Predicting Spotify Song Streaming Success

COSC 5931

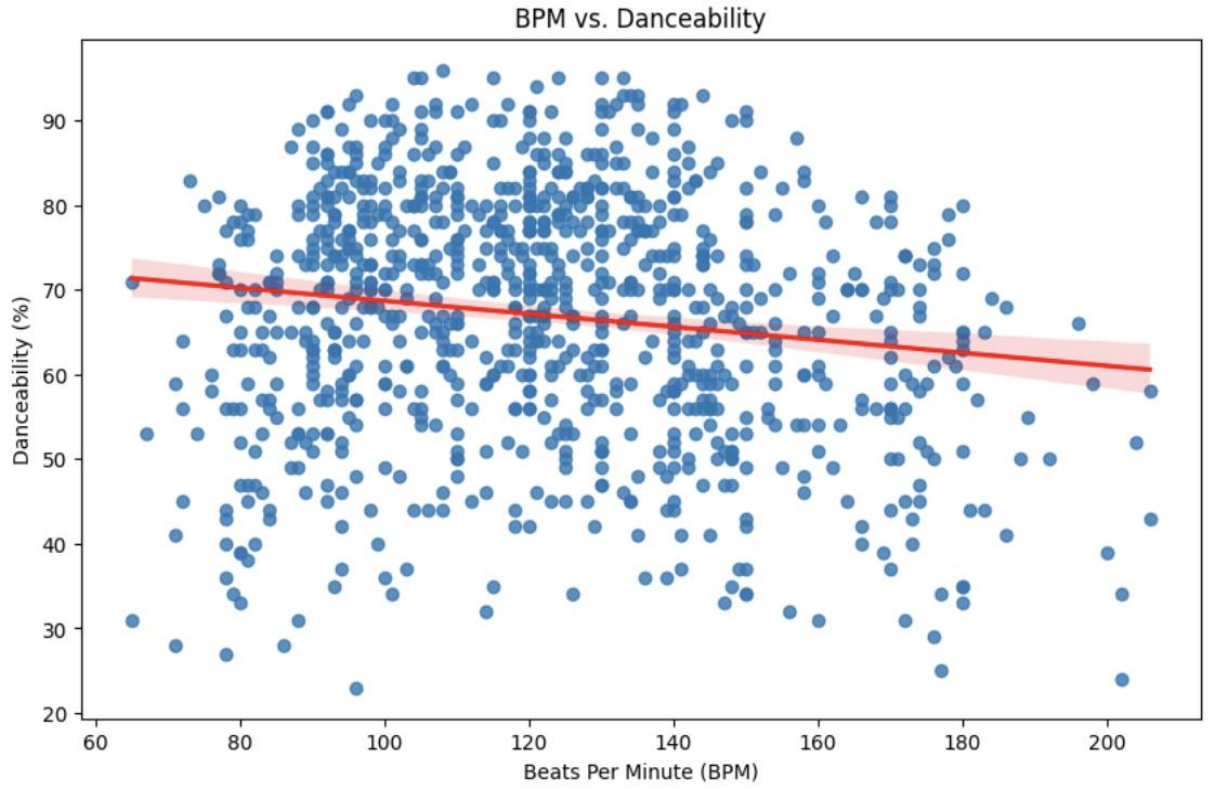
Kayley Reith & Violet Wang

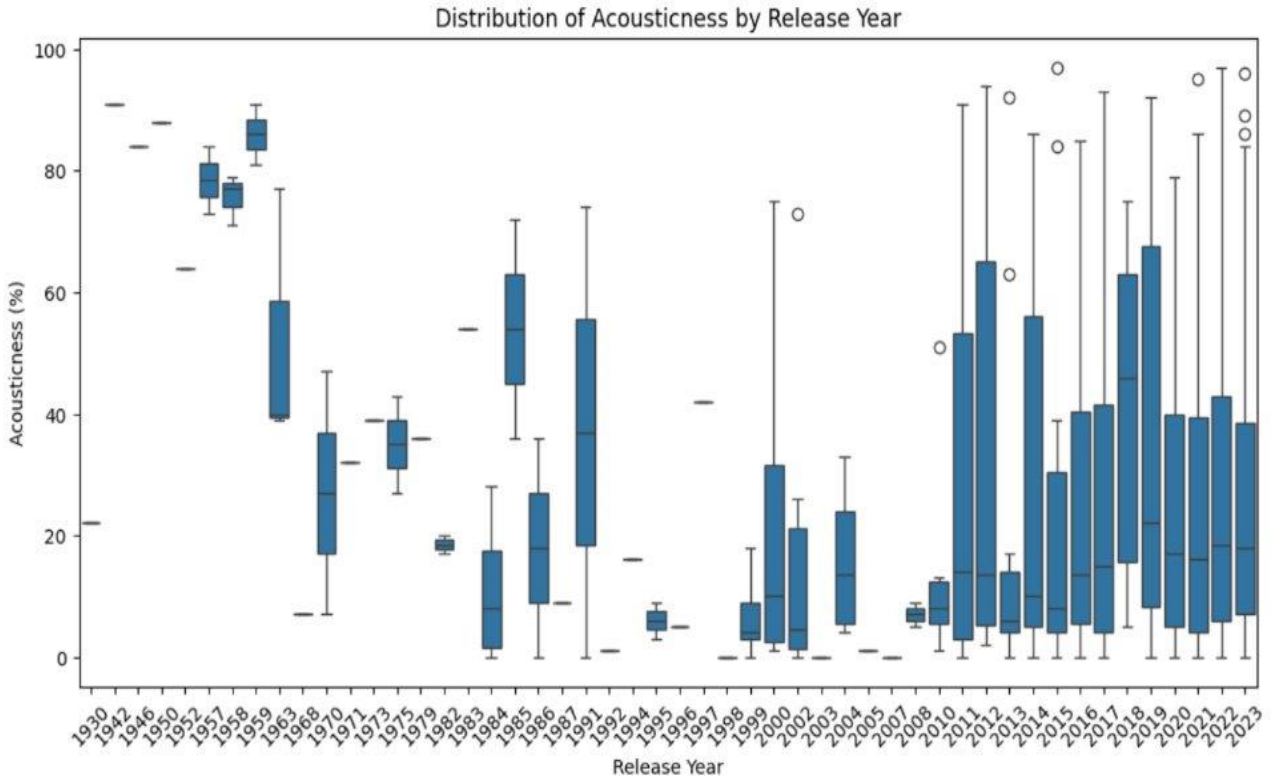
Magnitude of Problem

- **$14.69 billion** in revenue in 2023 (Iqbal 2023)
- 12.9% increase year-on-year (Iqbal 2023)
- **626 million** monthly active users (Singh 2024)

Project Goal: Develop a predictive model that identifies the characteristics of songs likely to achieve high streaming numbers based on their features

Exploratory Data Visualizations





Data Preprocessing

- Observations: 924 songs
- Features: 24
- Numerical & Categorical Data
- Missing values and normalized features for consistency (omit values)
- Feature selection (most impactful attributes)
- Disregard cover_urls and keys features

```
df.dtypes
```

```
track_name              object
artist(s)_name          object
artist_count             int64
released_year            int64
released_month           int64
released_day             int64
in_spotify_playlists     int64
in_spotify_charts        int64
streams                 object
in_apple_playlists       int64
in_apple_charts          int64
in_deezer_playlists     object
in_deezer_charts         int64
in_shazam_charts        object
bpm                      int64
key                     object
mode                    object
danceability_%           int64
valence_%                int64
energy_%                 int64
acousticness_%           int64
instrumentalness_%       int64
liveness_%               int64
speechiness_%            int64
cover_url               object
dtype: object
```

```
#missing values in each column
missing_values = df.isna().sum()
print(missing_values[missing_values > 0])
```

```
in_shazam_charts    50
key                 95
dtype: int64
```

