DSGA 3001

Modeling proposal

Group Amber

1. Exact (business problem)

Predict the probability that a potential borrower will default on a principal or interest payment for a prospective loan over the next 12 months based on the data available up to the statement date (after adjusting the timeline on the basis of the firm year (typically ending in March/April) instead of the fiscal year (January 1 - December 31))

1. Use scenario
   1. Reduce Banca Massiccia’s 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 to reduce risks and increase profits.
   2. Use the probability of borrower default to better calculate appropriate rates of interest and underwriting fees, adjust for the associated risk (based on the information of the borrower), and confirm if the loan is feasible according to the projected rate of interest being within allowed government guidelines.
2. Related data mining problem
   1. Supervised or unsupervised?
      1. We are predicting the value of a specific, labeled feature (i.e. the probability of default) with the help of ML methods that also incorporate the results of other existing models.
      2. Data is labeled (target variable is available) and we will be building a model that utilizes known statistical models. This cannot be done without human intervention, so **supervised learning** is preferred.
   2. Unit of analysis?
      1. Unit of analysis is the data for the individual borrowers (companies) due to each of them **having a default risk** associated with them.
   3. Potential target variables?
      1. Self-created feature ‘Probability of Default’ constructed using def\_date, based on our definition of default (12 months).
      2. If 12/31/99, firm has not defaulted
      3. Predicted probability of default will be calculated using different statistical and machine learning methods and comparing these values.
   4. Potential features
      1. Independent variables

Sorting the independent variables from the dependent variables. Some of the variable correlations are straightforward such as assets (total, tangible, fixed, current, intangible) and some will have hidden correlations or dependencies like industry type, location, etc..

* + 1. Removing the **long-term variables**

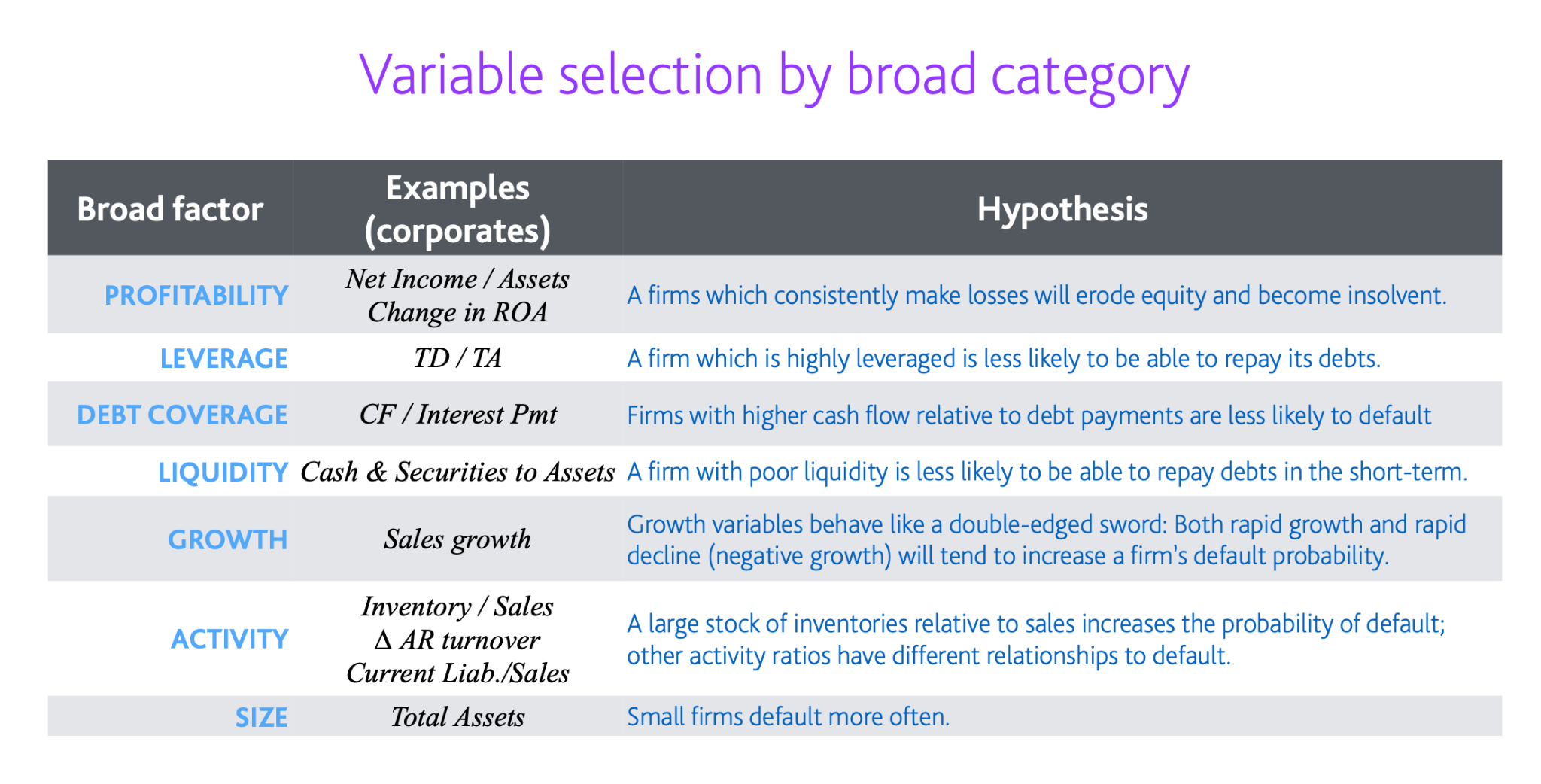
The long-term features such as long-term bank debts, liability, etc. will not have an impact on the probability of default within the next 12 months. So we can put them aside for now.

* + 1. Grouping by

Some features such as the financial structure of the firm, city, industry sector, etc. can be used to infer useful information about the probability of default by **aggregating individual entries** on these features.

* + 1. Self-defined variables

Some variables such as **profitability, leverage, liquidity, activity, etc.** can be calculated based on the independent variables and may have a powerful impact on the final calculations.



* 1. How do the results solve the business problem?

Our ‘result’ is the predicted probability of default with which we can adjust the rate of interest (based on the default risk) in order to filter out the ‘good’ and ‘bad’ candidates and thus, **reduce our portfolio default rate** (original business problem).

1. Our (tentative) approach:
   1. Use commonly available statistical models for predicting the probability of default (such as structural models like the Merton model, reduced form models such as the Duffie and Singleton model, etc.) to obtain neighborhood probability values and use these PD values as additional features to our own data-driven Neural Network. We will also use the commonly available statistical model results as a baseline for comparison.
   2. We can model the number of defaults with time, to see if there is a cascading effect in the loan market (some companies defaulting may result in others defaulting as well)

Some of the widely used statistical approaches that can be used as baselines: 