Task #1 – Algorithms of Machine Learning for data analsis.

Eng. Andrés Manuel Prieto Álvarez — andres.prieto@utp.edu.co

***Abstract*—Keywords.**

***Index Terms*—Song virality, Spotify, Top Spotify Songs dataset, Streaming platforms, Music consumption, Predictive modeling, Analytics, Machine learning, Data analysis.**

I. Introduction.

In recent years, the proliferation of digital streaming platforms, epitomized by Spotify's global dominance, has heralded a transformative shift in the music industry landscape. With millions of songs competing for listeners' attention, understanding the factors that distinguish one song's success from another assumes paramount importance. Against this backdrop, the present study assumes significance as it endeavors to decipher the underlying determinants contributing to song popularity and virality within the Spotify ecosystem. By leveraging advanced data analytics techniques, the research aims to provide empirically grounded insights for artists, record labels, and industry stakeholders. These insights hold promise in optimizing promotional strategies, tailoring content creation, and navigating the dynamic terrain of digital music consumption with precision and foresight. Thus, the research not only addresses a timely and pertinent issue but also stands poised to catalyze transformative change within the music industry.

In the contemporary landscape of music consumption, understanding the intricate dynamics governing the popularity and virality of songs emerges as a critical pursuit. This study delves into an analysis of the "Top Spotify Songs" dataset, a comprehensive reservoir of song features encompassing 24 distinct attributes. This dataset is pretty new, sourced from Kaggle as a bronze-tier resource, it was updated just 12 days ago until the writing point, without detracting from it for this, "Top Spotify Songs" provides a valuable lens into the multifaceted nature of song success within the digital streaming realm. Positioned within the domain of big data analytics and machine learning, this research aims to systematically dissect the underlying patterns driving song success on one of the world's foremost music streaming platforms. By leveraging advanced data analytics techniques and adopting a hypothesis-first followed by a deep analysis approach, the research aims to provide empirically grounded insights for artists, record labels, and industry stakeholders. These insights hold promise in optimizing promotional strategies, tailoring content creation, and navigating the dynamic terrain of digital music consumption with precision and foresight. Thus, the research not only addresses a timely and pertinent issue but also stands poised to catalyze transformative change within the music industry.

II. Problem Statement.

The specific problem addressed in this research revolves around the challenge of predicting song virality on Spotify, utilizing the recently uploaded "Top Spotify Songs" dataset. The importance of this problem stems from the evolving landscape of music consumption, where digital streaming platforms have become primary avenues for discovering and enjoying music. Understanding the factors that contribute to a song's popularity and virality is crucial for artists, record labels, and industry stakeholders seeking to optimize promotional strategies and content creation.

Of particular significance is the "Streams" attribute, denoting the total number of streams on Spotify, which emerges as the most pivotal feature within the dataset. "Streams" serves as a direct measure of a song's traction and reach among listeners, making it a key determinant of song virality. Additionally, other attributes such as "In Spotify playlist," "In Spotify charts," "In apple playlist," "In apple charts," "In Deezer playlists," "In Deezer charts," and "In Shazam charts" provide insights into the track's presence and performance across various streaming platforms, further enriching the predictive modeling process.

However, given that this dataset was uploaded just 12 days ago, there is a scarcity of existing research on its contents and implications. As such, there is a pressing need for rigorous analysis and exploration to unlock latent insights and trends within the dataset. Existing solutions, if any, are likely to be limited in scope and sophistication due to the novelty of the dataset and the lack of prior research. Thus, there is an opportunity to pioneer novel analytical approaches and methodologies to extract meaningful insights from this recently available data resource.

By addressing this problem through a combination of database analysis and machine learning modeling, with a specific emphasis on critical features such as "Streams" and others (See table I), this research seeks to bridge existing gaps and provide actionable insights for stakeholders in the music industry. Ultimately, the goal is to enhance our understanding of music consumption patterns and empower industry practitioners with the tools and knowledge needed to navigate the complexities of the digital music landscape effectively.

