Enhancing ML Accuracy:Vertical vs. Horizontal Dataset Expansion

Abstract

The aim of this project is to compare classification models for predicting whether the stock price of a given dataset will increase or decrease. The data explanation section provides insights into the columns of the dataset, while the data analysis phase involves identifying data imbalances, correlations between attributes, and employing moving averages to smooth the time series. Two distinct pipelines are introduced: Pipeline1 utilizes Moving Averages (MA) to add additional columns to the dataset, enhancing the features for training machine learning models. In contrast, Pipeline2 employs Monte Carlo simulation and Random Walks (MC/RW) to augment the dataset with more rows.The comparison of models involves evaluating performance on the original dataset, the Moving Average-enhanced dataset, and the Monte Carlo/Random Walk augmented dataset. The results indicate that the MC/RW dataset outperforms the MA dataset in predicting stock price changes on the validation set.

1 Introduction

Analyzing and predicting stock prices is a complex task that involves a combination of financial analysis, market trends, and sometimes even behavioral economics.This article addresses the widely acknowledged inherent unpredictability, dynamism, and nonlinearity of the stock market. The complexity of accurately estimating stock prices arises from a diverse array of macro and micro factors, including political influences, the status of the global economy, unforeseen events, and a company's financial performance. This study emphasizes the necessity of a sophisticated analytical approach, integrating market trends, behavioral economics, and financial research to effectively analyze and forecast stock values. Despite the intricate nature of the task, analysts employ a diverse range of techniques and resources to gain insights into potential future developments. The article provides an in-depth exploration of the challenges posed by various influencing factors and highlights the diverse methodologies utilized in the forecasting of future stock values.This project seeks to assess and compare the performance and characteristics of various machine learning models, focusing on two primary scenarios: classification and regression tasks applied at the same dataset. The following project is devided in to: i) Data Analysis, where, is explained the dataset, and an initial analysis of the dataset ii) ML Pipline, where it is explained the main procedure of the paper and the Results of the Project are illustrated at the last section.

2 Data

This section provides a thorough explanation of the project's initial phase, covering the description and analysis of the data. It offers a thorough explanation of the dataset, providing

details on its variables and organization. A systematic data analysis is then carried out, revealing important features, patterns, and trends present in the dataset. By means of this painstaking investigation, a fundamental comprehension of the data is built, laying the groundwork for later phases of the project, especially in relation to the creation and assessment of machine learning models.

2.1 Data Description

This project uses a dataset that provides a comprehensive exploration of Tesla's stock price evolution, encompassing the period from its initial public offering (2010-06-29) until 2017-03-17, indicating 1692 samples/days. Accessible on Kaggle, the dataset encompasses crucial details such as the date of each recorded data point, the opening price of Tesla's stock for a given day (labeled as "Open"), the highest and lowest recorded prices on that day ("High" and "Low" respectively), the closing price of the stock ("Close"), the volume of Tesla stocks traded during the day ("Volume"), and the adjusted closing price ("Adjusted Close"). The adjusted closing price takes into account any distributions or corporate actions that transpired before the opening of the subsequent day, offering a more nuanced perspective on the stock's performance. Nevertheless the dataset did not consists of any null or duplicated values, indicating a promising data.

2.2 Data analysis

This section is dedicated to the comprehensive analysis of the dataset. As depicted in Table 1, it provides an overview of each column, showcasing key statistical measures for attributes such as mean, minimum and maximum values, and standard deviation. For example, the mean opening price of Tesla over the given period is $132.44, with a minimum value of $16.14 and a maximum value of $287.67. Additionally, the standard deviation of the opening price is $94.31. These descriptive statistics offer a succinct yet informative summary of the dataset's numerical attributes, providing a foundation for further exploration and interpretation of the data.

