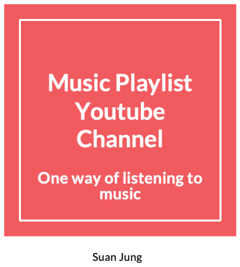
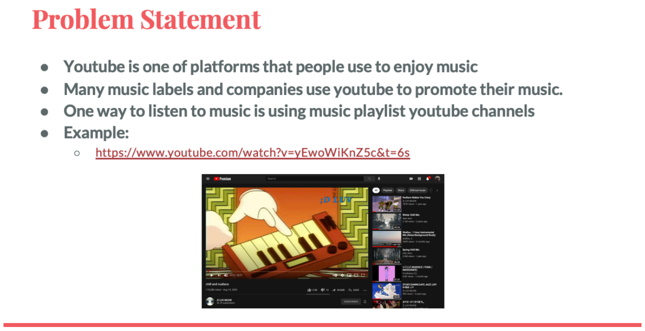
Final Results Presentation



Music Playlist Youtube Channel

One way of listening to music

Suan Jung



Problem Statement

- Youtube is one of platforms that people use to enjoy music
- Many music labels and companies use youtube to promote their music.
- One way to listen to music is using music playlist youtube channels
- Example:
  - https://www.youtube.com/watch?v=yEwoWiKnZ5c&t=6s

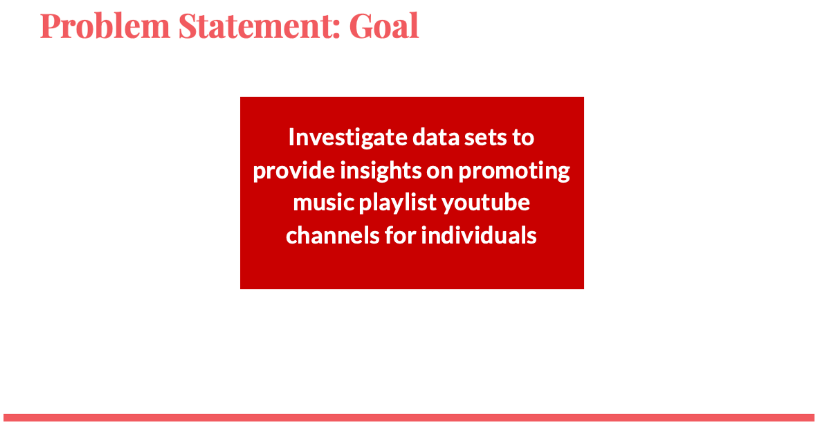
1. Problem Statement

There are different ways that people enjoy music. More and more people listen to music during this unsettling time using Spotify, Youtube, or Twitch.

As a person who loves listening to music, I recently found very interesting trends on youtube. Some people created youtube channels and post music playlists or music mix sets with eye-catching visual graphics suited to the music playlists.

For example https://www.youtube.com/channel/UCTdn10ESvSLuJHShEV4S6qA

Some of these music playlist youtube channels are owned by major music labels or companies to promote their music. However, there are also a lot of youtube channels that individuals own.

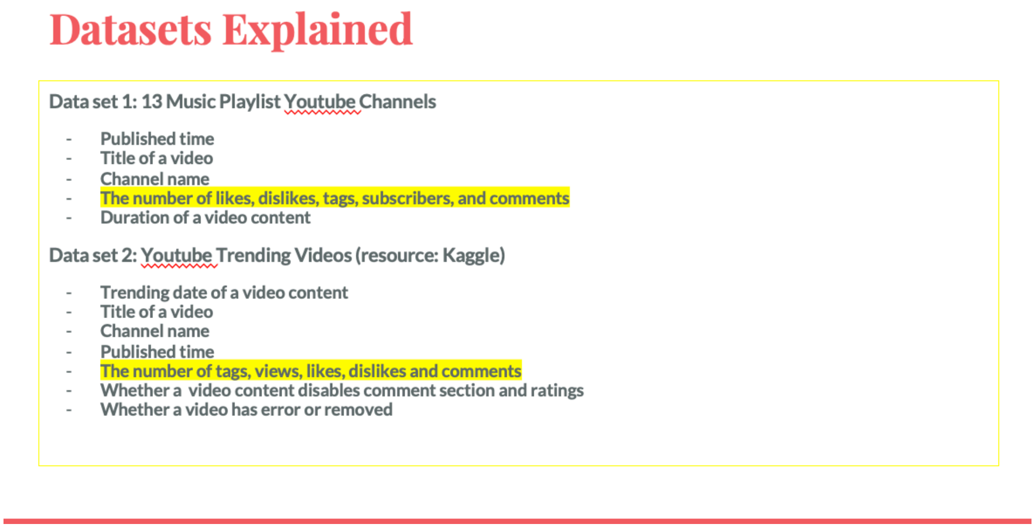


Problem Statement: Goal

Investigate data sets to provide insights on promoting music playlist youtube channels for individuals

The analysis and modeling aims to find a way to provide some insights in promoting music playlist youtube channels to those individuals.

