LastFM Recommendation System

RECOMMENDATION TOOLS
GROUP PROJECT – LASTFM RECOMMENDATION SYSTEM

Contents

INTRODUCTION

The following recommendation system was created to improve the possible suggestions that can be made to the users of the LastFM music platform.
For this purpose, we have used the dataset that was previously collected and supplied by our sponsor.

Data Exploration

DATA SETS

The datasets used to consist of the following files and variables:

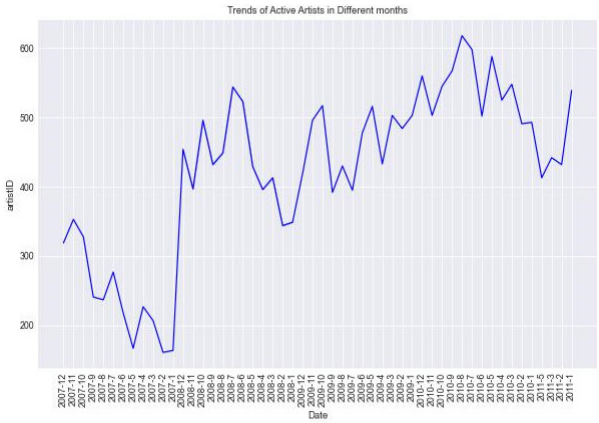
- Artists.dat: This dataset contains information about all the artists that are played in LastFM. These data tables contain four (4) variables; id, name, URL, pictureURL.
- tags.dat: The dataset contains two (2) variables with information about the music gender (tagValue) and the tag id (tagID).
- user_artists.dat: The user_artists dataset is composed by three (3) variables with information regarding the userID, artistID and weigh which can be translated to the times of an artist has been played.
- user_taggedartists.dat: This dataset contains six (6) variables regarding the user tagged artis. userID, artistID, tagID, day, month, and year.

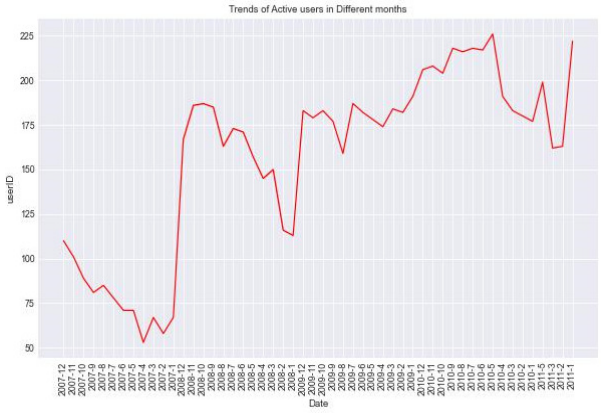
DATA CLEANING

The data exploration was done after reading all .dat files in our python notebook (ipynb), where we decided to create two new variables based the on the weight percentage localization. Our new variables rating 10 and rating 5 will be used to test the performance of our recommendation algorithms and thus choose the combination with the best performance.

