Cars 24 Problem Overview

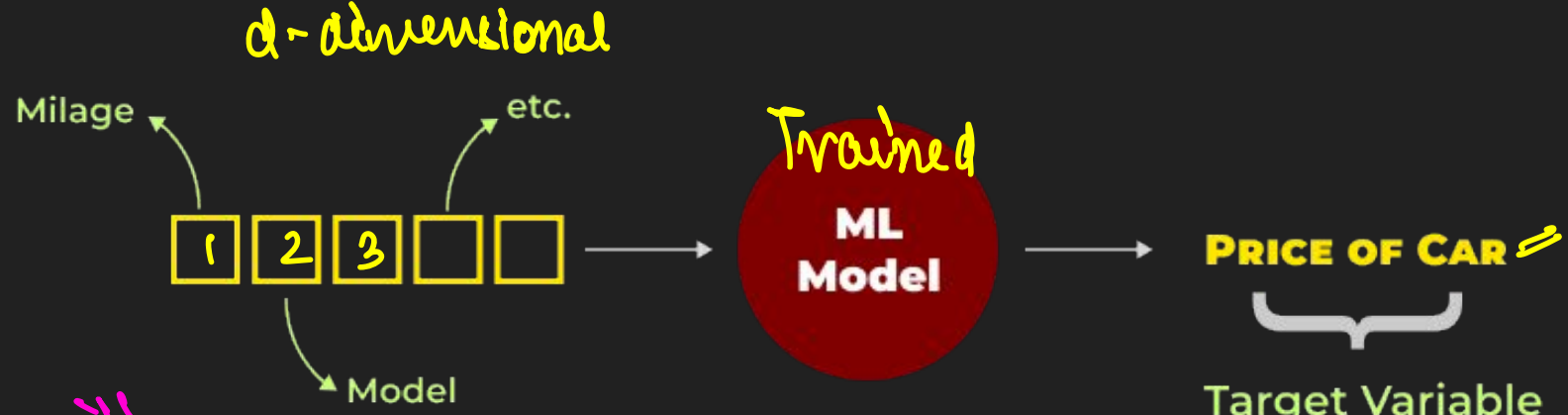
Data scientist at **Cars 24** → Sells pre-owned cars

→ To automate pricing the old car

Features :- Make, Model, Mileage, Odometer Reading, Service History, Year, AFM

PREDICTORS

Train an ML Mode such that



Model will always do correct predictions??

y_i → Real price of the car

$\hat{y}_i$ → Predicted price of the car

Target

Look at the Data . .

Brand

Specific Model

	selling_price	year	km_driven	mileage	engine	max_power	age	make	model	Individual	Trustmark Dealer
0	1.20	2012.0	120000	19.70	796.0	46.30	11.0	MARUTI	ALTO STD	1	0
1	5.50	2016.0	20000	18.90	1197.0	82.00	7.0	HYUNDAI	GRAND I10 ASTA	1	0
2	2.15	2010.0	60000	17.00	1197.0	80.00	13.0	HYUNDAI	I20 ASTA	1	0
3	2.26	2012.0	37000	20.92	998.0	67.10	11.0	MARUTI	ALTO K10 2010-2014 VXI	1	0
4	5.70	2015.0	30000	22.77	1498.0	98.59	8.0	FORD	ECOSPORT 2015-2021 1.5 TDCI TITANIUM BSIV	0	0

Very different range

Lots of categories

Labels present in data - SUPERVISED

Nature of label is continuous - REGRESSION

Need1) Lin Regression requires all features to be numerical

Categorical ⟶ Numerical

1) One-hot encoding (OHE)
2) Binning - (for ordinal data)
3) Target Variable Encoding

Need2) For good training, all the features should be somewhat similar range.

Mileage - 1 to 20
Odometer - 10K to 2L

Scale the data to bring in similar range

Extra Reference Material

EDA for Car24 Dataset

① Outlier Treatment
②