REPORT

Akshita Teegalapally

05.07.2019.

CONTENT

PANDA'S PROFILING

CONFUSION MATRIX
KS SCORE

PRINCIPAL COMPONENT ANALYSIS

BIAS -VARIANCE TRADEOFF

ENSEMBLE MODELS

HYPERPARAMETER TUNING

WHAT'S NEXT

CONTENT

CONTENT

OVERVIEW

PANDA'S PROFILING

Generates **profile reports** from a pandas DataFrame.

- Basic data type information
- Descriptive statistics
- Quantile statistics
- Histograms
- Correlations

Dataset info

Number of variables	285
Number of observations	295020
Total Missing (%)	30.0%
Total size in memory	641.5 MiB
Average record size in memory	2.2 KiB

Have meaning as a measurement(**quantitative** data).

- Discrete
- Continuous

Variables types

Numeric	263
Categorical	0
Boolean	17
Date	0
Text (Unique)	0
Rejected	5
Unsupported	0

Represents characteristics such as a person's gender, marital status etc,(**qualitative** data).

The extent to which a distribution **differs** from a **normal distribution**.

- positively skewed distribution
- negatively skewed distribution

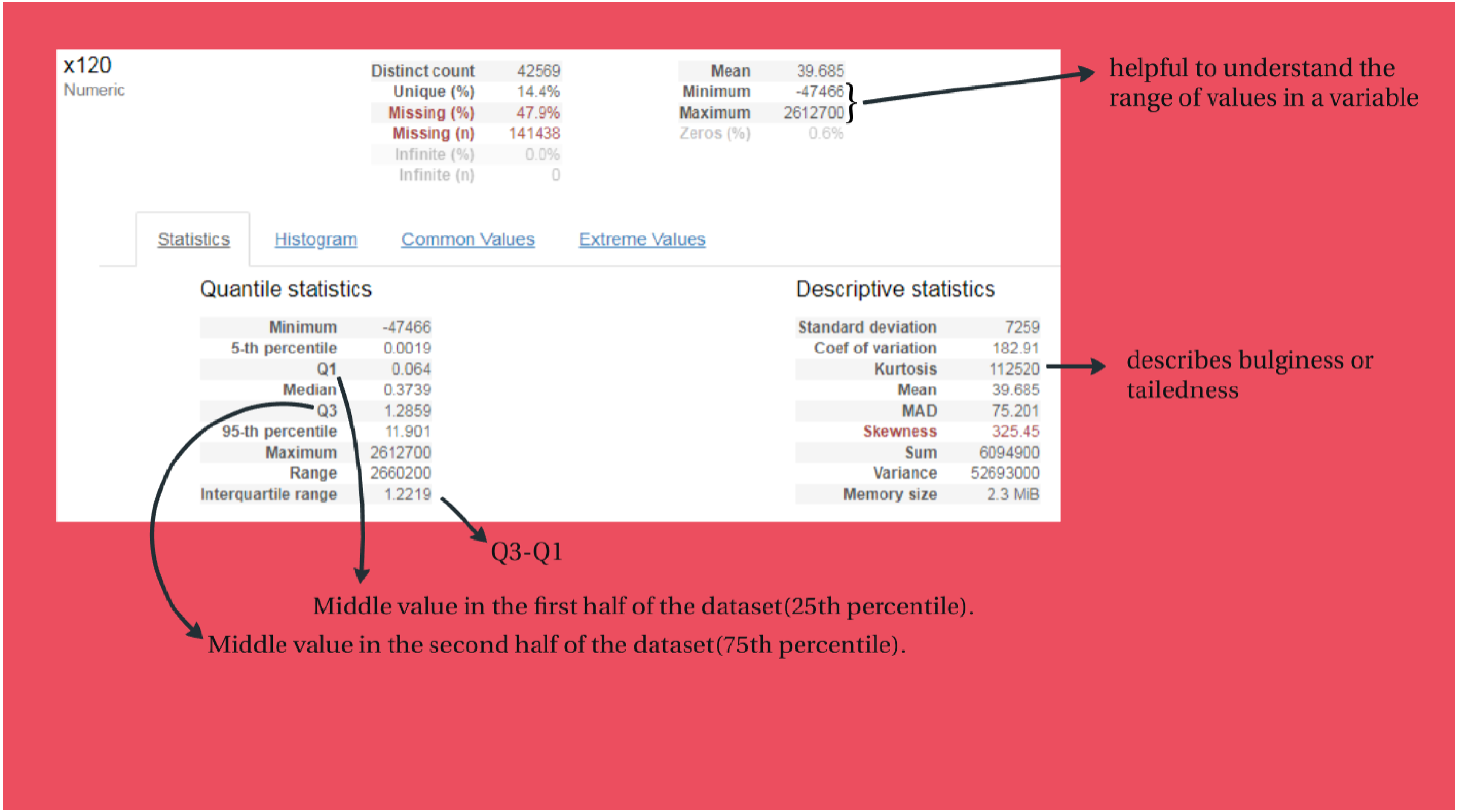
NaN values(Can be removed by **imputing** the missing values)- mean or median imputation

x120 is highly skewed (γ1 = 325.45) Skewed
x120 has 141438 / 47.9% missing values Missing
x121 is highly correlated with x120 (ρ = 0.9381) Rejected
x125 has 3980 / 1.3% zeros Zeros

