Customer Churn Prediction Challenge Submission

By Group 1

Presentation Flow

- Objective & Methodology
- Data Overview & Preprocessing
- Exploratory Analysis
- Modeling & Evaluation
- Feature Importance
- Takeaways

Objective & Methodology

Goal: Analyze the key factors contributing to customer churn and build prediction models to predict customer churn using KNIME workflow

Methodology:

1. Data Ingestion, Cleaning & Preprocessing
2. Exploratory Data Analysis to understand the key factors affecting customer churn
3. Predictive Modeling
4. Insight Generation & Recommendations

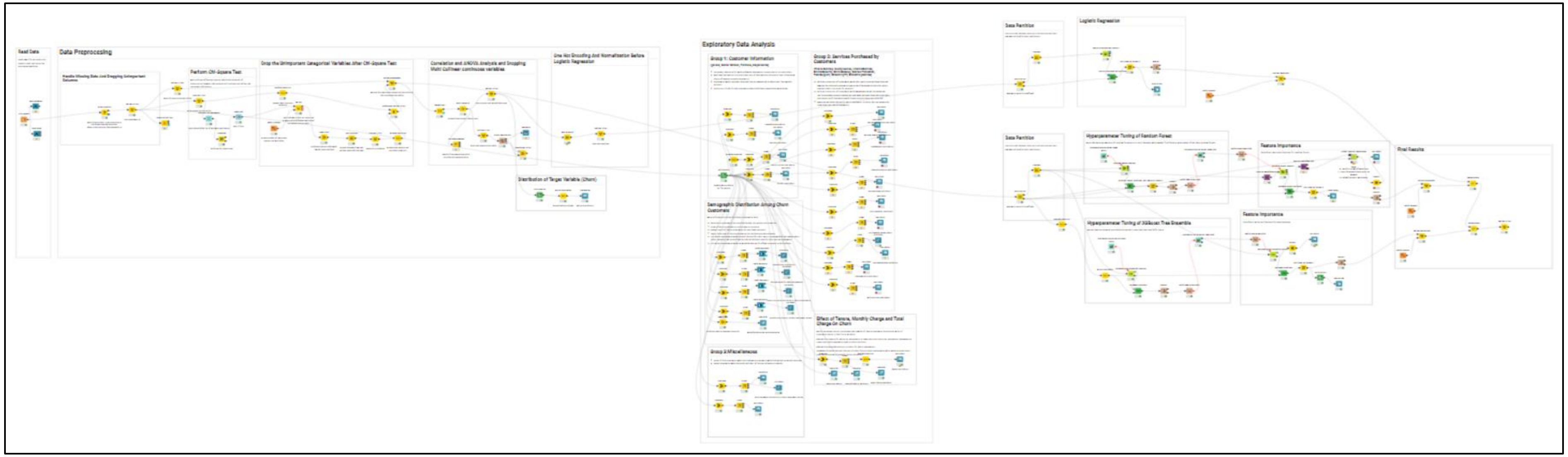


Figure 1: Snapshot of the full KNIME workflow broken down highlighting the process flow

Data Overview & Preprocessing

Data Overview:

1. The dataset contained 7,043 customer records
2. The target variable was Churn (Yes/No)
3. Features included:
    1. Customer demographics (e.g., gender, senior citizen)
    2. Service details (e.g., internet, phone services)
    3. Financial metrics (e.g., monthly & total charges)
    4. Contract information (e.g., contract type, payment method)

Data Preprocessing:

1. 11 records had null values in TotalCharges and were removed
2. The customerID column was excluded using the Column Filter node
3. One-hot encoding was applied to all categorical variables
4. Min-Max normalization was used to scale numerical features
5. A new feature was engineered: tenure_to_total_charges_ratio = tenure / TotalCharges

