Project Final Report

Team 20, Section C

Weihan Liu, Sai Maradugu, Jannat Mubarik, Allie Sutter, Nishtha Wakankar

Business Understanding

Business Problem and Motivation

To be a successful company, Spotify must retain its subscriber base, as user churn directly results in a decrease in revenue, an increase in marketing costs in an attempt to acquire new users, and a loss in long-term customer lifetime value. In the highly competitive market of music streaming, it is important to maintain and prioritize user loyalty. Therefore, the main business problem we are analyzing here is to identify which users are at a high risk of churning, understand the demographic, behavioral, etc. factors affecting churn, and formulate data-driven strategies and recommendations to reduce churn rate.

How a Data Mining Solution Addresses the Problem

A data mining solution helps Spotify identify and predict which users are more likely to churn and understand why. Using supervised learning tools such as logistic regression and random forest, Spotify can estimate the probability of each user's churn and rank them in order of importance. Analyzing the features of the data set helps to show which behaviors (listening time, ad frequency, etc.) drive churn, guiding targeted improvements. Through unsupervised learning tools such as clustering, users can be grouped into segments based on behaviors (heavy listeners, free users), allowing for personalized strategies for retention. These insights create business value by reducing churn, optimizing marketing efforts, and building long-term loyalty through data-focused individualized engagement.

Data Understanding

Our data is a 2025 Spotify analysis data set that was obtained from Kaggle. It is synthetically generated and each observation represents a Spotify user with demographic, subscription, and engagement information. Some of the key attributes include:

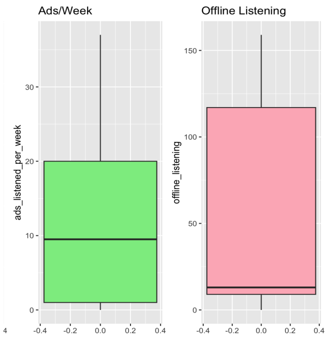
- **Demographics**: age, gender, country
- **Subscription**: subscription_type, device_type, ads_listened_per_week
- **Engagement**: listening_time, skip_rate, songs_player_per_day

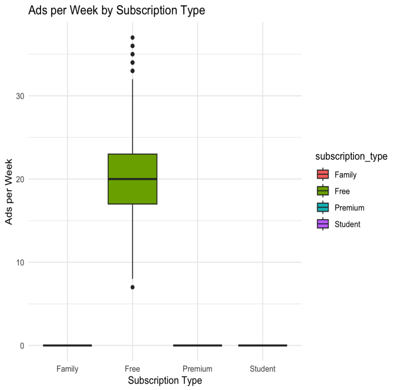
Because the data is synthetically generated, there may be some bias such as oversimplified or exaggerated churn patterns. Results demonstrate general trends rather than precise real-world outcomes.

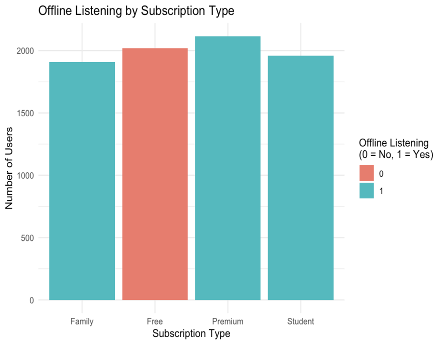
Data Preparation

1) **Missing Values & Duplicates** - There are no missing values in the dataset, hence no imputations are required. There are no duplicate users i.e. all user_ids are unique hence pertaining to individual unique users
2) **Sanity Checks** -
    a) Numerical - All numerical variables were within valid ranges: **age** values were reasonable (10–90), **skip_rate** stayed between 0 and 1, and both **listening_time** and **songs_played_per_day** were nonnegative with higher values reflecting heavy users.
    b) The categorical variables (gender, country, subscription_type, and device_type) were reviewed and found to contain only valid and consistent categories, with no evidence of typos, duplicate labels, or unexpected values.
3) **Outliers-**

For **subscription_type**: Premium, Family, and Student users consistently reported zero ads, while Free users experienced higher ad exposure. Thus, these are not true outliers but valid business-driven differences, and no cleaning is required. In addition, the **offline_listening** variable is strictly binary (0 = no offline listening, 1 = offline enabled) and should therefore be treated as a categorical feature rather than continuous in subsequent modeling. Offline listening is also consistent with the subscription type. This also tells us that the variables **subscription_type ads_listened_per_week**, **offline_listening** are highly correlated and will need to account for this during modeling.





