

# Final Project Report: Robert&Shadmun

---

**Group Members:** Shadmun Talukder Shahed (ID 1620656) and Robert Colburn (ID 1578752)

**Course:** COSC 6315

**Instructor:** Dr. Hardik Gohel

**Date:** 12/1/2025

## Abstract

The Spotify Analysis Dataset 2025 from Kaggle is a dataset that our project uses to determine the factors that lead to user churn (Zahid, 2025). The dataset, comprising 8000 entries with 12 features, underwent comprehensive preprocessing including data type conversions, one-hot encoding for categorical variables, and standardization of numerical features. Exploratory Data Analysis (EDA) revealed a significant class imbalance with 74.11% non-churned vs. 25.89% churned, and very weak linear correlations between individual features and churn. Supervised classification models such as Logistic Regression, Decision Tree, k-Nearest Neighbors, and Random Forest consistently showed poor performance in predicting churn, largely attributable to the class imbalance, with ROC AUC scores barely exceeding random chance (0.496-0.529). Random Forest identified listening time, songs played per day, skip rate, and age as the most influential features. Unsupervised k-Means clustering successfully segmented users into four distinct behavioral groups, offering valuable insights for targeted strategies despite the challenges in direct churn prediction. Future work will focus on feature enrichment and advanced class imbalance handling techniques.

## Introduction

Customer churn poses a significant challenge for subscription based services like Spotify, impacting revenue and growth. Understanding and predicting why users leave is crucial for developing effective retention strategies. Our project aims to analyze a simulated Spotify user dataset to identify factors contributing to churn and develop predictive models. The motivation stems from the business need to proactively identify at-risk users, allowing for targeted interventions to improve customer loyalty. The dataset, available on Kaggle Hub (nabihazahid/spotify-dataset-for-churn-analysis), provides a comprehensive view of user demographics, subscription types, and listening habits, alongside a binary target variable indicating churn status. The objectives include performing a thorough exploratory data analysis, implementing and evaluating various machine learning models for churn prediction and user segmentation, and identifying key features associated with churn.

## Methodology

The categories were placed in a dataframe with the user id category being dropped because it is irrelevant to predicting churn. The category of ads listened to per week was considered an outlier because only free tier users are exposed to ads, and free tier users only represent 25% of the dataset. The same can be said for offline listening because it is a feature only available to paid tier users. The columns were split into numerical and qualitative categories for analysis. The approach used is supervised learning using classification algorithms because classification is best for qualitative outputs, like a boolean true/false value (Lee, 2025). The classification models used include logistic regression, decision tree, k-nearest neighbors, and random forest classifier. The training and testing split uses a 80/20 ratio. A value of 70/30 was tried, but it gave the same results as an 80/20 split. The metrics used to evaluate the models are accuracy, precision, F1, and ROC AUC scores. After training and testing, feature importance of the dataset categories are evaluated.

For modeling, a `df_model` DataFrame was created:

- The `user_id` column was dropped as it serves as an identifier and `offline_listening` was also dropped from the model training set.
- Categorical features (`gender`, `country`, `subscription_type`, `device_type`) were transformed using one-hot encoding through `pd.get_dummies(drop_first=True)` to convert them into a numerical format while avoiding multicollinearity.
- Quantitative features (`age`, `listening_time`, `songs_played_per_day`, `skip_rate`, `ads_listened_per_week`) were scaled using `StandardScaler` to normalize their ranges, ensuring no single feature disproportionately influences distance based algorithms.

Exploratory Data Analysis (EDA):

- Descriptive Statistics: `df.describe()` provided statistical summaries for numerical columns (mean, median, std, min, max, quartiles). `value_counts()` and percentages were used for categorical and boolean features to understand their distributions.
- Target Variable Distribution: `df['is_churned'].value_counts()` and a bar plot revealed a class imbalance: 74.11% non-churned (False) and 25.89% churned (True).
- Quantitative vs. Target Variable: Grouped analysis (`df.groupby('is_churned')[quantitative_cols].agg([mean, "median"])`) showed minimal differences in feature means/medians between churned and non-churned groups.
- Correlation Analysis: Pearson correlations between all relevant features such as quantitative, boolean, and one-hot encoded categorical, and `is_churned` were calculated. A heatmap visualized correlations among quantitative and boolean

features. Cross-tabulations visualized relationships between pairs of categorical features.

