

**ALY 6980 – Capstone**

**Module 5 Assignment**

**Sponsor Project (Group Assignment)**

**Course Number (CRN): 90334**

**Term and Year: Summer 2023**

**Submitted to:**

**Prof. Valerie A. Atherley**

**Submitted by:**

**Amir Shabbir Shaikh**

**Tanisha Srivastava**

**Vatsal Shah**

**Abid Khan**

**Kunal Tutoo**

**Table of Content**

1. Introduction
   1. About the Sponsor
2. Problem Statement
   1. About the dataset
   2. Integration with Literature Review Article
   3. Project Phases and Timeline
3. Group Analysis Methodology
   1. Descriptive Statistics
   2. Exploratory Data analysis
   3. Visualization
4. Research Methodology
   1. Feature Selection
   2. Predictive modelling using Regression
5. Future Scope
6. Conclusion
7. References

**Introduction**

**About the Sponsor**

The Project cover’s vital aspects of the Sponsor Marathon Digital Holdings, services offered, and Individual Project Proposal based on the business requirement. Marathon Digital Holdings is a cryptocurrency mining company based in the United States. Formerly known as Marathon Patent Group, the company transitioned its focus to Bitcoin mining in 2020 and rebranded itself as Marathon Digital Holdings in 2021. Marathon Corporation is a NASDAQ-listed company that operates as a publicly traded company "MARA." The mission of Bitcoin is to make Bitcoin the world's most decentralized and secure monetary network by sustainably increasing computational power ("hash rate"). The primary business of Marathon Digital Holdings is to mine Bitcoin, which involves validating transactions and securing the Bitcoin network. Additionally, Marathon Digital Holdings has also made strategic investments in other cryptocurrency-related businesses, the company has invested in NYDIG, a financial services firm specializing in Bitcoin-related products and services. The company utilizes specialized computer hardware and sophisticated mining rigs to solve complex mathematical problems and earn Bitcoin rewards. By accumulating Bitcoin through mining, Marathon aims to build a portfolio of digital assets. It is worth noting that the cryptocurrency industry is highly dynamic and subject to rapid changes. In addition to Bitcoin value and mining difficulty, Marathon's financial performance is affected by various other factors. Fluctuations in Bitcoin's price can have a significant impact on the company's profitability.

**Problem Statement**

The purpose of this study concerns the analysis of the dataset provided by the sponsor. We have two components of the project to complete, the requirement of the Sponsor is to analyze the historical data to determine the variance and further gaining insights to the factors that drives those variances.  Secondly, if any portion of the variance falls beyond the normal distribution a model needs to be created to further predict the variance considering the hashrate. The above can be fulfilled by inheriting a variance score to check for hypothesis. Observed values are compared to predicted values to determine variability. It will provide the sponsor a basis for prediction to understand the correlation between parameters including miner's efficiency, energy price in comparison to power availability and how it further affects the variance score. Additionally, hashrate is a crucial factor to consider as it can help investors gauge the health and security of a cryptocurrency's network.

The proposed data analytics solution for maximizing daily mined blocks from the pool with the highest hash rate could potentially utilize various types of data points. Based on the data provided by the sponsor, the critical parameters under consideration are, Timestamp, height, version, bitcoins, transaction count, reward fees, confirmations, and pool name. By analyzing these data points, we can gain insights into the efficiency and profitability of mining activities within different pools, the overall network activity, and the factors influencing mining rewards. Utilizing this information, you can make data-driven decisions to maximize daily mined blocks from the pool with the highest hash rate, ultimately improving the efficiency and profitability of cryptocurrency mining operations.

**About the dataset**

The purpose of this study concerns the analysis of the dataset provided by the sponsor with 7th July 2023 Timestamp obtained from the official BTC website comprising of 156 records and 27 distinct variables. This dataset can provide valuable insights into the characteristics of individual blocks in the blockchain network, facilitating analysis and research related to mining, block validation, and overall network performance. The variables in the dataset represent various attributes and characteristics related to blocks in a blockchain network. Below is a short description of the important variable:

|  |  |
| --- | --- |
| **height** | The height or position of the block within the blockchain, indicating its sequence in the chain. |
| **version** | The version number of the block format used in the blockchain protocol. |
| **mrkl\_root** | An efficient means of verifying the integrity of transactions contained in a block is to use the Merkle root of those transactions. |
| **timestamp** | The timestamp when the block was mined, indicating the time of its creation. |
| **bits** | A numerical value representing the target difficulty for mining the block. |
| **nonce** | A random number used in the mining process for validating a block hash. |
| **hash** | The hash of the block header, a unique identifier for the block. |
| **prev\_block\_hash** | Blockchain's previous block hash, linking blocks together in a chain. |
| **next\_block\_hash** | Blockchain's next block hash, if available. |
| **size** | Block size in bytes. |
| **pool\_difficulty** | The difficulty level of the mining pool associated with the block. |
| **tx\_count** | An indication of how many transactions there are in a block. |
| **reward\_block** | The total reward (block reward + transaction fees) for mining the block. |
| **reward\_fees** | The block's transaction fees. |
| **confirmations** | The number of subsequent blocks added to the blockchain after this block. |
| **sigops** | The number of signature operations performed in the block. |
| **weight** | The block weight, a metric used in SegWit-enabled blocks. |
| **extras/pool\_name** | Additional information about the mining pool's name associated with the block. |

