Haneen AlHomoud

Using Linear Regression to Predict the Streams of a Song on Spotify MVP

Abstract

This project was conducted for the T5 Data Science Boot Camp, it aims to give a linear regression model to predict the number of streams of the top 200 songs on Spotify each year from 2016-2021 based on many features the song has.

Data

About Spotify:

Spotify is a digital music, podcast, and video service that gives you access to millions of songs and other content and currently it has around 365 million active users

Data Description:

The First Data set was gathered by Scraping the Spotify Charts website https://spotifycharts.com/regional using BeautifulSoup library, 1200 data point and the main features were gathered from the website including Song name, Artist name, Popularity and Streams

The second dataset was scraped from https://kworb.net/spotify/artists.html it shows the total streams per artist on Spotify, this dataset was used to get the Artist Rank

The third and final dataset gives all Songs/Audio Features Using Spotify API, many different continues features was gathered Including danceability, energy, loudness and many other features

Scope:

- Top 200 songs in the end of each year from 2016- 2021
- 1200 data point

Columns description:

Field Name	Description
Features	
popularity	represents the rank of the song based on the top 200 list
artist_rank	shows the artist rank based on his total streams on Spotify
danceability	how suitable a song is for dancing
energy	how energetic tracks feel fast, loud, and noisy.
Loudness	overall loudness of a track in decibels (dB)
speechiness	detects the presence of spoken words in a track
acoustics	a measure from 0.0 to 1.0 of whether the track is acoustic.
instrumentals	predicts whether a track contains vocal
liveness	presence of an audience in the recording
valence	positiveness conveyed by a track
tempo	estimated tempo of a track in beats per minute (BPM)
duration	the time of song measured in ms
Target	
streams	shows the number of streams for a song

Algorithms

1. Problem Understanding
2. Dataset Exploration (Data Cleansing)
    - Null Values
    - structural errors
    - Outliers
    - Duplicated rows
3. Exploratory Data Analysis (EDA)
4. Feature Engineering
5. Modeling
    - Linear Regression
    - Log Regression
    - Polynomial regression
    - Lasso regression
    - Ridge regression
6. Conclusion

Tools:

- Technologies: Python, Jupyter Notebook
- Libraries: Numby, Pandas, Matplotlib, Seaborn, sqlalchemy, sklearn, BeautifulSoup, requests, spotipy, sklearn,

Communication

- Charts:

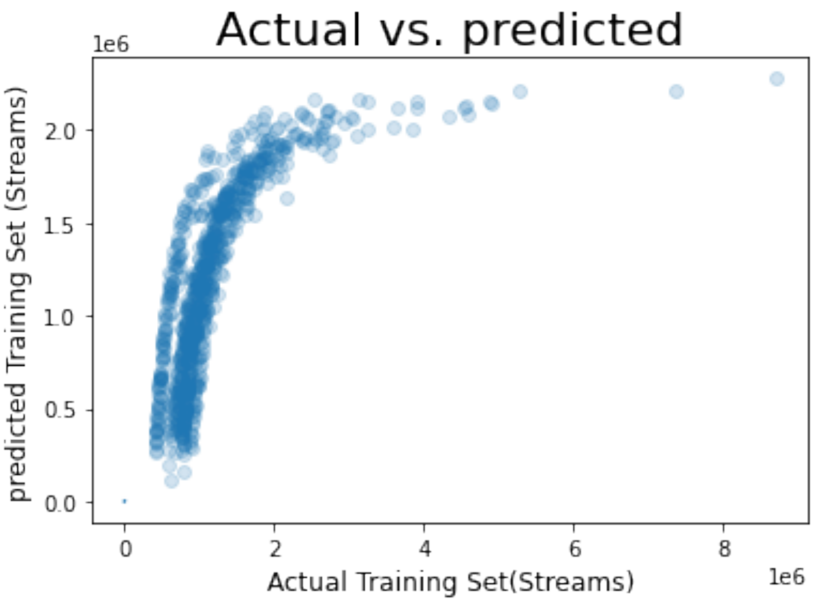


Fig a. Actual Vs Predicted values of the baseline model

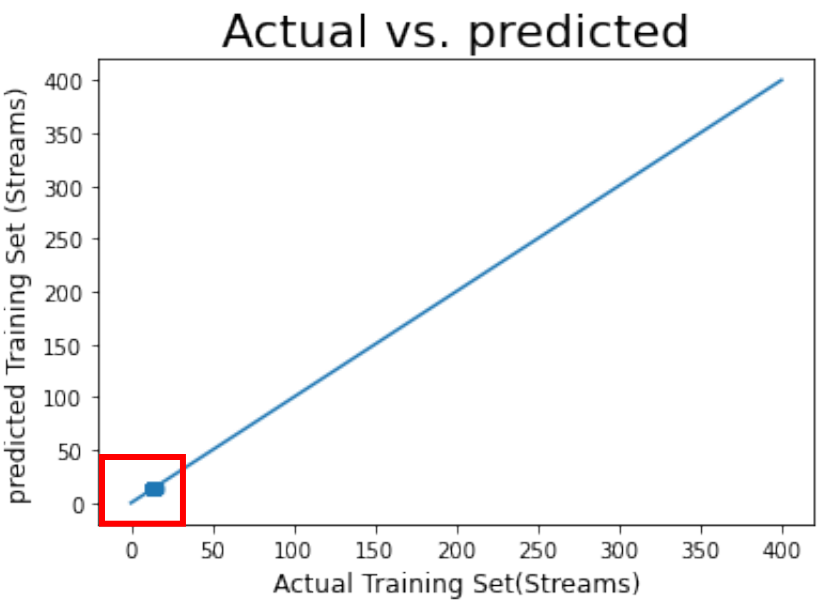


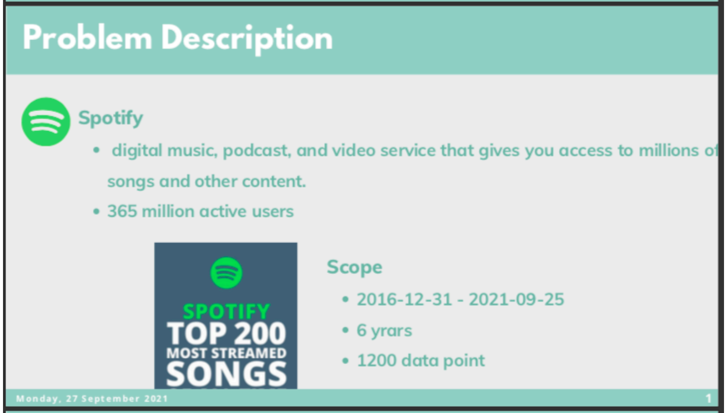
Fig b. Actual Vs Predicted values after improving the model

- **Presentation snips:**

Spotify

Using Linear Regression to Predict the Streams of a Song on Spotify

Prepared by: Haneen Alhomoud



Problem Description

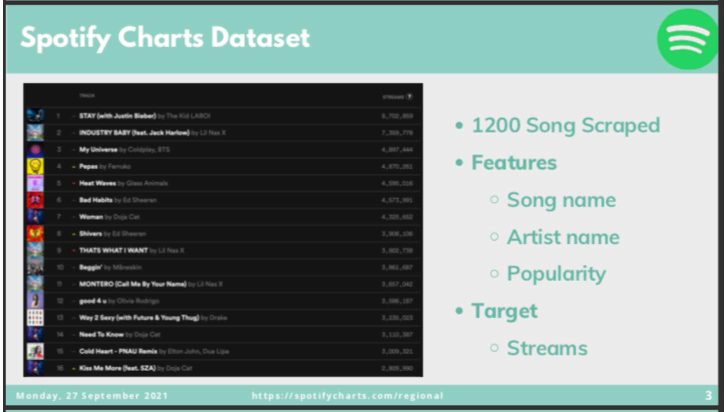
Spotify

- digital music, podcast, and video service that gives you access to millions of songs and other content.
- 365 million active users