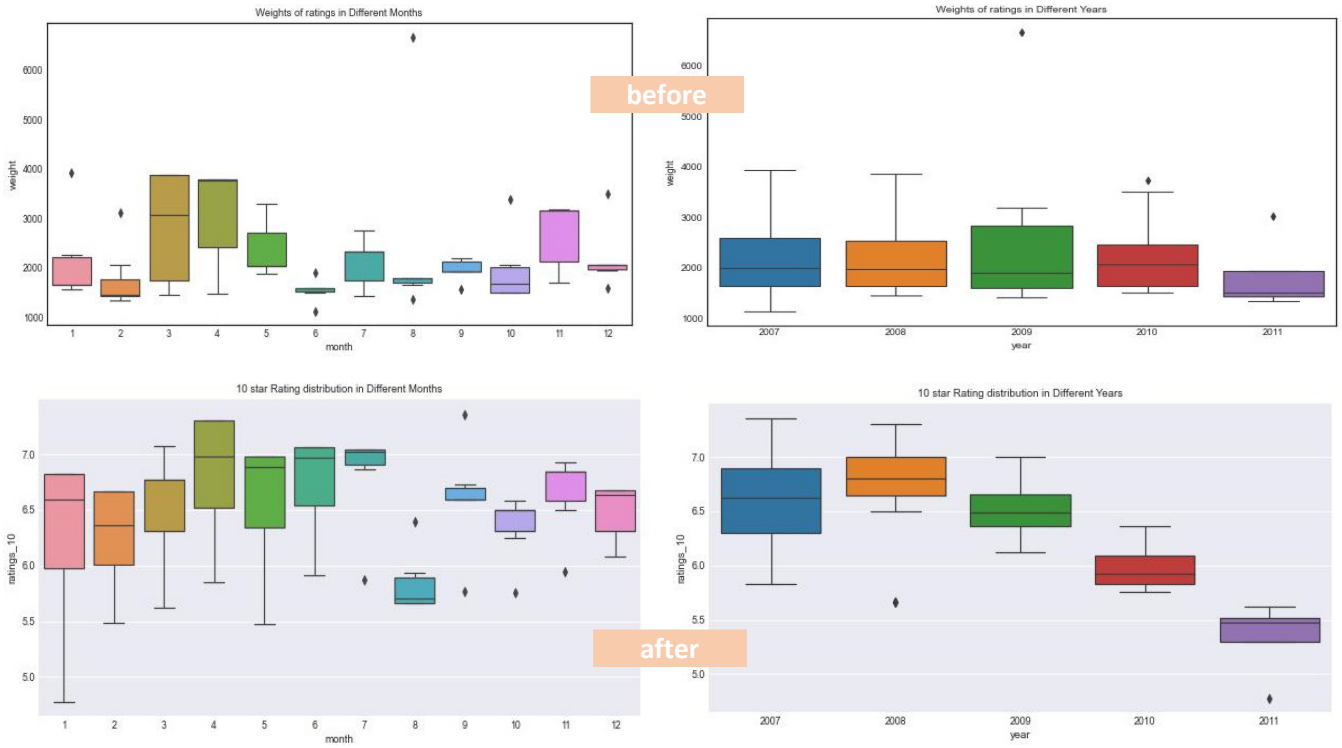
Description of the Data

In the charts bellowed we check the data based on the demographics and time.
In the first plot we can see the trend of artists and user activities between the years 2007 and 2011.

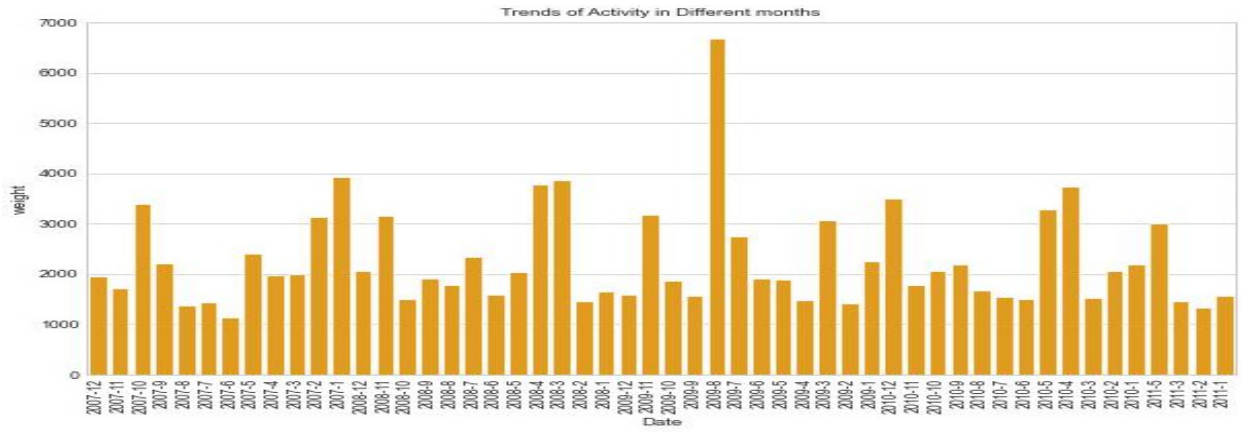




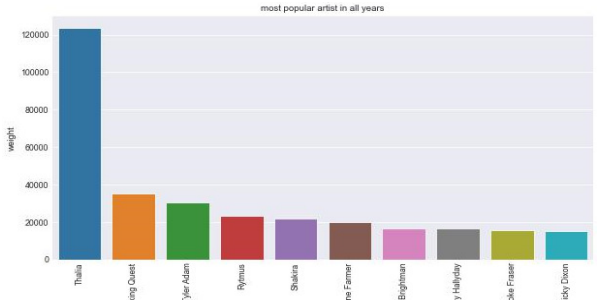
Our data is skewed and by exploring we can say we have the data from 2007 to 2011.
The weights are the ratings, but they don't show the popularity of artists in a good way so we will aggregate it, to use it in the evaluation process.
You can see the difference of distribution before the normalization of data. In the boxplots, shown below you can see the data normalized, we can interpret the data.

