COSC 5931: Introduction to Machine Learning

Final Project Report

Predicting Song Streaming Success

Kayley Reith & Violet Wang

12/10/24

Abstract:

The advancement of entertainment technology has reached new heights in popularity, profitability, and user experience. The music industry has digitized platforms like Spotify to offer users a personalized listening experience. Classifying which songs have low or high streaming success can dictate the algorithms Spotify uses to advertise and recommend songs to users and increase advertising efforts. The methods deployed to predict and classify song streaming success were multifaceted in supervised and unsupervised machine learning techniques consisting of; linear regression, logistic regression, bag-of-words, and principal component analysis. The results yielded some inconsistency with class imbalances and demonstrated higher performance for the majority class of low-streaming songs. Ultimately, this paper aims to uncover the significant features of high-streaming songs on Spotify.
Keywords: Spotify, Machine Learning, Supervised Learning, Unsupervised Learning

Literature Review:

The entertainment and music industry has transformed with the increase of digital technology to provide artists with a global presence and digital platforms to advertise and share their music with a widespread audience. One of the most used applications and streaming services in the United States is Spotify which has over 626 million monthly active users (Singh 2024). The increasing demand and popularity of this application has generated over $14.69 billion in revenue in 2023 and continues to improve annually with 12.9% profitable increase year-over-year (Iqbal 2023). The profitability and large audience indicate how significant this application is to our economy, entertainment industry, and the lives of everyday people. With all of this information in mind, determining which songs and artists will rise to the top in terms of streaming numbers and listeners can be difficult to navigate. Therefore, the goal of this project is to apply machine learning techniques to analyze Spotify's most streamed songs and predict the future streaming success of artists. This new information could help optimize listener experiences, improve advertising efforts, increase profits for undiscovered artists, and bridge the gap in listeners of this application.

This data source was found in a repository in the Kaggle dataset platform. The dataset is a comprehensive list of the most famous and streamed songs of 2023 and contains 924 song observations with 24 features of mixed categorical and numerical data. The first steps consisted of data management and preprocessing to ensure optimal quality data before performing and deploying models. The creation of exploratory visualizations helped identify data points that were outliers and missing data that was

omitted, like variables keys and songs within Shazam charts. The exploratory visualizations helped establish relationships like song distribution and release years (Appendix A), distribution of song acoustics by release year (Appendix B), and a correlation heatmap matrix of song features (Appendix C). Additionally, the data types of some column features were transformed from objects to numerical data types (Integers) to perform linear and logistic regression to ensure the data quality was consistent.

Gap in Knowledge:

Spotify's application-related measures, such as playlist inclusion and chart rankings, and song attributes, like genre, tempo, track names, and danceability, all have an impact on streaming success on platforms like Spotify. Previous studies have examined some of these factors separately, but they have not yet thoroughly examined their combined effects. Thus, it is imperative that we investigate both aspects and close the gap between application and song streaming in order to better understand the factors that most influence listener experience and optimize artist outcomes in the entertainment industry.

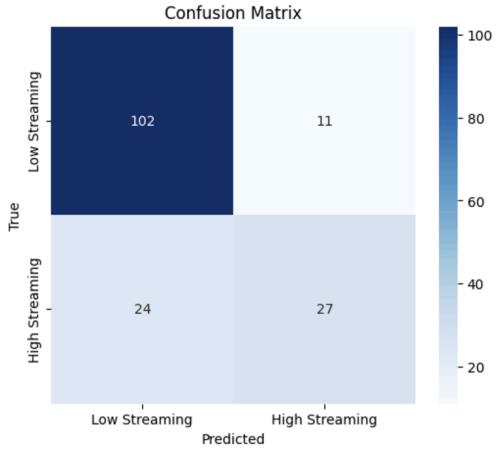
Methodology:

To fully examine which features and characteristics are the most significant in contributing to Spotify's high streaming success, several candidate methods and models were developed, compared, and evaluated to ensure optimal performance. A mixture of supervised and unsupervised techniques was analyzed to enable a comprehensive understanding of the patterns in the data.

