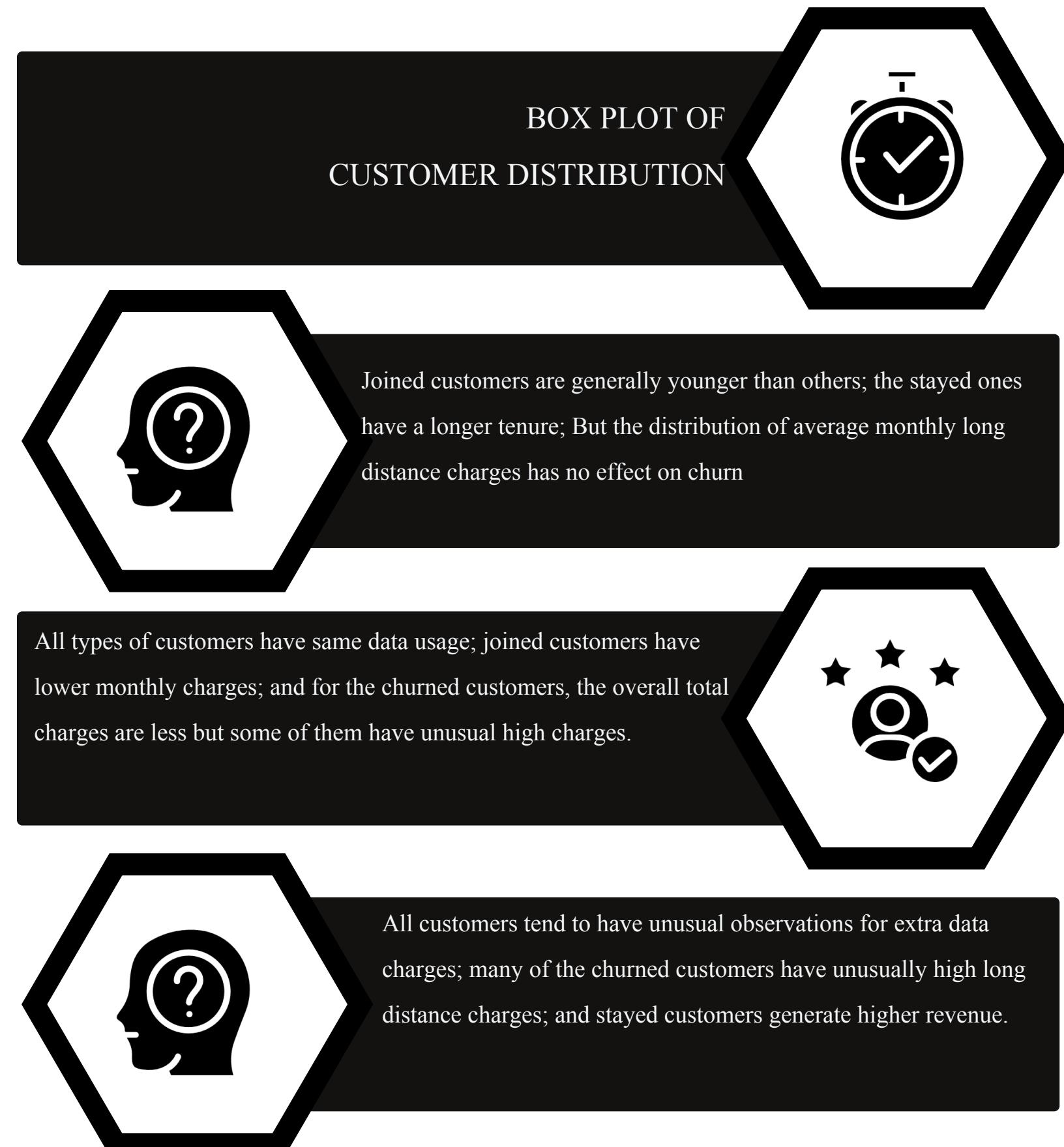
