Tongtong Zhao
Andras
DATA 1030
Oct 12th 2021

Midterm Report of Used Car Dataset

I. Introduction

As the Covide-19 keeps impacting all kinds of things globally, the used car's price cannot escape from this effect. For example, in China, the price has increased more than 40% compared to two years ago. Thus, the prediction of the price trend of the used car has become a popular topic now. Therefore, this project will benefit the people interested in the topic and the car dealers and customers in that field. The dataset is downloaded from a Machine Learning competition website ---"Tianchi," the competition is still ongoing; only a few EDA has been posted for this dataset, no publication is available yet. Those EDA did missing value replacement as well as visualization. There are two parts of this used car dataset: one is for training and one for testing. The only difference is that the training dataset has the target variable "Price," which does not exist in the test dataset. In addition, there are 30 features plus one target variable in the training dataset, and the test dataset has 30 features same as those in the training dataset. Besides, 150000 data points are stored in the training dataset, 50000 data points are in the test dataset, and all of them are desensitized. The 30 features are: ID, features of a used car(name, fuel type, power, registered date, created date(the date that car has been posted), kilometer that it has been driven, body type, brand, model, gearbox, repaired for damage or not, region code, offer type(request or providing), seller), v0 to v14 (15 features) are embedding vectors that come from car evaluation, customer experience and so on. Since the project's ultimate goal is predicting the price for the cars in the test dataset, which is considered a quantity, it is a regression problem.

II. Exploratory Data Analysis

The graphs below are generate for EDA purpose.

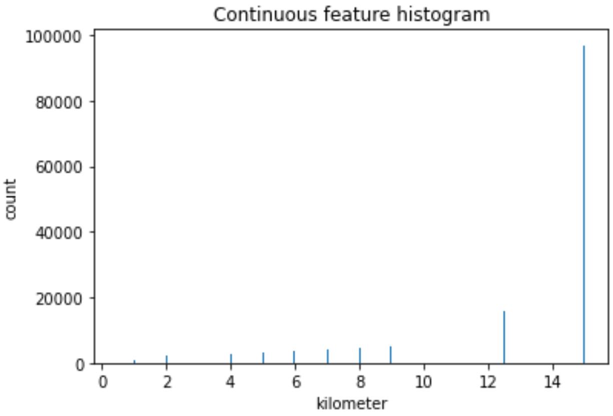

Figure 1 This histogram shows the distribution of used cars' driving distance. The x-label is variable name, and the y-label is the value counted for each "category" of the variable. From the figure, most of used cars' driving distance are located above 140000 meters range, which may has a positive correlation to the target variable ---"price."

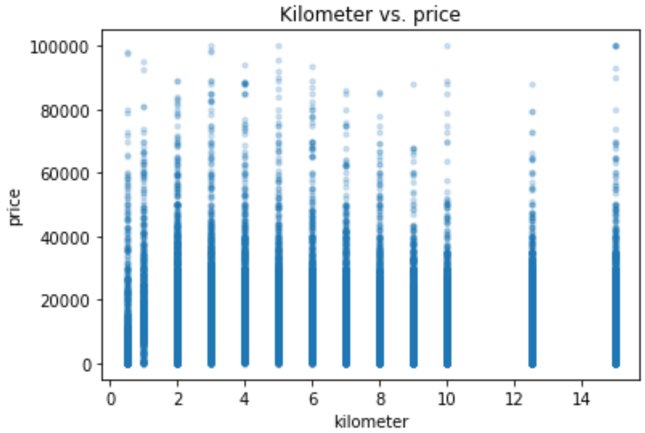

Figure 2 This scatter plot shows relationship between trading price and kilometer. The cars that have a high driving usually has a lower trading price. But there also some cars that have a low range in kilometer get a low value of price. Similarly, there is an outlier in the greater than 140000 meters range, which means that car has a 140k+ driving distance but has a trading value about 100k RMB. This is one thing that will be discussed and explored in the future analysis.

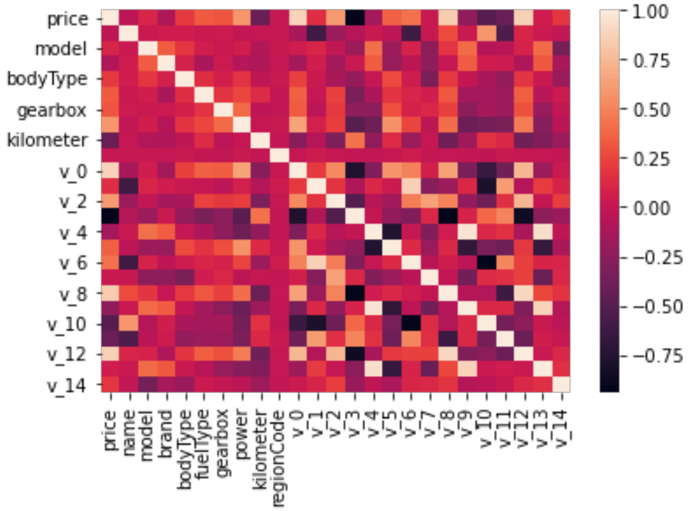

Figure 3 This heatmap shows the correlation between each feature to the target variable (price). V_0, V_8, V_12 and power are highly related to the price. Kilometer is not highly related to price as the hypothesis that stated in last figure. Those positive correlated features will be the important ones that need to pay extra attention in the future analysis.

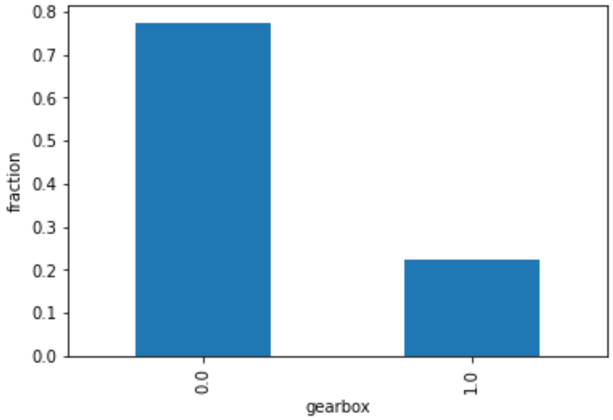

Figure 4 The bar plot shows the 2 types of gearbox, 0.0 stands for manual shift, 1.0 stands for automatic shift. To conclude the information from the graph, used car market has more manual shift cars than automatic shift cars.

III. Methods

3.1 Data Splitting and Preprocessing

After the EDA steps, the idea of how to split dataset is getting clear. The dataset will be split into 20% of testing and 80% of training data, the training data will use second split to get 50% validation dataset and 50% training dataset, which are finally 40% validation data, 40% training data and 20% testing data. The splitting method is based on the goal of try to get high accuracy of the prediction but not "over-use" the training data so that it may cause not enough data for testing, then fail the prediction in the separated cvs file, which is the real test set. The dataset is IID, which means each trading price are independent from next one, that is also important in the Machine Learning process. The data also has group structure, which are v_0 to v_14 features corresponds to each data points. There are features that related to the time, such as registration date and created data. In the next step, those two features may be generated to a new feature, like a time slot for each car. The StandardScalar are used on the continuous features (v_0 to v_14, power and kilometer) but except the SaleID, those features have lots of unique values, not follows certain pattern. The OneHotEncoder is applied to the categorical features (such as bodyType, fuelType, gearbox, notRepairedDamage), those features are categorized into bounded number of categories. There are 21 features in the preprocessed data.

IV. Reference

"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.

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.

V. Github repository

https://github.com/Christy9615/DATA1030FinalProj