UMBC

Prediction of Car Prices – Regression

Under the Guidance :

Dr. Chaojie Wang

Presented By :

Abdul Junaid Mohammed

Introduction/Problem

- Buying a used car can be a difficult task, especially when you are new to the country and don’t know the market value of the cars.
- To address this issue, I have undertaken a project to predict the prices of used cars using various data sources and regression algorithms.
- The project aims to help users get an estimate of the price of a car by providing basic details such as year, model, company, miles driven, etc.
- Through my analysis, I aim to identify useful insights and patterns that can help users make informed decisions when purchasing a used car.

Dataset

- The dataset is a collection of data scraped from the well-known website Craigslist. Dataset has been acquired from Kaggle.
- It has 426880 Rows and 26 columns. It has all the useful features which will be required for the purchase of used car. Here the Target Variable will be the Price Column.
- Of course not all the columns are going to be useful, but I found a good amount we can use.

Attributes/Characteristics

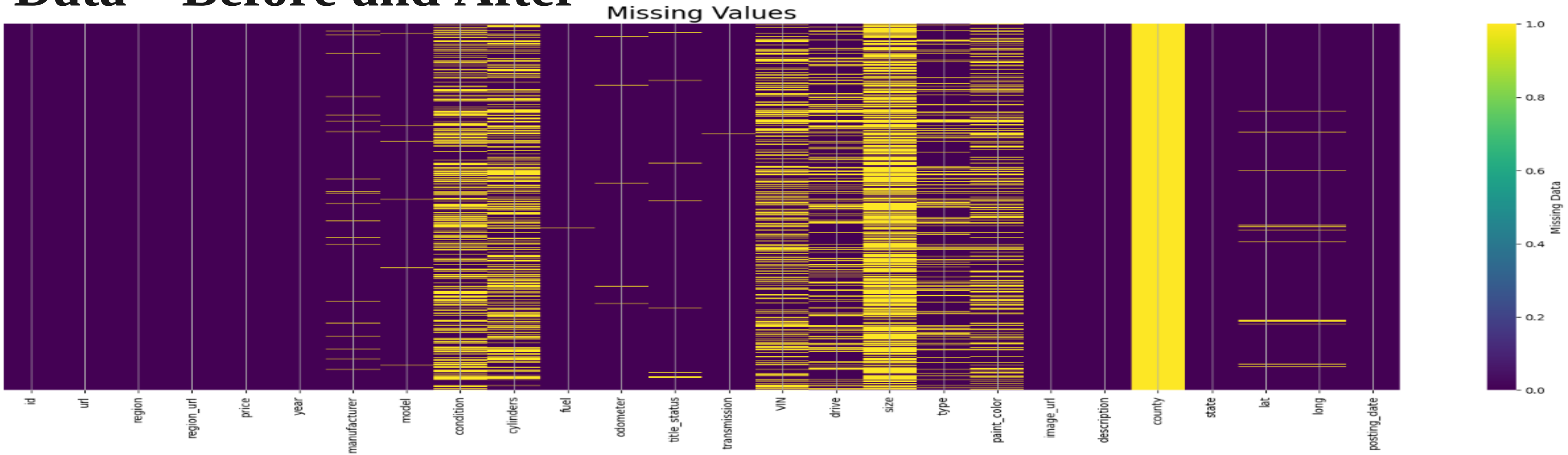
- Some of the few important attributes from 26 columns are

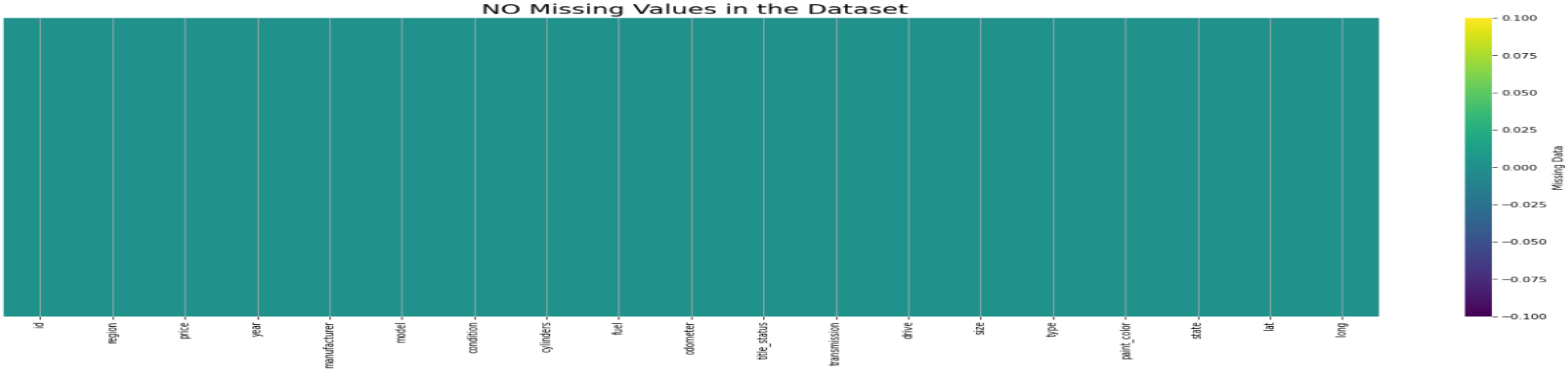
Attribute	Data Type	Description
PRICE	Integer Type	The Price of the used cars from the craigslist website
YEAR	Integer Type	The year in which car was manufactured
MANUFACTURER	String Type	The Car manufacturers such as Ford, Honda, Gmc
CONDITION	String Type	The Condition of a Car such as excellent, fair, new
CYLINDERS	String Type	The Cylinder type a car have such as four, six, eight cylinders
ODOMETER	Integer Type	The Odometer reading such as miles of a car driven
TYPE	String Type	The type of a car such as Sedan, Suv, Hatchback

Data Wrangling

- **Dropping Redundant Columns -** County, Id, Lat, Long, region, VIN, etc. and once removed those columns, they were left with around 14 attributes.
- **Handling Missing values -** Imputation(Mean – *Numerical Columns*, Mode- *Categorical Columns*), dropna(More than 75% Nan values)
- **Removal of Duplicates records**
- **Outlier Treatment -** Inter Quartile Range(IQR) – For columns Odometer, Price
- **Categorical Encoding** – Label Encoding – For categorical columns
- **Data Scaling -** Normalization (MinMaxScaler()) – In the range of 0 to 1.