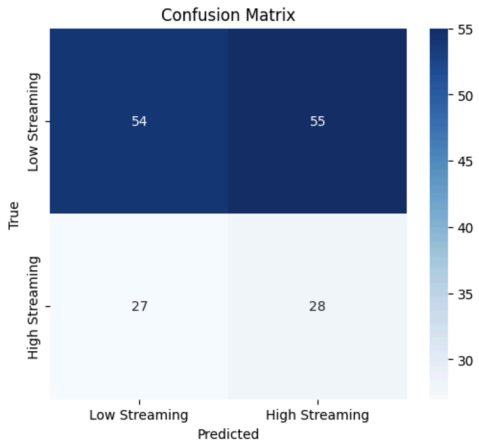
Through modeling the linear relationship between independent factors and the dependent variable of song streams, linear regression is a supervised technique that predicts the number of streams as a continuous outcome. To successfully deploy the linear model we dropped unnecessary columns that could not be converted to integers like track name, cover url, or artist name, this feature selection helped refine the process to only account for the most important variables. We then split and trained the data with 80% training and 20% testing, then ran the Ordinary Lease Squares (OLS) regression test to highlight the coefficient values of each variable to uncover which increases the streams most significantly.

Another supervised technique deployed was logistic regression, which predicts a class outcome. A binary classification approach was tackled in classifying songs into high vs. low streaming success as 0 and 1 values, based on whether the song streams were above or below the average of total streams. The environment was set up by importing necessary libraries, such as LogisticRegression from sklearn.linear_model, into the Jupyter Notebook workspace. The binary target variable was developed based on the average of streams and the data was split for 80% training and 20% testing based on the subset of feature variables that were selected which consisted of artist count, song release year, song release month, beats per minute (bpm), danceability, and energy. The model was then deployed to

determine class outcomes for high streaming success. An issue arose with class imbalance as indicated by the confusion matrix uneveness (Figure 1.1). The solution was to use class weighting to give more importance to the minority class of high streaming (Figure 1.2), which helped create a more uniform distribution between classes helping improve model performance.



(Figure 1.1 Unweighted Classes)

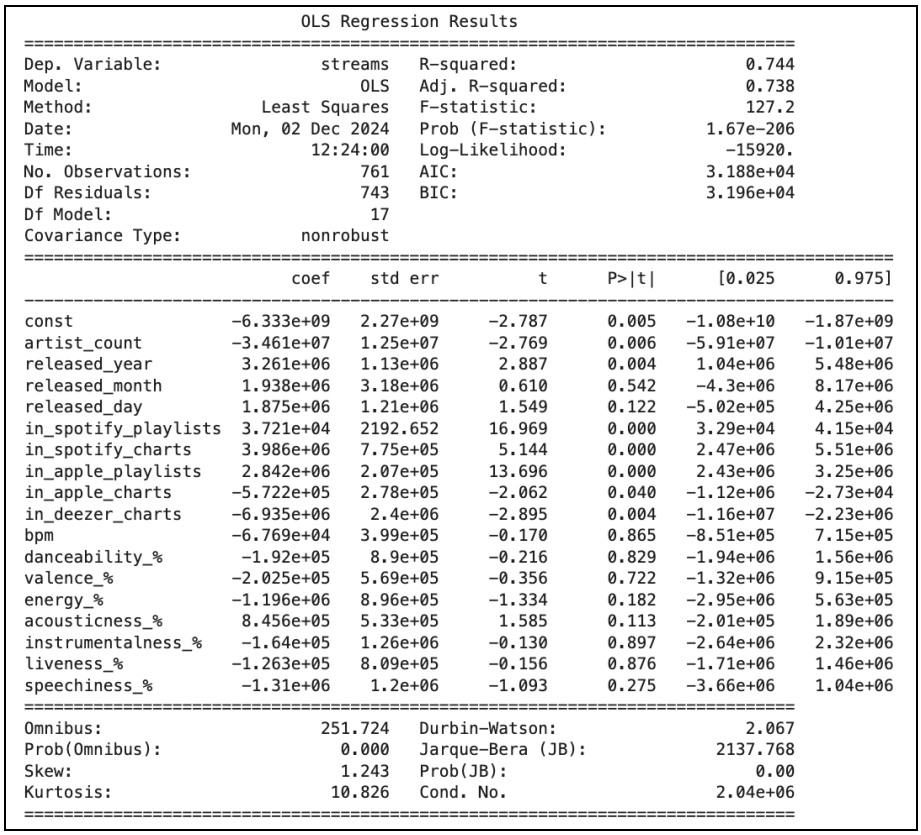


(Figure 1.2 Weighted Classes)

The final algorithmic approach was an unsupervised text-mining analysis using a bag-of-words model. This method looked at the unpositioned keywords in the track names to see if there were any common words, phrases, themes, or expressions that were associated with high streaming success. To determine which set of keywords classified songs as having a high or low streaming success based on the target variable we put in the logistic regression model, the environment was first set up by importing multiple libraries from Sklearn, vectorizing the words found in the track names to count the total number, and then splitting the data for training and testing the model before deployment.

