A logo for university computing

Auto-generated description

**Evaluation cover page**

|  |  |
| --- | --- |
| *Student's Full Name:* | Rosemary Dejesus Ramirez Cords. |
| *Student Number:* | 2024017. |
| *Module Title:* | Strategic Thinking. |
| *Evaluation Title:* | CA 3 Final Submission. |
| *Assessment Due Date:* | 08/11/2024. |
| *Presentation Date* | 08/11/2024. |

**Declaration**

By submitting this review, I confirm that I have read CCT's policy on academic misconduct and understand the implications of submitting work that is not mine or that does not appropriately reference material taken from a third party or other source.

I declare that it is my own work and that all third-party material has been properly referenced.

I further confirm that this work has not previously been submitted for evaluation by me or anyone else at CCT College Dublin or any other higher education institution.

Impact of Drought Events on Grains Pricing.

[Introduction 2](#_Toc181999951)

[Objectives 3](#_Toc181999952)

[Defining the Problem 4](#_Toc181999953)

[Scope 5](#_Toc181999954)

[Data Sources 6](#_Toc181999955)

[Ethical Considerations 6](#_Toc181999956)

[Develop Data Set 7](#_Toc181999957)

[Characterization of data and pre-processing 7](#_Toc181999958)

[Scaling and Normalization 11](#_Toc181999959)

[Training and Testing our Data Set 11](#_Toc181999960)

[Applying Modals 12](#_Toc181999961)

[Cross validation and Hyperparameter Tunning 22](#_Toc181999962)

[Conclusion 25](#_Toc181999963)

[References 26](#_Toc181999964)

[GitHub link 26](#_Toc181999965)

**Title:** Impact of Drought Events on Grains Pricing.

# Introduction

Agriculture is a complex sector that involves different driving parameters (environmental, economic, and social). Agricultural production is now known to be highly sensitive to climate change (Easterling et al., 2007).

Climate change affects all agricultural sectors in a multitude of ways that vary from region to region, reducing the predictability of seasonal weather patterns and increasing the frequency and intensity of extreme weather events, such as floods, cyclones, and heatwaves (Food and Agriculture Organization, FAO, 2011).

Climatic factors directly impact the supply and demand of grains in the market, consequently influencing prices in accordance with the principles of the Law of Supply and Demand.

Of all the categories of commodities, grain commodities prices play a critical role in everyone's daily life. Fluctuations in grain commodities prices pose a threat to consumers and lead to instability in the incomes and operations of farmers' households (Ayankoya et al., 2016).

To cope with anticipated changes in climatic conditions, can resort to – among others – the following measures: modify your crop rotation to optimize the use of available water, readjust planting dates based on temperature patterns and rainfall, use crop varieties better adapted to new weather conditions (for example, more resistant to heat and drought) and plant in tilled lands or small areas trees that reduce runoff and serve as windbreaks. Among the key measures that the EU and its States can provide to the agricultural community with more precise information on climate risks and adaptation options and providing support for advisory services as well as activities deformation (Climate change and European agriculture, The challenges ahead, available from <https://publications.europa.eu/resource/cellar/14d3648c-4078-46eb-90dd-c4e787a32fca.0011.02/DOC_1>)

Given the increasing volatility in grain commodity prices due to climate change and its associated impacts, this project aims to offer valuable tools for decision-making. By providing management with the option to analyse and compare historical price trends during weather events, this project seeks to help stakeholders in the agricultural sector make informed decisions that can mitigate risks and enhance their resilience to the ongoing and anticipated changes in climatic conditions. In this way, the project aspires to contribute forecasting values, enabling stakeholders to better navigate the uncertainties posed by climate change.

# Objectives

1. Examine the Impact of Weather Events on Stock Market Performance, analysing how some specific weather conditions influence in stock market returns over time.
2. Analyse the causal relationships between time series data points and external factors to identify key drivers of changes in the data, such as weather factors.
3. Test and compare different machine algorithm models as Linear Regression, SARIMA and Random Forest Regressor to assess their accuracy in forecasting or capturing the underlying structure of the data.

# Defining the Problem

Agricultural production is affected by different market factors, which affect supply and demand and in consequence pricing.