Linear Regression

OLS Regression Results

Dep. Variable:	streams	R-squared:	0.744
Model:	OLS	Adj. R-squared:	0.738
Method:	Least Squares	F-statistic:	127.2
Date:	Fri, 29 Nov 2024	Prob (F-statistic):	1.67e-206
Time:	14:58:04	Log-Likelihood:	-15920.
No. Observations:	761	AIC:	3.188e+04
Df Residuals:	743	BIC:	3.196e+04
Df Model:	17		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-6.333e+09	2.27e+09	-2.787	0.005	-1.08e+10	-1.87e+09
artist_count	-3.461e+07	1.25e+07	-2.769	0.006	-5.91e+07	-1.01e+07
released_year	3.261e+06	1.13e+06	2.887	0.004	1.04e+06	5.48e+06
released_month	1.938e+06	3.18e+06	0.610	0.542	-4.3e+06	8.17e+06
released_day	1.875e+06	1.21e+06	1.549	0.122	-5.02e+05	4.25e+06
in_spotify_playlists	3.721e+04	2192.652	16.969	0.000	3.29e+04	4.15e+04
in_spotify_charts	3.986e+06	7.75e+05	5.144	0.000	2.47e+06	5.51e+06
in_apple_playlists	2.842e+06	2.07e+05	13.696	0.000	2.43e+06	3.25e+06
in_apple_charts	-5.722e+05	2.78e+05	-2.062	0.040	-1.12e+06	-2.73e+04
in_deezer_charts	-6.935e+06	2.4e+06	-2.895	0.004	-1.16e+07	-2.23e+06
bpm	-6.769e+04	3.99e+05	-0.170	0.865	-8.51e+05	7.15e+05
danceability_%	-1.92e+05	8.9e+05	-0.216	0.829	-1.94e+06	1.56e+06
valence_%	-2.025e+05	5.69e+05	-0.356	0.722	-1.32e+06	9.15e+05
energy_%	-1.196e+06	8.96e+05	-1.334	0.182	-2.95e+06	5.63e+05
acousticness_%	8.456e+05	5.33e+05	1.585	0.113	-2.01e+05	1.89e+06
instrumentalness_%	-1.64e+05	1.26e+06	-0.130	0.897	-2.64e+06	2.32e+06
liveness_%	-1.263e+05	8.09e+05	-0.156	0.876	-1.71e+06	1.46e+06
speechiness_%	-1.31e+06	1.2e+06	-1.093	0.275	-3.66e+06	1.04e+06

