Gift This Course

A-Z Machine Learning using Azure Machine Learning (AzureML)

Hands on AzureML: From Azure Machine Learning Introduction to Advance Machine Learning Algorithms. No Coding Required.

BEST SELLER ★★★★★ 4.3 (215 ratings) 1,597 students enrolled

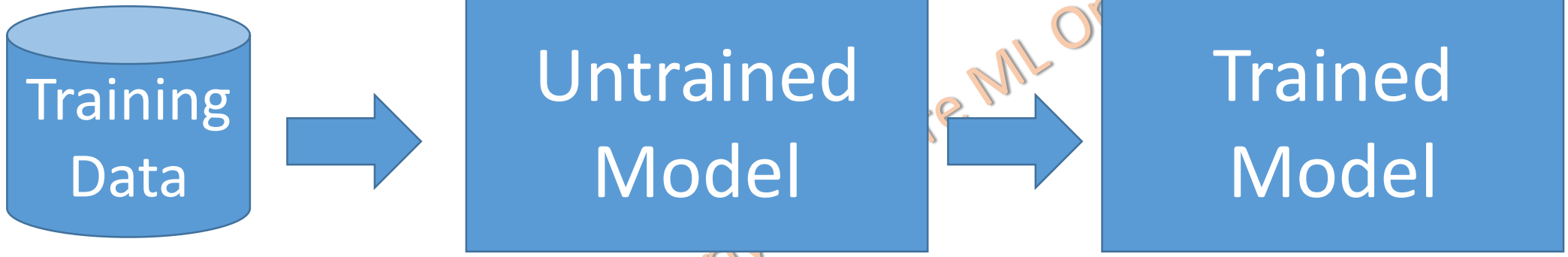
Created by Jitesh Khurkhuriya Last updated 3/2018 English English

Feature Selection

Loan Approval Data

- **Personal Information** – Name, Gender, Marital Status, Age etc
- **Financial Information** – Monthly Income, Additional Income, Credit History, Any other loan
- **Demographic Information** – Address, Race, Address type, Type of customer
- **Loan Provider Information** – Product type, existing customer, Group based on past behaviour
- **Application Details** – Date of Application, Advisor information, Campaign Information
- **Educational Details** – Level of education, Type of education, Institute
- **Employment Information** – Type of employment, employer category, Designation, Years of experience
- **Coapplicant Details** – all the above

Importance of Feature Selection



Garbage

Garbage

Why to Use Feature Selection?

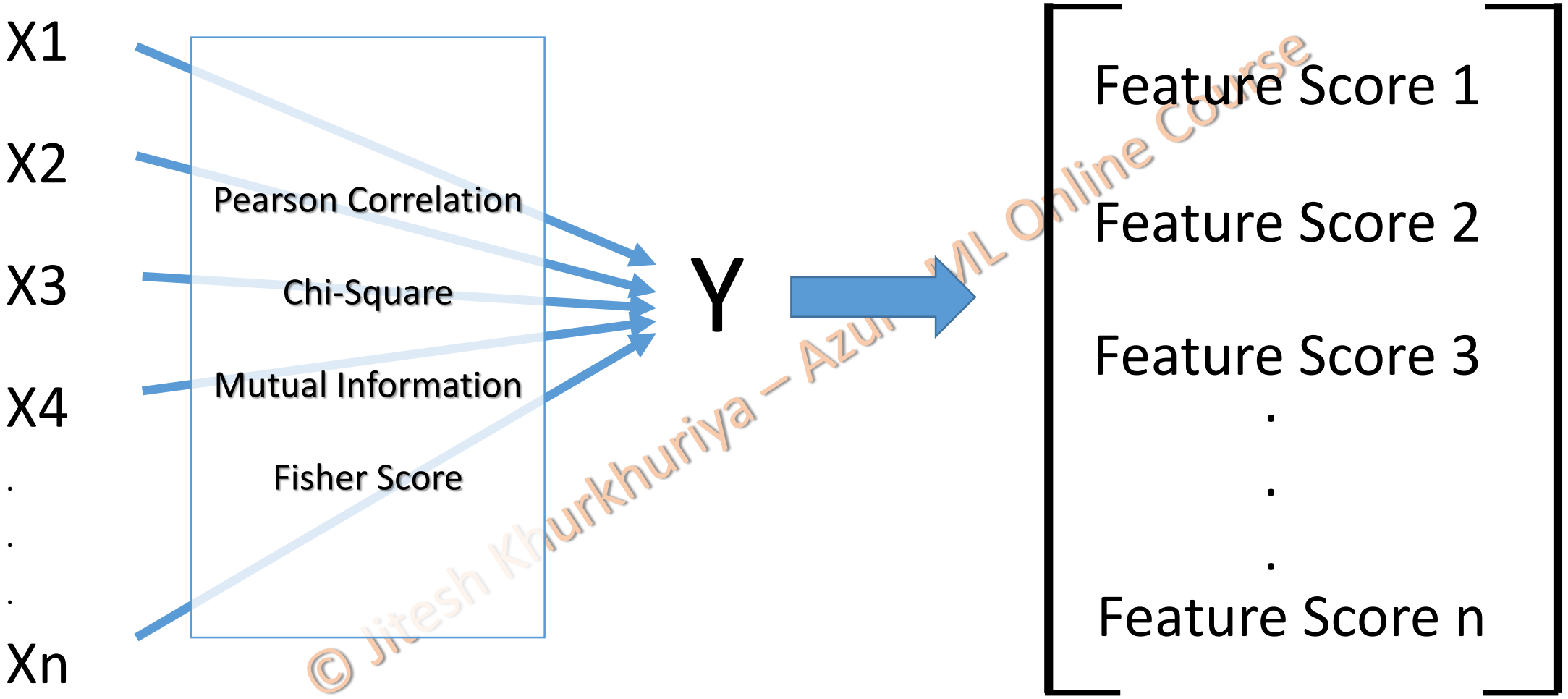
- Simplification of models to make them easier to interpret by researchers/users
- Shorter training times
- Improves Accuracy
- To avoid the curse of dimensionality,
- Enhanced generalization by reducing overfitting

Methods of Feature Selection

- **Filter Based Methods** – Uses Correlection with the outcome variables
- **Wrapper Methods** – Uses a subset of features by selecting the best set
- **Embedded Methods**

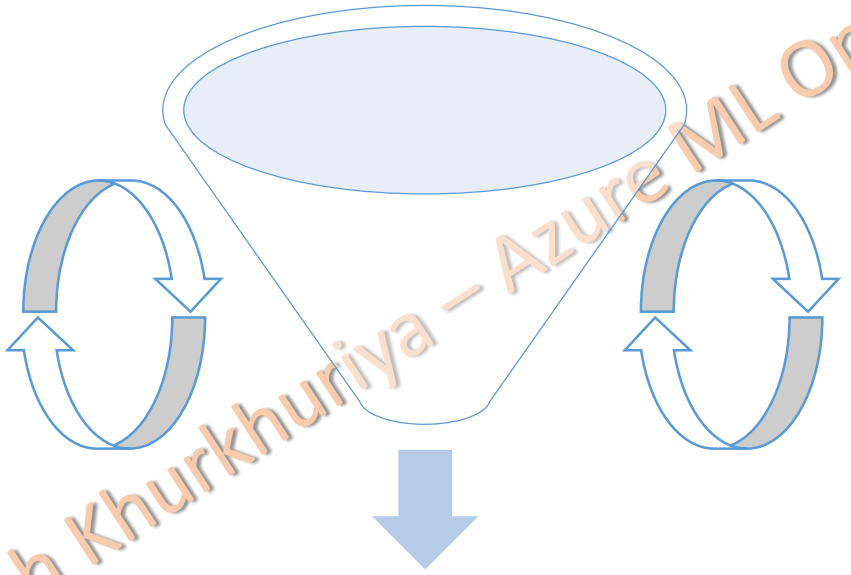
© Jitesh Khurkhuriya – Azure ML Online Course

