Luis Rosado

Machine Learning ITM 617

CAPSTONE PROJECT

INTRODUCTION:

In the era of personalized digital experience, understanding user behavior is paramount for optimizing content delivery and engagement strategies. The study we conducted was the application of machine learning techniques to predict the listening time of podcast subscribers by leveraging various attributes intrinsic to a given podcast. By analyzing factors such as publication day, episode length, advertising, episode sentiment our goal is to develop a predictive model that can give an accurate estimate of how long a subscriber/listener is likely to engage in a particular episode. Through a series of data cleaning, wrangling, feature selection, fine tuning and model evaluation our objective is to achieve a prediction with the highest possible accuracy through machine learning and statistical methods. This will provide valuable insights for content creators and distribution platforms whose aim is to maximize listener retention, satisfaction and to understand what content will captivate audiences.

DATA:

The dataset used in this study was sourced from Kaggle, specifically from the competition titled "Predict Podcast Listening Time". It consists of synthetically generated data designed to mimic real-world podcast listener behavior. The dataset was comprised of 750,000 rows and 12 columns. In our model we used an 80/20 split between training and testing data sets. Given the dataset's large size, this factor was carefully considered when selecting our models.

Since the data was synthetically generated rather than collected from real-world podcast platforms, this approach ensures that test labels remain private while still providing realistic data patterns. However, synthetic data can introduce artifacts, unintended distortions or inaccuracies that stem from its artificial nature. The competition organizers aim to continuously enhance the quality of synthetic data over time to minimize these imperfections, making the dataset more representative of actual user behavior.

To better understand artifacts in synthetic data, we can compare them to imperfections in a digitally altered photo. When an image is edited, certain elements may stand out as artificially generated. Likewise, in this dataset, artifacts manifest as patterns that deviate from real-world podcast listening behavior.

METHODOLOGY:

- Data Preparation and Cleaning

Prior to choosing and building a model the dataset underwent a systematic cleaning and wrangling process. The initial dataset was comprised of multiple columns, including categorical and numerical variables that were relevant to the podcast performance such as *Episode_Length_minutes*, H*ost_Popularity_percentage, Guest_Popularity_percentage* and the target variable *Listening_Time_minutes.* Filling the empty variables with the mean of the column most likely directly affected the outcome of the data but was the most logical way and is considered standard practice when cleaning a data set with a lot of missing values. Since there were a total of 146,030 missing rows of data it was a practical choice to use this method in our data cleaning.

- Missing Data Handling:

When searching the data, there were 3 columns with missing data; *Episode_Length_minutes*, *Guest_Popularity_percentage* and *Number_of_Ads*. The mean for the missing values was calculated and inputted to fill in the missing cells. This ensured that the missing values did not affect the model estimates. This also enhanced the clarity and coherence of the data, making it more accessible for more meaningful interpretation and analysis.

Redundant and non-informative columns, such as *id, Podcast_Name and Episode_Title,* were excluded from analysis. These columns and data were not valuable to interpretation or analysis. The *id* column was passed through in some processes to be able to identify what the other features of the row were once predictions were made.

- Encoding Categorical Variables:

Initially the categorical features, E*pisode_Sentiment, Publication_Day, Publication_Time* and *Genre* were transformed using One-Hot Encoding to create binary indicator variables. This approach was intended to incorporate categorical predictors without imposing artificial ordering that would have occurred if Label Encoding were to be used. However, through exploratory analysis it was discovered that the inclusion of encoded categorical variables significantly inflated multicollinearity. Evidence of this was seen with decreasing Variance Inflation Factor (VIF) scores. This, in turn, degraded model performance. Consequently, encoded variables were omitted from the final regression models.

- Model Selection and Rationale
    - Multiple Linear Regression:

The primary analytical method applied was Multiple Linear Regression (MLR). MLR was chosen because the primary research question aimed to model the relationship between multiple independent variables and a continuous dependent variable (*Listening_Time_minutes*). The assumptions of linearity, independence, homoscedasticity, and normality of residuals were examined to validate model suitability. MLR was particularly appropriate due to its interpretability:

- Coefficients provide clear insights into the magnitude and direction of how each of the independent variables effected the dependent variable.
- The R^2 statistic was utilized to measure the overall model fit.

