Norwegian School of Economics
Bergen, Fall 2022

«They'll just go to Moody's»

Investigating Corporate Credit Rating Updates Using Machine Learning Techniques

Synnøve Vigander Mjølhus & Henrik Solheim Holen

Supervisor: Håkon Otneim

Master Thesis, Economics and Business Administration

Major: Business Analytics

NORWEGIAN SCHOOL OF ECONOMICS

«They'll just go to Moody's»

Answered a Standard & Poor's employee in a scene in the movie *The Big Short (2015)*, when asked why they did not insist on higher standards in their credit rating assessments during the Great Financial Crisis of 2008.

Acknowledgements

This thesis was written during the fall of 2022 as part of our MSc in Economics and Business Administration at the Norwegian School of Economics (NHH).

We want to express our sincere gratitude to all those who have contributed to completing our master's thesis in Business Analytics.

First and foremost, we thank our supervisor, Håkon Otneim, for his guidance and valuable feedback throughout the research and writing process. We are truly grateful for the time and effort he dedicated to helping us challenge ourselves to make the thesis the best it could be.

We want to express our gratitude to each other for an exciting semester and a flawless partnership.

Finally, we would like to thank our families and friends for their support and encouragement during the challenges of completing a master's thesis.

Norwegian School of Economics

Bergen, December 2022

Synnøve Vigander Mjølhus

Henrik Solheim Holen

Abstract

Credit Rating Agencies («CRAs») play an important role in the global debt market. They influence the credit spread and thus the borrowing costs for major corporations. An inherent problem is the conflict of interest that arise when the CRAs are paid by issuers. This is not a recent concern, and numerous studies have looked into this and other issues with CRAs. In this master's thesis, we extend this area of research by applying machine learning («ML») models for predicting credit rating updates.

For this task, we construct a prediction model using financial ratios, for which we have 20 years of data for two major agencies; Moody's and Fitch. We also include ratings for an investor-paid agency: Egan-Jones. In the model, we change the *soft factor* in the CRAs' assessment with a new factor that both theoretically and, as will be shown, empirically explain rating updates; trailing stock returns. We apply the XGBoost algorithm to provide more accurate predictions of credit rating updates. Moreover, we analyse SHAP values to interpret different features' contributions to the predictions of rating updates.

We evaluate our approach on a dataset of credit ratings in the US and EU and obtain an accuracy of 84.25%. We find that the total return 12 months before the update is the most important when predicting, which suggests stale credit rating updates. Most excitingly, we find that for CRAs with an investor-paid model, the total return three months before the update is the most important when predicting. For the issuer-paid revenue model, twelve months' total stock return turned out to be important: This suggests that investor-paid revenue models are more proactive in updating credit ratings than issuer-paid agencies.

The model is applied to the rating downgrade of Wirecard in 2020, which allows for an interesting interpretation of local SHAP values. We also discuss the potential limitations of using ML in credit rating predictions, such as loss of interpretability, unreliable accounting data and the sensitivity of SHAP values.

Keywords – NHH, master's thesis, Business Analytics (BAN), Finance, Agency Theory, Corporate Bonds, Credit Ratings, Credit Rating Agencies, Machine Learning, XGBoost

Abbreviations

CRA	Credit Rating Agency
ML	Machine Learning
S&P	Standard & Poor's
Moody's	Moody's Investors Service
Fitch	Fitch Ratings
E-J	Egan-Jones Ratings Company
Big Three	Standard & Poor's, Moody's Investor Services and Fitch Ratings
KMV	Kealhofer, Merton & Vasicek - Model
EDF	Expected Default Probability
IOSCO	International Organization of Securities Commissions
SEC	Securities and Exchange Commission
NRSROs	Nationally Recognised Statistical Rating Organisation
Dodd-Frank	The Dodd-Frank Wall Street Reform and Consumer Protection Act
ESMA	European Securities and Markets Authority
EU	European Union
EEA	European Economic Area
IFRS	International Financial Reporting Standards
CRA3	Third Reform of CRA Regulation
CART	Classification and Regression Trees
XGBoost	Extreme Gradient Boosting Method
SHAP	SHapley Additive exPlanations
MCC	Matthew's Correlation Coefficients
EMH	Efficient Market Hypothesis
GFC	Great Financial Crisis

Contents

List of Figures

List of Tables

1 Introduction

The global corporate bond market amounted to USD 41 trillion in 2020 (ICMA, 2020). In this huge market, Credit Rating Agencies (hereafter «CRAs») play an important role as assessors of credit risk. Credit risk is the probability of default and any indication or perception of an increased probability of default will increase lenders' required rate of return, which means a higher borrowing rate. Investors use ratings for their assessments of risk. Some investors cannot invest in bonds with a certain level of default risk, called *non-investment grade bonds*. Obviously, the CRAs must be independent and unbiased in their assessments.

Nevertheless, CRAs have been criticised for being prone to conflict of interest. As long as the issuers pay for the rating, suspicion about client pressure will arise. Studies have shown how this is problematic. Moreover, CRAs have been criticised for being «too late» on rating actions, and not providing the market with anything it did not already know. A significant part of the criticism is directed towards CRAs being a «lagging indicator» instead of a leading indicator (Nye, 2014).

In this thesis, we use machine learning techniques to address these issues. We predict a multiclass classification problem on credit rating updates using financial accounting data. Unlike other studies, we predict credit rating updates of corporate bonds and not the alphanumeric credit rating level. Moreover, we investigate whether CRAs in fact *lag* the market, by including the trailing total returns of the issuer's stock for different periods. This has, to our knowledge, not been investigated before. Additionally, we also look for any differences between the two revenue models; the investor-paid model and the more traditional issuer-paid model.

For the technical part, the machine learning model, we use the Extreme Gradient Boosting algorithm (hereafter «XGBoost»). The XGBoost model is preferred in our multiclass classification problem because it is computationally superior and manages problems with multicollinearity when using financial data. Using the XGBoost model, we obtain good results, with an accuracy of 84.40%. By adjusting for the imbalance in the data of the three classes, we obtained a Matthews Correlation Coefficient of 66.23%.

We apply SHAP values to interpret the model's results. Using global and local SHAP values to interpret credit rating updates is unprecedented and a relatively new way to explain machine learning models. The SHAP values indicate that the *Trailing 12 months total return* feature is most important when predicting credit rating updates, providing tangible evidence of the allegations of CRAs' actions lagging the stock market. Moreover, we find it surprising that previous studies have not included the issuer's stock return in their machine learning models, as it has good predictive power. Maybe not surprising, but still interesting, we find that, by looking at SHAP values, positive values for an issuer's *Trailing 12 months total return* contribute to a credit rating update being predicted as an *Upgrade*, and negative values for *Trailing 12 months total return* predict a *Downgrade.*

To contribute to the endless dispute about which revenue model is preferred, we split the data set by revenue models. We obtain exciting results when investigating the two revenue models separately. The SHAP values for the investor-paid revenue model indicate that *Trailing three months total return* contributes the most to the predictions, which is a meaningful nine-month difference from the issuer-paid model.

Lastly, we use SHAP values in local explanations to investigate the Wirecard Scandal of 2020. As the last downgrade of the infamous German fintech company Wirecard is part of our data set, we use local SHAP values to see which features contribute most when predicting the update. The model predicted that the case belongs in the *Downgrade* class with a 94% probability, which turned out to be correct, as the bond was downgraded on the 19th of June 2020.

Moreover, we were curious to see if our model could outperform Moody's assessment prior to Wirecard's default in June 2020. Using the last quarter's financial data, we created a synthetic case one month before the last case. Our model's predicted probability of the synthetic case belonging to class *Affirmation* is 69%. The predicted probability for class *Downgrade* and *Upgrade* is 50% and 1%, respectively, meaning our model could not affirmatively predict a downgrade before Moody's. Although, it is essential to keep in mind that our model is trained on Moody's data, which could cause our results to be biased and closer or further from the actual unknown truth.

Nonetheless, this highlights our most significant challenge in this thesis and credit ratings in general. Credit ratings are established by private companies and cannot be benchmarked to an actual observable truth, which creates a paradox when using machine learning models to predict credit ratings.

1.1 Motivation

Surprisingly, no prior studies use SHAP values to interpret credit rating updates and detect possible lags. Understanding the mechanisms of credit ratings is highly important to prevent future credit rating scandals, and we believe this work makes a contribution to the field of study.

Moreover, as leverage has increased and credit quality worsen, good models for credit risk are essential for lenders and regulators. We think the demand for more and better credit risk models will increase and we believe machine learning techniques can bring promising progress in this respect.

Machine Learning (hereafter «ML») methods are known for their ability to predict accurately. Until recently, however, predictions have been hard to interpret due to the model complexity in ML. New model-agnostic methods, like SHAP values, have made it possible to interpret black-box models. By applying relatively new ML methods and new interpreting methods, such as SHAP values, we give new insights into credit rating updates.

1.2 Scope and Limitations

The global bond market is large and complex, and some limitations were necessary for accurate prediction and thorough analysis. The thesis scope was narrowed down to analysing the European and US bond market as there are many similarities between the two markets and economies. Countries that follow the EU regulations are included in the European data, comprising countries in the European Union, the European Economic Area, the United Kingdom, and Switzerland.

This thesis is limited to bonds issued by corporations, known as corporate bonds. We only look at companies with publicly available financial data, which primarily means companies listed on a stock exchange. The financial accounting data we use exists for corporations only and cannot be derived for other bond issuers, such as governments and municipalities, and are, consequently, not included.

A vital part of our analysis uses different trailing total returns as features. Hence the scope is limited to corporate bonds publicly listed on global stock exchanges. Financial corporations, such as institutions, banks, asset managers, and insurance companies, are excluded as they are covered by different regulations and have a separate credit rating assessment.

Unfortunately, due to access restrictions, we could not access data from Standard & Poor's (S&P). In Refinitiv, our primary data resource, the exportation of S&P ratings is now a separate chargeable service, meaning S&P is not part of the final data set used in this thesis. The final data set thus comprises Moody's, Fitch, and Egan-Jones. Egan-Jones has a different revenue model than the Big Three (Moody's, Fitch, and S&P). While not a major CRA, Egan-Jones is included to study potential differences in the revenue models and credit rating updates. Data from Egan-Jones is only included from 2016 and onwards, as they did not provide ratings to Refinitiv before 2016.

1.3 Structure

This thesis is structured in seven parts. Part 1 introduces the reader to the current challenges with CRAs and why it is interesting to investigate this topic. Part 2 explains CRAs' past and current role and importance. Part 3 reviews similar literature, relevant financial theory, and previous events and describes regulatory guidelines. Part 4 gives an understanding of the applied methodology before the data and data modelling is introduced in Part 5. Lastly, the results are analysed and discussed in Part 6 before concluding in Part 7.

2 Background

2.1 From Railroads to Ratings

The reputation of the CRAs is described as the agencies' most competitive advantage and stems from a long history of an oligopolistic market structure (Nye, 2014). Significant barriers to entry, reputation, international regulation, and high profits have protected the CRAs against newcomers and innovation.

Modern corporate bonds have existed since the early 17th century. They were first issued to the public in the Netherlands to fund the Dutch East India Company in 1623 (Gelderblom et al., 2013). Later, during the 19th century, corporate bonds became a critical funding instrument during the construction of the North American railroad system, as small local banks were unwilling and unable to give credit to US Railroad companies (Sylla, 2002).

Historically, debt obligations mainly existed between parties who knew each other on interpersonal levels or between financial institutions. As bonds gained popularity and geographical range, investors needed assurance of a corporation's creditworthiness and ability to pay back its debt obligations. The need for assurance led to the birth of Credit *Reporting* Agencies, such as The Mercantile Agency, founded in 1841, which collected and sold information on the creditworthiness of US businesses.

As the US corporate bond market was more advanced than in most countries, it was no surprise that a Credit *Rating* Agency was introduced by the American John Moody in 1909, roughly 60 years after the first corporate bond issuance. In the same period, the editor of *The American Railroad Journal*, Henry Poor, collected data from 1849-1862 on railroads' property, such as their earnings, assets, and liabilities. The data later turned into the *Manual of the Railroads of the United States*, which reported operating and financial data on most American railroads. John Moody and Henry Poor were pioneers in the credit rating industry, and their companies are still the two largest rating agencies in the world (Sylla, 2002).

2.2 Fundamentals of Credit Ratings

2.2.1 Corporate Bonds

A bond is a debt obligation typically issued by corporations or governments, such as federal governments, municipalities, or agencies. However, in this thesis, we only study corporate bonds, and for the sake of understanding later nomenclature, we define and explain the fundamentals of the financial instrument, and the following paragraphs aim to provide some necessary insights.

Today, corporations (issuers) frequently use bonds to raise finance investments and operations. Bonds are more attractive to issuers as bank financing is usually more expensive. Investors, i.e. the lenders, receive an interest (coupon) compensating for the time value of money and risk. When bonds reach their maturity date, the original investment (par) is returned to investors, and future payments terminate (Bodie et al., 2022).

Institutional investors, such as mutual funds, pension funds, and insurance companies, are typically large holders of corporate bonds. Institutional investors hold investments of significant sizes, as these, again, are aggregated by many smaller investors. Institutional investors are often mandated by law or internal standards to base some of their investment choices on credit ratings of reputable CRAs, as a practice to minimise risk for the smaller investors (Booth and de Bruin, 2019). Usually, these investors are required to invest in corporate bonds rated *investment grade* only. However, there are also debt funds that only invest in what can be classified as *non-investment grade* bonds, typically referred to as *high yield* or *junk bonds*, see Table 2.1. A bond being downgraded from investment grade to junk is called a *fallen angel.* To issuers, receiving a *fallen angel* label could be harmful as it could require institutional investors to sell their bonds as some investment mandates restrict what they are allowed to hold (Belloni et al., 2020).

In our data set, which consists of 37,873 credit rating updates, 56.63% are classified as investment grade, while 43.37% are classified as non-investment grade. Of the credit rating updates, 1.62% can be classified as fallen angels, while 1.64% of the updates can be classified as a *rising star*, meaning the credit rating was upgraded from non-investment grade to investment grade.

2.2.2 Corporate Bond Market

Even though corporate bonds have existed for a long time, there has been an increase in popularity in recent years. Since 2000, the aggregated financing market has increased substantially more than bank lending in Europe (Darmouni and Papoutsi, 2022).

In 2020, US corporate bond issuance hit an all-time high of USD 1.919 trillion (Rennison, 2020). Similarly, there has been an increasing surge of smaller issuers to the European bond market since 2008, making the nominal value of the market EUR 12.12 trillion (Darmouni and Papoutsi, 2022). The global corporate bond market is estimated to be USD 41 trillion, dominated by the US with USD 10.9 trillion and China following up with USD 7.4 trillion (ICMA, 2020). In 2015, bond financing represented 34.66% of US corporations' total debt, relative to only 19.70% in Europe (Darmouni and Papoutsi, 2022). Accordingly, it is appropriate to label the European financial system as *bank-based* and the American as *market-based.*

The rise of bond financing has been supported by policymakers, as diversification of sources of funds shields corporations from shocks in the banking industry. The other side of the coin is that the frequency and amplitude of rating scandals have increased, making it valuable that we contribute to understanding the mechanism of Credit Rating Updates.

2.2.3 Credit Ratings

Credit ratings are standardised measures to assess the *probability of default* by the issuer. Ratings are ranked using an alphanumeric letter grading scale, where A is the premier and D is the default. There are 21 to 24 combinations of the letters, each representing a level of credit risk. See Table 2.1 for a complete comparison.

There are three fundamental functions of credit ratings. A credit rating has an economic function to objectively measure an issuer's credit risk while eliminating essential information asymmetry between the issuer and investors, making it easier for issuers to access funding. Secondly, it offers a standardised comparison method for investors over a range of corporate bonds, helping them manage their portfolios. Lastly, they give market participants a uniform standard for discussing credit risk. The three functions aim to ensure that ratings should be independent and verifiable (OECD, 2010), although as we will discuss in Section 3, this has not always been the case.

Issuers can request public, private, or confidential ratings from the CRAs. After a public rating is assessed, the rating is often released as a stock exchange announcement and distributed directly through the CRAs' channels. A private rating, on the other hand, is distributed to parties designated by the issuer, and confidential ratings are for internal usage (by the issuer) only. Private and confidential ratings can later be reclassified and published as public ratings upon the issuer's request, creating an option to conceal unfavourable ratings. As an issuer is not obliged to disclose unfavourable ratings publicly, it creates an option for a potential bias and skewness in credit rating distribution.

In Figure 2.1 we can observe some skewness in our data, where the majority of our observations are investment grade (green bars), and the minority is junk bonds (blue bars).

Figure 2.1: Alphanumeric Value of Credit Rating Updates in the Analysed Data

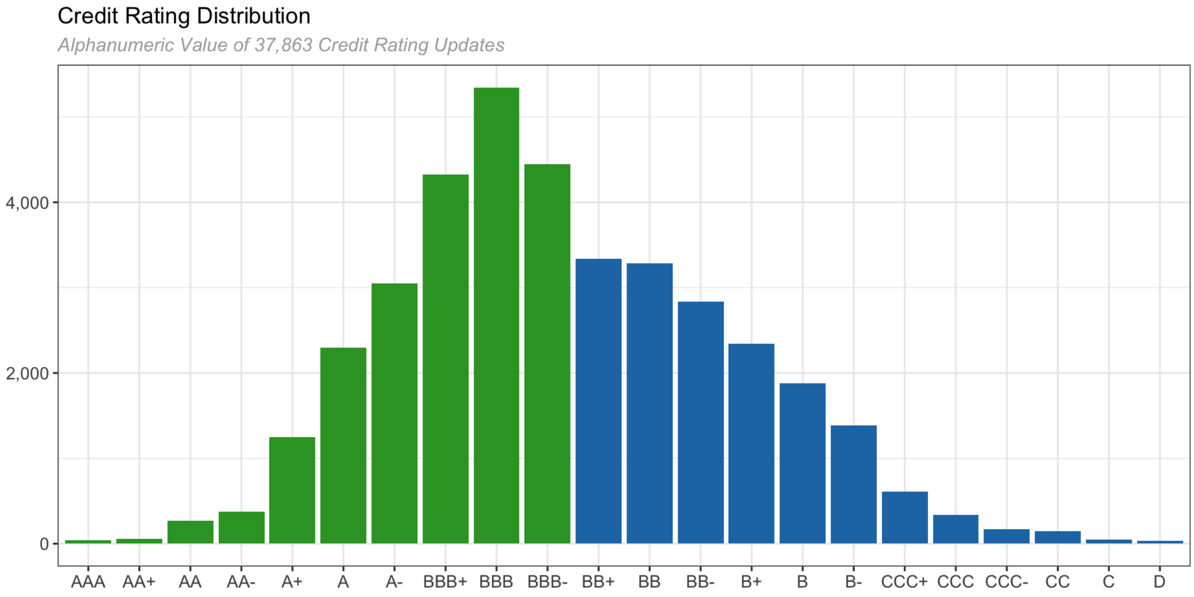


Table 2.1: Comparison of Alphanumerical Letter Grades in Credit Ratings

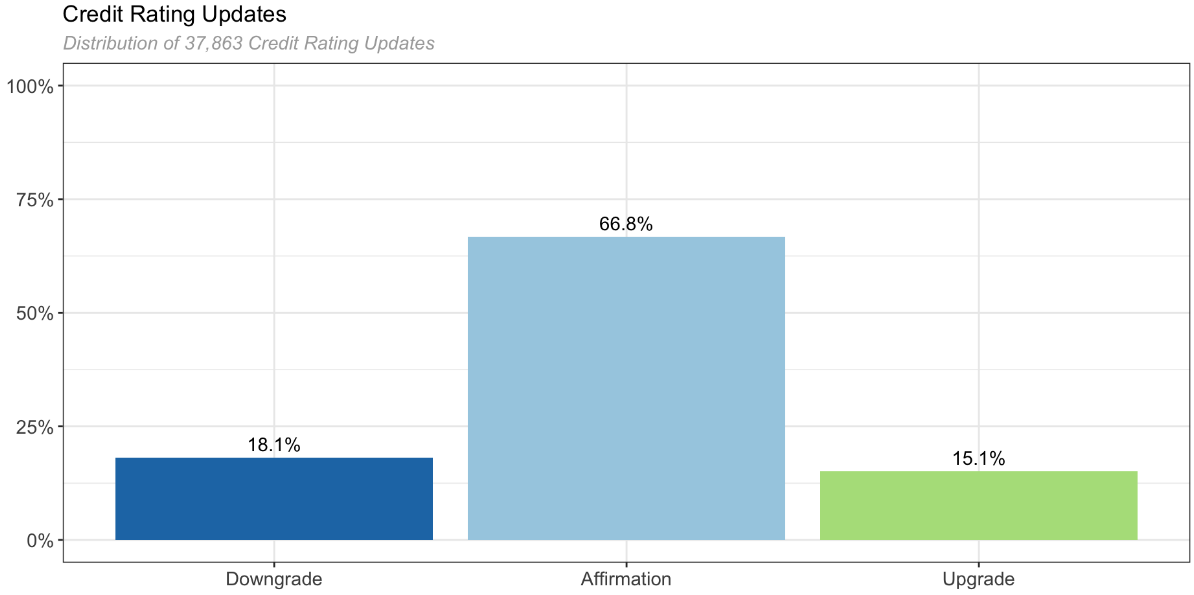
		S&P	**Fitch**	**Egan-Jones**	**Moody's**	**KMV**[1]
Investment Grade	Highest quality	AAA	AAA	AAA	Aaa	0.02%
	Superior quality	AA+ AA AA-	AA+ AA AA-	AA+ AA AA-	Aa1 Aa2 Aa3	0.05% 0.08% 0.13%
	Good quality	A+ A A-	A+ A A-	A+ A A-	A1 A2 A3	0.21% 0.32% 0.42%
	Medium quality	BBB+ BBB BBB-	BBB+ BBB BBB-	BBB+ BBB BBB-	Baa1 Baa2 Baa3	0.54% 0.70% 0.90%
Junk	Low, medium quality	BB+ BB BB-	BB+ BB BB-	BB+ BB BB-	Ba1 Ba2 Ba3	1.15% 1.48% 2.09%
	Low quality	B+ B B-	B+ B B-	B+ B B-	B1 B2 B3	2.95% 4.14% 6.66%
	Bad quality	CCC+ CCC CCC-	CCC+ CCC CCC-	CCC+ CCC CCC-	Caa1 Caa2 Caa3	10.64% 17.00% 17.95%
	Speculative	CC	CC	CC	Ca	20.00%
	Default	C D	C DDD DD D	C	C	20.00% >20.00%

[1] Expected Default Probability according to Moody's KMV model (Langohr and Langohr, 2009)

2.2.3.1 Rating Updates From time to time, CRAs update their outstanding ratings. In the analysis, we will not predict the alphanumeric letter grade of a corporate bond rating. Instead, we are more interested in the actions of the CRAs by investigating credit rating updates. Several studies have already utilised machine learning models to predict the alphanumeric letter grade, and Section 3.5 reviews several of these studies.