2. Datasets Explained



Datasets Explained

Data set 1: 13 Music Playlist Youtube Channels

- Published time
- Title of a video
- Channel name
- The number of likes, dislikes, tags, subscribers, and comments
- Duration of a video content

Data set 2: Youtube Trending Videos (resource: Kaggle)

- Trending date of a video content
- Title of a video
- Channel name
- Published time
- The number of tags, views, likes, dislikes and comments
- Whether a video content disables comment section and ratings
- Whether a video has error or removed

To investigate, I used two data sets.

One is the data set from specific music playlist youtube channels that have many viewers and subscribers. The other is the youtube trending videos.

The first dataset includes,

- published_ time
- title of a video
- channel_name
- the number of likes, dislikes, tags, subscribers, and comments
- duration of a video content

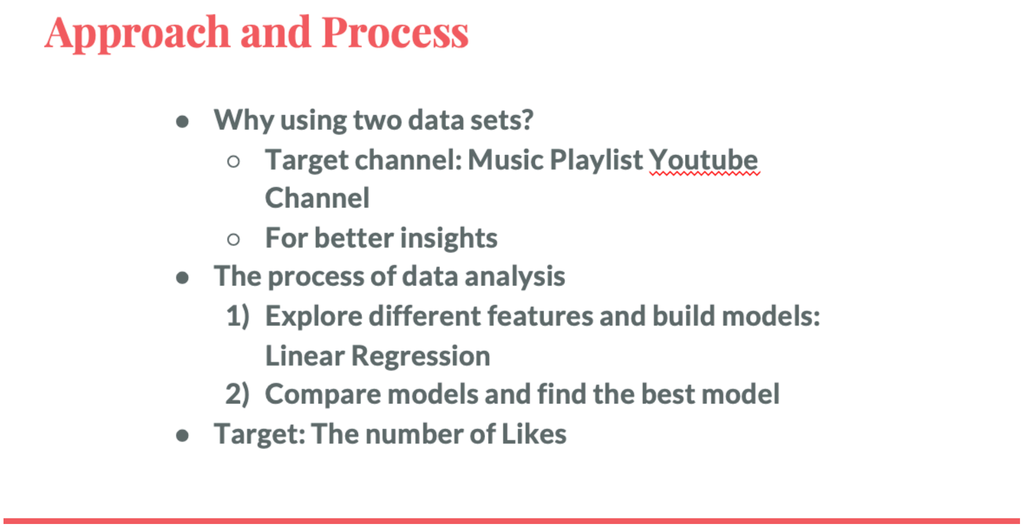
These data are from targeted Youtube channels, and I chose those channels based on their popularity. All music playlist youtube channels have more than 1 k views. Most of those Youtube channels are based in the US and Korea.

The second dataset includes,

- trending date of a video content
- title of a video
- channel name
- published time
- the number of views, likes, dislikes, comments, and tags
- whether the video content disabled comment section and ratings
- whether the video has an error or removed
- descriptions of videos and the ids of video contents, meaning the different ways to identify the channel or the video contents.

These data are from Kaggle, a famous online community of data scientists and machine learning practitioners. Youtube maintains a list of the top trending videos on the platform, and these data are based on that list. Also, the data set only includes data from the US and Canada. But I only included the video data categorized as 10 (aka Music) since my project targeted Youtube videos that are mainly focused on music content.

3. Approach and Process



Approach and Process

- **Why using two data sets?**
  - **Target channel: Music Playlist Youtube Channel**
  - **For better insights**
- **The process of data analysis**
  1) **Explore different features and build models: Linear Regression**
  2) **Compare models and find the best model**
- **Target: The number of Likes**

I decided to compare two data sets instead of one since the first data set only includes music playlist video contents. The second data set can help to understand the insights that I found from the first data set.

The process of analyzing data is separated into two ways.