**Integration with Literature Review Article**

Drawing inspiration from the insights presented in "Analysis of Bitcoin Cryptocurrency and Its Mining Techniques", our analysis will focus on harnessing the potential of predictive modelling to enhance Bitcoin mining efficiency. By leveraging regression and classification models, we aim to uncover the complex relationships between mining variables and their influence on both the hash rate and Bitcoin prices. This comprehensive analysis will allow me to gain a deeper understanding of the factors driving variance in the hash rate and prices, enabling me to provide valuable guidance to miners and stakeholders. Additionally, we will implement feature selection techniques, such as random forest, to identify the most significant variables that play a pivotal role in predicting hash rate variance and prices. By incorporating real-time data analysis and the power of blockchain technology, we seek to optimize mining operations and make informed decisions in the ever-evolving world of cryptocurrencies. Through this research, we aim to contribute to the advancement of Bitcoin mining techniques and provide valuable insights for stakeholders navigating the complexities

of the cryptocurrency market.

A diagram of a blockchain process

Description automatically generated

**Working mechanism of bitcoin blockchain**

**Project Phases and Timeline**

A diagram of a project

Description automatically generated

# **A. Project Initiation (Understanding the requirement, Data Collection and Analysis) – Week 1,2,3**

1. Our primary focus will be on analyzing and collecting the crucial information from the data sets our sponsor has provided. Python would be used for data cleansing, data scraping, descriptive and statistical analysis to illustrate the results. Defining the specific metrics related to hashrate and Bitcoin that will be analyzed.
2. Create a project timeline, setting milestones and deadlines for each phase.
3. Identifying reputable sources of historical hashrate and Bitcoin price data and developing data acquisition methods to retrieve relevant and accurate data.
4. Conducting exploratory data analysis to understand the distribution and trends of the data, performing data cleansing and pre-processing to remove any inconsistencies and outliers. And calculating descriptive statistics, including means, standard deviations, and correlations.
5. Visualizing the data through graphs and charts to gain insights into the relationship between hashrate and Bitcoin price.

# **B. Market Analysis (Hypothesis formulation, Statistical analysis) – Week 4,5**

1. Conducting market research to determine the external factors driving the bitcoin prices based on the geographical location, energy prices, technology and process efficiency.
2. Formulate hypotheses based on domain knowledge, initial data analysis and market research and establishing null and alternative hypotheses for statistical testing.
3. Application of appropriate statistical methods to test the formulated hypotheses through regression analysis to determine the strength and direction of the relationship between hashrate and Bitcoin price.
4. Assessing the significance of the results and interpreting the findings.

# **C. Model Building (Predictive Modelling, Machine Learning) – Week 6,7**

1. This section will involve developing models to make a comparison between miner efficiency and energy-price data to understand the hash rate variance.
2. For this study, I plan on developing multiple models (KNN Models, Random, Forest, Neural Network, and Logistic Regressions) to identify the significant variables which can be used to understand block variance and its relationship to the hash rate. In order to answer the business questions, I will use parameters such as accuracy, MSE score, AIC score, recall, and precision to measure the performance of the model.

# **D. Project Conclusion (Decision-making, Review, Presentation) – Week 8**

1. Incorporating insights gained from the analysis into discussions and implementations with the Sponsor. Analyzing the methodology, results, and incorporating any necessary revisions based on Sponsor feedback.
2. Summarizing the findings in a comprehensive report, including data visualizations and statistical analyses followed by actionable insights and recommendations based on the results.
3. Final step is creating a presentation for the Sponsor containing the key findings, outcomes, and lessons learned including information for future reference, data sources, methodologies, and analysis approaches.

**Tools used in the Project analysis.**

**1. Data gathering and understanding:** Microsoft Excel

**2. Exploratory Data Analysis**: Python

**3. Data Interpretation and Insights:** Predictive Modelling – Python

4**. Data Visualization**: Python, Tableau

1. **What techniques have you used to explore the data you are working with?**

Data exploration is a crucial phase in the data analysis process. To get insights and make wise decisions entails comprehending the traits, patterns, and connections inside the dataset.

The techniques used by our team are as follows:

We can obtain summary statistics using descriptive statistics like mean, median, mode, standard deviation, minimum, maximum, and quantiles. These statistics offer a fundamental comprehension of the distribution of the data and the primary tendencies. We designed a variety of plots and charts to show the data visually using data visualization. Bar charts, line charts, scatter plots, histograms, box plots, heat maps, and other common visualizations are available. Data patterns, trends, outliers, and correlations can all be found through visualization. Using correlation matrices or scatter plots, we examine the connections between the dataset's various variables. Using correlation analysis, determine if a variable is favourably, negatively, or unrelated.

**What did we find?**

We performed a thorough data visualization research on a Bitcoin dataset to gain insights into the market behaviour and historical trends of the cryptocurrency by combining data exploration approaches, and we sought to visually display essential elements of the Bitcoin data, revealing light on its price movements, trade volumes, and potential relationships with other pertinent variables. This investigation aims to present a clear and understandable explanation of Bitcoin's market fluctuations through time through various charts, plots, and graphs. The visualizations presented in this study to identify patterns and anomalies open the door for strategic planning and well-informed decision-making in the constantly changing world of bitcoin investing.