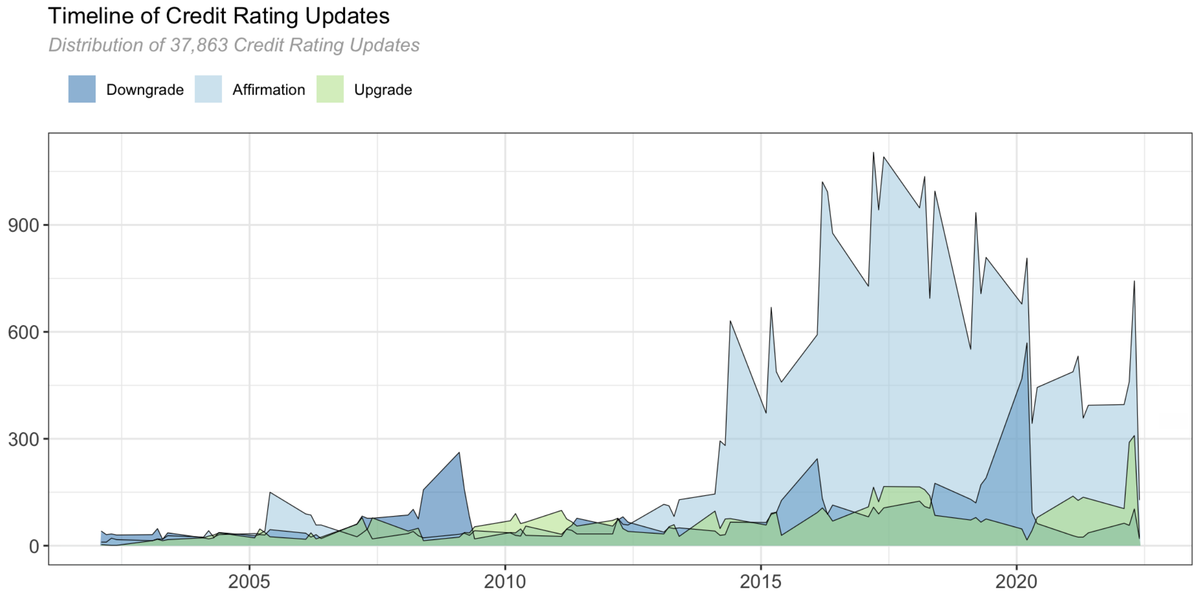
A credit rating receives an update when the CRAs opinion of its risk changes in the long-term perspective. A credit rating can either receive an update as (1) *Downgrade*, meaning the credit rating worsens, (2) *Affirmation*, meaning the rating does not change or (3) *Upgrade*, meaning the credit rating is upgraded. A bond receiving either an upgrade or downgrade can be critical for an investor who is a bondholder, as it confirms that the bond's risk has shifted.

Figure 2.2: Distribution of the Credit Rating Updates in the Analysed Data



As the credit risk of corporations normally is quite stable, most credit rating updates are *Affirmation*, which is reflected in the distribution of updates in the analysed data, illustrated in Figure 2.2. Rating updates tend to be cyclical, meaning the number of downgrades increases in recessions or crises, and vice versa for updates. In Figure 2.3 of our data set, we see an increase in the number of rating updates being *Downgrades* in 2008 and 2020, which were both years of economic recessions. In most cases, *Downgrades* and *Upgrades* are just changing the credit rating between one or two notches on the alphanumeric rating scale.

Figure 2.3: Timeline of the Credit Rating Updates in the Analysed Data



2.2.3.2 Rating Outlooks and Watchlist Along with credit ratings and long-term updates, the CRAs also announce *rating outlooks* and can put outstanding credit ratings on a *rating watchlist*. The CRAs do not always place ratings on the watchlist prior to a rating change, and likewise, for an upgrade or downgrade, they are not always preceded by a positive or negative outlook. The CRAs may announce that they issue an outlook instead of a rating update under uncertainty.

Rating Outlooks are opinions of the likely direction of credit ratings from a medium-term perspective. The CRAs use rating outlooks to express their opinion on the expectation of a rating change and, often, the likely direction of that rating. The outlooks focus mainly on special events, such as merges and recapitalisations, that call for additional monitoring.

Rating Watchlist represents an opinion of credit ratings that are under review for possible changes in a short-term perspective. When downgraded, affirmed, or upgraded, a credit rating is withdrawn from the watchlist.

2.2.4 Credit Risk

Credit risk is the likelihood that the borrowing corporation cannot fully meet its debt obligations, which can cause a default. Credit ratings describe the issuer's credit risk and guide investors to set a meaningful interest rate for compensation for the probability of default. The following paragraphs are included to give some theoretical background of credit risk, which is the underlying foundation of credit ratings.

2.2.4.1 Five C-s To give an understanding of the basics of credit risk, we introduce the *Five C-s*. The Five C-s are fundamental factors when lenders assess a borrower's creditworthiness. The fundamentals attempt to evaluate the likelihood of default and, subsequently, the risk of a financial loss for the lender.

1. *Character*, Creditworthiness and reputation of the borrower. Factors such as credit history, financial stability, the integrity of management
2. *Capacity*, Borrower's ability to repay the debt, factors such as business model, competitive advantages, and financial projections
3. *Capital*, Borrower's financial resources, equity cushions such as cash reserves
4. *Collateral*, Assets that the borrower is willing to pledge to the lender
5. *Condition*, Economic, market and business environment conditions, such as industry trends, interest rates, and state of the economy.

2.2.4.2 Probability of Default *Default* is the least desired situation for issuers and investors. Default occurs when a borrower can no longer fulfil the required payments on a loan or debt obligation due to a lack of resources or an unwillingness to pay. *The Probability of Default* is the likelihood that default occurs.

In credit risk theory, there are three main branches of models for assessing the probability of default: (1) *Structural models* utilise the direct relationship between capital structure and default risk, (2) *Accounting models* use accounting data to assess the probability of default, and (3) *Hybrid models* are a combination of the former two (Wagner, 2008).

2.2.4.3 Merton Model The *Merton model* is one of the most common structural models used to assess credit risk by evaluating the direct relationship between a corporation's capital structure and default risk (Merton, 1974). The model is commonly used by stock analysts, commercial loan officers, and others to assess a company's credit default risk. We mention it here because we will introduce equity volatility (stock returns) as an important feature in our model later.

According to Merton's model, a company's credit risk depends on its assets, liabilities, and their volatility. A version of the Black-Scholes option pricing formula determines the probability of default. In short, the model evaluates a company's structural credit risk by modelling its equity as a European call option on its assets.

Merton's Model - Value of Equity

The theoretical value of a company's equity E is given as:

$$E = V_0 \Phi(d_1) - Be^{-rT} \Phi(d_2) \tag{2.1}$$

$$d_1 = \frac{(\ln \frac{V_t}{B}) + (r + \frac{\sigma^2}{2} \Delta T)}{\sigma \sqrt{\Delta T}} \tag{2.2}$$

$$d_2 = d_1 - \sigma \sqrt{\Delta T} \tag{2.3}$$

where V_t is the value of the company's assets in period t. B is the value of the corporation's debt. Φ is a function of the cumulative standard normal distribution. r is the risk-free interest rate. σ is the standard deviation of the company's stock. The Merton Model assumes there are no dividends or share repurchases and that the issued debt is a zero-coupon bond.

Merton's Model - Expected Default Frequency

The Expected Default Frequency (hereafter «EDF») is the probability that a given firm will default within a year. Merton defines the one-year probability default of a firm as the probability that in one year, the asset value V_1 will be below threshold B. Threshold B is usually reflecting the liabilities payable within one year. Using the symmetry principles of the Gaussian distribution and arranging the terms of Formula 2.1, the EDF is given as:

$$EDF_{Merton} = P(V_1 \leq B) = \bar{\Phi}\left(\frac{\log{(V_0)} - \log(B) + (r - \frac{1}{2}\sigma^2)}{\sigma}\right) \tag{2.4}$$

The model assumes that a firm's assets are invested in a stock and bond portfolio that follows a log-normal distribution. This indicates that although the value of the assets is log-normally distributed, the returns on the assets are normally distributed. The model also assumes that the company's liabilities are known and constant and can always pay its debts off by liquidating its assets.

Considering these assumptions, the model determines the likelihood that a company will fail to make its debt payments. If this likelihood is high, the company is thought to be at a high default risk, and investors may be reluctant to lend to the company or buy its assets. Overall, Merton's model offers a helpful framework for assessing a company's or financial asset's credit risk. Still, it is vital to keep in mind that the model rests on a number of assumptions that makes it hard to implement in real life. In the next section, however, we will briefly look at one application of Merton's model.

2.2.4.4 Moody's KMV Model The *KMV model* was introduced in the late 1980s by the company KMV, later acquired by Moody's. Moody's KMV is an industry model derived from Merton's model. Many financial institutions use this model to assess the probability of default. The KMV model does the following:

1. Estimate the value and volatility of the firm's asset, preferably using stock returns
2. Calculates an index distance measure of default risk using Merton's model
3. Scale the distance to default to probabilities using a proprietary default database

Merton's model and KMV differ because the latter uses a large database. In short, the KMV tries to overcome some of the weaknesses of Merton's model by replacing the log-normal distribution with Moody's proprietary distribution. The KMV model also considers not only the volatility of a firm's assets but also the volatility of its liabilities. Merton's model assumes there are no payments before the bond expires, while KMV takes intermediate payments into account.

Merton's model assumes that default only happens at maturity date T, while KMV relaxes the assumption and allows for default before maturity. To arrive at this threshold, KMV uses empirically observed probabilities of default from the market.

Merton's model and KMV best assess publicly traded companies whose assets are valued based on market value. The models use equity value to assess default risk, meaning for the market value, it uses the price of the issuer's stock times the number of outstanding shares. These widely used structural models show the importance of the stock returns when assessing credit risk and indirect credit ratings, which are therefore used as essential indicators in our analysis.

2.2.5 Credit Rating Agencies

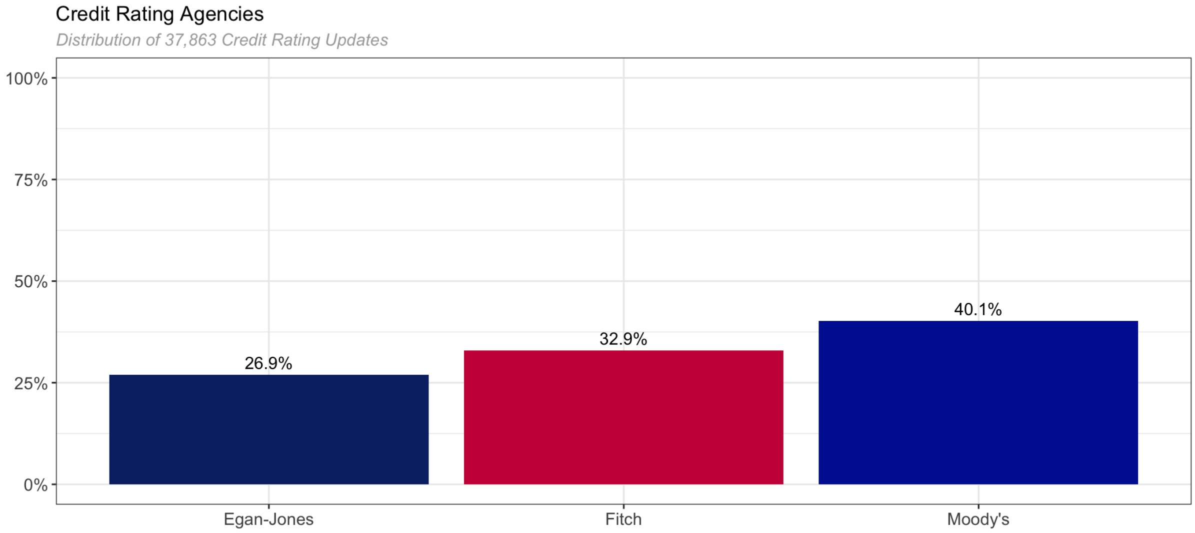
Credit Rating Agencies issue credit ratings and rating updates, where the opinions of only a handful of agencies are recognised. As explained in Section 2.1, some agencies have a long and strong history, and for a long time the market was duopolistic, with Standard & Poor's and Moody's as the only providers of credit ratings.

Today, the credit rating market in the United States and Europe resembles an oligopolistic market and has characteristics of high entry barriers, price rigidity, and non-price competition. The providers Standard & Poor's (hereafter «S&P»), Moody's Investors Service (hereafter «Moody's») and Fitch Ratings (hereafter «Fitch»), known as the *Big Three*, collectively have a market share of over 90% in the United States and Europe (Refinitiv, 2022). These three agencies are private for-profit companies and had estimated annual revenue of roughly $17 billion in 2020 (Refinitiv, 2022).

S&P is the largest provider of ratings in Europe, with a market share of 51.77% (ESMA, 2021). Following S&P are Moody's and Fitch, with European market shares of 30.12% and 10.30%, respectively. The Big Three dominate the US market, where S&P has the most significant market share of 43.9%, with Moody's share at 25.8%, and Fitch at 16% (SEC, 2022). The market share is calculated from the revenue of corporate credit rating activities.

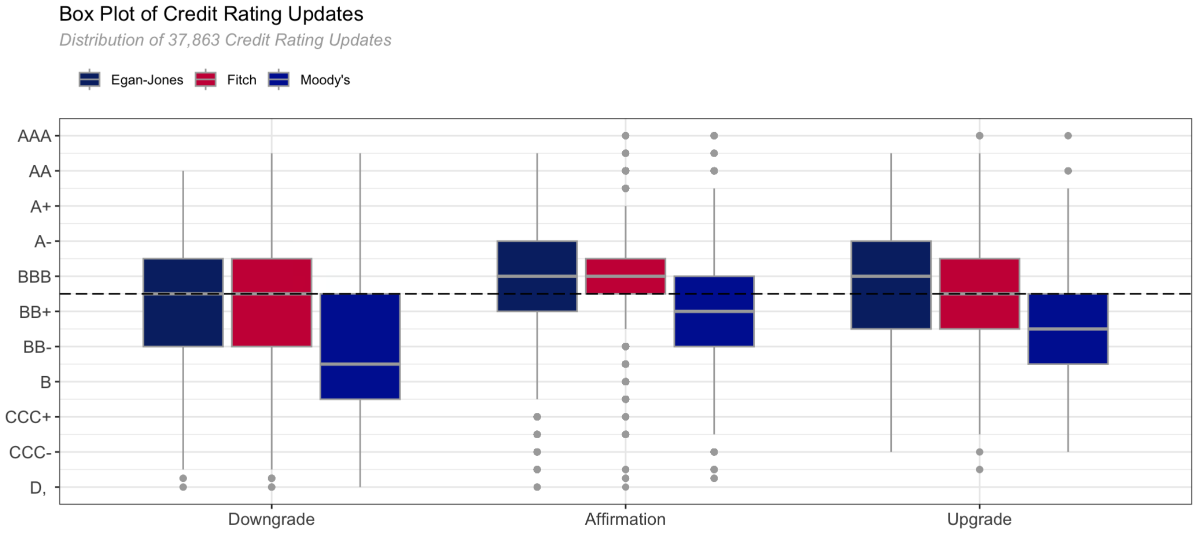
There are several smaller CRA behind the Big Three in Europe and the United States. Around the turn of the millennium, many new and smaller credit rating agencies were established. Of these smaller agencies, *Egan-Jones Ratings Company* (hereafter «E-J») was founded in 1995 and received recognition following the aftermath of the financial crisis. E-J was founded by Sean Egan, an investment banker and consultant, launching E-J as a challenger to the Big Three. E-J has been an outspoken critic of the Big Three, describing his company as a challenger that is not intimated by large corporations and has previously been known to give less favourable ratings to large corporations than their competitors (Lucchetti, 2008). E-J is different from the Big Three and operates with an investor-paid revenue model, and E-J argues that its revenue model reduces the infamous conflict of interest problem.

Figure 2.4: Share of Credit Rating Agencies in the Analysed Data



In Figure 2.5, the ratings in our data set are presented in a box plot, with the dotted line being the investment grade boundary. The box plot shows that Moody's ratings are relatively lower than Fitch and E-J. The box in a box plot represents the interquartile range, which is the range of the middle 50% of the data. The line within the box illustrates the median value of the data, which varies for each class and CRA. The whiskers of the box plot show the minimum and maximum values of our ratings, not including outliers.

Figure 2.5: Box Plot Analysis of the Credit Rating Updates in the Analysed Data



2.2.6 Revenue Models

The two primary revenue models used by the CRAs are *issuer-paid* and *investor-paid*. There is no consensus on which model is preferred or the better, and this will be discussed in Section 3.3.1 and investigated using machine learning models in Section 6.4.3.

Between 1900-1970, the investor-paid model dominated. Investors purchased ratings and relevant intelligence originally in journals from the CRAs. Between 1970 and 2001, most CRAs switched to an issuer-paid model, which is used by the Big Three today.

In an issuer-paid model, illustrated in Figure 2.6, the revenue stream comes from issuers engaging the CRAs for a rating on their bond. The rating is then made public, and investors can choose to invest in the bond based on the rating assessment. If an investor invests in the bond, the issuer gets financing. In an Investor-Paid model, illustrated in Figure 2.7, the investors pay the CRAs to access their ratings. The rating is assessed independently from the influence of the issuers.

Figure 2.6: Issuer-Paid Revenue Model

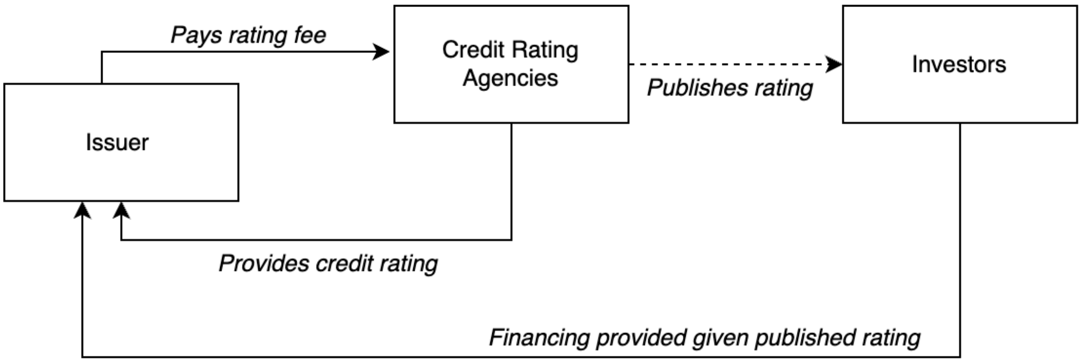
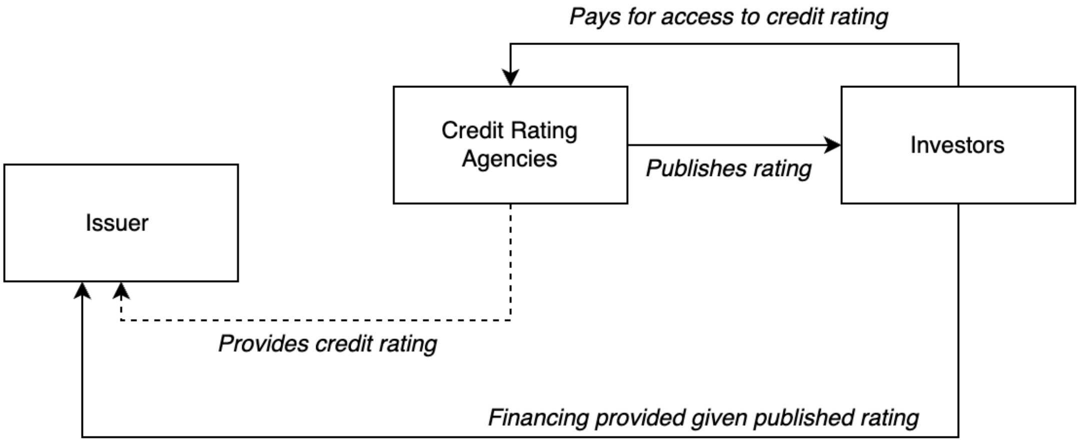


Figure 2.7: Investor-Paid Revenue Model



2.2.7 Assessment

In this section, we review the assessment methods of Moody's, Fitch and E-J to ensure that our machine learning model (1) has reasonable features, (2) is realistic, and (3) is a viable alternative to the assessments done by the CRAs. We find that the rating methodology of the CRAs shares similarities, most likely due to regulations, competition, and data availability.

In the issuer-paid model, a bond will not be rated unless instituted by an issuer hiring CRAs to assess its bond. Moody's evaluates qualitative and quantitative factors such as scale, business profile, financial accounting ratios, and the company's financial policy (Moody's, 2022). Moody's rating methodology varies depending on the bond class and sector, where a rating committee applies subjective judgments to emphasise each rating factor. A thorough review of Moody's assessment methods is provided in Appendix A1.

After reviewing a bond, analysts present the rating recommendations to a *rating committee*. The rating committee then assesses the factors most likely to affect credit risk in the sector. The rating committee compromise of five to nine analysts, where each committee member gets one vote, and the majority vote decides the ratings. The rating decision can take anywhere from 30 minutes to several days for more debatable ratings (Nye, 2014, p. 131). The assessment method for Fitch is like Moody's, and they stress that the most significant measures when they rate corporate bonds are debt service, short-term liquidity, leverage, and coverage ratios.

For CRAs with an investor-paid revenue model, the assessment is instigated regardless of the issuer's interest. Recall investors pay CRAs for rating potential investments, which means a rating can be published shortly after the bond issuance in the interest of investors. In investor-paid CRAs' assessment, there is usually no access to non-public information from the management of the issuers. Like the issuer-paid model, the assessment uses qualitative and quantitative methods.

E-J state that they mainly use the Five C-s, reviewed in Section 2.2.4.1, when assessing ratings. For corporate bond ratings, E-J uses a monthly quantitative screening method called Rating Change Anticipator (hereafter «RCA») to identify possible changes in credit quality. However, the details of the RCA method are not disclosed publicly. E-J does not disclose its ratings to the issuer unless requested (Egan-Jones Ratings Company, 2020). If the issuer disagrees with the rating, the issuers must provide a written report to support their objection.

In Section 6.4.3, we investigate if there is a lag difference between the two revenue models, and one element in the analysis is how assessments are instigated.

2.3 Regulation

International Regulation: IOSCO The International Organisation for Securities Commissions («IOSCO») is an association of securities regulators. One of their objectives is to issue the Code of Conduct for the CRAs. The Code of Conduct has four objectives: (1) Ensure quality and integrity of the rating process, (2) Ensure independence and reduce conflict of interests, (3) Promote transparency and timeliness of rating disclosures, and (4) Properly secure confidential information. This Code of Conduct is implemented in the local regulations; however, the local regulators only impose that the CRAs should apply it voluntarily without any enforcement mechanisms (European Commission, 2016).

Regulation in the European Economic Area: ESMA The European Economic Area («EEA») follow the same regulation imposed by the European Securities and Markets Authority («ESMA»). The supervision of CRAs in the United Kingdom is conducted by the Financial Conduct Authority («FCA»), and still, post-Brexit, follows EU regulations. One of the objectives of the European CRA regulation is to increase competition in the markets for credit ratings by encouraging issuers to use smaller credit rating agencies by requiring issuers to consider appointing at least one small CRA. Considering the Financial Crisis of 2008, the European Parliament and the Council introduced a new regulation on CRAs just months after the crisis hit (European Union, 2019). This regulation was again updated in 2013 when a new EU reform was introduced, called *Third Reform of CRA Regulation* (hereafter «CRA3»). CRA3 aims to address the over-reliance on external credit rating and potential conflict of interest (European Commission, 2016).

Regulation in the United States: SEC The Securities and Exchange Commission of the United States («SEC») has a long history of regulating CRAs. In 1975, the SEC introduced the concept of Nationally Recognised Statistical Rating Organisations («NRSRO»). Today there are nine CRAs with NRSRO status, including S&P, Moody's, Fitch, and Egan-Jones. In 2006, merely months before the Financial Crisis, the *Credit Rating Agency Reform Act* was introduced in the United States. The reform was intended to improve the quality of the CRAs to protect investors by promoting accountability, transparency, and competition. The reform granted the SEC authority to examine the CRAs with NRSRO status, making the CRAs required to disclose specific information, such as conflicts of interest, rating performance, and internal procedures.

3 Literature Review

Section 3 reviews several studies to assess the most critical and relevant problems with credit ratings and CRAs. Most literature criticises the CRAs' role in past events, such as the lack of competition, agency problems, and asymmetric information problems. Relevant financial literature is also reviewed in this section, as it is highly relevant for the feature selection in Section 5.2.2.

