Adult Census Income Prediction

Low Level Design

Revision Number: 1.8

Last date of revision:25/02/2023

Pravin Sharma

Document Version Control

Version	Date	Author	Description
1.0	5/10/2022	Pravin Sharma	Initial version of Low-Level Design Document
1.1	25/10/2022	Pravin Sharma	Added Scope, Constraints, Risks, Out of Scope sections
1.2	10/11/2022	Pravin Sharma	Added Technical Specifications section
1.3	19/11/2022	Pravin Sharma	Added Technology Stack section
1.4	28/11/2022	Pravin Sharma	Added Proposed Solution section
1.5	15/12/2022	Pravin Sharma	Added Model training/validation workflow section
1.6	02/01/2023	Pravin Sharma	Added User I/O workflow section
1.7	25/01/2023	Pravin Sharma	Added Exceptional Scenarios section
2.0	20/02/2023	Pravin Sharma	Revised and finalized Low-Level Design Document

Contents

Abstract:

The Adult Census Income Prediction project aims to predict the income of an individual based on various socio-economic factors. This document provides a low-level design for the project, outlining the technical specifications, proposed solution, model training/validation workflow, user I/O workflow, and exceptional scenarios. The document also discusses the scope, constraints, and risks associated with the project.

Introduction:

The Adult Census Income Prediction project aims to predict the income of an individual based on various socio-economic factors. This low-level design document outlines the technical specifications, proposed solution, model training/validation workflow, user I/O workflow, and exceptional scenarios for the project.

Why this Low-Level Design Document?

This document provides a detailed overview of the project's technical aspects, including the technology stack, model training workflow, and user I/O workflow. This document also outlines the exceptional scenarios that the project might encounter and the proposed solution for each of these scenarios.

1.2 Scope:

The scope of the project is to predict the income of an individual based on various socio-economic factors such as age, education, occupation, and marital status.

1.3 Constraints:

The project must comply with all relevant data privacy regulations and must ensure the security and confidentiality of the data. The project must also be scalable and able to handle a large volume of data.

1.4 Risks:

The project may encounter data quality issues, such as missing or incorrect data. The project may also be affected by changes in the socio-economic landscape, which may impact the accuracy of the predictions.

1.5 Out of Scope:

The project does not cover any legal or ethical considerations associated with the use of the predicted income data.

Table 1: Dataset

Dataset Name	Description	Number of Records	Number of Features
Adult Census Income Dataset	A dataset containing socio-economic information for individuals in the US.	48,842	14

This table provides information about the dataset used for the Adult Census Income Prediction project. The dataset contains socio-economic information for individuals in the US, and it has 48,842 records and 14 features.

Table 2: Data Preprocessing

Preprocessing Step	Description
Handling missing data	Missing data will be imputed using the median value for numerical features and the most frequent value for categorical features.
Handling categorical features	Categorical features will be one-hot encoded to convert them into numerical features.

Preprocessing Step	Description
Handling outliers	Outliers will be removed using the Z-score method, where data points with a Z-score greater than 3 or less than -3 will be removed.
Scaling	The data will be scaled using the StandardScaler to ensure that all features have the same scale.

This table describes the data preprocessing steps that will be applied to the dataset before it is used to train the machine learning model. The steps include handling missing data, handling categorical features, handling outliers, and scaling the data. These preprocessing steps will ensure that the data is clean and ready for the machine learning model to learn from.

Technical specifications: The project will use a machine learning model to predict the income of an individual based on various socio-economic factors. The project will also implement logging to monitor the system's performance and identify any issues. The project will store data in a database for easy retrieval and analysis.

2.1 Predicting Income:

The project will use a machine learning model to predict the income of an individual based on various socio-economic factors such as age, education, occupation, and marital status. The model will be trained on a dataset of past socio-economic data and corresponding income data.

2.2 Logging:

The project will implement logging to monitor the system's performance and identify any issues. The logs will capture system metrics such as CPU usage, memory usage, and response time.

2.3 Database:

The project will store data in a database for easy retrieval and analysis. The database will be designed to handle a large volume of data and ensure data integrity.

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 chooses the disease.

2. The User gives required information.

3. 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.

