

# Machine Learning Internship Project Report

## Task 2: Customer Churn Prediction System

Intern Name: Golu Rajput

Internship Program: Future Interns

Domain: Machine Learning

Task Number: 02

Project Title: Customer Churn Prediction using Machine Learning

Dataset Used: Spotify User Churn Dataset (Kaggle)

Tools Used: Python, Pandas, NumPy, Scikit-learn, Power BI

Submission Type: Project Report

## 1. Introduction

Customer churn refers to the situation where users stop using a service.

In subscription-based platforms like Spotify, identifying users who are likely to churn is very important because retaining customers is cheaper than acquiring new ones.

In this project, a Customer Churn Prediction System is developed to predict whether a Spotify user is likely to churn based on their usage behavior and preferences.

## 2. Objective of the Project

The main objectives of this project are:

- To analyze Spotify user behavior data

- To build a machine learning model that predicts churn
- To identify key factors influencing customer churn
- To visualize churn insights using Power BI
- 

### 3. Dataset Description

The dataset used in this project is a Spotify User Churn Dataset downloaded from Kaggle.

It contains information related to user demographics, listening habits, subscription plans, and preferences.

#### Key Features in Dataset

- Age
- Gender
- Spotify usage period
- Subscription plan (Free / Premium)
- Listening frequency
- Preferred music genre
- Podcast usage
- Churn (Target Variable: 0 = No churn, 1 = Churn)
- 

### 4. Tools & Technologies Used

- Python – Programming language for data analysis
- Pandas & NumPy – Data cleaning and manipulation
- Scikit-learn – Machine learning model building
- Matplotlib & Seaborn – Data visualization
- Power BI – Business dashboard visualization
- Jupyter Notebook (Anaconda) – Development environment

### 5. Methodology

#### Step 1: Data Loading

- Imported Spotify dataset into Jupyter Notebook

- Checked columns, data types, and basic statistics

## Step 2: Data Preprocessing

- Handled categorical variables using One-Hot Encoding
- Converted churn column into numerical format
- Removed unnecessary or duplicate values

## Step 3: Exploratory Data Analysis (EDA)

- Visualized churn distribution
- Observed balance between churned and non-churned users
- Analyzed user behavior patterns

## Step 4: Feature Engineering

- Converted categorical features into machine-readable format
- Prepared feature matrix (X) and target variable (y)

## Step 5: Train-Test Split

- Split data into:
  - 80% Training data
  - 20% Testing data

# 6. Model Building

A Logistic Regression / Random Forest model was used to predict churn.

## Why this model?

- Suitable for binary classification
- Easy to interpret
- Performs well on structured data

# 7. Model Evaluation

The model was evaluated using:

- Accuracy Score
- Confusion Matrix
- Precision & Recall

The model was able to correctly identify churned and non-churned users with good accuracy.