3.1 Efficient Market Hypothesis

Fama (1970) introduced the Efficient Market Hypothesis (hereafter «EMH»), and according to this theory, assets are priced according to all available information in the market at any given time. There are three underlying assumptions for market efficiency: (1) investors are rational, (2) if investors are not rational, their random trades will be cancelled, and (3) all arbitrage opportunities will be used (Fama, 1970). In an efficient market, it is impossible to profit from publicly available information.

EMH states that asset prices reflect all information. The hypothesis argues that, for example, stocks trade at a fair market value on a stock market. There are three forms of the EMH: (1) *Strong form* of EMH states that all information in a market is reflected in the asset price; this includes both public and private information. (2) *Semi strong* form states that all public information is reflected in the asset price. (3) *Weak form* states that all past prices of a specific asset are reflected in today's asset price.

According to Easley et al. (2010), abnormal stock returns can be explained by private information, violating the strong form of EMH. Their study creates a measure of private information and a portfolio consisting of long on stocks with a high value of private information and short on stocks with a low value, this portfolio yields significant abnormal returns. Based on the abnormal returns from the trading strategy of the private information measure, the study argues that information is an important determinant of asset returns, even in the presence of other explanatory factors.

Returns on financial assets are directly linked to credit ratings, as credit ratings are an information indicator. For example, the price of a corporate bond is both influenced and influences the credit rating. The issuers' stock prices can influence the credit ratings, or the credit ratings can influence the stock prices. Hite and Warga (1997) study how credit rating changes affect bond pricing performance, and their results show that credit ratings often trail the market's pricing of the bond. More specifically, on average, the prices of the bonds decrease prior to a rating downgrading.

3.2 Asymmetric Information Theory

The economic theory of asymmetric information was developed in the 1970s to explain market inefficiency (Akerlof, 1970). The theory suggests that an imbalance of information between buyers and sellers can cause market failure. Sellers are believed to hold more information on the product quality than the buyer, which can cause a discrepancy regarding the price and added value. Historically, CRAs were created to solve the asymmetric information problem between issuers and investors in the days before open financial accounting data was accessible to the public (Sylla, 2002).

3.2.1 Asymmetric Information between Management and CRAs

There is an asymmetric relationship between CRAs and the issuer's management, as the management can selectively disclose information to the CRAs in credit rating assessments. Kothari et al. (2009) and Lougee and Marquardt (2004) find evidence that managers of corporations choose to disclose information strategically to third parties, such as the CRAs. These studies find that managers often withhold bad news and emphasise the good news.

Ahn et al. (2019) study the asymmetric relationship between the issuer's management and the CRAs, using data on CRAs. They find evidence that issuers are more hesitant to give less optimistic information to CRAs in their private communications, which is reflected in credit rating levels. The study uses linear regression to review if CRAs react less to negative public information, like earnings announcements, when they have access to private information. The study reveals that S&P, which has an issuer-paid revenue model, is «relatively more likely to downgrade a firm's rating before an earnings announcement», while Egan-Jone, which has an investor-paid revenue model is «relatively less likely to downgrade a firm's rating after an earnings announcement».

3.3 Principal Agent Theory

In the principal agent dilemma, a conflict of interest can arise between a party and the agent appointed to act on their behalf (Eisenhardt, 1989). The principal agency problem can occur when CRAs act as agents for investors and issuers and favour issuers over investors. Investors want CRAs to give accurate ratings, and issuers expect the best possible ratings. When both parties aim to maximise their economic benefits while having different goals, it causes a conflict of interest (Berk and DeMarzo, 2019). Sinclair (2005) claims that an issuer-paid revenue model incentivises cooperation between issuers and CRAs.

3.3.1 Client Pressure and Conflicts of Interest in Credit Ratings

Morgenson (2008) investigates the incident where Moody's was pressured by its client, Countrywide Financial, to withdraw a rating. In 2005, Moody's assessed securities issued by Countrywide Financial, the largest mortgage lender in the US. However, these ratings were changed the next day as Countrywide meant the assessment was too harsh. The rating was changed regardless of any new and significant public information. Leading up to the financial crisis, Moody's had assigned a high-level rating to many mortgage backed securities, especially Collateralised Debt Obligations (CDOs) including Countrywide's mortgages. The erroneously high ratings resulted in significant investor losses.

Frost (2007) reviews the research evidence on the Enron scandal and the role of the CRAs. Prior to Enron's bankruptcy, executives at Enron and its potential acquirer Dynegy informed Moody's that downgrading Enron would jeopardize Dynegy's acquisition of Enron. It was later revealed that executives at influential financial services firms, who stood to gain from the acquisition, prompted Moody's to maintain an investment grade rating on Enrons bonds, despite signs of high-risk associated with its debt. On the other hand, Egan-Jones downgraded Enron to a *junk rating* 32 days before the bankruptcy, while S&P and Moody's downgraded only five days before bankruptcy. The difference in the time before downgrading could be due to S&P and Moody's having access to intelligence from management and favouring issuers over investors. On the other hand, Egan-Jones were uninfluenced and had little to lose when downgrading Enron.

3.3.2 Revenue Models

There is no consensus as to which revenue model is preferred, as discussed in Section 2.2.6. Researchers have done comparative studies on the two revenue models. We aim to contribute to the field of research by investigating the two revenue models and see if we can see any differences in rating updates.

Jiang et al. (2012) studies ratings from S&P and Moody's in the 1970s and 1980s after they switched models. They show that in the period before S&P switched to an issuer-paid revenue model, Moody's offered more favourable ratings than S&P. However, they find that this effect has disappeared after S&P's switch to the issuer-paid revenue model.

Bonsall (2014) study the impact of Moody's and S&P's adoptions of the issuer-paid model on corporate bond ratings. The results in his study show that after adopting the issuer-paid model, the ratings became more «informative». Bonsall argues that the issuer-paid revenue model allows for economic bonding between the CRAs and issuers through their contractual agreements, which gives the CRAs access to non-public information. Using a difference-in-difference model, the results show that the ratings became relatively more accurate and better to time default after adopting the issuer-paid model in the 1970s.

Kashyap and Kovrijnykh (2016) argues that credit rating errors are more pronounced when assessed by a CRA with an issuer-paid revenue model than CRAs with an investor-paid model. The study also finds that all CRAs produce more significant credit rating errors than socioeconomic optimal. According to their model, the CRAs need to be paid more to have incentives to assess more accurate ratings.

Kronlund (2019) finds evidence of a phenomenon they describe as *Rating Shopping*, which occurs in the issuer-paid revenue model. Rating Shopping is when issuers engage several CRAs, which pressures the CRAs to give the best rating, or they might risk losing the issuer's business. The study finds evidence that issuers frequently wait until after a bond is sold before publishing less favourable ratings.

Bonsall et al. (2022) provide evidence of systematic bias of CRAs with an investor-paid model. They hypothesise that E-J provides higher ratings relative to Moody's when more E-J clients are invested in the corporations close to the investment grade boundary. Their findings cannot prove that investor-paid ratings will resolve problems arising from conflicts of interest, but they could alter the nature of these conflicts in the rating process. They find evidence that E-J ratings are less accurate for firms more held by E-J subscribers.

Xia (2014) looks at S&P's ratings after E-J started coverage of the same issuers. The study finds that S&P ratings were «more responsive to credit risk and its rating changes incorporate higher information content» after E-J started covering the same issuer. Xia uses Moody's KMV model to calculate the Expected Probability of Default and found that E-J tracks this probability better than S&P. The study concludes that CRAs with issuer-paid revenue models have better quality than those with the investor-paid model.

To summarise, there are differences in opinions regarding the best practice of revenue models. CRAs have a motivation to maximise earnings for their shareholders as they are private, profit-driven businesses. The issuers' desire to receive excellent ratings does not coincide with the investors' desire to receive reliable ratings. Since issuers and CRAs have similar interests, the CRAs are coerced into providing issuers with higher ratings to increase their profits. This alignment is at the expense of the investors and well-functioning capital markets.

3.3.3 Market Conditions

In an oligopoly, the market is dominated by a handful of independent firms. As already mentioned (Section 2.2.5), the credit rating market is dominated by the Big Three, making up over 90% of the market in Europe and the US.

Before the CRA market became an oligopoly, it was a duopoly controlled by S&P and Moody's. Becker and Milbourn (2011) studied what happened when Fitch became a prominent agency in 1989, on how it affected S&P and Moody's ratings. The authors describe their findings as «relatively troubling». Fitch's entry increased competition, and the credit ratings' quality fell. In general, the level of the credit ratings increased, and the correlation between the ratings and market-implied yields fell, so the CRAs' ability to predict default declined.

Strong oligopolistic markets often generate antitrust action and regulations by the government (Posner, 1969). Using statistical testing to review the effects of the post-financial crisis regulations of oligopolistic market competition. Malewska (2021) could not conclude that stricter regulations improved the competition.

3.4 Event Study

Hemraj (2015) investigates CRAs in the wake of Enron and Lehman Brothers' default and how CRA agencies compromised their integrity to appease issuers and investors. Hemraj argues that the CRAs should have updated these issuers' investment grade credit ratings earlier, as their reputations were at stake. These scandals illustrate how the CRAs have compromised their integrity to satisfy the issuer. In Sections 3.4.1 and 3.4.2, we will elaborate on two events where the CRAs has be criticised for not providing reliable credit rating updates

3.4.1 The Financial Crisis of 2008

Scalet and Kelly (2012) reviews the role of the CRAs during and the period leading up the Great Financial Crisis of 2008 (hereafter «GFC»). The failures of CRAs partly caused the Bankruptcy of Lehman Brothers on the 15th of September 2008. Lehman Brothers sold high-risk debt products like Collateral Debt Obligations (CDOs) and Mortgage-Backed Securities (MBS), which CRAs had rated as investment grade (Baker Library, 2022).

During the first quarter of 2007, a whopping 53% of Moody's total revenue came from rating complex structured financial products. Mishkin and Eakins (2018) concludes that a conflict of interest arose as CRAs advised clients on how to structure these complex financial instruments at the same time as they were rating them.

In 2013, in the aftermath of GFC, S&P was sued by 19 states and the District of Columbia for issuing inflated ratings that falsely represented the security's real credit risk. The lawsuit included allegations that S&P wrongfully represented its ratings as objective, independent and uninfluenced by its business relations with the investment banks that issued the securities. The lawsuit was settled with a USD 1.375 billion payment and new governance regulations from SEC to increase transparency and reduce conflict of interest (Department of Justice, 2015).

GFC led to the establishment of the *Dodd–Frank Wall Street Reform and Consumer Protection Act of 2010* (hereafter «Dodd-Frank»). The Dodd-Frank Act's primary goal is to discourage issuers from rating shopping and encourage CRAs to assess more accurate ratings. Toscano (2020) looks at the effects of the Dodd-Frank Act on reducing the conflict of interest of CRAs. The study compares rating data on S&P versus E-J and uses these results to conclude whether Dodd-Frank has helped reduce conflict of interest. Toscano finds evidence of S&P ratings being more accurate than E-J ratings in the post-Dodd-Frank period, as seen by the lesser likelihood of significant changes in credit ratings and rating reversals. Additionally, she demonstrates that issuer-paid ratings are more worried than investor-paid ratings are about providing timely ratings in the post-Dodd-Frank era, safeguarding their status as top information providers. Critics of Dodd-Frank, on the other hand, believe the regulation constrains CRAs in the United States compared to non-US CRAs (Committee on the Judiciary, 2012).

Europe's legislative response to the GFC was considered controversial. In the CRA3, one of the goals was to deter investors from relying mechanically and solely on credit ratings. Edwards (2013) opinions that some of the provisions of the CRA3 are noteworthy, but that some are too extreme and fail to meet its objective. Edwards disagreed with two provisions. Notably, (1) the issuer's obligatory rotation of CRAs, and (2) providing issuers and investors with a civil cause of action against CRAs' breaches of regulations. Edwards claims that these provisions may impact the «willingness and ability» of the CRAs to provide a breadth of credit ratings.

To summarise, several local regulatory reforms have been implemented by the governments in response to the wrongdoings of the CRAs post-GFC. However, there is no consensus on international reform for the CRAs. In short, the CRAs disclaim all liability for harm done due to investors or institutions relying solely on their credit ratings.

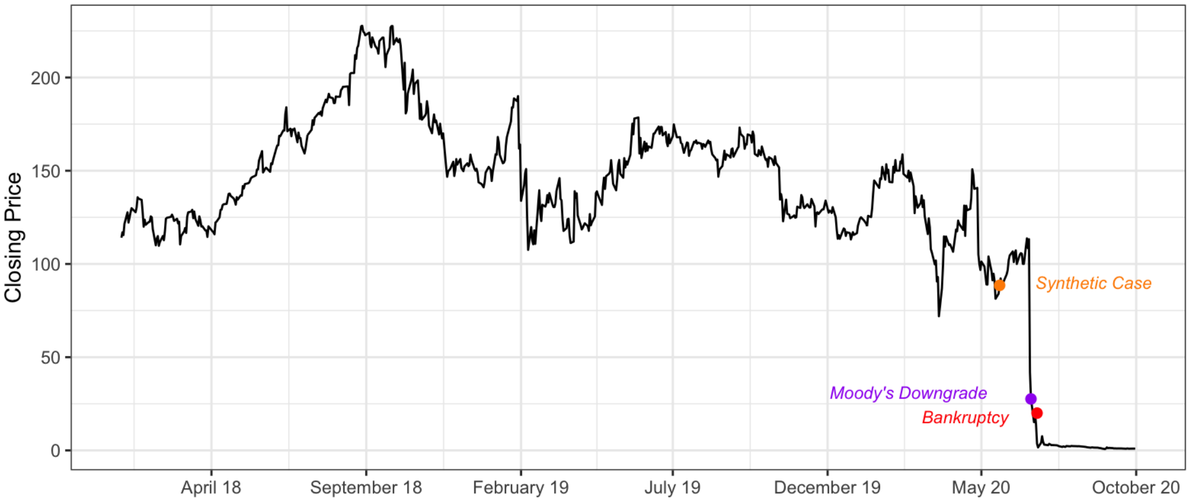
3.4.2 The Wirecard Scandal of 2020

Another event where the CRAs were criticised for being slow and lacking in diligence is the Wirecard Scandal of 2020, which is a case in our data set (See 3.1).

Wirecard was a German financial services company that provided payment processing and banking services. The company was once considered a significant player within the fintech industry. Still, it turned out to be a massive accounting scandal in which it was revealed that it had been reporting fake revenue and profits for several years, with audits signed by auditor Ernst & Young. Wirecard filed for insolvency in 2020, and its former CEO was arrested and charged with accounting fraud. While unrolling these malpractices, the CRAs did not issue a credit rating update.

Financial Times (2022) started an investigative series on Wirecard in 2015 after several allegations of malpractice. However, the allegations were persistently dismissed by Wirecard. In January 2019, the financial regulators of Germany BaFin investigated the Financial Times over allegations of market manipulation. In February 2019, BaFin introduced a two-month short-sell ban on the stock of Wirecard after increased short-selling activity and new allegations.

Figure 3.1: Closing Price of Wirecard Stock



In September 2019, Wirecard issued EUR 500 million worth of bonds, and Moody's rated these bonds as investment grade. In April 2020, KPMG published a report saying they struggled to verify that most of the Wirecard business was real; the same day, the shares fell 26 per cent, see Figure 3.1.

There were many red flags in the weeks leading up to Wirecard's bankruptcy. On June 5th 2020, the police launched a criminal investigation and raided Wirecard's Munich offices. Eleven days later, two banks informed Wirecard's auditor Ernst & Young that the reportedly EUR 1.9 billion they held for Wirecard did not exist. On June 18th, when Wirecard was supposed to publish its 2019 audit, they announced the billions were missing.

Moody's issued a credit rating update where they downgraded Wirecard from investment grade (*Baa3*) to non-investment grade (*B3*) on June 19th and put Wirecard under review on their outlooks list. The CEO of Wirecard resigned that day, and Wirecard filed for insolvency six days later, on June 25th.

It may appear at first glance that no credit rating agencies conducted credit proceedings against Wirecard, considering the long history of short calls and fraud allegations. Comparing Moody's actions to the stock price, shown in Figure 3.1, one can argue that Moody's pulled the trigger late on its credit rating update. In Section 6.4.4, we aim to investigate local SHAP values to see whether our model can detect the decline of Wirecard and predict a *Downgrade* better than Moody's did in June 2020.

Table 3.1: Observation of Wirecard Downgrade by Moody's

Issuer	WIRECARD AG
Country	Germany
Effective Date	2020-06-19
Credit Agency	Moody's
New Rating	B3
Previous Rating	Baa3
Update Type	Rating Downgrade
X1: TotalReturn1Mo	-70.82
X2: TotalReturn3Mo	-71.73
X3: TotalReturn6Mo	-77.01
X4: TotalReturn12Mo	-84.13
X5: EBITMargPct	23.45
X6: TotAssetstoTotEq	3.05
X7: RetainedEarntoTotAssets	0.24
X8: CAPEXtoNetCashFlowOp	0.15
X9: NetDebtToEBITDA	-1.92
X10: LTDebtPctofTotAssets	24.45
X11: QuickRatio	1.75
X12: CurrRatio	1.75
X13: PriceToBookValuePerShr	7.87

3.5 Predicting Credit Ratings using Machine Learning

There seems to be limited research on machine learning on credit rating due to a prevalent issue with obtaining large data sets of credit ratings (Golbayani et al., 2020). There are two reasons why obtaining large data sets on credit ratings is hard. Firstly, a corporation can have multiple bonds with the same financial data. Secondly, bonds are traded more frequently than financial statements are updated, which are usually quarterly or yearly. The two reasons cause some bonds to have the same feature values and some credit rating updates are lost. Moreover, one must be cautious in predicting credit ratings, as machine learning models are trained on data provided by CRAs and can potentially be biased.

However, some studies have been conducted using smaller data sets. See Appendix A2 for an overview of other studies. Studies using primarily using machine learning have shown that machine learning works well when predicting alphanumerical credit ratings using historical financial data.

3.6 Boosting Methods and Asymmetric Information

In this section, which is more technical, we look at previous studies using the methodology we apply in our thesis. The prevalence of using gradient boosting machines for classification has increased in the last couple of years (James et al., 2021). Several studies using machine learning models to explain asymmetric information in finance have shown that gradient boosting methods are suitable for explaining, making it the preferred method in our classification problem using financial data.

Park and Chai (2021) points out that, unlike various academic fields, in finance, machine learning methods are not often used for explaining, and most studies only utilise machine learning for predictions. They suggest using the XGBoost algorithm due to its computational superiority and much higher interpretability than, for instance, Support Vector Machines, which was previously used in similar studies. The main justification for using a gradient boosting method is that one can compare the level of features that «ultimately impact the formation of information asymmetry». We are interested in comparing the level of features that *ultimately impact the formation of credit rating updates*, which makes XGBoost an attractive methodology for our classification problem.

The study of Chang et al. (2018) looks at the construction of machine learning models for financial institutions, and how XGBoost can be implemented in credit risk assessment models for financial institutions. The study indicates that the XGBoost classifier performs better than other tree-based methods in assessing credit risk.

Nevasalmi (2020) utilises several machine learning methods to forecast stock returns in a multinomial classification problem. The study concludes that the gradient boosting machine is the best-performing machine learning method. The gradient boosting model performs best in terms of statistical evaluation, but more importantly, having the best economic predictive performance. The study tests this economic predictive performance by conducting the model's results in a real-life trading simulation. The trading strategy indicated by the model achieves a positive abnormal return in the trading simulation, which is in direct contrast to the efficient market hypothesis.

Gradient Boosting Methods excellent performance in interpreting asymmetric information, credit risk and stock return motivate the selection of the method in our master's thesis.

4 Methodology

This master's thesis aims to contribute to similar research by investigating patterns in credit rating updates and differences in revenue models. We use ML models due to their prediction abilities and accuracy. Moreover, ML has become increasingly prevalent in finance, especially in assessing financial risk, as classic statistical methods describe the relationship between the variables insufficiently (Mashrur et al., 2020). Furthermore, as reviewed in Sections 3.5 and 3.6, ML models can predict credit ratings accurately and create the foundation for the methodology in this thesis.

4.1 Terminology

In ML and statistics, many different terms are used to explain the same phenomena. Breiman (2001) reviews the terminology and differentiates between terms used in ML and statistics. The terminology used in this thesis is chosen as it is most frequently used in the field of ML.

Features are the *explanatory variables*, *predictors*, or *covariates*. **Target** is the *dependent* or *response variable*. **Cases** are *observations*; in this context, one case is a single credit rating update. Model **training** is when a model is *fitted* on cases in the training data.

4.2 Machine Learning Algorithm

Machine learning is the use and development of computer processes that can adapt and learn without specific instructions. These processes are done by analysing data patterns and drawing inferences from these patterns using algorithms and statistical models.

One often distinguishes between two types of ML, supervised and unsupervised. Supervised ML analyses identified and labelled input and output data, while unsupervised learning processes unidentified or raw data. Supervised learning can be further categorised into regression and classification (James et al., 2021).

4.2.1 Interpretable and Explainable Machine Learning

Previous studies using ML in credit ratings have predicted the alphanumeric letter grade and evaluated the model's prediction accuracy, recall Section 3.5. We take the ML techniques further and use them to interpret and explain credit ratings.

ML are often solely evaluated on their predictive performance, such as using a metric to assess the accuracy of the predictions. Nonetheless, not only should ML models be accurate, they should also be interpretable and explainable. The focus on interpretable ML has therefore recently increased (Molnar, 2022).

The terms *interpretability* and *explainability* are frequently interchangeable, and there are several definitions of these terms. Molnar (2022) defines interpretability as «the degree to which a human can understand the cause of a decision and consistently predict the model's result», and further does not define the term explainability. Other researchers call for a distinguishment between interpretability and explainability. Arrieta et al. (2020) define *interpretability* as a passive characteristic of a model that makes sense for humans and *explainability* as a process applied to a model after its predictions. In other words, an active characteristic refers to any action or process carried out by a model to elaborate or clarify its internal processes. The interpretability is a priori (from the earlier), while the explainability is a posterior (from the later).

Doshi-Velez and Kim (2017) state that there is no consensus on interpretability and explainability, and there are too many definitions. The authors argue that most demand the concept of interpretability as the problem is defined incompletely, and they define this incompleteness as a gap between the actual problem and the model formulation.

Complex ML models with low interpretability and explainability are described as *Black Box-models*, and easily interpretable models as *Glass Box-models*. Fortunately, there have been developed methods for explainability, such as SHAP values.

There is a fundamental trade-off between interpretability and model accuracy in ML (James et al., 2021). ML methods are less interpretable than other statistical methods, a linear model could have been utilised to in this thesis. Regression would most likely have multicollinearity due to the usage of financial data. We prefer CART to logistic and linear regression because it does not make simplifying results to explain reality.

4.3 Classification and Regression Trees (CART)

Classification and Regression Trees («CART») are *decision tree algorithms* that can be used in classification or regression predictive modelling problems and were introduced by Breiman et al. (1984). Decision Trees are easy to interpret, but many features and cases cause the algorithm to perform poorly. Decision Trees are susceptible to noise, causing varying results if changes to small training data are applied (James et al., 2021, p. 340).

4.3.1 Classification Trees

Because our target variable is non-numerical (Downgrade, Affirmation, Upgrade), we must use classification trees instead of regression trees with quantitative target variables. In classification, each case is predicted to belong to the most commonly occurring class of training cases in the region to which it belongs (James et al., 2021, p. 335).

Classification Trees

Recursive binary splitting is used to grow a classification tree. Classification error rate is used as a criterion for doing the binary splits, see Formula 4.1. The classification error rate counts as the fraction of the training case for a given region that does not belong to the typical class.

$$E = 1 - \max_k (\hat{p}_{mk}) \tag{4.1}$$

where $\hat{p}_{mk}$ represents the proportion of training cases in the m-th region from the k-th class. In addition, a *Gini index* or *Entropy* needs to be applied supplementary to grow a classification tree. The Gini Index, see Formula 4.2, measures the node purity by measuring the total variance across the K classes. If $\hat{p}_{mk}$ is close to 1 or 0, the index takes on a small value and indicates that a node mainly consists of cases from one class. Entropy, on the other hand, is given by the function in Formula 4.3.

