Machine Learning Model

Problem Statement

This machine learning project aims to develop a binary classification model to predict positive short-term (weekly) stock price movements (positive weekly return) for BHP (ASX:BHP). Our objective is to develop a model that has an economically viable predictive ability with respect to forming an investment strategy based on the models predictions, such that a trading strategy based on this model would generate a positive return over the back test in addition outperforming the underlying stock over the test period.

Data Availability

BHP stock price data is available from numerous sources. For reproducibility, this report sources stock price data from Yahoo Finance. These sources quote the open, high, low and closing prices of BHP stock returns (ASX:BHP) from as early as the 31st of December 1998 to the current date. Given that BHP was initially listed in 1985, the three-year gap amounts to a negligible timeframe for which data is inaccessible. Prices sourced form Yahoo finance are limited to one-minute intervals over that same period.

Using data from the date of listing to the present, we are limited to a maximum sample size of 2,366,715 returns over one-minute intervals or:

- 45,891 hourly returns
- 6,556 daily returns
- 1,342 weekly returns
- 308 monthly returns
- 103 quarterly returns

Sampling

At face value, return intervals less than one month are non-restrictive. Monthly and quarterly returns may be restrictive for some machine learning techniques, however there are there are other constraints, including material changes in company structure that may reduce the comparability of historical returns relative to future returns. This is discussed further in sect x.

Evaluation Metrics Considered

Accuracy measures the proportion of correctly predicted uptrends in the total predictions. High accuracy will indicate an overall effectiveness in classifying uptrends.

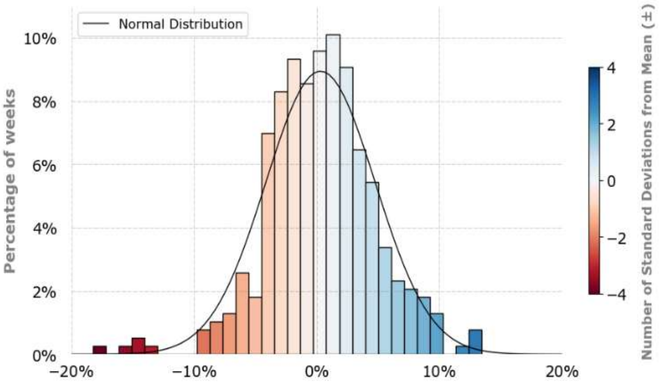
Precision focuses on the proportion of true positives out of all positive predictions. High precision is crucial to ensure only confident uptrend predictions are made.

F1 Score balances precision and recall, especially useful if a trade-off exists between these metrics.
Backtesting Profits: Although not a traditional classification metric, backtesting the model’s predictions in a simulated trading environment will evaluate its real-world utility. Success is measured by the net returns generated from trading based on the model's uptrend signals.

Recall (Sensitivity) assesses the proportion of actual uptrends that are correctly identified. Despite it's irrelivance in-sample, itallows for a clopudy yet useful comparison with other models developed on different stocks.

Figure 1: Positive skew and considerable tail risk in BHP returns

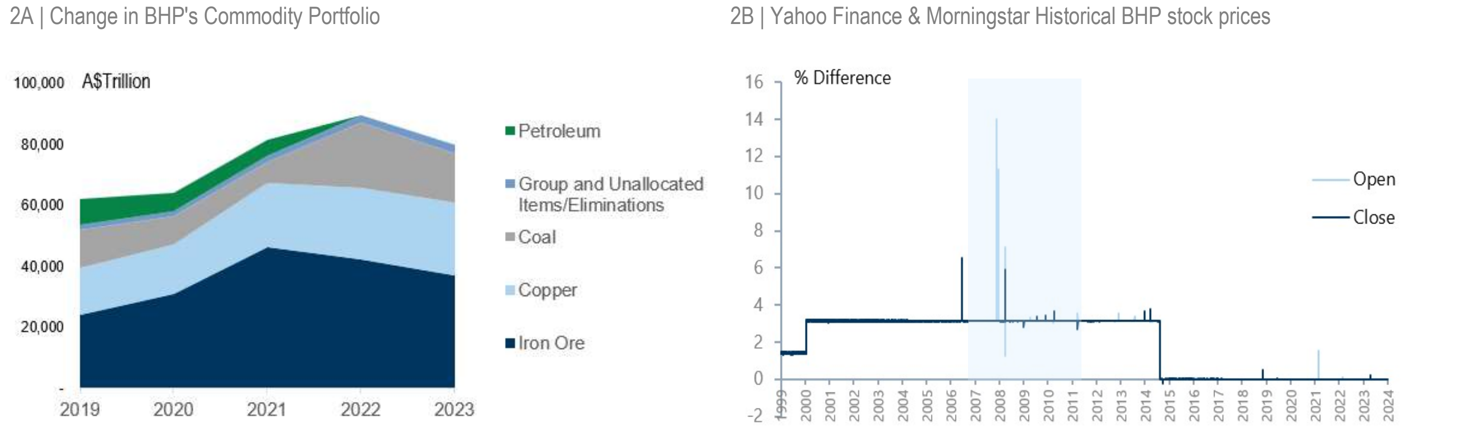
1A | BHP.AX weekly stock return distribution, 2019 - 2024



Source: Own Analysis, Stock Price Data from Yahoo finance

Data constraints and solidified evaluation criteria

Figure 2: BHP Portfolio and Data Reliability



2A | Change in BHP's Commodity Portfolio

2B | Yahoo Finance & Morningstar Historical BHP stock prices

Source: Statistics Canada

Reliability of Data

Reliance on Yahoo Finance stock price data is subject to errors. Yahoo finance sources data from third parties and is not a broker, qualities that raise concern over the reliability of this data. Using non-public data sourced from Morningstar Premium, we compare the open and closing prices and display the variance between the price quoted between the two sources below. We acknowledge that each provider has their own methodology for reporting their open and closing prices which may explain some of the consistent variability in the years before 2015. Its clear that the degree of variance drops significantly after 2015, albeit with some minor deviations. Whilst a comparison of two providers is still subject to uncertainty, a more robust comparison would be preferable in the absence of data limitations.

This raises an important question; are the determinant of BHP's returns the same today as it was in 2015? There are foreseeable differences if we assume that the companies underlying performance influences the stock price. Ie. Petroleum prices are unlikely to have any material influence on BHP's stock price beyond 2022. If left unaccounted for, there are likely to be significant differences between training and test performance.

Commodity production would provide an intuitive solution to this problem, however production figures are reported quarterly. There is therefore a trade-off between return sample size and our ability to include these variables – with both likely to have an unknown but significant impact on the model's predictive power.

Evaluation Criteria

Given the objective of accurately predicting positive returns in BHP's stock price, where the theoretical application of each model is to act on positive predictions alone, success will be primarily measured by the precision of the classification models estimates. Precision refers to the proportion of True positive predictions made relative to the total positive predictions made, thereby providing an indication of how reliable each positive prediction is. This metric also allows for comparative performance measurement between several different models. Despite addressing the objective of this project, it fails to account for practical utility within real world applications.

