Predicting Customer Churn in Telecom

A Machine Learning Approach

By Hans Darmawan - JCDS2602

Overview

CRISP-DM Approach

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. Evaluation
6. Deployment
7. Conclusion and Recommendations

1. Business Understanding

- Background
- Problem Statement
- Goals
- Analytic Approach
- Metric Evaluation
- Success Criteria

Business Understanding

Background

Customer Churn Definition

Customer churn in telecom refers to the rate at which customers stop using services, significantly impacting revenue and profitability.

Causes of Churn

Factors contributing to customer churn include dissatisfaction with service quality, pricing, customer support, and competitive offers.

Importance of Retention

Retaining existing customers is more cost-effective than acquiring new ones, making churn management crucial for financial health and growth.

Business Understanding

Problem Statement

Which customers are likely to churn?

Identifying At-Risk Customers: The primary question is to determine which customers are likely to churn.

What factors influence customer churn the most?

Influential Factors: Understanding the key factors that influence customer churn is crucial for developing effective strategies.

How can the company reduce churn and improve customer retention?

Strategies for Retention: Finding ways to reduce churn and improve customer retention is a central concern for the company.

Business Understanding

Goals

Develop Predictive Model

Create a model to classify customers as likely to churn or not, using historical data and machine learning techniques.

Identify Key Features

Analyze which factors most significantly impact customer churn to inform retention strategies.

Actionable Insights

Provide insights that enable the company to implement targeted strategies to reduce churn, such as personalized plans and improved customer support.

Business Understanding

Approaches

Rule Based

Widely used by companies.

Machine Learning Based

To overcome rule-based limitations: uncovering hidden patterns in customer behavior.

Business Understanding

Metric Evaluation

Customer Acquisition Cost (CAC)

CAC measures the total cost of acquiring a new customer, essential for evaluating marketing efficiency and profitability.

Customer Retention Cost (CRC)

CRC reflects the costs associated with retaining customers, including salaries and loyalty programs.

Recall Score

Recall is critical for assessing the effectiveness of churn prediction models, focusing on capturing at-risk customers for targeted retention efforts.

Business Understanding

Success Criteria

High Recall Performance

Achieve a recall rate of 80% or higher for effective churn prediction.

Reduce Customer Acquisition Cost and Retention Cost.

Achieve a total annual cost reduction rate of 20% or higher.

Interpretable Insights

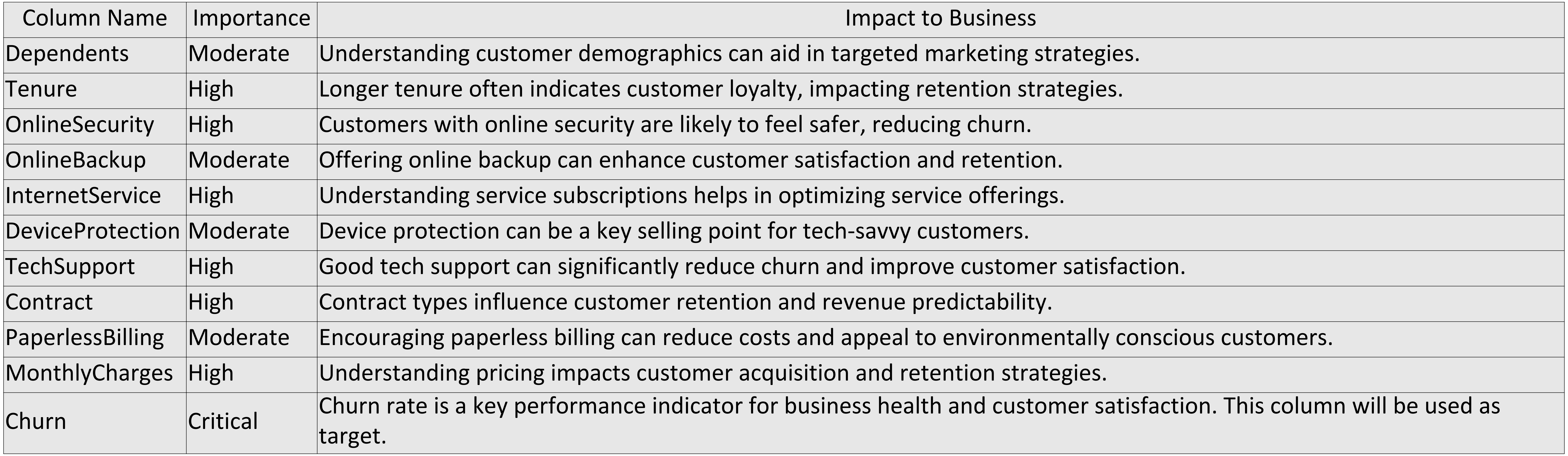
Provide insights that are interpretable and actionable for informed business decisions, ultimately reducing acquisition and retention costs.

2. Data Understanding

- Dataset Information
- Missing Values Checking
- Duplicated Values Checking
- Exploratory Data Analysis (EDA)

