Churn Prediction for Subscription-Based Services using Business Analytics techniques

BUSINESS ANALYTICS
WITH R

Debasmita Ray

01 BACKGROUND

- Objective
- Financial Impact
- Methodology
- Analysis Overview

02 ACTION PLAN

- Flowchart
- Data Collection
- Data Preprocessing
- Exploratory Data Analysis
- Feature Selection
- Model Building
- Model Training and Testing
- Model Evaluation
- Model Interpretation

03 WORKING

- Decision Tree
- Random Forest
- Logistic Regression
- Evaluating Models

04 CONCLUSION

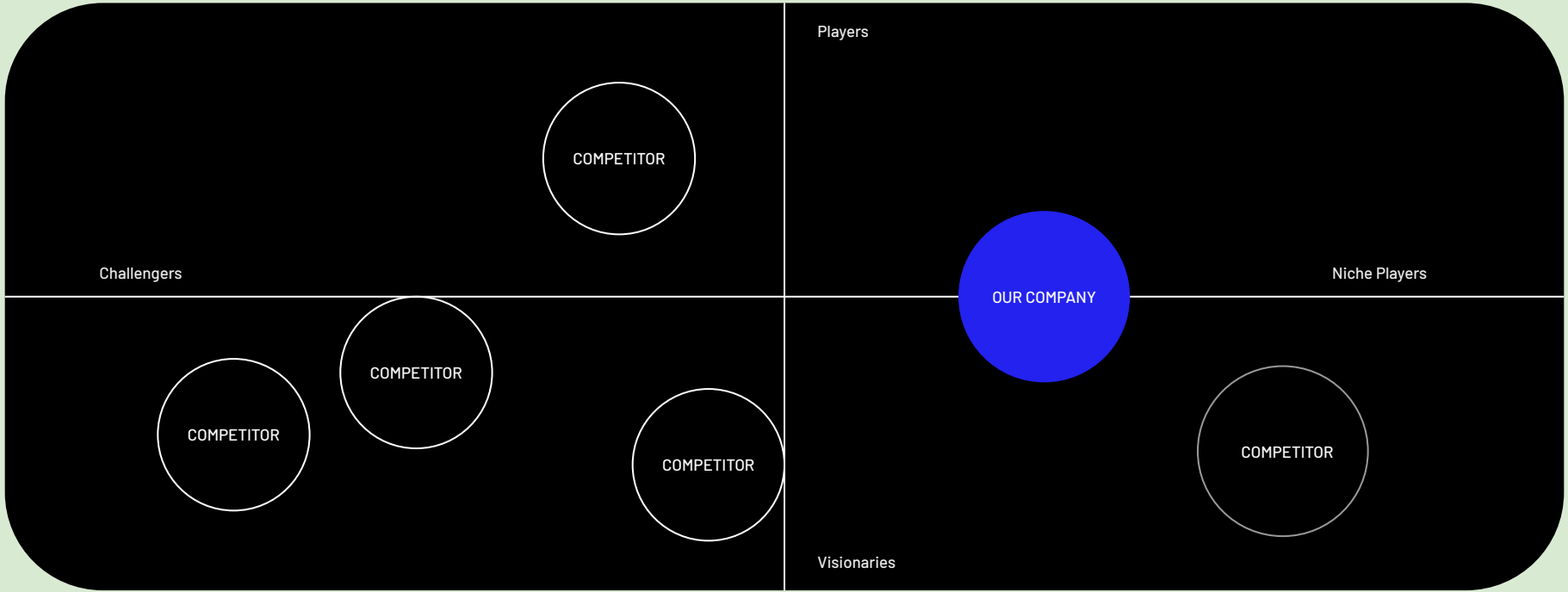
- Deployment and Report
- Key Drivers
- Actionable Insights
- Findings

01

Background

OBJECTIVE

Our primary goal is to enhance customer retention for a subscription-based company by developing an effective churn prediction model. Our primary objectives are twofold: to accurately identify customers at risk of churning and to uncover the key factors driving this churn. The end goal is not just to reduce churn rates, but to significantly improve customer loyalty and lifetime value. In the competitive landscape of subscription services, this project represents a critical step towards sustainable growth and increased market share.



FINANCIAL IMPACT

In today's competitive landscape, subscription-based businesses face a serious challenge: rising customer churn. With some companies experiencing monthly churn rates above 5%, this translates to a potential loss of over half their customer base annually. Studies show that reducing churn by just 5% can increase profits by up to 95%, making effective churn prediction essential. This project aims to develop a predictive model in R, analyzing historical customer data to identify those at risk of leaving. The resulting insights will support targeted retention strategies, helping businesses reduce churn and improve profitability.

METHODOLOGY

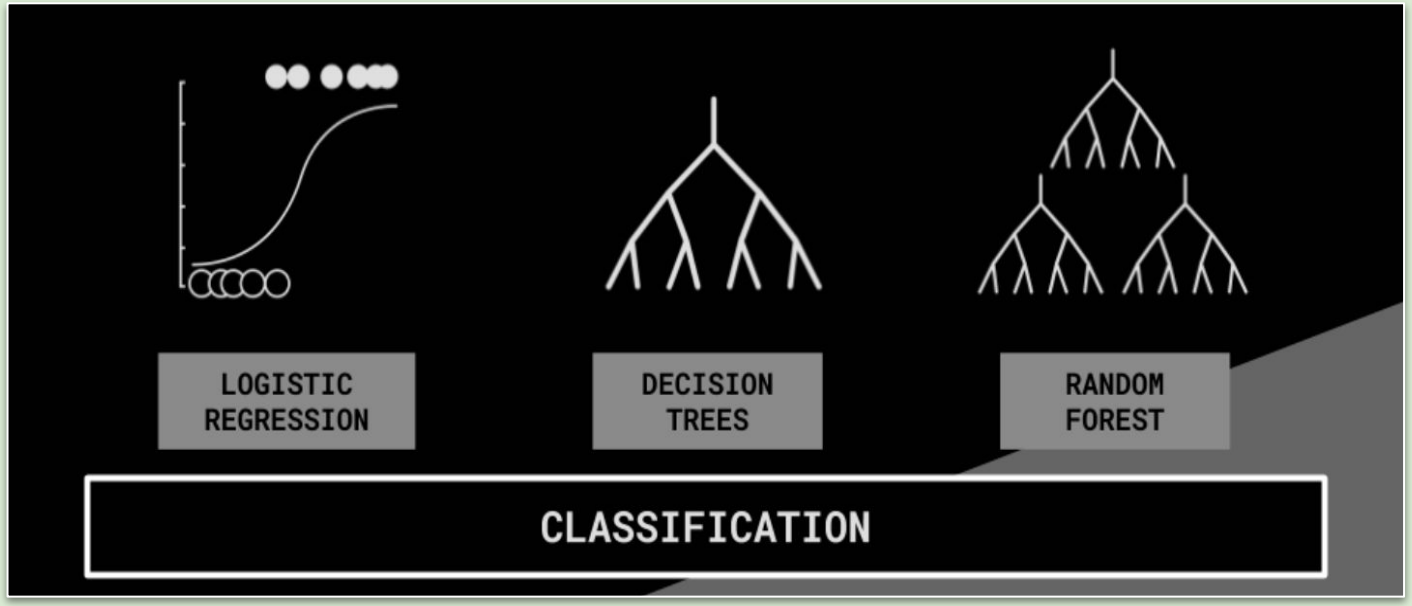
We are using Decision Tree, Random Forest, and Logistic Regression for our project to predict churning of customers for subscription-based products. Decision Tree, Logistic Regression, and Random Forest models are commonly used for churn prediction due to their strengths in classification tasks.

Why These Models:

- Logistic Regression: Offers insights into feature influence.
- Decision Trees: Provide interpretable decision paths.
- Random Forest: Captures complex relationships, boosts accuracy.

Outcomes:

- Actionable insights to identify churn risks.
- Effective strategies to improve retention and loyalty.
- Supports sustainable growth in a competitive market.



ANALYSIS OVERVIEW

The combination of Logistic Regression, Decision Tree, and Random Forest models provides a robust framework for churn prediction and actionable insights. While Logistic Regression offers interpretability, Decision Trees highlight clear decision paths, and Random Forest captures complex interactions for improved accuracy. By leveraging these insights, subscription-based businesses can proactively address churn, improve customer retention, and enhance profitability.

