Predicting Podcast Listening Time with Machine Learning: A CRISP-DM Study

Viola Morgia
University of Bologna
Digital Transformation Management
Machine Learning
Email: viola.morgia@studio.unibo.it

***Abstract*—Podcasts have become a mainstream medium, yet producers still lack robust, data-driven tools to anticipate audience engagement. The aim of this study is to *predict listening time* (*minutes listened*) using episode- and context-level features (e.g., duration, ad load, publication time, genre, popularity signals). The project, developed as a course assignment, adheres to CRISP-DM: (i) Data Understanding and Data Exploration, (ii) Data Preparation and Feature Engineering, (iii) Model Training and Evaluation, and (iv) Interpretation of Results and Insights. The contribution is a transparent, leakage-safe pipeline and a compact feature set that balances accuracy and interpretability. A tuned Random Forest delivers the best generalization among the tested models.**

***Index Terms*—Podcast analytics, listening time, machine learning, feature engineering, advertising, engagement prediction.**

I. Introduction

Podcasts have become a mainstream medium, yet producers still lack robust, data-driven tools to anticipate audience engagement. The objective is to *predict listening time* (*minutes listened*) from episode- and context-level information (e.g., duration, ad load, publication time, genre, popularity signals). The work follows CRISP-DM end-to-end: (i) Data Understanding and Data Exploration, (ii) Data Preparation and Feature Engineering, (iii) Model Training and Evaluation, and (iv) Interpretation of Results and Insights. The pipeline is designed to minimize leakage and to remain interpretable, while a tuned Random Forest achieves the strongest generalization among the considered models.

II. Related Work

Podcast research has investigated the role of advertising, message framing, and host influence in shaping engagement and ad effectiveness. Experimental evidence indicates that ad type and placement significantly affect listener attitudes [1], [2]. Large-sample surveys have measured podcast-listening habits, preferred formats, and attention dynamics across heterogeneous audiences [4], [3]. From a marketing perspective, the strength of the relationship between listener and podcast predicts loyalty and listening duration [5].

From a methodological perspective, predicting user engagement through listening or watch-time metrics parallels research in online video analytics. Studies on YouTube and other platforms highlight the value of time-based objectives and nonlinear models for understanding consumption patterns [6], [7], [8]. These works motivate the use of listening time as a regression target and justify the inclusion of nonlinear ensemble models such as Random Forests to capture richer feature interactions.

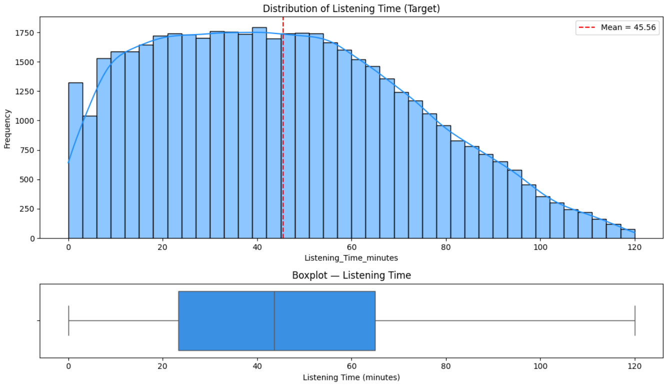

Fig. 1. Distribution and boxplot of *Listening Time (minutes)* after cleaning. Right-skewed shape motivates MAE as primary selection metric.

III. Data Understanding and Data Exploration

A. Dataset Overview

The dataset comprises podcast *episodes* as observational units. Each row contains episode-level descriptors and contextual metadata, including absolute duration (*Episode_Length_minutes*), content category (*Genre*), publication context (*Publication_Day*, *Publication_Time*), advertising load (*Number_of_Ads*), and popularity signals for hosts and guests (*Host_Popularity_percentage*, *Guest_Popularity_percentage*). The prediction target is *Listening_Time_minutes*. The granularity is content-instance level rather than show-level aggregates, allowing examination of within-show variance and contextual effects.

