Predicting Stock Price for Health Care and IT Industry Using Earning per share, Dividend per share, Free Cash Flow, and Earning Surprise

Jia Xi Zhang

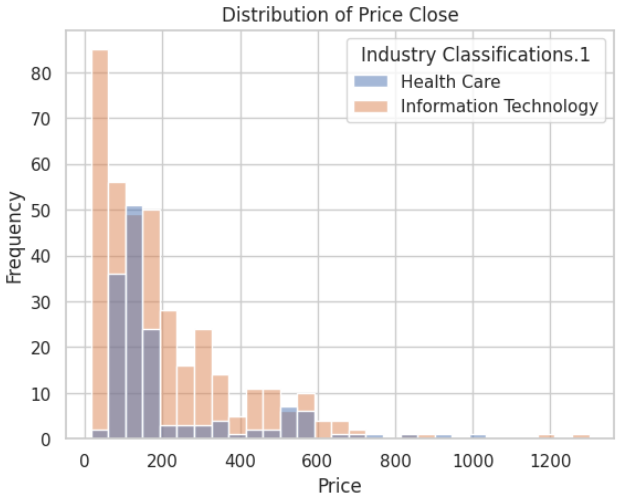
Problem Statement:

In the world of finance, valuation and forecasts are essential. While there are already many advanced models and studies about financial ratios EV/Sales, Price/Book value ratios, as a beginner, I decide to examine a little more on the relationship between price and financial fundamentals. Specifically, I want to learn about the relationship between price and earning per share, dividend per share, free cash flow, and earning surprise, and use these financial fundamentals to build a predictive model. Chosen model types for testing include 1. Multi-factor Regression, 2. K-Neighbors-Nearest, and 3. Random Forest. Since the influence of financial fundamentals like DPS and unexpected earnings shifts dramatically across different industries, a pooled data including all industries is likely to be less significant. Thus, this study will only be focusing on two industries, both are under the high growth and mild to high volatility category: Information Technology and Health Care.

When using models to predict the closing price of a company based on its financial fundamentals, the predicted closing price could be viewed as an estimate of what the market price might be if the market conditions and pricing were primarily influenced by these fundamentals. The rationale of using such models links to the concept of the Efficient Market Hypothesis (EMH). According to EMH, stock prices reflect all available information at any time in a fully efficient market .The prediction from the models could be viewed as the "fair value" based on the known fundamentals, under the assumption that the market efficiency is achieved, or all known information is factored into the stock prices. In reality, market could never be 100% efficient but securities often converge to their intrinsic value as time passes. These models are useful for valuation and investment decision because analysts can use such models to identify potentially undervalued or overvalued stocks. If the model’s predicted price is higher than the current market price, it might suggest that the stock is undervalued, resulting in a buying opportunity, vice versa.

