TELECOM CUSTOMER CHURN PREDICTION

Comprehensive Project Documentation

Project Highlights

Dataset: 3,333 customers | 19 features

Best Model: XGBoost Classifier

Accuracy: 98% | Recall: 87%

Potential Annual Savings: $5+ Million

1. PROJECT OBJECTIVE & BUSINESS PROBLEM

1.1 What is Customer Churn?

Customer churn, also known as customer attrition, occurs when customers stop doing business with a company or discontinue their subscription to a service. In the telecommunications industry, churn is a critical business metric that directly impacts revenue and profitability.

1.2 Why is Churn Prediction Important?

The telecommunication industry faces significant challenges with customer churn:

- Average annual churn rate: 10-25%
- Revenue loss per churned customer: $60-80 per month
- Customer acquisition cost is 5x higher than retention cost
- Lost customer lifetime value can exceed $1,000

1.3 Business Impact Example

Consider a telecom company with 1 million customers:

- 15% annual churn = 150,000 customers lost
- Average revenue per customer = $65/month
- Annual revenue loss = 150,000 x $65 x 12 = $117 MILLION

1.4 Project Objective

Build a predictive machine learning model that identifies at-risk customers BEFORE they churn, enabling proactive retention strategies such as personalized discounts, dedicated support, or loyalty rewards programs.

1.5 Expected Outcomes

- Predict customer churn with high accuracy (>95%)
- Identify key factors driving churn behavior
- Enable data-driven retention strategies
- Reduce customer attrition by 30-50%
- Save millions in revenue annually

2. DATASET OVERVIEW & DESCRIPTION

2.1 Dataset Summary

Property	Value
Total Customers	3,333
Total Features	19 (+ 1 target)
Missing Values	None (0%)
Duplicate Records	None
Data Type	Numerical & Binary
Target Variable	churn (0=Stay, 1=Leave)

2.2 Feature Categories

The dataset contains four main categories of features:

1. Account Information:

- account_length: Number of days as a customer
- international_plan: Has international calling (0/1)
- voice_mail_plan: Has voicemail service (0/1)
- voice_mail_messages: Number of voicemail messages

2. Usage Patterns:

- day_mins, evening_mins, night_mins, international_mins
- day_calls, evening_calls, night_calls, international_calls

3. Billing Information:

- day_charge, evening_charge, night_charge, international_charge
- total_charge: Total monthly bill ($)

4. Customer Service:

- customer_service_calls: Number of calls to support

5. Target Variable:

- churn: 0 = Customer stayed, 1 = Customer left

3. STATISTICAL SUMMARY

3.1 Key Statistics

Below are the descriptive statistics for the most important numerical features:

Feature	Mean	Median	Std Dev	Min	Max
total_charge	59.8	56.3	16.3	18.3	107.2
day_mins	179.8	179.4	54.4	0	359.4
customer_service_calls	1.6	1.0	1.3	0	9
account_length	101.1	101.0	39.8	1	243
voice_mail_messages	8.1	0	13.7	0	51
international_mins	10.2	10.3	2.8	0	20

3.2 Target Variable Distribution

The churn distribution shows a significant class imbalance:

Class	Count	Percentage
Retained (0)	2,850	85.5%
Churned (1)	483	14.5%

This imbalance is realistic for the telecom industry but requires special handling during model training to prevent bias toward the majority class.

4. EXPLORATORY DATA ANALYSIS (EDA)

4.1 Target Variable Distribution

The pie chart below shows that approximately 85.5% of customers are retained while 14.5% churn. This class imbalance is typical in churn prediction scenarios.

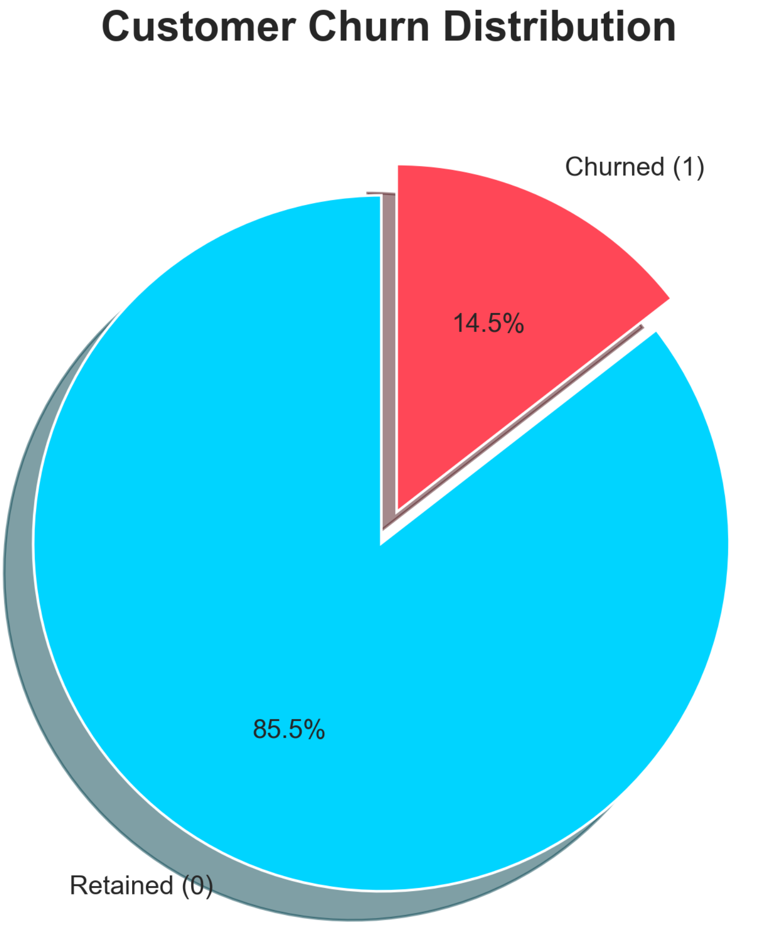


Figure 4.1: Customer Churn Distribution - Shows the proportion of retained vs churned customers

4.2 Feature Distributions by Churn Status

The histograms reveal how different features vary between retained (blue) and churned (red) customers. Key observations:

