Telco Company

Why Do Customers Leave?

An End-to-End Data Science Project to Predict Customer Churn

Daffa Kaisha P. C. | Data Science Portfolio Project

The High Cost of Customer Churn

Problem:
Customer churn is a critical revenue leakage point for subscription-based businesses like telecommunications companies.

Project Objectives:
- Perform Exploratory Data Analysis (EDA)
- Conduct robust Data Pre-processing.
- Build and tune multiple classification models.
- Evaluate and select the best-performing model.

Understanding the Data

Data source: Kaggle's "Telco Customer Churn" dataset, initially loaded into a MySQL database.

The dataset contains over 7.000 records of customer data and 21 features.

Dataset key information:

- Demographic: Gender, Senior Citizen Status.
- Account Info: Tenure, Contract Type, Monthly & Total Charges.
- Services: Phone, Internet, Online Security, etc.
- Target: Churn (Yes/No)

Asking the Right Questions

Goals: Exploring the data to find patterns and form hypothesis about what drives churn.

My EDA process included:

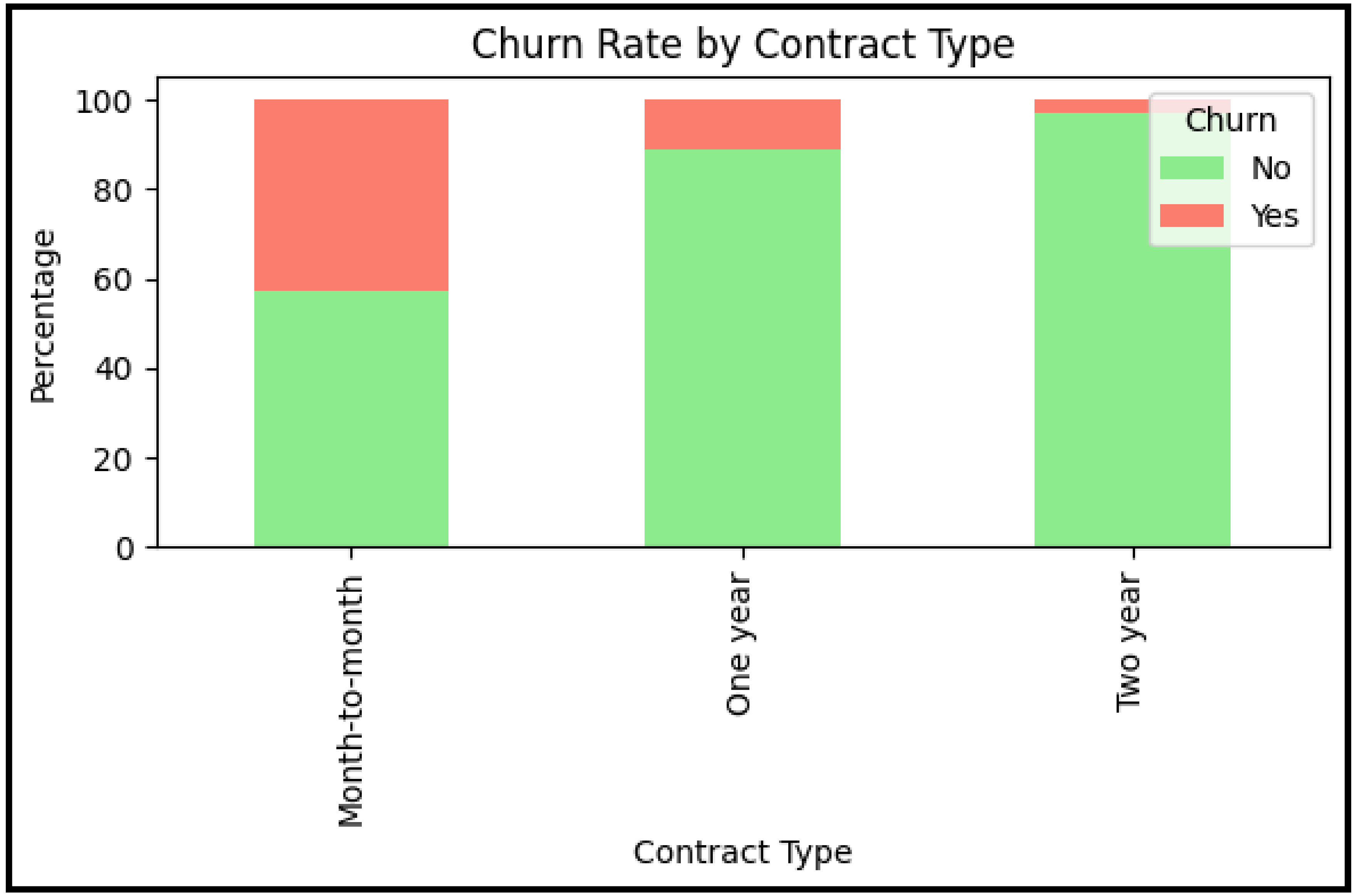
1.Distribution analysis:
Checked the distribution of key variables.