Filter Based Feature Selection



Wrapper Method

X1 X2 X3 X4 ..... Xn



Learning Algorithm

Classification
Regression

Subset
X2 X4

Performance

Accuracy
Precision
Root-Mean-Squared

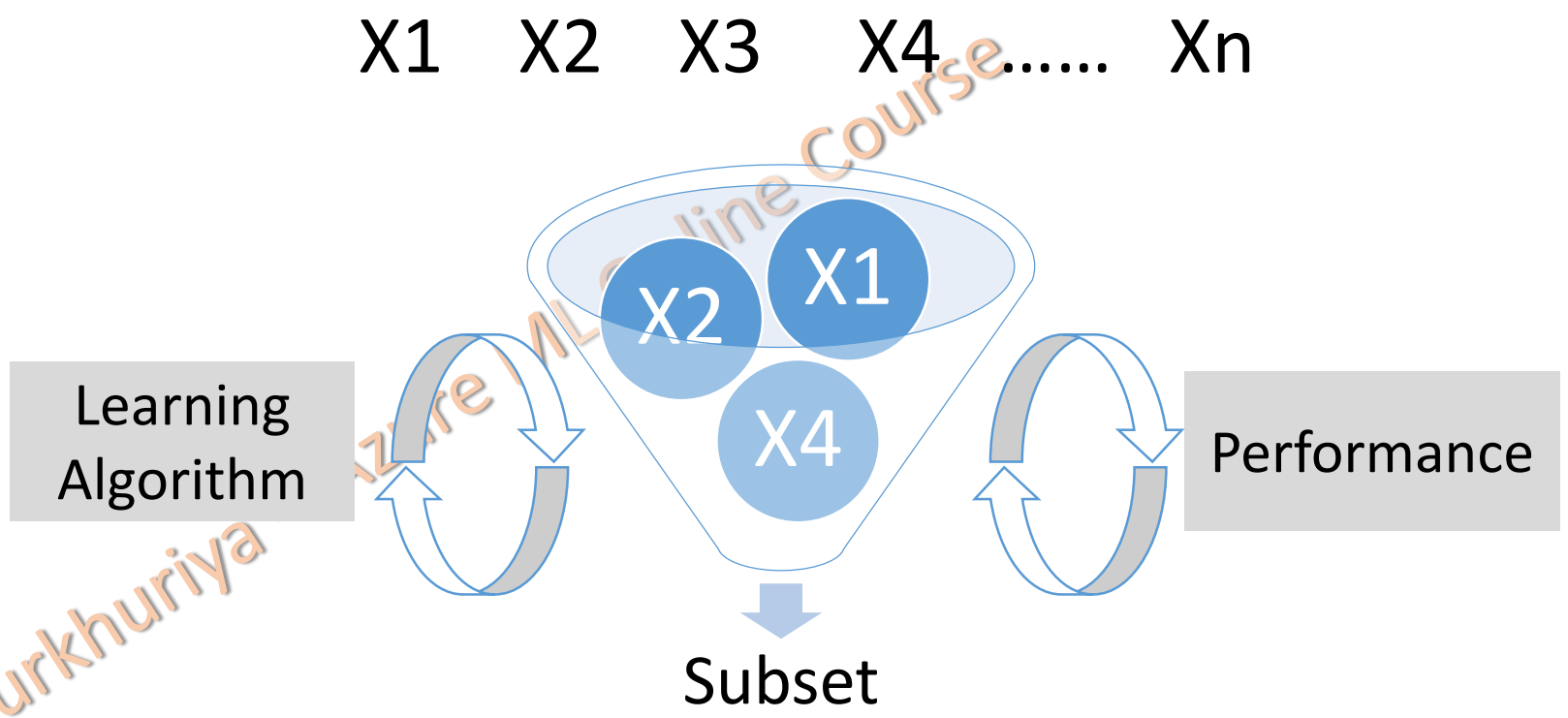
© Jitesh Khurkhuriya – Azure ML Online Course

Wrapper Method

Forward Selection

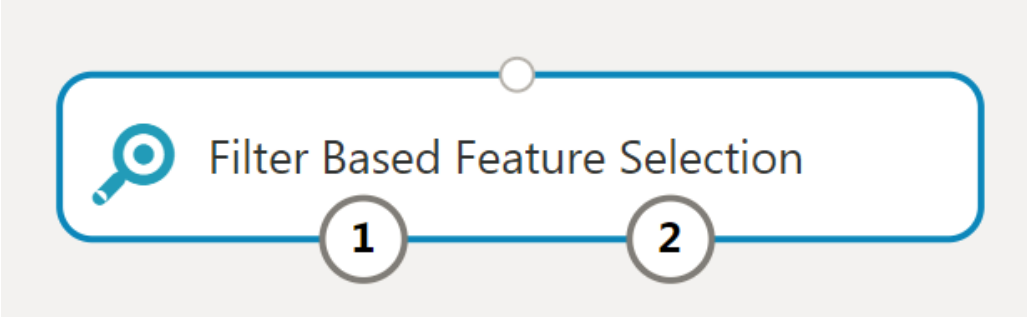
Backward Elimination

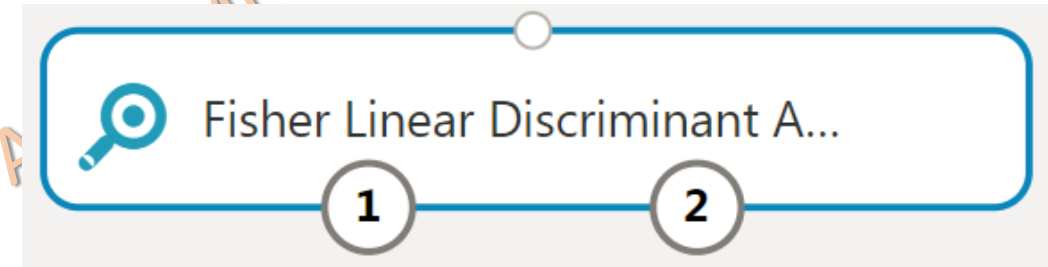
Recursive Feature Elimination

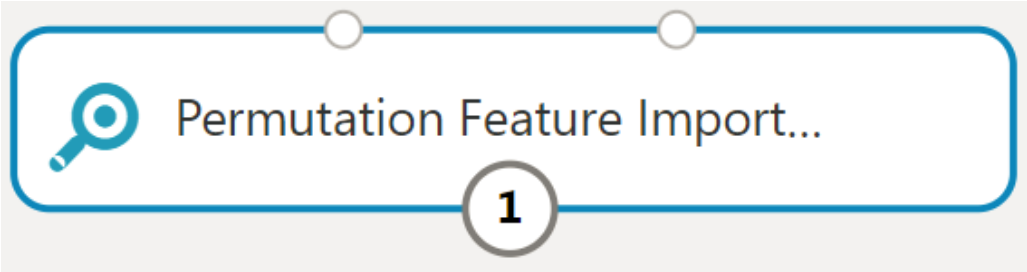


AzureML Feature Selection Methods

- Filter Based Feature Selection
- Fisher Linear Discriminant Analysis
- Permutation Feature Importance







Filter Based Feature Selection

Types of Feature Scoring Methods

- Pearson Correlation
- Mutual Information
- Kendall Correlation
- Spearman Correlation
- Chi Squared
- Fisher Score
- Count Based

