Enhancing Customer Retention with Machine Learning: A Telco Case Study

Analysis and Prediction

Katriel Ester Amanda
Niken Andika Putri

Project Contents

Business Understanding

- Business problem understanding
- Metrics evaluation

Data Understanding

- Data features
- Data cleaning
- EDA
- Data pre-processing

Modeling and Evaluation

- Benchmark model
- Tuned Model

Explainable Machine Learning

- Feature Importance

CLR

- Conclusion
- Limitation
- Recommendation

Business Understanding

Business Understanding

context

- The telecom industry is highly competitive, making it easy for customers to **switch providers** due to poor service, high costs, or inadequate support.
- High churn rates reflect customer dissatisfaction and can severely impact a company's financial health.
- Identifying the **key drivers of churn**, such as pricing, contract terms, service reliability, or customer engagement, is essential for implementing targeted retention strategies.

problem statement

- **Customer churn** is a critical challenge for industries that rely on long-term customer relationships, such as telecommunications.
- **Retaining existing customers** is significantly more cost-effective than acquiring new ones.
- For telecom companies, **high churn rates** can lead to substantial revenue losses and increased marketing costs.

primary objective

- Develop a predictive model to identify customers who are most likely to churn.

Metric Evaluation

		Predicted	
		Positive Would Churn (1)	**Negative** Would Not Chrun (0)
Actual	**Positive** Will Churn (1)	True Positive (**TP**) Model correctly predicts costumer who will churn	False Negative (**FN**) Model predicts costumer will not churn when actually they
	Negative Will Not Churn (0)	False Positive (**FP**) Model predicts costumer will churn when actually they don't	True Negative (**TN**) Model correctly predicts costumer who will not churn

Primary metric: F2-Score

- Focuses more on recall to minimize missed churners, which are costlier.

Additional metrics:

Precision: Measures the accuracy of positive predictions.

Recall: Captures how well the model identifies actual churners.

Accuracy: Overall correctness of predictions.

Data Understanding

Data Features

	dataFeatures	dataType	null	unique	sample unique
0	customerID	object	0	7043	[4815-GBTCD, 2782-JEEBU]
1	gender	object	0	2	[Female, Male]
2	SeniorCitizen	int64	0	2	[1, 0]
3	Partner	object	0	2	[Yes, No]
4	Dependents	object	0	2	[Yes, No]
5	tenure	int64	0	73	[59, 42]
6	PhoneService	object	0	2	[No, Yes]
7	MultipleLines	object	0	3	[No, Yes]
8	InternetService	object	0	3	[DSL, Fiber optic]
9	OnlineSecurity	object	0	3	[Yes, No]
10	OnlineBackup	object	0	3	[No, Yes]
11	DeviceProtection	object	0	3	[Yes, No]
12	TechSupport	object	0	3	[No internet service, No]
13	StreamingTV	object	0	3	[No internet service, Yes]
14	StreamingMovies	object	0	3	[No internet service, Yes]
15	Contract	object	0	3	[Month-to-month, One year]
16	PaperlessBilling	object	0	2	[Yes, No]
17	PaymentMethod	object	0	4	[Credit card (automatic), Bank transfer (autom...
18	MonthlyCharges	float64	0	1585	[20.45, 54.3]
19	TotalCharges	object	0	6531	[814.75, 92.05]
20	Churn	object	0	2	[Yes, No]

- The data contains of **21 features** and **7,043 entries**.
- **Target feature**: Churn status (Yes/No)
- **Explanatory features**:
  1. Demographic factors
  2. Customer account information
  3. Add-on services

Data Cleaning

	PaymentMethod	MonthlyCharges	TotalCharges	Churn
488	Bank transfer (automatic)	52.55	NaN	No
753	Mailed check	20.25	NaN	No
936	Mailed check	80.85	NaN	No
1082	Mailed check	25.75	NaN	No
1340	Credit card (automatic)	56.05	NaN	No
3331	Mailed check	19.85	NaN	No
3826	Mailed check	25.35	NaN	No
4380	Mailed check	20.00	NaN	No
5218	Mailed check	19.70	NaN	No
6670	Mailed check	73.35	NaN	No
6754	Bank transfer (automatic)	61.90	NaN	No

