**ROADMAP**

**PHASE 1: Understanding the Problem**

**✅ Steps:**

* **Read the competition overview**: Understand the target (listening\_time) and evaluation metric (likely RMSE or MAE).
* **Explore the data**:
  + Look at train.csv, test.csv, and sample\_submission.csv.
  + Note the features, types (categorical, numerical), missing values.

🛠 Tools: pandas, seaborn, matplotlib

**📍 PHASE 2: Exploratory Data Analysis (EDA)**

**✅ Steps:**

* Check distributions of features (histograms, value\_counts).
* Correlation with the target.
* Detect outliers and missing values.
* Identify categorical vs numerical features.

📌 Tip: Create an EDA notebook and share on Kaggle for feedback!

**📍 PHASE 3: Data Preprocessing**

**✅ Steps:**

* Handle missing values (imputation or removal).
* Encode categorical variables:
  + Label Encoding or One-Hot Encoding.
* Normalize or scale numerical features.
* Split train.csv into X\_train, y\_train and X\_valid, y\_valid.

🛠 Tools: scikit-learn, LabelEncoder, StandardScaler

**📍 PHASE 4: Baseline Modeling**

**✅ Steps:**

* Start with simple models:
  + Linear Regression
  + Decision Tree Regressor
* Use RMSE or MAE to evaluate.
* Try a few basic models and create a leaderboard for yourself.

**📍 PHASE 5: Advanced Models & Tuning**

**✅ Steps:**

* Try:
  + Random Forest
  + XGBoost / LightGBM
  + CatBoost (works well with categorical features!)
* Tune hyperparameters with GridSearchCV or Optuna.

🛠 Tools: xgboost, lightgbm, optuna

**📍 PHASE 6: Feature Engineering**

**✅ Steps:**

* Create new features based on domain logic.
  + e.g., interaction between user and podcast?
* Try feature importance from models to drop low-importance ones.

**📍 PHASE 7: Ensembling (optional but helpful)**

**✅ Steps:**

* Combine multiple models for better predictions:
  + Averaging
  + Stacking

**📍 PHASE 8: Submission**

**✅ Steps:**

* Predict on test.csv using the final model.
* Format predictions as sample\_submission.csv.
* Submit on Kaggle.
* Check your position on the leaderboard.

**📍 PHASE 9: Reflect & Share**

* Write a Kaggle notebook summarizing your approach.
* Note what you learned.
* Save your notebook for portfolio purposes!

**🧰 Suggested Folder Structure**

kotlin

KopyalaDüzenle

📁 podcast-prediction/

┣ 📜 eda.ipynb

┣ 📜 baseline\_model.ipynb

┣ 📜 advanced\_models.ipynb

┣ 📜 final\_submission.ipynb

┗ 📂 data/

┣ 📄 train.csv

┣ 📄 test.csv

┗ 📄 sample\_submission.csv