ELL409 Project: Old Car Price Prediction

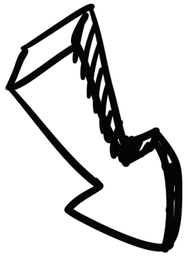
24th November, 2024

Bharat Agarwal: 2022EE11790
Harsh Agarwal: 2021EE1
Abhinav Verma:

Dataset

- Only 50000 datapoints and no missing data!

id	brand	model	model_year	milage	fuel_type	engine	transmission	ext_col	int_col	accident	clean_title	price
0	Ford	F-150 Lariat	2018	74349	Gasoline	375.0HP 3.5L V6 Cylinder Engine Gasoline Fuel	10-Speed A/T	Blue	Gray	None reported	Yes	11000
1	BMW	335 i	2007	80000	Gasoline	300.0HP 3.0L Straight 6 Cylinder Engine Gasoli...	6-Speed M/T	Black	Black	None reported	Yes	8250
2	Jaguar	XF Luxury	2009	91491	Gasoline	300.0HP 4.2L 8 Cylinder Engine Gasoline Fuel	6-Speed A/T	Purple	Beige	None reported	Yes	15000
3	BMW	X7 xDrive40i	2022	2437	Hybrid	335.0HP 3.0L Straight 6 Cylinder Engine Gasoli...	Transmission w/Dual Shift Mode	Gray	Brown	None reported	Yes	63500
4	Pontiac	Firebird Base	2001	111000	Gasoline	200.0HP 3.8L V6 Cylinder Engine Gasoline Fuel	A/T	White	Black	None reported	Yes	7850