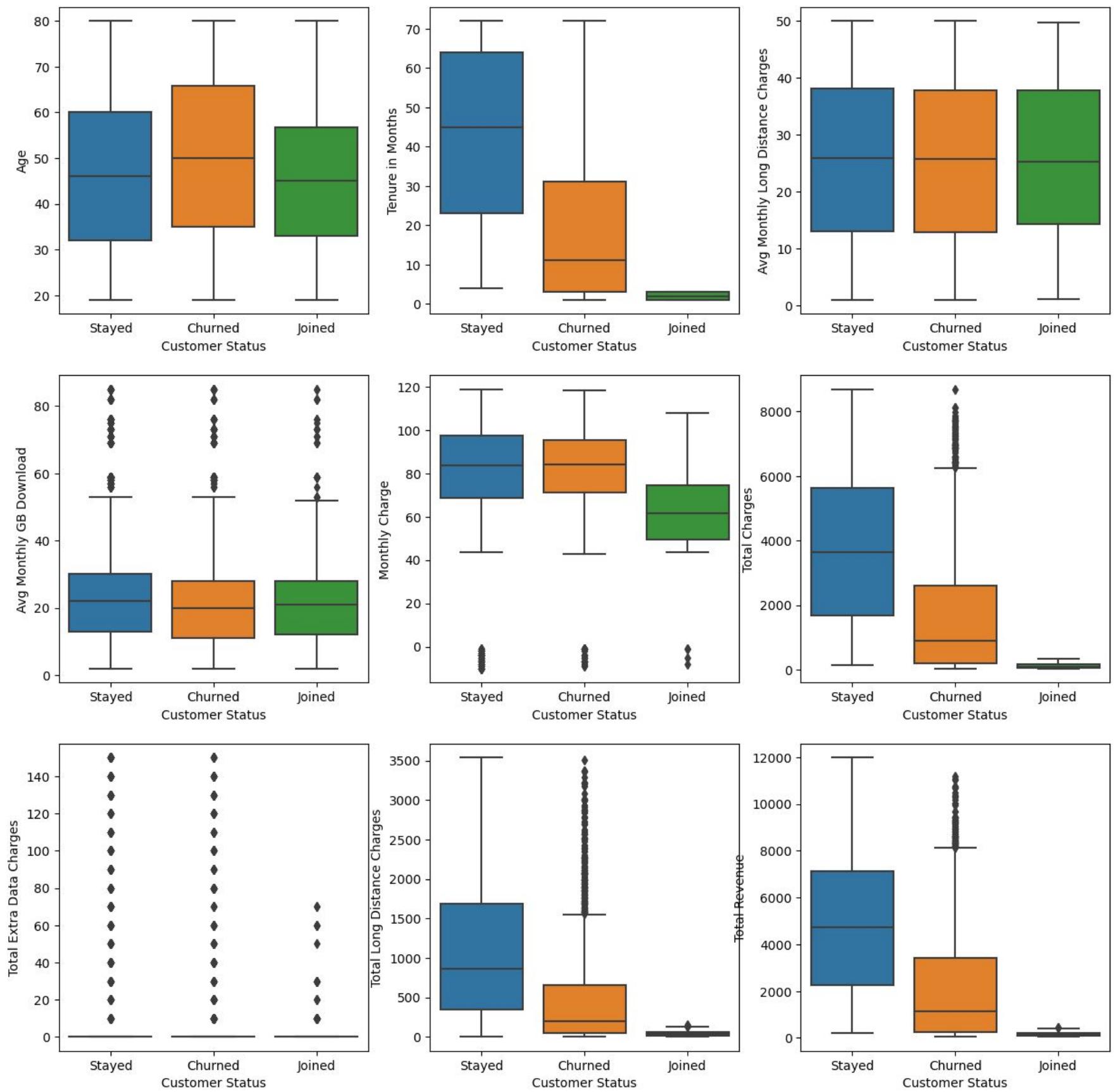