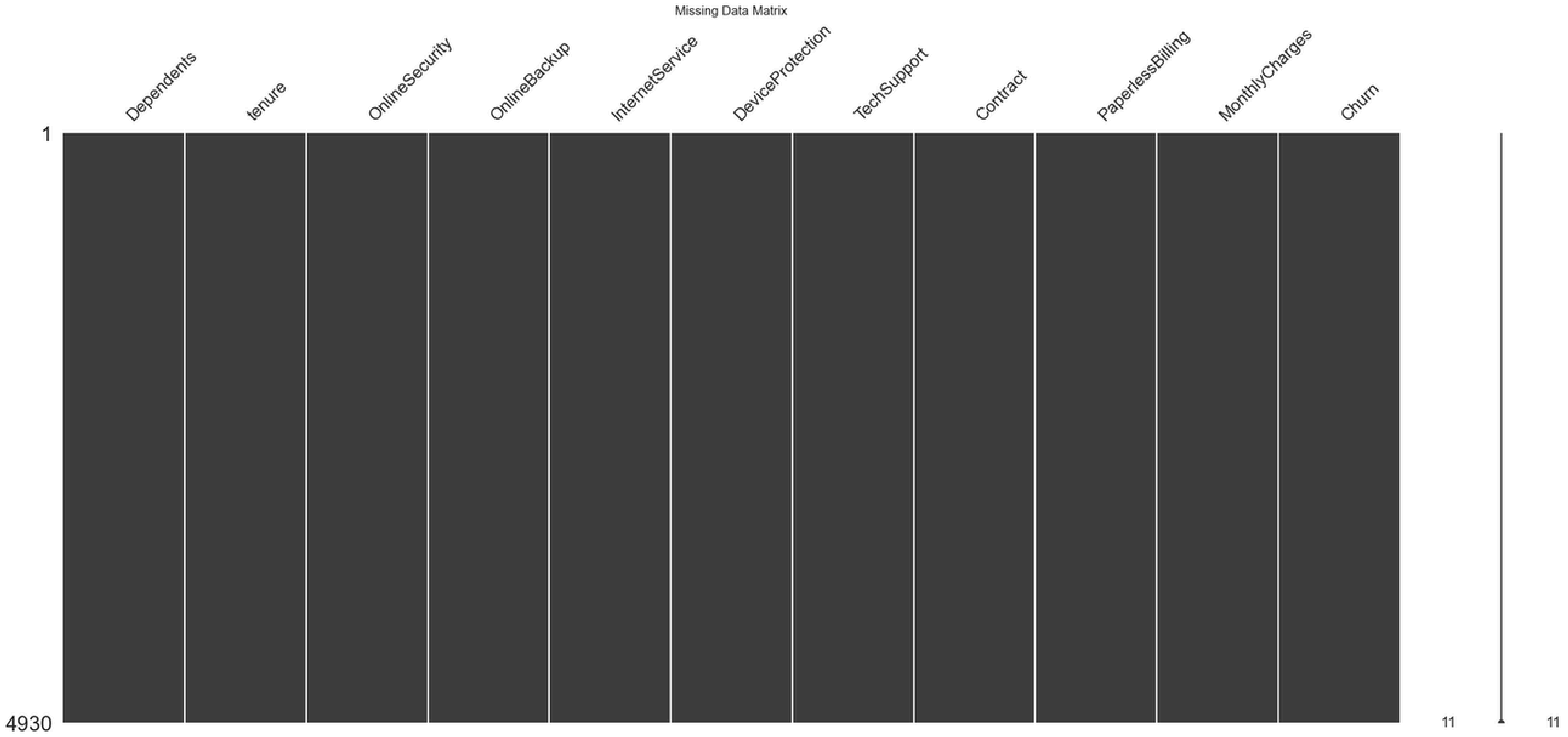
Data Understanding

Dataset Information



Column Name	Importance	Impact to Business
Dependents	Moderate	Understanding customer demographics can aid in targeted marketing strategies.
Tenure	High	Longer tenure often indicates customer loyalty, impacting retention strategies.
OnlineSecurity	High	Customers with online security are likely to feel safer, reducing churn.
OnlineBackup	Moderate	Offering online backup can enhance customer satisfaction and retention.
InternetService	High	Understanding service subscriptions helps in optimizing service offerings.
DeviceProtection	Moderate	Device protection can be a key selling point for tech-savvy customers.
TechSupport	High	Good tech support can significantly reduce churn and improve customer satisfaction.
Contract	High	Contract types influence customer retention and revenue predictability.
PaperlessBilling	Moderate	Encouraging paperless billing can reduce costs and appeal to environmentally conscious customers.
MonthlyCharges	High	Understanding pricing impacts customer acquisition and retention strategies.
Churn	Critical	Churn rate is a key performance indicator for business health and customer satisfaction. This column will be used as target.

Data Understanding



Missing Values Checking

- No missing values.
- No need to use imputer in pre-processing.

Data Understanding

Duplicated Values Checking

Number of duplicated rows: 77

Action: Retain Duplicates

Validity of Duplicates

Duplicated rows may represent legitimate repeated observations or transactions, crucial for accurate analysis, particularly in datasets involving transactional data or repeated measurements.

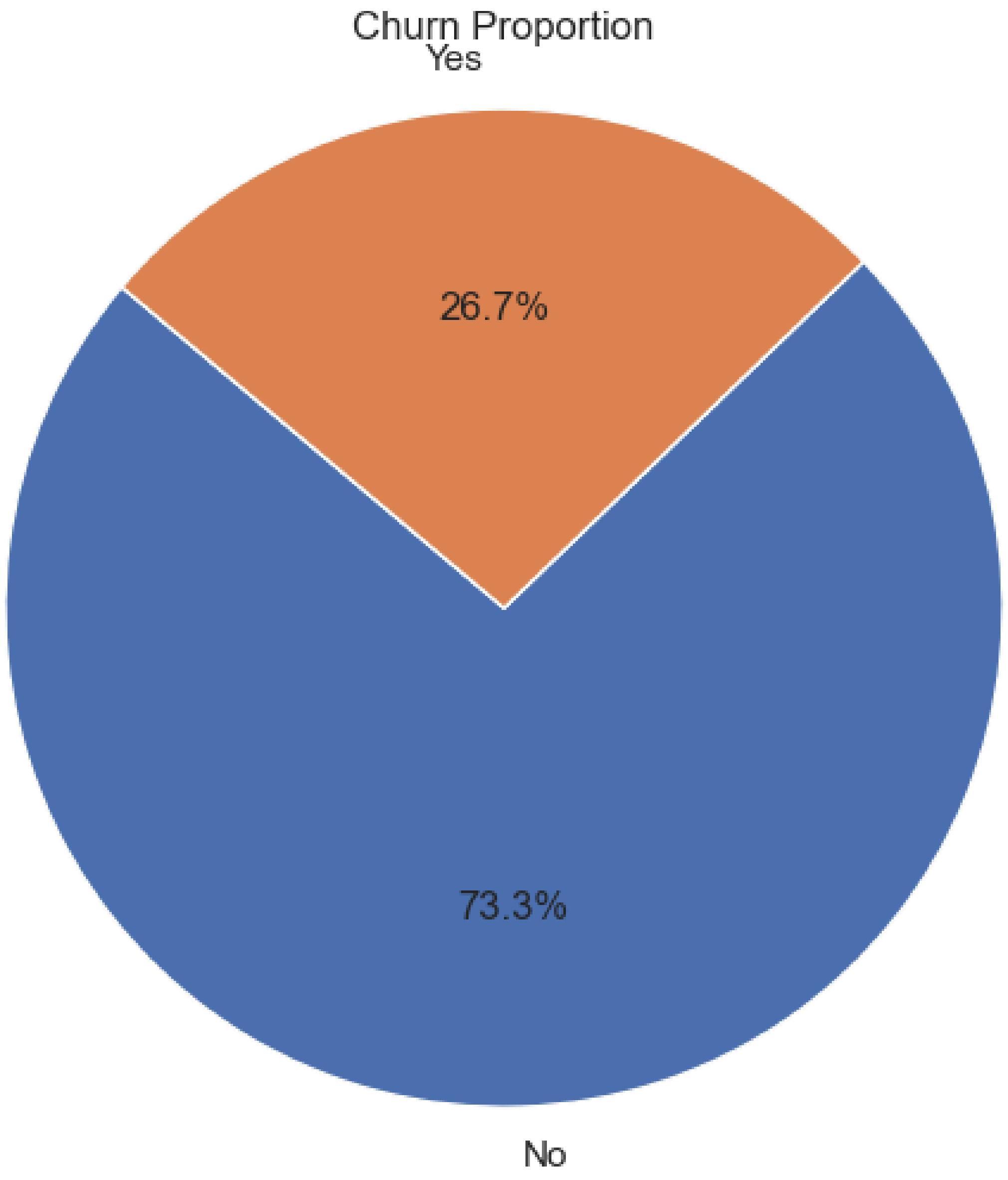
Impact on Data Integrity

Removing duplicates without understanding their context can distort data distributions and compromise data integrity, potentially leading to misleading results and affecting the validity of analyses.

Analytical Significance

Duplicates can provide meaningful information in certain analytical methods and visualizations, highlighting the importance of assessing their origin and significance before deciding to remove them.

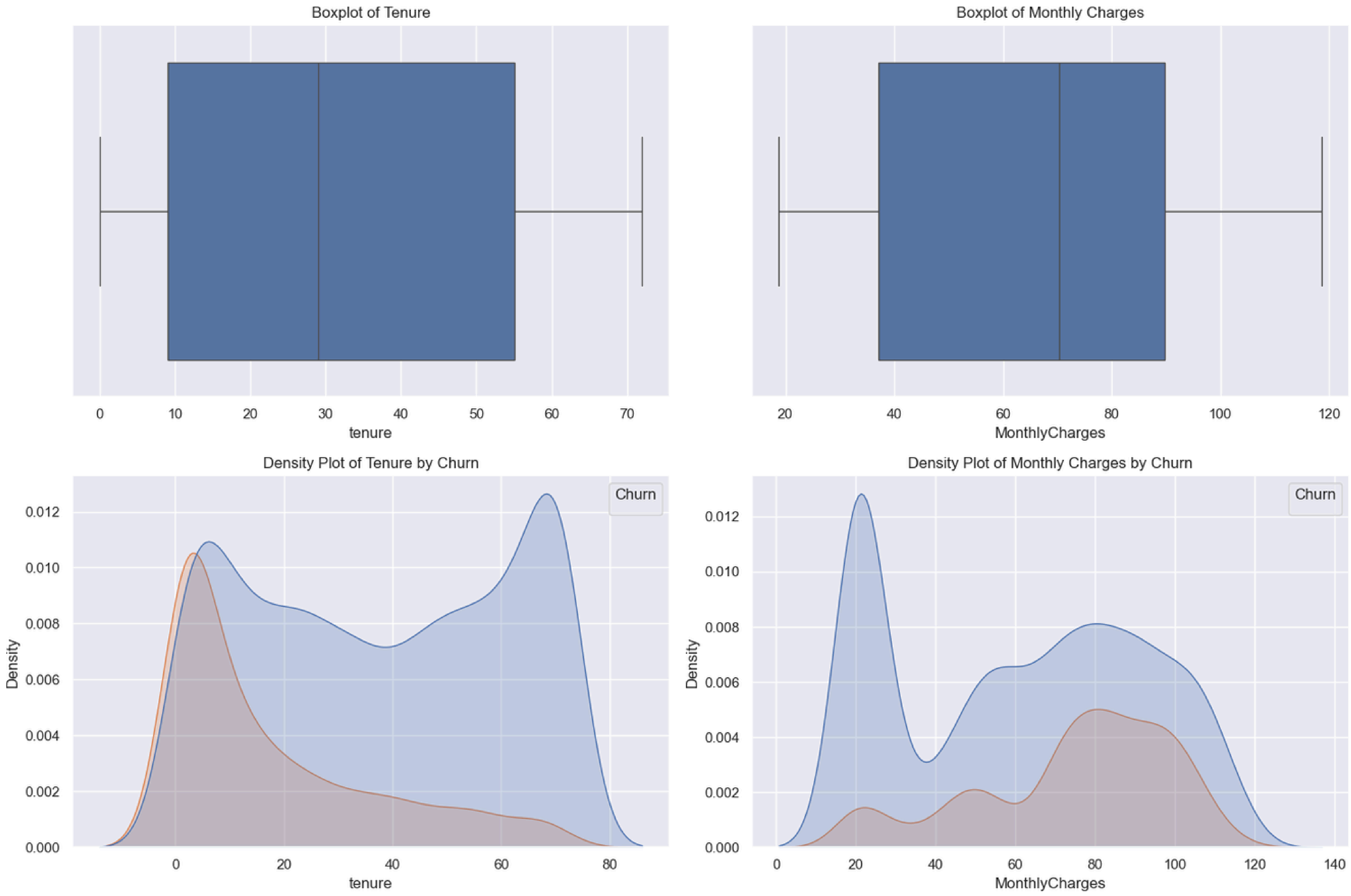
Data Understanding



EDA - Churn Proportion

- Churn: 26.7 %
- Not Churn: 73.3%
- Imbalanced Dataset
- Need to oversampling dataset

Data Understanding - EDA - Numerical Features Analysis



	count	mean	std	min	25%	50%	75%	max
tenure	4930.0	32.401217	24.501193	0.0	9.00	29.00	55.00	72.00
MonthlyCharges	4930.0	64.883032	29.923960	18.8	37.05	70.35	89.85	118.65

- ○ No outlier detected
- ○ No negative values
- ○ Not normally distributed
- ○ Action: Benchmark MinMax and Robust Scaler in Preprocessor

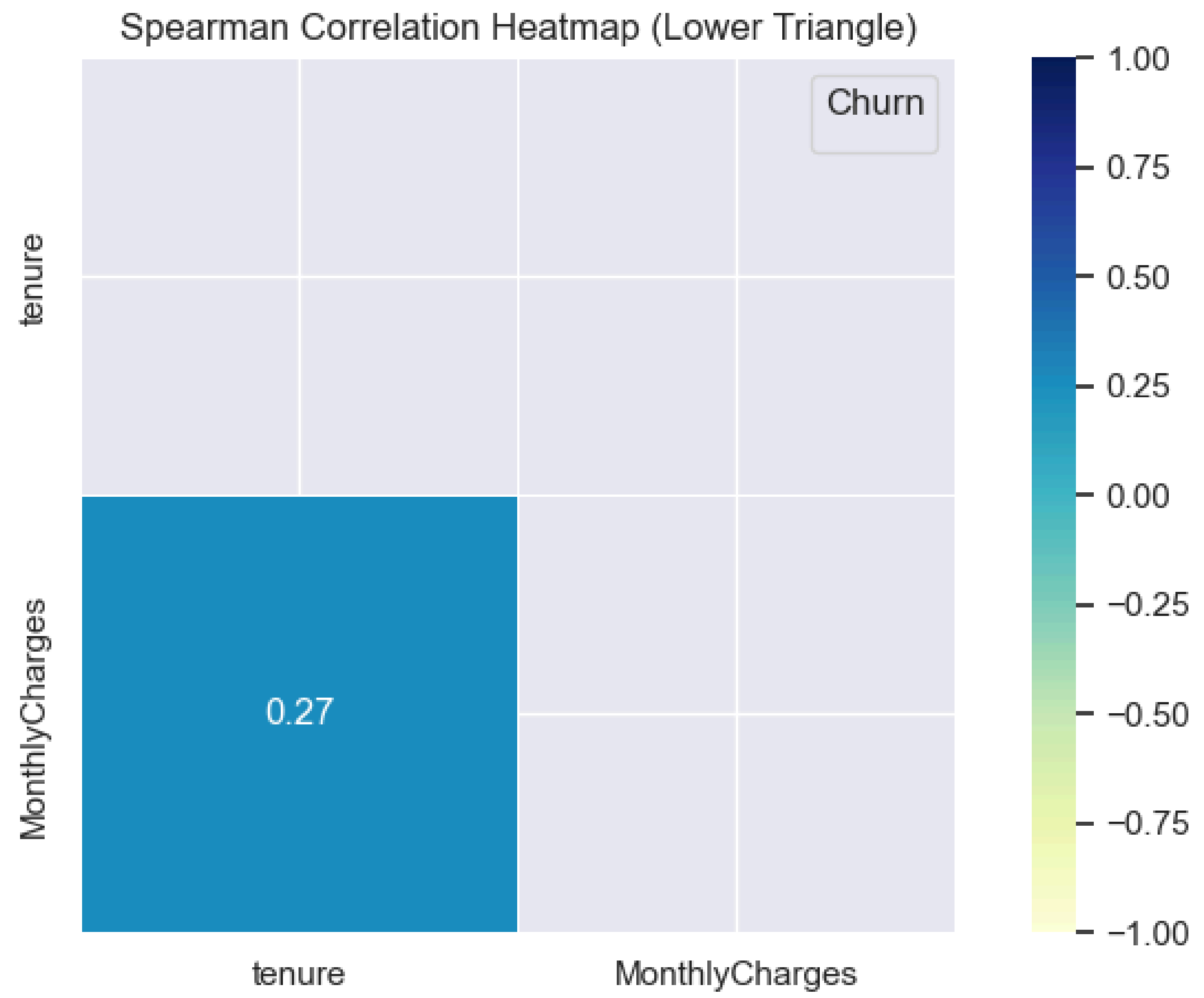
Data Understanding

	unique	top
Dependents	2	No
OnlineSecurity	3	No
OnlineBackup	3	No
InternetService	3	Fiber optic
DeviceProtection	3	No
TechSupport	3	No
Contract	3	Month-to-month
PaperlessBilling	2	Yes
Churn	2	No