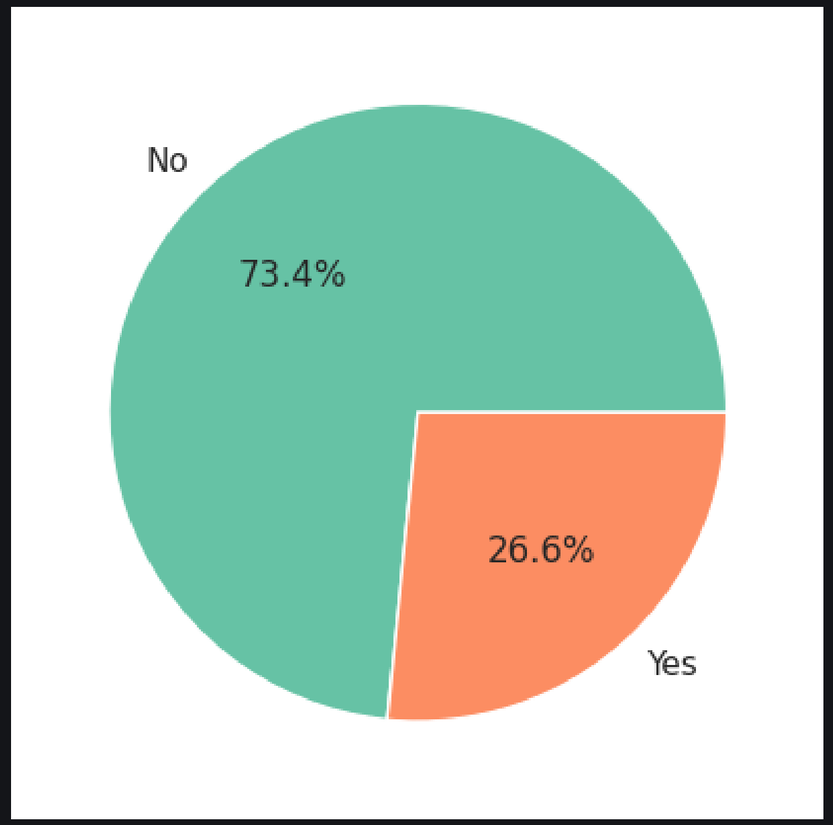
Missing values handling

- The **TotalCharges** variable has *less than 1%* missing values, which will be removed.

Outliers handling

- None of the numerical variables contain outliers.

EDA: Imbalanced Dataset



Problem of imbalanced dataset

- The Telco Customer Churn dataset is **imbalanced**, with a higher proportion of “No Churn” cases compared to “Churn” cases.
- Imbalanced data can lead to biased models that favor the majority class (No Churn), resulting in **poor performance** in identifying actual churners.

Challenges of imbalanced data

- Standard models tend to predict “**No Churn**” for most cases due to the skewed distribution.
- Metrics like **accuracy** can be misleading as they do not reflect the model’s ability to detect the minority class (Churn).
- Requires careful handling to **prevent underestimating** the risk of customer churn.

Solutions implemented

- **Resampling Techniques**: Applied SMOTE (Synthetic Minority Over-sampling Technique) to balance the classes by generating synthetic samples for the minority class.
- **Evaluation Metrics**: Used F2-score to focus more on recall (identifying churners) while avoiding a bias towards accuracy.
- **Threshold Optimization**: Optimized the decision threshold to maximize recall and reduce missed churners.

EDA: Churn by Demographic

Gender

Both male and female customers churn at similar rates, indicating that gender does not significantly impact customer retention.

