National College of Ireland

An Intelligent User-specific Music Recommendation Engine

- using Machine Learning

MSc Research Project
Data Analytics

Akshay Mungekar
Student ID: x18103952

School of Computing
National College of Ireland

Supervisor: Professor Noel Cosgrave

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Akshay Mungekar
Student ID:	x18103952
Programme:	Data Analytics
Year:	2019
Module:	MSc Research Project
Supervisor:	Professor Noel Cosgrave
Submission Due Date:	12/08/2019
Project Title:	An Intelligent User-specific Music Recommendation Engine
Word Count:	7561
Page Count:	28

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	5th December 2019

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

An Intelligent User-specific Music Recommendation Engine

Akshay Mungekar
x18103952

Abstract

The world has turned out to be more of digital as per the recent statistics. A huge amount of digital data generated from various multimedia sources is consumed and processed daily. In the online music streaming applications, users expect the music to be recommended as per their interests which is not the case with most of the current music applications. Thus, an intellectual music recommendation system (RS) focused towards generating user-specific music based on their usage pattern has been proposed. This system takes the users behaviour logs into consideration to recommend the music of their interest. Also, most importantly, it only keeps the legitimate music tracks to make the RS more trustworthy. Thus, a number of unique Machine Learning (ML) based approaches have been implemented to develop an effective RS. The aim of this research is to implement a unique and highly efficient user-based music recommendation system using various ML algorithms chosen from previous studies. The methods used to implement the proposed RS is based on the effective results obtained from the previous researches. Thus, integrating such approaches with the existing ML models like Gradient-boosting decision trees, SVM, Nave Bayes and Random Forest are expected to obtain highly accurate performance in terms of recommendations.

Index terms— Gradient boosting, Grid-search, Random-search, EDA, SVM, Legitimate, KKBOX, Parameter Tuning, Music, Usage pattern

1 Introduction

Recommendation systems (RS), which comes under the domain cognitive systems, are a widely adopted approach in the field of predicting preferences for applications like books, movies, blogs, music and many more such. The amount and frequency of new music and albums being released has increased rapidly in the recent times. Therefore, it seems practically out of scope for a user to listen to each and every music track due to the wide variety of music choices from all corners of the world. Adding to it, the rate at which new music albums are released is rising day by day with the new artists debuting frequently. Due to this, users are unable to keep a track of all the music that belongs to their favourite genres and artists. Thus, an efficacious recommendation system has been proposed to address the stated user-related concerns. This system takes the users listening behaviour and their interactions with the application into consideration and then recommends a list of most relevant music to the intended user. The proposed system has been built using a number of machine learning (ML) algorithms to satisfy the requirements Portugal et al. (2018). The main aim of using this system is to suggest music based on individuals interest. A number of supervised and deep learning techniques Patel and Wadhvani (2018) have been utilized to fulfil the challenges and to overcome the issues faced. For more significant recommendations, some of the improved ML techniques that yields better results have been used as compared to the existing conventional methods which are based on the collaborative

and content-based filtering approach. More often, such approach leads to inconsistent results and inaccurate recommendations. To meet the individual demands of each user, their listening pattern as well as their interactions are considered to generate effective recommendations with the help of this approach. It also tackles the sparse historical data related concerns that leads to inappropriate recommendations Wu (2019). Apart from that, it also captures the hidden interactions of the user with the application interface. Thus, the proposed methods have been performed on the KKBOX contributed data downloaded from the kaggle website. All of the proposed models are successfully implemented and also have been compared against each other along with the conventional approaches in terms of relevant evaluation metrics as per the business requirementsGluhih et al. (2016). Hence, after evaluating, GBDT has been found to be the best performing model with minimal error rate which outcomes effective music recommendations.

1.2 Motivation

The journey of applications focused on music has been for a quite long time now. With lots of advancements and updations, such applications have seen a lot many improvements. At the start, a user-specific online music player was not less than a dream as there were hardly any real-time music gadgets available. After many years of modifications and developments in the field of information technology, new user-friendly music streaming applications with various searchable options were developed. Later, a concept of recommendation was brought into picture with the arrival of machine learning (ML) technology. This technology made it possible to provide in-general recommendations to an individual which wasnt that perfect each time. The number of factors in the data and inappropriate utilization of its features were one of the major reason for the amount of inaccuracies in the generated recommendations. Some of the top applications such as KKBox, Deezer, TIDAL are still making use of such inappropriate approaches leading ineffective recommendations. Thus, a highly efficient recommendation engine proposed in this study makes it possible to address the challenges and to satisfy requirements of the applications. The proposed method sharply recommends appropriate music to the user and intelligently classifies the music as per every individual. Apart from that, it also overcame the issues with the traditional RS related to insufficient data. Hence, the proposed approach takes care of most of the critical things to direct more traffic and thus minimizing the churn rate.

1.3 Research Objective

Music, is often stated to be humans best friend. It has been regarded as one of the major source of entertainment for any kind of occasions. It is also held responsible for instant mood swings in humans. At the same time, it would be equally risky if the music played; is opposite to the users current emotion. In such cases, looking for the most appropriate music for an individual from a huge collection becomes a time-consuming task. Thus, to tackle this kind of challenges, more effective RS were needed to be developed to meet the desired goals. Various analysis and pre-processing of the data along with the use of effective data mining and modelling algorithms makes it possible to outcome more accurate music recommendations.

How intelligently can a legitimate music recommendation system be built in order to predict the chances of a user listening to a song repetitively from the available historical logs of the existing users using the improved ML approaches?

2 Related Work

2.1 Recommender systems and cognitive systems

Cognitive systems (CS) can be explained as the learning of the human brain related functions and activities. These systems are stated to be a key aspect in order to build an intelligent system. Haykin et al. (2014) It makes use of cognitive perception and entropic state for knowledge extraction and providing feedback. Perceptual attention is also determined based on the experimental outcomes. The study of CS states the explanation on state estimation and sparse coding related concepts. Thus, CS have been stated to influence the information flow based on the preceptor and entropic state of the system.

Moving on to the study about the recommender systems and its evolution phases with respect to the modest decision tree algorithms such as gradient boosting. This study focuses on the customer behaviour in terms of items that were clicked but havent purchased through ecommerce website. It has been stated to make use of classification techniques like gradient boosting and also has used collaborative filtering approach to generate recommendations Rawat et al. (2017). Further, a clustering algorithm like rough set algorithm has also been used as a similarity measure. This study has well described the flow right from extracting features to generating recommendations. A detailed understanding on predicting the item preferences on the basis of clicked and not clicked have also been specified. Lastly, valuable analysis has also been carried out in order to get more effective recommendations.

2.2 Machine learning (ML) based feature engineering and feature selection techniques

Feature selection (FS) is an important factor in the field of machine learning. It is one of the vital aspect when it comes to preprocessing of the data. It has been stated in this study that the purpose of FS is to eliminate the any irrelevant or ineffective features present in the data. In this study, the role of FS on water has been explained quite nicely by stating the removal of features leading to contamination of water Visalakshi and Radha (2014). It also accounts for some of the other tasks carried out like dimensionality reduction or data reduction in order to increase the predictive score as per the study. Apart from this, various criteria and concepts related to feature selection have also been mentioned. Thus, this ensures improved accuracy and accurate recommendations when applied to a large set of data.

Further, having the right FS technique is also equally important to obtain more appropriate recommendations. This study states some of the useful FS techniques in order to reduce noise, redundancy and removal of unnecessary features. It has highlighted a newest online based FS algorithm on top of the conventional techniques Devi and Sabrigiriraj (2018). The aim of this study is to provide a suitable FS technique that leads to better solutions to be used in the process of building a machine learning based recommendation engine. Algorithms such as SVM, Logistic Regression and Nave Bayes have been utilized to assess the FS algorithm performance. Later, the accuracy scores of all the created techniques have been compared and QSVM with apriori particle swarm optimization has found to be performing the best amongst them.

Moving on, other than feature selection, feature engineering (FE) is also equally important and useful when to comes to improving the accuracy and F1-scores. This study discusses the usage of gradient boosting (GB) classification technique for theft detection. It also states that GB improves both detection as well as false positive rate (FPR) with the help of stochastic features. GB also minimizes the complexity of the feature classifier as per the study Punmiya and Choe (2019). In this, GB has been stated to be used in the theft detection with minimized FPR and data storage capacity. This study has proposed a novel GBTD approach build from XGBoost, CATBoost and LightGBM, out of which LightGBM proved to be the best performer

from the numerical results. Thus, such improved decision trees and gradient boosting algorithms can be used as an effective technique to process the music data and select the most appropriate features to predict precise music preferences.

Apart from the feature selection and engineering, there are also other crucial factors that leads towards building a better predictive model. One such factor is the hyper-parameter tuning of the ML model parameters. The advantage of using this technique is it outputs high accuracy by considering the best possible parameters effective for the model Mantovani et al. (2016). Grid, random and Bayesian optimization are three of the popular parameter tuning techniques. The study states improvement to the existing scores of the implemented model.

2.3 Analysis on classification methods and recommendation techniques for effective classifications

In this study, a comparative analysis for classification of feature-based song genre using a number of classifiers has been carried out. Various classification methods have been utilized to categorize the music based on different genres. Later, a comparative study between these algorithms have been carried based on the evaluation metrics used. Methods such as Fast Fourier Transform (FFT) and Mel Frequency Cepstral Coefficients (MFCC) contributing towards making a feature-rich data Kumar et al. (2016). The music tracks have been classified into several categories like jazz, rock, silent and folk. To classify these genres, algorithms such as Decision trees, Logistic Regression (LR), SVM, KNN and Recurrent Neural Network (RNN) have been used and compared against each other based on their accuracy scores. The purpose of this approach is to suggest user-specific music related to similar genre. In this study, LR and MFCC have been stated as the methods with highest accuracy.