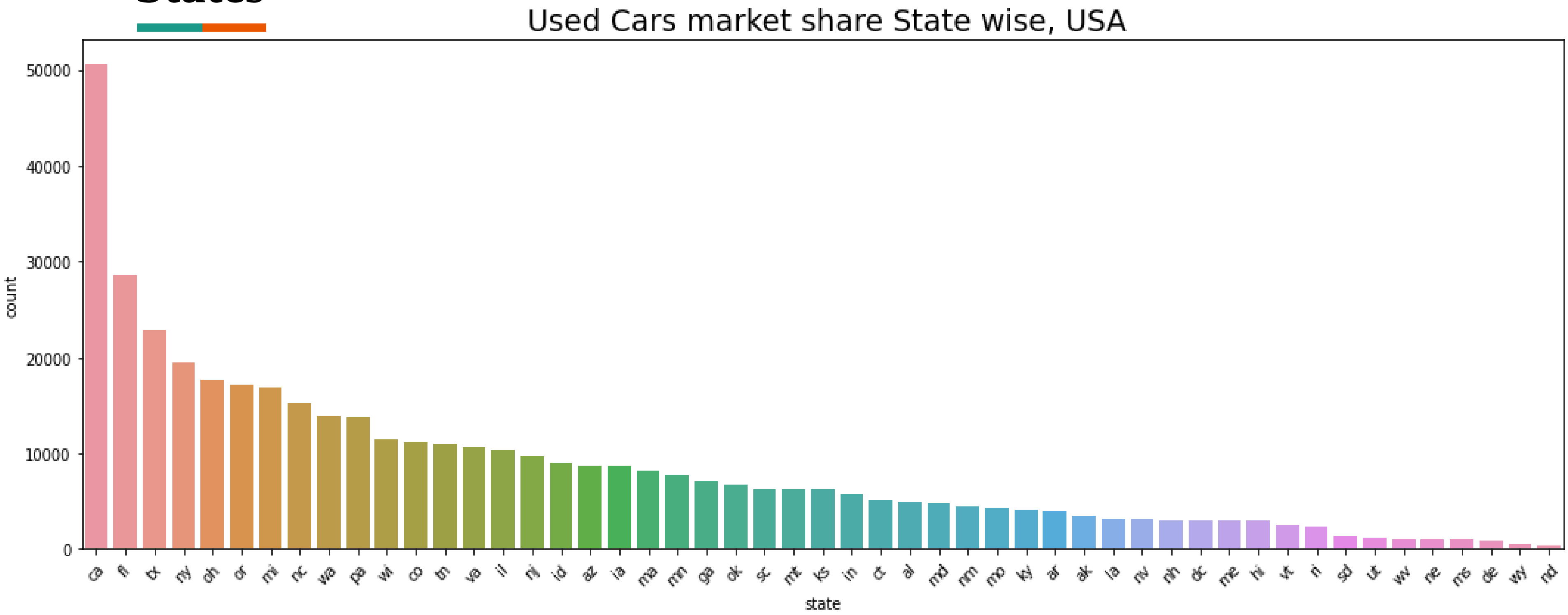
Data – Before and After


Missing Values

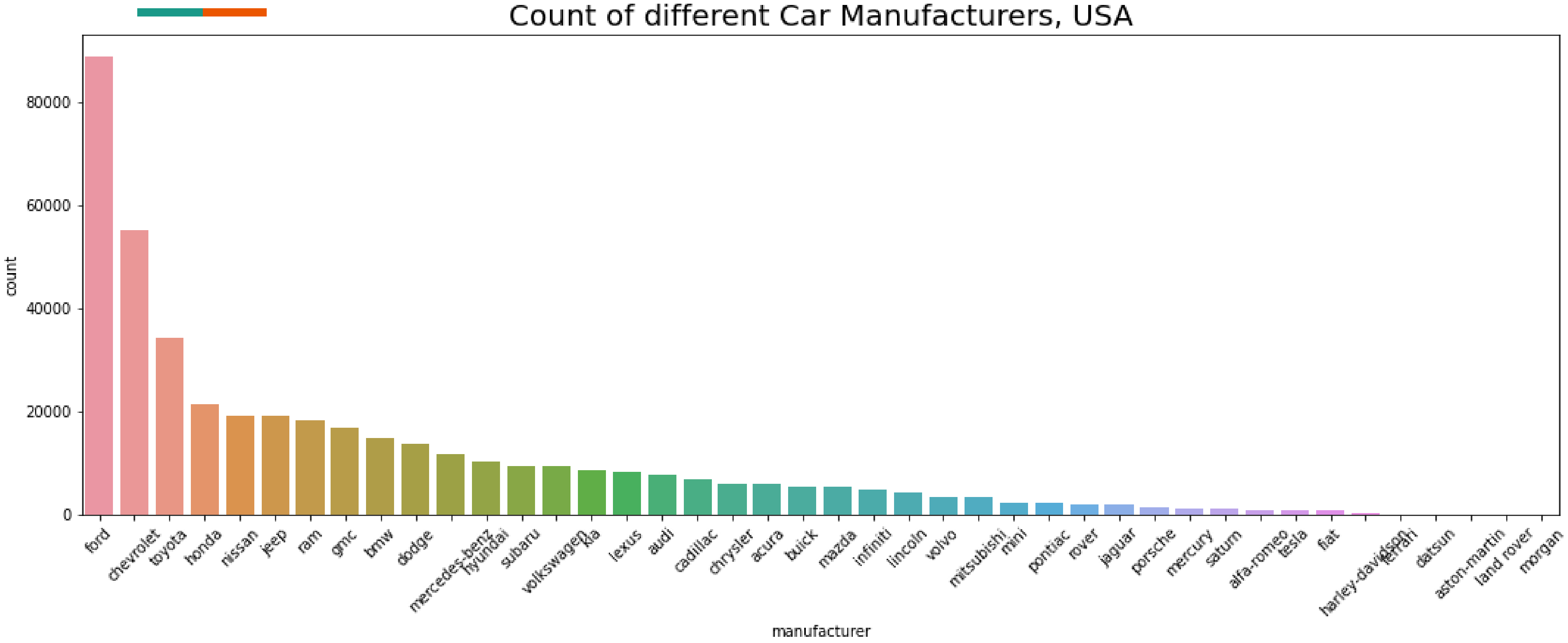

NO Missing Values in the Dataset

Potential Modeling Features

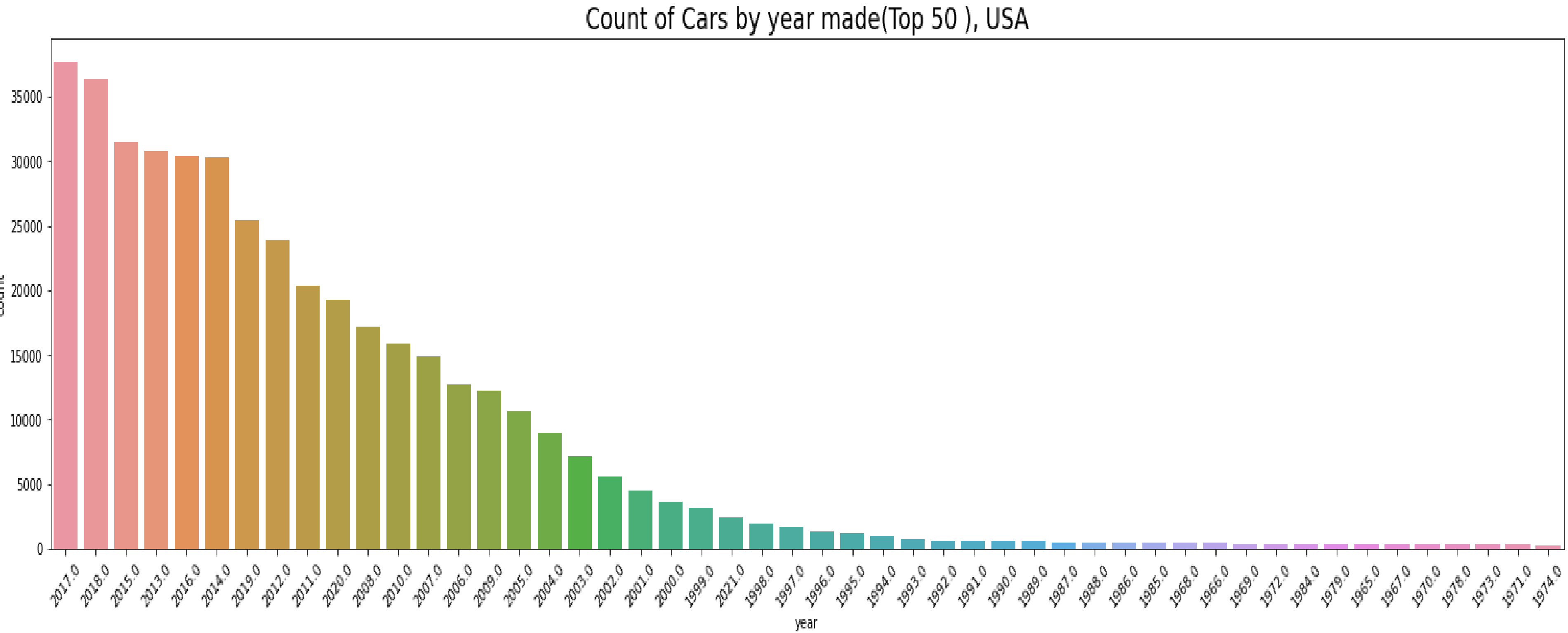
States


Used Cars market share State wise, USA
count
50000
40000
30000
20000
10000
0
state
ca
fl
tx
ny
oh
or
mi
nc
wa
pa
wi
co
tn
va
il
nj
id
az
ia
ma
mn
ga
ok
sc
mt
ks
in
ct
al
md
nm
mo
ky
ar
ak
la
nv
nh
dc
me
hi
vt
ri
sd
ut
wv
ne
ms
de
wy
nd

Car Manufacturers

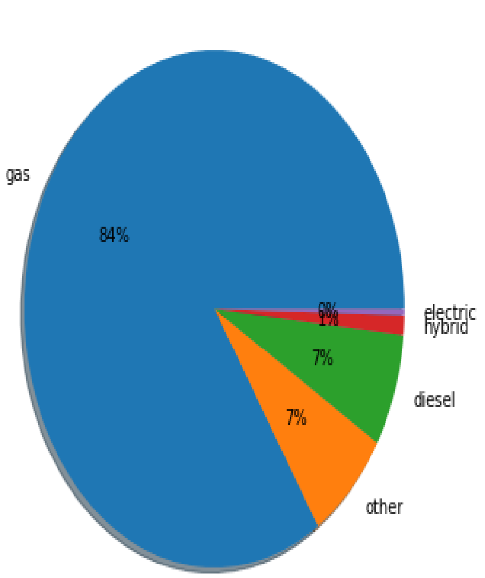


