A PROJECT ON

"HIB VISA ACCEPTANCE PREDICTION"

SUBMITTED IN
PARTIAL FULFILLMENT OF THE REQUIREMENT
FOR THE COURSE OF
DIPLOMA IN BIG DATA ANALYSIS

SUNBEAM INSTITUTE OF INFORMATION TECHNOLOGY
'Sunbeam IT Park', Phase 2 of
Rajiv Gandhi Infotech Park,
Hinjewadi, Pune, 411057,
MH-INDIA

SUBMITTED BY:

Shreeya Chavan (75845)

Satyam Gawade (75527)

UNDER THE GUIDENCE OF:

Mrs. Manisha H
Faculty Member
Sunbeam Institute of Information Technology, PUNE.

CERTIFICATE

This is to certify that the project work under the title ‘H1B Visa Acceptance Prediction’ is done by Shreeya Chavan & Satyam Gawade in partial fulfillment of the requirement for award of Diploma in Big Data Analysis Course.

Mrs. Manisha Hinge
Project Guide

Mrs. Pradnya Dindorkar
Course Co-ordinator

Date:

ACKNOWLEDGEMENT

A project usually falls short of its expectation unless aided and guided by the right persons at the right time. We avail this opportunity to express our deep sense of gratitude towards Mr. Nitin Kudale (Center Coordinator, SIIT-Pune) and Mrs. Pradnya Dindorkar (Course Coordinator, SIIT-Pune) and Project Guide Mrs. Manisha Hinge.

We are deeply indebted and grateful to them for their guidance, encouragement and deep concern for our project. Without their critical evaluation and suggestions at every stage of the project, this project could never have reached its present form.

Last but not the least we thank the entire faculty and the staff members of Sunbeam Institute of Information Technology, Pune for their support.

Shreeya Chavan
DBDA March 2023 Batch,SIIT,
Pune

Satyam Gawade
DBDA March 2023 Batch,
SIIT, Pune

TABLE OF CONTENTS

1. Introduction

1.1 Introduction And Objectives:

H1B visa is a non-immigrant temporary work visa that lets people with special skills to work in USA. To apply for this visa, the applicant must have a job offer from an employer in the USA. The employer then files H1B visa petition with the US Immigration Service (USCIS) for the employee with relevant attestations, including attestations about wages and working conditions. After the approval from USCIS, this petition allows employee to obtain a visa stamp and work in the U.S. for that employer.

1.2 Why this problem needs to be solved?

Our goal is to develop an algorithm that can predict the petition outcome based on Application information. This problem needs to be solved to reduce the lead time as the Employer can understand that the employee with certain attributes can get the Visa or not. Since, we are planning to forecast the visa petition status, the outcome of the analysis is divided into two classes (Denied or Certified). Considering it as a binary classification problem, we have applied multiple model on this dataset. The report initially represents the analysis of available Dataset by plotting relationship between various attributes.
In the next sections, the following have been described in brief:
Data preprocessing
Feature selection
Implement of different Classification Models on the available data
Model Evaluation
Model Selection
Deployment

1.3 Dataset Information.

H1B_RECORDS.csv:

It has 10 columns.

CASE_STATUS:
This feature defines the final decision of application. It is divided into seven classes, CERTIFIED, CERTIFIED-WITHDRAWN, WITHDRAWN, DENIED, PENDING QUALITY AND COMPLIANCE REVIEW – UNASSIGNED, REJECTED, INVALIDATED.

EMPLOYER_NAME:
Represents the company name who files the petition to recruit employee.

SOC_NAME:
States occupational name that is classified by the Standard Occupational Classification (SOC) System.

JOB_TITLE:
Title of the job.

FULL_TIME_POSITION:
Represents job status where, Y= Full time and N=Part time.

PREVAILING_WAGE:
Defines the average wage paid to similarly employed workers in the requested occupation in the area of intended employment

YEAR:
Year in which the visa petition filed

WORKSITE:
The employee's intended area of employment.

LON:
Longitude of the worksite.

LAT:
Latitude of the Worksite

2. Problem Definition and Algorithm:

2.1 Problem Definition

The problem is quite straightforward. Data from H1B Visa CSV from the US is given, and it is up to us to predict the acceptance. The data is unclean, we need to first preprocess the data, do Feature Engineering, find the important columns and fit an algorithm for the final Data and Predict whether one will get the H1B Visa or not and match with the Original Outcome. This of course is highly unlikely, but we must try to get as close as possible.

2.2 Algorithm

Logistic Regression Classifier

Logistic Regression is a statistical method used for binary classification problems, where the goal is to predict one of two possible outcomes based on input features. Despite its name, logistic regression is used for classification, not regression tasks. It's a fundamental algorithm in machine learning and is particularly well-suited for scenarios where the relationship between the features and the outcome is likely to be linear.

The primary objective of logistic regression is to estimate the probability that a given input belongs to a particular class. The sigmoid function maps any input value to a value between 0 and 1, which can be interpreted as the probability of belonging to the positive class.

Decision Tree Classifier

A Decision Tree Classifier is a machine learning algorithm used for both classification and regression tasks. It's a versatile and intuitive algorithm that creates a tree-like structure to make decisions based on input features. Each internal node of the tree represents a decision based on a specific feature, and each leaf node represents a class label or a numeric value.

The basic idea behind a decision tree is to divide the feature space into regions, where each region corresponds to a specific class label. To build the tree, the algorithm recursively selects the best feature to split the data based on certain criteria (such as information gain), then continues to split the data until a stopping criterion is met, like reaching a maximum depth or having a minimum number of samples in a leaf node.

Random Forest Classifier

A Random Forest Classifier is an ensemble learning method that combines multiple decision trees to improve predictive accuracy and control overfitting. It's a versatile and powerful algorithm that is widely used for classification tasks. Random Forests build a "forest" of decision trees and aggregate their predictions to make a final classification decision.

Gradient Boosting Classifier

Gradient Boosting Classifier is another powerful ensemble learning method used for classification tasks. Similar to Random Forests, it combines multiple weak learners (typically decision trees) to create a strong predictive model. However, the process by which Gradient Boosting builds its ensemble is different, focusing on iteratively correcting the errors of previous models.

Adaptive Boosting Classifier

Adaptive Boosting, often referred to as AdaBoost, is a popular machine learning algorithm that belongs to the ensemble learning category. Ensemble learning involves combining the predictions of multiple individual models to create a stronger, more accurate model.