2.Segment analysis
Visualized churn rates across different customer segments

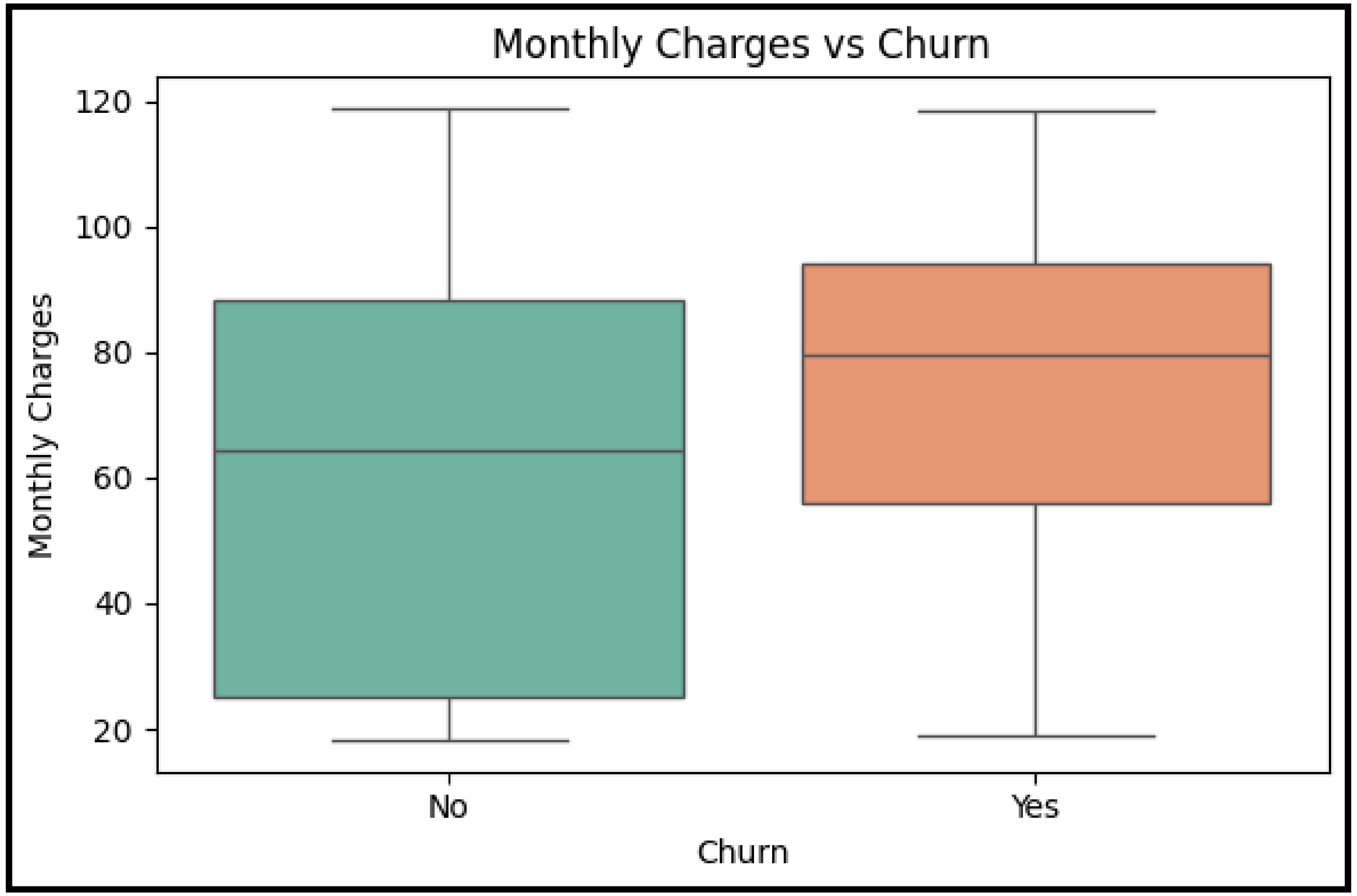
3.Statistical comparison:
Analyzed the average MonthlyCharges for customers who churned versus those who did not.

