MLOPS Project Report: Credit Risk Classification

Team: Prateek Majumder, Neha Roy Choudhury, Pranjal Grover, Anshuman Jha, Rahul Sanjay Trivedi

Introduction & Approach

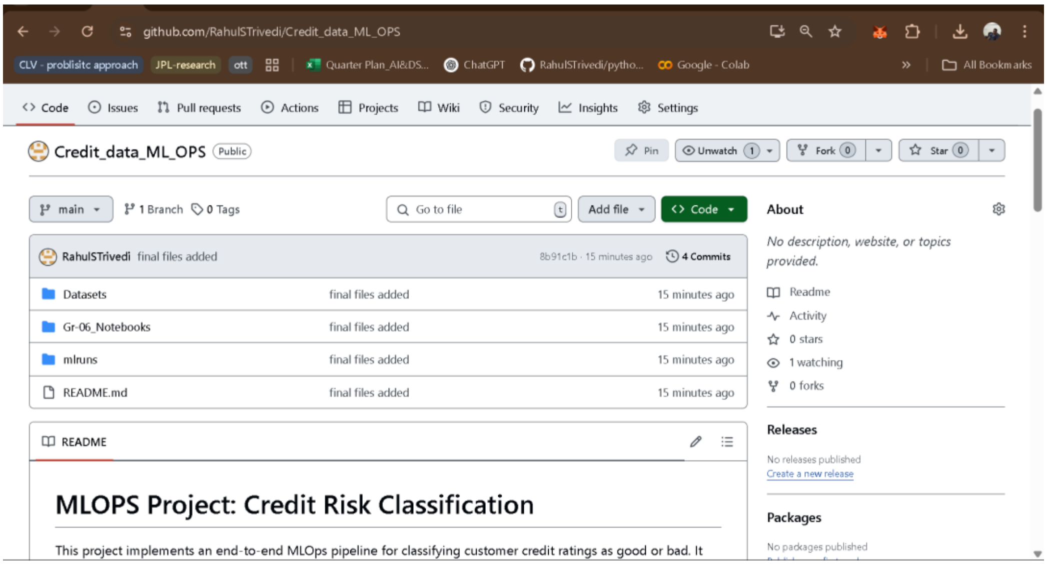
This project aims to classify customer credit ratings as good or bad using an end-to-end MLOps pipeline. The solution includes data ingestion, validation, model training, deployment, and monitoring.

Key Steps

- **Dataset Preparation:**
    - Defined dataset schema and validated data using Pydantic.
    - Stored the dataset in Parquet format and split it into training, testing, and production sets.
- **Version Control:**
    - All dataset versions and notebooks were tracked and organized in a GitHub repository.
- **ML Pipeline & Experimentation:**
    - Built an ML pipeline using scikit-learn and tracked multiple experiments using MLflow.
    - The best-performing model was identified as Random Forest based on cross-validation and test metrics.
- **Deployment:**
    - The selected model was deployed as a RESTful API using FastAPI.
- **User Interface:**
    - A Streamlit application was developed to interact with the API, allowing users to enter data and receive predictions.
- **Monitoring:**
    - Implemented data drift analysis (both numeric and categorical) to monitor production data for consistency.

Snapshots & Artifacts

GitHub Repository: https://github.com/RahulSTrivedi/Credit_data_ML_



github.com/RahulSTrivedi/Credit_data_ML_OPS

CLV - problisitc approach | JPL-research | ott | Quarter Plan_AI&DS... | ChatGPT | RahulSTrivedi/pytho... | Google - Colab | All Bookmarks

Credit_data_ML_OPS Public

1 Branch · 0 Tags

RahulSTrivedi final files added · 8b91c1b · 15 minutes ago · 4 Commits

Datasets	final files added	15 minutes ago
Gr-06_Notebooks	final files added	15 minutes ago
mlruns	final files added	15 minutes ago
README.md	final files added	15 minutes ago

README

MLOPS Project: Credit Risk Classification

This project implements an end-to-end MLOps pipeline for classifying customer credit ratings as good or bad. It

About

No description, website, or topics provided.

