Erasmus University Rotterdam
Erasmus School of Economics
Master Thesis Data Science and Marketing Analytics

Master Thesis

Eren Muller (492822)

Deciphering Crypto Market Dynamics: The Role of Financial Indicators and Marketing Sentiment

Supervisor:	Armen Arakelyan Badalyan
Second assessor:	Andreas Bayerl
Date final version:	August 2024

Abstract

This research is about findings correlations to future price movements to Bitcoin, Ethereum, and Solana; At the same time we are testing the Efficient Market Hypothesis posed by Eugeuene Fama in our research to see if it holds true. Findings show that Bitcoin and Ethereum follow the efficient market hypothesis making predicting and explaining influences of variables futile. However, it was found that Solana does not adhere to the weaker form of the EMH, allowing models to more accurately predict the price movements of Solana prices. We find S&P 500, Crypto returns (auto-correlation) and Total Value Locked (TVL) to be the three most important variables to help explain future price movements of Solana prices. Additionally, social media market sentiment on Reddit, was not found to be informative in predicting future price movements for any of the currencies.

Acknowledgements

The research process was both fun and engaging, providing me the opportunity to apply the theoretical knowledge gained from classes to real-life practical situations. It was a time-consuming and mentally taxing endeavor where I delved deeply into the complexities of machine learning models and their interpretation. This process also taught me the importance of organizing and indexing raw data, as information can become chaotic very quickly. Understanding how to construct a dataset was another significant lesson; I discovered the challenges of gathering, cleaning, and transforming data. A crucial lesson I learned was to never underestimate the level of programming engineering required to establish smooth and automated data pipelines that are ready for use. For any project moving forward, building robust data pipelines will be a priority.

I would like to acknowledge my thesis supervisor, Armen Arakelyan Badalyan, whose invaluable support and financial expertise were instrumental throughout the thesis process. His guidance in selecting variables, refining methodologies, and enhancing feature engineering was crucial. Additionally, his insights greatly improved the writing, flow, and structure of this thesis. I am deeply grateful for his contributions.

I would like to express my deepest gratitude to Nikilesh Jagan and Sai Rithvik V., true friends who have been a constant source of motivation and support, not just during the writing of this thesis, but in all aspects of my life. Your unwavering encouragement has been invaluable.

A special thank you to my little brother, Fabian Muller, whose relentless drive and commitment inspire me to maintain a strong work ethic and keep pushing forward. Your example has been a constant reminder to stay focused and dedicated.

Finally, I would like to extend my heartfelt thanks to my parents, whose love, guidance, and unwavering support have been the foundation of everything I have achieved. Your belief in me has been my greatest strength, and for that, I am eternally grateful.

Contents

Introduction

In 2024, the cryptocurrency market value had exceeded US $2.3 trillion, marking a significant milestone in the new financial model that competes with traditional financial systems. The legitimacy of blockchain technology is secure, despite consistent attacks on its credibility due to the criticism cryptocurrency has faced throughout its history. With major institutions like BlackRock and Fidelity entering the market, it is evident that blockchain's relevance is well-established. In light of this, the question now shifts from blockchain's long-term credibility and sustainability to understanding the use cases and major forces driving the prices.

According to Statista, as of June 2024, there are slightly over 10,000 unique currencies. Nonetheless, this thesis will focus on the Top 3 cryptocurrencies by market capitalization: Bitcoin, Ethereum, and Solana. The objective is to predict price movements and identify the underlying factors that contribute to future price fluctuations.

The motivation for this paper stems from the need to develop models that can predict cryptocurrency prices, enabling better decision-making and risk management. By exploring the interplay between technical indicators, market sentiment and macro-economic indicators, this research aims to improve our understanding of the driving forces behind cryptocurrency price dynamics, ultimately contributing to more informed investment strategies and enhanced market efficiency.

There have been numerous attempts at predicting stock market assets, where some researchers had positive result. What most methodologies have in common is time series forecasting using various machine learning algorithms, ranging from basic algorithms such as linear regressions, to more advanced models such as support vector machines or neural networks. However, a major view opposing the feasibility of predicting stock prices comes from Eugene Fama's Efficient Market Hypothesis (EMH).

The challenge stems from Fama's understanding that all available information of stocks is already reflected in current prices of the asset thereby making any future prediction futile. Keeping this in mind, opposing views have also been studied that claimed that not all markets are efficient, and that some markets are indeed prone to be predictable. With a split in sentiment regarding the predictability of stock/cryptos, this thesis will try to take part in the debate regarding EMH in crypto markets. This will be done by utilizing a wide range of machine learning algorithms trained on a multi-variable dataset.

The main research question guiding the thesis is: "Is it feasible to accurately predict the price movements of cryptocurrency assets? If so, which variables have the most significant influence on the decision-making process of predictive models?" This research question won't only answer concerns about EMH in the crypto currency market but will also help understand the forces driving future price movements of crypto assets.

The machine learning algorithms will be trained on a diverse set of variables to ensure ro-

bust predictive modeling. These variables encompass sentiment analysis metrics, including Reddit sentiment scores, Wikipedia page views, and Google Trends data, providing insight into the market's psychological factors. Additionally, macroeconomic indicators such as the S&P 500 index and gold prices will be integrated to capture broader economic influences on cryptocurrency markets.

Furthermore, currency-native data Total Value Locked (TVL) will be incorporated to reflect the inherent value within the crypto ecosystem. To complement these, we will also use trend-deterministic technical analysis variables derived from historical cryptocurrency prices, which will help in identifying patterns and trends within the market. This approach aims to leverage a wide spectrum of data sources to enhance the accuracy and reliability of the models.

To understand how machine learning algorithms leverage different features to effectively predict the price movements, SHAPley global interpretation methods will be used. This is a method that enables the possibility to investigate the decision making of the machine learning algorithm. This in turn will describe each features impact on the future price movements of the crypto assets and how they are correlated.

It was found that Bitcoin and Ethereum adhere to the weak form of EMH. Solana however, does not adhere to the weaker form of the EMH, allowing models to more accurately predict the price movements of Solana prices. We find S&P 500, Crypto returns (autocorrelation) and Total Value Locked (TVL) to be the three most important variables to help explain future price movements of Solana prices. Additionally, social media market sentiment on Reddit, was not found to be informative in predicting future price movements for any of the currencies.

The findings of this thesis are highly relevant for managers and investors. Understanding the variables that drive cryptocurrency prices can provide valuable predictive insights, enhancing decision-making processes and risk assessment capabilities. As the regulatory frameworks for cryptocurrencies continue to evolve, insights from this research could also assist in crafting informed regulations that ensure market stability and protect investors.

From an academic perspective, this study contributes to the literature by addressing the gap related to the modeling of cryptocurrency price movements. By applying machine learning and deep learning techniques, it extends the discussion of EMH to the relatively unexplored domain of digital currencies, offering a fresh perspective on market efficiency theories in the context of the 21st-century cryptocurrency landscape.

The thesis will be structured as follows: First, previous literature will be reviewed to understand the methodologies used to predict cryptocurrency and stock price movements. Predictive philosophy will also be examined to assess the feasibility of prediction attempts. Secondly, we will delve into the methodology of data collection, machine learning training, and the interpretation of the machine learning models. This will be followed by a detailed discussion of the results to provide a comprehensive understanding of the outputs generated by the machine learning algorithms. Lastly, we will address the limitations of this study and suggest areas for further research.

Literature Review

Research on predicting Bitcoin and other cryptocurrencies using machine learning or deep learning algorithms is relatively unexplored. Consequently, literature on the stock market will also be closely examined to serve as a proxy for understanding cryptocurrency research. The literature review will begin at the foundation of the thesis: Is it even possible to predict stock or cryptocurrency prices? A comparison of research both supporting and challenging the possibility of predicting stock and crypto prices will be explored.

Additionally, the literature review aims to gather extensive information on the variables related to the returns of these assets. An array of past empirical work, from exploring technical indicators to the influence of social media sentiment on prices, will be considered. At the forefront of the hypothesis is the expectation that marketing-related factors, such as social media sentiment, significantly affect asset returns; therefore, literature examining this influence will also be analyzed.

Lastly, we will want to explore various methodologies used in past research to shed light on model choices, deep learning approaches, and the engineering of feature sets. Numerous machine learning and deep learning models exist, with each having its own set of hyperparameter settings. Extracting this information can be helpful in minimizing hyperparameter tuning times.

Overall, this literature review not only aims to explore the philosophical and technical feasibility of predicting stock and cryptocurrency price movements but also the technical aspects of model architectures such as model choices, and hyperparamter settings.

2.1 Efficient market hypothesis

Eugene Fama's Efficient Market Hypothesis (EMH) asserts that stock prices will reflect all the available information, meaning that they almost always trade at their fair value (Fama, 1970). Fama argues that since that all new information is quickly incorporated into the stock prices, it is essentially impossible and futile to consistently predict market movements. This underpins the argument for passive index fund investing, which aims to match market returns rather than exceed them.

Fama (1995) further elaborates on why technical analysis, or charting, is ineffective in an efficient market. According to EMH, the market almost instantly adjusts to new information. This instantaneous adjustment renders efforts to predict future stock prices based on past data futile, as all known and predictable information is already embedded in current prices.

Fama (1995) provides empirical evidence for this through the evaluation of strategies such as the 5% filter rule. This strategy involves buying an asset if its closing price increases by 5% and holding it until the closing price decreases by 5%, at which point the investor

sells and simultaneously shorts the asset. The short position is held until a subsequent 5% rise, followed by buying and covering. Fama's analysis shows that such strategies do not yield higher returns compared to a traditional buy-and-hold portfolio; in fact, the latter often outperforms these filter strategies.

Moreover, Fama (1995) explains that while some analysts may anticipate the outcomes of new events and buy at lower prices expecting future increases, the existence of many proficient analysts leads to instantaneous price adjustments. This collective efficiency means that individual analysts cannot consistently outperform the market. For technical analysts to justify their methods, they must demonstrate an ability to consistently make better-than-chance predictions. He concludes that the competitive nature of the market, coupled with the rapid dissemination of information, ensures that stock prices follow a random walk, thus supporting the core tenets of EMH.

