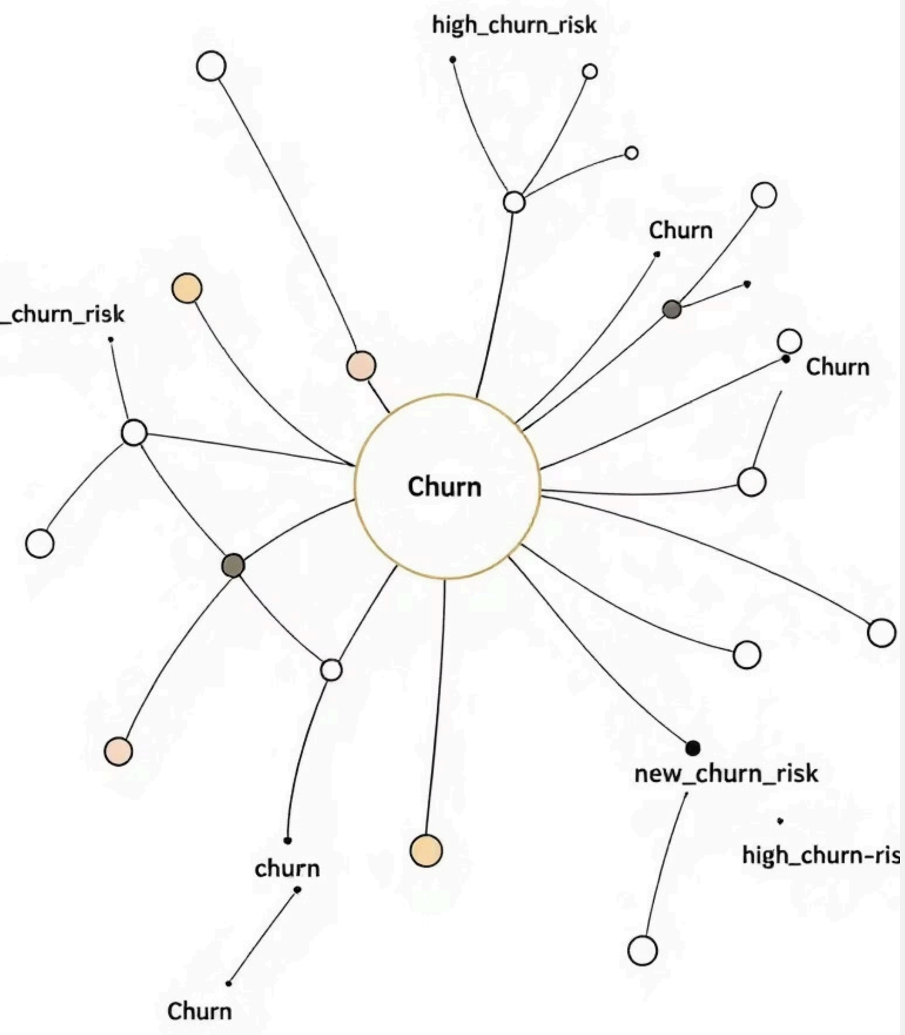


Feature Engineering Rationale

A new feature, **high_churn_risk**, was engineered based on the strong inverse correlation between MonthlyCharges and tenure.

- *Calculation:* high_churn_risk = MonthlyCharges / tenure.