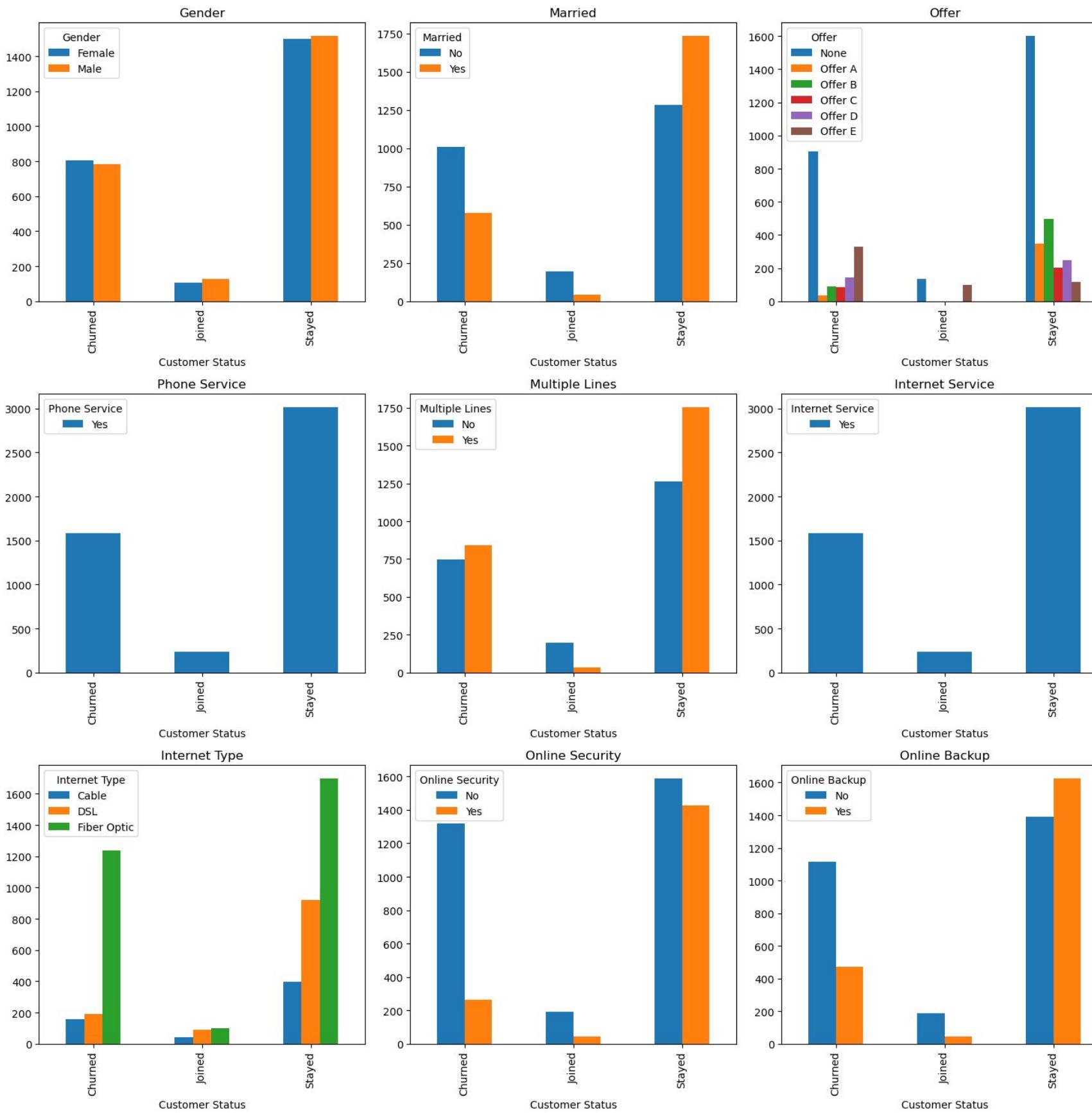
The customers with lower data usage, higher monthly charges & lower total charges have a higher churn rate.



