High level Design (HLD)

Adult Census Income Prediction (ACI)

Revision Number:1.0

Last Date of revision:15-Jan-22

Document Control:

Date issued	Version	Description	Author
15- Jan-22	1	Initial HLD ver-1	Rajeev Ranjan

Contents

Abstract:

The prominent inequality of wealth and income is a huge concern especially in the United States. The likelihood of diminishing poverty is one valid reason to reduce the world's surging level of economic inequality. The principle of universal moral equality ensures sustainable development and improve the economic stability of a nation. Governments in different countries have been trying their best to address this problem and provide an optimal solution. This study aims to show the usage of machine learning techniques in providing a solution to the income equality problem. The UCI Adult Dataset has been used for the purpose. Classification has been done to predict whether a person's yearly income in US falls in the income category of either greater than 50K Dollars or less equal to 50KDollars category based on a certain set of attributes.

1.Introduction:

1.1. Why this High-level design Document?

The purpose of this High-level Design document is to add necessary detail to the current project description to represent the suitable model for coding. This document is also intended to help detect contradictions prior to coding, and can be used as a reference manual for how the modules interact at high level.

The HLD will:

- Present all the design aspects and define them in detail
- Describe the user interface being implemented
- Describe hardware and software interfaces
- Describe performance requirements
- Include Design features and 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 database architecture, application architecture(layers), application flow, and technology architecture. The HLD uses non-technical to mildly technical terms which should be understandable to the administrators of the system.

1.3 Definitions

Term	Description
Database	Collection of all information monitored by system
IDE	Integrated Development Environment
AWS	Amazon web services

2.General Description

2.1 Product Perspective

Adult Census Income prediction (ACI) is a Machine Learning classification model-based modelling where we try to predict the income group of a person and take necessary actions.

2.2 Problem Statement

To create AI solution for classifying the person's income using Machine Learning

- Here the goal is to predict whether the person has an income more than 50k a year or not.
- This is binary classification problem where a person classified into >50k group or <=50k group

2.3 Proposed System

The solution proposed here is ACI (Adult Census Income prediction) is used to perform above use cases, if there any person with less than 50k income it categorizes into less category and vice versa. This study helps in knowing the economy of the country and it helps in solving income equality problem

2.4 Further Improvements:

This ACI can be added with more use cases like knowing your Savings, budgeting your expenditure. It can also be used to synchronize with Govt apps to make avail their schemes to low annual income people.

2.5 Technical Requirements

This document addresses the requirements for classifying economy of people in society. Technical requirements may include cloud platform for hosting the application, data bases like Cassandra to save our application data, and proper API connectivity for communication.

2.6 Data requirements

Data requirement completely depends on problem statement, in our case:

- The data should contain only numerical and categorical data
- The data can be anything in csv, tsv, xls , json, etc. There are different formats of data and based on requirements those data can be used.
- The data can’t be image format or image related document.
- Based on DSA (Data Sharing agreement) the data should be provided

2.7 Tools used

Python programming language and frameworks such as pandas, Numpy, Plotly, Scikit-learn are used to build model

- PyCharm is used as IDE
- For visualization and report we use Pandas profiling
- Cassandra is used to perform various CRUD operation
- Front end is done with HTML/CSS
- Backend development is done by python Flask
- GitHub is used as version control system

2.7.1 Hardware requirements

- PC of Ram minimum 4gb

Hard disk :1Tb

2.7.2 Software requirements

- Operating System: Windows/Linux/Mac
- Processor: Minimum of 1GHz
- System type: 64-bit Operating system

2.8 Constraints

The ACI prediction should be use friendly and as automated as possible. Users don’t require to know working of the application.

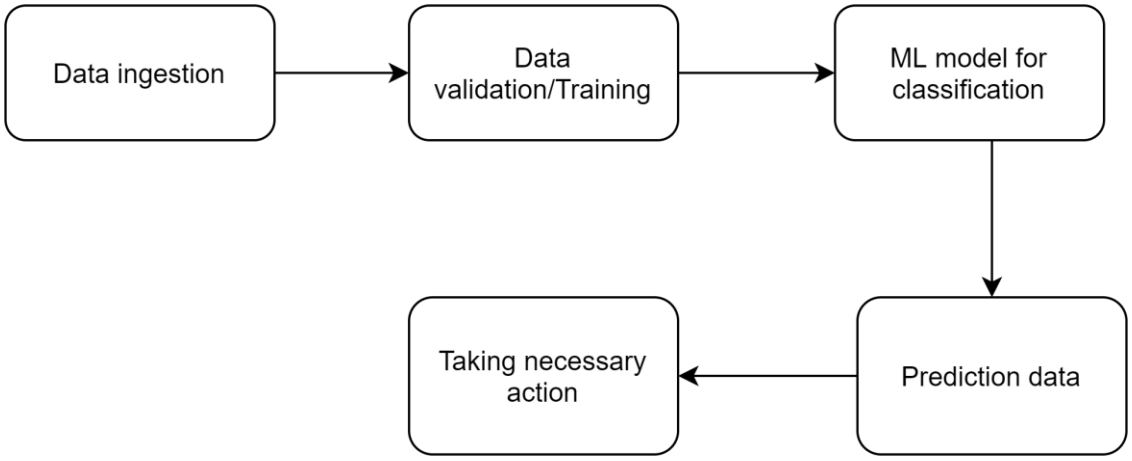
2.9 Assumptions

The main objective of the project is to implement use cases as previously mentioned (2.2 problem statement) for new dataset user provides. It is also assumed that this project works in user friendly manner.

