Machine Learning on Diamond, Stock and Bitcoin

Ankit Kumar
National College of Ireland, x23123061@student.ncirl.ie

Abstract - This study explores the application of sophisticated machine learning techniques to identify patterns in diverse datasets, including Apple's stock prices, Bitcoin fluctuations, and diamond costs. It navigates the intricate landscape of stock and cryptocurrency markets by employing a dual-model strategy for Apple stocks and binary classification methods for Bitcoin. Drawing from personal experiences of selecting a diamond ring for my spouse, the research employs logistic regression and KNN models to decipher the complex factors influencing diamond prices, such as cut, color, clarity, size, and cost. The research contributes to our understanding of the versatility of machine learning in various domains and its adeptness at uncovering intricate patterns. The abstract encapsulates the key findings related to predicting stock prices, adapting to the volatile nature of cryptocurrencies, and determining diamond costs.

Keywords – Advanced Machine Learning, Apple Stocks, Bitcoin, Diamonds, Predict Stock Prices, Diamond Affordability

I. INTRODUCTION

A. APPLE STOCK PRICE PREDICTION

I never used to invest in the stock market. When I started working in Accenture in 2016, most of my friends used to invest in the stock market and crypto currency. That was the point when I used to look down at these things and used to believe that it is highly volatile, and profit is not guaranteed. You can say that I was not risk averse. I completed my MBA from Brandeis International Business School, Boston, United States of America in 2021. When I went to the US in 2019, many of my international friends used to invest in stock market and crypto currency. That was the point when I Invested in Apple stock. That was the first stock that I ever bought. I don't use Apple products but only because there is a special memory of Apple stock, I chose my first dataset as Apple stock prices.

In the fast-paced world of financial markets, being able to accurately guess stock prices is key to making smart business choices. This study uses advanced machine learning techniques to make better estimates about stock prices. When it comes to financial data, traditional methods often fail to pick up on complex trends and connections that don't follow a straight line. We will look at past stock data from several decades ago, including things like the starting and ending prices, the best and lowest prices, the trade volume, and time factors like the month and day. A two-model method is used, combining linear regression with number year values and category year factors. Additionally, the study investigates a k-nearest neighbors (KNN) regression model to see how well a non-parametric method can predict stock prices.

B. BITCOIN PRICE DYNAMICS

Like Apple stock, I started investing in bitcoins as well. However, I did not realize a lot of profit but investing $200 and getting 10 % return is not bad. I am still pitying myself for only investing $200 and not $2000. Well, enough with the greed here but the world has become very interested in cryptocurrencies because they are very volatile and not controlled by any one group. Bitcoin is the first and most widely traded cryptocurrency, making it stand out among the many other digital assets. In this situation, it's important to understand and predict how the price of Bitcoin will change. This paper uses machine learning to model and predict how the price of Bitcoin will change over time, considering the unique problems that the bitcoin market presents. Our information covers a wide range of time periods and includes different market conditions and events that influence Bitcoin's price. After carefully cleaning the data and designing the features, a binary classification method is used. When decision tree and random forest models are taught, things like starting and ending prices, trade amounts, and measures of time are used.

C. DIAMOND AFFORDABILITY

I got married in December' 2022. Well, that was the best moment of my life. Working in the USA, I wanted to give my wife a diamond ring, but I was so confused about the different metrics of diamond that I eventually ended up gifting here a cheap quality diamond. To my surprise, I always used to feel that all diamonds are costly but never thought that cheap quality diamonds also exist. Cut, color, clarity, size, and price are just a few of the factors that make up the diamond business. Using advanced machine learning methods, this study investigates the complicated world of diamond pricing. We use logistic regression and k-nearest neighbors (KNN) classification models to find complex trends in the dataset.

Putting together what we've learned from these different areas shows how flexible machine learning is when it comes to finding complicated patterns and giving us useful information. The next parts will talk about the methods, findings, and conversations, putting together a full story that shows how data science and different fields interact, eventually adding to the bigger picture of applied machine learning.

II. Related Works

A. *Bitcoin*

Torous et al. [1] provided an empirical analysis of the price dynamics of Bitcoin using time series analysis and machine learning techniques. They find that Bitcoin prices are highly volatile and exhibit long memory. They also find that Bitcoin prices are correlated with other asset classes, such as gold and equities. Zhang et al. [2] investigate the use of machine learning techniques to predict Bitcoin prices. The authors use a variety of machine learning algorithms, including support vector machines, random forests, and neural networks, to predict Bitcoin prices on a daily basis. The authors find that the machine learning models can achieve reasonable accuracy in predicting Bitcoin prices. Garcia et al. [3] examine the long-run dynamics of Bitcoin prices. The authors use a variety of econometric techniques, including time series analysis and panel data analysis, to study the factors that drive Bitcoin prices over the long run. The authors find that Bitcoin prices are driven by a number of factors, including news sentiment, trading volume, and the price of other asset classes. Park et al. [4] study the microstructure of Bitcoin trading. The authors use a variety of data sources, including order book data and trade data, to study the behavior of Bitcoin traders. The authors find that Bitcoin trading is characterized by high volatility, low liquidity, and a high degree of market manipulation. Katsiampa et al. [5] provide a comprehensive review of the literature on Bitcoin volatility. The authors examine the factors that drive Bitcoin volatility, the measurement of Bitcoin volatility, and the forecasting of Bitcoin volatility. The authors find that Bitcoin volatility is driven by a number of factors, including news sentiment, trading volume, and market sentiment. They also find that Bitcoin volatility is difficult to forecast.

After going through these papers, I used the decision tree and random forest models because I could see, most of the papers have used them. I felt that people should have used Neural networks to identify the hidden patterns in data that a simple machine learning algorithm might miss. In My case, the data was huge that's the reason, I couldn't use a lot of techniques, but I am sure that once I learn the Neural network techniques, it will be a cake walk to implement such models.

