SEAS 8515 - Data Engineering for AI
Project 4

Due Date: June 2, 2024 (12:00 noon EST)

Project: Data Engineering and Machine Learning with Spark and MLflow on Loan Dataset

Overview

This project involves developing end-to-end data analysis and machine learning models using the Apache Spark libraries. The loan dataset comprises comprehensive information on borrowers and their loan details, including unique identifiers for loans and members, loan amounts, funding details, and loan terms. It tracks key metrics such as interest rates, installment amounts, and borrower attributes like employment title, length, home ownership, and annual income. The dataset also includes loan statuses, payment plans, and detailed credit histories, capturing metrics like FICO scores, delinquency records, credit inquiries, and utilization rates. Additionally, it records financial hardships, debt settlements, and other critical metrics to assess creditworthiness and loan performance, offering a holistic view of the lending and borrowing landscape.

Part 1: Data Exploration and Analysis

1. Explore the dataset and create FOUR analysis questions of varying complexity. Use Spark DataFrame API and SparkSQL to answer them (any other visualization tools can be used). Ensure the questions cover different aspects of the data.

Part 2: Data Cleaning and Feature Engineering

1. Address any defects in the dataset, such as missing data, class imbalance, and outliers.

2. Calculate and visualize the correlation between different features.

Part 3: Model Development and Evaluation

1. Select relevant features for predicting loan status. You do not need to use all features.

2. Develop at least two classification models (e.g., logistic regression, random forest, xgboost etc) using Spark libraries.

3. Evaluate the models using metrics like accuracy, precision, recall, ROC-AUC, and F1-score.

4. Determine feature importance and derive insights.

5. Answer at least one of the following question:

   - Discuss any model governance and relevant quality checks that should be implemented before using machine learning for this dataset.
   - Use any explainable AI tool to achieve local interpretability of the complex model developed. Describe the importance of model interpretability in the financial data space.

Part 4: Experiment Tracking with MLflow

1. Use MLflow on Databricks to track parameters, metrics, versions, models, and artifacts.

2. Include screenshots of the MLflow dashboard showing experiment details, model metrics, and artifacts. Add this to the documentation (not slides).

Submission Requirements

1. **Documentation**: Include well-commented source code and comprehensive documentation explaining your process.

2. **Presentation**: Prepare a concise presentation (maximum 12 slides) summarizing key aspects and outcomes of the project.

Submission Instructions

- Submit your code with all scripts, notebooks, and documentation.

- Include the presentation slides as a separate file or link.

This project will provide hands-on experience in data exploration, cleaning, feature engineering, model development, evaluation, and experiment tracking using Spark and MLflow.

Description of Features

- **id**: Unique identifier for the loan.
- **member_id**: Unique identifier for the borrower.
- **loan_amnt**: The total amount of the loan applied for.
- **funded_amnt**: The total amount of the loan funded by investors.
- **funded_amnt_inv**: The total amount of the loan funded by individual investors.
- **term**: The term of the loan in months.
- **int_rate**: The interest rate of the loan.
- **installment**: The monthly payment amount owed by the borrower.
- **grade**: The loan grade assigned by the lending platform.
- **sub_grade**: The loan sub-grade assigned by the lending platform.
- **emp_title**: The job title of the borrower.
- **emp_length**: The length of employment of the borrower in years.
- **home_ownership**: The home ownership status of the borrower.
- **annual_inc**: The annual income of the borrower.
- **verification_status**: The verification status of the borrower's income.
- **issue_d**: The date the loan was issued.
- **loan_status**: The current status of the loan.
- **pymnt_plan**: Indicates if the borrower is on a payment plan.
- **url**: URL for the loan listing.
- **desc**: Description of the loan provided by the borrower.
- **purpose**: The purpose of the loan.

- **title**: The title of the loan listing.
- **zip_code**: The first 3 digits of the borrower's zip code.
- **addr_state**: The state of the borrower's address.
- **dti**: The debt-to-income ratio of the borrower.
- **delinq_2yrs**: The number of 30+ days past-due incidences in the borrower's credit file for the past 2 years.
- **earliest_cr_line**: The date the borrower's earliest reported credit line was opened.
- **fico_range_low**: The lower bound of the borrower's FICO score range.
- **fico_range_high**: The upper bound of the borrower's FICO score range.
- **inq_last_6mths**: The number of inquiries in the borrower's credit file in the last 6 months.
- **mths_since_last_delinq**: The number of months since the borrower's last delinquency.
- **mths_since_last_record**: The number of months since the last public record.
- **open_acc**: The number of open credit lines in the borrower's credit file.
- **pub_rec**: The number of derogatory public records.
- **revol_bal**: The total credit revolving balance.
- **revol_util**: The revolving line utilization rate.
- **total_acc**: The total number of credit lines currently in the borrower's credit file.
- **initial_list_status**: The initial listing status of the loan.
- **out_prncp**: The remaining outstanding principal for the loan.
- **out_prncp_inv**: The remaining outstanding principal for the loan funded by investors.
- **total_pymnt**: The total amount of payments received to date.
- **total_pymnt_inv**: The total amount of payments received to date for the investors.
- **total_rec_prncp**: The total principal received to date.
- **total_rec_int**: The total interest received to date.

- **total_rec_late_fee**: The total late fees received to date.
- **recoveries**: The post charge-off gross recovery.
- **collection_recovery_fee**: The fee for collecting the recovery.
- **last_pymnt_d**: The date of the last payment received.
- **last_pymnt_amnt**: The amount of the last payment received.
- **next_pymnt_d**: The date of the next payment due.
- **last_credit_pull_d**: The date the borrower's credit file was last pulled.
- **last_fico_range_high**: The upper bound of the last FICO score range.
- **last_fico_range_low**: The lower bound of the last FICO score range.
- **collections_12_mths_ex_med**: The number of collections in the last 12 months excluding medical collections.
- **mths_since_last_major_derog**: The number of months since the last major derogatory event.
- **policy_code**: The public policy code indicating which policies apply to the loan.
- **application_type**: Indicates whether the loan is an individual or joint application.
- **annual_inc_joint**: The combined annual income of co-borrowers for joint applications.
- **dti_joint**: The combined debt-to-income ratio for joint applications.
- **verification_status_joint**: The verification status of the combined income for joint applications.
- **acc_now_delinq**: The number of accounts currently delinquent.
- **tot_coll_amt**: The total collection amounts ever owed.
- **tot_cur_bal**: The total current balance of all accounts.
- **open_acc_6m**: The number of open accounts in the last 6 months.
- **open_act_il**: The number of currently active installment accounts.
- **open_il_12m**: The number of installment accounts opened in the last 12 months.
- **open_il_24m**: The number of installment accounts opened in the last 24 months.

- **mths_since_rcnt_il**: The number of months since the most recent installment account opened.
- **total_bal_il**: The total current balance of all installment accounts.
- **il_util**: The ratio of total current balance to the credit limit on installment accounts.
- **open_rv_12m**: The number of revolving credit accounts opened in the last 12 months.
- **open_rv_24m**: The number of revolving credit accounts opened in the last 24 months.
- **max_bal_bc**: The maximum current balance on a bankcard.
- **all_util**: The total utilization across all credit lines.
- **total_rev_hi_lim**: The total high credit/credit limit for all revolving accounts.
- **inq_fi**: The number of personal finance inquiries.
- **total_cu_tl**: The number of credit union trade lines.
- **inq_last_12m**: The number of credit inquiries in the last 12 months.
- **acc_open_past_24mths**: The number of accounts opened in the past 24 months.
- **avg_cur_bal**: The average current balance of all accounts.
- **bc_open_to_buy**: The total available credit on bankcards.
- **bc_util**: The utilization rate on bankcards.
- **chargeoff_within_12_mths**: The number of charge-offs within 12 months.
- **delinq_amnt**: The delinquent amount.
- **mo_sin_old_il_acct**: The months since the oldest installment account was opened.
- **mo_sin_old_rev_tl_op**: The months since the oldest revolving account was opened.
- **mo_sin_rcnt_rev_tl_op**: The months since the most recent revolving account was opened.
- **mo_sin_rcnt_tl**: The months since the most recent account was opened.
- **mort_acc**: The number of mortgage accounts.

- **mths_since_recent_bc**: The number of months since the most recent bankcard account was opened.

- **mths_since_recent_bc_dlq**: The number of months since the most recent bankcard delinquency.

- **mths_since_recent_inq**: The number of months since the most recent inquiry.

- **mths_since_recent_revol_delinq**: The number of months since the most recent revolving delinquency.

- **num_accts_ever_120_pd**: The number of accounts ever 120+ days past due.

- **num_actv_bc_tl**: The number of currently active bankcard accounts.

- **num_actv_rev_tl**: The number of currently active revolving accounts.

- **num_bc_sats**: The number of satisfactory bankcard accounts.

- **num_bc_tl**: The number of bankcard accounts.

- **num_il_tl**: The number of installment accounts.

- **num_op_rev_tl**: The number of open revolving accounts.

- **num_rev_accts**: The number of revolving accounts.

- **num_rev_tl_bal_gt_0**: The number of revolving accounts with a balance greater than zero.

- **num_sats**: The number of satisfactory accounts.

- **num_tl_120dpd_2m**: The number of accounts 120+ days past due in the last 2 months.

- **num_tl_30dpd**: The number of accounts 30+ days past due.