To contribute to the discussion, Borges (2010) argues that the German and Spanish stock market follow strong forms of EMH. The authors claim this by stating two main reasons. First, there was no evidence of auto-correlation between prices, meaning that positive returns today likely mean positive returns tomorrow. Secondly, there was no evidence of mean reversion meaning that prices tend to deviate back to their historical means. However, the author mentions that stock markets in Spain, Greece, UK, and France does not follow weaker forms of EMH. This is due to the presence of auto-correlation and mean reversion. This opens the stage to consider perspectives opposing Fama and EMH.

2.2 Views opposing EMH

Behavioral finance challenges Eugene Fama's Efficient Market Hypothesis by arguing that psychological factors and irrational behavior of investors can lead to market inefficiencies. Scholars Kahneman and Tversky (2013), developed prospect theory, to demonstrate that cognitive biases such as overconfidence and loss aversion significantly influence investor decisions, often leading to predictable and systematic errors. Kahneman and Tversky (2013) mention that these biases cause stock prices to deviate from their true values, creating opportunities for superior returns through strategic trading, contrary to EMH's assertion that such opportunities are fleeting or non-existent. Behavioral finance thus provides a framework to understand why and how markets might not be entirely efficient.

Kang, Lee, and Park (2022) investigate the presence of the Efficient Market Hypothesis (EMH) in the cryptocurrency market. Their study involves testing 893 cryptocurrencies, and the results reveal that only a small fraction of these currencies adhere to the EMH. Specifically, Kang et al. (2022) find that only 54 cryptocurrencies (6%) follow the weak-form EMH, and just 24 (3%) adhere to the semi-strong-form EMH. These findings suggest that the cryptocurrency market demonstrates limited efficiency in information processing. Moreover, the study concludes that most cryptocurrencies do not incorporate past prices or new information into their market prices.

Le Tran and Leirvik (2020) also undertakes examining the Efficient Market Hypothesis (EMH) within the cryptocurrency market. Le Tran and Leirvik (2020) conclude that market efficiency is highly variable over time, particularly noting significant inefficien-

cies before 2017. Tran observes that, over time, the cryptocurrency market is becoming increasingly efficient. Among the cryptocurrencies tested, Litecoin emerges as the most efficient, while Ripple is identified as the least efficient.

2.3 Past attempts at predicting the stock market

In the study conducted by Kara, Boyacioglu, and Baykan (2011), an artificial neural network (ANN) and support vector machine (SVM) are employed to predict stock price movements on the Istanbul Stock Exchange. The independent variable in this research is a binary indicator reflecting whether the stock price will move up or down the following day. Kara et al. (2011) finds that the SVM, particularly with a polynomial activation function, outperforms all other algorithms, including ANN and backpropagation network (BPN), achieving an accuracy of 71.5%. The feature set for this study comprises various technical analysis (TA) indicators such as the Moving Average Convergence Divergence (MACD), Moving Average (MA), and the stochastic oscillator %K (K%).

In another study focusing on trend deterministic data for stock price prediction by Patel, Shah, Thakkar, and Kotecha (2015), a classification model is used to forecast the up or down movement of stock prices. Patel et al. mentions that this research incorporates a feature set consisting of binary variables indicating whether a technical indicator suggests an upward or downward trend. The highest performing model in this study is the random forest, which achieves an accuracy of 83.5%. However, a noted limitation of this approach is the binary nature of the technical indicators. The study suggests that incorporating additional levels to represent the degree of movement, such as 'slightly up', 'slightly down', and 'barely down', could enhance the model's accuracy.

In another study, Weng, Lu, Wang, Megahed, and Martinez (2018) attempt to predict short-term stock prices using ensemble methods. The feature set in this research is diverse, comprising historical stock prices, well-known technical indicators, sentiment scores derived from published newspaper articles, trends in Google searches, and the number of visits to Wikipedia pages. The study demonstrates impressive results, predicting the next day's stock prices with a mean absolute percentage error (MAPE) of less than 1.5%. The best-performing algorithms in this research are boosted decision trees, including XGBoost and AdaBoost.

Usmani, Adil, Raza, and Ali (2016) conduct a study to predict the Karachi Stock Exchange (KSE) using various machine learning algorithms. They employ a classification model to forecast whether the market will go up or down. The feature set for this study is extensive, including oil rates, gold and silver rates, interest rates, foreign exchange (FEX) rates, news and social media feeds, simple moving averages (SMA), and autoregressive integrated moving average (ARIMA) data. Usmani et al. (2016) find that the best performing model is the multilayer perceptron (MLP), a type of artificial neural network (ANN), alongside support vector regression (SVR).

Kumbure, Lohrmann, Luukka, and Porras (2022) conduct a comprehensive literature review on the application of machine learning and data used for stock market forecasting.

This review examines a total of 138 articles related to machine learning in stock markets, providing a detailed overview of the models, markets, and feature sets used in these studies. Kumbure et al. (2022) highlight that the most used machine learning methods are neural networks, support vector machines/support vector regression (SVM/SVR), and fuzzy theories. Additionally, they note that most of these papers incorporate technical indicators in their feature sets.

2.4 Past attempts at predicting the cryptocurrency market

Chen, Li, and Sun (2020) conduct a study to predict Bitcoin prices using various machine learning methods, including logistic regression and long short-term memory (LSTM) networks. Chen et al. (2020) finds that by utilizing 5-minute interval price data, the model achieves an accuracy of 66%, outperforming more complex neural network models. The feature set in this study is comprehensive, incorporating not only Bitcoin price data but also external factors such as gold spot prices, property and network data, as well as trading and market information.

Another study conducted by McNally, Roche, and Caton (2018) utilized deep learning models to predict the price movement of Bitcoin. Their dataset consisted of historic price data, including Open, High, Low, Close values, spanning from 2013 to 2016. Additionally, they incorporated on-chain data such as mining difficulty and hash rate. All data was standardized, as deep learning algorithms greatly benefit from such preprocessing. Their results showcased model accuracies ranging from 50.25% to 52.75%, with the LSTM model achieving the highest accuracy. Although their best model demonstrated 100% precision, the low recall value of 14.7% significantly limits the utility of this metric.

Mudassir, Bennbaia, Unal, and Hammoudeh (2020) employed four different machine learning algorithms to predict the price movement of Bitcoin. The models used included ANN, SANN, and SVM. These models relied on technical indicators and on-chain data, such as hash rate and mining difficulty, to predict price movements. Principal Component Analysis (PCA) was applied to their dataset of 700 technical indicators, with the first principal component (PC1) explaining 95% of the variance. The accuracies of these models ranged between 45% and 60%, with SANN being the best performing model and SVM achieving an accuracy of 54%. However, Mudassir et al. (2020) did not report additional evaluation metrics such as precision, recall, or specificity.

Phaladisailoed and Numnonda (2018) compared various machine learning models to evaluate their relative performances using a feature set that included historic Bitcoin prices, weighted prices, and volume data sampled at 1-minute intervals from 2012 to 2018. The data was standardized using MinMax scalers. The study employed two regression models—Theil-Sen Regression and Huber Regression—alongside two deep learning models, LSTM and GRU. The results indicated that the deep learning algorithms outperformed the regression models, with the LSTM and GRU models achieving RMSE values of 0.002 and 0.004, respectively. The authors noted unusually high R^2 values, with both models capturing 99.2% of the variance, which suggests exceptionally strong model fits

that may warrant further scrutiny for potential overfitting.

2.5 Social Media and Sentiment Analysis in the Role of Predicting Stock/Crypto Prices

The influence of social media on Bitcoin prices has been a topic of significant interest in recent research. A study by Mai, Shan, Bai, Wang, and Chiang (2018) employs textual analysis and vector error corrections to demonstrate a clear link between social media sentiment and Bitcoin price movements.Mai et al. (2018) shows that bullish posts on social media platforms are associated with higher future Bitcoin prices. This suggests that social media sentiment is a valuable predictor of Bitcoin price fluctuations, highlighting the impact of public opinion and social discourse on cryptocurrency markets.

In addition to social media, online search activity also correlates with Bitcoin price movements. Kristoufek (2013) analyzes Google Trends and Wikipedia page visits, finding strong correlations between these data points and Bitcoin prices. This research suggests that increased online searches and Wikipedia activity, reflecting public interest and awareness, can significantly influence Bitcoin market trends. Complementing these findings, Chan (2003) study stock market prediction through news sentiment reveals that positive newspaper headlines often lead to overvaluation of stocks, while negative headlines result in undervaluation. Chan (2003) further notes that this sentiment effect is more pronounced in smaller market capitalization stocks and that investors typically react slowly to sentiment changes. Together, these studies underscore the significant role of public sentiment, whether expressed through social media, search activity, or news headlines, in influencing financial markets.

The predictive power of social media sentiment on cryptocurrency prices has been further explored in recent studies. Kraaijeveld and De Smedt (2020) research the influence of Twitter sentiment on the returns of major cryptocurrencies. Kraaijeveld and De Smedt (2020) conclude that Twitter sentiments indeed have predictive power over cryptocurrency prices, utilizing a lexicon-based sentiment analysis. The study highlights that news disseminated through Twitter can rapidly alter investor sentiments, leading to immediate and significant price movements. This finding emphasizes the crucial role of real-time sentiment analysis in anticipating market trends and price fluctuations in the volatile cryptocurrency market.

Similarly, news sentiment shows a notable impact on Bitcoin prices. Rognone, Hyde, and Zhang (2020) study the effect of unscheduled news on Bitcoin compared to traditional currencies using intra-day data from January 2012 to November 2018. Rognone et al. (2020) finds that Bitcoin often reacts positively to news, whether positive or negative, indicating a high level of enthusiasm among investors towards Bitcoin, unlike traditional stock markets. However, specific negative news, such as reports of fraud and cyber-attacks, have adverse effects on Bitcoin prices. The study utilizes RavenPack's real-time news data and employs a Vector Auto-Regressive Exogenous (VARX) model for the analysis. In parallel, Khan et al. (2022) combines social media and news sentiment to predict stock market movements, using a dataset from Twitter and Yahoo Finance. Khan et al. (2022) demonstrate that their predictive model achieves an accuracy of 80% after filtering out

spam tweets, underscoring the significant impact of integrated sentiment analysis on market predictions. These studies collectively highlight the importance of sentiment analysis in understanding and forecasting market dynamics across various financial assets.

