TELECOM CHURN PREDICTION

(DOCUMENTATION)

BY FAIZAN QUAZI

Project overview:

Objective:

The central objective of this project is to develop a sophisticated predictive model tailored to the specific nuances of the telecom industry. This advanced model aims to accurately identify and forecast instances of customer churn – the phenomenon where customers discontinue using services or products. Recognizing the complexity of this challenge within the dynamic telecom sector, our project endeavors to leverage historical data and a diverse array of features to predict which customers are more likely to transition into a churn state.

Business Context:

In the pursuit of operational excellence and sustained business success, the strategic imperative to reduce customer churn aligns seamlessly with overarching business objectives. The significance of this alignment is articulated through several key dimensions:

1. Revenue Maximization:
The reduction of customer churn stands as a strategic pillar for revenue maximization. Retaining existing customers proves economically prudent, preserving established revenue streams and bolstering overall profitability.

2. Enhancement of Customer Lifetime Value (CLV):
Prioritizing the reduction of churn is intricately tied to the augmentation of Customer Lifetime Value. By fostering enduring customer relationships, businesses can elevate the overall value derived from each customer over their lifetime.

3. Market Competitiveness:
A low churn rate becomes synonymous with a positive brand image, fostering competitiveness in dynamic markets. Businesses that effectively mitigate churn are perceived as reliable and customer centric, gaining an advantageous position in the competitive landscape.

4. Operational Efficiency and CustomerCentric Strategies:
The quest to minimize churn necessitates a nuanced understanding of customer needs. Through the implementation of customer centric strategies, businesses not only reduce churn but also elevate service quality and operational efficiency.

5. Data Driven DecisionMaking:
The emphasis on reducing churn aligns seamlessly with the era of data driven decision making. Analyzing customer behavior provides invaluable insights, enabling strategic decisions, targeted marketing initiatives, and the development of personalized customer retention strategies.

6. Enhanced Customer Satisfaction:

Proactive churn reduction efforts directly contribute to heightened customer satisfaction. Satisfied customers are more likely to exhibit loyalty, advocate for the brand, and positively influence the business's reputation.

7. Sustainable Growth Foundation:
A reduced churn rate establishes a robust foundation for sustainable growth. Recognizing that customer acquisition costs exceed retention costs, this strategic focus contributes to the long term sustainability of the business.

8. Adaptability to Market Changes:
Businesses actively engaged in churn reduction demonstrate enhanced adaptability to market changes. The understanding of customer preferences and the timely addressing of concerns enable agile adjustments to products, services, and marketing strategies.

2. Data Sources:

Description:
The telecom churn dataset, a key component of our analytical endeavors, has been sourced from Kaggle. This dataset delves into customer dynamics within the telecom industry, focusing on parameters influencing churn behavior. Structured to encapsulate customer demographics, service tenure, and various telecom service subscriptions, it serves as a robust resource for in depth analysis.

Types of Data:
The dataset encompasses both categorical and numerical data types. Categorical variables include gender, partner status, dependency status, service contracts, and payment methods. Numerical variables, such as customer tenure, monthly charges, and total charges, provide quantitative insights into customer behavior.

Formats:
Presented in a tabular format, the dataset features rows representing individual customers and columns representing distinct attributes. This organized structure facilitates seamless analysis and modeling, ensuring clarity and coherence in data interpretation.

Acquisition Methods:

The dataset's origin lies in Kaggle, a widely recognized platform for data exploration. The selection of Kaggle ensures reliability and relevance, given its reputation for curated datasets spanning diverse domains.

Data Dictionary:

1. **customerID**: Unique identifier for individual customers.
2. **gender**: Customer's gender (Male/Female).
3. **SeniorCitizen**: Binary indicator of senior citizen status (1 for senior, 0 for nonsenior).
4. **Partner**: Presence of a partner (Yes/No).
5. **Dependents**: Presence of dependents (Yes/No).
6. **tenure**: Duration of customer relationship in months.
7. **PhoneService**: Availability of phone service (Yes/No).
8. **MultipleLines**: Presence of multiple phone lines (Yes/No/No phone service).
9. **InternetService**: Type of internet service subscribed (DSL/Fiber optic).

10 **OnlineSecurity**: Availability of online security service (Yes/No).

