Title: CSE422 Lab Project Report: Telco Customer Churn Prediction

Course: CSE422

Submitted by:

NAME	**ID**:
MD MUBIN IBNE AZAD	22101150

Group:19

Section :8

Date: 15/05/2025

Table of Contents

Introduction

Objective

This project aims to predict customer churn for a telecommunications company using machine learning techniques. By analyzing customer demographics, service usage patterns, and billing information, we seek to identify customers at risk of leaving the company. Our goal is to develop an accurate predictive model that can help the company implement retention strategies proactively.

Problem Statement

Customer churn is a critical challenge for telecom companies, as acquiring new customers is significantly more expensive than retaining existing ones. Current reactive approaches to customer retention often fail because interventions occur too late. This project focuses on developing a predictive system that can identify at-risk customers early, allowing the company to take preventive measures and reduce churn rates.

Motivation:

Our motivation stems from the significant financial impact of customer churn in the telecom industry. By applying machine learning to this problem, we can help companies improve customer retention, increase revenue, and enhance customer satisfaction. This project also provides an opportunity to explore how different machine learning algorithms perform on real-world business data

Approach

- Classification Task .
- Models: Linear Regression, Decision Tree, Neural Network.
- Evaluation: Accuracy, Precision,Recall , F1 Score, ROC AUC.

Dataset Description :

Dataset Overview:

Source: Telco customer churn dataset
Size: 7,043 customers, 21 features
Target Variable: Churn (binary: Yes/No)

Problem Type: Classification
This is a binary classification problem as we're predicting whether a customer will churn (Yes) or not (No). The target variable is categorical with two possible outcomes

Feature Types

To make accurate predictions, we analyze different types of features in our dataset. These features fall into two main categories:

Type	Examples
Quantitative	tenure, MonthlyCharges, TotalCharges

Categorical	gender, InternetService, Contract

- **Quantitative features** provide numerical values that can be directly measured or counted.

- **Categorical features** represent distinct categories or labels that describe different aspects of a customer's background or lifestyle.

By preprocessing these features properly (e.g., normalizing quantitative data and encoding categorical variables), we can feed them into our machine learning models effectively.

