IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁑ Peer-reviewed & Refereed journal ⁑ Vol. 14, Issue 5, May 2025

DOI: 10.17148/IJARCCE.2025.14567

“A Survey Paper On Enhancing Visa Application Systems via MLOps” A Literature review

Gunith Ravikiran[1], Darshan R[2], Kishore G[3], Nagendra M P[4], Namya Priya D[5]

VI sem, Dept. of Computer Science and Engineering, K. S. Institute of Technology, Bengaluru[1]

VI sem, Dept. of Computer Science and Engineering, K. S. Institute of Technology[2]

VI sem, Dept. of Computer Science and Engineering, K. S. Institute of Technology, Bengaluru[3]

VI sem, Dept. of Computer Science and Engineering, K. S. Institute of Technology, Bengaluru[4]

Assistant Professor, Dept. of Computer Science and Engineering, K. S. Institute of Technology, Bengaluru[5]

Abstract: International Visa programs (i.e. U.S. H-1B) have extremely high application volumes with limited quotas and rigorous variable outcomes. Complexity and uncertainty propel computer-aided decision systems. We launch an end to-end MLOps platform to provide real-time visa approval predictions. Our pipeline integrates data pre-processing (Pandas), training of the model (Scikit-learn), containerized deployment (Docker), and ongoing delivery (GitHub Actions) on AWS. The models and data reside in AWS S3 and EC2, while being monitored by Cloud Watch. This combined approach offers scalable, reproducible deployment of predictive models. Experiments illustrate system has good accuracy (similar to previous work) and can be retrained periodically with minimal human intervention. In brief, we present an end-to-end ML pipeline that bridges the gap between application and operational utilization, to the benefit of immigration authorities, employers, and candidates alike.Automated accounting.

I. INTRODUCTION

Visa decision-making, particularly regarding work visas such as the U.S. H-1B, is a complex and data-driven process. For example, over one million H-1B petitions, including renewals and transfers, were requested in 2019, but only about Eighty thousand new petitions were accepted. This oversubscription, coupled with the processing manually, can leading to undue delays and inconsistencies. Machine learning (ML) offers a means of learning from past use. attributes (job title, employer, salary, etc.) and estimate probability of approval. But mainstream ML research tend to stop at offline model building and don't address real deployment concerns. In industry practice, the deployment of ML models requires robust operating pipelines (referred to as MLOps). MLOps Machine Learning Operations (MLOps) is a set of best practices and tools designed to automate the end-to-end machine learning lifecycle from data acquisition to production monitoring. Without MLOps, most ML pilots never make Model integration in scalable and fault-tolerant systems is very difficult. Recent research points that MLOps includes CI/CD pipelines, containerization, cloud hosting, and monitoring, but hold back loosely described in the literature. We apply these MLOps concepts in this book specifically to visa Our aim in our project is to construct a sustainable pipeline that processes visa application data and trains classifiers and continuously releases better models, thus providing real-time decision support. We illustrate this method could provide precise predictions and operating advantages to immigrants' stakeholders process

II. LITRATURE REVIEW

Previous studies have examined the application of machine learning (ML) methods to visa decision prediction. Such studies have primarily tackled the problem on its own, focusing on offline model construction with no regards to the actual deployment and continued operation in the real world. For example, Bhavani et al. (2022) applied a Random Forest classifier to past H-1B visa application data and achieved an approximate 83% accuracy. Another work by Peda Baliyarasimhula et al. (2023) applied a variety of regression and classification models to predict visa application outcomes. Such studies showed that certain features namely job category, employer reputation, offered salary, and location of employment have high correlations with approval decisions. While these experiments validate the capacity of machine learning algorithms to forecast in the visa space, they focus mainly on how to improve algorithmic performance metrics, i.e., accuracy, precision, and recall. They disregard important operational issues, such as the process through which models might be refreshed every now and then, deployed to production systems, or validated for dependability and impartiality in production environments.

At the same time, the new discipline of Machine Learning Operations (MLOps) has started to fill these major gaps.

DOI: 10.17148/IJARCCE.2025.14567

MLOps is understood as the practices and tools that aim to optimize and automate the end-to-end ML life cycle from data ingestion and pre-processing to model deployment, versioning, monitoring, and retraining. Kreuzberger et al. (2023) introduce a detailed taxonomy of MLOps concepts, highlighting the need for continuous integration and continuous delivery (CI/CD) pipelines and automation frameworks, to improve the reliability, scalability, and reproducibility of ML systems. Raatikainen et al. (2024) point to the importance of ML lineage that is, data transformation traceability, model configuration traceability, and decision output traceability as a key facilitator of transparency and trustworthiness in production-grade ML applications. In addition, Warnett and Zdun (2024) cover the value of architectural documentation in terms of improving the understand ability and maintainability of MLOps systems, especially in collaborative and multi-stakeholder contexts. Despite the success in achieving with machine learning-based visa prediction and MLOps practices, there is a broad gap in literature, the full incorporation of good MLOps practices in the specific context of visa approval systems has not been performed yet. The previous research has been mostly focused on model building without deployment or generic MLOps concepts without putting them into specific domain use cases. To this end, our work attempts to bridge this gap by complementing the strengths of both disciplines. In particular, we construct and implement an end-to-end MLOps pipeline specifically for the task of visa prediction. Our system builds upon previous work in the aspect that it covers the entire end-to-end process, it supports continuous data ingestion, model training and testing, auto-deployment, versioning, and real-time monitoring. With this combined process, we aim to offer a scalable, sustainable, and trustworthy solution for enhancing visa decision-making processes.