11. **OnlineBackup**: Availability of online backup service (Yes/No).
12. **DeviceProtection**: Availability of device protection service (Yes/No).
13. **TechSupport**: Availability of tech support service (Yes/No).
14. **StreamingTV**: Availability of streaming TV service (Yes/No).
15. **StreamingMovies**: Availability of streaming movies service (Yes/No).
16. **Contract**: Type of service contract (Monthtomonth/One year).
17. **PaperlessBilling**: Preference for paperless billing (Yes/No).
18. **PaymentMethod**: Customer's chosen payment method.
19. **MonthlyCharges**: Monthly charges incurred by the customer.
20. **TotalCharges**: Cumulative charges incurred by the customer.
21. **Churn**: Churn status indicating customer departure (Yes/No).

3. Data Preprocessing:

Handling Missing Values:

Visualized missing values as a matrix to identify patterns. No peculiar pattern observed, and no missing data identified.
Identified indirect missingness, particularly in the 'TotalCharges' column, which had 11 missing values.

Outlier Treatment:

Conducted a comprehensive analysis using box plots on all numeric columns to detect outliers. No outliers were found in the dataset, as depicted by the box plots.

Data Manipulation:

Removed the 'customerID' column from the dataset.

Reason: The 'customerID' was deemed nonsignificant for model building, as it serves as a unique identifier with no inherent predictive value.

Handling Indirect Missingness:
Discovered 11 missing values in the 'TotalCharges' column.

Encoding Categorical Variables:
Utilized LabelEncoder to convert all categorical features into numerical format.

Standardizing Numeric Attributes:
Applied StandardScaler to standardize numeric attributes.

Feature Engineering:
No new feature creation.

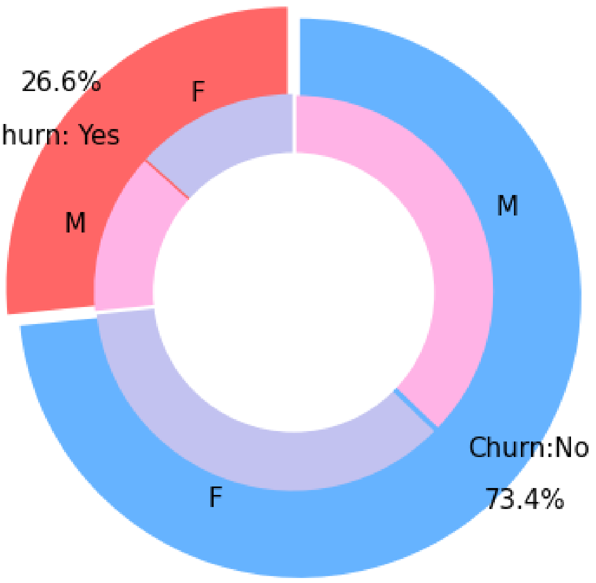
4.Exploratory Data Analysis (EDA):

Summary:

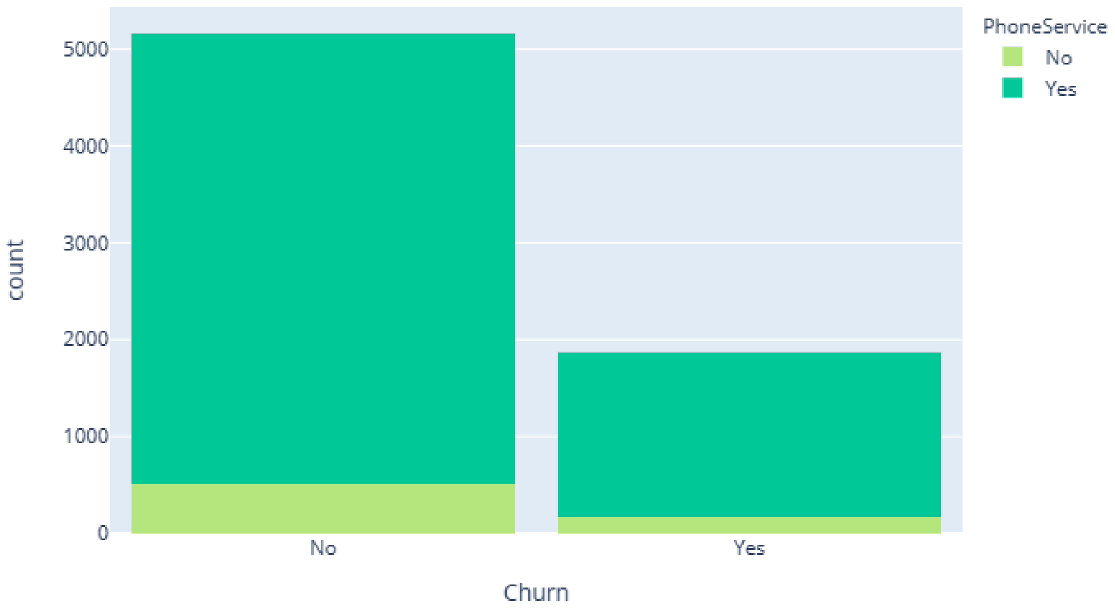
- Approximately 26.6% of customers opted to switch to another firm.
- Gender distribution among customers is relatively balanced, with 49.5% female and 50.5% male.
- Negligible difference observed in the churn rate between genders, indicating similar behavior.
- A significant proportion (75%) of customers with Month to Month contracts chose to switch providers, contrasting with 13% with One Year contracts and 3% with Two Year contracts.
- The majority of customers who moved out used Electronic Check as their payment method.
- Customers utilizing CreditCard automatic transfer, Bank Automatic Transfer, and Mailed Check as payment methods were less likely to churn.
- Fiber optic service users exhibited a high churn rate, suggesting potential dissatisfaction with this internet service.
- Customers with DSL service, while more numerous, displayed a lower churn rate compared to Fiber optic service users.
- Customers without dependents and partners were more likely to churn.
- A notable fraction of senior citizens churned.
- Churn was more prevalent among customers lacking online security, having Paperless Billing, and lacking TechSupport.
- Customers without phone service, especially new customers, had a higher likelihood of churning.
- Higher Monthly Charges correlated with an increased likelihood of churn, particularly among customers with charges between $60 and $120.

