IST. DATA

VEHICLE PRICE PREDICTION

via kaggle dataset

HALİL İBRAHİM KAYA

DATASET INFO

Selected dataset contains **13** features related to a car such as name, year, mileage, fuel type, engine, etc. The objective of this project is to predict vehicle prices using these features.

The dataset, which is quite messy, contains **8128** samples. The dataset has gone through various stages to create an appropriate model.

PROJECT STAGES

01 INITIAL OBSERVATIONS

02 TYPE CONVERSIONS-CREATING FEATURES

03 EDA & VISUALIZATIONS

04 PREPROCESSING

05 FEATURE SELECTION

06 MODEL SELECTION

07 EVALUATION

FIRST STAGES

After the initial observations, it was noticed that there were a lot of **duplicate** rows in the dataset, and these were removed. Subsequently, some columns with problematic types underwent type conversion with the help of **regex** and were similarly scaled using **domain knowledge**.

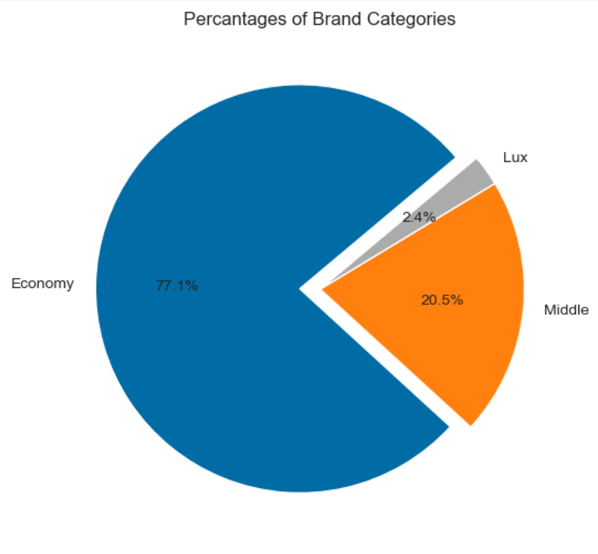
mileage	engine	max_power	torque
23.4 kmpl	1248 CC	74 bhp	190Nm@ 2000rpm
21.14 kmpl	1498 CC	103.52 bhp	250Nm@ 1500-2500rpm
17.7 kmpl	1497 CC	78 bhp	12.7@ 2,700(kgm@ rpm)
23.0 kmpl	1396 CC	90 bhp	22.4 kgm at 1750-2750rpm
16.1 kmpl	1298 CC	88.2 bhp	11.5@ 4,500(kgm@ rpm)