1. **What techniques are you using to “tackle” your sponsor’s business question?**

To effectively address the sponsor's business question and gain comprehensive insights into the cryptocurrency market dynamics, I will employ a robust combination of data analytics techniques and advanced modelling approaches.

1. **Descriptive Analysis and Hypothesis Testing:** The initial step involves performing descriptive analysis on historical hashrate, block timestamps, and miner hardware efficiency and energy prices. By calculating summary statistics and visualizing the data, we can gain an understanding of trends, patterns, and potential variations in the variables. Next, hypothesis testing will be conducted to examine if the observed block discovery time variance aligns with the expected bounds of a normal distribution under the constant hashrate assumption. This analysis will help identify any significant variance in the data that cannot be explained solely by a constant hashrate hypothesis.
2. **Exploratory Data Analysis (EDA):** In the EDA phase, the data will be visualized using various charts and graphs to explore relationships and correlations between different variables. This exploration will help identify potential insights that may shed light on the factors influencing variance in hashrate and Bitcoin prices.
3. **Feature Engineering:** To address the unexplained variance (as per part b of the business question), feature engineering techniques will be employed. New variables will be created to capture additional information or underlying patterns that might be influencing the variance.
4. **Predictive Modeling:** The heart of the analysis lies in predictive modeling. I plan to use various machine learning algorithms such as Regression, Decision Trees, Random Forest, Neural Networks, and other Classification Models to develop predictive models. These models will aim to understand the influence of different variables on the likelihood of variance in hashrate and Bitcoin prices. The models will be evaluated based on their accuracy and performance.
5. **Machine Learning Algorithms:** One approach involves using miner efficiency and energy-price data, along with assumptions about power availability, to explain the observed variance pattern. Decision Trees, Random Forests, and Gradient Boosting algorithms will be considered.
6. **Game-Theoretic Models:** Another crucial aspect is analyzing the data using game-theoretic models to investigate if miners deliberately manipulate the hashrate to influence variance. This will provide insights into strategic behaviors and their impact on the fluctuations in hashrate.
7. **Principal Component Analysis (PCA) and Regression:** Additionally, I plan to employ PCA to identify the underlying factors that drive the variance in hashrate and Bitcoin prices. Subsequently, regression-based predictive models will be created to make future predictions based on these factors.
8. **Model Evaluation and Interpretation:** In order to ensure robustness and generalizability of all predictive models, appropriate metrics and cross-validation techniques will be applied to their evaluation. In order to gain insight into the factors influencing variance in the cryptocurrency market, the results of this study will be interpreted.
9. **Recommendations:** Based on the insights gained from the analysis, actionable recommendations will be offered to Marathon Digital Holdings. These recommendations will be aimed at optimizing their mining operations, adapting to market dynamics, and making informed decisions in the cryptocurrency landscape.

By combining these data analytics techniques and modeling approaches, I am confident that the research will yield valuable insights into the intricacies of Bitcoin mining and its impact on price dynamics. The results will not only contribute to the sponsor's understanding of the cryptocurrency market but also provide valuable guidance for investors, miners, and other stakeholders in the digital currency ecosystem.

**Group Analysis Methodology**

**Techniques and Software Platform** Utilizing and implementing software like Python and Tableau are part of the research analysis process for gaining data insights. The empirical analysis methodology involves the following, Data Pre-processing and profiling, data cleaning, visualization, correlation effect and modelling. We suggest the analysis of the research be done by looking at distributions and descriptive statistics of various factors, such as mining pools, block size, reward and transaction fees, etc.

**Descriptive statistics**

**Data, Analysis and Statistics table:** The data includes 156 observations with 27 attributes with presence of no duplicate values. Missing data can lead to a lack of precision in the statistical analysis therefore we replace the missing information by 0. The descriptive statistic information of the Dataset is as below.

A screenshot of a computer

Description automatically generated

* **Count**: It is evident that there are 156 records in the dataset
* **Mean**: This displays the mean of each attribute
* **Std**: A close clustering of data around the mean is seen in the variable nonce, which has the lowest standard deviation. By contrast, the variable with the highest standard deviation is the height, suggesting that the data points are more heterogeneous.
* **Min** – For all variables in the data set, this display shows the lowest value.
* **Max** - For all variables in the data set, this display shows the highest value.
* **Quartiles (25%,50%,75%)** – They statistical measures that divide a dataset into four equal parts, each containing 25% of the data. Quartiles are used to understand the distribution and spread of data.

# **Exploratory Data analysis**

In this section, we cover all the relevant elements of the context, mainly the whys and how’s. Based on predetermined variables, you can calculate, model, or visualize an outcome to determine how likely it is to occur. We will begin by profiling the raw data set, which contains 156 observations with 27 attributes. There are two steps involved in data profiling, data cleaning, and EDA.

**Data Pre-processing & Profiling:** By conducting Data Profiling, we will improve the quality of the data and increase its productivity. There are records of crime incidents that occurred in Boston included in this dataset.

A screenshot of a computer

Description automatically generatedA table of code with text

Description automatically generated

In the first step, we inspect the shape of the data set, the variables in the dataset, and the attributes in the dataset. The variables with null values are also checked and visualized to understand how many missing values there are in the variable and whether dropping the variable would lead to inaccurate analysis. The figure below represents the count of missing values in each of the columns.