Feature	Description
Track name	Name of the song
Artist(s) name	Name of the artist of the song
Artist count	Number of artists contributing to the song
Released year	Year when the song was released
Released month	Month when the song was released
Released day	Day of the month when the song was released
In Spotify playlist	Number of Spotify playlists the song is included in
In Spotify charts	Presence and rank of the song on Spotify charts
Streams	Total number of streams on Spotify
In apple playlist	Number of Apple Music playlists the song is included in
In apple charts	Presence and rank of the song on Apple Music charts
In Deezer playlists	Number of Deeser playlists the song is included in
In Deezer charts	Presence and rank of the song on Deezer charts
In Shazam charts	Presence and rank of the song on Shazam charts
BPM	Beats Per Minute
Key	Chords scales (From major to minor)
Mode	Rhythmic relationship between long and short duration
Danceability (%)	Measures how suitable a song is for dancing based on a combination of musical elements
Valence (%)	A measure from 0 to 100 describing the musical positiveness conveyed by a track
Energy (%)	Perceptible measure of intensity and activity
Acousticness (%)	A confidence measure from 0.0 to 1.0 of whether the track is acoustic
Instrumentalness (%)	Predicts whether a track contains vocals
Liveness (%)	Refers directly to reverberation time
Speechiness (%)	Detects the presence of spoken words in a track

TABLE I: Features explanation

III. Implementation and Analysis of the Dataset.

A. Preliminary Analysis

This section provides a comprehensive overview of the preliminary steps taken to prepare the data and formulate hypotheses, laying the groundwork for the subsequent phases of analysis.

1) Data Preparation: Before delving into the analysis, a preliminary examination of the dataset was conducted to outline the initial steps. Given the objective of focusing on general trends rather than individual cases, certain features were deemed irrelevant for analysis and thus removed from the dataset. These included ”Track name,” ”Artist(s) name,” ”Released year,” ”Released month,” ”Released day,” as well as platform-related metrics such as ”In Spotify playlist,” ”In Spotify charts,” and others. Additionally, metadata features such as ”Key,” ”Mode,” ”Valence (%),” ”Acousticness (%),” ”Instrumentalness (%),” and ”Liveness (%)” were excluded due to their lack of relevance to the analysis until that research point. This pruning resulted in a subset of the dataset containing the following metrics of interest: Streams, BPM, Danceability (%), Energy (%), and Speechiness (%). Along with an extra feature that was taken accidentally which was ”Artist count”.

Once these metrics were defined, a second review of the dataset documentation was carried out to subsequently verify that the types of the selected features were compatible, since the ”Top Spotify Songs” dataset has values that are text, as well as values that are objectives. and values that are numeric. This review determined a table of values that allowed a series of hypotheses to be made based on the textual description of the features in the dataset. In the following table II you can see the types that were verified together with their description.

Feature	Description	Data type
Streams	Total number of streams on Spotify	str
BPM	Beats Per Minute	int
Danceability	Measures how suitable a song is for dancing based on a combination of musical elements	str
Energy	Perceptible measure of intensity and activity	float
Speechiness	Detects the presence of spoken words in a track	float
Artist count	Number of artists contributing to the song	int

TABLE II: Preliminiary subset features explanation

2) Hypothesis Generation: Based on the remaining metrics, several hypotheses were formulated to guide the preliminary analysis. The primary focus was on identifying potential correlations between these metrics and the target variable, Streams, which represents the total number of streams on Spotify. The hypotheses proposed were as follows:

- There is a correlation between BPM and Streams.
- There is a correlation between Artist count and Streams.
- There is a correlation between Danceability and Energy.
- There is a correlation between Speechiness and Streams.
- There is a correlation between Danceability and Streams.

