
CSE422 Lab Project Report

Project Title: Telecom Churn Prediction using Machine Learning
Course: CSE422
Instructor: Mahjabin Chowdhury & Mazed Hossain Parag
Date:05.13.2025
Your Name: Ayhan Arash Tasin-(24241122) & Nihad Hasan NIloy-(24241124)

Table of Contents

---

1. Introduction

This project aims to predict whether a customer will churn (leave the service) based on their interaction with a telecom company. The goal is to help the telecom company identify at-risk customers and take preventive measures to reduce churn. This is a **classification** problem where the target variable is churn (1 = Churned, 0 = Stayed). The motivation behind this project stems from the importance of customer retention in the telecom industry and the need for efficient customer management strategies.

2. Dataset Description

The dataset used is a telecom churn dataset that contains various customer attributes. The dataset includes both categorical and numerical features, representing customer demographics, service usage, and account information.

- **Features in the dataset**:
  The dataset consists of 21 features, including *Tenure*, *MonthlyCharges*, *TotalCharges*, *PaymentMethod*, *InternetService*, and several others related to customer services.

- **Problem Type**:
  This is a **classification problem** since the target variable *Churn* is categorical, with two classes (Churned vs Stayed).

- **Number of Data Points**:
  The dataset contains 7,043 data points, representing individual customer information.

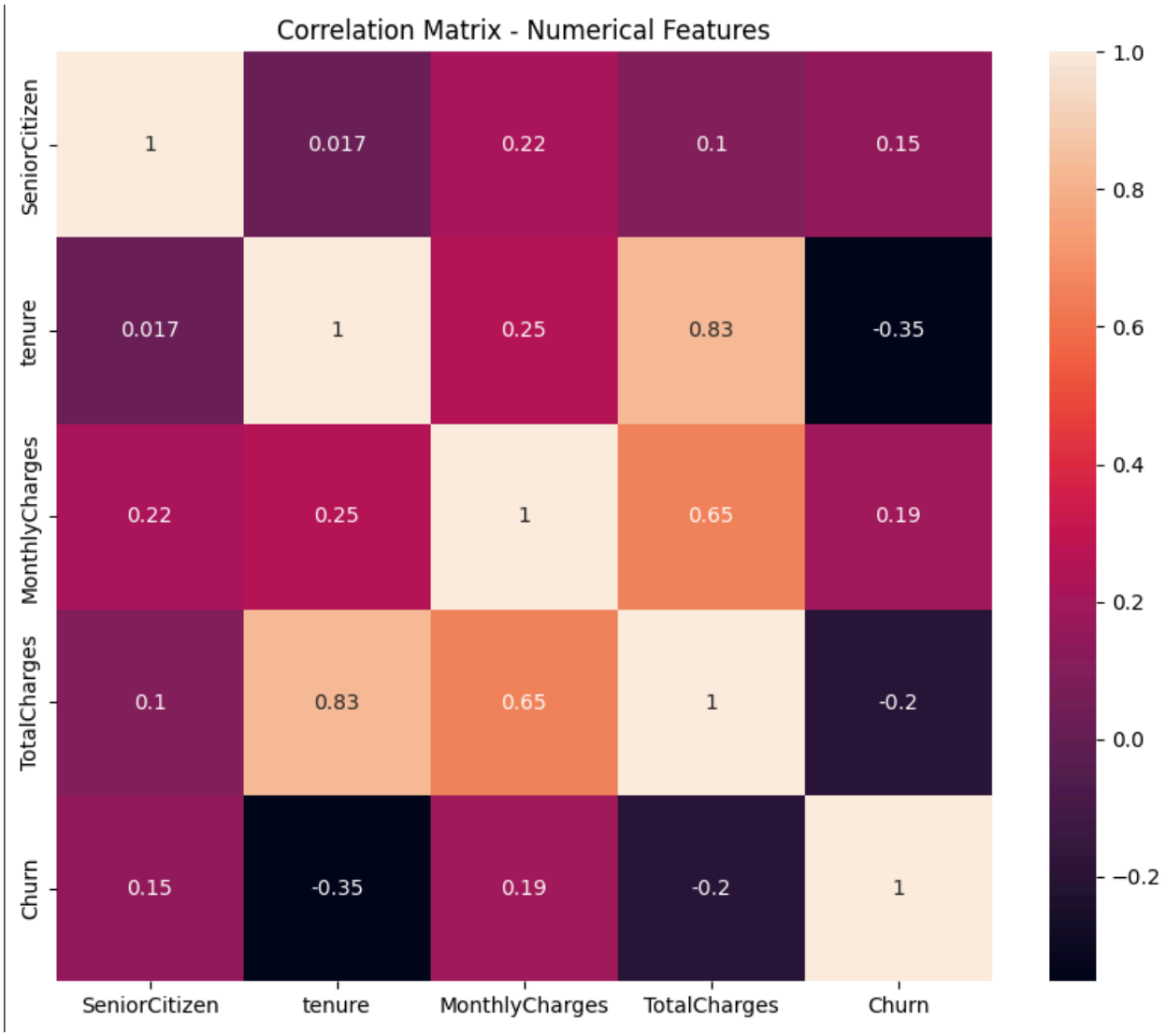
- **Feature Types**:

  - **Quantitative Features**: *Tenure*, *MonthlyCharges*, *TotalCharges*

  - **Categorical Features**: *PaymentMethod*, *InternetService*, *Contract*, and others.

Correlation of Features:

A correlation heatmap was plotted to explore the relationships between numerical features. The heat-map revealed high correlations between *Tenure* and *TotalCharges*, indicating that as customers stay longer, they tend to accumulate higher total charges.



The heatmap can show which features are strongly correlated with the target variable (Churn). For example, a strong correlation between MonthlyCharges and Churn may indicate that customers with higher charges are more likely to churn. Features like Tenure may have a negative correlation with Churn, implying that long-term customers

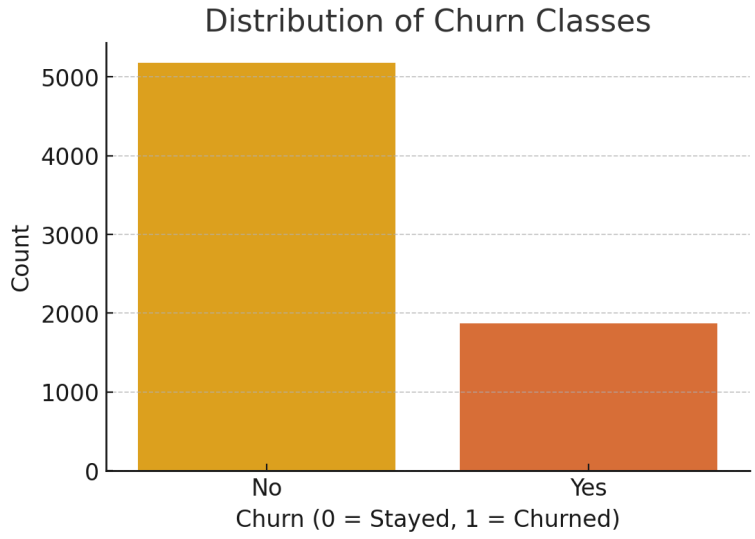
are less likely to leave. Understanding these relationships helps in identifying which features are most important for predicting customer churn.

3. Imbalanced Dataset

The target variable, *Churn*, is imbalanced, with a larger proportion of customers staying rather than churning.

- **Class Distribution**:
  The distribution of *Churn* shows:
    - *Stayed (0)*: 73.46%
    - *Churned (1)*: 26.54%

The imbalance in the dataset can lead to biased predictions. To visualize this, a bar chart is used to display the distribution of the classes:



Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was performed to gain a deeper understanding of the data, identify important relationships, and reveal patterns that might help improve model performance.

Correlation Analysis

A correlation matrix was generated to explore the relationships between numerical features. The correlation heatmap revealed that:

- **Tenure** and **TotalCharges** are highly correlated, with **TotalCharges** increasing as **Tenure** increases. This makes sense as longer-tenured customers tend to accumulate higher charges.