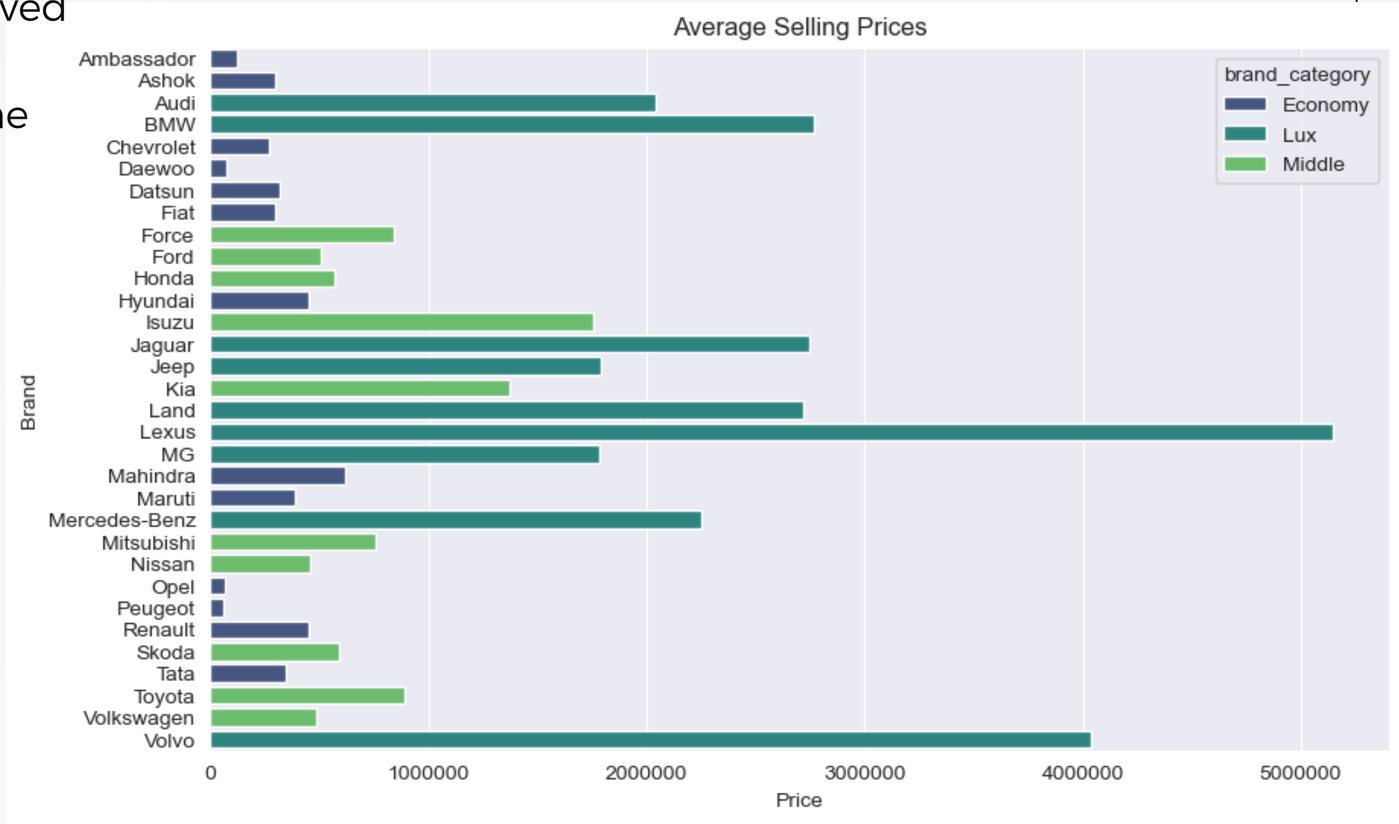
mileage	engine	max_power	seats	max_torque	rpm
23.40	1248.0	7040.00	5.0	190.00	2000.0
21.14	1498.0	100030.05	5.0	250.00	2500.0
17.70	1497.0	7080.00	5.0	124.54	2700.0
23.00	1396.0	9000.00	5.0	219.66	2750.0
16.10	1298.0	8080.02	5.0	11.50	4500.0

NOTE:
1KGM=9.81NM

FIRST STAGES

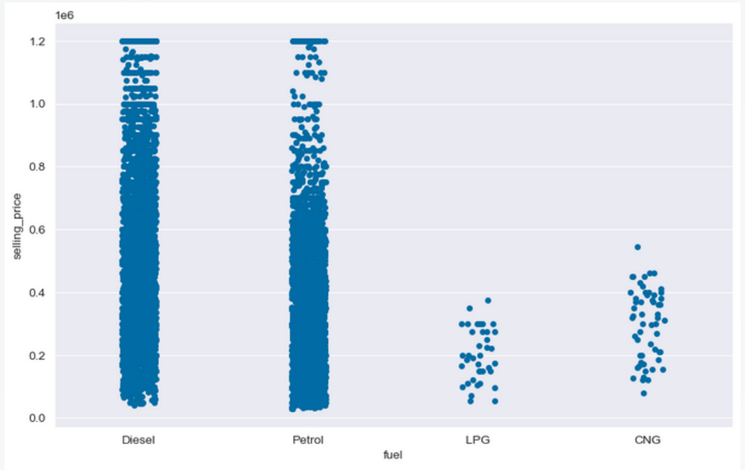
Car **brands** were extracted from the car names, and they were **categorized** into three groups: Luxury, Middle, and Economy, using some domain knowledge. It was observed during the EDA that these categories were consistent. Additionally, it was noted that a very small portion of the sample owned luxury cars.