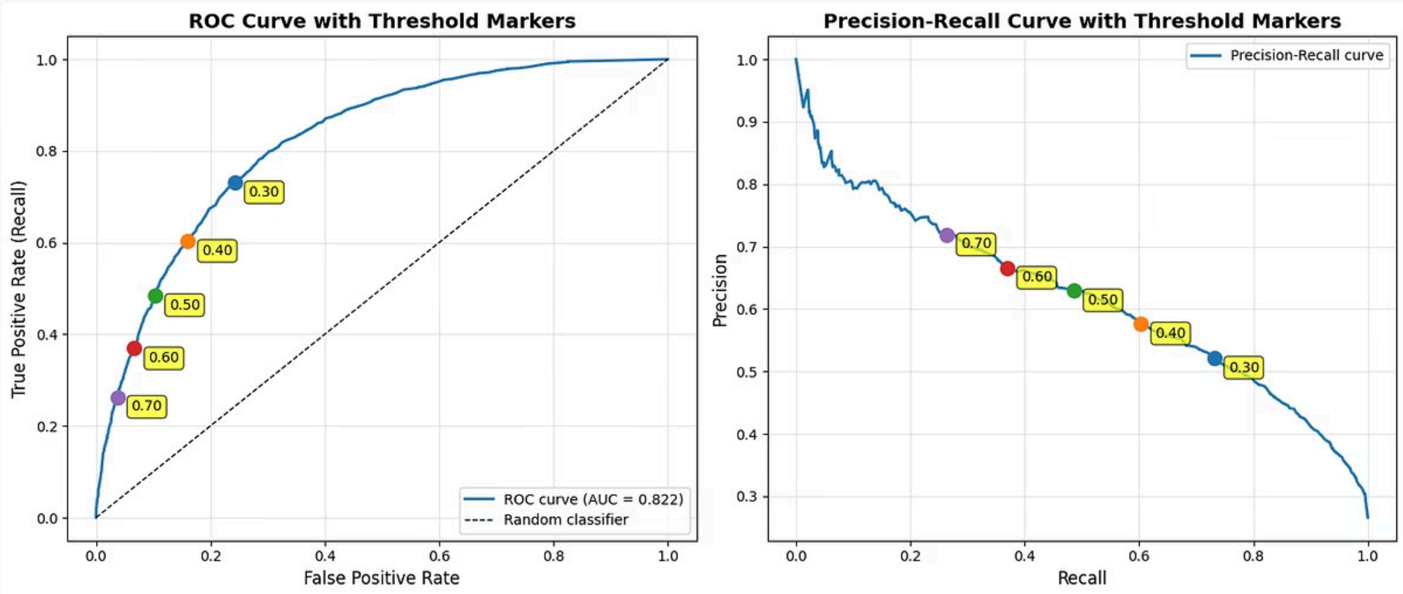
This ratio acts as a **"risk measurement,"** yielding a high value when the customer has high monthly charges but low tenure.

The high_churn_risk feature achieved the highest correlation with the target variable (Churn) at **0.386**.

Feature Selection

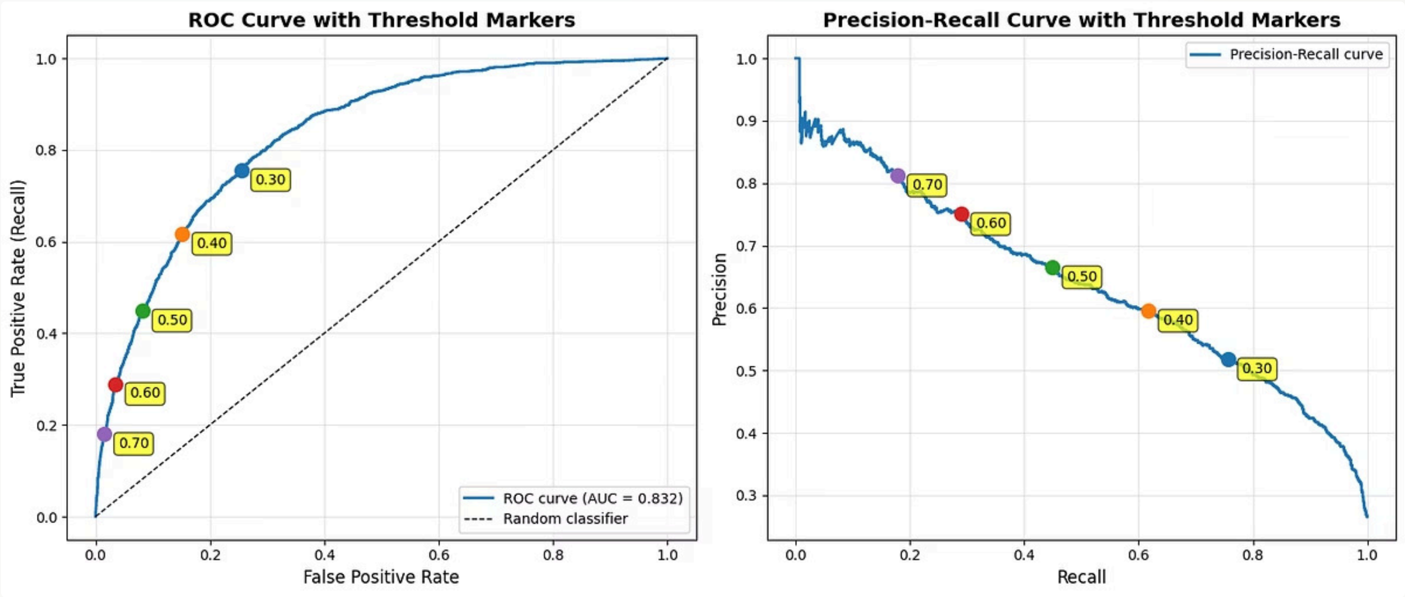
- **Engineered Features Used:** high_churn_risk.
- **Features Removed:**
  - MonthlyCharges and tenure were dropped as their combined information is captured in high_churn_risk.
  - Features with very low correlation to Churn were dropped: PhoneService, gender, and MultipleLines.