Senior Citizen

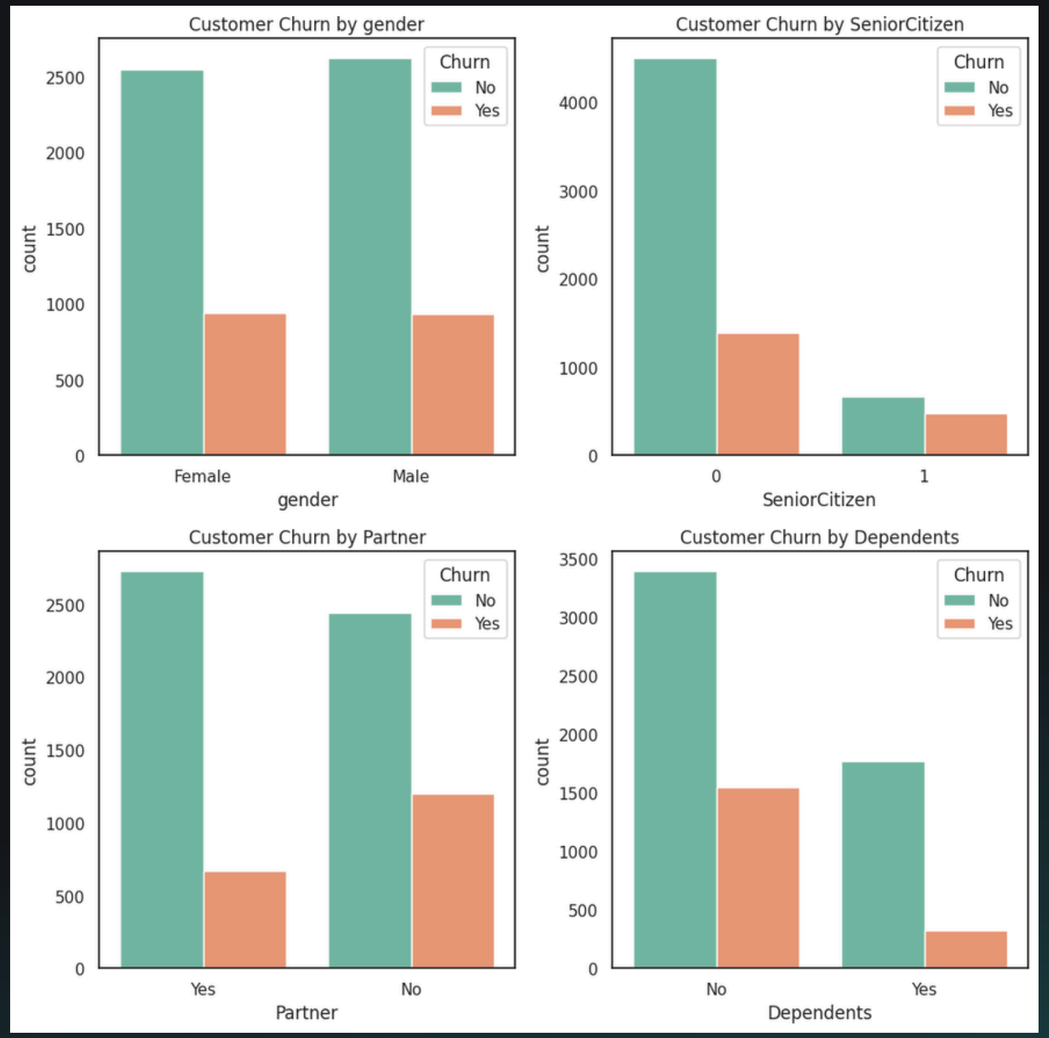
Senior citizens have a higher churn rate compared to non-senior customers.

Partner

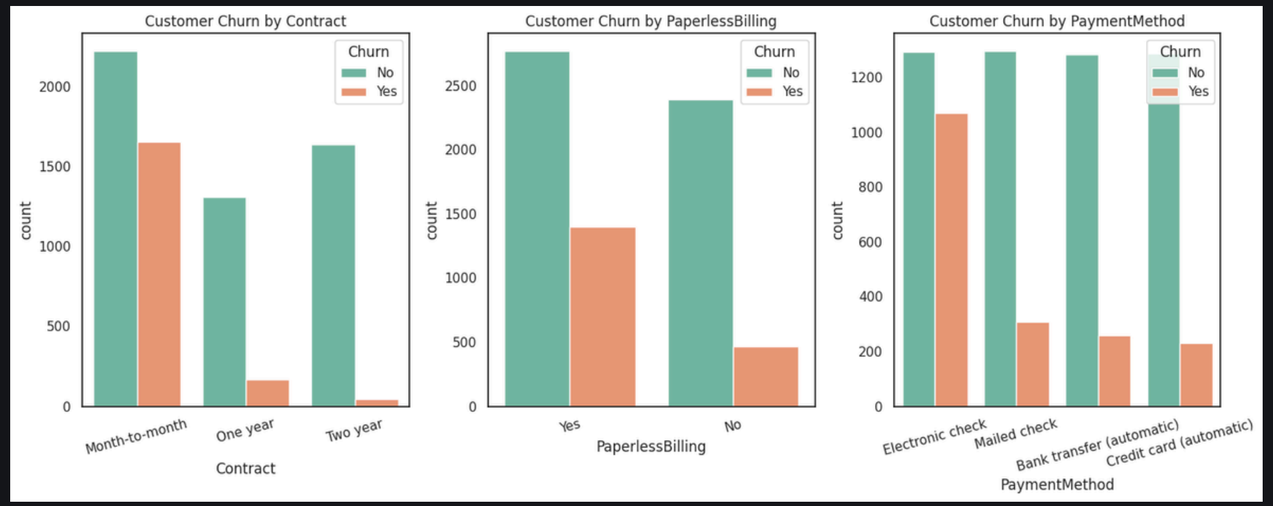
Customers without a partner churn at a higher rate compared to those with a partner.

