

CUSTOMER CHURN ANALYTICS

Telecommunication Industry, Mobile
Network Services

Composed by: Nguyễn Đức Minh



What is Customer Churn Rate?

Definition

The rate at which customers stop doing business (using product or service) with a company

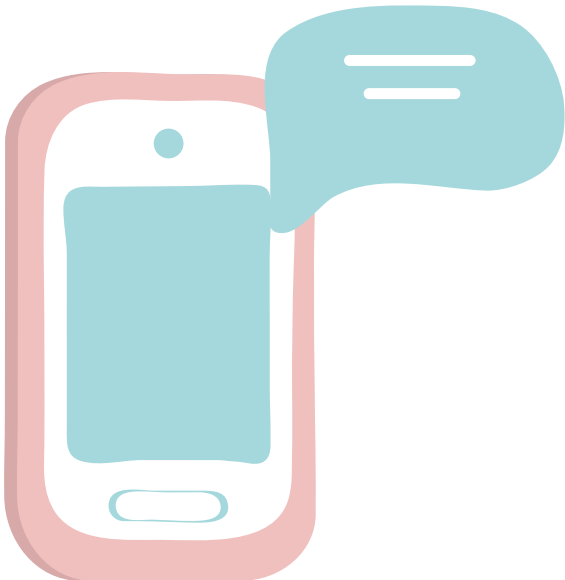
Type of Churn

#1 Voluntary Churn

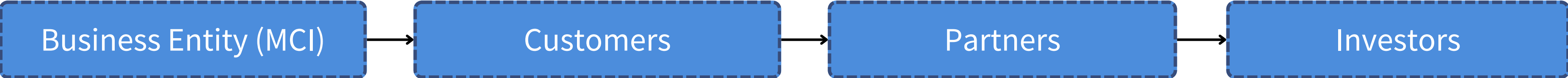
Choose to leave (**dissatisfaction, better offers**)
> **Focus of the Project**

#2 Involuntary Churn

Due to factors beyond customer control
(**payment failures**)



Importance of Churn Rate for Stakeholders



- **Revenue:** 10% cheaper to retain customers than acquire new ones & increase Customer Lifetime Value (CLV)
- **Competitive Adv.:** Reinforce customer **satisfaction** & **market position**, enhance **brand reputation**
- **Operational Efficiency:** Help optimizes **resource planning** for **product development**

- **Enhanced Experience:** Improved **personalized service** & **offerings**
- **Better Value:** Competitive **pricing** and **added benefits** from retention efforts

- **Business Stability:** Encourages **long-term collaborations**
- **Shared growth opportunities**

- **Financial Health Indicator:** Help define underlying **issues** & **probability** and **revenue projection**
- **Harmonious Management:** Reflects competent leadership, suggesting **sustainable growth** if favorable

Overview of Customer Profiles

CHURNED CUSTOMERS

NON-CHURNED CUSTOMERS

Demographics

More likely in states like NJ, TX, AR, MD, ME, MI, MS, PA, SC (esp. NJ & TX) with **longer account length**

- **Distributed** across various states, no centralized
- Varied account lengths, generally satisfied over time

Service Plans

- Majority use the **International Plan** (~ 79.5%)
- **Lower engagement** with **Voice Mail Plan**

- Not likely use the **International Plan**
- **Higher engagement** with **Voice Mail Plan**

Usage Patterns

- **Higher total day** minutes and charges
- Make **fewer but longer calls** during the day
- **Higher international** minutes and charges

- **Consistent usage** across day, evening, and night
- Regular **call durations**
- **Lower international** minutes and charges

Customer Service Interaction



- **Frequent calls** to customer service (4+ calls)
- Churn rate approaches **100% with 7+ calls**
- Indicates **unresolved issues** and **dissatisfaction**

- **Few** customer service calls (0-2 calls)
- Issues are **resolved promptly**
- Higher overall **satisfaction**



Charges and Billing

- Higher avg **charges per account day** (~ 10% more)
- Sensitive to **fluctuating international rate**
- Perceive **lower value** for money (cost-quality)

- **Lower average charges** per account day
- **Better value perception**
- **Less sensitive** to slight price variations

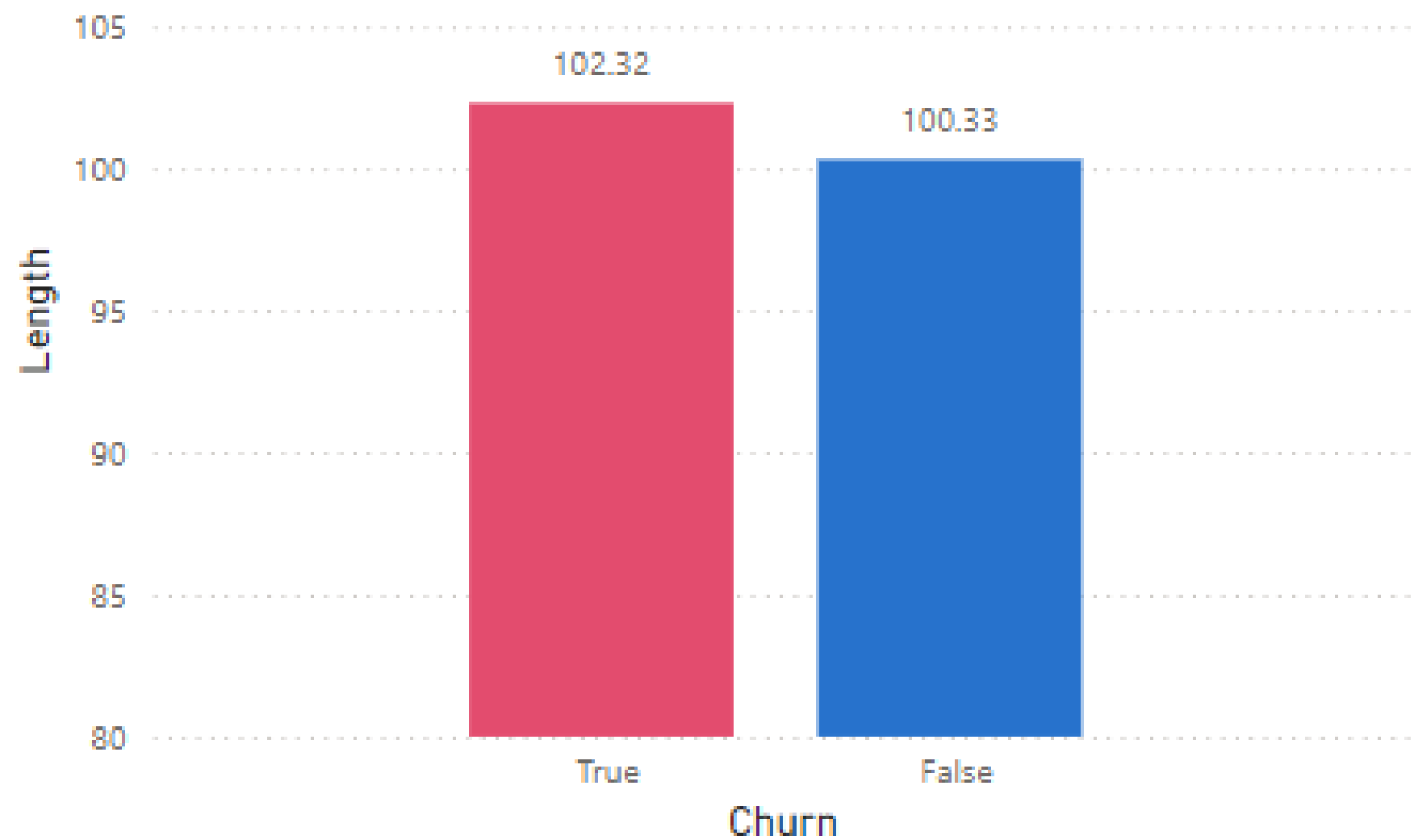


DEMOGRAPHICS AND CUSTOMER TENURE



Churned Customers Likely Have Longer Account Lengths...

Average Account Length by Churn



... Longevity Does Not Guarantee Loyalty

By Over 2%

Mean account length for churned customers is slightly higher compared to non-churned customers (102.32 > 100.33)

T-Test show no significant difference, yet, still indicating **churn can occur regardless of tenure**

→

Key Insights

Contrary to expectations, especially in service industry, churned customers are more inclined to a longer account length than non-churned ones.

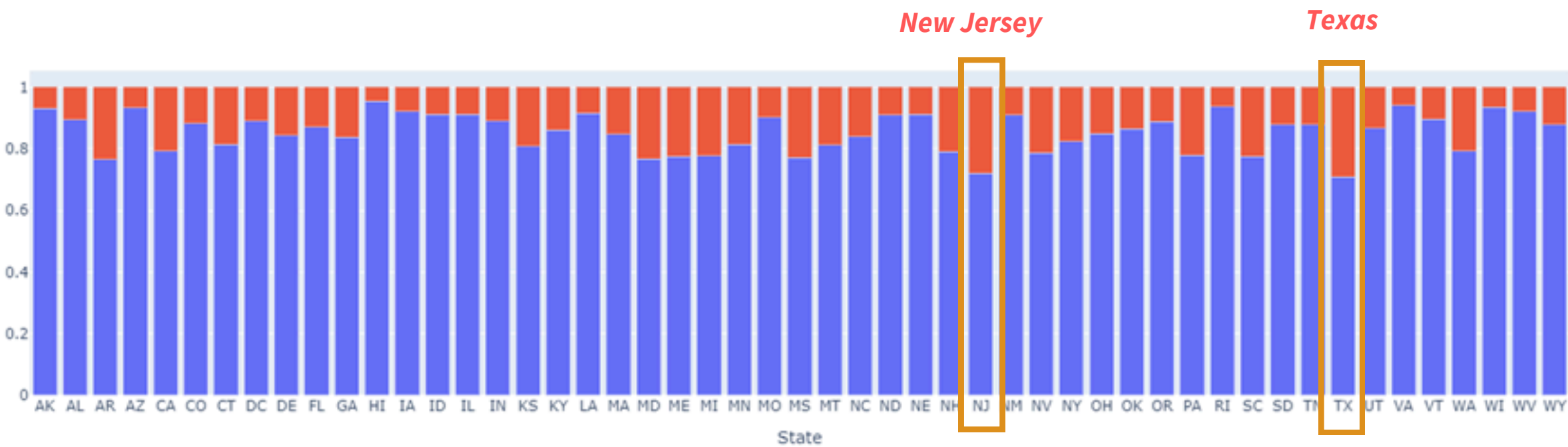
↓

This suggest that **even long-term customers are at risk of churning if their needs are not continuously met** while the number of substitute is on constant rise overtime

New Jersey and Texas Exhibit Exceptionally Higher Churn Rates

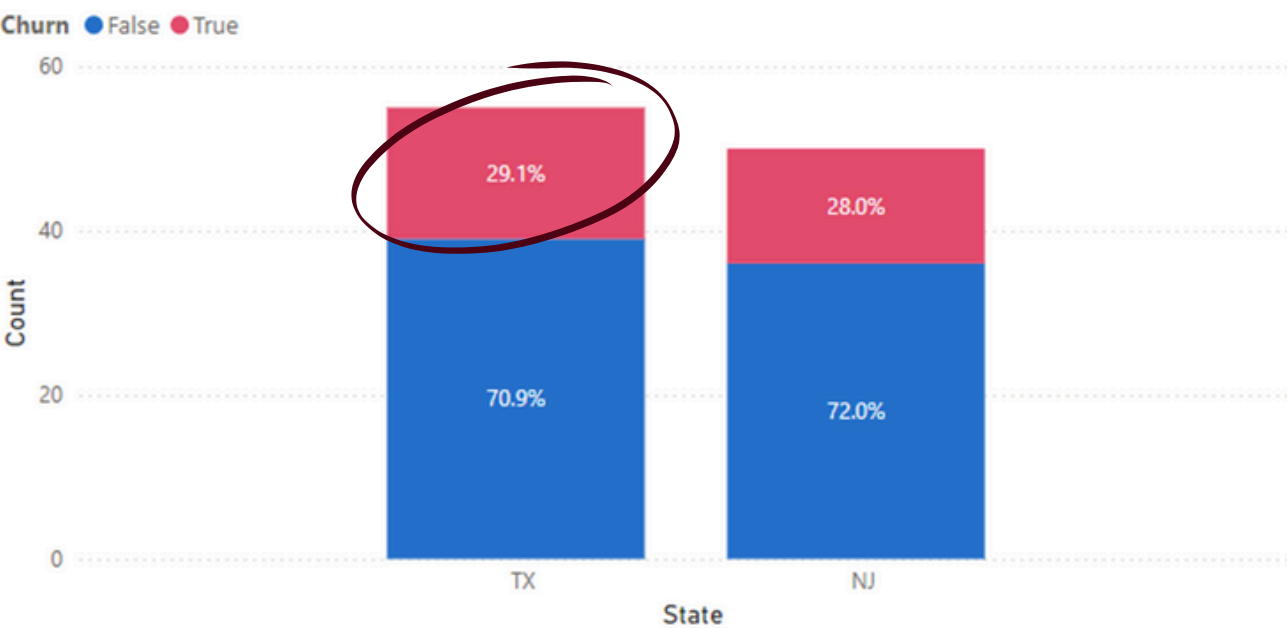
... Highlighting Potential Geographical Impact