Discussion & Analysis:

The linear regression results yielded from the OLS summary (Figure 2.1) showed that, when all other factors are held constant, songs added to Spotify playlists boost streams by about 37,210 streams per playlist entry. Additionally, songs that are featured on the Spotify charts boost the streaming capacity by roughly 3,986,000 streams per chart appearance, when all other factors are held constant. The coefficient values are depicted below.



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:	Mon, 02 Dec 2024	Prob (F-statistic):	1.67e-206
Time:	12:24:00	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

(Figure 2.1 OLS Summary)

The model evaluation metrics and diagnostics display an R-squared value of 0.744 explaining the model with 74.4% accuracy and explaining the variance in streams. The significant predictors of high streaming success are songs in Spotify playlists and release year as they have p-values less than 0.05, indicating their strong impact on streaming success. However, while the R-squared value is a good fit there is still room for improvement and the mean squared error (MAE) suggests how far the predictions are off in the model development. Typically, the mean squared error focuses on the absolute difference between the actual and predicted values (Appendix D) and the value detected was 200299013.12, indicating that on average the predicted number of high streams strays away from the observed value of streams by 200 million, while this number seems significantly large it falls under the average of streams listened which is 468 million.

The logistic regression model yielded more significant results once the class imbalance issue (Figure 2.2) was addressed with class weighting (Figure 2.3). The findings show that while the model predicts streaming success in the majority of cases with an accuracy of 76.83%, there is still an opportunity for improvement in the minority class. This could have occurred from the class imbalance impact affecting the model's ability to predict the song's high streaming class accurately. The low

streaming class had 81% precision with 90% recall and 85% for the F1-score. Meanwhile, the high streaming class had 71% precision with 53% recall and 61% F1-score, indicating a weaker performance for songs with high streaming predictions.

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

(Figure 2.2 Class Imbalance Report)

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

(Figure 2.3 Class Balance Report)

The bag-of-words model results indicated with 65.2% accuracy that these positive keywords (Figure 2.5), “dark," "summer," "remix," and "levitating" within song titles contribute the most to high streaming success. Meanwhile the negative keywords "feels," "thought," "baby," and "crazy" tend to correlate with low song streaming success (Figure 2.4). The results from this bag of words model can optimize up-and-coming artists' different songs on the Spotify platform by incorporating these keywords into their track names to gain more viewership and streams.

Top 10 keywords that correlate with low streaming success:

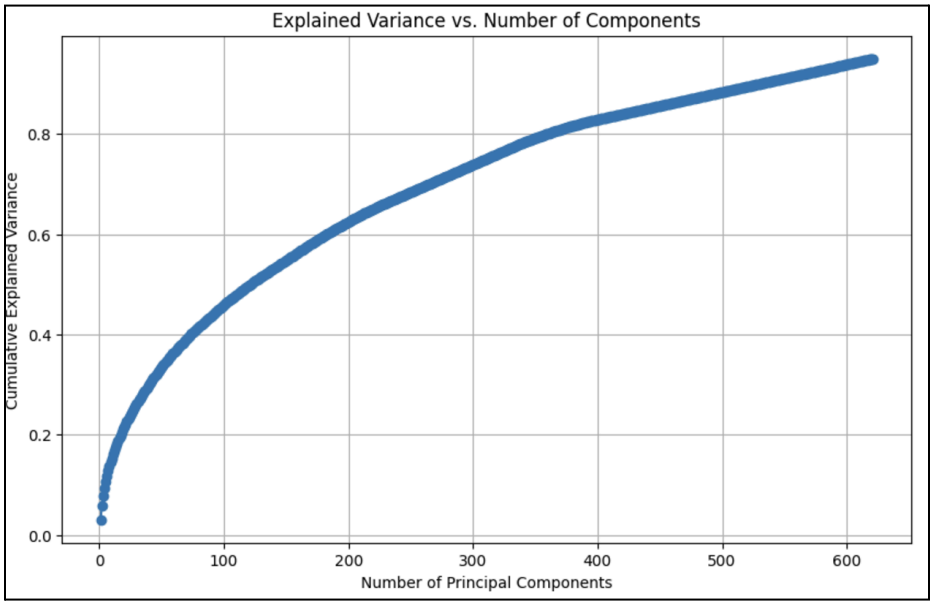
	Feature	Coefficient
376	feels	-0.485667
1046	thought	-0.519382
409	future	-0.530965
1059	toliver	-0.539307
241	crazy	-0.564033
331	el	-0.666652
1010	super	-0.675310
85	baby	-0.755150
1032	taylor	-0.780348
1195	½ï	-1.362818

(Figure 2.4 Negative Keywords)

```
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
```

(Figure 2.5 Positive Keywords)

The model was evaluated with a principal component analysis (PCA) where we specified the number of components to explain 95% variance or reduce to a fixed number of components (Figure 2.6). The PCA model predicts the majority class only achieves moderate accuracy by 60.37% of the low streaming class and fails to identify the instances of the minority class. After applying the PCA there still exists high dimensionality after reduction with 621 song observations that explain approximately 95% of the variance in the dataset.



(Figure 2.6 PCA)

Insights & Lessons:

One of the most persistent challenges was the class imbalance impact seen throughout the logistic regression model, bag-of-words model, and principal component analysis (PCA) as it skewed accuracy

results for the minority class of high streaming (1). The results of the low streaming class for all models tend to perform stronger than the minority class of songs with streams below the average of 400 million listens.

Evaluating all three models reveals how the most optimal model in predicting song streaming success was the logistic regression model. This model yielded the highest accuracy of 78% explaining the model after class weighting to address the class imbalance impact. The precision of predicting the positive cases correctly for the low streaming class was 81% accurate and the recall was 90% accurate in classifying the proportion of true positive cases by the model. While the model performs well for the majority class of low streaming classes some improvement could be made to address the high streaming metrics, with a weaker performance of 71% precision and 53% recall. While the logistic regression model has room for improvement it made significant strides in addressing, classifying, and predicting Spotify song streaming success. The results of this model can be applied to personalize Spotify user experience to recommend songs based on streaming success which was classified by the feature subsets containing song distribution year, energy percentages, and beats per minute.

Future Directions:

To enhance the predictive capabilities and insights derived from this project, several improvements can be pursued in future work. Incorporating additional features, such as regional popularity, user demographics, or listener engagement metrics like skip rates and completion rates, could provide a more nuanced understanding of factors influencing streaming success. Moreover, integrating external data, such as social media activity or marketing campaign efforts, might uncover hidden correlations between promotion strategies and streaming numbers.

Advanced modeling techniques could also improve performance. Methods such as Random Forests or Gradient Boosting (e.g., XGBoost, LightGBM) may provide a more robust alternative to linear and logistic regression. Additionally, exploring deep learning approaches like recurrent neural networks (RNNs) or transformers could help capture sequential patterns in song-release data or even lyrical structures. Addressing the class imbalance issue further, with techniques like Synthetic Minority Oversampling (SMOTE) or cost-sensitive learning methods, may enhance the model's ability to accurately predict high-streaming songs.

Refinement of text analysis is another promising area for improvement. Expanding beyond song titles to include lyrics or additional metadata and employing advanced natural language processing techniques like word embeddings (e.g., Word2Vec, GloVe) or transformer-based models (e.g., BERT) could provide deeper insights into linguistic patterns that drive streaming success. Additionally,

introducing time-series models to analyze trends over time may reveal temporal dynamics in streaming behavior that static models cannot capture.

