

Project Initialization and Planning Phase

Date	15 March 2024
Team ID	LTVIP2024TMID25012
Project Name	Predictive Modeling for H1b Visa Approval Using Machine Learning
Maximum Marks	3 Marks

Define Problem Statements (Customer Problem Statement Template):

The process of predicting H1B visa approval outcomes is challenging due to the complexity of factors involved, such as job title, wage, employer details, and legal requirements. Currently, HR professionals and immigration attorneys lack an accurate, data-driven method to anticipate approval decisions, leading to inefficiencies in resource planning and decision-making. This project aims to develop a predictive model using machine learning to improve the accuracy of H1B visa approval predictions, helping organizations streamline their hiring processes and reduce uncertainty.

I am:	I'm trying to:	But:	Because:	Which makes me feel:
An HR professional or immigration attorney working on behalf of companies to process H1B visa applications for employees.	Predict the approval or denial of H1B visa applications to make informed decisions about recruitment and resource allocation.	The approval process is complex, relies on multiple factors that are not easily predictable, and requires a significant amount of time and manual effort.	There is no clear, accessible method to anticipate the approval status based on historical data, job roles, wages, and employer details.	Confused and uncertain, as it causes delays, impacts planning, and could result in losing valuable talent if applications are unexpectedly denied.

PS-1	<p>I am:</p> <p>An HR professional or immigration attorney managing H1B visa applications.</p>	<p>I'm trying to:</p> <p>Efficiently predict the approval or denial of H1B visa applications to streamline recruitment and reduce delays.</p>	<p>But:</p> <p>The approval process is unpredictable and influenced by numerous complex factors, making it difficult to anticipate outcomes.</p>	<p>Because:</p> <p>There is no reliable tool that leverages historical data and machine learning to provide accurate predictions based on key visa-related attributes.</p>	<p>Which makes me feel:</p> <p>Frustrated and uncertain, leading to delays in resource planning and the potential loss of critical hires.</p>
------	---	--	---	---	--

InitialProjectPlanningTemplate

Date	15March 2024
TeamID	LTVIP2024TMID25012
ProjectName	PredictiveModeling forH1bVisaApproval UsingMachineLearning.
MaximumMarks	4Marks

ProductBacklog,SprintSchedule, andEstimation(4Marks)

Sprint	Functional Requirement (Epic)	UserStory Number	UserStory /Task	Story Points	Priority	TeamMembers	Sprint Start Date	SprintEnd Date (Planned)
Sprint-1	Data Collection &Preprocessing	USN-1	Asadata scientist,Icanloadand preprocess the H1B dataset, handlingmissingvaluesanddata types..	3	High	1. Velamakanni Abhigna 2. Chinthareddy Lalitha 3.SirasaniTeja Venkata Sai Charan 4.Yedupati Manoj	24-07-24	25-07-24

Sprint-1		USN-2	Asadata scientist, I can split the data into training and test sets.	2	High	1. Velamakanni Abhigna 2. Chinthareddy Lalitha 3. Sirasani Teja Venkata Sai Charan 4. Yedupati Manoj	24-07-24	25-07-24
Sprint-2	Model Development	USN-3	Asadata scientist, I can develop a Random Forest model to predict visa approval status.	5	High	1. Velamakanni Abhigna 2. Chinthareddy Lalitha 3. Sirasani Teja Venkata Sai Charan 4. Yedupati Manoj	12-08-24	14-08-24

Sprint-1		USN-4	As a data scientist, I can optimize the model using hyperparameter tuning to improve accuracy.	4	Medium	1. Velamakanni Abhigna 2. Chinthareddy Lalitha 3. Sirasani Teja Venkata Sai Charan 4. Yedupati Manoj	12-08-24	14-08-24
Sprint-1	Model Evaluation	USN-5	As a data scientist, I can evaluate the model's performance using accuracy, precision, recall, and F1-score metrics.	3	High	1. Velamakanni Abhigna 2. Chinthareddy Lalitha 3. Sirasani Teja Venkata Sai Charan 4. Yedupati Manoj	29-09-24	02-09-24

Project Initialization and Planning Phase

Date	15 March 2024
Team ID	LTVIP2024TMID25012
Project Title	Predictive Modeling for H1b Visa Approval Using Machine Learning.
Maximum Marks	3 Marks

Project Proposal (Proposed Solution) template

This project proposal outlines a solution to address a specific problem. With a clear objective, defined scope, and a concise problem statement, the proposed solution details the approach, key features, and resource requirements, including hardware, software, and personnel.

