Spotify User Behavior Project

Business Understanding

This project is designed to address a significant business challenge faced by Spotify. Despite being a market leader in revenue generation, Spotify experienced substantial operating losses in 2022, totaling approximately 659 million euros. Furthermore, these losses are projected to grow, leading to decreasing profit margins. The primary objective of this project is to leverage data mining techniques to explore and implement strategies that can enhance Spotify's revenue, ultimately aiming to maximize profits and mitigate financial losses.

One way to increase revenue is to target free users with marketing strategies with hopes of converting them to a premium user. As shown in the Statista visualization, Spotify's revenue from free, ad-supported users was only 1.4 billion vs 10.2 billion euros in 2022. However, the amount of ad-supported still surpassed premium users with ad-supported being (~228 Million) vs. Premium (~205 Million).[1]

Therefore, the goal is to analyze differences between these free and premium users, segment the ad-supported users and use machine learning models to predict the likelihood of these users converting. These models will help Spotify determine which segments and customers are worth targeting with marketing strategies. More data will be needed to determine whether these marketing strategies work and how effective they are. This project will serve as a solid foundation for secondary analysis like cost-benefit

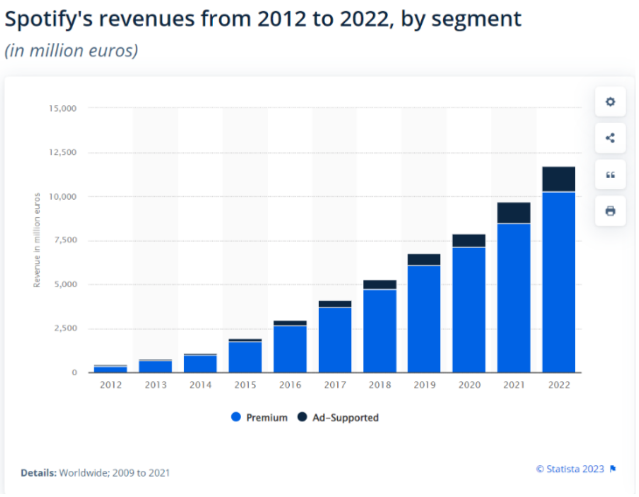


Figure 1 Spotify's Reveunes from 2012 to 2022 by segment, *Statista*

[1] Gotting, Marie Charlotte, "Spotify's Premium Subscribers 2013-2022." Statista, https://www-statista-com.proxy.lib.duke.edu/statistics/244995/number-of-paying-spotify-subscribers/ Accessed September 28, 2023

analysis and A/B testing on marketing strategies and/or churn prediction.

Data Understanding

I used the Spotify User Behavior Dataset (from Kaggle) to conduct customer analysis so I could potentially help Spotify identify the best segments for their marketing campaign.

The Spotify User Behavior Dataset is an exhaustive collection of data that provides valuable information about Spotify User's behavior patterns and preferences. This primary dataset contains demographic information, usage history, behavior patterns, and preferences of the users, which can be used for analyzing user interactions, music consumption habits, and engagement metrics within the Spotify music streaming platform.

Data impurity: This data was collected through an online Google form in a region in India.

Caveats include: The data, collected via an online survey, may exhibit response bias. The survey's dissemination in a potentially restrictive region might not represent the diversity of Spotify users. Furthermore, the non-randomized method of spreading the survey through word of mouth could affect its representativeness.

EDA

Gender Distribution： About 75% of Spotify users are **female**, suggesting a need for Spotify to amplify features they prefer. **Males** make up roughly 22%, indicating a potential area for exploration to increase engagement. The **'Others'** category, at nearly 3%, emphasizes the importance of inclusivity in catering to diverse user needs.

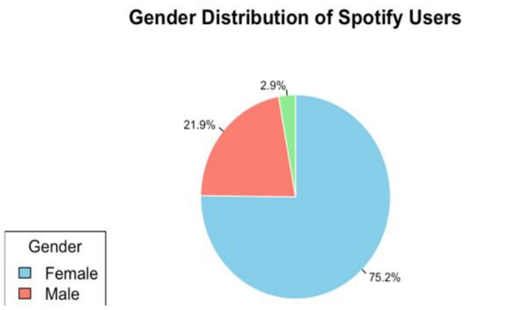


Figure 2 Gender Distribution of Spotify Users Pie Chart

Age Distribution： Over 80% of Spotify users are aged **20-35**, indicating a strong appeal among young adults. Teens **(12-20)** constitute about 14%, suggesting potential for increased engagement. The **35-60** and **6-12** groups have limited representation, at 4.4% and 0.6% respectively, highlighting areas for Spotify to enhance content for middle-aged and young listeners. The senior demographic **(60+)** is barely present at 0.2%, emphasizing a need to better understand and cater to their preferences.

