Predicting regular and special dividends using machine learning

Gaurav Soni (IIT Kanpur)

Prof. Harshal Rajan Mulay (IIT Kanpur)

Prof. Suman Saurabh (IIT Kanpur)

Prof. Parvati Neelakantan

Abstract:

Dividend payout is a crucial corporate finance decision. Predicting whether a stock will give a dividend or not is, therefore, of critical importance. In this research, we forecast equity dividends, special dividends, and equity dividend growth using machine learning techniques employing a large number of firm-year observations and sixty-nine features. We utilized logistic regression, decision trees, random forests, XGBoost, and Artificial neural networks for this purpose. Among these, tree-based models demonstrated the best performance. Random Forest and XGBoost are found to be the best models for the prediction task. Random forest gives 96.7% and 95.12% accuracies for equity dividends and equity dividend growth, respectively. While XGBoost gives 99.90% accuracy for special dividends prediction. Our models also exhibited excellent performance across all 10 performance metrics. Additionally, we analyzed feature importance for each model and target variable. To the best of our knowledge, our study shows the highest accuracies among all published works for equity dividends and equity dividend growth. Further we are the first to predict special dividends using machine learning techniques. And finally our models do not use any market related variables so these models can be used for unlisted firms as well.

1 Introduction

Dividend decisions are critical for both investors and firms. Investors exhibit varying preferences regarding dividend payments. While some favor companies that provide regular cash flows through dividends, others prefer firms that reinvest their earnings to generate long-term growth. Consequently, a company's dividend payout policy can significantly influence investor decision-making. From the firm's perspective, retaining profits is essential to meet capital needs and reduce reliance on costly external funding through equity or debt. This creates a strategic challenge to optimize the distribution of profits. The complexity of this decision underscores its importance in shaping corporate financial strategy.

The COVID-19 pandemic had a significant impact on dividend payouts, with many firms reducing or suspending their dividends to maintain liquidity during the height of the crisis. A study of 8,889 firms across G-12 countries revealed that the proportion of dividend suspensions and omissions was markedly higher during the pandemic. However, the majority of firms managed to maintain or even increase their dividends, likely as a means of signaling financial stability and positive prospects to investors (Ali, 2021).

However, after COVID-19, the situation has changed. Globally, dividend payouts reached a record $1.75 trillion US dollars in 2024, representing a 6.6% increase on a year-on-year basis, according to the Janus Henderson Global Dividend Index. This growth is well spread across sectors, including tech, banking, media, etc.. Globally, 88% of companies either raised or maintained their dividend payouts. The forecast for 2025 anticipates further growth of 5%, reaching a new record of $1.83 trillion US dollars, indicating sustained confidence in corporate earnings and a commitment to shareholder returns (Fund Society, 2025)

In India, companies are spending a large chunk of their profits on dividend payouts, rewarding their shareholders. The year 2024 witnessed a 12% year-on-year rise in dividend payouts, reaching a record high of INR 2.2 trillion, the highest in at least six years (Bhalerao, 2024). Many large Indian firms like Infosys, Vedanta, Coal India, and Hindustan Zinc have raised their dividends in recent years, with several of them reaching record-high amounts (Reuters, 2024; Shah, 2024)

Predicting whether a stock will pay a dividend is an important task. Some investors choose to invest in dividend-paying stocks. According to the Bird-in-Hand Theory, dividend-paying stocks are particularly favored by income-focused investors, such as retirees and conservative investors, who seek regular and predictable cash flows. These investors prioritize stability and lower risk, often preferring established companies that consistently distribute dividends. During market downturns, even the non-conservative investors prefer dividend-paying stocks because they are generally less volatile (WalletInvestor, 2022). As per Agency Theory, investors prefer dividends because paying dividends can reduce agency problems by limiting the discretionary cash available to managers (Jensen & Meckling, 1976). The importance of dividends is so significant that Benjamin Graham once stated, "The true investor will do better if he forgets about the stock market and pays attention to his dividend returns and the operating results of his companies" (Graham, 1947). While this sentiment might be considered outdated by some, but many investors still prefer dividends.

On the other hand, some investors do not prefer to invest in dividend-paying stocks. Reinvesting dividends can compound returns over time, appealing to long-term investors focused on wealth accumulation. .(WalletInvestor, 2022). The Tax Preference Theory posits that investors favor long-term capital gains due to tax advantages over receiving dividends, as dividends are taxed at the income tax rate, which is generally higher than the capital gains tax rate (Brennan, 1970; Litzenberger & Ramaswamy, 1979). According to Feldstein and Green (1979), businesses aiming to maximize the value of their shares could retain earnings instead of paying dividends and later distribute the funds to shareholders in a manner that allows for more favorable taxation as capital gains.

So, as discussed above, different investors have varying proclivities towards dividends, making it extremely important to predict a company's dividend policy in the future. Existing literature on predicting dividends using machine learning is scarce, particularly in the Indian context. Moreover, the existing studies primarily focus on listed companies. Further, they do not account for special dividends or dividend growth trajectories. Thus, in this paper, we employed various kinds of algorithms and methods to fill this gap.

We employ traditional statistical learning methods, advanced machine learning algorithms, and cutting-edge deep learning models on a large database to predict whether a company will change its dividend issuance. Further, dividends are classified as equity dividends and special dividends. Equity dividends are the regular periodic payouts. In contrast, special dividends are one-time distributions, typically resulting from extraordinary events, such as asset sales or exceptional earnings, and are not expected to recur regularly. In this study, we build separate models to predict both types of dividends. Additionally, we develop another model to predict whether a company's equity dividend will rise in the future. Our research provides a robust comparative analysis of these methods, offering novel insights into predictive accuracy and efficiency. The study also conducts a thorough examination of feature importance, identifying the factors that have the most significant impacts on equity dividend payment, special dividend decisions, and dividend growth trends. It provides useful information for investors, policymakers, and business decision-makers.

Our study is based on an extensive dataset comprising 55,708 unique listed and unlisted companies, spanning up to 37 years, resulting in a total of 283,162 firm-year observations. We utilize a total of 69 features, which can broadly be classified into categories such as liquidity, profitability, leverage, investments, expenses, dividend history, other firm characteristics, industry-level influences, and macroeconomic indicators. We used a diverse set of models, including Logistic Regression, Decision Tree, Random Forest, XGBoost, and Artificial Neural Networks, to achieve maximum predictive accuracy. For optimal model performance, hyperparameter tuning and regularization techniques likes Lasso, Ridge and Elastic Nets were used for each model.

We found that tree-based algorithms performed better than other models. For equity dividend prediction Random Forest classifier achieves the highest accuracy of 96.7%. For predicting the special dividend, XGBoost emerged as the best model, delivering an outstanding accuracy of 99.90%. For the dividend growth prediction, the Random Forest model again proved to be the best, with an accuracy of 95.12%. We also analyzed the ROC curves of all the models to showcase their discriminative power. Additionally, we considered 10 other performance metrics including precision, recall, specificity, sensitivity, and F1 score. The models perform extremely well in terms of all these metrics. These results highlight the robustness and predictive power of tree-based models.

Tree-based models demonstrated their superiority for our dataset due to their ability to effectively handle non-linear data. While the Logistic Regression is advantageous for inference, as it provides a direct insight into the nature and impact of each feature on the target variable, its performance was lower for prediction. However, it remains a reliable option, especially as we addressed multicollinearity and worked with a large-scale dataset.

Furthermore, we validated the necessity of machine learning models in this prediction problem. Dividend Smoothing Theory suggests that the dividends are highly autocorrelated, so simple Time Series Models may be sufficient to predict dividends. We compare the performance on two datasets, one including dividend history features and the other excluding them. Although dividend-history variables were found to be important for prediction, they were not the sole drivers of accuracy, as the other features also significantly contributed to the model's performance.

We contribute to the existing literature in several ways. First, our research is the first to forecast special dividends using machine learning models. Second, it is the first to categorize businesses according to whether or not their dividends will increase. Finally, our features do not include any market-related variables, allowing our models to be used for predicting dividends of unlisted companies as well.

The structure of the paper is organized as follows: section 2 reviews the relevant literature on dividend prediction. In section 3, we detail the data collection process and the criteria for sample selection. section 4 presents the methodology used in the study. The results of our analysis are discussed in section 5, and section 6 concludes with a summary of key findings along with industry applicability of this study and scope for future research.

2 Literature review

Numerous classical theories regarding dividend payout policies offer differing perspectives. Some argue that dividend policy has no effect on the value of the firm, while others advocate for or against paying dividends. This topic has always been a subject of debate among academics and practitioners. The classical theories about dividend policies are discussed in Section 2.1. Additionally, many research articles highlight the influence of various financial factors, such as profitability, liquidity, and leverage, as well as economic conditions, board structure, corporate governance, macroeconomic factors, and industry-specific dynamics on dividend policies. These elements, which can significantly shape a firm’s approach to dividends, are comprehensively reviewed in Section 2.2. Furthermore, Section 2.3 examines similar machine learning based studies, drawing comparisons to provide a holistic view of the theoretical and empirical underpinnings of dividend policy decisions.

2.1 Existing Theories about Dividend Policy

The **Dividend Irrelevance Theory**, developed by Modigliani and Miller (1961), hypothesizes that a firm's market value is based on its earning power and the risk of its underlying assets, not on how these earnings are distributed between retained earnings and dividends, in a perfect capital market free from taxes, transaction costs, and information asymmetry. According to this theory, the value of a company is determined solely by its earning power and investment decisions, not by how profits are distributed between dividends and retained earnings. Investors can create their own "homemade dividends" by selling shares if they desire cash, making the firm's dividend policy immaterial to shareholder wealth. This view assumes that any dividend payout will result in a corresponding drop in share price, leaving investors indifferent between dividends and capital gains

The **Bird-in-the-Hand Theory**, by Lintner (1965) and Gordon (1959), asserts that dividends are relevant to a firm's value because shareholders prefer the confidence of current dividends over the doubt of future capital gains. According to this theory, since current dividends are seen as less risky than possible future gains, investors place a higher value on companies that pay dividends. As a result, firms with higher dividend payouts may experience higher market valuations, since investors are willing to pay a premium for the perceived reduction in risk

Signaling Theory addresses the information asymmetry between company management and outside investors (Miller & Rock, 1985). It argues that dividend announcements serve as signals to the market about a firm's future prospects. When a company increases its dividend, it signals management's confidence in the firm's future earnings and financial health, often leading to a positive market reaction. Conversely, a reduction or omission of dividends may be interpreted as a sign of potential trouble. Thus, dividends are used as a communication tool to reduce information gaps and influence investor perceptions(Lotfi, T., 2019)

The Agency problem arises when there is information asymmetry and conflict of interest between managers (agents) and stakeholders (principals). The Agency Problem is the main focus of the Agency Cost Theory. It makes the assumption that managers might not always act in the best interests of shareholders and could invest in low-return projects or use extra money for personal gain. By paying dividends, businesses restrict managers' access to free cash flow, which limits their ability to perform such duties. Dividend payments thus serve as a mechanism to align management's actions with shareholder interests and reduce agency costs, potentially increasing firm value. and hence, supporting Dividend Payout (Jensen & Meckling, 1976)

Tax Preference Theory, introduced by Litzenberger and Ramaswamy (1979), implies that investors may favor capital gains over dividends because they have a more favorable tax treatment in many jurisdictions. Investors may favor businesses that retain earnings and offer returns in the form of capital appreciation because dividends are frequently taxed at a higher rate and may be subject to double taxation. As investors look to optimize after-tax returns, this preference may result in a higher valuation for companies with lower dividend payout ratios.

According to Graham and Dodd (1934), **Traditional Theory**, shareholders value current dividends more than they do uncertain future capital gains. This perspective holds that since the stock market strongly supports liberal dividend policies, businesses that pay out large dividends are probably worth more on the market. According to this theory, dividend policies are important in determining a company's market value, particularly in situations where investors are risk-averse and seek quick returns(Lotfi, T, 2019).

Walter's Model is a **dividend relevance theory** that examines the connection between a company's cost of capital and return on investment (ROI) in order to determine how valuable a company's dividend policy is. According to the model, a company should hold onto its earnings for reinvestment rather than distributing them as dividends if its return on investment (ROI) is higher than its cost of capital. On the other hand, the company ought to pay dividends if the return on investment is less than the cost of capital. Any payout ratio is ideal for businesses where return on investment (ROI) is equal to the cost of capital. Therefore, the model offers a framework for choosing the best dividend policy depending on the profitability and investment opportunities of the company. (Walter J. E., 1963)

Gordon's Model, also called the Dividend Growth Model, emphasizes the importance of dividend policy by asserting that a company's value is determined by the present value of its projected future dividends, which are assumed to grow at a constant rate. The model makes the case that higher dividend payouts result in higher share prices by taking into account both the growth rate of retained earnings and the investor's preference for current dividends. The idea that investors favor the certainty of dividends over the uncertainty of future capital gains is supported by Gordon's Model (Gordon, 1959), just like the Bird-in-the-Hand Theory.

According to the **Life Cycle Theory** of dividends, a company's dividend policy is directly related to where it is in the corporate life cycle. According to this theory, young businesses in their early stages of development usually keep their earnings and pay little to no dividends because they have a lot of investment opportunities and need a lot of money to expand. A companies accumulate more free cash flow and encounters fewer Profitable investment opportunities as it grows older and its growth prospects decrease. Businesses are more likely to give shareholders larger dividends at this stage of maturity. This theory is supported by empirical research, which demonstrates that while younger, high-growth companies prefer to reinvest profits, more stabilized, profitable companies with steady earnings typically pay dividends(Mueller, 1972).

The **Clientele Effect Theory** states that different groups of investors, or "clienteles," have varying preferences for dividend policies based on their personal financial needs, tax situations, and investment purpose. For instance, younger investors or those in higher tax brackets might favor companies that reinvest earnings for growth, preferring lower or no dividends, whereas retirees or income-focused investors might favor companies with high, consistent dividend payouts. Certain customers are attracted to a company by its dividend policy, and major adjustments to this policy may cause changes in the investor base, which could have an impact on the stock price of the company. The clientele effect features the importance for firms to maintain consistency in their dividend policies to retain their preferred investor base and avoid volatility in share prices resulting from abrupt changes in dividend distributions (Elton and Gruber 1970b).

Even though these theories offer thorough insights, they frequently contradict one another, and no single theory can conclusively say whether or not a company should pay dividends. In the end, the choice is based on the particular financial situation of the company, expansion prospects, shareholder preferences, and outside variables like tax laws and market conditions. Furthermore, it is clear that dividend policy is a complex topic impacted by both quantitative measurements and qualitative factors like agency conflicts and signaling effects. Businesses need to carefully consider these factors because a bad policy could not only fail to maximize shareholder value but also have unforeseen effects on investor relations and market perception.

Although these theories provide comprehensive information, they are contradictory to each other, and no theory can definitely state whether a company should or should not pay dividends. Ultimately, the decision depends on the specific financial condition of the company, growth opportunities, shareholder attitudes, and external factors such as tax policies and market conditions. In addition, it is evident that dividend policy is a multifaceted issue influenced by both quantitative measures and qualitative considerations such as agency conflicts and signaling effects. Companies must think very hard about these because a poor policy may not only not maximize shareholder value but also have unintended consequences for investor relations and market perception as well.

2.2 Impact of Key Features on Dividend Payout

Size:

Larger companies can raise money more easily compared to smaller ones, at a lower cost, and with fewer restrictions than small businesses. From this, it can be concluded that as a company grows, it does not have to rely solely on internal financing. As a result, large companies can afford to pay higher dividends to their shareholders. According to studies (Lloyd et al., 1985; Barclay et al., 1995; Reeding, 1997; Holder et al., 1998; Fama and French, 2001), the size of a company is positively correlated with payout levels and is one of the most important factors in determining a company’s dividend policy. Even during financial crises, large companies tend to maintain higher dividend payments. Due to dispersed ownership in larger firms, including banks, they may face higher monitoring costs. Dividends can help mitigate agency conflicts. Additionally, large companies generally have better access to capital markets, which enables them to maintain stable payouts.

Profitability:

One of the crucial factors that influences a business’s dividend decisions is its level of profitability. Dividends are paid based on the annual profits of the company; therefore, companies that are running at a loss generally do not pay dividends. The company’s earnings influence changes in dividend payments(Lintner, 1956). Jensen et al. (1992), Han et al. (1999), and Fama and French (2002) highlighted a positive relationship between profitability and dividend payouts. Similarly, Adaoglu (2000), Pandey (2001), and Aivazian et al. (2003) supported these findings. These studies are logical, as companies with profitable operations typically generate significant cash flows, which help sustain dividend payments. Myers (1984) and Myers and Majluf (1984) mentioned that when a firm wants to invest, it prefers internal financing first. If additional funds are needed, the firm then considers external financing, starting with debt, followed by equity, to minimize transaction costs and the effects of information asymmetry.

Growth and Investment Opportunities:

Organizations focused on growth opportunities typically reinvest their profits into new projects rather than allocating them for dividend payments. Distributing dividends limits the funds available for expansion efforts. Companies often rely on internal financing, which is more cost-effective and flexible than raising funds through debt or equity, particularly in markets influenced by taxes, agency issues, or issuance costs. According to the pecking order theory, firms prioritize using internal resources to fund their projects. When firms have numerous investment opportunities, they are more likely to retain earnings rather than distribute them as dividends. The decision to allocate profits for shareholder payouts or reinvest them for growth creates a complex interplay between shareholder returns and business expansion. Elston (1996) observed that in imperfect markets, dividends significantly influence corporate investment decisions. Rapidly growing companies generally maintain lower dividend payout ratios, as noted by Myers and Majluf (1984). However, D'Souza (1999) found a positive but statistically insignificant correlation between growth and dividend distribution.

Age:

Companies aged beyond the set point need less funding for big capital deals because their core operations mature to stability. Mature firms generally lack several expansion projects, thus accumulating surplus cash that becomes available for shareholder dividend distribution. According to the "maturity hypothesis" explained by Grullon et al. (2002), established firms during their low-growth period maintain stable operations while needing less money for investments, so they tend to distribute excess profits. Older enterprises tend to pay dividends since they demonstrate greater willingness, together with the capability to do so. Economic research conducted by Barclay et al. (1995), Grullon et al. (2002), and Deshmukh (2003 presents evidence that business lifespan connects in positively correlated with dividend distribution practices. Younger enterprises retain their profits mainly to finance upcoming growth initiatives and future financial needs because they currently exist in their expansion phases. Researchers use firm age as a substitute variable that indicates company growth prospects. Older companies tend more frequently to have liberal dividend policies compared to other firms with equal conditions, as long as other variables remain consistent. It's accepted that this connection may present exceptions because different mature firms show different payment choices, yet this connection continues to stand.

Cash:

Companies with large cash reserves often benefit from paying dividends in several ways. By distributing cash to shareholders, the funds available for management are reduced, encouraging companies to seek external financing more often. This process increases transparency since capital markets require companies to disclose more information. Dividends also make it easier to monitor management decisions, reducing the risk of funds being misused. Studies by Easterbrook (1984) and Moh'd et al. (1995) highlight that monitoring payments can lower agency costs. Holding on to excess cash can lead managers to invest in projects that reduce company value, as Jensen (1986) pointed out. Paying dividends helps solve this issue by limiting the control managers have over surplus funds, discouraging poor financial decisions. This approach fosters trust between managers and shareholders. Jensen also suggested that debt financing is an effective way to manage surplus cash, especially in bank-dependent financial systems. For example, Jordanian companies often rely on short-term bank loans, which require frequent fundraising. This creates pressure on managers to maintain transparency and ensures regular monitoring to prevent conflicts between managers and shareholders.

Ownership (percentage held by insiders) :

Managers as well as directors, as well as key executives' roles become directly impacted by firm performance as their ownership stake increases significantly. High executive ownership creates goal compatibility between senior leaders and shareholders, thus minimizing managerial conflicts. The authors Jensen and Meckling (1976) established that rising insider ownership reduces the necessity for dividends because it improves managerial oversight without any need to reassure shareholders. Insiders maintain company value as their priority because they share ownership with the business. Higher levels of insider ownership reduce the need for dividend payments because insider ownership works as an alternative method to control agency costs.

Leverage:

Financial leverage refers to the extent to which a firm relies on debt financing. Using debt enhances equity returns but introduces financial risks, notably from fixed fiscal obligations, which include interest payments. A business that fails to fulfill its debt obligations will face potential financial collapse, up to the point of being dissolved. Organizations that handle substantial debt prefer to conserve operational funds to fulfill debt requirements instead of sharing income with stockholders. Organizations with higher financial leverage need to keep cash reserves to fulfill debts, which reduces the resources available for dividend payments. When firms have high leverage, they normally minimize dividend distributions because it reduces their costs of external financing according to Rozeff (1982). In debt agreements, financial institutions frequently add agreements that forbid distributions to shareholders. A negative association exists between the utilization of leverage and corporate dividend distribution activities. Multiple studies, such as Jensen et al. (1992) and Agrawal and Jayaraman (1994), and Gugler and Yurtoglu (2003) demonstrate that debt and dividends exist in an inverse relationship.

Multiple assessment methods of leverage may produce different correlation outcomes. The research by Hariem Abdullah (2022) about Turkish financial companies demonstrates how the debt-to-equity ratio produces a positive statistical effect on dividend payout, besides its negative impact from total debt ratio measurements. However, the results were statistically insignificant.

Market to book value:
D'Souza (1999), shows a negative but insignificant relationship in the case of market-to-book value. The market-to-book value ratio impacts dividend policy by signaling a firm's growth prospects and financial health. High ratios generally lead to lower dividends as firms reinvest earnings, while low ratios often result in higher dividends to reward shareholders and maintain confidence.

Free cash flow:
The decision about dividend policy depends heavily on free cash flow according to agency theory. Firms with all profitable investment opportunities utilized face agency problems because of their excess free cash flow between managers and shareholders. Managers frequently invest excess funds into unprofitable and negative NPV projects, because of which shareholders become dissatisfied and capital allocation becomes inefficient (Jensen, 1986). Dividend payments function well as a solution to limit this issue because they take money out of the manager's discretion. Managed cash distribution to shareholders reduces the tendency of excessive control by managers and helps financial choices better serve shareholder needs. These payments facilitate agency conflict management by reducing decision-making freedom for managers.

The agency costs between principals and managers can be effectively reduced through debt financing, according to Jensen (1986). Debt incurs steadfast financial responsibilities on managers that require them to fulfill interest and principal payments while enforcing financial restraint. The financial environment of Jordan functions mainly through banks because this debt-based oversight functions effectively. The extensive use of short-term debt by Jordanian firms leads to bank scrutiny that causes regular monitoring of their operations. Frequent monitoring activities performed by financial institutions limit managerial freedom while potentially reducing the firm's requirement for dividend distributions as an agency control method.

Ownership:

The costs related to agency problems decrease when insider personnel (managers, directors, and executive officers) boost their ownership levels in the firm, thus creating shareholder-alignment (Jensen and Meckling 1976). Firms that are partially controlled by their insiders require less dividend distributions because these stakeholders share ownership risks with shareholders. The presence of insider ownership (INSD) is expected to show negative effects on dividend payments (Rozeff, 1982; Jensen et al., 1992; Holder et al., 1998).

P/E Ratio:

The price-to-earnings ratio (PER) serves as a standard valuation measure to demonstrate possible future company growth according to Ang and Peterson (1984) and Glen et al. (1995). The market determines a company's worth through its evaluation of upcoming profit projections. Organizations with elevated PER status exhibit promising expansion potential, which drives them to hold back their earnings instead of distributing dividends. The preference of these companies to invest profit growth for business expansion creates a negative relationship between PER and dividend payout. Higher PER values lead to decreased possibilities of dividend payout increases.

Risk:

According to Pruitt and Gitman (1991), a company's dividend policy can also be influenced by the risk it faces, especially the unpredictability of its yearly earnings. Companies with consistent earnings are generally in a better position to estimate future profits and, as a result, are more comfortable distributing a larger share of those profits as dividends. In contrast, firms with unstable earnings tend to be more conservative with payouts. Additionally, research by Rozeff (1982), Lloyd et al. (1985), and Collins et al. (1996) used the beta value as a measure of market risk and reported a significant negative link between beta and dividend payout. This means companies with higher market risk (higher beta) are less likely to pay substantial dividends. D'Souza (1999) supported this view with similar findings showing a negative and significant correlation between beta and dividend payments.

Earnings:

Given these considerations, this study proposes that earnings positively impact the dividend payouts of Australian companies. In line with the work of DeAngelo et al. (1992), a loss dummy variable is also introduced to assess how current losses affect dividend distribution. When a company reports a loss, it is generally expected to reduce dividend payouts, as this serves as a signal to investors that the firm's long-term earnings potential may have weakened (DeAngelo et al., 1992).

Privately held companies:

Private companies choose dividends infrequently, while their payouts mostly remain at lower amounts compared to public corporations. The distribution of ownership capital in private enterprises is highly concentrated thus minimizing agency concerns that would otherwise prompt managers to use dividends as signals of quality performance or control mechanisms. tờ La Porta et al., 2000; Michaely & Roberts, 2007) Public firms face higher pressure from separate ownership and management positions to distribute dividends to reassure shareholders of their financial stability. Private business owners can minimize dividend taxation by using tax-friendly methods to reward themselves, such as deducting payments from their income, which reduces dividends as an attractive distribution method (Enis & Ke, 2003).

Standalone company and group company:

Companies inside groups show higher preference for dividend distributions than independent entities do. The tax-free dividend transfer rules in Belgium enable effective surplus funding moves between affiliated entities resulting in this preferred method of moving funds between group companies. Group companies can utilize their internal financial systems because they allow sufficient movement of capital between subsidiaries so they can alleviate funding issues while supporting weaker units (Deloof, 2001). Within a group the nature of owner-ship arrangements between entities plays an essential role. Companies with minority shareholders distribute higher dividends to their shareholders as compared to fully owned subsidiaries. Rolling minority shareholders receive less protection because they struggle to exchange their shares in the public market thus increasing the chance of exploitation from controlling owners. To protect themselves against exploitation controlling owners usually want higher dividend distributions according to Faccio et al. (2001).

Block holder:

Blockholder ownership affects dividend payments in two possible ways. Shareholders with large ownership assets often eliminate the need for generous dividend distributions since their substantial shareholding makes management respect shareholder interests—this phenomenon is known as the substitution effect. Firms that have substantial blockholders tend to provide reduced dividend payments under this perspective. When blockholders possess substantial ownership shares smaller shareholders could ask for higher dividend payments to ensure their confidence and cultivate trust within the company structure. The outcome model proposed by La Porta et al. (2000a) supports the hypothesis that shareholders intensify pressure on companies to distribute profits. Blockholder ownership creates pressure on firms but this pressure intensifies until a specified threshold is reached. Past a specific threshold dominant shareholders tend to stay entrenched inside the firm since they choose to leave profits within the business rather than pay dividends for their personal gain. As the ownership percentage rises to extremely high numbers blockholders eventually start paying dividends because they personally experience consequences from corporate financial misuse (Morck et al., 1988).

Board:

Through effective oversight of management, the board directors, along with their sub-committees, reduce agency-related issues by controlling company performance. Features of the board membership directly influence the dividend payout ratio (DPR) scale. Organizations using staggered boards have proven effective in generating higher dividend payout ratios according to Jiraporn and Chintrakarn (2009). Borokhovich et al. (2005) discovered in their study of 177 Nigerian firms that the external director proportion generated a negative impact on dividend distributions. According to Al-Najjar and Hussainey (2009), outside directors in 400 non-financial companies were linked to reduced dividend payments.

Board Gender Diversity:

Through effective oversight of management, the board directors, along with their sub-committees, reduce agency-related issues by controlling company performance. Features of the board membership directly influence the dividend payout ratio (DPR) scale. Organizations using staggered boards have proven effective in generating higher dividend payout ratios according to Jiraporn and Chintrakarn (2009). Borokhovich et al. (2005) discovered in their study of 177 Nigerian firms that the external director proportion generated a negative impact on dividend distributions. According to Al-Najjar and Hussainey (2009), outside directors in 400 non-financial companies were linked to reduced dividend payments.

Board Meeting Frequency:

Board organizations demonstrate diverse viewpoints regarding their meeting intervals to maintain corporate governance. The argument exists that regular board meetings improve operational transparency and diminish agency conflicts which produces more effective administration (Allegrini & Greco, 2011). The frequency of board sessions grows independence within the board and enhances directors' ability to evaluate managerial performance (Conger, Finegold, & Lawler, 1998). Regular board oversight activities correlate to superior business results that yield increased dividend payments according to Ntim and Osei (2011).

CEO, Board Chair's Duality:

Studies about the connection between the dual roles of CEOs as board chairs and corporate dividend policies yield conflicting results. According to Abor and Fiador (2013), researchers provide diverse findings regarding this subject. CEO dual leadership proved to negatively impact dividend payments in the analyzed Chinese companies, according to Zhang (2008). Research conducted in Iran along with the U.S. and Malaysia did not establish any meaningful connection between these variables.

