**Stat 6021: Project 2 Proposal**

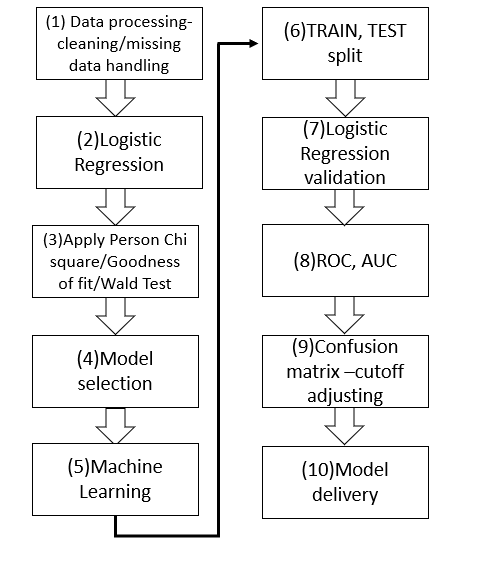
Due November 4

**Adult data set** is from UCI Machine Learning Repository (<https://archive.ics.uci.edu/ml/datasets/adult>). 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 concerns are whether they will always have a high income and simple model delivery.** There are total 14 predictors, consisting of eight categorical and six numerical types and are named as follows:

Response variable: Income: >50K, <50K

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Predictors | 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) |

The project goals are to (1)deliver the model which can predict whether a person can earn 50K income or not and (2) simplified model.

Figure 1 The workflow for model delivery 

Our group plans to conduct the following analytical tests to obtain the required results and our approaches are as follows:

1. **Data processing:** the data set consists of 32,560 rows with 14 attributes and a binary response variable. We need to perform data cleaning, data outlying check, mutate data, handling of missing data and to add data hander.
2. **Logistic regression**: apply logistic regression since the response variable is binary. We need to use logistic regression.
3. **Apply test statistic** to understand the deviance between the full model and any possible reduced model. Apply goodness of fit testing to achieve 95% confidence level.
4. **Test hypotheses** of a subset of the model parameters to determine statistically significant predictors.
5. **Machine learning**: since our goal is a model that can predict whether a person to earn an income of 50K, apply machine learning.
6. **Train and split** data in a proper ratio.
7. **Logistic Regression validation** – performance measurement is an essential task and we use ROC and AUC for checking our model’s performance.
8. We use ROC and AUC to check the model formation and interpret the true positive rate and the false negative rate.
9. **Apply confusion matrix** to determine the cut off value and meet our client’s concern – precisely predict whether a person has an income > 50K.
10. **Model delivery** –
    1. Summary of findings: our recommended model(s) including suggestions for fine-tuning the model(s) and further development.
    2. Summary of findings for the client: addressing client's main concerns and why we are recommending the model(s).