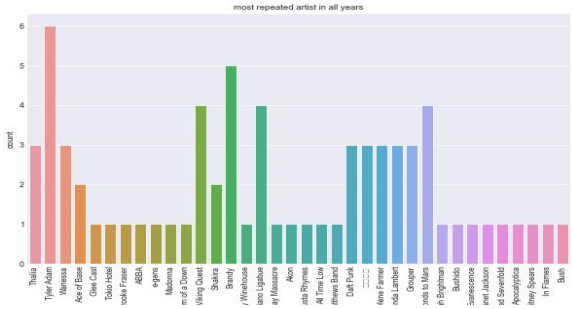


Below you can see the most active months of LastFM was the August of 2009.



In the next two charts you can see the most popular artist based on the number of times played by year, where we identified the top 10 artist and overall popularity of the artists.





Applying Models

Collaborative Filtering

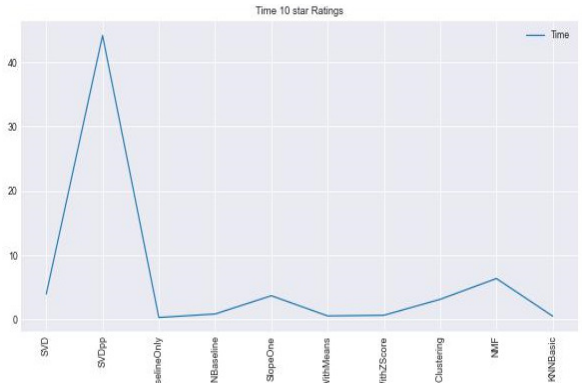
There are 10 different algorithms provided by the surprise package that we can use for collaborative filtering and here you can find a short explanation of each. The process of all algorithms is quite similar you can fit the model and use predict(for prediction of just one item) or test (to predict on the test set). You can evaluate the model using cross-validation and the RMSE index.

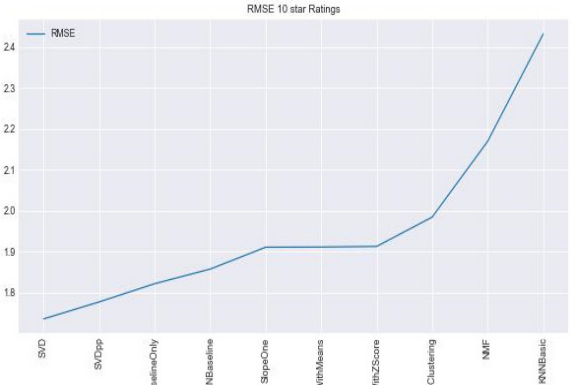
- **Baseline only:** The algorithm predicts a random rating based on the distribution of the training set, which is assumed to be normal.
- **KNNWithMeans:** A basic collaborative filtering algorithm, considering the mean ratings of each user.
- **SVD:** The famous SVD algorithm, as popularized by Simon Funk during the Netflix Prize. When baselines are not used, this is equivalent to Probabilistic Matrix Factorization
- **SVDpp:** Extension of SVD algorithm.
- **Coclustering:** the algorithm that takes account of similar clusters inside the dataset to calculate and propose a recommendation

MODEL CROSS-VALIDATION

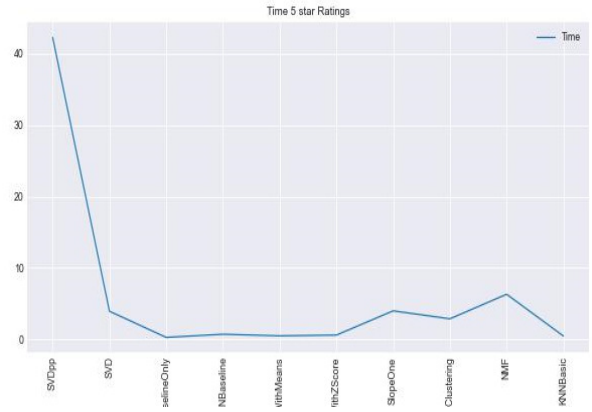
To perform the cross-validation, we benched all the algorithms in a for loop so we can have all the results in the same data frame to identify the better results. It is important to mention that we did this process twice, the first one for the ratings from 1 to 10 and the second one for the ratings from 1 to 5.

