USED CAR PRICE PREDICTION REGRESSION TECHNIQUES

BY GROUP 12
CHINMAY BHAT
RAM WARUTKAR
SAKSHI JAIN

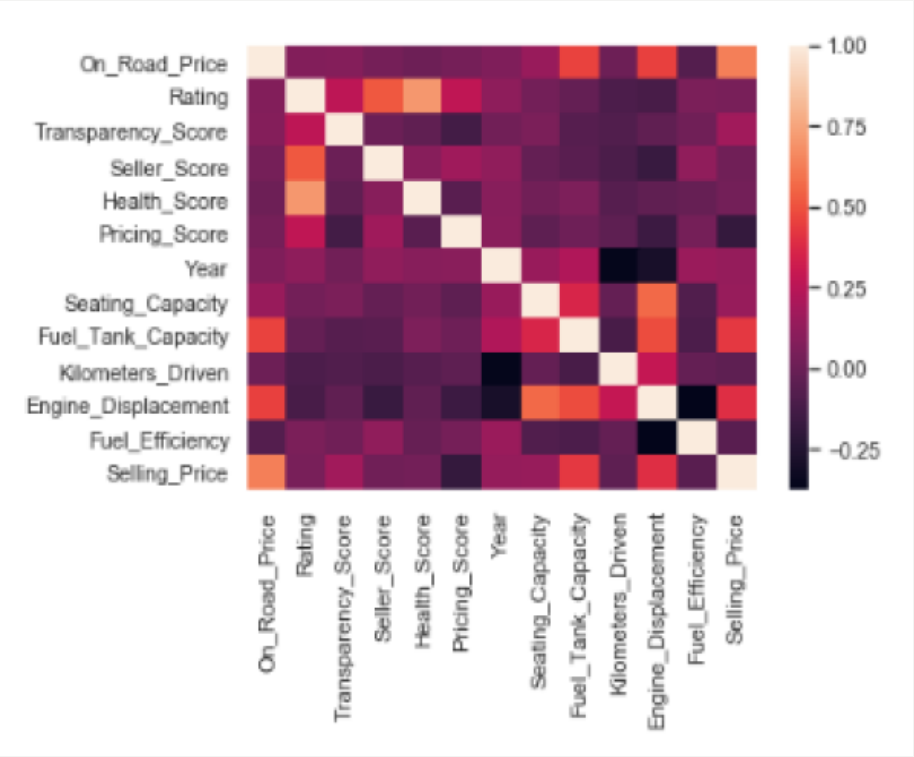
INTRODUCTION

- Sale of a used car market is growing rapidly
- Estimating the price of a vehicle by building machine learning algorithm based on certain important factors can automate this process and it can also speed up the process
- To automate this process we have tried to build an algorithm that will predict the price of used car based on certain important features

ANALYSIS OF DATA

- For this project we have scrapped the data from droom.in
- The dataset contains 597 records and 27 columns, out of which 14 columns are categorical and 13 are numeric
- The dataset contains 351 null values- which were filled before applying any machine learning algorithm
- The categorical features were encoded in order to apply macine learning algorithm

