Capstone Project Report

Used Car Price Prediction

By Parul Rana

Background

Vehicle shortage has caused prices for both new and used cars to hit record highs. And while inventory has improved in recent months, it is nowhere near pre-pandemic levels.
Car shoppers today are likely to face price hikes from either dealer-added (often non-negotiable) accessories or "market adjustments." Discounts are around but vary by brand. As of May 2023, brands offering the greatest discounts were Alfa Romeo, Volvo, Ram, Infiniti, Buick, Audi, GMC and Mercedes-Benz. On the other hand, popular brands such as Kia, Honda, Toyota, Dodge and luxury brands such as Land Rover and Cadillac had average transaction prices.

Having good knowledge about increasing ad decreasing price of used help us to find out which car is best to purchase and what are the features we should keep in mind while buying a new or used car. In this project, we will predict the used car prices by using machine learning algorithms. Data s from kaggle and the like is below:

https://www.kaggle.com/datasets/taeefnajib/used-car-price-prediction-dataset/data

Problem Statement
Lack of knowledge regarding the features put customers in loss either someone purchasing or selling a used car. Knowing features well help customers to make informed decision and prevent them making good negotiation during car purchase. We will find out, what is the price of a used car? What are the features one should look into before buying a car based on given features and factors such as brand or company, manufacturing year, mileage, Accident, Engine Type etc. in order to buy more used cars.

Dataset

Used Car Price Prediction Dataset is a comprehensive collection of automotive information extracted from the popular automotive marketplace website, https://www.cars.com. This dataset comprises 4,009 data points, each representing a unique vehicle listing, and includes nine distinct features providing valuable insights into the world of automobiles.

This dataset is a valuable resource for automotive enthusiasts, buyers, and researchers interested in analysing trends, making informed purchasing decisions or conducting studies related to the automotive industry and consumer preferences. Whether you are a data analyst, car buyer, or researcher, this dataset offers a wealth of information to explore and analyse.

Dataset Description:

- Brand & Model: Identify the brand or company name along with the specific model of each vehicle.
- Model Year: Discover the manufacturing year of the vehicles, crucial for assessing depreciation and technology advancements.
- Mileage: Obtain the mileage of each vehicle, a key indicator of wear and tear and potential maintenance requirements.
- Fuel Type: Learn about the type of fuel the vehicles run on, whether it's gasoline, diesel, electric, or hybrid.
- Engine Type: Understand the engine specifications, shedding light on performance and efficiency.
- Transmission: Determine the transmission type, whether automatic, manual, or another variant.
- Exterior & Interior Colors: Explore the aesthetic aspects of the vehicles, including exterior and interior color options.
- Accident History: Discover whether a vehicle has a prior history of accidents or damage, crucial for informed decision-making.
- Clean Title: Evaluate the availability of a clean title, which can impact the vehicle's resale value and legal status.
- Price: Access the listed prices for each vehicle, aiding in price comparison and budgeting.

Data Wrangling

Data is directly loaded taken from ‘kaggle’. We imported the necessary packages to run the large dataset with efficient memory called pandas package in a Dataframe.
Firstly, we see the statistical summary from that we get the idea of outliers via looking into how much data is disperse, outliers present in model year, mileages, price columns. We saw if the data types match to the data we have and looks like we have all data types in objective form except model year. We have changed the data types for ‘mileages’ and ‘prices’ by removing mi. and $ sign from the price. We see the ‘nunique’ value, which gave us numbers of distinct observations.

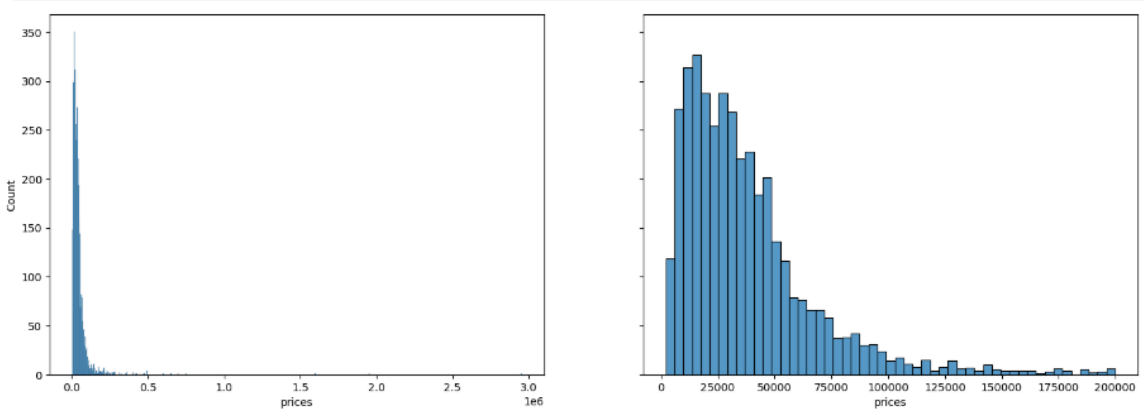
Next we did the value count which is use to count of all unique values. We also see if we have any null val ues in our dataset with percentage in (accident, clean title. Values with higher percentage have removed fro m dataset. We had (4009-rows × 12-columns).

```
#To get a concise summary of the dataframe.
car_cap.info()
```

