DatasetMemo

Ben Hertzberg

2022-10-03

Overview of Dataset

The dataset I've chosen includes information on used car listings from dealerships around the Unites States. I downloaded the data from Kaggle: https://www.kaggle.com/datasets/3ea0a6a45dbd4713a8759988845f1a58038036d84515ded5
resource=download&select=ca-dealers-used.csv

The set includes over 7 million observations, but I will be truncating it to the first 100,000 observations so it is easier to handle. There are 21 variables associated with each observation. One is the price, which is the value I will attempt to predict. Some variables, such as dealership name or VIN, I will not use in the model. I decided to use 14 of the variables provided as predictors. There are both quantitative and qualitative predictor variables. The target variable is quantitative, which means this will be a regression model.

In the last code chunk, I used a function to find out how many total missing data points there are in the first 100,000 observations, across all 21 variables. The function stated there were 117,411 missing data points, out of a total of 2,100,000 possible data points. Some initial thoughts on how to deal with this is to ignore the variable's affect on the target for an observation where the data is missing for that variable. Another idea is to reach further into the original data set and collect only observations without any missing data.

Overview of Research Question

I am interested in predicting the listing price of a used car. Specifically, I want to know how what the price would be given information about the type and condition of the car. The response variable I will be using is called price in this data set, and provides the price that a certain used car was listed for at a dealership. My goal will be best achieved using a regression approach because the target variable is quantitative. I think the make, model, year, and mileage of each observation will be especially useful in predicting the price. The goal of my model is predictive, because I am not testing a theory or visually emphasizing a trend but rather attempting to predict a target variable given some data.

Proposed Project Outline

I have my data set loaded already. A general timeline I'd like to follow is described below. Week 1: Github Prep and Data choice Week 2: Exploratory Data Analysis Week 3: Data Splitting Week 4 onward: Fit 4 models and narrate process Week 10: Tune Best model

Questions/Concerns

I am a bit worried about working with such a large data set, but I also know having more observations can be helpful. I think fine tuning my final model will also be tricky.

```
#headuc = head(us-dealers-used, 10)
#headuc
```

```
#headuc2 = head(us_dealers_used, 100000)

#sum(is.na(headuc2))
```