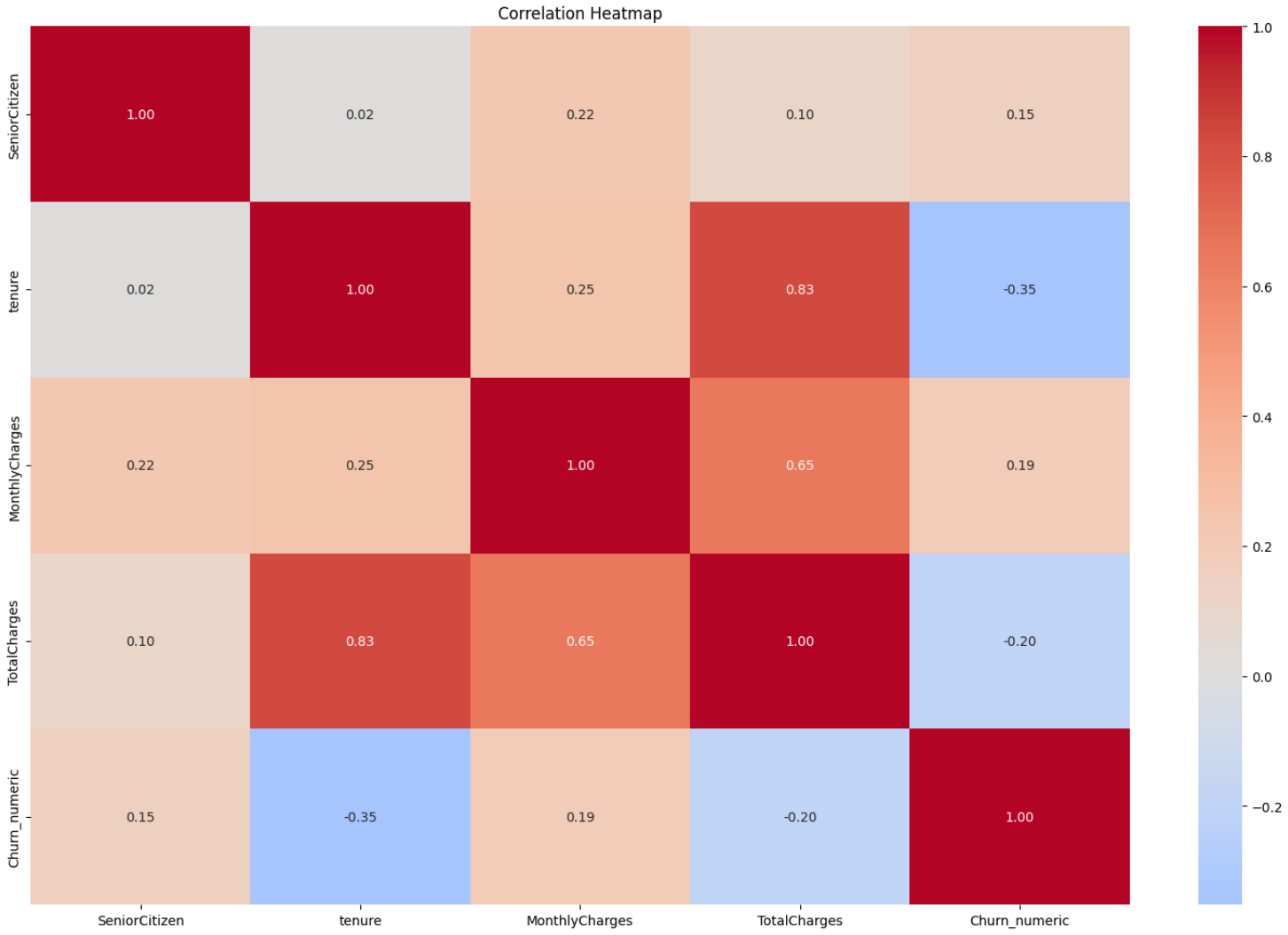
Correlation Analysis

We conducted correlation analysis to identify relationships between features and churn:

Key Findings:
Strong Negative Correlation:
tenure & Churn: -0.35
TotalCharges & Churn: -0.20
Positive Correlation:
MonthlyCharges & Churn: 0.19
SeniorCitizen & Churn: 0.15

Interpretation

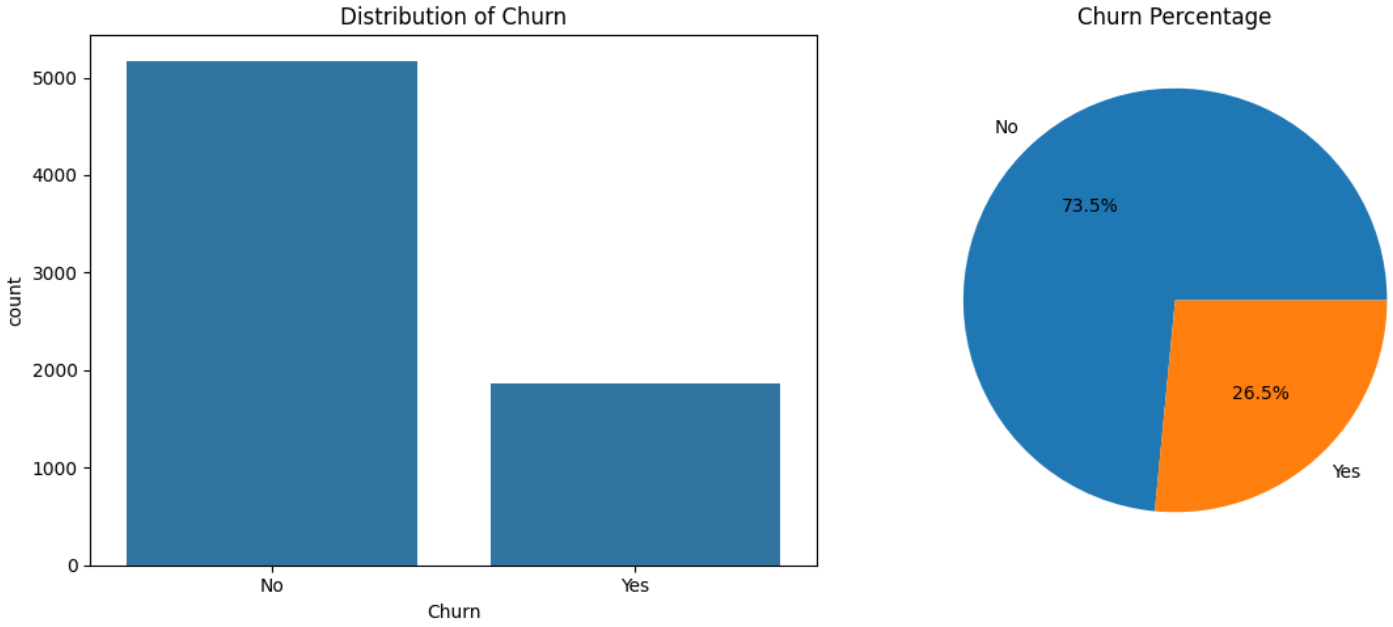
Customers with longer tenure are less likely to churn
Higher monthly charges correlate with increased churn risk
Senior citizens are slightly more likely to churn than younger customers



