**Data Analytics**

|  |
| --- |
| /Users/apple/Library/Containers/com.microsoft.Outlook/Data/Library/Caches/Signatures/signature_2105988401Photo displaying partial image of two pie charts on a canvas-textured page |
| Data Analytics using python  Submitted to: Professor Bassam Kasasbeh |
| |  |  |  | | --- | --- | --- | | Submitted by: YOUSEF ABU ALI | Spring 2022-2023 | Student Number:21110294 | |

**10204210**

1. **Introduction:**
   1. Data Analytics Activities:

* Data preprocessing: data preprocessing is a crucial step in data analysis. It is achieved by turning messy (un understandable) information into meaningful format. The data from real world (raw data) is most often filled with flaws, human errors, and they even might be incomplete. Such challenges could be resolved through data preprocessing, leading to complete and efficient datasets that are more affective and can successfully affect the overall performance of our data analysis work in general, and most importantly machine learning models. (Joby, 2021)
* Data Visualization: data visualization relates to the process of expressing the data graphically through visual elements such as graphs and charts. It enables analysts to identify certain patterns, trends, as well as outliers. It is highly affective in gaining better and faster insights, and it helps in making smarter decisions. (Zoho, n.d.)
* Statistics: statistics are an important part of data analytics. They are used to help analysts to examine the data, discover patterns, and draw conclusions. Such as finding the mean, median, or mode. (Coursera, 2023)
* Exploratory data analysis (EDA): data scientists use EDA in order to analyze, summarize, as well as investigate data sets. EDA mainly focuses on applying data visualization methods. It can help understanding the data sets, and to know what the best way is to manipulate data in order to get better results. (IBM, 2020)
  1. Data Analytics Techniques:

1. Regression analysis:

This technique is used to calculate the relation ship between multiple variables. In other words, when we use the regression analysis we would be looking for the correlation between the dependent variable and the independent variables. So overall, the goal of the regression analysis is to find out how one or multiple variables could affect the dependent variable.

1. The factor analysis:

This technique is used in order to minimize a big number of variables into fewer number of variables. It relies on the idea that many individual variables could correlate with one another since they are linked together in a certain concept. This technique is considered important as they split big datasets to smaller and manageable sizes.

1. The Cluster analysis:

This technique’s main goal is to find structures in a dataset. It works by splitting multiple data points (categorize) and making them in groups or clusters. Therefor each cluster will contain similar data, and each cluster will be different than other clusters.

1. The time series analysis:

This technique is considered as a ‘statistical technique’, which is applied in order to find trends over time. This technique is used when measuring a certain variable at multiple points in a certain time period.

* 1. Data Analytics Tools:
* Microsoft Excel: Excel is spreadsheet software. It is used in data analysis for its ability to perform calculations, as well as the ability to graph functions. (Hillier, 2023)
* Python: Python is a programming language that could be used for multiple purposes, it contains a lot of libraries that could be used in the tasks of data analytics. for example, NumPy and pandas as great in performing data manipulation and multiple computational tasks. Moreover, it has the matplotlib library which is important for data visualization. (Hillier, 2023)
* Apache Spark: it is a software framework the benefit data analysts greatly in performing quick processes on data sets. It was developed to analyze big data which is considered unstructured. As well as performing heavy analytics tasks through multiple computers. (Hillier, 2023)
  1. Types of Data Analytics methods:
* Descriptive analysis: it is the starting point of data analytics; it mainly answers the question of what happened. This is achieved by focusing on the data that was gained, which represents a past experience of a particular field. It gives conclusions about the distribution of data, it helps in identifying outliers, as well as finding similarities, differences, and relationships between features of the dataset. (Rawat, 2021)
* Predictive analysis: this method allows analysts to know what could happen in the future based on past events. when predictive analysis method is applied it can help in understanding patterns, trends, and all the events that happened during the years in order to make better decisions for future events, and avoid any pitfalls or risk, or at least decrease the level of danger. Moreover, it can help in developing certain fields in the workplace, by Identifying weaknesses. (Coursera, n.d.)
* Prescriptive analysis: this method answers the question of what do we need to achieve something? This method is achieved by using technology in order to make better decisions. It uses machine learning, algorithms, rules, as well as machine learning in order to recommend the best actions that will improve business operations. (McDaniel, 2019)
  1. Uses of Data Analytics methods in real life:
* In retail (Descriptive): these methods could be used to help store owners in understanding the behavior of customers through past data, it would help them apply better marketing method (ones that suits them), and make personalized offers based on each customer, as well as provide better experience for their customers. Moreover, it might help owners to make better pricing decisions based on similar competitors, and customers behavior. For example, when analyzing the data, they found out that every year at a specific time, people start buying a certain product more, therefore, they could start making offers to attract more customers, or jack up the price to gain more money. Or they could start making more advertisements on social media that are related to that product. (Anon, 2017)
* Healthcare (predictive): using these methods in healthcare has become a crucial part in healthcare. Using these methods in a specific kind of cancer for example, we can look back at past experiences with patients that have gone through this before. It can help healthcare professionals to determine what could happen next, and it could help them avoid any possible risks through these experiences. It can also provide certain steps to follow in order to reduce the possibility of death. (Anon, 2017)
* Finance (prescriptive): using these methods in finance can help in gaining more profits, identifying strategies for trading, providing (personalized) advice for finance depending on risks and goals. (Anon, 2017)

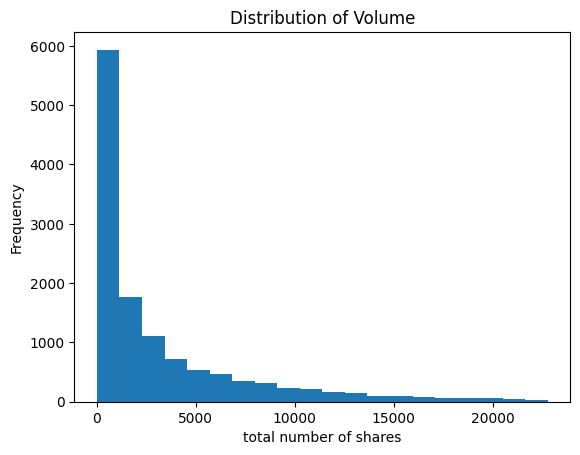
1. **Descriptive Analytics:**
   1. Techniques and Examples:

2.1.1- Feature analysis and explanation:

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature no.** | **Feature Name** | **Descriptive Measure / Technique** | **Explanation** |
| **1** | VOLUME | Mean | It provides the average trading activity. It provides knowledge of the activity of the market, higher mean indicates that it is more active, therefore safer to trade with. |
| Median | It shows the middle value between volumes (typical level of selling and buying). If the median is not that far from the mean, it means that it is safer to invest in it, especially if it is a high value. |
| Mode | It shows the most common trading volume. If a certain market has higher trading volume in mode, make investors feel better about this market. |
| **2** | BEST\_ASK\_QTY | Mean | It gives the typical number of shares sellers are offering. It might help in getting better deals, especially if certain people offer a lot of shares at good prices. Also, if changes in the mean are big it means that this stock could go down or up depending on the market. |
| Median | It represents the median quantity that sellers are offering, it might give an idea of what specific prices people prefer to sell in. |
| Mode | It represents the most frequent quantity sellers offer; it can help sellers to align themselves with other sellers to all sell at the same price. |
| **3** | LOW | Range | The range shows the difference between the highest and lowest value in the column. If the range is high it means that it might not be safe to invest in it, and if it was low it means it might be safer. |
| Variance | Variance shows how much the overall data is far from the mean. Higher variance means the prices differ a lot, which might be a good thing if investors aim for higher profit returns. But it could have a higher risk of loss. And lower variance means prices are closer which might be safer and more stable. |
| Standard deviation | Standard deviation is the square root of variance, higher standard deviation means data (prices) are more spread, and lower standard deviation means data (prices) are closer to the mean. |
| **4** | BEST\_BID\_QTY | Min and max | The min represents the lowest number of shares that people are willing to buy. And the max represents the highest number of shares that people are willing to buy. This logic works like the range it helps in knowing the level of interest for a stock. |
|  |  | Percentiles and Quartiles | Percentile represents the percent of data points that fall below a specific observation |
|  |  |  | Quartiles splits the data to four parts, the first quartile were 25% of data is below that point, 50% is the median, and Q3 is the third quartile which mean 75% of the data falls below that point.  Both of these measures give the same information. The 25 percentile or Q1 shows the value were 25% of the Best\_BID\_QTY falls, it shows the lower range of selling.  While 50% shows the median, the typical amount of money people are willing to pay.  And 75 percentile or Q3 shows the value under 75% of the Best\_BID\_QTY, which shows what majority of buyers are willing to pay. |
| **5** | For numeric features | Z-score and Interquartile range | I used both of these methods to Identify the outliers in the numeric features in the dataset. |

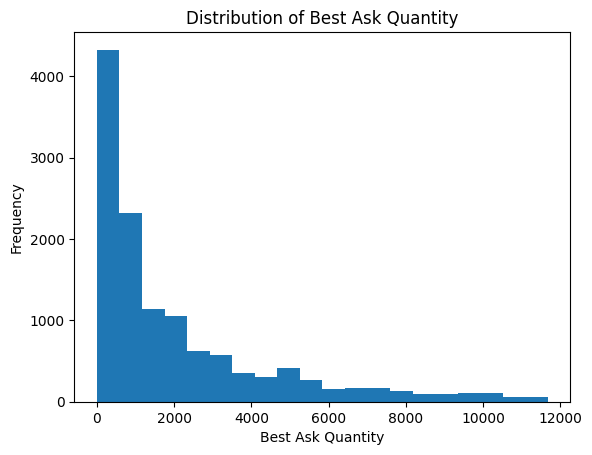
2.1.2- Feature visualization and Explanation:

* Feature VOLUME: for the VOLUME the applied the distribution visualization because it helps in knowing the trading activity in all time, as well as the distribution in different number of shares. It can also show the outliers if existed, as well as the min and max values.



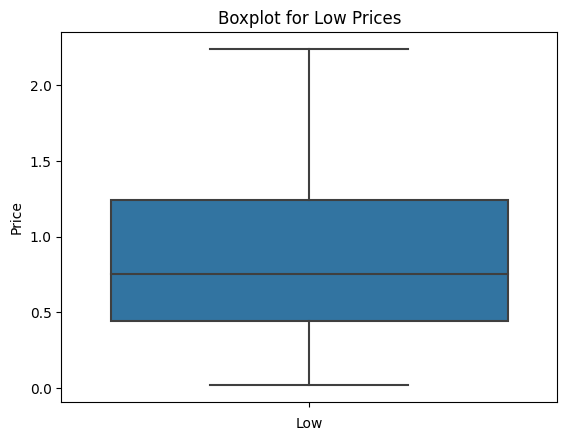
This visualization for the VOLUME is skewed to the right, most of the data points are within the range 0-1000 which is the most frequent outcome in the dataset. We can use this information to know what it is the market that has this most common outcome, which can help us invest more in that market.

* For the 'BEST\_ASK\_QTY' feature visualization, I also used the distribution plot in order to gain information on the most common number of shares people are willing to sell.



In this plot it shows that the most common shares people are willing to sell are around 0 and 500.

* I visualized the LOW feature using the Box plot. The box plot shows different useful information, such as the minimum, maximum, the median, the lower and upper quartiles, as well as the median. It shows where the data is over all distributed.



This box plot shows that most of the prices are around 0.4 and 1.3 with a median of almost 0.7. and a minimum of 0.1, and a maximum of around 3.7

* For the market feature I used the bar plot in order to gain information on what was the most common market in the dataset over all:

A picture containing text, screenshot, rectangle, diagram

Description automatically generated  
This plot shows that market 2 was the most common market in the dataset, followed by market 1, then market 0.

