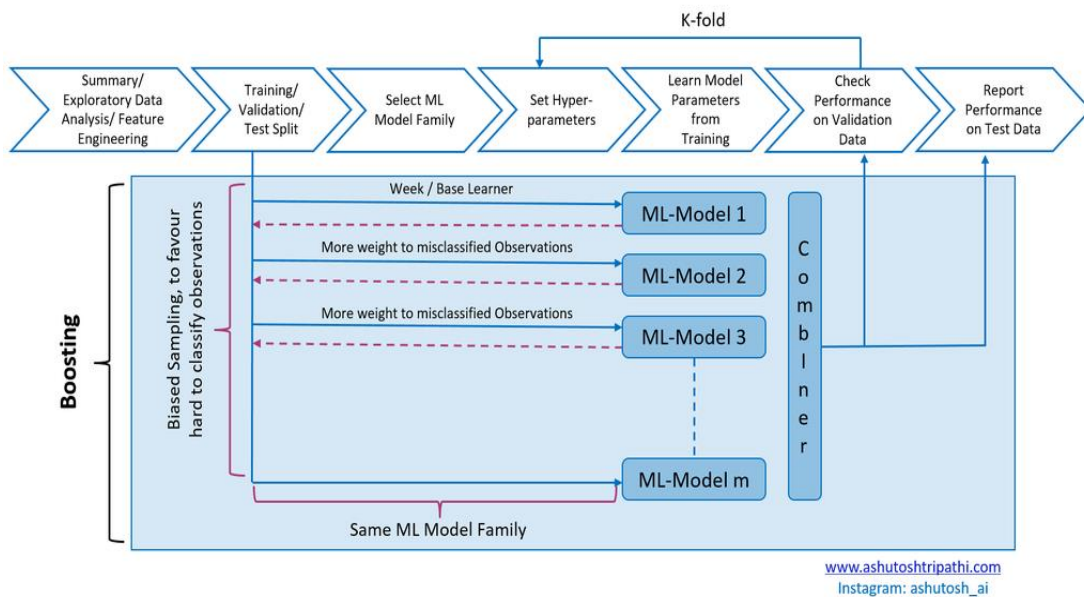


Data Science and Machine Learning | Yearly round-up 2019

Guys, I have consolidated all my ML and DS articles. In case you have missed it, here are the links in one place.

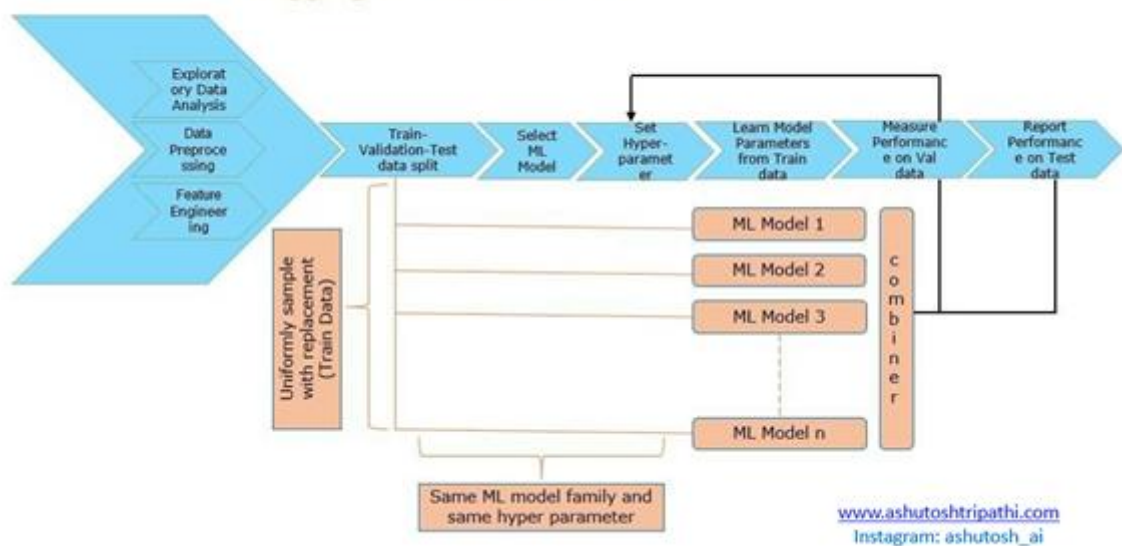
1. <https://ashutoshtripathi.com/2019/12/16/what-is-boosting-in-ensemble-learning/>