2.6 Conclusion

The literature review underscores the challenging nature of predicting stock and cryptocurrency prices, emphasizing the importance of robust predictions for effective trading strategies. Research on stock price prediction is extensive, utilizing various machine learning (ML) and deep learning (DL) methods. Studies such as those by Kara et al. (2011) and Patel et al. (2015) highlight the effectiveness of SVM and random forest models, respectively, in forecasting stock prices using technical indicators. Similarly, Chen et al. (2020) and Weng (2018) demonstrate the predictive power of logistic regression, LSTM networks, and ensemble methods for Bitcoin and stock prices, leveraging comprehensive feature sets that include market variables and sentiment scores. The review also notes the evolving efficiency of cryptocurrency markets, with studies like those by Kang et al. (2022) and Tran (2020) revealing limited adherence to the Efficient Market Hypothesis (EMH), indicating significant information processing inefficiencies.

The impact of sentiment analysis on market predictions emerges as a critical theme. Research by Mai et al. (2018) and Kraaijeveld and De Smedt (2020) establishes the predictive power of social media sentiment on cryptocurrency prices, while Kristoufek (2013) and Chan (2003) demonstrate similar effects for online search activity and news sentiment on Bitcoin and stock markets. Studies such as Rognone et al. (2020) and Khan et al. (2022) further validate the significant influence of real-time sentiment, integrating social media and news data to achieve high prediction accuracy. These findings collectively highlight the importance of incorporating diverse feature sets, including sentiment analysis, to enhance the predictability of financial markets and challenge traditional notions of market efficiency.

2.7 Hypotheses

In this section, we will test three key hypotheses that challenge the assumptions of the Efficient Market Hypothesis (EMH) as proposed by Eugene Fama. The aim is to determine whether the cryptocurrency market, specifically Bitcoin, Ethereum, and Solana, adheres to the weak form of EMH. By testing these hypotheses, we seek to identify whether certain variables can predict future price movements, thereby providing insights into the efficiency of the cryptocurrency market.

Null Hypothesis 1: Technical indicators do not play an important role in predicting future price movements of cryptocurrencies.

The weak form of the EMH asserts that all past trading information, including price and volume data, is already reflected in current asset prices. If this hypothesis holds true, it implies that technical analysis, which relies on historical data to forecast future price trends, should not provide any additional predictive power. However, if this null hypothesis is rejected, it would suggest that technical indicators can indeed forecast future price

movements, thereby challenging the weak form of EMH and indicating inefficiencies in the cryptocurrency market.

Conversely, however, the study conducted by McNally et al. (2018) did not provide sufficient evidence to reject the null hypothesis, as the highest accuracy they achieved was 53%. Given these conflicting findings, this thesis seeks to contribute to the ongoing debate by offering a fresh perspective on the role of technical indicators in predicting cryptocurrency price movements. By rigorously testing this hypothesis, the research aims to clarify whether technical analysis can indeed provide meaningful predictive insights in the context of the cryptocurrency market, thus adding valuable evidence to the discourse on market efficiency.

Null Hypothesis 2: Sentiment analysis from social media does not play a role in predicting future price movements of cryptocurrencies.

Market sentiment, particularly as expressed on social media platforms, can influence investor behavior and, consequently, asset prices. The weak form of EMH would argue that such sentiment is already factored into current prices, making it irrelevant for predicting future movements. However, if this null hypothesis is rejected, it would indicate that social media sentiment does have a significant impact on price dynamics, suggesting that the cryptocurrency market may be more susceptible to psychological factors and less efficient than the EMH would propose.

This is further evidenced in studies conducted by Mai et al. (2018) where they concluded that bullish posts on social media platofrms correlated with future Bitcoin price movements. Additionally, Kraaijeveld and De Smedt (2020) also conclude that twitter sentiment has predictive power in future cryptocurrency price movement. Based on the literature, there seems to be a lot of evidence to make grounds to reject the null hypothesis. This thesis will add to this debate, and try to contribute to see if sentiment on social media still holds predictive power in the current cyrpto currency market.

Null Hypothesis 3: Macroeconomic indicators do not play an important role in predicting future cryptocurrency prices.

Macroeconomic indicators, such as stock market indices and commodity prices, often reflect broader economic conditions that can affect a wide range of asset classes, including cryptocurrencies. According to the weak form of EMH, these external factors should not offer any predictive insight into future prices, as all relevant information is already embedded in current prices. However, if this hypothesis is rejected, it would imply that macroeconomic variables do influence cryptocurrency prices, indicating that the market is not fully efficient and is affected by external economic forces.

By rigorously testing these null hypotheses, this research will provide empirical evidence on whether the cryptocurrency market adheres to the weak form of the EMH. Rejecting any of these null hypotheses would suggest that the market is not entirely efficient, as certain variables would be shown to have predictive power over future price movements.

Data & Methodology

3.1 Data Collection

Collecting the data involves several steps, leveraging multiple data sources and usuing various Python scripts to gather and process the required information. This section outlines the approach taken to ensure the data collection.

The variable that will be predicted for all machine learning models will be a dummy indicator to represent the movement of the crypto price movements. If for any given day the price increases from the day before, then this variable should be 1, 0 otherwise. In other words we will be predicitng a binary variable for price movements. This means that any models we choose must be classificiation models.

Financial market data was obtained from Yahoo Finance. This includes historical prices and volumes for a range of assets such as Bitcoin (BTC), Ethereum (ETH), Solana (SOL), Gold, Nvidia, VIX and S&P 500. We used the requests library from python to download from the yahoo website the different data sets per asset.

To gauge public interest into cyrpto, data from Trends and Wikipedia is incorporated. For Google Trends, the focus is on terms such as 'bitcoin', 'ethereum', 'solana', 'crypto', and 'blockchain'. This data is accessed directly from google trends, which allows us to download historical search interest data for these terms. Similarly, Wikipedia page views for the same terms are collected from https://pageviews.wmcloud.org/. The aggregation of this data involves summing the total page views and total search interest over a daily time frame. This provides a measure of the general public's engagement and interest in crypto related topics.

Reddit posts data from the direct subreddits of Bitcoin, Ethereum, and Solana were gathered using the Reddit API, focusing on the top 100 posts at the time of collection. Python automation was required to systematically retrieve the top 100 posts during same time intervals, as the API only allows for current data. The retrieved information includes the Post title, post body, number of upvotes, and the date posted.

The total value locked (TVL) data for each cryptocurrency is obtained from Defi Lama. TVL represents the total capital held within a blockchain's DeFi ecosystem, providing insight into the level of trust and engagement from the community. This data is crucial for understanding the financial health and adoption of each cryptocurrency.

The final step involves merging all the collected data on the date field. This step integrates the financial market data, Google Trends and Wikipedia data, Reddit sentiment scores, TVL data, and technical indicators into a single cohesive dataset. The merged dataset is then ready for feature selection and modeling.

The final shape of the dataset spans from February 2, 2024, until July 17, 2024, with a total of 168 rows of data.

3.2 Feature Engineering

3.2.1 Technical Indicators:

Technical indicators are derived from historical price data. We use a Python function to calculate several key indicators, which are shown in table 3.1. Additionally, our approach requires creating binary indicators for each technical indicator. This means that for any given technical indicator, a signal will be extracted based on specific logic. For example, if the price is higher than the 21-day moving average (21MA), it is considered a bullish/buy signal. The table below presents both the calculation of each technical indicator and the corresponding logic used to determine the signal. //

Indicator	**Formula**	**Trend Det. (1/0)**
21D MA	$MA_{td} = \frac{1}{21}\sum_{i=0}^{20} P_{t-i}$	1 if $P_t > MA_{td}$, 0 else
%K	$K_{td} = \frac{P_t - LL}{HH - LL} \times 100$, $LL = \min(P_{t-9}, \ldots, P_t)$, $HH = \max(P_{t-9}, \ldots, P_t)$	1 if $K_{td} > K_{td-1}$, 0 else
%D	$D_{td} = \frac{1}{3}\sum_{i=0}^{2} K_{td-i}$	1 if $D_{td} > D_{td-1}$, 0 else
RSI	$RSI_{td} = 100 - \frac{100}{1+\frac{U_{td}}{D_{td}}}$, where $U_{td} = \frac{1}{14}\sum_{i=0}^{13}\max(\Delta P_{t-i}, 0)$ and $D_{td} = \frac{1}{14}\sum_{i=0}^{13}\max(-\Delta P_{t-i}, 0)$	$\begin{cases} -1 & RSI_{td} \geq 70 \\ 1 & RSI_{td} \leq 30 \\ 0 & \text{else} \end{cases}$
Momentum	$Momentum_{td} = P_t - P_{t-10}$	1 if $Momentum_{td} > 1$, 0 else
MACD	$MACD_{td} = EMA_{12,td} - EMA_{26,td}$, where $EMA_{12,td} = P_t \cdot \frac{2}{13} + EMA_{12,t-1} \cdot \frac{11}{13}$, $EMA_{26,td} = P_t \cdot \frac{2}{27} + EMA_{26,t-1} \cdot \frac{25}{27}$	1 if $MACD_{td} > MACD_{td-1}$, 0 else
CCI	$CCI_{td} = \frac{TP - SMA_{TP}}{0.015 \cdot MD}$, $TP = \frac{H_t + L_t + P_t}{3}$, $SMA_{TP} = \frac{1}{20}\sum_{i=0}^{19} TP_{t-i}$, $MD = \frac{1}{20}\sum_{i=0}^{19} \lvert TP_{t-i} - SMA_{TP}\rvert$	$\begin{cases} -1 & CCI_{td} \geq 100 \\ 1 & CCI_{td} \leq -100 \\ 0 & \text{else} \end{cases}$
Bollinger	$Bollinger_{td} = \begin{cases} UB = SMA_{20} + 2 \cdot STD_{20} \\ LB = SMA_{20} - 2 \cdot STD_{20} \end{cases}$, where $SMA_{20} = \frac{1}{20}\sum_{i=0}^{19} P_{t-i}$, $STD_{20} = \sqrt{\frac{1}{20}\sum_{i=0}^{19}(P_{t-i} - SMA_{20})^2}$	$\begin{cases} -1 & P_t > UB \\ 1 & P_t < LB \\ 0 & \text{else} \end{cases}$

Table 3.1: The table displays all technical indicators used in the feature set. A total of eight indicators were employed. Additionally, the table shows how these variables, based on absolute values, were transformed into binary, trend-deterministic features.