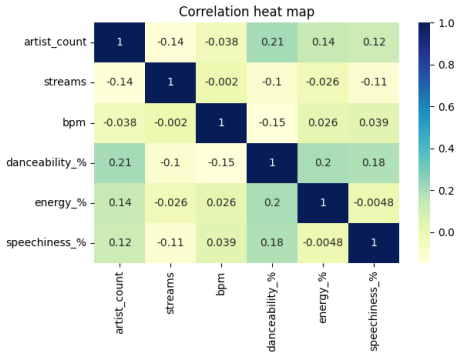

Fig. 1: Correlation heat map

Once we had the hypotheses, what was done was to prepare a python experiment, casting all data types to float for ease. With this, a simple correlation matrix was generated to determine whether or not there was a relationship between the data and, if so, if so. was the same one that posed the hypotheses, of these it was determined that the hypotheses were really negated because by analyzing (Figure 1) it was easy to determine that ”streams” really did not have signifcant correlations with the selected features. However, the negative correlation of ”speechines (%)” stood out above the others when analyzing the plots corresponding to each of the relationships (See Figure 2) from which the basis was taken to propose the following experiment.

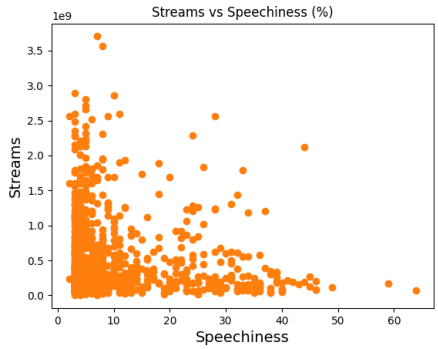

Fig. 2: Streams vs Speechiness (%)

3) Insights: Upon analyzing the correlations between the selected metrics and Streams using correlation matrices, most of the hypotheses were refuted. However, an intriguing observation emerged regarding the relationship between Speechiness and Streams. Despite the lack of significant correlations in other areas, a noteworthy correlation was found between these two variables. Further exploration of this relationship is warranted in subsequent analyses.

B. Exploratory Linear Regression Analysis – "Streams" and "Speechiness (%)".

1) Feature Selection: Building upon the insights gained from the preliminary analysis, a more focused examination was conducted to explore the relationship between Streams and Speechiness in greater detail. To facilitate this investigation, all previous features were removed, leaving only Streams and Speechiness in the subdataset.

2) Linear Regression Modeling: A linear regression model was then employed to quantify the relationship between Streams and Speechiness. This approach allowed for the estimation of how changes in Speechiness corresponded to changes in the number of Streams. By fitting a linear regression model to the data, the goal was to discern any discernible patterns or trends. See Figure 3

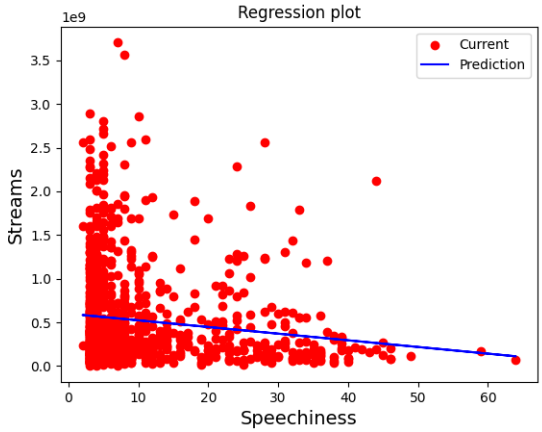

Fig. 3: Streams vs Speechiness

3) Implications: The observed negative correlation between Speechiness and Streams underscores the importance of considering lyrical content in the context of song popularity and listener preferences. While further investigation is warranted to fully understand the underlying mechanisms driving this relationship, the initial findings offer valuable insights for artists, record labels, and industry stakeholders. Understanding the impact of speech content on streaming metrics can inform content creation strategies and promotional efforts, ultimately enhancing the success of music releases on digital streaming platforms.

This subsection outlines the methodology employed and the key findings derived from the exploratory linear regression analysis, providing valuable insights into the relationship between Streams and Speechiness in the context of song virality on Spotify.

C. Deeper understanding – Taking all features in one correlation.

Due to the performance of the previous hypotheses, the need to add new hypotheses to the experimental study was generated. For this purpose, a separate set of tests was designed for this purpose. This set of tests is expanded in this section.

1) Objective: The objective of this experiment was to gain a comprehensive understanding of the relationship between each feature of the dataset and the target variable, Streams (See Figure 4). By conducting a complete correlation analysis of all the dataset, insights were sought to identify potential predictors of song virality on Spotify.

