Understanding and Attaining an Investment Grade Rating in the Age of Explainable AI

Ravi Makwana* Dhruvil Bhatt† Kirtan Delwadia‡ Agam Shah§
Bhaskar Chaudhury¶

This Version: August 8, 2023

Abstract

Specialized agencies issue corporate credit ratings to evaluate the creditworthiness of a company, serving as a crucial financial indicator for potential investors. These ratings offer a tangible understanding of the risks associated with the credit investment returns of a company. Every company aims to achieve a favorable credit rating, as it enables them to attract more investments and reduce their cost of capital. Credit rating agencies typically employ unique rating scales that are broadly categorized into investment-grade or non-investment-grade (junk) classes. Given the extensive assessment conducted by credit rating agencies, it becomes a challenge for companies to formulate a straightforward and all-encompassing set of rules which may help to understand and improve their credit rating. This paper employs explainable AI, specifically decision trees, using historical data to establish an empirical rule on financial ratios. The rule obtained using the proposed approach can be effectively utilized to understand as well as plan and attain an investment-grade rating. Additionally, the study investigates the temporal aspect by identifying the optimal time window for training data. As the availability of structured data for temporal analysis is currently limited, this study addresses this challenge by creating a large and high-quality curated dataset. This dataset serves as a valuable resource for conducting comprehensive temporal analysis. Our analysis demonstrates that the empirical rule derived from historical data, yields a high precision value, and therefore highlights the effectiveness of our proposed approach as a valuable guideline and a feasible decision support system.

*Group in Computational Science and HPC, DA-IICT, Gandhinagar, India; *Email:* ravird796@gmail.com

†University of California, Irvine, CA, USA; *Email:* dhruvilbhattlm10@gmail.com

‡University of Southern California, Los Angeles, CA, USA; *Email:* kirtandelwadia@gmail.com

§Georgia Institute of Technology, GA, USA; *Email:* ashah482@gatech.edu

¶Group in Computational Science and HPC, DA-IICT, Gandhinagar, India; *Email:* bhaskar_chaudhury@daiict.ac.in

1 Introduction

The credit rating of a corporation is an assessment made by a credit rating agency to evaluate the corporation's ability to meet its financial obligations (Gonzalez et al., 2004). This rating is used by various parties, including debt issuers, investors, regulators, and other corporations, to make decisions. A favorable credit rating not only attracts investors and potential partners but also generally results in lower interest rates for a corporation. Additionally, a higher credit rating is typically associated with a lower likelihood of default in the future (Matthies, 2013). Conversely, a lower credit rating can negatively impact a corporation's stock price and capital cost (Kisgen and Strahan, 2010). Ultimately, a corporation's credit rating reflects its credibility and reputation, and therefore, companies strive to achieve a high rating. Rating analysts focus on various quantitative and qualitative factors such as industry risk, country risk, the competitive position of a company, historical cyclicality of revenues, profit margins and their volatility, operating efficiency, leverage, capital structure, liquidity, diversification, and many more (Kernan et al., 2013) for rating a company. Thus, replicating an end-to-end model which predicts corporate credit rating is a challenging task.

Prediction of corporate credit rating started with the work of Altman (1968), who applied discriminant analysis to predict corporate defaults. Huang et al. (2004), Novotná (2012), Oelerich and Poddig (2006) and Wei et al. (2010) applied logistic regression, while Hung et al. (2013), Karminsky and Khromova (2016) and Mizen and Tsoukas (2012) applied ordered probit model to predict credit rating. Multivariate discriminant analysis have also been performed by Novotná (2012) and Pinches and Mingo (1973). But it is difficult to use these statistical methods to accurately model complex financial systems due to the limitations of the model and the statistical assumptions (Wang and Ku, 2021). Hence, black-box machine learning techniques have become a widely accepted choice for credit rating prediction compared to simple statistical methods. One or other variants of SVM (Chen and Shih, 2006; Guo et al., 2012; jae Kim and Ahn, 2012; Kim and Sohn, 2010) and

neural networks (Feng et al., 2020; Golbayani et al., 2020b; Hájek, 2011; Wang and Ku, 2021; Yijun et al., 2009) are also investigated for the prediction of corporate credit rating. Golbayani et al. (2020a), Huang et al. (2004), Lee (2007) and Wallis et al. (2019) compared various statistical and machine learning methods for prediction. Existing research in this domain also achieves higher accuracy for credit rating prediction by incorporating various ML techniques like feature selection (Hajek and Michalak, 2013), balancing minority classes (Hájek and Olej, 2014), ensembling schemes (Abellán and Castellano, 2017; Hájek and Olej, 2014; Marqués et al., 2012; Nguyen et al., 2021), and fusion models (Wu and Hsu, 2012). Further, Golbayani et al. (2020a) provides an extensive comparative study of corporate credit rating prediction models. So far, most of the work is focused on achieving higher prediction accuracy, rather than explaining the model and generating inferences from it.

The interpretability of a predictive model is a crucial and a highly desired characteristic in the financial domain. The opaque nature of black-box machine learning algorithms limits their acceptance by financial practitioners and rating analysts (Martens et al., 2009). Also, these algorithms can not be utilized by corporations to make clear strategies to enhance their ratings. Huang et al. (2004) and Kim (2005) tried to partially solve the problem of interpretability by giving a rank of importance to features in the credit rating prediction problem. Bussmann et al. (2021) and de Lange et al. (2022) used Shaply values to figure out crucial financial features in predicting credit rating and credit scores, respectively. But the bounds on those financial features to achieve better credit rating are not provided. Roeder et al. (2022) linked credit risk measured by CDS (Credit Default Swaps) spreads with the sentiments extracted from company's analysis reports created by financial experts. Company wise relevant topics were also enlisted using NLP techniques for the better interpretability. But here, the focus was on textual information rather than financial ratios. Other researchers tried to build a model which provides interpretability by stating a set of rules based on financial ratios. Martens et al. (2009) used Ant Colony Optimization (ACO) for binary classification. ACO generates an interpretable and comprehensive set of rules which can be

used by humans to determine a firm's credit rating. Wu and Hsu (2012) came up with a novel fusion approach called the enhanced decision support model. The data-set with predicted labels from the relevance vector machine was fed to the decision tree as a training data-set to map the complex classification method into a set of interpretable rules. Novotná (2012) applied decision trees for credit rating prediction of European companies.

Fully explainable prediction models are more likely to be used for practical applications in comparison to partially explainable or black box models (Martens et al., 2009). Therefore, in this work, we aim at using explainable AI (Artificial Intelligence) techniques for prediction purposes. Angelov et al. (2021) presented an analytical overview of different explainable AI techniques, which suggest that decision trees offer a high interpretability of the results, as well as deliver a prediction accuracy comparable to other Deep Learning models. Predictions based on random forest can deliver moderately better accuracy than the decision tree model, however, the results obtained would be less interpretable due to the opaque nature of the model. Hence, in our work, we build a fully explainable data driven prediction model based on decision trees and we focus on binary classification with the two most important classes (investment and junk). Derived rules from the decision tree should remain stable with time. Hence, we test temporal stability using time based train-test split of the dataset. Feature selection is also done by blending correlation analyses with financial domain knowledge, to reduce computational costs and add better interpretability. Our work uniquely combines feature selection leveraging domain knowledge and emphasizes the long-term temporal stability of predictions. In order to further enhance the interpretability of our findings, we enhance the basic visualization of decision trees by adding distribution. This approach enhances transparency and facilitates informed decision-making by incorporating both expert domain knowledge and visual representations that capture the distribution of data.