	Churn
Churn	1.000000
high_churn_risk	0.386065
MonthlyCharges	0.192858
PaperlessBilling	0.191454
SeniorCitizen	0.150541
PaymentMethod	0.107852
MultipleLines	0.038043
TotalCharges	0.012891
PhoneService	0.011691
gender	-0.008545
StreamingTV	-0.036303
StreamingMovies	-0.038802
InternetService	-0.047097
Partner	-0.149982
Dependents	-0.163128
DeviceProtection	-0.177883
OnlineBackup	-0.195290
TechSupport	-0.282232
OnlineSecurity	-0.289050
tenure	-0.354049
Contract	-0.396150



Algorithm Selection Justification (Random Forest)

- **Model Tested 1:** Random Forest Classifier.
- **num of estimators:** >10k samples so n_estimators=200 was chosen.
- **Evaluation:** 5-fold *Stratified* Cross-Validation was used to generate probability scores (churn_scores).
- **Performance Summary :**
  - **ROC AUC:** 0.822.
  - At TH=0.40: Recall = 0.604, Precision = 0.577, F1 = 0.590.



Algorithm Selection Justification (Logistic Regression)

- **Model Tested 2:** Logistic Regression (max_iter=1000).
- **Evaluation:** 5-fold Stratified Cross-Validation.
- **Performance Summary :**
  - **ROC AUC:** 0.832.
  - At TH=0.40: Recall = 0.617, Precision = 0.595, F1 = 0.606.
- **Justification:** Logistic Regression was selected for the final financial analysis due to its slightly higher AUC (0.832 vs 0.822).

Model Performance Metrics (Threshold Analysis)

Since the target curn value is imbalced and the positive values are more rare, the **Precision-Recall Curve** is crucial.

The shape of the curve shows a trade-off: as Recall increases (moving left along the X-axis), Precision drops.