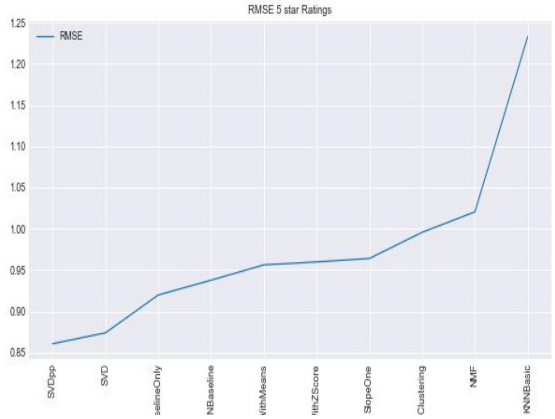
The results of 1 to 10





The results of 1 to 5





After obtaining the RMSE, MAE, FIT_TEST, and TEST_TIME we clearly see that the best performance is made in the population segmented in ratings from 1 to 5. That being said, we will base our prediction calculations on our models, using this dataset and the SVD++, SVD, BaselineOnly, and KNNBaseline algorithms.

Content-based Recommendation

For the content-based system, we aggregate the data of tag value based on the granularity of artistID and after that, we applied NLP and Factorization techniques (like deleting stop words and tokenization and stemming) to have our TFIDF matrix, then we applied the content base model on the data and matrix and the RMSE of our content base model was 0.968, and F score is 0.223.

Hybrid Model

There are many kinds of hybrid model applications based on the algorithms that we can use and the steps that we choose to apply each algorithm and the way that we choose to combine the results of algorithms. Here we used weighted and mixed approaches.

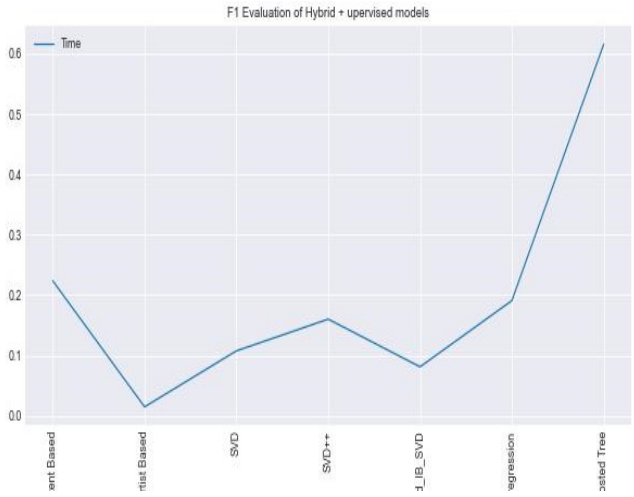
On weighted approach, we combine the results of our 4 different models which were content base, item (artist) base, SVD, and SVDpp, and the best RMSE belonged to the combination of SVDpp and content base that was equal to 0.88 then we can see the significance of the contribution of this approach to the final results.

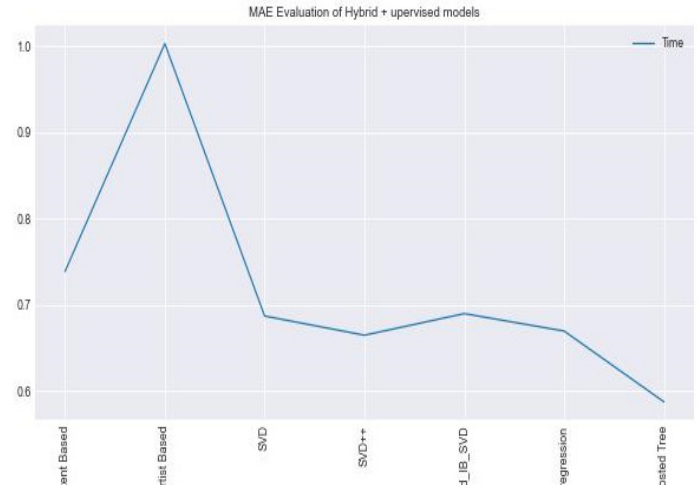
In the second step, we used a mixed approach, and we did a combination of other machine learning algorithms with collaborative filtering algorithms. We select Gradient boosted tree and linear regression and the RMSE of each one consecutively was 0.977 and 0.86.

