Low Level Design (LLD)

Adult Census Income Prediction

Document Version Control

Date Issued	Version	Description	Author
20th Aug 2021	1.1	Initial HLD-V1.0	Umang Tank

Document Version Control

Abstract

1 Introduction

1.1 Why this Low-Level Design Document?

1.2 Scope

1.3 Constraints

1.4 1.3 Risks

1.5 1.4 Out of Scope

2 Technical specifications

2.1 Dataset

2.2 Logging

2.3 Database

3 Technology stack

4 Proposed Solution

5 Model training/validation workflow

6 User I/O workflow

7 Exceptional scenarios

1 Introduction

1.1 Why this Low-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.

An Adult Census Income contains the information, such as:

- Age
- Capital-gain
- Capital-loss
- Education-year
- Working-hours
- Sex
- Race
- Country
- Occupation
- Work-class
- Education
- Merital-status
- relation

This project shall be delivered in two phases:

Phase 1: All the functionalities with Scikit-learn packages.

Phase2: Integration of UI to all the functionalities.

1.2 Scope

This software system will be a Web application This system will be designed to predict whether income of people is more than 50K or not. . The models can be applied to the data collected in coming years to predict the income. This system is designed to predict the income based on some information like age, working-hours, sex, race, occupation ,capital-gain/loss etc.

1.3 Constraints

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

1.4 Risks

Document specific risks that have been identified or that should be considered.

1.5 Out of Scope

Delineate specific activities, capabilities, and items that are out of scope for the project.

2 Technical specifications

2.1 Dataset

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, Transportmoving, 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)

2.1.1 Insurance Premium dataset overview

To create the income prediction model, we obtained the data set through the Kaggle site. The data set includes 14 attributes, the data set is separated into two-part the first part called training data, and the second called test data; training data makes up about 80 percent of the total data used, and the rest for test data The training data set is applied to build a model as a predictor of Income. the test set will use to evaluate the Classification model.

Some of the records in the dataset are following

	age	workclass	education_level	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	income
0	39	State-gov	Bachelors	13.0	Never-married	Adm-clerical	Not-in-family	White	Male	2174.0	0.0	40.0	United-States	<=50K
1	50	Self-emp-not-inc	Bachelors	13.0	Married-civ-spouse	Exec-managerial	Husband	White	Male	0.0	0.0	13.0	United-States	<=50K
2	38	Private	HS-grad	9.0	Divorced	Handlers-cleaners	Not-in-family	White	Male	0.0	0.0	40.0	United-States	<=50K
3	53	Private	11th	7.0	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0.0	0.0	40.0	United-States	<=50K
4	28	Private	Bachelors	13.0	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0.0	0.0	40.0	Cuba	<=50K

2.1.2 Input schema

Feature name	Datatype	Size	Null/Required
Age	int	2	Required

2.2 Predicting income

- The system displays the form where the all features are available if user can all features correctly then machine can able to predict the income of that person.

2.3 Logging

We should be able to log every activity done by the user.

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

2.4 Database

System needs to store every request into the database and we need to store it in such a way that it is easy to retrain the model as well.

1. The User gives required information.
2. The system stores each and every data given by the user or received on request to the database. Database you can choose your own choice whether MongoDB/ MySQL. Here we use MySQL.

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

age	capital_gain	capital_loss	education_year	working_hours	occupation	sex	merital	country	race	work_class	education	relation
22	0	0	16	60	Adm-clerical	Male	Divorced	India	Amer-Indian-Eskimo	Federal-gov	10th	Husband
40	0	0	16	60	Adm-clerical	Male	Divorced	India	Amer-Indian-Eskimo	Federal-gov	10th	Husband
28	22000	0	12	40	Transport-moving	Male	Never-married	United-States	White	Private	Bachelors	Unmarried
32	22000	0	18	34	Prof-specialty	Female	Married-civ-spouse	India	Other	Federal-gov	Doctorate	Not-in-family
20	0	0	10	20	Sales	Female	Never-married	India	Amer-Indian-Eskimo	Private	Preschool	Unmarried
65	10000	0	16	40	Protective-serv	Male	Married-spouse-absent	Jamaica	Black	Self-emp-inc	Some-college	Other-relative

2.5 Deployment

1. FLASK

3 Technology stack

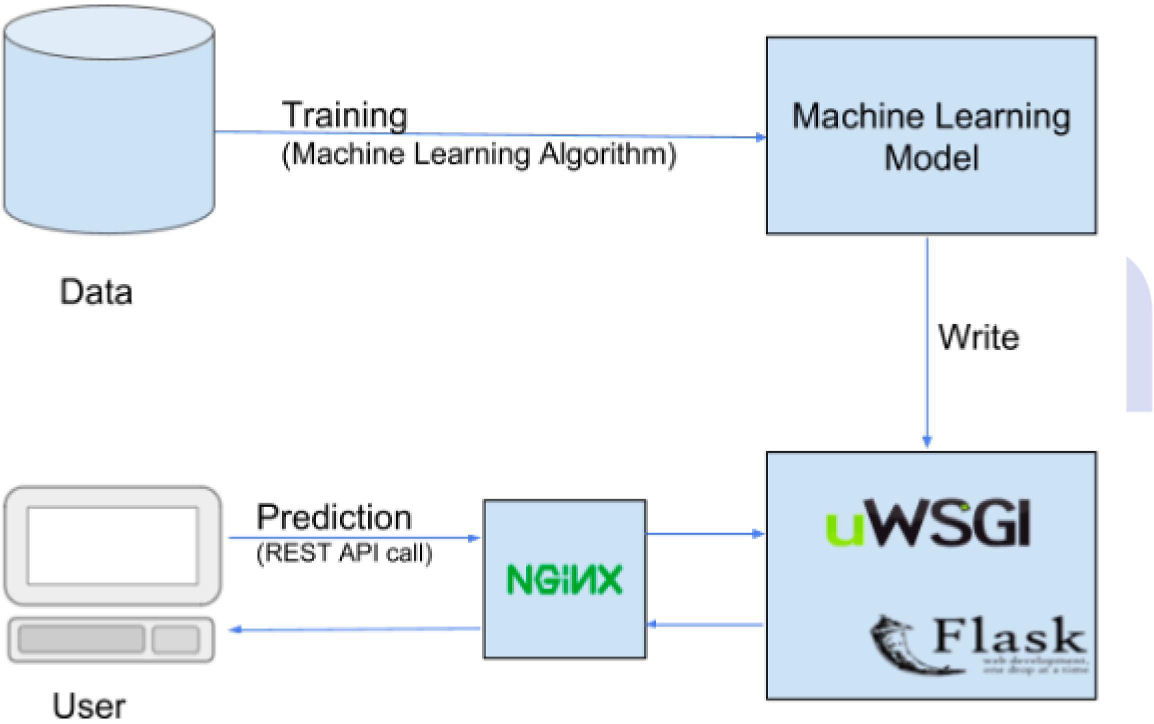
Front End	HTML/CSS/JS/React
Backend	Python Flask
Database	MySql
Deployment	Heroku

4 Proposed Solution

AdaBoostClassifier gives better accuracy as compare to other so in this project we use AdaBoostClassifier algorithm to predict income. However, drawing a baseline in the form of some Machine Learning algorithm would be helpful.

1. Actual model: AdaBoostClassifier

5 Model training/validation workflow



6 User I/O workflow