FOCUS AREA

Understanding Customers

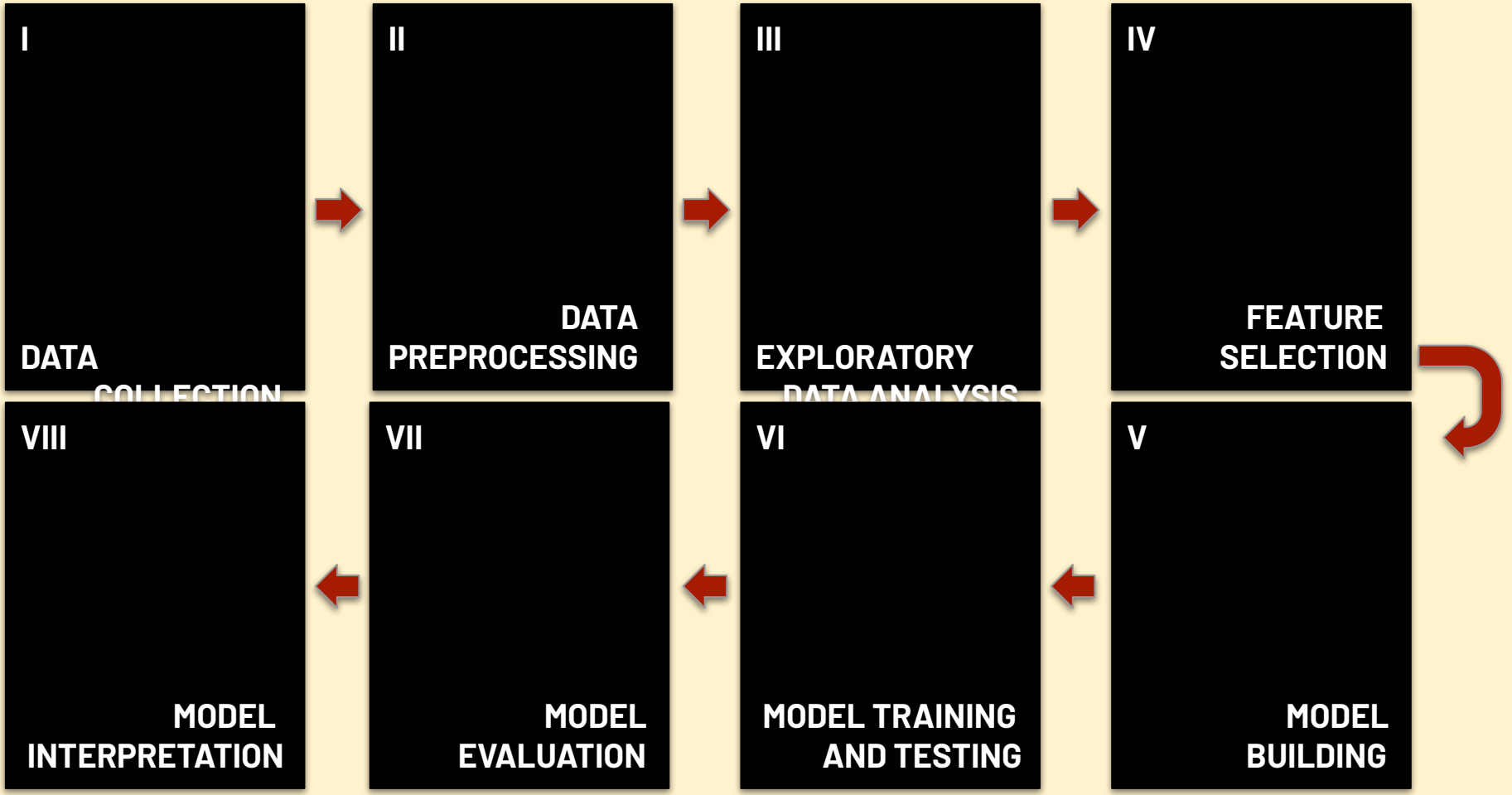
Identifying Key Predictors

Predicting Outcomes

02

Action Plan

FLOWCHART


I
DATA
II
DATA
PREPROCESSING
III
EXPLORATORY
IV
FEATURE
SELECTION
V
MODEL
BUILDING
VI
MODEL TRAINING
AND TESTING
VII
MODEL
EVALUATION
VIII
MODEL
INTERPRETATION

DATA COLLECTION

We have taken Customer Subscription Data from Kaggle Dataset repository.
This data is about a subscription-based digital product offering financial advisory that includes newsletters, webinars, and investment recommendations. The offering has a couple of varieties, annual subscription, and digital subscription. The product also provides daytime support for customers to reach out to a care team that can help them with any product-related questions and sign-up/cancellation-related queries.

The data set contains the following information:

- Customer Sign-up and Cancellation Dates: Detailed records of customer start and end dates at the product level, providing timing of cancellations.
- Subscription Details: Information about the subscription plan, payment frequency, and contract renewal status.
- Customer Demographics: Demographic data such as age, location, and potentially other customer attributes, offering context on the types of customers who are more likely to cancel.
- Product Pricing Information: Details on product pricing, which may correlate with churn likelihood.

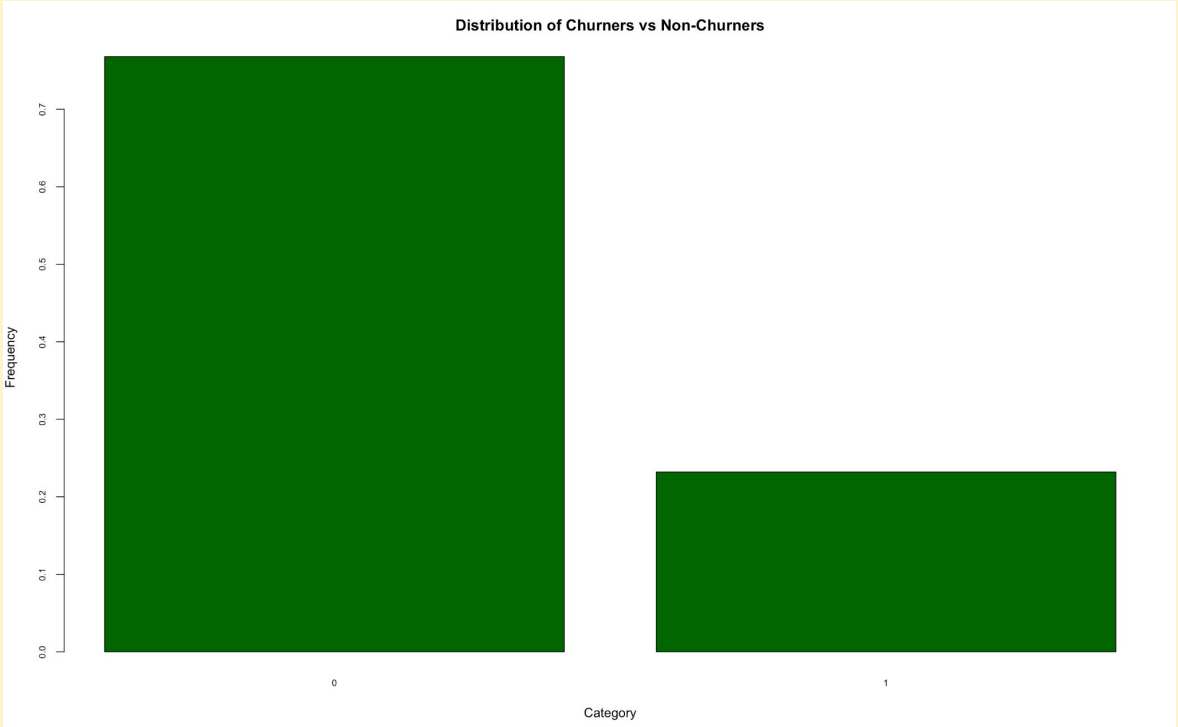
DATA PREPROCESSING

Effective data preprocessing is essential for building a reliable and accurate churn prediction model. The preprocessing steps applied to this dataset include handling missing values, encoding categorical variables, and scaling numerical features. By transforming the data into a clean and structured format, we lay a solid foundation for accurate analysis and model training, for predictions in the churn prediction model.

Understanding the data	Identifying the type of the variables and their features
Checking for missing data	No missing data
Encoding Categorical Variable	Binary encoding of the categorical variables
Scaling Feature	Normalization of data
Feature Engineering	Creating new variables that better capture customer behavior, Customer Tenure
Target Variable	Defining a target variable, that says, churn or not

