Homework 3

You can submit in groups of 2.
Due 3/27, 1pm.

All assignments need to be submitted via github classroom:
https://classroom.github.com/g/keeKxv7A
and as PDF via gradescope.

The goal of this homework is to provide a realistic setting for a machine learning task. Therefore instructions will not specify the exact steps to carry out. Instead, it is part of the assignment to identify promising features, models and preprocessing methods and apply them as appropriate.

The overall goal is to predict whether a payment by a company to a medical doctor or facility was made as part of a research project or not.
Please use the data from 2017 provided here:
https://www.cms.gov/OpenPayments/Explore-the-Data/Dataset-Downloads.html
The dataset download contains a description of the files.

You will have to join multiple files to generate the dataset. You do not have to use the whole dataset, and it's recommended that you strongly subsample the data while developing your solution.

While the modelling process is likely iterative, please lay out the following tasks in the given order to facilitate grading.

Task 1 Identify Features

Assemble a dataset consisting of features and target (for example in a dataframe or in two arrays X and y). What features are relevant for the prediction task?
What features should be excluded because they leak the target information?
Show visualizations or statistics to support your selection.

Task 2 Preprocessing and Baseline Model

Create a simple minimum viable model by doing an initial selection of features, doing appropriate preprocessing and cross-validating a linear model. Feel free to generously exclude features or do simplified preprocessing for this task. As mentioned before, you don't need to validate the model on the whole dataset.

Task 3 Feature Engineering

Create derived features and perform more in-depth preprocessing and data cleaning. Does this improve your model? In particular, think about how to encode categorical variables.

Task 4 Any model

Use any classification model we discussed (trees, forests, gradient boosting, SVM) to improve your result. You can (and probably should) change your preprocessing and feature engineering to be suitable for the model. You are not required to try all of these models. Tune parameters as appropriate.

Task 5 Feature Selections

Identify features that are important for your best model. Which features are most influential, and which features could be removed without decrease in performance? Does removing irrelevant features make your model better?

Task 6 An explainable model

Can you create an “explainable” model that is nearly as good as your best model?
An explainable model should be small enough to be easily inspected - say a linear model with few enough coefficients that you can reasonable look at all of them, or a tree with a small number of leafs etc.