High Level Design

Adult Census Income Prediction

Content:

Abstract:

Outstanding inequality of wealth and income is a major concern especially in the United States. Opportunities Poverty reduction is one of the most important reasons for reducing global poverty a growing level of economic inequality. The goal of the universe moral equality ensures sustainable development and improvement the stability of the national economy. Almost every country across the globe are trying their best to deal with this problem again to provide the right solution. This study aims to show use machine learning and data mining techniques in providing a solution to the income equity problem. UCI Adult Dataset used for a purpose. Separation is done predicting the annual income of a person in the US falls into income category greater than 50K Dollars or less equivalent up to the $ 50K category based on a specific set of attributes.

Introduction:

Why this Low Level Design?

The intention of LLD or a low level-degree design report(LLDD) is to present the inner logical layout of the actual program code for adult census income prediction. LLD describes the class diagrams with the methods and family members between classes and program specifications. It describes the modules so that in future if any programmers want to bring any modification they can code the program from the record.

Scope:

The motive of this low level design report is to present a detailed description of a machine learning model to describe whether he is earning good or not. Its going to provide an explanation for the motive and the functions of the machine, the interfaces of the model, what the system will do. This document is intended for stakeholders and developers so they can infer from here and could do give a upgraded baseline to the model and could be proposed to the better control of its approval.

The objective of this project is simple, its to check whether the person's income is greater than 50K or not based on different attributes like their age, workclass, hours of week, etc.

Problem Statement:

To check whether a person based on his status is able to get a fortunate income to sustain in the American society. This is basically a simple binary classification where a person is classified into the >50K group or <=50K group. The Adult dataset is from the Census Bureau and the task is to predict whether a given adult make more than $50000 a year based attributes such as education, hours per week, etc.

Proposed Solution:

Every person is very calculative when it comes about career choice and it also indirectly points on the state of income in that specific line of profession. So for that they need a proper understanding to help them pave way for more clearer path when deciding with a career choice and need to do analysis based on the income. So we have proposed a model which has gathered a dataset related to the same.

The dataset has recorded random individuals based on their age, years of experience, work class, martial status, their highest education qualification, their contribution for work in terms of hours per week, these attributes are taken into consideration in building a desired model which will help individuals. We have implemented the above use case in building the income prediction

Further Improvements:

Adult census income prediction is basically a binary classification which displays two possible outcomes either they are above 50K or less than 50K. Possible further improvements can be done like transform from binary to multi class classification, giving us a clearer idea about their income range, the range can be factioned in a range like 10k-20k, 20k-30k, 30k-40k and so on, this will give a more clearer picture when computing with income prediction.

Algorithms used:

There are several sets of algorithms are used to find an optimal result to ensure and maintain best possible prediction accuracy. For this algorithms used were Random forest classifier, Decision Tree, Logistic Regression, SVM.

Among these brute force approach was applied to check out the best possible output, upon test regressively we have found that Random Forest gave the best accuracy, so we chose the same and did hyperparameter tuning to get more accuracy.

Dataset:

The dataset we have obtained is from:

https://www.kaggle.com/datasets/overload10/adult-census-dataset

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	age	workclass	fnlwgt	education	education	marital-st	occupatio	relationsh	race	sex	capital-gai	capital-los	hours-per	country	salary
2	39	State-gov	77516	Bachelors	13	Never-ma	Adm-cleri	Not-in-fa	White	Male	2174	0	40	United-St	<=50K
3	50	Self-emp-	83311	Bachelors	13	Married-c	Exec-man	Husband	White	Male	0	0	13	United-St	<=50K
4	38	Private	215646	HS-grad	9	Divorced	Handlers-	Not-in-fa	White	Male	0	0	40	United-St	<=50K
5	53	Private	234721	11th	7	Married-c	Handlers-	Husband	Black	Male	0	0	40	United-St	<=50K
6	28	Private	338409	Bachelors	13	Married-c	Prof-spec	Wife	Black	Female	0	0	40	Cuba	<=50K
7	37	Private	284582	Masters	14	Married-c	Exec-man	Wife	White	Female	0	0	40	United-St	<=50K
8	49	Private	160187	9th	5	Married-s	Other-ser	Not-in-fa	Black	Female	0	0	16	Jamaica	<=50K
9	52	Self-emp-	209642	HS-grad	9	Married-c	Exec-man	Husband	White	Male	0	0	45	United-St	>50K
10	31	Private	45781	Masters	14	Never-ma	Prof-spec	Not-in-fa	White	Female	14084	0	50	United-St	>50K
11	42	Private	159449	Bachelors	13	Married-c	Exec-man	Husband	White	Male	5178	0	40	United-St	>50K
12	37	Private	280464	Some-coll	10	Married-c	Exec-man	Husband	Black	Male	0	0	80	United-St	>50K
13	30	State-gov	141297	Bachelors	13	Married-c	Prof-spec	Husband	Asian-Pac	Male	0	0	40	India	>50K
14	23	Private	122272	Bachelors	13	Never-ma	Adm-cleri	Own-chil	White	Female	0	0	30	United-St	<=50K
15	32	Private	205019	Assoc-acc	12	Never-ma	Sales	Not-in-fa	Black	Male	0	0	50	United-St	<=50K
16	40	Private	121772	Assoc-voc	11	Married-c	Craft-rep	Husband	Asian-Pac	Male	0	0	40	?	>50K
17	34	Private	245487	7th-8th	4	Married-c	Transport	Husband	Amer-Ind	Male	0	0	45	Mexico	<=50K
18	25	Self-emp-	176756	HS-grad	9	Never-ma	Farming-f	Own-chil	White	Male	0	0	35	United-St	<=50K
19	32	Private	186824	HS-grad	9	Never-ma	Machine-	Unmarrie	White	Male	0	0	40	United-St	<=50K
20	38	Private	28887	11th	7	Married-c	Sales	Husband	White	Male	0	0	50	United-St	<=50K