EXPLORATORY DATA ANALYSIS

Distribution of Non Churners VS Churners

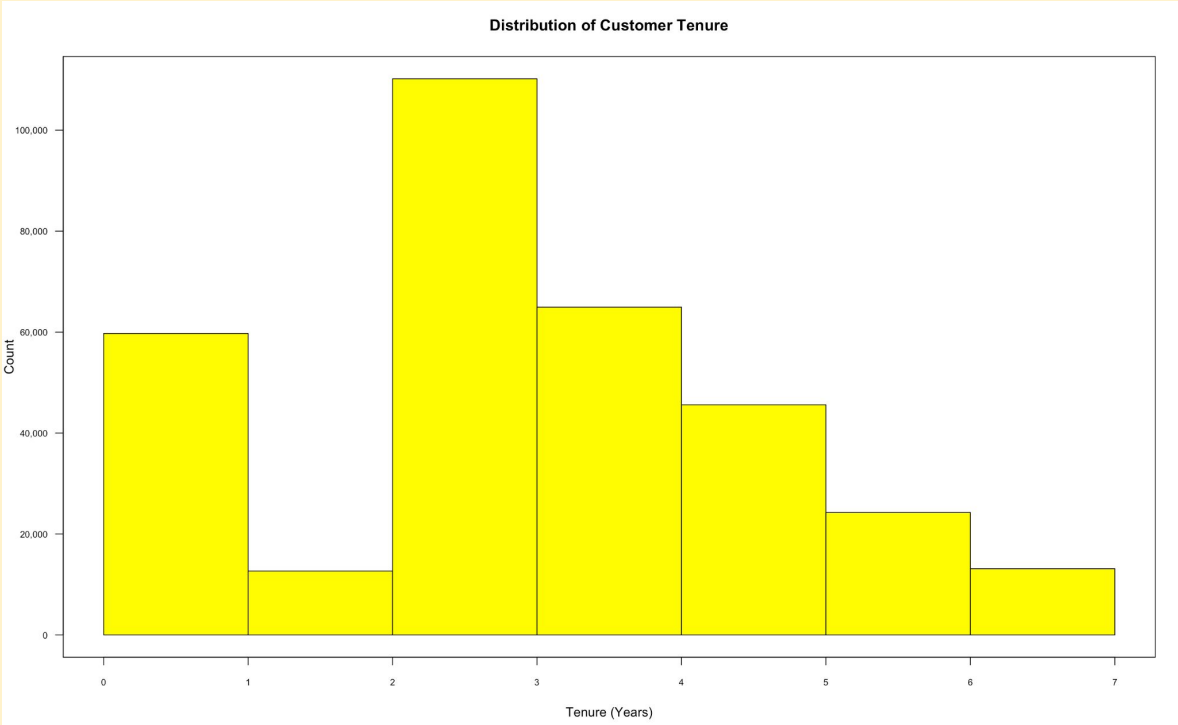


#Number of individuals who did not churn (0)
253826
Freq: 0.7679782

#Number of individuals who churned (1)
76686
Freq: 0.2320218

EXPLORATORY DATA ANALYSIS

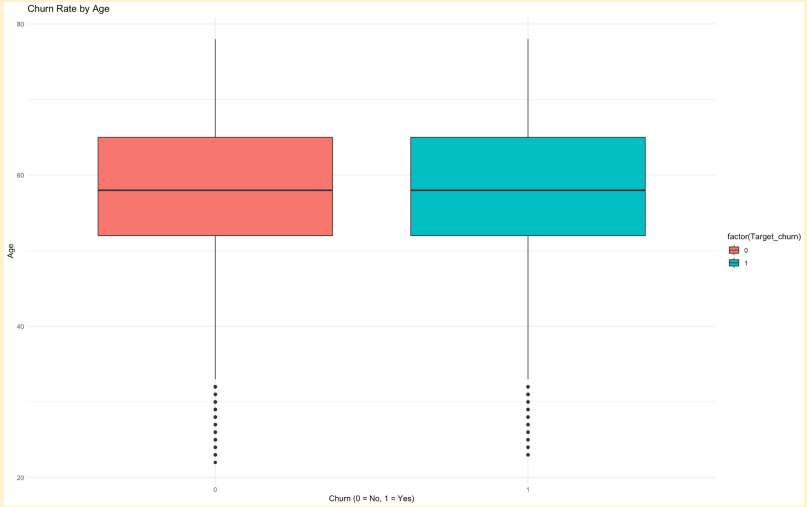
Customer Tenure

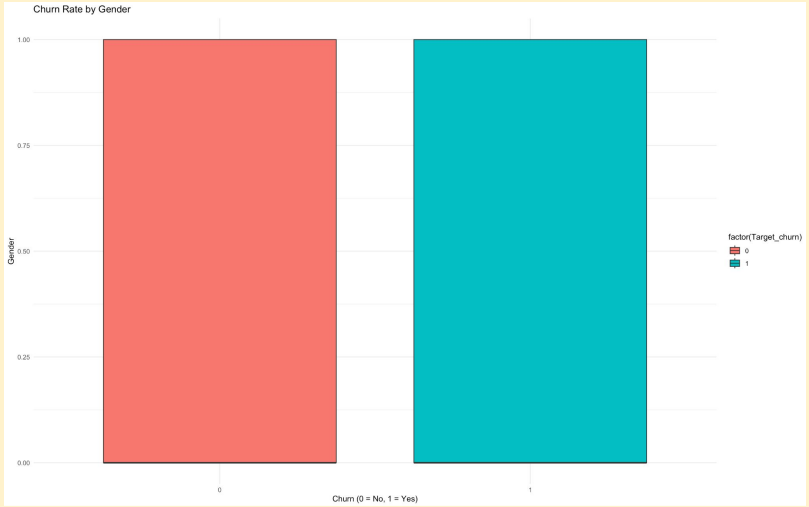


*This is a **Distribution of Customer Tenure** representing the number of customers who have been active for a certain range of years (tenure).*

EXPLORATORY DATA ANALYSIS

Churn Rate by Demographic Feature (Age & Gender)





Age Distribution:

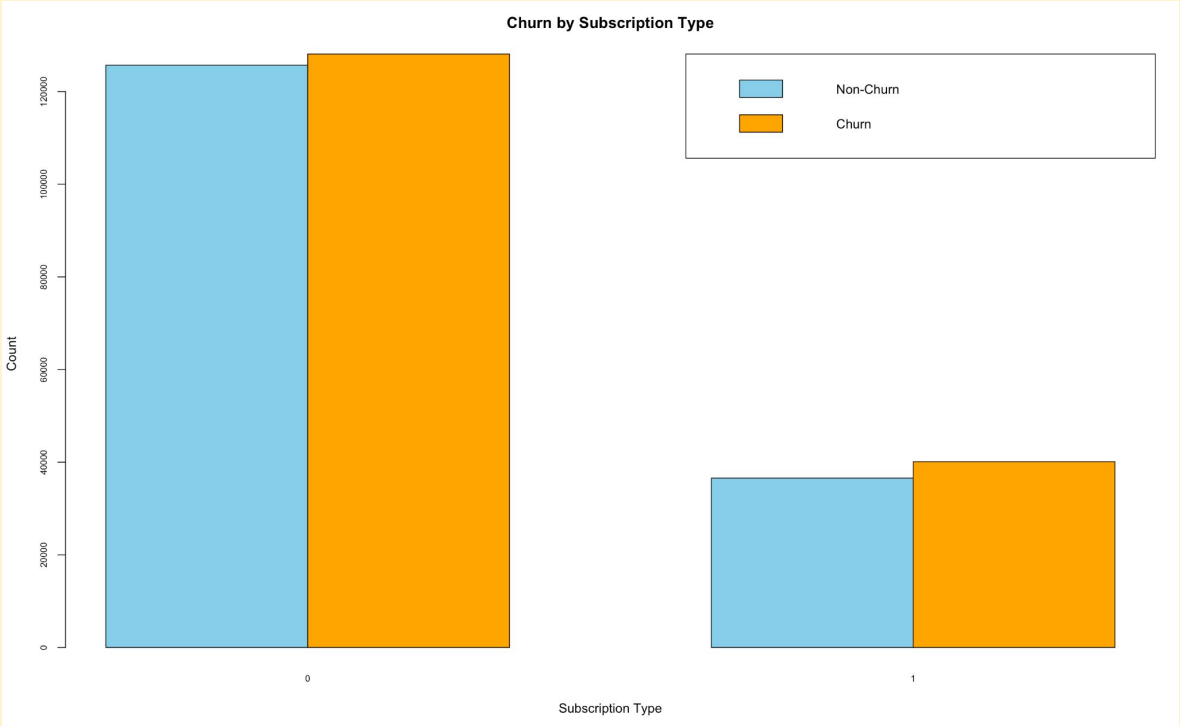
Both churned (1) and non-churned (0) customers have a similar median age, suggesting that age might not significantly influence churn directly.

Gender Distribution:

The chart does not show significant gender-specific differences in churn rates. This suggests gender may not be a strong predictor of churn.

EXPLORATORY DATA ANALYSIS

Subscription Type



Subscription Type 0 (i.e., Monthly):

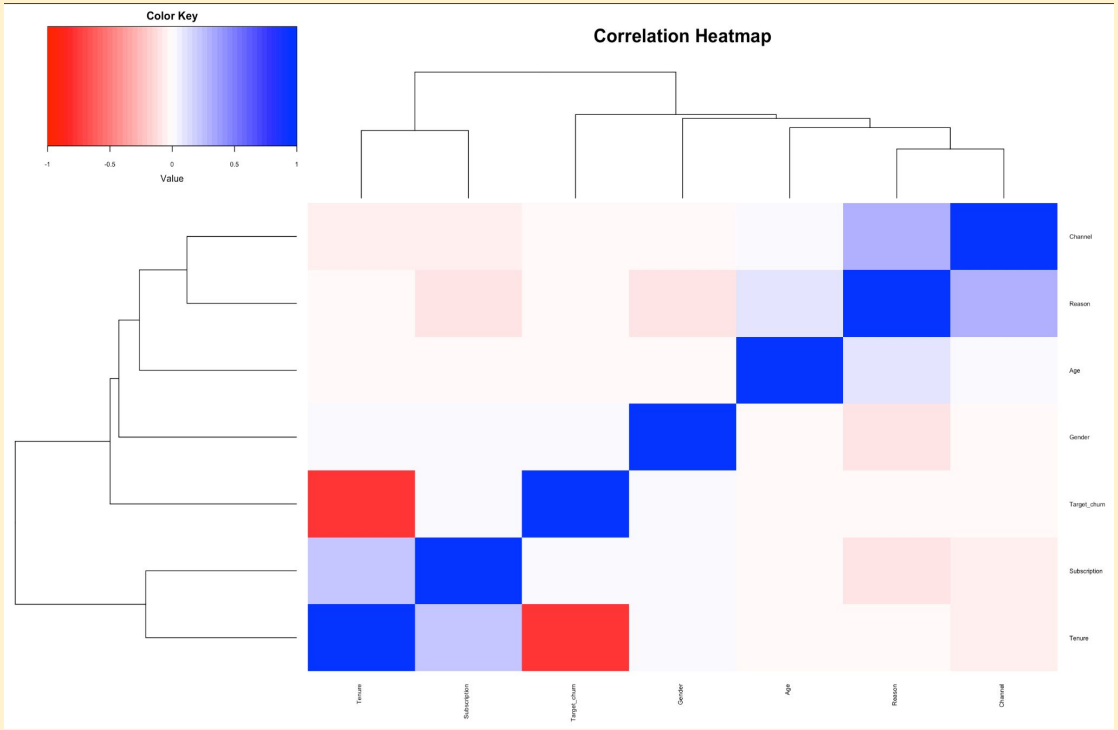
- *The monthly subscription has a larger customer base compared to the annual subscription.*
- *Despite the larger customer base, there is a higher proportion of churn among monthly subscribers.*
- *Monthly customers may lack long-term commitment, making them more susceptible to churn. Strategies such as offering discounts for upgrading to annual plans, loyalty programs, or additional value-added services could help reduce churn.*