The customers with lower extra data charges, lower long distance charges & lower revenue are more likely to churn.







BAR PLOT OF CUSOMER DISTRIBUTION ON DIFFERENT ATTRIBUTES



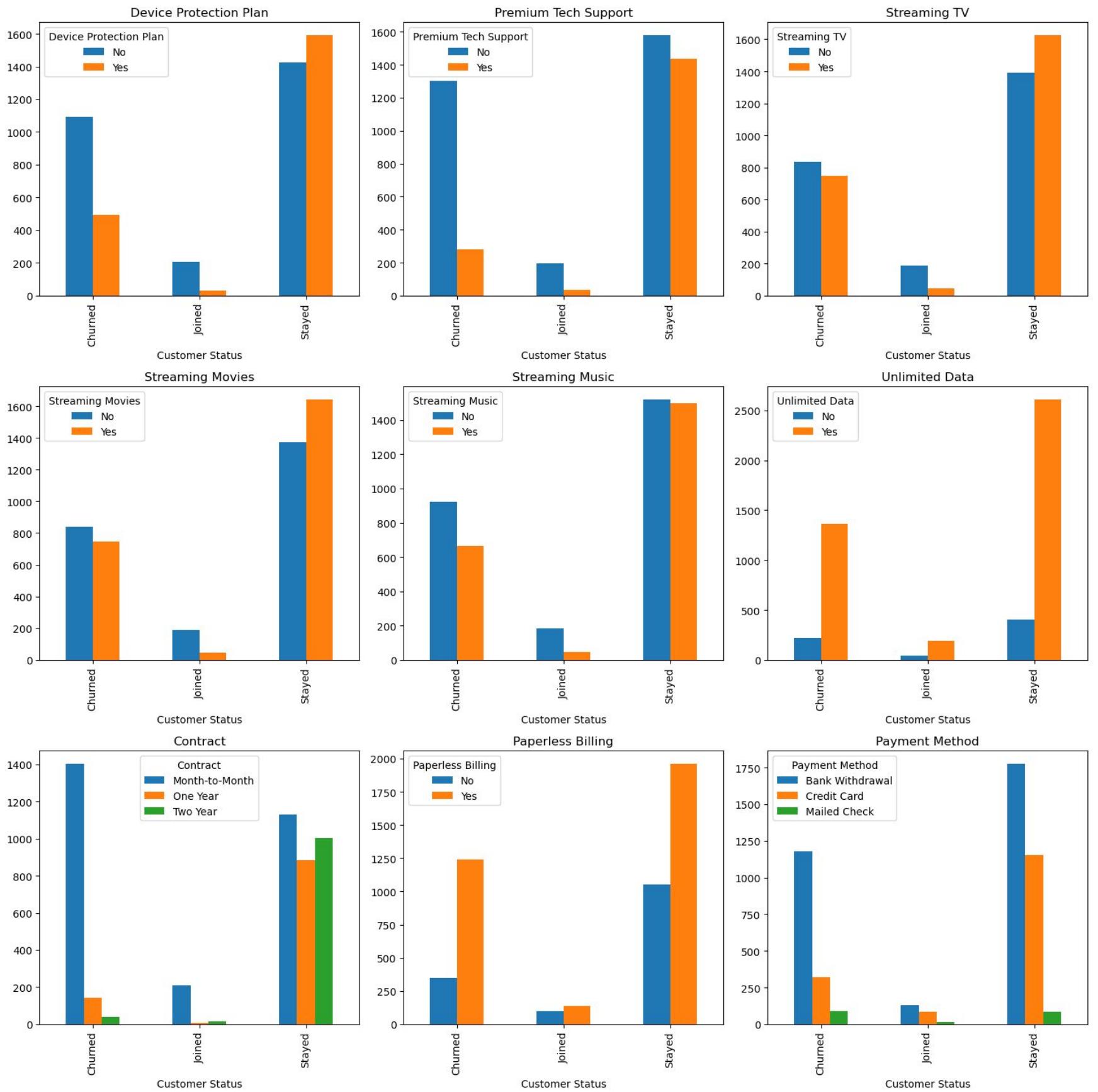
Gender distribution for all are almost same; married customers are more likely to stay; and certain offers might be more associated with churned or joined customers.

All types of customers have same phone service & same internet service; but a single line have a higher join rate.



Certain statuses are more likely to churn or join; but customers not getting online security & online backup has high churn rate.





BAR PLOT OF CUSOMER DISTRIBUTION ON DIFFERENT ATTRIBUTES



Customers without a device protection plan and without premium tech support are more likely to churn

Customers who have streaming TV, Movies, Music & unlimited data are more likely to stay.



