ECO482: Machine Learning Applications in Macro Economic Finance

Research Project Part 2

Title: ***Drivers of Dividend Reduction An Analysis of Insurance Companies***

Group 1

Introduction

Insurance companies play an essential role in the financial system by providing products and services that offer a sense of security to their customers. In the United States, 92.1% of the population, approximately 304 million people, had insurance coverage at some point during 2022 (Keisler-Starkey, 2023). Thus, it is crucial for insurance companies to maintain sufficient liquidity to fulfill claims promptly when customers file them. Historically, there have been significant instances where insurance companies have denied claims to customers during times of crisis. During the COVID-19 pandemic in 2020, National Farmers Union (NFU) Mutual faced scrutiny when it was unable to honor business interruption claims from policyholders (Case, 2024). This underscores the importance of robust financial planning and adequate risk assessment within the insurance sector.

Dividends on common stock serve as a valuable indicator for assessing a company's financial health on a quarterly basis and can signal stability and security to its policyholders. This paper utilizes dividend payout as a key metric for evaluating a company's financial condition where a consistent decline in dividends over multiple quarters may suggest that the company is experiencing financial distress. In such cases, consumers can make more informed decisions, such as considering alternative insurance providers. The core offerings of insurance companies primarily encompass automobile, property and casualty, and life insurance products. Beyond financial metrics such as liquidity, shareholder equity, and profitability ratios, the dividends issued by these companies are particularly sensitive to catastrophic events (like natural disasters and road accidents). In the United States, recurring natural disasters such as hurricanes, floods, and wildfires result in substantial financial losses and tragic fatalities. Additionally, road accidents, which occur frequently, contribute to increased claims and financial strain on insurers.

In prior studies, researchers have explored the use of machine learning models to predict dividend yields. Chen investigates the reversal of the correlation between dividend yield and dividend growth predictability during periods marked by significant dividend cuts, employing financial indicators to forecast dividend changes, with a particular focus on the likelihood of dividend reductions (Chen, 2017). Similarly, Ivascu applies machine learning models such as Random Forest, XGBoost, Logistic Regression, and Decision Tree Classifiers to estimate the probability of dividend payouts (Ivașcu, 2023). Building on these studies, our research specifically targets the insurance sector while incorporating additional models, including K-Nearest Neighbors (KNN), Linear Discriminant Analysis (LDA), and Bayesian Classifiers. Furthermore, we enhance the analysis by integrating variables uniquely pertinent to the insurance industry, such as automobile accidents and natural disaster losses, to develop a domain-specific framework for predictive analysis.

Thus, our research aims to explore whether financial ratios, car accidents, and natural disasters can accurately predict the likelihood of a dividend payout reduction in insurance companies. Our study concluded that financial ratios remain the most dominant factor in predicting dividend payouts in the subsequent quarter, particularly the price to earnings ratio and net return on assets.

Data and Methodology

The data used to address our research question was obtained from multiple reliable sources. We selected 18 publicly traded Insurance companies listed on the New York Stock Exchange such as Allstate, Cigna International, MetLife etc. The dataset has a panel data structure, covering the period from 2000 to 2022, with quarterly observations for each company. Financial metrics for these insurance companies were sourced quarterly from FactSet, a widely recognized financial data provider. These metrics encompass a range of ratios like profitability, liquidity, shareholder,

and debt management ratios. This includes the net income return on assets, net income on shareholder equity, dividend payout ratio, return on total capital, debt-to-equity ratio, and debt-to-capital ratio, among others. Additionally, data on road accidents was sourced from the National Highway Traffic Safety Administration, providing quarterly totals of road accidents in the U.S. from 2000 to 2022. Data on natural disasters was retrieved from the National Centers for Environmental Information. This data reports the type and date of disasters, the number of fatalities, and the financial loss incurred (measured in billions). In total, our dataset consists of approximately 1,400 observations and includes 30 predictors.

