Luca Ordronneau

OCEAN Token Sentiment Analysis

Part II

Ocean Data Challenge

Discord :
lucanew#3793

GitHub :
lucaordronneau

Version of
July 21, 2023

Contents

Introduction

This report stems from the second challenge of a competition series that focuses on uncovering the intricate relationship between social media engagement and the fluctuating price of the OCEAN token. Building on the knowledge and experience gathered from the initial challenge, this subsequent stage has allowed for a more complex exploration of the task at hand, effectively pushing boundaries of data analysis and machine learning techniques.

The aim of this report is to present a thorough and sophisticated analysis conducted to understand the impacts of social media, particularly Twitter, on the OCEAN token's price. The chosen approach deviates significantly from the one applied in the first challenge, introducing an innovative concept known as the **Impact Score**. This score measures the influence of individual or groups of tweets over specific periods (an hour, a day, or a week) on the future price of the OCEAN token, enabling trading strategies to be created and providing valuable information for potential investors.

Rather than focusing on a classification model that categorises tweets as 'bullish', 'neutral' or 'bearish', this report goes further by implementing a regression model that combines Twitter volume metrics, tweet sentiment and user influence. This analysis employed advanced machine learning techniques, specifically, a Gradient Boosting model with hyper-parameters tuning and cross-validation steps. The crux of the research centers on the construction of the Impact Score, which is targeted by these machine learning models, as well as the creation of new features.

Throughout this report, we delve into the methodology, the challenges faced, and the insights garnered, contributing to the broader understanding of cryptocurrency market dynamics.

Chapter 1

Methodology

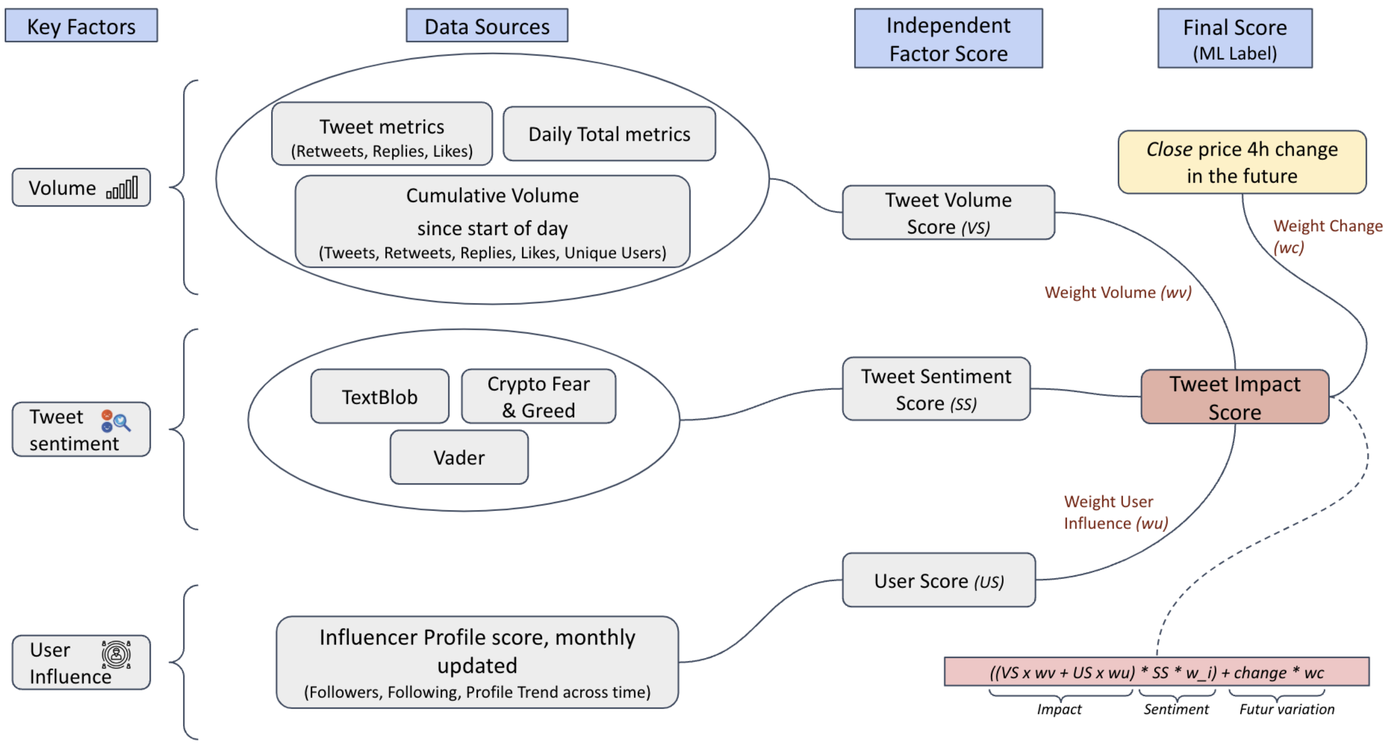


Figure 1.1: Social Media Engagement Impact score on OCEAN Token Price creation

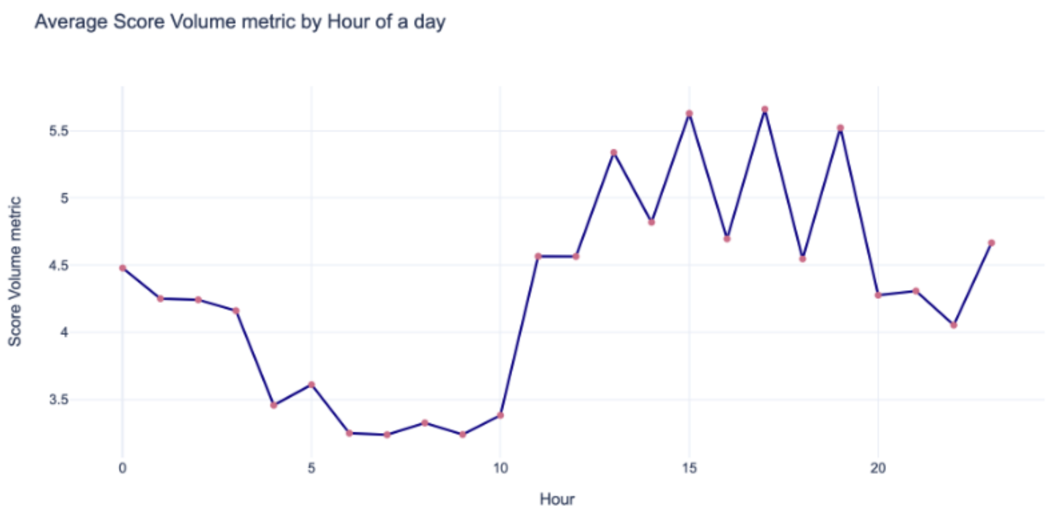
1.1 Key Factors

In our pursuit to comprehend the multifaceted relationship between social media engagement and the OCEAN token's price, we have pinpointed three pivotal factors: **Volume metrics**, **User Influence**, and **Tweet Sentiment**. The subsequent sections delve into each of these variables, elucidating their relevance and impact within the larger cryptocurrency market context. The pipeline shown in 1.1 represents all the stages involved in creating the Impact Score, which originates from the Keys Factors identified and which we will detail below

1.1.1 Volume engagement score

The factor of 'Volume' encapsulates a range of tweet metrics, all of which contribute to understanding the depth of social media engagement around the OCEAN token. The metrics we consider include the immediate engagement a tweet garners (retweets, replies, likes), as well as a time-bounded accumulation of these metrics from the start of the day when the tweet was posted. We also incorporate an aggregate measure of total engagement for the day, irrespective of the time series. Together, these metrics form an amalgamated 'Volume' score that has shown strong correlation with the OCEAN token price.