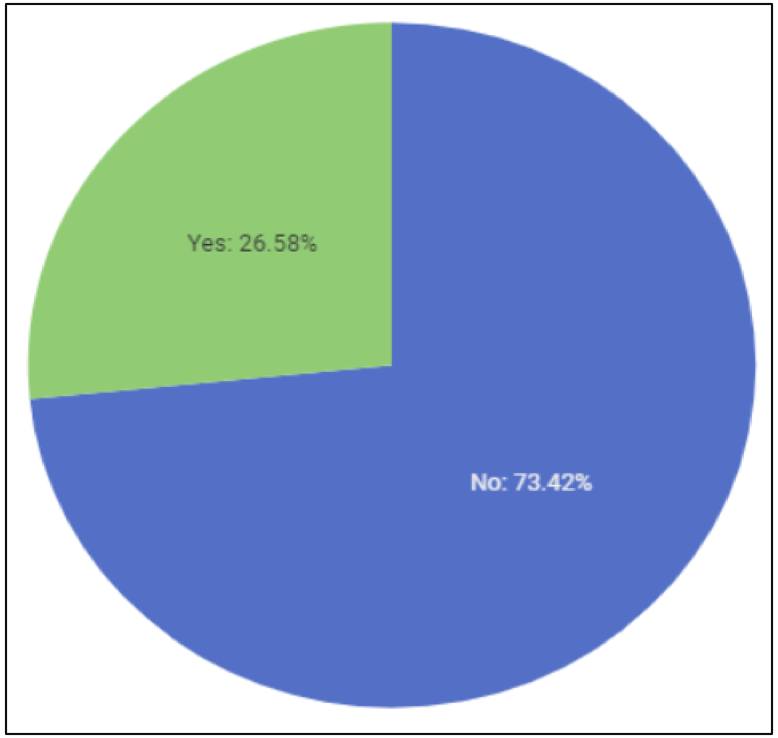


Figure 2: In the final cleaned dataset, 26.58% rows had churn values as 'Yes' while 73.42% rows had churn values as 'No'

Exploratory Analysis (1/2)

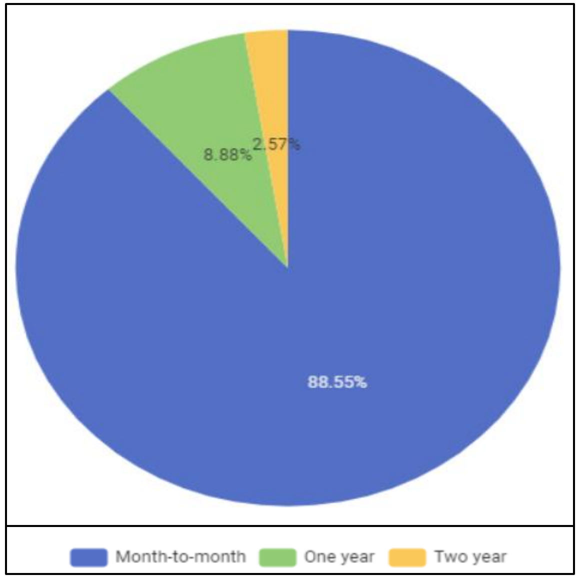


Figure 3: Break-up of churn customers contract type-wise

Key finding: Over 88% of churned customers were on month-to-month contracts, suggesting that customers with longer-term commitments are more loyal

