Indicate And Identify The Rationality Of Used Car Prices

Chuheng (Kevin) Yu, Shih-han (Elle) Juan, Sifan (Ellen) Zhu
Weiming (Kevin) Wang, Ziyuan (Ryan) Li

BA780 Team A3

Agenda

- Background
- Project Goal

- Exploratory Analysis
- Manufacturer Analysis

- Regression Analysis
- Conclusion

Background

WHY MATTERS?

- High demand from customers
- Global used car market size -> **$2.75** trillion by 2030
- The average used car price in the US has increased by **30%**

Source: https://www.globenewswire.com/en/news-release/2022/09/22/2521333/0/en/Used-Car-Market-Size-is-projected-to-reach-USD-2-75-trillion-by-2030-growing-at-a-CAGR-of-6-17-Straits-Research.html

Primary Goal

Identify and evaluate the price of used cars based on their features in the U.S.

Data Source Overview

craigslist

- From *Kaggle.com*, scrapped from *craigslist.org*
- 04/04/2021 to 05/04/2021
- Area: within the U.S.
- 426,880 rows and 26 columns
- Data types:
  - **Descriptive:** *manufacturer, model, condition, cylinders, fuel, title_status, transmission, drive, size, type, paint_color, state*
  - **Numerical:** *price, year, odometer*
  - **Geographical:** *lat, long*
  - **Other strings:** *id, url, region, region_url, VIN, image_url, description, county, posting_date*

Data Cleaning

- drop_na for rows with **<= 10%** missing
  - *year, manufacturer, model, fuel, odometer, title_status, transmission, lat, long, type*
- Impute mode for columns with **<= 20%** missing
  - *drive, paint_color*
- Fill in 'unknown' for columns with **> 20%** missing
  - *condition, cylinders*
- $20 < price < $400,000
- Drop outliers using IQR
- Prevent down payments
  - Drop **<= 10%** average price by model year

Exploratory Analysis

Exploratory on the whole market and each variable

Geographical Distribution

- More used car listings on the east coast
- Size of the marker
  --> Number of listing
- The density of colour
  --> The average price

The whole market

- The average price kept increasing
- Used vehicles with fewer years
  --> higher price

- Most of the price is concentrated in 0-30,000

Amount and Color

- White cars are the most numerous used cars, followed by the black ones
- The number of black cars is only half of that of white cars

Price and Fuel Type

- Price of each fuel type is different
- Diesel cars --> highest average price
- Gas and hybrid cars have cheaper price

Price and Odometer

- Used cars with high price have small odometers
- Most used cars have low odometers
- Buyers hope the condition of cars is new

Price and Vehicle Type

- The most common vehicle types SUVs and sedans both have an average price around $15,000
- pickup truck --> highest price
- mini-van --> lowest price

Price and Engine

- 4wd, rwd vehicles
  --> balanced distribution
- fwd --> more skewed to the right.
- Customer looking for a lower price
  --> front-wheel drive cars

Price and Title Status

- clean title status and old model year --> costly
- Years around late 1990 is the base of the curve, and the price starts to climb as time moves forwards.
- Customers prefer a clean title status

Deeper into Manufacturers

Exploration on the relationship between price, manufacturers and a third variable

Price, Manufacturer and Fuel Type

Diesel

Gas

Electric

Price, Manufacturer and Fuel Type

Hybrid

Other

- Most expensive cars have a "Green Energy" (electric/hybrid) fuel type
- Majority of manufacturers produce gas fuel vehicles
- Hybrid and electric average prices between manufacturers have a high variance

Price, Manufacturer and Drive Type

- Nissan has rwd with highest average price
- Nearly 60% manufacturers have 4wd as their average most expensive drives; rwd and fwd account for 35% and 5%

Price, manufacturer and vehicle type

- Sedan is more recommended with both lower price and odometer
- Sedans with Saturn is considered the "newest" one with a low price

Preprocessing

Previously Prepared Data Set

⬇

Drop 'Posting_Date'

⬇

Sampling

⬇

OneHotEncoder

```
df['posting_date'].max(),df['posting_date'].min()
```

```
('2021-05-04T23:59:45-0400', '2021-04-04T00:10:40-0700')
```

	region_SF bay area	region_abilene	region_akron / canton	region_albany	region_albuquerque	region_altoona-johnstown
0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0

Regression Analysis

	MSE	Sqrt MSE
Linear Regression	1101598622349.81	1049570.68
Support Vector Regression	165773315.66	12875.30
Polynomial Regression	131425814.14	11464.11
Random Forest Regression	120513702.26	10977.87

Feature Importance

Conclusion & Future Studies:

- Recommendation for a cheaper used car:
  - Brand: Saturn
  - Year: 2002
  - Odometer: 100,0000 Miles +
  - Color: Yellow
  - Drive: fwd drive type
  - Transmission: Manual
  - Title: Non-clean title status

- Future Studies
  - Data Set Balancing
  - Model Tuning

Thank you!
Any Questions?

Reference:

- https://instamotor.com/blog/first-car-study
- https://www.globenewswire.com/en/news-release/2022/09/22/2521333/0/en/Used-Car-Market-Size-is-projected-to-reach-USD-2-75-trillion-by-2030-growing-at-a-CAGR-of-6-17-Straits-Research.html
- https://www.cargurus.com/Cars/price-trends/
- https://www.progressive.com/answers/car-down-payment/#:~:text=One%20rule%20of%20thumb%20for,and%20how%20car%20loans%20work.
- https://www.checkmarket.com/sample-size-calculator/