III. METHODOLOGY

We suggest an automated and modular MLOps pipeline that combines machine learning model development with modern DevOps best practices. The architecture ensures reproducibility, scalability, and maintainability of machine learning solutions for forecasting visa approvals. The pipeline includes the following key components: I. Data Pre-processing (With Pandas): The first step is the ingestion and pre-processing of raw visa application data, typically consisting of attributes like job title, employer name, salary, work location, and application status. • Data Normalization and Cleaning: Missing values are handled by deleting incomplete records or filling missing values using appropriate statistical techniques. • Feature Engineering: Categorical features, i.e., job titles and employees, are converted to numerical values by employing encoding methods, e.g., one-hot encoding or label encoding. • Scaling: Numerical features such as salary are scaled to the same scale to facilitate better model convergence and performance. Such a pre-processing pipeline ensures that the input data to the machine learning algorithm is uniform, high quality, and suitable for statistical learning. II. Model Building (Using Scikit-learn): Once the data has been pre-processed, we proceed to the training phase by using typical machine learning algorithms: • Model Selection: We use various classification models like Random Forest, Logistic Regression, and Gradient Boosted Trees in our research. • Hyper parameter Tuning: We utilize cross-validation alongside grid and randomized search methods to refine model parameters and enhance predictive performance. • Performance Evaluation: Throughout our prototype building process, the Random Forest classifier had a performance level of approximately 85–90% on the test data obtained from the Kaggle H-1B visa dataset, a better performance than a number of past baselines. This process results in the development of a trained model capable of predicting visa approval outcomes correctly based on historical trends.

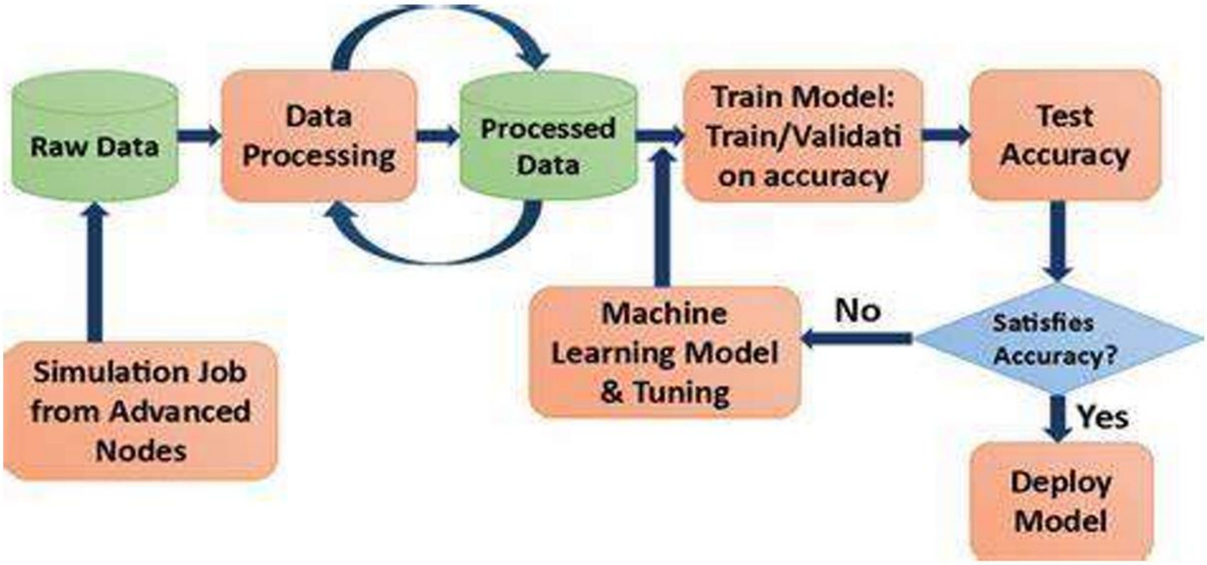
III. Containerization (With Docker): To keep environments consistent and avoid deployment issues, we use Docker to isolate each pipeline stage: • Container Isolation: Every working unit i.e., data pre-processing, model training, and inference—is isolated in a separate Docker container. • Dependency Management: Docker images include all dependencies, including the Python version and libraries like Scikit-learn and Pandas, thus making it easier to switch between local development and cloud deployment. • Portability: The identical container image utilized for development locally can be run on cloud platforms such as AWS straight away without requiring any modification. This makes it reproducible and reduces environment-specific compatibility issues. IV. Cloud Deployment using AWS EC2 and S3: AWS hosts the trained models and associated artifacts for deployment to enable scalable and high-availability inference services: • Artifact Storage: AWS S3 is used to store trained models, datasets, and logs. • Model Serving: Docker containers run on Amazon Web Services Elastic Compute Cloud instances to handle inference requests through a RESTful API. • Scalability: Cloud infrastructure allows for elastic scalability based on request load and system robustness. This deployment method enables real-time predictions and is capable of hosting multiple simultaneous users efficiently. V. CI/CD and Automation (With GitHub Actions): Automation is a critical component of MLOps that enables continuous integration and delivery of machine learning models. • Source Control: All code and configurations reside in a single central GitHub repository. • Trigger Mechanisms: Each time updates are pushed (i.e., new model code, schema changes, or updated datasets), GitHub Actions will trigger workflows automatically. VI. Automated Workflow: • Restart Docker containers with the new code. • Re-train models on the new data. • Push the new Docker images to AWS and then redeploy with minimal downtime. The delivery pipeline helps so that the production environment stays