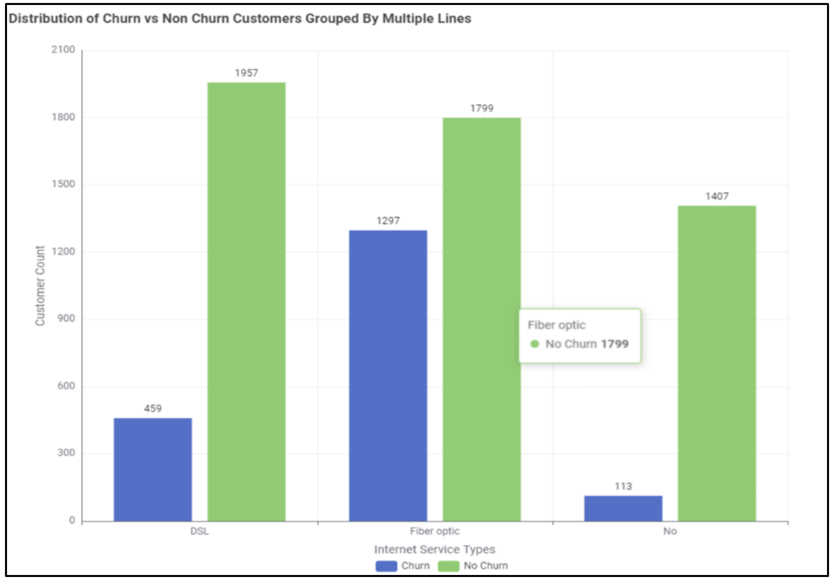
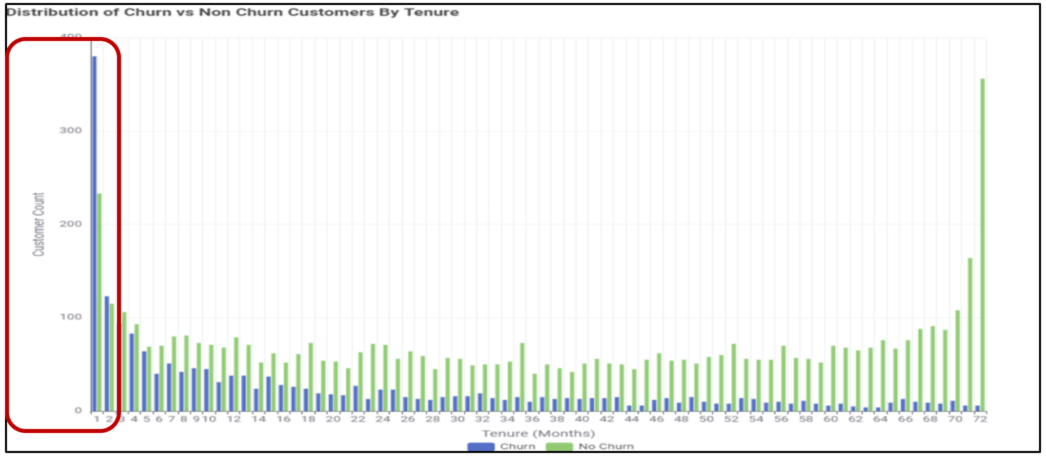


Figure 4: Break-up of churn customers internet service type-wise

Key finding: Churn is significantly higher among customers using Fiber Optic internet service. This may indicate customer dissatisfaction related to speed, stability, or cost of the service.



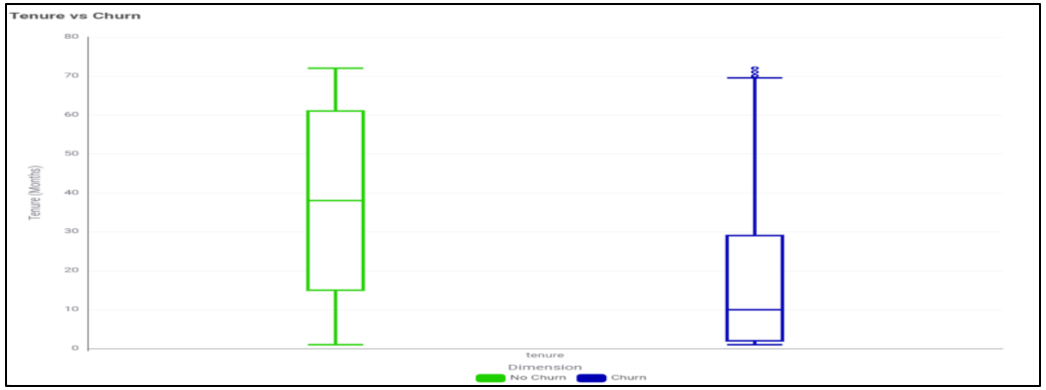


Figure 5 & 6: Bar chart and Box plot representing the tenure for churn and non-churn customers

Key finding: Both the bar chart and box plot confirm that churned customers generally have shorter tenure. Newer customers are more likely to leave, highlighting a critical onboarding period.