© Jitesh Khurkhuriya – Azure ML Online Course

Pearson Correlation

Standard Deviation

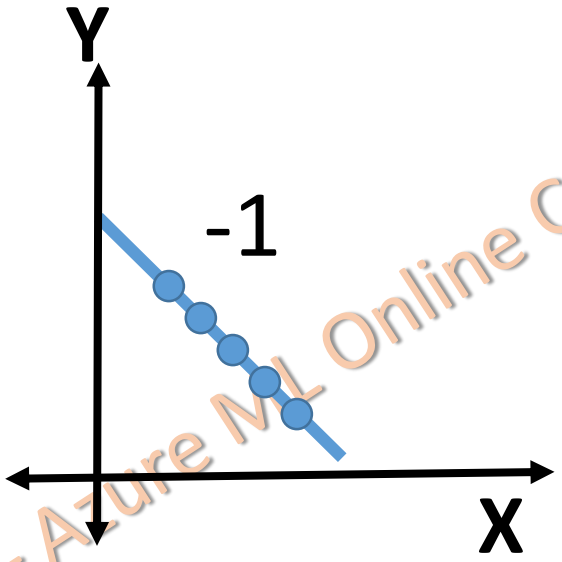
$$\sigma_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(xi - \bar{x})^2}$$

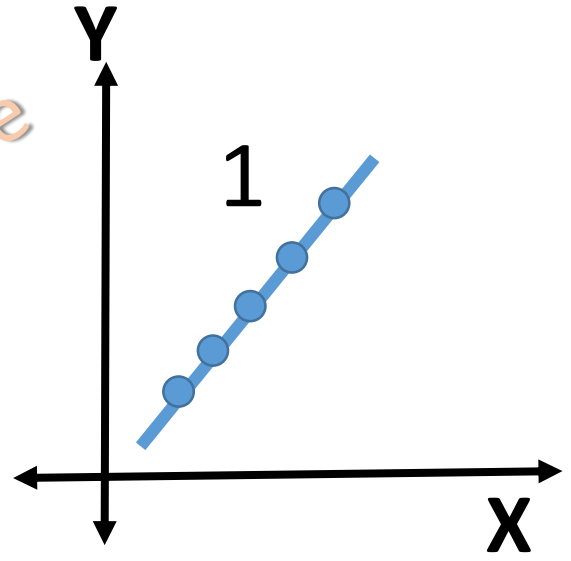
© Jitesh Khurkhuriya – Azure ML Online Course

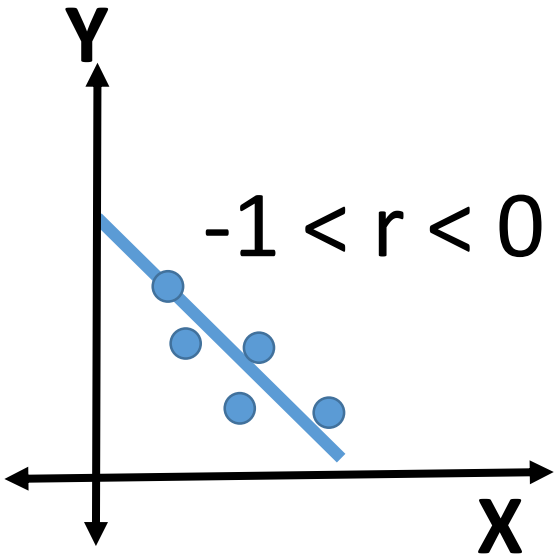
Pearson Correlation Coefficient

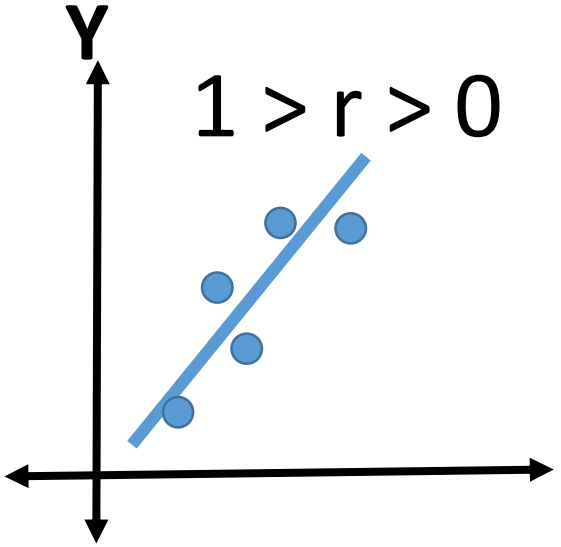
- Measure of linear correlation between two variables

$$r = \frac{\sum_{i=1}^{n}(xi - \bar{x})(yi - \bar{y})}{(N-1)\,\sigma_x\,\sigma_y}$$









© Jitesh Khurkhuriya – Azure ML Online Course

Pearson Correlation Coefficient

Engine-size	Horsepower	**Price**
130	111	**13495**
130	111	**16500**
152	154	**16500**
109	102	**13950**
136	115	**17450**
136	110	**15250**
136	110	**17710**
136	110	**18920**
131	140	**23875**
108	101	**16430**
108	101	**16925**
164	121	**20970**
164	121	**21105**
164	121	**24565**

$$\sigma_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(xi - \overline{x})^2}$$

Engine Size $\sigma_x = 18.9$

HP $\sigma_x = 14.39$

Price $\sigma_y = 3267.78$

Engine Size – r

engine-size	horsepower	price
130	111	**13495**
130	111	**16500**
152	154	**16500**
109	102	**13950**
136	115	**17450**
136	110	**15250**
136	110	**17710**
136	110	**18920**
131	140	**23875**
108	101	**16430**
108	101	**16925**
164	121	**20970**
164	121	**21105**
164	121	**24565**

Engine Size, σ_x = **18.9** Price, σ_y = **3267.78**

$$r = \frac{\sum_{i=1}^{n}(xi - \bar{x})(yi - \bar{y})}{(N-1)\,\sigma_x\,\sigma_y}$$

$$= \frac{519{,}985}{13 * 18.9 * 3267.78} = \boxed{0.65}$$

Horsepower – r

engine-size	horsepower	price
130	111	**13495**
130	111	**16500**
152	154	**16500**
109	102	**13950**
136	115	**17450**
136	110	**15250**
136	110	**17710**
136	110	**18920**
131	140	**23875**
108	101	**16430**
108	101	**16925**
164	121	**20970**
164	121	**21105**
164	121	**24565**

HP, σ_x = **14.39**

Price, σ_y = **3267.78**

$$r = \frac{\sum_{i=1}^{n}(xi - \bar{x})(yi - \bar{y})}{(N-1)\,\sigma_x\,\sigma_y}$$

$$= \frac{286{,}400}{13 * 14.39 * 3267.78} = 0.47$$

Feature Selection

engine-size	horsepower	price
130	111	**13495**
130	111	**16500**
152	154	**16500**
109	102	**13950**
136	115	**17450**
136	110	**15250**
136	110	**17710**
136	110	**18920**
131	140	**23875**
108	101	**16430**
108	101	**16925**
164	121	**20970**
164	121	**21105**
164	121	**24565**

$$r = \frac{\sum_{i=1}^{n}(xi - \bar{x})(yi - \bar{y})}{(N-1)\,\sigma_x\,\sigma_y}$$

Horsepower r = 0.47

Engine Size r = 0.65

Pearson Correlation Coefficient

- Advantages
  - Identifies the correlation as well as the extent
  - Identifies the positive or negative correlation
  - Easiest to apply and understand among all the other correlation techniques

- Disadvantages
  - Outliers affect the value of “r”
  - Not good for non-linear relationship
  - Might not be good to find out causal

Engine	HP	**Price**
130	111	**13495**
130	111	**16500**
152	154	**16500**
109	102	**13950**