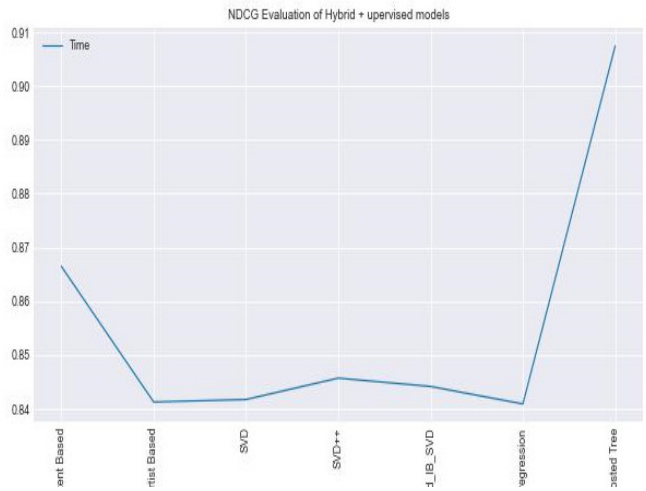
Models Evaluation

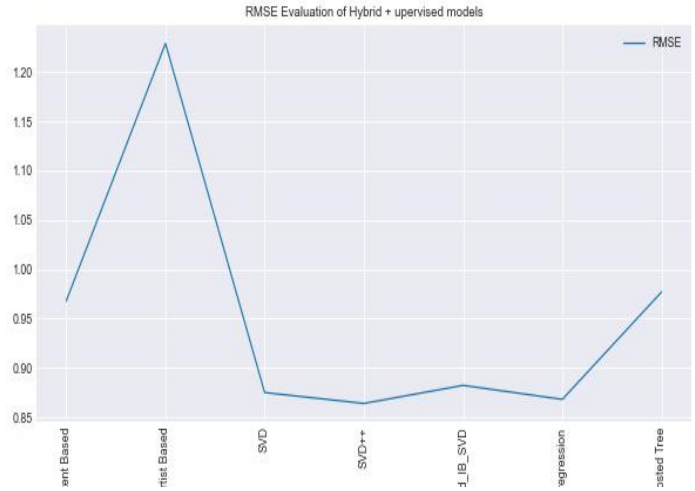
Based on all the above we select these models and evaluate them based on different scores.

Below are the plots related to these evaluations. The best performing model is SVDpp with RMSE of 0.86 and the worst one is item (artist) base with RMSE of 1.23.









Analyzing the Results

Here we want to analyze the performance of our models based on the similarity of users' preference, the tag values, and the artist's weights. We want to discuss about which model is performing better to suggest the user the most similar artists and which model is better for having the more variety recommendation list. Here are the steps we took for analyzing:

Similarity and Distance Matrixes

In order of achieving these objectives we create 4 different matrixes as follow:

1. Matrix of cosine similarity of Users
2. Matrix of cosine distance of Users
3. Matrix of cosine similarity of Artists
4. Matrix of cosine distance of Artists
5.

User Selection

We randomly select user number 13 to apply the analysis.

Getting the top 10 Recommendation

We use the function to get the top 10 recommendations for selected models which are **SVD, SVDpp, Baseline and content base** for user number 13. And below you can see the result of this recommendation. Kindly note that the tag values are more than what we are showing in the below tables.

