Predicting The Popularity of Songs on Spotify

Sean Rubin, David Schaffer,Malek Kheirddin
Gerardo Lopez Rodriguez

What’s Our Goal

The purpose of this project is to utilize the Spotify tracks dataset to develop a model that can predict the popularity of a song. This model will assist future artists and musicians by providing insights into what factors contribute to a song's popularity.

Data Table

We first went to look for data that would have the most popular streamed songs on Spotify

We gathered this dataset from Kaggle

https://www.kaggle.com/datasets/maharshipandya/-spotify-tracks-dataset/data

track_id	artists	album_name	track_name	popularity	duration_ms	explicit	danceability	energy	key	loudness	mode	speechiness	acousticness	instrumentalness	liveness	valence	tempo	time_signature	track_genre
3nqQXoyQOWXiESFLl...	Sam Smith;Kim Petras	Unholy (feat. Kim...	Unholy (feat. Kim...	100	156943	False	0.714	0.472	2	-7.375	1	0.0864	0.013	4.51e-06	0.266	0.238	131.121	4	dance
2tTmW7RDtMQtBk7m2...	Bizarrap;Quevedo	Quevedo: Bzrp Mus...	Quevedo: Bzrp Mus...	99	198937	False	0.621	0.782	2	-5.548	1	0.044	0.0125	0.033	0.23	0.55	128.033	4	hip-hop
4uUG5RXrOk84mYEfF...	David Guetta;Bebe...	I'm Good (Blue)	I'm Good (Blue)	98	175238	True	0.561	0.965	7	-3.673	0	0.0343	0.00383	7.07e-06	0.371	0.304	128.04	4	dance
5ww2BF9slyYgNOk37...	Manuel Turizo	La Bachata	La Bachata	98	162637	False	0.835	0.679	7	-5.329	0	0.0364	0.583	1.98e-06	0.218	0.85	124.98	4	latin
6Sq7ltF9Qa7SNFBsV...	Bad Bunny;Chencho...	Un Verano Sin Ti	Me Porto Bonito	97	178567	True	0.911	0.712	1	-5.105	0	0.0817	0.0901	2.68e-05	0.0933	0.425	92.005	4	latin
1IHWl5LamUGEuP4oz...	Bad Bunny	Un Verano Sin Ti	Tití Me Preguntó	97	243716	False	0.65	0.715	5	-5.198	0	0.253	0.0993	0.000291	0.126	0.187	106.672	4	latin
5Eax0qFko2dh7Rl2l...	Bad Bunny	Un Verano Sin Ti	Efecto	96	213061	False	0.801	0.475	7	-8.797	0	0.0516	0.141	1.73e-05	0.0639	0.234	98.047	4	latin
5IgjP7X4th6nMNDh4...	Chris Brown	Indigo (Extended)	Under The Influence	96	184613	True	0.733	0.69	9	-5.529	0	0.0427	0.0635	1.18e-06	0.105	0.31	116.992	4	dance
4h9wh7iOZ0GGn8QVp...	OneRepublic	"I Ain’t Worried ...	I Ain't Worried	96	148485	False	0.704	0.797	0	-5.927	1	0.0475	0.0826	0.000745	0.0546	0.825	139.994	4	piano
4LRPiXqCikLlN15c3...	Harry Styles	As It Was	As It Was	95	167303	False	0.52	0.731	6	-5.338	0	0.0557	0.342	0.00101	0.311	0.662	173.93	4	pop
3k3NWokhRRkEPhCzP...	Bad Bunny;Bomba E...	Un Verano Sin Ti	Ojitos Lindos	95	258298	False	0.647	0.686	3	-5.745	0	0.0413	0.08	1.34e-06	0.528	0.268	79.928	4	latin
6Xom58OOXk2SoU711...	Bad Bunny	Un Verano Sin Ti	Moscow Mule	94	245939	True	0.804	0.674	5	-5.453	0	0.0333	0.294	1.18e-06	0.115	0.292	99.968	4	latin
6xGruZOHLs39ZbVcc...	Joji	Glimpse of Us	Glimpse of Us	94	233456	False	0.44	0.317	8	-9.258	1	0.0531	0.891	4.78e-06	0.141	0.268	169.914	3	pop
1xzi1Jcr7mEi9K2Rf...	Beyoncé	RENAISSANCE	CUFF IT	93	225388	True	0.78	0.689	7	-5.668	1	0.141	0.0368	9.69e-06	0.0698	0.642	115.042	4	dance
2QjOHCTQ1Jl3zawyY...	The Neighbourhood	I Love You.	Sweater Weather	93	240400	False	0.612	0.807	10	-2.81	1	0.0336	0.0495	0.0177	0.101	0.398	124.053	4	alt-rock
3JvKfv6T31zO0ini8...	Tom Odell	Long Way Down (De...	Another Love	93	244360	True	0.445	0.537	4	-8.532	0	0.04	0.695	1.65e-05	0.0944	0.131	122.769	4	chill
31i56LZnwE6uSu3ex...	Bad Bunny	Un Verano Sin Ti	Neverita	93	173119	False	0.876	0.498	10	-7.511	1	0.0478	0.0706	0.0	0.143	0.428	122.016	4	latin
7dSZ6zGTQx66c2GF9...	KAROL G	PROVENZA	PROVENZA	93	210200	False	0.87	0.516	1	-8.006	1	0.0541	0.656	0.00823	0.11	0.53	111.005	4	reggae
4Dvkj6JhhA12EX05f...	Harry Styles	Harry's House	As It Was	92	167303	False	0.52	0.731	6	-5.338	0	0.0557	0.342	0.00101	0.311	0.662	173.93	4	pop
0mBP9X2gPCuapvpZ7...	Charlie Puth;Jung...	Left and Right (F...	Left and Right (F...	92	154486	False	0.881	0.592	2	-4.898	1	0.0324	0.619	1.32e-05	0.0901	0.719	101.058	4	dance

