Data 144 Final Project : Predicting Used Car Prices in India.

Adel Abdalla

INTRODUCTION

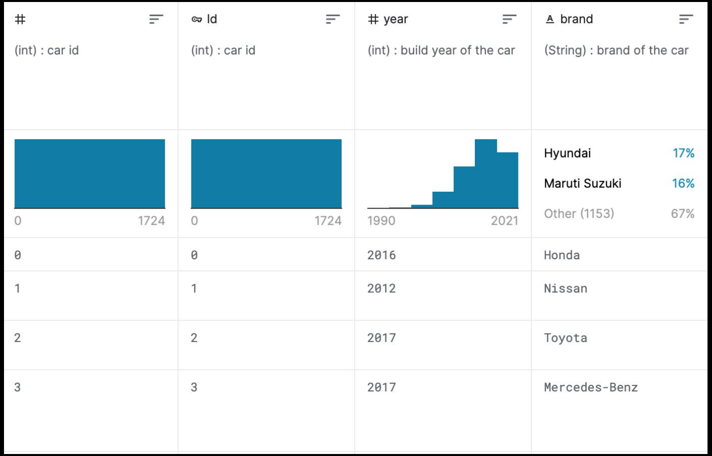
- The Indian used car market is expected to grow in double digits every year from now till 2026.
- By knowing specifications, such as fuel type, age, and distance travelled of the car, one can accurately set a budget and make their searching process more efficient.

RESEARCH QUESTION

What variables predict the price of a used car in India best (with this dataset)?

DATASET

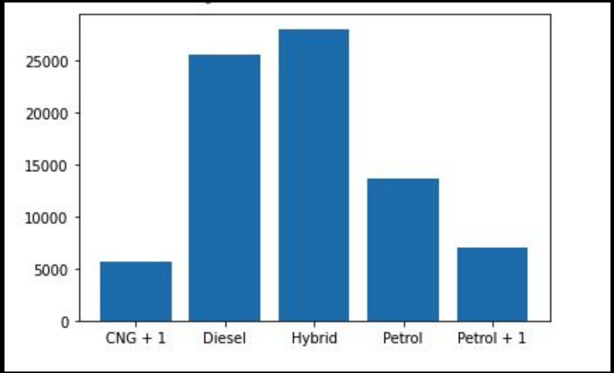
- Link: https://www.kaggle.com/sanjeetsinghnaik/used-car-information
- A private database, the collector scraped various online car re-selling websites and noted the prices of the cars, distance travelled, fuel type, and more.
- We applied log transformation to the distance travelled initially, but it didn't make a difference in the accuracy so we reverted.



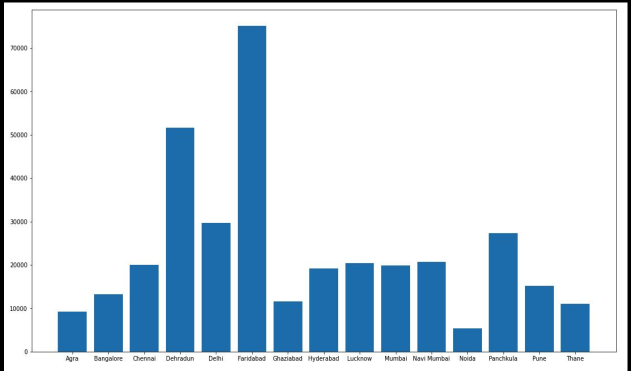
#	Id	# year	brand
(int) : car id	(int) : car id	(int) : build year of the car	(String) : brand of the car
0 1724	0 1724	1990 2021	Hyundai 17% Maruti Suzuki 16% Other (1153) 67%
0	0	2016	Honda
1	1	2012	Nissan
2	2	2017	Toyota
3	3	2017	Mercedes-Benz

