Telecom Churn Prediction

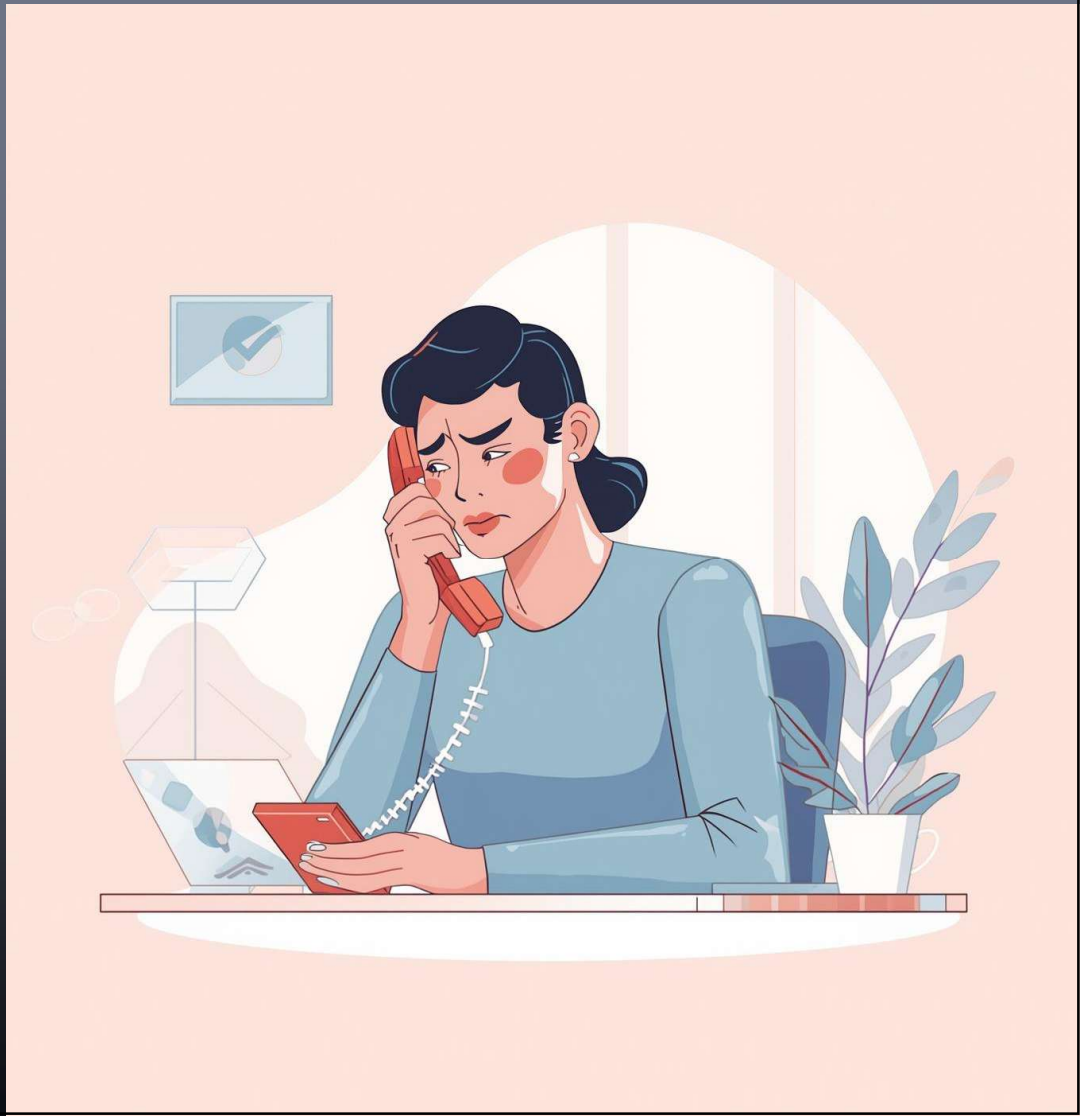
Predicting customer churn using machine learning

Presented by:

Vrushali Oak
Data Scientist

Understanding Churn

Customer churn refers where clients discontinue their service.
Understanding churn is essential as it impacts revenue and growth. Identifying the reasons behind customer dissatisfaction helps businesses **improve services** and develop effective retention strategies.