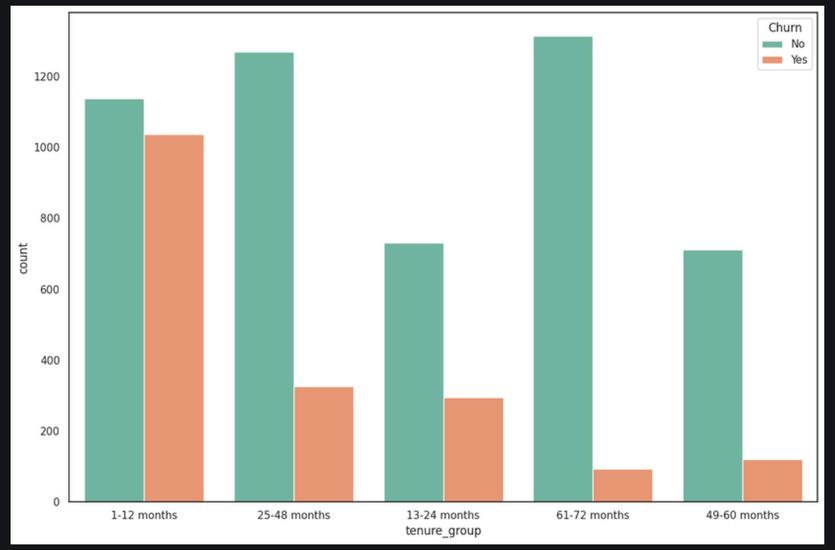
Dependents

Customers with dependents churn less frequently than those without dependents.



EDA: Churn by Costumer Account Information (1)





1. Contract

Customers with month-to-month contracts leave the company more often.

2. PaperlessBilling

Customers who use paperless billing tend to leave at a higher rate than those who receive paper bills.

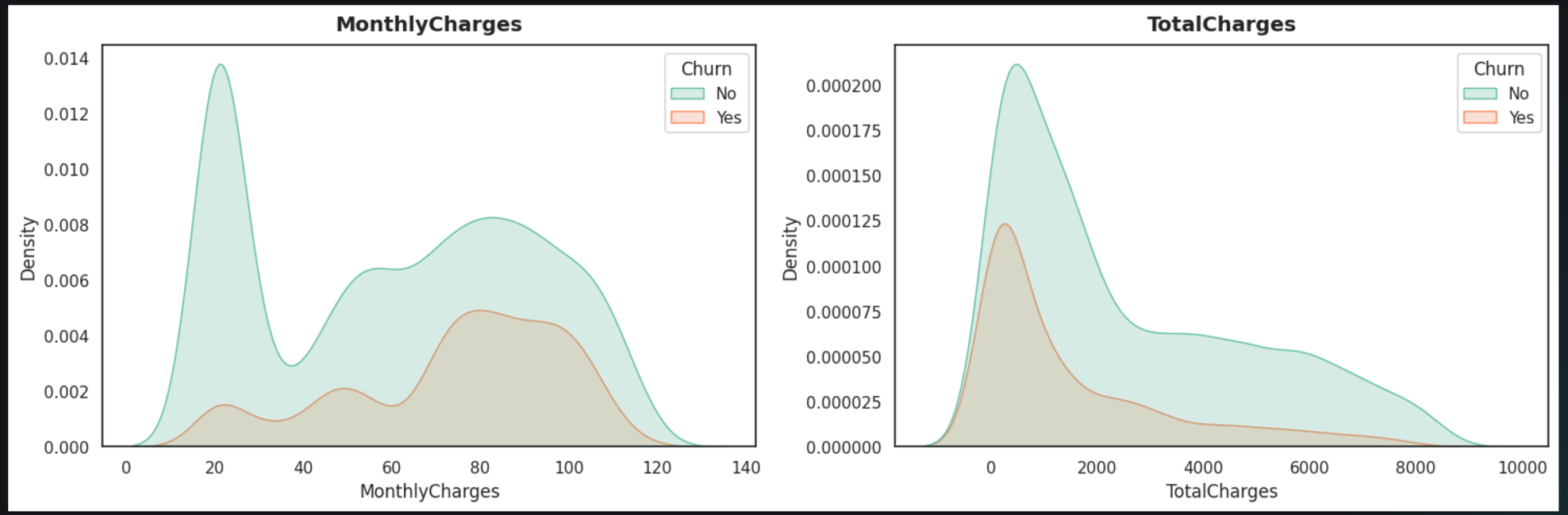
3. PaymentMethod

Customers who pay via electronic checks have the highest churn rate.

4. Tenure

Customers with loner tenures (> 25-48 months) have significantly lower churn rates.

EDA: Churn by Costumer Account Information (2)



Monthly Charges

The churned customers are more evenly spread across mid-to-high price ranges, meaning churn risk increases as monthly charges rise.

Total Charges

Customers with higher total charges (are much less likely to churn, reinforcing the idea that longer-tenured customers or have multiple bundled services are more loyal.

EDA: Churn by Add-on Services

PhoneService

Costumers who have phone service make up the majority of the dataset, and while some still churn, their churn rate is lower.

MultipleLines

Costumers with multiple lines churn slightly more than those with a single line.