AdaBoost specifically focuses on classification tasks and works by training a series of weak learners (usually decision trees, often referred to as "stumps") sequentially. A weak learner is a model that performs slightly better than random guessing. In each iteration of AdaBoost, the algorithm gives more weight to the misclassified instances from the previous iteration, thereby forcing the subsequent weak learner to focus on the instances that were previously difficult to classify correctly.

Extreme Gradient Boosting Classifier

XGBoost is another powerful ensemble learning algorithm that is particularly effective for classification and regression tasks. It's an improved version of traditional gradient boosting that has gained popularity due to its high performance and scalability. XGBoost is known for its ability to handle large datasets, feature importance visualization, and flexibility in handling various types of data.

Deep Learning Neural Network Classification

An Artificial Neural Network (ANN) classifier is a type of machine learning model used for classification tasks. ANNs are inspired by the structure and functioning of the human brain, consisting of interconnected nodes (neurons) organized into layers. Each node in a layer is connected to nodes in the previous and/or subsequent layers, and these connections have associated weights.

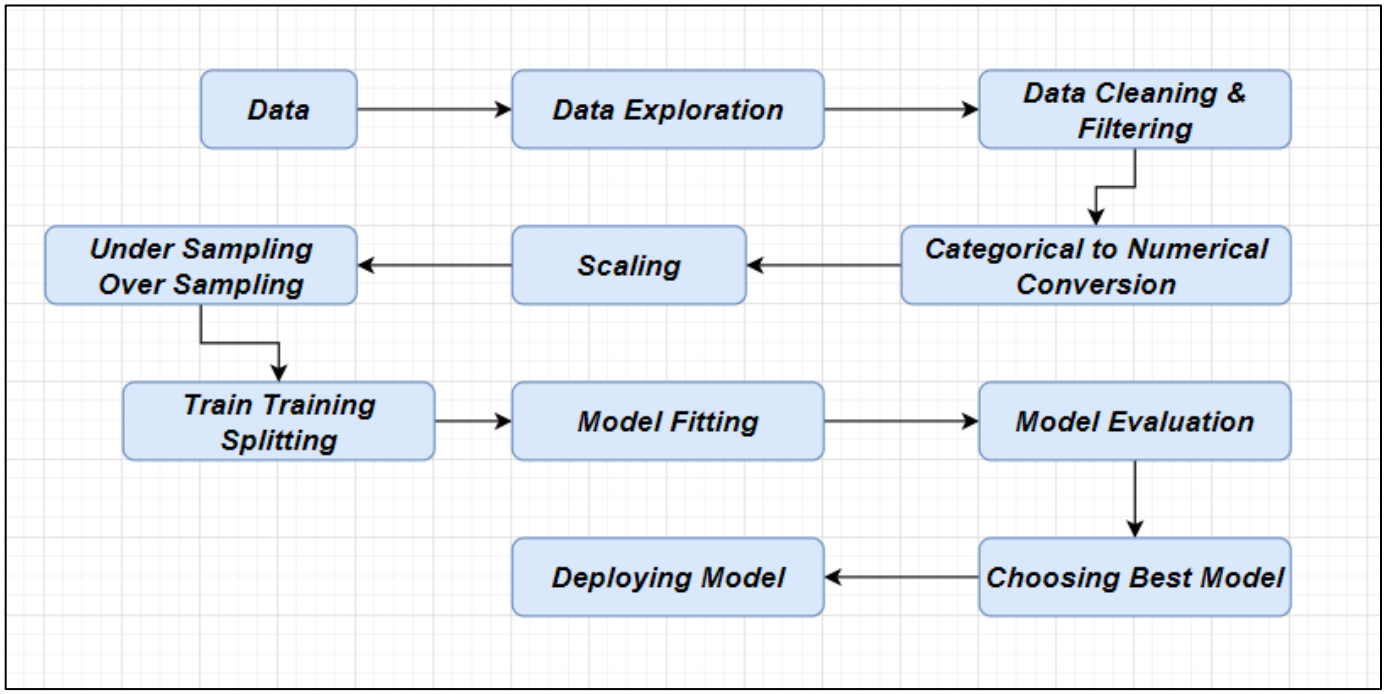
The basic idea behind an ANN classifier is to learn patterns and relationships in the input data by adjusting the weights of the connections between nodes. This learning process involves presenting the network with a dataset where each example is associated with a label or class. The network then adjusts its weights iteratively through a process called training, aiming to minimize a certain objective function, often a measure of prediction error, like cross-entropy loss.

The architecture of an ANN classifier typically includes an input layer that receives the feature values, one or more hidden layers that process and transform the features, and an output layer that produces the classification results. The number of hidden layers and the number of nodes in each layer are parameters that can be adjusted based on the complexity of the task and the characteristics of the data.

3. Experimental Evaluation:

3.1 Methodology:

The objective of this project is to predict the approval off Visas for working in US. The Methodology is explained with help of the following Flow Process



3.2 Exploratory Data Analysis

The Percentage of Case Status whether Certified or Denied can be visualized by the Pie Chart which Shows that 30% comes in Denied Case while 70% comes in Certified Case

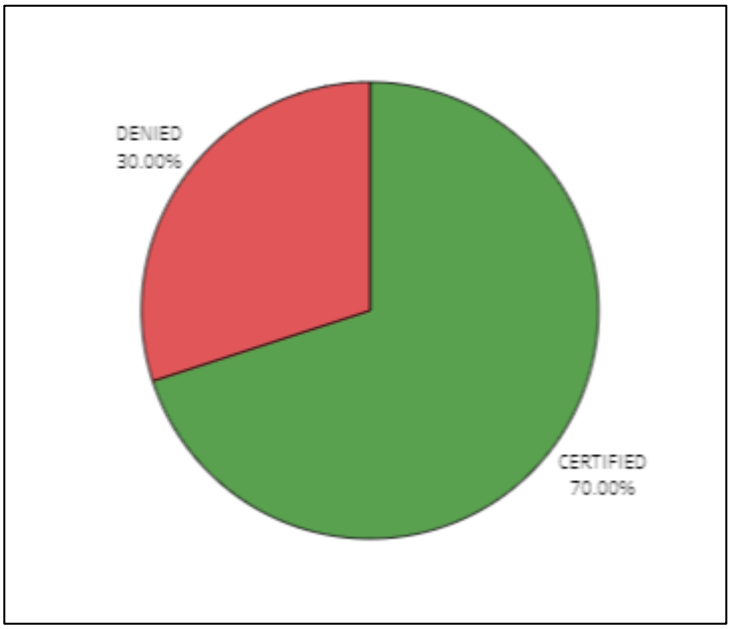


Fig 1: Pie- chart Showing Case Status Percentage

The Employer wise count of certified vs Denied can be visualized by the Vertical bar plot, and from this we can see various companies with their count of certified vs denied.