Business Problem

- Customer churn leads to significant revenue loss in telecom
- Retaining existing customers is cheaper than acquiring new ones
- Objective: Predict customers likely to churn so retention teams can act early

Dataset Overview

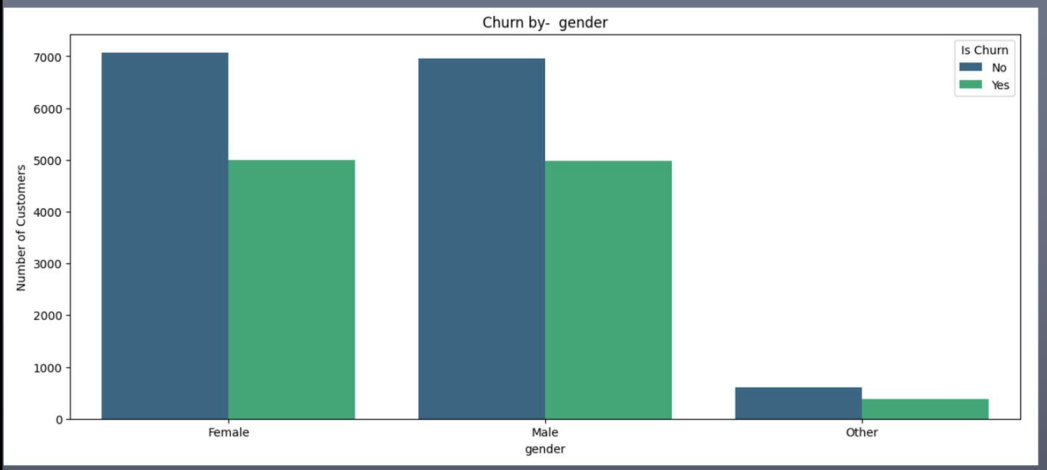
- 25,000 customers, 36 features
- Includes numerical, binary, and categorical variables
- Target variable: is_churn (1 = churned, 0 = retained)
- Churn distribution:
  - 58.6% retained
  - 41.4% churned
  - This is moderate imbalance, not extreme