A study on recommendation system (RS) is also important in order to develop an efficient music recommendation engine. A comparative performance-based analysis on the algorithms chosen to build the RS has been carried out, to know the advantages and drawbacks associated. A personalized recommendation method to predict music preferences has been proposed in this study. With this, it becomes easier to detect the current mood of an individual. Both implicit and external feedback is essential in the process of generating more accurate recommendations Patel and Wadhvani (2018). Thus, to satisfy this purpose, a graph based technique has been undertaken. Other than this, to address the recently updated data, context aware music recommendation method has been utilized. Hence, the proposed technique has shown positive results irrespective of the users active status.

2.4 Interpretation of user interactive behavioural patterns for effective recommendations

The study describes about the role of the user based recommendation systems to identify the behavioural pattern of a specific user. It states the usage and learning of users historical data as well as the application content. There arise problems in generating recommendations if the user is a first-time visitor or an unidentified user Gluhih et al. (2016). An adaption process using the similarity criteria provides more accurate recommendations as per the study. To achieve such higher accuracy, it makes of functions such as utility and similarity estimation. This adaption process based on feature similarity provides an efficient solution and has been stated to address the recommendation issues for unknown users. It makes use of iteration formula to re-evaluate the relative frequency scores for every iteration. Thus, this process makes it possible to provide user specific recommendations without much of external data.

Contextual information also plays a major role to build a music recommendation engine. A weighted combination of users context based information like weather, location etc have been undertaken to develop a constructive system to predict most favourite music as per the user

preferences. To achieve this, it makes use of a rating algorithm to detect the closeness of musics context as per the study Dolatkia and Azimzadeh (2016). This algorithm when applied on the combination of context, allows the users to select a variety of music as per their interest based on their current conditions. The experimental outcomes obtained from this study has also shown effective recommendations.

Moving further, understanding the selection behaviour or interacting pattern of an individual in ecommerce systems is also essential. In this study, a set of complex network tools has been utilized to analyse the movie based data. Average similarity of movies oftenly watched, dissimilarity between movies if watched after a long gap and also the movie stickiness on demand are some of the measures undertaken to predict a user behaviour Wang et al. (2018). A user interaction simulation technique to predict the behaviour has been proposed in this study . These technique is capable enough to address the requirements of complex networks as per the study. Adding to it, various content based recommendation methods has been described and also stated the integrated recommendation method as the top performer. Thus, the proposed network-based model has been said to uncover the interactive patterns of users in the field of recommendation systems.

2.5 Conventional approaches followed to build music recommendation system

Various approaches and strategies have been followed to build an efficient music recommendation system based on different parameters like artists mood similarity. It works on learning users mood to further recommend accurate music from the huge music database. Also, the interactions with the user interface by varying inputs have been evaluated in this study. With this, it makes it more efficient to understand the user behaviour to recommend music and also highlights how vital is the role of mood when it comes to listening to music. This study focuses on both the users as well as artists mood for more effective recommendations. An interactive recommender engine has been proposed in this study in order to predict music preferences based on the input and the artist mood. Thus, a hybrid recommender system, MoodPlay, which assumes a powerful link between user emotions and music has been stated to be a proven recommendation method Andjelkovic et al. (2019). To calculate the active duration of a user on the application, measures like log analysis have been performed to assess the user satisfaction level. The outcomes obtained by considering combination of musical features are also higher and provides improved recommendations. Thus, utilizing such system leads to constructive results and is advantageous to use it in the process of developing a recommendation system.

Acoustic information and listening feedback also contributes developing an automated music recommendation system. To address the concern of sparse historical data, users listening pattern has been targeted to predict the music as per users interest which would also target problems such as cold start. Also, the already existing data can be updated with the recently launched songs without any listening history. Therefore, a Codeword Bernoulli Average (CBA) model has been proposed to evaluate any relation between the recently added songs and the listening behaviour Borges and Queiroz (2018). It takes the listening feedback in terms of partial or complete listening of a music to predict the user preferences as per the study. The proposed model has been stated to predict the binary values corresponding to users listening feedback. Thus, it has been successful to predict the same with good results as per the study and hence it could be utilized while building a music recommendation system as it also supports addition of new music to the existing database.

There is also another study on a personalized music recommendation system using improved KNN algorithm. It has preferred improved KNN over KNN due to presence of high error rate in KNN. The improved KNN has been built upon the conventional KNN as well as a baseline algorithm Li and Zhang (2018). This approach decreases the amount of inaccuracies that leads

to increased error rate and inaccurate outcomes. A performance comparison of the proposed algorithm has been carried out with other recommendation algorithms like centered KNN, SVD and NMF. The study also suggests improvement in the quality of the data to increase the predictive power of the recommendation system and decrease the error rate.

Further, it is also important to assess problems like cold start that arose in case of newly updated music. To address this problem, a hybrid music recommendation system has been proposed in this study. Music playlist and all of its characteristics also needs to be undertaken to build this system Vall and Widmer (2018). Adding to it, music granularities and the relation between the music and the playlist has also been considered. It has also mentioned the use of collaborative filtering and content-based pattern extraction technique with respect to listening behaviour. Basically, the hybrid system has been built upon the traditional system. This approach is split into two process, first which predicts users music preferences and the second which fetches the next song recommendation. Thus, this hybrid approach helps to assess the recent interactions of an individual with the application for generating accurate recommenations.

A hybrid filtration technique to filter out unwanted records also plays an essential role in building an efficient recommendation system. It highlights the fact that not considering some useful musical attributes leads to high amount of inaccuracies in the developed system, despite of high accuracy score at times. Music genes plays a vital part in the process of developing such system as per the study. Weights have been assigned to all the music with these genes and arranged in the highest order of priority. In this way, it makes use of useful musical genes and the likes of collaborative filtering to build an improved music recommendation system Wu (2019). Further, the presence of different layers like presentation, data access layer and couple more in the architecture have been mentioned to perform each of the respective tasks. A detailed flow-chart of the entire hybrid system has been represented which includes the overall flow from the users listening history to generating the recommendations. Thus, this hybrid approach has proved to perform better than many of the singular approach and hence said to produce accurate recommendations as per the study.

2.6 Study of efficient ML methods in the context of recommendation systems

Gradient boosting decision trees (GBDT) are an effective method in the context of recommendation system. In this study, a GBDT model for consumption of electricity in commercial buildings has been proposed. It also highlights the fact that to generate effective predictions, huge amounts of data is required as it will pave the way for a power efficient model which ultimately leads to minimal energy usage Touzani et al. (2018). Thus, GBDT algorithms have been utilized to fulfil the required purpose. In order to evaluate the performance of the predictive models, new techniques have been experimented. M&V, SVM and Gaussian mixture has been stated to be some of the techniques to be undertaken to build a hyper-parameter based electricity consumption model. GBDT based models have been proven to be the best contributor as it showed increased R-squared accuracy and Root Mean Square Error (RMSE) in comparison to the mostly used algorithms such as random forest and regression. Thus, this method signifies to be a preferred choice to develop a music recommendation engine.

EXtreme Gradient Boosting (xgboost), a GBDT method, is an improved and effective technique for developing a RS. In this study, the online usage pattern of an individual has been understood to generate user based recommendations. Thus, xgboost derived from GBDT has been used to generate accurate recommendations and also to extract important features required for effective performanceXu et al. (2018). Considering all this, a high performing RS based on the individuals usage pattern has been proposed. In this, the recommendation problem has been transformed to a binary classification problem. Thereafter, a users usage data has been utilized to predict their buying behaviour. The presence of inconsistencies due to sparse his-

torical records, resources and noise has been addressed. Thus, the stated approach has proven to have achieved increased performance in terms of standard evaluation measures, but it also highlights a scope to resolve the cold start problem.

Next is one more GBDT based algorithm, LightGBM, which is another popular technique in the field of classification for RS. The study states the effectiveness of using such decision tree methods. It highlights the concerns related to low scalability and efficiency scores even when evaluated with high dimensional huge size data. This concerns are addressed by estimating the information gain which can calculating from the scanned data instances. For this, LightGBM has been proposed as it evaluates the information gain using only a specific set of data instances. This approach has been built upon Exclusive Feature Bundling (EFB) and Gradient-based 1-way Sampling techniques (GOSS)Wang et al. (2017). GOSS makes it possible to consider only certain instances that provides high information gain. To minimize the amount of features, the mutually exclusive features are bundled using the EFB technique as per the study. Thus, this LightGBM based approach has been said to perform much better as compared to the traditional GBDT algorithms when measured in terms of accuracy.

Moving on to the next study, an advanced and improved random forest algorithm has been utilized in order to develop a multi-dimensional RS. The presence of issues in the conventional context-based RS has been mentioned in this study. Adding to it, an improved multiple feature analysis method has also been mentioned in this study Li et al. (2018). The approach followed by the study is firstly to calculate the context weights and then recommend best items as per the weights. The overall flow of the proposed approach has been explained in this study along with processes such as initialization, weighing, preference predictions and many such. It also has proved to reduce the MAE and RMSE score as per the experimented results.

