

Customer Churn Prediction

Leveraging data analytics and machine learning to address customer attrition

Project **Team 14**

Akshada Girish Malpure {agmalpur}

Rishi Dange {radange}

Rucha Mahesh Kulkarni {rkulkar5}



What is Churn Rate?

Customer churn, also known as customer attrition, is when customers stop doing business with a company over a certain period.

$$\text{Churn Rate} = \frac{\text{Customers lost in a period}}{\text{Total customers at the beginning of the period}} * 100\%$$

- Helps Enable targeted retention strategies.
- Improves forecasting accuracy and strategic planning.
- Enhances customer engagement by personalizing retention efforts.



Why is it relevant?

A high churn rate can indicate that customers are dissatisfied with the products or services, or that the business isn't marketing to the right audience.

- **Optimize Revenue Impact:** High churn rates can reduce recurring revenue and growth.
- **Max Cost Efficiency:** Retaining customers is generally more cost-effective than acquiring new ones.
- **Increase Customer Lifetime Value:** Lower churn increases customer loyalty and the lifetime value of each customer.
- **Business Insights based Tweaking:** Understanding churn helps identify service gaps, improving customer experience and satisfaction.

Our Agenda

The goal of this project is to accurately predict customer churn using machine learning and derive actionable insights to strengthen customer retention strategies.

By identifying high-risk customers, we aim to support proactive, cost-effective retention efforts that enhance customer satisfaction, loyalty and the overall service experience.

Ultimately, we plan to address the following questions:

- Which customers are at the highest risk of churning?
 - What are the primary features affecting the churn rate?
 - How can churn prediction insights support customer loyalty initiatives?
-



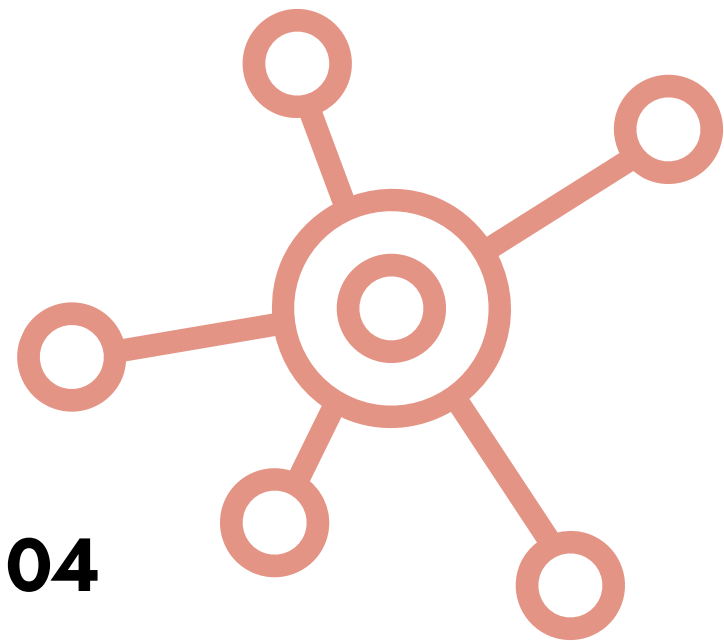
The Data

Description of the Data we worked upon.

Dataset source: Kaggle, Telco Customer Churn

Key features:

- **Personal information** including details like customer gender, age, and dependents.
- **Services subscribed** such as phone lines, internet, and security features.
- **Billing information** which includes monthly charges, payment methods, tenure, and the overall cost of services.



The primary target feature in this dataset is **Churn**, a binary variable indicating whether or not a customer has discontinued their services.

Telco Customer
Churn Data
[\[LINK\]](#)

7043
Customer Records

21
Features of Classification

70- 30
Test Train Split

Data Preprocessing

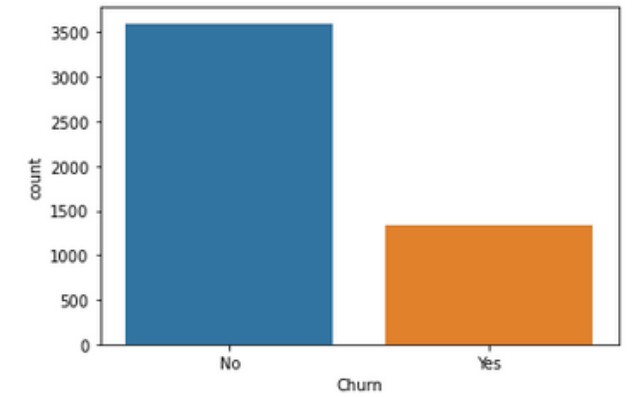
Data Cleaning:

- Handling Missing Values: Used mean/mode imputation for missing numerical and categorical values.
- Removing Irrelevant Features: Dropped the non-informative nominal columns like customerID.