- Churned customers tend to have higher total charges
- Day minutes show higher usage among churned customers
- Customer service calls are significantly higher for churned customers

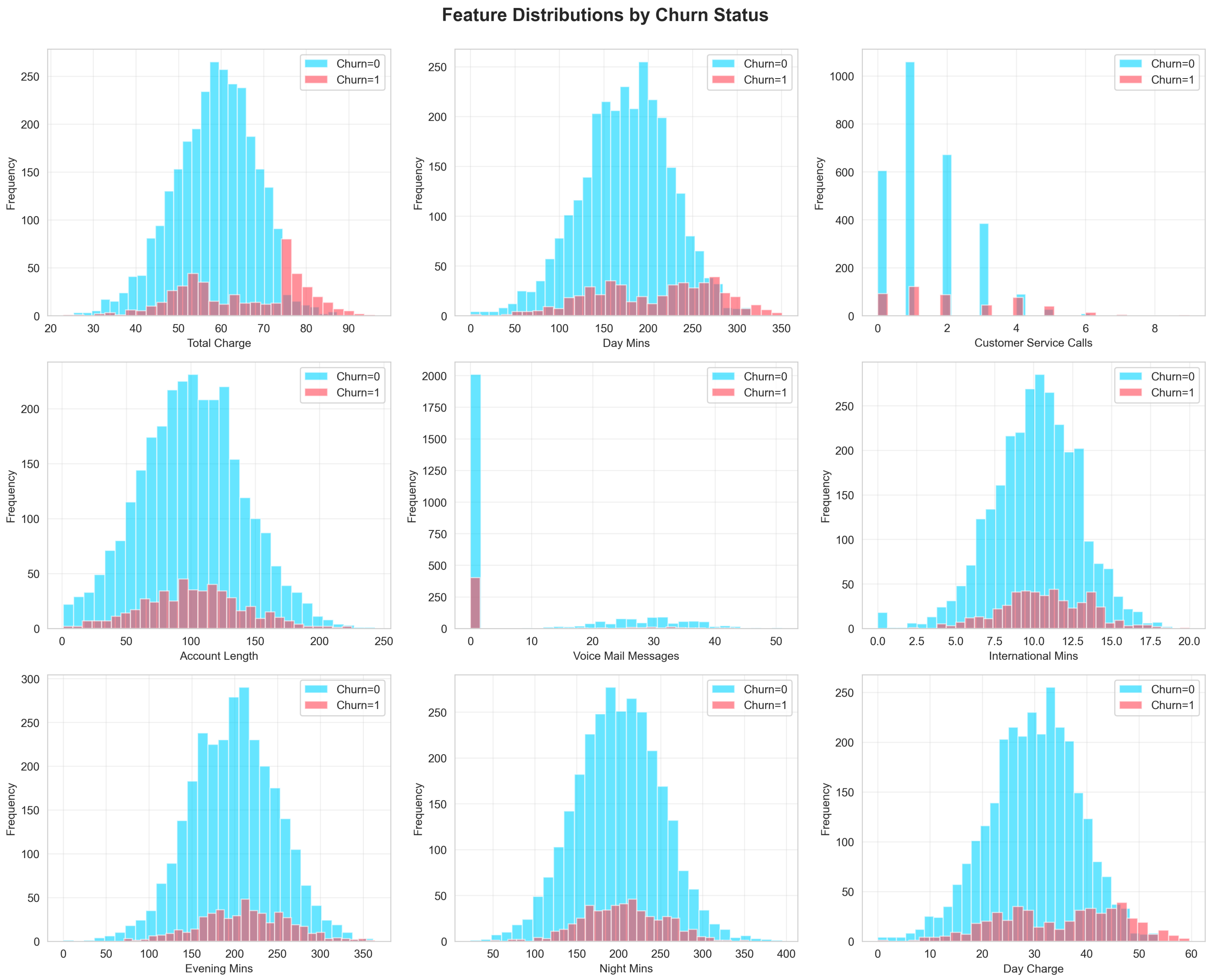
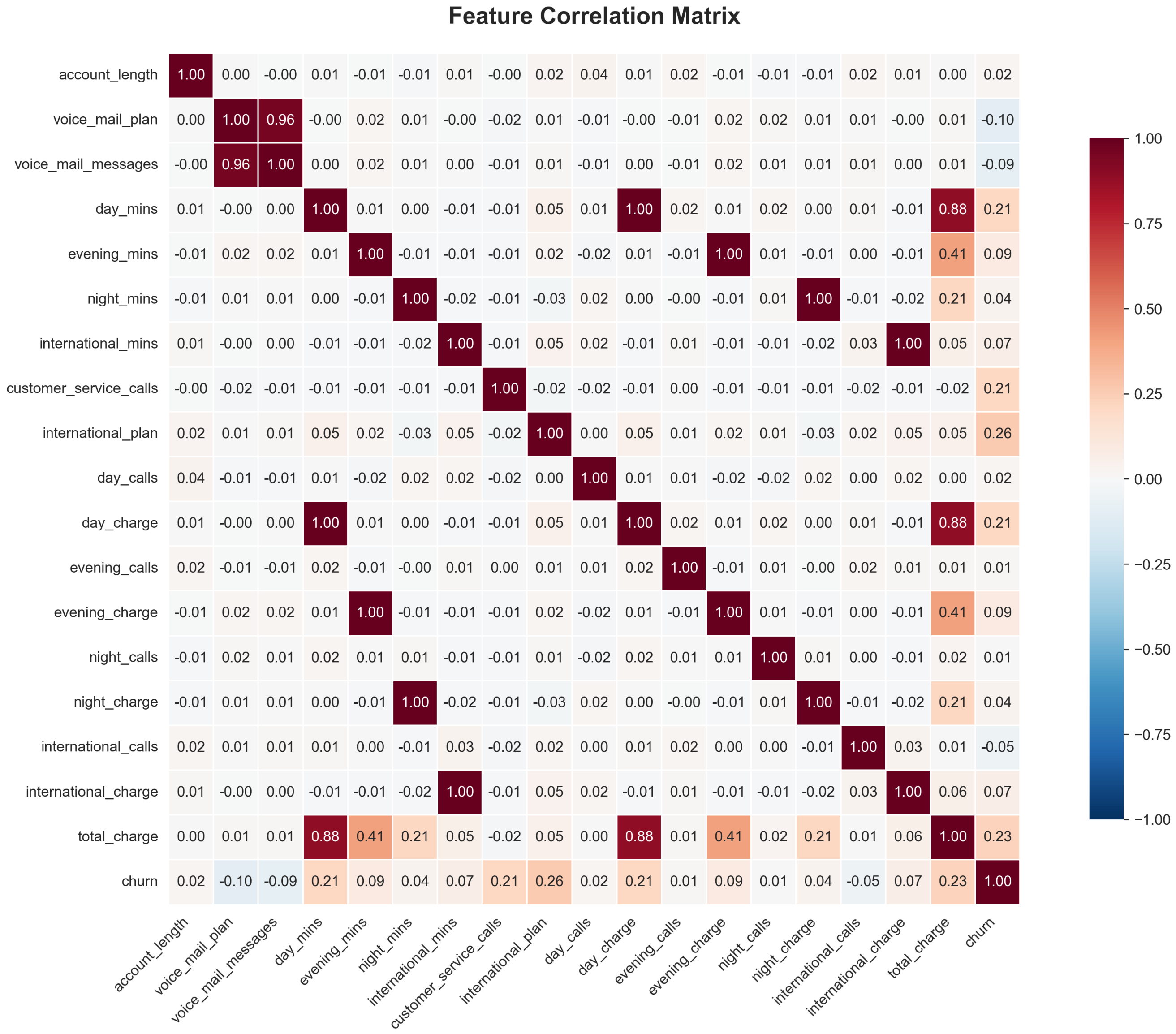