Attributes	**Mean**	**(Max – Min)**	**Std**
Open	132.44	(287.67 – 16.14)	94.31
High	134.77	(291.42 – 16.63)	95.69
Low	130	(280.4 – 14.98)	92.86
Close	132.43	(286.04–15.8)	94.31
Volume	4.27074e+06	(37163900.0–118500.0)	4.29597e+06
Adj Close	132.43	(286.04–15.8)	94.31

Table 1: Data details

The provided figure (1) visually represents the various attributes from the initial year of 2010 to the concluding year of 2017. However, the figures presented do not explicitly con-

vey that the Open, Close, High, Low, and Adj Close attributes exhibit analogous behaviors. It is essential to note that this observation can be substantiated through two distinct approaches. Firstly, employing correlation formulas, such as Pearson, can be insightful. It is worth mentioning that while Pearson correlation assesses point-to-point relationships between two variables, Cross-correlation offers a more comprehensive evaluation of the time-series behavior.The figure (1) provides indications of a high Pearson correlation score for the time-series, suggesting a close relationship between the attributes. Additionally, the visual analysis of the figure suggests a strikingly similar behavior among these attributes, further supported by the application of cross-correlation.

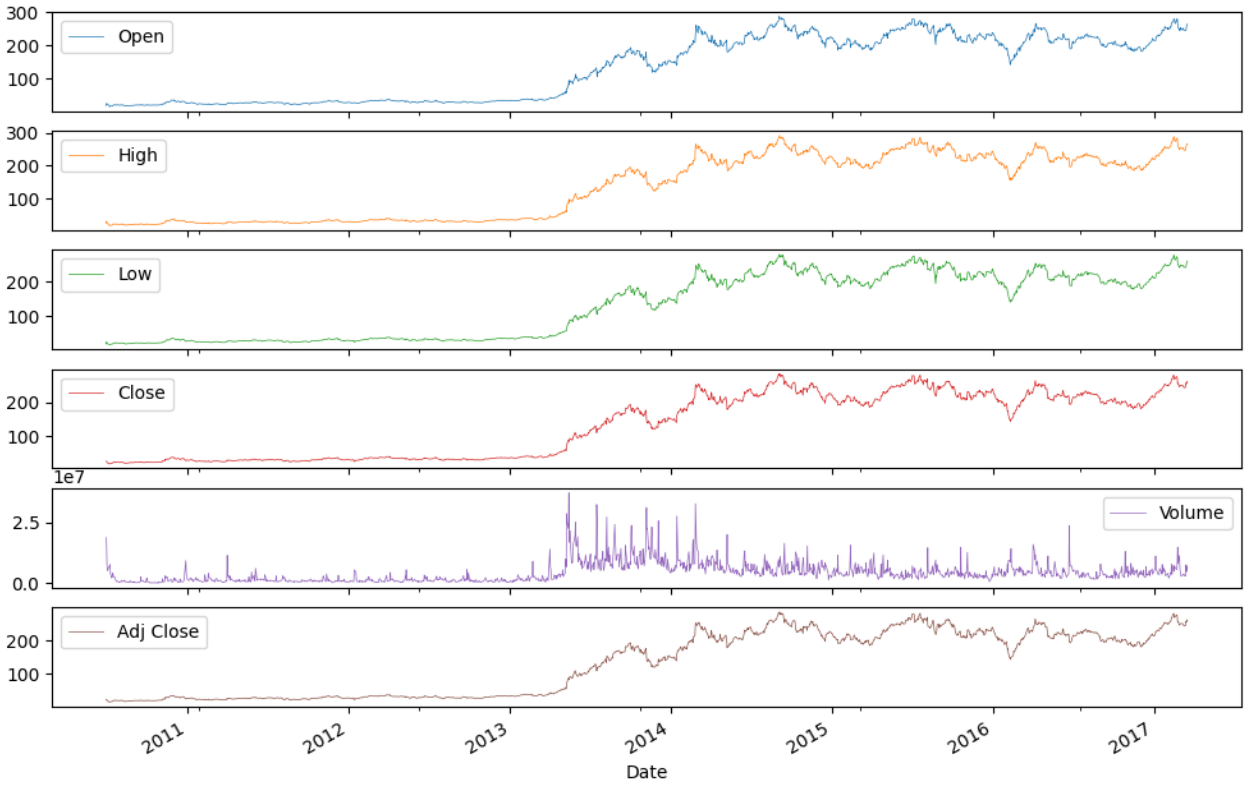


Figure 1: Attributes over Years

The first part of the analysis is a combination of Pearson correlation and Cross-Correlation. Figure 3 and Table 2 demonstrate a comparison of the two methods. To begin with, Pearson Correlation Coefficient uses the following formula

$$\rho_{XY}(\tau) = \frac{\text{cov}(X_t, Y_{t+\tau})}{\sigma_X \sigma_Y}$$

This formula measures linear correlation between two sets of data from point-to-point. If the coefficient is close to 1 or -1 indicates a linear relationship, however a score close to 0,

indicates a non-linear relationship. For instance, the opening and highest price of Tesla stock market, have a linear relationship, indicating, if the open price rises, then the highest price will rise. In the same way of thinking, we can see that High attribute and Volume do not have a linear relationship, indicating that if one of them rise, then we do not know the behavior of the other. On the other hand,cross-correlation is a statistical tool used to quantify the degree