A. Surya Prakash 2211CS010023
Section-1

Spotify Churn Analysis Report

1. Introduction

This project delves into Big Data Analysis using PySpark to investigate and predict user churn on a digital music streaming platform, analogous to Spotify. The primary goal is to analyze user demographics, subscription tiers, and detailed behavioral metrics (such as listening time and skip rates) to understand the factors driving user attrition. By processing this large dataset with PySpark's scalable capabilities, the project aims to not only identify high-risk user segments but also to develop a foundational machine learning model capable of predicting which users are likely to churn, thereby providing actionable intelligence for targeted user retention efforts.

2. Dataset Overview

The dataset, spotify_churn_dataset.csv, is a collection of user data designed for Spotify user churn analysis. It comprises 12 distinct columns spanning both categorical features, such as gender, country, subscription_type, and device_type, and numerical features, including metrics like age, listening_time, songs_played_per_day, and skip_rate. Crucially, the dataset includes the binary target variable, is_churned, which is essential for developing a predictive model to identify users who have left the service. The data was loaded into a PySpark DataFrame, underwent initial inspection, and was confirmed to contain no missing values, simplifying the subsequent data cleaning and feature engineering pipeline.

3. Key Findings

3.1. Data Characteristics & Overall Churn

----The dataset contains 8,000 user records across 12 features, with no missing values.

---- Overall churn rate for the entire user base is approximately 25.89%.

3.2. Exploratory Data Analysis (EDA) Insights

User Segmentation (Subscription & Device)

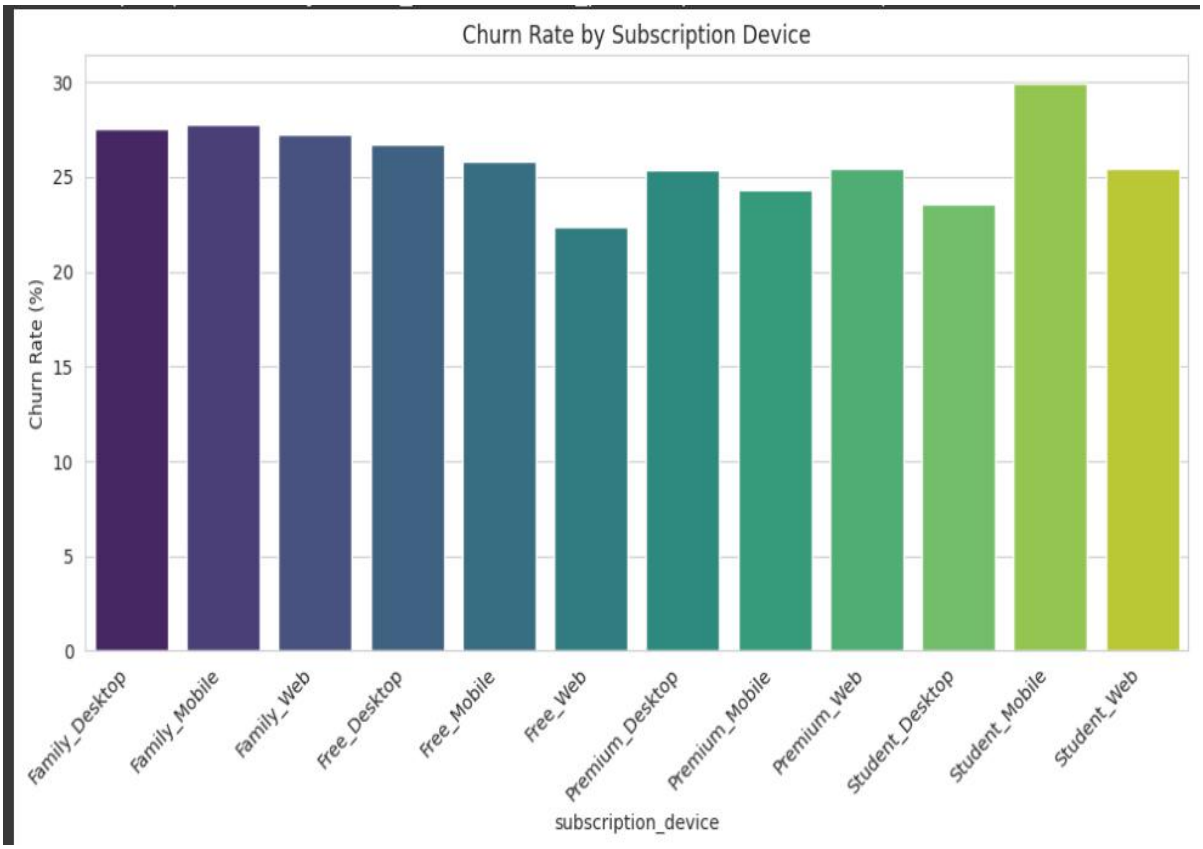
The 'Student_Mobile' segment exhibits the highest propensity to churn,with a churn rate of 29.92%, notably above the overall average.