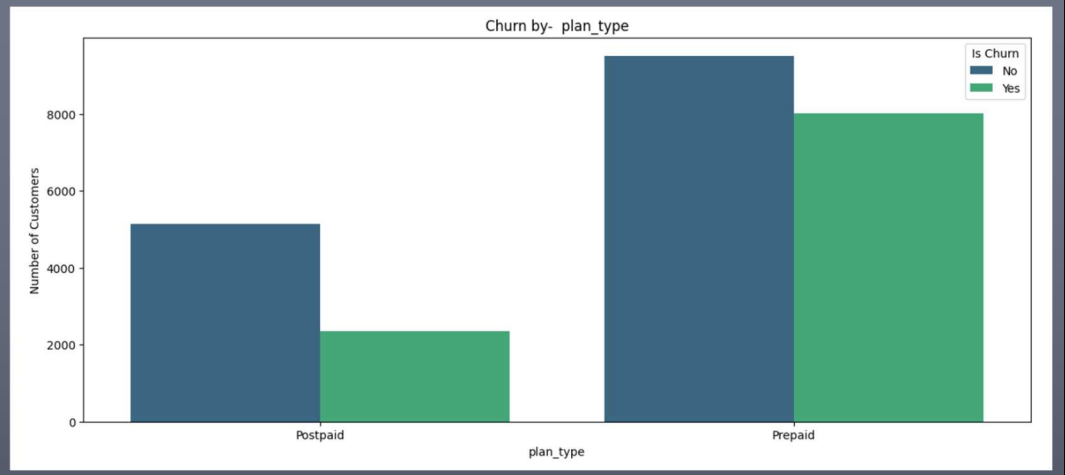
Data Understanding & Cleaning

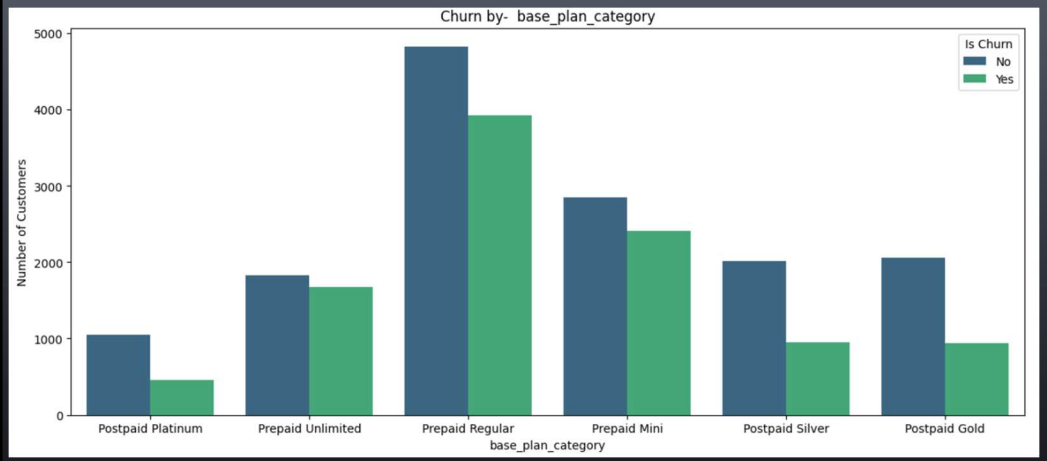
- Removed non-informative and leaky columns:
  - customer_id (identifier)
- Corrected data types
- Verified missing values and distributions- **No missing values** and **no duplicate records** after validation
- Separated numerical, categorical and binary features

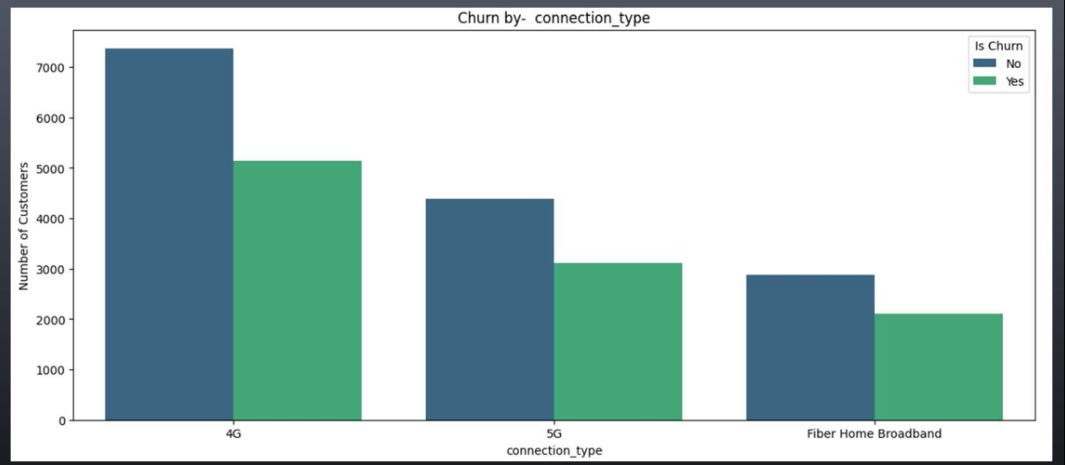
Exploratory Data Analysis (EDA)

