A Music Recommendation System Scenario Using CRISP-DM

Author

elisa.ancarani4@studio.unibo.it

Project Report for the Data Mining, Text Mining and Big Data Analytics Course

September 16, 2023

Premise

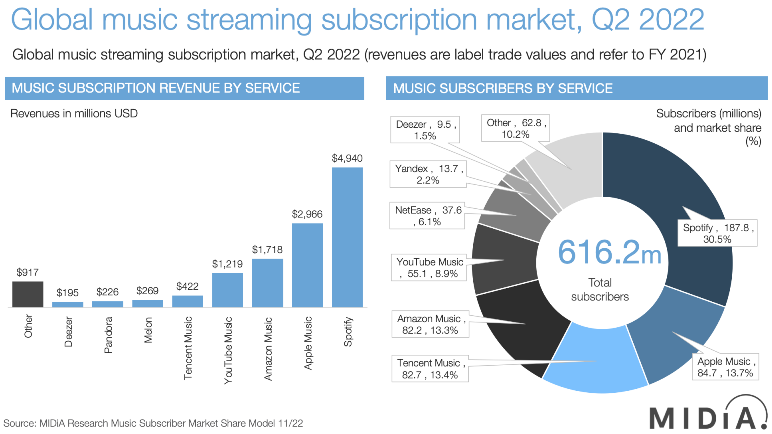
This project is inspired by a Kaggle competition dataset provided by KKBox [1]. However, it departs from the traditional competition approach and adopts the point of view of a consultant, tasked with creating a music recommendation system for KKBox [1]. Guided by the principles of the Cross-Industry Standard Process for Data Mining (CRISP-DM), this work focuses on business solutions for KKBox. Even though this is a sample project, it strives for a genuine application of the methodology to showcase a potential real-world scenario. The main CRISP-DM documentation that is followed for the conduct of this project is that of: [2].

Contents

Chapter 1: Business Understanding

1 Introduction

Nowadays streaming platform are the main mode of music consumption. IFIP research finds that streaming is the most popular way of listening music and streaming platform represented the 65% of music revenue in 2021. According to [3] revenue from music streaming has increased every year since 2012. Moreover, in 2022 music streaming apps generated $43.3 billion revenue for music industry, a 15% increase on the year prior. Midia Research [4] provided an overview of the global music streaming subscription market, counting a total of 616.2 million in 2022.



Source: MIDiA Research Music Subscriber Market Share Model 11/22

Moreover, advancements in technology led to a growth in music production as hardware and software for music recording are more accessible [5]. Thus, making also easier for emerging artists to produce and release music. Hence, the use of recommendation systems and their optimisation is a today's challenge.

Recommendation Systems [6] are AI-based algorithms that aims to anticipate user preferences by predicting their interests and suggesting product items that are likely to be of interest to them. These systems are able to deal with a large amount of data (big data) and to filter relevant information based on the user preferences and interest.

Music recommendation system are a type of recommendation systems whose aim is to enhance music and artists discovery and customise users' listening experiences by predicting the preference that a user would give to a song [7]. This project applies the CRISP-DM methodology to guide the data mining and machine learning process for the implementation of a music recommendation system for KKBox. The streaming platform offers over 30 million Asia-Pop track, Asia's leading music streaming service. The company was founded in Taiwan in 2004, and currently serves Taiwan, Japan, Hong Kong, Singapore, and Malaysia. Its market share in Taiwan exceeds 60%, and its current music library exceeds 40 million songs and the platform accounts over 12 million users [8].

2 Business Objectives

In this project consists in the implementation of a music recommendation system for KKBox. The main objective of this system is to predict the likelihood of a user repeatedly listening to a song after the first observable listening event within a specific time window. The projects wants to achieve the following objectives:

- **Enhance User Engagement**: by recommending songs that users are more likely to listen again, KKBOX can the dwell time of users on the platform, improving user engagement.
- **Retention and Loyalty**: KKBOX can create a more engaging music experience, which leads to increased user loyalty and retention. Satisfied users are more likely to continue using the service, resulting in higher customer retention rates and reduced costumer churn. This can have a positive impact on the overall growth of the streaming platform.
- **Increase the Market Share**: A good service, as well as preventing users from leaving, can attract new potential users and increase market share.

2.1 Business Success Criteria

In a hypothetical real-world scenario for KKBox, business success criteria would be built upon quantifiable metrics to demonstrate the value proposition of the project to stakeholders. Considering that specific numerical targets cannot be set, due to the unavailability of actual data, here are the aspirational goals, positioned as if presenting to KKBox stakeholders:

- **Loyalty Enhancement**: The success of the recommendation system would be gauged by a hypothetical increase in user retention rates. This would mean achieving a significant percentage growth in users consistently returning to the platform, suggesting enhanced user satisfaction.
- **Boosted Engagement**: A key objective would be to increase user engagement metrics. This translates into longer sessions on the platform, more frequent visits and increased interaction with the platform's features – all pointing towards a positive and engaging user experience and a reduced customer churn.
- **Augmented Song Consumption**: To affirm that the music recommendations resonate with users, a hypothetical target would be an increase in the average number of songs played per user within a specific timeframe.

Employing the CRISP-DM methodology ensures methodical progression through each phase, fostering clarity and efficiency. With this approach, the project gains a deeper understanding of user behaviour, iteratively refines models performance and identifies potential risks at an early stage, culminating in a robust and effective recommendation system. Moreover, the tool can also potentially allow to detect music trends. For instance, when numerous users repeatedly tune into the same song, trends can be identified and the audience segmented, paving the way for future targeted recommendations.

3 Assess the Situation

3.1 What sort of data are available for analysis?

The dataset available for this project was provided by KKBox to build a better recommendation system for the 11th ACM International Conference on Web Search and Data Mining (WSDM 2018). In particular, the company provides four different datasets. A **training set** which contains information about a triggered event: user and song ids, system tab from which the event was triggered, entry point of the song and target variable. The target=1 means there are recurring listening event(s) triggered within a month after the user's very first observable listening event, target=0 otherwise. There are three additional dataset, **members**, **songs** and **song extra info**. The first contains anonymous information on subscribers, such as origin, age, gender and subscription information. The second and the third contain information about the songs like song id, artist, composer, song name and International Standard Recording Code which can be used as song identifiers.

3.2 Resource Inventory

Hardware Resources: The hardware required will depend on the size of the data and model complexity. In this project I rely on Google Colab to train the algorithms on the data, taking into account the limitations of the GPU provided by the service.

Identify Data Sources and Knowledge Stores: Kaggle is an online platform for data science and machine learning (which includes many of its subfields such as nlp, computer vision,..), where users can create notebooks with Python or R code. Kaggle also provides cloud-based storage for a large collections of datasets which user can access directly from their notebooks. The datasets can be publicly available or private, depending on the dataset owner's settings. However, the exact details of how and where the platform keeps the data are not provided.
The service hosts data science competitions with real-world datasets provided by companies and organisations. In honour of one of those competitions, namely the 11th ACM International Conference on Web Search and Data Mining (WSDM 2018), KKBox provided four datasets which information about subscribers, songs and interactions of a user with a song. The data are publicly available and accessible to anyone. All the data are already provided by the Company, including demographic information about the users of the platform. Thus, there's no need to purchase any external data. Moreover, since the data are publicly available, there are no security issues preventing access to required data.

Identify Personnel Resource To be determined in a real case. This phase leads to questions on the access to the personnel needed such as business and data experts, database administrators and other support staff.

3.3 Requirements, Assumptions and Constraints

Requirements: The access to the data is a fundamental requirement for the project. Moreover, data has to adhere privacy regulations and laws. Furthermore, in order to apply machine learning models, a significant amount of data is also required. This ensures the models are trained adequately, leading to more accurate and reliable predictions and recommendations.

Assumptions: Concerning the **econonimical** aspect, in this specific case there are no economic factors that may influence the project, as it is just a sample machine learning project that applies the CRISP-DM methodology.
With regard to the **data**, it is assumed to be complete, accurate and compliant with the privacy regulations. Any issues have already been resolved during the processing of these, a part not dealt with in this project.
Being a streaming music service, there is an opportunity to increase engagement with customer data. Therefore, the **distribution of the results** is expected to take place within the KKBox company. This is due to the fact that are handled sensitive user data as well as sensitive business data. Therefore, quantitative results should therefore not leave the company.

Constraints: To be determined in a real case. In this context, there are no financial, legal or accessibility constraints.

3.4 Risk, Opportunities, Costs and Benefits

As with any predictive modelling project, especially in a dynamic domain such as music streaming, there are inherent risks, opportunities and costs that must be weighed against the potential benefits.

Risks:

- Predictive models may not work as expected, resulting in failure to achieve the desired performance. Moreover, the regular fine-tuning of the model is a necessary but time-consuming task.
- There could be already existent and better working algorithm than the one developed.
- Data might be incomplete and/or biased. For example, data could be skewed if it only captures listening habits from 8 pm to 12 pm, which might primarily represent younger audiences. Age could introduce bias because of different user behaviors. Additionally, the non-representativity of the data is also a risk.

Opportunities:

- Meeting the machine learning objectives could lead to business growth.

Benefits:

- A well-functioning recommendation system can improve the user experience, increasing user engagement and loyalty.
- Successfully achieving machine learning objectives can increase user interaction, leading to potential revenue growth.
- Developing a proprietary recommendation system or integrating a market-leading solution can enhance the platform's competitive edge.
- An in-depth data analysis can provide valuable insights into user behavior, emerging music trends, and potential new business opportunities

Costs : Costs have to be determined in a real-case.

After outlining risks, opportunities, costs and benefits, I assess the project's internal and external factors through a SWOT analysis which summarises the strengths, weaknesses, opportunities and threats in a structured manner.

Strengths	Weaknesses	Opportunities
Clear Business Objectives	Predicting user behaviour is challenging. External factors influencing song replays can affect the model's predictions, such as emotions, bad/good memories.	A successful music recommendation system can enhance user experience, longer dwell times, and higher customer satisfaction.
There's access to different data such as, listening events, users, and songs, which provides a solid foundation for implementing predictive models.	Data Quality Issues	An efficient system can give a competitive edge over other streaming platforms, attract new users and retain existing ones.
KKBox likely has established data governance protocols to ensure data security, privacy, and compliance with regulations.	The analysis of users' personal data may raise potential privacy issues that need to be addressed by data protection measures.	Reduce customer churn by creating a more personalised user experience

Threats
The music streaming sector is highly competitive and other platforms also invest in recommendation systems, making it crucial for KKBOX to stay ahead in the race.
User preferences and behaviors can evolve over time.
Changes in privacy and data protection regulations may impose restrictions on the use of data and the implementation of the model, affecting the progress of the project.

