




Project – AI for Finance



Zanolo, Luca

238838

Contents

Introduction	2
Forecasting Model	2
Features	3
Data Extraction.....	3
Data Cleaning	3
Unification and Feature Selection.....	6
Example: Pair Correlation Matrixes and VIFs before and after feature reduction.	7
Models	9
Econometric Analysis	9
Comments	9
Neural network analysis.....	11
Comments	11
Example: predictive capabilities	12
Conclusion.....	13

Introduction

This project focuses on implementing artificial intelligence (AI) models to forecast the stock returns of a company that belongs to a specific sector and country using quarterly data. The dataset is built acquiring and combining data from Refinitiv Eikon and World Bank Data and it's subjected to a cleaning and preparation process to ensure its consistency for model training.

The project employs two distinctive approaches. The first method is an econometric approach using a Logit model and the second approach uses a Multi-Layer Perceptron (MLP) Neural Network.

This report describes each stage of the project, from data acquisition and cleaning to model application and analysis. In the conclusion section, a comparison and commentary on both methodologies are provided.

It's important to note that all the results presented in this report depend on the choice of the country and sector. The project structure is designed for seamless testing of models across different country and sector configurations. It offers flexibility to modify various parameters that influence the entire pipeline, from data download and feature reduction to model training and prediction.

Thus, the outcomes and insights derived are specific to the chosen domain and may vary if other countries or sectors are selected. The used dataset includes a pool of 21 companies that belongs to the industrial sector of United States of America.

Forecasting Model

To predict stock returns, a range of 17 variables has been used. The predictive model considers both macroeconomic indicators, sourced from the World Bank, and financial ratios and stock information of specific companies, collected from Eikon. This mixed data approach aims to incorporate macroeconomic indicators, company-specific data and financial ratios.

The forecasting model used can be presented as follows:

$$Y_{t+1} = \alpha_0 + \sum_{i=1}^L (\alpha_i * StockIndex_{t-i}) + \sum_{i=1}^L (\gamma_i * MacroEconomic_{t-i}) + \sum_{i=1}^L (\varepsilon_i * CompanyVar_{t-i})$$

where:

- Y_{t+1} : the predicted binary outcome for the stock's closing value at time t+1. It takes the value of 1 if positive, and 0 if negative.
- α_0 : baseline value of the prediction when all other terms are zero.
- $\alpha_i, \gamma_i, \varepsilon_i$: coefficients associated with the lagged values of the corresponding group of variables. They determine the extent to which past values influence the prediction for Y_{t+1} .
- $StockIndex_{t-i}$: Represents a set of stock-related variables at a lag of i periods.
- $MacroEconomic_{t-i}$: a set of macroeconomic variables at a lag of i periods.
- $CompanyVar_{t-i}$: a collection of company-specific metrics at a lag of i periods.
- L: Represents the number of lags considered in the model.

Features

From the World Bank data, the model uses two macroeconomic indicators ($MacroEconomic_{t-i}$):

- Real GDP ("NY.GDP.MKTP.KD"): Annual percentage growth rate of GDP. A measure of the economic output of a country.
- CPI ("FP.CPI.TOTL.ZG"): Consumer Price Index, which reflects the annual percentage change in the cost for the average consumer to acquire a basket of goods and services.

For Eikon data, the model incorporates information about a company's stock ($StockIndex_{t-i}$), including:

- Close: closing price of a company's stock at the end of the trading day, or in this analysis at the end of the quarter. This variable is not directly used, from it, it's computed the binary dependent variable of the model. The dependent variable is obtained transforming the value of Close into returns and then to 1 if the return is positive, or to 0 if the return is negative.
- Volume: number of shares of a company's stock that change hands during a given period.

The model also considers various financial ratios and balance sheet/income statement variables ($CompanyVar_{t-i}$), divided into six groups:

- General Company Data: Total Liabilities, EBIT Percentage, Total Asset, and Total Debt.
- Financial Leverage Ratios: Debt-Equity Ratio, Equity Multiplier, and Total Debt Ratio.
- Asset Turnover Ratios: Asset Turnover Ratios and Reinvestment Rate.
- Profitability Ratios: Return on Equity (ROE), Return on Assets (ROA), and Profit Margin.
- Market Value Ratio: Price-Earnings Ratio, Market-to-Book Ratio, and Price-Sales Ratio.
- Dividends: Dividend Payout Ratio and Earnings Retention Rate.

These variables bring a broad range of information, from company profitability and leverage to market valuation ratios and dividend policies. The goal is to capture a comprehensive view of the financial health and performance of the companies.

Data Extraction

The data retrieval procedure is a dynamic process that aims to extract data to create a dataset according to certain parameters. The initial parameters involve the country and sector where the analysis is to be conducted. It's possible to specify the size of the group of companies to be considered for the dataset and specify the maximum number of requested companies. Other parameters to specify include the start and end date of the period of the analysis, the percentage of missing values tolerated, and the maximum number of missing values in a row that can be handled.

Once these parameters are defined, the download procedure can start. It will download macroeconomic data and then specific data for the stock and all the other variables. After some simple preliminary operations, the data will be combined into a single object to undergo a process of verification and cleaning. Both during the download and during the cleaning phase, the data must meet certain requirements to be accepted into the dataset, the failure to meet any of these requirements causes the immediate rejection of the current data and moves on to the data of the next company in the group.

Data Cleaning

The input of this phase is a structure containing the macroeconomic and company data merged in a single Pandas Dataframe. These data will undergo the following operations to evaluate, make them consistent, and optimized for the subsequent phases.

The cleaning phase is divided into 6 stages (0 - 5):

0. The columns are renamed according to a predefined mapping to have a standard and easily recognizable nomenclature of the columns. Then, any zeros are replaced with the nearest machine-representable value to zero to avoid subsequent errors due to the use of the value 0 in certain functions and operations.
1. In this stage, the period defined by the oldest and most recent quarter is identified. The goal of this stage is to ensure that the quarters represented within this period are complete, avoiding having unmanaged missing quarters for some companies. Any additional quarters will contribute to the missing value count.