Boosting in Ensemble Learning



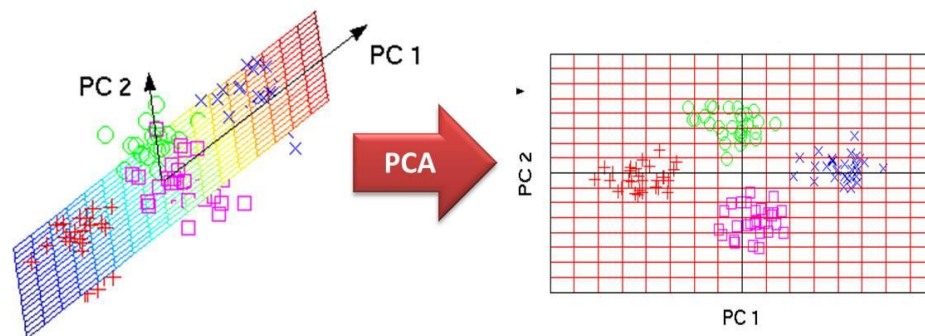
2. <https://ashutoshtripathi.com/2019/12/09/what-is-bagging-in-ensemble-learning/>

Bagging In ML Framework



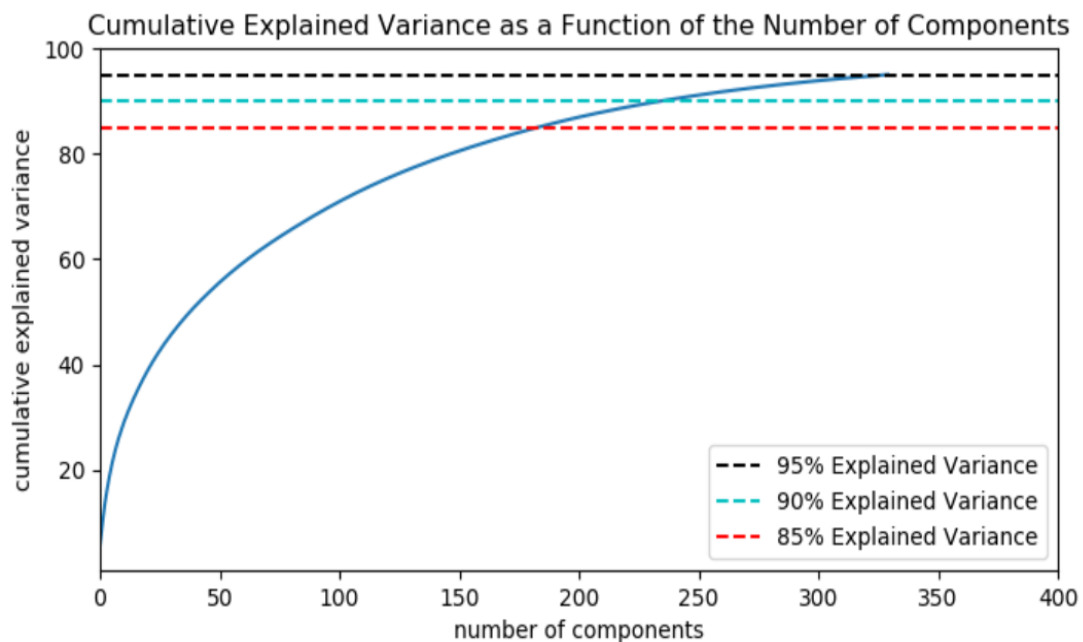
3. <https://ashutoshtripathi.com/2019/07/11/a-complete-guide-to-principal-component-analysis-pca-in-machine-learning/>

