iNeuron

High Level Design

1.1.   Adult Census Income Predictor

Written By	Shubham Rawat
Document Version	0.1
Last Revised Date	22 – Aug-2022

Document Control

Change Record:

Version	Date	Author	Comments
0.1	19 – Aug - 2022	Shubham Rawat	Initial HLD

Contents

Document Version Control

Abstract

1. Introduction

2.-- General Description

3.-- Design Details

Abstract

Adult Census Income prediction is based on data collected by Ronny Kohavi and Barry Becker (Data Mining and Visualization, Silicon Graphics).This data has extracted from the 1994 Census bureau database.

It is a Classification Problem, Random Forest Classifier , Logistic Regression used as a final Model for Prediction. Parameter update feature given with model and validation data downloading is possible.

In this Project we Perform CI/CD pipeline to automated the Process of Deployment on Heroku Cloud Platform.

1. Introduction

1.1. What is High-Level design document?

The goal of HLD or a high-level design document (HLDD) is to give the internal logical design of the actual program code for Adult Census Income System. HLD describes the class diagrams with the methods and relations between classes and program specs. It describes the modules so that the programmer can directly code the program from the document.

1.2. Scope

The HLD document 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

Adult Census Income Prediction is Used to Predict weather Person is earning more then or less than $50k per year. We receive a 35k row data set with lots of error and missing data which we handled properly.

2.2. Problem Statement

This data was extracted from the 1994 Census bureau database by Ronny Kohavi and Barry Becker (Data Mining and Visualization, Silicon Graphics).

 A set of reasonably clean records was extracted using the following conditions:

> Age is between 16 to 100
>
> Different Group of working Class
>
> There Education Status
>
> Marriage status and there Respective Genders.

The prediction task is to determine whether a person makes over $50K a year

2.3. Proposed Solution

In this project, we examine the correlations between several characteristics, including educational attainment, job characteristics, gender, nationality, race, age, etc. To get the best solution, the User's this year's wage is Predicted.

We deploy a working model using docker into the Heruko cloud Platform using CI/CD Pipeline

2.4. Further Improvements

We can Store our data into some database like MongoDB or MySql. Currently our data is running dynamically stored into Ram. We Can use some Cloud Platform like AWS to store our trained model so that we can use it again and again. UI can be improved as I don't have much UI skill (CSS). User to Software connection can be improved.

2.5. Tools Used

- PyCharm is used as IDE.
- For Visualization Matplotlib, Seaborn Libraries used
- Heroku for Deployment
- Front end development is done using HTML/CSS
- Flask is used back-end development
- GitHub is used as version control system