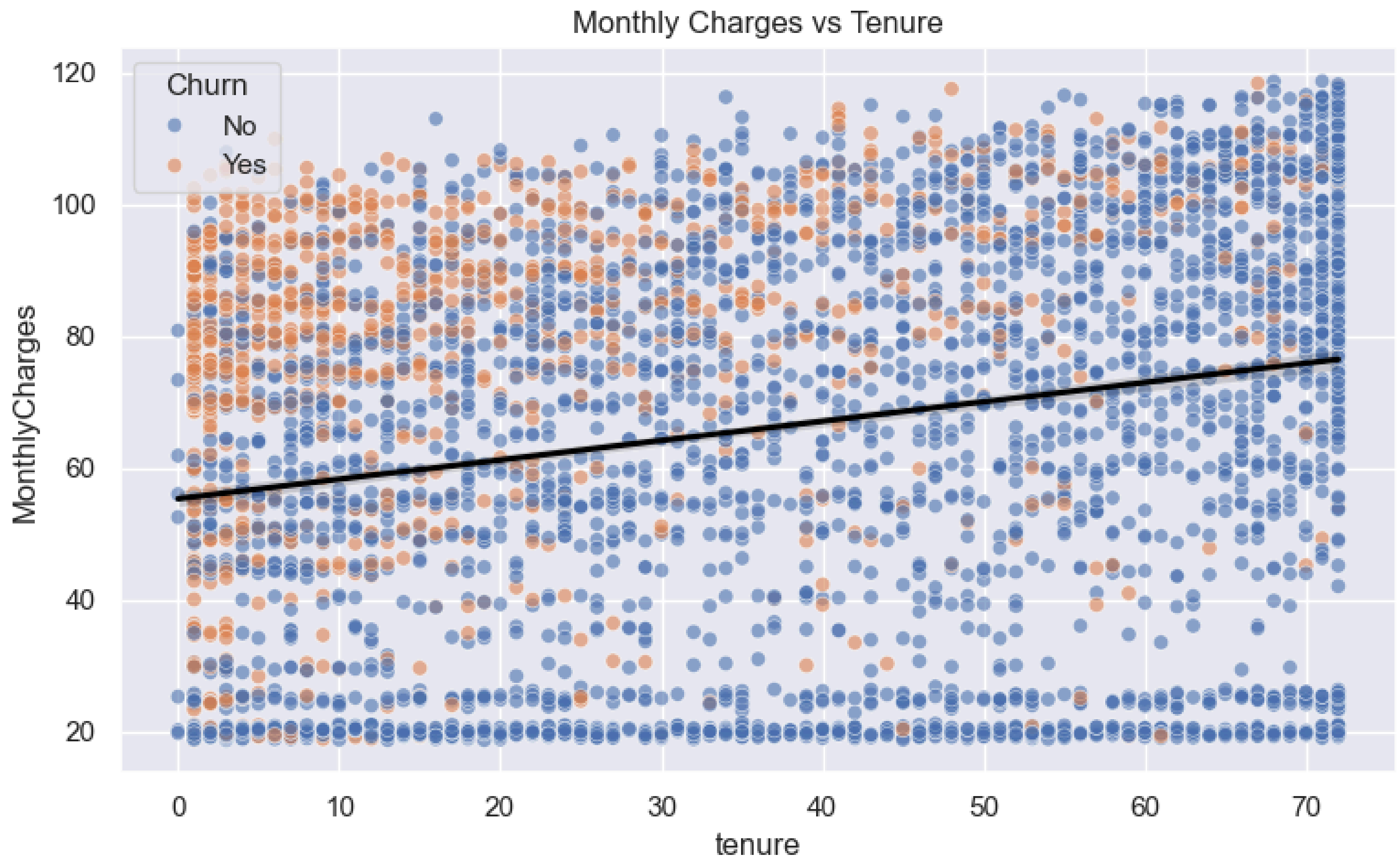
EDA - Categorical Feature Analysis

- All categorical columns are nominal, **except Contract columns (ordinal).**
- Actions: To use ordinal encoder for Contract columns and one hot encoder for the rest while pre-processing.

Data Understanding



Spearman Correlation Heatmap (Lower Triangle)



Monthly Charges vs Tenure

Correlation Analysis

- Low monotonic relationship
- Directly proportional

3. Data Preparation

- Feature Engineering
- Binning
- Target Labeling
- Define X and y
- Train-Test Split
- Data Transformation Setup

Data Preparation

Feature Engineering

Total Charges = tenure x Monthly Charges

Binning

Equal frequencies based on quarter quantile on tenure, monthly charges, and total charges.

