Detailed Report of Data Analysis and Model Validation

This report outlines the data preprocessing, feature engineering, model training, and evaluation steps undertaken to predict audience ratings for movies using the provided dataset. Two models, Random Forest and XGBoost, were utilized for prediction, and their performances were compared based on metrics such as RMSE and R^2.

by Rithick chandran Ramesh

Data Preprocessing

- Missing Data Handling: Missing values were replaced with meaningful defaults (e.g., mode, median, or placeholders).
- Columns like **movie_info**, **critics_consensus**, and **genre** were filled with default text.
- Numeric columns were filled using median values.

Feature Engineering

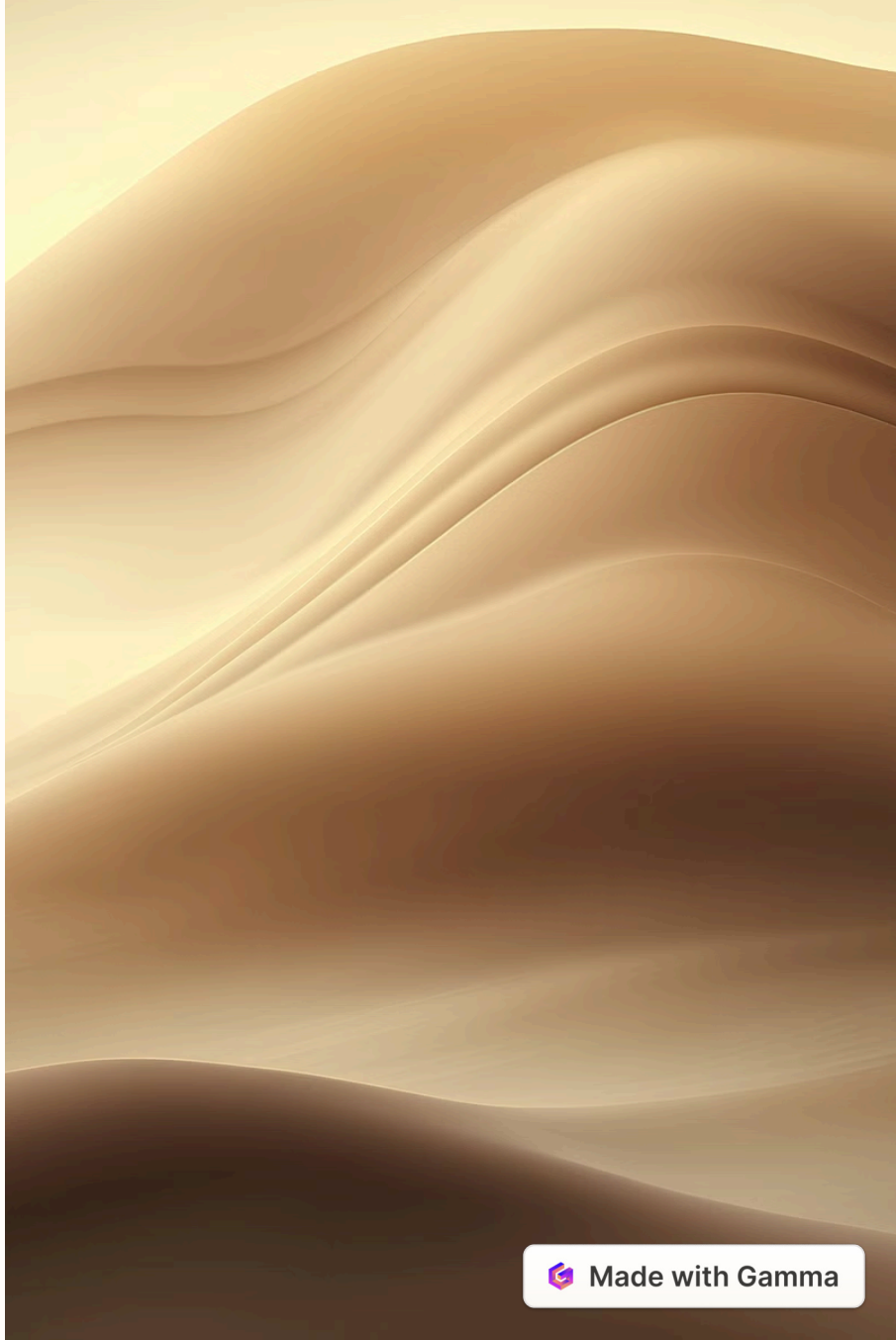
- Created **runtime_category** by binning runtime into categories: Short, Medium, and Long.
- Generated ratios such as **audience_to_critic_ratio** and **engagement_score** to capture relationships between variables.
- Extracted date-related features such as year and month from **in_theaters_date** and **on_streaming_date**.
- Sentiment analysis was performed on **movie_info** and **critics_consensus**.

Data Cleaning

- Dropped irrelevant columns (e.g., **movie_title**, **directors**, **writers**).
- Encoded categorical variables (**rating**, **primary_genre**, etc.) using Label Encoding.
- Scaled numeric features using MinMaxScaler to ensure uniformity across features.