Year Make

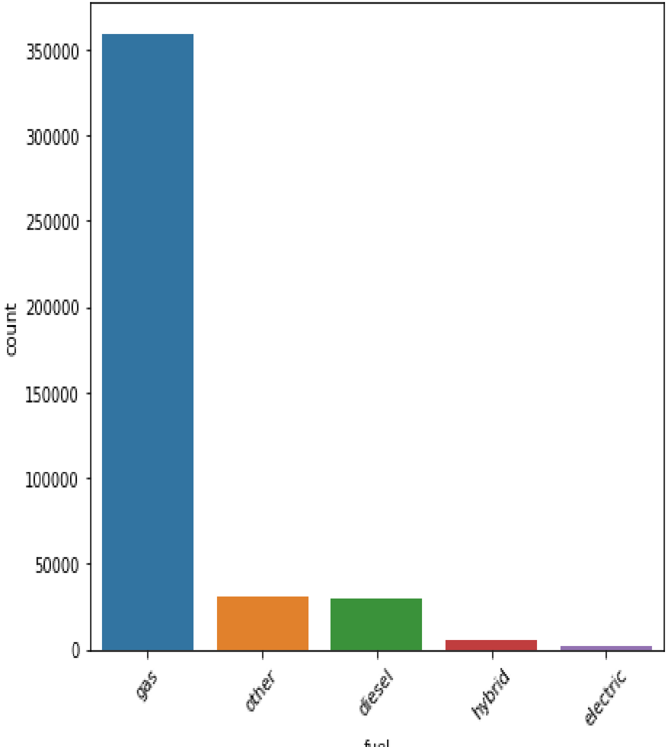

Count of Cars by year made(Top 50), USA
count
35000
30000
25000
20000
15000
10000
5000
0
2017.0
2018.0
2015.0
2013.0
2016.0
2014.0
2019.0
2012.0
2011.0
2020.0
2008.0
2010.0
2007.0
2006.0
2009.0
2005.0
2004.0
2003.0
2002.0
2001.0
2000.0
1999.0
2021.0
1998.0
1997.0
1996.0
1995.0
1994.0
1993.0
1992.0
1991.0
1990.0
1989.0
1987.0
1988.0
1986.0
1985.0
1968.0
1966.0
1969.0
1972.0
1984.0
1979.0
1965.0
1967.0
1970.0
1978.0
1973.0
1971.0
1974.0
year