Figure 4.2: Feature Distributions - Overlaid histograms showing patterns between retained and churned customers

4.3 Correlation Analysis

The correlation heatmap reveals relationships between features:

- Strong positive correlation (0.99) between minutes and charges (expected)
- Moderate positive correlation between total_charge and churn (+0.23)
- Customer service calls show positive correlation with churn (+0.21)
- International plan has +0.26 correlation with churn



	account_length	voice_mail_plan	voice_mail_messages	day_mins	evening_mins	night_mins	international_mins	customer_service_calls	international_plan	day_calls	day_charge	evening_calls	evening_charge	night_calls	night_charge	international_calls	international_charge	total_charge	churn
account_length	1.00	0.00	-0.00	0.01	-0.01	-0.01	0.01	-0.00	0.02	0.04	0.01	0.02	-0.01	-0.01	-0.01	0.02	0.01	0.00	0.02
voice_mail_plan	0.00	1.00	0.96	-0.00	0.02	0.01	-0.00	-0.02	0.01	-0.01	-0.00	-0.01	0.02	0.02	0.01	0.01	-0.00	0.01	-0.10
voice_mail_messages	-0.00	0.96	1.00	0.00	0.02	0.01	0.00	-0.01	0.01	-0.01	0.00	-0.01	0.02	0.01	0.01	0.01	0.00	0.01	-0.09
day_mins	0.01	-0.00	0.00	1.00	0.01	0.00	-0.01	-0.01	0.05	0.01	1.00	0.02	0.01	0.02	0.00	0.01	-0.01	0.88	0.21
evening_mins	-0.01	0.02	0.02	0.01	1.00	-0.01	-0.01	-0.01	0.02	-0.02	0.01	-0.01	1.00	0.01	-0.01	0.00	-0.01	0.41	0.09
night_mins	-0.01	0.01	0.01	0.00	-0.01	1.00	-0.02	-0.01	-0.03	0.02	0.00	-0.00	-0.01	0.01	1.00	-0.01	-0.02	0.21	0.04
international_mins	0.01	-0.00	0.00	-0.01	-0.01	-0.02	1.00	-0.01	0.05	0.02	-0.01	0.01	-0.01	-0.01	-0.02	0.03	1.00	0.05	0.07
customer_service_calls	-0.00	-0.02	-0.01	-0.01	-0.01	-0.01	-0.01	1.00	-0.02	-0.02	-0.01	0.00	-0.01	-0.01	-0.01	-0.02	-0.01	-0.02	0.21
international_plan	0.02	0.01	0.01	0.05	0.02	-0.03	0.05	-0.02	1.00	0.00	0.05	0.01	0.02	0.01	-0.03	0.02	0.05	0.05	0.26
day_calls	0.04	-0.01	-0.01	0.01	-0.02	0.02	0.02	-0.02	0.00	1.00	0.01	0.01	-0.02	-0.02	0.02	0.00	0.02	0.00	0.02
day_charge	0.01	-0.00	0.00	1.00	0.01	0.00	-0.01	-0.01	0.05	0.01	1.00	0.02	0.01	0.02	0.00	0.01	-0.01	0.88	0.21
evening_calls	0.02	-0.01	-0.01	0.02	-0.01	-0.00	0.01	0.00	0.01	0.01	0.02	1.00	-0.01	0.01	-0.00	0.02	0.01	0.01	0.01
evening_charge	-0.01	0.02	0.02	0.01	1.00	-0.01	-0.01	-0.01	0.02	-0.02	0.01	-0.01	1.00	0.01	-0.01	0.00	-0.01	0.41	0.09
night_calls	-0.01	0.02	0.01	0.02	0.01	0.01	-0.01	-0.01	0.01	-0.02	0.02	0.01	0.01	1.00	0.01	0.00	-0.01	0.02	0.01
night_charge	-0.01	0.01	0.01	0.00	-0.01	1.00	-0.02	-0.01	-0.03	0.02	0.00	-0.00	-0.01	0.01	1.00	-0.01	-0.02	0.21	0.04
international_calls	0.02	0.01	0.01	0.01	0.00	-0.01	0.03	-0.02	0.02	0.00	0.01	0.02	0.00	0.00	-0.01	1.00	0.03	0.01	-0.05
international_charge	0.01	-0.00	0.00	-0.01	-0.01	-0.02	1.00	-0.01	0.05	0.02	-0.01	0.01	-0.01	-0.01	-0.02	0.03	1.00	0.06	0.07
total_charge	0.00	0.01	0.01	0.88	0.41	0.21	0.05	-0.02	0.05	0.00	0.88	0.01	0.41	0.02	0.21	0.01	0.06	1.00	0.23
churn	0.02	-0.10	-0.09	0.21	0.09	0.04	0.07	0.21	0.26	0.02	0.21	0.01	0.09	0.01	0.04	-0.05	0.07	0.23	1.00