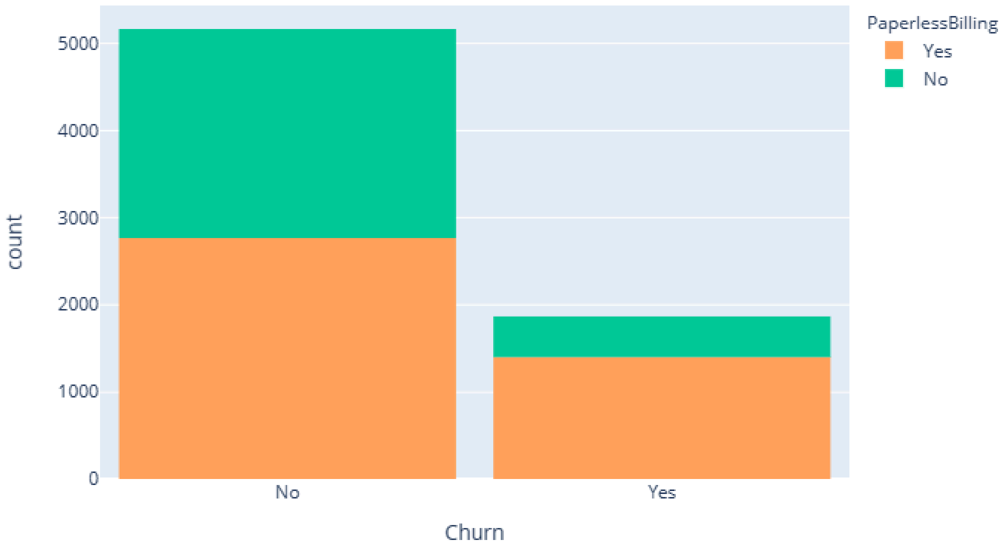
Visualizations:

Churn Distribution w.r.t Gender: Male(M), Female(F)