After these first two stages, the first quality check on the data is carried out, considering the total number of cells without value and the number of columns that have more than the accepted number of missing values in a row. If the dataset does not have statistics that meet the imposed constraints, it will be discarded.

2. This stage involves a backward interpolation operation on the data. If the data has passed the previous checks, they will have at most an acceptable number of missing values, which will then be valued.
3. In this stage, the percentage change of some predefined variables is calculated.

After this stage, another checkpoint perform a check on the number of missing values, the expectation is to have only one row of missing values, introduced in stage number 3. If not, the data is discarded.

4. In this stage, a drop of the missing row is performed.
5. In the 5th and final stage, the data is standardized. The values are changed to ensure that each column has a mean of 0 and a variance of 1. This operation is useful to improve the performance of predictive modeling algorithms.

The last checkpoint checks that the dataset represents enough quarters, if the check is positive, then the prepared data will be added to the final dataset.

The procedure for analyzing the Group of companies will continue with this process until reaching the requested number of companies or until the Group of companies is exhausted. At the end of this phase, it will be possible, in a successive section, to examine reports related to cleaning. Each refers to a company and contains the details of the dataset for each stage of the cleaning phase. The details are both at the level of the entire dataset but also at the level of the single column.

The parameters used to retrieve a dataset from a group of 224 companies that belongs to the industrial American sector are the following:

- Missing Tolerance: 10% of total number of cells.
- Requested Companies: 22
- Minimum number of quarters: 100
- Maximum number of missing values in a row accepted: 8

Here is an example of generated reports during download phase. The first and second tables are related to a discarded company, the third regard an accepted company. The columns Rows # and Missing Rows # and Missing Rows % are referred to the entire dataset. All the others are specific for the column analyzed in the row.

Report for MATX.N - Used: False

	Variable	Component	Stage	Period	Rows #	Missing Rows #	Missing Rows %	Missing Cells #	Missing Cells %	Max Value	Min Value	Mean Value	Variance	Standard Deviation
20	CPI	MATX.N	0 - Ren. & Replace 0	1989-06-30 - 2023-03-31	136	93	68.382	0	0.000	8.003000e+00	-3.560000e-01	2.770000e+00	2.450000e+00	1.565000e+00
41	CPI	MATX.N	1 - Expand & Standard Na.N	1989-06-30 - 2023-03-31	136	93	68.382	0	0.000	8.003000e+00	-3.560000e-01	2.770000e+00	2.450000e+00	1.565000e+00
19	GDP	MATX.N	0 - Ren. & Replace 0	1989-06-30 - 2023-03-31	136	93	68.382	0	0.000	2.095269e+13	9.674724e+12	1.532368e+13	1.147626e+25	3.387664e+12
40	GDP	MATX.N	1 - Expand & Standard Na.N	1989-06-30 - 2023-03-31	136	93	68.382	0	0.000	2.095269e+13	9.674724e+12	1.532368e+13	1.147626e+25	3.387664e+12
12	ROA	MATX.N	0 - Ren. & Replace 0	1989-06-30 - 2023-03-31	136	93	68.382	6	4.412	3.885000e+01	2.830000e-01	5.499000e+00	3.740400e+01	6.116000e+00
33	ROA	MATX.N	1 - Expand & Standard Na.N	1989-06-30 - 2023-03-31	136	93	68.382	6	4.412	3.885000e+01	2.830000e-01	5.499000e+00	3.740400e+01	6.116000e+00
11	RCE	MATX.N	0 - Ren. & Replace 0	1989-06-30 - 2023-03-31	136	93	68.382	6	4.412	8.453900e+01	6.110000e-01	1.445100e+01	1.969070e+02	1.403200e+01