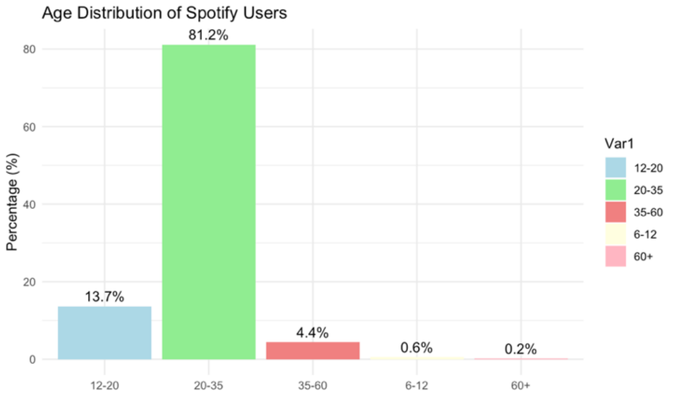


Figure 3 Age Distribution of Spotify Users Histogram

Preferred Music Genres:

Melody dominates as the top genre among Spotify users, highlighting its broad appeal. **Pop and Rap** closely follow, indicating strong interest in both. Given their shared popularity, Spotify could consider crossover playlists or artist collaborations in these genres. **Classical and Melody** genres see moderate preference, suggesting potential for unique playlists blending both styles. **Kpop, Rock, and Electronic/Dance** also have decent representation, opening avenues for themed challenges or spotlighting artists in these categories. While other genres have lower counts, they cater to niche audiences. Showcasing emerging artists from these genres could appeal to dedicated listeners and attract new ones.

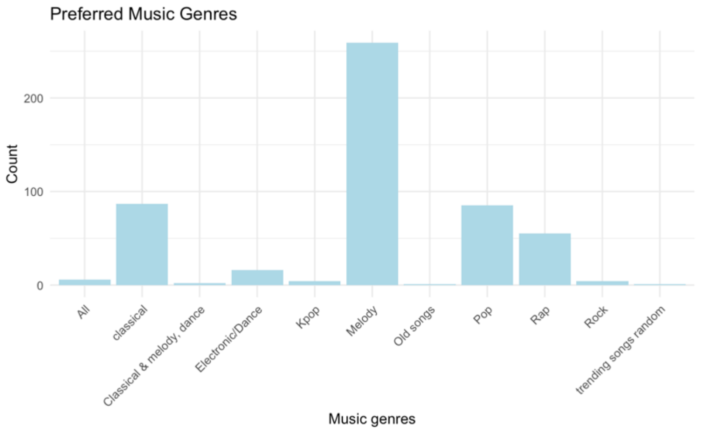


Figure 4 Preferred Music Genres Histogram

Subscription Plans by Gender:

A significant majority of **female** Spotify users lean towards the **Free (ad-supported) subscription**. Their overwhelming preference for this plan indicates potential barriers or specific motivations against upgrading to Premium. Spotify might consider targeted surveys to understand these choices better.

Male users display **a more even distribution between Free and Premium subscriptions**. While a notable number prefer the Free version, a considerable portion also sees value in the Premium offerings. Insights into this balanced trend could refine Spotify's strategies for this demographic.

The **"Others"** category, though lesser in count for both plans, underscores the importance of inclusivity. Ensuring diverse and tailored experiences for all users remains crucial for Spotify.

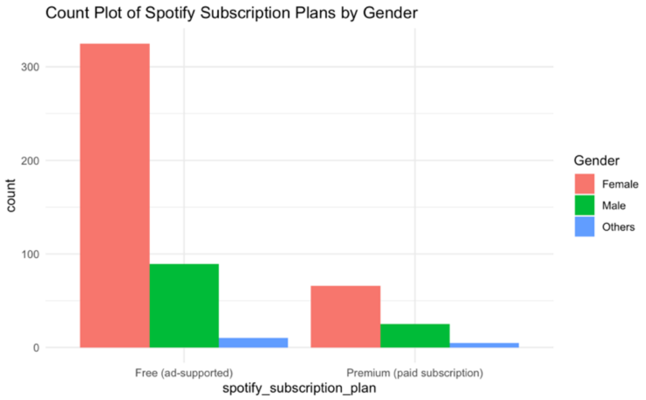


Figure 5 Count Plot of Spotify Subscription Plans by Gender

Modeling: Ad-supported Customer Segmentation (PCA and k-means)