- Visualizations: Histograms and box plots were used to visualize distributions of quantitative features, and heatmaps for correlations and categorical relationships. It was noted that `ads_listened_per_week` is primarily relevant for free-tier users.

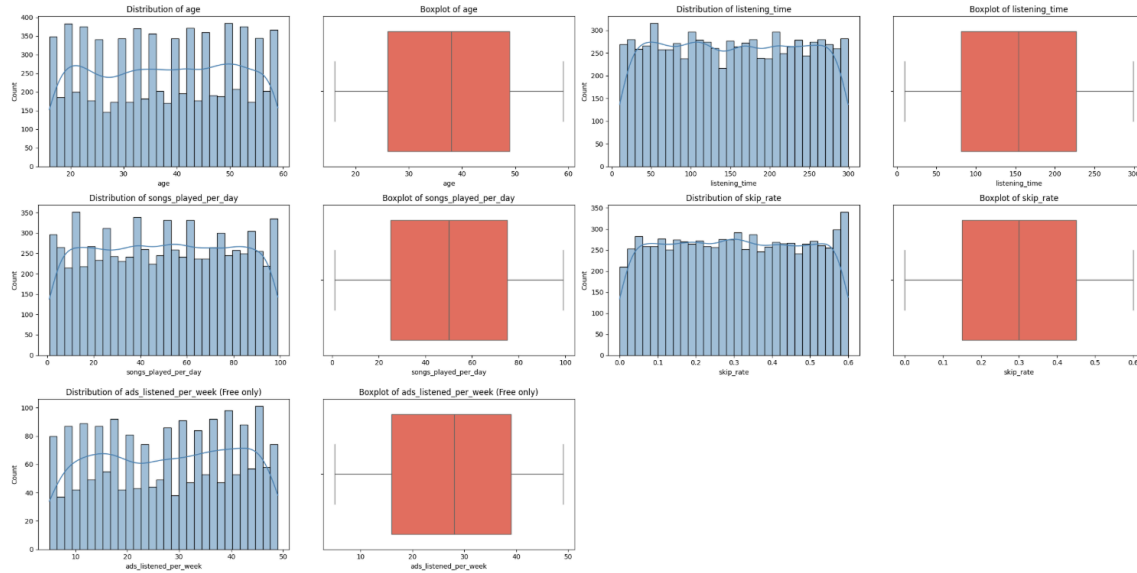


Figure: Graphs of key user behavior features

### Modeling Strategies:

- Supervised Learning (Classification)
  - Objective: To predict the `is_churned` status (binary classification).
  - Models Implemented: Logistic Regression, Decision Tree, k-Nearest Neighbors (k-NN), and Random Forest.
  - Train-Test Split: The dataset was split into an 80% training set and a 20% testing set. `random_state=42` ensured reproducibility, and `stratify=y` maintained the proportion of churned users in both sets to account for class imbalance.
  - Features Used: All preprocessed features from `df_model` such as scaled numerical and one-hot encoded categorical except `is_churned` were used as predictors.
- Unsupervised Learning (Clustering)
  - Objective: To segment users into distinct groups based on behavioral patterns.
  - Model Implemented: k-Means Clustering.

- Features Used: A subset of key numerical features such as skip\_rate, songs\_played\_per\_day, age, and listening\_time. These were scaled using StandardScaler prior to clustering.
- Number of Clusters: The model was configured to identify 4 clusters (n\_clusters=4).

