Homework for invited lecture on Feature Engineering

In this homework, you will exercise a very short cycle of FE for what is known as a*feature drill-down*.

The provided tab-delimited file (`ch6_cell28_dev_feat1_filtered.tsv.gz`, 381 columns, 44960 rows including header row, 3.1Mb) is an intermediate file in a case study for FE over graph data. It contains 379 features for a regression task (plus a label column).

Over this data, use random forest feature importances (but beware scikit-learn implementation has some issues: https://explained.ai/rf-importance/index.html) and choose one feature (let's call it F) to do a drill-down by threshold the feature.

The drill-down you will be doing, therefore, is to produce a new feature F' that is 0 or 1 depending whether the value of the original feature F is below a threshold you have to pick.

With F', you can choose to add F' to the feature vector or replace F with F'. Evaluate the new feature vector and decide whether the new feature helps or not.

Be particularly careful on the process you follow to pick the threshold and to evaluate the impact of the new feature.

Deliverables:

- A Jupyter notebook showing your work.
- A write-up of up to 200 words (you will be penalized if you write more than 200 words!) concluding whether the new feature worked or not and why you think that is the case. Consider also future recommendations (further drill-down, drop, discretize, etc).