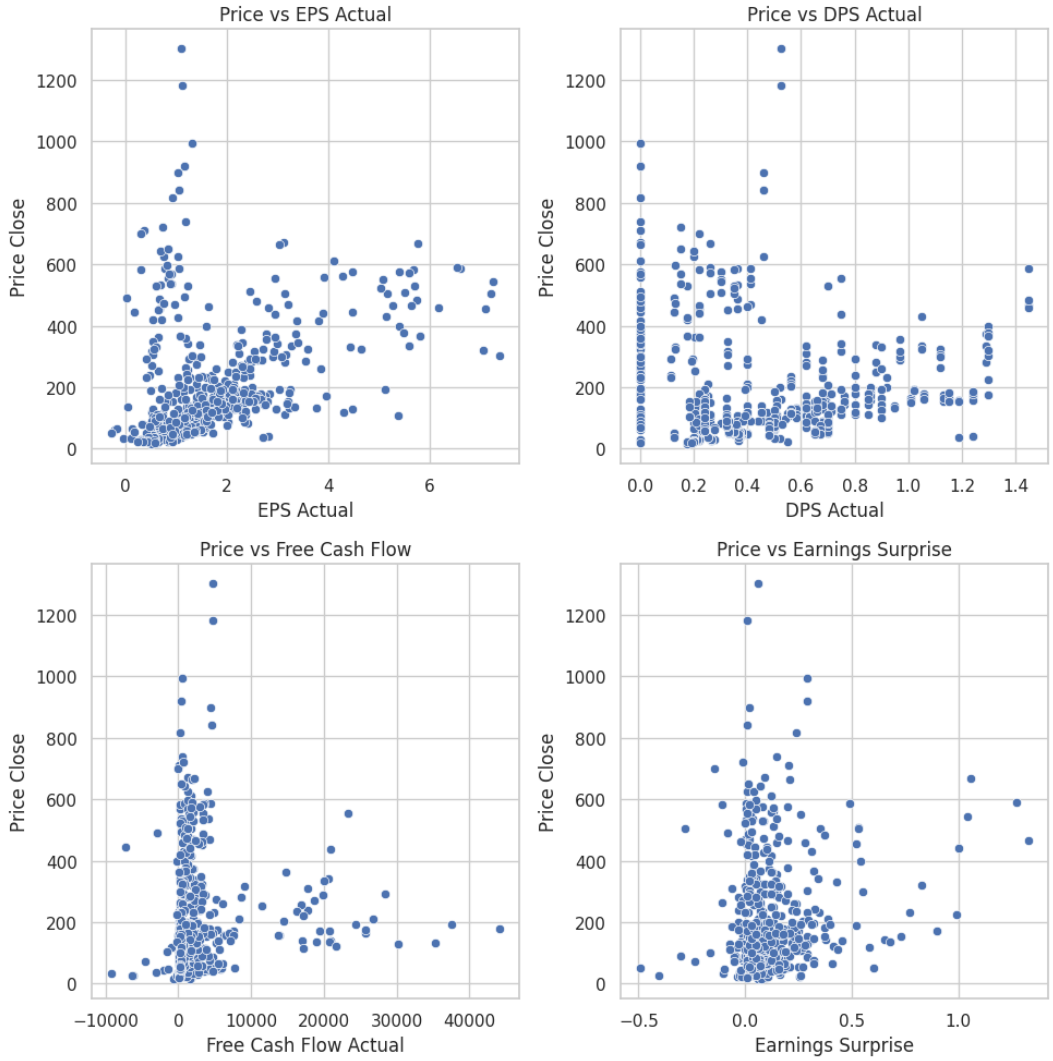
Data Description:

Data for this project is sourced from Capital IQ in xlsx format, displaying information on Company Name, Exchange:Ticker, Tickers, Industry Classifications, past numbers of Fiscal Quarters, DPS Actual, EPS Actual, EPS Forecast, Free Cash Flow Actual, and Price Close for S&P 500 companies in the industrial, health care, and information technology sectors in the past 19 quarters. The raw dataset is in wide format, after converting it to a long dataset, filtering to keep only the health care and IT industries, dropping NAs, and creating the new "Earnings Surprise" column by subtracting EPS Forecast from EPS Actual, there are 529 observations left. For this study, the main variable of interest is the dependent variable of the model, "Price Close".



The above box plot displays the distribution of the closing prices for stocks in health care and information technology industries respectively. It shows that while health care industry's median closing price is slightly lower than the IT sector, its interquartile range is much smaller. The IT sector has a much larger interquartile range and include many extreme outliers, reflecting high variance in company sizes, types, and success rates, from startups to tech giants like Google and Apple. The histogram shows that the price distributions for both industries are right skewed. This is because while a stock can only lose a maximum of 100%, its potential for upside gains is unlimited, creating a longer "tail" on the right side of the distribution. The concentration of stocks is very high at the lower price ranges, indicating a large number of small to mid-sized companies in the Health Care sector. The distribution is more balanced for the IT sector, with a

significant presence of stocks across a range of prices, including a mix of both rapidly growing startups and well-established tech giants. The first scatterplot below of the datapoints below shows a weak positive relationship between price close and EPS. The scatterplots for price vs dividends per share, free cash flow, and earning surprise have data points that are more dispersed, with a cluster around lower and 0 DPS and price values and some outliers at higher price levels. There are no clear linear relationship visible for these three graphs



Model and Interpretations

Three types of models are chosen, the multiple regression model which lies on the traditional econometrics side, and the KNN and random forest models, which lie on the machine learning side. Train-test-split is applied to the data where 20% of the data will be the test set. The

models will be fitted using the train set. In the end, the mean squared error of the model's prediction will be compared. The model with the lowest mean squared error for both test and train set will be chosen to predict stock prices based on the 4 inputs for the most recent quarter in the dataset. By comparing the actual price with the model's predicted price (which should be the stocks' fair value theoretically), the top 5 undervalued and overvalued stocks will be identified.