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}) \tag{4.2}$$

$$D = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk} \tag{4.3}$$

Recall, $\hat{p}_{mk}$ represents the proportion of training cases in the m-th region that are from the k-th class, meaning $0 \leq \hat{p}_{mk} \leq 1$. Subsequently $-\hat{p}_{mk} \log \hat{p}_{mk}$ will always be more significant than zero. The Entropy will take on a value near zero if the $\hat{p}_{mk}$ are all near zero or one, and the m-th node is pure.

4.3.2 Ensemble Methods

Ensemble methods are preferred because we have an extensive data set of $37,863$ cases. Ensemble methods combine many simple models to obtain a more robust model (James et al., 2021, p. 340). The simpler models are often known as *weak learners*, as these alone may predict mediocrely. A single decision tree could be the weak learner in CART algorithms, and combining many decision trees is an ensemble method. Examples of ensemble methods include *Bagging*, *Boosting*, *Random Forests* and *XGBoost*.

The advantage of using *Ensembled CART Algorithms* is that they are non-parametric, can handle different data types, are categorical and numerical and are robust against overfitting, outliers, and noise (Nikulski, 2020). Moreover, the multicollinearity of features does not influence the accuracy and predictive performance of the model, and features do not need to be removed or altered to decrease the correlation between them.

However, in terms of performance, having too many unnecessary features adds complexity and should be avoided. The final number of features and how these have been chosen is further described in Section 5.2.2.

4.3.3 Boosting

Boosting is an approach to improve a decision tree's predictions and can be applied to multiple statistical learning methods for regression or classification. Boosting makes predictions for T rounds using a sequential algorithm on the training sample and improves the performance of the boosting algorithm iteratively. The improvement derives from the information on the prior round's prediction accuracy (James et al., 2021; Freund and Schapire, 1997).

Boosting Algorithm

```
1. Set $\hat{f}(x) = 0$ and $r_i = y_i$ for all $i$ in the training set.
2. For $b = 1, 2, \ldots, B$ repeat:
    (a) Fit a tree $\hat{f}^b$ with $d$ splits ($d + 1$ terminal nodes) to the training data $(X, r)$
    (b) Update $\hat{f}$ by adding in a shrunken version of the new tree: $\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}^b$
    (c) Update the residuals: $r_i \leftarrow r_i + \lambda \hat{f}^b$
3. Output the boosted model: $\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^b(x)$
```

Where:

B: *Number of trees*

λ: *Shrinkage parameter, which controls the rate at which the boosting learns*

d: *Number of splits in each tree, which controls the complexity of the boosted ensemble*

4.3.4 Gradient Boosting Methods

Gradient Boosting is a boosting method that is versatile and can be used in both regression and classification problems. This method generates an ensemble of weak learners, typically decision trees, and utilises this ensemble as a prediction model. The model is built in several stages, like other boosting methods, while it generalises the models by using a loss function as an optimiser. The gradient boosting method consists of three components: (1) a loss function that penalises error in the prediction, (2) a weak learner and (3) an additive model (Hastie et al., 2009).

Gradient Boosting Algorithm

```
1. Initialize $f_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma)$
2. For $m = 1, 2, \ldots, M$:
    (a) For $i = 1, 2, \ldots, N$ compute generalised residuals:
        $r_{i,m} = -[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}]_{f=f_{m-1}}$
    (b) Fitting a regression tree to the residual targets $r_{im}$ giving terminal regions:
        $R_{jm}$, $j = 1, 2, \ldots, J_m$.
    (c) For $j = 1, 2, \ldots, J_m$ compute:
        $\gamma_{jm} = \arg\min_{\gamma} \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma)$
    (d) Update $f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$
3. Output model $\hat{f}(x) = f_M(x)$
```

Where:

L: *Different loss criteria*

r: *Generalised (pseudo) residuals*

M: *Number of iterations*

J_m: *Dimensions of each of the individual trees*

For classification, Lines 2 (a) to (d) are repeated K times for each iteration m, using a gradient vector g_{km}. The result in Line 3 is K different tree expansions $f_{kM}(x)$ where $k = 1, 2, \ldots, K$.

4.3.5 eXtreme Gradient Boosting Model (XGBoost)

The eXtreme Gradient Boosting model was created to be used on large, complicated data sets and was first introduced by Chen and Guestrin (2016). It was built on previous CART models, specifically the Gradient Boosting Method, to enhance speed and performance while introducing regularisation parameters to reduce overfitting. We chose the XGBoost algorithm because of its speed and performance, and as Chang et al. (2018) found in their study, it is superior to other algorithms when assessing risk in finance.

The XGBoost method applies the process of various CART as weak learners and bundles them while increasing the performance of each tree formed by minimising its objective regularisation function. The XGBoost method has an excellent ability to predict, especially in the case of classification, because this method has concepts such as sparsity awareness algorithms for each tree formed. The XGBoost algorithm is known for training an optimal model with low tendencies to overfit by regulating the complexity of the tree.

XGBoost has primarily been used for predictions, as it is a black-box model. However, the usage of XGBoost for analytical and explanatory models is increasing (Chang et al., 2018). The algorithm's prevalence increases as it outputs each feature's accuracy contribution scores (gains) and the frequency through the entire tree growth process (feature importance), which are later used to explain the models.

4.3.6 Multiclass Classification Problem

When there are more than two possible outcomes in a classification problem, it becomes a multiclass classification. When we predict *Downgrade*, *Affirmation* or *Upgrade*, we have a multiclass classification problem. There are two approaches for handling a multiclass classification problem: One-versus-All and One-versus-One.

In the *One-versus-All approach*, a classifier for each potential class value is created, with a positive outcome when the prediction belongs to this class and a negative for any other class. Meaning for N class instances, there are N binary classifier models. In this case, a credit rating update rating can be either a Downgrade, Affirmation or Upgrade, meaning a total of three binary classifier models.

On the other hand, in the *One-versus-One approach*, a classifier on every potential pair of classes is created. For example, whether a credit rating update belongs to Downgrade or Upgrade, Downgrade or Affirmation, and Upgrade or Affirmation. With N class instances, there are $N \times \frac{N-1}{2}$ binary classifier models, meaning, in this case, there are three binary classifier models.

4.4 Model Tuning

The parameters of a model must be estimated before it can be used for prediction. Some parameters can be determined directly from the training data the model is fitted on, but other parameters must be set before the model is fitted, as they cannot be approximated from the training data. These parameters are *tuning parameters* or *hyperparameters* and can considerably impact the model. Selecting these hyperparameters must be done cautiously. Many of these parameters regulate the complexity of the model, and not correctly tuning these may lead to overfitting. However, performing this tuning is computationally costly, especially in our case with over $30,000$ observations, therefore, we utilise a new method, ANOVA racing, to save computational time.

4.4.1 Grid Search

Complex machine learning models, such as XGBoost, have several hyperparameters, up to 35 different parameters. Therefore, trying several different values, referred to as candidates, is time- and computationally consuming. There are several approaches to minimize testing all candidates. Resampling techniques like cross-validation or bootstrap are frequently used to evaluate a candidate set of values and select the best one based on a pre-defined criterion. However, the model tuning procedure can be sped up by adaptively choosing candidate values so those sub-optimal settings can be eliminated.

A common approach to quickly find candidates is to perform a grid search. A set of values for each parameter is defined in a regular grid search before models are fitted on different combinations. There are several methods of building various types of regular grids. A benefit of using a regular grid is that the linkages and patterns between the tuning parameters and the model metrics are simple to comprehend. However, it requires many models to be fitted and evaluated, which, as already mentioned require tremendous computational time.

Other irregular grids and iterative methods exist that utilise the candidates' results to decide which parameter value should be tested next. However, it requires using other models or processes, such as Bayesian optimisation or a Gaussian process model, that again introduces more assumptions in the modelling work.

4.4.2 ANOVA Race

Tuning hyperparameters in a grid search is computationally complex. An alternative to the grid search is a set of techniques referred to as *racing methods*. The tuning search evaluates every model configuration on a subset of the resamples. Racing methods perform a limited set of resamples for each grid candidate before conducting a statistical test to determine which ones should be eliminated or given greater attention.

One of these tuning methods is the ANOVA racing method. The ANOVA racing method performs an Analysis of Variance test on the candidates to investigate the statistical significance between the differently-tuned hyperparameters before deciding which to continue testing and which to drop. These racing methods are based on the works of Kuhn (2014), allowing the usage of parallel processing and quickly discarding candidates that are not viable. Performing a grid search saves considerable computational complexity.

4.5 Evaluating Features

Features are the explanatory variables of a ML model. Discussing and evaluating features is part of the explainability of ML and is a central part of our analysis. *Global explanations*, also known as data set-level explanations, describe a model's behaviour over all cases summarised. Global explanations allow for investigation of the model's behaviour on average and in general. A *local explanation*, also known as an instance-level explanation, explains how a model predicts a single case, specifically, how the features contribute to the model's prediction (Biecek and Burzykowski, 2021).

This thesis uses two methods for evaluating the features, SHAP values and Feature Importance, both model-agnostic evaluation methods. SHAP Values are essential to our analysis. We will use global SHAP values to identify patterns in the data and the overall behaviour of the model and local SHAP values to understand why a particular prediction was made, like in the Wirecard case. The predictions of SHAP values lay the foundation for our conclusions in Section 6.

4.5.1 Global Explanation: Feature Importance

Feature Importance opens up for ranking the features by their contribution to the ML model. The plots are easy to interpret and give insights into models with low explainability.

There are various methods to measure feature importance, however, we are only looking at the methods embedded in the XGBoost algorithm. There are three options for measuring model-specific feature importance directly from the XGBoost algorithm: (1) *Weight* is determined by how frequently it is utilised to make a split in the model. It is a measure of how often the feature is used to determine the structure of the model. (2) *Cover* measures the same as Weight but is weighted by the number of training data cases used in the splits. (3) *Gain* measures the increase in accuracy that a feature contributes to the model by being included, versus not being included.

A weakness is that the ordering of the feature importance is often different for all three options to calculate feature importance (Lundberg et al., 2018). Another weakness of feature importance is that the importance is undervalued when two or more features are highly correlated.

Feature Importance

```
1. Initialise $f_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma)$
2. For $m = 1, 2, \ldots, M$:
    (a) For $i = 1, 2, \ldots, N$ compute generalised residuals:
        $r_{i,m} = -[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}]_{f=f_{m-1}}$
    (b) Fitting a regression tree to the residual targets $r_{im}$ giving terminal regions:
        $R_{jm}$, $j = 1, 2, \ldots, J_m$.
    (c) For $j = 1, 2, \ldots, J_m$ compute:
        $\gamma_{jm} = \arg\min_{\gamma} \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma)$
    (d) Update $f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$
3. Output model $\hat{f}(x) = f_M(x)$
```

There are various ways to calculate and approximate feature importance. Instead of trying out the various model-agnostic methods of feature importance, we have decided to focus solely on SHAP values as a model-agnostic method, as it can be used for explanations both locally and globally.

4.5.2 Local and Global Explanation: SHAP Values

Shapley Values is a solution concept in cooperative game theory, introduced by Lloyd S. Shapley (1952). In a cooperative game, meaning competition between groups of players, the Shapley Values quantify each player's contribution. The idea is to fairly allocate the payoff amongst the players in the game when assuming collaboration between the players.

SHapley Additive exPlanations (hereafter «SHAP») is the concept of using Shapley values to assess the features of a machine learning model. Lundberg and Lee (2017) contextualized the concept of Shapley values from economics as a unified approach to machine learning. When there is a need to understand the model's choices in its predictions, SHAP values are essentially utilised for sophisticated models, such as gradient boosting models. The use of SHAP values to demystify black-box models has increased in recent years. The SHAP values can be used for local interpretation, but by aggregating the values, SHAP can also be used for global interpretation.

When explaining an ML model, SHAP values are used to assess a feature's significance in relation to other features. The SHAP values consider the impact of including versus not including that predictive feature on the loss function. Additionally, SHAP values explain if the relationship between the feature and the target is positive or negative.

Transferring the concept of Shapley values to a machine learning model means the game is the prediction outcome of the model, and the players are the model's features. The Shapley values quantify each player's contribution to the game, while in a machine learning context, SHAP quantifies each feature's contribution to the model's predictions.

4.5.2.1 Properties of Shapley Values Lloyd S. Shapley (1952) pointed out that Shapley values have four desirable properties: (1) *Efficiency*, the sum of the values of all players equals the value of the total game. (2) *Symmetry*, all players have an equal chance to be part of the game. (3) *Dummy player*, a player's contribution is zero if the player does not contribute to the game, and (4) *Additivity*, arithmetic summation of pair of games is possible.

Shapley Values

Shapley values are founded on the idea that when determining the significance of a single player, the results of every possible coalition of the players must be considered. Aas et al. (2021) formulates the concept of Shapley values as follows; considering a cooperative game with M players, where all players' objective is to maximise payoff. Let subset $S \subseteq \mathcal{M} = 1, \ldots, M$, where $|S|$ is the number of players — assuming a contribution function every for player $v(S)$ that represents the quantified subsets of the players, also known as the contribution of coalition S. The function $v(S)$ describes the total expected sum of payoff the members of the coalition S can achieve by cooperation.

Assuming all the players collaborate, the Shapley value fairly allocates the total gains to the players. The amount that player j receive is given in Formula 4.4, which is the weighted mean over contribution function differences for all subsets S of players not containing player j.

$$\phi_{j(v)} = \phi_j = \sum_{S \subseteq \mathcal{M}\{j\}} \frac{|S|!(\mathcal{M} - |S| - 1)!}{\mathcal{M}!}(v(S \cup \{j\}) - v(S)), \quad j = 1, \ldots, \mathcal{M} \tag{4.4}$$

SHAP uses the idea of Shapley values by taking the marginal contribution of a feature for a given model. Introducing a new feature m adds computational cost, which is a disadvantage in using the method for large prediction models as the SHAP formula requires training 2^F models.

Another downside with Shapley values is that they implicitly assume that features are independent. However, there are methods for approximating SHAP values that handle high computational costs and adapt extensions for handling dependent features. Two popular approximation methods are the *Tree-Based SHAP* and the *Kernel SHAP*.

The Kernel SHAP approximation of Aas et al. (2021) uses a kernel estimator to reduce computational time and has an extension to handle dependent features. However, this approximation method «suffers greatly from the curse of dimensionality, which inhibits its use in multivariate problems» (Aas et al., 2021).

4.5.2.2 Tree-Based SHAP Tree-Based SHAP has the advantage of being the currently fastest computational way to approximate SHAP values. Lundberg et al.

(2018) developed this approximation, which is today the most widely used approximation of SHAP values due to it being bundled in popular algorithms, such as XGBoost.

The Tree-Based SHAP algorithm reduces the complexity from $O(TL2^F)$ to $O(TLD^2)$, where T is the number of trees in the model, L is the maximum number of leaves in the tree, D is the maximum depth of a tree and F is the number of explained features.

This algorithm uses the structure of tree-based models and the *Additivity property* of the Shapely values to calculate an approximate SHAP value quickly. Three-Based SHAP has two approximation methods for feature perturbation: (1) *Tree-dependence* and (2) *Interventional*.

Tree-Dependence Feature Perturbation using Conditional Expectation The tree-dependence method for feature perturbation takes advantage of the conditional distribution from the tree structure of the underlying CART model. This is the standard and most widely used method, as it is superior in computational time. However, the usage of conditional distribution can introduce the causality problem. Tree-dependence approximation uses the following perturbation for each feature x:

$$v(S) = E[v(x)|x_S = x_S^*] \tag{4.5}$$

Interventional Feature Perturbation using Marginal Expectation The interventional method for feature perturbation breaks the dependencies between the features according to the rules of causal inference dictated by Janzing et al. (2019). This method handles the problem of causality. However, using the marginal distribution of a background data set can provide unlikely data points to the model. Interventional approximation uses the following perturbation for each feature x:

$$v(S) = E[v(x_S, x_S^*)] \tag{4.6}$$

There is currently no consensus on which method is preferable. It discusses whether the attribution method should be «true to model» with tree-dependence or «true to data» with interventional feature perturbation. H. Chen et al. (2020) argue that the preferable method is application dependent.

4.6 Data Splitting

4.6.1 Test and Train Data

The principle of data splitting is to subset the original data, and best practices in supervised machine learning call for partitioning the data into independent sets. The training set is a subset used to train a model to learn possible underlying relationships and patterns. The test set is a subset used to test the trained model by estimating the unbiased accuracy of the model. A common practice is having a validation set for assessing a model's performance across different algorithms and hyperparameters. Best practices call for the test set not to be utilised in the selection, as using the test set ahead of time will lead to overfitting and bias (James et al., 2021).

4.6.2 Stratification

Machine learning models perform better with balanced data sets in a classification problem (Kuhn and Silge, 2022). Stratification ensures that the distribution of one or more specific features is approximately equal in the different subsets. By using stratification, each group within a population receives a proper representation, resulting in a more accurate estimation of a model's parameters. However, this presupposes that the population can be divided into homogeneous subgroups.

4.6.3 Resampling using Bootstrap

Bootstrap is the technique of sampling with replacement, leading to subsets of the same size as the novel data set (James et al., 2021, p. 209). The cases excluded from the bootstrap subset are called Out-of-Bag cases. An alternative to bootstrap is cross-validation, which draws cases without replacement, leading to smaller data subsets than the original data set.

4.7 Model Performance Metrics

There are several ways to assess model performance; the best performance measure depends on the data structure. As we have a classification problem, the performance measures of our model are reviewed in proportions of correctly predicted cases.

4.7.1 Confusion Matrix

The confusion matrix is an accessible and valuable performance measure for ML classification and works well with multiclass classification. The confusion matrices show the predicted values against the actual values, and the left-to-right diagonal values show the correctly classified cases.

The matrix rows represent the predicted classes, while the columns represent the actual classes. The matrix cells contain the number of cases that were predicted to belong to a given predicted class and belong to a given actual class.

Figure 4.1: Confusion Matrices

		Actual	
		P	N
Pred.	P	TP	FP
	N	FN	TN

		Actual		
		A	B	C
Pred.	A	▨		
	B		▨	
	C			▨

Correctly classified cases are True Positives (TP) and True Negatives (TN). Falsely classified cases are False Positives (FP) and False Negatives (FN). From the Confusion Matrix in Figure 4.1, it is possible to derive multiple metrics, given in Table 4.1.

In a multiclass classification problem, instead of operating with positives and negatives, each class's individual TN, TP, FP, and FN are calculated and compile different metrics like *Precision*, *Recall* and the *F1 Score*. The confusion matrices for multiclass problems, shown to the right in Figure 4.1, have the predicted values for each class vertically and the actual value for each class horizontally, and the correctly classified values on the diagonal.

4.7.2 Measures for Accuracy

Table 4.1: Accuracy Measures

Measure	**Formula**
Sensitivity (*True Positive Rate*) measures the rate of True Positives to the total of actual positives and is important when identifying the positives, and preventing false negatives is crucial.	$Sens = \frac{TP}{TP+FN}$
Precision measures the true positives to the total predicted positives and is valuable when the occurrence of false positives is unacceptable.	$Pr = \frac{TP}{TP+FP}$
Specificity (*True Negative Rate*) is the rate of true negatives to the total of actual negatives and is used when you want to cover all true negatives and prevent false alarms.	$Spec = \frac{TN}{TN+FP}$
F1 Score considers both precision and recall and is useful when dealing with uneven class distribution. It separates different types of errors (false positives and false negatives).	$F_1 = \frac{2 \times Sens \times Pr}{Sens+Pr}$

4.7.2.1 Accuracy Accuracy is perhaps one of the most used measures to assess the performance of a model. The accuracy measures the total correct predictions over the total prediction and is given in Formula 4.7. The accuracy measures work well in classification, but it does not work optimally when dealing with imbalanced data sets.

$$Accuracy = \frac{TN + TP}{TP + FP + TN + FN} \tag{4.7}$$

4.7.2.2 Cohen's Kappa As we are dealing with a highly imbalanced data set, recall 2.2, the Accuracy measure is not the most fitting. Cohen's Kappa is used when the data set is imbalanced and normalises the accuracy by weighing the *observed accuracy* against the *expected accuracy.* The goal is to remove the possibility of the classifier and a random guess being the same and measure the number of predictions a model makes that cannot be explained by a random guess. The expected accuracy's presence adjusts for the data set's imbalance.

$$K_{Cohen's} = \frac{p_0 - p_e}{1 - p_e} = \frac{\text{Observed Accuracy} - \text{Expected Accuracy}}{1 - \text{Observed Accuracy}} \tag{4.8}$$

4.7.3 Receiver Operating Characteristic (ROC)

The Receiver Operating Characteristic (hereafter «ROC») is a curve visualising the performance of a given classification model at all classification thresholds. The x-axis shows the proportion Specificity (True Negative Rate), and the y-axis shows the Sensitivity (True Positive Rate). The Sensitivity and Specificity pair associated with each point on the ROC curve corresponds to a specific decision threshold.

4.7.4 Area Under the Curve (AOC)

The Area Under the Curve (hereafter «AUC») is a metric used to evaluate a model's performance. As the name implies, the AUC is the area under the ROC curve. The AUC metric evaluates the model's ability to distinguish between the different classes. A perfect classifier would have an AUC of 1.0, while a classifier that makes random predictions would have an AUC of 0.5.

In a multiclass setting, the AUC is averaged over One-versus-All, and this average is not possible to plot. However, the N classes One-Versus-All is possible to plot and is a good illustration of the model's performance for each different class.

4.7.5 Matthews Correlation Coefficient (MCC)

The Matthews Correlation Coefficient (hereafter «MCC») is a measure of the quality of the classification and is particularly relevant in this multiclass classification problem. It considers True Positives, False Positives, True Negatives and False Negatives and produces only high scores if the prediction returns reasonable rates for all four of these measures. MCC can therefore be described as a balanced measure that works well for imbalanced class distributions.

Matthews Correlation Coefficient

In a multi-class classification problem, the MCC has a generalised calculation for all classes:

$$MCC = \frac{\sum_k \sum_l \sum_m C_{kk}C_{lm} - C_{kl}C_{mk}}{\sqrt{\sum_k (\sum_l C_{kl})(\sum_{k'|k' \neq k} \sum_{l'} C_{k'l'})}\sqrt{\sum_k (\sum_l C_{lk})(\sum_{k'|k' \neq k} \sum_{l'} C_{l'k'})}} \tag{4.9}$$

Where:

$\sum_i C_{ik}$: *Number of times class k truly occurred*

$\sum_i C_{ki}$: *Number of times class k was predicted*

$\sum_k C_{kk}$: *Total number of samples correctly predicted*

$\sum_i \sum_j C_{ij}$: *Total number of samples*

4.7.6 Multiclass Averaging

Moreover, in the multiclass classification setting, some evaluation metrics must be individual, and while some of these measures have unique multiclass implementations, most use micro, macro, or macro-weighted averaging.

Macro and Macro-Weighted Averaging Macro averaging uses multiple sets of binary predictions, calculates the metric for each binary case, and averages the results in multiple One-versus-All approaches. Formula 4.10 shows an example of how metric Precision Pr is macro-averaged.

$$Pr_{macro} = \frac{Pr_1 + Pr_2 + \ldots + Pr_k}{k} \tag{4.10}$$

Where Pr_1 is the precision calculated in a One-versus-All approach, which predicts either one as the first class and the two others as zero. Macro averaging give all classes an equal weight, which might not be realistic if there is a class imbalance (Vaughan, 2022). An alternative to handling the class imbalance is to use macro-weighted averaging, where the frequency of that class in the test data set calculates the weights.