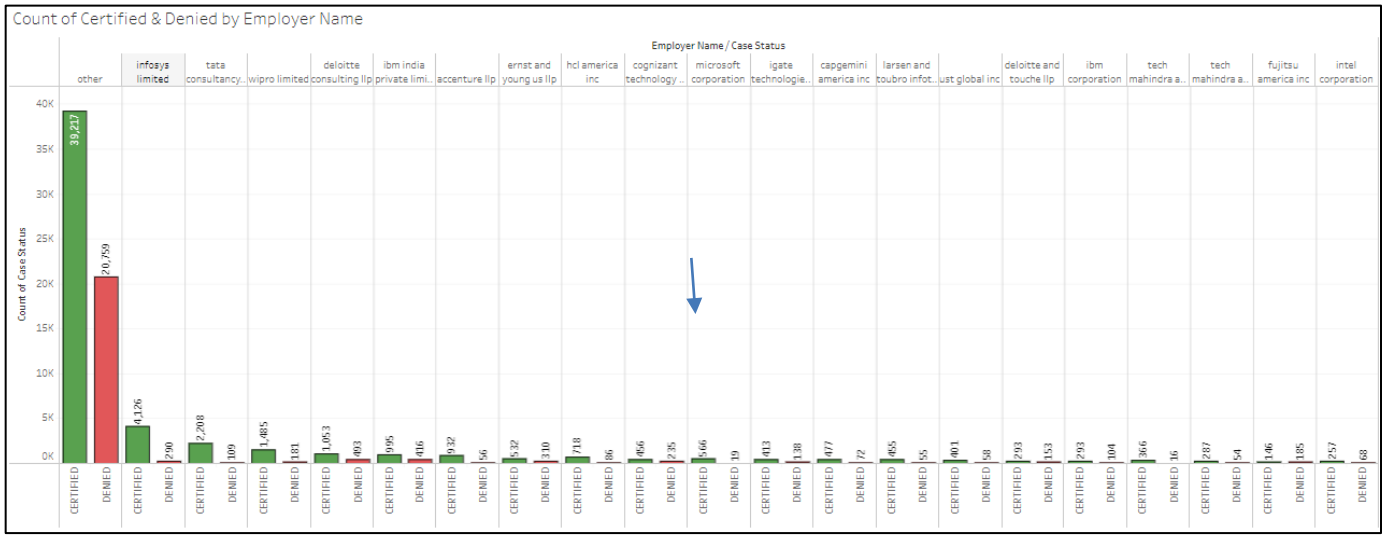


Fig 2: Count of Certified and Denied by Employer Name

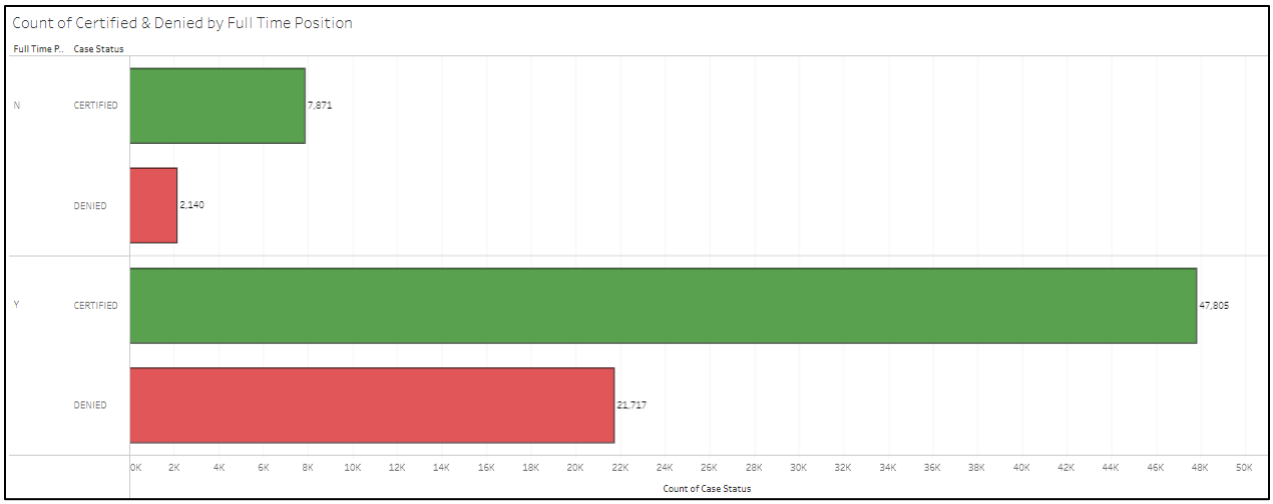


Fig 3: Count of Certified and Denied by Full Time Position

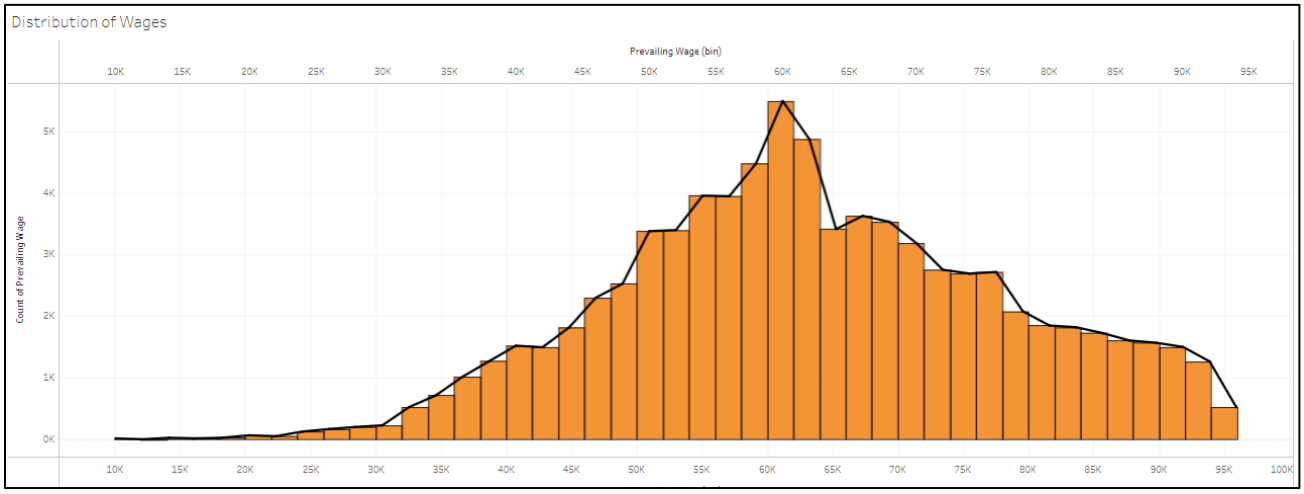


Fig 4: Distribution of Wages

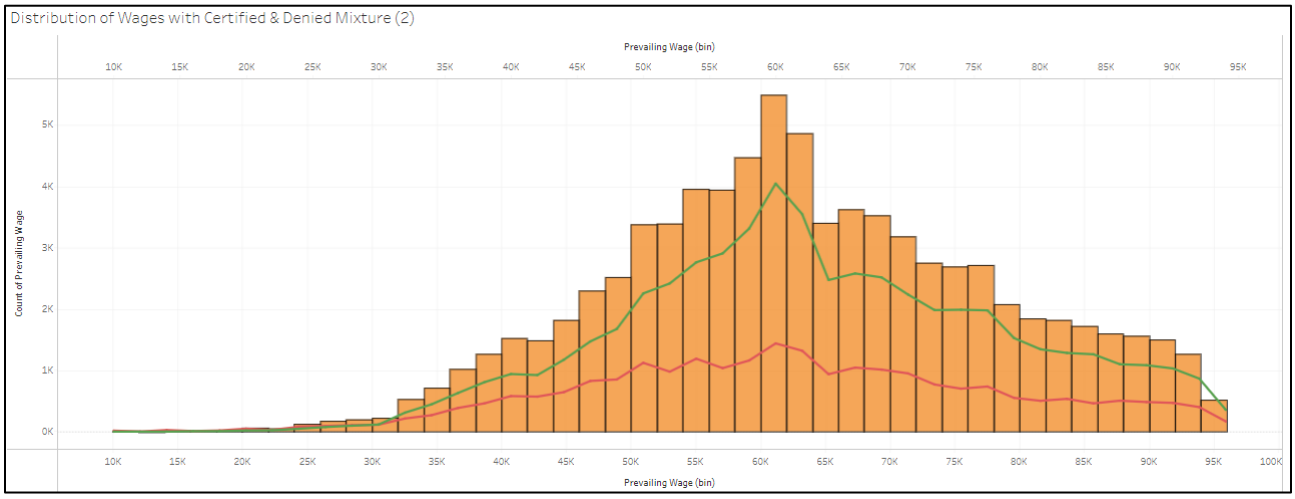


Fig 5: Distribution of Wages with Certified & Denied Mixture

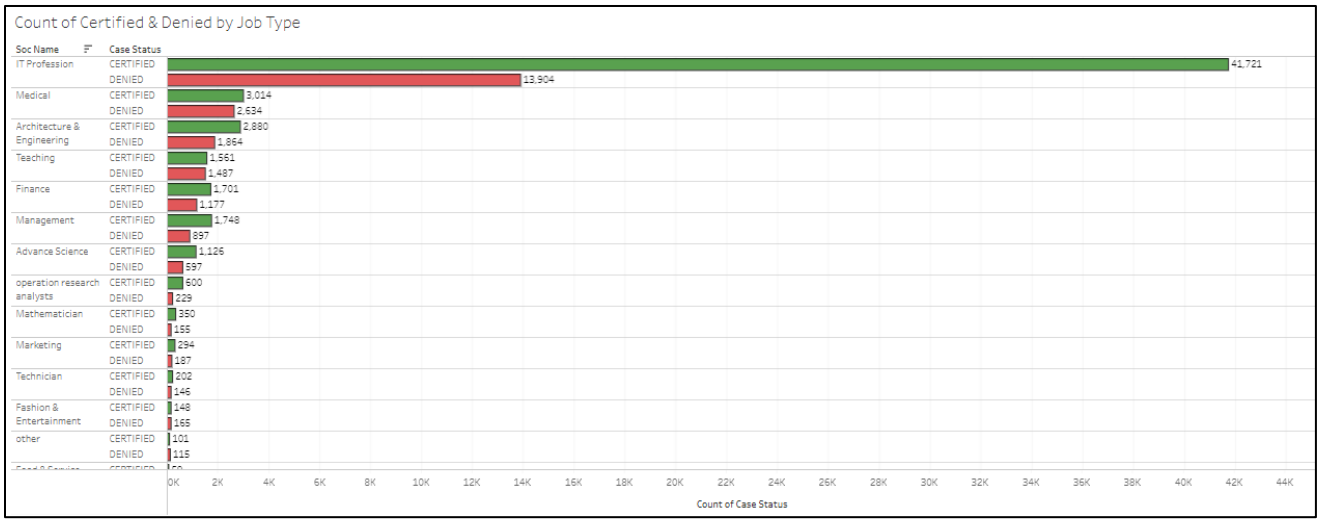


Fig 6: Count of Certified and Denied by Job Type

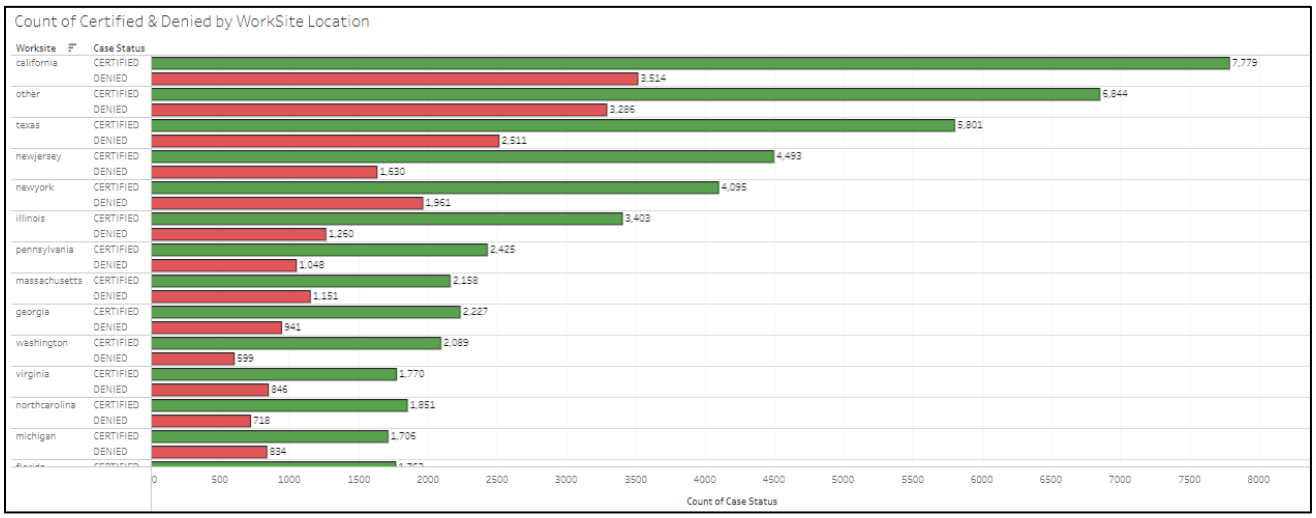


