Term Project Problem Statement and Data Descriptions

Sept. 8, 2022

To help you gain experience working with real-world data sets, this project will require that you mine actual transaction data to solve a problem of interest to your client, Banca Massiccia. a large Italian bank that is interested in understanding how reduce its portfolio default rate on individual loans. *In particular, the firm wishes to get better predictions of the one-year probability of default (PD) for a prospective borrower.*[1]

Context overview

Imagine that you have been approached by Banca Massiccia, a large Italian bank. Banca Massiccia has been making loans to businesses for many years, and has experimented with different approaches to underwriting these loans, including statistical models of default. The Head of Loan Origination has asked you to assist the firm in optimizing its underwriting using the power of machine learning. The goal is to produce estimates of the probability of default (PD) for prospective borrowers so the bank can use risk-based pricing to set interest rates and underwriting fees for borrowers. One way to do this is to leverage the data that Banca Massiccia has accumulated over the past several years.

The basics of a transaction flow is as follows:

1. A potential borrower applies for a loan by providing information about the finances of the firm in the form of fionancial statement data.
2. The loan officer assigned to the borrower analyzes the financial data using a number of human and automated approaches to try to determine how likely it is that the borrower will default.
3. The bank officer then determines what the appropriate interest rate and underwriting fees would be for the loan to be profitable, in consideration of the default probability.
4. If the bank officer concludes that the loan may be made at a rate that is within the bank's guidelines, the loan is made, and the customer's information is transferred to the loan monitoring group, which continues to monitor both changes in the financial status of the borrower, and the payment status of the loan.
5. If the borrower defaults before the loan is repaid, the borrower's case is transferred to the "work-out" group to try to recover the shortfall through various legal remedies.

Specific Business need: Predicting Probability of Default

Banca Massiccia would like to be able to better predict the probability that a potential borrower will default on a principal or interest payment for a prospective loan over the next 12 months.

The bank has asked you to design the data mining task, mine the data, and describe your results. (See the *Term Project Instructions* documents for the details of the deliverable and documentation.)

[1] This data and problem are from real-world Italian companies. The data is real-world data. However, some of the details of the data have been modified for confidentiality.

Data Overview

The IT department at Banca Massiccia has agreed to provide you with information on previous borrowers, including information about the company type, industry, etc. as well as financial statement data on an annual basis for each borrower. Because the firm is located in the EU, where privacy and data sharing is heavily regulated, some of the data that Banca Massiccia collects may not be shared with your team. The Data Dictionary section, below, provides the definitions of the variables in the training data set.

Data Conventions

The data in both the training and holdout samples will have identical structures and will conform with the following conventions:

- Each row is one firm-year
- Annual observations
- 44 variables in total
- All quantities (except ratios) are reported n €
- Only firms with > €1.5MM in assets are included in data
- Only non-finance/insurance firms are included
- There is no information in the training data regarding defaults that occurred after 12/31/2012
- The value for the default date (`def_date`) is `NA` in the holdout sample

The holdout sample is drawn from a future time period and will include both firms that in the training data and those that are not.

The training and holdout samples will have identical structures and will conform with the following conventions:

A data dictionary is given on the next page.

`train.csv` (financial statements and default dates behavior)

Variable name (column name, feature, etc.)	Description
id	Firm identifier
HQ_city	City of main branch
legal_struct	Legal structure of the firm
ateco_sector	Industry sector code (see ATECO sector definition doc)
fs_year	Year of the financial statement
asst_intang_fixed	Intangible assets
asst_tang_fixed	Tangible Assets
asst_fixed_fin	Financial assets
asst_current	Current assets
AR	Accounts receivable
cash_and_equiv	Cash & equivalent holdings
asst_tot	Total assets
eqyt_tot	Total equity
eqty_corp_family_tot	Total equity for entire group ("family")
liab_lt	Long-term liabilities
liab_lt_emp	Long-term liab to employees
debt_bank_st	Short-term bank debt
debt_bank_lt	Long-term bank debt
debt_fin_st	Short-term debt other
debt_fin_lt	Long-term debt other
AP_st	Short-term accounts payable
AP_lt	Long-term accounts payable
debt_st	Short-term debt
debt_lt	Long-term debt other
rev_operations	Operating revenue
COGS	COGS (Cost of goods sold)
prof_operating	Operating profit
goodwill	Goodwill
inc_financing	Financial income
exp_financing	Financial expenses
prof_financing	Financial profit
inc_extraord	Extraordinary income
taxes	Taxes
profit	Net profit
days_rec	Days recievables
ebitda	Earnings before interest, taxes, depreciation, and amortization
roa	Return on assets
roe	Return on equity
wc_net	Net working capital
margin_fin	(Equity - Fixed assets)
cf_operations	Operating cashflow