Model 1: Multiple Regression Model

A multiple linear regression model is chosen because it shows both the direction and magnitude of the predictors' impact on the outcome variable. Meanwhile, it is also the most basic and simplistic model comparing to the models involving machine learning. Both of the above reasons makes multiple regression model a good model to begin the study with. Before running there regression, a correlation matrix is built to prevent multicollinearity. The result is shown as below:

	DPS Actual	EPS Actual	Free Cash Flow Actual	Earnings Surprise	Price Close
DPS Actual	1.000000	0.371816	0.002433	0.092480	-0.070849
EPS Actual	0.371816	1.000000	0.050787	0.481748	0.363034
Free Cash Flow Actual	0.002433	0.050787	1.000000	0.017287	0.058753
Earnings Surprise	0.092480	0.481748	0.017287	1.000000	0.166960
Price Close	-0.070849	0.363034	0.058753	0.166960	1.000000

Table 1

All correlations between all variables are below 0.5; the highest correlation is 0.483, between EPS Actual and Earning Surprise. Multicollinearity is unlikely to occur.

Table 2 on next page shows the statistical result of the multiple regression model. According to the model, if all of the predictors are 0, the stock price will be 133.12, and this intercept is statistically significant at 1%. Among the independent variables, only the coefficient for earnings per share actual and dividend per share actual are statistically significant and they are significant at 1%. Holding all else constant, each dollar increase in EPS Actual is related to $69.68 increase in stock price, while each dollar increase in DPS Actual comes with $139.88

decrease in stock price. The coefficient for EPS aligns with traditional financial theory and studies. EPS is a key indicator of a company's profitability. Since stock prices could represent the present value of the future cash flow a company will generate, EPS is a critical figure for investors assessing a company's value; a higher EPS suggests that a company is efficient at making money, managing costs, and has a solid approach to using its capital, all of which are good signs for potential profitability.[1] In rapid growth sectors like technology or healthcare, EPS usually signals strong future earnings and growth potential. Thus, high eps tends to push up stock prices. The negative coefficient for DPS might seem counterintuitive but also makes sense. Generally, dividends are seen as a positive return on investment for shareholders. Not only do they constitute part of the earnings, but a consistent stream of dividends also implies stability and low risk, which are usually worth paying a premium for, making the stock price higher. In high growth industry like Health care and IT however, growth is often highly emphasized, and the key driver for growth is efficient investment and research and development. Thus, Companies that pay high dividends might be perceived as having fewer opportunities for profitable reinvestment. Conversely, many high price and big market cap technology firm have very low or even 0 DPS as shown in the scatterplot of Price vs DPS too. Free Cash Flow has an extremely small and insignificant coefficient, suggesting that it has neither positive or negative relationship with closing price for firms in the health care and IT industry.

Parameter	Coefficient	Std. Error	t-value	P-value
Intercept	133.1204	20.711	6.428	< 0.001
Industry Cls.1 IT	26.5538	18.660	1.423	0.155
EPS Actual	69.6811	7.585	9.187	< 0.001
DPS Actual	-139.8841	24.988	-5.598	< 0.001
FCF Actual	-0.000032	0.002	-0.020	0.984
Earnings Surprise	-42.7512	51.988	-0.822	0.411

Table 2

[1] Lei, Mozaffari, Zhang, "Quarterly Metrics Impact on Stock Price"

Table 3 shows the feature importance calculated by the model. EPS Actual has the most importance, followed by DPS Actual. Notice how Free Cash Flow has a negative importance. This means that the model performance improves after a feature's values are permuted, meaning that that the FCF Actual is misleading the model rather than helping it make correct predictions. This could occur because in many cases, free cash flow are used for financial activities that do not directly return values to shareholders. “For instance, debt repayment, reinvestment in growth opportunities (e.g., R&D, capital expenditures), and acquisitions. Thus, its impact on price is highly related to the firm’s efficiency in allocating the cash. For example, free cash flow could indicate low reinvestment and thereby lower growth potentials.. in many cases, the outcome of free cash flow allocation decisions take time to materialize, this further leads to a weaker direct relationship between FCF and price.”[2]

Multiple Regression model’s mean square error is 27168.05 for train data and 28847.06 for test data. The R squared for the train and test model predictions are 19.8% and 12.1% respectively. Overall both lower than the baseline MSE of 33735.74 but the R squared numbers, which represent the percentage of variance in the dependent variable that could be explained by the model are not on the high end. Potential reasons for this might include nonlinear relationship between some predictors and the dependent variable, lack of relevant predictors, and omitted variable bias.