Rating analysts use different methodologies, norms, and rules to assess the credit rating of corporations belonging to different business sectors (Kernan et al., 2013). Less attention has been paid to the fact that the business sector of a company also plays an important role in

the interpretation of financial ratios. In the present work, we focus on building a prediction model using a decision tree for only one such sector, which can be further extended to all sectors. However, the lack of data for any particular sector is a major obstacle (Golbayani et al., 2020a), as large datasets are either paid or proprietary. Therefore, with extensive work, we create our own novel dataset which is being made publicly available under a Creative Commons license. The key contributions of this work are as follows:

- Present the largest open, quality curated dataset of corporate credit rating with attached financial ratios.
- Feature selection by combining domain knowledge and correlation analysis.
- Development of an interpretable model with visualization in contrast to a black box model. Visualization used in this study enables to better understand the underlying patterns and factors influencing credit ratings.
- Devising a set of time-independent, simple if-else rules based on financial ratios to attain investment grade rating with an attached precision value.
- Temporal validation of derived rules on financial ratios.

The rest of the paper is organized as follows. Section 2 describes an end-to-end dataset creation pipeline. The feature selection process is presented in Section 3. Section 4 provides the details of the computational experiments and the important results, followed by the conclusion and future work in Section 5.

2 Dataset creation

To the best of our knowledge, corporate credit rating datasets for US companies are either paid or proprietary (Golbayani et al., 2020a; Hajek and Michalak, 2013; Hájek and Olej, 2014; Kim, 2005; Martens et al., 2009). Open datasets are collectively large (Nguyen et al., 2021),

but filtering those datasets for a particular business sector and a particular year makes the dataset too small. This poses a major challenge for performing detailed temporal analysis for a single business sector. Few datasets (Wang and Ku, 2021) do not include the business sector of a company. As credit rating agencies follow different methodologies for different sectors while assigning rating, business sector information is vital for robust analysis.

The majority of the time for running a machine learning end-to-end pipeline is spent on preparing a quality data (Roh et al., 2021). The lack of a large dataset makes the credit rating problem more challenging (Golbayani et al., 2020a). The three largest credit rating agencies are Standard & Poor's Global Ratings (S&P), Moody's Corporation, and Fitch Ratings, and each credit rating agency has a unique, discrete ordinal rating scale. For example, the rating scale of the S&P is: $\{AAA, AA+, AA, AA-, A+, A, A-, BBB+, BBB, BBB-, BB+, BB, BB-, B+, B, B-, CCC+, CCC, CCC-, CC, C, D\}$ – a total of 22 grades that are ordered from AAA, the most promising one to D, the most risky one. S&P broadly classifies the companies with rating higher than $BB+$ as investment grade companies and others as junk grade companies. Manually mapping all rating scales to S&P 22-grade rating scale is tedious and time-consuming. Bulk collection of financial ratios is only possible by company ticker, or Central Index Key (CIK) since they are unique identifiers of a company. The absence of them in the credit rating dataset makes it difficult to collect the financial ratios. Therefore, we build an end-to-end pipeline to tackle all the foregoing issues and provide a large, quality curated open-source dataset.

2.1 Primary datasets and data extraction

The end-to-end data Extraction, Transform and Load (ETL) pipeline is depicted in Fig. 1. Five different datasets are utilized in order to generate a large open-source corporate credit rating dataset for US companies. Sample data-points for all five datasets are depicted in the upper part of Fig. 1. A detailed description of these five primary datasets is provided below.

- **D1: Credit rating -** Joffe and Partnoy (2018) described various projects which aim

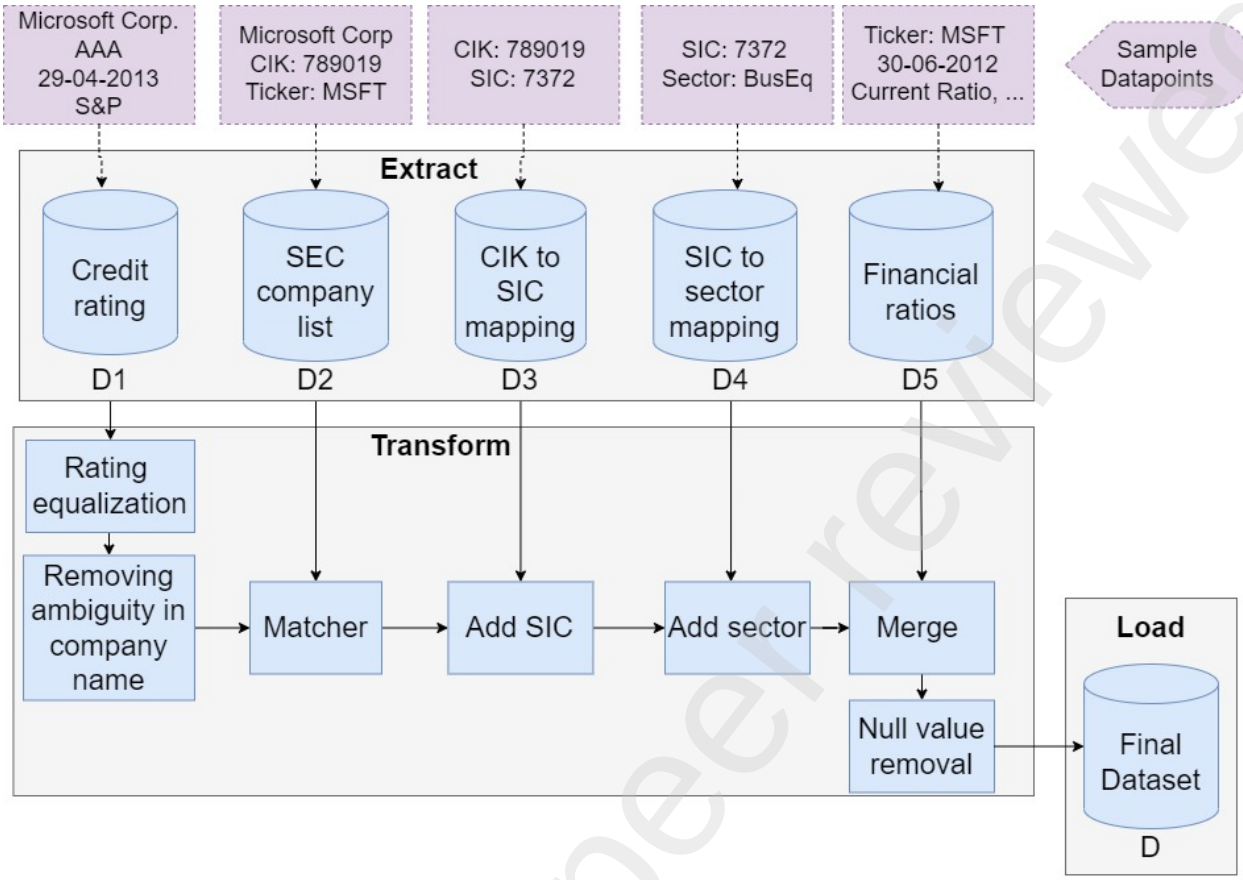


Figure 1: Data ETL (Extraction, Transform and Load) pipeline to create corporate credit rating dataset with attached financial ratios. Sample data points are provided for all primary datasets at the top of the figure.

to make credit rating data publicly available. For our work, we use a corporate credit rating dataset from data-world [1]. This dataset contains credit rating for various entities like US public, sovereign, financial, corporate, residential mortgage-backed securities, etc. Credit ratings of corporate entities ranging from the year 2010 to 2016 are collected for our research. As depicted in Fig. 1 company name (ex. Microsoft), credit rating (ex. AAA), rating date (ex. 29-04-2013), and rating agency (ex. S&P) are retrieved from dataset D1.

- **D2: Security and Exchange Commission (SEC) company list -** Company's name (ex. Microsoft), ticker (ex. MSFT), and CIK (ex. 789019) of all publicly traded

[1] See https://data.world/muni-finance/credit-ratings-history-data

US companies are gathered from SEC's EDGAR (Electronic Data Gathering, Analysis, and Retrieval system) API[2] (as depicted in Fig. 1). This step is mandatory because financial ratios can be gathered easily via the company ticker or company CIK.

