

**USED CAR PREDICTION**

Submitted by:

SAUNAK MUKHERJEE

**ACKNOWLEDGMENT**

This includes mentioning of all the references, research papers, data sources, professionals and other resources that helped you and guided you in completion of the project.

**INTRODUCTION**

With the development of modern transportation infras

tructure or the increasing needs for commuting and trav

elling, personal vehicles have more and more pop

ular among people. In a meantime, the used car market

has also grown rapidly to fulfifill the demands from the low

and middle-income groups. Yet, to set the prices of used

cars is never an easy job, as a lot of conditions of a used car

can signifificantly affffect its price, including but not limited

to its basic performance, how long it has been used, how

far it has been driven, etc. Thus, we are also interested in to finding of the relationship between the prices and those potential conditions, and we want to apply the some data mining and machine learning techniques to the develop a model to predict of the prices, which can provide references for used car sellers and buyers.

We analyzed the “CARS DATASET (Audi, BMW,

, Skoda, VW)” and predicted the prices

of used cars. The dataset is available on Kaggle via

link to the data. This data set has 5966 rows,

each representing one used car, and 6 columns with the

information of the brands, models, years, prices, transmis

sions, mileages, fuel types, road taxes, miles per gallon

(mpg), and engine sizes.

**Exploratory Data Analysis:**

We have to performed some preliminary exploratory data

analysis. This was to helpful for the understanding of the data more thoroughly and ready for the more advanced analyses in the next steps. We used to Pandas and Numpy [1] libraries in Python.

**Distribution of Categorical Features**

The relationships between price and categorical features

are visualized using the violin plots shown in Figure 4.

**Data Preprocessing and Feature Engineering**

During data collection, there are often some “out of range”

values which will cause inaccuracy during the data analysis.

Data preprocessing with enhances the quality of the data set

by data cleaning or the transformation. After preprocessing,

the final data set can be lead to the more precise and reliable

conclusion under the data mining algorithms. In this report,

based on our data set, we did the following steps:

1. model varies a lot and mpg makes no sense for electric and hybrid cars, so we dropped these two columns.

2. Applied log-transformations to make the data not so skewed.

**Modeling and Analysis**

After getting familiar with the data and doing some nec

essary transformations, we could start which are applying some regression models to predict the prices of used cars based to the attributes. To prevent the potential issue of over-

fifitting, we used 80% of the data to train the models and

the rest 20% to test the models. The model performances

were measured and compared by the metric of R-square

for the test set, which quantififies how much of the varia

tion in the prices can be explained by the models.

**CONCLUSION**

We created a function which is predict car price.

our function is to “car\_predict\_price”

We pass the values into our function and it will predict.

We got the accuracy 99.99% of the model