Churn Rate by State (100% Stacked)



Filter by Top States

New Jersey & Texas - States w/ Highest Churn Rates



- AR, MD, ME, MI, MS, PA, SC have moderately higher proportion of churned users, **exceeding the threshold of 20%**
- States like **New Jersey and Texas** have churn rates **exceeding 29%**, significantly higher than the average



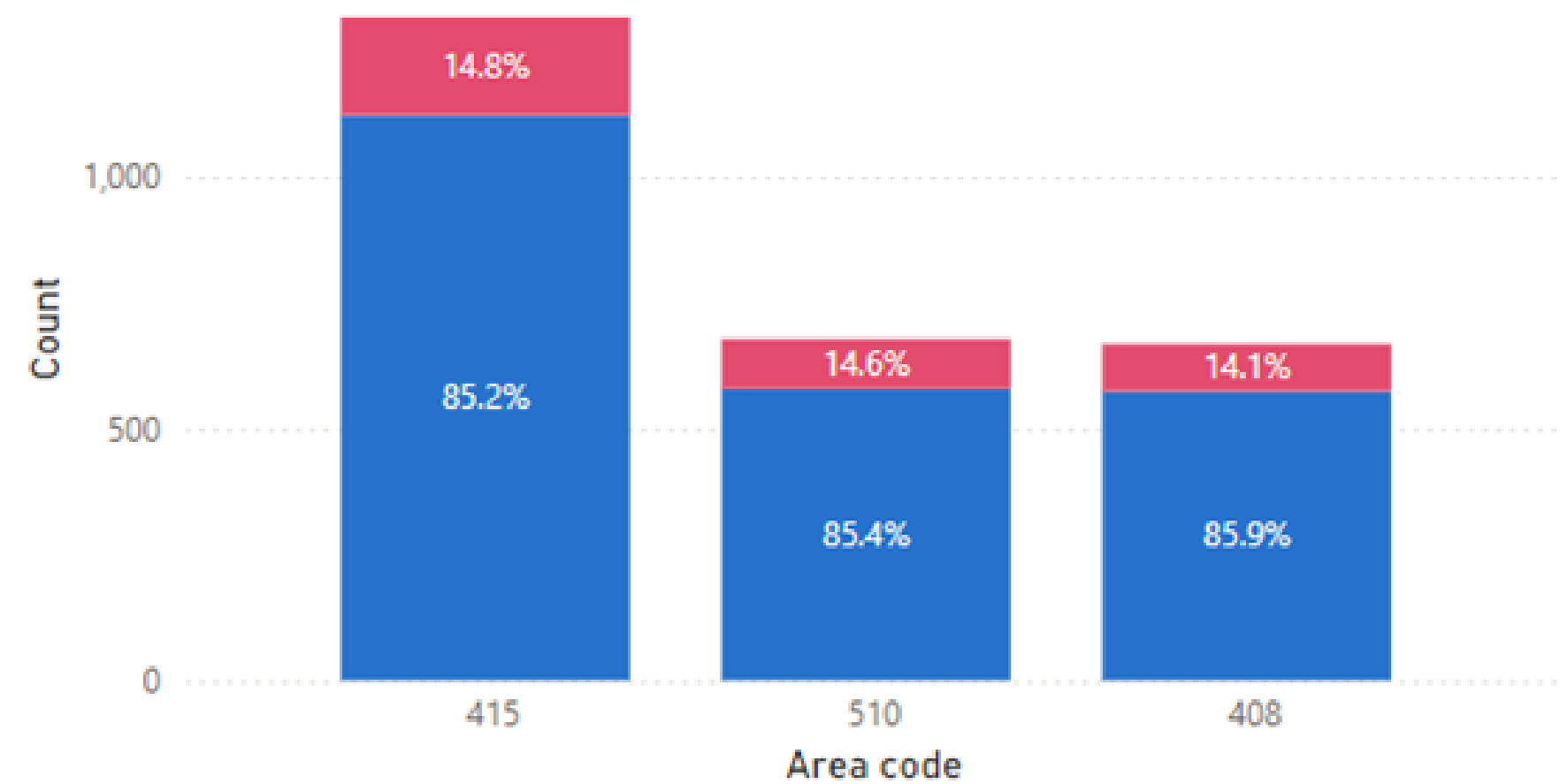
Key Insights

This pattern of separation can likely be attributed to **regional service issues, emerging local competition, or demographical differences**, which require further local product investigation.

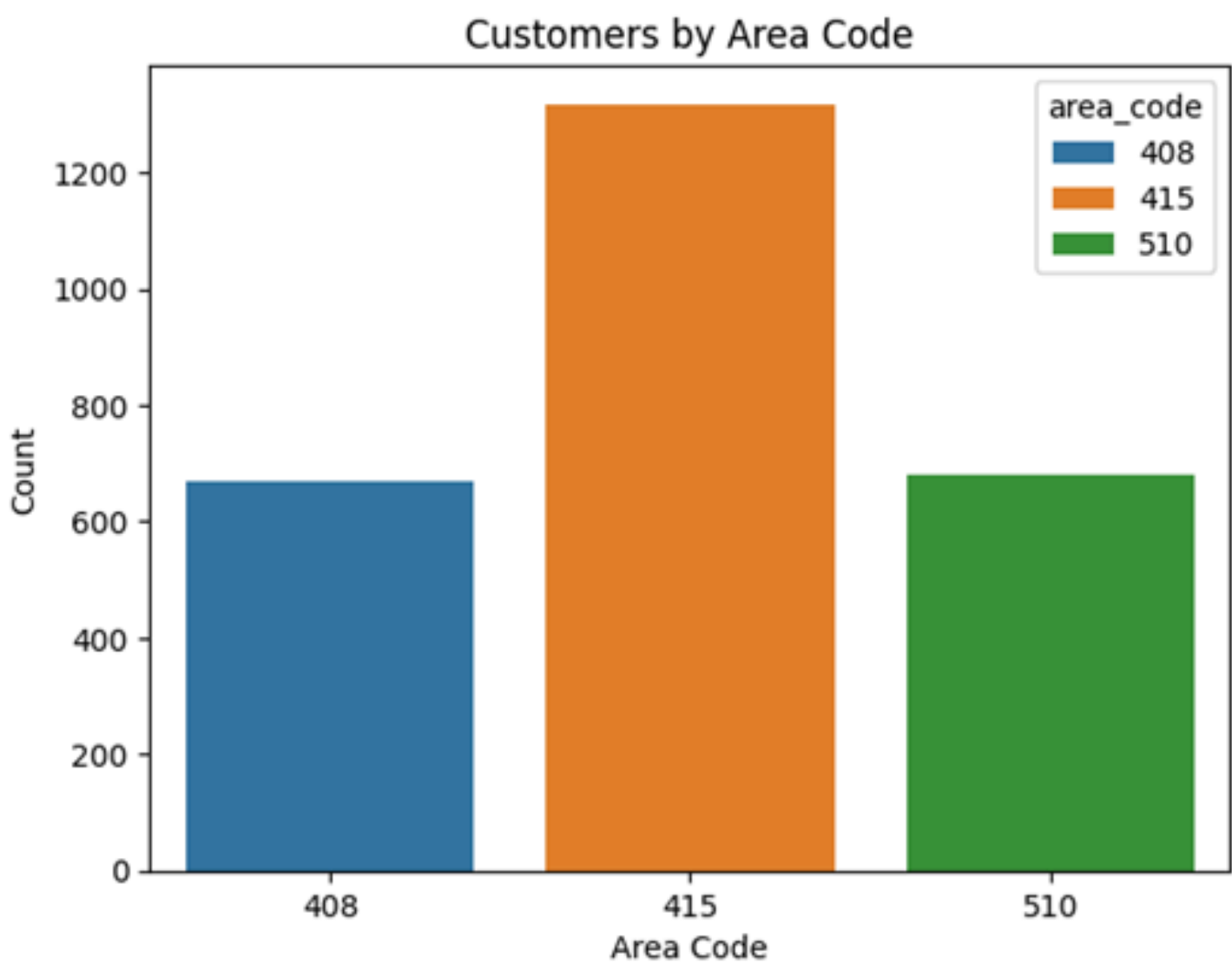
No noticeable difference between Area Codes ... Despite Area Code 415 Being Dominant In Users

Area Code by Churn

Churn ● False ● True



- **No discernible relationship** between **area code** and **state**.
- The distribution is pretty **constant across all states**, with 415 area code taking up the lion's share of users.



➔

Key Insights

There is **no difference in figures for Area Code** at all
>> This feature might be **redundant for churn rate analysis**, which can be removed from predictive models

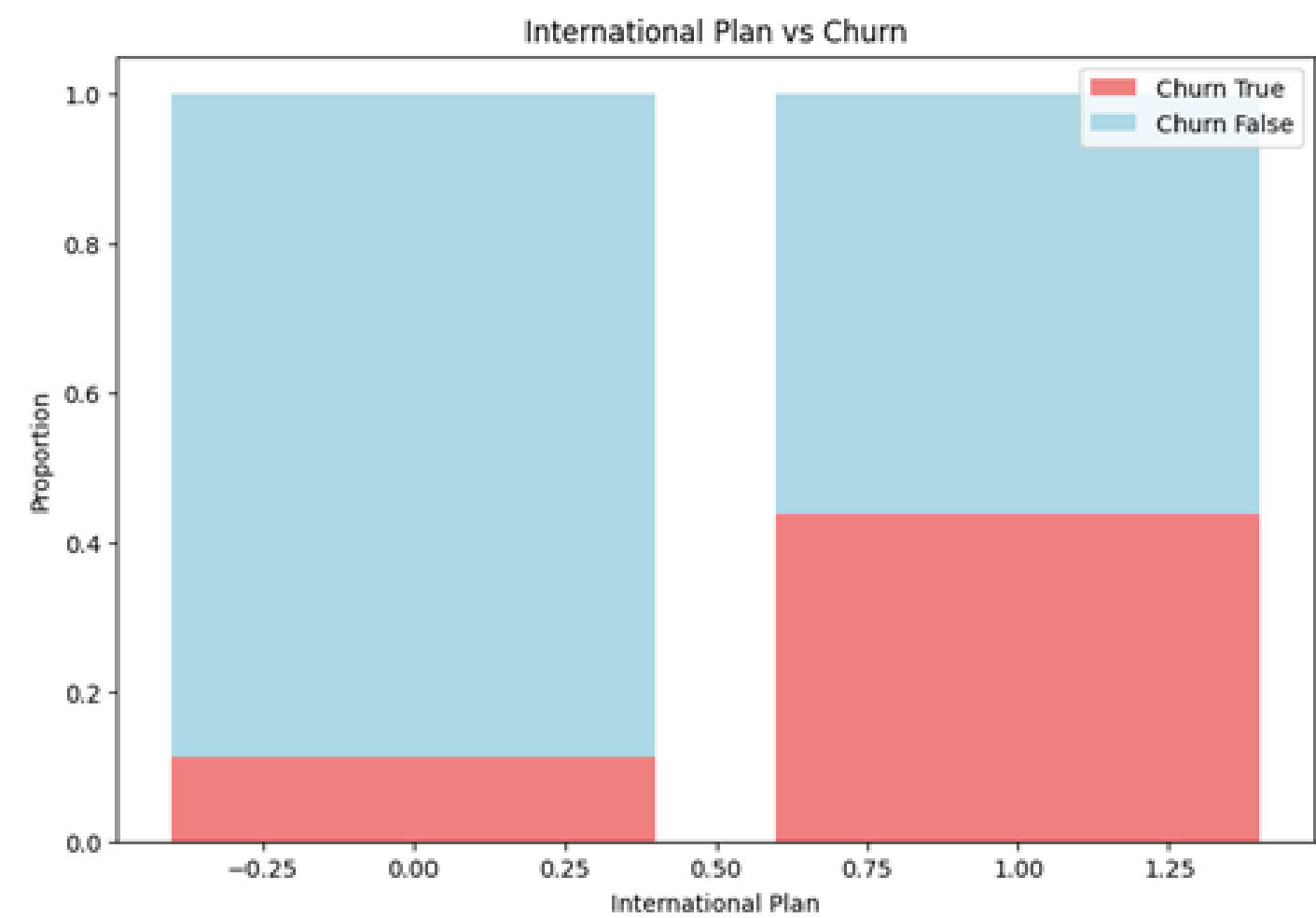


SERVICE PLAN AND VALUE-ADDED OFFERINGS



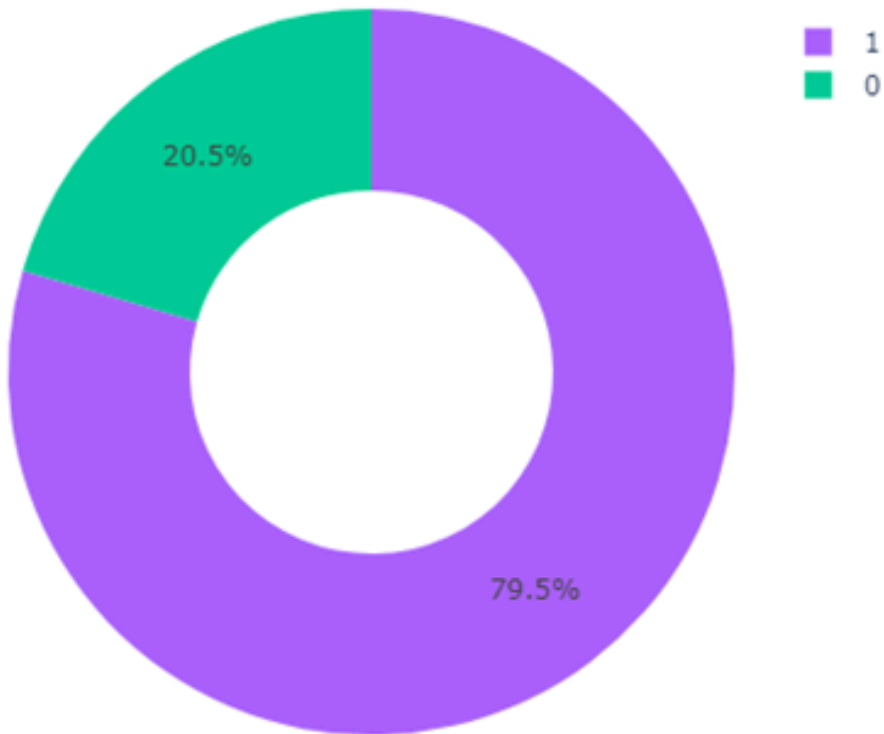
Users with International Plan are Nearly 60% More Likely to Churn