4.Correlation analysis:
Created a heatmap to check for multicollinearity between features.

Key Insights from EDA



Month-to-month contract type has the Highest Churn Rate!



Churned customers tend to have higher Monthly Charges

What the Data Revealed

Exploratory Data Analysis Key Insights:

1.Imbalance is real

The dataset is imbalanced. Most of customers (73%) did not churn.

2. Contract is King

Month-to-month customers churn the most.

3.Price Matters

Churned customers tend to have higher Monthly Charges

4.Service Type is a Clue

Fiber optic users show a higher churn rate.

Engineering a Smarter Dataset

Built a comprehensive pre-processing function to handle all transformation systematically.

Key Steps:

1.Encoding
Converted binary columns and one-hot encoded multi-category features.
2. Outlier Handling
Capped outliers in numerical columns using the IQR method.
3. Feature Engineering
Created new features to add predictive power (TotalSpent, HasInternet, IsLongContract).

Building a Robust Modeling Pipeline

The Challenge: The model needs balanced, consistently scaled data to learn effectively.

The Solution:

I constructed a Pipeline that automatically performs three crucial steps during training:

1.StandardScaler():

Scale numerical features.

2. SMOTE():

Over-samples the minority class (Churn = ‘Yes’)

3. Classifier():

Trains the machine learning model (Logistic Regression, Decision Tree, Random Forest, XGBoost).

Finding the Best Model

Hyperparameter tuning using Optuna to find the best performing model.