An important indicator highlighted in this analysis is the financial losses resulting from natural disasters. Notably, *figure 1* illustrates significant spikes in the years 2005, 2012, 2017, and 2022, all of which correspond to severe tropical cyclones or hurricanes, known to be among the most destructive types of natural disasters. Moreover, the response variable in this analysis is binary, indicating whether the dividend payout percentage decreased in the subsequent quarter (True) or remained unchanged (False). As depicted in *figure 2*, the distribution of dividend reductions is balanced between the two categories. This absence of class imbalance in the dataset enhances the predictive capability of our machine learning models, as no single class dominates the predictions.

Figure 1: Financial Losses

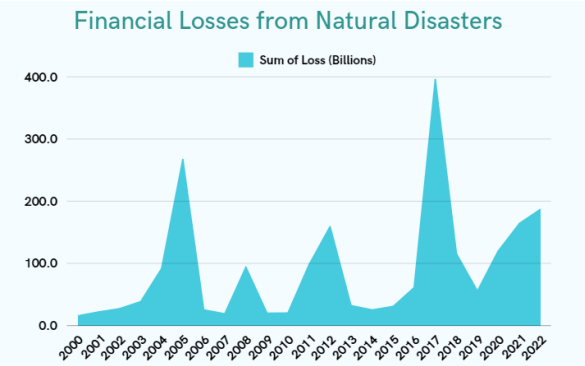
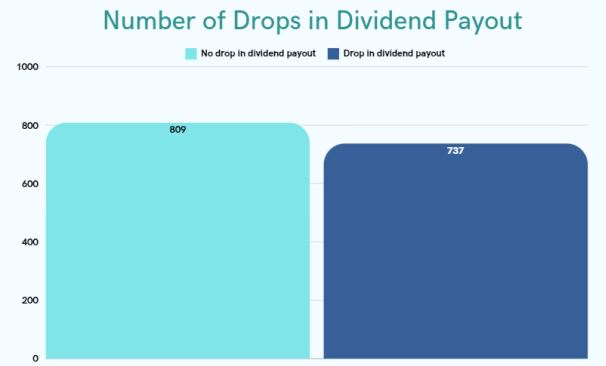


Figure 2: Distribution of Dividend Drop



Our objective is to develop a model with strong out-of-sample performance which ensures accurate predictions of dividend drops. Testing accuracy is crucial, as the model's financial value lies in correctly forecasting future drops. Moreover, a low false positive rate is critical to minimize the financial risk to investors and policy holders, making Area under the Curve (AUC) a key evaluation metric.

Beyond accuracy, we aim to assess the relative importance of financial ratios, accident rates, and natural disaster damages in predicting dividend drops. To achieve this, we will employ a diverse set of models each reflecting distinct assumptions about the data-generating process:

- Linearity: Whether the relationship between features and response is linear or non-linear
- Distributional assumptions: Whether predictors follow a Gaussian distribution which aids in predicting the response label.
- Distance or similarity assumptions: Whether the distance between points in the feature space significantly determines our response label.

This comprehensive approach will provide valuable insights into the underlying data-generating mechanism while helping identify the model with the strongest out-of-sample performance.

The models we will use are as follows:

Linear	Nonlinear	Distributional assumptions	Distance or similarity assumptions
-Logistic regression -Linear Discriminant analysis	-Random forest -XGboosting -Decision tree	- Linear Discriminant analysis -Naive Bayes classifier	-K neighbours classifier -Naive Bayes classifier

We begin by splitting the data into training (80%) and testing (20%), scaling the features to ensure compatibility with the logistic regression and K-neighbors models. We use lagged features, assuming a quarterly decision lag in response to the publishing of financial information. Using grid search, we then tune hyperparameters by initializing the models and perform stratified five-fold cross-validation on the training set. Stratification ensures that class distributions are similar over the k folds, while k=5 balances out-of-sample performance and overfitting risk. The grid search identifies optimal hyperparameters that maximize average cross-validation accuracy (Refer to Appendix A for models and respective hyperparameters).

After determining the optimal hyperparameters, we reinitialize and fit the models on the training set. Using these models, predictions are made on the test set. For logistic regression, we build separate models using Ridge and Lasso penalties. For the random forest model, we apply recursive feature elimination based on the feature importance matrix. This helps identify key predictors, reducing overfitting and improving test accuracy. Finally, we calculate the training and testing accuracy of all the models and compute predicted probabilities for the positive class to determine the Area Under the Curve (AUC).

