Text

Description automatically generated

**ALY 6980 – CAPSTONE PROJECT**

**PROF. VALERIE ATHERLEY**

**MODULE 12: CAPSTONE PROJECT**

**Date: 08/22/2023**

**SUBMITTED BY: GROUP 3**

**Vishaka Mohan Mohana**

**Sakthi Pradeepa Muthukumar**

**Shraddha Reddy Panyala**

**Richa Umesh Rambhia**

**Mohan Samala**

**INTRODUCTION**

The landscape of finance has been revolutionized by the cryptocurrency industry, with Bitcoin at its forefront, introducing a realm of secure and decentralized digital currency. As the popularity of Bitcoin continues to surge, so does the challenge of grasping its volatile market dynamics. Pioneering the cryptocurrency mining sector, Marathon Digital Holdings comprehends the vital significance of Bitcoin's hashrate volatility, a pivotal gauge of the network's computational prowess.

Through a data-centric initiative, the aim is to delve profoundly into the intricacies of the Bitcoin market, delving into historical data to identify the factors influencing fluctuations in hashrate. The goal is to adeptly predict future hashrate changes with remarkable precision and dependability by crafting a robust predictive model. Such a model holds the potential to greatly enhance Marathon Digital Holding's mining operations, enabling strategic adjustments, efficient resource allocation, and well-informed decision-making to respond to evolving hashrate patterns.

Valuable insights will be extracted from the provided data, encompassing variables like block height, version, timestamp, pool difficulty, transaction count, reward fees, and more, utilizing comprehensive exploratory data analysis (EDA) methodologies. This EDA approach will furnish a comprehensive understanding of the interconnections among these factors and their impacts on variations in hashrate.

The ultimate objective of this project is to contribute to the enduring advancement and stability of the Bitcoin network, concurrently empowering Marathon Digital Holdings to navigate the dynamic market landscape with data-driven insights and informed judgments.

**PROBLEM STATEMENT**

Marathon Digital Holdings, a key participant in the cryptocurrency mining sector, is faced with the difficulty of understanding and forecasting fluctuation in Bitcoin's intraday hashrate, which is a vital measure of the network's processing capacity. The project's goal is to create a strong prediction model that reliably anticipates hashrate changes using in-depth data analysis and relevant inferences drawn from past data.

The primary problem statement revolves around two key challenges:

1. **Understanding Bitcoin Hashrate Volatility**: The first difficulty is to obtain a thorough grasp of the causes behind Bitcoin hashrate variations. Extensive exploratory data analysis (EDA) is required to find patterns, correlations, and relationships between parameters like as block height, version, timestamp, pool difficulty, transaction count, reward fees, and others. The investigation should give insights into the influence of factors on hashrate changes and aid in the identification of major drivers of computational power shifts within the Bitcoin network.
2. **Developing an Accurate Predictive Model**: The second problem is to create a predictive model that predicts future hashrate fluctuations with high accuracy and reliability. The model must adapt to changing market conditions and optimize its predicting performance over time by leveraging the power of reinforcement learning techniques. To produce meaningful and actionable forecasts for Marathon Digital Holdings, the model should take into consideration past hashrate data, market trends, and other important variables.

**OBJECTIVES**

The primary objective of this study is to develop a dependable and precise prediction model to forecast alterations in Bitcoin's hashrate. Marathon Digital Holdings' aim is to attain a comprehensive comprehension of hashrate fluctuations, signifying the collective computational capacity employed for mining and transaction processing on the Bitcoin network. Achieving this objective will empower Marathon Digital Holdings to adopt data-guided decisions, enhance mining operations, and elevate their overall effectiveness in the realm of Bitcoin trading.

**METHODOLOGY**

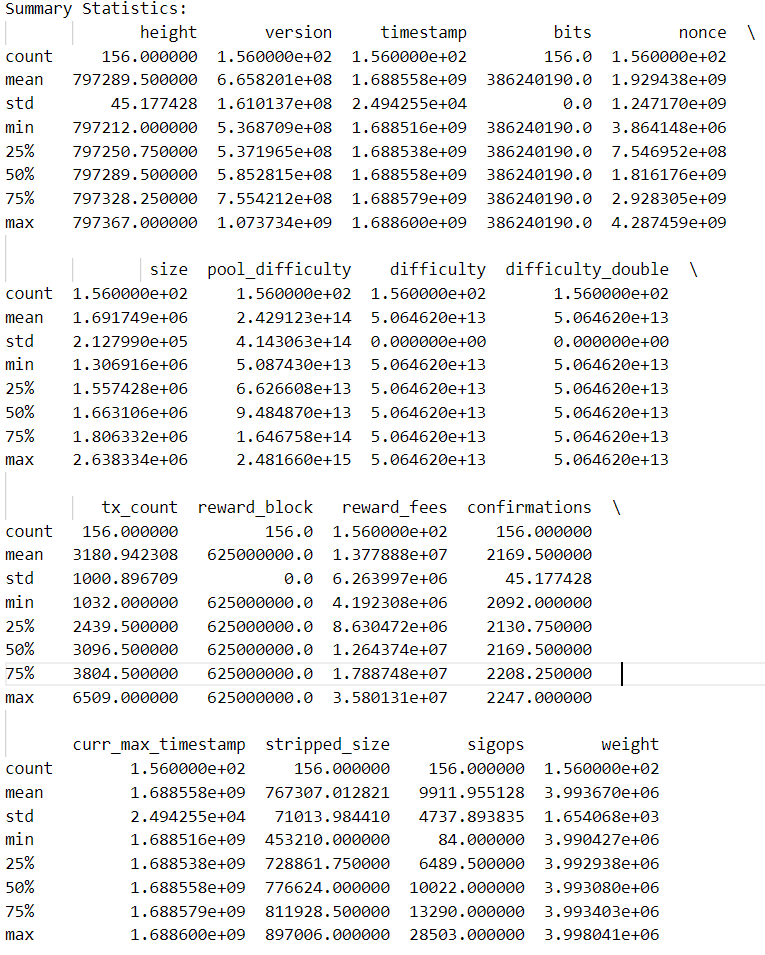
To realize the research objectives and address the stated issue, the following approach will be embraced:

1. Data Collection: Gather historical Bitcoin block header data, encompassing parameters like block height, version, timestamp, pool difficulty, transaction count, reward fees, and more. Secure data from reputable cryptocurrency sources and blockchain archives.
2. Data Refinement: Clean and preprocess collected data, addressing missing values, anomalies, and irregularities. Ensuring data integrity is essential for accurate analysis and modeling.
3. Exploratory Data Analysis (EDA): Conduct comprehensive EDA to uncover insights into Bitcoin hashrate dynamics. Examine interrelations and correlations among diverse parameters, visualize temporal trends, and pinpoint influential factors shaping hashrate fluctuations.
4. Model Exploration: Explore diverse machine learning algorithms encompassing regression and reinforcement learning to construct the predictive model. Appraise each model's performance, selecting the one offering utmost precision and dependability.
5. Model Training: Train the chosen prediction model utilizing historical hashrate data and pertinent features. Utilize historical data to fine-tune model parameters and optimize performance.
6. Model Validation: Validate the prediction model utilizing varied assessment metrics and cross-validation techniques to ensure robustness and generalizability.
7. Prediction and Projection: Apply the trained model to anticipate forthcoming hashrate variations. Regularly monitor and enhance the model to adapt to evolving market dynamics and enhance predictive prowess over time.

**EXPLORATORY DATA ANALYSIS**

1. **Summary Statistics**

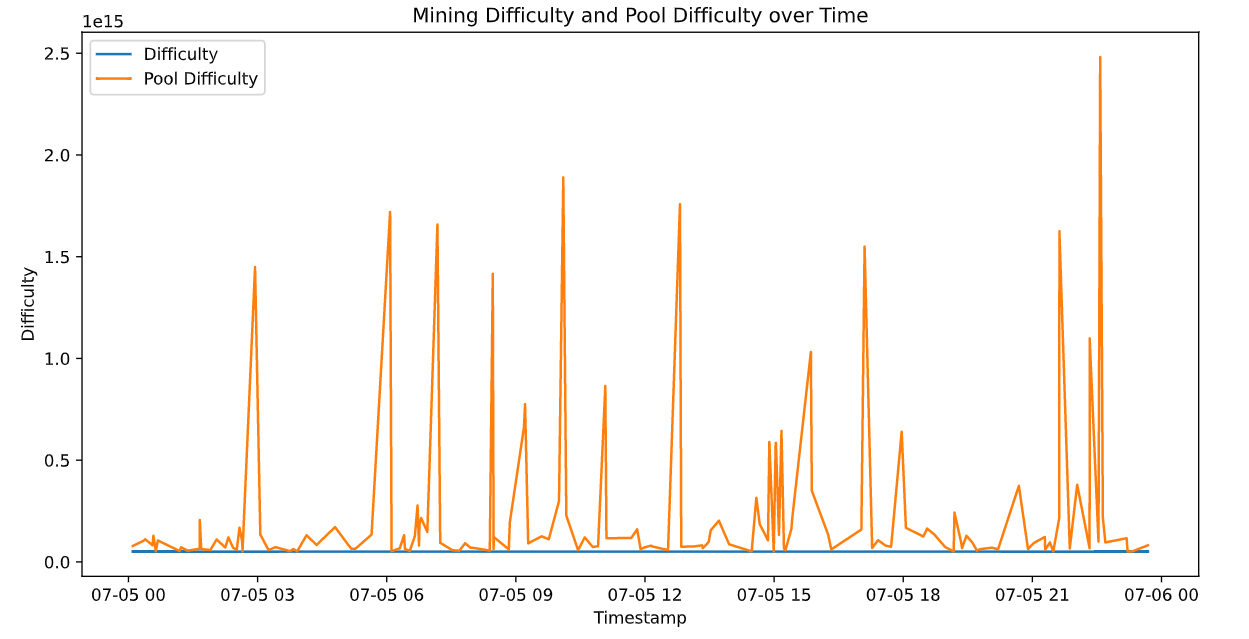
The figure below displays the summary statistics of the dataset using the describe() function. The function calculates descriptive statistics for all numerical columns in the Data Frame. The statistics include count, mean, standard deviation, minimum, 25th percentile (Q1), median (50th percentile or Q2), 75th percentile (Q3), and maximum for the numerical variables of the dataset as shown below.