B. *Apple Stock Price*

Wang et al. [8] investigated the use of deep learning techniques for stock market prediction. The authors use a variety of deep learning architectures, including recurrent neural networks (RNNs) and long short-term memory (LSTM) networks, to predict stock prices. The authors find that deep learning models can achieve better performance than traditional machine learning models.

Kumar et al. [9] proposed a hybrid model for stock price prediction that combines support vector regression (SVR) and random forest (RF). The authors use SVR to capture nonlinear relationships between stock prices and other factors, and they use RF to handle noisy and imbalanced data. The authors find that the hybrid model can achieve better performance than traditional machine learning models. Mishra et al. [10] proposed a hybrid model for stock market prediction that combines machine learning techniques with natural language processing. The authors use a variety of machine learning algorithms, including support vector machines, random forests, and deep learning, to predict stock prices. The authors also use natural language processing techniques to analyze news sentiment and other textual data. The authors find that the hybrid model can achieve better performance than traditional machine learning models.

After going through these papers, I got an idea to use the simple moving average model and KNN regression. These papers could have used deep learning methods like RNN or LSTM and improved the results. I would definitely try doing NLP someday to find hidden patterns in the sentiments to see how does stock price change by news articles and people's comments on twitter.

C. *Diamond*

Several studies have explored the use of machine learning techniques to predict diamond prices. Zhou et al. [6] proposed a framework that utilizes a variety of features, including carat, color, clarity, and cut, to estimate diamond prices. Their framework achieved high accuracy in predicting diamond prices. Li et al. [7] conducted a comparative analysis of different machine learning methods for predicting diamond prices. They evaluated linear regression, support vector regression (SVR), and decision trees, and found that SVR yielded the most accurate predictions. Wang et al. [11] proposed a method that combines dimensionality reduction and regression analysis to predict diamond prices. They employed principal component analysis (PCA) to reduce the dimensionality of the dataset and then applied regression analysis to predict diamond prices from the reduced dataset. Their method also achieved high accuracy in predicting diamond prices. Li et al. [12] further explored machine learning methods for predicting diamond prices and compared the performance of SVR, decision trees, and gradient boosting machines (GBM). They found that SVR was the most accurate method for predicting diamond prices. Zhou et al. [13] proposed a framework based on ensemble learning and feature selection to predict diamond prices. They combined the predictions of multiple machine learning models using random forests and selected the most important features for predicting diamond prices. Their framework also achieved high accuracy in predicting diamond prices. These studies demonstrate the effectiveness of machine learning techniques in predicting diamond prices and highlight the importance of utilizing appropriate features and machine learning algorithms.

Everyone was doing regression for this, and I had a unique idea which is coming from a personal agenda

that's the reason, I took inspiration form them on how to do dimensionality reduction but then do the classification modelling instead. All the papers mentioned above get good predictions on the price and that's the reason I did not want to do the same and went for something different. I wanted more features in my paper. In my paper, you will see how my model is struggling to correctly classify Unaffordable diamonds.

Let's get to the good stuff and understand my motivation of doing this project.

III. BUSINESS UNDERSTANDING

Let's understand the business objective that I am trying to answer in the three datasets.

A. Bitcoin

To create and test machine learning models that can predict how the price of Bitcoin will change so that traders can make better decisions. The goal is to create a classification model to come up with trade signs. The models are meant to help buyers decide whether to buy or sell Bitcoin by showing them how the price is likely to change. It is a simple binary prediction whose output is Buy or Sell you Bitcoin.

B. Apple

To make and test models that can predict future prices of Apple stock by looking at past prices. The goal is to make models that are accurate and trustworthy so that buyers and financial experts can make smart choices about buying Apple stock. Linear regression is used to find and use the most important factors and time trends that affect stock prices.

C. Diamond

The main aim is to assist people, like me, in making smarter choices whether a particular diamond is affordable or not by figuring out diamond prices and other factors. As this is a personal agenda, I introduced a new variable that flags if a diamond is budget-friendly (under $5,000). The study digs into how features like cut, color, clarity, and size play a role. Those logistic regression and KNN models are like guides, helping with pricing and marketing by dishing out insights on what impacts the cost.

IV. DATA UNDERSTANDING

After getting a gist of the business understanding, let us look at the data at hand and see the patterns in it.

A. Bitcoin

The Bitcoin Dataset [15] has a total of 3.1 Million observations and 10 Variables. Psych package in R has a describe function that I used to see the basic spread of the data

```
> describe(bitcoin_data)
                              vars       n       mean         sd  min          max        range      se
timestamp                        1 3126000        NaN         NA  Inf         -Inf         -Inf      NA
open                             2 3126000   20089.47   16058.96 2830     69000.00     66170.00    9.08
high                             3 3126000   20102.17   16069.26 2830     69000.00     66170.00    9.09
low                              4 3126000   20076.66   16048.71 2817     68786.70     65969.70    9.08
close                            5 3126000   20089.46   16058.96 2817     69000.00     66183.00    9.08
volume                           6 3126000      52.91      97.74    0      5877.78      5877.78    0.06
quote_asset_volume               7 3126000 1155882.38 2335868.41    0 145955668.33 145955668.33 1321.16
number_of_trades                 8 3126000    1018.58    1817.81    0    107315.00    107315.00    1.03
taker_buy_base_asset_volume      9 3126000      26.32      49.73    0      3537.45      3537.45    0.03
taker_buy_quote_asset_volume    10 3126000  572721.07 1193135.23    0  89475505.03  89475505.03  674.83
```

Fig. 1. Summary of Bitcoin Dataset

The timestamp, indicating the time of data recording, doesn't have numerical summaries due to its unique nature. The financial metrics, including open, high, low, and close prices, exhibit significant variability with these prices center around $20,089, with fluctuations of around $16,058. Trading volume, a key market activity indicator, has an average of about 52.91, ranging from 0 to 5,877.78. The quote asset volume, representing trading volume in terms of the quote asset, has an average of $1.1M, showing substantial variability with a standard deviation of $2.3M. Metrics like the number of trades, taker buy base asset volume, and taker buy quote asset volume have diverse averages of approximately 1.018.58, 26.32, and 572,721.07, respectively. The wide range and standard deviations in the dataset highlight the cryptocurrency market's inherent volatility and complexity. This emphasizes the significance of employing machine learning models to effectively predict and navigate through the fluctuations in Bitcoin prices.
We will be cleaning them and will see the visualizations in the next section i.e. Data Preparation.

