CUNY The City University of New York | School of Professional Studies

DATA 621 – Business Analytics and Data Mining

Homework #1 Assignment Requirements

Overview

In this homework assignment, you will explore, analyze and model a cigarette consumption data set containing information from a 1970 study of cigarette usage in the United States. This small data set has been modified in two ways. First, it contains only three predictor variables of the original seven. Second, five states have been removed and placed into a separate evaluation file.

Your objective is to build a multiple linear regression model on the training data to predict the per capital sales of cigarettes (measure in packs) for the five states in the evaluation data set. You can only use the variables given to you (or variables that you derive from the variables provided). Below is a short description of the variables of interest in the data set:

- Age: median age of persons living in the state (predictor variable)
- Income: per capita personal income for the state (predictor variable)
- Price: weighted average price (in cents) of a pack of cigarettes in the state (predictor variable)
- Sales: per capita sales of cigarettes (measured in packs) in the state (response variable)

Deliverables:

- A write-up submitted in PDF format. Your write-up should have four sections. Each one is described below. You may assume you are addressing me as a fellow data scientist, so do not need to shy away from technical details.

- Assigned predictions (cigarette sales) for the evaluation data set.

- Include your R statistical programming code in an Appendix.

Write Up:

1. DATA EXPLORATION (25 Points)

Describe the size and the variables in the cigarette training data set. Consider that too much detail will cause a manager to lose interest while too little detail will make the manager consider that you aren't doing your job. Some suggestions are given below. Please do NOT treat this as a check list of things to do to complete the assignment. You should have your own thoughts on what to tell the boss. These are just ideas.

a. Mean / Standard Deviation / Median
b. Bar Chart or Box Plot of the data
c. Is the data correlated to the target variable (or to other variables?)
d. Are any of the variables missing and need to be imputed "fixed"?

2. DATA PREPARATION (25 Points)

Describe how you have transformed the data by changing the original variables or creating new variables. If you did transform the data or create new variables, discuss why you did this. Here are some possible transformations.

a. Fix missing values (maybe with a Mean or Median value)
b. Create flags to suggest if a variable was missing
c. Transform data by putting it into buckets
d. Mathematical transforms such as log or square root (or use Box-Cox)
e. Combine variables (such as ratios or adding or multiplying) to create new variables

3. BUILD MODELS (25 Points)

Build a multiple linear regression model using all three predictor variables. Try leaving each of the predictor variables out one at a time. Do any of these new models seem to be better than the full model? Explain. Try using each of the three predictor models as a single predictor in a simple linear regression model. How well does each one do? How much variation does each predictor seem to account for? Explain.

Be sure to explain how you can make inferences from the models, discuss multi-collinearity issues (if any), and discuss other relevant model output. Discuss the coefficients in the models, do they make sense? Are you keeping the model even though it is counter intuitive? Why? The boss needs to know.

4. SELECT MODELS (25 Points)

Which of your seven models would you use? Why? Decide on the criteria for selecting the best multiple linear regression model. Will you select a model with slightly worse performance if it makes more sense or is more parsimonious? Discuss why you selected your model.

For the multiple linear regression model, will you use a metric such as Adjusted R^2, RMSE, AIC, etc.? Be sure to explain how you can make inferences from the model, discuss multi-collinearity issues (if any), and discuss other relevant model output. Using the evaluation data set, evaluate the multiple linear regression model based on (a) mean squared error, (b) R^2, (c) F-statistic, and (d) residual plots.