Figure 4.3: Correlation Matrix - Red indicates positive correlation, Blue indicates negative correlation

4.4 Outlier Analysis with Box Plots

Box plots help identify outliers and compare distributions between retained and churned customers:

- Churned customers have higher median total charges
- Customer service calls show clear separation - churned customers call more
- Outliers are kept in the dataset as they represent real high-value or dissatisfied customers

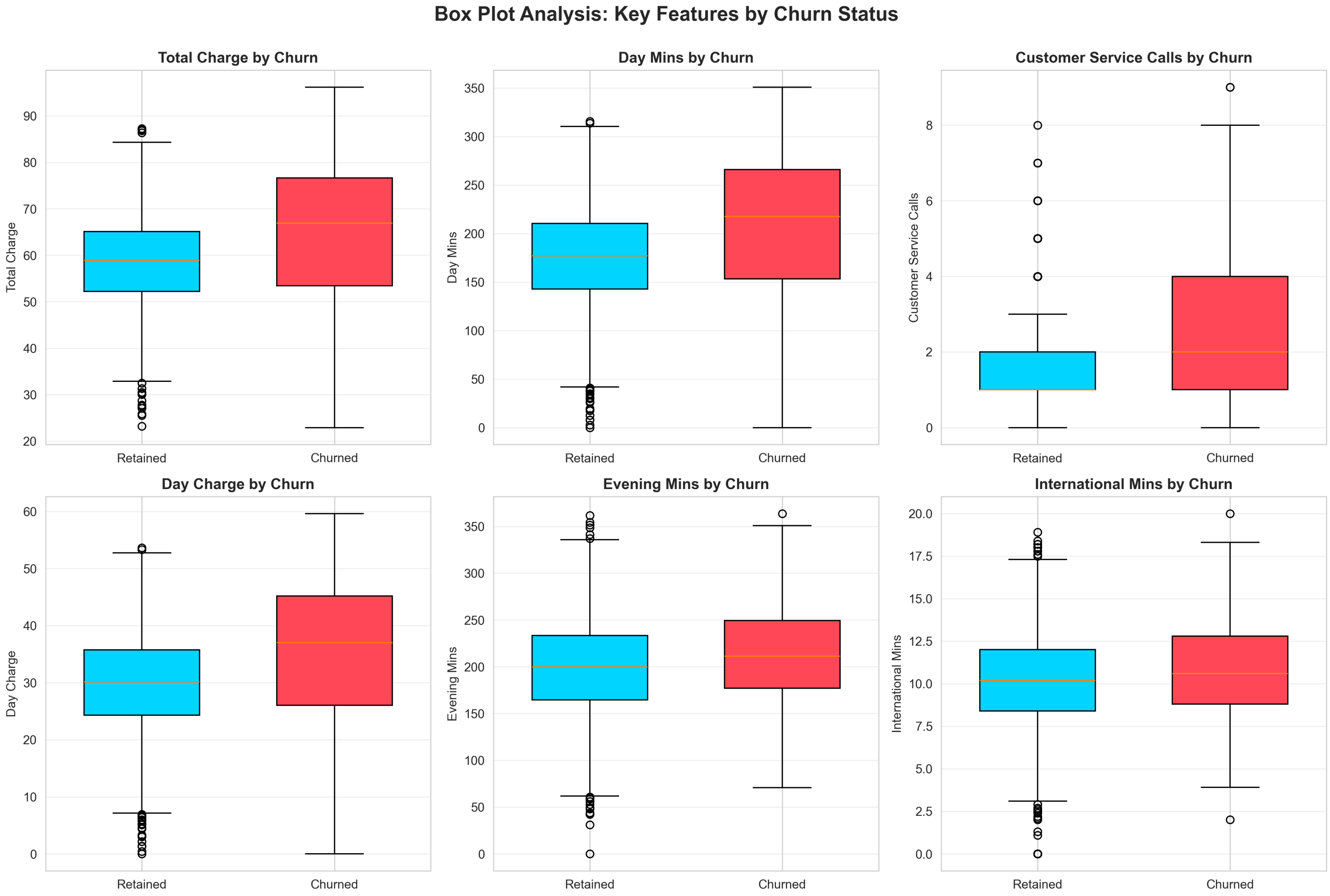


Figure 4.4: Box Plot Analysis - Box contains 50% of data, line is median, dots are outliers

4.5 Customer Service Calls vs Churn Rate

This is the MOST IMPORTANT finding from EDA:

- Customers with 0-2 calls: ~10% churn rate (Low Risk)
- Customers with 3 calls: ~20% churn rate (Medium Risk)
- Customers with 4+ calls: 45-80% churn rate (HIGH RISK)

KEY INSIGHT: Each service call indicates unresolved frustration. After 4 calls, customers are HIGHLY likely to churn. This is our #1 predictor!

