PREDICTING LISTENING TIME

Agenda

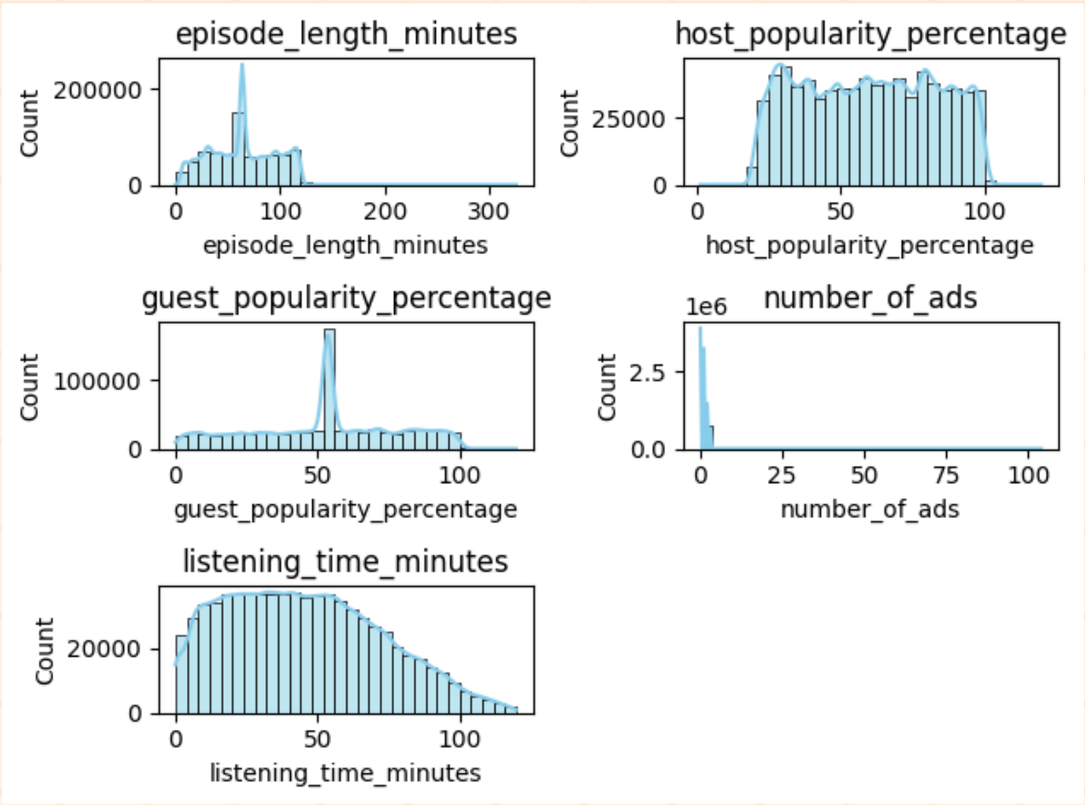
- Scope
- EDA & Data Manipulation
- Model settings
- Results

Scope

The scope of this Kaggle challenge is to establish a model which is going to be able to predict the average listening time for podcast episodes based on different features. The data set has been provided by Kaggle.