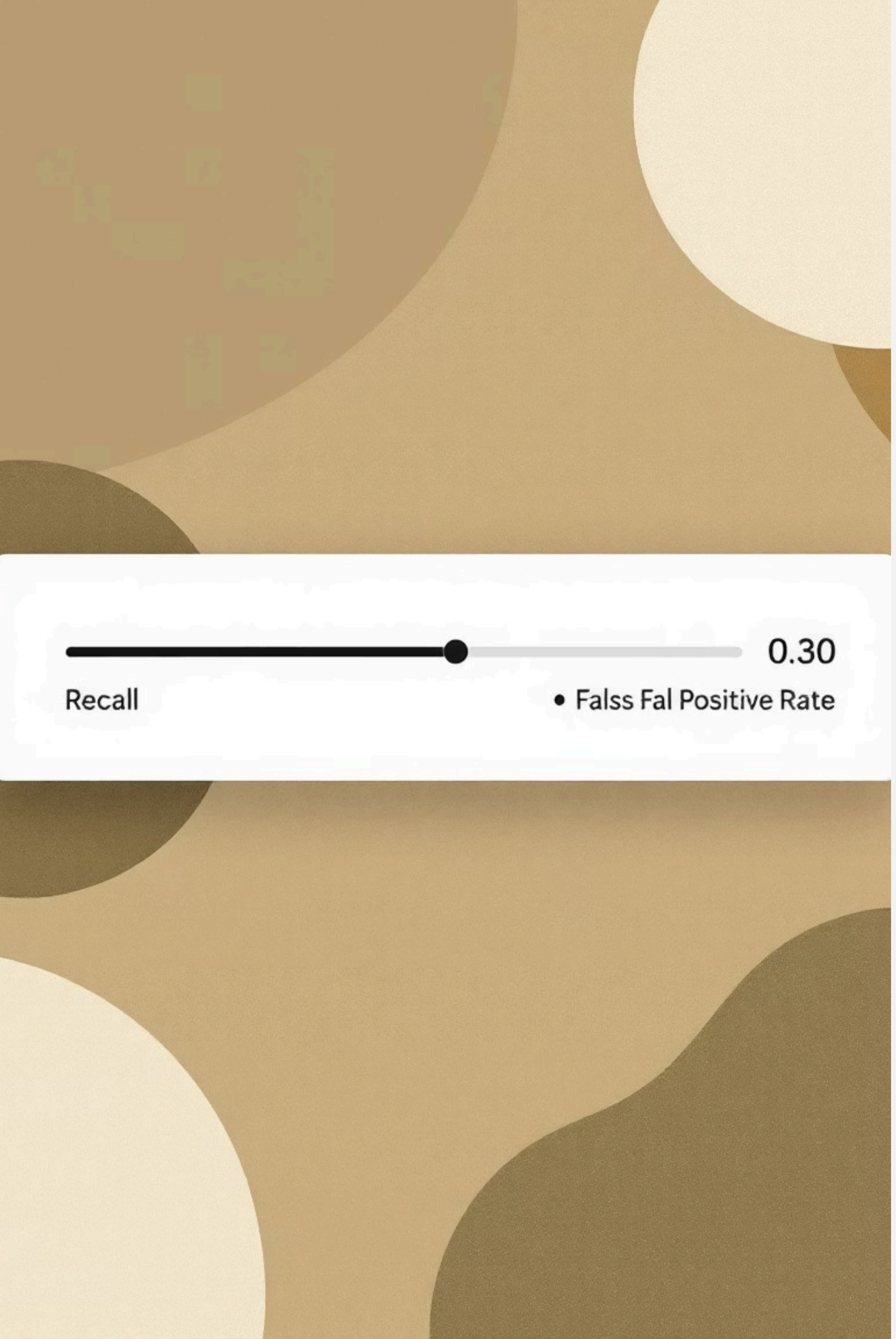
Example Metrics at various thresholds (Logistic Regression):

Threshold	Recall	Precision	FPR	F1
0.30	0.756	0.518	0.255	0.615
0.50	0.449	0.665	0.082	0.536

Optimization for Business Value (The Value Score)

- **Cost/Benefit Definition:**
  - Net Value per Saved Customer: 4,100.30$.
  - Cost per False Positive (Wasted Discount): 300$.
- **Optimization Function:** Value Score = (Recall × 4,100.30$) - (FPR × 300$).
- **Analysis:** We calculated the Value Score for Logistic Regression across multiple thresholds.

Threshold	Recall	FPR	Value Score	Rank
0.30	**0.756**	**0.255**	**3023.439**	**1**
0.40	0.617	0.152	2486.145	2
0.50	0.449	0.082	1816.059	3



Optimal Threshold Selection

The optimal threshold is 0.30.

This threshold maximizes the financial outcome, accepting a higher False Positive Rate (25.5%) to ensure a high True Positive Rate (Recall of 75.6%).