Customers with a yearly contract have less churn rate; paperless billings have no effects on churn rate; and customers using credit card & mailed cheques are rare to churn.



FEATURE ENGINEERING

Key steps include feature encoding for categorical variables, label encoding to convert labels to numerical format, dummy variable introduction for binary representation, feature scaling for magnitude uniformity and train-test splitting for model evaluation on unseen data.

Feature Encoding



Gender values, Paperless Billing, Unlimited Data Usage, Movies- Music- TV Streaming, Premium Tech Support, Device Protection Plan, Online Backup & Security, Multiple Lines, Marital Status are converted into numerical representations.

Label Encoding



Customer statuses (Churned, Joined & Stayed) are encoded into numerical representations, (0,1,2) facilitating the incorporation of this information into machine learning models.

Dummy Introducing



Dummy variable is introduced for some columns like Payment Method, Contract, Internet Type, Offer & City

Feature Scaling



Dependents, age, referrals, tenure, monthly charges etc. columns are scaled by normalization

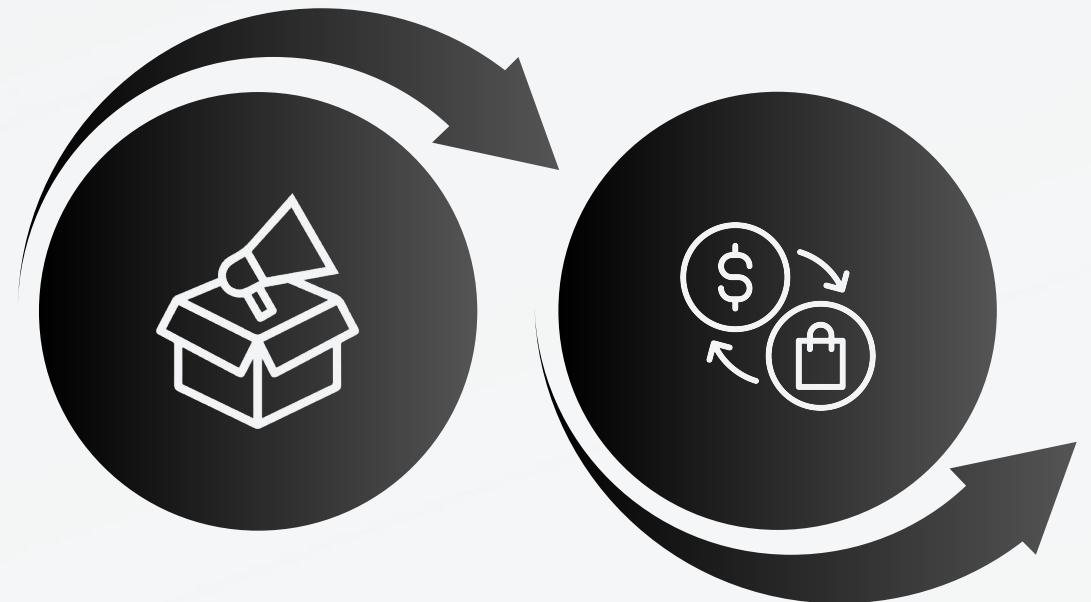
Data Splitting



80-20 Split for Train & Test

Naïve Bayes

The Naïve Bayes algorithm employs Bayes' theorem and assumes feature independence given the churn status. It calculates the probability of a customer churning based on observed features, considering the prior probability of churn and individual feature probabilities.