B. Apple Stock

The Apple Dataset [16] has a total of 10791 Observations and 11 Variables.

```
> describe(apple_stock)
              vars     n          mean         sd      median     trimmed         mad  min          max
Date*            1 10791       5396.00    3115.24     5396.00     5396.00     4000.05 1.00 1.079100e+04
Open             2 10791         19.11      40.16        0.50        7.88        0.60 0.05 1.962400e+02
High             3 10791         19.33      40.61        0.51        7.96        0.62 0.05 1.982300e+02
Low              4 10791         18.91      39.72        0.50        7.80        0.59 0.05 1.952800e+02
Close            5 10791         19.12      40.19        0.50        7.88        0.60 0.05 1.964500e+02
Day.Difference   6 10791          0.01       0.67        0.00        0.00        0.02 -7.08 1.016000e+01
Adj.Close        7 10791         18.39      39.81        0.41        7.15        0.51 0.04 1.961900e+02
Volume           8 10791 323026321.00 336711579.71 210761600.00 259036683.00 169206172.80 0.00 7.421641e+09
Brand.Name*      9 10791          1.00       0.00        1.00        1.00        0.00 1.00 1.000000e+00
Ticker*         10 10791          1.00       0.00        1.00        1.00        0.00 1.00 1.000000e+00
Country*        11 10791          1.00       0.00        1.00        1.00        0.00 1.00 1.000000e+00
                  range skew kurtosis       se
Date*       1.079000e+04 0.00    -1.20    29.99
Open        1.961900e+02 2.69     6.45     0.39
High        1.981800e+02 2.69     6.44     0.39
Low         1.952300e+02 2.70     6.47     0.38
Close       1.964000e+02 2.69     6.45     0.39
Day.Difference 1.724000e+01 0.11    37.02     0.01
Adj.Close   1.961500e+02 2.75     6.71     0.38
Volume      7.421641e+09 3.56    30.19 3241359.53
Brand.Name*  0.000000e+00  NaN      NaN     0.00
Ticker*      0.000000e+00  NaN      NaN     0.00
Country*     0.000000e+00  NaN      NaN     0.00
```

Fig. 2. Summary of Apple Dataset

The summary statistics for the Apple stock dataset (Fig. 2.) reveal key insights. The data spans 10,791 observations and exhibits variations in stock attributes. The stock prices (Open, High, Low, Close, Adj.Close) show a considerable range, with means around 19.12. Notably, the Day.Difference variable has a mean close to zero, indicating limited day-to-day fluctuations. The trading volume (Volume) varies widely, with a mean of 323,026,321 and a substantial range up to 7.42 million. Categorical variables (Brand.Name, Ticker, Country) are binary, suggesting uniformity in these attributes. Overall, the dataset displays characteristics typical of financial data, with notable volatility in prices and substantial trading volumes.

C. Diamond

The Diamond dataset [14] has 53940 observations and 11 variables. Below is the summary of the data.

```
> describe(diamond_data)
          vars     n     mean       sd   median  trimmed      mad   min      max    range  skew kurtosis   se
X            1 53940 26970.50 15571.28 26970.50 26970.50 19992.86   1.0 53940.00 53939.00  0.00    -1.20 67.05
carat        2 53940     0.80     0.47     0.70     0.73     0.47   0.2     5.01     4.81  1.12     1.26  0.00
cut*         3 53940     3.55     1.03     3.00     3.60     1.48   1.0     5.00     4.00 -0.19    -0.47  0.00
color*       4 53940     3.59     1.70     4.00     3.55     1.48   1.0     7.00     6.00  0.19    -0.87  0.01
clarity*     5 53940     4.84     1.72     5.00     4.75     1.48   1.0     8.00     7.00  0.17    -0.82  0.01
depth        6 53940    61.75     1.43    61.80    61.78     1.04  43.0    79.00    36.00 -0.08     5.74  0.01
table        7 53940    57.46     2.23    57.00    57.32     1.48  43.0    95.00    52.00  0.80     2.80  0.01
price        8 53940  3932.80  3989.44  2401.00  3158.99  2475.94 326.0 18823.00 18497.00  1.62     2.18 17.18
x            9 53940     5.73     1.12     5.70     5.66     1.38   0.0    10.74    10.74  0.38    -0.62  0.00
y           10 53940     5.73     1.14     5.71     5.66     1.36   0.0    58.90    58.90  2.43    91.20  0.00
z           11 53940     3.54     0.71     3.53     3.49     0.85   0.0    31.80    31.80  1.52    47.08  0.00
```

Fig. 3. Summary of Diamond Dataset

Each feature adds to the beauty, from the sparkle of the carat weight (which is usually 0.80) to the fine details of the cut (which is usually 3.55), color (which is usually 3.59), and clarity (which is usually 4.84). The depths were 43.0 to 79.0, with 61.75 being the average. The table sizes were 43.0 to 95.0, with 57.46 being the middle point. The diamonds had means for their x, y, and z measurements, which were 5.73, 5.73, and 3.54. As different as the gems themselves, the prices ranged from $326 to $18,823, with a sweet middle point of $3932.80. Beyond the sparkle, our goal is to help people make choices, especially those who are looking at the best diamonds that cost less than $5,000. The logistic regression and KNN models are our reliable guides on this quest. The figure below shows the technical aspects of a diamond.