Dimensionality Reduction & Principal Component Analysis



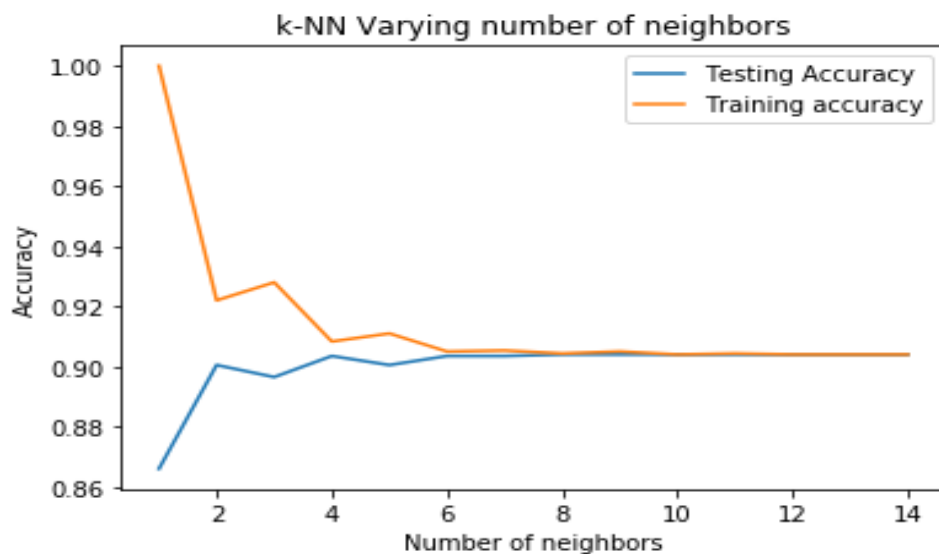
4. <https://ashutoshtripathi.com/2019/07/15/step-by-step-approach-to-principal-component-analysis-using-python/>

KNN in Machine Learning



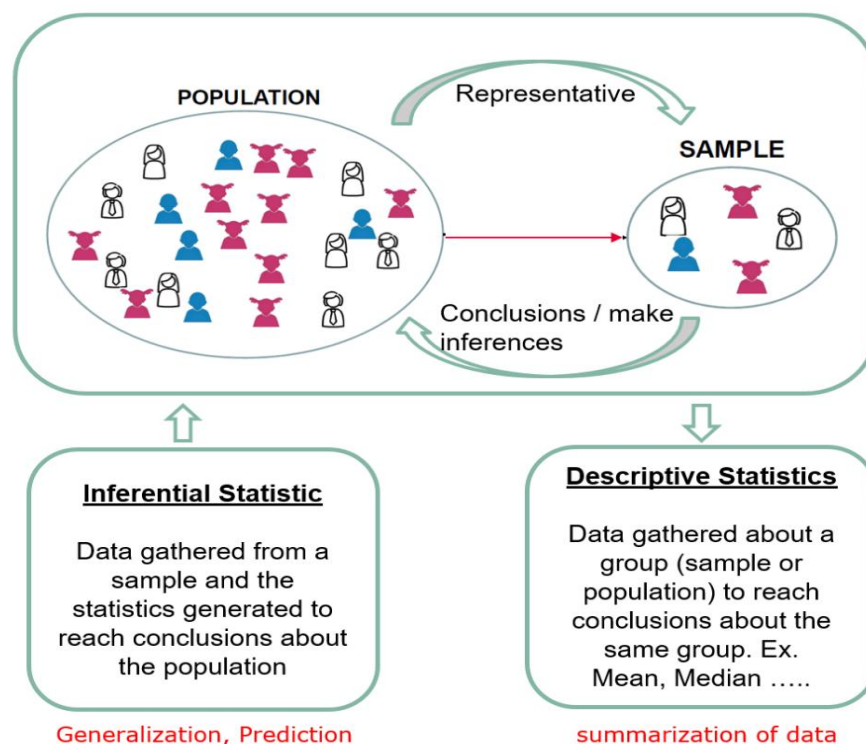
5. <https://ashutoshtripathi.com/2019/08/05/a-complete-guide-to-k-nearest-neighbors-algorithm-knn-using-python/>