Table 1: SWOT Analysis

4 Data Mining Goals

The goal of this project is to implement a predictive system which predicts the likelihood of a user of repeatedly listening to a song after their first observable listening event within a specific time window. Specifically, if there are recurring listening event(s) triggered within a month after the user's first observable event, in the training set the target variable is marked as 1, and 0 otherwise. The project will follow the CRISP-DM methodology to ensure a structured and systematic approach to data mining.

- **Trends & Insights** Trend identification helps to understand which songs are most listened to. A lot of insights can be gained through statistics and exploratory data analysis (EDA).
- **Predictive Model Development** Use historical events related to user-song interactions to predict the probability of a user repeatedly listening to a song.

4.1 Data Mining Success Criteria

The data mining success criteria are the benchmarks and measures which are essential for evaluating the performance of the data mining models and the overall success of the project. They may vary depending on the specific goals and objectives. The success of this project is determined by the accuracy and the ROC-AUC curve.

Methods for model assessment The models performance is assessed both the accuracy and the ROC-AUC score which evaluates the model ability to distinguish between positive and negative classes. Speed and Efficiency have also to be considered. Given the limited computational resources, the aim is to use a relatively lean model that can make accurate predictions in a relatively short time.

Benchmarks for evaluating success: Exact benchmarks for evaluating the success of the project have to be determined in a real case.

- Quantitative Benchmarks: These could be - Achieve an accuracy of above 70% and maintain an ROC-AUC score of at least 0.75
- Time and Efficiency Benchmarks: These could be - Model training time should be under 3 hours for the complete dataset.

Subjective Measurements:

- Interpretability: The ability to understand and interpret the model's decisions, especially for stakeholders who might not be technically inclined.
- Usability: The ease with which end-users, especially non-technical ones, can deploy and use the model.
- Relevance: The results from the model should be contextually relevant and actionable.
- Arbiter of Success: To be determined in a real case. Usually, a stakeholders or a review board, should be in charge of deciding if the subjective success criteria are met.

Deployment Considerations: To be determined in a real case. However, a model should be deployable in common environment and should be built in such a way that it can be easily updated with new data. Moreover, with the growth of the data, the model and its infrastructure should be able to scale accordingly without a significant performance degradation.

5 Producing a Project Plan

The project plan gives a comprehensive description of the steps to perform in order to achieve the business goal. It includes the selection of the tools and techniques. When creating a project plan one should answer these questions:

- Have you discussed the project tasks and proposed plan with everyone involved?
- Are time estimates included for all phases or tasks?
- Have you included the effort and resources needed to deploy the results or business solution?
- Are decision points and review requests highlighted in the plan?
- Have you marked phases where multiple iterations typically occur, such as modeling?

To answer these questions, there's the need to be in a real-world scenario. This phase should also be followed by an assessment of the tools and the techniques.

6 Assessing Tools and Techniques

In a real world scenario. These are the questions one should positively answer in order to go to the next data understanding phase.

From a business perspective:

- What does your business hope to gain from this project?
- How will you define the successful completion of our efforts?
- Do you have the budget and resources needed to reach our goals?
- Do you have access to all the data needed for this project?
- Have you and your team discussed the risks and contingencies associated with this project?
- Do the results of your cost/benefit analysis make this project worthwhile?

From a data mining perspective:

- How specifically can data mining help you meet your business goals?
- Do you have an idea about which data mining techniques might produce the best results?
- How will you know when your results are accurate or effective enough? (Have we set a measurement of data mining success?)
- How will the modeling results be deployed? Have you considered deployment in your project plan?
- Does the project plan include all phases of CRISP-DM?
- Are risks and dependencies called out in the plan?

Chapter 2: Data Understanding

KKbox has a huge pool of data on user interactions that have the potential to uncover patterns, preferences and pathways for business growth. In this phase of the CRISP-DM methodology, the primary objective is to examine the raw data, derive insights from it and identify potential pitfalls or gaps. At the end of this phase, the goal is to have a comprehensive overview of the existing data, thus laying the foundation for the next steps. This complete understanding is crucial for subsequently developing robust predictive models.

7 Dataset Description

The main goal is to predict whether a user will listen to a song multiple times after the first observable listening event. Several datasets, detailed below, are available to support this. These datasets include attributes related to user behavior, song details, and user-song interactions. Although the methods for collecting the data aren't specified, it's likely sourced from user interactions with the streaming platform, possibly captured using various database and logging tools. All datasets are in tabular format, stored as .csv files.

The `training set` has 7,377,418 entries and six feature columns. The target variable is numeric (binary) and indicates a recurring listening events within a month after the initial event. The other columns contain symbolic (object) values. These attributes - such as user ID, song ID, the system tab under which the event was triggered, the visual layout the user interfaced with, and the user's entry point for initiating a song play (like album or online playlist) - provides insights into user behaviours and song interactions.
The `members` dataset, with its 34,403 records and 7 attributes, provides information about users, including demographics and registration details.
Next, the `song` dataset has 2,296,320 entries, detailing info about the songs. A song is defined by its duration, genres, artist, composer, lyricist, and language. The length of the song is a numeric attribute and language is a floating-point number, while the rest of the features are symbolic. Along with this information is another dataset, `song-extra-info`, which contains the name of a song and its isrc code.
There's minimal overlap in the coding schemes of these data sources, mostly due to the nature of the attributes each dataset focuses on. The overlapping variables, such as msno and song id, are the keys for merging and relating datasets. These features maintain consistent coding schemes, ensuring smooth data integration.In summary, this extensive dataset collection provides a comprehensive overview of KKBox's user behaviors, song preferences, and interactions, laying the groundwork for building a predictive model for song repetitions. The tables below provide detailed information on the attributes of each dataset.

Column Name	**Description**	**Type**
msno	User ID	Object
song_id	Song ID	Object
source_system_tab	Tab where the event occurred	Object
source_screen_name	Name of the layout user saw	Object
source_type	Entry point for song play	Object
target	Indicates recurring listening events	Int64

Table 2: Training Data (The same applies to the test data except for the target variable)

Column Name	**Description**	**Type**
msno	User ID	Object
city	City	Int64
bd	Age	Int64
gender	Gender	Object
registered_via	Registration method	Int64
registration_init_time	Registration time (format %Y%m%d)	Int64
expiration_date	Expiration date (format %Y%m%d)	Int64

Table 3: KKBox Members Data

Column Name	**Description**	**Type**
song_id	Song ID	Object
song_length	Song length in ms	Int64
genre_ids	Genre IDs	Object
artist_name	Artist Name	Object
composer	Composer	Object
lyricist	Lyricist	Object
language	Language	Float64

Table 4: Song Data

Column Name	**Description**	**Type**
song_id	Song ID	Object
name	Song Name	Object
isrc	International Standard Recording Code	Object

Table 5: Song Extra Info

Which attributes (columns) from the database seem most promising? song id, song length, language, source system type, source screen tab and source screen name. Of course, among the promising attributes is the target variable itself.

Is there enough data to draw generalizable conclusions or make accurate predictions? Yes. The data is really numerous.

Are there too many attributes for your modeling method of choice? No

Are you merging various data sources? If so, are there areas that might pose a problem when merging? Yes. We are merging the training data with the user and song info. There are no particular problem. Some features might be excluded.

Have you considered how missing values are handled in each of your data sources? Yes.

Did you compute basic statistics for the key attributes? What insight did this provide into the business question? These are calculated in the next step.

Are you able to prioritize relevant attributes? If not, are business analysts available to provide further insight? The attributes provided are all considered relevant, any attributes to be discarded will be analysed at a later stage. In a real-life case, business analysts can be consulted for even deeper insights into the data.

8 Data Exploration

Although the provided datasets are well-organized, they're not raw data and don't require preprocessing for exploration. For this project, the training data is merged with the member and song datasets, resulting in consolidated attributes. The 'song extra info' dataset isn't integrated due to its limited additional insight, except for the ISRRC column, which hasn't been officially verified so it may contain errors.

This new merged data now contains aggregated, detailed and comprehensive information on the users, tracks and artists they listen to. Exploration is useful for uncovering patterns as well as for detecting anomalies or errors in the data.

The training data contains 7377418 records, balanced between recurring and non-recurring listening events triggered within a month after the user's first observable listening event. Figure 1 shows the balance of the target variable in the training set. There are no duplicate values of these records.

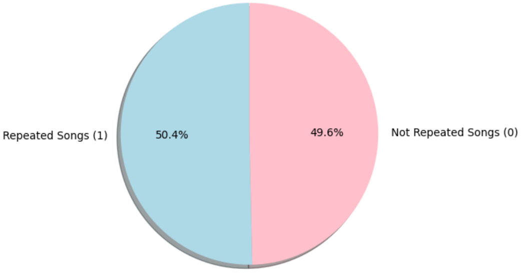


Figure 1: This pie chart illustrates the balance of the records between recurring and non-recurring listening events within a month of the user's first listening event.

Table 6 provides some statistics about the unique and null values of the data. It is worth to notice that in this dataset, there 359966 different songs a well as there are more than 40,000 different artists. Moreover, the data accounts for distinct 30755 users. While these numbers represent distinct records, certain attributes like lyricist, composer, and artist can contain multiple names separated by a delimiter in a single record. To obtain the exact count of individual lyricists, composers, and artists, it's necessary to parse these names using a regex expression. Accordingly, the distinct counts for composers, artists, and lyricists are 82740, 42740, and 38887, respectively. Also, in this dataset, some columns have an extremely high number of null values.

The correlation between features is calculated through the Person Coefficient. The results are shown in Figure 2. It can be shown that the variables are not particularly correlated with each other. The only variables that show a rather significant positive correlation are city and age (bd). In contrast, the feature 'registered via', indicating the user's registration device, shows a significant negative correlation with the variable 'registration init time' indicating the start of the subscription.

