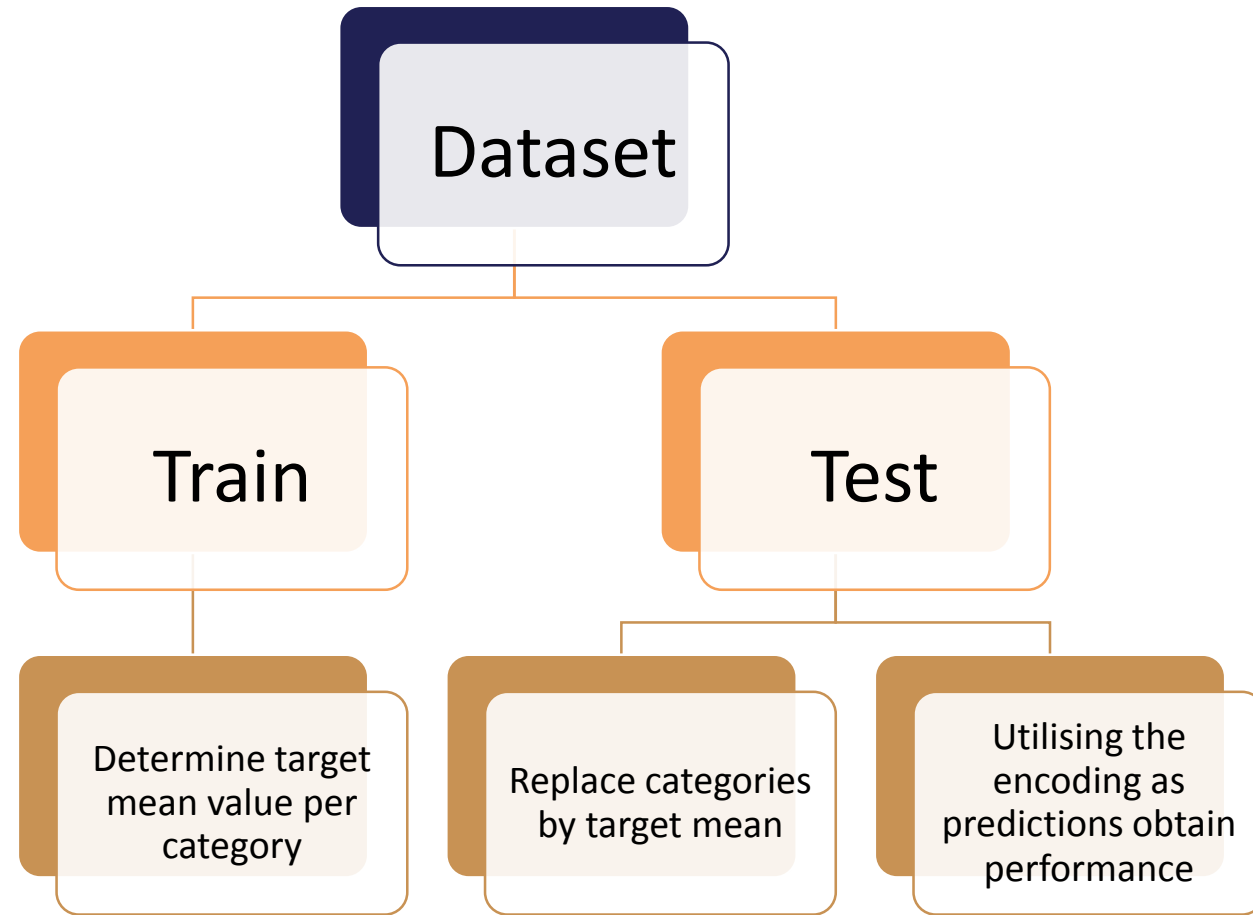


# KDD 2009: Target mean encoding

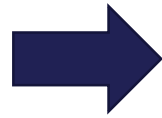
# Method description



# Mean encoding: example

## Train Set

Colour	Target
Red	1
Red	0
Red	1
Blue	1
Blue	1
Green	0
Green	0
Yellow	1
Yellow	0



Colour
0.66
0.66
0.66
1
1
0
0
0.5
0.5

{

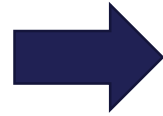
Red: 0.66,  
Blue: 1,  
Green: 0  
Yellow: 0.5