of similarity between two time series as a function of a time lags. However Cross-correlation does not, indicate connection, suggesting that changes in one time series do not always result in changes in the other. For instance, the Open and Close prices, have similar behaivior, according to cross correlation, but the rise of the one, do not necessarily mean the rise of the other, hence, using Pearson Correlation coefficient, we can clarify this. In conclusion, comparing the results from Pearson correlation and cross-correlation yields meaningful insights. While cross-correlation captures similarities in behavior over time, Pearson correlation helps discern the nature of the relationship, especially in cases where a linear correlation may not necessarily translate into a causal or predictive connection between the variables.

Attribute 1	Attribute 2	Score
Open	Close	0.99
Open	High	0.99
Open	Low	0.99
High	Low	0.99
High	Close	0.99
High	Volume	0.41
Low	Volume	0.39
Close	Volume	0.40
Low	Close	0.99

Table 2: Pearson Correlation Coefficient

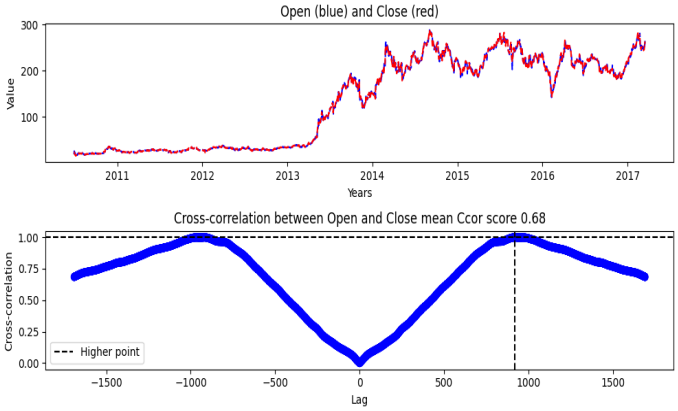


Table 3: Cross-Correlation

Merging the results of Cross-correlation and Pearson correlation leads to the conclusion that all attributes, except of 'Volume,' exhibit similar behavior and a linear correlation. This analysis is pivotal in enhancing model accuracy, providing valuable insights for predictive modeling and decision-making processes.

In this phase of the analysis, the creation of the target column is integral for classification models. This column involves a comparison where it evaluates whether the closing price of the next day is higher than that of the current day. To be more precise, when the closing price of the current day surpasses the closing price of the next day, the target column is assigned the value of one; otherwise, it is assigned zero. This binary target column serves as a crucial element for training and evaluating classification models. Following this crucial step, an observation reveals that the target column is balanced, with 826 instances where the current day's closing price is higher than that of the next day, and 866 instances with the

reverse occurrence. This balance in the target column lays a foundation for robust training and evaluation of classification models.

