# CONFIDENCE BASED ON TREE VARIANCE

It is basically the VARIANCE of the observations calculated across all the trees for each row of the dataset.

Suppose predicted values are -> 6 7 8 9 7 8 , so the mean is 45/6= 7.5.

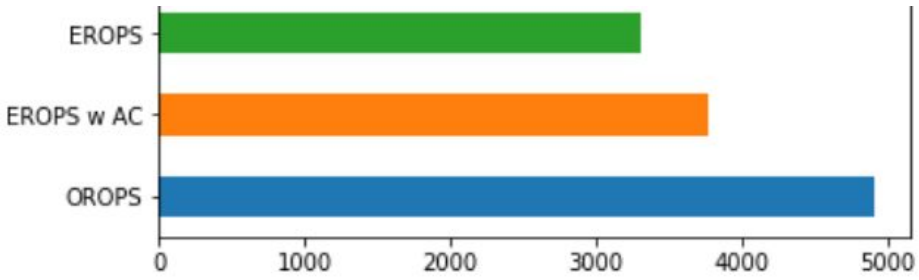
Now variance for the row would be $((6-7.5)^2+ (7-7.5)^2+... )/ 6)$.

This would be the variance for predictions of A ROW IN YOUR DATASET.

- **Why is it useful?**

  -If the variance of a particular group in a variable is **HIGH** then we can claim that **WE DIDN'T HAVE ENOUGH OBSERVATIONS OF THAT PARTICULAR GROUP TO TRAIN WITH AND THUS WE HAVE HIGH VARIANCE.**

  **SEE BELOW IMAGE**




[28]:
```
fld = ["Enclosure","SalePrice","preds_mean","preds_std"]
enc_summary = x[fld].groupby("Enclosure",as_index = False).mean()
enc_summary
```

[28]:

| | Enclosure | SalePrice | preds_mean | preds_std |
|---|---|---|---|---|
| 0 | EROPS | 9.849178 | 9.845488 | 0.271443 |
| 1 | EROPS AC | NaN | NaN | NaN |
| 2 | EROPS w AC | 10.623971 | 10.577112 | 0.265684 |
| 3 | NO ROPS | NaN | NaN | NaN |
| 4 | None or Unspecified | NaN | NaN | NaN |
| 5 | OROPS | 9.682064 | 9.688057 | 0.223017 |

In the above image the OROPS category of Enclosure has the highest number of observations and thus it is noted that its STANDARD DEVIATION OF PREDICTIONS IS LOWEST.

- **For a single row we can look at this example->**

>We take a single row and run it through the trees and look at how **CONFIDENT** we are for that particular row in our forest. If the variance is high, then we will assume that that row **DOES NOT HAVE A VERY CONFIDENT PREDICTION** pertaining to our forest.

Another example can be like-> You are predicting the **CREDIT RISK EVALUATION.**

**Suppose you model it as a RISK between 0->1.**

**Now 0 is least and 1 is the highest risk.**

**So suppose your model gives you an OVERALL low risk for a particular person but the VARIANCE is high.**

**That means that the person is a LOW RISK BUT YOU ARE NOT SURE AS IT HAS HIGH VARIANCE.**

**This is where variance of predictions comes into play.**