Fig 7: Count of Certified and Denied by Location

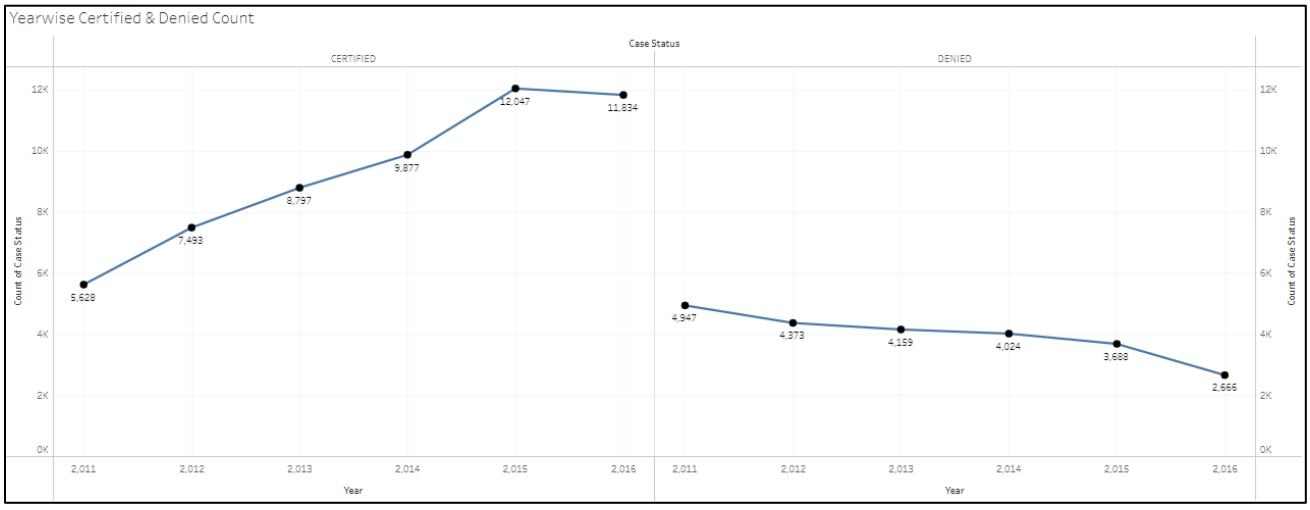


Fig 8: Year wise Certified and Denied Cases

From the visualization we can come to conclusions about the Features present in our Dataset.

Our Dataset is Imbalanced having 30% of Denied Cases and 70% of Certified Cases, hence solution to that we need to perform under sampling or over sampling and then check the results

From the plotting in Fig 2, 3 it can be derived that the Employer name and full time