

# MLOps Certification Course

---

## Capstone Project

## Problem Statement

In many industries, salary structures are influenced significantly by the level of experience an individual possesses. Accurate salary predictions based on such a variable can help organizations in budget planning and ensuring equitable salary distributions. This project aims to develop a predictive model that estimates salaries using years of experience as the primary predictor. The challenge is to build and compare multiple regression models to identify which model provides the most accurate and reliable predictions.

To enhance the development process and ensure model reliability post-deployment, the project will integrate MLflow and WhyLabs. MLflow will be used for managing the machine learning lifecycle, including experiment tracking, model versioning, and deployment. WhyLabs will be utilized to monitor the model in production to track performance metrics and detect any deviations or anomalies in real-time.

## Objectives

1. Model Development: Develop three types of regression models:
2. Linear Regression: To establish a baseline for performance comparison.
3. Random Forest Regressor: To assess performance improvements using an ensemble method.
4. Decision Tree Regressor: To evaluate a non-linear regression approach.
5. Experiment Tracking: Use MLflow for tracking experiments, including parameters, metrics, models, and artifacts to facilitate model comparisons and reproduce results.
6. Model Deployment: Deploy the best-performing model into a production-like environment using MLflow.
7. Model Monitoring: Implement WhyLabs to monitor the deployed model's performance in production, detecting any operational drifts or anomalies that may affect predictions.

## Dataset

The dataset will be sourced from Kaggle and is titled "Salary Dataset." It is available in the LMS.

## Expected Outcomes

The project aims to deliver:

1. A comparison of model performances to identify the most effective regression technique for the given data.

2. A fully tracked and documented ML project via MLflow, with models ready for deployment.
3. An operational model in production with ongoing performance monitoring through WhyLabs, ensuring the model's reliability and accuracy over time.

### **Implementation Steps**

1. Set Up Environment: Install and configure all necessary libraries and tools.
2. Data Collection and Preprocessing: Acquire and preprocess data to ensure quality inputs for model training.
3. Model Training and Experiment Tracking: Train models while tracking experiments using MLflow.
4. Model Deployment: Deploy the best-performing model using MLflow's deployment capabilities.
5. Model Monitoring: Monitor the model in production using WhyLabs to ensure sustained performance and accuracy.

By addressing these objectives, the project seeks to not only provide accurate salary predictions but also leverage advanced tools to enhance the transparency and maintainability of the machine learning lifecycle.