High Level Design (HLD)

Adult Census Income Prediction

Revision Number:
Last date of revision:

Document Version Control

Date Issued	Version	Description	Author
07-02-2024	1	Initial HLD-V1.0	Rushikesh Shinde

Contents

Abstract

Predicting **income** levels from demographic information is a crucial task in socioeconomic research and policymaking. This abstract presents an overview of the Adult Census Income Prediction dataset, a widely used benchmark dataset in the machine learning community. The dataset comprises anonymized features such as age, education, occupation, and marital status, along with a binary income class label indicating whether an individual earns more than **$50,000 annually**.

The objective of this dataset is to develop predictive models that accurately classify individuals into income categories based on their demographic attributes. Various machine learning algorithms, including decision trees, support vector machines, and neural networks, have been applied to this dataset for predictive modelling. Additionally, feature engineering techniques and data preprocessing methods have been explored to enhance model performance.

In this abstract, we provide an overview of the dataset's characteristics, including the distribution of features and the class imbalance in the income labels. Furthermore, we summarize key findings from prior research on this dataset, highlighting the performance of different machine learning algorithms and feature selection strategies.

Understanding and accurately predicting income levels from demographic data have significant implications for social welfare programs, targeted marketing strategies, and economic policy formulation. By leveraging the Adult Census Income Prediction dataset, researchers and practitioners can develop robust predictive models that contribute to informed decision-making in various domains.

1. Introduction

1.1. Why this High-Level Design Document?

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

The HLD will

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

1.2. Scope

The HLD document presents the structure of the system, such as the database architecture, application architecture (layers), application flow (Navigation), and technology architecture. The HLD uses non-technical to mildly technical terms which should be understandable to the administrators of the system.

1.3. Definition

- ACIP – Adult Census Income Prediction

2. General Description

2.1. Product Perspective

An individual's annual income results from various factors. Intuitively, it is influenced by the individual's education level, age, gender, occupation, etc.

2.2. Problem Statement

To create an AI solution for classifying individuals based on their education, age, gender, education level, occupation etc. Into income class less than or more than 50K.

- This will help organization like government to rectify the income of an individual in the country.

2.3. Proposed Solution

The solution proposed here is a data science model based on machine learning can be implemented to perform above mention use cases. In first use case, we will take input from a user with income less than 50K and see whether proposed solution is going to predict income class or not. And in second use case, we will take input from user with more than 50K income and check our solution whether it is performing or not in right way.

2.4. Future Improvements

Collaborating with domain experts from sociology, economics, or public policy fields could provide valuable insights for refining the predictive model. Integrating domain-specific knowledge into feature engineering and model development processes can lead to more contextually relevant and accurate predictions.

Considering the temporal dynamics of demographic trends and economic factors could improve the relevance and accuracy of predictions over time. Incorporating time-series analysis or including additional temporal features may capture evolving patterns in income distribution.

2.5. Data Requirements

This intermediate level data set was extracted from the census bureau database. There are 32561 instances of data set, mix of continuous and discrete.

The data set has 15 attribute which include age, sex, education level and other relevant details of a person. The data set will help to improve your skills in Exploratory Data Analysis, Data Wrangling, Data Visualization and Classification Models.

- **Age**: the age of an individual.
- **Work class**: The type of work or employment of an individual.
- **Final Weight**: The weights on the CPS files are controlled to independent estimates of the civilian noninstitutional population of the US. These are prepared monthly for us by Population Division here at the Census Bureau. We use 3 sets of controls.
- **Education**: The highest level of education completed.
- **Education-Num**: The number of years of education completed.
- **Marital-Status**: The marital status.
- **Occupation**: Type of work performed by an individual.
- **Relationship**: The relationship status.
- **Race**: The race of an individual.
- **Sex**: The gender of an individual.
- **Capital-gain**: The amount of capital gain (financial profit).
- **Capital-loss**: The amount of capital loss an individual has incurred.
- **Hours-per-week**: The number of hours works per week.
- **Native country**: The country of origin or the native country.
- **Income**: The income level of an individual and serves as the target variable. It indicates whether the income is greater than $50,000 or less than or equal to $50,000, denoted as (>50K, <=50K).

2.6. Tools Used

Python programming language and frameworks such as NumPy, Pandas, Scikit-learn, Matplotlib, Plotly, Flask etc are used to build the whole model.

