



Customer Churn Analysis- Insights and Predictions

Group 4:

Aditya Chowdhuri

KPV Abhishek

Ankit

Khyati

Navigating the Challenges of Customer Churn

Customer churn, a critical metric in the telecom sector, signifies the loss of clients or customers.

Our analysis aims to unravel the underlying patterns leading to churn and develop predictive models to identify at-risk customers.

Leveraging data-driven insights, we explore ways to enhance customer retention strategies.



Dataset Snapshot: A Window into Customer Dynamics

01

"Our dataset encompasses 3,150 customers, each described through 14 key features."

02

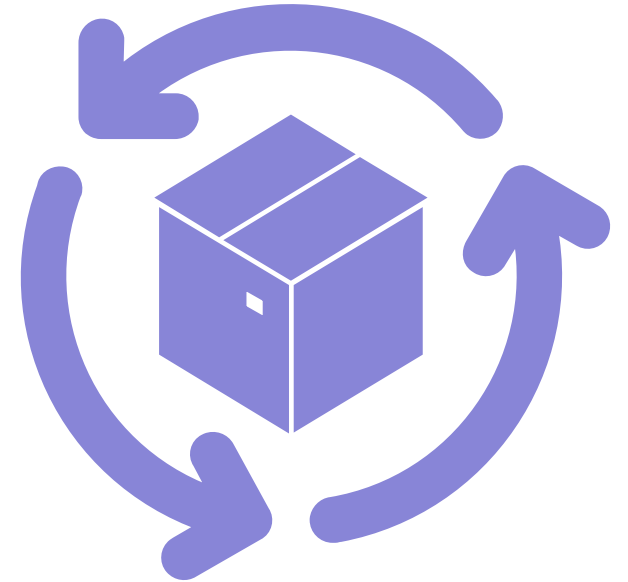
"Primary focus: Churn Status
(0: No Churn, 1: Churn),
reflecting customer retention and loss."

03

"Features range from Call Failures and Usage Patterns to Demographics and Tariff Plans, providing a holistic view of customer behavior."

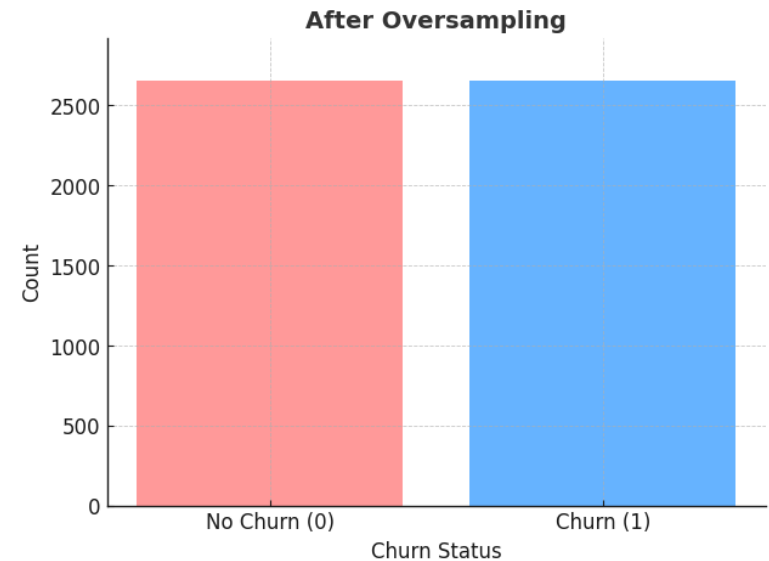
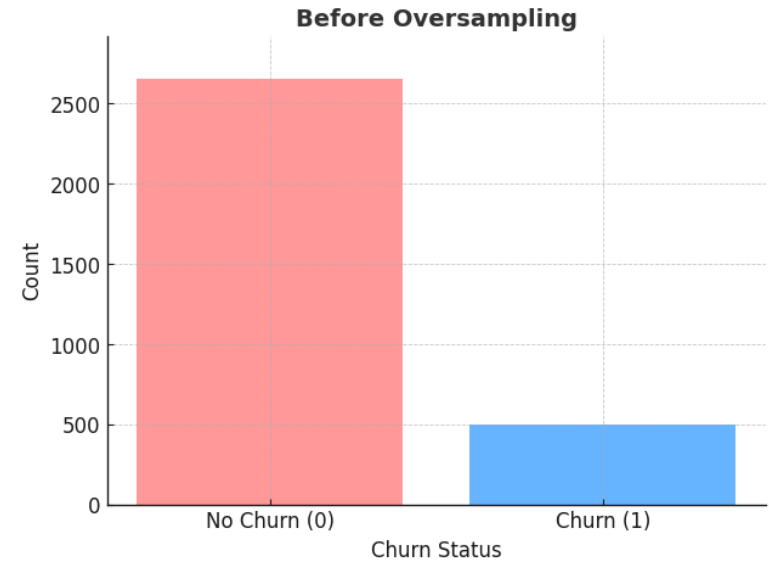
Data Preprocessing