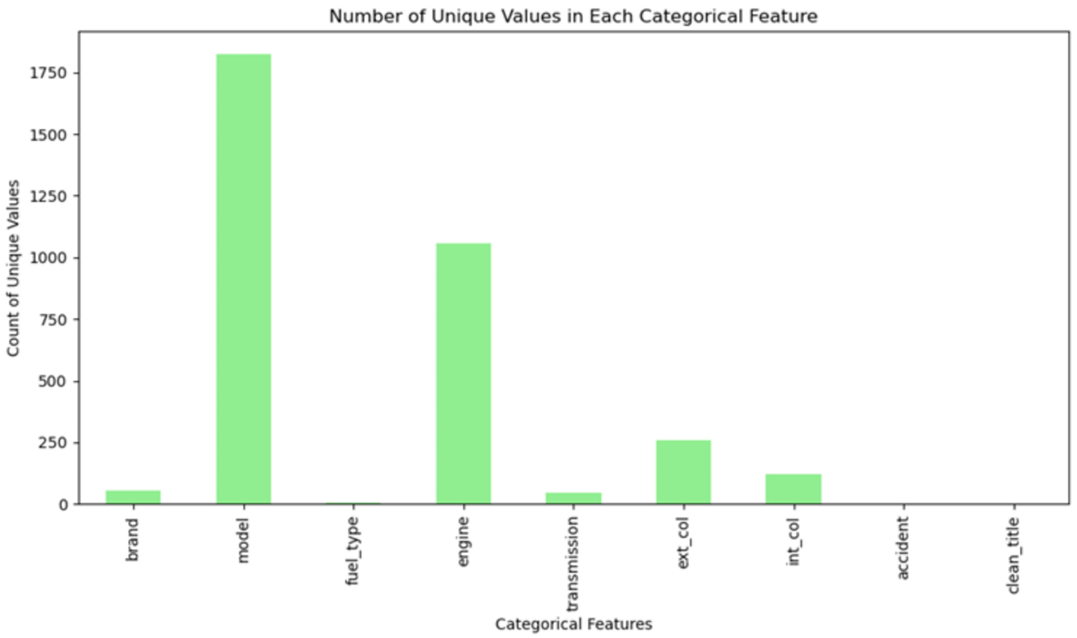


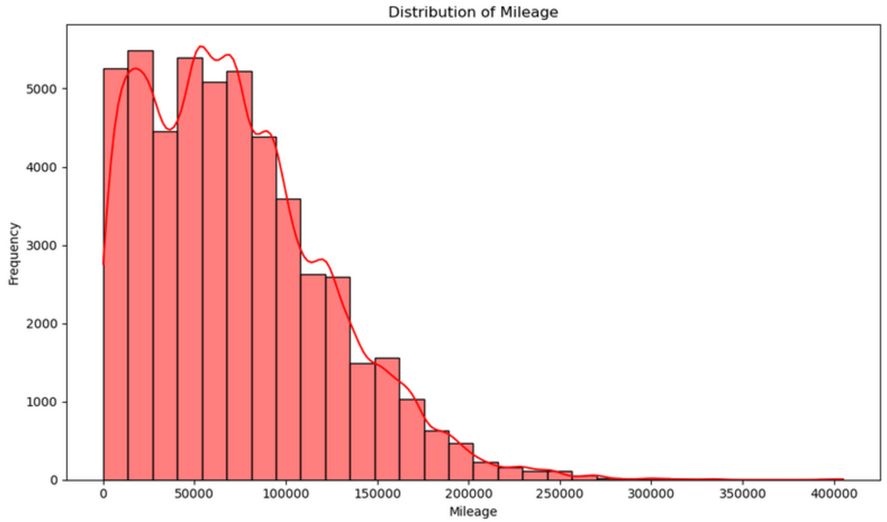
	milage	damage	age	ext_common	ext_neutral	int_common	int_neutral	Automatic	Manual	Semi	popular_engine	Luxury	Affordable	gasoline	price
0	74349	0	6	0	1	0	1	1	0	0	1	0	1	1	11000
1	80000	0	17	1	0	1	0	0	1	0	1	0	0	1	8250
2	91491	0	15	0	0	0	1	1	0	0	0	0	0	1	15000
3	2437	0	2	0	1	0	0	0	0	1	0	0	0	0	63500
4	111000	0	23	1	0	1	0	1	0	0	0	0	1	1	7850

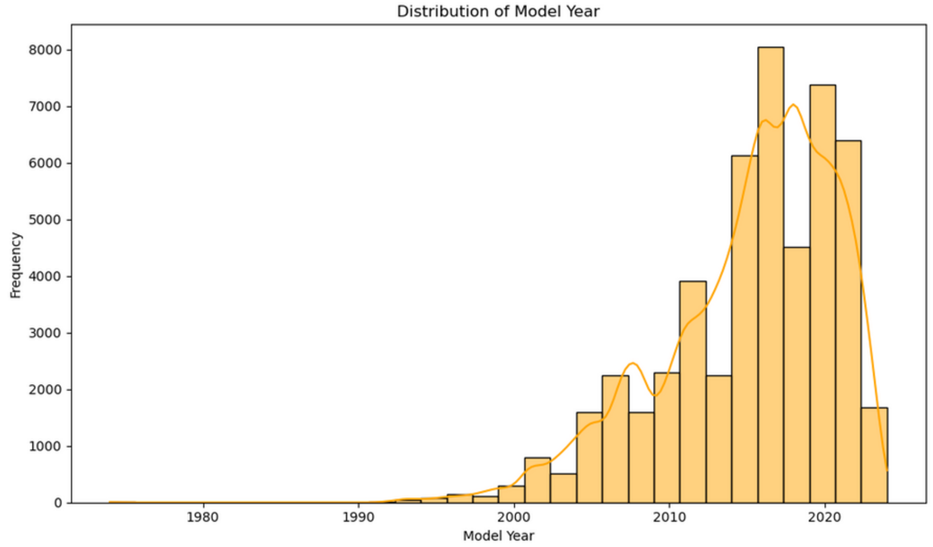
- Brand
- Model.
- Model Year
- Mileage
- Fuel Type
- Engine
- Transmission
- Exterior and Interior Color
- Accident History
- Clean Title

Why difficult?

