Capstone Project Summary

(This document contains no confidential information and should not be restricted in the CEU ETD system.)

Modelling Probability of Default Using Data on Hungarian Companies

Abylaikhan Shaken
Department of Economics and Business
Central European University
Vienna, Austria
(email: shaken_abylaikhan@student.ceu.edu)

Supervisor:
Dr. Ibolya Schindele
Senior Lecturer, Department of Economics
Central European University

COPYRIGHT NOTICE

AUTHOR'S DECLARATION

I, the undersigned, Abylaikhan Shaken, candidate for the MSc degree in Finance declare herewith that the present thesis titled "Modelling Probability of Default Using Data on Hungarian Companies" is exclusively my own work, based on my research and only such external information as properly credited in notes and bibliography.

I declare that no unidentified and illegitimate use was made of the work of others, and no part of the thesis infringes on any person's or institution's copyright.

I also declare that no part of the thesis has been submitted in this form to any other institution of higher education for an academic degree.

Vienna, 08 June 2025

Abylaikhan Shaken

Table of Contents

1. Project Motivation

Financial distress can occur long before formal bankruptcy or default is reported. Early detection is key for banks, lenders, consultants, and policymakers to manage credit risk, prevent contagion, and allocate capital efficiently. However, conventional scoring models (e.g. Altman Z-score, Ohlson O-score, Zmijewski model) rely on static formulas developed decades ago in the U.S. context, using historical data from listed firms.

In this project, I propose a modern, data-driven alternative tailored to the Hungarian SME and corporate landscape. The aim was to build a predictive model that reflects recent economic conditions, local firm behavior, and industry-specific financial structures, while using flexible machine learning methods.

2. Data Source and Structure

The dataset consisted of firm-level annual financial statements covering the years 2014–2023, with over 18,000 firm-year observations. The data was provided in raw Excel format with one sheet per year, and required extensive cleaning and reshaping.

Key cleaning steps included:
- Standardizing column names across years
- Dropping highly incomplete variables (e.g. net working capital, bonds)
- Removing COVID years (2020–2021) due to extreme volatility
- Filtering to keep firms with at least two consecutive years of data
- Winsorizing financial ratios at the 1st and 99th percentiles to manage outliers

The final dataset included hundreds of variables per firm, ranging from standard balance sheet and income statement items to derived ratios and company-level attributes (like firm age).

3. Constructing a Custom Distress Label

One of the most important methodological innovations was defining the distress label. Instead of using bankruptcy or default flags—which are rare, delayed, and often unavailable—I created a composite financial health score based on four widely accepted ratios:
- Profit to Assets
- EBITDA Margin
- Revenue Growth
- Equity to Assets

Each ratio was standardized (converted to z-scores) and combined into a single score.

Firms in the bottom 30% were labeled as "distressed", the rest as "healthy". This approach increases statistical power and reflects a multi-dimensional view of financial weakness.

Compared to traditional scores (Altman, Ohlson, Zmijewski), this method offers:
- Adaptability: Tuned to local Hungarian data
- Transparency: Fully explainable and customizable
- Continuity: Captures early-stage financial stress rather than binary default

4. Machine Learning Pipeline

To predict distress, I used only non-leaky features—meaning I excluded any variable that was part of the target definition. Instead, predictors included variables such as:
- Cash and liquid assets
- Receivables and inventories
- Liabilities structure (short/long-term)
- Firm age

After a train-test split by year (train: 2014–2019; test: 2022–2023), I trained three models:
- Logistic Regression (baseline)
- Random Forest (handles nonlinearity and feature interaction)
- XGBoost (optimized gradient boosting)

Finally, I combined them using a Voting Ensemble (soft voting on probabilities). I also optimized the decision threshold based on F1-score, rather than defaulting to 0.5, to better handle class imbalance. The optimal threshold was found to be 0.27.

5. Results

The Voting Ensemble model showed the strongest performance:
- F1-score: 0.48
- Precision: 0.35
- Recall: 0.72
- ROC AUC: 0.728

The model was particularly strong in identifying true positives—important in financial applications where missing a distressed company is more costly than flagging a healthy one.

Additionally, I evaluated the statistical difference between distressed and healthy groups

using t-tests and visual comparisons of each ratio. All four ratios showed statistically significant mean differences, supporting the validity of the custom label.

6. Practical Output: Dashboard

To make the model accessible to non-technical users (e.g., analysts, consultants), I developed a user-friendly Streamlit dashboard. Users can enter six financial metrics and receive:

- A probability of distress
- Visual risk indicator (low/medium/high)
- Optional breakdown of top contributing variables

This tool allows real-time prediction and could be integrated into broader risk assessment systems.

7. Limitations and Future Improvements

- The model does not use macroeconomic variables (e.g., inflation, interest rates), which may affect firm behavior
- Feature engineering could be enhanced with year-over-year trends or sector indicators
- Labeling could be improved using expert-reviewed credit events or clustering-based approaches

8. Conclusion

This capstone demonstrates how modern data science can enhance financial risk analysis. By combining machine learning with sound economic reasoning and local data, the project creates a scalable, explainable, and robust tool for identifying financially distressed firms—valuable for lenders, advisors, and regulators.