**CS703 1.0 Business Understanding**

*Madeline Lin*

**Task 1.1: Determine Business Objectives**

**Deliverable: Project Scope Document – Part 1**

**Background/Problem management wants to address:**

Music is an important element of people’s mundane lives. As the largest streaming service provider, Spotify is founded with a group of engineers back in 2006 and Spotify differentiates with other competitors (i.e. Apple Music, Amazon Music, etc.) with its heavy investment on research and development.

One of the charms that Spotify offers is the customization of the application based on the user’s tastes and preferences. The customized features include recommendation system, user’s top mixes or on repeat songs, user’s favorite artists and user’s preferred genres. Among those, recommendation system is an interesting topic to explore further for two reasons. First, listening to new songs that matched people’s taste can certainly delight their days. It is worthwhile to research about this topic to see whether we are able to build a model with stronger performance. Second, compared to user’s top mixes or on repeat songs and other topics, the recommendation system topic requires not only the user’s individual streaming data but the generalized Spotify music whole dataset, which will probably serve better for the purpose of this course as we want to conduct a project with the goal of providing broader and more generalized benefits instead of only focusing on the individuals.

**Business Goals:**

The business goals for this project are to explore: 1) the behind algorithms of music recommendation systems; and 2) how to provide a better music recommendation system.

**Business Success Criteria:**

The business success criteria for this project is to measure the accuracy the recommendation system will be based on the performance of models we’ve built. Details of the quantitative success criteria are elaborated in the “Data-Mining Success Criteria” part.

**Constraints:**

I personally think there will not be much constraints for this project in terms of staffing, technology, timeframes or money.

**Organizational and Business Impact:**

If we are able to provide a better recommendation system, it will bring a positive impact on Spotify. More users will subscribe as they enjoy using this app by having new good songs every week. The increasing subscriptions will increase the revenue of organization, which benefit the shareholders and employees.

**Task 1.2: Assess the Situation**

**Deliverable: Project Scope Document – Part 2**

**Inventory of Resources:**

1) Data

1.1) Individual dataset

1.2) Global Weekly Top 100 Songs

1.3) Generalized dataset\*

\*Previously I found one dataset containing all songs from Spotify with the audio features and popularity from 1921 to 2020 from Kaggle. I was planning to find a more updated dataset (till 2022). There is one dataset (i.e. The Million Playlist Dataset – dataset for music recommendation and automatic music playlist continuation) from Spotify R&D Research page. However, this dataset only has the information of track, artist, album, URI and duration but does not have the audio features of each song. Hence, it would be better for me to utilize the dataset from Kaggle.

2) Hardware

Macbook

3) Software

Python

4) People

Myself – Individual Project

**Requirement, Assumptions, Constraints:**

***Requirement***

A well-organized plan should be drafted in advance for this project. I have consolidated a plan table in the “Project Plan” part.

In terms of the definitions of acceptable finished work, I think finally I will submit a package which includes 1) Python codes 2) Word documentation 3) Slide 4) My thoughts/Tips for this project if applicable. I also want to regard this project as my milestone I’ve completed through this certificate program and put it as an individual piece of demo in my Github page.

***Assumptions***

I assume that there will not be much issue in terms of the legal and security obligations for the public data I’ve pulled and its usage as the global weekly top 100 songs dataset and generalized dataset are used among other data enthusiasts previously.

Meanwhile, it should be assumed that three datasets (i.e. individual dataset, global weekly top 100 songs dataset, generalized dataset) are factually based and there will not be much bias on it. There may be slight adjustments for the data but overall it should be justified.

***Constraints***

I personally think there will not be constraints for this project in terms of staffing, technology, timeframes or money.

**Risks, Contingencies:**

I do not foresee any risks related to incapability of getting data as I have already downloaded the necessary data for this project.

There may be some risks related to running the models. For instance, I may be unable to run the statistical models due to the code error. Or, owing to the limitation of the datasets, there may be some nuances of the results that we are not able to interpret fully. Those risks are quite normal in data science.

My contingency plans for dealing these risks are that: first, being more patient with the debugging process and doing more research to figure out the errors; second, accepting the results which we may not be able to explain due to the uncontrollable factors.

**Terminology:**

There will be a lot of terminologies for this project when it comes to the definitions of different statistical models, the criteria of determining the optimal model, explanations of some Python functions and packages and so forth. I think I will list out these terminologies in the documentation. For now, I can give one example of Decision Tree as below.

Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.

Decision tree learning employs a divide and conquer strategy by conducting a greedy search to identify the optimal split points within a tree. This process of splitting is then repeated in a top-down, recursive manner until all, or the majority of records have been classified under specific class labels.

I used the regression decision tree to solve the numerical data. I used the tree function and put all x variables in the right hand of the equation and y variable in the left hand.

**Costs and Benefits:**

As this is an individual data science project using the public data, fortunately there will be not much cost. There might be some costs with regard to searching materials which are not free, or some costs regarding outsourcing help as needed if I am unable to figure out the issue. Overall, the costs should not be material.

As for the benefits, if we are able to build a good recommendation system, for our own sake we will be able to enjoy better tailored music. Honestly speaking, nowadays I am depressed with the Discover Weekly as I am only able to have one or two or sometimes none of the songs among 30 songs per week. If I can explore new great songs by myself, that would be fabulous. I plan to run a Github page of myself and post my project there. People who are interested in this may be potentially utilize my models to find the songs they love. It is beneficial to the society. Additionally, as I’ve mentioned in the “Organizational and Business Impact” section, if we are able to provide a better recommendation system, more users tend to subscribe as they enjoy using this app by having new good songs every week. The increasing subscriptions will increase the revenue of organization, which benefit the shareholders and employees.

**Task 1.3: Determine the Data-Mining Goals**

**Deliverable: Data Mining Scope Document**

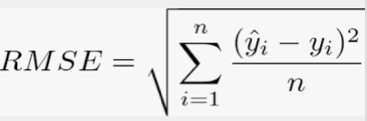
**Data-Mining Goals:**

For the content-based filtering method, the data-mining goal is to build a statistical model with the best performance.

For the collaborative filtering method, the data-mining goal is to find the variable with the least distance of the target variable.

**Data-Mining Success Criteria:**

For the content-based filtering method, the results will be measured by the RMSE (root mean square error).



RMSE is a frequently used measure of the differences between values predicted by a model or an estimator and the values observed. I will use the above formula of the RMSE as my indicator to determine the optimal prediction model.

For the collaborative filtering method, I will apply neighborhood collaborative filtering to use the similarity metrics method. Calculate the distance using all audio features available in the generalized dataset and find the neighbor songs which have relatively less distance. So the results will be measured by the distance of the neighbor songs with a target song.

**Task 1.4: Produce a Project Plan**

**Deliverable: Data Mining Project/Resource Plan**

**Project Plan:**

Based on my quick glance of the CRISP-DM and also the instructions of the following weeks’ tasks, we will have a detailed documentation for each ongoing phase of the plan. Therefore, I only created a table below to show the rough plan of this project.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Tasks** | **Start /End Date** | **Required Resources** | **Inputs** | **Outputs** |
| **Data Preparation**  1) Understand the data  2) Data transformation  3) Data validation  4) Data cleaning | Jan 29  ~Feb 18 | JSON  Excel  Python | Manual observation and evaluation | Data ready for modeling |
| **Data Modeling**  1) Content-based modeling  1.1) Simple regression  1.2) Regression with interactions  1.2) Stepwise regression  1.3) Decision Tree  1.4) Random Forest  *(to be finalized)*  2) Collaborative modeling  2.1) Neighborhood collaborative filtering/ Similarity metrics method | Feb 19  ~Mar 12 | Python | 1) Understand the model  2) Multiple modelling techniques  3) Interpret the results | 1) Content-based modeling  1.1) Produce the optimal model  2) Collaborative modeling  2.1) Find the variable which has the least distance with the target variable |
| Model Evaluation  and Conclusion | Mar 13  ~Apr 1 | Word | 1) Understand the models  2) Ability to evaluate the performance of each model  3) Ability to interpret the model | Analysis and Conclusion |
| Documentation,  Presentation Preparation and Wrap up | Apr 2  ~Apr 7 | Word | 1) Technical skills  2) Writing skills  3) Presentation skills | Report and Slide |

**Initial Assessment of Tools and Techniques**

At this point, I think I have grasped a clear understanding of the tools and techniques that are needed for completing my project. However, this is just a high-level overview of my project and I haven’t put it down to earth yet. If there are any issues which may cause a change in the tools and resources needed, I will find an alternative way and ask for the professor’s approval first.