Results

As observed in *table 1*, the random forest model achieves the highest testing accuracy, likely due to its ability to mitigate overfitting by aggregating predictions from multiple decision trees trained on bootstrap samples. This is valuable given our dataset of 1,418 observations which increases the risk of overfitting, especially for the decision tree and XGBoost classifier.

Logistic regression ranks second, indicating a reasonably strong linear relationship between features and the response. In contrast, the K-Neighbours classifier is second last likely due to the

ineffectiveness of the distance metric in the high-dimensional feature space, where points tend to appear equidistant. This issue may stem from the violation A KNN's assumption that features follow a normal distribution. This makes sense since financial time series data tends to exhibit non-normality (**include citation**).

Table 1: Model Accuracy and AUC

Model	Average CV Training Accuracy	Testing Accuracy	Area under Curve
Random Forest	0.637527	0.640845	0.693452
Logistic Regression (Lasso)	0.562571	0.630282	0.655605
Logistic Regression (Ridge)	0.604056	0.626761	0.655506
XGBoost Classifier	0.659573	0.626761	0.672569
Decision Tree Classifier	0.598752	0.559859	0.575273
Bayesian Classifier	0.539687	0.556338	0.572073
K-Nearest Neighbors (KNN)	0.541400	0.542254	0.574628
Linear Discriminant Analysis	0.527305	0.538732	0.559772

Table 2: Evaluation Metrics for Random Forest Model

Metric	Value
TPR	0.714286
FPR	0.430556
AUC	0.693452
Accuracy	0.640845

Out of our top two models, our random forest model has the highest AUC of 69% with a false positivity rate of 43%, as seen in *table 2*. This suggests that the random forest model can sufficiently discriminate between cases when the dividend drops and when it doesn't. However,

on an absolute scale, 43% is still high as it implies that close to half of all predictions are false positives.

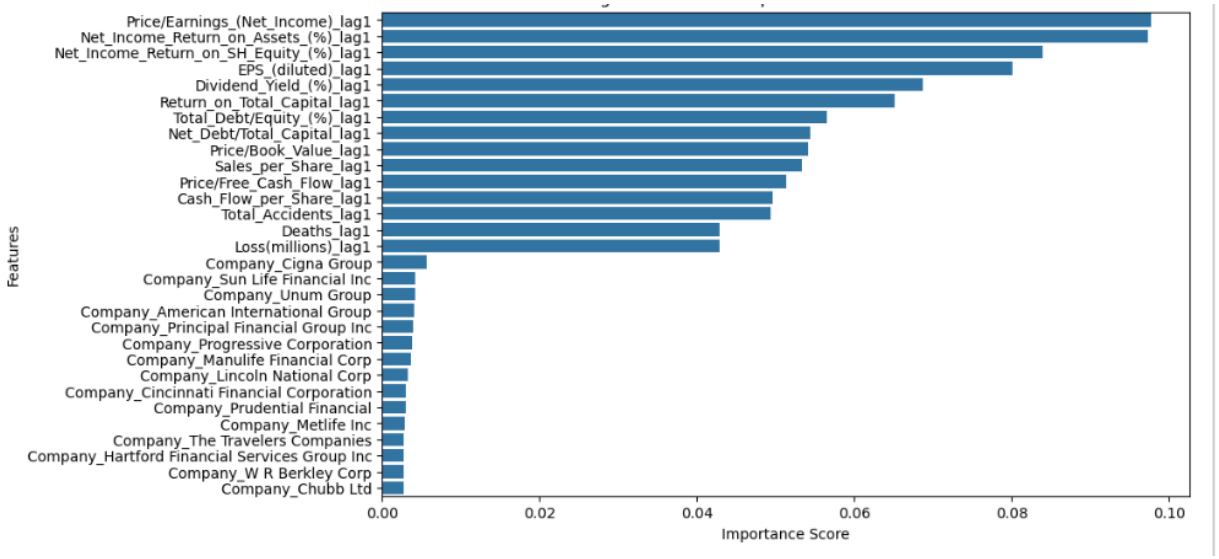
Table 3: Evaluation Metrics for Logistic – Lasso Model