Micro Averaging In micro averaging, the averaging is done with the entire data set as an aggregated result, and just the metric is calculated, rather than averaging k metrics.

Each case gets an equal weight in micro averaging, while each class has an equal weight in macro averaging. Formula 4.11 shows an example of how metric Precision Pr is micro-averaged.

$$Pr_{micro} = \frac{TP_1 + TP_2 + \ldots + TP_k}{(TP_1 + TP_2 + \ldots + TP_k) + (FP_1 + FP_2 + \ldots + FP_k)} \tag{4.11}$$

Other Averaging Methods However, some metrics do not require multiclass averaging, such as Accuracy and Kappa and MC. The standard averaging method works poorly for other measures, such as the AUC.

4.7.6.1 Hand & Till AUC Averaging Hand and Till (2001) introduced a new method that preserves insensitivity when there is an imbalanced class distribution to compute a multiclass AUC more accurately. Unfortunately, this approach to ROC and AUC does not make sense in a visual interpretation.

Hand & Till AUC Averaging

The Hand & Till averaging method uses a One-versus-One approach, where N pairs of classifiers (i, j) are considered. The probability that a case belonging to class j has a lower probability for class i than a random case belonging to class i is defined as $\hat{A}(i|j)$. Since $\hat{A}(I|j) = \hat{A}(j|i)$, one can define $\hat{A}(I, j) = \frac{1}{2}(\hat{A}(i|j) + \hat{A}(j|i))$. The overall AUC of a multiclass classifier is the average value $\hat{A}(i, j)$ is given in Formula 4.12.

$$\underset{\text{Hand \& Till}}{\text{AUC}} = \frac{2}{N(N-1)} \sum_{i<j} \hat{A}(I, j) \tag{4.12}$$

5 Data and Modeling

5.1 Data

5.1.1 Credit Rating Updates on Corporate Bonds

The credit rating updates on the corporate bonds are retrieved from Ratings Monitor in Refinitiv Workspace. The data set contains information on European and US bonds from 01.01.2002 to 16.09.2022, with ratings from Moody's, Fitch, and Egan-Jones.

Each case in the data set is an update on a credit rating of a corporate bond. An update can have one of three actions: (1) *Affirmation*, (2) *Downgrade*, or (3) *Upgrade*. The cases contain information on the issuer, date, sector, previous rating, previous rating date, and the bond's duration (short- or long-term) and the issuer's country (See Table X).

As the models require financial data, the data is limited to only include publicly listed issuers. The data set contains bonds issued by listed companies in the following sectors: Manufacturing, Service Company, Electric Power, Telephone, Transportation, Consumer Goods and Gas Distribution (See Table 5.1). Issuers in sectors such as Banking and Financial Services are excluded, as these issuers typically have other implications for their credit ratings.

Table 5.1: Issuer's sector distribution from the Analysed Data

Sector	**Frequency**
Manufacturing	39.79%
Service Company	22.92%
Electric Power	12.55%
Energy Company	8.87%
Consumer Goods	6.92%
Telephone	3.91%
Transportation	3.87%
Gas Distribution	1.16%

5.1.2 Financial Ratios and Returns

The financial data used for the features are retrieved from Refinitiv Workspace. The credit ratings' data set, financial ratios and returns were matched based on the Equity Reuters Instrument Code. The financial ratios are gathered by trailing the previous quarter of the date of each credit rating update. Missing data were imputed with yearly data where quarterly data was unavailable.

5.2 Model Specification

5.2.1 Target Selection

The selected target is a multinomial ordinal variable with three class instances: (1) *Downgrade*, (2) *Affirmation* or (3) *Upgrade*. The target variable is imbalanced, as most credit rating updates are *Affirmations*, as corporate credit ratings seldom change.

Upsampling Having an imbalanced target variable can be problematic. Upsampling with different methods, such as Synthetic Minority Oversampling Technique (SMOTE) and Random Over-Sampling Examples (ROSE), did not pose significant improvements. Other performance measures, such as Cohen's Kappa, are utilised to assess the imbalance.

5.2.2 Feature Selection

Availability and reliability were prioritised in the feature selection process. Occasionally, some financial ratios in Refinitiv were deficient and could not be included. Qualitative and quantitative methods and financial theory were applied in the feature selection process. The features were selected using investigating the CRA rating methodologies and reviewing the feature selection in similar research papers and financial theory. Initially, the data set had 31 features based on the most used financial ratios, and in the final selection, there were 11 complete features.

A thorough analysis of the assessment methods by the CRAs was conducted. CRAs use different ratios and weighting of factors depending on the issuer's sector. The data set in this thesis comprises mainly manufacturing, service, and electric power companies, with 39.79%, 22.92% and 12.55%, respectively (See Table 5.1). The compound of sectors in the data set subsequently shaped the final selection of variables.

The literature review of similar research papers in Section 3 tells of success using historical financial data as features in credit rating prediction. The features most frequently observed in similar research were evaluated in the feature selection.

The most traditional way to assess credit risk is to calculate the solvency and liquidity of a debtor (Berk and DeMarzo, 2019). Liquidity ratios like *Current Ratio*, *Quick Ratio* and *Operating Cash Flow* are included to explain the corporation's ability to pay debt obligations and its safety margin. *Long-Term Debt to Assets* and *Net Debt to EBITDA* are included as variables to explain the corporation's solvency.

Table 5.2: Final Feature Selection

Feature	Description
$X1^{t=[0,1]}$	Total Return[2], Last Month
$X1^{t=[0,3]}$	Total Return, Trailing 3 Months
$X1^{t=[0,6]}$	Total Return, Trailing 6 Months
$X4^{t=[0,12]}$	Total Return, Trailing 12 Months
$X5$	EBIT[3] Margin (%)
$X6$	Total Assets to Total Equity
$X7$	Retained Earnings to Assets
$X8$	CAPEX[4] to Net Cash Flow from Operations
$X9$	Net Debt to EBITDA[5]
$X10$	Long-Term Debt to Total Assets
$X11$	Quick Ratio
$X12$	Current Ratio
$X13$	Price to Book

Return Features Unlike the papers discussed in the literature review in Section 3.5, the model includes the total returns of the issuer's stock as features. There are four different trailing time horizons included as features. The trailing total return $X_{[1,4]}^{t,T}$ is given by the stock price p_t, and any dividends or share repurchases denoted as D, in a trailing time from t to T:

$$X_{[1,4]}^{t,T} = \frac{(p_T - p_t) + D}{p_t} \tag{5.1}$$

[2] Total Return: Trailing Total Return of Issuer's Stock
[3] EBIT: Earnings Before Interest and Taxes
[4] CAPEX: Capital Expenditures
[5] EBITDA: Earnings Before Interest, Taxes, Depreciation, and Amortization

The total return features incorporate price changes and any relevant dividends trailing over a period for the last n months, where the chosen time horizon is $n = 1$, 3, 6 and 12 months to the date of the credit rating update.

5.2.3 XGBoost Algorithm

In the selection of the machine learning algorithm, multiple CART algorithms were tried. The chosen machine learning algorithm is XGBoost, with *Softmax* as the objective function. Other CART algorithms, besides XGBoost, did not stipulate desirable performance measures. XGBoost's interpretability is substantially lower than bagging and decision trees, but its computational time is a lot better, as well as accuracy.

A fundamental assumption for including total return features is that stock returns can, in some way, justify the absence of the qualitative assessment in the model. Recall the Efficient Market Hypothesis, which states that all available information is incorporated in stock returns. One of the main components of the structural models of Merton's and Moody's KMV, introduced in Section 2.2.4.3, is stock return volatility. The stock price picks up any indications of lower future income, like loss in market shares or possible disruptions in the corporations' market segment, which is like what CRAs review in their qualitative assessments. For example, in the Enron and Wirecard scandal, the stock market detected the downfall of these issuers before the CRAs.

Tuning of Hyperparameters The hyperparameters of the XGBoost machine learning model are set before training. The hyperparameters provide a framework for the model to make predictions and determine the overall behaviour of the model. Tuning these hyperparameters is an essential step in the modelling, as it can significantly impact the model's performance.

The goal of the hyperparameter tuning is to find the combination of hyperparameters that results in the best performance for the model. In this case, this measure uses Hand & Till-weighted as an indicator of model performance. The tuning was done using the ANOVA Racing Method to reduce computational processing time.

Table 5.3: Tuning of Hyperparameters

Parameter	**Value**	**Description**
Max depth	*6*	Maximum depth of a tree. A higher value leads to increased model complexity and may lead to overfitting.
Gamma	*0*	Minimum loss reduction is necessary for new leaf split in the tree. Large values lead to more conservative models.
Column subsample	*0.5385*	Subsample ratio of columns for each split. Subsampling occurs once every time a new split is evaluated.
Min. child weight	*7*	The constructing process will stop dividing if the tree partition step yields a leaf node with the sum of instance weight less than the minimum child weight, and a large value leads to more conservative models.
Subsample	*1*	Subsample ratio of the training instances $[0, 1]$. Subsampling occurs once in every boosting iteration.
Estimators	*1 646*	The number of fitted boosted trees.
ETA	*0.0862*	Learning Rate $[0, 1]$. Lower values make boosting process more conservative but more robust against overfitting. Low value utilises speed.
Early stopping	*50*	If the model does not improve for 50 consecutive iterations, it will revert to the iteration with the best AUC.
Features	*7*	The number of features will be randomly sampled at each split in the tree.

5.2.4 Training, Test and Validation Sets

The final data set contains 30,375 complete cases. The data is split into training and test data sets. The training data is further split into resampling folds set using the Bootstrap technique, used as a validation set for tuning the hyperparameters. Separate training and test sets reduce bias and decrease the chance of overfitting.

Stratification The data set has a high degree of *class imbalance*, as most credit rating updates are *Affirmations*. The training and test split was done using stratification to ensure that the test and train data were approximately equal. The stratification is done on the target to ensure balance in the data splitting.

6 Analysis and Results

We investigate credit rating updates using ML techniques while using the literature review and event study to discuss our findings. One hypothesis is that the CRAs are slow in updating a credit rating. Sections 2 and 3 examine other concerns, for example, client pressure. The analysis aims to see if a ML model can identify any of these concerns.

Explainability is particularly relevant in this analysis, as it does not aim to find the ML method with the highest accuracy – instead, see other studies in Appendix A2. Instead, the model in this analysis aims to look at rating updates and use SHAP values for discussion.

The ML model is used as an explanatory model, with changes in the training data used to train the model between the identified effects. When the model is retrained on new data, the test and training data are filtered only to include cases that help explain the identified effects. For example, when investigating the investor-paid revenue model, only cases issued by E-J are included.

Table 6.1: Data Set Used in Analysis, Credit Rating Updates by Agency

Data Set	Moody's	E-J	Fitch
Training Data	12,134	8,158	10,004
Test Data	3,070	2,048	2,459

6.1 The Benchmarking Paradox of Credit Ratings

When evaluating predictive models, the *factual truth* is often established. The problem with using models to predict or explain credit ratings is that the truth is benchmarked on the *truth of the CRAs*, and the factual truth is unknown. We cannot compare the predictive value to an observable and testable truth. With credit ratings, we must trust the CRAs to provide the *truth*.

The hypothesis is that the CRAs' truth occasionally deviates from the factual truth. Established empirically by similar studies, as outlined in Section 3, the truth of CRAs is occasionally biased caused by client pressure and agency problems. Consequently, the models are trained and tested on potentially biased data, and it is impossible to validate them unbiasedly. The estimated accuracy of the model is based on the truths of the CRAs, and the accuracy could be closer or further from the actual unknown truth.

6.2 Model Performance

The predictive ability of the machine learning model is assessed by examining a confusion matrix and the ROC curves. The model is trained on the training data set before it is evaluated on a separate test data set. The test data set is only used for assessing the model's performance.

6.2.1 Confusion Matrix

The Confusion Matrix, in Table 6.2, displays the model's predictions of the test data set. The confusion matrix shows that the model predicts correctly in most cases. The confusion between *Affirmation* and the two other classes, *Upgrade* and *Downgrade*, is expected. While confusion between *Upgrade* and *Downgrade*, and vice versa, occurs in a total of 46 cases, which indicates the model's predictive ability is good.

Table 6.2: Confusion Matrix of Test Data

		Truth		
		Downgrade	Affirmation	Upgrade
Pred.	Downgrade	889	136	22
	Affirmation	438	4,789	436
	Upgrade	25	136	706

From the confusion matrix, several measures are calculated. Most of these measures are calculated in a One-versus-All setting and averaged using Macro-weighted averaging. Macro-weighted is chosen as it accounts for class imbalance by computing the average of the One-versus-All binary metrics, weighted by the number of samples of each Action class. See Table 6.3 for a complete overview of these test metrics.

Table 6.3: Multiclass Measures

Measure	**Weighting**	**Value**
Accuracy		84.25%
Cohen's Kappa		65.66%
Matthews Correlation Coefficient		66.67%
AUC	Hand & Till	89.47%
F1 Score	Macro	83.56%
Precision	Macro	84.14%
Recall	Macro	84.25%
Sensitivity	Macro	84.25%
Specificity	Macro	75.95%
Negative Predicted Value	Macro	88.19%
Positive Predicted Value	Macro	84.15%

The model obtains a test accuracy of 84.25% and a Training Accuracy of 86.08%. The slight difference between the Accuracy of the train and test data set can be interpreted as the model had little signs of overfitting. Overfitting is when the model fits the training data well but cannot generalize and make accurate predictions for data it has not seen.

Alternative measures to the Accuracy are Cohen's Kappa, which has a score of 65.66%, and Matthews Correlation Coefficient has a score of 66.67%. Both measures are adjusted for the imbalanced class distributions. Compared to similar machine learning studies on credit rating prediction, listed in Appendix A2, the model's Accuracy and Cohen's Kappa scores are significantly better.

6.2.1.1 Weighted Averaged Measures Since this is a multiclass classification problem, the rest of the metrics used to assess the model performance are averaged from three One-versus-All classifiers to a single metric using Macro-Weighted averaging. Macro-Weighting is chosen as it is a good measure for an imbalanced class distribution.

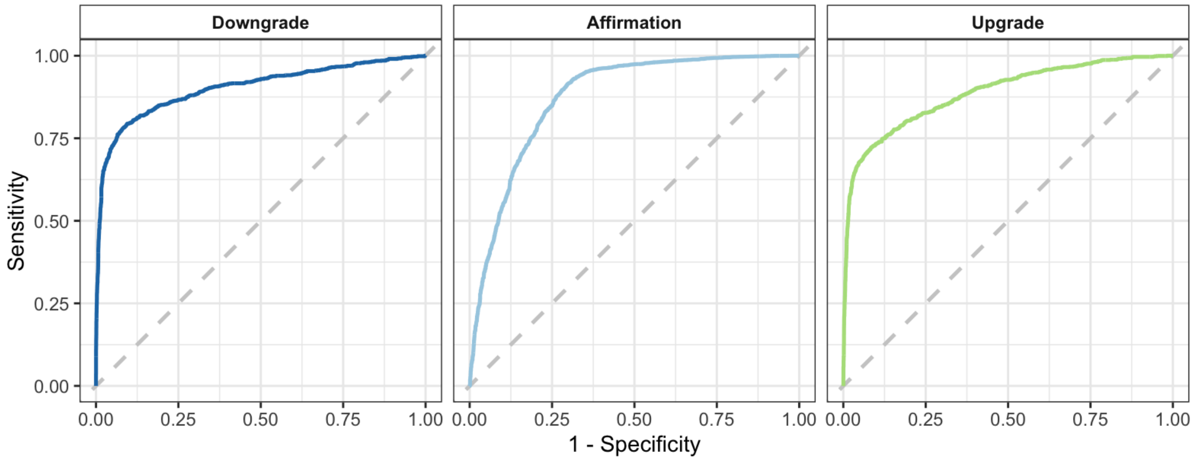
The model's Precision and Recall are 84.14% and 84.25%, respectively. The F1 score, which is a weighted harmonic mean of Precision and Recall, obtains a score of 83.56%. An F1 score of 83.56% can be considered good if it beats the model's baseline score. The baseline score of the model is if the model predicts one outcome for all cases. For example, predicting the *Affirmation* class for all cases in the test set would give a baseline score of 66.78%. Therefore, it can be concluded that the model predicts better than if it just predicted *Affirmation*.

The macro-weighted average Sensitivity (True Positive Rate) and Specificity (True Negative Rate) are 85.24% and 75.95%, respectively. The average Sensitivity is how good the model is at correctly predicting *Updates that belong* to its class, while average Specificity is how good the model is at predicting *Updates that do not* belongs to its class correctly. In other words, the model is better at predicting True Negatives, meaning it is better at predicting when something *should not* be classified as a *Downgrade* than when it *should* be classified as a *Downgrade*.

6.2.1.2 ROC and AUC The ROC curves are plotted in Figure 6.1. Every curve represents each One-versus-All classifier. The curves indicate that the model has the best success in predicting *Downgrades*. Looking at the separate AUC for the ROC curves, *Downgrade* has a value of 90.20%, beating *Affirmation* with 87.30% and *Upgrade* with 88.40%. The model average AUC of all classes is 89.47%, averaged using the Hand & Till method.

The Specificity (True Positive Rate) and the Sensitivity (True Negative Rate) show signs of most cases being *Affirmation*. The trained model predicts *Affirmation* in most cases, see Figure 6.1, and consequently, the Sensitivity is high for *Affirmation*, and the Specificity rate is high for the smaller classes, *Downgrade* and *Upgrade*. Thus, The ROC curve looks different for the three classes.

Figure 6.1: ROC Curves, One-versus-All



The ROC curve plots can be used for diagnosing a model. To interpret a ROC curve, you can look at the AUC supplementary to get an idea of how well the classifier is performing.

It is possible to look at the curve itself to see how the TPR and TFR change as the classification threshold are varied. A steep curve indicates that the classifier is able to distinguish between positive and negative cases very well, while a shallow curve indicates that the classifier is not able to distinguish between the two classes very well.

By looking at the ROC plots in Figure 6.1 the steepest curve is the *Downgrade* and *Upgrade*, which lines up with the AUC scores for each class (Table 6.4). In this multiclass classification, the three class instances are mutually exclusive. Therefore, each class have a multinomial probability distribution, which means that there are $N = 3$ probabilities of the prediction falling into each of the three classes sum up to one. Followingly, there is no threshold for our classification, but the model rather picks the class with the highest likelihood.

Table 6.4: Metrics per Class Instances

	Downgrade	**Affirmation**	**Upgrade**
AUC	90.20%	87.30%	88.40%
Sensitivity	65.75%	94.63%	60.65%
Specificity	97.46%	65.26%	97.49%

In statistical hypotheses testing, a *Type I Error* is a wrongful rejection of an actual true null hypothesis (a false positive). A *Type II Error* is the inability to reject a false null hypothesis (a false negative). Type I Error, in this case, is updating a rating when it should not have been updated, and Type II Error is not updating a rating when it should have been.

Downgrade The AUC for *Downgrade (versus-All)* is 90.20%, which can be interpreted as the model accurately predicts *Downgrades.* Sensitivity in the *Downgrade* is 65.75%, and Specificity is 97.46%, meaning the models predict negatives almost perfectly. A false positive for a *Downgrade* is when a credit rating update is predicted as a *Downgrade* when it should have been an *Upgrade* or *Affirmation.* Receiving a false positive on a *Downgrade* could be devastating for the issuer, especially if it causes the bond to be a fallen angel (falling from investment to non-investment grade). Due to client pressure instigated by the relationship between CRAs and bond issuers, it is crucial that CRAs do not issue a downgrade under uncertainty. On the other hand, a false negative is not downgrading a risky bond that should be downgraded.

Defining the threshold for *Downgrades* is complicated. CRAs face a conflict of interest, as bond issuers are interested in avoiding false positives, and investors and regulators are interested in avoiding false negatives. Having a low threshold for false positives and giving to the pressure of bond issuers, CRAs may lead to good client relationships and future revenue. However, this increases the risk of scandals like the ones during the Financial Crisis in 2008 or the Wirecard Scandal in 2020. The optimal threshold for downgrades depends on the model user. CRAs operating with an investor-paid revenue model likely uses a higher threshold for false negatives as they look out for the investors' best interest. Hypothetically, if regulatory authorities should use this model, the threshold for false negatives will likely be high compared to CRAs.

Affirmation *Affirmation (versus-All)* differs from *Upgrades* and *Downgrades*, likely because it is an extensively larger class, in terms of cases compared to the others. Alternatively, an explanation could be that *Affirmation* lies ordered between *Upgrades* and *Downgrades*. The AUC for *Affirmation* is 87.30% and has a lower degree of Specificity than *Upgrades* and *Downgrades* at 65.26%. The high Sensitivity at 94.63% is caused by the true cases being predominantly *Affirmation*.

Upgrade The AUC for *Upgrade (versus-All)* is slightly lower than for *Downgrades*, at 88.40%. The model does not predict *Upgrades* as well as *Downgrades*. The Sensitivity is 60.65%, and the Specificity is 97.49%, meaning the model is better at predicting true negatives. A false positive in the class *Upgrade* predicts an upgrade when it should not be upgraded, and a false negative does not give an upgrade where it should have been given. Again, regulatory bodies and investors are likely more critical of false positives than issuers.

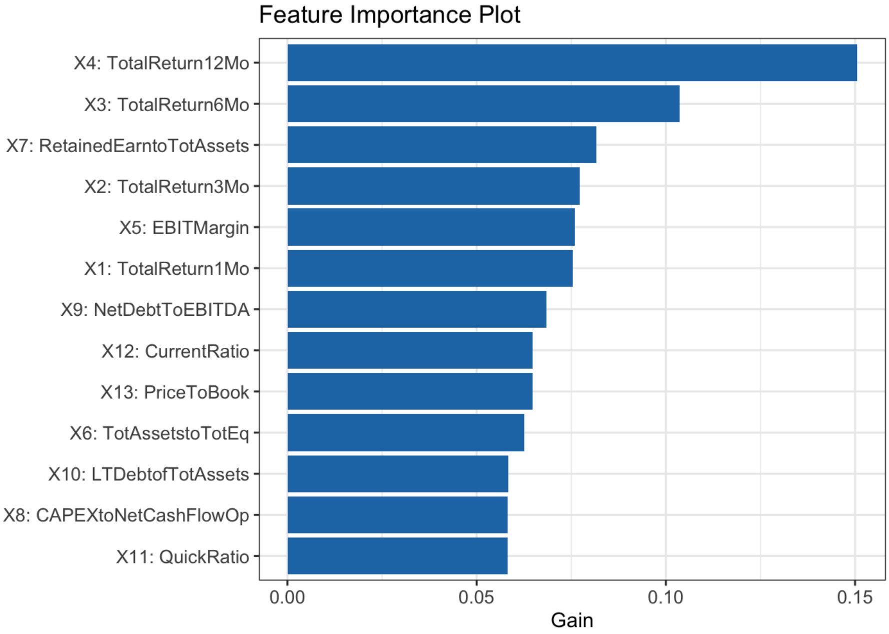
6.3 Explanatory Results

6.3.1 Feature Importance

The feature importance plot is a graphical representation of the importance of each feature in the model. The features are ranked in order from most important to least important. The absolute values of the features in the plot are not meant for interpretation, since the output is normalised, summing the total value to 1.

Based on previous studies and the CRAs' rating methodologies, the initial assumption was that financial ratios are the most important features when predicting credit ratings. Surprisingly, when looking at the feature importance plot by *Gain* in Figure 6.2, the most important feature in the model, when predicting, is $X4: \textit{Total Return } 12 \textit{ months}$, followed by $X3: \textit{Total Return } 6 \textit{ months}$, meaning the return variables have a significant predictive power, which, to our knowledge, has previously not been tested in machine learning. When two features are highly correlated, the importance is undervalued, which might suggest that the importance of the four total return features $X1$, $X2$, $X3$ and $X4$ are underestimated.