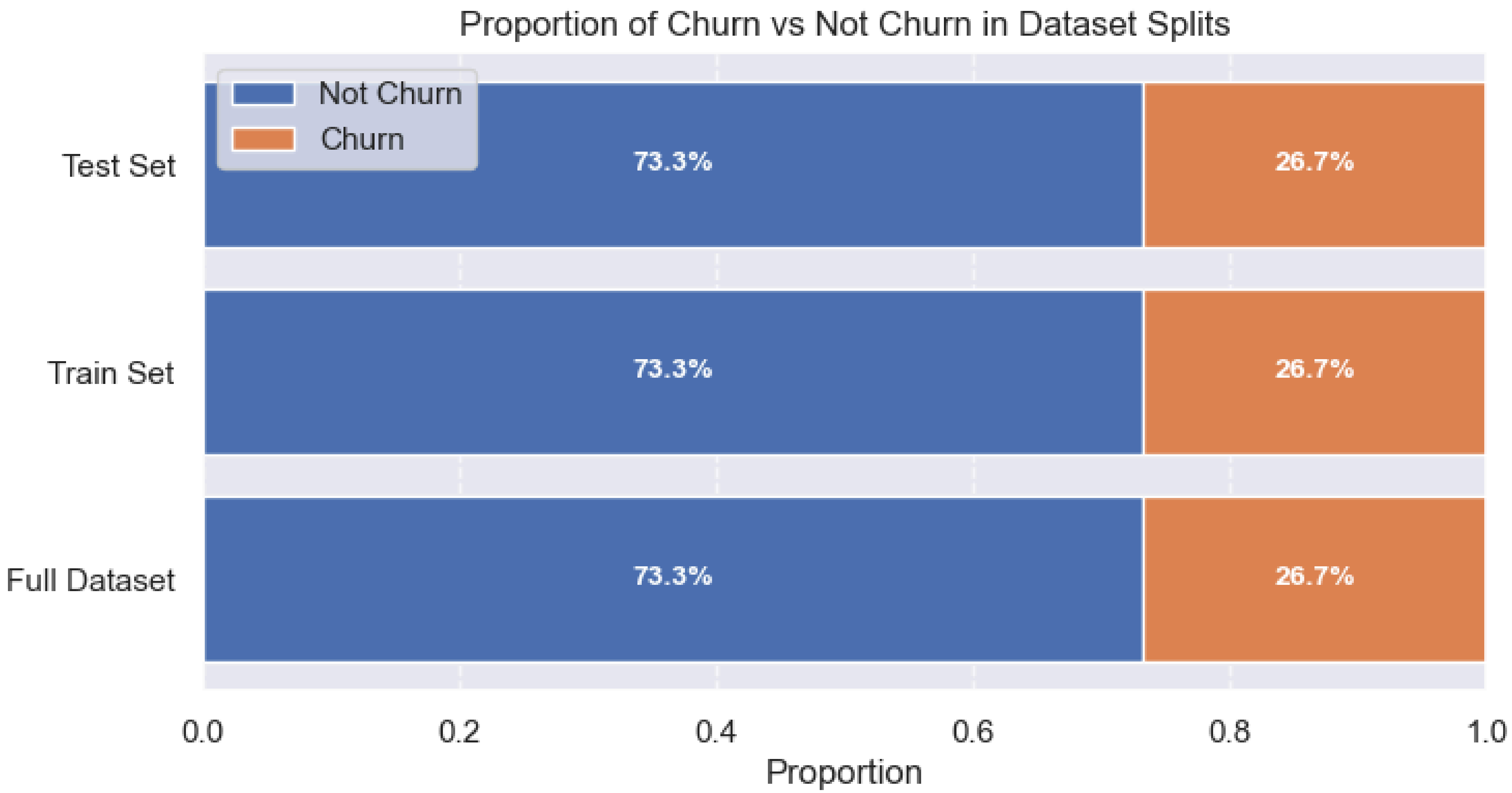
Target Labeling

Churn → 1, Not Churn → 0

Define Features and Target

Features (X): All columns except Churn
Target (y): Churn

Data Preparation



Train-Test Split

- Train Size: 80 %
- Test Size: 20 %
- Stratify on target

Data Preparation

Data Transformation Setup

Step 1

Identify Columns

Extract categorical columns (excluding Churn and Contract), numeric columns, and binary columns (with exactly 2 unique values) to categorize the data appropriately for preprocessing.

Step 2

Define Ordinal Categories

Specify the ordinal categories for the Contract column to ensure that the model understands the inherent order in these categories during training.

Step 3

Binary Mapping Function

Create a function to map binary values ('No' to 0 and 'Yes' to 1) to convert categorical binary features into a numerical format suitable for machine learning algorithms.

Step 4

Numeric Transformer Pipeline

Define a pipeline for scaling numeric features using a specified scaler (e.g., RobustScaler or MinMaxScaler) to standardize the range of numeric data and improve model performance.

Step 5

Preprocessor Creation

Create a preprocessor that combines transformations for numeric, binary, ordinal, and categorical features, allowing for a streamlined and efficient data preparation process.

4. Modeling

- Model Initialization and Benchmarking
- Hyperparameter Tuning
- Learning Curve
- Threshold Tuning

Modeling

Model Benchmark Initialization

Step 1

Define Base Models and Stacking Classifier

Create a list of base models (e.g., Logistic Regression, Random Forest, XGBoost) and define a stacking classifier using these models with Logistic Regression as the meta-model.

Step 2

Compile Models Dictionary

Construct a comprehensive dictionary of various machine learning models, including traditional classifiers and the stacking classifier.

Step 3

Set Up Cross-Validation and Scoring

Establish a Stratified K-Fold cross-validation strategy and define scoring metrics (e.g., recall) for evaluating model performance.

Step 4

Evaluate Models with Different Scalers

Iterate over different scalers and models, creating pipelines that include preprocessing and model training, and collect performance scores through cross-validation.

Step 5

Identify and Display Best Model and Scaler

Determine the best-performing model and scaler from the results, create a final pipeline using these selections, and display the optimized pipeline configuration.

Modeling

Best Model Benchmark Initialization

- Metric: Recall
- Model: AdaBoost
- Scaler: Robust
- Mean Score: 0.54
- Std Dev: 0.04

Modeling

AdaBoost Algorithm

AdaBoost, short for Adaptive Boosting, is an ensemble learning algorithm designed to enhance the performance of weak classifiers by combining them into a stronger model. It works by sequentially training a series of classifiers, each focusing on the misclassified samples from the previous ones, thereby improving the model's ability to handle difficult cases. The final prediction is made by aggregating the weighted outputs of all the classifiers, making AdaBoost effective for various classification tasks while also providing insights into feature importance.

Step 1

Initialize Weights

Assign equal weights to all training samples to ensure each sample contributes equally to the learning process.

Step 2

Train Base Classifier

Train a simple base classifier (e.g., a shallow decision tree) on the weighted training data and evaluate its performance to identify misclassified samples.

Step 3

Update Weights

Adjust the weights of the training samples by increasing the weights of misclassified samples and decreasing the weights of correctly classified ones, emphasizing difficult cases for the next classifier.

Step 4

Combine Classifiers

Train multiple base classifiers iteratively, combining their predictions by weighting them according to their accuracy to create a strong ensemble model.

Step 5

Make Predictions and Evaluate

Use the combined model to make predictions on new data, and assess its performance using various metrics, while also providing insights into feature importance.

AdaBoost

Parameters

estimator

To specify the type of base classifier used in the AdaBoost ensemble. The default classifier estimator in AdaBoost is a decision tree classifier with a maximum depth of 1

n_estimators

Controls the number of weak classifiers to be combined in the ensemble. More classifiers potentially improving performance but also increasing the risk of overfitting.

learning_rate

Adjusts the contribution of each weak classifier to the final model. A larger learning rate can speed up learning but may lead to overfitting.

Modeling

Searching The Best Model

Step 1

Identify Categorical Features for SMOTENC

Extract the indices of categorical features from the training dataset to prepare for the application of SMOTENC.

Step 2

Define Base Pipeline Steps

Create a base pipeline that includes preprocessing (with the best scaler), a placeholder for a resampler, and a classifier (AdaBoostClassifier).

Step 3

Set Parameter Distributions

Define a parameter distribution for the randomized search, including choices for resampling techniques, hyperparameters for AdaBoost, and options for the base classifier.

Step 4

Configure RandomizedSearchCV

Set up a RandomizedSearchCV instance to perform hyperparameter tuning, using recall as the scoring metric and stratified cross-validation.