~~x121~~
Highly correlated

This variable is highly correlated with x120 and should be ignored for analysis

Correlation 0.9381



x120
Numeric
Distinct count 42569
Unique (%) 14.4%
Missing (%) 47.9%
Missing (n) 141438
Infinite (%) 0.0%
Infinite (n) 0
Mean 39.685
Minimum -47466
Maximum 2612700
Zeros (%) 0.6%
helpful to understand the range of values in a variable
Statistics
Histogram
Common Values
Extreme Values
Quantile statistics
Minimum -47466
5-th percentile 0.0019
Q1 0.064
Median 0.3739
Q3 1.2859
95-th percentile 11.901
Maximum 2612700
Range 2660200
Interquartile range 1.2219
Descriptive statistics
Standard deviation 7259
Coef of variation 182.91
Kurtosis 112520
Mean 39.685
MAD 75.201
Skewness 325.45
Sum 6094900
Variance 52693000
Memory size 2.3 MiB
describes bulginess or tailedness
Q3-Q1
Middle value in the first half of the dataset(25th percentile).
Middle value in the second half of the dataset(75th percentile).

CONFUSION MATRIX KS SCORE

A confusion matrix gives a better idea of what the classification model is getting right and what **types of errors** it is making.

The KS Test is a **Goodness of Fit** Test.

CONFUSION MATRIX

KS SCORE

Random Forest

	Class 1 Predicted	Class 2 Predicted
Class 1 Actual	275024	205
Class 2 Actual	18300	170

Logistic Regression

	Class 1 Predicted	Class 2 Predicted
Class 1 Actual	275229	0
Class 2 Actual	18470	0

Class 1 - 0's
Class 2 - 1's

- **True Positive (TP)** : Observation is 0, and is predicted to be 0.
- **False Negative (FN)** : Observation is 1, but is predicted 0.
- **True Negative (TN)** : Observation is 1, and is predicted to be 1.
- **False Positive (FP)** : Observation is 0, but is predicted 1.

- **Accuracy**
- **Error rate**
- **True Positive Rate**
- **True Negative Rate**
- **False Negative Rate**
- **False Positive Rate**

Precision: When it predicts 0, how often is it correct?
= TP/(TP+FP)

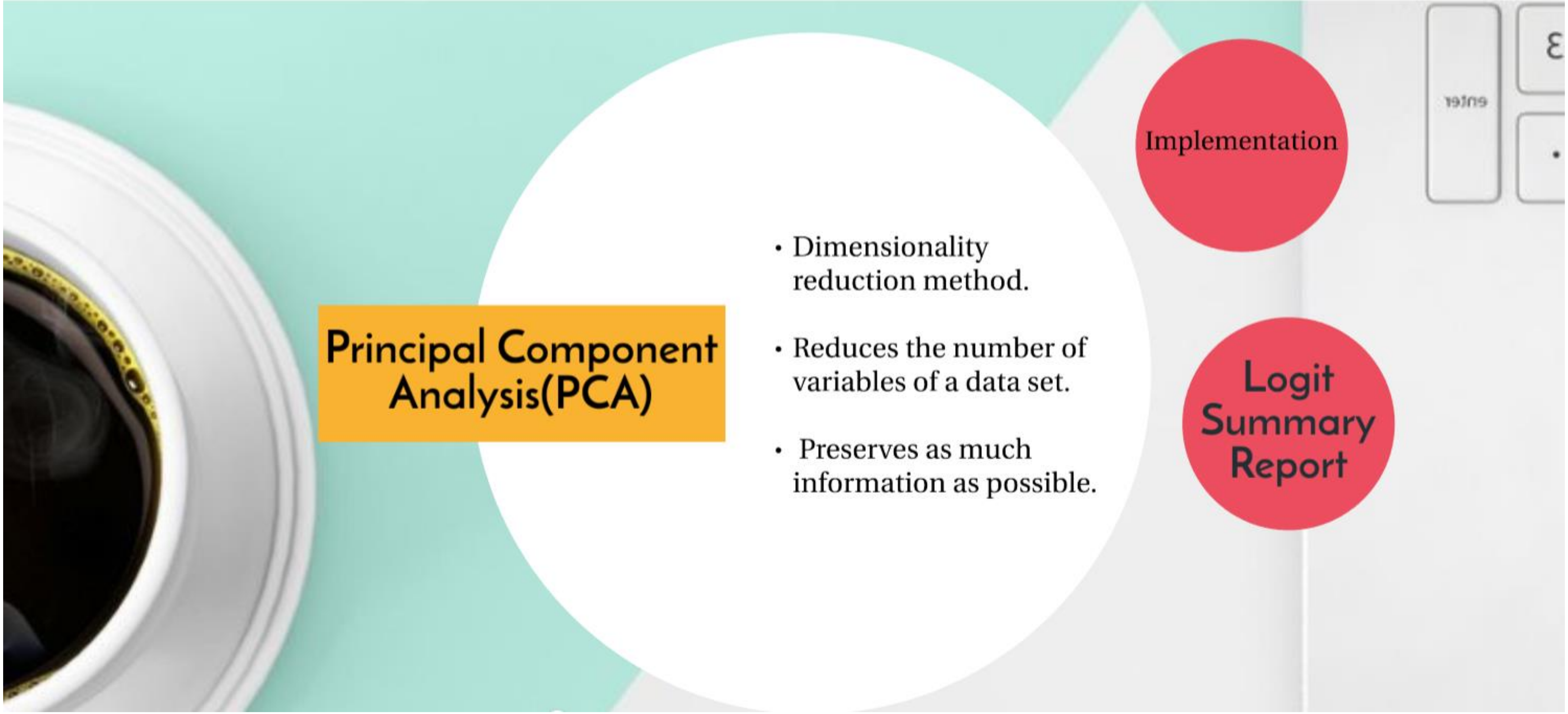
Prevalence: How often does the 0 condition actually occur in our sample?
= actual 0/total