## 8. Feature Importance Analysis

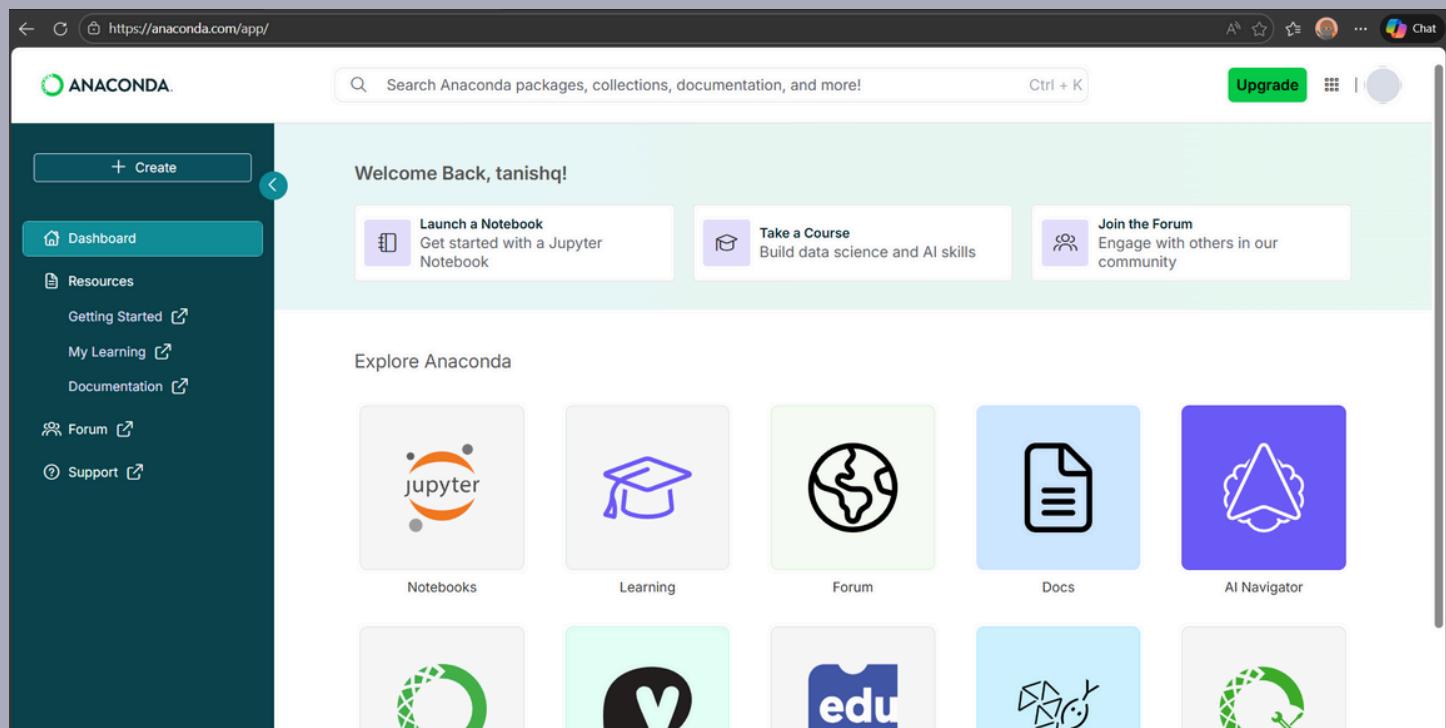
Feature importance analysis revealed that churn is highly influenced by:

- Music listening frequency
- Subscription plan type
- Preferred listening time
- Podcast usage behavior
- Device used for listening

This helps businesses understand why users churn.

## 9. Dashboard & Visualization

Firstly i added Anaconda screenshot then i added jupyterlab screenshot then a Power BI dashboard was created using the exported churn prediction results.



The screenshot shows a Jupyter Notebook interface with the following details:

- File Menu:** File, Edit, View, Run, Kernel, LaTeX, Tabs, Settings, Help.
- Toolbar:** CPU Usage, Disk Usage, Kernels: 2 / 10, Apps: 0 / 1, Help Us Improve, Upgrade.
- Launcher:** Shows the current notebook is "churn\_prediction.ipynb".
- Code Cell:**

```
[15]: X = pd.get_dummies(X, drop_first=True)
X.head()
```