- Initial preprocessing involved standardizing values, with specific actions like replacing certain codes
- Categorical columns were transformed into a string format
- Implemented oversampling of the minority class



Addressing Data Imbalance Through Oversampling

- Imbalance between the churned and retained customer classes.
- Employed oversampling techniques
- creating duplicate entries of the minority class to match the quantity of the majority class



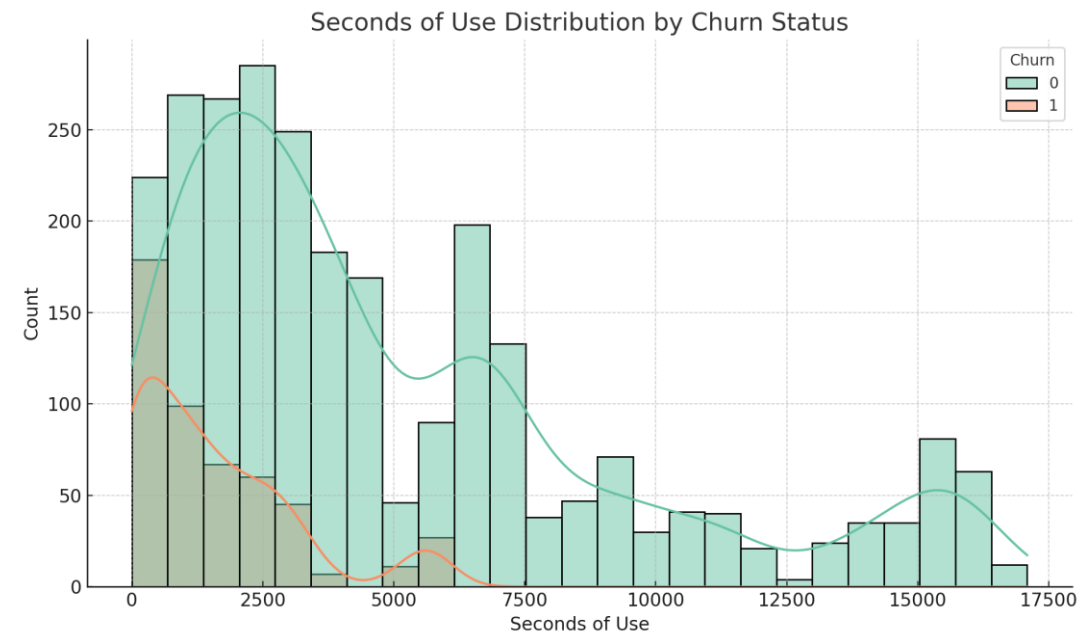
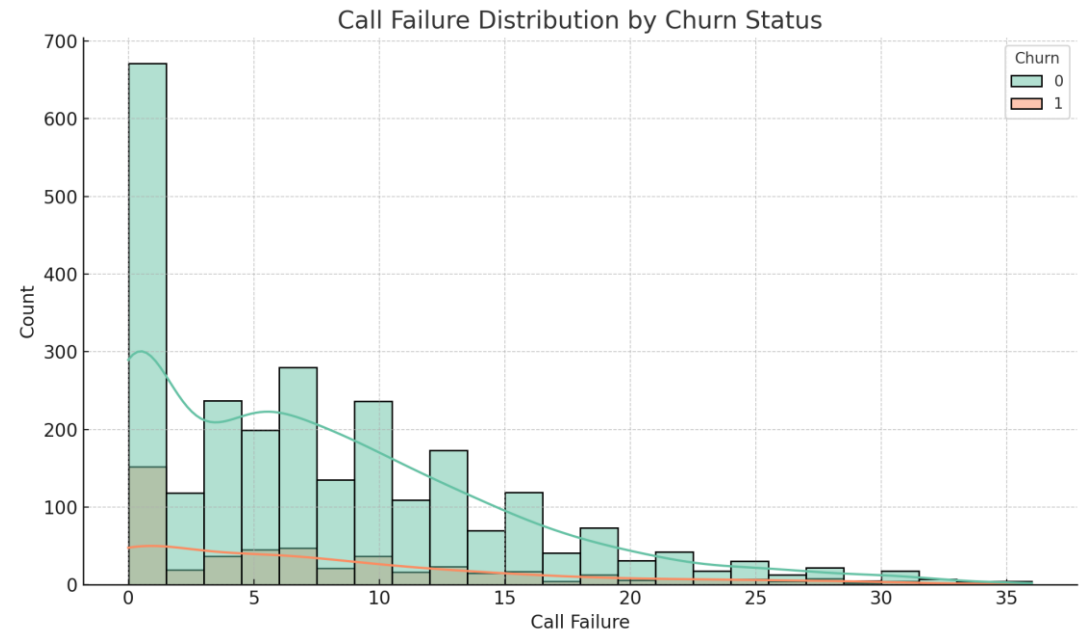
Understanding Customer Behavior