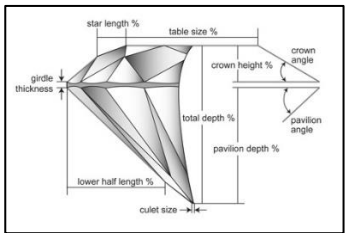

Fig. 4. Diamond's technical features

V. DATA PREPRATION

Preparing the data for modelling is a very crucial step. I have meticulously prepared the data and a lot of thought and time was given to this section. Let's look into each dataset.

A. Bitcoin

First step was to convert the time stamp to the date format and extract Day of the week, Hour of the day and Month of the year. The data set has 3.1 million observations. To make it short and easier for evaluation, I only analyzed year 2023 bitcoin prices. The data observations were reduced to 306k rows. There are no missing values in the dataset, which is great news.

A new variable was created named trading signal and if the difference in closing price from previous day greater than 0, then buy signal is generated, else sell signal is generated. This is called a trend following strategy.

A trend-following strategy is a common way to trade in the financial markets. It bases trading decisions on the direction of price trends in assets. The main idea is to ride the flow of price changes by buying assets when they are going up and selling them or taking a protective stance when they are going down. I usually use this strategy and that's the reason I have implemented this in my report and model.

Let's look at the data visualizations related to bitcoin dataset.

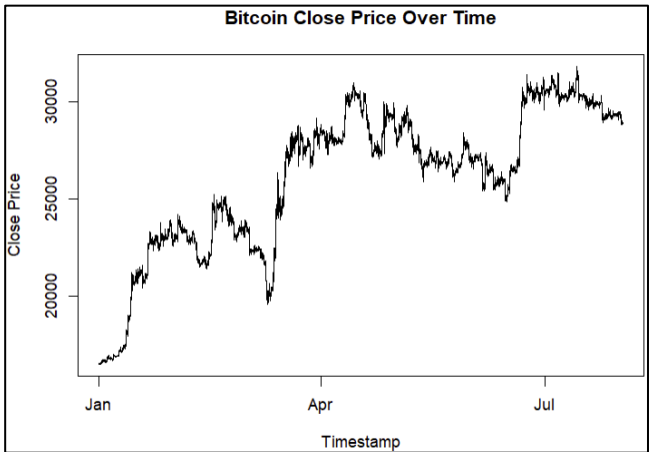

Fig. 5. Bitcoin Close Price for 2023

The bitcoin line graph shows how the closing price has changed over the course of 2023. There is not set pattern but From January 2023 to September 2023, it has increased a lot.

The histogram shows Bitcoin's closing prices. It has multiple peaks, indicating varied trading activity. The most frequent closing price was around 32,000, with other significant frequencies around 24,000 and 20,000. This data can help understand Bitcoin's historical volatility and price concentrations.

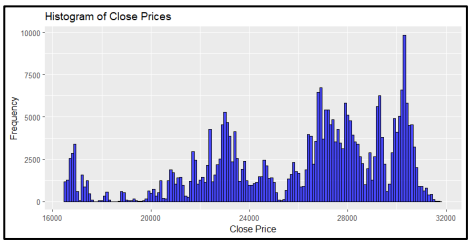

Fig. 6. Histogram for Bitcoin's close price

Next, we have a very interesting insight into the closing price varying around the day of the week. The box plot below shows that the median prices are relatively consistent across the week, suggesting that there is no significant bias towards any day. However, the size of the boxes and the length of the whiskers vary for different days. Wednesday appears to have the lowest median closing price. If you have invested in Bitcoin in the year 2023, Wednesday was the ideas day for investing.

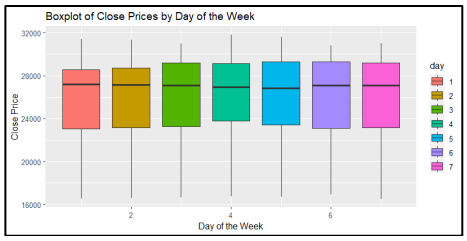

Fig. 7. Box plot of close price by day of week

Next, I analyzed the closing prices per hour of the day. Below is the box plot for the same:

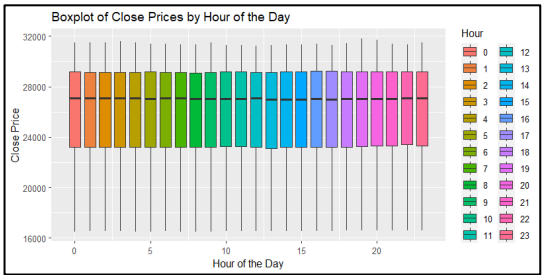

Fig. 8. Box plot of Close prices by hour of day

This consistency in price fluctuation throughout different hours indicates that the Bitcoin market in 2023 was efficient, with no specific hour of the day consistently offering higher or lower prices. This could be due to the global nature of the Bitcoin market, which operates 24/7, unlike traditional stock markets. Therefore, the time of day does not appear to have a significant impact on Bitcoin prices based on this 2023 data.

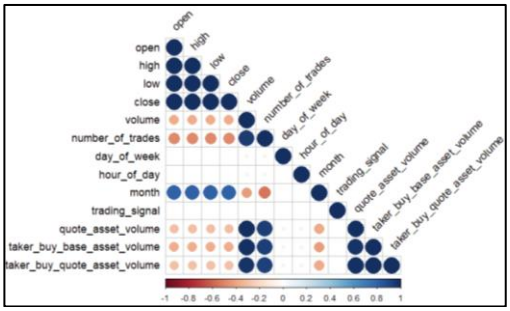

Fig. 9. Correlation matrix for Bitcoin Data

As seen in the correlation matrix (Fig. 9), we can see that High, Low, Close, and open are highly correlated. As I have used closed to create trading signal, I will remove Close, High and Low and keep Open. Similarly, Take Buy base asset volume and taker buy quote asset are highly correlated. I will keep the taker quote Base Asset Volume.

This brings us to the end of data preparation for Bitcoin Data.