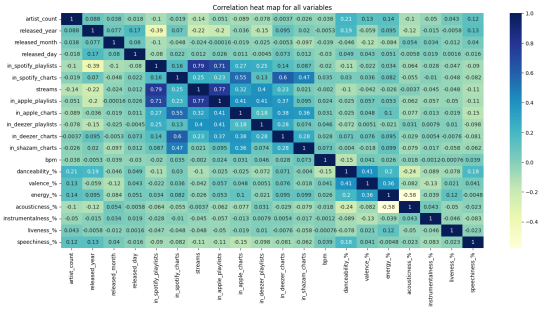

Fig. 4: Correlation of all features

2) Methodology: A correlation matrix was computed to assess the strength and direction of the relationship between each feature and Streams (See Figure 4). This analysis aimed to uncover patterns and correlations that could inform subsequent experiments and hypothesis generation. The correlation analysis revealed distinct patterns among the dataset features. Features were categorized into several groups based on their correlation with Streams. To generate these groups, the behavior of each of the "plots" that were generated by contrasting "streams" against each feature is carefully analyzed. Subsequently, they were labeled in a table and the groups with the lowest correlation with "streams" were discarded. The result of this process can be seen in the following table III

Groups	Description
Dispersed	Features such as BPM, Danceability (%), Valence (%), Energy (%), and Acousticness (%) exhibited a dispersed distribution of values, indicating a wide range of variation.
Concentrated to the left	Features including Instrumentalness (%), Liveness (%), and Speechiness (%) displayed a concentration of values towards the lower end of the scale.
Not so much variation	Features such as Artist count, Released month, and Released day showed limited variation across the dataset.
Predictable	Certain features, namely In Spotify playlist, In Spotify charts, In apple playlist, In apple charts, In Deezer playlists, In Deezer charts, and In Shazam charts, demonstrated a predictable relationship with Streams, suggesting a direct influence on song virality.
Special case (Concentrated to the right)	Released year exhibited unique characteristics, warranting further investigation into its relationship with Streams.

TABLE III: Preliminary subset features explanation

3) Implications: From the previous table, all groups were discarded except 1, the "predictable" group, precisely from that group a mathematical relationship significant enough to be able to predict something can be obtained; That "something" will depend on how strong the correlation is between the features and the objective, in this case the objective being "streams". This triggers a new hypothesis that is quite easy to understand: Can the group of "predictable" features predict whether a song will be a virtual hit?

The identification of distinct feature groups provides valuable insights into the factors influencing song virality on Spotify. Features categorized as dispersed or concentrated to the left may offer potential avenues for exploring correlations with Streams. Conversely, features categorized as predictable demonstrate a clear relationship with Streams, highlighting their importance in predicting song success.

D. Neuronal Network Analysis.

1) Objective: The objective of this experiment was to test several a neural network architectures (See table IV) to predict the behavior of songs on Spotify based on selected features including "In Spotify playlist", "In Spotify charts", "In apple playlist", "In apple charts", "In Deezer playlists", "In Deezer charts", "In Shazam charts", and "Released day". The neural network architecture was designed to perform regression tasks with the goal of predicting the number of Streams for a given song.

Epochs	AF	Input	FL	HL	DO	HL	DO	Output
5	linear	8	16	16	0.3	16	0.3	1
1000	linear	8	16	16	0.3	16	0.3	1
10000	linear	8	16	16	0.3	16	0.2	1
100000	linear	8	16	16	0.2	16	0.1	1
5	relu	8	16	16	0.3	16	0.3	1
1000	relu	8	16	16	0.3	16	0.3	1
10000	relu	8	16	16	0.3	16	0.2	1
100000	relu	8	16	16	0.2	16	0.1	1
5	mish	8	16	16	0.3	16	0.3	1
1000	mish	8	16	16	0.3	16	0.3	1
10000	mish	8	16	16	0.3	16	0.2	1
100000	mish	8	16	16	0.2	16	0.1	1

TABLE IV: Neuronal Network Architectures

A neural network schema was developed specifically to predict the behavior of songs on Spotify. The selected architecture consisted of 8 input nodes corresponding to the chosen features, followed by a sequence of two hidden layers, each comprising 16 nodes with dropout regularization. The output layer consisted of a single neuron for regression prediction.