3.2.2 Reddit Sentiment Analysis:

In the previous step, gathering the Reddit data was discussed. However, this data is raw and requires text processing to extract aggregated sentiment per day. //

To ensure a comprehensive sentiment analysis, the titles and texts were combined into a single column for each post. The combined text was then preprocessed using the Natural Language Toolkit (nltk) in Python. This preprocessing involved several steps to clean the text and prepare it for analysis. First, the text was converted to lowercase to ensure uniformity. Punctuation and special characters were removed, and the text was tokenized into individual words. Common stop words such as "the," "and," and "is" were removed to reduce noise. Finally, lemmatization was performed to convert words to their base forms, such as converting "running" to "run." This preprocessing step ensured that the text was clean and ready for sentiment analysis.

For sentiment analysis, we utilized VADER (Valence Aware Dictionary and sEntiment Reasoner), a lexicon and rule-based tool specifically designed to handle sentiments expressed in social media contexts. VADER employs a lexicon-based approach, where each word in its predefined list has an associated sentiment score indicating whether it is positive, negative, or neutral. Additionally, VADER applies rules and heuristics to handle punctuation, capitalization, degree modifiers, and conjunctions. For example, exclamation marks increase the intensity of the sentiment, uppercase words indicate stronger sentiment, and words like "very" or "kind of" modify the intensity of the sentiment.

The combined title and text of each Reddit post were passed through VADER for sentiment analysis. VADER first split the text into individual sentences, then matched each word in the sentences against its lexicon to retrieve sentiment scores. The sentiment scores of words in a sentence were aggregated, with adjustments made for the rules and heuristics VADER applies. The sentence-level sentiments were then combined to produce the final sentiment scores for the entire post.

VADER provides four key sentiment scores for each post: positive, neutral, negative, and compound. The positive score represents the proportion of positive words in the text, the neutral score represented the proportion of neutral words, and the negative score represented the proportion of negative words. The compound score, a normalized measure that sums the overall sentiment of the text, ranged from -1 (extremely negative) to +1 (extremely positive). This comprehensive approach allowed us to capture the nuanced sentiment expressed in Reddit posts effectively.

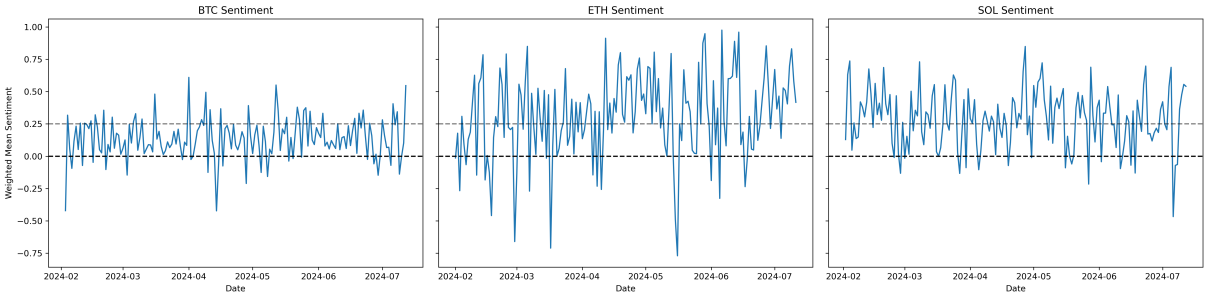

Figure 3.1: Diagram showing each Subreddits Sentiment Score over Time

3.3 Machine Learning Models

3.3.1 Random Forest

Random Forest is an ensemble learning method that operates by constructing a multitude of decision trees during training and outputting the mode of the classes (classification) or mean prediction (regression) of the individual trees. The principle behind Random Forest is to reduce the risk of overfitting by averaging multiple deep decision trees, trained on different parts of the same training set.
Mathematically, for a given input $\mathbf{x}$, the prediction of the i-th tree $h_i(\mathbf{x})$ in a random forest is:

$$f(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^{N} h_i(\mathbf{x})$$

where N is the number of trees in the forest.
Key parameters include:

- Number of trees (N)
- Maximum depth of each tree

3.3.2 Support Vector Classification (SVC)

Support Vector Classification (SVC) aims to find the optimal hyperplane that maximizes the margin between the classes. This hyperplane is defined by the support vectors, which are the data points closest to the hyperplane.
The optimization problem for SVC can be written as:

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 \quad \text{subject to} \quad y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1, \ \forall i$$

where $\mathbf{w}$ is the weight vector, b is the bias, $\mathbf{x}_i$ are the input vectors, and y_i are the class labels.
Key parameters include:

- Kernel type (linear, polynomial, radial basis function)
- Regularization parameter (C)

3.3.3 Gradient Boosting Classification

Gradient Boosting Classification is a sequential ensemble technique that builds models in a stage-wise manner. It optimizes for a loss function by adding weak learners to the model, typically decision trees.
The Prediction function for Gradient Boosting is:

$$\hat{y} = F_M(x) = F_0(x) + \eta \sum_{m=1}^{M} h_m(x)$$

Where:

- $\hat{y}$ is the predicted value.

- $F_0(x)$ is the initial prediction.
- M is the total number of iterations (trees).
- η is the learning rate.
- $h_m(x)$ is the prediction from the m-th weak learner (tree).

Key parameters include:

- Learning rate (η): The learning rate is a hyperparameter that controls the contribution of each weak learner (base model) to the final ensemble model. It scales the output of each weak learner before adding it to the accumulated model.
- Number of estimators: This parameter specifies the number of weak learners (usually decision trees) to be included in the ensemble. It defines how many iterations the boosting process will run.
- Maximum depth of each estimator: This parameter sets the maximum depth of the individual decision trees. It controls the complexity of the trees.

3.4 Training Method and Model Evaluation

The holdout method was utilized to train and evaluate the model. This involves splitting the data into training and testing sets. The model is then evaluated on its predictive capabilities using the testing set to assess its performance on new, unseen data. A wide range of evaluation metrics will be used to thoroughly examine the model's efficiency in predicting price movements. Some, but not all, of these metrics include accuracy, precision, recall, and ROC-AUC. Figure 6.1 in the appendix shows a more condensed visual on evaluation metrics. Finally, we will compare the models and their metrics to make it easier to choose the best models for further investigation.

Binary classification models generally provide a probability of a predicted classification. These probabilities range from zero to one. Generally, a threshold is drawn at 0.5, where anything over is rounded to 1, and anything less is rounded down to 0. This brings the question of where this threshold should be drawn, or how different thresholds affect the models' predictive capabilities. The Receiver Operating Characteristic (ROC) is there to help understand how different thresholds impact model performances. ROC graphs plot the True Positive Rate (TPR) against the False Positive Rate (FPR) across different threshold settings. A perfect model will have an ROC curve that reaches the top left, as there exists a threshold with perfect classification. Figure 6.2 in the appendix shows ROC curves and AUC visuals and its distinction between good and bad models.

Metric	Description	Formula
Accuracy	Measures how often the classifier correctly predicts price movements.	$\frac{\text{True Positives + True Negatives}}{\text{Total Predictions}}$
Precision	Indicates reliability of predictions for price increases.	$\frac{\text{True Positives}}{\text{True Positives + False Positives}}$
Recall (Sensitivity)	Measures ability to catch all real price increases.	$\frac{\text{True Positives}}{\text{True Positives + False Negatives}}$
F1 Score	Harmonic mean of Precision and Recall.	$\frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision + Recall}}$
Specificity	Identifies actual negative movements (price decreases).	$\frac{\text{True Negatives}}{\text{True Negatives + False Positives}}$
False Positive Rate	Proportion of incorrect predictions of price increase.	$\frac{\text{False Positives}}{\text{True Negatives + False Positives}}$
False Negative Rate	Proportion of missed predictions of actual price increase.	$\frac{\text{False Negatives}}{\text{True Positives + False Negatives}}$

Table 3.2: Performance Metrics Definitions

3.5 Global and Local Interpretation

SHAP (SHapley Additive exPlanations) values is a powerful tool used to explain predictions made by machine learning models. SHAP values quantify through bootstrapping, the contribution of each input feature to a specific prediction, providing clear insights into the model's decision-making process. This method enhances the local interpretability by breaking down how each feature influences the prediction. Cooper (2024)

Global interpretability aims to describe the overall behavior of a machine learning model across the entire dataset. In the context of cryptocurrency price prediction, SHAP values can be aggregated to understand which features most influence the model's predictions on average. A common approach for this is to use bar plots showing the mean absolute SHAP value for each feature.

Another method are Beeswarm plots. Beeswarm plots offer a detailed display of SHAP values, providing insights into not only the relative importance of features but also their actual correlation with the predicted outcome of the depdent varibale. Refer to Figure 6.3 in the appendix for a clear visual explanation. Mean absolute SHAP values offer several advantages over traditional feature importance measures. They are more theoretically sound and directly reflect the impact of features on predictions. Additionally, they are expressed in intuitive units, making it easier to grasp their significance. For example, a bar plot might reveal that trading volume has the highest mean absolute SHAP value, indicating that is has the highest influence on price movement predictions. Conversely, features with lower mean absolute SHAP values, such as a particular minor technical indicator, might be less influential to the overall output of the model. Figure 6.4 in the Appendix shows a visual.

Local interpretability focuses on understanding how individual predictions are made by a model. In the realm of cryptocurrency price prediction, SHAP values are invaluable for this purpose, as they detail the contributions of each feature to a specific prediction. Two effective visualization methods for this are waterfall plots and force plots.

These plots provide an in-depth breakdown of how each feature contributes to a single prediction. For instance, a waterfall plot might explain the factors leading to a model's prediction that the price of Bitcoin will increase. The plot illustrates the additive effects of various features, such as market sentiment and recent transaction volume, showing how they combine to influence the final prediction.

By using SHAP values, it is possible to gain a comprehensive understanding of both the global trends affecting cryptocurrency price movements and the specific factors driving individual predictions, thereby enhancing model transparency and interpretability. Figure 6.4 in the appendix shows a visual explaining the concept.