KNN Implementation in Python



6. <https://ashutoshtripathi.com/2019/04/18/basic-statistics-for-data-science-part-1/>

Types of Statistics



7. <https://ashutoshtripathi.com/2019/08/09/variance-standard-deviation-and-other-measures-of-variability-and-spread/>

$$\text{Variance} = \sum_{i=1}^n (x_i - \mu)^2 / n$$

Variance

$$\text{Standard Deviation} = \sqrt{\sum_{i=1}^n (x_i - \mu)^2 / n}, \mu \text{ is mean}$$

Standard Deviation

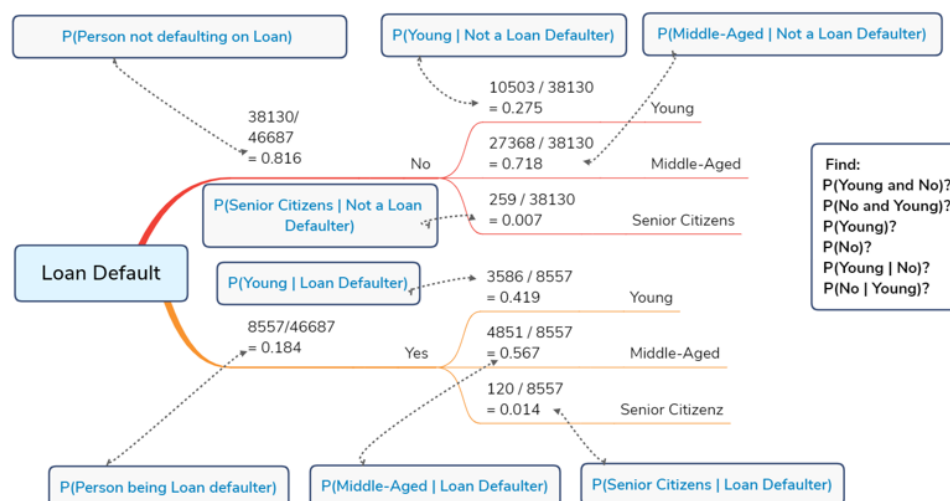
$$\text{Mean} = \sum_{i=1}^n x_i / \sum_{i=1}^n f_i \quad (\text{Sum of all scores} / \text{sum of frequencies})$$

Mean

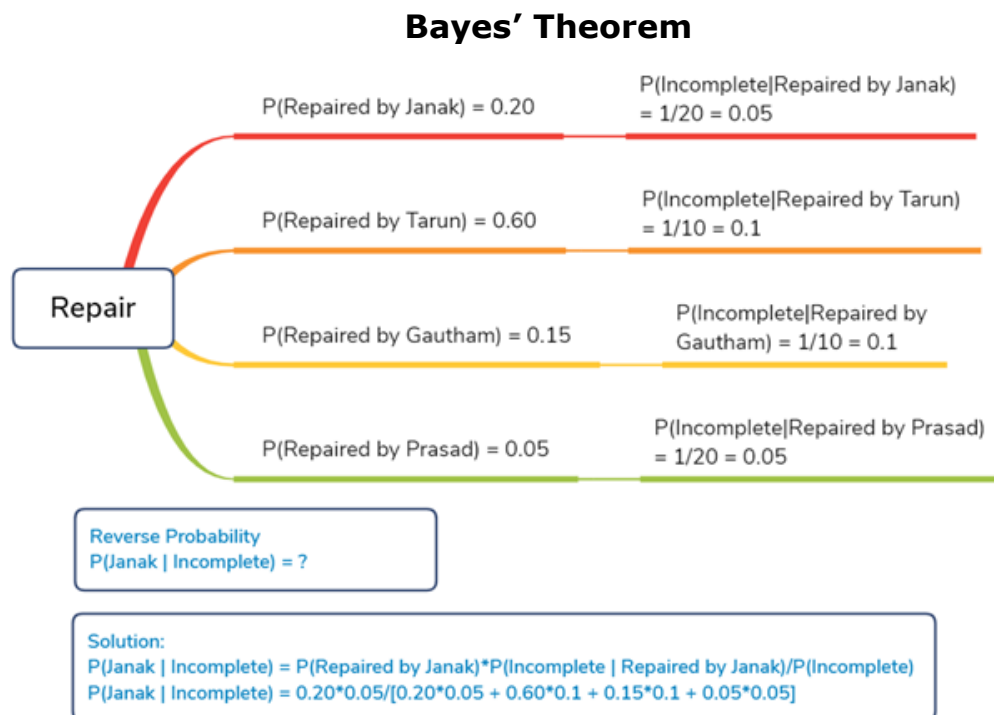
8. <https://ashutoshtripathi.com/2019/08/15/conditional-probability-with-examples-for-data-science/>

