**CS703 5.0 Modeling Evaluation**

*Madeline Lin*

**Task 5.1: Evaluating Results**

**Deliverable 1: Result Assessment**

As I’ve mentioned in the 4.0 Data Modeling assignment, I would like to reserve the step of manual inspection in the 5.0 Modeling Evaluation assignment.

* Content-based filtering method
* For content-based filtering method, I will fit the optimal prediction model to the global weekly top 100 songs dataset and add one new column pred\_msPlayed. I will choose the top 20 songs with the largest predicted streaming time and listen to them and determine how many songs I want to add to my library.
* The optimal prediction model is the simple linear regression model.

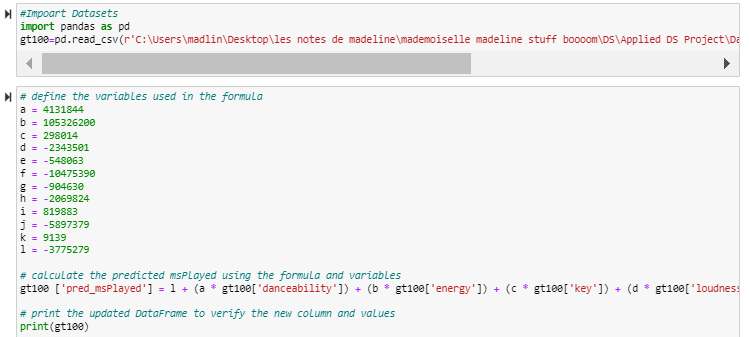
*y = -3,775,279 + 4,131,844 \* danceability + 105,326,200 \* energy + 298,014 \* key + -2,343,501 \* loudness -*

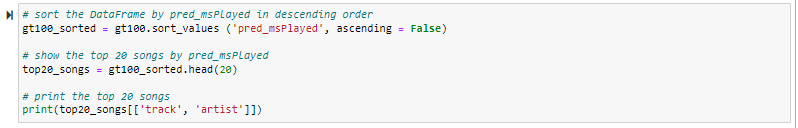
*548,063 \* mode -10,475,390 \* speechiness - 904,630 \* acousticness -2,069,824 \* instrumentalness +*

*819,883 \* liveness + -5,897,379 \* valence + 9,139 \* tempo + e*

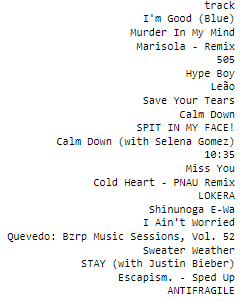
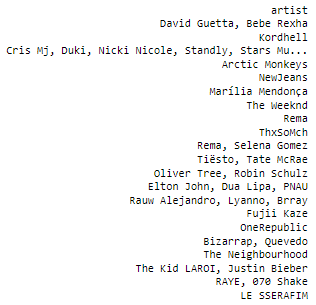
I used the above formula and calculated the pred\_msPlayed for each track. Then, I sorted in descending order and showed the top 20 songs. Those songs can be regarded as my potential liked songs.

*Code*





*Output*

\*I added an additional song because *Calm Down* appeared twice.

* Based on the simple linear regression model, the 20 songs shown above are identified to be my potential liked songs. To determine whether each song is a good fit for my music library, I listened to each one, rated it on a scale from 1 to 10 (where 1 means I do not like it at all, 5 means I am okay with it, 10 means I really like it), and indicated whether I would like to add it to my library.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Track | Artist | Scale | Added to my library |
| 1 | I’m Good (Blue) | David Guetta, Bebe Rexha | 6 | No |
| 2 | Murder In My Mind | Kordhell | 9 | Yes |
| 3 | Marisola – Remix | Cris Mj, Duki, Nicki Nicole, etc. | 4 | No |
| 4 | 505 | Arctic Monkeys | 3 | No |
| 5 | Hype Boy | NewJeans | 3 | No |
| 6 | Leao | Marilia Mendonca | 1 | No |
| 7 | Save Your Tears | The Weekend | 5 | No |
| 8 | Calm Down | Rema | 3 | No |
| 9 | SPIT IN MY FACE! | ThxSoMch | 3 | No |
| 10 | Calm Down (with Selena Gomez) | Rema, Selena Gomez | N/A | N/A (duplicate as #8) |
| 11 | 10:35 | Tiesto, Tate McRae | 8 | Already in my library |
| 12 | Miss You | Oliver Tree, Robin Schulz | 7 | Yes |
| 13 | Cold Heart – PNAU Remix | Elton John, Dua Lipa, RNAU | 10 | Yes! |
| 14 | LOKERA | Rauw Alejandro, Lyanno, Brray | 6 | No |
| 15 | Shinunoga E-Wa | Fujii Kaze | 4 | No |
| 16 | I Ain’t Worried | OneRepublic | 4 | No |
| 17 | Quevedo: Bzrp Music Sessions, Vol. 52 | Bizarrap, Quevedo | 6 | No |
| 18 | Sweater Weather | The Neighbourhood | 2 | No |
| 19 | Stay | The Kid LAROI, Justin Bieber | 9 | Already in my library |
| 20 | Escapism. – Sped Up | RAYE, 070 Shake | 7 | Yes |
| 21 | ANTIFRAGILE | LE SSERAFIM | 5 | No |