Results

4.1 Results Overview

The average accuracy across all models stands at 57.8%, indicating moderate predictive power, with performances surpassing the 50% threshold . Notably, the models trained

Currency	Model	Accuracy	Precision	Recall	Specificity	FPR	FNR
BTC	RF	60.60%	55.00%	73.30%	50.00%	50.00%	26.70%
	SVM	57.60%	52.60%	66.70%	50.00%	50.00%	33.30%
	XGB	57.60%	53.30%	53.30%	61.10%	38.90%	46.70%
ETH	RF	54.50%	50.00%	73.30%	38.90%	61.10%	26.70%
	SVM	63.60%	60.00%	60.00%	66.70%	33.30%	40.00%
	XGB	54.50%	50.00%	66.70%	44.40%	55.60%	33.30%
SOL	RF	66.70%	84.60%	55.00%	84.60%	15.40%	45.00%
	SVM	63.60%	75.00%	60.00%	69.20%	30.80%	40.00%
	XGB	75.80%	92.90%	65.00%	92.30%	7.70%	35.00%

Table 4.1: All Model type and crypto type comparisons across 7 evaluation metrics

on Solana data demonstrate superior performance, with an average accuracy of 63.1%. This is nearly 10% higher than those observed for Bitcoin and Ethereum models. Such a discrepancy give evidence to suggest that Solana's market might not follow weaker forms of EMH compared to Bitcoin and Ethereum, likely due to the significantly larger number of participants in the BTC and ETH markets, which generally enhances market efficiency. [1]

Highlighted Model Performance

Bitcoin (BTC): The Random Forest model shows a promising balance with an accuracy of 60.6%, precision at 55.0%, and a high recall of 73.3%. However, its F1 score and specificity remain moderate, reflecting a potential trade-off between identifying true positives and avoiding false positives.

Ethereum (ETH): The SVM model stands out with balanced metrics: 63.6% accuracy, 60.0% precision, and 60.0% recall. It exhibits higher specificity (66.7%) compared to other models, suggesting better generalization in distinguishing non-events.

Solana (SOL): The XGBoost model excels with the highest accuracy of 75.8% and precision at 92.9%. Its robust performance is also seen in its impressive specificity (92.3%)

[1] Additional two models where trained: Long Short-Term Memory (LSTM) and Hidden Markov Model (HMM). The results were not promising; details are available upon request.

and low false positive rate (7.7%), indicating it is particularly effective at predicting true positives without many errors. However, the drawback is the low Recall at 65%

Implications of Performance Metrics

The varied performance across different models and cryptocurrencies illustrates the complexity of predictive modeling in financial contexts. Accuracy alone does not tell the full story; metrics such as recall, specificity, and F1 score are crucial for understanding the practical implications of a model in real-world scenarios. For instance, a high recall rate indicates a model's effectiveness in capturing actual positive cases (true positives), crucial for strategies that prioritize capturing all potential gains or signals at the expense of increased false positives.

In the subsequent section, we will delve deeper into the specifics of these models' performances, focusing on their precision-recall trade-offs and the implications of their specificity and false negative rates. We will also examine the Receiver Operating Characteristic (ROC) curves for each model. ROC curves provide a comprehensive visualization of a model's trade-off between true positive rates and false positive rates across different thresholds, offering further insights into their predictive capabilities and robustness. However, we need to reduce the number of models so we can more carefully analyze them. Moving forward, we will select the top 3 models moving forward which are Random Forest, Support Vector Machiens, and Xtreme Gradient Boosting algorith.

4.2 Detailed Analysis

4.2.1 Random Forest

RF	Accuracy	Precision	Recall	F1 Score	Specificity	FPR	FNR
BTC	60.6%	55.0%	73.3%	62.9%	50.0%	50.0%	26.7%
ETH	54.5%	50.0%	73.3%	59.5%	38.9%	61.1%	26.7%
SOL	66.7%	84.6%	55.0%	66.7%	84.6%	15.4%	45.0%

Table 4.2: Performance of the Random Forest Across all Curencies

From table 4.1, it's evident that Solana has the highest accuracy, supported by an 85% precision rate. However, the model's low recall indicates that many price increases are not detected. When the model does predict price increases, they are relatively reliable, with a false positive rate (FPR) of 15%. Although Solana's model excels in predicting price increases, it is less effective at predicting price decreases, evidenced by a 45% false negative rate (FNR).

In contrast, the BTC and ETH models are better at predicting price decreases, with recall values of 73%. Unlike the Solana model, the BTC and ETH models are relatively unreliable at predicting price increases, with FPRs ranging from 50% to 69%. However,

they offer a tradeoff with lower FNRs of 26%, indicating more reliable predictions of price decreases.

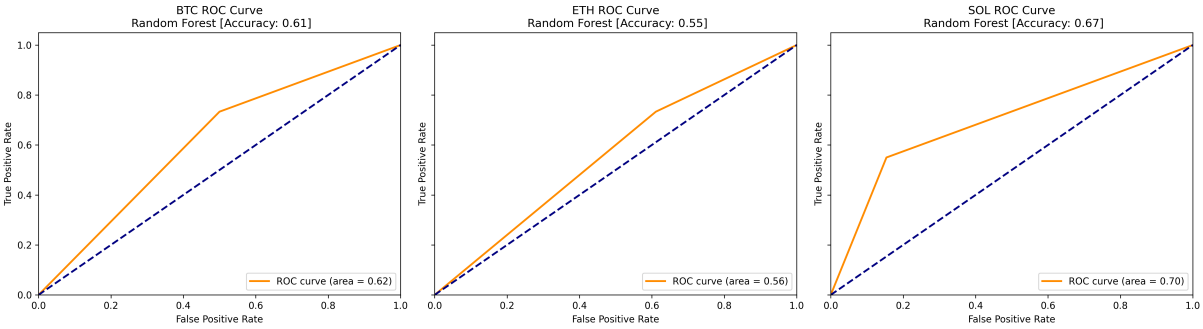

Figure 4.1: This figure plots the ROC curves for the Random Forest models across all currencies.

The ROC curves indicate that the most effective classifier based on ROC is the Random Forest Solana model, which boasts an AUC of 70%. This suggests that the model can correctly classify whether the price will rise or fall 70% of the time, which is 20% better than random chance. This performance highlights that some variables in the dataset possess predictive capabilities. In contrast, the models for Bitcoin and Ethereum show lower AUCs at 62% and 56%, respectively. These lower scores render the models less robust and credible, as they correctly predict the price direction only 56-62% of the time. This analysis addresses questions related to the Efficient Market Hypothesis (EMH), suggesting that Solana's market does not follow the weaker form of EMH based on the higher model performances.

4.2.2 Support Vector Machine

SVM	Accuracy	Precision	Recall	F1 Score	Specificity	FPR	FNR
BTC	57.6%	52.6%	66.7%	58.8%	50.0%	50.0%	33.3%
ETH	63.6%	60.0%	60.0%	60.0%	66.7%	33.3%	40.0%
SOL	63.6%	75.0%	60.0%	66.7%	69.2%	30.8%	40.0%

The performance of these models varied significantly across the cryptocurrencies. The Ethereum and Solana SVMs achieved identical accuracy rates of 63.6%, suggesting a moderate level of predictive reliability. In contrast, the Bitcoin SVM displayed a notably lower accuracy of 57.6%, indicating less predictive consistency compared to the other two. The distinction in performance is particularly evident in the context of false positive rates, where the Bitcoin model reported a high rate of 50%. This high rate implies that half of the upward price predictions made by the Bitcoin SVM are incorrect, rendering these predictions largely unreliable and akin to random chance.

Additionally, the Ethereum and Solana models exhibit lower false positive rates at 30%, enhancing their credibility in predicting price increases. However, despite these lower rates of misprediction, both models suffer from significant drawbacks in terms of recall,

at only 60%. This low recall indicates that a substantial number of actual price increases are missed by the models, being incorrectly classified as price drops. This issue is reflected in the high false negative rates observed in the Ethereum and Solana models. On the other hand, the Bitcoin SVM, while less accurate overall, shows lower false negative rates, suggesting that its predictions of price declines are relatively more reliable. However, it also suffers from the lowest specificity, meaning it frequently misses identifying actual negative outcomes. This varied performance across different metrics highlights the challenges in deploying SVMs for predictive accuracy in cryptocurrency markets, underscoring the need for further refinement and adjustment of the models to improve their predictive capabilities.

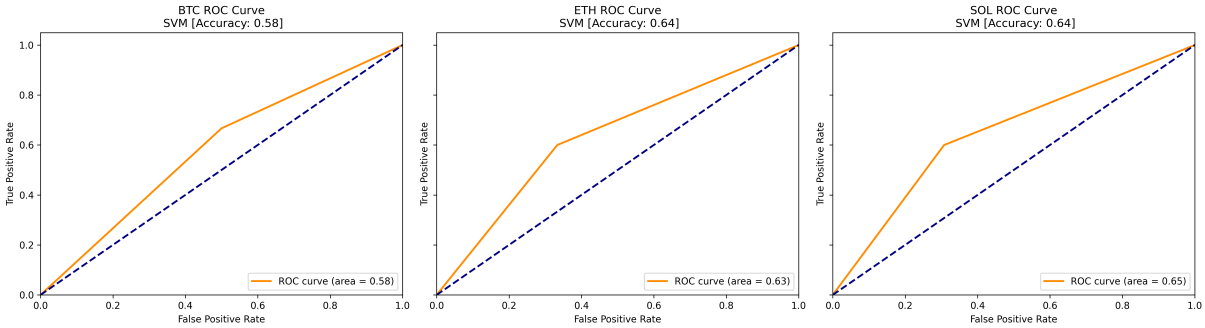

Figure 4.2: This figure plots the ROC-AUC curves for the SVM models across all currencies.