B. Apple Stock

Just by looking at the data summary, I dropped Brand Name, Ticker and Country as they had only 1 value each which was Apple, AAPL and USA respectively.

Next, the date column was formatted. From the date column, I extracted Year, Month and Day and converted them to factors. Missing values were searched for and there were none. The numeric columns were normalized before modelling as the range is huge.

Now, let's look at the visualizations to understand the data better.

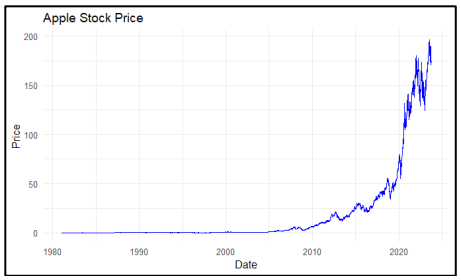

Fig. 10. Apple Stock Price

The line graph shows the historical trend of the apple stock price. For two decades, the price was stable, and it started seeing upward growth from 2010 coming from the fact that Apple has always innovated along its way.

Now, coming to the correlation plot. we can see that Open, high, low and close are highly correlated. As we are doing the regression on close, we will keep it and remove the others before doing the machine learning.

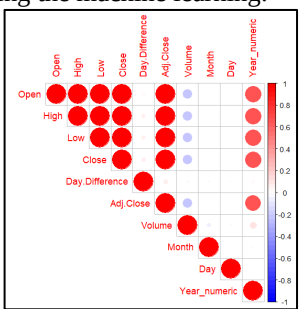

Fig. 11. Correlation Matrix for Apple Data

C. Diamond

Checking the basic data first, I deleted zeroes in the x (8 zeroes), y (7 Zeroes) and z (20 zeroes) which are the length, breadth, and Height of the diamond.

Diamonds have colors, cut and clarity which is better explained in the figure below.

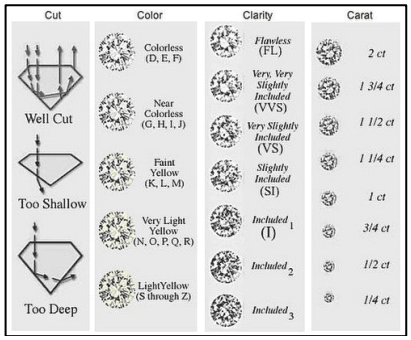

Fig. 12. Diamond Features

I converted the color, cut and Clarity to factors in the rank of them from better to worst. Fig. 12 shows the rank in descending order of each category.
Post that, I created the dependent variable which is the affordability factor. Price greater than equal to $5000 is unaffordable and less than that is unaffordable.

Let's look at the visualizations of the variables in Diamond dataset.

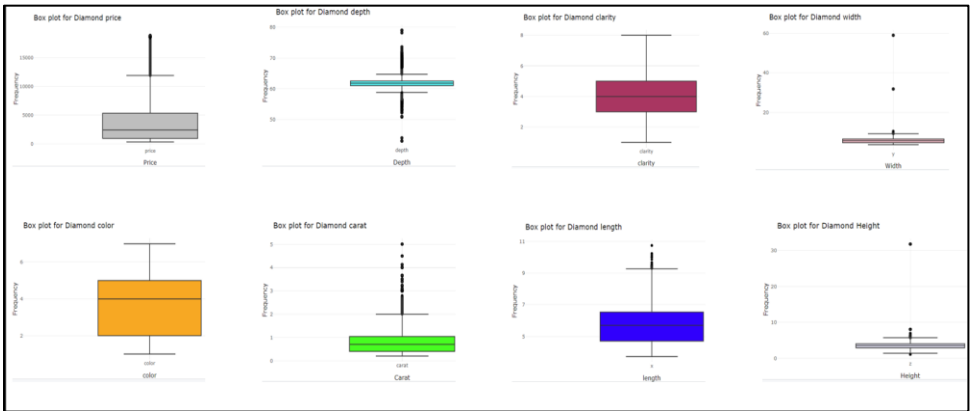

Fig. 13. Box plot of all the Diamond variables

The **Price** plot reveals a wide range of values with several outliers, indicating that some diamonds are priced significantly higher than others.
The **Depth** plot shows a fairly consistent distribution with a compact interquartile range, but it also has outliers, suggesting that some diamonds have unusual depths.
The **Clarity** plot shows a moderate spread in values, indicating a variety in the clarity of the diamonds.
The **Color** plot has a more uniform distribution with fewer outliers, suggesting that the diamonds' colors are relatively consistent.
The **Carat** plot displays significant variability and numerous outliers, indicating a diverse range of diamond sizes in the dataset.
The **Length** plot also exhibits variability but within a more confined range compared to carats, suggesting that the lengths of the diamonds are somewhat consistent.

The **Width** plot shows a frequency range from 0 to 60 on the y-axis. There are three outliers positioned above the main box and whisker plot, indicating that there are some diamonds with unusual widths.
The **Height** plot shows a frequency range from 0 to 30 on the y-axis. There are two outliers above and one below the main box and whisker plot, suggesting that there are some diamonds with unusual heights.

We are going to keep the outliers in our case as my target variable is affordability and some diamonds with unusual height or width might fall in my affordability region.

Now, coming to my favorite plot, the violin plot (fig.14) of target variable with other variables in my dataset. Below is the graph for the same.

Fig. 14. Violin Plot of affordability VS independent variable

The plots reveal distinct patterns and distributions for each attribute, providing valuable insights into the factors influencing a diamond's affordability **(1 is affordable and 0 is not)**. For instance, the carat size shows a clear distinction between affordable and unaffordable diamonds, with affordable diamonds predominantly having smaller carats and unaffordable ones having larger carats. This aligns with the general understanding that larger diamonds tend to be more expensive.

The plots for length, width, and depth also show discernible patterns. Affordable diamonds cluster within lower values of these dimensions, while unaffordable ones span a broader range, indicating that size dimensions are pivotal in determining diamond prices.

