Customer Churn Analysis for a Music Streaming Service

Aastha Shah, Aashray Saini, Laura Gil Ortiz, Saswati Devi

Table of Contents

1. Executive Summary

One of the key challenges faced by major companies that are in the growth or maturity stage and are aiming for sustainable growth is “Customer Churn”. We chose this problem as the main objective of our group project as this is a real and urgent problem many complex organizations are grappling with. We selected a dataset for an imaginary music streaming service, similar to Spotify or Apple Music. We did further analysis and data modeling to generate predictive business insights and provided potential solution recommendations on the basis of those insights.

2. Background

Today, there is a rising competition amongst music streaming services such as Spotify, Apple music, Youtube music, Amazon music etc. Customers have multiple options to choose from and the switching cost is relatively low or non-existent. If a customer does not like a service, they either downgrade to a cheaper/free version or move on to another service. Therefore, customer churn is an imminent problem that these music services struggle with on a daily basis.

Data scientists and executives in a company constantly scrutinize available data to solve this complex problem to predict the probability of churn so that it does not affect a company’s profit and provide a better customer experience. This problem is also particularly time-sensitive, because if it is identified before a customer leaves the service, then curative measures can be undertaken. The company incentivizes the dissatisfied customer using various services, such as discounts, coupons, or offers, to avoid churn. Moreover, companies have a practical strategic reason for doing so. According to research[1], it is 5 times more costly to acquire a new customer than to retain an existing one.

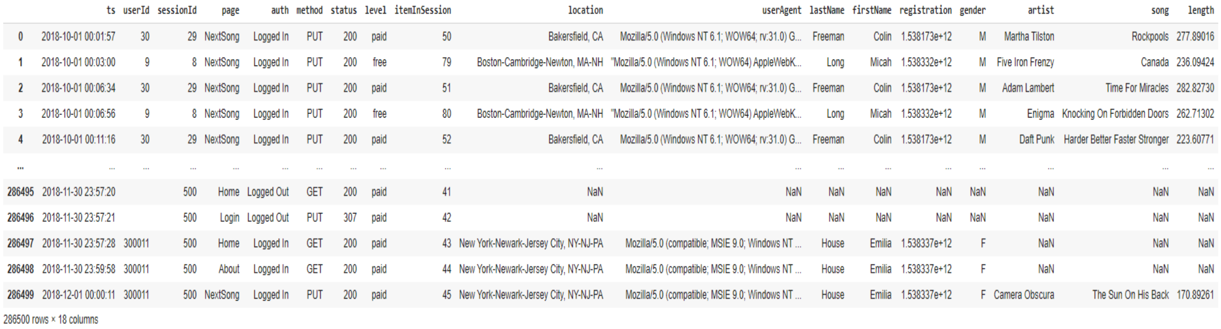
To generate quantitative and qualitative insights, music services gather real-time consumer behavior data. Companies gather data from customers in a variety of ways. The customers allow user data collection by accepting terms and conditions of a service. The more the users interact with the service the more data it generates to enhance customer experience by providing personalized content to increase customer satisfaction, resulting in higher engagement and revenue increase.

3. Data Description and EDA

For our project, the data (mini_audiofy_event_data.json) includes both macro and micro information about customers. Macro information such as customer demographics, location, account status (upgrade/cancellation), account type (free/premium), and device type are asynchronously monitored when a user logs in their account. Similarly, micro information such as timestamp, session ids, pages visited, and actions taken before canceling the services are also collected in real-time. Therefore, based on this information, we did exploratory data analysis to describe and clean up the data and built a supervised model to predict whether a customer will cancel the service or not, and how the company can improve to avoid this from happening.