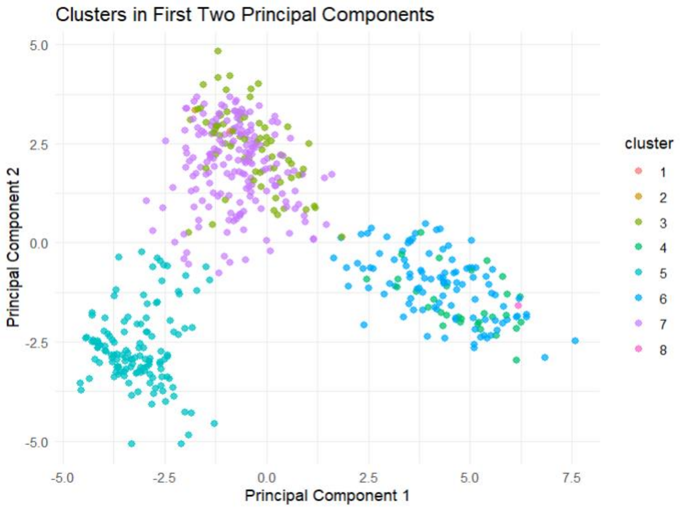


Figure 6 Clusters

CLUSTER	COLOR DESCRIPTION	AVERAGE RATING	WILLINGNESS TO SUBSCRIBE (%)	OBSERVATIONS/DESCRIPTION
1	Soft Red/Salmon	3.00	0	Predominantly female listeners who exclusively use smartphones, smart speakers, voice assistants, and wearable devices. They rely on free subscriptions and favor music, particularly of the melody genre, for relaxation and stress relief at night.

2	Mustard/Dark Gold	3.00	100	Exclusively female users who employ smartphones and wearables for listening. Despite having free accounts, they prefer music, especially the melody genre, to achieve relaxation and stress relief during nighttime.
3	Lime/Green-Yellow	3.80	26.2	A predominantly female group using mainly smartphones and computers for listening. While most have free accounts, a significant portion shows willingness to pay, favoring music, especially melody and pop genres, mostly at night for relaxation.
4	Medium Green/Mint	4.00	0	Majority female users, specifics about their devices are unknown. About 61.3% rely on free accounts but show willingness to subscribe, with most preferring podcasts and listening primarily in the afternoon.
5	Teal/Turquoise	2.88	62.7	Largely female audience, with device details missing. An overwhelming majority prefer free subscriptions and are not inclined to pay, exclusively enjoying music, particularly during nighttime.
6	Bright Sky Blue	3.39	66.7	A group with a female majority, device specifics are not provided. Most use free accounts, but a significant portion is willing to pay, leaning towards podcasts as their preferred content, especially in the afternoon.
7	Lavender/Light Purple	3.65	21.8	Predominantly female listeners, device details are absent. A vast majority rely on free subscriptions and are not willing to upgrade, mainly enjoying music at night. Relatively satisfied
8	Bright Pink/Fuchsia	2.74	60.9	Exclusively female users, with device information missing. They use premium subscriptions, indicating a willingness to pay, and have a preference for music as their primary content.

We utilized K-Means clustering and Principal Component Analysis (PCA) to extract actionable insights from the dataset. We chose K-Means, because of its efficiency and simplicity in segmenting the customers. (More details of the PCA analysis can be found in the appendix)

Key Insights:

Bridging the Gap Between Willingness and Subscription: A noteworthy discovery unveiled an obvious divergence between user willingness to subscribe and the actual prevalence of premium subscriptions. These differences need to be further explored.