- PyCharm is used as IDE.
- Virtual Studio Code is also used as IDE.
- For visualization of the plots, Matplotlib, Seaborn and Plotly are used.
- AWS is used for deployment of the model.
- Tableau/Power BI is used for dashboard creation.
- Python, Flask is used for backend development.
- GitHub is used as Version Control System.

2.7. Constraints

The system must be correct enough that it must not mislead any prediction and as automated as possible and user should not require any knowledge of the working.

2.8. Assumptions

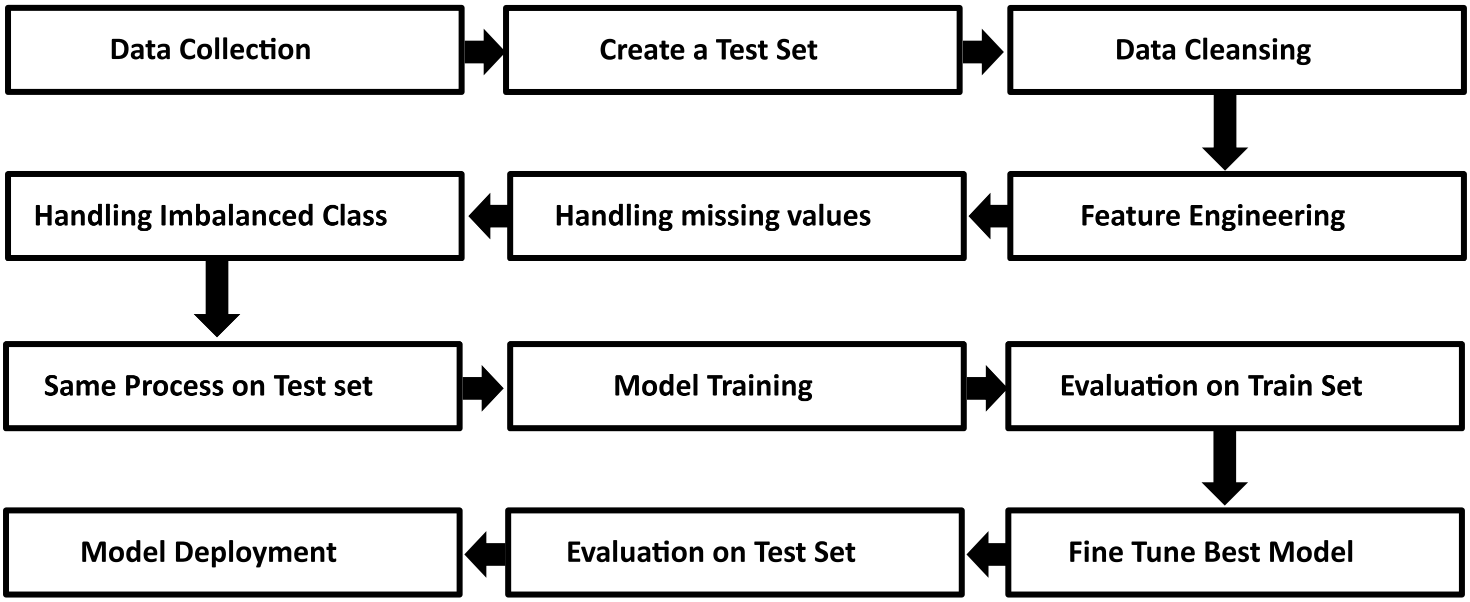
The main objective of the project is to implement the use cases as previously mentioned for new dataset that comes through organizations which has this solution install in their campus to capture people’s income.