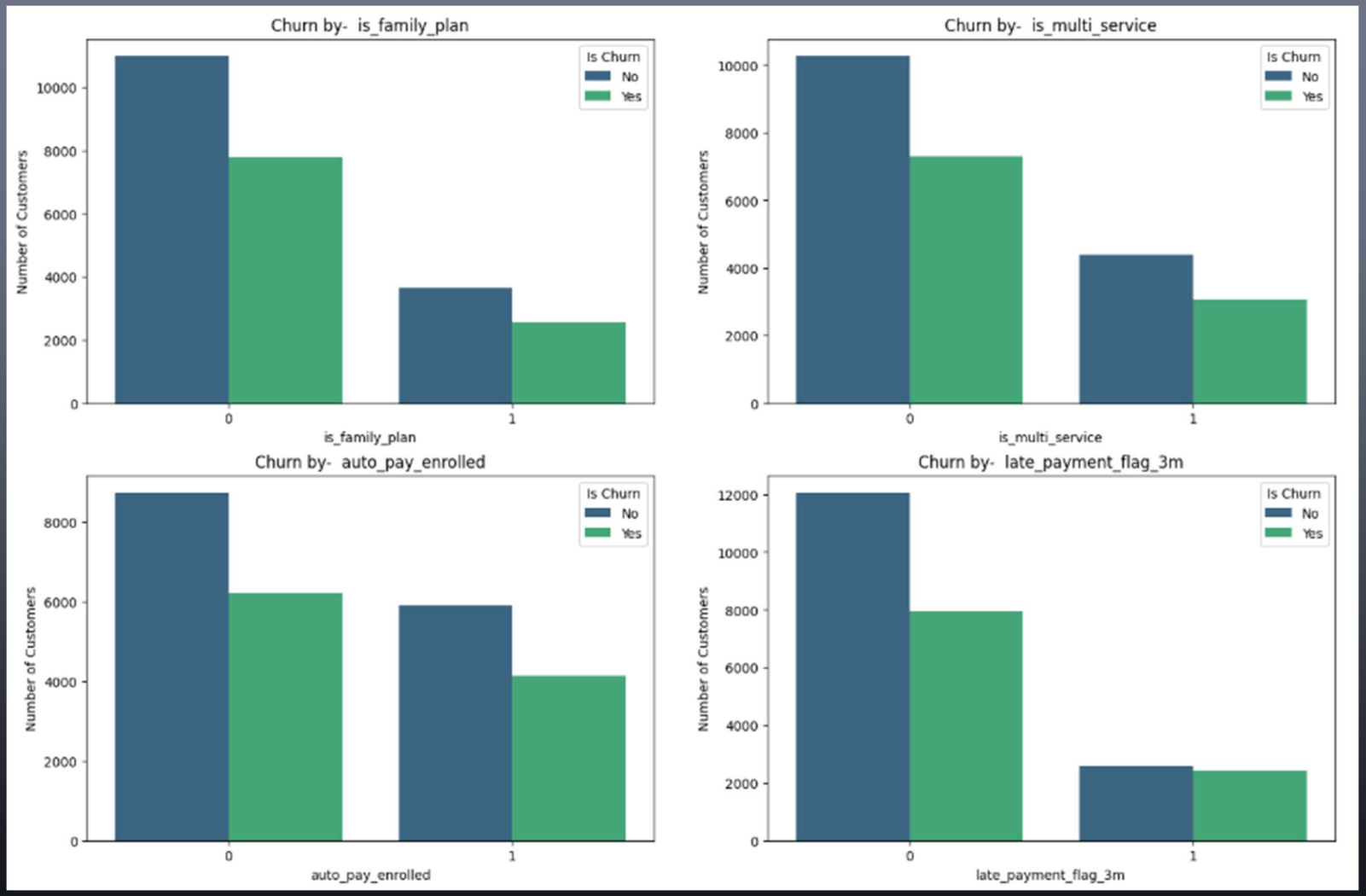
- Performed correlation heatmap on continuous numeric features only
- Identified:
  High correlation between monthly_charges and arpu
  Redundant pricing variables
- Droped columns -
  'monthly_charges',
  'num_complaints_3m',
  'total_charges'
- Used
  Bar plots for- Binary and Categorical features
  Box plots for- Numerical features