32	RCE	MATX.N	1 - Expand & Standard Na.N	1989-06-30 - 2023-03-31	136	93	68.382	6	4.412	8.453900e+01	6.110000e-01	1.445100e+01	1.969070e+02	1.403200e+01
9	asset_turnover	MATX.N	0 - Ren. & Replace 0	1989-06-30 - 2023-03-31	136	93	68.382	6	4.412	1.366000e+00	4.460000e-01	7.910000e-01	5.500000e-02	2.350000e-01
30	asset_turnover	MATX.N	1 - Expand & Standard Na.N	1989-06-30 - 2023-03-31	136	93	68.382	6	4.412	1.366000e+00	4.460000e-01	7.910000e-01	5.500000e-02	2.350000e-01
0	close	MATX.N	0 - Ren. & Replace 0	1989-06-30 - 2023-03-31	136	93	68.382	0	0.000	1.206200e+02	1.003500e+01	2.467000e+01	2.923710e+02	1.709900e+01
21	close	MATX.N	1 - Expand & Standard Na.N	1989-06-30 - 2023-03-31	136	93	68.382	0	0.000	1.206200e+02	1.003500e+01	2.467000e+01	2.923710e+02	1.709900e+01
17	dividend_payout_ratio	MATX.N	0 - Ren. & Replace 0	1989-06-30 - 2023-03-31	136	93	68.382	5	3.676	7.562310e+02	3.563000e+00	7.039400e+01	9.296937e+03	9.642100e+01
38	dividend_payout_ratio	MATX.N	1 - Expand & Standard Na.N	1989-06-30 - 2023-03-31	136	93	68.382	5	3.676	7.562310e+02	3.563000e+00	7.039400e+01	9.296937e+03	9.642100e+01
18	earnings_retention_rate	MATX.N	0 - Ren. & Replace 0	1989-06-30 - 2023-03-31	136	93	68.382	5	3.676	9.640000e-01	-6.562000e+00	2.960000e-01	9.300000e-01	9.640000e-01
39	earnings_retention_rate	MATX.N	1 - Expand & Standard Na.N	1989-06-30 - 2023-03-31	136	93	68.382	5	3.676	9.640000e-01	-6.562000e+00	2.960000e-01	9.300000e-01	9.640000e-01
3	ebit_margin	MATX.N	0 - Ren. & Replace 0	1989-06-30 - 2023-03-31	136	93	68.382	0	0.000	3.839500e+01	-6.561750e+02	6.866000e+00	3.338376e+03	5.777900e+01
24	ebit_margin	MATX.N	1 - Expand & Standard Na.N	1989-06-30 - 2023-03-31	136	93	68.382	0	0.000	3.839500e+01	-6.561750e+02	6.866000e+00	3.338376e+03	5.777900e+01
7	equity_multiplier	MATX.N	0 - Ren. & Replace 0	1989-06-30 - 2023-03-31	136	93	68.382	0	0.000	4.810000e+00	1.862000e+00	2.716000e+00	5.620000e-01	7.490000e-01
28	equity_multiplier	MATX.N	1 - Expand & Standard Na.N	1989-06-30 - 2023-03-31	136	93	68.382	0	0.000	4.810000e+00	1.862000e+00	2.716000e+00	5.620000e-01	7.490000e-01
15	market_to_book_ratio	MATX.N	0 - Ren. & Replace 0	1989-06-30 - 2023-03-31	136	93	68.382	92	67.647	4.727000e+00	9.420000e-01	2.715000e+00	9.390000e-01	9.690000e-01
36	market_to_book_ratio	MATX.N	1 - Expand & Standard Na.N	1989-06-30 - 2023-03-31	136	93	68.382	92	67.647	4.727000e+00	9.420000e-01	2.715000e+00	9.390000e-01	9.690000e-01
14	price_earnings_ratio	MATX.N	0 - Ren. & Replace 0	1989-06-30 - 2023-03-31	136	93	68.382	92	67.647	1.171000e+00	3.710000e-01	7.620000e-01	3.400000e-02	1.830000e-01
35	price_earnings_ratio	MATX.N	1 - Expand & Standard Na.N	1989-06-30 - 2023-03-31	136	93	68.382	92	67.647	1.171000e+00	3.710000e-01	7.620000e-01	3.400000e-02	1.830000e-01
16	price_sales_ratio	MATX.N	0 - Ren. & Replace 0	1989-06-30 - 2023-03-31	136	93	68.382	93	68.382	1.192000e+01	1.634000e+00	7.526000e+00	6.098000e+00	2.469000e+00
37	price_sales_ratio	MATX.N	1 - Expand & Standard Na.N	1989-06-30 - 2023-03-31	136	93	68.382	93	68.382	1.192000e+01	1.634000e+00	7.526000e+00	6.098000e+00	2.469000e+00
13	profit_margin	MATX.N	0 - Ren. & Replace 0	1989-06-30 - 2023-03-31	136	93	68.382	5	3.676	2.932900e+01	4.820000e-01	6.715000e+00	2.597600e+01	5.097000e+00
34	profit_margin	MATX.N	1 - Expand & Standard Na.N	1989-06-30 - 2023-03-31	136	93	68.382	5	3.676	2.932900e+01	4.820000e-01	6.715000e+00	2.597600e+01	5.097000e+00
10	reinvestment_rate	MATX.N	0 - Ren. & Replace 0	1989-06-30 - 2023-03-31	136	93	68.382	6	4.412	8.152700e+01	-4.172000e+00	8.808000e+00	2.087120e+02	1.444700e+01
31	reinvestment_rate	MATX.N	1 - Expand & Standard Na.N	1989-06-30 - 2023-03-31	136	93	68.382	6	4.412	8.152700e+01	-4.172000e+00	8.808000e+00	2.087120e+02	1.444700e+01