3.1 Available Data

Our dataset for the project comes in a JSON format and contains 286500 rows × 18 columns. The data is at a session level with each session having a corresponding timestamp.



	ts	userId	sessionId	page	auth	method	status	level	itemInSession	location	userAgent	lastName	firstName	registration	gender	artist	song	length
0	2018-10-01 00:01:57	30	29	NextSong	Logged In	PUT	200	paid	50	Bakersfield, CA	Mozilla/5.0 (Windows NT 6.1; WOW64; rv:31.0) G...	Freeman	Colin	1.538173e+12	M	Martha Tilston	Rockpools	277.89016
1	2018-10-01 00:03:00	9	8	NextSong	Logged In	PUT	200	free	79	Boston-Cambridge-Newton, MA-NH	"Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebK...	Long	Micah	1.538332e+12	M	Five Iron Frenzy	Canada	236.09424
2	2018-10-01 00:06:34	30	29	NextSong	Logged In	PUT	200	paid	51	Bakersfield, CA	Mozilla/5.0 (Windows NT 6.1; WOW64; rv:31.0) G...	Freeman	Colin	1.538173e+12	M	Adam Lambert	Time For Miracles	282.82730
3	2018-10-01 00:06:56	9	8	NextSong	Logged In	PUT	200	free	80	Boston-Cambridge-Newton, MA-NH	"Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebK...	Long	Micah	1.538332e+12	M	Enigma	Knocking On Forbidden Doors	262.71302
4	2018-10-01 00:11:16	30	29	NextSong	Logged In	PUT	200	paid	52	Bakersfield, CA	Mozilla/5.0 (Windows NT 6.1; WOW64; rv:31.0) G...	Freeman	Colin	1.538173e+12	M	Daft Punk	Harder Better Faster Stronger	223.60771
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
286495	2018-11-30 23:57:20		500	Home	Logged Out	GET	200	paid	41	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
286496	2018-11-30 23:57:21		500	Login	Logged Out	PUT	307	paid	42	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
286497	2018-11-30 23:57:28	300011	500	Home	Logged In	GET	200	paid	43	New York-Newark-Jersey City, NY-NJ-PA	Mozilla/5.0 (compatible; MSIE 9.0; Windows NT ...	House	Emilia	1.538337e+12	F	NaN	NaN	NaN
286498	2018-11-30 23:59:58	300011	500	About	Logged In	GET	200	paid	44	New York-Newark-Jersey City, NY-NJ-PA	Mozilla/5.0 (compatible; MSIE 9.0; Windows NT ...	House	Emilia	1.538337e+12	F	NaN	NaN	NaN
286499	2018-12-01 00:00:11	300011	500	NextSong	Logged In	PUT	200	paid	45	New York-Newark-Jersey City, NY-NJ-PA	Mozilla/5.0 (compatible; MSIE 9.0; Windows NT ...	House	Emilia	1.538337e+12	F	Camera Obscura	The Sun On His Back	170.89261

286500 rows × 18 columns

3.2 EDA and Data Preparation

First we started by importing the dataset into our workspace and cleaning the data by removing outliers and null values. Subsequently, with the updated data, we found a total of 225 unique users. Additionally, 165 and 195 users were found to be paid and free users respectively, which signifies that a few users moved from paid to free subscription.

After conducting preliminary study on unique users, we define the “numbers of churned users” by calculating the number of unique users who visited the “Cancellation Confirmation” page as the age visit signifies that the particular user canceled their subscription. The following subsections outline the steps we followed to do EDA analysis step-by-step.

3.2.1 Data Description

As mentioned before, the dataset has information for 225 unique users for activity duration from

10/1/2018 to 12/03/2018. Out of the 225 users:

1) 46% were female and 56% were male as seen in Figure 1.
2) The top 5 states where users lived were California, Texas, New York, New Jersey, Pennsylvania, Florida and Arizona, as seen in Figure 2.