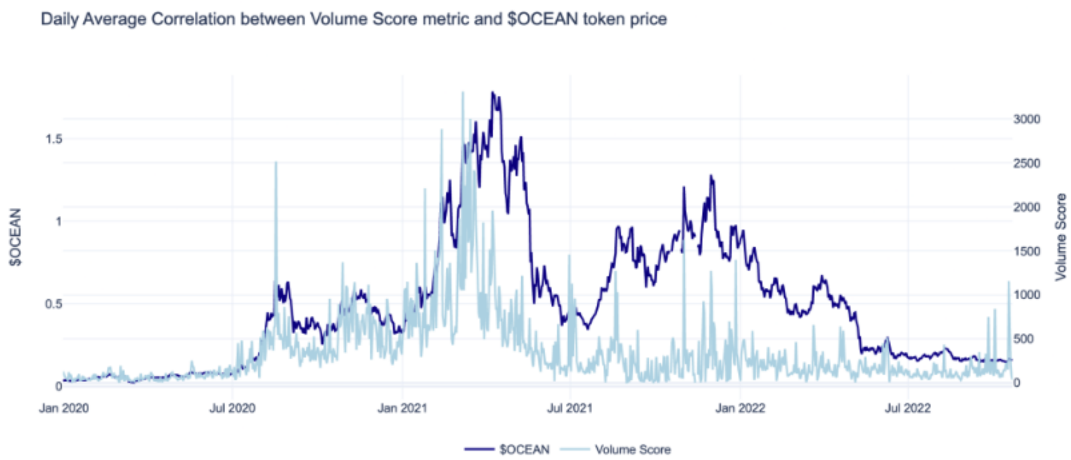
The graph below shows the average Score Volume metrics, one of the features of the final Score Volume. This score is directly linked to likes, retweets and replies. It can be seen that activity over the course of a day starts at 10am.



Below is a table outlining the individual correlation coefficients for each contributing element:

Metric	Correlation Coefficient
Tweet Metric	0.33
Cumulative Tweet Metrics Since Start of Day	0.46
Total Tweet Volume Metric of the Day	0.6
Score Volume	**0.59**

Table 1.1: Correlation of Volume Components with OCEAN Token Price



The final 'Score Volume' is a weighted average of the Tweet Metric, Cumulative Tweet Metrics since the start of the day, and the Total Metrics for the day. The weights for each component are derived from their respective correlation with the token price. Further elaboration on this calculation will be provided in the subsection 1.2.1.

Data Sources

The primary source of data utilized for this analysis is derived from the dataset explicitly provided for the challenge. We employed a variety of data manipulation techniques such as aggregations and calculations of cumulative volume to transform and derive insights from the original dataset.

1.1.2 User influence score

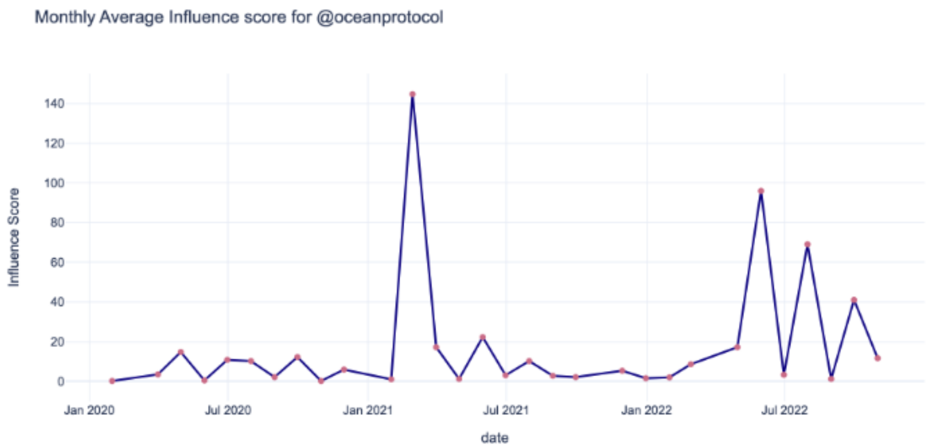
The 'User Influence' factor underscores the role of the tweet's author in shaping the OCEAN token's price. The ripple effects of a single tweet can significantly sway the token's value, particularly when the tweet originates from a highly influential person such as a celebrity, tech entrepreneur, or a reputed crypto analyst. These influential individuals can, through their positive or negative comments, steer the price trajectory by influencing their followers or the broader public's purchase or sale decisions.

Quantifying this 'user influence' required a meticulous examination of various elements. We exploited the metrics associated with the tweet published, as well as a compilation of the user's tweet metrics over a month, including the number of tweets published. In addition, time trends were also taken into account, as a user's influence can vary. As a result, the user's score, ranging from 0 (not influential) to 1 (very influential), summarises the user's dynamic influence and is recalibrated each month to take account of potential variations in influence.

The average daily 'User Influence' score demonstrating a correlation of 0.54 with the OCEAN token's price underscores the significance of influential voices in shaping the token's market value. This metric was calculated by averaging the 'User Influence' scores of all users in a given day, effectively capturing the overall influence trend in the cryptocurrency community for that day. The score's correlation with the OCEAN token price suggests that on any given day, shifts in the average influence of the users on social media can substantially sway the token's price.

Despite these encouraging results, our study could benefit from a broader scope of data, specifically the followers/following ratio over time for each user. Regrettably, access to this information via the Twitter API entails a premium, thus posing a constraint to our current analysis. However, including this metric in the future would undeniably enrich the 'User Influence' score as the followers/following ratio is a potent indicator of a user's sway.

In the figures below, the influence score has not yet been normalised between 0 and 1. We can see that the official account of the Ocean Protocol project (@oceanprotocol) is in second place in terms of influence among all the tweets that have the OCEAN cashtags.



Data Sources

Analogous to the 'Volume' score, our 'User Influence' metric also primarily relies on the dataset furnished for the challenge, necessitating various manipulative and aggregative transformations to distill meaningful insights.

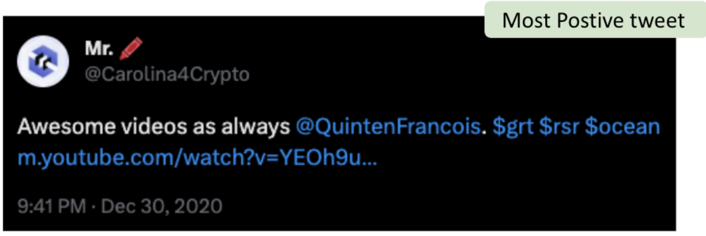
*Our analysis, though comprehensive, could be further enriched by incorporating additional data sources. Notably, accessing follower and following data through the Twitter API could bolster the granularity of our influence scores. Furthermore, harnessing paid APIs like **Social Blade**, which provides user statistics across platforms such as YouTube, Twitch, Instagram, and Twitter, could offer a broader perspective on user influence. Assigning scores to influential profiles across these platforms by Social Blade would have significantly enhanced the accuracy and robustness of our impact score for the machine learning model. These enhancements represent potential avenues for future refinement of our analysis.*

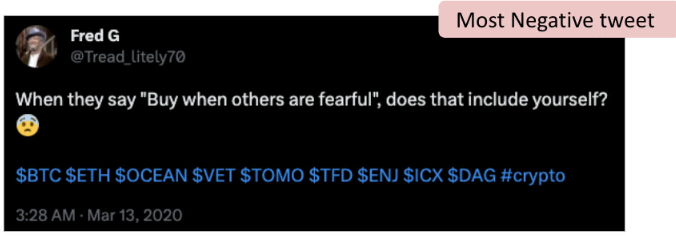
1.1.3 Sentiment score

The 'Sentiment Score' offers an additional layer of analysis, capturing the emotional nuances present in the cryptocurrency discourse. To craft this score, we amalgamated several data sources.