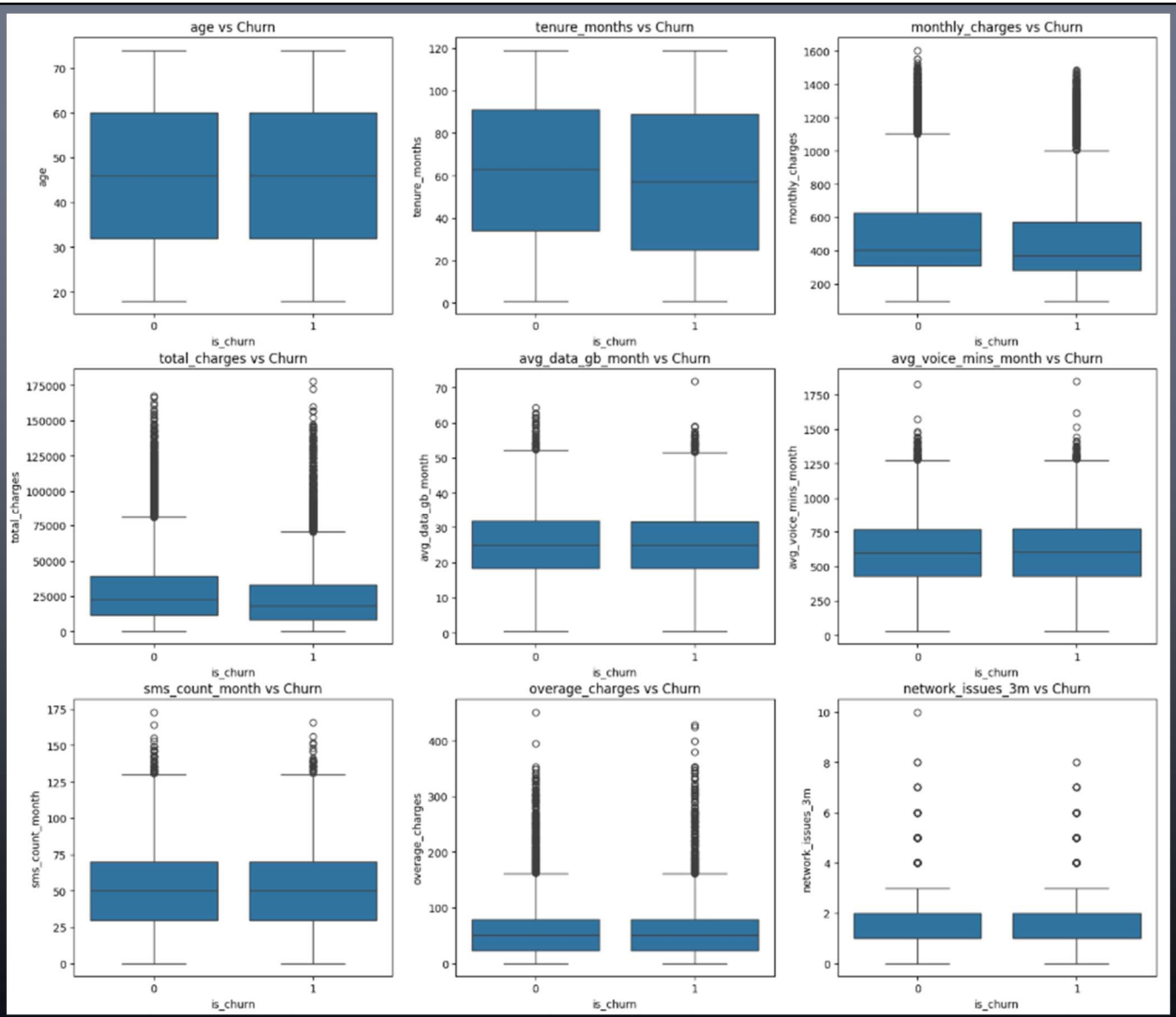

Churn by- gender
Is Churn
No
Yes
Number of Customers
0
1000
2000
3000
4000
5000
6000
7000
Female
Male
Other
gender

