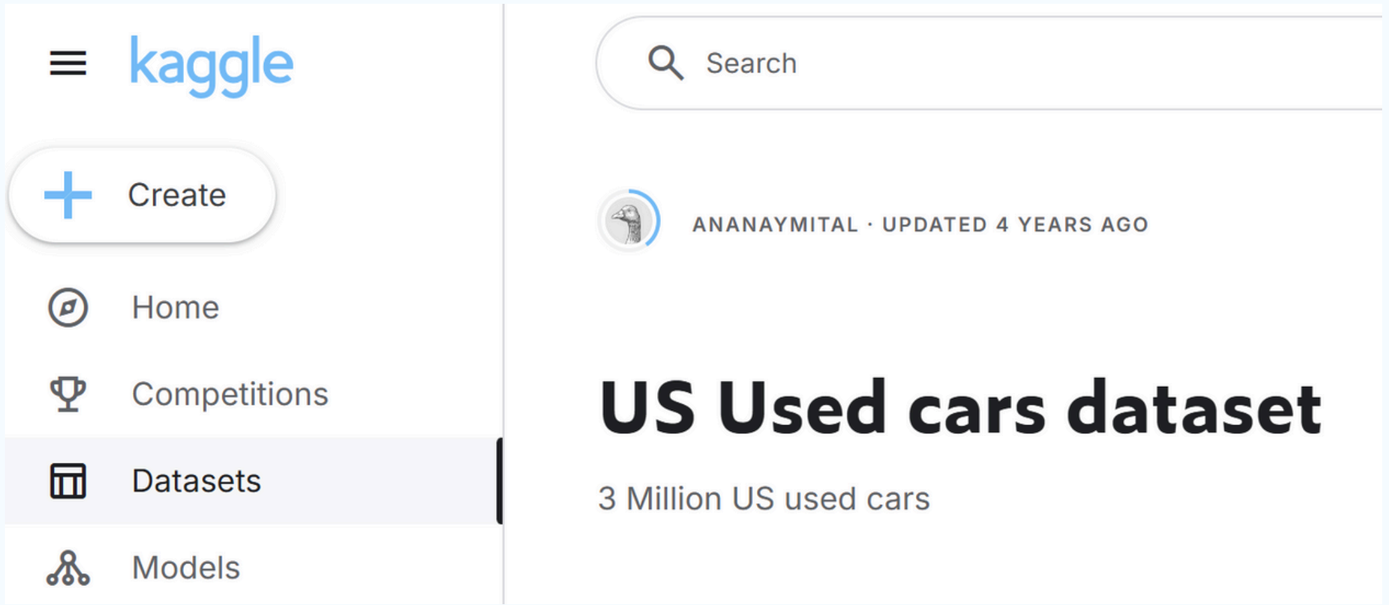
QI - Week2
Group 2 Presentation

Ji-Soo Kim, Tae-Yoon Kim, Jun-Beom Lee, Jin-Joo Yang, Sung-Hyun Kim

Table of Contents

0. Recap

Price Prediction of Private Sales Vehicles with Used Car Datas



1. Week 1 - FeedBacks

Q. Consider external Economic factors?

$$\text{Current Value} = \text{Past Price} \times \frac{\text{Current CPI}}{\text{Past CPI}}$$

$$\text{Normalized Price} = \frac{\text{Price}}{\text{CPI}}$$

- Consumer Price Index (CPI)?
  - Indicator that measures the average price level of “shopping baskets” including the items and services that consumers buy
  - How much prices have risen or fallen over time

TABLE II. MACROECONOMIC VARIABLES

variable	description
new/used car price index	The price index for each month [20].
new/used car registration & sales volume	Number of cars are registered and sold in each month [21].
2-year fixed interest rate/ floating rate, Inflation, Petrol/Diesel price, CPI, GDP, NZD/Yen index, Unemployment rate	Macroeconomic factors monthly data [22].

But how can we convert past prices to present values…?

[Huayi Jing]

2. Our Directionality

Helping owners to predict their used car sale price “without the dealer’s opinion”

Our Goal?

1. Fair Trade Tools for used car owners

2. Prevent dealer fraud

3. Build user-friendly systems (future works)

Major strategy?

1. Analyze various characteristics to increase the reliability of predictions

2. Select user-centric data
ex) year, odometer, condition

3. Consider macroeconomic factors affecting car prices?
ex) inflation, CPI, economic state of the target

3. Preceding Research Analysis 2

Proceedings of the 5th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/59/20251746

Predicting Vehicle Prices Using Machine Learning: A Case Study with Linear Regression

Jiahao He[1,a,*]

[1]*University of Washington, Seattle, 98105, United State*
a. hejh0612@gmail.com
**corresponding author*

Abstract: With the development of electronic vehicles, accurately predicting the price of vehicles is essential for both consumers and business. Thus, this study aims to explore the application of machine learning in vehicle price prediction, specifically focusing on the use of linear regression, a widely adopted technique in this domain. Utilizing a comprehensive dataset containing variables such as make, model, year, and mileage, the research develops a predictive model through rigorous data cleaning, feature engineering, and model tuning processes. The model's performance will be analyzed by R-squared value and Mean Squared Error(MSE). The model's predicting result will be visualized by scatter plot. The study also addresses potential biases and applies regularization techniques to enhance the model's accuracy. Additionally, a comparative analysis with a Decision Tree model evaluates the relative performance and nuances of each approach. This research not only underscores the practical value of predictive modeling in the automotive industry but also offers insights into integrating more advanced machine learning techniques to further improve prediction accuracy. Additionally, in this research, the drawbacks of linear regression model will be indicated and the future works to improve the performance of the model will be provided.

Keywords: Machine learning, Linear regression, Vehicle pricing, Data science, Python.

3. Preceding Research Analysis 2

Main concept in data preprocessing

2.1.4. Convert categorical date to numerical data

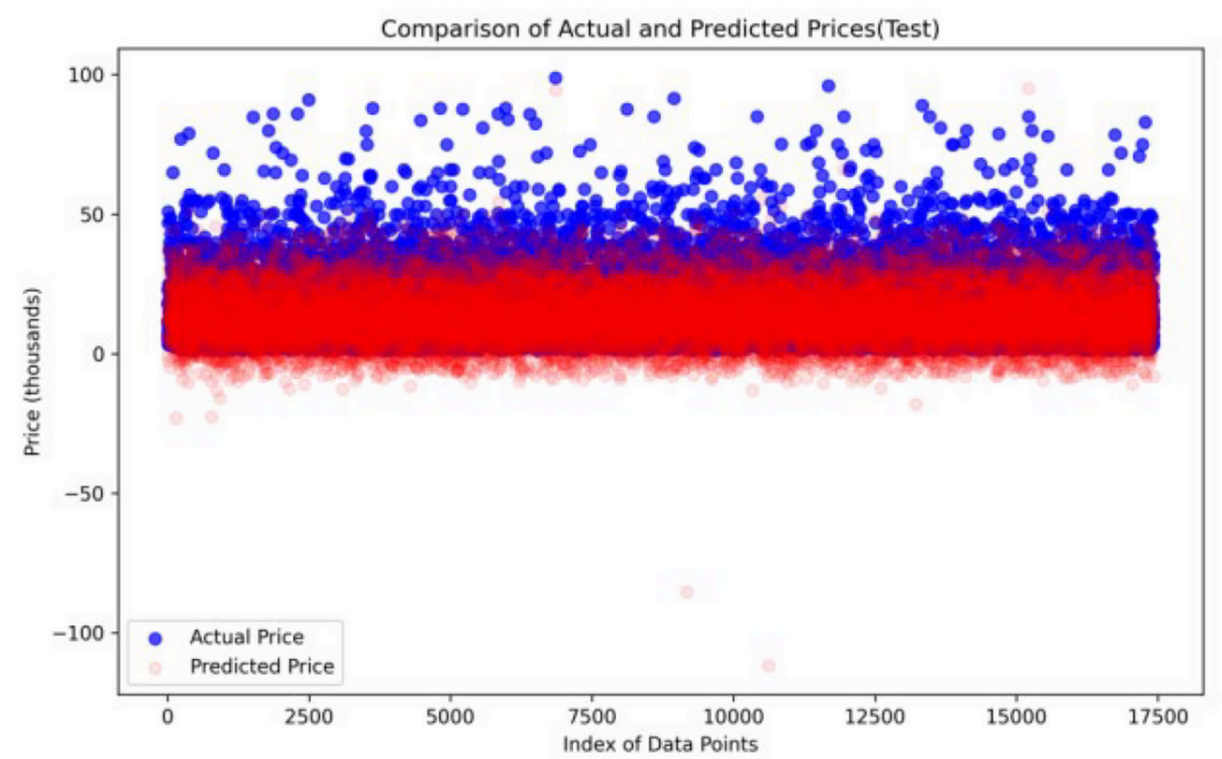
This paper used two strategies for converting, one is label-encoding[4], and another is mixed strategy.

