Used Vehicle Price Prediction

- Final Project for FA21-D590 Applied Data Science
  Group Members: Xinyu Zhou, Yunyang Zhao
  Code: Python (for EDA) & R (Web App)
  Web App URL: https://gixq9a-yunyang-zhao.shinyapps.io/ShinyApp/
  GitHub repository: https://github.com/zhaoy426/usedcar.git

Introduction

1. Background and Objectives

About 40 million used vehicles are sold each year [1]. The value of a used vehicle is not simply linearly decreased with year. The value of a car decreases by 20 percent of its initial value in the first year and decreases less fast after the first year [2]. There are many other factors affect a used vehicle's price as well. Mileage and condition are the most important factors that affect a used vehicle's price, while options, location, and color also play a role [3]. Therefore, developing a machine learning (ML) model to predict the price of used vehicles can help customers to find a desired vehicle with a reasonable market price. Creating a web application allows users to interactively explore data, visualize data and obtain used vehicle's price prediction based on customized feature selection.

2. Data Description

The "used-car-price-prediction" data from Kaggle was used for the project [4]. The data was originally collected from Craigslist, which is the world's largest collection of used vehicles for sale. Austin Reese creates the data for a school project. This data is updated every few months. The most recent data was updated 7 months ago (Version 10). There is only one csv data file with a size of 1.45 GB. The data contains 426,880 rows and 26 columns with 6 numeric variables, 19 categorical variables and 1 date time variable. The "price" variable is our target.

3. Tasks

Our team members are: Yunyang Zhao and Xinyu Zhou. Yunyang is responsible for exploratory data analysis using data visualization techniques and feature engineering to fill missing values and variable transformation. Xinyu is responsible for correlation analysis and building ML models to predict vehicle price. Yunyang developed three tab panels, including "Vehicle Data Table", "Vehicle Price by Year / Manufacturer" and "Correlation between Odometer and Price". Xinyu developed two tab panels, including "Regression models and train" and "Predictions". "About" tab was designed and developed by both Yunyang and Xinyu.

Project Design

1. Exploratory Data Analysis (EDA) and Feature Engineering

In order to make good use of the data, we first removed some missing values, and then trimmed the available data. We simply kept the factors that are relevant for predicting used vehicle prices, such as model, year, condition, mileage, transmission, etc. Removed the release date, URL, VIN, etc. It seems that the car production and sales grew to large numbers during the 1960s, according to Wikipedia [5]. At the same time, in order to make the model more accurate, we cut the year, odometer and price in an interval. After cleaning the data, we digitally coded all the non-numeric factors, and then normalized them in order to eliminate the influence of different factors on the model due to their numerical size.

2. Data Visualization

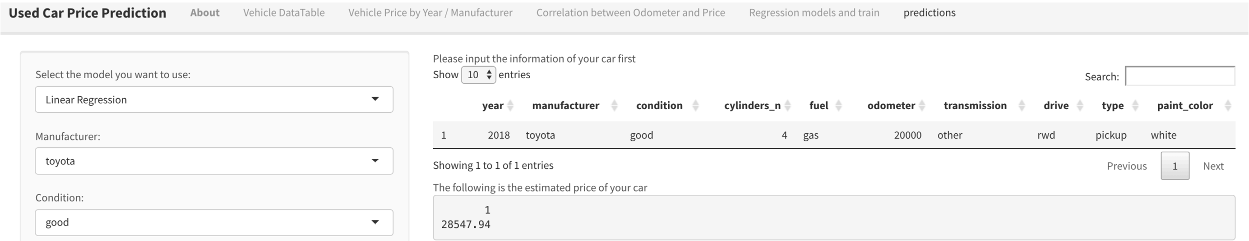
In EDA, apart from box plots, bar charts, missing matrix, we also visualized the geographic data with plotly and folium. A Web App url was created on https://www.shinyapps.io/. The web app was designed to present data interactively. Using R Studio, we created a data listing table, bar charts showing vehicle price by year/manufacturer, dot plots showing correlation between odometer and price with respect of fuel type.

3. ML modelling for Used Vehicle Price Prediction

This project uses three models to explore and predict used vehicle prices. We used 80% for the training set and 20% for the test set.

1) Linear regression model
   Linear regression is a regression analysis that models the relationship between one or more independent and dependent variables using a least-square function called a linear regression equation. When we predict the price of a used vehicle relative to other factors such as year, mileage, color, etc. We need to take the price of the used vehicle as the dependent variable, and all the other relevant parameters as the independent variable, fitting an equation about the price of the used vehicle and other factors.
2) Support vector machine
   Support vector machine is mainly used to do binary classification problems, but although people to support vector machine research, it can also solve more and more problems. We mainly use support vector regression to fit our curve. However, the computational complexity of support vector regression is much greater than that of linear regression. Therefore, if the data dimension is too large, using SVR can consume a lot of time.
3) Ridge regression model
   Ridge regression is a branch of linear regression. In the process of predicting used vehicle price, we find that each factor has a large difference in the impact on second-hand car price. This can easily lead to over-fitting or inaccurate results due to specific loss values as we train the training set.