We also consider the recall of the classification model; that is, the proportion of true positive predictions among all actual positive observations in the dataset. This metric accounts for several important factors: namely the useability of the classification model and the models predictive performance relative to the underlying stocks performance. Intuitively, a model's precision can be inflated merely by changing the decision boundary.

Performance Metrics

Performance metrics: combination of both precision and recall

-F beta Score
-F1 score (beta=1)
-F2 score (beta = 2)

$$F_{beta} = (1+\beta^2)\frac{precision * recall}{\beta^2 * precision + recall}$$

Data Availability

The F1 and F2 scores offer standardized beta values that indicate the relative importance of recall and precision in model evaluation. In contrast, the F beta score allows for a customizable definition of relative importance, which can vary based on specific use cases. For instance, in investment decision-making, a model exhibiting high precision but low recall may be less valuable when considered in isolation; however, aggregating predictions from multiple models across a broader stock sample can alleviate this limitation. The critical takeaway is that low recall does not inherently imply low utility, whereas low precision does. Additionally, it is essential to consider scenarios with few positive predictions, as these can skew performance metrics on the test set.

Constraints

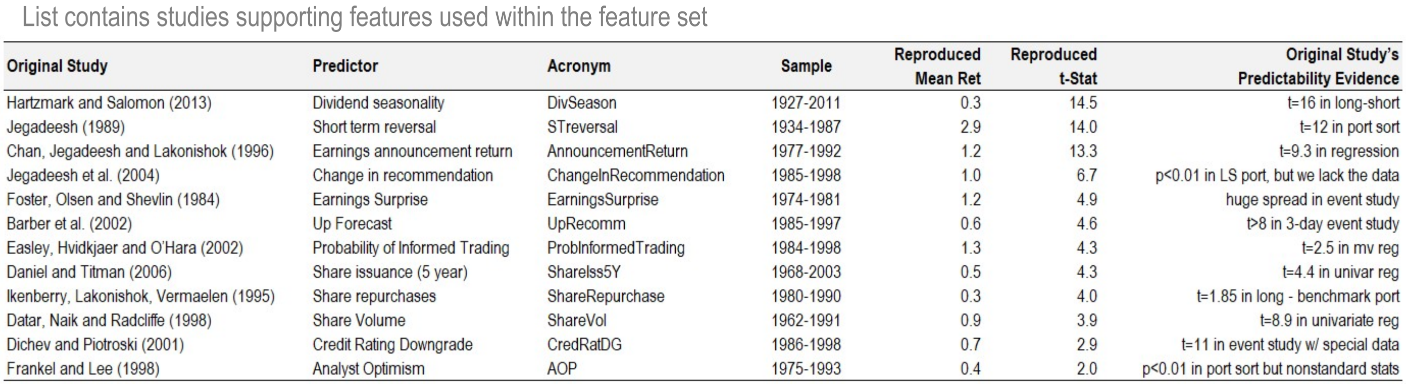
Overfitting primarily pertains to the training dataset, occurring when a model is excessively flexible and captures both signal and noise. This leads to inaccurate predictions and poor performance on out-of-sample data, even during backtesting. To mitigate overfitting, techniques such as cross-validation for assessing generalization error, regularization to manage parameter influence, and ensemble methods to reduce variance are recommended.

Conversely, test set overfitting arises when researchers iteratively modify a strategy during backtesting to achieve favorable results. This iterative process can result in misleadingly positive outcomes, indicating a potentially flawed research methodology. Researchers should critically evaluate their processes to identify any biases that may have influenced their strategy testing. Adjusting performance metrics, such as using an adjusted Sharpe ratio, and employing Monte Carlo simulations to create representative datasets are effective strategies for addressing this issue.

Moreover, many machine learning models necessitate that features are scaled consistently for optimal performance, typically achieved through normalization or standardization. High correlation among features may render some redundant, and dimensionality reduction techniques can effectively condense these features into a lower-dimensional space. This reduction not only decreases storage requirements but also enhances computational efficiency. In scenarios where datasets contain numerous irrelevant features, dimensionality reduction can improve predictive performance, especially in cases with a low signal-to-noise ratio.

Data Collection

Figure 3: Asset Pricing Research: statistically significant features



List contains studies supporting features used within the feature set

Original Study	Predictor	Acronym	Sample	Reproduced Mean Ret	Reproduced t-Stat	Original Study's Predictability Evidence
Hartzmark and Salomon (2013)	Dividend seasonality	DivSeason	1927-2011	0.3	14.5	t=16 in long-short
Jegadeesh (1989)	Short term reversal	STreversal	1934-1987	2.9	14.0	t=12 in port sort
Chan, Jegadeesh and Lakonishok (1996)	Earnings announcement return	AnnouncementReturn	1977-1992	1.2	13.3	t=9.3 in regression
Jegadeesh et al. (2004)	Change in recommendation	ChangeInRecommendation	1985-1998	1.0	6.7	p<0.01 in LS port, but we lack the data
Foster, Olsen and Shevlin (1984)	Earnings Surprise	EarningsSurprise	1974-1981	1.2	4.9	huge spread in event study
Barber et al. (2002)	Up Forecast	UpRecomm	1985-1997	0.6	4.6	t>8 in 3-day event study
Easley, Hvidkjaer and O'Hara (2002)	Probability of Informed Trading	ProbInformedTrading	1984-1998	1.3	4.3	t=2.5 in mv reg
Daniel and Titman (2006)	Share issuance (5 year)	ShareIss5Y	1968-2003	0.5	4.3	t=4.4 in univar reg
Ikenberry, Lakonishok, Vermaelen (1995)	Share repurchases	ShareRepurchase	1980-1990	0.3	4.0	t=1.85 in long - benchmark port
Datar, Naik and Radcliffe (1998)	Share Volume	ShareVol	1962-1991	0.9	3.9	t=8.9 in univariate reg
Dichev and Piotroski (2001)	Credit Rating Downgrade	CredRatDG	1986-1998	0.7	2.9	t=11 in event study w/ special data
Frankel and Lee (1998)	Analyst Optimism	AOP	1975-1993	0.4	2.0	p<0.01 in port sort but nonstandard stats