position plays significant role in Certifying or Denying the Visa for personal, where most of the individual who are employed as full time Position yes have more chance to get the approval similar with the Employer who are IT firms have more persons applying for the visas and getting approval

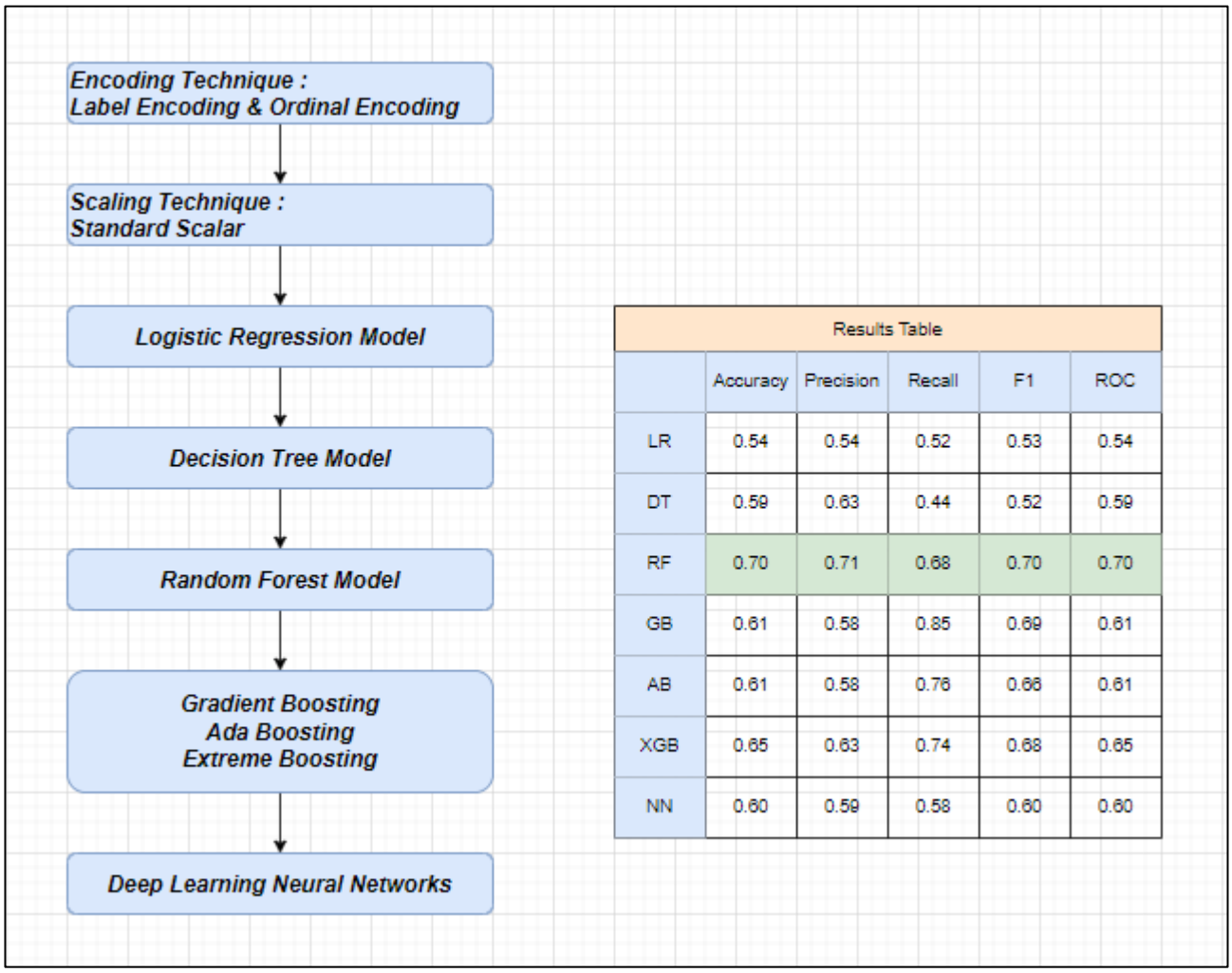
From the plotting 4, 5 we can see the distribution of the wages of the persons in our data set as well as distribution of wages for persons in certified and denied category

