Podcast Listening Time Analysis: Model Construction and Evaluation (Part 2)

Quantifying Listener Engagement through Statistical and Machine Learning Models

Table of Contents

[1. Business Problem 2](#_Toc200917424)

[2. Modeling Objective 2](#_Toc200917425)

[3. Model Development Framework 3](#_Toc200917426)

[Baseline Linear Regression Models 3](#_Toc200917427)

[Factor Re-leveling for Interpretability 3](#_Toc200917428)

[Train-Test Data Partitioning 3](#_Toc200917429)

[Model Diagnostics and Residual Analysis 3](#_Toc200917430)

[4. Feature Engineering and Transformation 4](#_Toc200917431)

[Binarization of Categorical Predictors 4](#_Toc200917432)

[Box-Cox Transformation of Response Variable 4](#_Toc200917433)

[5. Enhanced Linear Modeling 5](#_Toc200917434)

[Regression with Transformed Target 5](#_Toc200917435)

[Stepwise Feature Selection 5](#_Toc200917436)

[6. Regularization Techniques 6](#_Toc200917437)

[7. Advanced Modeling with XGBoost 7](#_Toc200917438)

[Variable Importance 7](#_Toc200917439)

[Partial Dependence Plots 7](#_Toc200917440)

[8. SHAP-Based Interpretability 8](#_Toc200917441)

[9. Residual Diagnostics for Model Robustness 9](#_Toc200917442)

[10. Comparative Evaluation of Models 11](#_Toc200917443)

[11. Final Remarks and Business Implications 12](#_Toc200917444)

[Strategic Recommendations: 12](#_Toc200917445)

# 1. Business Problem

With the rapid evolution of digital media consumption, podcast platforms are increasingly turning to data analytics to uncover the behavioral patterns that drive listener engagement. With monetization and content curation dependent on listener behavior, the ability to accurately predict Listening\_Time\_minutes—the total time a listener engages with an episode—is a competitive advantage.

The central business question is:

"Which podcast episode features significantly impact total listening duration, and how can this be accurately predicted and interpreted through statistical and machine learning models?"

Accurate forecasting enables:

* Tailoring episode structure to maximize retention.
* Scheduling releases for peak engagement.
* Optimizing advertisement placement without deterring audience attention.

# 2. Modeling Objective

The objective of this analysis is to develop and evaluate a suite of predictive models for Listening\_Time\_minutes using structured episode-level data. The focus lies on:

* Maximizing predictive accuracy.
* Improving model interpretability.
* Identifying key drivers of listener behavior.
* Ensuring generalization across unseen data.

# 3. Model Development Framework

## Baseline Linear Regression Models

Two preliminary models were developed:

* Model 1 (lm1): Utilized only numeric predictors.
* Model 2 (lm2): Included both numeric and categorical variables without preprocessing.

The initial models offered basic insights but suffered from issues such as non-standard factor references and possible multicollinearity.

## Factor Re-leveling for Interpretability

To improve interpretability and control baseline effects, categorical variables (Genre, Publication\_Day, Publication\_Time, and Episode\_Sentiment) were re-leveled. The most frequent category was set as the reference level, allowing regression coefficients to reflect deviation from the dominant baseline.

## Train-Test Data Partitioning

A stratified random sampling approach was employed to partition the data into training (75%) and test (25%) subsets. The means of the response variable across partitions were closely aligned, ensuring representativeness and minimizing sampling bias.

## Model Diagnostics and Residual Analysis

A full linear model (lm\_full) was fitted on the training data. Key observations included:

* Residual plots indicated heteroscedasticity and slight non-linearity.

Figure 1: Residual vs Fitted Plot from Linear Model

A graph of a number of values

AI-generated content may be incorrect.

* Variance Inflation Factor (VIF) values suggested multicollinearity among factor-expanded predictors.
* These insights motivated the use of regularization and transformation strategies.

# 4. Feature Engineering and Transformation

## Binarization of Categorical Predictors

Categorical predictors were transformed into binary indicators using one-hot encoding via dummyVars() (with full-rank encoding to eliminate redundancy). This allowed linear models to handle categorical inputs without introducing perfect multicollinearity.

## Box-Cox Transformation of Response Variable

To stabilize variance and normalize residuals, a Box-Cox transformation was applied to the target variable. The transformation took the form:

Figure 2: Histogram of Box-Cox Transformed Target

A graph showing a curve

AI-generated content may be incorrect.

The optimal lambda ) was selected by maximizing the Box-Cox log-likelihood profile. This transformation improved residual symmetry and reduced prediction error.

# 5. Enhanced Linear Modeling

## Regression with Transformed Target

Two refined models were fit:

* lm\_full\_bin: Response = untransformed target
* lm\_full\_bin\_bc: Response = Box-Cox transformed target

Predictions from the Box-Cox model were inverse-transformed for evaluation. This model exhibited superior predictive accuracy (lower RMSE) and more normal residuals.

