Springboard Capstone Project 1 – Project Proposal

# What is the problem you want to solve?

# Who is your client and why do they care about this problem? In other words, what will your client do or decide based on your analysis that they wouldn't have done otherwise?

Project Goals:

This capstone project is based on a dataset called 'Adult Data Set' from UCI's Machine Learning Repository.

This project aims to solve following problems:

1. Predict whether a person's income exceed $50K/yr based on the census data
2. Identify various segments which may exist in the population based on the data collected

Target Applications of this Data Analysis:

Census data is almost like a gold mine for business users. Primary reasons being, it is a true population data which captures the information about the total universe of data instead of a mere sample.

Such census data when combined with income classification based on 50K+ threshold would form a very valuable source of population demographics. Once we analyze this data and establish potential relationships between some of the key demographic parameters which influence the income category, this information can be very effectively used in making many economic, social, marketing, advertising, sales promotion and many other similar business decisions.

I envisage that this research project would be very useful to search marketers, social media advertisers, automotive manufacturers (or similar big ticket products manufacturers and marketers), media and print advertisers, etc.

# *What data are you using? How will you acquire the data?*

A little more information about the dataset:

· Dataset Information: The dataset is based on the 1994 US Census Database. A set of reasonably clean records was extracted using the following conditions: ((AAGE>16) && (AGI>100) && (AFNLWGT>1)&& (HRSWK>0))

Prediction task is to determine whether a person makes over 50K a year.

Attribute Information:

Listing of attributes:

- >50K, <=50K.
-
- age: continuous.
- workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
- fnlwgt: continuous.
- education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
- education-num: continuous.
- marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
- occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
- relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
- race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
- sex: Female, Male.
- capital-gain: continuous.
- capital-loss: continuous.
- hours-per-week: continuous.
- native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

# Briefly outline how you'll solve this problem. Your approach may change later, but this is a good first step to get you thinking about a method and solution.

Project Plan / Approach to achieve the goals outlined above:

a. As in any good data analysis project, I first plan to do a good deal of EDA (Exploratory Data Analysis) of the dataset. The objective is to thoroughly understand the data and then based on that prepare the data analysis plan/framework.

b. Based on the EDA, I will identify the key hypothesis which should answer following questions

- Overall income distribution of the population
- Which attributes may have greater influence on determining the income threshold of 50K/yr
- Any other interesting facts

c. Use regression methods to determine which attributes have higher impact/influence

d. Plan how to use the visualization techniques learned to present the findings

e. Use Supervised learning methods to identify patterns, segments and attributes which drive the results

f. Use classification methods to classify income segments

g. Use ML techniques to predict the income level and compare the accuracy of the different models

h. Use Unsupervised learning methods to identify patterns, segments and attributes which drive the results

# What are your deliverables? Typically, this includes code, a paper, or a slide deck.

Key Deliverables, Collaterals

I envision following documents to document and present the result of this effort

- iPython Notebooks containing all the codes, analysis and results
- A summary/executive finding report
- Presentation slides