The dataset contains 32,561 entries with a total of 15 columns representing different attributes of the people. Here's the list;

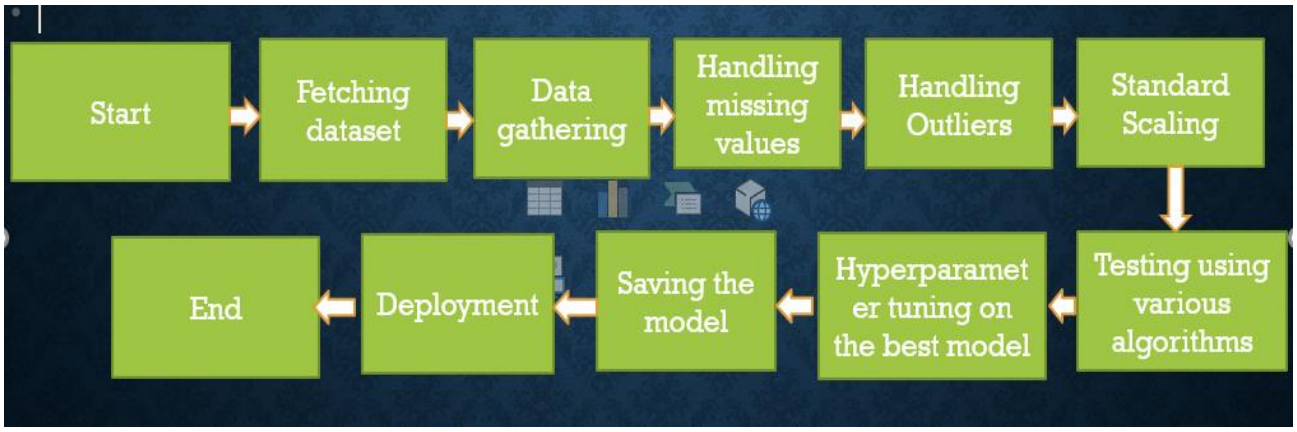
1. Age: Discrete (from 17 to 90)
2. Work class (Private, Federal-Government, etc): Nominal (9 categories)

3. Final Weight (the number of people the census believes the entry represents): Discrete
4. Education (the highest level of education obtained): Ordinal (16 categories)
5. Education Number (the number of years of education): Discrete (from 1 to 16)
6. Marital Status: Nominal (7 categories)
7. Occupation (Transport-Moving, Craft-Repair, etc): Nominal (15 categories)
8. Relationship in family (unmarried, not in the family, etc): Nominal (6 categories)
9. Race: Nominal (5 categories)
10. Sex: Nominal (2 categories)
11. Capital Gain: Continous
12. Capital Loss: Continous
13. Hours (worked) per week: Discrete (from 1 to 99)
14. Native Country: Nominal (42 countries)
15. Income (whether or not an individual makes more than $50,000 annually): Boolean (≤$50k, >$50k)

Tools Used:

Python programming language and frameworks such as Numpy, Pandas, Scikit-learn are used to build the whole model. The IDE used is Pycharm


Architecture:



Technology Stack:

Front End	HTML, CSS
Back End	Python
Deployment	Heroku

User-I/O Workflow: