Group 15 Project Proposal

1. Write your group number and the names of all members of your group.

Group 15:

-Nadeem Karachi
-Sara Shiha
-Hung Nguyen
-Elinnoel Nuñez
-Nimet Ozkan

2. Provide a detailed description of the data your group has selected, including

a. The source and inspiration for selecting this particular data set.

- **https://archive-beta.ics.uci.edu/dataset/2/adult**
- **The Adult dataset, also known as the "Census Income" dataset contains 48842 instances and 14 attributes. Based on the extraction done by Barry Becker from the 1994 Census database, the goal of this dataset is to predict whether income exceeds $50K/yr based on the given data. We would like to understand the possible factors that may affect one's earning potential of $50k/yr.**

b. The size of the data (number of observations and number of variables).

- **Observations: 48842**
- **Variables: 14**

c. Description of all variables (similarly to the descriptions you encounter for R data sets).

- **Age: Adult age**
- **Workclass: Private, self-emp-not-inc, federal-gov, etc.**
- **Fnlwgt: Adult final weight**
- **Education: Adult level of education**
- **Education-num: Adult's education number**
- **Marital Status: Married, divorced, widowed, etc.**
- **Occupation: Adult's job, tech-support, craft-repair, etc.**
- **Relationship: Wife, husband, own-child, etc.**
- **Race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, other, black**
- **Sex: Male or female**
- **Capital-gain: Money gained**
- **Capital-Loss: Money loss**
- **Hours-per-week: Hours worked per week**
- **Native-Country: Adult birthplace**

3. Decide on main questions you would like to ask for your data. The most critical steps would typically be

a. Determining the variable(s) you will use as response, and potential variables you will use as predictors.

- **Based on given attributes we will predict whether an individual's income will exceed $50K/yr (qualitative).**

b. Figuring out if you prioritize prediction or inference as your final goal(or maybe you'd like to attempt both).

- **We will be using prediction to determine our objective.**

c. Formulating the data question(s) clearly.

- **Since the response variable we are predicting is categorical/binary, we can use decision trees and logistic regression for our data set to answer our question.**

4. Outline the models and methods you anticipate using in order to answer those questions and addressing those tasks. Minimal requirements are:

- <u>Steps we'll go about creating our models and retrieving data</u>

a. Make sure to use at least two distinct models from the ones covered in this course (two out of: linear regression, logistic regression, decision trees, random forests, neural networks), pointing to reasons and advantages of each model over the others.

- **Logistic regression**
    - **We decided to choose a logistic regression method since it is easy to implement, train data, interpret and it is efficient.**
- **Decision trees**

b. Make sure to perform model comparison via resampling methods introduced in this course (mainly cross-validation).

5. How do you plan on distributing the workload across the group members? E.g. "Members 1 & 2 will be implementing and testing model A , while members 3 & 4 will focus on model B (both models should be mentioned in 2(a))"

Or

"Members 1 & 2 will work on data question A, while other members will deal with data question B (both questions should be formulated in 1(c)). (edited)

- **For the most part the group will all work together step by step on each section, and later on split evenly between models, which would include implementing and testing. Then we will reconvene and analyze the models against each other for our final conclusions.**