EDA

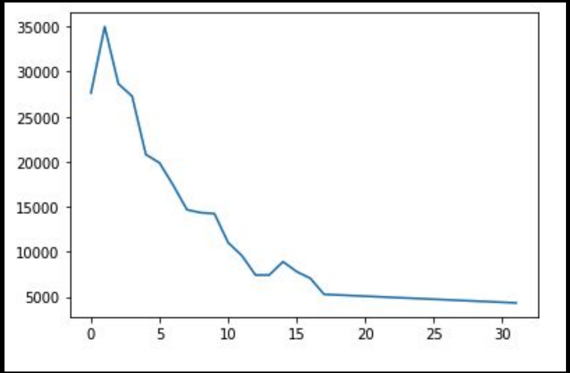
- We began EDA by exploring the average Price of a car against a variety of variables, including: City of sale, Fuel type, and Car Age.



Average Price of Car Per Fuel Type



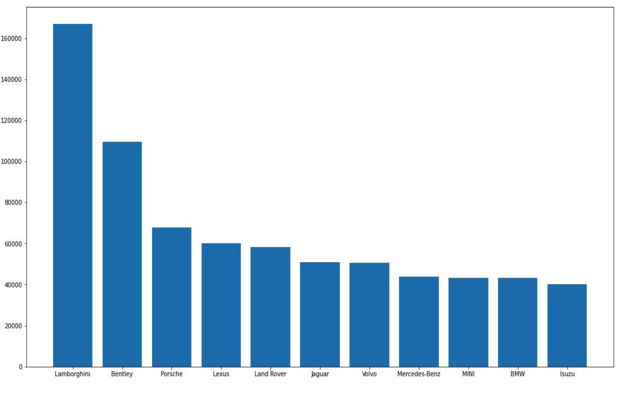
Average Sale Price Per City



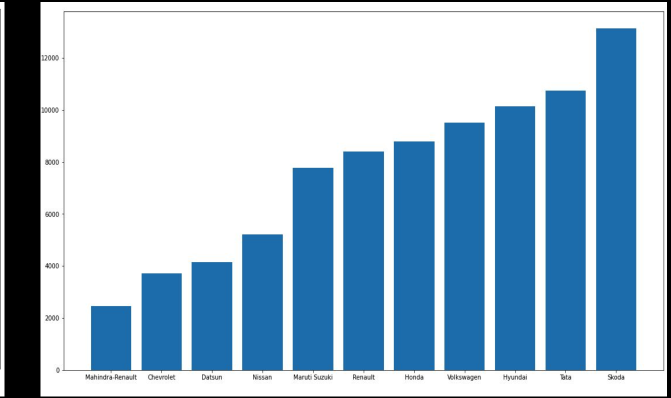
Average Sale Price Per Age of Car

EDA

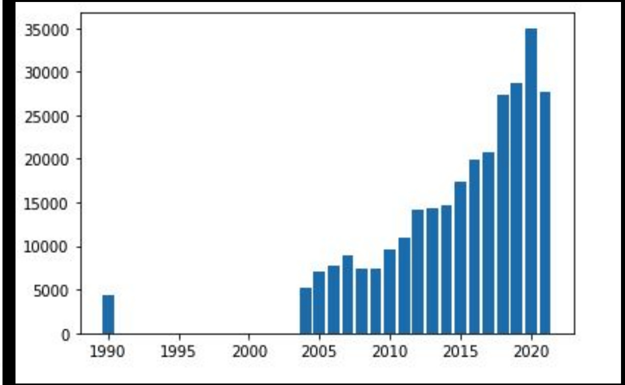
- We also analyzed the Year of the car model, and the actual Brand of the car against its average price.
- Because there were many brands, we could not simply display one bar plot for all brands, thus we had two: one for the top 11 most expensive brands, and one for the top 11 cheapest brands.