Fuel Type

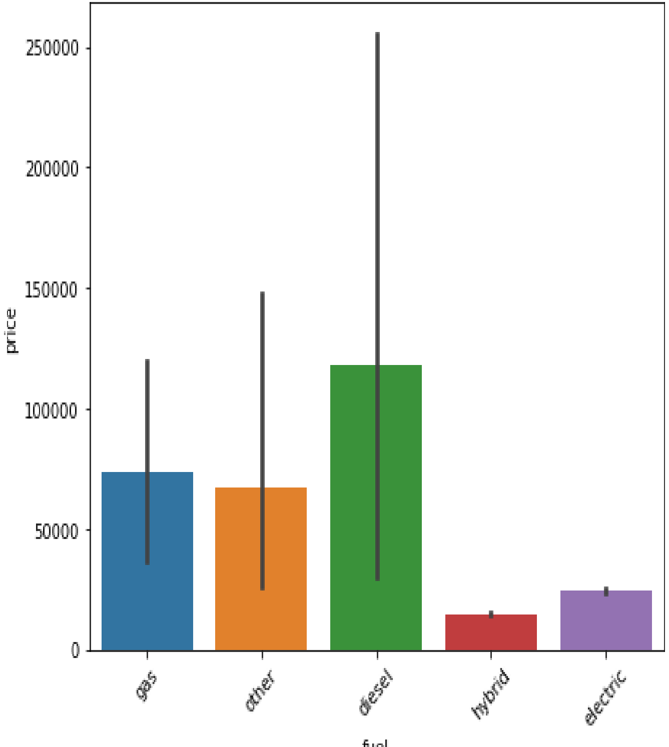
Percentage of different fuel types used in cars