- **D3: CIK to SIC mapping -** Standard Industrial Classification Codes (SIC Codes) are used to identify the primary line of business of a company. Dataset D3 contains two data fields, namely CIK (ex. 789019) and SIC (ex. 7372) for all publicly traded US companies collected from SEC's website.

- **D4: SIC to sector mapping[3]-** This dataset maps various SIC code to the corresponding business sector of a company. For example, SIC-7372 maps to *Business equipment (BusEq)* sector. (See Fig. 1)

- **D5: Financial ratios -** Financial ratios are sourced from SEC's website. It contains company's ticker (ex. MSFT), financial ratios (Current ratio, Net profit margin, etc.) along with the date on which these financial ratios are calculated (ex. 30-06-2012).

2.2 Data transformation

A detailed description of all blocks in the *Transform* part in Fig. 1 is provided below.

- **Rating equalization:** All rating agencies use their unique pool of predefined ratings. We manually map the rating scale of all rating agencies to S&P rating scale (total of 22 ratings) by visiting their respective web pages. For example, Moody's '$Aa1$' rating is equivalent to S&P's '$AA+$' rating. Datapoints having ratings that can not be converted into S&P rating scale are dropped. S&P broadly classifies the companies with rating higher than '$BB+$' as investment grade companies and others as junk grade companies. Since the aim of our current research is binary classification, we

[2]See https://www.sec.gov/files/company_tickers_exchange.json
[3]See http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/Data_Library/det_12_ind_port.html

add a binary representation of rating in the dataset, where '0' represents junk grade company while '1' represents investment grade company.

- **Removing ambiguity in company name:** Same company is listed with distinct names in dataset D1. For example, Microsoft is present in D1 with three distinct names: 'Microsoft Corporation', 'MICROSOFT CORPORATION', and 'Microsoft Corp'. It is an absolute requirement to identify such ambiguities and mark those companies as identical. Firstly, company names are sorted and converted to lowercase English letters. Removal of common words like 'corp', 'corporation', 'inc', 'ltd', and 'limited' from the end of the company name helps reduce such ambiguous instances from the dataset. Even after this step, not all ambiguities are eliminated. To solve this, we introduce the notion of a dot product between two strings. The string is mapped into a corresponding 26 dimensional frequency vector, where the i^{th} element of the frequency vector denotes the occurrence of i^{th} English character in the string. And a dot product between two strings is equivalent to the dot product between their respective frequency vectors. If $Cos(\theta)$ between two frequency vectors is greater than a predefined threshold, then the two strings can be considered as the same. An optimal threshold on $Cos(\theta)$ is found to be 0.92 through extensive manual check. An example of computing the value of $Cos(\theta)$ between two strings is provided in A.

- **Matcher:** Company CIK and company ticker are absent in dataset D1, and bulk collection of financial ratios is not possible without company ticker or company CIK. These attributes can be added in dataset D1 from dataset D2. We encounter many instances of identical companies having distinct names in dataset D1 and dataset D2. Therefore, hard string matching on company name is not an efficient technique to merge both datasets. To tackle this issue, firstly, the company names are converted to lowercase English letters. Removal of common words like 'corp', 'corporation', 'inc', 'ltd', and 'limited' from the end of the company names across both datasets helps to

build an efficient matcher. After trying several approaches, we construct a flexible matcher according to which two companies, one from D1 and the other from D2 are identical if and only if these two constraints are satisfied:

1. The first word of both companies should be identical.
2. $Cos(\theta)$ between the two company names is greater than 0.92 (as mentioned previously).

Companies with zero and multiple matches are dropped to maintain the quality of the dataset.

- **Add SIC:** SIC codes of companies are attached with the help of dataset D3. There exist many classification schemes to assign a sector to a corporation based on its SIC code. For instance, the Fama-French three-factor model uses a classification scheme with 48 sectors. SIC code is added to our dataset in addition to the sector because it provides the flexibility to utilize any fine-grained classification scheme.

- **Add sector:** Sector of companies is added from dataset D4.

- **Merge:** Ticker-wise financial ratios are sourced from dataset D5 ranging from the year 2009 to 2021 for all companies presented in D1. If a company is rated on date X, then its latest historical financial ratios are considered for merging credit rating with financial ratios. Companies for which the financial ratios are not present in dataset D5 are dropped.

- **Null value removal:** Four financial ratios are dropped due to a very high number of null values. Eventually, datapoints with at least one null feature (financial ratio) are dropped.

2.3 Dataset attributes and comparison with existing dataset

In the *Load* part of Fig. 1, our final corporate credit rating dataset with attached financial ratios is depicted as the dataset D[4]. Each row (datapoint) has a total of 25 attributes, out of which 16 attributes are financial ratios that can be utilized as input features for training ML and AI algorithms. All of these attributes are presented in Table 1, and the financial ratios are highlighted in bold text.

Rating agency	Corporation
CIK	Ticker
Rating	Binary rating
Sector	Rating date
SIC code	**Asset Turnover**
Current Ratio	**Long-term Debt/Capital**
Debt/Equity Ratio	**Gross Margin**
Operating Margin	**EBIT Margin**
EBITDA Margin	**Pre-Tax Profit Margin**
Net Profit Margin	**Free Cash Flow Per Share**
ROE - Return On Equity	**Return On Tangible Equity**
ROA - Return On Assets	**ROI - Return On Investment**
Operating Cash Flow Per Share	

Table 1: 25 attributes of a datapoint from our corporate credit rating dataset, out of which 16 financial ratios are highlighted in bold

We benchmark our dataset with an already existing largest open-source dataset from Kaggle, named as *Corporate Credit Rating*[5]. Comparison is shown in Table 2. The proposed dataset has 16 financial ratios which are less compared to the *Corporate Credit Rating dataset.* As explained in Section 3 some of these features are highly correlated, so even obtaining a reduced feature set out of these 16 features would not affect the results significantly. Also, addition of more features may lead to a loss of interpretability. Hence, more financial ratios will not add much value to our current research goal.

We validate few random samples of the dataset D1 with the *Corporate Credit Rating*

[4]Our novel dataset is on Kaggle: https://www.kaggle.com/datasets/kirtandelwadia/corporate-credit-rating-with-financial-ratios; DOI: https://doi.org/10.34740/kaggle/ds/2277577

[5]See https://www.kaggle.com/datasets/agewerc/corporate-credit-rating

Dataset	Proposed	Corporate Credit Rating
Timeframe	2010 – 2016	2010 – 2016
#Datapoints	7805	2027
Companies covered	678	593
#Rating Agency	7	5
#Sector	12	12
Rating Scale	S&P 22-grades	S&P 10-grades
CIK and SIC	Provided	Not provided
#Financial ratios	16	25

Table 2: Comparison of our dataset with the largest publicly available dataset, *Corporate Credit Rating*

dataset manually. Also, validation for a few random samples of scrapped financial ratios is done with manually calculated financial ratios obtained from SEC fillings.

There are 12 business sectors in our dataset. It is necessary to generate different prediction models for different sectors, as discussed in Section 1. So the current scope of our research is limited to one such sector, which is *Business Equipment*. This sector is selected due to the availability of high number of datapoints and a relatively easy methodology followed by rating analysts to rate corporations from this sector. All analysis and results presented in further sections are only for the *Business Equipment* sector, but can be extrapolated for all sectors.

3 Feature selection

Huang et al. (2004) concluded that a relatively small list of financial variables largely determines the rating results. Additionally, the interpretability of results calls for a smaller set of features. Therefore, a reduction in the number of features is essential.

In traditional ML problems, feature selection is done on the basis of computational algorithms like Mutual information, Information gain, Correlation-based filter, PCA, Chi-squared test, etc. All of these algorithms tend to make use of computational techniques to filter out features rather than making use of financial knowledge. Even few techniques produce a new and smaller set of features by combining financial ratios. These techniques

may reduce the number of features, but generated new features do not carry any financial relevance. We utilize feature correlation as well as findings of rating analysts and domain knowledge to find our reduced set of features.