Board Size:

A board of appropriate size will enable successful business operations according to FRC (2012). The research by Ntim and Osei (2011) concludes that bigger boards enhance management monitoring which reduces agency problems and improves dividend levels. According to the substitution theory the addition of directors beyond optimal quantity results in defective coordination and communication which weakens oversight (Lipton & Lorsch, 1992). Research findings about the link between board size and dividends display conflicting outcomes because Mansourinia et al. (2013) and Kiel and Nicholson (2003) discovered a positive association between these variables.

Audit Committee Size:

According to Razaee (2008) independent audit committees need to be formed by companies to make financial reports presented to shareholders more reliable. The experience and expertise of multiple committee members allow larger audit committees to provide better oversight according to Kyereboah-Coleman and Biekpe (2006). According to Chen (2010) dividends serve as an anti-agency tool in organizations showing poor governance structures.

Gross Domestic Product (GDP):

The expansion of GDP shows a typical positive impact on business distributions to stockholders. The rise in Indonesia's property and real estate sector economic activity delivered better firm performance through higher return on assets (ROA) values which triggered businesses to raise their dividend payments. The better consumer demand along with improved corporate profitability enabled companies to make larger earnings distributions to shareholders (Romus et al., 2020). Enhanced corporate earnings in Ghana operated through GDP growth to enable businesses to give larger dividend distributions to shareholders (Mahama, 2023).

Statistical significance between GDP growth and dividend payments does not exist constantly. Although GDP growth influences dividend decisions in certain circumstances, its effects do not necessarily extend equally to all business sectors throughout the world.

Interest Rates:

The payment of dividends tends to decrease when interest rates rise. Higher interest rates increase borrowing costs and diminish available company profits so shareholders receive less dividend payments. The rate hikes instituted by Bank Indonesia caused detrimental effects to company operational performance, which resulted in diminished dividend payouts according to Romus et al. (2020). Bangladeshi businesses increased their dividend distribution while interest rates remained at lower levels because lower borrowing expenses benefited them. The debt-to-equity ratio provided the basis for Turkish researchers to establish that firms paid higher dividends at high interest rates perhaps because of financial or tax-related advantages (Abdullah, 2022).

Inflation:

Inflation affects dividend payments differently depending on the situation. In Ghana, companies were able to increase their dividends even during inflation because they raised their prices to keep profits steady despite higher costs (Mahama, 2023). However, in Kenya, high inflation reduced the actual value of company earnings, which led to lower dividend payouts (Mundati, 2013). These differences show how important it is for firms to manage inflation well and how much a company's industry flexibility can influence its dividend policy.

Market Capitalization:

Market capitalization shows an association which typically works against dividend payments. The payout ratios in Ghana's market remained low because large corporations preferred to invest their earnings into business growth rather than distribute dividends (Mahama, 2023). Under the lifecycle theory companies grow more mature and tend to keep earnings for future development instead of sharing them with their stockholders.

Exchange Rates:

The way in which exchange rates fluctuate produces dissimilar impacts on companies' payment of dividends. The alterations of exchange rates in Ghana demonstrated no relevant or statistical correlation with corporate dividend payments (Mahama, 2023). Real estate companies in Indonesia experienced reduced profitability when their industry costs rose after the rupiah depreciated because import materials became more expensive (Romus et al., 2020).

Industry:

Each sector of business maintains its own dividend payment practice due to specific requirements like stable cash flow and investment requirements, and future growth potential. Utilities, together with consumer staples sectors, maintain steady income from mature markets thus tend to distribute higher dividends to shareholders through regular payments. Indian utility firms reduced their dividend payouts due to their high capital investment requirements, yet manufacturers who had excess cash gave larger distributions (Labhane & Das, 2015). Property and real estate organizations in Indonesia distribute moderate dividend payments through influences from economic conditions and market demand patterns (Romus et al., 2020). The way businesses distribute dividends depends on specific policies and tax regulations, which vary between different industries. The financial institutions of Turkey leveraged high individual dividends through tax shield benefits but the manufacturing organizations preferred debt repayment to dividend distribution according to Abdullah (2022). The Belgian business groups employed tax-free intercompany dividends to optimize capital distribution between their group entities according to Michel (1979).

2.3 Related Machine Learning Studies on Dividend Payout Policies

Kumar and Sinha (2024) analyzed dividend payout determinants using data from 3,162 non-financial firms listed on the Bombay Stock Exchange (BSE) over the period 2006–2022. The study incorporated firm-level characteristics, macroeconomic indicators, and the effects of major crises, including the 2007–2008 financial crisis and the 2019–2020 COVID-19 pandemic. They used Various machine learning techniques, including Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Logistic Regression, Random Forest, Linear Discriminant Analysis (LDA), and Decision Trees. The highest accuracy of 90.77% was achieved by Random Forest Classifier,

Won et al. (2012) analyzed dividend policy prediction by utilizing companies' data listed in the Korea Exchange (KRX) market for the years between 1980 and 2000. A total of 137 firms with more than 15 years of dividend history were targeted to avoid any incompleteness or inconsistency in the data. Current dividend, historical and current stock price were some key predictor variables used. They used Classification and Regression Trees (CART), Chi-square Automatic Interaction Detector (CHAID), Quick, Unbiased, Efficient Statistical Tree (QUEST), and Genetic Algorithm Knowledge Refinement (GAKR) in their study. They obtained the best mean accuracy of 76.16% based on five-fold cross-validation.

Longinidis and Symeonidis (2013) examined information for 246 companies traded on the Athens Exchange (ATHEX) from 2007 to 2009. Corporate governance matrices and financial ratios were employed as predictors. The techniques used were Logistic Regression, Neural Networks, and Decision Trees. They obtained the best accuracy of 93% on the whole dataset using Decision Trees, which was better than that of Neural Networks (82.52%). Net profit after tax emerged as the most powerful predictor in their work.

McMillan (2014) examined dividend growth forecasting with panel data regression over 15 nations from 1973 to 2023. The research emphasized the importance and influence of different microeconomic variables on dividend growth forecasting.

Ivașcu (2023) studied 2,059 listed firms from 56 nations. The techniques used were Logistic Regression, CART, Random Forest, XGBoost, and Random Forest with cost-sensitive learning. The research attained an accuracy of 82.7% with a recall of 55.4% and 94.1% for non-payers and payers, respectively. The results identified firm size and beta as the most significant determinants.

Vodwal and Negi (2023) examined the determinants of dividend decisions by analyzing data from 919 listed Indian non-financial firms over the period 1999 to 2019. Using Lasso regression, the study identified key predictors of dividend payouts, achieving an accuracy of 75.87%. They consider features included firm-specific factors such as profitability, size, free cash flows, liquidity, Tobin's Q, Altman's Z-score, and minority interest, along with Macroeconomic features.

Konak et al. (2024) investigated the factors influencing dividend decisions by analyzing data from 37 companies listed on the Borsa Istanbul Stock Exchange (BIST) over the period 2011 to 2021. The study considered 26 features, encompassing both company-specific and macroeconomic determinants, to identify internal and external factors affecting dividend rates. They employed a hybrid Genetic Algorithm and Artificial Neural Network (GA-ANN) model, achieving an R-squared value of 0.88 and the lowest Mean Squared Error of 0.075.

X. Wang et al. (n.d.) explored dividend forecasting using machine learning models by leveraging data from all US-listed firms between January 2002 and December 2022. They employed three nonlinear and non-parametric tree-based models: Random Forest (RF), Gradient Boosted Trees (GB), and Extreme Gradient Boosting Trees (XGB). Monthly analyst forecasts of Dividend Per Share (DPS) were obtained from the I/B/E/S summary file, alongside actual DPS values. Additionally, they included quarterly firm fundamentals, monthly stock prices and, and macroeconomic indicators. The study compared the forecast accuracy of machine learning models and analyst predictions for DPS across various forecasting horizons, including 1 month, 3 months, 6 months, 12 months, and 2 years. Performance was evaluated using Mean Squared Error (MSE), demonstrating the effectiveness of machine learning approaches in dividend prediction.

Elyasiani et al. (2019) conducted a study on dividend payout prediction using data from 314 companies listed on the Tehran Stock Exchange (TSE), encompassing 1,725 firm-year observations during the period 2009–2016. The analysis incorporated predictors at the country, industry, and firm levels to assess their influence on dividend decisions. Using logistic regression, the study achieved a maximum R-squared value of 0.54 for predicting Dividend Per Share (DPS) and an accuracy of 89% for Dividend Propensity.

Yaseen and Dragotă (2021) investigated the factors influencing dividend payout ratios by analyzing a dataset of 11,248 companies from 70 countries for the period 2008–2014. The study encompassed firm-level, macroeconomic, and sociocultural variables to identify key determinants of dividend decisions. Various predictive models were developed, achieving a highest accuracy of 81.3% using Decision Tree classifier.

Bhat (2022) examined dividend omission and its determinants using financial data comprising 12,942 firm-year observations from 2013 to 2018. The study revealed that 55% of firms omitted dividends during this period. They used dimensions such as size, growth, efficiency, profitability, liquidity, and financial; employed, including Logistic Regression, Naïve-Bayes, Decision Tree Ensembles (Decision Tree, Random Forest, Gradient Boosting Trees), Support Vector Machines, and Artificial Neural Networks (Probabilistic Neural Network with RNSprop). Among these, the Artificial Neural Network achieved the highest accuracy of 82.36%, offering a robust framework for understanding the dynamics of dividend omission.

In comparison to the existing literature, our study uses a much larger dataset than previous studies and excludes market-related variables to generalize the results for unlisted firms as well. Furthermore, our accuracy is significantly higher compared to these studies. Additionally, we predicted special dividend payouts and equity dividend growth.

3 Data

For our study, we have collected firm-level, industry-level, and country-level data from various online sources. The majority of data is collected from Prowess Dx, which has a collection of data of a large number of Indian firms. From Prowess Dx we have collected firm-level data like standalone financial data, Equity Ownership Pattern, Board Meeting data, Board of Directors, BSE & NSE Stock Trading Data. We have collected some country-level variables from the World Bank and the IMF.

Financial data of firm contains features like revenue, R&D expense, net property plant and equipment, Gross property plant and equipment, depreciation, interest expense, COGS, Tax expense, non-current liability, current liability, PAT, EBITDA, total capital, paid-up capital, long term and short-term borrowings, total dividend, equity dividend, interim dividend, final plus special dividend. Among the dividend-related variables are our primary target variables.

Equity ownership data contains the percentage of promoters and institutional investors. Board meeting data includes the date, purpose, announcement date, etc., from which we have created a useful variable: the number of board meetings in a year from the board meeting date. The board of directors' data contains the directors' names, the committee name, executive directors, independent directors, and the number of other owners. From that, we extracted data on the proportion of independent and executive directors and the total other ownership for a company in a particular year. Stock trading data contains information on market capital, P/E ratio (profit by earnings), EPS (earnings per share), etc. From this data, we have taken the mean of market capital, P/E, and EPS for a particular company and year. Additionally, we have extracted data that indicates whether the company is listed on the BSE or not.

In the country level variables we have collected data from the World Bank and the IMF of the Indian economy. The data contains exchange rate, Final consumption expenditure as percentage of GDP, Gross Capital formation as percentage of GDP, Domestic credit to private sector (% of GDP), Market capitalization of listed domestic companies (% of GDP), Government expenditure percent of GDP (% of GDP), Real GDP growth (Annual percent change), GDP per capita, Inflation rate, end of period consumer prices (Annual percent change), Unemployment, total (% of total labour force).

The data contains 55708 unique companies, out of which only 9460 companies are listed on the BSE. And out of that, 5534 companies are listed currently. So, the stock trading data, market capitalization, P/E, and EPS, the equity ownership data, and the board meeting data have very limited information, so we did not consider the data of Stock, board meetings, and equity ownership.

As the data was collected from different sources and different data frames, I have merged them based on year and company code on the firm’s financial data. Initially data has 6,84,806 data instances, which contain 54,871 different Indian companies. We have processed the raw data thoroughly and removed all kinds of ambiguities from the data, filtered it. The final merged data, which contains all the discussed variables than contains 4,21,300 rows in it. Still, the data has many null values in it.

We analyzed panel data containing 54,871 firms, with data availability ranging from a maximum of 37 years to a minimum of 1 year. Since firms with only 1 year of data are not useful for our study, we excluded those firms from the dataset. The bar chart in Figure 1 visually represents the availability of firms' data. The height of each bar indicates the frequency of firms, while the X-axis represents the total number of years for which data is available. This visualization highlights the distribution of data availability across firms.

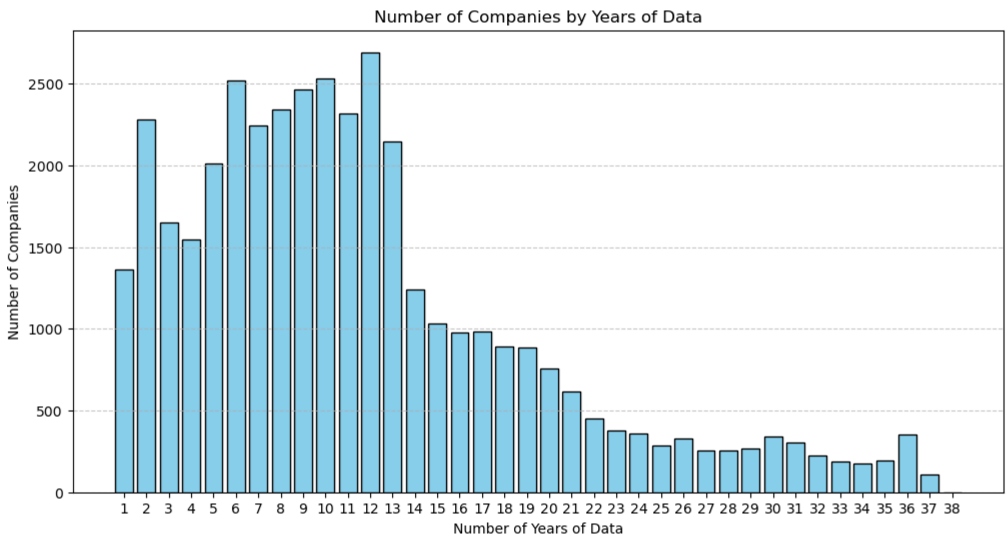


Figure 1: Cumulative frequency of firm data availability

The financial dataset has a raw feature that was not directly usable for analysis, so we created several derived features to enhance their utility. These included Leverage, Profitability, Asset and Investment, Liquidity, Dividend, and Expense-related Variables, which are mentioned in Table 1 below. These derived features allowed for a more structured and comprehensive analysis tailored to the study's objectives.

Table 1: Aggregated Variables and Formula

Sr. No.	Variable Name		Formula
	Leverage		
1	TOTALDEBT	=	ST_BORR + LT_BORR
2	LEV1	=	TOTALDEBT / TA
3	LEV2	=	NCL / TA
4	STLEV	=	ST_BORR / TA
5	LTLEV	=	LT_BORR / TA
6	MAT	=	LT_BORR / TOTALDEBT
7	DELTATOTALDEBTRAT	=	(TOTALDEBT - LAG_TOTALDEBT) / TA
8	DEBTISSUANCE	=	1 if DELTATOTALDEBTRAT > 0.05, else 0
9	DELTAPAIDUPCAP	=	PAIDUP_CAP - LAG_PAIDUPCAP
10	MEDLEV1_IND	=	MEDIAN(LEV1) by NIC2DIG and FYEAR
11	MEDLEV2_IND	=	MEDIAN(LEV2) by NIC2DIG and FYEAR
12	OVERLEV1	=	1 if LEV1 > MEDLEV1_IND, else 0
13	OVERLEV2	=	1 if LEV2 > MEDLEV2_IND, else 0
	Profitability		
14	ROA	=	PAT / TA
15	ROE	=	PAT / TOT_CAP
16	EBITDAMARGIN	=	EBITDA / REV
17	EBITMARGIN	=	(EBITDA - DEP) / REV
18	PATMARGIN	=	PAT / REV
19	EBITBYTA	=	(EBITDA - DEP) / TA
20	ROCE	=	(EBITDA - DEP) / (TA – CL)
21	PATGROWTH	=	(PAT - LAG_PAT) / LAG_PAT
22	REVGROWTH	=	(REV - LAG_REV) / LAG_REV
23	DUMMY_REVGROWTH	=	1 if REVGROWTH > 0, else 0
24	DUMMY_PATGROWTH	=	1 if PATGROWTH > 0, else 0
25	EARNINGVOL	=	Rolling standard deviation of ROA by CODE
26	REVVOL	=	Rolling standard deviation of REV by CODE
	Asset and Investment:		
27	SIZE	=	LOG(TA)
28	TANG	=	NPPE / TA
29	DEPRATIO	=	DEP / TA
30	ASSETMAT	=	(NPPE / DEP) * (NPPE / TA) + (CA / COGS) * (CA / TA)
31	CAPEX	=	(GPPE - LAG_GPPE) / TA
32	DELTACAPEX	=	(CAPEX - LAG_CAPEX) / TA
33	LOG_CAPEX	=	LOG(CAPEX), if CAPEX > 0, else NaN
34	MEDTANG_IND	=	MEDIAN(TANG) by NIC2DIG and YEAR
35	MEDCAPEX_IND	=	MEDIAN(CAPEX) by NIC2DIG and YEAR
36	OVERTANG	=	1 if TANG > MEDTANG_IND, else 0
37	OVERCAPEX	=	1 if CAPEX > MEDCAPEX_IND, else 0
38	LOGASSGR	=	SIZE - LAG_SIZE
39	MEDLOGASSGR_IND	=	MEDIAN(LOGASSGR) BY NIC2DIG AND FYEAR
	Liquidity:		
40	CASHRAT	=	TOTALCASH / TA
41	TOTALCASHEQURAT	=	(TOTALCASH + ST_INV)/TA

42	CASHFLOWRAT	=	(PAT + DEP) / TA
43	CR	=	CA / CL
44	NWCBYTA	=	(CA - CL) / TA
45	CFVOL_IND	=	Standard deviation of CASHFLOWRAT by NIC2DIG and YEAR
46	CASHFLOWVOL	=	Rolling standard deviation of CASHFLOWRAT by CODE
47	MEDCASHFLOWRAT_IND	=	MEDIAN(CASHFLOWRAT) by NIC2DIG and YEAR
48	MEDTOTAL-CASHEQURAT_IND	=	MEDIAN(TOTALCASHEQRAT) by NIC2DIG and YEAR
49	OVERCASHFLOWRAT	=	1 if CASHFLOWRAT > MEDCASH-FLOWRAT_IND, else 0
50	OVERTOTALCASHEQURAT	=	1 if TOTALCASHEQRAT > MEDTOTAL-CASHEQRAT_IND, else 0
51	DELTACASHEQURAT	=	(CASHEQU - LAG_CASHEQU) / TA
52	FCF	=	PAT + DEP + (CA - LAG_CA) - (CL - LAG_CL) - (GPPE - LAG_GPPE)
	Dividend		
53	EQUITY_DIVDUMMY	=	1 if EQUITY_DIV ≠ 0, else 0
54	SPE_DIVDUMMY	=	1 if SPE_DIV ≠ 0, else 0
55	EQUITY_PAYOUTRAT	=	EQUITY_DIV / PAT
56	EQUITY_DIVBYTA	=	EQUITY_DIV / TA
57	EQUITY_DIVBYOP	=	EQUITY_DIV / (EBITDA - INT)
58	SPE_PAYOUTRAT	=	SPE_DIV / PAT
59	SPE_DIVBYTA	=	SPE_DIV / TA
60	SPE_DIVBYOP	=	SPE_DIV / (EBITDA - INT)
61	EQUI-TYDIVGROWTH_DUMMY	=	1 if EQUITY_DIV > LAG_EQUITYDIV, else 0
62	DIVIN5YEAR	=	Sum of EQUITY_DIVDUMMY within a 5-year rolling window
	Expense:		
63	TAXRATE		TAX / (PAT + TAX)
64	SGABYTA	=	(REV - COGS - EBITDA) / TA
65	RNDBYREV	=	RND / REV
66	RNDBYTA	=	RND / TA
67	RND_DUMMY	=	1 if RND ≠ 0, else 0
68	LIFECYCLE	=	RNS / TOT_CAP
69	INTRATE	=	INT / TOTALDEBT

Note: LAG is used for taking the previous year's value of the particular feature.

3.1 Descriptive statistics

We carefully selected the features after performing exploratory data analysis. Missing values were imputed using KNN imputation, as discussed in the following section. The final dataset includes both numerical and binary features, for which we have provided descriptive statistics in Table 2. The table shows the descriptive statistics of the dataset in the form of measures of central tendency and variability. The mean gives the average value, while the median gives the middle point of the data distribution. Quartile 1 (Q1) and Quartile 3 (Q3) represent the 25th and 75th percentiles, giving the spread of the middle 50% of the data. The standard deviation also captures the variation in the dataset, indicating how much the data varies from the mean, and the range gives the difference between the minimum and maximum values.

Table 2: Summary statistics for Numerical features

Variable name	mean	std_dev	Q1	median	Q3	range
Leverage:						
LOG_STLEV	0.1900	0.1990	0.0714	0.1557	0.2534	6.5145
LOG_LTLEV	0.2086	0.2395	0.0602	0.1493	0.2968	10.7038
MAT	0.6152	0.3223	0.3642	0.6499	0.9183	35.6667
MEDLEV1_IND	0.3201	0.1107	0.2574	0.3128	0.3663	1.3071
DELTATOTALDEBTRAT	-0.0229	5.1851	-0.0282	0.0094	0.0726	3152.3077
LOG_DELTAPAIDUPCAP	0.0105	0.1003	0.0000	0.0000	0.0000	39.1791
Profitability:						
ROCE	0.1428	9.1353	0.0335	0.1163	0.2170	4740.2500
ROE	16.0883	882.642	-0.0098	0.1936	1.5291	448727.00
ROA	-0.0403	25.5459	-0.0022	0.0177	0.0599	15158.781
EBITBYTA	0.0723	3.6497	0.0171	0.0643	0.1145	1760.8670
EBITMARGIN	-0.1351	135.461	0.0173	0.0627	0.1396	75721.500
EBITDAMARGIN	0.1114	142.661	0.0330	0.0943	0.2012	90862.000
PATMARGIN	-2.0548	141.026	-0.0058	0.0169	0.0664	54321.125
PATGROWTH	1.8457	860.378	-0.7103	-0.0348	0.5526	469882.00
LOG_EARNINGVOL	0.0578	0.1531	0.0090	0.0226	0.0526	8.9549
REVGROWTH	15.7117	816.697	-0.0862	0.0995	0.3711	264139.00
LOG_REVVOL	4.4264	2.1756	2.9129	4.4919	5.9202	14.5350
Asset and Investment:						
SIZE	6.2812	2.2530	4.8621	6.3297	7.7465	18.3924
TANG	0.2537	0.2363	0.0590	0.1926	0.3859	13.7617
MEDTANG_IND	0.2003	0.1433	0.0726	0.1820	0.2904	0.8901
LOG_CAPEX	-4.1522	1.6380	-4.9967	-4.0147	-3.1039	20.4262
MEDCAPEX_IND	0.0107	0.0110	0.0016	0.0080	0.0164	0.5420
DELTACAPEX	-0.0478	4.6586	-0.0001	0.0000	0.0000	1796.1771
LOGASSGR	0.1183	0.4676	-0.0303	0.0643	0.2065	22.5419
MEDLOGASSGR_IND	0.0608	0.0404	0.0353	0.0592	0.0825	1.6902
LOG_DEPRATIO	0.0302	0.0386	0.0099	0.0222	0.0397	6.5993
ASSETMAT	22.3388	753.173	1.4445	3.2640	7.2036	346109.11
Liquidity:						
CASHFLOWRAT	0.0352	1.9630	0.0102	0.0446	0.0924	1391.9405
LOG_CASHFLOWVOL	0.0527	0.1137	0.0111	0.0249	0.0530	6.2415
MEDCASHFLOWRAT_IND	0.0481	0.0246	0.0303	0.0442	0.0615	0.3790
LOG_CFVOL_IND	0.3453	0.4083	0.1305	0.2180	0.3743	3.8388
TOTALCASHEQURAT	0.0965	0.1534	0.0089	0.0326	0.1111	2.0990

MEDTOTAL-CASHEQURAT_IND	0.0413	0.0336	0.0219	0.0357	0.0465	0.4817
NWCBYTA	0.0775	3.2040	0.0168	0.1618	0.3477	609.0000
CR	8.0556	141.231	1.0611	1.4737	2.5728	27500.000
Corporate Governance:						
AGE	22.0812	16.9581	11.0000	19.0000	28.0000	174.0000
LIFECYCLE	146.751	5249.46	0.0850	2.6948	14.2141	2014634.0
OWNER_PRIVATE	0.7611	0.4264	1.0000	1.0000	1.0000	1.0000
OWNER_GOV	0.0241	0.1533	0.0000	0.0000	0.0000	1.0000
Expense:						
RNDBYTA	0.0014	0.1955	0.0000	0.0000	0.0000	103.0002
SGABYTA	0.2256	2.7059	0.0227	0.0909	0.2177	1029.0000
INTRATE	0.4276	22.3713	0.0658	0.0962	0.1332	8355.9810
TAXRATE	0.3481	11.1231	0.0767	0.2922	0.4621	5803.5000
Dividend:						
DIVIN5YEAR	0.6636	1.4823	0.0000	0.0000	0.0000	5.0000
EQUITY_PAYOUTRAT	0.0640	1.1659	0.0000	0.0000	0.0000	401.0000
EQUITY_DIVBYTA	0.0050	0.0441	0.0000	0.0000	0.0000	7.1739
TOTAL_OTH_OWN	2.9702	10.0837	0.0000	0.0000	0.0000	204.0000
PROPO_INDEP	60.0693	40.7788	11.1111	66.6667	100.0000	100.0000
PROPO_EXECU	30.4109	36.6573	0.0000	0.0000	66.6667	100.0000
Macroeconomic:						
GDP	1819.9	354.007	1559.8640	1915.552	2050.16	1786.6760
GDP_GR	5.8458	3.9314	5.5000	6.8000	8.0000	15.5000
INF_CPI	5.8499	2.0522	4.6000	5.3000	6.7000	10.1000
REPORATE	6.1006	1.2426	5.0833	6.2500	7.3125	4.0000

Table 3 presents the binary features of the data in terms of 1s and 0s counts per feature. The mean is the fraction of 1s, indicating how frequently the positive class is present. The standard deviation provides a measure of how much the fraction of 1s can change across the dataset. The median is indicative of the majority class, which indicates whether 1s or 0s are more predominant in general.

Table 3: Summary Statistics for the Binary feature of the data

Variable Name	count_0	count_1	mean	median	mode	std_dev
Leverage:						
OVERLEV1	174562	108600	0.3835	0	0	0.4862
DEBTISSUANCE	227776	55386	0.1956	0	0	0.3967
Profitability:						
DUMMY_PATGROWTH	158392	124770	0.4406	0	0	0.4965
DUMMY_REVGROWTH	115814	167348	0.5910	1	1	0.4917
Asset and Investment:						
OVERCAPEX	158199	124963	0.4413	0	0	0.4965
OVERTANG	149020	134142	0.4737	0	0	0.4993
Liquidity/Cash:						
OVERCASHFLOWRAT	154256	128906	0.4552	0	0	0.4980
OVERTOTALCASHEQURAT	146648	136514	0.4821	0	0	0.4997
Corporate Governance:						
LISTED	232021	51141	0.1806	0	0	0.3847
OWNER_PRIVATE	67650	215512	0.7611	1	1	0.4264
OWNER_GOV	276346	6816	0.0241	0	0	0.1533
Expense:						
RND_DUMMY	260822	22340	0.0789	0	0	0.2696
Industry:						
INDUSTRY_46.0	223587	59575	0.2104	0	0	0.4076
INDUSTRY_77.0	270305	12857	0.0454	0	0	0.2082
INDUSTRY_20.0	270633	12529	0.0442	0	0	0.2056
Dividend:						
EQUITY_DIVDUMMY	244666	38496	0.1360	0	0	0.3427
SPE_DIVDUMMY	282311	851	0.0030	0	0	0.0547
EQUITYDIVGROWTH_DUMMY	262306	20856	0.0737	0	0	0.2612

4 Methodology

4.1 Data Processing

For our task to become accurate and dependable, data processing is essential. Removing unnecessary features that introduce noise or redundancy it helps make sure the model only includes relevant predictors. Transforming features enhances their predictive power and aligns them with the specific requirements of the model. Proper data processing subsequently enhances accuracy and model performance.