Climatic factors in agriculture are difficult for producers to handle because they cannot be controlled by them, such as temperature, precipitation, humidity, and seasonal changes, directly impact crop growth, yield, and overall productivity. These factors often determine the success or failure of a harvest, influencing the amount of produce available in the market. For example, an unexpected drought or excessive rainfall can lead to poor harvests, thereby reducing the supply of crops, which, in turn, increases prices due to scarcity. On the other hand, favorable weather conditions can lead to bumper crops, which might result in lower prices due to an oversupply of produce.

However, approximately 90% of natural disasters registered in Europe since 1980 can be attributed directly or indirectly to meteorological causes and climatic, and represent around 95% of the losses economic, it caused by natural disasters. The global losses derived from climatic phenomena and meteorological events have experienced a notable increase for the last 25 years. Although social changes and economic development are the factors that have most influenced, however, still It is too early to determine by what percentage the increase in losses can be attributed to the climate change of anthropogenic origin (The impacts of climate change in Europe: indicator-based on evaluation, 2011, available from <https://www.miteco.gob.es/content/dam/miteco/es/calidad-y-evaluacion-ambiental/publicaciones/impactos%20cambio%20climatico_tcm30-185070.pdf>)

By recognizing the diverse effects of climatic factors on the stock market, we can develop sustainable solutions and strategies aimed at mitigating the potential disruptions to market stability. Climate-related events, such as temperature fluctuations, extreme weather, and changing precipitation patterns, can significantly affect the performance of agricultural stocks and commodity prices, leading to volatility in the market. Understanding these relationships enables the creation of more resilient financial systems and investment strategies.

# Scope

Defining the project scope is identifying all the work that the project will accomplish to achieve its final goal. It is used to develop and confirm a common understanding of the project scope among key project stakeholders The project team has identified the activities that will be necessarily to support the project. (Project Scope Management, 2016, available from <https://www.pm4dev.com/resources/free-e-books/7-project-scope-management/file.html>)

It is imperative to fully understand the complex correlation between weather patterns and grain prices, as this relationship plays a critical role in shaping the agricultural market and the broader economy. Awareness of factors such as temperature variations, precipitation levels, frequency and intensity of natural disasters, and seasonal timing is essential for stakeholders within the agricultural sector to better anticipate and respond to price fluctuations driven by changing weather conditions. These weather-related factors directly influence the supply and demand of grain commodities, which, in turn, impacts market prices.

Compare how prices developed during climate event (drought) using prices and weather data between 2000 and 2024 and implementing machine learning techniques. Including corn, oat, wheat, rice, soybean and soybean oil as a grain, prices as a dependent column and drought as a climatic factor and independent column.

**Timeline:**

This project is structured across two semesters, with key tasks to be completed throughout the months of March to November. The timeline helps us track progress and ensure all steps are covered. Below is a breakdown of each task and its purpose:



The yellow highlights in the table represent completed tasks within the corresponding months. Each phase builds on the previous one, ensuring a logical and methodical approach to completing the project.

# Data Sources

A practical approach to defining data is that data are numbers, characters, images, or other method of recording, in a form which can be assessed to make a determination or decision about a specific action. Many believe that data on its own has no meaning, only when interpreted does it take on meaning and become information. By closely examining data we can find patterns to perceive information, and then information can be used to enhance knowledge (Denis Howe, 1993-2005).

To create the data set, data was collected from various sources. The main data set <https://www.kaggle.com/datasets/guillemservera/grains-and-cereals-futures?select=individual_data> (obtained from Kaggle ), which has the follow license to use it <https://creativecommons.org/licenses/by-nc/4.0/> contains data on cereal prices from the years 2000 to 2024. This data set was enriched with data referring to the climate <https://www.kaggle.com/datasets/pavansanagapati/usdroughtdata>, (obtained from Kaggle), which has the follow license to use it <https://creativecommons.org/publicdomain/zero/1.0/>.

# Ethical Considerations

Ethics concerns questions about how people should act and what constitutes truthful behaviour (Lewis,1985).