F-Measure = (2*recall*precision)/(recall+precision)

decile	min_scr	max_scr	bad	goods	total	bad_rate	cumbad	good_rate	cumgoods	ks
0	0.089650	0.252964	5299	24071	29370	0.286898	0.286898	0.087458	0.087458	0.199440
1	0.074313	0.089650	2974	26396	29370	0.161018	0.447916	0.095906	0.183364	0.264552
2	0.063608	0.074313	2195	27175	29370	0.118841	0.566757	0.098736	0.282100	0.284657
3	0.055081	0.063608	1899	27471	29370	0.102815	0.669572	0.099811	0.381911	0.287661
4	0.046968	0.055081	1617	27753	29370	0.087547	0.757120	0.100836	0.482747	0.274373
5	0.039242	0.046967	1363	28006	29369	0.073795	0.830915	0.101755	0.584502	0.246413
6	0.033738	0.039242	1156	28214	29370	0.062588	0.893503	0.102511	0.687013	0.206490
7	0.028903	0.033737	1018	28352	29370	0.055116	0.948619	0.103012	0.790026	0.158594
8	0.022953	0.028903	634	28736	29370	0.034326	0.982945	0.104408	0.894433	0.088512
9	0.008636	0.022953	315	29055	29370	0.017055	1.000000	0.105567	1.000000	0.000000

Ks = Cumulative % Event- Cumulative % Non-Event

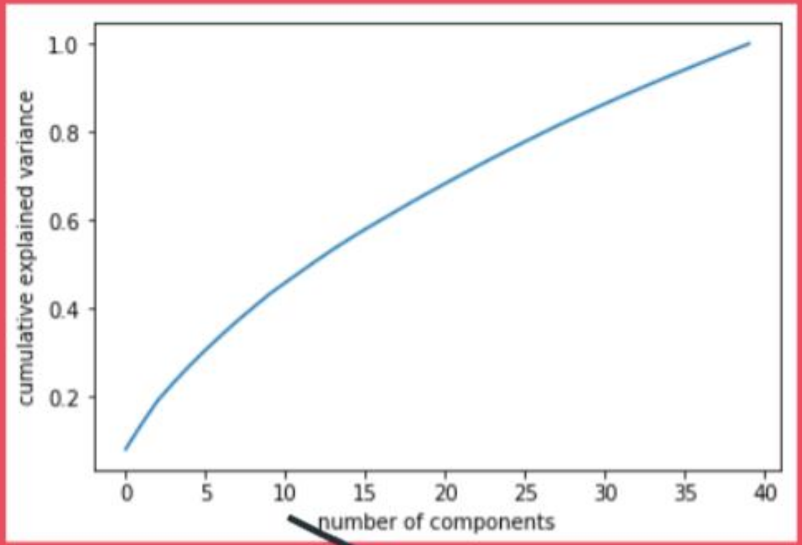
It is essential for the max KS Score to be a part of the first 3 deciles.

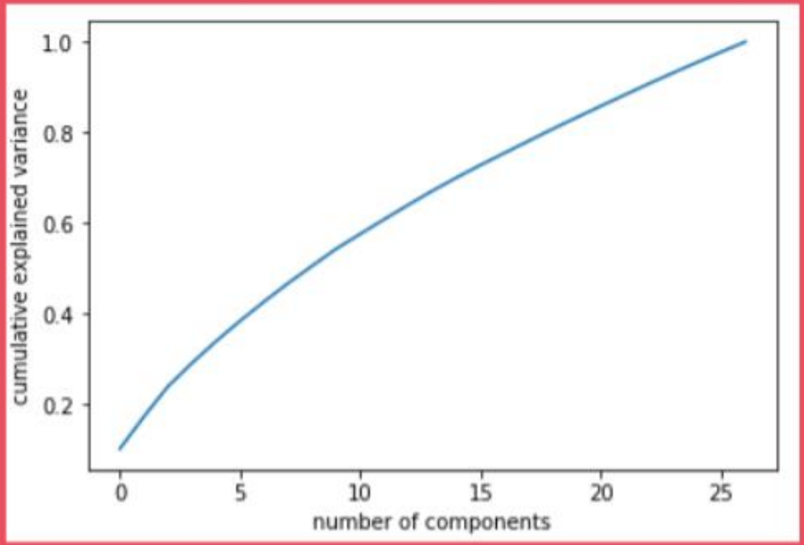

Principal Component Analysis(PCA)
• Dimensionality reduction method.
• Reduces the number of variables of a data set.
• Preserves as much information as possible.
Implementation
Logit Summary Report