Subscription Type 1 (i.e., Annual):

- *The annual subscription has fewer customers overall but exhibits a significantly lower churn rate compared to monthly subscriptions.*
- *This suggests that customers on annual plans are more committed, likely due to the upfront investment. Ensuring these customers remain satisfied through excellent customer service, periodic engagement, and loyalty benefits will help retain them.*

EXPLORATORY DATA ANALYSIS

Feature Correlation



Tenure and Churn: *There is a strong negative correlation between Tenure and Target_Churn. This suggests that customers with longer tenure are less likely to churn. Businesses should focus on strategies that increase tenure, such as loyalty programs or long-term benefits.*

Subscription and Target_Churn: *A slight correlation exists between Subscription Type and Target_Churn, indicating that the type of subscription (e.g., monthly vs. annual) has a minor impact on churn rates. Annual subscriptions are associated with reduced churn, emphasizing the value of promoting longer-term plans.*

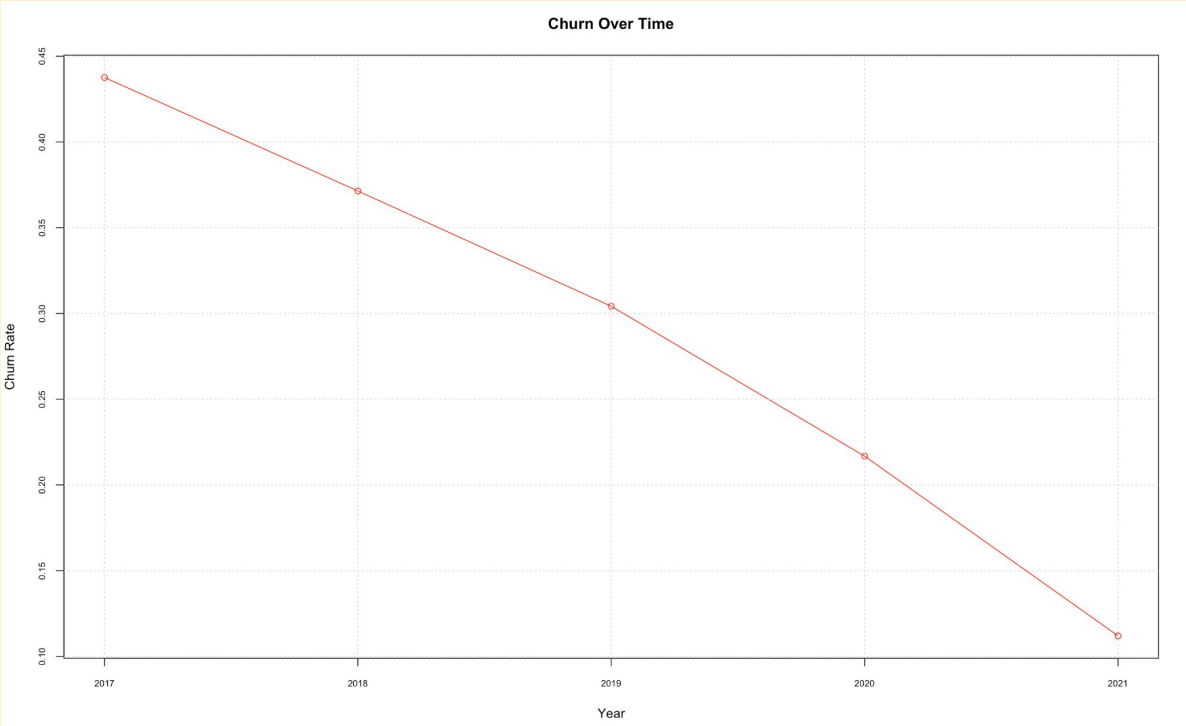
Age and Churn: *The correlation between Age and Target_Churn is minimal, indicating age has a negligible effect on churn prediction. However, age-specific campaigns may still improve customer engagement.*

Channel and Churn: *The correlation between Channel and Target_Churn is weak. This suggests the channel through which customers sign up (e.g., online, phone) has little influence on churn. Businesses can focus more on other predictors like tenure or subscription type.*

Reason and Churn: *Weak correlation between Reason and Target_Churn suggests that while reasons for signup may not directly affect churn, they could be explored further for indirect insights or segmentation.*

EXPLORATORY DATA ANALYSIS

Churn over time



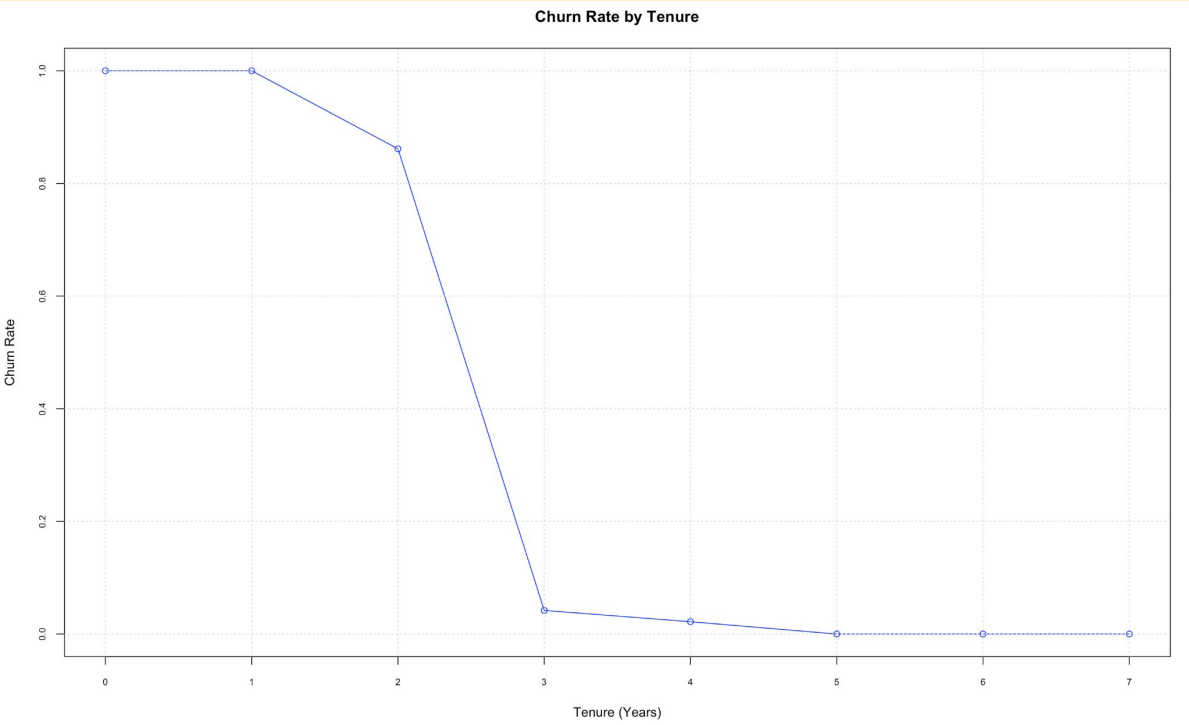
Declining Trend in Churn Rates: *The churn rate has consistently decreased from 2017 to 2021. This indicates effective retention strategies or other business improvements contributing to customer satisfaction.*

Impact of Retention Efforts: *A decline in churn might reflect initiatives such as better customer support, loyalty programs, improved product offerings, or incentivized long-term subscriptions.*

Annual Analysis: *Observing the consistent decline suggests a need to investigate what measures were introduced each year to replicate successful strategies in the future.*

Future Focus: *Although the churn rate has decreased, maintaining or further reducing churn requires continuous analysis of customer needs and behaviors.*

Churn over tenure



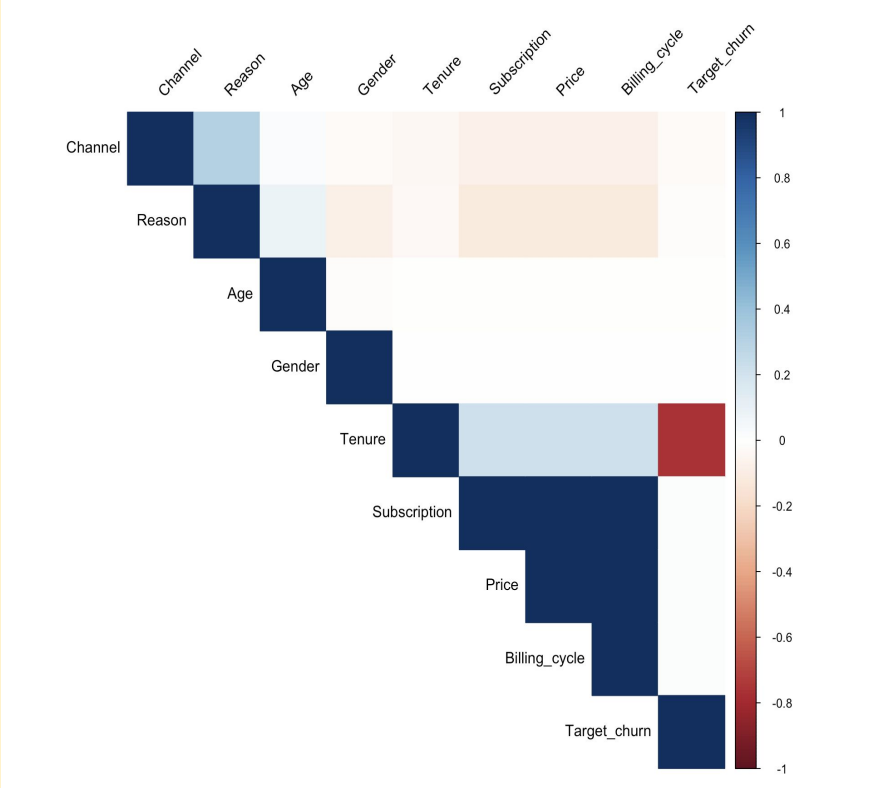
High Churn in Early Tenure: *Customers with a tenure of 0 to 2 years show the highest churn rates. This indicates that new customers are more likely to leave, suggesting challenges in the onboarding experience or unmet initial expectations.*

