TEAM 223

Forecasting ISA's Stock Prices

July 2022

Contents

List of Tables

List of Figures

1 Business Problem

1.1 Overview of the company

ISA, a company of the Ecopetrol Group, is a multi-Latin company with more than 54 years of experience and trajectory operating in the Electric Energy, Roads and Telecommunications and ICT businesses, which contributes to the quality of life of millions of people in Colombia, Brazil, Chile, Peru, Bolivia, Argentina and Central America, through the work of its more than 4000 employees in its subsidiaries.

According to the 2021 integrated management report of GRUPO ISA, the electric power business is one of the most important business, it represented 78 percentage of the company's income for that year, and it is present in 398 municipalities through a network in operation with more than 43.000 km which positions it as the largest transmitting agent in high voltage at the national level.

Within the international electric energy sector, through its affiliates and subsidiaries, expands, operates and maintains high voltage power transmission systems that contribute to the competitiveness of countries and ensure high standards of quality and reliability to users. It is currently the largest international electricity transporter in Latin America.

In Colombia, ISA was incorporate in the city of Medellin(Colombia) as mixed public utilities corporation of commercial character, national order, and governed by 142 and 43 laws of 1994 and was called Interconexión Eléctrica S.A. (ISA S.A. E.S.P.). As of December 2021, ISA's shareholder composition was as follows: 51% Ecopetrol, 9% EPM and 40% private investors.

1.2 Defining the Problem

In 2004, ISA entered in the stock market with public and private investors. The ISA's stocks and bonds are traded in the Colombian Stock Exchange(symbol:ISA) and has American Repository Receipt (ADRs) Level 1 that are traded in the Americas Over-the-Counter (symbol:IESFY). All stocks that have public capital, have the same rights, they are ordinary, normative and dematerialized, there are not statutory restrictions on their transferals.

The price of the ISA share is a key element to evaluate company value and ensure its economic solvency. It fluctuates depending on internal and external factors that increase volatile and risky such as:

- The supply or the number of shares available, - The demand or the amount that is bought and sold at each price, - Market's environment (political, regulatory, legal and social factors), - Macroeconomics trends (GDP, PPI, CPI,Exchange Rate or Interest Rate), - The industry of energy sector, - Profits and business management of the company, - And "irrational" or completely random factors.

Given the above, stakeholders seek to monitor, analyze, evaluate, and estimate the price behavior, identifying in-depth the relationship with the environment, and analyzing the factors that influence the company's share price, in order to allow them to anticipate impacts, make decisions before important changes, have good risk management and palliate market risk.

However, estimating the price of stocks is not easy enough. Doing a linear or non-linear regression on the future value of these asset prices can lead to models that appear to work on historical data but actually have no predictive value. Regression assumes that an observed variable is related to predictor variables but random walks or integrated moving averages usually intervene in this scenario. This means that the final value will be the sum of independent random numbers. Even so, a monitor with a baseline forecast or naive forecast is a first approximation tool to keep track of prices that informs possible considerable daily changes.

1.3 Potential Audience

Sustainability at ISA and its companies is a business approach that allows for managing opportunities, impacts, and the comprehensive risk system and building reputation, in order to create value for stakeholders, maintain its competitive advantage and contribute to the development of the societies where it is present.

ISA's stakeholders are workers, the State, investors, suppliers, customers, and society, with whom the company shares common interests. ISA seeks long-term relationships with its stakeholders that promote a growth framework for them.

However, the result of the share price analysis may be a more valuable indicator for stakeholders related to stock transactions in the Colombian stock market, especially those related to the energy sector.

2 Exploratory Data Analysis

2.1 First glance at the Datasets

The table 1 presents different data sets that allow the analysis to be performed. The dataset provided by ISA involved stocks from the Colombian Stock Exchange, a forex market symbol, and macroeconomics indexes. In order to broaden the perspective of the analysis, we explored those metrics that influence the share price, however, the project was focused on stocks from the Colombian equity market. The relevance of these data will be evaluated in the following stages of the project.