Readme
Activity
0 stars
1 watching
0 forks

Releases

No releases published
Create a new release

Packages

No packages published

Model Pipeline Flow

Model Pipeline Flow

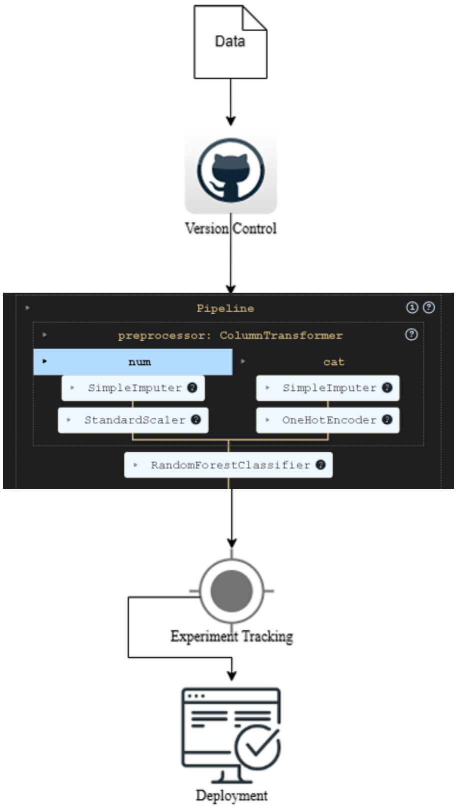
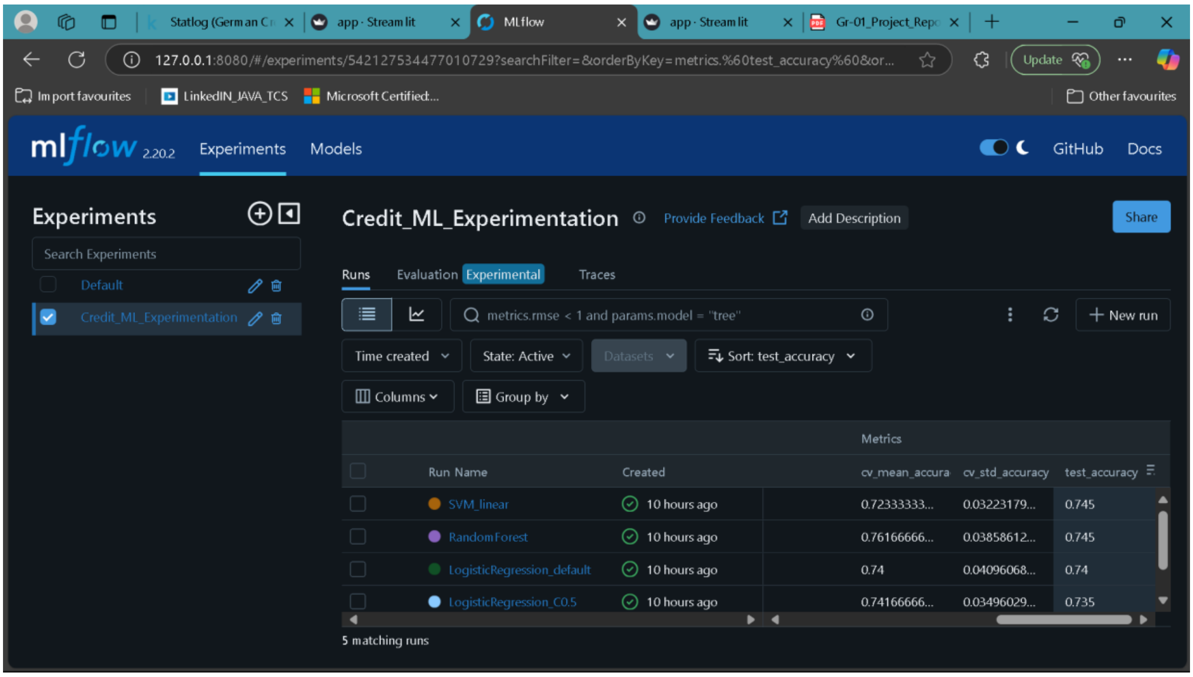