I determined a total of six out of 20 songs are my liked songs. Among these six songs, two are already in my library, four more songs are newly added in my library.

I have a high standard when adding songs to my library. As shown in the scale column, I set a threshold of 7 for determining whether a song is worthy of being added to my library. Only songs that I rate 7 or higher are added. I gave 10 to *Cold Heart*, because I really enjoy listening to this song and it indeed delights my day. I gave 9 to *Murder In My Mind*. It is great but not quite on the same level as *Cold Heart*. I gave 7 to *Miss You* and *Escapism. – Sped Up*. I think they are decent satisfactory and can pass my threshold. I gave 6 to *I’m Good (Blue)* and *Quevedo*. I think they did not meet the criteria for being added to my library. While they may be enjoyable in small listening doses, I do not have a strong desire to listen to them frequently. For the rest of the songs, I honestly do not quite enjoy that much, so I gave the ratings 5 or below. I am glad that out of 20 songs selected by the model, two of them were already existed in my library. This is a strong evidence that the model is pretty effective and accurate in capturing my musical preference.

Additionally, while I was listening to the songs, I noticed that the majority of the songs selected by the model were from EDM genre, known for their high energy and rhythm. Furthermore, all the songs were American pop music, with only one K-pop song appearing as the final entry. While Global Weekly Top 100 songs from Spotify Chart proved to be a good dataset for testing my model, but there might be some limitations which I will address in the Reviewing the Process - Process Evaluation Report part.

* Overall, in my opinion, a recommendation system model with 30% accuracy is quite satisfactory. Based on my personal experience, out of 30 songs that Spotify suggests in its Discover Weekly section every Monday, I am only able to add a maximum of three songs to my library. Therefore, I think that the prediction model that I built has a better performance than Spotify’s current algorithm.
* Collaborative filtering method
* For collaborative filtering method, as discussed above, I will listen to those 10 songs which were attained based on two similarity metrics and determine how many songs I want to add to my library.

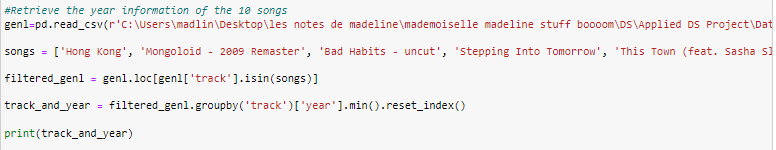
|  |  |  |
| --- | --- | --- |
|  | Euclidean Distance | Cosine Similarity |
| *The 5 most similar tracks to Soft* | | |
| 1 | Hong Kong | Citgo |
| 2 | Mongoloid – 2009 Remaster | Hello, Dolly! |
| 3 | Bad Habits – uncut | Love, Life and Money |
| 4 | Stepping Into Tomorrow | Tadow |
| 5 | This Town (feat. Sasha Sloan) | Good Vibrations – Remastered |

