Low-Level Design

US VISA PREDICTION SYSTEM

|  |  |
| --- | --- |
| Written By | Jenna Raveenkumar |
| Document Version | 0.1 |
| Last Revised Date | 08– June -2024 |

Table of Contents

[1. Introduction 3](#_Toc175008626)

[1.1 Abstraction 3](#_Toc175008627)

[1.2 Purpose 3](#_Toc175008628)

[1.3 Scope 3](#_Toc175008629)

[2. Architecture 4](#_Toc175008630)

[2.1 Overall Workflow 4](#_Toc175008631)

[2.2 Internal Workflow 5](#_Toc175008632)

[3. Technical Specifications 6](#_Toc175008633)

[3.1 Data Description 6](#_Toc175008634)

[3.2 Data Validation 7](#_Toc175008635)

[3.3 Data Insertion into Db 7](#_Toc175008640)

[3.4 Data Pre-processing 7](#_Toc175008643)

[3.5 Data Transformation 7](#_Toc175008648)

[3.6 Model Training 7](#_Toc175008650)

[3.7 Model Pusher 8](#_Toc175008653)

[3.8 Prediction 8](#_Toc175008656)

[3.9 Deployment 8](#_Toc175008660)

# Introduction

## 1.1 Abstraction

Business communities in the United States are facing high demand for human resources, but one of the constant challenges is identifying and attracting the right talent, which is perhaps the most important element in remaining competitive. Companies in the United States look for hard-working, talented, and qualified individuals both locally as well as abroad.

The Immigration and Nationality Act (INA) of the US permits foreign workers to come to the United States to work on either a temporary or permanent basis. The act also protects US workers against adverse impacts on their wages or working conditions by ensuring US employers' compliance with statutory requirements when they hire foreign workers to fill workforce shortages. The immigration programs are administered by the Office of Foreign Labor Certification (OFLC).

OFLC processes job certification applications for employers seeking to bring foreign workers into the United States and grants certifications in those cases where employers can demonstrate that there are not sufficient US workers available to perform the work at wages that meet or exceed the wage paid for the occupation in the area of intended employment.

In FY 2016, the OFLC processed 775,979 employer applications for 1,699,957 positions for temporary and permanent labor certifications. This was a nine percent increase in the overall number of processed applications from the previous year. The process of reviewing every case is becoming a tedious task as the number of applicants is increasing every year.

The increasing number of applicants every year calls for a Machine Learning based solution that can help in shortlisting the candidates having higher chances of VISA approval. OFLC has hired your firm EasyVisa for data-driven solutions. You as a data scientist have to analyze the data provided and, with the help of a classification model:

* Facilitate the process of visa approvals.
* Recommend a suitable profile for the applicants for whom the visa should be certified or denied based on the drivers that significantly influence the case status.