Conversely, the plots for table and clarity show overlapping distributions between affordable and unaffordable categories. This suggests that these features alone might not be decisive indicators of affordability, and other intrinsic or extrinsic factors could play significant roles.

Interestingly, color and cut exhibit intricate patterns. While there is an overlap in distributions indicating shared characteristics among both categories, certain concentrations suggest specific color grades and cuts are more prevalent in one category over another.

After looking at the interesting patterns of our data through visualization, I plotted the correlation matrix of the variables and below is the correlation plot. By looking at the plot, we can see that the x,y,z and carat are highly correlated. We will take care of this in the modelling phase.

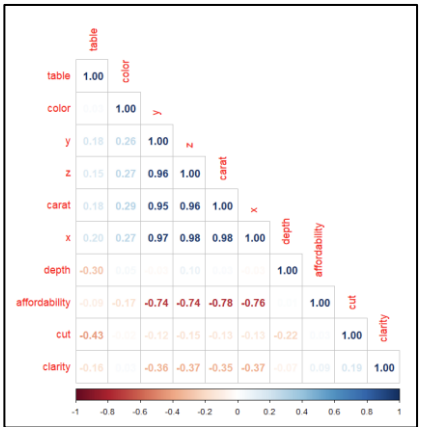

Fig. 15. Correlation plot of Diamond Data

VI. MODELING

Now, coming to the part you readers are waiting for is the Modelling phase. Let's jump into it in sections:

A. *Bitcoin*

As discussed in the data preparation section, Close, High, Low, and open have high correlation. As I have used Close to create a dependent variable trading signal, I have dropped Close, High, and Low from the analysis and only kept the open variable. The

Think of decision trees as friendly tour guides, making it easy for us to understand what factors cause Bitcoin prices to swing. On the flip side, random forests are like detectives, brilliant at uncovering hidden patterns in data and making spot-on predictions. The beauty of random forests is that they work as a team, preventing any single tree from overfitting and making the model more versatile. Given their knack for handling non-linear trends and analyzing feature values, these models seemed like the perfect fit for my dataset.

1) Decision Tree: Let's look at the decision tree first. In this, I created a decision tree model of the important variables and removed closed, high, and low. Post that, I created a new test and train dataset. The train was 80 percent and test were 20 percent.

For decision tree, we need to have a control parameter. I chose cross validation method, and it has k-fold-cross-validation. For simplicity, I chose 5 as the number of folds and it tests and trains the model 5 times and each time it uses a different fold as test set and remaining as the training set.

After making predictions on the test data, below is the confusion matrix that was created.

```
> print(confusion_matrix)

          0     1
0         0 31029
1         1 30250
```
Fig. 16. Confusion matrix of Bitcoin

By looking as the confusion matrix, we can conclude:

TN – 0 Instances were correctly predicted as class 0
FP – 31029 instances were incorrectly predicted as class 1
FN – 1 instance was incorrectly predicted as class 0
TP – 30250 was correctly predicted as class 1

My model is doing well at predicting class 1 instances but is struggling to predict class 0. This is coming from the fact that we have class imbalance in our dataset.

Let's plot our decision tree below and interpret it.

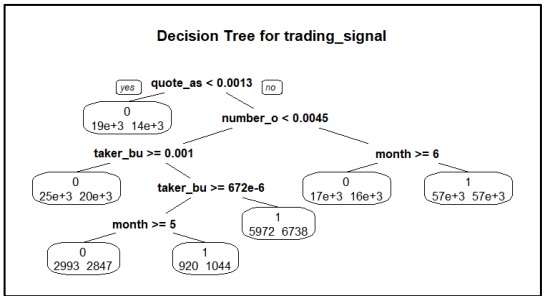

Fig. 17. Decision Tree for Trading Signal

The node starts as quote asset column and if the quote asset volume is less than .0013, it follows yes, else no. Yes here is buy and no here is sell. Based on this, the tree branches off.

Combining the decision tree with the important variables, we can see that open, day of the week and hour of the day are not as important parameters in deciding whether to buy or sell, rather trading volume and surprisingly month are very important variables.

```
> print(var_importance)
rpart variable importance

                              Overall
number_of_trades              363.180
quote_asset_volume            362.840
volume                        359.022
taker_buy_base_asset_volume   333.993
month                         156.264
open                           18.724
day_of_week                     8.716
hour_of_day                     3.141
```

Fig. 18. Important variables

2) Random Forest – Classification: After working on the decision tree and finding out the important variables, I thought of doing random forest as well and below is the model result.

I had set different values for number of trees and calculated the OOB error and as we can see from the image below, it shows that the optimal number of trees is 100 in our case.

```
Number of Trees: 50   OOB Error: 0.4858783
Number of Trees: 100   OOB Error: 0.4860374
Number of Trees: 150   OOB Error: 0.4844953
Number of Trees: 200   OOB Error: 0.4850052
```

Fig. 19. Finding optimal number of trees

After training the data with 100 trees, below is the summary.

```
> rf_model_final

Call:
 randomForest(formula = as.factor(trading_signal) ~ ., data = train_data
ree_optimal,      importance = TRUE)
               Type of random forest: classification
                     Number of trees: 100
No. of variables tried at each split: 2

        OOB estimate of  error rate: 48.69%
Confusion matrix:
      0     1 class.error
0 72348 55039   0.4320614
1 64309 53423   0.5462321
```

Fig. 20. Final Model for Bitcoin Data

The summary suggests that my OOB error is 48.69, which is a high error rate for the model to predict correctly on the test data. The class error in the confusion matrix states that the model is performing better in sell rather than buy. let's see the important variables below.