**Figure 1. Summary Statistics**

1. **Time Series Analysis**

The below graph displays the time series analysis and creates a line plot to visualize the changes in Bitcoin mining difficulty and pool difficulty over time.

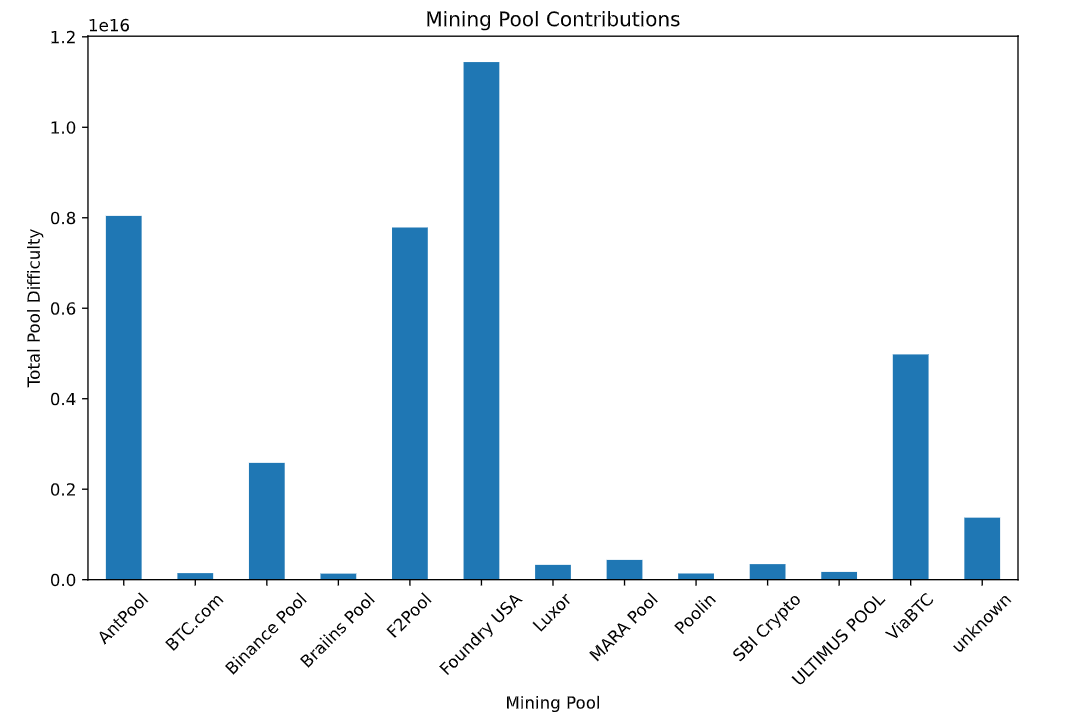


**Figure 2. Time Series Analysis**

The line plot shows two lines representing mining difficulty and pool difficulty over time. The x-axis represents the timestamps in chronological order, indicating the progression of time. By observing the plot, it is possible to identify patterns and anomalies in mining difficulty and pool difficulty over different time periods, which can be further investigated to gain insights into the behavior of miners and mining pools.

1. **Mining Pool Contributions**

The next graph performs an analysis of mining pool contributions and creates a bar chart to visualize the total pool difficulty contributed by each mining pool.



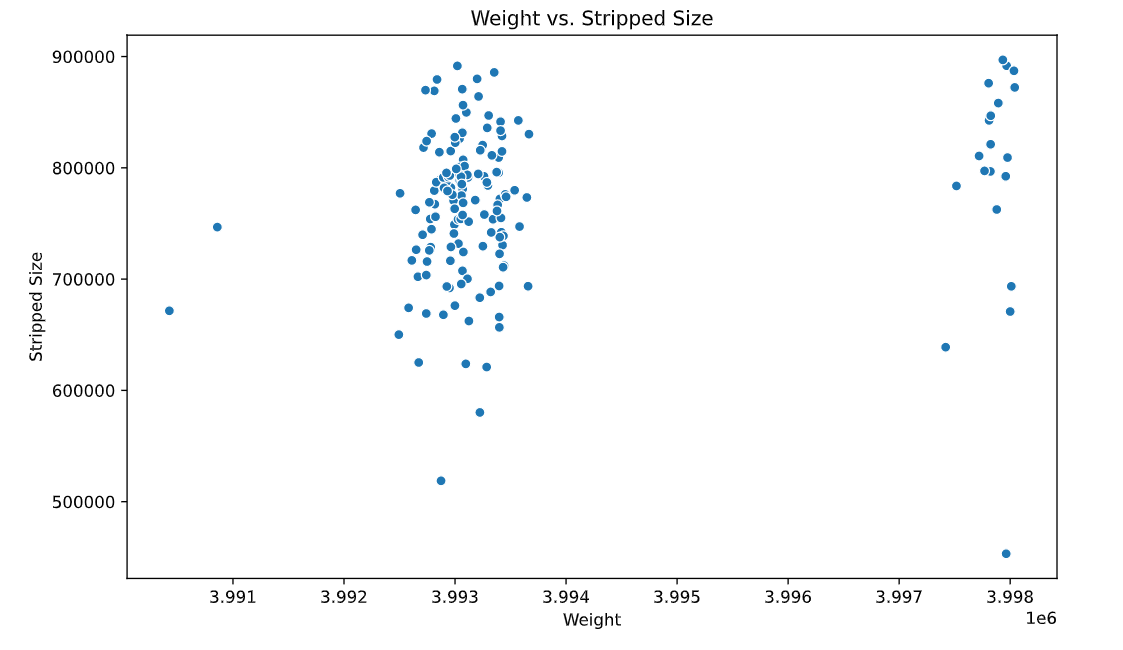
**Figure 3. Mining Pool Contributions**

The bar chart provides an overview of the total pool difficulty contributed by each mining pool in the dataset. Some mining pools contribute significantly more to the network's overall mining effort, as indicated by higher bar heights. These pools are likely large and powerful, with a substantial share of the total mining hash rate.

On the other hand, some pools have lower contributions, suggesting they might be smaller or less powerful pools with a relatively smaller share of the mining hash rate. The chart can help identify which mining pools play a more dominant role in the Bitcoin network's mining activities. By comparing the pool contributions over different time periods, it is possible to see if any pools' dominance has shifted over time or if new pools have emerged with substantial contributions.

1. **Weight and Stripped Size Analysis**

This graph creates a scatter plot to analyze the relationship between the 'weight' and 'stripped\_size' columns in the dataset.



**Figure 4. Weight & Stripped Size Analysis**

The scatter plot shows the relationship between the 'weight' and 'stripped\_size' columns for each data point in the dataset. Each point on the plot represents a specific Bitcoin block entry, and its position is determined by the corresponding values of 'weight' and 'stripped\_size'.

The plot does not show a clear linear or well-defined pattern, indicating that there might not be a strong linear correlation between 'weight' and 'stripped\_size'. It is essential to consider the context of 'weight' and 'stripped\_size' in the dataset to draw more meaningful insights. Further analysis and domain knowledge are required to interpret the specific implications of this relationship for the Bitcoin blockchain and mining network.

1. **Plotting Bitcoin Pool Difficulty over time using the Current Max Timestamp**

Plotting the Bitcoin pool difficulty over time using the curr\_max\_timestamp as the X-axis and pool\_difficulty as the Y-axis. This can help visualize the changes in mining difficulty and the contribution of different pools to the overall mining effort.