A graph of a number of people

Description automatically generated

**Data Cleaning:** Data cleansing is the most crucial part of EDA, which involves checking for duplicate values and cleaning them. A total of 9 null values were found for "Extras/pool link". Furthermore, in order to check the unique values in each variable, we check the value counts of each attribute. Our data set does not contain any duplicate records nor missing data, as indicated by the output. The dataset in now ready for visualization and interpreting the trends in occurrence of incidents.

1. **How are you using those techniques?**

To effectively utilize predictive/regression modeling and analyze Bitcoin hashrate, a systematic step-by-step approach is crucial. The process begins with data collection and cleaning. Historical data on Bitcoin hashrate and related variables are gathered from reliable sources such as blockchain explorers, cryptocurrency exchanges, or research databases. Ensuring data quality is paramount, involving handling missing values, outliers, and inconsistencies in the dataset. Missing values can be imputed or removed based on the context, while outliers can be identified and appropriately treated to prevent undue influence on the model's results.

Next, time series analysis comes into play, focusing on studying temporal patterns and trends in the hashrate data. Techniques like Autoregressive Integrated Moving Average (ARIMA) or Seasonal-Trend decomposition can be employed to model the time series data and extract meaningful insights. The subsequent step involves regression analysis, aimed at modeling the relationship between the hashrate (dependent variable) and predictor variables (independent variables). Relevant predictor variables are chosen based on domain knowledge and correlation analysis from the exploratory data analysis (EDA) phase. Multiple linear regression or polynomial regression can be utilized to model the relationship effectively. To incorporate more complex relationships, machine learning models are employed. Selecting suitable algorithms like XGBoost, Support Vector Machines (SVM), Neural Networks and Random Forest can capture non-linear relationships between variables. Supervised learning techniques are used, where the model is trained on labeled data with the hashrate as the target variable.

To assess the model's performance and prevent overfitting, k-fold cross-validation is implemented. This technique tests the model's generalizability on unseen data by splitting the dataset into k subsets and evaluating the model on different combinations. During the evaluation phase, appropriate metrics like Mean Absolute Percentage Error (MAPE) , Mean Squared Error (MSE), R-squared, or Root Mean Squared Error (RMSE) are used to assess the model's performance. This helps in comparing the performance of different models and identifying the one that best predicts the hashrate. Additionally, model interpretability is critical for transparency and understanding. Analyzing feature importance reveals which predictors have the most significant impact on the hashrate. Techniques like SHAP (Shapley Additive explanations) or LIME (Local Interpretable Model-agnostic Explanations) can be employed to interpret individual predictions and comprehend the model's decision-making process.

**Visualization**

Performing exploratory data analysis (EDA) is essential for achieving project goals. Analyzing data visually, identifying patterns, and understanding relationships between variables are all part of EDA. By exploring the distribution, range, and central tendencies of the variables, we can gain a better understanding of the data and its characteristics. EDA also allows us to detect any missing values or data quality issues that may require resolution before building predictive models based on the dataset.

1. **Distinct Pool Counts**

A graph of different colored bars

Description automatically generated

The histogram depicting the occurrences of distinct pools up to July 7th, 2023, provides valuable insights into the dominance of different mining pools in the context of bitcoin mining. On that specific day, 'Foundry USA' stands out as the most dominant pool with approximately 55 occurrences, indicating its strong mining capabilities and popularity among miners. Following closely are 'Antpool' with 28 occurrences and 'F2Pool' with 20 occurrences, signifying their competitive positions in the mining landscape. Additionally, Sponsor pool 'MARApool' has a count of 5, suggesting its presence among the mining competitors. This visualization allows for a quick comparison of the market share of different pools based on their occurrences, which can serve as a foundation for further analyses and comparisons across various parameters related to bitcoin mining performance.

1. **Reward Fees in Bitcoin Mining Pools**

**A graph of different colored squares

Description automatically generated**

The presented visualization categorizes "reward fees" into four distinct bins, ranging from 0-10 million to 30-40 million. The highest number of reward fees falls in the 10-20 million range, indicating it is the most common fee range associated with mining blocks. The second-highest number of reward fees lies in the 0-10 million range. This suggests that some transactions carry a higher premium, encouraging miners to include them in their candidate blocks to maximize earnings. The visualization provides valuable insights into the reward structure and revenue potential for miners, shedding light on the relationship between transaction fees and the urgency or value of transactions processed on the Bitcoin network.

1. **Comparing Signature Operations (SigOps) Across Bitcoin Mining Pools**

A graph of different colored bars

Description automatically generated

The visual representation comparing the number of signature operations (SigOps) among different Bitcoin mining pools clearly shows that the "Luxor" pool utilizes the highest number of SigOps, with approximately 14,000. In contrast, the "Binance" Pool has the lowest number of SigOps, around 8,000. This observation suggests that the Luxor pool might be employing more costly CPU-based hardware for its mining operations compared to the Binance pool. The significant difference in the number of SigOps indicates varying mining strategies and hardware capabilities among the pools, with Luxor opting for more resource-intensive operations, potentially aiming for higher mining efficiency or rewards.