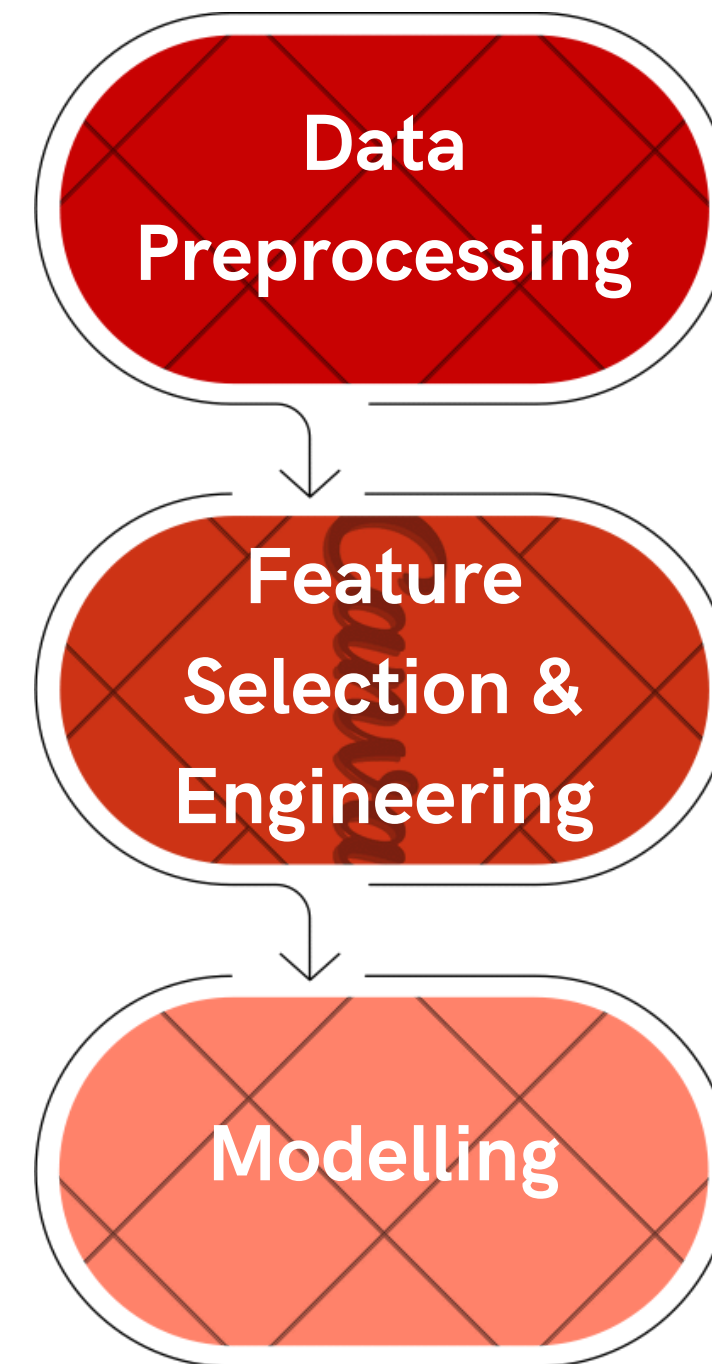
Encoding and Scaling:

- Label Encoding: Converted Yes/No labels to 1/0 for model compatibility.
- Standard Scaling: Scaled numeric columns like tenure and totalCharges.

Class Imbalance ? Solution: Applied SMOTE to balance churn vs. non-churn cases.



Plan of Action



Identifying and processing the features that impact churn prediction

Test various machine learning models for optimal performance

Data Augmentation

Feature Importance Assessment: Used Heatmap & domain knowledge to identify key features.

Selected Features: Tenure, Total Charges, Payment Method, and Contract Type.

Feature Engineering: Created tenure groups and churn likelihood per tenure group

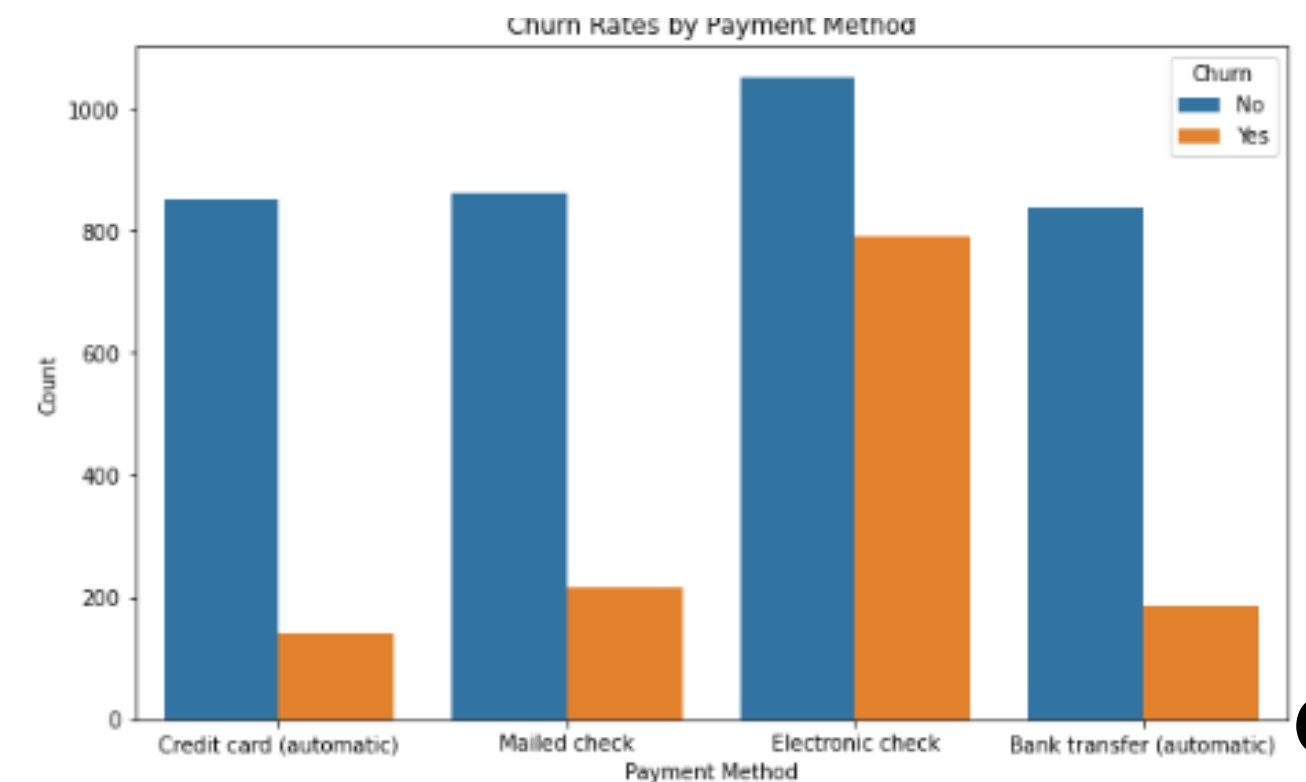
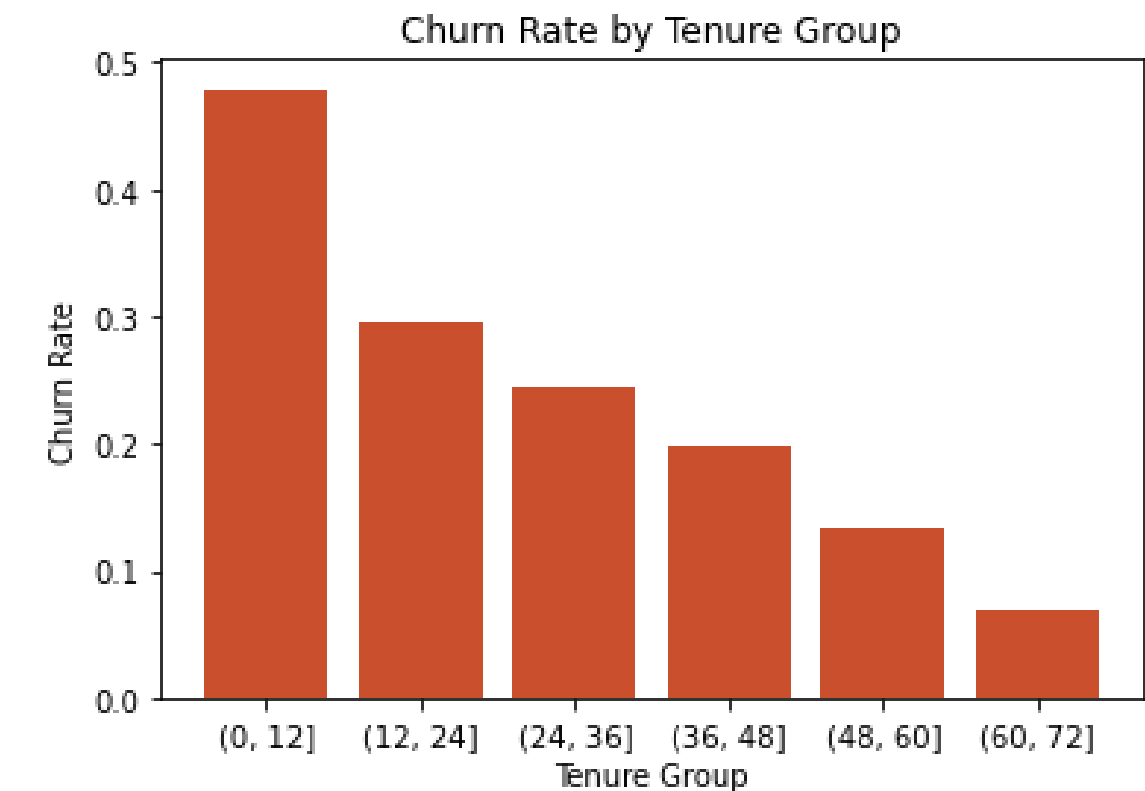
Hyperparameter Tuning:

Goal: To improve model performance and prevent overfitting.