3.3 Feature Significance (Numerical & Categorical)

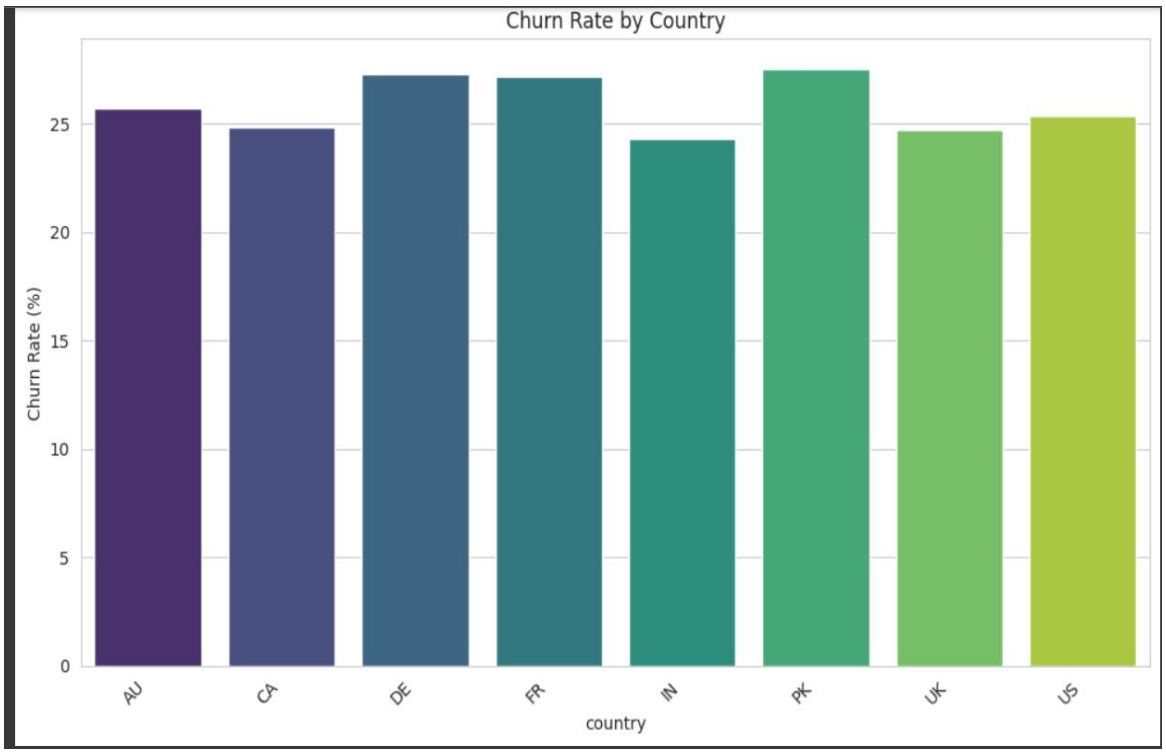
Numerical Usage Metrics (e.g., age, listening_time, skip_rate, songs_played_per_day) are not strong individual predictors of churn.

