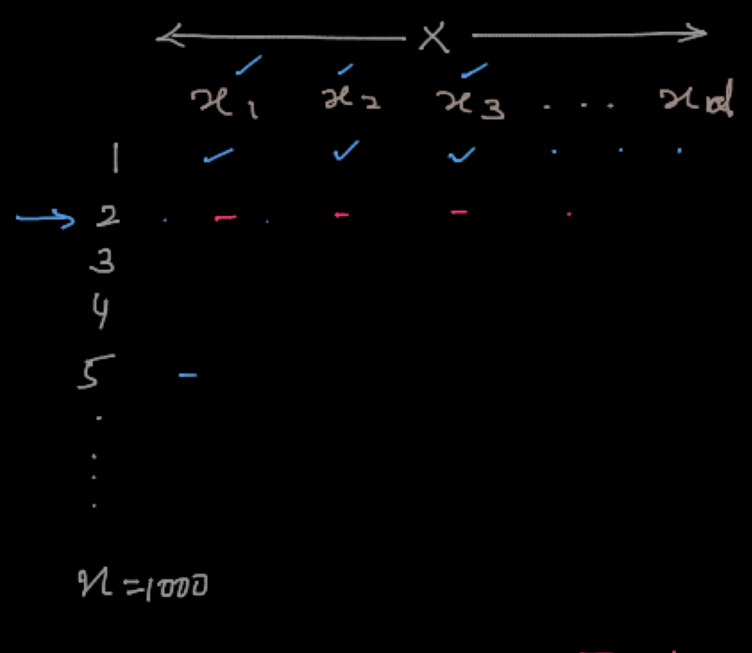
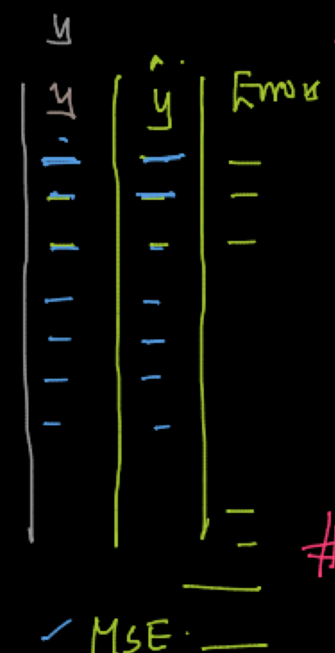
Model Validation Techniques.





Create the model

model. = LinearRegression( ).

Train the model

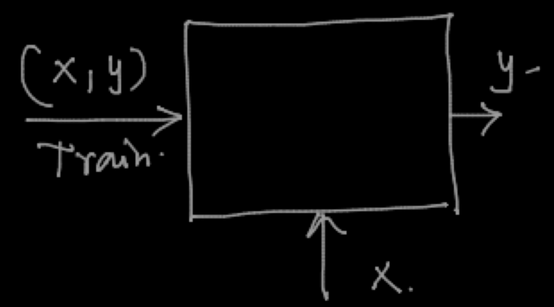
model.fit(X, y)

— Estimate the model parameters (β's)

Predict the outcomes.

model.predict(X).

↳ $\hat{y}$



✓ MSE.

✓ RMSE = $\sqrt{MSE}$

How well the model performs on 'UNSEEN DATA'.

	x	y	ŷ	
→	1200	50	48	2
→	1500	60	61	1 80-/-
→	1750	65	63	2
	1800	70	89	19 L →
	2000	90	70	20 L