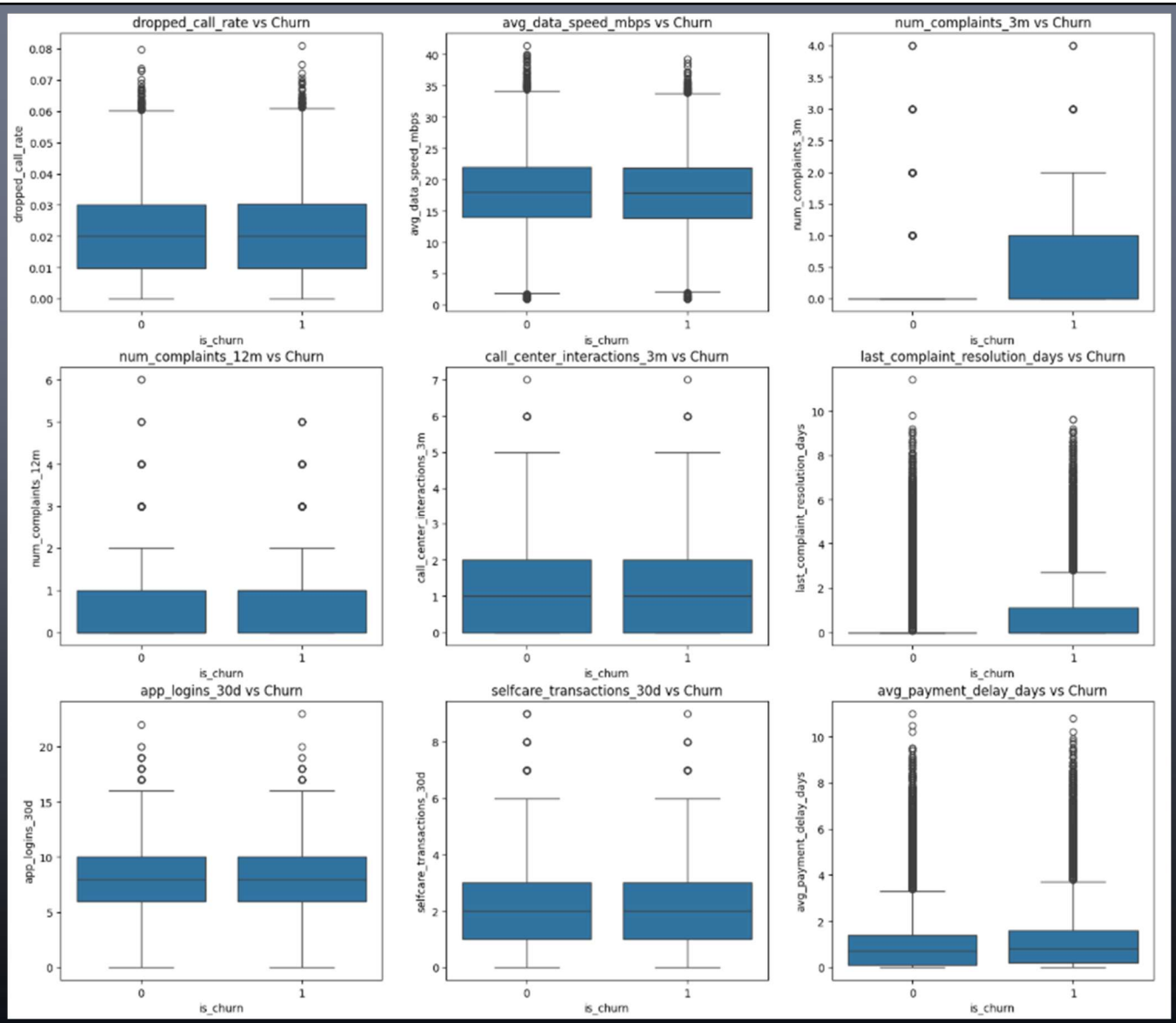

Churn by- plan_type
Is Churn
No
Yes
Number of Customers
0
2000
4000
6000
8000
Postpaid
Prepaid
plan_type


Churn by- base_plan_category
Is Churn
No
Yes
Number of Customers
0
1000
2000
3000
4000
5000
Postpaid Platinum
Prepaid Unlimited
Prepaid Regular
Prepaid Mini
Postpaid Silver
Postpaid Gold
base_plan_category


Churn by- connection_type
Is Churn
No
Yes
Number of Customers
0
1000
2000
3000
4000
5000
6000
7000
4G
5G
Fiber Home Broadband
connection_type


