Introduction to Machine Learning

First Semster Test
4th year Statistics and Data Science

Ayoub Asri

18 December 2025

First Semester Test

Context

In many emerging economies, small and mid-sized businesses form the backbone of local commerce and employment. However, their stability is often precarious. Conventional financial metrics like sales or income fail to account for their overall robustness, preparedness for unexpected difficulties, or their integration into supportive financial networks. These businesses frequently operate with thin margins, face difficulties securing loans, and are highly sensitive to personal, economic, or environmental disruptions.

To address this gap, we propose the development of a multidimensional Business Vitality Score (BVS). This indicator moves beyond simple profitability to evaluate a firm's strength across several interconnected areas: its cushion of reserves and tangible resources, its responsible management of existing obligations, its capacity to withstand disruptions, and its ability to obtain and use formal financial tools.

The core task involves creating predictive models that can assign this score—categorized as Fragile, Stable, or Robust—using a variety of anonymized operational and demographic data. This information may include a company's primary trade sectors, involvement in cross-border trade, owner characteristics, workforce size, and regional operating environment. The underlying data is drawn from a sample of businesses operating in several coutnries.

The implications of such a tool are significant. For lenders, it provides a more nuanced lens for risk evaluation. For policymakers and support organizations, it enables the identification of enterprises in need of specific interventions, fostering more targeted and inclusive economic strategies.

Ultimately, this effort seeks to reframe the conversation around small business success from one focused purely on earnings to one that equally values endurance and opportunity. By creating models to diagnose a business's current vitality, we contribute to building analytical frameworks that can help these enterprises secure a more prosperous and resilient future.

Data Description

The file "LOS.csv" contain the dataset and the variables used are :

ID: Unique identifier for each business record

country: Country where the business is located (A/B/C/D)

owner_age: Age of the business owner in years

attitude_stable_business_environment: Owner attitude: Country will have a stable business environment in the future

attitude_worried_shutdown: Owner attitude: Worried that the business will shut down

compliance_income_tax: Business complies with or acts in accordance with income tax regulations

perception_insurance_doesnt_cover_losses: Owner perception: Insurance does not cover the kinds of losses the business suffers

perception_cannot_afford_insurance: Owner perception: Cannot afford insurance payments

personal_income: Total monthly personal income of the owner before tax and other deductions

business_expenses: Approximate monthly or annual expenses of the business in local currency

business_turnover: Approximate annual turnover/revenue of the business in local currency

business_age_years: Number of years the business owner has been running this business

motor_vehicle_insurance: Business has or uses motor vehicle insurance

has_mobile_money: Business uses mobile money account

current_problem_cash_flow: Currently facing cash flow problems in business operations

has_cellphone: Business has access to or uses a cell phone

owner_sex: Gender/sex of the business owner

offers_credit_to_customers: Whether the business offers goods or services on credit to customers

attitude_satisfied_with_achievement: Owner attitude: Satisfied with what has been achieved so far in the business

has_credit_card: Business uses credit card for business purposes

keeps_financial_records: Whether the business keeps financial records

perception_insurance_companies_dont_insure_businesses_like_yours: Owner perception: Insurance companies do not insure businesses like this one

perception_insurance_important: Owner perception: Insurance is important for the business

has_insurance: Whether the business has any kind of insurance

covid_essential_service: Whether the business was considered an essential service provider during COVID-19

attitude_more_successful_next_year: Owner attitude: Believes the business will be more successful in the next year

problem_sourcing_money: Faced problem with sourcing money when starting or taking over the business

marketing_word_of_mouth: Business uses word of mouth as a method of marketing or advertising

has_loan_account: Business has a loan account or short-term loan from formal financial institution

has_internet_banking: Business uses internet banking services

has_debit_card: Business uses bank debit card (e.g. Visa Electron)

future_risk_theft_stock: Business is likely to face risk of theft of business stock in the future

business_age_months: Number of additional months (beyond full years) the business has been operating

medical_insurance: Business has or uses medical aid or medical scheme insurance

funeral_insurance: Business has or uses funeral plan or cover insurance

motivation_make_more_money: Motivation to start the business: To make more money or provide for family

uses_friends_family_savings: Business uses informal financial product: friends and family savings or borrowing

uses_informal_lender: Business borrows from informal money lender

Target: Target variable: Financial Health Index score or classification of the business

Questions

Q1. Read the file "fin_health.csv" and inspect the dataset.

Q2. perform some descriptive analysis on the different variables of the dataset and study their relationship to the target variable

Q3. split the data (20% for test) use a seed of 2026 (i.e., run the code below before using the function that splits the data)

```
set.seed(2026)
```

Q4. Propose any recipe and justify the use of every step function

Q5. Fit any number of different models from your choosing. (Boosting are acceptable)

Q6. Justify the choice of hyperparameters for each model proposed.

Q7. Inspect the results of the model on testing set? what are the metric you used ? justify !

Q8. plot the ROC curve ? what is the value of the AUC ?

Q9. Rank the model proposed from best to worst. Use any ploting tool to precise the difference between these models.

Q10. Take the best 03 models and tune the values of all their hyperparameters using a different Grid method for each model.

Q11. Is there any change in the ranking after the tuning.

Q12. Take the best model from question 11 and use all iterative tuning methods.

Q13. Precise the metrics of the best model on testing set.

Q14. Study the metrics of each category and plot the ROC curve.

Final Considerations

- your work must be completed before deadline. Work must be sent before Saturday January 3rd 2026 23:59:59.
- Students are asked to

1. Send a code file
2. Send a report file
3. The code (with results if possible) and the report must be sent as separate files. All files accepted are .r .rmd .docx or .pdf (and possible .ipynb if the jupyter notebook contains R code)
4. The code must **not** be commented at any level, and can be accompanied with rmarkdown text explain some novel ideas (if applicable).
5. The report must be different than a rendering of an rmarkdown document. The report must explain the problem, thought process, solutions and potential innovations proposed. The report may contain some summary results or plots from your results.

- Students **can not use external data** related to this dataset.
- Any work sent after the deadline or outside of the section of our Moodle's course will not be considered for grading.

Good Luck.