1200- 50

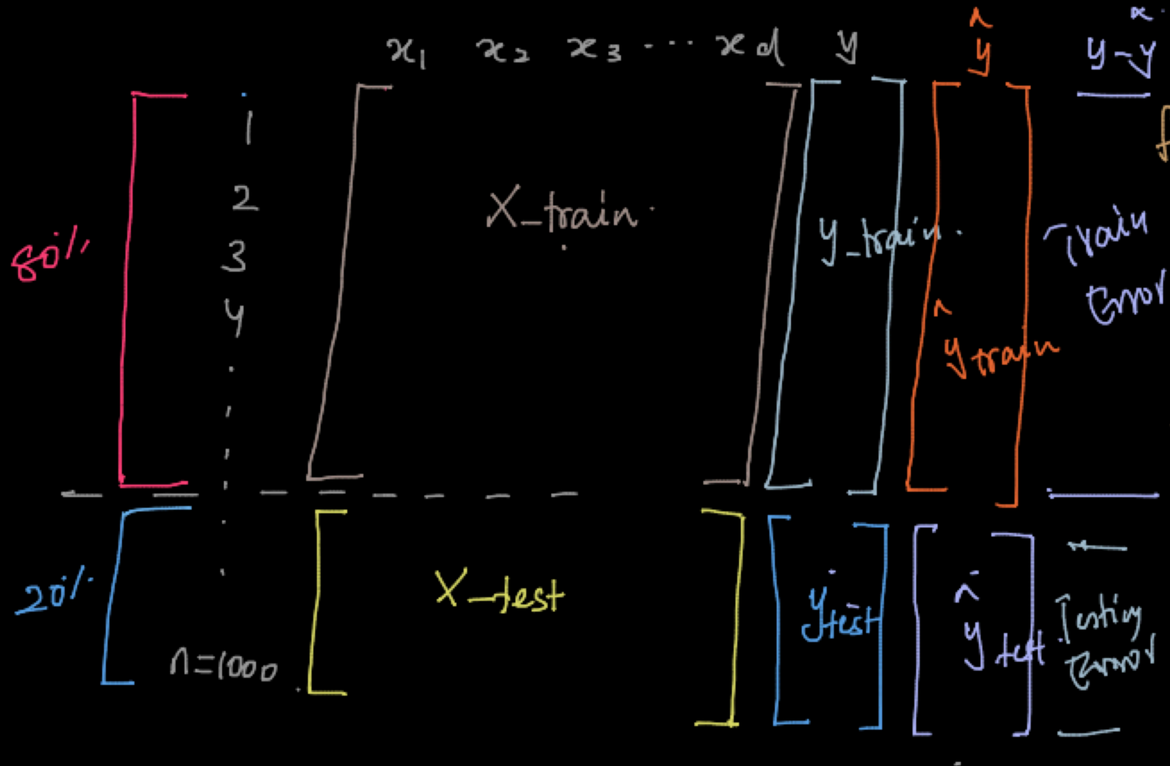
3

2

Validate the Model -

```
X_train, y_train, X_test, y_test = train_test_split(X, y, 0.20)
```

↳ Python Code -




```
# Create the Model
final_model = LinearRegression()

# Train the model
final_model.fit(X_train, y_train)

# Predict for training data (Seen data)
final_model.predict(X_train) → ŷ train

# Predict for unseen data.
final_model.predict(X_test)
↳ ŷ test
```

Training Error ↓
Testing Error ↓

Training Error ↓
Testing Error ↑
} → overfitting

Training Error ↑
Testing Error ↑
} underfitting

Errors → Regression

Ideal

- Low Training Error
- Low Testing Error

Overfitting

→ Low training Error ←
→ High testing Error

Underfitting

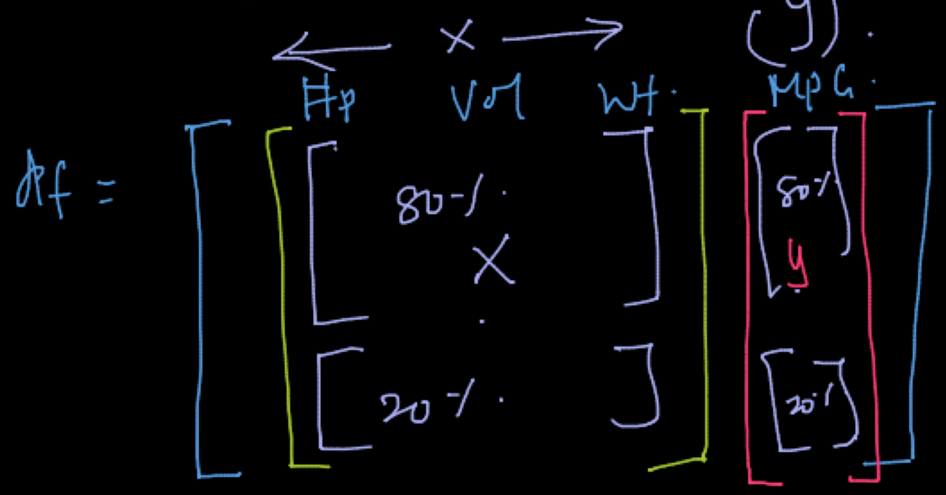
High Training Error
High Testing Error.

Accuracy → Classification

High traing Acc
High testing Acc.

