Proposal for Capstone Project 1

Introduction

Census Income Data Set:

http://archive.ics.uci.edu/ml/datasets/Census-Income+%28KDD%29

This data set contains weighted census data extracted from the 1994 and 1995 Current Population Surveys conducted by the U.S. Census Bureau. The data contains 41 demographic and employment related variables.
The instance weight indicates the number of people in the population that each record represents due to stratified sampling. There are 199523 instances in the data file and 99762 in the test file. The data was split into train/test in approximately 2/3, 1/3 proportions.

Problem to resolve:

To predict income class of US population (Whether a person makes over 50k a year). It's an imbalanced classification and a classic machine learning problem.

Clients and Audiences

Bank, Insurance company, Medical research institution may have interested in this kind of problem. As you know, machine learning is being extensively used to solve imbalanced problems such as cancer detection, fraud detection etc.

Approach

1. Downloading training set and test set from UCI website.
2. Be familiar with the variables in the data set.
3. Cleaning data to prepare the data set for analyzing.
4. Resolving imbalanced data set problem choosing "Under sampling the majority class" or "Over sampling the minority class" or "Synthetic sampling" way.
5. Applying Naïve Bay, SVM, Decision Trees, Random Forest, AdaBoost, logistical regression algorithms to build classifier for the prediction of income class.

Project Deliverables

The project deliverable will contain following:

- This document explaining approach of the project.
- Presentation created on this project.
- Python code for data wrangling with comments and details.
- Graphs and visualizations created for this project.
- Python code for model building with comments and details.