Content Base			Base Line		
	Artist name	Tag values		Artist name	Tag values
1	Britney Spears	pop dance rock alternative seen live...	1	A Day to Remember	post-hardcore hardcore pop punk scream...
2	Christina Aguilera	pop legend talent electronic dance ...	2	Britney Spears	pop dance rock alternative seen live...
3	Demi Lovato	disney pop dance rock female vocalis...	3	Christina Aguilera	pop legend talent electronic dance ...
4	Depeche Mode	electronic industrial new wave "80s" ...	4	Depeche Mode	electronic industrial new wave "80s" ...
5	Katy Perry	pop pop rock alternative rock electro ...	5	Glee Cast	glee musical pop party cover roman...
6	Miley Cyrus	electro pop teen pop disney country po...	6	Lady Gaga	pop electronic dance female vocalist ...
7	Panic! At the Disco	summer <3 love at first listen these g...	7	Rihanna	en live hit nan pop dance rnb electr...
8	Taylor Swift	pop singer-songwriter acoustic country...	8	SHINee	pop kpop k-pop korean heart myfavo...
9	The Beatles	classic rock dance rock "60s" pop b...	9	倖田來未	performer entertainer jpop taboo j-c...
10	Vanessa Hudgens	teen pop disney female vocalists amazi...	10	浜崎あゆみ	pop nan jpop it could be so much better\n...
SVD			**SVDpp**		
1	A Day to Remember	post-hardcore hardcore pop punk scream...	1	Autechre	ambient experimental idm minimal ele...
2	Autechre	ambient experimental idm minimal ele...	2	Britney Spears	pop dance rock alternative seen live...
3	Britney Spears	pop dance rock alternative seen live...	3	John Frusciante	awesome guitar jams indie beautiful fa...
4	Depeche Mode	electronic industrial new wave "80s" ...	4	Lady Gaga	pop electronic dance female vocalist ...
5	John Frusciante	awesome guitar jams indie beautiful fa...	5	Nine Inch Nails	nter nan industrial industrial pop melan...
6	Lady Gaga	pop electronic dance female vocalist ...	6	Rihanna	en live hit nan pop dance rnb electr...
7	Nine Inch Nails	nter nan industrial industrial pop melan...	7	SHINee	pop kpop k-pop korean heart myfavo...
8	SHINee	pop kpop k-pop korean heart myfavo...	8	Savoy	erican norwegian nan
9	倖田來未	performer entertainer jpop taboo j-c..	9	Sergey Lazarev	pop hot <3 russian male vocalist s...
10	浜崎あゆみ	pop nan jpop it could be so much better\n.	10	Teen Angels	p pop-rock spanish nan latin casi ange..

Getting Preferences of the Most Similar Users to User Number 13

Here we used the cosine similarity matrix to get the most similar users to the user number 13, additionally we used the base table and aggregate it by artists to get the most popular artist for each of these users.
Bellow we can see the table of top 10 artists names for the users that have most cosine similarity to the user number 13.

	user 57	user 398	user 417	user 238	user 484	user 1568	user 1780	user 1706	user 593	user 1701
1	Avril Lavigne	Air	Alicia Keys	Goodphellas	Depeche Mode	30 Seconds to Mars	Adam Lambert	Atari Teenage Riot	Avril Lavigne	Built to Spill
2	Black Veil Brides	Björk	Beyoncé	Marilyn Manson	Dolphin	Gusttavo Lima	Ashley Tisdale	BoA	Axel Fernando	Doug Stanhope
3	Bring Me The Horizon	Blur	Black Eyed Peas	Mini k Bros	Marilyn Manson	Jorge & Mateus	Glee Cast	Cansei de Ser Sexy	Camila	Eve 6
4	Escape The Fate	Daft Punk	Britney Spears	Neilos	Muse	Justin Bieber	Ke$ha	Capsule	David Archuleta	Frightened Rabbit
5	Falling In Reverse	Gorillaz	Christina Aguilera	Pink Puffers	Nautilus Pompilius	Ke$ha	Kelis	Daft Punk	Demi Lovato	Jimmy Eat World
6	Gloria	Los Hermanos	Girls Aloud	SUPERiO	Phoenix	Lady Gaga	Miley Cyrus	Dragon Ash	Jonas Brothers	Modest Mouse
7	Lady Gaga	Massive Attack	Jennifer Lopez	Spiral69	Rammstein	Restart	Nicole Scherzinger	Fatboy Slim	Michael Jackson	Our Lady Peace
8	Taylor Swift	Radiohead	Lady Gaga	Squartet	Дельфин	Ricky Martin	Paramore	Rhymester	Paramore	Say Anything
9	The Pretty Reckless	Silverchair	Leona Lewis	Tubax	Линда	Simon Curtis	Pixie Lott	she	Shakira	The Who
10	nevershoutnever!	The Beatles	P!nk	Zu	Мумий Тролль	nevershoutnever!	The Pussycat Dolls	安室奈美恵	Taylor Swift	Third Eye Blind