Given MLR's sensitivity to multicollinearity, which can inflate standard errors and undermine model reliability, VIF was calculated for all predictor variables. Any variable exceeding the accepted threshold (>10) would have been reviewed for modification or removal; however, none surpassed this limit.

- Ridge Regression & Lasso Regression

To address potential multicollinearity that remained despite variable exclusion, Ridge & Lasso Regression models also called L2 and L1regularization, were implemented as secondary modeling strategies with an attempt to overcome the overfitting problem that we observed in the Multiple Linear Regression model. Ridge Regression penalizes the function by squaring coefficient magnitudes, thus stabilizing our estimates when predictors are highly correlated. Lasso regression is designed to add a penalty based on the absolute values of the coefficients, effectively shrinking some (most of them) in our case to 0.

- Model Evaluation

Both the MLR and Ridge Regression models were trained using a 80/20 train-test split to ensure robust out-of-sample evaluation. Key metrics for performance evaluation included:

- R^2 Score: To assess the proportion of variance explained by the model.
- Mean Squared Error (MSE): To gauge prediction error magnitude
- Root Mean Squared Error (RMSE): To further gauge prediction error magnitude depending on data values
- VIF Scores: Continuously monitored in the MLR to assess multicollinearity.

The Ridge model's performance was compared directly to the MLR to determine the benefit of regularization in terms of predictive accuracy and model stability.

In preprocessing the attempt to encode the categorical variables was ultimately not incorporated into the final models as they negatively impacted multicollinearity and R2 performance. This finding underscored the importance of iterative feature engineering and the effects of adding complexity to the regression models.

RESULTS:

Within this study various models were run and tested. MLR, Ridge and Lasso Regression were run as part of an ensemble model to develop a more refined method and result. The findings of this analysis reveal key insights, offering a deeper understanding of the dataset's underlying trends. Through rigorous exploration and statistical examination, we unveiled meaningful relationships that not only validate initial hypotheses but also introduce nuanced perspectives. The following section presents a detailed breakdown of the results, highlighting significant observations, emergent correlations, and the broader implications of these discoveries.

In Multiple linear regression model performed quite well with a coefficient of determination of 0.759. This tells us that 75.9% of the variance in the dependent variable (Listening_Time_minutes) can be explained by the independent variables. Along with this it resulted in a 13.3 RMSE which is low enough

to give us confidence in our model. The result of the MLR gave us a MLR line of $Y = \beta_o + \beta_1X_1 + \beta_2X_2 + ... + \beta_pX_p$. Our model resulted in a MLR line with values of –2.49 for β_o, -0.008, Guest Popularity factor (β_1X_1), 0.04 ,Host Popularity factor (β_2X_2) , -1.87, Number of Ads (β_3X_3) and 0.754 Episode Length in minutes (β_4X_4). With this formula we can accurately predict the expected Listening time in minutes based on the Guest Popularity, Host Popularity, Number of ads and total Episode length of a selected Podcast.

The coefficient of determination, also called R-squared, was around the same for the other 2 models, Ridge and Lasso Regression. The lowest was the Lasso Regression with an R squared of 0.748. It is left to our discretion as analysts to decide whether we should focus on RMSE or MSE. It made more logical sense that, as a measure compared to the units of the target variable (minutes), RMSE would be more meaningful for interpretation. In this study the minutes listened seemed more meaningful to 13.3 (RMSE) than 177(MSE). Accuracy measures were run as well but resulted in a value almost identical to that of R-squared.

DISCUSSION:

The resulting outcome gave us a good idea of how we can confidently predict the Listening time of the given Podcast when equipped with the other features of the podcast. The models yielded a good level of fit and accuracy that would make for a really good application if the data was real-world data. Iterations and refinement of data developed from artificial means would improve upon the data getting closer to data that would accurately mimic the behavior of an actual listener. This coupled with ingenuity and advanced analysis could lead to machine learning models that can not only produce the data but also predict, through the interpretation of the fabricated data, human behavior.

The use of synthetically generated data in this dataset sparked curiosity about how real-world data might perform under the same model. Now that we have developed an ensemble model utilizing three regression methods, MLR, Lasso and Ridge Regression, its application could prove invaluable to podcast content creators. Understanding the predictability of past podcast data could enable them to refine their focus areas, optimize marketing strategies, and enhance content creation with greater precision.