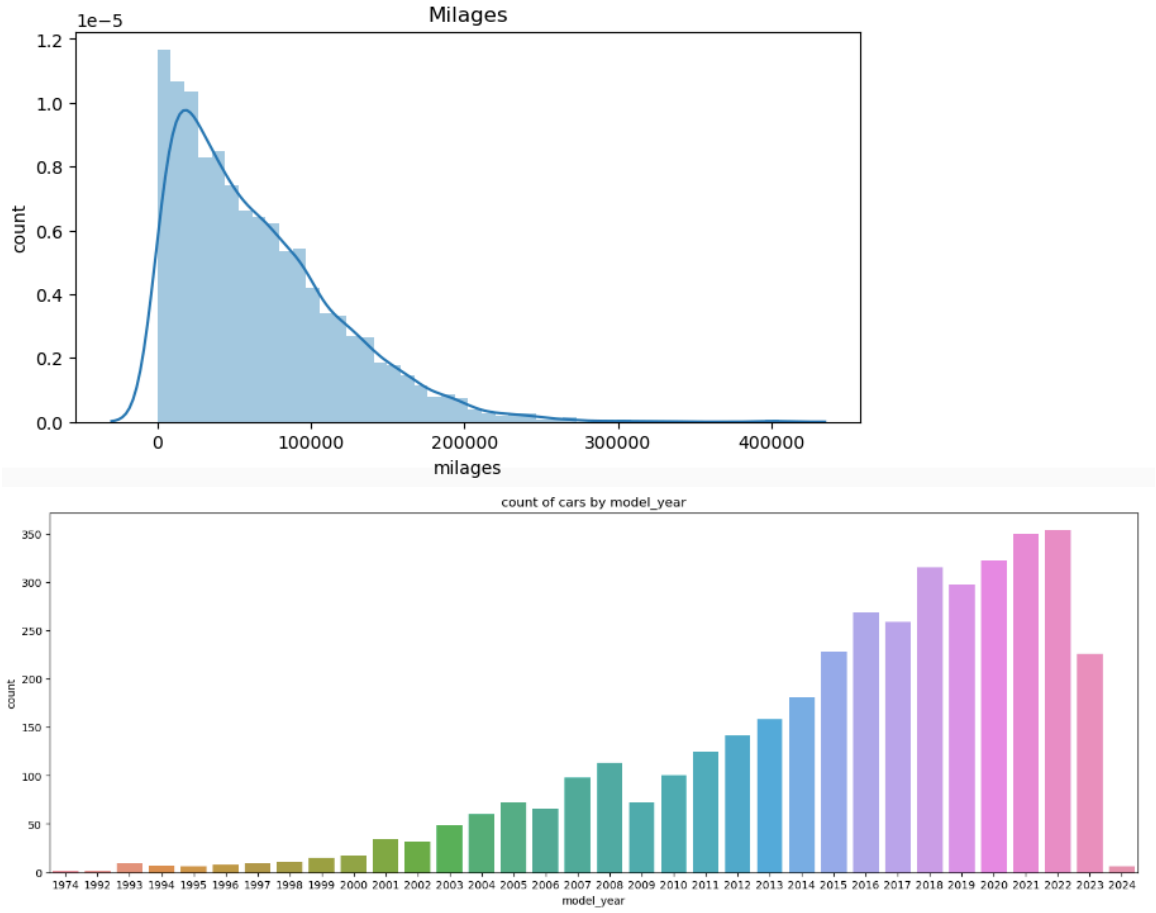
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4009 entries, 0 to 4008
Data columns (total 12 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   brand         4009 non-null   object
 1   model         4009 non-null   object
 2   model_year    4009 non-null   int64 
 3   milage        4009 non-null   object
 4   fuel_type     3839 non-null   object
 5   engine        4009 non-null   object
 6   transmission  4009 non-null   object
 7   ext_col       4009 non-null   object
 8   int_col       4009 non-null   object
 9   accident      3896 non-null   object
 10  clean_title   3413 non-null   object
 11  price         4009 non-null   object
dtypes: int64(1), object(11)
memory usage: 376.0+ KB
```

Exploratory Analysis

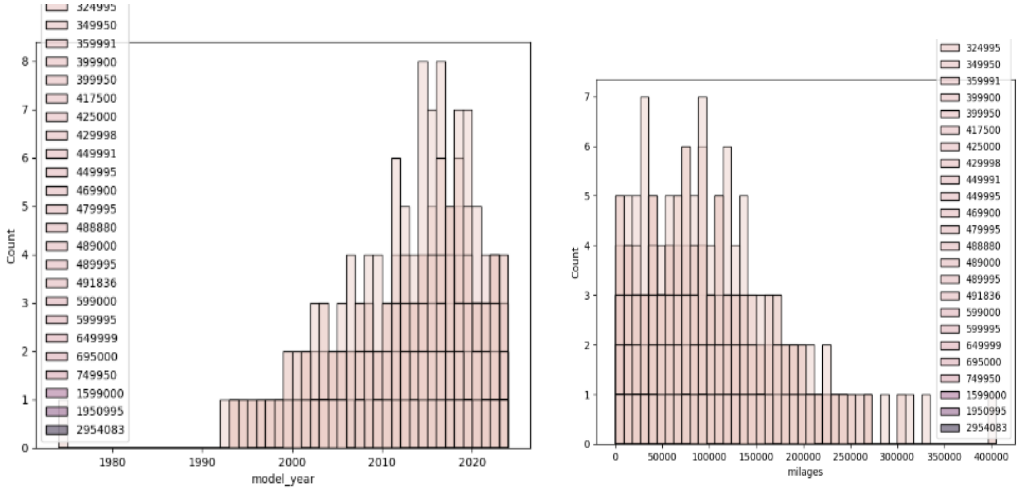
After wrangling, we will get some insight from the data and see how much outliers present what is relationship between each variable, which variable affecting prices. After getting insight from my data we explore it more in a graphical way. Explored data and their relationships using statistical analysis by selecting an important/interesting features.
My target is Price so I more keen to look relationship with other columns.



This graph shows that we have car value started from 2000 to 2954083. Next, I want to see what are the important feature which could possibly make a car value up and low.