Churn Proportion of International Plan Users



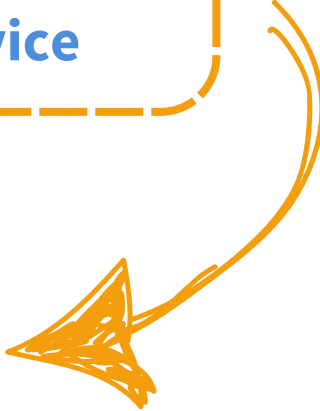
- Among **customers with International Plan**, nearly **50%** have churned – **a substantial increase compared to merely 10%** for those without the service

Churned International Plan Users vs. Total Churns



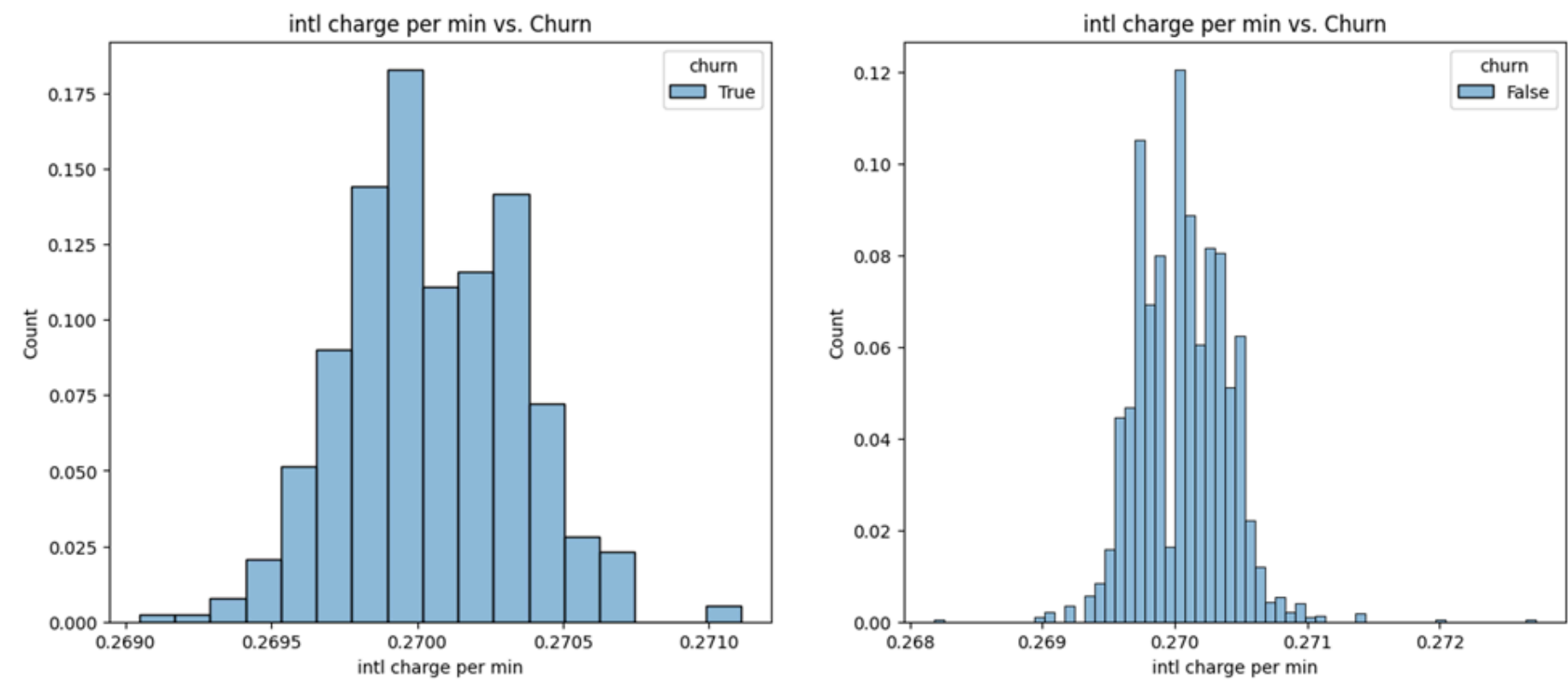
- ... While compared with all churned users, **a whopping 79.5% have actually used the service**

WHAT ARE THE CATALYSTS BEHIND THIS ALARMING FIGURE?



Users with International Plan are Nearly 60% More Likely to Churn

Charge per min Distribution by Churn Type



Churned customers have **slightly higher total international minutes and charges**, suggesting they use the service more but may be **dissatisfied with the cost-to-quality ratio**



Churned customers also make **fewer international calls**, which could indicate reduced service usage before churning, potentially **due to immediate dissatisfaction** with the service, triggering their decision to leave

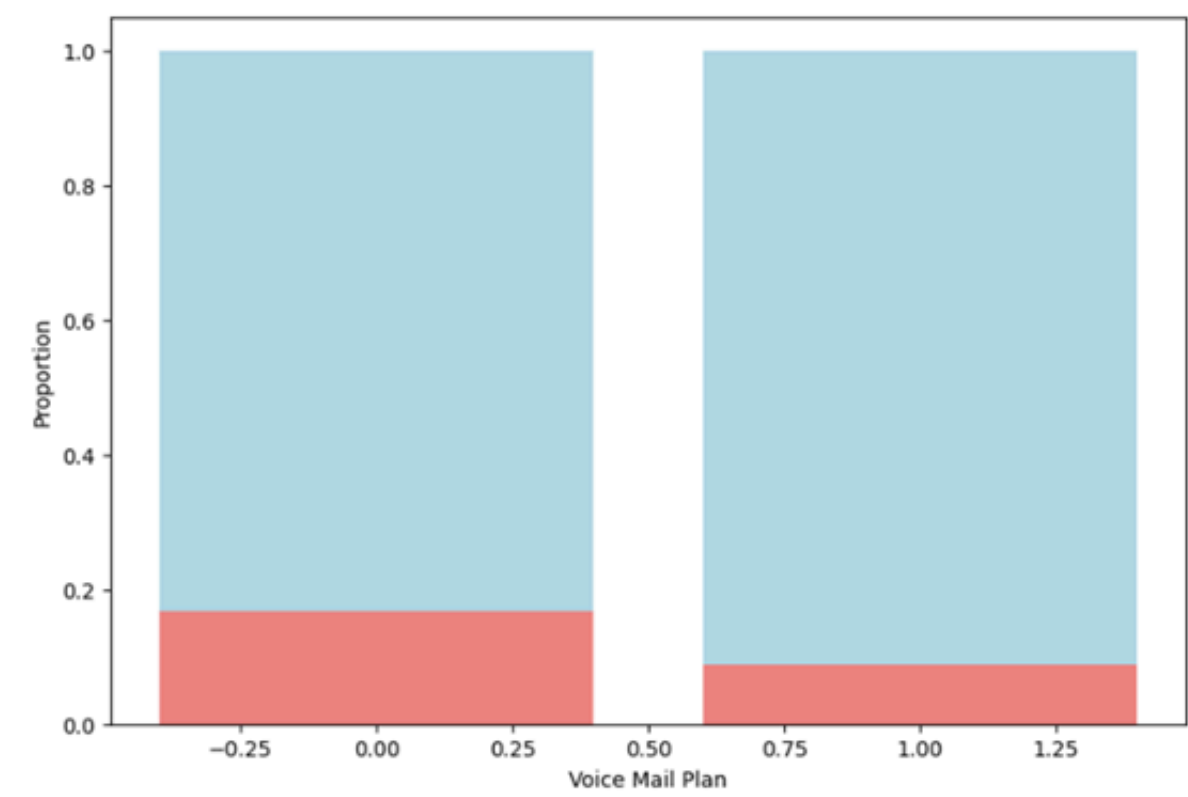


Key Insights

- Churned users** make fewer international calls but **accumulate more total minutes and incur higher charges** compared to non-churned users. Their **narrower range of charges per minute** indicates **greater price sensitivity**, suggesting that even minor unusual price variation (i.e. for longer calls) could lead them to switch providers.
- International services are better suited for short calls**, as longer calls lead to **disproportionate costs**. Even minor price variations can negatively impact customer experience and drive churn.

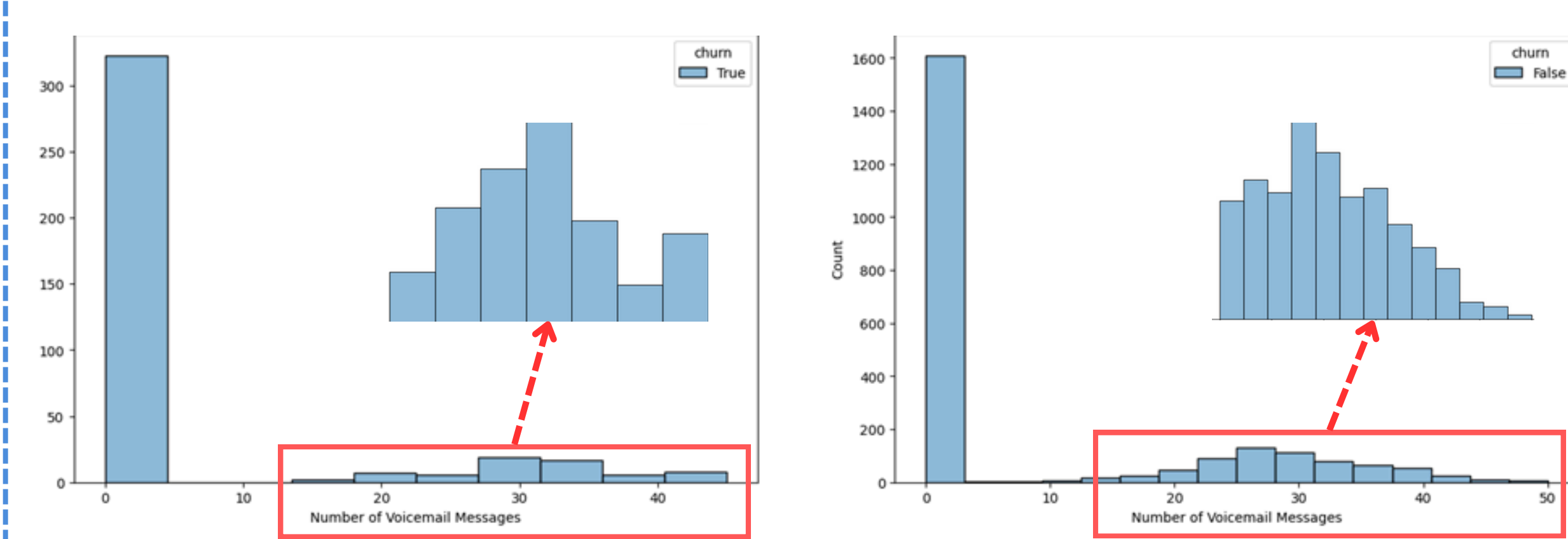
Voice Mail Plan Usage is Proven to Decrease Churn Rate by Almost 50% ...

Churn Proportion of International Plan Users



Only **34.7% of Churned Users** used the service, the Voice Mail Plan for which has a **reduced churn rate (around solely 10%)**.

Distribution of Voice Mail Messages Sent by Churn Types



Overall, **Voice Mail Plan users tend to be more engaged** compared to other services, sending around 30-40 messages on usual. Especially for **those with 45 messages onward, no churn was recorded**

➔

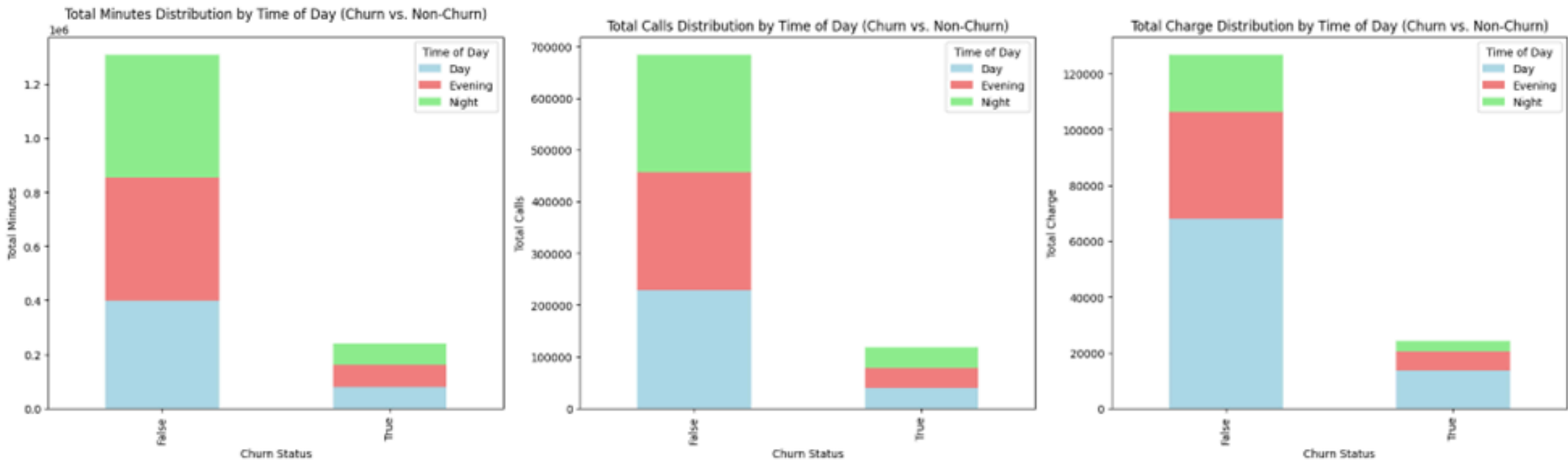
Key Insights

- **Voicemail service usage does not seem to be a major contributor** to churn, but those who use the service much **more frequently** (non-churned customers) tend to be **more satisfied or engaged**
- Users of this service tend to **form a more positive image of the brand**, resulting in better loyalty.

USAGE PATTERNS AND BEHAVIORS

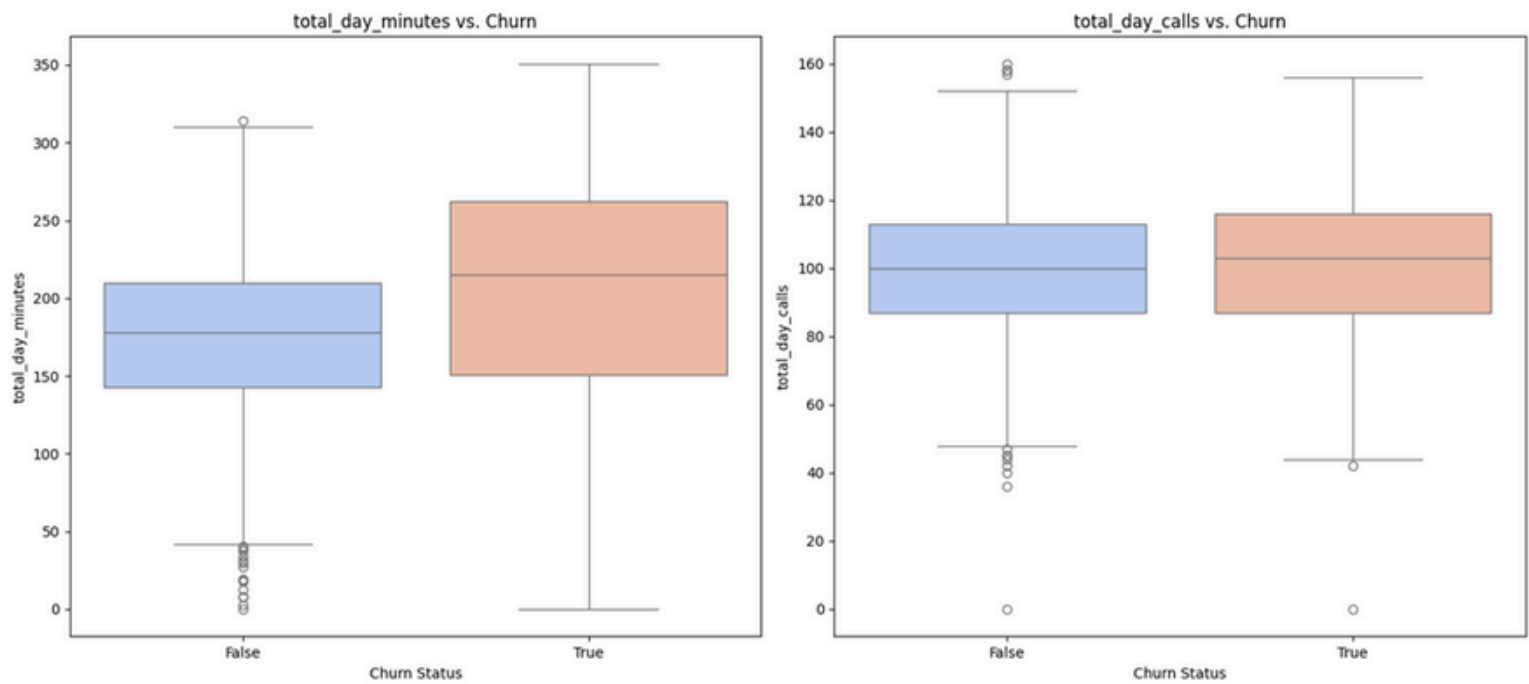
Churned Customers Are “Heavier” Daytime Users, Leading to Higher Charges Without Perceived Added Value

Breakdown of Minutes/Calls/Charges by Time



Overall, each dimension of **Day, Evening, and Night** shares **relatively equal proportion** for both Churn and non-Churn groups. There is **no significant difference** between churned and non-churned users in terms of **evening and night calls** as well as **evening and night total minutes**.

Total Minutes/Calls Distribution & Stats

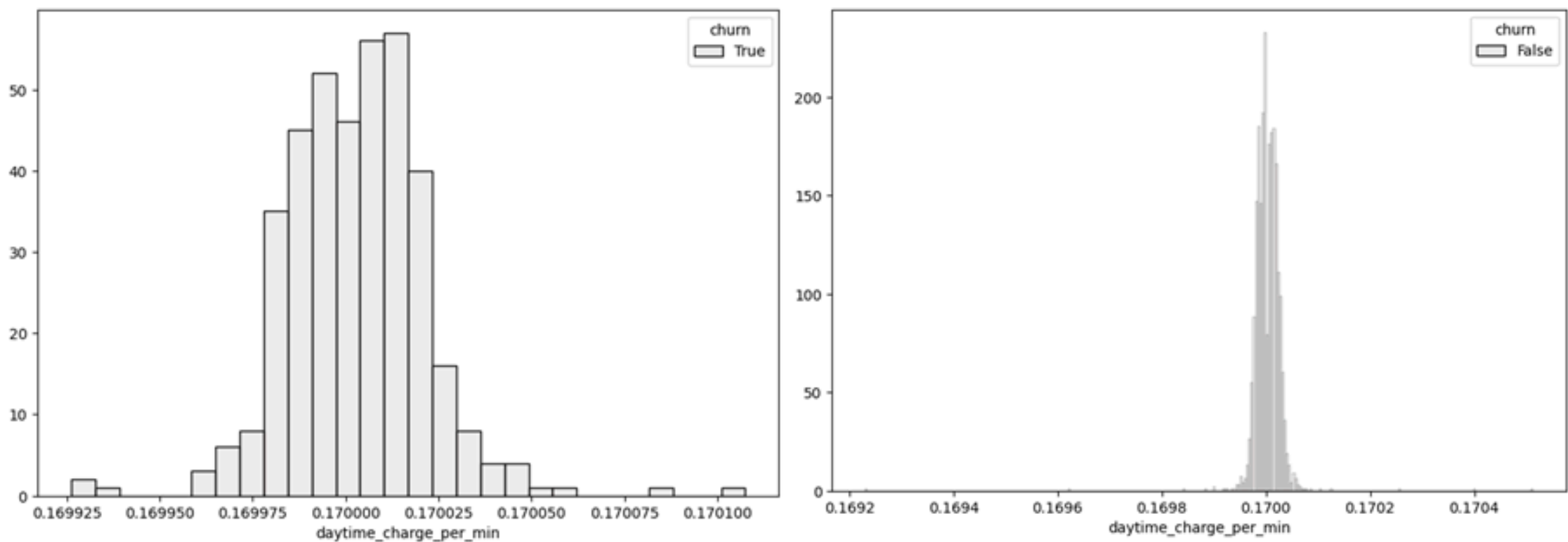


Churned users make **a similar number of daytime calls** (~100) but have notably **higher total day minutes** (205.18 vs. 175.1) and incur higher charges.

Churned users are making **fewer but longer calls during the day** (or in other words, **heavy daytime users may be more likely to churn**, potentially due to dissatisfaction with the **service quality** and probably pricing during peak hours for longer call duration)

Churned Customers Are “Heavier” Daytime Users, Leading to Higher Charges Without Perceived Added Value

Daytime Charge per min Distribution by Churn Type



The **distribution of daytime charge per minute** ffor churned users is more concentrated between **0.1699** and **0.1701**



Churned users are **charged very similarly** for daytime usage, with little deviation in the rate per minute



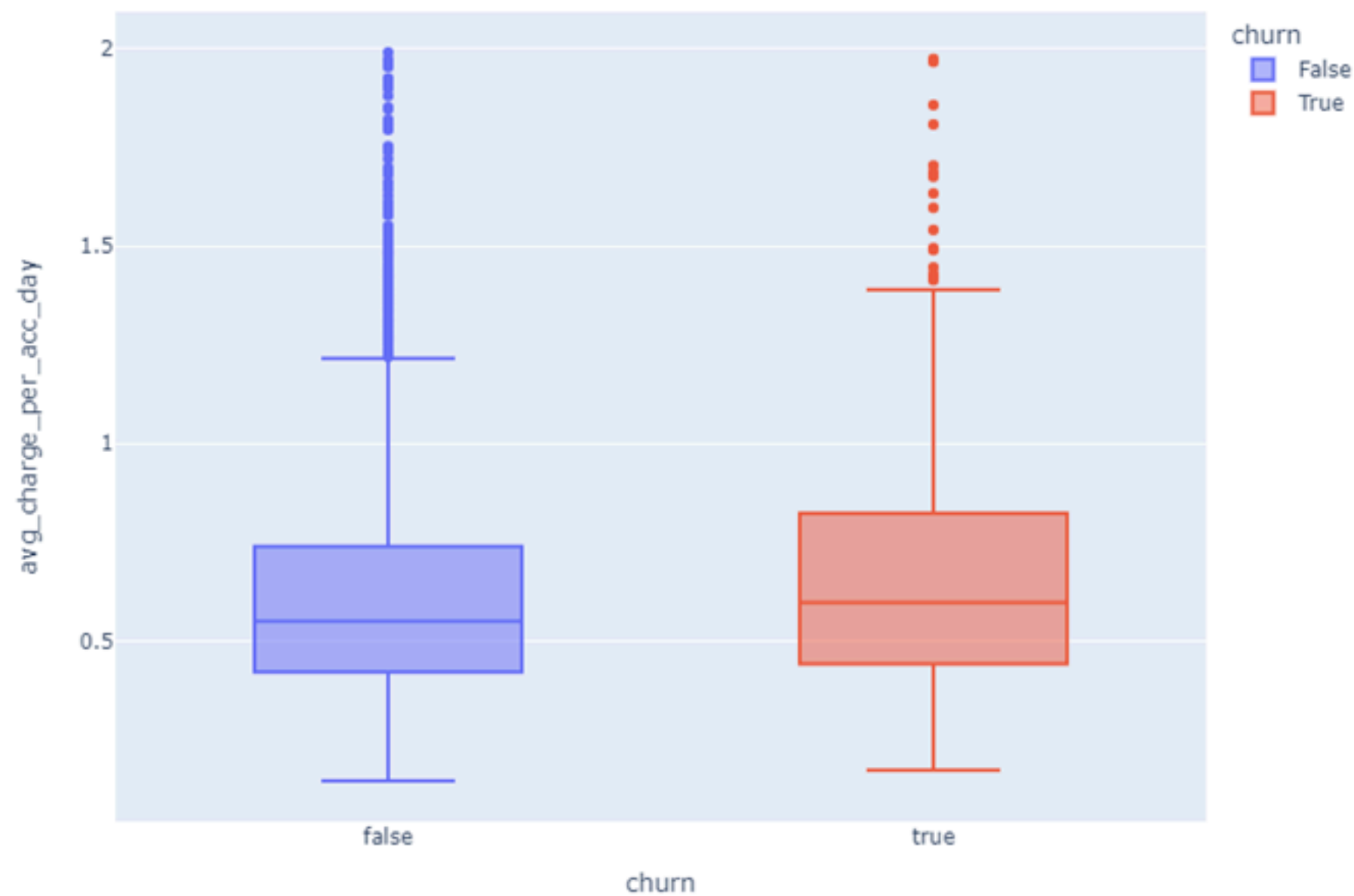
Key Insights

Pricing variability is not the sole factor for churned users, but possibly its effect **in relation with other issues** (i.e., **service dissatisfaction or call quality** during the day as I mentioned previously) could be contributing to their decision to churn.

Churned Customers Incur 10% Higher Average Charges/Account Day, Indicating Possible Dissatisfaction with Cost-to-Value Ratio

By Around 10%

Average Charge per Account Day Distribution & Stats



Despite similar pattern in distribution of spending behavior, the **average for churned users** is **still slightly higher** than non-churned ones

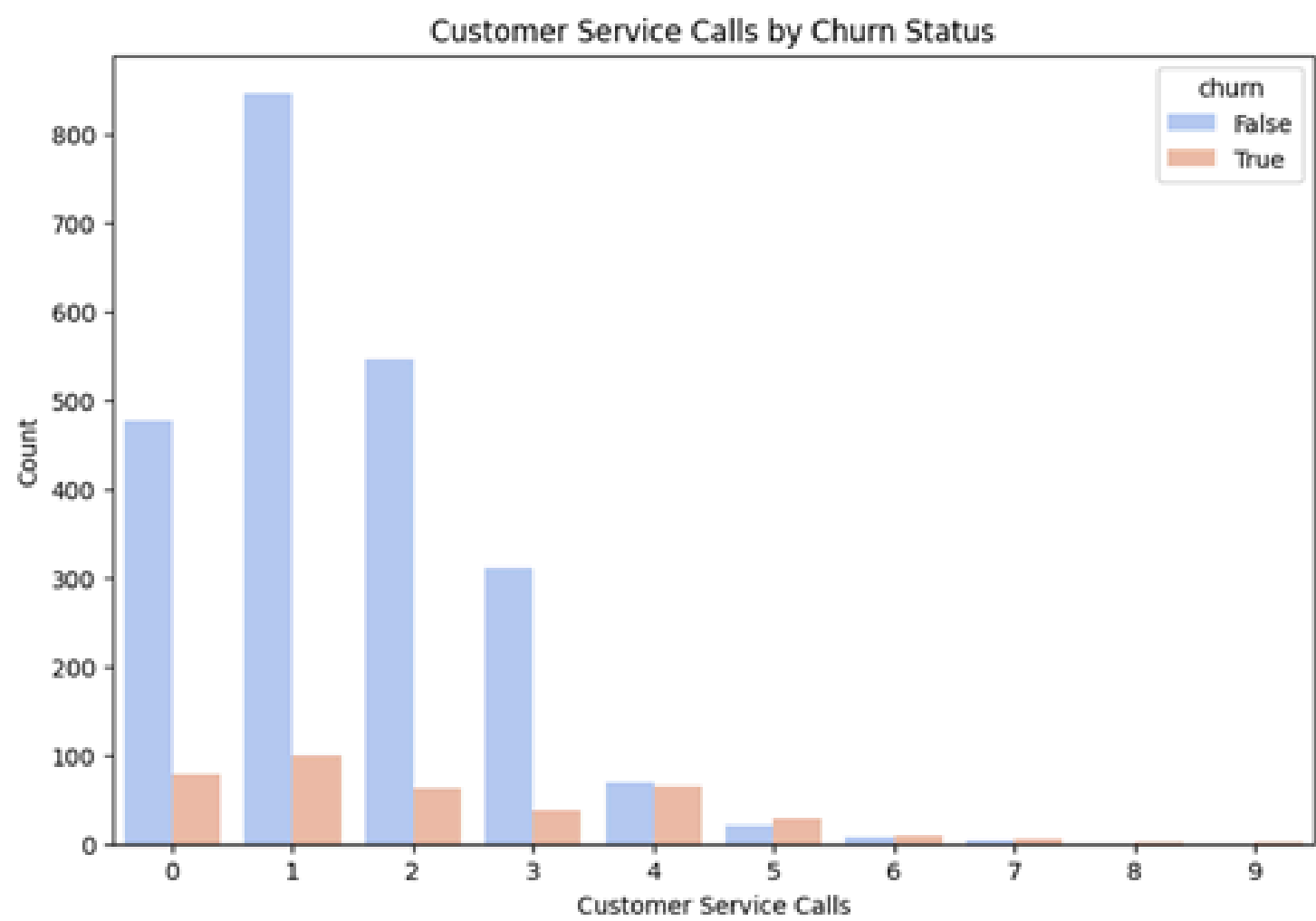


Key Insights

- Churned Users may feel they are **not receiving value commensurate** with their spending.
- Could be due to perceived overpricing or **lack of suitable range of calling plans**

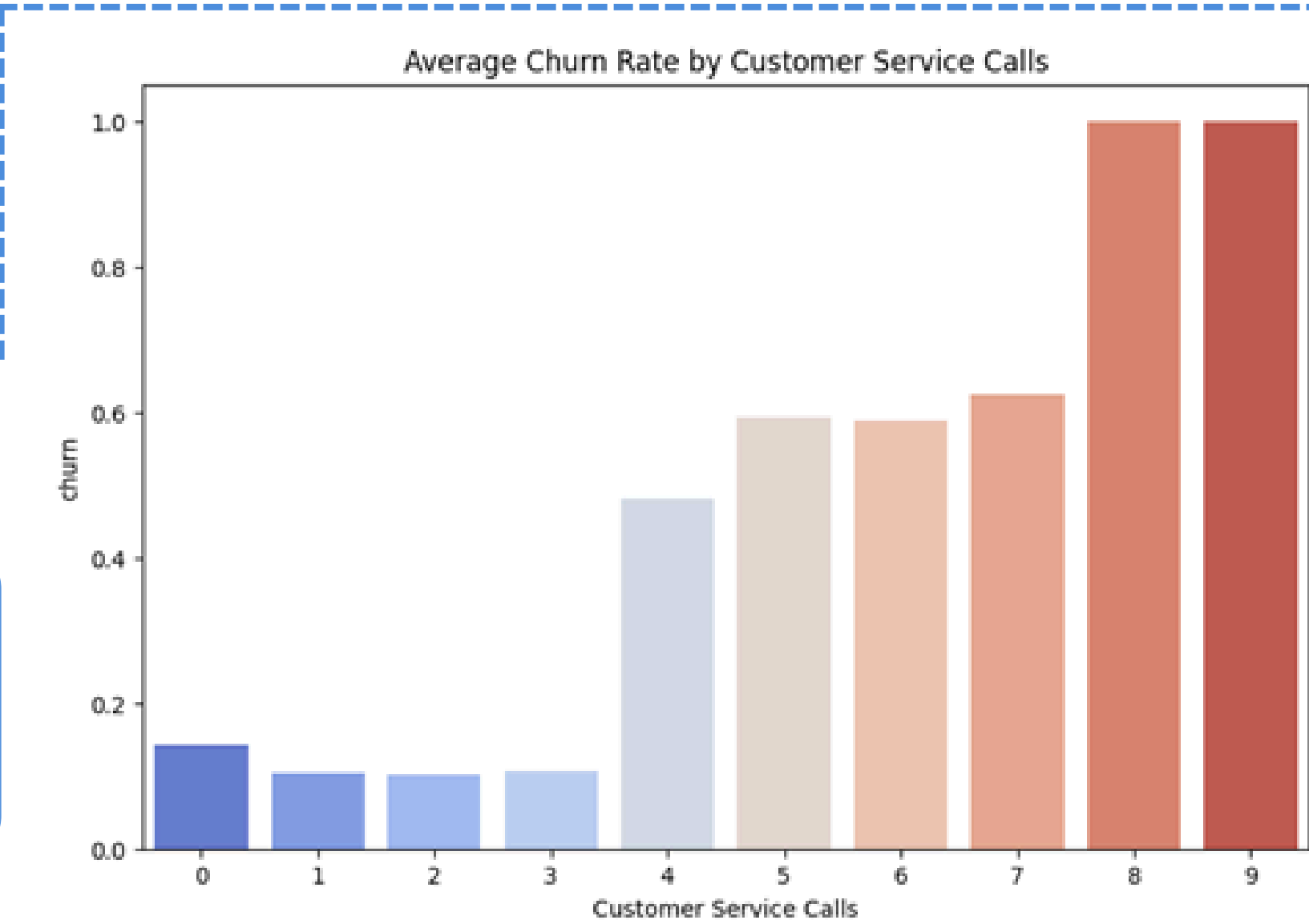
CUSTOMER SERVICE INTERACTION

Frequent Customer Service Calls Strongly Predict Churn, Especially Beyond Three Calls



Customer Service Call Frequency

- Most **non-churned users** made between **0 and 2 calls** to customer service, while churned users are **overrepresented in categories with more calls**.
- As the frequency of customer service calls increases, the likelihood of churn rises significantly.

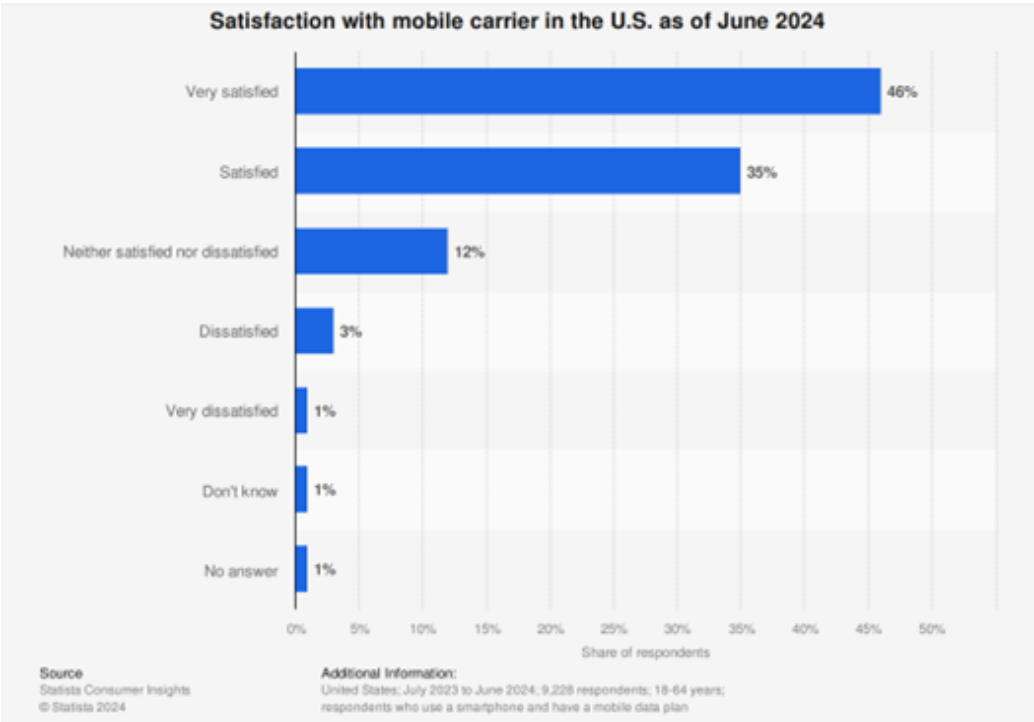


Churn Rate Correlation

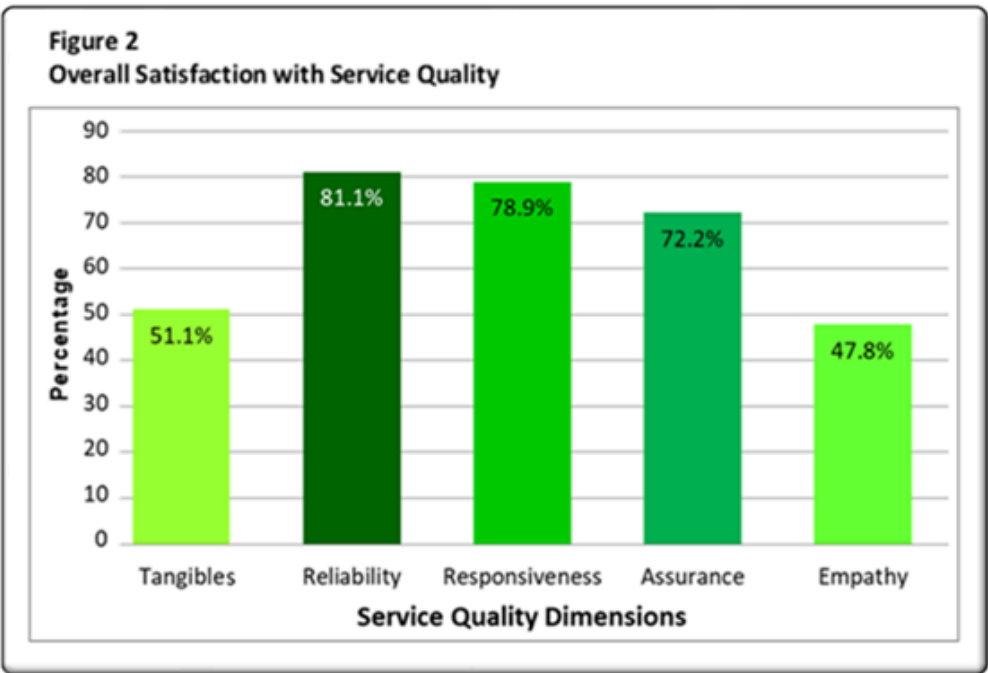
- Churn rate remains relatively **low for users making 0 to 3 calls**, but **increases drastically from 4** customer service calls onwards
- By **7+ calls**, the churn rate approaches **100%, highlighting severe dissatisfaction** among high-call customers