2.6. Constraints

Its Fully Automated so User doesn't need to know any prior knowledge before using it.
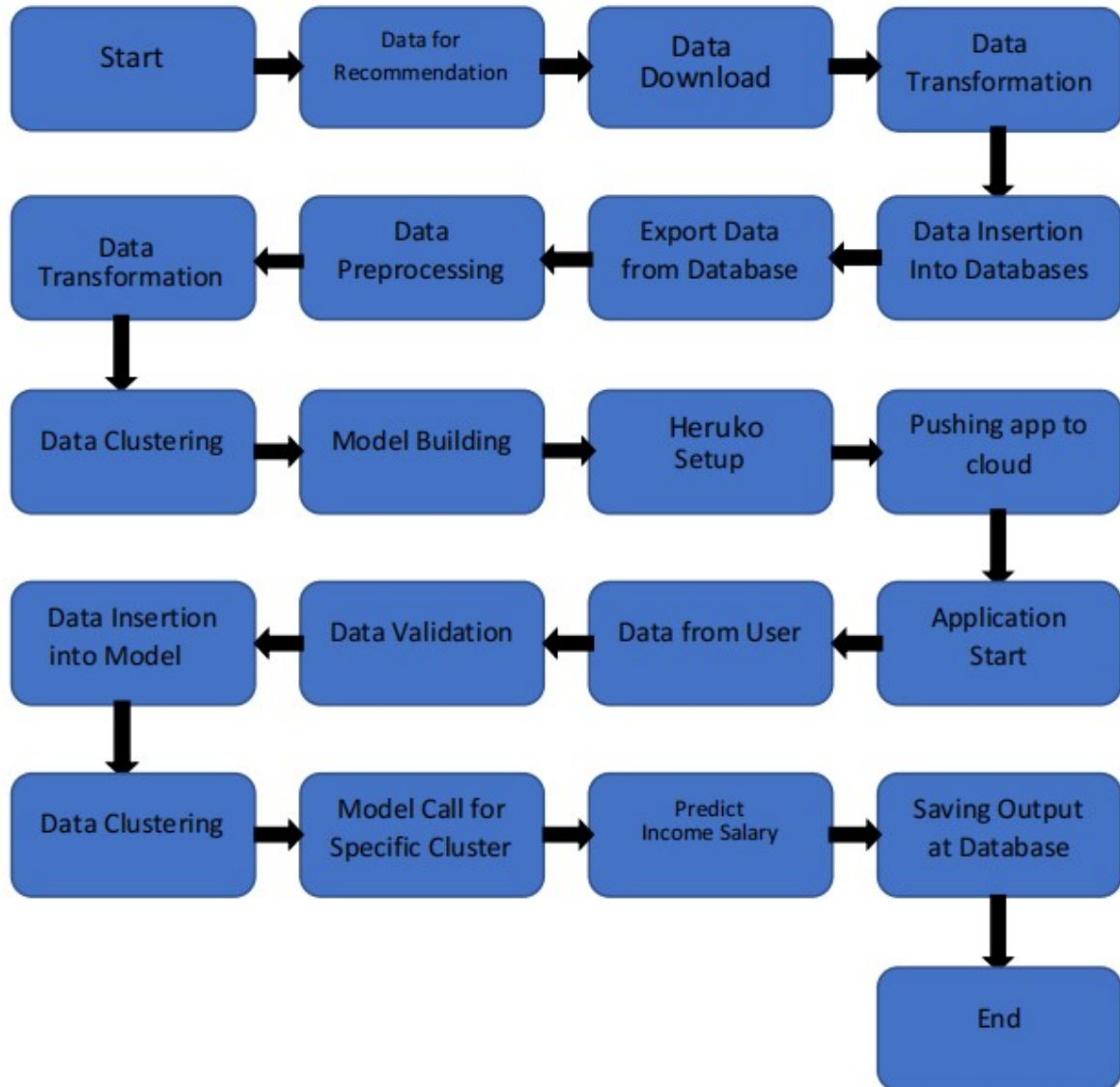
Just Fill the values and Predict the result.

2.7. Assumptions

The main objective of the Project is to Provide a method which can Predict wether a person is earning more than $50k or not. Machine Learning is used for Prediction and semi ML-Ops is used for completely automatize the Process.

3. Design Details

3.1. Process Flow

2. Architecture

3.2. Event log

The system should log event so that the user will know what process is running internally.

    1. The System identifies at what step logging required

    2. The System should be able to log each and every system flow.

    3. Developer can choose logging method. You can choose database logging/ File logging as well.

    4. System should not hang even after using so many loggings. Logging just because we can easily debug issues so logging is mandatory to do.

3.3. Error Handling

We have Applied Standard Processor to Prevent Error. Try and Exception is used to prevent error in system. Every error message is stored into Log file.

4. Performance

4.1. Re-usability

The Code is Written in the modular format so it is easy to understand and modified.We don't need to modified full code instead we can only touch segment of code.

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 the 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 until that function is finished.

4.4. Deployment

It can be deploy over any cloud platform.It is deploy over Heroku Continuous Integration and Deployment (CI/CD) pipelines using Git as a single source of truth.
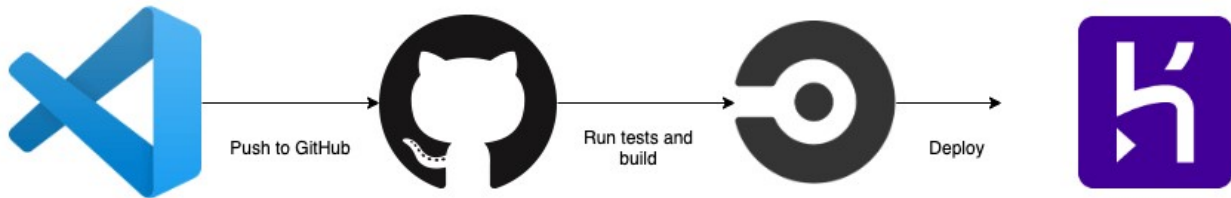
We used Docker to deploy our flask application. Docker provides the flexibility to ship the application easily to cloud platforms.

5. Conclusion

Adult Census Income Predictor Project will Predict wether a Person is earning more then $50K or not.

User need some basic information like Age, Race,Gender, Job Description, working Hour per week etc.