Stat 220 Lab 8

1 Introduction

In this lab, you are tasked with building and refining predictive models for Spotify song popularity. Predicting popularity is critical for understanding trends in music consumption, artist reach, and marketing strategies. You will explore various transformations and interactions in the model to enhance predictive power and will assess model performance on a test set.

You have been provided with pre-split training and test datasets. Your task is to iteratively refine your models on the training data and then evaluate your models on the test data to assess out-of-sample performance.

Data Acquisition and Initial Exploration

To get started, load and explore the data.

1. Download the training dataset from: `https://richardson.byu.edu/220/spotify_train.csv`
2. Download the test dataset from: `https://richardson.byu.edu/220/spotify_test.csv`
3. Load the datasets into Python and perform an initial exploration. Examine variable distributions, detect outliers, and familiarize yourself with each variable's potential relationship to the target variable `popularity`.

Variable	Description
`popularity`	Target variable, indicating the popularity score of a song
`duration_ms`	Song duration in milliseconds
`explicit`	Indicates if the song is explicit (1 = Yes, 0 = No)
`danceability`	Danceability score (0 to 1)
`energy`	Energy score (0 to 1)
`loudness`	Loudness of the song in dB
`mode`	Modality of the song (1 = Major, 0 = Minor)
`speechiness`	Speechiness score (0 to 1)
`acousticness`	Acousticness score (0 to 1)
`instrumentalness`	Instrumentalness score (0 to 1)
`liveness`	Liveness score (0 to 1)
`valence`	Positivity score (0 to 1)
`tempo`	Tempo of the song in BPM
`time_signature`	Time signature (integer, typical values are 3, 4, or 5)

Table 1: Data dictionary for Spotify dataset.

Regression Analysis

Your objective is to build and refine a predictive model for song popularity. Through an iterative process, explore a variety of transformations, higher-order terms, and interactions with the goal of achieving an R^2 of at least 0.125 on the test set. Use the training set for fitting models, and evaluate performance on the test set to gauge generalizability.

1. **Data Preparation and Initial Model**

   (a) **Transformations of the Target Variable (y-variable)**: Start by exploring potential transformations of the response variable, `popularity`. Common transformations to try include:

   - Log transformation: log(popularity)
   - Square root transformation: $\sqrt{\text{popularity}}$
   - Inverse transformation: 1/popularity

   Determine if any transformation improves the model's fit on the training set, and document the resulting R^2 values.

   (b) **Initial Linear Regression Model**: Begin with a linear regression model using the untransformed predictors in the training set to predict `popularity`. Calculate R^2 on the training and test sets to establish a baseline for comparison.

2. **Transformations of Predictor Variables**

   (a) For continuous predictors, test various transformations that may improve linearity or normalize skewed distributions. Consider:

   - Log transformation: $\log(X + 1)$, adding 1 to handle zero values.
   - Square root transformation: $\sqrt{X}$
   - Square transformation: X^2
   - Inverse transformation: $1/(X + \epsilon)$, where ϵ is a small constant added to avoid division by zero.

   (b) For each transformation, refit the model on the training data and evaluate R^2 on the test set. Document any improvements achieved.

3. **Higher-Order Terms and Interactions**

   (a) **Higher-Order Terms**: Add squared and cubic terms for continuous predictors where non-linear relationships are suspected. For instance, include terms such as `danceability`^{2}, `loudness`^{3}, and so on.

   (b) **Interaction Terms**: Explore two-way interactions between predictors (e.g., `energy * danceability`, `loudness * valence`). Interaction terms can capture combined effects of predictors on `popularity`.

   After adding higher-order terms and interactions, evaluate the model on the training and test sets. Document any changes in R^2.

4. **Iterative Model Refinement and Selection**

   (a) Use stepwise selection or other methods to reduce the model's complexity by removing predictors that do not contribute significantly. Aim to strike a balance between model complexity and predictive performance.

(b) Fit the refined model on the training set, and calculate R^2 on both the training and test sets. This will help you monitor any signs of overfitting or underfitting.

5. **Final Model Evaluation**

   (a) Ensure that the final model achieves an R^2 of at least 0.125 on the test set, or as close to this target as possible.

   (b) Evaluate your model for overfitting by comparing the training and test R^2 values. A large discrepancy may indicate overfitting, while similar R^2 values suggest better generalizability.

   (c) Document all variables included in your final model. Attempt to interpret each variable's role in predicting song popularity in context. For example, discuss how and why features like `danceability` or `acousticness` may contribute to popularity predictions.

Note: Keep a log of each model variation you try, including the transformations, higher-order terms, and interactions added. You do not need to report each variation, but if you forget to mark it down, you may try it again.

Expected Outcomes

Your report should summarize the steps you took to optimize the model and should clearly state the final in-sample and out-of-sample R^2 values. Additionally, include a discussion of any overfitting or underfitting observed, and provide interpretations of all final model variables in the context of song popularity.

Discussion and Reporting

Your report should address the following:

1. **Model Development**

   (a) Describe several model variations you tested and the rationale for adding transformations, interactions, and/or stepwise selection.

   (b) Summarize any important transformations that improved model fit

   (c) DO NOT report every single model attempted.

2. **Model Evaluation**

   (a) Report the final in-sample R^2 and out-of-sample R^2 for your best model.

   (b) Justify why the selected model is the best, based on statistical performance and interpretability.

3. **Interpretation**

   (a) Interpret the coefficients of your final model in the context of song popularity.

   (b) Discuss which features are most influential for predicting popularity and any interactions that provide insights.

Deliverables

Your final submission will consist of a Python Notebook or some other form of a report. Your report should be well-organized and formatted as a report, results, and visualizations. Make sure to answer the questions directly in the report. For easier grading, and to make sure we don't miss anything when grading, label the questions answered directly i.e. have dedicated sections to the four parts above and be clear what question you are answering in each section.