Call Failure vs Churn:

- Higher call failure rates are associated with increased churn.
- Improving call reliability can significantly reduce customer churn.

Usage (Seconds of Use) vs Churn:

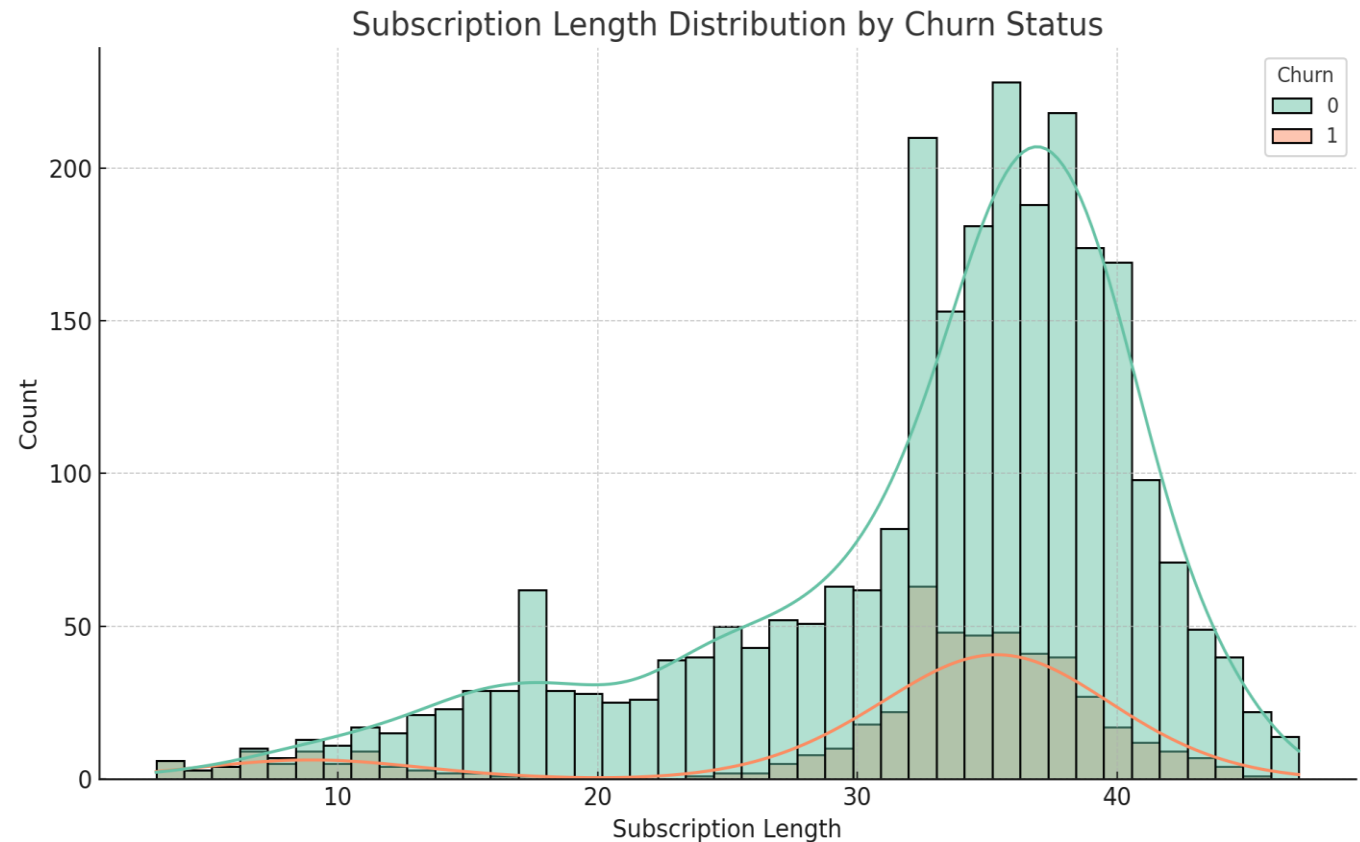
- Lower usage correlates with higher churn rates.
- Increasing customer engagement could reduce churn.



