IT3385 MACHINE LEARNING OPERATIONS

Assignment (45%)

AY2024 Semester 2

Table of Contents

Introduction

Machine Learning Operations (MLOps) streamlines the process of deploying Machine Learning models to production and then maintaining and monitoring them. MLOps is a collaborative function, often consisting of data scientists, ML engineers, and DevOps engineers. The word MLOps is a compound of two different fields i.e. machine learning and DevOps from software engineering.

MLOps can encompass everything from the data pipeline to machine learning model production. In some places, you will see MLOps implementation is only for the deployment of the machine learning model but you will also find enterprises with the implementation of MLOps across many different areas of ML Lifecycle development, including Exploratory Data Analysis (EDA), Data Preprocessing, Model Training, etc.

In this assignment, you will be working with **one/two teammate(s)** to build machine learning models and deploy them using a web application. The implementation by your team should consider the various aspects of the MLOps lifecycle, i.e. from integrating with model generation (software development lifecycle and continuous integration/continuous delivery), orchestration, and deployment, to health, diagnostics, governance and business metrics. Normally machine learning models are built so that they can be used to predict an outcome (binary value i.e. 1 or 0 for Classification, continuous values for Regression, labels for Clustering etc.

Tasks Involved in Assignment

Below is the list of tasks to be performed in this assignment

S/N	Task	Individual / Team
1.	Perform an Exploratory Data Analysis on a Given Dataset	Individual
2.	Train, validate and develop a Machine Learning pipeline using PyCaret.	Individual
3.	Build and deploy a front-end web application with real-time prediction	Individual
4.	Set up development and deployment environment according to MLOps Lifecycle	Team

Please note that for individual components, each member is required to work on ONE dataset given in Annex A. You should not use the same dataset as your teammate(s). Each of you will create an ML Pipeline for your modelling workflow.

The base marks of this assignment are **45 marks,** and it constitutes **45%** of your total ICA marks for this competency unit.

Copy work from other people or the internet is strictly prohibited. If found, it will be considered a case of plagiarism and is subject to disciplinary actions

Late Submission Penalty

Late submission of assignment may subject to penalty as shown below.

No. of Calendar Days	Penalty
If the submission is <=5 working days	Cap at 50% (of the base marks)
If the submission is > 5 working days	0 marks will be awarded

Submission Format and Mode

Below are the required deliverables for this assignment.

1. An URL to your GitHub repository which contains the codes for model training by the team. You may also zip the entire repository and submit it via Brightspace.
2. A demo video by the team
3. A readme file for the deployment guide & folder structure and an URL to the web app

Please be reminded to submit all the deliverables via Brightspace **by 2 Mar 2025 (Sunday) at 2359hrs.**

Please refer to **Annex B** for detailed assessment rubrics.

Task 1: Perform an Exploratory Data Analysis on a Given Dataset (5 marks)

Exploratory Data Analysis (EDA) is an approach that is used to analyse the data and discover trends, and patterns, or check assumptions in data with the help of statistical summaries and graphical representations.

You may make use of Python libraries and visualisation libraries to assess the relationship (correlation, distribution etc) among the variables.

Task 2: Train, validate and develop a Machine Learning pipeline using PyCaret. (10 marks)

You may train and validate your models in Integrated Development Environment (IDE) or Jupiter Notebook either on your local machine or on the cloud. You are required to use PyCaret to develop a machine-learning pipeline and train your models.

1. The training environment should be initialised with some additional **preprocessing tasks** such as scaling and normalisation, automatic feature engineering and binning continuous data into intervals etc based on your modelling and dataset which subsequently creates a transformation pipeline.