Metric	Value
TPR	0.750000
FPR	0.701389
AUC	0.655605
Accuracy	0.630000

Despite having an AUC of 65% (not far behind random forest) the logistic (Lasso) regression has a very high false positivity rate of 70% as seen in *table 3*. This means, the model incorrectly predicts dividend drops 70% of the time. This is likely because a linear decision boundary cannot sufficiently capture class separation. Despite its overall high accuracy, the high number of false positives is concerning as an investor using this model would be exposed to a high risk of financial loss. Thus, the random forest model is the best model to use as it has the highest accuracy and AUC.

Profitability ratios emerge as the strongest predictors of dividend drops, serving as a reliable proxy for a firm's financial performance as observed in *figure 3* (refer Appendix B for ratio category details). Valuation ratios rank second, suggesting firm size influences dividend decisions (cite supporting research). Efficiency ratios highlight the role of revenue generation efficiency, while leverage ratios indicate that a firm's debt management significantly impacts dividend policies. This aligns with the notion that higher debt necessitates reinvestment to cover future interest payments.

Figure 3: Feature Importance – Random Forest Model



Natural disaster damages and accident rates show low importance, suggesting our data poorly models claims frequency. While these events may affect insurance company performance, their impact is likely underrepresented due to uneven geographic distribution and limited insurer presence in affected areas. Additionally, quarterly and national aggregation may dilute localized event effects.

Conclusion

In conclusion, we found that financial ratios are good predictors of the likelihood that an insurance company will drop its dividend payments in the upcoming quarter, with our best model correctly predicting dividend drops 71% of the time which is 21% higher than random chance. The dividend drops decisions unexplained by the model can be attributed to behavioral factors like investor preferences which are not directly measurable with our current data and models.

Bibliography

Case, P. (2024, June 27). *NFU Mutual served with Covid group action*. Retrieved from Farmers Weekly: https://www.fwi.co.uk/news/nfu-mutual-served-with-covid-group-action

Chen, R.-S. (2017). Dividend cuts and predictability. *Journal of Economics and Finance, 42*. doi:http://dx.doi.org/10.1007/s12197-017-9395-9

Ivașcu, C.-F. (2023). Understanding Dividend Puzzle Using Machine Learning. *Computational Economics*, 161-179. doi:https://doi.org/10.1007/s10614-023-10439-7

Keisler-Starkey, K. (2023, September 12). *Health Insurance Coverage in the United States: 2022*. Retrieved from United States Cencus Bureau: https://www.census.gov/library/publications/2023/demo/p60-281.html

Appendix

Appendix A: Models and hyperparameters

Models	Hyperparameters/optimized values
Logistic regression	• C: 100 • Penalty: L1 (lasso) • Solver: 'liblinear'
Decision tree classifier	• Criterion: 'gini' • Max_depth: 10 • Min_samples_leaf: 1

	• Min_samples_split: 2
Random forest	• Max_depth: 10 • Max_features: 2 • n_estimators: 200
K Nearest Neighbours classifier	• n_neighbours: 8
XGboost classifier	• Colsample_bytree: 0.8 • Learning_rate: 0.2 • Max_depth: 3 • N_estimators: 300 • Subsample: 0.8
Linear Discriminant analysis	• Shrinkage: none • Solver: 'lsqr'
Naive Bayes classifier(gaussian)	• Var_smoothing: 1 10-9

Appendix B: table with financial ratios and their type label

Type	Ratios
profitability	• Net income returns on assets • Net income return on Shareholder equity • Return on total capital
Leverage	• Total debt to equity ratio • Net debt to total capital ratio
Valuation	• Price to Earnings ratio • Price to Book value ratio • Price to free cash flow ratio • Dividend yield
Efficiency	• Sales per share • Earnings per share (diluted)
Liquidity	• Cash flow per share