© Jitesh Khurkhuriya – Azure ML Online Course

Chi-Square

Chi Squared Test of Independence

- Developed by Karl Pearson

- Evaluates the relationship between two categorical variables

- Steps to Evaluate the Independence
  - Define Hypothesis – Null and Alternate
  - Define Alpha
  - Calculate the Degrees of Freedom
  - State Decision Rule
  - Calculate Test Statistics
  - Results
  - Conclusion

© Jitesh Khurkhuriya – Azure ML Online Course

Chi Square Test of Independence

Flight Status	Weather
Delayed	Rainy
Delayed	Rainy
Delayed	Rainy
Ontime	Rainy
Delayed	Rainy
Ontime	Sunny
Delayed	Rainy
Delayed	Rainy
Ontime	Sunny
Delayed	Rainy
Delayed	Overcast
Delayed	Overcast
Delayed	Overcast
Delayed	Overcast
Delayed	Overcast

Step 1

Null Hypothesis – There is no relationship between Flight Status and Weather

Alternate Hypothesis – There is relationship between Flight Status and Weather

Step 2

alpha = 0.05

Chi Square Test of Independence

	Rainy	Sunny	Overcast	
Delayed	36	16	13	**65**
On time	11	84	40	**135**
	47	**100**	**53**	

Step 3

Calculate the degrees of freedom

(no of Rows – 1) * (no of Columns – 1)

(2-1)*(3-1)

2

Chi Square Test of Independence

Step 4

State Decision Rule using chi square degrees of freedom table

Table 3-1 Critical Values of the χ^2 Distribution

					P		
df	**0.995**	**0.975**	**0.9**	**0.5**	**0.1**	**0.05**	**0.025**
1	.000	.000	0.016	0.455	2.706	3.841	5.024
2	0.010	0.051	0.211	1.386	4.605	5.991	7.378
3	0.072	0.216	0.584	2.366	6.251	7.815	9.348
4	0.207	0.484	1.064	3.357	7.779	9.488	11.143
5	0.412	0.831	1.610	4.351	9.236	11.070	12.832
6	0.676	1.237	2.204	5.348	10.645	12.592	14.449
7	0.989	1.690	2.833	6.346	12.017	14.067	16.013
8	1.344	2.180	3.490	7.344	13.362	15.507	17.535
9	1.735	2.700	4.168	8.343	14.684	16.919	19.023

Reject the Null Hypothesis if the X square value is greater than 5.991

Chi Squared Test of Independence

Step 5 Calculate Test Statistics

$$f_e = \frac{f_c \ f_r}{n}$$

(Delayed, Rainy)

= (47*65)/200

= 15.275

Actual

	Rainy	Sunny	Overcast	
Delayed	36	16	13	**65**
On time	11	84	40	**135**
	47	**100**	**53**	

Expected

	Rainy	Sunny	Overcast	
Delayed	15	33	17	**65**
On time	32	67	36	**135**
	47	**100**	**53**	

Chi Squared Test of Independence

Step 6 Calculate Results

$$\sum \frac{(f_o - f_e)^2}{f_e} = 55.6$$

o – Original/Actual

e – Expected

	Rainy	Sunny	Overcast
Delayed	28.12	8.38	1.04
On time	13.54	4.03	0.5

Chi Squared Test of Independence

Step 7 Conclusion

Reject the null hypothesis as X-Squared value (55.6) is greater than 5.991

The weather and Flight Status are correlated

Kendall Correlation

Kendall Correlation Rank Correlation

- Named after Maurice Kendall
- Measure of Rank Correlation

© Jitesh Khurkhuriya – Azure ML Online Course

Kendall Correlation Rank Correlation

engine-size	horsepower	**price**
130	111	**13495**
130	111	**16500**
152	154	**16500**
109	102	**13950**
136	115	**17450**
136	110	**15250**
136	110	**17710**
136	110	**18920**
131	140	**23875**
108	101	**16430**
108	101	**16925**
164	121	**20970**
164	121	**21105**
164	121	**24565**

Kendall Correlation Rank Correlation

	engine-size	price
i	108	16430
	108	16925
	109	13950
k	130	13495
	130	16500
j	131	23875
	136	17450
	136	15250
	136	17710
	136	18920
	152	16500
	164	20970
	164	21105
	164	24565

$x_i = 108$ $y_i = 16{,}430$

$x_j = 131$ $y_j = 23{,}875$

$x_i < x_j$

$y_i < y_j$

→ Concordant Pair

$x_i > x_k$

$y_i < y_k$

Discordant Pair

$x_i < x_k$

$y_i > y_k$

Kendall Correlation Rank Correlation

$$\text{Kendall Coefficient Tau, } T = \frac{\text{Number of Concordant Pairs} - \text{Number of Discordant pairs}}{n * (n-1)/2}$$

1. T is always in the range of +1 to –1
2. Positive value indicates positive correlation
3. Negative value indicates negative correlation
4. If X and Y are independent, T will be zero

Spearman’s Rank Correlation

Spearman's Rank Correlation

- Named after Charles Spearman
- Measure of Rank Correlation
- Relationship through monotonic function

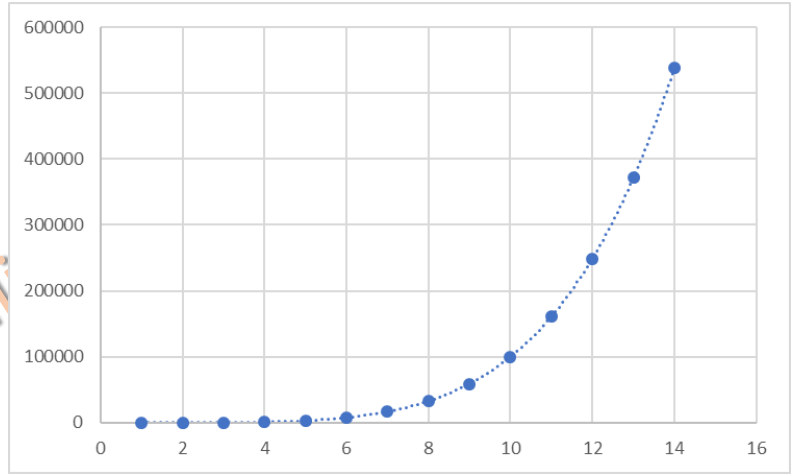
© Jitesh Khurkhuriya – Azure ML Online Course

Spearman's Rank Correlation

X	Y (X^5)
1	1
2	32
3	243
4	1024
5	3125
6	7776
7	16807
8	32768
9	59049
10	100000
11	161051
12	248832
13	371293
14	537824

$$X1 < X2$$

$$f(x1) < f(x2)$$



Best used for non-linear correlation

$$\text{rho} = 1 - \frac{6\sum d^2}{n^3 - n}$$

Calculate rho for Engine Size

engine-size	price
108	16430
108	16925
109	13950
130	13495
130	16500
131	23875
136	17450
136	15250
136	17710
136	18920
152	16500
164	20970
164	21105
164	24565

© Jitesh Khurkhuriya – Azure ML Online Course

Calculate rho for EngineSize

engine-size	Rank-x		price
108	**1**		16430
108	**1**		16925
109	**2**		13950
130	**3**		13495
130	**3**		16500
131	**4**		23875
136	**5**		17450
136	**5**		15250
136	**5**		17710
136	**5**		18920
152	**6**		16500
164	**7**		20970
164	**7**		21105
164	**7**		24565

1. Calculate the rank of the feature w.r.t. other values in the column

Calculate rho for EngineSize

engine-size	Rank-x		price	Rank-y
108	**1**		16430	**4**
108	**1**		16925	**6**
109	**2**		13950	**2**
130	**3**		13495	**1**
130	**3**		16500	**5**
131	**4**		23875	**12**
136	**5**		17450	**7**
136	**5**		15250	**3**
136	**5**		17710	**8**
136	**5**		18920	**9**
152	**6**		16500	**5**
164	**7**		20970	**10**
164	**7**		21105	**11**
164	**7**		24565	**13**

