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 2– Capstone Report. |
| *Assessment Due Date:* | 19/05/2024. |
| *Presentation Date* | 19/05/2024. |

**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](#_Toc167054084)

[Objectives 3](#_Toc167054085)

[Defining the Problem 3](#_Toc167054086)

[Scope 4](#_Toc167054087)

[Data Sources 5](#_Toc167054088)

[Ethical Considerations 5](#_Toc167054089)

[Develop Data Set 6](#_Toc167054090)

[Characterization of data and pre-processing 6](#_Toc167054091)

[Scaling and Normalization 10](#_Toc167054092)

[Training and Testing our Data Set 11](#_Toc167054093)

[Applying Modals 12](#_Toc167054094)

[Conclusion 16](#_Toc167054095)

[References 17](#_Toc167054096)

[GitHub link 18](#_Toc167054097)

**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>)

That said, this project seeks to offer for management the option to compare how prices developed during these events for decisions making.

# Objectives

1. Understand the relationship between grain price movements and weather events.
2. Predict future grain prices, assisting traders and stakeholders in decision-making.
3. Minimize risk in trading by predicting futures values.
4. Offer alternative solutions to producers developing price forecast.

# 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.

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 acknowledging the diverse influences of climate factors on both production and prices, we can strive to formulate sustainable solutions and strategies aimed at lessening the impact on the agricultural sector.

# 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 for both farmers and consumers to grasp the correlation between weather patterns and grains prices. Awareness of factors such as temperature variations, precipitation levels, occurrences of natural disasters, and the timing of seasons allows stakeholders within the agricultural sector to forecast and adjust to price fluctuations resulting from diverse weather conditions more effectively.

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:**

As this project spans two semesters where we must develop the following steps regarding complete it:

* Develop the project proposal.
* Find the necessary data.
* Work in cleaning data set it is necessary.
* Implement machine learning techniques.
* Develop conclusions and advice post results.

# 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.

A screenshot of a computer code

Description automatically generated



**Fig1: Display library to import data set and also other libraries which we will use in the future processing data.**

Regarding to combine two data set df1 and df2 we use the next function:

A screenshot of a computer

Description automatically generated

**Fig.2: Join “df1” and “df2”in just one data set called df.**

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.3: 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. 4: Display null values.**

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

A screenshot of a graph

Description automatically generated

**Fig.5: Display Histograms.**

To identify outliers in the features, we use boxplot

**A screenshot of a computer

Description automatically generated**

**Fig. 6: 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.7: 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.



**Fig. 8: Transform categorical to numerical.**

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.

A white rectangular object with black text

Description automatically generated

**Fig.9: Define X and Y to split data.**

# 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

**A screenshot of a computer program

Description automatically generated**

**Fig.10: Scaling data.**

# 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 11: 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

**Fig. 12: 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. 13: 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**

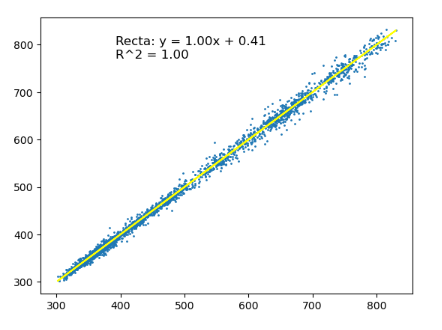
**A graph showing the difference between a graph and a graph

Description automatically generated with medium confidence**

**Fig. 14 and 15: Actual and predicted values.**

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

**The Linear Regression graphic.**

****

**Fig.16: Linear Regression modal.**

**Error**

**A screenshot of a computer

Description automatically generated**

**A blue graph with text

Description automatically generated**

**Fig. 17. Error graphic.**

**A computer error message

Description automatically generated**

**Fig.18: 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.

**A graph and chart with numbers

Description automatically generated with medium confidence**

**Fig.19: Important features by Random Forest.**

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

# Conclusion

After processing our data, performing Exploratory Data Analysis (EDA), and applying the algorithmic model, we can conclude that we achieved a promising score.

However, it is crucial to employ additional processing techniques to prevent errors like underfitting or overfitting, trying to avoid the noise in the training data rather than the underlying pattern, which can result in poor generalization to new, unseen data.

For future analysis, I recommend improving feature processing. Based on this analysis, I believe the "volume" feature is irrelevant to our research, as our goal is to predict prices influenced by potential climatic changes. This adjustment will help ensure that our model focuses on the most impactful variables related to climatic factors.

Also use other complementary metrics and diagnostic tools to evaluate the model, handling temporal and imbalanced data, scalability issues, preprocessing overhead, and potential interpretation difficulties.

# References

* 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/CA-2-Capstone-Report