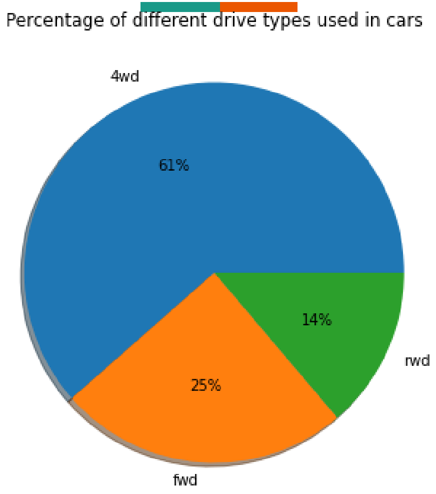
Number of cars by fuel type

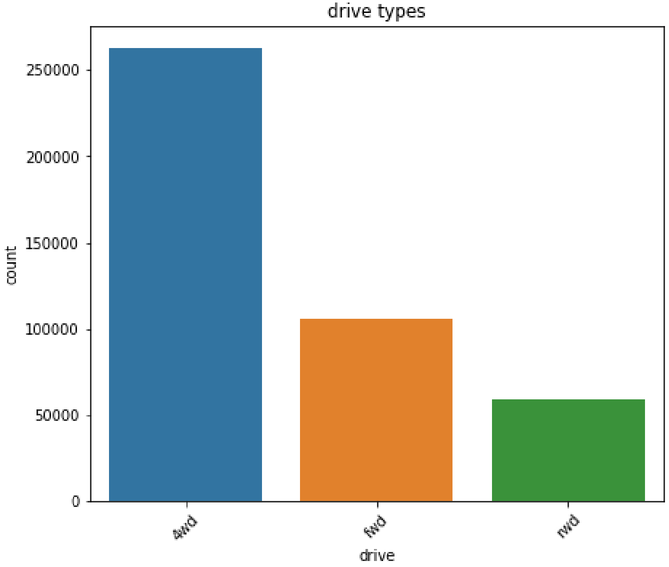


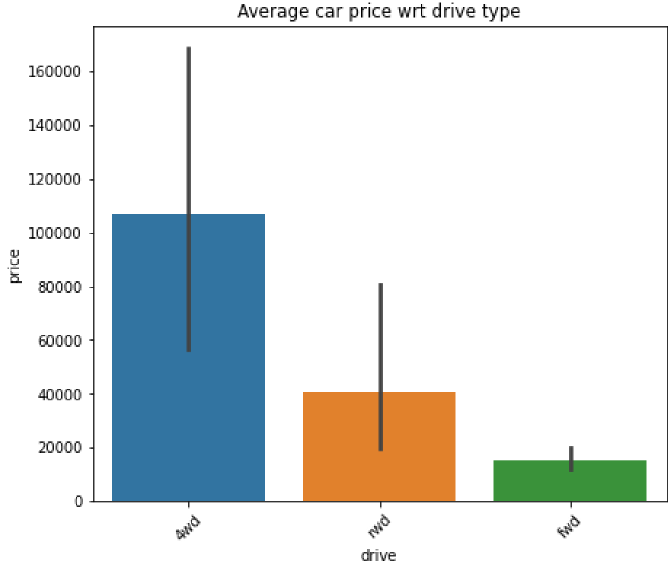
Average car price wrt fuel type



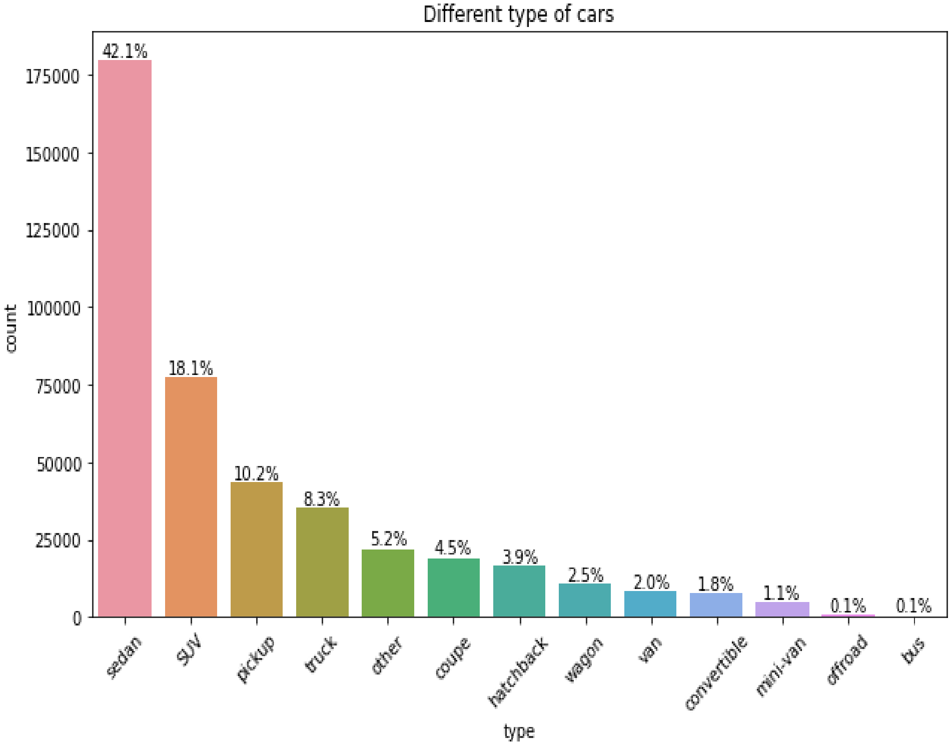
Drive