Sharp Decline After Year 2: *The churn rate significantly drops after the second year, indicating that customers who stay beyond this period are more likely to remain loyal.*

Stable Long-Term Customers: *From 3 years onward, churn rates approach zero. Long-tenured customers demonstrate strong loyalty, possibly due to established trust and satisfaction with the service.*

EXPLORATORY DATA ANALYSIS

Correlation Analysis



Correlation Matrix:

	Channel	Reason	Age	Gender	Tenure	Subscription	Price	Billing_Cycle	Target_Churn
Channel	1.000000	0.313242	0.025687	-0.028264	-0.044096	-0.070334	-0.070334	-0.070334	-0.029692
Reason	0.313242	1.000000	0.080208	-0.086914	-0.034887	-0.115126	-0.115126	-0.115126	-0.014857
Age	0.025687	0.080208	1.000000	-0.013819	-0.005180	-0.003768	-0.003768	-0.003768	-0.000276
Gender	-0.028264	-0.086914	-0.013819	1.000000	0.005142	0.007181	0.007181	0.007181	0.001320
Tenure	-0.044096	-0.034887	-0.005180	0.005142	1.000000	0.229900	0.229900	0.229900	-0.765264
Subscription	-0.070334	-0.115126	-0.003768	0.007181	0.229900	1.000000	1.000000	1.000000	0.015494
Price	-0.070334	-0.115126	-0.003768	0.007181	0.229900	1.000000	1.000000	1.000000	0.015494
Billing_Cycle	-0.070334	-0.115126	-0.003768	0.007181	0.229900	1.000000	1.000000	1.000000	0.015494
Target_Churn	-0.029692	-0.014857	-0.000276	0.001320	-0.765264	0.015494	0.015494	0.015494	1.000000

Key Insights:

- *Tenure is the most critical variable for churn prediction and should be a focus in modeling efforts.*
- *Variables with low correlation, such as Gender and Age, might not add significant value to predictive models and could potentially be dropped to simplify analysis.*
- *The subscription type and billing cycle could be further explored for interactions, but their direct impact on churn is minimal.*

FEATURE SELECTION

Correlation Analysis helps us to assess the relationship between each feature and the target variable and understand the importance of each variable in the dataset. We need to divide the variables into the ones that we want to retain and the ones we are going to remove.

MULTICOLLINEARITY

Subscription (categorical) effectively captures the relationship between price and billing cycle. For example, “annual” vs. “monthly” already defines different pricing and billing intervals. Product ID as well relates the same with Subscription.

STRONG OR WEAK CORRELATION

Age and Gender have minimal correlation, where as compared to them Channel and Reason have relatively impactful correlation. Tenure has strong correlation, suggesting that customers with higher tenure are significantly less likely to churn.

FEATURE RETENTION

Subscription, Price, and Billing_cycle, these three variables are highly multicollinear and provide redundant information. Retaining one (e.g., Subscription) while dropping others reduces redundancy.

DROPPED VARIABLES

Price can be redundant since it is implicitly represented by the Subscription type. Billing cycle often aligns perfectly with Subscription type (e.g., annual = 12, monthly = 1), making it redundant. Product ID can also be dropped for the similar r

MODEL BUILDING

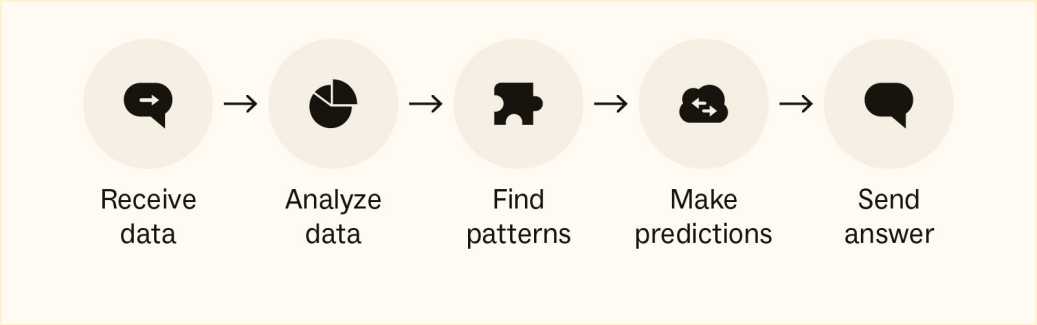
We are choosing Random Forest, Decision Tree, and Logistic Regression as our classification models because they offer a balance of interpretability, performance, and practicality compared to more complex models like neural networks, bagging, or XGBoost. While advanced models like neural networks and XGBoost often yield higher accuracy, they require extensive computational resources, hyperparameter tuning, and lack the interpretability needed for actionable business insights. By leveraging Random Forest, Decision Tree, and Logistic Regression, we ensure interpretable and resource-efficient solutions tailored for churn prediction.

Decision Tree *: Focused on interpretability and identifying primary churn indicators. Simplicity and interpretability make it highly usable, especially for scenarios needing clear decision-making.*

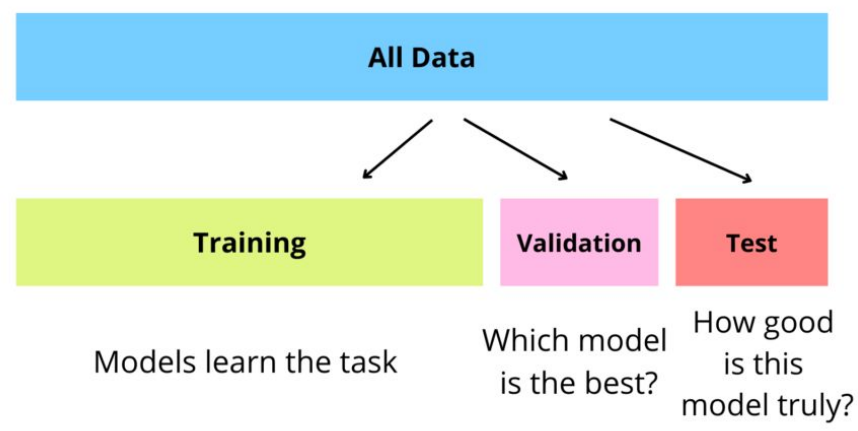
Random Forest *: Balances accuracy and capturing complex data relationships, often preferred for its robustness. Most suitable for complex datasets with high accuracy and robustness.*

Logistic Regression *: Emphasized for statistical clarity and understanding feature impact. Lightweight and efficient, ideal for binary classification tasks with simpler data.*

MODEL BUILDING PROCESS



MODEL TRAINING AND TESTING



In this phase, we evaluate multiple algorithms, adjusting parameters to optimize performance for accurate and reliable predictions. The goal is to select the best-performing model that balances prediction accuracy with interpretability, providing actionable insights for targeted retention strategies.

➔ *Split the data into training and validation sets. Typically, 80% of the data is used for training the model, and 20% is used for validating it.*
➔ *Train the selected classification model (Logistic Regression, Decision Tree, Random Forest) on the training data, using the selected features to predict the churn outcome.*

TRAINING SET

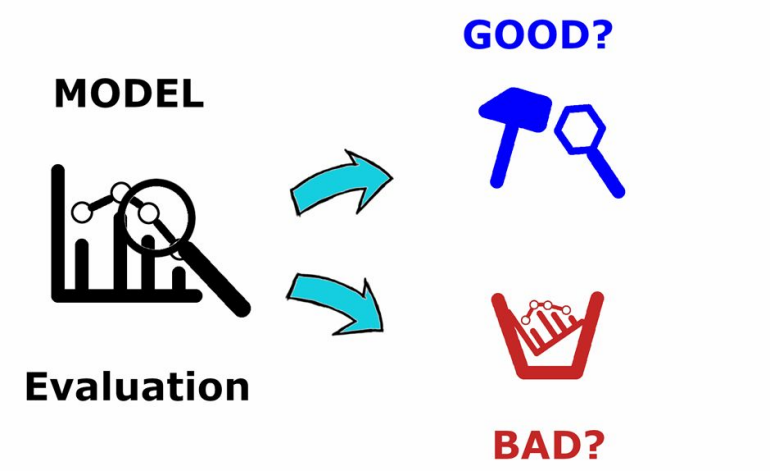
The subset of data used to train a machine learning model

VALIDATION SET

The intermediary subset of data used during the model development process to fine-tune hyperparameters

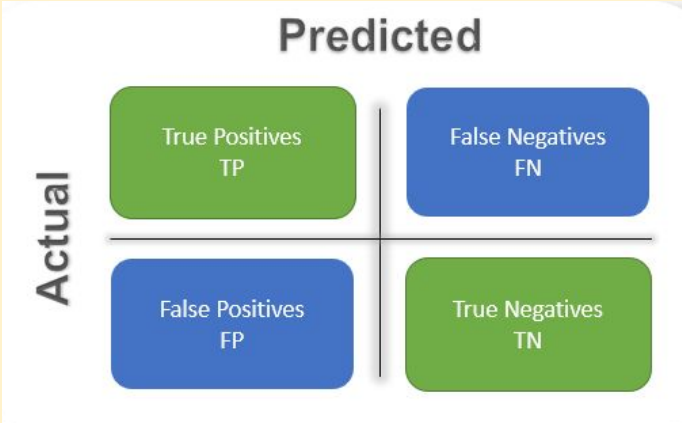
MODEL EVALUATION

Model evaluation is a crucial step in our churn prediction project, where we assess the performance and reliability of our trained model using various metrics such as accuracy, precision, recall, F1-score, and AUC-ROC curve used to measure the model's effectiveness in predicting customer churn. By comparing these metrics across different models and configurations, we can identify the most accurate and balanced model for deployment.