* Similarly, I scaled each song from 1 to 10 and determined whether they should be added to my library in the below table.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Track | Artist | Scale | Added to my library |
| *Euclidean Distance* | | | | |
| 1 | Hong Kong | Gorillaz | 1 | No |
| 2 | Mongoloid – 2009 Remaster | DEVO | 1 | No |
| 3 | Bad Habits – uncut | Maxwell | 5 | No |
| 4 | Stepping Into Tomorrow | Donald Byrd | 1 | No |
| 5 | This Town (feat. Sasha Sloan) | Kygo, Sasha Sloan | 5 | No |
| *Cosine Similarity* | | | | |
| 1 | Citgo | Chief Keef | 2 | No |
| 2 | Hello, Dolly! | Paul Anka | 1 | No |
| 3 | Love, Life And Money | Little Willie John | 1 | No |
| 4 | Tadow | Masego FKJ | 8 | Already in my library |
| 5 | Good Vibrations – Remastered | The Beach Boys | 3 | No |

I felt disappointed that none of the songs met the criteria of being added to my library based on either Euclidean Distance or Cosine Similarity metrics, except that one song *Tadow* which was already in my library. I am unsure why the collaborative filtering method is performing poorly. My initial thought is due to the nature of the generalized dataset. This dataset consists of 170,654 pieces songs stored in Spotify database, spanning from year 1921 to 2020. Given that I was born in 1990s, it is possible that I may not fully appreciate songs from the 1990s or earlier. Therefore, I retrieved the year information for these 10 songs as shown in the below two screenshots. As expected, six out of 10 songs are from earlier decades. Another issue, from my perspective, may be that using a single song as a target and then finding the most similar songs based on two similarity metrics could not provide a comprehensive view of the algorithm’s performance. Hence, I am considering testing the algorithm with multiple target songs to evaluate its robustness and reliability. This approach can provide the algorithm with more comprehensive information of the audio features, which may improve the performance. I will address this approach in the Reviewing the Process – Process Evaluation Report part.

*Code*



*Output*



* The business goals for this project are: 1) explore the behind algorithms of music recommendation systems; and 2) optimize the recommendation system to benefit Spotify users.

From the current standpoint, only the content-based filtering method meets my business goals. With regard to the collaborative filtering method, I will revise and see if it can meet the business goals. As content-based filtering method outperforms Spotify Discover Weekly, I think that I optimized the recommendation system algorithm. By sharing my work on the public platform, I can benefit Spotify users by providing them with access to my code.

**Deliverable 2: Model Approval**

In conclusion, the simple linear regression model worked pretty well. It has 30% accuracy, outperforming the Spotify Discover Weekly. The model meets the project business criteria.

At present, the two similarity metrics unfortunately did not perform as well as the simple linear regression model. However, I am still in the process of making revisions, and the model approval is still pending.

**Task 5.2: Reviewing the Process**

**Deliverable: Process Evaluation Report**

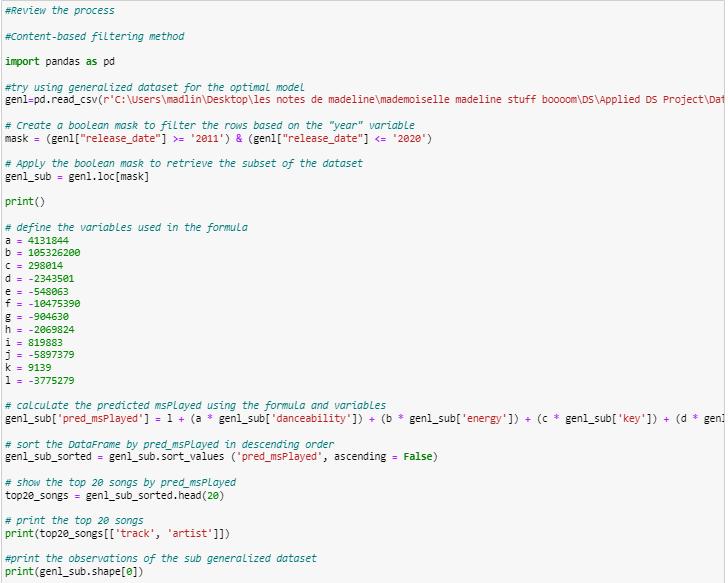
At this stage, it is time to evaluate the process and the steps taken for this project. I have completed the first five phases of the project, and the Modeling Evaluation phase is currently ongoing.