Source: *Chen, Zimmerman 2020 Journal of Financial Economics*

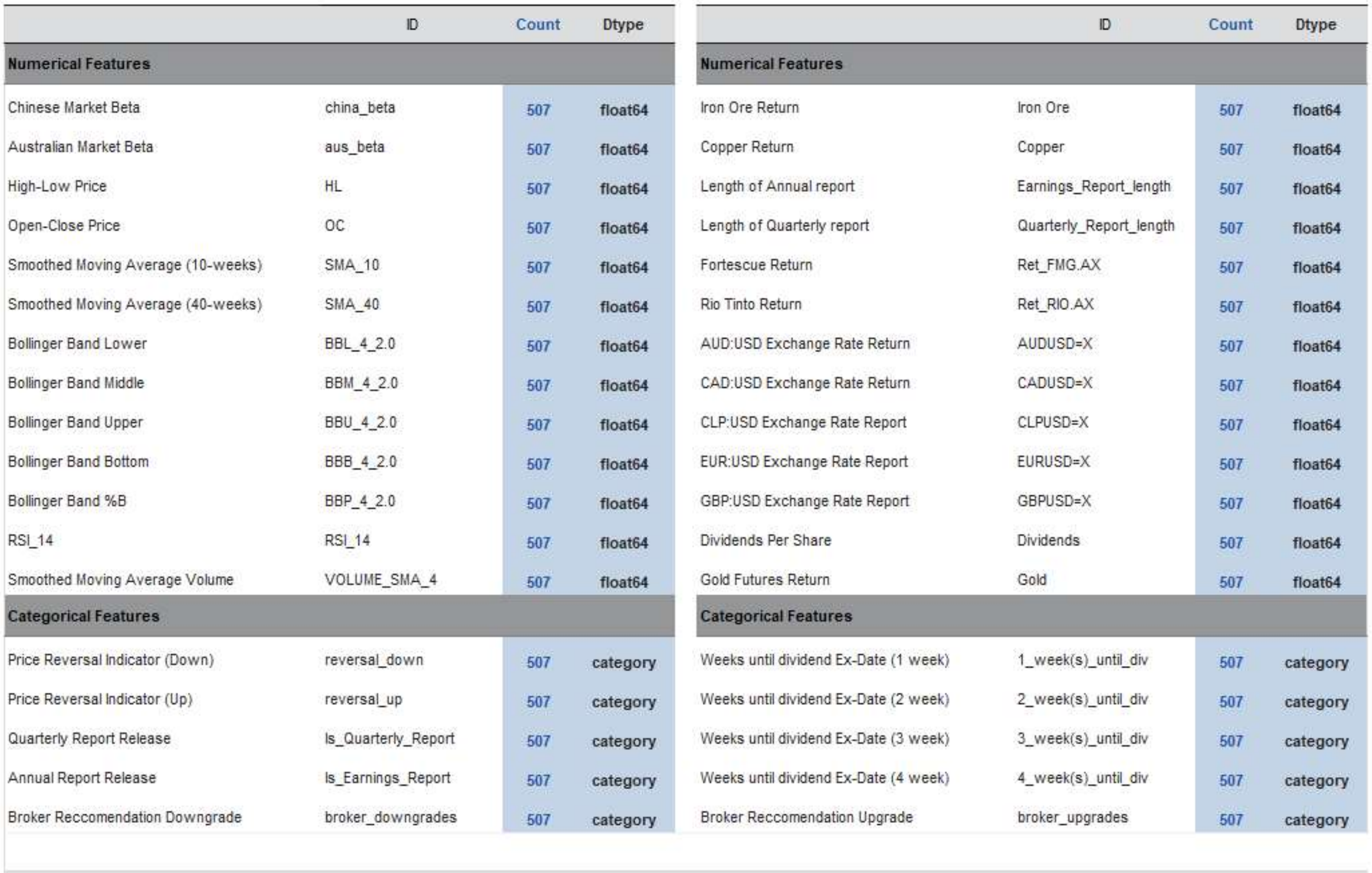
Empirical Evidence

The literature as it stands provides broad coverage in the context of factors that explain stock price movements. More often than not, this research is applied to a basket of securities from which portfolios are formed or market indices. There is a lack of single stock coverage particularly in the ASX market. The applicability of this research when applied to a single company is diminished and is at a high risk of redundancy should the stock be delisted or undergo material transformation. Continuity and breadth of applicability is out of scope for this project, so we do not rely heavily on the outstanding literature in our determination of firm-specific factors. Instead, we follow a rational approach using domain knowledge to incorporate industry and company specific nuances that in our view, have predictive capacity.

The following factors are those that have either been documented to have some predictive power in the empirical literature or, factors that on a reasonable basis, have the capacity to explain some of the variation in BHP's stock returns.

Chen and Zimmermann (2020) replicate studies on 319 cross-sectional stock return predictors from earlier meta-analyses, providing accompanying code and a comparison of t-statistics with the results from the original studies. For the factors that were statistically significant in both the original and reproduced analyses, we evaluate the variability in t-statistics, the magnitude of statistical significance, and the relevance of sample periods, particularly in relation to BHP. The included predictors meet all these criteria.

Feature Summary



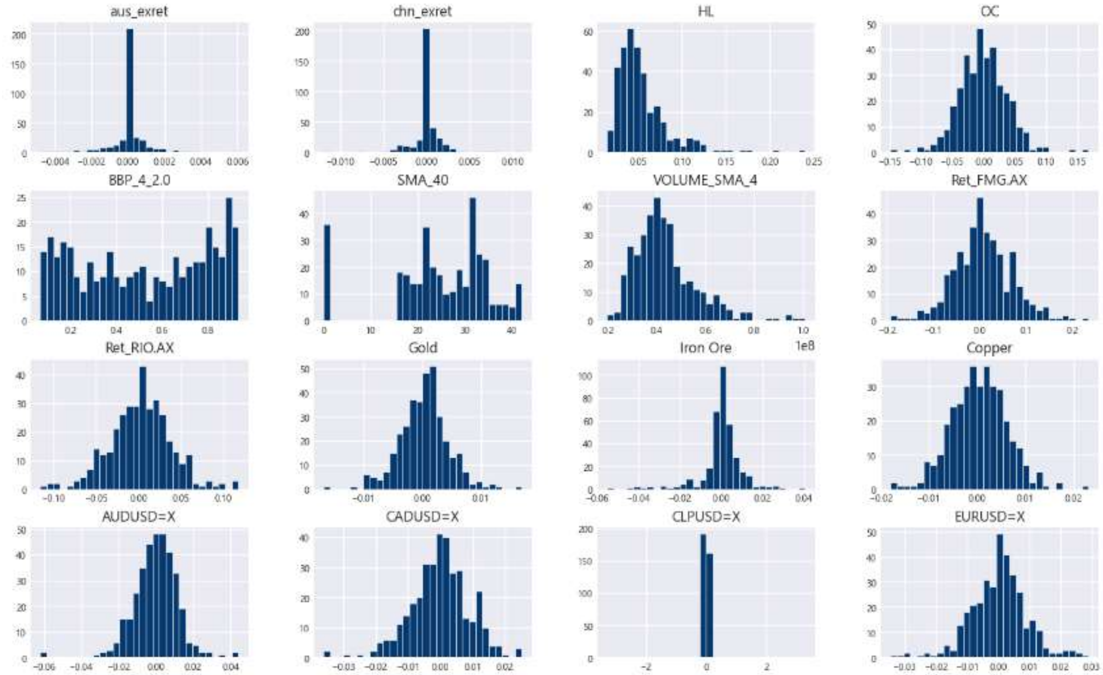
	ID	Count	Dtype
Numerical Features			
Chinese Market Beta	china_beta	507	float64
Australian Market Beta	aus_beta	507	float64
High-Low Price	HL	507	float64
Open-Close Price	OC	507	float64
Smoothed Moving Average (10-weeks)	SMA_10	507	float64
Smoothed Moving Average (40-weeks)	SMA_40	507	float64
Bollinger Band Lower	BBL_4_2.0	507	float64
Bollinger Band Middle	BBM_4_2.0	507	float64
Bollinger Band Upper	BBU_4_2.0	507	float64
Bollinger Band Bottom	BBB_4_2.0	507	float64
Bollinger Band %B	BBP_4_2.0	507	float64
RSI_14	RSI_14	507	float64
Smoothed Moving Average Volume	VOLUME_SMA_4	507	float64
Categorical Features			
Price Reversal Indicator (Down)	reversal_down	507	category
Price Reversal Indicator (Up)	reversal_up	507	category
Quarterly Report Release	Is_Quarterly_Report	507	category
Annual Report Release	Is_Earnings_Report	507	category
Broker Reccomendation Downgrade	broker_downgrades	507	category