InternetService

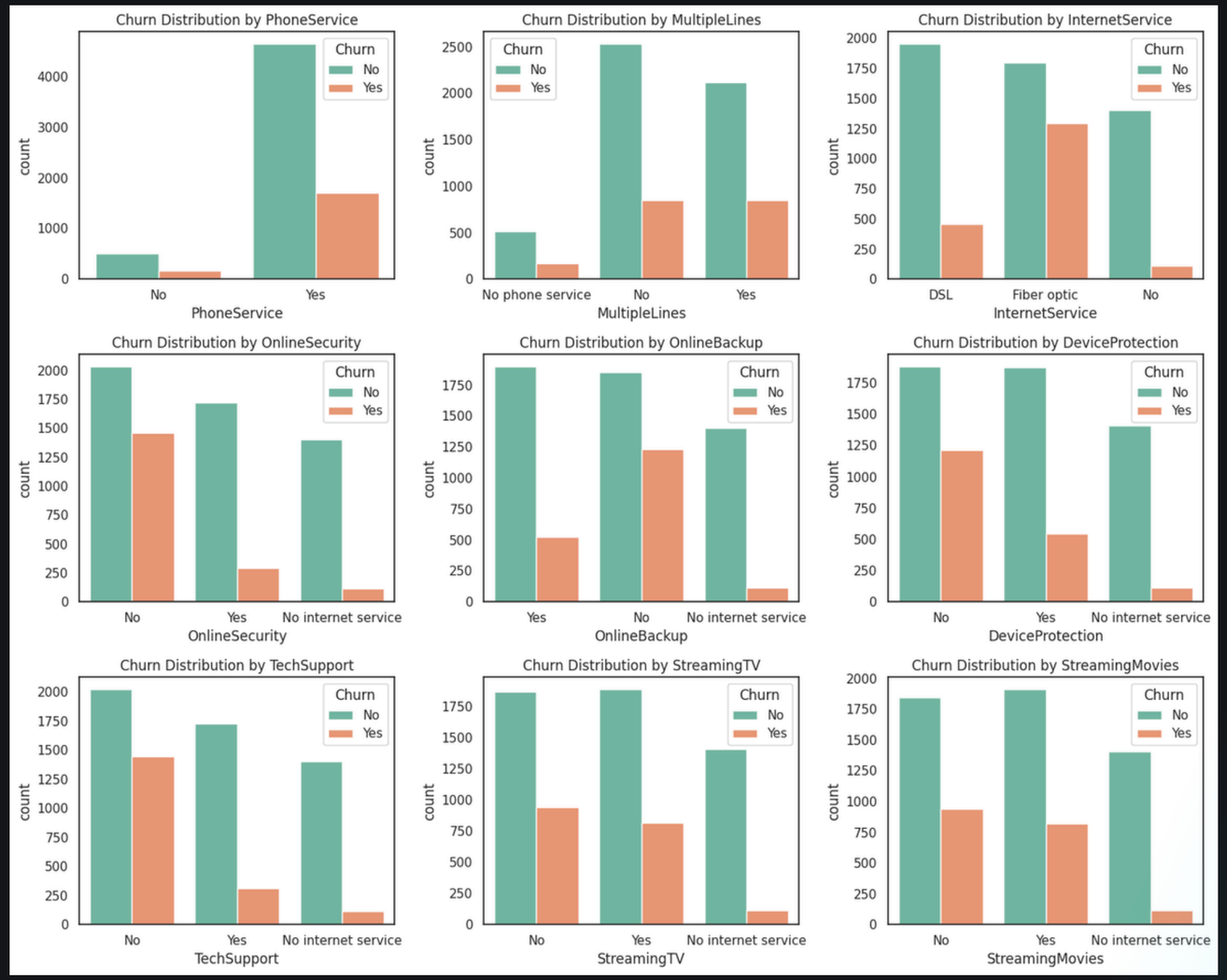
Fiber optic users have the highest churn rate, while DSL users churn less.

OnlineSecurity

Customers without online security services churn more.

OnlineBackup

Customers without online backup services have a higher churn rate compared to those who have it.



DeviceProtection

Customers without device protection are more likely to churn compared to those who have it.

TechSupport

Customers without tech support have a significantly higher churn rate.

StreamingTV

Customers without streaming TV services tend to churn at a higher rate than those with it.

StreamingMovies

Customers without streaming movie services have a higher churn rate compared to those who have access.

Data Preprocessing (1)

Handling Missing & Redundant Data

- Dropped **highly correlated feature** (TotalCharges).
- Removed tenure (after binning) to **prevent multicollinearity**

Feature Engineering

- **Binned tenure** into yearly intervals for better interpretability.
- Created a **new interaction feature** between Contractt and MonthlyCharges to capture cost impact.

Feature Encoding

- **Binary Encoding**: Convert. Yes/No values to 0/1.
- **One-Hot Encoding**: Applied to most categorical variables.
- **Ordinal Encoding**: Contract transformed as (0: Month-to-month, 1: One-year, 2: Two-year) to reflect increasing commitment levels.