4.1.1 Exploratory Data Analysis and Feature Selection

For data analysis and feature selection, exploratory data analysis (EDA) proved to be the most useful technique. During this process, we examined box plots, correlation matrices, and VIF scores. Due to the large number of features, visualizing all of them individually was impractical. So, we performed EDA by grouping the variables into categories. We classified all variables into the following groups: Leverage, Profitability, Liquidity/Cash, Miscellaneous, Dividend, Industry Dummies, Country-Level Variables, and Other Variables.

VIF Score:

In regression analysis, the Variance Inflation Factor (VIF) is a statistic used to detect multicollinearity. It gauges how much the correlation with other model predictors inflates the variance of a regression coefficient. Severe multicollinearity is indicated by a high VIF value (typically >10), which may affect the model coefficients' validity.

$$VIF_i = \frac{1}{(1 - R_i^2)}$$

Where R_i^2Is R-squared value for the i^{th} Predictor

For each category, we began by plotting a boxplot to visualize the distribution, skewness, and outliers in the data. For numerical features, we applied Winsorization, a statistical technique used to replace outlier values with the nearest non-outlier values. From the boxplots, if a feature exhibited a skewed distribution, we considered transforming the variable to potentially improve results. For some numerical features, we applied log transformations, ensuring non-negativity before applying the log transform. For variables with negative entries, we scaled them by adding the minimum value plus 1 to ensure positivity. To avoid taking the log of 0, we used the log1p function, which adds 1 to the cell value before transforming it into a logarithmic scale. Finally, we analyzed the correlation matrix of the grouped features, including the log-transformed variables.

Features with a high correlation (greater than 90%) were identified, and one feature from each highly correlated pair was removed based on its correlation with the target variable. We assumed that a higher correlation with the target variable indicates better predictability, so features with lower correlation were excluded. As a result, only the log-transformed version or the original version of a feature was retained in the dataset.

In the subsequent step of feature removal, we calculated the VIF scores for all features and removed variables with high multicollinearity (VIF > 15). Multicollinear features were systematically eliminated one at a time, ensuring that features with strong self-correlation were not inadvertently excluded.

In the Industry Dummies category, we had 70 different binary variables. Because of the large number of variables, it was not feasible to analyze them using boxplots or a correlation matrix. Therefore, we built separate logistic regression models for each target variable. Based on feature importance and model accuracy, we found that only "INDUSTRY_46.0"(Wholesale trade, except of motor vehicles and motorcycles and motorcycles), "INDUSTRY_77.0" (Leasing of Nonfinancial Intangible Assets) contributed significantly to predicting and "F_EQUITY_DIVDUMMY" "F_EQUITYDIVGROWTH_DUMMY". For predicting "F_SPE_DIVDUMMY", the dummies "INDUSTRY_46.0" and "INDUSTRY_20.0" (Manufacture of Food Products) made the most significant contributions. Consequently, we selected these specific industry dummies for their respective target variables.

For the leverage variables analyzed during exploratory data analysis, we have prepared boxplot for each numerical variable to observe distribution of them and from that we have transformed some of the highly skewed features to log transform. we removed "LEV1", "LEV2", "STLEV", "LTLEV", "LOG_MAT", "LOG_MEDLEV1_IND", "LOG_MEDLEV2_IND" and "DELTAPAIDUPCAP" because their correlation with other features is high. Additionally, "LOG_LEV1" and "LOG_DELTATOTALDEBTRAT" were excluded due to their high VIF scores. "LOG_LEV2" "OVERLEV2" and "MEDLEV2_IND" were removed to ensure that only one measurement of leverage, either "LEV1" or "LEV2," was retained.

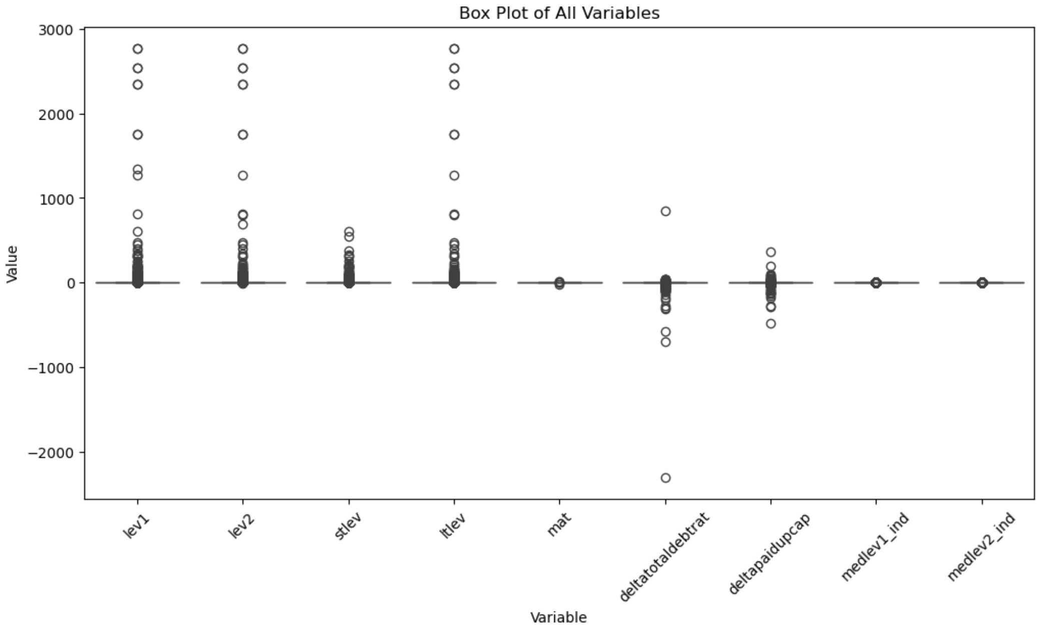


Figure 2: Box plots of leverage-related variables.

We have plotted box plots and correlation matrices for the profitability variable, also found some of the skewed features, and added a log transform of that feature "LOG_EARNINGVOL", "LOG_REVVOL". Removed highly correlated features.

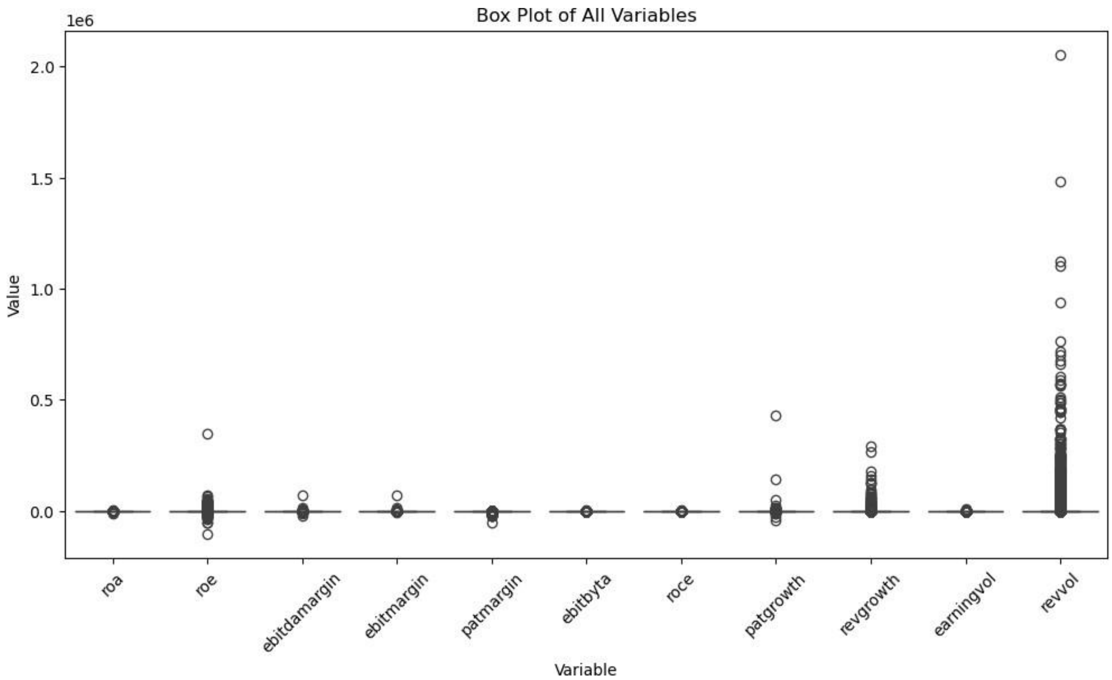


Figure 3: Box plots for profitability-related variables

For Asset and Investment variables, we plotted boxplots and introduced log transformations for the variables "TANG", "DEPRATIO" and "MEDTANG_IND" After reviewing the correlation matrix, we removed "LOG_TANG", "DEPRATIO", "LOG_MEDTANG_IND", and "CAPEX".

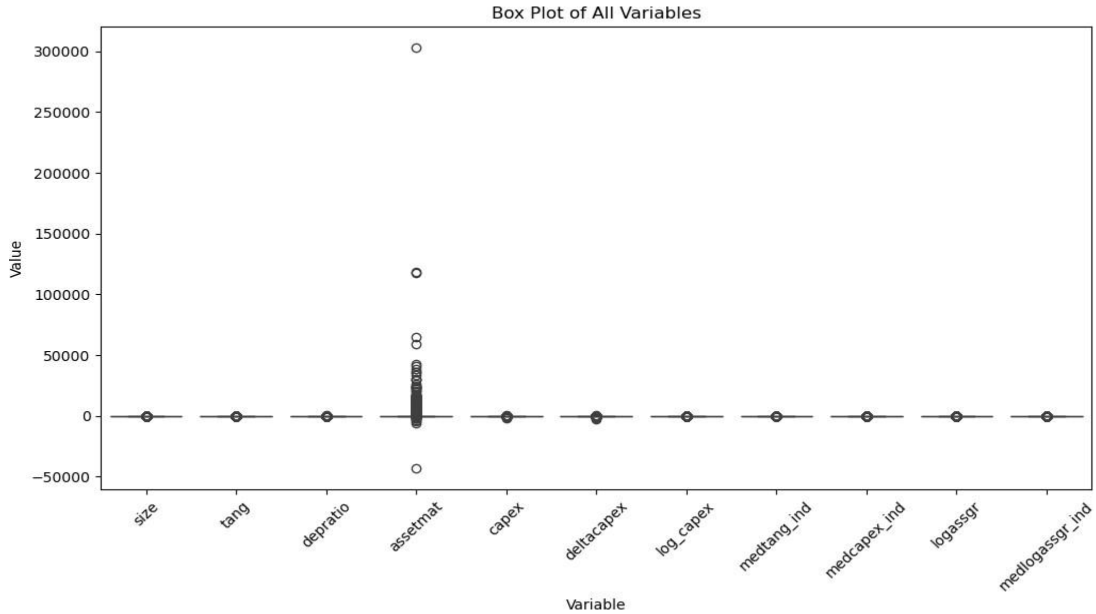


Figure 4: Box plots for Asset and Investment Related Variables

For Liquidity variables, we introduced log transformations for "CFVOL_IND", "CASHFLOWVOL", "MEDTOTALCASHEQURAT_IND", "NWCBYTA", "DELTACASHEQURAT" and "CR" After examining the correlation matrix, we removed "LOG_MEDTOTALCASHEQURAT_IND", "CFVOL_IND", "CASHFLOWVOL", "LOG_NWCBYTA", "LOG_DELTACASHEQURAT", "LOG_CR", "CASHRAT" and "DELTACASHEQURAT".

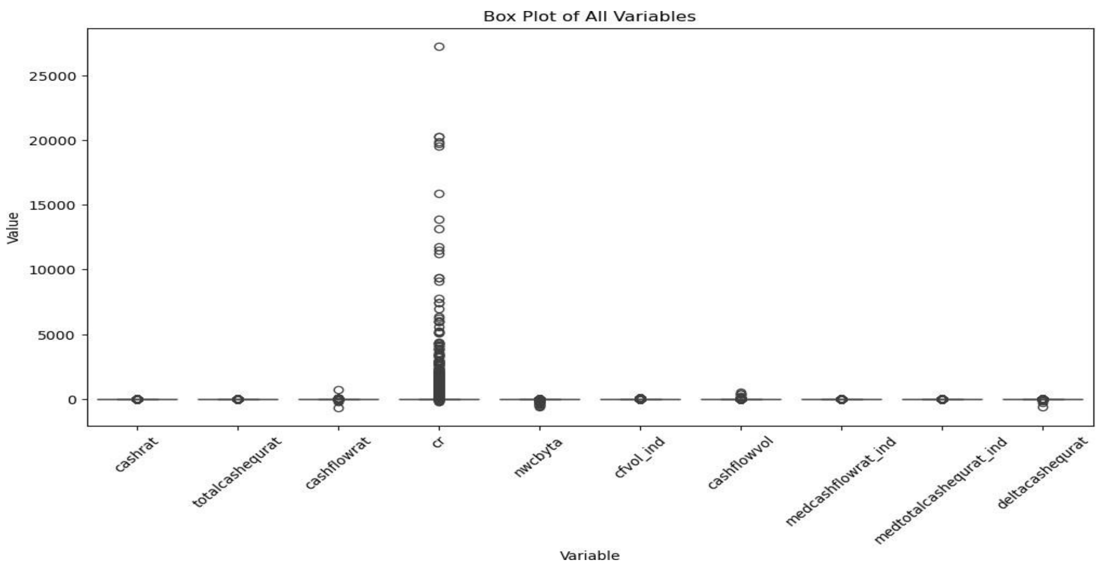


Figure 5: Boxplots for Liquidity Variables

For Dividend-related variables, we removed "SPE_PAYOUTRAT", "SPE_DIVBYTA", "SPE_DIVBYOP", and "EQUITY_DIVBYOP" based on the correlation matrix.

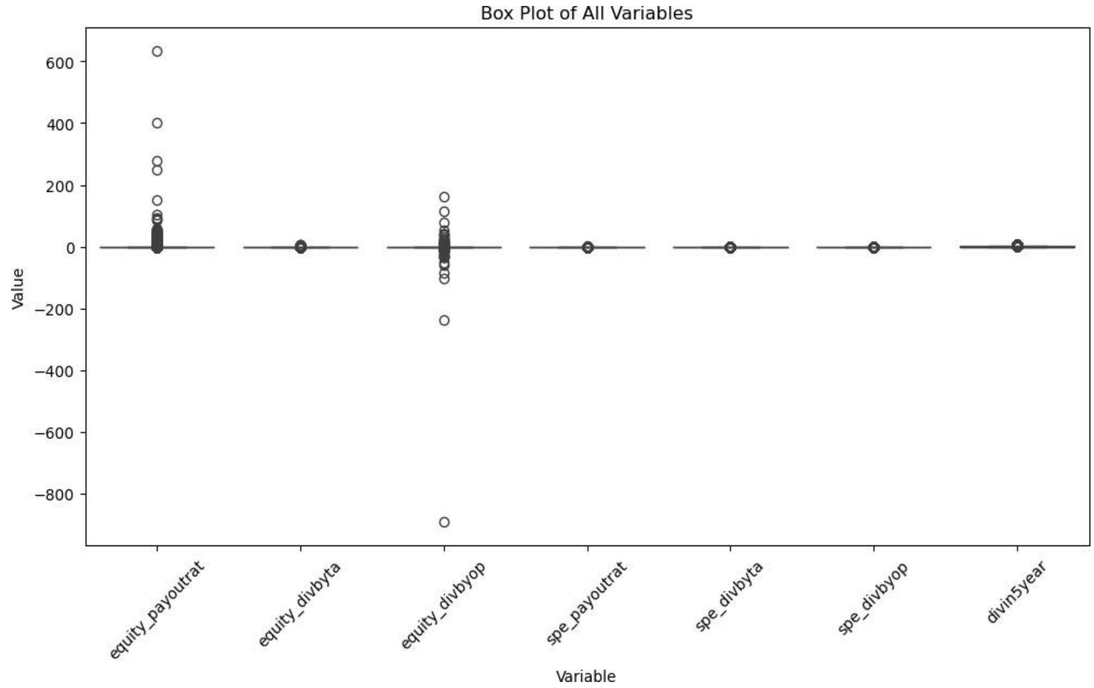


Figure 6: Box plots for Dividend-Related Variables

For Firm-related variables, we removed "OWNER_GROUP" due to its high correlation with "OWNER_PRIVATE." Also, we applied log transformations to "RNDBYREV", "RNDBYTA" and "INTRATE", Based on the correlation matrix, we removed "LOG_RNDBYREV", "LOG_RNDBYTA", "LOG_INTRATE", and "RNDBYREV".

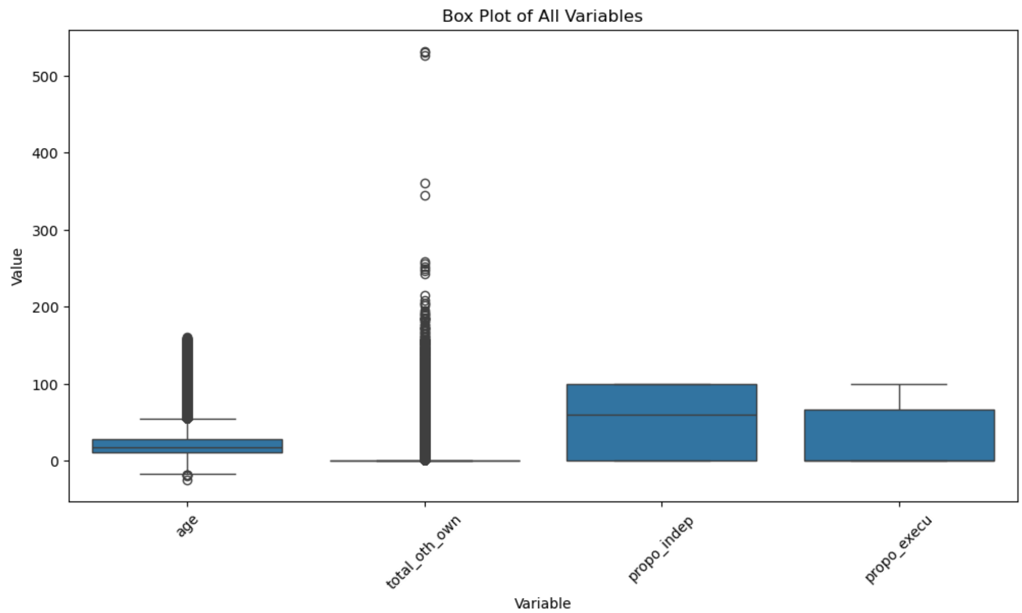


Figure 7: Box plots for firm-related variables

For Country-level variables, we checked for correlations among the features and removed "CONS_PCT_GDP," "EXC_RATE," "UNEMP_RATE," "CREDIT_PRIVATE_GDP," "INV_PCT_GDP," "EXPND_PCT_GDP," and "MRK_CAP_PCT_GDP."

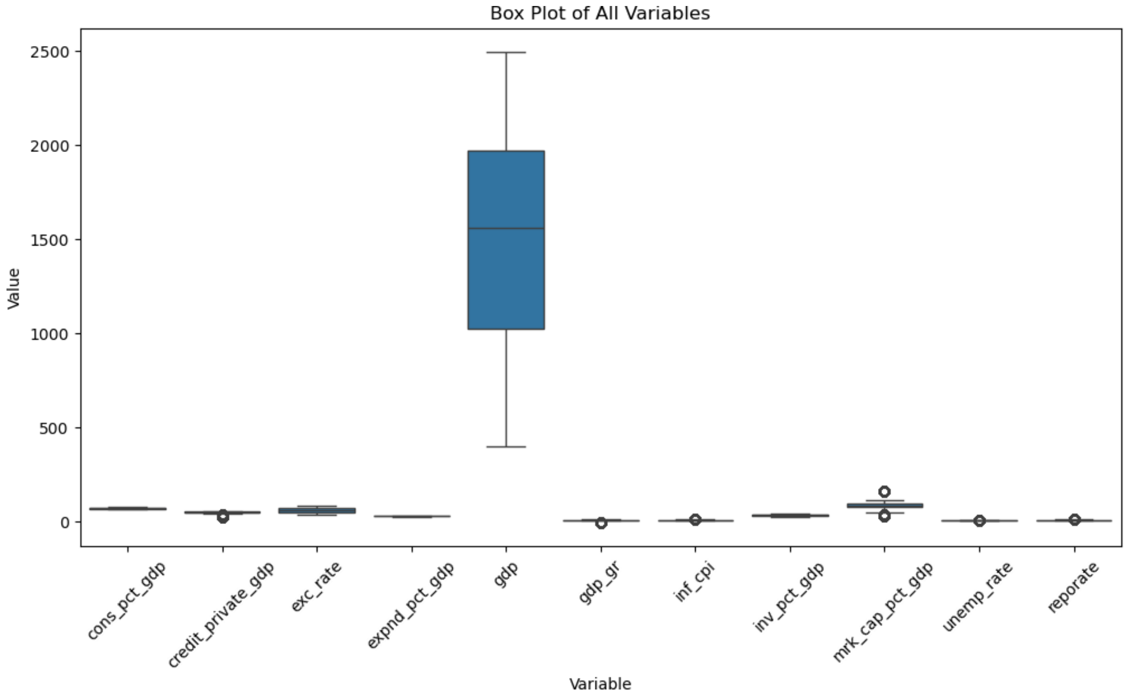


Figure 8: Box plots for Country-level variables

For expense expense-related variable, we have made a boxplot and added a log transform of highly skewed features. Added LOG_RNDBYREV, RNDBYTA, and INTRATE. And by observing Correlation matrices and VIF score, we have removed 'LOG_RNDBYREV', 'LOG_RNDBYTA', 'LOG_INTRATE','RNDBYREV'.

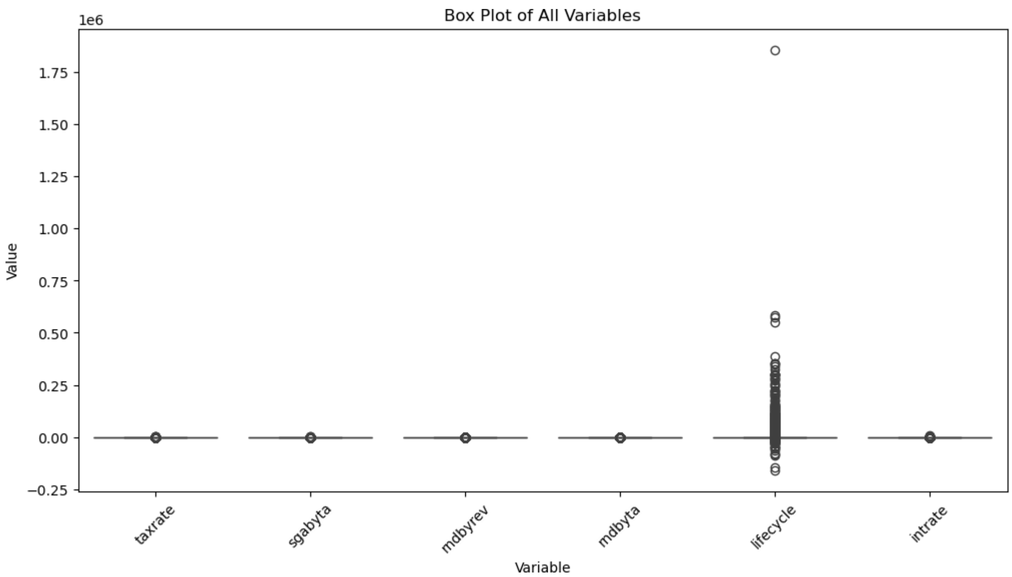


Figure 9: Box plots for expense-related variables

We have mansion Correlation among the variables to the target variables in Table 4 below. Along with the VIF score of each variable. Which helps us to understand the final data properly.

Table 4: Correlation with Target variable and VIF score for each Feature

Variable	**Correlation With Dividend Dummies**			**VIF Score**
	Equity	**Special**	**equity growth**	
Leverage:				
LOG_LTLEV	-0.1583	-0.0246	-0.1117	1.7749
LOG_STLEV	-0.1217	-0.0231	-0.0899	1.5508
MAT	-0.0902	-0.0072	-0.0545	1.5630
MEDLEV1_IND	-0.0491	-0.0057	-0.0484	2.4656
OVERLEV1	-0.1620	-0.0289	-0.1201	1.5717
DELTATOTALDEBTRAT	0.0024	0.0003	0.0016	1.0327
DEBTISSUANCE	-0.0548	-0.0121	-0.0475	1.2177
LOG_DELTAPAIDUPCAP	-0.0289	-0.0046	-0.0184	1.0409
Profitability:				
ROCE	0.0029	0.0006	0.0022	1.3744
ROE	0.0087	0.0000	0.0100	1.0410
ROA	0.0019	0.0003	0.0014	8.0626
EBITBYTA	0.0073	0.0016	0.0056	1.7001
EBITMARGIN	0.0008	0.0001	0.0006	12.4454
PATMARGIN	0.0059	0.0009	0.0042	2.2252
PATGROWTH	-0.0006	-0.0004	-0.0003	7.2444
DUMMY_PATGROWTH	0.1223	0.0235	0.1137	1.1801
LOG_EARNINGVOL	-0.0677	-0.0080	-0.0465	2.5281
REVGROWTH	-0.0012	0.0049	0.0008	1.0182
DUMMY_REVGROWTH	0.1060	0.0205	0.0901	1.2377
LOG_REVVOL	0.2405	0.0565	0.1918	3.1714
Asset and Investment:				
SIZE	0.2636	0.0638	0.2046	3.6649
TANG	-0.0208	0.0003	-0.0195	2.5868
MEDTANG_IND	0.0496	0.0148	0.0246	3.0647
OVERTANG	0.0161	-0.0003	0.0114	2.0102
LOG_CAPEX	0.0581	0.0103	0.0394	1.5251
MEDCAPEX_IND	0.1351	0.0221	0.0845	3.0045
OVERCAPEX	0.1086	0.0188	0.0799	1.4680
DELTACAPEX	0.0027	0.0006	0.0012	1.0262
LOGASSGR	0.0177	0.0005	0.0237	1.1821
MEDLOGASSGR_IND	0.1181	0.0142	0.0822	2.5658
LOG_DEPRATIO	-0.0353	-0.0024	-0.0217	1.4679
ASSETMAT	-0.0079	-0.0011	-0.0056	1.0035
Liquidity:				
CASHFLOWRAT	0.0153	0.0030	0.0115	2.0782
LOG_CASHFLOWVOL	-0.0734	-0.0083	-0.0500	2.3492

MEDCASHFLOWRAT_IND	0.1191	0.0198	0.1004	3.2591
OVERCASHFLOWRAT	0.2603	0.0482	0.1933	1.3413
LOG_CFVOL_IND	-0.0377	-0.0093	-0.0221	1.2336
TOTALCASHEQURAT	0.0810	0.0187	0.0703	1.6625
MEDTOTALCASHEQURAT_IND	-0.0010	0.0003	0.0168	2.5036
OVERTOTALCASHEQURAT	0.1086	0.0233	0.0828	1.4627
NWCBYTA	0.0234	0.0033	0.0169	1.2292
CR	-0.0063	-0.0016	-0.0035	1.0789
Corporate Governance:				
AGE	0.2226	0.0461	0.1348	1.2362
LISTED	0.2322	0.0532	0.1698	1.4113
LIFECYCLE	0.0026	-0.0007	0.0042	1.1078
OWNER_PRIVATE	-0.1876	-0.0509	-0.1452	1.3166
OWNER_GOV	0.0854	0.0431	0.0627	1.1574
TOTAL_OTH_OWN	0.3009	0.0723	0.2139	1.5378
PROPO_INDEP	0.0236	0.0005	0.0156	1.1931
PROPO_EXECU	0.1822	0.0391	0.1288	1.2897
Expense:				
RND_DUMMY	0.2572	0.0734	0.1916	1.2204
RNDBYTA	0.0018	0.0006	0.0015	2.6924
SGABYTA	-0.0030	-0.0011	-0.0016	1.7724
INTRATE	0.0025	-0.0003	0.0029	1.0010
TAXRATE	0.0025	0.0003	0.0017	1.0003
Dividend:				
EQUITY_DIVDUMMY	0.7720	0.1242	0.4274	5.1306
SPE_DIVDUMMY	0.1183	0.2255	0.0689	1.0307
EQUITYDIVGROWTH_DUMMY	0.5378	0.1092	0.3345	2.1438
DIVIN5YEAR	0.7396	0.1300	0.4496	3.9669
EQUITY_PAYOUTRAT	0.0811	0.0157	0.0364	1.0424
EQUITY_DIVBYTA	0.1839	0.0514	0.0862	1.1587
Macroeconomic:				
GDP	-0.0237	-0.0016	0.0098	3.0414
GDP_GR	0.0343	0.0051	0.0339	1.3316
INF_CPI	0.0513	0.0084	0.0308	1.3054
REPORATE	0.0483	0.0085	0.0112	2.6259
Industry:				
INDUSTRY_46.0	-0.0745	-0.0165	-0.0522	1.8299
INDUSTRY_77.0	-0.0577	-0.0098	-0.0422	1.3463
INDUSTRY_20.0	0.0936	0.0114	0.0663	1.1424

4.1.2 Data Imputation

We reduced the dataset to 69 features after addressing multicollinearity and identifying redundant features from the exploratory data analysis. However, the dataset still has missing values, which need to be addressed. since most machine learning algorithms cannot handle missing data. since most machine learning algorithms cannot handle missing data, To solve this, we employed the K-Nearest Neighbour (KNN) imputation technique [1]. This method operates row-by-row, leveraging the non-missing features in each row to identify the k nearest neighbours based on a chosen distance metric, such as Euclidean distance. After finding the nearest neighbours, all missing values in a row are filled in at one time by taking either the mean or weighted mean of the feature values corresponding to the neighbours. Every missing feature is filled in separately, applying the same neighbour set obtained from the full features. This is done to make sure that imputation does not create sequential dependency within the same row, which maintains data integrity.