3. Design Details

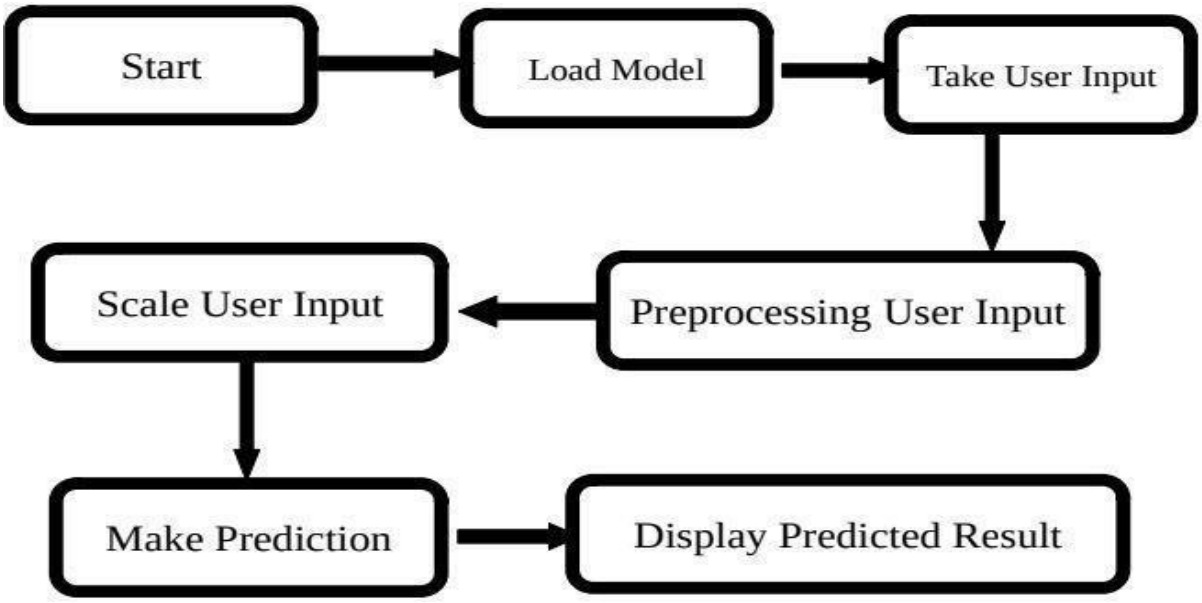
3.1. Process Flow

For predicting the income for an individual, we will be using machine learning model. Below is the flow diagram.

3.1.1.1. Model Training and Evaluation



3.1.1.2. Deployment Process



3.2. Event Log

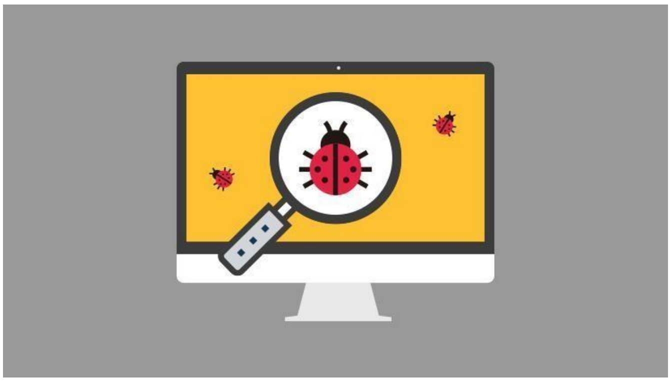
The system should log every event so that the user will know what process is running internally.

Initial Step-By-Step Description

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

3.3. Error Handing

Should errors 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. Performance

The machine learning based Adult Census Income Prediction solution will used for predicting an individual's income for classifying income class. Also, model retraining is very important to 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 for this 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

Microsoft
Azure

amazon
web services

5. Conclusion

In conclusion, the Adult Census Income Prediction dataset serves as a valuable resource for exploring the complexities of predicting income levels based on demographic attributes. Through extensive research and analysis, the dataset has provided insights into the factors influencing income distribution and has facilitated the development of predictive models aimed at addressing socioeconomic challenges.

The findings from various studies utilizing this dataset underscore the importance of feature engineering, model selection, and evaluation techniques in accurately predicting income levels. While machine learning algorithms have shown promising results, there is still room for improvement, particularly in addressing class imbalance, enhancing model interpretability, and ensuring ethical considerations are prioritized.

Moving forward, future research should focus on implementing advanced techniques such as ensemble learning, hyperparameter tuning, and domain-specific knowledge integration to further enhance predictive accuracy and model performance. Additionally, efforts should be made to promote transparency, fairness, and accountability in predictive modelling processes to mitigate biases and ensure equitable outcomes.

By continuing to leverage the Adult Census Income Prediction dataset and incorporating advancements in machine learning methodologies, researchers and practitioners can contribute to the development of more robust and effective models for income prediction. Ultimately, these efforts can inform policymaking, aid in resource allocation, and contribute to the advancement of socioeconomic well-being for diverse populations.

6. Reference

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