Step 5

Fit Model and Output Results

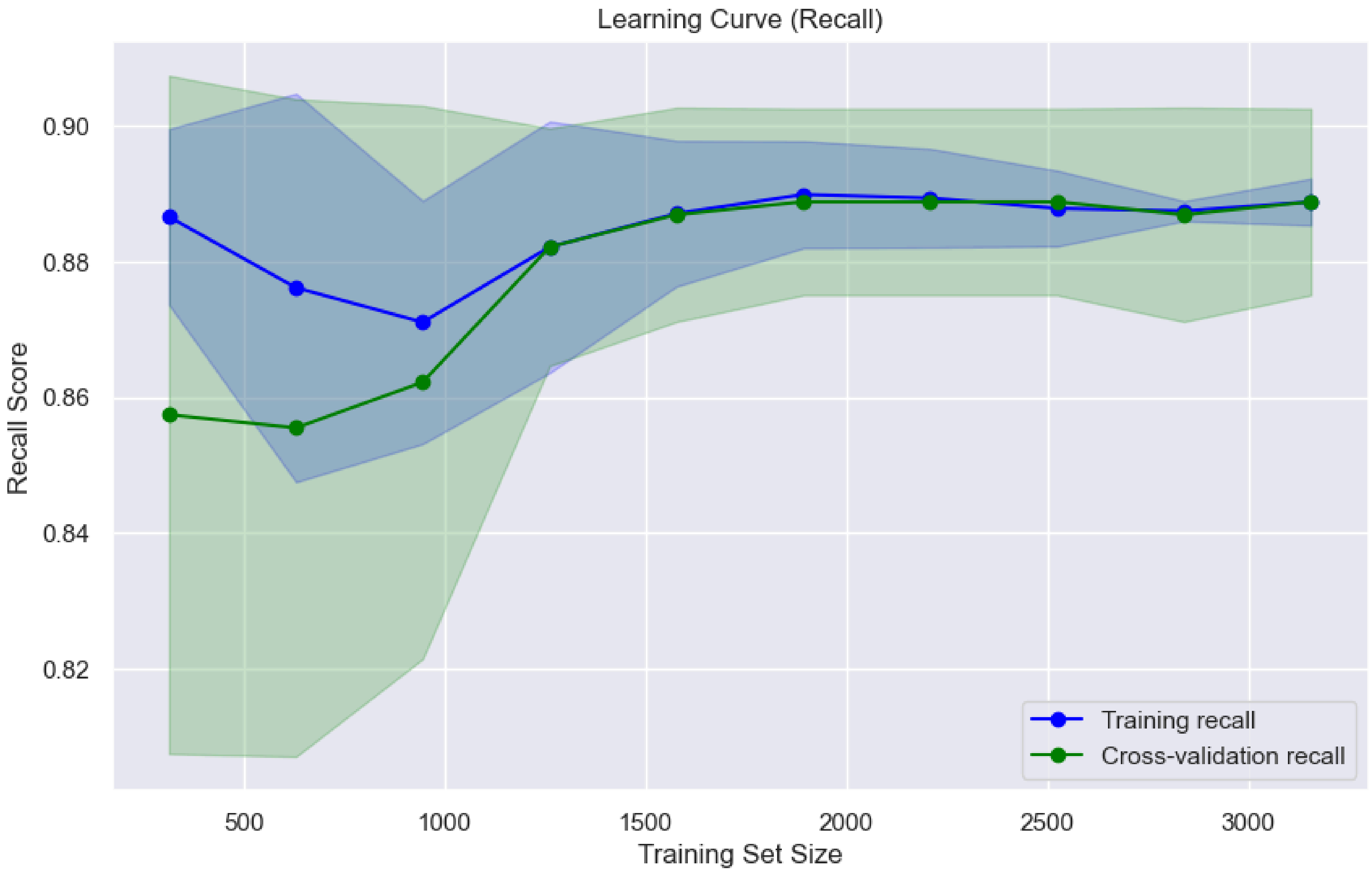
Fit the RandomizedSearchCV on the training data to find the best hyperparameters, and display the best pipeline with the tuned parameters.

Modeling

Best AdaBoost Model After Hyperparameter Tuning

- Best Hyperparameters:
  - 'resampler': RandomOverSampler(random_state=42),
  - 'classifier__n_estimators': 250,
  - 'classifier__learning_rate': 0.01,
  - 'classifier__estimator': DecisionTreeClassifier(max_depth=1),
- Recall Score: 0.92 → Increases around 70% From Base Model

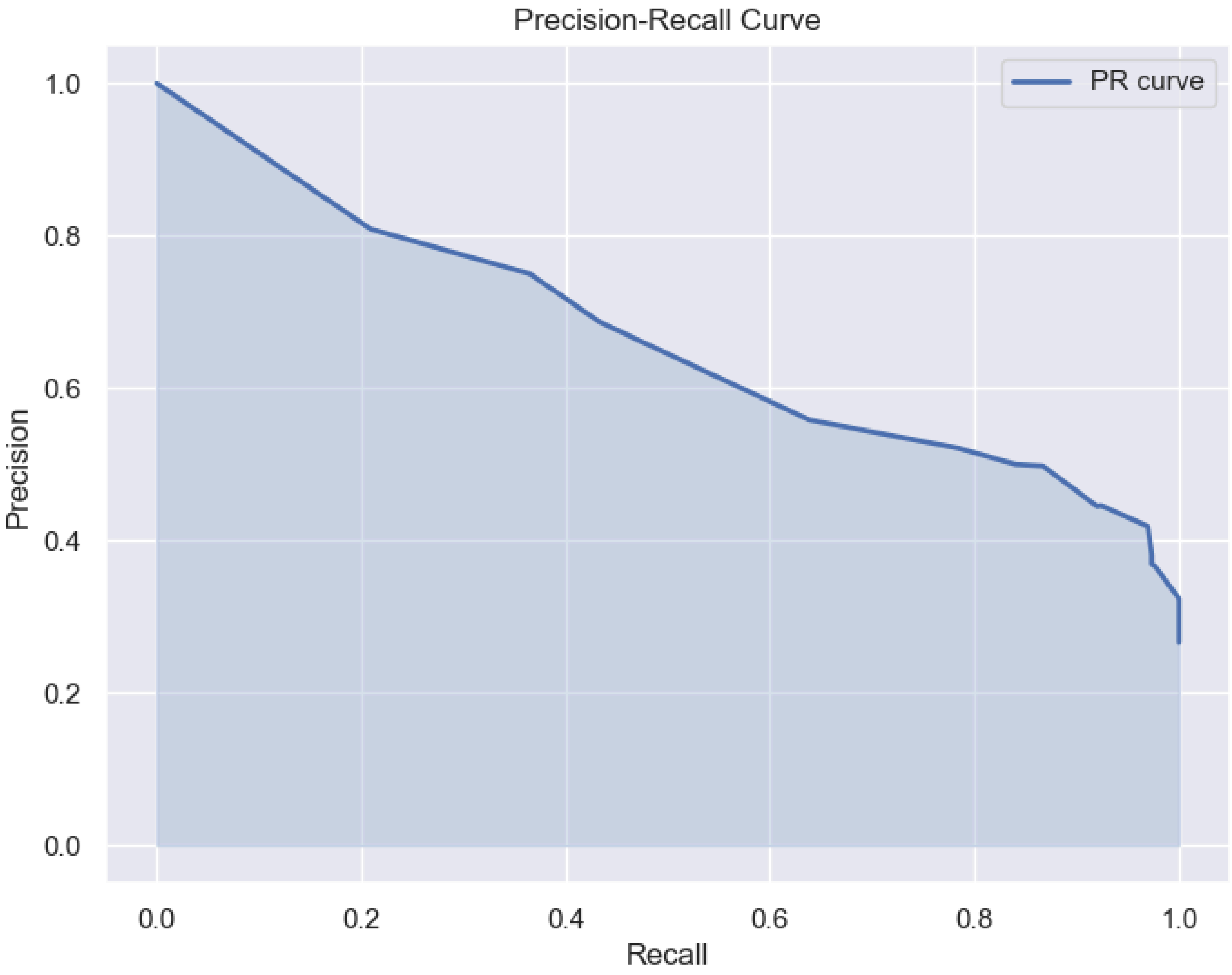
Modeling



Recall Learning Curve

- Training Recall Stability:
  The training recall scores remained consistently high (around 0.88 to 0.89) across varying training sizes, with low variability indicated by a narrow standard deviation.
- Cross-Validation Improvement:
  Initially lower cross-validation recall scores improved and converged towards the training recall as training size increased, showing reduced variability with more data, indicating a well-fitted model.
- Model Performance and Capacity:
  The model demonstrated good performance by minimizing false negatives, and recall scores stabilized at larger training sizes, suggesting that the model's capacity was reached with the current features and algorithm, with no significant improvements from additional data.