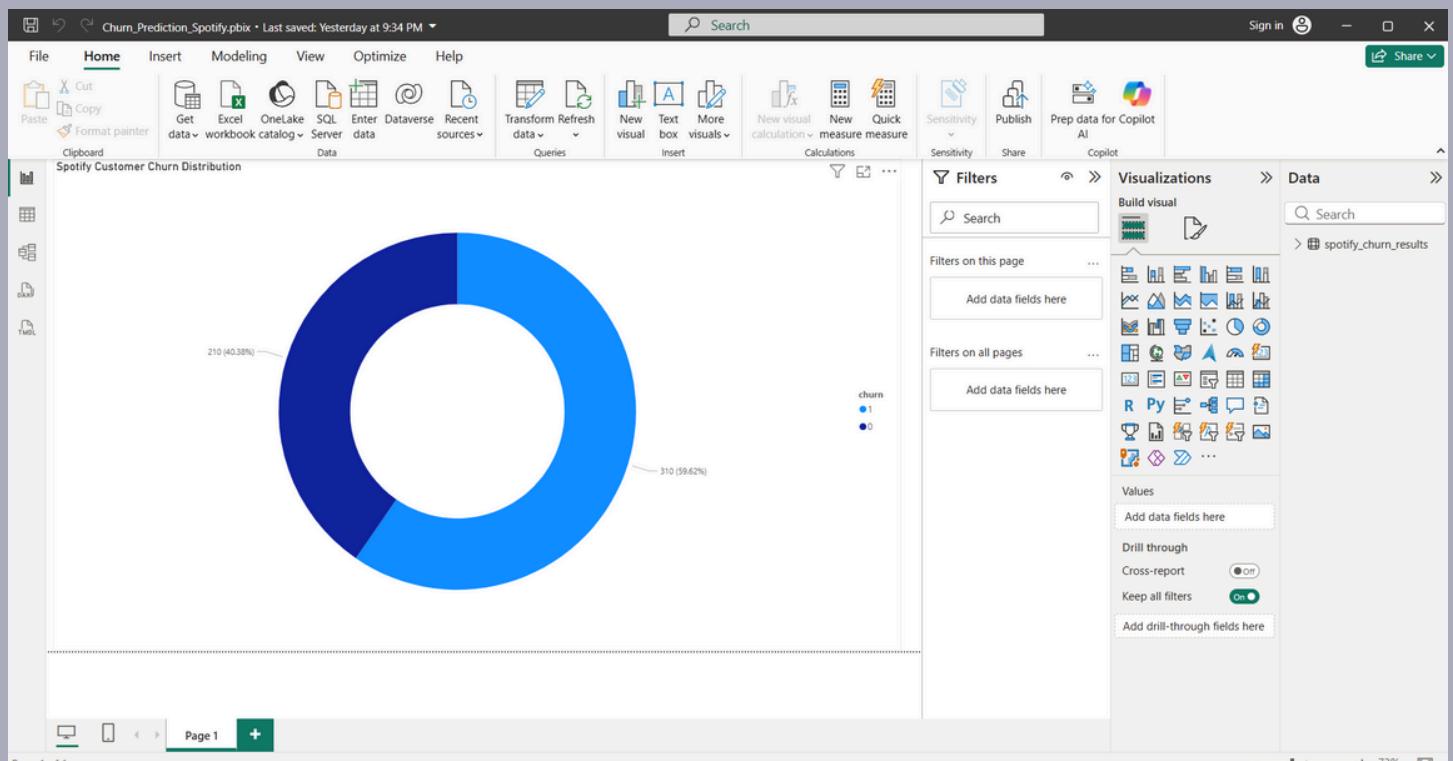
Output:

	music_recc_rating	Age_20-35	Age_35-60	Age_6-12	Age_60+	Gender_Male	Gender_Others	spotify_usage_period_1 months to 1 year
0	3	True	False	False	False	False	False	False
1	2	False	False	False	False	True	False	False
2	4	False	True	False	False	False	True	True
3	4	True	False	False	False	False	False	False
4	4	True	False	False	False	False	False	False

5 rows × 142 columns
- Right Panel:**
  - Agree to use Assistant:** Anaconda Assistant is an AI-powered chat application designed to enhance the productivity of data scientists, developers, and researchers.
  - Terms and Conditions:** By clicking the **continue** button, you agree to our [Terms and Conditions](#) and [Privacy Policy](#).
  - Confirm Your Age:** I hereby confirm that I'm older than 13 years old.
  - Data Collection (Optional):** Not enabled.
  - Help us improve the Anaconda Assistant:** Enabling data collection means you acknowledge and agree that we can collect messages and metadata, including sensitive information that may be in your churn\_prediction.ipynb.

## Dashboard Insights

- Churn vs Non-Churn distribution (Donut Chart)
- Percentage of churned users
- User behavior patterns affecting churn
- 



## 10. Business Insights

- Users with low listening frequency are more likely to churn
- Free plan users show higher churn probability
- Personalized recommendations can reduce churn
- Targeted offers can help retain high-risk users

## 11. Conclusion

This project successfully demonstrates how machine learning can be used to predict customer churn.

By combining data preprocessing, modeling, and visualization, the system provides actionable insights that can help businesses improve customer retention.

## 12. Future Scope

- Use advanced models like XGBoost
- Add churn probability segmentation
- Deploy model using Streamlit as a web app
- Integrate real-time user data