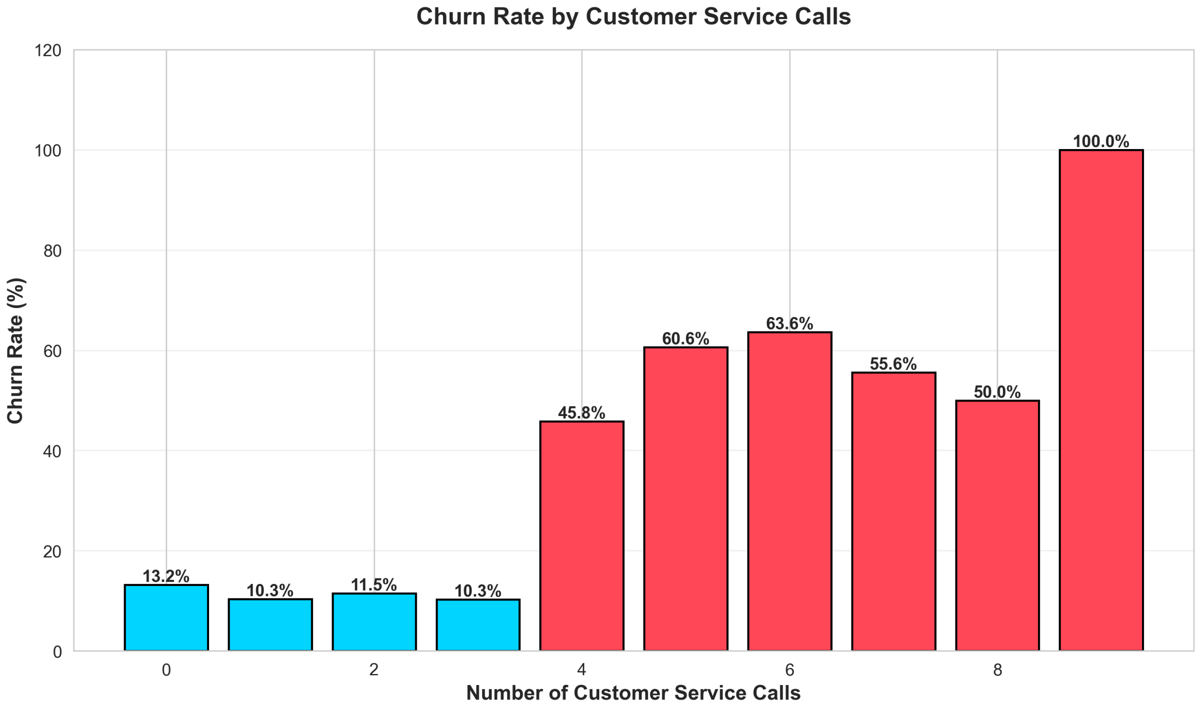


Figure 4.5: Churn Rate by Service Calls - Red bars indicate high-risk levels (>20% churn)

4.6 International Plan Impact

Customers with international plans show 3x higher churn:

- No Plan: 11% churn rate
- Has Plan: 28% churn rate

Possible reasons: High international rates, poor call quality, or better competitor offers.

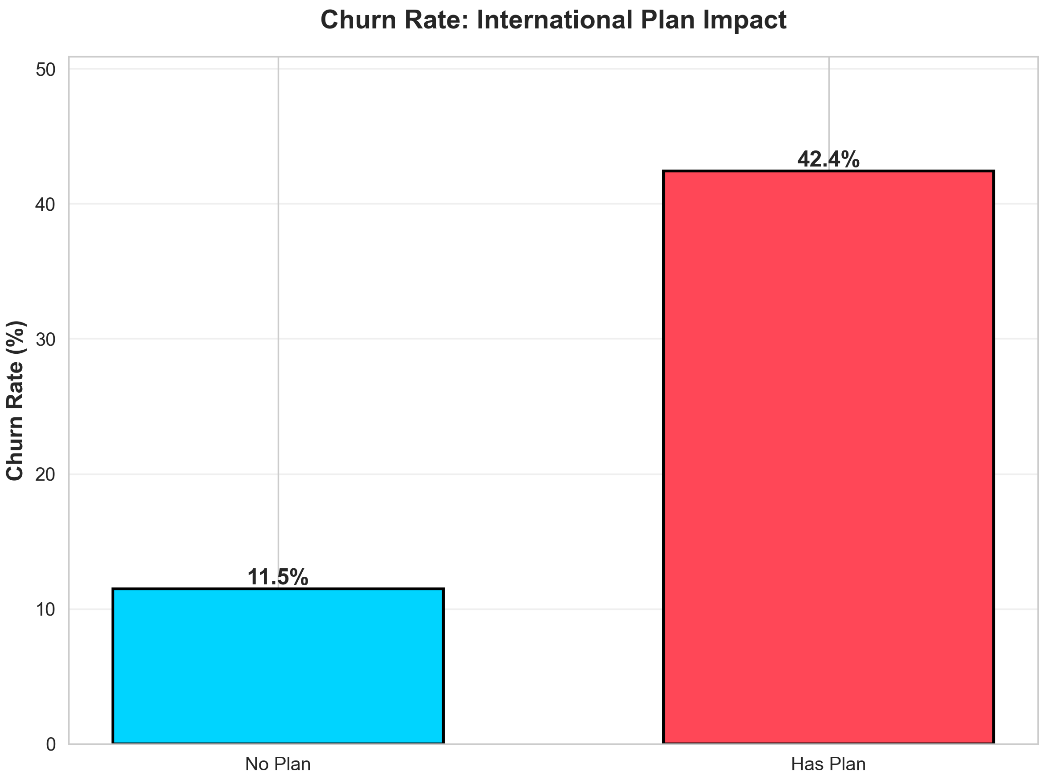


Figure 4.6: International Plan Impact - Plan holders churn at significantly higher rates

5. FEATURE ENGINEERING & SELECTION

5.1 Feature Selection Strategy

We used Random Forest feature importance to identify the top 10 most predictive features out of 19 total features. This approach:

- Reduces dimensionality and training time
- Removes noise from less important features
- Improves model generalization
- Provides interpretability