A graph of a graph

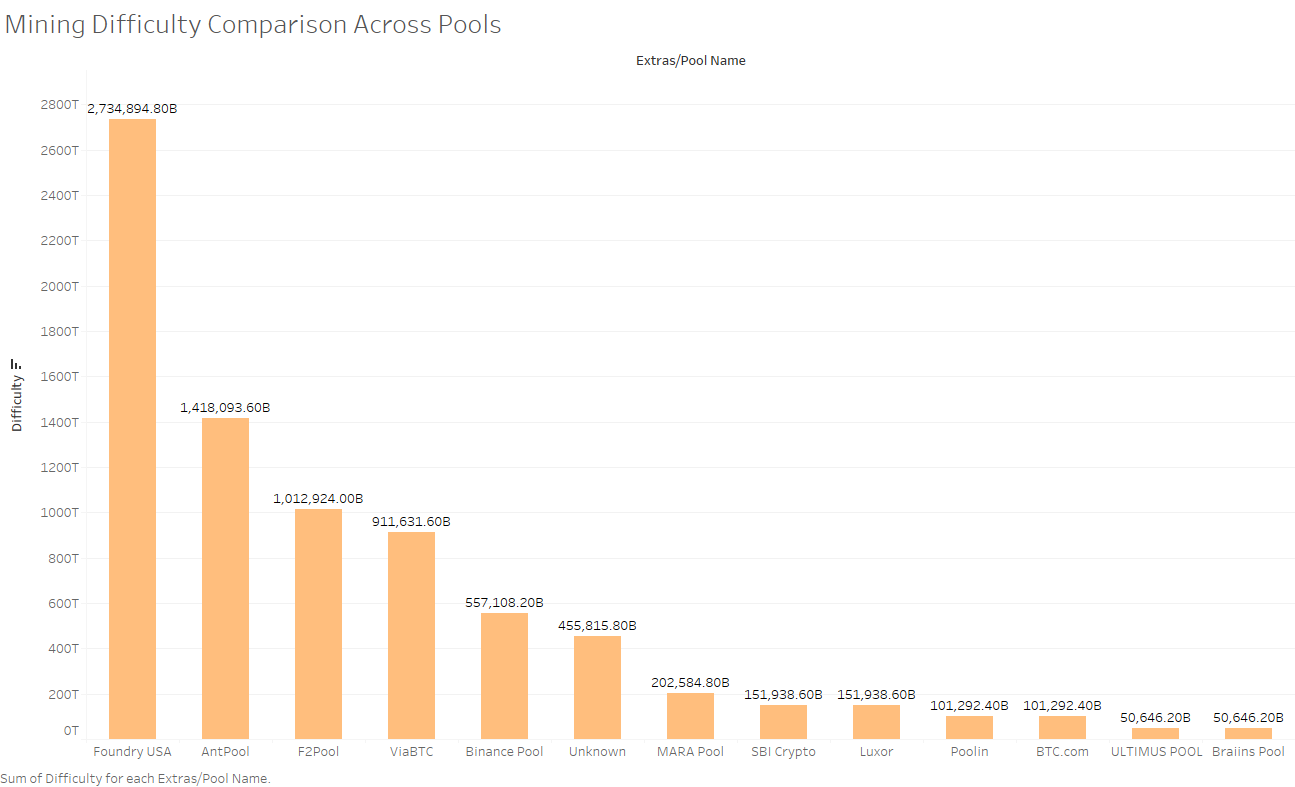
Description automatically generated

**Figure 5. Plotting Bitcoin Pool Difficulty over time**

The graph shows the trend of mining difficulty for different pools over time. As the X-axis represents time, and the Y-axis represents pool difficulty, we can observe how the mining difficulty has evolved. The line chart exhibits fluctuations in pool difficulty over time. This indicates that the mining difficulty of different pools is not constant and can vary based on factors like pool size, hash rate, and network conditions. By analyzing the line chart, we can identify periods when certain pools have higher difficulty values than others. Pools with higher difficulty have a more significant contribution to the overall mining effort during those periods. Overall, the visualization provides a comprehensive overview of the changes in mining difficulty fordifferentpools over time. It helps identify patterns, trends, and significant events in the mining ecosystem, which can be valuable for miners, researchers, and stakeholders in the cryptocurrency space.

1. **Mining Difficulty Comparison Across Pools**

**(**X-axis: extras/pool\_name (Dimension - Categorical); Y-axis: difficulty)



**Figure 6. Mining Difficulty Comparison Across Pools**

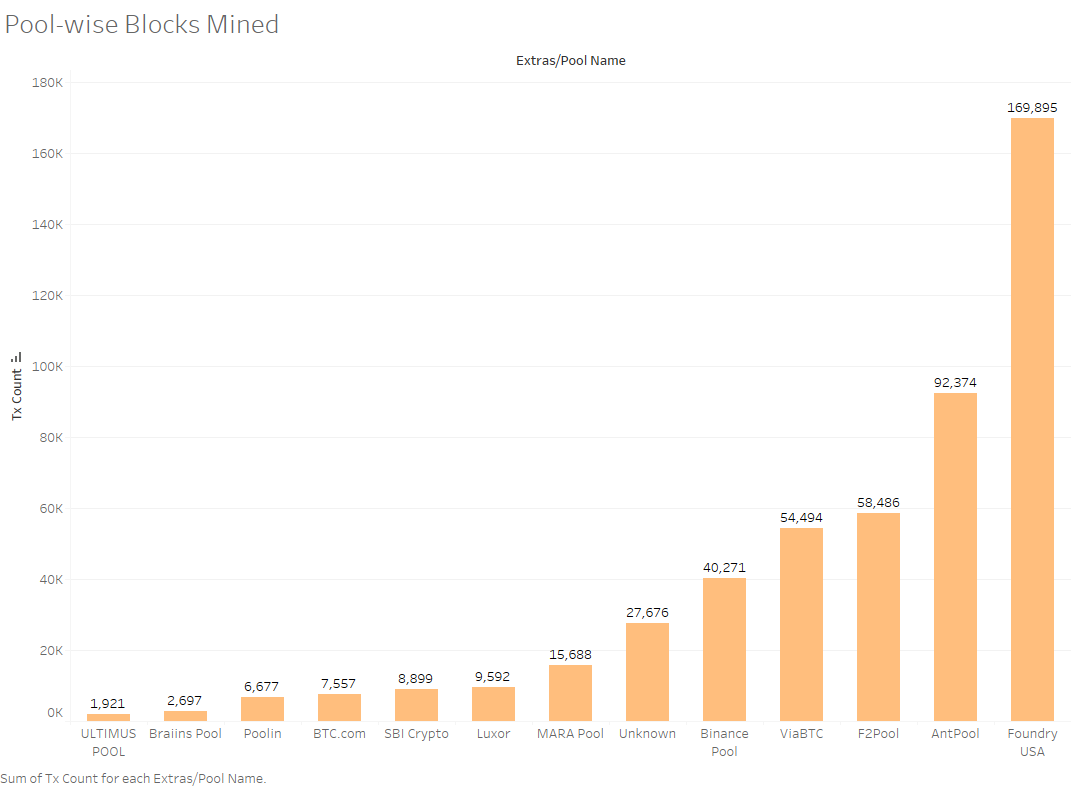
The graph compares the mining difficulty of different pools, with the X-axis representing the pool names (categorical dimension) and the Y-axis representing the difficulty.

The Y-axis shows the range of difficulty values across various pools. We can observe that the difficulty values are spread across different scales, indicating variations in mining power and hash rates among the pools. By looking at the bars representing each pool, we can identify which pools have the highest and lowest difficulties. Pools with higher difficulties are likely to be more dominant in terms of mining power and contribution to the network's security.

The graph also shows the distribution of mining difficulty across different pools. It provides insights into how mining power is distributed among various pools, highlighting potential centralization or decentralization of mining efforts.

1. **Pool-wise Blocks Mined**

(X-axis: extras/pool\_name (Dimension - Categorical); Y-axis: tx\_count (Measure – Count))



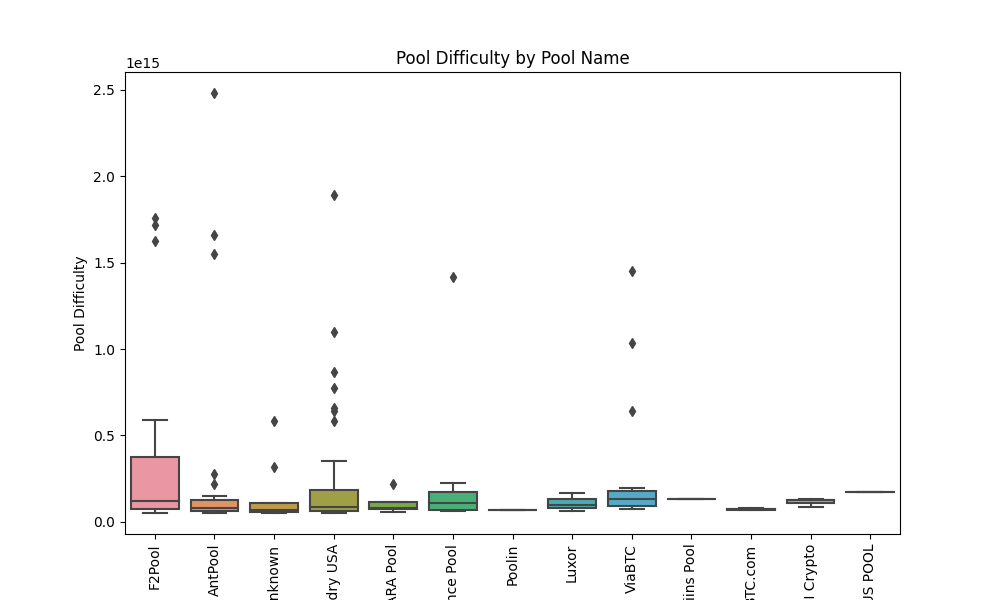
**Figure 7. Pool-wise Blocks Mined**

The graph compares the number of blocks mined by different pools, with the X-axis representing the pool names (categorical dimension) and the Y-axis representing the count of blocks mined (measure - count). The Y-axis shows the count of blocks mined by each pool, which directly reflects the mining performance and efficiency of each pool. Pools with higher block counts are more successful in mining and validating transactions. Some pools have a significantly higher number of blocks mined compared to others. These dominant pools have a larger share of the total mining power and contribute more to the blockchain's security. The graph provides insights into the distribution of block mining activities among different pools. It indicates how evenly or unevenly blocks are being mined across various pools.

Thus, the visualization allows for easy comparison between the block counts of different pools, making it straightforward to identify the most and least productive pools.

1. **Pool Difficulty by Pool Name**

The box plot visualizes the distribution of pool difficulty for different mining pools. Each box represents the distribution of pool difficulty for a specific pool, and the X-axis displays the pool names (categorical variable), while the Y-axis represents the pool difficulty (measure variable).

