|  |  |
| --- | --- |
|  |  |

**Introduction**

Nowadays, vehicles become an indispensable part of out lives. Buying a new car is sometime too expensive, especially for the young people. Therefore, buying used cars have become a better choice. Unlike new vehicles, used cars usually do not have a certain price. The price of a vehicle normally depends on several key factors and my analysis is to use these key factors to predict the selling price.

The data set supported my analysis is from Kaggle, a platform for data scientists and machine learning practitioners to find and publish data set, explore and build model and join computations to solve data science challenges. The data set has been collected in 2015, and it contains historical car auction information. The data set contains total 558,838 vehicles’ information and 16 variables: year, make, model, trim, body, transmission, vin, state, condition, odometer, color, interior, seller, mmr, and selling price.

**Data Preparation**

1. Data cleaning:

After loading the data set into R, we first delete empty rows and duplicated rows. After the cleaning process, 422,920 rows of data remained. Next, in order to do the prediction analysis, we want to remove irrelevant columns. For example, vin number is only used to identify vehicle. It’s not helpful for our prediction. Finally, only 7 useful variables were kept: year, make, model, body, condition, odometer and selling price. The following Table1 shows details of the seven kept variables.

|  |  |
| --- | --- |
| ***Year made*** | Year that vehicle was made. E.g.: 2014 |
| ***Make*** | Vehicle’s make. E.g.: Honda, Ford, Mercedes |
| ***Model*** | Vehicle’s Model. E.g.: CR-V |
| ***Body*** | Vehicle type. E.g.: SUV |
| ***Condition*** | Vehicle condition rate from 1 – 5 points |
| ***Odometer*** | Miles vehicle has been driven |
| ***Selling price*** | Final selling price |

Table 1

1. Data cutting:

Because this data set is large, we want to cut the following four variables into several small chunks: condition, odometer, year made, and selling price. For example, we cut the variable condition into six pieces: score 0 to 1 goes to category “terrible”, score 1 to 2 goes to category “poor”, score 2 to 3 goes to category “neutral”, score 3 to 4 goes to category “good”, score 4 to 4.5 goes to category “great” and score 4.5 to 5 goes to category “excellent”.

1. Preparing train and test dataset:

Data will be randomly separated into 75% training set and 25% testing set.

**Descriptive Analysis**

The following three graphs help us to have a better idea how the data distribute.

Figure1 shows the number of vehicles in different conditions. The graph is nearly normal distributed, means the condition of most of the vehicles are acceptable.

图表, 条形图

描述已自动生成

Figure1: The distribution of vehicle condition.

Figure2 shows the distribution of vehicles’ selling price. The graph tells us that the data is right skewed, means we don’t have that many luxury vehicles. Most vehicle price fall on the low-price interval: 0-30000.

图表, 直方图

描述已自动生成

Figure2: The distribution of vehicle selling price

Figure3 shows the distribution of vehicle’s age. The graph tells us that the data is left skewed, which means that we are buying used vehicles mainly from the past 10 years.

图表, 条形图

描述已自动生成

Figure3: The distribution of vehicles’ age

**Predictive Analysis**

For predict the selling price, there are three methods I use in my analysis: Naïve Bayes, KNN and Random Forest.

Cross validation is used to evaluate the model performance in train set. For KNN and Random Forest model, the number of cross validation folders is fixed to 3. For Naïve Bayes model, the number of cross validation folders is set to be 10. The reason is that Naïve Bayes model’s accuracy is unusually high after we run it first time. To prevent any mistake, we set up cross validation 10 times.

Naïve Bayes has the highest overall accuracy (99.98%). The overall accuracy of each model is summarized in Table 2.

|  |  |
| --- | --- |
| Model | Accuracy |
| Naïve Bayes | 0.9998 |
| KNN | 0.6820 |
| Random Forest | 0.5772 |

Table 2: Model Accuracy

After finishing the predictive analysis, I personally created a table which has two of my vehicles’ information (see table 3). I wanted to see whether the model can predict the price of my vehicle correctly and to see if I spend too much on my vehicles.

The result for Jeep Compass is 25000-30000 USD and result for Chevy Silverado is also 25000-30000. The real price I bought for Jeep Compass is 25300 and the real price I bought for Chevy Silverado is 32000. I might spend a little bit more on the Chevy, but the model does tell me that I paid just right for the Jeep.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Year made | make | model | body | condition | odometer | Selling price |
| recently made | Jeep | compass | SUV | very good | 30000-50000 miles | |
| recently made | Chevrolet | Silverado | crew cab | excellent | 10000-30000 miles | |

Table 3: my own vehicles information

**Conclusion**

By considering the performances of model evaluation, prediction, and confusion matrix, Naïve Bayes is the most appropriate model to predict the selling price for used vehicles. Moreover, I deem that no matter what, vehicle’s overall safety condition should be the most considerable factor when buying used car, because life only once.