1. **Average Reward Fees for various mining pools**

A graph of different colored bars

Description automatically generated

We have discovered a fascinating insight from the graph above. Compared to Foundry USA, Braiins Pool, and ULTIMUS Pool have substantially higher average reward fees. The combined prize of the Braiins Pool and ULTIMUS Pool is around 23 million, but Foundry USA's reward fee is just about 14 million. even though Foundry USA mined a far higher quantity of blocks. The bar graph offers an intriguing look at different mining pools' typical reward fees. Surprisingly although mining less blocks than Foundry USA, Braiins Pool and ULTIMUS Pool have much higher average reward fees. While Foundry USA's incentive fees are around 14 million, those for Braiins Pool and ULTIMUS Pool are more than 23 million. Even while Foundry USA mines significantly more blocks than the other two pools, its average reward fees are very different from those of the other two pools. This demonstrates that due to the reputation and effectiveness of the pool, users charge greater fees for their transactions.

1. **Comparing Block size Across Bitcoin Mining Pools**

A graph with different colored bars

Description automatically generated

The visual representation comparing the block size obtained among different Bitcoin mining pools provides valuable insights into the distribution of block sizes and the performance of each mining pool. The observation that "Binance" and "Antpool" obtained the highest number of blocks with approximately 1.8 MB in size suggests that these pools are consistently generating larger blocks, possibly indicating a higher computational power and efficient mining operations. On the other hand, the "Ultimus" pool's lowest number of block sizes received, around 1.3 MB, may indicate relatively lower computational power or a different mining strategy.

1. **Pool Difficulty for each mining Pool**

A graph of pool difficulty

Description automatically generated

The graph illustrates how the pool difficulty varies between mining pools. A cryptocurrency network's pool difficulty gauges how tough it is to discover a workable block solution. An increased pool difficulty means the pool attempts to solve more challenging mathematical puzzles to mine new blocks. F2Pool, ViaBTC, and AntPool are the top three mining pools regarding pool difficulty. The top three pools, F2Pool, ViaBTC, and AntPool, indicate a competitive atmosphere in the mining industry. Mining pools are rewarded with cryptocurrency when they are effectively mine a new block. They compete to solve cryptographic riddles. Marathon Digital Holding can use the information from this graph to guide strategic decisions. The distribution of pool difficulty might help provide information about the mining operations' competitive environment. If Marathon Digital Holding runs its mining pools, consider looking at the tactics used by these pools to increase their effectiveness and profitability.

1. **Correlations Between Variables**

**A table of numbers and a number of squares

Description automatically generated**

Correlation matrices show how variables are correlated. A matrix format is used to display the correlation between all possible pairs of values. By using a correlation matrix, we can identify patterns in a large dataset and make decisions accordingly. Moreover, we can visualize our results and identify the variables that are more correlated with one another. The correlation plot that was created for this dataset gives the relationship between the highest performing variables. By changing numbers to colors of varying shades, as shown on the correlation plot with color coding, the basic idea is to replace numbers with colors. Negative correlation occurs more frequently in lighter cells, that is, when one variable is positively influenced, it impacts the corresponding correlated variable negatively. Visualizing linear relationships between variables is much easier than looking at tables of numbers. The presented heatmap provides valuable insights into the relationships between important variables in the dataset after a thorough cleansing process. It reveals a strong positive correlation (close to 1) between "reward fees" and "sigops," indicating that mining operations with higher SigOps tend to yield higher rewards. However, there are no significant correlations between "reward fees" and "pool difficulty" or "timestamp," suggesting that "reward fees" may not be a decisive factor in determining patterns within this dataset. Overall, the heatmap efficiently helps identify patterns and relationships, enhancing our understanding of the underlying data structure and facilitating further analysis.

1. **Why are you using those techniques?**

Predictive modelling plays a crucial role in the world of Bitcoin price forecasting and trading due to several significant reasons.

* **Market Complexity:** The cryptocurrency market, particularly Bitcoin, is known for its complexity and susceptibility to various influencing factors. Market sentiment, macroeconomic developments, regulatory changes, and technical advancements can all impact Bitcoin's price. Predictive modelling offers a way to make sense of this intricate landscape and identify potential price fluctuations based on data-driven analysis.
* **Data-Driven Decision Making:** Predictions in Bitcoin trading are driven by predictive modelling, which relies on historical price data, trade volumes, and other relevant variables. By employing data analytics methods, traders and analysts can base their trading decisions on factual information rather than subjective judgments or emotions.
* **Pattern Recognition:** One of the key advantages of predictive algorithms is their ability to identify trends and patterns in past price data that might not be immediately apparent to human traders. These algorithms can uncover recurring price trends and identify potentially lucrative trading opportunities.
* **Predicting Price Movements:** Forecasting future price movements is a primary objective of predictive modelling in Bitcoin trading. By making accurate predictions, traders can make informed decisions about when to enter or exit positions, potentially maximizing profits and minimizing losses.
* **Time-Series Analysis:** Bitcoin price data is inherently sequential and time-dependent, making time-series analysis essential. This type of analysis helps to uncover trends, seasonality, and other patterns in past price data, providing insights into potential future price movements.
* **Machine Learning Techniques:** Various machine learning techniques are employed to predict Bitcoin prices. Regression models, such as linear regression and multiple regression, can capture the relationship between historical prices and other relevant factors like trade volume or market sentiment. Advanced models, such as support vector machines (SVM), decision trees, random forests, and neural networks, can uncover more intricate patterns in the data.
* **Evaluating Trading Strategies:** Predictive modelling allows traders to back test and assess the effectiveness of their trading strategies using historical data. This enables them to fine-tune and optimize their strategies based on past performance, enhancing their overall trading approach.
* **Risk Management:** Effective risk management is vital in trading, and predictive modelling plays a crucial role in this aspect. By being aware of probable market changes and the associated risks, traders can employ methods like stop-loss orders and position sizing to protect their capital and manage risk effectively.