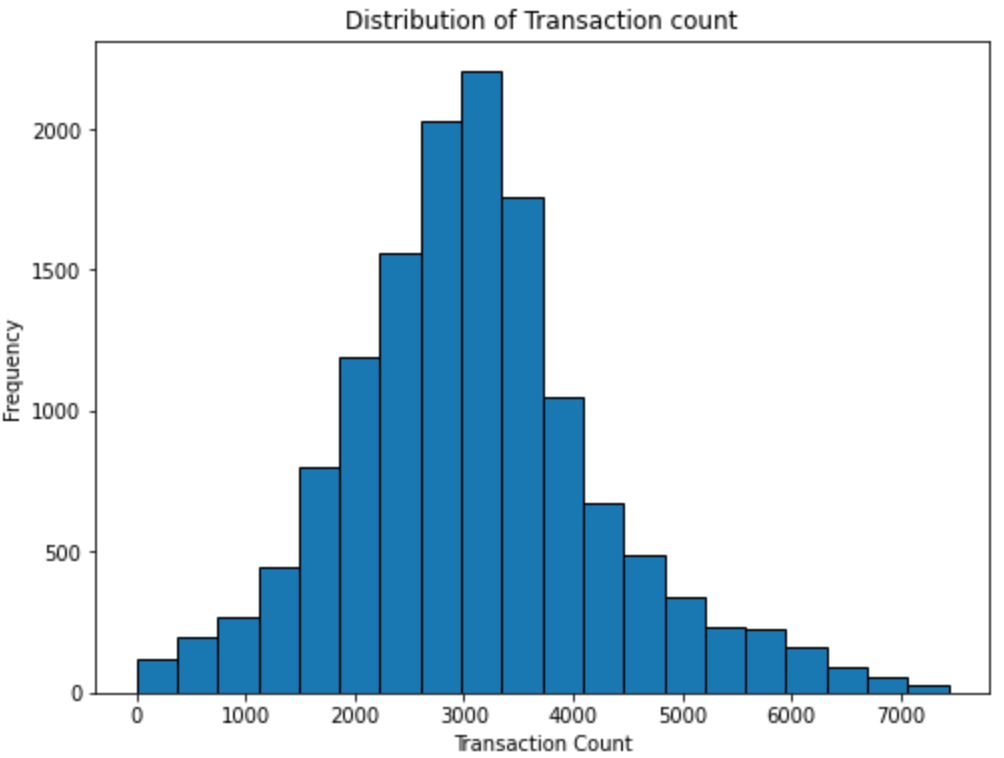


**Figure 8. Pool Difficulty by Pool Name**

The length of each box indicates the interquartile range (IQR) of pool difficulty for the respective pool. The middle line within the box represents the median difficulty value for that pool. The whiskers extending from the boxes show the range of pool difficulty, excluding any outliers. The width of the boxes gives an idea of how spread out the difficulty values are for each pool. A wider box indicates higher variance in pool difficulty, while a narrower box suggests less variation.

Pools with boxes closer to lower difficulty values may be less successful or less active in mining, indicating relatively lower mining power. Pools with boxes closer to lower difficulty values may be less successful or less active in mining, indicating relatively lower mining power. Overall, the box plot of pool difficulty by pool name provides a comprehensive view of the distribution of mining difficulty across different pools. It helps identify dominant pools, understand the spread of difficulty within each pool, and detect any unusual patterns or outliers in mining difficulty, enabling better analysis of the mining ecosystem.

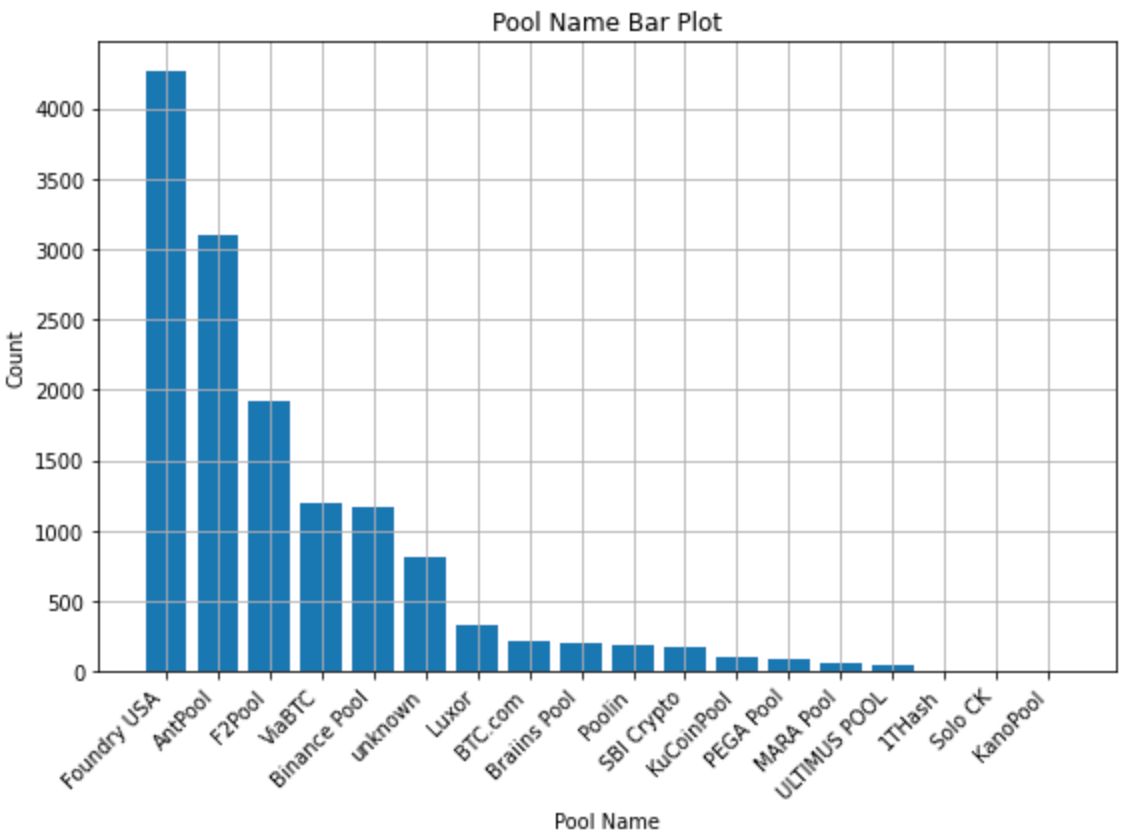
1. **Distribution of Transaction count**



**Figure 9: Distribution of Transaction Count**

**Observation:** A distribution plot of the dataset's transaction count is shown in the first figure. The distribution has a peak between 2500 and 3000 transaction counts and follows a normal distribution. As we move farther away from this peak, the frequency of transactions diminishes, indicating that the bulk of transactions occur inside this area. To analyze transaction patterns and spot any anomalous or outlier transactions that can impact the general functionality of the Bitcoin network, it is essential to comprehend the distribution of transaction counts.

1. **Pool Name Bar Plot**

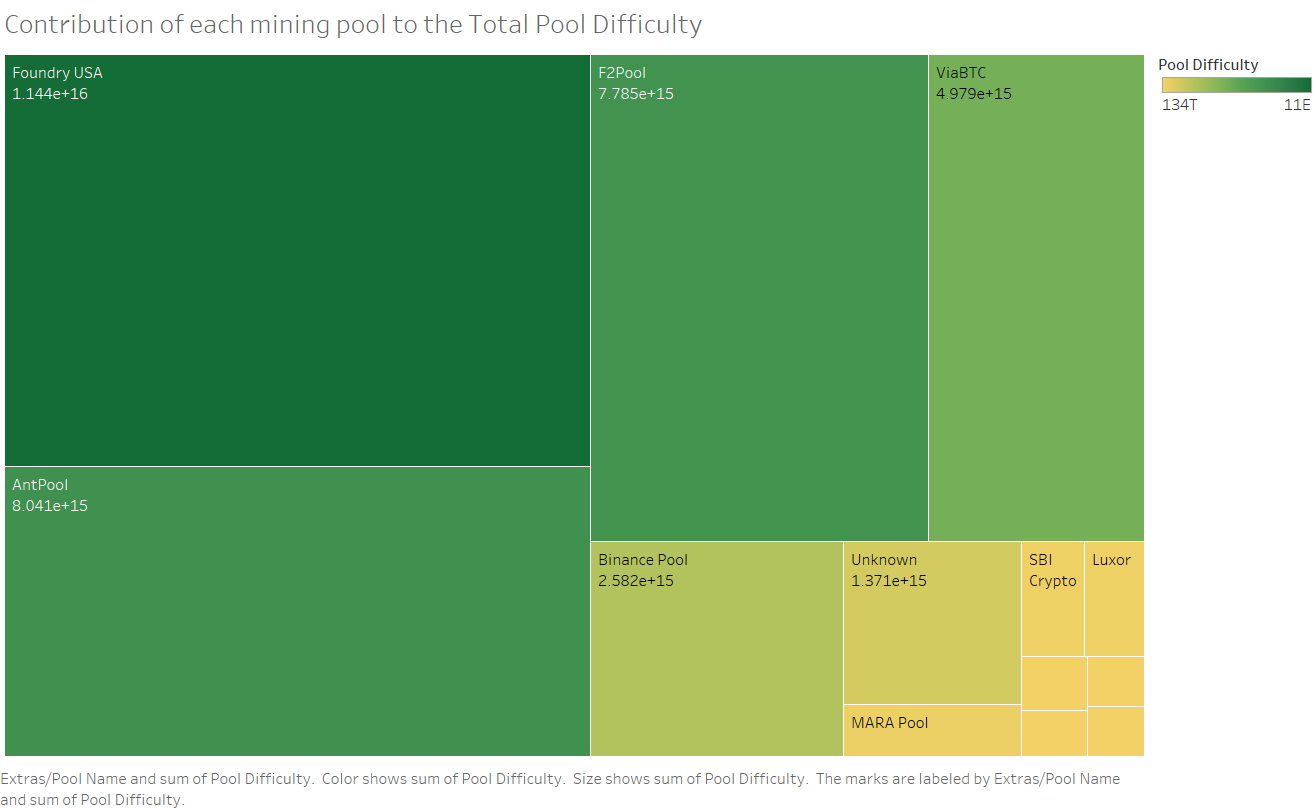


**Figure 10: Pool Name Bar Plot**

**Observation:** The frequency of each pool name in the dataset is shown in the second figure, which is a bar graph. We may infer from the plot that the "Ultimus" pool has the fewest transactions, while the "Foundry USA" pool has the most transactions. Understanding which pools are actively contributing to the Bitcoin network and handling a larger volume of transactions with this knowledge is helpful. For investors and miners in the bitcoin industry, pool name analysis can assist in determining the popularity and efficacy of various mining pools.