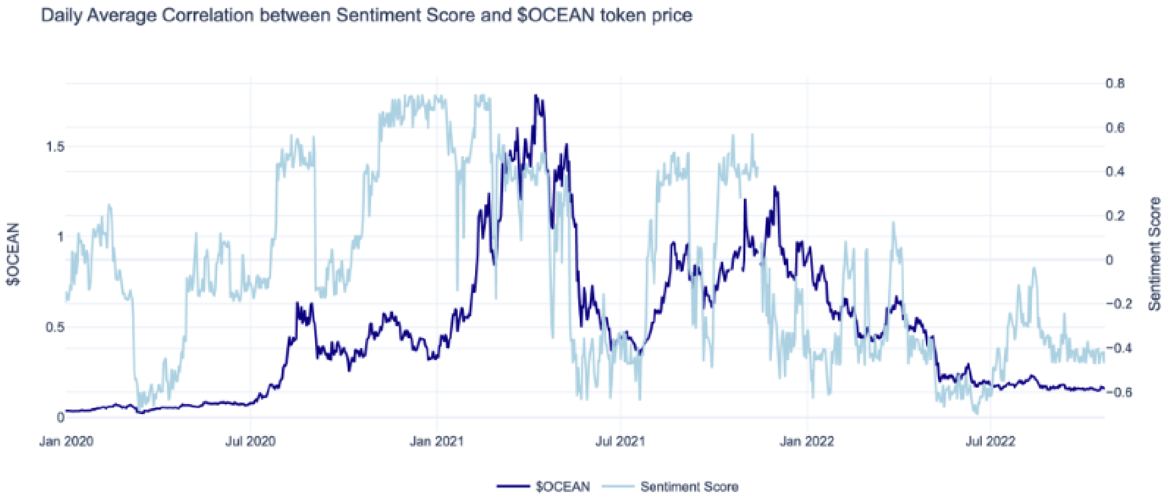
The 'Fear & Greed' crypto index serves as the foundation of our sentiment analysis. This index is a composite measure encompassing volatility (25%), market momentum (25%), social media sentiment (15%), market dominance (10%), and trends (10%). It encapsulates the overarching market sentiment, offering a broad context for our sentiment score.

In conjunction with the market sentiment, we leveraged two powerful Python libraries, Vader and TextBlob, to analyze the sentiment within tweets. Vader, or Valence Aware Dictionary for sEntiment Reasoning, excels at understanding the sentiment in social media text, accounting for both the polarity (whether the sentiment is positive or negative) and the intensity of the sentiment. On the other hand, TextBlob provides a straightforward API for diving deep into common natural language processing (NLP) tasks, including part-of-speech tagging, noun phrase extraction, and sentiment analysis. We calculated sentiment scores using TextBlob's polarity and subjectivity metrics, specifically the product of the two, which helped gauge the tweet's emotional intensity and the strength of the sentiment expressed.





Through the integration of the Fear & Greed index, Vader, and TextBlob outputs, we forged a comprehensive sentiment score. This metric exhibited an encouraging daily average correlation of 0.58 with the OCEAN token price, solidifying its pivotal role in our 'Impact Score'. The sentiment score ranges from -1 to 1, representing very negative to very positive sentiment respectively, further refining the final 'Impact Score' for each tweet.



Tweets text pre-processing

As for the Part I of the data challenge, the same technique has been employed to pre-process the text:

Twitter data is known for its **lack of structure and high levels of noise**. Consequently, the collected Twitter data required extensive preprocessing to make it useful for sentiment analysis. A set of **18 preprocessing techniques** is applied to the tweets to reduce the noise in the text, thereby yielding better-quality data for sentiment analysis.

Firstly, **tokenization** and **normalization** are applied by removing from the tweets: emojis, special characters, punctuation, URLs, superfluous whitespace, and user mentions (for example, *@account*).

Furthermore, we decided **to expand acronyms** which are abundant in the finance world (for example, *FED* becomes *Federal Reserve*), as well as **contractions** (for example, *we're* becomes *we are*).

An important technique in tweet cleaning is the **management of hashtags**, represented by the # symbol, allowing for the indexing of keywords or topics on Twitter. This function was created on Twitter and allows users to easily follow topics of interest. We have chosen the approach of removing hashtag sequences and keeping hashtags containing

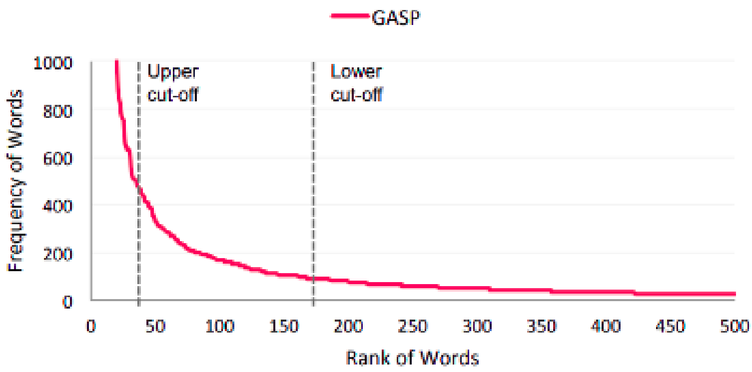


Figure 1.2: Distribution of the frequency of ranks of the first 500 terms of the GASP dataset.

an English word included in the English dictionary of the spell-check library *PyEnchant*. This retains some linguistic value of the tweet.

For illustration, consider the following Tweet as an example:

I #love bitcoin #BTC #BUY #ETH #trade becomes *I love bitcoin.*

Also, we removed **sequences of the same word** (for example, *buy buy buy* becomes *buy*) and **sequences of the same character** in a word (for example, *buuuuuy* becomes *buuy*).

Among the final steps of tweet cleaning, we removed words occurring only once in the entire dataset (**TF1**) and stopwords, thanks to **Zipf's law** Fig. 1.2, which states that given a corpus of natural language, the frequency of a word is inversely proportional to its rank in the frequency table, i.e., the most frequent word will appear approximately twice as often as the second most frequent word, three times as often as the third most frequent word, etc

Finally, a step of **lemmatization** is performed aimed at removing inflectional endings only and to return the base or dictionary form of a word. Moreover, tweets containing **fewer than four tokens** are omitted from the dataset as they are not suitable for sentence-level sentiment analysis.

An example of the application of the techniques above is presented:

RT @bitcoin https://twitter.com/FT/status.com Bitcoin ETF rejected but buuuy #NOW !!! Ask yourself why you aren't buying #Bitcoin lol, tomorrow it'll reach 80000 #BTC $BTC $ETH

↓

rt bitcoin etf rejected but buuy #now !! ask yourself why you aren't buying #bitcoin lol, tomorrow it'll reach 800 #btc

↓

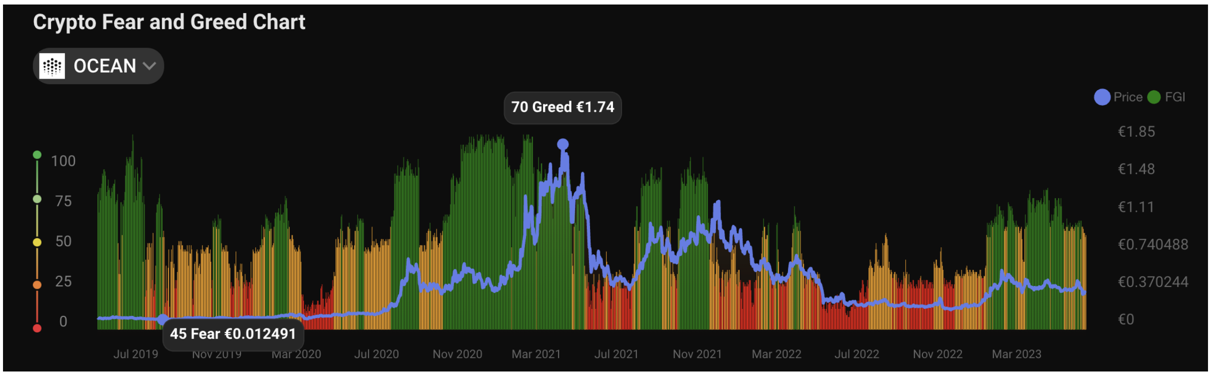
rt bitcoin exchange traded fund rejected but buuy now ask yourself why you are not t buying bitcoin lol tomorrow it ll reach

↓

bitcoin exchange traded fund rejected buuy ask not buying bitcoin lol tomorrow reach

Data Sources

In order to calculate the sentiment score 1 data source was used: Crypto Fear and Greed from Coin Stats.



