Podcast Listening Time

Can we predict the listening time of a podcast based on specific listening factors?

Viola Morgia – Machine Learning Digital Transformation Management

Project objective

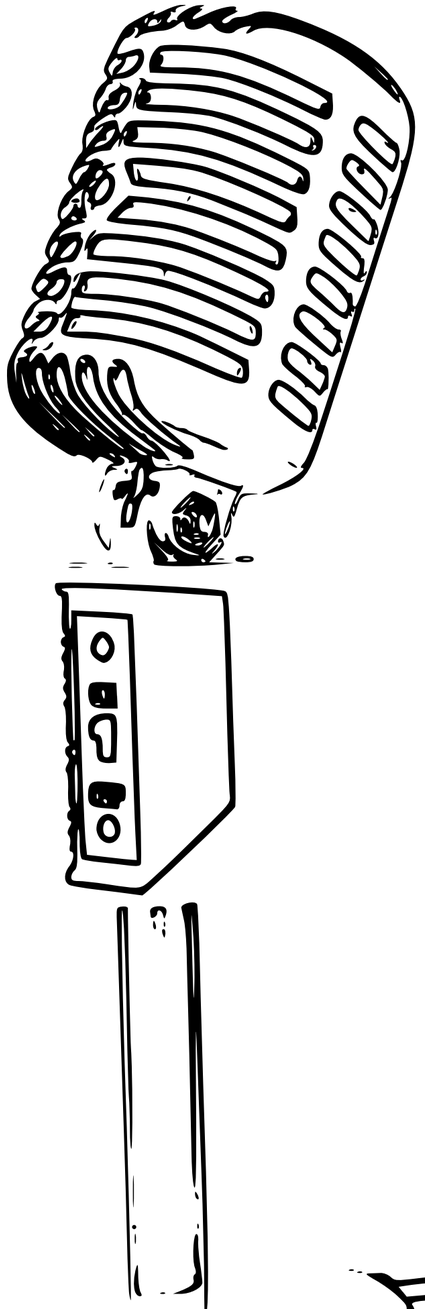
The objective is to predict listening time (minutes listened) from episode and context-level information(e.g., duration, ad load, publication time, genre, popularity signals) through CRISP-DM method:

1.Data Understanding and Data Exploration

2. Data Preparation and Feature Engineering

3. Model Training and Evaluation

4. Interpretation of Results and Insights



0. Business Understanding

The aim of the project is to provide practical insight for creators and advertisers about what drives engagement, predicting how long an episode will be listened to.

Key factors

- Podcasts is a rapidly growing medium lacking predictive analytics
- Producers need data-driven tools to anticipate engagement
- Understand behavioral patterns behind engagement

1.Data Understanding

The dataset comprises podcast episodes as observational units. Each row represents one podcast episode. The data combines structural, contextual, and popularity information used to model listening time.

Main characteristics

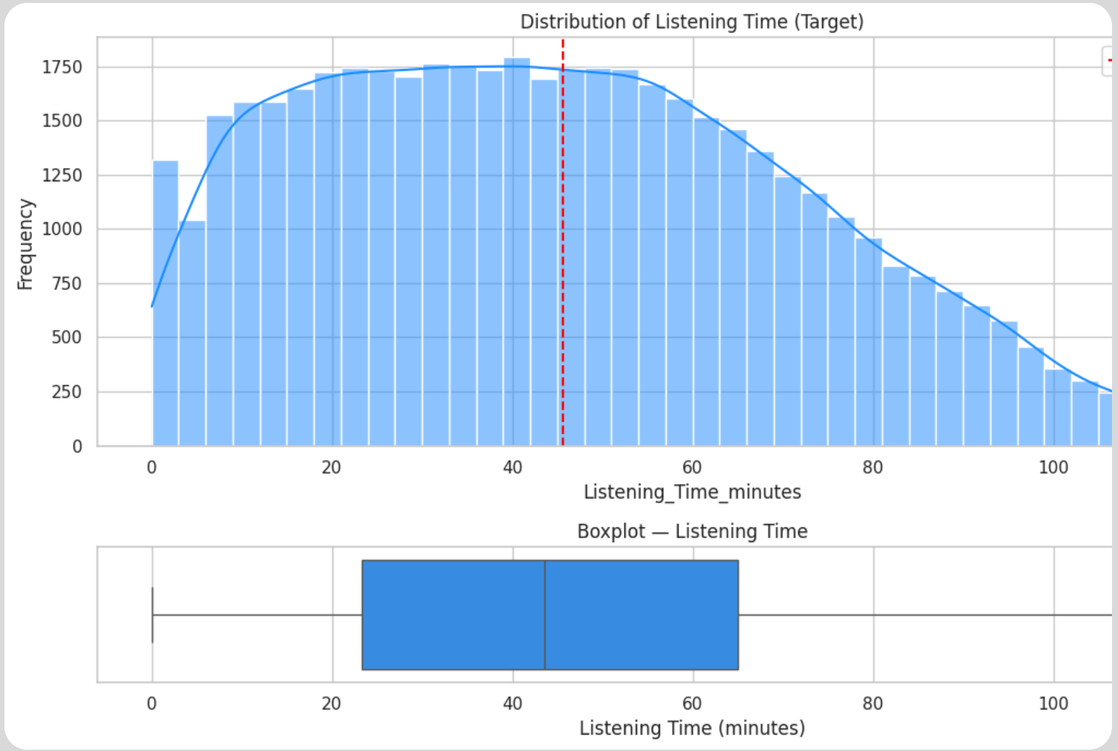
- Source: kaggle
- Over 47,000 cleaned observations after preprocessing
- Target variable: Listening_Time_minutes