2. Models are trained and evaluated according to multiple **performance metrics** using k-fold cross-validation.

3. Analyse the performance of a trained model on the test set using the plot_model or evaluate_model function.

4. Use the final model to generate predictions on unseen data with the predict_model function.

5. Save the entire pipeline

6. Enable experiment logging during the training of models.

7. Proper documentation is required for each step

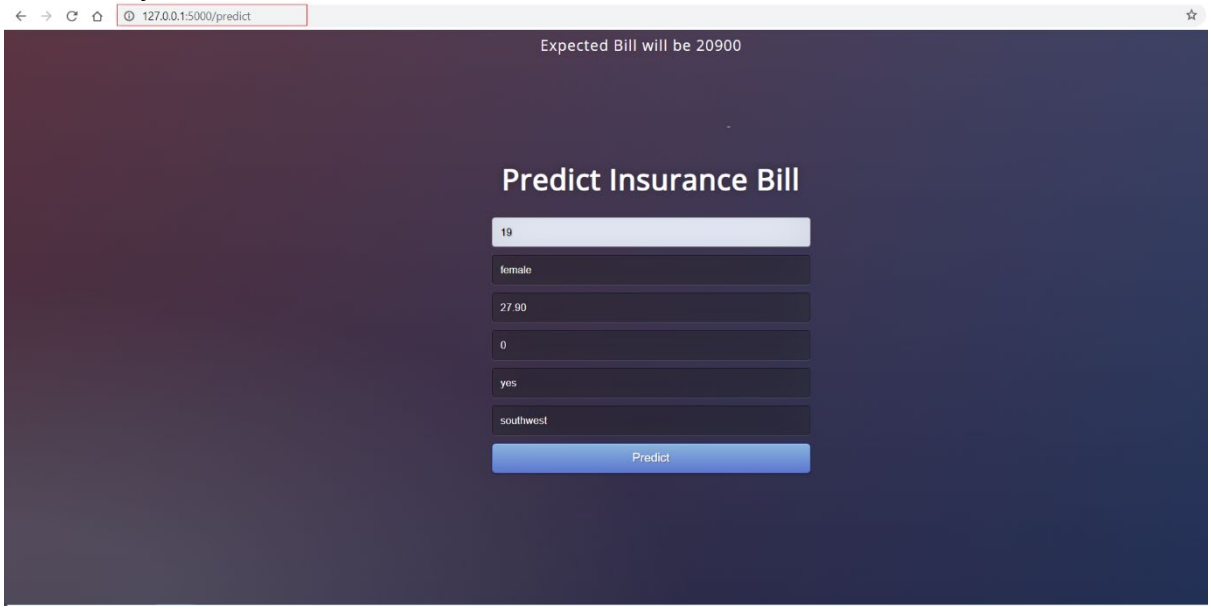
8. Register the model with MLFlow

Task 3: Build and deploy a front-end web application with real-time prediction (15 marks)

Once your machine learning pipeline and model are ready, you can start building a web application that can connect to them and generate predictions on new data in real time. There are two parts to this application:

- Front-end (designed using HTML) to get inputs from users
- Back-end (developed using Flask in Python) – connect the model and return predictions based on a given input.

Below is an example of a real-time prediction of insurance bills based on inputs entered by users.



The web application should be tested on a local machine and deployed to the cloud or any platform as a service (PaaS) that enables the deployment of web apps based on a managed container system, with integrated data services and a powerful ecosystem e.g. Heroku or Render.

Task 4: Set up development and deployment environment according to MLOps Lifecycle (15 marks)

The development and deployment environment for your team should be created according to the best practices for ML Model productisation. Below are some of the recommendations for ML Model productivisation

- Follow a standard ML project folder structure
- Use Poetry for dependency management
- Use Hydra for managing configuration files and minimising hard coding
- Use DVC for data version control

The environment should support collaborative work, easy check-in & out of codes and quick deployment. The team is required to do a 15-minute presentation for the demonstration of the entire MLOps lifecycle using their environment and ML web application.

Annex A: Datasets

Dataset 1

Business Objective: You are tasked to train an ML model to predict the selling price of a property in Melbourne based on its characteristics e.g. distance to the Central Business District (CBD), land size, number of bedrooms, building area, age of the property etc.

Data Source: **01_Melbourne_Residential.csv**

Data Column	Description
Suburb	Suburb of the residential property
Address	Address of the residential property
Rooms	Number of rooms
Type	br - bedroom(s); h - house,cottage,villa, semi,terrace; u - unit, duplex; t - townhouse; dev site - development site; o res - other residential
Price	Price in dollars
Method	S - property sold; SP - property sold prior; PI - property passed in; PN - sold prior not disclosed; SN - sold not disclosed; NB - no bid; VB - vendor bid; W - withdrawn prior to auction; SA - sold after auction; SS - sold after auction price not disclosed. N/A - price or highest bid not available.
Seller	Name of Real Estate Agent
Date	Date sold
Distance	Distance of property to CBD in km
Postcode	Postal code of the property

Bedroom2	Scraped # of Bedrooms (from different source)
Bathroom	Number of Bathrooms
Car	Number of parking spots for cars
Landsize	Land Size
BuildingArea	Building Size
YearBuilt	Year in which the property was built
CouncilArea	Governing council for the area
Latitude	Latitude of the property
Longitude	Longitude of the property
Region	Region of the property
Propertycount	Number of properties that exist in the suburb.

Dataset 2

Business Objective: You are tasked to train an ML model to predict the selling price of a used car in India based on its characteristics e.g. kilometers driven, fuel type, transmission, mileage etc.

Data Source: **02_Used_Car_Prices.xlsx**

Data Column	Description
Brand_Model	The brand and model of the car.
Location	The location in which the car is being sold or is available for purchase.
Year	The year or edition of the model.
Kilometers_Driven	The total kilometres driven in the car by the previous owner(s) in KM.
Fuel_Type	The type of fuel used by the car. (Petrol / Diesel / Electric / CNG / LPG)
Transmission	The type of transmission used by the car. (Automatic / Manual)
Owner_Type	Whether the ownership is Firsthand, Second hand or other.
Mileage	The standard mileage offered by the car company in kmpl or km/kg.
Engine	The displacement volume of the engine in cc.
Power	The maximum power of the engine in bhp.
Seats	The number of seats in the car.
Price	The price of the used car in INR Lakhs.