CORRELATION MATRIX



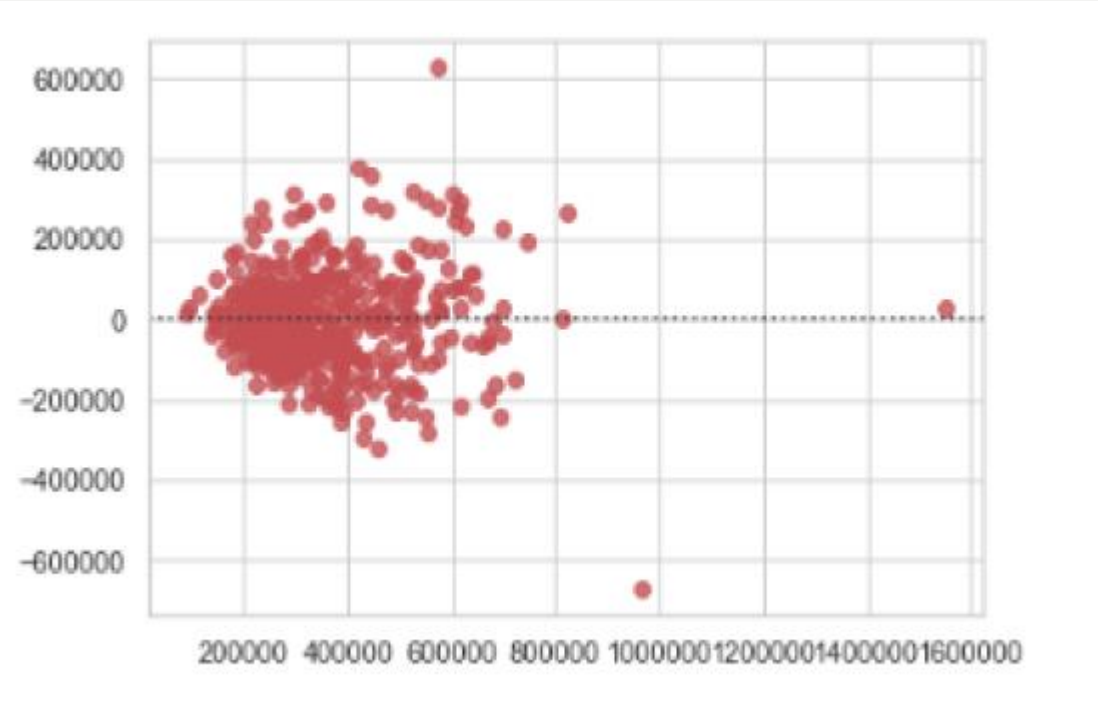
VARIANCE INFLATION FACTOR

	VIF Factor	features
0	22.662306	On_Road_Price
1	662.381627	Rating
2	101.576398	Transparency_Score
3	36.542658	Seller_Score
4	61.791701	Health_Score
5	50.259027	Pricing_Score
6	6.936224	Make
7	3710.573985	Year
8	2054.911625	Seating_Capacity
9	183.311208	Fuel_Tank_Capacity
10	3.407388	Kilometers_Driven
11	224.364644	Engine_Displacement
12	52.974772	Fuel_Efficiency
13	5.012817	Petrol
14	1.037067	Petrol + CNG
15	31.648849	Automatic
16	414.259298	Manual
17	327.039362	Hatchback
18	2.600399	MUV
19	152.686514	SUV
20	28.771236	Sedan
21	1.204940	Leather
22	1.306401	Leatherette
23	1.040072	Other
24	341.787776	First Owner
25	5.230322	Fourth Owner
26	101.117255	Second Owner
27	14.491587	Third Owner
28	57.160646	Dealer

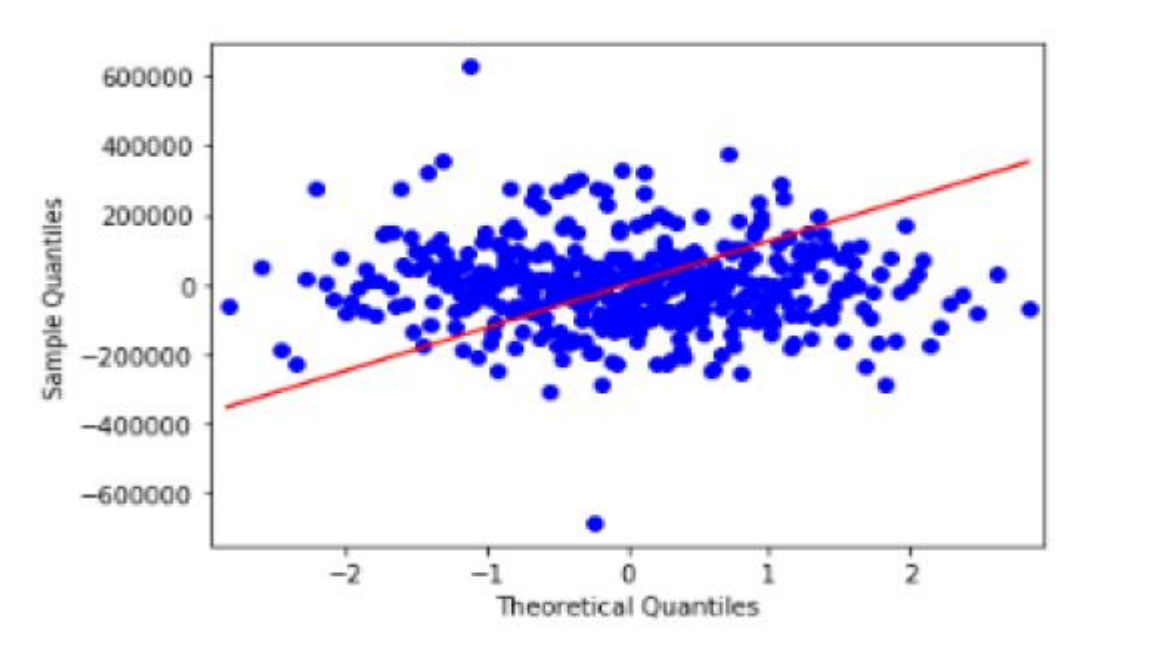
TECHNIQUES APPLIED

- Multiple Linear Regression
- Decision Tree
- Random Forest
- Boosting
- Support Vector Machines
- K Neighbor Regression

MULTIPLE LINEAR REGRESSION



The residual are distributed randomly.



So the graph suggest that the residuals are not normally distributed

COMPARISION OF ALL THE MODELS

Algorithm	R- Square	Training RMSE	Testing RMSE
Multiple Linear Regression	59 %	123425	125658
Regression after removing the features	57 %	123350	126329
Decision Tree	54 %	129416	167871
Random Forest	79 %	88353	152659
Boosting	65 %	113401	159081
Support Vector Machine (kernel linear)	56 %	0.08	0.08
Support Vector Machine- Kernel rbf	56%	0.08 (On scaled data)	0.08 (On scaled data)
K Neighbor Regression	61 %	119713	160414

CONCLUSION

- Ideally there is no model which can predict the selling price accurately. The random forest model is working quite good on the training dataset but not on testing dataset.
- We cannot opt for linear regression because the residuals are not normally distributed which suggest that the error is not constant across the full range of the data.
- All the models are overfitting the data except for support vector regression which has same RMSE for training as well as for testing dataset.
- For the SVR the data was scaled in range of 0 to 1, in according to that the RMSE of 0.08 is high. If this error is acceptable to the seller party or the buyer party. Then they can opt for this model at a cost of underestimated or overestimated selling price by 80,000.


Thank you