For mixed strategy, for ordinal data, where categories have a natural order, such as 'condition' (from 'salvage' to 'new'), this paper will apply manual mapping. This involves assigning each category a unique integer based on its relative ranking. For nominal categorical variables, such as 'color', this paper uses one-hot encoding. This method transforms each category into a new binary column, ensuring that the model treats each category distinctly without imposing any ordinal relationship[5].

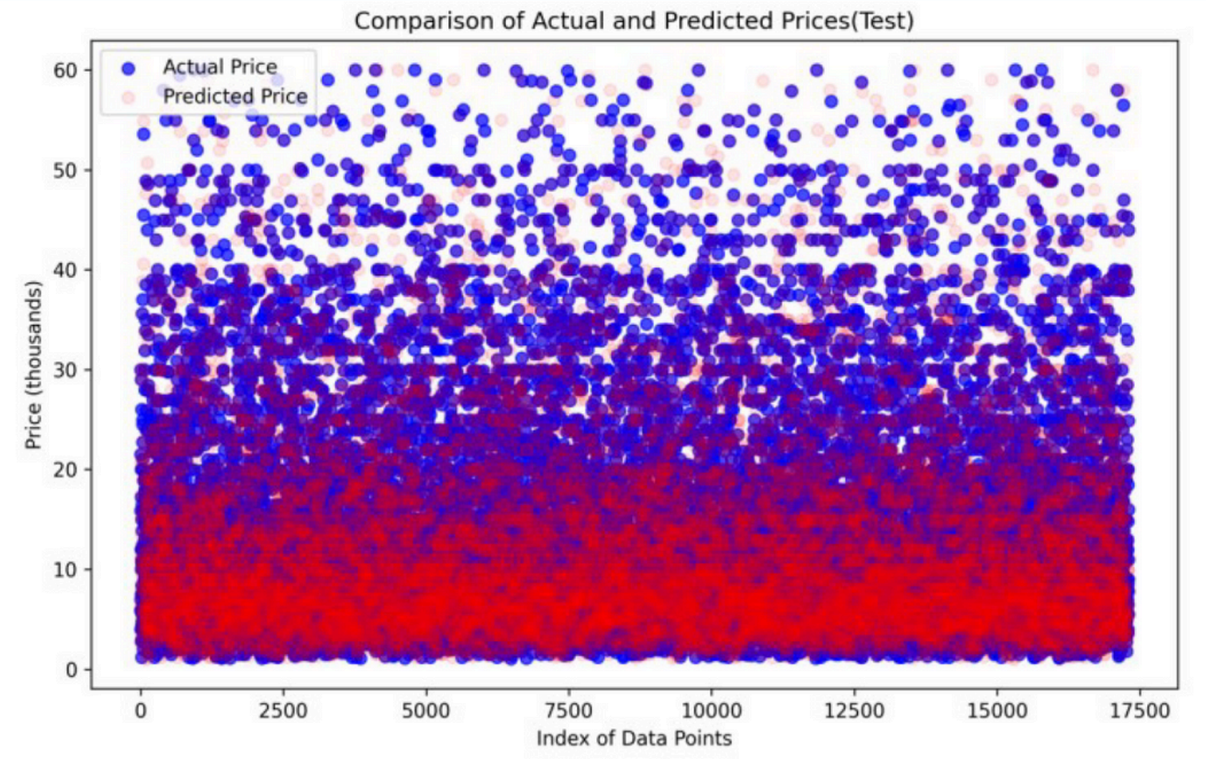
Mixed stategy - using one-hot encoding & manual mapping(label-encoding)