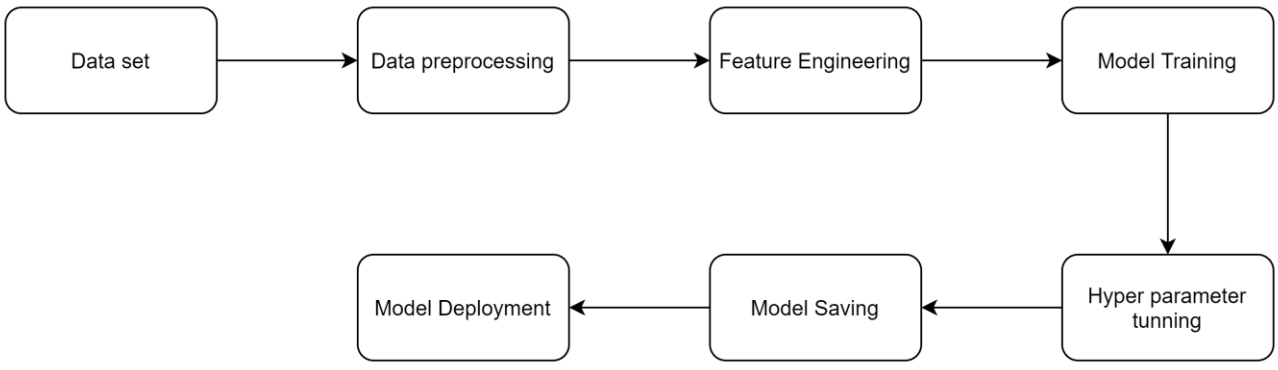
3.Design Details

3.1 Process Flow

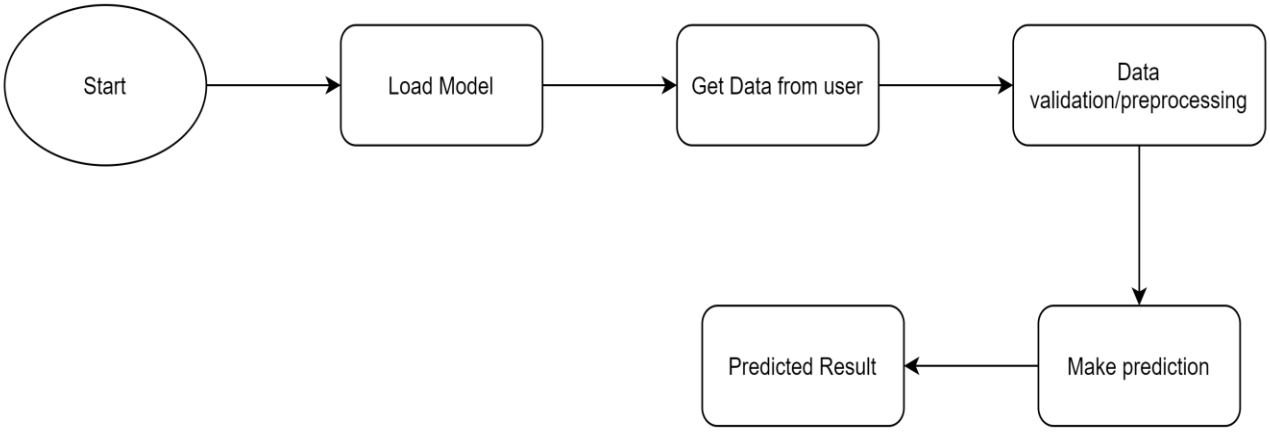
Below is the process flow diagram is as shown below



3.1.1 Model Training/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 is required.
2. The system should be able to log each and every system flow.
3. Developer can choose logging method. We can choose Database logging/File logging as well.
4. System should not hang even after using so many loggings. Logging is used because we can easily debug issues.

3.3 Error Handling

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

4.Performance

The adult census income prediction should be accurate in classification of salary of a person. So that it doesn't mislead the authorities and users. And also Model retraining approach used t improve performance.

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 of the project will be using python 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.Dashboards

6.Conclusion

The solution proposed here is ACI (Adult Census Income prediction) will classify persons, and if there any person with less than 50k income it categorizes into low category and vice versa. This study helps in knowing the economy of the country and it helps in solving income equality problem