A SVM-based RS is another technique to be studied as it has gained positive results in the field of recommendations. It has followed a conventional approach which is the collaborative filtering. The time consumption and concerns related to user rating predictions have been highlighted in this study. Such concerns are said to be addressed by the proposed SVM based algorithm Ren and Wang (2018). Apart from that, unwanted suggestions are filtered out using a hyperplane. The preferences of a user are predicted based on the ratings evaluated in this study. This evaluation has been carried out using multiple steps such as evaluating performance, choosing parameters and then also performing comparative study with respect to various measures to test the accuracy. Hence, the recommendations generated using this approach has showed effective results.

3 Methodology

The study focuses on building a user-based music RS build upon their usage pattern and interactions. This system targets both the new and subscribed users, and studies their interactive behaviour based on the type of users. Thereafter, the system recommends more accurate and legitimate music after successfully learning their behaviour using most of the powerful ML techniques. The study follows a Cross Industry Standard Process for Data Mining (CRISP-DM) approach to build such a system as shown in figure 1.

[1]

[1]https://www.researchgate.net/figure/CRISP-DM-process-model_fig3_261307514

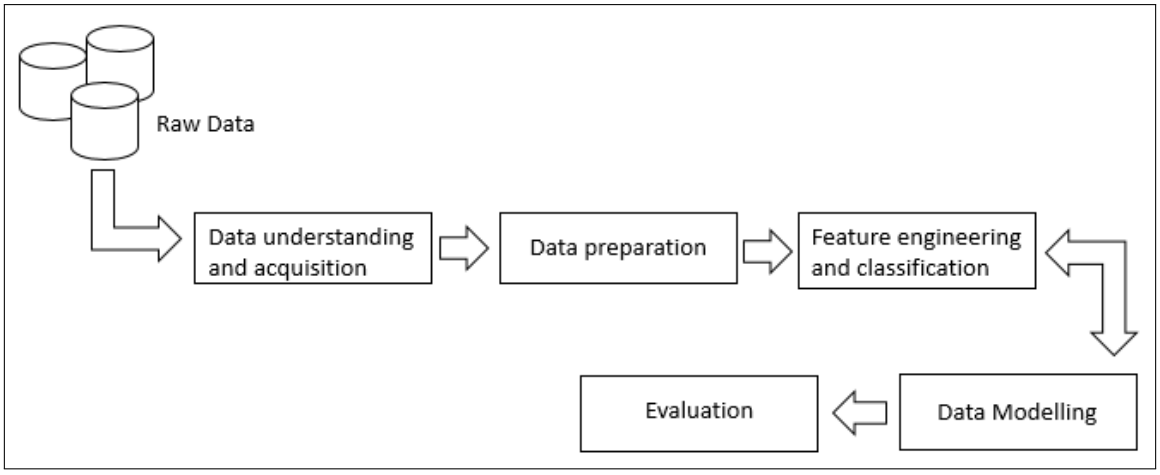

Figure 1: CRISP-DM

3.1 Business Understanding

Music has been one of the major source of income in the field of entertainment. Due to this, there have been a lot of competitors competing for the top spot to attract more and more traffic. The continuous advancements in the types and sources of music right from its existence, has made it a more diverse field with ample amount of scope. At the start, the only digital source of music was through the transistor which used to play random tracks without any music choices. Soon after that, DVD player were brought into picture which addressed the issue of choice-based music up to certain extent. Much later, with the rise of internet, the distribution and consumption of music began to grow rapidly throughout. Such interesting statistics have led to an increase in the number of music related applications. With hundreds of music applications being available online, it becomes difficult for a user to keep a track of their favourite music each time in each of the applications being visited on a daily basis. In most of the applications, the music tracks being recommended are often based on the recent ongoing trends and not on anyones personal taste thereby leading to visitor churn which is a huge loss to the business. To address such limitations , a personalized RS that works on individuals behavioural patterns is essential to generate on-point recommendations as per the expectations. Various previous researches in the similar context has been studied Patel and Wadhvani (2018) and a novel recommendation approach from the gained knowledge has been implemented. The implemented system meets several requirements that makes it a more effective and a more profitable system to the business.

3.2 Data understanding and acquisition

In this study, the data used to carry out the research is taken from the Kaggle website. The data available on this website is officially provided by the KKBOX music streaming organization. It thus provided several datasets like that of the songs, subscribed members and on the additional songs information. In Kaggle, apart from the unofficial data provider entities, many licensed companies use this website to put up a paid challenge with the intent to get a descent contribution from the online machine learning (ML) community which will further help to bring improvements to their online products. Moving on to the study, the downloaded datasets have been merged into a single dataset based on the effective attributes which is further used to train the recommendation model. Such effective utilization and processing of the fetched data contributes towards increasing the recommendation ability. The major data file, songs data, provided in this study contains around 40 thousand records; using which the over-

all music tracks can be classified into different categories. The other two datasets helped to perform the classification of music as well to consider only the legitimate tracks in order to get quality and trustworthy recommendations Kumar et al. (2016). Thus, in this way, the data is being pre-processed and different ML models are applied on this data using a uniquely defined approach.

3.3 Data Preparation

1. Figure 2 shows Songs dataset, that contains information about the length, artist, genre and more such related attributes.

2. Figure 4 shows Members dataset, that contains personal information of the KKBOX subscribed members such as the city, sex, birth date and subscription details.

3. Figure 3 shows Songs additional details dataset, that contains extra information about the given songs.

3.3.1 Raw Data Tables

song_id	song_length	genre_ids	artist_name
CXoTN1eb	247640	465	å¼µä¿¡å“² (Jeff Chang)
o0kFgae9C	197328	444	BLACKPINK
DwVvVurfp	231781	465	SUPER JUNIOR
dKMBWoZ	273554	465	S.H.E

Figure 2: Songs data

song_id	name	isrc
LP7pLJoJFI	æˆ‘â€˘	TWUM71200043
ClazTFnk6r	Let Me Lo	QMZSY1600015
u2ja/bZE3i	åŽŸè«‘æˆ‘	TWA530887303
92Fqsy0+p	Classic	USSM11301446

Figure 3: Songs extra information data

msno	city	bd	gender	registered_via	registration_init_time	expiration_date
XQxgAYj3k	1	0	male	7	20110820	20170920
UizsfmJb9	1	0	male	7	20150628	20170622
D8nEhsIOE	1	0	female	4	20160411	20170712
mCuD+tZ1	1	0	male	9	20150906	20150907

Figure 4: Members data

The effective attributes from all the three datasets are examined and combined into one single dataset as showed in Figure 5, which has been further utilized to build the recommendation system using mentioned the ML algorithms. [2]

3.3.2 Merged Data

The attributes retained from the three datasets are the ids of the user and song, source type, screen name and the system tab, isrc code and finally the target attribute. All these attributes are merged into one final dataset based on which an efficient RS has been developed as per the pre-defined approach. The pre-processing measures such as classification of music and removal of unwanted records have also been carried out. The user listening pattern has also been fetched from the provided data attributes. In this way, the proposed RS has been developed upon the merged data obtained from the existing independent datasets.

[2]https://www.kaggle.com/c/kkbox-music-recommendation-challenge/data

msno	song_id	source_system_tab	source_screen_name	source_type	target
FGtllVqz18	BBzumQN	explore	Explore	online-playlist	1
Xumu+NIjS	bhp/MpSN	my library	Local playlist more	local-playlist	1
Xumu+NIjS	JNWfrrC7z	my library	Local playlist more	local-playlist	1
Xumu+NIjS	2A87tzfnJ1	my library	Local playlist more	local-playlist	1

Figure 5: Merged data

3.4 Feature Classification

Data preparation is the initial stage in the process of developing the RS. Apart from this, another stage that is equally important after preparing the data is the feature classification stage. Such pre-processing and classification activity is carried out using n number of steps. Initially, the missing or the null values are identified from the given datasets. After that, the rows consisting of the NaN values are replaced with a pre-defined constant. Further, the classification of songs into several musical genres based on the albums, artist has been carried out. Once the music tracks have been categorized, the next step is to sample the data into train-test data sets. Thereafter, there is also a requirement to remove the non-legitimate tracks to make it a more trustable system. At the end, the available features are ranked as per their importance and on their predictive capability towards predicting the target attribute. Thus, such classification and pre-processing of data is must to gain a high performance ML model.

3.5 Modelling

The different ML algorithms like Gradient Boosting, Naive Bayes, Random Forest and Support Vector Machines have been described in this study; which has been utilized in the implementation of the proposed RS.

1. **Gradient boosting**
   Gradient boosting is an effective supervised ML algorithm build on top of decision tree that serves the purpose to solve classification related problemsPunmiya and Choe (2019). It is often regarded as a high performing method with high predictive capability.

2. **Random forest**
   Random forest, one more decision tree based ML technique, is also used for the same purpose to classify the features based on its importance Li et al. (2018). It is basically a combination of multiple decision trees used to predict the important features and then using such them to build a high performance recommendation model.

3. **Naive Bayes**
   Naive Bayes is also a supervised ML technique along with Random forest. It is a probabilistic ML model used to serve the purpose of classification of music and then recommending the same to each the users Haykin et al. (2014). It works on several assumptions like the independence between the independent variables which means the features are not related to each other in any of the aspect.

4. **Support Vector Machine**
   Support Vector Machines is also another supervised ML algorithm used for classification and recommendation of songs in this studyRen and Wang (2018). Gaussian radial basis kernel function (RBF) is used to implement it. It groups the music into several different categories in the form of a 2-D hyperplane.