Predictive modelling is an indispensable tool in Bitcoin price forecasting and trading. It empowers traders and analysts to navigate the complexities of the cryptocurrency market, make data-driven decisions, identify profitable opportunities, and manage risk effectively. By harnessing the power of predictive algorithms and data analytics, market participants can gain a competitive edge and achieve success in the dynamic and ever-evolving world of Bitcoin trading.

1. **How will it be delivered to the sponsor?**

The findings and results of the research, including the analysis of Bitcoin mining efficiency, variance in hashrate, and price dynamics, will be delivered to the sponsor in a comprehensive and actionable report, python code file and an interactive Tableau dashboard. The report will be structured to present the research methodology, data analysis techniques, and the results of the predictive models in a clear and organized manner.

**Project Report:** Providing an overview of the key findings and recommendations, the report will begin with an executive summary. The sponsor will be able to gather the most important insights quickly without having to delve into the technical details. Next, the report will outline the research objectives, scope, and the methodology employed to address the sponsor's business question. It will detail the data sources used, the data preprocessing steps, and the specific analytics techniques applied, such as descriptive analysis, exploratory data analysis, and predictive modeling. The predictive models' performance metrics, such as MAE, RMSE, and MAPE, will be reported to demonstrate the accuracy and reliability of the models. This will allow the sponsor to assess the predictive power of the models and make informed decisions based on the results. Additionally, the report will include actionable recommendations based on the research findings. These recommendations will be tailored to the sponsor's specific business needs and objectives. They will provide guidance on optimizing mining operations, adapting to market fluctuations, and making strategic decisions in the cryptocurrency landscape.

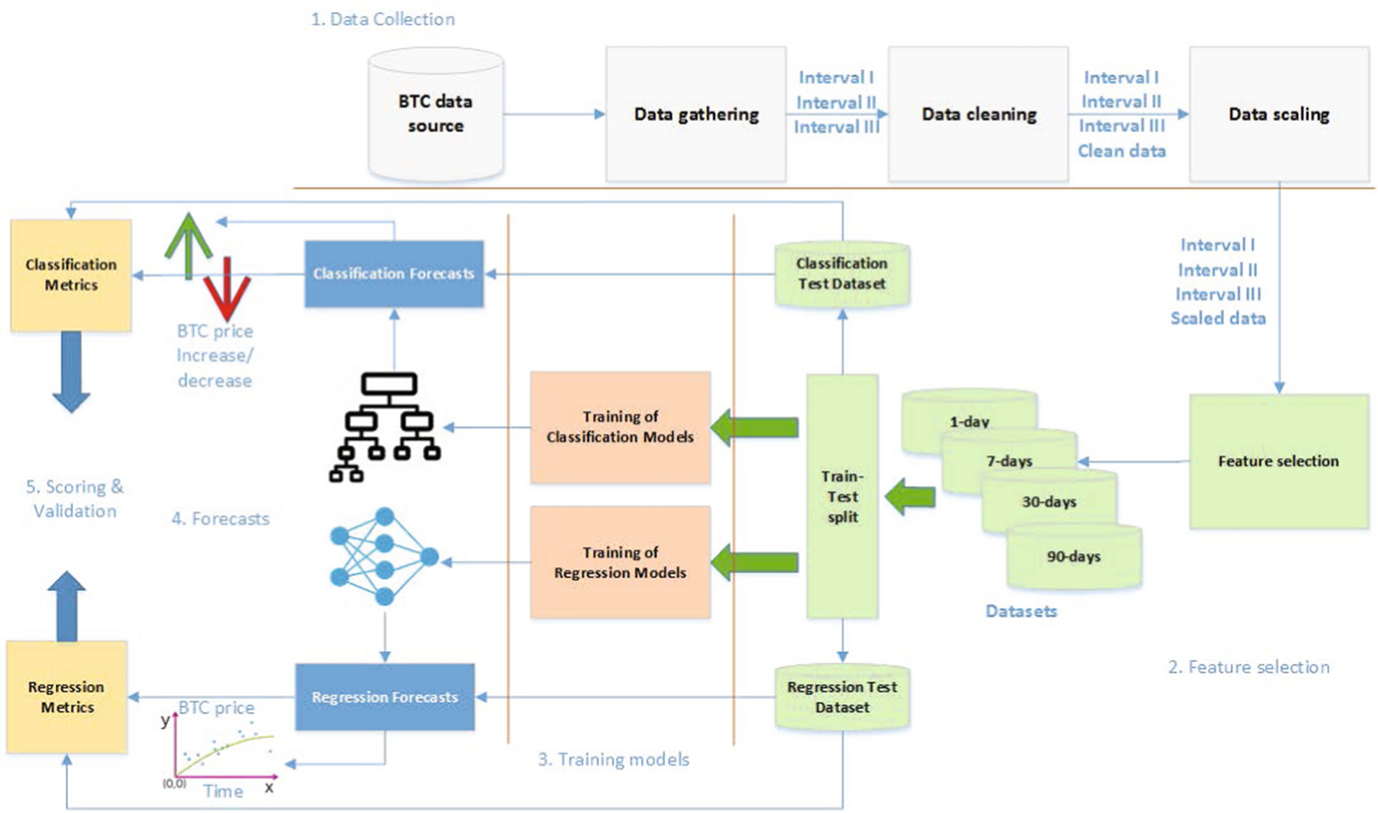
In addition to the comprehensive report, we will also provide the sponsor with Python code and Tableau dashboards as part of the deliverables. These tools will enhance the sponsor's ability to explore and interact with the data, gain deeper insights, and make informed decisions based on the research findings.