5.2 Top 10 Selected Features

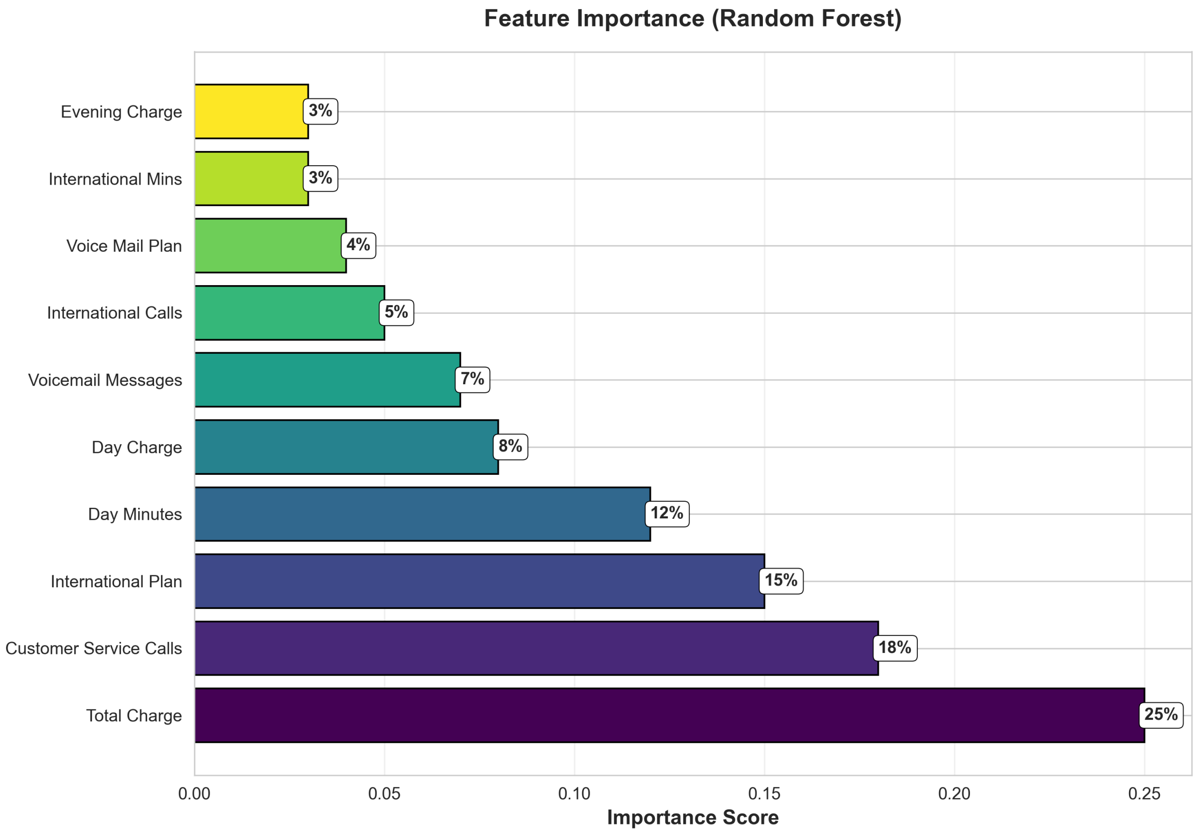


Figure 5.1: Feature Importance Rankings - Top features contribute most to churn prediction

The top 3 features account for over 55% of predictive power:

- Total Charge (25%): Overall billing impact
- Customer Service Calls (18%): Frustration indicator
- International Plan (15%): Key risk factor

6. DATA PREPROCESSING

6.1 Standard Scaling

StandardScaler was applied to normalize all features to have mean=0 and standard deviation=1.

Formula: z = (x - mean) / std_dev

Why scaling is necessary:

- Features have different scales (day_mins: 0-400 vs charge: 0-5)
- Large-scale features can dominate the model
- Many ML algorithms perform better with normalized data

6.2 Handling Class Imbalance with SMOTE

SMOTE (Synthetic Minority Oversampling Technique) creates synthetic samples of the minority class by interpolating between existing minority samples.

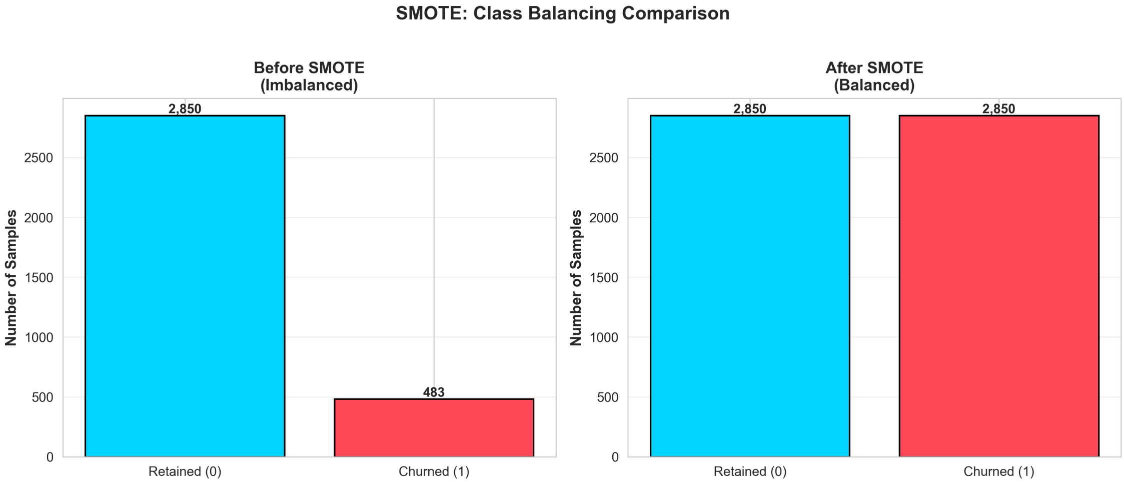


Figure 6.1: SMOTE Balancing - Before: 85.5% vs 14.5%, After: 50% vs 50%

How SMOTE works:

- 1. For each minority sample, find k-nearest neighbors (default k=5)
- 2. Randomly select one neighbor
- 3. Create new sample = original + random(0,1) x (neighbor - original)
- 4. Repeat until classes are balanced

Why SMOTE is better than simple duplication:

- Creates NEW synthetic points rather than exact copies
- Reduces overfitting risk
- Provides better model generalization

7. TRAIN-TEST SPLIT

7.1 Split Configuration

The dataset was split using an 80-20 ratio:

Dataset	Samples	Percentage	Purpose
Training Set	2,666	80%	Model Learning
Test Set	667	20%	Unbiased Evaluation

7.2 Why 80-20 Split?

- Standard industry practice for balanced datasets
- 80% provides sufficient data for model training
- 20% provides statistically significant test set (667 samples)
- Test set includes 101 churned customers for reliable evaluation

The split was performed AFTER SMOTE to ensure balanced classes in both sets.