Based on the table above we highlight in different colors the cells where we see that the content base model has most in common with the results of top 10 popular artists of top 10 similar users to user number 13. The similarity numbers of models are Content-base equal to 7, Baseline equal to 5, SVD equal to 3, and SVDpp equal to 2.
This doesn't mean that our RMSE is wrong, or the content base model is a better model to use. These results are limited to the top 10 of the top 10. If we go deeper into the cosine matrix, we definitely will find other artists as well.
The above chart will help us to analyze the recommendation system and decide which model to use based on the strategies of the company that we will discuss at the end but now we just need to explain the final steps that we applied in our analysis part first.

Getting a similar artist using the cosine similarity matrix

As we want to go deeper, we will now look into similar artists based on our recommendation.

	Britney Spears	Christina Aguilera	Demi Lovato	Depeche Mode	Katy Perry	Miley Cyrus	Panic! At the Disco	Taylor Swift	The Beatles	Vanessa Hudgens
1	De/Vision	Adriana Calcanhotto	Camouflage	Ashlee Simpson	Arash	Behemoth	Funkadelic	Ashlee Simpson	Band of Horses	Britney Spears
2	"Destinys Child"	Azax Syndrom	Cock Robin	Cassie	Blaze	Chino XL	Grabaż i Strachy Na Lachy	Cassie	Clap Your Hands Say Yeah	Bryan Adams
3	Dixie Chicks	Duran Duran	Cut Copy	David Cook	Clock DVA	Duff McKagan	Jacek Kaczmarski	Christina Aguilera	Devendra Banhart	Erykah Badu
4	Kate Miller-Heidke	Enigma	Deacon Blue	Jennifer Lopez	Demi Lovato	Editors	Joy Electric	Corona	Eels	Faith Evans
5	Morcheeba	Finch	Dido	Kat DeLuna	Luca Turilli	Morten Harket/Earth Affair	Kabaret Starszych Panów	David Cook	Elo da Corrente	Limp Bizkit
6	Noisettes	José González	Emperor	Kate Voegele	Psycroptic	Nightrage	Nadja Benaissa	E-Rotic	Flagelo Urbano	Live
7	Sneaky Sound System	Kenny Chesney	Faithless	Katharine McPhee	NaN	Shimon Masato	Ted Nugent	Jennifer Lopez	Modest Mouse	Mariah Carey
8	The Crystal Method	Marc Almond	Gorillaz	Kelly Clarkson	NaN	The Cardigans	The Mighty Boosh	Katy Perry	Sparklehorse	Olafur Arnalds
9	Vanessa Hudgens	Panic! At the Disco	Jean-Michel Jarre	Leona Lewis	NaN	You Me At Six	The Pussycat Dolls	Leona Lewis	Sunset Rubdown	Red Hot Chili Peppers
10	Weird Al Yankovic	Toxic Holocaust	Kylie Minogue	P!nk	NaN	NaN	Uriah Heep	Satanic Warmaster	Thom Yorke	Tamia

The purpose of the above table is to show we can add variety to our recommendation system by combining the recommended artists and the most similar artist that has a high frequency in the artist matrix. We can create these matrixes using the cosine similarity or the method provided in the content base recommendation class. But since not all the algorithms have this method, here we did it manually and below is the result using the method.

Getting the similar artist using the content base get similar item method

We just add this table to have a better comparison between these two methods.