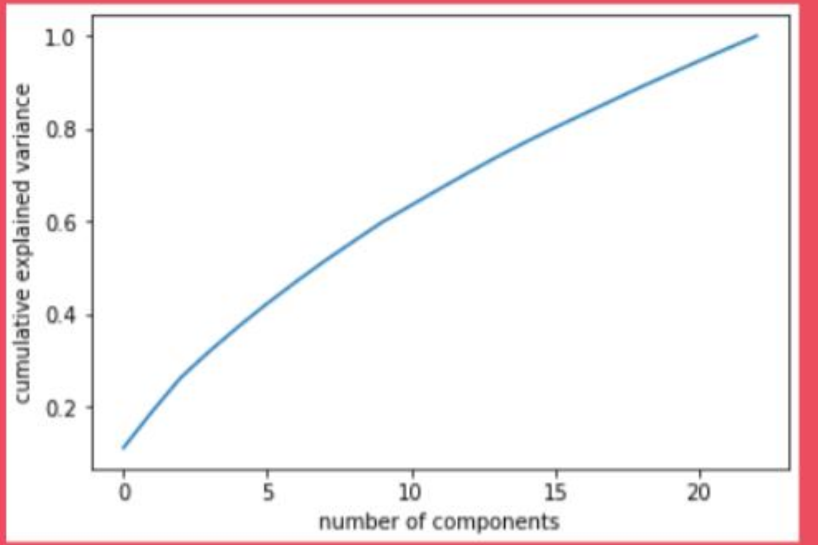
```
Explained Variance  [9.99551972 7.0276068  6.58436534 5.09290934 4.74719725 4.44374507
 4.20370786 4.02006222 3.7777198  3.72907838 3.22468744 3.20358061
 3.13814442 3.04581342 2.87192644 2.74403792 2.63544861 2.61045328
 2.60665216 2.49402689 2.47400078 2.43334731 2.42395064 2.35002955
 2.28312841 2.2517985  2.22177336 2.19417243 2.10816432 2.03697764
 2.02117955 1.98906683 1.92823992 1.90460071 1.88617449 1.8452145
 1.83942556 1.820248   1.78015372 1.76821275]


Explained Variance Ratio  [0.03582611 0.02518847 0.0235998  0.01825409 0.01701499 0.01592735
 0.015067   0.01440878 0.01354017 0.01336583 0.01155798 0.01148233
 0.01124779 0.01091686 0.01029361 0.00983523 0.00944602 0.00935643
 0.00934281 0.00893913 0.00886736 0.00872165 0.00868797 0.00842302
 0.00818323 0.00807094 0.00796332 0.00786439 0.00755612 0.00730097
 0.00724435 0.00712925 0.00691123 0.0068265  0.00676046 0.00661365
 0.0065929  0.00652416 0.00638046 0.00633766]


Cummulative Sum  [  9.99551972  17.02312653  23.60749187  28.7004012   33.44759845
  37.89134352  42.09505138  46.1151136   49.8928334   53.62191178
  56.84659922  60.05017983  63.18832426  66.23413768  69.10606412
  71.85010203  74.48555065  77.09600393  79.7026561   82.19668299
  84.67068376  87.10403107  89.52798171  91.87801126  94.16113967
  96.41293818  98.63471154 100.82888396 102.93704828 104.97402592
 106.99520548 108.98427231 110.91251223 112.81711295 114.70328744
 116.54850193 118.38792749 120.20817549 121.98832921 123.75654196]
```

```
Cumulative Sum  [ 9.99364422 17.00298478 23.58113182 28.67269935 33.41935356 37.85697215
 42.06067598 46.08143886 49.86030488 53.59093908 56.81593119 60.0190498
 63.15966631 66.21029096 69.0712568  71.81342894 74.44756237 77.06467706
 79.64497486 82.14614715 84.62336977 87.07038363 89.43498709 91.76719681
 94.04959767 96.29701493 98.50443399]
```

- Import the dataset.
- Scale the data.
- Apply PCA. Take the initial number of components as 40.
- Calculate the **explained variance**, **explained variance ratio** and **cumulative sum**.
- Consider the elements whose **cumulative sum** adds upto **100**(27 components).
- Again perform PCA on these components and calculate the explained variance, explained variance ratio and cumulative sum .
- Fit a logistic regression model and find the ROC-AUC Score.
- Get the Logit Summary report.







First 10 components contain approximately 40% of the variance

ROC-AUC Score

	Train	Test	Diff
282	0.70	0.56	0.14
40	0.67	0.47	0.20
27	0.64	0.57	0.07
23	0.70	0.59	0.11