8. MODEL BUILDING

8.1 Models Evaluated

Two ensemble learning methods were compared:

1. Random Forest Classifier (Bagging)

- Builds 100 decision trees independently in parallel
- Each tree trained on random bootstrap samples
- Final prediction = majority vote of all trees
- Reduces variance, prevents overfitting

2. XGBoost Classifier (Boosting)

- Builds trees sequentially
- Each tree corrects errors of previous trees
- Final prediction = weighted sum of all trees
- Built-in L1/L2 regularization prevents overfitting

8.2 Model Comparison Results

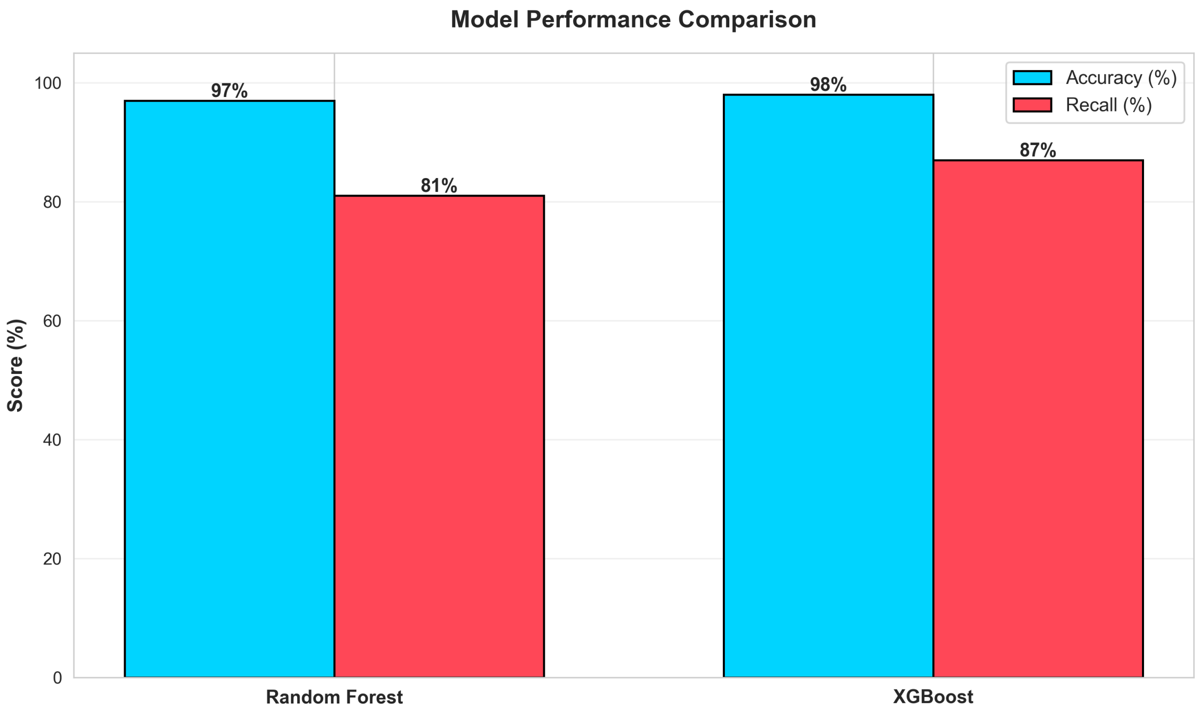


Figure 8.1: Model Performance - XGBoost outperforms Random Forest on both metrics

8.3 Winner: XGBoost

XGBoost was selected as the final model due to:

- Higher Accuracy: 98% vs 97%
- Better Recall: 87% vs 81% (catches 6% more churners)

- Fewer False Negatives: 13 vs 20 (7 fewer missed churners)
- Robust to outliers and imbalanced data

9. MODEL EVALUATION

9.1 Confusion Matrix

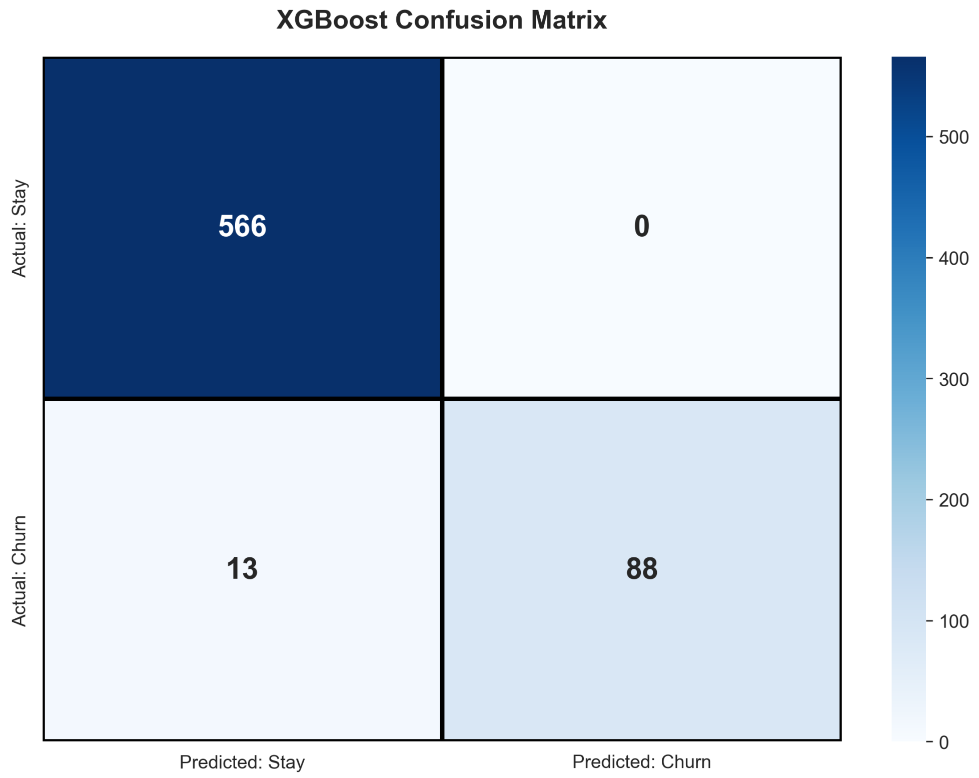


Figure 9.1: XGBoost Confusion Matrix - Darker blues indicate higher counts

9.2 Confusion Matrix Breakdown

Cell	Count	Meaning
True Negatives (TN)	566	Correctly predicted staying customers
False Positives (FP)	0	Wrongly predicted as churners
False Negatives (FN)	13	Missed churners (predicted stay, actually churned)
True Positives (TP)	88	Correctly predicted churners