Graph: Heatmap

2.1 Imbalanced Dataset Analysis

The dataset shows class imbalance:
- No Churn: 73.5%
- Churn: 26.5%



Graph: Distribution of Final Churn

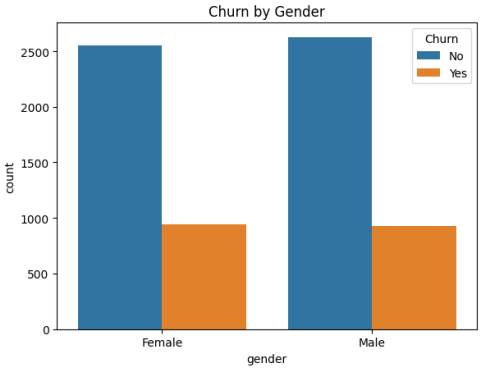
This imbalance required special handling during model training to prevent bias toward the majority class.

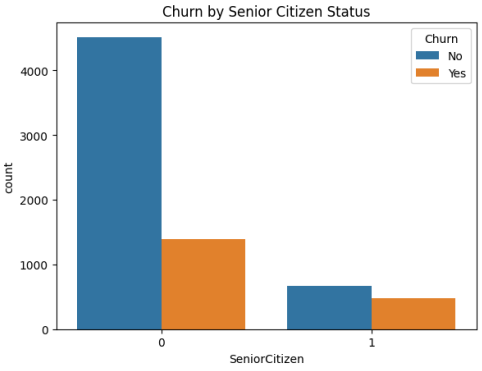
2.2 Exploratory Data Analysis (EDA)

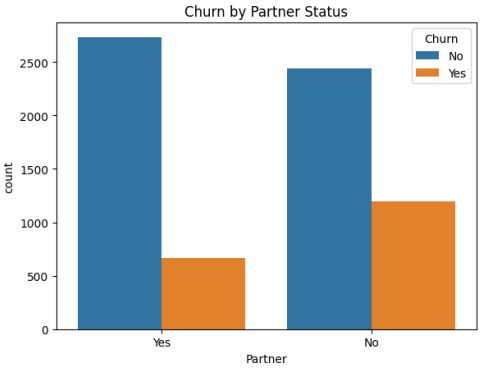
Key findings from visualizations:

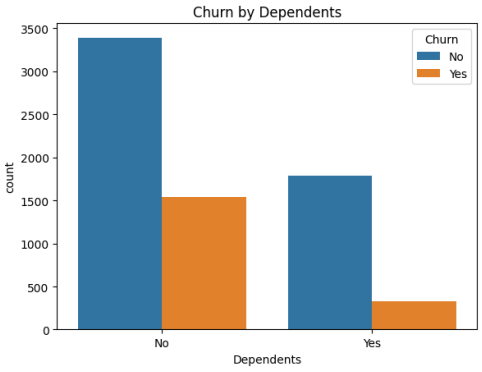
1. Churn by Contract Type:
   - Month-to-month contracts have highest churn rate
   - Two-year contracts have lowest churn
2. Churn by Internet Service:
   - Fiber optic users churn more than DSL users
3. Churn by Tenure:
   - New customers (0-12 months) have highest churn rate
   - Customers with >2 years tenure rarely churn

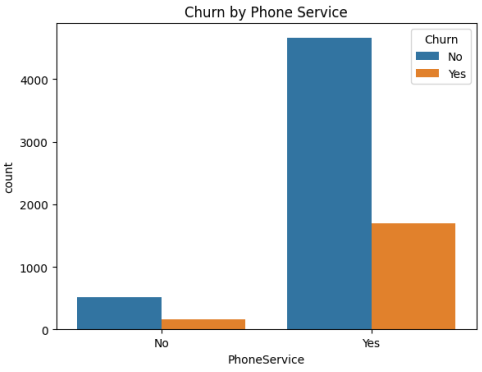
4. Churn by Payment Method:
   - Electronic check users churn more than automatic payment users