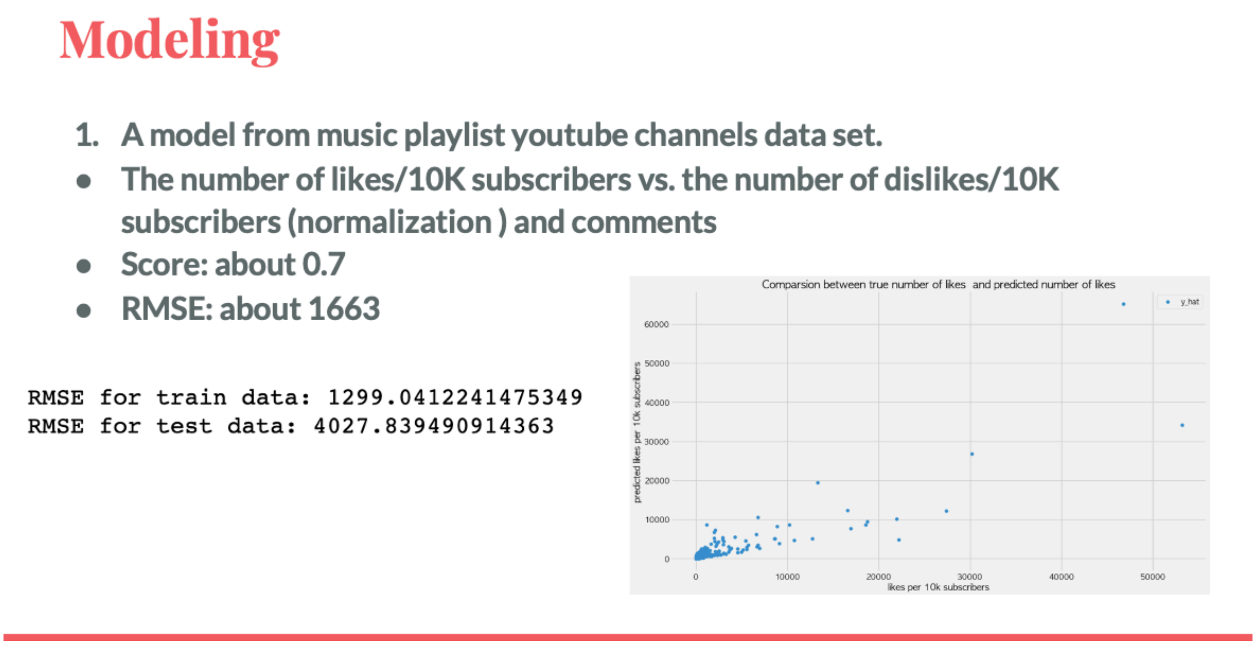
One, explore different features in both datasets and build models.

- As mentioned above, both datasets have different features. Since the goal is to find the insights to promote youtube channels, I decided to use linear regression models. A linear regression model can show how different features can be related to each other. Also, a linear regression is an excellent model to find which feature can be an indicator to promote youtube channels, and it can show the prediction based on the data as well.

Second, compare those models and find the best model that shows the association and the prediction.

- As there are many different features in both datasets, I decided to compare every feature and find the reliable model.
- Initially, I decided to choose the number of likes as an indicator of youtube popularity because youtube counts the number of views regardless of how long a viewer watches a video. However, the number of likes is a good indicator of how much viewers are willing to dedicate to channels.

4. Modeling



Modeling

1. A model from music playlist youtube channels data set.

- The number of likes/10K subscribers vs. the number of dislikes/10K subscribers (normalization) and comments
- Score: about 0.7
- RMSE: about 1663

```
RMSE for train data: 1299.0412241475349
RMSE for test data: 4027.839490914363
```

Comparsion between true number of likes and predicted number of likes
y_hat
predicted likes per 10k subscribers
0
10000
20000
30000
40000
50000
60000
likes per 10k subscribers
0
10000
20000
30000
40000
50000

Three models have the most accuracy on association and prediction after exploring linear regression models with various features.

This model is from the music playlist youtube channels dataset.

This model compares the number of likes per 10,000 subscribers as a target and the number of dislikes per 10,000 subscribers and comments.

Since subscribers' value in this dataset is relatively higher than other feature values, I computed feature values into proportions of 10,000 subscribers.

