Hannah Cronin

Advanced Data Mining Assignment 1:

Questions:

QA1: What is the main purpose of regularization when training predictive models?

Regularization works to reduce the magnitude of each variable. This is especially important when there are many variables, and their values may have wide ranges. This helps to avoid the model from overfitting to the noise of the data and avoid underfitting to optimize model performance.

(Regularization for Linear Regression Models, Slide 5)

QA2: What is the role of a loss function in a predictive model? And name two common loss

functions for regression models and two common loss functions for classification models.

Loss functions reflect how well our algorithm reflects our data. Loss function values are the differences between expected and predicted values. Two common loss functions for regression models are: MAE (mean absolute error) and RMSE (root mean square error). Two common loss functions for classification models are: cross entropy loss/log loss and hinge loss.

QA3: Consider the following scenario. You are building a classification model with many hyper

parameters on a relatively small dataset. You will see that the training error is extremely small.

Can you fully trust this model? Discuss the reason.

In this scenario, I would be hesitant to fully trust this model. First, with a small dataset, there’s a larger susceptibility that the model is overfitted and too attuned to the noise of the data. This is supported by the fact that the training error is extremely small. If the training error is too small, it suggests the model is overfitting. However, since we do not know how the model performed against the test or validation data, it’s hard to say for certain with just the training error. Also, since this model uses several hyperparameters (chosen by the modeler), I would not trust any initial runs/results until I could do further testing to verify that optimal hyperparameter values were chosen.

QA4: What is the role of the lambda parameter in regularized linear models such as Lasso or

Ridge regression models?

The lambda parameter measures association between features/variables. A lambda of 0 means no association and 1 means they are perfectly associated. Lambda also helps balance reducing loss on the training model data and minimizing magnitude of coefficients in model. A lambda of closer to 0 means that reducing training loss is prioritized while a lambda closer to 1 means coefficients may end up close to 0.

LINK TO GITHUB WHERE R CODE IS STORED:

QB1. The optimal lambda for my model was quite small and included all the variables accordingly.

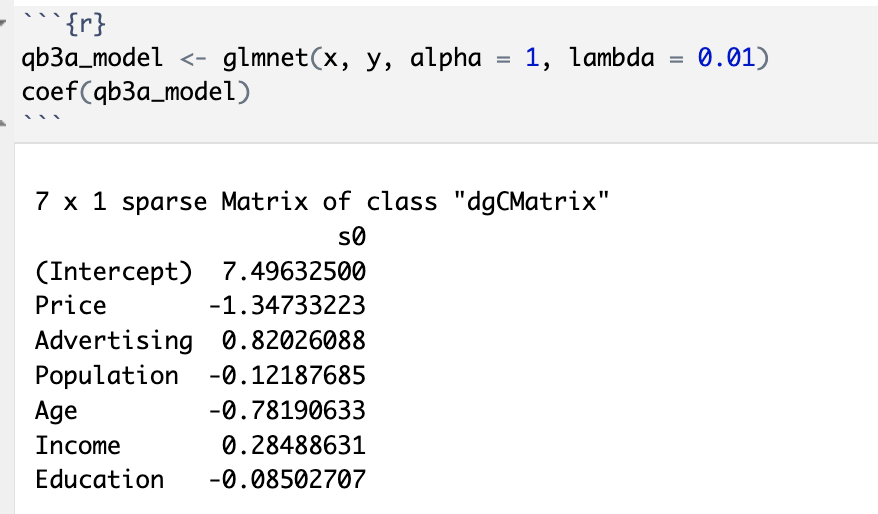
Graphical user interface, text, application

Description automatically generated

QB2. The coefficient for price is -1.35016474 in the optimal model (see screen print above).

QB3.

When lambda = 0.01



When the model has a lambda of 0.01, we can see that all features are still included.

When lambda = 0.1

Graphical user interface, text, application, email

Description automatically generated

When the model has a lambda of 0.1, the number of attributes drops to 4 instead of 6.

Based on these changes, it is expected that as lambda rises and the amount of weight placed on the coefficients increase, the number of attributes decreases. Less variables stay in the model (with non-zero coefficients) as lambda rises.

QB4.

Text

Description automatically generated with medium confidence

The best value for lambda is 0.01510377. Two variables are also not included when alpha = 0.6.