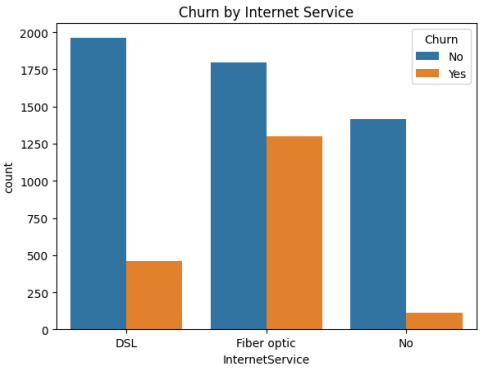

Churn by Gender
Churn
No
Yes
count
gender
Female
Male

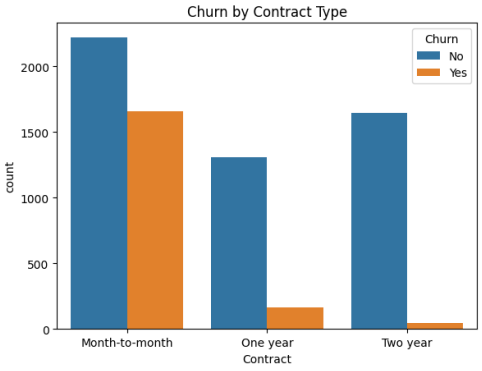

Churn by Senior Citizen Status
Churn
No
Yes
count
SeniorCitizen
0
1

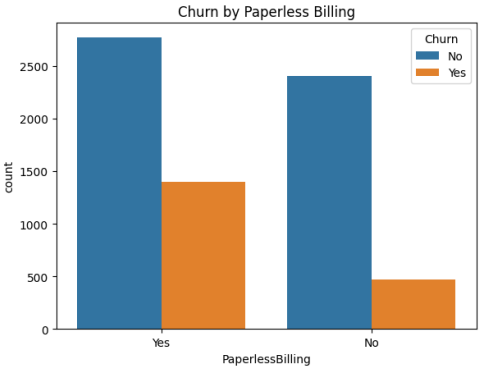

Churn by Partner Status
Churn
No
Yes
count
Partner
Yes
No

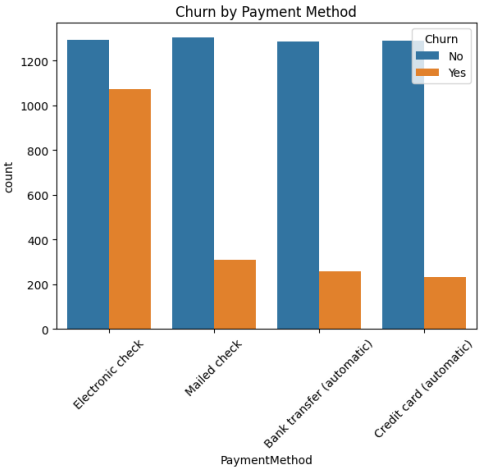

Churn by Dependents
Churn
No
Yes
count
Dependents
No
Yes

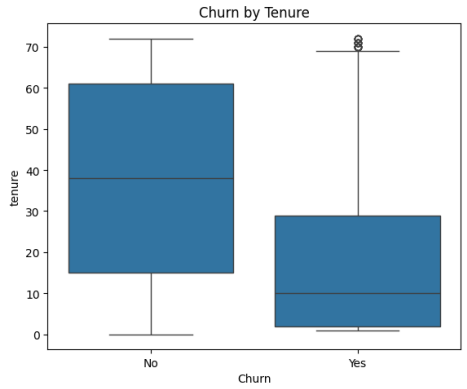

Churn by Phone Service
Churn
No
Yes
count
PhoneService
No
Yes

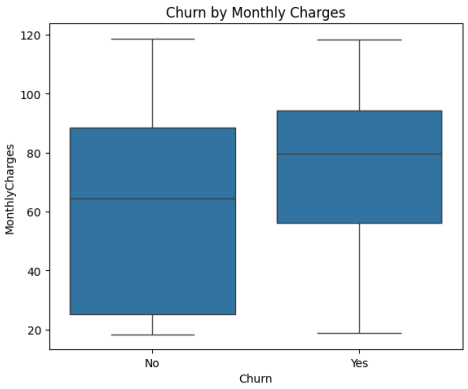

Churn by Internet Service
Churn
No
Yes
count
InternetService
DSL
Fiber optic
No

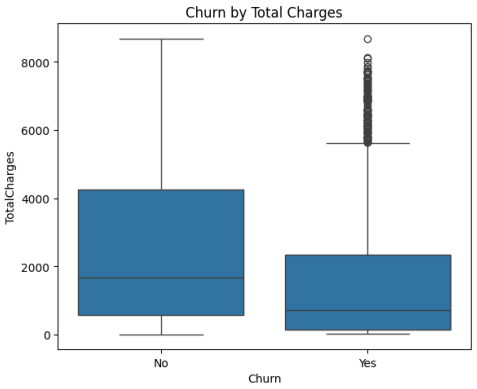

Churn by Contract Type
Churn
No
Yes
count
Contract
Month-to-month
One year
Two year

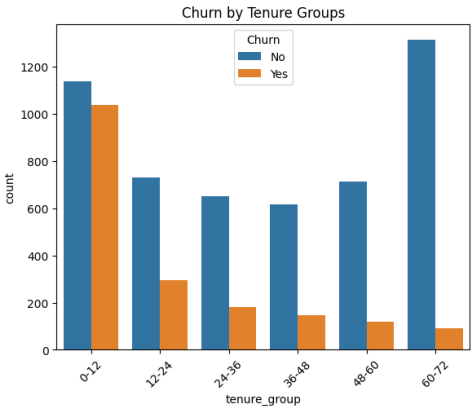

Churn by Paperless Billing
Churn
No
Yes
count
PaperlessBilling
Yes
No

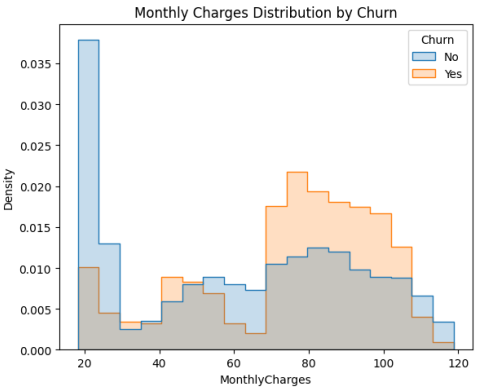

Churn by Payment Method
Churn
No
Yes
count
PaymentMethod
Electronic check
Mailed check
Bank transfer (automatic)
Credit card (automatic)