- **num_tl_90g_dpd_24m**: The number of accounts 90+ days past due in the last 24 months.

- **num_tl_op_past_12m**: The number of accounts opened in the past 12 months.

- **pct_tl_nvr_dlq**: The percentage of trade lines never delinquent.

- **percent_bc_gt_75**: The percentage of bankcards with a utilization rate greater than 75%.

- **pub_rec_bankruptcies**: The number of public record bankruptcies.

- **tax_liens**: The number of tax liens.
- **tot_hi_cred_lim**: The total high credit/credit limit.
- **total_bal_ex_mort**: The total balance excluding mortgages.
- **total_bc_limit**: The total bankcard credit limit.
- **total_il_high_credit_limit**: The total installment high credit limit.
- **revol_bal_joint**: The total revolving balance for joint applications.
- **sec_app_fico_range_low**: The lower bound of the FICO score range for the secondary applicant.
- **sec_app_fico_range_high**: The upper bound of the FICO score range for the secondary applicant.
- **sec_app_earliest_cr_line**: The earliest credit line opened for the secondary applicant.
- **sec_app_inq_last_6mths**: The number of inquiries in the credit file for the secondary applicant in the last 6 months.
- **sec_app_mort_acc**: The number of mortgage accounts for the secondary applicant.
- **sec_app_open_acc**: The number of open accounts for the secondary applicant.
- **sec_app_revol_util**: The revolving line utilization rate for the secondary applicant.
- **sec_app_open_act_il**: The number of currently active installment accounts for the secondary applicant.
- **sec_app_num_rev_accts**: The number of revolving accounts for the secondary applicant.
- **sec_app_chargeoff_within_12_mths**: The number of charge-offs within 12 months for the secondary applicant.
- **sec_app_collections_12_mths_ex_med**: The number of collections in the last 12 months excluding medical collections for the secondary applicant.
- **sec_app_mths_since_last_major_derog**: The number of months since the last major derogatory event for the secondary applicant.
- **hardship_flag**: Indicates if the borrower is under financial hardship.
- **hardship_type**: The type of hardship.

- **hardship_reason**: The reason for the hardship.
- **hardship_status**: The status of the hardship.
- **deferral_term**: The deferral term of the hardship.
- **hardship_amount**: The amount of the hardship.
- **hardship_start_date**: The start date of the hardship.
- **hardship_end_date**: The end date of the hardship.
- **payment_plan_start_date**: The start date of the payment plan.
- **hardship_length**: The length of the hardship.
- **hardship_dpd**: The number of days past due during the hardship.
- **hardship_loan_status**: The loan status during the hardship.
- **orig_projected_additional_accrued_interest**: The originally projected additional accrued interest.
- **hardship_payoff_balance_amount**: The payoff balance amount during the hardship.
- **hardship_last_payment_amount**: The last payment amount during the hardship.
- **disbursement_method**: The method of loan disbursement.
- **debt_settlement_flag**: Indicates if the borrower is in debt settlement.
- **debt_settlement_flag_date**: The date the debt settlement flag was set.
- **settlement_status**: The status of the debt settlement.
- **settlement_date**: The date of the debt settlement.
- **settlement_amount**: The amount agreed upon for the debt settlement.
- **settlement_percentage**: The percentage of the original balance paid in the debt settlement.
- **settlement_term**: The terms of the debt settlement.

Grading Rubric and Criteria

Criteria	Excellent (90-100)	Good (80-89)	Satisfactory (70-79)	Needs Improvement (0-69)
Design and Planning	Comprehensive design with clear understanding of structured data processing and machine learning principles.	Solid design with good understanding of concepts.	Basic design with some gaps in understanding structured data.	Inadequate design, lacking comprehension of structured data processing.
Effective Use of Spark MLlib and Data Processing Skills	Extensive use of Spark MLlib with machine learning principles, and excellent data processing and analysis using Apache Spark and SQL.	Good use of Spark MLlib and understanding of machine learning principles, with competent use of Spark APIs.	Basic use of Spark MLlib and adequate data processing skills with limited application of advanced features.	Poor utilization of Spark MLlib, ineffective data processing, and little to no application of machine learning principles.
Code Quality and Implementation	High-quality code, efficient, well-organized, and fully integrated.	Good coding with minor issues in efficiency or organization.	Functional code but lacking in efficiency or organization.	Poorly written or non-functional code, major integration issues.
Scalability and Reliability	Excellent scalability, handling large data volumes efficiently.	Good scalability for moderate data volumes.	Sufficient for small to moderate data volumes, some scalability issues.	Poor scalability, struggles with large data volumes.
Documentation and Presentation	Comprehensive and clear documentation and presentation.	Good documentation and effective presentation.	Basic documentation and presentation, lacking some clarity.	Poor documentation and unclear presentation.