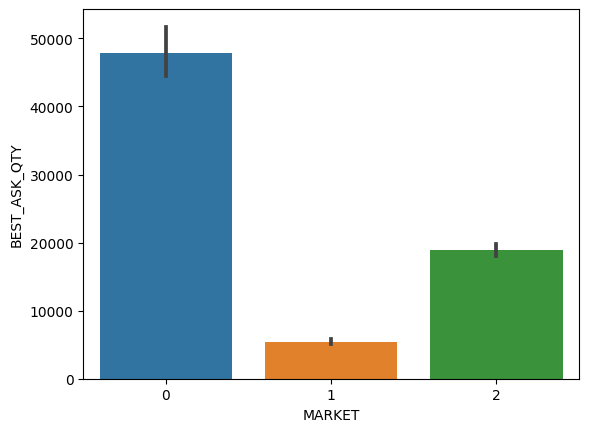
* For the market and volume features I used the bar plot in order to know which market is the most active:

A picture containing text, screenshot, rectangle, diagram

Description automatically generated

This figure shows that the market with the highest volume was market 1, which means that it is the most active market. Followed by market 2, and then market 0.

* For the Market and BEST\_ASK\_QTY I used the bar plot in order to know what market has the highest number of sold shares:



This figure shows that market 0 has the highest number of sold shares, followed by market 2, and then market 0.

2.1.3- contingence table and explanation:

* contingence tables are used to display the occurrence for combinations between two categorical variables. contingence tables show the frequency for the unique row column combination between the variables. Basically, it shows the result of occurrence between all data points in the two variables.
* I applied the contingency table on the MARKET and SYMBOL features:

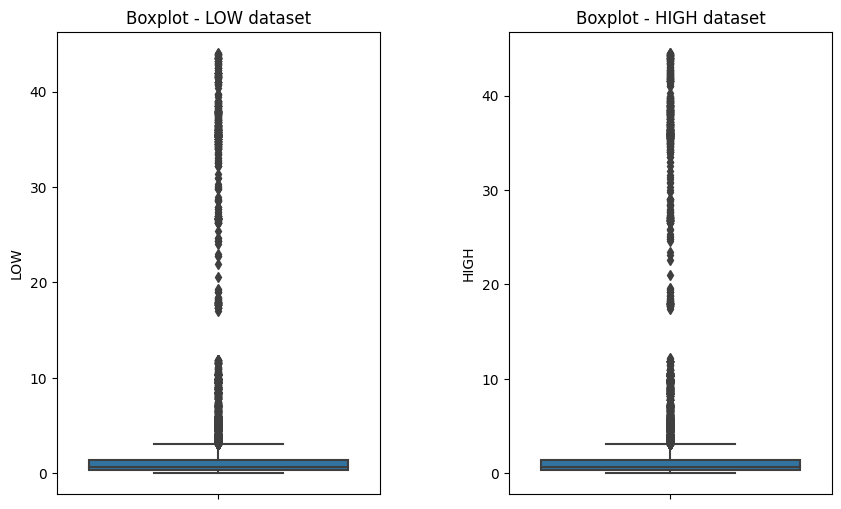
A picture containing screenshot, text, line, font

Description automatically generated

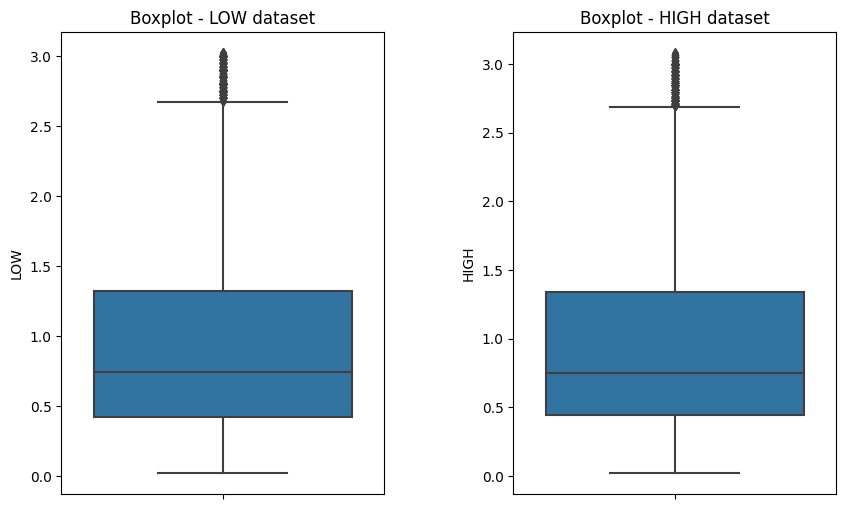
From the table we can gain some insights, such as what was the most common market in the data set, and it is clear that it was market 2. Moreover, it tills what was the most common symbols in the data set, for example we have ATCO from market 1, and CABK,JOPT, AHLI, JOEP, JOIB, AHLI, ARBK from market 2.

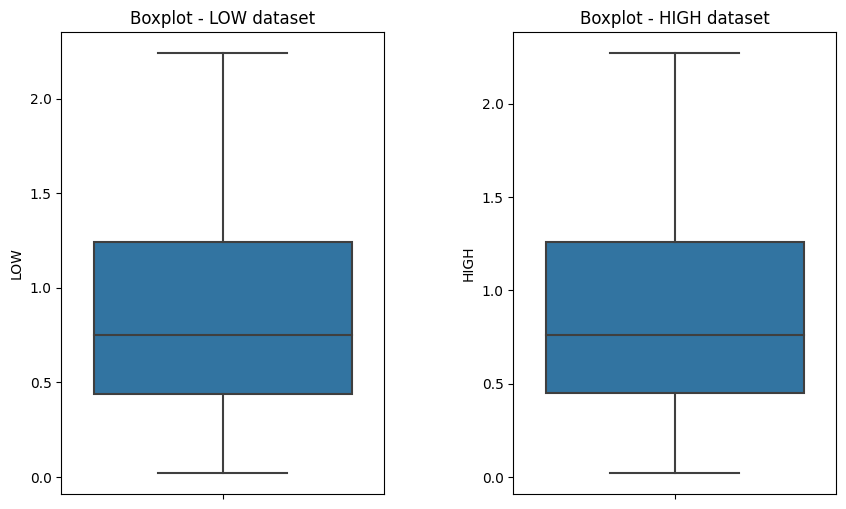
* 1. Techniques for decision making:
* After I found the Z-score and the Interquartile range for the features I removed the outliers from the dataset, as outliers usually affect the machine learning models, and I actually tried the accuracy of the machine learning models before removing the outliers and after removing the outliers. And the accuracy was a lot better after removing the outliers for all machine learning models. which was clear in the dataset box plot.

