CSCI316 Big Data Mining Techniques and Implementation

Project Specification Autumn 2023

You are required to prepare a set of presentation slides which must include (1) the full name and student number of each student in the group, the contribution (in percent) of each group member, (2) a description of the task, (3) description of your approach and how the methodology was implemented; (4) the strengths and weaknesses of the approach or implementation; (5) your results and an analysis of the results; (6) a brief summary and a conclusion. The summary should state new and interesting things that you learned and discovered while working on this project. The conclusion should summarize your main findings and statements about possible future work (e.g., how you plan to improve your models and approach in future).

Below is the recommended structure of your slides:

- Introduction (define the problem and the goal)
- Methods (approach, and discussion of strengths and weaknesses)
- Implementation (methods, key-issues, how these were addressed and sample codes)
- Results (include illustrative Figures and Tables and explanations)
- Discussion and Conclusions

The Data Set

The project uses the loan data set for credit risk analysis. The data set is available in the following link.
(https://www.kaggle.com/datasets/rameshmehta/credit-risk-analysis)
This data set has different types of features such as categorical, numeric & date. The target variable is the default (index). In financing, a default can occur when a borrower is unable to make timely payments, misses payments, avoids or stops making payments. An explanation of the features in the appendix of this document.

The Task

Definition of the task:
You are to implement an end-to-end data mining project to analyse the provided dataset. The objective is to implement a workflow to predict the targe variable of the data (i.e., classification or regression). This workflow must include two stages, as illustrated in Figure 1.

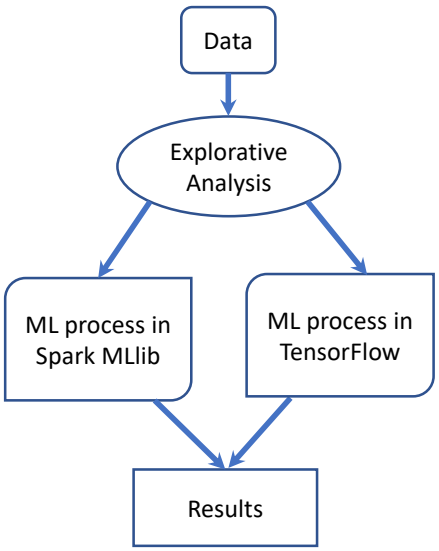

Figure 1. The workflow stages

Stage1: Data exploration

This stage includes the use of Spark's **DataFrame**, **Pandas API on Spark** or **RDD** APIs in Python to explore the data. Understand the dataset by querying a few important statistic measures of the data. Visualise the data and explain your findings. It is important that you demonstrate an in-depth understanding on the data that you are analysing. (Note. You *cannot* use Pandas and Scikit-Learn in this stage.)

Stage2: Predictive analysis

This stage includes two machine learning (ML) processes. Each process must include ***at least three*** kinds of ML models (such as decision tree, random forest, naïve Bayesian, feedforward network, etc.). The models must be evaluated with common metrics (such as accuracy, precision, recall and ROC).

Specific requirements of Process One:

- This process is built with the ML library of Spark (i.e., pyspark.ml and pyspark.mllib)

Specific requirements of Process Two:

- This process is built with TensorFlow and Keras.

In your slides, you must explain the detailed pipeline design and evaluation outcomes, as well as any other interesting findings or lessons learned. Any claim that you make in the slides must be supported by the implementation in your submitted Python source codes.

The requirements of deliverables are in the Introduction document.

Appendix (data dictionary)

LoanStatNew	Description
addr_state	The state provided by the borrower in the loan application
annual_inc	The self-reported annual income provided by the borrower during registration.
annual_inc_joint	The combined self-reported annual income provided by the co-borrowers during r egistration
application_type	Indicates whether the loan is an individual application or a joint application with t wo co-borrowers
collection_recover y_fee	post charge off collection fee
collections_12_mt hs_ex_med	Number of collections in 12 months excluding medical collections
delinq_2yrs	The number of 30+ days past-due incidences of delinquency in the borrower's cre dit file for the past 2 years
desc	Loan description provided by the borrower
dti	A ratio calculated using the borrower＊s total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the b orrower＊s self-reported monthly income.
dti_joint	A ratio calculated using the co-borrowers' total monthly payments on the total de bt obligations, excluding mortgages and the requested LC loan, divided by the co-b orrowers' combined self-reported monthly income
earliest_cr_line	The month the borrower's earliest reported credit line was opened
emp_length	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
emp_title	The job title supplied by the Borrower when applying for the loan.*
fico_range_high	The upper boundary range the borrower＊s FICO at loan origination belongs to.
fico_range_low	The lower boundary range the borrower＊s FICO at loan origination belongs to.
funded_amnt	The total amount committed to that loan at that point in time.