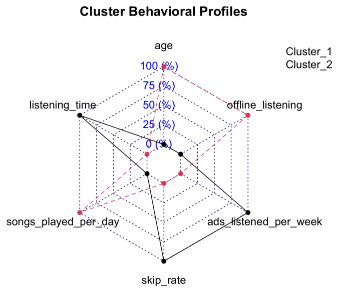


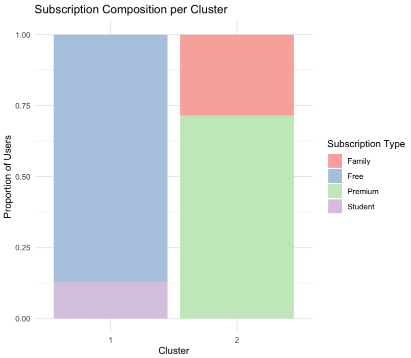
Summary statistics of data:

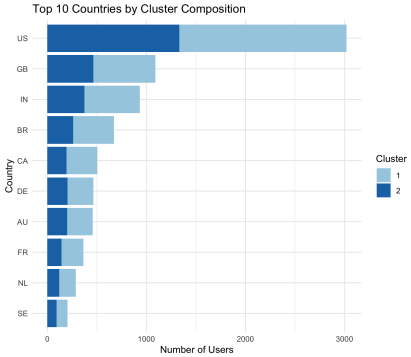
- Overall churn rate of individuals in our data set is 53.88%
- Churn rates by subscription type: Free (72.6%), Student (57.8%), Family (34.1%), Premium (29.6%)
- Churn rates by device type: Web (57%), Desktop (54.2%), Mobile (53.1%)
- Average listening time = 148.8 hrs/week, average songs played per day = 32.12, average skip rate =28.74%