2. Calculate the rank of the predicted feature w.r.t. other values in the column

Calculate rho for EngineSize

engine-size	Rank-x		price	Rank-y	D	D^2
108	1		16430	4	-3	9
108	1		16925	6	-5	25
109	2		13950	2	0	0
130	3		13495	1	2	4
130	3		16500	5	-2	4
131	4		23875	12	-8	64
136	5		17450	7	-2	4
136	5		15250	3	2	4
136	5		17710	8	-3	9
136	5		18920	9	-4	16
152	6		16500	5	1	1
164	7		20970	10	-3	9
164	7		21105	11	-4	16
164	7		24565	13	-4	16

3. Calculate the Rank-difference (RankX and RankY) between price and engine-size and its square

Calculate rho for EngineSize

engine-size	Rank-x		price	Rank-y	D	D^2
108	1		16430	4	-3	9
108	1		16925	6	-5	25
109	2		13950	2	0	0
130	3		13495	1	2	4
130	3		16500	5	-2	4
131	4		23875	12	-8	64
136	5		17450	7	-2	4
136	5		15250	3	2	4
136	5		17710	8	-3	9
136	5		18920	9	-4	16
152	6		16500	5	1	1
164	7		20970	10	-3	9
164	7		21105	11	-4	16
164	7		24565	13	-4	16

$$\text{rho} = 1 - \frac{6\sum d^2}{n^3 - \text{n}}$$

$$= 0.6$$

Calculate rho for HP

HP	Rank-x		price	Rank-y	D	D^2
101	1		16430	4	-3	9
101	1		16925	6	-5	25
102	2		13950	2	0	0
110	3		15250	3	0	0
110	3		17710	8	-5	25
110	3		18920	9	-6	36
111	4		13495	1	3	9
111	4		16500	5	-1	1
115	5		17450	7	-2	4
121	6		20970	10	-4	16
121	6		21105	11	-5	25
121	6		24565	13	-7	49
140	7		23875	12	-5	25
154	8		16500	5	3	9

$$\text{rho} = 1 - \frac{6\sum d^2}{n^3 - n}$$

$$= 0.49 \quad < \quad 0.6$$

Select Engine Size Over HP

Correlation Coefficients

Pearson Correlation

$$r = \frac{\sum_{i=1}^{n}(xi - \bar{x})(yi - \bar{y})}{(N-1)\,\sigma_x\,\sigma_y}$$