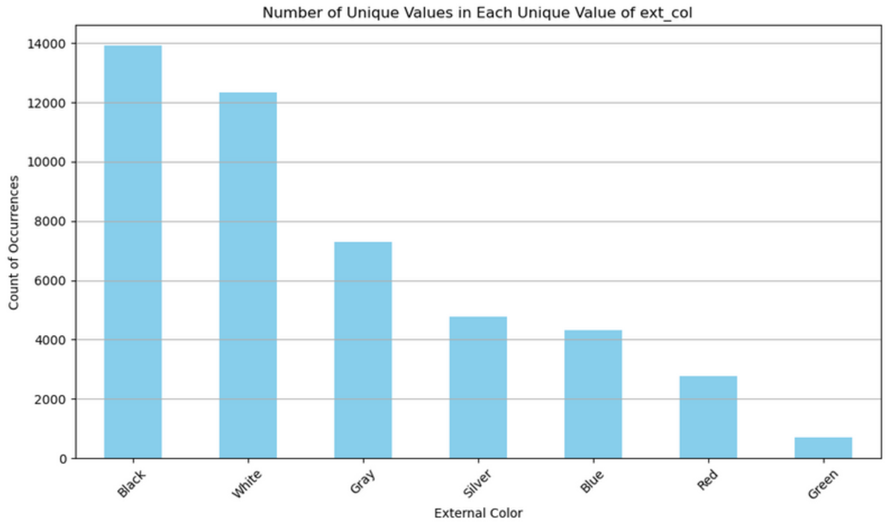
- brand 53
- model 1825
- fuel_type 7
- engine 1059
- ext_col 258
- int_col 123
- transmission 46