Model 2 & 3: K-Nearest Neighbors Regression Model & Random Forest Model

Feature	Importance
EPS Actual	0.497141
DPS Actual	0.119190
Free Cash Flow Actual	-0.000259
Earnings Surprise	0.000674
Industry Classifications.1	0.002885

Table 3

[2] Ibid

KNN is a straightforward algorithm that makes predictions based on the 'nearest' data points in the feature space. For KNN, the class is assigned based on the majority vote of the neighbors while the prediction is the average of the target values of the neighbors. KNN makes predictions based on the similarity of instances in the feature space. When applied to this study, KNN can effectively take into account the natural clustering of industry or companies based on shared characteristics or performance metrics. Unlike multiple regression model, KNN does not assume any underlying relationship or distribution in the data. This could be helpful because the relationship between the financial fundamentals and close price might not necessarily linear, they could be quadratic or logarithmic. For this model, through grid search technique, 5 nearest neighbor is chosen among [to yield the most accurate predictions for closing price. The resulting mean squared error for train and test data are 11693.31 and 10583.99, these numbers drastically improved from the 33735.74 mean square error of the baseline model. The mean squared errors also half the mean squared error for the multiple regression model. At the meantime, R squared for the KNN model also boosted to 65.5% for train model and 67.7% for test model. The model performed better for test set than train set which is unusual. One potential reason might be that my test data (20% of dataset) is much smaller than my train data (80% of dataset), so it might have fewer outliers or less variance in the target variable, making it easier for the model to predict accurately.

Well-suited for datasets with a large number of features, Random Forest can assess the importance of each feature in predicting the stock price, thus focusing on only the most relevant indicators. Like what was done to the KNN model, a grid with different values for the number of trees (n_estimators) and the maximum depth of these trees (max_depth) is set up. The results showed that the best performance came from using 150 trees and a maximum depth of 9 per tree. For the train set, random forest model's mean squared error is extremely low compared to the other models, only 2663.65, R squared also reaches 92.1%. The performance on test data, however, is much worse, the mean squared error rises to 15605.38 while the R squared dropped to 52.5%. The significant disparity in performance between the training data and test data suggests potential overfitting, which occurs when a model learns the detail and noise in the training data to an extent that it negatively impacts the performance of the model on new data.

Overall, taking both mean squared error level and the consistency between training and testing performance into consideration, KNN model will be chosen as the model used for identifying over and undervalued stocks.

Result and Model Limitation

Results

Data on independent variables from the latest quarter are used as inputs for the KNN model, allowing for the prediction of each stock's "fair price". Closing price is then subtracted from this "fair" price, if the difference is positive, then the stock is undervalued, otherwise, overvalued. The 5 most undervalued stocks are printed in table 4 below.

Company Name	Ticker	Industry	Price Close	Predicted Price	Price Difference
Broadcom Inc. (NasdaqGS:AVGO)	AVGO	Information Technology	160.68	733.142	+572.462
Intuitive Surgical, Inc. (NasdaqGS:ISRG)	ISRG	Health Care	262.07	586.524	+324.454
Thermo Fisher Scientific Inc. (NYSE:TMO)	TMO	Health Care	107.95	386.936	+278.986
Cisco Systems, Inc. (NasdaqGS:CSCO)	CSCO	Information Technology	46.98	152.278	+105.298
HP Inc. (NYSE:HPQ)	HPQ	Information Technology	23.46	122.224	+98.764