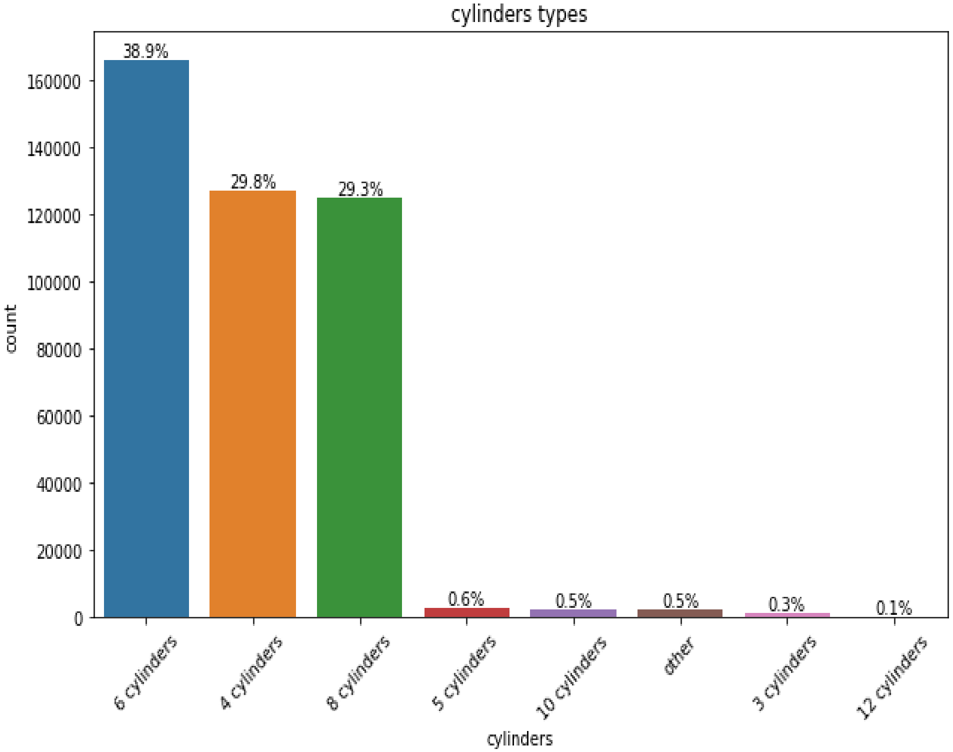

Percentage of different drive types used in cars


drive types

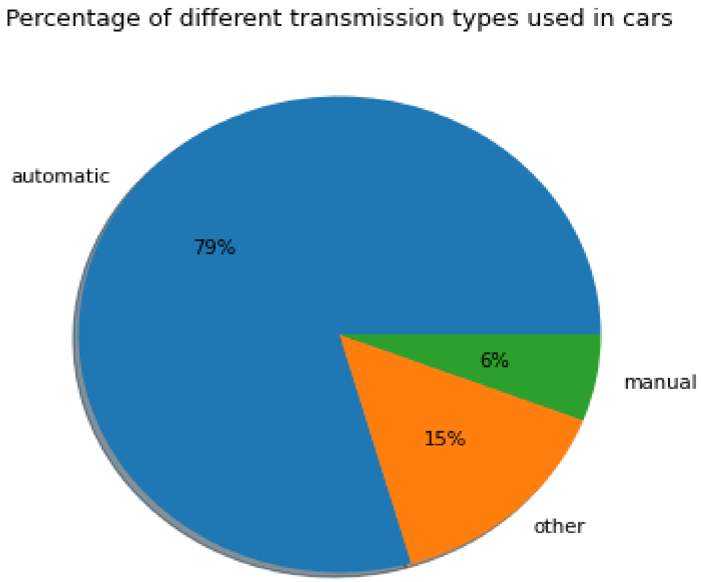

Average car price wrt drive type

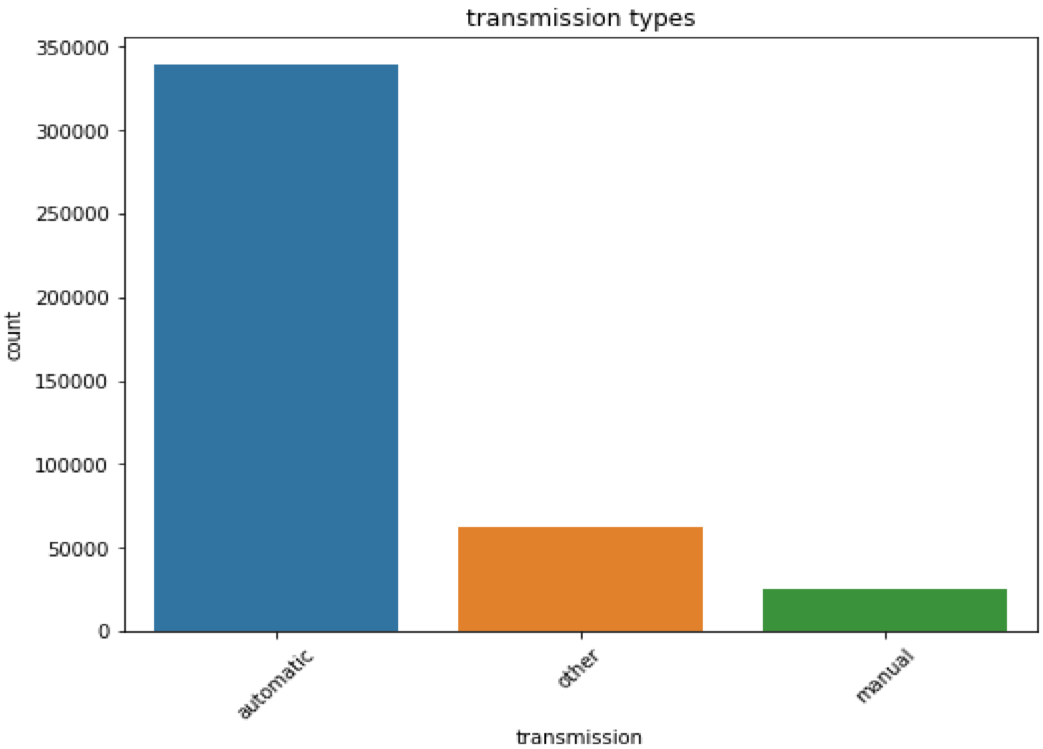
Different type of cars and Cylinders types