Close Correlation Between Customer Service and Satisfaction

In the **large U.S. market**, mobile plan users generally report **high satisfaction**,



(source: Statista)

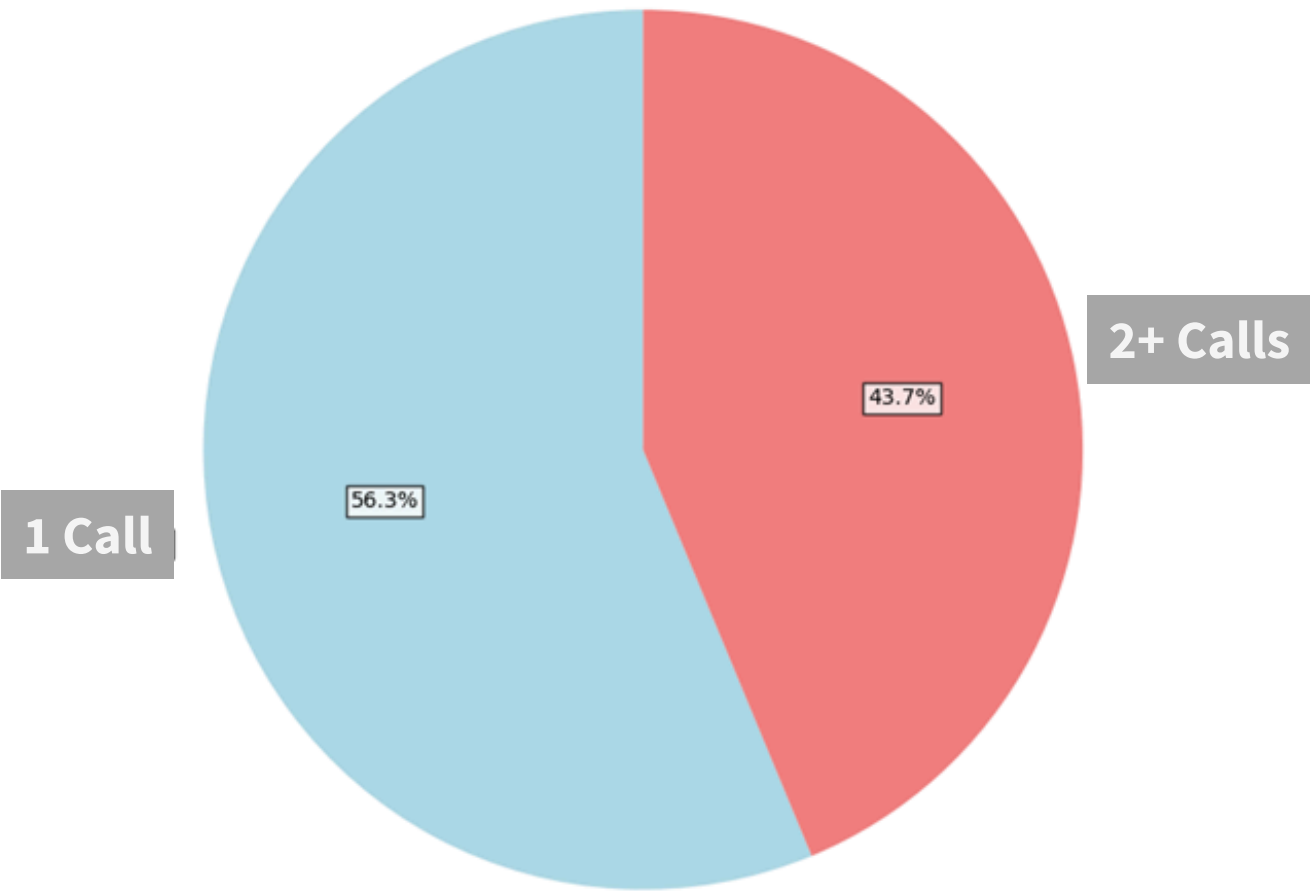


(source: William Chua, 2019)

Direct linkage between customer satisfaction and customer service (i.e. responsiveness), we can infer a generally **favorable market of telecommunication to date**.

High Customer Service Interaction Indicates Unresolved Issues

... Our business shows the **opposite trend**, proving significant issues with customer satisfaction



Key Insights

- 1st Call Resolution dominates, however, 2+ calls still take place quite often, **highlighting a gap in customer service**
- High interaction with customer service (**4+ calls**) indicates unresolved issues leading to **frustration and dissatisfaction**

➔ **Strong Predictor of Churn**

Machine Learning Modelling

A

Preliminary Feature Engineering (Appendix)

- I. Feature Generation
- II. Preliminary Feature Evaluation
 - Tackling Multicollinearity
 - Information Value Model

B

Model Construction

- I. Overview of Development Pipeline
- II. Why LightGBM and XGBoost are Shortlisted? (Appendix)
- III. Comparison & Final Model Selection (Appendix)

C

Final Model Deployment & Interpretation

- I. Evaluation Metrics
- II. Feature Importance & Insights

B. Model Construction

Development Pipeline Overview

1. **Shortlisting most appropriate ML Models:** Consider Ensemble Learning Models, such as XGBoost and LightGBM, to tackle class imbalance and prevent overfitting with decision tree algorithms.
2. **Assessing the necessity of Resampling to tackle Class Imbalance:** Placeholder for SMOTE upsampling method if needed.
3. **Generating 2 separate versions of both training and holdout datasets:** Two final datasets after EDA, with different feature sets, will be tested to evaluate the necessity of specific features (voice_mail_plan, intl_charge_per_min_cate, total_eve_calls, account_length).
4. **Tuning Hyper-Parameters with Grid Search CV:** Grid Search will be used for hyperparameter tuning due to time and efficiency constraints, instead of Bayesian Search.
5. **Fitting Optimized Models with each dataset version.**
6. **Comparing and Assessing Performance Metrics:** Evaluate models using metrics like ROC AUC or precision-recall due to class imbalance.
7. **Preliminary Feature Adjustment for Final Deployment:** Refine the final subset of features for deployment.
8. **Final Model Implementation with BayesSearch CV & Interpretation:** Run Bayes Search CV for the final model for maximal optimization of hyper-params before deployment

C. Final Model Deployment & Interpretation

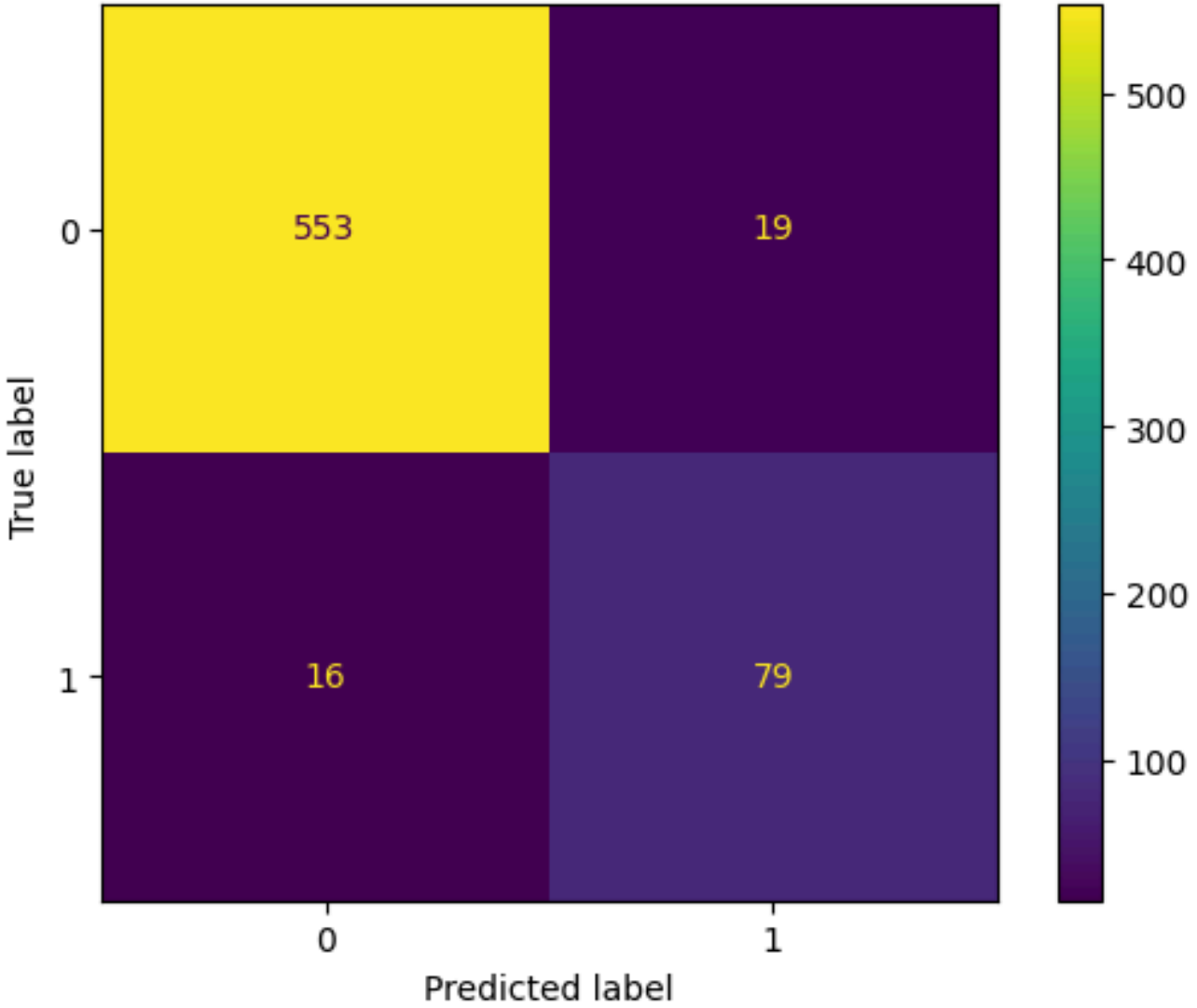
Evaluation Metrics

LGBMClassifier

```
LGBMClassifier(colsample_bytree=0.8, learning_rate=0.05946682769647679,
               max_depth=4, min_child_weight=3, n_estimators=200,
               random_state=42, reg_alpha=0, reg_lambda=0,
               scale_pos_weight=5.871134020618556)
```

- This model shows excellent performance in predicting customer churn, with a high overall **accuracy of 95%**. It excels in identifying non-churners (class 0) with **97% precision** and **98% recall**, ensuring accurate classification of most non-churners.
- For churners (class 1), the model achieves **81% precision** and an **82% F1-score**, balancing precision and recall effectively.
- With a **recall of 83%**, there's minimal room for improvement in capturing potential churners.

Overall, the model handles imbalanced data well, and slight optimization in reducing false negatives could further enhance its impact on customer retention efforts

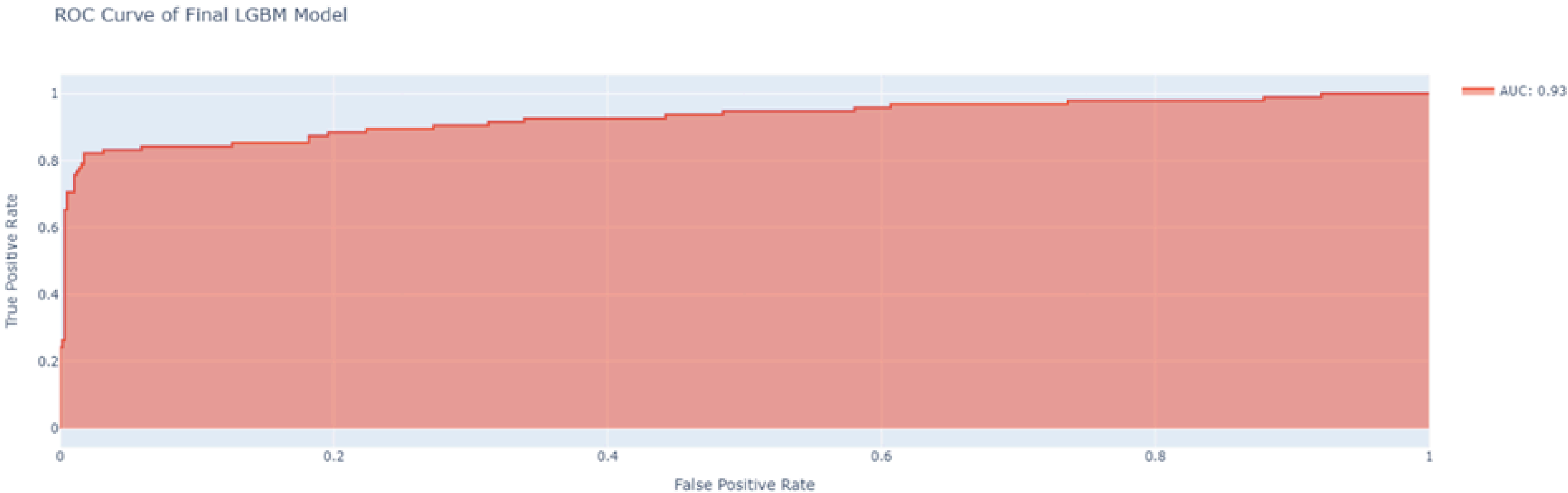


C. Final Model Deployment & Interpretation

Evaluation Metrics (ROC Curve)