Attribute	Data Type	Unique Values	Null Values	% Null Values
lyricist	object	33888	3178798	0.430882
gender	object	2	2961479	0.401425
composer	object	76064	1675706	0.227140
source_screen_name	object	20	414804	0.056226
genre_ids	object	572	118455	0.016056
source_system_tab	object	8	24849	0.003368
source_type	object	12	21539	0.002920
language	float64	10	150	0.000020
song_length	float64	60266	114	0.000015
artist_name	object	40582	114	0.000015
bd	int64	92	0	0.000000
registration_init_time	int64	3811	0	0.000000
registered_via	int64	5	0	0.000000
msno	object	30755	0	0.000000
city	int64	21	0	0.000000
song_id	object	359966	0	0.000000
target	int64	2	0	0.000000
expiration_date	int64	1395	0	0.000000

Table 6: Overview of Unique and Null Values present in the training set

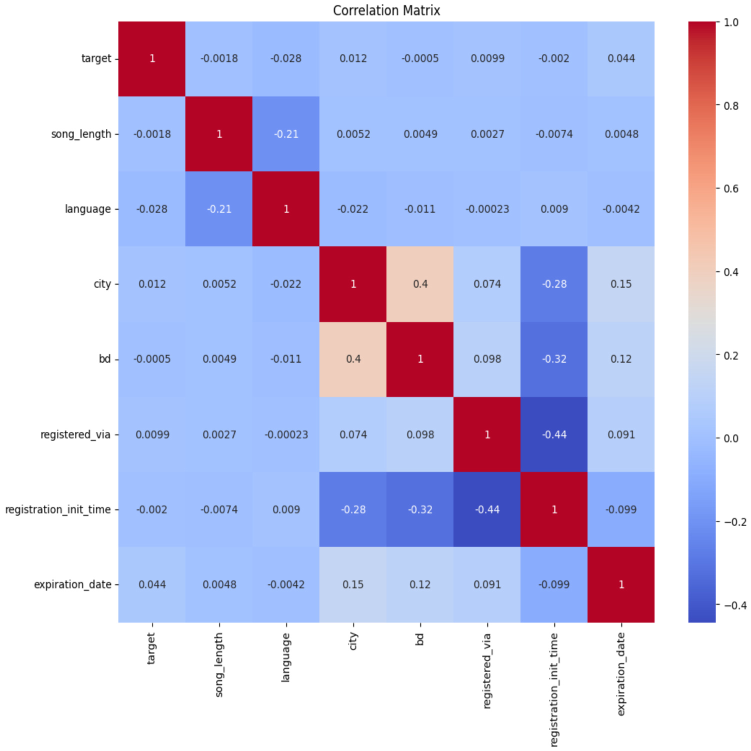


Figure 2: Pearson Correlation

After this preliminary step, an exploratory data analysis (EDA) is performed on the data. Figure 3 shows the frequency of the songs languages present the data.

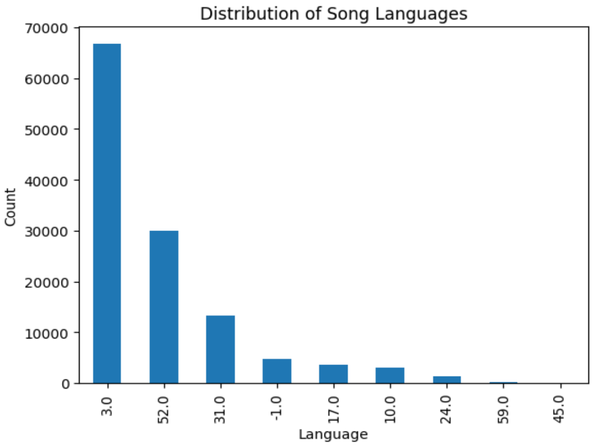


Figure 3: Song Language Frequency

Language mapping is not provided directly, but by performing an analysis on song titles and their respective languages, it can be easily derived that:

- **3.0**: Mandarin Chinese
- **52.0**: English
- **31.0**: Korean
- **-1.0**: Mixed Languages (E.g Italian, French,..)
- **17.0**: Japanese
- **59.0, 10.0, 24.0**: Chinese Variations
- **45.0**: Taiwanese
- **38.0**: Hindi

It is important to analyse the user's main sources of music consumption within the platform. Figure 4 shows the main sources from which the users plays the songs.

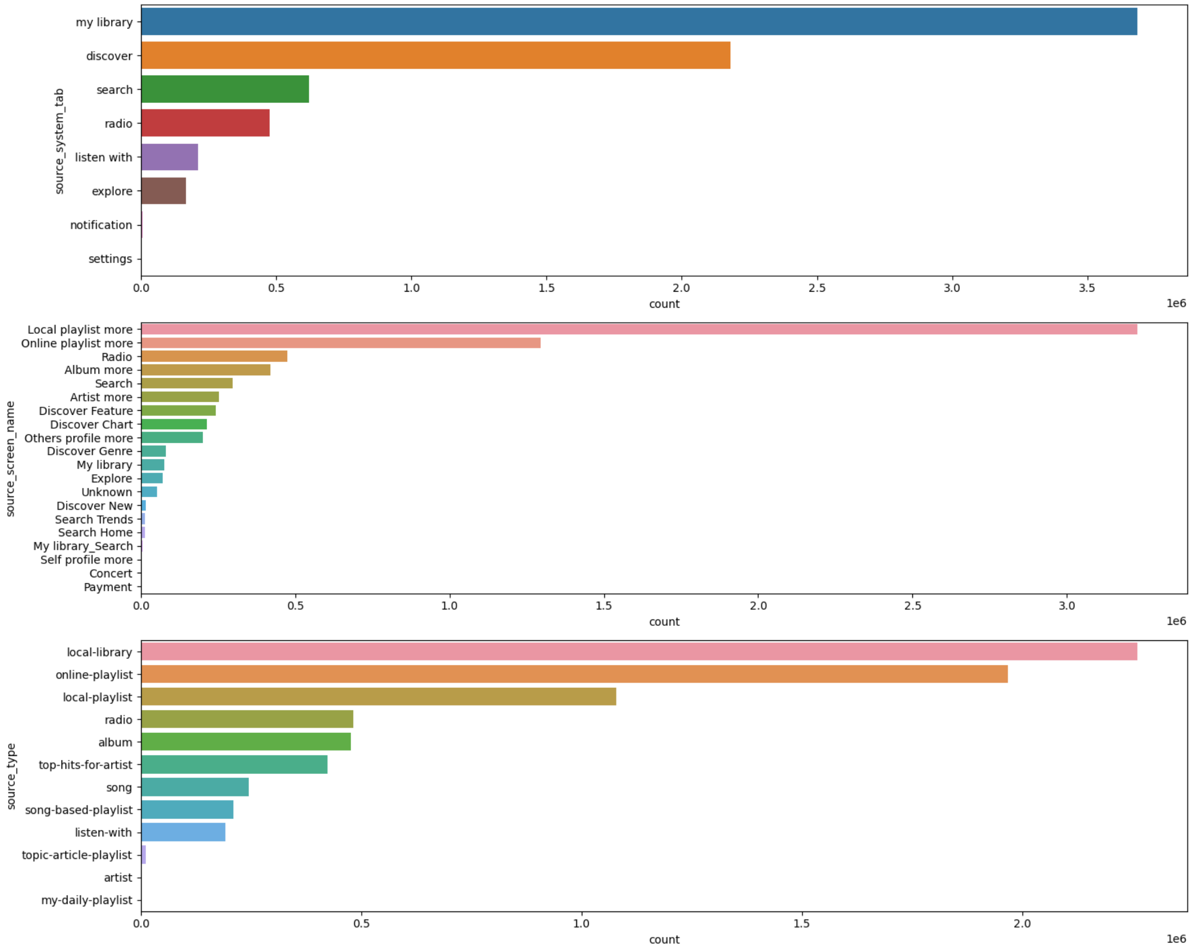


Figure 4: Sources from which the event (song play) is triggered

The songs mainly played by users are already in their library, secondly we have playback in the discovery section. This underlines the importance of music recommendation, as the discovery section, where songs are recommended to users, is much more used than the search section. The explore section where many hits are present is generally little used. It is evident that users listen to their favourite songs and want to discover them by similarity (discovery section). The source screen name represents the name of the layout a user sees, still the main resource of music consumption are the local library and the online playlist. The distribution of the source type confirms these two as the main sources from which a song is played.

Data records present events concerning whether or not the user has listened to a song again. The average length of songs, whether repeated by a user or not, is about 6 minutes and they both have a standard deviation of about 1.7 minutes. Figure 5 shows a histogram showing the distribution of song lengths (measured in milliseconds).

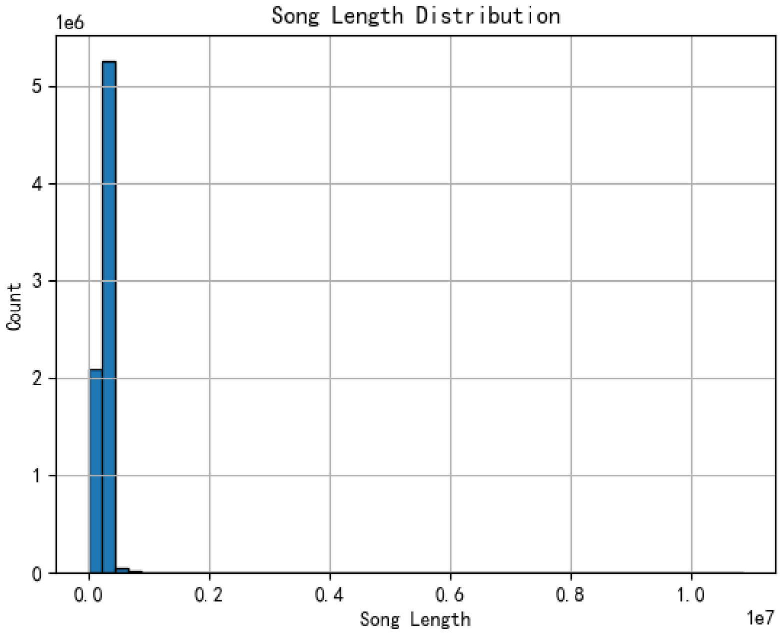


Figure 5: Distribution of the length of the songs

It is also interesting to analyse the gender distribution of users. In the dataset, males slightly outnumber females by a difference of 179,307 users, though the disparity isn't big. Figure 6 shows the distribution of the target variable between male and female users and highlights a near-equal distribution between the two genders.

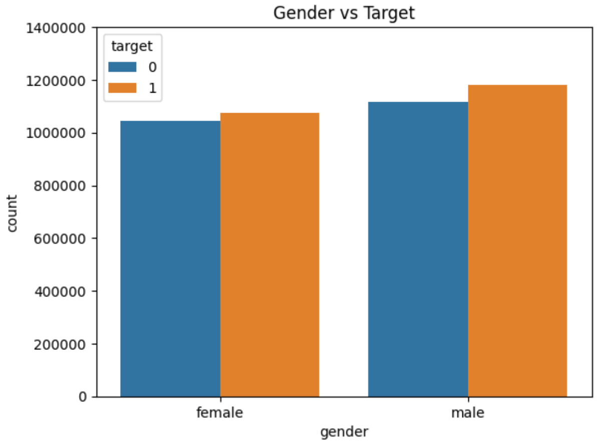


Figure 6: This figure illustrates the target variable's distribution among male and female users.

An analysis of the age of the users shows that there are numerous outliers. About 40 per cent of the age values are completely wrong, there are users who are over a thousand years old and users who are zero.

