Overview

The goal of this study is to compare the performance of four different machine learning models when applied to the same dataset using the same train/test splits. The main objective is to determine whether these models yield varying results. This analysis is conducted based on the following assumptions:

1. Model architecture: Each machine learning model has its unique structure, influencing how it handles and learns from input data.

2. Hyperparameter settings: Machine learning models possess hyperparameters that need to be optimized to achieve optimal performance on a particular dataset. The selection of hyperparameters can significantly impact a model's performance.

3. Regularization: Various models employ different techniques to prevent overfitting, which can also impact their performance.

4. Noise in the data: Models may differ in their ability to handle noise within the dataset, potentially leading to dissimilar outcomes.

In summary, this study investigates how these factors influence the performance of different machine learning models when applied to the same dataset.

Hypothesis Summary

For this study, we have selected four machine learning models which are Logistic Regression (LR), Support Vector Machine (SVM), Random Forest Classifier (RFC), and Decision Tree Classifier (DTC). We will be testing these models to determine their scores in predicting the

closing price of the stock using relevant features. For our study, we have formulated the following hypotheses:

Null Hypothesis: The performance of the four machine learning models will be identical.

Alternative Hypothesis: At least one out of the four machine learning models will differ from the others.

Throughout the study, we will be testing these hypotheses and evaluating the performance of each machine learning model accordingly.

Data Set Overview

For this project, the primary dataset used consists of a CSV file containing stock prices recorded at different times. The file encompasses various details, including the date, opening price, high price, low price, closing price, and volume of the stock. The dataset covers the time span from January 2015 to June 2017.

The data utilized for this analysis was collected from Kaggle's official website (www.kaggle.com). In our study, we will focus on three input features: the opening, high, and low price. As for the closing price, it will be transformed into multiple classes to facilitate our analysis.

To enhance the accuracy of our analysis, we will employ principal component analysis (PCA), a statistical technique that simplifies the data reducing the noise. By utilizing PCA, we can retain the most significant information from the original data. This approach will enable us to concentrate on the most relevant part of the data and enhance the precision of our analysis.

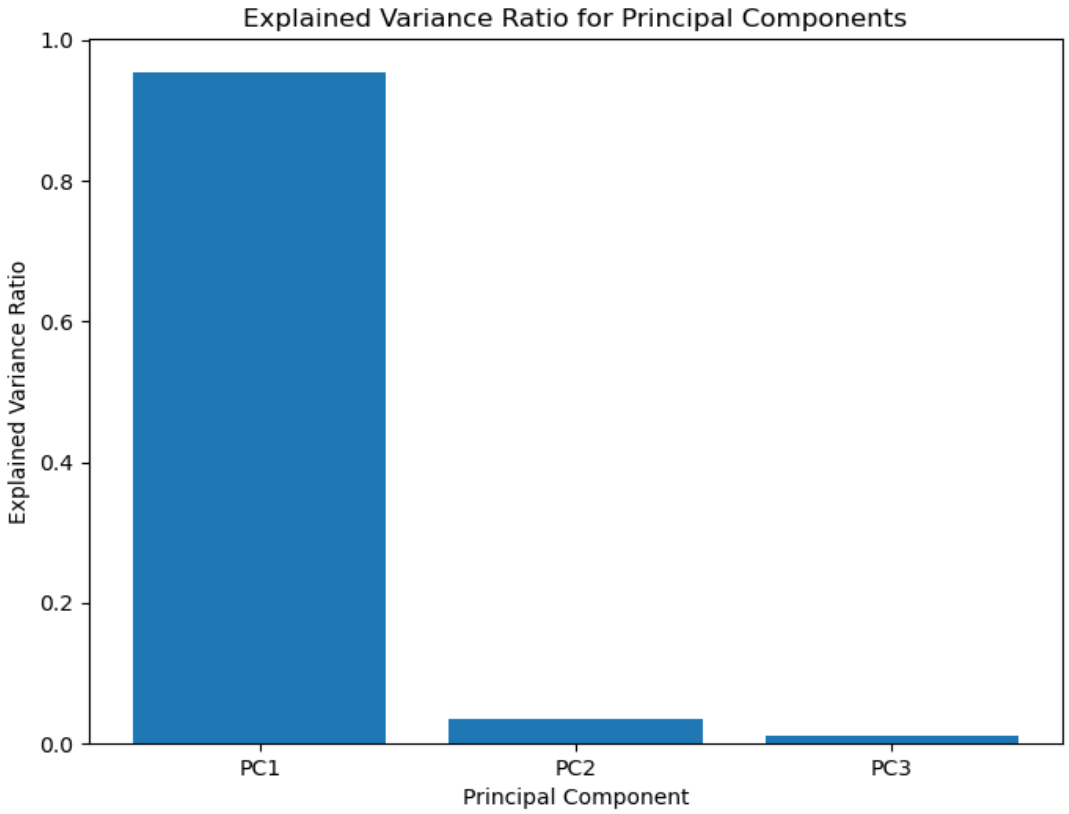
Analytics Performed

For our analysis, we will use the Scikit-Learn libraries in Python. We will utilize the aforementioned four machine learning models to train and test our dataset. To ensure robust evaluation, we will employ cross-validation with multiple folds.

Results

After performing Principal Component Analysis (PCA), we generated the subsequent results.

Principal Component	Variance Explained
Component 1 (PC1)	95.40%
Component 2 (PC2)	3.48%
Component 3 (PC3)	1.12%
Total	**100.00%**



Based on the provided results, we can conclude that the first principal component captures 95.40% of the total variance in the dataset. Including the second component further increases the coverage to 98.88% of the variance. Since the third principal component explains only 1.12% of the variance, we will focus on the first two components for our subsequent analysis.

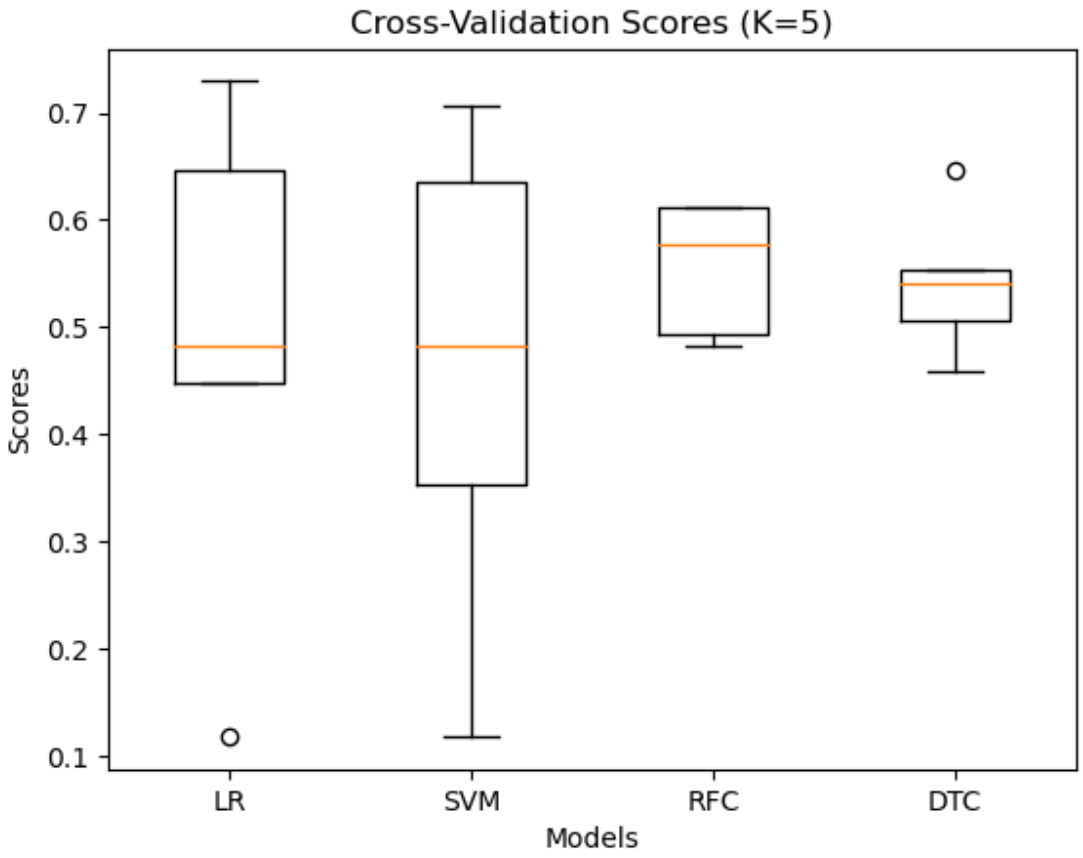
For training and testing our machine learning models, we divided the dataset into an 80% training set and a 20% testing set. The features used for training and testing are the opening, high, and low price. The target variable is classified based on the following criteria.

Criteria	Classification
Closing Price > Opening Price	Class 2 (Positive)
Closing Price = Opening Price	Class 1 (Neutral)
Closing Price < Opening Price	Class 0 (Negative)

Initially, we implemented a 5-fold cross-validation approach. By executing the Python code, we obtained a summary of the generated outputs as presented below.

Model	Min	Q1	Median	Q3	Max
LR	11.76%	44.71%	48.24%	64.71%	72.94%
SVM	11.76%	35.29%	48.24%	63.53%	70.59%
RFC	48.24%	49.41%	57.65%	61.18%	61.18%
DTC	45.88%	50.59%	54.12%	55.29%	64.71%

The table highlights variations in the results among different models. To provide a clearer representation of these differences, the table is visualized in the form of a box plot below.

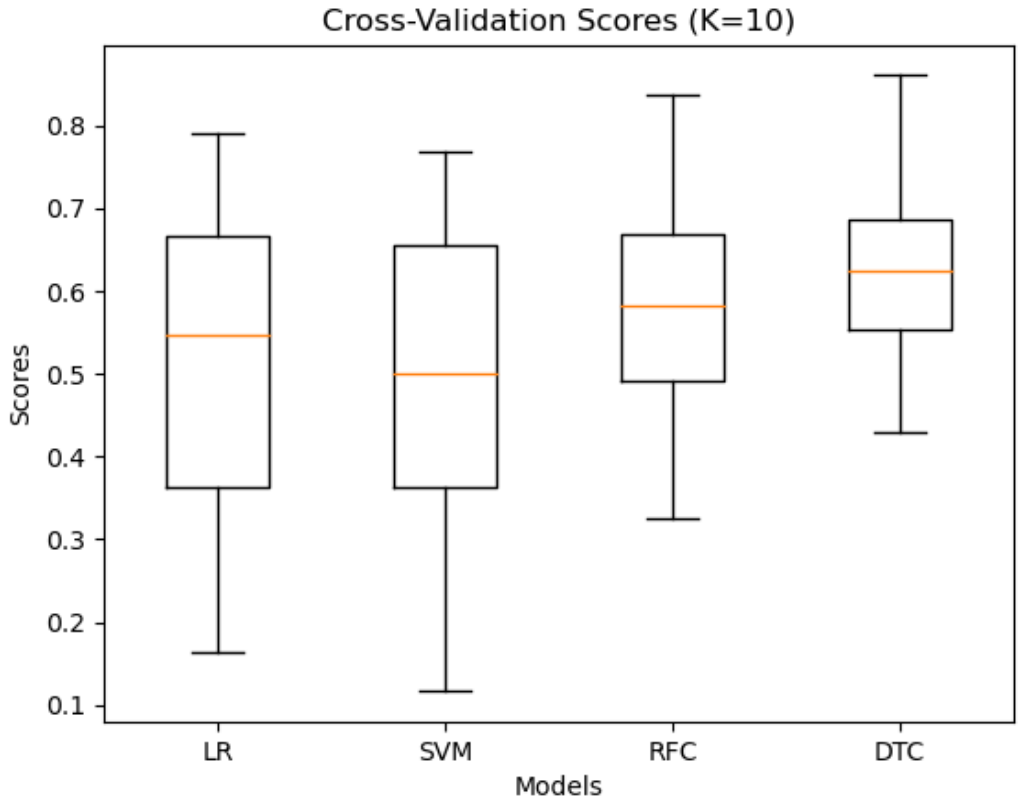


Both the SVM and LR models exhibit identical median scores and third quartiles. On the other hand, the RFC and DTC models demonstrate slight differences. There is a significant difference between the minimum and maximum values of the prediction scores.

To validate our analysis results, we decided to increase the number of folds in our cross-validation to 10. The following summary provides an overview of the results obtained with the increased folds.

Model	Min	Q1	Median	Q3	Max
LR	16.28%	36.31%	54.76%	66.61%	79.07%
SVM	11.63%	36.31%	50.00%	65.46%	76.74%
RFC	32.56%	49.09%	58.33%	66.90%	83.72%
DTC	42.86%	55.36%	62.38%	68.60%	86.05%

The scores obtained from the 10-fold cross-validation further reinforce the analysis from the 5-fold analysis. The median score of the SVM model is lowest among the models. While LR and SVM models exhibit identical first quartiles, they are notably distinct from the other models. On the other hand, the third quartile shows almost uniform results across all models. The maximum values have less differences compared to the differences among minimum values across all models. These findings are visualized in the box plot below.



As mentioned earlier, the box plot reveals subtle variations in the median scores and interquartile ranges among the four models. It is evident that the scores are not identical across the models, indicating some degree of differentiation among them.

Conclusion and Future Work

Based on the analysis and findings presented above, we reject the null hypothesis and conclude that the scores among the different machine learning models exhibit slight variations. These variations can be attributed to differences in model architecture, hyperparameter settings, regularization techniques, and noise handling, which impact their predictions.

It is important to note that this research utilized a limited set of metrics and visualizations. Future researchers may consider incorporating additional metrics to enhance the analysis. Moreover, our study focused on a specific time frame of two years to simplify calculations and processing time, but including more data could potentially yield more accurate results. Additionally, other machine learning algorithms such as Naive Bayes were not included in this study, providing an opportunity for future researchers to explore additional models. Lastly, our study employed a limited number of features, which could be expanded upon in future research endeavors.