Report for ADP.OQ - Used: True

	Variable	Component	Stage	Period	Rows #	Missing Rows #	Missing Rows %	Missing Cells #	Missing Cells %	Max Value	Min Value	Mean Value	Variance	Standard Deviation
20	CPI	ADP.OQ	0 - Ren. & Replace 0	1989-06-30 - 2023-03-31	136	5	3.676	0	0.000	8.003000e+00	-3.560000e-01	2.770000e+00	2.450000e+00	1.565000e+00
41	CPI	ADP.OQ	1 - Expand & Standard Na.N	1989-06-30 - 2023-03-31	136	5	3.676	0	0.000	8.003000e+00	-3.560000e-01	2.770000e+00	2.450000e+00	1.565000e+00
62	CPI	ADP.OQ	2 - BRll Interpolation	1989-06-30 - 2023-03-31	136	0	0.000	0	0.000	8.003000e+00	-3.560000e-01	2.770000e+00	2.450000e+00	1.565000e+00
83	CPI	ADP.OQ	3 - Pct Change & Standard Na.N	1989-06-30 - 2023-03-31	136	1	0.735	0	0.000	8.003000e+00	-3.560000e-01	2.770000e+00	2.450000e+00	1.565000e+00
104	CPI	ADP.OQ	4 - Drop Na.N	1989-09-30 - 2023-03-31	135	0	0.000	0	0.000	8.003000e+00	-3.560000e-01	2.754000e+00	2.432000e+00	1.560000e+00
125	CPI	ADP.OQ	5 - Standardization	1989-09-30 - 2023-03-31	135	0	0.000	0	0.000	3.378000e+00	-2.001000e+00	-0.000000e+00	1.007000e+00	1.004000e+00
19	GDP	ADP.OQ	0 - Ren. & Replace 0	1989-06-30 - 2023-03-31	136	5	3.676	0	0.000	2.095269e+13	9.674724e+12	1.532368e+13	1.147626e+25	3.387664e+12
40	GDP	ADP.OQ	1 - Expand & Standard Na.N	1989-06-30 - 2023-03-31	136	5	3.676	0	0.000	2.095269e+13	9.674724e+12	1.532368e+13	1.147626e+25	3.387664e+12
61	GDP	ADP.OQ	2 - BRll Interpolation	1989-06-30 - 2023-03-31	136	0	0.000	0	0.000	2.095269e+13	9.674724e+12	1.532368e+13	1.147626e+25	3.387664e+12
82	GDP	ADP.OQ	3 - Pct Change & Standard Na.N	1989-06-30 - 2023-03-31	136	1	0.735	1	0.735	1.482000e+00	-7.050000e-01	5.750000e-01	2.070000e-01	4.550000e-01
103	GDP	ADP.OQ	4 - Drop Na.N	1989-09-30 - 2023-03-31	135	0	0.000	0	0.000	1.482000e+00	-7.050000e-01	5.750000e-01	2.070000e-01	4.550000e-01
124	GDP	ADP.OQ	5 - Standardization	1989-09-30 - 2023-03-31	135	0	0.000	0	0.000	2.002000e+00	-2.826000e+00	0.000000e+00	1.007000e+00	1.004000e+00
12	ROA	ADP.OQ	0 - Ren. & Replace 0	1989-06-30 - 2023-03-31	136	5	3.676	4	2.941	1.507000e+01	2.288000e+00	7.081000e+00	1.594000e+01	3.993000e+00
33	ROA	ADP.OQ	1 - Expand & Standard Na.N	1989-06-30 - 2023-03-31	136	5	3.676	4	2.941	1.507000e+01	2.288000e+00	7.081000e+00	1.594000e+01	3.993000e+00
54	ROA	ADP.OQ	2 - BRll Interpolation	1989-06-30 - 2023-03-31	136	0	0.000	0	0.000	1.507000e+01	2.288000e+00	7.242000e+00	1.633100e+01	4.041000e+00
75	ROA	ADP.OQ	3 - Pct Change & Standard Na.N	1989-06-30 - 2023-03-31	136	1	0.735	0	0.000	1.507000e+01	2.288000e+00	7.242000e+00	1.633100e+01	4.041000e+00
96	ROA	ADP.OQ	4 - Drop Na.N	1989-09-30 - 2023-03-31	135	0	0.000	0	0.000	1.507000e+01	2.288000e+00	7.202000e+00	1.624100e+01	4.030000e+00
117	ROA	ADP.OQ	5 - Standardization	1989-09-30 - 2023-03-31	135	0	0.000	0	0.000	1.959000e+00	-1.224000e+00	-0.000000e+00	1.007000e+00	1.004000e+00

Unification and Feature Selection

After downloading the data and generating a dataset consisting of the desired number of companies, it has the form of a dictionary of data tables, with one Pandas Dataframe allocated per company. During this data preparation phase, all data is consolidated by handling a multi-level index comprising the date and the corresponding name for each company.

Afterthat, it is possible to extract one of the companies available in the dataset. The chosen company will subsequently be utilized as a test case to evaluate the predictive capabilities of the model. Once the model is trained, it can be deployed to assess the probability of an increase or decrease in the company's stock returns. The data related to the selected company will be removed from the dataset and thus won't be included in the training set.

The extracted company from the used dataset is shown in the table below.

Company		Country	Date	Sector	Country Code	Income
Identifier						
WSO.N	Watsco Inc	United States of America(US)	1994-06-16	Industrials	US	1.480257e+08

The analysis proceeds by analyzing the available features from the standpoint of the variance inflation index (VIF) and correlation. Features with a variance inflation factor (VIF) score exceeding 5.0 are iteratively eliminated. After that, the remaining features are reduced based on the identified correlation values for each pair of features.

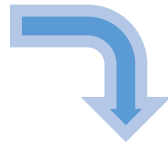
Lastly, it is possible to manually exclude any remaining features, by manually typing the name of the feature. The manually excluded features can be chosen, for instance, by considering the p-values that will be shown after the training of the model.

Having identified the features to utilize, several lagged versions of the dataset are then produced. These lagged versions can help the models capture autocorrelation in the data, where past observations might have an impact on the future ones. The creation of these lagged versions is managed on a company-by-company basis, ensuring that during data shifts, values aren't incorrectly assigned to other companies. The input for each model will hence be a dictionary of lagged versions of the dataset. Each version will be identified within the dictionary using the index of the applied lag.

Example: Pair Correlation Matrixes and VIFs before and after feature reduction.