* Let’s Get Started
* This assignment provides an overview of my proposed project for this course. It includes a brief description of the project, the reasons for my choice of the topic, the datasets that I will be working with, and the methods that I plan to use.
* As this is the very preliminary phase, I was able to quickly come up with the idea of working on a cool data project about Spotify, which has always been my passion. I think that nothing was overlooked, the steps were properly executed, sufficient time was given to this assignment, and no issues or problems that could impact the final product be corrected before deployment.
* 1.0 Business Understanding
* This assignment provides a comprehensive overview of my project from the business perspective. It includes a detailed project scope, which encompasses the background/program management, business goals, business success criteria, constraints, and organizational and business impact. Moreover, it covers the assessment of the project’s situation, which includes the inventory of resources, requirement, assumptions and constraints, risks, contingencies, terminology, costs and benefits. Finally, it includes the data mining scope that details the data mining goals, success criteria, and resource plan that details project plan and initial assessment of tools and techniques.
* I remembered that the due date of this assignment is at the end of January, when my workload was particularly heavy. However, I was still able to complete the assignment on time. There was nothing being overlooked, the steps were properly executed, and no issues or problems that could impact the final product be corrected before deployment. However, not sufficient time was given to this assignment. I conducted some research in terms of the Spotify recommendation system algorithms and searched for appropriate datasets for statistical modeling. While I was satisfied with the datasets I found, I wish I could delve deeper into public perceptions of the Spotify recommendation system by reading more research papers.
* 2.0 Data Understanding
* This assignment is about understanding the data for my project. It includes how I gathered the data, the descriptions of the data, explorations of the data, and the verification of the data quality. In my case, I have three datasets that will be used for my project. They are individual dataset, global weekly top 100 songs dataset and generalized dataset. For each dataset, I provided the source and the format. Moreover, I described the fields of the data, including the audio features, general features and other features. Additionally, I summarized a table to show the range or distribution of the quantifiable variables. At last, I checked the data quality of all these three datasets.
* I requested the extension of the due date of this assignment as I was swamped with my work at that time. Thanks to the professor’s approval, I had an extra day to work on it, so I was able to complete it without rushing. I ensured nothing was overlooked, all steps were properly executed, and no issues or problems that could impact the final product be corrected before deployment.
* 3.0 Data Preparation
* This assignment is about preparation of the data for my project. It includes the steps of selecting data, cleaning data, constructing data, integrating data and formatting data. For selecting the data, I elaborated the rationales of the inclusion and exclusion of the datasets. For cleaning data, I listed out the detail and action used to clean my data. For example, I removed unnecessary rows, removed zero value, renamed some columns, etc. For constructing data, I created some new variables, transferred the data across from the datasets, pulled the audio feature data from Spotify Web API. For integrating data, I merged some sub datasets. Finally, for formatting data, I ensured all three datasets were consistent in terms of the attribute naming and data format.
* Data preparation is a very crucial phase in the data analysis project. This process is significantly necessary because the raw data may contain errors, inconsistencies, and missing values that can affect the accuracy and validity of the analysis. Without proper data preparation, the results of the analysis can be inaccurate and misleading. We do not want garbage in, garbage out. Therefore, it is essential to invest sufficient time and effort in the data preparation phase to ensure the accuracy and reliability before the data modeling.
* I dedicated a significant amount of time to this phase. I ensured the datasets were ready for the modeling. As I did not encounter any issues with the data quality during the modeling phase, I am confident that I did not overlook anything, executed all steps properly, and there were no issues or problems that could impact the final project to be corrected before deployment.
* 4.0 Data Modeling
* This assignment is about modeling of the data for my project. It includes selecting modeling techniques, designing tests, building models, and assessing models. For selecting modeling techniques, which was more theory level based, I introduced different types of models including simple linear regression model, linear regression model with interactions, stepwise model, decision tree model, random forest model, and neural network model as well as four similarity metrics, which are Euclidean distance, Pearson Correlation coefficient, Cosine similarity, and Jaccard similarity. For designing tests, I split the data into training, testing and validation sets. For building models, I firstly used the min-max method to scale my audio features data from 0 to 1, and then I conducted the train/validation/test data split. After that, I provided the model descriptions, including the reasons of the inclusions of the variables of each model, as well as how the model should be interpreted in my case. Next, I built the models and documented the rationales of the codes of each model. Finally, for assessing models, in terms of content-based filtering method, I presented a summary table of the performance of each model and interpreted the results of the optimal model (i.e. simple linear regression model) based on the RMSE. However, for the collaborative filtering method, as it is unsupervised learning method, the assessment approach involves manually inspection of the results, and I reserved this evaluation for the next phase. Moreover, for revising parameter settings, I performed some revisions of the linear regression models by adding or removing some interactions. Unfortunately, they did not have a smaller RMSE compared to the simple linear regression model. I also tried the Manhattan distance instead of Euclidean distance by adjusting the p value to one, four out of five songs remain the same except *Stepping Into Tomorrow* was replaced by *When I say I do*.
* Data modeling involves selecting the appropriate modeling techniques, preprocessing the data, building the model, tuning the hyperparameters, and evaluating the performance of the model. It requires a deep understanding of the data and the business problem, as well as expertise in selecting and implementing the appropriate modeling techniques.
* From my perspective, data modeling phase is the core of a data project as it involves creating quantitative models that help in predicting or classifying the target variable. The effectiveness and success of a data project heavily rely on the accuracy and efficiency of the data modeling phase. To ensure that I can have ample time for this phase, I started working on it well in advance and devoted my majority of my time to it. Although I encountered several issues when coding, I was able to resolve them through research. However, one issue that I could not resolve - when I used Pearson correlation coefficient similarity metric to find the similar songs of my target song, there was not enough memory available to perform the request of creating a distance matrix of size (170654, 170654). I attributed this issue to the nature of the dataset and made a decision not to use this metric. Apart from that, I followed the CRISP-PM guide and executed all steps correctly. There were no issues or problems that could impact the final project to be corrected before deployment.
* 5.0 Modeling Evaluation
* This assignment focuses on the post phase of data modeling. When it comes to modeling evaluation, it looks more broadly at which model best meets the business needs. In my data modeling assignment, I’ve addressed that I would reserve the manual inspection for the Modeling Evaluation phase since it is not technically focused.
* For content-based filtering method, I fit the optimal model to the global weekly top 100 songs dataset and selected the top 20 songs which have the largest predicted streaming time. I listened to these 20 songs and determined six songs that I like. For collaborative filtering method, I used two similarity metrics and found 10 songs similar to *Soft*. I listened to these 10 songs and determined only one song that I like. Nothing has been overlooked and all steps have been properly executed so far. Sufficient time was given to this assignment. However, as I mentioned earlier, the content-based filtering model has some limitations, and the two similarity metrics used showed poor performance. To address this, I would like to make the following adjustments to see whether I can make any improvement.
* Content-based filtering method