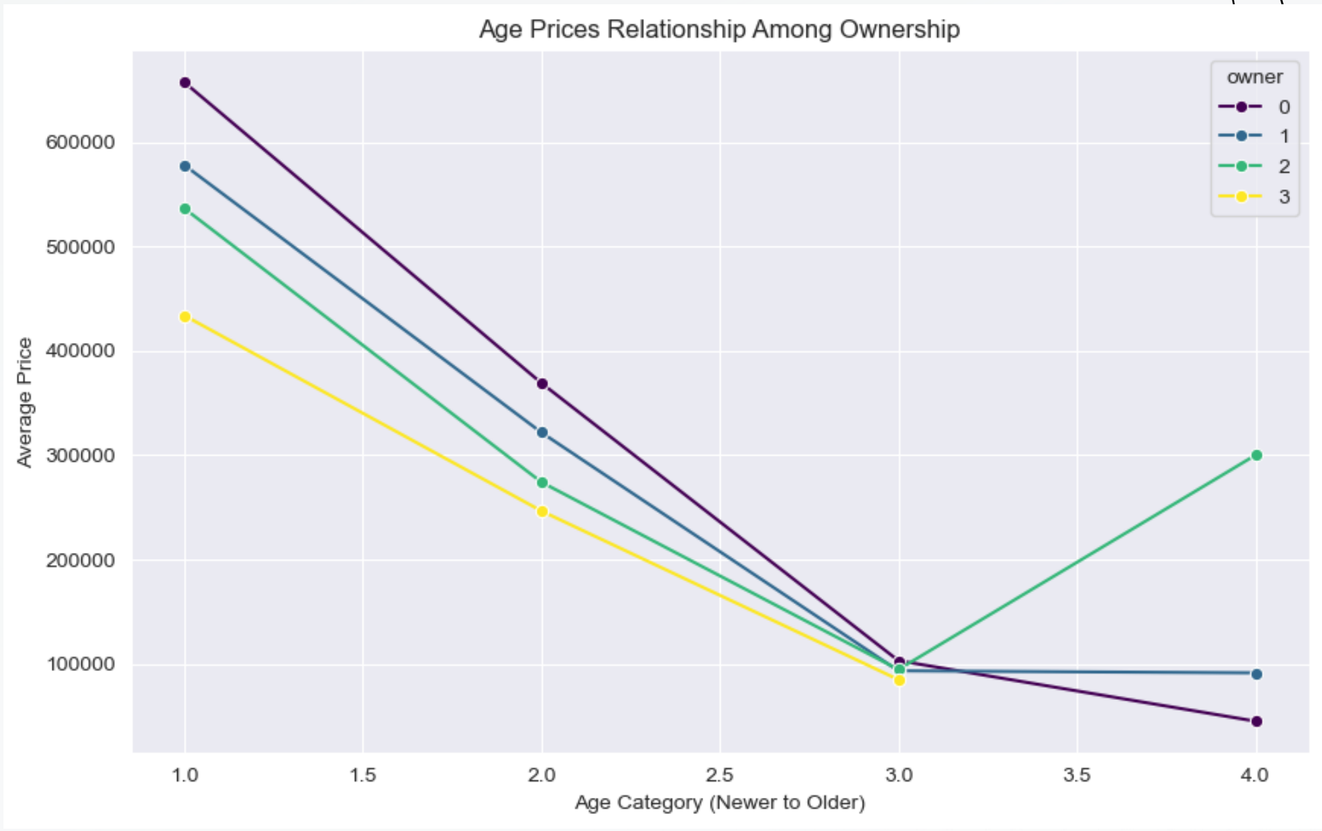




FIRST STAGES

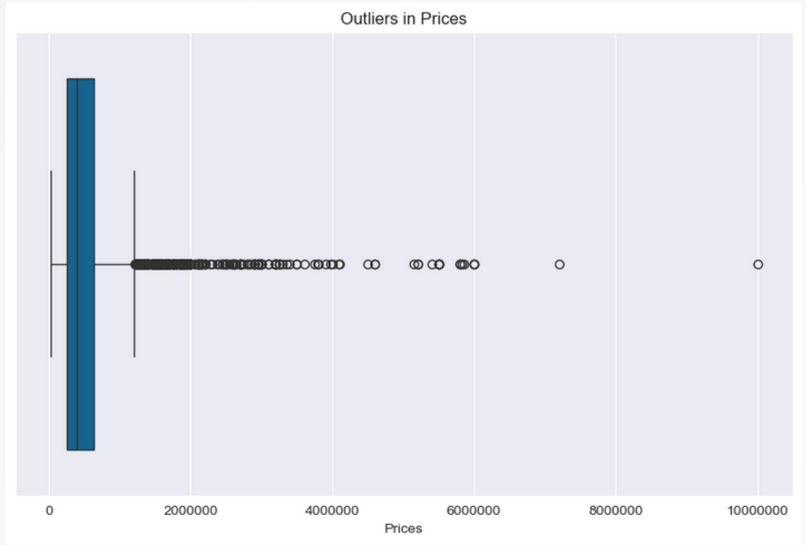
The year the dataset was created was subtracted from the car's manufacturing year to determine the car's **age**, which was then **categorized**. When examining **ownership status** with these categories, a meaningful relationship emerged. Additionally, when examining fuel types, it was noticed that there were very few vehicles in the LPG and CNG categories, so these were combined into the '**Other**' category.