1. **Contribution of each mining pool to the Total Pool Difficulty**

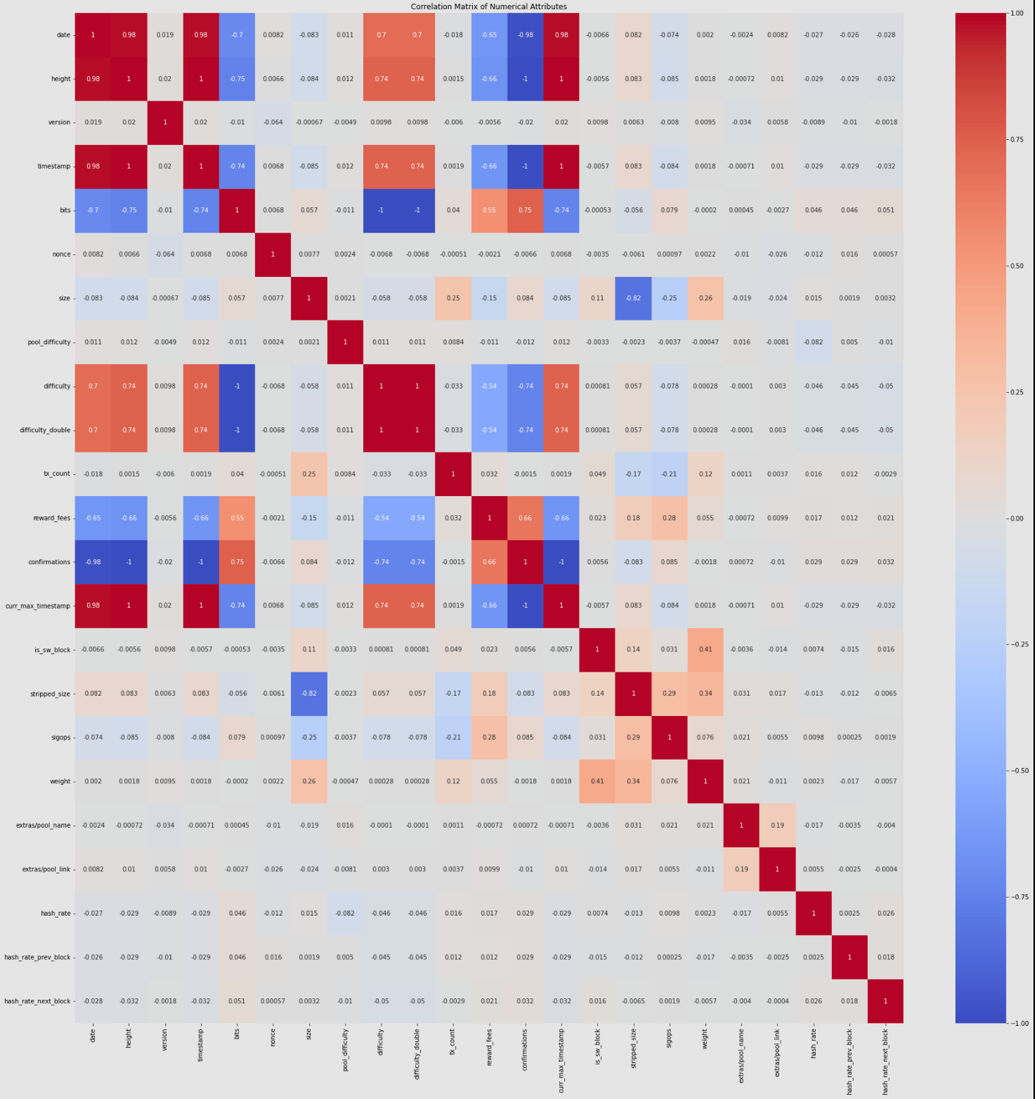
**T**reeMap visualization to represent the contribution of each mining pool to the total Pool Difficulty.



**Figure 11. TreeMap – Pool Contributions**

The TreeMap visualization will represent the contribution of each mining pool to the total pool\_difficulty as a hierarchical layout of nested rectangles. The size of each rectangle will be proportional to the percentage of pool\_difficulty contributed by the corresponding mining pool to the total pool\_difficulty.

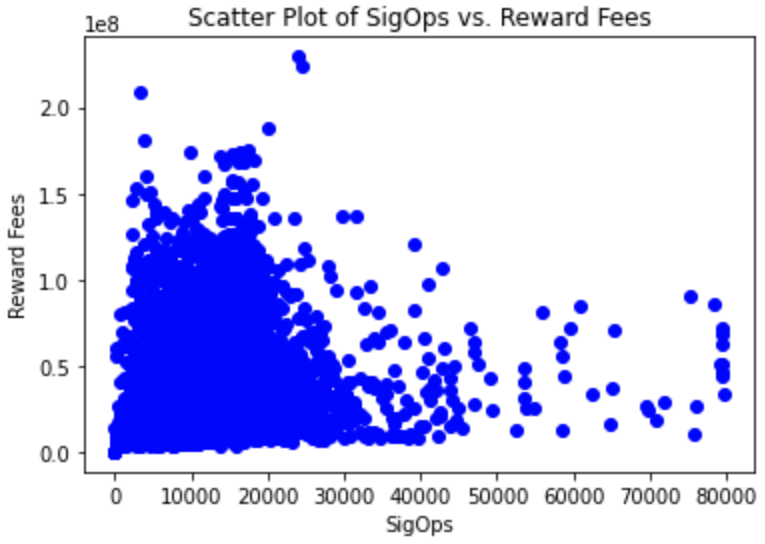
1. **Correlation Matrix across Numerical Variables**



**Figure 12: Correlation Matrix across Numerical Attributes**

**Observation:** A correlation matrix of the dataset's numerical properties makes up the third graphic. The coefficient correlation between any two qualities is shown in each matrix cell. A strong positive association is indicated by a correlation coefficient of 1, a strong negative correlation by a correlation coefficient of 1, and no correlation by a correlation coefficient of 0. We can see from the figure that the "sigops" and "stripped\_size" qualities strongly positively correlate with the "reward\_fees" property. This shows that incentive fees are likely to rise along with the number of signature procedures and stripped size. Additionally, there is a significant positive connection between the "size" and "transaction\_count" attributes, showing that larger transactions often have more transactions. The "size" attribute has the least association with the "stripped\_size" attribute, while the "height" attribute has the least correlation with "confirmations," respectively.

1. **Scatter Plot of SigOps vs. Reward Fees**



**Figure 13: Scatter Plot of SigOps vs Reward Fees**

**Observation:** The scatter plot of "sigops" and "reward\_fees" reveals an interesting relationship between the two attributes. We can observe that the "sigops" values range between 0 to 30000. As the number of signature operations (sigops) increases, we can notice a corresponding increase in the reward fees associated with each transaction.

The scatter plot shows a positive correlation between "sigops" and "reward\_fees," which suggests that transactions with a higher number of sigops tend to have higher reward fees. This relationship is consistent with the economic incentives in Bitcoin mining. Miners are economically motivated to include transactions with higher fees in their blocks, as it allows them to earn more rewards for their mining efforts.

1. **Bitcoin Blockchain Analysis and Insights**

Hashrate is a measure of the computational power used to mine or process cryptocurrency transactions, such as Bitcoin. It represents the speed at which a mining device or network can perform cryptographic calculations known as hash functions. A higher hashrate indicates more computational power and, therefore, a higher likelihood of successfully mining a block. The following dashboard provides insights using the factors such as ‘timestamp’, ‘bits’, ‘nonce,’ ‘tx-count’ that usually influence the hashrate.

A screenshot of a computer screen

Description automatically generated

**Figure 14. Bitcoin Blockchain Analysis & Insights**

1. Trend of block creation over time: By analyzing the trend of block creation over time using 'timestamp' and 'height,' we can identify patterns and fluctuations in the rate at which new blocks are mined. Sudden spikes or dips in the block creation rate may indicate changes in mining activity or network congestion. Additionally, observing the growth of the blockchain over time can provide insights into its overall health and adoption.
2. Number of Transactions mined per pool: The visualization of the number of transactions mined per pool using 'tx\_count' and 'extras/poolname' allows us to assess the distribution of transaction processing across different mining pools. We can identify which pools are actively participating in the mining process and their relative shares in the network. This insight can help us understand the decentralization or concentration of mining power among different pools.
3. Block Rewards for Miners in each Mining Pool: Analyzing block rewards for miners in each mining pool using 'Reward Block' and 'extras/poolname' helps us understand which pools are receiving higher rewards for successfully mining blocks. This insight can indicate the profitability and efficiency of different mining operations. It may also highlight whether certain pools consistently generate higher rewards, suggesting their competitiveness and expertise in mining.
4. Target Difficulty trend: By comparing the bars representing overall difficulty with the trend line showing pool\_difficulty, you can assess how the mining pool's difficulty level aligns with the network-wide difficulty. Significant differences between these two lines might indicate that the mining pool is using different settings or strategies. The alignment of the trend line with the bars can indicate how efficiently the mining pool's difficulty is adjusted. If the trend line closely matches the bars, it suggests that the mining pool's difficulty settings are effectively reflecting the overall network conditions.