Average Price for the 11 most expensive brands



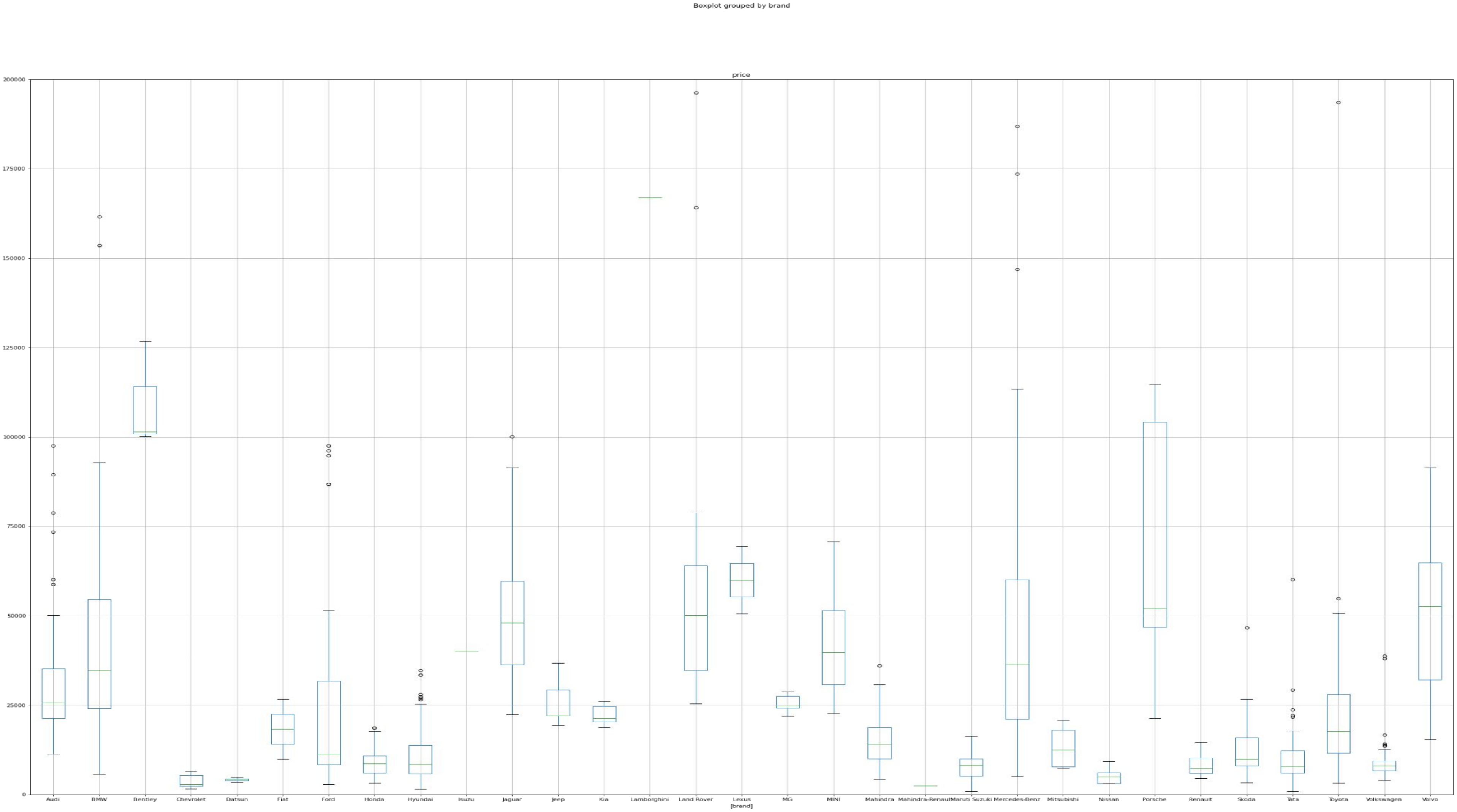
Average Price for the 11 least expensive brands



Average Sale Price of Car per year of its model

EDA

- Lastly, to compare the spread of Pricing across all of the brands, we created a boxplot grouped by brand of car, allowing us to compare the distribution of each car brand.
- Due to its large size, we’ve shown it on the next slide.