- **MonthlyCharges** has a moderate correlation with **TotalCharges**, which is expected since monthly charges contribute to the total charges over time.

The correlation matrix was visualized using the `seaborn` library:

- sns.heatmap(df.corr(), annot=True, cmap='coolwarm')

The heatmap helps to identify which features are highly correlated, and we can decide whether to drop redundant features (e.g., **TotalCharges**) to reduce multicollinearity.

Distribution of Key Features

For further insights, the distributions of key numerical features were plotted. The following observations were made:

- **Tenure**: Most customers have relatively short tenures, with a few long-term customers. The distribution is right-skewed.

- **MonthlyCharges**: The distribution shows a higher concentration of customers paying lower monthly charges, indicating that many customers are on basic plans.

- **TotalCharges**: Similar to MonthlyCharges, this feature also shows a skewed distribution with most values concentrated on the lower end.

Class Distribution by Churn

The target variable, **Churn**, was analyzed by plotting a bar chart to visualize the

imbalance in the dataset:

The **Stayed** class (0) significantly outnumbers the **Churned** class (1), with **73.46%** of customers staying and **26.54%** churning. This imbalance needs to be addressed, possibly through oversampling or other balancing techniques.

Relationship Between Categorical Features and Churn

The relationships between categorical features and the target variable were analyzed using **count plots** to determine how different categories affect churn rates. The following observations were made:

- **InternetService**: Customers with fiber optic internet services tend to have a higher churn rate than those with DSL or no internet.

- **Contract Type**: Customers on month-to-month contracts have a much higher churn rate compared to those on one- or two-year contracts.

Relationship Between Numerical Features and Churn

Numerical features were analyzed by comparing the distributions of these features for customers who stayed vs. those who churned. Box plots were used to visualize the spread of data for each group:

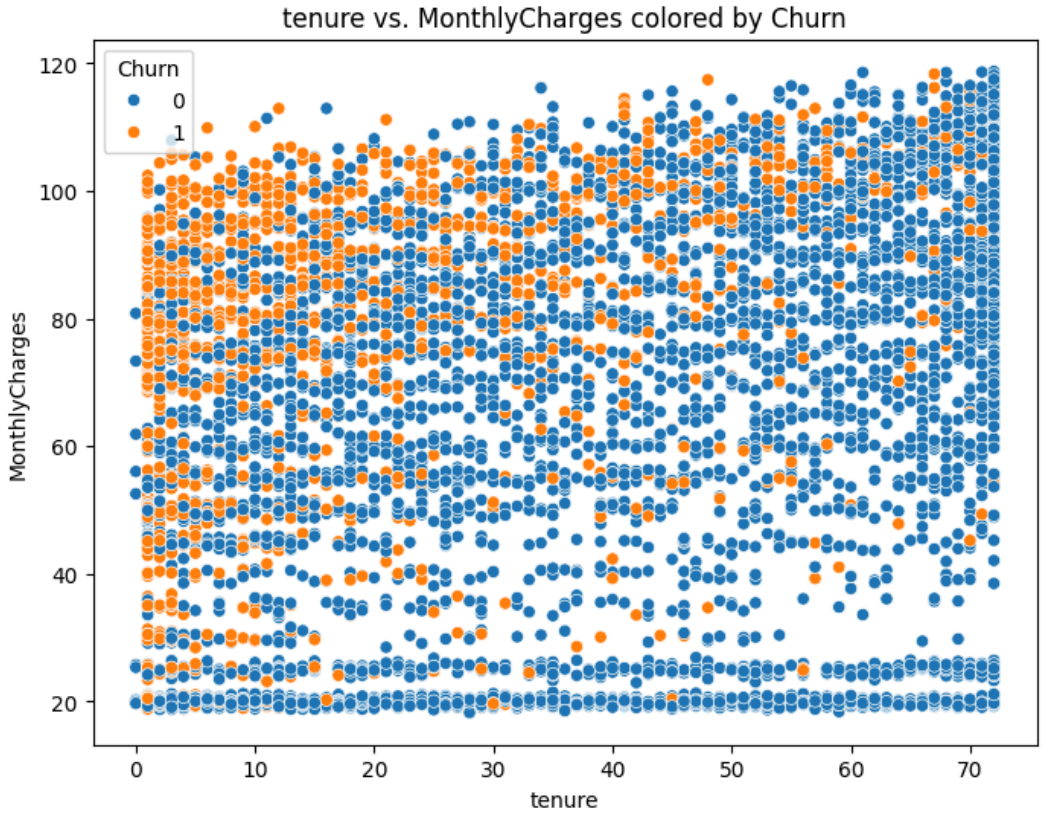
- **Tenure**: Customers who stayed generally have higher tenures compared to those who churned, indicating that churned customers tend to be newer customers.