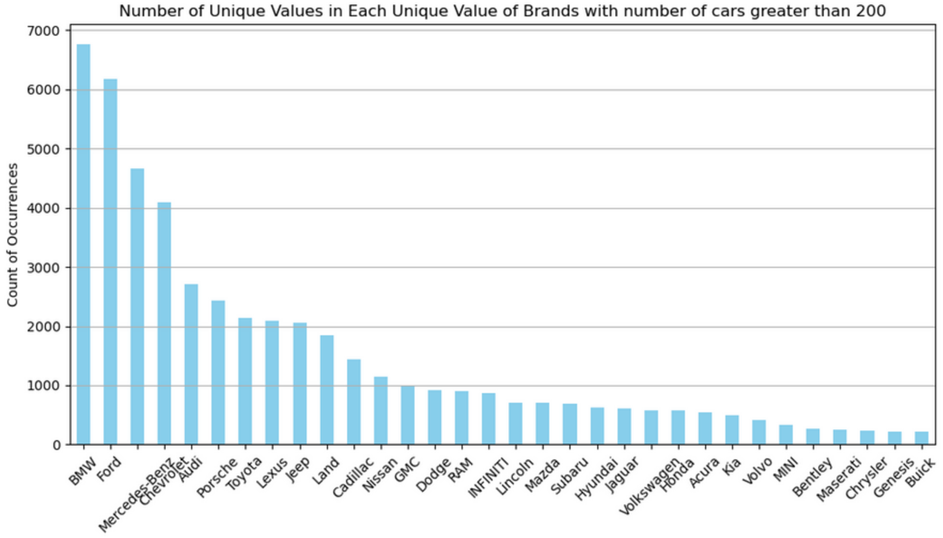


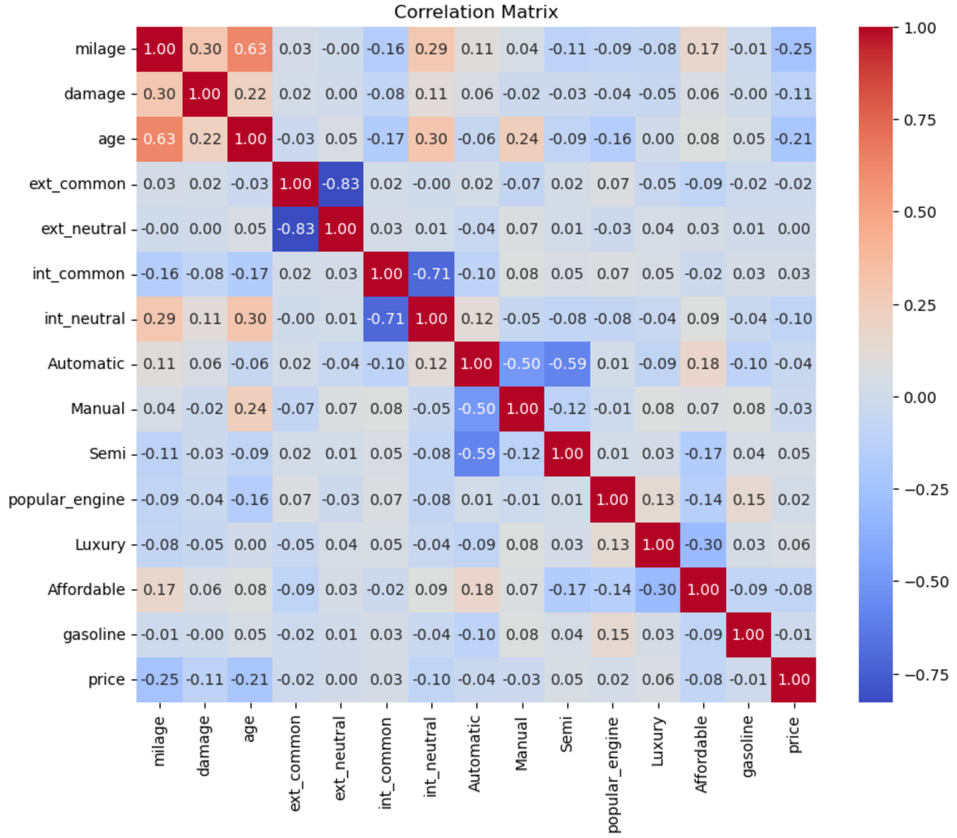




Exploratory Data Analysis







Feature Engineering

- Thought Process: Using Domain Knowledge to reduce the number of categories!

A. Brand: We hypothize that price depends on the brand name and the market segment it caters to. Thus we categorize the brands into three groups:

1. Luxury Cars: Audi, BMW, Rolls-Royce
2. Sports Cars: Porshe, Chevrolet
3. Affordable Cars: Nissan, Hyundai, Suzuki

B. Type of Fuel: We realize that the type of fuel helps the customer approximate their expenses on car after they buy it. Since the type of Fuel is an important sentiment, we categorize them into gasoline and diesel.

C. The type of transmission: We have 46 different types of transmissions which we map to three distinct categories:

1. Automatic: 6-Speed A/T
2. Semi Automatic: Transmission w/Dual Shift Mode
3. Manual: 6-Speed Manual

Feature Engineering

D. Color: Some Colors are inherently more in demand that others. We assume that popular colors like white and black will have a different price ranges than those which are unusual. In fact we know that internal colors have less importance than the outer appearance!

E. Age: Since we have been given the model number. We can simply subtract 2022 to obtain its age

F. Engines: Some engines are Popularly used in cars. While others might be rare. The ones which are rare might have higher prices

CAT Boost

Random Forests

Models Implemented

Support Vector Regression

XG Boost

Hyper Parameter Tuning

We performed grid search for tuning our parameters to enhance the RMSE:
This enabled us to get the following hyper parameters:

1. For Random Forest Regressor:

n_estimators=100, max_features='sqrt', max_depth=12,
min_samples_split=20,
min_samples_leaf=20,
bootstrap=True,
random_state=42,
n_jobs=-1