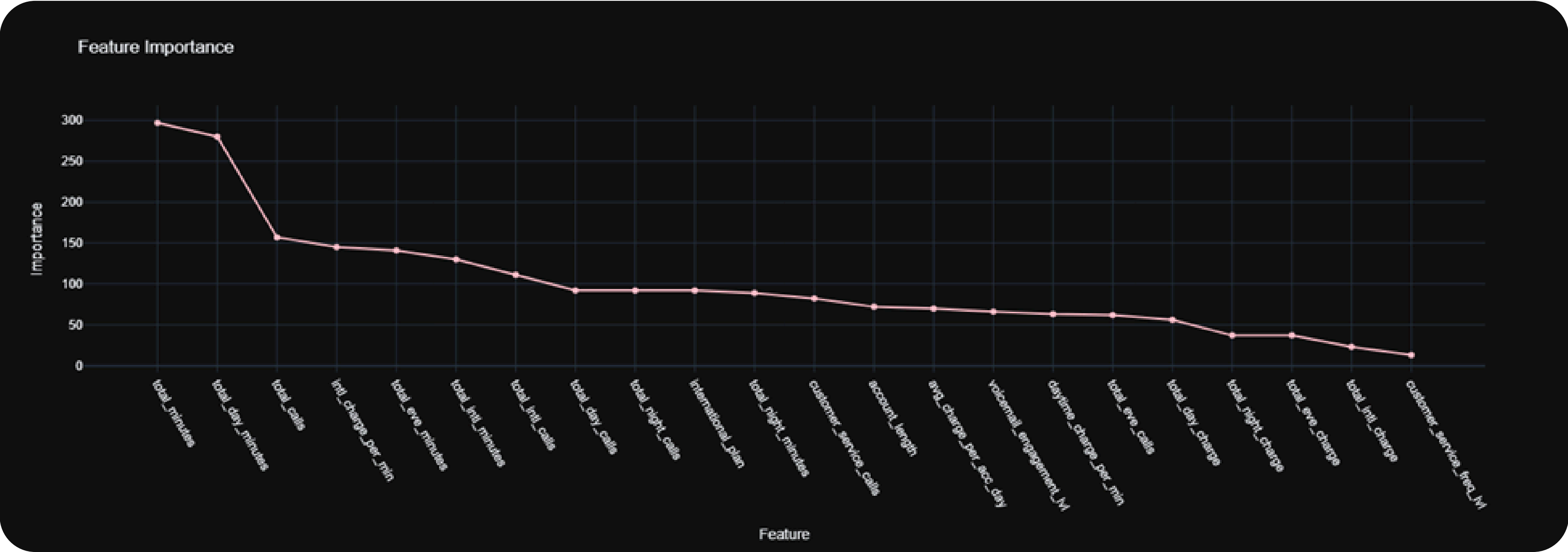
Note: AUC and Precision-Recall Score are important metrics to evaluate models with Class Imbalance

The ROC curve for the final LightGBM model shows impressive classification performance, with an AUC score of 0.93. The curve being close to the top-left corner means the model does a great job distinguishing between churners and non-churners even in an imbalanced dataset



C. Final Model Deployment & Interpretation

Feature Importance



C. Final Model Deployment & Interpretation

Feature Importance Interpretation (Model & EDA-based)

- 1. Usage Patterns and Behaviors:** High daytime usage (**total_day_minutes, 280**) is the strongest churn predictor. Overall high usage (**total_minutes, 297**) and significant evening usage (**total_eve_minutes, 141**) are linked to churn, indicating high-usage customers are sensitive to service quality and pricing. Frequent international calling (**total_intl_minutes, 130**) also contributes to churn. High call volumes during the day (**total_day_calls, 92**) and overall (**total_calls, 157**) suggest that customers reliant on voice services may churn if their needs aren't met.

2. Charges and Billing: Higher international call charges (**intl_charge_per_min, 145**) and daytime call costs (**daytime_charge_per_min, 63**) increase churn risk, highlighting price sensitivity. Daily charges (**avg_charge_per_acc_day, 70**) correlate with churn, where perceived value for money impacts retention.

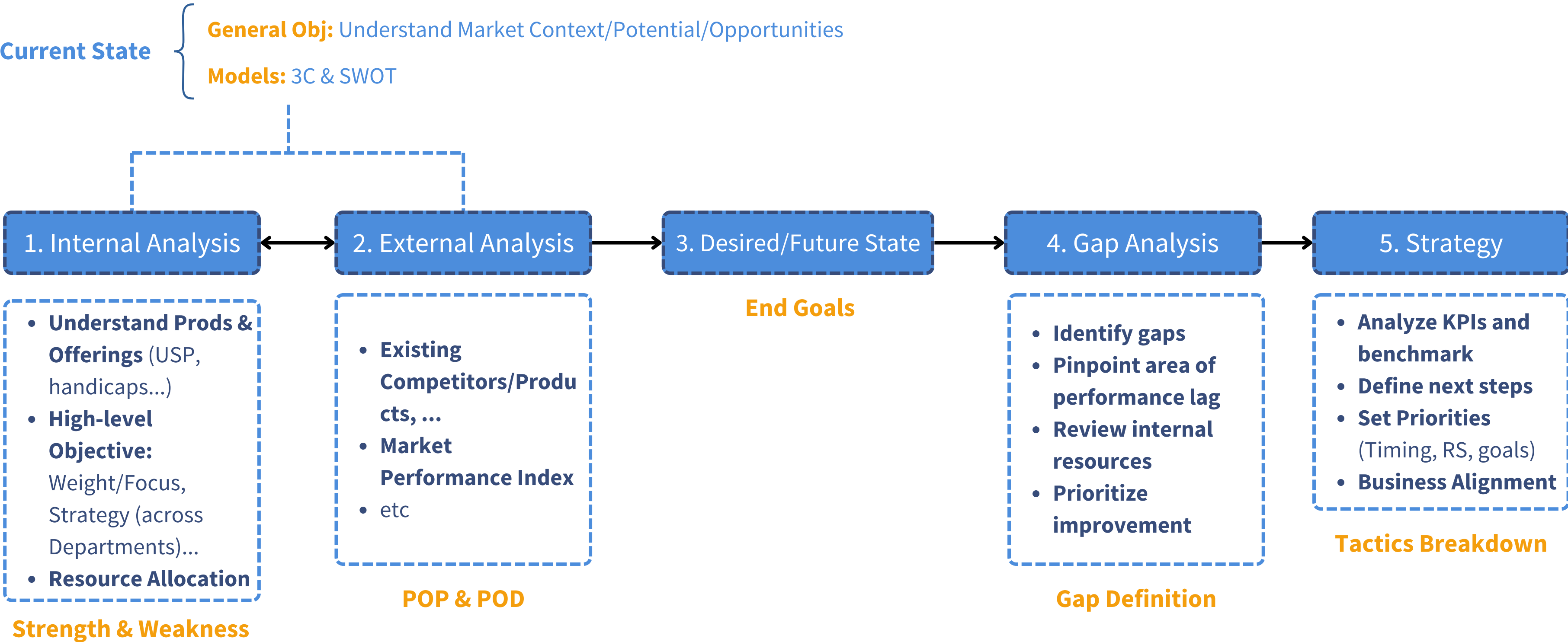
3. Value-Added Services: Having an international plan (**international_plan, 92**) and frequent international calls (**total_intl_calls, 111**) predict churn, emphasizing the importance of service quality and pricing. Voicemail engagement (**voicemail_engagement_lvl, 66**) shows that value-added services can enhance loyalty

4. Customer Service Interaction: Frequent customer service calls (**customer_service_calls, 82**) are linked to churn, meaning unresolved issues drive customers away.

5. Account Tenure: The length of time with the company (**account_length, 72**) affects churn, indicating both new and long-term customers need tailored retention strategies, as loyalty isn't guaranteed if service changes no longer meet their needs.

Planning & Recommendation Thinking Process

Disclaimer: Since this is a fictional telecom business with undefined specifics, providing a precise recommendation for improvement is unfeasible. However, I will depict my step-by-step problem solving approach and give a high-level yet most practical recommendations possible



Recommendation List

High Priority

Proposed Solution	Step-by-Step Approaches	Rationales
Revamp International Calling Plans to Reduce Churn Among International Plan Users	<div><div>1. More Personalized Pricing Models/Plans: Tiered pricing or bundled international minutes to offer cost savings for longer calls, or at least fixed expected price without much variability</div><div>2. Transparent Communication: Clearly communicate rates, fees, and any changes to avoid unexpected charges.</div><div>3. Quality Assurance: Invest more in network infrastructure to improve international call quality, reducing dropped calls and poor connections.</div><div>4. Customer Feedback Integration: Regularly solicit feedback from international plan users to identify pain points and areas for improvement, ready for iterative improvement overtime.</div></div>	<div><div>• By Addressing cost Concerns through Better Pricing, the company can improve customer satisfaction and perceived value, which is currently a major point of dissatisfaction. Enhancing call Quality also reduces frustration, boosting the overall service experience.</div><div>• Transparency in Communication fosters Trust, helping customers feel More valued and Less likely to churn.</div><div>• Revamping International plans positions the company more competitively, making it an attractive choice over competitors with Better offerings.</div></div>
Enhance Customer Service Efficiency to Resolve Issues Promptly	<div><div>1. Training Programs & Quality Monitoring: Provide advanced training for customer service representatives based on call monitoring and feedbacks to handle a wide range of issues effectively, the goal is to optimize personnel's capability, avoiding over 1st resolution call.</div><div>2. Empowerment: (This take into account the hierarchical structure of the department) Allow representatives the authority to make decisions that can quickly resolve customer problems.</div><div>3. Resource Allocation: Ensure sufficient staffing during peak times.</div><div>4. Proactive Outreach: Proactively reach out to customer with potential issues (unusual usage behavior) to prevent forthcoming problems</div></div>	<div><div>• Resolving issues efficiently helps prevent escalation, reducing the need for multiple customer service calls and minimizing frustration.</div><div>• High-quality service also strengthens the company's reputation, bolstering its public image. Proactively addressing unresolved issues can stop minor problems from escalating into bigger ones, helping retain customers.</div><div>• Offering effective, reliable support fosters loyalty and encourages long-term relationships, which is crucial for maintaining customer retention.</div></div>
Promote Voice Mail Plan to Increase Customer Engagement and Reduce Churn	<div><div>1. Marketing Campaigns: Highlight the benefits of the Voice Mail Plan through targeted advertising and communications.</div><div>2. Creating incentives: Offer promotions like free trials or discounted rates for new subscriber.</div><div>3. Feature Enhancements: Add functionalities such as voicemail-to-text or personalized greetings, ...</div><div>4. User Education: Provide tutorials and support to help customers maximize the service.</div></div>	<div><div>• Active users of the Voice Mail Plan show higher satisfaction and lower churn rates, enhancing loyalty. Offering unique and valuable improved services sets the company apart from competitors, driving differentiation by UX.</div><div>• Encouraging customers to engage with additional services helps strengthens relationships and increases reliance on its offerings. Also creates upselling opportunities, as customers become more receptive to upgrades and extra services.</div></div>

Recommendation List

Medium Priority

Solution	Step-by-Step Approaches	Rationales
Implement Region-Specific Strategies to Reduce Churn in High-Risk States (NJ and TX)	<div><div>1. Local Market Research: Conduct analyses to identify specific issues such as network coverage gaps or service outages to enhance local service quality.</div><div>2. Competitive Analysis: Assess competitors' offerings in these states to understand customer preferences.</div><div>3. Customized Marketing Campaigns: Develop region-specific promotions highlighting improvements and personalized services.</div><div>4. Community Engagement: Participate in local events and sponsor community programs to build brand presence and loyalty, building deeper positive image with the potential cohorts</div></div>	<div><div>• Addressing regional needs and resolving state-specific issues, the company can reduce churn in high-risk areas. Tailored services that meet local demands enhance customer satisfaction and demonstrate a commitment to the community, which strengthens the company's market position and brand reputation.</div><div>• Offering services that compete with or surpass competitors' offerings helps retain customers who might otherwise switch.</div><div>• Engaging with the community fosters a positive image, further enhancing brand loyalty and strengthening customer relationships.</div></div>
Cater to Heavy Daytime Users with Tailored Plans	<div><div>1. Usage Analysis: Identify customers who make longer daytime calls and assess their needs.</div><div>2. Custom Plans: Create plans offering unlimited or increased minutes during daytime hours at reasonable costs.</div><div>3. Value Bundles: Include additional benefits such as priority network access or discounted rates on other services.</div><div>4. Feedback Loop: Gather feedback from these users to refine and improve the plans.</div></div>	<div><div>• Tailored plans help address dissatisfaction among heavy daytime users who may feel they aren't getting sufficient value. By offering unique, personalized plans, the company can gain a competitive edge, attracting new customers while retaining those seeking better options.</div><div>• Satisfying the needs of high-usage customers not only increases loyalty but also ensures a steady stream of revenue.</div><div>• Aligning offerings with customer usage patterns enhances perceived value, leading to greater satisfaction. Ultimately, meeting these specific needs plays a crucial role in reducing churn, making customers less likely to explore competitor options.</div></div>