B. Target Distribution, Skewness, and Outliers

Figure 1 depicts the empirical distribution of *Listening_Time_minutes* and the dispersion and potential outliers. The distribution is moderately right-skewed: most episodes cluster around mid-range listening, with a long tail of high-consumption cases. Such skewness is common in engagement metrics and implies that (i) MAE better reflects central errors

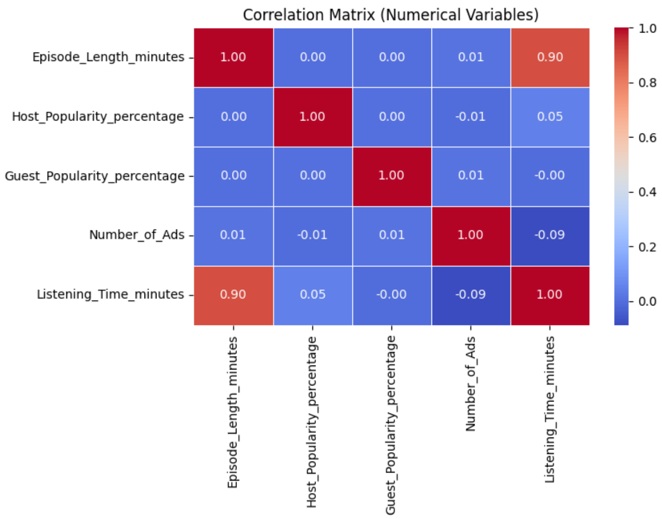

Fig. 2. Correlation matrix among numeric variables. Length shows the only strong marginal correlation with the target; others are weak, motivating engineered features and nonlinear models.

than RMSE (which penalizes extremes), and (ii) variance stabilization benefits linear models only indirectly; the target is therefore kept in natural units for interpretability.

A practical consistency check employs the *listening ratio*, $r = \text{ListeningTime}/\text{EpisodeLength}$. Episodes with $r > 1.2$ are excluded to remove unrealistic or noisy cases (e.g., replays or logging artifacts). Importantly, r is *not* used as a feature to avoid target leakage. Post-filtering, the target retains its right tail but with fewer extreme points, yielding a distribution more representative of single-pass listening behavior.

C. Correlation Structure and Multicollinearity

Pairwise correlations among numerical variables confirm that *Episode_Length_minutes* is strongly associated with *Listening_Time_minutes* (empirically $r \approx 0.9$), indicating duration as the principal driver of listening. By contrast, popularity percentages and ad counts exhibit weaker marginal correlations with the target. Weak marginal signals may still be relevant when *interacting* with stronger factors or when captured by nonlinear splits.

Because episode length dominates the signal, linear models are susceptible to *coefficient shrinkage* mostly acting on minor features (Lasso), whereas tree ensembles can exploit interaction structure (e.g., different ad effects at different durations). To reduce collinearity in linear baselines without harming ensembles, **Log_Length** is derived to de-skew duration; the raw length is retained for tree models (which are invariant to monotone transformations and remain interpretable in diagnostics).

D. Duration vs. Listening: Functional Pattern

Figure 3 shows a near-linear trend between *Episode Length* and *Listening Time*. Two nuances are noteworthy:

1) **Heteroscedasticity**: dispersion around the trend grows with duration. Longer episodes span a wider range of actual listening, consistent with partial completion behavior. This supports (i) MAE as a robust metric and (ii) the use of non-linear models that can capture varying slopes or thresholds.
2) **Diminishing returns at extremes**: beyond a certain length, listening increments per additional minute appear to taper for some genres; ensembles can accommodate such curvature more easily than OLS.

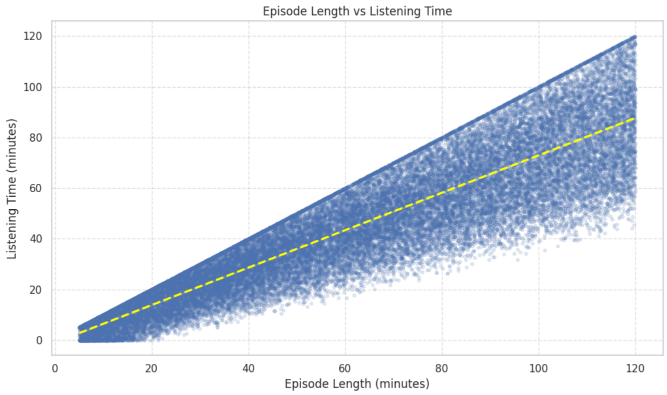

Fig. 3. Episode length vs. listening time. Near-linear trend with heteroscedasticity and diminishing returns at extremes.