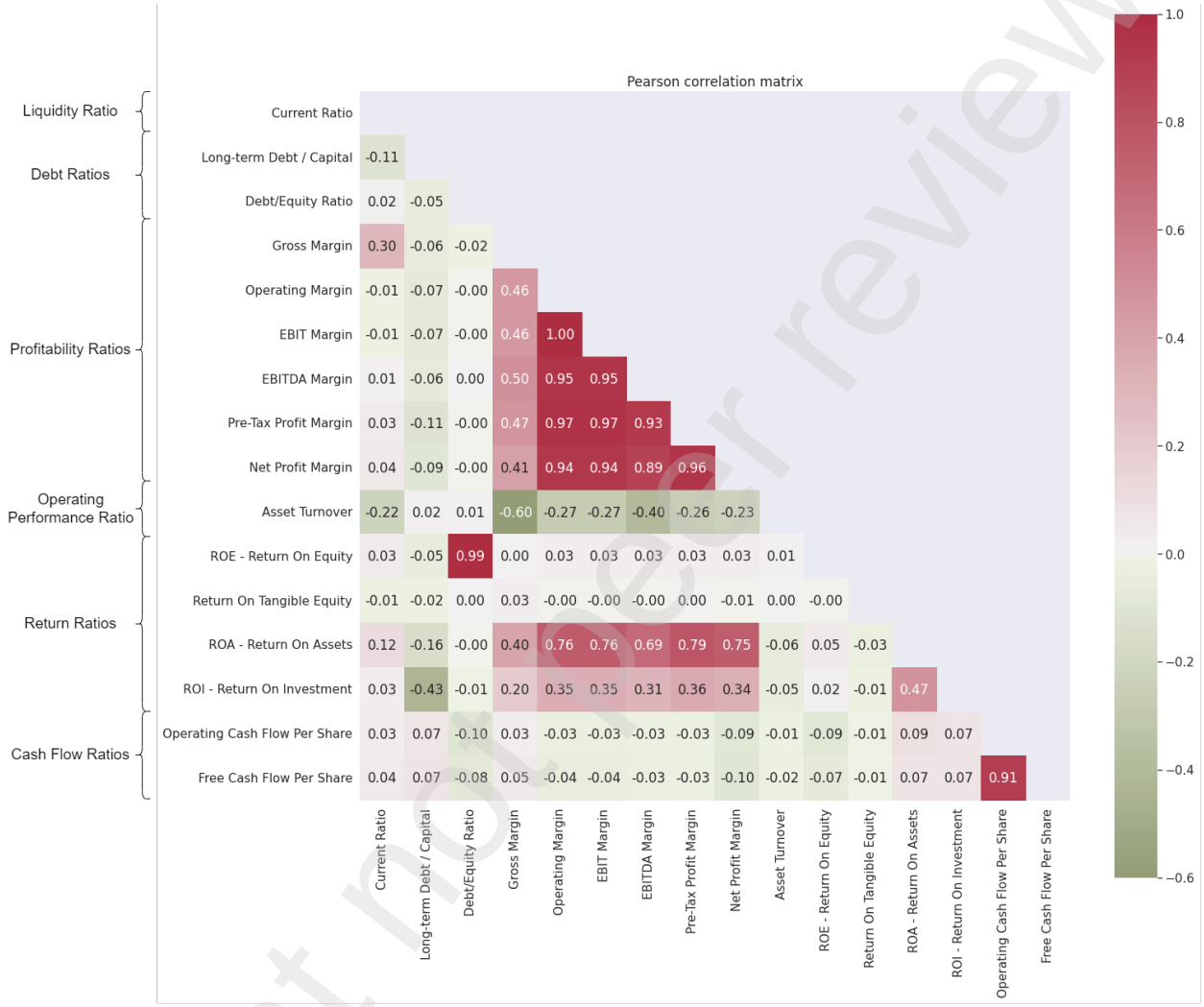


Figure 2: Feature correlation matrix for *Business Equipment* sector, where cell (i, j) represents the pearson correlation coefficient between the i^{th} and j^{th} feature. Each of the financial ratios is further classified in the left part of the figure

In Fig. 2 a feature correlation matrix is depicted, where a cell (i, j) represents the Pearson correlation coefficient between the i^{th} and j^{th} feature. All financial ratios are further categorized in the same figure. The current ratio and asset turnover are not significantly correlated with any other feature, hence retained. Operating cash flow per share and Free cash flow per share are highly correlated, and they are classified as cash ratios in the finance

domain. Hence, only operating cash flow per share is retained. Debt/Equity Ratio and Long-term Debt/Capital are classified as debt ratios in finance, therefore only the Debt/Equity ratio is selected, and the latter is dropped. The debt/Equity ratio is highly correlated with Return on Equity, but since they provide a different kind of financial information, both are retained. All profit margins namely gross, operating, EBITDA, EBIT, pretax, and net are highly correlated. Gross profit can be obtained by subtracting the cost of goods from revenue, and it is considered to be the initial stage profit of a corporation. It is also not very highly correlated with other profit margins. EBITDA can be viewed as intermediary stage profit, which is profit before interest, tax, depreciation, and amortization. Also, EBITDA plays a pivotal role in determining credit rating (Kernan et al., 2013). Net profit is the final stage profit of a corporation after subtracting all expenses. These three ratios are sufficient to capture the overall profitability of a corporation, hence they are selected. Finally selected 9 financial ratios (features) are provided in Table 3.

Current Ratio	Debt/Equity Ratio
Gross Margin	EBITDA Margin
Net Profit Margin	Asset Turnover
ROE - Return on Equity	ROI - Return on Investment
Operating Cash Flow Per Share	

Table 3: Reduced set of 9 financial ratios for *Business Equipment* sector after performing correlation analysis with domain knowledge

4 Results and Discussion

In this section, we analyze the performance of explainable AI technique for corporate credit rating prediction. Since we aim at generating simple if-else conditions for attaining investment grade rating, decision trees are used for credit rating prediction purpose. This section also aims at finding the optimal decision tree and financial parameters for achieving the best results. These parameters are enlisted below, and they need to be tuned before training the decision tree. For each parameter, there are multiple choices from which the optimal choice

has to be filtered.

- **Choice of decision tree algorithm:** ID3(Quinlan, 2004) and CART (Breiman et al., 2017) algorithms are considered for our experiments. Hence, the set of all possible choices for this parameter are:

$$S_{decision\_tree} = \{ID3, CART\} \tag{1}$$

- **Feature set for training decision tree:** Our experimentation with choosing the feature set involves two choices. First, all the available numeric features in the dataset can be used, or second, the reduced set of features can be used (See Table 3). Possible choices for this parameter are:

$$S_{feature\_set} = \{All\ features, Reduced\ features\} \tag{2}$$

- **Training data window size:** Huang et al. (2004), Lee (2007), Nguyen et al. (2021), Novotná (2012) and Wu and Hsu (2012) combined the credit rating data of all available years and randomly divided it into a training-testing dataset. This traditional method for obtaining a training-testing dataset is not coherent with the credit rating prediction problem because of an absolute requirement to test the model for samples that are outside the time window of the training data. Hence, training a model on historical years and testing it on subsequent years ensures the temporal stability of the model, which is one of the major aims of our research. Initially, training of the decision tree is done on the dataset of a single year and tested on subsequent years. For example, a decision tree can be trained on 2013 year data and tested on 2014, 2015, and 2016 data, or it can be trained on 2014 year data and tested on 2015 and 2016 data. But it resulted in poor accuracy and high instability, as described in the next subsections. As a solution, data from multiple years are utilized for training purpose, with more weight assigned to data points of recent years (described in detail in the subsections below).