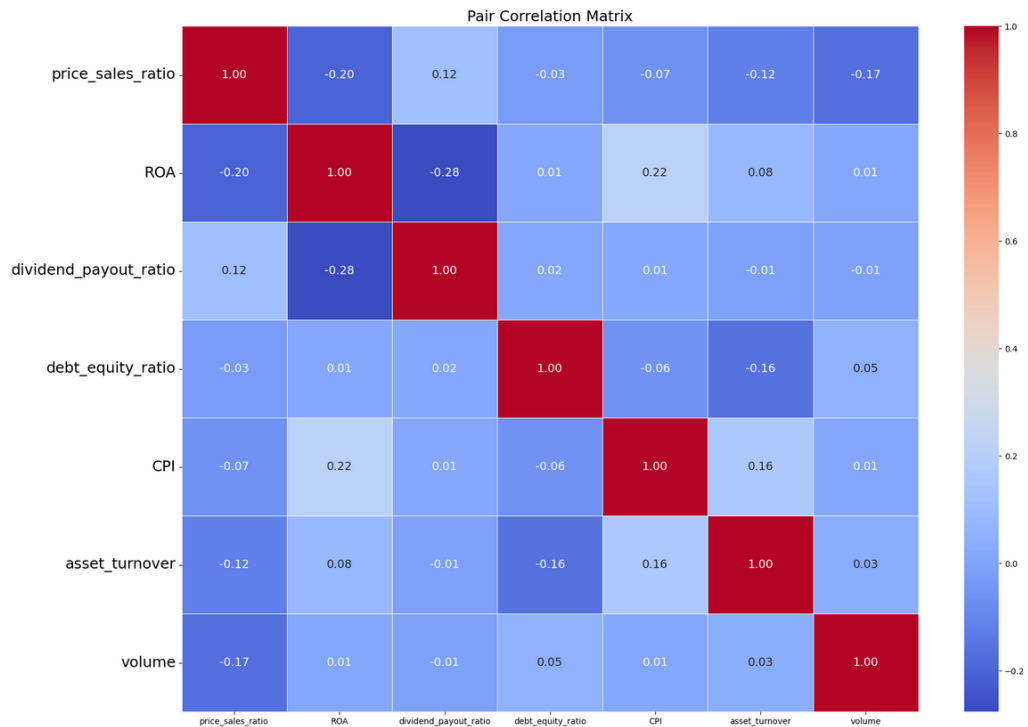
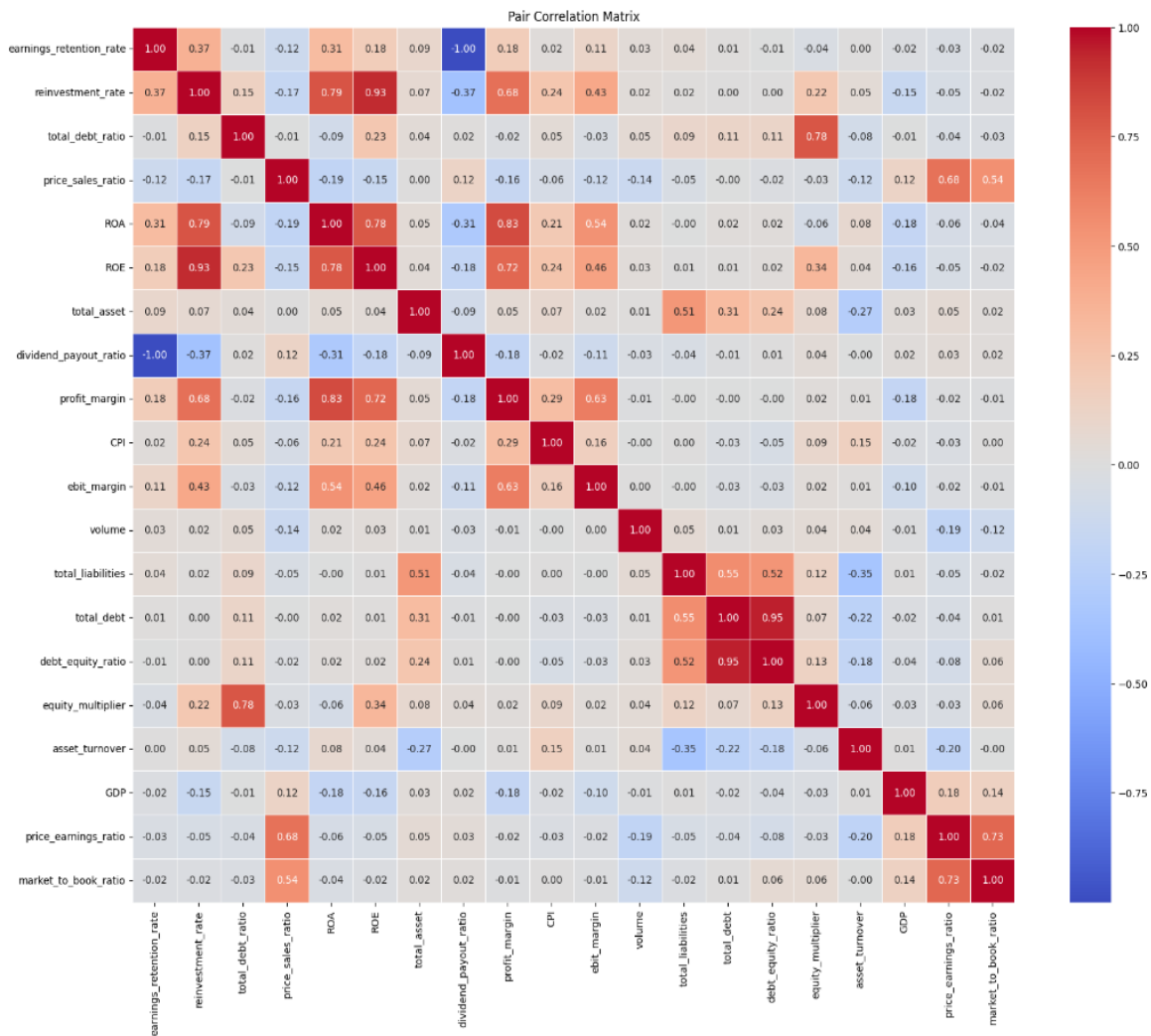
The two tables of this page display the VIF values for each feature in the used dataset. The table on the left calculates the VIF for each feature initially available in the dataset, while the table on the right demonstrates the VIF scores exclusively for the selected features.

	Features	VIF
0	earnings_retention_rate	71056.979590
1	reinvestment_rate	11.883414
2	total_debt_ratio	2.898919
3	price_sales_ratio	2.000865
4	ROA	5.889516
5	ROE	14.440953
6	total_asset	1.473899
7	dividend_payout_ratio	71143.794839
8	profit_margin	4.141751
9	CPI	1.167370
10	ebit_margin	1.666726
11	volume	1.047435
12	total_liabilities	1.907429
13	total_debt	13.928426
14	debt_equity_ratio	13.651235
15	equity_multiplier	3.975606
16	asset_turnover	1.370458
17	GDP	1.085799
18	price_earnings_ratio	3.526830
19	market_to_book_ratio	2.699915



	Features	VIF
0	ROA	1.2
1	dividend_payout_ratio	1.1
2	price_sales_ratio	1.1
3	CPI	1.1
4	debt_equity_ratio	1.0
5	volume	1.0
6	asset_turnover	1.1

The two matrices displayed on the next page represent the outcome of the feature reduction phase. The initial matrix show the correlation value for each pair of all the available features in the dataset. The second matrix shows the correlation for each pair of selected features. This last group of features is obtained filtering pairs with correlation greater than 0.4.