Conditional Probability Visualization using Probability Tree

		Age			Total
		Young	Middle-Aged	Senior Citizens	
Loan Default	No	10503	27368	259	38130
	Yes	3,586	4,851	120	8557
	Total	14089	32219	379	46687

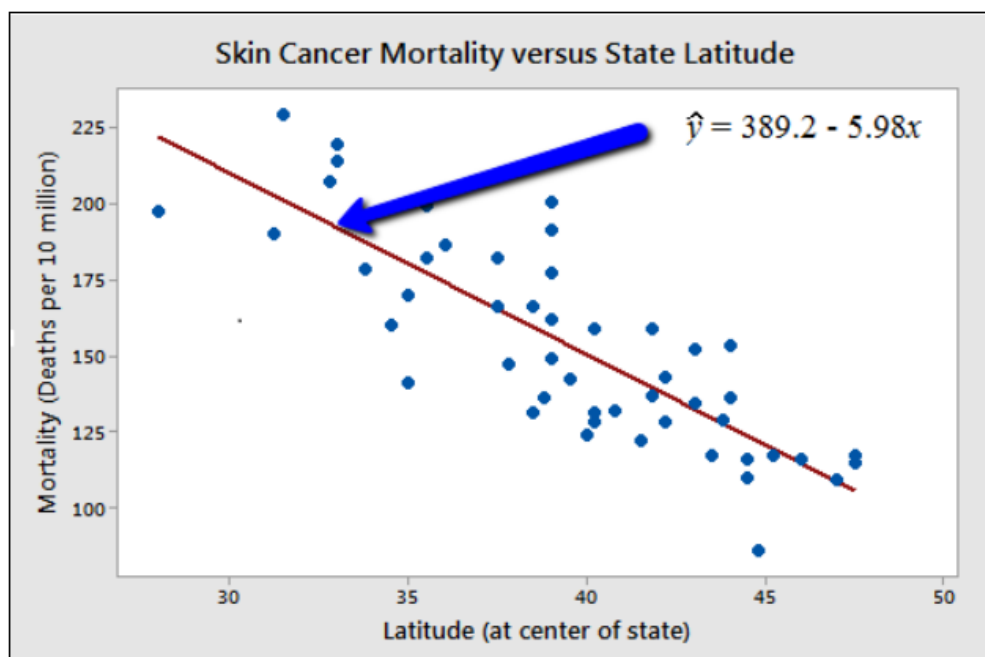


9. <https://ashutoshtriplathi.com/2019/08/20/bayes-theorem-with-example-for-data-science-professionals/>

