High Level Design (HLD)

Adult Census Income Prediction

Revision Number**: 2.0**

Last date **of** revision: 23/02/2023

Document Version Control:

Version	Date	Author	Description
1.0	5/10/2022	Pravin Sharma	Initial draft

ADULT CENSUS INCOME PREDICTION PROJECT

TABLE OF CONTENTS:

Abstract:

This High-Level Design Document outlines the approach for an adult census income prediction project. The goal of this project is to develop a machine learning model that predicts whether an individual's income exceeds $50,000 per year based on their demographic and employment data. The model will be trained on a dataset of adult census data and will be deployed as a web service. This document describes the problem statement, proposed solution, technical requirements, data requirements, design details, deployment, and KPIs.

1 Introduction

1.1 Why this High-Level Design Document?

This High-Level Design Document provides a comprehensive overview of the adult census income prediction project. It describes the problem statement, proposed solution, technical requirements, data requirements, design details, deployment, and KPIs. This document serves as a reference for the development team and stakeholders to understand the scope, constraints, assumptions, and solution design.

1.2 Scope:

The scope of this project is to develop a machine learning model that predicts whether an individual's income exceeds $50,000 per year based on their demographic and employment data. The model will be trained on a dataset of adult census data and will be deployed as a web service.

1.3 Definitions:

Machine Learning: A type of artificial intelligence that allows software applications to learn from the data and become more accurate in predicting outcomes without human intervention.

Web Service: A software system designed to support interoperable machine-to-machine interaction over a network.

2 General Description:

2.1 Product Perspective:

The adult census income prediction project is a machine learning model that predicts whether an individual's income exceeds $50,000 per year based on their demographic and employment data. The model will be trained on a dataset of adult census data and will be deployed as a web service.

2.2 Problem Statement:

The problem is to develop a machine learning model that predicts whether an individual's income exceeds $50,000 per year based on their demographic and employment data.

2.3 Proposed Solution:

The proposed solution is to develop a machine learning model using Python and scikit-learn libraries. The model will be trained on a dataset of adult census data and will use classification algorithms such as logistic regression, decision trees, and random forests to predict the income level of an individual. The trained model will be deployed as a web service using Flask.

2.4 Further Improvements:

Future improvements to the adult census income prediction project could include incorporating additional data sources, refining the feature selection process, and evaluating the model's performance on a regular basis.

2.5 Technical Requirements:

The technical requirements for the adult census income prediction project include the following:

Python 3.7 or higher

scikit-learn library

Flask web framework

Pandas data analysis library

Numpy numerical computing library

2.6 Data Requirements:

The data requirements for the adult census income prediction project include a dataset of adult census data with demographic and employment information for each individual.

2.7 Tools used:

Python: Programming language used for machine learning model development

scikit-learn: Library for machine learning model development

Flask: Web framework used for deploying the machine learning model as a web service

Pandas: Library for data analysis and manipulation

Numpy: Library for numerical computing

2.8 Hardware Requirements:

The hardware requirements for the adult census income prediction project are minimal and include a computer with at least 4GB of RAM and a processor capable of running Python and scikit-learn.

2.9 Constraints:

The constraints for the adult census income prediction project include:

The availability and quality of the dataset

The performance limitations of the hardware used for training and deploying the machine learning model

The time and budget constraints for development and deployment

Privacy and security considerations for the handling of sensitive personal data

2.10 Assumptions:

The assumptions for the adult census income prediction project include:

The dataset is representative of the population and contains accurate and complete data

The machine learning algorithms used are effective in predicting income levels

The deployment environment meets the technical requirements and can handle the expected traffic

3 Design Details:

3.1 Process Flow:

The process flow for the adult census income prediction project includes the following steps:

Data preparation: Cleaning and preprocessing the dataset to prepare it for training

Feature selection: Identifying the relevant features to use in the machine learning model

Model training: Using scikit-learn to train a classification algorithm on the dataset

Model evaluation: Testing the performance of the trained model on a validation set

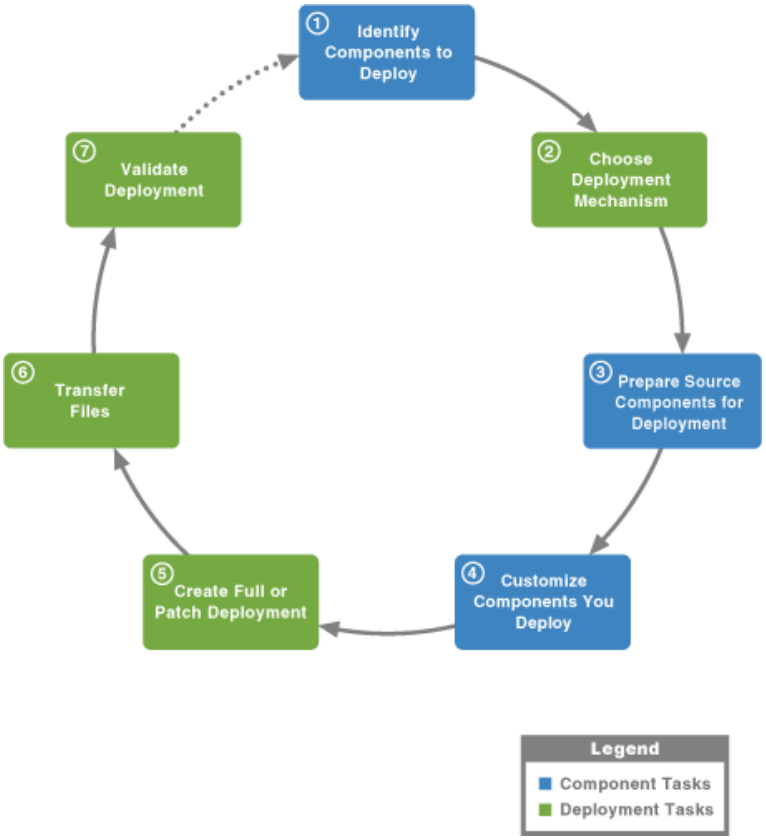
Model deployment: Using Flask to deploy the trained model as a web service

3.1.1 Model Training and Evaluation:

The machine learning model will be trained using classification algorithms such as logistic regression, decision trees, and random forests. The model will be evaluated using metrics such as accuracy, precision, recall, and F1-score.