Exploratory Analysis (2/2)

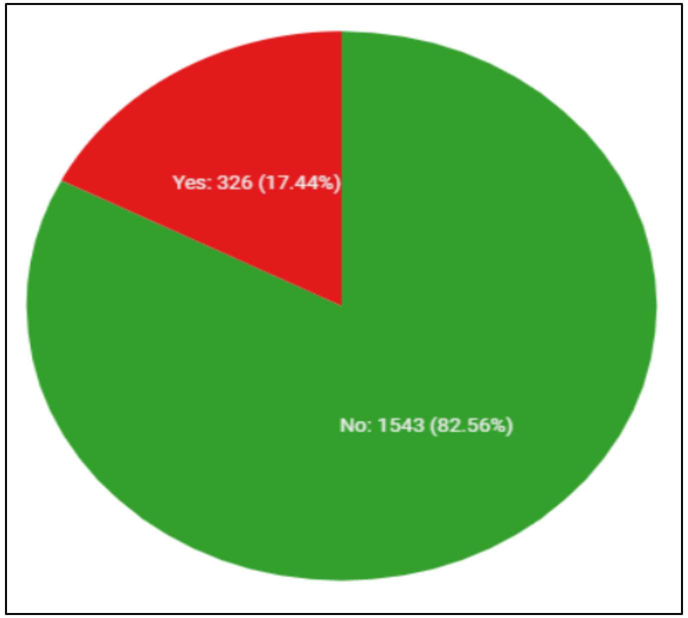


Figure 7: Break-up of churn customers dependent-wise

Key finding: A majority of churned customers have no dependents. Customers with fewer personal obligations may exhibit lower attachment or usage frequency, increasing churn risk.

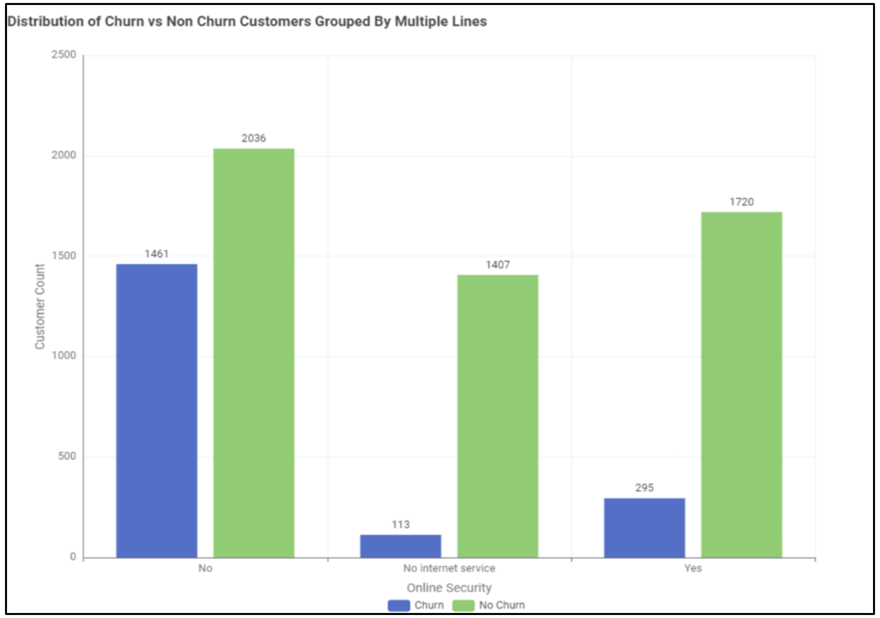


Figure 8: Break-up of customers by number of lines

Key finding: Customers with no additional lines are more likely to churn. This aligns with the dependent insight: fewer household members = fewer services = weaker customer stickiness.