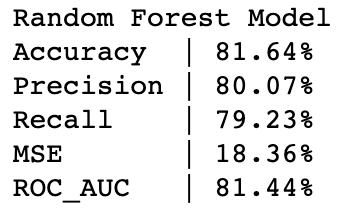
Overall, the dashboard provides a comprehensive view of the network's health, mining ecosystem, and the challenges faced by miners. By monitoring these metrics, we can make data-driven decisions related to mining strategies, network optimization, and overall engagement with the Bitcoin blockchain.

**PREDICTIVE MODELING**

1. **Random Forest**

In the context of the project focused on analyzing fluctuations in Bitcoin mining hash rate, the utilization of the Random Forest model holds significant potential for generating insights and predictions.

It's important to acknowledge that while the Random Forest model is a formidable asset, it operates in tandem with other factors in analysis. Domain expertise, a profound comprehension of the Bitcoin mining ecosystem, and further exploration of the features influencing hash rate fluctuations are integral to extracting meaningful insights from the model's predictions.



**Figure 15. Random Forest Model Accuracy**

1. **Accuracy (81.64%):**

* Accuracy quantifies the proportion of correctly predicted instances compared to the total instances within the dataset. In this context, the Random Forest model achieved an accuracy of about 81.64%, reflecting a notably high score.
* Such a high accuracy score signifies the model's effectiveness in correctly categorizing instances into their respective classes.

1. **Precision (80.07%):**

* Precision gauges the ratio of correctly predicted positive instances to the total instances predicted as positive.
* In the model, a precision rate of 80.07% indicates that approximately 80.07% of instances the model identified as positive (predicting an increase in hash rate) were indeed accurate.
* Elevated precision is particularly valuable when the consequences of false positives are significant, highlighting the model's cautious approach when forecasting positive instances.

1. **Recall (79.23%):**

* Recall, synonymous with sensitivity or true positive rate, assesses the proportion of correctly predicted positive instances relative to all instances within the actual positive class.
* A recall score of 79.23% indicates that the model correctly recognized around 79.23% of all actual positive instances.
* Robust recall is crucial when the implications of false negatives are substantial, indicating the model's competence in capturing genuine positive instances.

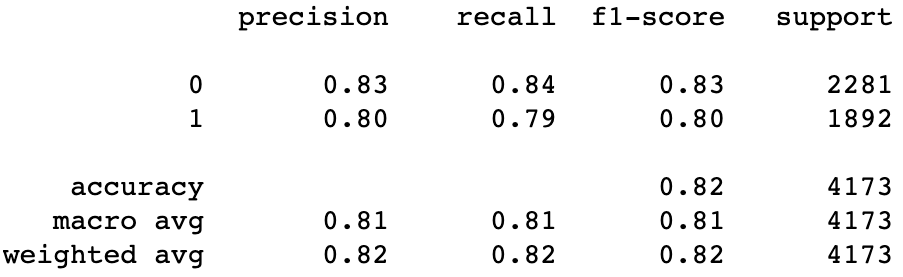
1. **Mean Squared Error (MSE) (18.36%):**

* MSE calculates the average squared deviation between predicted and actual values.
* Lower MSE values signify that the model's predictions closely align with the real values.
* A MSE of 18.36% suggests that, on average, the model's predictions bear a close resemblance to the true values of the target variable.

1. **Receiver Operating Characteristic Area Under Curve (ROC\_AUC) (81.44%):**

* ROC\_AUC quantifies the model's ability to differentiate between positive and negative classes.
* A value of 1 denotes perfect differentiation, while 0.5 indicates chance-level differentiation.
* An ROC\_AUC score of 81.44% underscores the model's proficient capability in distinguishing between the two classes.

**Observations:**

* The overall performance of the Random Forest model appears promising.
* Scores such as accuracy, precision, recall, and ROC\_AUC surpassing 80% suggest the model's propensity for reasonably accurate predictions.
* A balanced alignment between precision and recall implies a lack of excessive bias toward either false positives or false negatives.
* The MSE of 18.36% underscores the model's adeptness at delivering numerically accurate predictions.  
  

**Figure 16. Random Forest detailed Accuracy**

Let's break down the confusion matrix results:

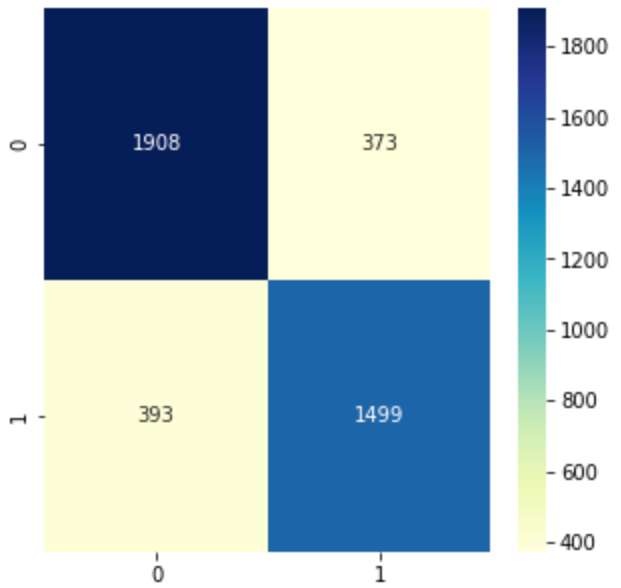
- **True Negative (TN)**: This corresponds to instances correctly predicted as "0" (no hash rate increase). In the dataset, the count of such instances is 1908.

- **False Positive (FP)**: This represents instances inaccurately predicted as "1" (hash rate increase) when they are genuinely "0". The dataset has 373 instances falling into this category.

- **False Negative (FN)**: These are instances incorrectly predicted as "0" when they are actually "1" (hash rate increase). The dataset contains 393 instances of this nature.

- **True Positive (TP)**: These instances are accurately predicted as "1" (hash rate increase). There are 1499 such instances in the dataset.

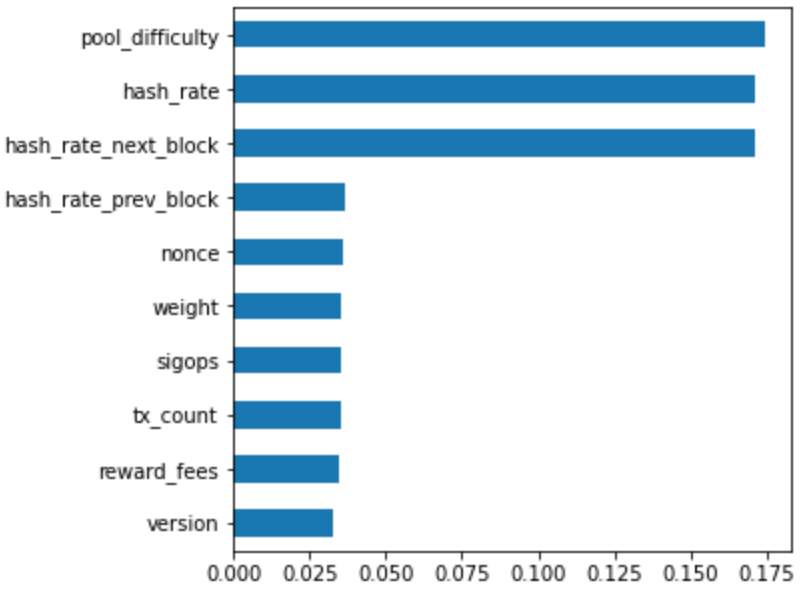
The confusion matrix figures offer a comprehensive view of the model's performance in distinguishing between hash rate increase and no increase scenarios.



**Figure 17. Random Forest Confusion Matrix**

Upon analyzing the plot, certain attributes emerge as highly influential according to the Random Forest model:

1. **pool\_difficulty**: The difficulty level of the mining pool also exhibits notable importance. This suggests that the mining pool's difficulty level strongly contributes to predicting hash rate increases.
2. **hash\_rate:** As anticipated, the current hash rate holds considerable significance. This aligns with the expectation that the hash rate directly signifies mining activity and performance.
3. **hash\_rate\_next\_block:** This attribute appears to exert a significant impact on predicting the target variable. It implies that the hash rate of the subsequent block holds a pivotal role in determining the likelihood of a hash rate increase.
4. **hash\_rate\_prev\_block:** Similarly, the hash rate of the previous block is influential. This indicates that the hash rate of the prior block might impact the hash rate increase in the current block.



**Figure 18. Feature importance using Random Forest**

1. **ARIMA Model**

The AutoRegressive Integrated Moving Average (ARIMA) model is a powerful time series forecasting technique used to analyze and predict patterns in sequential data. It combines autoregressive (AR) and moving average (MA) components with different ways to handle non-stationary time series data.

Data Preprocessing: Prior to modeling, the raw hashrate data was preprocessed. This involved converting the 'timestamp' column to datetime format and setting it as the index. The data was then resampled to a monthly frequency to obtain a smoother representation of the hashrate trends.

The ARIMA model was trained on the preprocessed hashrate data using the specified parameters. The model was then used to generate forecasts for the next month's hashrate. The forecasted value was integrated with the historical data to create a continuous visualization of hashrate trends.

The visualization generated from the ARIMA model showcases both historical and forecasted hashrate data. The historical data line demonstrates past trends, while the forecasted data point marks the prediction for the upcoming month. This graphical representation aids in understanding the model's ability to capture underlying patterns.