## 1.2 Purpose

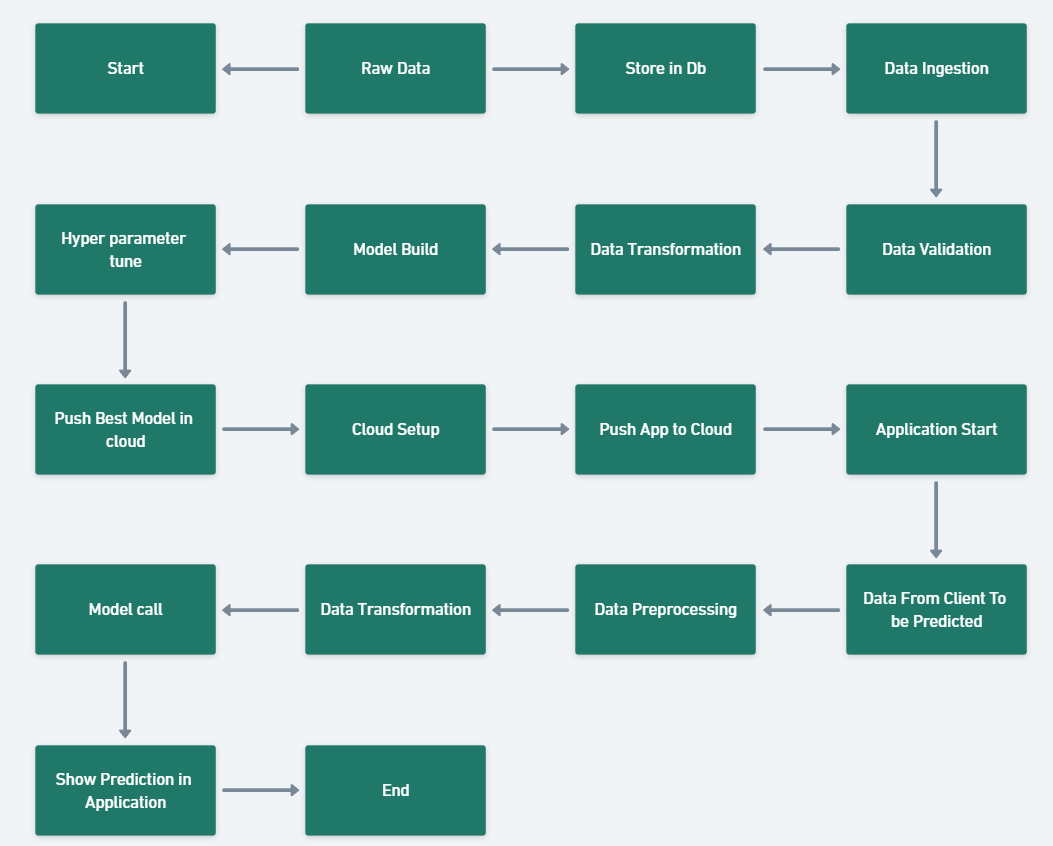
The main aim is to outline a high-level design for a Machine Learning-based system that predicts visa approval chances for applicants. This system will assist in expediting the visa certification process by identifying candidates with a higher probability of approval

## 1.3 Scope

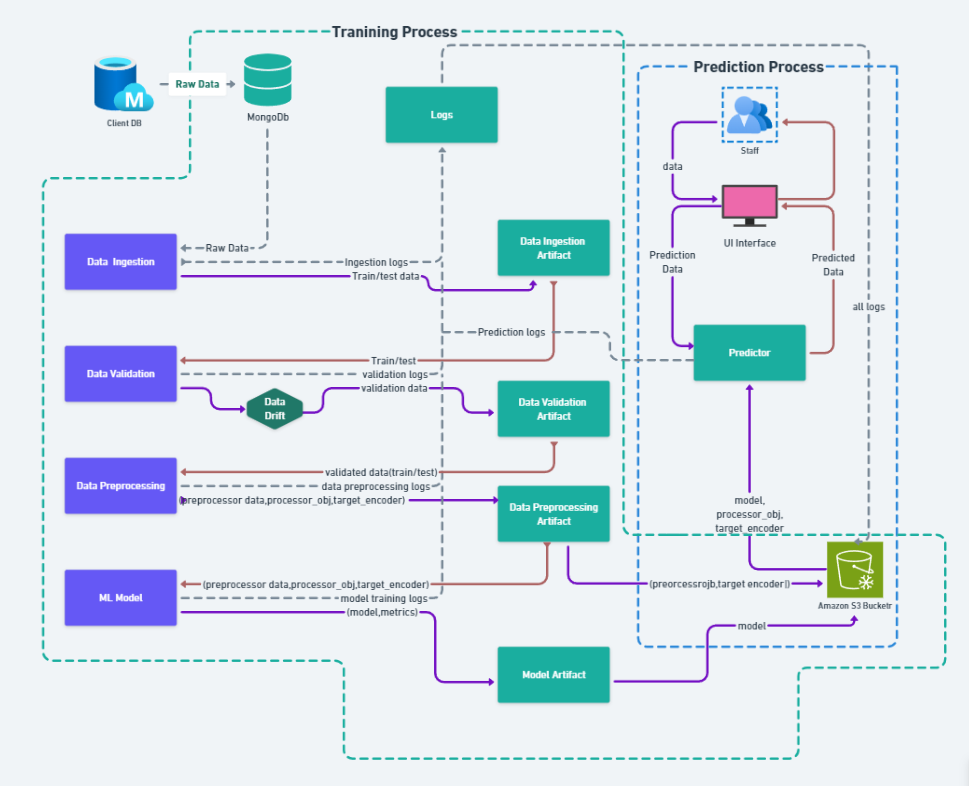
The system will analyze historical visa application data to identify the key factors that influence visa approval. It is going to use a classification model to predict the likelihood of visa approval for new applicants. The scope includes data pre-processing, model training, evaluation, and deployment as a web application for easy access.