Boxplot grouped by brand
price
200000
175000
150000
125000
100000
75000
50000
25000
0
Audi
BMW
Bentley
Chevrolet
Datsun
Fiat
Ford
Honda
Hyundai
Isuzu
Jaguar
Jeep
Kia
Lamborghini
Land Rover
Lexus
MG
MINI
Mahindra
Mahindra-Renault
Maruti Suzuki
Mercedes-Benz
Mitsubishi
Nissan
Porsche
Renault
Skoda
Tata
Toyota
Volkswagen
Volvo
[brand]

Feature Engineering

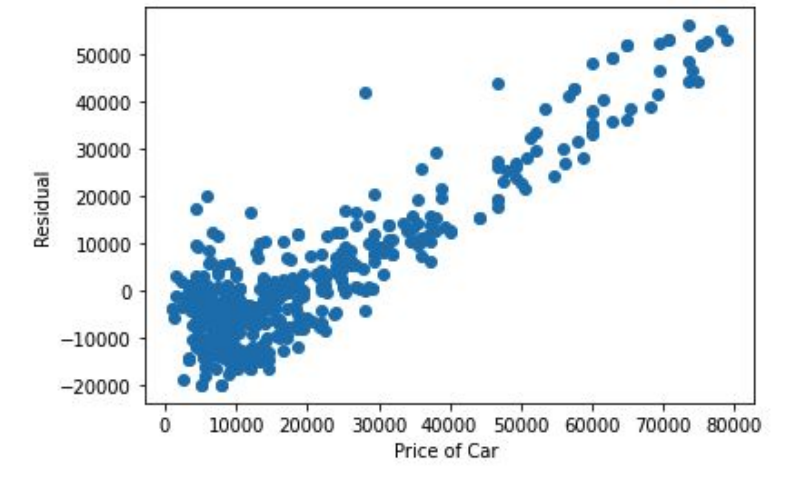
- The results of EDA allowed us to settle on which variables we were to use in our prediction model.
- Note we needed to convert the Price variable from Rupees to USD for interpretability reasons.
- Our data set was formatted well with no Null values, or unpredictable (negative) values in the price and year columns.
- Initially we had: Year, Distance Travelled (kms), Fuel type, and Car Age.
- After trying multiple models, we decided that we needed to include more features.
- We one hot encoded the Brand of the car, City of where the car was sold, and the Model name of the car.
- This increased the accuracy by about 25%!

Machine Learning models.

- Much of our time was spent on finding the best ML algorithm to the pricing of used cars.
- Because our desired predicted value was continuous, we needed to experiment with different regressors.
- We analyzed the performance of our regressors based on their residual plots and accuracy (in which a prediction is accurate if it is within $3000).
- We split the data set, with 570 rows in the test set, and 1155 training set rows.
- While we tested many different models, we'll show you the top 4 performing ones!

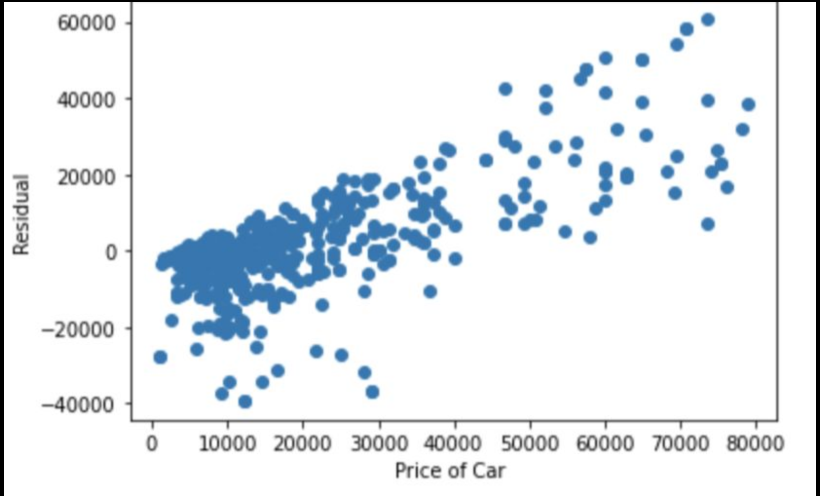
Multiple Linear Regression.