Kandell's Correlation

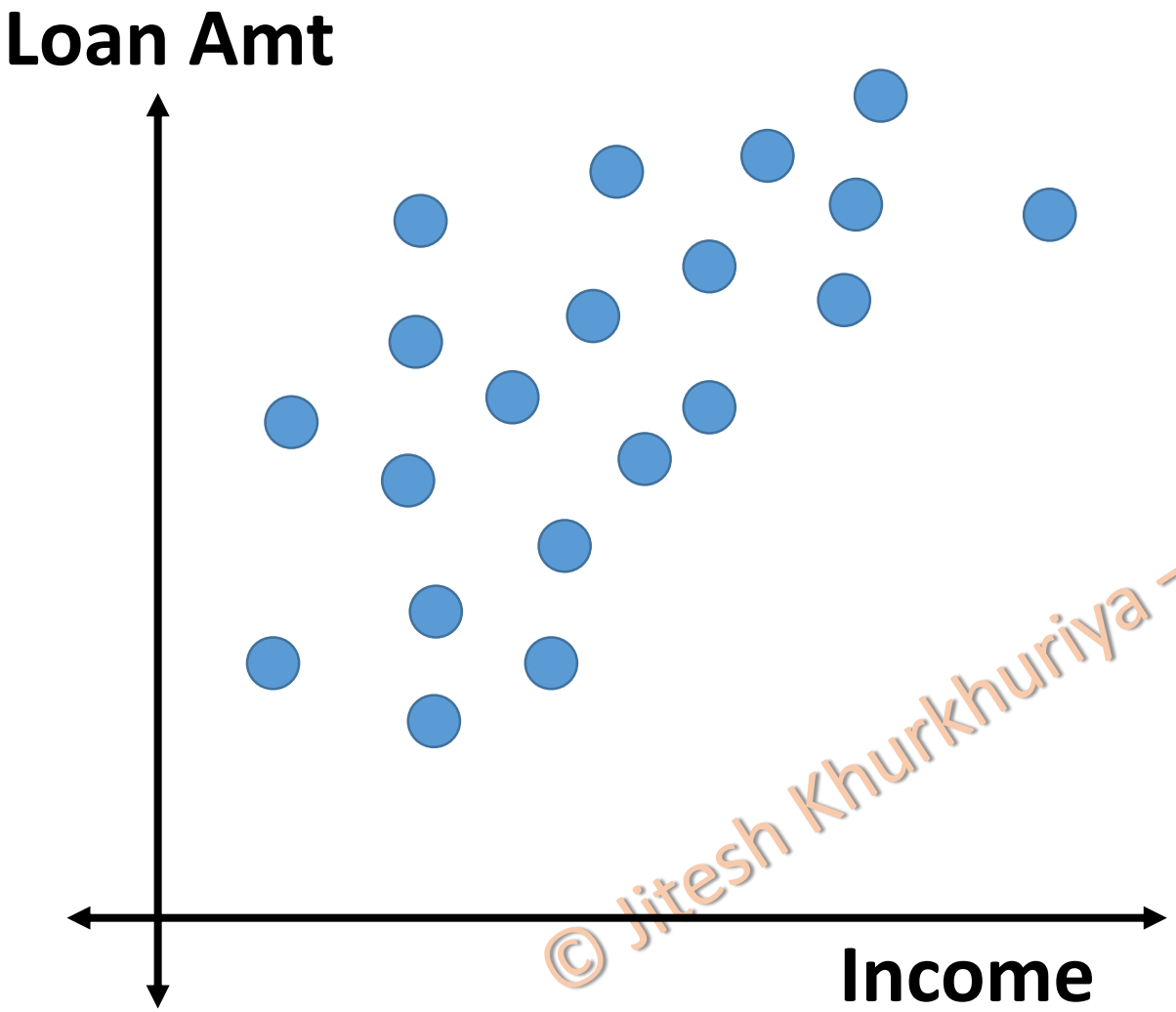
$$T = \frac{C - D}{C + D}$$

Spearman's Correlation

$$rho = 1 - \frac{6\sum d^2}{n^3 - n}$$

Fisher Linear Discriminant Analysis

Loan Approval Dataset

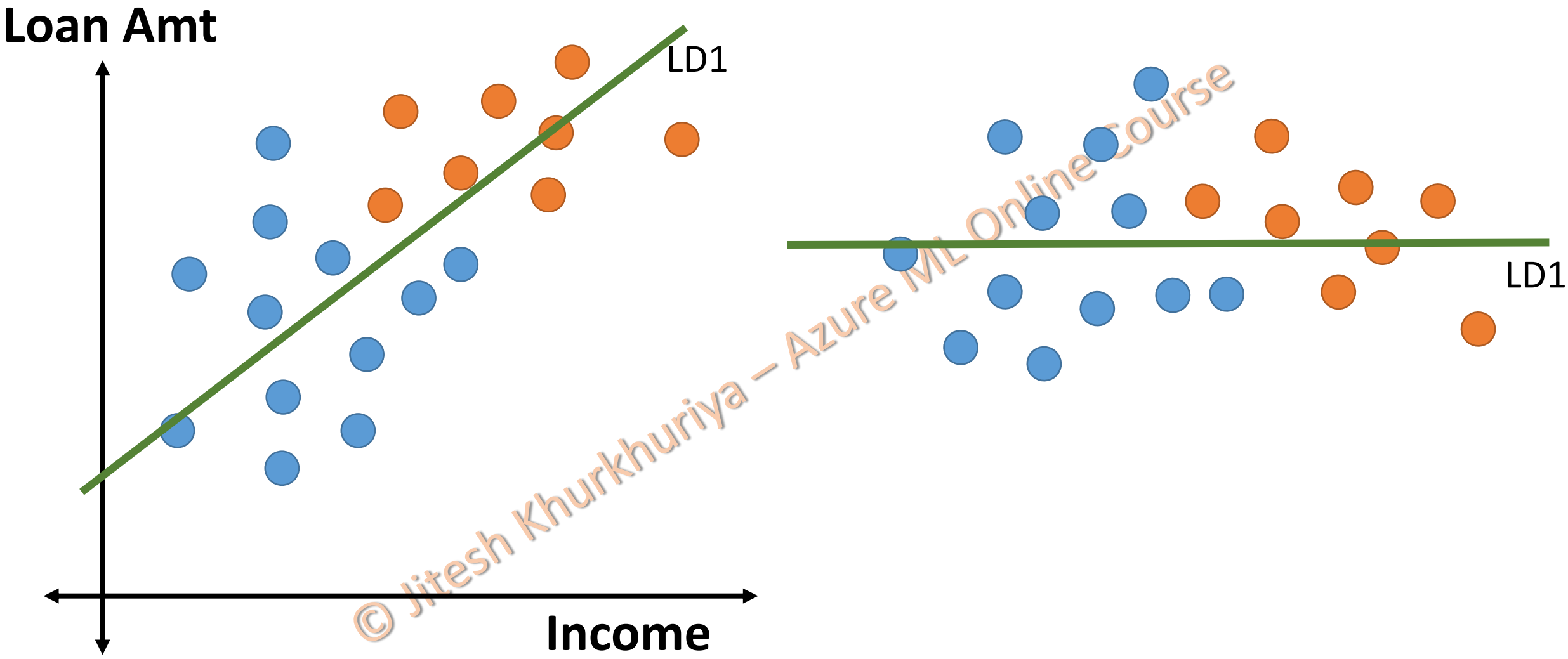


Loan Approved?