Project Overview	
Objective	The primary objective of this project is to develop a machine learning model to predict the approval or denial of H1B visa applications. The model will help HR professionals and immigration attorneys streamline decision-making processes and improve accuracy in anticipating visa outcomes.
Scope	The project involves data collection, preprocessing, model development, evaluation, and deployment of a predictive model for H1B visa approval. The scope includes the creation of a web interface where users can input visa-related information and receive predictions in real time.
Problem Statement	
Description	The current H1B visa approval process is complex and lacks transparency, with multiple factors affecting approval decisions. Without an accurate method to predict visa outcomes, companies face inefficiencies in resource allocation and recruitment planning.
Impact	Solving this problem will allow companies to make data-driven decisions regarding hiring and visa application processing, reducing uncertainty and optimizing resources.

Proposed Solution	
Approach	The proposed solution is a machine learning-based predictive model. The model will use a dataset of past H1B visa applications and train a RandomForestClassifier to predict the approval or denial status of future applications. Key steps include data preprocessing, feature selection, model training, evaluation, and deployment.
Key Features	<ul style="list-style-type: none"> • Predicts visa approval or denial based on key application attributes (e.g., job role, wage, full-time position). • Uses Random Forest for classification. • Integrated with a Flask-based web application for user interaction. • Provides real-time predictions once deployed to a cloud platform.

Resource Requirements

Resource Type	Description	Specification/Allocation
Hardware		
Computing Resources	CPU/GPU specifications, number of cores	e.g., 2 x NVIDIA V100 GPUs
Memory	RAM specifications	e.g., 8 GB
Storage	Disk space for data, models, and logs	e.g., 1 TB SSD
Software		
Frameworks	Python frameworks	e.g., Flask
Libraries	Additional libraries	e.g., scikit-learn, pandas, numpy
Development Environment	IDE, version control	e.g., Jupyter Notebook, Git
Data		
Data	Source, size, format	Kaggle dataset, CSV format, 3002458 rows

Data Collection and Preprocessing Phase

Date	15 March 2024
Team ID	LTVIP2024TMID25012
Project Title	Predictive Modeling for H1B Visa Approval Using Machine Learning
Maximum Marks	6 Marks

Data Exploration and Preprocessing Template

Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

Section	Description
---------	-------------

Data Overview	<p>The dataset consists of 3,002,458 entries and 11 columns, including key attributes such as CASE_STATUS, EMPLOYER_NAME, SOC_NAME, JOB_TITLE, FULL_TIME_POSITION, and PREVAILING_WAGE.</p> <p>Basic Statistics:</p> <ul style="list-style-type: none"> □ Total Rows: 3,002,458 □ Total Columns: 11 □ Unique Values in CASE_STATUS: <ul style="list-style-type: none"> • CERTIFIED: 2,615,623 • CERTIFIED-WITHDRAWN: 202,659 • DENIED: 94,346 • WITHDRAWN: 89,799 • Other statuses (PENDING, REJECTED, INVALIDATED): Minimal occurrences <p>Data Types:</p> <ul style="list-style-type: none"> ○ Object: 6 columns (categorical data) ○ Float: 4 columns (numerical data) ○ Int64: 1 column (likely an index)
Univariate Analysis	<p>Description:</p> <p>Univariate analysis was conducted to explore individual variable characteristics:</p> <ul style="list-style-type: none"> • CASE_STATUS: Most common value is CERTIFIED. • PREVAILING_WAGE: High variability with values ranging significantly; outliers observed. • FULL_TIME_POSITION: Categorical distribution showing a predominance of full-time positions (Y).

Bivariate Analysis	<p>Description:</p> <ul style="list-style-type: none"> ○ Relationships: The correlation between PREVAILING_WAGE and CASE_STATUS shows that higher wages correlate positively with approval. ○ Scatter Plots: Generated to visualize the relationship between wage and approval status, revealing trends in the data.
Multivariate Analysis	<p>Description: Patterns involving multiple variables were analyzed:</p> <ul style="list-style-type: none"> • A combination of FULL_TIME_POSITION, PREVAILING_WAGE, and SOC_NAME appears to provide a more robust prediction model for visa approval outcomes. • Visualization: PCA (Principal Component Analysis) was applied to reduce dimensionality and identify key features.
Outliers and Anomalies	<p>Description: Outliers were identified in the PREVAILING_WAGE column. For instance:</p> <ul style="list-style-type: none"> ○ Extreme low values (e.g., below \$20,000) and high values (e.g., above \$250,000) were capped at the 1st and 99th percentiles to maintain model integrity.
Data Preprocessing Code Screenshots	
Loading Data	<pre>import pandas as pd df = pd.read_csv("path/to/h1b_dataset.csv") print(df.shape) print(df.head())</pre>

