Customer Churn Prediction Ops Pipeline

IF4054 Final Project
Semester I 2024/25
Ver 20 Dec 2024

INTRODUCTION

You're a MLOps/Dataops specialist working for a telco and are responsible to develop a working end-to-end pipeline and workflow for customer churn prediction using Apache Spark, Apache Airflow, and MLflow. The project focuses on constructing scalable and reproducible workflows for data preprocessing, model training, evaluation, and deployment tracking.

Student needs to:

1) Come up with a pipeline and workflows for data cleanup (watch for nulls and empty string etc.)
2) Deliver drift simulation workflow
3) Monitor the drift that happens and adjust model for retraining automatically.

DATASET AND METHOD

Dataset to be used: https://www.kaggle.com/datasets/blastchar/telco-customer-churn/code

- Churn is the target / dependent variable
- **Use Population Stability Index (PSI)** that is commonly used in **data drift detection**, especially in industries like finance, banking, and insurance.

TOOLS:

Apache Spark: For big data processing and feature engineering.
Apache Airflow: For orchestrating and scheduling pipeline workflows.
MLflow: For experiment tracking, model management, and deployment.
Containerization/Orchestration: Docker/Kubernetes
CI/CD and Repository: Gitlab

TEAMS:

Student form a team of 3-4 member for each team.

DELIVERABLES:

These should be submitted before 10th of January 2025 via MsTeams

1) Pipeline and workflow artefact using forementioned tools (ZIP file)
2) Project Report (including each role within Team) (PDF Document)