Midterm Report
Of
Used Car Trading Price Prediction

Tongtong Zhao

Brown University DSI

Oct 14th 2021

https://github.com/Christy9615/DATA1030FinalProj

I. Project Introduction

- Dataset Overview --- Used Car Trading Price Prediction
  - Comes from Tianchi Competition Website
  - Two datasets provided: Training & Test
    - Features intro, size of two dataset

Categorical	seller/ OfferType/ bodyType/ fuelType/ gearbox/ notRepairedDamage regDate/ creatDate/ regionCode/ model/ brand
Continuous	'power', 'kilometer', 'v_0', 'v_1', 'v_2', 'v_3', 'v_4', 'v_5', 'v_6', 'v_7', 'v_8', 'v_9', 'v_10', 'v_11', 'v_12', 'v_13','v_14', 'price'

BROWN

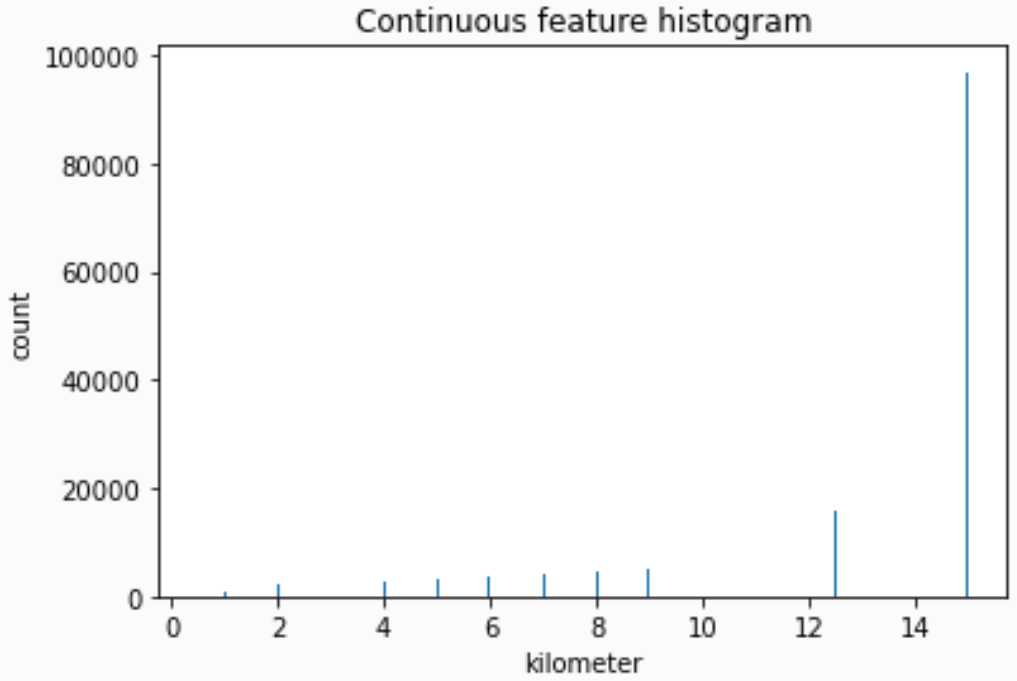
I. Project Introduction(continue)

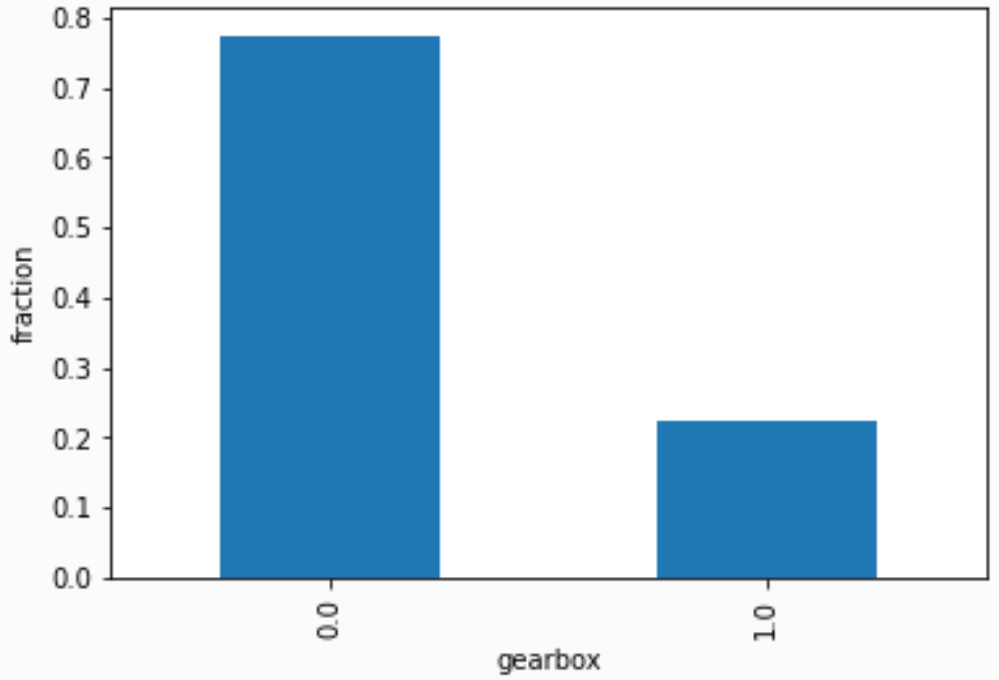
- Project Goal: Predict the trading price of used cars in the test dataset
  - Why is it important/interesting?
- Problem Type: Regression or Classification (Difference)?
  - Ans: Regression
    - The target variable (price) is a quantity, it's numerical