100% of the variance

90%

```
                           Logit Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:               236016
Model:                          Logit   Df Residuals:                   235988
Method:                           MLE   Df Model:                           27
Date:                Fri, 17 May 2019   Pseudo R-squ.:                  -2.091
Time:                        16:34:21   Log-Likelihood:            -1.6319e+05
converged:                       True   LL-Null:                       -52795.
                                        LLR p-value:                     1.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
x1             0.0031      0.001      2.415      0.016       0.001       0.006
x2            -0.0080      0.002     -5.005      0.000      -0.011      -0.005
x3             0.0101      0.002      6.235      0.000       0.007       0.013
x4             0.0054      0.002      2.916      0.004       0.002       0.009
x5            -0.0163      0.002     -8.585      0.000      -0.020      -0.013
x6             0.0044      0.002      2.262      0.024       0.001       0.008
x7             0.0008      0.002      0.374      0.709      -0.003       0.005
x8            -0.0095      0.002     -4.597      0.000      -0.014      -0.005
x9             0.0027      0.002      1.195      0.232      -0.002       0.007
x10            0.0152      0.002      7.067      0.000       0.011       0.019
x11            0.0119      0.002      5.145      0.000       0.007       0.016
x12            0.0224      0.002      9.535      0.000       0.018       0.027
x13            0.0314      0.002     13.395      0.000       0.027       0.036
x14            0.0019      0.002      0.801      0.423      -0.003       0.007
x15            0.0188      0.002      7.639      0.000       0.014       0.024
x16           -0.0279      0.003    -11.063      0.000      -0.033      -0.023
x17            0.0077      0.003      2.669      0.008       0.002       0.013
x18            0.0069      0.003      2.597      0.009       0.002       0.012
x19           -0.0132      0.003     -4.584      0.000      -0.019      -0.008
x20           -0.0062      0.003     -2.263      0.024      -0.012      -0.001
x21           -0.0046      0.003     -1.625      0.104      -0.010       0.001
x22            0.0010      0.003      0.380      0.704      -0.004       0.006
x23           -0.0098      0.003     -3.475      0.001      -0.015      -0.004
x24           -0.0048      0.003     -1.649      0.099      -0.010       0.001
x25            0.0059      0.003      2.006      0.045       0.000       0.012
x26           -0.0105      0.003     -3.499      0.000      -0.016      -0.005
x27            0.0040      0.003      1.278      0.201      -0.002       0.010
x28            0.0019      0.003      0.671      0.502      -0.004       0.008
```

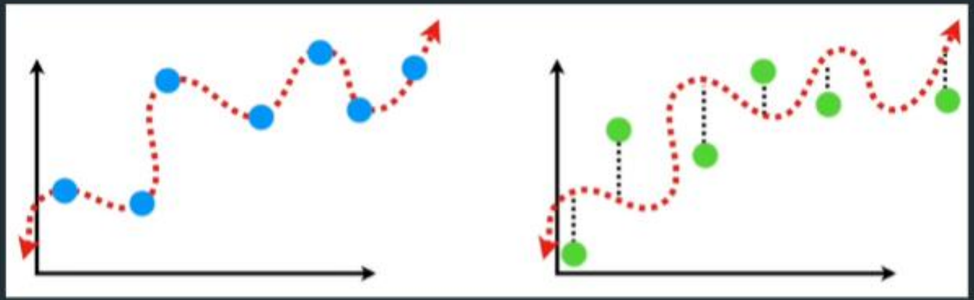
z & p-value : Provide the z-value and 2-tailed p-value used in testing the null hypothesis that the coefficient (parameter) is 0. Coefficients having p-values less than alpha are statistically significant (i.e., you can reject the null hypothesis and say that the coefficient is significantly different from 0).

Log-Likelihood : maximized value of the log-likelihood function. Used to help compare nested models.

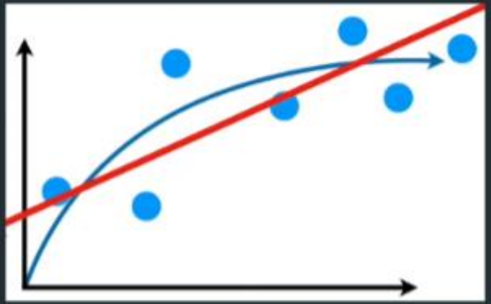
LL-Null : result of the maximized log-likelihood function when only an intercept is included

coef : Values for the logistic regression equation for predicting the dependent variable from the independent variable. They are in log-odds units.

std err : These are the standard errors associated with the coefficients. The standard error is used for testing whether the parameter is significantly different from 0; by dividing the parameter estimate by the standard error you obtain a z-value.



Bias - Refers to an estimator that is too general and does not learn relationships from a data set that would allow it to make better predictions.



Variance - Learning relationships that are specific to the training set but will not generalize to new observations well.

High Bias - Underfitting
High Variance - Overfitting

The bias-variance tradeoff is the tradeoff between underfitting and overfitting.

Bias variance tradeoff is to get an **optimal bias and variance** for the model- Should be general enough to make good predictions on new data but specific enough to pick up as much signal as possible.

How is it accomplished?