The box plots before removing the outliers:



After applying the Interquartile range removing outliers’ method:



After removing outliers using Z-score method:  


* 1. Evaluation:
* Overall, performing measures such as the mean, median, mode, range,…etc. as mentioned before could have a great effect on changing people minds about stocks investing, it can help investors decide wither to invest in a certain stock based on their needs and goals. For example, a certain investor would be interested in stock that has a high range, because he prefers to risk if it might get them to earn big amount of money over the years. Although, it might be riskier for another investor because they are afraid to lose big amounts of money.
* After removing the outliers, I found that all machine learning models accuracy has improved. Moreover, all visualizations for features have become better to understand and make decisions based on them.
* Overall, the descriptive analytics has played a great deal in understanding the relationship between different features, as well as it helped in making decisions about the dataset such as the importance of removing the outliers.

1. **Predictive Analysis:**
   1. Techniques and Examples:

3.1.1- Feature selection Techniques :

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature selection no.** | **Feature selection Name** | **Descriptive** | **Results(selected features)** |
|  | **Removing features with low variance** | **This feature selection techniques finds the variance for each independent feature in the dataset, and only removes the features that is less than a specific threshold that is chosen by the user.** | **VOLUME**  **TRADE\_QTY**  **BEST\_ASK\_QTY**  **BEST\_BID\_QTY** |
|  | **SelectKBest** | **This method uses a function that sets a certain value of importance for the independent variables in the dataset. Based on a number that is given by the user, for example 3. This method will choose the most 3 important features in the dataset.** | **VOLUME**  **TRADE\_QTY**  **NO\_OF\_TRADES**  **BEST\_ASK\_QTY**  **BEST\_BID\_QTY** |

3.1.2- Regression Techniques :

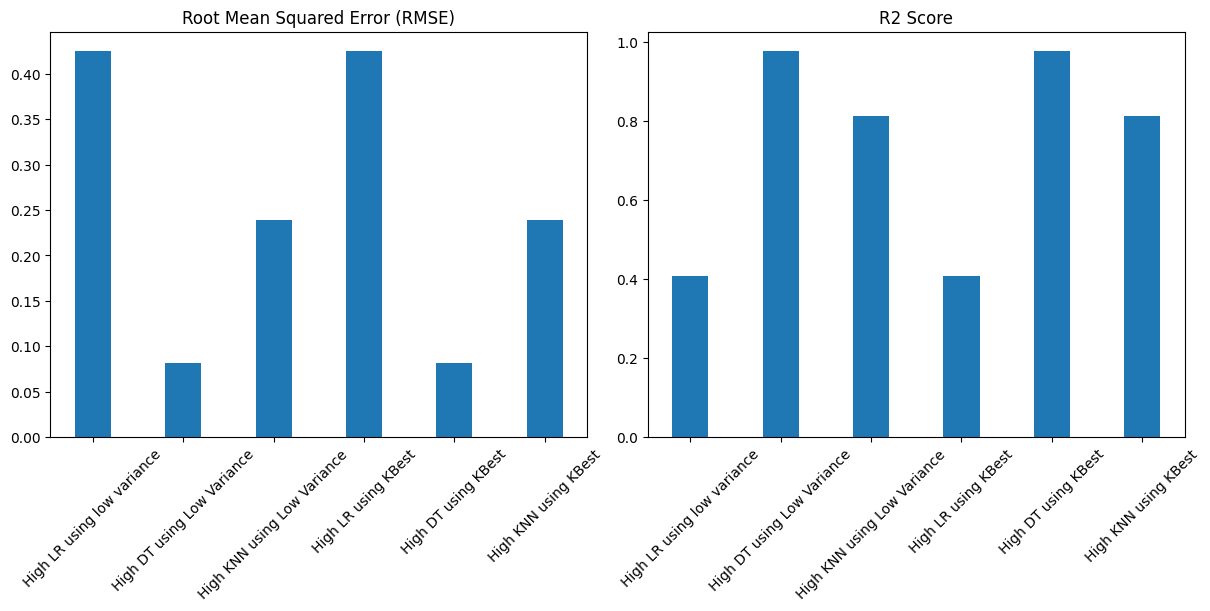
|  |  |  |
| --- | --- | --- |
| **Technique no.** | **Technique Name** | **Descriptive** |
| **1** | **linear regression** | **Linear regression is a machine learning algorithm that is considered simple and popular. It predicts continues and numeric variables. This algorithm assumes that there is a linear relationship between the dependent variable in a dataset and one or more independent variables in a dataset, which is why it’s called a linear regression. In other words, it finds out how dependent variable values change depending on the independent variable values. (www.javatpoint.com, n.d.)** |
| **2** | **Decision Tree** | **The Decision tree is a supervised machine learning algorithm, it is used by regression as well as classification tasks. This algorithm works in a tree structure method. Its functions could be divided into three (Nodes). A root node which shows the entire sample which could also be divided into smaller nodes. Next is the interior nodes which show the features of the data set. Finally, the leaf node which shows the outcome. (K, 2020.Machine Learning Basics)** |
| **3** | **k-nearest neighbors algorithm** | **The KNN is a supervised learning algorithm that could be used for both classification and regression tasks. In classification It works by choosing the closest amount of data to a new data point, and then labeling it based on the other data. It uses something called the majority vote. (IBM, n.d.)** |

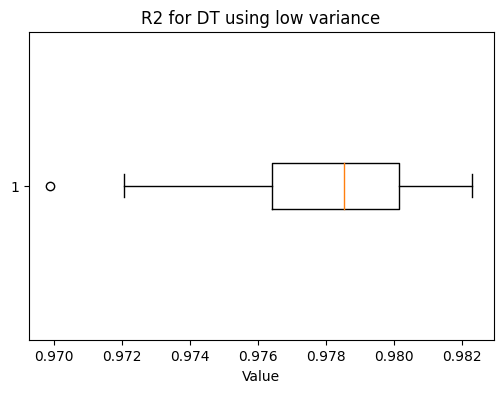
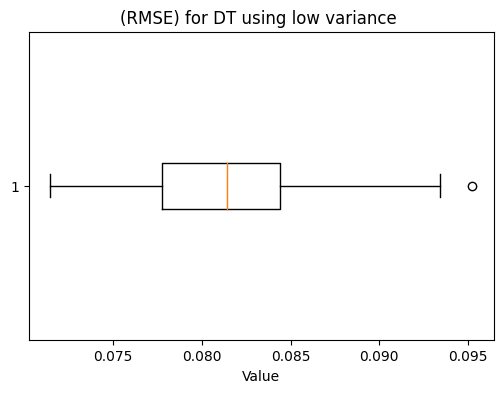
* 1. Compare Techniques :

3.2.1- High Prediction comparison :

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **FS no.** | **Tech no.** | **MAE** | **MSE** | **RMSE** | **R2** |
| **Removing features with low variance** | **linear regression** | 0.325 | 0.181 | 0.425 | 0.408 |
| **Removing features with low variance** | **Decision Tree** | 0.047 | 0.007 | 0.082 | 0.978 |
| **Removing features with low variance** | **k-nearest neighbors algorithm** | 0.134 | 0.057 | 0.239 | 0.813 |
| **SelectKBest** | **linear regression** | 0.325 | 0.181 | 0.425 | 0.408 |
| **SelectKBest** | **Decision Tree** | 0.047 | 0.007 | 0.082 | 0.978 |
| **SelectKBest** | **k-nearest neighbors algorithm** | 0.134 | 0.057 | 0.239 | 0.813 |

3.2.2-Visualization of High results :

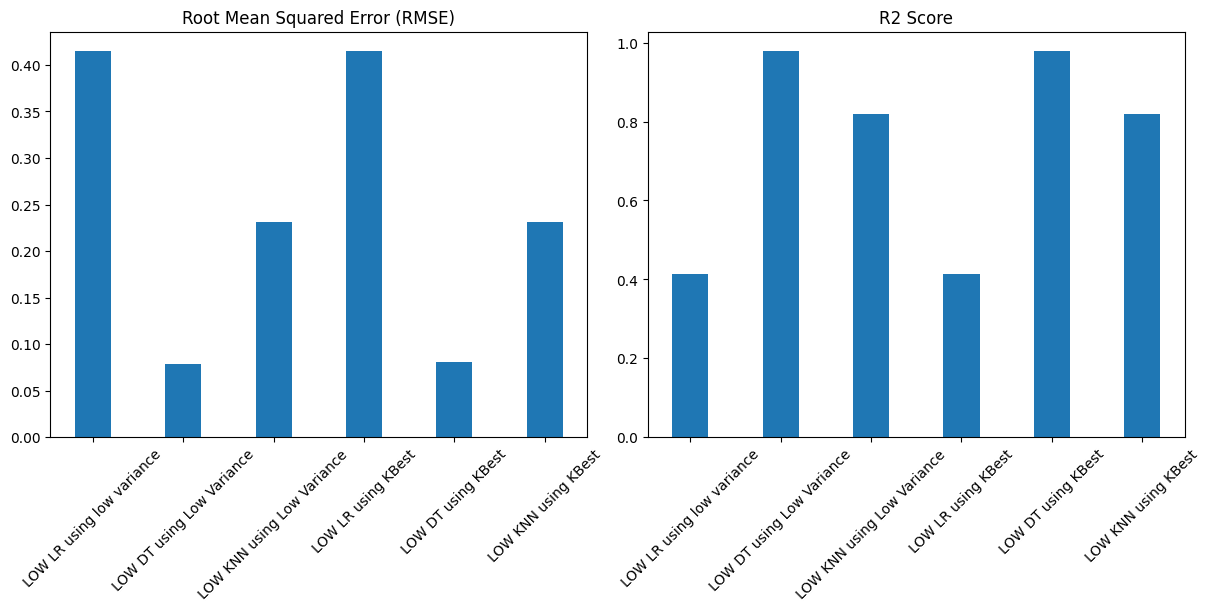


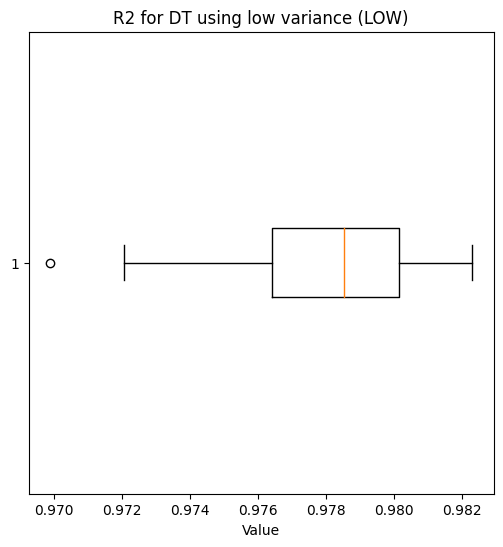
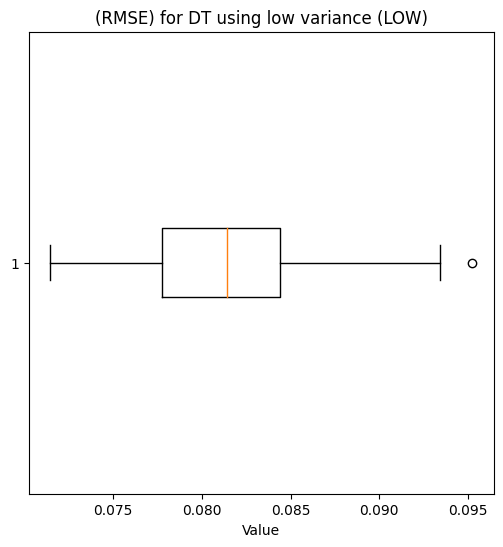