Figure: Flowchart of the ML pipeline

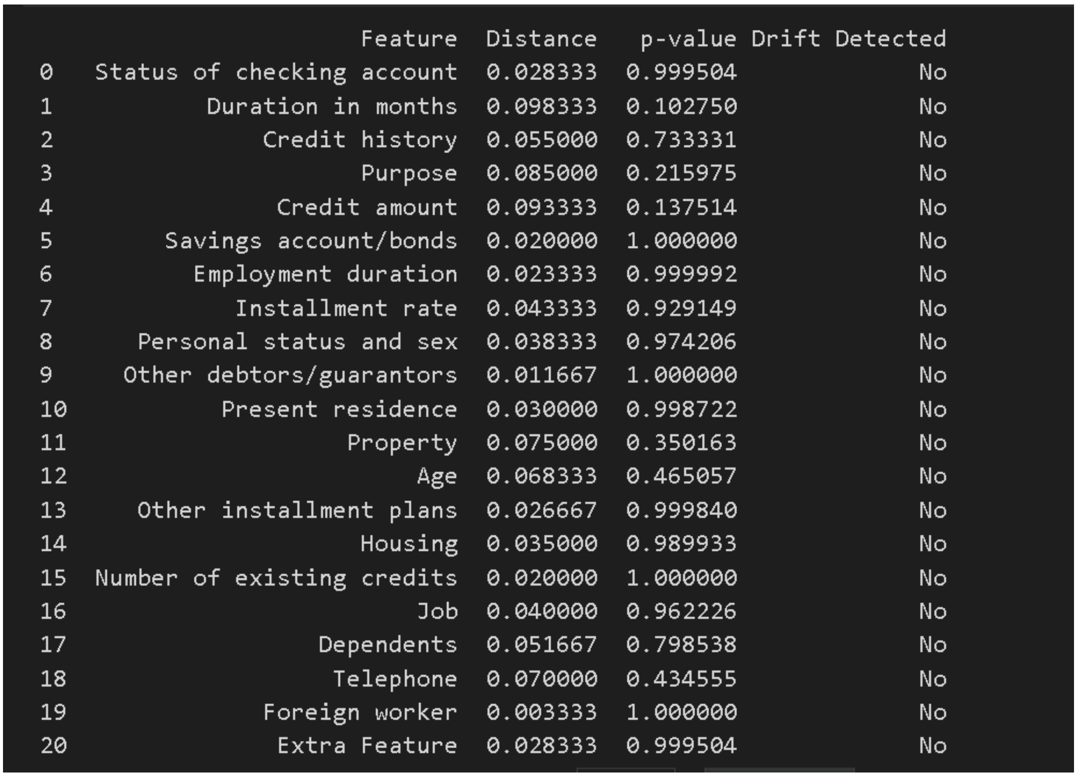
Experiment Tracking



			Metrics		
□	Run Name	Created	cv_mean_accura	cv_std_accuracy	test_accuracy
□	SVM_linear	10 hours ago	0.72333333...	0.03223179...	0.745
□	RandomForest	10 hours ago	0.76166666...	0.03858612...	0.745
□	LogisticRegression_default	10 hours ago	0.74	0.04096068...	0.74
□	LogisticRegression_C0.5	10 hours ago	0.74166666...	0.03496029...	0.735