Figure 6.2: Feature Importance Plot, Measured using Gain



The hyperparameter tuning affects the features' ranking, and the feature ranking is slightly different when using *Cover*, *Gain* or *Weight* as the ranking measure. SHAP values also provide a ranking of the features. We find it more meaningful to rank the features after their contribution, and the rest of the analysis only uses SHAP values for feature ranking.

6.3.2 SHAP Values

The global SHAP values intuitively explain the models' predictions and quantify the magnitude and direction (positive or negative) of the features' effect on a prediction. Each feature has a SHAP value that measures the feature's contribution to the prediction. Since SHAP values can be used as local and global explanations, giving it a common foundation, the following results will be discussed using SHAP values only.

The SHAP values are calculated using the Tree-Based SHAP algorithm, with Tree-Dependence as feature perturbation. Test using the Interventional feature perturbation to handle the independence assumption did not yield any significant differences in results. It is important to note that the independence assumption is breached, according to Janzing et al. (2019). However, in the context of finance, ensuring independence is difficult. SHAP values are not approximated directly in a multiclass setting; instead, they are implemented in a One-versus-All classifier approach.

The absolute mean of the SHAP values is displayed in Figure 6.3. The x-axis shows the absolute mean of the SHAP value, which is the feature's average impact on the model's output for a given class. The plot shows that for class *Affirmation*, *X4:Total Return 12 months* has a SHAP value of 0.2, while for class *Downgrade* it is 0.5 and 0.45 for *Upgrade*. In all the One-versus-All classifiers, the SHAP values are greatest for *X4: Total Return 12 months* and most important for *Downgrades* and *Upgrades*. For *Affirmation*, features *X3: Total Return 6 months* and *X4: Total Return 12 months* are equally important.

The individual absolute mean SHAP for each class are depicted in Figure 6.4 and emphasizes the importance of the *X4: Total Return 12 months* feature, in all classes. However, for *Downgrades*, the contribution is more substantial than the classes' remaining features.

Figure 6.3: Plot of SHAP Values Summarised

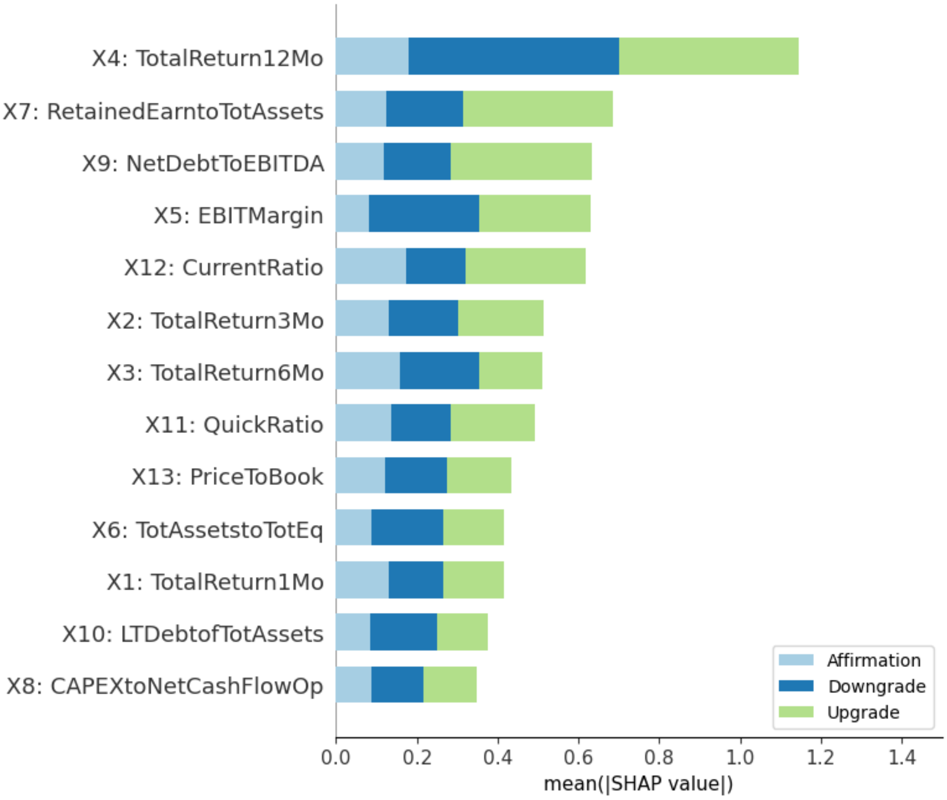
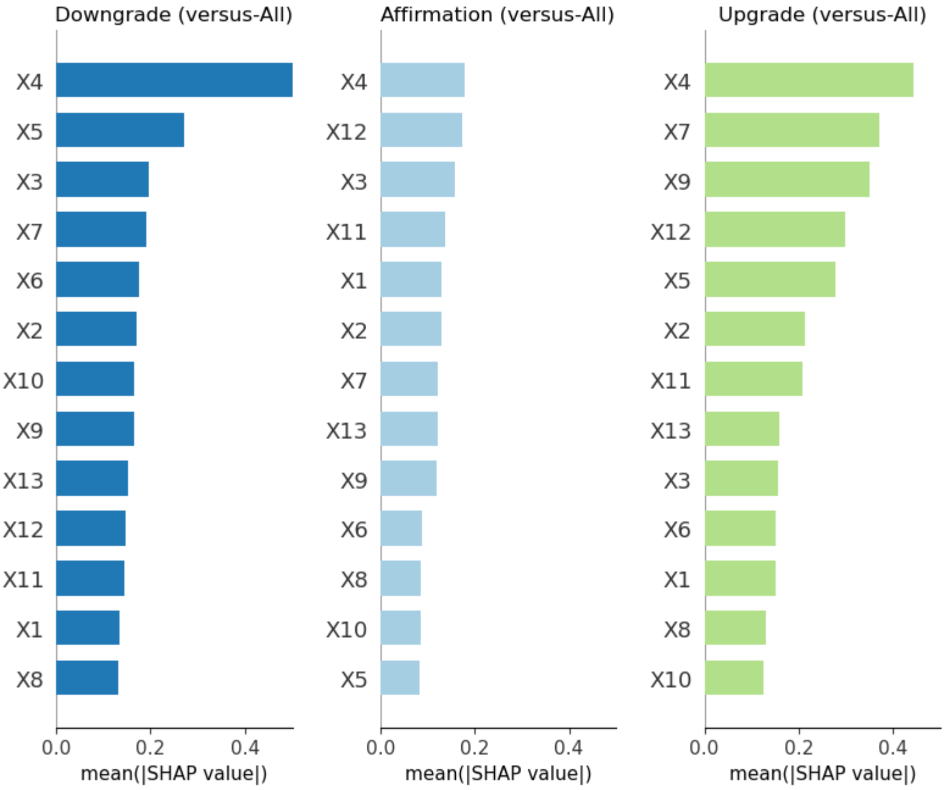


Figure 6.4: Plot of SHAP Values



6.4 Analysis

6.4.1 Investigating the Total Return Features

Recall from Section 5.2.2 that the total stock return is included as a proxy for the qualitative assessment by the CRAs. Moreover, as discussed in Section 3.1, this is justified using the Efficient Market Hypothesis and the study of Easley et al. (2010). The EMH states that a price of an asset contains all available information. However, this assumption requires that public and private information is reflected in the total return of the issuer's stock, which is only valid in the strongest form of the EMH.

The discrepancy between a changing total stock return and an unchanged bond rating could indicate delays caused by hesitation due to, for example, client pressure. However, the delay could alternatively be caused by systematic reasons or an argument favouring the CRAs as they are cautious and thorough in their assessments.

6.4.2 Investigating the Upgrades and Downgrades Classes

Feature X4: Total Return 12 months is the most important feature when predicting all classes. Suppose the delay of up to 12 months is caused by the CRAs' thorough diligence work, which is systematically time-consuming. In that case, the feature should be equally important in predicting *Downgrades* and *Upgrades*. All cases are summarised in three beeswarm plots in Figures 6.5, 6.6 and 6.7 to investigate a possible difference between classes.

The SHAP beeswarm summary plots combine the feature importance with their contribution. Each point on the plot represents a specific case. Overlapping points are jittered through the y-axis, allowing a better illustration of the SHAP values per feature. The features are ordered according to importance on the y-axis, while the x-axis represents the SHAP value. The colour of the point is the value of the feature, going from low values in blue and high values in red. The cases should be roughly centred in the plot since it shows the impact of including the feature versus not including the feature in the model. The plot illustrates the relation between the SHAP values and the corresponding feature value for all the cases labelled to a particular class.

A positive SHAP value pushes the model above the expected output of the average model, while a negative SHAP value pushes the model below the expected average model output. The SHAP value is the average marginal contribution of a feature value across all possible sets of features.

Figure 6.5: Plot of SHAP Beeswarm: Downgrade (versus-all)

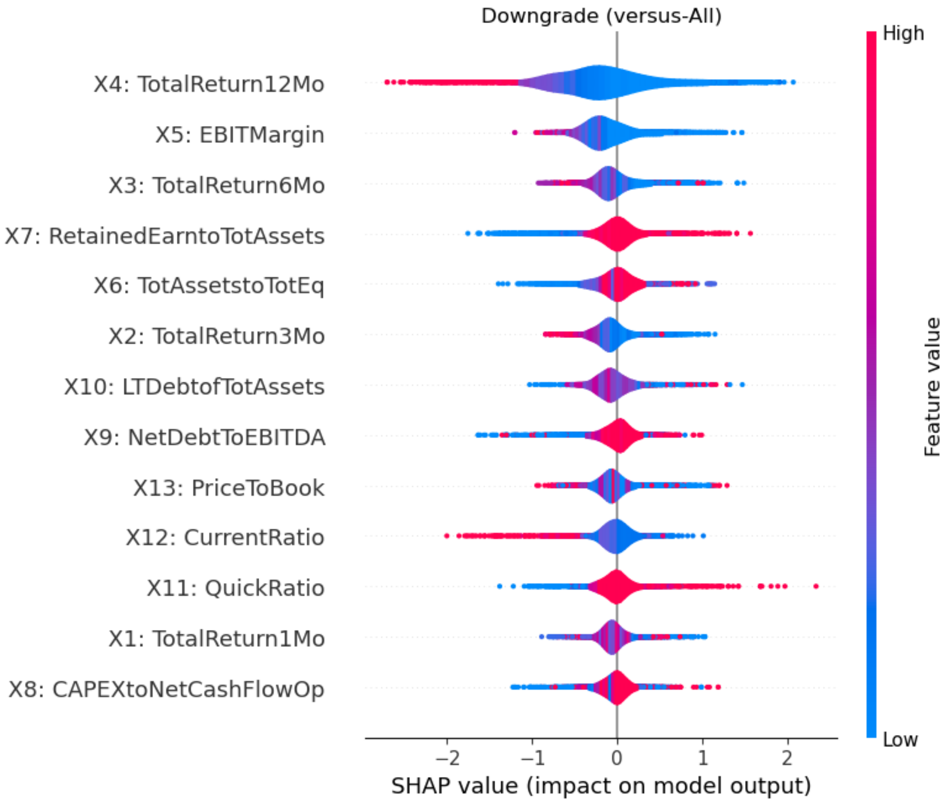


Figure 6.6: Plot of SHAP Beeswarm: Affirmation (versus-all)

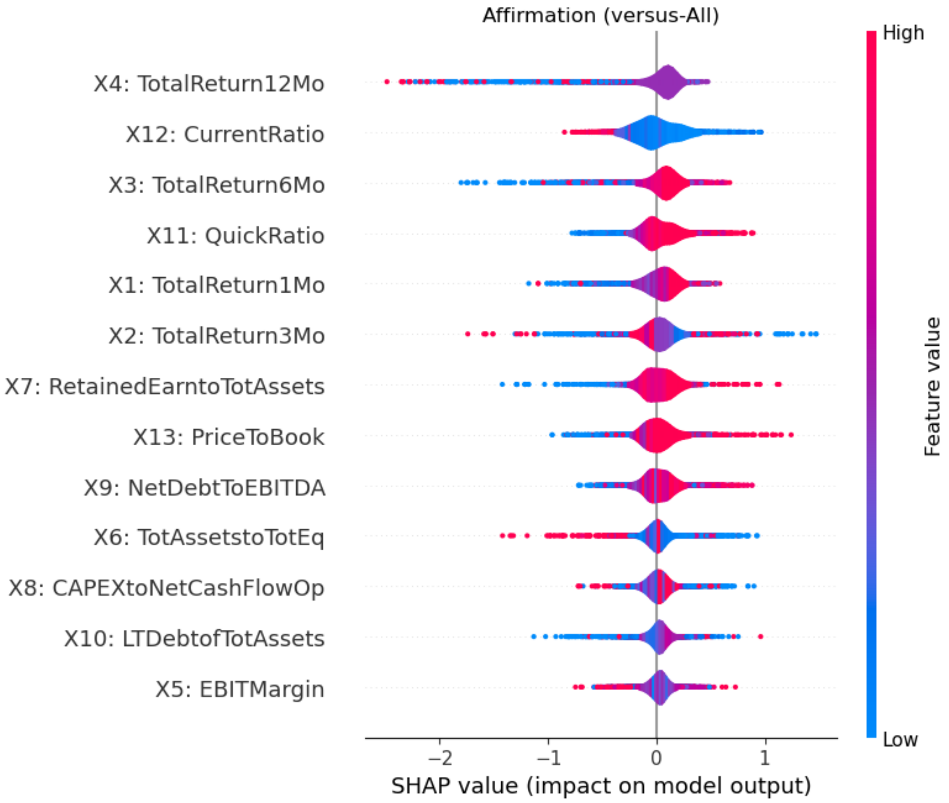
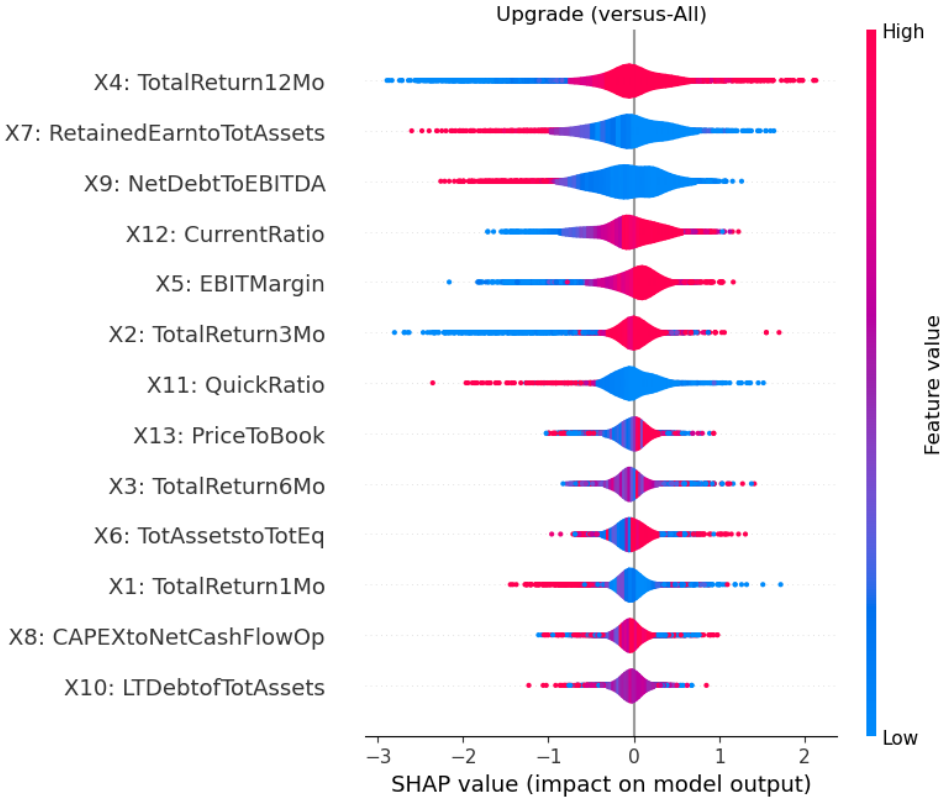


Figure 6.7: Plot of SHAP Beeswarm: Upgrade (versus-all)



The SHAP Summary Plot for *Downgrade (versus Affirmation and Upgrade)* in Figure 6.5 shows that low feature values for Total Returns ($X1$, $X2$, $X3$ and $X4$), *X5: EBIT Margin*, *X12: Current Ratio*, and high feature values of *X7: Retained Earnings to Assets*, *X9: Net debt to EBITDA* and *X11: Quick Ratio* results in high SHAP values and means it contributes in predicting a *Downgrade*.

The SHAP summary plot for *Upgrade (versus Downgrade and Affirmation)* in Figure 6.7 shows that when predicting an *Upgrade*, high feature values of Total Returns ($X1$, $X2$, $X3$ and $X4$), *X12: Current Ratio* and *X5: EBIT Margin* yields negative SHAP values. Low feature values of *X7: Retained Earnings to Total Assets*, *X9: Net Debt to EBITDA* and *X11: Quick Ratio* yield positive SHAP values.

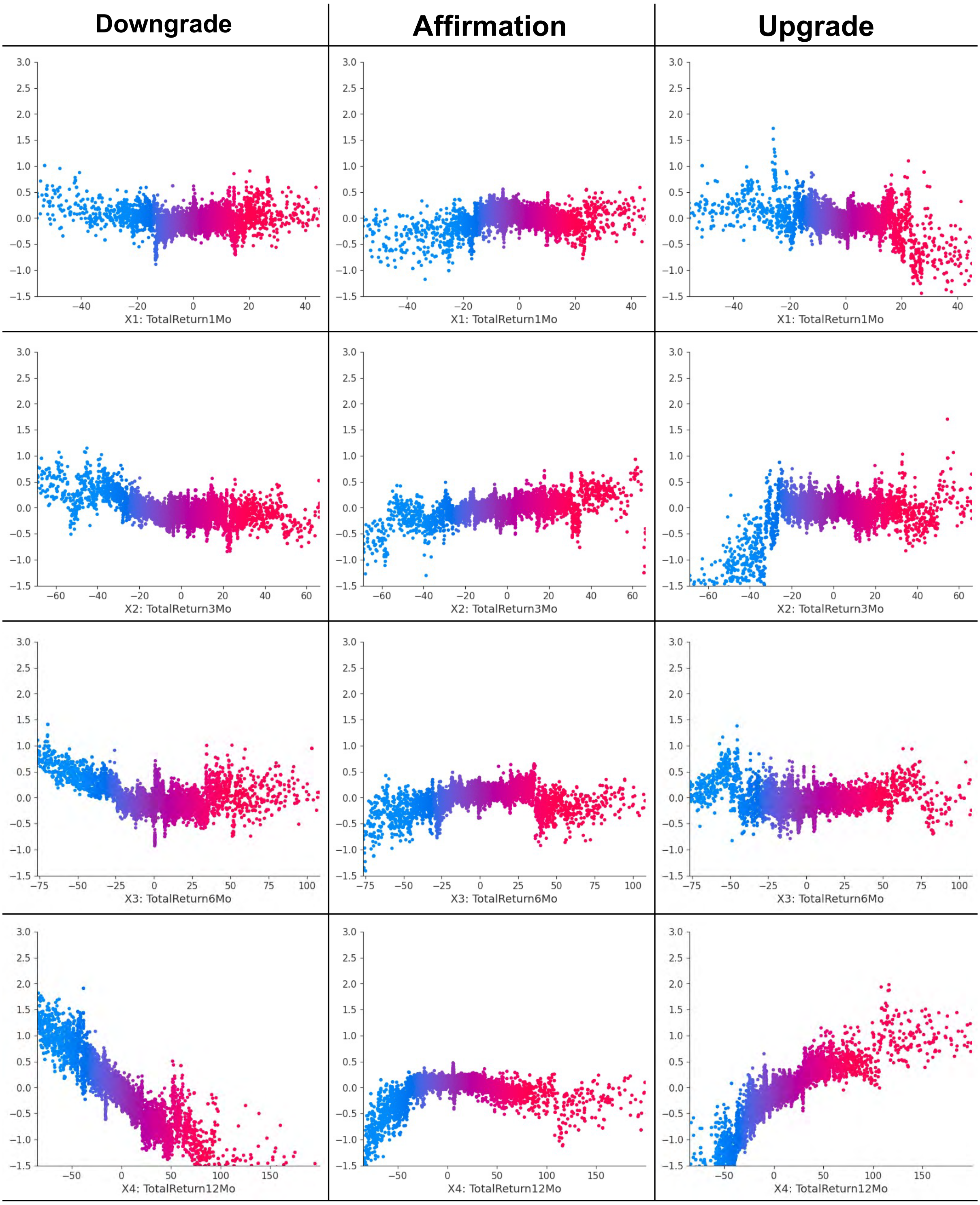
The plots indicate that negative values of *X4: Total Return 12 months* contribute more to predicting a *Downgrade* than positive values of *X4: Total Return 12 months* predict an *Upgrade*. This could be interpreted as the CRAs being more reluctant to upgrade a credit rating when the total stock return is positive than downgrading a credit rating when the total stock return is negative. This is contrary to the initial assumption that CRAs favour issuers over investors.

Closer Look at the Trailing 12 Months Total Return Feature The feature *X4: Total Return 12 months* is further investigated as it is the feature with the most considerable contribution, measured as the highest absolute mean SHAP value.

SHAP dependence plots show the effect of a single feature on the model prediction, which in this case, is the relationship between the feature $X4$ and the credit rating update action. Each point denotes a single prediction, with the feature value on the x-axis, the SHAP value on the y-axis, and the colour representing the feature value.

The plot is a global explanation method, as it considers all cases and plots them. To investigate this relationship further, one must look at the dependence plots. The following dependence scatter plots show the feature for Total Return, $X1$, $X2$, $X3$ and $X4$. The plot shows the feature's marginal effect on the model's predicted outcome. The plots identify any linear, monotonic, or more complex relationship effect between the feature and the target.

Figure 6.8: Plot of SHAP Dependence of Total Return Features



The dependence plots in Figure 6.8 show that the feature *X4: Total Return 12 months* can be described as a linear effect between the feature and the credit rating update action. For *Downgrades*, the *X4* feature has a negative linear relationship, meaning increasing values for returns results in lower SHAP values, and vice versa for *Upgrades.* These linear relationships are expected; however, these linear relationships are not as prominent in the other Total Return features ($X1$, $X2$, $X3$).

The spurious non-linear relationship between the Total Return features ($X1$, $X2$, $X3$) and the target variable could be because the total stock return trailing for a period less than 12 months contains less information or that short total return periods are volatile.

6.4.3 Investigating Revenue Models

The ML model is trained on data from March 2016 to September 2022. Updates from E-J were unavailable to gather from Refinitiv before this period. Two models are trained, split by revenue models. The first model uses data from E-J only, representing the investor-paid revenue model. The second model is trained with data from Moody's and Fitch, representing the issuer-paid revenue model.

The results are shown in Figure 6.9 and 6.10, with SHAP values for each corresponding model. The SHAP values are low for the class *Affirmation* for both models, revealing that the features poorly describe this class. However, this could be caused by relatively more *Downgrades* and *Upgrades* in this limited period.

The most important feature in the issuer-paid revenue model is undoubtedly *X4: Total Return 12 months*, and is similar to the previous model. Surprisingly, and perhaps the most exciting discovery in the analysis is that the E-J model indicates that the *X2: Total Return 3 months* contributes the most. The individual SHAP plots per class are in Appendix A3.

The nine-month trailing total return difference between the two revenue models could indicate that investor-paid revenue agencies, represented here by E-J, update their ratings earlier. E-J updates its corporate ratings monthly using a quantitative screening method, RCA, to identify possible changes in credit quality, recall Section 2.2.7. Not involving the issuers allows for more frequent assessments, which could explain why their assessments are closer to market fluctuations.