Here are three screenshots of the prediction site to compare the prediction results：


Used Car Price Prediction
About
Vehicle DataTable
Vehicle Price by Year / Manufacturer
Correlation between Odometer and Price
Regression models and train
predictions
Search:
Previous
1
Next

Select the model you want to use: Linear Regression

Manufacturer: toyota

Condition: good

Please input the information of your car first

Show 10 entries

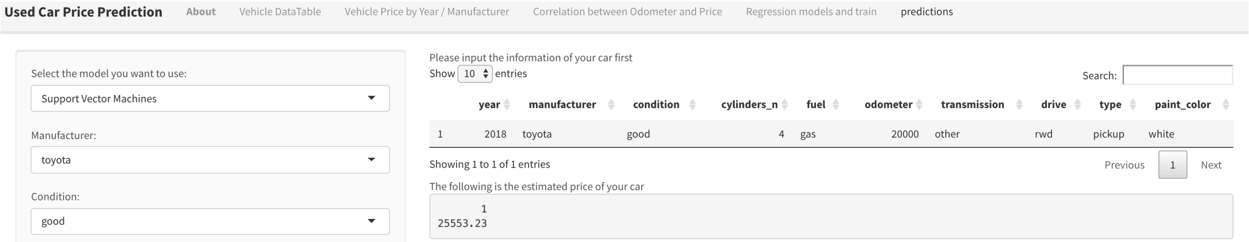
	year	manufacturer	condition	cylinders_n	fuel	odometer	transmission	drive	type	paint_color
1	2018	toyota	good	4	gas	20000	other	rwd	pickup	white

Showing 1 to 1 of 1 entries

The following is the estimated price of your car

```
       1
28547.94
```

Figure1：Results of linear regression for Toyota brand vehicles in 2018


Used Car Price Prediction
About
Vehicle DataTable
Vehicle Price by Year / Manufacturer
Correlation between Odometer and Price
Regression models and train
predictions
Search:
Previous
1
Next

Select the model you want to use: Support Vector Machines

Manufacturer: toyota

Condition: good

Please input the information of your car first

Show 10 entries

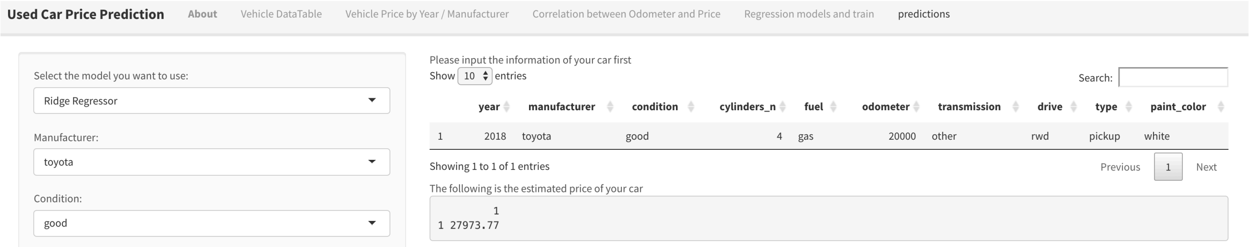
	year	manufacturer	condition	cylinders_n	fuel	odometer	transmission	drive	type	paint_color
1	2018	toyota	good	4	gas	20000	other	rwd	pickup	white

Showing 1 to 1 of 1 entries

The following is the estimated price of your car

```
       1
25553.23
```

Figure2：Results of Support vector regression for Toyota vehicles in 2018


Used Car Price Prediction
About
Vehicle DataTable
Vehicle Price by Year / Manufacturer
Correlation between Odometer and Price
Regression models and train
predictions

Select the model you want to use: Ridge Regressor

Manufacturer: toyota

Condition: good

Please input the information of your car first

Show 10 entries

Search:

	year	manufacturer	condition	cylinders_n	fuel	odometer	transmission	drive	type	paint_color
1	2018	toyota	good	4	gas	20000	other	rwd	pickup	white

Showing 1 to 1 of 1 entries

The following is the estimated price of your car

```
         1
1 27973.77
```

Figure3：Results of Ridge regression for Toyota brand vehicles in 2018

This is a 2018 Toyota (the rest of the details are in the screenshot). The linear regression model predicts $28547.94, the support vector regression model predicts $25553.23, and the Ridge regression model predicts $27973.33. It can be seen that their results are relatively close. It is quite normal for a car to fluctuate by $2000 ~ $3000 due to too many artificial factors. After market research, the result of our prediction is almost the same as the market price, so we have completed our scheduled task.

Challenges

One of the main challenges we faced is coding with Shiny in R. This is our first time using R to build web application. The DataCamp course [6] provided us the basic coding skills to build an interactive dashboard with R. And the example code showing in the R shiny gallery is useful for us to scaffold our shiny dashboard app. We have encountered barriers whiling debugging R code. To fix those issues, we searched solutions for the errors online.

References

1. Edmunds. “Used Vehicle Outlook, https://static.ed.edmunds-media.com/unversioned/img/industry-center/insights/2019-used-vehicle-outlook-report-final.pdf” Page 1. Accessed June 29, 2021.
2. Used Cars Price Prediction. https://www.kaggle.com/c/1056lab-used-cars-price-prediction. Accessed Dec 11, 2021.
3. Just What Factors into The Value of Your Used Car? https://www.investopedia.com/articles/investing/090314/just-what-factors-value-your-used-car.asp#citation-2. Accessed Dec 11, 2021.
4. Used Cars Dataset. (2021). Vehicles listings from Craigslist.org. Retrieved from https://www.kaggle.com/austinreese/craigslist-carstrucks-data.
5. Wikipedia page of Automotive industry in the United States. https://en.wikipedia.org/wiki/Automotive_industry. Accessed Dec 12, 2021.
6. DataCamp: Building Dashboards with shinydashboard. https://app.datacamp.com/learn/courses/building-dashboards-with-shinydashboard. Accessed Oct 26, 2021.
7. R Shiny Gallery. https://shiny.rstudio.com/gallery/. Accessed Dec 11, 2021.