Architecture Design
Salary Prediction

Revision Number – 1.1

Last Date of Revision: 26 – 11 - 2022

Nitesh
Sharma

Document Version Control

Date	Version	Description	Author
26 – 11 - 2022	1.0	Abstract Introduction Architecture	Nitesh
26– 11- 2022	1.1	Architecture Design	Nitesh

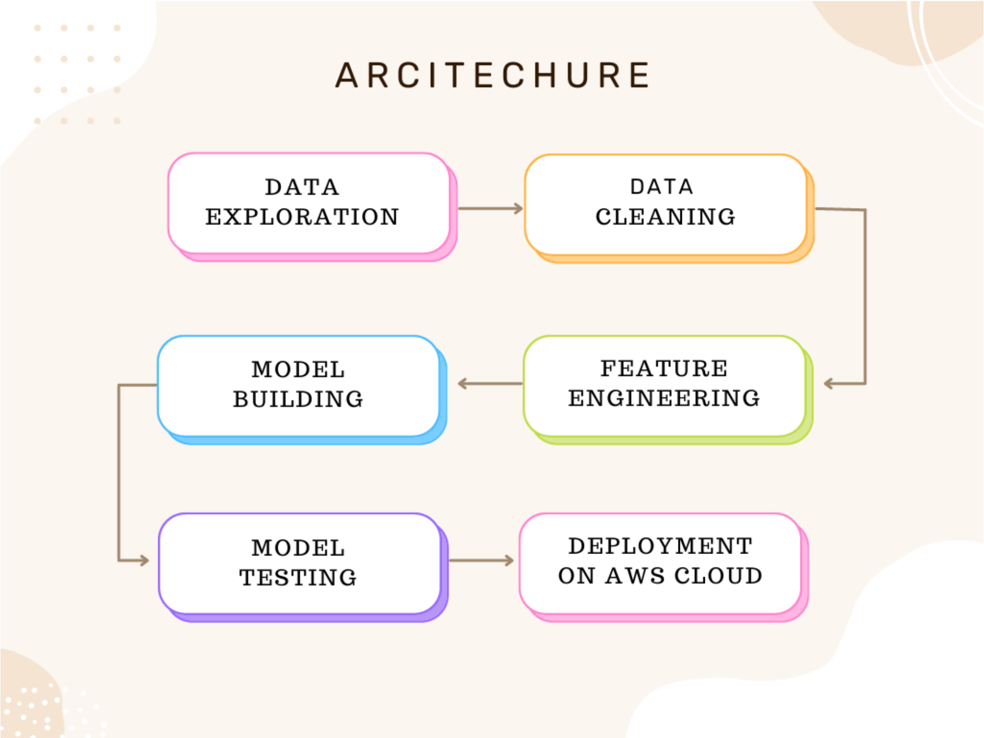
Contents

1. Introduction

1.1 Why this Architecture Design Document ?

The main objective of the Architecture design documentation is to provide the internal logic understanding of the Salary prediction code. The Architecture design documentatıon ıs desıgned ın such a way that the programmer can directly code after reading each module description in the documentation.

2. Architecture



3. Architecture Design

3.1 Data Collection

The data for these project is collected from the Kaggle Dataset, using kaggle API.

3.2 Data Description

Adult Census dataset is 30K+ dataset publicly available on kaggle. dataset contain 15 columns named age, workclass, fnlwgt, education, education-num, marital-status, occupation, relationship, race, sex, capital-gain, capital-loss, hours-per-week, country, salary.

3.3 Exploratory data analysis (EDA)

Explore the data to get insight information from the data. which helps us in training an model more efficiently and effectively. find out if there is any outlier or having missing values in any columns in the dataset, and handling it in the next steps.