Scope

- 2016-12-31 - 2021-09-25
- 6 yrars
- 1200 data point

Monday, 27 September 2021 1

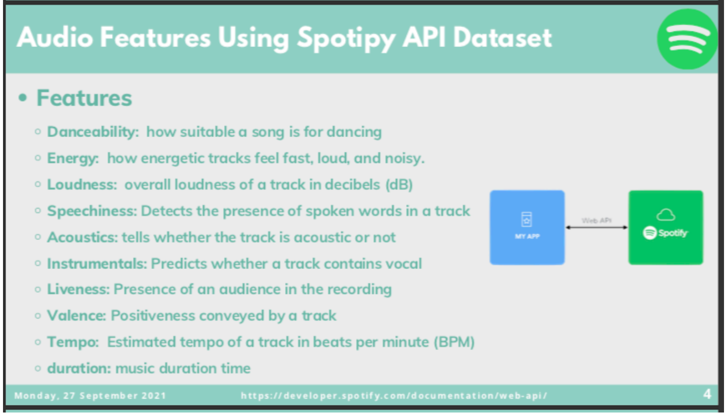


Spotify Charts Dataset

	TRACK	STREAMS
1	**STAY (with Justin Bieber)** by The Kid LAROI	8,702,859
2	**INDUSTRY BABY (feat. Jack Harlow)** by Lil Nas X	7,359,779
3	**My Universe** by Coldplay, BTS	4,897,444
4	**Pepas** by Farruko	4,870,261
5	**Heat Waves** by Glass Animals	4,595,016
6	**Bad Habits** by Ed Sheeran	4,573,991
7	**Woman** by Doja Cat	4,325,652
8	**Shivers** by Ed Sheeran	3,908,106
9	**THATS WHAT I WANT** by Lil Nas X	3,902,738
10	**Beggin'** by Måneskin	3,861,687
11	**MONTERO (Call Me By Your Name)** by Lil Nas X	3,657,042
12	**good 4 u** by Olivia Rodrigo	3,586,187
13	**Way 2 Sexy (with Future & Young Thug)** by Drake	3,235,023
14	**Need To Know** by Doja Cat	3,110,387
15	**Cold Heart - PNAU Remix** by Elton John, Dua Lipa	3,009,321
16	**Kiss Me More (feat. SZA)** by Doja Cat	2,809,990

- 1200 Song Scraped
- Features
  - Song name
  - Artist name
  - Popularity
- Target
  - Streams

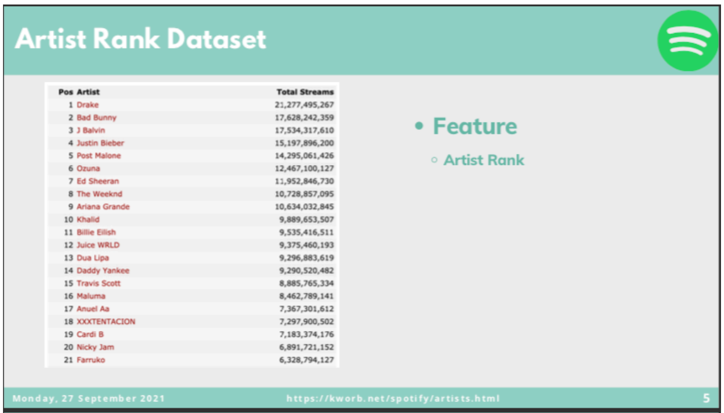
Monday, 27 September 2021 https://spotifycharts.com/regional 3