A few tree-based machine learning algorithms, such as Decision Trees, Random Forest, and XGBoost, can handle missing values without requiring imputation. However, their performance tends to be low when missing values are not treated properly. Thus, imputed data will generally provide improved results even for such algorithms, as it helps achieve more accurate and reliable model predictions.

4.1.3 Data Oversampling

The target variables is highly imbalanced. F_EQUITY_DIVDUMMY has 245,070 zeros and 38,092 ones, indicating that 86.54% of companies do not pay dividends. F_EQUITYDIVGROWTH_DUMMY has 262,046 zeros and 21,116 ones, showing that only 7.46% of companies increase their equity dividends. Similarly, F_SPE_DIVDUMMY has 282,310 zeros and only 852 ones, highlighting an extremely imbalanced distribution where just 0.3% of companies issue a special dividend.

Since the target variables are highly imbalanced, this can bias the machine learning model to predict only the majority class, leading to deceptively high accuracy a scenario that is undesirable. To address this issue, we employed **SMOTE (Synthetic Minority Oversampling Technique)**. SMOTE generates synthetic samples for the minority class, helping to balance the dataset and make the model more generalizable.

4.1.4 Data Scaling

Algorithms like Logistic Regression and Neural Networks associate weights or coefficients with predictors. Hence, scaling is essential before training the model. For this purpose, we have used standard scaling.

$$z_j^i = \frac{(x_j^i - \overline{x^i})}{\sigma^i}$$

Where $\overline{x^i}$ is the mean value of the ith feature, σ^i is standard deviation and z_j^i is standard scaled features.

Tree-based algorithms operate on the principle of entropy and do not require feature scaling. Therefore, we have used the features directly for models such as Decision Tree, Random Forest, and XGBoost.

4.2 Machine Learning Models

In this study, we have employed several machine learning models to predict F_EQUITY_DIVDUMMY", "F_EQUITYDIVGROWTH_DUMMY" and "F_SPE_DIVDUMMY", predicting using 69 predictor variables that belong to different categories like leverage, Profitability, Liquidity, Asset and Investment Related Variables, Industry Level features, Dividend, Macroeconomic Features, and Firm related variables. We have used a basic model like logistic regression, incorporating Lasso, Ridge, and Elasticnet Regularization with Hyperparameter tuning. We have also used Tree based models like Decision Tree, Random Forest, XGBoost, and among them, we hyperparameter tune the model that performs best. We have also used a deep learning base model and used Artificial Neural Network (ANN) for the task. Each model is described below with its mathematical formulation.

4.2.1 Traditional Statistical Methods

Traditional Statistical methods is classical approaches used in data analysis and modelling, relying on well-established statistical theories. Logistic regression is a very popular statistical model widely used for binary class classification. Here in this section, we have used logistic regression and different variations of logistic regression, including lasso, ridge, and elasticnet regularization, with appropriate hyperparameter tuning. Logistic Regression is a foundational classification algorithm widely used for binary classification tasks due to its simplicity and interpretability.

4.2.1.1 Logistic Regression:

In this section, we applied Logistic Regression on processed data to predict the target variables. logistic regression is a basic classification algorithm that is frequently used for binary classification tasks. moreover, it is simplistic and interpretable. Logistic Regression stabilized the relationship between a dependent binary variable and one or more independent variables by estimating probabilities using a sigmoid function. It makes the assumption that the log-odds of the dependent variable and the independent variables have a linear relationship.

The mathematical formula for Logistic Regression is given below:

$$P(Y = 1 \mid X) = \frac{1}{1 + e^{-z}}$$

where:

$$z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

Here, P(Y=1|X) represents the probability of the target variable Y being 1, X_i are the independent variables, β_0 is the intercept, and β_i are the coefficients for each independent variable.

4.2.1.2 Lasso Regularization:

Lasso (Least Absolute Shrinkage and Selection Operator) regularization is a technique used for Feature selection, reducing overfitting, which leads to enhancing the predictive power of logistic regression while performing feature selection as well. It brings in an L1-norm penalty to the loss function, which encourages sparsity in the model by shrinking some feature coefficients to zero. Lasso is especially helpful in high-dimensional datasets, Lasso is especially useful for nullifying many unnecessary or superfluous features in high-dimensional datasets.

The objective function for logistic regression with Lasso regularization is:

$$L(\beta) = -\frac{1}{N}\sum_{i=1}^{n}[y_i \log(y_i) + (1 - y_i)\log(1 - y_i)] + \lambda \sum_{j=1}^{p} |\beta_j|$$

Where, $L(\beta)$ is the loss function. y_i is the actual label for the ith sample. $\hat{y}_i$ is the predicted probability. N is the number of samples. β_j represents the model coefficients for the jth feature. λ is the regularization parameter, and p is the total number of features.

The regularization parameter λ controls the trade-off between fitting the data and penalizing large coefficients. A higher value of λ results in more coefficients being reduced to zero, simplifying the model.

Lasso regularization not only prevents overfitting but also improves model interpretability by identifying the most relevant features, making it a valuable tool in predictive modelling.

4.2.1.3 Ridge Regularization:

Ridge regularization, also recognized as L2-norm regularization, is a method used to address overfitting in logistic regression by adding a penalty term to the loss function. Contrasting Lasso, which promotes sparsity, Ridge regularization shrinks the coefficients towards zero without forcing them to become exactly zero, making it suitable for datasets where all features may have some predictive power.

The objective function for logistic regression with Ridge regularization is:

$$L(\beta) = -\frac{1}{N}\sum_{i=1}^{n}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)] + \lambda\sum_{j=1}^{p}\beta_j^{\,2}$$

Where, $L(\beta)$ is the loss function. y_i is the actual label for the ith sample. $\hat{y}_i$ is the predicted probability. N is the number of samples. β_j represents the model coefficients for the jth feature. λ is the regularization parameter, and p is the total number of features.

The regularization parameter λ controls the strength of the penalty. Greater values of λ cause the coefficients to shrink more significantly, which lowers the complexity of the model and avoids overfitting. Ridge regularization is particularly effective when multicollinearity is present, as it reduces the sensitivity of the model to correlated features, ensuring more stable predictions

4.2.1.4 Elastic Net Regularization:

Elastic Net regularization is a strong technique that has the advantages of both Lasso (L1) and Ridge (L2) regularization. making it a versatile method for addressing overfitting and feature selection. it works exceptionally well when the features are highly correlated. Elastic Net introduces a combination of L1 and L2 penalties to the model, resulting in a balanced approach that makes use of the powers of both methods.

The objective function for logistic regression with Elastic Net regularization is:

$$L(\beta) = -\frac{1}{N}\sum_{i=1}^{n}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)] + \lambda\,[\alpha\sum_{j=1}^{p}|\beta_j| + (1 - \alpha)\sum_{j=1}^{p}\beta_j^{2}]$$

Where, $L(\beta)$ is the loss function. y_i is the actual label for the ith sample. $\hat{y}_i$ is the predicted probability. N is the number of samples. β_j represents the model coefficients for the jth feature. λ is the regularization parameter controls the overall strength of the penalty. α is the mixing parameter $(0 < \alpha < 1)$ that determines the balance between L1 and L2 penalties, and p is the total number of features.

Elastic Net behaves similarly to Lasso and imposes the L1 penalty alone when α=1. It behaves similarly to Ridge and imposes the L2 penalty alone when α=0. it behaves similarly to Ridge, imposing only the L penalty. Intermediate values of α achieve a balance between them.

Elastic Net regularization is especially useful when features are correlated or there are more features than observations. Not only does it prevent overfitting, but it also automatically selects features, resulting in a more stable and interpretable model.

4.2.2 Classical Machine Learning Models

In this section, we investigate classical machine learning models, including Decision Tree, Random Forest, and XGBoost. The Decision Tree algorithm divides the data into subsets according to feature values, acting as a rule-based model. It is Simple and interpretable, making it a popular choice for classification tasks. Building on the Decision Tree, Random Forest combines multiple decision trees to create an ensemble model, improving prediction accuracy and robustness. By sequentially optimizing decision trees to reduce errors, the gradient boosting algorithm XGBoost advances ensemble learning and achieves excellent performance on challenging datasets. Together, these models show adaptability and efficiency in managing structured data and a range of classification problems.

4.2.2.1 Decision Tree:

The Decision Tree algorithm creates a tree model to predict by recursively partitioning the dataset into subsets based on the feature thresholds in a tree. Starting with the root node representing the entire dataset, the algorithm will evaluate all of the attributes and select the optimum one that splits the data most well according to some criteria, say Information Gain or Gini Impurity in case of classification tasks or Mean Squared Error (MSE) for regression tasks. This process involves the determination of a threshold value in the selected feature and partitioning the set into two subsets based on the feature value being less than or equal to the threshold and the feature value being more than the threshold. The algorithm continues this recursive partitioning for each subset, creating new nodes and reducing the separation of the target variable with each split. The algorithm stops when it encounters a termination criterion, say maximum tree depth, minimum node size, or pure class distribution. The final nodes are known as leaf nodes and store the predictions, i.e., the majority class in the case of classification problems or the mean value for regression problems. This scheme enables Decision Trees to be able to learn non-linear relationships of data and form an understandable model that can potentially be represented as a series of decision rules.. Each split aims to maximize the purity of the resulting subsets using metrics such as Gini Impurity or Information Gain. The predicted class is assigned at a leaf node after the tree has been traversed from the root to the leaf node.

The mathematical criterion for splitting using Gini Impurity is given as:

$$G = 1 - \sum_{i=1}^{c} {p_i}^2$$

Where

G is the Gini Impurity, c is the number of classes, and p_i is the proportion of samples belonging to class i in the subset.

Alternatively, Information Gain is calculated as:

$$IG = H_{parent} - \sum_{k=1}^{K} \frac{n_k}{n} H_k$$

where:

IG is the Information Gain, H_{parent} is the entropy of the parent node, H_k is the entropy of the kth child node, n_k is the number of samples in the kth child node, and n is the total number of samples in the parent node.

This algorithm is effective for both classification and regression tasks and can handle nonlinear relationships in the data.

4.2.2.2 Random Forest:

Random Forest is an ensemble learning technique that builds multiple decision trees and combines their outputs to make more accurate and reliable predictions. Each decision tree is first constructed using a randomly chosen subset of the data—a procedure called bootstrapping—in which replacement samples are taken. In addition, for every split in a tree, only a randomly chosen subset of the features are used to avoid similarity among the trees. This diversity minimizes the danger of overfitting and enhances the capability of the model to generalize from existing data to new data. When all the trees have been constructed, the predictions from the trees are combined to yield the final output. For regression tasks, the mean of the outputs of the trees is employed, and for classification tasks, the most frequent vote of all the trees decides the predicted class. Random Forest maintains its accuracy, is noise-resistant, and works well with large datasets and high-dimensional features by combining the strength of multiple trees. Its ability to strike a balance between simplicity and complexity makes it a popular and adaptable machine learning algorithm.

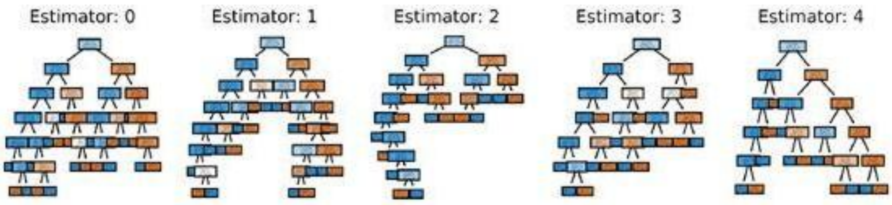


Figure 10: Random Forest Visualization Adapted from Codementor (2020).

The formula for the aggregated prediction in Random Forest for classification is:

$$\hat{y} = \text{mode}\{h_1(\text{x}), h_2(\text{x}), \dots, h_T(\text{x})\}$$

where:

$\hat{y}$ is the final predicted class, $h_t(\text{x})$ is the prediction of the t^{th} tree, and T is the total number of trees in the forest.

4.2.2.3 XGBoost:

XGBoost, short for eXtreme Gradient Boosting, is a robust ensemble learning algorithm based on the gradient boosting architecture. It builds a sequence of decision trees in sequence, each tree attempting to decrease the errors of the previous ones. In contrast to Random Forest, where trees are constructed separately, XGBoost is concerned with additive corrections and utilizes each tree to reduce the residual errors of the model. The algorithm learns to minimize a loss function and exploits both first and second-order derivatives to achieve higher split accuracy and stable predictions. XGBoost is performance-optimized with advanced features such as regularization to prevent overfitting, parallel processing for speed, and memory-friendly usage for processing large data. Its ability to balance performance, speed, and accuracy makes XGBoost a preferred choice for many complex machine learning tasks, particularly in structured data problems.

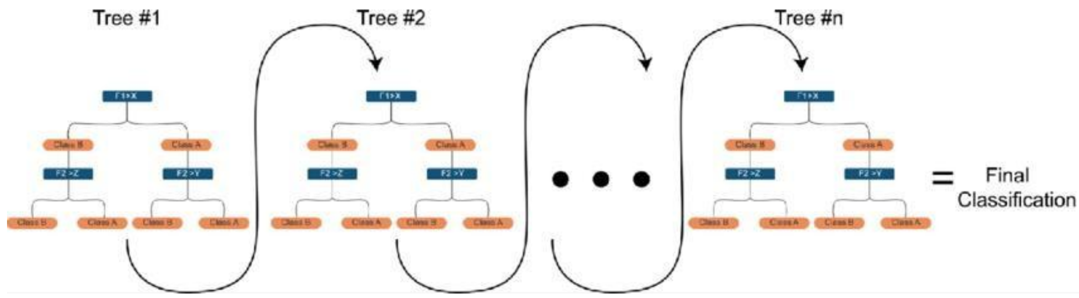


Figure 11: XGboost Visualization Adapted from Abualdenien & Borrmann (2022)

The formula for updating predictions in XGBoost is:

$$y_i^{(t)} = y_i^{(t-1)} + \eta f_t(x_i)$$

where:

$y_i^{(t)}$ is the updated prediction at iteration t. $y_i^{(t-1)}$ Is the prediction from the previous iteration. η is the learning rate, and $f_t(x_i)$ is the t^{th} tree's prediction for input x_i.

XGBoost is highly regarded for its predictive power and scalability, making it a preferred choice in many machine learning competitions.

4.2.3 Modern Deep Learning Methods

4.2.3.1 Artificial Neural Networks (ANN):

Artificial Neural Networks (ANN) are computational models that draw inspiration from the architecture and operation of the human brain. ANNs are composed of interlinked layers of nodes (neurons), where every node receives input data via weighted connections, applies an activation function, and sends the result to the next layer. ANNs can learn complex, non-linear associations in data and are highly adaptable. ANNs usually consist of an input layer (features), one or more hidden layers (learning patterns), and an output layer (prediction). The model's weights are tuned by means of algorithms such as backpropagation, where the error between predicted and target values is reduced by iteratively changing the weights.

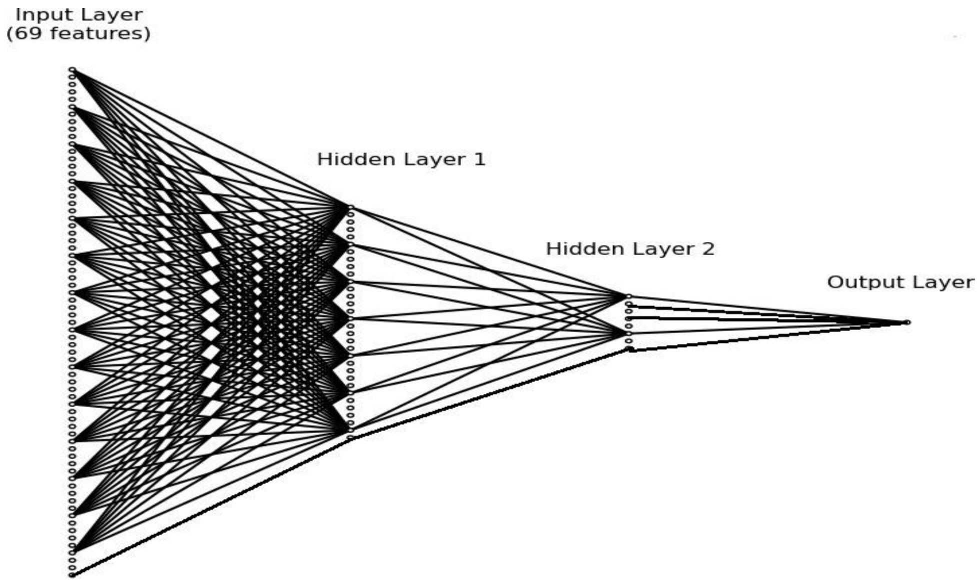


Figure 12: ANN Feedforward Neural Network Architecture

The forward propagation for a single neuron can be represented mathematically as:

$$z = \sum_{i=1}^{n} w_i x_i + b$$

$$a = f(z)$$

where:

x_i These are the input features. w_i are the corresponding weights. b is the bias term. z is the weighted sum, and $f(z)$ is the activation function (e.g., sigmoid, ReLU, tanh).

During training, the model minimizes the error using a loss function, such as Mean Squared Error (MSE) for regression or Cross-Entropy Loss for classification, given by:

$$L = -\frac{1}{N}\sum_{i=1}^{n}[y_i \log(y_i) + (1 - y_i)\log(1 - y_i)]$$

Where, y_i is the actual value, $\hat{y}_i$ is the predicted value, and N is the number of samples.

ANNs are powerful for handling large, complex datasets and have been successfully applied in tasks like image recognition, natural language processing, and predictive modelling.

4.3 Evaluation Metrics for Model Performance

For evaluating the performance of our models, we have considered various performance metrics. Additionally, we have analyzed confusion matrices for all the models, which are discussed in the results section.

Considering,

$$tp = true\ positives$$

$$tn = true\ negatives$$

$$fp = false\ positives$$

$$fn = false\ negatives$$

Below are the definitions of all the attributes we considered:

Accuracy:

Accuracy is the measure of true predictions out of total predictions. Measured by taking a ratio of both true positive and true negative with total predictions.

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn}$$

Precision:

precision measures the proportion of true positive predictions among all the positive predictions made by the model. The higher the precision lower the risk of false positive prediction.

$$Precision = \frac{tp}{tp + fp}$$

Recall:

Recall is the proportion of true positives predicted by the model concerning the actual positives. The higher the recall lower the false negatives.

$$Recall = \frac{tp}{tp + fn}$$

F1score:

Precision and recall are not sufficient to determine a model's performance. F1score is the harmonic mean of precision and recall, providing a balanced metric when precision and recall are equally important.

$$F1score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Error Rate:

It is a measurement of false predictions. Error rate is the proportion of incorrect predictions among the total predictions.

$$Error\ rate = \frac{fp + fn}{tp + tn + fp + fn}$$

Specificity:

It is also known as the True Negative Rate. It measures the proportion of actual negative cases that we correctly identify by the model. It is similar to recall but for negative classes.

$$Specificity = \frac{tn}{tn + fp}$$

Sensitivity:

It is another name for recall, also known as the True Positive Rate. Which measures how well the model identifies actual positives.

$$Sensitivity = \frac{tp}{tp + fn}$$

AUC-ROC:

AUC-ROC stands for Area Under Receiver Operating Characteristic Curve. It evaluates the model’s ability to distinguish between positive and negative classes. The ROC curve plots the true positive rate (sensitivity) against the false positive rate (1-specificity).

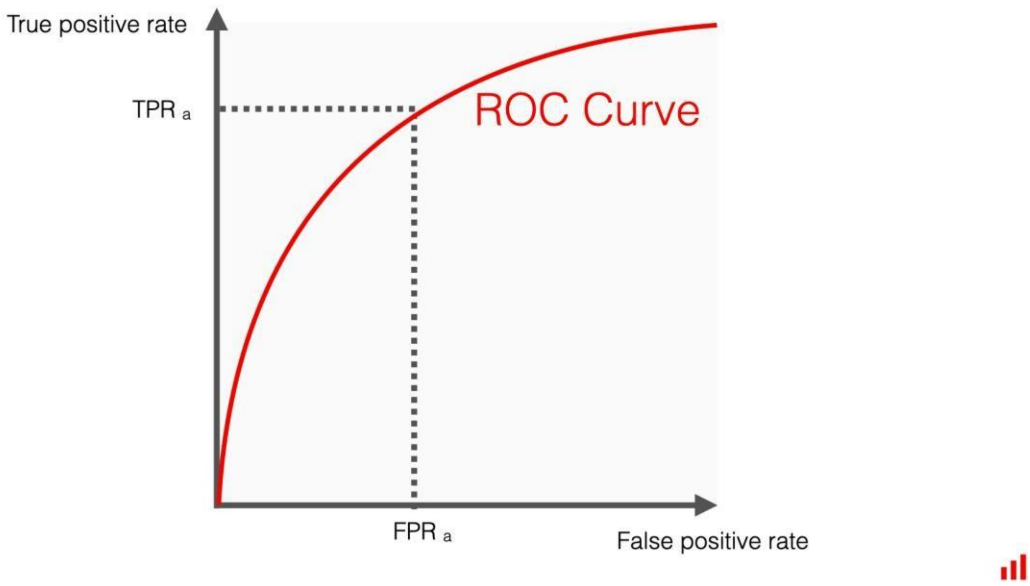


Figure 13: ROC curve Adapted from EvidentlyAI (2025).

G-Mean (Geometric Mean):

G-Mean evaluates the balance between sensitivity and specificity. It is particularly useful for imbalanced datasets.

$$GMean = \sqrt{Sensitivity \times Specificity}$$

Balanced Accuracy:

It is the average of sensitivity and specificity. Addressing the imbalance in the class distribution.

$$Balance\ Accuracy = \frac{Sensitivity + Specificity}{2}$$

Youden's Index:

Youden's index evaluates the effectiveness of a diagnostic test by combining sensitivity and specificity into a single metric.

$$youden's\ index = Sensitivity + Specificity - 1$$

4.4 Hyperparameter Tuning

As described in the data processing section, much of the effort was put into prepping the data so that it gives the best predictive performance. We did feature selection using in-house exploratory data analysis to choose the most useful features and eliminate the non-essential ones. We also transformed the chosen features so that it prepare the data properly for the model. To handle missing values, we employed advanced imputation techniques, such as KNN imputation, to successfully address null values in the dataset. Outliers present in the data were treated using Winsorization, which helped to maintain the integrity of the dataset while mitigating the impact of extreme values.

In addition, hyperparameter tuning was recognized as being among the most important steps in constructing the model. This step, explained in the next section, was necessary for maximizing the performance of the model and obtaining the optimal trade-off between accuracy and generalizability. Collectively, these preprocessing and tuning procedures guaranteed the construction of a strong and stable predictive model.

4.4.1 Logistic regression with Elastic Net Regularization

Hyperparameter optimization is important in order to best tune the performance of a logistic regression model under Elastic Net regularization. It aims at finding the optimal combination of hyperparameters to optimize predictive accuracy with the retention of model interpretability. The two important hyperparameters that are optimized during Elastic Net are the penalty mixing ratio (α) and the penalty coefficient (λ). The parameter λ controls the overall strength of regularization, with higher values enforcing greater shrinkage of coefficients to manage the bias-variance trade-off. The parameter $\alpha\alpha\alpha$ determines the balance point between L1 (Lasso) and L2 (Ridge) penalties: 0 corresponds to pure Ridge regularization, 1 to pure Lasso, and intermediate values draw benefits from both approaches.

To perform this process, a grid of possible values for λ and α was established to systematically evaluate combinations. The grid consisted of $1/\lambda$ values of 0.1, 1, and 10 to cover different levels of regularization strength, as well as α values of 0.3, 0.5, and 0.7 to cover different mixing ratios of L1/L2. A grid search with cross-validation was then performed to evaluate the model's performance for each combination. This helped the chosen hyperparameters generalize well to unseen data, minimizing the possibility of overfitting. Consequently, the model attained a good balance between regularization and feature selection, improving its predictive accuracy and stability.

4.4.2 Decision Tree

Hyperparameter optimization of a Decision Tree model involves optimizing max depth, min samples split, and min samples leaf to get the best trade-off between model complexity and predictive power. Maximum depth is the depth at which a decision tree continues to divide. min samples split Controls the number of splits. Higher values make the model more conservative. min samples leaf, which means the minimum sample at the leaf node. The max depth parameter, tested with values None, 10, 30, and 50, determines the maximum depth of the tree, helping to prevent overfitting or underfitting. The minimum number of samples required to split an internal node is defined by the min samples split, which is experimented with values 2, 10, and 20 to balance computational cost and tree growth. The min samples leaf, varied with values 1, 2, 5, and 10, ensures each leaf node has enough samples to maintain stability and avoid noisy predictions. Using grid search with cross-validation, these parameters were systematically evaluated to find the combination that optimizes model performance while maintaining generalization, leading to a robust and interpretable decision tree.

4.4.3 Random Forest

Hyperparameter tuning of a Random Forest model means setting n estimators, max depth, min samples split, and min samples leaf in a way that it balances model accuracy and computational complexity. The n estimators parameter, representing the number of trees in the forest, was tested with values 50, 100, and 200 to evaluate the trade-off between computational time and ensemble stability. The max depth parameter, which was tried with values None, 10, and 20, controls the depth of individual trees and helps control overfitting. The min samples split, which was tried with values 2, 5, and 10, controls the minimum number of samples needed to split an internal node and impacts tree complexity and cost. Lastly, the min samples leaf was varied with values 1, 2, and 4, ensuring that leaf nodes have a sufficient number of samples to provide predictive stability. Using grid search with cross-validation, the parameters were experimented with systematically to acquire the optimal configuration that enhances both model generalization and predictive accuracy, resulting in a stable Random Forest model.

4.4.4 XGBoost

Hyperparameter tuning of the XGBoost model includes the tuning of n estimators, max depth, and learning rate to balance predictive performance and computational cost. The n estimators, or boosting rounds, were changed using values 50, 100, and 200 to study the impact of more iterations on model performance. The max depth, having test values 3, 6, and 10, controls the depth of each tree, with a balance between generality and complexity. The learning rate, having tested values 0.01, 0.1, and 0.2, determines the update step size when performing an update on weights while boosting, enabling fine control over predictions. Grid search with cross-validation was used to systematically try out these combinations of parameters so that the final model maximizes on computational efficiency, prediction accuracy, and generalization to new data, leading to an appropriately calibrated XGBoost model.

4.4.5 Artificial Neural Network

We have used an Artificial Neural Network (ANN) model with two hidden layers, Hyperparameter tuning involves optimizing learning rate, neurons first layer, neurons second layer, and dropout rate to achieve a balance between predictive performance and model generalization. To assess the effect on convergence speed and accuracy, the learning rate, which establishes the step size during weight updates, was adjusted with values of 0.01 and 0.001. The number of neurons in the first hidden layer, neurons first layer, was tested with values 32, 16, and 8, while the second layer, neurons second layer, was evaluated with values 16, 8, and 4 to identify configurations that best capture patterns in the data. To avoid overfitting, the dropout rate, which creates regularization by randomly deactivating neurons during training, was adjusted between 0.2, 0.3, and 0.4. Grid search with cross-validation was employed to systematically explore combinations of these parameters, resulting in a fine-tuned ANN model that achieves an optimal trade-off between complexity, generalization, and predictive accuracy.

4.5 Cross-validation

Cross-validation is a resampling technique used to evaluate the performance of a machine learning model by dividing the dataset into multiple subsets or "folds." The model is trained on some folds and tested on the remaining fold(s), and this process is repeated several times (e.g., in k-fold cross-validation, the data is split into k parts). Each fold serves as a test set once, ensuring the model's performance is tested on all data points. This helps assess model generalization, reduce overfitting, and ensure robustness by providing an averaged performance metric across folds. In our project, we used 10-fold cross-validation for all models except Artificial Neural Networks (ANN), for which we opted for 5-fold cross-validation due to the computational complexity of ANN. Consequently, all performance metrics derived in this study are aggregates across folds, minimizing bias and ensuring accuracy. Additionally, cross-validation is particularly valuable in projects with limited or imbalanced data, as it ensures the model is exposed to diverse subsets, improving reliability and robustness in real-world scenarios

5 Results and Discussion

In this section we discuss the results of the machine learning models. To evaluate the contribution of non-dividend-related variables to the model's accuracy, we analyzed the predictions under two conditions: one using all selected variables (including dividend-related variables) and another excluding dividend-related variables. This comparison allowed us to determine whether the model's accuracy heavily relies on the inclusion of dividend-related variables or whether other collected data also plays a significant role in improving prediction accuracy.