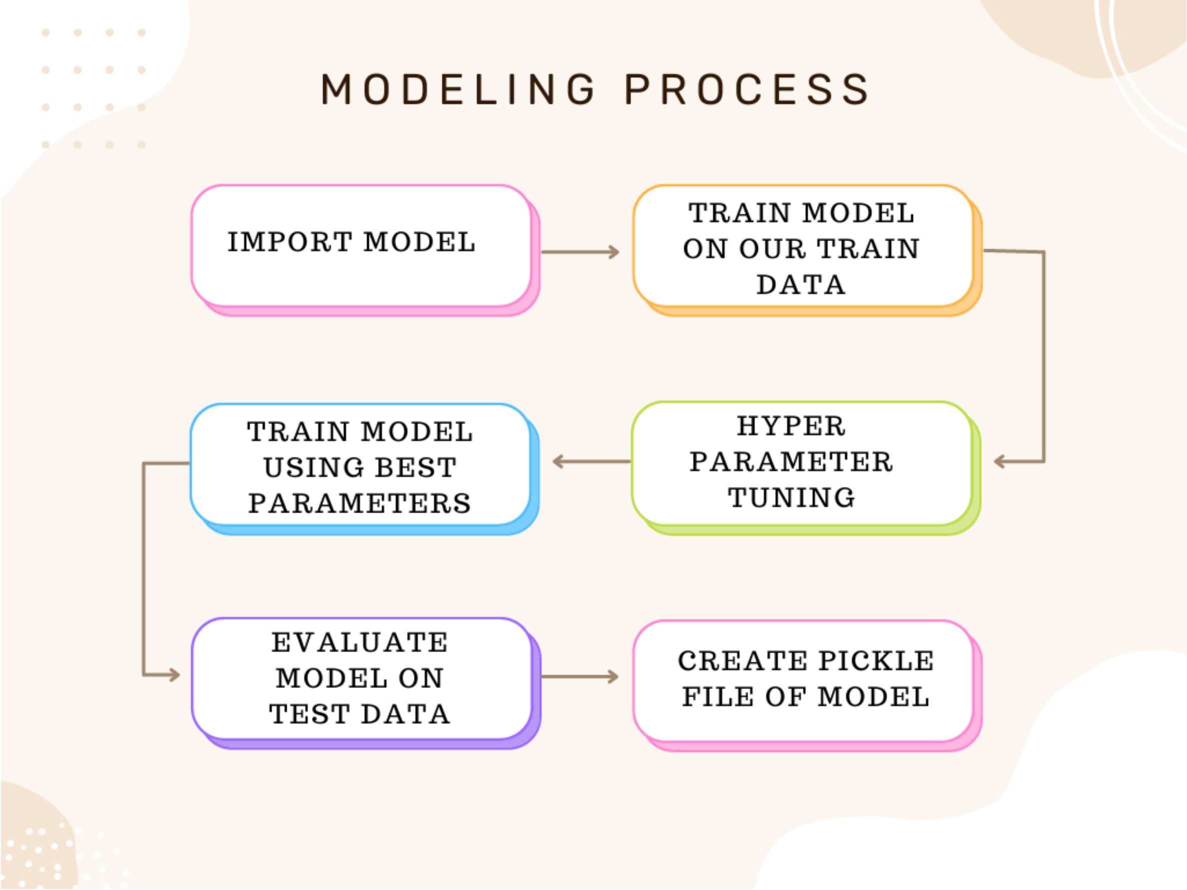
3.4 Data Preprocessing

- Checked for info of the Dataset, to verify the correct datatype of the Columns.
- Checked for Null values, because the null values can affect the accuracy of the model.
- Converted all the desired columns into Datetime format.
- Performed One - Hot encoding on the desired columns.
- Checking the distribution of the columns to interpret its importance.

Now, the info is prepared to train a Machine Learning Model.

3.5 Modeling Process

After preprocessing the data, we randomly spread our data into train and test data. using train data to train our model and test data to evaluate our trained model. we use logistic Regression on train data to predict the Salary is <=50K or >50k. create a pickle file of our train model with our data preprocessing steps. upload pickle file in aws S3 bucket.


MODELING PROCESS
IMPORT MODEL
TRAIN MODEL ON OUR TRAIN DATA
HYPER PARAMETER TUNING
TRAIN MODEL USING BEST PARAMETERS
EVALUATE MODEL ON TEST DATA
CREATE PICKLE FILE OF MODEL

3.6 UI Interface

- UI is created using streamlit framework. take input from user unsing streamlit objects like slider, container, expender etc.

3.7 Deployment

Using ci-cd pipeline, we config github action to create docker image of continuous changes and deploy that docker image to aws beanstalk. so that user can access the project from internet.