UNIVERSITY OF EDINBURGH
Business School

Predictive Analytics and Modelling of Data

CMSE11428 (2020-2021)

Dr Xuefei Lu
The University of Edinburgh Business School

Random forests

other tree-based approaches

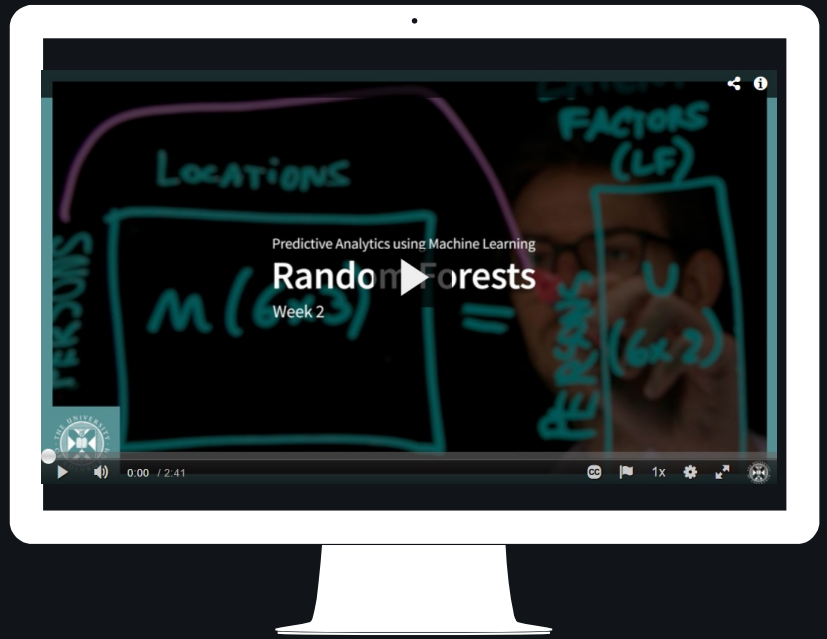
Bagging (data-based)

- 'Bootstrap aggregation'
- Rationale:
  - Create a number of bootstraps
  - Build various models for these samples
  - Average (regression)/majority vote (classification) of outcome
- Reduces variance and bias (overfitting)

Boosting (model-based)

- Use many weak learners for an aggregate prediction
  - Weak learner: classifier that barely outperforms random guess
- Decision trees are great weak learners:
  - Can be pruned to be very small (and hence underfitting)
  - Can be created to stop after being very small and shallow still
- Rationale:
  - Create a lot of 'weak' models and aggregate their results
  - Again take majority vote/average of outcome

Random Forests



Please watch the following video

https://media.ed.ac.uk/media/Random+Forests/1_5z6sh1wx/141757871

Random forests

- Applies bagging
- Combines this with randomly selecting subset of variables to use for splits:
  - Not necessarily the best split will be made
  - Nevertheless, there will be a lot of variety in the different trees
  - Counters the deterministic way decision trees are built
- Very strong predictive algorithm
- Variable importance calculated by averaging the reduction of Gini impurity of a variable across all trees

- 3 - Building different tree-based models.ipynb+ churn_ibm.csv
- 4 - Applying different tree-based models.ipynb + bank-full.csv

Activity: Building different tree-based models

UNIVERSITY OF EDINBURGH
Business School