Group Project

2022-11-02

Group Data Project Description

We will consider the data set given in the kaggle competition Give Me Some Credit. You can find it here: https://www.kaggle.com/c/GiveMeSomeCredit. You will, as a group, build a learner that you will then evaluate on the test data via Kaggle. The evaluation metric is the area under the curve (AUC) of the receiver operating characteristic (ROC) curve. The analysis is a classification problem, but also involves missing data problems which are typically solved with regression approaches. A write up of your analysis is due on the 22nd of November, with an approximately 5 minute presentation given the 17th in class.

Beyond describing the learner you used and its performance, discuss your imputation strategy, outlier detection, and data visualization, and anything else you find valuable.

Some R packages that you might find useful that don't appear in the notes:

rfImpute: random forest based imputation

mice: Multivariate Imputation by Chained Equations

stray: high dimensional anomaly detection

but there are many others.