**High Level Design (HLD)**

**Adult Censes Income Prediction**



**Revision Number - 1.0**

**Last Date of Revision - 05/07/2022**

**SRIPHANI**

# Document Version Control

|  |  |  |  |
| --- | --- | --- | --- |
| **Date Issued** | **Version** | **Description** | **Author** |
| 05-07-2022 | 1.0 | HLD document building | Sriphani |
|  |  |  |  |
|  |  |  |  |

**Contents**

Document Version Control…………………………………………………………………………2

Abstract……………………………………………………………………………………………….4

1. Introduction……………………………………………………………………………………….4
   1. Why this High-Level Design Document?..............................................................4
   2. Scope……………………………………………………………………………………5
2. General Description……………………………………………………………………………...5
   1. Product Perspective & Problem Statement…………………………………………5
   2. Tools Used……………………………………………………………………………...6
3. Design Detail……………………………………………………………………………………...6
   1. Functional Architecture………………………………………………………………...7
   2. Optimization………………………………………………………………………….….8
4. KPI……………………………………………………………………………………………….…9
   1. KPIs (Key Performance Indicators) ……………………………………………..…...9
5. Deployment………………………………………………………………………………...........10

# 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 and data mining 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 50K Dollars 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 (HLD) Document is to add the necessary detail to the current project description to represent a suitable model for coding. This document is also intended to help detect contradictions before coding and can be used as a reference manual for how the modules interact at a high level.

**The HLD will:**

* Present all of the design aspects and define them in detail
* Describe the user interface being implemented
* Describe the hardware and software interfaces
* Describe the performance requirements
* Include design features and the 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 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.

# 2 General Description

## 2.1 Product Perspective & Problem Statement

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.

Approach:

The classical machine learning tasks like Data Exploration, Data Cleaning, Feature Engineering, Model Building and Model Testing.

## 2.2 Tools used

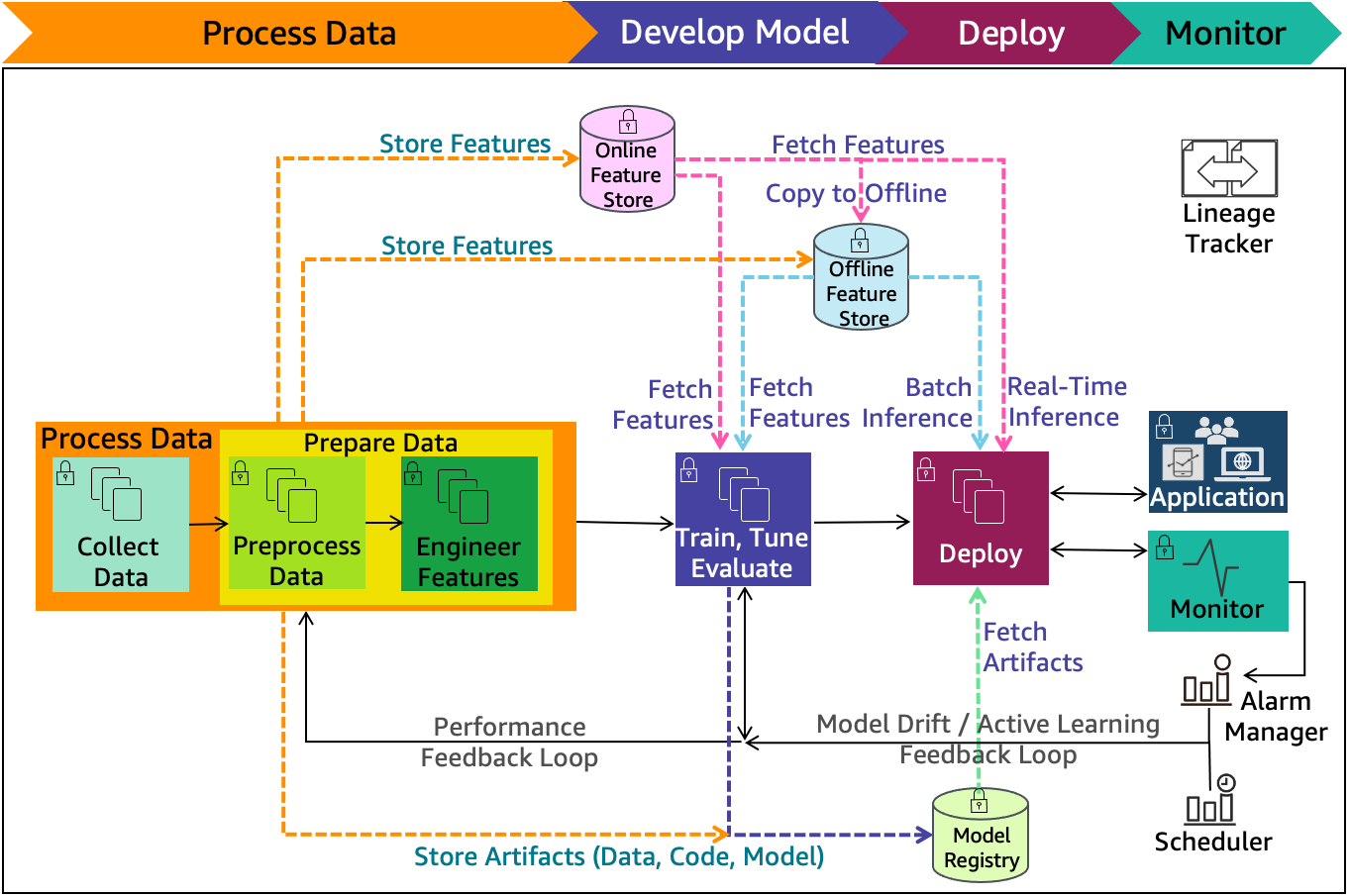
Machine Learning tools and libraries works such as NumPy, Pandas, Seaborn, Matplotlib, MS-Excel, Pandas-Profiling, evalml, scikit-learn, Jupyter Notebook and Python Programming Language are used to build the whole framework.



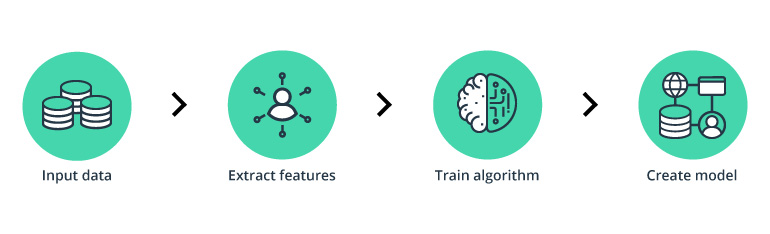


**3 Design Details**

## 3.1 Functional Architecture



**How MACHINE LEARNING Works**



## 3.2 Optimization

* It has become one of the most popular research directions and plays a significant role in many fields, such as machine translation, speech recognition, image recognition, recommendation system, etc.
* Optimization is one of the core components of machine learning.
* The essence of most machine learning algorithms is to build an optimization model and learn the parameters in the objective function from the given data.
* In the era of immense data, the effectiveness and efficiency of the numerical optimization algorithms dramatically influence the popularization and application of the machine learning models.
* In order to promote the development of machine learning, a series of effective optimization methods were put forward, which have improved the performance and efficiency of machine learning methods.
* From the perspective of the gradient information in optimization, popular optimization methods can be divided into three categories: first-order optimization methods, which are represented by the widely used stochastic gradient methods
* Variational inference is a useful approximation method which aims to approximate the posterior distributions in Bayesian machine learning.
* It can be considered as an optimization problem.
* For example, mean-field variational inference uses coordinate ascent to solve this optimization problem
* As the amount of data increases continuously, it is not friendly to use the traditional optimization method to handle the variational inference.
* Thus, the stochastic variational inference was proposed, which introduced natural gradients and extended the variational inference to large-scale data

# 4 KPI

* Any machine learning model after deployment in a business becomes a product or service.
* This product/service is directly interacting with the user and a company’s business is at the behest of its user’s experience.
* There are ML models which work internally(not directly with the user as a product or service) to improve user’s experience or create brand awareness only to support companies’ main objectives.
* Improving model performance doesn’t always support the business objective. It seems intuitive that improving the model based on the performance metrics will yield a favourable outcome.
* However, it is witnessed by data scientists that it is not always the case.
* The differentiator here is the business KPI’s, oftentimes optimising the model have no result on the KPI’s which means the improvements are no good for a business stakeholder of the company, in fact, it would be viewed as a waste of resources and time which is equivalent to money in the harsh business world.
* KPIs lay like a wall in front of the model, in order to reach the business objective it must affect and cross the KPIs.

## 4.1 KPIs (Key Performance Indicators)

Key indicators displaying a summary of the Adult income prediction and its relationship with different metrics

1. Percentage of People Having >50K group or <=50K group.
2. Age Distribution including Gender
3. Gender Distribution Based on >50K group or <=50K group.

# 5 Deployment

* Heroku is a Platform-as-a-Service tool by Salesforce.
* Heroku is backed by AWS and all Heroku applications/services are hosted on AWS.
* AWS provides the infrastructure and handles all the load-balancing, resource utilization, networking, logging, monitoring and **Heroku acts as a middle-man to provide a scalable, automated rapid deployment platform with all cloud capabilities**.
* Using Flask will provide UI to test and it can be integrated with enterprise-level applications.