The ROC curves for the Support Vector Machine models illustrate different levels of effectiveness in distinguishing between price increase and decrease scenarios. The Solana SVM model leads with an AUC of 65%, which suggests that it ranks a randomly chosen positive instance (price increase) higher than a randomly chosen negative instance (price decrease) 65% of the time. This indicates a moderate discriminative ability above random guessing. The Ethereum model follows with an AUC of 63%, showing a similar but slightly lower capability. Meanwhile, the Bitcoin model displays the least effectiveness with an AUC of 58%, pointing to its relatively weaker performance in predicting price directions under varying threshold settings. These AUC values shed light on the potential inefficiencies in cryptocurrency markets, particularly reflecting how information is assimilated into prices, which is a core consideration of the Efficient Market Hypothesis (EMH)

4.2.3 eXtreme Gradient Boosting

XGB	Accuracy	Precision	Recall	F1 Score	Specificity	FPR	FNR
BTC	57.6%	53.3%	53.3%	53.3%	61.1%	38.9%	46.7%
ETH	54.5%	50.0%	66.7%	57.1%	44.4%	55.6%	33.3%
SOL	75.8%	92.9%	65.0%	76.5%	92.3%	7.7%	35.0%

The Extreme Gradient Boosting (XGB) models tailored to cryptocurrency price predictions exhibit varied performance across Bitcoin, Ethereum, and Solana, with Solana's

model significantly outperforming the others. The Solana XGB model demonstrates an impressive overall accuracy of 75.8%, indicating a high degree of reliability in its predictions. This model shines particularly in predicting upward price movements, boasting a remarkable precision rate of 92.99%. Such high precision, coupled with a low false positive rate of only 7.7%, suggests that the up predictions made by the Solana model are predominantly accurate, making it a robust tool for traders focusing on positive trends.

However, the model is not without its limitations, as evidenced by a recall rate of only 65%. This suggests that while the predictions it makes are reliable when they occur, the model fails to capture a substantial portion of actual upward movements, incorrectly predicting them as negative. This is further supported by a false negative rate of 35%, indicating that many potential gains are overlooked. In comparison, the Bitcoin XGB model lags significantly behind, with an accuracy of just 57.6% and suffering from high false positive and false negative rates of 34 and almost 47%, respectively. This indicates a general unreliability in its predictive accuracy. The Ethereum model fares similarly poorly in terms of upward price movement predictions, with a precision of only 50% and a very high false positive rate of 55.6%. This lack of reliability in predicting positive trends further underscores the challenges faced by these models in achieving the high level of performance exhibited by the Solana XGB model.

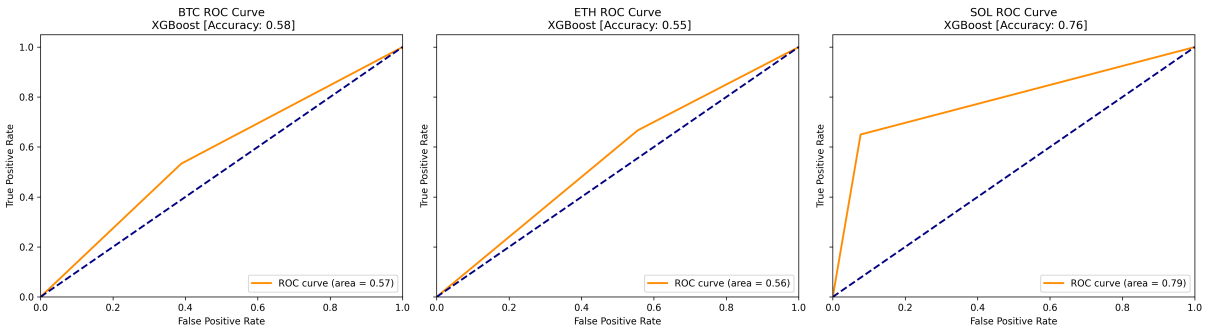

Figure 4.3: This figure plots the ROC-AUC curves for XGBoost models across all currencies.

The ROC curves for the XGBoost models depicts a disparity in their ability to differentiate between scenarios of price increases and decreases. The Solana XGB model demonstrates superior performance with an AUC of 0.79, indicating that it ranks a randomly chosen positive instance (price increase) higher than a randomly chosen negative instance (price decrease) 79% of the time. This high discriminative ability substantially exceeds random guessing, suggesting a strong predictive capability. In contrast, the Bitcoin and Ethereum models yield AUCs of 0.57 and 0.56, respectively. These lower values indicate much weaker performances in distinguishing between the price movements, highlighting potential limitations in their predictive accuracy under different threshold settings. This variance in AUC values among the cryptocurrencies further contributes to the discussion on market efficiency. A case can be made that Bitcoin and Ethereum exhibit characteristics of weaker forms of the Efficient Market Hypothesis (EMH), whereas Solana appears to not align with the weaker form of EMH.

We will employ SHAP Bee Swarm plots and Waterfall plots to elucidate the reasons

behind the model's classification decisions and to explore how various variables interact to shape the predicted price. Initially, we will delve into the most complex graph to understand a model's global interpretation. Our focus will be on Solana's XGBoost model, which demonstrated the highest accuracy and overall metrics in our analysis.

4.2.4 Global Interpretation

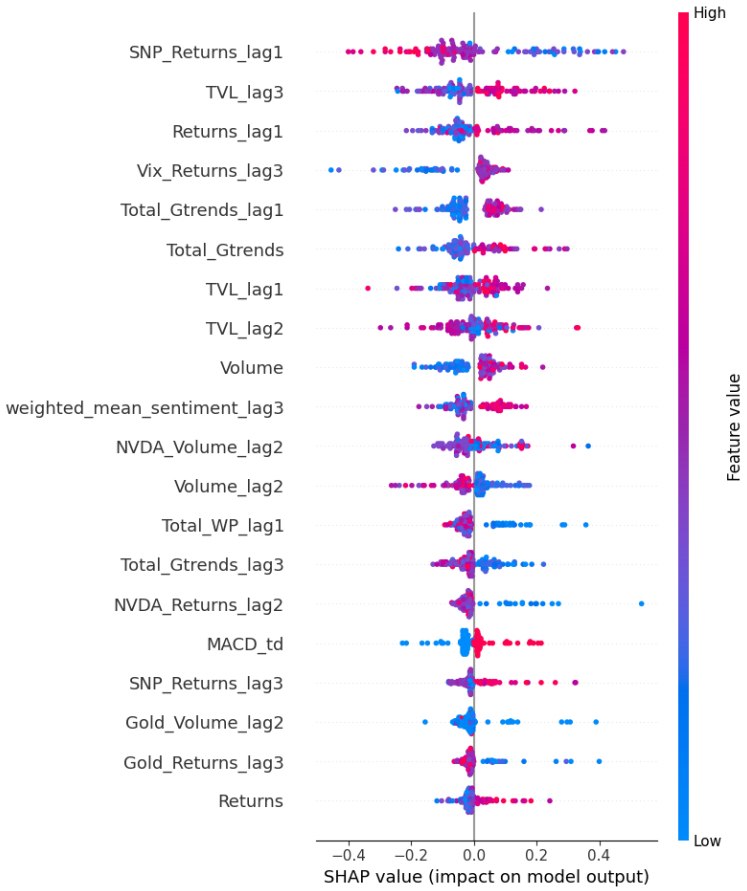

Figure 4.4: This figure depicts the Beeswarm plot of the XGBoost model for Solana modelling.

To effectively address this plot, our analysis will concentrate on two key aspects: 1) the importance of each variable in predicting future returns, and 2) the correlation between each variable and future prices, as indicated by the model.

Starting off, the S&P returns at lag 1, TVL (Total Value Locked) lagged by three days, and the returns of Solana itself are the three most important variables for predicting Solana price movements. The significance of S&P returns at lag 1 is clearly illustrated in the graph. There is a strong negative correlation between the S&P returns and subsequent Solana price movements. This correlation is visually represented by the red dots on the left side of the swarm plot, indicating negative contributions to the model. In other words, when the S&P experiences losses, Solana's price is likely to drop the next day, suggesting that broader market trends impact Solana.

Interestingly, TVL changes from three days prior play a crucial role and are positively correlated with Solana prices. According to the model, if the TVL saw a significant increase three days ago, it is likely to be reflected in higher Solana prices. This relationship is visually represented by clusters showing higher SHAP values corresponding to higher percentage increases in TVL, emphasizing the importance of TVL as an indicator of future price movements

Additionally, the returns of Solana itself are also an important predictor, although the swarm plot shows a less clear separation in values. However, the general trend indicates that higher returns are associated with higher prices in subsequent days. This suggests a momentum effect where positive returns in Solana tend to lead to continued price increases, albeit with some variability.

Tying the debate to the second research question regarding marketing-related variables to explain next-day price movements is mildly confirmed by the swarm plot. We see that Google Trends data is the fifth most important variable in explaining price movements. In this example, there is a clear positive correlation: higher Google Trends interest in crypto-related terms translates to positive price movements in Solana's price.

Secondly, the Reddit data also significantly impacts the model's predictions. The strongest variable is the 3rd lag, indicating how Reddit data from three days ago reflects tomorrow's price movement. Based on the XGB model, there is a positive relationship with price movements. When the Reddit sentiment scores increase three days prior, the model predicts upward price movements for Solana.

These findings support the notion that marketing-related variables, such as Google Trends and Reddit sentiment, play a role in predicting future price movements of Solana.

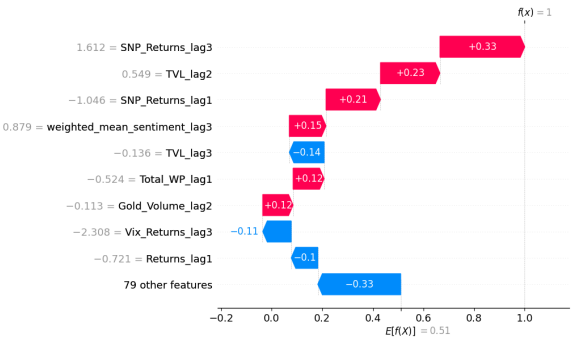

Figure 4.5: Waterfall plot for a future day instance of upward price movement. Model predicted up movement. Driven by High SNP and TVL percent changes

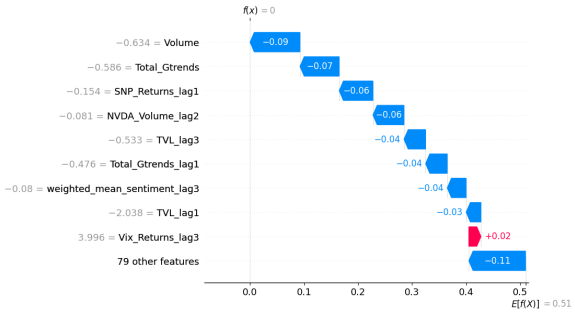

Figure 4.6: Waterfall plot depicting a future day where prices went down. Opposite of previous example, model predicted down movement. Driven by low solana volme, low views on google trends the day before, and lower Nvidia returns.

