NOVA IMS Information Management School

CREDIT RISK MODELING

GROUP PROJECT DESCRIPTION

Afshin Ashofteh; PhD, PGDip, MBA, MSc

E-mail | Academic homepage | Nova Research | ResearchGate | Kaggle | Google Scholar | Discussion Group

TABLE OF CONTENTS

Contents

PROJECT DESCRIPTION

Project Description

A. PURPOSES OF THE PROJECT

The purpose of the project is to practice with the course material and to strengthen your skills of problem formulation and solution, research, cooperation with others, and professional oral and written communication. Every member is expected to carry an equal share of the group's workload. As such, it is in your interest to be involved in all aspects of the project. Even if you divide the work rather than work on each piece together, you are still responsible for each part. The group project will be graded as a whole: its different components will not be graded separately. Your exams may contain questions that are based on aspects of your group projects. It is recommended that each group establish ground rules early in the process to facilitate joint work including a problem-solving process for handling conflicts. In the infrequent case where you believe that a group member is not carrying out his or her fair share of work, you are urged not to permit problems to develop to a point where they become serious. If you cannot resolve conflicts internally after your best efforts, they should be brought to our attention, and we will work with you to find a resolution. You may be asked to complete a peer evaluation form to evaluate the contribution of each of your group members (including your own contribution) at the conclusion of each project. If there is a consensus that a group member did not contribute a fair share of work to the project, we will consider this feedback during grading.

B. TASKS

- Make a preliminary statistical analysis of the credit dataset.
- Develop a logistic regression model to predict the probability of default.
- Develop a machine learning model to predict the probability of default.
- Develop a deep learning model to predict the probability of default.
- Evaluate the models results (predictive power)

You are free to select the appropriate software (R or Python) for making the calculations & preparing the final report.

C. GROUP SIZE

Standard (and recommended) group size is 4 or 5. You will organize your own groups and register it in Moodle (Group Choice).

PROJECT DESCRIPTION

D. SUPPORT MATERIALS

1-As a support material, read the following article which was presented in 2018 but is published in 2023 by Springer (proceeding publication was postponed because of COVID-19), Big Data for Credit Risk Analysis: Efficient Machine Learning Models Using PySpark - 2023

https://www.researchgate.net/publication/374860505_Big_Data_for_Credit_Risk_Analysis_Efficient_Machine_Learning_Models_Using_PySpark

2-Joining discussion groups is very important in learning process. Review Q/As of the following link:

ML Credit Risk Prediction FALL22 | Kaggle

E. PROJECT MILESTONES/REPORTS

- All figures and tables should be numbered and have descriptive captions. All figures should be referred to by number in the text, and any items listed in a bibliography or reference section should be explicitly referred to in the text. Failure to cite and reference properly may result in accusations of plagiarism. Please use the **Harvard Referencing System**.
- A single written numbered technical report (maximum 10 pages) and your code file must be submitted by December 23 in PDF format in Moodle (Report Delivery). In the case of any difficulty contact me by Email.

F. EVALUATION - GRADING CRITERIA

- Data (if different from original one by sampling etc.), Code (in Kaggle) and report (PDF file in Moodle) are delivered before deadline.
- Primary Statistical Analysis
  - Levels of the response variable are defined.
  - Outlier detection and treatment of missing values are done.
- logistic regression model, Machine Learning model, Deep Learning model
  - The full model is fitted, and the Final model is developed properly.
  - Variable Selection is done.
  - New variables and measures are created.
- Model Evaluation

  - Train and Test data are allocated properly.
  - Confusion Matrix and AUC are presented and discussed.
- Summary Conclusion
  - Quality of Outputs, Tables and Charts are good.
  - Summary and conclusion are available and reasonable.
- Figs and Tables
  - Figs and Tables have captions with description.
  - Figs and Tables are referenced in the text.
- References
  - References are mentioned with Harvard Referencing System.
- Report
  - The coherence and cohesion of the report is good.
  - The description of analysis outcomes are correct and complete.
- Discussion Group (Optional)
  - Group members activities in sharing knowledge, asking and replying to questions in this discussion group of the Kaggle competition.

CRITERIA CHECKLIST

The project will be graded according to:

- Topic In-depth Analysis, Real Potential for a Contribution and Originality: 40%
- Technical Merit and Total Mass of the Work Done: 40%
- Written Report (Clarity, Completeness, Professional Polish, and General Effectiveness) 20%