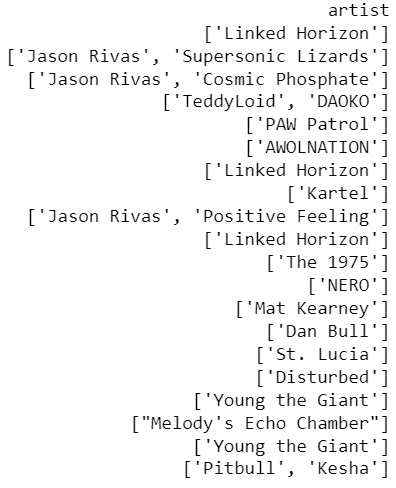
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Track | Artist | Scale | Added to my library |
| 1 | I’m Good (Blue) | David Guetta, Bebe Rexha | 6 | No |
| 2 | Murder In My Mind | Kordhell | 9 | Yes |
| 3 | Marisola – Remix | Cris Mj, Duki, Nicki Nicole, etc. | 4 | No |
| 4 | 505 | Arctic Monkeys | 3 | No |
| 5 | Hype Boy | NewJeans | 3 | No |
| 6 | Leao | Marilia Mendonca | 1 | No |
| 7 | Save Your Tears | The Weekend | 5 | No |
| 8 | Calm Down | Rema | 3 | No |
| 9 | SPIT IN MY FACE! | ThxSoMch | 3 | No |
| 10 | Calm Down (with Selena Gomez) | Rema, Selena Gomez | N/A | N/A (duplicate as #8) |
| 11 | 10:35 | Tiesto, Tate McRae | 8 | Already in my library |
| 12 | Miss You | Oliver Tree, Robin Schulz | 7 | Yes |
| 13 | Cold Heart – PNAU Remix | Elton John, Dua Lipa, RNAU | 10 | Yes! |
| 14 | LOKERA | Rauw Alejandro, Lyanno, Brray | 6 | No |
| 15 | Shinunoga E-Wa | Fujii Kaze | 4 | No |
| 16 | I Ain’t Worried | OneRepublic | 4 | No |
| 17 | Quevedo: Bzrp Music Sessions, Vol. 52 | Bizarrap, Quevedo | 6 | No |
| 18 | Sweater Weather | The Neighbourhood | 2 | No |
| 19 | Stay | The Kid LAROI, Justin Bieber | 9 | Already in my library |
| 20 | Escapism. – Sped Up | RAYE, 070 Shake | 7 | Yes |
| 21 | ANTIFRAGILE | LE SSERAFIM | 5 | No |