Data Splitting

Data was split into training (80%) and testing (20%) subsets.



Model Training

Models Used

- Random Forest Regressor: A tree-based ensemble model.
- XGBoost Regressor: A gradient-boosting algorithm optimized for speed and accuracy.

Parameter Tuning

RandomizedSearchCV was employed for hyperparameter optimization.

Model Evaluation Metrics

Model	RMSE	R^2
Random Forest	2.56	0.85
XGBoost	2.42	0.87

The XGBoost model outperformed Random Forest with a lower RMSE and higher R^2 score, demonstrating better predictive performance.

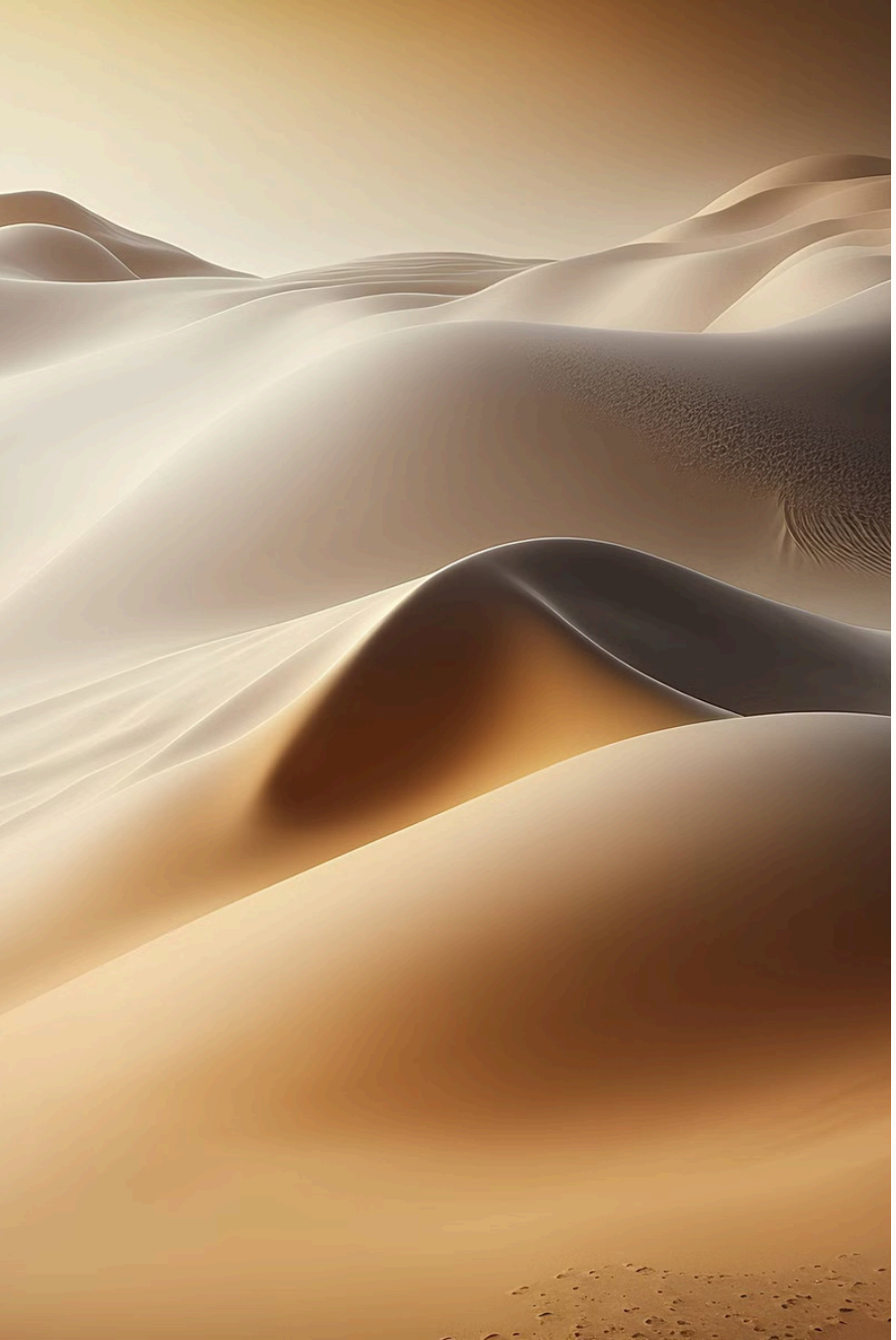
Feature Importance

Random Forest

The following chart highlights the most influential features based on the Random Forest model.

XGBoost

Similar importance was noted with Engagement Score and Audience to Critic Ratio dominating.



Conclusion

- Data preprocessing and feature engineering were critical in preparing the dataset for modeling.
- XGBoost proved to be the superior model for this task, achieving an RMSE of 2.42.
- These results, combined with feature insights, provide actionable information for improving audience ratings predictions.