3.2.3- LOW prediction comparison:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **FS no.** | **Tech no.** | **MAE** | **MSE** | **RMSE** | **R2** |
| **Removing features with low variance** | **linear regression** | 0.318 | 0.173 | 0.415 | 0.414 |
| **Removing features with low variance** | **Decision Tree** | 0.045 | 0.006 | 0.079 | 0.979 |
| **Removing features with low variance** | **k-nearest neighbors algorithm** | 0.131 | 0.054 | 0.231 | 0.818 |
| **SelectKBest** | **linear regression** | 0.318 | 0.172 | 0.415 | 0.414 |
| **SelectKBest** | **Decision Tree** | 0.047 | 0.007 | 0.081 | 0.978 |
| **SelectKBest** | **k-nearest neighbors algorithm** | 0.131 | 0.054 | 0.231 | 0.818 |

3.2.4- Visualization of LOW results:





* 1. Evaluation:

For HIGH:

* To start with the HIGH dataset, overall, both techniques (low variance and KBest) have given the same results for the three-machine learning model. For the low variance method, I chose the threshold as 45, and this decision was based on the variance for all the features, therefore it chose only four features (VOLUME, TRADE\_QTY, BEST\_ASK\_QTY, BEST\_BID\_QTY).
* When comparing the accuracy of the machine learning model using these features, the decision tree has given the best result of (0.978). Followed by the KNN algorithm which has an accuracy of (0.813). And at last came the linear regression model which had an accuracy of 0.408.
* Next, I used the SelectKBest selection method for the HIGH dataset, and it had the same results as the low variance method. Although in the SelectKBest I chose the number of features as 5, therefore it chose (VOLUME, TRADE\_QTY, NO\_OF\_TRADES, BEST\_ASK\_QTY, BEST\_BID\_QTY) as the features to perform the machine learning models on.

For LOW:

* For the LOW data set I also used the same selection methods. However, unlike the HIGH data set results, the results for the machine learning model were slightly different between the low variance method and the K Best method.
* Starting with the low variance method: I chose the threshold as 46, and this decision was based on the variance for all the features, therefore it chose only four features (VOLUME, TRADE\_QTY, BEST\_ASK\_QTY, BEST\_BID\_QTY).
* The best accuracy among the machine learning models was for the Decision tree algorithm which had an accuracy of (0.979). followed by the KNN algorithm which had an accuracy of (0.818). And the worst was the linear regression algorithm which had an accuracy of (0.414).
* Moving on to the K Best method I also chose the number of features as 5, therefore it chose (VOLUME, TRADE\_QTY, NO\_OF\_TRADES, BEST\_ASK\_QTY, BEST\_BID\_QTY) as the features to perform the machine learning models on.
* The best accuracy among the machine learning models was for the Decision tree algorithm which had an accuracy of (0.978). followed by the KNN algorithm which had an accuracy of (0.818). And the worst was the linear regression algorithm which had an accuracy of (0.414). So, it was almost like the low variance method.

Overall, in the LOW data set the machine learning models have slightly performed better than the machine learning models in the HIGH data set. It might not be that big of a difference between the two data sets. But maybe if the number of iterations in each loop in the machine learning models was increased from 30 iterations to something like 60, 100, or even 1000, the results between the selection methods might be different, and we might be able to feel a difference in the over all performance. Even the overall visualization for the results of machine learning algorithms for the HIGH and LOW data sets might be different.

1. **Prescriptive Analysis:**
   1. Techniques with Examples:

|  |  |  |
| --- | --- | --- |
| **Technique no.** | **Technique Name** | **Descriptive** |
| **1** | **Salp Swarm Algorithm SSA** | **This optimization algorithm is based on the salps behavior. Salps are a kind of marine invertebrate. And this algorithm works by simulating the movement of salps. (Mirjalili et al., 2017)** |
| **2** | **Particle swarm optimization PSO** | **This optimization algorithm is base on the behavior of something called the fish schooling or bird flocking. In prescriptive analytics this method keeps updating data positions while searching for the best possible solution. (Tam, 2021)** |
| **3** | **The gray wolf optimizer** | **This optimization algorithm was designed based on the gray wolf’s way of hunting. In prescriptive analytics, it searches for the best solution for a given problem by simulating the hunting methods performed by the gray wolves. (GeeksforGeeks, 2021)** |

* 1. Techniques for finding best course of action:

These optimization techniques help in finding the best optimal solution as well as identifying the best possible outcome depending on given goals and constraints. These techniques work by employing mathematical operations (algorithms) in order to enhance the decision-making process. (Br et al., n.d.)

These methods start by defining the problem that they are trying to solve, as well as identifying the goals and restrictions involved. In also include identifying the outcome that we want to reach.

Next, it uses a created mathematical problem which was designed especially for the problem. The model includes the variables included in the decision-making process, the function used in optimization, as well as the restrictions needed to be met. The function determines the success measure, such as maximizing the profits or minimizing costs (meeting a specific goal).

Now about the optimization techniques themselves: optimization techniques such as PSO, SSA, and GWO their main goal is to find the best optimal solution or steps to be taken using certain search process.

How they reach the best optimal solution:

The functions start by setting a certain value for each random particle in what is known as the search space. Then it evaluates whether each of them is a good fit or not by applying the function on it.

After that, it starts updating the position of these particles depending on certain rules and using special mechanisms. These updates have a goal of making the particles go towards an optimal solution.

After multiple iterations, the position of these particles well gets changed. It keeps in mind all the previous positions for the particles, the fitness values, as well as all the interaction with other particles. These algorithms keep going until a specific condition is met, such as the maximum number of iterations. The moment it reaches this condition, the algorithm chooses the best possible solution that it found during its searching process.