1.2 Impact score: Target creation

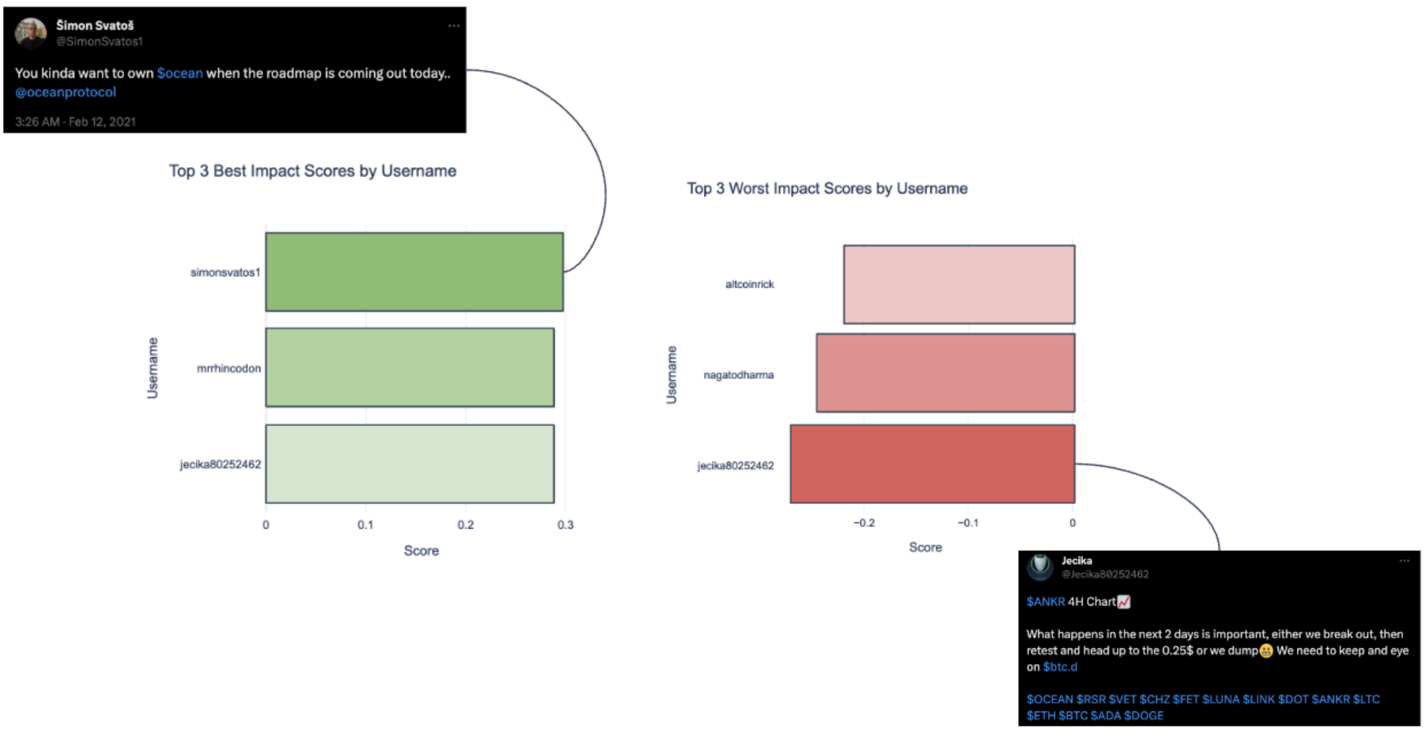
The cornerstone of our analysis is the creation of an 'Impact Sentiment Score', which is envisioned as the target for our machine learning model, to be elucidated in subsequent sections. This score is a forward-looking measure, inherently using future data such as likes, retweets, replies, and daily totals, which only become available after a tweet has been published.

The 'Impact Sentiment Score' is an amalgamation of two primary components. The first is the 'Impact Section', a combination of the 'User Influence Score' and the 'Volume Score'. These metrics were included as they encapsulate how a tweet's engagement (Volume) and the status of the user who tweeted it (User Influence) can potentially sway the price of the OCEAN token.

The second component is the 'Sentiment Section', comprised solely of the 'Sentiment Score'. This sentiment component helps assess the emotional tone of the tweet, which can significantly influence market sentiment and, by extension, the token's price.

To tether our score to the token's actual price, we incorporated the change in price four hours after the publication of the tweet into our 'Impact Sentiment Score'. For example, if a tweet is published at 12pm, we gauge the change in the token's price between 12pm and 4pm. The selection of a four-hour window was guided by the Granger Causality test, which yielded a p-value of 0.0037, implying a statistically significant causality (Refer to the dedicated section for a detailed explanation of p-value interpretation).

The 'Impact Sentiment Score', thus, serves as a predictive indicator of OCEAN token's future performance, accounting for the synergistic effects of tweet engagement, user influence, sentiment, and future price change. A higher score suggests a stronger positive market performance. The score ranges between -1 and 1.



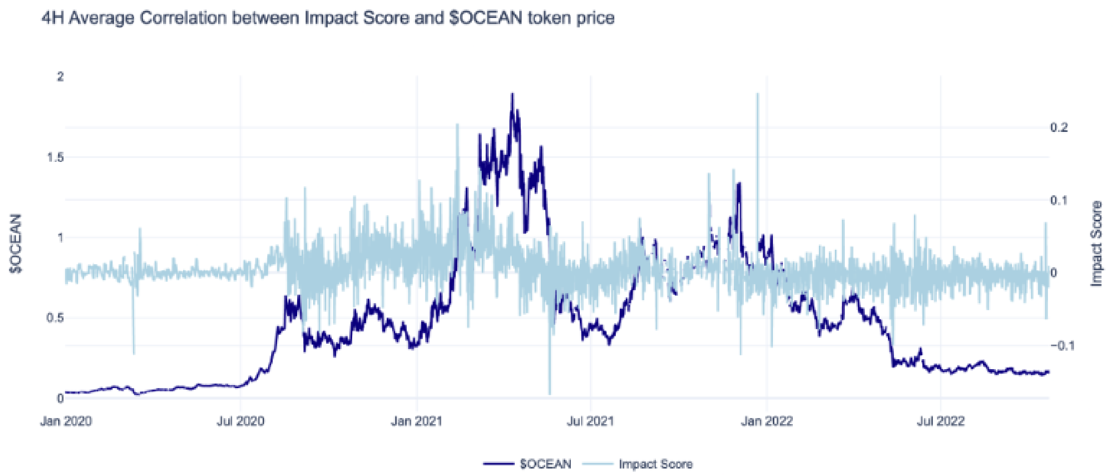
The formula for the 'Impact Sentiment Score' is as follows:

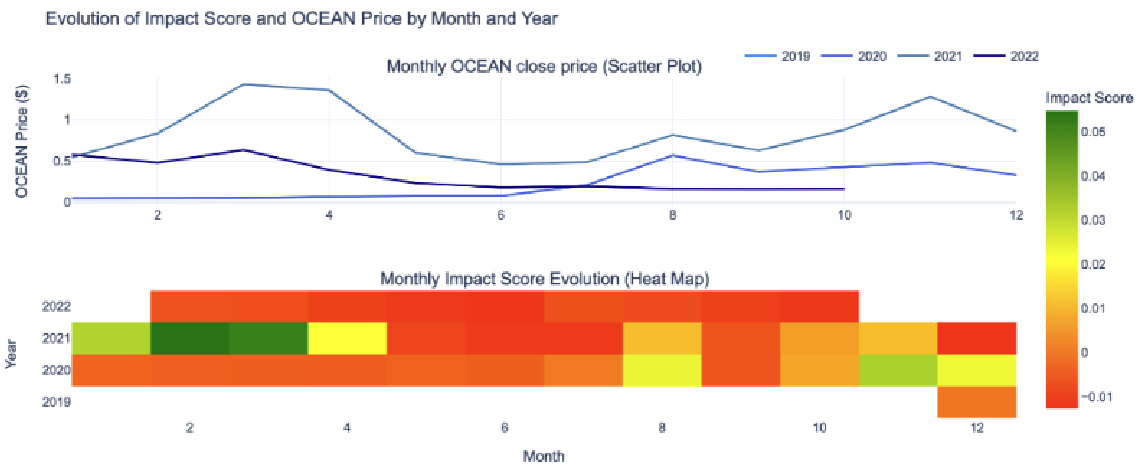
$$IS = ((VS \times W_{VS}) + (UIS \times W_{UIS})) \times SS \times W_{IS} + \Delta P_{4h} \times W_{\Delta P_{4h}}$$

where:

IS : Impact Sentiment Score
VS : Volume Score
W_{VS} : Weight of the Volume Score
UIS : User Influence Score
W_{UIS} : Weight of the User Influence Score
SS : Sentiment Score
W_{IS} : Weight of the Impact Score
ΔP_{4h} : Change in price 4 hours after the tweet
$W_{\Delta P_{4h}}$: Weight of the 4-hour future change

It's important to note that this model hinges on our assumptions and interpretations of the market dynamics. Future enhancements could introduce more factors or different weighting mechanisms to better capture the nuances of the cryptocurrency market.





1.2.1 Weights determination

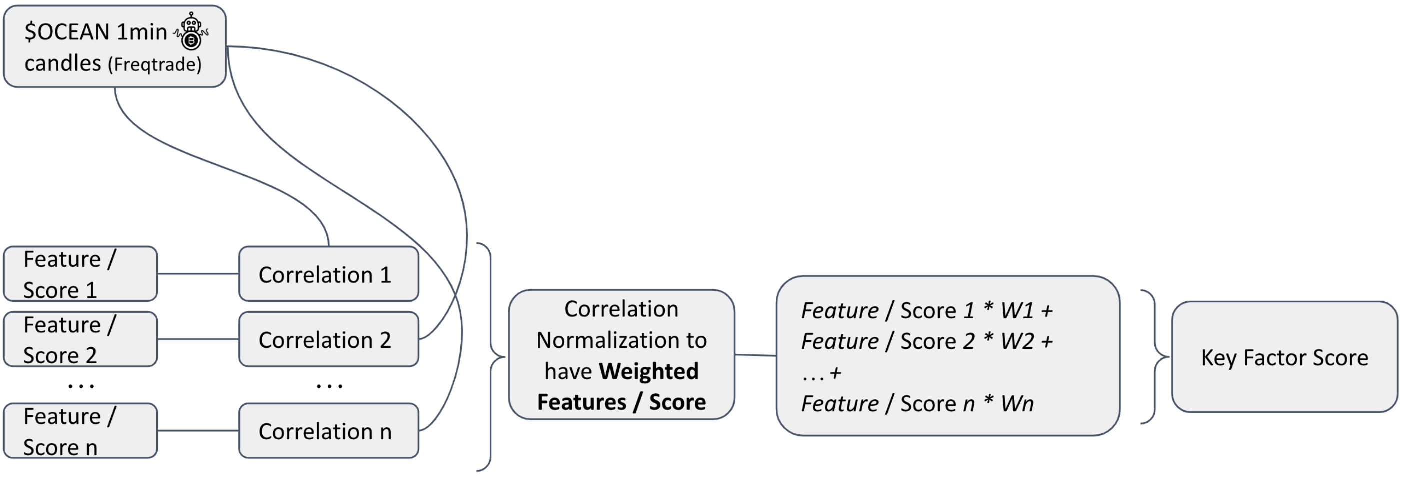


Figure 1.3: Determining feature weights using OCEAN price correlation

The strength of this weighted average strategy goes beyond just the determination of the final Impact Sentiment Score. It is a fundamental part of calculating the individual key factor scores as well, particularly in handling complex features that themselves are composed of multiple components.

For instance, the Sentiment Score is calculated by taking a weighted average of several sentiment indicators: TextBlob, Vader, and the Fear & Greed Index. Each of these sources provides a unique perspective on market sentiment, and their individual contributions are weighted based on their correlation with the price of the OCEAN token. By integrating these distinct elements in this manner, we not only capture a more comprehensive understanding of market sentiment but also ensure the relevance of each component.

This approach allows us to leverage the advantages of different sentiment analysis techniques and indices while mitigating their individual limitations. The resulting score thus becomes a balanced and robust measure of market sentiment that is attuned to the fluctuations in the price of the OCEAN token. This is one of the many examples where the flexibility of our strategy has been instrumental in developing a comprehensive and responsive analytical framework.

1.2.2 Granger Causality Test

The Granger Causality Test is a statistical hypothesis test for determining whether one time series is useful in forecasting another. In the context of your cryptocurrency analysis, it's used to check if the past values of the impact score 'Granger cause' the price of OCEAN.

In my case, the p-values indicate the probability that the null hypothesis (that the impact score does not Granger cause the price) is true. A common threshold for significance is 0.05, below which the null hypothesis can be rejected.

At lag 4 (4 hours), the p-values is less than 0.05 (0.043), which indicates that you can reject the null hypothesis. This suggests that the impact score Granger-causes the price of the cryptocurrency, meaning that the past values of the impact score can help predict the future values of the price.