3.6 Evaluation

In this research, there are several performance measures evaluated to test the predictive power of the implemented system. Measures such as the accuracy, recall, F1 score, precision, sensitivity and the specificity are evaluated. The accuracy is directly proportional to high percentage value. A high precision states the presence of low false positive rate and a high recall rate corresponds to value above 0.5. A F1-score close to 1 is often considered as a better one. Sensitivity is referred to as the amount of true positive values correctly classified by the RS whereas specificity is the amount of true negative values correctly classified by the RS. Apart from all this, a correlation matrix and a confusion matrix is also evaluated to check out the correlations between the independent and the dependent variables.

4 Implementation

The architecture followed in the overall implementation of the project is represented in the below figure 6. From the architecture, the initial stage is the data collection stage which explains downloading of the datasets from the online source. All these datasets are merged into a single data file which is further split into train and test data sets. Thereafter, once the data is merged, various pre-processing measures are carried out and then later the data is classified to bring the data in an appropriate format such that the proposed ML models can be effectively applied on it. At the end, a comparative analysis activity has been performed to know the top performing ML based approach using various metrics like the plots and graphs.

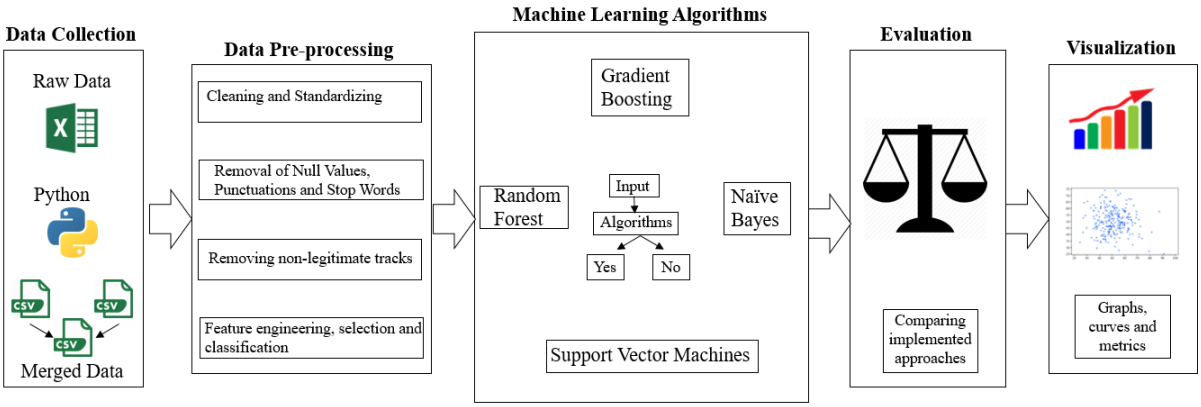

Figure 6: Project architecture

4.1 Data Collection

The data is obtained from the Kaggle data platform as a comma separated file. This data is uploaded officially by the KKOBX music streaming organization with the intent of challenging the open-source ML community by placing a paid challenge on it.

4.2 Data Pre-processing

1. **Data quality**
   The raw data downloaded from kkbox platform has been pre-processed using many of the built-in python libraries like sklearn, numpy, pandas, missingno, tqdm and few more.

The raw files have been downloaded and read using the pandas library. After that, all the data files are merged into one single file and then the attributes with missing values are handled using the missingno library. Also, the attributes containing NaN values are replaced with an explanatory constant, say -5. After handling the missing values, the attributes consisting of factors are encoded into binary levels using the LabelEncoder library. Thus, carrying out all the above steps improves the quality of the data and also eliminates unwanted noise from the data.

2. **Exploratory analysis of Data**
   EDA has been carried out on the raw as well as the merged data. It is an important aspect which helps to gain more clarify on the provided data. It basically investigates the data so as to discover effective patterns and trends using which can be considered in the development of the proposed system.

   BEFORE MERGING :