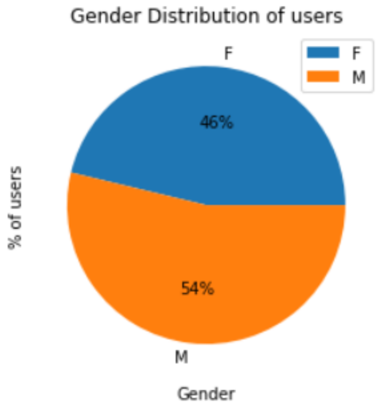


Fig 1

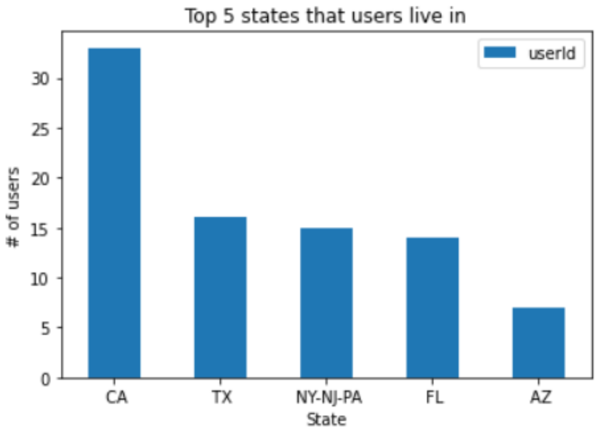


Fig 2

3.2.2 Defining the Target Variable

For the purposes of our project, we defined our target variable to be the number of users who have churned, i.e., users who visited the page “Cancellation confirmation”. Therefore, the churn rate was calculated to be 23%. The other pages that were available to the user through app are elaborated in Fig. 4 below. Next, we did a comparative analysis between churned and non-churned users.

3.2.3 Comparison of Churned v/s Non-churned users

The following tables highlight the distinctions between churned and non-churned users on the basis of various factors.

1. **Comparison based on the Tier (free/paid)**

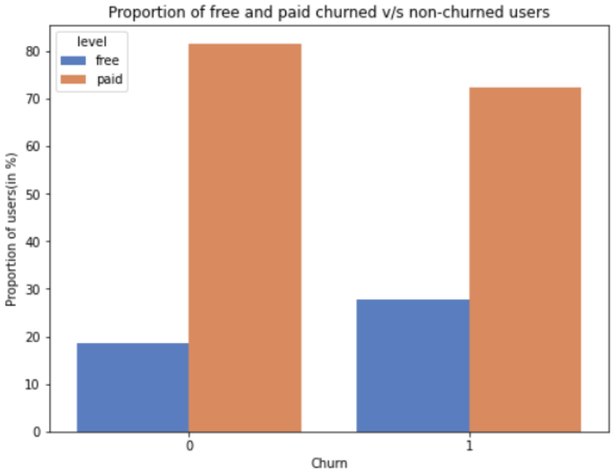


Fig 3

There are two important takeaways from Figure 3: 1) For the users in the free tier, the proportion of churned users is more than non-churned users, and 2) for the users in the paid tier, the proportion of churned users is less than the non-churned users.

2. Comparison of pages visited

To learn further about the user behavior when it comes to churn risks, we analyzed their behaviors w.r.t. pages visited as shown in Figure 4.

