High Level Design Document

Adult Census Income Prediction

Revision Number : 1.0

Last date of revision : 20/06/2022

Ritesh Naik

Darshan Naik

Document Version Control

Date Issued	Version	Description	Author
20-June-2022	1.0	Initial Release	Ritesh Naik

Contents

Abstract

Inequality in wealth and income is a major source of worry, particularly in the United States. One reasonable motivation to lessen the world's rising level of economic disparity is the possibility of reducing poverty. The notion of universal moral equality promotes long-term development and improves a country's economic stability. Governments in several countries have been working hard to address this issue and find the best answer possible. The goal of this project is to demonstrate how machine learning and data mining techniques can be used to solve the problem of income inequality. The Adult Dataset from UCI was utilized for this.

Based on a set of attributes, classification has been done by building a web app to forecast whether a person's annual income in the United States fits into the income categories of greater than 50K Dollars or less than 50K Dollars.

1. Introduction

The goal of this High-Level Design (HLD) Document is to provide all of the pertinent information regarding this project. This document gives overall understanding about the project.

1.1 Purpose of the Document

The purpose of this document is to

1. Different design techniques should be described.
2. Describe various methods of analysis based on a range of Use Cases.
3. Describe the system's third-party components and tools.
4. This project's whole process flow will be discussed

1.2 Objective of HLD

- To give a comprehensive overview of the entire system.
- To provide an overview of the problem's perspective and statement, as well as data requirements, tools used, and other topics.
- The goal is to provide a module-by-module breakdown of the complete system.

1.3 Scope of HLD

1. This HLD covers every aspect of the system.

2. General Description

2.1 Product Perspective & Problem Statement

Using classification-based Supervised Machine Learning algorithms, the Adult Census Income Prediction identified the person's income category as either greater than 50K Dollars or less than 50K Dollars..

The goal is to determine whether or not a person earns more than $50,000 per year. This is essentially a binary classification problem in which a person is assigned to one of two groups: >50K or 50K.

.

2.2 Data Requirement

In this project, we are using datasets provided by the "iNeuron" company to estimate the price. The dataset consist of 10 columns including the Target variable

- ★ **Age** : an integer value – user's age
- ★ **Capital Gain** : an integer value b/w [0-99999]
- ★ **Capital Lose** : an integer value b/w [0-3456]
- ★ **Hours per Week** : an integer value b/w [1-99]
- ★ **Education** : Preschool, 1st- 4th, 5th- 6Th, 7th-8th, 9th, 10th, 11th, 12ht, HS-grad, Somecollege, Bachelors, Masters, Assoc-voc, Assoc-acdm, Prof-school, Doctorate.
- ★ **Work-Class** : Private, Self-emp-not-Inc., Local-gov, State-gov, Self-emp-Inc., Federalgov, Without-pay, Never-worked
- ★ **Marital Status** : Married-AF-spouse, Married-civ-spouse, Married-spouse-absent, Never-married, Separated, Widowed
- ★ **Occupation** : Adm-clerical, Armed-Forces, Craft-repair, Exec-managerial, Farmingfishing, Handlers-cleaners, Machine-op-inspect, Other-service, Priv-house-serv, Profspecialty, Protective-serv, Sales, Tech-support, Transport-moving.
- ★ **Race** : Asian-Pac-Islander, Black, Other, White
- ★ **Gender** : Male, Female
- ★ **Relationship** : Not-in-family ,Other-relative, Own-child, Unmarried, Wife
- ★ **Native Country** : 'United-States', 'Cuba', 'Jamaica', 'India', 'Mexico', 'Puerto-Rico',

