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

**Class 5 – 8 Sept 2018**

Project 4 – Dataset selection

Supervised- Classification, Regression

Unsupervised - Clustering

Special Topics

Natural Language Processing – Corpus (dataset?)

* Python library – Vader

NLP - Topic Modelling – group like words in ‘topics’, e.g. Google News

* Topic Keywords

Example:  
Topic 1 - Sushi, sashimi etc

Topic 2 – Beer, steak etc

Words can be repeated in multiple topics depending on association

Topic identities can be vague or abstract, depending on the resulting topic groups

Time Series – dealing with time, e.g. stock prices rising and falling

Using historical y to predict future y (autoregression)

CURSE OF DIMENSIONALITY

* More columns in dataset, the more rows you need

COMBINE Supervised with Unsupervised (Unsupervised is data prep)

<https://plot.ly/python/ipython-notebook-tutorial/#installation>