High Level Design (HLD)

Adult Census Income Prediction

Document Version Control

Date Issued	Version	Description	Author
17/08/2021	1	Initial HLD-V1.0	Umang Tank
24/08/2021	1.1	Model Training and Evaluation-V1.1	Umang Tank

Contents

Abstract

Census Income prediction

Various parameters affected income prediction. Machine Learning for income prediction can make human life more efficient. This study demonstrates how different model of classification can forecast Income, And we compare the result with different model.

1 Introduction

1.1 Why this High-Level Design Document?

The purpose of this document is to present a detailed description of the Adult Census income prediction. It will explain the purpose and features of the system, the interfaces of the system, what the system will do, the constraints under which it must operate and how the system will react to external stimuli. This document is intended for both the stakeholders and the developers of the system and will be proposed to the higher management for its approval.

The HLD will:

- Present all of the design aspects and define them in details.
- Describe the user interface being implemented - Describe the hardware and software interfaces
- Describe the performance and requirements
- Include design features and the architecture of the project - List and describe the non-functional attributes like:
    - Security
    - Reliability
    - Maintainability
    - Portability
    - Reusability
    - Application compatibility
    - Resource utilization
    - Serviceability

1.2 Scope

The HLD documentation presents the structure of the system, such as the database architecture, application architecture, application flow (Navigations),

and technology architecture. The HLD uses non-technical to mildly-technical term which should be understandable to the administrator of the system.

2 General Description

2.1

Adult income Prediction ML model Predict whether income of that person is more than 50K or not.

2.2 Problem Statement

Product Perspective

The Goal is to predict whether a person has an income of more than 50K a year or not. This is basically a binary classification problem where a person is classified into the >50K group or <=50K group.

2.3 Proposed Solution

Using ML model we have to model Predict whether income of that person is more than 50K or not.

2.4 Data Requirement

Name	Description
Age	Age
Work Class	: Working Class (Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Neverworked)
Education-level	Level of 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	: Number of educational years completed
Marital-status	Marital status (Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AFspouse)
Occupation	Work 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	: Relationship Status (Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried)
race	Race (White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black)
sex	Sex (Female, Male)
Capital-gain	Monetary Capital gains
Capital-loss	Monetary Capital Losse
Hours-per-week	Average Hours Per Week Worked
Native-Country	Native Country (United-States, Cambodia, England, PuertoRico, 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)

2.5 Tools used

Python programming language and frameworks such as NumPy, Pandas, Scikit-learn, Flask, Heroku, Git.

- VS-code is used as IDE.
- For visualization of the plots, Matplotlib, Seaborn and Plotly are used.
- Heroku is used for deployment of the model.
- MySQL is used to retrieve, insert, delete, and update the database
- Frontend development is done using HTML/CSS
- Python Flask is used for backend development.
- GitHub is used as version control system.

2.6 Constraints

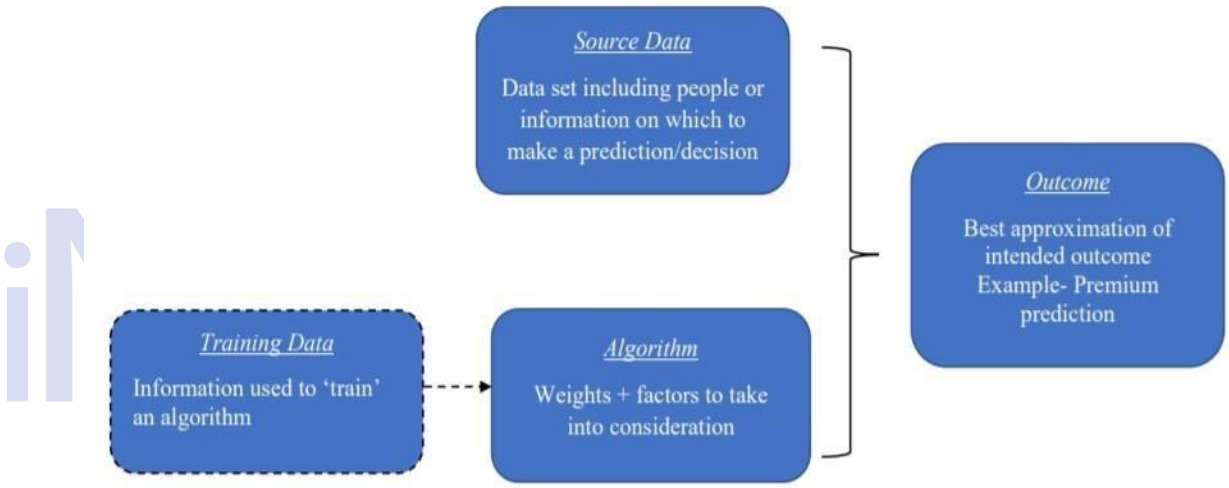
The Adult Census income Prediction must be user friendly, as automated as possible and users should not be required to know any of the workings.