	ID	Count	Dtype
Numerical Features			
Iron Ore Return	Iron Ore	507	float64
Copper Return	Copper	507	float64
Length of Annual report	Earnings_Report_length	507	float64
Length of Quarterly report	Quarterly_Report_length	507	float64
Fortescue Return	Ret_FMG.AX	507	float64
Rio Tinto Return	Ret_RIO.AX	507	float64
AUD:USD Exchange Rate Return	AUDUSD=X	507	float64
CAD:USD Exchange Rate Return	CADUSD=X	507	float64
CLP:USD Exchange Rate Report	CLPUSD=X	507	float64
EUR:USD Exchange Rate Report	EURUSD=X	507	float64
GBP:USD Exchange Rate Report	GBPUSD=X	507	float64
Dividends Per Share	Dividends	507	float64
Gold Futures Return	Gold	507	float64
Categorical Features			
Weeks until dividend Ex-Date (1 week)	1_week(s)_until_div	507	category
Weeks until dividend Ex-Date (2 week)	2_week(s)_until_div	507	category
Weeks until dividend Ex-Date (3 week)	3_week(s)_until_div	507	category
Weeks until dividend Ex-Date (4 week)	4_week(s)_until_div	507	category
Broker Reccomendation Upgrade	broker_upgrades	507	category

Note: *Methodology and rationale for feature selection in appendix section 1.*

Figure 4: Distribution of Categorical Variables

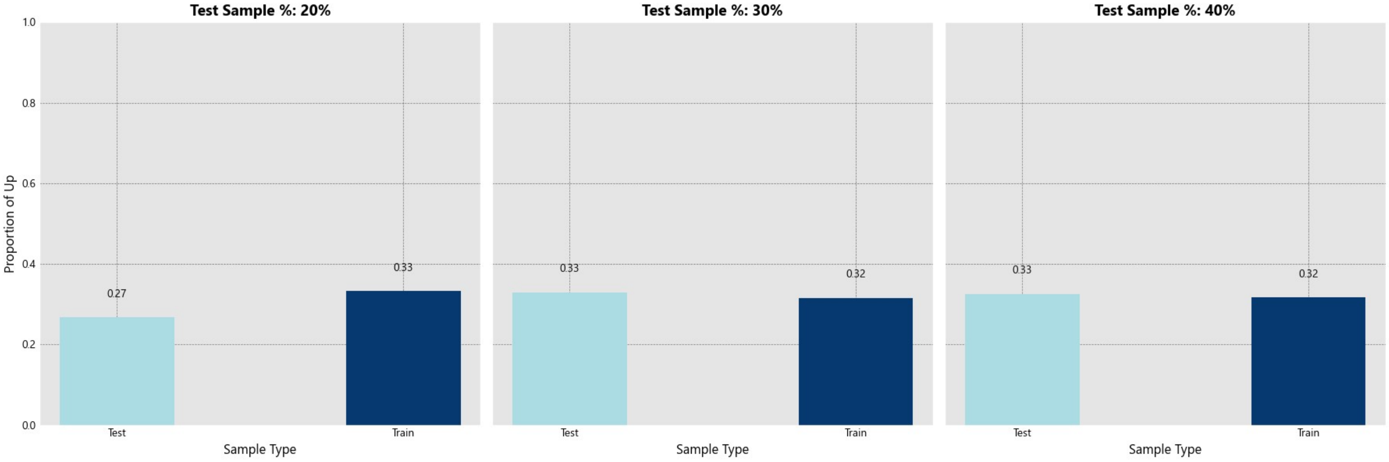
Test set categorical variables



Data Preprocessing

Figure 5: Proportion of positive return weeks in 3 different sample size %

Percentage of observations within each test set split: % total, 2015-2024



Source: Own Research, Yahoo Finance

1. Train and Test Set Split: 70/30

Given that there is no universal rule as to the best train and test split ratio, we instead consider test sample size proportions of 20%, 30% and 40% and select that which contains the closest to an even distribution of classes. Whilst this introduces lookahead bias from an implementation standpoint, it allows for direct comparisons to be drawn when evaluating the model. Note that whilst we attempt to balance the proportion of target classes, the cumulative returns are still unknown.

2. Train & Validation Set Split: 70/30

The rationale for this split was to remain consistent with the Train / Test split to approximate the expected variance in the Train / Test samples.

3. Feature selection and extraction to reduce dimensionality

Note this is only applied to the training set

Each of the 36 Features are scaled via:

- Normalisation - Min/Max [0,1]
- Standardisation - $\mu = 0$ & $\sigma = 1$

Reducing Dimentionality over training set

We beign by fitting a Logistic Regression model to the training set to compare in sample performance with out of sample (validation set) performance. We then implement two sequential feature selection algorithms to assess feature importance with the objective of reducing dimentionality.

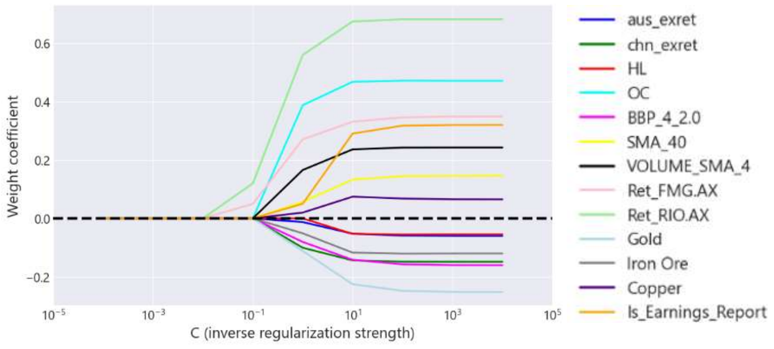
4. Fitting a preliminary logistic regression model

Logistic regression model to the training and validation set and compare the in-sample performance with the out of sample (validation) performance.

