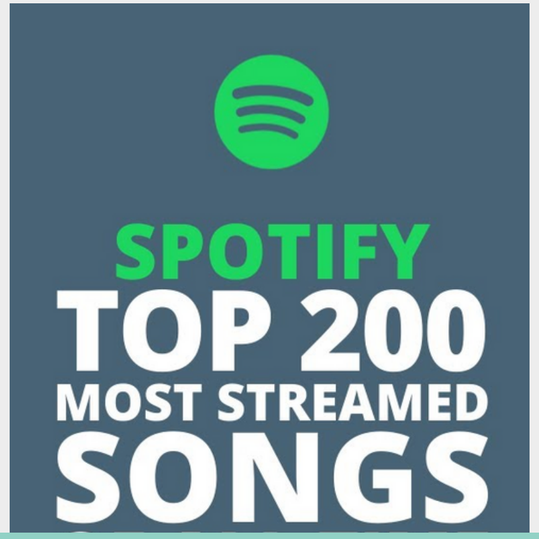
Using Linear Regression to Predict the Streams of a Song on Spotify

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

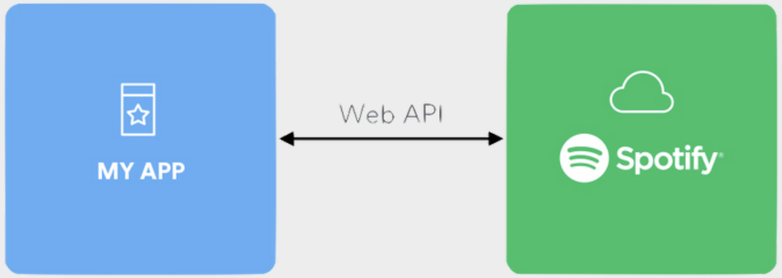
Spotify Charts Dataset

	TRACK	STREAMS
1	**STAY (with Justin Bieber)** by The Kid LAROI	8,702,859
2	**INDUSTRY BABY (feat. Jack Harlow)** by Lil Nas X	7,359,778
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

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



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

Dataset

12 Features

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

Target

- streams

Dataset Exploration (Data Cleansing)

Outliers

Target: Streams

Duplicated rows

The same song is on top 200 list more than once

Null Values

None

Modeling - Linear Regression

Baseline Model

Training Score: **0.517**

Validation Mean Score: **0.524**

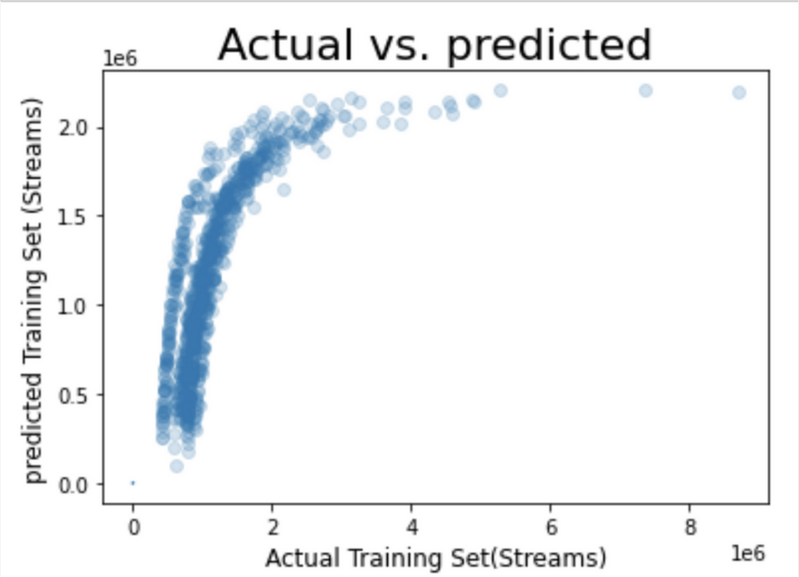
Experiment#1 (log Experiment to Handle Skewness of the Streams)

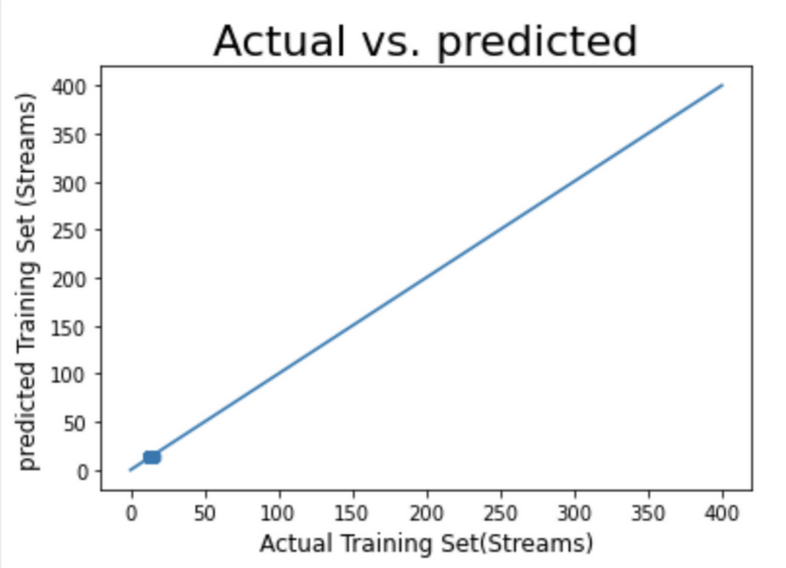
Training Score

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	λ = 10	λ = 1
Training Score	Training Score	Training Score
0.656 → 0.693	0.656 → 0.719	0.656 → 0.735
Validation Mean Score	Validation Mean Score	Validation Mean Score
0.432 → 0.673	0.432 → 0.695	0.432 → 0.704

Experiment#4 (Ridge)

λ = 100	λ = 10	λ = 1
Training Score	Training Score	Training Score
0.656 → 0.790	0.656 → 0.793	0.656 → 0.796
Validation Mean Score	Validation Mean Score	Validation Mean Score
0.432 → 0.744	0.432 → 0.742	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