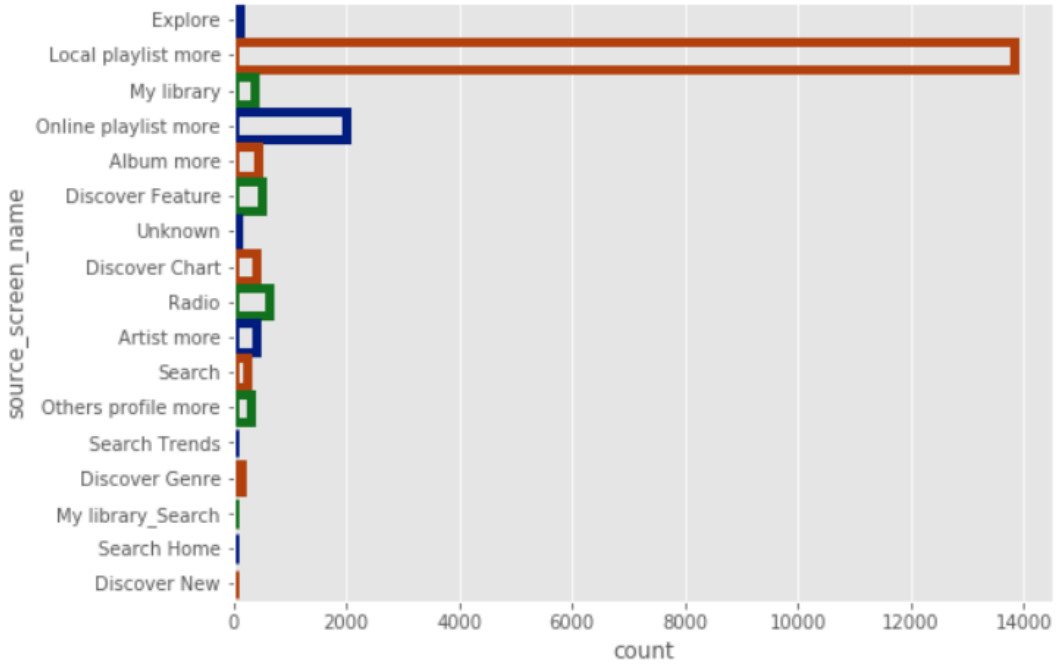

Figure 7: Source screen name count

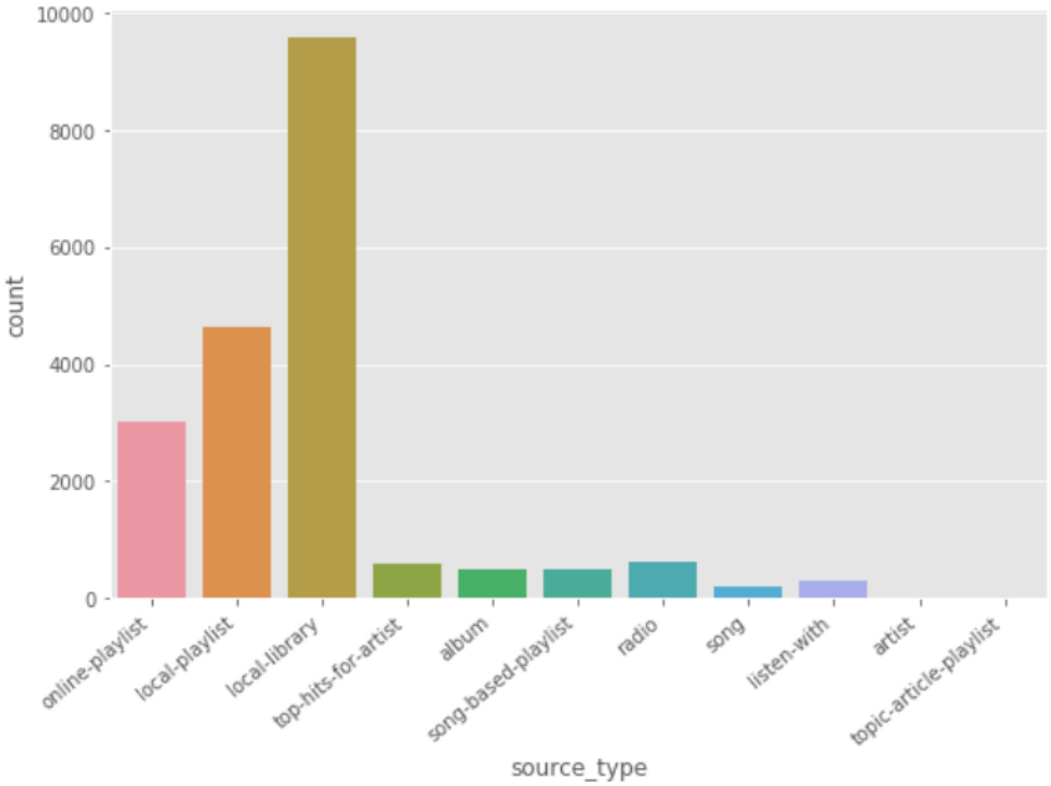

Figure 8: Source type count

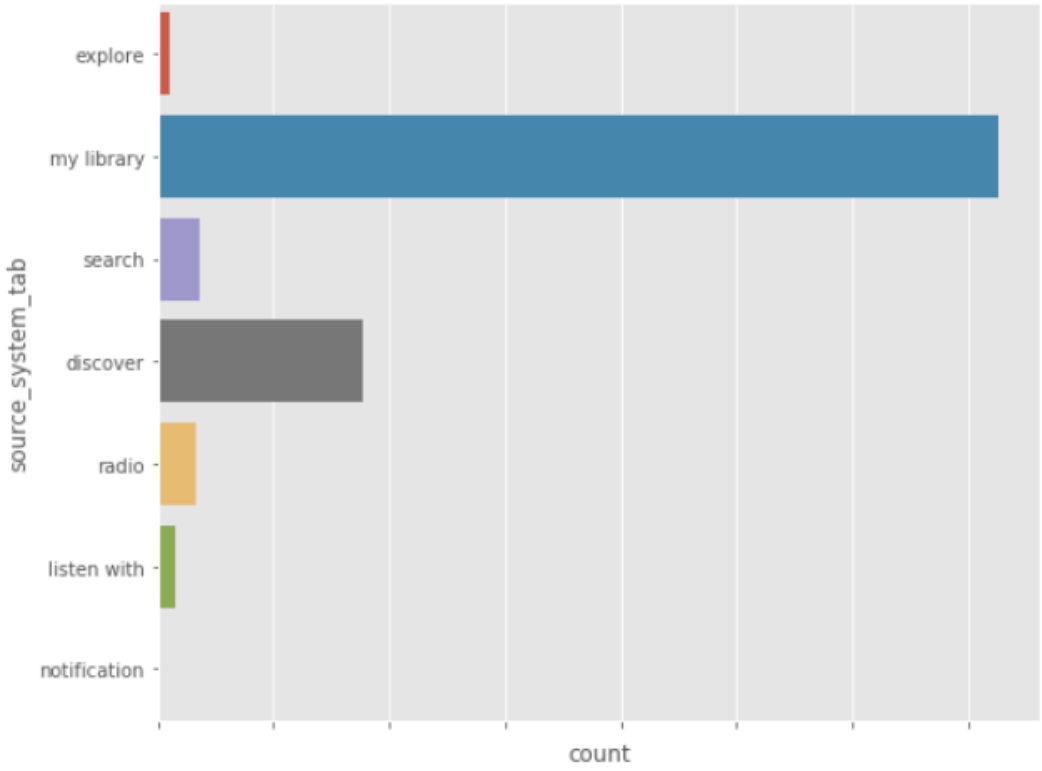

Figure 9: Source system tab count

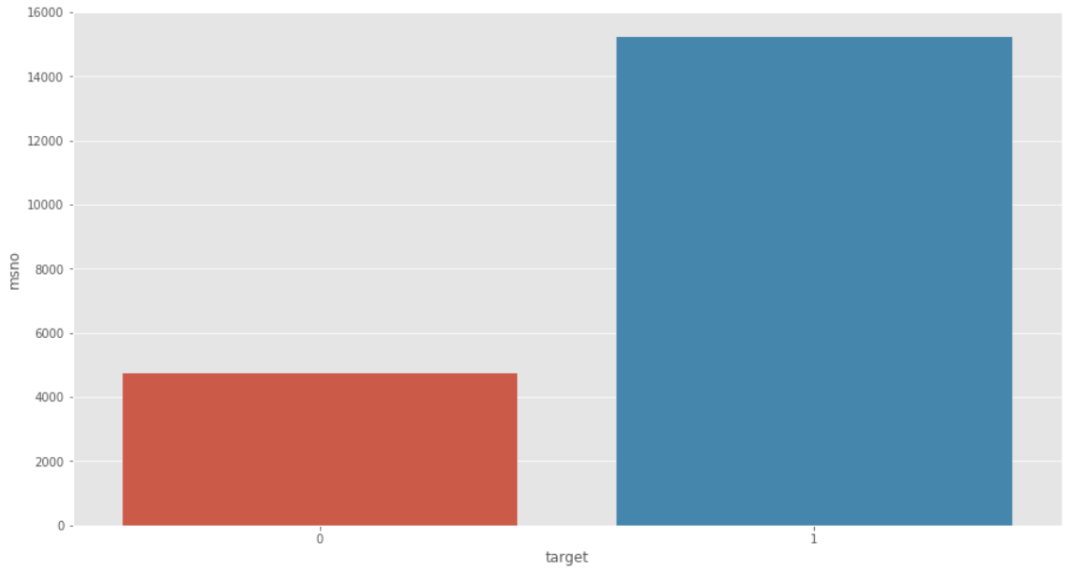

Figure 10: Target distribution count

From the above figures, initially the exploration has been carried out on the untransformed data. The total count of data attributes like the source screen name, source type, source system tab and the target variable distribution has been represented using bar plots. This visualization specifies the amount of distribution in each of the attribute. It further helps to determine the amount of bias present in the data as shown in the Figure 7.

AFTER MERGING :