If the results show that excluding dividend-related variables significantly reduces the model's accuracy, it would indicate that these variables have a substantial impact on the model's performance. In such a case, we can consider exploring additional methodologies, such as time series analysis, to gain deeper insights into the relationship between dividend-related variables and the predictive outcomes. This approach will help us better understand the overall contribution of different variables to the model's accuracy and guide further improvements.

5.1 Results of Equity Dividend Dummy

In Table 5, we show the results of the 8 models discussed above. The table shows the accuracy and the F1 scores of each model in two specifications: one without dividend history variables and one with dividend history variables. All the models are trained with hyperparameter tuning, and we have achieved the best-performing models for each category of algorithm. For the equity dividend dummy after applying SMOTE, we have 4,90,140 data instances.

Out of all the models, the tree-based models give the best accuracy and F1 score in both specifications. Among these, the random forest method gives the best accuracy and the F1 score for both specifications. For the model that includes dividend history variables, the accuracy and F1 score are 96.73% and 96.72%, respectively. When we exclude dividend history variables, the accuracy and F1 score reduce very slightly to 93.62% and 93.43%, respectively.

Table 5: Performance of each model for the equity dividend dummy

F_EQUITY_DIVDUMMY	With Dividend-Related Variables		without Dividend-Related Variables	
	Accuracy	F1 score	Accuracy	F1 score
Statistical models				
Simple Logistic Regression	91.81%	91.57%	81.42%	81.75%
Logistic Regression with Lasso Regularisation	91.81%	91.57%	81.42%	81.76%
Logistic Regression with Ridge	91.81%	91.57%	81.42%	81.76%
Logistic Regression with Elastic Net	91.81%	91.57%	81.42%	81.76%
Classical Machine learning models:				
Decision Tree	95.99%	95.96%	89.96%	89.80%
Random Forest	96.73%	96.72%	93.62%	93.43%
XGBoost	96.21%	96.19%	92.69%	92.42%
Deep Learning				
Artificial Neural Network	93.11%	93.09%	86.45%	86.73%

After tree-based models, Artificial neural network gives the next best accuracy and F1 score of 93.11% and 93.09%, respectively. Statistical models perform the worst. There is a negligible performance difference between different regularization methods for logistic regression (Lasso, Ridge, and Elastic net). Among these, Elastic net performs slightly better compared to the other two and the model without regularization.

Now we present additional performance matrices in Table 6. Here we only report the performance of the best models from each category, viz., Logistic Regression with Elastic Net, Random Forest, and Artificial Neural Network. We discussed each of these three individually in subsequent sections.

Table 6: Performance summary of the best-performing models for the equity dividend dummy

Equity Dividend Dummy	With Dividend-Related variables			Without Dividend-Related Variables		
Performance metrics	Logistic regression	Random Forest	ANN	Logistic regression	Random Forest	ANN
Accuracy	0.9181	0.9673	0.9311	0.8142	0.9362	0.8645
Error Rate	0.0819	0.0327	0.0689	0.1858	0.0638	0.1355
Precision	0.9433	0.9696	0.9330	0.8030	0.9629	0.8500
Recall	0.8898	0.9649	0.9289	0.8326	0.9073	0.8853
F1 Score	0.9157	0.9672	0.9309	0.8176	0.9343	0.8673
Specificity	0.9465	0.9697	0.9333	0.7957	0.9651	0.8437
G-Mean	0.9177	0.9673	0.9311	0.8140	0.9358	0.8643
Balanced Accuracy	0.9181	0.9673	0.9311	0.8142	0.9362	0.8645
Youden's Gamma	0.8363	0.9347	0.8622	0.6284	0.8724	0.7290

5.1.1 Logistic Regression with Elastic Net for Equity Dividend Prediction

The best mode was arrived at after hyperparameter tuning. There are two hyperparameters, C and L1 Ratio. C represents the regularization strength. A higher value of C corresponds to lower regularization strength. The L1 ratio in Elastic Net is the proportion of L1 (Lasso) penalty relative to the combined L1 and L2 (Ridge) penalties. For the logistic regression model including dividend-history variables, the best parameters identified are C = 10 and L1 ratio = 0.3. For the model excluding dividend-related variables, the optimal parameters are C = 0.01 and L1 ratio = 0.3. This combination effectively prevents overfitting while ensuring good generalization to unseen data.

For logistic regression with elastic net, we achieved an accuracy of 91.81%, including dividend-history variables, and 81.42% when it was excluded. This indicates that the dividend-history variable significantly influences the prediction performance.

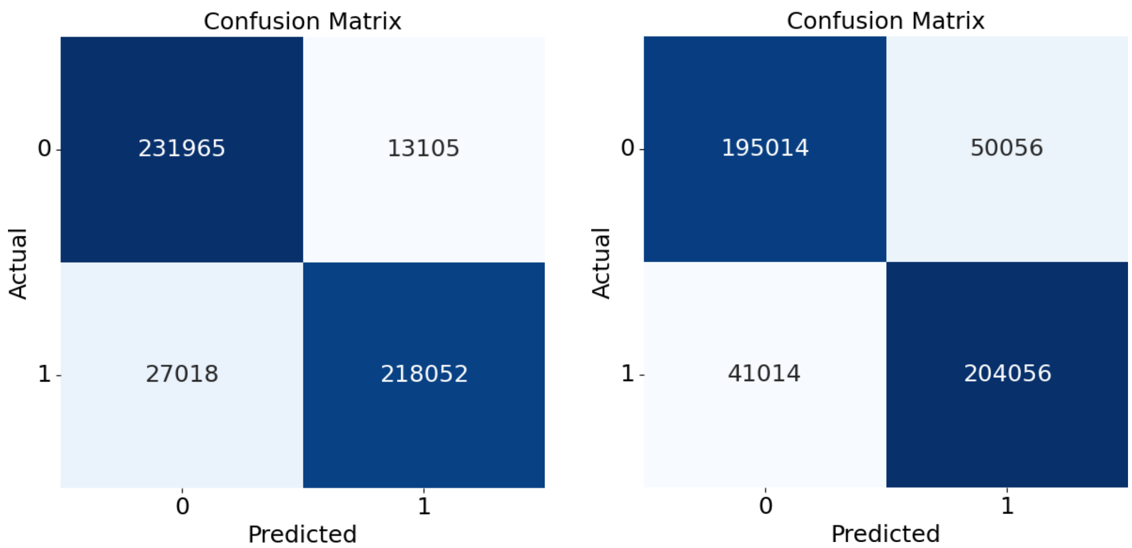


Figure 114: Confusion matrices from the logistic regression model for the equity dividend dummy

To arrive at the performance metrics described in Table 6, we have analysed the confusion matrices of the predictions, which directly give an idea about Type 1 and Type 2 errors. The confusion matrix is shown in Fig. 14. We now jointly look at the results from Table 6 and Fig. 14. For the model considering dividend-related variables, we observed that the Type 1 Error (False Positive) is 13105 and the precision is 0.9433. Whereas the Type 2 Error (False Negative) is 27018, and the recall is 0.8898. This indicates that the model predicts class 0 more accurately than class 1, meaning it has higher precision than recall. The overall F1 score of the model is 0.91118.

Opposite results are observed for the model excluding dividend-related variables. It exhibits a Type I error of 50,056 (with a precision of 0.8030) and a Type II error of 41,014 (with a recall of 0.8326). This model performs significantly worse than the previous model in terms of overall accuracy. However, given that the dataset is imbalanced with very few observations in the positive class, recall is more critical than precision. Since the model without dividend-related variables achieves higher recall relative to precision, it is preferable to the model with dividend-related variables in this context.

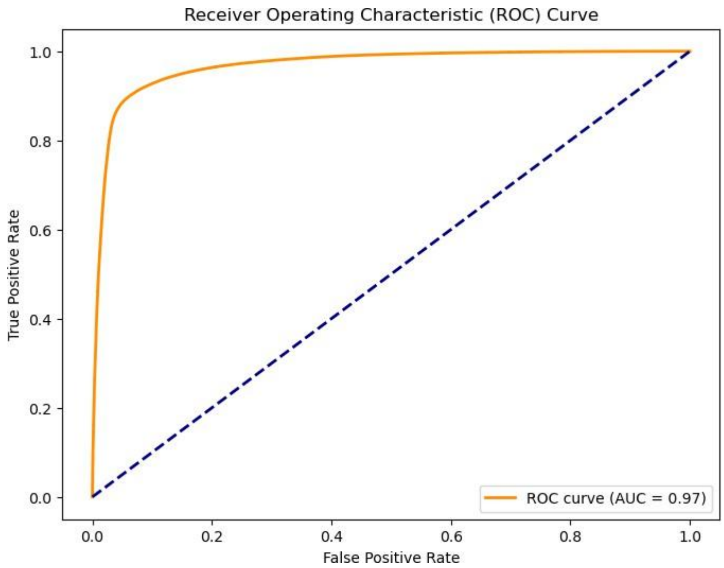


Figure 15: ROC curve from the logistic regression model for the equity dividend dummy

Further, we plot the ROC curve of the model with dividend-related variables in Fig. 15. The Area under the ROC curve is 0.97. This implies that the discriminative power of the model is good.

We present the feature importance of the top variables can be observed in Figure 16 below. Feature importance is determined by the coefficient values of each variable in the trained model.

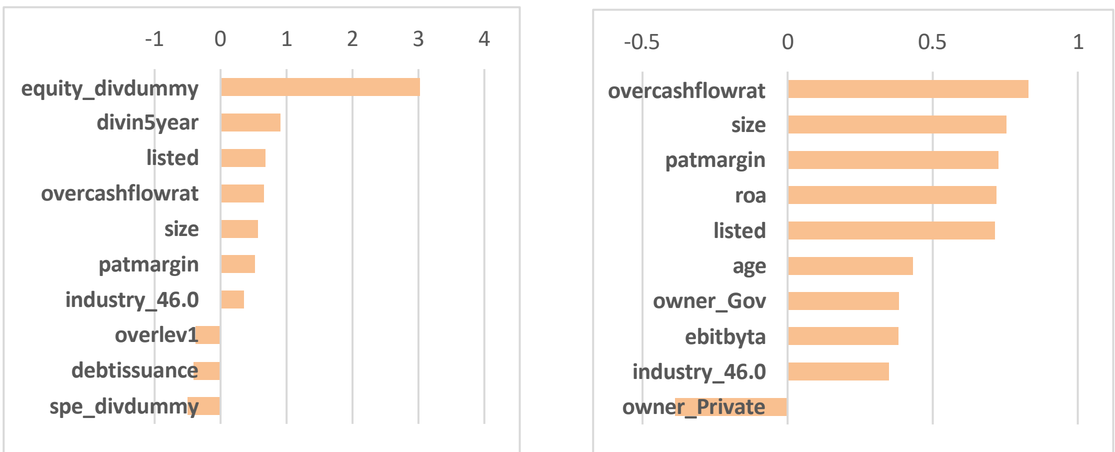


Figure 16: Feature importance from the logistic regression model for the equity dividend dummy

To predict F_EQUITY_DIVDUMMY, EQUITY_DIVDUMMY, and DIVIN5YEAR emerge as the most important features. Following these, OVERCASHFLOWRAT, LISTED, SIZE, PATMARGIN, and INDUSTRY 46.0 are some of the common important features across both models. We discuss the theoretical justification of important features of the best model in section 5.1.4.

5.1.2 Random Forest for Equity Dividend Prediction

Random forest Model performs best among all the tree-based models. There are four hyperparameters we tuned: maximum depth of trees, minimum samples per leaf, minimum sample required to split a node, and number of estimators(trees) used in this bagging method. The model, including dividend-related variables, gives the best performance at no limit on the maximum depth of the trees, a minimum of two samples required per leaf node, a minimum of two samples required to split an internal node, and 200 estimators (trees) in the forest. For the model excluding dividend-related variables with the 200 estimators(trees) in the algorithm, each leaf node requires at least one sample, splitting an internal node requires at least two samples, and no restriction for the depth of the trees gives the best accuracy.

For Random Forest with optimal hyperparameters, we achieved accuracies of 96.73%, including dividend-related variables, and 93.62% accuracy excluding them. It indicates a small reduction in accuracy when we exclude dividend-related variables.

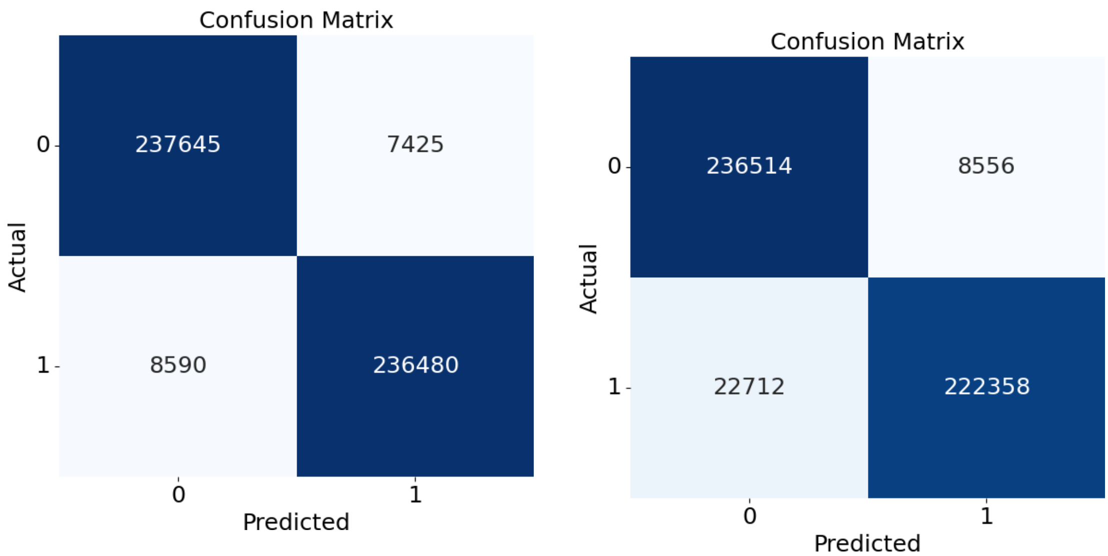


Figure 17: Confusion matrices from the Random Forest model for the equity dividend dummy

For the evaluation of all the performance metrics given in Table 6, we have analysed the confusion matrices of the predictions. Which also gives us a sight on Type 1 and Type 2 errors. By analysing the results in Table 6 and Fig. 17 jointly, for the model considering dividend-related variables, we observed that the Type 1 Error (False Positive) is 7,425 with the precision of 0.9696, whereas the Type 2 Error (False Negative) is 8,590 with the recall of 0.9649. This indicates precision is slightly better than the recall. We can expect to predict both classes with equal exactness.

The model excluding dividend-related variables, Type 1 Error is 8,556 with precision 0.9629, and Type 2 Error is 22,712 with recall 0.9073. It has higher precision than recall, this indicates the model predicts class 0 more accurately than class 1. In the previous model, We observed similar precision and recall in the previous model; however, in this model, we found a larger difference between precision and recall, with precision being larger than recall, making it unreliable for predicting Class 1 compared to Class 0.

The area under the ROC curve for the model is 0.97, which represents excellent discriminative power of the model. show in Figure 18 below.

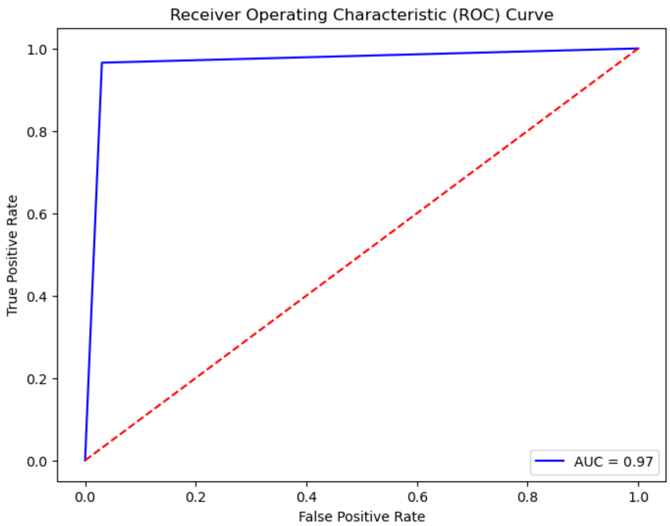
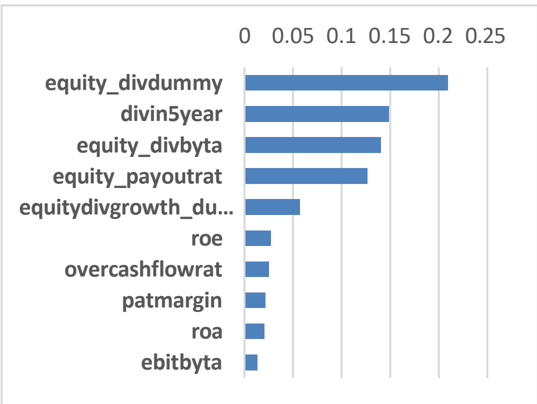


Figure 18: ROC curve from the Random Forest model for the equity dividend dummy

We presented feature importance for both models in Figure 19 below. Feature importance is determined by the entropy reduction at a split based on any feature.



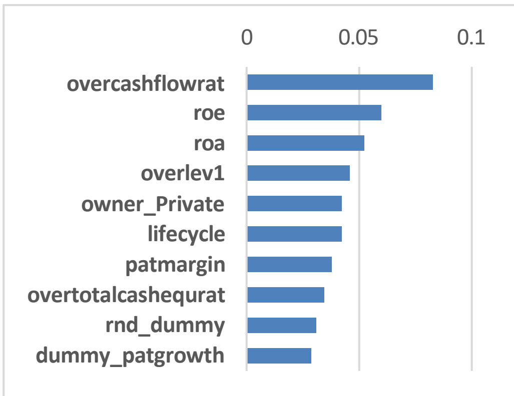


Figure 19: Feature importance from the Random Forest model for the equity dividend dummy

For predicting F_EQUITY_DIVDUMMY, EQUITY_DIVDUMMY, DIVIN5YEAR, EQUITY_DIVBYTA, EQUITY_PAYOUTRAT, and EQUITYDIVGROWTH_DUMMY are the most important dividend-related features. Meanwhile, OVERCASHFLOWRAT, ROE, ROA, and PATMARGIN are common valuable features across both models.

5.1.3 ANN for equity dividend prediction

We implement an ANN model with two hidden layers and tuned the hyperparameters using grid search, as discussed in Section 4.4. Additionally, we used 5-fold cross-validation to obtain a generalized model. The best model arrived after the model was run for 30 epochs for each set of hyperparameters, and the parameters of the ANN were updated every 32 passes, meaning we set the batch size to 32 data instances. We also designed the model with the condition that neurons in the forward architecture have fewer neurons than the previous layer, which is the suggested architecture for any ANN model. We have tuned four hyperparameters: Learning rate, which is the rate of updating the weights in ANN; number of neurons in the first layer, number of neurons in the second layer, and dropout rate, which is the proportion of neurons dropped at every epoch.

For the dataset including dividend-related variables, after performing hyperparameter tuning, we identified the best hyperparameters for our model as follows: Learning Rate = 0.001, Neurons in the first layer = 32, Neurons in the second layer = 8, and Dropout Rate = 0.2. For the case excluding dividend-related variables, we determined the best parameters to be as follows: Learning Rate = 0.001, Neurons in the first layer = 32, Neurons in the second layer = 16, and Dropout Rate = 0.2.

For the ANN model, including dividend-related variables, we achieved an accuracy of 93.11% and an F1 score of 93.09%. The model without dividend-related variables achieved an accuracy of 86.45% and an F1 score of 86.83%. This shows the reduction in accuracy after excluding dividend-related variables

The confusion matrices shown in the figure provide insights into Type 1 Error and Type 2 Error. Now we are jointly looking at Table 6 and Fig. 20. For the model that includes dividend-related variables, the Type 1 Error is found to be 16,347 with precision 0.933, and the Type 2 Error is 17,430 with recall 0.9289, suggesting the model has slightly lower recall than the precision. In contrast, the model that does not consider dividend-related variables has a Type 1 Error of 38,299 with precision 0.85, and a Type 2 Error of 28,108 with recall 0.8853. This indicates that this model has higher recall than precision. In contrast to the previous model, here we got a higher Recall, making it more reliable to predict class 1 than class 0.

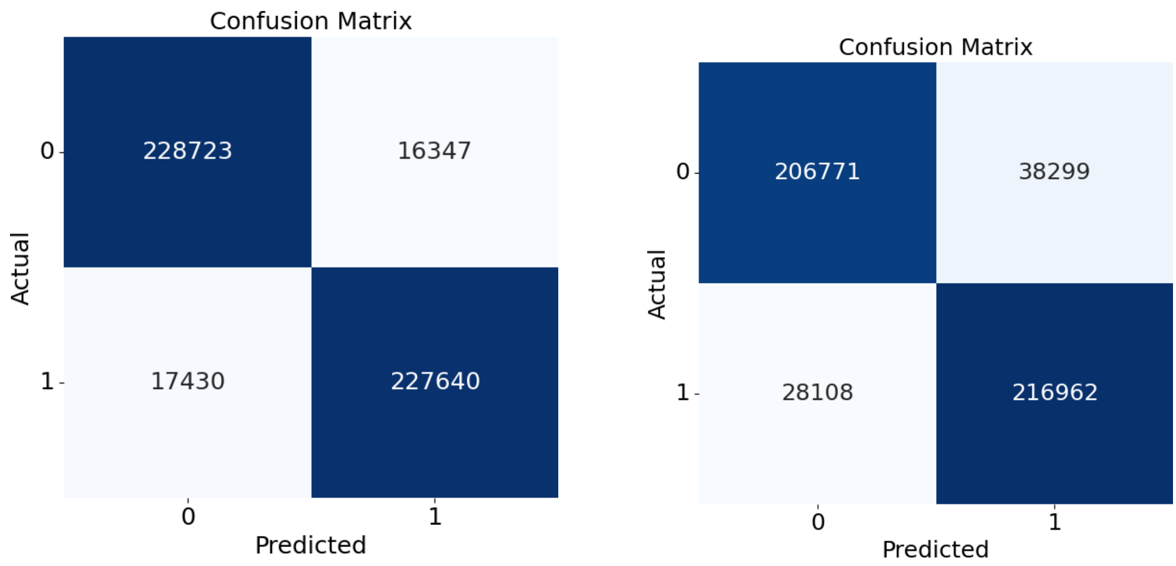


Figure 20: Confusion matrices from the ANN model for the equity dividend dummy

The area under the ROC curve for the model is 0.9778, which represents excellent discriminative power of the model, and it is also the highest among all the models.

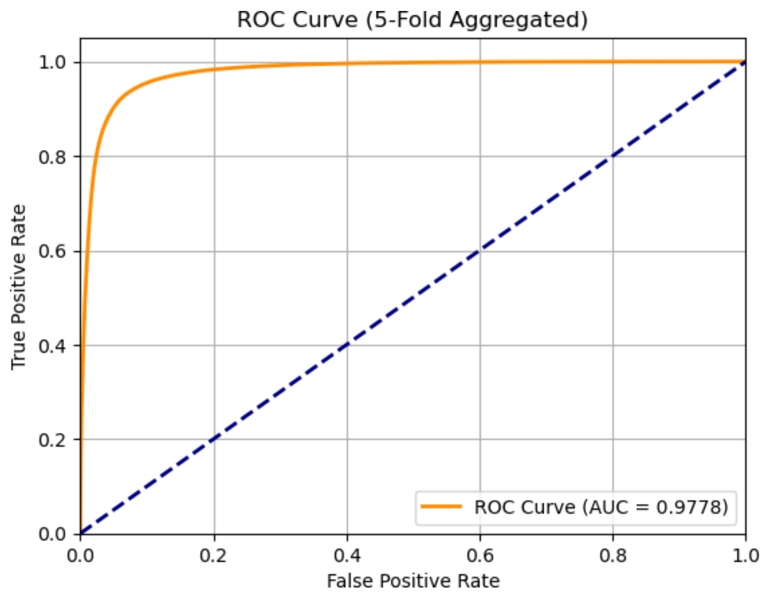


Figure 21: ROC curve from the ANN model for the equity dividend dummy

5.1.4 Discussion on Feature importance for equity dividend dummy

Among all the models used to predict equity dividends, the Random Forest model demonstrates superior performance, excelling across all parameters and providing the highest predictive power. Several important features influencing equity dividend decisions were identified. Firms that have consistently paid dividends over the past five years are more likely to continue paying dividends in the future. Conversely, firms that distributed substantial dividends in the most recent year and increased their dividend payouts last year are less likely to give dividend payments in the future. Additionally, firms with higher net profit margins, return on equity, return on assets, and higher EBITDA are more inclined to pay dividends in the future. The firm that generates higher cash flow and has a higher cash balance in the industry is more likely to pay dividends in the future. Firms spending money on research and development are more likely to pay dividends. Standalone private firms are less likely to pay dividends than government or private group firms.

5.2 Results for Special Dividend Dummy

Table 7 presents the performance outcomes for the 8 models mentioned earlier. The accuracy and F1 scores for each model are detailed under two configurations: one excluding dividend history variables and one including them. Each model underwent hyperparameter tuning to deliver the optimal performance for its respective algorithm category. The special dividend dummy, after SMOTE application, includes 5,64,620 data instances.

Among all the models, the tree-based models perform the best in terms of accuracy and F1 score for both setups. XGBoost stands out by delivering the highest accuracy and F1 score in both cases. With dividend history variables included, the accuracy and F1 score are both 99.90%. When we leave out dividend history variables, the results remain nearly identical, with an accuracy of 99.921% and an F1 score of 99.91%.

Table 7: Performance of each model for the special dividend dummy

F_SPE_DIVDUMMY	With Dividend-Related Variables		without Dividend-Related Variables	
	Accuracy	F1 score	Accuracy	F1 score
Statistical models				
Simple Logistic Regression	91.20%	91.35%	87.22%	87.50%
Logistic Regression with Lasso Regularisation	91.24%	91.42%	87.22%	87.50%
Logistic Regression with ridge	91.24%	91.42%	87.22%	87.51%
Logistic Regression with Elastic Net	91.24%	91.42%	87.22%	87.50%
Classical Machine learning models:				
Decision Tree	99.68%	99.68%	99.55%	99.55%
Random Forest	99.86%	99.86%	99.86%	99.86%
XGBoost	99.90%	99.90%	99.91%	99.91%
Deep Learning				
Artificial Neural Network	99.39%	99.40%	99.13%	99.14%

Following the tree-based models, the artificial neural network delivers the second-best accuracy and F1 score of 99.39% and 93.40%, respectively. Statistical models lag behind in performance. There is only a slight difference in how well the regularization methods for logistic regression (Lasso, Ridge, and Elastic Net) perform. Elastic Net shows a slight edge over the other two methods and the model without any regularization.

In Table 8, we provide additional performance matrices. The table focuses on the best-performing models from each category: Logistic Regression with Elastic Net, XGBoost, and Artificial Neural Network. Each of these three is discussed individually in the sections that follow.

Table 8: Performance summary of the best-performing models for the special dividend dummy

Special Dividend Dummy	With Dividend-Related variables			Without Dividend-Related Variables		
Performance metrics	Logistic regression	XGBoost	ANN	Logistic regression	XGBoost	ANN
Accuracy	0.9125	0.9990	0.9939	0.8722	0.9992	0.9913
Error Rate	0.0875	0.0010	0.0061	0.1278	0.0008	0.0087
Precision	0.8960	0.9995	0.9883	0.8560	0.9996	0.9837
Recall	0.9332	0.9985	0.9997	0.8950	0.9987	0.9992
F1 Score	0.9142	0.9990	0.9940	0.8750	0.9992	0.9914
Specificity	0.8917	0.9995	0.9881	0.8494	0.9996	0.9834
G-Mean	0.9122	0.9990	0.9939	0.8719	0.9992	0.9913
Balanced Accuracy	0.9125	0.9990	0.9939	0.8722	0.9992	0.9913
Youden's Gamma	0.8249	0.9980	0.9879	0.7444	0.9983	0.9826

5.2.1 Logistic Regression with Elasticnet for Special Dividend Prediction

The best mode is achieved after hyperparameter tuning. There are two hyperparameters, C and L1 Ratio. C represents the regularization strength. A higher value of C corresponds to lower regularization strength. The L1 ratio in Elastic Net is the proportion of L1 (Lasso) penalty relative to the combined L1 and L2 (Ridge) penalties. For the logistic regression model including dividend-related variables, the best parameters identified are C is 1 and L1 ratio is 0.5. For the model excluding dividend-related variables, the optimal parameters are C is 0.1 and L1 ratio is 0.7.