- Reduce dimensionality of the data
- Ensemble methods

ENSEMBLE MODELS

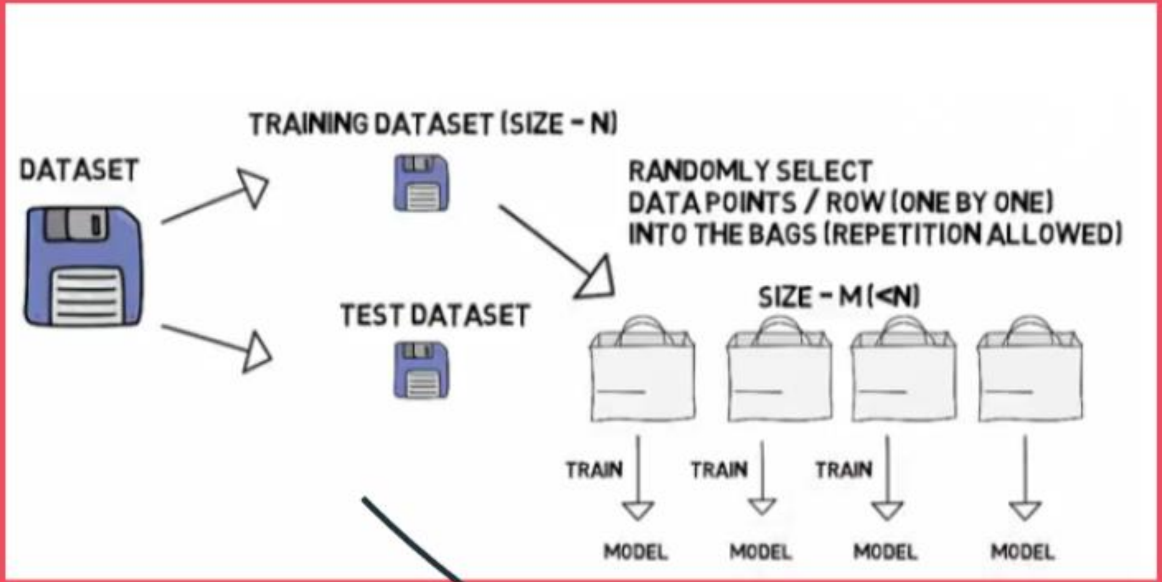
- Better accuracy
- Avoids overfitting
- Reduces bias and variance errors

BAGGING

BOOSTING

RANDOM FOREST

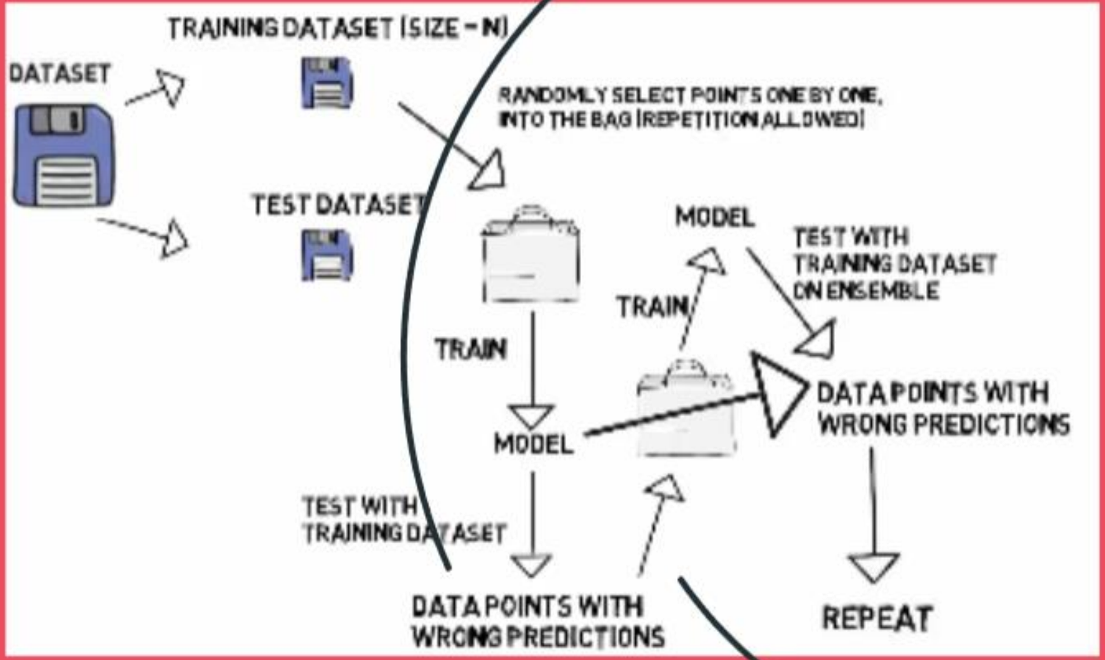
BAGGING



- Split the dataset into training and testing.
- Select subsets of training dataset into bags.
- Take the vote of their output.

Multiple models of same learning algorithm trained with subsets of dataset, randomly picked from the training data.

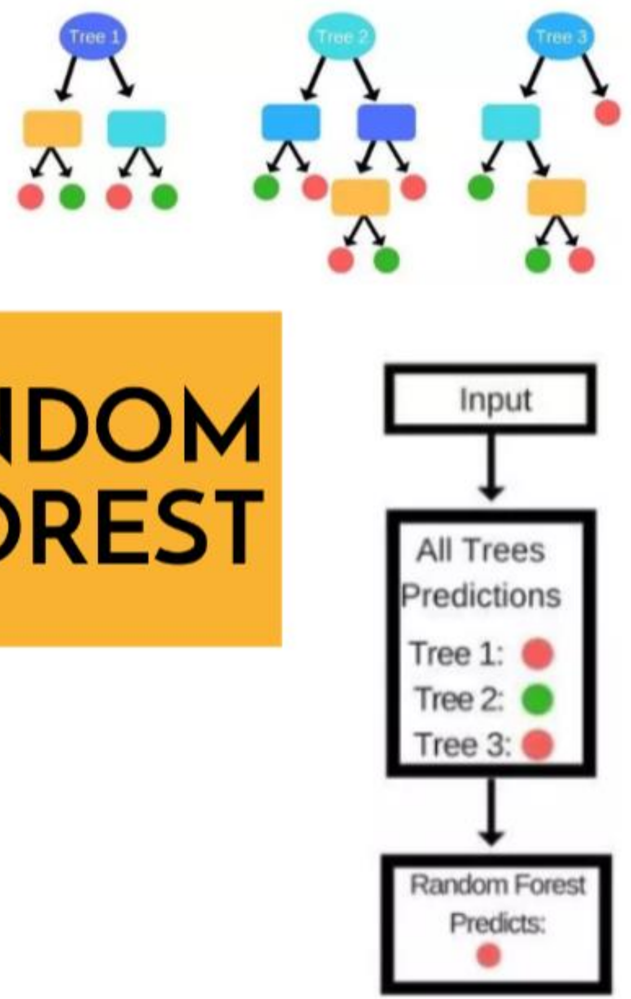
BOOSTING



Selecting datapoints which give wrong predictions in order to improve accuracy.