Modeling



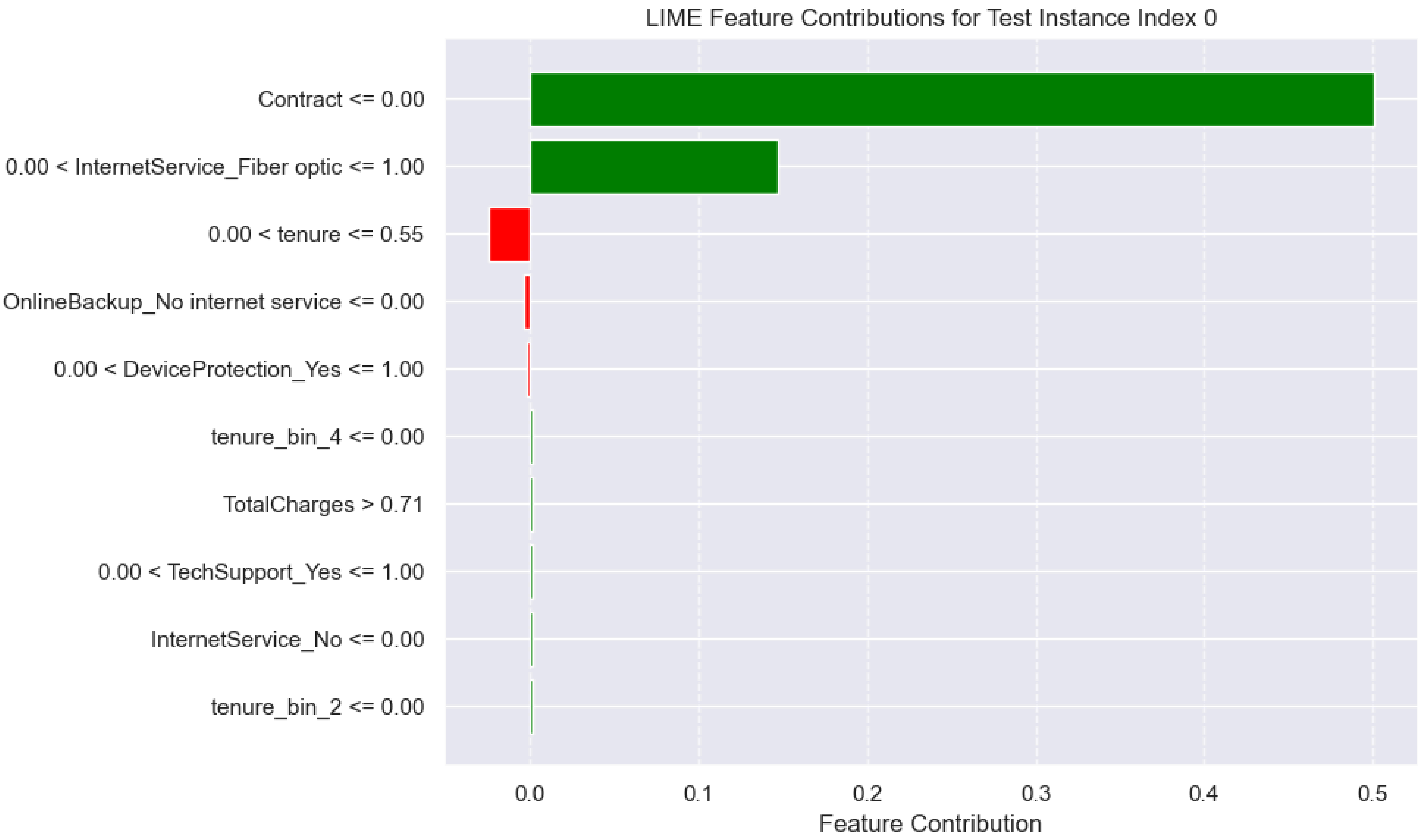
Threshold Tuning

- ○ Best threshold by Recall score: 0.1192
- ○ Recall score at this threshold: 1.0000
- ○ Decision: “No Go”; Too good to be true

5. Evaluation

- ○ Feature Importance Using LIME Analysis
- ○ Rule Based vs Machine Learning Based

