Stat 6021: Project 2 Proposal

Due November 4

Adult data set is from UCI Machine Learning Repository (???). The data set is meant for binary class classification to predict whether the income of adults is greater than 50K or not. Our client’s main concern whether they will always have a high income. There are a total 14 predictors, consisting of eight categorical and six numerical types and are named as follows:

* age (numerical)
* workclass (categorical)
* fnlwgt (numerical)
* education (categorical)
* education-num (numerical)
* marital-status (categorical)
* occupation (categorical)
* relationship (categorical)
* race (categorical)
* sex (categorical)
* capital-gain (numerical)
* capital-loss (numerical)
* hours-per-week (numerical)
* native-country (categorical)

1. The data needs to be processed by data cleaning, data outlying check and handling of missing data. Adopt null/hypothesis tests to identify statistical significant predictors.

2. The initial model considered: since our client’s primary concern is to ensure the model will predict whether an adult can get the income higher than 50K or not. The objective is to apply Logistic Regression on the data to split test and train sets.  Receiver Operating Characteristic (ROC) curve and the Area under the ROC Curve (AUC) is used to assess how well our logistic regression model performs in classifying outcomes. Apply confusion matrix to adjust threshold to see how the True positive and false negative varied.

3. Summary of findings: our recommended model(s) including suggestions for fine-tuning the model(s) and further development.

4. Summary of findings for the client: addressing client's main concerns and why we are recommending a model(s).