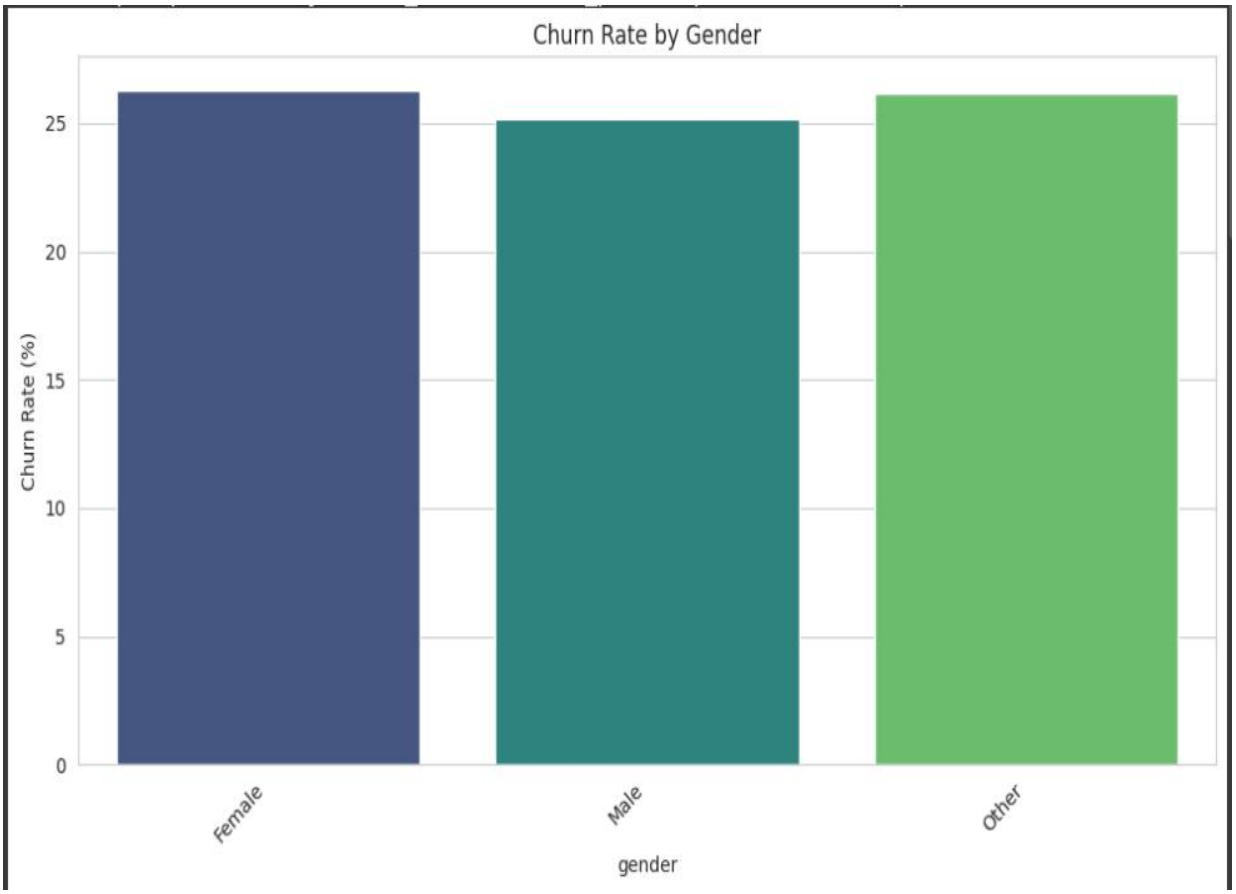
The mean values for these features are similar between churned and non-churned users, with absolute differences of less than 5%.Data Visualizations

The following visualizations illustrate the insights obtained from the analysis:

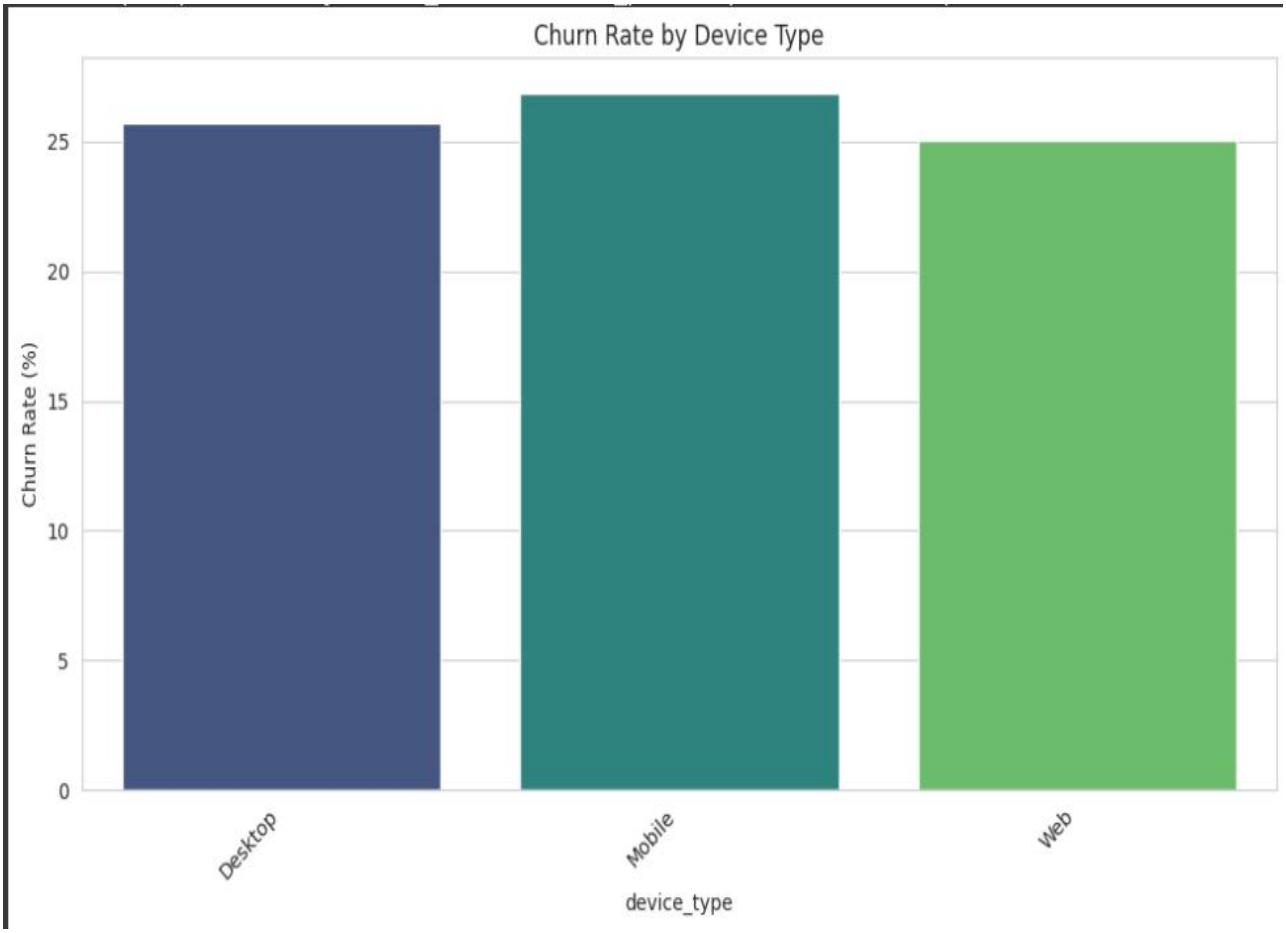


A. Surya Prakash 2211CS010023
Section-1





A. Surya Prakash 2211CS010023
Section-1



Conclusion

This big data analysis project successfully leveraged PySpark to explore and model user churn behavior using the Spotify dataset. The Exploratory Data Analysis (EDA) identified crucial segments, particularly highlighting the 'Student_Mobile' users who demonstrated a notably higher churn rate (29.92%) compared to the overall average (25.89%).

Therefore, the key conclusion is that future efforts should focus on designing targeted retention strategies for the high-risk 'Student_Mobile' segment and exploring more sophisticated feature engineering or non-linear machine learning models to improve the accuracy and actionability of the churn prediction system.