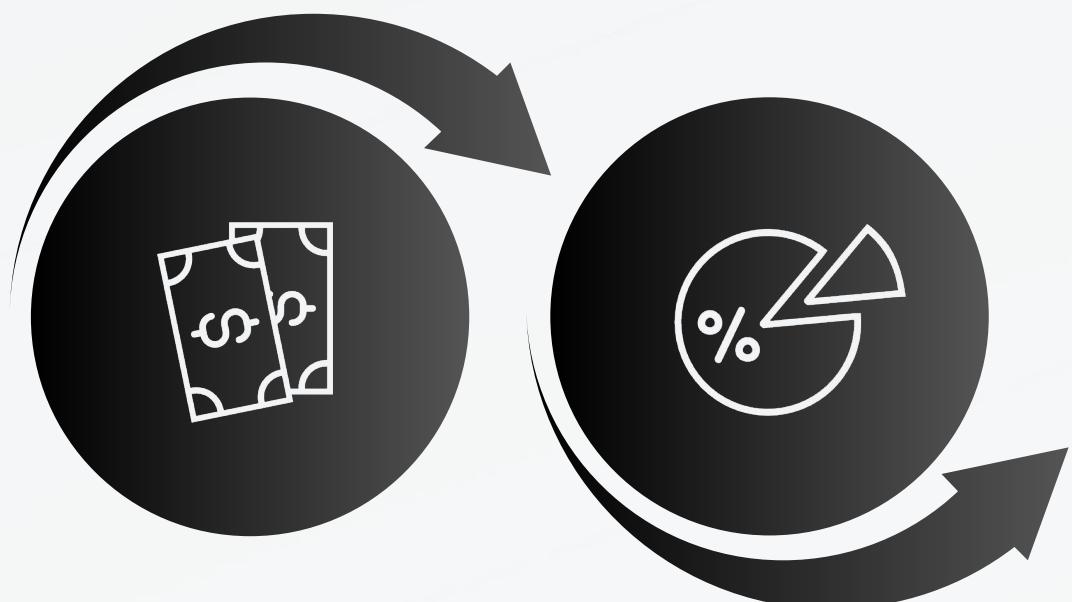


Logistic Regression

Logistic regression is employed by assembling a dataset with numerically encoded labels and utilizing the 'multinomial' option for multiclass classification. The interpretability of logistic regression allows for understanding the impact of individual features on the likelihood of churn, as indicated by the model's coefficients.

Random Forest

Random Forest for a three-class churn classification involves using an ensemble of decision trees to predict whether customers will Stay, Churn or Join. Features like usage patterns, complaints, and customer behavior are considered in the training process.



Decision Tree

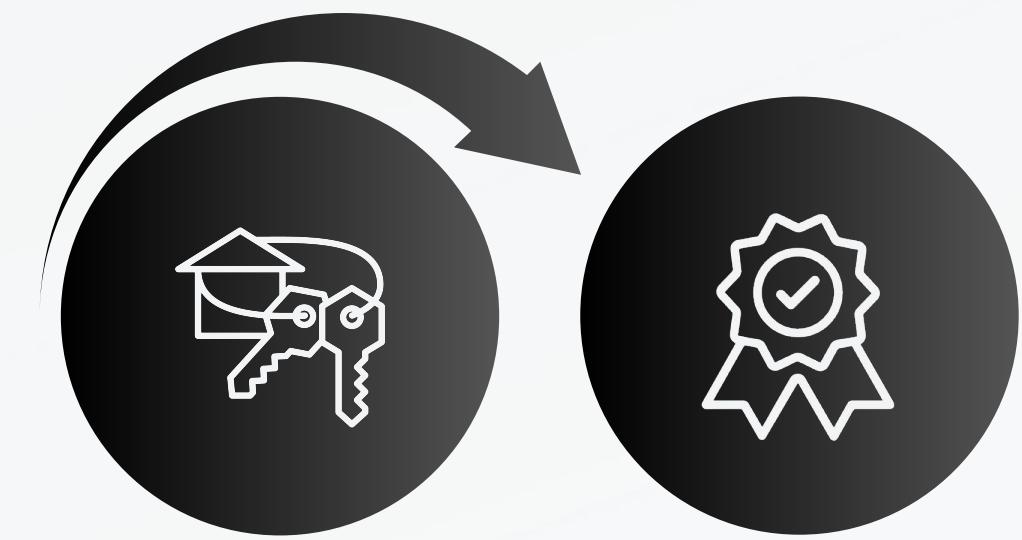
Decision Tree recursively splits the dataset based on features to create a tree-like structure. Each leaf node represents a class. Decision Tree, the algorithm recursively splits the dataset into Stayed, Churned or Joined based on features to create a tree-like structure. Each leaf node represents a class.