Wherever data is used to predict and support decision-making processes, those decisions can affect people in many ways (Barocas & Selbst, 2016). Although the growing field of data science has brought many new possibilities for problem solving and developing new insights based on data analysis (Saltz & Dewar, 2019), the topic of ethical challenges and the “appropriate” way of using data has only recently been starting to receive the attention it deserves. Since an overall compliance regarding to what is considered ethical vs. unethical seems to be lacking (Asadi-Someh et al., 2016), the field of data science requires a more thorough investigation.

The idea of ethics involves not only human rights but also the rights of data derived from people as well as how to best handle this abundance of information for the greater good.

# Develop Data Set

According to describes in the capstone proposal we want to predict futures prices of grain by weather condition; to begin our research, we processed the data in the following manner to ensured that our data was properly prepared and that our models were reliable and accurate, providing a solid foundation for our research findings.

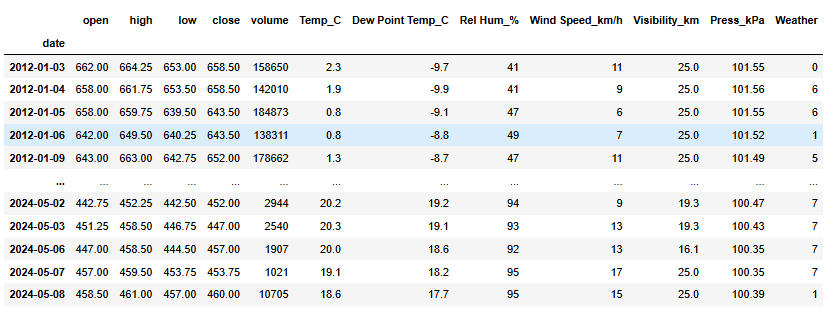
# Characterization of data and pre-processing

Exploratory Data Analysis or (EDA) is understanding the data set by summarizing its main characteristics and often plotting them visually. This step is very important especially when we arrive at modelling the data to apply Machine learning. Plotting in EDA consists of Histograms, Box plot, Scatter plots and many more. Through the process of EDA, we can also refine the problem statement or definition of our problem.(McQuaid, D. (2024b). file:///C:/Users/Dell/Downloads/Feature%20Scaling%20or%20Normalization%20(3).pdf.).

Also, define the characteristics of our data (number of columns, rows, null values, etc). In the following part we will explain about it.

These data is a csv document which we display as the name of “df1”(corn data set) and “df2”(weather data set) in our Jupyter Notebook.

We need to combine the two data set df1 and df2 and also put Date as index for our present analysis, so we got the next DataFrame:



**Fig.1: Join data set with date as index.**

In order to know how many columns, rows and which data types we have we display the function .info

**A screenshot of a computer

Description automatically generated**

**Fig.2: Display rows, columns and data types.**

In this case our data have 5956 rows (observations) and 13 columns (features) with object as a data type which means string values.

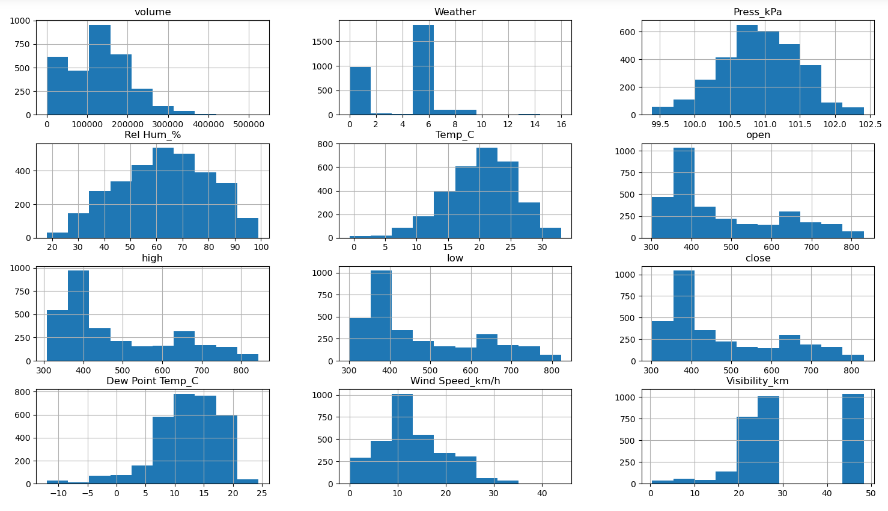
In regards to know if we have null values in our data set, we use the function .isnull().sum()