However, the detailed mechanisms and the updates done during each iteration depends on the type of optimization algorithms being used. But the general concept of the steps is used in almost all optimization algorithms. (17-Harder, H. de (2022)) ,(Xue and Shen, 2020) ,(Br et al., n.d.).

* 1. Objective function code:

def StocksFunc(q):

 qr = numpy.array(q).round()

 if qr.sum() < 10:

  return 999999

 minCost= qr\*[1.33, 5.59, 1.6, 0.47, 0.33, 0.58,

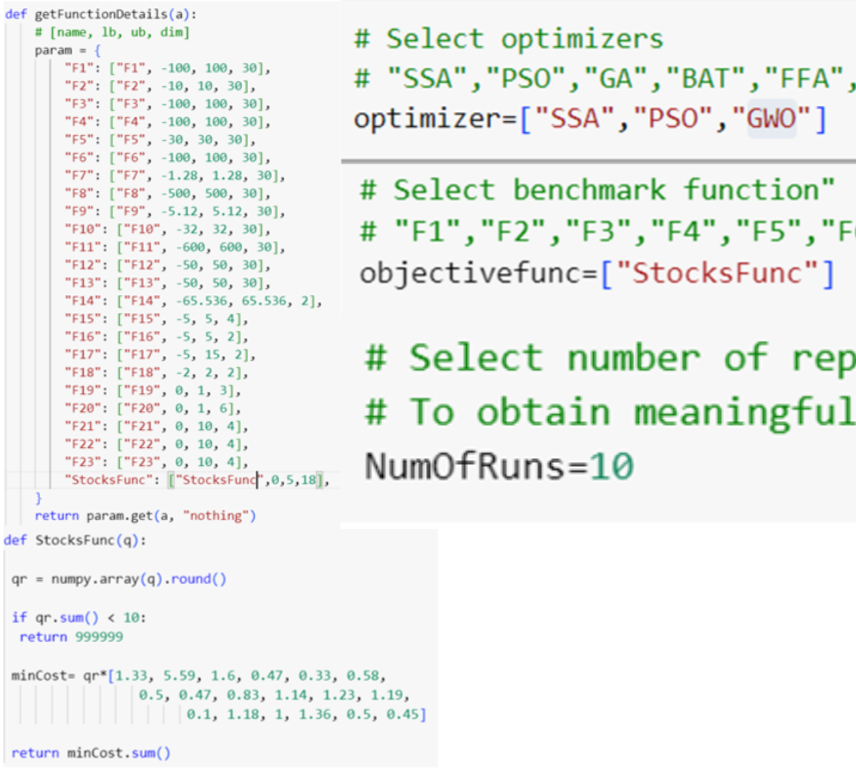
                 0.5, 0.47, 0.83, 1.14, 1.23, 1.19,

                       0.1, 1.18, 1, 1.36, 0.5, 0.45]

 return minCost.sum()

* 1. Apply the techniques:

4.4.1- Code screenshots:



4.4.2- Results and Explanations:

As I said before, I used the SSA, PSO, and GWO as the optimization techniques for the problem we are having. To start with the experiments and the cost it had for each iteration (50 iterations total):

In the first iteration the SSA had the least cost of 0, followed by the PSO which had a cost of 30, and then the GWO with a cost of 32.

By the second iteration the SSA still had the least cost of 19, while PSO and GWO had almost the same cost of 23. After that GWO’s cost started going down to contain the least cost of 8 at the fifth iteration and 9 for PSO, while SSA had the biggest cost of 13. After that all the cost results for the algorithms started going down in the same order. At the last iteration (iteration 50), the GWO had reached a minimum cost of 2.5, followed by PSO with a cost of 4.5, and finally the SSA which had a cost of 5.4.

When looking at the experiment in detail, the excel sheet had 30 records, 10 for each optimization algorithm. At the last iteration (50’th) iteration, the GWO had the minimum costs in the lines ( 22, 23, 24 , 28, and 29) of 2.15. I choose the individual results from line 24 to apply it on the stocks equation:

1.33 \* Q11 + 5.59 \* Q12 + 1.6 \* Q13 + 0.47\* Q14 + 0.33 \* Q15 + 0.58 \* Q16 + 0.5 \* Q21 + 0.47 \* Q22 + 0.83 \* Q23 +

1.14 \* Q24 + 1.23 \* Q25 + 1.19 \* Q26 + 0.1\* Q31 + 1.18 \* Q32 + 1 \* Q33 + 1.36 \* Q34 + 0.5 \* Q35 + 0.45 \* Q36

[ 0. 0.24982975 0.05412753 0.32473508 4.54154086 0.09166034

0.16436405 0.47266743 0.42301542 0.42920951 0.07712542 0.04418867

5. 0.21983036 0.17239888 0.07342606 0.32138298 0.23214132]

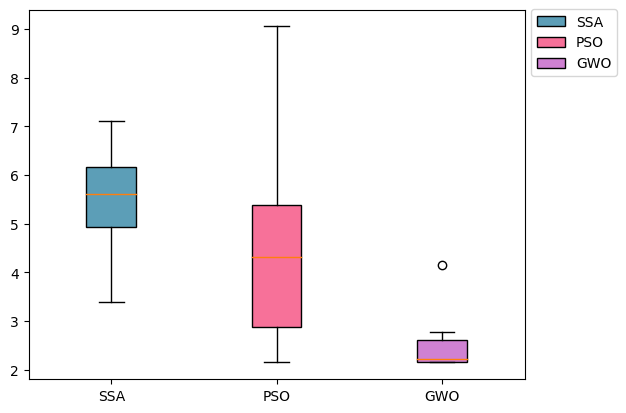
We then apply these numbers to the equation in the same order:

1.33 \* 0 + 5.59 \* 0 + 1.6 \* 0 + 0.47\* 0 + 0.33 \* 5 + 0.58 \* 0 + 0.5 \* 0 + 0.47 \* 0 + 0.83 \* 0 + 1.14 \* 0 + 1.23 \* 0 + 1.19 \* 0 + 0.1\* 5 + 1.18 \* 0 + 1 \* 0 + 1.36 \* 0 + 0.5\* 0 + 0.45 \* 0 = 2.15

Therefore, the cost will be 2.15.