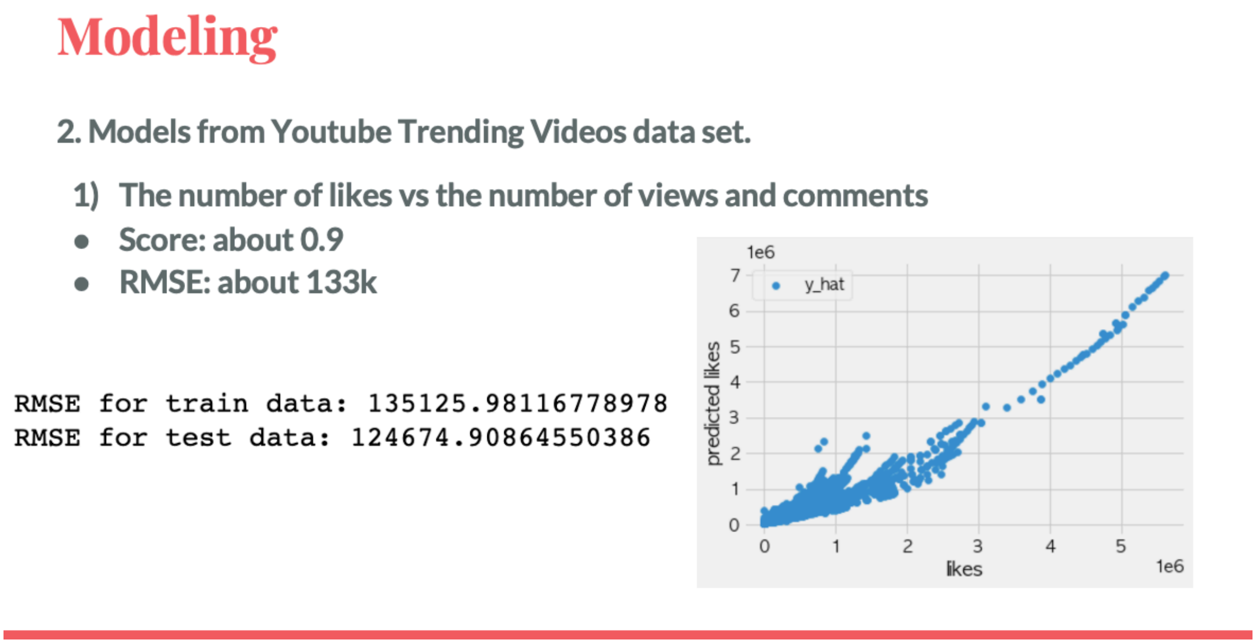
The model has about a 0.7 score in correlation. Features are correlated to each other if the correlation value is close to either 1 or -1. Accodring to the model, the number of dislikes per 10,000 subscribers and the number of comments are highly associated with the number of likes per 10k subscribers.

This model's RMSE value is about 1,663, meaning that there is about 1,663 number difference between the actual number of likes and the predicted number of likes based on the model.

Considering the average number of likes per 10,000 subscribers is about 824, RMSE values are nearly two times larger than the average. Therefore, the difference is not ideal, but it is reliable enough to find the insights.

The scatter plot in the graph demonstrates the difference between the actual likes per 10,000 subscribers from the dataset and the predicted likes per 10,000 subscribers. If the model predicts close to the actual value, the plots show the linear line.

The plots show that the model has a small error in predicting likes per 10,000 subscribers. This graph explains the RMSE values.



Modeling

2. Models from Youtube Trending Videos data set.

1) The number of likes vs the number of views and comments

- Score: about 0.9
- RMSE: about 133k

```
RMSE for train data: 135125.98116778978
RMSE for test data: 124674.90864550386
```

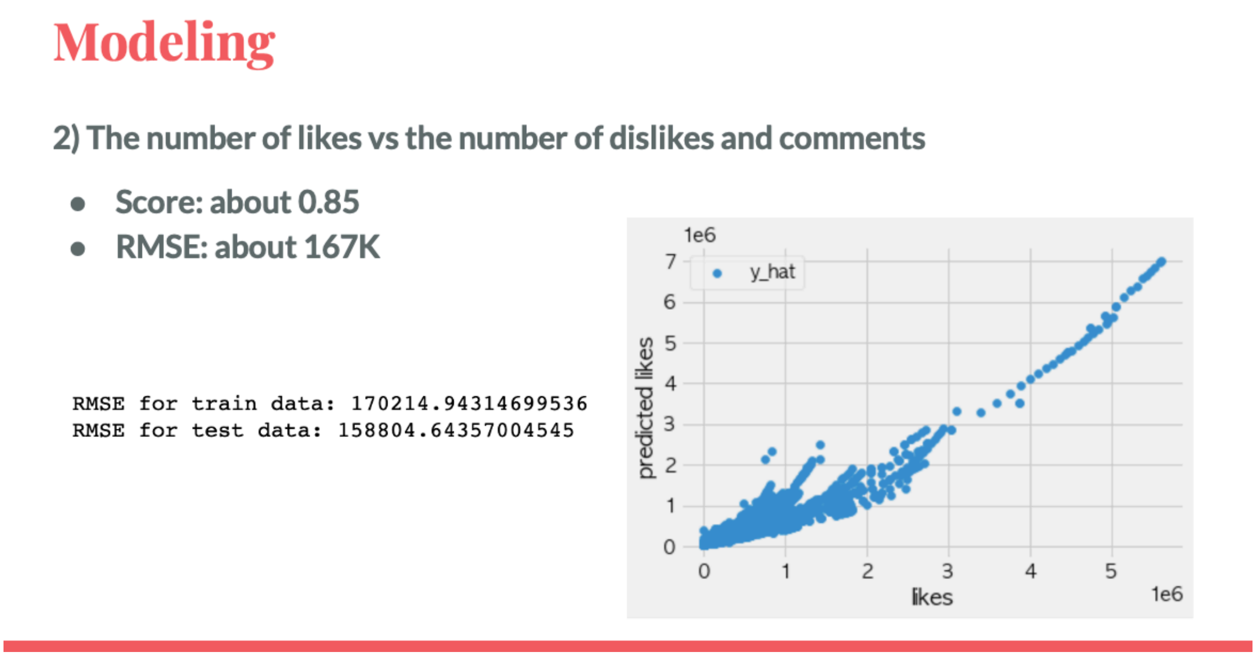
There are different models built using Youtube trending videos dataset. These models are built before normalizing the number of viewers as I did from other models. As a justification, the values of features are comparable, unlike the other data set. Also, the average of each feature in this data set is similar to each other.

