Probability of default model enrichment with macroeconomic and industry data: a start-up case

MS in Finance Capstone Project, June 2020

Public Project Summary

Central European University

Project Overview

The project was a collaboration with a Hungarian start-up focused on invoice financing for small and medium-sized enterprises (SME-s). The start-up's vision is to enable SME-s to increase their liquidity at the ease of clicking a button, at low cost, full transparency, completely online. The probability of default (PD) model developed so far needs further improvements – adding macroeconomic and industry data to the already collected company financial data in order to increase model accuracy.

The start-up currently operates in Hungary, but plan to expand internationally in the future.

Project Objectives

The objectives of the project are:

1. Compile a list of macroeconomic indicators and sources to be used
2. Download macroeconomic data
3. Feature engineer industry variables
4. Clean and prepare all data for modelling

All parts of the project are executed in line with the processes and systems used by the start-up. All analysis was done in Python.

In addition, I consulted the business on how to better use and apply the Agile methodology for project management.

Macroeconomic data

The project started with academic research of papers about PD models with macro data from other European countries. After I gathered some ideas of potential indicators, I started looking through online sources that we can get the data from. The main idea that I had was to **get the data in an automatic way** into Python rather than download it manually. Also, my focus was on **reusability** and **easy replication** of the method used to other countries and periods of time.

After a thorough research of the available sources and their advantages and disadvantages, we decided to use the **World Bank** database which offers a wide range of indicators in all areas of life (not only economics and finance, but also environment, public and private sector, education, etc.)

The World Bank collects statistical data for almost all countries around the world and there is a readily available Python package for getting the data from World Bank database directly into Python (pandas-datareader). The user can choose the period over which data will be downloaded, the granularity (yearly, monthly or quarterly), which specific indicators to download, and the country/countries of interest.

In this specific project, a list of 790 macroeconomic indicators for Hungary were downloaded over the period 2004 – 2018.

Industry data

Due to time constraints, the capstone project included industry feature generation based on the company's internal database of available companies, rather than the usage of an external data source.

All companies in the database were grouped by industry using their respective NACE code. NACE (Nomenclature of Economic Activities) is a European industry standard system for classifying business activities. The final industry features were obtained as ratios between the specific data point and the respective industry average.

Industry feature generation produced a list of 153 indicators.

Data cleaning/preparation

The next step in the process was cleaning the collected data. The focus was on missing values.

The strategy followed for **macroeconomic indicators** was removing any indicator with missing data for the recent years or more than 20% missing values in the time series. For the indicators that are left, any missing values that remain are filled out by inferring the mean from the rest of the time series. 427 macro indicators are kept after data cleaning.

As for the **industry indicators**, a similar strategy is followed. The difference here is that inference is performed by industry groups, rather than on the entire feature. Similarly, many of the variables are dropped during the data cleaning process and only 42 variables are left at the end.

Principal Component Analysis (PCA)

Both the macroeconomic and industry feature generation processes produced long lists of indicators with possibly strongly correlated variables. To make the PD modelling process easier, I suggested performing a Principal Component Analysis in order to reduce the dimensionality of the data.

PCA is a common machine learning algorithm for data transformation in the pre-modelling phase. By performing orthogonal projections of the data vectors and solving an eigenvalue/eigenvector problem, PCA creates new uncorrelated variables (principal components) that successively maximize the explained variance in the dataset. Those new variables can be directly used as

independent variables in a standard linear/logistic model. Ultimately, PCA increases interpretability of the data but at the same time minimizes information loss.

The first step of the PCA for both the macroeconomic and industry indicators was to standardize the data around mean = 0 and standard deviation = 1. This ensures that all variables follow the standard normal distribution and is a prerequisite step to any PCA.

The maximum number of principal components is equal to the minimum between the number of observations and the number of variables in a dataset. For the dataset with macroeconomic variables, this makes a maximum of 15 PC-s (15 years of data), while for the industry variables, the maximum number of PC-s is 42 (number of features in the final dataset after cleaning).

For macroeconomic indicators, the analysis showed that 96.6% of the variance in the data is explained by PC9, so only the first 9 principal components were kept. After that, the individual explained variance drops to below 1%.

A similar approach was followed for industry features. 95% of the variance is explained by PC24, after which the individual explained variance falls significantly and below 1%. That is why only PC1 to PC24 were kept for the regression analysis.

The principal component analysis helped greatly in reducing the dimensionality of the data. By including this step in the capstone project, we managed to reduce the number of potential regressors by more than 75% while still keeping more than 95% of the initial information in the dataset.

Final Steps

The produced macro and industry principal components are merged with the original company data. The final prepared datasets are ready to be used for further PD modelling.

Key outcomes and Lessons learnt

I believe the project was a successful collaboration between both parties. It went beyond its objectives and produced a solution which can be applied not only for the current need, but also for future fine-tunings (including recent statistical data) and new countries added to the portfolio.

Personally, it was a rewarding experience for me as well since I was able to perform analysis that I hadn't had the chance to do before on real data. It was also interesting to try credit risk modelling from the perspective of a start-up business.