RESULTS: Training Precision: 0.884

Validation Precision: 0.714

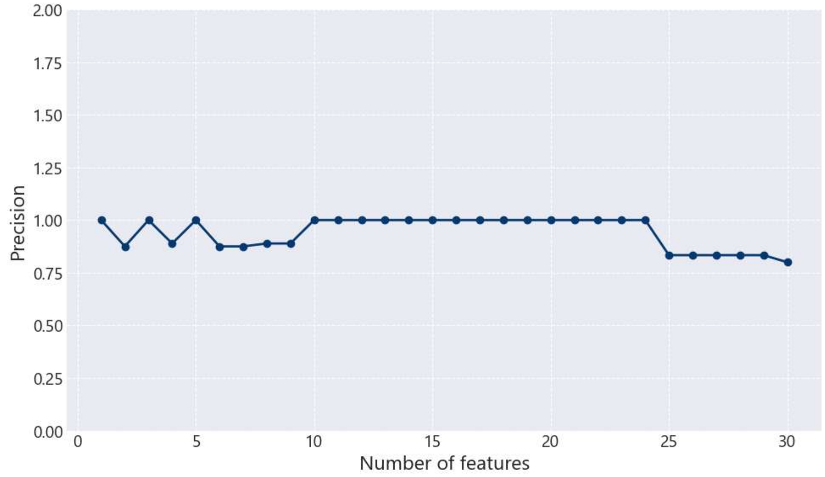
The precision rates are very high and only differ marginally. The difference does not inherantly suggest the model is overfitting.



Data Preprocessing

Figure 6:

Percentage of observations within each test set split: % total, 2015-2024



Source: Own Research, Yahoo Finance

1. Train and Validation Set Split: 70/30

Note - previously stated: Given that there is no universal rule as to the best train and test split ratio, we instead consider test sample size proportions of 20%, 30% and 40% and select that which contains the closest to an even distribution of classes. For the validation set, we adopted the same proportions for consistency.

2. Select minimum value of K to maximises precision

The mdoel is re-estimated for the subset of K that yields to highest precision, and we compare the out of sample performance on the validation set with the base 30 param model.

All features:

Training precision: 0.759

Validation precision: 0.783

Best Subset - K = 9:

Training precision: 0.807

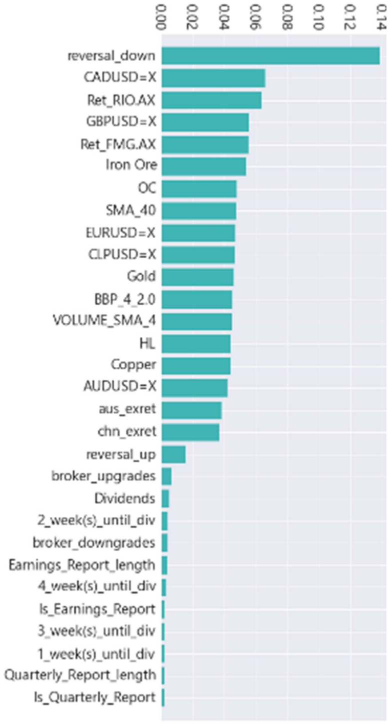
Validation precision: 0.764

In the preceding code section, we used the complete feature set and obtained 75.9 percent precision on the training dataset and approximately 76.4 percent precision on the validation dataset, which indicates that our model already generalizes well to new data. it also suggests that there is very little explainitory power in the other variables not included in the best subset.

3. Assessing feature importance with Random Forest

Using a random forest, we can measure the feature importance as the averaged impurity decrease computed from all decision trees in the forest, without making any assumptions about whether our data is linearly separable or not. As a result we fit this model on the `entire training set`