Binary Classification

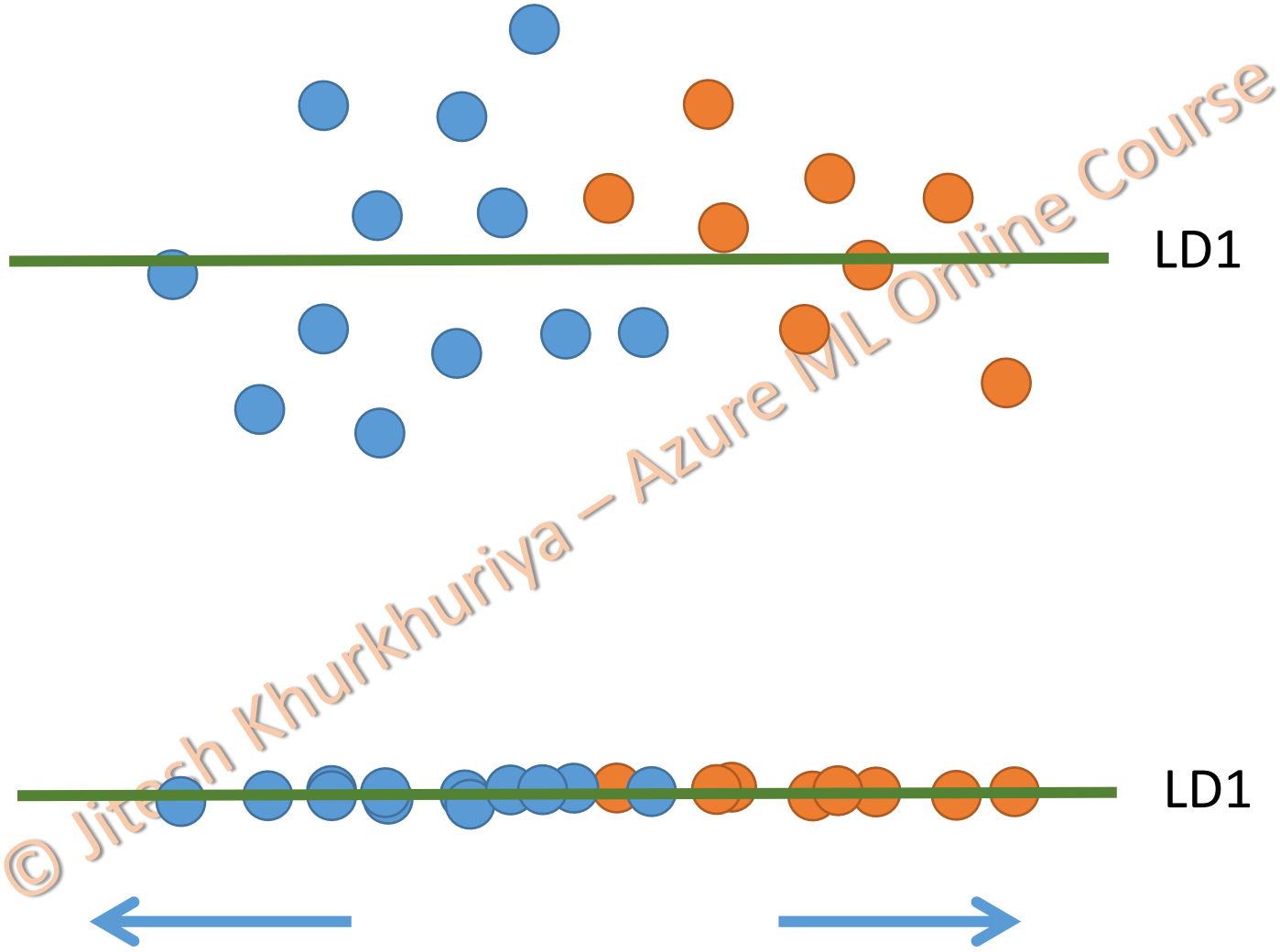
Yes or No

Linear Discriminant Analysis



© Jitesh Khurkhuriya – Azure ML Online Course

Linear Discriminant Analysis

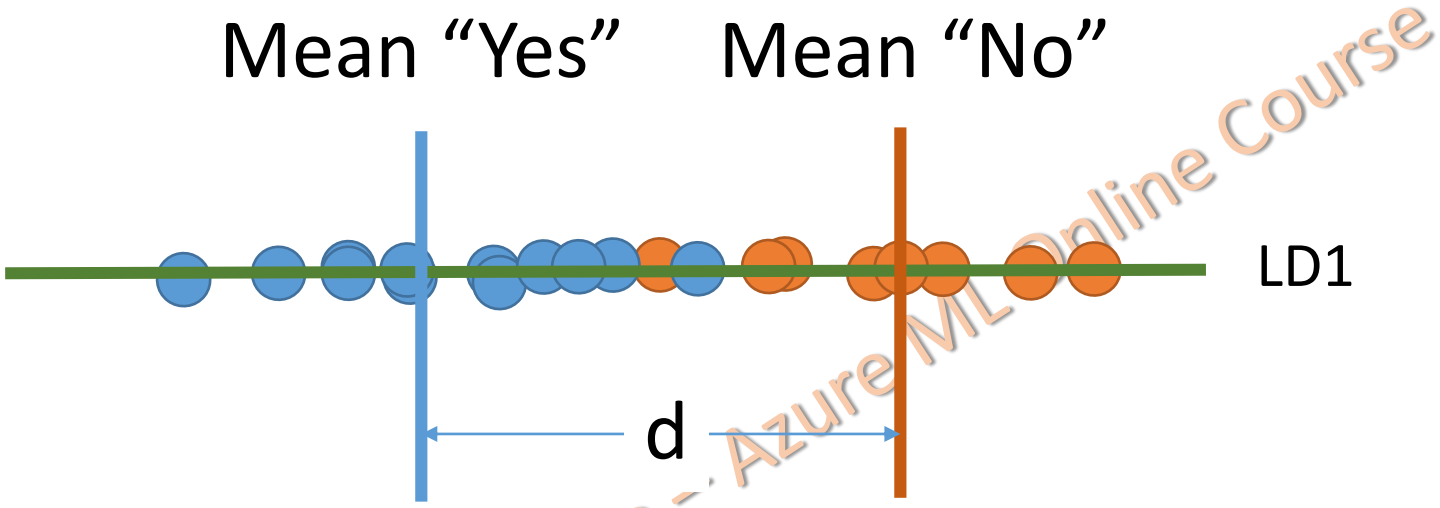


Sir Ronald Fisher

- British statistician and biologist who used mathematics to combine Mendelian genetics and natural selection.

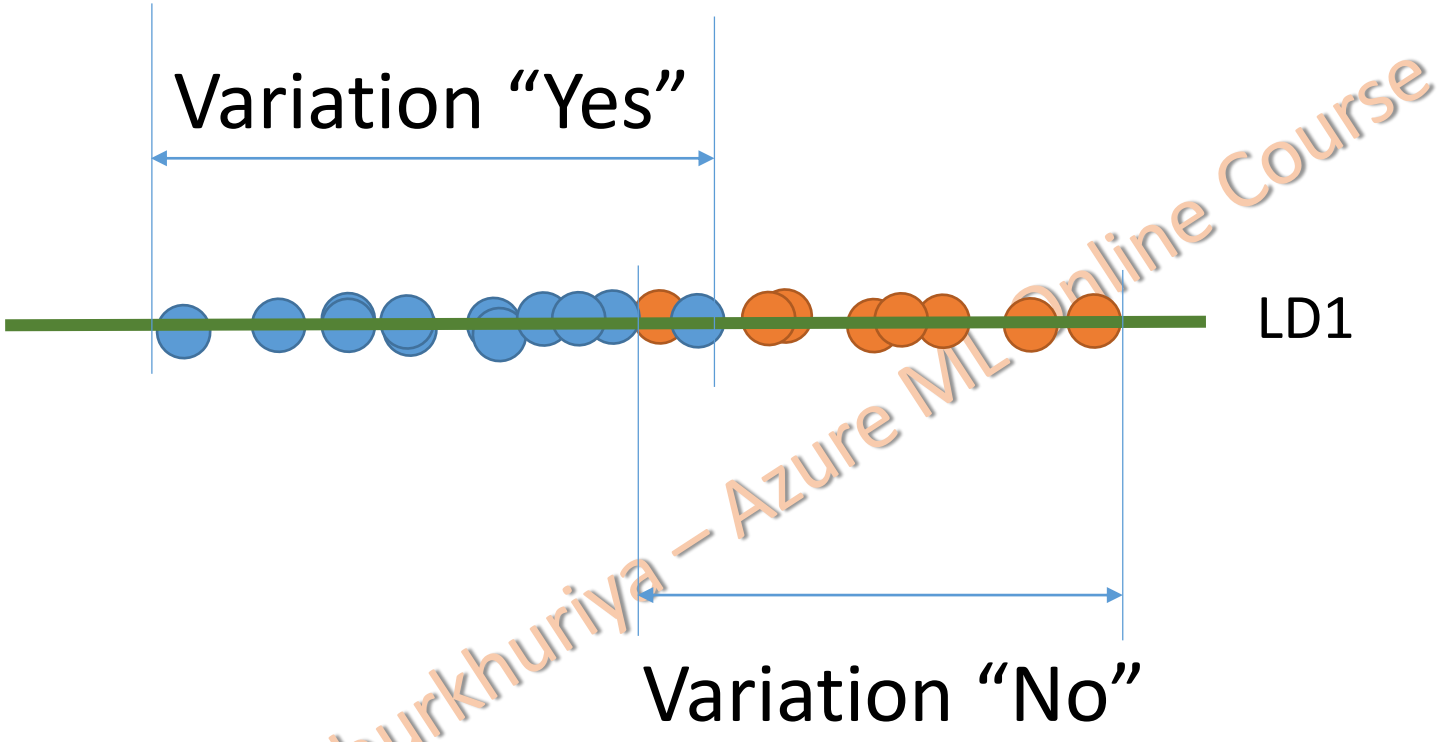
- Developed Analysis of Variance or ANOVA

How LDA is Created?



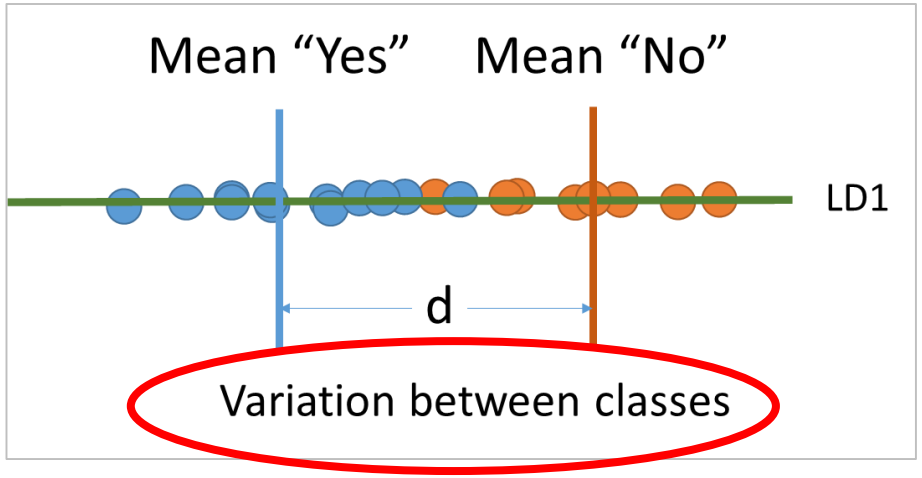
Variation between classes

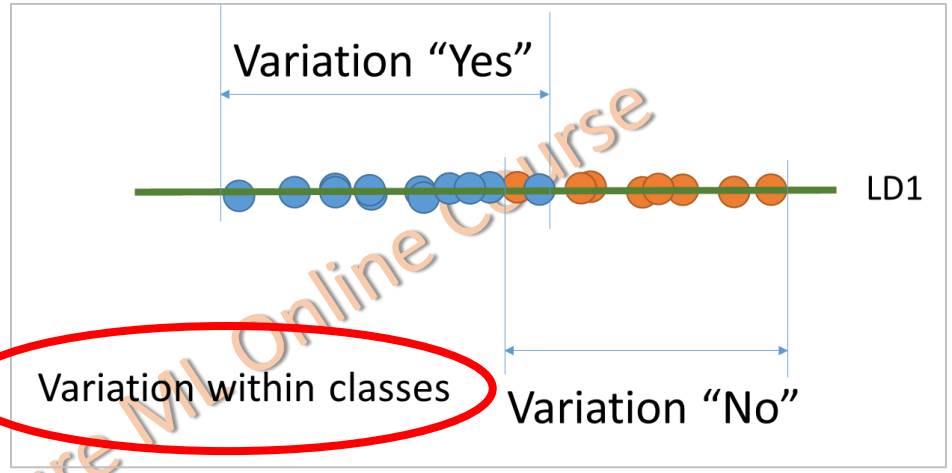
How LDA is Created?