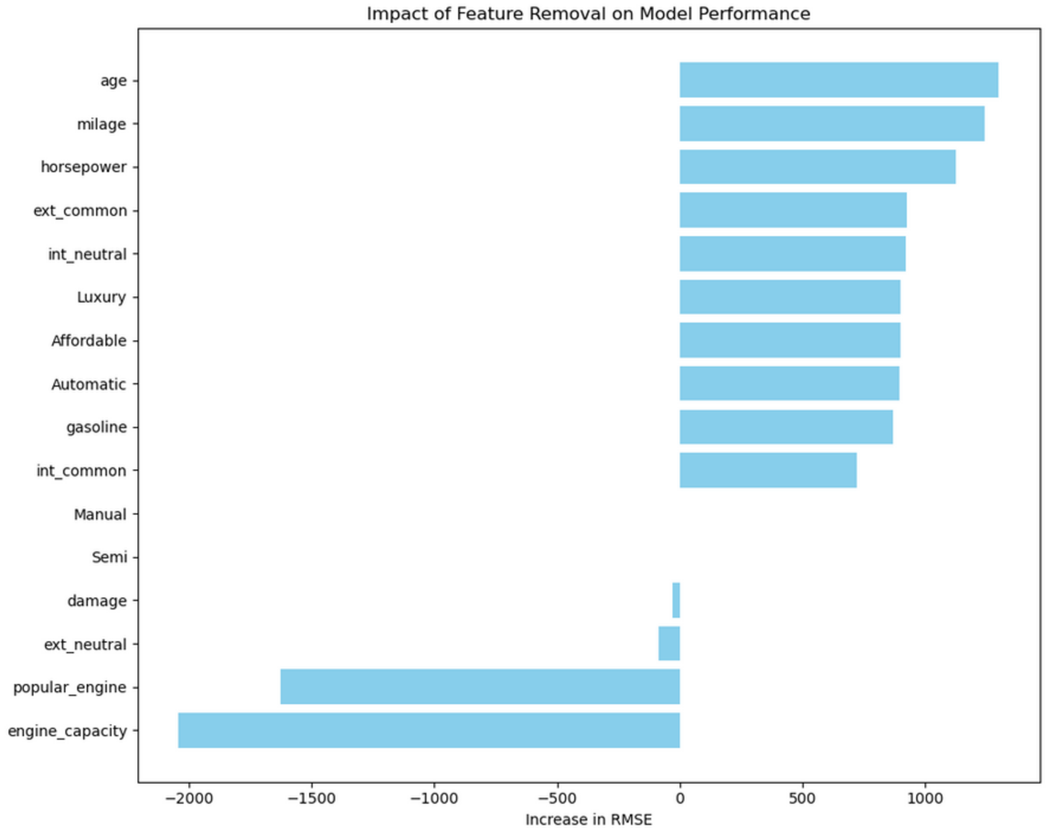
2. Cat Boost:

params = 'iterations': 179, 'depth': 5, 'learning_rate': 0.17338474251588606, 'random_strength': 1.0456923128361446e-06, 'bagging_temperature': 0.719450631818423 2, 'border_count': 247, 'l2_leaf_reg': 18.037542579305022

3. XG Boost:

params = 'n_estimators': 387, 'max_depth': 6, 'learning_rate': 0.022195589115042 1, 'subsample': 0.956049799110561, 'colsample_bytree': 0.6020390043948215, 'gamma': 5.356447556257337e-08, 'min_child_weight': 6, 'reg_alpha': 1.2820640194401 97e-05, 'reg_lambda': 5.1415598885152124e-05

Ablation Studies



The next step to understand if what we are thinking is correct is ablation studies. We remove one feature at a time and measure it impact.

As was easily predictable, mileage and age are amongst the top features to determine the price

Results

RMSE Values thus obtained are as follows on validation set

- CatBoost RMSE: 64,230.44
- XGBoost RMSE: 63,620.39
- Random Forest RMSE: 62,523.12
- SVR: 70,123.08

Thank you!