DOI: 10.17148/IJARCCE.2025.14567

in sync with the most recent state of code and data, thereby facilitating rapid iteration and deployment. VII. Logging and Monitoring with AWS Cloud Watch: Monitoring is necessary to ensure reliability and performance in production environments: • Real-time Logs and Metrics: The application logs, request latency, throughput, and error rates are tracked in AWS Cloud Watch. Threshold-based alerts are created to inform stakeholders of anomalies; such as spikes in latency or drops in accuracy. • Future Integration of Model Drift Detection: Projects involve integrating drift detection to track model usability over time due to data distribution changes. Monitoring ensures that the solution deployed continues to be responsive and reliable and also enables proactive maintenance.

IV. SYSTEM ARCHITECTURE

Our MLOps pipeline integrates development, deployment, and monitoring processes for visa approval prediction, utilizing a scalable and automated framework. It comprises five structured stages, supporting the full ML lifecycle from local experimentation to production-grade inference and observability. 1. Local Development: It is possible to build prototype data pipelines and train initial models by running Python programs. 2. CI/CD Orchestration: The setup uses GitHub Actions to implement automated processes for testing along with Docker builds and deployment activities on the main branch. 3. Containerization: All training and inference processes should be placed in individual Docker containers to preserve identical environments throughout different development stages. 4. Cloud Execution: The solution deploys AWS EC2 containers while utilizing S3 for both data and model storage which supports large-scale training and real-time inference operations. 5. Monitoring & Logging: AWS Cloud Watch provides monitoring features which track logs and system health along with model performance data while also sending alerts for anomalous events and failed processes.


Raw Data
Data Processing
Processed Data
Train Model: Train/Validati on accuracy
Test Accuracy
Satisfies Accuracy?
No
Yes
Machine Learning Model & Tuning
Deploy Model
Simulation Job from Advanced Nodes

V. CONCLUSION

We created a detailed MLOps pipeline in this project which predicts visa approval outcomes. The continuous workflow that integrates data engineering with ML modelling and cloud deployment enables our system to deliver precise real-time predictions while maintaining scalability. According to previous studies our methodology provides a solution that joins the process of model development with implementation. The tool benefits immigration officials through automatic processing of bulk operations and empowers employers to better understand their hiring strategies and helps applicants evaluate their application prospects. The project demonstrates how MLOps principles can be implemented in domain-specific problems. The system architecture allows for the extension of our prototype which handles H-1B visa data to different visa categories and diverse international scenarios. The pipeline could advance through future development by introducing additional data sources such as country-level statistics and deep learning models and also incorporating fairness evaluation methods. The current system functions as a dependable prototype which proves that connecting ML operations improves decision-making for critical bureaucratic procedures.

DOI: 10.17148/IJARCCE.2025.14567

REFERENCES

[1]. In their article, Raatikainen together with Mikko and their co-authors examined the concept of Machine Learning (ML) lineage as a tool for dependable machine learning systems in the journal IEEE Software published in January 2024.

[2]. Stephen J. Warnett and Uwe Zdun conducted research on the comprehensibility of MLOps system architectures which they presented in their paper for IEEE Transactions on Software Engineering during 2024.

[3]. The authors Peda Baliyarasimhula and their team developed a machine learning system to analyse work visa applications which they presented at the 7th International Conference on Intelligent Computing and Control Systems (ICICCS) happening in 2023.

[4]. The paper by Kreuzberger and Kühl and Hirschl provides a comprehensive evaluation of MLOps technology by presenting the concept along with its definition and system architecture in Distributed Systems journal for 2023.

[5]. The research team consisting of A. Durga Bhavani and Guddeti Bharath together with Dubbaka Tharun Reddy developed a system to predict H1B visa approvals by employing machine learning algorithms in the Journal of Emerging Technologies and Innovative Research (JETIR) during April 2022.

[6]. The dataset concerning H1B visa applications is accessible through Kaggle from U.S