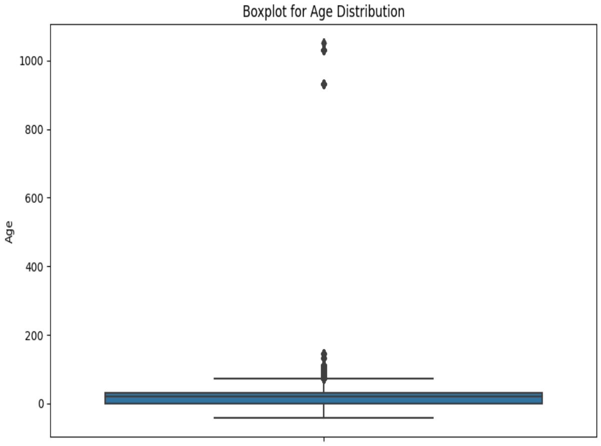


Figure 7: Age Boxplot

The listening behaviour of users in the dataset within the platform is shown in Figure 8. The peak on the x-axis means that more than 14000 users out of a total of 30755 have listened to 100 songs. Figure 9 shows the log distribution of listening events per song. It should be noted that many songs are heard very few times. Therefore, the songs that are listened to several times by users must be very popular.

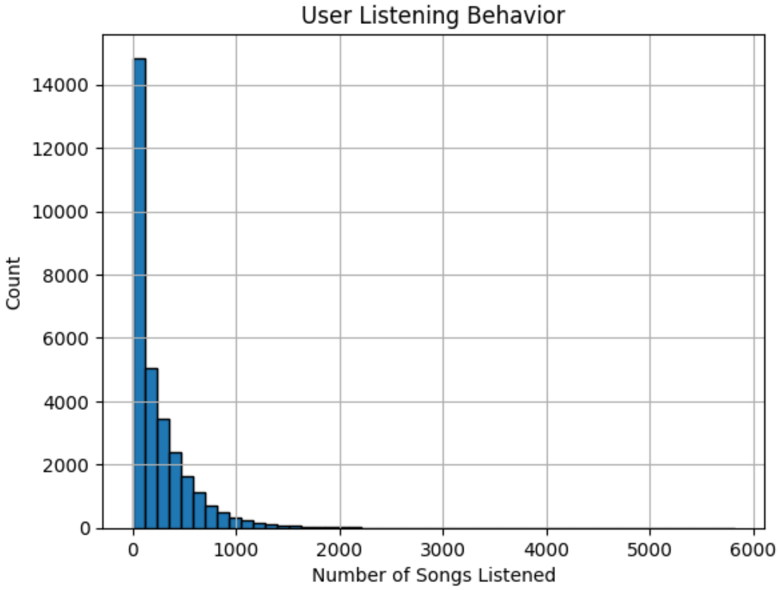


Figure 8: This histogram shows the distribution of the number of songs users have listened to.

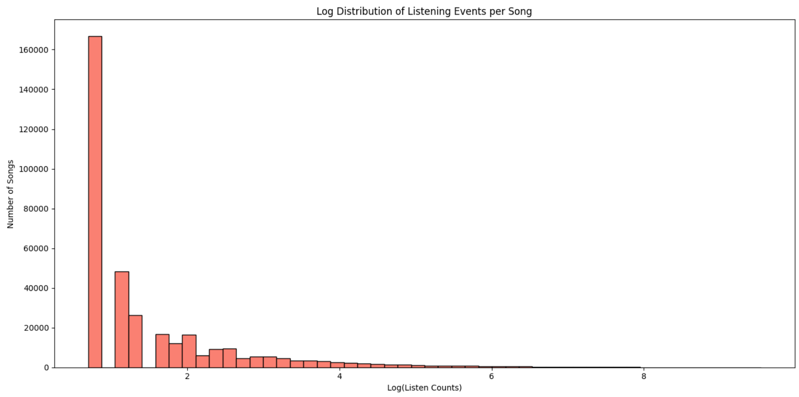


Figure 9: This histogram shows the distribution of the listening events per song.

Figure 10 shows that users mainly register on the platform via method 9 and 7, emphasising the importance of focusing on these methods. Insights on the user registration within the platform are illustrated in Figure 11. The distribution of user registration within the platform is fairly even across all months, with a higher concentration in December. Furthermore, the streaming platform registers the highest number of users between the years 2016 and 2017.

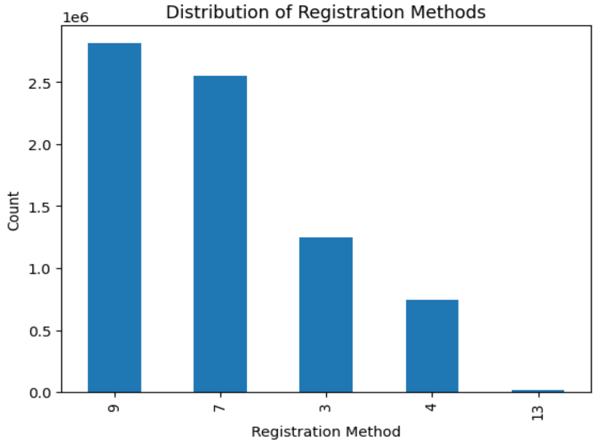
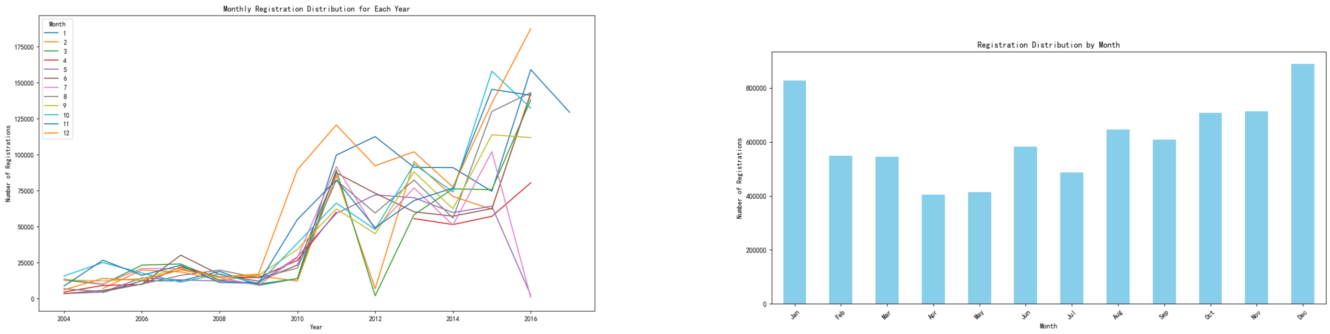


Figure 10: This histogram illustrates the main methods used by users to register.

Figure 12 illustrate an histogram of the user provenance. It shows that users of the platform are mainly concentrated in cities 1 and 13.



(a) Trend of user registrations each month across different years. It shows the monthly patterns of registration over time

(b) Monthly distribution of user registrations, showing the number of users who registered in each month.

Figure 11: Overall caption for the two subfigures

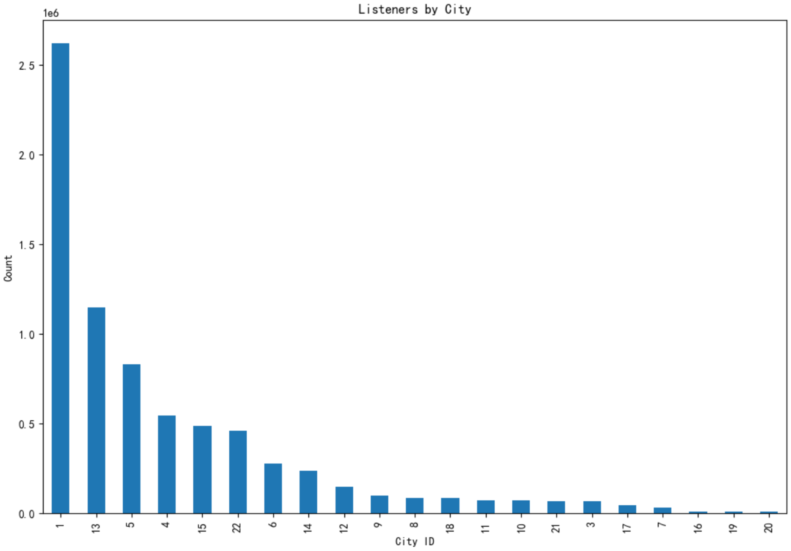


Figure 12: Distribution of user provenance within the platform.

What sort of hypotheses have you formed about the data? This data presents a detailed view of user behaviour within the platform. The songs mainly listened to by users are mainly songs in Mandarin Chinese, English and Korean. Being an Asian streaming service platform, it is understandable that there is a focus on content in these languages. Users on the platform listened to approximately between 100-200 songs and the main methods of listening are the library section and the discovery section. Surprisingly, users search little for songs and prefer to rely on algorithms that recommend songs considering those already in the library. Additionally, the main type of sources from which a user listens to a song are local-library and online-playlists. Surprisingly few users listen to songs directly from albums or by looking at the top hits. This emphasises that in order to increase user engagement, a good music recommendation algorithm is of paramount importance.

Identification of particular subsets of the data The main two subsets of the training data are members, songs and songs extra info.

Has this exploration altered the goals? No, after this exploration phase, this data was confirmed as suitable for achieving the data mining goals.

Which attributes seems promising for further analysis? The attributes that needs further analysis are:

- age (bd): Given the very high number of outliers, this attribute certainly needs further analysis.
- source system tab, source type, source screen name: provide interesting insights on the user behaviour within the platform
- genre: it may be interesting to analyse the genres most listened to
- language: With Mandarin Chinese, English, and Korean being the primary languages of songs that users listen to, analyzing patterns within these specific language categories might yield insights into users' preferences.

Further analysis on attributes containing missing values must also be carried out.

9 Data Quality Report

Have you identified missing attributes and blank fields? If so, is there meaning behind such missing values? In this dataset, there are features (see Table 6) that contain many missing values. Probably, being such a large amount are simply data that are not available in the songs present on KKBOX. The main features, such as user ID, song ID, and target variable do not contain missing values.

Are there spelling inconsistencies that may cause problems in later merges or transformations? After an analysis of the data, it was found that there are no spelling inconsistencies. Regex expressions were also used to analyze the names of singers, songwriters, and composers.

Have you explored deviations to determine whether they are "noise" or phenomena worth analyzing further? The age column contains about 40% outliers. There are users who record completely incorrect values that cannot stick to the facts such as 0 or 1,000 years. The age values considered correct in this analysis are 16 to 99 years. All other values are replaced with the median (27) years.