## Results & Evaluation

The results of the model training is as follows (rounded to the third significant digit):

Model	Accuracy	Precision	Recall	F1	ROC AUC
Logistic Regression	0.741	0.549	0.741	0.631	0.496
Decision Tree	0.613	0.624	0.613	0.618	0.509
k-Nearest Neighbors	0.688	0.626	0.688	0.646	0.527
Random Forest	0.735	0.597	0.735	0.631	0.529

Figure: Summary of results acquired from classification models.

The logistic regression model has the best accuracy score, and logistic regression does best when it comes to recall. The k-nearest neighbors model has the highest precision and F1 score. For ROC AUC, the best model is random forest. Using the feature importances of the random forest model, the most important features that influence the model are age, skip rate, songs played per day, and listening time.

## Discussion & Insights

Model accuracy represents the percentage of correct predictions out of the total testing set. Precision measures the correct amount of predicted positives. Recall is the proportion of actual positives that were correctly identified as positive (Google, 2025). The F1 score is the balance between the precision and recall scores. ROC AUC measures the area under a curve of the model and the probability that a model will rank a positive example higher than a negative one (Google, 2025). Logistic regression's accuracy of 0.741 means that out of the entire testing set, the sum of its true positives and true negatives is greater than the other models. The k-nearest neighbors model's predicted positives had the highest percentage of actually being positive with a ratio of 0.626. All of the models have a similar F1 score, but the k-nearest neighbors model scores highest with 0.646. This

means the model has the best ratio of precision and recall. For ROC AUC, the highest value is better, with the value being from 0.0 to 1.0. A value of 0.5 means a model predicts an example correctly 50% of the time. All of the models hover around 0.5, but random forest performs best with a value of 0.529.

The models may have been biased to predict non-churners because the dataset is heavily skewed towards non-churners, with 74% of users not churning and 26% churning. This factor could explain why the confusion matrices had a significant amount of false negatives. In the random forest model, the most important features were age, skip rate, songs played per day, and listening time. These categories are all numerical. The qualitative categories such as subscription type, device type, and gender do not have much of an impact when it comes to model training.

## **Conclusion & Future Work**

Out of our selected models of logistic regression, decision tree, k-nearest neighbors, and random forest, the random forest classifier demonstrated the strongest overall performance. Its ability to capture nonlinear relationships and reduce overfitting through ensemble learning made it particularly effective for churn prediction. As a result, future prediction efforts can reliably use the random forest model as a baseline, or as a benchmark for comparing more advanced methods.

Feature importance analysis revealed that listening time, songs played per day, skip rate, and age are the most influential variables in determining user churn. These features directly reflect user engagement and satisfaction, making them strong behavioral indicators. In contrast, categorical attributes such as gender, country, subscription type, and device type showed minimal predictive value. This suggests that future data collection efforts should prioritize high quality numerical and behavioral metrics over demographic or qualitative variables, as the former contribute much more meaningfully to model accuracy and actionable insights.

Regarding future improvements, one key area of feedback involved addressing the imbalance between churn and non-churn users. Since churners represent a minority class, traditional models tend to struggle in identifying them accurately. To mitigate this, techniques such as downsampling the majority non-churn class can be implemented to create a more balanced dataset. This adjustment can help models better learn the patterns associated with churn and improve recall for the minority class.

## References

- Google. (2025, November 3). *Classification: Accuracy, recall, precision, and related metrics | machine learning | google for developers*. Google. <https://developers.google.com/machine-learning/crash-course/classification/accuracy-precision-recall>
- Google. (2025, November 3). *Classification: Roc and AUC | machine learning | google for developers*. Google. <https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc>
- Lee, F. (2025, November 17). *Classification vs regression*. IBM. <https://www.ibm.com/think/topics/classification-vs-regression>
- Zahid, N. (2025, August 28). *Spotify Analysis Dataset 2025*. Kaggle. <https://www.kaggle.com/datasets/nabihazahid/spotify-dataset-for-churn-analysis/data>