Figure 4.5 presents an example from the training set where the actual next-day price was 1, and the predicted value was also 1. It is important to note that the model achieved 100% accuracy in predicting outcomes on the training data, indicating that local interpretation has some validity. As explained in the methodology, the model begins with an average price prediction of 0.51 (the mean of binary outcomes 1 and 0 in the training

dataset). The analysis shows that 79 other features in the dataset (not mentioned in the waterfall plot) influence the price prediction downward.

We see Low TVL (Total Value Locked) percentage changes day-over-day reduce the overall prediction, highlighting a positive correlation between TVL and prices. The weighted-mean-sentiment for a 3-day lag of Reddit sentiment pushes the model's prediction upwards due to high sentiment values 3 days prior, indicating its importance and a correlation with price.

As detailed above and shown in the swarm plot (Figure 4.4), the two most influential variables are introduced last. Initially, the S&P return lag 1, which represents negative returns, enhances the model's prediction. Similarly, positive returns from the third lag of the S&P also increase the model's prediction, demonstrating the complex interactions between different variables. The relationships are not one-sided where one variable has a distinct correlation, but rather the aggregation and interaction of many variables contribute to the output. The waterfall plot breaks down the contributions by variable. Lastly, TVL lag 2 also positively influences the model due to its high percentage change value, further illustrating the dynamic interplay among various factors.

Figure 4.6 illustrates an example where the next-day prices declined instead of increasing. This figure provides an opportunity to understand which variable setups lead the model to predict downward movements in prices. In this example, trading volume emerges as the most influential variable. This significance is attributed to significant drops in volume on the previous day. Similar trends are observed with drops in Google Trends interest and declines in the returns for the S&P and Nvidia. These variables collectively indicated to the model that the next-day prices for Solana were likely to decrease, which indeed it did.

Conclusion

The primary objective of this thesis was to predict future price movements of Bitcoin, Ethereum, and Solana. A secondary goal was to extract insights from the models to identify which features are most important in forecasting these price movements. This research ultimately aimed to test the Efficient Market Hypothesis (EMH) within the context of the modern, current day cryptocurrency market.

One of the key findings was that Solana was easier to predict in terms of accuracy and recall compared to Ethereum or Bitcoin. This may serve as evidence supporting the Efficient Market Hypothesis (EMH). The study suggests that Solana does not follow the weaker form of EMH, while Bitcoin and Ethereum adhere to weaker forms. These conclusions were drawn based on the efficiency of machine learning algorithms in predicting next-day prices. The accuracy for Bitcoin and Ethereum ranged from 52%-64%, while models for Solana showed a higher accuracy range of 64%-76%. This significant gap in predictive power could be attributed to the principles of the Efficient Market Hypothesis. Additionally, Bitcoin and Ethereum have trading volumes that are 9 times and 5 times higher than that of Solana, respectively, which introduces more traders and market makers, thereby making these markets more efficient and harder to predict.

The analysis identifies a strong correlation between next-day Solana prices and the S&P 500 as well as TVL. Although tree-based algorithms, particularly XGBoost, are challenging to interpret, one workaround was to use global and local interpretation methods, such as SHAP values. Through the SHAP analyis, we determined that the S&P 500 and TVL are the two most influential variables for predicting next-day Solana prices. S&P 500 returns on a 1-day lag are inversely correlated with Solana prices; a drop in S&P 500 returns predicts a likely drop in Solana prices two days later. Conversely, an increase in TVL tends to correlate positively with future Solana prices.

The third key finding relates to sentiment-based variables such as Google Trends and Reddit sentiment. The swarm plot highlighted the importance of the Google Trends variable and its first lag. We observe a clear correlation where higher interest on Google Trends also correlates with higher returns the next day for Solana, with the model achieving an accuracy of 76% and a precision of 92%. However, recall does diminish overall performance. A similar pattern is observed with the Reddit data; Reddit sentiment also correlates positively with higher next-day prices, but only at the third lag. Notably, in the swarm plot, Reddit sentiment ranks lower on the y-axis compared to Google Trends, suggesting that within the entire model, Google Trends data is more informative of next-day prices compared to Reddit sentiment data.

Revisiting our first null hypothesis, we cannot reject the null hypothesis. We do not find enough evidence to suggest that technical indicators have predictive power. This was derived from the fact that technical indicators did not have high SHAP values indicating low importance.

In regards to our second hypothesis testing social media sentiment scores on future price movements, we find enough evidence to reject the null hypothesis. Social media sentiment scores do play an important factor in predicting the future price movements according to the XGBoost model.

Lastly, our last null hypothesis positing that Macroeconomic indicators do not have predictive power, we do find enough evidence to reject it. Through the SHAP value analysis, we find that S%P 500 played the most important role in predicting future price movement.

In conclusion, this thesis provides valuable insights into the predictability of cryptocurrency prices, particularly highlighting the varying levels of efficiency across different assets within the market. The findings indicate that while Solana exhibits more predictable price movements, potentially challenging the Efficient Market Hypothesis, Bitcoin and Ethereum align more closely with the principles of market efficiency, rendering them less predictable. The importance of macroeconomic indicators and social media sentiment further underscores the complex interplay of factors influencing cryptocurrency markets. These insights contribute to the broader understanding of market dynamics and may serve as a foundation for future research aimed at refining predictive models and exploring the nuances of market efficiency in the rapidly evolving cryptocurrency landscape.

Limitations and Future Research

One of the major limitations of this paper is the small dataset size. I began collecting daily Reddit data in February 2024, which to date comprises approximately 190 rows of data. Many deep learning frameworks, such as LSTMs, perform better with larger datasets. This could explain the low performance scores of LSTMs observed in this study. In future research, it would be interesting to assess the performance of deep learning models with datasets that are three to four times larger.

A second limitation is the binary nature of the indicators used. Expanding beyond the simple buy and sell options to include a 'hold' or neutral option could provide more nuanced insights into model performance. Similarly, the technical indicators were also binary; introducing a third category could yield more complex and informative results.

Another challenge was the analysis of Reddit data. The free Reddit API allows downloading of post data but not comments. Omitting comments can lead to significant information loss, as some argue that the true sentiment of a post is better reflected in its comments rather than in the top 100 posts. The number of votes can serve as a proxy for sentiment analysis, with higher vote counts potentially indicating stronger sentiment in the comments. Future methodologies could aim to include comment scraping for a more detailed analysis of sentiments.

Lastly, we faced technical compute limitations. Hyperparameter tuning for neural networks, and especially for XGBoost, requires significant computational resources and time. I did not have the computing power necessary to fully optimize the XGBoost model; exploring all parameters with possible tunings could take weeks with my current setup. Future research could benefit from acquiring more compute power to evaluate results with a perfectly tuned XGBoost model.

5.0.1 Managerial and Academic Implications

From a managerial perspective, the findings of this thesis provide actionable insights for investors and financial analysts seeking to navigate the cryptocurrency market. The identified importance of macroeconomic indicators, such as the S&P 500, and the predictive power of social media sentiment suggest that managers should incorporate these variables into their decision-making frameworks to enhance forecasting accuracy and risk management strategies.

For academic research, this study contributes to the ongoing debate about the applicability of the Efficient Market Hypothesis in the context of digital currencies. By demonstrating the varying degrees of market efficiency across different cryptocurrencies, this research opens new avenues for future studies to explore the unique characteristics and behaviors of these emerging financial assets, potentially leading to more refined theories and models in the field of financial economics.

Appendix

6.1 Introduction to Cryptocurrencies and Blockchain

Cryptocurrencies and blockchain technology not only represent a shift in the way we think about money and financial transactions, but it also revolutionizes applications associated with peer-to-peer technology. The era of blockchain technology began in 2008 by an anonymous person or group known as Satoshi Nakamoto, whose whitepaper titled "Bitcoin: A Peer-to-Peer Electronic Cash System" laid the foundation for the first cryptocurrency. Blockchain technology in big is celebrated for its transparency, security, and immutability of transactions, presenting a robust alternative to conventional financial systems.

How Cryptocurrencies Work

At the core of cryptocurrencies is blockchain technology, a distributed ledger that records all transactions across a network of computers. Each block in the blockchain contains a list of transactions and a reference to the previous block, creating a secure chain of data, hence the name blockchain. This structure prevents tampering and ensures that once data is recorded, it cannot be altered without consensus from the network.

Initially, cryptocurrencies like Bitcoin used a mechanism called Proof of Work (PoW) to validate transactions and secure the network. PoW requires participants, known as miners, to solve complex mathematical puzzles, consuming significant computational power and energy. As of 2023, Bitcoin's network consumed approximately 120 terawatt-hours (TWh) of electricity annually, comparable to the energy consumption of a small country.

The Shift to Proof of Stake

Due to the high energy consumption and scalability issues associated with PoW, many cryptocurrencies are transitioning to Proof of Stake (PoS). PoS reduces the need for computational power by allowing validators to create new blocks and verify transactions based on the number of coins they hold and are willing to "stake" as collateral. This method is more energy-efficient and can handle a higher volume of transactions, making it a more sustainable option for the future.

Introduction to Ethereum and Solana

Ethereum, introduced in 2015 by Vitalik Buterin, expanded the capabilities of blockchain by enabling smart contracts and decentralized applications (DApps). These features allow developers to build and deploy a wide range of applications on the Ethereum network, from financial services to gaming and beyond. Ethereum transitioned to PoS in 2022 with its "Ethereum 2.0" upgrade, significantly reducing its energy consumption and improving transaction speeds. Solana, another significant player in the cryptocurrency space, was

founded by Anatoly Yakovenko in 2017. It is known for its exceptionally fast transaction speeds, capable of handling up to 65,000 transactions per second and growing, addressing scalability issues that have plagued other blockchain networks. Solana's innovative consensus mechanism, Proof of History (PoH), works in conjunction with PoS to further enhance speed and efficiency.