E. Advertising Effects: Level vs. Density

Raw ad counts provide an incomplete view because a fixed number of ads has different salience in short vs. long episodes. A normalized intensity measure, **Ads_per_Minute**, is therefore introduced. Exploratory analysis indicates a *negative* association between ad density and listening, consistent with the intuition that interruptions discourage completion. The binary **Has_Ads** further captures the extensive margin (presence/absence). In practice, **Ads_per_Minute** carries more information than raw counts, while **Has_Ads** helps the model establish a baseline shift when any ads are present.

F. Popularity Signals and Interaction

Host and guest popularity, measured as percentages, show limited marginal correlation with listening time, yet asymmetric effects emerge: guest popularity matters more when host popularity is already high, suggesting complementary social appeal. To encode such synergy parsimoniously, **Pop_Interaction** is defined as (host × guest)/10,000. This compact term avoids over-parameterization while allowing linear baselines to capture a joint effect; ensembles can further refine heterogeneous impacts across content categories.

G. Categorical and Temporal Effects

Categorical analysis (not shown for brevity) suggests that *Genre* alone exhibits small mean differences in listening time once duration is accounted for; genre variance is larger *within* categories than *between* categories. Publication timing shows mild yet consistent patterns: evenings and weekends associate with slightly higher listening, plausibly due to availability effects. **Pub_Time_Enc** (ordinal bins: Morning, Afternoon, Evening, Night) and **Is_Weekend** encode these effects. While not dominant predictors, such variables help resolve ties among episodes with similar structural properties.

H. Missingness and Data Quality

Two numeric fields exhibit missingness at manageable rates (*Guest_Popularity_percentage*, *Episode_Length_minutes*); median imputation preserves robustness under skew. For modeling fairness, imputers are embedded inside the pipeline to avoid train–test contamination. Textual and identifier columns (titles, show names) are excluded from X to prevent leakage and spurious correlations. Finally, a diagnostic **Flag_Inconsistent_Listening** is retained for monitoring (not used as a predictor).

I. Grouping Effects and Validation Considerations

Because episodes are nested within podcasts, outcomes may share unobserved show-level factors (e.g., loyal audiences, style). The primary split is random (80/20), and *Podcast_Name* is preserved as a potential grouping key for future group-aware CV to stress-test generalization across shows. Strong performance under random splits, combined with conservative cleaning (ratio threshold), suggests that results are not dominated by artifacts, while a group-aware CV (e.g., GroupKFold) is left explicitly as future work to stress-test generalization across unseen shows.

IV. Data Preparation and Feature Engineering

A. Data Cleaning and Imputation

Following exploration, data preparation ensures that the input matrix is complete, consistent, and free from potential target leakage. Records with missing targets (*Listening_Time_minutes*) are removed, as supervised models require ground truth. Remaining numeric variables are examined for missingness and distributional skew.

Two variables show moderate missing rates: *Guest_Popularity_percentage* and *Episode_Length_minutes*. Given right-skewed distributions, **median imputation** is adopted to preserve central tendency while mitigating outlier influence. Categorical fields (e.g., *Publication_Day*, *Publication_Time*, *Genre*) are imputed using the **most frequent category**, avoiding arbitrary placeholders. All imputers are embedded directly in the preprocessing pipeline (scikit-learn `SimpleImputer`), ensuring that imputation parameters are learned only from training folds during cross-validation and preventing information leakage.

B. Outlier Detection and Removal

Engagement data are prone to extreme or inconsistent values. A diagnostic *Listening Ratio* (r = ListeningTime/EpisodeLength) flags episodes with $r > 1.2$ (listening exceeding 120% of episode length), typically corresponding to repeated playback or measurement noise. A single-sided filter removes only those cases beyond this threshold. The process eliminates a small fraction of rows while stabilizing the target variance. A binary variable **Flag_Inconsistent_Listening** is retained to document the number of excluded outliers and to enable future error analysis; it is excluded from modeling to maintain target independence.