	Britney Spears	Christina Aguilera	Demi Lovato	Depeche Mode	Katy Perry	Miley Cyrus	Panic! At the Disco	Taylor Swift	The Beatles	Vanessa Hudgens
1	Ashlee Simpson	Arctic Monkeys	Britney Spears	Carrie Underwood	Ashley Tisdale	Britney Spears	Carrie Underwood	Ashley Tisdale	Lady Gaga	Christina Aguilera
2	Ashley Tisdale	Ashley Tisdale	Duran Duran	Demi Lovato	Britney Spears	Christina Aguilera	Christina Aguilera	Britney Spears	MGMT	Demi Lovato
3	Demi Lovato	Britney Spears	Gary Numan	Kylie Minogue	Christina Aguilera	Demi Lovato	Jonas Brothers	Demi Lovato	Paramore	Katy Perry
4	Lady Gaga	Kaiser Chiefs	INXS	Lady Gaga	Jonas Brothers	Jennette McCurdy	Katy Perry	Jonas Brothers	Paul McCartney	Kylie Minogue
5	Lindsay Lohan	Katy Perry	Martin L. Gore	Madonna	Katy Perry	Katy Perry	Lady Gaga	Lady Gaga	Queen	Lady Gaga
6	Maroon 5	Lady Gaga	Mesh	Mariah Carey	Lady Gaga	Lady Gaga	Lily Allen	Miley Cyrus	Red Hot Chili Peppers	Madonna
7	Miley Cyrus	MGMT	New Order	Miley Cyrus	Miley Cyrus	Lindsay Lohan	Miley Cyrus	Panic! At the Disco	Simple Plan	Michael Jackson
8	Prima J	Maroon 5	Simple Minds	Shakira	Paramore	Taylor Swift	Robert Pattinson	Simple Plan	The Smiths	Miley Cyrus
9	Samantha Mumba	Simple Plan	Split Enz	Taylor Swift	Vanessa Hudgens	Vanessa Hudgens	Shakira	Taylor Swift	The Turtles	Shakira

As we can see the results of the content base model method suggest artists with the most similarity to the calculation, we did use cosine matrix of the artist. This is because the cosine similarity is just based on the ratings but when we add tag values, we will get more similar artists as you can see the number of colored cells in the last table is much more than the previous one.

Analyzing the Performance of the Models

As we can see there are different variables that we can discuss for analyzing the performance of the models, if we go with RMSE the best performance belongs to SVD but if we want to have better models, we can mix the Content-base and SVD model and use the ML algorithms like regression to predict better and create boosted recommendation system.

Model Selection and Corporate Strategy

Based on the company's strategy we can propose the recommendation system to have more similar results or suggest more variety artists to the users, here are our suggestions:

1. The company can use a Content-base model if they want to recommend to their users the most similar artists. This is very good if the company's users are risk averter(risk averter users have some specifications like their age is greater than 30 …)
2. The company can use SVDpp to build their recommendation system, in this way they can be sure that their model is highly accurate, and the recommendations have diversity and the users can find new artists that they may like. This strategy is good for the company if they want to convince more artists to come to their platform since they can say that their recommendation system will enable new artists and albums to introduce more to users, especially if it is in a special genre or category.
3. The third but not last strategy that the company can take is that they can use the combination of 2 models and apply them:
    a. To two different segmentation of the customers and using A/B test to see how these two selected models will perform in reality.
    b. Or using 4 different combinations of this recommendation system based on the two types of customers (risk averter or not) and two types of artists(novel and new or not)

Suggestions

Our suggestions to improve the recommendation system are:

- The NLP we did in the model can be more advance if the company can create a dictionary for analyzing the words the results of content base approach can improve.
- There are so many other data that the company can inject into the tables which will help to the improvement of the model. We suggest that the company also add some other variables that can help to know the customers' behavior more accurately.
- The system of weights is not a good way of rating and there are better systems that can help the company to have better ratings, and this will boost the recommendation system.

Conclusions

We can say that based on the time and data that we had the above results are a good sample for getting to know the company's customers and choosing a recommendation system. However, if we had more time we can go deeper and analyze the cosine similarity matrix in deep. We can also have some pipelines of data to aggregate the similarity matrixes after model prediction. The above results are the thing that we did step by step and if a company was satisfied with our results we can provide these steps as a pipeline to make the analysis part easier and deeper.

Reference

1. http://surpriselib.com/
2. https://gist.github.com/susanli2016/e0cdcf1bca69a2b144fd8c04f30b522f
3. https://analyticsindiamag.com/singular-value-decomposition-svd-application-recommender-system/
4. https://buomsoo-kim.github.io/recommender%20systems/2020/10/22/Recommender-systems-collab-filtering-14.md/
5. https://pypi.org/project/scikit-surprise/