Have you conducted a plausibility check for values? Take notes on any apparent conflicts (such as teenagers with high income levels). Yes.

Have you considered excluding data that has no impact on your hypothesis. Depending on the hypothesis in question, certain columns might not be relevant. For instance, if the hypothesis is only related to user behavior and song preferences, columns like `registration init time` and `expiration date` might not be directly relevant. Moreover, these were not even present in the original training set but resulted from the merging of the training data with user data (members). Therefore, they are disregarded for the training of the model.

Are the data stored in flat files? If so, are the delimiters consistent among files? Does each record contain the same number of fields? All the data are stored in a tabular format, in .csv files (flat files structure) and all use the same structure and delimiters. In the datasets each record contains the same number of fields. Each row has entries for fields. All rows consistently have values (or placeholders for missing values) for each of these fields.

Chapter 3: Data Preparation

10 Selecting Data

- All rows from the training were included and no particular subset of user is excluded from the analysis.
- No filters are applied to the attributes in this study. All the attributes selected are treated equally.

11 Including or Excluding Data

Is a given attribute relevant to you data mining goals? For this analysis, we merged the members and the song datasets with the training data to have more elements available to achieve data mining objectives. However, 3 attributes were excluded: ISRC (International Standard Recording Code), Registration Time and Expiration Date. The former was excluded from the merging of the database songs because it is not a verified attribute and may contain errors. Furthermore, it is only a code that can be used as a song identifier and Multiple songs could share one ISRC since a single recording could be re-published several times. The other two attributes are excluded from the merging of the dataset member as they are not relevant to to predict the chances of a user listening to a song repetitively after the first observable listening event within a time window was triggered. However, they are still used for exploratory data analysis and, given their formatting, they require some pre-processing, which is done in the next chapter.

Does the quality of a particular data set or attribute preclude the validity of your results? No, all the data and data attributes (except for the three mentioned above) are necessary to achieve the data mining objectives, as they answer different questions and, together, provide specific insight into user behaviors within the platform.

Is it possible to save such data? Yes

Are there any constraints on using particular fields such as gender or race? There are no info regarding the ethnicity of the users but there are info regarding their gender. However, this is not an attribute that can lead to particular constraints/gender biases.

12 Cleaning Data

The training set is merged with the dataset of members and songs. Therefore, at the end of the cleaning process the training set contains contains 18 columns, resulting from the merging and the reformatted date column. The features are: msno, song id, source system tab, source screen name, source type, target, song length, genre ids, artist name, composer, lyricist, language, city, bd, gender, registered way, registration init time, expiration date.

This dataset doesn't contain duplicate values, so their handling is not needed.

Data Problem	Data Preprocessing Decision
Missing data	Several categorical columns such as lyricist, gender, composer, source screen name, genre ids, source system tab, source type, and artist name exhibited missing values. Given the substantial occurrence of these missing entries across the columns, they were labeled as "Unknown". This strategy ensures that the information gap is transparently addressed, allowing subsequent analytical stages to operate without missing data-induced disruptions. For the song length numerical column the median value was used as an imputing measure. The reason behind using the median lies in its capability to provide a middle ground without being overly influenced by potential outliers or extreme values, thus ensuring that the integrity of the dataset remains intact. For the language column, also numeric, a different approach is used. Since it is rather easy to find a language from a title, I implemented a mapping function that uses regex expression to map the language of the song based on the title. In the training set there are 150 songs with null language, but the unique values are 53 (therefore the titles are repeated). Performing a manual mapping control analysis is therefore a straightforward task.
Data Error	Considering the presence of noise within the age column, we identified that approximately 40% of the data in this category qualified as outliers (users below 16 and above 99). To create a more robust representation, these outliers were substituted with the median age value. The choice of median, given its resistance to extreme values, provides a central tendency that is more representative and less prone to influence by the skewed distribution.
Format Data	The column that records data regarding the platform user's start and end of subscription has the following formatting for years, months and days: "20120102." However, it can be very uninformative, so this column is divided into three different columns that track the registration for years, months, and days. Mind that these attributes are not considered for model training, however they can provide indications of the churn rate of users within the platform, so they are used for an EDA.

Table 7: This table presents the problem in the data and solution found

13 More Detailed EDA

An initial EDA offers preliminary understanding of the data. However, as CRISP-DM is an iterative process, revisiting EDA after handling missing values allows to further discover new patterns previously obscured by gaps and to have enhanced visualisations based on a fuller data representation.

Given that 40% of the gender column consists of missing values. Figure 13 presents an updated histogram displaying the distribution of the target variable among users. In this visualization, the label "Unknown" represents those missing gender records.

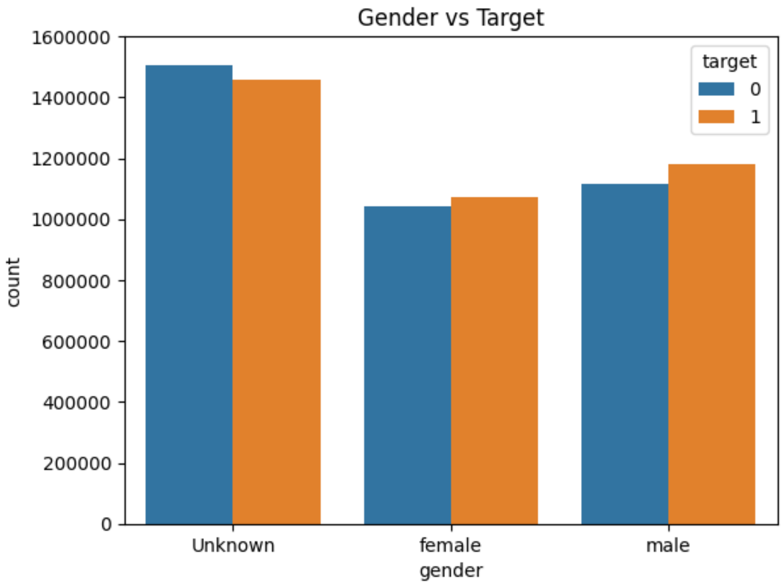


Figure 13: Updated distribution of gender

The distribution of the user age after pre processing is shown in figure 14. The platform is mainly used by users between 16 and 35 years old. Figure 15 illustrates the age distribution of users across cities.

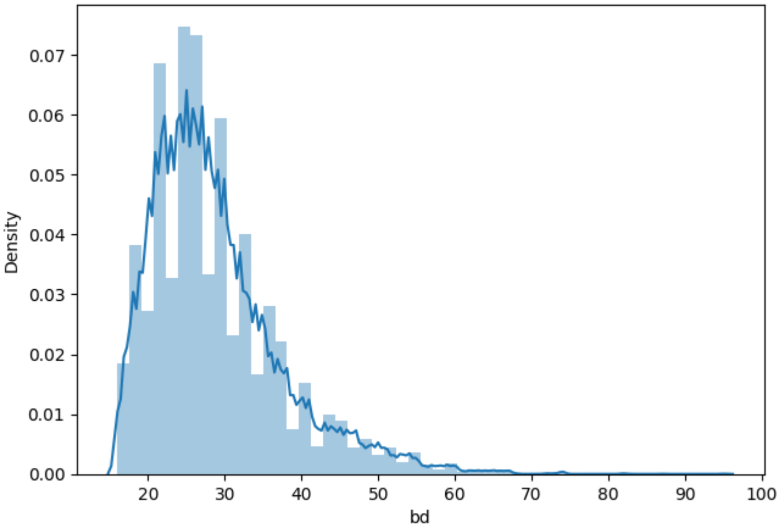


Figure 14: Distribution of user age

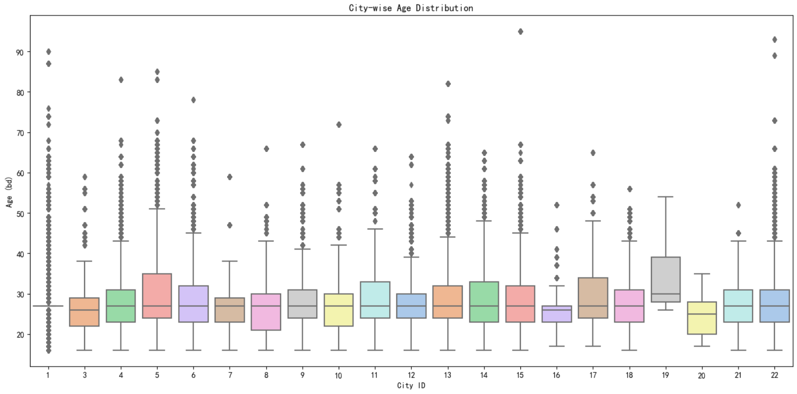


Figure 15: Distribution of the user age across the various cities

The plots in Figure 16, 17, 18 show the sources of users' music consumption in relation to the target variable. Analysing all these figures, it can be seen that the songs that are most frequently played are in the user's library. In the discovery section, on the other hand, the user plays songs but many times they are no longer listened to from there. This is probably because either they are added to the library section or the user did not like them. To conclude, the main sources of music consumption for a user are the local library, where the user tends more frequently to play a song than not to play it, the online-playlist and the discovery section.

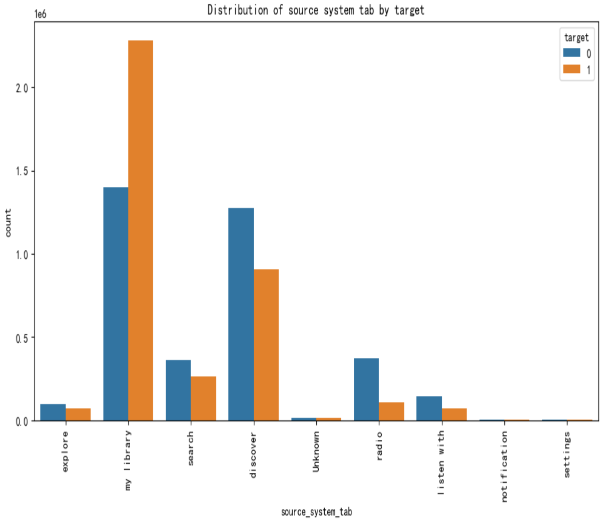


Figure 16: Distribution of the source system tab which represents the name of the tab where the event was triggered with respect to the target variable.

