Project #1: Diamonds

Background

ICE is an online clearinghouse of information for the wholesale diamond industry. More than 300 jewelers use the ICE database and online marketplace to purchase and sell diamonds. A dealer wishing to sell a diamond posts its asking price and physical characteristics to ICE's database. Prospective buyers contact the seller directly. Dealers pay a fee to have their diamonds listed on ICE. Dealers who use ICE range from large corporate jewelers to pawnshops that may have only one diamond for sale.

The price of a diamond is primarily determined by its physical characteristics, known in this business as the 4 C's: color, carat, cut and clarity. Larger diamonds, colorless diamonds, and diamonds with fewer imperfections sell for more. The weight of a diamond is measured in carats. One carat equals 0.2 grams. A 0.05-carat diamond is relatively small; larger ones exceed three carats.

A diamond's clarity is graded on a scale from flawless to imperfect. A gemologist examines the diamond under 10x magnification to determine the number and position of flaws. A gem that is free from internal inclusions (imperfections) receives the highest grade of IF—internally flawless. Slight external "blemishes" are allowed on IF-graded diamonds. A score of VVS1 (very, very slightly included) means any inclusions and/or external blemishes are very hard to locate. VVS2 is a slightly lower grade than VVS1. Grades of VS1 and VS2 (very slightly included) mean any inclusions and/or external blemishes are hard to locate. SI means slightly included—a diamond is graded as SI1 or SI2 if inclusions or external blemishes are easy to locate under magnification. On some diamonds, inclusions or external deficiencies can be seen by the naked eye. Those whose flaws are possible but hard to locate without magnification, but are obvious under 10x magnification, are graded I1 for imperfect.

Colorless diamonds are more highly valued than tinted ones. To grade a diamond's color, it is compared to a master set of gems that are internationally accepted as the industry's benchmarks. Diamonds are graded on a scale from D for completely colorless to Z, the most yellowish. Diamonds with D color grades are expected to sell for more, all else the same.

The cut of a diamond means many things, including its proportions, polish, and symmetry. These factors determine the appearance, brilliance, and expected life of a gem. Two common measurements are used to determine a stone's cut. Depth is the height of a stone relative to its width. Table is the width of the flat topside surface of the stone relative to the overall width. Values of 60 for both depth and table are considered optimal. For example, a diamond whose height relative to its width is less than 60 is shorter and thicker than the desired optimum and a diamond whose table exceeds 60 has an undesirably large face. Thus, the value of a diamond is expected to decline as its table and/or depth move away from sixty.

Grading the physical characteristics of a diamond requires the subjective opinion of an expert gemologist. Because small differences in appraisals can significantly influence a gem's

These data are real, although the name of the online diamond marketplace has been changed. Because of the sensitivity of these data, the price variable has been changed by an undisclosed, but constant, amount for each diamond. Further information about diamond ratings may be found on the website of the International Gemological Institute. These data were initially collected and studied by John Kejr, Bradley P. Pepper, statistical analyst, and Chris M. Olsen.

value, independent laboratories are often used to verify a diamond's attributes. Most diamonds come with registration documents and are inscribed with the evaluating laboratory's name and registry number. Diamonds evaluated by the "J" laboratory, the most prestigious evaluators in the industry, typically command a higher price. Gems without an independent evaluation are less valued in the marketplace due to the uncertainty surrounding their true physical characteristics.

Your Task

You have data containing diamonds in the ICE database during a one-week period in March of a recent year; the data set is named "Diamonds." The variables in that data set are those described above along with the wholesale asking price of the diamond in U.S. dollars. An additional column stipulates the training and validation partitions.

Your task is to use the data in the data file named "Diamonds" to develop a predictive model designed to predict the wholesale asking price of any diamond that might be listed on ICE in the future.

Once that model has been developed, you will apply it to new data obtained from ICE during the first week in April. That data set is called "New Diamonds" and will be made available later in the course.