Data Science Notes

# SNS Heat maps and Correlations

1. Correlation maps only tell you how much two variables are linearly correlated which indicates that a direct relationship between them may exist. However, this could not be true. Remember that correlation does not imply causation.

2. As a rule of thumb, you should remove highly correlated features.

3. You define the threshold for the correlation based on experimentation and domain knowledge. In some cases, features that are correlated > 0.5 can be considered highly correlated and in some cases, 0.9 is a good threshold.

4. You want to remove features that are highly correlated.

This can lead to overfitting since model can't distinguish between highly correlated features but is using them to train (so basically same feature is given more importance than required)

5. You want to maximize correlation between features and the target variable though, higher the correlation, more useful that feature is, in order to predict the target variable.

6. Heatmaps are a quick way to look at 1 on 1 relationships.

For small number of features, you might be able to look at multi-variable linear dependencies. Again, such features should be removed too but they are hard to detect from heatmaps.

7. Unsupervised learning techniques can help look at the multi-variable dependencies. Techniques like PCA which makes sure the features are orthogonal means that they are independent of each other.

8. Dimensionality reduction based unsupervised learning techniques are not making predictions, they are helping reduce/ remove the linear dependency among the features. Some techniques do help make predictions based on the feature separation (Like LDA).