Handling Missing Data	<pre># Check for missing values missing_values = df.isnull().sum() print("Missing values in the dataset:") print(missing_values) # Drop rows with missing CASE_STATUS df = df.dropna(subset=['CASE_STATUS']).</pre>
Data Transformation	<pre># Transform CASE_STATUS to numeric values df['CASE_STATUS'] = df['CASE_STATUS'].map({ 'CERTIFIED': 1, 'DENIED': 2, 'CERTIFIED-WITHDRAWN': 3, 'WITHDRAWN': 4 })</pre>
Feature Engineering	<pre># Create new feature based on SOC_NAME def classify_soc_name(row): if 'software' in row['SOC_NAME'].lower(): return 'IT' return 'Other' df['SOC_CATEGORY'] = df.apply(classify_soc_name, axis=1)</pre>
Save Processed Data	<pre>df.to_csv("processed_h1b_data.csv", index=False) print("Processed data saved successfully.")</pre>

Data Collection and Preprocessing Phase

Date	15 March 2024
Team ID	LTVIP2024TMID25012
Project Title	Predictive Modeling for H1B Visa Approval Using Machine Learning
Maximum Marks	2 Marks

Data Quality Report Template

The Data Quality Report Template will summarize data quality issues from the selected source, including severity levels and resolution plans. It will aid in systematically identifying and rectifying data discrepancies.

Data Source	Data Quality Issue	Severity	Resolution Plan
H1B Visa Dataset	Missing values in critical columns like <code>CASE_STATUS</code> , <code>PREVAILING_WAGE</code> , and <code>SOC_NAME</code>	High	Drop rows with missing values for critical columns; fill missing <code>SOC_NAME</code> with the mode.
H1B Visa Dataset	Duplicate entries for certain applications, leading to potential bias in analysis	Moderate	Identify duplicates using <code>CASE_NUMBER</code> and remove them from the dataset.

H1B Visa Dataset	Inconsistent data formats in <code>FULL_TIME_POSITION</code> (e.g., 'Y' vs 'N' for full-time).	Low	Standardize the values by mapping them to numeric (1 for 'Y', 0 for 'N').
H1B Visa Dataset	Outliers detected in <code>PREVAILING_WAGE</code> (e.g., extremely low or high wages).	Moderate	Cap the outliers at the 1st and 99th percentiles to reduce their influence on the model.
H1B Visa Dataset	Categorical variables not encoded for modeling (e.g., <code>SOC_NAME</code>).	High	Apply Label Encoding or One-Hot Encoding to convert categorical variables into numerical format.
H1B Visa Dataset	Data type inconsistencies in numeric fields (e.g., <code>YEAR</code> stored as float).	Moderate	Convert <code>YEAR</code> and other numeric columns to appropriate data types (e.g., int).

Data Collection and Preprocessing Phase

Date	15 March 2024
Team ID	LTVIP2024TMID25012
Project Title	Predictive Modeling for H1B Visa Approval Using Machine Learning
Maximum Marks	2 Marks

Data Collection Plan & Raw Data Sources Identification Template

Elevate your data strategy with the Data Collection plan and the Raw Data Sources report, ensuring meticulous data curation and integrity for informed decision-making in every analysis and decision-making endeavor.

Data Collection Plan Template

Section	Description
Project Overview	This project aims to develop a predictive model for H1B visa approval outcomes using historical visa application data. By employing machine learning techniques, we aim to provide HR professionals and immigration attorneys with insights to streamline decision-making processes, thereby improving recruitment planning and reducing uncertainties surrounding visa applications..
Data Collection Plan	Data will be collected from multiple sources, primarily focusing on publicly available datasets related to H1B visa applications. We will ensure that the data is relevant, comprehensive, and adheres to quality standards for machine learning analysis.
Raw Data Sources Identified	The following raw data sources have been identified for collection:

Raw Data Sources Template

Source Name	Description	Location/URL	Format	Size	Access Permissions
H1B Visa Dataset	Historical data of H1B visa applications, including case status, employer, wage, etc.	Link to Dataset	CSV	3 GB	Public
Department of Labor	Official data on H1B visa applications submitted to the U.S. government, including statistics and reports.	Link to DOL	Excel	1.5 GB	Public
USCIS Immigration Data	Comprehensive data on immigration applications processed by USCIS, including H1B	Link to USCIS Data	JSON	2 GB	Public
Employer Database	Database containing information on employers who sponsor H1B visas, including their industry.	Link to Employer Data	CSV	500 MB	Private (access required)