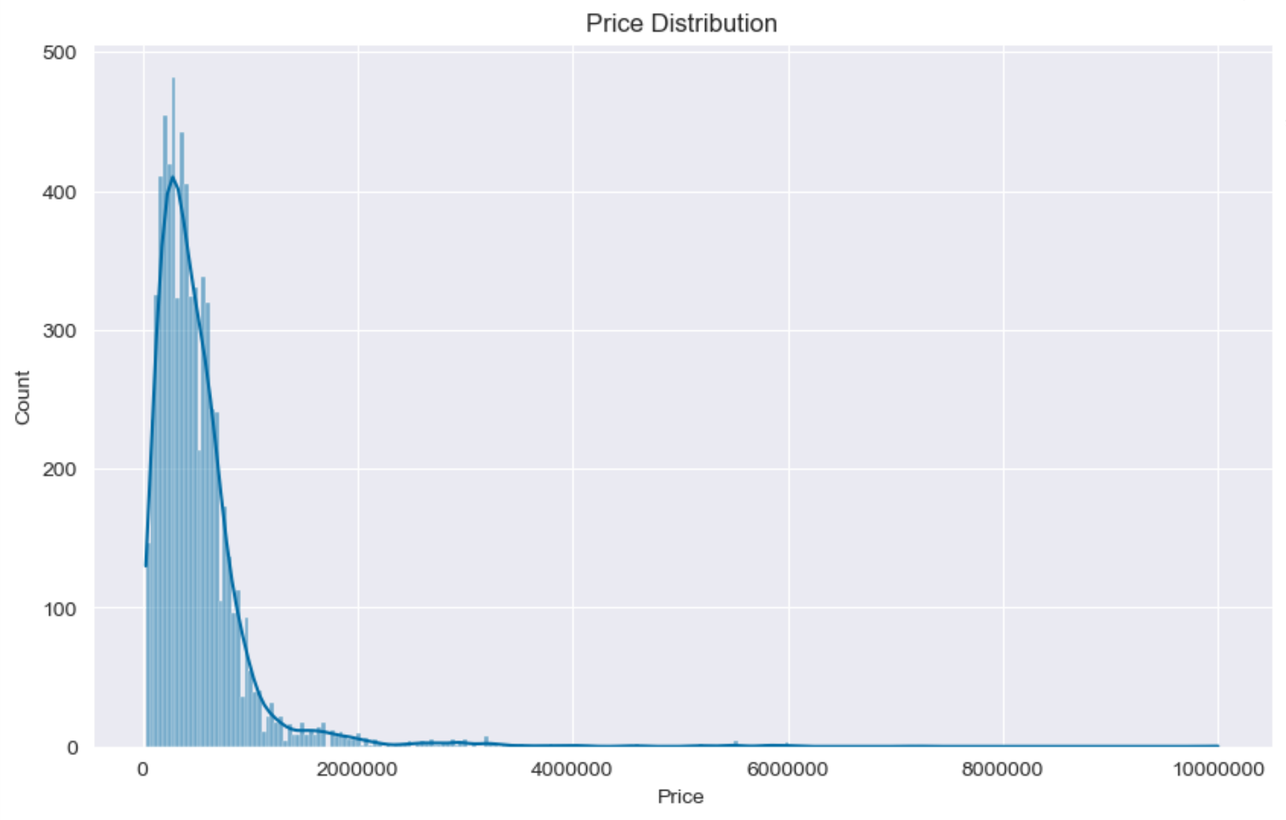




FIRST STAGES

The price distribution was observed to be **right-skewed**, supporting previous findings. Also **outliers** were identified and **suppressed** using the **IQR** method. Additionally, outliers in other numeric columns were investigated.

