TikTok Music Popularity Prediction

Team #20 - Chuqi Fang, Jianjun Lei, Yaqi Jia, Charlaine Jo

1. Business Understanding

Identify and motivate the business problem

As one of the fastest-growing social media platforms worldwide, TikTok has taken the globe by storm in just a few years. By 2024, TikTok has surpassed 1 billion monthly active users globally, covering all age groups and regions. The platform's short video format, rich music library, and unique algorithmic recommendations have quickly attracted a large number of content creators and users. Due to the unique nature of TikTok's short videos, the choice of music becomes a key factor in determining the success of video content. Viral music can significantly boost video reach and user interaction. Therefore, predicting which songs have the potential to go viral is of great importance to both TikTok creators and the platform itself.

How Data Mining Solutions Solve Business Problems

We will develop and train a predictive model using historical data from TikTok's 2021 popular songs. By experimenting with different machine learning and data mining methods, we aim to find the most accurate model. The model will analyze various musical attributes, such as artist popularity, song energy, and tempo, to predict whether a song will go viral. The target variable of the model is the song's popularity (#track_pop), defined as a score from 0 to 100, with higher scores indicating greater popularity. By this model, TikTok can provide creators with valuable tools to help them choose the right music to increase their video appeal. This not only improves content quality and platform engagement but also allows TikTok to secure rights to trending music in advance, optimize its content recommendation system, and enhance advertising effectiveness. These measures will help TikTok continuously maintain its competitive edge, increase user retention, and drive direct business revenue, further strengthening its global influence and market leadership.

Business Use Cases

- Music Licensing Strategy Optimization, TikTok can secure trending song licenses in advance and analyze historical data to identify which music attributes or genres are more likely to go viral. This allows for more strategic licensing decisions, saving significant costs.
- Improved Ad Targeting, By predicting music trends, advertisers can select the most suitable songs for ad backgrounds or collaborate with trending tracks for campaigns. This increases ad relevance and appeal, boosting conversion rates and ad revenue.

2. Data Understanding

Data Sources

we will use a publicly available dataset from Kaggle titled “TikTok Popular Songs 2021”. This dataset contains information on songs that have been popular on TikTok, along with various musical attributes that may correlate with a song’s potential to go viral. The dataset includes both song metadata and features related to musical composition and popularity metrics. TikTok popular songs 2021 (kaggle.com)

Data Overview (Variable Description in Appendix)

18 columns (number of features of song)

190 observations (number of songs in the data)

Target variable

track_pop: Track Popularity, a numerical value ranging from 0 to 100, where higher numbers indicate greater popularity.

3. Data Preparation

1.Checking for Missing (Null) Values

We checked the dataset for missing values and found none, so no imputation was needed.

2.Removing Irrelevant Columns

We removed non-numeric columns (track_name, artist_name, album) as they were not useful for prediction.

3.Converting Categorical Variables

Categorical variables "mode" and "key" were converted into numeric format for use in modeling.

4.Data Standardization

We standardized the dataset to ensure all features have the same weight in the model. It helps improve stability and consistency of the models.

5.Data Splitting

We split the data into 80% training and 20% testing sets, ensuring reproducibility with a random seed.

4. Modeling

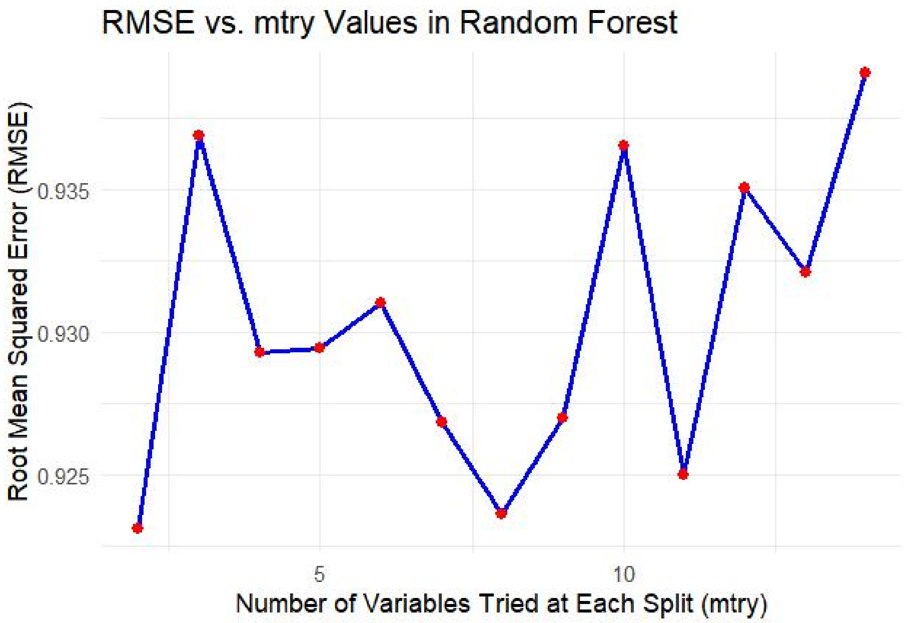
Types of Models We Built and Analysis for Each Model:

- **PCA + Linear Regression:**

This method reduces dimensionality, making the model easier to interpret. However, linear assumptions may not capture complex relationships and principal components may not always align with meaningful features. In our specific business problem, this model provides insights into which features are most predictive of song popularity and helps TikTok focus on the characteristics of successful songs (e.g., high energy, danceability).

- **Random Forest:**

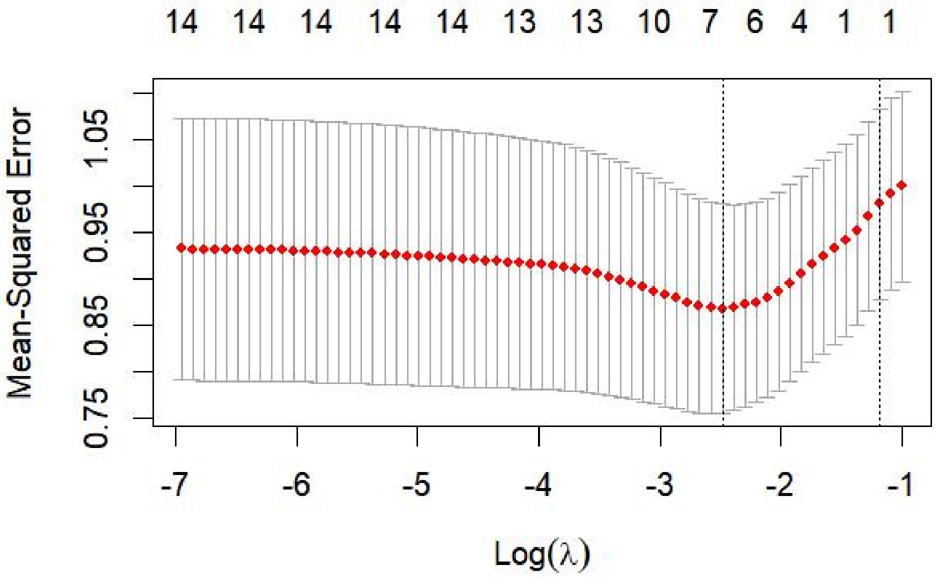
We test mtry from 2 to the number of features through 5-fold cross-validation.



The plot shows that the model has the best performance when mtry = 2 because it leads to the lowest RMSE.

The Random Forest model, which utilizes an ensemble approach, effectively captures complex interactions between features, leading to more robust predictions. The ability to average the results from multiple decision trees reduces variance and enhances the model's reliability in predicting song popularity; however this model is difficult to interpret. In our specific business problem, this model captures complex interactions and non-linear patterns that simpler models might miss, leading to more accurate predictions. By providing feature importance, it allows TikTok to identify key attributes of popular songs.

- **LASSO Regression:**



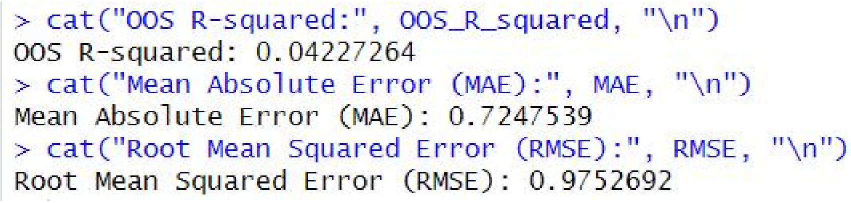
This plot helps us to find the best lambda, which is 0.083.

This model prevents overfitting, but only works well with linear regression. In our specific business problem, the function of this model is similar to PCA because they are both linear regression models, providing insights about feature selection and predictions.