Hence, the notion of window size is introduced. Window size of k represents that data of total k historical years are utilized for training the decision tree. The value of this parameter is assigned from the set:

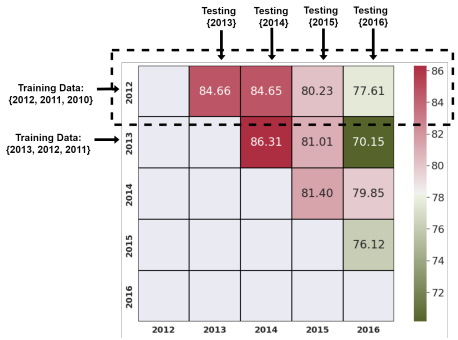
$$S_{window\_size} = \{1, 2, 3, 4, 5\} \tag{3}$$

- **Depth of a decision tree :** The depth of a decision tree is an important parameter that could lead to either overfitting or underfitting the ML model. Increasing the depth of the decision tree might improve prediction accuracy, but causes the problem of overfitting and generates a gigantic set of rules. Since one of the major aims of our work is to develop an explainable AI model that could provide a simple and concise set of rules to achieve an investment grade rating, it is essential to determine the optimum depth of the decision tree which balances accuracy and interpretability. The depth of decision tree is selected from the set given below:

$$S_{depth} = \{1, 2, 3, 4, 5, 6, 7, 8\} \tag{4}$$

4.1 Prediction accuracy metric

In order to evaluate the performance of a decision tree for a specific set of parameters, we introduce a notion of an accuracy matrix. An accuracy matrix aids to simplify the temporal analysis of financial rules and helps in testing the validity of past financial rules for different window sizes. Accuracy matrices of decision trees trained on different parameters are shown in Fig. 3. We use the concept of a rolling window for training purpose. The first row of the accuracy matrix in Fig. 3a shows the prediction accuracy values (in percentage) of the decision tree trained on the 2010-2012 data for the *Business Equipment* sector and tested for the years after 2012. In a more generic sense, for the accuracy matrix shown in Fig. 3a having window size 3, the row labeled as the year X contains training data from the years X, $X-1$ and $X-2$. The datapoints of the year X are assigned the highest weight while



(a) Accuracy matrix for decision trees trained for the following choice of parameters: (i) Decision tree algorithm: CART (ii) Feature set: reduced feature set (iii) Window size: 3 (iv) Depth: 4

(b) Accuracy matrix for decision trees trained for the following choice of parameters: (i) Decision tree algorithm: ID3 (ii) Feature set: all features (iii) Window size: 2 (iv) Depth: 2

Figure 3: Accuracy matrices for decision trees trained for different sets of parameters. A row labeled as X in an accuracy matrix consists of weighted training datapoints from the years {X, X-1, X-2,..., X-(k-1)} where k is the window size. A column labeled as year Y represents that testing is done on dataset of year Y.

training, followed by the datapoints of year $X-1$ and $X-2$. The ordering of the importance of the datapoint for different years is done with an aim to better capture the recent trend, and at the same time utilize previous years' data. The accuracy matrix is upper triangular, since the historical model is only tested for subsequent years. It is evident from Fig. 3 that going down in a particular column mostly increases the accuracy. Hence, current year's credit rating prediction is best done by the model trained on the most recent historical years. Our temporal analysis concludes that historical models lose their performance after a few years, therefore, they must need to be re-trained. To find the optimal value of four parameters, we need to compare accuracy matrices. As per Fig. 3, the dimension of the accuracy matrix is a function of window size. It is a cumbersome task to compare matrices of different dimensions. This forces to quantify the performance of a set of parameters by a single number. Hence, for a given set of parameters, the mean value of this accuracy matrix over the different temporal training-testing sets is computed. Since the consistency/uniformity of the accuracy values is also vital while analysing the performance of decision tree parameters, the standard deviation of the accuracy matrix is also calculated.

4.2 Estimating the optimal choice of parameters

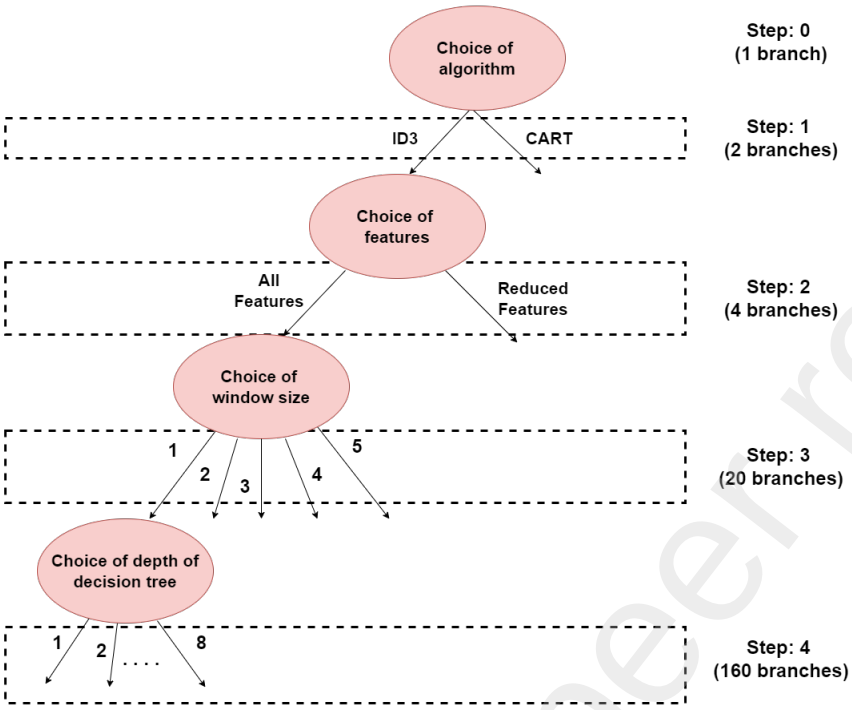


Figure 4: Visualisation of the elimination approach for estimating the optimal choice of parameters

A total of 160 accuracy matrices can be constructed by spanning the space of tunable parameters. We aim at finding the optimum value of each parameter and then determine the empirical rule for achieving an investment grade rating. We define an empirical rule as a set of time independent simple if-else conditions, which if followed could help a corporation attain an investment grade rating. Rather than searching exhaustively for optimum parameters, we are using a step-wise elimination technique. Fig. 4 is a visualisation of the elimination technique used for estimating the best parameters for training the decision tree. The root node of Fig. 4 contains all the parameters required to generate 160 distinct accuracy matrices. Half of these use ID3 as the underlying algorithm, and the rest are trained on the CART algorithm. We first aim at determining the better of the two algorithms on the basis of prediction accuracy analysis.

The mean prediction accuracy value across all the 80 accuracy matrices that use the

ID3 algorithm for training is $75.872 \pm 5.122\%$. Similarly, for the CART algorithm the mean accuracy value is $76.132 \pm 4.76\%$. The accuracy values for both the algorithms are comparable, and eliminating one of these at step-1 (See Fig. 4) could possibly lead to not finding the optimal set of parameters. Hence, we do not eliminate any algorithm in step-1.