Label stategy - using only label-encoding

3. Preceding Research Analysis 2



Distribution using linear regression



Distribution using decision tree

Using the mixed strategy -> both models achieved high accuracy at low and medium prices, but the accuracy declined when the price was high.

3. Preceding Research Analysis 2

Table 1: Performance Matrix of Linear Regression.

	Mixed Strategy	Label Strategy
Training MSE	43.73	54.26
Training R-squared	0.69	0.62
Test MSE	44.02	54.23
Test R-squared	0.70	0.63

Table 2: Performance Matrix of Decision Tree.

	Mixed Strategy	Label Strategy
Training MSE	0.02	0.02
Training R-squared	1.00	1.00
Test MSE	23.92	26.40
Test R-squared	0.84	0.82

Overfitting

-> recommend using the linear regression method at low and medium prices and consider other methods when the price is high.

4. Feature Selection - Not Selected

Null

- bed
- bed_height
- bed_length
- cabin
- combine_fuel_economy
- is_certified
- is_cpo
- is_oemcpo
- owner_count
- vehicle_damage_category

Carmax

- back_legroom
- engine_type
- transmission_display

ID&URL

- vin
- dealer_zip
- listing_id
- sp_id
- trimId
- main_picture_url
- sp_name

Duplication

- make_name
- exterior_color
- trim_name
- horsepower
- savings_amount
- wheel_system_display

Biased

- salvage
- theft_title

Paper

- description
- latitude
- longtitude
- city

Confirmed

- is_new
- maximum_seating
- franchise_dealer

Survey

- seller_rate

4. Feature Selection - Selected

Vehicle Performance

- city_fuel_economy
- highway_fuel_economy
- power_HP
- power_rpm
- torque_torque
- torque_rpm

Vehicle size

- height
- length
- width
- front_legroom
- wheelbase

Vehicle Specs

- fuel_type
- body_type
- engine_cylinders
- fuel_type
- franchise_make
- engine_displacement
- model_name
- wheel_system
- transmission
- major_options
- mileage