Audio Features Using Spotipy API Dataset

- Features
  - Danceability: how suitable a song is for dancing
  - Energy: how energetic tracks feel fast, loud, and noisy.
  - Loudness: overall loudness of a track in decibels (dB)
  - Speechiness: Detects the presence of spoken words in a track
  - Acoustics: tells whether the track is acoustic or not
  - Instrumentals: Predicts whether a track contains vocal
  - Liveness: Presence of an audience in the recording
  - Valence: Positiveness conveyed by a track
  - Tempo: Estimated tempo of a track in beats per minute (BPM)
  - duration: music duration time

Monday, 27 September 2021 https://developer.spotify.com/documentation/web-api/ 4



Artist Rank Dataset

Pos	Artist	Total Streams
1	Drake	21,277,495,267
2	Bad Bunny	17,628,242,359
3	J Balvin	17,534,317,610
4	Justin Bieber	15,197,896,200
5	Post Malone	14,295,061,426
6	Ozuna	12,467,100,127
7	Ed Sheeran	11,952,846,730
8	The Weeknd	10,728,857,095
9	Ariana Grande	10,634,032,845
10	Khalid	9,889,653,507
11	Billie Eilish	9,535,416,511
12	Juice WRLD	9,375,460,193
13	Dua Lipa	9,296,883,619
14	Daddy Yankee	9,290,520,482
15	Travis Scott	8,885,765,334
16	Maluma	8,462,789,141
17	Anuel Aa	7,367,301,612
18	XXXTENTACION	7,297,900,502
19	Cardi B	7,183,374,176
20	Nicky Jam	6,891,721,152
21	Farruko	6,328,794,127

- Feature
  - Artist Rank

Monday, 27 September 2021 https://kworb.net/spotify/artists.html 5

Dataset

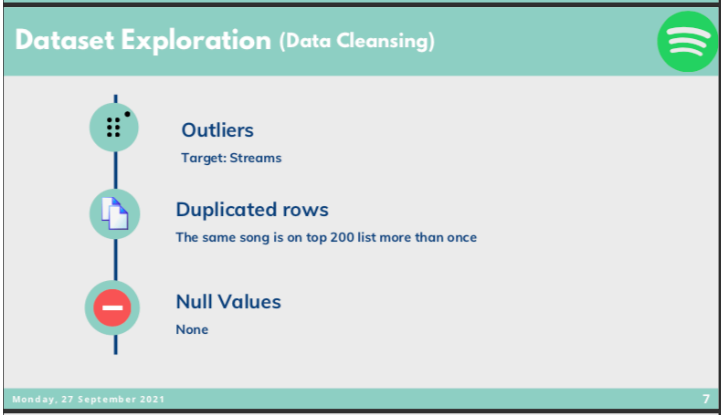
12 Features

- popularity
- artist_rank
- danceability
- energy
- loudness
- speechiness
- acoustics
- instrumentals
- liveness
- valence
- tempo
- duration

Target

- streams

Monday, 27 September 2021 6



Dataset Exploration (Data Cleansing)

Outliers

Target: Streams

Duplicated rows

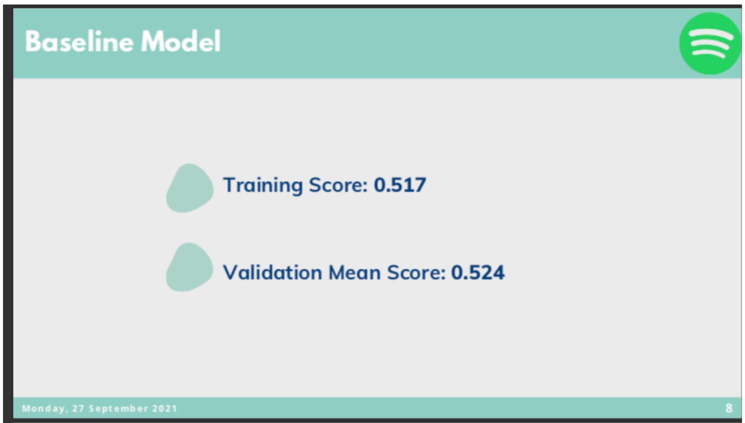
The same song is on top 200 list more than once