From plotting 6, 7 we can observe that location is an important feature to consider as counts can be visualized for the same, as well as Job type plays important role as most personal from IT Fields are applying for visas for most of the time

From Figure 8, we observe the count of people applied for visas in each year and the count of certified and denied for the same.

4. Results and discussion:

V01 Models



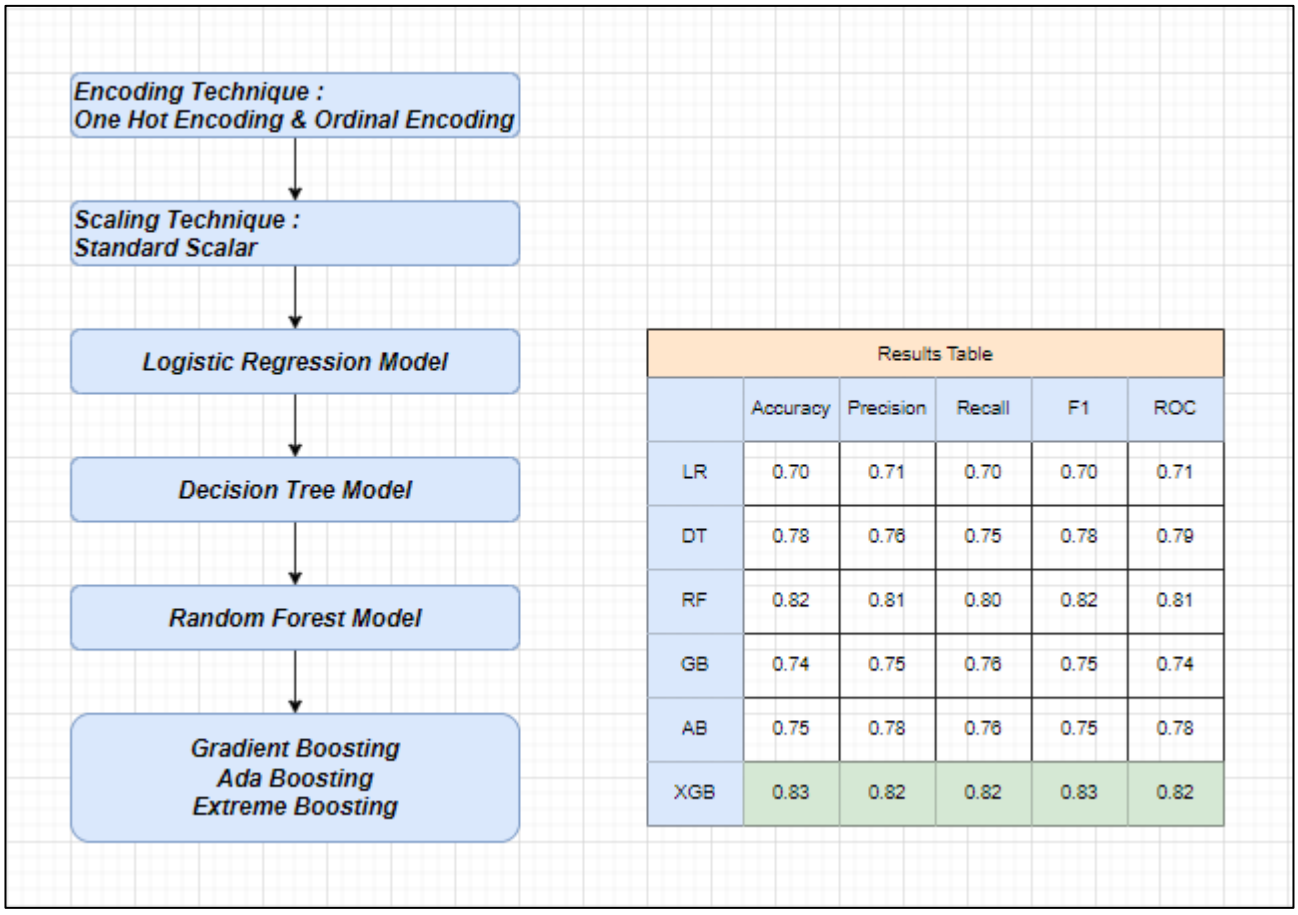
Results Table

	Accuracy	Precision	Recall	F1	ROC
LR	0.54	0.54	0.52	0.53	0.54
DT	0.59	0.63	0.44	0.52	0.59
RF	0.70	0.71	0.68	0.70	0.70
GB	0.61	0.58	0.85	0.69	0.61
AB	0.61	0.58	0.76	0.66	0.61
XGB	0.65	0.63	0.74	0.68	0.65
NN	0.60	0.59	0.58	0.60	0.60