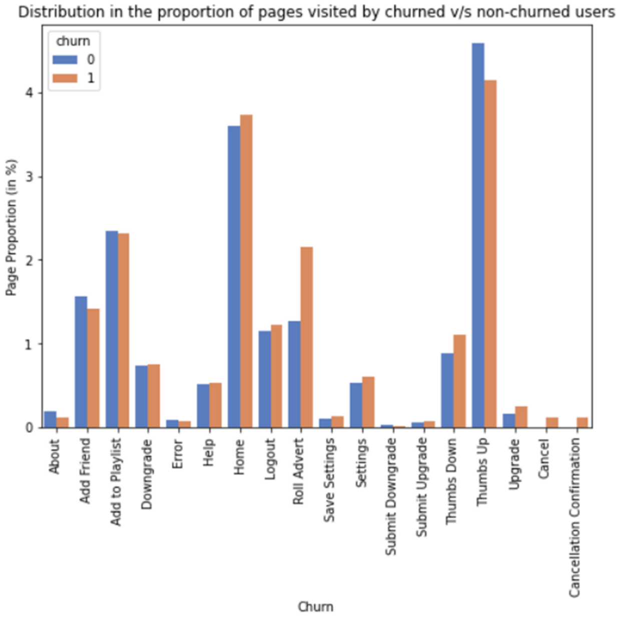


Fig 4

As shown in Figure 4, churned users visited the pages “Roll Advert”(Roll Advertisement) and “Thumbs Down” more than non-churned users. As seen from section 1, since many of the churned users belonged to the free tier, naturally they would be seeing a lot of advertisements. So, they might churn as they might get bored of these ads. Additionally, churned users have disliked more songs (through higher “thumbs down” clicks). So this could mean that they are not getting their preferences and that is why they might churn.

In contrast, we see that the non-churned users have given more “Thumbs up” to songs than churned users. This further validates our hypothesis that churned users probably are not able to find songs that match their preferences.

3. Comparison of session activity of the users

Next, we analyzed how the sessions differ between churned and non-churned users. As shown in Fig. 5, on an average, churned users have fewer sessions than non-churned users. Similarly, from Fig. 6, on average, the number of items(activities) in each session is lower for churned users as compared to non-churned users.

Finally, from Fig. 7, avg session duration is also shorter for churned users compared to non-churned users.

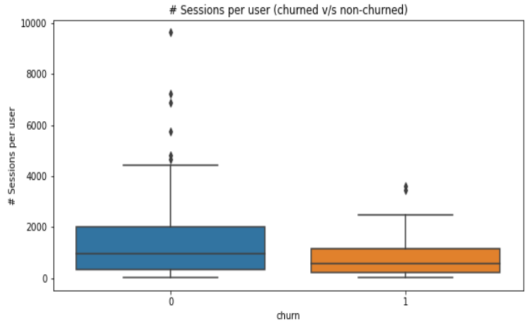


Fig 5

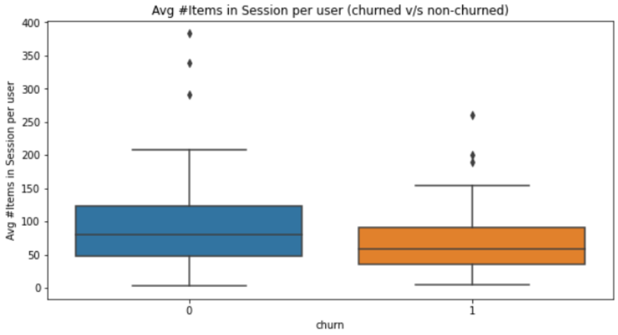


Fig 6

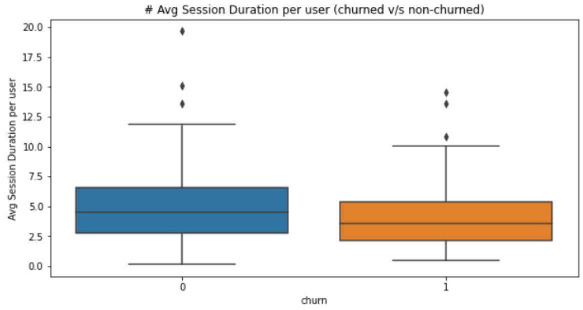


Fig 7

4. Comparison of user engagement

After analyzing various session parameters , we continued to observe the kind of songs and artists that churned and non-churned users listen to.