- We began our search for the best regressor with the most familiar to us—Linear regression.
- On the right you'll find the residual plot for our linear regression.
- As you can see, the residuals follow a linear pattern, meaning as the price of the car increases, our regression model performs more and more poorly.
- The accuracy score here was about **23%**.



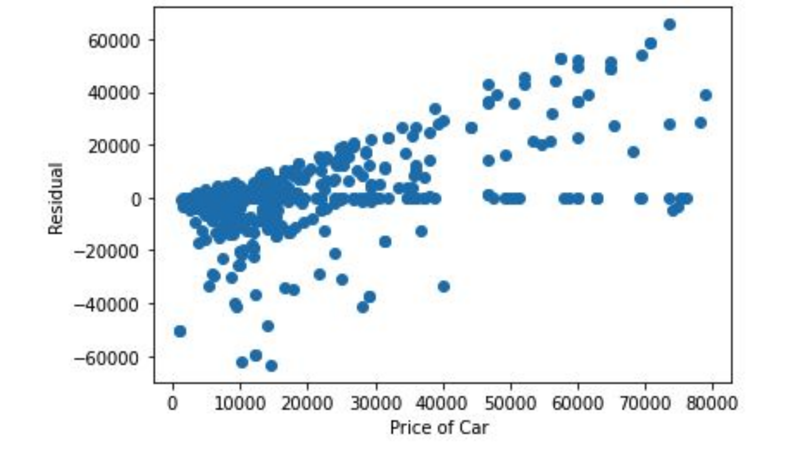
Random Forest

- Next, we tried a Random Forest Regressor.
- The Residuals on the right follow less of a linear pattern than before, as they are more spread out overall.
- However, this regressor still performed poorly for higher car prices.
- The accuracy score here was about **<u>40%</u>**.



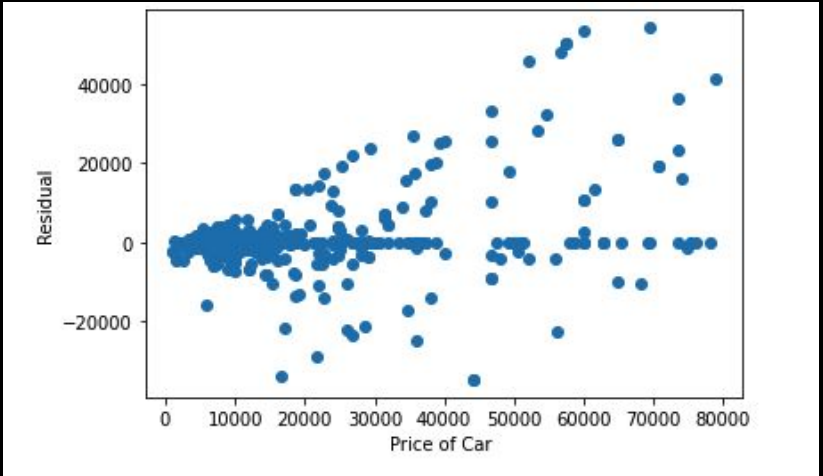
Decision tree (prior to one hot feature engineering).

- We moved onto a decision tree regressor.
- As we can see on the residual plot, this regressor performed much better.
- Not only was the linear pattern minimized, but the residuals at zero were quite populated.
- The decision tree's performance in comparison to all the other regressors proved to be the best so far.
- The accuracy score here was about **<u>51%</u>**.



Decision tree (After one hot feature engineering).

- We continued using a decision tree as the previous iteration of the tree proved to hold high accuracy.
- Having said this, we one hot encoded variables such as model name and brand to boost our accuracy score (as explained previously).
- As you can see on the residual plot to the right, the residuals at zero are heavily populated indicating that many predictions were correct (within the given 3000$ range).
- Accuracy score **76%**.



Conclusion

- Through EDA and Feature engineering, we have concluded that the following variables most impact the Pricing of used cars in India:

Year of Model, Car Age, City of Sale, Fuel Type, Distance Travelled (kms), Brand, and Model Name

Note that the Brand and Model Name are highlighted, as they had the largest impact on accuracy(25% increase) once added to the regression model.