Variation within classes

Fisher LDA



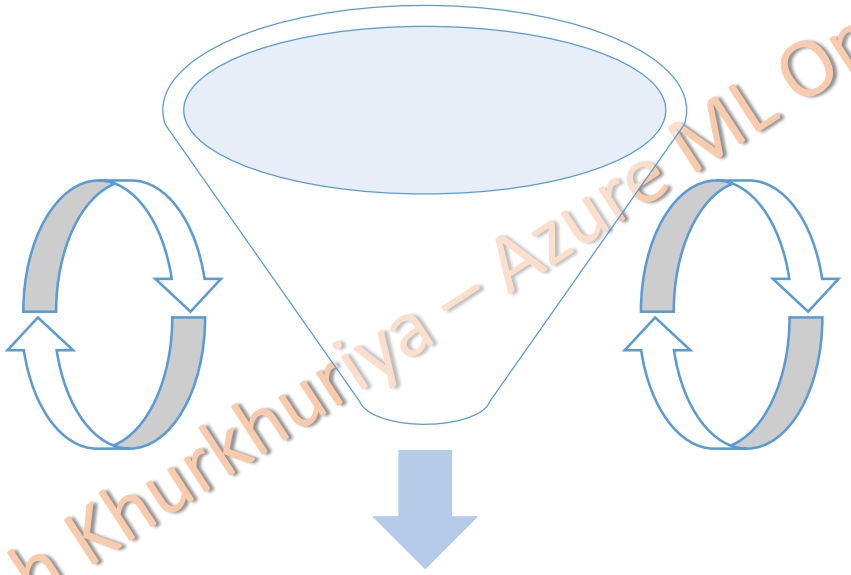


$$S - \text{Separation} = \frac{\text{Variation Between Classes}}{\text{Total Variation Within Classes}}$$

Permutation Feature Importance

Wrapper Method

X1 X2 X3 X4 ..... Xn



Learning Algorithm

Classification
Regression

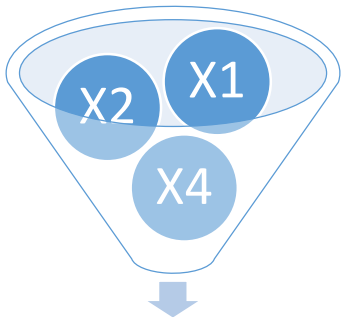
Subset
X2 X4

Performance

Accuracy
Precision
Root-Mean-Squared

© Jitesh Khurkhuriya – Azure ML Online Course

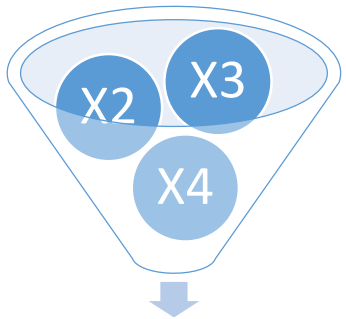
X1 X2 X3 X4 ..... Xn



Subset 1

Training 1

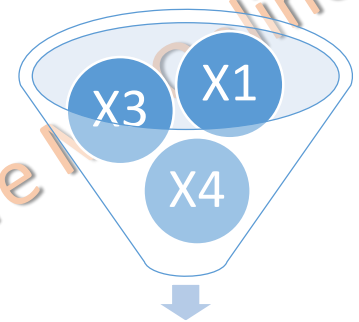
Performance 1



Subset 2

Training 2

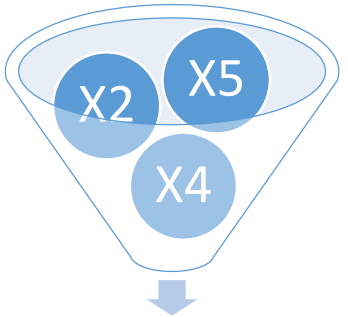
Performance 2



Subset 3

Training 3

Performance 3



Subset 4

Training 4

Performance 4

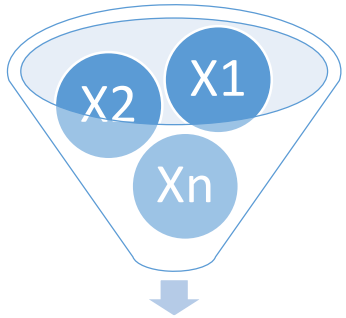
© Jitesh Khurkhuriya – Azure ML Online Course

Permutation Feature Importance

X1 X2 X3 X4 ..... Xn

Training

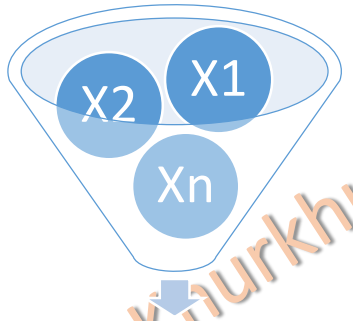
Shuffle



Test Data 1

Performance 1

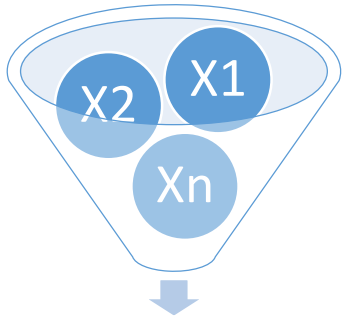
Shuffle



Test Data 2

Performance 2

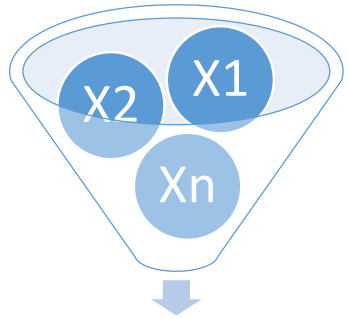
Shuffle



Test Data 3

Performance 3

Shuffle



Test Data 4

Performance 4

Example using a dataset

Engine Size	HP	Colour	Price
130	111	Black	**13495**
130	111	Red	**16500**
152	154	Red	**16500**
109	102	Black	**13950**
136	115	Grey	**17450**
136	110	Black	**15250**
136	110	Black	**17710**
136	110	Grey	**18920**
131	140	Red	**23875**
108	101	Grey	**16430**
108	101	Red	**16925**
164	121	Black	**20970**
164	121	Red	**21105**
164	121	Red	**24565**

Trained Model

Shuffle and Compare

Trained Model

Engine Size	HP	Colour	Price
130	111	Black	**13495**
130	**111**	Red	**16500**
152	**154**	Red	**16500**
136	110	Black	**17710**
136	110	Grey	**18920**
131	**140**	Red	**23875**
108	**101**	Grey	**16430**
164	121	Red	**24565**

Shuffle and Compare

Trained Model

Engine Size	HP	Colour	Price
130	111	Black	**13495**
130	**154**	Red	**16500**
152	**111**	Red	**16500**
136	110	Black	**17710**
136	110	Grey	**18920**
131	**101**	Red	**23875**
108	**140**	Grey	**16430**
164	121	Red	**24565**

Engine Size	HP	Colour	Price	Price Predicted
130	111	Black	**13495**	
130	**154**	Red	**16500**	
152	**111**	Red	**16500**	
136	110	Black	**17710**	
136	110	Grey	**18920**	
131	**101**	Red	**23875**	
108	**140**	Grey	**16430**	
164	121	Red	**24565**	

Performance Measure

Classification

- Accuracy
- Precision
- Recall
- Average Log-Loss

Regression

- Mean Absolute Error
- Root Mean Squared Error
- Relative Absolute Error
- Co-Efficient of Determination
- Relative Squared Error

© Jitesh Khurkhuriya – Azure ML Online Course

Thank You and have a Great Time..!!