In step-2, we consider both of the decision tree algorithms and compare the performance

Window Size	**Mean Accuracy**
1	$73.038 \pm 6.243\%$
2	$75.483 \pm 4.956\%$
3	$78.252 \pm 4.420\%$
4	$77.813 \pm 3.996\%$
5	$77.275 \pm 2.563\%$

Table 4: Mean accuracy of accuracy matrices as a function of window size.(Accuracy is averaged over all possible choices of depth and both of the algorithms. The decision trees are trained only on the reduced feature set. Hence, for each window size, the mean accuracy is calculated over 16 accuracy matrices.)

of the decision trees trained using all the numeric features with the decision trees trained on a reduced feature set. Decision trees trained on all features and reduced feature sets attain a mean prediction accuracy of $75.632 \pm 5.447\%$ and $76.372 \pm 4.435\%$ respectively. A reduced feature set not only trains the decision tree better in terms of mean accuracy but also trains the model to have a more uniform predictive ability, due to its lesser standard deviation. Fewer features will also add more interpretability. For the reasons stated above, we eliminate the choice of using all the features for training the subsequent decision trees. Our result is also aligned with the work of Huang et al. (2004), that a relatively small set of features largely determines the credit rating. Table 4 presents the performance of the decision tree trained on different window sizes. From the accuracy values shown in the table, the performance of the decision trees trained on the dataset of window sizes 3, 4, and 5 show the most optimal results. We eliminate window size 1 and 2 at step-3 of Fig. 4 for analysing the performance of subsequent decision trees.

At the last step of Fig. 4, we determine the best choice for the depth of the decision

Depth of Decision Tree	Mean Accuracy
1	$73.918 \pm 2.254\%$
2	$78.064 \pm 2.917\%$
3	$77.535 \pm 3.575\%$
4	$78.659 \pm 4.179\%$
5	$78.745 \pm 4.797\%$
6	$78.163 \pm 4.479\%$
7	$78.299 \pm 3.529\%$
8	$78.855 \pm 3.546\%$

Table 5: Mean accuracy of accuracy matrices as a function of depth. (Accuracy is averaged over window size {3, 4, 5} and both of the algorithms. The decision trees are trained only on the reduced feature set. Hence, for each depth of the decision tree, the mean accuracy is calculated over 6 accuracy matrices.)

tree. According to the accuracy results presented in Table 5, the decision trees restricted to a depth of 5 show the best mean accuracy result. The performance of the decision trees having a maximum depth of 4, 6, 7, and 8 also show comparable results with the ones having a depth of 5. We aim to establish simple if-else rules for investment grade rating, and since a decision tree of depth greater than 5 would make the rules complicated and less interpretable, it would not be an optimal choice of the depth value for our goal. Hence, we only consider the depth value of 4 and 5 for the construction of the desired decision tree.

After following the above methodology, we arrive at the following choices of the parameters for constructing the optimal decision tree:

- **Choice of decision tree algorithm:** {ID3 algorithm, CART algorithm}.
- **Feature set for training decision tree:** {Reduced set of features}.
- **Training data window size**: {3, 4, 5}
- **Depth of decision tree:** {4, 5}

Different filtered choices of these parameters lead to the construction of 12 accuracy matrices from the initial pool of 160 accuracy matrices. Finally, we fine-tune the parameters based on the exhaustive analysis of these twelve accuracy matrices. Table 6 presents the analysis

Algorithm	Window Size	Depth of Decision Tree	Accuracy
CART	**3**	**4**	$\mathbf{80.339 \pm 4.012\%}$
CART	3	5	$79.037 \pm 4.428\%$
CART	4	4	$78.544 \pm 4.391\%$
CART	4	5	$77.973 \pm 5.321\%$
CART	5	4	$78.773 \pm 3.239\%$
CART	5	5	$77.768 \pm 4.173\%$
ID3	3	4	$78.583 \pm 5.383\%$
ID3	3	5	$79.575 \pm 5.653\%$
ID3	4	4	$78.594 \pm 5.292\%$
ID3	4	5	$79.345 \pm 5.344\%$
ID3	5	4	$77.122 \pm 2.757\%$
ID3	5	5	$78.773 \pm 3.866\%$

Table 6: Mean accuracy values of the accuracy matrices for different choices of the filtered parameters. Best performing choices of parameters are highlighted in bold. The decision trees are trained on the reduced feature set.

of the 12 filtered accuracy matrices, concluding that **Decisision tree algorithm: CART**, **Feature set: Reduced**, **Window size: 3**, **Depth of decision tree: 4** leads to an optimal tuning of parameters which balances both the accuracy and the interpretability.

4.3 Empirical rule and visualisation of decision tree

The accuracy matrix for optimal parameters is already presented in Fig. 3a. A total of 4 decision trees could be constructed for the same choice of optimal parameters that differ from each other in terms of the training dataset.To better illustrate, each of the first 4 rows in Fig. 3a correspond to a decision tree with a different training dataset. Fig. 5 presents the visualisation of the decision tree trained on 2013-2015 dataset for the optimal choice of parameters. As per Fig. 5, there are 7 leaf nodes that have the majority of the investment grade datapoints (i.e have a major portion of green color in the pie chart). Equivalently, it could also be stated that there are 7 different paths (i.e set of if-else conditions) that lead to a major proportion of the investment grade rating. Similar analysis across all 4 decision trees leads to a collection of 21 such leaf nodes or unique paths leading towards investment

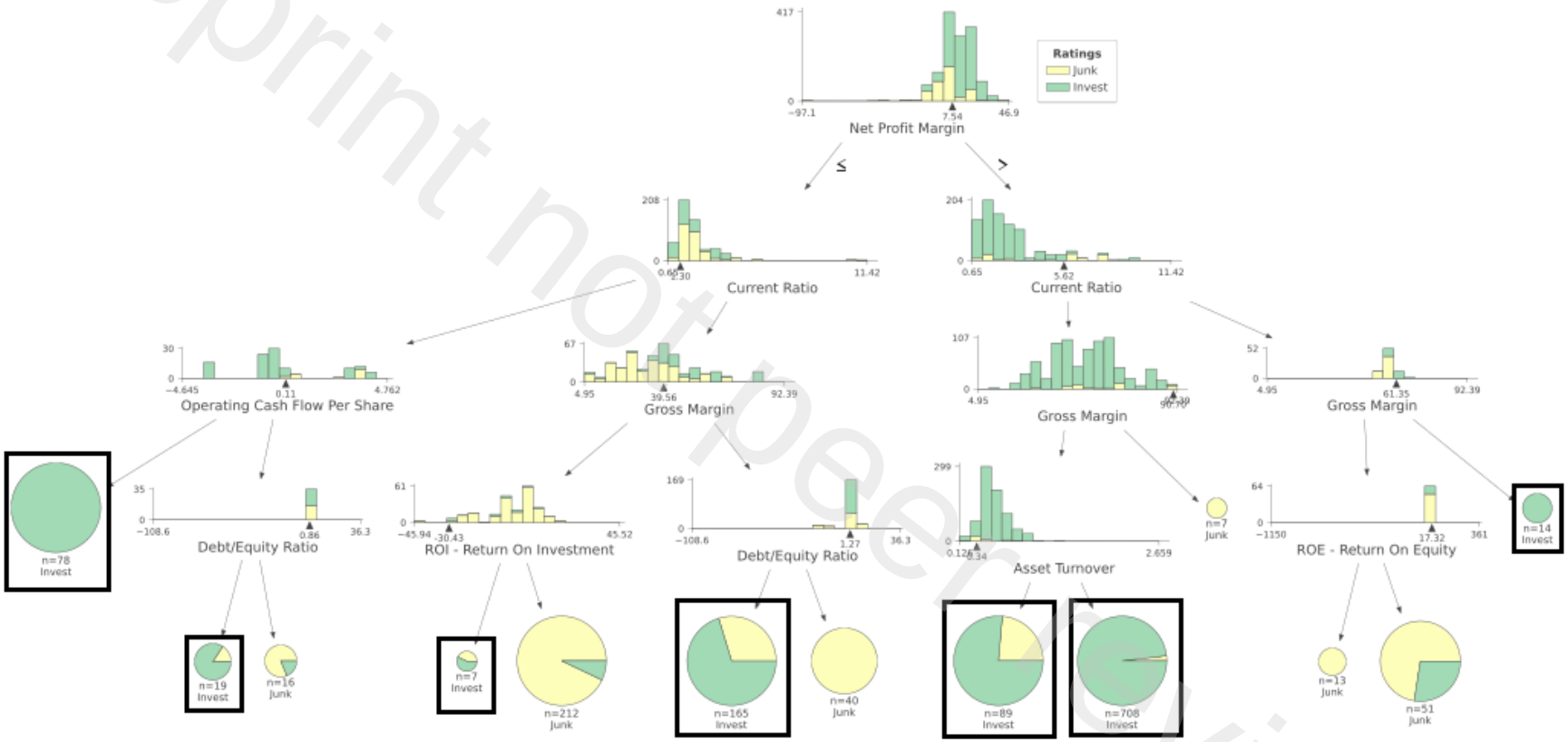


Figure 5: Visualisation of the decision tree trained on 2013-2015 dataset for the optimal choice of parameter, leaf nodes with a majority of investment grade datapoints are highlighted in black box