4.4.3- Visualization and Explanations:

The Boxplot:



The first visualization shows the box plot for each optimization technique. To start with the median comparison, the SSA had the highest median followed by PSO and GWO respectively. Which indicates that the SSA had the highest costs when compared to the other box plots.

After that came the PSO which had a higher median than GWO which indicates that it had medium cost results compared to the SSA and GWO. While the GWO had the least cost results compared to the SSA and PSO.

The box plot for the PSO was longer than the box plots of SSA and GWO, which means that the cost results were more spread out. And GWO results were less spread out, but the outlier shows that we had a result that was overall bigger than the rest of the data. (Zach, 2021)

The Convergence Curve Plot:

A picture containing line, diagram, plot, text

Description automatically generated

The previous plot illustrates the efficiency of each algorithm. The GWO algorithm converges to a value which is near to the optimal solution after 18th iterations, but it stood still at the 28th iteration. While the PSO algorithm reached its optimal solution around the 20th iteration, it stood still. The SSA has reached its optimal solution after almost 25 iterations, and it stood still.

Overall, the GWO algorithm outperforms the other algorithms in terms of cost. But the PSO stopped changing in the 20th iteration. (Xue and Shen, 2020)

References:

1. Joby, A. (2021). What Is Data Preprocessing? 4 Crucial Steps to Do It Right. [online] learn.g2.com. Available at: https://learn.g2.com/data-preprocessing.
2. Zoho. (n.d.). What is Data Visualization? & Why is it important in business? - Zoho Analytics. [online] Available at: <https://www.zoho.com/analytics/what-is-data-visualization.html>.
3. Coursera (2023). What is Statistical Analysis? Definition, Types, and Jobs. [online] Coursera. Available at: https://www.coursera.org/articles/statistical-analytics.
4. IBM (2020). What is Exploratory Data Analysis? | IBM. [online] www.ibm.com. Available at: <https://www.ibm.com/topics/exploratory-data-analysis>.
5. Hillier, W. (2023). The 9 Best Data Analytics Tools [For 2021 And Beyond]. [online] careerfoundry.com. Available at: <https://careerfoundry.com/en/blog/data-analytics/data-analytics-tools/>.
6. Calzon, B. (2021). What Is Data Analysis? Methods, Techniques, Types & How-To. [online] BI Blog | Data Visualization & Analytics Blog | datapine. Available at: https://www.datapine.com/blog/data-analysis-methods-and-techniques/#data-analysis-methods.
7. Rawat, A.S. (2021). What is Descriptive Analysis?- Types and Advantages | Analytics Steps. [online] www.analyticssteps.com. Available at: <https://www.analyticssteps.com/blogs/overview-descriptive-analysis>.
8. Coursera. (n.d.). What Is Predictive Analytics? Definition, Examples, and More. [online] Available at: <https://www.coursera.org/articles/predictive-analytics>.
9. McDaniel, S. (2019). What is Prescriptive Analytics? Definitions and Examples | Talend. [online] Talend Real-Time Open Source Data Integration Software. Available at: <https://www.talend.com/resources/what-is-prescriptive-analytics/>.
10. Anon, (2017). Top 14 Areas For Data Analytics Application. [online] Available at: <https://www.digitalvidya.com/blog/data-analytics-applications/>.
11. www.javatpoint.com. (n.d.). Linear Regression in Machine learning - Javatpoint. [online] Available at: <https://www.javatpoint.com/linear-regression-in-machine-learning>.
12. K, G.M. (2020). Machine Learning Basics: Decision Tree Regression. [online] Medium. Available at: <https://towardsdatascience.com/machine-learning-basics-decision-tree-regression-1d73ea003fda>.
13. IBM (n.d.). What is the k-nearest neighbors algorithm? | IBM. [online] www.ibm.com. Available at: <https://www.ibm.com/topics/knn>.
14. Mirjalili, S., Gandomi, A.H., Mirjalili, S.Z., Saremi, S., Faris, H. and Mirjalili, S.M. (2017). Salp Swarm Algorithm: A bio-inspired optimizer for engineering design problems. Advances in Engineering Software, 114, pp.163–191. doi: <https://www.sciencedirect.com/science/article/abs/pii/S0965997816307736\>
15. Tam, A. (2021). A Gentle Introduction to Particle Swarm Optimization. [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/a-gentle-introduction-to-particle-swarm-optimization/>.
16. GeeksforGeeks. (2021). Grey wolf optimization - Introduction. [online] Available at: <https://www.geeksforgeeks.org/grey-wolf-optimization-introduction/>
17. Harder, H. de (2022). Why Every Data Scientist Should Learn Mathematical Optimization. [online] Medium. Available at: <https://towardsdatascience.com/why-every-data-scientist-should-learn-mathematical-optimization-3ac102663456>
18. Xue, J. and Shen, B. (2020). A novel swarm intelligence optimization approach: sparrow search algorithm. Systems Science & Control Engineering, 8(1), pp.22–34. doi: <https://www.tandfonline.com/doi/full/10.1080/21642583.2019.1708830>
19. Br, April 18, on V. in B.D. on, 2019 and Pst, 11:50 A. (n.d.). Prescriptive analytics: A cheat sheet. [online] TechRepublic. Available at: <https://www.techrepublic.com/article/prescriptive-analytics-a-cheat-sheet/>.
20. Zach (2021). How to Compare Box Plots (With Examples). [online] Statology. Available at: <https://www.statology.org/compare-box-plots/>.
21. Xue, J. and Shen, B. (2020). A novel swarm intelligence optimization approach: sparrow search algorithm. Systems Science & Control Engineering, 8(1), pp.22–34. doi:https://doi.org/10.1080/21642583.2019.1708830.