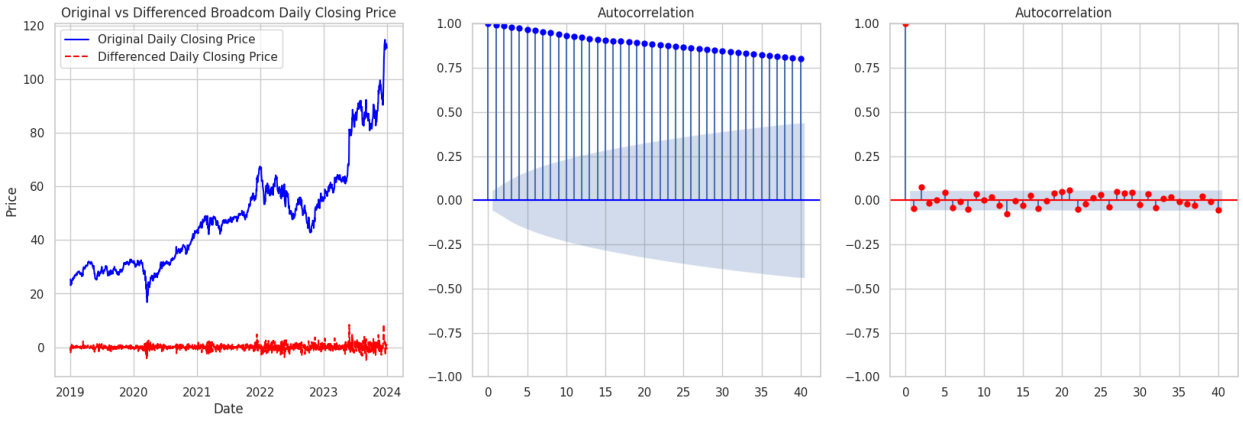
Table 4: Most Undervalued

Company Name	Ticker	Industry	Price Close	Predicted Price	Price Difference
Juniper Networks, Inc. (NYSE:JNPR)	JNPR	Information Technology	698.57	275.144	-423.426
Medtronic plc (NYSE:MDT)	MDT	Health Care	528.07	174.142	-353.928
Microchip Technology Incorporated (NasdaqGS:MCHP)	MCHP	Information Technology	420.72	137.398	-283.322
Microsoft Corporation (NasdaqGS:MSFT)	MSFT	Information Technology	555.79	312.058	-243.732
Oracle Corporation (NYSE:ORCL)	ORCL	Information Technology	275.28	124.834	-150.446

Table 5: Most Overvalued

Limitation

The biggest limitation and challenge of applying the KNN model on the dataset used for this study is that KNN model does not account for time dependencies or auto-correlation between observations, which is critical in panel data where past outcomes might influence future results. KNN relies on distance metrics to find the nearest neighbors. With time-series, data often show trends or seasonality. In this case, the distance metric could become skewed, leading to poorer performance. It is believed that stock prices often demonstrate momentum effects; and trends in growth or recession tend to persist over time. Taking the most undervalued stock Broadcom as an example, the below graphs demonstrate the difference between the original stock price and the differenced stationary stock price. The original stock price has an apparent



upward trend, leading to strong autocorrelation shown in graph 2. The autocorrelation in the original data remains positive and quite high for many lags, indicating that past values have a strong influence on future values. In contrast, differenced data usually fluctuate around a mean with much less trend and has low signs of autocorrelation. KNN treats each input as an independent observation and doesn't consider the order of data. Since KNN is unable to capture the interdependencies between past and present values. In addition, if the scale of the data changes over time, which is usually the case for non-stationary data, the model could become biased towards more recent data or outliers, affecting its predictive accuracy.

Conclusion and Next Steps

Key findings:

1. Model Comparison: The study tested multiple regression, K-Nearest Neighbors, and Random Forest models to predict stock prices based on financial fundamentals including EPS, DPS, Free Cash Flow, and Earnings Surprise. Random Forest model yielded the most accurate prediction for train data while KNN model emerged as the most effective for test data. Both Random Forest and KNN model outperformed multiple regression due to their reliance on local patterns and ability to model highly non-linear and complex relationships.
2. Effectiveness of Financial Metrics: Among the financial metrics, EPS Actual and DPS Actual were found to be statistically significant predictors of stock prices in the multiple regression model. EPS has a positive relationship with stock price in the health care and IT industry as it reflects profitability while DPS has a negative relationship with stock price in high growth

sector has high DPS are often linked to low level of reinvestment. Free Cash Flow and earnings surprise showed statistically insignificant relationship, suggesting it may not be a strong predictor in the Health Care and IT sectors.

3. Performance Issues: The Random Forest model showed potential overfitting, as evidenced by its excellent performance on the training data but poor performance on the test data. In addition, both Random Forest and KNN model are time-insensitive; failing to address for autocorrelation could lead to bias during prediction

Next Steps:

1. **Inclusion of more relevant predictors:** Future study could eliminate insignificant financial fundamentals and test some other financial fundamentals to the model. For instance, including debt metrics and beta value can help measuring volatility, which is highly related to stock price as riskier stocks often need to be compensated by a risk premium. Other financial metrics such as market capitalization which allows controls for company size and ROE which measures efficiency could also be incorporated
2. **Address for knn's inability to capture autocorrelation:** Further research could consider incorporating lagged variables of important predictors. This might help capture some time dependencies. One could also choose to integrate time series analysis techniques or models that account for temporal dynamics and autocorrelation. One case in point is ARIMA (Autoregressive Integrated Moving Average), which is specifically designed to handle such data.

Word count: 2772

Work Cited

Lei, B., Zhang, J. X., & Mozaffari, S. (2024). Quarterly metrics impact on stock price