Key Insights and Performance Results

1

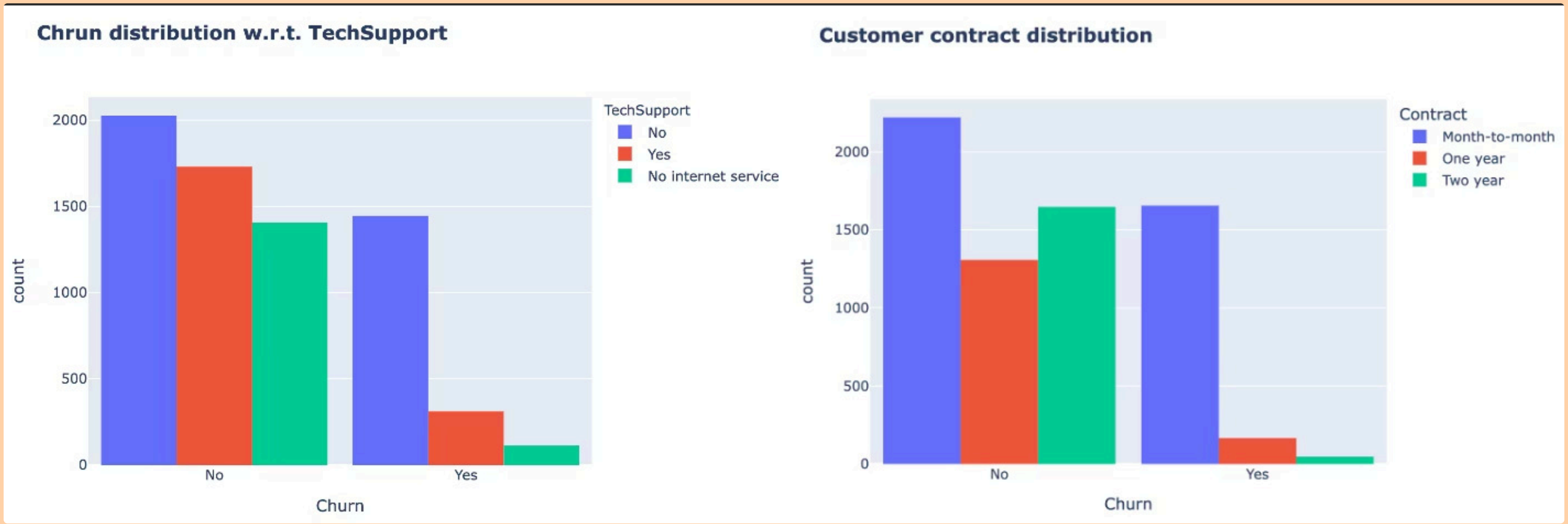
Key Insight 1 (The Churn Trigger):

The single best predictor of churn is the high_churn_risk feature, which combines high Monthly Charges with short Tenure → Watch new customers closely: Combine high monthly charges + short tenure = high risk churn

2

Key Insight 2 (The Retention Shield):

Long-term **Contract** agreements and technical add-ons (Security, Tech Support) are the most effective retention tools.



- **Final Model:** Logistic Regression.
- **Optimal Threshold:** 0.30.
- **Performance at Optimal TH:** Recall of 75.6% and an FPR of 25.5%.

Business Impact and ROI with hard numbers

Metric	Calculation	Result
Baseline Annual Loss	Losing all 1869 churners 1869 times 4,400.30$ (CLV)	**8,224,760.70$**
Churners saved	(Recall x Churners) → 0.756 times 1869	**1413 retentioned customer**
Model's Annual Loss	(Missed churners x CLV) → 456 times 4,400.30$	**2,006,695.21$**
Total saving	Baseline Annual Loss - Model's Annual Loss	**5,793,576.29$**

- The model successfully identifies 75.6% of actual churners (**True Positives**) who can be saved, providing a saved value of 4,100.30$ each.
- The model incurs a controlled cost by incorrectly flagging 25.5% of stable customers (**False Positives**) who receive an unnecessary 300$ discount.
- The model ensures maximum revenue impact by explicitly balancing these benefits and costs.

Recommendations and Next Steps

01

Deployment Recommendation:

Implement the Logistic Regression model using the **0.30 probability threshold**. This threshold is proven to maximize net financial value.

02

Integration:

Ensure smooth, automated integration of the model's output (positive prediction client IDs) into the existing ML discount system as required.

03

Marketing Action:

Use the insights regarding high-risk customers (short tenure, high charges, lack of contract/support) to tailor specific retention campaigns that go beyond just temporary discounts.

04

Monitoring:

Continuously monitor the real-world performance metrics (Recall and FPR) to validate the optimal threshold and ensure the financial gains are realized.

Thank You!

We appreciate your time and attention. We are open to any questions you may have.