Null Values

None

Monday, 27 September 2021 7

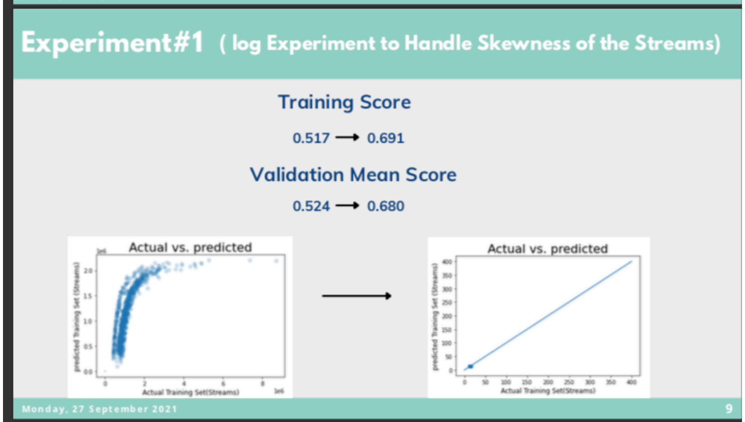
Modeling - Linear Regression



Baseline Model

Training Score: **0.517**

Validation Mean Score: **0.524**



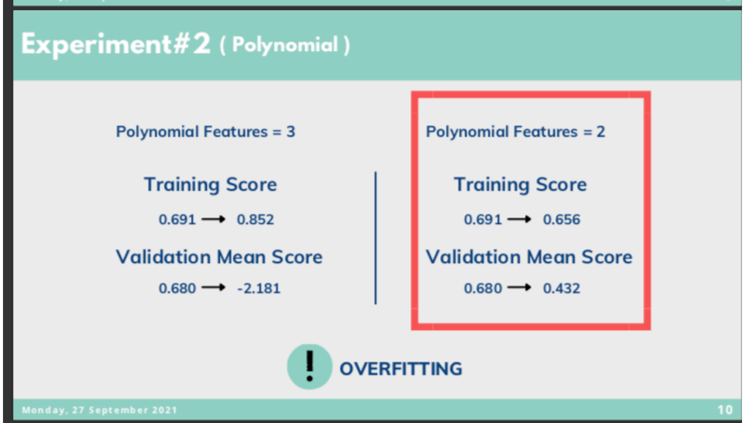
Experiment#1 (log Experiment to Handle Skewness of the Streams)

Training Score

0.517 → 0.691

Validation Mean Score

0.524 → 0.680



Experiment#2 (Polynomial)

Polynomial Features = 3

Training Score

0.691 → 0.852

Validation Mean Score

0.680 → -2.181

Polynomial Features = 2

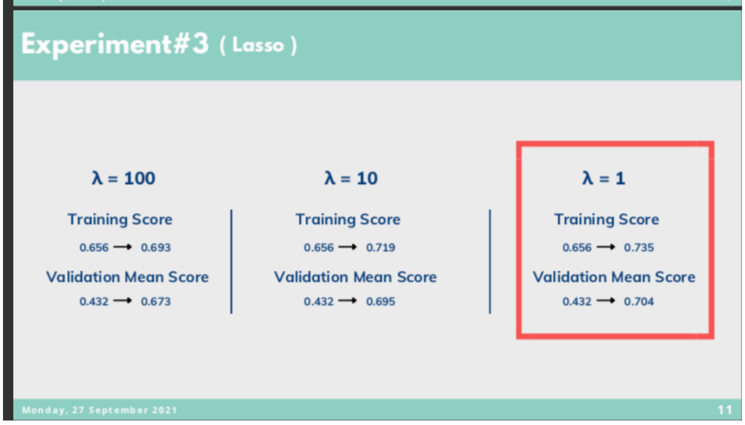
Training Score

0.691 → 0.656

Validation Mean Score

0.680 → 0.432

! OVERFITTING



Experiment#3 (Lasso)

λ = 100

Training Score

0.656 → 0.693

Validation Mean Score

0.432 → 0.673

λ = 10

Training Score

0.656 → 0.719

Validation Mean Score

0.432 → 0.695

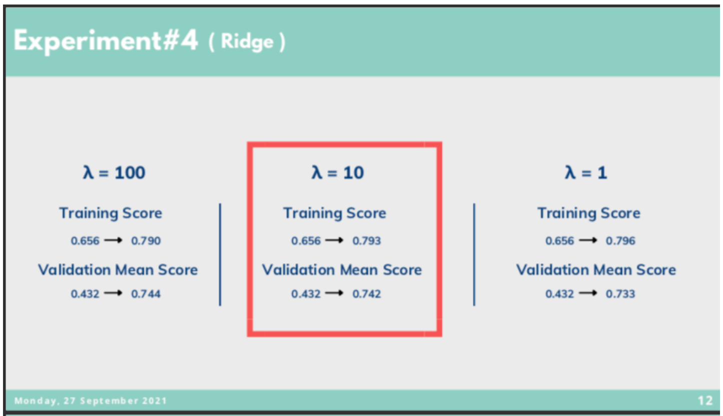
λ = 1

Training Score

0.656 → 0.735

Validation Mean Score

0.432 → 0.704



Experiment#4 (Ridge)

λ = 100

Training Score

0.656 → 0.790

Validation Mean Score

0.432 → 0.744

λ = 10

Training Score

0.656 → 0.793

Validation Mean Score

0.432 → 0.742

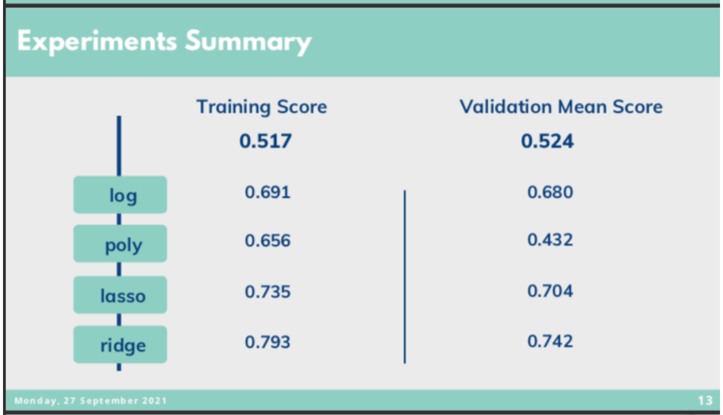
λ = 1

Training Score

0.656 → 0.796

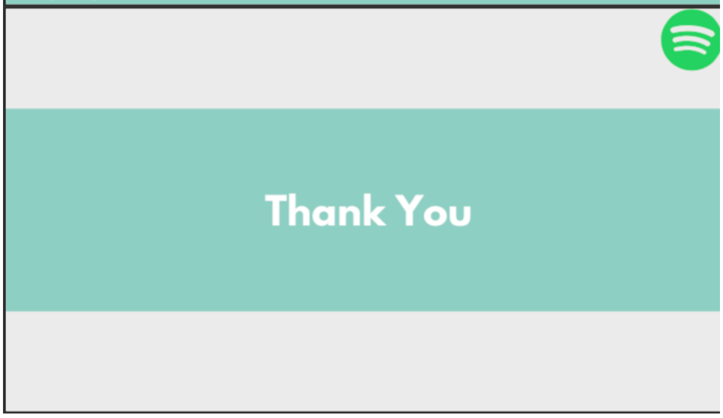
Validation Mean Score

0.432 → 0.733



Experiments Summary

	Training Score	Validation Mean Score
	0.517	**0.524**
log	0.691	0.680
poly	0.656	0.432
lasso	0.735	0.704
ridge	0.793	0.742



Thank You