only showing top 20 rows

Cleaning The Data

- Dropped rows with missing values
- Removed duplicate entries
- Lessened the dataset to the 10,000 entries

Narrowing Down Our Search

To get an even more accurate model we decided to look at just the 75th percentile

```
test = final['popularity'].quantile(0.75)
print(test)
```

```
71.0
```

Most Popular Songs

We searched for what the most popular songs were

Track_Name	Popularity
Unholy (feat. Kim...	100
Quevedo: Bzrp Mus...	99
I'm Good (Blue)	98
La Bachata	98
Me Porto Bonito	97
Tití Me Preguntó	97
Efecto	96
Under The Influence	96
I Ain't Worried	96
As It Was	95
Ojitos Lindos	95
Moscow Mule	94
Glimpse of Us	94
CUFF IT	93
Sweater Weather	93
Another Love	93
Neverita	93
PROVENZA	93
As It Was	92
Left and Right (F...	92

only showing top 20 rows

Most Seen Genres

We went to see what genres the most streamed songs

Genre	Genre_Count
dance	239
k-pop	185
alt-rock	147
latino	131
electro	119
indie-pop	112
hip-hop	109
edm	103
pop	93
hard-rock	92
country	68
rock	68
emo	62
folk	57
blues	56
british	55
disco	51
alternative	50
singer-songwriter	48
latin	45

only showing top 20 rows

Create And Train The Model

```
# Define the deep learning model
nn_model = tf.keras.models.Sequential([
    tf.keras.layers.Dense(units=80, activation="relu", input_shape=(X_train_scaled.shape[1],)),
    tf.keras.layers.Dense(units=30, activation="relu"),
    tf.keras.layers.Dense(units=1, activation="sigmoid")
])

# Compile the Sequential model
nn_model.compile(loss="binary_crossentropy", optimizer="adam", metrics=["accuracy"])


# Train the model
history = nn_model.fit(X_train_scaled, y_train, epochs=3, validation_split=0.2, verbose=1)


#
model_loss, model_accuracy = nn_model.evaluate(X_test_scaled, y_test, verbose=2)
print(f"Loss: {model_loss:.4f}, Accuracy: {model_accuracy*100:.2f}%")
```

Find The Average Popularity

Avg_Popularity	track_genre	artists	album_name	duration_ms	explicit	danceability	energy	key	loudness	mode	speechiness	acousticness	instrumentalness	liveness	valence	tempo	time_signature	track_genre
86.0	pop	Taylor Swift	folklore	261922	False	0.532	0.623	5	-9.208	1	0.0331	0.538	7.28e-05	0.0925	0.403	89.937	4	pop

Plug The Data Into The Model

```
my_song = pd.DataFrame({
    "duration_ms": [261922],
    "danceability": [0.532],
    "energy": [0.623],
    "key": [5],
    "loudness": [-9.208],
    "mode": [1],
    "speechiness": [0.0864],
    "acousticness": [0.538],
    "instrumentalness": [7.28e-05],
    "liveness": [0.0925],
    "valence": [0.403],
    "tempo": [89.937],
    "time_signature": [4]
})
```

```
# Apply the same scaling used during training
my_song_scaled = scaler.transform(my_song)
```

```
# Make predictions
predicted_popularity = nn_model.predict(my_song_scaled)

# Print the predicted popularity
print(f"Predicted popularity: {predicted_popularity.item()*100:.1f}%")
```

```
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 12ms/step
Predicted popularity: 16.2%
```