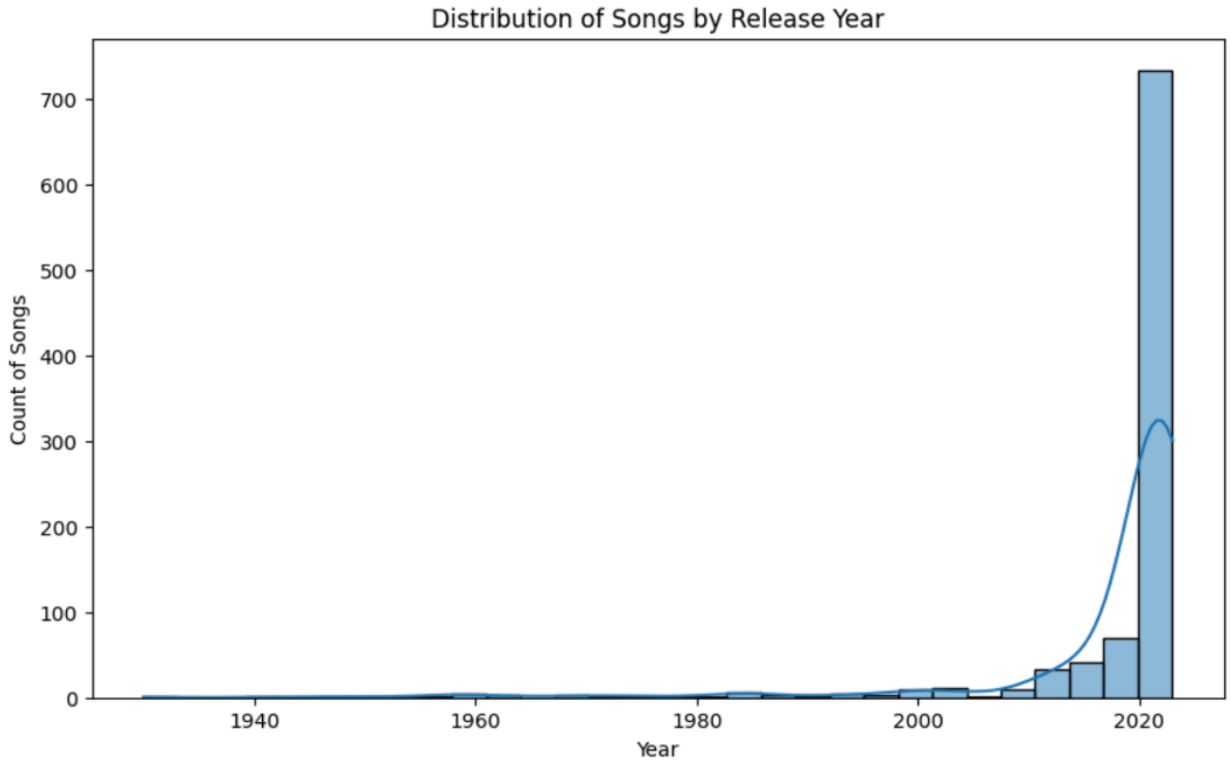
Future research could also focus on cross-platform analysis, integrating data from other streaming services such as Apple Music or YouTube Music, to examine trends across different platforms and their collective influence on song success. Understanding individual user behaviors, including playlist preferences and listening habits, could further refine predictive models by uncovering micro-patterns in how audiences interact with music. Finally, enhancing model evaluation by considering alternative metrics, such as ROC-AUC, precision-recall curves, or the Matthews correlation coefficient, may provide a more comprehensive assessment of model performance, particularly for imbalanced datasets.

Conclusion:

This project effectively identified key factors influencing song streaming success on Spotify, such as playlist inclusion and chart rankings, by employing a combination of linear regression, logistic regression, and text-mining approaches. The linear regression model explained 74.4% of the variance in streams, while logistic regression achieved a balanced accuracy of 76.83% after addressing class imbalance. Text analysis further revealed actionable insights, identifying keywords in song titles that correlate with high or low streaming success.

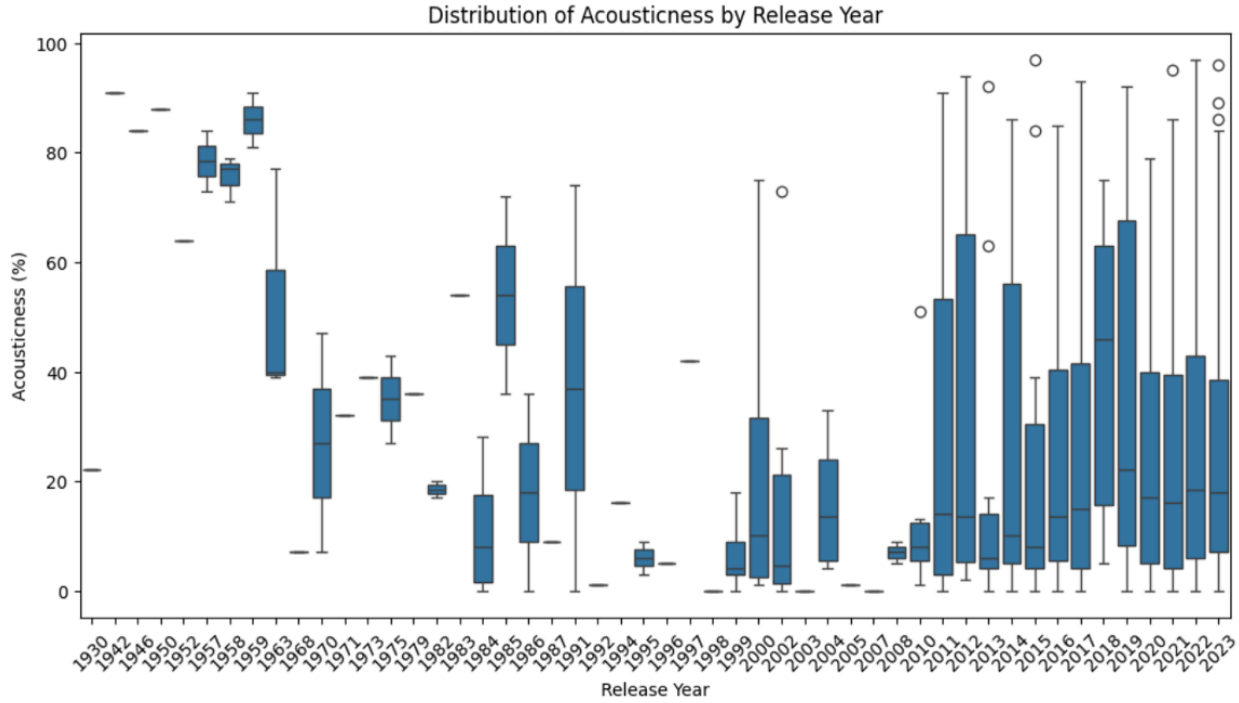
However, the models also exhibited limitations, particularly in accurately predicting songs with high streaming success and in managing the high dimensionality of the data. Future improvements, such as the inclusion of additional features, adoption of advanced modeling techniques, refined text analysis, and cross-platform comparisons, present opportunities to deepen understanding and enhance predictive accuracy. By building upon these findings and addressing existing challenges, the project has the potential to further assist artists, platforms, and listeners in optimizing the dynamics of digital music streaming.