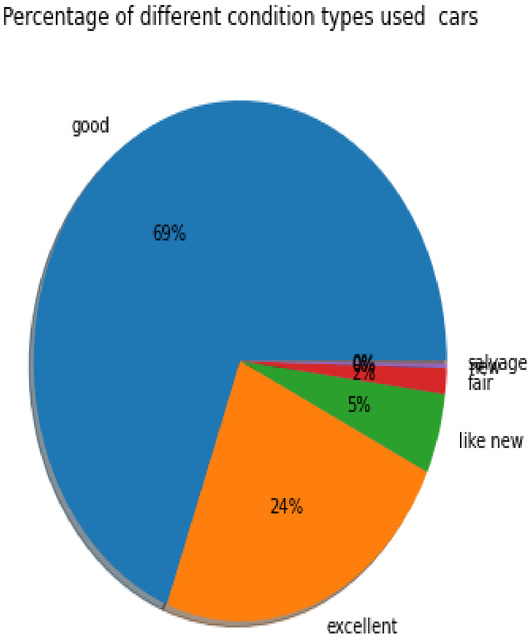


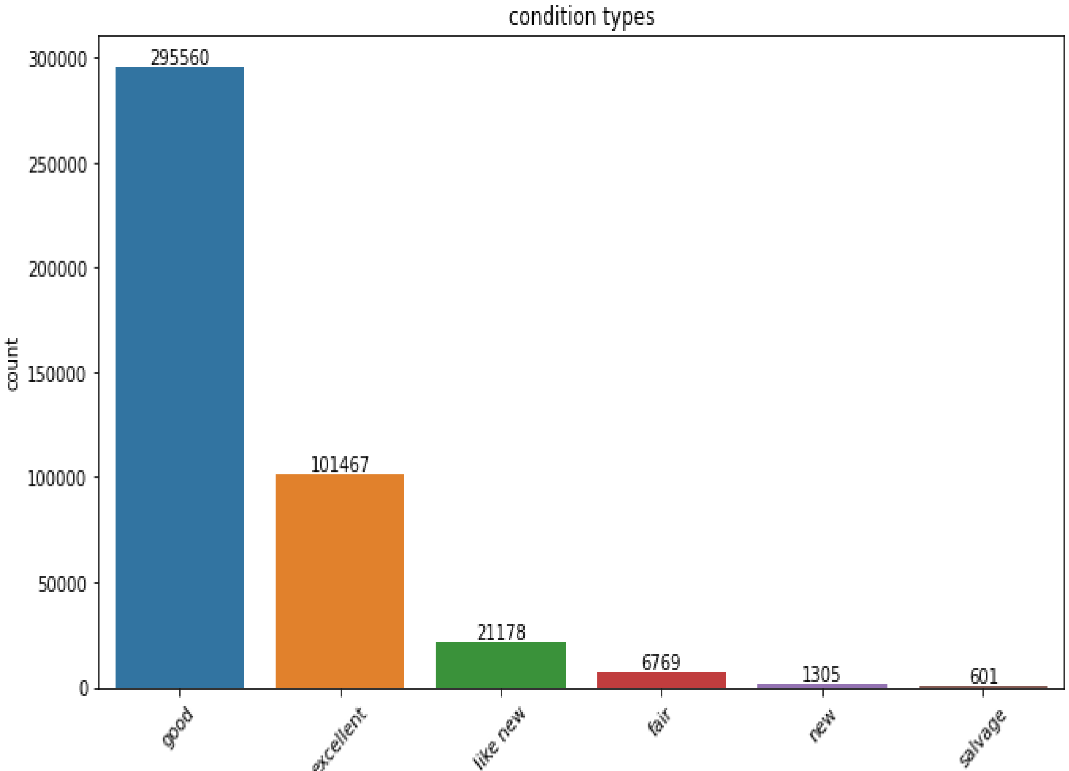
Transmission



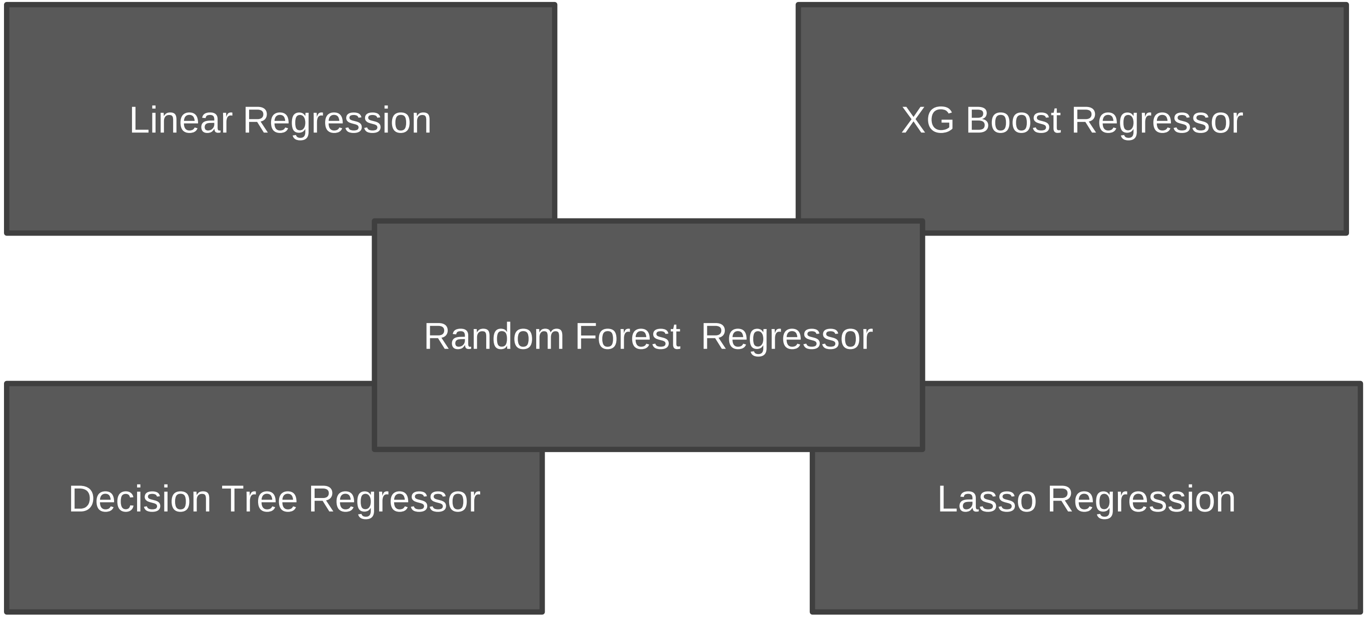


Condition


Percentage of different condition types used cars
good
69%
excellent
24%
like new
5%
fair
2%
new
0%
salvage
0%


condition types
count
300000
250000
200000
150000
100000
50000
0
295560
101467
21178
6769
1305
601
good
excellent
like new
fair
new
salvage