Analyzing Call Failures and Usage Patterns

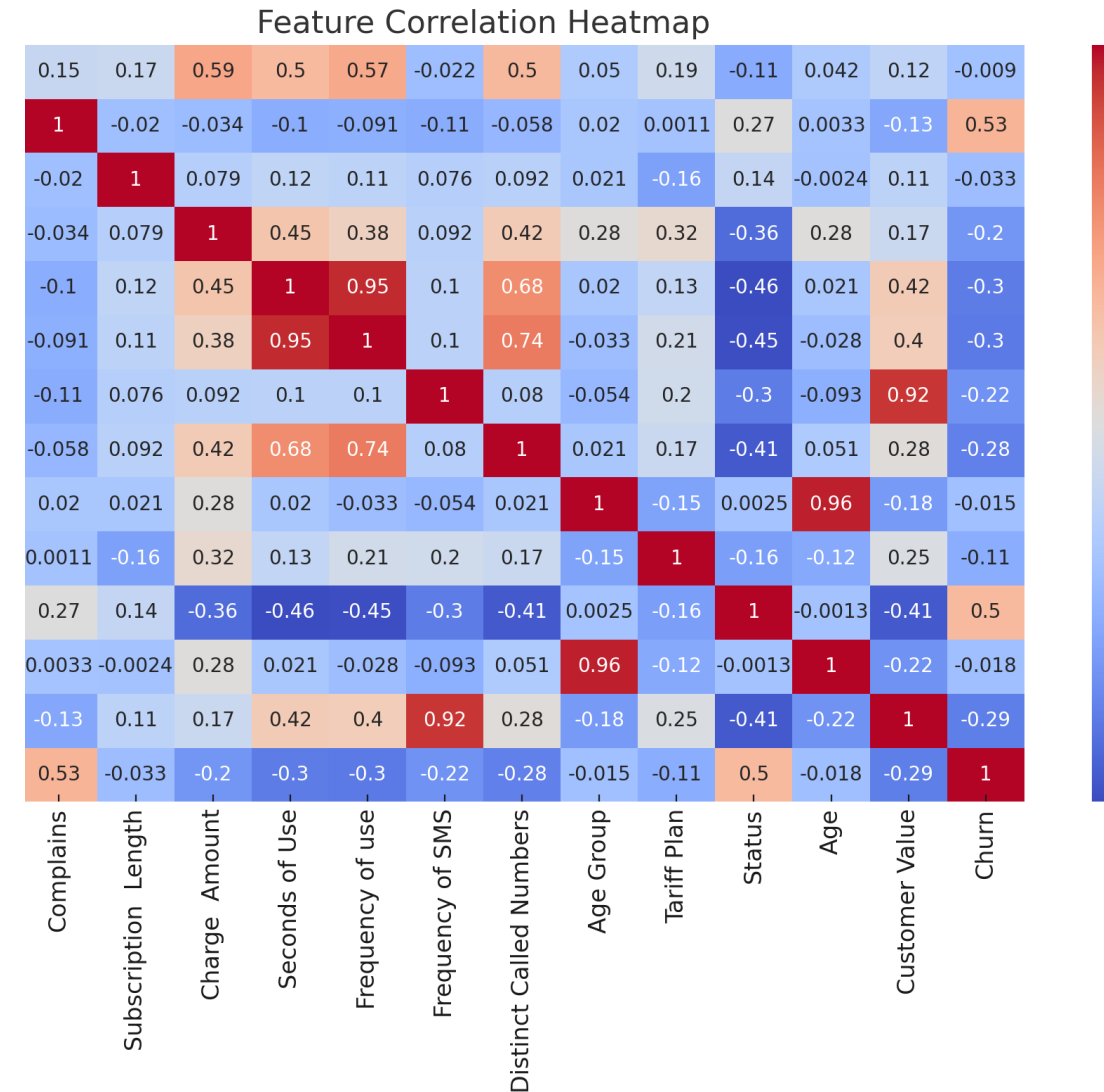
The histogram reveals that customers with shorter subscription lengths are more prone to churn. This suggests that churn risk decreases as customers' tenure with the service increases.