Appendix A



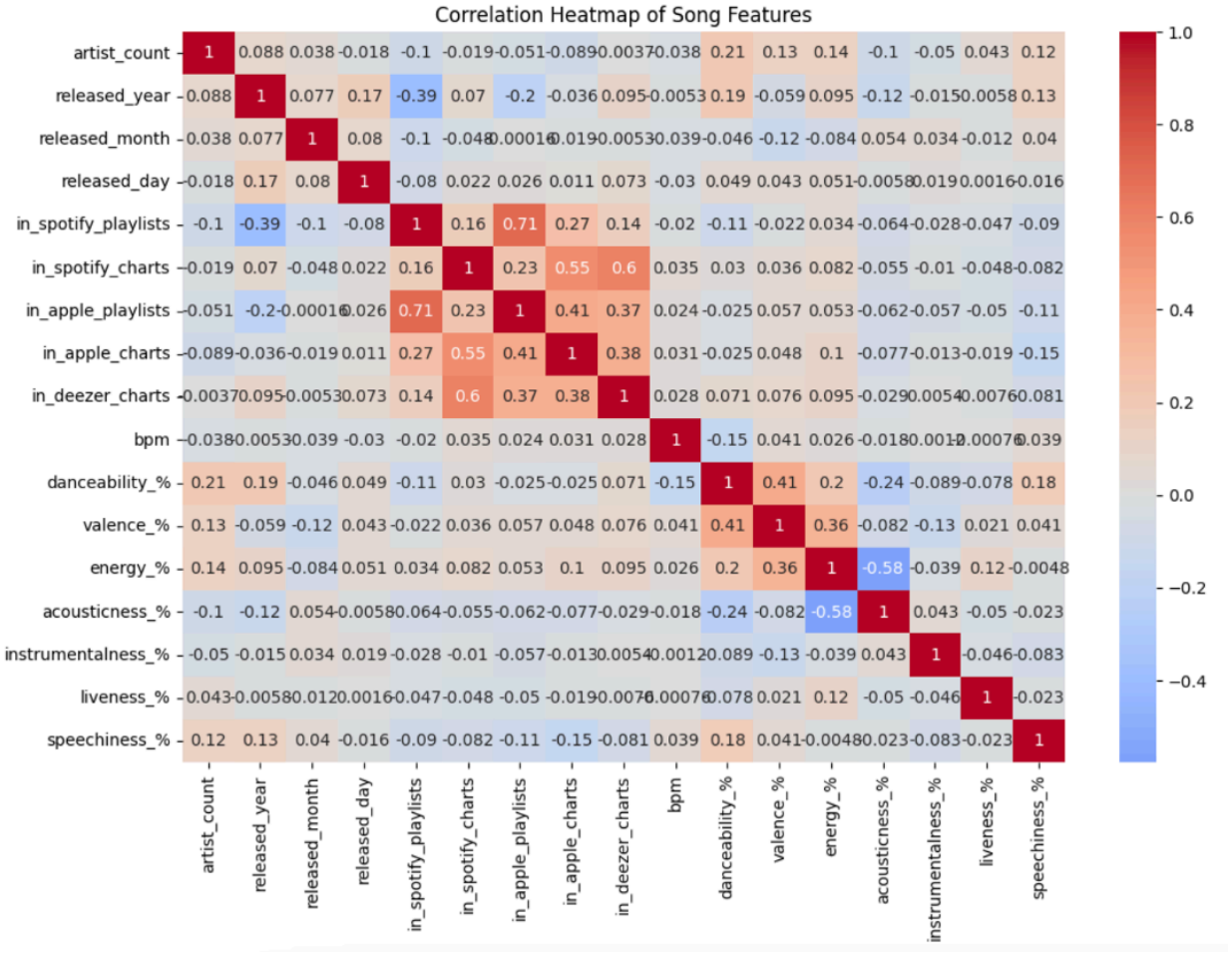
The distribution shows within the historical period (1940-1990) a low number of songs per year with a relatively flat distribution, indicating fewer than 50 songs per year. Meanwhile, the modern era (2000-2020)demonstrates a dramatic exponential increase in the number of songs, with a sharp rise from 2015 onwards and reaching a capacity of around 700+ songs in the most recent year 2020.