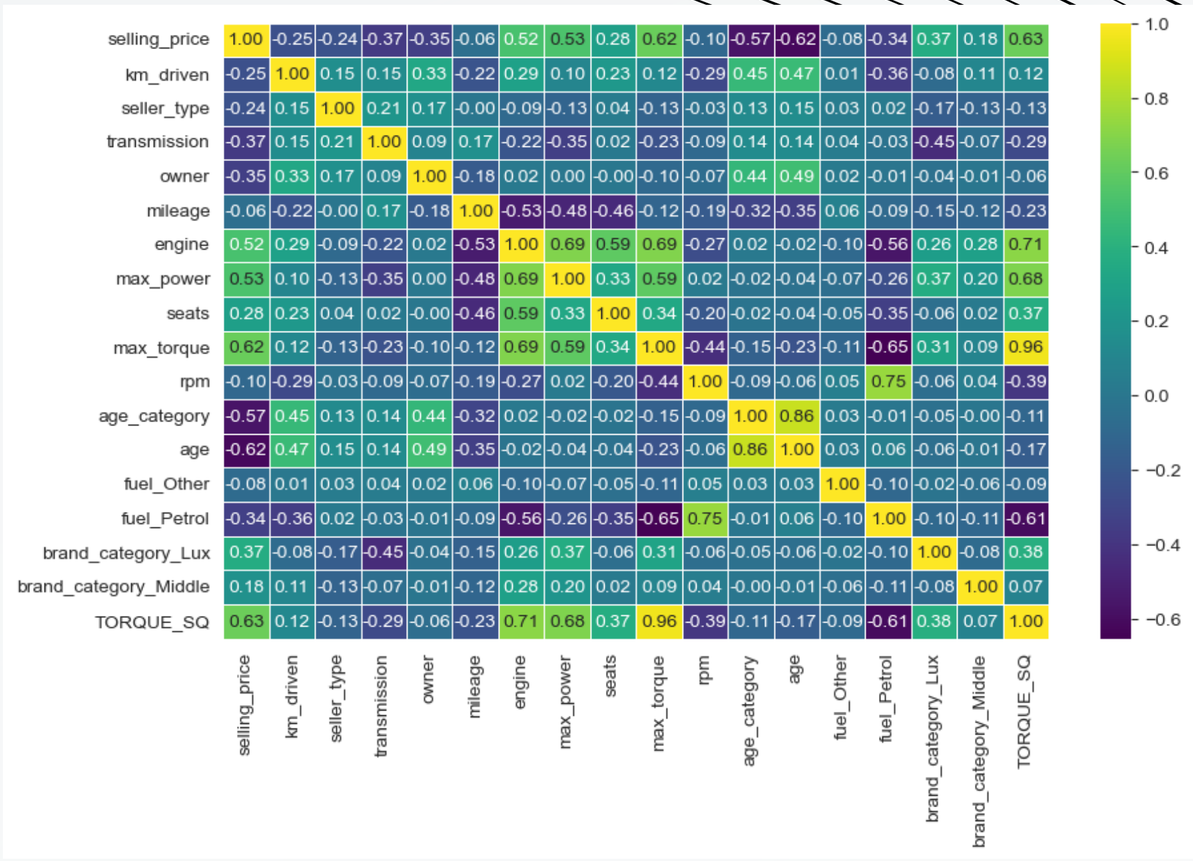




PREPROCESSING

Before selecting the model, the following preprocessing steps were applied too:

- **Null** values were filled with the median.
- A pairplot was observed. (*As a result, the 'torque' column was **squared**, resulting in a minor improvement in correlation.*)
- Label **encoding** or one-hot encoding was applied to necessary features.
- **Correlations** were observed.
- **Scaling** was performed using Standard Scaler.
- Dataset splitted into **train and test**
- **Log** transformation applied to y_train (*%5 r2 improvement*)