Color

- interior_color
- listing_color

Date information

- year
- daysonmarket
- listed_date

Past uses

- fleet
- isCab

Accident status and condition

- frame_damaged
- has_accidents

5. Data Preprocessing

```
[28]: print(df[['fuel_tank_volume','front_legroom','height']])
```

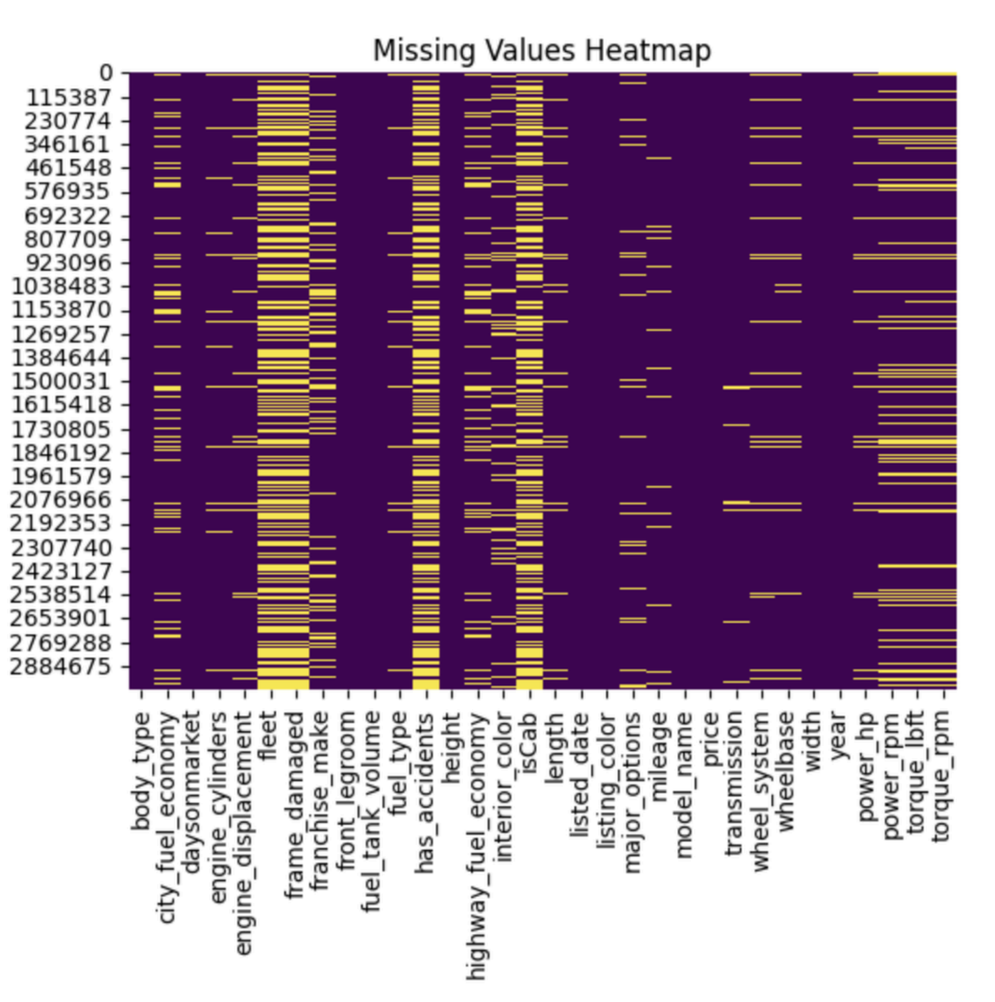
```
         fuel_tank_volume front_legroom height
0                    12.7          41.2   66.5
1                    17.7          39.1     68
2                    15.9          43.3   58.1
3                    23.5            39     73
4                    17.7          39.1     68
...                   ...           ...    ...
3000035              14.9          40.9   65.4
3000036              19.4            41   70.7
3000037              16.5          44.3   58.2
3000038              14.8          41.5   55.7
3000039              14.5            43   68.1

[3000040 rows x 3 columns]
```

- changed values into string type, then split by space and extract the first value
- ex) “40 inches” -> “40"

power	price	torque
177 hp @ 5,750 RPM	23141.0	200 lb-ft @ 1,750 RPM
246 hp @ 5,500 RPM	46500.0	269 lb-ft @ 1,400 RPM
305 hp @ 6,000 RPM	46995.0	290 lb-ft @ 4,000 RPM
340 hp @ 6,500 RPM	67430.0	332 lb-ft @ 3,500 RPM
246 hp @ 5,500 RPM	48880.0	269 lb-ft @ 1,400 RPM

- Splitted feature ‘power’, ‘torque’ into ‘power_hp’, ‘power_rpm’ / ‘torque_lbft’, ‘torque_rpm’ and made new feature