Appendix B



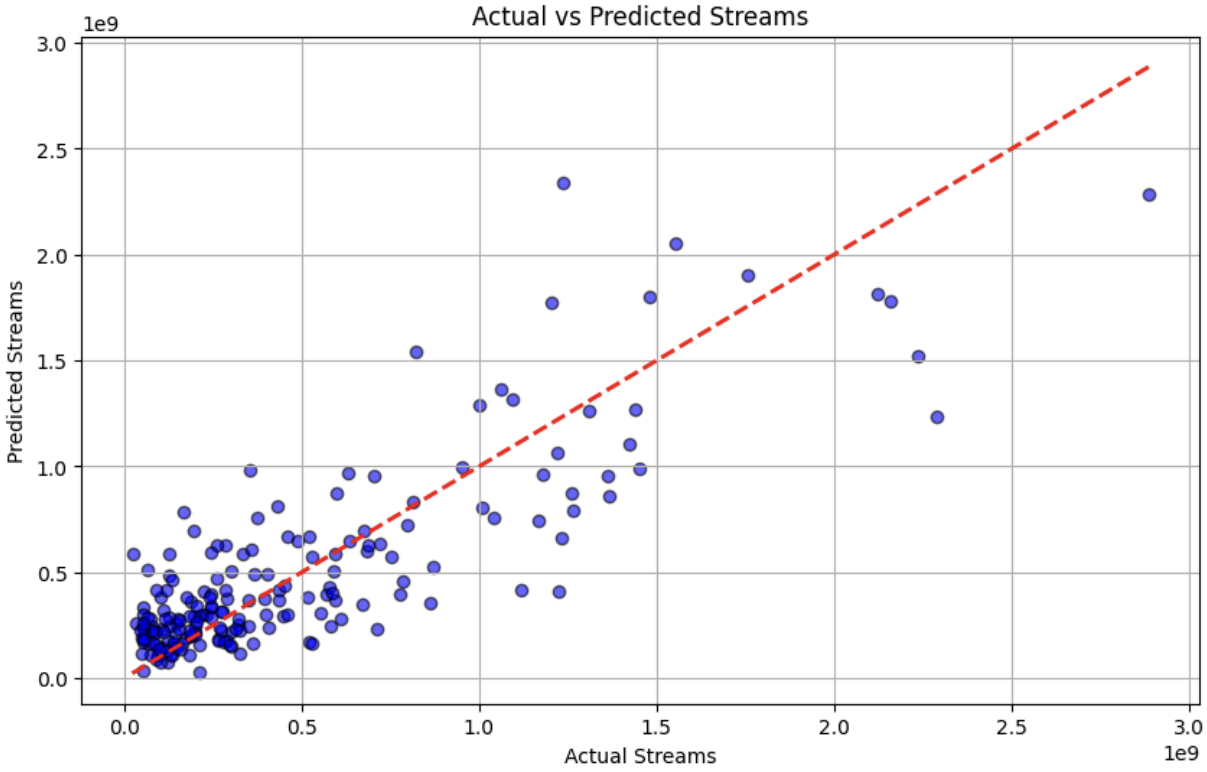
This box plot displays the distribution of "acousticness" which is a musical song attribute measured as a percentage across different release years, from approximately 1930 to 2023. Within the earlier years (1930s-1950s) there is consistently high acoustics values, between 75-90% with less variation in the measurements. However, the mid-century transition (1960s-1970s) depicts a decline in acoustics values with larger variations between the measurements. Lastly, the modern era (1980s-2023) generally has lower acoustics values, ranging from 0-40% with greater variation in measurements. While there are some slight fluctuations in the graph, it generally is maintained at lower levels compared to the earlier years' acoustic metrics.

