

# Adult Census Income Prediction

## High-Level Document



Written By	Gurjeet Singh
Document Version	0.1
Description	Initial HLD V0.1
Date Issued	14 November, 2022

# Content

## Abstract

1. Introduction
  - 1.1 Why this High Level Design Document?
  - 1.2 Scope
2. General Description
  - 2.1 Definitions
  - 2.2 Product Description
  - 2.3 Problem Statement
  - 2.4 Proposed Solution
  - 2.5 Further Improvement
  - 2.6 Data Requirements
  - 2.7 Tools Used
  - 2.8 Hardware Requirements
  - 2.9 Constraints
  - 2.10 Assumptions
3. Design Details
  - 3.1 Process Flow
  - 3.2 Deployment Process
  - 3.3 Event Log
  - 3.4 Error Handling
4. Performance
  - 4.1 Re-usability
  - 4.2 Application Compatibility
  - 4.3 Resource Utilization
  - 4.4 Deployment
5. KPI's
6. Conclusion

## **Abstract**

Data has always been the backbone of many important decisions. When an assumption is backed up by facts and numbers, the chances of incorrectness and bad decisions decrease and in today's world, countless decisions in private and public sectors are based on Census data. Census data is the backbone of the democratic system of government, highly affecting the economic sectors. Census-related figures are used to distribute the federal funding by the government into different states and localities.

The above introduction had an aim to increase the awareness about how the income factor actually has an impact not only on the personal lives of people, but also an impact on the nation and its betterment.

We will now have a look on the data extracted from the 1994 Census bureau database, and try to find insights about how different features have an impact on the income of an individual and also do some predictive analysis using the modern Data Science and Machine Learning techniques to predict whether an individual's income will be greater than \$50,000 per year based on several attributes from the census data.

# Introduction

## 1.1 Why this High-Level Design Document?

The purpose of this High-Level Design (HLD) Document is to add necessary detail 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 a reference manual for how the modules interact at a high level.

The HLD will:

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

## 1.2 Scope

The HLD documentation presents the structure of the system, such as 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.

# General Description

## 2.1 Definitions

Term	Description
ACIP	Adult Census Income Predictor
Database	Collection of the Information monitored by the system
Cloud	A data centre full of servers connected to the internet performing a service
IDE	Integrated Development Environment
UI	User Interface
Heroku	A cloud service

## 2.2 Product Description

The ACIP is a Machine Learning based classification model which helps us to do predictive analysis on the Income of a person using certain parameters.

## 2.3 Problem Statement

To create an AI based solution for predictive analysis of a person's annual income and also deploy it in the form of a UI.

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.

## 2.4 Proposed Solution

Using all the standard techniques used in the life-cycle of a Data Science project starting from Data Exploration, Data Cleaning, Feature Engineering, Feature Selection, Model Training, Model Testing and Model Deployment and also building a frontend where a user can fill their information in the form input and get the output instantly.

## 2.5 Further Improvements

The ACIP can be easily embedded inside any website or an application and can be used to find out whether a person earns more than 50K\$ annually or not and can be used by various governmental/non- governmental/private agencies around the world. This can also be improved further by feeding the model with more data, this can data from various UCI repositories, from various banks, data scraped from internet, etc.