2) Methodology: The neural network model was trained and evaluated using the designed architecture without aiming to compare different neural network architectures. The focus was on assessing the predictive capabilities of the model under various configurations of activation functions and training epochs. The dataset was split into 80% training and 20% evaluation sets, and the model's performance was evaluated using various metrics and visualizations.

Following each training session, the model's performance was assessed using metrics and visualizations such as loss through epochs, residual histograms, actual vs. predicted values, and predicted values vs. residuals. These insights provided a comprehensive understanding of the model's behavior and predictive accuracy. [1]

It is also important to keep in mind that this process was automated, that is; The machine was not supervised during the creation, training or analysis of the architectures; only the respective results were analyzed

[1]Note: You can find the referring plots at attached zip or running the code by yourself downloading it from GitHub, the repository is also attached in at references section. In this section just few plots are being showed

3) Implications: The results of the neural network analysis offer valuable insights into the predictive capabilities of the model for understanding the behavior of songs on Spotify. By leveraging selected features, the model can provide valuable predictions regarding the potential virality of songs on the platform. This has implications for artists, record labels, and industry stakeholders seeking to optimize promotional strategies and content creation.

IV. Test and Results

A. Preliminary Analysis

From a general point of view, it can be stated that the preliminary analysis was not very fruitful; Actually, from this we obtained more than anything the denial of the hypotheses that were raised; Well, even experiment #2 did not show a good way to predict how listened to a song will be, from this point of view its usefulness was mostly exploratory, which is fine for a preliminary analysis.

As mentioned above, when we carefully review the heat map (See figure 5) representing the correlation, we see that the negative or "cold" values according to the color scheme for "currents" marked with red, represent that the hypotheses are indeed false, since there are It's really not a significant correlation. With the most significant correlation we see being "Artist count", but that specific characteristic doesn't tell us much about how listened to a song may or may not be, the second most significant correlation turns out to be "Speechiness", which is the percentage of how many words are there basically in a song, from there it was deduced that there could be a relationship; That is why experiment 2 was carried out.

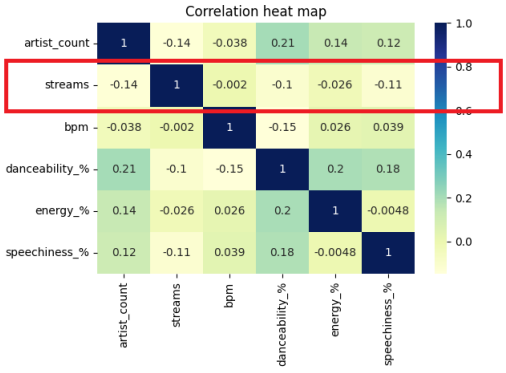

Fig. 5: Assertion #0 - No correlation

Upon delving deeper into the analysis, we uncovered a notable correlation between the variables of "Danceability," "Energy," and "Speechiness" and the number of artists involved (See figure 6). It is indeed surprising to find that an increase in the number of artists appears to correspond with a slight elevation in the speechiness and energy of the lyrics, alongside a marginal enhancement in the danceability of the song. This observation aligns with the notion that the presence of multiple voices in a song may lead to dynamic shifts in lyrical delivery, ranging from varying speeds to fluctuations between low and high energy levels.

