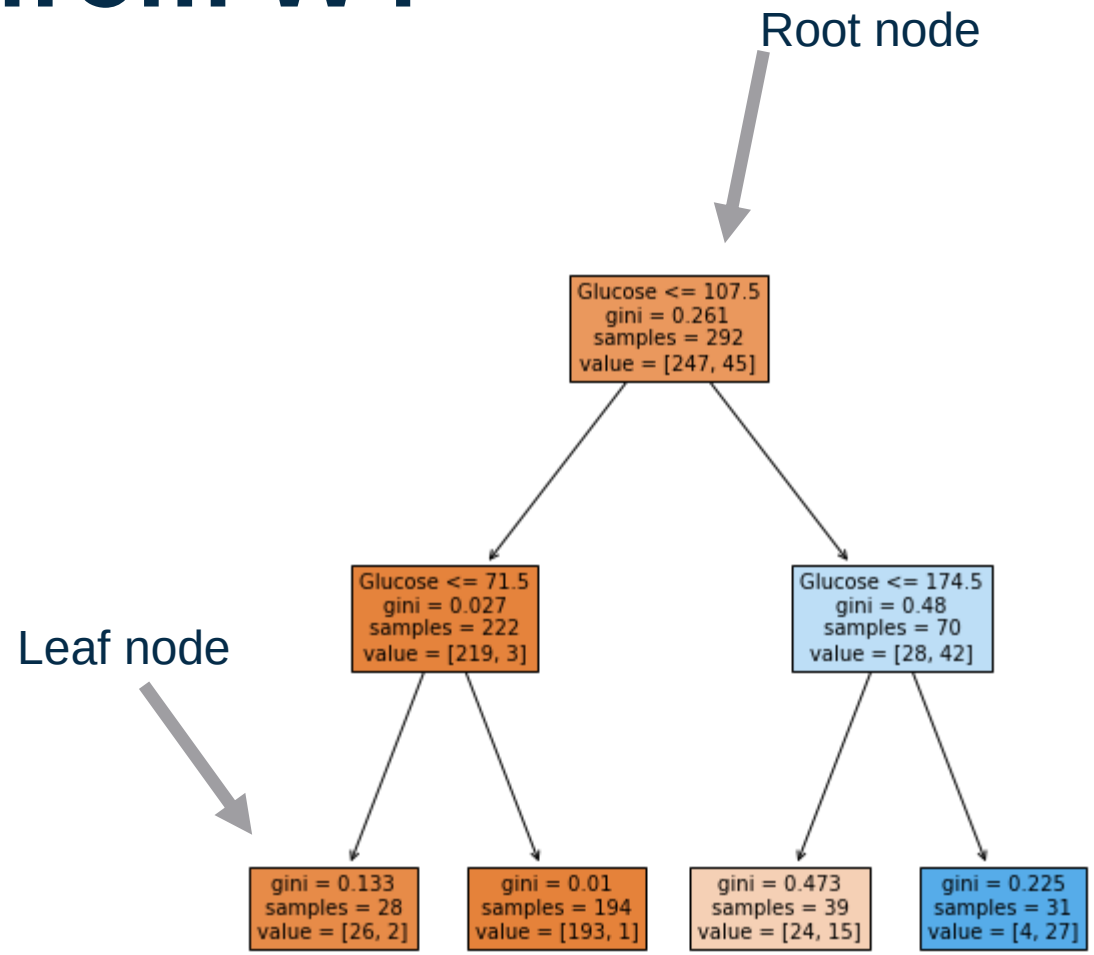
MSDS600 W5: Automation and Data Science

REGIS UNIVERSITY

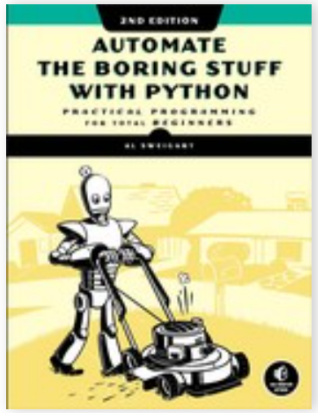
Review from W4

- Decision trees, random forests, and feature selection
- DT - Tries all possible splits with all features by default
- Gini criterion or Entropy to measure purity of splits, choose split with lowest Gini or Entropy
- RF – many DTs, adds bootstrapping (sampling with replacement) for each tree, add random subsampling of features for each split
- Feature importances – calculated from decrease in Gini or Entropy for each feature in the tree or forest; we can remove the least important features
- Correlations – can remove features with weak correlation to the target (or use in combo with feature importance)



Automation with Python

- We can put Python code in .py files (scripts) that can run repetitive tasks (e.g. generating the same report, moving files around computers, etc).
- We can also put code in functions or classes within these files to make it more organized and easier to maintain.



Automate the Boring Stuff with Python, 2nd Edition

4.5 ★★★★½ 7 REVIEWS

by Al Sweigart

Publisher: No Starch Press

Release Date: November 2019

ISBN: 9781593279929

Topic: Python

Available through the library:
https://learning.oreilly.com/library/view/automate-the-boring/9781098122584/
There are other related books from the same author and publisher

Automatic Data Science

- We can think of automating each step in the CRISP-DM or other data science process models, and we can think about automating the overall process. Turns out it's really hard!
- Every dataset is different, so data cleaning and feature engineering are hard to automate. Deployment and evaluation against business needs can also be difficult to automate.
- auto EDA, auto ML, auto AI

Automatic Data Science

- We saw how there are auto-EDA packages:
  - pandas-profiling, autoviz, sweetviz (and many others including packages in R)
- There are also auto-modeling (machine learning) packages and tech:
  - Python packages: pycaret, TPOT, H2O, more
  - Cloud solutions: Azure, AWS, GCP, IBM Watson

Machine learning takes lots of time and effort

It's best to try several different models (no free lunch theorem) and to optimize each model. To code this by hand is tedious, time-consuming, and repetitive.

Often we'll use cross-validation to score our model and optimize hyperparameters (e.g. the max_depth of decision trees). This breaks the data into equal parts (e.g. in groups of 4 for 4-fold CV), then takes each of the 4 splits separately. For each of the splits (rows, on the right), it fits to the training data and evaluates performance on the test data. Then we can average the test data score over all the folds to get an overall score.

AutoML at its base automates this selection of a best model. Some AutoML software add in other steps including data cleaning, preparation (e.g. feature engineering, encoding strings to numbers), and feature selection.

https://en.wikipedia.org/wiki/Cross-validation_(statistics)

Will autoML and autoDS replace data scientists?

- Maybe eventually, but probably not for a while.
- Data scientists should have expertise or at least a good knowledge base in a few areas (programming, statistics and ML to name a few), and based on that knowledge we can deploy the right tool for the situation.
- AutoML is just another tool to use in your toolkit.
- Data oddities will always be there, such as the ‘0’ values for tenure in the churn data that were likely missing values.

Will autoML and autoDS replace data scientists?

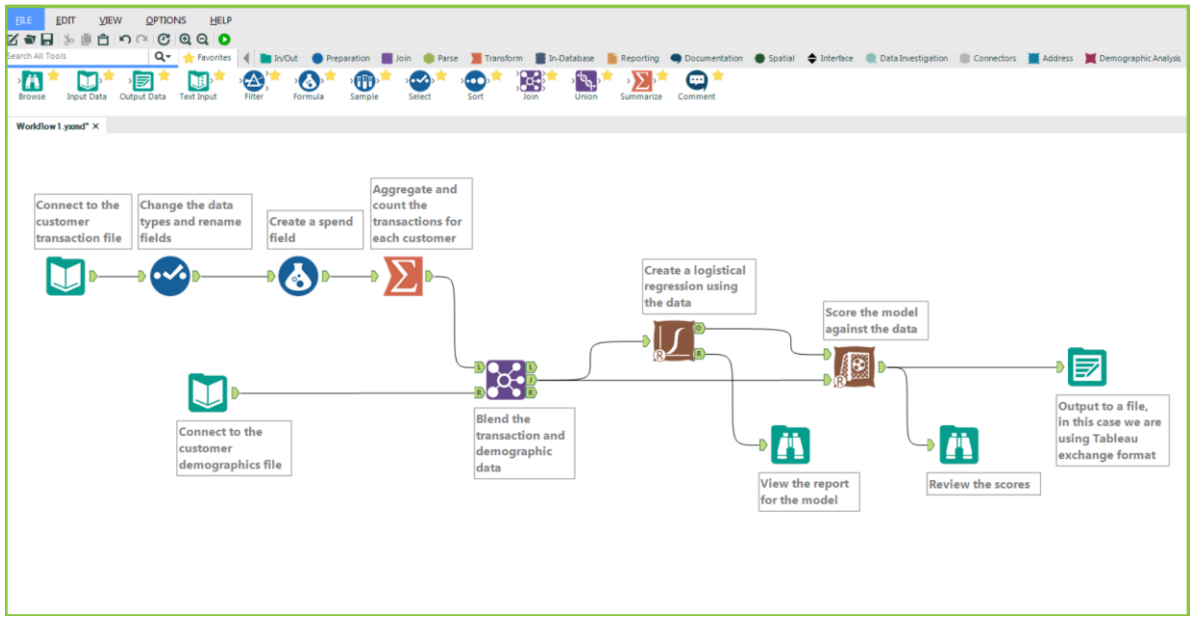
- AutoDS and ML solutions don’t always get all the details right, so it’s important to know how it’s working in order to avoid pitfalls and caveats.
- We also need to evaluate solutions against business requirements.
- Deployment of a solution and monitoring are also important, and not easy to automate with a one-size-fits-all solution.
- New DS techniques are being developed all the time, and keeping up with the latest methodology may be hard to automate.

DS GUIs

Data science GUIs can make doing data science easier with no-code solutions. Many allow use of R and Python code as well. Some GUIs are:

- Alteryx
- Orange
- RapidMiner
- Cloud solutions (AWS, GCP)
- H2O

Many of these also include autoDS and autoML features. Most of their features are not really “automating” data science but are making it easier to do without writing code. It does automate repetitive processes like a Python script. The drawbacks are that licensing can be expensive, it can be harder to integrate with other systems, and it can be harder to customize and use cutting-edge techniques.



https://blog.revolutionanalytics.com/2017/03/alteryx-integrates-with-microsoft-r.html

Cloud tools

- Google Cloud Platform (GCP) autoML suite – image, video, text, tabular data
- AWS – similar tools (e.g. Rekognition, shown on right)
- Azure – a little behind GCP and AWS at the moment
- IBM Watson – auto AI

https://aws.amazon.com/rekognition/?blog-cards.sort-by=item.additionalFields.createdDate&blog-cards.sort-order=desc

AutoDS tools for the discussion

A lot of people like to make lists of things on GitHub. Here are a few for autoML and autoEDA:

- Tools, projects, and commercial products
- AutoEDA packages

Some other automation techniques we didn't look at but are useful:

- cronjobs in linux
- Automated reporting: reportlab package in Python and others
- Automated dashboarding