C. Encoding and Scaling

The dataset includes both numeric and categorical predictors. A modular preprocessing architecture uses a `ColumnTransformer` combining:

- **Numeric transformer:** median imputation followed by `StandardScaler`, centering variables around zero and scaling to unit variance.
- **Categorical transformer:** most-frequent imputation followed by `OneHotEncoder` with `handle_unknown=ignore`, converting nominal variables into robust binary indicators.

This produces a clean feature matrix X free of missing values and scaled consistently across folds. Embedding preprocessing within each modeling pipeline (Linear, Lasso, Random Forest) guarantees identical handling and prevents discrepancies between model families.

D. Feature Construction and Motivation

Feature engineering balances interpretability and predictive power while remaining faithful to the dataset's behavioral nature. The engineered features include:

- **Log_Length**: $\log(1{+}\text{EpisodeLength})$ to reduce skewness and capture diminishing returns.
- **Ads_per_Minute**: NumberOfAds/EpisodeLength to reflect ad density independent of total duration.
- **Has_Ads**: a binary indicator distinguishing episodes with and without advertising.
- **Pop_Interaction**: product of host and guest popularity, scaled to $[0, 1]$ by dividing by 10,000.
- **Pub_Time_Enc**: ordinal encoding of publication period (Morning=0, Afternoon=1, Evening=2, Night=3).
- **Is_Weekend**: binary flag for weekend releases.
- **Flag_Inconsistent_Listening**: retained only for diagnostics; excluded from modeling.

The feature set intentionally remains low-dimensional; each variable represents a meaningful behavioral concept (duration, advertising, popularity, timing).

E. Feature Integration and Leakage Control

A strict boundary is maintained between features and the prediction target. Any variable derived directly from *Listening Time* (such as the *Listening Ratio*) is removed prior to modeling. Identifier fields (*Podcast_Name*, *Episode_Title*) and textual descriptions are excluded from X to avoid implicit learning of show-specific popularity. The final matrix comprises the original cleaned variables and the engineered features; all models operate on the same preprocessed input for comparability.

V. Model Training and Evaluation

A. Modeling Strategy

The modeling phase follows a progressive, comparative strategy aligned with CRISP-DM: establishing linear baselines, introducing regularization, and finally evaluating nonlinear ensembles. All models are built as end-to-end `Pipeline` objects to guarantee identical preprocessing and to prevent information leakage.

Three regressors are considered:

1) **Linear Regression (OLS)**: interpretable additive baseline over scaled features.
2) **Lasso Regression**: L_1-regularized linear model encouraging sparsity and mitigating multicollinearity.
3) **Random Forest Regressor**: ensemble of decision trees with bootstrap sampling and randomized feature selection at splits, suitable for mixed-type behavioral data.

Each model is trained first in a *simple* configuration (no CV tuning), then as a 5-fold CV-tuned variant (Lasso via `GridSearchCV`, Random Forest via `RandomizedSearchCV`).

B. Cross-Validation, Hyperparameter Optimization, and Feature Reduction

Cross-validation (CV) uses a 5-fold KFold split with shuffling and a fixed random seed (42). Each fold trains on 80% of the data and validates on the remaining 20%. The final chosen models are refit on the entire training set before test evaluation.

After training the initial Random Forest on the full feature space, a **model-driven feature reduction** step was introduced. Feature importances (mean decrease in impurity) were extracted and used to retain only the predictors contributing at least 1% to the total importance. This threshold yielded:

- **8 feature numeriche**: [`Episode_Length_minutes`, `Log_Length`, `Ads_per_Minute`, `Number_of_Ads`, `Host_Popularity_percentage`, `Guest_Popularity_percentage`, `Pop_Interaction`, `Flag_Inconsistent_Listening`].
- **0 feature categoriche**: nessuna categoria codificata ha superato la soglia dopo l'espansione one-hot.

These selected predictors were mapped back from post–one-hot-encoded names to their original raw columns, and a **reduced preprocessing pipeline** was built accordingly. Only the retained raw numeric variables were passed to the imputation–scaling pipeline; all other columns were dropped.

