# Project report for Adult Income Dataset.

BY

Mohammad Arshadulla Noor & Amina Mahnoor

Electronics and Communication Engineering, Final Year Student, Muffakham Jah College of Engineering and Technology, Mount Pleasant, 8-2-249, Road No. 3, Banjara Hills, Hyderabad, Telangana, 500034

Introduction:

The adult income dataset that involves predicting personal income levels as above or below 50,000 per year based on personal details such as relationship and education level. This means that techniques for imbalanced classification can be used whilst model performance can still be reported using classification accuracy, as is used with balanced classification problems.

Approach:

In this, you will discover how to develop and evaluate a model for the imbalanced adult income classification dataset.

We evaluate:

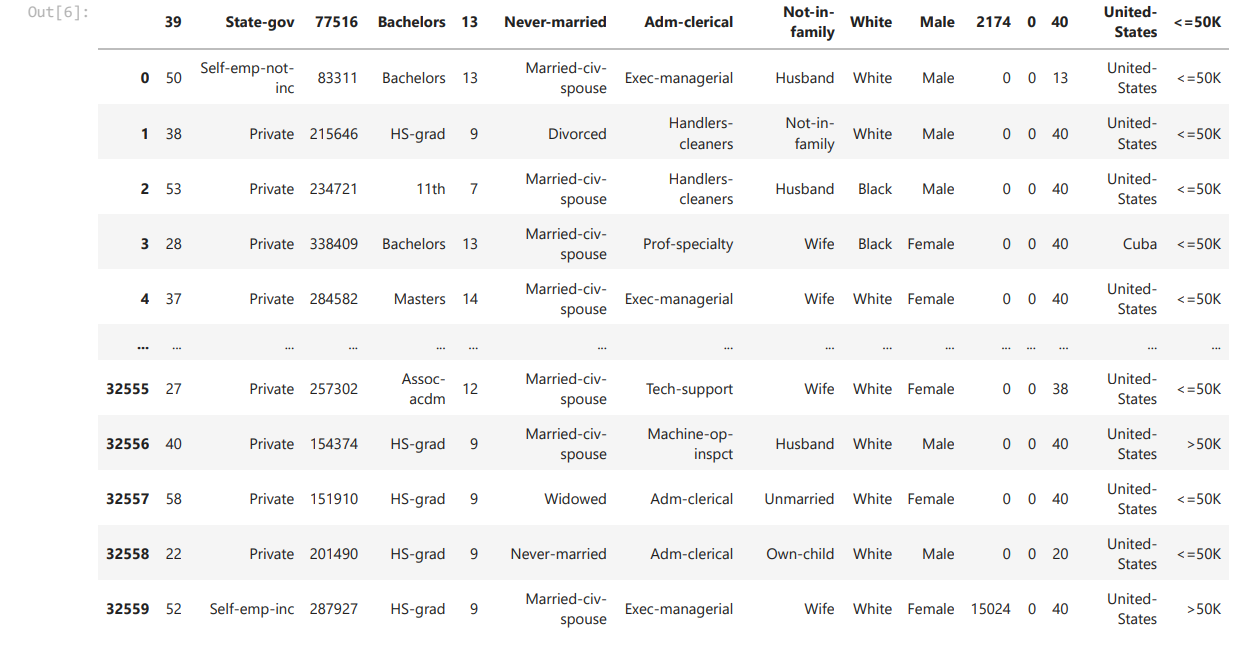
* How to load and explore the dataset and generate ideas for data preparation and model selection.
* How to fit a final model and use it to predict class labels for specific cases.

This is divided into five parts:

1. Adult Income Dataset
2. Explore the Dataset
3. Model Test
4. Evaluate Model
5. Make Prediction on New Data

In this project, we use a standard imbalanced machine learning dataset referred to as the “Adult Income” dataset.

The dataset is given below:



The dataset provides 14 input variables that are a mixture of categorical, ordinal, and numerical data types. The complete list of variables is as follows:

* Age.
* Workclass.
* Final Weight.
* Education.
* Education Number of Years.
* Marital-status.
* Occupation.
* Relationship.
* Race.
* Sex.
* Capital-gain.
* Capital-loss.
* Hours-per-week.
* Native-country.

The dataset contains missing values that are marked with a question mark character (?).

There are a total of 32,560 rows of data, and 15 with columns of data.

After data preprocessing, data visualization is done using the given dataset to check to analyze the income for different categories.

There are two class values ‘*>50K*‘ and ‘*<=50K*‘, meaning it is a binary classification task. The classes are imbalanced, with a skew toward the ‘*<=50K*‘ class label.

* **‘>50K’**: majority class, approximately 25%.
* **‘<=50K’**: minority class, approximately 75%.

Given that the class imbalance is not severe and that both class labels are equally important, it is common to use classification accuracy to report model performance on this dataset.

Model Building is done using different types of classification models.Using predefined train and test sets, reported good testing classification accuracy of about 87.30 percent by using XGB Classifier. This might provide a target to aim for when working on this dataset.

Conclusion:

By looking above values we can say that education-num, marital status, relationship and capital gain are the features that are important for income prediction.

Code: