Adult Income Classification

Problem Statement: Predict whether income exceeds $50K/yr based on census data. Also known as the "Census Income" dataset

Data Set Characteristics:	Multivariate	**Number of Instances:**	48842
Attribute Characteristics:	Categorical, Integer	**Number of Attributes:**	14
Associated Tasks:	Classification	**Missing Values?**	Yes

Data Set Information:

Dataset: adult.csv.zip

Attribute Information:

Listing of attributes:

Target Column: Salary

Column Name	Description
age	continuous.
work class	Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
Fnlwgt (Final Weight)	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.

Build a robust machine learning model that would determine whether census income is above 50 K or not.

Evaluation Criteria

- Data Cleaning
- EDA + Insights
- Feature Engineering
- Modelling + Hyper Parameter Tuning

All the above criteria will be checked in the Jupyter Notebook (quality of Analysis) as well as on achieving a top-scoring model.

Follow PEP guidelines to write your python code to maintain coding standards.

2. Problem Description:

- Predict which customers are potentially interested in a caravan insurance policy.
- Describe the actual or potential customers; and possibly explain why these customers buy a caravan policy.

Prediction

We want you to predict whether a customer is interested in a caravan insurance policy from other data about the customer. Information about customers consists of 86 variables and includes product usage data and socio-demographic data derived from zip area codes. The data was supplied by the Dutch data mining company Sentient Machine Research and is based on a real-world business problem. The training set contains over 5000 descriptions of customers, including the information on whether or not they have a caravan insurance policy. A test set contains 4000 customers.

For the prediction task, the underlying problem is to find the subset of customers with a probability of having a caravan insurance policy above some boundary probability. The known policyholders can then be removed and the rest receives a mailing. The boundary depends on the costs and benefits such as the costs of mailing and the benefits of selling insurance policies. To approximate this problem, we want you to find the set of 800 customers in the test set that contains the most caravan policy, owners.

Description

The purpose of the description task is to give a clear insight into why customers have a caravan insurance policy and how these customers are different from other customers. Descriptions can be based on regression equations, decision trees, neural network weights, linguistic descriptions, evolutionary programs, graphical representations, or any other form. of solutions (e.g. minimize a loss function, maximize comprehensibility, minimize response time, etc.)

The descriptions and accompanying interpretation must be comprehensible, useful, and actionable for a marketing professional with no prior knowledge of computational learning technology. The value of a description is inherently subjective.

Dataset **ticdata2000.txt**
Dataset to train and validate prediction models and build a description (5822 customer records). Each record consists of 86 attributes, containing sociodemographic data (attributes 1-43) and product ownership (attributes 44-86). The sociodemographic data is derived from zip codes. All customers living in areas with the same zip code have the same

sociodemographic attributes. Attribute 86, "CARAVAN: Number of mobile home policies", is the target variable.

Dataset: ticeval2000.txt
Dataset for predictions (4000 customer records). It has the same format as TICDATA2000.txt, only the target is missing. Participants are supposed to return the list of predicted targets only. All datasets are in tab-delimited format.
The meaning of the attributes and attribute values is given below.

Dataset: tictgts2000.txt
Targets for the evaluation set.

Information about dataset dictionary.txt

Note: All the variables starting with M are zipcode variables. They give information on the distribution of that variable, e.g. Rented house, in the zip code area of the customer.

One instance per line with tab-delimited fields.

Evaluation Criteria

- Data Cleaning
- EDA + Insights
- Feature Engineering
- Modelling + Hyper Parameter Tuning

All the above criteria will be checked in the Jupyter Notebook (quality of Analysis) as well as on achieving a top-scoring model.

Follow PEP guidelines to write your python code to maintain coding standards.