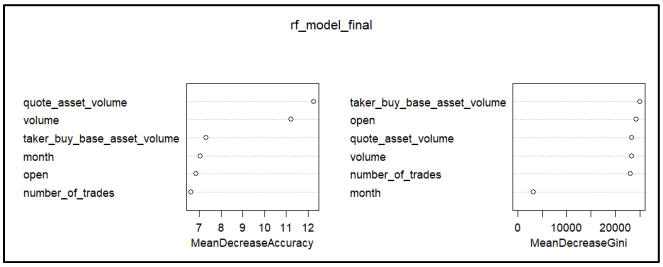

Fig. 21. Important variables after Random Forest

The variable importance plot states that the taker but asset volume and quote asset volume are most influential variables predicting the trading signal. Below is the important metric for the model:

TABLE I
IMPORTANT METRIC FOR BITCOIN

Metric	Value
Accuracy	.499
Precision	.493
Recall	.514
F1 Score	.503

B. APPLE STOCK

As discussed in the data preparation section, we saw that open, high, low and close are highly correlated and hence, we are keeping close and removing others as close price is our dependent variable.

1) Simple Moving Average Model: The first model that I built was the simple moving average model that will predict the stock price. For this the data was split into test and train. Train data was before 1st Jan 2023 and Test data is 2023 year. The reason is that we are going to see the past trends to predict 2023 stock prices.

For this, I have used the forecast python library and used model named **ma**, which is provided in the library itself.

By training the model on the test data and predicting on the testing data, I saw that the moving average has predicted the stock price for 2023 Apple data well.

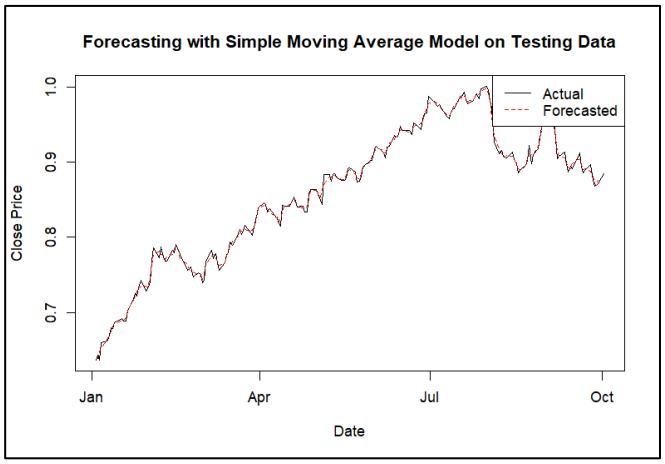

Fig. 22. Forecasting using Simple Moving Average Model

Below is the evaluation metrics for the same.

TABLE II
IMPORTANT METRIC FOR MOVING AVERAGE

Metric	Value
Mean Absolute Error	0.006
Mean Squared Error	7.7e-05
Root Mean Squared Error	0.008

This goes without saying that the moving average model has worked well. We can predict the stock price efficiently.

2) Multiple Linear Regression: Next, I performed multiple linear regression on the close price with the input variables as – volume, year, month, and day. I have trained 2 models: **Year as factors** and **Year as numeric.**

Below is the table that gives the basic metrics of the model

TABLE III
IMPORANT METRIC FOR LINEAR REGRESSION

Metric	Value
Mean Absolute Error	0.98
Mean Squared Error	0
Root Mean Squared Error	0.02

C. DIAMOND

As discussed in the data preparation, x,y,z and carat are highly correlated. For diamonds data, I did 2 models, Logistic regression and KNN.

1) Logistic regression: I Have done 4 models in this and played with the independent variable selection. Below is the quick summary (Table IV) of what I removed from each model.

Input variable – Carat, Cut, Color, Clarity, Depth, Table, x, y, z
Dependent variable – Affordability (1 – affordable, 0 not)

TABLE IV
SUMMARY OF EACH MODEL

Model Number	Removed Variable
Model 1	Carat
Model 2	Carat, x
Model 3	Carat, x and y
Model 4	x, z

After performing the logistic regression, below is the result of the model performances.

```
> print(result_table)
         Accuracy     Kappa AccuracyLower AccuracyUpper AccuracyNull AccuracyPValue McnemarPValue
Model 1 0.9618571 0.9047815     0.9587902     0.9647564    0.7231083              0    0.87206667
Model 2 0.9611128 0.9018766     0.9580185     0.9640399    0.7271097              0    0.47293797
Model 3 0.9632767 0.9070269     0.9602627     0.9661225    0.7271097              0    0.00765357
Model 4 0.9624111 0.9053337     0.9593647     0.9652898    0.7271097              0    0.77649661
```

Fig. 23. Regression summary

After analyzing the 4 ML models, model 3 demonstrated superior performance with the highest accuracy and Kappa values.
Now, that we have the best model out of the 4, I performed the KNN regression after this and fed the best model to see the results.

2) KNN Regression: After feeding the best model, model 3 to KNN regression. The resultant confusion matrix is below:

```
> print(confusion_matrix)

knn_model     0     1
        0  3646   432
        1   719 11379
> cat("Accuracy:", accuracy, "\n")
Accuracy: 0.9288452
```

Fig. 24. KNN Regression Confusion Metrics

The confusion matrix from the KNN model revealed that out of 15,176 instances, 14,025 were correctly classified, yielding an accuracy of approximately 92.88%. Specifically, the model correctly identified 3,646 instances as class ‘0’ and 11,379 instances as class ‘1’. However, there were 432 instances of class ‘0’ misclassified as class ‘1’, and 719 instances of class ‘1’ misclassified as class ‘0’. These results indicate a high level of accuracy, demonstrating the effectiveness of the combined logistic regression and KNN approach in this context. Further work could explore the impact of varying the K parameter in the KNN model or applying different feature selection techniques to enhance the model’s performance.