* The above are the top 20 songs with the largest streaming time based on the optimal model. While I was trying listening, I noticed the majority of the songs selected by the model were from EDM genre, known for their high energy and rhythm. Furthermore, all the songs were American pop music, with only one K-pop song appearing as the final entry. While Global Weekly Top 100 songs from Spotify Chart proved to be a good dataset for testing my model, but there might be some selection bias if we used this dataset. To elaborate further, I am self-aware that my music preference may not be fully aligned with that of most people. The Global Weekly Top 100 songs primarily includes pop music in English lyrics. My music tastes lean towards genres such as EDM, K-Indie, R&B and Soul music. As it is true that EDM is generally liked by many people, it makes sense that the top 20 songs selected by my model were predominantly from EDM genre. To mitigate this bias, I plan to try the revised generalized dataset. As I’ve mentioned earlier in the collaborative filtering method section, songs from the earlier ages may not align my music taste. To address this, I retrieved a subset of songs from the year 2011 to 2020. I am curious to know if the model can be improved by switching to a dataset with a broader range of genres.

*Code*

­ 

*Output*

Similarly as what I did when I used the Global Weekly Top 100 Songs dataset, I listened to each one, rated it on a scale from 1 to 10, and indicated whether I would like to add it to my library.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Track | Artist | Scale | Added to my library |
| 1 | Jiyuu no Tsubasa | Linked Horizon | 2 | No (J-pop) |
| 2 | Aww Sh\*t – Instrumental Mix | Jason Rivas Supersonic Lizards | 2 | No |
| 3 | Greece 2000 – Extended Mix | Jason Rivas Cosmic Phosphate | 2 | No |
| 4 | Me!Me!Me!, Pt. 1 | TeddyLoid DAOKO | 4 | No (J-pop) |
| 5 | PAW Patrol on a Roll | PAW Patrol | 3 | No |
| 6 | Kill Your Heroes | AWOLNATION | 3 | No |
| 7 | Guren no Yumiya | Linked Horizon | 2 | No (J-pop) |
| 8 | Soda | Kartel | 3 | No |
| 9 | Gita | Jason Rivas Positive Feeling | 6 | No |
| 10 | Shinzo wo Sasageyo! | Linked Horizon | 3 | No (J-pop) |
| 11 | Sex | The 1975 | 5 | No |
| 12 | Promises | NERO | 4 | No |
| 13 | Runaway | Mat Kearney | 3 | No |
| 14 | Boom Boom Boom Boom Boom B… | Dan Bull | 7 | Yes |
| 15 | Elevate | St. Lucia | 3 | No |
| 16 | Hell | Disturbed | 2 | No |
| 17 | I Got | Young the Giant | 3 | No |
| 18 | Some Time Alone, Alone | Melody’s Echo Chamber | 2 | No |
| 19 | My Body | Young the Giant | 2 | No |
| 20 | Timber (feat. Ke$ha) | Pitbull Kesha | 6 | No |