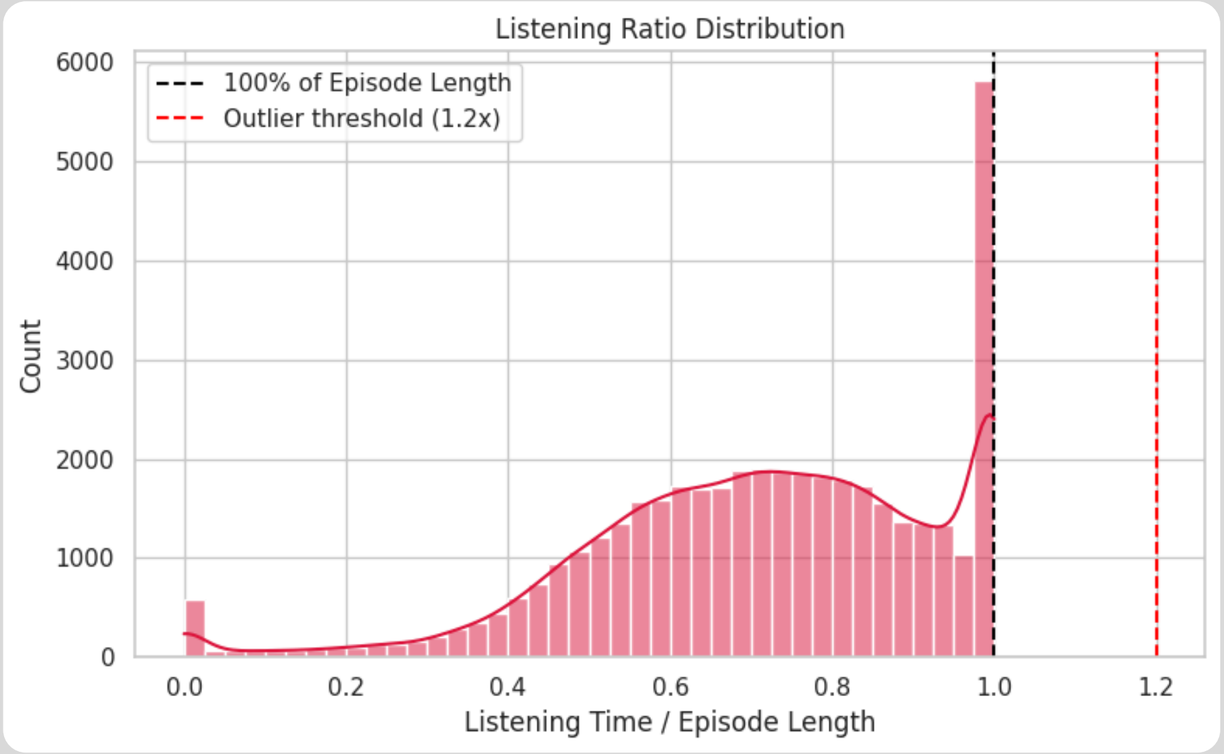
Columns

- Podcast_Name: Name of the podcast show each episode belongs to.
- Episode_Title: Title of the individual podcast episode.
- Episode_Length_minutes: Total duration of the episode in minutes.
- Genre: Content category or genre of the podcast episode.
- Host_Popularity_percentage: Popularity score of the podcast host, expressed as a percentage (0–100).
- Guest_Popularity_percentage: Popularity score of the guest, if present (0–100).
- Publication_Day: Day of the week the episode was published.
- Publication_Time: Time period of publication within the day (Morning, Afternoon, Evening, Night).
- Number_of_Ads: Number of advertisements contained in the episode.
- Episode_Sentiment: Overall sentiment of the episode's content (Positive, Neutral, Negative).
- Listening_Time_minutes (target variable): Average listening duration of the episode, measured in minutes.