3 Machine Learning Models

This section centers on the employed machine learning models and their respective contributions to the project. The primary objective is to assess and compare the accuracy of machine learning models in accordance with their designated use cases. Initially, a meticulous comparison of various machine learning classification models is conducted, employing relevant evaluation metrics.Following this, two distinct datasets are generated: one employs moving averages to expand attribute dimensions, while the other utilizes Monte Carlo/Random Walks simulations to augment dataset samples. The objective is to discern the impact of these techniques on ML model accuracy. If models using moving averages achieve higher scores, it suggests that, for a classification problem, the quantity of attributes outweighs the significance of sample size. Conversely, if models with Monte Carlo simulations exhibit superior performance, it implies that, in this context, the number of samples holds greater importance than the number of attributes.

Classification models
In predicting whether the stock price will be higher or lower than the current day, a comparative analysis involves the utilization of five distinct classification models. The initial exploration of the dataset begins with a Decision Tree model to discern its behavioral patterns. Following this, comprehensive training and testing procedures are undertaken for a Random Forest, a Support Vector Machine, an XGBoost Classifier, and a Gaussian Naive Bayes model. The evaluation of each model is conducted based on key metrics such as accuracy, F1 score, and recall, while the overall error is quantified through the mean square error. This approach enables a thorough evaluation of each classification model's performance subtleties and predictive skills within the given stock price prediction scenario.

3.1 Evaluation Metrics

This paragraph briefly explains the evaluation matrices that are used in this project.

Confusion Matrix: A matrix that displays the performance of an algorithm

Accuracy: The number of correctly predicted sample points out of all the data points is accuracy. $\frac{TP+TN}{TP+FP+FN+TN}$

Precision: The percentage of positive identifications that were actually correct is precision. $\frac{TP}{TP+FP}$

Recall: The percentage of real positive values that are correctly identified is referred to as recall. $\frac{TP}{TP+FN}$

F1-score: The weighted average of Precision and Recall is the F1 Score. $2 * \frac{(Recall * Precision)}{(Recall + Precision)}$

3.2 Monte Carlo/Random Walks vS Moving Average

Moving Average

The moving average is a mathematical calculation used to analyze datasets by generating a sequence of averages from various subsets of the complete dataset. Specifically, the simple moving average (SMA) is the unweighted mean of the preceding N-data points. This computation involves taking the average of a designated set of values within a specified window. As new data points are introduced, the moving average is continuously recalculated along the data series, providing a smoothed representation of trends or patterns within the dataset.

$$\text{SMA} = \frac{\text{Sum of values over specified period}}{\text{Number of values in the set}} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

,

Where:

N Moving window,

x_i ith value in the set.

Monte Carlo/Random Walks

Monte Carlo simulation is a mathematical technique, used for investigating and analyzing complex systems by simulating a variety of possible outcomes for unknown occurrences. Monte Carlo simulations provide forecasts based on estimated ranges of values, introducing volatility into the evolution of the modeled situations, in contrast to traditional approaches that depend on fixed values. Monte Carlo Simulation is used in finance to generate an updated equity curve by introducing random changes into previous stock data, such as when assessing the stability of a trading strategy. This method makes it possible to evaluate how well a trading strategy can tolerate and adjust to the intrinsic unpredictability brought forth by random fluctuations. More details of Monte Carlo and Random Walks can be found here.

4 Methodology & Results

This section elucidates the methodological approach employed in this paper to address the underlying problem, providing a comprehensive blueprint of the implemented code.

