Used Car data exploration, price analysis, and insights for used luxury car standards

Group Member: Yiyuan Cui (A15438228), Naiwen Shi (A15554690)
Email: y2cui@ucsd.edu, nashi@ucsd.edu

Abstract

In this project, we were able to explore the data we found on Kaggle, and we were able to find out what are the important features for us do decide if a car can be considered luxury, based on the features correlated to the label we created. And we were able to classfy if a car is indeed a luxury car by our machine learning algorithms. On top of that, we performed a detailed exploration on the data, and see what are the deciding factors for the selling price of a used car. In the end, we tried to explore the price difference between similar car brands like Toyota, Honda, Nissan, or BMW, Mercedes-Benz, Audi.

Keywords: Random Forest Classifier, KNN Classifier, Gaussian Naïve Bayes Classifier

Introduction

Nowadays, people tend to buy luxury cars if they have what it takes to do so. Because luxury cars are able to provide a more premium driving experience to people who spend a lot of times in car. For example, luxury cars will have more expensive car paint to make the car look nicer, or leathre interior with multiple choice of colors; luxury cars tends to spend more budge on eliminate round noise or wind noise when you drive the car, and it provides more sophisticated entertainment system to drive or more advance automatic driving abilities.

Thus, we want to use this dataset to provide a detailed analysis on what are the features that decides the selling price of a used car. As for similar car brands(Toyota, Honda, Nissan or BMW, Mercedes-Benz, Audi), we want to prove which one of them have the lowest selling price on used market, and what are the distribution of their car prices.

By using machine learning algorithms such as Random Forest Classifier, K-Nearest Neighbor Classifier, and Naïve Bayesian Classifier. we are able to find out some of the most important features that are correlated with deciding if a car should be considered luxury. And we are able to achieve a 82% accuracy with our current model.

Data Pre-Processing

The data we used have 558837 rows and 13 columns. Some of the important features we focused on are odometer, condition, selling price, make, interior color, paint color, body. The data needed a lot of cleaning. For example, listing of body type were consist of 7 different body type which are ‘SUV’, ‘Sedan’, ‘Convertible’, ‘Coupe’, ‘Wagon’, ‘pick-up truck’, ‘Van’. But they have 20 different ways of calling Coupe, and 32 different ways of decribing a pick up truck. After cleaning up the data, we used the cleaned version to do the data exploration.

As for using machine learning algorithms to classify luxury car brands, we selected 8 different car brands to do the classification. Consist of 4 luxury brands and 4 economical brands(Luxury: BMW, Mercedes-Benz, Audi, Infiniti. Economy: Toyota, Nissan, Honda, Kia). We intentionally included Infiniti which is normally considered as lower priced luxury car brand to make the model more robust. Since the car price kind of fall between Luxury average and Economy average. And we labeled the data with binary values on luxury being 1, and economy being 0 to achieve a better performance with our machine learning algorithms.

In order to perform machine learning algorithms and achieve the best possible performance from the algorithm, we first dicovered that the data is unbalanced, with 50330 luxury data, and 119966 economy elections. So we performed oversample to make sure the binary class have same number of data. On top of that, we group several close colors into groups and hopefully get better results from more samples in each color category. For example, we grouped grey and silver interior to be the same group called grey group.
We also drop several features to make sure the model is learning the right perameters(model, trim, make, year, state ,selling price). And we implemented one hot encoding to categorical features like interior color, paint color, as well as transmission type.

Exploratory Data Analysis

As we all know, luxury car brands tend to have a higher selling price than economical car brands. Thus we want to start our data exploration with finding relationships between selling price and other deciding features we could use in the dataset.

In the following figures, we will show some deciding features like Odometer readings(**Fig1**), number of years since certain vehicle was produced(**Fig2**), condition of the car(**Fig3**), as well as interior color or paint color of the car(**Fig4**), all compared to selling price.

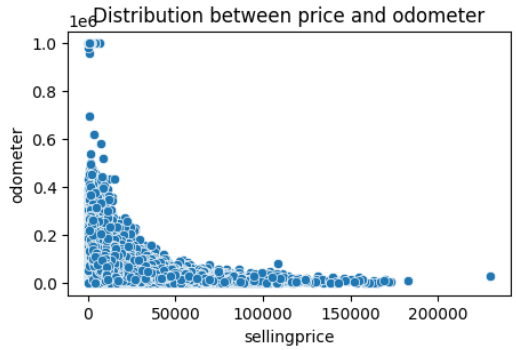

Fig1: This scatterplot shows realationship between selling price and odometer reading. When odometer readings are low, selling price will have a higher value.

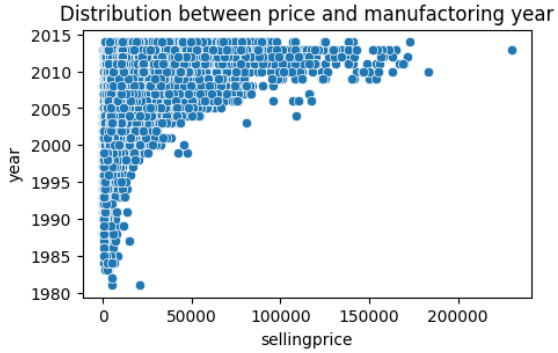

Fig2: This scatterplot shows realationship between selling price and production year of the vehicle. When the car's ages are larger, selling price will have a lower value. On the contrary, selling price goes up, when the year value are close to recent years.

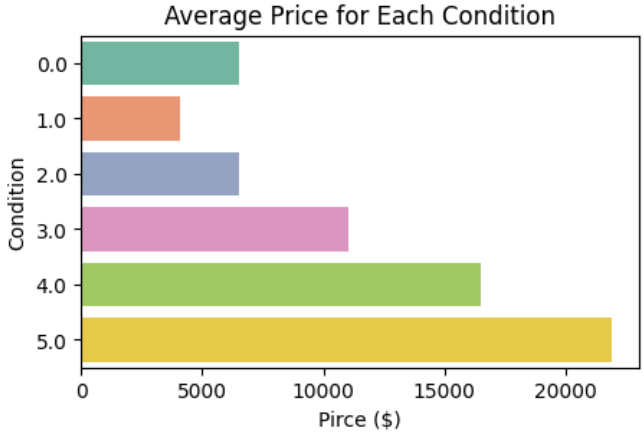

Fig3: This scatterplot shows realationship between selling price and condition of the vehicle ranges from 0,as the worst to, to 5, as the newest condition. When the car's condition are bad, selling price will have a lower value. On the other hand, selling price goes up, when the condition is as good as brand new cars.

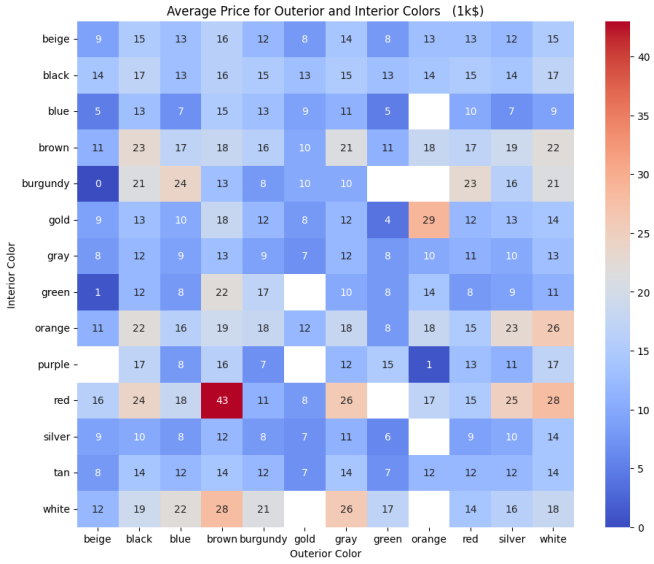

Fig4: This heatmap shows average price for every signle finds of combination between interior and outerior color that exists in the dataset.

In the figure above, we saw that common colors like black or white interior tends to have a higher resale values with all kinds of paint colors of the car. And gray and silver interior tends to have a lower resale value compares to other interior colors. This discovery closely corresponds to our machine learning algorithm we will talk about later. For the highest value we saw in the heatmap, red and brown combinations, the car we could think about is Porsche Caynne at 43k average price. And this make a lot if sense to us become some of the luxury brands like to use red interior which is also a premium option when purchasing a car.

After finding all those relationships, we wanted to address the myth about lowest selling price and distribution between three major (Fig5) Japanese car manufactors(Nissan, Honda, Toyota) and (Fig6) Germany car manufactors(BMW, Mercedes-Benz, Audi).

In Fig 5 we could clearly say that the Nissan's car price didn't really changed in the last 5 years, but Toyota and Honda's car price have a clear tendency of growing in the past five years. As for the outliers, Nissian have the well known GTR listed between 60k- 87k. Honda is the most well behaved between the three car brands. And Toyota also produce famous sports car like Supra or family used seven seater, Sienna, that could explain the outliers.