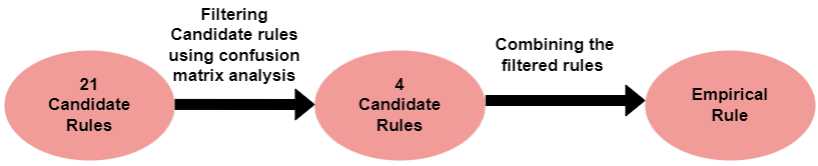


Figure 6: Steps to devise the empirical rule from an initial pool of 21 candidate rules for the *Business Equipment* sector

Empirical Rule
Net Profit Margin $>$ 7.92 &
Return on Investment $>$ 4.39 &
Operating Cash Flow Per Share $>$ -1.58 &
0.701 $\leq$ Current Ratio $\leq$ 4.16

Table 7: Empirical Rule for achieving investment grade rating for the *Business Equipment* sector

grade rating. In order to generate a single empirical rule for achieving investment grade rating, these 21 paths (referred to as 21 **candidate rules**) are analysed.

A confusion matrix for each of the 21 candidate rules is computed across the entire dataset. With a threshold value of 70% for accuracy and 90% for precision value, 4 candidate rules are filtered for further analysis (See Fig. 6). Combining these 4 candidate rules leads to the construction of a single empirical rule for achieving an investment grade rating. This empirical rule is shown in Table 7.

The performance of the empirical rule is analysed across the entire dataset, and the results are presented in Table 9. The mean precision value is close to 95%, while the mean accuracy is approximately 74%. Since one of the primary aims of our research is to devise

Financial Parameter	**Mean**	**Median**
Net Profit Margin	13.189	13.46
Return on Investment	10.85	10.61
Operating Cash Flow Per Share	0.283	0.198
Current Ratio	2.32	1.96

Table 8: Mean and median values of the financial ratios presented in empirical rule over the data consisting of investment grade companies for the *Business Equipment* sector

Year	Accuracy	Precision
2012	83.59%	96.67%
2013	74.23%	98.63%
2014	79.67%	98.48%
2015	67.83%	93.39%
2016	68.66%	87.30%
Mean across all years	73.86%	**94.79**%

Table 9: Accuracy and precision analysis of the empirical rule for the *Business Equipment* sector over different years. Results from the year *2010* and *2011* are not included due to lack of datapoints in these years

an empirical rule, which if followed could help a company attain an investment grade rating with a high probability. Precision value is a better metric for estimating the performance of the empirical rule than the accuracy value. In other words, the precision value of the empirical rule suggests the percentage chance of a company achieving an investment grade rating if the company follows the given empirical rule. Table 9 also depicts that the precision values remain consistently greater than 93% across all the years (except for 2016). Attaining a mean accuracy value of close to 74% also suggests the quality of the derived empirical rule considering that no complex black-box model is used for prediction purpose.

The obtained empirical rule has four financial parameters: net profit margin, return on investment (ROI), operating cash flow per share, and current ratio. Net profit margin measures the net profit generated as a percentage of the revenue and therefore higher net profit margin is better for the company. Return on investment (ROI) measures the benefit an investor will receive on the investments made. A higher ROI is considered to be a good investment. The current ratio of a company measures its ability to pay off its short term obligations in a year. A higher current ratio is generally considered better, but a very high current ratio shows inefficiency in the use of available funds. The threshold values for each of these parameters given in Table 7 are less than their mean and median values(See Table 8) calculated over investment grade companies. In conclusion, the requirements proposed by the empirical rule are attainable.

The devised empirical rule is a set of time-invariant, concise, and explainable rules that would help corporations to identify critical financial ratios and their thresholds to attain investment grade rating. By following the proposed methodology, corporations can form clear strategies to allocate their resources judiciously to achieve investment grade rating.

5 Conclusions and future work

We propose a data-driven explainable AI-based methodology to establish empirical rules on financial ratios to understand and achieve investment class credit rating. The paper provides a methodology for the creation of a large and quality curated credit rating dataset with attached financial ratios. Our dataset is also being made publicly available[6], which is almost four times larger than the existing largest open dataset. In this paper, emphasis has been laid on the interpretability of the system designed to determine the credit rating, and hence decision tree algorithm has been used. Interpretability is also considered in the feature selection process, and selected features indeed improve the performance of the model. The model obtains an average accuracy of 80.34% over a distinct training-testing dataset with the CART algorithm, using window size 3 and maximum depth 4. The highest accuracy obtained is as high as 86.31%, when the testing is done on the data associated with a particular year just succeeding the years on which the model is trained on. The experiments show that the best results are obtained with a window size of 3. Quite interestingly, S&P generally observes the financial ratios of the past 3 years of a corporation to determine its credit ratings (Kernan et al., 2013). Hence, experimental results are aligned with the rating methodology used in practice. The model helps to determine an empirical rule, following which a company can have higher odds of attaining investment grade credit rating. The obtained empirical rule demonstrates a mean precision value of 94.79%, indicating that firms have a 0.94 probability of achieving an investment grade rating if they adhere to this rule. This high precision value highlights the effectiveness and reliability of the rule in guiding companies towards attaining

[6] https://doi.org/10.34740/kaggle/ds/2277577

the desired credit rating. In addition to achieving remarkable accuracy, this study places a strong emphasis on embedding interpretability at each stage of the experiment. The focus on interpretability ensures that the results and findings are easily understandable and provide insights into the underlying factors and decision-making process.

Based on our work, future researchers can extend the research in three directions. (i): Extrapolate the current research for multiple sectors and realize the problem as multiclass classification rather than binary. (ii): Study the performance of various machine learning algorithms and ensembling techniques on our new dataset and achieve the best possible accuracy. Here, interpretability will not be an important factor. (iii): Quantify business risk (country risk, industry risk and competitive position) information utilized by a rating analyst to rate a corporation, and extend the number of features in our dataset. We believe it is important to achieve higher accuracy without compromising the interpretability aspect while predicting corporate credit ratings.

A Example of finding $cos(\theta)$ between two strings

Let string s1='microsoft' and string s2='amazon'. Let $f1$=frequency vector($s1$) and $f2$=frequency vector($s2$). Both frequency vectors are presented in Table 10

	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	...	z
f1	0	0	1	0	0	1	0	0	1	0	0	0	1	0	2	0	0	1	...	0
f2	2	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	...	1

Table 10: $f1$ and $f2$ are frequency vectors of $s1$ and $s2$ respectively, where i^{th} element represents the frequency of i^{th} English alphabet in a respective string

So dot product between $s1$ and $s2$,

$$\begin{aligned} < s1, s2 > &\equiv < f1, f2 > \\ &= \sum_{i=1}^{26} f1(i) \times f2(i) \\ &= 3 \end{aligned}$$

Similarly $< s1, s1 >= 11$ and $< s2, s2 >= 8$

As per the definition,

$$\begin{aligned} cos(\theta) &= \frac{< s1, s2 >}{\sqrt{< s1, s1 >< s2, s2 >}} \\ &= \frac{3}{\sqrt{11 \times 8}} \\ cos(\theta) &= 0.32 \end{aligned}$$

References

Joaquín Abellán and Javier G. Castellano. 2017. A comparative study on base classifiers in ensemble methods for credit scoring. *Expert Systems with Applications*, 73:1–10.