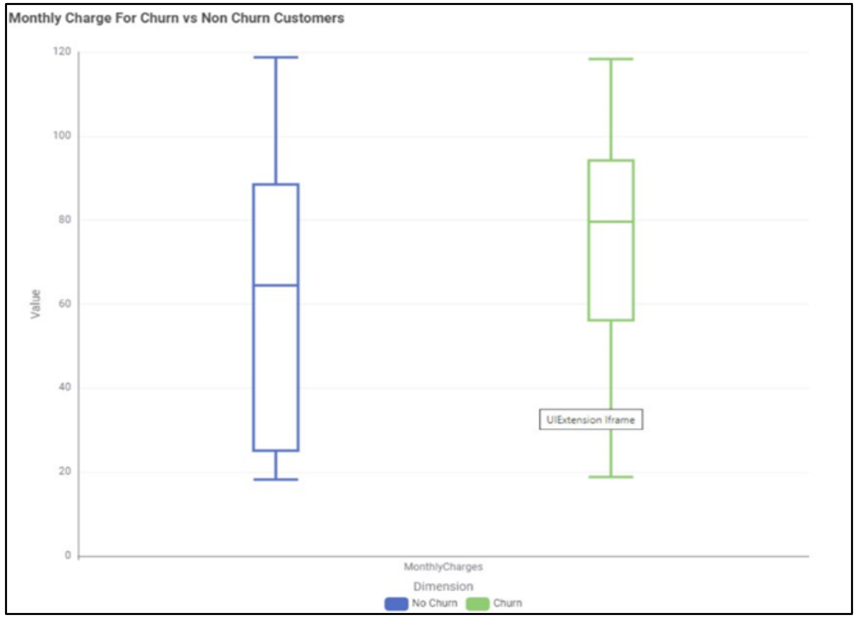


Figure 9: Box plot for monthly charges for customers

Key finding: Churned customers tend to incur higher monthly charges compared to those who stayed. This may indicate price sensitivity, where higher billing leads to dissatisfaction or switching behavior.

Modeling & Evaluation

Objective: Focus on maximizing the F1 Score to balance recall (catching as many churners as possible) and precision (minimizing false positives)

Models Evaluated:

1. Logistic Regression Learner
2. Random Forest Learner
3. XGBoost Tree Ensemble Learner

Evaluation Criteria:

Metric	Why it Matters
F1 Score	Balances recall & precision – our key goal
Accuracy	Overall correctness of predictions
Precision	Reduces false positives

Based on our evaluation, Logistic Regression produced the highest F1 Score, making it the preferred model for deployment in a churn prediction scenario.

Rows: 9 | Columns: 5

Table Statistics

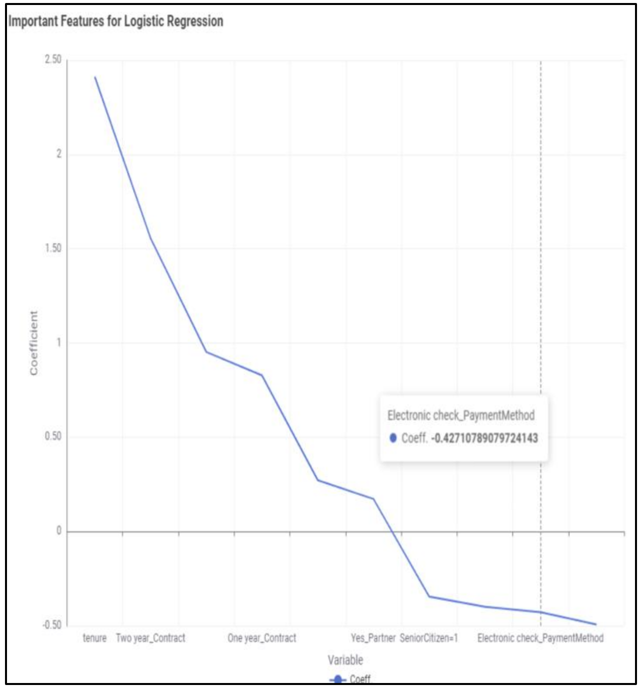
#	RowID	Recall Number (double)	Precision Number (double)	F-measure Number (double)	Accuracy Number (double)	Model String
1	Row0	0.503	0.797	0.616	?	Logistic Regression
2	Row1	0.907	0.714	0.799	?	Logistic Regression
3	Row2	?	?	?	0.736	Logistic Regression
4	Row3	0.481	0.807	0.603	?	Random Forest
5	Row4	0.908	0.684	0.78	?	Random Forest
6	Row5	?	?	?	0.717	Random Forest
7	Row6	0.499	0.727	0.592	?	XGBoost Tree
8	Row7	0.882	0.736	0.802	?	XGBoost Tree
9	Row8	?	?	?	0.733	XGBoost Tree

