Class 3 – EDA -1st Sept 2018

Project 4 dataset - considerations

1. Shape of data – minimum 1500 rows 10 useable columns (room for manoeuvre)

* Avoid ‘curse of dimensionality’ – ‘tall’ and ‘slim’ (not too few features)

1. Types of features that describe problem statement
   1. Avoid features that have the same value for all rows
2. Even sampling?
3. Kaggle data + problem statements
   1. Can use Kaggle dataset to solve another problem statement

EDA

Intra-feature corr. – 0.8 is red flag, 0.6 is tolerable, depending on number of features you can afford to drop.

Class Imbalance:

Jitesh Khurkhuriya – Azure ML Online Course

Smote – Tomek

Oversampling + Undersampling combined

Smote – create synthetic observations between two datapoints

<http://contrib.scikit-learn.org/imbalanced-learn/stable/auto_examples/combine/plot_smote_tomek.html>

Smote, Smoteenn, Smote-Tomek

The Art and Science of Learning from Data 4th Ed

Testing between to two groups