**A screenshot of a computer

Description automatically generated**

**Fig. 3: Display null values.**

To visualize distribution of all variables in the data we use histograms



**Fig.4: Display Histograms.**

To identify outliers in the features, we use boxplot

**A diagram of a graph

Description automatically generated with medium confidence**

**Fig. 5: Display boxplot.**

To visualise the relationship between the features and the response we can use scatterplots, here some examples:

A graph with blue dots

Description automatically generatedA graph with blue dots

Description automatically generated

A graph with blue dots

Description automatically generated A graph of a number of blue dots

Description automatically generated

**Fig.6: Display scatterplots, features vs target.**

When we want to apply our modals, we need to replace categorical values for numerical values, as we see before in function .info the feature “Weather” has that condition.

For this case we use a Nominal encoder because it doesn’t have any range of importance between each category.

Also, we need to define our dependant(Y) and independent(X) values.

# 

# Scaling and Normalization

Some features, such as latitude or longitude, are bounded in value. Other numeric features, such as counts, may increase without bound. Models that are smooth functions of the input, such as linear regression, logistic regression, or anything that involves a matrix, are affected by the scale of the input. Tree-based models, on the other hand, couldn’t care less. If your model is sensitive to the scale of input features, feature scaling could help. As the name suggests, feature scaling changes the scale of the feature. Sometimes people also call it feature normalization. Feature scaling is usually done individually to each feature. Next, we will discuss several types of common scaling operations, each resulting in a different distribution of feature values.( McQuaid, D.(2024a)file:///C:/Users/Dell/Downloads/Feature%20Scaling%20or%20Normalization%20(3).pdf.)

So, before to apply our models, normalization is necessary to perform our points of data and have the same measure on it. In this case we use RobustScaler because as our boxplot show we could identify outliers in the data set

# Training and Testing our Data Set

Regarding apply the model we need to split the data for train and test. In this first case we use for test size 30% .

**A screenshot of a computer code

Description automatically generated**

**Fig 7: Splitting data.**

For test size of 30% we will use 931 observations and 11 features.

# Applying Modals

Machine learning algorithms that learn from input/output pairs are called supervised learning algorithms because a “teacher” provides supervision to the algorithms in the form of the desired outputs for each example that they learn from.( Müller, A.C. and Guido, S. (2016). https://www.nrigroupindia.com/ebook/Introduction%20to%20Machine%20Learning%20with%20Python%20(%20PDFDrive.com%20)-min.pdf. Available at: <http://safaribooksonline.com/>.)

Linear Regression modal was a good option regarding to predict prices of corn by weather conditions, because our variables have continuous values.

A screenshot of a computer code

Description automatically generated

A screenshot of a computer code

Description automatically generated

**Fig. 8: Linear Regression score.**

According to our result, we got a high R2 score which can be indicative of a good model fit and strong predictive power.

**Cross validation**

A screenshot of a computer program

Description automatically generated

**Fig. 9: Cross validation score.**

Trying to avoid underfitting and overfitting in our modal we apply cross-validation, technique which estimate how well our model generalizes our data set.

The mean accuracy suggest that the model predict 98.40% of the time across the ten different folds of the cross-validation while the low Standard Deviation indicates that the model's performance is stable.

**Predictions and True Values**

**A graph of a graph

Description automatically generated**

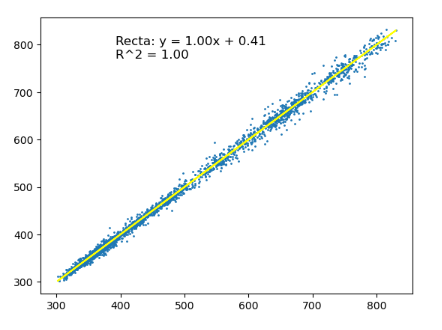
**A graph showing the difference between a graph and a graph

Description automatically generated with medium confidence**

**Fig. 10 and 11: Actual and predicted values.**

These graphics represent actual and predicted prices done after Linear Regression.

