Chapter 1: Introduction

*<after implementation/evaluation are finished>*

Chapter 2: Research Design

The proposed research design and methodology are presented in this chapter to cover the goals of this study. In the first section, it will display an overview of the data collection procedures followed by the identification and clarification of the problem that will guide the project to validate its relevance for the data science community and stakeholders involved. The research objectives and expected outcomes will be presented in the third section, followed by the ethical consierations that were observed during the process. *Among the goals of this research are: design, implement, and analyse different algorithms in which were found to be the most effective ones for this case study, based on the data available and its nature.*

Primary Data

The primary data for this study will be obtainded from the application of several time series forecasting models to historical values of coffee traded in the global commodity stock market based on the Composite Indicator Price (I-CIP) calculated and published by the International Coffee Organisation. To study the dynamics and to attempt a prediction of future movements of the i-cip prices in the global market based on the secondary data gathered from this organisation.

Experimentation with these models, and their different aspects characterize the main method of data collection for this project, as it generates new insights and predictive outcomes from the existing historical dataset. Similar to the Along with the forecasted values, the error results obtained from the modelling (mse, rmse, mea) are also considered as primary data in the context of this study since they represent a key new information used for measuring the performance.

Data Collection: Secondary Data

In order to obtain the primary data via experimentation, secondary data was collected from the International Coffee Organisation’s public database available at their website under the page “Public Market Information”(ICO, 2023). In this page, the daily indicator prices are published, as well as current observations on supply, demand and trade of coffee beans globally, such as the exports and imports of coffee per country. To shorten the scope of data collection for this reaserch, it was decided to only source at least one year worth of daily composite indicator prices (I-CIP), which is considered a benchmark for the coffee industry that takes an average of prices of distinct groups of coffee beans from different countries compiled to represent the sector’s situation. (among the data manipulation tecniques to be deployed on the secondary are: *data pre-processing, imputation of missing values, general eda combined to statistical test to check for normality and stationarity, data transformation- rolling average…?*)

Problem Identification and Clarification

Having the power to predict the future has been a longstanding quest across many areas of knowledge. One of the main challenges in forecasting future values is the high fluctuation and non-linearity characteristics of historical data, combined with the uncertainty and dynamics of . This research derives from this curiosity of looking into future values of one commodity that has been traded for centuries and occupies a solid position as a major player in the global market.

Comodity price research has been extensively conducted on various items such as copper, oil and sugar, as Faith (2021) points, however it was identified a certain gap in studies targeting the future prices of coffee regarding their composite index. When it comes to the coffee industry, there is a broad amount of studies focused on a more agricultural perspective by using machine learning and data science to predict and classify leaf diseases in coffee plants (Fatih, 2021) (De Oliveira Aparecido et al., 2020; Martinez et al., 2022), or to predict crop yields, for example. (Ansarifar et al., 2021) (Fianu, 2022)

This research takes from existing frameworks used to predict prices in different commodity markets and stock markets, employing time series models to analyse data provided by the International Coffee Organization’s Indicator Price (ICO-ICIP), which aggregates coffee prices from multiple markets, which provide a more stable, representative of the physical market conditiosn and less speculation and complex volatility as compared to the future prices/contracts from the Intercontinental Exchange (ICE). (Fatih, 2021; Hwase and Fofanah, 2021; Zhu, 2022).

Research Objectives

The core objective is to evaluate various time series models to determine which algorithm most effectively predicts future coffee prices. This endeavour not only seeks to contribute to academic knowledge but also aims to provide insights that could benefit economists, traders, and policy makers in the coffee industry in the short term. The research questions guiding this study include: What is the most effective time series model for predicting ICO’s I-CIP coffee prices? How can the accuracy in coffee price forecasts be optimized? (*do neural networks provide a superior forecasting capability compared to other machine learning models, specialluy in terms of accuracy and reliability across evolving market conditions?)*

The research objectives for this study are as follows:

* From the secondary data obtained from the international coffee organisation public database, perform data cleaning techniques to achieve a comprehensive dataset to identify patterns (such as overall trend and seasonality), statistical properties (like normality or stationarity of the data) through exploratory data analysis prior to modelling.
* Use feature engineering techniques to apply different predictive machine learning algorithms in time series field to forecast future prices of coffee trade based on the Composite Indicator Price (I-CIP) and the historical trends observed in exploratory phase. Considering a hybrid machine learning modelling, the intention is to compare how the changes in parameters impact the results in at least three different types of proposed models: Regression (linear regression), Autoregressor (such as SARIMA) and a deep learning approach (via Long-Term Short Memory) to visualise predictions for each model and compare the supervised learning from traditional statistical models to deep learning approaches and see which one provides more accuracy to the prices forecasting.
* Create an artifact to compare how different data preparantion and parameters adjustments can influence the forecasting results in the respective models and evaluate which predictions are the closest to accuracy, considering the different architecture and weights each model has, better results could be achieved through experimentation and hyperparameter tuning to understand and document how they behave with such changes. For measuring the performance of the forecasting of linear regression (LR), seasonal autoregressive integrated moving average (SARIMA) models and artificial neural network (ANN) there are three forecast performance measures that will be compared: Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) as based in literature are the key values to measure performance in machine learning.

