**CPF Diageo Documentation**

**Contents**

1. Problem in hand
2. Business Requirement
3. Data Requirement
4. Data Collection
5. Data Understanding
6. Data Preparation

* Data One view
* Data transformation
* Exploratory Data Analysis

1. Independent Forecast
2. Feature Selection

* Variable Importance
* Multicollinearity

1. Model Building
2. Model Evaluation

**Problem In hand:**

Our objective is to forecast the price of commodity (broken rice) for next 3 months based on past data. Detailed Information about the project is pointed out below:

|  |  |
| --- | --- |
| **Client** | Diageo |
| **Commodity** | Broken rice |
| **Mandi covered** | (Gondia, Gangavathi, Burdawan, Nizamabad) |
| **Duration** |  |

**Business Requirement:**

With an amazing array of premium brands, Diageo is India's leading beverage alcohol firm. It produces best-in-class beverages with broken rice as its raw material. The purchase of broken rice is one of the company's major expenses because changes in commodity prices have a significant impact on the firm's revenue, the company wants to know the commodity's future price index in order to determine the optimum procurement and inventory schedule. Diageo wishes to utilize an analytical technique to determine the elements that influence the price of broken rice in the market, such as export/import, competition, raw material price, machinery, substitute products, by products, exchange rates, etc.

**Data Requirement:**

The price of broken rice (dependent variable) in the market is influenced by several factors (independent variables). We classified them into various categories as macroeconomic factors, trade, price, weather, and other independent elements. To figure out how these factors affect the BR price, we'll need historical data on them. Categories and included variable list are shown in below table:

|  |  |
| --- | --- |
| **Category** | **Variables** |
| **Price**  *(Competition Commodity Prices,*  *Maize, Paddy)* | Vietnam broken rice price, Thailand broken rice price, etc. |
| Indian Broken rice price, Paddy price, etc. |
| **Macro**  *(CPI, WPI, labor wages, Inventory level)* | Urea, dap, injection pumps, Agricultural machinery, rice mill machinery, Molasses, Rice products etc. |
| Inventory level rice stock, MSP paddy. |
| **Other**  *(Fuel price, Substitute Crop,**Exchange rates, Futures)* | Diesel price, Maize price, etc. |
| USD INR exchange, Futures rough rice investing, etc. |
| **Trade**  *(Export & Import)* | Broken rice import, Broken rice export, etc. |
| **Weather** | Day length, Average temperature, Average humidity. |

**Data Collection:**

Data collection is one of the major steps in project as entire statistical analysis and business decisions are data driven. We conducted extensive workshops with Diageo to get an understanding of what factors can influence the commodity. After discussion on the same, Diageo provides data for some independent factors, while the remainder is gathered from authorized and credible data sources following extensive research and study of the variables by the Deloitte team. The following table provides a description of these factors and data sources:

|  |  |  |
| --- | --- | --- |
| **Subcategory** | **Variable Description** | **Data Reference** |
| Price | Broken Rice | Directorate of Marketing & Inspection (DMI), Ministry of Agriculture and Farmers Welfare, |
| Price | Vietnam Broken Rice | World Bank |
| Price | Thailand Broken Rice | World Bank |
| WPI | Paddy | Reserve Bank of India |
| CPI | Fuel and Light | Reserve Bank of India |
| CPI | Transportation and communication | Reserve Bank of India |
| GDP | Agriculture | Reserve Bank of India |
| MSP | Paddy | Food Corporation of India |
| Exchange | Exchange rate with INR to major import/export countries | Investing |
| Labour wage | Average Daily Wage Rates (in Rs.) in Rural India for Men | Reserve Bank of India |
| Rainfall | Rainfall | India Meteorological Department |
| Inventory level | Rice Stock in Central Pool | Food Corporation of India |
| Temperature | Average temperature | World Bank |
| Day length | Length of Daylight | Time and Date |
| Futures | Rough Rice | Investing |
| Price | Paddy Price | Directorate of Marketing & Inspection (DMI),Ministry of Agriculture and Farmers Welfare, |
| Crude Oil | Crude oil Price | Investing |
| Diesel | Diesel used in Machinery | Index Mundi |
| Substitute Crop- Maize | Maize Price | Index Mundi |
| imports | Broken Rice imports- India | Ministry of commerce and Industry |
| exports | Broken Rice exports- India | Ministry of commerce and Industry |
| Covid positivity rates | Covid positivity rates in Indian states | Covid-19 India |
| Food Inflation Index | Food Price Index | FAO |
| Fertilizer WPI | Urea | Reserve Bank of India |
| Fertilizer WPI | Di ammonium phosphate | Reserve Bank of India |
| Fertilizer WPI | Ammonium sulphate | Reserve Bank of India |
| Fertilizer WPI | Pottasium Chloride | Reserve Bank of India |
| Fertilizer WPI | Nitrogenous fertilizer, others | Reserve Bank of India |
| Fertilizer WPI | Ammonium phosphate | Reserve Bank of India |
| Fertilizer WPI | Superphospate/Phosphatic fertilizer, others | Reserve Bank of India |
| Fertilizer Machinery WPI | Injection pump | Reserve Bank of India |
| Irrigation WPI | Pneumatic tools | Reserve Bank of India |
| Irrigation WPI | Water pump | Reserve Bank of India |
| Irrigation WPI | Centrifugal Pumps | Reserve Bank of India |
| Agricultural Machinery WPI | Manufacture of agricultural and forestry machinery | Reserve Bank of India |
| Rice Mill Machinery WPI | Rice mill machinery | Reserve Bank of India |
| Raw Material WPI | Rice products | Reserve Bank of India |
| Raw Material WPI | Manufacture of starches and starch products | Reserve Bank of India |
| Substitute WPI | Molasses | Reserve Bank of India |
| Substitute | Maize Price | Reserve Bank of India |
| Imports | Maize/Maize seed imports- India | Ministry of commerce and Industry |
| Exports | Maize/Maize seed exports- India | Ministry of commerce and Industry |
| NCDEX | Maize | Investing |
| Humidity | Average humidity | Time and Date AS |
| Temperature | Average temperature | Time and Date AS |