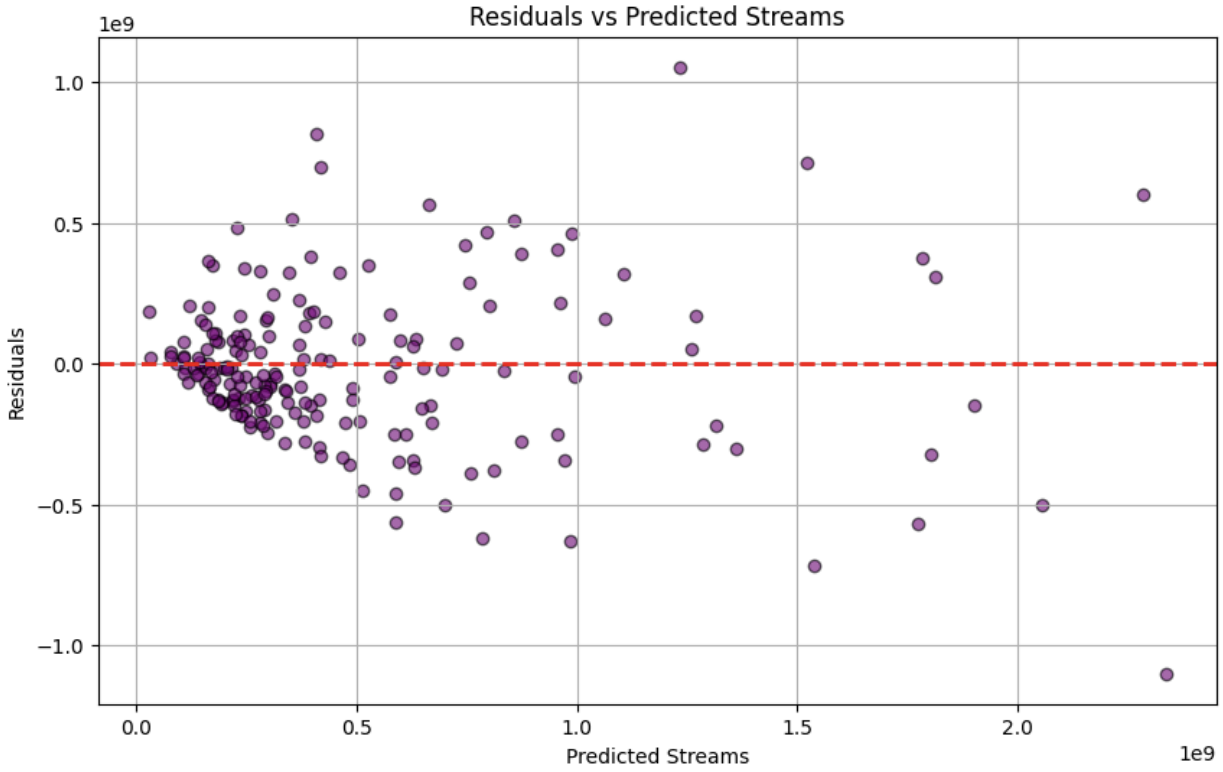
Appendix C



The heatmap correlation matrix has unveiled some unique relationships between song and platform attributes that contribute to high streaming success. Some of the correlations to highlight are a strong positive correlation between music in Spotify playlists and music in Apple playlists with a 71% correlation suggesting that songs in Spotify playlists are likely also included in Apple playlists, indicating a shared popularity or playlisting strategy across platforms. A somewhat strong negative correlation between energy and acoustics at -58% indicates how higher-energy songs tend to have lower acoustics.

Appendix D


Actual vs Predicted Streams
1e9
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Predicted Streams
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Actual Streams
1e9


Residuals vs Predicted Streams
1e9
1.0
0.5
0.0
−0.5
−1.0
Residuals
0.0
0.5
1.0
1.5
2.0
Predicted Streams
1e9

References

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.