1.Data Exploration: Target

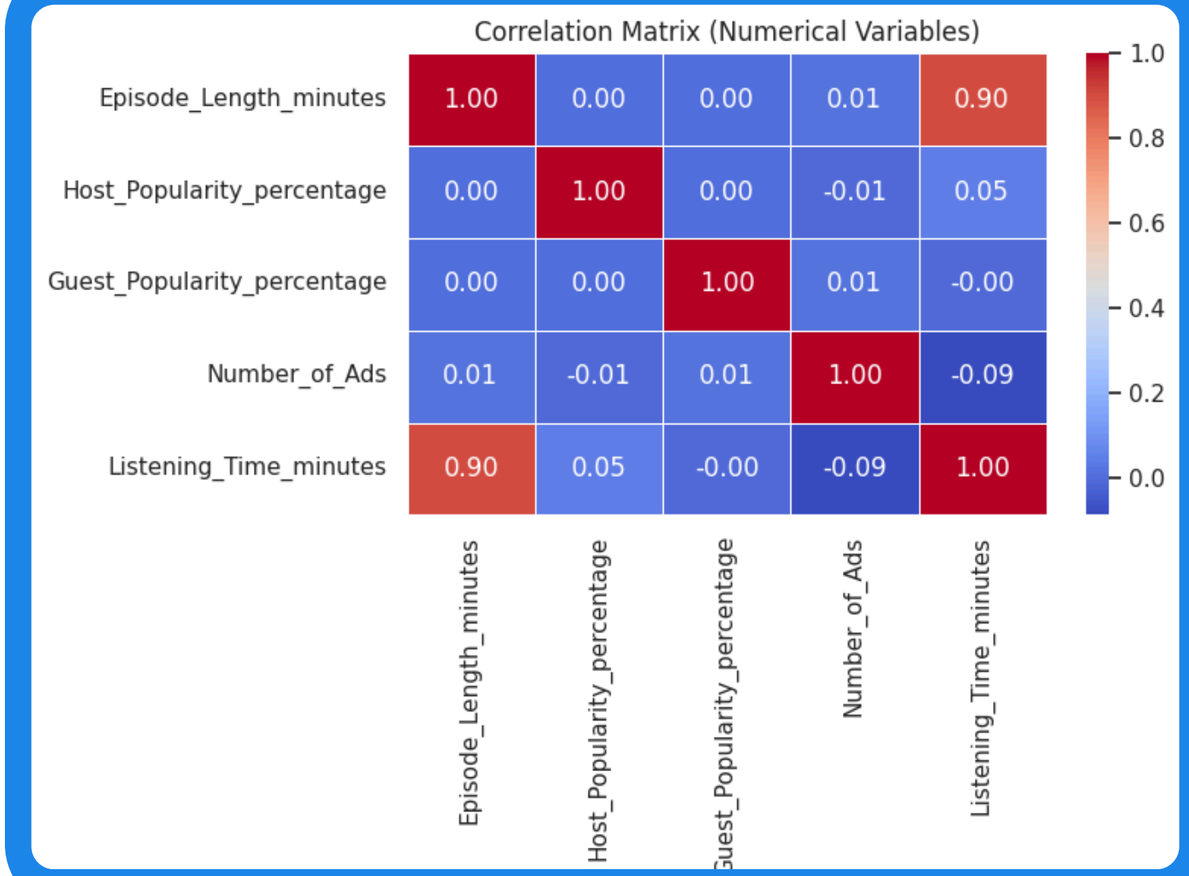
Exploring how users engage with episodes helps identify behavioral trends and outliers. The **listening time** shows a **right-skewed pattern**, with most users listening to less than the full episode. The **listening ratio** confirms that a **large share of sessions end before completion**, while only a few exceed 100% (outliers or replayed segments).





1.Data Exploration: Correlation

Exploring relationships among numeric features helps identify redundancy and potential drivers of listening behavior.



The correlation matrix reveals moderate associations between episode length, advertising load, and listening time, while most other features remain weakly correlated, suggesting nonlinear interactions captured later by Random Forest.

2. Data Preparation

Before modeling, the dataset was cleaned, encoded, and split to ensure fair evaluation and reproducibility.