A graph with a line

Description automatically generated

**Figure 19. Monthly Hashrate Analysis**

The above plotted graph provides a visual representation of the monthly hashrate trend over time. If the hashrate data shows a consistent upward or downward trend, it indicates a general increase or decrease in computational power in the network. Seasonal patterns, such as fluctuations that occur at specific times of the year, can also be identified. Sudden spikes or drops in the hashrate could be indicative of events that affected mining activity, such as hardware changes or network upgrades. The visualization helps in identifying long-term trends, potential anomalies, and patterns that can guide further analysis and decision-making related to cryptocurrency mining or blockchain network behavior.

**A graph with a line and red dots

Description automatically generated**

**Figure 20. ARIMA Model Forecast**

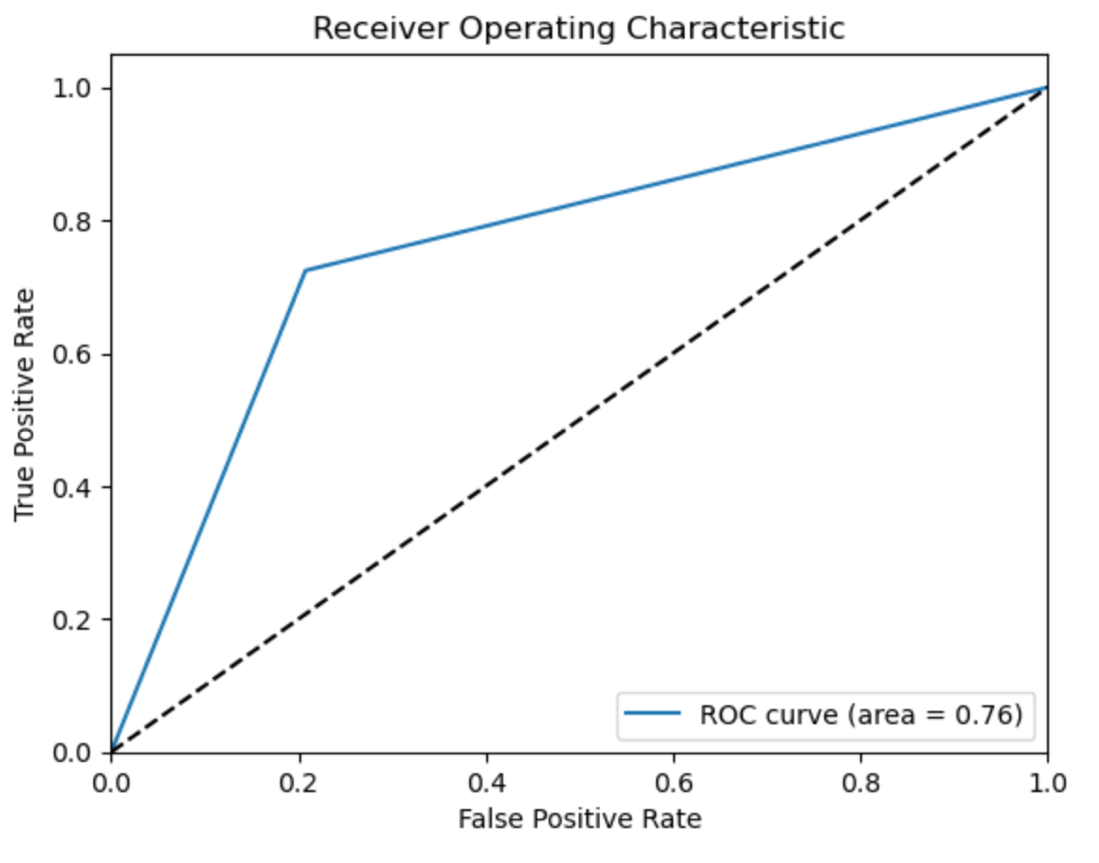
The line plot shows the historical monthly hashrate data as a continuous line and the forecasted hashrate as a red circle at the predicted point. The red circle indicates the forecasted hashrate for the next month based on the ARIMA model's simple average prediction. This visualization provides a clear representation of the model's forecasted value in the context of the historical data, aiding in understanding the model's performance and the predicted trend.

1. **Gradient Boost**

The Gradient Boosting Classifier is a powerful machine learning algorithm used for classification tasks. It's an ensemble method that combines the strengths of multiple weak learners (usually decision trees) to create a strong predictive model.

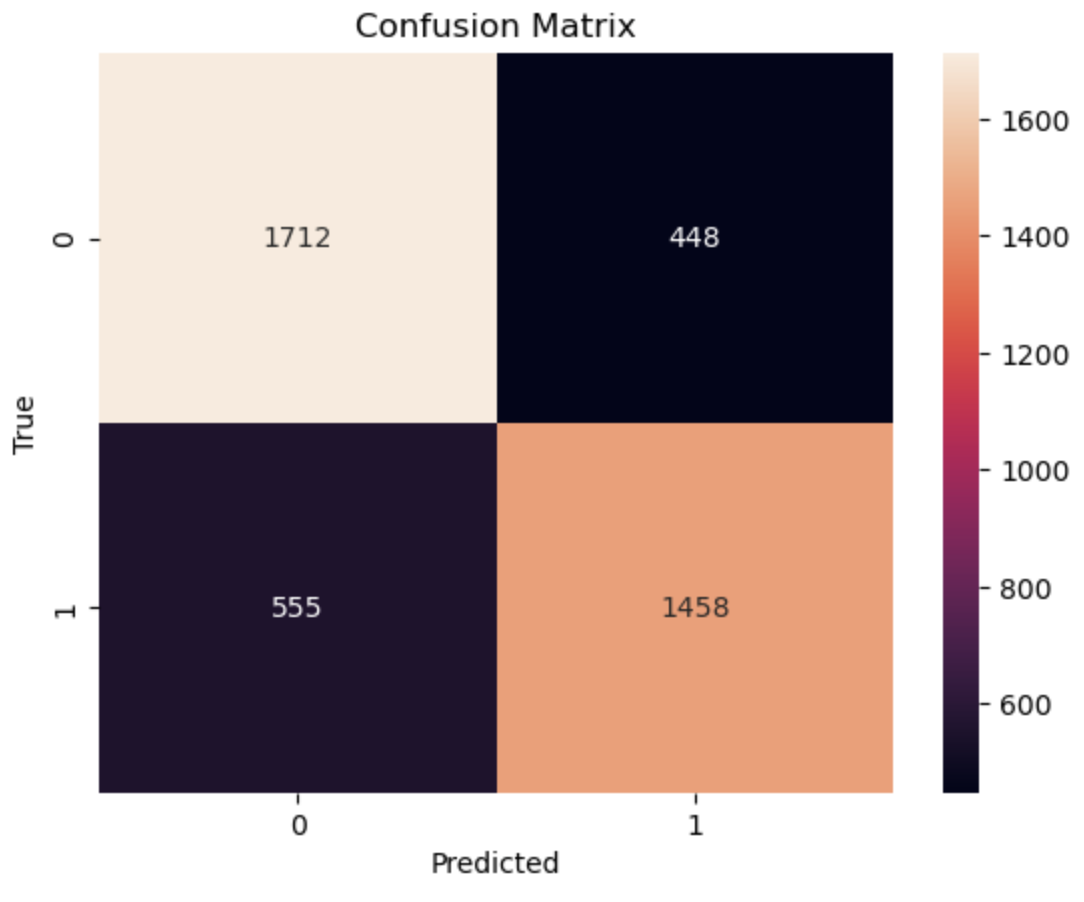
Gradient Boosting builds a series of decision trees sequentially, where each subsequent tree aims to correct the errors made by the previous ones. Each tree is constructed using a subset of the data, and the algorithm assigns more weight to the instances that were misclassified by the previous trees.

The Gradient Boosting Classifier model is a sophisticated machine learning algorithm used to predict whether there will be an increase in the hashrate or not.Hashrate, in the context of cryptocurrency mining,represents the computationla power dedicated to verifying transactions and adding them to the blockchain

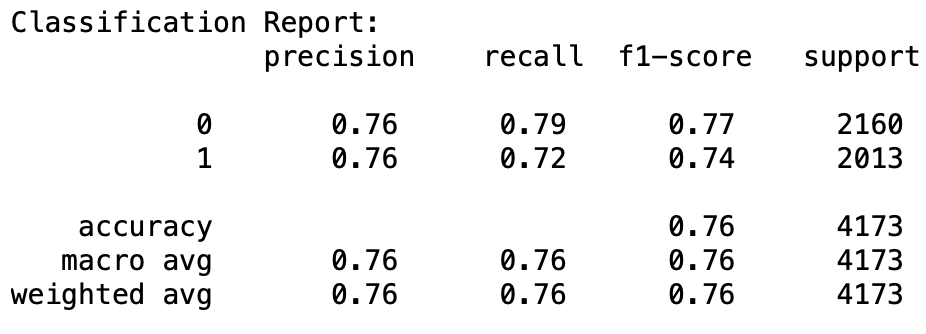


**Figure 21: ROC Curve for Gradient Boost Classifier**

The ROC curve visually shows how well your model can distinguish between the positive and negative classes. An AUC-ROC value of 0.76 means that the model has a moderate ability to differentiate between the two classes. The ROC curve helps you understand the trade-off between true positives and false positives. As the threshold for classifying an instance as positive changes, the trade-off between these two rates also changes.

****

**Figure 22: Confusion Matrix using Gradient Boost Classifier**

****

**Figure 23: Classification Report**

The model's accuracy is 75.96%, indicating that it correctly predicted the outcome (whether hashrate increases or not) for approximately 76% of the total instances in the test dataset.

For the "hashrate does not increase" class (0), the precision is 0.76. This means that out of all instances the model predicted as not increasing, 76% were correct predictions. For the "hashrate increases" class (1), the precision is also 0.76, indicating that 76% of the instances predicted as increases were accurate.

These averages provide an overall measure of the model's performance, considering both classes. The macro average is calculated by averaging the precision, recall, and F1-score for each class independently. The weighted average takes into account the class distribution in the dataset.