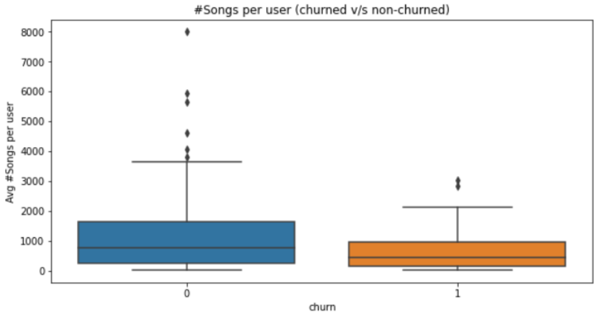


Fig 8

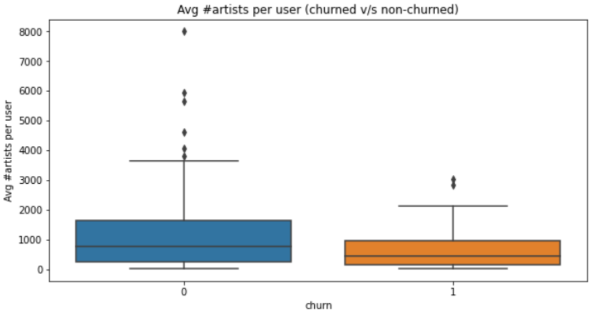


Fig 9

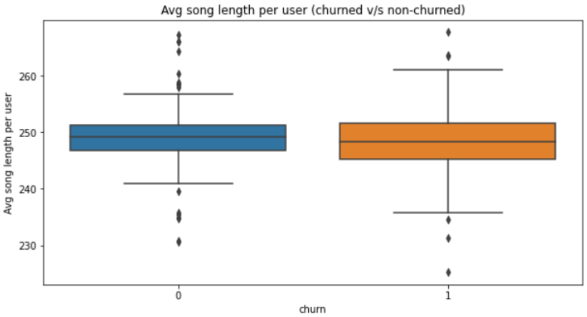


Fig 10

From Fig. 8 and 9, we can see that churned users listen to less number of songs on an average than non-churned users. We can also see that on an average, churned users listen to a smaller variety of artists as compared to non-churned users. This might be because they might not be liking what they listen to and so might lose interest in listening to other artists.

However, from Fig. 10 we can see that there is not much of a difference between the average song length of songs listened by churned users v/s non-churned users.

5. Comparison of demographic data of the users

We also observed some key differences in the demographics of churned v/s non-churned users.
As shown in Fig. 11–
1) Among the churned users, the proportion of males is more than that of females
2) Among non-churned users, the proportion of females is more than that of males.

So we can say that males are at a higher risk of churning as compared to females.

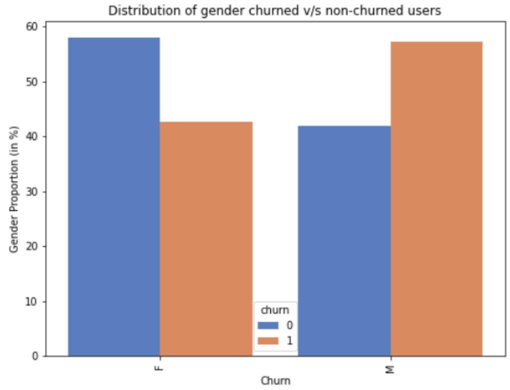


Fig 11

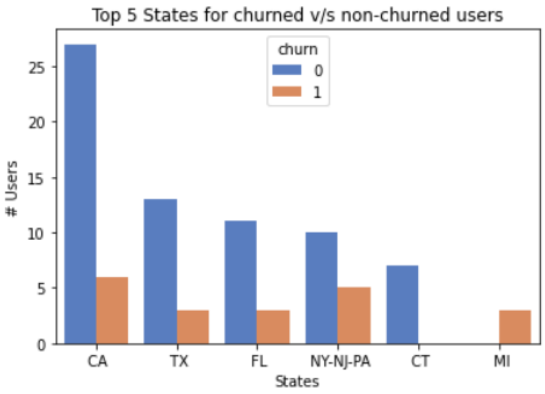


Fig 12

Also, Fig. 12 displays the top 5 states for churned and non-churned users. As shown in the figure, most of the states have both churned as well as non-churned users, except Connecticut that has all non-churned users and Michigan that has all churned users.