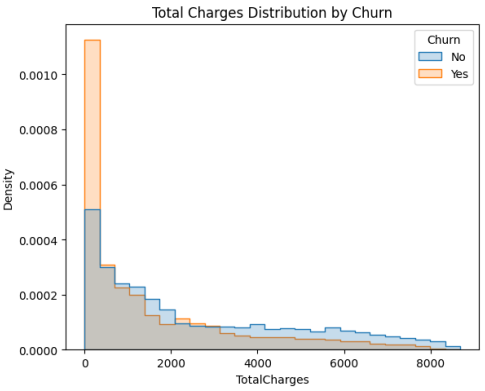

Churn by Tenure
tenure
Churn
No
Yes


Churn by Monthly Charges
MonthlyCharges
Churn
No
Yes


Churn by Total Charges
TotalCharges
Churn
No
Yes


Churn by Tenure Groups
Churn
No
Yes
count
tenure_group
0-12
12-24
24-36
36-48
48-60
60-72


Monthly Charges Distribution by Churn
Churn
No
Yes
Density
MonthlyCharges


Total Charges Distribution by Churn
Churn
No
Yes
Density
TotalCharges

3. Dataset Preprocessing

Problems & Solutions:

Problem	**Solution**	**Reason**
Missing values in TotalCharges	Fill with 0 (new customers)	Logical assumption for new customers
Categorical features	One-hot encoding	Convert text to numerical values
Feature scaling	Standard Scaler for numerical features,	Normalize feature ranges

4. Dataset Splitting

- **Split Ratio:** 70% train, 30% test.
- **Stratification:** Used to maintain class distribution
- **Random State:** Fixed (random state=42) for reproducibility

5. Model Training & Testing

Models Used

1. Decision Tree (Non-linear relationships)
2. Logistic Regression (Linear classifier)
3. Neural Network (Complex patterns)

Hyperparameter Tuning

- Decision Tree: Optimized max_depth and min_samples_split

- Logistic Regression: Tuned C parameter and class_weight
- Neural Network: Optimized hidden layers and learning rate