Removal of missing targets and outliers

Median imputation for numeric, mode for categorical

Definition of listening ratio and removal of episodes exceeding 120% of their duration
r > 1.20

Categorical variables one-hot encoded and numeric variables standardized within a unified preprocessing pipeline to avoid data leakage.

The dataset split into 80% training and 20% testing sets

Consistency flag for
Listening mins > Episode Lenght

2. Feature Engineering

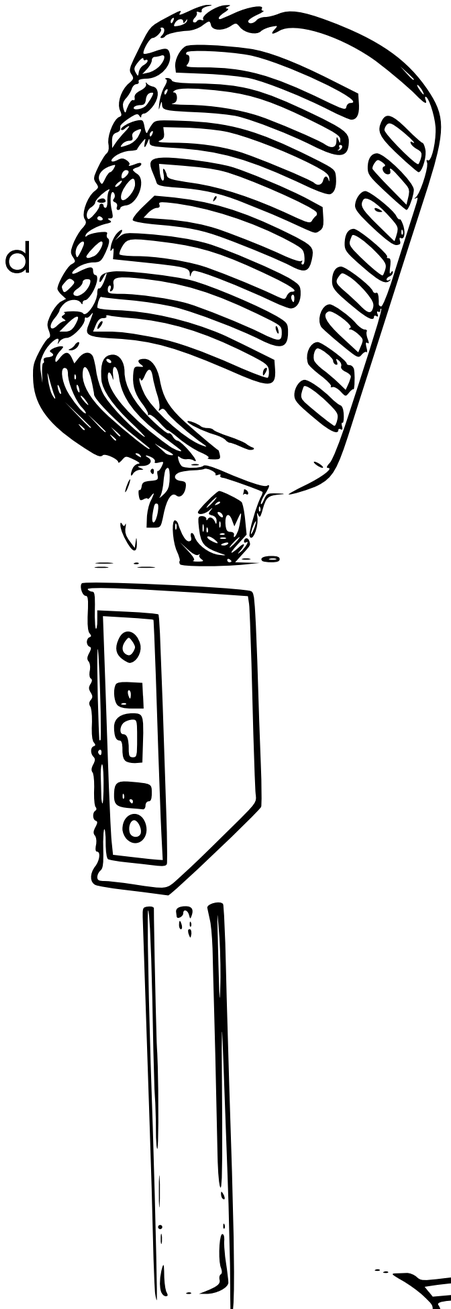
The engineered features add behavioral meaning while keeping the dataset small and interpretable. I added a total of 6 feature to the analysis.

Log_Length
reduces skewness, captures diminishing returns

Pop_Interaction
host × guest popularity

Ads_per_Minute
Has_Ads
ad density and presence

Pub_Time_Enc
Is_Weekend
temporal context



3. Model training and Evaluation

Each model uses the same preprocessing steps, ensuring a fair comparison across linear and nonlinear approaches. With MAE, RMSE, R^2 as evaluation metrics.

Linear Regression

Simple interpretable baseline, but with limited explanatory power. Fails to capture nonlinear behavior in listening trends

Lasso Regression

Adds L1 regularization to control multicollinearity and improves feature sparsity. Anyways, remains restricted to linear relationships and underfits complex patterns.

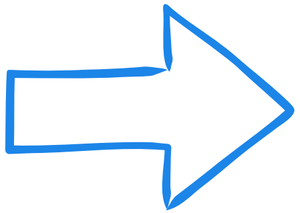
Random Forest Regressor

Nonlinear ensemble capturing interactions and performs best across all metrics. It effectively models nonlinear and interaction effects in behavioral data.

3. Feature Selection, Cross-Validation and Hyperparameter Optimization

The goal is to refine the models to improve generalization and remove redundant predictors with:

- **Cross-Validation (5-Fold)**: ensures stable performance across data splits.
- **Lasso Optimization (GridSearchCV):** tested alpha values to balance bias–variance trade-off.
- **Random Forest Optimization (RandomizedSearchCV):** explored 20 parameter combinations (n_estimators, max_depth, leaf size, features).
- **Feature Selection (SelectFromModel)**: retained only features with ≥ 1% importance, based on Random Forest importances.



Results

Refinement phase improved interpretability and reliability, with fewer features, more trustworthy patterns.

- Most relevant predictors: Episode_Length_minutes, Log_Length, Ads_per_Minute, Host_Popularity_percentage.
- Slightly higher MAE but stronger generalization, the model now reflects realistic listening behaviors.
- Lasso remained limited by linearity instead Random Forest captured non-linear effects more effectively.

3. Model Results

The **Linear Regression** provided a simple baseline but captured only limited variance.