The primary strategy involves the vertical expansion of the dataset through the utilization of a moving average technique, specifically employing sliding windows of 5, 20, 30 and 50 days. This results in the creation of three new columns for each attribute, exemplified by the addition of rolling averages for opening prices over the aforementioned periods. The resulting augmented dataset encompasses a substantial sample size and boasts 1643 attributes across 31 distinctive features. This methodology serves as the foundation for subsequent analyses and findings presented in this research endeavor.In Figure 2, the Moving Average for all attributes is visually depicted, providing a comprehensive overview of the analytical process. Complementing this visualization, Algorithm 1 serves as a detailed blueprint delineating the procedural steps undertaken in the calculation of the Moving Average. Together, these graphical and algorithmic representations contribute to the clarity and transparency of the methodology employed in this study for moving average analysis.

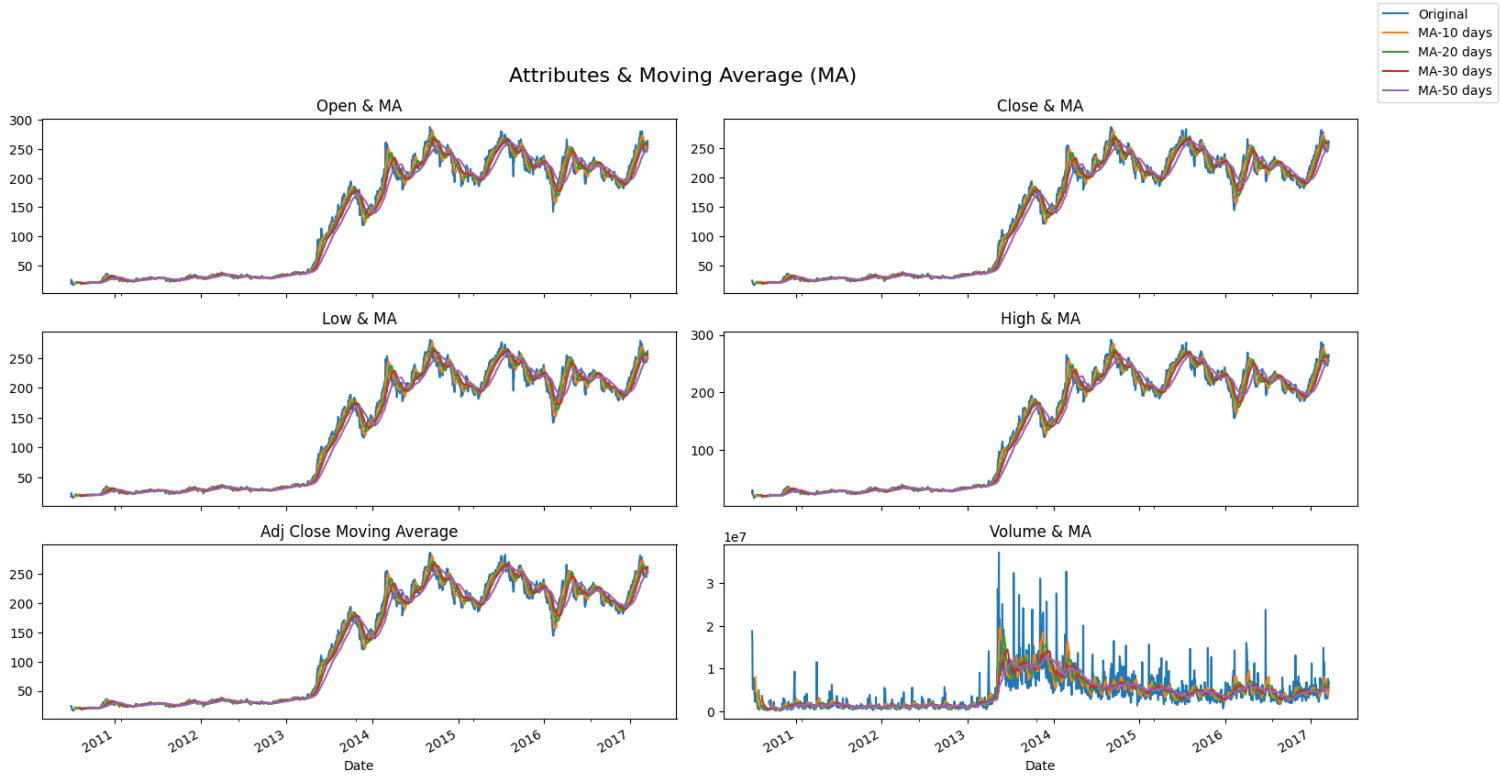


Figure 2: Moving Average

Following the dataset update, this project proceeds to partition the dataset into training and validation sets, subsequently training the specified machine learning (ML) algorithms to evaluate their performance. Each ML algorithm is trained on 1314 samples and 30 features and then validated on 329 samples. The performance metrics of each algorithm are presented in Table 6, offering a quantitative assessment. Concurrently, Figure 7 illustrates the Receiver

Algorithm 1 Moving Average Algorithm

Require: List of Rolling Step (days) & List with Features

1: **for** $Step_i \ \forall$ `i` $\in$ $RollingSteps$ **do**
2: **for** $Feature_i \ \forall$ `i` $\in [Features]$ **do**
3: Calculate Feature's Moving Average
4: Add to the Dataset
5: **end for**
6: **end for**
7: Delete the NA values from the update Dataset

Operating Characteristic (ROC) curve for each algorithm, providing a statistical analysis of their classification effectiveness. Notably, the proximity of the curve to the dashed line indicates the model's correctness, with the Random Forest Algorithm demonstrating a higher capacity for correct classifications compared to the Support Vector Machine. This observation underscores the importance of considering additional metrics, such as ROC curves, beyond accuracy scores to comprehensively assess the capabilities of different ML algorithms.

	RF	DT	XGB	Gaussian	SVM
Acc	0.55	0.54	0.55	0.5	0.53
Precision	0.57	0.56	0.57	0.51	0.52
Recall	0.53	0.53	0.53	0.63	0.95
F1-Score	0.55	0.54	0.55	0.56	0.67

Table 4: Moving Average Metrics of ML

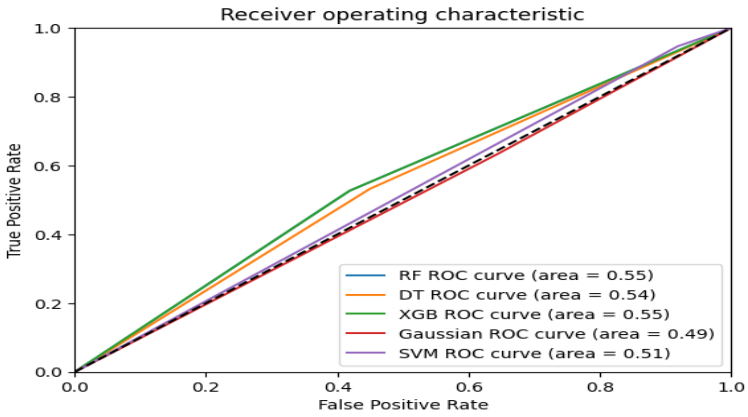
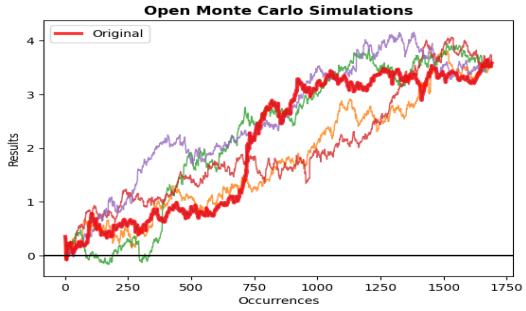