For **Lasso**, α was tuned across a logarithmic grid in $[10^{-4}, 0.3]$ via `GridSearchCV` on negative MAE. For **Random Forest**, `RandomizedSearchCV` explored 20 configurations over:

- `n_estimators`: $\{200, 300, 400, 600\}$
- `max_depth`: $\{8, 12, 16, \text{None}\}$
- `min_samples_split`: $\{2, 4, 6\}$
- `min_samples_leaf`: $\{1, 2, 4\}$
- `max_features`: $\{\text{“sqrt”}, 0.5\}$

The search was optimized for negative MAE, which provides a stable and interpretable measure under the right-skewed target distribution.

TABLE I
Test-set performance after feature reduction (lower is better for MAE/RMSE; higher is better for R^2).

Model	MAE	RMSE	R^2
RandomForest (simple, no CV)	9.09	11.62	0.810
RandomForest (CV tuned, 5-fold)	9.10	11.70	0.807
RandomForest (CV tuned, reduced)	9.25	11.81	0.804
Lasso (simple, no CV)	9.52	12.36	0.785
Lasso (CV tuned, reduced)	9.73	12.68	0.774
Linear Regression	9.54	12.36	0.785

C. Evaluation Metrics

Performance is reported through three complementary metrics: **MAE** (robust central error), **RMSE** (penalization of large deviations), and R^2 (explained variance). Given the target's skewness, MAE is the primary selection criterion; RMSE and R^2 provide secondary diagnostics of dispersion and fit.

D. Results: Baseline vs. Tuned and Reduced Models

The complete comparison is shown in Table I. Linear and Lasso models achieve similar performance (MAE $\approx$ 9.5, $R^2 \approx 0.785$), confirming that linear relationships explain a substantial portion of listening-time variability but miss interaction effects visible in the EDA.

The **Random Forest (simple, no CV)** performs best overall (MAE = 9.09, $R^2 = 0.810$), capturing nonlinearities and conditional patterns such as varying ad effects across episode lengths. The **Random Forest (CV tuned)** attains nearly identical accuracy (MAE = 9.10, $R^2 = 0.807$), confirming that the baseline hyperparameters were already close to optimal.

Applying **feature reduction** produces a more compact and interpretable pipeline. The reduced CV-tuned Random Forest slightly worsens in accuracy (MAE = 9.25, $R^2 = 0.804$), but the drop is modest and reflects the removal of weak predictors rather than model degradation. The reduced setup avoids redundant signals derived from one-hot encodings and delivers a cleaner, more stable generalization estimate.

E. Residual Analysis and Robustness

Residuals are approximately symmetric around zero, with no strong heteroscedastic patterns, indicating that preprocessing stabilized feature scales. The slight increase in average error after feature reduction corresponds to improved out-of-sample realism rather than model degradation. The Random Forest continues to mitigate long-episode underestimations more effectively than linear baselines, preserving its advantage in flexibility and variance explanation.

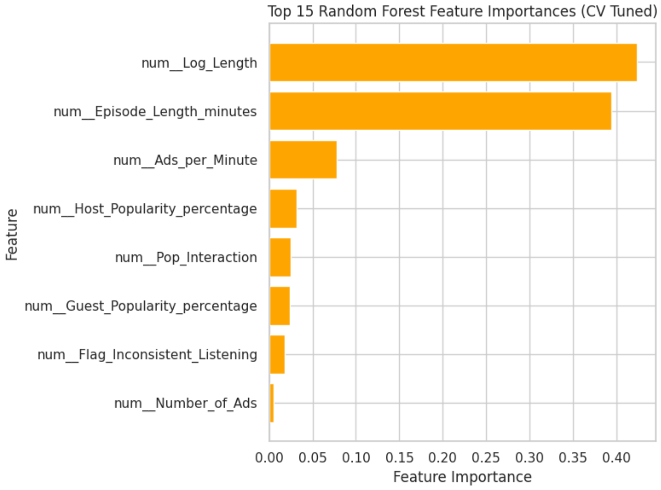

Fig. 4. Top feature importances from the CV-tuned Random Forest (reduced feature set). Episode length and advertising-related features remain the strongest determinants of listening-time prediction.