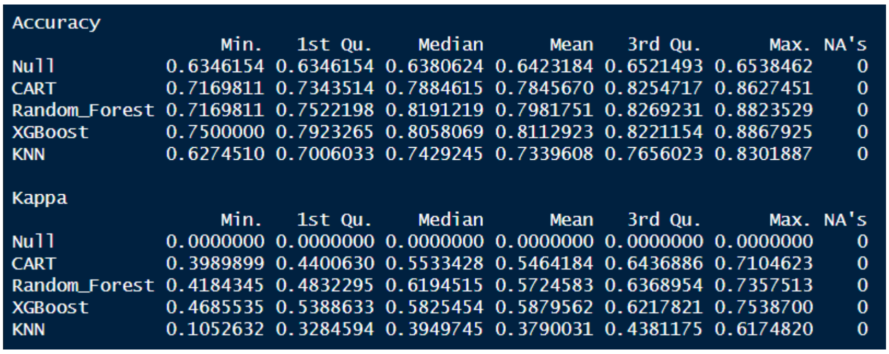
High Satisfaction but Low Willingness (Cluster 4): We identified a specific segment, Cluster 4, characterized by high user satisfaction but a marked absence of willingness to subscribe. This finding underscores the importance of channeling marketing efforts more judiciously.

While PC1 leans more towards capturing variance in user preferences and behaviors around content and subscription, PC2 emphasizes timing, device, and mood in relation to listening habits, alongside some demographic and usage details. These components enable a nuanced understanding of different aspects of user interaction and preferences with the platform, offering a foundation for further analysis and potential recommendation or personalization strategies.

Modeling

Core Task: Use classification model in deciding which segment to target based on their probability of converting through their indication of willingness to subscribe and other features.

Model Selection: For model selection, we chose to include the Null model to benchmark the 4 models' performance. We chose 4 other classification models and proceeded to compare these potential models with 10 k-fold cross validation to choose the model with the best Out-of-Sample (OOS) performance. We ended up choosing the Random Forest model for classification because it had the highest median accuracy and Kappa score.



```
Accuracy
                   Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
Null          0.6346154 0.6346154 0.6380624 0.6423184 0.6521493 0.6538462    0
CART          0.7169811 0.7343514 0.7884615 0.7845670 0.8254717 0.8627451    0
Random_Forest 0.7169811 0.7522198 0.8191219 0.7981751 0.8269231 0.8823529    0
XGBoost       0.7500000 0.7923265 0.8058069 0.8112923 0.8221154 0.8867925    0
KNN           0.6274510 0.7006033 0.7429245 0.7339608 0.7656023 0.8301887    0

Kappa
                   Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
Null          0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000    0
CART          0.3989899 0.4400630 0.5533428 0.5464184 0.6436886 0.7104623    0
Random_Forest 0.4184345 0.4832295 0.6194515 0.5724583 0.6368954 0.7357513    0
XGBoost       0.4685535 0.5388633 0.5825454 0.5879562 0.6217821 0.7538700    0
KNN           0.1052632 0.3284594 0.3949745 0.3790031 0.4381175 0.6174820    0
```

Figure 7 Screenshot of Random Forest model performance in R

Using the Random Forest model, predictions on the test set were generated and evaluated. The customer segment and the accuracy of the model would give us good indicators of which segment to target and how likely they were to switch from being a free user to a paid user. The main pattern we sought to mine from the data was to predict the probability of users being willing to subscribe to a premium service.

Solving the Business Problem: The Random Forest model is designed to improve user targeting for marketing campaigns. By predicting the likelihood of a user being willing to subscribe, we can focus our marketing resources more efficiently, leading to higher conversion rates and greater ROI.

Model Evaluation: After running some predictions on the test data set, we generated a confusion matrix and a ROC curve to evaluate how well the model was predicting potential subscribers. Results were evaluated by generating a confusion matrix and drawing an ROC curve determining the True positive rate vs. False positive rate of the predictions. The ROC curve shows that the model performs quite well in predicting customers who are willing to subscribe.

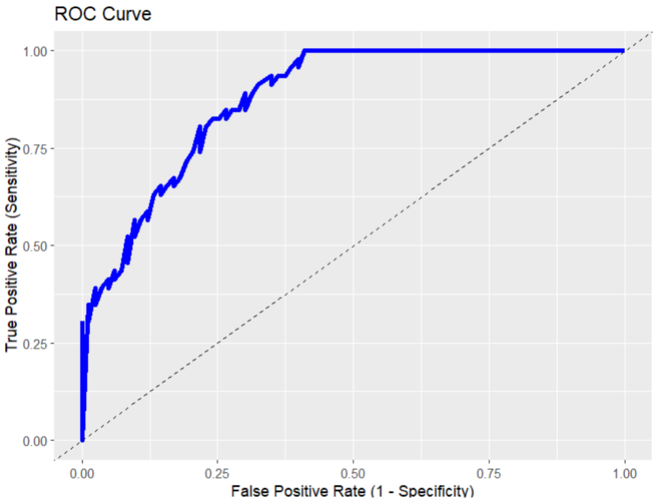


Figure 8 ROC Curve

To develop a business case, we would project the expected improvement in conversions (subscriptions) using the model predictions. These projections can be compared to historical conversion rates to estimate the potential ROI. If projecting ROI becomes challenging due to unpredictable user behavior or other external factors, alternative measures could include user engagement metrics or feedback post-campaign.