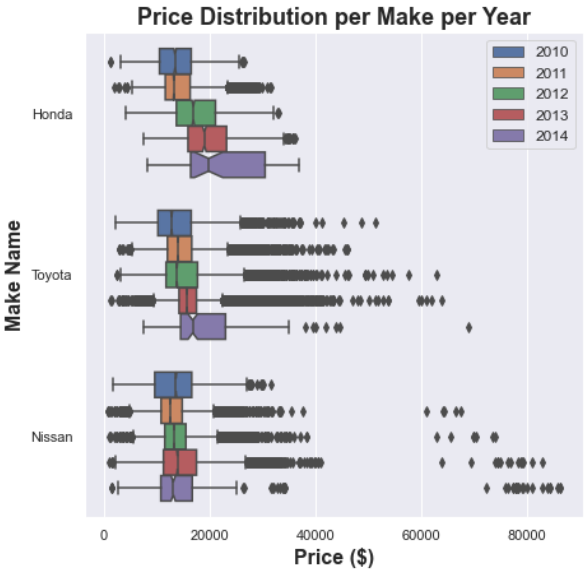

Fig5: This boxplot shows price distribution for Honda, Toyota, Nissan

Fig6: This boxplot shows price distribution for BMW, Mercedes-Benz, Audi.

In Fig 6, we could clearly see the price ranking of the three car brands are as follows: Mercedes-Benz, BMW, Audi. As Audi being the lowest among the three car brands. The reason why 2014 Audi's price have lower values is because the number of luxury models are too low, and most of those data are for audi A3s in the data. And the outliers for Audi are Q5, and A6s which are in much higher selling prices. As for the outliers in Mercedes-Benz and BMW, they are having crazy luxury cars like S class or 7 series. And super performance cars like M series or AMGs. Thus the plot make a lot of sense.

Machine Learning Algorithms and Performance

Based on our exploratory data analysis, we found something very interesting which worth further discussion. The interior colour of the car seems to have an impact on the selling price of the car, i.e., luxury or economy. Thus, we want to drop the label of selling price, and perform machine learning on this dataset with binary label 1 as luxury, and 0 as economy. In other words, we want to see what features count toward making a car luxury. Instead of only looking at selling price, we want to use our models to see the feature importance contributing to the luxury standard of a car.

We tested on Serveral different machine learning algorithms such as Random Forest Classifier, K-Nearest Neighbor, Gaussian Naïve Bayes Classifier. And compared performance on each single one of them to see which one outperforms the other algorithms. We will list the top three condidates for the machine learning algorithms

1. **Gaussian Naïve Bayes**

A naïve classifier we used is GaussianNB in sklearn. GaussianNB is a classsifier that makes predictions by substituting the parameters with the new input value of the variable aqnd as a result, the Gaussian function will give an estimate for the new input value's probability.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \text{for } -\infty < x < \infty$$

2. **K Nearest Neighbors(KNN)**

The second model we implemented was KNN in sklearn. KNN is one of the easiest algorithms we learned by finding the shortest eucalidian distrance between the new data point to all other training data points. And most times, KNN provide an accpetable results to the classification problem we arise. But the disadvantage of KNN is that it does not work well with high dimensional data. Or if the data was too large, it could overfit easily.

3. **Random Forest Classifier**

The final machine learning algorithm we used was Random Forest Classifier. This type of classifier could be used to solve for regression or classification problems. This algorithm is made up of a collection of decision trees, and each tree in the ensemble is comprised of a data sample drwan from a training set with replacement, called the bootstrap sample. On top of that, Random Forest Classifier has the abilty to rank the feature importance, which would be very benefical to our purpose of using machine learning. And the result of this classifier was the best among all three machine learning algorithms we implemented.

Result

In this section, we will show the results of the models, and the feature important we collected from random forest classifier.