There's a visible trend indicating that long-term customers tend to be more loyal, highlighting the value of nurturing long-term customer relationships for better retention.



Deep Dive into Feature Correlations and Churn

- Certain features, such as Call Failure, Complains, and Subscription Length, show notable correlations with Churn. For instance, Call Failure has a positive correlation, indicating that an increase in call failures is associated with a higher likelihood of churn.
- Subscription Length exhibits a negative correlation with Churn, suggesting that longer subscription durations are typically associated with lower churn rates. This underscores the importance of customer loyalty and long-term engagement in retention strategies.
- The heatmap also reveals inter-correlations between other features, like the positive correlation between Call Failure and Blocked Calls, which might indicate underlying service quality issues affecting customer satisfaction.



Standardizing Data for Model Consistency

Definition of Standard Scaling:

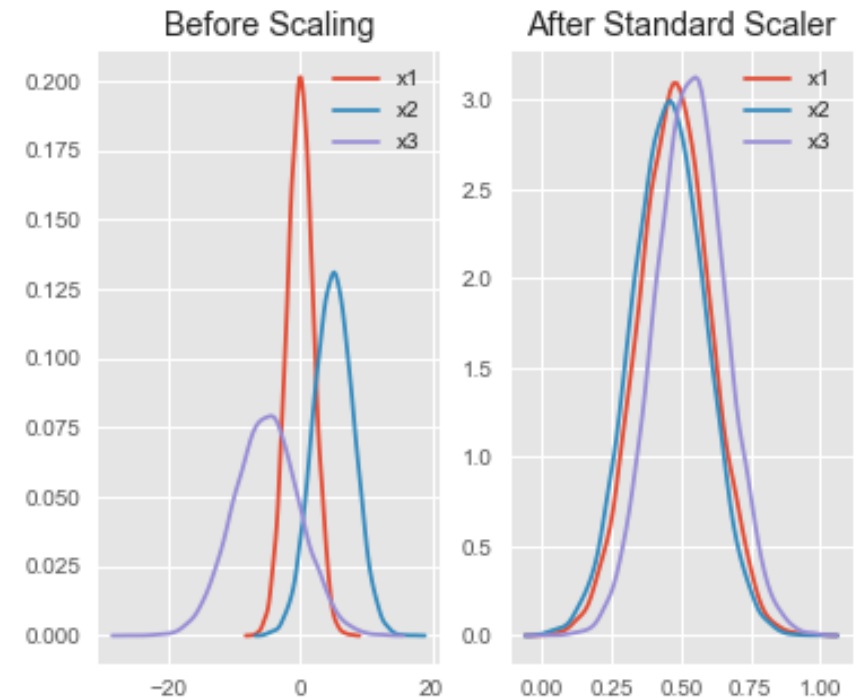
- A preprocessing technique used to normalize the range of independent variables or features of data.
- Ensures each feature contributes equally to model predictions, preventing any single feature from dominating due to its scale.