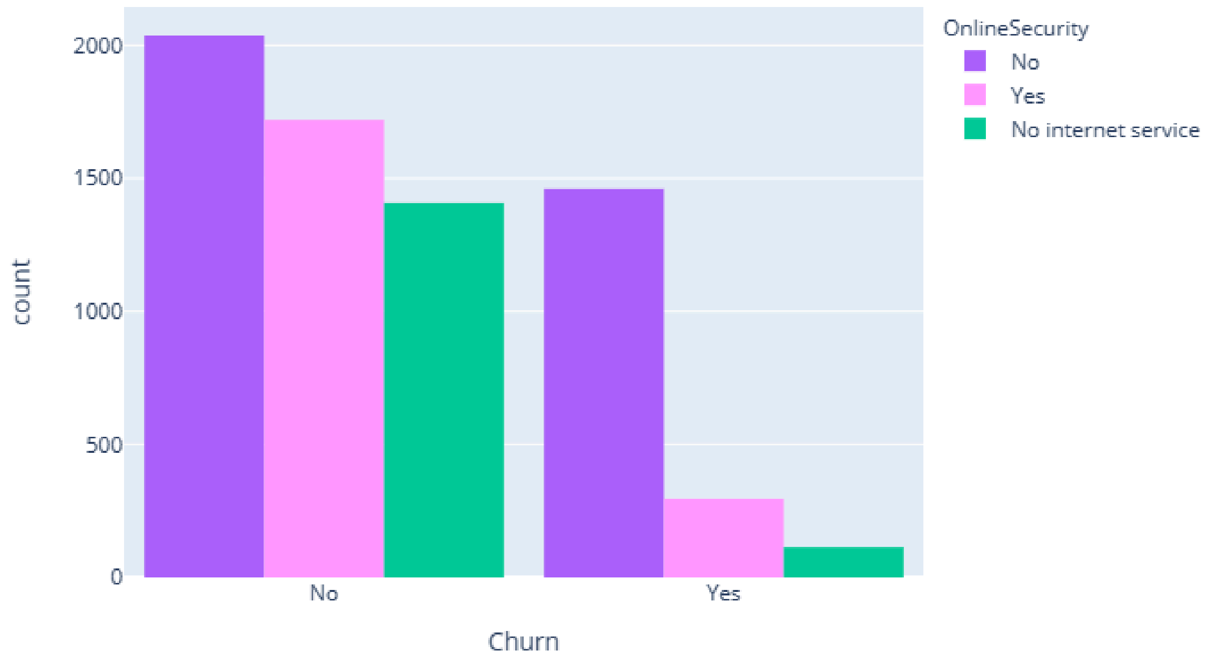


Chrun distribution w.r.t. Phone Service

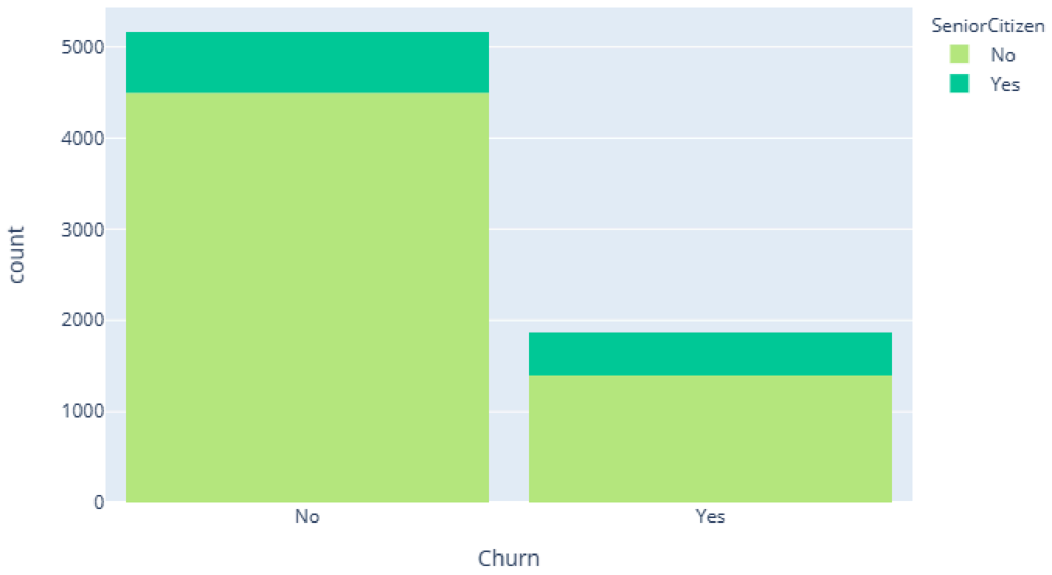