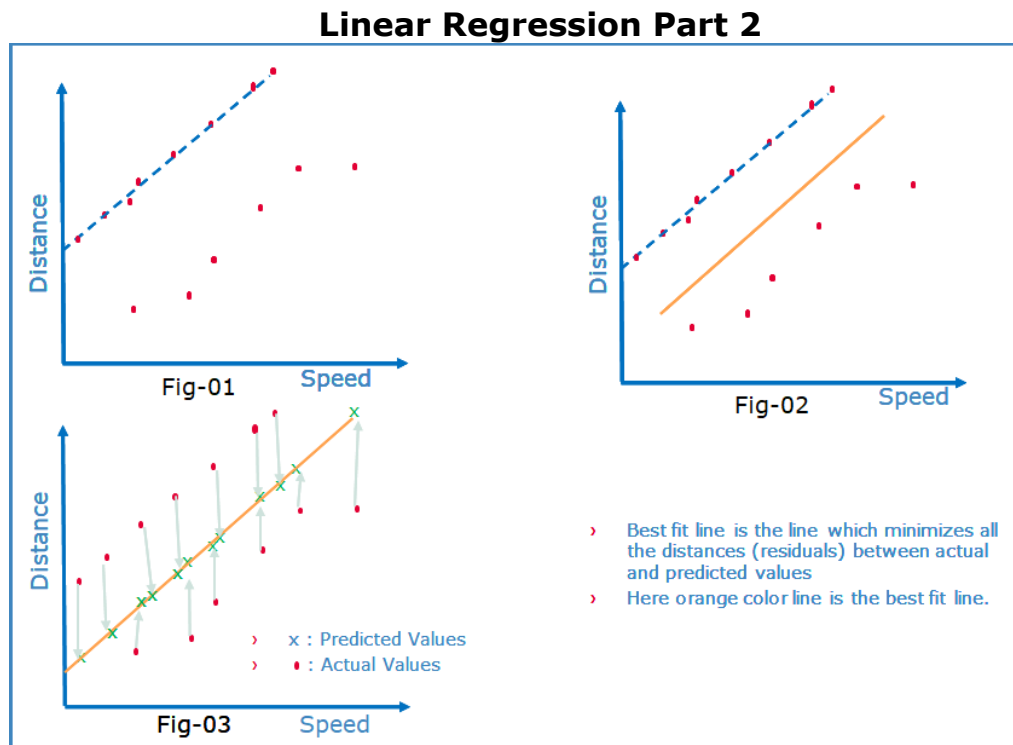


10. <https://ashutoshtriplathi.com/2019/01/16/what-is-linear-regression-part1/>

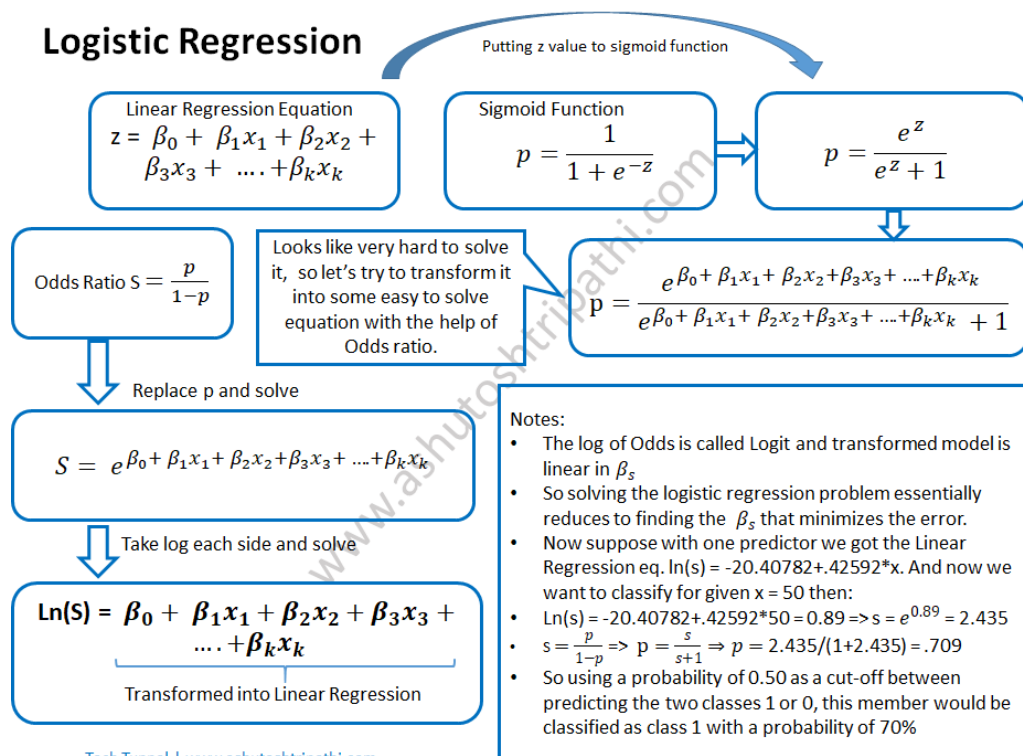
Linear Regression Part 1



11. <https://ashutoshtripathi.com/2019/01/06/what-is-linear-regression-part2/>



12. <https://ashutoshtripathi.com/2019/06/17/logistic-regression-with-an-example-in-r/>



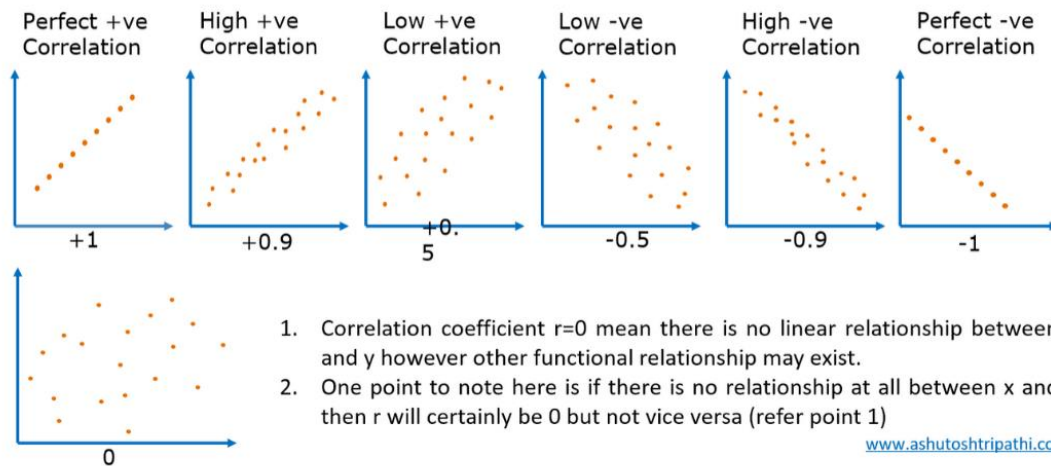
13. <https://ashutoshtripathi.com/2019/01/15/covariance-and-correlation/>

Correlation Coefficient

Correlation coefficient r is a number between -1 to +1 and tells us how well a regression line fits the data and defined by

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \quad \text{where,}$$

- s_{xy} is the covariance between x and y
- s_x and s_y are the standard deviations of x and y respectively.



14. <https://ashutoshtripathi.com/2019/01/22/what-is-the-coefficient-of-determination-r-square/>

Coefficient of Determination (R Square)

$$R^2 = \frac{SSR}{SST}$$

Where,

$$SSR = \sum_i (\hat{y}_i - \bar{y})^2$$

$$SST = \sum_i (y_i - \bar{y})^2$$

- SSR is Sum of Squared Regression also known as variation explained by the model
- SST is Total variation in the data also known as sum of squared total
- y_i is the y value for observation i
- \bar{y} is the mean of y value
- \hat{y}_i is predicted value of y for observation i

www.ashutoshtripathi.com

15. <https://ashutoshtripathi.com/2019/06/07/feature-selection-techniques-in-regression-model/>

Feature Selection Techniques

```
> MultilinearReg = lm(mpg ~ ., data = mtcars)
> summary(MultilinearReg)
```

Call:
lm(formula = mpg ~ ., data = mtcars)