3. Design Details

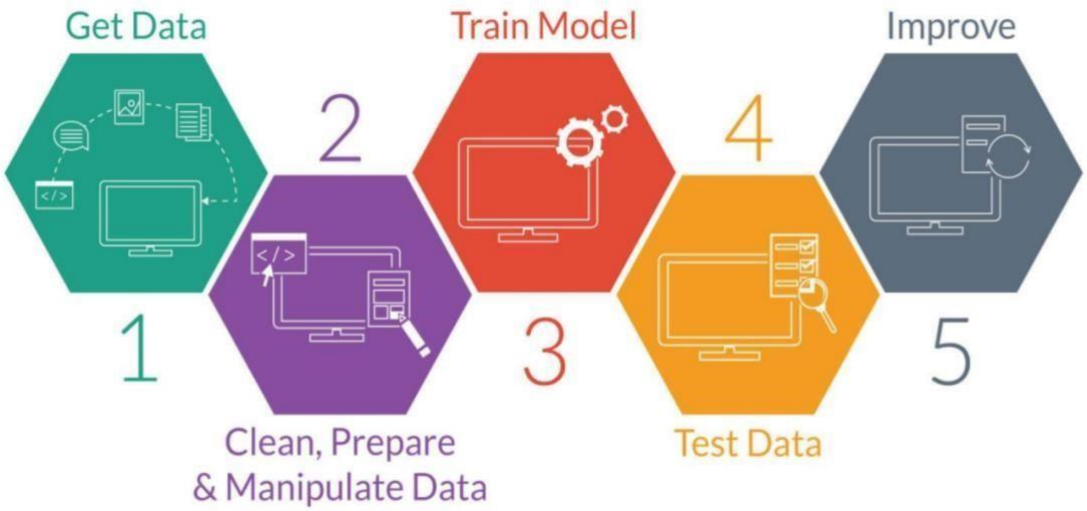
3.1 Process Flow

For predicting the Income, we will use Classification model. Below is the process flow diagram is as shown below.

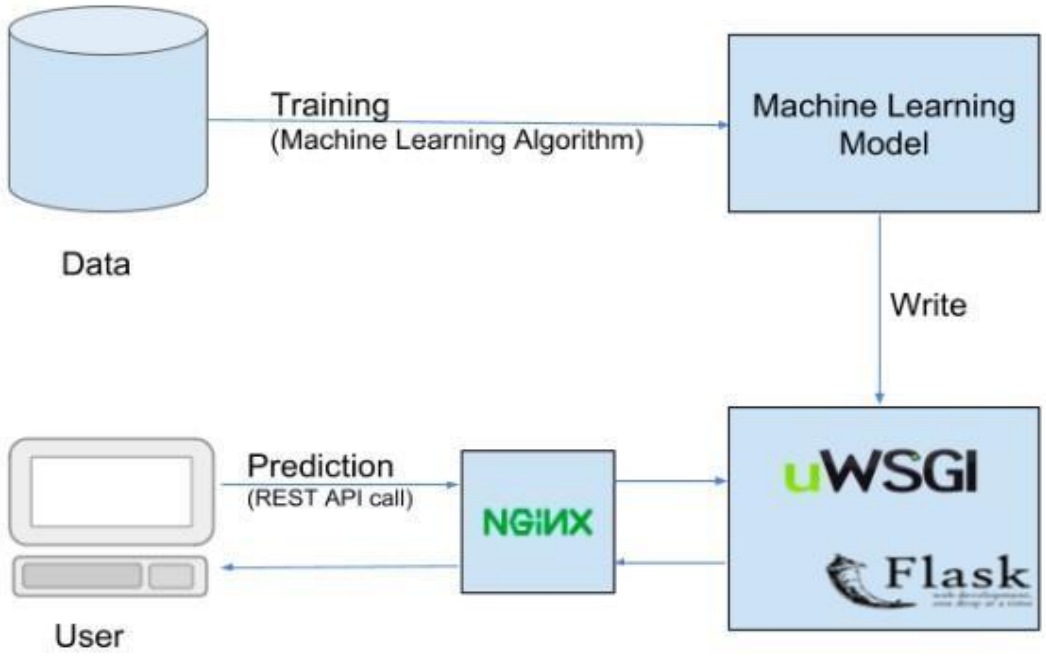
Proposed Methodology



3.1.1 Model Training and Evaluation



3.1.2 Deployment Process



3.2 Event Log

The system should log every event so that the user will know that process is running internally.

Initial Step-By-Step Description:

1. The system identifies at what step logging required
2. The system should be able to log each and every system flow
3. Developer can choose logging method. You can choose database logging / File logging as well.
4. System should not hang even after using loggings. Logging just because we can easily debug issues so logging is mandatory to do.

3.3 Error Handling

Should error be encountered, an explanation will be displayed as to what went wrong? An error will be defined as anything that falls outside the normal and intended usage.

4.1 Reusability

The code written and the components used should have the ability to be reused with no problems.

4.2 Application Compatibility

The different components for this project will be using as an interface between them. Each component will have its own task to perform, and it is the job of the python to ensure proper transfer of information.

4.3 Resource Utilization

When any task is performed, it will likely use all the processing power available until that function is finished.

4.4 Deployment

5 Conclusion

Background In this project, five Classification models are evaluated for individual Adult income prediction data. It has been found that AdaBoostClassifier model which is built is the best performing model.