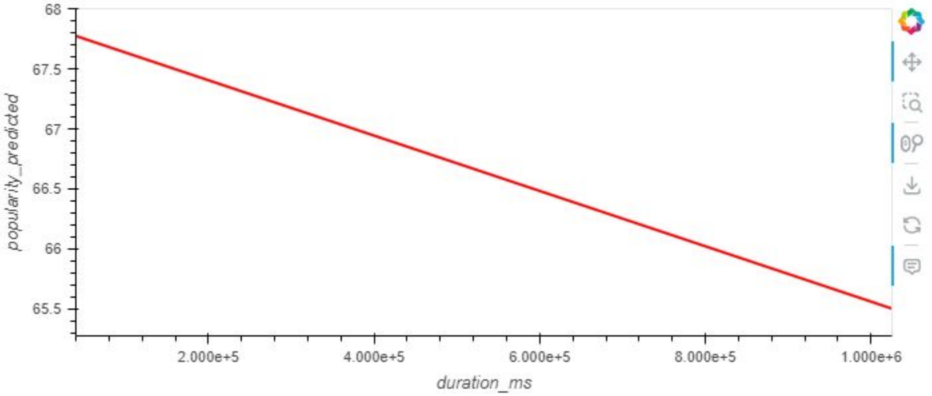
```
# Make y_pred
y_pred = (nn_model.predict(X_test) > 0.5).astype(int).flatten()

# Print classification report
print("\nClassification Report:")
print(classification_report(y_test, y_pred))
```

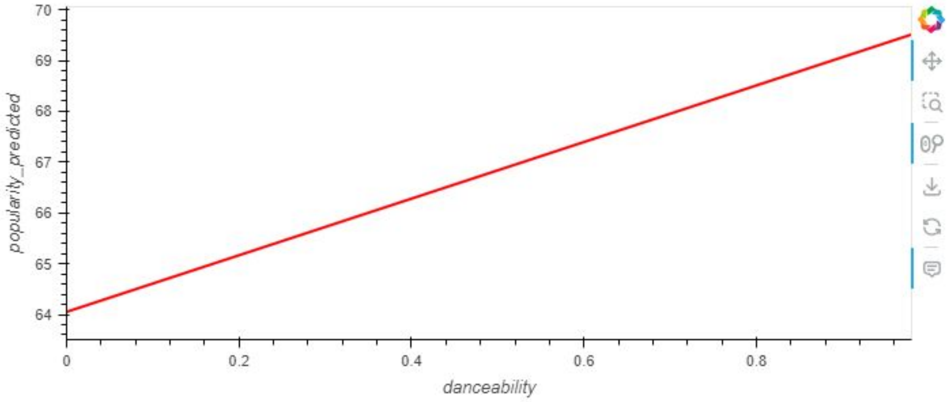
```
63/63 ━━━━━━━━━━━━━━━━━━━━ 0s 419us/step

Classification Report:
              precision    recall  f1-score   support

           0       0.76      1.00      0.86      1512
           1       0.00      0.00      0.00       488

    accuracy                           0.76      2000
   macro avg       0.38      0.50      0.43      2000
weighted avg       0.57      0.76      0.65      2000
```

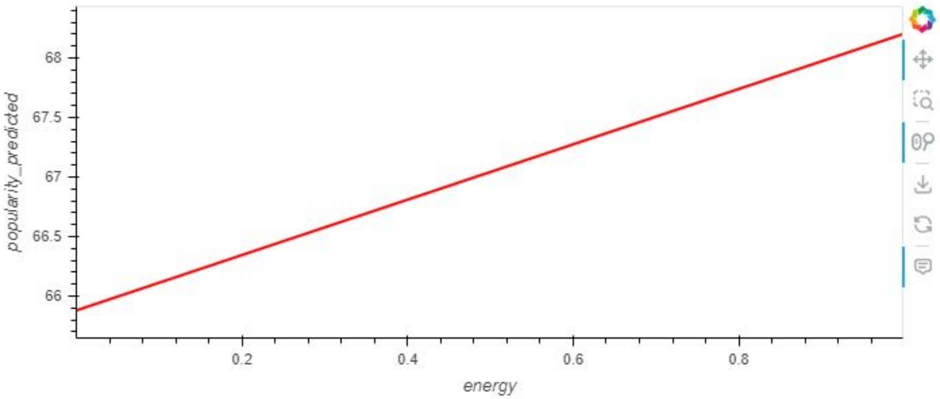
Predict Song Popularity

With this model, we can enter theoretical statistics of a song and see a predicted popularity and how successful the song will be on Spotify

Some Things We've Learned







Some Things We’ve Learned

- The longer a song is the less popular it will be on average
- The higher energy a song is the more popular it will be
- Dance is the most abundant genre of music between the most popular songs
- This model is a good tool to steer a song in the right direction but it does not account for unmeasurable “It Factors” and cannot pump out hits scientifically otherwise someone would have discovered it before