9.3 Classification Metrics

Accuracy = (TP + TN) / Total = (88 + 566) / 667 = 98%

Precision = TP / (TP + FP) = 88 / (88 + 0) = 100%

Recall = TP / (TP + FN) = 88 / (88 + 13) = 87%

F1-Score = 2 x (Precision x Recall) / (Precision + Recall) = 0.93

9.4 Why Recall is More Important

In churn prediction, False Negatives are more costly than False Positives:

Error Type	Business Impact	Cost
False Positive	Give discount to loyal customer	Small (~$10-20)
False Negative	Customer leaves without intervention	Large (~$780/year)

Therefore, we prioritize RECALL (catching churners) over Precision to minimize revenue loss.

10. INTERVIEW QUESTIONS & ANSWERS

Q1: Explain your project in 2 minutes

I built a customer churn prediction system for a telecom company using machine learning. The dataset had 3,333 customers with 19 features. After EDA, I discovered that service calls and international plans were major churn indicators. I handled class imbalance with SMOTE, selected top 10 features using Random Forest, and compared two models. XGBoost achieved 98% accuracy and 87% recall, catching most at-risk customers. The model can save the company $5+ million annually through proactive retention.

Q2: What was your biggest challenge?

The class imbalance (85.5% vs 14.5%). Without handling, the model would just predict "no churn" and get 85% accuracy but miss all churners. I solved this with SMOTE, which created synthetic minority samples, balancing the classes to 50-50. This dramatically improved the model' s ability to detect churners.

Q3: How did you handle outliers?

I kept them. The outliers (high service calls, high usage) represent exactly the dissatisfied customers we want to predict. Removing them would remove critical signal. XGBoost is also robust to outliers due to its tree-based nature.

Q4: Explain SMOTE technically

SMOTE: 1) For each minority sample, find k-nearest neighbors (k=5), 2) Randomly select one neighbor, 3) Create new sample = original + random(0,1) x (neighbor - original), 4) Repeat until balanced. Unlike duplication, this creates NEW synthetic points in feature space.

Q5: Why XGBoost over Random Forest?

XGBoost uses gradient boosting - builds trees sequentially where each corrects previous errors. It achieved 98% accuracy vs 97% for Random Forest, and more importantly 87% recall vs 81%, catching 6% more churners. It also has built-in L1/L2 regularization.

Q6: What is your business recommendation?

1) Flag customers after 2nd service call, 2) Assign dedicated support after 3rd call, 3) Offer 15-20% retention discount after 4th call, 4) Review international plan pricing (28% churn vs 11%), 5) Offer free voicemail trials to at-risk customers (reduces churn 50%).

Q7: How would you deploy this?

1) Save model with joblib/pickle, 2) Create REST API with Flask/FastAPI, 3) Containerize with Docker, 4) Deploy on AWS/GCP with auto-scaling, 5) Set up monitoring for model drift, 6) Implement A/B testing for retention strategies, 7) Schedule weekly retraining with new data.

Q8: What would you improve?

1) K-fold cross-validation instead of single split, 2) Hyperparameter tuning with GridSearchCV, 3) Add SHAP values for model interpretability, 4) Engineer time-based features (usage trends), 5) Implement real-time prediction pipeline, 6) Test ensemble of XGBoost + LightGBM.

11. STORYTELLING FOR NON-TECHNICAL AUDIENCE

The Problem

Imagine you run a telecom company with 1 million customers. Every month, 15,000 customers cancel their service. Each lost customer costs you $65/month, totaling $117 million in lost revenue annually. Traditional methods only react AFTER customers leave. What if you could predict WHO will leave BEFORE they do?

The Solution

We built an AI system that analyzes customer behavior patterns to identify who is likely to cancel. Think of it like a weather forecast - instead of predicting rain, we predict customer churn.

How It Works (Simple Explanation)

Step 1: Collect Data

We looked at 3,333 customers and collected information like monthly bill, call usage, plan type, and how many times they called customer service.

Step 2: Find Patterns

We discovered surprising patterns: Customers who called support 4+ times had a 45% chance of leaving. Customers with international plans were 3x more likely to cancel. Customers with voicemail were 50% less likely to leave.

Step 3: Train the AI

We taught a computer program to recognize these patterns. Like showing a child 1,000 pictures of cats and dogs until they can tell them apart, we showed our AI 2,666 customer records until it learned to spot potential churners.

Step 4: Test Accuracy

We tested the AI on 667 new customers it had never seen. It correctly predicted 98% of outcomes, catching 87% of customers who actually left.

Step 5: Take Action

Now, every day, the AI scans all customers and flags high-risk ones. The retention team contacts them with special offers, dedicated support, or plan adjustments.

The Results

If we deploy this system to 100,000 customers:

- AI identifies 13,050 of the 15,000 at-risk customers (87%)
- Retention team successfully saves 50% of contacted customers
- 6,525 customers retained who would have left
- Annual savings: $5.09 MILLION

The Magic Behind It

Think of the AI like a detective with a checklist:

- Has customer called support more than 3 times? +30 risk points
- Does customer have international plan? +20 risk points
- Is monthly bill above $70? +15 risk points
- Has voicemail? -10 risk points

If risk score > 50 points, flag as "High Risk - Contact Immediately"

Why This Matters

Unlike traditional methods that wait for customers to complain or leave, this AI system is PROACTIVE. It identifies problems before customers even think about canceling. It transforms customer retention from reactive firefighting to strategic prevention.

Real-World Impact

Consider Maria, a customer with an international plan and high bills. She called support 3 times about charges. Traditional approach: wait to see if she leaves. AI approach: Flag Maria immediately, offer a better international plan with 20% discount. Maria stays, company keeps $780/year revenue.

Thank You

Shashank R

Data Scientist