Lab 07

Rohit Kulkarni

3/2/2022

Exercises

Part I: Data Prep & Modeling

1. Read through the Spotify documentation page to learn more about the variables in the dataset. The response variable for this analysis is target, where 1 indicates the user likes the song and 0 otherwise. Let's prepare the response and some predictor variables before modeling.

- If needed, change target so that it is factor variable type in R.
- Change key so that it is a factor variable type in R, which takes values "D" if key==2, "D#" if key==3, and "Other" for all other values.
- Plot the relationship between target and key. Briefly describe the relationship between the two variables.

2. Fit a logistic regression model with target as the response variable and the following as predictors: `acousticness, danceability, duration_ms, instrumentalness, loudness, speechiness, and valence`. Display the model output.
3. We consider adding `key` to the model. Conduct the appropriate test to determine if key should be included in the model. Display the output from the test and write your conclusion in the context of the data.

Use the model you selected in Exercise 3 for the remainder of the lab.

4. Display the model you chose in Exercise 3. If appropriate, interpret the coefficient for `keyD#` in the context of the data. Otherwise, state why it's not appropriate to interpret this coefficient.

Part II: Checking Assumptions

In the next few questions, we will do an abbreviated analysis of the residuals.

5. Use the `augment` function to calculate the predicted probabilities and corresponding residuals.
6. Create a binned plot of the residuals versus the predicted probabilities.
7. Choose a quantitative predictor in the final model. Make the appropriate table or plot to examine the residuals versus this predictor variable.
8. Choose a categorical predictor in the final model. Make the appropriate table or plot to examine the residuals versus this predictor variable
9. Based on the residuals plots from Exercises 6 - 8, is the linearity assumption satisfied? Briefly explain why or why not.

Part III: Model Assessment & Prediction

10. Plot the ROC curve and calculate the area under the curve (AUC). Display at least 5 thresholds (n.cut = 5) on the ROC.
11. Based on the ROC curve and AUC in the previous exercise, do you think this model effectively differentiates between the songs the user likes versus those he doesn't?
12. You are part of the data science team at Spotify, and your model will be used to make song recommendations to users. The goal is to recommend songs the user has a high probability of liking.

- Choose a threshold value to distinguish between songs the user will like and those the user won't like. What is your threshold value? Use the ROC curve to help justify your choice.

13. Make the confusion matrix using the threshold chosen in the previous question.
14. Use the confusion matrix from the previous question to answer the following:

- What is the proportion of true positives (sensitivity)?
- What is the proportion of false positives (1 - specificity)?
- What is the misclassification rate?