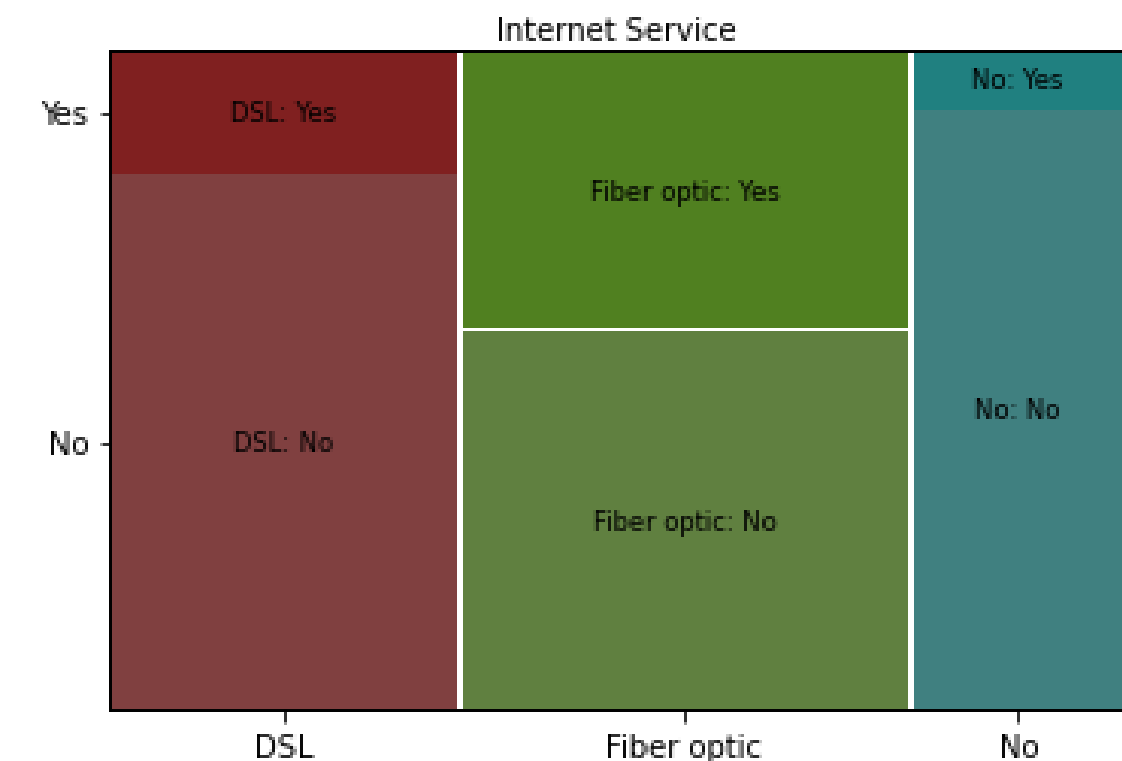
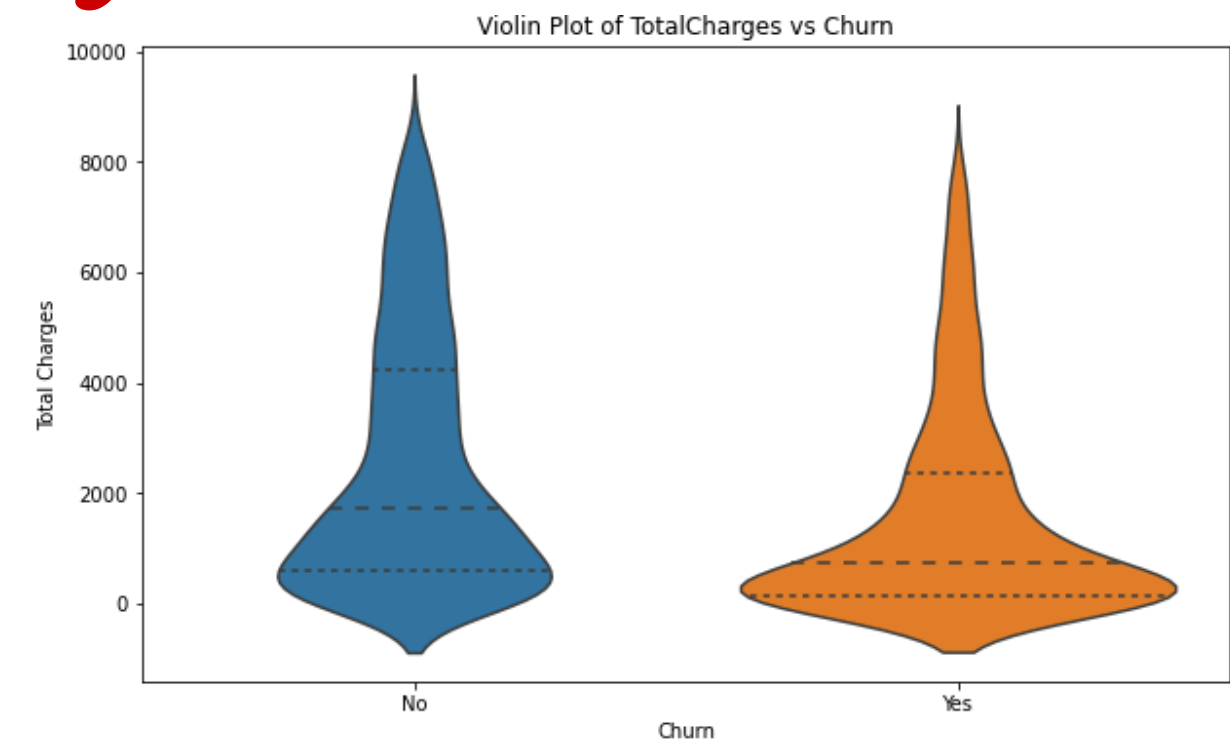
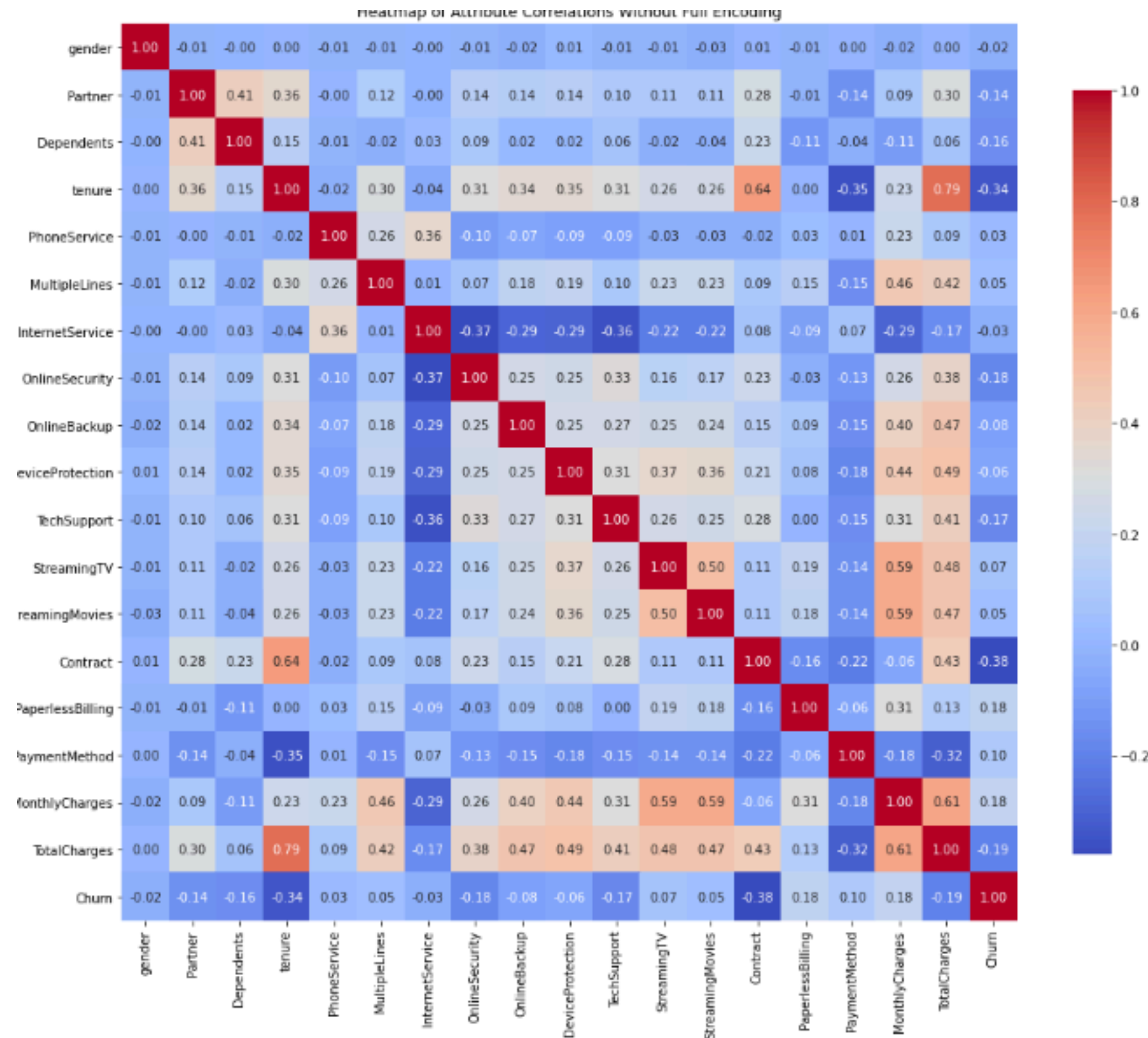
Techniques Used: Used grid search and randomized search for optimal parameters in:

Random Forest, LightGBM, Gradient Boosting

Outcome: Enhanced model accuracy and robustness across the selected models.



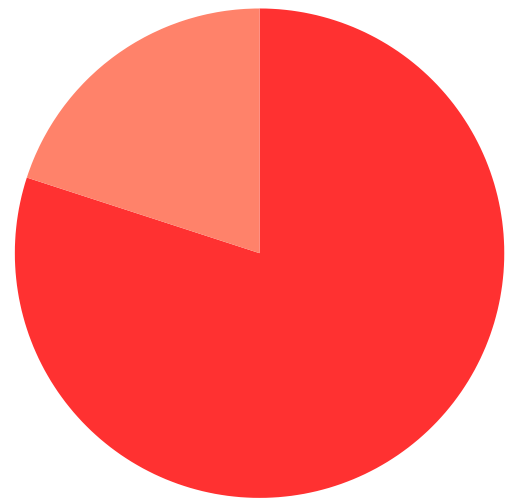
Visual Analysis



Model Training

Data Split: Training and evaluation sets created using stratified sampling to preserve class balance.

Evaluation Data
20%



Training Data
80%

Accuracy, precision, recall,
F1 score, and AUC-ROC

Evaluation Metrics

- Naive Bayes
- KNN
- Decision Tree
- XGBoost
- LightGBM
- Logistic Regression
- Random Forest
- Gradient Boosting
- SVM

MODELS TRAINED

Each model captures a different aspect of churn behavior, improving prediction accuracy, like non-linear patterns, independent features, high dimensional data, scalability, robustness, and complex feature interactions.

Tweaking Performance

- From our initial results, Random Forest, LightGBM, and Gradient Boosting models showed promising performance, achieving accuracies of around 84%.
- To further enhance the predictive power, we implemented hyperparameter tuning by adjusting the following:
 - Learning rate
 - Number of estimators
 - Max depth and min samples split
 - Number of leaves.
- This fine-tuning allowed for accuracy to improve to approximately **85%** and contributed to **model robustness**, enabling better identification of customers likely to churn.

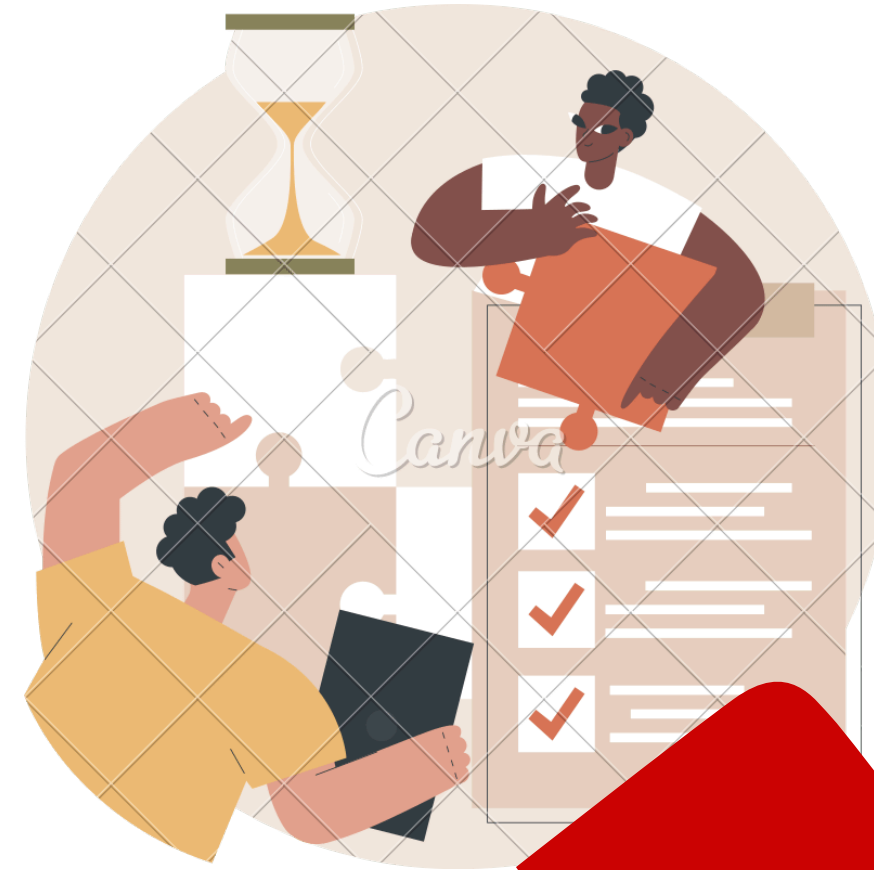
	Model	Accuracy	Precision	Recall	F1 Score	AUC-ROC
0	Logistic Regression	0.760417	0.755102	0.770833	0.762887	0.760417
1	Random Forest	0.853472	0.873715	0.826389	0.849393	0.853472
2	Gradient Boosting	0.845833	0.856734	0.830556	0.843441	0.845833
3	Support Vector Machine	0.820833	0.827195	0.811111	0.819074	0.820833
4	Gaussian Naive Bayes	0.756944	0.750678	0.769444	0.759945	0.756944
5	K-Nearest Neighbors	0.772917	0.720045	0.893056	0.797272	0.772917
6	Decision Tree	0.798611	0.807143	0.784722	0.795775	0.798611
7	XGBoost	0.840278	0.856105	0.818056	0.836648	0.840278
8	LightGBM	0.850694	0.869693	0.825000	0.846757	0.850694

	Model	Accuracy	Precision	Recall	F1 Score	AUC-ROC
0	Random Forest (Tuned)	0.853472	0.858956	0.845833	0.852344	0.853472
1	LightGBM (Tuned)	0.848611	0.865889	0.825000	0.844950	0.848611
2	Gradient Boosting (Tuned)	0.845833	0.856734	0.830556	0.843441	0.845833

But Why these Models?		
Model	Merits	Demerits
Logistic Regression	Simple and interpretable for linear relationships.	Assumes linear relationships, struggles with complex patterns.
Random Forest	Robust with complex feature interactions.	Can be slow with large data; less interpretable due to ensemble nature.
Gradient Boosting	High accuracy for non-linear patterns.	Computationally intensive and prone to overfitting without tuning.
LightGBM	Fast, scalable, and handles categorical data well.	Prone to overfitting with small data; requires careful parameter tuning.
XGBoost	High accuracy and efficient on large data.	Computationally expensive, complex to tune for optimal results.
Decision Tree	Interpretable with key feature insights.	Easily overfits, especially with complex churn data.
KNN	Effective with similarity-based predictions.	Inefficient with large data, sensitive to irrelevant features.
Naive Bayes	Fast and good for independent features.	Assumes feature independence, which may not hold in churn data.
SVM	Finds optimal boundaries in high-dimensional data.	Slow with large datasets, sensitive to feature scaling, less interpretable.

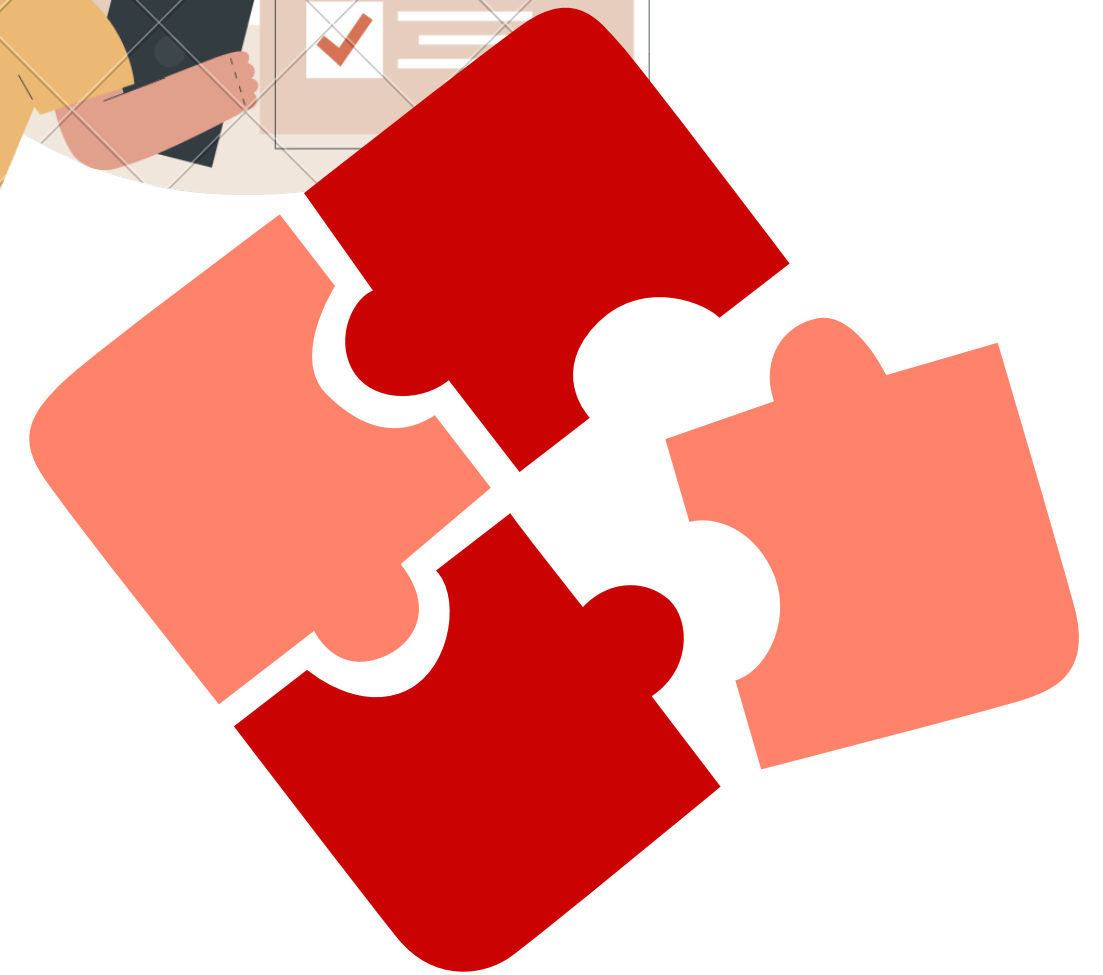
Future Scope

1. Personalized Retention Strategies
2. Integration with Behavioral Data
3. Automated Machine Learning
4. Real-Time Prediction



Challenges

- a. Quality of data available
- b. Changing needs and dynamic behavior
- c. Scalability with increasing customer corpus
- d. Inconsistencies and uneven trends in the behavioral pattern



Conclusion

12

1

Summary of Project: Our study demonstrates the power of machine learning in predicting customer churn with actionable insights to reduce attrition.