- **Neural Network:**

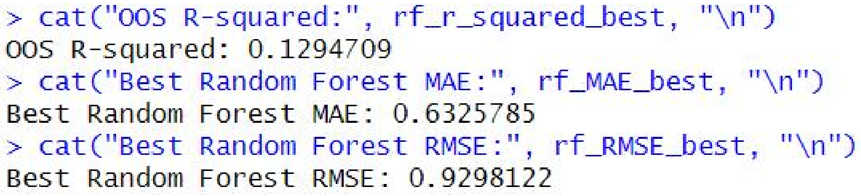
This model can address complex, non-linear relationships, but is computationally expensive and causes overfitting on small datasets. In our specific business problem, It aims to provide an accurate prediction so TikTok could focus on songs that will go popular.

5. Evaluation:



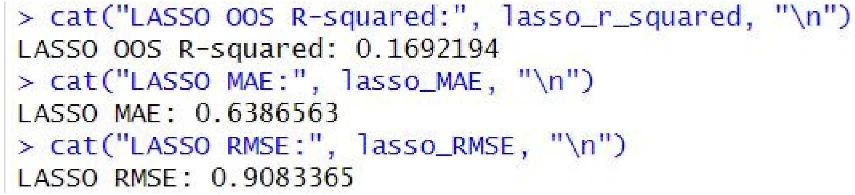
```
> cat("OOS R-squared:", OOS_R_squared, "\n")
OOS R-squared: 0.04227264
> cat("Mean Absolute Error (MAE):", MAE, "\n")
Mean Absolute Error (MAE): 0.7247539
> cat("Root Mean Squared Error (RMSE):", RMSE, "\n")
Root Mean Squared Error (RMSE): 0.9752692
```

(PCA Evaluation)



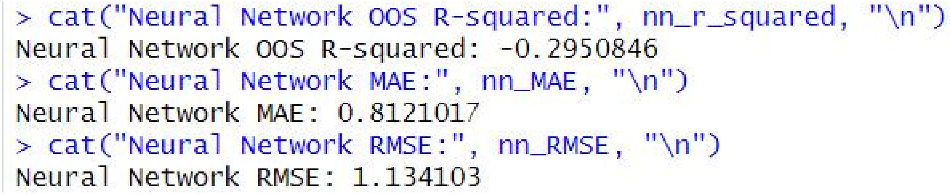
```
> cat("OOS R-squared:", rf_r_squared_best, "\n")
OOS R-squared: 0.1294709
> cat("Best Random Forest MAE:", rf_MAE_best, "\n")
Best Random Forest MAE: 0.6325785
> cat("Best Random Forest RMSE:", rf_RMSE_best, "\n")
Best Random Forest RMSE: 0.9298122
```

(Random Forest Evaluation)



```
> cat("LASSO OOS R-squared:", lasso_r_squared, "\n")
LASSO OOS R-squared: 0.1692194
> cat("LASSO MAE:", lasso_MAE, "\n")
LASSO MAE: 0.6386563
> cat("LASSO RMSE:", lasso_RMSE, "\n")
LASSO RMSE: 0.9083365
```

(Lasso Evaluation)