For logistic regression with elastic net, we achieved an accuracy of 91.25%, including dividend-history variables, and 87.22% when it was excluded. This indicates that the dividend-history variable significantly influences the prediction performance.

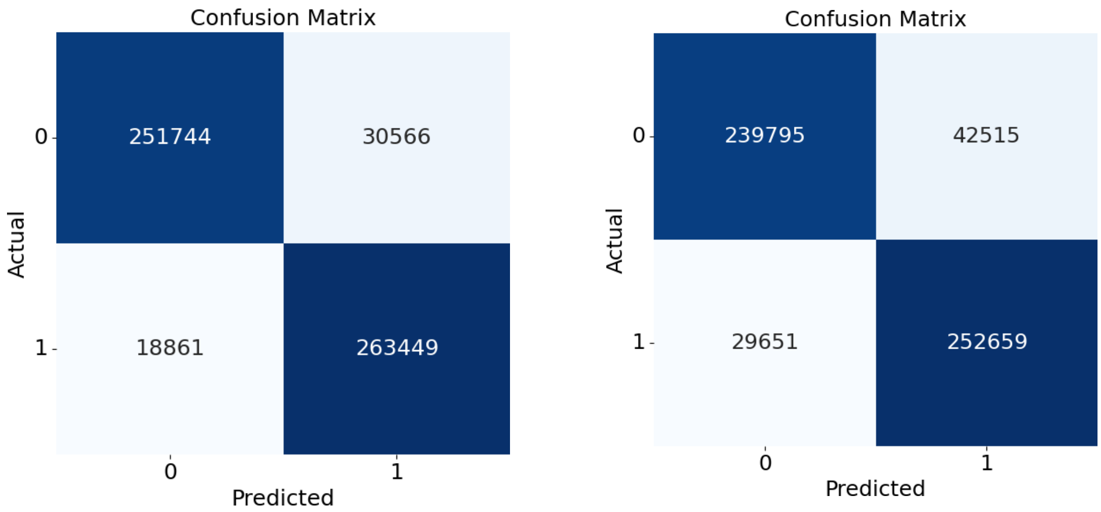


Figure 22: Confusion matrices from the logistic regression model for the special dividend dummy

To arrive at the performance metrics described in Table 8, we have examined the confusion matrices of the predictions, which directly give insights into Type 1 and Type 2 errors. The confusion matrix is presented in Fig. 22. We now jointly review the results from Table 8 and Fig. 22. For the model with dividend-related variables, we found that the Type 1 Error (False Positive) is 30566, with a precision of 0.896. Whereas the Type 2 Error (False Negative) is 18861, and the recall is 0.9332. This shows that the model predicts class 1 more effectively than class 0, with recall being higher than precision. The overall F1 score of the model is 0.9142.

Opposite results are observed for the model excluding dividend-related variables. It shows a Type I error of 42,515 (with a precision of 0.8560) and a Type II error of 41,014 (with a recall of 0.8950). This model performs much worse than the previous one regarding overall accuracy. For both logistic regression models, we observed that recall is better than precision, which is a desirable result in the case of an imbalanced data set.

The area under the ROC curve suggests higher discriminative power, and the ROC curve is shown in the figure below. The AUC-ROC score of 0.97, showcasing the discriminative power of the model.

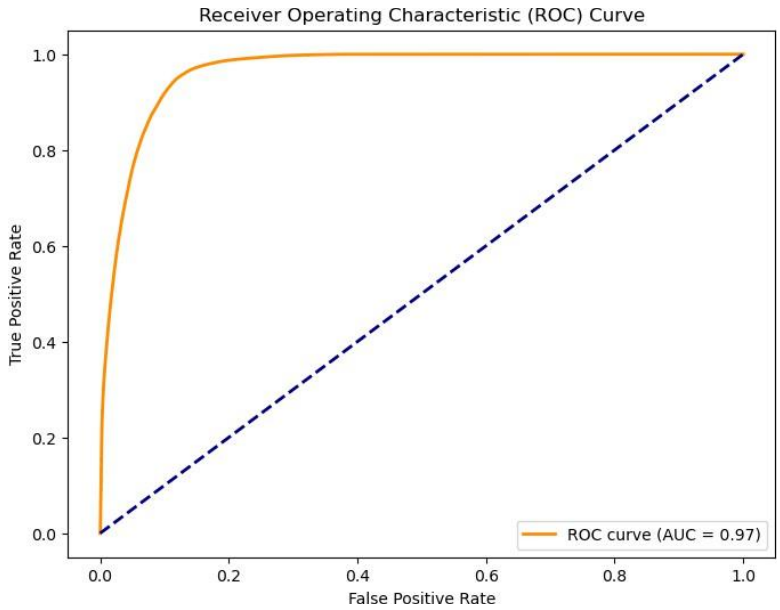


Figure 23: ROC curve from the logistic regression model for the special dividend dummy

Furthermore, Figure 12 below illustrates the feature relevance of the top factors. The coefficient values of every variable in the trained model indicate the significance of a feature.

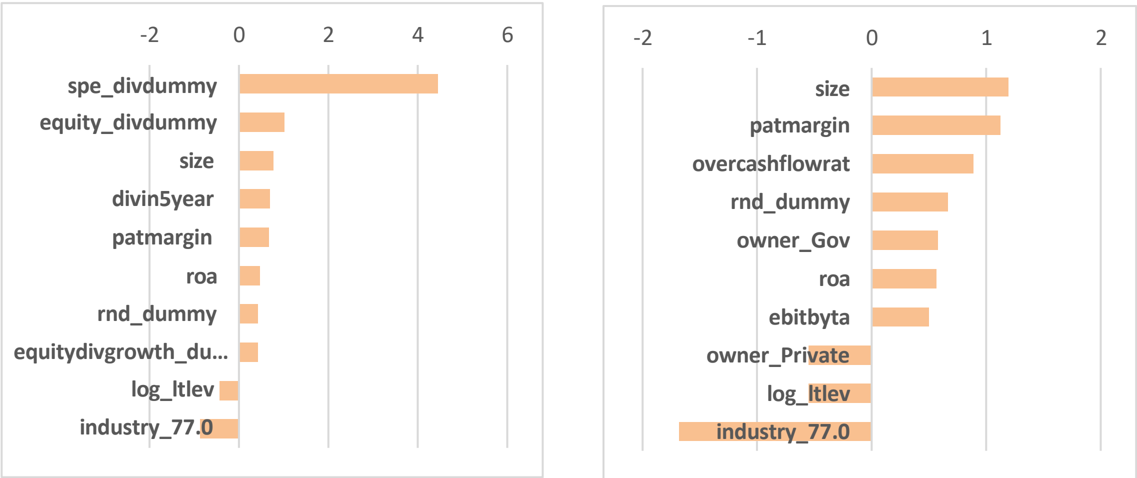


Figure 24: Feature importance from the logistic regression model for the special dividend dummy

For predicting the F_SPE_DIVDUMMY, our feature SPE_DIVDUMMY plays a significant role, and the next important feature is DIVIN5YEAR, suggesting that dividend-related variables have a significant contribution to the model's performance. Additionally, SIZE, PATMARGIN, OVERCASHFLOWRAT, RND_DUMMY, LOG_LTLEV, and INDUSTRY_77.0 are common important features among both models, indicating consistency between these models.

5.2.2 XGboost for special Dividend Prediction

XGBoost Model performs best among all the tree-based models. There are three hyperparameters we tuned: maximum depth of trees, learning rate, and number of estimators(trees) used in this boosting algorithm. Both model, including and excluding dividend-related variables, gives the best performance at a maximum depth of the trees is 10, a learning rate of 0.2, and 200 estimators (trees) in the forest.

For XGBoost with optimal hyperparameters, we achieved accuracies of 99.90%, including dividend-related variables, and 99.91% accuracy excluding them. It indicates no such performance difference for both specifications.

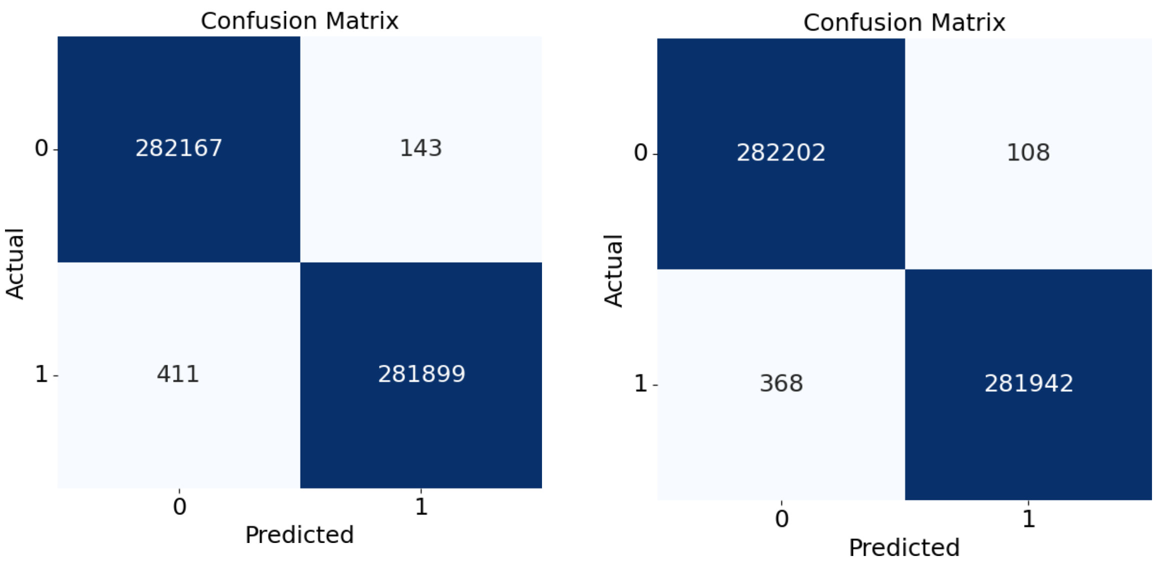


Figure 25: Confusion matrices from the XGBoost model for the special dividend dummy

For the evaluation of all the performance metrics given in Table 8, we have analyzed the confusion matrices of the predictions, which also provide insights into Type 1 and Type 2 errors. Looking at the results in Table 8 and Fig. 25 jointly, for the model considering dividend-related variables, we observed that the Type 1 Error (False Positive) is 143 with a precision of 0.9995, whereas the Type 2 Error (False Negative) is 411 with a recall of 0.9985. Both precision and recall are extremely high, but from the Type 1 and Type 2 errors, it is evident that the model predicts class 0 better than class 1. This suggests that, unlike other models for the special dividend dummy, precision slightly outweighs recall for both specifications.

The model excluding dividend-related variables has a Type 1 Error of 108 with a precision of 0.9996 and a Type 2 Error of 368 with a recall of 0.9987. Compared to the previous model, we observe similar precision and recall values; however, the Type 1 Error is notably lower than the Type 2 Error. This suggests slightly better precision than recall for this model.

The AUC ROC score of the model is extremely close to 1. This suggests excellent discriminative power. The curve is shown in the figure below.

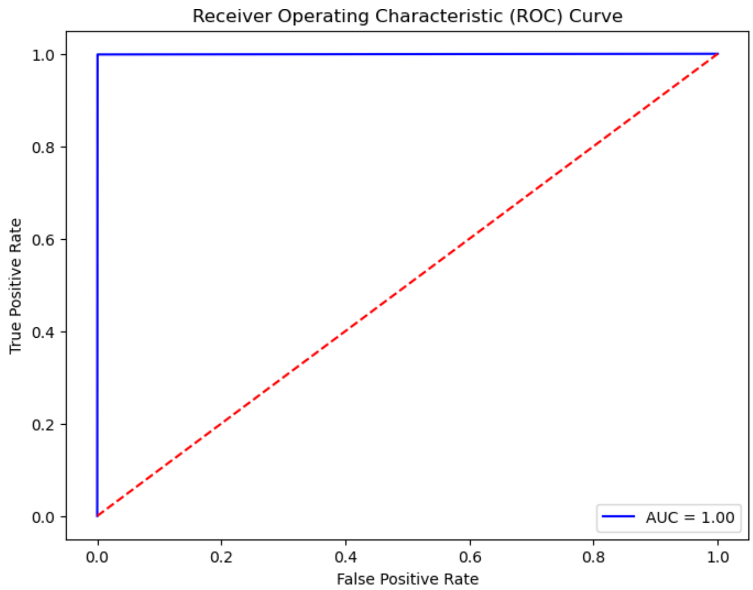


Figure 26: ROC curve from the XGBoost model for the special dividend dummy

We present feature importance for both models in Figure 27 below. Feature importance is determined by the entropy reduction at a split based on any feature.

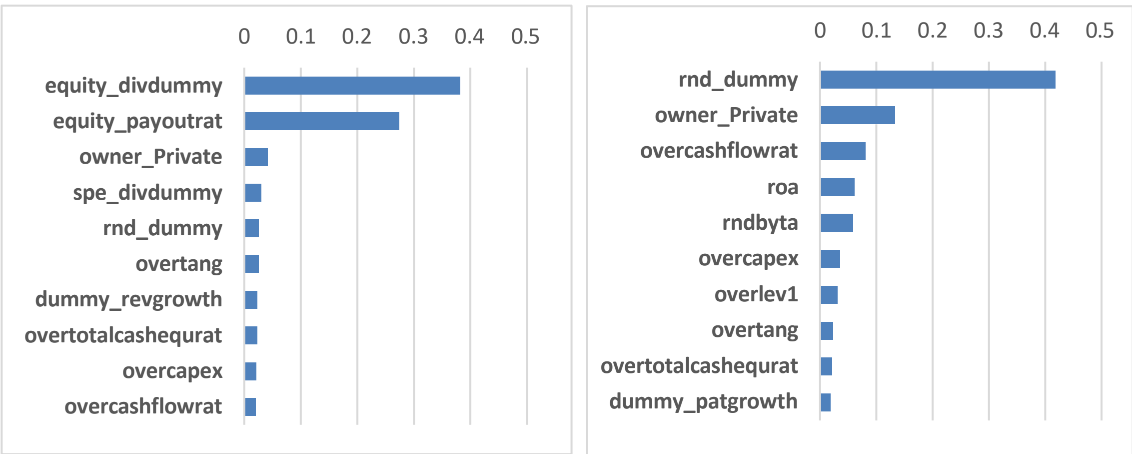


Figure 27: Feature importance from the XGBoost model for the special dividend dummy

To predict F_SPE_DIVDUMMY, EQUITY_DIVDUMMY, and EQUITY_PAYOUTRAT are the most important features. Following these, features like OWNER_PRIVATE, RND_DUMMY, OVERTOTALCASHEQRAT, OVERTANG, and OVERCAPEX are consistent across both models.

5.2.3 ANN for equity dividend prediction

As discussed in Section 4.4, we used grid search to adjust the hyperparameters of an ANN model with two hidden layers. In addition, we obtained a generalized model by using 5-fold cross-validation. We set the batch size to 32 data instances and trained the model for 30 epochs for each set of hyperparameters, updating the ANN's parameters every 32 passes. In addition, as is recommended for any ANN model, we built the model with the condition that neurons in the forward architecture have fewer neurons than the layer before it. We have tuned four hyperparameters: Learning rate, which is the rate of updating the weights in ANN; number of neurons in the first layer, number of neurons in the second layer, and dropout rate, which is the proportion of neurons dropped at every epoch.

After performing hyperparameter tuning, for the dataset including dividend-related variables, we identified the best hyperparameters for our model as follows: Learning Rate is 0.001, 32 Neurons in the first layer, 16 Neurons in the second layer, and Dropout Rate is 0.2. For the case excluding dividend-related variables, we determined the best parameters to be as follows: Learning Rate is 0.001, 32 Neurons in the first layer, 16 Neurons in the second layer, and Dropout Rate is 0.2.

For the model that includes dividend-related variables, we achieved an accuracy of 99.39% and an F1 score of 99.40%. In contrast, the model without dividend-related variables achieved an accuracy of 99.13% and an F1 score of 99.14%. These results are lower than those of the previous model, highlighting the importance of dividend-related variables in predicting special dividend policies using an Artificial Neural Network.

The confusion matrices shown in Fig. 28 provide insights into Type 1 Error and Type 2 Error. Referring to the performance metrics in Table 8, for the model that includes dividend-related variables, the Type 1 Error is 3,354 with a precision of 0.9883, and the Type 2 Error is 76 with a recall of 0.9997. Conversely, the model that excludes dividend-related variables has a Type 1 Error of 4,687 with a precision of 0.9837 and a Type 2 Error of 220 with a recall of 0.9992. Both models demonstrate very high recall rates compared to precision, indicating strong effectiveness in classifying class 1 instances but reduced accuracy in identifying class 0 instances.

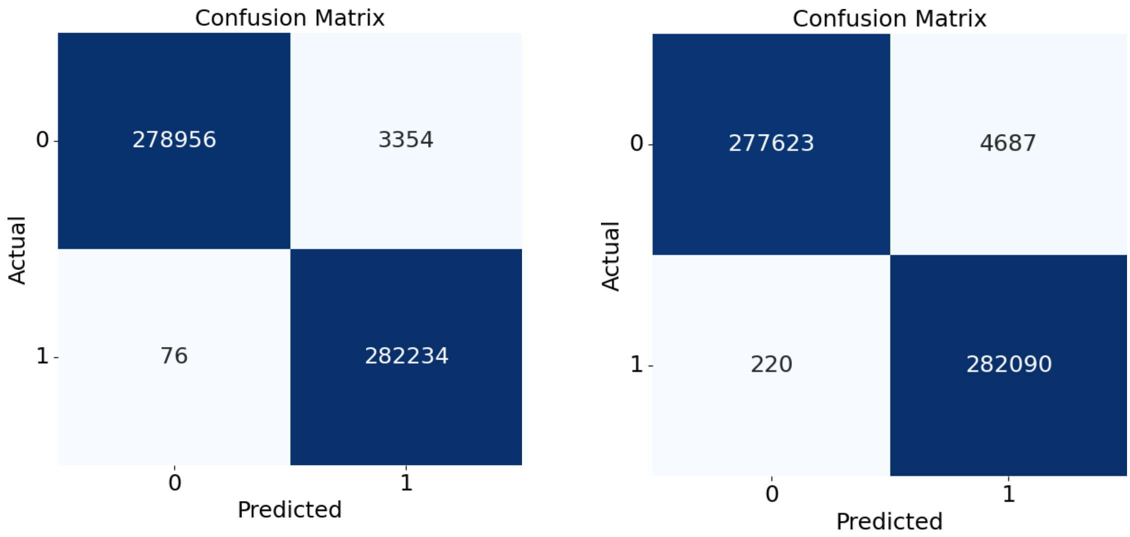


Figure 28: Confusion matrices from the ANN model for the special dividend dummy

The area under the AUC ROC curve is also very high for the ANN model. The curve is shown in Figure 29 below.

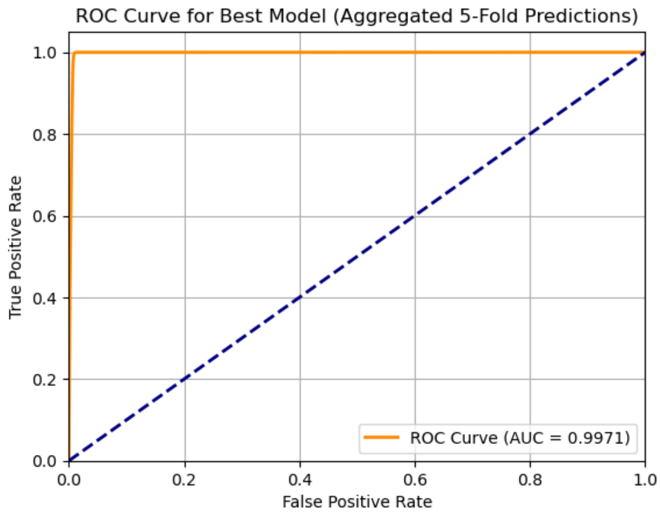


Figure 29: ROC curve from the ANN model for the special dividend dummy

5.2.4 Discussion on Feature importance for equity dividend dummy

Among the various models used to predict special dividends, XGBoost demonstrates superior performance, excelling across all parameters and providing the highest predictive power. Insights from the XGBoost model reveal several key patterns. Firms that paid dividends in the previous year are more likely to distribute special dividends. Similarly, firms that issued special dividends in the past year exhibit a higher likelihood of continuing to do so in the future. Conversely, firms that distributed exceptionally high dividends in the previous year are less likely to issue special dividends moving forward. Furthermore, privately-owned firms exhibit a lower propensity to pay special dividends. Firms investing in research and development are likely to issue special dividends. Additionally, firms with cash balances exceeding industry averages and those generating extraordinary cash flows relative to their peers are more likely to distribute special dividends. Lastly, firms experiencing revenue growth compared to the previous year are more likely to issue special dividends in the future. Asset-heavy firms in an industry are more likely to pay special dividends. Firms with higher capital expenditures compared to the industry median are also more likely to pay special dividends.

5.3 Results: Equity Dividend Growth Dummy

Table 9 illustrates the results of the 8 models we discussed earlier. It highlights the accuracy and F1 scores under two settings: one that excludes dividend history variables and one that incorporates them. All models have been fine-tuned with hyperparameter optimization to ensure the best results for each algorithm type. After applying SMOTE, the equity dividend growth dummy comprises 5,24,092 data instances.

The tree-based models provide the best accuracy and F1 score across both setups. In particular, the random forest method achieves the highest accuracy and F1 score for both cases. With dividend history variables included, the accuracy and F1 score are 95.12% and 95.01%. Even without dividend history variables, the accuracy and F1 score remain almost the same at 95.52% and 95.39%.

Table 9: Performance of each model for the equity dividend growth dummy

F_EQUITYDIVGROWTH_DUMMY	With Dividend-Related Variables		without Dividend-Related Variables	
	Accuracy	F1 score	Accuracy	F1 score
Statistical models				
Simple Logistic Regression	83.94%	83.79%	79.31%	79.85%
Logistic Regression with Lasso Regularisation	83.92%	83.70%	79.30%	79.85%
Logistic Regression with Ridge	83.94%	83.79%	79.31%	79.85%
Logistic Regression with Elastic Net	83.96%	83.80%	79.31%	79.86%
Classical Machine learning models:				
Decision Tree	93.71%	93.60%	92.67%	92.57%
Random Forest	95.12%	95.01%	95.52%	95.39%
XGBoost	94.65%	94.51%	94.91%	94.75%
Deep Learning				
Artificial Neural Network	87.82%	87.77	85.58%	86.01%

The artificial neural network comes next to the tree-based models, providing an accuracy and F1 score of 87.82% and 87.77%, respectively. Statistical models perform the least effectively. The performance difference among the regularization techniques for logistic regression (Lasso, Ridge, and Elastic Net) is very small. Elastic Net does slightly better compared to the others and the model without regularization.

Table 10 includes additional performance matrices. It highlights the performance of the top models from each category, namely Logistic Regression with Elastic Net, Random Forest, and Artificial Neural Network. We discuss these three models separately in the subsequent sections.

Table 10: Performance summary of the best-performing models for the equity dividend growth dummy

Dividend Growth Dummy	With Dividend-Related variables			Without Dividend-Related Variables		
Performance metrics	Logistic regression	Random Forest	ANN	Logistic regression	Random Forest	ANN
Accuracy	0.8396	0.9512	0.8782	0.7931	0.9552	0.8558
Error Rate	0.1604	0.0488	0.1218	0.2069	0.0448	0.1442
Precision	0.8463	0.9728	0.8458	0.7781	0.9817	0.8353
Recall	0.8298	0.9284	0.9122	0.8202	0.9276	0.8864
F1 Score	0.8380	0.9501	0.8777	0.7986	0.9539	0.8601
Specificity	0.8493	0.9740	0.8469	0.7661	0.9827	0.8252
G-Mean	0.8395	0.9509	0.8789	0.7927	0.9548	0.8553
Balanced Accuracy	0.8396	0.9512	0.8795	0.7931	0.9552	0.8558
Youden's Gamma	0.6791	0.9024	0.7591	0.5862	0.9103	0.7116

5.3.1 Logistic Regression with Elastic Net for Equity Dividend Growth Prediction

The best mode is to arrive after hyperparameter tuning. There are two hyperparameters, C and L1 Ratio. C represents the regularization strength. A higher value of C corresponds to lower regularization strength. The L1 ratio in Elastic Net is the proportion of L1 (Lasso) penalty relative to the combined L1 and L2 (Ridge) penalties. For the logistic regression model including dividend-history variables, the best parameters identified are C = 0.01 and L1 ratio = 0.7. For the model excluding dividend-related variables, the optimal parameters are C = 10 and L1 ratio = 0.3. This combination effectively prevents overfitting while ensuring good generalization to unseen data.

For logistic regression with elastic net, we achieved an accuracy of 83.96%, including dividend-history variables, and 79.31% when it was excluded. This indicates that the dividend-history variable significantly influences the prediction performance.

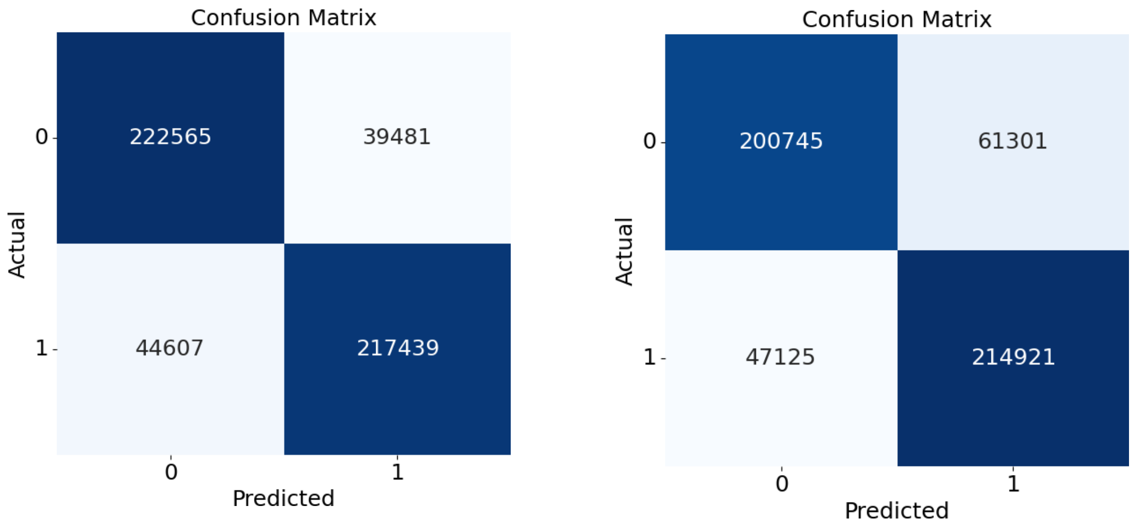


Figure 30: Confusion matrices from the logistic regression model for the equity dividend growth dummy

To arrive at the performance metrics described in Table 10, we analyze the confusion matrices of the predictions, which directly reflect Type 1 and Type 2 errors. The confusion matrix appears in Fig. 30. We now look at the results from Table 10 and Fig. 30 together. For the model considering dividend-related variables, we noted that the Type 1 Error (False Positive) is 39481 and the precision is 0.8463. Whereas the Type 2 Error (False Negative) is 44607, and the recall is 0.8298. Precision and recall are well balanced, and the overall F1 score of the model is 0.8380.

Opposite results are observed for the model excluding dividend-related variables. It displays a Type I error of 61,301 (with a precision of 0.7781) and a Type II error of 47,125 (with a recall of 0.8202). This model performs worse than the earlier model in terms of overall accuracy. Nevertheless, due to the imbalance in the dataset, where positive class observations are limited, recall takes precedence over precision. Given that the model without dividend-related variables achieves superior recall in relation to precision, it is deemed more appropriate in this situation than the model with dividend-related variables.

Further, we plot the ROC curve of the model with dividend-related variables in Fig. 31. The Area under the ROC curve is 0.91. This implies a slightly discriminative power of the model

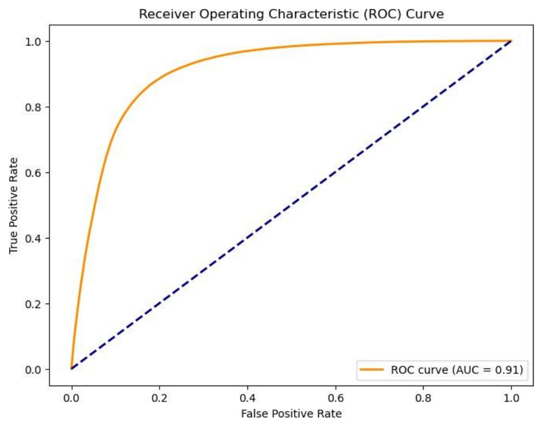


Figure 31: ROC curve from the logistic regression model for the equity dividend growth dummy

Furthermore, we present the feature importance of the top variables in Figure 32 below. Feature importance is determined by the coefficient values of each variable in the trained model.