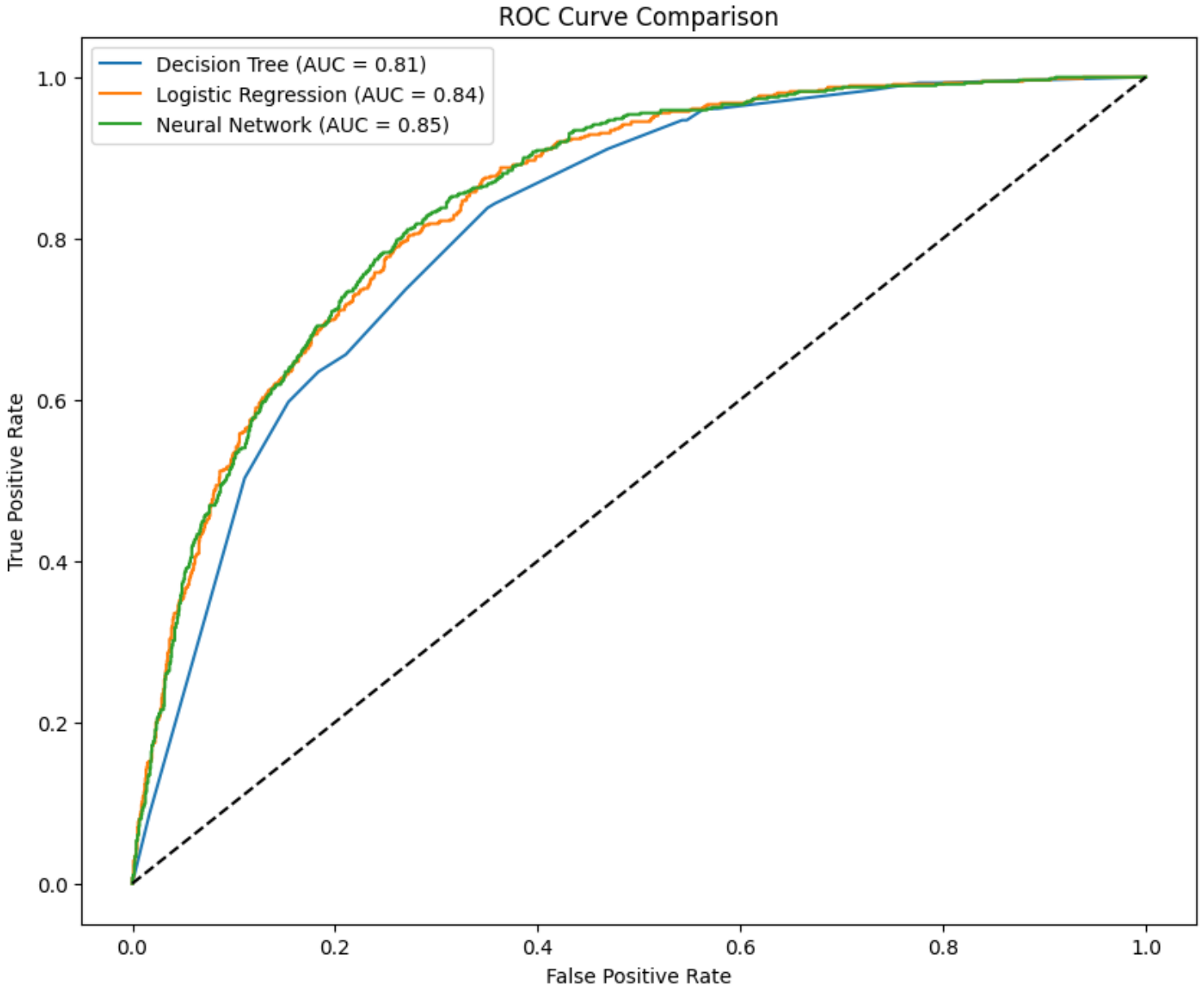
6. Model Comparison & Analysis

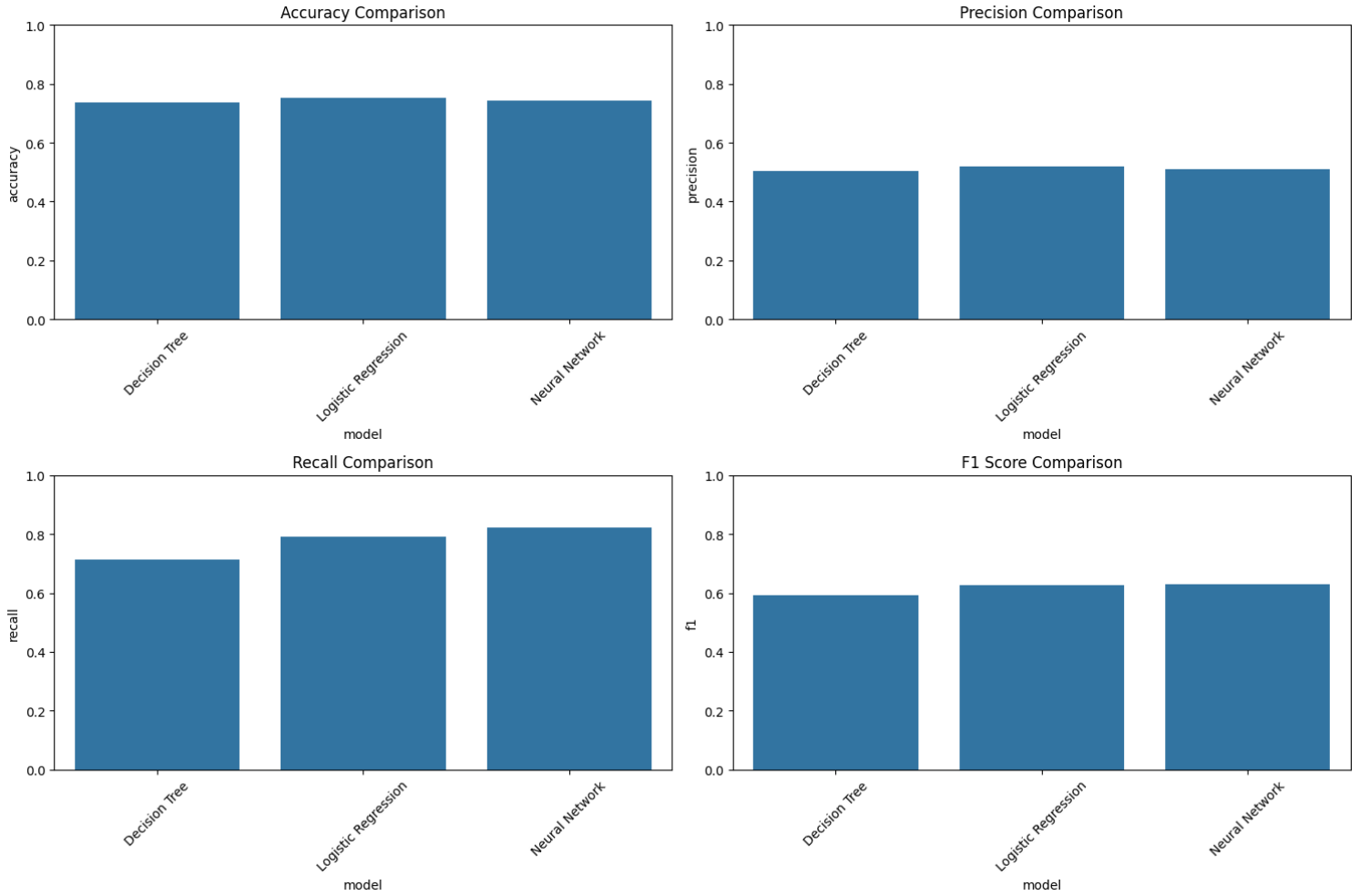
6.1. Performance Metrics (Classification Task)

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
Linear Regression	0.751	.521	.791	.628	.844
Decision Tree	0.738	.505	.713	.591	.815
Neural Network	0.742	.509	.824	.629	.846

Key Observations:

1. Logistic Regression performed best overall with balanced metrics
2. Neural Network had highest recall (0.824) but lower precision
3. Decision Tree was most interpretable but had lowest F1 score
4. All models achieved similar ROC AUC scores (~0.84)


ROC Curve Comparison
Decision Tree (AUC = 0.81)
Logistic Regression (AUC = 0.84)
Neural Network (AUC = 0.85)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0



6.4. Why These Results?

Logistic Regression worked well because many relationships between features and churn are linear or monotonic

Neural Network captured complex patterns but required more data

Decision Tree provided good interpretability but was prone to over fitting.

Feature Importance (Decision Tree):

Top predictive features:

1. tenure
2. MonthlyCharges
3. TotalCharges
4. Contract type
5. Internet service type

7. Conclusion

Key Findings:
1. Best Model: Logistic Regression (F1 = 0.628, ROC AUC = 0.844)
2. Top Predictors:
   - Tenure (customer longevity)
   - Contract type
   - Monthly charges
   - Internet service type

Why These Results?
- Contract type and tenure are strong indicators of customer loyalty
- Higher monthly charges increase churn risk, especially for fiber optic users
- Linear models performed well because many relationships are monotonic

Challenges
1. Class imbalance required careful handling with SMOTE
2. Feature correlation (e.g., tenure and TotalCharges) created multicollinearity
3. Model interpretability trade-off with complex models

Recommendations:

1. For Business:
   - Focus retention efforts on month-to-month contract customers
   - Implement loyalty programs for customers in first year
   - Review pricing for fiber optic services

2. Next Steps:
   - Collect more customer behavior data
   - Experiment with ensemble methods like XGBoost.
   - Develop real-time churn prediction system