Models

This project uses two artificial intelligence models to estimate directional trends in stock returns. These two models are the Logit from Statsmodels library and a Multi-Layer Perceptron (MLP) neural network from the scikit-learn library.

The first method is an econometric approach, Logit model was chosen for its capabilities in predicting binary outcomes and its ability to offer probabilities, which can be crucial in decision-making processes. In contrast, the MLP excels in detecting non-linear patterns within the data, offering a versatile modeling approach. This ability of the MLP to capture complex relationships in data underscores the neural network's adaptability, especially in situations with intricate datasets like the stock market.

The subsequent sections detail tables that assess the performance of both the Logit and MLP models. The performance metrics used are accuracy, precision, recall, AUROC and specificity. These metrics evaluate multiple aspects of the model's effectiveness, from its overall accuracy to its ability to distinguish both positive and negative outcomes and the balance between true positive and false positive results across different thresholds.

Moreover, the table incorporates p-values for each predictor within the model. These p-values denote the statistical relevance of each predictor, with lower values suggesting that the predictor's observed impact is less likely due to random variability. This metric can also be useful to adjust the manually excluded features during the previous phase of multicollinearity analysis.

Econometric Analysis

Here are the results obtained from training the Logit model on the chosen dataset, which comprises American companies from the industrial sector. The following table contains the score for each model trained.

Lag	Split	Accuracy	Precision	Recall	AUROC	Specificity	p_val(ROA)	p_val(dividend_payout_ratio)	p_val(price_sales_ratio)	p_val(CPI)	p_val(debt_equity_ratio)	p_val(volume)	p_val(asset_turnover)
4	4	0.854000	0.861000	0.844000	0.854000	0.864000	0.003000	0.294000	0.000000	0.148000	0.012000	0.000000	0.978000
2	4	0.852000	0.851000	0.843000	0.852000	0.861000	0.008000	0.084000	0.000000	0.769000	0.014000	0.000000	0.857000
3	4	0.851000	0.848000	0.848000	0.851000	0.853000	0.004000	0.148000	0.000000	0.646000	0.013000	0.000000	0.895000
1	4	0.851000	0.846000	0.850000	0.851000	0.853000	0.005000	0.116000	0.000000	0.575000	0.011000	0.000000	0.958000
2	3	0.795000	0.816000	0.779000	0.796000	0.813000	0.158000	0.161000	0.000000	0.775000	0.041000	0.000000	0.883000
4	3	0.793000	0.828000	0.763000	0.795000	0.827000	0.086000	0.390000	0.000000	0.113000	0.041000	0.000000	0.864000
3	3	0.793000	0.820000	0.774000	0.794000	0.813000	0.134000	0.251000	0.000000	0.445000	0.030000	0.000000	0.888000
1	3	0.785000	0.810000	0.770000	0.786000	0.802000	0.122000	0.210000	0.000000	0.272000	0.027000	0.000000	0.781000
3	2	0.776000	0.820000	0.705000	0.776000	0.846000	0.206000	0.318000	0.000000	0.510000	0.414000	0.000000	0.390000
2	2	0.774000	0.820000	0.699000	0.773000	0.848000	0.261000	0.208000	0.000000	0.918000	0.415000	0.000000	0.404000
1	2	0.771000	0.822000	0.695000	0.772000	0.848000	0.273000	0.252000	0.000000	0.441000	0.400000	0.000000	0.408000
4	2	0.766000	0.810000	0.700000	0.767000	0.833000	0.159000	0.433000	0.000000	0.203000	0.402000	0.000000	0.397000
1	1	0.738000	0.761000	0.668000	0.736000	0.804000	0.499000	0.948000	0.000000	0.590000	0.354000	0.000000	0.590000
2	1	0.736000	0.754000	0.668000	0.733000	0.798000	0.449000	0.911000	0.000000	0.750000	0.363000	0.000000	0.608000
3	1	0.733000	0.754000	0.660000	0.730000	0.800000	0.347000	0.930000	0.000000	0.677000	0.357000	0.000000	0.585000
4	1	0.725000	0.750000	0.657000	0.724000	0.790000	0.233000	0.777000	0.000000	0.378000	0.356000	0.000000	0.606000

Comments

The table presents the performance metrics of a predictive model across different configurations. Each configuration is characterized by two main parameters: "Lag" (number of previous time points considered) and "Split" (dataset partitioning strategy). The performance of the model under each configuration is evaluated with the previously presented metrics.

The model exhibits the best performance at the fourth split of the fourth lag. This combination results in the highest accuracy of 85.4%, precision of 86.1%, and recall of 84.4%. The AUROC and Specificity values are 85.4% and 86.4%, respectively.

The variables "price_sales_ratio" and "volume" consistently have p-values close to 0 across all configurations, indicating their significant contribution to the predictive capability of the model. These predictors are likely to have a strong

association with the target variable. On the contrary, variables like “dividend_payout_ratio”, “CPI”, and “asset_turnover” have relatively higher p-values in most configurations. This suggests that these predictors might not be as impactful for the prediction.

In the first split of each lag there’s a noticeable drop in the performance metrics, suggesting that considering too few previous time points and using first data partition might not be ideal for this model. But even in configurations where the overall performance drops, the p-values for predictors “price_sales_ratio” and “volume” remain consistently low, re-emphasizing their significance.