Table 5: ROC Curve

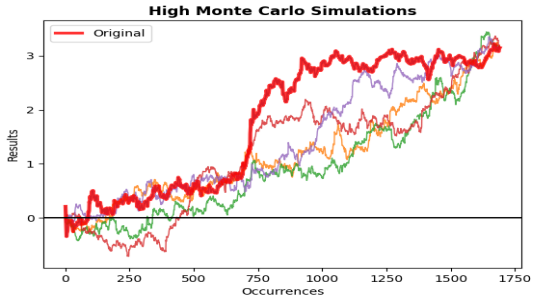
The table presents the performance metrics of various machine learning algorithms, including Random Forest (RF), Decision Tree (DT), XGBoost (XGB), Gaussian, and Support Vector Machine (SVM). Accuracy scores range from 0.5 to 0.55, with Random Forest achieving the highest accuracy. Precision values show similar patterns, while Recall is notably high for SVM. F1-Score, a balance between precision and recall, reveals varied performance across the algorithms, emphasizing the need for a comprehensive evaluation.

The secondary approach involves augmenting the dataset by increasing the number of samples, signifying a dataset with more rows while retaining the same number of features. Leveraging Monte Carlo simulation and Random Walks, this project generates random walks for five simulations, indicating that for each feature, five distinct simulations are created. Sub-

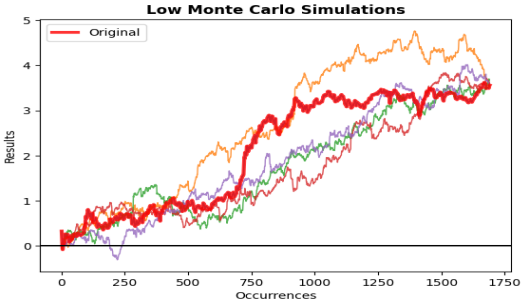
sequently, these simulations are incorporated into the dataset as extensions of the respective features. Figure 3 visually presents the Monte Carlo Simulations for all features, with each simulation distinguished by a unique color, providing a comprehensive representation of the varied trajectories resulting from the simulations.



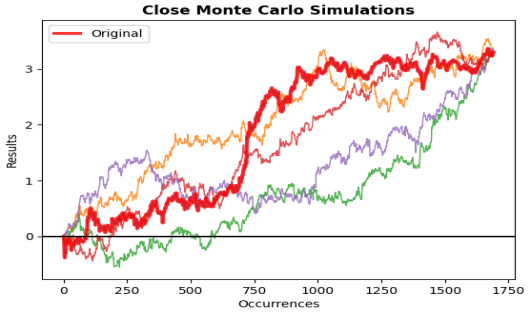
(a) MC 'Open' attribute



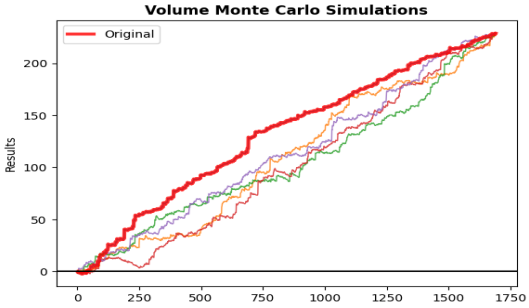
(b) MC 'High' attribute



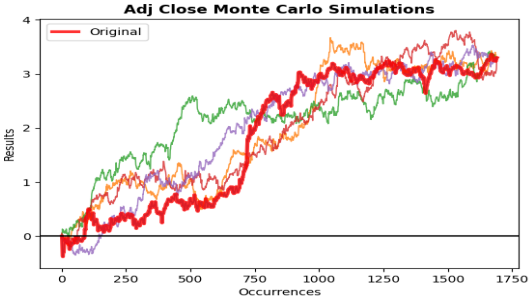
(c) MC 'Low' attribute



(d) MC 'Close' attribute



(e) MC 'Volume' attribute



(f) MC 'Adj Close' attribute

Figure 3: Monte Carlo Simulations

Upon updating the dataset with the augmented samples from Monte Carlo simulations and Random Walks, this project proceeds to train from snatch the same machine learning algorithms. The performance of these algorithms is then compared utilizing ROC curves, as conducted previously. This iterative evaluation provides insights into how the algorithms perform with the expanded dataset, enabling a comprehensive assessment of their effective-

ness in handling the increased sample size and diverse scenarios introduced by the Monte Carlo simulations and Random Walks.