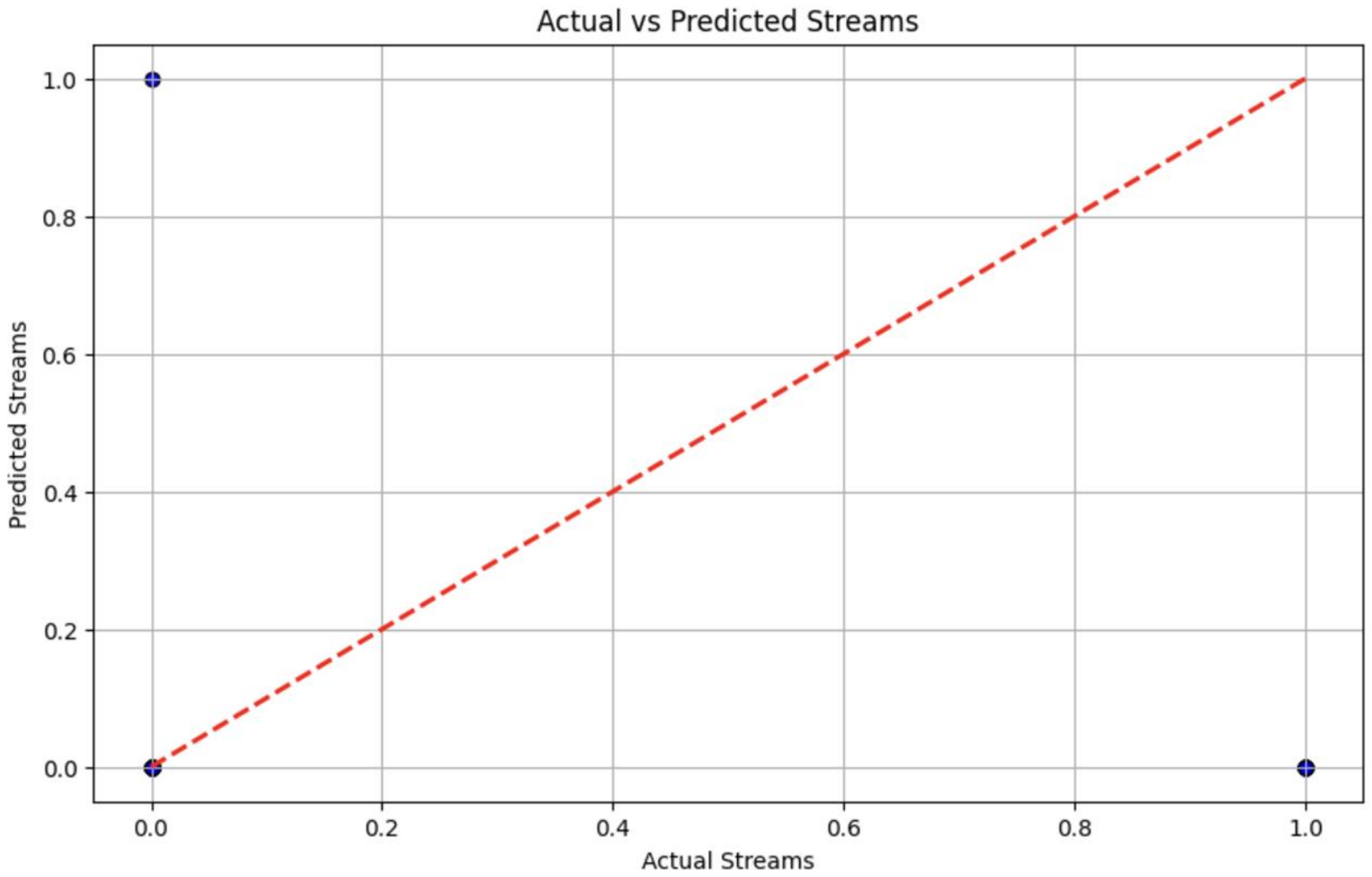
Omnibus:	251.724	Durbin-Watson:	2.067
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2137.768
Skew:	1.243	Prob(JB):	0.00
Kurtosis:	10.826	Cond. No.	2.04e+06

Results Analysis