Figure 7: Feature importance scores

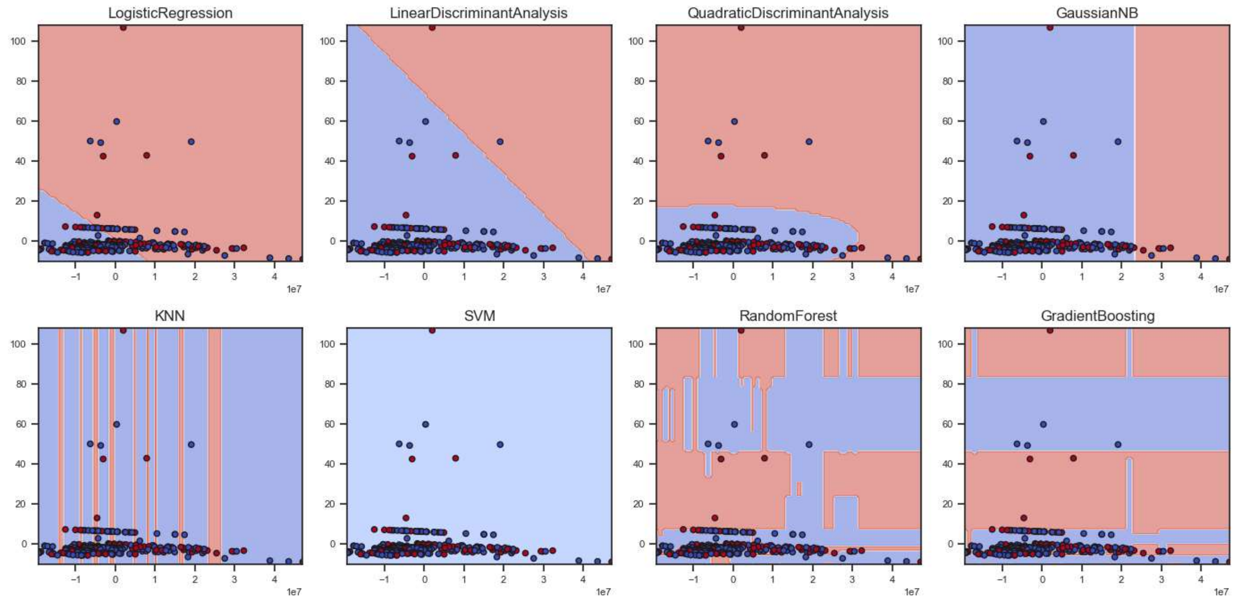


Note: *Feature importance values are normalized so that they sum up to 1.0*

Exploratory Analysis

Figure 8: Descision Boundary of Various Classification Models - 2D

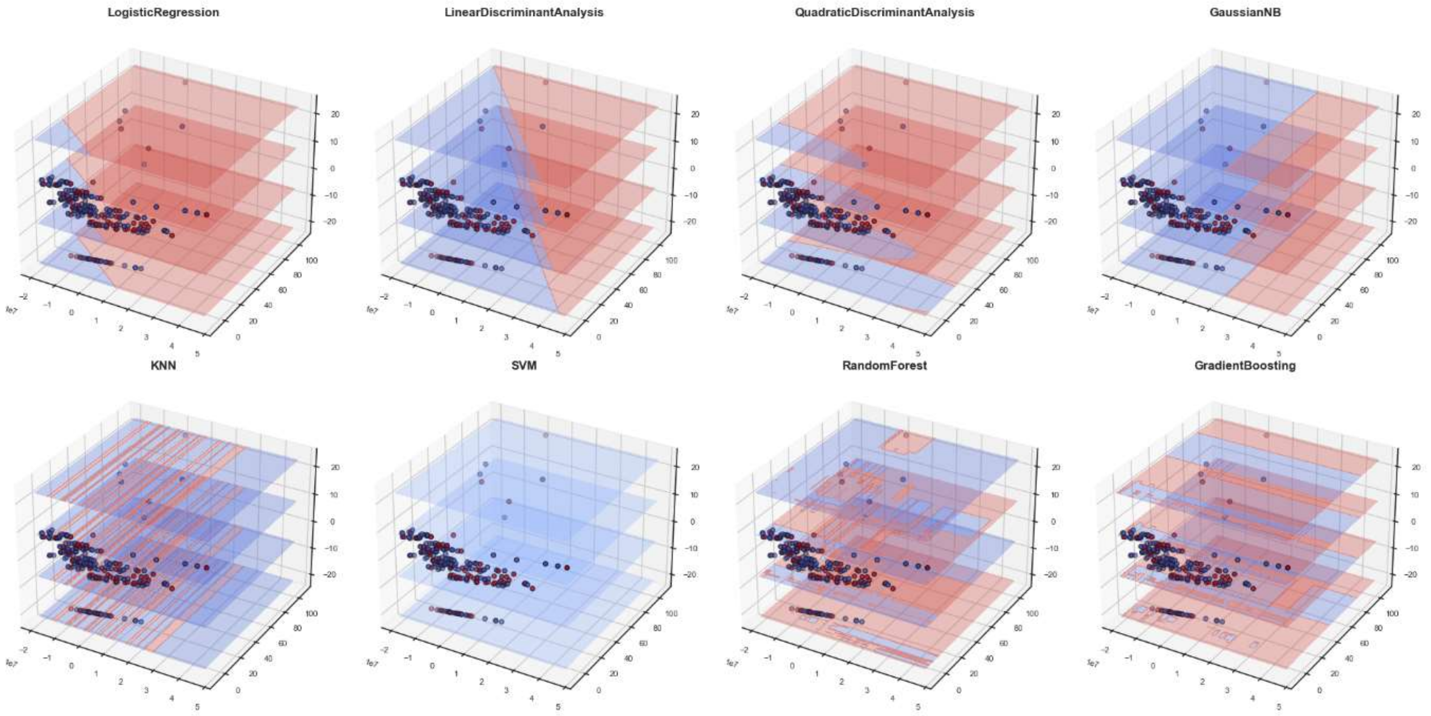
2 Dimentional Principal Component Analysis (2) on Test + Validation Set, BHP



Source: Own Research, Yahoo Finance

Figure 9: Descision Boundary of Various Classification Models - 3D

3 Dimentional Principal Component Analysis on Test + Validation Set, BHP



Source: Own Research, Yahoo Finance

Algorithm Selection & Implementation

Recall that our objective is to accurately predict positive weekly returns in BHP's stock price. This is an example of a binary classification problem in machine learning. The following algorithms can be used in binary classification problems. Given the level of dimenionality we observed in the previous section, it is likely that support vector machines and random forrest models will suit the characteristics of the dataset. Note that these two models are very computationally expensive relative to the other models. Given that the dataset is relatively small (510 observations), there is very little sacrificed to capture the bennefit of these models. In practice, the SVM model is the most computationally expensive with a runtime of roughly 10 minites. Note that we apply all models to the test set for comparative performance measurement.

Binary Classification Models

1 - Logistic regression

2 - Linear Discriminant Analysis

3 - Quadratic Discriminant Analysis

4 - Gaussian Naive Bayes

5 - K Nearest Neighbours

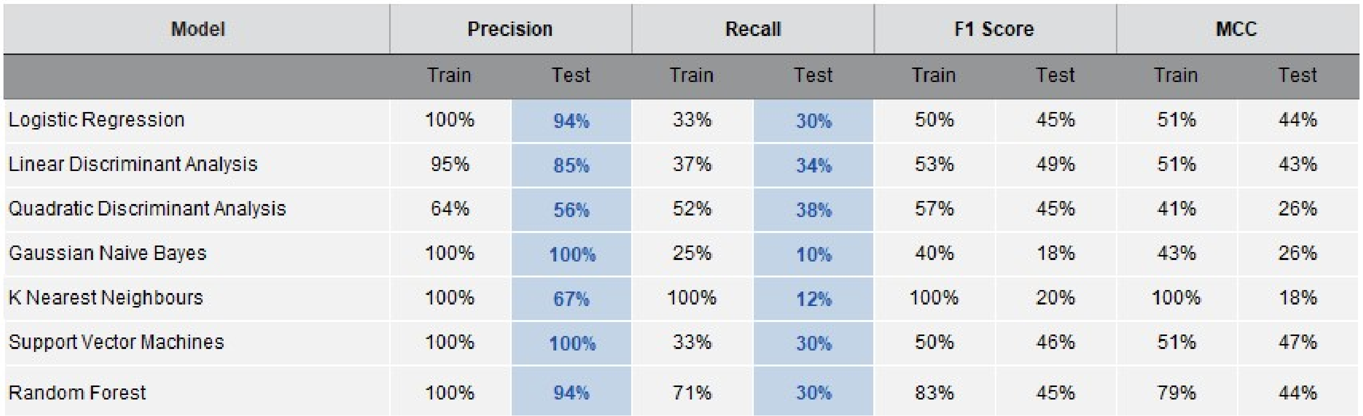
6 - Support Vector Machines

7 - Random Forrest

The table below shows the performance of the different machine learning models. Most models exhibit high precision but struggle with recall on the test set, indicating that while they accurately identify positive weekly returns, they miss a significant number of them (false negatives). Both Logistic Regression, Random Forrest and SVM achieve high precision (94-100%) but have low recall (30-34%). Random Forest, while having one of the highest recall % on the training set (71%), suffers a drop to 30% on the test set, suggesting the model has overfit.

Given our objective, the SVM model has exhibited the highest precision over both training and test sets (100%) and based on the performance over the test set, is our preffered model. Both KNN and Random forrest models see a significant decline of precision suggesting that they have overfit. On the basis of precision alone, SVM is undoutebly the most highly performing model. Whilst there is a sacrifice of precision, in practice, we could reduce the confidence threshold which may result in a worthwhile tradeoff.