Recommendation List

Medium Priority

Solution	Step-by-Step Approaches	Rationales
Implement Loyalty and Rewards Programs to Recognize and Retain Customers	<div><div>1. Points Systems: Allow customers to earn points for usage, on-time payments, or engagement, redeemable for discounts or gifts.</div><div>2. Tiered Memberships: Offer different levels of benefits (Offer premium or bonus services or upgrades at no additional cost ...) based on tenure or spending.</div><div>3. Exclusive Offers: Provide loyal customers with early access to new services or special promotions.</div><div>4. Personalized Rewards: Tailor rewards to individual customer preferences based on usage patterns and interests.</div><div>5. Predictive Analytics/AB Testing (Optional): Use existing data regarding a particular service to identify long-term customers showing signs of dissatisfaction. Also, applying AB Testing for offerings' modification to reassess change in behaviors before going live >> Reach out for special offerings and compensation</div></div>	<div><div>• Making customers feel valued, which significantly reduces the likelihood of churn. By offering incentives, customers are motivated to use more services and remain engaged with the company.</div><div>• Unique loyalty programs further differentiate the business from its competitors, setting it apart in the market. Engaged and loyal customers contribute to sustained revenue growth, not only through continued patronage but also through upselling opportunities.</div><div>• Long-term customers may feel neglected if they perceive that new customers receive better deals, so offering rewards helps prevent complacency. Demonstrating continued value and appreciation reinforces the customers' decision to stay, ensuring their loyalty over time.</div></div>

Recommendation List

Low Priority

Solution	Step-by-Step Approaches	Rationales
Reassess Pricing Strategies to Enhance Perceived Value and Fairness	<div><div>1. Price Sensitivity Testing: Conduct research to understand how price changes affect customer satisfaction and churn.</div><div>2. More Pricing Options: Introduce more varying plans where customers can select features that fit their budget and needs.</div><div>3. Transparent Billing: Simplify bills to clearly show charges, reducing confusion and mistrust.</div><div>4. Regular Plan Reviews: Periodically adjust pricing structures to remain competitive and fair, while keeping contact of customers for notifying</div></div>	<div><div>• Fair pricing enhances customer satisfaction and reduces churn by addressing overpricing concerns.</div><div>• Transparent pricing builds trust, and flexible options meet diverse needs.</div><div>• Regular reviews ensure competitiveness, preventing customer loss to competitors.</div></div>

Recommendation Caveats

1

FOR HIGH PRIORITY RECOMMENDATIONS

- CS training is attached with **long-term vision**, with **consistent investment**
- Suitable for **company with some establishment** in the industry, with minimal impact from external disruption

2

FOR MEDIUM PRIORITY RECOMMENDATIONS

- Attention to **coordination across departments** since adjusting service packages might call for strategy replanning, operational complexity and additional resources

3

FOR LOW PRIORITY RECOMMENDATIONS

- Involve **significant changes to pricing structures**, leading to customer confusion and potential dissatisfaction if not handled carefully.
- Research **entails locality**, which requires more **integral and comprehensive data collection** for analytics

APPENDICES

A. Preliminary Feature Engineering

I. Feature Generation

During EDA process, a number of new attributes displaying interactivity among other attributes have been generated (using logical function or binning, discretizing methods)

Attribute	Business Name	Logic
intl_charge_per_min	International Charge per Minute	total_intl_charge / total_intl_minutes
intl_charge_per_min_cate	International Charge Category	Categorized as 0 (≤ 0.25), 1 (0.25 - 0.35), 2 (> 0.35)
total_minutes	Total Minutes Used	total_day_minutes + total_eve_minutes + total_night_minutes
total_calls	Total Calls Made	total_day_calls + total_eve_calls + total_night_calls
total_charge	Total Charges Incurred	total_day_charge + total_eve_charge + total_night_charge
daytime_charge_per_min	Daytime Charge per Minute	total_day_charge / total_day_minutes
avg_charge_per_acc_day	Average Daily Charge per Account	(total_day_charge + total_eve_charge + total_night_charge) / account_length
voicemail_engagement_lvl	Voicemail Engagement Level	Categorized as 0 (0 messages), 1 (1-10 messages), 2 (11-20 messages), 3 (> 20 messages)
customer_service_freq_lvl	Customer Service Frequency Level	Categorized as 0 (0-3 calls), 1 (4-6 calls), 2 (> 6 calls)

APPENDICES

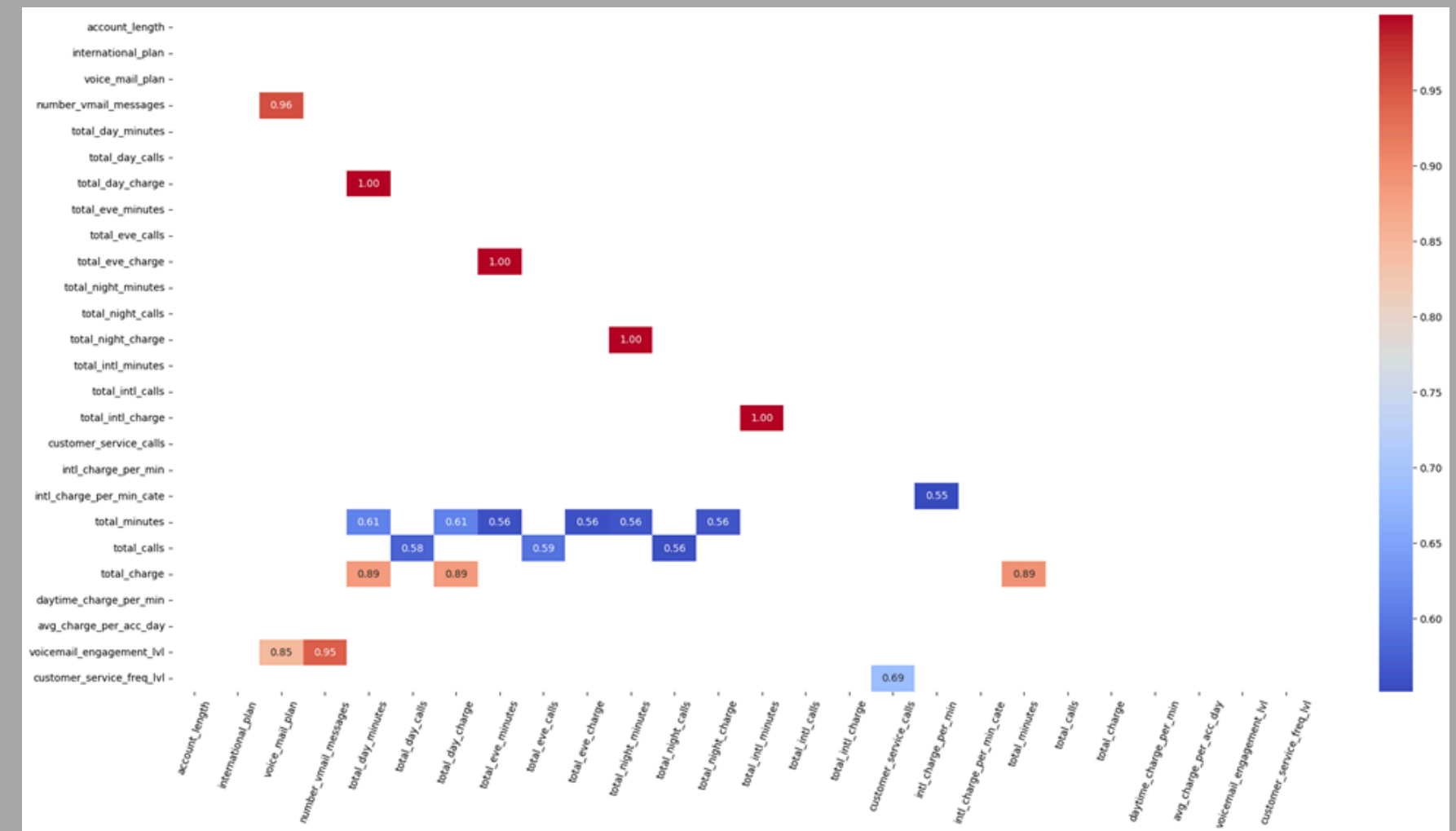
A. Preliminary Feature Engineering

II. Preliminary Feature Evaluation & Selection

1. Tackling Multicollinearity

Attributes with noticeably high correlational value with others might be dropped

- **voice_mail_plan** and **number_vmail_messages** together generate **voicemail_engagement_lvl**. Given that **voice_mail_plan** displays a discrete Yes/No pattern of voicemail usage, the **number_vmail_messages** will be dropped first while the **voice_mail_plan** can be assessed later via the Information Value Model.
- The same is also true for **customer_service_calls** and **customer_service_freq_lvl**, however, the multicollinearity is less severe for these two.
- **total_charge** will be dropped since it displays high multicollinearity with **total_day_minutes**, **total_day_charge**, and **total_minutes**.



APPENDICES

A. Preliminary Feature Engineering

II. Preliminary Feature Evaluation & Selection

2. Information Value Model

- **State** and **area code** provide little value to the prediction and will likely be dropped unless they show interactive influence on the target attribute.
- **customer_service_calls** is crucial for the model, while **voice_mail_plan** is a weak predictor and may be dropped after further testing.
- I will monitor **intl_charge_per_min_cate**, **total_eve_calls**, and **account_length**, as they currently appear to be unuseful predictors.

** Note: For better attributes' influence judgement, I would use **2 versions of dataset**, one with all attributes as aforementioned and the other with a more “refined” subset as follows:

Information Value	Predictive Power
< 0.02	Not Useful
0.02 to 0.1	Weak Predictor
0.1 to 0.3	Medium Predictor
0.3 to 0.5	Strong Predictor
> 0.5	Suspicious

	Variable	IV
0	total_day_minutes	0.655953
1	total_day_charge	0.655953
2	customer_service_calls	0.563057
3	total_minutes	0.535617
4	customer_service_freq_lvl	0.535085
5	international_plan	0.429781
6	total_eve_charge	0.094372
7	total_eve_minutes	0.093106
8	total_intl_minutes	0.081906
9	total_intl_charge	0.081906
10	total_intl_calls	0.080006

	Variable	IV
10	total_intl_calls	0.080006
11	voicemail_engagement_lvl	0.079856
12	daytime_charge_per_min	0.063135
13	intl_charge_per_min	0.056950
14	total_night_charge	0.041289
15	total_night_minutes	0.039182
16	avg_charge_per_acc_day	0.039180
17	total_night_calls	0.030197
18	total_day_calls	0.029164
19	total_calls	0.024541

APPENDICES

B. Model Construction

II. Why LightGBM and XGBoost are Shortlisted?

- 1. Scalability and Robustness for Large Datasets:** As the customer base in the telecommunications industry grows, datasets may scale from medium to large. LightGBM and XGBoost are robust algorithms designed to handle large-scale data efficiently, making them suitable for this context. Also, these 2 models are efficient in handling high-dimensional datasets
- 2. Effective Handling of Class Imbalance:** The target variable `churn` exhibits significant class imbalance. Both models offer parameters like `scale_pos_weight` (XGBoost) and `is_unbalance` or `scale_pos_weight` (LightGBM) to address this issue, minimizing bias toward the majority class.
- 3. Prevention of Overfitting through Regularization:** Built-in regularization techniques in both models help prevent overfitting, enhancing the model's ability to generalize to new data.
- 4. Capturing Complex Nonlinear Relationships:** As powerful gradient boosting algorithms, they excel at modeling complex nonlinear patterns in data, often outperforming other classification methods.
- 5. Efficient Training with Parallel and Distributed Computing:** Support for parallel computing and distributed training accelerates the training process, making them efficient for large datasets.
- 6. Insightful Feature Importance Scores:** Both models provide feature importance metrics, helping identify the most influential factors contributing to customer churn, which is valuable for strategic decision-making.

APPENDICES

B. Model Construction

III. Comparison & Final Model Selection

Dataset Ver 01

	Model	Accuracy	ROC AUC
0	XGBoost	0.958021	0.933382
1	LightGBM	0.959520	0.933346

Dataset Ver 02

	Model	Accuracy	ROC AUC
0	XGBoost	0.938531	0.918421
1	LightGBM	0.952024	0.925653

Dataset Version 1 vs. Dataset Version 2:

Version 1 (with more features) slightly outperformed Version 2, particularly with XGBoost. LightGBM, however, produced similar results across both versions, showing consistency.

Model Comparison (XGBoost vs. LightGBM):

LightGBM outperformed XGBoost in efficiency and accuracy, especially with Version 2. LightGBM also required less training time, making it more scalable for larger datasets.

LightGBM was **more consistent** and **better at capturing complex relationships**, making it the preferred model. LightGBM is officially chosen as the final model.

Note: ROC AUC was chosen as the primary evaluation metric due to its reliability for imbalanced datasets.

- As expected along the EDA process as well as with the **Information Value Model**, **geographical values bare little to almost no influence** at all on the target variables. Apart from that, ``voice_mail_plan`` and ``intl_charge_per_min`` also have little impact, which can be officially dropped for the final model before deployment.
- Otherwise, ``total_eve_calls`` and ``account_length`` do have moderate positive impact on the model, which mean they can be officially kept