Explore and Analyze Used Cars Market in the U.S.

1.Problem

Over the past year, as the pandemic disrupted supply chains and caused shortages in critical auto components, resulting in a lack of new vehicles, which caused the used car price surging.

According to the U.S. Bureau of Labor Statistics' Consumer Price Index, used car prices are up a staggering 39.8% since March of 2020, while the U.S. inflation is only up 6.3%. JPMorgan analysts say prices will continue to rise and will stay high for longer than expected.

It is difficult for people to use previous data and experience to judge whether the price of a used car is a good deal. So, our team try to create a method determine whether the price of a used car is appropriate.

2.The Dataset

The dataset we used in this project was download from Kaggle. The provider of the dataset scraped data from Craigslist every few months, it contains all relevant information that Craigslist provides on car sales. We selected data from the beginning of 2020 and the same period in 2021.

There are 426880 rows and 26 columns in dataset of 2021 and 539759 rows and 25 columns in dataset of 2020.

We first to drop some unnecessary features, like 'id', 'url', 'VIN', 'image_url', 'description','county','posting_date','lat','long', etc.

Then we found there were some outliers in numerical variables. In the dataset of 2021, the target variable price ranges between 0 to $3.7 billion. We transformed the price value into log data and removed the outliers of price by using 3-sigma rules. We finally got the price variable ranged from $320 to $105,000. The year variable ranged from 1900 to 2022, we only kept values from 1991 to 2021. For the odometer variable, the max value was about 10 million miles. According to the Federal Highway Administration, the Americans drive an average of 13,476 miles per year. So, we dropped odometer values higher than 40,000.

For missing values, there were some variables missing lots of values, we didn’t want to drop them in order to keep important features and enough data for our model. So, we kept all the values and dealt with them after exploratory data analysis.

我们对 2020 年的数据做了相同的处理，以便对比这两年同期的数据。

3.Exploratory Data Analysis

In this study, we will look at the various features of used cars and try to understand the patterns in the data.

3.1 Target Variable

Price: The price of used car prices, given in US dollar.

F: price distribution

分析 price 变量

3.2 Other features of Used Car:

odometer: The distance that the car has been drove after it being bought.

year: The year in which the car was manufactured.

manufacturer: Manufacture of the car.

model: The exact model of the car.

condition: The condition of the used car, including *excellent, good, fair, like new, salvage, new*.

cylinders： The number of cylinders in the car engine.

fuel： The fuel type of the car, including *diesel, gas, electric, hybrid and other*.

title_status: Including *clean, lien, rebuilt, salvage, parts only and missing.*

transmission: The transmission of the car, including *automatic, manual and other.*

drive: Including 3 types of drive transmissions: *4WD, FWD and RWD.*

size: The size of the car, including compact, full-size, mid-size, sub-compact

type: The generic type of the car.

State: The state that the car belongs to.

Paint_color: The color of the car.

结合参考文章、EDA single 和 multi 的图，分析每个变量及他们和 price 的关系。

Odometer

Year

There are 51240 cars for the last 5 years, which is 16.78% of the entire market.

There are 160951 cars for the last 10 years, which is 52.72% of the entire market.

Manufacture & Models

Type & Transmissions & Fuel

Condition

Drive & Title Status & Size & Type & Color

State (map 图)

3.3 Comparison Between 2021 and 2020

4. Data Preprocessing

4.1 Missing Value

各分类变量有不同程度的缺失。对于 missing value 最简单的做法是 drop 掉所有的 missing value 或者 remove missing value 过多的 column。但这些变量可能会对二手车价格产生影响，为了保证不丢失重要的变量和足够的数据，我们将所有的分类变量中的缺失值补充为 ‘UNKNOWN ‘

4.2 OneHot Encoding

There are 12 categorical variables and 3 numerical variables. We need to transform these categorical variables into numerical variables. Most of the current projects about used car prices usually use LabelEncoder to solve this problem. But LabelEncoder is suitable when the variables are equidisatant. For the categorical variables in our dataset, it’s better to use OneHot Encoding. 解释 One Hot Encoding。说明 model 变量过多，如果使用 OneHot Encoding 将会产生过多的 columns，影响运行速度。因此我们使用 fuzzy matching 对 model 变量进行处理。Fuzzy matching is to approximately match strings and determine how similar they are. 我们根据 fuzzy matching 计算出的相似分数将 model 变量进行分组，最后一共得到 70 个 value。

最后我们对所有的分类变量进行 One Hot Encoding，共得到 185 个 columns。

4.3 Normalization

To reduce the scale of odometer to prevent from dominating the prediction model. We used MiniMaxScaler to normalize odometer variable.

5. Model

In this section, different machine learning algorithms are used to predict price/target-variable.

1. Linear Regression
2. Linear Regression after Log Price
3. Decision Tree Regressor

4. Random Forest Tree Regressor
5. XGBoost

分别描述一下模型，最后说一下各模型的结果

6.Final Goal

我们最终选取 Random Forest 模型进行价格预测。通过对比一辆二手车的 listing price 和预测价格，来判断这辆二手车的价格是 good deal 还是 bad deal。

如果 listing price 小于预测价格，是 good deal
如果 listing price 大于预测价格，是 bad deal
如果。。。等于。。。，是平均价格

7.Conclusion