## Stepwise Feature Selection

To balance complexity and performance:

* Backward selection (AIC) pruned less relevant variables from the full model.
* Forward selection (BIC) built a compact model starting from the null.

These reduced models offered similar accuracy to the full model, with improved interpretability.

# 6. Regularization Techniques

To further address multicollinearity and overfitting, penalized regression methods were employed using glmnet:

* Ridge Regression ( = 0): Shrinks coefficients, retains all predictors.
* Lasso Regression = 1): Performs variable selection by setting some coefficients to zero.
* Elastic Net ( = 0.5): Combines ridge and lasso penalties.

Cross-validation determined optimal .min and .1se values. Elastic Net with .min outperformed others on test RMSE, striking a good balance between model complexity and predictive power.

Figure 3: Cross-Validation Plot for λ Selection in Elastic Net

A graph of a function

AI-generated content may be incorrect.

# 7. Advanced Modeling with XGBoost

To model non-linear relationships and interactions, an xgbTree model was trained using 5-fold cross-validation and a comprehensive hyperparameter grid:

* max\_depth = 7, eta = [0.001–0.1], nrounds = [50–300]
* colsample\_bytree = 0.6, subsample = 0.6, gamma = 0

## Variable Importance

Top-ranked predictors based on XGBoost’s internal gain metric:

Figure 4: XGBoost Feature Importance Plot

A graph with numbers and a bar

AI-generated content may be incorrect.

* Episode\_Length\_minutes
* Number\_of\_Ads
* Host\_Popularity\_percentage
* Guest\_Popularity\_percentage
* Episode\_Number

## Partial Dependence Plots

PDPs revealed key non-linear trends:

Figure 5: Partial Dependence Plot for ‘Episode\_Length\_minutes’

A graph with a line

AI-generated content may be incorrect.

* Longer episodes drove higher engagement.
* More ads generally reduced listening time.
* Host popularity showed diminishing returns.

# 8. SHAP-Based Interpretability

SHAP (SHapley Additive exPlanations) values were computed to explain individual predictions from the XGBoost model.

Figure 6: SHAP Summary Plot

A graph with numbers and lines

AI-generated content may be incorrect.

* SHAP Summary Plot: Quantified each feature’s average impact on predictions.

Figure 7: SHAP Dependence Plot (e.g., Host\_Popularity\_percentage)

A graph with dots and lines

AI-generated content may be incorrect.

* Dependence Plots: Highlighted interactions and nonlinear relationships, e.g., between Host\_Popularity\_percentage and Episode\_Number.
* Both training and test SHAP analyses confirmed model robustness and consistency.

These insights offer actionable guidance for podcast producers to fine-tune content based on measurable impacts.

# 9. Residual Diagnostics for Model Robustness

For the XGBoost model:

* Residual vs Fitted Plot: Displayed no clear patterns, suggesting low model bias.

Figure 8: Residual vs Fitted Plot

A graph showing a line of black dots

AI-generated content may be incorrect.

* Q-Q Plot: Confirmed approximate normality of residuals.

Figure 9: Q-Q Plot

A line graph with a red line

AI-generated content may be incorrect.

* Histogram: Suggested symmetric residual distribution with light tails.

Figure 10: Histogram of residuals

A graph of a histogram

AI-generated content may be incorrect.

Overall, residual behavior supported the reliability of predictions across the full response range.

# 10. Comparative Evaluation of Models

| **Model** | **RMSE (Test)** |
| --- | --- |
| Linear Regression (Full) | 10.22638 |
| Box-Cox Transformed | 10.40447 |
| Stepwise AIC | 10.22638 |
| Stepwise BIC | 10.22708 |
| Ridge Regression | 12.24826 |
| Lasso Regression | 14.48427 |
| Elastic Net (λ.min) | 10.22876 |
| XGBoost | 9.870905 |

Figure 11: RMSE Comparison Bar Chart

A graph with blue lines

AI-generated content may be incorrect.

# 11. Final Remarks and Business Implications

This study systematically explored predictive modeling approaches for podcast listener engagement. Key conclusions include:

* Feature transformations and one-hot encoding significantly improved model accuracy and residual normality.
* Elastic Net regression offered a balance of generalization and parsimony.
* XGBoost delivered the highest accuracy, and SHAP values made this model interpretable and actionable.

## Strategic Recommendations:

* Optimize episode length and minimize unnecessary ads to retain listeners.
* Prioritize morning releases and high-performing genres like Health and True Crime.
* Leverage host popularity effectively but avoid over-reliance.

Future work may include real-time deployment, integration with podcast production tools, and A/B testing of content strategies informed by model outputs.

Prepared by: Supreet Mutsuddi  
GitHub Repository: https://github.com/Supreet1982/Podcast