}

# Mean encoding: example

## Test Set

Colour	Target
Red	1
Red	0
Blue	1
Blue	1
Blue	0
Green	0
Yellow	1



Colour
0.66
0.66
1
1
1
0
0.5

{


Red: 0.66,  
Blue: 1,  
Green: 0  
Yellow: 0.5

}

# Mean encoding: example

Test Set

Colour	Target
Red	1
Red	0
Blue	1
Blue	1
Blue	0
Green	0
Yellow	1



Colour
0.66
0.66
1
1
1
0
0.5



Performance Metric

{

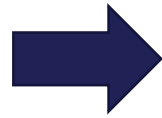
Red: 0.66,  
Blue: 1,  
Green: 0  
Yellow: 0.5

}

# Mean encoding: Numerical variables

## Train Set

Price	Bins	Target
1	[1-5]	1
2	[1-5]	0
3	[1-5]	1
6	[6-10]	1
6	[6-10]	1
7	[6-10]	0
8	[6-10]	0
12	[11-15]	1
14	[11-15]	1

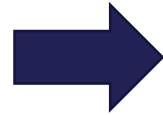


Bins
0.66
0.66
0.66
0.5
0.5
0.5
0.5
1
1

{  
[1-5]: 0.66,  
[6-10]: 0.5,  
[11-15]: 1  
}

# Mean encoding: example

Price	Bins	Target
1	[1-5]	1
3	[1-5]	0
4	[1-5]	1
8	[6-10]	1
8	[6-10]	0
13	[11-15]	0
14	[11-15]	1



