**Questions & Answers**

1. **[Python] Difference between isnull and isna**

isnull = isna

pandas.**isnull**(*obj*)

Detect missing values (NaN in numeric arrays, None/NaN in object arrays)

pandas.**isna**(*obj*)

Detect missing values (NaN in numeric arrays, None/NaN in object arrays)

[**pandas.isnull**](https://pandas.pydata.org/pandas-docs/stable/generated/pandas.isnull.html#pandas.isnull)

alias of isna

1. **[Statistic] Why variance is taking square of difference between mean and the value**

<http://www.leeds.ac.uk/educol/documents/00003759.htm>

1. **[Python, R] What is apply function**

<https://www.datacamp.com/community/tutorials/r-tutorial-apply-family>

<https://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.apply.html>

1. **[Python] what is axis=0**

<https://stackoverflow.com/questions/22149584/what-does-axis-in-pandas-mean/22149930>

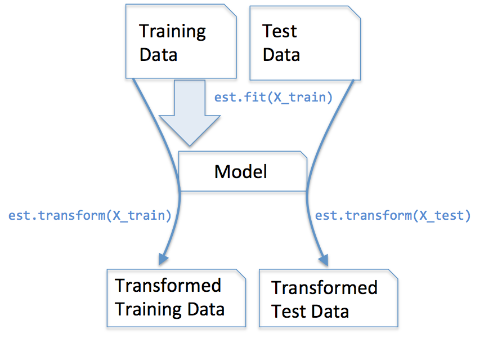
1. **[Python] What is difference between fit\_transform and transform method?**

In scikit-learn estimator api,

**fit()** : used for generating learning model parameters from training data

**transform() :** parameters generated from fit() method,applied upon model to generate transformed data set.

**fit\_transform()** : combination of fit() and transform() api on same data set



<https://datascience.stackexchange.com/questions/12321/difference-between-fit-and-fit-transform-in-scikit-learn-models>

1. **[Python] What is random\_state?**

<https://stackoverflow.com/questions/28064634/random-state-pseudo-random-numberin-scikit-learn>

But I am still wondering why we get different result first time.??

1. **[ML] What, when, where and how to use feature scaling and its impact**

<https://en.wikipedia.org/wiki/Feature_scaling>

<https://www.quora.com/When-should-you-perform-feature-scaling-and-mean-normalization-on-the-given-data-What-are-the-advantages-of-these-techniques>

<https://medium.com/greyatom/why-how-and-when-to-scale-your-features-4b30ab09db5e>

<http://dungba.org/why-should-we-implement-feature-scaling-mostly-all-the-time/>

1. **[ML, Python] What is ‘C’ parameter in SVM.**

**C** : float, optional (default=1.0)

Penalty parameter C of the error term.

The C parameter trades off misclassification of training examples against simplicity of the decision surface. A low C makes the decision surface smooth, while a high C aims at classifying all training examples correctly by giving the model freedom to select more samples as support vectors.

<https://stats.stackexchange.com/questions/31066/what-is-the-influence-of-c-in-svms-with-linear-kernel>

1. **Notes form Meetup 5.**

* ggplot - all featuires against all the featues..
  + Grey area denotes confidence interval
  + Dark line denotes trend
* Individual feature vs feature correlation assessment helps find high collinearity which will help remove the redundancy of features. Feature vs dependent variable correlation assessment helps find significance of the feature to the model; higher the correlation, more does it impact the dependent variable and higher is the need for it to be kept in the final set of features.
* After finding feature vs feature correlation, there are various techniques to resolve it. VIF (Variance inflation factors (VIF) is an example.
* VIF to solve multi collinearity. Generally used for problems with few features. PCA, LDA etc techniques are generally used for problems with a lot of features.
* Tools like Gretl, Knit, AzureML, SAS help build a quick model to assess the initial results and improve upon the model.
* Azure.notebook is a good cloud based option.
* While evaluating backward elimination and removing the features..check not only p value but also f value.
* AIC (akaike) and BIC works for classification problem similar to R-sq for regression to carry out model performance evaluation.