Dataset 3

Business Objective: This dataset contains the data about kernels belonging to three different varieties of wheat: Kama, Rosa and Canadian, 70 elements each, randomly selected for the experiment. You are tasked to train an ML model to determine if type of wheat based on attributes of its kernel given in the dataset.

Data Source: **03_Wheat_Seeds.csv**

Below are the geometric parameters of wheat kernels.

Data Column	**Description**
Area	Area of wheat kernel
Perimeter	Perimeter of wheat kernel
Compactness	Compactness is calculated using the following formula. Compactness = 4*pi*Area/Perimeter^2
Length	Length of kernel
Width	Width of kernel
AsymmetryCoeff	Asymmetry coefficient of kernel
Groove	Length of kernel groove
Type	1- Kama, 2 - Rosa and 3 - Canadian,

Annex B: Assessment Rubrics

Individual Component

Criteria	Advanced	Proficient	Functional	Developing	Not Competent
Task 1 - Perform an Exploratory Data Analysis on a Given Dataset **(5 marks)**	Intense data exploration by looking at data in many different ways. All the visualisations address the required analysis concisely and clearly. *(4.1 – 5 marks)*	Plenty of data exploration and investigation on the relationship among the variables. Most of the visualisations address the required analysis concisely and clearly. *(3.1 – 4 marks)*	Some data exploration and investigation on the relationship among the variables. At least half of the visualisations address the required analysis concisely and clearly. *(2.1 – 3 marks)*	Limited data exploration and investigation on the relationship among the variables. Only a few visualisations address the required analysis concisely and clearly. *(1.1 – 2 marks)*	No evidence of data exploration and investigation on the relationship among the variables. The visualisation did not address the required analysis concisely and clearly. *(0 – 1 mark)*
Task 2 - Train, validate and develop a Machine Learning pipeline using PyCaret. **(10 marks)**	Performed thorough and comprehensive data cleansing and transformation Able to choose the best performing model by performing thorough tuning for the best hyper-parameter values Uses the comprehensive performance metric for the model evaluation *(8.1 – 10 marks)*	Performed most of the required data cleansing and transformation Able to choose the best performing model by performing tuning for the good hyper-parameter values. Uses the appropriate performance metric for the model evaluation *(6.1 – 8 marks)*	Performed some of the required data cleansing and transformation Able to choose the best performing model using PyCaret function Uses some performance metric for the model evaluation *(4.1 – 6 marks)*	Performed little data cleansing and transformation Able to choose the best performing model using PyCaret function Uses inappropriate performance metric for the model evaluation *(2.1 – 4 marks)*	No evidence of data cleansing and transformation Did not choose the best performing model using PyCaret function Did not use performance metric for the model evaluation *(0 – 2 marks)*
Task 3 - Build and deploy a front-end web application with real-time prediction	Source codes are working flawlessly, very organised and with detailed comments	Source codes are working fine, organised and properly commented. Codes are clear.	Source codes are working with few bugs, organised and minimal comments. Codes are unclear and	Source codes are working with some bugs, disorganised and minimal comments. Codes	Source codes are buggy, not working, disorganised and without comments. Hardly

(15 marks)	providing explanation for complex steps. Codes are well structured and follows good programming practices. ML application was deployed and generated real-time predictions for single and batch request. *(12.1 – 15 marks)*	Follows good programming practices. ML application was deployed and generated real-time predictions for single request. *(9.1 – 12 marks)*	difficult to follow. Follows some good programming practices. ML application was deployed and generated real-time predictions for single request with some bugs. *(6.1 – 9 marks)*	are unclear and difficult to follow. Follows some good programming practices. ML application was deployed and no prediction generated. *(3.1 – 6 marks)*	follows any good programming practices. ML application was not deployed *(0 – 3 marks)*

Group Component

Task 4 - Set up development and deployment environment according to MLOps Lifecycle					
Criteria	**Advanced**	**Proficient**	**Functional**	**Developing**	**Not Competent**
Set up development and deployment environment according to MLOps Lifecycle **(10 marks)**	All the tools and libraries are properly integrated and functional for supporting CI/CD operations in ML model deployment. *(8.1 – 10 marks)*	Majority of the tools and libraries are properly integrated and functional for supporting CI/CD operations in ML model deployment. *(6.1 – 8 marks)*	Some of the tools and libraries are properly integrated and functional for supporting CI/CD operations in ML model deployment. *(4.1 – 6 marks)*	Some of the tools and libraries are properly integrated but not functional for supporting CI/CD operations in ML model deployment. *(2.1 – 4 marks)*	None of the tools and libraries are used. *(0 – 2 marks)*
Presentation **(5 marks)**	Presentation is very clear, well organised, and compelling. *(4.1 – 5 marks)*	Presentation are mostly clear, reasonably well crafted and organised. *(3.1 – 4 marks)*	Presentation are mostly clear and reasonably organised. *(2.1 – 3 marks)*	Presentation are unclear and not organised. *(1.1 – 2 marks)*	Presentation are confusing, clearly unpolished and disorganised *(0 – 1 mark)*