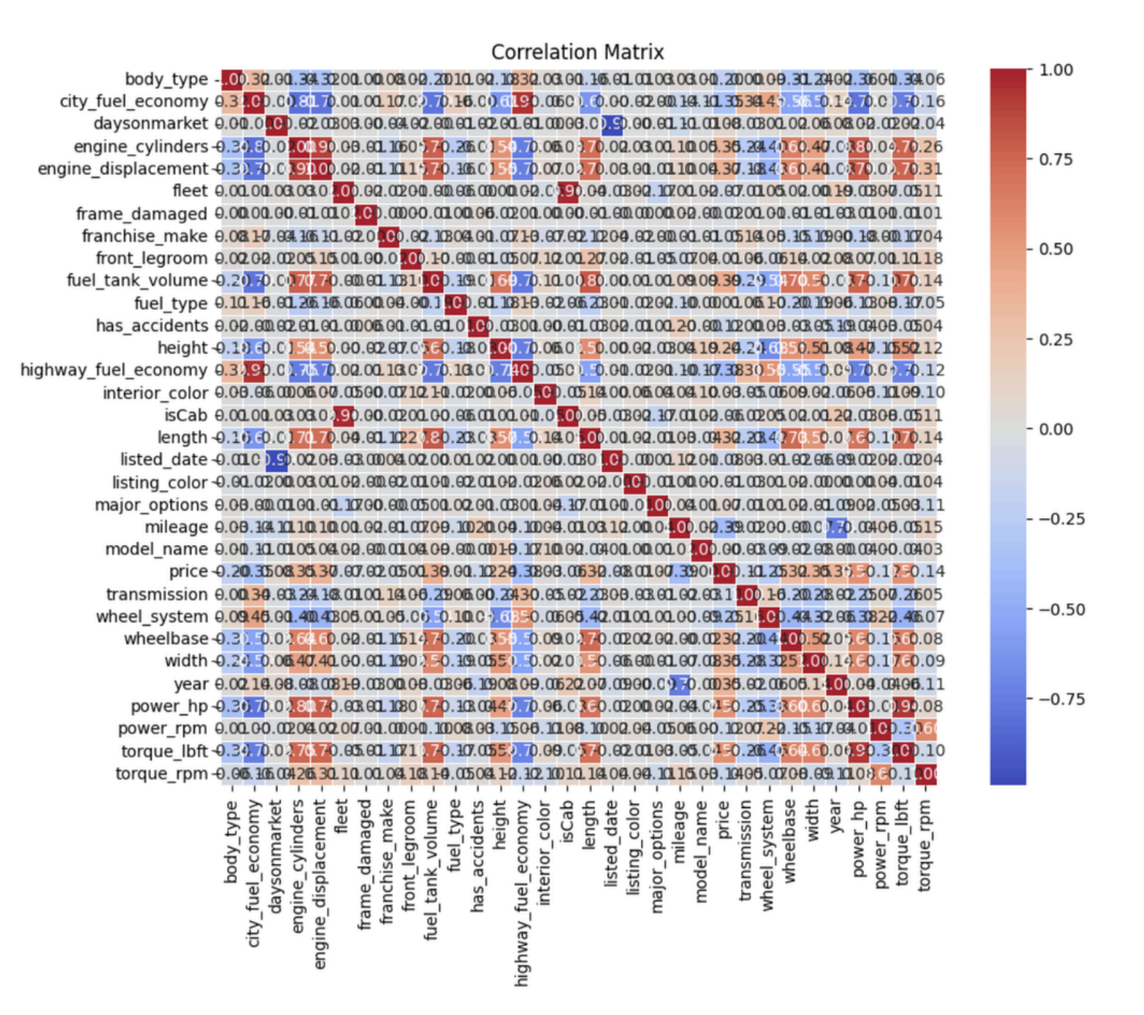
6. Future Work



body_type	13543
city_fuel_economy	491285
daysonmarket	0
engine_cylinders	100581
engine_displacement	172386
fleet	1426595
frame_damaged	1426595
franchise_make	572635
front_legroom	0
fuel_tank_volume	0
fuel_type	82724
has_accidents	1426595
height	0
highway_fuel_economy	491285
interior_color	383986
isCab	1426595
length	159269
listed_date	0
listing_color	0
major_options	200048
mileage	144387
model_name	0
price	0
transmission	64185
wheel_system	146732
wheelbase	159269
width	0
year	0
power_hp	172386
power_rpm	481427
torque_lbft	517793
torque_rpm	517794