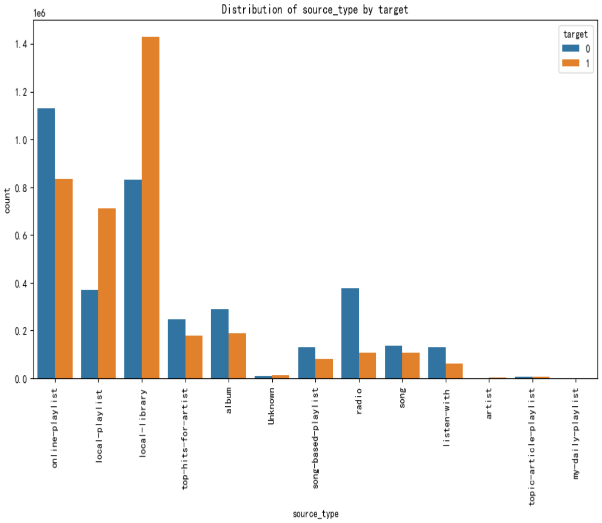


Figure 17: Distribution of the the entry point a user first plays music on mobile apps with respect to the target variable.

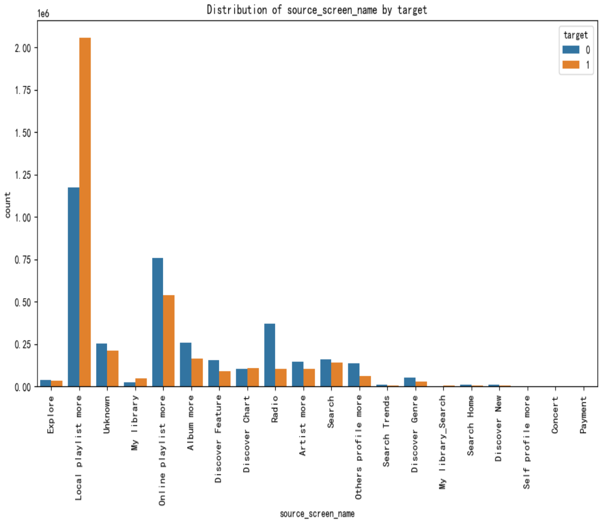


Figure 18: Distribution of the source screen name, specifically the name of the layout a user sees, with respect to the target variable.

Table 8: This table represents the top 10 artists how much their songs were listened to within the platform.

Artist	Listen
Various Artists	303616
Jay Chou	186776
Mayday	182088
JJ Lin	115325
Hebe	104946
aMEI	82799
Eason Chan	76035
Nine One One	70445
G.E.M.	67296
BIGBANG	61927

The 10 most listened to artists on the platform are shown in Table 8. Whilist, Figure 19 shows the 10 most popular genres of music in relation to the target variable. The genre that is clearly more listened to than others is 465. Moreover, it is also the most listened in all the cities, as shown in Figure 20.

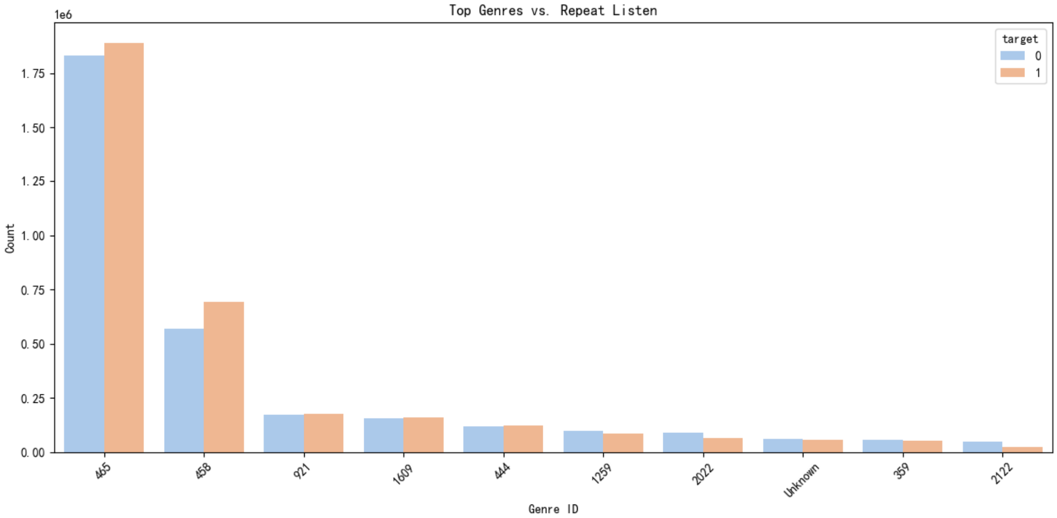


Figure 19: The 10 most listened genres of music in relation with the target variable

Figure 21 visualizes the total listens for artists associated with the top 10 most frequently played songs on the KKBOX platform. Each bar represents an artist, with the height indicating the cumulative number of times their songs from the top 10 list were played by users. It provides a quick overview of which artists dominate user listening behavior within the most popular tracks.

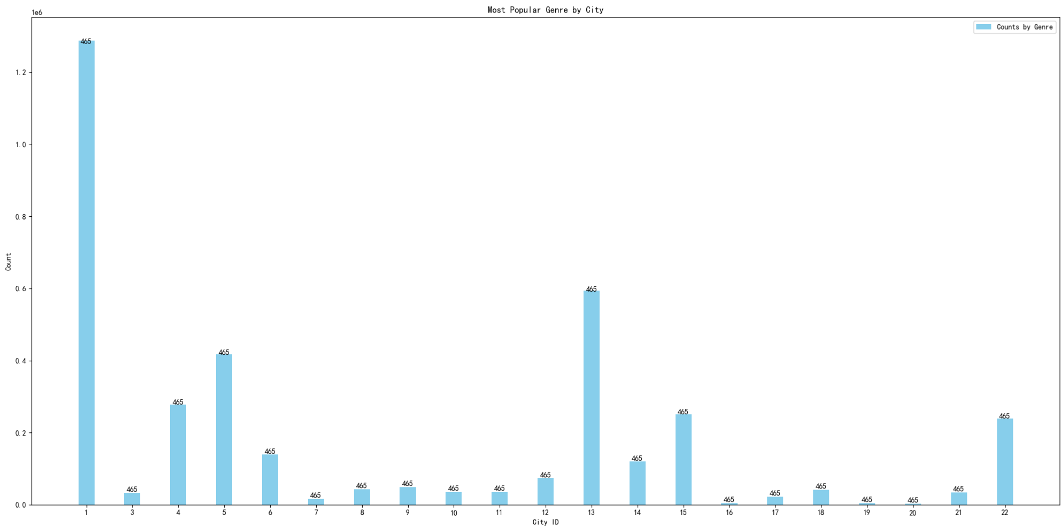


Figure 20: Most popular genre and its listening frequence grouped by city

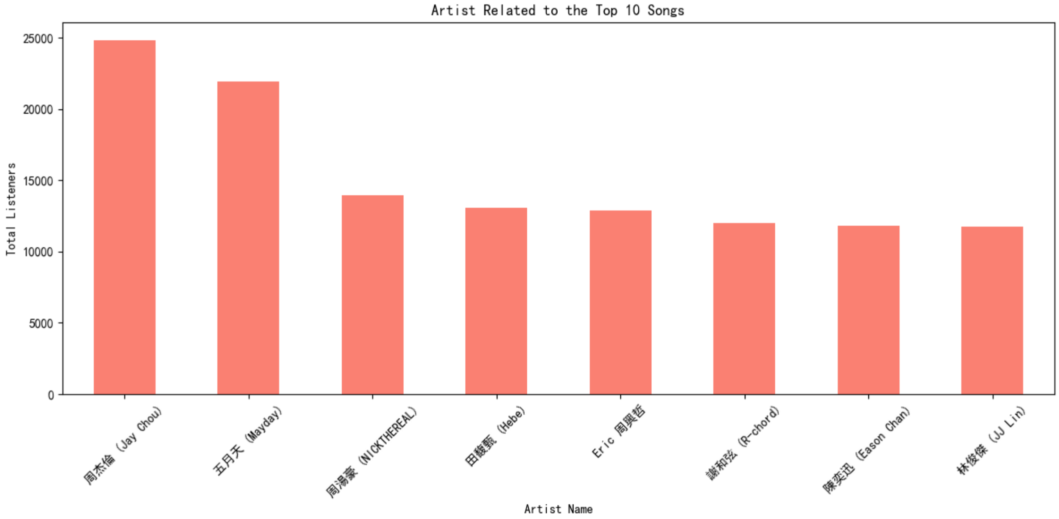


Figure 21: Total Listens of Artists Associated with the Top 10 Most Played Songs on KKBOX

Chapter 4: Modeling

The modeling techniques employed by this KKBox project are driven by the data mining goals:

14 Select the right modelling technique

Improved music recommendations: It is developed a predictive model which involves the use of historical events related to user-song interactions to predict the probability of a user listening to a song. Feed-forward Neural Networks are particularly suitable for this task as they are able to capture non-linear relationship between features, given their ability to integrate a mix of continuos and categorical variables. XGBoost, like the Feed Forward Netowrks, is able to capture non-linearity patterns in data. Additionally, this model is also able to naturally handle categorical features and provide insights into influential factors. Moreover, they both are able to handle large-scale dataset.

Does the model require the data to be split into test and training sets? Yes. Given the fact that this is a supervised learning approach, it's important to split the data into training, validation and testing sets. The training set allows the model to learn patterns, the validation set helps to adjust hyperparameters and monitor the model during training to avoid overfitting, and the testing set is used to evaluate the model's performance on unseen data to ensure that it is generalising well.

Do you have enough data to produce reliable results for a given model? With 7377418 records, we have a sufficient number of records for training deep and machine learning models. However, the reliability of results does not only depends on the volume of data but also by its quality and relevance to the problem.

Does the model require a certain level of data quality? Can you meet this level with the current data? In order to produce meaningful results, the quality of the data has to be good. In this case, the data presents good quality, the data are numerous and balanced, meaning that the target variable is fairly spread between 0 and 1. However, some pre-processing had to be carried out in order to be given as input to the models and meet the desired level of quality. I mainly handled excessive noise in the column age and missing data in both numerical and categorical features.

Are your data the proper type for a particular model? If not, can you make the necessary conversions using data manipulation nodes? The data were not initially formatted suitably for certain models, but all necessary conversions were made to prepare it as input. For instance, for the feed-forward network, we encoded the categorical features using a Label Encoder and standardized the numerical features (excluding the target variable) using a Standard Scaler. For XGBoost, only the encoding of categorical features was necessary, as its main components are decision trees that are insensitive to feature scaling. It is also important, mainly for FFNN, to handle missing values as XGBoost is inherently able to manage them.