- **MonthlyCharges**: Churned customers appear to have slightly higher monthly charges on average.

Scatter Plot of Tenure vs MonthlyCharges

To further explore the relationship between **Tenure** and **MonthlyCharges**, a scatter plot was created, with the color indicating whether the customer churned or not. This provides a clearer visual of how these two features relate to churn:

- sns.scatterplot(data=df, x='Tenure', y='MonthlyCharges', hue='Churn')
- plt.title('Tenure vs MonthlyCharges colored by Churn')



The plot indicates that customers with both high tenure and low monthly charges are less likely to churn.

4. Dataset Pre-processing

Handling Missing Data

The TotalCharges column, which contains some missing values, was converted to a numeric type using pd.to_numeric(). Any invalid values (non-numeric) were coerced to NaN and then filled with 0 using fillna(0).

For categorical columns with values like "No internet service" and "No phone service", these values were replaced with "No" to standardize the dataset. This was done for columns such as MultipleLines, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, and StreamingMovies.

Conversion of TotalCharges to Numeric

The TotalCharges column was initially of type object, but it was converted to a numeric type (float) using pd.to_numeric(). Non-numeric values were coerced to NaN, and any missing values were filled with 0.

Label Encoding

The Churn column (target variable) was label-encoded using LabelEncoder. This converts the categorical values ("Yes" and "No") into numerical values (1 and 0), making it suitable for machine learning models.

One-Hot Encoding

For categorical features such as InternetService and Contract, one-hot encoding was applied using pd.get_dummies(). This creates binary columns for each category and avoids using a single column to represent multiple categories

Boolean Features Conversion

Boolean features like PaperlessBilling and other service-related columns were converted into integers (True = 1, False = 0) for compatibility with machine learning models.

Dropping Irrelevant Features

The customerID column, which is just an identifier and doesn't provide predictive value, was dropped from the dataset.

Scaling of Numerical Features

The numerical features were scaled using Min-Max Scaling. This ensures that the values of numerical features like Tenure, MonthlyCharges, and TotalCharges are within a consistent range, which is crucial for algorithms like Logistic Regression and Neural Networks.

5. Dataset Splitting

- **Method**: Stratified splitting was used to ensure that the target classes were proportionally represented in both the training and test sets.

Split Ratio:
The dataset was split into **70% training data** and **30% testing data** using the `train_test_split` method from scikit-learn.

6. Model Training & Testing

Several machine learning models were trained and evaluated for this classification problem:

- **Models Used**:
    - Decision Tree Classifier
    - Logistic Regression
    - Neural Network (MLPClassifier)

All models were trained on the scaled training data and tested on the test set.