Residuals:

	Min	1Q	Median	3Q	Max
	-3.6074	-0.9126	-0.2565	0.8726	4.1079

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-16.48709	25.58316	-0.644	0.530
cyl	1.17550	1.50475	0.781	0.449
disp	0.01744	0.02903	0.601	0.558
hp	-0.01745	0.02987	-0.584	0.569
drat	3.62707	2.62587	1.381	0.190
wt	-3.22226	2.73952	-1.176	0.261
qsec	0.81826	0.82806	0.988	0.341
vs-v-shaped	-0.90970	3.43725	-0.265	0.795
am	1.46786	2.55847	0.574	0.576
gear	4.03646	2.65074	1.523	0.152
carb	-1.39833	1.52270	-0.918	0.375

Residual standard error: 2.775 on 13 degrees of freedom
(8 observations deleted due to missingness)
Multiple R-squared: 0.8889, Adjusted R-squared: 0.8034
F-statistic: 10.4 on 10 and 13 DF, p-value: 0.0001112

Eliminate "vs" due to highest p value

Here we see that none of the p-value is <.05 hence it seems no variable is significant. But lets follow backward elimination by removing highest p-value variable and see what happens

16. <https://ashutoshtripathi.com/2019/06/10/what-is-stepaic-in-r/>

StepAIC in R

```
> mtcars = read.csv(file = "C:/Ashutosh/Blog/TT/Examples/mtcars.csv", header=TRUE, sep=",")
> mtcars$X=NULL
> sum(is.na(mtcars))
[1] 8
> mtcars = na.omit(mtcars)
> sum(is.na(mtcars))
[1] 0
> MultilinearReg = lm(mpg ~ ., data = mtcars)
> library(MASS)
> library(car)
> stepAIC(MultilinearReg, direction = "both")
Start: AIC=56.28
mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
```

	Df	Sum of Sq	RSS	AIC
- vs	1	0.5395	100.67	54.412
- am	1	2.5355	102.67	54.883
- hp	1	2.6289	102.77	54.905
- disp	1	2.7792	102.92	54.940
- cyl	1	4.7007	104.84	55.384
- carb	1	6.4959	106.63	55.792
- qsec	1	7.5216	107.66	56.022
<none>			100.14	56.283
- wt	1	10.6566	110.79	56.711
- drat	1	14.6965	114.83	57.570
- gear	1	17.8613	118.00	58.223

Step: AIC=54.41
mpg ~ cyl + disp + hp + drat + wt + qsec + am + gear + carb

	Df	Sum of Sq	RSS	AIC
- am	1	2.0140	102.69	52.888
- hp	1	2.0903	102.77	52.906
- disp	1	2.2765	102.95	52.949
- cyl	1	4.1825	104.86	53.389
- qsec	1	7.6645	108.34	54.173
<none>			100.67	54.412
- wt	1	10.1655	110.84	54.721
- carb	1	11.0852	111.76	54.919
- drat	1	14.2201	114.89	55.583
+ vs	1	0.5395	100.14	56.283
- gear	1	22.7382	123.41	57.300

"vs" has lowest AIC value which means the amount of information loss by removing "vs" is minimum.

Minus sign before each variable tells that stepAIC method has checked, what is the info loss by removing each variable one by one.

Step2: it will remove "vs" and run the stepAIC with remaining set of variables.

Plus sign in front of "vs" tells that in subsequent iteration, it has also checked by adding the removed variable again if it increases the AIC

17. <https://ashutoshtrpathi.com/2019/06/13/what-is-multicollinearity/>

Multicollinearity



Thank You

Wish you all a Merry Christmas and a very Happy New Year

If you like my Posts on Machine Learning, Please connect with me on

Follow my blog: <https://ashutoshtrpathi.com/>

LinkedIn: <https://www.linkedin.com/in/ashutoshtrpathi1/>

Instagram: https://www.instagram.com/ashutosh_ai/

Medium Articles: <https://medium.com/@ashutosh.optimistic>