1. **Principal Component Analysis**

In order to perform the principal component analysis, the hash rate was first calculated using the total number of blocks mined and the time taken to mine blocks in seconds. This was considered the target variable, with other numerical attributes being considered as the independent variables. The categorical columns were either dropped or converted to numerical using one-hot encoding as required. PCA simplifies the data's complexity into linear combinations of attributes. PCA helps reduce the dimensionality of your data while retaining the most important information, making it easier to visualize and interpret underlying patterns and relationships within your dataset.

The first graph below highlights important insights into the cumulative explained variance and helps us determine how many principal components are needed to retain a certain amount of the original data's variability. The red dashed horizontal line represents a common threshold of explained variance, often set at 95%. The curve crosses the red line at, x = 10, meaning that using the first 10 principal components captures approximately 95% of the total variability in the data.

**A graph with a line

Description automatically generated A graph with blue dots

Description automatically generated**

**Figure 24. Explained Variance Plot Figure 25. Scree Plot for PCA**

The second graph shown above is a scree plot helps us to determine the number of principal components to retain in a PCA analysis based on the explained variance. Based on the "elbow point" on the curve, which indicates a significant drop in the explained variance, the number of principal components to retain can be determined. In this case, the elbow point is again at x=10.

**A graph of a number of blue bars

Description automatically generated with medium confidence**

**Figure 26. Explained Variance Ratio by PC**

The third graph above provides a visual representation of the variance explained by each individual principal component, allowing us to understand the contribution of each component to the total variance. The first bar (PC1) represents the percentage of variance explained by the first principal component. The second bar (PC 2) represents the additional percentage of variance explained by adding the second principal component, and so on. The height of each bar indicates the importance of each principal component in capturing the variability of the original data. Steeper increases in the beginning indicate that the initial components capture a significant portion of the variance.

**A screenshot of a computer code

Description automatically generated**

**Figure 27. Principal Component Loadings**

The fourth image above represent the loadings which are the coefficients that describe the relationships between the original variables and the principal components (PCs) obtained through PCA. Each loading value indicates the strength and direction of the correlation between a specific variable and a particular principal component. Variables with high absolute loadings (positive or negative) on a particular PC contribute significantly to that component. Loadings close to zero indicate minimal contribution.

1. PC1 seems to be influenced by attributes related to block height, timestamp, difficulty, size, reward fees, and confirmation. This suggests that these attributes are positively correlated with each other and tend to increase together.
   * The most influential variable is "height," with a high negative loading.
   * "timestamp\_unix," "difficulty," "difficulty\_double," and "confirmations" also have relatively high negative loadings.
   * "bits" and "tx\_count" have relatively high positive loadings.
2. PC2 is influenced by attributes like nonce, size, tx\_count, stripped\_size, and sigops. This indicates that larger transactions and certain operational attributes tend to decrease the value of PC2.
   * "size" and "weight" have high positive loadings, indicating that they contribute significantly to this component.
   * "total\_blocks\_mined" and "time\_taken\_to\_mine\_blocks\_seconds" have negative loadings, implying an inverse relationship.
3. PC3 is influenced by attributes like bits, pool\_difficulty, and weight. It suggests that higher values of these attributes are associated with lower values of PC3.
   * "weight" and "stripped\_size" have high negative loadings.
   * "sigops" has a high positive loading, indicating its strong association with this component.
4. PC4 to PC10: These components show more complex relationships and patterns, but they represent combinations of different attributes. For example, PC4 involves contributions from attributes like nonce, pool\_difficulty, and difficulty, indicating that these attributes are likely associated with each other.

**PROJECT ROADMAP**

The project roadmap as shown below outlines the key milestones and activities to achieve this research's objectives. Throughout the project roadmap, iterative feedback loops and continuous improvement will be emphasized to enhance the model's predictive capabilities and ensure its relevance in the ever-changing cryptocurrency landscape.

**A screenshot of a project

Description automatically generated**

**Figure 28. Project Roadmap**

Through this well-defined roadmap, the project aims to enhance understanding of Bitcoin hashrate fluctuations, contribute to the cryptocurrency ecosystem, and support Marathon Digital Holdings in making informed decisions for long-term success in the dynamic Bitcoin market.

**PROJECT CONSTRAINTS**

Based on the sponsor’s project, some constraints include:

1. **Data Availability and Quality:** The availability and quality of historical hash rate, block timestamps, miner hardware efficiency, and energy-price data may pose constraints on the accuracy and reliability of the analysis. Incomplete or noisy data could lead to potential biases or limitations in the findings.
2. **Time Constraints:** As the project has deadlines or limitations which could impact the analysis's depth and scope. Completing a thorough investigation within the allocated time could be a challenge.
3. **Model Complexity and Interpretability:** The requirement to create an explainable and predictable model for estimating the unexplainable variance in hash rate may impose constraints on the model's complexity. Thus, striking a balance between complexity and interpretability can be challenging.
4. **Assumptions and Limitations of Models:** When developing predictive models using assumptions about power availability, energy-prices, or game-theoretic models for hash rate manipulation, it is important to acknowledge the inherent assumptions and limitations of these models.
5. **Model Validation and Generalization:** It Ensures the validity and generalizability of predictive models to different scenarios and time limits may present constraints.

**ETHICAL CONSIDERATIONS:**

The "Bitcoin Hashrate Predictive/Regression Modeling and Analysis" project involves studying historical hashrate, block timestamps, and miner hardware data to understand Bitcoin mining hashrate variations. While the project aims to unveil variance drivers, ethical implications tied to cryptocurrency data analysis and modeling must be taken into account. Key ethical factors include:

1. **Data Privacy and Security**: The project involves handling historical hashrate and miner data, which may contain sensitive information about individuals and organizations involved in Bitcoin mining. It is crucial to ensure the privacy and security of the data throughout the analysis process, adhering to data protection laws and guidelines.
2. **Transparency and Informed Consent**: If the project involves accessing data from miners or other stakeholders in the Bitcoin ecosystem, it is essential to seek informed consent from these parties. Transparency about the purpose of data collection, analysis, and any potential outcomes should be provided to ensure that participants are fully aware of their involvement and the project's implications.
3. **Fairness and Non-Discrimination**: The analysis should be conducted in a way that ensures fairness and avoids discrimination against any specific group of miners or stakeholders. Bias in data collection or analysis that disproportionately affects certain participants should be minimized to maintain integrity and impartiality.
4. **Interpretability and Explainability**: As the project aim to create a predictive model for hashrate, it is important to ensure the model's interpretability and explainability. Stakeholders in the Bitcoin ecosystem should be able to understand how the model works and what factors influence its predictions. Clear communication of model assumptions and limitations is essential to avoid misinterpretation and potential misuse.
5. **Accountability and Impartiality**: The project should strive for objectivity and impartiality in its analysis and findings. Researchers should be accountable for the methods used and the interpretations made. Any potential conflicts of interest or biases should be disclosed and managed to maintain credibility and trustworthiness.
6. **Environmental Impact**: The project involves analyzing energy consumption data of Bitcoin mining. Ethical considerations should be given to the environmental impact of mining activities, especially in regions where energy sources may not be sustainable or have adverse ecological consequences.
7. **Use of Predictive Models**: If the project creates predictive models to explain the hashrate variance, caution should be exercised to avoid potential misuse of the models for manipulation or market speculation.
8. **Stakeholder Engagement**: Throughout the project, stakeholders within the Bitcoin ecosystem, including miners and industry experts, should be engaged in the process. Their feedback and insights can provide valuable context to the analysis and ensure that the research aligns with the interests of the broader community.

**CONCLUSION**

In conclusion, this data-driven effort aimed to understand and predict the fluctuations in Bitcoin's intraday hash rate, a critical measure of the network's processing capacity. Through comprehensive exploratory data analysis (EDA) and the development of robust predictive models, valuable insights were gained into the factors influencing hash rate changes. Summary statistics unveiled numerical characteristics of the dataset, while time series analysis revealed patterns and fluctuations in mining difficulty and pool contributions. Mining pool behaviors were unveiled through visualization, demonstrating varying degrees of dominance and decentralization. The relationship between attributes such as weight and stripped size was explored, reflecting nuanced connections within the blockchain ecosystem. Notably, the report demonstrated predictive modeling prowess through Random Forest, ARIMA, and Gradient Boost, enabling accurate predictions and trend analyses. Principal Component Analysis shed light on underlying attribute relationships. In sum, this report underscores the significance of data exploration in revealing valuable insights for stakeholders in the dynamic realm of cryptocurrency mining.

**REFERENCES**

1. Lumerin Protocol (2022, July 25). *Hashrate Explained: A Critical Aspect for Bitcoin’s Security.* Medium. Retrieved from: <https://medium.com/lumerin-blog/what-is-hashrate-a80271af40ac>
2. Oliveira, Marcio (2021, September 12). *Applying Linear Regression on Bitcoin’s historical data.* Medium. Retrieved from: <https://medium.com/coinmonks/applying-linear-regression-on-bitcoins-historical-data-94d986010283>
3. Nilshmeier. (2022). Bitcoin EDA - Exploratory Data Analysis. *Kaggle*. Retrieved from: <https://www.kaggle.com/code/nilshmeier/bitcoin-eda-exploratory-data-analysis/notebook>
4. GeeksforGeeks. (2022b). Bitcoin Price Prediction using Machine Learning in Python. *GeeksforGeeks*. Retrieved from: <https://www.geeksforgeeks.org/bitcoin-price-prediction-using-machine-learning-in-python/>
5. Chen, J. (2023) *Analysis of bitcoin price prediction using machine learning*, *MDPI*. Available at: <https://www.mdpi.com/1911-8074/16/1/51> (Accessed: 24 July 2023).