Figure 6.9: Plot of SHAP Values for Egan-Jones

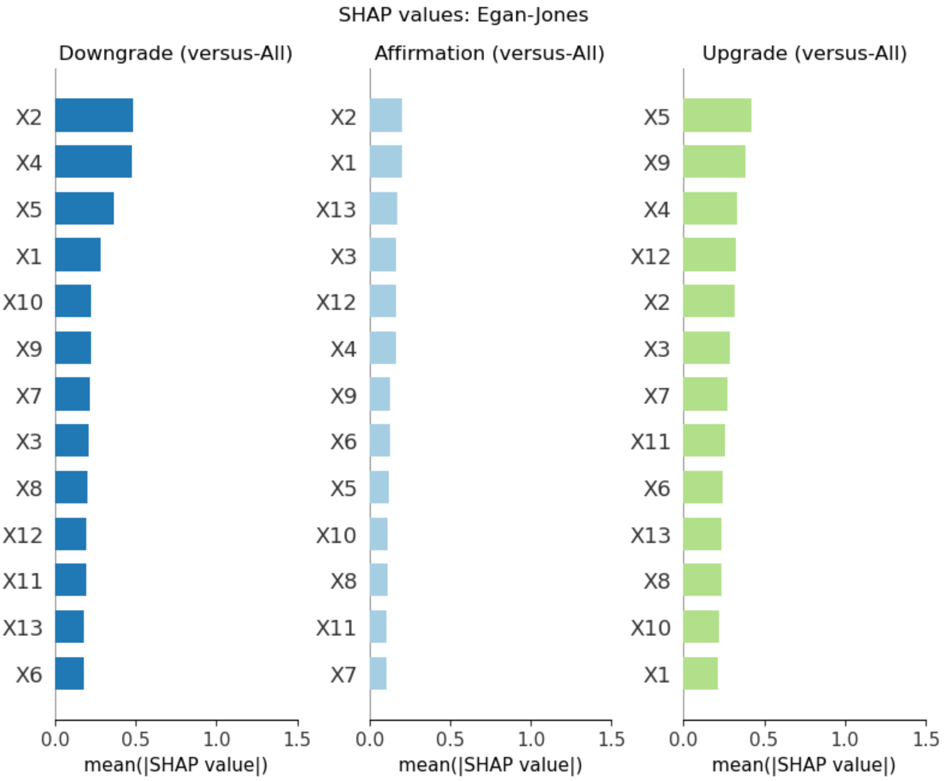
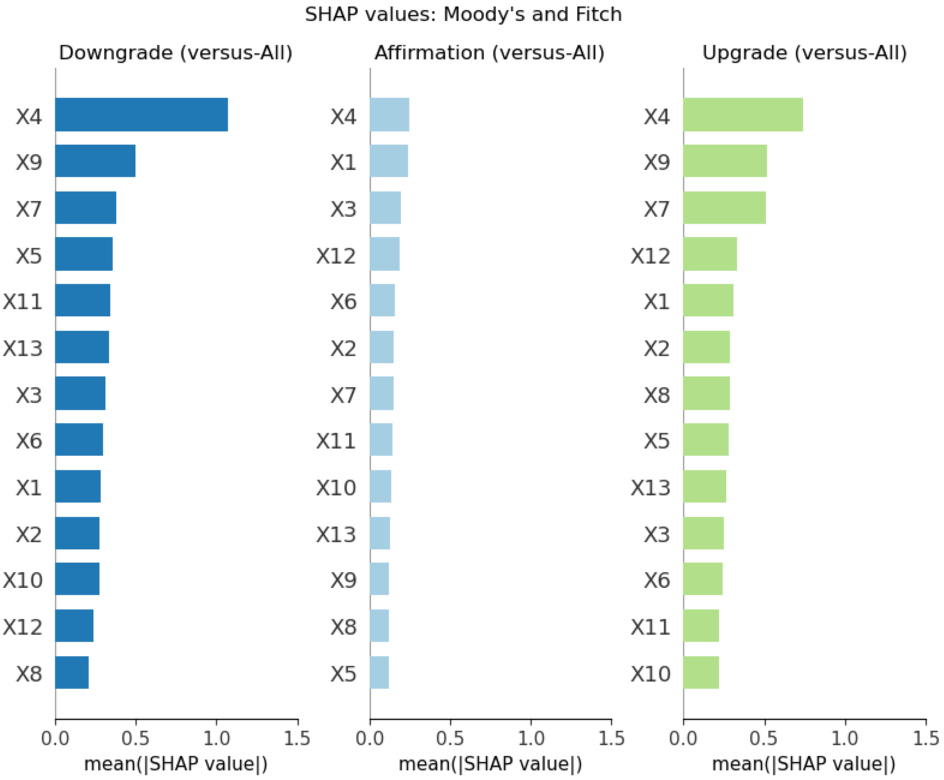


Figure 6.10: Plot of SHAP Values for Moody's and Fitch



6.4.4 Investigating The Wirecard Scandal of 2020

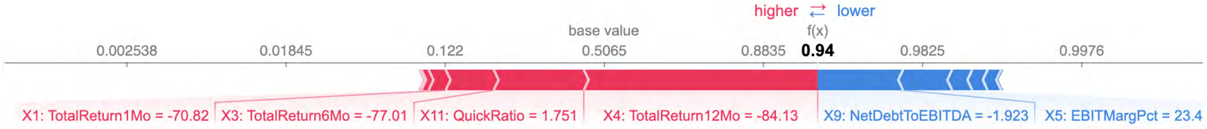
Interestingly, the data set contains cases of Wirecard's credit rating updates, enabling in-depth and exciting analysis of the months leading up to the bankruptcy in June 2020. Moody's downgraded Wirecard from investment grade (*Baa3*) to junk status (*B3*) on the 19th of June 2020. Moody's was criticised for its slow action, as there were indications of corporate decline prior to the unfolding of the scandal.

We create a synthetic case to investigate local SHAP values and determine if our model would predict a *Downgrade*. The synthetic case is an Update on the 19th of May 2020, a month prior to the actual case observed on the 19th of June 2020. The financial ratios are equal because the cases are within the same quarter (Q2 2020). The only difference in feature values is the trailing total return features, resembling a ceteris paribus analysis.

6.4.4.1 Case: 19th of June 2020 The model's train data is filtered to include Moody's observations from 2016 to 2020. The local SHAP values of the Wirecard update on the 19th of June 2020 are summarised in a force plot in Figure 6.11.

The force plot explains a single prediction of the model by illustrating how the features contributed to the predictive outcome of a given case. The plot's name derives from the plot illustrating each feature's force on the predicted output.

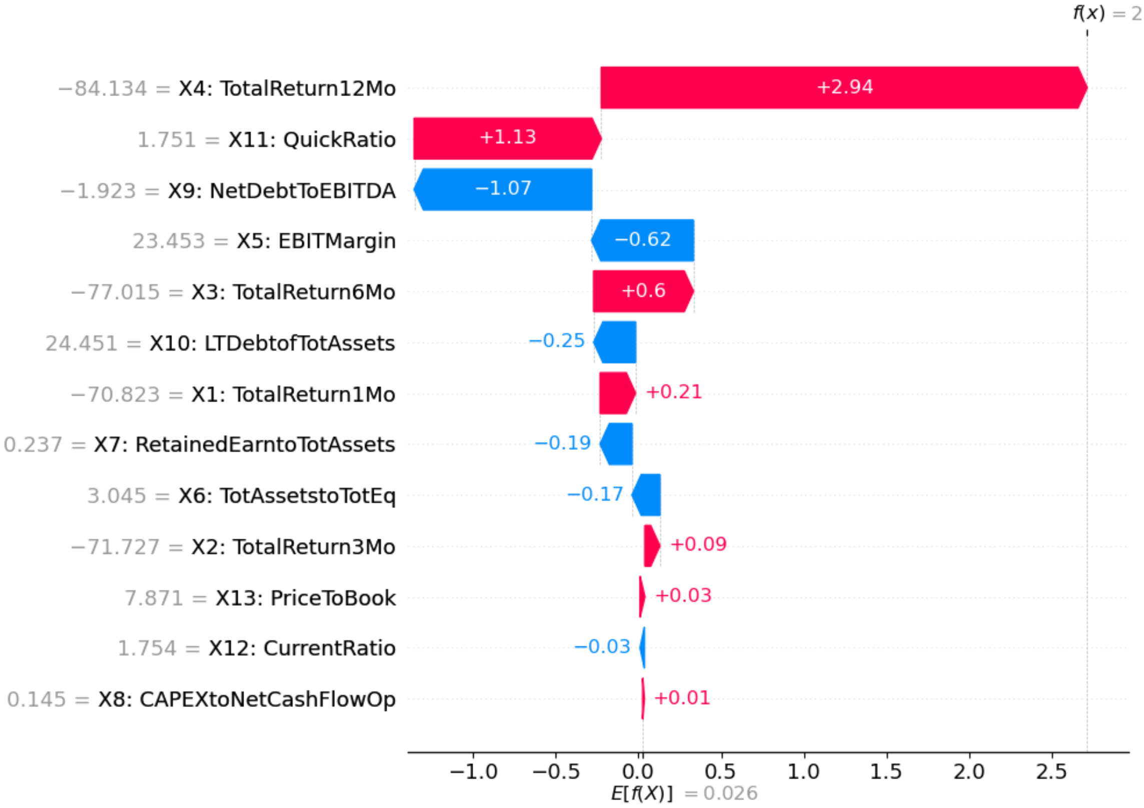
Figure 6.11: Plot of SHAP Force - Class: Affirmation



The model predicts that this case belongs to the *Downgrade* class with a 94% probability, which is a true positive, as the bond was downgraded on the 19th of June 2020. The predicted probability for class *Affirmation* and *Upgrade* is 9% and 1%, respectively. The base value for *Downgrade* is 50.65%, which is equal to the class probability in a feature-less model. In other words, the base value is the average of all the model's predicted outputs. The base value for *Affirmation* is 76.91% and 51.69% for *Upgrade*.

The force plot's most important features for the given prediction are coloured blue or red. Red denotes features that push toward the given class, while blue denotes features that push in the other classes' direction. The features with the most significant impact on the score are located on the line between the red and blue scores, and the bars' sizes illustrate that feature's total impact on the prediction.

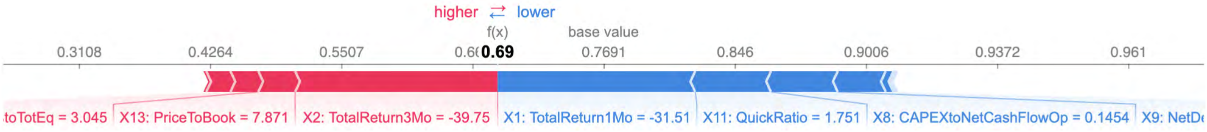
Figure 6.12: Plot of SHAP Waterfall - Class: Downgrade



An alternative visualisation of 6.11, a waterfall plot, is shown in Figure 6.12. The waterfall plot is an alternative visualisation to the force plot. The waterfall plot provides SHAP values for all model features instead of the predicted probabilities. The features are ranked in order of importance. The feature importance shows that feature *X4: Total Return 12 months* is the most important and pushes the case to the class *Downgrade*. Features $X12$ and $X11$ have positive SHAP values of 2.94 and 1.13, respectively, which contribute to predicting the case to be a *Downgrade*. On the other hand, feature *X9: Net DEBT to EBITDA* pushes the predicted case away from class *Downgrade* with a negative value of 1.07.

6.4.4.2 Synthetic Case: 19th of May 2020 A synthetic case has been created to explore the claim that Moody's should have downgraded the corporate bond on Wirecard earlier. The synthetic case tests to see if Moody's, according to the model, should have downgraded Wirecard on the 19$^{\text{th}}$ of May 2020. The only difference between the case on the 19$^{\text{th}}$ of June 2020 and the 19$^{\text{th}}$ of May 2020 is the Total Return features ($X1$, $X2$, $X3$ and $X4$).

Figure 6.13: Plot of SHAP Force, Synthetic Case - Class: Affirmation



The force plot in Figure 6.13 shows that the model's predicted probability of the synthetic case belonging to class *Affirmation* is 69%. The predicted probability for class *Downgrade* and *Upgrade* is 50% and 1%, respectively. Accordingly, the model classifies the synthetic case as *Affirmation*.

Even though the value of the features *X2: Total Return Last 3 Months* with -39.75% and *X1: Total Return Last 1 Month*, with -31.51% are both negative, they force the case to belong in the *Affirmation* case differently. The $X2$ feature pushes the case towards the *Affirmation*, while the return for the last three months (shorter-term $X3$) pushes it toward a *Downgrade* or *Upgrade*.

Figure 6.14: Plot of SHAP Force, Synthetic Case - Class: Downgrade

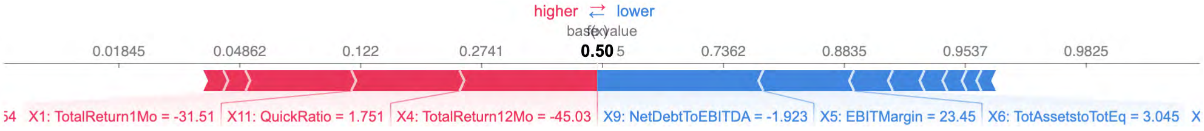


Figure 6.14 illustrates that the synthetic case has a 50% probability of belonging to class *Downgrade*. The negative *X4: Total Return Last 12 Months* is pushing the case towards a Downgrade, while $X9$: *Net Debt to EBITDA* is pushing it towards *Affirmation* or *Upgrade*.

The model predicts that the synthetic case belongs to classes *Affirmation*, *Downgrade*, and *Upgrade* with a 69%, 50% and 1% probability, respectively. In conclusion, the ML model was aligned with Moody's assessment. Remember that the data used to predict the class is based on data from Moody's, making it possible that the actual probabilities were further or closer to the predicted values.

6.5 Discussion

The models have demonstrated the ability to predict the CRAs' credit rating updates. Remember, the models are just predicting the actions of the CRAs rather than the factual truth of credit risk. It is essential to remember that not having complete insight into the rating processes and the Benchmarking Paradox may not give fully reliable predictions.

6.5.1 Critics

6.5.1.1 Fallacies The McNamara fallacy, also known as the quantitative fallacy, states that deciding solely on quantitative observations or metrics and ignoring all other possible explanations leads to flawed reasoning Fischer (1970). We hope to have overcome this fallacy by including other studies and thorough background analysis. However, it is worth noting that most studies in this context field are primarily quantitative.

Goodhart's law is an economics principle that when a measure becomes a target, it ceases to be a good measure Goodhart (1984). In the context of credit ratings, when investors begin using credit ratings as the decisionmaker in their investment decisions, issuers may begin taking actions to raise their credit ratings, even if doing so is not in the best interests in terms of issuers' underlying business or financial health.

6.5.1.2 Machine Learning and Explainability Due to its accuracy, a machine learning model was used to predict credit ratings. However, when using ML methods, some explainability is lost. Moreover, Rosé et al. (2019) argue that ML models alone are insufficient to use as explanatory models and require interdisciplinary understanding.

A potential pitfall with using an ML model as an explanatory model is what is typically referred to as the «Rashomon effect». The effect occurs when different models with the same performance typically have identical predictive performances and base their predictions on entirely different effects and relations from the same data (Breiman, 2001).

6.5.1.3 SHAP Values Slack et al. (2019) criticises post-hoc explanation techniques such as SHAP values and proves that these techniques can easily be modified intentionally to create misleading interpretations. Kumar et al. (2020) criticises that SHAP values «do not provide explanations which suit human-centric goals of explainability». We have considered this criticism by conducting a thorough literature review and being cautious in reaching conclusions from the SHAP values alone.

6.5.1.4 Garbage In, Garbage Out We use financial ratios as features in the ML model. However, this requires the financial ratio to be reliable and correct. If the model were to be used by an analyst to predict rating actions, it is essential to remember the *Garbage in, garbage out* problem. As observed in the Wirecard scandal, the main problem was the incorrect accounting numbers approved by auditors at Ernst & Young. The ML cannot detect any inaccurate accounting data, and training on these data will produce inaccurate results. In the case of Wirecard, the balance was falsely high by at least EUR 1.9 billion, and the model will consequently not assign a downgrade.

The trailing total return features were included to reduce reliance on financial ratios to detect any indications of decline that financial data cannot explain. However, it is not unproblematic to use these features, Even though the total return indicates the market consensus, some extreme cases can cause falsely low or high stock returns. For instance, in January 2021, the GameStop stock rocketed overnight due to a short squeeze by social media users and hedge funds. Despite years of financial losses and a record-breaking net loss of USD 673 million (in 2019), the share price increased by 1500% over a period of two weeks.

The SHAP dependence plot (6.8) in Section 6.4.3 explains that high values for the return features contribute to predicting an upgrade. Due to the high SHAP value of the return features, the 1500% increase in the GameStop stock would likely be predicted to be an upgrade in our XGBoost model, despite declining financial ratios and no substantial operational improvements.

To summarise, ML models alone cannot be used uncritically as the algorithm blindly trusts the inputs, and deceptive inputs can cause deceptive and inaccurate predictions.

6.5.2 Use Cases of the Thesis

We have used this XGBoost model to explain credit rating updates with SHAP values. However, we have also demonstrated that it can be used in predicting the outcome of an update using financial data and total returns with an 84.25% accuracy and an MCC of 66.67%.

An alternative application of our model is that institutional investors can use it (1) to classify the bonds in their funds, (2) to be prepared to sell poorly performing bonds and possible fallen angels before CRAs downgrade them. Moreover, our model and results highlight the importance of the issuer's total returns in bond performance and the model can be used in local explanations to understand the mechanism of anomalies like the Wirecard scandal and GameStop debacle.

6.5.3 Further Research

Including data from the most established CRA, S&P, could yield more accurate results in further research. Due to licensing limitations, we could not collect historic rating action from S&P; therefore, both the methodology and ratings from S&P are excluded from the data set.

We utilised the feature results from the study of Yang et al. (2020), which uses Principal Component Analysis and added four more variables with surprisingly high importance, meaning there could be other unconsidered features with even higher importance. For further research, we suggest that an extensive set of potential features are collected and then use a Principal Component Analysis to pick the best-performing features in model accuracy.

Furthermore, we suggest that the same ML model could be applied to sector-specific data sets if large sets are obtained. Our methods are not sector-based, unlike the methods of the CRAs, and if sectors were considered in the model, it could potentially increase accuracy. Alternatively, we suggest checking the accuracy and feature ranking between the different sectors and seeing if there is a significant difference between the models. We suggest bundling several models together with sector-specific features if there is a difference in model performance across the different sector models.

We did not consider rating watchlists and outlooks. A more precise analysis would include these rating outlooks, but the data is not publicly available. We would also like to point out that many investors are not following these outlooks as closely as they follow the credit ratings. These outlooks are neither used where credit ratings are part of regulatory risk assessment, such as in the regulations in Basel III. Also, looking at rating outlooks, preferably in the same model, these outlooks contain much financial information, so perhaps the lag would not be so sizeable.

Lastly, we suggest including advanced mechanisms of credit risk in the models, such as predicting Credit Default Swap Spreads or including a momentum or an age feature. There is some momentum in the credit rating's direction, meaning an upgrade and vice versa will often follow an upgrade. Additionally, there could be an age effect on the credit ratings, meaning how long an issuer has had a specific rating is a factor for their current rating. Both the momentum and the age feature should be considered in further research.

7 Conclusion

Machine learning techniques for predicting credit rating updates performed well, we obtained an accuracy of 84.4%. Additionally, local and global SHAP values proved to be powerful tools for understanding machine learning models' behaviour and providing transparent and interpretable explanations of its predictions.

Training models on historical financial data make it possible to capture complex patterns and relationships that may not be immediately apparent to analysts. The results of this study demonstrate that machine learning models work well in predicting credit rating updates, particularly when incorporating a wide range of relevant features and utilising advanced techniques such as ensemble methods.

Local SHAP values explain the prediction for a specific instance, while global SHAP values explain the model as a whole. Local SHAP values can be more helpful in understanding why a particular prediction was made, like in the Wirecard case. While global SHAP values can help identify patterns in the data and the overall behaviour of the model like the *Trailing 12 months Total Return* feature importance.

We included four total return features as proxies to justify the absence of qualitative assessments in the machine learning model and to see if different time horizons have different importance in predictions. The SHAP values indicated that the trailing 12 months total return feature was most important when predicting credit rating updates, providing tangible evidence of the allegations of CRAs trailing the market. Additionally, we investigated the two revenue models, the investor and issuer-paid model, to see if there were any differences in lag. The global SHAP values indicated that the total returns trailing three months were most important to investor-paid, Egan-Jones, and the 12-month returns were most important for issuer-paid agencies Moody's and Fitch.

Lastly, we used SHAP values in local explanations to investigate the Wirecard Scandal of 2020. The model predicted that the case belongs in the *Downgrade* class with a 94% probability. Moreover, we were curious to see if our model could outperform Moody's assessment prior to Wirecard's default in June 2020. Using the last quarter's financial data, we created a synthetic case one month before the last case. Our model's predicted probability of the synthetic case belonging to class *Affirmation* is 69%. The predicted probability for class *Downgrade* and *Upgrade* is 50% and 1%, respectively, meaning our model could not affirmatively predict a *Downgrade* before Moody's. However, the results provide some viability to our model.

Although, it is essential to note that the success of ML models in this context is highly dependent on the quality and quantity of the data used for training and testing. In order to achieve optimal performance, it was necessary to pre-process and clean the data carefully and to pay attention to issues such as class imbalance and overfitting.

Overall, using machine learning in credit rating prediction can significantly enhance the efficiency and effectiveness of credit risk analysis, and this study provides valuable insights into the design and implementation of such models. Further research could explore using additional data sources and machine learning techniques and integrate these models into real-world credit risk management systems.

«There are two superpowers in the world today, in my opinion. There's the United States, and there's Moody's Bond Rating Service. The United States can destroy you by dropping bombs, and Moody's can destroy you by downgrading your bonds. And believe me, it's not clear sometimes who's more powerful.»

Pulitzer Prize winning-journalist Thomas Friedman, February 1996

References

Aas, K., Jullum, M., and Løland, A. (2021). Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artificial Intelligence*, 298:103502.

Ahn, M., Bonsall, S. B., and Buskirk, A. V. (2019). Do managers withhold bad news from credit rating agencies? *Review of Accounting Studies*, 24:972–1021.

Akerlof, G. A. (1970). The Market for Lemons: Quality Uncertainty and the Market Mechanism. *Quarterly Journal of Economics*, 84:488–500.

Arrieta, A. B., Díaz-Rodríguez, N., Ser, J. D., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., and Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115.

Baker Library (2022). Lehman Brothers 1850-2008 - Bankruptcy. https://www.library.hbs.edu/hc/lehman/exhibition/bankruptcy.

Becker, B. and Milbourn, T. (2011). How did Increased Competition Affect Credit Ratings? *Journal of Financial Economics*, 101:493–514.

Belloni, M., Helmersson, T., Jarmuzek, M., Mosk, B., and Nikolic, F. (2020). Understanding What Happens When “Angels Fall”. https://www.ecb.europa.eu/pub/financial-stability/fsr/focus/2020/html/ecb.fsrbox202011_03~578f4f74dc.en.html.

Berk, J. and DeMarzo, P. (2019). *Corporate Finance.* Pearson Education, 5th edition.

Biecek, P. and Burzykowski, T. (2021). *Explanatory Model Analysis.* Chapman and Hall. https://ema.drwhy.ai.

Bodie, Z., Kane, A., and Marcus, A. J. (2022). *Essentials of Investments.* McGraw Hill, 12th edition.

Bonsall, S. B. (2014). The impact of issuer-pay on corporate bond rating properties: Evidence from Moodys and SPs initial adoptions. *Journal of Accounting and Economics*, 57:89–109.

Bonsall, S. B., Gillette, J., Pundrich, G., and So, E. C. (2022). Conflicts of Interest in Subscriber-Paid Credit Ratings. *SSRN Electronic Journal.*

Booth, A. and de Bruin, B. (2019). Stakes Sensitivity and Credit Rating: A New Challenge for Regulators. *Journal of Business Ethics*, 169:169–179.