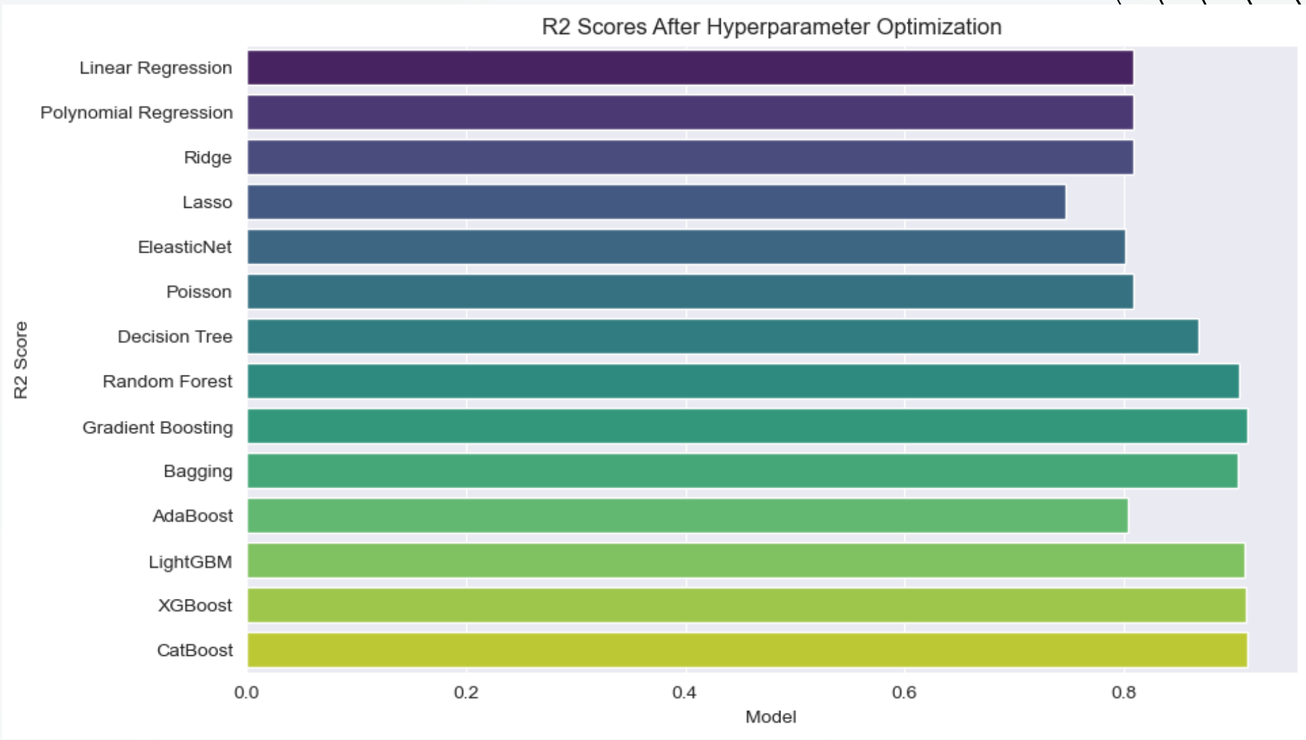


- A new feature, '**Horse Power**', was derived to enhance the efficiency of the rpm feature with low correlation.

- *HP = Torque * RPM / 5252*

MODEL SELECTION

Model selection, **hyperparameter optimization**, and **cross-validation** were **automated** using a function. Parameter **grids** were created for each model, and model-grid pairs were sent to the function. Once optimization processes were completed, the models were **saved** in a dictionary based on their cross-validation **R2** scores. To make more accurate predictions, the top 5 models were combined to create a **Voting Regressor**.

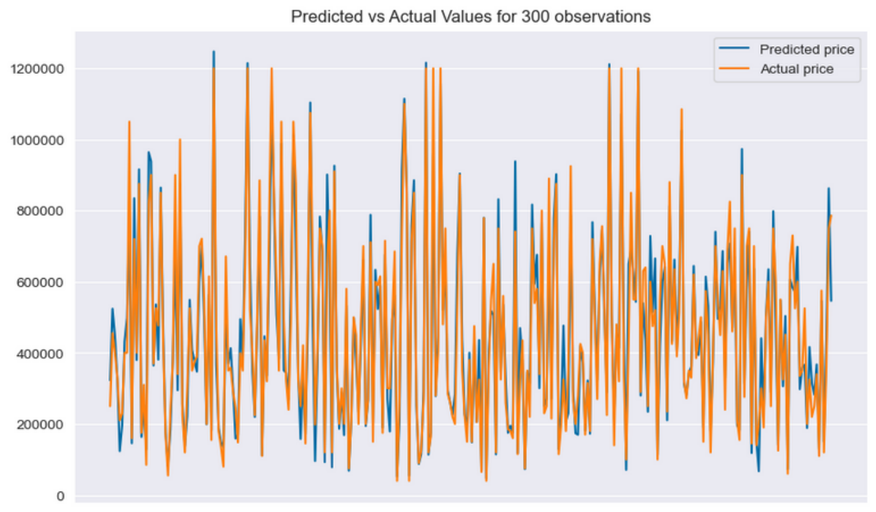


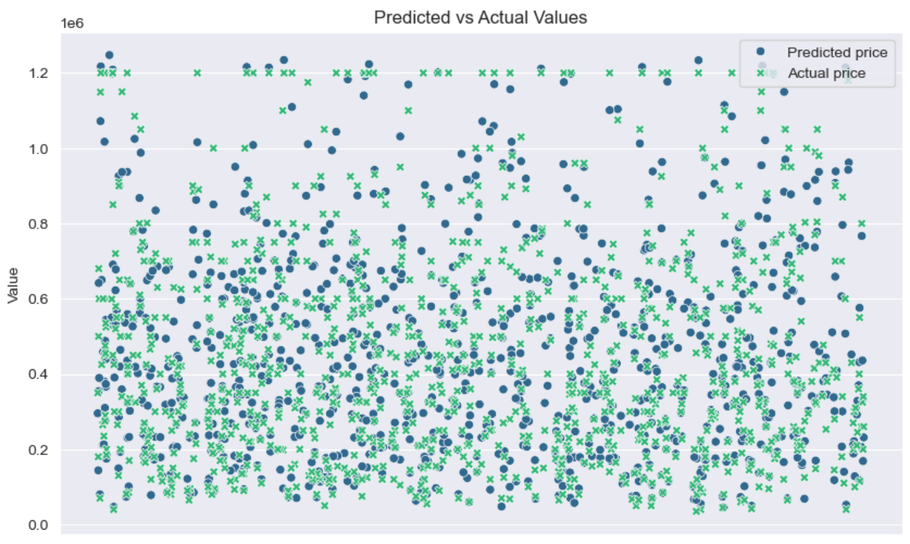
NOTE***:*** *Before above,* ***p-values*** *were examined using* ***statmodels****. Since no issues were encountered, an advanced* ***feature selection*** *algorithm called* ***RFE*** *was used to* ***reduce*** *the number of features. However, as it did not positively impact performance, it was abandoned.*

EVALUATION

After training the selected model and testing it on the test data, **promising** results were obtained.

R2	0.90
MSE	7694885587.66
MAE	58107.69
RMSE	87720.49
MAPE	15.33
Confidence Interval	76149.30 - 97933.92



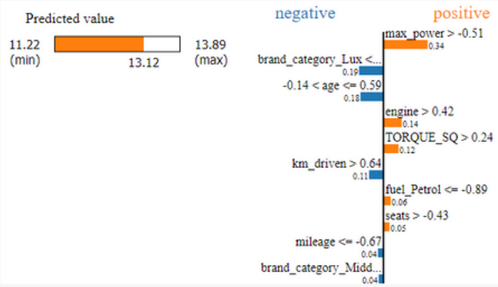


NOTE: ***Confidence intervals checked with the help of stats library for %95 confidence***

X-AI

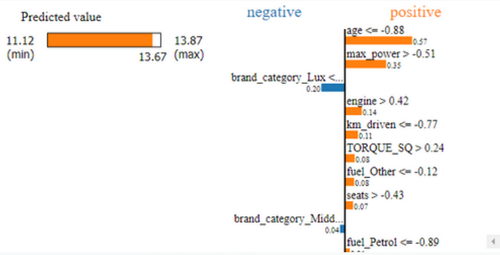
Finally, the decisions made by the model based on specific samples were observed using the **LIME** library, which provides **explanations** for model predictions(feature importance for 10 feats.).

SAMPLE 1



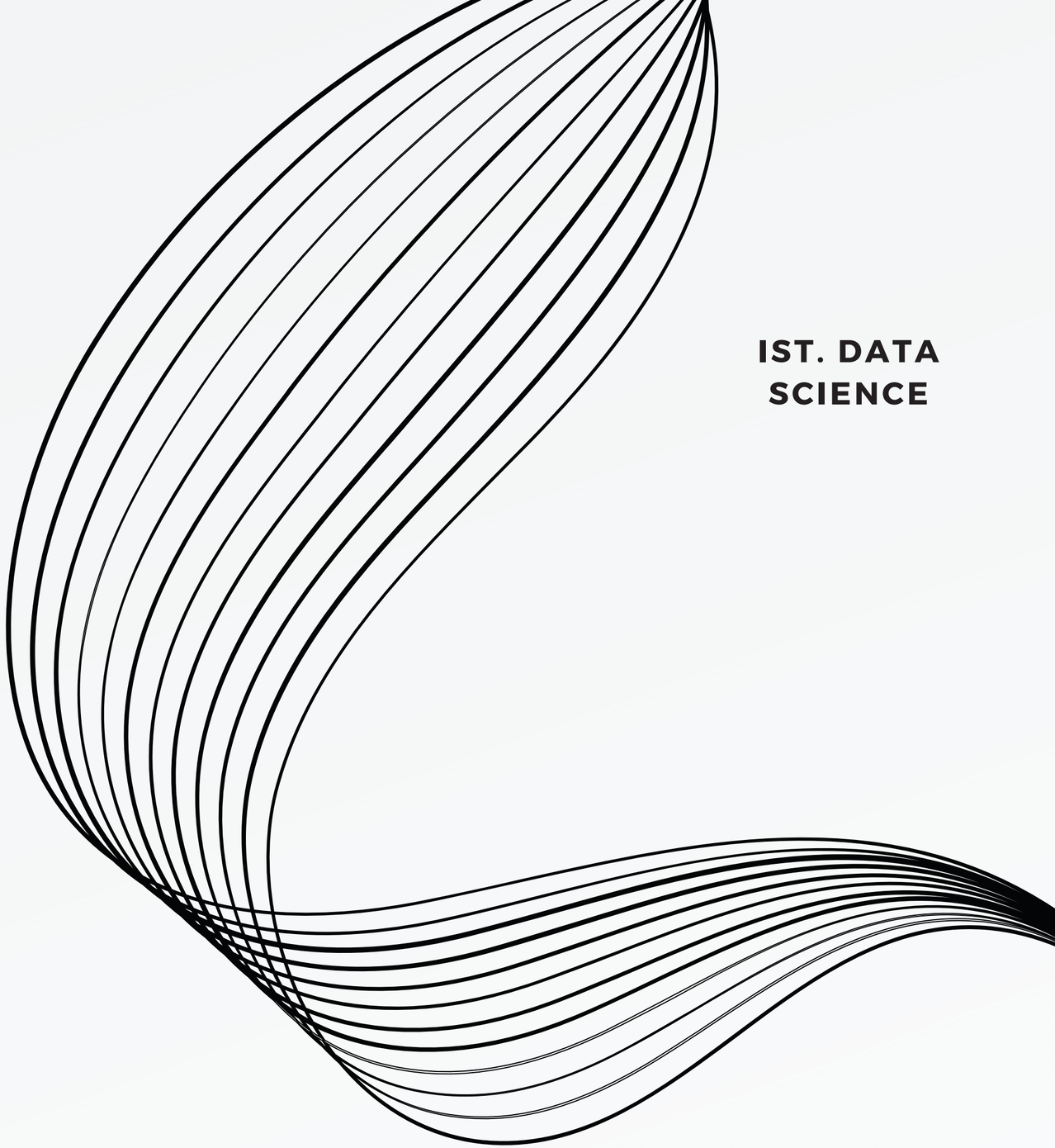
Feature	Value
max_power	1.66
brand_category_Lux	-0.16
age	0.59
engine	1.74
TORQUE_SQ	1.70
km_driven	1.79
fuel_Petrol	-0.89
seats	2.65
mileage	-1.92
brand_category_Middle	-0.51

SAMPLE 2



Feature	Value
age	-1.12
max_power	1.61
brand_category_Lux	-0.16
engine	1.74
km_driven	-1.32
TORQUE_SQ	0.94
fuel_Other	-0.12
seats	0.59
brand_category_Middle	-0.51
fuel_Petrol	-0.89

THANKS FOR LISTENING



IST. DATA SCIENCE

For more info, please check the related notebook