Data Preprocessing (2)

Scaling & Standardization

- Standardized MonthlyCharges and tenure using **StandardScaler** to bring numeric features to a comparable scale.

Stratified Train-Test Splitting

- **Ensure balanced churn distribution** between training & test sets.
- Applied **80-20 stratified split** to preserve the ratio of churned vs. non-churned customers.

Pipeline Integration for Consistency

- All transformations were combined into a **single preprocessing pipeline** to ensure reproducibility.
- This prevents data leakage and maintains consistent transformations across training & testing.

Machine Learning

Model Benchmarking

Cross Validation and Test Benchmarking Performance

- **Best Model:** **Logistic Regression** achieves the highest F2 Score (0.7154) with the lowest variance (0.0131), making it the most effective and stable model.
- **Performance Trends:** Tree-based models (Decision Tree, Random Forest: 0.5150) perform the worst, while boosting models (LightGBM, XGBoost) fail to outperform simpler models, suggesting **hyperparameter or dataset limitations.**
- **Stability & Variability:** **Logistic Regression** remains the most consistent choice.

Model	CV F2 Score (Mean)	CV F2 Score (Std Dev)	Test F2 Score (Mean)	Test F2 Score (Std Dev)
Logistic Regression	**0.7154**	0.0131	**0.7154**	0.0131
K-Nearest Neighbors	0.6427	0.0162	0.6427	0.0162
LightGBM	0.5864	0.0138	0.5864	0.0138
XGBoost	0.5633	0.0296	0.5633	0.0296
Decision Tree	0.5333	0.0179	0.5333	0.0179
Random Forest	0.5150	0.0392	0.5150	0.0392

conclusion

Logistic Regression is the best-performing model with strong and consistent results.

Hyperparameter Tuning

Best Model: XGBoost

- **Best F2 Score (0.7130) & Highest Recall (87.17%):** Ensures minimal false negatives, aligning with the project's goal.
- **Handles Imbalanced Data:** Adjusts weight for minority class (churners), improving recall.
- **Captures Complex Patterns:** Boosting method learns intricate relationships, outperforming simpler models.
- **Regularization (L1 & L2):** Reduces overfitting, enhancing generalization to new data.

Model	Test F2 Score	Test Recall	Test Precision	Comments
XGBoost	**0.7130**	**0.8717**	0.4127	**Best recall, minimizing false negatives.**
Logistic Regression	0.6922	0.7674	0.4974	Higher precision but more false negatives.
KNN	0.6744	0.7775	0.4434	Decent recall, but lower precision.
LightGBM	0.6063	0.6177	0.5648	High precision but poor recall.