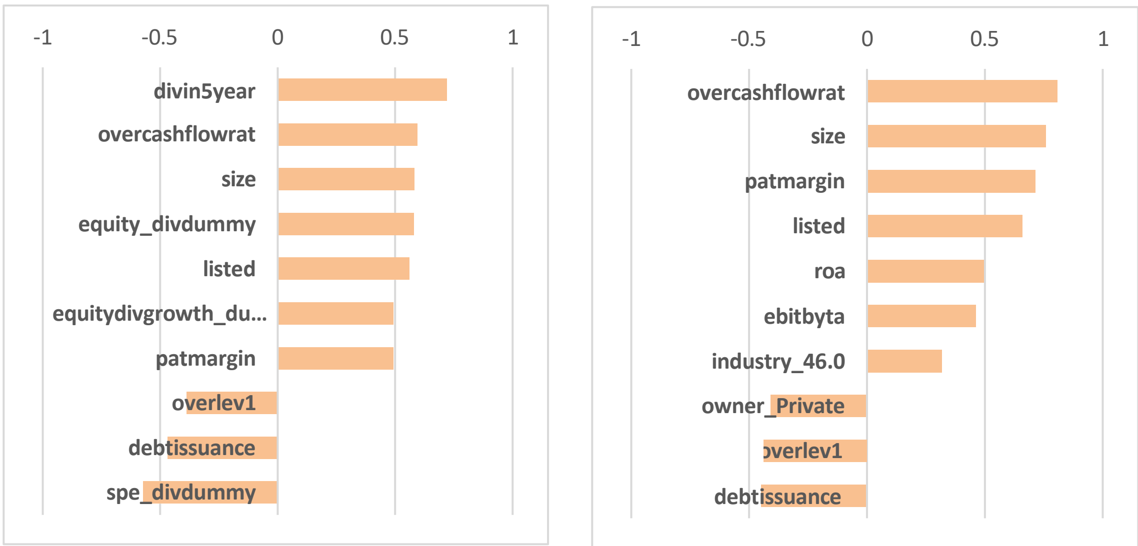


Figure 32: Feature importance from the logistic regression model for the equity dividend growth dummy

To predict a F_EQUITYDIVGROWTH_DUMMY, DIVIN5YEAR, EQUITY_DIVDUMMY, EQUITYDIVGROWTH_DUMMY emerges as the most important dividend-related variable. Following this, features like OVERCASHFLOWRAT, SIZE, LISTED, PATMARGIN, OVERLEV1, and DEBTISSUANCE are commonly important features across both models

5.3.2 Random Forest for Equity Dividend Growth Prediction

Random forest Model performs best among all the tree-based models. There are four hyperparameters we tuned: maximum depth of trees, minimum samples per leaf, minimum sample required to split a node, and number of estimators(trees) used in this bagging method. The model, including dividend-related variables, gives the best performance at no limit on the maximum depth of the trees, a minimum of one sample required per leaf node, a minimum of two samples required to split an internal node, and 200 estimators (trees) in the forest. For the model excluding dividend-related variables with the 200 estimators(trees) in the algorithm, each leaf node requires at least one sample, splitting an internal node requires at least two samples, and no restriction for the depth of the trees gives the best accuracy.

For Random Forest with optimal hyperparameters, for the model including dividend-related variables, we achieved accuracy and F1 score of 95.12% and 95.01%, respectively, while for the model excluding dividend-related variables, accuracy and F1 scores are 95.52% and 95.39%, respectively. Suggesting no difference in performance due to dividend-related variables.

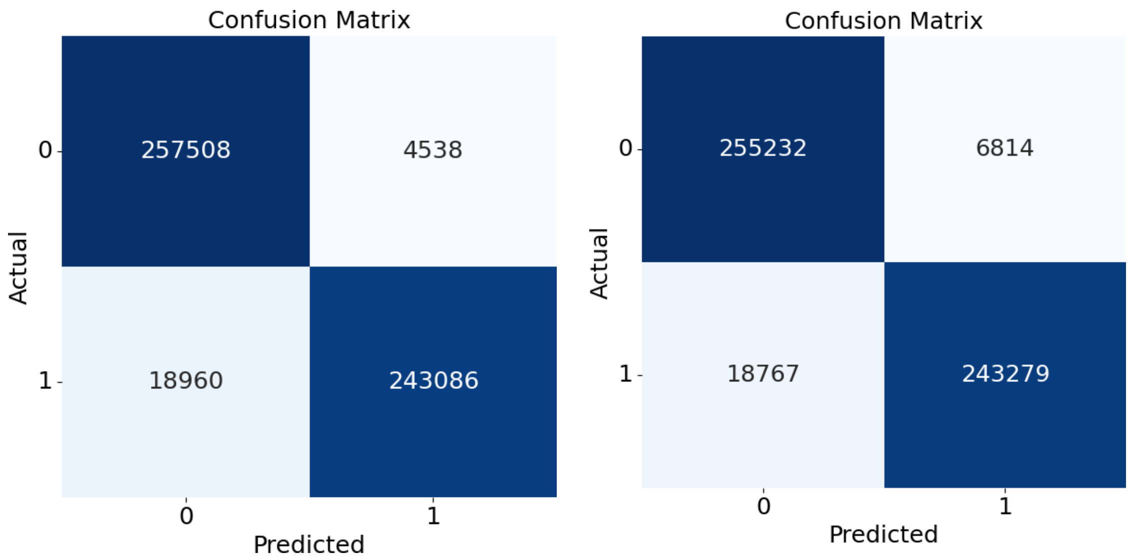


Figure 33: Confusion matrices from the Random Forest model for the equity dividend growth dummy

For the evaluation of all the performance metrics given in Table 10, we have analysed the confusion matrices of the predictions, which also provide insights into Type 1 and Type 2 errors. Considering the results in Table 10 and Fig. 33 jointly, for the model considering dividend-related variables, we observed that the Type 1 Error (False Positive) is 4,538 with a precision of 0.9728, whereas the Type 2 Error (False Negative) is 18,960 with a recall of 0.9817. This indicates that precision slightly exceeds recall.

The model excluding dividend-related variables reports a Type 1 Error of 6,814 with a precision of 0.9817 and a Type 2 Error of 18,767 with a recall of 0.9276. Precision is higher than recall, indicating that the model predicts class 0 more effectively than class 1. Although the previous model showed comparable precision and recall, this model highlights a more pronounced difference between the two, making it less suitable for accurately predicting Class 1.

The area under the ROC curve for the model is 0.95, which represents excellent discriminative power of the model. show in Figure 34 below.

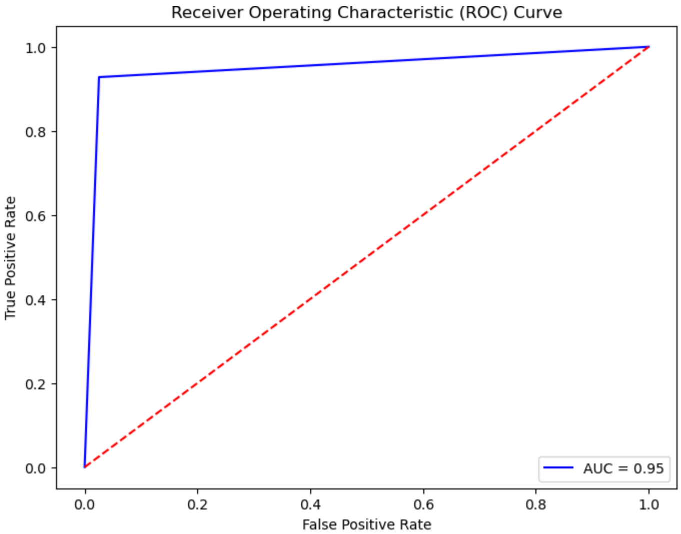
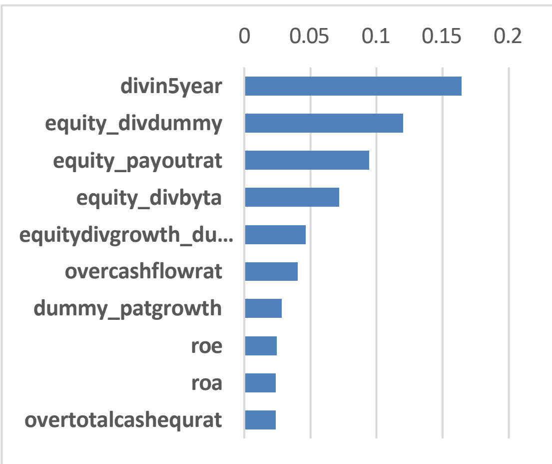


Figure 34: ROC curve from the Random Forest model for the equity dividend growth dummy

We presented feature importance for both models in Figure 19 below. Feature importance is determined by the entropy reduction at a split based on any feature.



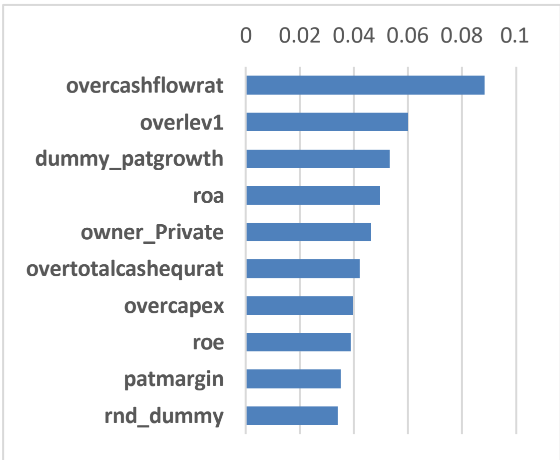


Figure 35: Feature importance from the Random Forest model for the equity dividend growth dummy

For predicting F_EQUITYDIVGROWTH_DUMMY, DIVIN5YEAR, EQUITY_DIVDUMMY, EQUITY_DIVBYTA, and EQUITYDIVGROWTH_DUMMY are the most important dividend-related features. Additionally, OVERCASHFLOWRAT, ROA, ROE, DUMMY_PATGROWTH, and OVERTOTALCASHEQRAT are the common important features across both models.

5.3.3 ANN for Equity Dividend Growth Prediction

We employe grid search to modify the hyperparameters of a two-layered ANN model. we used 5-fold cross-validation to develop a generalized model. For each set of hyperparameters, as explained in Section 4.4. Additionally, we trained the model for 30 epochs, changing the ANN's parameters every 32 passes, with a batch size of 32 data examples. Furthermore, we constructed the model with the stipulation that neurons in the forward architecture have fewer neurons than the layer preceding it, as is advised for any ANN model. We have tuned four hyperparameters: Learning rate, which is the rate of updating the weights in ANN; number of neurons in the first layer, number of neurons in the second layer, and dropout rate, which is the proportion of neurons dropped at every epoch.

For the dataset including dividend-related variables, after performing hyperparameter tuning, we identified the best hyperparameters for our model as follows: Learning Rate is 0.001,32 Neurons in the first layer, 16 Neurons in the second layer, and Dropout Rate is 0.2. For the case excluding dividend-related variables, we determined the best parameters to be as follows: Learning Rate is 0.001, 32 Neurons in the first layer, 16 Neurons in the second layer, and Dropout Rate is 0.2.

For the ANN model, including dividend-related variables, we achieved an accuracy of 87.82% and an F1 score of 87.77%. The model without dividend-related variables achieved an accuracy of 85.58% and an F1 score of 86.01%. this shows the reduction in accuracy after excluding dividend-related variables

The confusion matrices in Fig. 36 provide a clear understanding of Type 1 Error and Type 2 Error. Analyzing Fig. 36 and Table 10 together, for the model that includes dividend-related variables, the Type 1 Error is 16,328 with a precision of 0.8458, and the Type 2 Error is 17,550 with a recall of 0.8864. This indicates the model has a higher recall than precision. On the other hand, the model that excludes dividend-related variables shows a Type 1 Error of 45,805 with a precision of 0.8353 and a Type 2 Error of 29,760 with a recall of 0.8864. Both models exhibit similar trends in precision and recall, with recall being higher. This makes recall a more valuable metric for evaluating models on imbalanced datasets.

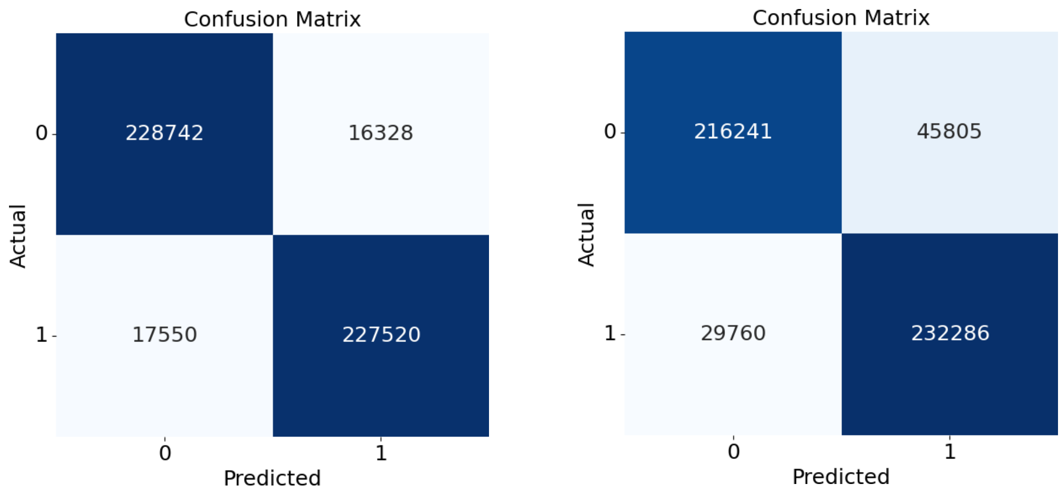


Figure 36: Confusion matrices from the ANN model for the equity dividend growth dummy

The ROC curve is shown in Figure 40 below. Area under ROC curve is also high, 0.9476, suggesting good classification power.

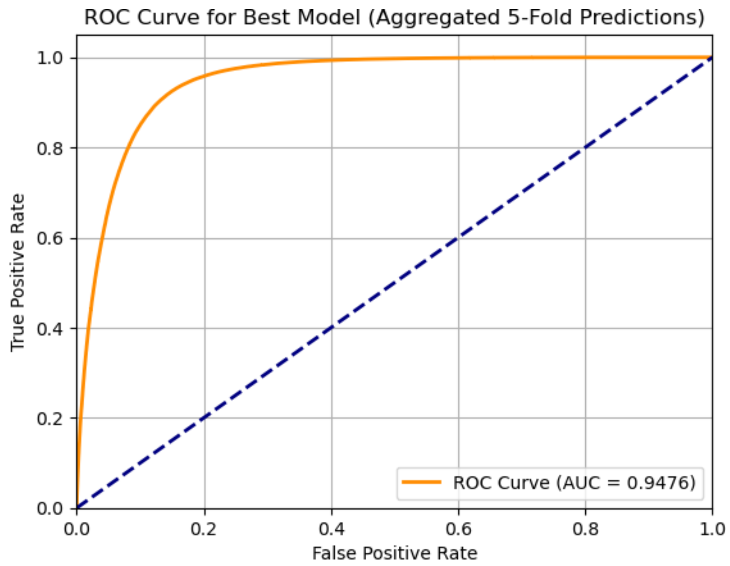


Figure 37: ROC curve from the ANN model for the equity dividend growth dummy

5.3.4 Discussion on feature importance for equity dividend growth dummy

The Random Forest model demonstrates the best performance, allowing us to identify key features influencing dividend growth decisions. The model indicates that the frequency of dividend payments by a firm over the past five years significantly impacts its likelihood of increasing dividends. Specifically, if a firm distributed dividends in the previous year and increased its dividend during that period, it is more likely to continue raising dividends in the future. Conversely, firms that paid substantial dividends in the last year are less likely to further increase them. Additionally, firms with excess cash balances compared to industry averages are more inclined to raise dividends. Similarly, firms generating extraordinary cash flows relative to industry standards are more likely to increase their dividends. Firms that have exhibited growth in net profits are also more likely to enhance their dividend payments. Lastly, profitable firms with higher returns on equity and assets demonstrate a greater propensity for increasing dividends.

6 Conclusion

In this paper, we aim to provide solutions for three classification problems: whether a company will pay a regular dividend, whether it will offer a special dividend and whether there will be growth in the dividends compared to last year. We are able to predict regular dividends, special dividends and growth in dividends with accuracies of 96.73%, 99.90% and 95.12% respectively. For equity dividends, we have the highest prediction accuracy among all published works. Further, as per best of our knowledge, we are the first ones to predict special dividends and dividends growth. Further, we do not use any market-related features and hence our models can be used for unlisted firms as well.

We find that tree-based models work best for all three prediction problems. Random Forests were the best to predict equity dividends and dividend growth, while XGBoost worked best for predicting special dividends. This is primarily due to the working principles of these models, which make them well-suited for handling non-linear data. Since many features in the dataset do not follow a standard distribution and include numerous binary variables, Logistic Regression, which assumes a linear relationship between features and the target variable, tends to perform poorly in comparison. Tree-based models, on the other hand, work on the principle of entropy reduction. They iteratively split the data to minimize entropy and maximize information gain, making them better suited for the non-linear relationships and diverse feature types present in the dataset. This approach aligns well with the objectives of the study, as it improves the model's predictive accuracy.

We developed two separate specifications: one including dividend history variables and the other excluding them. Theories suggest that dividends are highly autocorrelated and hence simple time series models might be suffici9ent to predict them. But we find that other non-dividend history features have a significant predictive power. From the results, we observed that if we build models with only non-dividend history variables, maximum loss in accuracy is just 10% compared to models with dividend history variables. Further, in many cases there was no significant difference in performance across the two specifications. This implies that machine learning models are better in predicting dividends than time series models even though there is significant autocorrelation. Additionally, the consistency in feature importance rankings across all machine learning models in both specifications for all target variables highlights the robustness of our findings.

Our models leverage financial statement data, which is readily available for most Indian companies. This accessibility enables seamless deployment by investors. Consequently, the models can support more informed decision-making in stock selection and trading. Furthermore, they can assist investment advisors in better addressing client requirements and aid fund managers in selecting securities in alignment with their investment mandates.

References

Abdullah, H. (2022). Profitability and leverage as determinants of dividend policy: Evidence of Turkish financial firms. MPRA Paper No. 112134.

Abor, J., & Fiador, V. (2013). Does corporate governance explain dividend policy in Sub-Saharan Africa? International Journal of Law and Management, 55(3), 201–225. https://doi.org/10.1108/17542431311327637

Abualdenien, J., & Borrmann, A. (2022). Ensemble-learning approach for the classification of levels of geometry (LOG) of building elements. Advanced Engineering Informatics, 51, 101497. https://doi.org/10.1016/j.aei.2021.101497

Adaoglu, C. (2000). Instability in the dividend policy of the Istanbul Stock Exchange (ISE) corporations: Evidence from an emerging market. Emerging Markets Review, 1, 252–270.

Agrawal, A., & Jayaraman, N. (1994). The dividend policies of all-equity firms: A direct test of the free cash flow theory. Managerial and Decision Economics, 15, 139–148.

Aivazian, V., Booth, L., & Cleary, S. (2003). Do emerging market firms follow different dividend policies from U.S. firms? The Journal of Financial Research, 26, 371–387.

Ali, H. (2021). Corporate dividend policy in the time of COVID-19: Evidence from the G-12 countries. Finance Research Letters, 46, 102493. https://doi.org/10.1016/j.frl.2021.102493

Allegrini, M., & Greco, G. (2011). Corporate boards, audit committees and voluntary disclosure: Evidence from Italian listed companies. Journal of Management & Governance, 17, 187–216. https://doi.org/10.1007/s10997-011-9168-3

Al-Najjar, B., & Hussainey, K. (2009). The association between dividend payout and outside directorships. Journal of Applied Accounting Research, 10(1), 4–19. https://doi.org/10.1108/09675420910963360

Ang, J. S., & Peterson, P. P. (1984). The leasing puzzle. Journal of Finance, 39, 1055–1065.

Barclay, M. J., Smith, C. W., & Watts, R. L. (1995). The determinants of corporate leverage and dividend policies. Journal of Applied Corporate Finance.

Bhalerao, M. (Ed.). (2024). Shareholders got record-high dividends in FY24-Will the trend continue? Livemint. https://www.livemint.com/companies/news/shareholders-got-record-high-dividends-in-fy24-will-the-trend-continue-11718878404106.html

Bhat, R. (2022). Predicting Dividend Omission Behaviour of Indian Firms Using Machine Learning Algorithms. Quarterly Journal of Finance. https://www.researchgate.net/publication/339324945_Predicting_Dividend_Omission_Behaviour_of_Indian_Firms_Using_Machine_Learning_Algorithms

Blas. (2019, July 3). The Intelligent Investor by Benjamin Graham. The Rabbit Hole. https://blas.com/intelligent-investor/

Borokhovich, K. A., Brunarski, K. R., Harman, Y., & Kehr, J. B. (2005). Dividends, corporate monitors and agency costs. Financial Review, 40(1), 37–65. https://doi.org/10.1111/j.0732-8516.2005.00092.x

Brennan, M. J. (1970). Taxes, market valuation and corporate financial policy. National Tax Journal, 23(4), 417–427. https://doi.org/10.1086/ntj41792223

Carter, D. A., Simkins, B. J., & Simpson, W. G. (2003). Corporate governance, board diversity, and firm value. Financial Review, 38(1), 33–53. https://doi.org/10.1111/1540-6288.00034

Chen, Y. (2010). Agency issues and the effects of corporate governance mechanisms on agency costs in Chinese listed companies (PhD thesis, Cardiff University). http://orca.cf.ac.uk/54376/

Codementor. (2020). Visualizing decision trees with Python (Scikit-learn, Graphviz, Matplotlib). https://www.codementor.io/@mgalarny/visualizing-decision-trees-with-python-scikit-learn-graphviz-matplotlib-154mszcto7

Collins, M. C., Saxena, A. K., & Wansley, J. W. (1996). The role of insiders and dividend policy: A comparison of regulated and unregulated firms. Journal of Financial and Strategic Decisions, 9(2), 1–9.

Conger, J. A., Finegold, D., & Lawler, E. E. (1998). Appraising board room performance. Harvard Business Review, 76(1), 139–148. https://hbr.org/1998/01/appraising-boardroom-performance

DeAngelo, H., DeAngelo, L., & Skinner, D. J. (1992). Dividends and losses. The Journal of Finance, 47(5), 1837–1863.

Deloof, M. (2001). Belgian intragroup relations and the determinants of corporate liquid reserves. European Financial Management, 7(3), 375–392.

Deshmukh, S. (2003). Dividend initiations and asymmetric information: A hazard model. Financial Review, 38, 351–368.

Dong, M., Robinson, C., & Veld, C. (2005). Why individual investors want dividends. Journal of Corporate Finance, 12(1), 121–158. https://doi.org/10.1016/j.jcorpfin.2004.04.006

D'Souza, J. (1999). Agency cost, market risk, investment opportunities and dividend policy-An international perspective. Managerial Finance, 25(6), 35–43.

Easterbrook, F. H. (1984). Two agency costs explanations of dividends. American Economic Review, 74, 650–659.

Elston, J. A. (1996). Dividend policy and investment: Theory and evidence from US panel data. Managerial and Decision Economics, 17, 267–275.

Elton, E. J., & Gruber, M. J. (1970). Marginal stockholder tax rates and the clientele effect. The Review of Economics and Statistics, 52(1), 68. https://doi.org/10.2307/1927599

Elyasiani, E., Jia, J., & Movaghari, H. (2019). Determinants of dividend payout and dividend propensity in an emerging market, Iran: an application of the LASSO. Applied Economics, 51(42), 4576–4596. https://doi.org/10.1080/00036846.2019.1593315

Enis, C. R., & Ke, B. (2003). The impact of the 1986 tax reform act on income shifting from corporate to shareholder tax bases: Evidence from the motor carrier industry. Journal of Accounting Research, 41(1), 65–88.

Erhardt, N. L., Werbel, J. D., & Shrader, C. B. (2003). Board of directors diversity and firm financial performance. Corporate Governance: An International Review, 11(2), 102–111. https://doi.org/10.1111/1467-8683.00011

Faccio, M., Lang, L. H. P., & Young, L. (2001). Dividends and expropriation. American Economic Review, 91(1), 54–78.

Fama, E. F., & French, K. R. (2001). Disappearing dividends: Changing firm characteristics or lower propensity to pay? Journal of Financial Economics, 60, 3–43.

Fama, E. F., & French, K. R. (2002). Testing trade-off and pecking order predictions about dividends and debt. The Review of Financial Studies, 15, 1–33.

Feldstein, M., & Green, J. (1979). Why do companies pay dividends? https://doi.org/10.3386/w0413

Financial Reporting Council (FRC). (2012). The UK Corporate Governance Code. https://www.frc.org.uk/getattachment/e322c20a-1181-4ac8-a3d3-1fcfbcea7914/UK-Corporate-GovernanceCode-(September-2012).pdf

Glen, J. D., Karmokolias, Y., Miller, R. R., & Shah, S. (1995). Dividend policy and behavior in emerging markets. Discussion Paper No. 26, International Financial Corporation.

Gordon, M. J. (1959). Dividends, earnings, and stock prices. The Review of Economics and Statistics, 41(2), 99. https://doi.org/10.2307/1927792

Graham, B., & Dodd, D. L. F. (1934). Security analysis. New York, NY: McGraw-Hill Education.

Grullon, G., Michaely, R., & Swaminathan, B. (2002). Are dividend changes a sign of firm maturity? Journal of Business, 75, 387–424.

Gugler, K., & Yurtoglu, B. B. (2003). Corporate governance and dividend pay-out policy in Germany. European Economic Review, 47, 731–758.

Han, K. C., Lee, S. H., & Suk, D. Y. (1999). Institutional shareholders and dividends. Journal of Financial and Strategic Decisions, 12, 53–62.

Holder, M. E., Langrehr, F. W., & Hexter, J. L. (1998). Dividend policy determinants: An investigation of the influences of stakeholder theory. Financial Management, 27, 73–82.

How to explain the ROC AUC score and ROC curve? (n.d.). https://www.evidentlyai.com/classification-metrics/explain-roc-curve

Ivașcu, C. (2023). Understanding dividend puzzle using machine learning. Computational Economics, 64(1), 161–179. https://doi.org/10.1007/s10614-023-10439-7

Jensen, G. R., Solberg, D. P., & Zorn, T. S. (1992). Simultaneous determination of insider ownership, debt, and dividend policies. Journal of Financial and Quantitative Analysis, 27, 274–263.

Jensen, M. C., & Meckling, W. H. (1976). Theory of the firm: Managerial behavior, agency costs and ownership structure. Journal of Financial Economics, 3(4), 305–360. https://doi.org/10.1016/0304-405x(76)90026-x

Jensen, M. C. (1986). Agency costs of free cash flow, corporate finance, and takeovers. American Economic Review, 76, 323–329.

Jensen, M. C., & Meckling, W. H. (1976). Theory of the firm: Managerial behavior, agency costs and ownership structure. Journal of Financial Economics, 3, 305–360.

Kiel, G. C., & Nicholson, G. J. (2003). Board composition and corporate performance: How the Australian experience informs contrasting theories of corporate governance. Corporate Governance: An International Review, 11(3), 189–205. https://doi.org/10.1111/1467-8683.00318

Konak, F., Bülbül, M. A., & Türkoğlu, D. (2024). Feature selection and hyperparameters optimization employing a hybrid model based on genetic algorithm and artificial neural network: Forecasting dividend payout ratio. Computational Economics, 63(4), 1673–1693. https://doi.org/10.1007/s10614-023-10530-z

Kumar, A., & Sinha, P. (2024). Changing dividend payout behavior and predicting dividend policy in emerging markets: Evidence from India. Investment Management and Financial Innovations, 21(1), 259–274. https://doi.org/10.21511/imfi.21(1).2024.20

Kyereboah-Coleman, A., & Biekpe, N. (2006). The link between corporate governance and performance of the non-traditional export sector: Evidence from Ghana. Corporate Governance: The International Journal of Business in Society, 6(5), 609–623. https://doi.org/10.1108/14720700610706090

La Porta, R., Lopez de Silanes, F., Shleifer, A., & Vishny, R. W. (2000). Agency problems and dividend policies around the world. Journal of Finance, 55(1), 1–33.