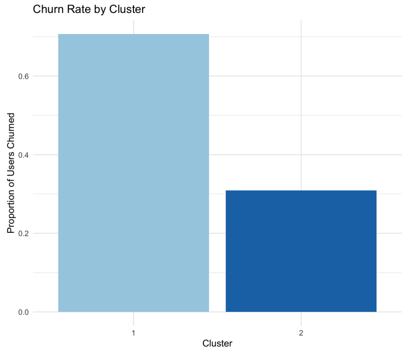
Modeling

Clustering



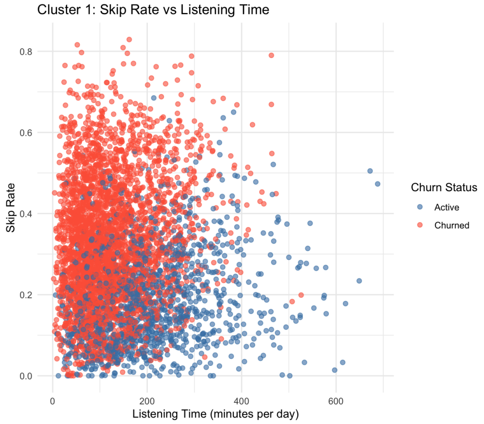


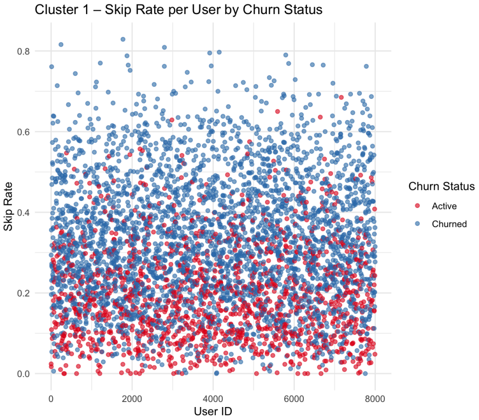




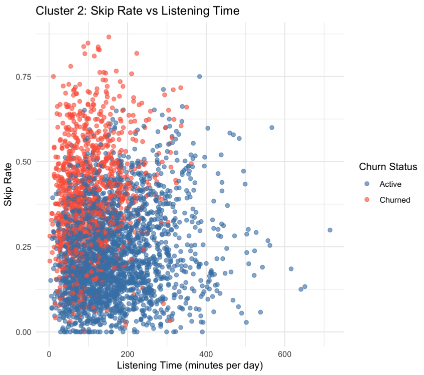
The k-means clustering grouped Spotify users into two distinct segments based on listening behavior, engagement, and subscription type. These clusters differ sharply in listening intensity, skip behavior, offline usage, and churn patterns, highlighting two fundamentally different user journeys: content explorers and committed subscribers.

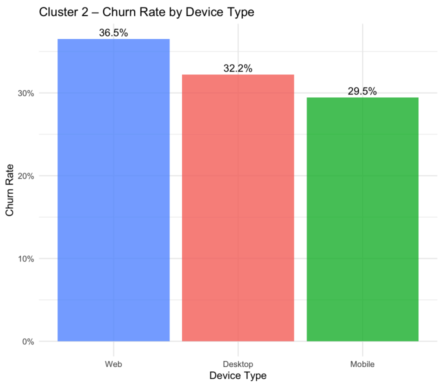
1. **Cluster 1 (Free/Student)** are price-sensitive, easily dissatisfied users; they represent Spotify’s conversion and retention challenge. Their high skip behavior signals recommendation mismatches and opportunity for algorithm improvement.

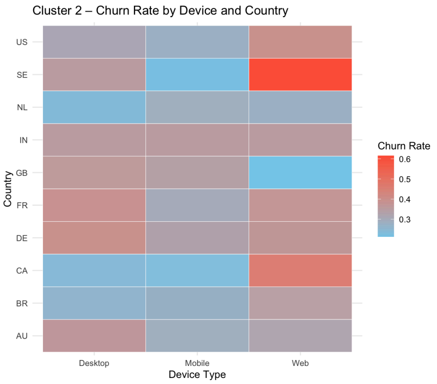


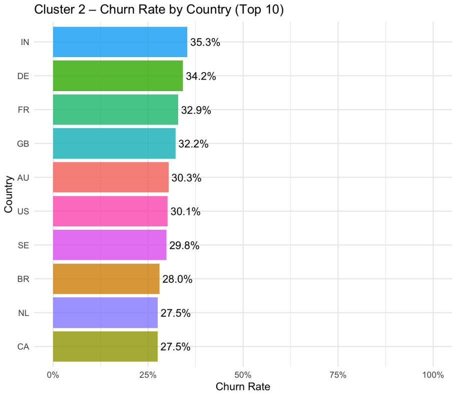


2. **Cluster 2 (Premium/Family)** are satisfied, high-value customers with steady engagement; they represent Spotify's retention and upsell opportunity. Their churn is less behavioral and more situational (e.g., subscription fatigue, payment cycle).









Approach

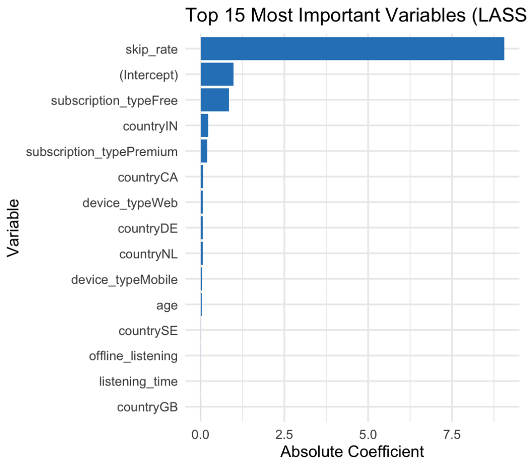
We developed six models: logistic regression, logistic regression with interactions, LASSO logistic regression, classification tree, random forest, and neural network, with a null model as a performance baseline.

The logistic regression model uncovered key churn predictors, such as age, listening time, skip rate, and offline listening ($p < 0.001$). Older users and those with more listening time or offline engagement were less likely to churn, although high skip rates increased the risk. Users from Germany and India were more likely to churn, whereas Premium subscribers remained more loyal. The model's deviance decrease (7949.9 → 4747.9) and AIC of 4793.9 suggest a strong fit, with engagement behaviors as the primary churn drivers.