Results in huge accuracy but also tends to overfit and increase the variance.

RANDOM FOREST



FEATURE SELECTION

FEATURE SELECTION

What is it?

Process of automatically or manually selecting those features which contribute the most to the prediction variable or output.

Why do we use it?

- Reduces Overfitting
- Improves accuracy
- Reduces training time

IMPLEMENTATION

```
Index(['x1', 'x2', 'x3', 'x4', 'x5', 'x6', 'x7', 'x8', 'x10', 'x16', 'x17',
       'x19', 'x20', 'x26', 'x27', 'x32', 'x33', 'x35', 'x36', 'x37', 'x41',
       'x42', 'x43', 'x48', 'x54', 'x55', 'x56', 'x61', 'x72', 'x82', 'x84',
       'x85', 'x86', 'x87', 'x88', 'x89', 'x90', 'x91', 'x92', 'x95', 'x96',
       'x98', 'x100', 'x101', 'x102', 'x104', 'x105', 'x117', 'x118', 'x119',
       'x120', 'x121', 'x125', 'x147', 'x148', 'x149', 'x150', 'x167', 'x168',
       'x170', 'x172', 'x175', 'x176', 'x177', 'x184'],
      dtype='object')
```

	variable	importance
0	5	0.041856
1	27	0.037222
2	23	0.033504
3	47	0.033305
4	19	0.033206
5	22	0.030952
6	1	0.029930
7	11	0.025000
8	29	0.024801
9	18	0.023474
10	42	0.023166
11	13	0.022217
12	0	0.022173
13	33	0.022097
14	56	0.021227
15	21	0.020935

49	44	0.008158
50	49	0.008122
51	46	0.008014
52	6	0.007696
53	58	0.007535
54	28	0.007404
55	10	0.007155
56	43	0.007063
57	31	0.006799
58	62	0.006584
59	25	0.006370
60	8	0.006139
61	61	0.005580
62	52	0.005446
63	7	0.004758
64	15	0.004573

```
d=np.cumsum(b)
(d<=0.9999).argmin()
```

```
64
```

- Import the dataset Dev_Sample1.
- Fit a RandomForest Classifier model.
- Print the **selected features**.
- Print the **feature importance** for each of the **65** selected features.
- Print all the features whose **cumulative sum** adds upto **99** (64 such variables).

```
['x1', 'x2', 'x3', 'x4', 'x5', 'x6', 'x7', 'x8', 'x9', 'x10', 'x11', 'x12', 'x13', 'x14', 'x15', 'x17', 'x18', 'x19', 'x20', 'x
21', 'x22', 'x23', 'x24', 'x25', 'x26', 'x27', 'x28', 'x29', 'x30', 'x31', 'x32', 'x33', 'x34', 'x35', 'x36', 'x37', 'x38', 'x3
9', 'x40', 'x41', 'x42', 'x43', 'x44', 'x45', 'x46', 'x47', 'x48', 'x49', 'x50', 'x51', 'x52', 'x53', 'x54', 'x55', 'x56', 'x5
7', 'x58', 'x59', 'x60', 'x61', 'x62', 'x63', 'x64', 'x65']
```

	Train	Test	Diff	
282	95.63	17.70	77.94	
65	95.62	14.87	80.76	
64	95.62	16.28	79.34	99% of variance
49	95.61	16.48	79.13	90%
39	95.5611	15.8476	79.71	80%
30	95.5611	15.8476	79.71	70%
24	95.4879	15.9689	79.52	60%
18	94.9697	15.3449	79.62	50%
13	94.7503	13.7908	80.96	40%

Fit a RandomForest model for the 64 variables(**training set**).

For the **test set**, import the OOT_Sample and consider the same 64 variables in this dataset.

Finally, find the **KS train** and **KS test** score.

Repeat the above steps for various % of variance.