**The Linear Regression graphic.**

****

**Fig.12: Linear Regression model.**

**Error**

**A screenshot of a computer

Description automatically generated**

**A blue graph with text

Description automatically generated**

**Fig. 13: Error graphic.**

**A computer error message

Description automatically generated**

**Fig.14: Error score**

Mean squared error indicates that the predictions are very close to the actual values on average. The very high r square value suggests that the model captures almost all the variability in the target variable.

**Fig.15: Important features by Random Forest.**

We apply Random Forest regarding to find the most important features, for this case low, high and open.

**A graph with a number of blue bars

Description automatically generated**

**A graph of a graph

Description automatically generated with medium confidence**

**Time Series Forecast**

Time series forecasting is an important area of machine learning that is often neglected. It is important because there are so many prediction problems that involve a time component. These problems are neglected because it is this time component that makes time series problems more difficult to handle (Brownlee, 2017).

Our goal in this analysis is to Forecast Returns values from our stock market data. To begin analysing a time series, we started by examining the “close” points in the data to create a plot and identify any observable trends. This is important because, when applying time series algorithms, we need to detect seasonality. Identifying seasonality helps us determine the periodicity of the observations (e.g., weekly, daily, etc.). Additionally, analysing the trend allows us to assess whether the data is stationary. Stationarity is crucial for time series modelling, as it indicates that the statistical properties of the series (such as mean and variance) do not change over time. A stationary time series is essential for many modelling techniques, ensuring that predictions are reliable. So, as we can see the data point it’s not stationary because it’s changing over time.

**A graph showing a line of blue color

Description automatically generated with medium confidence**

**Fig. 16:** Close visualization.

Since the plot uses yearly data, I decided to analyze the seasonality directly from the data frame. Upon inspection, it's clear that the data follows a weekly pattern, which was expected, as the stock market operates on business days (Monday to Friday).

**A table with numbers and numbers

Description automatically generated**

**Fig. 17:** Stock market data frame.

To confirm the observation from the previous plot regarding non-stationary data, we can calculate the p-value using the Augmented Dickey-Fuller (ADF) test. A p-value of 0 indicates that the data is stationary, while a p-value greater than 0 suggests non-stationarity. As we can see, the p-value we obtained is greater than 0, indicating that the data is indeed non-stationary.

****

**Fig. 18:** ADF Test.

Calculating returns from closing prices is a common practice in time series analysis. The return is typically calculated as the percentage change in the closing price from one period to the next. By transforming raw closing prices into returns, we often achieve stationarity, as returns tend to exhibit more stable statistical properties (e.g., constant mean and variance) over time. This transformation is crucial because stationarity is a key assumption for many time series models, enabling more accurate and effective forecasting of financial data.

**A group of numbers and digits

Description automatically generated**

**Fig.19:** Returns values.

For our time series analysis, we will focus only on the "Returns" and the exogenous variables. Since the other features such as 'open', 'high', 'low', 'volume', and 'close' are no longer needed, we have decided to drop them. As a result, our updated data frame now looks like this:

A table with numbers and text

Description automatically generated

**Fig.20:** Returns and exogenous data frame.

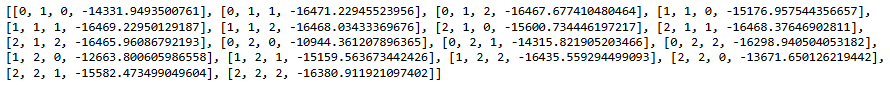
Now we ca recalculate the Augmented Dickey-Fuller (ADF) test, checking if we get a p value of 0, which means that we have a stationary data.



**Fig. 21:** ADF test.

Since our data has become stationary, we can now proceed with modelling. The first step is to evaluate various ARIMA model configurations and calculate their AIC (Akaike Information Criterion) values. We will assess different combinations of parameters for ARIMA (**p**: the number of autoregressive terms, **d**: the degree of differencing, **q**: the number of moving average terms.

After testing multiple configurations, we found that the ARIMA(0,1,1) model has the lowest AIC value, indicating it is the optimal choice for our data.