14.1 Modelling Assumptions

Null data, both numerical and categorical, were processed at the previous step. Further pre-processing is needed in this step to make them fit the models. The categorical data is encoded using the LabelEncoder function, which convert scategories into integers, assigning unique numbers to distinct categories in a column, making them interpretable for the algorithms. Scaling of the numerical data is performed with StandardScaler which standardize features by removing the mean and scaling to unit variance, making it centered around zero. For the FFNN neural network, both steps are performed as they require numeric features to be scaled in order that they all can evenly contribute and that gradient-based optimization methods converge faster. Whilist, for XGBoost, only the encoding of the categorical data is performed as it is naturally robust to differences in feature scales due to its tree-based structure, but it still necessitates the conversion of categorical variables into a suitable numeric format for optimal performance.

15 Generating the test design

What data will be used to test the model? Have you partitioned the data into train/test sets? I have partitioned the dataset into train, validation and test, using a 80-20 split.

How might you measure the success of the supervised models? For this task, a combination of the accuracy metric and the AUC-ROC curve would be quite insightful to evaluate the goodness of the model. The former represents the percentage of the total correct predictions, whereas the latter is a graphical representation of the performance of the binary classifier. An AUC-ROC value of 1 indicates perfect classificatio, while 0.5 indicates that your model is like random guessing.

How many times are you willing to rerun a model with adjusted settings before attempting another type of model? This depends on the task complexity, time and budget constraints. After 3-5 unsuccessful repetitions with modified hyperparameters, one might consider switching to a different architecture or model.

16 Building Model

16.1 Parameter Settings

The algorithms used for this task is a feed forward neural network and XGBoost. Every algorithm has parameters that can influence its performance.

The parameters settings for the feed forward neural network are:

- **Data Encoding and Pre-processing:**
  - **Categorical Data:** Encoded using the `LabelEncoder` from the `sklearn` library. This converts categorical columns to numeric representations.
  - **Numerical Data:** Standardized using `StandardScaler` from the `sklearn` library. This scales the numerical columns to have a mean of 0 and standard deviation of 1.
- **Embedding Layer:**
  - For each categorical column, the number of unique categories determines the vocabulary size.

  - The embedding size is determined as the minimum of half the number of unique categories and 50.
  - Each categorical column has its own embedding layer which gets flattened for further processing.

- **Dense Layers:**
  - **First Dense Layer:** 256 neurons with ReLU activation, specifically for numerical data.
  - **Second Dense Layer:** 256 neurons with ReLU activation.
  - **Third Dense Layer:** 128 neurons with ReLU activation.
  - **Fourth Dense Layer:** 128 neurons with ReLU activation.
  - **Fifth Dense Layer:** 64 neurons with ReLU activation.
  - **Final Output Layer:** 1 neuron with sigmoid activation for binary classification.

- **Dropout and Batch Normalization:**
  - Dropout Layers (for regularization): Two dropout layers with dropout rates of 0.2 to prevent overfitting.
  - Batch Normalization: Used after concatenation of embeddings and after some of the dense layers to stabilize and accelerate the training process.

- **Optimization and Loss:**
  - **Optimizer:** Adam optimizer with an initial learning rate of 0.0005.
  - **Loss Function:** Binary Crossentropy, suitable for binary classification tasks.

- **Batch Generator:**
  - Shuffles the dataset and divides it into batches to feed into the network. This ensures that in every epoch, the model encounters data in a different sequence.

- **Batch Size:** 128

- **Training Configurations:**
  - **Epochs:** Maximum of 15 epochs, but with early stopping in place, the model may stop training early if the validation loss doesn't improve for 3 consecutive epochs.
  - **Callbacks:**
    - `EarlyStopping`: Stops training if the validation loss doesn't improve for 3 consecutive epochs and restores the best weights.
    - `ReduceLROnPlateau`: Reduces the learning rate by a factor of 0.2 if the validation loss doesn't improve for 2 consecutive epochs. The minimum learning rate is set to 0.0001.

The parameters settings for the XGBoost are:

- **Data Encoding and Pre-processing:**
  - **Categorical Data:** Encoded using the `LabelEncoder` from the `sklearn` library. This method converts categorical columns to numeric representations.
  - Note: Unlike the Feed Forward Neural Network, there was no scaling for the numerical columns in the XGBoost model since tree-based models are not sensitive to feature scale.

- **Model Parameters:**
  - **Learning Rate (lr):** Set to 0.2, this parameter determines the contribution of each tree to the final prediction.
  - **Max Depth:** Set to 15, it specifies the maximum depth of a tree.
  - **Min Child Weight:** Set to 5, it defines the minimum sum of weights needed in a child node. It is used to control over-fitting; higher values prevent more partitioning.
  - **N_estimators:** Set to 250, this is the maximum number of gradient boosted trees to be trained (i.e., the number of boosting rounds).
  - **Objective:** Set to 'binary:logistic', indicating that the task is binary classification and the logistic regression is to be used for prediction.

- **Training Configurations:**
  - **Evaluation Metric:** Logarithmic loss (logloss) is used as the evaluation metric.
  - **Early Stopping:** The training stops if the performance on the validation set doesn't improve for 10 consecutive rounds.
  - The model is verbose, meaning it prints out progress during the training.

16.2 Models

Feed Forward Neural Network: Implemented to capture complex non-linear patterns from the KKBox dataset. It uses multiple layers and neurons to achieve high-dimensional data representation.

XGBoost: Ensemble machine learning method which uses gradient boosted trees. Particularly suited for structured data like the KKBox dataset.

16.3 Model Descriptions

Can you draw meaningful conclusions from this model? Yes, these models accurately predict the chances of a user listening to a song repetitively after the very first observable listening event.

Are there new insights or unusual patterns revealed by the model? No, the models didn't revealed any unusual pattern.

Where there executions problems for the model? How reasonable was the processing time? Considering the large amount of data, the processing time is quite reasonable, each model took approximately two hours to train.

Did the model have difficulties with data quality issues, such as a high number of missing values? No. The issues on the data were handled in the previous steps.

Were there any calculation inconsistencies that should be noted? No

16.4 Model Assessment

Improved Music Recommendations: The FFNN and XGBoost models both provided reasonable results, making it difficult to choose one over the other. By delving into their applications, we hope to exploit the strengths of both models, embracing the recommendations on which they agree and examining the areas in which they differ. With a little effort and applied business knowledge, further rules can be developed in order to resolve the differences between the two techniques. The propensity of the models to discern intricate patterns from listener behaviour indicates a promising path for the creation of real-time personalised music recommendations. These predictive insights could improve the KKBox user experience by aligning song suggestions to user inclinations in real time.

16.5 Adjust the parameters of existing models.

The FFNN model initially contained only a first dense layer of 128 neurons and three intermediate dense layers of 256, 128 and 64 respectively. Furthermore, there was no regularisation technique. At the beginning the performance was not that promising. Perhaps, due to the large amount of data, the model was too simplistic. Also, a batch size of 64 required too many computational resources and 256 deteriorated the results somewhat, 128 turned out to be a good trade off. The structure of the network was then gradually refined by adding a few layers and regularisation techniques to arrive at the final structure explained in 12.1. XGBoost, on the other hand, had fewer parameters for the setting. Initially a learning rate of 0.01 and a number of trees of 100 with a max depth of 5 were set. However, these parameters did not prove ideal, the learning rate was raised (0.01, 0.05, 0.1, 0.2, 0.3) along with the number of trees and the max depth (5, 10, 15, 20). A very good trade off between result and time taken for training was obtained in 12.1.

Are you able to understand the results of the models? Yes, I am able to understand the results of the models in terms of the performance metrics, data inputs, and the outcomes they generate. The data is clear, and the results align with expectations and the data mining objectives.

However, it's essential to note that while the results themselves are transparent, the underlying mechanisms of machine learning and deep learning models, FFNN and XGBoost in this case, often operate as 'black boxes.' This means that while we can interpret the outputs and evaluate their accuracy and relevance, the inner decision-making processes of these models can be challenging to decipher in detail.

Do the model results make sense to you from a purely logical perspective? Are there apparent inconsistencies that need further exploration? The model results do align with the expectations and make logical sense, especially considering the data we provided and the objective.

From your initial glance, do the results seem to address your organization's business question? Yes, at an initial glance, the results do appear to address our organization's business question. The primary objective was to improve music recommendations for users, and the achieved accuracy and AUC scores suggest that our models can predict user listening behaviors with reasonable confidence. The higher AUC score, particularly from the FFNN, indicates a good separation between positive and negative classes, thus potentially leading to more relevant music recommendations for users. However, it's crucial to remember that while these metrics are promising, the real test lies in implementing these recommendations and monitoring user engagement and satisfaction in a real-world scenario.

Have you used analysis nodes and lift or gains charts to compare and evaluate model accuracy? While I have focused on accuracy and AUC as our primary evaluation metrics, I have not delved into using analysis nodes or lift/gains charts in the current evaluation.

Have you explored more than one type of model and compared the results? Yes

Are the results of your model deployable? To be determined in a real case. Some of the many factors that can affect the decision are: the accuracy obtained, the models already on the market, the nature of the black-box nature of the algorithms, the computational complexity of the models and the acceptance of the stakeholders.

Chapter 5: Evaluation

The evaluation of the results should be made using the business success criteria established in the initial phases of the projects. This is the key to ensuring that your organization can make use of the results you've obtained.

17 Evaluating the Results

Are your results stated clearly and in a form that can be easily presented? The results are calculated clearly. It is achieved an accuracy of 73% in the FFNN (best model) and an accuracy of 0.72% in the XGBoost. In the former, we have also an uplift in the AUC curve. The AUC curve, which will be presented in the next section, distinctly exhibits the models' ability to rank positive instances higher than negative ones.

Can you rank the models and findings in order of their applicability to the business goals? The Feed Forward Neural Network (FFNN) is the best model. Here's the evaluation of its results using the business success criteria:

- **Loyalty Enhancement:** The FFNN's higher AUC of 0.81 suggests it's better at ranking true positive instances, which can directly lead to better recommendations. Better recommendations can increase user satisfaction, thus potentially increasing user loyalty.
- **Boosted Engagement:** Thanks to its ability to make good predictions, the FFNN model can lead to more engaging user sessions, suggesting its applicability to achieve this goal.
- **Augmented Song Consumption:** The performance of the FFNN can also lead to users playing more songs that resonate with their preferences, possibly increasing the number of songs played.

The second best model is XGBoost and according to the business success criteria:

- **Loyalty Enhancement**: The performance of the XGBoost are good. However, it has a slightly lower AUC with respect the best model, implying room for improvement in terms of loyalty-enhancing recommendations.
- **Boosted Engagement**: XGBoost reveals which features most influence engagement, the FFNN has demonstrated slightly superior predictive accuracy for our dataset in forecasting user behavior. Given its inherent feature importance capabilities, XGBoost can give insights into what drives engagement.
- **Augmented Song Consumption**: The XGBoost's ability to make accurate recommendations is good. Yet, when compared to FFNN, it might be a little less effective in increasing song consumption.