One of the most perplexing and impractical metrics within the dataset was the Guest & Host Popularity percentage feature, as validating its percentage posed a significant challenge. Given that the data was synthetic, there was no clear methodology or explanation for how this factor was determined, rendering it unreliable. Consequently, we exercised discretion in excluding outliers exceeding 100% popularity, based on the assumption that popularity correlates with likability. To improve the quality of synthetic data, it would be beneficial to prioritize features that are both practical and commonly used in real-world podcast analysis, ensuring they contribute meaningful insights.

Cross validation models were run on the original data file after cleaning and encoding categorical features as another method for splitting the data to see if we could get improved regression scores of the three models (Ridge, Lasso, Mutiple Regression) with less bias. Outliers were again identified through scatter plot identification and 3 std deviations or greater from the mean. Again, all the null values were replaced with the mean, preventing any skewing of the data. The cross-validation (3-fold through 7-fold) models were used to see if the results would be affected by the number of folds. However, each time the results were the same as before with the best scores being tied between the ridge and multiple regression models at 0.757 using the 5-fold model.

We observed the average percentage of Listening Time over Episode Length by Genre and found that in all genres, listeners only tune into about 70% of the total length of an episode with the Technology and True Crime genres improving to 71% (not much of an increase). Additionally, we found that the average episode is 64 minutes with 1.36 advertisements. This average is decreased to 1.31 for episodes with a positive Episode Sentiment.

LIMITATIONS:

- Data Source

Our data was sourced from a Kaggle competition and is synthetically generated from a deep learning model trained on a Podcast Listening Time Prediction dataset. Due to the method of origin, this data did have various artifacts that were generated such as outliers with features that contextually did not make sense (one example is guest popularity percentage rankings over 100%) and 19.5% of the rows contained null values.

- Outliers

When initially examining the data, we created scatterplots and ran descriptive statistics to identify any outliers. We then determined to label any data point that was 3 standard deviations or more away from the mean value of a feature an outlier and were able to identify a total of 54 out of 750,000 data points and removed any row affected from the dataset to avoid skewing the data

- Missing data:

The dataset included 146,030 rows across 3 numerical columns with missing data: *Episode_Length_minutes, Guest_Popularity, Number_of_Ads.* We were able to limit the impact on our analysis by filling these missing values in with the mean value for each feature affected. We did notice that while this provided more clarity, our scatterplots now had a line of data points at the mean that ran perpendicular to the axis of the columns that were filled.

- Scope:

One of our major limitations was that our target variable was continuous while most of our data was categorical. Due to the combination of datatypes available, we felt that we were limited to exploring fewer models than we initially hoped to.

CONCLUSION:

This study leveraged machine learning techniques to estimate podcast listening duration based on key attributes, aiming to assist content creators and platforms in maximizing engagement. The

dataset, derived from a Kaggle competition, comprised 750,000 synthetically generated records that simulated listener behavior. Extensive data preprocessing addressed 19.5% rows including at least 1 missing value. They were accounted for through mean substitution and outlier removal. Key features influencing listening time included episode length, host and guest popularity, and advertisement count. Due to multicollinearity concerns identified during exploratory analysis, non-informative columns and problematic categorical variables (such as genre and sentiment) were excluded from the ML models and only included in the descriptive statistics.

The modeling approach centered on Multiple Linear Regression (MLR), supplemented by Ridge and Lasso regression to counteract multicollinearity and overfitting risks. MLR demonstrated strong predictive capabilities ($R^2 \approx 0.76$, RMSE ≈ 13.3 minutes), explaining roughly 76% of the variance in listening duration. Ridge and Lasso models yielded comparable results, reinforcing the reliability of regularized methods. Cross-validation (5-fold model) further supported these findings, with minimal bias across different data splits.

Key takeaways indicated that episode length, advertisement frequency, and host/guest popularity significantly influenced listening duration. However, the synthetic nature of the dataset introduced challenges, notably in the ambiguous definition of "popularity" and implausible values (e.g., popularity exceeding 100%). These constraints underscore the importance of real-world data for validating and refining the findings.

In conclusion, while the models provide valuable insights for engagement prediction, future research should enhance data realism and explore additional predictive factors to more accurately represent listener behavior.

Resources:

Walter Reade and Elizabeth Park. Predict Podcast Listening Time. https://kaggle.com/competitions/playground-series-s5e4, 2025. Kaggle.