Fig 1: Version 01 Techniques & Models evaluation

From the Version 01 Data preprocessing and Model Making the Max F1 Score we could obtain was 70% with Random Forest Hyper Parameter Tuned Model, but with later improvements we could achieve more accuracy in Version 02 Models

V02 Models



Results Table

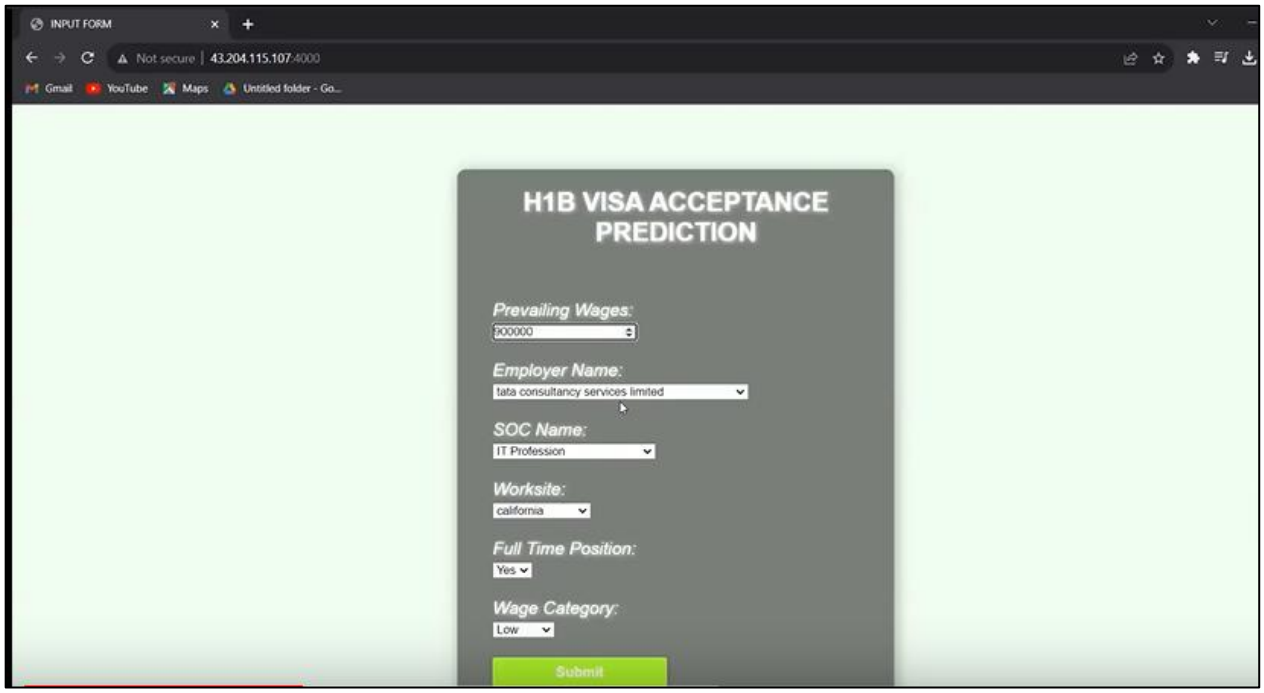
	Accuracy	Precision	Recall	F1	ROC
LR	0.70	0.71	0.70	0.70	0.71
DT	0.78	0.78	0.75	0.78	0.79
RF	0.82	0.81	0.80	0.82	0.81
GB	0.74	0.75	0.76	0.75	0.74
AB	0.75	0.78	0.76	0.75	0.78
XGB	0.83	0.82	0.82	0.83	0.82

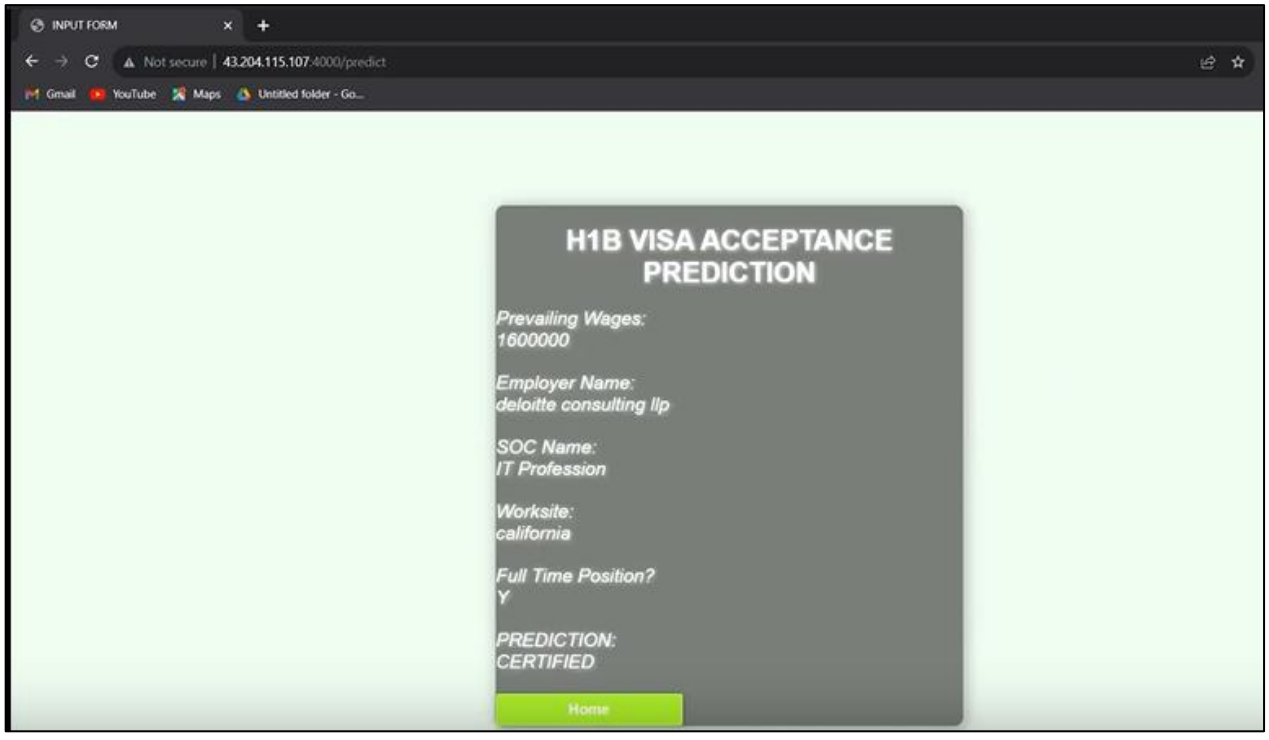
Fig 2: Version 01 Techniques & Models evaluation

From the Version 02 Data preprocessing and Model Making the Max F1 Score we could obtain was 83% with Extreme Gradient Boosting Model, we have used this model for Deploying

5. GUI:

GUI is made using Flask framework. **Flask** is a micro web framework written in Python. It is classified as a micro framework because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions. However, Flask supports extensions that can add application features as if they were implemented in Flask itself.

