PROJECT PROPOSAL: Adult Income >=50K Prediction

Group Members:

Sharook Shaik - 11706345

Nikhil Chagi - 11684249

SAIRAM SIRIPURAM - 11653118

Kavya Challa - 11656795

Introduction:

Individual income is the primary measure for any country to understand growth and development. In order to calculate this census data is collected every year and stored. This census data can be used in multiple ways to analyze the country and implement new schemes, plans and plan the upcoming development. Analyzing data greater than $50,000 helps in many ways including implementing future schemes, advertising strategies, budget plans and not limited.

This project helps to get insights of the income dataset, a subset of the vast dataset. This dataset is helpful for the assessment of the income trends and create a robust machine learning model which will help to predict the income is weather >=50K or not for given variables. One of the major application of this dataset is easy credit card approvals.

Objective:

With the help of this project, our goal is to understand various data patterns in the census data and predict weather the person / adult is earning income >= (greater than or equal to) 50,000 or not.

About the Data:

Overview:

The dataset represents the census data with filters age > 16. This dataset is curated to understand the income trends which suites the current problem.

Source : https://archive.ics.uci.edu/dataset/2/adult

Number of instances : 32560

Number of features : 15

Sample:

```
]: data.head()
```

]:

	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
0	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
1	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
2	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
3	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K
4	37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	United-States	<=50K

Data Dictionary

Column	Description
age	Continuous variable representing the age of the person
workclass	Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
fnlwgt	continuous.
education	Education of the person available values are : 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	Id for education index. continuous.
marital-status	Person married 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.
class	>50K, <=50K

Challenges:

Since the data is real time collected data, it is expected to have NA values, outliers and many other anomalies. The challenge is to clean and preprocess the data without deviating from the actual values or lose of data.

Objectives

I. understand trends in the various individual features

Dataset contains the various univariate columns which can be further analyzed such as age, understanding age column let us bring out the patterns over the age and income. Which age people are generally having income > $50000

II. Understanding external factors effecting the income

External factors such as type of government, hours per week, education gives us overview of the income trends that are affecting. This helps us to understand and implement the upcoming schemes or contributing to the major changes in the country government.

Suggested Approach:

Cleaning of data: This step involves either removing or replacing the null values, detection of outliers, adjusting the outliers, removing unwanted columns etc.

Data Exploration: This is the key step in solving the problem. This involves creating of graphs, analyzing various trends, univariate analysis, bivariate analysis etc.

Model Selection: There are various models out in market but it is important to select the right one which will fit our data. Selection of right mode involves model capabilities, time complexity, accuracy and many other metrics.

Testing: Once the model is trained, it is important to test with unknown values and test to the boundaries. We will test with unknown data and see how it is behaving.

Conclusion:

By analyzing and understanding this census data gives us a clarity on what are the external and internal factors affecting the income for an individual. Since income plays a very important role in the country economy and individual this project serves as a key for multiple applications over the time.