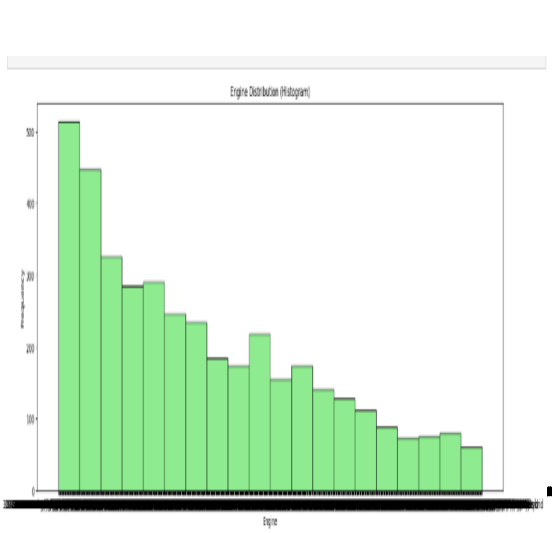


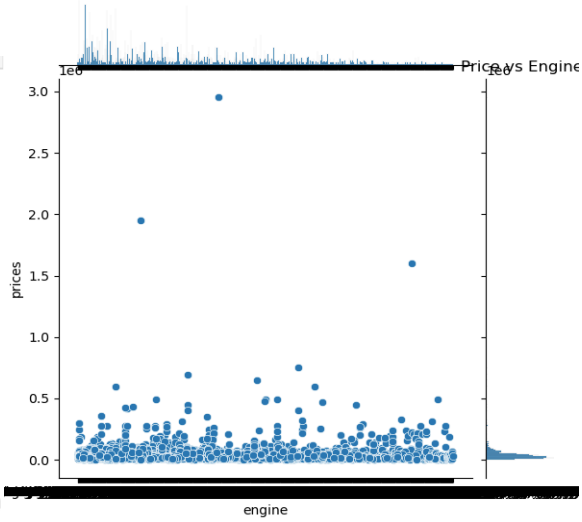
These two features are highly correlate to price. If milage incraeses price will increase, same for the model year too.New and updated model have high price
We will see the how both mileage and model year increase or decrease with price.



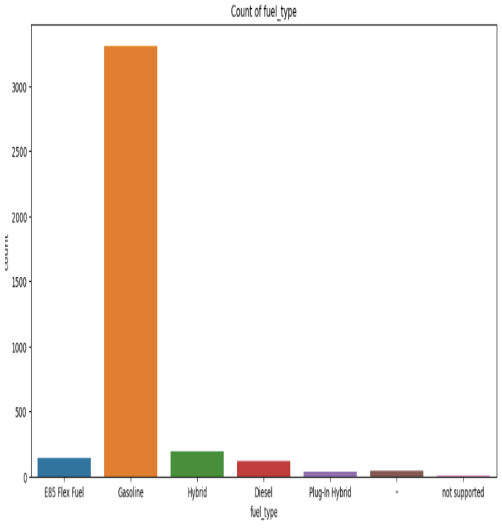
We have graphs, which shows the highest car mileage is 32000 mi. On the other hand Most of the cars manufactured in 2022. Last Spike of car manufacture was in 2008 and gradually increased 2015 to 2022. Cars with the model year 2022 have the highest count and lowest in 1974 and dropped after 2023 especially in 2024.

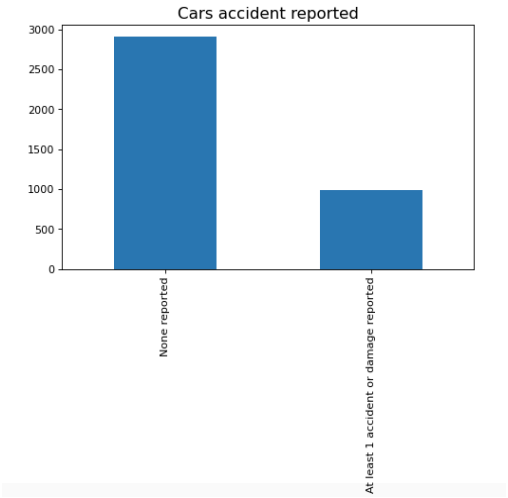
We have only two things which is very important features but there are other features too which can be important features to by a car or sell a car.



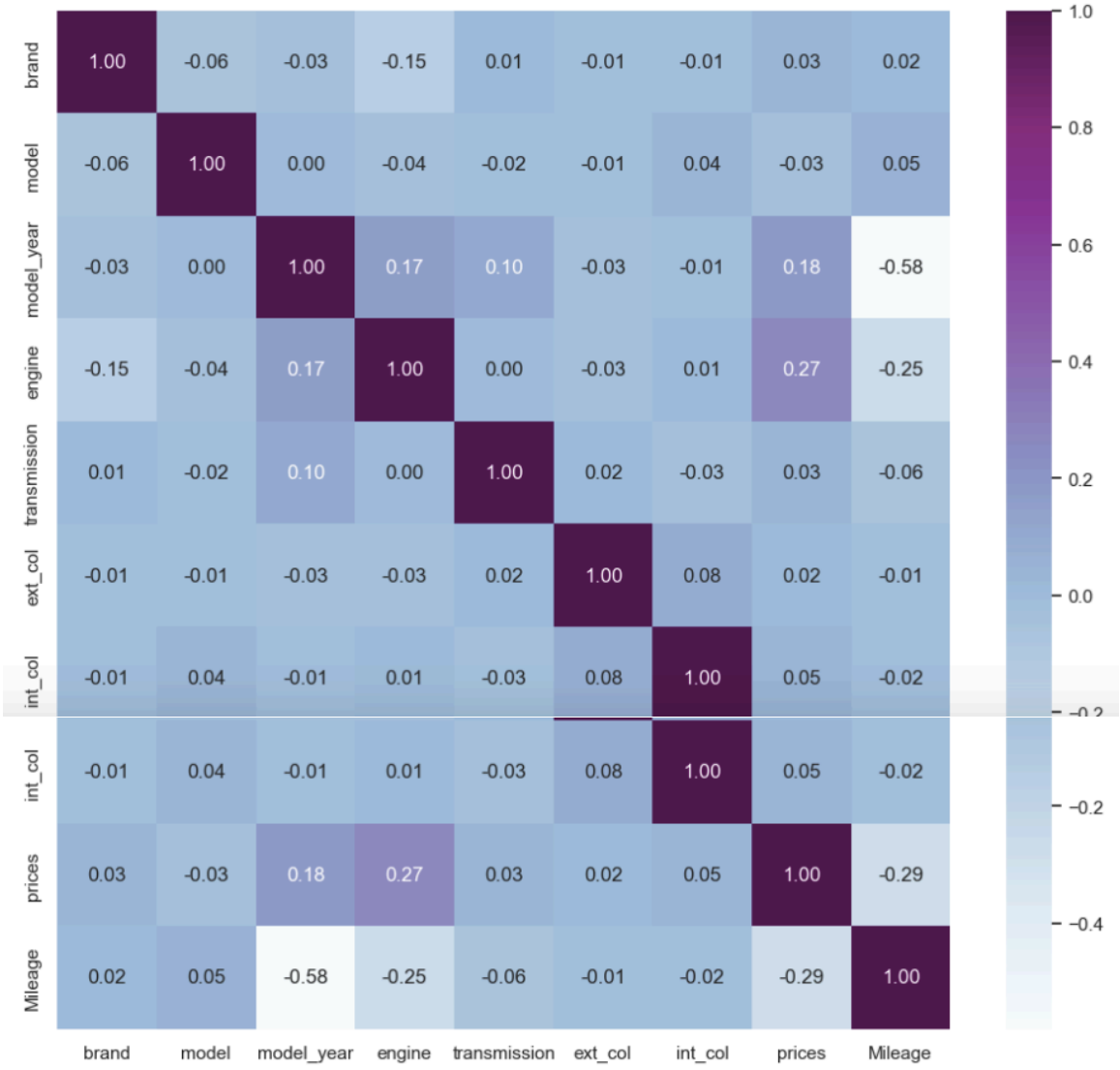


Engines are distributed equally that means they have all kind of cars on demand from good to bad to average. Car engine is also reflecting price of the car, good engine has high price value. Cylinder Engine Gasoline Fuel car is more popular in terms of engine.





In type of fuel, gasoline is preferable than other fuel type. In these two features we have some missing data as we can see in the accident column we have most of the cases that is not reported as accident and only few accident reported. These columns might not be necessary for approaching my target and I can further drop it. We have some other correlation matrix found in our data set, which could be necessary for machine learning model.

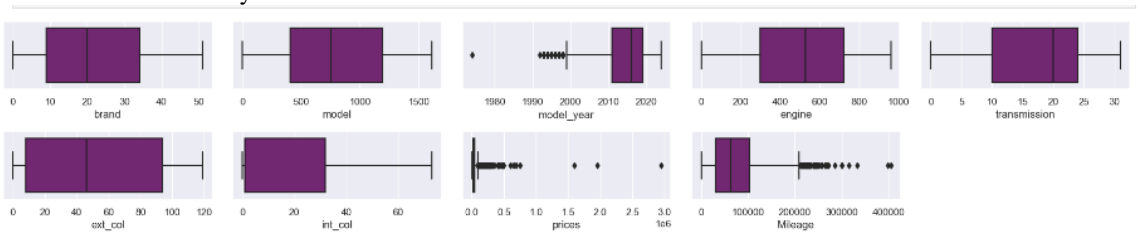


We have explored our data and now we have good idea what we need to keep and what we need to drop further to predict our target.

Machine Learning

For this, we did some pre-processing to build a good model. After data analysis I get to know that these columns mention here 'Accident, 'Clear title', 'fuel type' has a lot of missing values soI dropped them from my data. To correction in outliers in 'mileage','prices','model year' columns, I used IQR method to

remove outliers from my data.



Except [mileage',’prices’,’model year’] these columns, our all other columns are categorical columns and need to turn them into the numerical values. Label Encoder does easy job to turn categorical to numeric in a way that machine-learning algorithms can take values easily. It is a simple and efficient way to convert categorical data into numerical data that can be used for analysis and modelling.

- After feature engineering I split the data into train and test, I used multiple classifier to get best model. I used regression algorithms. I also did the cross validation for every algorithm. I applied cross-validation with a split in 5 fold to avoid overfitting. After cross validations, the models are performing well and providing the best accuracy score. After all the tests, cross validations and tunings, the XG boost is performing well with the accuracy score is 86.5% with a cross validation mean score of 81% for 5 cross validations.

After selecting model, I have done hyperparameter tunning; hyperparameter optimization is the process of finding the right combination of hyperparameter values to achieve maximum performance on the data.

Predictions

After cross –validation with 5-fold split is used to avoid overfitting. Hyper parameter tunning with GridSearchCV help the process plays a vital role in the prediction accuracy of a machine-learning algorithm.

XG boost is performing well with the accuracy score is 86.5% with a cross validation mean score of 81% for 5 cross validations.

Future Improvements

- For future improvements I would like to set different 'criterion', 'max_depth','max_features', 'n_estimators'.

Conclusion

In this report, we have looked at the prices of used car, what are the features will be greatly important to it. Knowledge of those features makes one to take informed decision without any second thought. Using all the important features and machine learning algorithms, we are able to predict the price of used cars. After Cross validation for XG boost algorithm provide us best accuracy of 86%.