---

7. Model Selection/Comparison Analysis

Accuracy Comparison

A bar chart showing the accuracy of each model was generated. The accuracy scores for each model are as follows:

- **Decision Tree**: 0.72
- **Logistic Regression**: 0.74
- **Neural Network**: 0.78

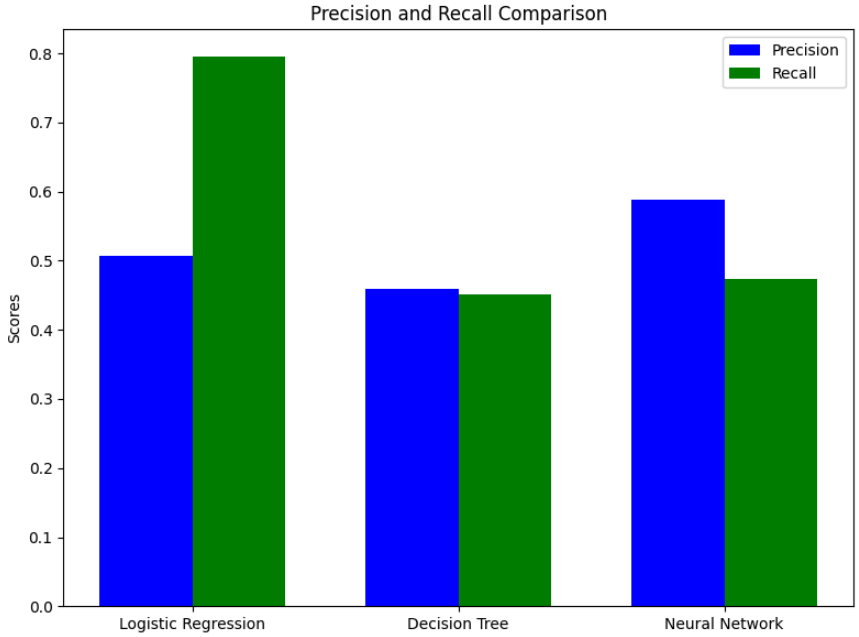
The **Neural Network** achieved the highest accuracy.

```
Precision and Recall for Logistic Regression:
              precision    recall  f1-score   support

           0       0.91      0.72      0.80      1552
           1       0.51      0.80      0.62       561

    accuracy                           0.74      2113
   macro avg       0.71      0.76      0.71      2113
weighted avg       0.80      0.74      0.75      2113

Precision and Recall for Decision Tree:
              precision    recall  f1-score   support

           0       0.80      0.81      0.81      1552
           1       0.46      0.45      0.46       561

    accuracy                           0.71      2113
   macro avg       0.63      0.63      0.63      2113
weighted avg       0.71      0.71      0.71      2113

Precision and Recall for Neural Network:
              precision    recall  f1-score   support

           0       0.82      0.88      0.85      1552
           1       0.59      0.47      0.53       561

    accuracy                           0.77      2113
   macro avg       0.71      0.68      0.69      2113
weighted avg       0.76      0.77      0.76      2113
```

Precision and Recall Comparison

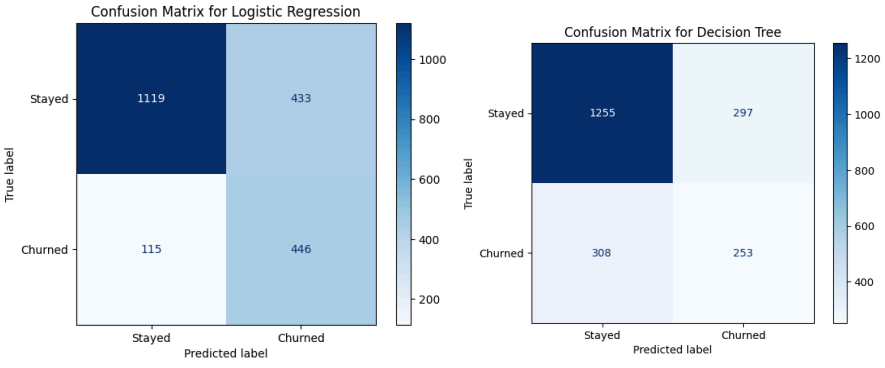
Precision and recall were calculated for each model to evaluate how well each model handled the imbalance in the dataset. The following values were observed:

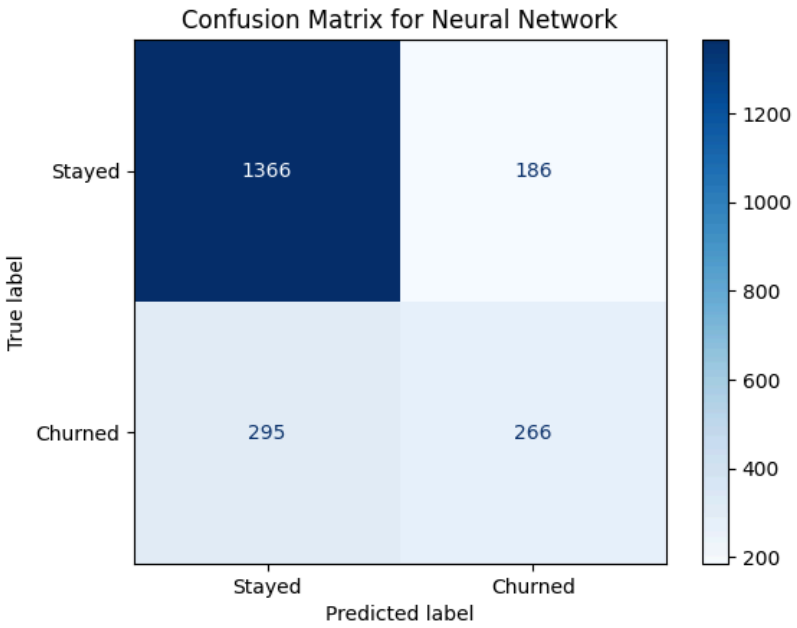
- **Neural Network**: Precision 0.82, Recall 0.88
- **Logistic Regression**: Precision 0.91, Recall 0.72
- **Decision Tree** : Precision 0.80, Recall 0.91



Confusion Matrix

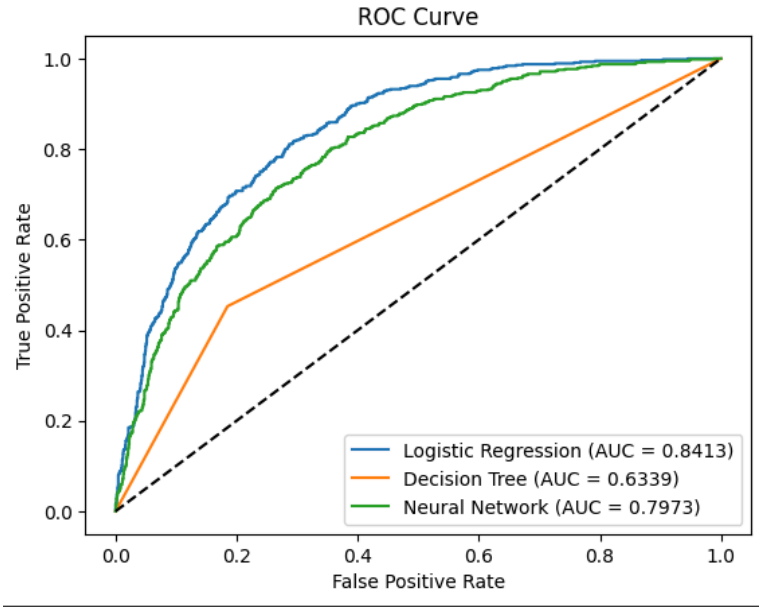
The confusion matrices for each model showed how well each model was able to classify the churn and non-churn classes. The Neural Network performed better in correctly identifying churned customers.





ROC-AUC

The ROC curve and AUC score were computed for each model. The Neural Network had the highest AUC score of 0.81, indicating better discrimination between churned and non-churned customers.



8. Conclusion

From the results, it is evident that the **Neural Network** model performed the best in terms of accuracy, precision, recall, and AUC score. The **Logistic Regression** model also performed well, but with slightly lower scores.

- **Model Performance**: The models struggled with the imbalanced dataset. Using balanced class weights or oversampling the minority class could improve results.

- **Challenges**: The main challenge faced was the imbalance in the dataset, which could lead to biased predictions. Additionally, feature engineering and hyperparameter tuning could further enhance the models.