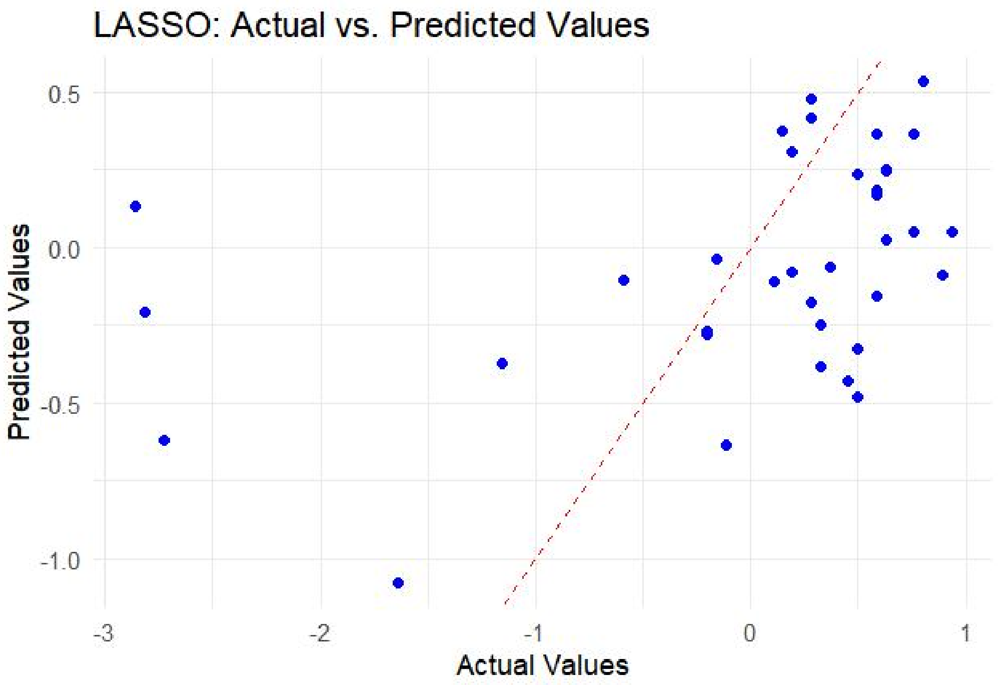


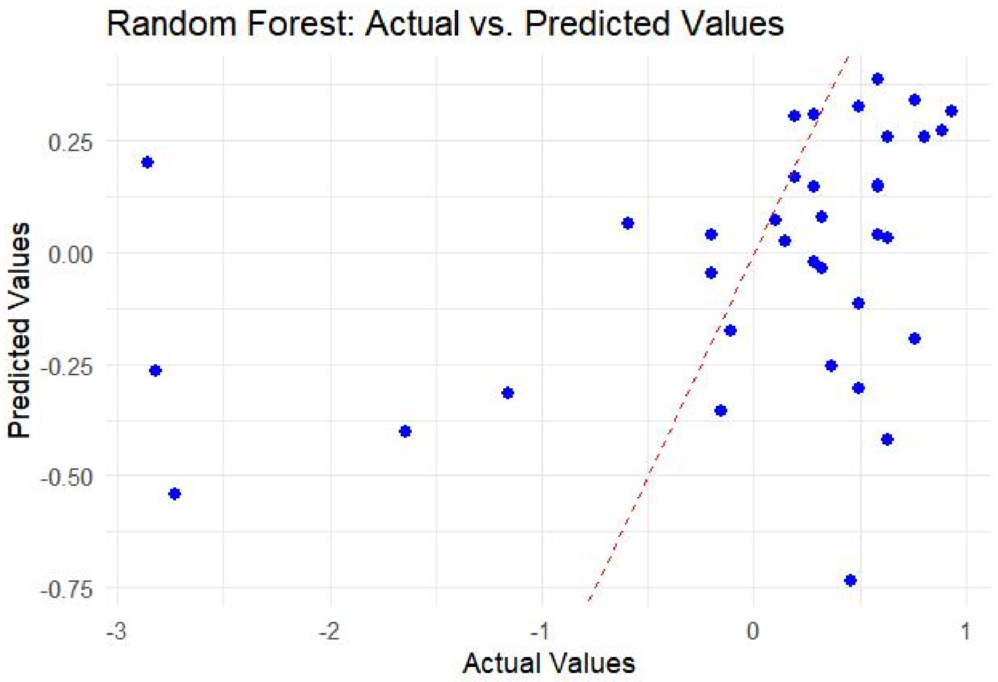
```
> cat("Neural Network OOS R-squared:", nn_r_squared, "\n")
Neural Network OOS R-squared: -0.2950846
> cat("Neural Network MAE:", nn_MAE, "\n")
Neural Network MAE: 0.8121017
> cat("Neural Network RMSE:", nn_RMSE, "\n")
Neural Network RMSE: 1.134103
```

(Neural Network Evaluation)

Among these models, Lasso Regression has the highest out-of-sample (OOS) R-squared(0.169), indicating it captures more variance than the other models; it also shows relatively low MAE(0.64) and RMSE(0.91), suggesting it makes reasonably accurate predictions.

Random Forest has the second highest OOS R-squared(0.129); and relatively low MAE(0.63) and RMSE(0.93).





From the plots, we have different insights. The Random Forest plot shows less spread in predictions, and more points are clustered closer to the diagonal line compared to LASSO, indicating that Random Forest makes more accurate predictions, capturing the relationships between features and the target variable better than LASSO. Furthermore, Random Forest is better at capturing non-linear interactions and feature dependencies, as shown by its closer alignment to the ideal diagonal.

These 2 models play different but complementary roles in solving our business problem - predicting the popularity of TikTok songs. LASSO Regression provides both feature selection and interpretability, helping us understand which attributes (e.g., energy, danceability) are most important. This model offers insights for targeted marketing and feature-based song recommendations. Random Forest provides reliable predictions and highlights feature importance. This model is helpful for TikTok to make accurate predictions and robust decision-making in a dynamic environment.