Evaluation Metrics

- *Accuracy: Proportion of correctly classified instances (useful for balanced datasets).*
- *Precision: Ratio of true positives to all predicted positives (important when false positives need minimization).*
- *Recall (Sensitivity): Ratio of true positives to all actual positives (important for minimizing false negatives).*
- *F1 Score: Harmonic mean of precision and recall, balancing both.*
- *AUC-ROC: Evaluates the model's ability to distinguish between classes at various thresholds.*
- *Lift Chart: Compare the model's predictions against a random baseline to measure improvement in targeting.*
- *Confusion Matrix helps us understand how well the model performs across all classes.*



MODEL INTERPRETATION

Interpreting the results of the model is crucial for understanding which factors are driving churn.

DECISION TREE

➔ Feature Importance:
In a Decision Tree, feature importance is determined by the features that
split the data most effectively at each level of the tree. The higher a feature appears in the tree, the more critical it is for making accurate predictions.

➔ Sensitivity Analysis:
Decision Trees allow easy interpretation of how changes in key variables affect churn.

RANDOM FOREST

➔ Feature Importance:
In a Random Forest model, feature importance is derived from the average importance of each feature across all decision trees in the forest.

➔ Sensitivity Analysis:
Sensitivity analysis in Random Forest can be done by evaluating how predictions change when important features are varied.

➔ Random Forest also allows you to conduct partial dependence analysis, which shows the relationship between a feature and the probability of churn while averaging out the effects of other features.

LOGISTIC REGRESSION

➔ Feature Importance:
In Logistic Regression, feature importance is reflected by the magnitude of the coefficients for each predictor. The sign and size of these coefficients indicate how each feature impacts the likelihood of churn.

➔ Sensitivity Analysis:
In Logistic Regression, sensitivity analysis involves examining how changes in feature values affect churn probability, leveraging the model's linear nature.

03

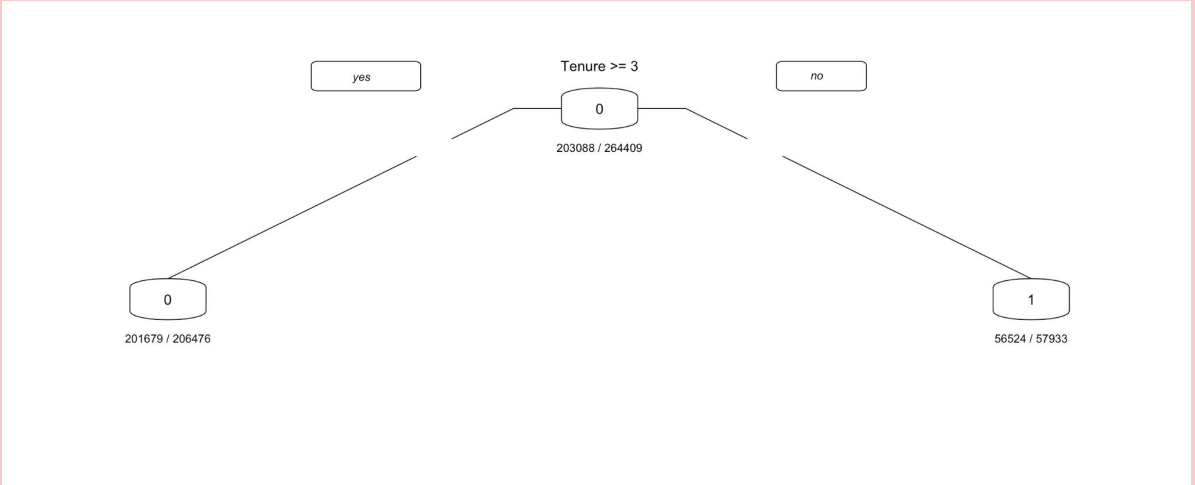
Working

DECISION TREE

Root Node (Tenure >= 3):
The model starts with the variable Tenure as the first decision point.
If a customer's tenure is greater than or equal to 3 years, they are directed to the left branch; otherwise, they go to the right branch.
Left Branch (Tenure >= 3 → No Churn):
For customers with tenure >= 3 years, the majority are predicted not to churn (class 0).
This is evident as the leftmost leaf node shows a high number of non-churners (201,679 out of 206,476 total cases).
Right Branch (Tenure < 3 → Churn):
For customers with tenure < 3 years, the right branch shows a higher likelihood of churn (class 1).
The final leaf node shows that a significant portion of customers in this category are predicted to churn (56,524 out of 57,933 cases).

This simple tree structure effectively highlights how tenure influences churn, making it a key feature for predictive modeling and customer retention strategies.

DECISION TREE FOR CHURN PREDICTION



DECISION TREE

In a deeper decision tree, to address the issue of overfitting, pruning is performed by selecting the Complexity Parameter (CP) value associated with the minimum cross-validation error (xerror). Upon analyzing the CP table, it is observed that the second entry has the lowest xerror value, which aligns with the tree previously built. Therefore, the earlier tree already represents the optimal model for prediction. As a result, no further pruning is required, and the existing tree can be utilized for subsequent predictions effectively.

	CP	nsplit	rel error	xerror	xstd
1	0.8987948664	0	1.00000	1.00000	0.0035392
2	0.0000114153	1	0.10121	0.10121	0.0012695
3	0.0000081538	17	0.10099	0.10168	0.0012724
4	0.0000054359	29	0.10090	0.10218	0.0012755
5	0.0000044475	35	0.10086	0.10231	0.0012763
6	0.0000040769	46	0.10081	0.10240	0.0012768
7	0.0000027179	54	0.10078	0.10241	0.0012769
8	0.0000000000	78	0.10072	0.10249	0.0012774

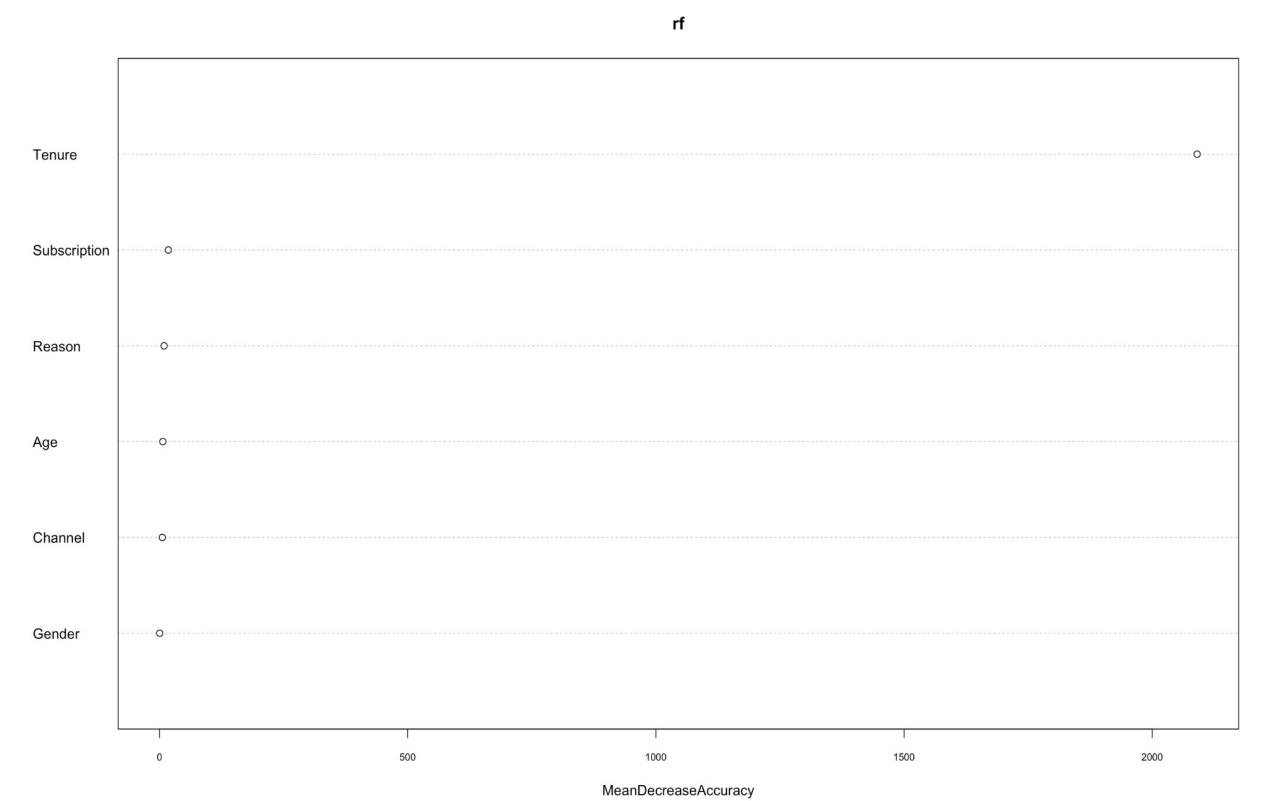
Overfitting occurs when a model learns the training data too well, including its noise and irrelevant patterns, leading to poor generalization of new data. While the model achieves high accuracy on the training set, its performance on validation or test sets deteriorates. Overfitting often results from overly complex models, insufficient data, or lack of regularization. Techniques like pre-pruning or post-pruning helps us to handle overfitting problem.