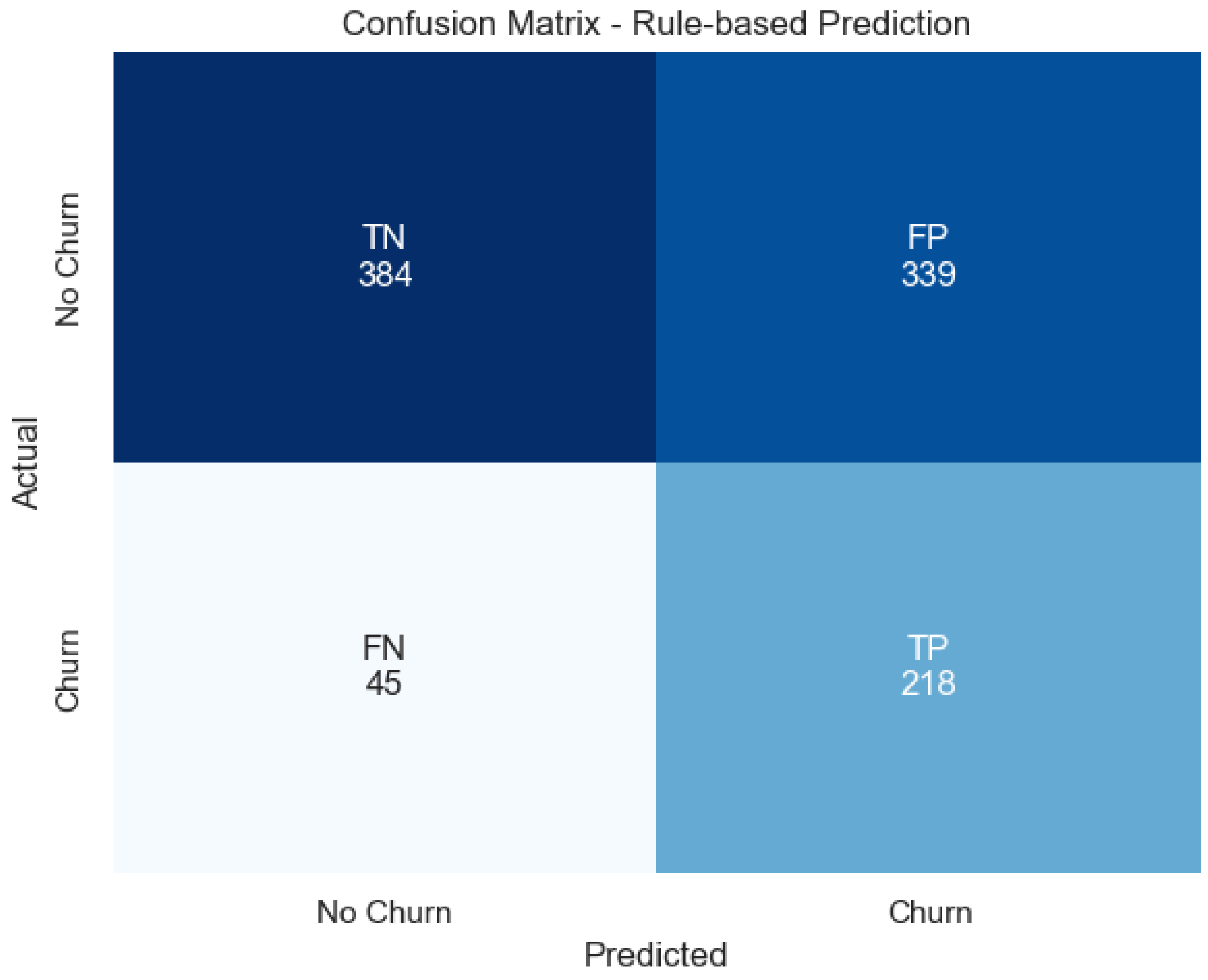
Evaluation

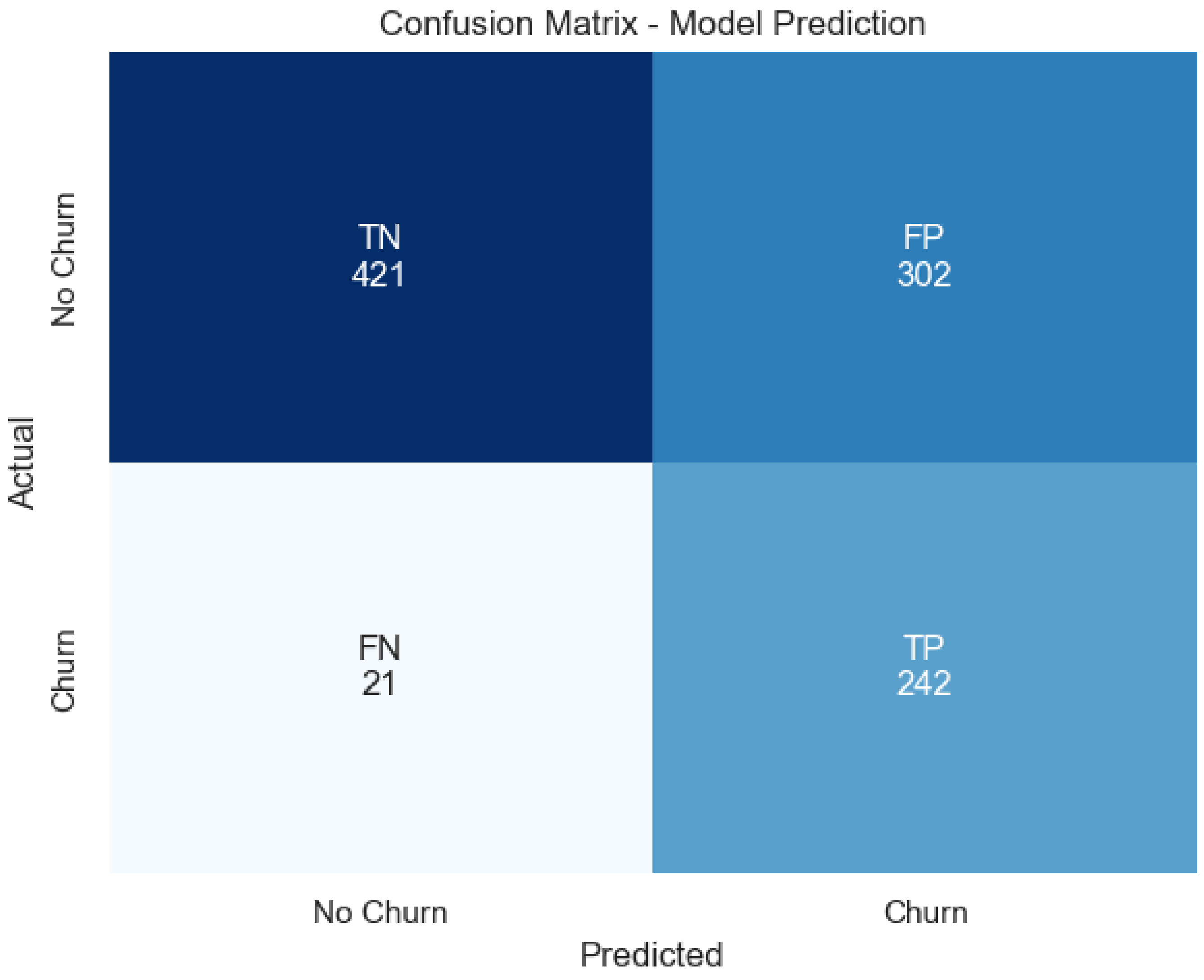


Feature Importances - LIME

- Key Positive Contributors:
  - Contract: Most significant positive impact, crucial for favorable predictions.
  - InternetService_Fiber optic: Also positively influences the prediction.
- Key Negative Contributor:
  - Tenure (0.00 < tenure <= 0.55): Has a negative contribution, suggesting that this range decreases the likelihood of the predicted outcome.

Evaluation





Recall Score Rule: 0.8289
Recall Score ML: 0.9202
Differences: 9.13%

Rule-Based Scenario:
Loss from FN: $9,000
Cost from FP: $16,950
Total Monthly Loss: $25,950
Total Annual Loss: $311,400

Machine Learning Scenario:
Loss from FN: $4,200
Cost from FP: $15,100
Total Monthly Loss: $19,300
Total Annual Loss: $231,600

Differences: $79,800
Percentages savings:

Rule Based vs Machine Learning Based

- With Machine Learning Model, recall score increases 9.13 % compared to the rule based.
- The company also can saves 25.63% based on Machine Learning Model instead of rule based.

6. Deployment

- Joblib Deployment
- Model Limitations

Deployment

Joblib Deployment File

models\best_tuned_pipeline.joblib

Limitations 1 - Data Dependence and Bias

The model's accuracy is heavily reliant on the quality and representativeness of the training data, which may contain biases that affect predictions and limit its generalizability.

Limitations 2 - Pattern Recognition and Adaptability

The model's ability to identify complex patterns or rare events is constrained by the selected features. Additionally, it may not adapt well to changes in customer behavior or external factors unless regularly updated.

Limitations 3 - Overfitting and Interpretability

There is a risk of overfitting or underfitting if the model is not properly tuned. Furthermore, the model may have limited interpretability, and the evaluation metrics used might not capture all aspects of its performance effectively.

7. Conclusion and Recommendations

- Conclusion
- Business Recommendations
- Model Recommendations

Conclusion and Recommendations

Conclusions

- **Contract has the strongest positive contribution to churn prediction** — customers with certain contract types are more likely to churn.
- **InternetService_Fiber optic also positively contribute** to churn risk.
- **Tenure has a negative contribution** — shorter tenure customers are more likely to churn.
- **The Machine Learning Scenario significantly increases recall score around 9%** and reduces False Negatives and False Positives losses, **saving nearly $80,000 annually**, a quarter of the total losses under the Rule-Based system.

Conclusion and Recommendations

Business Recommendations

- **Contract Optimization:** Promote contract types that enhance customer retention and satisfaction, offering incentives to encourage selection.
- **Internet Service Offerings:** Invest in expanding fiber optic infrastructure and promote its benefits to attract more customers.
- **Customer Retention Strategies:** Develop targeted strategies for customers with short tenures, including personalized outreach and loyalty programs.
- **Feature Awareness Campaigns:** Educate customers on the advantages of additional features like online backups and device protection.
- **Customer Feedback Loop:** Implement feedback mechanisms to refine service offerings and address customer pain points.
- **Analyze Customer Segments:** Conduct analysis to understand diverse customer needs, enabling tailored marketing and services.
- **Retention Metrics Monitoring:** Regularly track key metrics related to contracts and service usage to identify trends and adjust strategies.

Conclusion and Recommendations

Model Recommendations

- **Balance Precision and Recall:**
  Instead of focusing solely on maximizing recall, the company can tune the model to achieve a better balance between precision and recall. This can be done by adjusting the classification threshold or using evaluation metrics like the F1-score that consider both precision and recall.

- **Regular Model Retraining:**
  Customer behavior and market conditions change over time. Regularly retraining the model with new data ensures that it adapts to recent trends and maintains accuracy.

“It is the greatest truth of our age: Information is not knowledge.”
— Caleb Carr

Thank You :D