****

**Fig. 22:** AIC values.

Once we have defined our AIC value, we can proceed to calculate SARIMAX, being “Returns” our target and “Temp, Weather, Wind Speed, etc.” our exogenous variables.

This model is used when the time series shows seasonality. This model is similar to **ARIMA** models, but we add a few parameters to include for the seasons. We write **SARIMA** as **ARIMA(p, d, q)(P, D, Q)m** => **p:** the number of autoregressive, **d:** degree of differencing, **q:** the number of moving average terms Iqbal, M. (2024).

A screenshot of a computer

Description automatically generated

**Fig. 23:** SARIMAX results.

The model includes external factors (e.g., Temp\_C, Rel\_Hum\_%, Wind Speed, Visibility, Pressure, Weather, Dew Point Temp), each with a coefficient indicating its impact on "Returns." Also, we can see Significant Predictors: Some variables, such as Relative Humidity and Visibility, have low p-values, suggesting they have a meaningful effect on "Returns."

Relative Humidity has a positive effect, meaning higher humidity correlates with higher "Returns." Visibility has a negative effect, indicating that increased visibility is associated with lower "Returns."

Since Visibility\_km has the highest absolute coefficient, it can be considered the exogenous variable with the most impact on "Returns" in this model.

In the follow image we plot the SARIMAX Test result.

The residuals oscillate around zero without clear patterns, indicating stability, though there are some extreme values at the edges. The histogram aligns closely with a normal distribution, showing only minor differences at the extremes. The Q-Q plot confirms a good fit to normality, particularly in central quantiles, with only slight deviations at the ends. In the correlogram, only the first lag shows significant autocorrelation, while the rest of the residuals are nearly independent. Overall, the model shows a strong fit, with acceptable normality and independence in the residuals.

A collage of graphs and diagrams

Description automatically generated

**Fig. 24:** SARIMAX plot.

In the forecast step, we will try to predict the Returns values, specifically, we will predict Return values over the next 156 steps, equivalent to 3 years. The graph below already shows a strong fit compared to historical data, but recognizing the weekly seasonality can further enhance the model's accuracy. By explicitly modeling this weekly cycle, we can better account for consistent patterns that repeat each week, leading to a more precise forecast.

A graph of a blue line

Description automatically generated

**Fig. 25:** SARIMAX plot.

The SARIMAX model's forecast indicates that the central prediction remains fairly stable over time, the model predicts stability in the data around zero, with no strong trend, but growing uncertainty over the long term. The widening of the confidence interval suggests that while the model doesn’t predict major deviations, it’s less certain about the exact values further into the future.

To evaluate the model performance, we calculate the R-squared score of the dataset to test the authenticity of the model.

We can also evaluate the model using the score method, which for regressors returns the R 2 score. The R 2 score, also known as the coefficient of determination, is a measure of goodness of a prediction for a regression model and yields a score between 0 and 1. A value of 1 corresponds to a perfect prediction, and a value of 0 corresponds to a constant model that just predicts the mean of the training set responses (Müller, A.C. and Guido, S. 2016).

R Squared gives an indication of how well a model fits a given dataset. It indicates how close the regression line (i.e. the predicted values plotted) is to the actual data values. The R squared value lies between 0 and 1 where 0 indicates that this model doesn't fit the given data and 1 indicates that the model fits perfectly to the dataset provided Iqbal, M. (2024).



**Fig. 26:** Evaluation results.

The negative R-squared value is concerning because it indicates the model performs worse, this suggests that the model may not be suitable for the data, and there might be issues with the features, the model choice, or how the data is prepared.

# Cross validation and Hyperparameter Tunning

Cross-validation is a statistical method of evaluating generalization performance that is more stable and thorough than using a split into a training and a test set. In crossvalidation, the data is instead split repeatedly and multiple models are trained. The most commonly used version of cross-validation is k-fold cross-validation, where k is a user-specified number, usually 5 or 10 (Müller, A.C. and Guido, S. 2016).

Because grid search with cross-validation is such a commonly used method to adjust parameters, scikit-learn provides the GridSearchCV , which implements it in the form of an estimator. To use the GridSearchCV, you first need to specify the parameters you want to search over using a dictionary. GridSearchCV will then per‐ form all the necessary model fits (Müller, A.C. and Guido, S. 2016).

Trying to optimize our machine learning model, we apply hyperparameter tuning and cross-validation to enhance performance and ensure the reliability of results. Hyperparameter tuning helps identify the best combination of model parameters, maximizing accuracy and efficiency. Cross-validation provides a robust method for validating model performance by splitting the data into multiple subsets, minimizing overfitting and delivering a more reliable estimate of how the model will perform on unseen data.

A white background with black text

Description automatically generated

**Fig.27:** Evaluation result with cross validation and hyperparameter.

Together, these techniques refine the model's predictive capabilities while offering greater confidence in the results. improve and get a better score, but for our investigation is still a poor result having still a R2 score negative.

Saying that we decided to explore additional models after observing poor performance from the initial one. By trying alternative approaches, we aim to find a more effective model that better fits the data and delivers improved results.

**Random Forest Regressor**

To build a random forest model, you need to decide on the number of trees to build (the n\_estimators parameter of RandomForestRegressor or RandomForestClassifier). Let’s say we want to build 10 trees. These trees will be built completely independently from each other, and the algorithm will make differ‐ ent random choices for each tree to make sure the trees are distinct (Müller, A.C. and Guido, S. 2016).

We used a time series split for cross-validation, and although the model achieved low Mean Squared Error (MSE) and Mean Absolute Error (MAE) scores, the negative R2 value indicates potential issues. A negative R2 suggests that the model may be poorly suited to the dataset, possibly due to overfitting or an inability to capture meaningful patterns in the data. This discrepancy implies that the model may fit closely to specific data points but fails to capture the underlying trend or seasonality, resulting in weak performance on unseen data.

****

**Fig.28:** Evaluation result.

**Random Forest Regressor with PCA**

**A white background with black text

Description automatically generated**

**Fig.29:** Evaluation result.

Here, we are using Principal Component Analysis (PCA) as a dimensionality reduction technique in combination with K-fold cross-validation to improve our model. However, as seen in the previous graphic, the results have not improved as expected. The reason for using PCA in this context is that Random Forest models, although powerful, can sometimes struggle with high-dimensional data, especially when features are highly correlated or redundant. PCA helps by transforming the original features into a smaller set of uncorrelated components, potentially reducing overfitting, improving computational efficiency, and helping the model generalize better. However, if the dataset does not have significant multicollinearity or if too much variance is lost during the PCA transformation, it might lead to a decline in performance.

**Random Forest Regressor with PCA**

**A computer code with numbers and letters

Description automatically generated**

**Fig.30:** Evaluation result.

The improvements in R2 and the reduction in MAE show that our model now captures patterns in the data more effectively and generalizes better. By using RandomizedSearchCV for hyperparameter tuning, we identified a more optimal set of parameters than those found through the previous GridSearchCV. This resulted in better predictive accuracy and a significant boost in R2

With Time Series Split for cross-validation combined with RandomizedSearchCV, we’ve developed a model that achieves a realistic R2 of 0.57, indicating it explains a reasonable portion of variance in future data. It also maintains low error metrics (MSE and MAE), providing predictions that closely match actual values. Additionally, by training only on past data relative to each validation set, we reduce the risk of overfitting and prevent data leakage, making our model more robust for forecasting unseen data. .

With the follow graphic we can understand better how our modal prediction, it provides a rough approximation of actual returns, performing decently in capturing trends but falling short during periods of high volatility. It suggests that while the model can guide general expectations, it may not be reliable for precise forecasting, especially when the market experiences rapid changes.

To reduce noise and highlight the underlying trends in the data, we applied a 5-period smoothing. By minimizing short-term fluctuations, this approach allows us to better observe and analyze the true patterns in returns, making it easier to assess model performance and identify meaningful discrepancies between actual and predicted values.