Accurate predictions help TikTok promote songs with higher popularity potential, leading to more streams and better user engagement. This can lead to higher revenue through increased ad impressions, subscriptions, or song purchases. However, calculating ROI precisely can be challenging due to the complexity of isolating the model’s impact from other factors and the long-term nature of some benefits. Streaming business models are also complex, making it hard to directly link predictions to revenue growth.

As an alternative, TikTok can track user engagement metrics like streams, shares, or playlist additions, which are easier to measure and closely tied to revenue. A/B testing can also help assess the impact of improved recommendations.

The low R-squared values across all models can be attributed to several factors. First, insufficient data limits the models' ability to capture meaningful patterns. Second, the lack of interaction terms may prevent the models from fully understanding complex relationships between features. Additionally, the target variable (song popularity) is inherently difficult to predict due to its dynamic nature. Finally, market behavior and user preferences are often influenced by numerous random or hard-to-quantify factors, further complicating accurate predictions.

6. Deployment

Result deployment

The model can be integrated into TikTok to instantly analyze song popularity. The platform can quickly act on high-potential tracks to save on licensing costs.

Deployment Issue

Model accuracy and Data Updates: Due to the rapid and unpredictable changes in music and fashion trends on TikTok, a strategy for regularly updating its models is essential. This includes retraining them with the latest data to ensure they remain valid and forward-looking.

Cost issues: TikTok will need to expand its computing and storage capabilities to handle the massive amount of data (e.g., millions of songs and user interaction data). Real-time data processing, model training, and large-scale predictions will incur high cloud computing costs, especially as the platform continues to grow. Developing the predictive model will require significant investment in skilled personnel, including data scientists, machine learning engineers, and software developers. Initial development, testing, and optimization will incur substantial costs.

Ethical Considerations

Bias and Fairness: The model may unintentionally favor certain music genres, languages, or well-known artists, which could marginalize lesser-known or niche artists. This lack of diversity in the promoted music could lead to a negative experience for both artists and users who prefer niche music.

Data Protection and Privacy: When processing user data, the model must comply with privacy regulations in different regions, such as the EU's GDPR and the US's CCPA. TikTok needs to ensure that data mining and usage do not violate users' privacy, and proper safeguards must be in place to protect sensitive information.

Risks and Mitigation

- **Data Security Breaches**

  Risk: Handling large volumes of user data and song interaction records makes TikTok a target for potential data breaches. Unauthorized access to personal data could result in significant financial and reputational damage.

  Mitigation: When processing user data, the model must comply with privacy regulations in different regions, such as the EU's GDPR and the US's CCPA. TikTok needs to ensure that data mining and usage do not violate users' privacy, and proper safeguards must be in place to protect sensitive information.

- **User Fatigue from Algorithmic Recommendations**

  Risk: If creators all use the model to predict and choose the most popular music, and everyone uses the same background music, TikTok users may receive repetitive music recommendations. This could lead to "algorithmic fatigue," where users become less engaged with the platform due to the lack of variety.

  Mitigation: Limit access to the model to a specific group of creators. Introduce diversity into the recommendation algorithm by offering a mix of personalized, random, and trending songs.

Appendix I: Variable Descriptions

Song Information:

track_name	The name of the song
artist_name	The name of the performing artist or group
album_name	The name of the album the song is part of
artist_pop	The popularity score of the artist

Musical Attributes:

Danceability	Suitability of the song for dancing (0-1 scale)
Energy	Intensity and activity level of the song
Loudness	Overall loudness of the song (measured in dB)
Key	Musical key of the song (0-11)
Mode	Major (1) or minor (0) key
Speechiness	Measure of spoken words in the track
Acousticness	Likelihood of the song being acoustic
Instrumentalness	Likelihood the song has no vocals
Liveness	Presence of a live audience in the recording
Valence	Measure of positivity in the song's sound
Tempo	Speed of the song (measured in BPM)
Time_signature	Beats per bar
Duration_ms	Length of the song (in milliseconds)

Appendix II: Team Member Contribution

Peter Lei	Data search Report writing: Business Understanding, Data Understanding, Data Preparation, Deployment Slides
Chuqi Fang	Data search Data preparation Model building Model tuning Report writing: Modeling, Evaluation Slides
Yaqi Jia	Data search Model tuning Report writing: Overall review and revision
Charlaine Jo	Data search