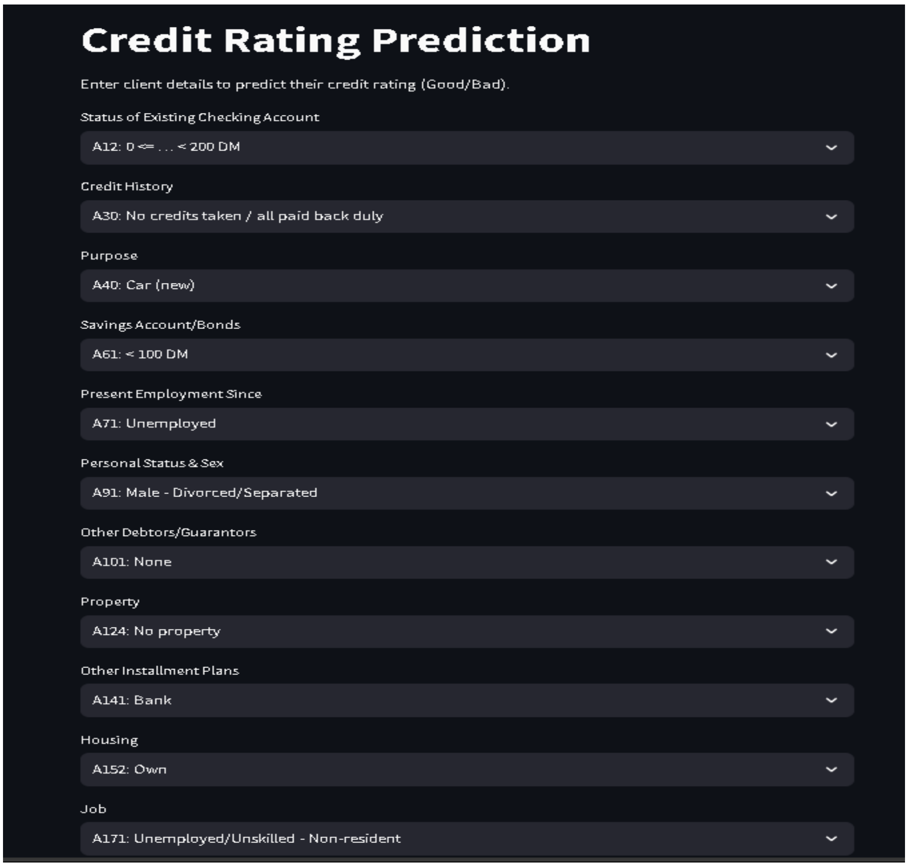
5 matching runs

Data Drift Results



```
                        Feature  Distance   p-value Drift Detected
0   Status of checking account  0.028333  0.999504             No
1           Duration in months  0.098333  0.102750             No
2               Credit history  0.055000  0.733331             No
3                      Purpose  0.085000  0.215975             No
4                Credit amount  0.093333  0.137514             No
5        Savings account/bonds  0.020000  1.000000             No
6          Employment duration  0.023333  0.999992             No
7             Installment rate  0.043333  0.929149             No
8         Personal status and sex  0.038333  0.974206          No
9      Other debtors/guarantors  0.011667  1.000000             No
10           Present residence  0.030000  0.998722             No
11                    Property  0.075000  0.350163             No
12                         Age  0.068333  0.465057             No
13     Other installment plans  0.026667  0.999840             No
14                     Housing  0.035000  0.989933             No
15  Number of existing credits  0.020000  1.000000             No
16                         Job  0.040000  0.962226             No
17                  Dependents  0.051667  0.798538             No
18                   Telephone  0.070000  0.434555             No
19              Foreign worker  0.003333  1.000000             No
20               Extra Feature  0.028333  0.999504             No
```

UI Predictions



Credit Rating Prediction

Enter client details to predict their credit rating (Good/Bad).