From all these insights into the churned users and their behavior, we now move on to creating our model to predict Customer Churn.

4. Model Development & Evaluation

EDA done in the prior section helped us not only clean the dataset of any missing or Null values, but also generate preliminary hypotheses to inform feature engineering.

Feature Engineering

For feature engineering, considering we are predicting individual user's propensity to churn, we decided to aggregate features (listed below) at the user-level based on the EDA. However, in the dataset, there is a class imbalance problem, i.e., the ratio of churned users to non-churned user is 52 to 123 (23%). Therefore, for tackling class imbalance we chose oversampling for balancing the class using SMOTE (Synthetic Minority Oversampling Technique) which oversamples minority classes at the boundary of classes.

Feature	**Explanation**
Num. of sessions	Number of sessions the user had
Avg. session duration	Average duration of each session in minutes
Gender	Gender of the user
Avg. daily number of songs	Number of songs listened daily on average

Avg songs per session	Number of songs listened per session on average
Avg daily thumbs up	Number of thumbs up given to songs daily on average
Avg daily thumbs down	Number of thumbs down given to songs daily on average
Avg daily friends	Number of friends added daily on average
Avg num of “Add to Playlist”	Number of songs added to playlist on average
Active days	Days since user has been active
Latest level of user before they churn	Last clicked page before canceling the subscription

Moreover, we selected Logistic Regression, which works great with smaller datasets, which was optimal for the project dataset. For model evaluation, we chose K-Fold cross-validation. Subsequently, the model outcomes helped us prioritize features that had the highest impact on it. Next, we segmented “at-risk churn” levels of the users and the impact of said features on those segments.

In addition, it is known that in a classification problem, there are 2 types of errors: a) false positives, and b) false negatives. Depending on the business needs and cost of errors, the discrimination threshold value is chosen. In this scenario, false positive (FP) is predicting that a customer will churn when they are not going to and incentivizing them to stay whereas false negative (FN) is predicting someone will not churn when they are actually going to. Considering our business goal is to correctly identify people who are going to churn, FN is costlier than FP, Hence, we are focusing on optimizing the false negatives. The precision, recall, and F1 scores are illustrated in Figures 13 and 14. Also, in fig. 14 it is evident that 0.33 seems a good cut-off point to ensure good recall, as well as precision and F1.