6.2 Figures and Tables

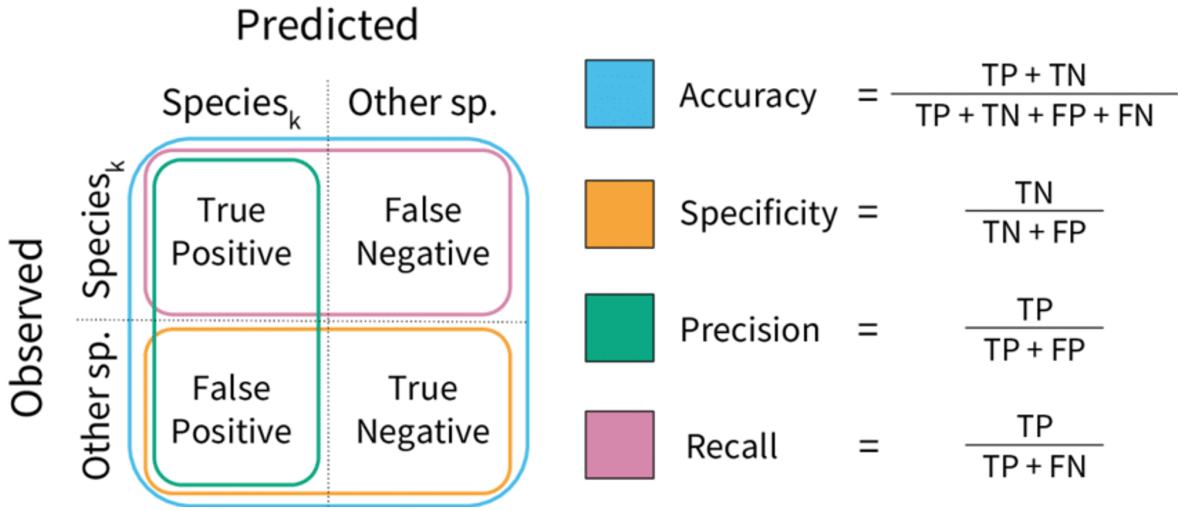

Figure 6.1: Easy-to-understand visual of a classification matrix. This image was sourced from AlmenBetter.

Figure 6.2: ROC-AUC Visualization. Diagram sourced from Riccardo Di Sipio.

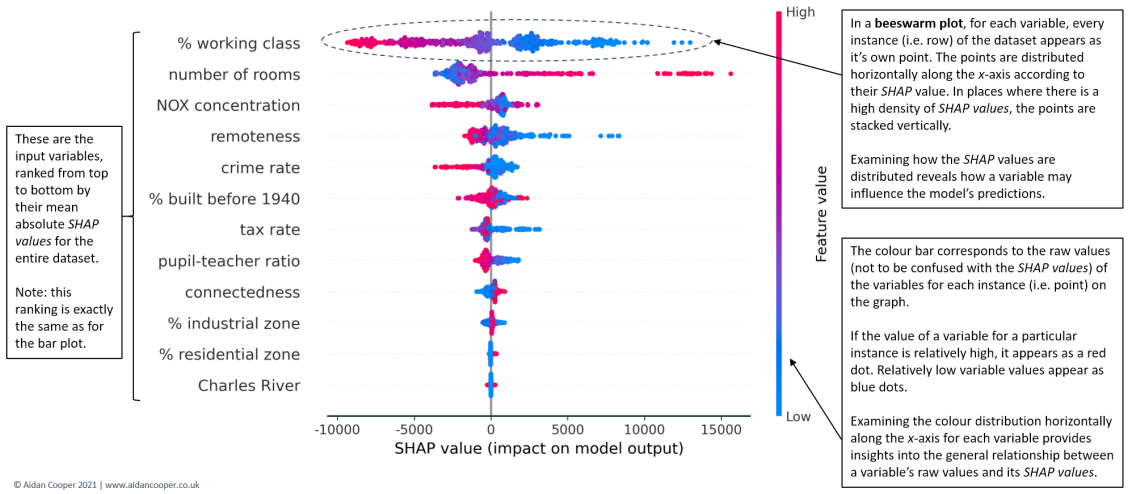

Figure 6.3: Educational illustration of a SHAP Bee Swarm Plot. Credit: Aidan Cooper.

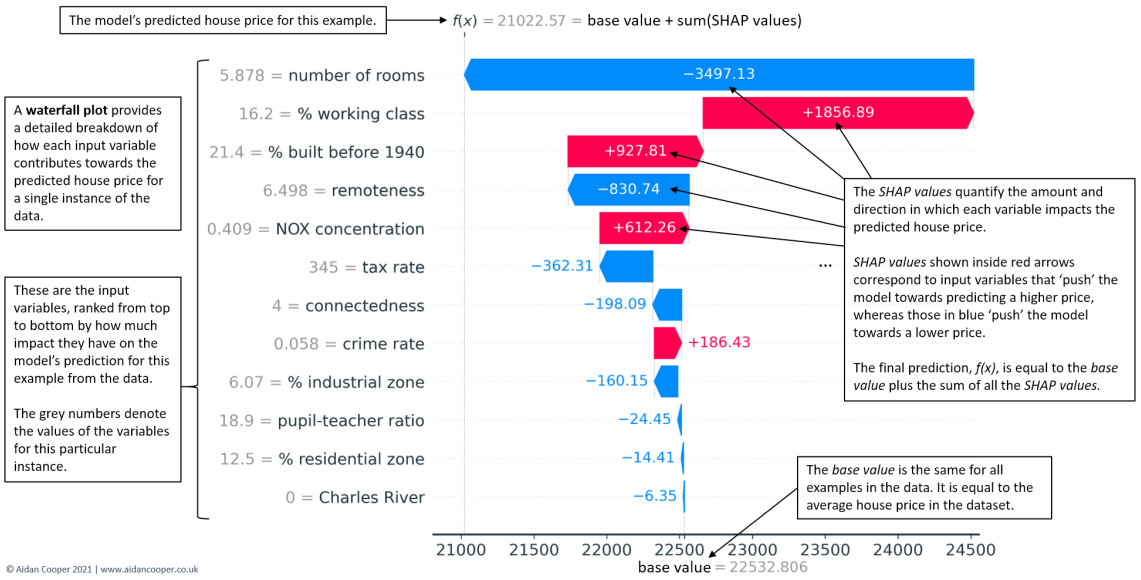

Figure 6.4: Educational illustration of a Waterfall Plot. Credit: Aidan Cooper.

References

Borges, M. R. (2010). Efficient market hypothesis in european stock markets. *The European Journal of Finance*, *16*(7), 711–726.

Chan, W. S. (2003). Stock price reaction to news and no-news: drift and reversal after headlines. *Journal of financial economics*, *70*(2), 223–260.

Chen, Z., Li, C., & Sun, W. (2020). Bitcoin price prediction using machine learning: An approach to sample dimension engineering. *Journal of Computational and Applied Mathematics*, *365*, 112395.

Cooper, A. (2024, Apr). *Explaining machine learning models: A non-technical guide to interpreting shap analyses.* Author. Retrieved from `https://www.aidancooper.co.uk/a-non-technical-guide-to-interpreting-shap-analyses`

Fama, E. F. (1970). Efficient capital markets. *Journal of finance*, *25*(2), 383–417.

Fama, E. F. (1995). Random walks in stock market prices. *Financial analysts journal*, *51*(1), 75–80.

Kahneman, D., & Tversky, A. (2013). Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part i* (pp. 99–127). World Scientific.

Kang, H.-J., Lee, S.-G., & Park, S.-Y. (2022). Information efficiency in the cryptocurrency market: The efficient-market hypothesis. *Journal of Computer Information Systems*, *62*(3), 622–631.

Kara, Y., Boyacioglu, M. A., & Baykan, Ö. K. (2011). Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the istanbul stock exchange. *Expert systems with Applications*, *38*(5), 5311–5319.

Khan, W., Ghazanfar, M. A., Azam, M. A., Karami, A., Alyoubi, K. H., & Alfakeeh, A. S. (2022). Stock market prediction using machine learning classifiers and social media, news. *Journal of Ambient Intelligence and Humanized Computing*, 1–24.

Kraaijeveld, O., & De Smedt, J. (2020). The predictive power of public twitter sentiment for forecasting cryptocurrency prices. *Journal of International Financial Markets, Institutions and Money*, *65*, 101188.

Kristoufek, L. (2013). Bitcoin meets google trends and wikipedia: Quantifying the relationship between phenomena of the internet era. *Scientific reports*, *3*(1), 3415.

Kumbure, M. M., Lohrmann, C., Luukka, P., & Porras, J. (2022). Machine learning techniques and data for stock market forecasting: A literature review. *Expert Systems with Applications*, *197*, 116659.

Le Tran, V., & Leirvik, T. (2020). Efficiency in the markets of crypto-currencies. *Finance Research Letters*, *35*, 101382.

Mai, F., Shan, Z., Bai, Q., Wang, X., & Chiang, R. H. (2018). How does social media impact bitcoin value? a test of the silent majority hypothesis. *Journal of management information systems*, *35*(1), 19–52.

McNally, S., Roche, J., & Caton, S. (2018). Predicting the price of bitcoin using machine learning. In *2018 26th euromicro international conference on parallel, distributed and network-based processing (pdp)* (pp. 339–343).

Mudassir, M., Bennbaia, S., Unal, D., & Hammoudeh, M. (2020). Time-series forecasting of bitcoin prices using high-dimensional features: a machine learning approach. *Neural computing and applications*, 1–15.

Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015). Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. *Expert systems with applications*, *42*(1), 259–268.

Phaladisailoed, T., & Numnonda, T. (2018). Machine learning models comparison for bitcoin price prediction. In *2018 10th international conference on information technology and electrical engineering (icitee)* (pp. 506–511).

Rognone, L., Hyde, S., & Zhang, S. S. (2020). News sentiment in the cryptocurrency market: An empirical comparison with forex. *International Review of Financial Analysis*, *69*, 101462.

Trends, G. (2024). *Google trends.* Google. Retrieved from `https://trends.google.com/trends/`

Usmani, M., Adil, S. H., Raza, K., & Ali, S. S. A. (2016). Stock market prediction using machine learning techniques. In *2016 3rd international conference on computer and information sciences (iccoins)* (p. 322-327). doi: 10.1109/ICCOINS.2016.7783235

Weng, B., Lu, L., Wang, X., Megahed, F. M., & Martinez, W. (2018). Predicting short-term stock prices using ensemble methods and online data sources. *Expert Systems with Applications*, *112*, 258–273.

Wikipedia. (2024). *Wikipedia page views data source.* Retrieved from `https://pageviews.wmcloud.org/?project=en.wikipedia.org`