The **Lasso Regression** improved stability through regularization yet remained restricted to linear effects.

The **Random Forest** performed best, effectively modeling nonlinear interactions.
After cross-validation and feature selection, the results became slightly less accurate but more realistic and generalizable.

Model	MAE ↓	RMSE ↓	R^2 ↑
RandomForest (simple, no CV)	9.09	11.62	0.810
RandomForest (CV tuned, 5-fold)	9.10	11.71	0.807
Lasso (simple, no CV)	9.52	12.36	0.785
Lasso (CV tuned, 5-fold)	9.53	12.40	0.783
Linear Regression	9.54	12.36	0.785

4. Interpretation of Results and Insights

Feature selection showed that only a **small set of variables meaningfully** explains listener engagement, resulting in a simpler yet more reliable model.

The **Random Forest** provided a more **realistic view** of the underlying behavior, capturing nonlinear relationships where listening time depends primarily on structural aspects such as episode length and advertising load, together with social influence factors.
In contrast, **linear models failed** to capture these interaction effects, oversimplifying the dynamics of user attention.

4. Key Insights

Episode duration (Episode_Length_minutes, Log_Length)

is the strongest predictor, listening time increases with length but shows diminishing returns.

Advertising variables (Ads_per_Minute, Number_of_Ads)

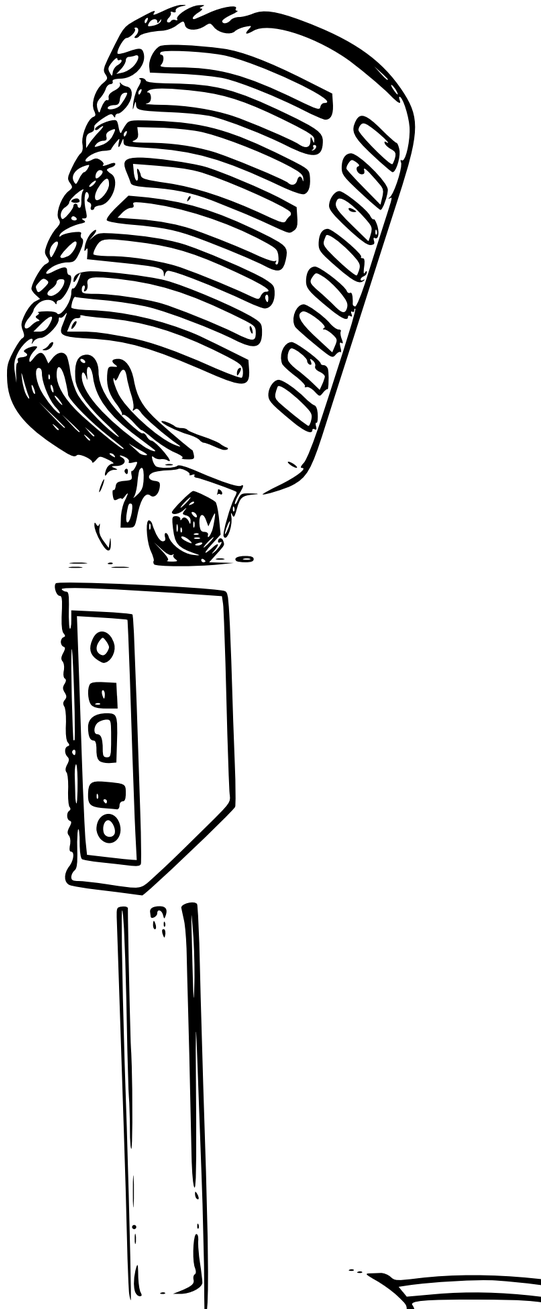
have a clear negative impact, confirming that frequent ads reduce engagement.

Popularity factors (Host_Popularity, Guest_Popularity, Pop_Interaction)

contribute moderately, suggesting social influence effects between well-known hosts and guests.

Temporal context (publication day/time)

plays a smaller yet consistent role, evening or weekend releases perform slightly better.



Conclusions

The project successfully predicted podcast listening time using metadata features.
Among all models, the **Random Forest** achieved the **best balance** between accuracy and interpretability, outperforming linear baselines by capturing nonlinear patterns.
After **cross-validation** and **feature selection**, results became **slightly less accurate** but more **realistic** and generalizable.
Episode length and **ad density** emerged as the **strongest behavioral drivers**, while popularity and timing had smaller but consistent effects.
The refined model provides a clear, trustworthy view of listening behavior and sets the foundation for future work with richer textual or audio data.

Thank you for your attention