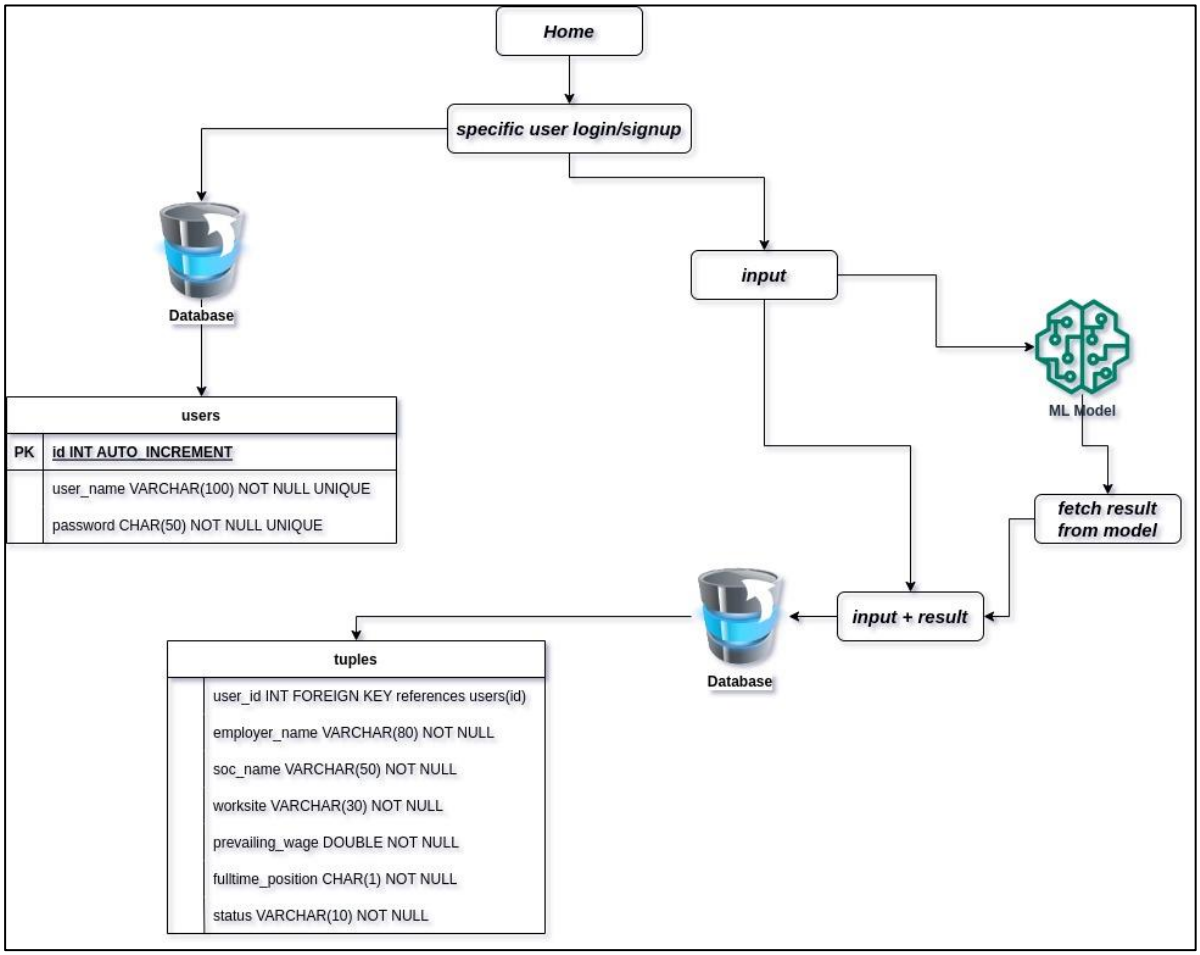




6. Future work And Conclusion

6.1Future Work:

This Model can be used by various Employers or Employees or the H1B Visa People for their use and this Concept can be further taken ahead and modified as per Individual Requirements. The Diagram below shows the proper use with Database Connectivity



6.2 Conclusion:

The conclusion of an H1B visa prediction machine learning project would typically summarize the outcomes, findings, and implications of the project. Here's a conclusion for such a project:

In this machine learning project, we aimed to predict the likelihood of H1B visa approval based on a variety of features related to the applicant's profile and the job position. Through extensive data preprocessing, feature engineering, and model selection, we were able to develop a predictive model that achieved promising results.

Our analysis revealed that certain features, such as the applicant's education level, job location, and employer's track record, had significant impacts on the H1B visa approval outcome. The chosen machine learning algorithm, after hyper parameter tuning, demonstrated good performance in terms of accuracy, precision, recall, and F1-score.

However, it's important to note that the prediction model is not without its limitations. The model's performance might vary when applied to new, unseen data due to changes in immigration policies, economic conditions, and other factors that might influence visa approval trends. Additionally, the model might not capture the nuances of certain individual cases that involve unique circumstances not well-represented in the training data.

From a practical standpoint, the insights gained from this project can be valuable to both prospective H1B visa applicants and immigration authorities. Applicants can utilize the model's predictions to make more informed decisions regarding their visa application process. Immigration authorities can use the model's findings to identify potential areas of improvement in the visa approval process and potentially streamline decision-making.

As with any predictive model, ongoing monitoring and updates are necessary to ensure its relevance and accuracy. Future iterations of this project could explore incorporating more recent data, considering external factors like policy changes, and potentially integrating other advanced techniques such as natural language processing for analyzing application statements.

In conclusion, this project highlights the potential of machine learning in predicting H1B visa outcomes. While the model's predictions can offer valuable insights, they should be considered as part of a larger decision-making process that takes into account individual circumstances and the broader immigration landscape.