Price
0.66
0.66
0.66
0.5
0.5
1
1



Performance Metric

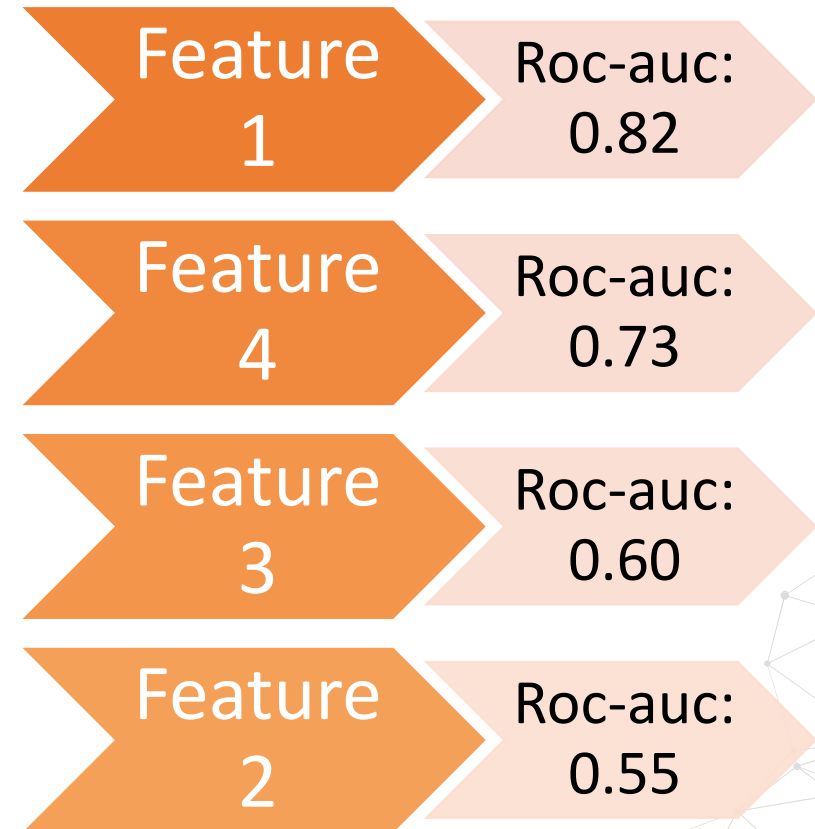
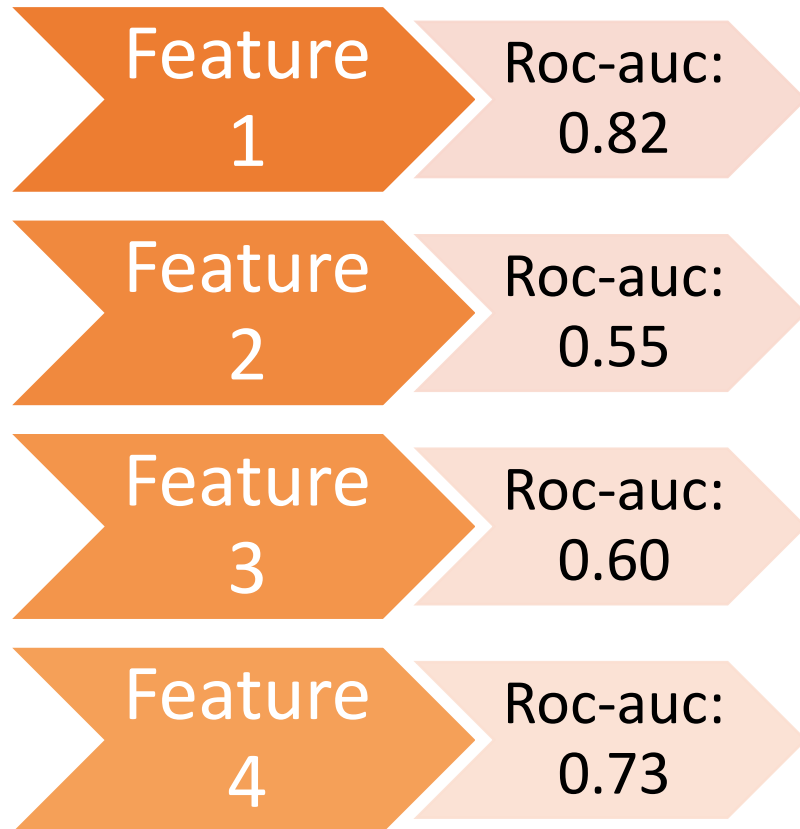
Test Set