In summary, the table provides comprehensive insights into the performance and behavior of the Logit model under various configurations. It highlights the key predictors driving the current model’s performance and suggests areas for further refinement.

Neural network analysis

The following table contains the scores of the MLP model, trained on the same dataset.

Lag	Split	Accuracy	Precision	Recall	AUROC	Specificity	p_val(ROA)	p_val(dividend_payout_ratio)	p_val(price_sales_ratio)	p_val(CPI)	p_val(debt_equity_ratio)	p_val(volume)	p_val(asset_turnover)
3	2	0.798000	0.803000	0.787000	0.798000	0.808000	0.121000	0.747000	0.000000	0.842000	0.028000	0.000000	0.560000
2	4	0.788000	0.773000	0.799000	0.788000	0.778000	0.317000	0.696000	0.000000	0.196000	0.000000	0.000000	0.062000
4	4	0.779000	0.774000	0.785000	0.779000	0.772000	0.566000	0.960000	0.000000	0.399000	0.000000	0.000000	0.087000
1	2	0.741000	0.764000	0.700000	0.741000	0.782000	0.124000	0.731000	0.000000	0.883000	0.030000	0.000000	0.540000
1	4	0.741000	0.714000	0.783000	0.742000	0.700000	0.532000	0.753000	0.000000	0.984000	0.000000	0.000000	0.055000
3	4	0.740000	0.733000	0.740000	0.740000	0.739000	0.408000	0.788000	0.000000	0.537000	0.000000	0.000000	0.066000
2	3	0.738000	0.747000	0.747000	0.738000	0.729000	0.056000	0.917000	0.000000	0.487000	0.010000	0.000000	0.122000
4	3	0.737000	0.754000	0.740000	0.737000	0.735000	0.120000	0.726000	0.000000	0.260000	0.007000	0.000000	0.182000
1	3	0.736000	0.746000	0.752000	0.735000	0.718000	0.113000	0.863000	0.000000	0.716000	0.005000	0.000000	0.112000
3	3	0.718000	0.734000	0.724000	0.718000	0.712000	0.057000	0.804000	0.000000	0.754000	0.006000	0.000000	0.138000
2	2	0.700000	0.724000	0.641000	0.700000	0.758000	0.113000	0.634000	0.000000	0.539000	0.034000	0.000000	0.531000
1	1	0.700000	0.701000	0.663000	0.699000	0.735000	0.222000	0.642000	0.000000	0.980000	0.054000	0.000000	0.097000
4	2	0.689000	0.723000	0.618000	0.689000	0.760000	0.183000	0.843000	0.000000	0.307000	0.026000	0.000000	0.649000
2	1	0.650000	0.644000	0.609000	0.648000	0.688000	0.230000	0.752000	0.000000	0.704000	0.059000	0.000000	0.096000
3	1	0.639000	0.621000	0.640000	0.639000	0.637000	0.222000	0.700000	0.000000	0.654000	0.056000	0.000000	0.101000
4	1	0.625000	0.622000	0.597000	0.625000	0.652000	0.366000	0.615000	0.000000	0.709000	0.056000	0.000000	0.134000

Comments

This model uses a Multi-Layer Perceptron (MLP) with two hidden layers, each containing 30 neurons. The activation function 'ReLU' (Rectified Linear Unit) was chosen due to its efficiency and ability to handle non-linearity. The 'lbfgs' solver is an optimizer preferred for smaller datasets (like this one). The learning rate adjustment is set to "adaptive" allowing the model to change the learning rate during training for better convergence. Shuffling is turned off, meaning the training data order remains consistent across iterations.

The table presents the evaluation metrics of MLP model across different lags and splits.

The model's performance varies across configurations, with accuracy ranging between 62.5% to 79.8%. The model appears to have its best performance in terms of accuracy, precision, recall, AUROC, and specificity at the second split of the third lag. This implies that considering three previous time points and using the fourth split provides the most reliable results. However, it's notable that there's not a drastic change in the performance metrics. The difference in accuracy, for instance, is about 17% between the best and worst score, suggesting that the model has a relatively consistent predictive capability across different settings and a good level of robustness. Let's now consider the predictors' significance.

- "ROA": the significance of this predictor is inconsistent across configurations, with p-values ranging from 0.056 to 0.566.
- "dividend_payout_ratio": overall, this predictor does not display strong predictive power due to its consistently high p-values.
- "price_sales_ratio": this predictor show excellent predictive power across all configurations, also evidenced by low p-values.
- "CPI": despite having high p-values in some configurations, the Consumer Price Index (CPI) showcases significant predictive power in several model settings.
- "debt_equity_ratio": This predictor is also statistically significant, confirmed by its low p-value.
- "Volume": this predictor also demonstrates strong predictive capabilities across all configurations, with consistently low p-values.
- asset_turnover: the significance of this predictor is somewhat varied, but it is evident that it is significant in several configurations.