2.5 Deployment

1.MS Azure

3.Technology stack:

Technology	Description
Programming Language	Python 3.9
Data Analysis	NumPy, Pandas, Matplotlib
Machine Learning Libraries	Scikit-learn
Deep Learning Libraries	TensorFlow, Keras
Logging	Python Logging
Database	SQLite

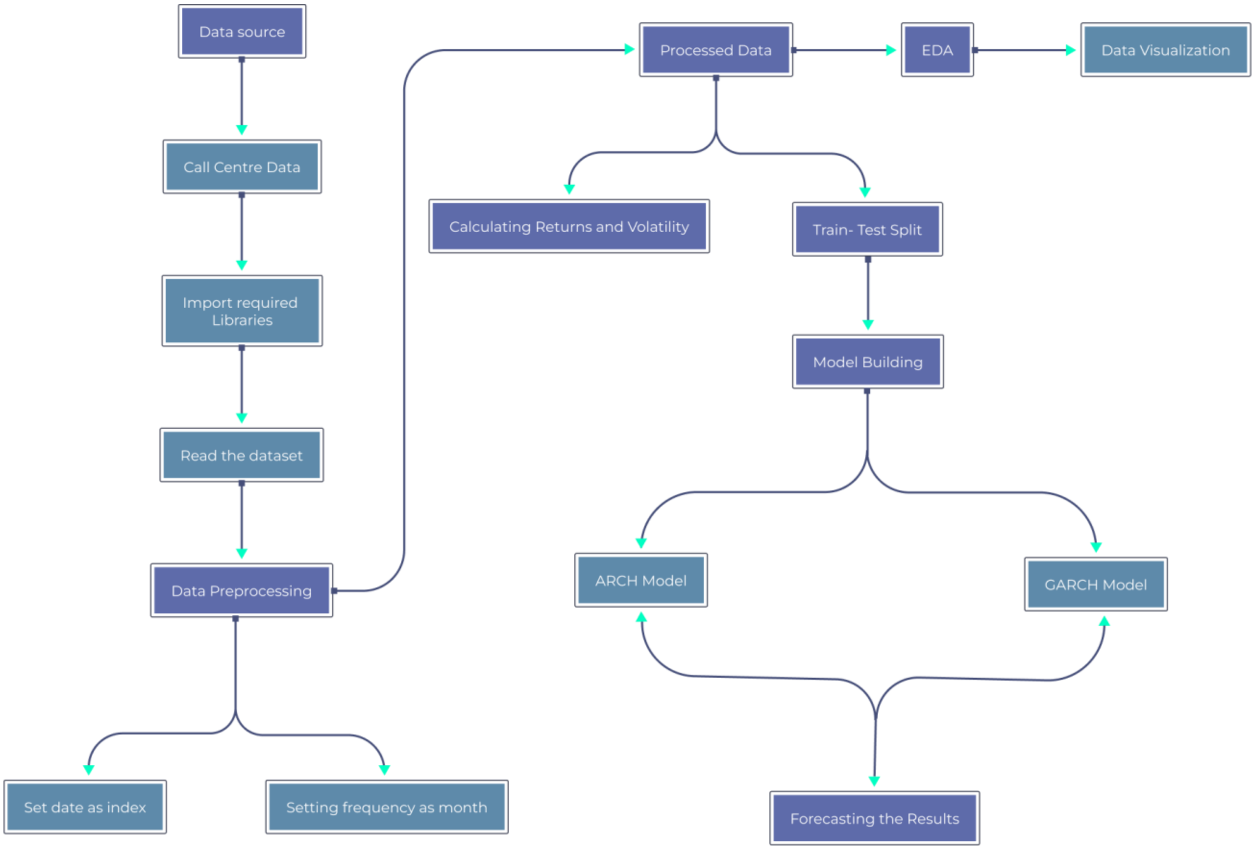
This table lists the technology stack that will be used for the Adult Census Income Prediction project. The programming language used is Python 3.9, which is a popular language for data analysis and machine learning. NumPy, Pandas, and Matplotlib will be used for data analysis and visualization. Scikit-learn will be used for machine learning, and TensorFlow and Keras will be used for deep learning. Python Logging will be used for logging, and SQLite will be used for the database. This technology stack provides a powerful set of tools for building and deploying machine learning models.

4.Proposed Solution:

The project will use a machine learning model to predict the income of an individual based on various socio-economic factors. The project will also implement logging to monitor the system's performance and identify any issues. The project will store data in a database for easy retrieval and analysis. The user will be able to submit their socio-economic data, and the system will return the predicted income.