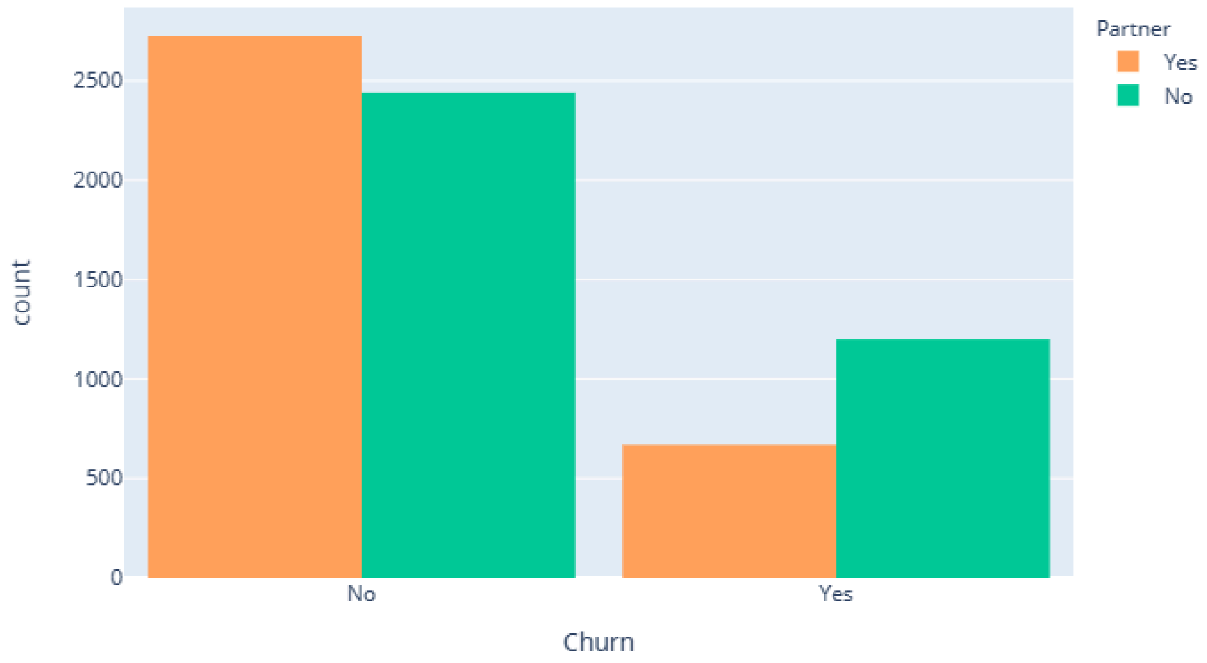
Chrun distribution w.r.t. Paperless Billing



Churn w.r.t Online Security

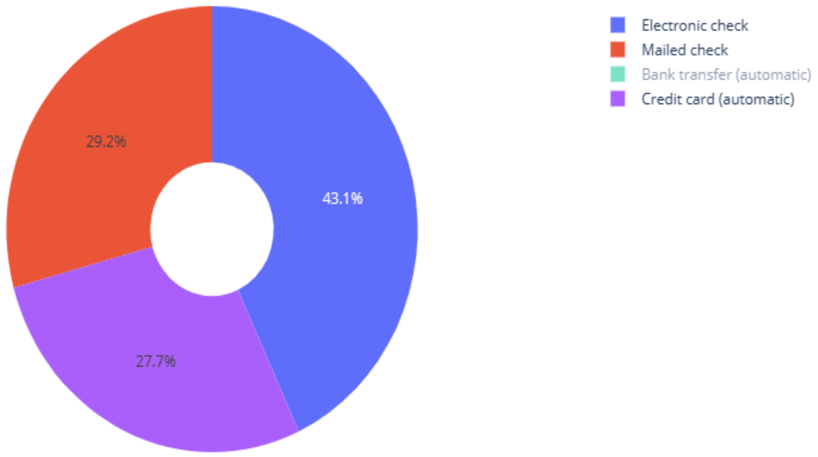


Chrun distribution w.r.t. Senior Citizen

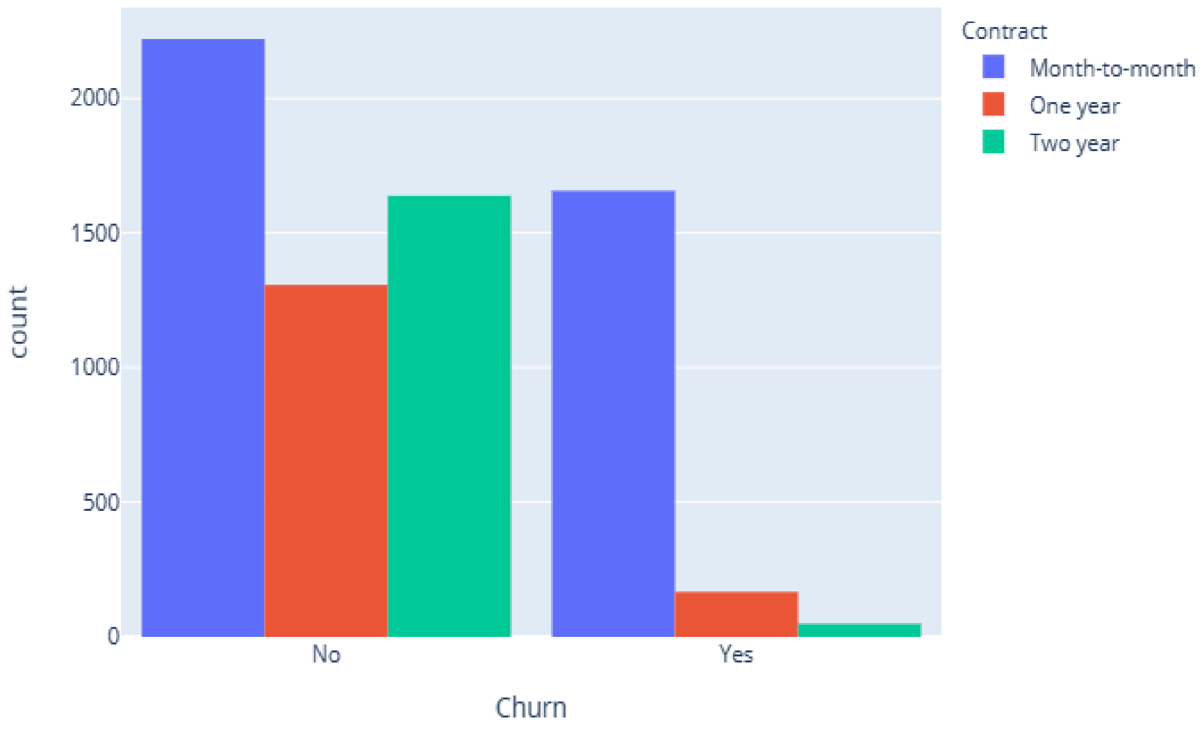


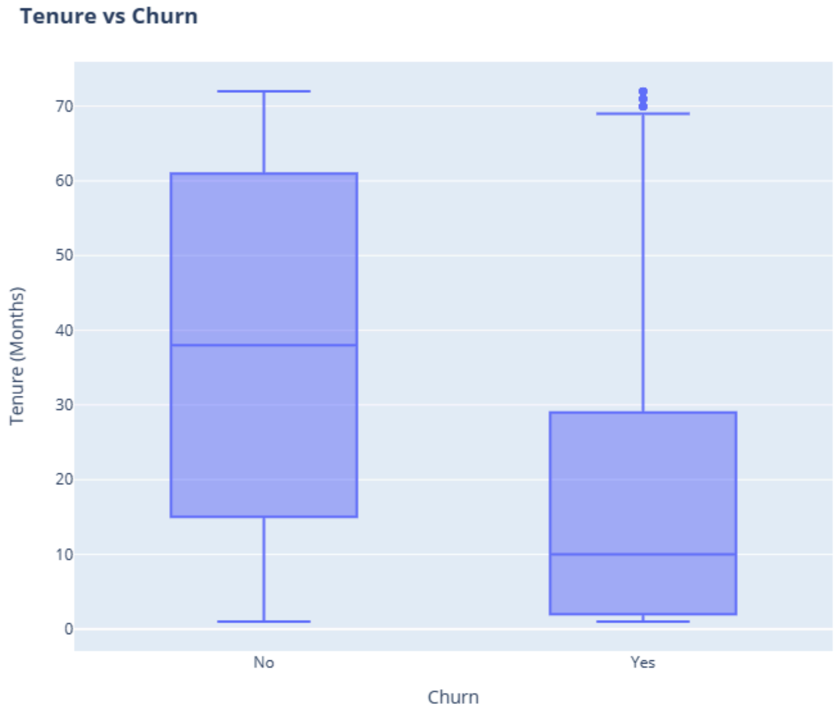
Chrun distribution w.r.t. Partners

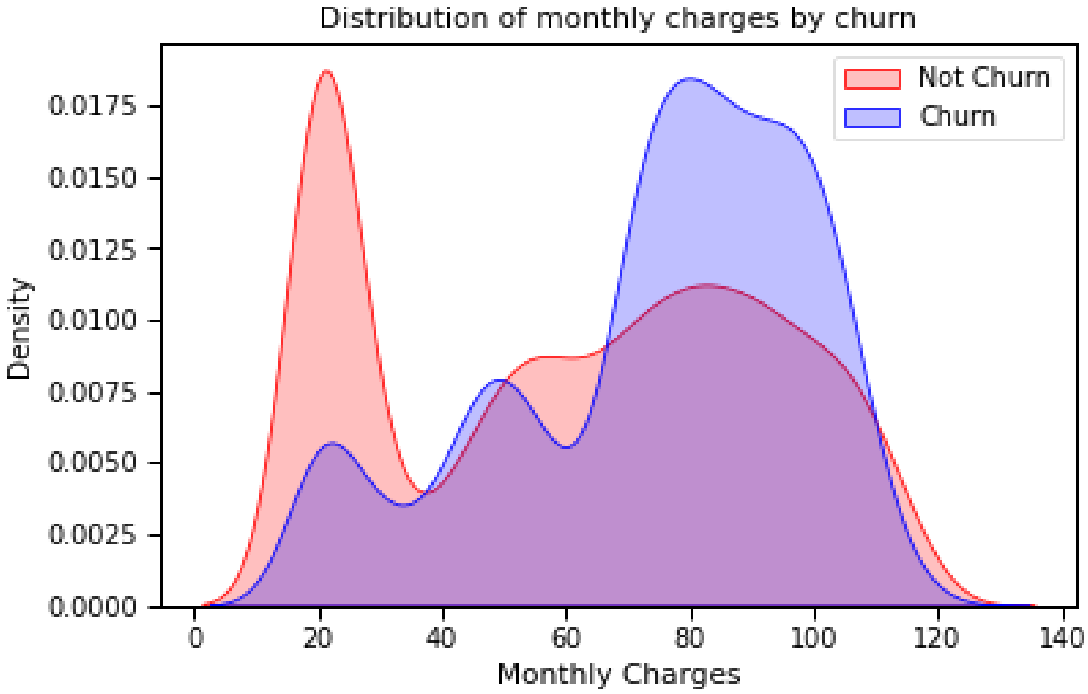
Payment Method Distribution