Lakshmishree. (2024). Highest dividend-paying stocks in India 2024. Lakshmishree.com. https://lakshmishree.com/blog/highest-dividend-paying-stocks-in-india

Lintner, J. (1965). The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. The Review of Economics and Statistics, 47(1), 13. https://doi.org/10.2307/1924119

Lintner, J. (1956). Distribution of incomes of corporations among dividends, retained earnings, and taxes. American Economic Review, 46, 97–113.

Lipton, M., & Lorsch, J. W. (1992). A modest proposal for improved corporate governance. Business Lawyer, 48(1), 59–77. https://www.hbs.edu/faculty/Pages/item.aspx?num=4179

Litzenberger, R. H., & Ramaswamy, K. (1979). The effect of personal taxes and dividends on capital asset prices. Journal of Financial Economics, 7(2), 163–195. https://doi.org/10.1016/0304-405x(79)90012-6

Lloyd, W. P., Jahera, S. J., & Page, D. E. (1985). Agency cost and dividend payout ratios. Quarterly Journal of Business and Economics, 24(3), 19–29.

Longinidis, P., & Symeonidis, P. (2013). Corporate dividend policy determinants: Intelligent versus a traditional approach. Intelligent Systems in Accounting, Finance & Management, 20(2), 111–139. https://doi.org/10.1002/isaf.1338

Lotfi, T. (n.d.). Dividend policy, signaling theory: A literature review. https://ssrn.com/abstract=3359144

Mahama, I. (2023). Relationship between macroeconomic variables and dividend payout of some listed firms on the Ghana stock exchange market: A random effect analysis approach. Asian Journal of Probability and Statistics, 21(3), 1–10. https://doi.org/10.9734/ajpas/2023/v21i3463

Mansourinia, E., Emamgholipour, M., Rekabdarkolaei, E. A., & Hozoori, M. (2013). The effect of board size, board independence and CEO duality on dividend policy companies: Evidence from Tehran stock exchange. International Journal of Economy, Management and Social Sciences, 2(6), 237–241. https://ru.scribd.com/document/251461981/The-Effect-of-Board-Size-Board-Independence-and-CEO-Dualityon-Dividend-Policy-of-Companies-Evidence-from-Tehran-Stock-Exchange

McMillan, D. G. (2014). Stock return, dividend growth and consumption growth predictability across markets and time: Implications for stock price movement. International Review of Financial Analysis, 35, 90–101. https://doi.org/10.1016/j.irfa.2014.07.011

Michaely, R., & Roberts, M. R. (2007). Corporate dividend policies: Lessons from private firms. Working Paper, Cornell University.

Michel, A. (1979). Industry influence on dividend policy. Financial Management, 8(3), 22. https://doi.org/10.2307/3665034

Miller, M. H., & Modigliani, F. (1961). Dividend policy, growth, and the valuation of shares. The Journal of Business, 34(4), 411. https://doi.org/10.1086/294442

Miller, M. H., & Rock, K. (1985). Dividend policy under asymmetric information. The Journal of Finance, 40, 1031–1051. http://dx.doi.org/10.1111/j.1540-6261.1985.tb02362.x

Moh'd, M. A., Perry, L. G., & Rimbey, J. N. (1995). An investigation of dynamic relationship between agency theory and dividend policy. The Financial Review, 30, 367–385.

Morck, R., Shleifer, A., & Vishny, R. (1988). Management ownership and firm value: An empirical analysis. Journal of Financial Economics, 20(1), 293–315.

Mueller, D. C. (1972). A life cycle theory of the firm. Journal of Industrial Economics, 20(3), 199. https://doi.org/10.2307/2098055

Mundati, Z. W. (2013). The effects of macroeconomic variables on the dividend payout of firms listed at the Nairobi securities exchange. http://chss.uonbi.ac.ke/sites/default/files/chss/MUNDATI%20Z%20WANJIRU-D61-73241-2009_0.pdf

Myers, S. C. (1984). The capital structure puzzle. Journal of Finance, 39, 575–592.

Myers, S. C., & Majluf, N. S. (1984). Corporate financing and investment decisions when firms have information that investors do not have. Journal of Financial Economics, 13, 187–221.

Ntim, C. G., & Osei, K. A. (2011). The impact on corporate board meeting on corporate performance in South Africa. African Review of Economics and Finance, 2(2), 83–103. https://mpra.ub.uni-muenchen.de/45814/1/MPRA_paper_45814.pdf

Pandey, I. M. (2001). Corporate dividend policy and behaviour: The Malaysian experience. Working Paper No. 2001-11-01, Indian Institute of Management Ahmedabad.

Pruitt, S. W., & Gitman, L. W. (1991). The interactions between the investment, financing, and dividend decisions of major US firms. Financial Review, 26(33), 409–430.

Razaee, Z. (2008). Corporate governance and ethics. Hoboken, NJ: John Wiley & Sons, Inc.

Redding, L. S. (1997). Firm size and dividend payouts. Journal of Financial Intermediation, 6, 224–248.

Reuters. (2024). India's Hindustan Zinc nearly doubles dividend payout. Reuters. https://www.reuters.com/markets/commodities/indias-hindustan-zinc-nearly-doubles-dividend-2024-08-20

Romus, M., Anita, R., Abdillah, M. R., & Zakaria, N. B. (2020). Selected firms environmental variables: Macroeconomic variables, performance and dividend policy analysis. IOP Conference Series Earth and Environmental Science, 469(1), 012047. https://doi.org/10.1088/1755-1315/469/1/012047

Rozeff, M. S. (1982). Growth, beta and agency costs as determinants of dividend payout ratios. The Journal of Financial Research, 5, 249–259.

Society, F. (2025, March 18). Global dividend payouts reach record 1.75 trillion in 2024. Funds Society. https://www.fundssociety.com/en/news/markets/global-dividend-payouts-reach-record-1-75-trillion-in-2024

Vodwal, S., & Negi, V. (2023). Applications of Machine Learning and Determinants of Dividend Decision : Evidence from Indian Firms. Indian Journal of Finance, 17(5), 8. https://doi.org/10.17010/ijf/2023/v17i5/171154

Walter, J. E. (1963). Dividend policy: Its influence on the value of the enterprise. The Journal of Finance, 18(2), 280–291. https://doi.org/10.1111/j.1540-6261.1963.tb00724.x

Wang, X., Barbopoulos, L. G., Rzayev, K., University of Edinburgh, Business School, Edinburgh, UK, & Koc University, Istanbul, Turkey. (n.d.). Dividend forecasts via machine learning. Abstract.

Why do some investors prefer dividend stocks over growth stocks? - WalletInvestor Magazin - Investing news. (n.d.). Walletinvestor.com. https://walletinvestor.com/magazine/why-do-some-investors-prefer-dividend-stocks-over-growth-stocks?

Won, C., Kim, J., & Bae, J. K. (2012). Using genetic algorithm based knowledge refinement model for dividend policy forecasting. Expert Systems With Applications, 39(18), 13472–13479. https://doi.org/10.1016/j.eswa.2012.06.001

Yaseen, H., & Dragotă, V. (2021). Forecasting the dividend policy using machine learning approach: decision tree regression models. In Eurasian studies in business and economics (pp. 19–39). https://doi.org/10.1007/978-3-030-71869-5_2

Zhang, Y. (2008). Information asymmetry and the dismissal of newly appointed CEOs: An empirical investigation. Strategic Management Journal, 29(8), 859–872. https://doi.org/10.1002/smj.689

Appendix

A. Definition of all variables used in our study

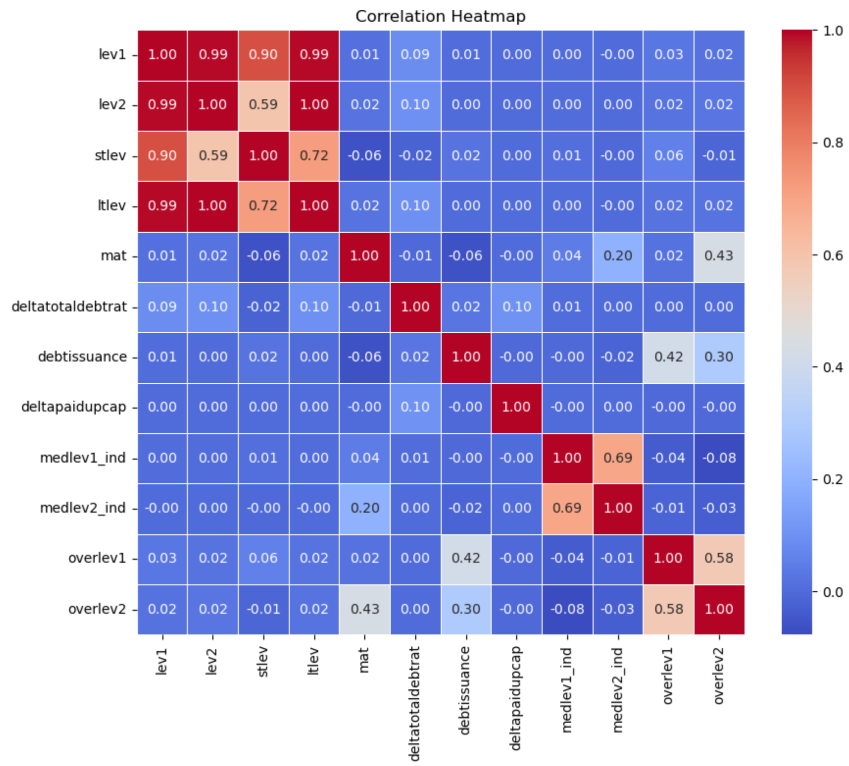
Variable	Description
Leverage	
TOTALDEBT	Total debt
LEV1	Leverage 1
LEV2	Leverage 2
STLEV	Short-term leverage
LTLEV	Long-term leverage
MAT	Maturity
DELTATOTALDEBTRAT	Delta total debt ratio
DEBTISSUANCE	Debt issuance
DELTAPAIDUPCAP	Delta paid-up capital
MEDLEV1_IND	Industry median leverage 1
MEDLEV2_IND	Industry median leverage 2
OVERLEV1	Over leverage 1
OVERLEV2	Over leverage 2
Profitability	
ROA	Return on assets
ROE	Return on investment
EBITDAMARGIN	Ebitda margin
EBITMARGIN	Ebit margin
PATMARGIN	Pat margin
EBITBYTA	EBIT by total assets
ROCE	Return on capital employed
PATGROWTH	Pat growth
REVGROWTH	Revenue growth
DUMMY_REVGROWTH	Dummy revenue growth
DUMMY_PATGROWTH	Dummy pat growth
EARNINGVOL	Earning volatility
REVVOL	Revenue volatility
Asset and Investment	
SIZE	Log of total assets
TANG	Tangibility

DEPRATIO	Depreciation ratio
ASSETMAT	Asset maturity
CAPEX	Capital expenditure
DELTACAPEX	Delta capital expenditure
LOG_CAPEX	Log of capex
MEDTANG_IND	Industry median of tangibility
MEDCAPEX_IND	Industry median of capex
OVERTANG	Over tangibility
OVERCAPEX	Over capex
LOGASSGR	Log of asset growth
MEDLOGASSGR_IND	Industry median of log asset growth
Liquidity	
CASHRAT	Cash ratio
TOTALCASHEQURAT	Total cash equivalent ratio
CASHFLOWRAT	Cash flow ratio
CR	Current ratio
NWCBYTA	Net Working Capital by Total Asset
CFVOL_IND	Cash flow volatility at the industry-year level
CASHFLOWVOL	Cash flow Volatility
MEDCASHFLOWRAT_IND	Industry median of cash flow ratio
MEDTOTALCASHEQURAT_IND	Industry median of Total cash equivalent ratio
OVERCASHFLOWRAT	Binary variable: Over cashflow ratio
OVERTOTALCASHEQURAT	Binary variable: Total cash equivalent ratio
DELTACASHEQURAT	Delta Cash equivalent ratio
FCF	Free cash flow
Dividend	
EQUITY_DIVDUMMY	Binary variable: Equity Dividend
SPE_DIVDUMMY	Binary variable: Special Dividend
EQUITY_PAYOUTRAT	Equity dividend payout ratio
EQUITY_DIVBYTA	Equity Dividend by Total Asset
EQUITY_DIVBYOP	Equity Dividend by Operational margin
SPE_PAYOUTRAT	Special dividend Payout ratio
SPE_DIVBYTA	Special dividend by Total Asset
SPE_DIVBYOP	Special dividend by operational margin
EQUITYDIVGROWTH_DUMMY	Binary variable: for growth in equity Dividend
DIVIN5YEAR	Dividend Given In last 5 Years
Company	
AGE	Age of the firm
LISTED	Binary Variable: BSE listing of Firm.
OWNER_GOV	Binary variable: Government owned Firm

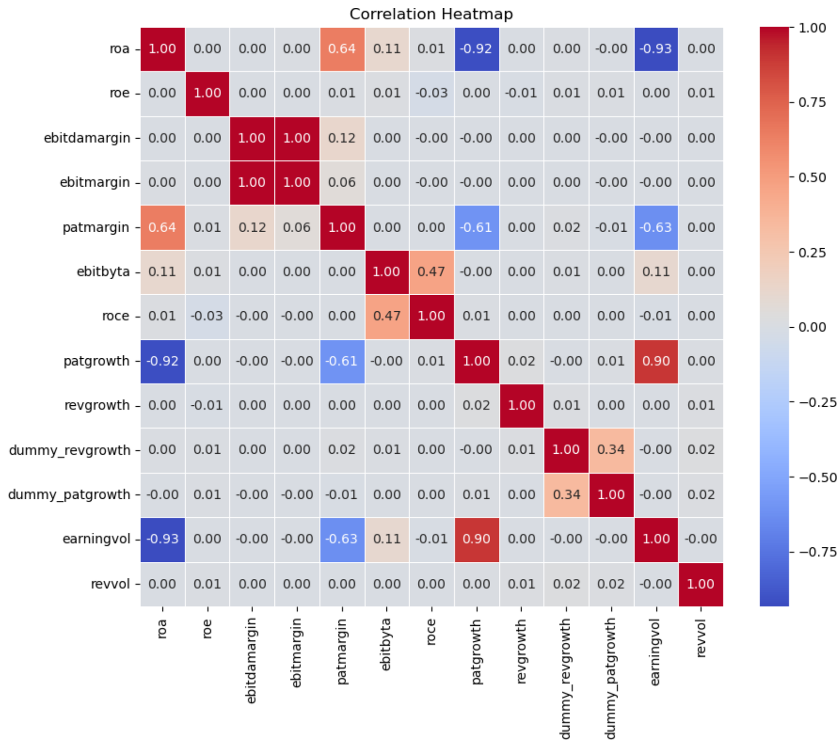
OWNER_GROUP	Binary variable: The Firm is owned By Group Company
OWNER_PRIVATE	Binary variable: Private owned Firm
TOTAL_OTH_OWN	Total Other Directorship in a Given Year
PROPO_INDEP	Proportion of Independent Directors
PROPO_EXECU	Proportion of Executive Directors
Expense	
TAXRATE	Tax rate
SGABYTA	SA& G Expense by total Asset
RNDBYREV	R&D expenditure by Revenue
RNDBYTA	R&D expenditure by Total Asset
RND_DUMMY	Is company spending for R&D or not
LIFECYCLE	Life cycle
INTRATE	Interest rate
Macroeconomic	
EXC_RATE	Exchange rate
CONS_PCT_GDP	Final consumption expenditure (% of GDP)
INV_PCT_GDP	Gross capital formation (% of GDP)
CREDIT_PRIVATE_GDP	Domestic credit to private sector (% of GDP)
MRK_CAP_PCT_GDP	Market capitalization of listed domestic companies (% of GDP)
EXPND_PCT_GDP	Government expenditure, percent of GDP (% of GDP)
GDP_GR	Real GDP growth (Annual percent change)
GDP	GDP per capita, current prices (U.S. dollars per capita)
INF_CPI	Inflation rate, end of period consumer prices (Annual percent change)
UNEMP_RATE	Unemployment, total (% of total labour force)
REPORATE	Exchange rate
INDUSTRY_46.0	Wholesale trade, except motor vehicles and motorcycles, industry dummy
INDUSTRY_77.0	Leasing of non-financial intangible assets, industry dummy
INDUSTRY_20.0	Manufacture of chemicals and chemical products, industry dummy

B. Correlation matrices for all set of variables:

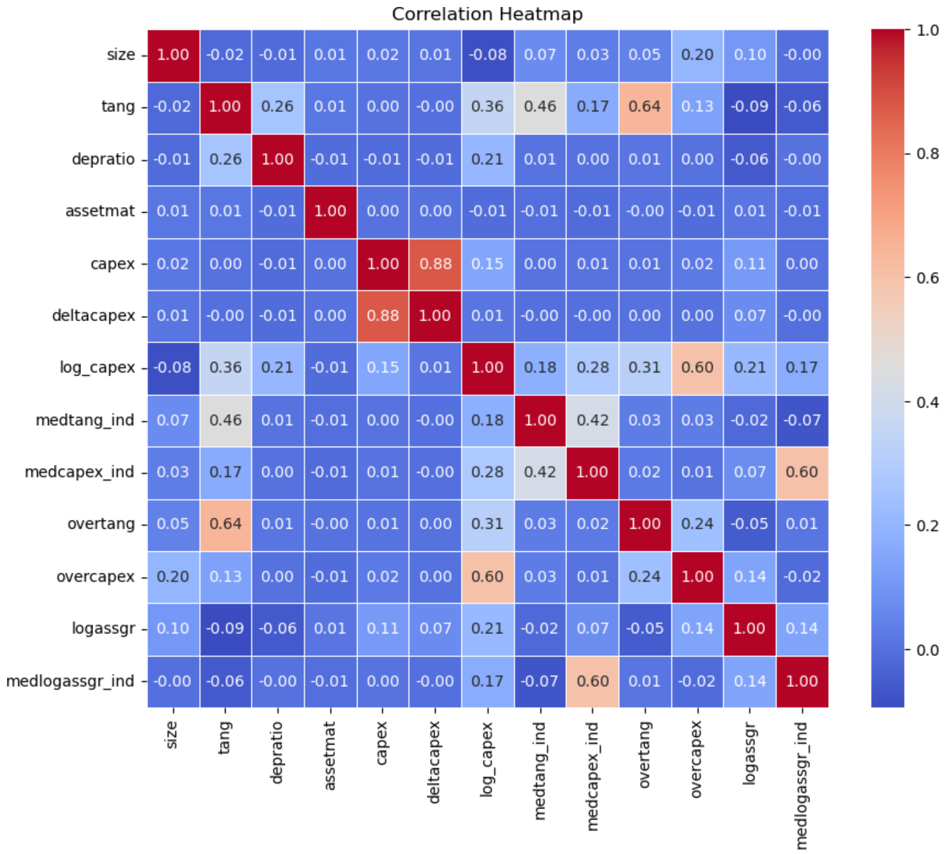
i. Leverage related variables:



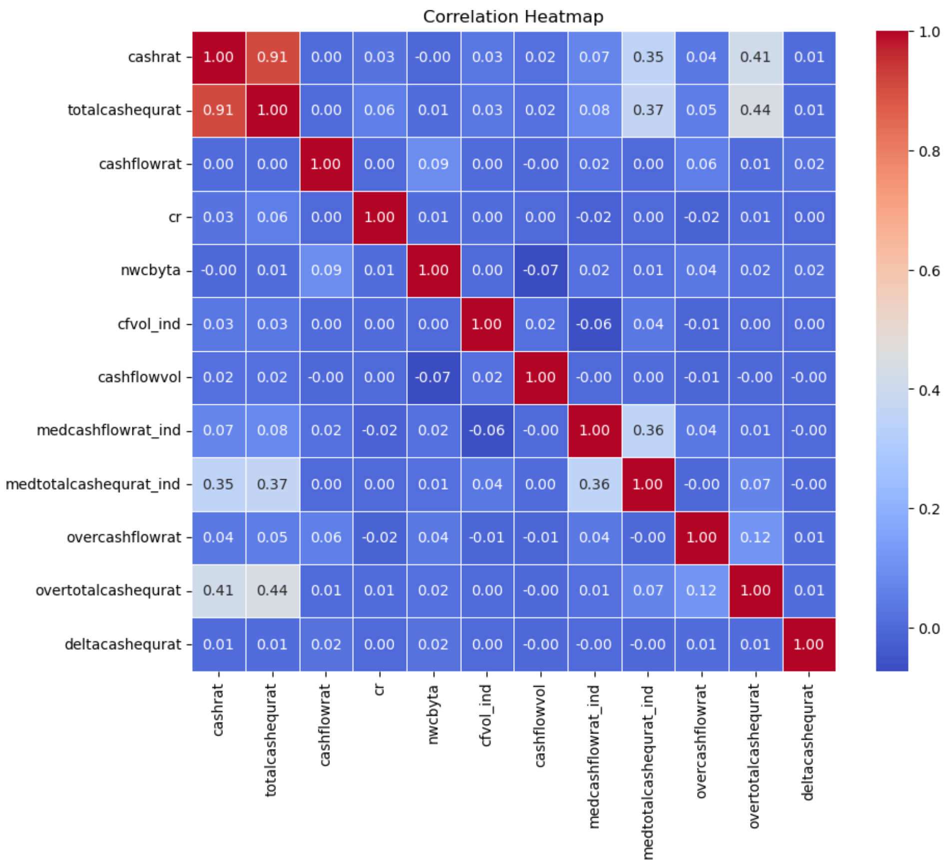
ii. Profitability variables



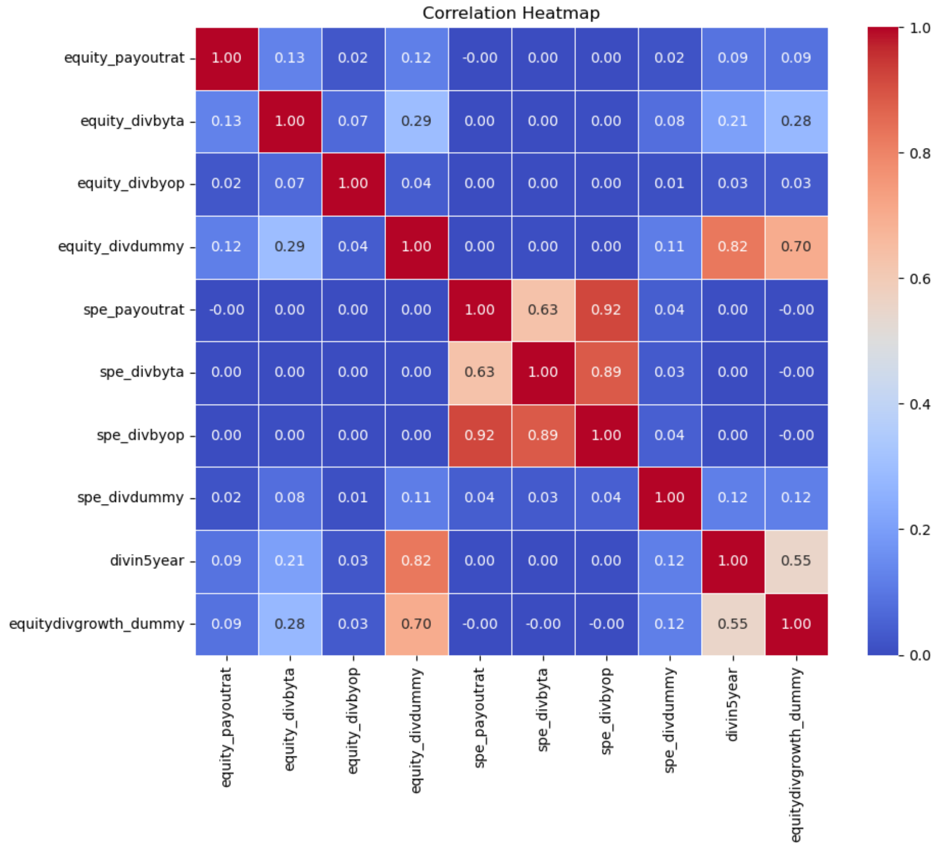
iii. Asset Investment Retreated Variables:



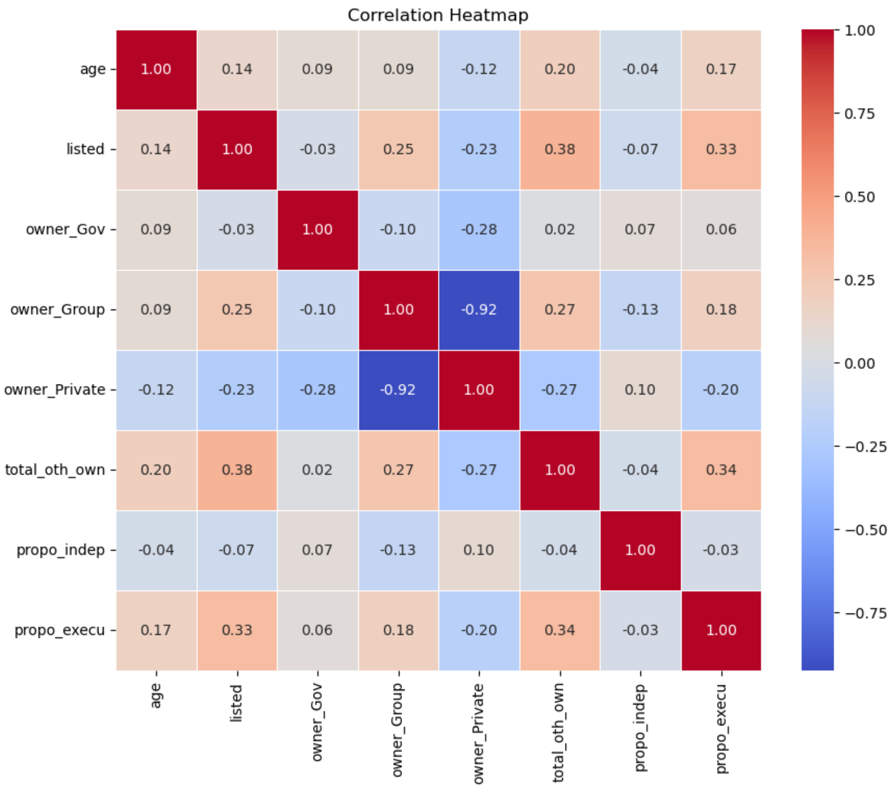
iv. Liquidity-related variables



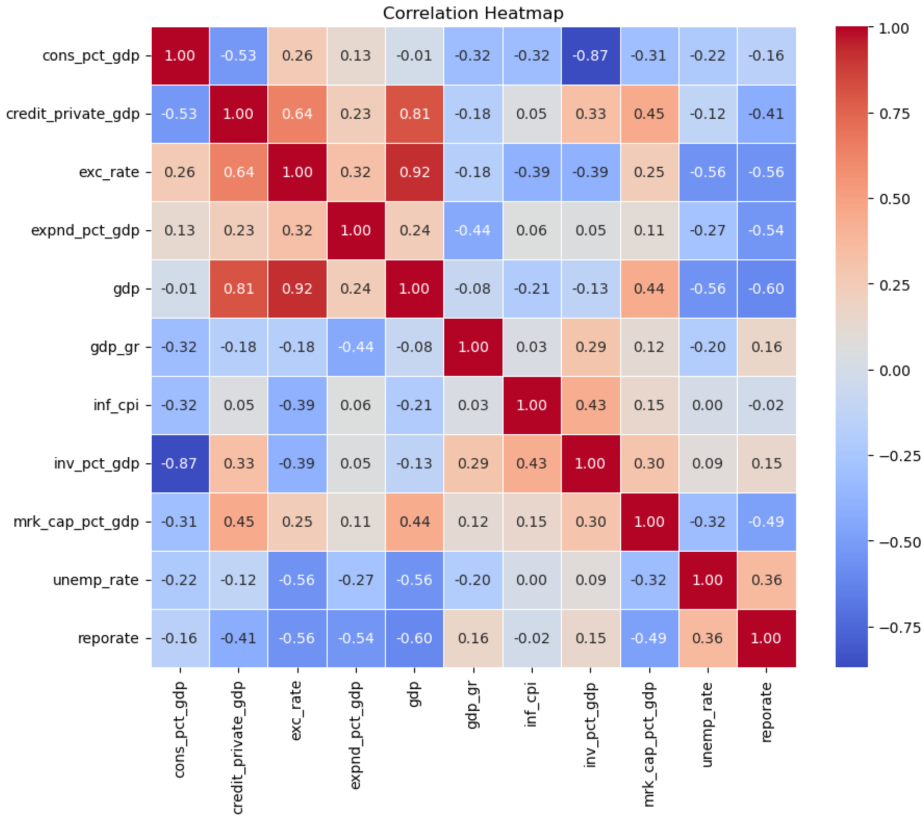
v. Dividend-related variables:



vi. Firm-related variables



vii. Country-level variables



viii. Expense Related Variables