# Architecture

## 2.1 Overall Workflow



## 2.2 Internal Workflow



# Technical Specifications

## 3.1 Data Description

The dataset contains 25480 records and the different attributes of the employee and the employer. The detailed data dictionary is given below

* case\_id: ID of each visa application
* continent: Information of continent the employee
* education\_of\_employee: Information of education of the employee
* has\_job\_experience: Does the employee has any job experience? Y= Yes; N = No
* requires\_job\_training: Does the employee require any job training? Y = Yes; N = No
* no\_of\_employees: Number of employees in the employer's company
* yr\_of\_estab: Year in which the employer's company was established
* region\_of\_employment: Information of foreign worker's intended region of employment in the US.
* prevailing\_wage: Average wage paid to similarly employed workers in a specific occupation in the area of intended employment. The purpose of the prevailing wage is to ensure that the foreign worker is not underpaid compared to other workers offering the same or similar service in the same area of employment.
* unit\_of\_wage: Unit of prevailing wage. Values include Hourly, Weekly, Monthly, and Yearly.
* full\_time\_position: Is the position of work full-time? Y = Full Time Position; N = Part Time Position
* case\_status: Flag indicating if the Visa was certified or denied

## 3.2 Data Validation

### Validate the Number of Columns: Validate the number of columns based on the DSA. Ensure that the same number of columns is received every time.

### Validate Column names: Validate each column name to ensure it matches the schema file for the proper process.

### Validate Column Type: Validate each column Type to ensure it aligns with the schema file for the correct process. This step is crucial for maintaining data integrity and accuracy in our system.

### Detect Data Drift: Using Evidently AI to detect data drift between training and test data is essential for proper training and ensuring correct data distribution.

## 3.3 Data Insertion into Db

### Insert Raw Data into DB: The entire raw data will be stored in a MongoDB database for training and testing purposes.

### Store Best Model in DB: After successful training, the best model and pre-processing objects will be stored in an AWS S3 bucket for better predictions.

## 3.4 Data Pre-processing

### Remove Unwanted Columns: Remove unwanted columns which is not useful training.

### Encoding The Data: Encode the categorical data and target feature to improve prediction accuracy.

### Power Transformation: Some features in the dataset are skewed, so we need to transform the features.

### Scaling Down the features: Scaling down all the features is useful for improving model performance.

## 3.5 Data Transformation

### Derive New Feature (Company Age): We derive new features based on the requirements, which are useful throughout the prediction process. And build a custom functional Transformational process for the Custom pipeline.

## 3.6 Model Training

Evaluate the AUC-ROC and F1-Score of all the models. Then find the top three highest-scored models. Hyper-tune these three models and find the top model among them. Finally, push this model to the S3 bucket.

### Getting the model Object: Getting the stored model object for training.

### Initiate Training Process: Start the training process using transformation data and the model object.

## 3.7 Model Pusher

### Model Pusher Process: Upload the best model into cloud storage for prediction.

### Initiate Model Pusher Process: Start the model pusher process to store the best models and pre-processing objects in the S3 Bucket.

## 3.8 Prediction

The prediction process starts from the pre-processing pipeline. After completing the training pre-processing, we will store the pre-processing object and the target encoded object in the S3 Bucket and retrieve the best model from the S3 Bucket. Using these three objects, we will make the prediction.

### Data into Dataframe: Convert the prediction data into Dataframe for prediction.

### Data pre-processing: The Converted Dataframe passed through the pre-processing pipeline.

### Prediction Process: Please initiate the prediction using the best model.

## 3.9 Deployment

We deploy the application in the AWS Cloud using AWS EC2 and a CI/CD pipeline.