2

Takeaway Message: Predictive analytics allows businesses to proactively retain customers through targeted interventions.

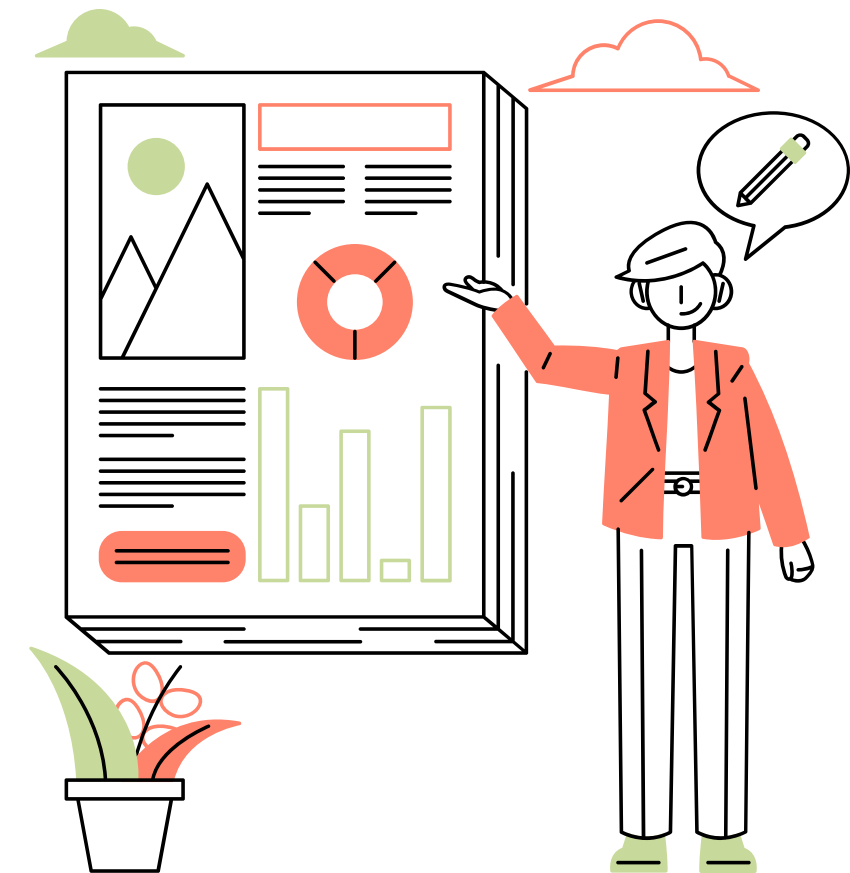
3

Final Thought: With continuous monitoring and model improvements, the company can maintain an effective churn prediction system and drive long-term customer loyalty.


4

Customer Retention Strategies:

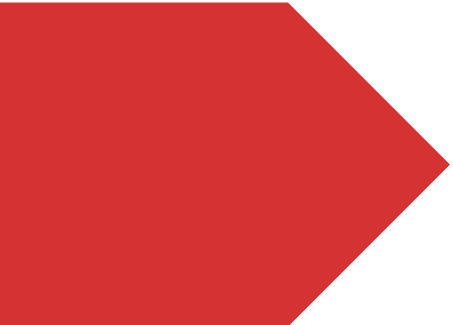
- Early engagement for customers with short tenure.
- Incentives for customers who use automated payment methods.
- Enhanced paperless billing experience.
- Targeted marketing and educational campaigns to build loyalty.



References

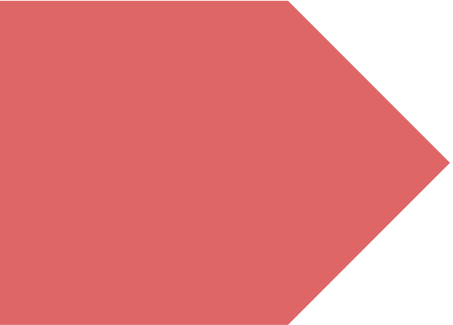


Ibrahim AlShourbaji et al. An efficient churn prediction model using gradient boosting machine and metaheuristic optimization. 2023. DOI: 10.1038/s41598-023-41093-6.



Hemlata Jain, Ajay Khunteta, and Sumit Srivastava. Telecom churn prediction and used techniques, datasets and performance measures: a review. 2021. DOI: 10.1007/s11235-020-00727-0.

Jain et al. (2021) emphasized the need for balancing imbalanced datasets and extracting domain-specific features to improve results.



Sharmila K. Wagh et al. Customer churn prediction in telecom sector using machine learning techniques. 2024. DOI: 10.1016/j.rico.2023.100342.

Wagh et al. (2024) highlighted the importance of demographic and usage features in their analysis using methods like logistic regression and random forests by emphasizing how feature engineering enhances model performance.

“Without data, you’re just another person with an opinion.”

-W Edwards Deming

THANK YOU