dtype: int64

6. Future Work

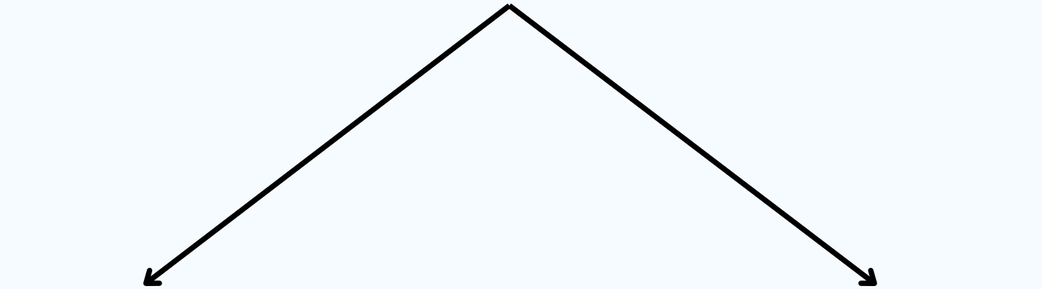


Specify feature selection by considering the correlation of features with correlation heatmap

6. Future Work

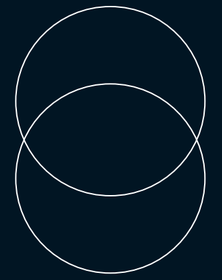
Handling date information

[listed_date] + [daysonmarket]
-> the date when the used car was sold



Sold used cars
-> **1 (True)**

Unsold used cars
-> **0 (False)**



Q & A

Thank you for listening!

Appendix

A. *Feature Engineering*

The dataset in its raw state contains some features that are not necessarily relevant in predicting the price of a vehicle. Such features include but are not limited to posting date, image URL, listing URL, latitude, longitude, and VIN. The features that were selected and used to predict used car prices include year, manufacturer, condition, number of cylinders, title status, drive type, body type, and mileage. The selection also considered leveraging data biases.

[Jewel Donkor Apeko, 2023]

Additionally, we perform the following preprocessing steps on the data set helping us narrowing down the features:

(1) Keep only listings for cars sold by private owners and filter out those sold by dealerships
(2) Keep only listings for cars being sold, and filter out all request for purchase listings
(3) Filter out cars manufactured before 1863 and after 2017, and derive the car's age
(4) Filter out all cars with unrealistic Power values
(5) Filter out listings which don't have an associated price
(6) Filter out all cars listed as unavailable
(7) Filter out invalid registration dates
(8) Convert boolean (true/false) fields to numeric (0/1) based
(9) Filter out all data with value as 'NA' (Not Available)

H. *Benchmarking Real-Life Car Price Estimation Tools*

This research benchmarks the car price estimation tools used by Kelley Blue Book (KBB) and Edmunds on their respective websites [12][14]. They use a proprietary editorial process, which is a holistic view of market trends. It is notable to point out that both models produce results in the form of a range.

1) *Kelley Blue Book Price Estimation Tool*

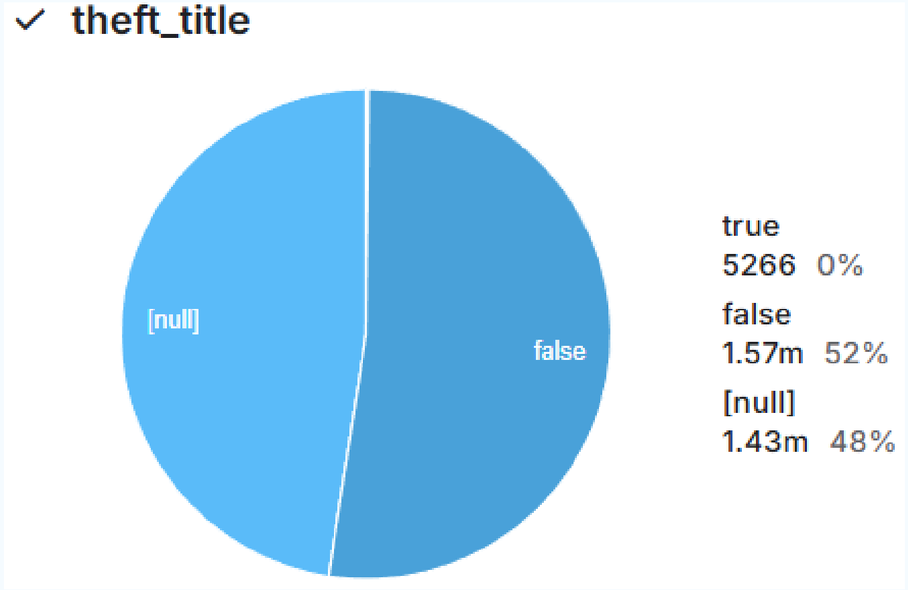
Table II compares benchmark values from the Kelly Blue Book website with the proposed Bayesian network model. Some details about the vehicles have been omitted for brevity.