Models



Linear Regression

XG Boost Regressor

Random Forest Regressor

Decision Tree Regressor

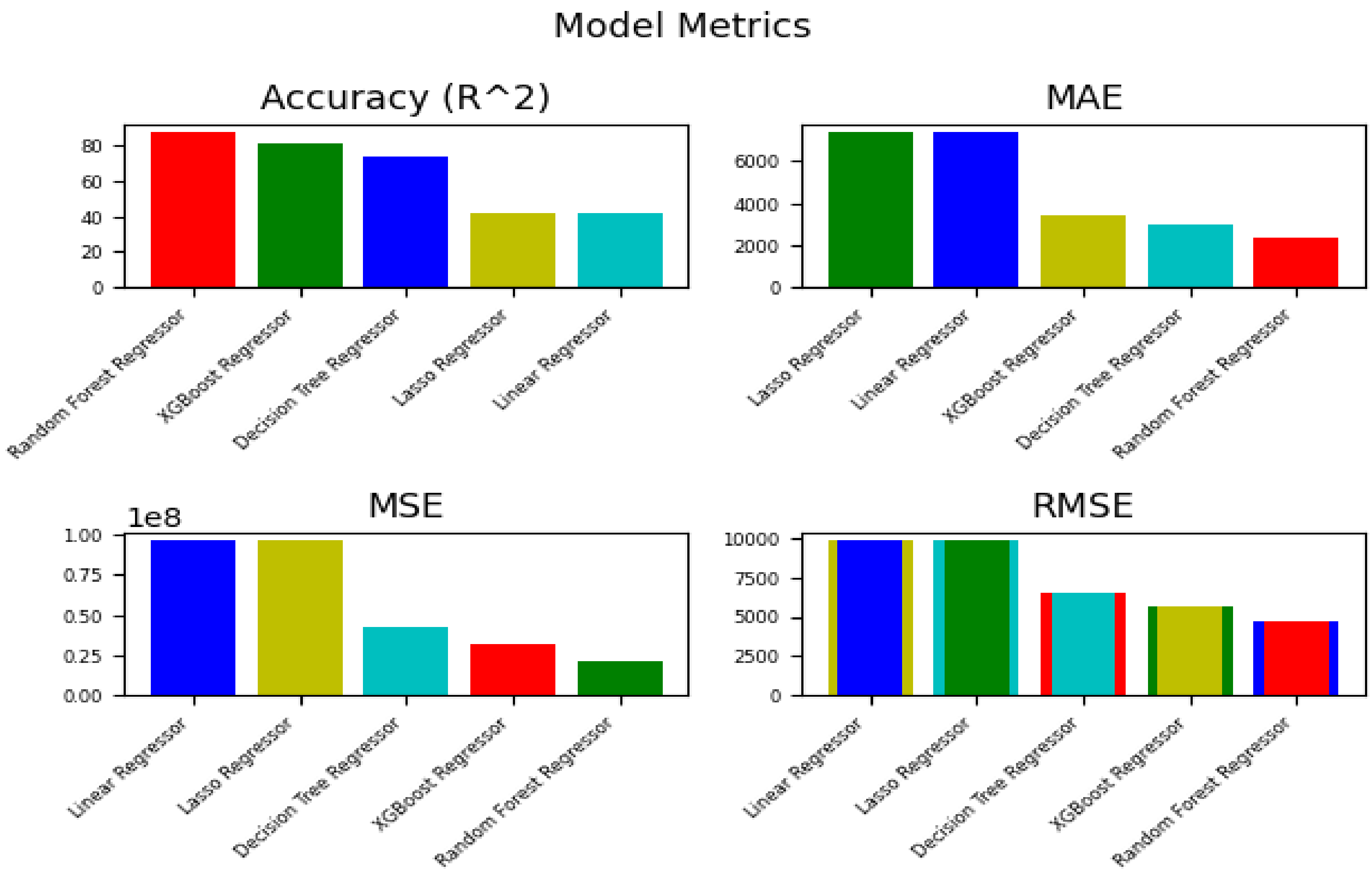
Lasso Regression

Model Building and Deployment Steps

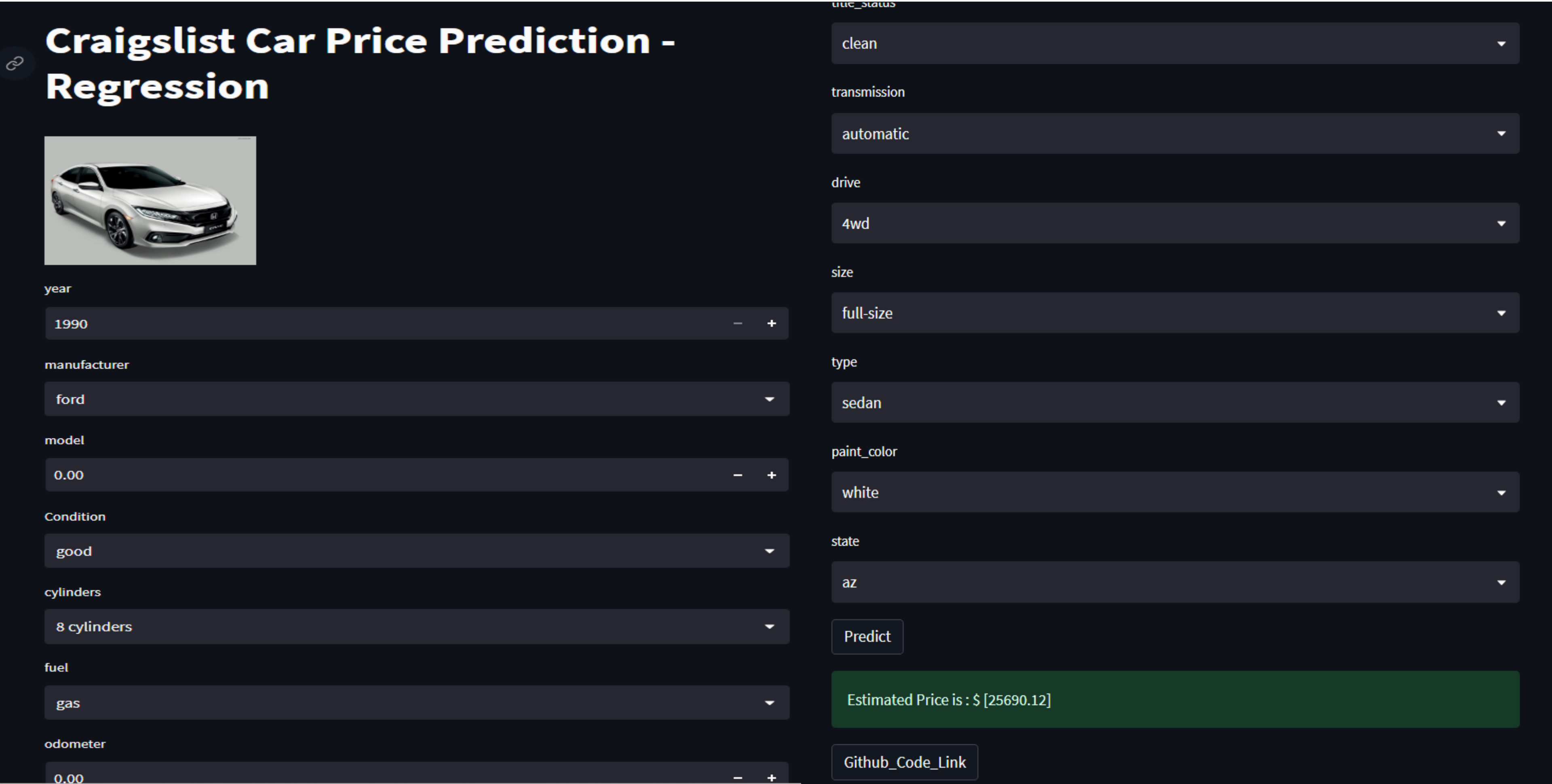
- Splitting data into : **Train = 70 %** **Test = 30 %**
- Label Encoding of categorical Columns
- First built models with default parameters
- Hyperparameter Tuning – Randomized SearchCV - To improve the Accuracy
- Study models for any Underfitting or Overfitting
- Used the Random Forest model with best accuracy score as – **R**2** (Coefficient Of Determination) – 86%.
- Best Model is saved as Pickle file.
- Used Streamlit for deploying the model as WebApp.

Results

Model	Accuracy	MAE	MSE	RMSE
Random Forest Regressor	86.78	2382.29	21701516.81	4658.48
XGBoost Regressor	80.78	3466.68	31542312.17	5616.25
Decision Tree Regressor	73.90	3008.64	42549793.55	6523.02
Lasso Regressor	41.52	7326.61	96013084.67	9798.62
Linear Regressor	41.51	7325.84	96015990.90	9798.77


Model Metrics
Accuracy (R^2)
0
20
40
60
80
Random Forest Regressor
XGBoost Regressor
Decision Tree Regressor
Lasso Regressor
Linear Regressor
MAE
0
2000
4000
6000
Lasso Regressor
Linear Regressor
XGBoost Regressor
Decision Tree Regressor
Random Forest Regressor
MSE
1e8
0.00
0.25
0.50
0.75
1.00
Linear Regressor
Lasso Regressor
Decision Tree Regressor
XGBoost Regressor
Random Forest Regressor
RMSE
0
2500
5000
7500
10000
Linear Regressor
Lasso Regressor
Decision Tree Regressor
XGBoost Regressor
Random Forest Regressor

Graphical User Interface



Craigslist Car Price Prediction - Regression

year

1990

manufacturer

ford

model

0.00

Condition

good

cylinders

8 cylinders

fuel

gas

odometer

0.00

title_status

clean

transmission

automatic

drive

4wd

size

full-size

type

sedan

paint_color

white

state

az

Estimated Price is : $ [25690.12]

References

- https://www.census.gov/quickfacts/CA
- https://dagshub.com/blog/ci-cd-for-machine-learning-test-and-and-deploy-your-ml-model-with-github-actions/
- https://towardsdatascience.com/what-and-why-behind-fit-transform-vs-transform-in-scikit-learn-78f915cf96fe
- https://washingtondc.craigslist.org/

Thank You!