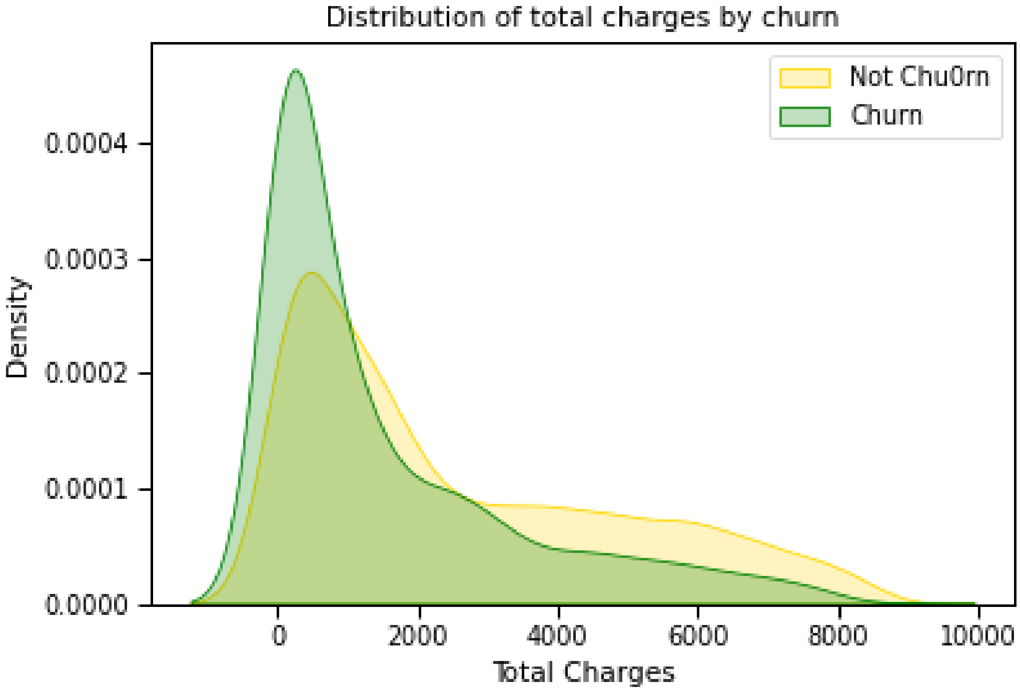


Customer contract distribution




Tenure vs Churn
Tenure (Months)
0
10
20
30
40
50
60
70
No
Yes
Churn


Distribution of monthly charges by churn
Not Churn
Churn
Density
0.0000
0.0025
0.0050
0.0075
0.0100
0.0125
0.0150
0.0175
0
20
40
60
80
100
120
140
Monthly Charges