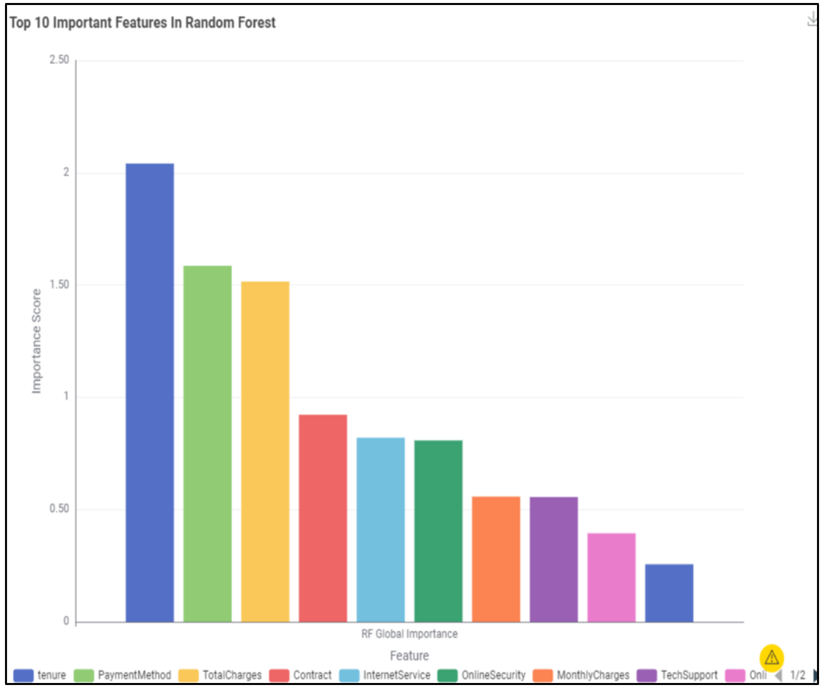
Figure 10: Table comparing the 3 models on 3 metrics

Feature Importance

Key features as obtained from all the models

- Month-to-Month Contracts → Highest churn risk
- Short Tenure → Newer customers more likely to leave
- Electronic Check Payments → Strong churn indicator
- Fiber Optic Internet → Associated with higher churn
- Lack of Tech Support / Online Security → Increases churn
- Higher Monthly Charges → Price-sensitive churners





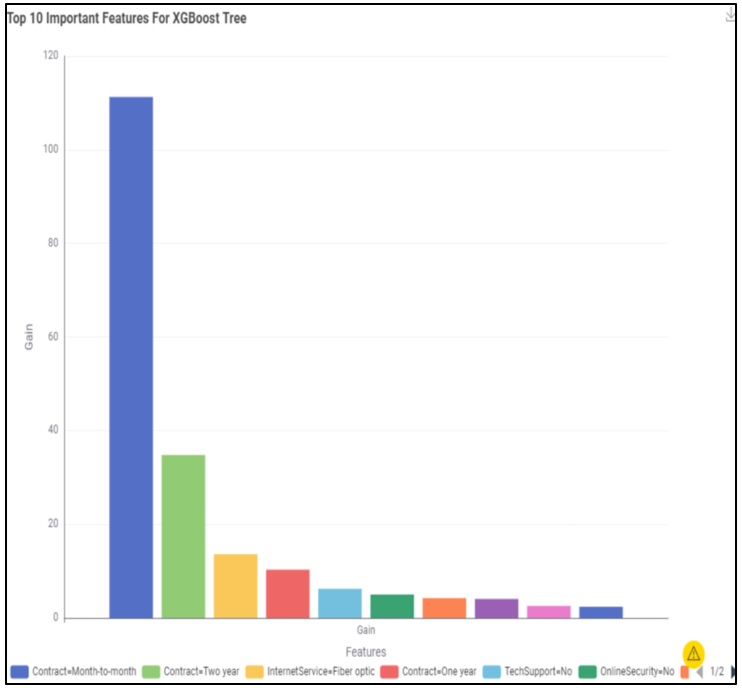


Figure 11,12,13: Top features from the 3 models

Takeaways

Key Finding	Suggestion to Improve Retention
Majority of churners are on month-to-month contracts	Promote annual contracts with loyalty discounts or bonus benefits
Churners typically have low tenure	Focus on early onboarding campaigns during the first 3 months
Electronic check is the most churn-prone payment method	Encourage AutoPay methods (bank transfer/credit card) via incentives
Fiber optic users show higher churn	Investigate service quality or pricing issues and improve value
Lack of Tech Support and Online Security linked to churn	Bundle or upsell support features to increase perceived value
Churners have higher monthly charges	Offer bundled plans or personalized pricing offers for high-cost users
Churners tend to have no dependents or extra lines	Create family/multi-line incentives to deepen customer engagement

Thank You