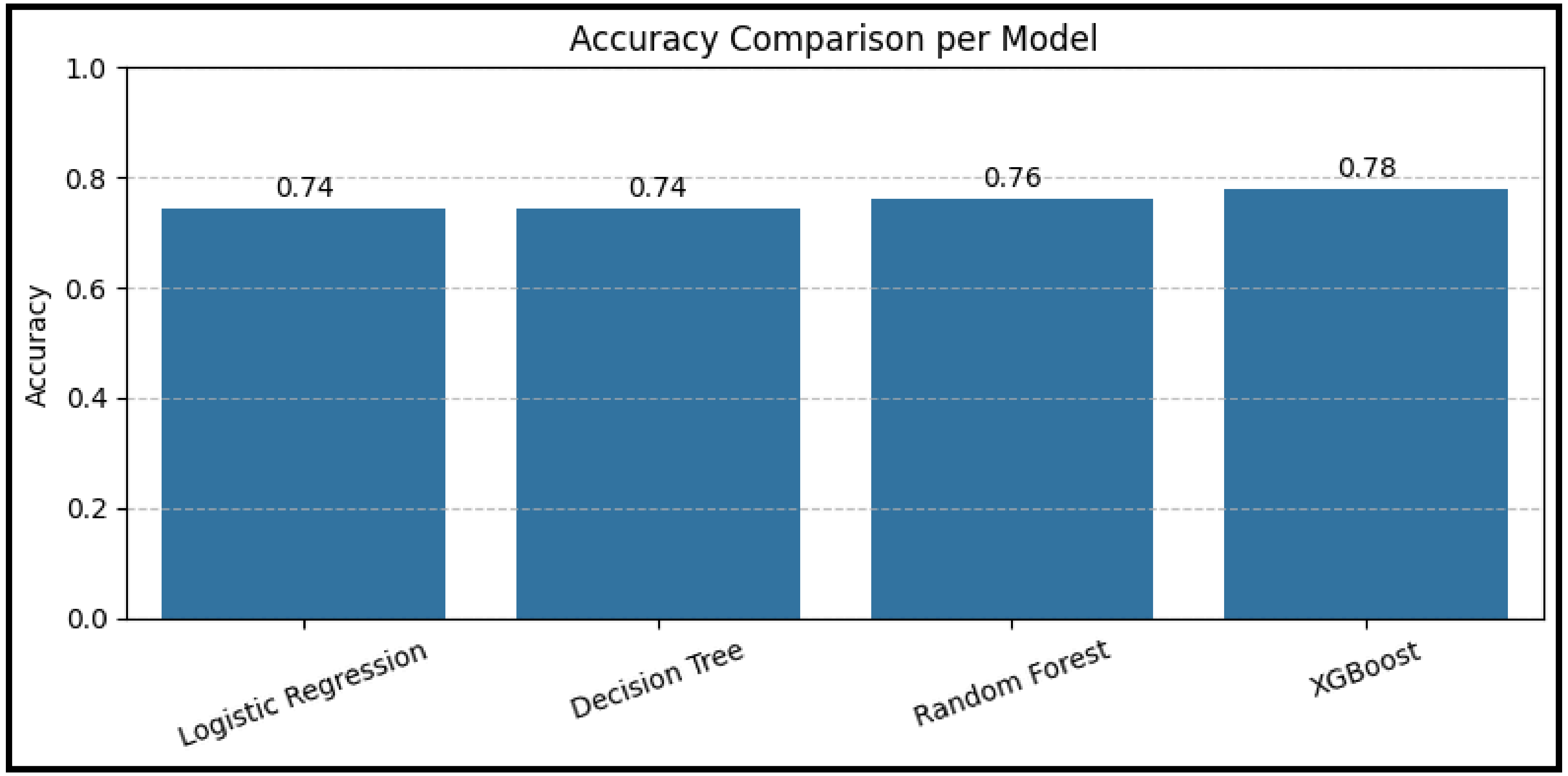
The Process:

1. Defined a “search space” of hyperparameters for all model used.
2. Used Optuna to run 25 trials for each model, and automatically finds the best settings based on AUC-ROC Score.

This ensures each model is performing at its peak.

The Winning Model

After hyperparameter tuning, XGBOost emerged as the best model with a final accuracy of 78% on the test set.



More importantly, the model achieved a Recall of 0.70 for the 'Churn' class, meaning it successfully identified 70% of the customers who were actually at risk of leaving.

From Prediction to Actionable Strategy

Key Actions:

1.Targeted Retention:

Use the model to identify high-risk customers for personalized incentives.

2. Service Review:

Investigate why Fiber Optic customers are churning (price, reliability, support?)

3. Loyalty Programs:

Develop programs to move customers from monthly to long-term contracts.

Key Learnings and Future Work

Key Takeaway:

This project highlighted the power of a structured workflow: from deep EDA and feature engineering to automated hyperparameter tuning with Optuna.

Next Step:

The next step would be to deploy this model using a framework like Flask or FastAPI to create a REST API for real-time churn predictions.

The Tools I Used and How

MySQL

Used for initial data storage and profiling.

Python

Built a robust pipeline using Pandas for data wrangling, Scikit-learn for modeling, applying SMOTE to correct class imbalance and Optuna for automated hyperparameter tuning.

Thanks for Reading!

I'm passionate about using data to solve real-world problems. I'd love to connect or hear your feedback!

Daffa Kaisha Pratama Chandra
daffakpc21@gmail.com
linkedin.com/in/daffakaisha/