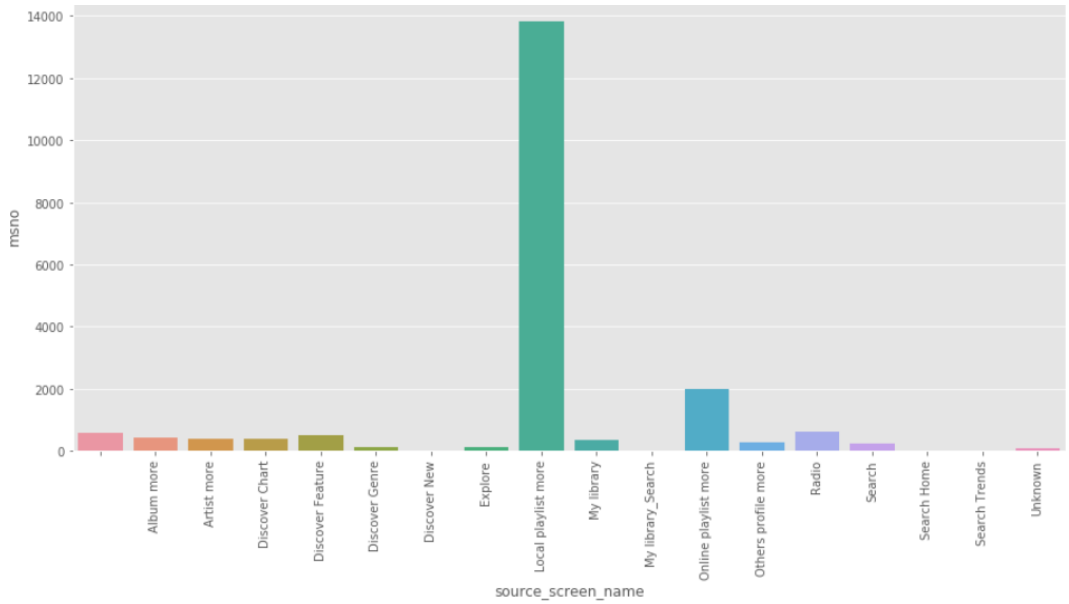

Figure 11: Source screen name count by members

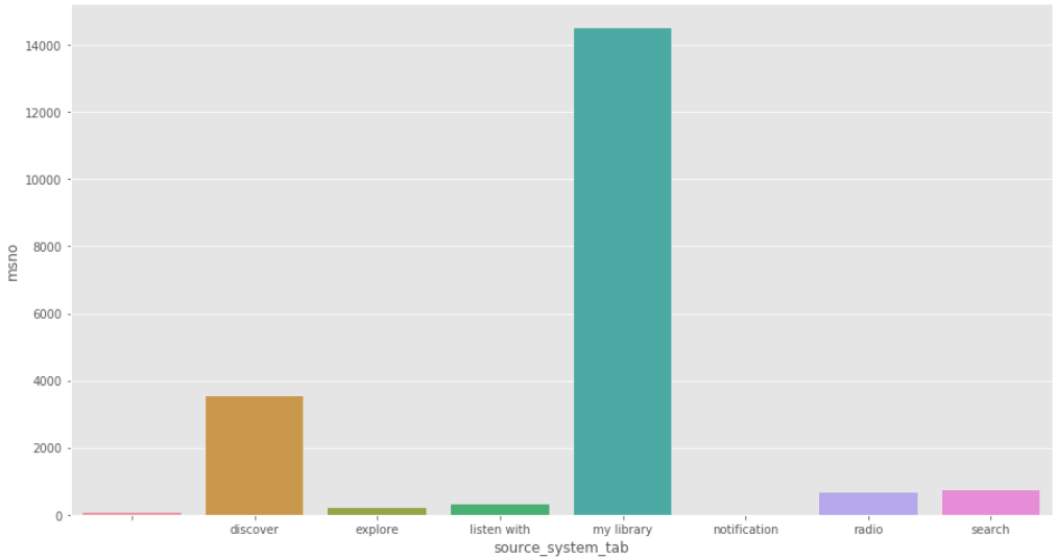

Figure 12: Source system tab count by members

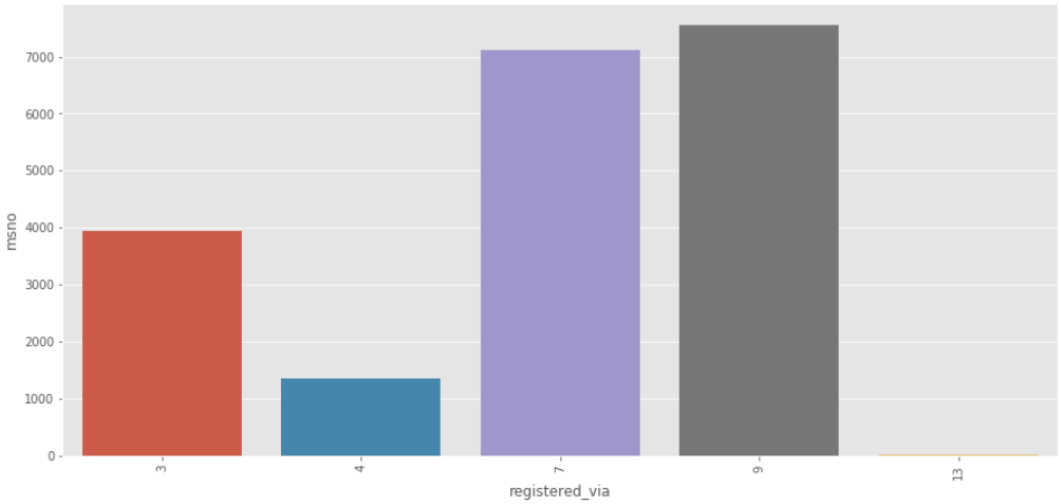

Figure 13: Registered source count by members

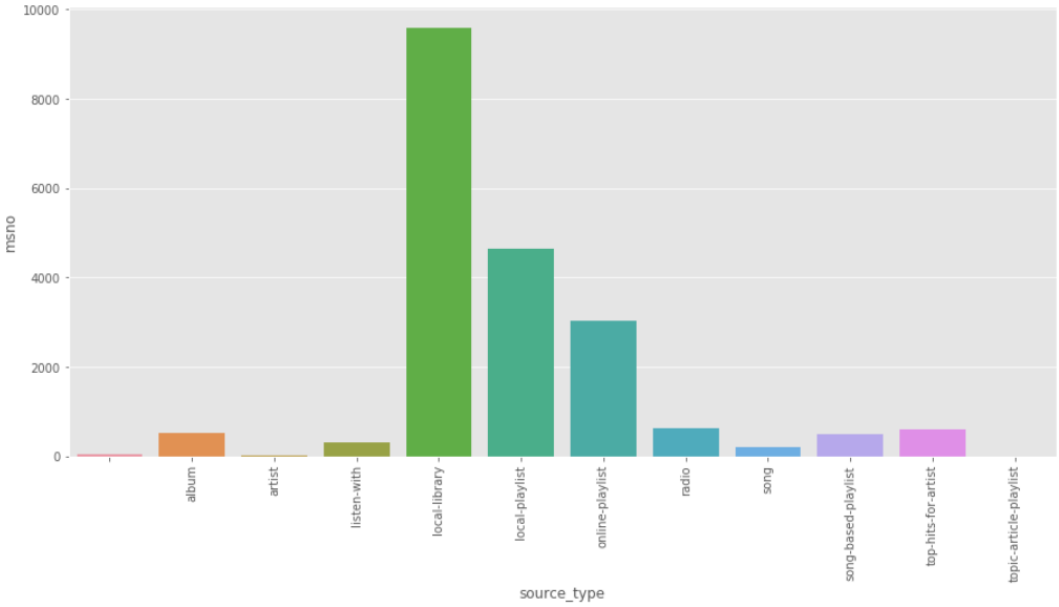

Figure 14: Source type members count by members

After the completion of initial exploration, the next analysis has been carried out on the merged data obtained from the raw data. The relation of various attributes with the subscribed members has been represented in the form of bar plots. The count of the above explored attributes with respect to the subscribed members have been visualized to better understand the relations between the members and the music attributes to learn the average usage pattern of each user as shown in the Figure 8.

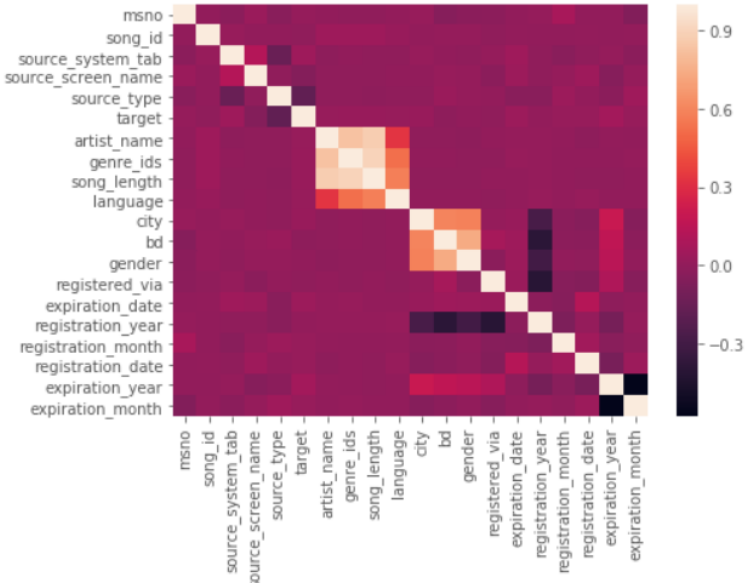

Figure 15: Correlation matrix

As shown in the figure 9, a correlation check has also been carried out in this research. The correlation matrix is used to estimate the linear historical relationship between multiple features in the given data.

3. **Feature Extraction and classification**
   After the data-processing activity has been carried out, the next stage is the extraction and classification of features as per required by the ML models. Features from the processed data are grouped based on the defined criteria essential for an effective RS. The features to be selected are determined using a random forest based feature importance technique which outputs a list of all the features ordered by their importance level as shown in the figure 10. Thus, these levels or score helps to decide the importance or effectiveness of an attribute with respect to the target. Later, the features with high scores are considered to categorize the music tracks based on several parameters such as the genres, albums. Thereafter, to ensure the legitimacy of the music in the database, the tracks with missing isrc code has been filtered out. At the end, the necessary operations are performed to mine the user behaviour from the provided data logs.

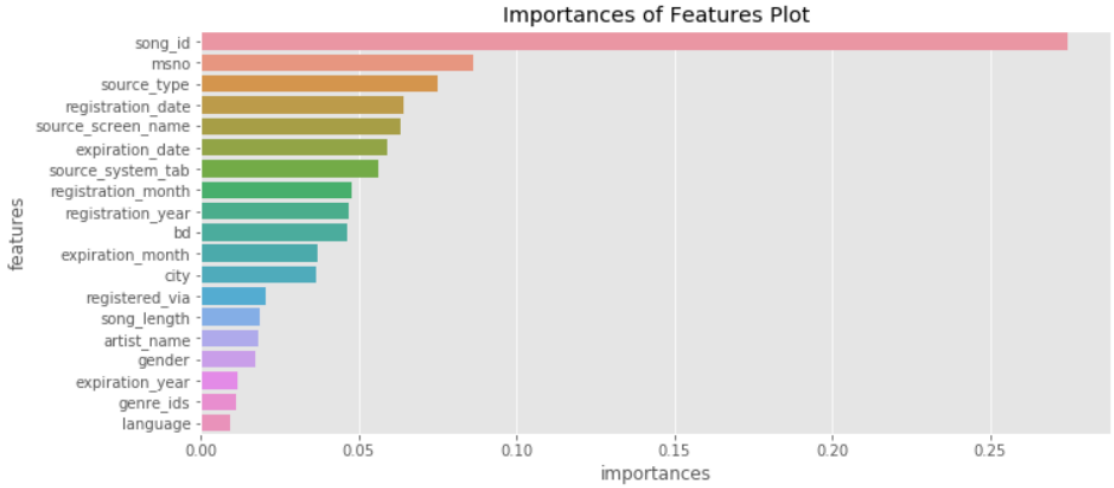

Figure 16: Feature importance

4.3 Machine Learning Algorithms

The research consists of implementing the stated ML algorithms with the intent to attain the best-performing method obtained from comparing the evaluated results. The ML algorithms implemented in this research have already provided effective results in the field of RS. The preprocessing carried out on the raw data differs a bit from model to model as this ensures best possible results in terms of appropriate recommendations. Apart from the feature importance, techniques such as parameter tuning integrated with grid and random search achieves high performance. Also, non-legitimate music tracks are discarded to eliminate any piracy related concerns and reduce customer churn. Furthermore, approaches like content-based filtering are used to predict the preferences of an individual from the given data files. Thus, the stated ML techniques are utilized to obtain highly accurate recommendations. Initially, Gradient boosting tree (GBDT) was preferred to implement as it is a proven method in the field of RS. Hence, GBDT based models, which are the LightGBM and the XGBoost are initially implemented in this research. Random forest (RF) is the next model implemented after the GBDT, as it has showed high performance with minimal error rate in similar areas. Other than this, RF is also used for feature selection and classification to determine the important features. Nave Bayes is one more classification technique implemented in this research. Another classification algorithm, SVM, is implemented due to its precise recommendations obtained in previous studies. Apart from these standard classification algorithms, some additional improvements by

means of effective pre-processing techniques also led to huge improvements in the results. It has the capability to learn and interpret the listening pattern or historical logs from the supplied data into the RS. In this way, it stores and manipulates listening logs of the users obtained by merging the data to generate on-point and exact recommendations. Hence, all the proposed ML methods have been implemented with the intent of building the user-specific RS.

5 Evaluation and Results

The metrics utilized for evaluating the model performances are as follows,

1. Accuracy is the ratio of total count of correct predictions to the total count of input data.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{1}$$

2. Precision is the ratio of total count of correct positive outcomes to the total count of positive outcomes predicted by the model.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

3. Recall is the total count of correct positive outcomes to the total count of relevant samples.

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

4. F1-Score is the harmonic Mean between precision and recall. It determines out the robustness of the model.

$$F1 - Score = \frac{2 * Precision * Recall}{(Precision + Recall)} \tag{4}$$

5. Specificity defines the model's ability to determine the negative results

$$Specificity = \frac{TN}{TN + FP} \tag{5}$$

6. ROC curve is defined as the plot of True Positive Rate against the False Positive Rate.

7. A confusion matrix is defined as a method for classifying the performance of a machine learning model as shown in the figure 11.

The evaluation of the five implemented ML methods has been carried out using the essential measures and metrics.