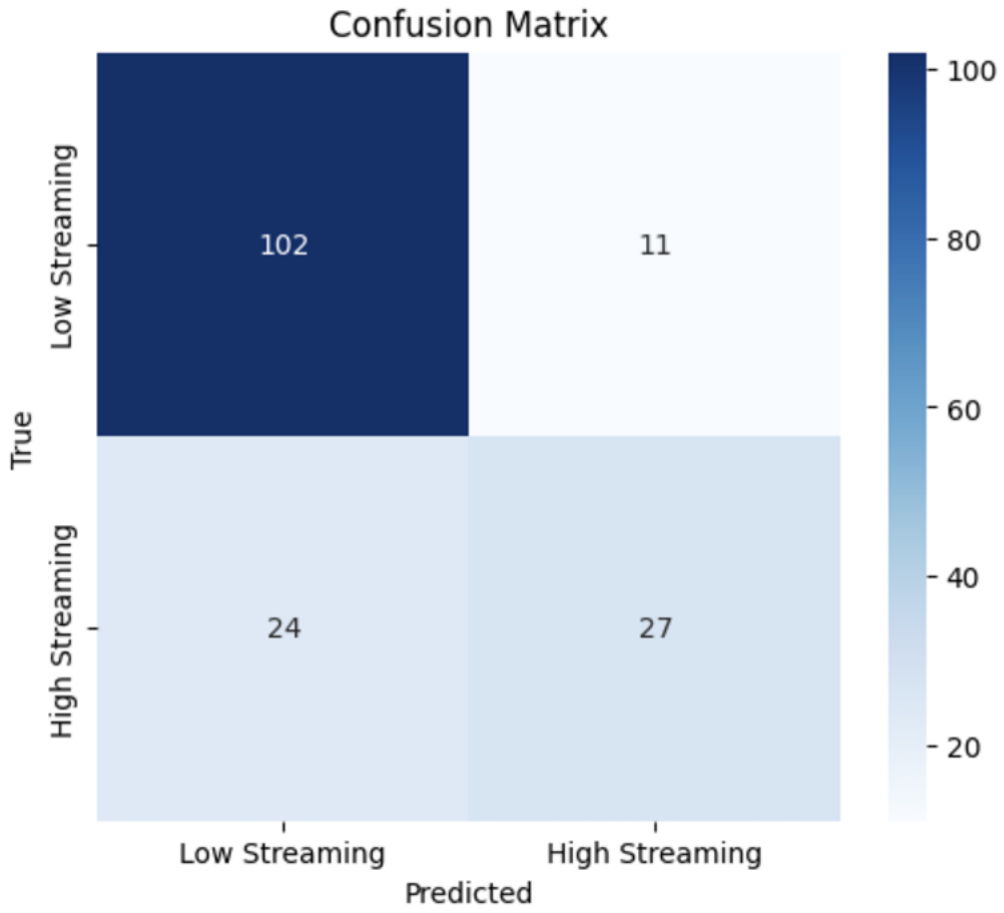
Model Fit: The R-squared value of 0.744 shows that 74.4% of the variance in streams is explained by the model, and the model is statistically significant with a F-statistic of 127.2 (p-value < 0.0001).

Significant Predictors: Variables like in_spotify_playlists, in_apple_playlists, and released_year are significant (p-values < 0.05), indicating their strong impact on streams, while other variables like bpm and danceability_% are not significant.

Linear Regression



Logistic Regression



- Binary Classification (0,1) for high vs low streaming based on the average of song streams
- Class imbalance issue persists

Logistic Regression

```
Classification Report:
              precision    recall  f1-score   support

           0       0.67      0.50      0.57       109
           1       0.34      0.51      0.41        55

    accuracy                           0.50       164
   macro avg       0.50      0.50      0.49       164
weighted avg       0.56      0.50      0.51       164
```

```
Class distribution (high vs low streaming):
high_streaming
0    560
1    256
Name: count, dtype: int64
Accuracy: 0.7865853658536586
Confusion Matrix:
[[102  11]
 [ 24  27]]
Classification Report:
              precision    recall  f1-score   support

           0       0.81      0.90      0.85       113
           1       0.71      0.53      0.61        51

    accuracy                           0.79       164
   macro avg       0.76      0.72      0.73       164
weighted avg       0.78      0.79      0.78       164
```

Deployed class weighting to give more importance to the minority class of high streaming.

Logistic Regression

Results

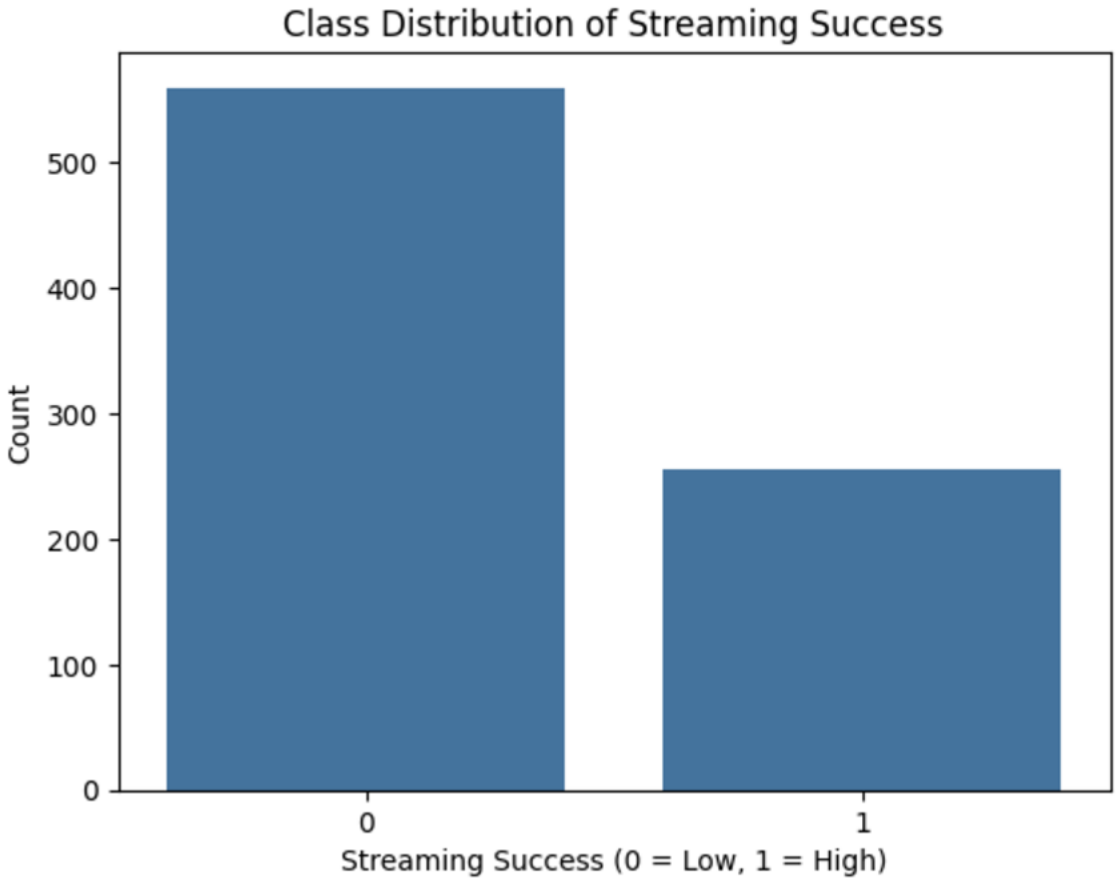
- **Accuracy:** The model achieves an accuracy of 78.65%, correctly predicting streaming success for most instances, but with room for improvement in the minority class.

- **Class Imbalance Impact:** There is a significant class imbalance (651 low vs. 301 high), affecting the model's ability to predict the high streaming class accurately.

- **Precision & Recall:** Precision for low streaming (0) is 81%with high recall 90% while precision for high streaming (1) is 71% with recall at 53%, indicating weaker performance for high streaming predictions.

- **F1-Score:** The F1-score is 85% for low streaming (class 0) and 65% for high streaming (class 1), highlighting the model's struggle with the high streaming class.

Bag of Words Model



Top 10 keywords that correlate with low streaming success:

	Feature	Coefficient
376	feels	-0.485667
1046	thought	-0.519382
409	future	-0.530965
1059	toliver	-0.539307
241	crazy	-0.564033
331	el	-0.666652
1010	super	-0.675310
85	baby	-0.755150
1032	taylor	-0.780348
1195	½ï	-1.362818

Top 10 keywords that correlate with high streaming success:

	Feature	Coefficient
271	dark	1.139508
599	lost	0.945983
568	levitating	0.935201
546	la	0.886033
1005	summer	0.810141
1029	talking	0.785155
264	damn	0.754994
561	leave	0.709468
541	know	0.684657
127	bl	0.656986

Bag of Words Model

Results

Accuracy:

65.2%, but the model struggles with identifying high-streaming success songs, as shown by precision and recall of 0 for high streaming.

Class Imbalance:

The model is biased towards predicting low streaming success (class 0), as indicated by a recall of 98% for class 0, while it fails to predict high streaming success.

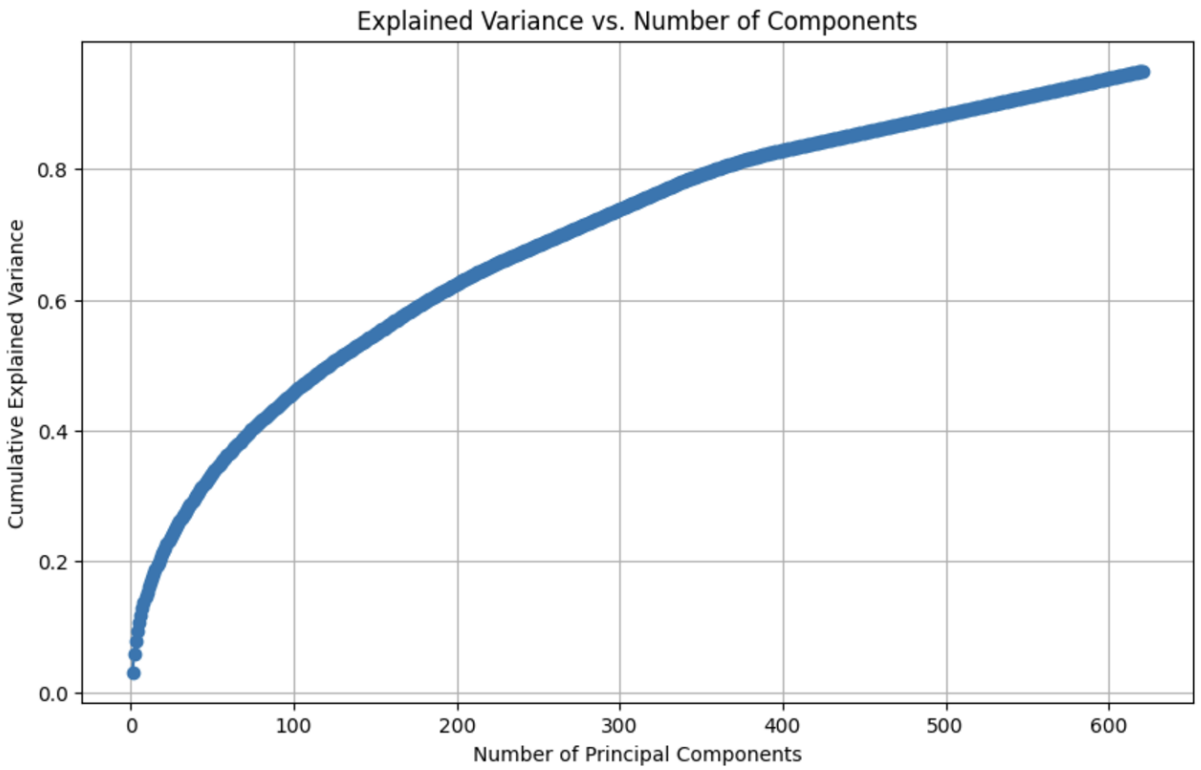
Positive Keywords for High Streaming:

Keywords like "dark," "summer," "remix," and "levitating" are positively associated with high streaming success, suggesting themes of energy, vibe, and popular trends.

Negative Keywords for Low Streaming:

Keywords like as "feels," "thought," "baby," and "crazy" correlate with low streaming success, potentially reflecting less popular or niche themes.

Principal Component Analysis



Number of components retained to explain 95% variance: 621

- **Model Predicts Majority Class Only:** The PCA-based model has a moderate accuracy of 60.37% by predicting the majority class (0) well
- Fails to identify any instances of the minority class (1), resulting in low precision and recall
- **Imbalance Skews Results**: Class imbalance heavily influences the model's performance, where the negative class (0) dominates the outcomes with high recall of 96%

Key Takeaways

- **Playlist presence** is a critical predictor of streaming success
- **Class imbalance** significantly affects model performance, highlighting the need for balanced datasets
- **Future improvements** could include advanced models like neural networks or ensemble methods to better handle imbalanced data
- **Insights into song features** provide valuable guidance for artists and producers aiming for **high streaming success**
- **Logistic Regression** was the optimal model with the highest accuracy of 78% after class weighting

Sources

- Iqbal, Mansoor. "Spotify Revenue and Usage Statistics (2024)." Updated October 1, 2024. Business of Apps.
- Singh, Shubham. "Spotify Statistics (2024) – User Growth, Top Artists & More." September 17, 2024. Demand Sage.