Deployment

To optimize Spotify's marketing operations and increase premium subscriptions, the clustering model will be combined with the predictive classification model to better target Cluster 6. The data-informed deployment strategy accentuating Cluster 6 is outlined below.

1. Deployment Process and Implementation

Data Mining Insights: Our meticulous clustering analysis reveals "Cluster 6 (Bright Sky Blue)" as a pivotal segment, with a striking 66.7% willingness rate to subscribe. Insights like these guide our deployment.

Marketing Message Personalization: Using insights from Principal Component Analysis (PCA), particularly the 'ContentPodcast' feature, we can craft targeted marketing messages to emphasize premium podcast benefits, such as ad-free listening, high-quality audio, and offline listening. Additionally, we could introduce targeted discount incentives for podcast listeners, offering limited-time promotional pricing for Premium subscriptions. This approach combines the value proposition of premium features with monetary incentives, potentially boosting conversions. Continually optimize this marketing strategy by employing methods like A/B testing to refine messages, ensuring the best outcome for both user experience and subscription revenue growth.

Predictive Analysis Integration: The chosen Random Forest model will be a linchpin, forecasting the likelihood of users from Cluster 6 migrating to premium subscriptions. These predictions will be integrated into our outreach strategy to zero in on high-conversion ad-supported users.

2. Deployment Concerns for Spotify

Scalability: As we upscale our efforts targeting Cluster 6, Spotify should ensure that backend operations, including customer service and podcast content delivery, can handle increased traffic and engagement.

Model Validity Over Time: These Models will need regular tuning to maintain their predictive accuracy. Continuous monitoring of model performance against real-world outcomes is key especially since user preferences can change at any time.

3. Ethical Considerations

Data Privacy: Respecting users' data privacy is non-negotiable. All personal information should be anonymized, and data usage must conform to prevailing regulations and standards.

Transparent Communication: Marketing messages should avoid overpromising. Users must be clearly informed of what premium subscriptions entail.

4. Risks and Mitigation

Data Bias: Potential response bias may be present in the data due to how the data was collected

Over-reliance on Cluster 6: Banking predominantly on one segment might alienate other potential segments. A diversified approach, albeit with a focus on Cluster 6, can provide a balanced strategy.

Overfitting of the Model: An overfit model might not generalize well in real-world scenarios. Cross-validation techniques and using diversified data subsets will be crucial.

Potential Backlash from Over-targeting: Bombarding Cluster 6 users with excessive marketing messages can be counterproductive. Frequency capping and ensuring message variety prevent this.

5. Feedback Loop and Evolution

Continuous monitoring of campaign outcomes against model predictions will fuel iterative improvement. This feedback mechanism is essential to finetune both our model and marketing strategies.

6. Adaptive Recalibration:

With a steady stream of incoming user data, periodic model recalibrations will keep our strategy aligned with evolving user behaviors and market dynamics.

Conclusion

Effective deployment is not just about leveraging data-mined insights but seamlessly integrating them into a holistic marketing framework. With marketing focus on Cluster 6 steering the strategy, we should see an increase in premium subscriptions which will increase both revenue and profit generation for Spotify.

References

Spotify. "Spotify's Revenues from 2012 to 2022, by Segment (in Million Euros)." Statista, Statista Inc., 31 Jan 2023, https://www-statista-com.proxy.lib.duke.edu/statistics/245125/revenue-distribution-of-spotify-by-segment/

“Spotify User Behavior Dataset.” *Www.kaggle.com*, Meera Ajayakumar, July 2023, www.kaggle.com/datasets/meeraajayakumar/spotify-user-behavior-dataset.

Spotify. "Operating Income of Spotify Worldwide from 2013 to 2022 (in Million Euros) ." Statista, Statista Inc., 31 Jan 2023, https://www-statista-com.proxy.lib.duke.edu/statistics/813758/spotify-operating-income/