Benefits of Standard Scaling:

- Improves algorithm convergence: Helps in faster and more stable convergence of gradient descent algorithms.
- Enhances model performance: Particularly crucial for models like SVM and KNN, which are sensitive to the scale of input features.

Implementation Details:

- Each feature was standardized to have a mean of 0 and a standard deviation of 1, ensuring consistency across the dataset.



Stratified Sampling for Reliable Models

Understanding Stratified Sampling:

- Dividing the population into homogeneous subgroups and taking a random sample from each subgroup.
- Ensures that each subgroup is appropriately represented in the sample.

Significance in Train-Test Split:


- Maintains the original distribution of the target variable (churn) in both training and testing sets.
- Crucial for dealing with imbalanced datasets, ensuring that the minority class is adequately represented.



Multi Collinearity Checks

- Variance inflation factor was used to check for multi-collinearity
- Columns with multi-collinearity are:
 - Seconds of Use
 - Frequency of use
 - Frequency of SMS
 - Age Group
 - Age
 - Customer Value

VIF	
Call Failure	2.00349
Complains	0.06192
Subscription Length	1.24054
Charge Amount	2.11582
Seconds of Use	13.2333
Frequency of use	14.0467
Frequency of SMS	5.82229
Distinct Called Numbers	2.78006
Age Group	48.3787
Tariff Plan	0
Status	2.60682
Age	47.6355
Customer Value	10.3552



Models Employed in Churn Prediction

Decision Tree Classifier

K-Nearest Neighbors (KNN)

Logistic Regression

Random Forest Classifier

Support Vector Machine (SVM)

XGBoost Classifier

Decision Tree Classifier

Train-Test Split	Test Accuracy
90:10	97.55%
85:15	97.49%
80:20	97.27%
75:25	96.61%
70:30	97.74%
65:35	96.88%
60:40	97.08%
55:45	96.32%
50:50	95.82%

K Nearest Neighbors Classifier

Train-Test Split	Test Accuracy
90:10	95.10%
85:15	94.98%
80:20	94.54%
75:25	93.52%
70:30	93.97%
65:35	94.03%
60:40	93.17%
55:45	92.85%
50:50	92.35%

Logistic Regression

Train-Test Split	Test Accuracy
90:10	80.79%
85:15	82.31%
80:20	81.54%
75:25	81.78%
70:30	81.98%
65:35	82.36%
60:40	82.49%
55:45	82.13%
50:50	82.22%

Random Forest Classifier

Train-Test Split	Test Accuracy
90:10	98.31%
85:15	98.62%
80:20	98.40%
75:25	98.12%
70:30	98.18%
65:35	98.01%
60:40	97.83%
55:45	97.45%
50:50	97.48%

Support Vector Classifier

Train-Test Split	Test Accuracy
90:10	79.10%
85:15	80.93%
80:20	79.38%
75:25	79.07%
70:30	79.41%
65:35	79.45%
60:40	79.24%
55:45	79.33%
50:50	79.47%

XGBoost Classifier

Train-Test Split	Test Accuracy
90:10	97.93%
85:15	98.24%
80:20	98.12%
75:25	98.04%
70:30	97.74%
65:35	97.63%
60:40	96.99%
55:45	97.70%
50:50	96.84%

Comparative Analysis of Model Accuracy

Model	Test Accuracy	Split Ratio
Random Forest	98.62%	85:15
Gradient Boosting (XGBoost)	98.24%	85:15
Decision Tree	97.74%	70:30
K-Nearest Neighbors	95.1%	90:10
Logistic Regression	82.49%	60:40
Support Vector Classifier	80.93%	85:15

Metrics for Assessing Churn Prediction Models

- Accuracy: Measures the overall correctness of the model.
- Precision, Recall, F1-Score: Evaluates the model's performance in terms of its ability to correctly predict positive (churned) cases.
- ROC-AUC Score: Reflects the model's ability to distinguish between the classes.