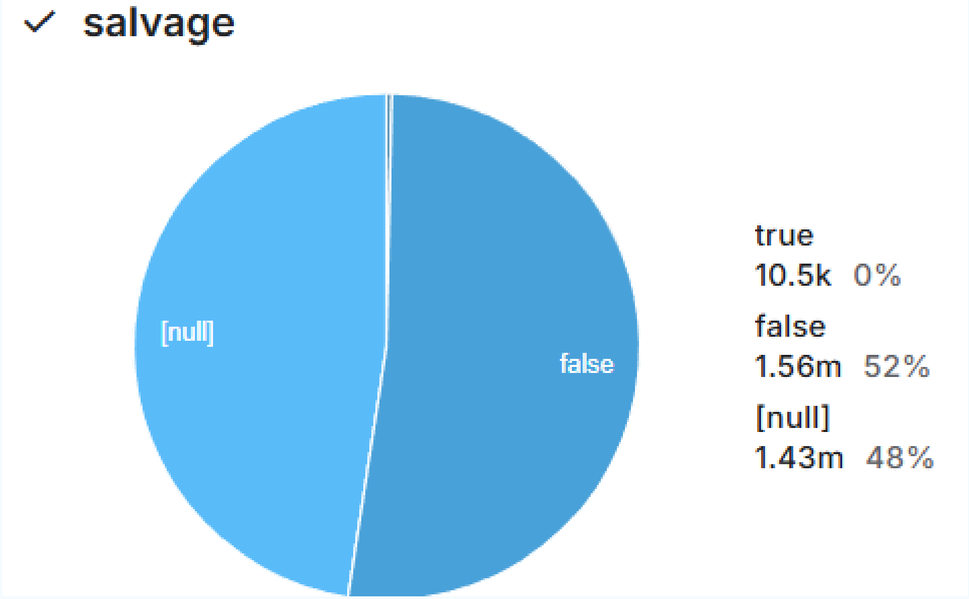
TABLE V. KBB PRIVATE PARTY AND BN MODEL ESTIMATE RANGES

No.	Vehicle Details	BN Range	KBB Range
	Luxury Vehicles		
1	2013 BMW 3 Series, Clean Title, Sedan, 4-cylinder engine, Rear wheel drive	12001 – 15000	11106 - 14471
2	2013 Cadillac CTS, Clean Title, Sedan, 4-Cylinder engine, Rear wheel drive	9001 – 12000	9543 – 12136
	Non-Luxury Vehicles		
3	2013 Nissan Sentra, Clean Title, Sedan, 4-cylinder engine, front-wheel drive	3001 – 6000	3672 – 5443
4	2010 Toyota 4Runner, Clean Title, SUV, 6-cylinder engine, Rear wheel drive	18001 – 21000	16812 – 19663

[Jewel Donkor Apeko, 2023]

Appendix


theft_title
[null]
false
true
5266 0%
false
1.57m 52%
[null]
1.43m 48%


salvage
[null]
false
true
10.5k 0%
false
1.56m 52%
[null]
1.43m 48%