After listening to these 20 songs, I found that they are mostly electronic and rock music, with some diversity in language, including four Japanese songs. Interestingly, the algorithm predicted that I would like Linked Horizon, Jason Rivas, and Young the Giant, as it suggests three songs by Linked Horizon, two songs by Jason Rivas, and two songs by Young the Giant out of 20 songs. Unfortunately, I am not a fan of any of these three artists. Out of the 20 songs, only one song kind of caught my eye. I am not sure why the optimal model did not perform as well as in adjusted generalized dataset compared to Global Weekly Top 100 songs. One potential reason is that generalized dataset is too large, which it may become difficult for me to find meaningful patterns and relationships between variables. Larger datasets often have a lot of noise or irrelevant data that may make it more challenging to identify the patterns. However, as I checked the size of the adjusted generalized dataset, it only includes 440 observations which should not be too large.

I will still use Global Weekly Top 100 songs as the test dataset in this case.

* Collaborative filtering method

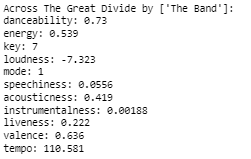
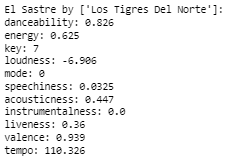
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Track | Artist | Scale | Added to my library |
| *Euclidean Distance* | | | | |
| 1 | Hong Kong | Gorillaz | 1 | No |
| 2 | Mongoloid – 2009 Remaster | DEVO | 1 | No |
| 3 | Bad Habits – uncut | Maxwell | 5 | No |
| 4 | Stepping Into Tomorrow | Donald Byrd | 1 | No |
| 5 | This Town (feat. Sasha Sloan) | Kygo, Sasha Sloan | 5 | No |
| *Cosine Similarity* | | | | |
| 1 | Citgo | Chief Keef | 2 | No |
| 2 | Hello, Dolly! | Paul Anka | 1 | No |
| 3 | Love, Life And Money | Little Willie John | 1 | No |
| 4 | Tadow | Masego FKJ | 8 | Already in my library |
| 5 | Good Vibrations – Remastered | The Beach Boys | 3 | No |

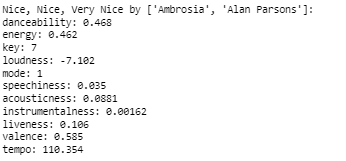
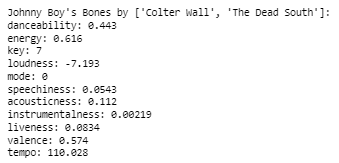
* The above 10 songs were selected as five most similar songs as *Soft* based on two similarity metrics, respectively. As mentioned earlier, I plan to test the algorithm with multiple target songs instead of using one single target song. I used five K-indie songs as the target songs. They are *Soft*, *Polaroid*, *Vacance in September*, *Soothe (feat. Jumbo)*, and *bath*. In the below codes, these five target songs were defined as a list of strings. The individual dataset then was filtered to only include rows corresponding to those target songs. After that, I calculated the average value of audio features and used these values to find the five most similar songs in the generalized dataset using two similarity metrics. Note that I did not use the adjusted generalized dataset as it did not provide better results in the content-based filtering method section.

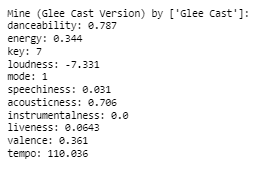
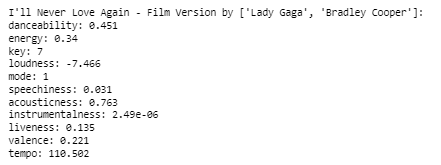
*Code*