DEEPER DECISION TREE

RANDOM FOREST

Random Forest is an ensemble learning method used for classification, regression, and other tasks. It works by building multiple decision trees during training and combining their outputs to improve performance and reduce overfitting.
After we have run the Random Forest model, this plot shows the feature importance from the Random Forest model based on the MeanDecreaseAccuracy metric, which reflects how much each feature contributes to the model's predictive accuracy.
Here, Tenure is the most important feature, having the highest impact on model accuracy. Other features like Subscription, Reason, Channel, Age, and Gender have lower importance. This highlights that customer tenure plays a critical role in predicting churn, while demographic factors like Age and Gender contribute minimally to the model's performance.

VARIABLE IMPORTANCE PLOT



LOGISTIC REGRESSION

LOGISTIC REGRESSION SUMMARY

The logistic regression model identifies key predictors of customer churn. Significant variables include Channel, Reason, Tenure, and Subscription, indicating their strong influence on churn behavior.

For instance, longer Tenure significantly reduces the likelihood of churn, while certain Subscription types increase it. Variables like Age and Gender are not statistically significant, suggesting they have minimal impact on churn.

```
Call:
glm(formula = Target_churn ~ ., family = "binomial", data = train.df)

Coefficients:
              Estimate Std. Error  z value            Pr(>|z|)
(Intercept)  8.1718228  0.1010478   80.871 <0.0000000000000002 ***
Channel     -0.8930515  0.0447631  -19.951 <0.0000000000000002 ***
Reason       0.3750693  0.0265926   14.104 <0.0000000000000002 ***
Age         -0.0009559  0.0012963   -0.737               0.461
Gender       0.0258283  0.0247998    1.041               0.298
Tenure      -3.8995008  0.0248582 -156.870 <0.0000000000000002 ***
Subscription 1.6902297  0.0256019   66.020 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 286399  on 264408  degrees of freedom
Residual deviance:  50175  on 264402  degrees of freedom
AIC: 50189

Number of Fisher Scoring iterations: 9
```

EVALUATING MODELS

This table summarizes the performance evaluation based on the Confusion Matrices of three machine learning models— Decision Tree, Random Forest, and Logistic Regression—used for predicting customer churn. The metrics are evaluated on both training and validation datasets to assess model performance and potential overfitting.

Measures	Decision Tree		Random Forest		Logistic Regression	
	Training data	Validation data	Training data	Validation data	Training data	Validation data
Accuracy	97.6%	97.5%	97.6%	97.5%	97.5%	97.5%
Error Rate	2.3%	2.4%	2.3%	2.4%	2.4%	2.4%
Sensitivity	92.1%	91.8%	92.1%	91.8%	91.8%	91.4%
Specificity	99.3%	99.3%	99.3%	99.3%	99.3%	99.3%
False Positive Rate	0.69%	.68%	.69%	.67%	.66%	.66%
False Negative Rate	7.8%	8.1%	7.8%	8.1%	8.1%	8.5%
Precision	97.5%	97.6%	97.5%	97.6%	97.6%	97.6%
F1 score	94.7%	94.6%	94.7%	94.6%	94.6%	94.4%

Accuracy: *All models achieve high accuracy (~97.5%-97.6%) on both training and validation datasets, indicating that they are effective at predicting churn and non-churn classes.*

Error Rate: *The error rates (~2.3%-2.4%) are consistent across all models, suggesting minimal misclassification overall.*

Sensitivity (Recall): *The values (~91.4%-92.1%) are slightly lower than the overall accuracy, indicating that while the models are accurate, there is some room for improvement in detecting churners.*

Specificity: *All models achieve extremely high specificity (~99.3%), meaning they are highly effective at identifying non-churn customers.*

False Positive Rate (FPR): *The FPR values (~0.66%-0.69%) are low, indicating that very few non-churners are mistakenly classified as churners.*

False Negative Rate (FNR): *The FNR (~7.8%-8.5%) is relatively higher than the FPR, showing that some churners are being missed, which could be critical for retention strategies.*

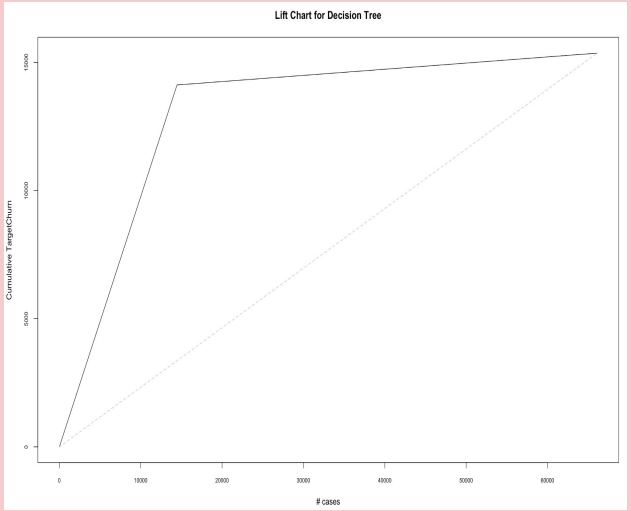
Precision: *Precision values (~97.5%-97.6%) indicate that when the models predict a customer will churn, they are correct in most cases, minimizing false positives.*

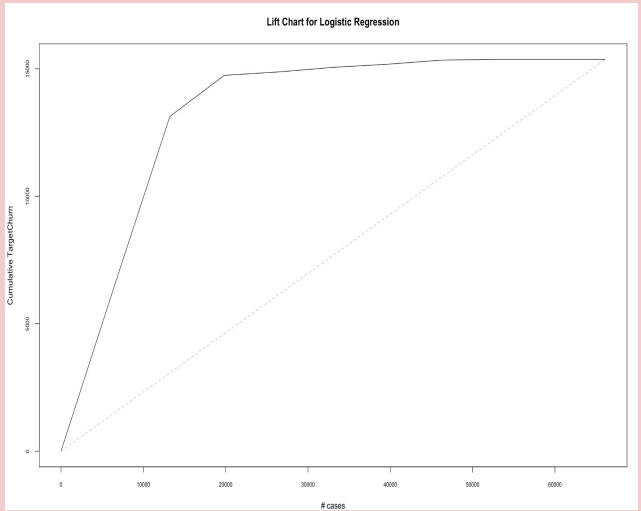
F1 Score: *The F1 scores (~94.4%-94.7%) provide a balanced measure of precision and recall, confirming that the models perform well overall in predicting churn.*

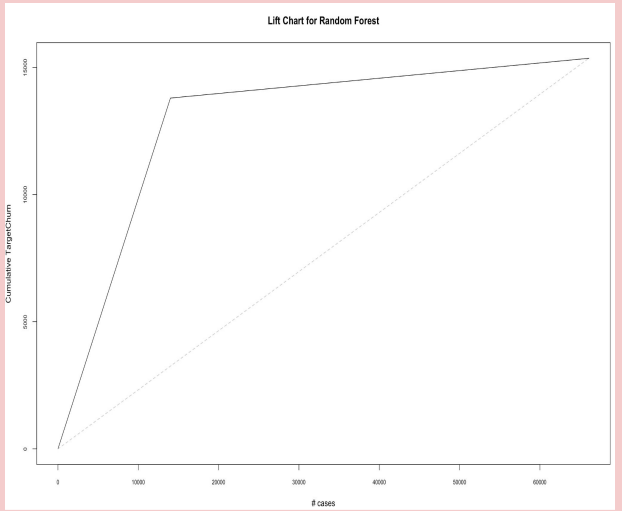
EVALUATING MODELS

LIFT CHART :

The lift charts for the Decision Tree, Random Forest, and Logistic Regression models compare their ability to predict customer churn against a random baseline. The Decision Tree demonstrates the highest lift, with a steep curve that rises quickly and effectively identifies churners, especially among the top-ranked cases. While Random Forest performs well, its curve is slightly less steep, indicating it may not prioritize high-risk churners as effectively as the Decision Tree. Logistic Regression has the least steep curve, showing relatively lower effectiveness in identifying churners compared to the Decision Tree and Random Forest. Overall, the Decision Tree emerges as the best-performing model, making it the most effective choice for predicting churn and informing customer retention strategies.





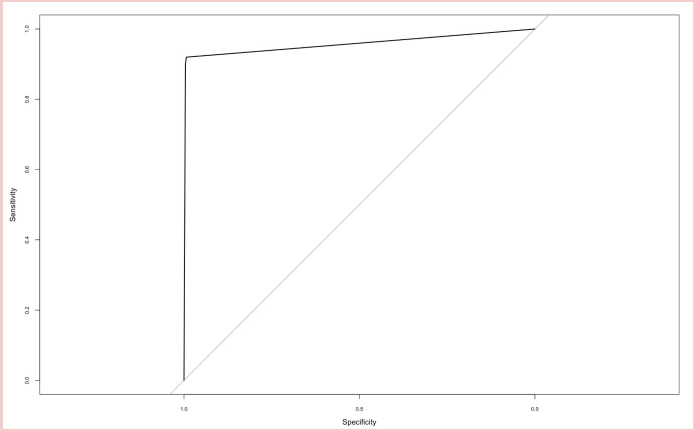


EVALUATING MODELS

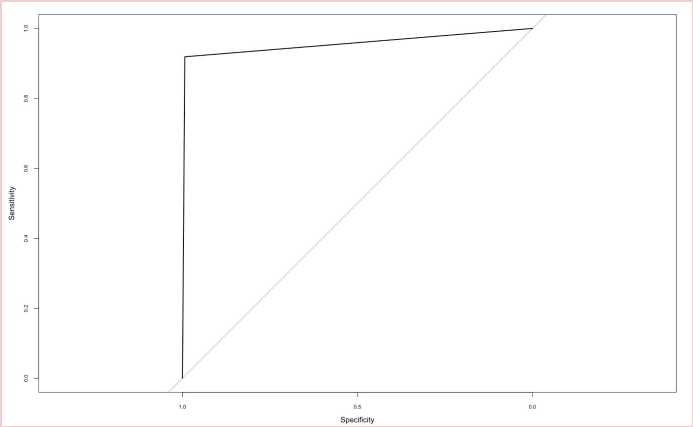
AUC-ROC (Area Under the Curve - Receiver Operating Characteristic):

The AUC-ROC is to evaluate the performance of a binary classification model. It provides a comprehensive measure of a model's ability to distinguish between the positive and negative classes, considering all possible classification thresholds.