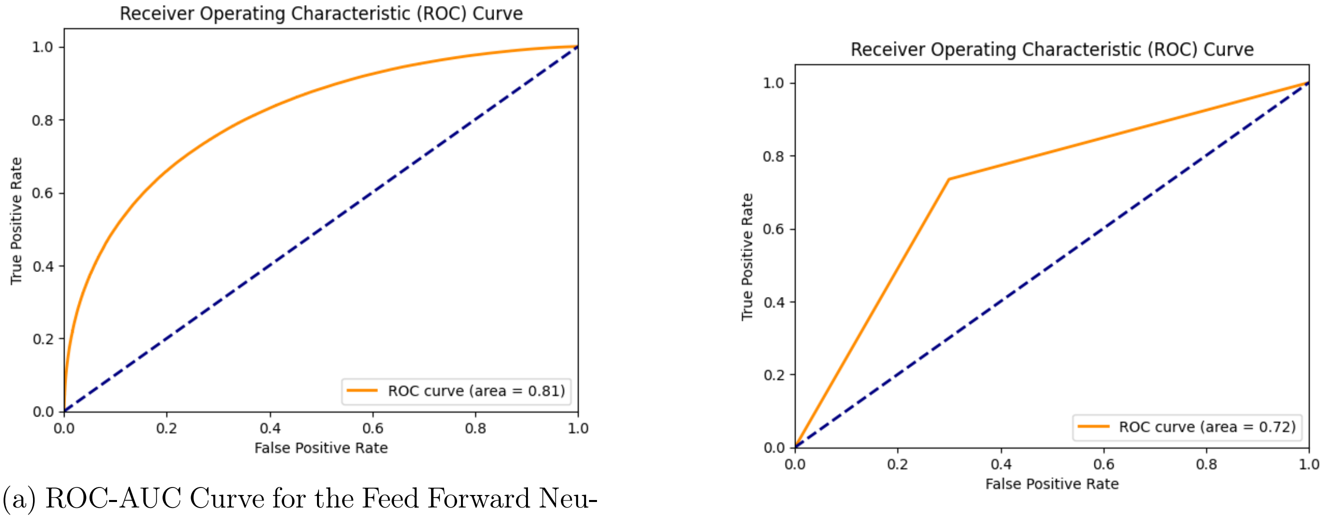
In general, how well do these results answer your organization's business goals? To be determined in a real case. The results seem to be in line with our objectives, satisfying and engaging users but in real-world scenarios one should always interface with the stakeholders to see if the results obtained their meet expectations.

What additional questions have your results raised? How might you phrase these questions in business terms? The following questions may emerge:

- How can the model performance be further improved?
- Are there specific user segments that don't resonate well with the current recommendations?
- How quickly do user music preferences evolve over time and how frequently should the recommendation system adapt to these changes to remain relevant?

18 Model Performance Visualization

The AUC curve provides a clear representation of our model's performance. An AUC closer to 1 indicates a superior model, and from the curve, we can deduce the FFNN model's superiority in terms of distinguishing between positive and negative instances.



(a) ROC-AUC Curve for the Feed Forward Neural Network

(b) ROC-AUC Curve for XGBoost

Figure 22: Overall caption for the two subfigures

19 Review Process

In this phase, I summarize the activities and decisions for each phase, including data preparation steps, model building, etc. A list of questions it is considered to facilitate this step.

19.1 Data Exploration

Did this stage contribute to the value of the final results? Yes, it provided really meaningful insights on the features of the data.

Are there ways to streamline or improve this particular stage or operation? Not in particular.

What were the failures or mistakes of this phase? How can they be avoided next time? Not in particular.

Are there alternative decisions or strategies that might have been used in a given phase? Note such alternatives for future data mining projects. Not in particular.

Are your results stated clearly and in a form that can be easily presented? Yes, the results from the analysis are presented in plots and commented coincisely.

19.2 Data Preparation

Did this stage contribute to the value of the final results? Yes, in this phase we mainly handled missing values, which were a lot, and errors in data (outliers). This phase was crucial for the next steps

Are there ways to streamline or improve this particular stage or operation? Not in particular, this phase resulted to be quite straightforward also thanks the already organised data.

What were the failures or mistakes of this phase? How can they be avoided next time? No particular error was found.

Were there any surprises (both good and bad) during this phase? There were a surprisingly large number of missing values.

Are there alternative decisions or strategies that might have been used in a given phase? Note such alternatives for future data mining projects. An alternative to the handling of missing, incorrect numeric values and the handling of missing values in categorical columns can probably be found. Furthermore, in this task it was decided to retain all features. However, there are techniques to perform feature selection such as chi-squared, fisher score and others.

Are your results stated clearly and in a form that can be easily presented? Yes.

19.3 Modelling

Did this stage contribute to the value of the final results? Yes. The choice of the model and their implementation is crucial for the success of this projects.

Are there ways to streamline or improve this particular stage or operation? Since I manually adjusted hyperparameters, in a real-world scenario do some **automated hyperparameter tuning** such as Bayesian optimization or GridSearch could help streamline the process. Moreover, given the numerous amount of data, do using **parallel Processing** or **distributed computing platforms** like Apache Spark can significantly reduce the model training time.

What were the failures or mistakes of this phase? How can they be avoided next time? Not in particular.

Were there dead ends, such as particular models that proved fruitless? Are there ways to predict such dead ends so that efforts can be directed more productively? No.

Are there alternative decisions or strategies that might have been used in a given phase? Note such alternatives for future data mining projects. Yes, it would be possible to opt for **model alternatives**, LightGBM for example, as well as different **label encoding** such as one hot encoding and different **embedding strategies**.

Are your results stated clearly and in a form that can be easily presented? Yes, However, one must always consider the black-box nature of deep and machine learning projects.

Chapter 6: Deployment

The Deployment stage in the CRISP-DM framework involves the transition of the models implemented during the modeling step into real-world applications. This final phase ensures the integration of these solutions into a production environment for tangible business benefits.

20 Planning for Deployment

It has to be determined in a real case. However, in real-world scenarios it should one should consider that high-performance servers are essential, especially for deep learning models. Moreover, with the continuous influx of user data and music metadata, a scalable database system (like SQL or NoSQL databases) is essential. Since user personal data are handled, a secure infrastructure also extremely important in these kind of systems. Finally, in order integrate recommendation models into the existing platform, one might need middleware solutions that can help the application interact with the models seamlessly.

21 Deployment Planning

To be determined in a real case. However, it would involve the integration of the models into the KKBox application, requiring dedicated serves with GPU support for efficient processing. A scalable database would be established to manage ever-growing user and music metadata. tarting with a month-long testing phase in a controlled environment, progressively refine the model after testing, followed by the deployment into the live platform. The continous collaboration between team developers, stakeholders, data scientists, software developers and user experience teams would ensure a seamless introduction of this enhanced recommendation feature to KBox's users.

22 Planning Monitoring and Maintenance

For each model or finding, which factors or influences (such as market value or seasonal variation) need to be tracked? To be determined in a real case.

How can the validity and accuracy of each model be measured and monitored? To be determined in a real case. However, possible solutions could be use offline evaluation which involves holding out a portion of the data, using the model to make predictions on this set, and then comparing the predictions to the actual outcomes. Metrics such as Mean Average Precision can be user. Moreover, online A/B testing can offer real-world performance insights. It segments users into two groups: one experiencing recommendations from the old system and the other from the new model (FFNN or XGBoost). By comparing their listening behaviour within the platform, it is possible to evaluate the real-world performance of each model. User feedback loops also ensure the models remain aligned with evolving user preferences.

How will you determine when a model has "expired"? Give specifics on accuracy thresholds or expected changes in data, etc. To be determined in a real case. Consider that if there is a significant change in user behaviour patterns or music trends, it may be a signal to re-evaluate the effectiveness of the model and potentially update or replace it.

What will occur when a model expires? Can you simply rebuild the model with newer data or make slight adjustments? Or will changes be pervasive enough as to require a new data mining project? To be determined in a real case. When the KBox music recommendation model expires, it can initially be updated with more recent data to capture user preferences. However, if music consumption patterns have changed dramatically or if new features become essential for prediction, slight adjustments or feature engineering may be required. In the case of profound changes in the music landscape or platform functionality, a new data mining project might be justified.

Can this model be used for similar business issues once it has expired? This is where good documentation becomes critical for assessing the business purpose for each data mining project. Shifting the focus of the problem, these models can also be used to predict user churn rates. In any case, the adaptability of the model is based on comprehensive documentation that provides clarity on the capabilities and limitations of the model and the specific business context for which it was initially designed.

Chapter 7: Final Report

The KBox music recommendation project provided a comprehensive insight into the musical preferences of listeners on the platform. This project revealed the necessity of each step of the CRISP-DM methodology. Starting from the beginning, both a graphical (EDA) and statistical analysis of the available data are carried out to spot the both data's strengths and weaknesses as well as a deeper understanding of the problem. One of the main issues encountered was in the pre-processing step, especially with missing or inconsistent data entries. It became clear that domain-specific expertise could have provided more intuitive strategies for tackling these issues in the data. The selection of the models was also challenging. The integration of both FFNN and XGBoost showcased promising result. The FFNN deemed a more intricate fine-tuning and hyperparameter tuning. Similarly, the hyperparameter optimization for XGBoost added layers of complexity to the implementation.

After the data mining results have been deployed, a next step could be, if possible, to get feedbacks from customers or business partners. The goal should be to establish whether the project was worthwhile.

The results of these interviews can be further summarized along with your own impressions of the project in another final report. This step has to be determined in a real case. Yet, these efforts are all important to ensure the model's robustness and adaptability to the dynamic world of music preferences

Chapter 8: * Bibliography

[1] Kkbox music recommendation challenge. URL https://www.kaggle.com/c/kkbox-music-recommendation-challenge.

[2] URL https://www.ibm.com/docs/it/spss-modeler/saas?topic=understanding-business-overview.

[3] URL https://www.businessofapps.com/data/music-streaming-market/.

[4] URL https://midiaresearch.com/blog/music-subscriber-market-shares-2022.

[5] . URL https://online.suu.edu/degrees/business/master-music-technology/tech-impact-music-industry/#:~:text=Early%20digital%20recording%20hardware%20and,the%20music%20industry%20at%20first.

[6] URL https://www.nvidia.com/en-us/glossary/data-science/recommendation-system/.

[7] . URL https://www.eliftech.com/insights/all-you-need-to-know-about-a-music-recommendati

[8] URL https://ikala.cloud/en/cases/kkbox/#:~:text=KKBOX%20holds%2060%25%20market%20share,has%20exceeded%2040%20million%20tracks.