In order to achieve these goals, this research uses an exploratory and descriptive approach with a literature review built to support the experimental process carried out.

Ethical considerations:

Given that the data used in this study are publicly available and aggregated without any personal identifiers, the main ethical consideration is ensuring the accuracy and reliability of the information presented. There is the acknowledgement that any misinterpretation of data, overstatements or even assumptions regarding the models' predictive competences could lead to misguided decisions by stakeholders relying on this research, or impact on the future studies that could derive from the insights presented here. In addition, is important to give the appropriate credits to the source of secondary data gathered for this study to the International Coffee Organisation and their commitment to distribuite accurate and reliable data.

*An internal application for ethics approval for preliminary research was completed and approved by the institution/thesis supervisor (add doc in appendix??)*

Chapter 4: Methodology

Data Collection of Secondary Data

One of the restrictions from accessing the data from this channel was that the ICO only makes public the data from the current month up until the same period from the year before, considering the time constraints present in undertaking the research, the secondatry data was collected from the period of ‘February of 2023’ to ‘February of 2024’ due to the availably of free data in ICO’s website .

An attempt to gather data from a longer period of at least 2-3 years (after the events of COVID-19) to further investigate the patterns and cyclical behaviours to increase training dataset was not possible due to the time sensitivity between making the request for additional data, having the risk being denied. This also added to a new risk of a possible delay to conclude the study, between collecting, processing the data and performing the experimentation artifact. To obtain further historical data, the ICO has a private plan for subscribing to their entire database which was not judged necessary at this stage due to financial constraints . a direct contact was made to enquire about access for academic purposes, which was met with a limitation that the data provided would not be kept at another’s institution database for data protection and ethical concerns, which is why it was stablished one year of data would be used for the final experiement collected from open source channels. This adjustments were kept to ensure the practical boundaries to execute the project.

Data Description

As mentioned in the previous chapter, the historical data utilized in this project comprises daily price records of the ICO Composite Indicator Price (I-CIP) collected from the International Coffee Organization's public database. This dataset includes coffee prices of four coffee beans groups, such as ‘Colombian Milds’, ‘Other Milds’, ‘Brazilian Naturals’, and ‘Robustas’ as well as the composite indicator price (I-CIP) compliling these four groups based on different weights. The period of data ranges from February 2023 to February 2024 and the data's comprehensiveness allows for an in-depth analysis of market trends, seasonality, and price volatility over a year’s time. It would be preferable to have a slight longer range, in order to observe more than one cycle to see with more detail the ups and downs of the prices over the years, as a general rule of thumb for time series is that the more historical data used, the more a forecast can be improved, as stated by Svolba (2022). However, in terms of fast changing and dynamic environments, such as the stock market/ commodities market, having to many years back in historical inputs could also make pose as a challenge for accurate forecasting because it could mask the recent history fluctuations, which are more relevant to models, other than relying on older data (for example, great economic recessions from past decades could influence the predictions by downgrading the overall average of prices)

The initial data was collected as available in the ICO’s database, month by month and the pre-processing techniques will be presented on the next section.

Data pre-processing

Once the secondary data was gathered, each csv file contained daily values of i-cip prices divided by month, in total there was 13 files stored in a random sequence and imported to a jupyter notebook. The first data pre-processing technique used was to combine all csv files into the same dataframe in order to advance the analysis, which had to me manually sorted to chronological order to respect the sequence (daily) needed to perfom any type of time series modelling. The dataset was then scanned to display basic features from heading (to observe names of columns, shape and data types) and identifying missing values.

Exploratory data analysis was executed at the early stage through simple visualisations to verify the presence of outliers and to better understand how the data is distributed via histograms and boxplots of values per month.

It was observed that not all months had the same features which caused a misplacing of values that had to be addressed in the processing stage, it was also observed that the data types were not in the correct form for applying time series. One of the most crucial characteristic of this kind of modelling is that the date should be as the index position and in the correct *datetime* as data type, followed by numeric values as independent variables. The dataset did not present any missing values at the first scaning (from the 279 observations no null values were identified), however to follow the time series principles, besides having the correct datatypes, there cannot be any missing dates despite the frequency of each case. For the ICO’s data, it shows a weekly frequency based on business days, with data published from Monday to Friday, meaning weekends and holidays values are not included in the original calculations. By comparing the weekdays and business days present in the range between feb23 and feb24 and the ones expected to have in the dataset, three dates were perceived as missing from the desired sequence to respect the modelling requirements, which were added via data imputation after comparing three different techniques: forward fill (uses previous data to fill null values), backward fill (fill missing value with the next datapoint) and linear interpolation (gets the average between 2 points adjacent to the missing value).

Linear interpolation was the one elected to fill the null values from the new dates added, despite all methods displaying a similar curve, the linear interpolation has a straighfoward approach and helps to mantain the overall trend, and it’s use is also indicated when the missing values are in the middle of the dataset instead of the extremities to avoid bias (Koech, 2022).

* Normalisation/standartisation of data was made to ensure that the different scales would not affect or mislead the comparative analysis as well as minimising the effects of the variance/volatility. In addition, a separate dataset was created with the standardised i-cip prices to be compared with the original values.
* Once identified the non-stationarity present via statistical tests (adf/knss), data transformation was applied as attempt to make the dataset stationary and enable a more precise modelling. So differentiation was applied to stabilize the variance and mean of the time series, and also compared to the original patterns found in the i-cip prices prior to applying seasonal decomposition of data.

*( Add more features, describe final dataset after processing and models used?)*

Statistical Tests

-Adf/knsss for stationarity combined with visuals of seasonal decomposition made possible to understand the original data is not stationary and presents both trend and seasonality

- from Shapiro-wilk test for normality combined with visualisations of distribution it was identified a slightly skewed pattern.

- Granger Causality test to estimate forecastbility (calculates the prediction of how easy it is to forecast next values) For GC test data must be stationary. If the data series have trends or unit roots, the results of the Granger causality tests can be misleading, requering further investigations. Based on the relatively high p-values for each of the lags analysed (0.38 and p=0.46), there is no evidence to reject the null hypothesis, and it can be said that there is no statistical evidence of Granger causality from the tested lags of the first time series to the second time series.

- <to be added in this section considering interpretations from EDA>

Machine Learning

Model Selection

Model Evaluation

Limitations and Considerations

[*limitations that resulted from the chosen methods and other considerations to have interpreting the results of the analysis*.]

### References

Ansarifar, J., Wang, L., Archontoulis, S.V., 2021. An interaction regression model for crop yield prediction. Sci Rep 11, 17754.

De Oliveira Aparecido, L.E., De Souza Rolim, G., Da Silva Cabral De Moraes, J.R., Costa, C.T.S., De Souza, P.S., 2020. Machine learning algorithms for forecasting the incidence of Coffea arabica pests and diseases. Int J Biometeorol 64, 671–688.

Fatih, C., 2021. Estimating and forecasting Volatility in Coffee Prices: Evidence of ARIMA-GARCH models.

Fianu, E.S., 2022. Analyzing and Forecasting Multi-Commodity Prices Using Variants of Mode Decomposition-Based Extreme Learning Machine Hybridization Approach. Forecasting 4, 538–564.

Hwase, T.K., Fofanah, A.J., 2021. Machine Learning Model Approaches for Price Prediction in Coffee Market using Linear Regression, XGB, and LSTM Techniques. IJSRST 10–48.

ICO, 2023. Public Market Information | International Coffee Organization [WWW Document]. URL https://icocoffee.org/resources/public-market-information/ (accessed 12.3.23).

Koech, D.K., 2022. A Complete Guide on How to Impute Missing Values in Time Series in Python [WWW Document]. Engineering Education (EngEd) Program | Section. URL https://www.section.io/engineering-education/missing-values-in-time-series/ (accessed 11.19.23).

Martinez, Fredy, Montiel, H., Martinez, Fernando, 2022. A Machine Learning Model for the Diagnosis of Coffee Diseases. IJACSA 13.

Svolba, G., 2022. Determining the best length of the history of your timeseries data for timeseries forecasting. MLearning.ai.

Zhu, Y., 2022. Stock price prediction using the RNN model. In: 2022 International Conference on Applied Physics and Computing (ICAPC). Presented at the 2022 International Conference on Applied Physics and Computing (ICAPC), IEEE, Ottawa, ON, Canada.