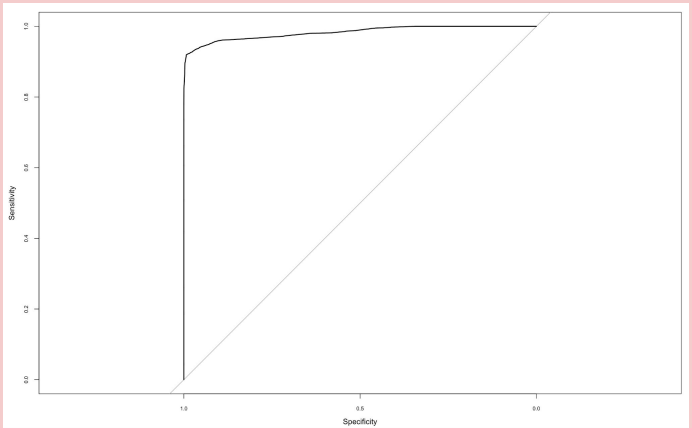
ROC for Random Forest



ROC for Decision Tree



ROC for Logistic Regression



EVALUATING MODELS

AUC-ROC (Area Under the Curve - Receiver Operating Characteristic):

DECISION TREE

Area under the curve: 0.9561

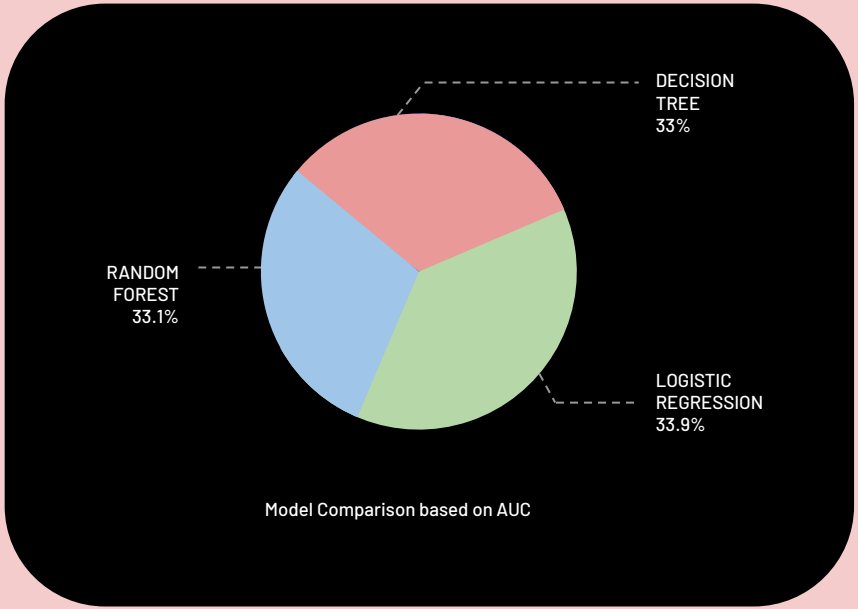
RANDOM FOREST

Area under the curve: 0.9577

LOGISTIC REGRESSION

Area under the curve: 0.9829

Based on the Lift Chart and AUC values, Logistic Regression is the better model. The lift chart for Logistic Regression shows strong performance in identifying high-probability churners early on, and its results remain consistent. Additionally, the AUC value of 0.9829 for Logistic Regression is higher than both the Decision Tree (0.9561) and Random Forest (0.9578), showing it does a better job of telling churners and non-churners apart. These results make Logistic Regression the best choice for this analysis.



04

Conclusion

DEPLOYMENT AND REPORT

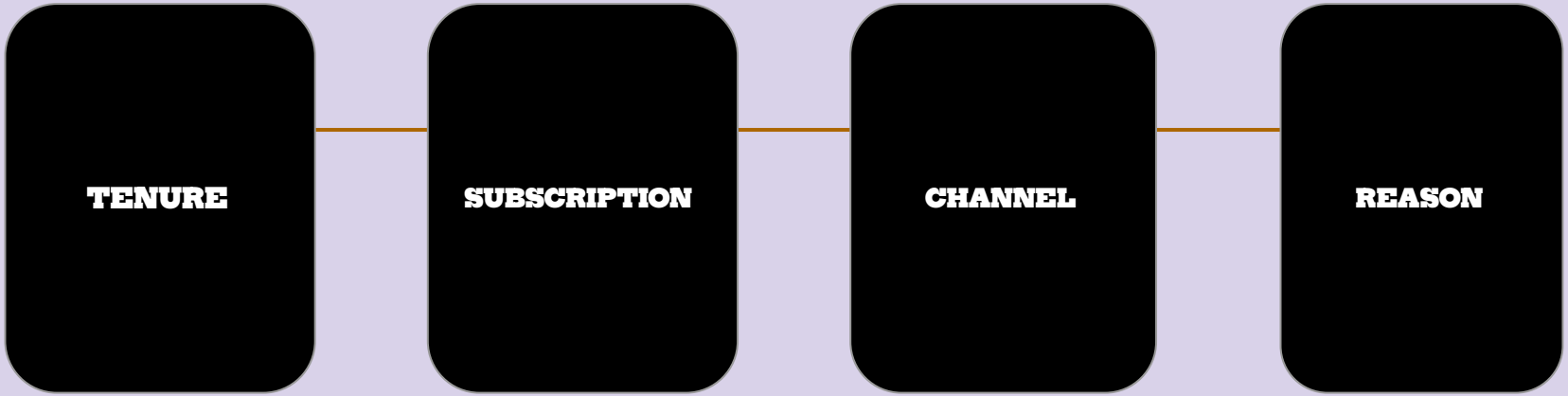
The AUC-ROC, accuracy, and F1 scores summarize the overall effectiveness of the models in predicting churn. AUC-ROC evaluates the model's ability to distinguish between churners and non-churners across all thresholds. Accuracy reflects the proportion of correctly classified cases, while F1 balances precision and recall.

If the focus is on identifying churners accurately (reducing false negatives), recall should be prioritized. High recall ensures that most actual churners are correctly identified, even if some non-churners are mistakenly flagged as churners. This approach is suitable when missing a churner (false negative) has significant consequences, such as losing valuable customers.

Alternatively, if minimizing false positives is more critical (predicting churn when it's not), the focus should shift to precision. High precision ensures that when the model predicts churn, it is likely correct, reducing unnecessary interventions or costs.

Depending on the company's requirements, whether reducing false negatives or false positives is more critical, the model with either high recall or high precision can be selected.
For example:
We can use the model with high recall when the cost or impact of missing churners is high
We can use the model with high precision when the cost of acting on false churn predictions is high, such as offering incentives or discounts to customers who were not at risk of leaving.

KEY DRIVERS



Significant Predictor of Churn

Contribute to Churn, but to a lesser degree

Contribute to Churn, but to a lesser degree

Contribute to Churn, but to a lesser degree

ACTIONABLE INSIGHTS

Pricing Recommendations:

- *Consider introducing tiered pricing with discounts for longer billing cycles to reduce churn.*

Retention Strategies:

- *Focus on retaining customers in their first three years, as churn rates are highest during this period.*
- *Offer targeted retention campaigns based on tenure and subscription type.*

Customer Segmentation:

- *Prioritize interventions for monthly subscribers and customers nearing the end of their first or second year of tenure.*
- *Use predictive models to identify high-risk churn segments for proactive outreach.*
- *These insights provide a roadmap for reducing churn while optimizing customer retention strategies and pricing models.*

SENSITIVITY ANALYSIS

A sensitivity analysis reveals that subscription type and pricing play a crucial role in influencing customer tenure and churn rates. Higher prices are associated with shorter customer tenure and increased churn, particularly for customers on monthly subscriptions. Conversely, longer billing cycles, such as yearly subscriptions, tend to extend customer tenure and reduce churn rates. This suggests that offering pricing incentives or discounts for annual subscriptions could encourage longer commitments, enhancing customer retention and minimizing churn.

This project focused on predicting customer churn for subscription-based services using Decision Tree, Random Forest, and Logistic Regression models. Through thorough data preprocessing, exploratory data analysis, and model evaluation, we identified key factors contributing to churn and developed predictive models.

Among the models, Logistic Regression proved to be the most optimal based on both the lift chart and the AUC value (0.9829), demonstrating its ability to accurately distinguish between churners and non-churners.

By using these insights from this analysis, subscription-based businesses can prioritize high-risk customers, implement targeted retention efforts, and reduce churn. This project highlights the importance of predictive analytics in improving customer loyalty and maximizing profitability.

THANK YOU