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{All Samples}}$$

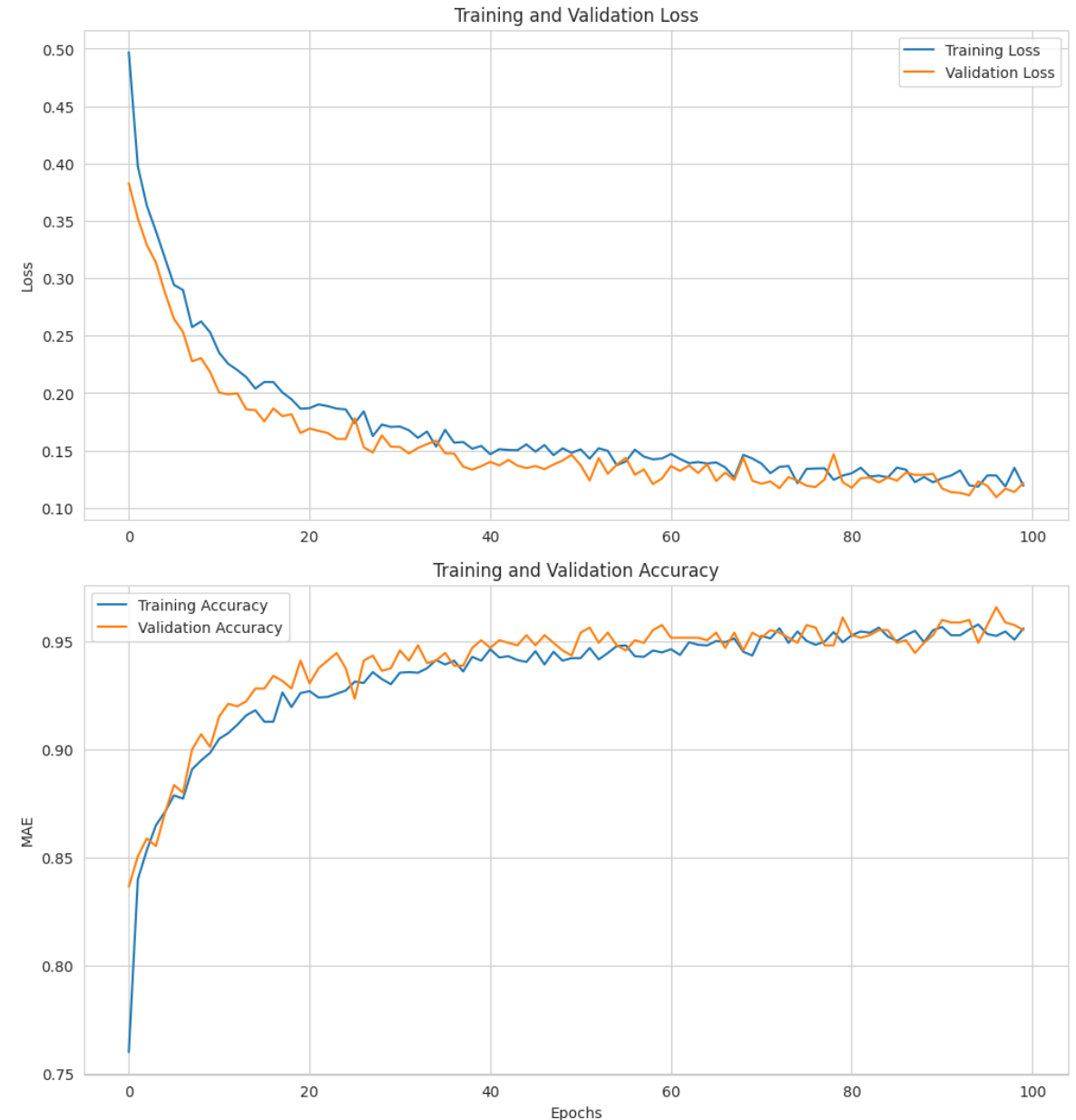
$$\text{Precision} = \frac{\text{True Positives}}{\text{Total Predicted Positives}}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