Chapter 2

ML model creation

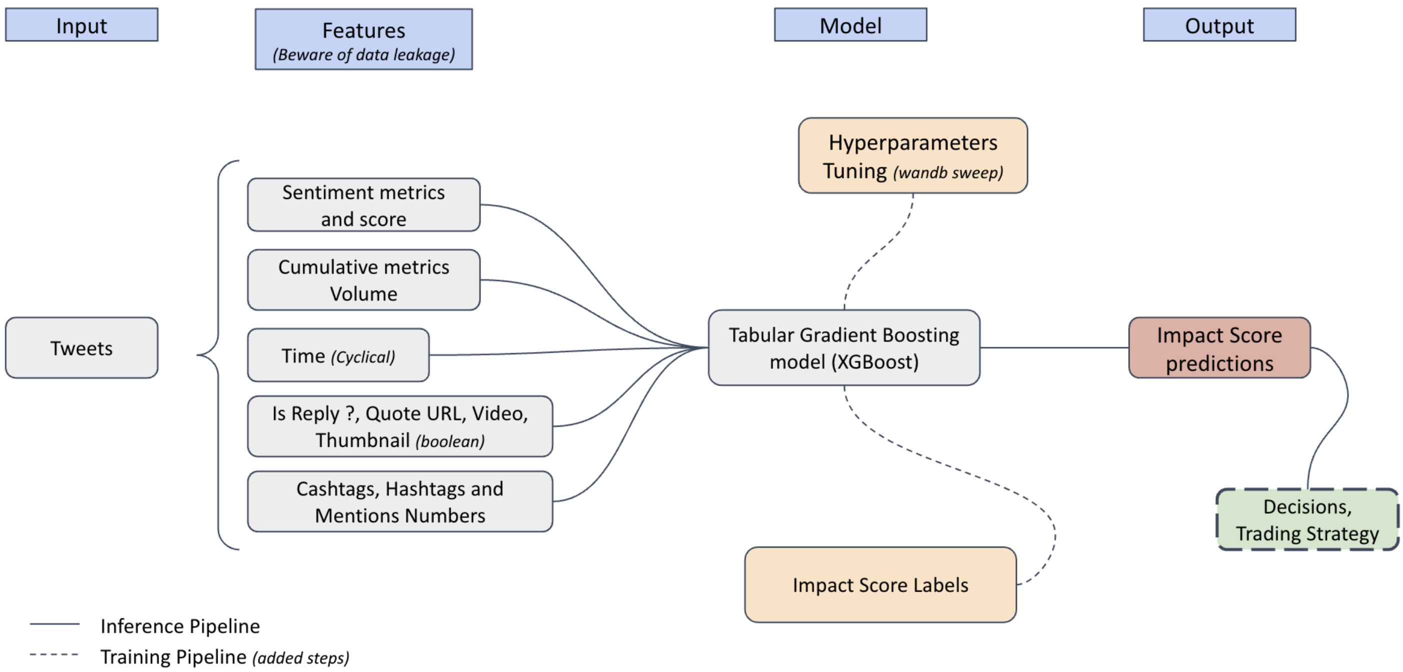


Figure 2.1: ML Model Pipeline

2.1 Methodology & Reflection

The challenge's open-ended question - *Develop an ML model that can provide information on the relationship between social media engagement and the price of the OCEAN token* - required a thoughtful interpretation. As both a Data Scientist and investor, I decided to frame this question from my perspective, considering the insights and predictions I would value from an ML model. This led to the creation of a regression model with the Impact Sentiment Score as the target.

As elaborated in previous sections, the Impact Sentiment Score is a potent amalgam of multiple elements: user influence, sentiment, tweet engagement, and future price variation. It serves as a predictive indicator of the OCEAN token's performance, with a higher score implying a favorable market response. The score lies within the range of -1 and 1, providing a standardized measure of potential future performance.

During the model's design, careful attention was paid to feature selection and engineering. We ensured that only features available at the time a tweet is published were included, thus preventing data leakage and ensuring our model's predictive integrity. Further details regarding this crucial aspect of model design will be discussed in the upcoming Feature Engineering section 2.2.1.

2.2 Impact score Tabular model

The creation of a tabular regression model emerged as the most fitting choice due to its ability to seamlessly assimilate a diverse set of features, thus paving the way for a more comprehensive analysis. These features encompass tweet metrics such as the number of hashtags, sentiment of the tweet, and the timestamp.

The decision to adopt a regression approach was guided by the quest for precision. In contrast to a classification model that might pigeonhole results into fixed categories, a regression model offers a continuous spectrum of outcomes. This granularity of output proves indispensable when we're interested in discerning the nuanced impacts of an individual tweet or a conglomerate of tweets over a particular duration.

2.2.1 Feature Engineering

In the process of model development, the creation of useful features is a pivotal step as it significantly contributes to the model's predictive capabilities. One such feature is the encoding of time-related information such as hour, month, day, and day of the week. Since these time variables are cyclical (e.g., after 23:59, the hour returns to 00:00), we used a sin-cosine transformation to preserve this cyclical nature and help the model understand these patterns.

Furthermore, to prevent potential data leakage, we ensured to only utilize information available at the moment a tweet is published. Features encapsulating the cumulative volume of tweets, retweets, replies, likes, and unique users since the start of the day were integrated into the model.

Also, has been added the previous 4h price change of the OCEAN token.

Additionally, every tweet entered into the pipeline undergoes sentiment analysis using Vader and TextBlob libraries. This step contributes additional features to the model, providing deeper insights into the sentiment conveyed in each tweet, thereby augmenting the predictive power of the model. This comprehensive feature engineering process was crucial in enhancing our model's capacity to predict the future performance of the OCEAN token accurately.