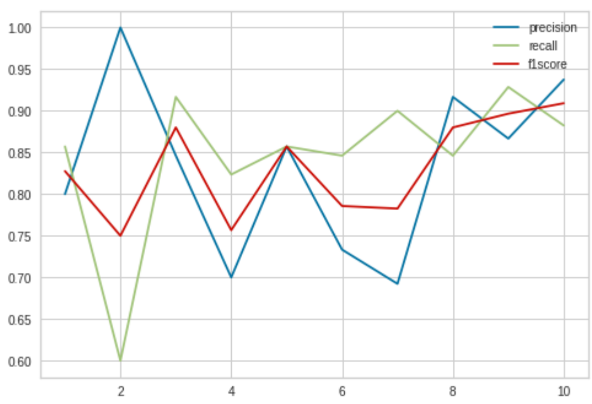


Fig 13

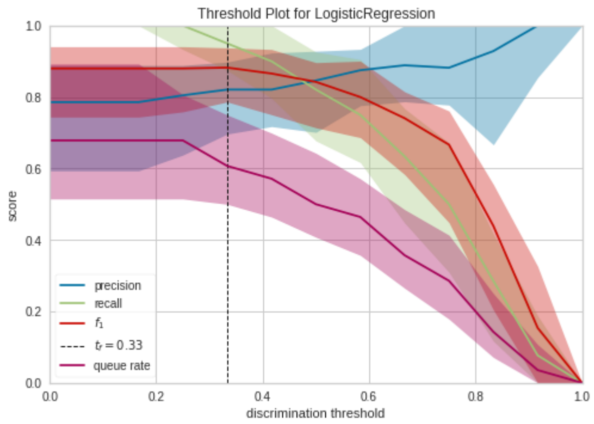


Fig 14

Fig 15

5. Model interpretation

In our predictive model, we not only assigned the label, but we also assigned a risk score, which is the probability of a customer churning. We then segmented those probabilities into deciles which are represented by the values of 1, 2, 3, 4, 5, with 1 bein the segment of customers who are at the highest risk of churning and specifically 4 and 5 are the customer segments that are at the lowest risk of churning.

Fig. 16 represents that when a segment of customers have not been active for long, it probably means that they are unhappy with the product or there is no product-market fit. Similarly, figures 17 and 18 denote that customers who are more likely to churn have higher daily avg. thumbs down which indicates that they are not able to quickly discover songs of their preferences and could be a noteworthy customer pain point to investigate.

Finally, Fig. 19 ranks the features that are important for predictions. The most important feature is active days, which denotes that the users that are inactive for long are more likely churn (class 1 in customer segment) and vice versa. Similarly, other features that important factors in users' likelihood of churn are daily_avg_thumbs_down, level_paid, and daily_avg_roll_advert.

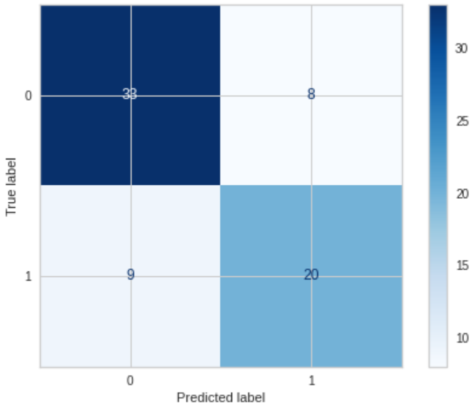
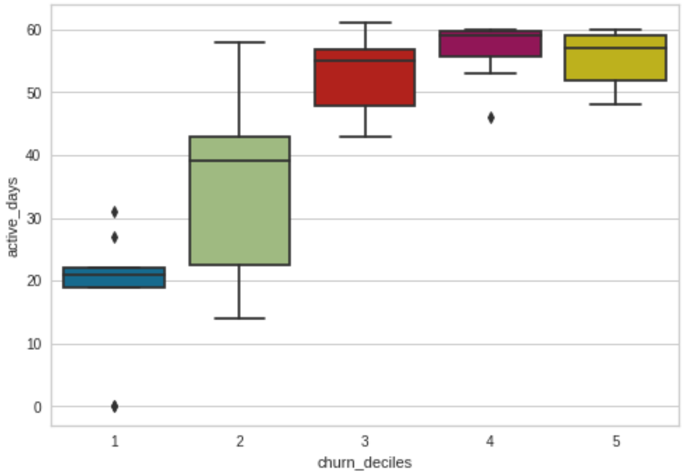