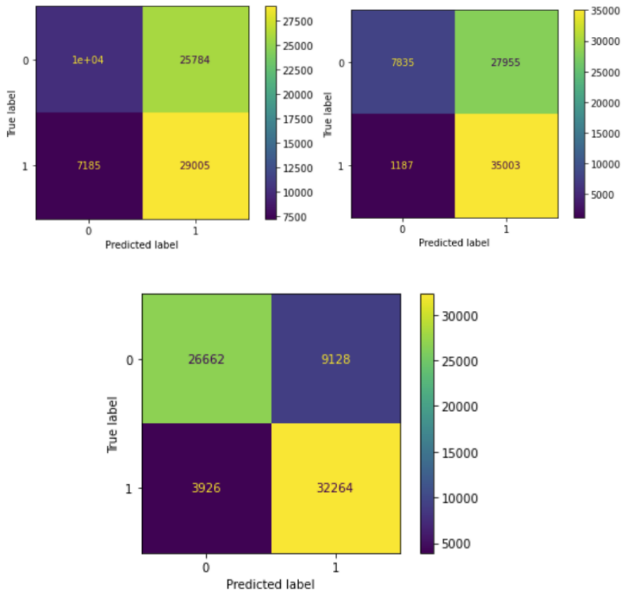

Fig7: The top left confusion matrix for three different classifiers

The top left confusion matrix is for GaussianNB classifier, and we were able to achieve 54% accuracy based on the model. So the performance is just as we expeceted, since the data distribution are not really Normal distribution. And we used one hot encoding, which clearly does not satisfy normal distribution as we discussed in class. The top right confusion matrix is for KNN classifer, and we were able to achieve 68.1% with k being set to 2. And we are already see improvement in our model. What stands out is the Random Forest Classifer at the bottom of **Fig7.** We were able to achieve 81.8% accuracy with this classifer.

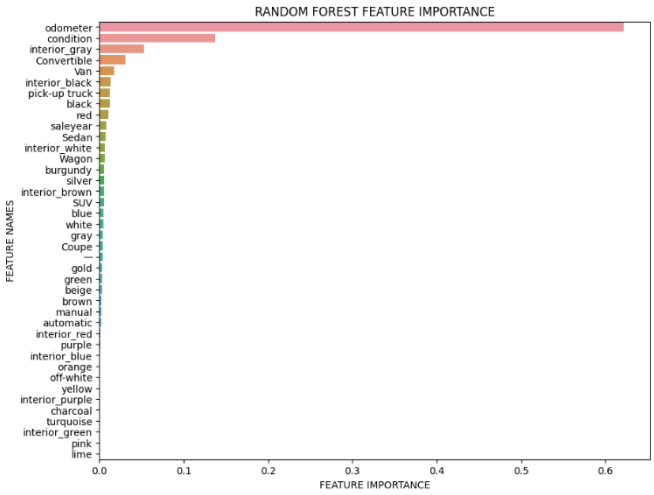

Fig8: The feature importance we got from random forest classifer.

And by using random forest classifer, we saw that odometer, condition, interior color as gray, convertible body type are being ranked the top 4 features for deciding if a car is luxury or not. We think odometer are closly representing the selling price, since those two have strong correlations. Same applies for conditions. But for interior color as gray, which represents gray or silver interior color, this result is very surprising. But in the end, the result makes sense. Based on our eariler discussion, gray or silver interiors tends to have a low price point with every single type of combination with car paint color. And that is exactly what random forest is doing with classifying luxury or economy. Based on my own experience working at Triton Transit, the 2014 honda civic have the gray interior, which represents economy car brands.

Future works

In this project, we only considered 4 luxury brands, and 4 economy car brands. We could include more car brands to make the model more robust, therefore improve the model accuracy. And getting more recent data might shift our results significantly. We could also fine tune the random forest classifer to get a better results.

Summary

In summary, we did data exploration, we found out how each siginifiant features are correlated to selling price. We addressed the interesting topic of choosing the best bang for the buck car brands when considering German made, or Japanese made. We also justify the results we obtain from comparing color combination of the car with average price by using machine learning algorithm. The stand out one we got is Random Forest Classifer, which also provides the feature importance ranking that matches our earlier discovery.

Source Code

https://github.com/yiyuancui/car-price-analysis

Reference

1. View of exploring the factors influencing the choice of young generation while buying cars: A factor analysis approach. (n.d.). Retrieved December 11, 2022, from https://www.ijcms.in/index.php/ijcms/article/view/268/257
2. Random Forest feature importance plot. Random Forest Feature Importance Plot in Python - AnalyseUp.com. (n.d.). Retrieved December 11, 2022, from https://www.analyseup.com/learn-python-for-data-science/python-random-forest-feature-importance-plot.html
3. Gutoskey, E. (2022, June 16). The 3 best (and 3 worst) car colors for resale value. Mental Floss. Retrieved December 11, 2022, from https://www.mentalfloss.com/posts/best-car-color-resale-value
4. Gorzelany, J. (2022, October 12). Here's how the color of your car will affect its resale value. Forbes. Retrieved December 11, 2022, from https://www.forbes.com/sites/jimgorzelany/2022/06/16/heres-how-the-color-of-your-car-will-affect-its-resale-value/?sh=21aee2284415
5. By: IBM Cloud Education. (n.d.). What is Random Forest? IBM. Retrieved December 11, 2022, from https://www.ibm.com/cloud/learn/random-forest