BROWN

II. EDA

- Sample graph of continuous and categorical features (using histogram to visualize continuous features, barplot for categorical)

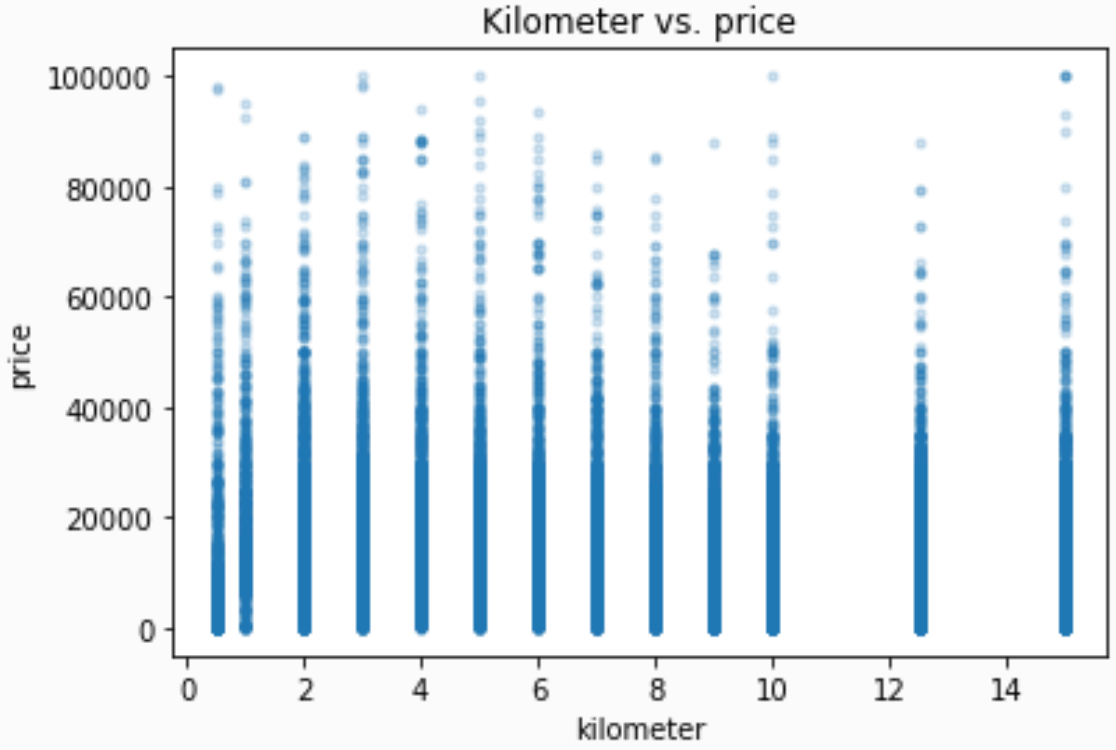




BROWN

II. EDA(continue)