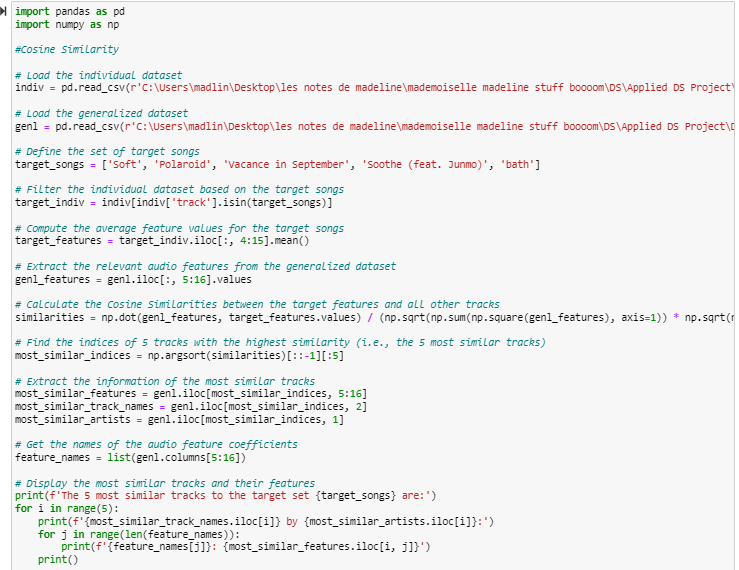
*Output*

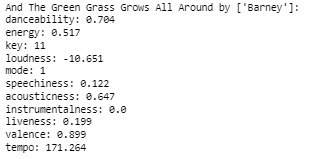
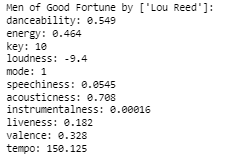


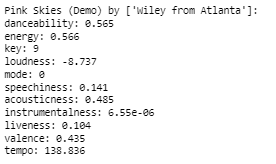
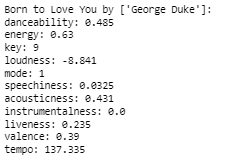
 

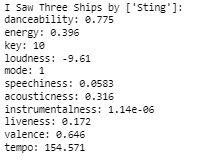
*Code*



*Output*



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Track | Artist | Scale | Added to my library |
| *Euclidean Distance* | | | | |
| 1 | Across The Great Divide | The Band | 3 | No |
| 2 | El Sastre | Los Tigres Del Norte | 4 | No |
| 3 | Johnny Boy’s Bones | Colter Wall The Dead South | 4 | No |
| 4 | Nice, Nice, Very Nice | Ambrosia Alan Parsons | 5 | No |
| 5 | Mine (Glee Cast Version) | Glee Cast | 8 | Yes |
| *Cosine Similarity* | | | | |
| 1 | And The Green Grass Grows All Around | Barney | 1 | No |
| 2 | Men of Good Fortune | Lou Reed | 2 | No |
| 3 | Pink Skies (Demo) | Wiley from Atlanta | 8 | Yes |
| 4 | Born to Love You | George Duke | 6 | No |
| 5 | I Saw Three Ships | Sting | 2 | No |

I was able to find one out of five songs in both Euclidean Distance and Cosine Similarity metrics. Also, there is one song *Born to Love You* which I gave a rating of 6, which almost met my criteria for adding it to my library. Using multiple target songs slightly improved the accuracy of the similarity metric compared to the previous result where only one song matched my taste out of the total 10 recommended songs. By using more than one target song, the algorithm had more information to identify and patterns and similarities in the data, which improved its ability to recommend similar songs.

**Task 5.3: Determining the Next Steps**

**Deliverable 1: Possible Actions**

I attempted to enhance the performance of the models for both content-based filtering method and collaborative filtering method; however, I faced difficulty in improving the results when I switched to the generalized dataset as the test dataset for the content-based filtering method. Fortunately, for the collaborative filtering method, the result improved slightly when I adjusted the size of target songs. As the simple regression model has 30% accuracy in song recommendations, which outperforms the Spotify Discover Weekly, I believe this model is ready for deployment.

There should be certain steps repeated to improve the results, and I have done those in the Reviewing the Process section and tried to improve. While the results did not achieve my expectations, they were also not disappointing. There is no need to undertake a new data-mining project, as I have at least one model that demonstrates excellent performance, despite not all models performing well.

Regarding the possible actions to be taken, given that we will have the deployment plan assignment in the subsequent phase, my first step is to conduct a thorough inspection of my models and the generalized dataset to eliminate any possible errors or mistakes that may have been made during the modeling/evaluation processes. Following that, I will assess whether there is any further room for me to update the models based on the model performance or changes in available data.

**Deliverable 2: Final Decision**

From the current perspective, my current final decision is using the simple linear regression model and the global weekly top 100 songs dataset as the test dataset base for our recommendation system algorithm.

Note I also attached the codes for this assignment in the submission. The three datasets used in the codes are the same as the ones I had previously sent to you.