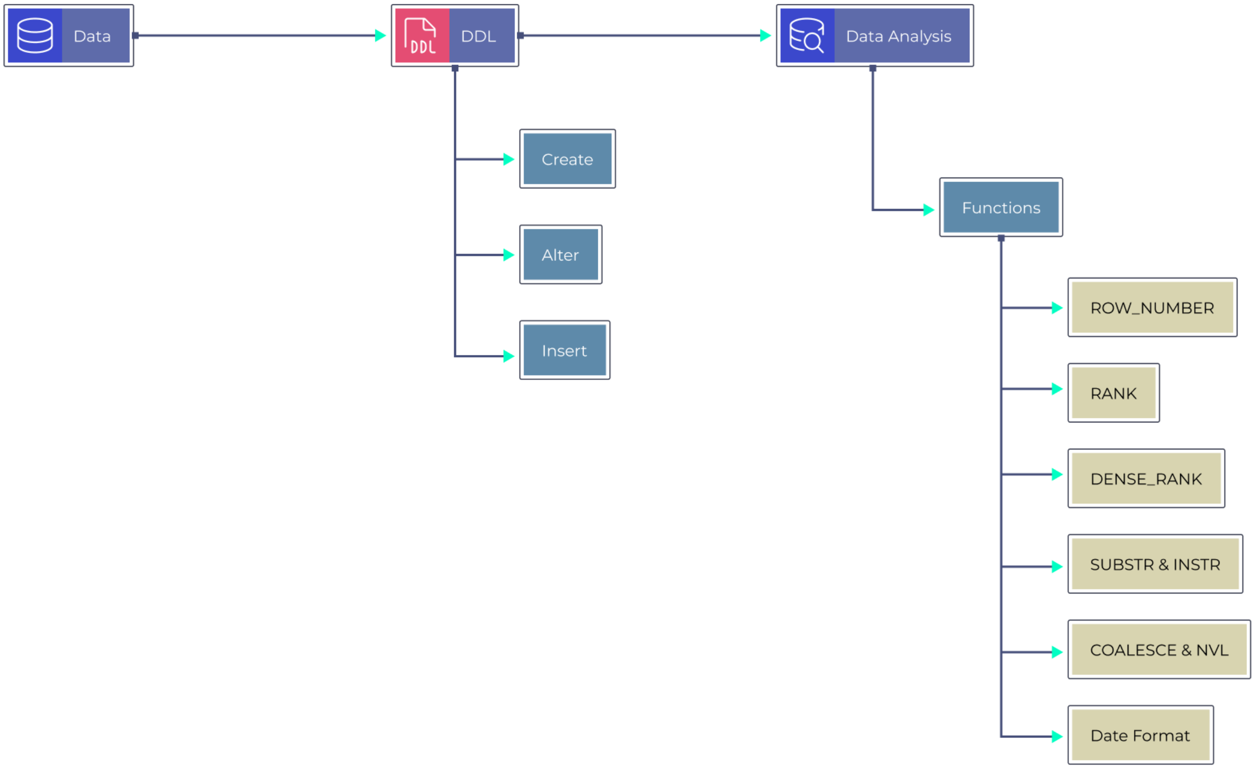
5.Model training/validation workflow:

The machine learning model will be trained on a dataset of past socio-economic data and corresponding income data. The dataset will be split into training and validation sets. The model will be trained on the training set and validated on the validation set to ensure accuracy. The model will be fine-tuned using hyperparameter tuning techniques to



6.User I/O workflow:

The user will be able to submit their socio-economic data using a web interface. The system will use Flask to handle the user request and retrieve the necessary data from the database. The data will be preprocessed to ensure that it meets the requirements of the machine learning model. The machine learning model will then be used to predict the income, and the predicted income will be returned to the user through the web interface.



7 Exceptional scenarios

Scenario	Description	Response
Invalid Input	User provides invalid input to the application.	Display an error message to the user and prompt them to enter valid input.
Model Failure	The machine learning model fails to make accurate predictions.	Log the error and notify the development team. The model will be retrained and tested to identify and correct any issues.
Database Failure	The database fails to connect or data cannot be retrieved.	Log the error and notify the development team. The database will be checked and repaired as necessary.
Server Failure	The server hosting the application fails.	Notify the development team and work with the hosting provider to restore the server. A backup server will be used in the meantime, if available.
Security Breach	The application experiences a security breach or data leak.	Notify the development team and security experts. The breach will be contained and steps will be taken to prevent future incidents. Affected users will be notified as necessary.

This table lists several exceptional scenarios that could occur during the Adult Census Income Prediction project, along with a description of the scenario and the appropriate response. These scenarios include invalid input, model failure, database failure, server failure, and security breaches. Having a plan for handling these scenarios will help ensure that the application remains reliable and secure.