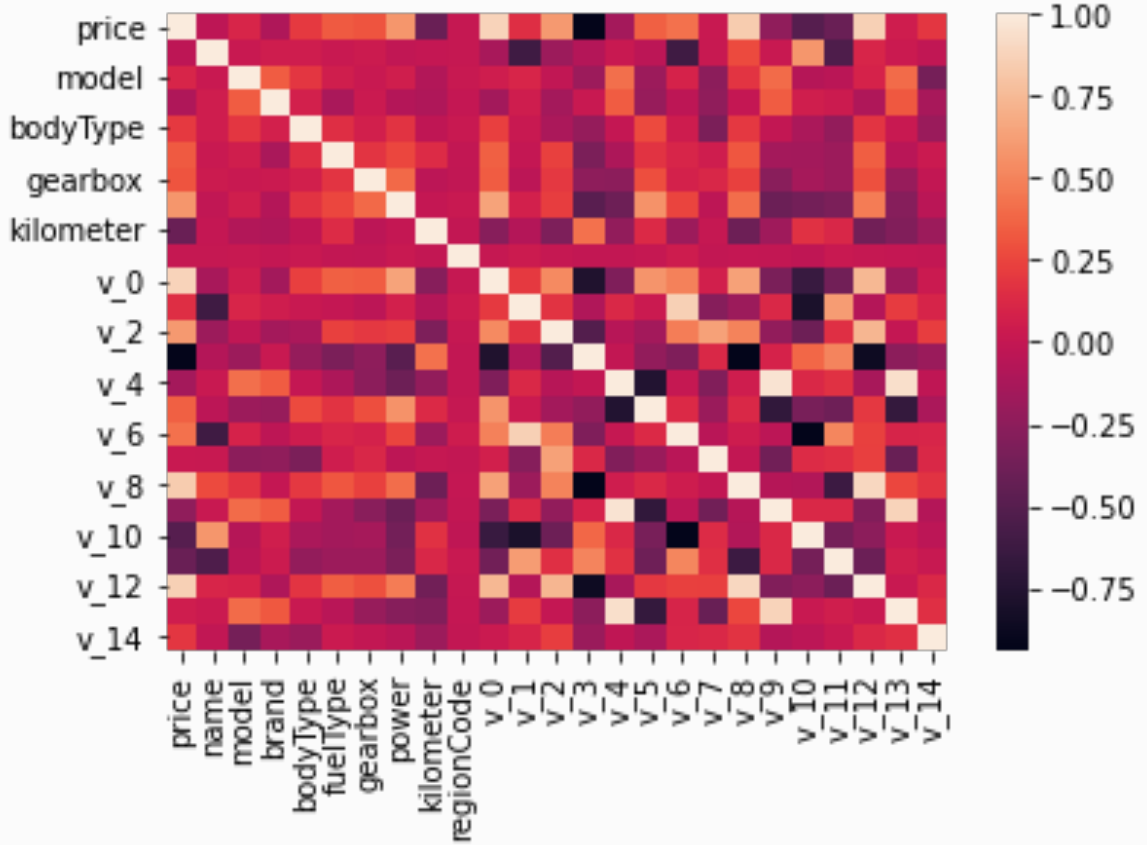
- Sample graph of Continuous vs. Continuous (Scatter plot)



BROWN

II. EDA(continue)

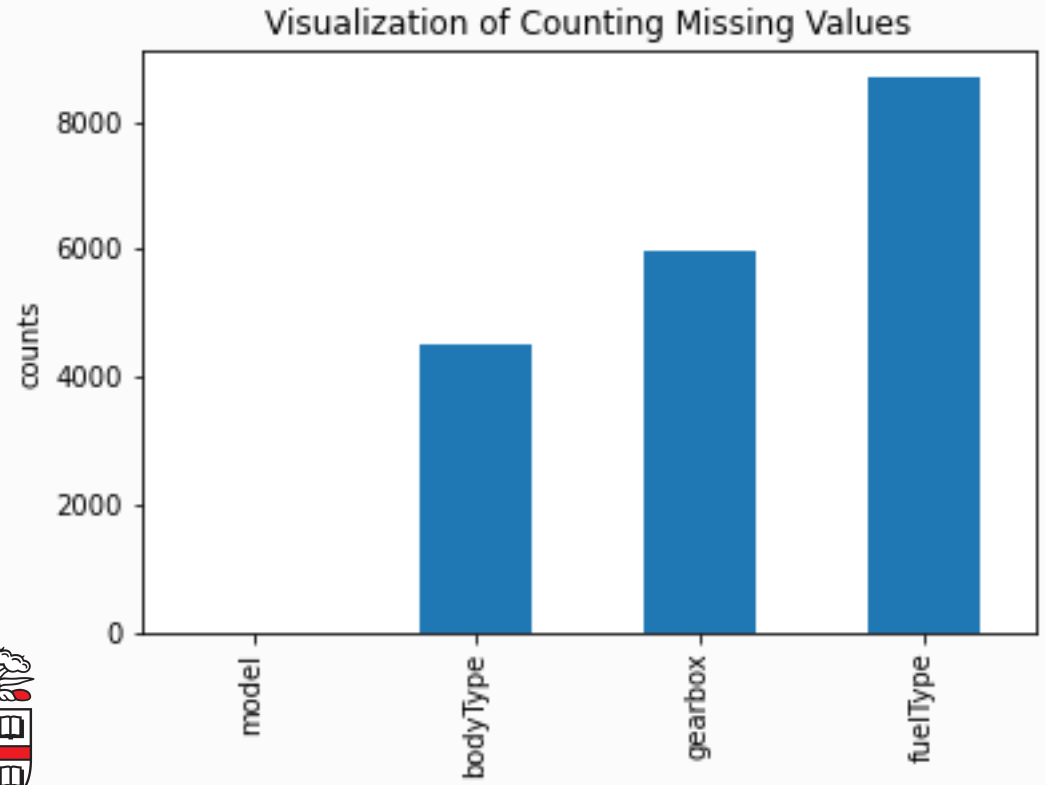
- Surprise me: Generate heatmap to see the correlation between each feature to price