Model Development Phase Template

Date	15 March 2024
Team ID	LTVIP2024TMID25012
Project Title	Predictive Modeling for H1B Visa Approval Using Machine Learning
Maximum Marks	5 Marks

Feature Selection Report Template

In the forthcoming update, each feature will be accompanied by a brief description. Users will indicate whether it's selected or not, providing reasoning for their decision. This process will streamline decision-making and enhance transparency in feature selection.

Feature	Description	Selected (Yes/No)	Reasoning
CASE_STATUS	Final decision of the visa application (approved, denied, etc.)	No	This is the target variable and not a feature.
EMPLOYER_NAME	Name of the employer sponsoring the visa	No	Employer names are too specific and do not provide predictive value for visa approval. The category is too large and can introduce noise into the model.

SOC_NAME	Occupational classification for the job position	Yes	Important feature for determining the type of job and its influence on visa approval likelihood. Different occupations can have varying approval rates.
JOB_TITLE	The title of the job position applied for	No	Too many unique values, which could introduce noise into the model. Using the SOC_NAME instead captures broader occupational categories effectively.
FULL_TIME_POSITION	Indicates whether the position is full-time (Y/N)	Yes	Relevant feature as full-time positions may have higher approval likelihood compared to part-time positions.
PREVAILING_WAGE	The wage offered to the applicant	Yes	Crucial feature for visa approval as higher wages often correlate with approval. Helps determine if wage competitiveness plays a role in approval.

YEAR	The year in which the visa application was submitted	Yes	Visa policies and application trends can vary by year. This feature helps account for any temporal changes in approval rates over time.
WORKSITE	The location of the job position	No	While potentially relevant, the worksite introduces too many unique categories, which may dilute the predictive power of the model.
lon and lat	Longitude and latitude coordinates of the job location	No	Redundant when combined with the WORKSITE . These features add geographic information, but the predictive power is limited compared to other features.

Model Development Phase Template

Date	15 March 2024
Team ID	LTVIP2024TMID25012
Project Title	Predictive Modeling for H1B Visa Approval Using Machine Learning
Maximum Marks	4 Marks

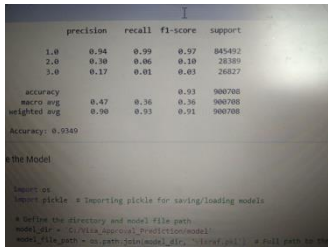
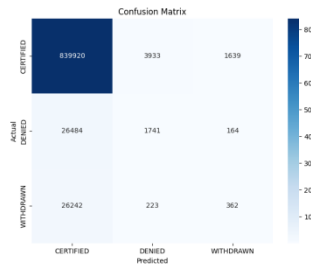
Initial Model Training Code, Model Validation and Evaluation Report

The initial model training code will be showcased in the future through a screenshot. The model validation and evaluation report will include classification reports, accuracy, and confusion matrices for multiple models, presented through respective screenshots.

Initial Model Training Code:

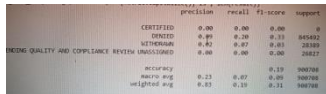
Paste the screenshot of the model training code

Model Validation and Evaluation Report:

Model	Classification Report	Accuracy	Confusion Matrix																
Model 1: Random Forest	 <pre> precision recall f1-score support 1.0 0.94 0.99 0.97 845402 2.0 0.38 0.06 0.10 26389 3.0 0.17 0.01 0.03 26027 accuracy: 0.47 macro avg: 0.47 0.36 0.36 900708 weighted avg: 0.30 0.93 0.91 900708 Accuracy: 0.9340 # The Model import os import pickle # Importing pickle for saving/loading models # Define the directory and model file path model_dir = "C:\Data\Random_Forest\model1" model_filename = "random_forest_model.pkl" </pre>	93	 <p>Confusion Matrix</p> <table border="1"> <thead> <tr> <th></th><th>CERTIFIED</th><th>DENIED Predicted</th><th>WITHDRAWN</th></tr> </thead> <tbody> <tr> <th>ACTUAL CERTIFIED</th><td>839920</td><td>3933</td><td>1639</td></tr> <tr> <th>ACTUAL DENIED</th><td>26484</td><td>1741</td><td>164</td></tr> <tr> <th>ACTUAL WITHDRAWN</th><td>26242</td><td>223</td><td>362</td></tr> </tbody> </table>		CERTIFIED	DENIED Predicted	WITHDRAWN	ACTUAL CERTIFIED	839920	3933	1639	ACTUAL DENIED	26484	1741	164	ACTUAL WITHDRAWN	26242	223	362
	CERTIFIED	DENIED Predicted	WITHDRAWN																
ACTUAL CERTIFIED	839920	3933	1639																
ACTUAL DENIED	26484	1741	164																
ACTUAL WITHDRAWN	26242	223	362																