{

[1-5]: 0.66,  
[6-10]: 0.5,  
[11-15]: 1

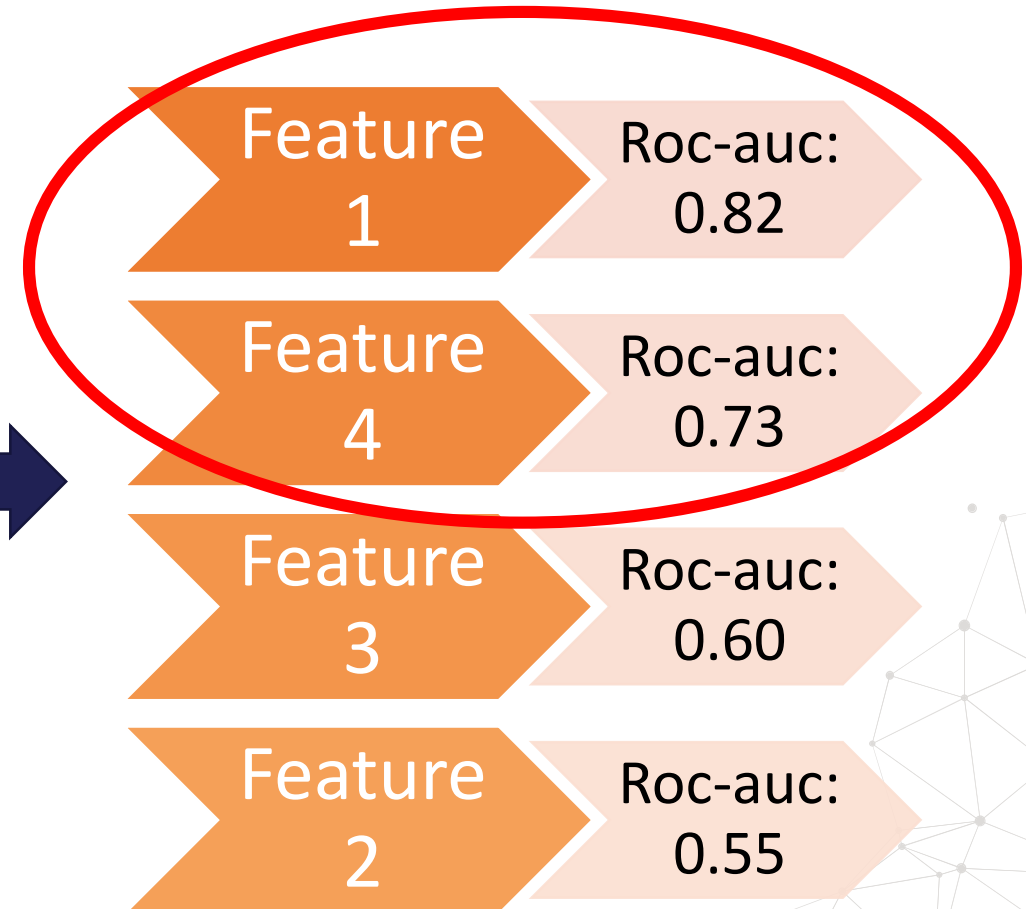
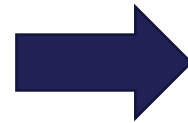
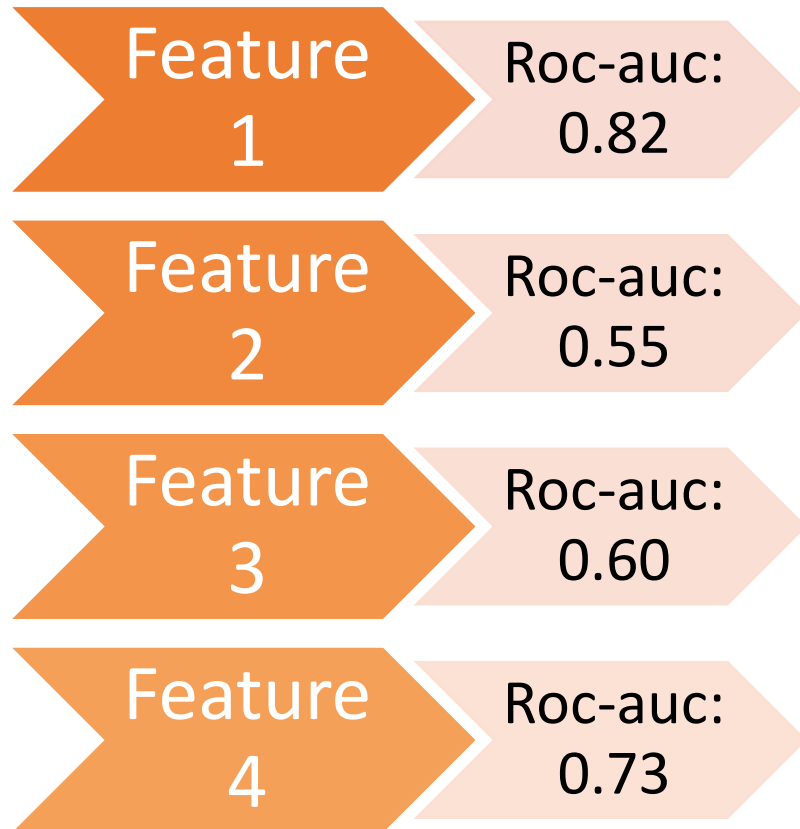
}

# Rank features





# Select features



# Performance metric



- We can use any performance metric we like
  - ✓ Roc-auc, accuracy, Precision, Recall, etc
  - ✓ MSE, RMSE, R2, etc
- Different metrics may lead to different selected features

# THANK YOU

[www.trainindata.com](http://www.trainindata.com)