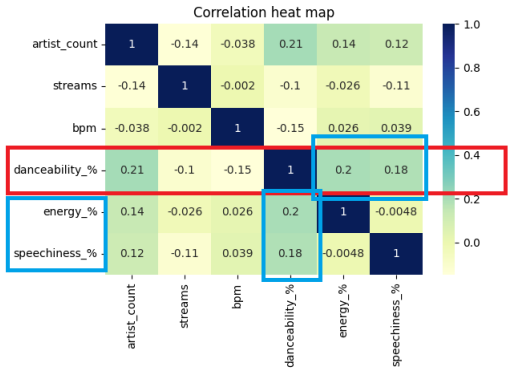

Fig. 6: Assertion #1

Furthermore, an intriguing observation emerged from the analysis, revealing a nuanced relationship between danceability, energy, and speechiness. The findings suggest that songs characterized by higher danceability also tend to exhibit a slight propensity towards increased energy levels and speechiness (See figure 7). This implies that the ability of a song to facilitate movement on the dance floor may be influenced by its energetic qualities and the presence of spoken words within the lyrics. Such insights shed light on the multifaceted nature of song attributes and their interplay in shaping the listener experience.

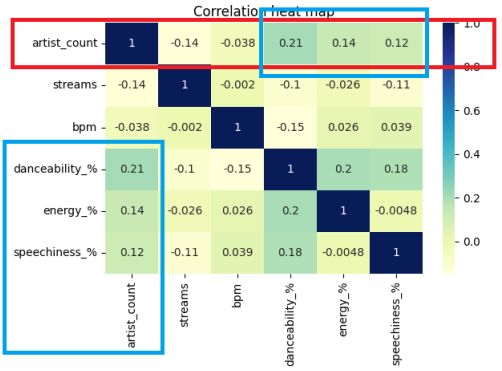

Fig. 7: Assertion #2

B. Exploratory Linear Regression Analysis – "Streams" and "Speechiness (%)".

The results of the linear regression analysis (See Figure 3) revealed an intriguing observation: a significant negative correlation between Speechiness and Streams. Contrary to initial expectations, the analysis indicated that songs with fewer spoken words tended to garner higher numbers of streams. This finding suggests that songs characterized by a higher degree of instrumental or non-vocal elements may be more appealing to listeners, leading to increased streaming activity.

C. Deeper understanding – Taking all features in one correlation.

The correlation analysis offers valuable insights into the relationship between dataset features and song virality on Spotify. The identification of feature groups and their respective correlations with Streams provides a foundation for hypothesis generation and further experimentation. This analysis sets the stage for subsequent investigations aimed at uncovering the underlying mechanisms driving song popularity and virality.

Derived from this analysis by groups, it is interesting to see the case of "Released year" (See Figure 8). This feature is easy to interpret, but it is the only one that presented a diametrically different behavior from the other groups; In this feature we see that there is a fairly strong relationship that indicates that songs from recent years are or have been listened to much more than songs from previous decades. This makes sense when you think about the complicated digitalization process; But the trend that this particular feature marks indicates that in order to determine how listened to a song will be, the decade in which it is released also seems to have a small impact, being a factor against older pieces of music.

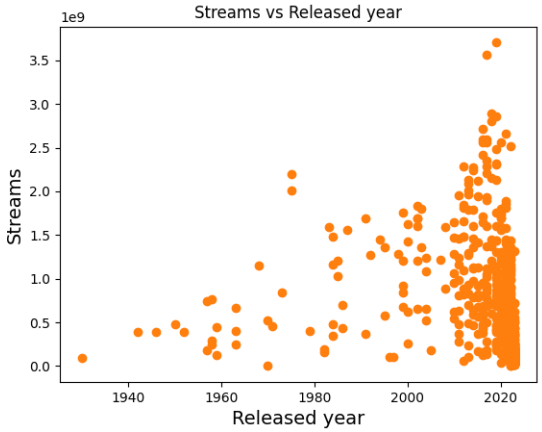

Fig. 8: Streams vs Released year

During the correlation analysis, distinct patterns emerged among the dataset features (See Figure 9), leading to their categorization into several discernible groups. Features such as BPM, Danceability (%), Valence (%), Energy (%), and Acousticness (%) exhibited a dispersed distribution of values, indicating a wide range of variation. Conversely, features including Instrumentalness (%), Liveness (%), and Speechiness (%) displayed a concentration of values towards the lower end of the scale, suggesting a more uniform distribution. Some features, such as Artist count, Released month, and Released day, showed limited variation across the dataset, indicating a more static or consistent nature. Additionally, certain features such as In Spotify playlist, In Spotify charts, In apple playlist, In apple charts, In Deezer playlists, In Deezer charts, and In Shazam charts demonstrated a predictable relationship with Streams, implying a direct influence on song virality. This categorization provides valuable insights into the dataset's structure and the potential relationships between features, offering a foundation for subsequent analysis and hypothesis generation. Along with the behavior those groups present seem to be a relation this report is leaving, future studies over "Top Spotify Songs" might deliver a better understanding of those unseen relations