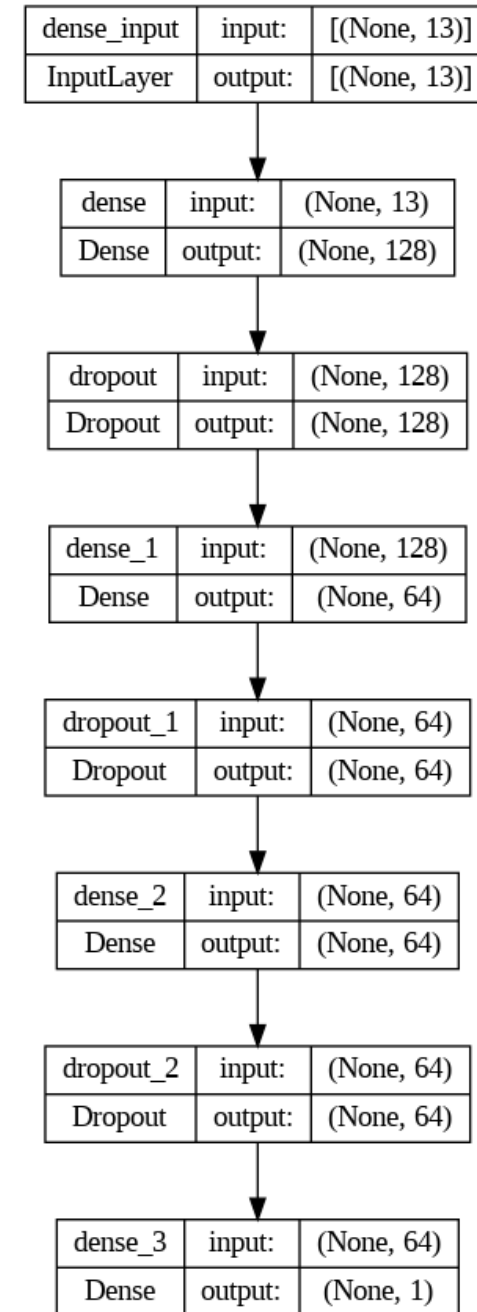
$$\text{F1 Score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}$$

Neural Network Classifier

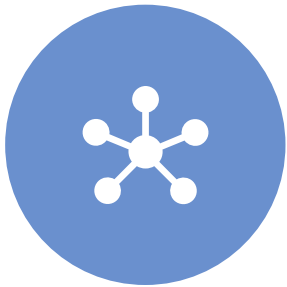
- A deep neural network was trained over 100 epochs to predict Customer Churn
- Parameters
 - Optimizer: Adam
 - Loss function: Binary cross-entropy
 - Trainable Parameters: 14273
- **Test Accuracy: 95.76%**



Model Architecture



The Power of Ensemble Models and Data Preprocessing



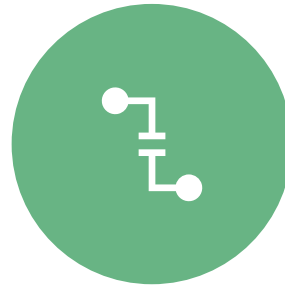
Ensemble Models: Both Random Forest and XGBoost are ensemble models that combine multiple decision trees for predictions, often resulting in higher accuracy.



Data Preprocessing: Balancing addressed class imbalance issues through oversampling, Scaling normalized feature scales improving model stability, Stratified Sampling preserved the original churn distribution



Impact: Comprehensive preprocessing enhanced pattern capture, leading to high accuracies.



No Hyperparameter Tuning: High accuracies achieved without extensive hyperparameter tuning, showcasing the effectiveness of ensemble models and robust preprocessing in churn prediction.

Key Findings



Customer Retention:

Key factors impacting churn identified: Call Failures. Subscription length, complaints
Improvements in service quality could significantly reduce churn



Model Performance:

Random Forest turned out to be the model best suited to this application with an accuracy of 98.62%

Thank you





Appendix



Performance metrics for Decision Tree

index	precision	recall	f1-Score	support
Non-Churn (0)	0.997	0.959	0.978	399
Churn (1)	0.961	0.997	0.979	398

Performance metrics for Random Forest

index	Precision	recall	f1-Score	support
Non-Churn (0)	1	0.972	0.986	399
Churn (1)	0.973	1	0.986	398

Performance metrics for XGBoost

index	precision	recall	f1-Score	support
Non-Churn (0)	0.996	0.966	0.980	531
Churn (1)	0.967	0.996	0.981	531

Link to Collab
Notebook