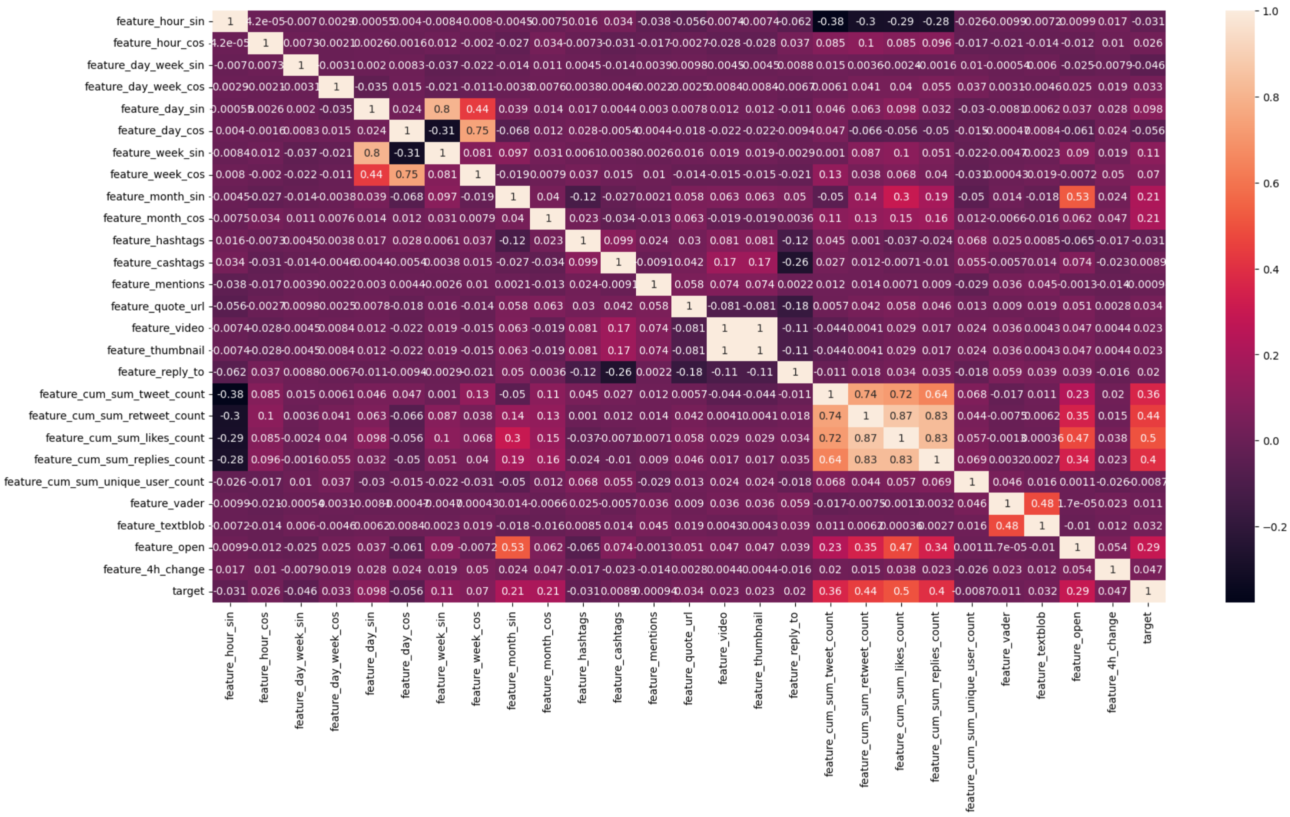


Figure 2.2: ML model feature correlation matrix

2.2.2 Hyperparameters Tuning & Cross Validation

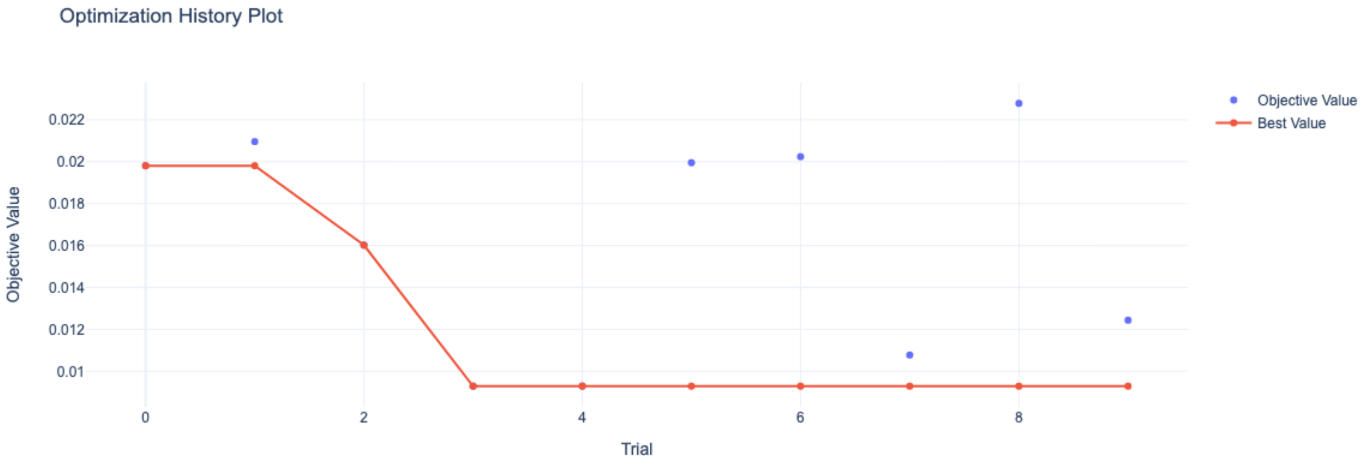
In the process of optimizing the XGBoost regression model, we deployed a dual strategy of **hyperparameters tuning** and **cross-validation** to ensure both a highly effective and generalizable model. This dual approach utilized the **Optuna** library's capabilities, a powerful tool grounded in a Bayesian approach for hyperparameter optimization.

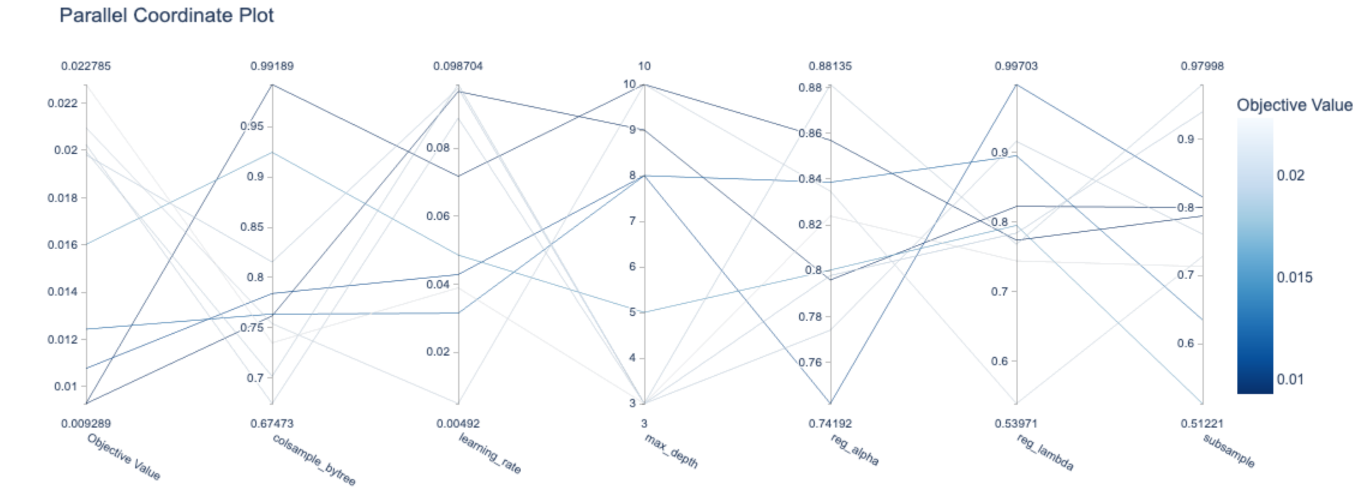
The hyperparameters tuning component of the strategy was designed to tune the model's hyperparameters to refine the model's predictive capabilities and performance (see attached explanation for a more detailed understanding of this method). Hyperparameters tuning is a pivotal step in model optimization because it allows the learning algorithm to adjust itself more accurately to the data, potentially leading to improved results.

Nevertheless, while pursuing hyperparameters tuning, we remained aware of the risk of overfitting, a condition where the model, in attempting to minimize the loss function, becomes overly complex and begins to learn the noise in the data rather than the underlying pattern. Overfitting can result in a misleadingly high accuracy on the training data but poor generalization to unseen data.

To address this risk, we combined hyperparameters tuning with 5-fold cross-validation during the hyperparameter search process. Cross-validation helped us to avoid overfitting by training and testing the model on different subsets of the data, ensuring that the model was not overly fitted to a particular set of data and thereby enhancing its generalizability.

The sweep plot included at the end of the report visually depicts this complex process and its impact on the model's performance, providing a comprehensive overview of the hyperparameters tuning and cross-validation process. Our integrated approach of simultaneous hyperparameter tuning and cross-validation thus provided a careful balance, optimizing the model's predictive capabilities while preserving its ability to generalize to unseen data.



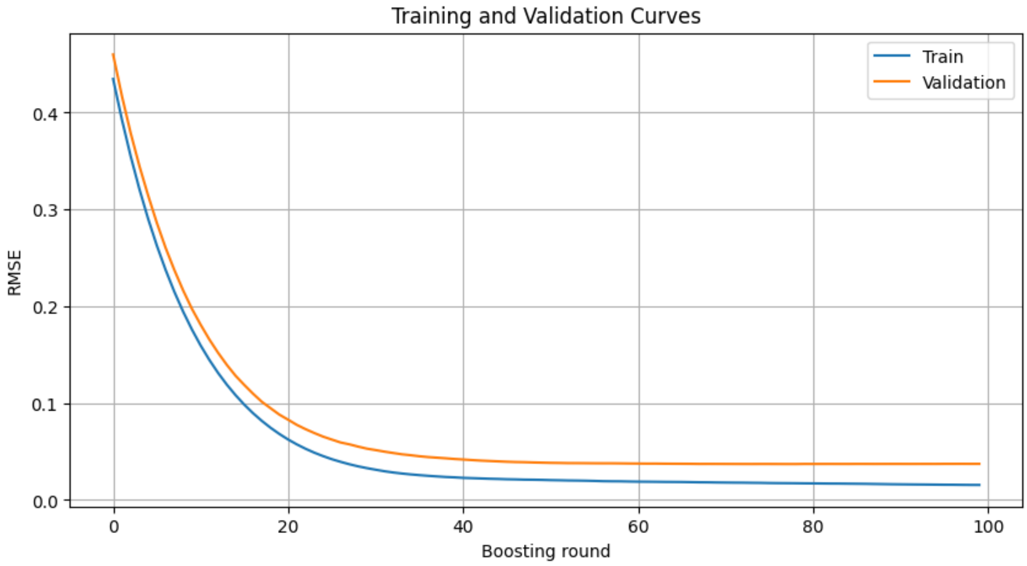


2.2.3 Model Training

The decision to utilize the XGBoost model was primarily based on its adaptability and robustness when dealing with diverse and complex data, such as our target impact score, which combines social media impact and future price performance. XGBoost leverages gradient boosting and decision trees, enabling it to effectively capture complex relationships between a broad range of input features. This is crucial for our task, where inputs encompass a mix of social media metrics and market indicators.