The first model compares the number of likes and the number of views and comments. The model has about 0.9 scores in correlation, meaning the number of views and comments are highly associated with the number of likes.

This model's RMSE value is about 133,000, meaning that there is about a 133,000 number difference between the actual number of likes and the predicted number of likes.

Considering the average number of likes is about 194,000, so the difference is small, meaning that the model can show a good prediction on the number of likes.

As the plots show linear line in the graph, I concluded that the model accuracy on predicting likes value is acceptable.



Modeling

2) The number of likes vs the number of dislikes and comments

- Score: about 0.85
- RMSE: about 167K

```
RMSE for train data: 170214.94314699536
RMSE for test data: 158804.64357004545
```

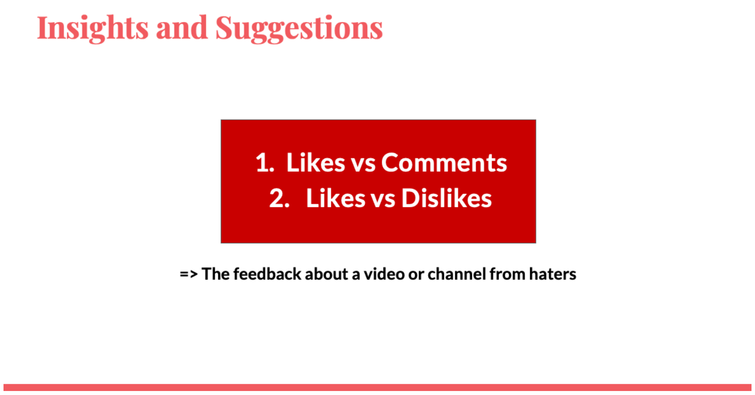
The second model compares the number of likes and the number of dislikes and comments.

The model has about a 0.85 score in correlation, meaning the number of dislikes and comments are highly associated with the number of likes.

This model's RMSE value is about 167,000, meaning that there is a 167,000 likes difference between the actual number of likes and the predicted number of likes.

Therefore, the accuracy error of this model considered being small as well. Also, as the plots show slight linear line in the graph, I concluded that the model accuracy on predicting likes value is acceptable.

5. Insights and Suggestions



Insights and Suggestions

1. Likes vs Comments
2. Likes vs Dislikes

=> The feedback about a video or channel from haters

Based on the analysis and models, I concluded that

- To make viewers like the video content, encouraging people to write comments about video content is essential.
- The number of comments was the most highly associated with the number of likes. This result may seem obvious

because if people like the video, they would be more likely to write comments about it and press the like button on the screen.

- The interesting insight is the association between the number of likes and dislikes. Surprisingly, both are associated with each other even though these features indicate opposite opinions about video contents. In my opinion, to dislike video content, at least people need to be interested in video content to have an opinion. If viewers don't care about the video, likely, viewers are not going to express their opinion by pressing likes or dislikes on the screen. Therefore, disliking a video can be another way to express interest in the video.

My suggestion for YouTubers who own their music playlist youtube channel, it is a good idea to get feedback about a video from the comments, especially the video content with many dislikes. It could be a good way to find what should be improved to promote the channels.