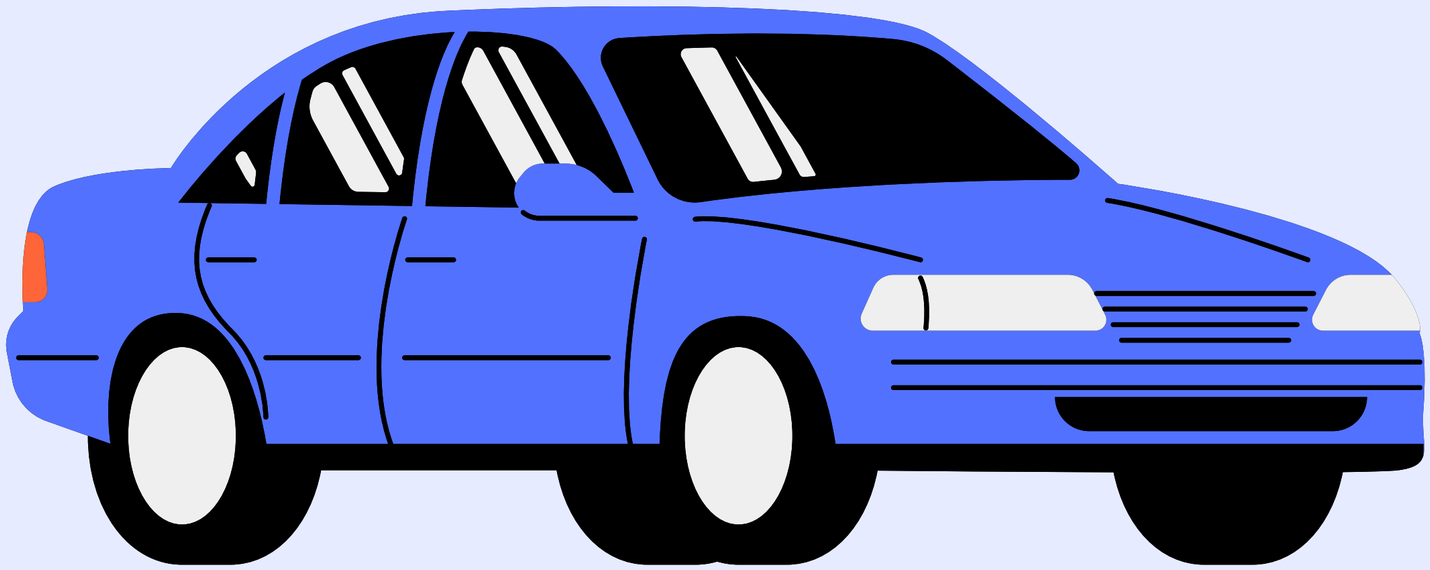
WHAT DRIVES CAR PRICES?

Can we accurately predict a car's price from its features?
What features of the car are most influential on its cost?
Which model most effectively predicts the price?



ANDREW GUO

THE DATA

OVERVIEW

Source: Vehicle Sales Data (Kaggle)