In the above table 1, the performance of five different ML algorithms is represented with respect to its accuracy. Out of all these, LightGBM using random search provided the best performance in terms of appropriate recommendations. It is followed by algorithms like SVM, Random Forest and Naive Bayes in which Nave Bayes provided the least performance.

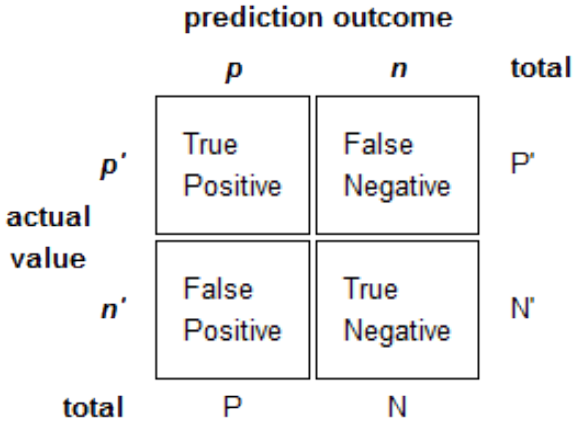

Figure 17: Confusion matrix

Algorithms	**Accuracy**
Using Random Forest based feature importance	
LightGBM	80.51%
XGBoost	83.78%
Naive Bayes	82.27%
SVM	76.03%
Random Forest	82.72%
Using Grid search	
LightGBM	75.86%
Using Random search	
LightGBM	86.16%
XGBoost	81.03%

Table 1: Accuracy scores for the implemented ML appraoches

5.1 Experiment 1: LightGBM (using random forest feature importance)

Recommendations	**Precision**	**Recall**	**F1-Score**	**Support**
Positive predictions (Yes)	0.82	0.95	0.88	4594
Negative predictions (No)	0.68	0.32	0.43	1406
Weighted Average	0.79	0.81	0.78	6000

Table 2: Metrics for LightGBM

Figure 18: AUC - ROC Curve

From the above table 2, the analysis is carried out using lightgbm ML technique. It is based on the random forest based feature importance approach. This experiment gained 80.51% accuracy. The precision score is quite balanced for both positive and negative type of classification. Recall and F1-Score have high positive classifications but low negative classifications. The specificity score is 0.55. Also the AUC score is 0.78 from the figure 12 which is regarded as a good score.

5.2 Experiment 2: XGBoost (using random forest feature importance)

Recommendations	Precision	Recall	F1-Score	Specificity
Positive predictions (Yes)	0.87	0.93	0.90	4594
Negative predictions (No)	0.70	0.55	0.61	1406
Weighted Average	0.83	0.84	0.83	6000

Table 3: Metrics for XGBoost

Figure 19: AUC - ROC Curve

This experiment, as shown in the table 2, is carried out using XGBoost based on the feature importance approach. It gained around 83.78% accuracy. The precision score is satisfying for both types of classifications but it is not the case with the recall and f1-score. The specificity score obtained is 0.55. The AUC score for this model is 0.87 from the figure 13 which is stated as an excellent score.

5.3 Experiment 3: Multinomial Naive Bayes (using random forest feature importance)

Recommendations	Precision	Recall	F1-Score	Support
Positive predictions (Yes)	0.88	0.89	0.88	3032
Negative predictions (No)	0.64	0.61	0.62	968
Weighted Average	0.82	0.82	0.82	4000

Table 4: Metrics for Naive Bayes

Figure 20: AUC - ROC Curve

From the above table 3, Multinomial Naive Bayes model is built in this experiment which is also based on feature importance approach. It gained the accuracy around 82.27%. The precision score along with the recall and f1-score shows better performance towards positively classified values but struggles with the negatively classified values. The specificity score obtained is 0.55. The AUC score obtained from this approach is 0.85 as shown in the figure 13 which is considered a very good score.

5.4 Experiment 4: SVM (using random forest feature importance)

Recommendations	Precision	Recall	F1-Score	Support
Positive predictions (Yes)	0.76	1.00	0.86	4566
Negative predictions (No)	0.17	0.00	0.00	1434
Weighted Average	0.62	0.76	0.66	6000

Table 5: Metrics for SVM

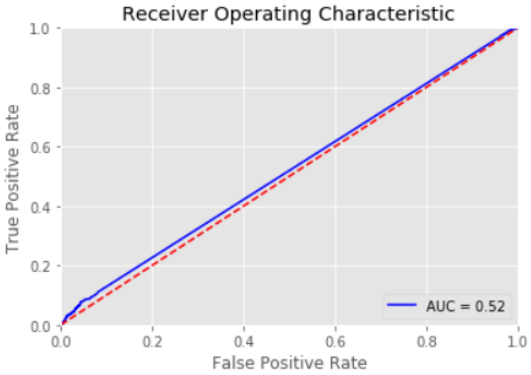

Figure 21: AUC - ROC Curve

This experiment, as shown in the table 5, is carried out using SVM, again based on the feature importance approach. It provided 76.03% of accuracy score. The precision, recall and the f1-score is much better for positively classified values but worse for the negative predictions. The specificity score obtained is 0.00. The AUC score obtained is 0.52 which is a average score as seen in the figure 15.

5.5 Experiment 5: Random Forest (using random forest feature importance)

Recommendations	**Precision**	**Recall**	**F1-Score**	**Support**
Positive predictions (Yes)	0.88	0.89	0.89	3039
Negative predictions (No)	0.65	0.62	0.63	961
Weighted Average	0.82	0.83	0.83	4000

Table 6: Metrics for Random Forest

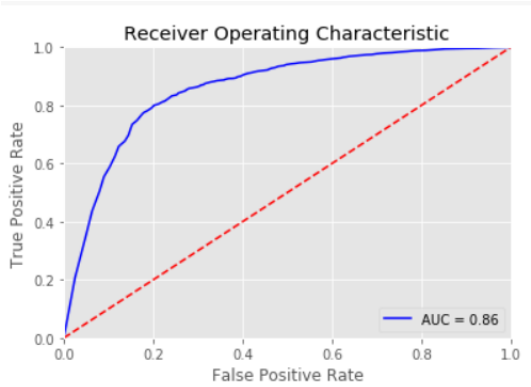

Figure 22: AUC - ROC Curve

This experiment, shown in table 6, is carried out using random forest based on its own rated importance features. It provided accuracy around 82.72%. The scores for precision, recall and f1 are very good for the positive predictions. It performs reasonably well for the negatively classified values. The specificity score obtained is 0.62. Also, the AUC gained from this is 0.86 seen from the figure 16 which is quite excellent.

5.6 Experiment 6: LightGBM (using grid search)

Recommendations	Precision	Recall	F1-Score	Support
Positive predictions (Yes)	0.76	1.00	0.87	4552
Negative predictions (No)	0.00	0.00	0.00	1448
Weighted Average	0.58	0.76	0.66	6000

Table 7: Metrics for LightGBM

Figure 23: AUC - ROC Curve

In this appraoch, shown in table 7, a grid-search based approach is utlized to build a lightgbm model. The accuracy gained by this technique is around 75.86%. All of the scores for positively predicted values are good enough but it doesn not predict the negative classifications at all. The specificity score obtained is 0.00. The AUC score is 0.72 as from the 17 which is a acceptable score.

5.7 Experiment 7: LightGBM (using random search)

Recommendations	Precision	Recall	F1-Score	Support
Positive predictions (Yes)	0.87	0.97	0.92	4600
Negative predictions (No)	0.84	0.51	0.63	1400
Weighted Average	0.86	0.86	0.85	6000

Table 8: Metrics for LightGBM

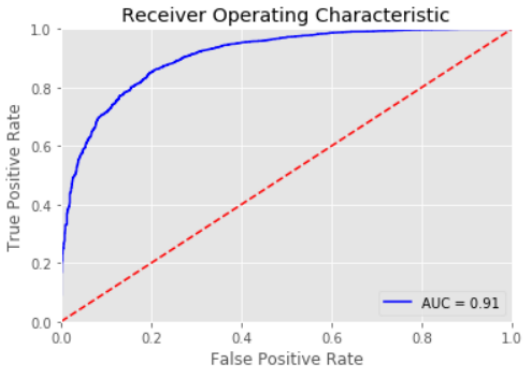

Figure 24: AUC - ROC Curve

As shown in the table 8, the experiment is performed using the random-search technique in combination to the lightgbm ML model. It gained accuracy score of around 86.16%. From the results, it yielded high performance for the positive classifications whereas it performed quite good for the negative predictions in terms of precision, recall and f1-score. The specificity score obtained is 0.51. The AUC score obtained from this approach is the best amongst all which is 0,91 and can be seen from the figure 18.

5.8 Experiment 8: XGBoost (using random search)

Recommendations	**Precision**	**Recall**	**F1-Score**	**Support**
Positive predictions (Yes)	0.82	0.95	0.88	4600
Negative predictions (No)	0.68	0.32	0.43	1400
Weighted Average	0.79	0.81	0.78	6000

Table 9: Metrics for XGBoost

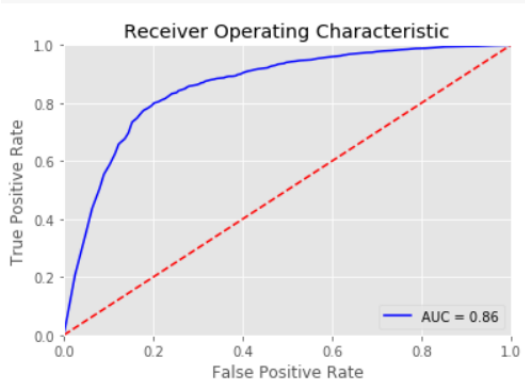

Figure 25: AUC - ROC Curve

As shown in the table 9, the above values are obtained from the grid-search based experiment carried out using the xgboost model. The accuracy score gained is around 81.03%. The precision, recall and f1-score provided the top scores in terms of positive predictions but provided below average scores for negatively predicted values. The specificity score obtained is 0.45. The AUC score gained from this approach is 0.85 which is considered as an excellent score as shown in the figure 19.