XGBoost

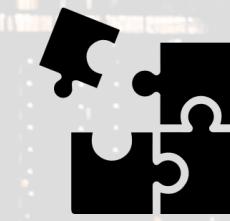
XGBoost is applied to a three-class churn classification by first preparing a dataset with relevant features and numerically encoding the class labels. The XGBoost algorithm, known for its ensemble of decision trees, is then employed to learn the patterns in the data.

SVM

SVM for three-class churn classification uses "one-vs-one" or "one-vs-the-rest," creating binary classifiers for class pairs or treating each class individually. Trained on labeled data with features reflecting customer behavior, SVM seeks decision boundaries maximizing class separation.

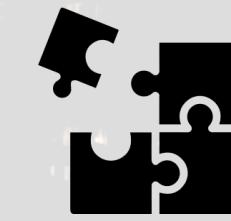


CONFUSION MATRIX



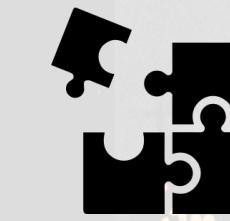
Logistic Reg

244	4	100
18	13	19
73	0	496



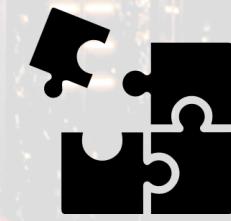
Naïve Bayes

169	92	87
18	15	17
359	64	146



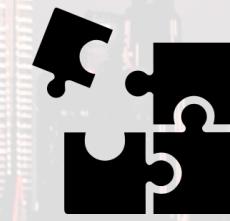
Decision Tree

214	22	112
30	20	0
71	0	498



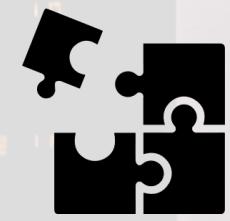
Random Forest

219	0	129
40	9	1
28	0	541



XGBoost

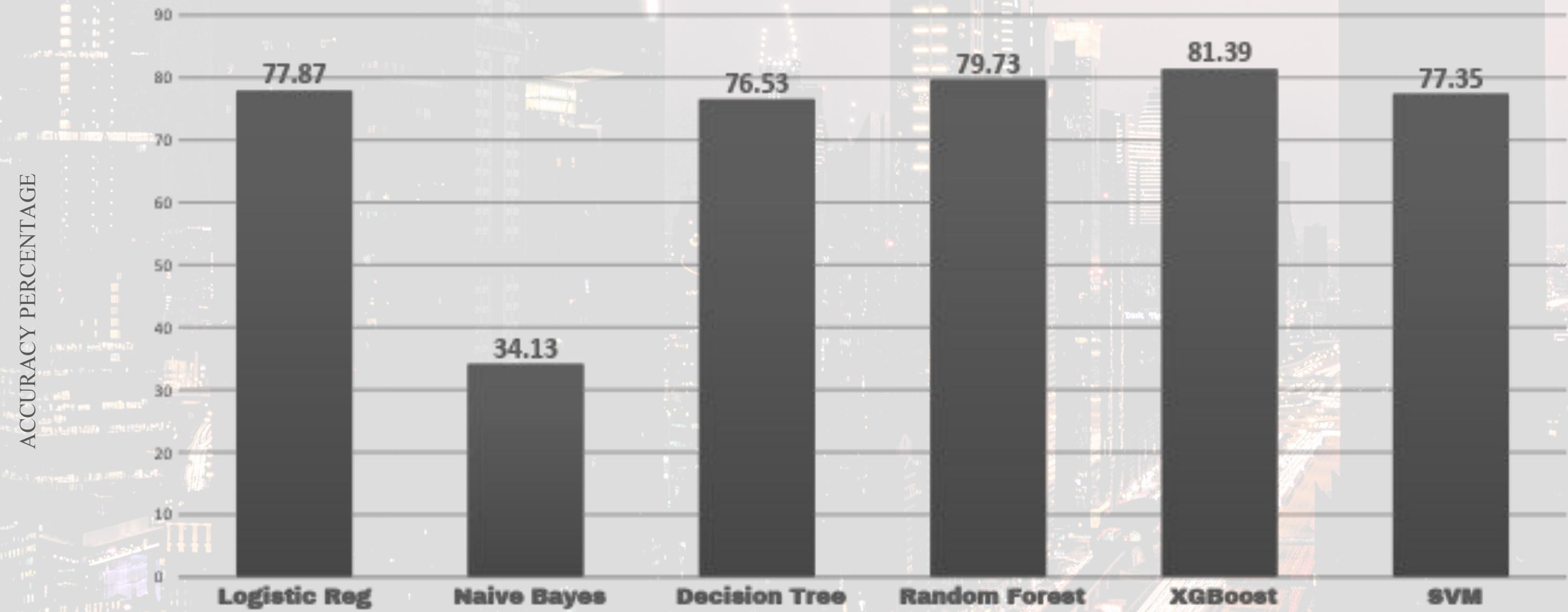
231	0	107
25	25	0
38	0	531



SVM

241	0	107
32	0	18
62	0	507

ACCURACY EVALUATION



CHALLENGES

Inaccuracies, missing values or outdated information can compromise the model's accuracy and reliability. Also, no real life data follows any particular distribution; so modelling on that dataset may violate model assumptions.



Selecting relevant features is crucial for accurate predictions. Understanding the business domain and identifying meaningful features can be highly challenging.



Striking the right balance to prevent overfitting or underfitting of models is challenging. Proper model tuning and regularization techniques are necessary to achieve optimal performance on unseen data.