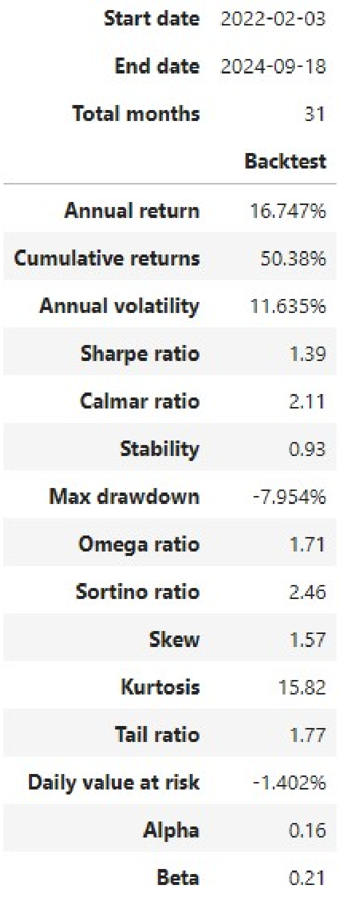
Figure 8: Test Set Results



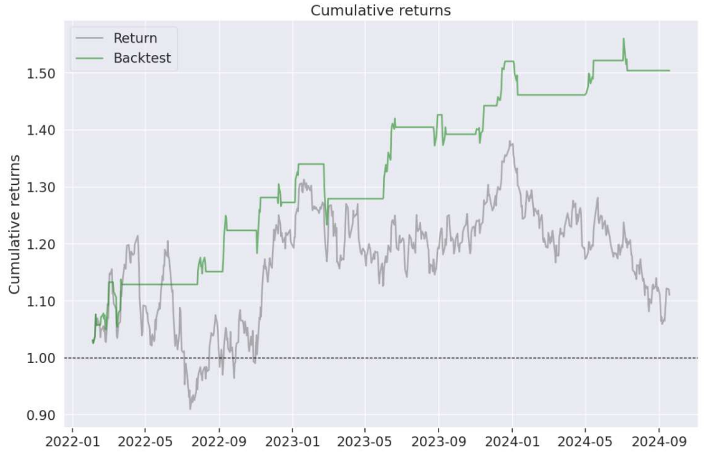
Model	Precision		Recall		F1 Score		MCC	
	Train	Test	Train	Test	Train	Test	Train	Test
Logistic Regression	100%	94%	33%	30%	50%	45%	51%	44%
Linear Discriminant Analysis	95%	85%	37%	34%	53%	49%	51%	43%
Quadratic Discriminant Analysis	64%	56%	52%	38%	57%	45%	41%	26%
Gaussian Naive Bayes	100%	100%	25%	10%	40%	18%	43%	26%
K Nearest Neighbours	100%	67%	100%	12%	100%	20%	100%	18%
Support Vector Machines	100%	100%	33%	30%	50%	46%	51%	47%
Random Forest	100%	94%	71%	30%	83%	45%	79%	44%

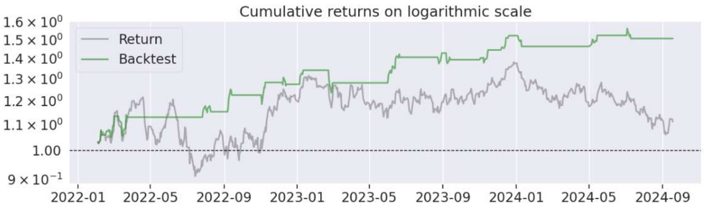
Backtest

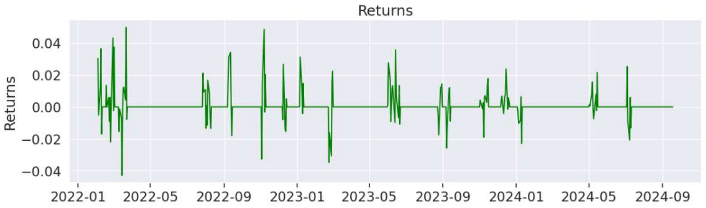
The SVM Model outperforms the underlying model with an average annual return of 16.74%. Note that the model also outperforms on a risk adjusted basis with a sharpe ratio of 1.39. This suggests that our model achieved the underlying objective of this report.

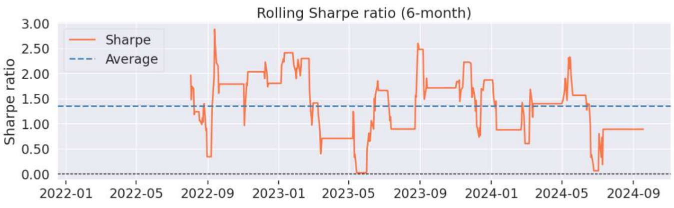


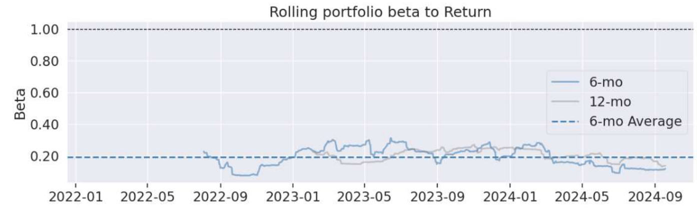
Start date	2022-02-03
End date	2024-09-18
Total months	31
	Backtest
Annual return	16.747%
Cumulative returns	50.38%
Annual volatility	11.635%
Sharpe ratio	1.39
Calmar ratio	2.11
Stability	0.93
Max drawdown	-7.954%
Omega ratio	1.71
Sortino ratio	2.46
Skew	1.57
Kurtosis	15.82
Tail ratio	1.77
Daily value at risk	-1.402%
Alpha	0.16
Beta	0.21

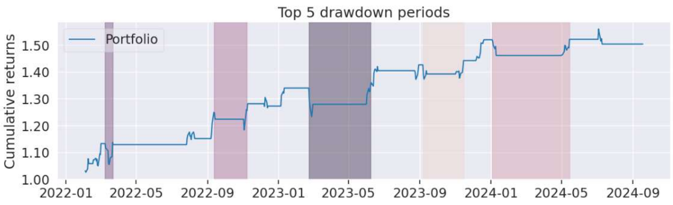


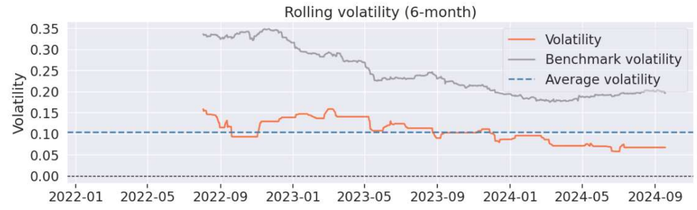










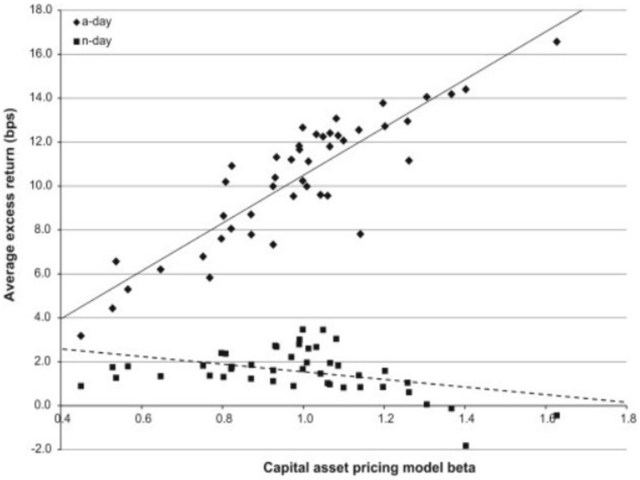


Methodology - Feature Construction

Systematic Risk Exposure and Macroeconomic news releases - Aus & China Adj Beta

Stock betas are a well-known predictor of risk premia. However, most studies find no direct relation between beta and average excess returns across stocks. Over time, expected returns should depend positively on market risk, most often proxied for by some measure of expected market volatility, but such a relation has not yet been conclusively found. Savor and Wilson (2014) show that for an important subset of trading days stock market beta is strongly related to returns, and a robustly positive risk–return trade-off also exists on these same days. Specifically, on days when news about inflation, unemployment, or Federal Open Markets Committee (FOMC) interest rate decisions is scheduled to be announced (announcement days or a-days), stock market beta is economically and statistically significantly related to returns on individual stocks.

Classification of news vs non-news day return



We construct two separate news adjusted betas, one for the Australian market and one for the Chinese market. In each case, we collect the dates where material macroeconomic data were released. We then classify each period within our sample relative to whether a news event was released during that period.

In addition, 1-year rolling betas are estimated using an Australian market index (AORD) and a Chinese market index (Shanghai composite).

These beta values are then cross multiplied by the economic news

Peer Returns

We include the returns of both Rio Tinto (RIO) and Fortescue Metals (FMG), which are broadly considered to be the most appropriate Australian peer group of BHP. These companies are diversified miners with similar commodity portfolios.

Commodity Prices

Commodities produced by different mining companies are generally indifferentiable, with prices typically determined relative to the prevailing spot rate. Whilst quality specifications and legacy offtake agreements introduce a degree of distortion, top line revenue can be simplified as a function of commodity production volume and the realised weighted average price of each respective commodity produced. Companies produce at a given cost, and the market price at the time the commodities are sold determine the margin achieved. Commodity prices therefore have a deterministic impact of the companies bottom line and in turn, the stock price.

The following three commodities have accounted for over 90% of BHP's revenue over the sample period:

1. Iron ore
2. Copper
3. Metallurgical Coal

Market data for these commodity prices are difficult to source. As an alternative approach, yahoo finance provides daily futures prices for each of these commodities which provides a viable alternative.

Market sensitive announcements

Both Chan, Jegadeesh and Lakonishok (1996) and Foster, Olsen and Shevlin (1984) found evidence of abnormal returns following earnings announcements, whose findings have since been successfully reproduced by Chen and Zimmermann (2020). This is an intuitive factor that holds true due to the release of material information which is subsequently acted upon and therefore reflected in the stock price. This same logic applies to other. Predicting directionality based on the individual announcement is out of scope, so instead we opt to recognise the periods where 'market sensitive announcements' are released.

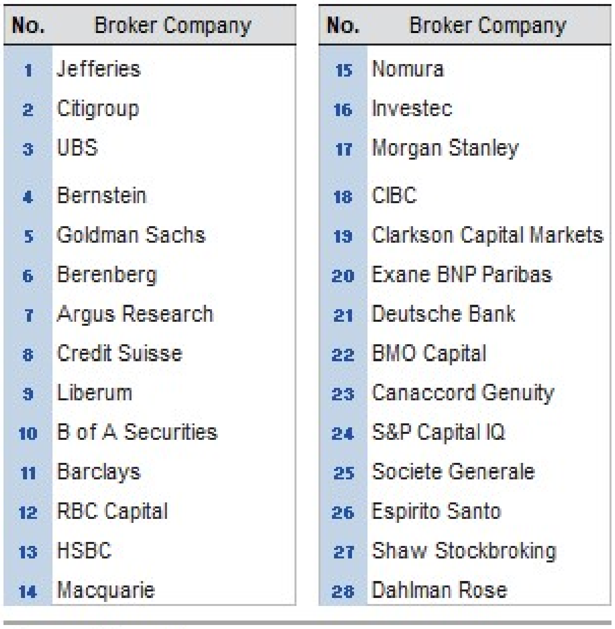
NEWS is another influencing factors considered for market performance. NEWS can be of different category but in this model only business, financial, political and international event based NEWS were included.

Analyst recommendations

Jegadeesh et al. (2004) and Frankel and Lee (1998) both find evidence that analyst recommendations have explanatory power in relation to stock price movements. We include two binary variables: analyst upgrades and analyst downgrades using a sample of 28 sell-side research providers.

Methodology - Feature Construction

Brokers Included Within Broker Reccomendation Feature



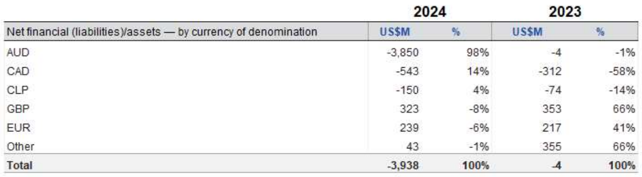
No.	Broker Company	No.	Broker Company
1	Jefferies	15	Nomura
2	Citigroup	16	Investec
3	UBS	17	Morgan Stanley
4	Bernstein	18	CIBC
5	Goldman Sachs	19	Clarkson Capital Markets
6	Berenberg	20	Exane BNP Paribas
7	Argus Research	21	Deutsche Bank
8	Credit Suisse	22	BMO Capital
9	Liberum	23	Canaccord Genuity
10	B of A Securities	24	S&P Capital IQ
11	Barclays	25	Societe Generale
12	RBC Capital	26	Espirito Santo
13	HSBC	27	Shaw Stockbroking
14	Macquarie	28	Dahlman Rose

Source: Yahoo Finance

FX Pairs

BHP operates internationally, with transactions denominated in several different currencies. As an exporter, commodities are typically sold at spot rates which are often denominated in USD. A material proportion of operating costs including direct mining costs, processing costs, G&A expenses alongside capital expenditures are settled in the currency of the projects domicile nation. Adverse fluctuations in FX rates may increase the effective cost of foreign denominated expenses and therefore have the capacity to influence the company's performance. Additionally, the dual listing on the ASX and NYSE mean that arbitrage opportunities may be possible if these two prices are not in equilibrium. The price change of AUD:USD and CAD:USD exchange rates are included in the predictor set due to the significance of BHP's exposure to these prices.

BHP FX Exposure



	2024		2023	
Net financial (liabilities)/assets — by currency of denomination	US$M	%	US$M	%
AUD	-3,850	98%	-4	-1%
CAD	-543	14%	-312	-58%
CLP	-150	4%	-74	-14%
GBP	323	-8%	353	66%
EUR	239	-6%	217	41%
Other	43	-1%	355	66%
Total	**-3,938**	**100%**	**-4**	**100%**

Source: Company Reports

Dividend Month Premium

Companies have positive abnormal returns in months when they are predicted to issue a dividend. Abnormal returns in predicted dividend months are high relative to other companies and relative to dividend-paying companies in months without a predicted dividend, making risk-based explanations unlikely.

Given this, we include four class variables:

1 week until dividend date
2 weeks until dividend date
3 weeks until dividend date
4 weeks until dividend date

Short term reversal

The short-term reversal effect refers to the tendency whereby stocks that have outperformed (underperformed) the market in the short-term will underperform (outperform) in the following period. While there is evidence to suggest that mean-reversion trading strategies underperform in practice due to trading costs and the need to frequently trade/rebalance, the contexts are not directly comparable, and the underlying predictive utility remains.