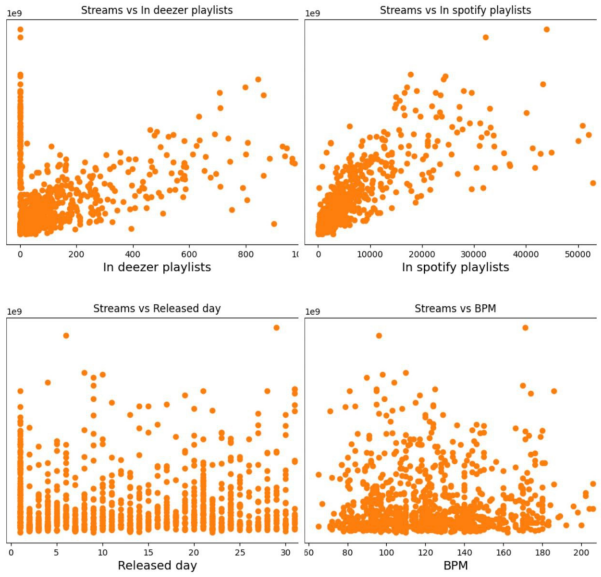

Fig. 9: Groups

D. Neuronal Network Analysis.

1) Difficulties: During the training process, it was observed that the model's performance deteriorated significantly as the number of training epochs increased, particularly around 10,000 and 100,000 epochs. This deterioration was evidenced by a sharp increase in the loss criteria, suggesting that the model struggled to generalize well to the data with higher epoch counts. This indicates a potential issue with overfitting and the lack of sufficient data to support the complexity of the neural network architecture (See Figure 10).

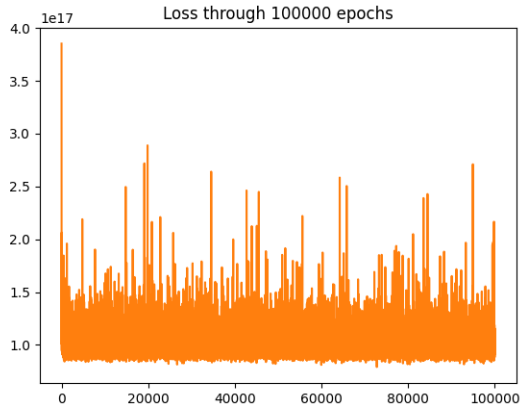

Fig. 10: Loss through 100000 epochs on linear - 16 x 16 x 16

An analysis of the performance of different activation functions revealed that the ReLU activation function exhibited the most promising results among the tested functions on loss through. The ReLU activation function demonstrated better stability and convergence compared to the Mish and linear activation functions. Conversely, the linear activation function exhibited the worst performance, particularly when trained for 100,000 epochs (See Figure 10). But contradictorily it has the best r2 (See Table V).

Epochs	AF	MAE	MSE	R2
5	linear	220597692.06	120390814422784112.00	0.56
1000	linear	201365060.52	89296573375715472.00	0.67
10000	linear	203227751.90	97686496965461808.00	0.64
100000	linear	196095259.94	82448820483056912.00	0.70
5	relu	219139710.47	135771922525852928.00	0.50
1000	relu	228234206.08	133880319975235968.00	0.51
10000	relu	214552697.13	125689641037188240.00	0.54
100000	relu	-	-	-
5	mish	226924229.03	127529536381550480.00	0.53
1000	mish	202578819.10	93475528415206848.00	0.65
10000	mish	200116363.02	100724239956045088.00	0.63
100000	mish	-	-	-

TABLE V: Testing metrics

During the testing process with 100,000 epochs, a peculiar issue was encountered where the neural network model produced NaNs (Not-a-Number) values. This issue resulted in blank plots and compromised the reliability of the model's predictions (See Figure 11). Further investigation is warranted to identify the root cause of this problem and address it effectively.