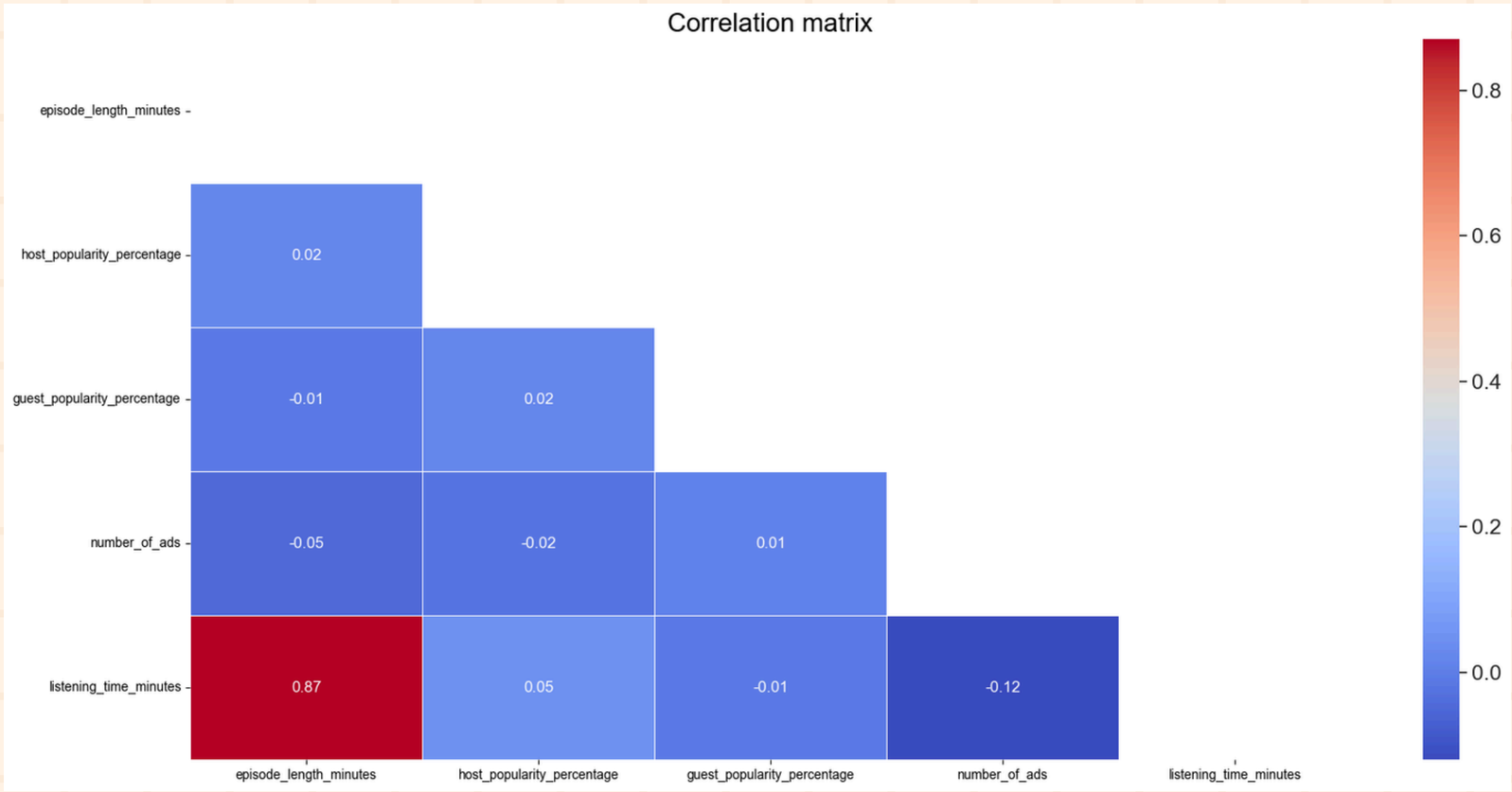
EDA & Data Manipulation

- Categorical features holding less predictive power removed (Episode Title, Id, etc.)
- Listening Time right positively skewed
- Episode length, Host popularity and Guest popularity fairly uniform with some outliers