6 Discussion

In this research, a music RS is implemented for the KKBOX streaming organization. There have been several ML-based approaches followed to build the system. These approaches have their own processing and modelling of the data. Certain approaches work on tuning the parameters whereas others work on importance of features to obtain the best parameters. In KKBOX, the earlier system made use of conventional techniques like collaborative filtering with factorization of matrix which included a lot of problems like the cold-start, false recommendations, incorrect mood detection that leads to customer churn etc. Therefore, all these challenges are efficiently addressed in this research. Importantly, the data exploration and feature engineering proved to be the decisive factor for obtaining a high accuracy score after a thorough analysis of the dataset. A couple of feature selection packages like Boruta and MRMR have also been tried but were unfortunately not applicable on these data due to its limitations and incompatibility with the execution environment. Model selection and hyper parameter tuning are also equally important steps. There were other unsuccessful approaches such as recognizing user emotions due to the absence of acoustic data. Thus, the implemented approaches still yields better performance than the traditional recommendations using the same set of ML algorithms irrespective of the limitations which can be seen in the below figure.

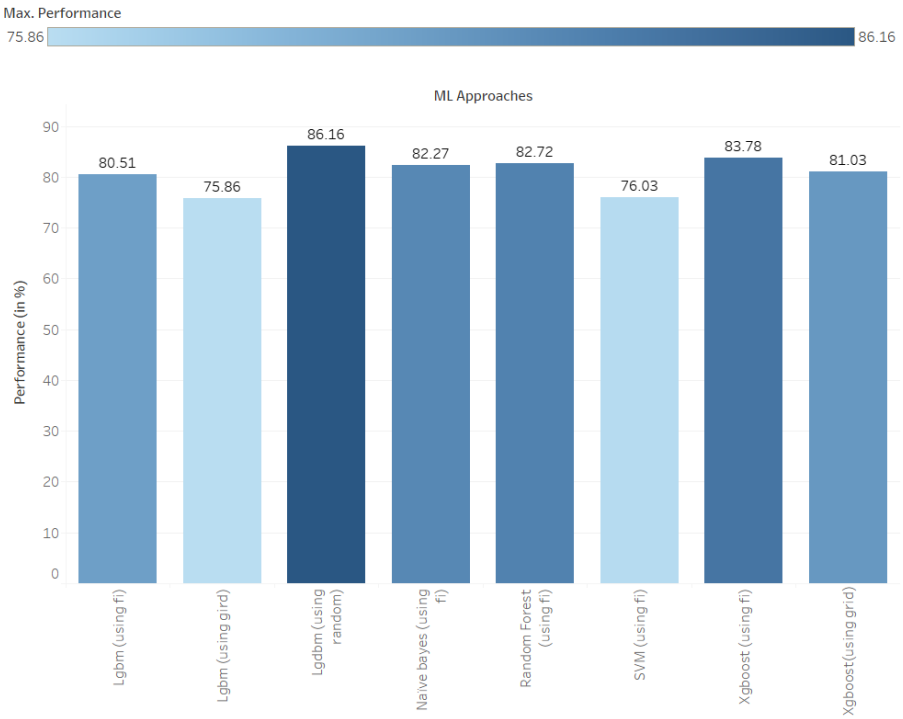

Figure 26: Model performance comparison

7 Conclusion and Future Work

In this study, a user-specific RS has been built based on the useful researches from previous studies, with the intent of improving the conventional approaches. The data pre-processing activity which includes improving the data quality, transforming the data into required form and classification of most important features responsible for accurate recommendations has been carried out. The developed system also overcomes the drawbacks like the cold-start, unofficial music tracks and scalability issues present in the conventional RS. Thus, this approach successfully meets the desired challenges of uncovering the usage pattern which helps to build a RS as per every individual. Therefore, out of all the algorithms, Light gradient boosting using randomized search technique proves to be the top-performer in terms of correctly recommended music to the kkbox music application users.

Additional information on the music listening time by the users is also an important point to be considered to make it more robust. Thus, there is enough scope left for improvement in this project in terms of more data collection, rapid addition of new music and to determine the exact churn time for a track listened by a particular user. To address the complexity and hidden patterns in the data, deep learning techniques can be undertaken to serve the purpose. Such methods would lead to increase in the classification score of the true negative values, specificity score, which is not that high in this research.

References

Andjelkovic, I., Parra, D. and ODonovan, J. (2019). Moodplay: Interactive music recommendation based on artists mood similarity, *International Journal of Human-Computer Studies* **121**: 142 – 159. Advances in Computer-Human Interaction for Recommender Systems. **URL:** *http://www.sciencedirect.com/science/article/pii/S1071581918301654*

Borges, R. C. and Queiroz, M. (2018). Automatic music recommendation based on acoustic content and implicit listening feedback., *Revista Musica Hodie* **18**(1): 31 – 43.

Devi, S. G. and Sabrigiriraj, M. (2018). Feature selection, online feature selection techniques for big data classification: - a review, *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*, pp. 1–9.

Dolatkia, I. and Azimzadeh, F. (2016). Music recommendation system based on the continuous combination of contextual information, *2016 Second International Conference on Web Research (ICWR)*, pp. 108–114.

Gluhih, I. N., Karyakin, I. Y. and Sizova, L. V. (2016). Recommender system providing recommendations for unidentified users of a commerial website, *2016 IEEE 10th International Conference on Application of Information and Communication Technologies (AICT)*, pp. 1–3.

Haykin, S., Amiri, A. and Fatemi, M. (2014). Cognitive control in cognitive dynamic systems: A new way of thinking inspired by the brain, *2014 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL)*, pp. 1–7.

Kumar, D. P., Sowmya, B. J., and Srinivasa, K. G. (2016). A comparative study of classifiers for music genre classification based on feature extractors, *2016 IEEE Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER)*, pp. 190–194.

Li, G. and Zhang, J. (2018). Music personalized recommendation system based on improved knn algorithm, *2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, pp. 777–781.

Li, X., Wang, Z., Wang, L., Hu, R. and Zhu, Q. (2018). A multi-dimensional context-aware recommendation approach based on improved random forest algorithm, *IEEE Access* **6**: 45071–45085.

Mantovani, R. G., Horvth, T., Cerri, R., Vanschoren, J. and d. Carvalho, A. C. P. L. F. (2016). Hyper-parameter tuning of a decision tree induction algorithm, *2016 5th Brazilian Conference on Intelligent Systems (BRACIS)*, pp. 37–42.

Patel, A. and Wadhvani, R. (2018). A comparative study of music recommendation systems, *2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*, pp. 1–4.

Portugal, I., Alencar, P. and Cowan, D. (2018). The use of machine learning algorithms in recommender systems: A systematic review, *Expert Systems with Applications* **97**: 205 – 227.
URL: *http://www.sciencedirect.com/science/article/pii/S0957417417308333*

Punmiya, R. and Choe, S. (2019). Energy theft detection using gradient boosting theft detector with feature engineering-based preprocessing, *IEEE Transactions on Smart Grid* **10**(2): 2326–2329.

Rawat, M., Goyal, N. and Singh, S. (2017). Advancement of recommender system based on clickstream data using gradient boosting and random forest classifiers, *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pp. 1–6.

Ren, L. and Wang, W. (2018). An svm-based collaborative filtering approach for top-n web services recommendation, *Future Generation Computer Systems* **78**: 531 – 543.
URL: *http://www.sciencedirect.com/science/article/pii/S0167739X17300389*

Touzani, S., Granderson, J. and Fernandes, S. (2018). Gradient boosting machine for modeling the energy consumption of commercial buildings, *Energy and Buildings* **158**: 1533 – 1543.
URL: *http://www.sciencedirect.com/science/article/pii/S0378778817320844*

Vall, A. and Widmer, G. (2018). Machine learning approaches to hybrid music recommender systems, *CoRR* **abs/1807.05858**.
URL: *http://arxiv.org/abs/1807.05858*

Visalakshi, S. and Radha, V. (2014). A literature review of feature selection techniques and applications: Review of feature selection in data mining, *2014 IEEE International Conference on Computational Intelligence and Computing Research*, pp. 1–6.

Wang, D., Zhang, Y. and Zhao, Y. (2017). Lightgbm: An effective mirna classification method in breast cancer patients, *Proceedings of the 2017 International Conference on Computational Biology and Bioinformatics*, ICCBB 2017, ACM, New York, NY, USA, pp. 7–11.
URL: *http://doi.acm.org/10.1145/3155077.3155079*

Wang, X., Li, H. and Zeng, A. (2018). Quantifying users selection behavior in online commercial systems, *Physica A: Statistical Mechanics and its Applications* **512**: 86 – 95.
URL: *http://www.sciencedirect.com/science/article/pii/S0378437118309580*

Wu, D. (2019). Music personalized recommendation system based on hybrid filtration, *2019 International Conference on Intelligent Transportation, Big Data Smart City (ICITBS)*, pp. 430–433.

Xu, A. L., Liu, B. J. and Gu, C. Y. (2018). A recommendation system based on extreme gradient boosting classifier, *2018 10th International Conference on Modelling, Identification and Control (ICMIC)*, pp. 1–5.