CITS4009 Computational Data Analysis

CITS4009 Computational Data Analysis ⌄

Lab 08 - Project (Classification)

How to get started

This lab aims to help you to get started on Project.

Data Preparation

The whole purpose of data manipulation for Project is to prepare data ready for modelling. Assuming you have dealt with incomplete, invalid, missing and duplicate data problem already, here are some of the fundamental steps:

- To prepare data for binary classification, we need to discretise one **meaningful** numerical column into binary. So first thing first, what do you want to predict? Earnings? Educational vs. Non-Educational?

- Incorporate **at least one other meaningful data** to help your investigation (e.g. country median age).

- We then need to split the data (ensuring re-producibility) to **test and training set**.

Build Single Variable Models

Follow Lecture Slides to

- Explain what a null model would look like for classification

- Find all categorical variables and numerical variables

- Expand the dataset with columns built from single variable models

- Plot AUC and select the best performing variables according to AUC

Previous
Lab 07 - Single Variable Models and Naïve Bayes model

Last updated 9 days ago