Status of Existing Checking Account

A12: 0 <= . . . < 200 DM

Credit History

A30: No credits taken / all paid back duly

Purpose

A40: Car (new)

Savings Account/Bonds

A61: < 100 DM

Present Employment Since

A71: Unemployed

Personal Status & Sex

A91: Male - Divorced/Separated

Other Debtors/Guarantors

A101: None

Property

A124: No property

Other Installment Plans

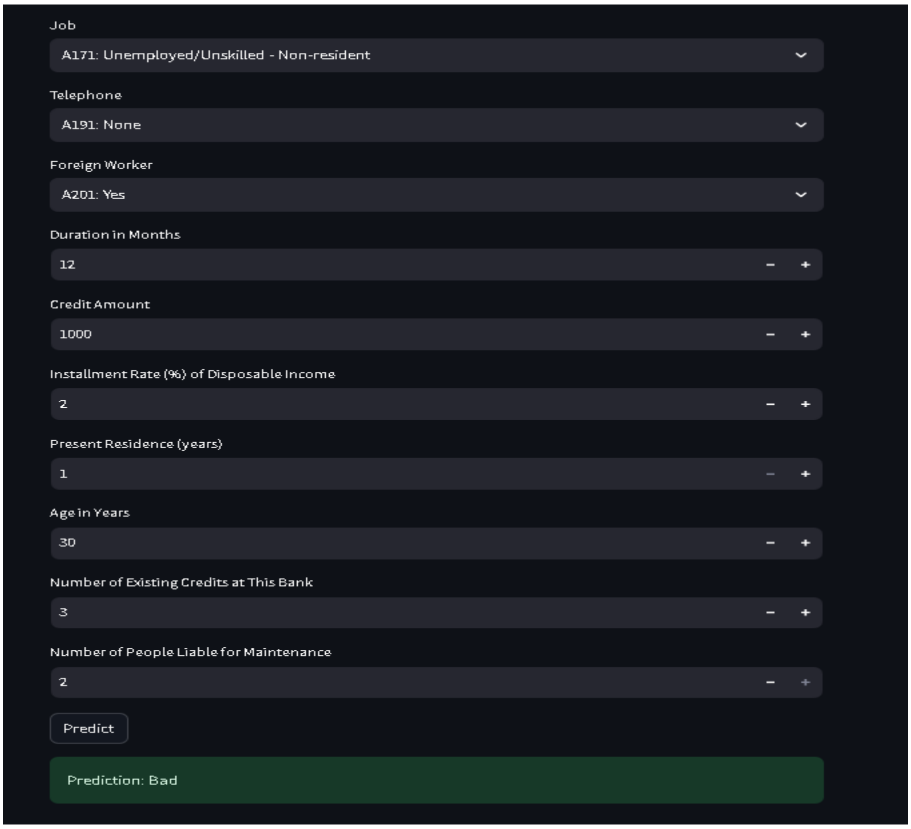
A141: Bank

Housing

A152: Own

Job

A171: Unemployed/Unskilled - Non-resident



Job

A171: Unemployed/Unskilled - Non-resident

Telephone

A191: None

Foreign Worker

A201: Yes

Duration in Months

12

Credit Amount

1000

Installment Rate (%) of Disposable Income

2

Present Residence (years)

1

Age in Years

30

Number of Existing Credits at This Bank

3

Number of People Liable for Maintenance

2

Prediction: Bad

Key Results & Inferences

- **Model Performance:**
  - Random Forest was the best-performing model, achieving a cross-validation accuracy of **76%** and test accuracy of **74%**.
  - Other models (e.g., Logistic Regression, Decision Tree, SVM) showed slightly lower accuracy, confirming Random Forest as the best candidate.
- **Data Drift Monitoring:**
  - **Numeric Drift:** No significant drift detected. P-values were well above **0.05**, indicating stability between training and production data.
  - **Categorical Drift:** Most features remained stable.

This project demonstrates a complete MLOps pipeline, from data validation and model training to deployment and monitoring. The results indicate that the deployed model is robust, with strong performance and minimal data drift in production.