Census Income Prediction

Overview

Predict whether income exceeds $50K/yr based on census data.

Here, the prediction task is to determine whether a person makes over 50K a year or not.

Data Dictionary

The dataset contains several parameters which can be useful for census income prediction.

Train File

CSV file containing samples for whom 'wage_class' is known.

Variable	Possible values
Person_Id	Discrete
age	continuous
work_class	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-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
race	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
sex	Female, Male

capital_gain	continuous
capital_loss	continuous
hours_per_week	continuous
naive_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
wage_class	>50K, <=50K

Test File

CSV file containing samples for whom 'wage_class' is to be predicted.

Submission File Format

Variable	Description
Person_Id	Person Id
wage_class	Predicted wage_clss

Public and Private LeaderBoard

Test file is further divided into Public (25%) and Private (75%).

- Your initial responses will be checked and scored on the Public data.
- The final rankings would be based on your private score which will be published once the competition is over.

Evaluation Criteria

Your model performance will be evaluated on the basis of **ROC-AUC score**.

Rubrics

Component	Weightage
Data Cleaning and Data Visualization	25%
Model Building and Evaluation	60%
Pipeline and Deployment (Dashboard/Webapp)	15%