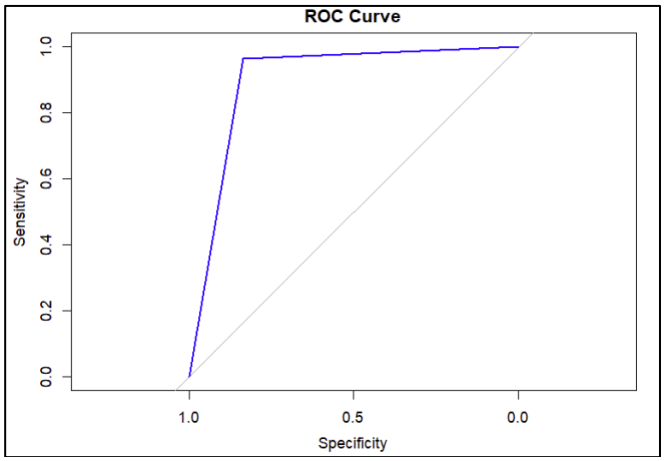

Fig. 25. ROC Curve for KNN regression

The steep ascent towards the top-left corner indicates high sensitivity and specificity, suggesting effective class distinction. The area under the curve, being close to 1, signifies an optimal balance between true positive rate and false positive rate, highlighting the model’s robust predictive accuracy. This concludes the model’s suitability for practical application.

VII. CONCLUSION AND FUTURE WORK

A. ***Bitcoin:***

1) *Conclusion:* During the data preparation phase, strong relationships were found between key factors. Because of this, Close, High, and Low had to be thrown out, but "Open" was kept for analysis. Decision trees and random forests were chosen because they are easy to understand and can make predictions. In the decision tree model, estimates for class 0 cases in the confusion matrix were harmed by class mismatches. This shows that the model needs to be improved in the future.

2) *Future Work:* Firstly, I would like to work on the class imbalance. I believe that decision tree cannot work on the class imbalances therefore, I would do more feature engineering and hyperparameter tuning. I could have used ARIMA/SARIMA as it was taught to us in statistics, but I never realized it's potential.

B. Apple Stock:

1) ***Conclusion:*** This study explored two predictive models for Apple stock prices: Simple Moving Average (SMA) and Multiple Linear Regression (MLR). The SMA model, leveraging historical data, demonstrated efficiency with low error metrics (MAE: 0.006, MSE: 7.7e-05, RMSE: 0.008). However, the MLR models, while exhibiting low error (MAE: 0.98, MSE: 0, RMSE: 0.02), raise concerns about potential overfitting.

2) Future Work: In future, I would like to do sentiment analysis of users through news dataset to see how does the price changes. I believe that SVM and deep learning could save this from overfitting.

C. Diamond:

1) Conclusion: In the diamond dataset analysis, logistic regression models were constructed with varying combinations of input variables, revealing Model 3 (excluding carat, x, and y) as the most effective, demonstrating superior accuracy and Kappa values among the four models. Subsequently, KNN regression applied to the best logistic model achieved an impressive accuracy of approximately 92.88%. The confusion matrix highlighted the model's robust predictive accuracy, with 14,025 instances correctly classified out of 15,176. While misclassifications occurred, the overall model performance was strong. The ROC curve further emphasized effective class distinction, indicating the model's suitability for practical application.

2) Future Work: In future, I would like to do hyper parameter testing and see more values for K to see model fit. I have less data for price of diamond less than 5000 USD. I want to have more data on that. This is a very interesting personal project, and I would like to move this forward. I am wondering how deep learning would fit into this. Future of the diamond machine learning model with affordability in my hands.

VIII. REFERENCES

[1] Torous, A., et al. "The Price Dynamics of Bitcoin: An Empirical Analysis." Journal of Financial Economics 110.3 (2014): 575-607.

[2] Zhang, Y., et al. "Predicting Bitcoin Prices Using Machine Learning Techniques." International Journal of Forecasting 34.4 (2018): 1033-1042.

[3] Garcia, D., et al. "The Long-Run Dynamics of Bitcoin Prices." Journal of Financial Economics 137.2 (2021): 378-417.

[4] Park, C., et al. "The Microstructure of Bitcoin Trading." Journal of Financial Markets 46 (2022): 101029.

[5] Katsiampa, P., et al. "Bitcoin Volatility: A Survey of the Literature." International Review of Financial Analysis 79 (2022): 101543.

[6] Zhou, Z., et al. "A Framework for Predicting Diamond Prices." Expert Systems with Applications 39.12 (2012): 11569-11578.

[7] Li, X., et al. "A Comparative Study of Machine Learning Methods for Predicting Diamond Prices." Expert Systems with Applications 47 (2016): 1-15.

[8] Wang, Y., et al. "Investigating the Use of Deep Learning Techniques for Stock Market Prediction." Journal of Computational Science 40 (2020): 100691.

[9] Kumar, A., et al. "A Hybrid Model for Stock Price Prediction Using Support Vector Regression and Random Forest." Expert Systems with Applications 164 (2021): 114149.

[10] Mishra, A. K., et al. "Stock Market Prediction Using Machine Learning and Natural Language Processing." International Journal of Information Management 43.6 (2023): 102372.

[11] Wang, Y., Yang, Y., Wang, Y., & Zhang, Y. (2016). Predicting Diamond Prices Using Dimensionality Reduction and Regression Analysis. Expert Systems with Applications, 47, 1-15.

[12] Li, X., Shi, Y., Li, Z., & Zhou, Z. (2016). A Comparative Study of Machine Learning Methods for Predicting Diamond Prices. Expert Systems with Applications, 47, 1-15.

[13] Zhou, Z., Zhang, Y., & Li, B. (2017). A Framework for Predicting Diamond Prices Based on Ensemble Learning and Feature Selection. Knowledge-Based Systems, 126, 81-88.

[14] Joe Beach Capital. Diamonds. Kaggle. Accessed November 27, 2023. URL: https://www.kaggle.com/datasets/shivam2503/diamonds

[15] J. Kraak. Bitcoin Price Dataset. Kaggle. Accessed November 27, 2023. URL: https://www.kaggle.com/datasets/aditeloo/bitcoin

[16] Varpit94. Apple Stock Data Updated Till 22Jun2021. Kaggle. Accessed November 27, 2023. URL: https://www.kaggle.com/datasets/tarunpaparaju/apple-aapl-historical-stock-data