**Data Understanding:**

It is preferable to comprehend the data before statistically analyzing it. Obtaining information on each independent variable allows us to draw specific conclusions about the impact of each variable in our model. Below is a list of variables and their impact on commodity.

|  |  |  |
| --- | --- | --- |
| **Category** | **Variables** | **Understanding** |
| Competition Commodity Prices | Vietnam broken rice price, Thailand broken rice price, etc. | Vietnam and Thailand are key competitors in the broken rice trade, the pricing of BR in these countries will probably affect the price of BR in India. |
| Paddy Price | Indian Broken rice price, Paddy price, etc. | Broken rice is byproduct of Paddy, so impact of paddy price is obvious to Broken rice. |
| Fertilizers and Machinery | Urea, dap, injection pumps, Agricultural machinery, rice mill machinery, etc. | Fertilizers and machinery are used for better production of paddy crop so change in their prices will impact commodity price |
| Substitute product | Maize and Molasses | Maize and molasses are used as substitute of broken rice in breweries so change in their price will impact. |
| Raw Material | Starch and Rice Products | Usually broken rice (brewer's rice) is used for starch production. (Change in price index of final product will impact the price of raw material used for its production and vice versa) |
| Inventory and MSP | Inventory level rice stock, MSP paddy. | MSP issued by GOI and stock in inventory affects the procurement policy of a company. |
| Fuel price | Diesel price, Crude oil price, etc. | Fuels are used for transporting and other purpose so they can impact end price of broken rice in market. |
| Exchange rates, Futures | USD INR exchange, Futures rough rice investing, etc. | Currency exchange rates can impact merchandise trade, economic growth, |
| Procurement | Production and arrival of broken rice in Mandis | Quantity of broken rice produced and arrived in a particular mandi will impact price of BR locally. |
| COVID | COVID positivity rates in Indian states | Impact of pandemic on Paddy producers. |
| Export & Import | Broken rice import, Broken rice export, etc. | Export/ import of a particular commodity directly affects its local market price. |
| Weather | Day length, Rainfall, Average temperature, Average humidity. | Weather is an important aspect in crop production, so it is important to include weather variables in our data. |

**Data Preparation:**

The process of taking raw data and preparing it for ingestion into an analytics platform is known as data preparation. The data must be cleansed, structured, and transformed into something that analytics tools can understand to analyze the data properly. The actual procedure might comprise a wide range of steps, such as consolidating/separating fields and columns, treating null values and outliers, changing formats, deleting unnecessary or junk data, and making corrections to data. In this project we have prepared the data in efficient manner and end to end process is explained in three major steps below:

1. **Data One View:** As previously stated, we obtained some data from the Diageo team and the rest was gathered by the Deloitte team after extensive research. We acquired data from several sources and collated it as per our requirements. We used simple methods like aggregation (converted weekly and daily data to monthly data) and formatting to prepare a data one view which comprises of all the above-mentioned variables in the desired format.