Churn by- is_family_plan
Is Churn
No
Yes
Number of Customers
0
2000
4000
6000
8000
10000
0
1
is_family_plan
Churn by- is_multi_service
Is Churn
No
Yes
Number of Customers
0
2000
4000
6000
8000
10000
0
1
is_multi_service
Churn by- auto_pay_enrolled
Is Churn
No
Yes
Number of Customers
0
2000
4000
6000
8000
0
1
auto_pay_enrolled
Churn by- late_payment_flag_3m
Is Churn
No
Yes
Number of Customers
0
2000
4000
6000
8000
10000
12000
0
1
late_payment_flag_3m


age vs Churn
age
is_churn
tenure_months vs Churn
tenure_months
is_churn
monthly_charges vs Churn
monthly_charges
is_churn
total_charges vs Churn
total_charges
is_churn
avg_data_gb_month vs Churn
avg_data_gb_month
is_churn
avg_voice_mins_month vs Churn
avg_voice_mins_month
is_churn
sms_count_month vs Churn
sms_count_month
is_churn
overage_charges vs Churn
overage_charges
is_churn
network_issues_3m vs Churn
network_issues_3m
is_churn


dropped_call_rate vs Churn
dropped_call_rate
is_churn
avg_data_speed_mbps vs Churn
avg_data_speed_mbps
is_churn
num_complaints_3m vs Churn
num_complaints_3m
is_churn
num_complaints_12m vs Churn
num_complaints_12m
is_churn
call_center_interactions_3m vs Churn
call_center_interactions_3m
is_churn
last_complaint_resolution_days vs Churn
last_complaint_resolution_days
is_churn
app_logins_30d vs Churn
app_logins_30d
is_churn
selfcare_transactions_30d vs Churn
selfcare_transactions_30d
is_churn
avg_payment_delay_days vs Churn
avg_payment_delay_days
is_churn

Key Insights from EDA

- High overage charges → higher churn probability
- Frequent complaints and call center interactions → churn risk
- Payment delays strongly associated with churn
- Service quality issues (low speed, dropped calls) impact churn

Feature Selection Decisions

- Dropped ‘monthly_charges’ due to high correlation with arpu
- Kept arpu as it reflects actual realized customer cost
- Dropped cumulative and leakage-prone features (monthly_charges', 'num_complaints_3m')
- Retained behavior-driven features impacting churn

Feature Engineering

- Created High Overage Risk Flag to capture billing shock behavior
- Engineered behavioral signals instead of raw extremes
- Focused on business meaning, not just statistical transformation

Outlier Handling Strategy

- Did NOT blindly remove outliers
- Applied soft capping only where outliers likely represent noise:
  - Billing spikes
  - Measurement errors
- Preserved extreme values where they represent real churn behavior

Data Preparation

- One-hot encoded categorical variables
- Feature scaling applied for Logistic Regression
- Stratified train-test split to maintain churn ratio

Models Used

- Logistic Regression (baseline & interpretable)
- Random Forest (non-linear, ensemble)
- XGBoost (gradient boosting)
- Purpose:
  Compare simplicity vs performance vs generalization

Evaluation Metrics

- ROC-AUC (primary metric – ranking churn risk)
- Precision, Recall, F1-score
- Train vs Test performance to detect overfitting

Model Performance

	Model	ROC-AUC
0	Logistic Regression (unscaled)	0.647326
1	Logistic Regression (scaled)	0.647234
2	Random Forest	0.647614
3	XGBoost (baseline)	0.635404
4	XGBoost (tuned)	0.654467

Model Performance Comparison

- Logistic Regression: stable but limited non-linear learning
- Random Forest: better than LR, moderate generalization
- XGBoost: highest CV and test ROC-AUC

XGBoost performed best in ranking high-risk churn customers

Why XGBoost Performed Best

- Captures non-linear churn behavior
- Learns feature interactions automatically
- Focuses on hard-to-predict churn cases
- Handles mixed feature types and imbalance effectively

Final Model Selection

- Selected XGBoost as primary model
- Logistic Regression retained as interpretable baseline
- Acknowledged mild overfitting and discussed regularization improvements

Business Impact

- Customers ranked by churn probability
- Enables targeted retention campaigns
- Improves marketing efficiency
- Reduces revenue leakage due to churn

Conclusion

- Built an end-to-end churn prediction pipeline
- Combined business understanding with ML modeling
- Delivered a practical, deployable churn scoring solution