PROJECT DESCRIPTION

CHECKLIST

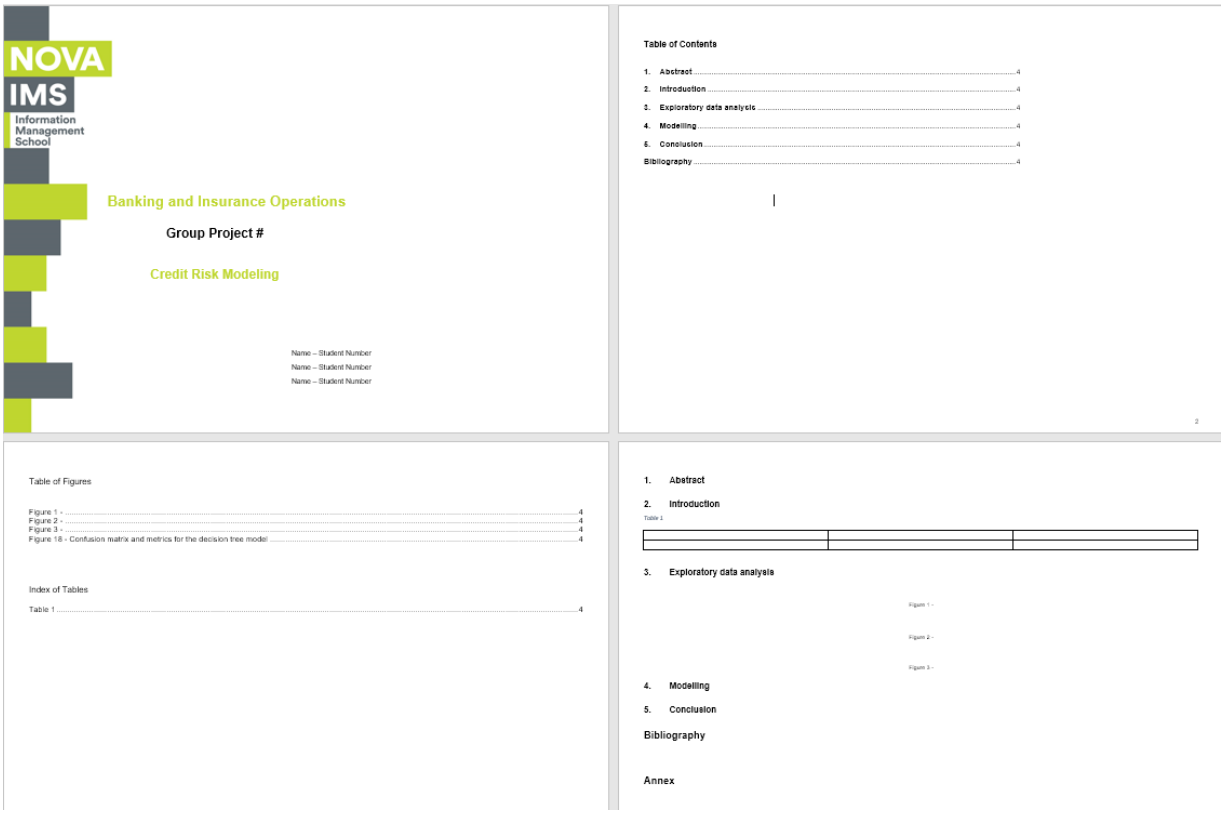
TITLE	DESCRIPTION	YES	NO
Delivery	Meet the deadline		
	Language of report in English		
	Final Clean Data is attached		
	Software code is attached		
	The report is attached both in Word and Pdf format		
	The report has complete group information		
Primary Statistical Analysis	Response variable (Label Variable) is reproduced in binary format properly		
	Outliers are treated		
	Missing values are imputed		
	Feature engineering is done		
	The full model is fitted and described		
Model Evaluation	Train, Validation and Test datasets are chosen properly		
	The final model is developed based on appropriate features		
	The Confusion matrix is presented		
	AUC is presented		
Outputs, Tables and Figures	The quality of Outputs, Tables, and Charts is good and they are meaningful		
	Figures and Tables have captions		
	Outputs, Tables, and Charts are referenced in the text		
Summary Conclusion	The summary and conclusion section exist at the end of the report		
References	Harvard Referencing System is obeyed in the reference section		
Extra activity	Some other optional algorithms are checked for solving this problem		

Analytical Base Table

You are provided with a consumer loan dataset (dataset.csv) which contains more than One million rows and 28 columns (variables):

Attribute	Description
Id	A unique LC assigned ID for the loan listing.
loan_amnt	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
funded_amnt	The total amount committed to that loan at that point in time.
funded_amnt_inv	The total amount committed by investors for that loan at that point in time.
term	The number of payments on the loan. Values are in months and can be either 36 or 60.
int_rate	Interest Rate on the loan
installment	The monthly payment owed by the borrower if the loan originates.
grade	LC assigned loan grade
emp_title	The job title supplied by the Borrower when applying for the loan.
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.
home_ownership	The home ownership status provided by the borrower during registration. Our values are: RENT, OWN, MORTGAGE, OTHER.
annual_inc	The self-reported annual income provided by the borrower during registration.
verification_status	Indicates if income was verified by LC, not verified, or if the income source was verified
issue_d	The month which the loan was funded
loan_status	Current status of the loan
purpose	A category provided by the borrower for the loan request.
addr_state	The state provided by the borrower in the loan application
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 borrower's self-reported monthly income.
delinq_2yrs	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years
earliest_cr_line	The month the borrower's earliest reported credit line was opened
inq_last_6mths	The number of inquiries in past 6 months (excluding auto and mortgage inquiries)
open_acc	The number of open credit lines in the borrower's credit file.
pub_rec	Number of derogatory public records
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.
total_acc	The total number of credit lines currently in the borrower's credit file
out_prncp	The remaining outstanding principal for the total amount funded.
total_pymnt	Payments received to date for total amount funded

Report template



Check the template of the report, which is provided on the Moodle. It is only a suggestion, and you might like your creativity to make a better one or add more sections and subsections.