Breiman, L. (2001). Statistical Modeling: The Two Cultures. *Statistical Science*, 16:199–231.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression trees.* CRC Press.

Chang, Y. C., Chang, K. H., and Wu, G. J. (2018). Application of eXtreme gradient boosting trees in the construction of credit risk assessment models for financial institutions. *Applied Soft Computing Journal*, 73:914–920.

Chen, H., Janizek, J. D., Lundberg, S., and Lee, S.-I. (2020). True to the Model or True to the Data? *arXiv*.

Chen, T. and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *arXiv*.

Committee on the Judiciary (2012). House of Representatives Hearing: Dodd-Frank Act's Effects on Financial Services Competition. https://www.govinfo.gov/content/pkg/CHRG-112hhrg74976/pdf/CHRG-112hhrg74976.pdf.

Darmouni, O. and Papoutsi, M. (2022). The Rise of Bond Financing in Europe. *European Central Bank*.

Department of Justice (2015). Justice Department and State Partners Secure $1.375 Billion Settlement with S&P for Defrauding Investors in the Lead Up to the Financial Crisis. https://www.justice.gov/opa/pr/justice-department-and-state-partners-secure-1375-billion-settlement-sp-defrauding-investors.

Doshi-Velez, F. and Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. *arXiv*.

Easley, D., Hvidkjaer, S., and O'Hara, M. (2010). Factoring Information into Returns. *The Journal of Financial and Quantitative Analysis*, 45:293–309.

Edwards, H. (2013). CRA3 and the liability of rating agencies: inconsistent messages from the regulation on credit rating agencies in Europe. *Law and Financial Markets Review*, 7:186–191.

Egan-Jones Ratings Company (2020). Methodologies for Determining Credit Ratings. https://www.egan-jones.com/media/032h3y52/main-methodology-04-28-2020.pdf.

Eisenhardt, K. M. (1989). Agency Theory: An Assessment and Review. *The Academy of Management Review*, 14:57.

ESMA (2021). Report on CRA Market Share Calculation. *European Securities and Markets Authority*.

European Commission (2016). Study on the State of the Credit Rating Market – Final Report. https://ec.europa.eu/info/sites/default/files/state-of-credit-rating-market-study-01012016_en.pdf.

European Union (2019). Regulation (EC) No 1060/2009 of the European Parliament and of the Council of 16 September 2009 on credit rating agencies. *Official Journal of the European Union*, 302.

Fama, E. F. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. *The Journal of Finance*, 25:383.

Financial Times (2022). Inside Wirecard. https://www.ft.com/wirecard.

Fischer, D. H. (1970). *Historians' fallacies; toward a logic of historical thought*. Harper & Row.

Freund, Y. and Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, 55:119–139.

Frost, C. A. (2007). Credit Rating Agencies in Capital Markets: A Review of Research Evidence on Selected Criticisms of the Agencies. *Journal of Accounting, Auditing & Finance*, 22:469–492.

Gelderblom, O., Jong, A. D., and Jonker, J. (2013). The Formative Years of the Modern Corporation: The Dutch East India Company VOC, 1602–1623. *The Journal of Economic History*, 73:1050–1076.

Golbayani, P., Florescu, I., and Chatterjee, R. (2020). A comparative study of forecasting Corporate Credit Ratings using Neural Networks, Support Vector Machines, and Decision Trees.

Goodhart, C. A. E. (1984). Problems of Monetary Management: The UK Experience. *Monetary Theory and Practice*, pages 91–121.

Guo, X., Zhu, Z., and Shi, J.-L. (2012). A Corporate Credit Rating Model Using Support Vector Domain Combined with Fuzzy Clustering Algorithm. *Mathematical Problems in Engineering*, 2012:1–20.

Hand, D. J. and Till, R. J. (2001). A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. *Machine Learning*, 45:171–186.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer US, 2nd edition.

Hemraj, M. (2015). *Credit Rating Agencies*. Springer International Publishing.

Hite, G. and Warga, A. (1997). The effect of bond-rating changes on bond price performance. *Financial Analysts Journal*, 53.

ICMA (2020). Bond Market Size. https://www.icmagroup.org/market-practice-and-regulatory-policy/secondary-markets/bond-market-size/.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2021). *An Introduction to Statistical Learning*. Springer US, 2 edition.

Janzing, D., Minorics, L., and Blöbaum, P. (2019). Feature Relevance Quantification in Explainable AI: A Causal Problem. *arXiv*.

Jiang, J. X., Stanford, M. H., and Xie, Y. (2012). Does it matter who pays for bond ratings? Historical evidence. *Journal of Financial Economics*, 105:607–621.

Kashyap, A. K. and Kovrijnykh, N. (2016). Who Should Pay for Credit Ratings and How? *Review of Financial Studies*, 29:420–456.

Kim, K. J. and Ahn, H. (2012). A corporate credit rating model using multi-class support vector machines with an ordinal pairwise partitioning approach. *Computers & Operations Research*, 39:1800–1811.

Kothari, S. P., Shu, S., and Wysocki, P. D. (2009). Do Managers Withhold Bad News? *Journal of Accounting Research*, 47:241–276.

Kronlund, M. (2019). Do Bond Issuers Shop for Favorable Credit Ratings? *SSRN Electronic Journal*.

Kuhn, M. (2014). Futility Analysis in the Cross-Validation of Machine Learning Models. *arXiv*.

Kuhn, M. and Silge, J. (2022). *Tidy Modeling with R: A Framework for Modeling in the Tidyverse*. O'Reilly Media, Inc. https://tmwr.org.

Kumar, I. E., Venkatasubramanian, S., Scheidegger, C., and Friedler, S. A. (2020). Problems with Shapley-value-based explanations as feature importance measures. *arXiv*.

Langohr, H. M. and Langohr, P. T. (2009). *The Rating Agencies and Their Credit Ratings: What They Are, How They Work, and Why They are Relevant*. Wiley.

Lloyd S. Shapley (1952). A Value for N-Person Games. *RAND Corporation*.

Lougee, B. A. and Marquardt, C. A. (2004). Earnings Informativeness and Strategic Disclosure: An Empirical Examination of "Pro Forma" Earnings. *The Accounting Review*, 79:769–795.

Lucchetti, A. (2008). Tiny Firm Gives Ratings Giants Another Worry. *The Wall Street Journal*. https://www.wsj.com/articles/SB120251672233155415.

Lundberg, S. M., Erion, G. G., and Lee, S.-I. (2018). Consistent Individualized Feature Attribution for Tree Ensembles. *arXiv*.

Lundberg, S. M. and Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*.

Malewska, A. (2021). Failed Attempt to Break Up the Oligopoly in Sovereign Credit Rating Market after Financial Crises. *Contemporary Economics*, 15:152–163.

Mashrur, A., Luo, W., Zaidi, N. A., and Robles-Kelly, A. (2020). Machine Learning for Financial Risk Management: A Survey. *Institute of Electrical and Electronics Engineers*, 8.

Merton, R. C. (1974). On the Pricing of Corporate Debt: The Risk Structure of Interest Rates. *The Journal of Finance*, 29:449–470.

Mishkin, F. S. and Eakins, S. G. (2018). *Financial Markets and Institutions, Global Edition*. Pearson Education, 9th edition.

Molnar, C. (2022). *Interpretable Machine Learning*. Independently published, 2nd edition. https://christophm.github.io/interpretable-ml-book/.

Moody's (2022). Research and Ratings: List of Rating Methodologies. https://www.moodys.com/researchandratings/home.

Morgenson, G. (2008). Debt Watchdogs: Tamed or Caught Napping? *The New York Times*. https://www.nytimes.com/2008/12/07/business/07rating.html.

Nevasalmi, L. (2020). Forecasting multinomial stock returns using machine learning methods. *The Journal of Finance and Data Science*, 6:86–106.

Nikulski, J. (2020). The Ultimate Guide to AdaBoost, random forests and XGBoost. https://towardsdatascience.com/the-ultimate-guide-to-adaboost-random-forests-and-xgboost-7f9327061c4f.

Nye, R. P. (2014). *Understanding and Managing the Credit Rating Agencies: A Guide for Fixed Income Issuers.* Euromoney Institutional Investor, 1st edition.

OECD (2010). Competition and Credit Rating Agencies. https://www.oecd.org/daf/competition/46825342.pdf.

Park, M. and Chai, S. (2021). *A Machine Learning-based Model for the Asymmetric Prediction of Accounting and Financial Information.* Springer.

Posner, R. A. (1969). Oligopoly and the Antitrust Laws: A Suggested Approach. *Stanford Law Review*, 21:1562.

Refinitiv (2022). Refinitiv Workspace. https://workspace.refinitiv.com/.

Rennison, J. (2020). US corporate bond issuance hits $1.919 tn in 2020. *Financial Times.* https://www.ft.com/content/a59c2a9d-5e0b-4cbc-b69e-a138de76a776.

Rosé, C. P., McLaughlin, E. A., Liu, R., and Koedinger, K. R. (2019). Explanatory learner models: Why machine learning (alone) is not the answer. *British Journal of Educational Technology*, 50:2943–2958.

Scalet, S. and Kelly, T. F. (2012). The Ethics of Credit Rating Agencies: What Happened and the Way Forward. *Journal of Business Ethics*, 111:477–490.

SEC (2022). 2021 Staff Report on Nationally Recognized Statistical Rating Organizations. https://www.sec.gov/ocr/ocr-reports-and-studies.html.

Sinclair, T. J. (2005). *The New Masters of Capital: American Bond Rating Agencies and the Politics of Creditworthiness.* Cornell University Press.

Slack, D., Hilgard, S., Jia, E., Singh, S., and Lakkaraju, H. (2019). Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods. *arXiv.*

Sylla, R. (2002). *An Historical Primer on the Business of Credit Rating.* Springer US.

Toscano, F. (2020). Does the Dodd-Frank Act reduce the conflict of interests of credit rating agencies? *Journal of Corporate Finance*, 62:101595.

Vaughan, D. (2022). Multiclass Averaging. *Tidymodels: Yardstick.* https://yardstick.tidymodels.org/articles/multiclass.html.

Wagner, N. (2008). *Credit Risk: Models, Derivatives, and Management.* Chapman and Hall.

Wallis, M., Kumar, K., and Gepp, A. (2019). Credit Rating Forecasting Using Machine Learning Techniques. *Managerial Perspectives on Intelligent Big Data Analytics.*

Xia, H. (2014). Can investor-paid credit rating agencies improve the information quality of issuer-paid rating agencies? *Journal of Financial Economics*, 111:450–468.

Yang, S., lonut Florescu, and Islam, M. T. (2020). Principal Component Analysis and Factor Analysis for Feature Selection in Credit Rating. *arXiv.*

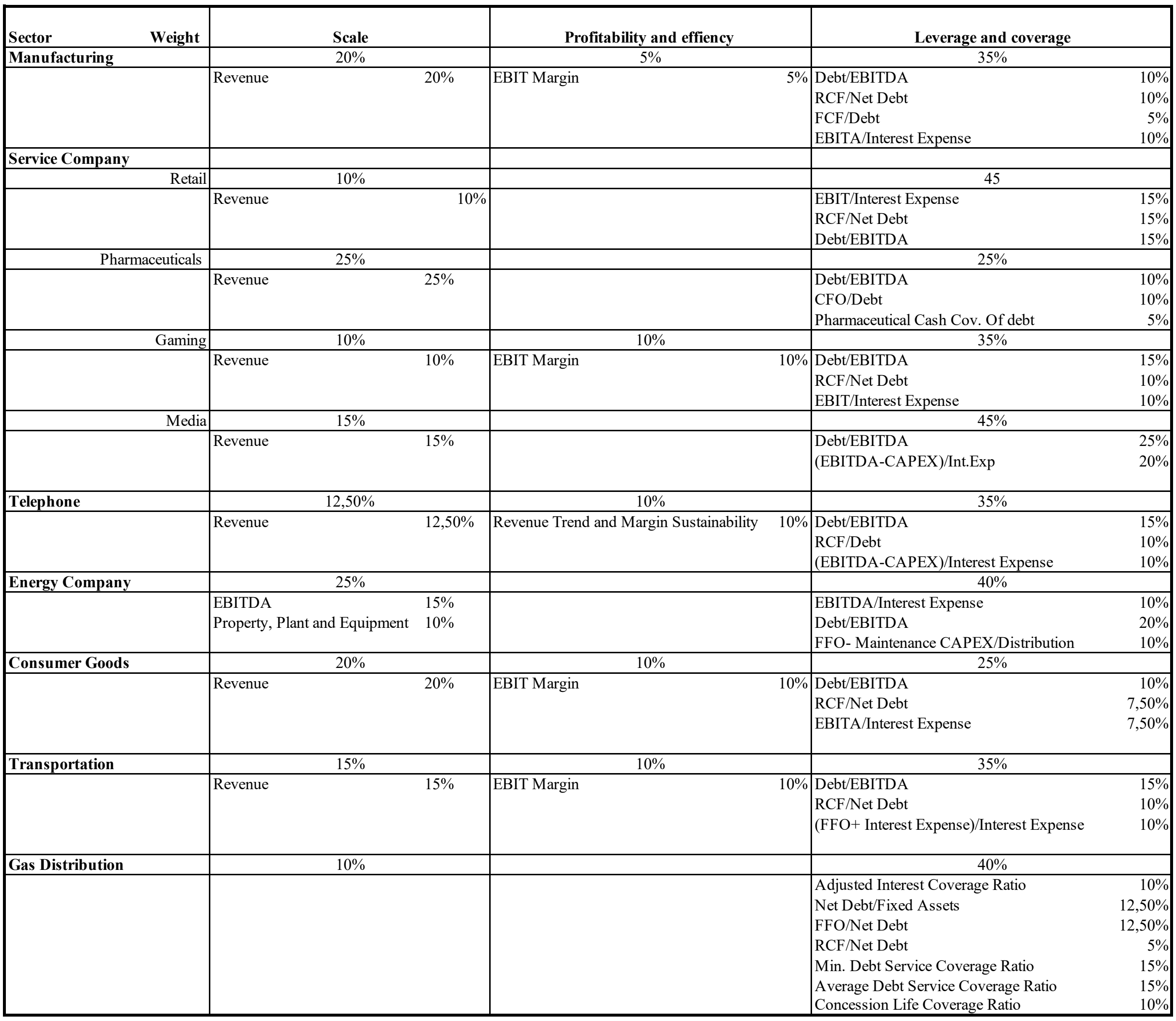
Ye, Y., Liu, S., and Li, J. (2008). A multiclass machine learning approach to credit rating prediction. *arXiv.*

Zhong, H., Miao, C., Shen, Z., and Feng, Y. (2014). Comparing the learning effectiveness of BP, ELM, I-ELM, and SVM for corporate credit ratings. *Neurocomputing*, 128:285–295.

Appendix

A1 Moody's Methodology

Figure A1.1: Assessment of Moody's Methodology



Sector	Weight	Scale		Profitability and effiency		Leverage and coverage	
Manufacturing		20%		5%		35%	
		Revenue	20%	EBIT Margin	5%	Debt/EBITDA	10%
						RCF/Net Debt	10%
						FCF/Debt	5%
						EBITA/Interest Expense	10%
Service Company							
	Retail	10%				45	
		Revenue	10%			EBIT/Interest Expense	15%
						RCF/Net Debt	15%
						Debt/EBITDA	15%
	Pharmaceuticals	25%				25%	
		Revenue	25%			Debt/EBITDA	10%
						CFO/Debt	10%
						Pharmaceutical Cash Cov. Of debt	5%
	Gaming	10%		10%		35%	
		Revenue	10%	EBIT Margin	10%	Debt/EBITDA	15%
						RCF/Net Debt	10%
						EBIT/Interest Expense	10%
	Media	15%				45%	
		Revenue	15%			Debt/EBITDA	25%
						(EBITDA-CAPEX)/Int.Exp	20%
Telephone		12,50%		10%		35%	
		Revenue	12,50%	Revenue Trend and Margin Sustainability	10%	Debt/EBITDA	15%
						RCF/Debt	10%
						(EBITDA-CAPEX)/Interest Expense	10%
Energy Company		25%				40%	
		EBITDA	15%			EBITDA/Interest Expense	10%
		Property, Plant and Equipment	10%			Debt/EBITDA	20%
						FFO- Maintenance CAPEX/Distribution	10%
Consumer Goods		20%		10%		25%	
		Revenue	20%	EBIT Margin	10%	Debt/EBITDA	10%
						RCF/Net Debt	7,50%
						EBITA/Interest Expense	7,50%
Transportation		15%		10%		35%	
		Revenue	15%	EBIT Margin	10%	Debt/EBITDA	15%
						RCF/Net Debt	10%
						(FFO+ Interest Expense)/Interest Expense	10%
Gas Distribution		10%				40%	
						Adjusted Interest Coverage Ratio	10%
						Net Debt/Fixed Assets	12,50%
						FFO/Net Debt	12,50%
						RCF/Net Debt	5%
						Min. Debt Service Coverage Ratio	15%
						Average Debt Service Coverage Ratio	15%
						Concession Life Coverage Ratio	10%

A2 Overview of Machine Learning Studies on Credit Ratings

Table A2.1: Overview of Machine Learning Studies on Credit Ratings

Authors (Year)	**Region**	**Features**	**Cases**	**Method**	**Results**
Kim and Ahn (2012)	South-Korea	14	Unspecified	Ordinal Multiclass Support Vector Machine (OMSVM), DAGSVM, ECOC, Weston & Watkins, Crammer & Singer	OMSVM: 67.98% Accuracy
Guo et al. (2012)	South-Korea, China	14	Unspecified	OMSVM, DAGSVM, Crammer & Singer	OMSVM
Wallis et al. (2019)	United States	27	Unspecified	Multinominal Logistic Regression, LDA, RDA, Artificial Neural Network (ANN), SVM, Gaussian Process Classifier, Random Forest (RF), Gradient Boosted Machine (GBM)	RF: 64.6% Accuracy ANN: 31.5% Cohen's Kappa
Ye et al. (2008)	Unspecified	33	2,541	Ordered Probit, Bagged Decision Tree, Multiclass SVM, Multiclass Proximal SVM	Multiclass SVM: 64% Accuracy
Zhong et al. (2014)	Unspecified	8	1,198	SVM, Extreme Learning Machine (ELM), Incremental ELM, Backpropagation (BP)	SVM is the best approach on Moody's data BP is the best approach on S&P data
Yang et al. (2020)	Unspecified	20	Unspecified	Principal Component Analysis	7 final features

A3 SHAP Values for Different Revenue Models

Figure A3.1: Plot of SHAP Values per Class: Egan-Jones

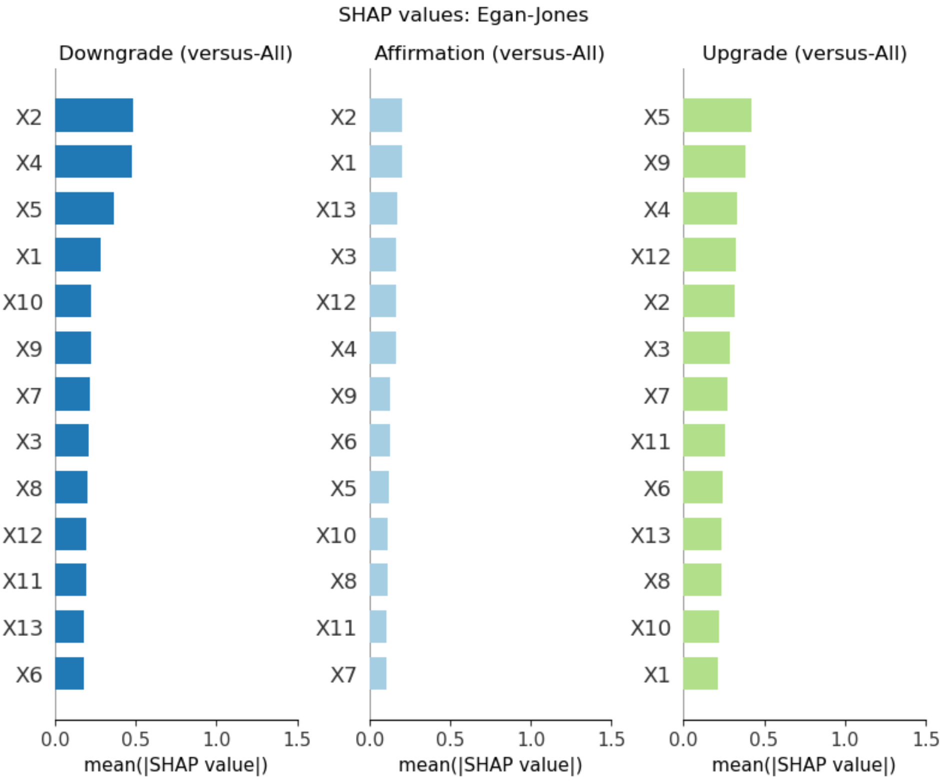
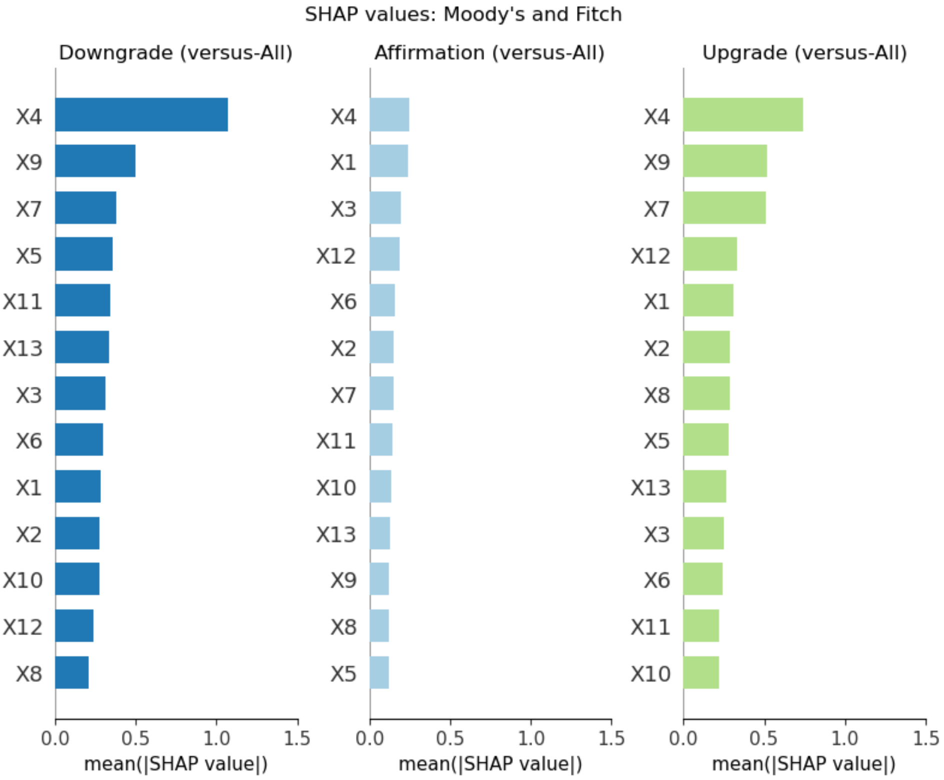


Figure A3.2: Plot of SHAP Values per Class: Moody's and Fitch


SHAP values: Moody's and Fitch
Downgrade (versus-All)
X4
X9
X7
X5
X11
X13
X3
X6
X1
X2
X10
X12
X8
Affirmation (versus-All)
X4
X1
X3
X12
X6
X2
X7
X11
X10
X13
X9
X8
X5
Upgrade (versus-All)
X4
X9
X7
X12
X1
X2
X8
X5
X13
X3
X6
X11
X10
0.0
0.5
1.0
1.5
mean(|SHAP value|)