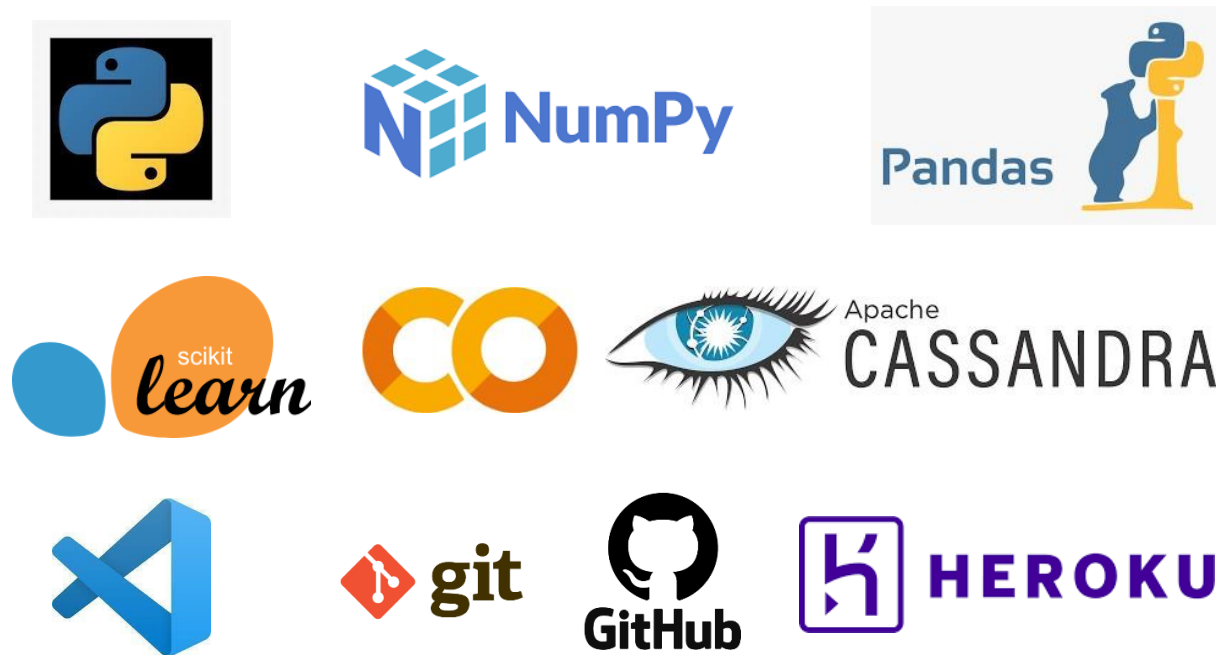
## 2.6 Data Requirements

Data requirement completely depend on our problem statement. We need the dataset from the census of 1994, to train our model. Each record in the dataset must have the following features which are important to determine one's income. The dataset is downloaded from Kaggle and contains 15 columns and 32000+ rows.

- Age: Age of the person whom the income has to be determined.
- Hours per week: The number of hours the person work per week.
- Capital Gain/Loss: whether the person had any capital gain or loss or none.
- Country: country in which the person lives.
- Educational Level: highest level of education the person has attained.
- Marital Status: whether the person is married or single
- Occupation: occupation of the person.
- Workclass: working class of the person.
- Gender: gender of the person.
- Race: race to which the person belongs.

These are the set of parameters required in order to predict one's annual income.

## 2.7 Tools Used



- Google Colab and Vscode were used as an IDE.
- For visualization of the plots Matplotlib, Seaborn and Plotly were used.
- Heroku was used for deployment of the model.
- Apache Cassandra was used to storage and retrieval of data.
- Flask and HTML were used for building the web application and server to run the code.
- Github was used as version control system.
- Numpy and Pandas were used to clean and interpret the data.
- Scikit learn was used to cross validate and compare different models.

## 2.8 Hardware Requirements

- Windows Server, Linux, or any operating system that can run as a webserver, capable of delivering HTML5 content.
- Minimum 1.10 GHz processor or equivalent.
- Between 1-2 GB of free storage
- Minimum 512 MB of RAM
- 3 GB of hard-disk space

## 2.9 Constraints

The front-end must be user friendly and should not need any one to have any prior knowledge in order to use it.