5. Data Splitting:

TrainTest Split:

The dataset was divided into training and testing sets using the `train_test_split` function from the scikit learn library.
A test size of 20% was chosen to allocate a representative portion for model evaluation.
A fixed random state of 0 was set to ensure consistent splits for reproducibility.

TrainValidation Split:

Within the training set, an additional split was performed to create a training subset (80%) and a validation subset (20%).
This enabled the evaluation of models on a separate validation set during training to finetune hyperparameters and assess generalization performance.

6. Model Selection:

Algorithm Choice:

The choice of algorithms was guided by the need for a diverse set of models suitable for the customer churn prediction task.

The selected algorithms, including Logistic Regression, KNearest Neighbors, Support Vector Classifier, Bernoulli Naive Bayes, Decision Tree Classifier, Random Forest Classifier, AdaBoost Classifier, and LightGBM Classifier, were chosen for their distinct characteristics in handling classification problems.

Hyperparameter Tuning:

Hyperparameters were meticulously tuned to optimize the performance of each algorithm.

Specific hyperparameters, such as regularization strength, tree depth, and learning rate, were adjusted for each model to achieve an optimal balance between bias and variance.

Tuning was conducted through a systematic grid search or random search approach, depending on the specific requirements of each algorithm.

8. Model Evaluation:

Metrics Used:

Accuracy: Represents the overall correctness of the models' predictions, showcasing the percentage of accurately classified instances.

F1 Score: Balances precision and recall, yielding a comprehensive score considering both false positives and false negatives.

Validation Score: Signifying the mean accuracy derived from kfold crossvalidation, ensuring robustness in model assessment.

Confusion Matrix: Delivers a detailed breakdown of true positives, true negatives, false positives, and false negatives, offering insights into model performance.

Performance Summary:

The following table encapsulates the evaluation metrics for each individual model, presenting their respective strengths and areas for enhancement:

SNO.	model_name	accuracy	f1_scores	val_score	bias	variance	confusion_matrix
0	LogisticRegression	0.810628	0.81526861	0.8007926	0.805025	0.810628	[[813 220] [172 865]]
1	KNeighborsClassifier	0.77343	0.78959175	0.7509576	0.841749	0.77343	[[721 312] [157 880]]
2	SVC	0.5975845	0.61381548	0.7346317	0.592534	0.5975845	[[575 458] [375 662]]
3	BernoulliNB	0.763285	0.77502296	0.7489711	0.776637	0.763285	[[736 297] [193 844]]
4	DecisionTreeClassifier	0.7956522	0.7986673	0.7266811	0.998067	0.7956522	[[808 225] [198 839]]
5	RandomForestClassifier	0.8565217	0.85890736	0.7927013	0.998067	0.8565217	[[869 164] [133 904]]
6	AdaBoostClassifier	0.8193237	0.82539683	0.8030652	0.826528	0.8193237	[[812 221] [153 884]]
7	LGBMClassifier	0.8608696	0.86311787	0.7948312	0.90662	0.8608696	[[874 159] [129 908]]

Voting Accuracy:

The voting classifier, an ensemble of all models, achieved an accuracy of approximately **84.15%**, underscoring the efficacy of amalgamating diverse models for predictive purposes.

.

Recommendations for Future Work and Improvement:

1. Continuous Model Monitoring:
Regularly monitor and update the predictive model with new data to ensure its continued relevance and effectiveness in adapting to evolving customer behaviors.

2. Feature Exploration:
Explore additional features or external factors that may influence churn. Consider incorporating dynamic factors that reflect changes in the telecom industry and customer preferences.

3. Dynamic Industry Adaptation:
Acknowledge the dynamic nature of the telecom industry and be prepared to adapt the model to changing market conditions and technological advancements.

4. Customer Feedback Integration:
Integrate customer feedback into the model training process to capture subjective insights and enhance the model's understanding of customer sentiments.

5. Collaboration Across Departments:
Foster collaboration between data science teams and other departments such as marketing and customer service to ensure a holistic approach in addressing churn and implementing effective retention strategies.