High training Acc
→ Low testing Acc.

Low training Acc
Low testing Acc



df['MPG'] = y

Regression → Metric - Errors - MSE, RMSE, MAE

Classification → Accuracy ;

y	$\hat{y}$	
Yes	Yes	✓
Yes	No	✗
No	No	✓
No	Yes	✗

$\frac{2}{4} \Rightarrow 50\%$

0.20

80 - 20 ✓ ✓

75 - 25

66 - 33 :

$X\_train, y\_train, X\_test, y\_test$ = train-test-split (X, y, 20%)

↳ % of testing data

	Sqft	Bed	price	
1	1000	1	53	
2	1500	2	65	→ testing
3	1800	3	73	
4	1400	2	55	→
5	1700	1	67	

X_test: 1500 2 — 65 y_{test}

1400 2 — 55

	x_1	x_2	x_3	..	y
1					
2					
3					

20 · D_1

20 · D_2

20 · D_3

$n = 100$ ←

K-Fold Validation.

$K = 10$ Folds

k – hyperparameter.

→ $K = 5$; 20% → testing.

$n = 100$; $k = 5$;

$D_1 = 20$.

← Training → ← Testing →.

→ D_1 D_2 D_3 D_4 D_5 → MSE_1

→ D_1 D_2 D_3 D_5 D_4 · MSE_2

D_1 D_2 D_4 D_5 D_3 MSE_3

D_1 D_3 D_4 D_5 D_2 MSE_4.

D_2 D_3 D_4 D_5 D_1 MSE_5.

→ MSE

10:20 AM.

LOOV

— Leave one out Validation

n is Very Small →

	x_1	x_2	x_3	y
1				
2				
3				
4				
5				
6				
7				
8				
→ 9				$y - \hat{y} = E_2$
10				? y → Error

Training → 1 to 9 — Testing → 10 → E_1

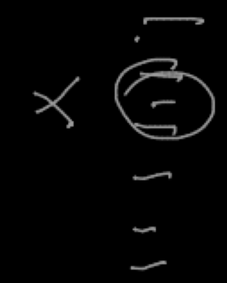
1-8, 10 → 9th → E_2

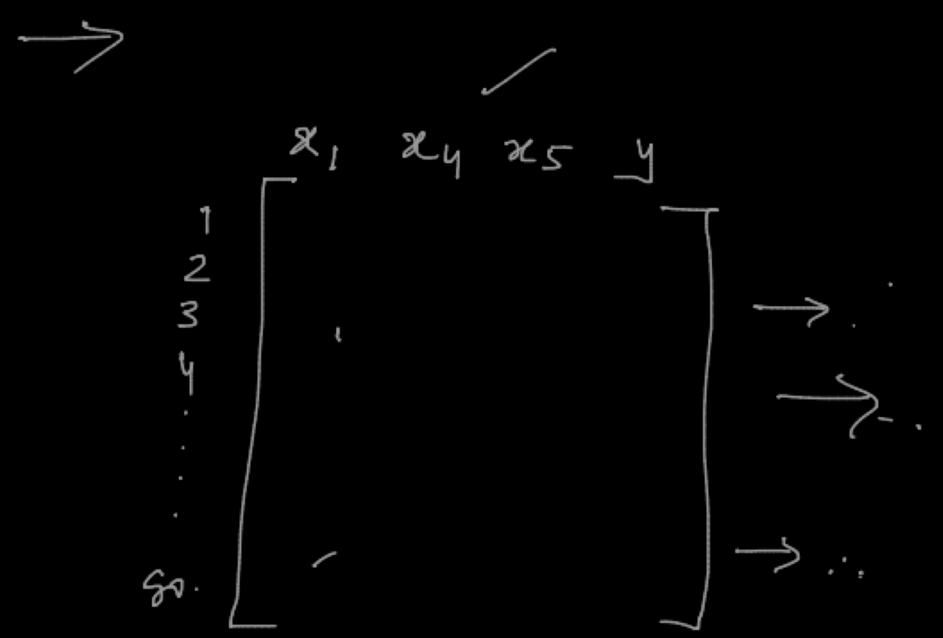
1-7, 9-10 → 8th

1-6, 8, 10 → 7th

Feature Engineering

x_1 x_2 x_3 y





deployment → O

Data drift

50 48.3

20% 80% → y.

10% 90% → overfit

100 50% — → (=) → underfitting

50% —