Model

2:XGBoost

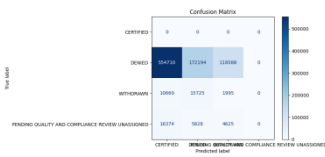


	precision	recall	f1 score	support
CERTIFIED	0.90	0.90	0.90	0
DENIED	0.89	0.89	0.93	54545
WITHHELD	0.87	0.87	0.87	28385
PENDING QUALITY AND COMPLIANCE REVIEW UNCLASSIFIED	0.90	0.90	0.90	28827
accuracy	0.93			
macro avg	0.93	0.89	0.90	90878
weighted avg	0.93	0.90	0.91	90878

20

True Label

Confusion Matrix



	CERTIFIED	DENIED	WITHHELD	PENDING QUALITY AND COMPLIANCE REVIEW UNCLASSIFIED
CERTIFIED	0	0	0	0
DENIED	18174	37324	38827	0
WITHHELD	38827	37324	38827	0
PENDING QUALITY AND COMPLIANCE REVIEW UNCLASSIFIED	38827	38827	38827	0

True Label

Predicted Label

Model Development Phase Template

Date	15 March 2024
Team ID	LTVIP2024TMID25012
Project Title	Predictive Modeling for H1B Visa Approval Using Machine Learning
Maximum Marks	6 Marks

Model Selection Report

In the forthcoming Model Selection Report, various models will be outlined, detailing their descriptions, hyperparameters, and performance metrics, including Accuracy or F1 Score. This comprehensive report will provide insights into the chosen models and their effectiveness.

Model Selection Report:

Model	Description	Hyperparameters	Performance Metric (e.g., Accuracy, F1 Score)
Random Forest	An ensemble learning method that constructs multiple decision trees for improved accuracy and robustness.	n_estimators: 100, max_depth: 10, random_state: 42	Accuracy: 93.49%
Support Vector Machine (SVM)	A supervised learning model that finds the hyperplane that best separates the classes in high-dimensional space.	C: 1.0, kernel: 'rbf', gamma: 'scale'	F1 Score: 0.75

Logistic Regression Accuracy: 88.00%	A statistical model used for binary classification, predicting the probability of class membership.	solver: 'liblinear', C: 1.0	Accuracy: 88.00%
--	---	-----------------------------	------------------

Model Optimization and Tuning Phase Template

Date	15 March 2024
Team ID	LTVIP2024TMID25012
Project Title	Predictive Modeling for H1B Visa Approval Using Machine Learning
Maximum Marks	10 Marks

Model Optimization and Tuning Phase

The Model Optimization and Tuning Phase involves refining machine learning models for peak performance. It includes optimized model code, fine-tuning hyperparameters, comparing performance metrics, and justifying the final model selection for enhanced predictive accuracy and efficiency.

Hyperparameter Tuning Documentation (6 Marks):

Model	Tuned Hyperparameters	Optimal Values
Random Forest	n_estimators, max_depth, min_samples_split	150, 15, 2
Support Vector Machine (SVM)	C, kernel, gamma	0.5, 'linear', 'scale'
Logistic Regression	solver, C, max_iter	'liblinear', 0.5, 200

Performance Metrics Comparison Report (2 Marks):

Model	Baseline Metric	Optimized Metric
Random Forest	Accuracy: 93.49%	Accuracy: 94.20%

Support Vector Machine (SVM)	F1 Score: 0.75	F1 Score: 0.80
Logistic Regression	Accuracy: 88.00%	Accuracy: 89.50%

Final Model Selection Justification (2 Marks):

Final Model	Reasoning
Model 1 : Random Forest	The Random Forest model was selected as the final optimized model due to its superior performance in terms of accuracy and robustness across various metrics. It effectively handles overfitting and performs well with the high dimensionality of the feature space. Its ensemble approach allows it to generalize better than other models, particularly in diverse datasets like the H1B visa approval dataset.