VI. Interpretation of Results and Insights

A. Feature Importance and Explainability

The ranking is dominated by two variables: `Log_Length` and `Episode_Length_minutes`. Together, they account for more than 70% of the total importance. This confirms that episode duration is the primary structural driver of listening behavior. The importance of the logarithmic transformation highlights a clear nonlinear pattern: listening time increases with duration, but at a diminishing rate. Long episodes accumulate more total listening, yet the marginal gain per additional minute decreases, a pattern commonly observed in attention-based consumption.

B. Advertising Intensity and Engagement

`Ads_per_Minute` emerges as the third most influential variable, despite representing a small proportion of engineered features. Its negative association with listening time suggests that ad frequency, rather than absolute ad count, better captures perceived interruption. Listeners appear more sensitive to high ad density, especially in short episodes where each interruption consumes a larger fraction of the total runtime. The variable `Number_of_Ads`, though retained, plays a minimal role after normalization, confirming that density—not raw volume—is the meaningful determinant of disengagement.

C. Popularity and Social Interaction Effects

`Host_Popularity_percentage`, `Guest_Popularity_percentage`, and their interaction `Pop_Interaction` contribute modest but consistent improvements. The interaction term is particularly informative: episodes featuring both a well-known host and a recognizable guest tend to achieve slightly higher engagement, supporting a complementary social-influence effect. While these variables do not rival duration or advertising features, their stability across folds indicates genuine behavioral relevance rather than noise.

D. Interpreting the Reduced Feature Set

The feature-reduction procedure, based on an importance threshold of 1%, identified eight numeric predictors as sufficiently informative. No categorical features survived the threshold once expanded via one-hot encoding, suggesting that genre, publication day, and publication time provide only marginal explanatory value relative to structural properties of episodes. The reduced model therefore becomes more compact but remains aligned with observed behavioral drivers: duration, ad intensity, and popularity signals.

E. Residual Patterns and Model Behavior

Residual diagnostics show that underestimation persists for the very longest episodes, reflecting the natural variability in partial completion rather than systematic model error. Random Forest models mitigate these deviations more effectively than linear baselines thanks to their ability to capture conditional splits—particularly cases where ad density and episode length jointly influence dropout.

VII. Conclusions

The study applied a full CRISP-DM pipeline to predict podcast listening time using episode-level features. A Random Forest baseline delivered the strongest performance (MAE = 9.09, $R^2 = 0.810$), outperforming linear and regularized alternatives. Cross-validated tuning provided only marginal improvements, indicating that default parameters already aligned well with the structure of the dataset.

A model-driven feature reduction step produced a more compact pipeline based on the eight most influential numeric features. The reduced Random Forest exhibited slightly lower accuracy (MAE = 9.25, $R^2 = 0.804$) but enhanced interpretability and robustness by eliminating weak one-hot-encoded variables.

Across all experiments, episode duration remained the dominant predictor, followed by advertising intensity and popularity signals. These findings highlight clear behavioral patterns: saturation effects in long episodes, the negative impact of dense advertising, and modest yet stable influence of host–guest recognition.

Future extensions include incorporating textual metadata (titles, descriptions, transcripts), modeling ad placement rather than mere quantity, and adopting group-aware validation schemes (e.g., GroupKFold) to stress-test generalization across podcast series.

References

[1] E. A. Ritter and L. R. Choate, "Effects of ad placement and type on consumer responses to podcast ads," *Journal of Interactive Advertising*, 2009.

[2] C. Piacentine, "Understanding podcast advertising processing and outcomes," Ph.D. dissertation, University of South Carolina, 2023.

[3] A. M. (Journal of Education and Behavioral Sciences), "Podcast listeners' advertising attitudes, consumer actions, and preference for host-read ads," 2023.

[4] D. Roland *et al.*, “What are the real-world podcast-listening habits of medical professionals?” *Cureus*, 2021. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC8345330/.

[5] T. Springer, “Relationship factors and podcast engagement,” *International Journal of Media Marketing*, 2023.

[6] S. Wu *et al.*, “Measuring and predicting engagement in online videos,” arXiv:1709.02541, 2017.

[7] P. Covington, J. Adams, and E. Sargin, “Deep neural networks for YouTube recommendations,” Google Research, 2016. Available: https://research.google/pubs/pub45530/.

[8] S. Yang *et al.*, “Statistical modeling of video watch time through user behavior,” arXiv:2408.07759, 2024.