In addition, XGBoost is recognized for its strong performance across numerous machine learning tasks. It is particularly proficient at handling common challenges in social media and market data, including noise, missing values, and outliers. The ability to handle these complexities is pivotal for delivering reliable predictions in our context.

Model training is an integral part of the machine learning process. This stage involves learning the underlying patterns in the data and adjusting the model's parameters accordingly. For our XGBoost model, this entailed a series of systematic adjustments to find the optimal balance that would result in the most accurate predictions of the target impact score. The accompanying plot shows an illustration of the training process:



2.2.4 Evaluation & Interpretation

After undertaking diligent training and fine-tuning of our model, as described earlier, we proceeded to evaluate its performance. Our model yielded a Root Mean Square Error (RMSE) of 0.037 on the test set. This metric is particularly noteworthy when considering the range of the target variable, which spans from -1 to 1. The RMSE can be interpreted as a measure of the standard deviation of the prediction errors. An RMSE of 0.037 means that on average, our predictions deviate from the actual values by approximately 3.7% of the full scale. Considering the scale of our target, this indicates a strong predictive performance, showcasing the model's proficiency at accurately predicting the complex, multi-dimensional impact score.

Interestingly, when we aggregate the predictions over one-hour intervals, the RMSE decreases further to 0.03. This implies that our model's predictions are even more accurate when considered in an aggregated timeframe, capturing the overall trend of the impact score in the next four hours. This suggests that our model is quite adept at recognizing the cumulative influence of numerous tweets within a given hour, rather than solely the isolated impact of a single tweet. Therefore, it provides a more comprehensive understanding of the holistic influence of social media engagement on the future price performance of the OCEAN token.

2.3 Challenges & Other tested approaches

A significant challenge I faced during this project involved the reinterpretation of results from the first part of the challenge. My aim was to offer a tangible, useful metric to the end user - a metric that could effectively illuminate the relationship between social media engagement and the OCEAN token's price, ultimately aiding decision-making processes. The development of the Impact Sentiment Score presented its own set of challenges, as it was essential to make it both comprehensive and interpretable for users. This wasn't a linear process; rather, it involved an iterative approach, requiring continuous fine-tuning and adjustments to ensure the score's validity and user-friendliness.

A substantial challenge faced during this project was related to data acquisition. The necessity of accessing various data sources posed a considerable hurdle, especially when each API required a distinct set of permissions, including Twitter account credentials and associated bank details. This repetitive process was not only time-consuming but also presented security concerns, as these steps needed to be completed on multiple platforms. The availability of a data marketplace, where data could be retrieved anonymously, would have streamlined this process significantly. A platform like the OCEAN Marketplace, which rates available datasets based on quality and offers customizable filter options, could have provided a more efficient solution. Having a centralized resource for data acquisition in a dentralized way would not only have saved time but could have potentially enriched the model with a broader range of high-quality, reliable data.

Throughout this project, I experimented with various methodologies, some of which proved less successful than others. One such approach involved the utilization of a Hugging Face fine-tuned sentiment classification model in a regression task, with the Impact Score serving as the target. This model was combined with a Gradient Boosting model, which used the outputs from the Hugging Face model trained on the tweets as its inputs. Despite the theoretical appeal of this method, it fell short in practice. The initial training sessions yielded disappointing results, with a mean squared error (MSE) of 0.3 on the validation set for a metric range of -1 to 1. Given these suboptimal results, I ultimately decided to abandon this approach. While the technique was conceptually engaging and appeared promising, it did not yield the desired outcomes in a practical setting.

2.4 Actionable Insights & Recommendations

2.4.1 Actionable Insights

The Impact Sentiment Score, serving as an integrative predictive indicator of the OCEAN token's future performance, furnishes actionable insights for diverse users, facilitating more informed and strategic decision-making processes.

- **Investment Decisions**: Primarily, investors in the OCEAN token can utilize the Impact Sentiment Score to guide their investment strategies. A high score, signifying a likely positive market response, could be interpreted as an opportune moment to purchase or retain the token. In contrast, a low score, suggestive of a less favorable market outcome, may indicate an advisable period to sell or delay buying the token.
- **Market Trend Forecasting**: In addition to immediate investment choices, the Impact Sentiment Score can be employed for anticipating longer-term market trends. By monitoring the Score's progression over time, investors and market analysts can discern patterns linking social media sentiment with market fluctuations. This predictive capacity can signal prospective changes in the OCEAN token's price, providing an advantage for pre-emptive strategy adjustment.
- **Risk Evaluation**: The Impact Sentiment Score can be integral to risk assessment processes as well. Particularly negative or volatile scores might signify an elevated risk level associated with investing in the OCEAN token at a given moment.
- **Strategic Planning**: On an institutional level, cryptocurrency organizations or financial entities dealing with the OCEAN token can leverage the Impact Sentiment Score for strategic planning. For instance, if the Score consistently projects positive market responses, an enterprise may decide to augment its emphasis on OCEAN token-related services or products.

In essence, the Impact Sentiment Score, by synthesizing key elements like user influence, sentiment, tweet engagement, and future price variation, provides a robust tool for multiple stakeholders to harness in their respective decision-making processes, optimizing their strategies based on anticipatory market insights.

2.4.2 Ocean Market: Value the model output

In order to expand the utility and marketability of our predictive model, we propose its integration with the Ocean Market platform. This would enable monetization of our model's outputs, providing a valuable resource for both users and developers. To elaborate, users would have the capability to apply the model to existing market datasets or their personal data, yielding predictions formulated by our advanced model.

To ensure transparency and foster continuous improvement, we propose the addition of a user feedback and rating system. This mechanism would allow users to evaluate the model's performance, providing crucial insights into the confidence and quality of the predictions generated. Through this interactive approach, we can dynamically improve our model, offering a consistently high-value tool for the Ocean Market community.

Conclusion

This technical report has embarked on an ambitious exploration of the nexus between social media engagement, specifically via Twitter, and the OCEAN token's price volatility. We pioneered the concept of the Impact Score, an innovative metric that synthesizes tweet volume, sentiment, user influence, and future price variation to offer a robust predictive indicator of the OCEAN token's market performance.

Beyond traditional classification models, we implemented an advanced regression approach, the Gradient Boosting model. Enriched with hyperparameters tuning and cross-validation techniques, this model yielded a nuanced understanding of the intricate relationship we sought to elucidate.

Despite the challenges encountered, such as data accessibility and the trial of various unsuccessful modeling approaches, we persevered. The experiences served to refine our methodologies and enhance our model's performance, leading to a more accurate and reliable Impact Score.

Ultimately, the key output of this journey is the Impact Score, a tool yielding potent actionable insights for multiple stakeholders, from individual investors to large financial institutions. It can serves to inform investment decisions, anticipate market trends, evaluate risk, and guide strategic planning, thereby exemplifying the profound potential of harnessing social media data in the realm of cryptocurrency market dynamics.