1. **Data Transformation:** After collating the data together, we have developed some new features which can aid us in better understanding and delivering better results. In this manner we constructed lag (some independent variables will have impact on dependent at a later point of time e.g. rainfall), logs, ratios and difference of the variables. Many variables, such as WPI paddy, rainfall, dependent, and others, had null values that were imputed using the backward fill or forward fill methods. Various outliers in data are effectively managed using the inter quartile range technique.
2. **Exploratory Data analysis:** The primary goal of EDA is to assist in the analysis of data prior to making any assumptions. It can aid in the detection of evident errors, as well as a better understanding of data patterns, the detection of outliers or unusual events, and the discovery of interesting relationships between variables. To determine the trend, seasonality, and patterns in all the variables, we used univariate and bivariate visualizations. We also conducted correlation analysis to determine how these variables are related to each other.

Various events which led to price fluctuations in the dependent variable were identified which gave us thorough understanding of what factors could impact the prices. To capture any aberrations, various smoothening techniques have also been used.

**Independent Forecast:**

In our project, we not only predicted the dependent variable, but also predicted our independent variables to train our model so that it may detect unobserved fluctuations and patterns in the values of independent variables. To forecast these factors, we employed a variety of methods, which are detailed below:

1. **ARIMAX:** ARIMAX is an Auto Regressive Integrated Moving Average with Explanatory Variable.This method is suitable for forecasting when data is stationary/non stationary, and multivariate with any type of data pattern, i.e., level/trend /seasonality/cyclicity.
2. **UCM:** UCM stands for Unobserved Components Model,decomposes a time series into trend, seasonal, cyclical, and individual components and allows for exogenous variables. UCM is an alternative to ARIMA models and provides a flexible and formal approach to smoothing and decomposition problems.
3. **Holt Winter’s Method:** Holt-Winters uses exponential smoothing to encode lots of values from the past and use them to predict values for the present and future value by computing the combined effects of trend, value, and seasonality.
4. **Holt’s Linear Method:** Holt's linear trend method is a valuable extension of exponential smoothing that helps deal with trending data. It eliminates the initialization issue faced when fitting the original model and simplifies the optimization process.
5. **Moving Average Method:** The moving average is a statistical method used for forecasting long-term trends. The technique represents taking average of fixed number of items in the time series which move through the series by dropping the top items of the previous averaged group and adding the next in each successive average.

We examined all these model results and determined that the model with the lowest MAPE throughout a validation period is the best model, and only the forecasts provided by the best model are used for independent prediction.

**Feature Selection:**

Feature selection methods involve assessing the relationship between each independent variable and the dependent variable using statistics and selecting those input variables that have the strongest relationship with the dependent variable. It is desirable to reduce the number of independent variables to both reduce the computational cost of modeling and, in some cases, to improve the performance of the model. As a result, deciding on a suitable strategy for feature selection might be difficult. For this, we used two methods, which are explained below:

1. **Variable Importance:** Feature importance scores play an important role in a predictive modeling project, including providing for [dimensionality reduction](https://machinelearningmastery.com/dimensionality-reduction-for-machine-learning/) and [feature selection](https://machinelearningmastery.com/rfe-feature-selection-in-python/) that can improve the efficiency and performance of a model on the problem.Variable importance refers to a technique for assigning scores to independent features to a predictive model that indicates the relative importance of each variable when making a prediction. The relative scores can highlight which variables may be most relevant to the dependent, and the converse, which variables are the least relevant. In this project, We used various decision tree approaches such as random forest, gradient boosting methods and Light gradient boosting method for assigning relative importance scores which is based on the reduction in the criterion used to select split points, like Gini or entropy.
2. **Multicollinearity:** Multicollinearity is the state where two variables are highly correlated and contain similar information about the variance within a given dataset. To detect collinearity among variables, simply create a correlation matrix and find variables with large absolute values (VIF). if the VIF is between 5–10, multicollinearity is likely present, and we should consider dropping the variable.

In this scenario, we first pick variables by calculating relative variable importance using various tree algorithms, then we eliminate variables with high VIF values (greater than 5) and conclude the subset of variables to be used in our model.

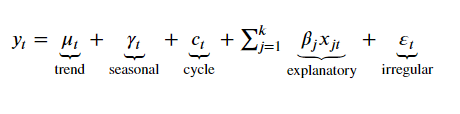
**Model Building:**

Our dataset comprises dependent values from July’18 to October’21. To keep all variables in a comparable range, we used the minmax scaler to scale our dataset. After that, we divided the data into training (July’18 to March’21) and OOT periods (April’21 to October’21), and then fed the selected variable list into the models. We are now ready to forecast the dependent value for the OOT period, and we applied two machine learning algorithms to do so, as described below:

1. **UCM:** UCM stands for Unobserved Components Model, used to derive the stylized facts of business cycle. This algorithm specifically decomposes the time series into different components such as Trend, Seasons, Cycles, and regression effects due to predictor series.