**Python Code:** The Python code will be well-documented and organized, allowing the sponsor to replicate the data analysis and modeling process. It will include all the data preprocessing steps, descriptive and exploratory analysis, feature engineering, and the implementation of predictive models using various machine learning algorithms. We will use popular Python libraries such as Matplotlib, NumPy, Seaborn, Scikit-learn and Pandas to perform the data manipulation, visualization, and modeling tasks. The code will be commented to explain each step and ensure readability. By providing the Python code, the sponsor can easily re-run the analysis on updated data or apply similar techniques to other datasets in the future.

**Tableau Dashboards:** The Tableau dashboards will be interactive and visually appealing, presenting the key findings and insights in a user-friendly manner. The dashboards will include dynamic visualizations, such as line charts, scatter plots, heatmaps, and bar charts, to showcase trends, patterns, and correlations in the data. The sponsor can interact with the dashboards, apply filters, and drill down into specific details to explore the data from different angles. The Tableau dashboards will be linked to the Python code, allowing for seamless integration between the analysis and visualization components. The sponsor can access the Tableau dashboards through a web link or Tableau Reader, enabling them to explore the data and findings at their convenience.

Overall, the combination of the comprehensive report, Python code, and Tableau dashboards ensures that the sponsor receives a complete package of insights, analysis, and tools to navigate the complexities of the cryptocurrency market and make informed decisions for their business.

**Mock-up Results**

The purpose of this study is to use machine learning to forecast time-series BTC prices. There are successive moments in time represented by the values of a time series. By analyzing time-series data, a time-series forecast can be made for the future. In this research methodology, we adopt a Machine Learning (ML)-based time-series forecast approach to analyze cryptocurrency price movements, particularly focusing on Bitcoin (BTC). The process commences with the construction of a comprehensive dataset, which entails the gathering and cleaning of BTC price technical indicators from open data sources. Subsequently, data scaling and normalization are performed to ensure the uniformity and comparability of the collected data. To address the underlying interdependencies in cryptocurrency price time series, feature selection assumes paramount importance. We employ the random forest (RF) method to identify high-ranking features from different datasets, each representing various forecast horizons and time intervals. This selection of features is pruned based on variance inflation factor (VIF) and Pearson cross-correlation to retain only the most relevant explanatory features. Technical market indicators as well as statistical data on the blockchain's operations are included.

The datasets are then divided into training and validation sets for subsequent ML model training. For the regression models, such as Artificial Neural Networks (ANN) and Long Short-Term Memory (LSTM), missing data cases are imputed using linear interpolation or the most frequently occurring value. The training sets undergo fivefold cross-validation to effectively train the stacked ANN models.

To ensure robust and accurate predictions, appropriate feature scaling methods are applied to the ML models. The use of ML methods in time-series forecasting differentiates our approach from traditional model-based methods. The training data effectively captures statistical trends and underlying features using machine learning models, facilitating accurate predictions for previously unseen cases. In the context of this research, ML is employed for both classification and regression tasks to forecast cryptocurrency prices. By adhering to this rigorous methodology, we aim to unravel the complex dynamics of cryptocurrency price movements, leveraging the power of ML to provide valuable insights into market trends and underlying factors influencing BTC price fluctuations.

**Feature selection**

In this research methodology, feature selection emerges as a pivotal step during the data pre-processing phase, aimed at enhancing model performance in predicting cryptocurrency price movements, specifically focusing on Bitcoin. The process involves iterative extraction and pruning of features using various approaches to ensure the selection of the most relevant and influential variables.

To begin, we employ an ensemble method based on random decision forests to determine feature importance. This ensemble approach utilizes multiple decision trees, which allows for more accurate and robust forecasts compared to individual decision trees. Subsequently, the reduced feature set undergoes examination to identify multi-collinearity and cross-correlations. To assess multi-collinearity, we utilize the Variance Inflation Factor (VIF) method, which quantifies the extent to which a feature is linearly correlated with a combination of other features. Additionally, Pearson correlation is employed to assess cross-correlations between features.

Iteratively, we arrive at a subset of features that are relatively high in importance, have low cross-correlations, and are not multi-collinear. This subset of features is deemed suitable for the specific time intervals under consideration. Moreover, when predicting for different forecasting horizons, such as nth day ahead, the feature selection process is repeated to tailor the feature set to the specific forecasting period. This ensures that the selected features are optimal for forecasting prices over varying time frames.

**Predictive Modelling using Regression**

In the context of our research methodology, the evaluation of regression models for price forecasting entails the utilization of specific performance metrics to assess the reliability and accuracy of predictions. The key metrics employed for this evaluation are the mean absolute percentage error (MAPE), mean absolute error (MAE) and root mean squared error (RMSE).

The MAE measures the average absolute difference between the predicted price and the actual price of Bitcoin. A lower MAE value indicates a more accurate model, where the predicted prices closely align with the actual market prices. For instance, an MAE of 5 signifies that the model's predictions deviate by approximately USD 5 from the true price. The MAPE, on the other hand, quantifies the prediction errors in terms of percentages. It is particularly useful for understanding the relative accuracy of the model's forecasts, regardless of the price scale.

The RMSE, another crucial metric, indicates the dispersion or spread of forecast errors. It captures the variability in prediction accuracy, where a higher RMSE value suggests occasional erratic predictions even if the model exhibits lower MAE or MAPE. Thus, a comprehensive evaluation of regression models should consider all three metrics to gain a holistic understanding of their performance.

By employing these performance metrics, we aim to rigorously assess the forecasting abilities of our regression models and identify the most effective approach for predicting Bitcoin prices. The models with low MAE, MAPE, and RMSE values are deemed desirable, as they demonstrate higher accuracy and precision in their price predictions. This evaluation process is essential in guiding the selection of the most optimal regression model, ultimately contributing to reliable and informed decision-making in the dynamic and ever-changing cryptocurrency market.

**Future Scope**

In the upcoming weeks, our objective is to conduct a thorough analysis of the cryptocurrency market using a range of models, including Neural Networks, Regression, Classification Models , Decision tree and Random Forest. Each model offers unique strengths, enabling us to explore complex relationships between variables and their impact on hash rate variance and Bitcoin prices. Through comparative analysis, we will determine the most suitable model for the data and the specific problem at hand. Additionally, we will employ feature importance techniques to identify crucial factors driving hash rate variance and Bitcoin prices. This combined approach of traditional statistical models and advanced deep learning will provide valuable insights for investors and stakeholders, enabling informed decision-making in the dynamic cryptocurrency market. Ultimately, this research will equip us with essential knowledge and insights to navigate and capitalize on opportunities in the evolving world of digital currencies.

**Conclusion**

In this project, we explore and analyze the intricate relationship between Bitcoin mining, hash rate variance, and price dynamics in the cryptocurrency market. Through extensive data analysis, regression modeling, and predictive techniques, we are gaining valuable insights that shed light on the factors influencing these interconnected aspects of the industry. Our analysis of historical data provides crucial insights into the variations in hash rate and its impact on Bitcoin prices. By employing a diverse set of classification models, we are gaining a comprehensive understanding of the complex relationships between different variables and their influence on hash rate variance and price movements. Through feature importance techniques, we identify the key factors that play a significant role in predicting hash rate variance and Bitcoin prices, aiding investors and stakeholders in making informed decisions.

Our research highlights the importance of market complexity in cryptocurrency, where technical, blockchain-based, sentiment/interest-based, and asset-related components all come into play. By leveraging advanced techniques like Random Forest and Neural Networks, we can capture high-dimensional statistical trends and underlying features, enabling more accurate price forecasts. In the dynamic and rapidly evolving cryptocurrency market, the comprehensive analysis performed in this project will serve as a valuable guide for investors, miners, and stakeholders. The findings contribute to a deeper understanding of the market's complexities, offering valuable patterns and trends that can lead to more strategic and informed decision-making. By combining traditional statistical models with advanced machine learning techniques, this research opens up new possibilities for predicting and understanding the dynamic behavior of cryptocurrencies.

The outcomes of this project provide a solid foundation for further research and advancements in the field of cryptocurrency mining and price forecasting. The results contribute to a more comprehensive understanding of the cryptocurrency market and hold great promise for future developments in this exciting and fast-paced industry. The findings will continue to shape and inform the decisions of stakeholders in the ever-changing world of digital currencies.

Top of Form

Bottom of Form

**References**

Marathon Digital Holdings Announces Bitcoin Production and Mining Operation Updates for June 2023. (2023, July 5). GlobeNewswire. Retrieved from https://www.globenewswire.com/news-release/2023/07/05/2699990/0/en/Marathon-Digital-Holdings-Announces-Bitcoin-Production-and-Mining-Operation-Updates-for-June-2023.html

Marathon Digital Holdings | Setting the pace for the Bitcoin mining industry. (n.d.). Marathon Digital Holdings. Retrieved from https://www.mara.com/

Bitcoin Network Hash Rate. (n.d.). YCharts. Retrieved from https://ycharts.com/indicators/bitcoin\_network\_hash\_rate

Waskom, M. L. (2023, March 8). Seaborn: statistical data visualization. PyData. Retrieved from <https://seaborn.pydata.org/>

*BTC.com Blockchain Explorer - Playground*. (n.d.). BTC.com Professional Data Service for Global Blockchain Enthusiasts. <https://explorer.btc.com/btc/adapter?type=api-doc>

Bitcoin price today, BTC to USD live, marketcap and chart | CoinMarketCap. (n.d.). CoinMarketCap. <https://coinmarketcap.com/currencies/bitcoin/>