Observations: 558836 rows

Condition: condition of the vehicle on scale of 1-100

FEATURES

Categories: 16 variables cut down to 9 key predictors

Numeric: Odometer, Year

Categorical: Body, Condition, Make, Color, State, Interior, Transmission

PREPROCESSING

Body types simplified to 10 generic styles

Removed estimated_value category for more robust data

Test size of 25%

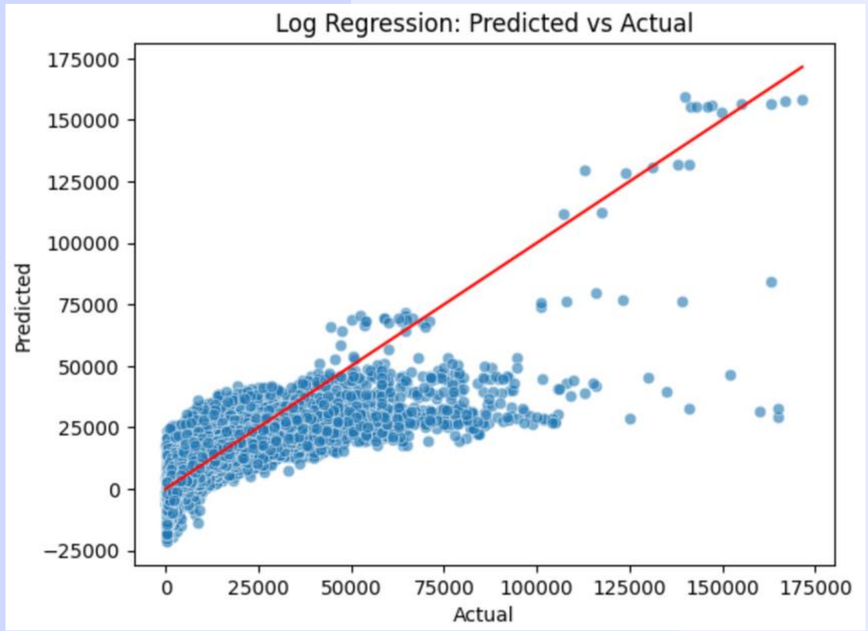
	year	make	body	transmission	state	condition	odometer	color	interior	estimated_value	sellingprice
0	2015	Kia	SUV	automatic	ca	5.0	16639.0	white	black	20500.0	21500.0
1	2015	Kia	SUV	automatic	ca	5.0	9393.0	white	beige	20800.0	21500.0
2	2014	BMW	Sedan	automatic	ca	45.0	1331.0	gray	black	31900.0	30000.0
3	2015	Volvo	Sedan	automatic	ca	41.0	14282.0	white	black	27500.0	27750.0
4	2014	BMW	Sedan	automatic	ca	43.0	2641.0	gray	black	66000.0	67000.0

https://www.kaggle.com/datasets/syedanwarafridi/vehicle-sales-data

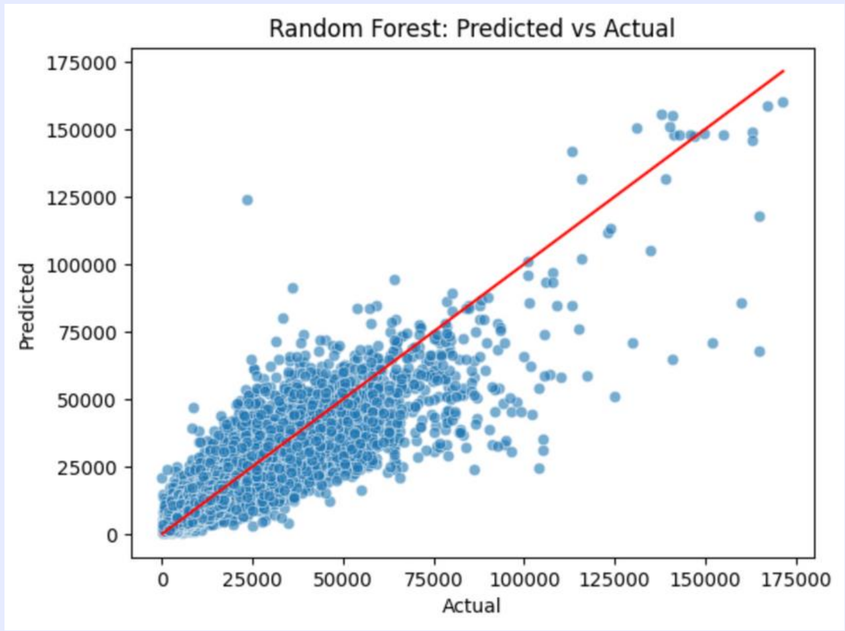
MODELS

LOG REGRESSION
RMSE: 5443.38
R2: 0.68

Some negative predictions, accurately follows linear trends but struggles with non-linear relationships



RANDOM FOREST
RMSE: 3956.30
R2: 0.83
Mean residual: -21.18
Std of residuals: 3956.26
MAE: 2362.32



SUMMARY

Random Forest outperforms Linear Regression

- Lower RMSE and R2
- Both models have relatively high RMSE given that the vast majority of cars sell for under 50k

Influential Features

- Odometer, make, and body type were the most powerful predictors
- Transmission, interior material, and color were the least

odometer	0.407009
make	0.227538
body	0.124598
year	0.102823
condition	0.046594
state	0.042194
color	0.024956
interior	0.021544
transmission	0.002744

Residuals (Random Forest Only)

- Mean: slightly underestimates price but is very close to 0
- Standard Deviation: very high, almost 10% error per prediction
- MAE: lower than STD, not as affected by outliers

Conclusion: Random Forest was very effective at predicting car selling prices but struggled with outliers at extremely high or low costs