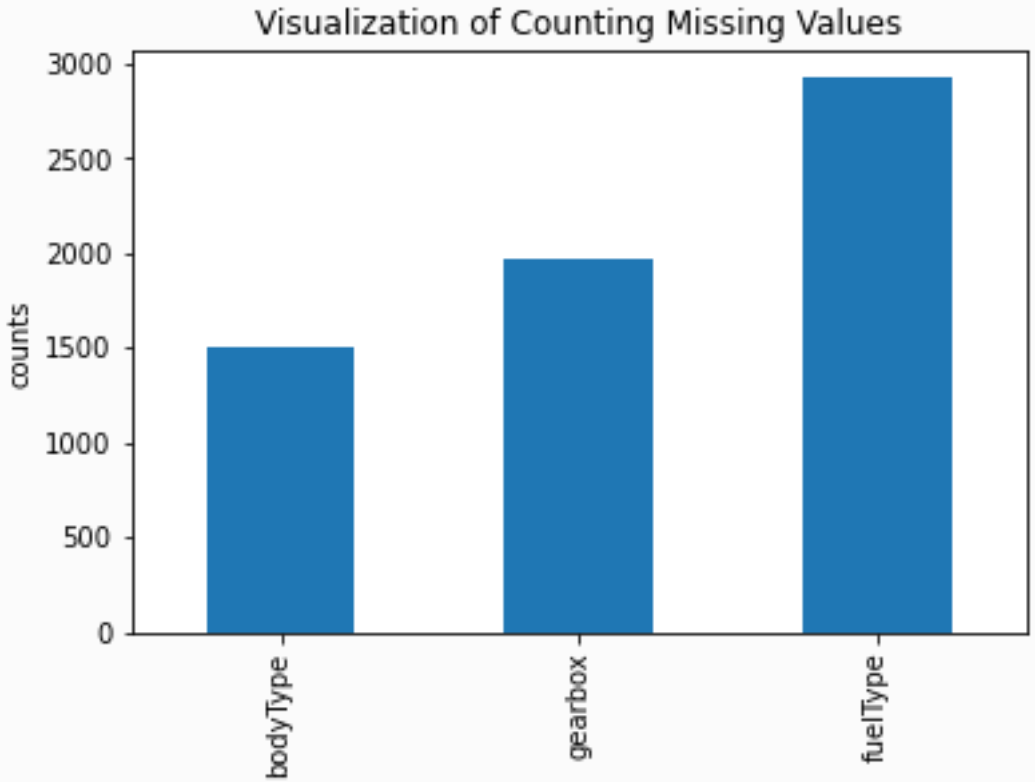


BROWN

III. Data Splitting and Preprocessing

- Missing value:





BROWN

III. Data Splitting and Preprocessing (continue)

- Splitting train dataset
  - Using Pareto Principle to split, 80%(train+val), 20% test
  - Method used: GroupShuffleSplit, v0_v14 are considered as group structure in the dataset.

```
other: [     0      1      3 ... 149997 149998 149999] test: [     2      5     29 ... 149982 149987 149992]
number of train dataset + val dataset 120925
number of test dataset 29075
Validation: [     0      4      5 ... 120921 120922 120923] TRAIN: [     1      2      3 ... 120917 120920 120924]
number of val_data 60442
number of train_data 60483
```

BROWN

III. Data Splitting and Preprocessing(continue)

- Preprocessing train dataset
  - Using StandardScalar for the continuous features, OneHotEncoder is applied to the categorical features
  - 28 features, 150,000 data points in the processed data
  - Using Label encoder, tried to preprocess the target variable, but still debugging

IV. Reference

1. “天池_二手车交易价格预测数据分析.” 开发者的网上家园, www.cnblogs.com/cgmcoding/p/13279789.html.

2. 零基础入门数据挖掘 - 二手车交易价格预测赛题与数据-天池大赛-阿里云天池. tianchi.aliyun.com/competition/entrance/231784/information.

BROWN

Appreciate Your Time and Patience