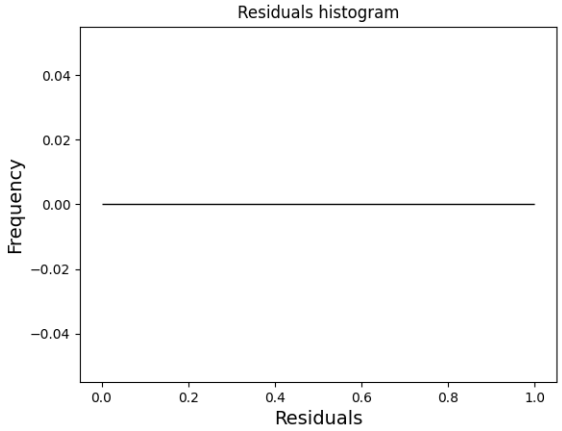

Fig. 11: Caption

Overall, the neural network experiment highlighted the challenges of training a model with limited data and the importance of selecting appropriate activation functions and training epochs. Despite encountering issues such as overfitting and NaNs, the experiment provides valuable insights into the performance of neural network models for predicting song behavior on Spotify. Further optimization and refinement of the model architecture and training process may be necessary to improve its performance and reliability.

2) Positive Findings: The neural network experiment demonstrated robust performance when utilizing the Rectified Linear Unit (ReLU) activation function. Compared to other activation functions such as Mish and linear, ReLU exhibited superior stability and convergence during training (Figure 12). This suggests that ReLU is well-suited for capturing the underlying patterns in the data and producing reliable predictions.

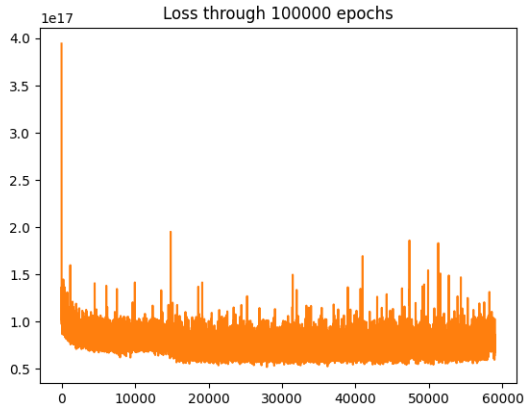

Fig. 12: Loss through 100000 epochs on ReLu - 16 x 16 x 16

Despite encountering challenges, the neural network experiment showcased the potential for improved predictive accuracy with further optimization. By fine-tuning the model architecture and training parameters, significant enhancements in performance can be achieved. This highlights the versatility and adaptability of neural network models in capturing complex relationships within the dataset.

The experiment provided valuable insights into the prediction of song behavior on Spotify using neural network models. By leveraging selected features, the model demonstrated the ability to predict song virality with a reasonable degree of accuracy. This has significant implications for the music industry, offering a data-driven approach to inform promotional strategies and content creation.

V. Conclusions

In conclusion, this report has provided a comprehensive analysis of song virality prediction on Spotify using a combination of exploratory data analysis, correlation analysis, and neural network modeling. Through meticulous examination of dataset features and rigorous experimentation with machine learning techniques, valuable insights have been gleaned into the factors influencing song behavior on digital streaming platforms. Despite encountering challenges such as data limitations and model performance issues, the findings offer promising avenues for further research and development. By leveraging advanced analytical approaches and refining predictive models, stakeholders in the music industry stand to benefit from enhanced understanding and predictive capabilities in promoting and optimizing content on platforms like Spotify. Moving forward, continued exploration and innovation in music analytics hold the potential to revolutionize the way songs are discovered, consumed, and propelled to success in the digital

References

AndresMpa. (2024). big_data: Lab_1 [GitHub repository]. Retrieved from https://github.com/AndresMpa/big_data/tree/main/task/lab_1