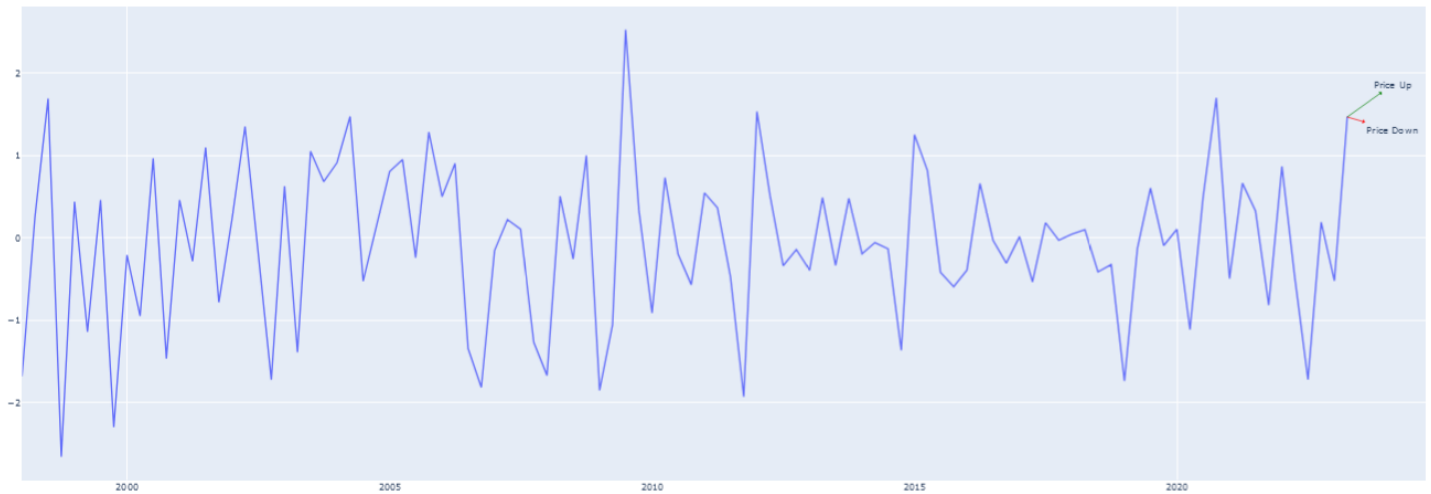
In the context of the project, certain predictors consistently demonstrate strong predictive power, such as `price_sales_ratio`, `debt_equity_ratio`, and `volume`. Others, like `dividend_payout_ratio`, seem less significant to the model. This analysis implies that for optimal results, stakeholders should consider the Lag 3 and Split 2 configuration.

For future model refinement or feature selection, it may be worth investigating predictors with inconsistent significance levels more deeply, either by gathering more data or trying different metrics and ratios.

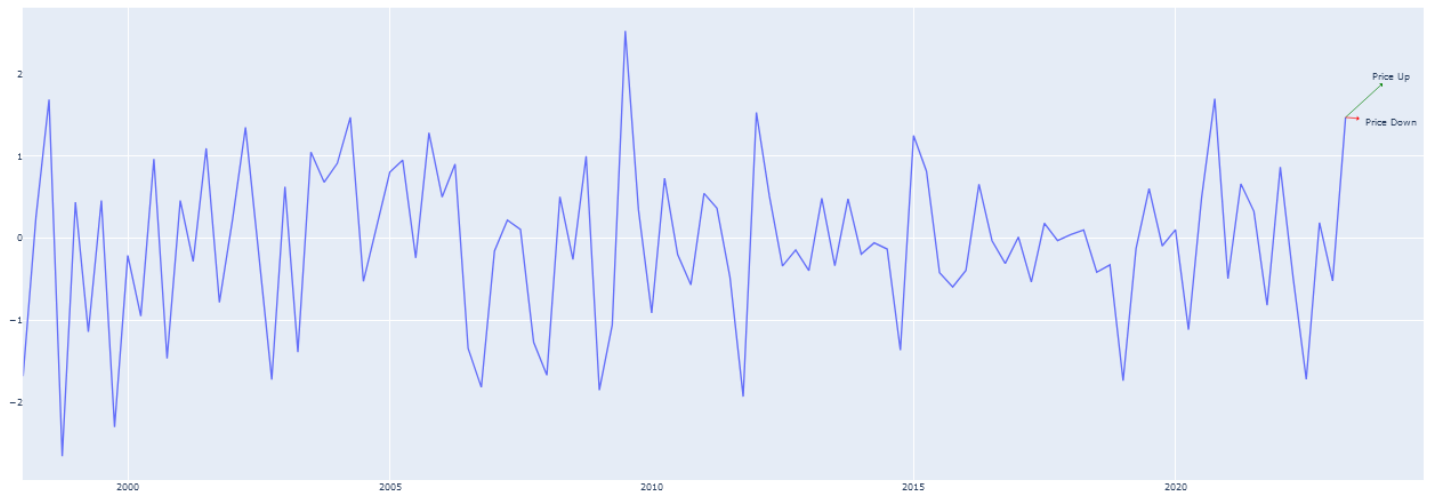
Example: predictive capabilities

Considering the chosen tester from the used dataset, the prediction of the model for the next future quarter (next intended as the first future quarter not available in the dataset) is shown in Figure. The first figure show the average prediction of logit models, the second figure show the average prediction with MLP.

Price Up Probabilities: 73.88% - Price Down Probabilities: 26.12%



Price Up Probabilities: 93.75% - Price Down Probabilities: 6.25%



Conclusion

In the context of this project, which aims to use AI models to predict stock returns using data sourced from Refinitiv Eikon and World Bank Data, both the Logit model and the Multi-Layer Perceptron (MLP) model have demonstrated quite good performance.

The Logit model works well with certain factors like “price_sales_ratio” and “volume”. These factors had a strong influence on predictions. This model did best with the fourth lag and the fourth split of data, reaching an accuracy of 85.4%. But some factors, like “dividend_payout_ratio”, “CPI”, and “asset_turnover” didn’t seem to help much. Also, there’s about a 17.3% difference between the model’s best and worst accuracy scores, showing that it’s quite stable.

The MLP model was steady across different setups. Its accuracy scores also have a low variance, with about a 17% difference from best to worst. Some factors, like “price_sales_ratio”, “debt_equity_ratio”, and “volume”, were important for it, just like in the Logit model. But there are differences too, like the “dividend_payout_ratio” factor being more useful in the Logit model than in MLP.

The two models gave different views. For instance, the Logit model valued the “dividend_payout_ratio” a lot, but the MLP model didn’t. This shows that financial data can be complex and different models can see it differently.

In conclusion, the project demonstrates that both the Logit model and the MLP model can be useful tools for predicting stock returns. By using both, it’s possible to make a little bit better decision about stock investment.