Interpreting complex models for stakeholders can be challenging. In industries where decisions impact customer relationships, explaining predictions is crucial for gaining trust and acceptance.



FUTURE WORK

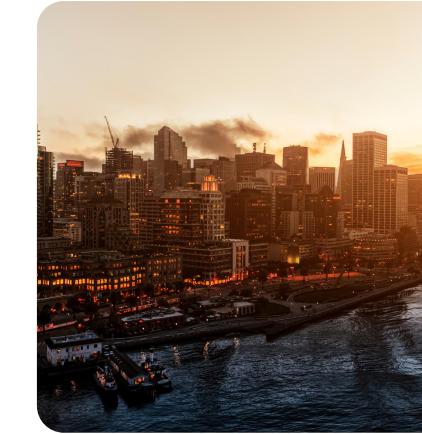
Future works involve refining feature engineering, optimizing model hyperparameters, exploring advanced techniques like ensemble methods & deep learning and fostering collaboration with business stakeholders for continuous model improvement aligned with strategic goals. Ethical considerations, bias assessment and user-friendly interfaces are also essential aspects to address in the evolving landscape of customer behavior prediction.

Handling Imbalanced Data



Implement techniques to address class imbalance in the dataset, especially if the occurrence of churn is significantly lower than retention.

Business Strategy Integration



Collaborate closely with business stakeholders to align the model's predictions with actionable strategies for customer retention and engagement.

Cost Benefit Analysis



Conduct a thorough cost-benefit analysis to understand the economic impact of implementing churn prevention strategies based on the model's predictions.

CONCLUSION

The work conclusively empowers businesses with actionable insights, derived from machine learning models, to strategically mitigate customer churn and foster sustained growth.

Actionable Insights



Business Impact



Empowering
Decision-Making





OUR TEAM



Soham Paul

2nd Year MS-QMS Student

Indian Statistical Institute, Hyderabad

+91 93303 87993

[soham-paul-96889019b](https://www.linkedin.com/in/soham-paul-96889019b)

Hyderabad, India



Anirban Dey

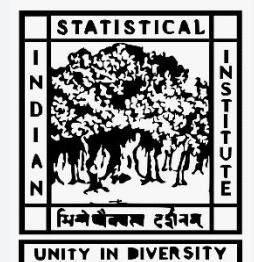
2nd Year MSc BDA Student

RKM Vivekananda Educational & Research Institute

+91 91635 68623

[anirban-dey-26b081260](https://www.linkedin.com/in/anirban-dey-26b081260)

Kolkata, India



ISI HYD RKMVERI

