Abstract:

This report details the analysis conducted to predict customer churn within the telecom industry. A machine learning model, specifically Logistic Regression, was employed to identify customers at high risk of churn. The analysis leveraged the "Telco Customer Churn" dataset and involved data preprocessing, model building, performance evaluation, and customer segmentation. The key drivers of churn were identified, and actionable recommendations are provided to help telecom companies proactively retain valuable customers and improve overall profitability.

Introduction:

Customer churn, the phenomenon of customers discontinuing their service, poses a significant challenge to businesses, particularly in highly competitive sectors like the telecom industry. Retaining existing customers is often more cost-effective than acquiring new ones, making churn prediction and prevention a critical business priority. This project aims to develop a predictive model for customer churn in the telecom sector and provide actionable insights to mitigate it.

Tools Used:

The primary tools used in this analysis were:

- **Python:** A versatile programming language used for data manipulation, statistical analysis, and machine learning. Key libraries included:
    - **Pandas:** For data loading, cleaning, and transformation.
    - **Scikit-learn:** For building and evaluating the Logistic Regression model.
    - **Matplotlib:** For data visualization.

Steps Involved in Building the Project:

The project followed a structured approach, encompassing the following key steps:

1. **Data Acquisition and Understanding:**
    - The "Telco Customer Churn" dataset from Kaggle was utilized.
    - The dataset was explored to understand its structure, data types, and identify any missing values.
2. **Data Preparation:**
    - Missing values in the TotalCharges column were handled.
    - Categorical features were converted into a numerical format using one-hot encoding.

- The target variable, 'Churn', was converted to a binary representation (1 for 'Yes', 0 for 'No').

3. **Model Building and Evaluation:**
    - The dataset was split into training and testing sets.
    - A Logistic Regression model was trained on the training data.
    - The model's performance was evaluated on the testing data using metrics such as accuracy, precision, recall, F1-score, and AUC-ROC.
    - The ROC curve was analyzed to assess the model's ability to discriminate between churned and non-churned customers.
4. **Key Drivers of Churn:**
    - The coefficients of the Logistic Regression model were analyzed to determine the features that most significantly influence customer churn. This identified factors that increase or decrease the likelihood of a customer terminating their service.
5. **Customer Segmentation:**
    - Customers were segmented into three risk categories (High, Medium, and Low) based on their predicted churn probabilities.
    - The characteristics of each segment were analyzed to identify common traits and behaviors.
6. **Recommendations:**
    - Based on the analysis, actionable strategies were formulated to reduce customer churn. These recommendations target high-risk customers, incentivize long-term contracts, promote valuable services, and suggest implementing loyalty programs.

Conclusion:

This project demonstrates the effectiveness of machine learning in predicting customer churn within the telecom industry. By identifying high-risk customers and understanding the key drivers of churn, telecom companies can proactively implement targeted retention strategies. The recommendations provided offer a roadmap for reducing churn, improving customer loyalty, and ultimately enhancing profitability. Continuous monitoring and refinement of the predictive model and retention efforts are essential for sustained success in a dynamic market.