3.1.2 Deployment Process:

The trained model will be deployed as a web service using Flask. The deployment process will include setting up a RESTful API, handling HTTP requests and responses, and integrating the model with the API.



3.2 Event Log:

An event log will be maintained to track the performance of the machine learning model over time. The event log will include metrics such as accuracy, precision, recall, and F1-score, as well as information on any errors or issues encountered during deployment.

3.3 Error Handling:

Error handling will be implemented to handle any errors or exceptions that occur during the deployment process. This will include logging errors and returning appropriate error messages to the client.

3.4 Performance:

Performance of the machine learning model will be monitored using metrics such as response time, throughput, and error rate. Performance testing will be conducted to identify and mitigate any performance bottlenecks.

3.5 Reusability:

The machine learning model and deployment code will be designed to be reusable for other similar projects. The code will be modular and well-documented to facilitate reuse.

3.6 Application Compatibility:

The machine learning model will be compatible with a variety of applications and platforms. The model will be deployed as a web service using RESTful APIs, making it accessible to a wide range of clients.

3.7 Resource Utilization:

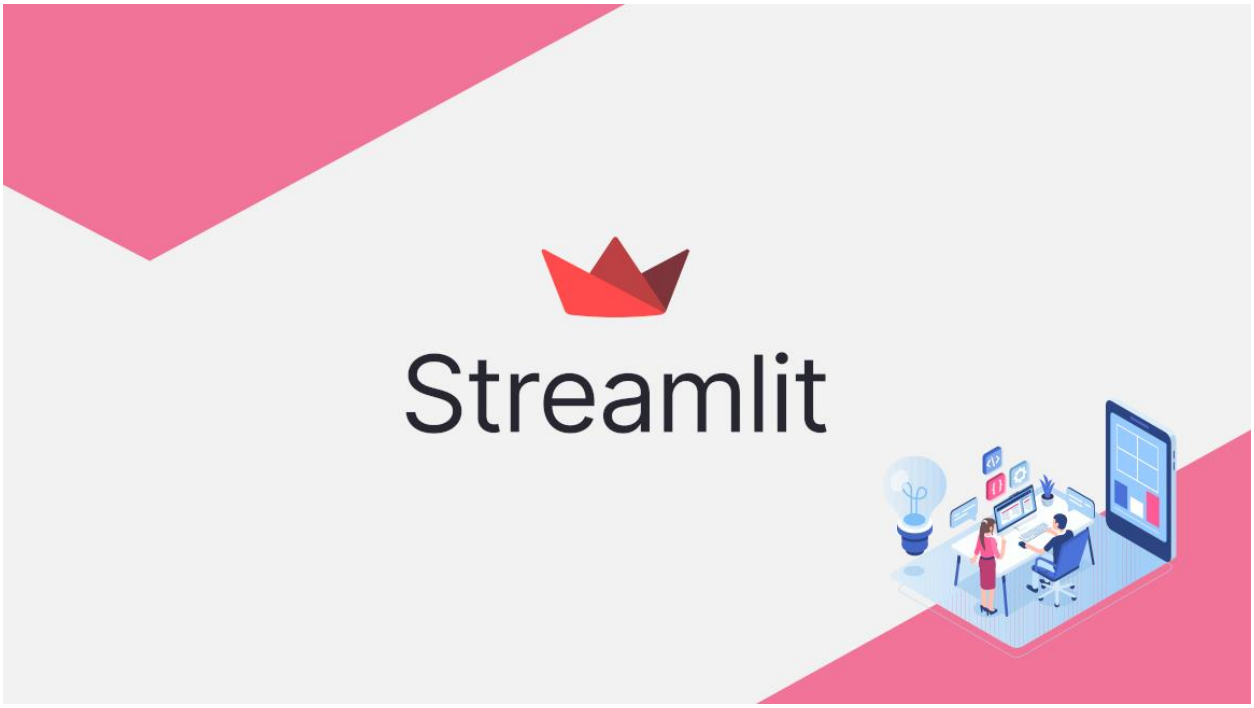
The machine learning model and deployment code will be designed to minimize resource utilization, including CPU, memory, and network resources.

4 Deployment:

The machine learning model will be deployed as a web service using Flask. The deployment process will include setting up a server environment, configuring the web server and application server, and deploying the code.

5 Dashboards:

Dashboards will be created to visualize the performance of the machine learning model over time. The dashboards will include metrics such as accuracy, precision, recall, and F1-score, as well as information on any errors or issues encountered during deployment.



4.1 KPIs (Key Performance Indicators):

The KPIs for the adult census income prediction project include:

Model accuracy: The percentage of correct predictions made by the machine learning model

Response time: The time it takes for the model to respond to a prediction request

Throughput: The number of prediction requests the model can handle per

6.Conclusion:

The Adult Census Income Prediction project aims to develop a machine learning model that predicts an individual's income level based on demographic and employment data. The project will use a supervised

learning approach and evaluate various classification algorithms to determine the most suitable algorithm. The expected outcomes of the project include developing an accurate predictive model, identifying important features, providing insights into contributing factors, and demonstrating the potential of machine learning in predicting income levels for policy-making and decision-making purposes