EDA & Data Manipulation

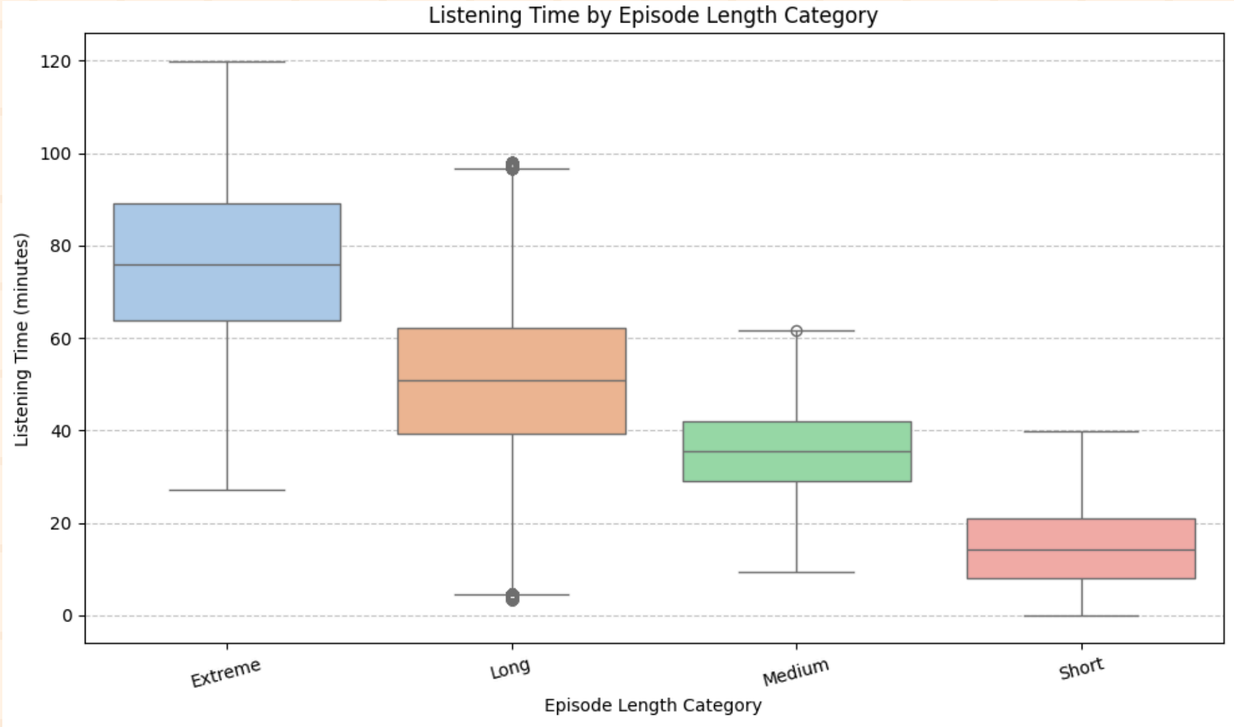
Correlation for the numerical values:



The numerical values stemming from the data frame carry weak positive and negative correlation. The strongest correlation derives from Listening Time Minutes vs Episode Length Minutes which is expected as the longer is the episode, the longer is the average time listened to.

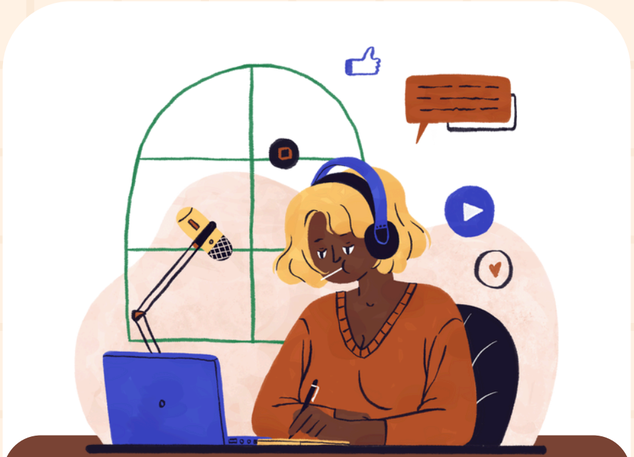
EDA & Data Manipulation

- New Features implemented from the original features: Popularity Difference, is weekend?, Ad Density, Episode Length Label.
- Once Episode Length Label was established outliers have been removed.