Fig 16

Fig 17



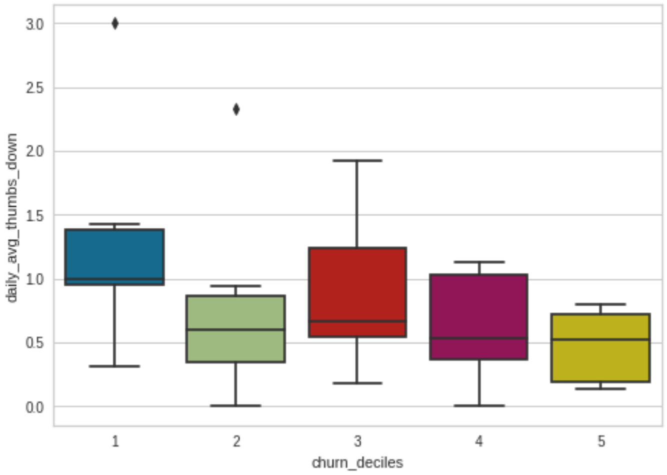
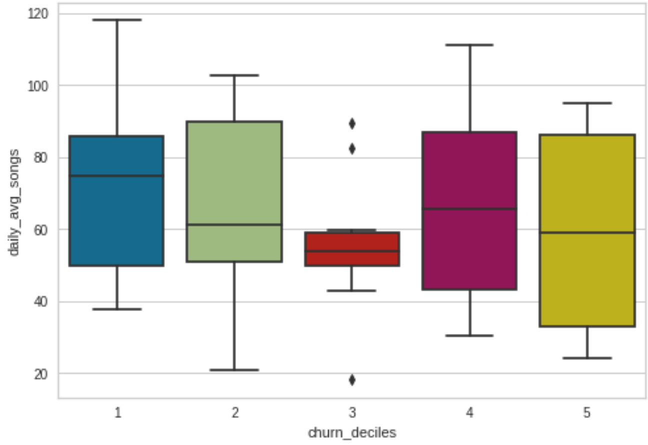
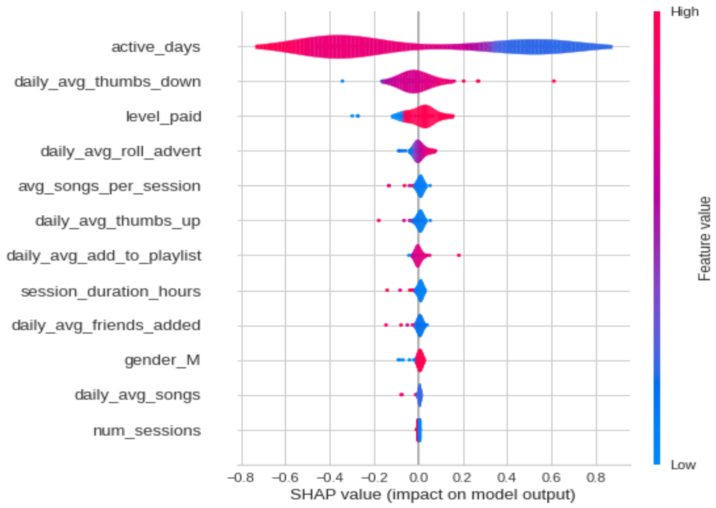


Fig 18

Fig 19





6. Recommendations

Our analysis and predictive model generated qualitative insights that informed our potential solutions recommendation plan. We then created a roadmap to prioritize a phased action plan.

Short-term:

1. For high-risk churn customers who have been inactive for a while, we recommend sending email notifications as gentle reminders for renewed engagement as well as discounted pricing to upgrade.

2. For customers that have higher avg thumbs down rate, we would like to enable surveys/ratings to gather more information for a better user preferences alignment. We also would recommend to improve UI features to not surface the songs disliked by the user.
3. In addition, it would also be beneficial to provide or start looking into service bundling or discounted options (student, family plan etc.) to upgrade free users into paid user category.
4. To increase engagements of inactive users who are likely to churn, we recommend referral links to enlist their family and friends into the music streaming service.

Long-term:

1. To validate the effectiveness of short-term solutions implementation (treatment effect) by implementing the changes to selected group of customers predicted to churn.
2. Finally, for making a product-level change to prevent churn, we recommend optimal pricing strategy for all tier levels that aligns with customers' willingness to pay (WTP). Additionally, the company should focus on continuous improvement on their backend algorithm to enhance customer experience. This could also be achieved through A/B testing and procuring high quality music catalog to increase their value proposition.