LoanStatNew	Description
funded_amnt_inv	The total amount committed by investors for that loan at that point in time.
grade	LC assigned loan grade
home_ownership	The home ownership status provided by the borrower during registration. Our val ues are: RENT, OWN, MORTGAGE, OTHER.
id	A unique LC assigned ID for the loan listing.
initial_list_status	The initial listing status of the loan. Possible values are W, F
inq_last_6mths	The number of inquiries in past 6 months (excluding auto and mortgage inquiries)
installment	The monthly payment owed by the borrower if the loan originates.
int_rate	Interest Rate on the loan
is_inc_v	Indicates if income was verified by LC, not verified, or if the income source was ve rified
issue_d	The month which the loan was funded
last_credit_pull_d	The most recent month LC pulled credit for this loan
last_fico_range_hi gh	The upper boundary range the borrower ∗s last FICO pulled belongs to.
last_fico_range_lo w	The lower boundary range the borrower ∗s last FICO pulled belongs to.
last_pymnt_amnt	Last total payment amount received
last_pymnt_d	Last month payment was received
loan_amnt	The listed amount of the loan applied for by the borrower. If at some point in tim e, the credit department reduces the loan amount, then it will be reflected in this value.
loan_status	Current status of the loan

LoanStatNew	Description
member_id	A unique LC assigned Id for the borrower member.
mths_since_last_delinq	The number of months since the borrower's last delinquency.
mths_since_last_major_derog	Months since most recent 90-day or worse rating
mths_since_last_record	The number of months since the last public record.
next_pymnt_d	Next scheduled payment date
open_acc	The number of open credit lines in the borrower's credit file.
out_prncp	Remaining outstanding principal for total amount funded
out_prncp_inv	Remaining outstanding principal for portion of total amount funded by investors
policy_code	publicly available policy_code=1new products not publicly available policy_code=2
pub_rec	Number of derogatory public records
purpose	A category provided by the borrower for the loan request.
pymnt_plan	Indicates if a payment plan has been put in place for the loan
recoveries	post charge off gross recovery
revol_bal	Total credit revolving balance
revol_util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
sub_grade	LC assigned loan subgrade
term	The number of payments on the loan. Values are in months and can be either 36 or 60.
title	The loan title provided by the borrower

LoanStatNew	Description
total_acc	The total number of credit lines currently in the borrower's credit file
total_pymnt	Payments received to date for total amount funded
total_pymnt_inv	Payments received to date for portion of total amount funded by investors
total_rec_int	Interest received to date
total_rec_late_fee	Late fees received to date
total_rec_prncp	Principal received to date
url	URL for the LC page with listing data.
verified_status_joint	Indicates if the co-borrowers' joint income was verified by LC, not verified, or if the income source was verified
zip_code	The first 3 numbers of the zip code provided by the borrower in the loan application.
open_acc_6m	Number of open trades in last 6 months
open_il_6m	Number of currently active installment trades
open_il_12m	Number of installment accounts opened in past 12 months
open_il_24m	Number of installment accounts opened in past 24 months
mths_since_rcnt_il	Months since most recent installment accounts opened
total_bal_il	Total current balance of all installment accounts
il_util	Ratio of total current balance to high credit/credit limit on all install acct
open_rv_12m	Number of revolving trades opened in past 12 months
open_rv_24m	Number of revolving trades opened in past 24 months

LoanStatNew	Description
max_bal_bc	Maximum current balance owed on all revolving accounts
all_util	Balance to credit limit on all trades
total_rev_hi_lim ?	Total revolving high credit/credit limit
inq_fi	Number of personal finance inquiries
total_cu_tl	Number of finance trades
inq_last_12m	Number of credit inquiries in past 12 months
acc_now_delinq	The number of accounts on which the borrower is now delinquent.
tot_coll_amt	Total collection amounts ever owed
tot_cur_bal	Total current balance of all accounts

---The End---