The Power of Digital Storytelling

Enhances engagement through voice and sound.



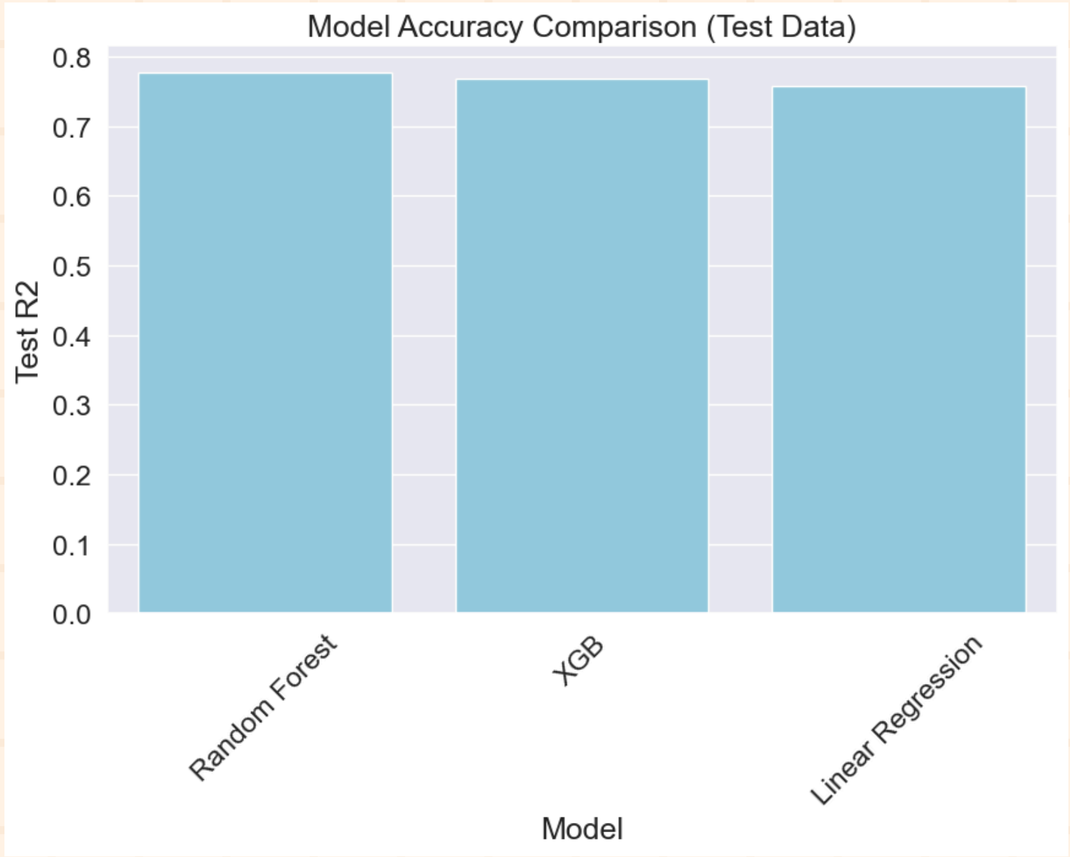
Allows creators to craft immersive narratives.

Provides a personal and intimate connection with the audience.

Digital storytelling through podcasts is unique because it relies on voice and sound to create an immersive experience, making it a powerful way to connect with audiences personally.

Model Setting

Models selected to complete the task:
RandomForest, XGB and LinearRegression



RandomForest performing slightly better than XGB without further Feature Engineering but XGB turned out to be the best choice on further tests reaching roughly 79% in precision and MAE on Train and Test data of respectively: 9.28 and 9.33 vs RandomForest which presented: 3.57 and 9.36 (probably overfitting)

Conclusion

The XGB Model performs fairly and is able to predict listening time with a MAE of roughly 9.33 and R2 score of 0.780142