Edward I. Altman. 1968. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, 23(4):589–609.

Plamen P. Angelov, Eduardo A. Soares, Richard Jiang, Nicholas I. Arnold, and Peter M. Atkinson. 2021. Explainable artificial intelligence: an analytical review. *WIREs Data Mining and Knowledge Discovery*, 11(5):e1424.

Leo Breiman, Jerome H Friedman, Richard A Olshen, and Charles J Stone. 2017. *Classification and regression trees*. Routledge.

Niklas Bussmann, Paolo Giudici, Dimitri Marinelli, and Jochen Papenbrock. 2021. Explainable machine learning in credit risk management. *Computational Economics*, 57(1):203–216.

Wun-Hwa Chen and Jen-Ying Shih. 2006. A study of taiwan's issuer credit rating systems using support vector machines. *Expert Systems with Applications*, 30(3):427–435. Intelligent Information Systems for Financial Engineering.

Petter Eilif de Lange, Borger Melsom, Christian Bakke Vennerød, and Sjur Westgaard. 2022. Explainable ai for credit assessment in banks. *Journal of Risk and Financial Management*, 15(12).

Bojing Feng, Wenfang Xue, Bindang Xue, and Zeyu Liu. 2020. Every corporation owns its image: Corporate credit ratings via convolutional neural networks. In *2020 IEEE 6th International Conference on Computer and Communications (ICCC)*, pages 1578–1583.

Parisa Golbayani, Ionuct Florescu, and Rupak Chatterjee. 2020a. A comparative study of forecasting corporate credit ratings using neural networks, support vector machines, and decision trees. *ArXiv*, abs/2007.06617.

Parisa Golbayani, Dan Wang, and Ionut Florescu. 2020b. Application of deep neural networks to assess corporate credit rating.

Fernando Gonzalez, François Haas, Ronald Johannes, Mattias Persson, Liliana Toledo, Roberto Violi, Martin Wieland, and Carmen Zins. 2004. Market dynamics associated with credit ratings: A literature review. *SSRN Electronic Journal*.

Xuesong Guo, Zhengwei Zhu, and Jia Shi. 2012. A corporate credit rating model using support vector domain combined with fuzzy clustering algorithm. *Mathematical Problems in Engineering*, 2012.

Petr Hajek and Krzysztof Michalak. 2013. Feature selection in corporate credit rating prediction. *Knowledge-Based Systems*, 51:72–84.

Zan Huang, Hsinchun Chen, Chia-Jung Hsu, Wun-Hwa Chen, and Soushan Wu. 2004. Credit rating analysis with support vector machines and neural networks: a market comparative study. *Decision Support Systems*, 37(4):543–558. Data mining for financial decision making.

Ken Hung, Hui Wen Cheng, Shih-shen Chen, and Ying-Chen Huang. 2013. Factors that affect credit rating: An application of ordered probit models1. *Romanian Journal of Economic Forecasting*, 16:94–108.

Petr Hájek. 2011. Municipal credit rating modelling by neural networks. *Decision Support Systems*, 51(1):108–118.

Petr Hájek and Vladimír Olej. 2014. Predicting firms' credit ratings using ensembles of artificial immune systems and machine learning – an over-sampling approach. *IFIP Advances in Information and Communication Technology*, 436:29–38.

Kyoung jae Kim and Hyunchul Ahn. 2012. A corporate credit rating model using multi-class support vector machines with an ordinal pairwise partitioning approach. *Computers & Operations Research*, 39(8):1800–1811. Special Issue: Advances of Operations Research in Service Industry.

Marc D. Joffe and Frank Partnoy. 2018. Making credit ratings data publicly available. *SSRN Electronic Journal*.

Alexandr Karminsky and Ella Khromova. 2016. Extended modeling of banks' credit ratings. *Procedia Computer Science*, 91:201–210.

Peter Kernan, Andrew D Palmer, and Marta Castelli. 2013. Corporate methodology. Technical report.

Hong Sik Kim and So Young Sohn. 2010. Support vector machines for default prediction of smes based on technology credit. *European Journal of Operational Research*, 201(3):838–846.

Kee S. Kim. 2005. Predicting bond ratings using publicly available information. *Expert Systems with Applications*, 29(1):75–81.

Darren J. Kisgen and Philip E. Strahan. 2010. Do Regulations Based on Credit Ratings Affect a Firm's Cost of Capital? *The Review of Financial Studies*, 23(12):4324–4347.

Young-Chan Lee. 2007. Application of support vector machines to corporate credit rating prediction. *Expert Systems with Applications*, 33:67–74.

A.I. Marqués, V. García, and J.S. Sánchez. 2012. Exploring the behaviour of base classifiers in credit scoring ensembles. *Expert Systems with Applications*, 39(11):10244–10250.

David Martens, Tony Van Gestel, Manu Backer, Raf Haesen, Jan Vanthienen, and Bart Baesens. 2009. Credit rating prediction using ant colony optimization. *Journal of the Operational Research Society*, 61.

Alexander B. Matthies. 2013. Empirical research on corporate credit-ratings: A literature review. SFB 649 Discussion Paper 2013-003.

Paul Mizen and Serafeim Tsoukas. 2012. Forecasting us bond default ratings allowing for previous and initial state dependence in an ordered probit model. *International Journal of Forecasting*, 28(1):273–287. Special Section 1: The Predictability of Financial Markets Special Section 2: Credit Risk Modelling and Forecasting.

Cuong V. Nguyen, Sanjiv R. Das, John He, Shenghua Yue, Vinay Hanumaiah, Xavier Ragot, and Li Zhang. 2021. Multimodal machine learning for credit modeling. In *2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC)*, pages 1754–1759.

Martina Novotná. 2012. The use of different approaches for credit rating prediction and their comparison. *Proceedings of the 6th International Conference on Managing and Modelling of Financial Risks*, 165:448 – 457.

Andreas Oelerich and Thorsten Poddig. 2006. Evaluation of rating systems. *Expert Systems with Applications*, 30(3):437–447. Intelligent Information Systems for Financial Engineering.

George E. Pinches and Kent A. Mingo. 1973. A multivariate analysis of industrial bond ratings. *The Journal of Finance*, 28(1):1–18.

J. Ross Quinlan. 2004. Induction of decision trees. *Machine Learning*, 1:81–106.

Jan Roeder, Matthias Palmer, and Jan Muntermann. 2022. Data-driven decision-making in credit risk management: The information value of analyst reports. *Decision Support Systems*, 158:113770.

Yuji Roh, Geon Heo, and Steven Euijong Whang. 2021. A survey on data collection for machine learning: A big data - ai integration perspective. *IEEE Transactions on Knowledge and Data Engineering*, 33(4):1328–1347.

Mark Wallis, Kuldeep Kumar, and Adrian Gepp. 2019. *Credit Rating Forecasting Using Machine Learning Techniques*, pages 180–198. IGI Global, United States.

Mingfu Wang and Hyejin Ku. 2021. Utilizing historical data for corporate credit rating assessment. *Expert Systems with Applications*, 165:113925.

Yong Wei, Shouxiang Xu, and Fanhua Meng. 2010. The listed company's credit rating based on logistic regression model add non-financial factors. pages 172 – 175.

Tsui-Chih Wu and Ming-Fu Hsu. 2012. Credit risk assessment and decision making by a fusion approach. *Knowledge-Based Systems*, 35:102–110.

Liu Yijun, Cai Qiuru, Luo Ye, Qian Jin, and Ye Feiyue. 2009. Artificial neural networks for corporation credit rating analysis. In *2009 International Conference on Networking and Digital Society*, volume 1, pages 81–84.