A graph of a graph showing a number of times

Description automatically generated with medium confidence

**Fig.31:** Actual vs Predicted plot.

Trying to support that our modal could have problem predicting during high volatility periods we make a residual plot which generally hover around zero, indicating that the model does not consistently overpredict or underpredict returns, which is a positive sign. The residuals appear random and lack any clear pattern, suggesting that the model's errors are unbiased and do not systematically trend in one direction over time. This randomness implies that the model isn’t missing any obvious, predictable patterns in the data. However, there are occasional spikes in the residuals, particularly during certain time periods, which highlight moments when the model struggled to accurately capture the dynamics of returns. These spikes suggest that the model may have difficulty predicting returns during periods of high market volatility, leading to larger prediction errors during those times.

A graph of a graph showing a number of times

Description automatically generated with medium confidence

**Fig.32:** Residuals plot.

# Conclusion

In this investigation, we explored the impact of weather conditions on stock market returns, focusing on factors like Relative Humidity and Visibility. Our analysis revealed that higher Relative Humidity is positively correlated with stock market returns, while increased Visibility tends to be associated with lower returns. These findings were reinforced by the results from the SARIMAX model, which accounted for both seasonal patterns and the influence of weather variables. However, when we used Random Forest to assess feature importance, weather conditions did not emerge as significant predictors of stock market behavior. This suggests that while weather may have some influence, other factor, likely play a more prominent role in shaping stock market performance.

Also, there were applied three distinct models—Linear Regression, SARIMA, and Random Forest Regressor with PCA—to predict and forecast stock market returns, using cross-validation and hyperparameter tuning to optimize performance. The Linear Regression model yielded promising results, with a high R² close to 1, indicating that it captured most of the variability in the data. Its low mean error suggests that the predicted values were close to the actual returns, making it suitable for short-term predictions. On the other hand, the SARIMA model indicated stability around zero with no strong trend, but the widening confidence intervals highlighted growing uncertainty over longer forecasting horizons. The negative R² suggested that the model performed poorly, likely due to issues with feature selection or the data preparation process, which affected its ability to predict accurately.

The Random Forest Regressor with PCA showed more promise, capturing more complex relationships between the features and returns. However, this model exhibited a higher mean error and an R² of around 0.57, indicating room for improvement. Residual analysis revealed issues during periods of high volatility, suggesting that the model may not be entirely reliable in turbulent market conditions.

Moving forward, further fine-tuning, feature engineering, and model adjustments are necessary to improve accuracy, particularly in high-volatility environments, and to better capture the influence of external factors like weather on market dynamics.

For future research, it would be valuable to expand the scope by incorporating a broader range of weather variables, considering longer time horizons, and experimenting with different machine learning models. These steps could provide a deeper understanding of the complex relationship between weather and stock market trends. Additionally, segmenting the analysis by geographical regions and incorporating external economic variables may help improve model accuracy and produce more targeted insights.

# References

* Brownlee, J. (2017). *Introduction to Time Series Forecasting With Python*. Machine Learning Mastery.
* Iqbal, M. (2024). ARIMA Times Series Model. Lecture 6 Week 1. CCT Moodle: CCT College.
* McKinney, W. (2018). *Python for Data Analysis Data Wrangling with Pandas, NumPy, and IPython*. [online] file:///C:/Users/Dell/Downloads/Python%20for%20Data%20Analysis.%20Data%20Wrangling%20with%20Pandas,%20NumPy,%20and%20IPython%20(2017,%20O%E2%80%99Reilly)%20(3).pdf. Available at: <http://oreilly.com/catalog/errata.csp?isbn=9781491957660>.
* McQuaid, D. (2024a). *Feature Scaling or Normalization*. file:///C:/Users/Dell/Downloads/Feature%20Scaling%20or%20Normalization%20(3).pdf.
* McQuaid, D. (2024b). *What is Exploratory Data Analysis?* file:///C:/Users/Dell/Downloads/Feature%20Scaling%20or%20Normalization%20(3).pdf.
* Müller, A.C. and Guido, S. (2016). *Introduction to Machine Learning with Python A GUIDE FOR DATA SCIENTISTS*. [online] https://www.nrigroupindia.com/e-book/Introduction%20to%20Machine%20Learning%20with%20Python%20(%20PDFDrive.com%20)-min.pdf. Available at: <http://safaribooksonline.com/>.

# GitHub link

<https://github.com/Rosma28/CA3_Strategic_Thinking>