## 2.10 Assumptions

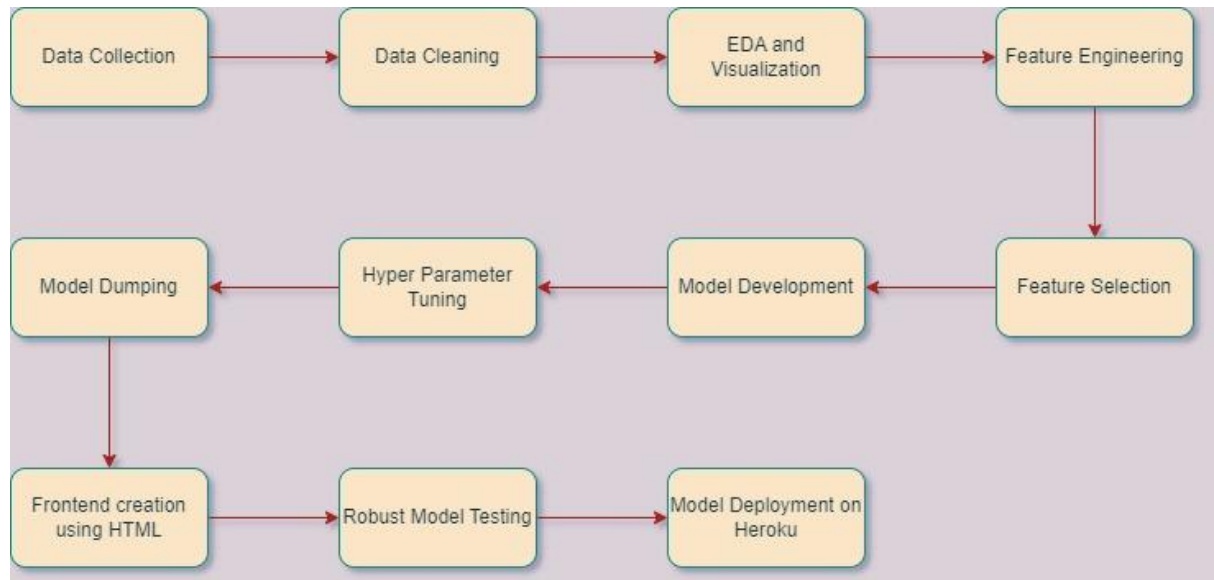
The main objective of this project is to implement the use case as previously mentioned (2.3 problem statement) for new dataset that comes through the UI. It is assumed that all aspects of this project have the ability to work together as the designer is expecting and also the data on which our model is trained is as correct as possible.



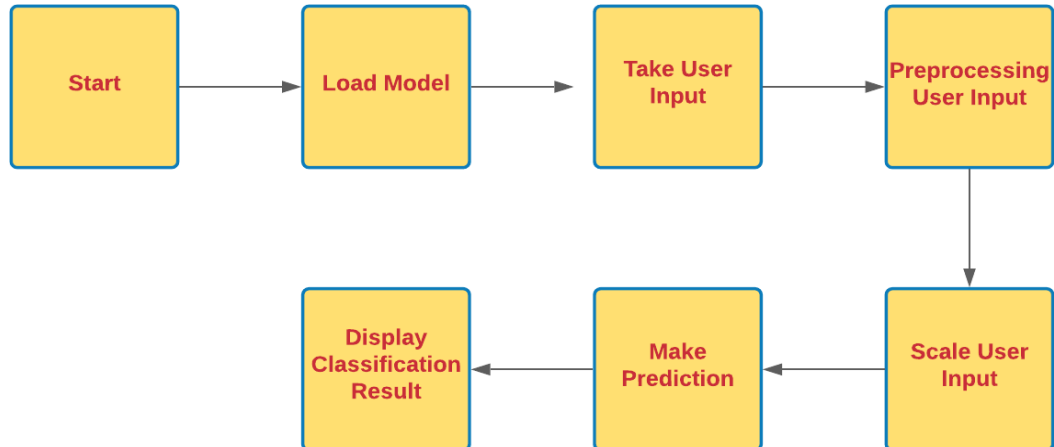
### 3. Design Details

#### 3.1 Process Flow

The process flow diagram is shown below:



#### 3.2 Deployment Process



#### 3.3 Event Log

The system should log every event so that the user will know what process is running internally. Initial step-by-step description:

The system identifies at what level logging is required

- The system should be able to log each and every system flow

- Developer can chose logging method. You can chose database logging/ File logging as well
- System should not hang even after so many loggings. Logging just because we can easily debug issues so logging is mandatory to do.

### 3.4 Error Handling

The project is designed in such a way that, at any step if error occurs then our application should not terminate rather it should catch the error and display that error with proper explanation as to what went wrong during process flow.

## 4. Performance

The ACIP tool is used to predict whether a person earns above or below 50K dollars per annum or not. So this is made keeping in mind that if it will be used by various governmental/ non-governmental/ private agencies then it is supposed to be as accurate as possible. So that it doesn't mislead authorities. Also model retraining is very important to further enhance its 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 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 to it until finished.

#### 4.4 Deployment



### 5. KPIs (Key Performance Indicators)

5.1 Key Performance Indicators of ACIP.

5.2 Latency or the amount of time the application takes to display results for some specific input.

5.3 The processing power our application takes to run.

5.4 The memory and RAM our application takes to run on a web server.

### 6. Conclusion

The ACIP will give the income predictions of a person instantly and has the potential to help various organizations, agencies and companies around the world in various tasks.