The key components of the model are:

* **Trend:** Defined as the natural tendency of the series to increase/ decrease/ remain constant over a period in absence of any other influencing variable.
* **Seasonality:** A seasonal pattern exists when there is a consistent pattern of variation influenced by seasonal factors (e.g., month of the year, or day of the week, etc.)
* **Cyclicity:** The cyclical component is intended to capture cyclical effects at time frames much longer than captured by the seasonal component
* **Regression effects:** Impacts observed on predictor series due to explanatory variables
* **Irregular:** White noise error term



1. **ARIMAX:** ARIMAX is an Auto Regressive Integrated Moving Average with Explanatory Variable. It explicitly caters to a suite of standard structures in time series data, and as such provides a simple yet powerful method for making skillful time series forecasts.  
   The key aspects of the model are:

* **AR:** *Autoregression*. A model that uses the dependent relationship between observation and some number of lagged observations.
* **I:** *Integrated*. The use of differencing of raw observations (e.g. subtracting an observation from observation at the previous time step) to make the time series stationary.
* **MA:** *Moving Average*. A model that uses the dependency between an observation and a residual error from a moving average model applied to lagged observations.

**Model Evaluation:**

Model evaluation is the process of determining which model is better. Therefore, it is critical to consider the model outcomes according to every possible evaluation method. Various metrics are used to compare models and determine which model is the best. We used two of these metrics, and they are listed below:

1. **Accuracy:** Accuracy refers to the closeness of a measured value to a standard or known value.

**Accuracy = (100 – MAPE) %**

1. **MAPE:** Mean absolute percentage error (MAPE) is a measure of how accurate a forecast system is. It measures this accuracy as a percentage and can be calculated as the average absolute percent error for each time period actual values divided by actual values.

**Error (E1) % = (F1 – A1) / A1**

Where A1 represents actual, F1 represents forecast and E1 represents error for month1.

**MAPE = Average of E1%, E2% & E3%**

Below are the model summaries for both models:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Rundate** | **Variables** | **coef** | **std err** | **p-value** | **MAPE** |
| 3/31/2021 | starch\_products | 0.42 | 0.19 | 0.03 | 0.60 |
| 3/31/2021 | futures\_rough\_rice\_lag1 | 0.14 | 0.14 | 0.31 | 0.60 |
| 3/31/2021 | arrival\_paddy\_burdawan\_lag1 | 0.08 | 0.08 | 0.31 | 0.60 |
| 3/31/2021 | inventory\_level\_unmilled\_paddy\_lag1 | -0.06 | 0.18 | 0.75 | 0.60 |
| 3/31/2021 | paddy\_price\_lag2 | -0.09 | 0.28 | 0.76 | 0.60 |
| 3/31/2021 | broken\_rice\_exports\_lag1 | -0.10 | 0.14 | 0.46 | 0.60 |
| 3/31/2021 | rice\_mill\_machinery | -0.06 | 0.07 | 0.41 | 0.60 |

**Summary of results For Burdwan Mandi Using ARIMAX**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Rundate** | **Variables** | **coef** | **std err** | **p-value** | **MAPE** |
| 3/31/2021 | starch products | 1.17 | 0.14 | 0.00 | 1.46 |
| 3/31/2021 | Paddy\_price\_lag2 | 0.41 | 0.17 | 0.02 | 1.46 |
| 3/31/2021 | inventory\_level\_unmilled\_paddy\_ lag2 | -0.29 | 0.13 | 0.03 | 1.46 |
| 3/31/2021 | Maize\_index | -0.31 | 0.17 | 0.06 | 1.46 |
| 3/31/2021 | arrival\_paddy\_burdwan\_lag2 | 0.09 | 0.15 | 0.52 | 1.46 |
| 3/31/2021 | Ammonium\_sulphate\_lag3 | -0.05 | 0.14 | 0.69 | 1.46 |

**Summary of results For Burdwan Mandi Using UCM**

Afterwards, we compare both the models on the basis of MAPE and select the model with least MAPE for OOT period. As per the summary for both models for burdawan mandi, it is clear that only the UCM model's p-values are within the specified range of 0 to 0.6. Consequently, variables are only significant for the UCM model.

We find that UCM is considerably superior for all mandis. UCM also has a better interpretability than ARIMAX.