The interaction model included pairwise terms while preserving the same major predictors. Notable interactions include countryUS × skip_rate (reduced churn sensitivity to skipping) and favorable ads_listened_per_week effects across various countries. Premium web users experienced more turnover than expected, as indicated by the interaction between Premium and Web devices. While deviation improved (4560.1 versus 4747.9), AIC rose to 4986.1, indicating possible overfitting.

The classification tree highlighted offline listening, skip rate, and listening time as top predictors, resulting in eight terminal groups with approximately 78% accuracy and 21.6% misclassification. Users with limited offline listening (<78 mins) and high skip rates (>0.25) saw the most churn, while heavy offline listeners with longer sessions (>148 mins) had the lowest risk (~5%). This shows that increasing offline use and deeper involvement may lower churn.

The random forest model obtained good predictive performance by using all available variables but identified skip rate, listening time, and offline listening as the most essential for predicting churn. The model has a low out-of-bag error rate, and the confusion matrix showed balanced accuracy across classes, implying that it generalizes effectively without overfitting.



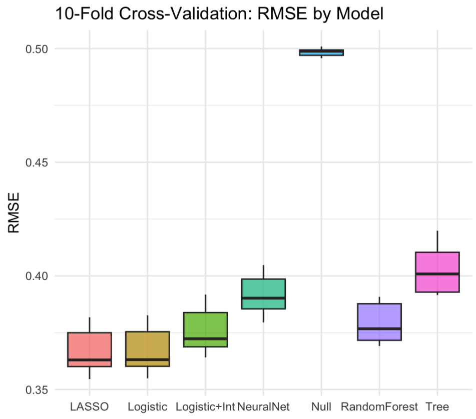
The neural network efficiently predicted non-linear relationships with eight hidden neurons. It caught nuanced behavioral patterns that influence churn, albeit at a lower interpretability than tree-based models.

The modeling framework targets Spotify’s main business challenge: reducing customer churn rate. Spotify can use the predictive model to retain customers before they disengage the app.

The model helps Spotify to identify at-risk users (users with high predicted churn rate) early so that Spotify can start retention campaigns such as discount, email, and etc. Also, the LASSO model helps Spotify with which behavioral factors have a stronger correlation with churn rate, for example a low listening time and a free usage. They can produce targeted playlists and marketing strategies to intervene early. Compared to the Null model, our best model, LASSO (AUC = 0.889, RMSE ≈ 0.37, Accuracy ≈ 0.80) shows a big improvement in prediction. In all, LASSO solves transforming behavioral data into churn rate, enabling Spotify to target at-risk users and intervene early.

Evaluation

We employed a 10-fold cross-validation using four metrics, AUC, RMSE, Log-loss, and Accuracy. AUC measures a model's ability to rank churner above non-churner, which is important for marketing strategy. RMSE measures the deviation between predicted churn rate and actual outcomes, testing how well the model performs in predictive precision. Log-loss penalized overconfidence, incorrect prediction, checking if the model provides reliable estimates. Among the 6 predictive models, LASSO performs the best with AUC = 0.889, RMSE = 0.37 and Accuracy = 0.80. Although both LASSO and standard logistic regression achieve good performance in prediction, LASSO also selects the most influential variable automatically, making the result more interpretable. Features like skip rate, free-tier subscription, and shorter listening time all show strong positive correlation with churn rate. While Random Forest and Neutral Network's performance is acceptable, they act like a black



boxes. They offered limited transparency about features' influence. For Spotify's marketing and strategy group, it is important to understand why a user might churn.

With the model, Spotify can target churn-prevention efforts more effectively, reducing waste in unnecessary marketing campaigns. For example, it can identify the top 30% of users with high churn-risk using AUC and personalized retention campaigns. While the LASSO model is a powerful tool for prediction, it does not by itself explain the different types of users' behavior. Therefore, we extended our analysis using unsupervised learning via clustering.