	RF	DT	XGB	Gaussian	SVM
Acc	0.73	0.65	0.71	0.73	0.49
Precision	0.73	0.65	0.71	0.75	0.49
Recall	0.72	0.65	0.72	0.7	0.25
F1-Score	0.72	0.65	0.72	0.73	0.34

Table 6: Monte Carlo Metrics of ML

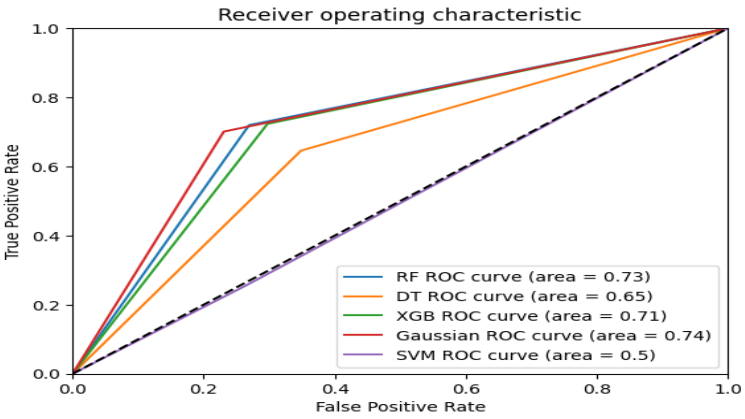


Table 7: ROC Curve

The table presents the updated performance metrics of various machine learning algorithms after the dataset augmentation with Monte Carlo simulations and Random Walks. Notable changes are observed in the accuracy (Acc) and precision scores, where Random Forest (RF) maintains the highest values. Gaussian and SVM, however, exhibit reduced accuracy, precision, recall, and F1-Score values compared to the initial evaluation. This suggests that while some algorithms perform well with the expanded dataset, others may struggle, emphasizing the need for a nuanced understanding of algorithmic responses to diverse data scenarios.

Expanding the dataset through increased samples using Monte Carlo and Random Walks proves advantageous for ML models, resulting in enhanced accuracy compared to both the Moving Average dataset and the feature expansion strategy. Notably, Random Forest achieves a significant improvement with a 73% accuracy and demonstrates a favorable ROC curve, contrasting with its performance with the Moving Average dataset. Other algorithms also exhibit improved performance with the Monte Carlo modification, highlighting the effectiveness of this approach. However, it's worth noting that the SVM model maintains a consistent performance, showing no significant improvement with the increased dataset samples.

5 Conclusion

In summary, this project conducts a comprehensive comparison of accuracy and performance across various ML algorithms, considering variations in the training dataset. The expansion strategies involve vertical expansion, accomplished by increasing features through Moving Average, and horizontal expansion, achieved by increasing samples using Monte Carlo. The consistent training and validation of identical algorithms on these distinct datasets reveal that augmenting the number of samples has a more substantial impact on improving accuracy and model performance compared to increasing the number of features. Notably, SVM consistently demonstrates lower performance across both scenarios. This analysis underscores the significance of dataset characteristics in influencing model outcomes and highlights the varying responses of algorithms to different types of dataset modifications.In conclusion, this study underscores the pivotal role of dataset characteristics in influencing machine learning model performance, revealing that increasing the number of samples through horizontal expansion has a more pronounced impact on accuracy than vertical expansion of features, with SVM exhibiting consistent challenges in both scenarios.