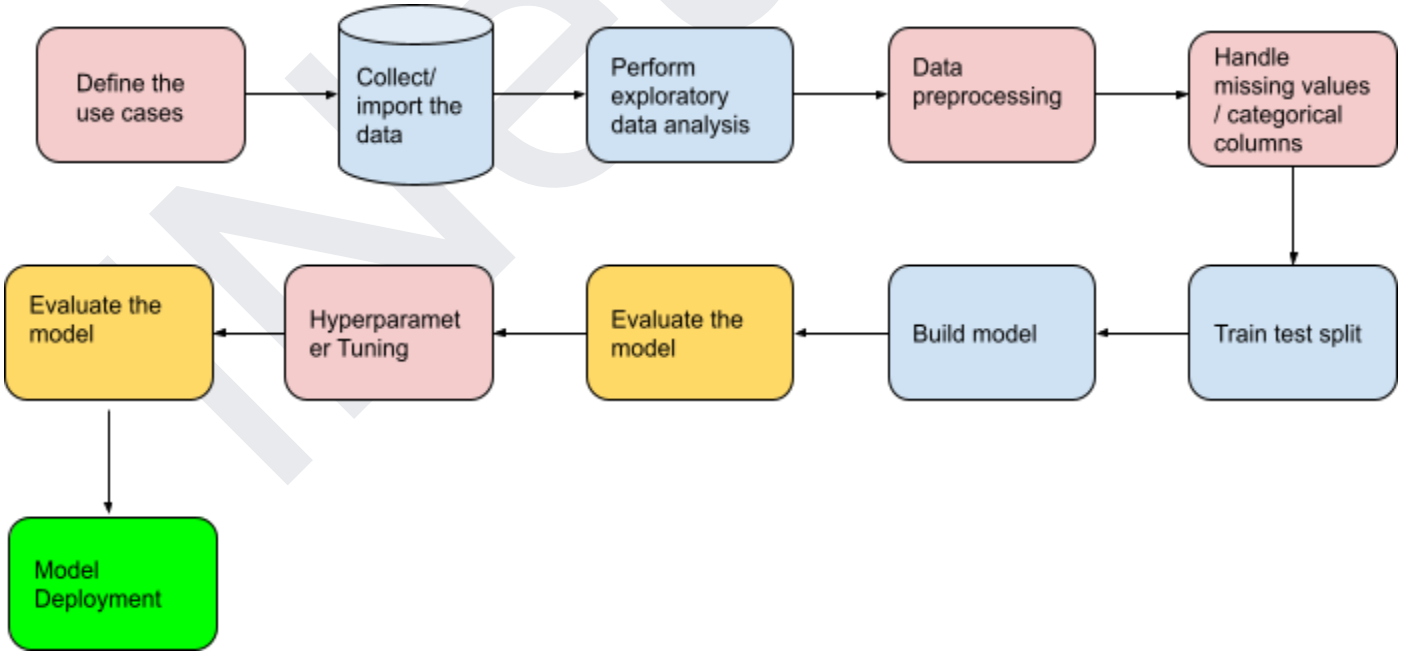
'Honduras', 'England', 'Canada', 'Germany', 'Iran', 'Philippines', 'Italy', 'Poland', 'Columbia', 'Cambodia', 'Thailand', 'Ecuador', 'Laos', 'Taiwan', 'Haiti', 'Portugal', 'Dominican-Republic', 'El-Salvador', 'France', 'Guatemala', 'China', 'Japan', 'Yugoslavia', 'Peru', 'Outlying-US(Guam-USVI-etc)', 'Scotland', 'Trinadad&Tobago', 'Greece', 'Nicaragua', 'Vietnam', 'Hong', 'Ireland', 'Hungary', 'Holand-Netherlands'

2.3 Tools Used

- Google collab is used as IDE.
- Pandas and NumPy are used for Data Manipulation & Pre-processing and mathematical functions respectively.
- Exploratory data analysis is automated by dataprep.
- For visualization of the plots, Matplotlib, Seaborn, Plotly are used.
- GitHub is used as version control system
- Heroku is used to deploy the model so that others can make use of it.

3. Design Details

3.1 Process flow



3.2 Error handling/ Exception handling

This project has been developed in such a way that the entire script is tested and run numerous times to ensure that no errors arise during the process flow.
We've also used the filterwarnings class from the warnings module to remove any unnecessary warnings to reduce confusion.

3.3 Deployment Process

1. Created a github_repository uploaded the below files
    a. Model.pkl
    b. App.py
    c. Requirements.txt
    d. Income_index.html
    e. Income_result.html
    f. Procfile

2. Login to Heroku platform & connect github, in the deployment method
3. Select the branch as main & then Deploy branch

4. Steps

1. Import the dataset
2. Dataset seems to have contained “?” instead of null values. Lets deal with it
3. Lets deal with the categorical features
4. Drop unwanted columns
5. Split the data as training & test dataset
6. Build the model. I have built a Decision tree , Random Forest & Neural Network models for this.
7. Evaluate the model. I am using accuracy as a metric to evaluate the model performance.
8. Finalize the model
9. Deploy the model

5. Conclusion

Based on several variables, the Adult Income Prediction algorithm will predict a user's income category. The algorithm will be trained on all of the different qualities of users, allowing it to accurately estimate whether or not an individual's salary is greater than $50,000.