Deployment

The goal of the deployment is twofold:

1. Predict churn risk (propensity) for each user using behavioral features.
2. Segment users (via clustering) to design differentiated retention interventions.

Our solution intends to integrate the churn prediction model and the clustering framework into a unified deployment strategy aimed at reducing churn and improving user retention across Spotify's customer base. The deployment approach differs by cluster, reflecting the contrasting user motivations and engagement behaviors identified through the analysis.

Cluster 1: Free and Student Users

Cluster 1 primarily consists of free and student-tier listeners who are price-sensitive and less committed, but they also exhibit a clear interest in the quality and relevance of music recommendations. Their churn appears to be driven more by dissatisfaction with their immediate listening experience. From a business standpoint, it may not be practical to attempt large-scale improvements to the free-tier recommendation engine, which operates differently from the

Premium recommendation systems. However, the existing data can be used to intelligently identify high-risk users for more targeted interventions :- Users whose predicted churn probability exceeds 0.7 are auto-flagged for proactive engagement**.** Once identified, these users can receive short Premium trial offers or curated playlist refreshes designed to re-spark engagement. Since users in this cluster are already sensitive to music quality and curation, but hesitate to upgrade due to cost ; Offering time-limited Premium trials or dynamically refreshed playlists **exposes them to the improved algorithmic experience of the paid tier.** This not only addresses the satisfaction gap, but also serves as a low-cost conversion strategy. Global algorithmic improvements should focus on refining specific playlist categories such as "Daily Mix" or "Discover," rather than the entire free recommendation framework. By connecting churn prediction probabilities with user-level behavioral indicators, these targeted interventions can improve satisfaction at scale while maintaining cost efficiency.

Cluster 2: Premium and Family Users

Cluster 2 represents users who demonstrate steady engagement levels, longer listening sessions, and low skip rates. While this group has a lower overall churn rate compared to free users, they churn despite satisfactory listening experiences, suggesting that the underlying causes are tied to perceived value, pricing, subscription fatigue, or changes in lifestyle and consumption patterns.

To address this, the future focus should be on enhancing the richness of behavioral and lifecycle data used in churn prediction. Beyond the existing engagement features, new data points such as subscription tenure, renewal history, offline listening ratios, device diversity, and payment method stability would add significant predictive power. For instance, users with shorter tenures or repeated failed renewals are often early indicators of churn risk, while those who engage across multiple devices or use offline mode more frequently are generally more stable. Tracking these signals can help differentiate between users at temporary risk (e.g., payment failures) and those exhibiting long-term disengagement.

In terms of immediate deployment, the churn prediction model can serve as an early warning trigger system. Users with a churn probability above a defined threshold (e.g., 0.7) are automatically flagged for retention campaigns, while moderate-risk users (above 0.4) are placed into engagement or loyalty reinforcement programs. Retention strategies for high-risk Premium users should center around value reinforcement rather than aggressive discounting like high-risk cases P(>0.7) automated campaigns can offer loyalty incentives "Extend your Premium for 3 months at a discounted rate" or pause options "Take a break, your playlists will wait for you"

For lower risk cases, reminders of listening milestones, usage insights (You listened to 1,200 minutes this month), or prompts highlighting underused Premium features such as offline listening or for Family plan users, communications can emphasize shared features such as collaborative playlists and multi-device continuity to strengthen household-level engagement.

From a regional standpoint, the analysis indicates that churn behavior among Premium users varies modestly by country and device type, with web-only listeners in certain markets (for example, Sweden and India) exhibiting higher churn rates. Targeted actions in these cases could involve promoting mobile app adoption or offering localized payment flexibility.

Overall, for Cluster 2 rather than addressing content dissatisfaction, the goal is to sustain long-term value perception and minimize voluntary churn through habit reinforcement, flexible billing, and targeted retention messaging.

Appendix A: Individual Contribution

Weihan: Modeling (Random Forest and Neural Network) and Evaluation

Sai: Data Preparation and Modeling (Logistic Regression)

Jannat: Modeling (Logistic Regression, Logistic Regression with Interactions, Classification Tree, LASSO) and Evaluation

Allie: Business Understanding, Data Understanding, and Summary Statistics of Data Preparation

Nishtha: Performed k-means clustering and Deployment Strategy