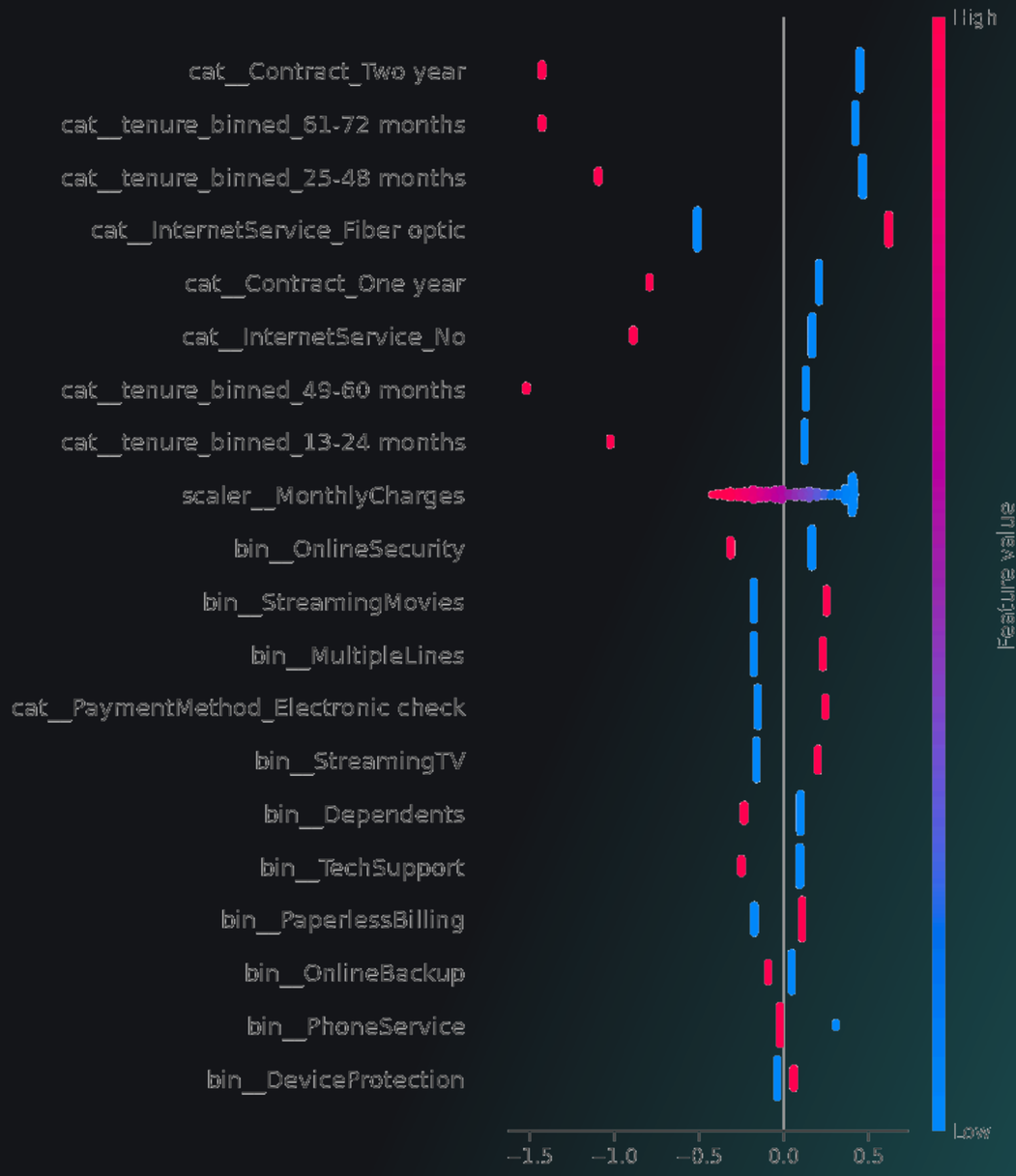
best hyperparameters

- colsample_bytree: 1.0, learning_rate: 0.01, max_depth: 3
- n_estimators: 50, subsample: 0.8

Explainable ML

Key Drivers of Churn

- **Top Influential Features:** Features such as **Contract_Two year, tenure_binned, and InternetService_Fiber optic** are the most influential factors affecting the churn prediction model.
- **Contract Type:** Customers with a **“Two-year contract” have a significant negative impact** on the churn prediction, suggesting that they are less likely to churn.
- **Tenure:** Longer tenure categories (e.g., “61-72 months,” “49-60 months”) show a negative impact on churn, indicating that **longer-tenured customers are less likely to leave**.
- **Internet Service:** Customers using **"Fiber optic" service tend to have a higher likelihood of churn** compared to those with “No” or other internet services.



Feature Importance

Key Drivers of Churn

Tenure & Contract Type

- Customers with **longer tenure are less likely to churn**, while those under 12 months have a higher churn risk.
- **Long-term contracts (e.g., two-year contracts) reduce churn**, whereas month-to-month plans increase churn risk.

Monthly Charges & Price Sensitivity

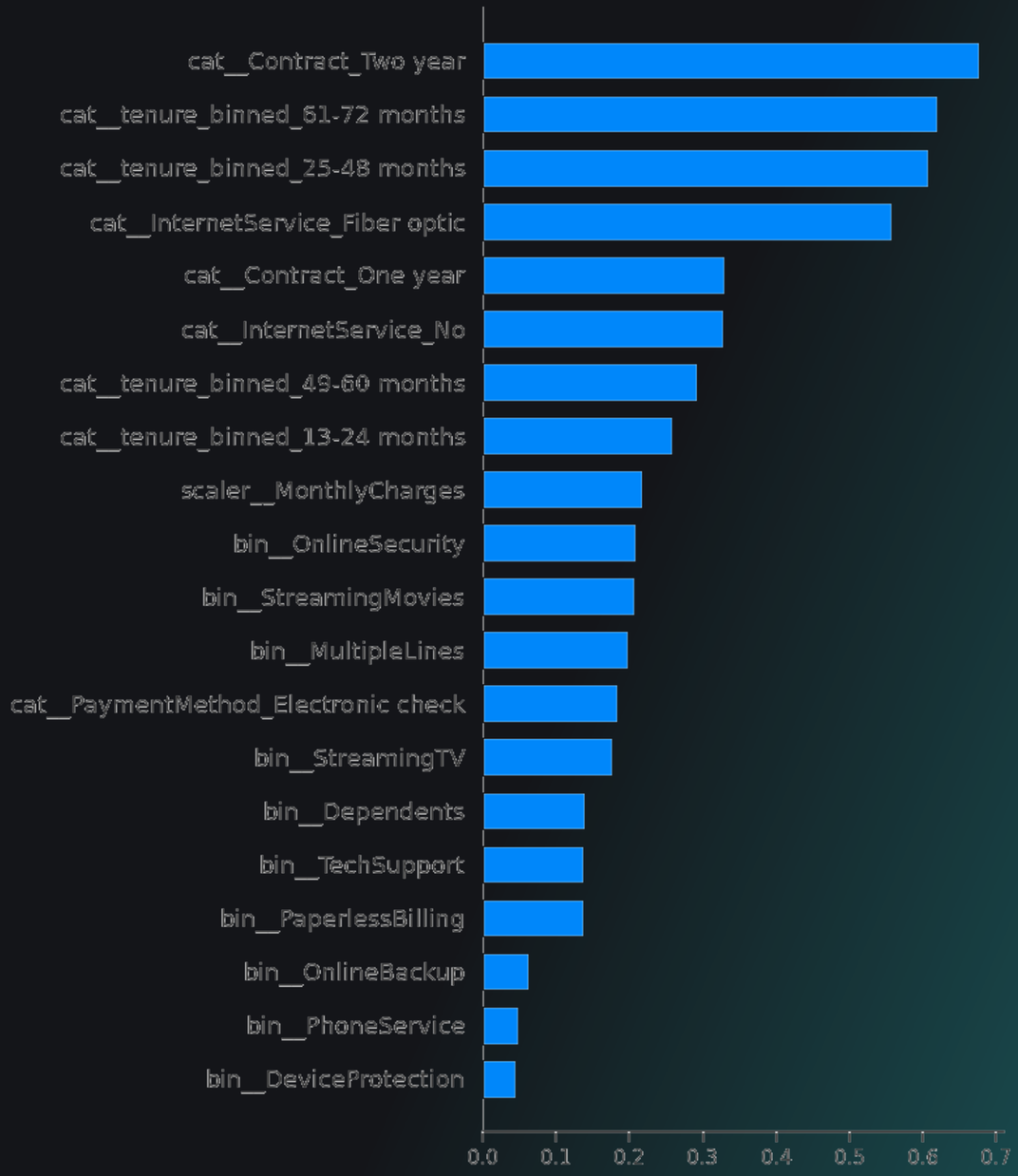
- **Higher monthly charges are linked to higher churn rates**, indicating that cost is a major factor in customer retention.

Internet Service Type & Performance Issues

- **Fiber optic users churn more often**, possibly due to pricing, performance concerns, or competitive alternatives.

Service Add-Ons & Customer Loyalty

- Security, backup, tech support, and streaming services reduce churn, suggesting that **customers who use value-added services stay longer**.



Conclusion

XGBoost identified as the best model, achieving an **F2-score of 0.71** with strong recall and precision.

Contract type, tenure, and service add-ons emerged as key churn predictors, highlighting the importance of long-term commitments and additional services in retention.

Limitation

1. **Imbalanced dataset:** Despite using techniques like SMOTE, class imbalance may still affect model performance.
2. **Feature engineering constraints:** Some **behavioral factors** influencing churn (e.g., customer satisfaction, service interactions) were not available in the dataset.
3. **Model generalization:** While XGBoost performed well, results may vary when applied to new data; **ongoing monitoring is required.**

Recommendations

1. **Address Class Imbalance:** The dataset remains imbalanced despite SMOTE application. Exploring **alternative sampling techniques**, such as ADASYN or stratified bagging, may improve model performance.
2. **Enhance Feature Engineering:** Incorporate **behavioral and interaction-based features**, such as customer service call frequency, complaint history, and service downgrade requests, to provide deeper insights into churn predictors.
3. **Expand Data Sources:** If feasible, integrate **customer sentiment analysis** from reviews or support tickets to detect dissatisfaction signals early and enhance predictive accuracy.
4. **Leverage AI-Driven Churn Prevention:** Use model insights to **create an early warning system**, triggering personalized retention offers or interventions before a customer reaches the churn tipping point.

Business Recommendations

Long-Term Contracts Reduce Churn

- **Insight:** Customers with two-year contracts churn less.
- **Recommendation:** Offer discounts, loyalty rewards, and bundled benefits to encourage long-term contracts.

New Customers (<12 Months) Are High-Risk

- **Insight:** Early-stage customers churn the most.
- **Recommendation:** Implement a “Welcome Program” with onboarding support, first-time discounts, and proactive outreach.

High Monthly Charges Increase Churn

- **Insight:** Price-sensitive customers are more likely to leave.
- **Recommendation:** Provide tiered pricing, retention discounts, and flexible payment options to retain them.

Fiber Optic Customers Churn More

- **Insight:** Service issues or competition impact fiber users.
- **Recommendation:** Improve service reliability, address complaints, and offer exclusive retention incentives.

Value-Added Services Improve Retention

- **Insight:** Customers with Online Security, Tech Support, or Streaming churn less.
- **Recommendation:** Promote bundled packages, free trials, and upselling strategies for these services.

Customers with Dependents Are More Loyal

- **Insight:** Family-oriented customers stay longer.
- **Recommendation:** Create family-friendly bundles, exclusive promotions, and a loyalty program for families.

thank you