HYPERPARAMETER TUNING

- n_estimators
- min_samples_leaf
- min_samples_split
- max_depth

SEQUENTIAL SEARCH

n_estimators = 10
KS Train = 95.6159415
KS Test = 16.27516194

n_estimators = 50
KS Train = 95.6525185
KS Test = 24.04592496

n_estimators = 100
KS Train = 95.6525185
KS Test = 25.27075898

n_estimators = 150
KS Train = 95.6525185
KS Test = 26.1604969

n_estimators = 200
KS Train = 95.6525185
KS Test = 26.59958834

n_estimators = 250
KS Train = 95.597653
KS Test = 19.5740889

max_depth = 5
KS Train = 28.65076584
KS Test = 27.34485196

max_depth = 10
KS Train = 39.43488532
KS Test = 28.98570368

max_depth = 15
KS Train = 63.20627907
KS Test = 29.14166036

max_depth = 20
KS Train = 85.07566869
KS Test =28.14792709

max_depth = 25
KS Train = 94.55520843
KS Test = 27.43155107

max_depth = 30
KS Train = 34.20437401
KS Test = 24.35787465

min_samples_split = 0.01
KS Train = 35.50041912
KS Test = 28.84704322

min_samples_split = 0.03
KS Train = 31.24529453
KS Test = 28.47146815

min_samples_split = 0.05
KS Train = 29.65419493
KS Test = 27.61061783

min_samples_split = 0.1
KS Train = 27.95946049
KS Test = 26.56488689

min_samples_split = 0.3
KS Train = 25.53928218
KS Test = 24.21921419

min_samples_split = 0.5
KS Train = 24.42977978
KS Test = 23.60679718

min_samples_leaf = 0.01
KS Train = 29.86146461
KS Test = 27.46040233

min_samples_leaf = 0.1
KS Train = 25.21618533
KS Test = 24.08055374

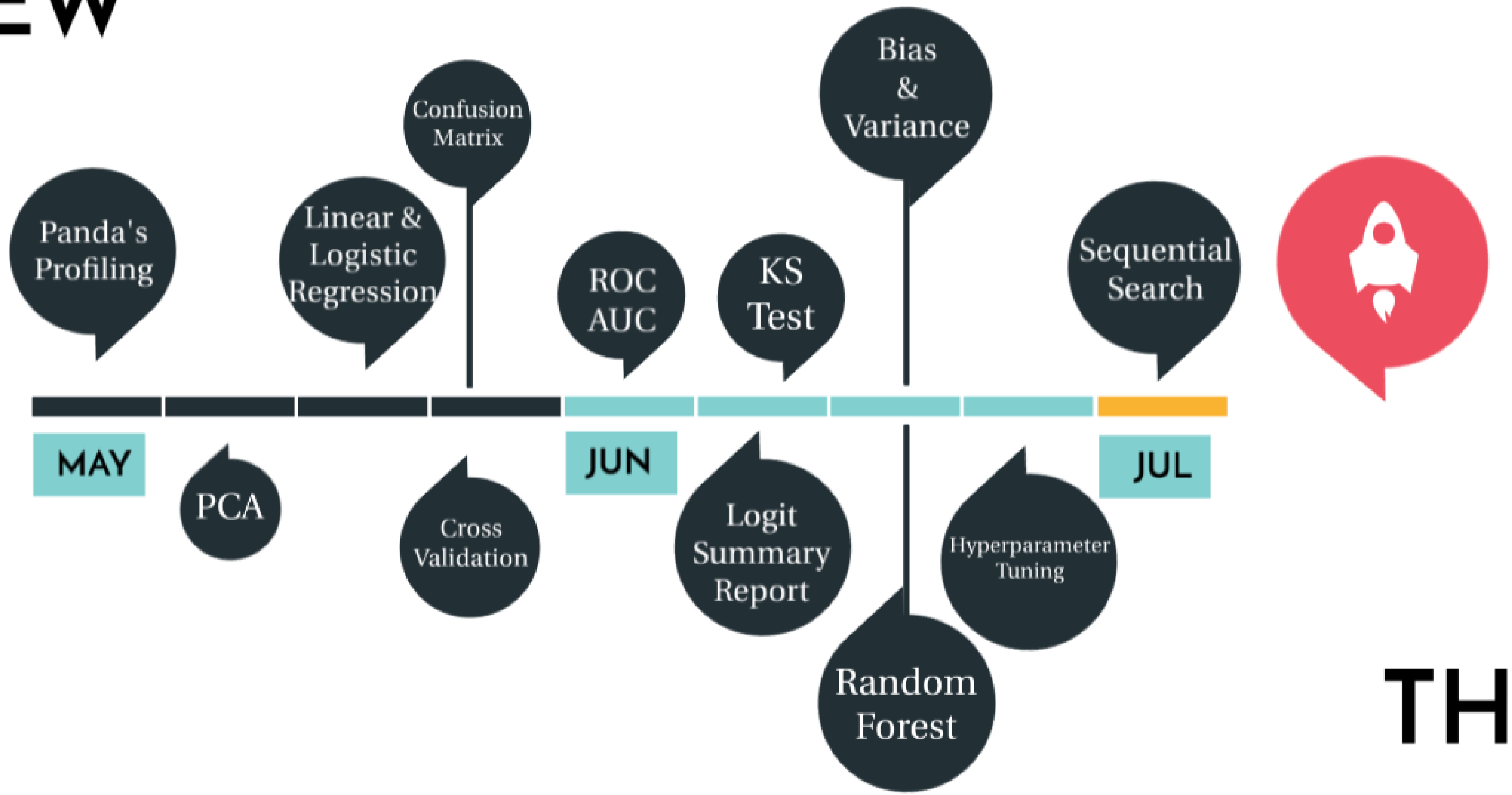
min_samples_leaf = 0.05
KS Train = 26.45736494
KS Test = 25.20135608

min_samples_leaf = 5
KS Train = 34.7079174
KS Test = 28.93948352

min_samples_leaf = 7
KS Train = 34.67743657
KS Test = 29.00877742

min_samples_leaf = 10
KS Train = 34.65914807
KS Test = 28.80082307

OVERVIEW



THANK YOU