Table 1: Dataset identified to provide an answer on ISA's stock price

Name	Reference	Description	Subject	Key Words
Local Stocks Prices	(Bolsa de Valores de Colombia, 2022b)	•Daily prices for each local stock since 2017-06-01 to 2022-07-01 in the Colombian Stocks Exchange •Symbol, date, quantity stocks, volume, close price, highest price, mean price, lowest price, relative and absolute changes variables. •Missing values for highest price and lowest price in some stocks	Predict price using technical indicators	Equity market, Colombian stocks exchange, Local Market, Daily Prices
ISA Stocks Prices	(Bolsa de Valores de Colombia, 2022a)	•Daily prices for ISA's stock since 2003-01-01 to 2017-05-31 in the Colombian Stocks Exchange •Symbol, date, quantity stocks, volume, close price, highest price, mean price, lowest price, relative and absolute changes variables.	Predict price using technical indicators	Equity market, Colombian stocks exchange, Local Market, Daily Prices

Table 1 continued from previous page

Stocks and Indexes	ISA Company	•Daily prices for stocks and indexes relevant for ISA since 2006-04-25 to 2022-05-23 in the Colombian Stocks Exchange •Symbol, date, open price, volume, highest price, lowest price, relative change variables. • Symbols include Petroleowti, Ecopetrol, USD-COP, Celsia, ISA, GEB, COLCAP, T1y, T4y, T5y, T10y, T15y.	Predict price using technical indicators	Equity market, Colombian stocks exchange, Local Market, Daily Prices, Dolar Index, TES

Local stock prices: It is a dataset containing the behavior of the shares of 29 indexes in a time range from June 1, 2017 to July 01, 2022. Through 34.346 records and 9 columns it is possible to know the historical information of the stock market behavior.

ISA share prices: It is a dataset detailing the movement and performance of ISA shares in the stock market over a historical period from July 3, 2003 to February 25, 2022. The data has 4508 records and 10 columns.

Stock Indices: It is a dataset with 43.736 elements provided in an xlsx file that details in 13 sheets of the file the price of the indices that possibly affect ISA shares. In each sheet the data is distributed in 8 columns. The data in this dataset has a time range from April 24, 2012 to May 23, 2022.

2.2 Data preparation

The 3 datasets, Local Stocks Prices, ISA Stocks Prices and ISA dataset were integrated into a single dataset to combine the different records into common variables. All data-sets representing time serie with prices and details of daily transactions were merged. As a result we obtained a data set of 84589 records in 11 columns: Date, Index (symbol), closing price (close),open price (open), mean price (mean), maximum price (max), minimum price (min), variance absolute (varab), variance relative (var), quantity of trade stocks(quantity), and volume in COP(vol). Following subsections describe the process.

2.2.1 Data Cleaning

Format ISA dataset was provided and stored as excel files (XLSX) using the international standard Unicode UTF-8. Each sheet is a symbol, some of them in COP and others in dollars., it was consolidated in one dataframe using a function. The other two datasets were provided and stored as comma-separated values files (CSV).

Consistency

- Data type consistency. Each date was transformed into a data-time format and assigned as an index, remaining columns were converted to text (object) or float format accordingly.
- Unit consistency. Another transformation was the interpretation of the engineering format units to float data type. This data type has a decimal notation, followed by a letter that represents the units of the data, among the units found in the data, were the letters M and K. As values are in the same units, for example, "Petroleo WTI", "TES 1año", "TES 4años", "TES 5años", "TES 10años" and "TES 15años" columns are valued in dollars. These values were adjusted with the time serie of the dollar price in pesos, using the closing value as a reference.

- Categorical consistency. The column "symbols" was standardized according to a unique key, for instance, "TES 1año" was transformed to "T1Y". Spaces had been deleted and transformed to upper case.
- Column name consistency. Finally, the equivalent titles were unified in order to perform an adequate data base merge.
- Merging. Finally, the three datasets were consolidate as one dataframe using the columns.

Relevance

- Differences in temporality. The excel file provided by ISA contains thirteen spreadsheets. Of these spreadsheets the EMBI was not used since it contains the Emerging Markets Bonds Index. This exclusion was made because these data did not have the same temporality as the other spreadsheets.
- Duplicates. After an analysis of the duplicates, dataframe drop duplicates, letting the first value found.
- Nulls and 0 values. First, Min and max was replaced it for the close value. Second, Open was replace for the close value of last day. Third, Volumen was replace for the quantity multipling close and quantity for the division between volumen and close. Finally, Variation an absolute variation was calculated using the difference between the close price of the one day before.
- Missing dates. There are no operative days for Colombian Exchange. Missing dates had been identified using a function to subtract those no operative days. Only 5 days left from 2005 and 2004, but those data can't be found in bvc.com.co.
- Dropping data. Mean price was dropped due to the number of missing data, also no operative days and close price lower or equal to 0 was deleted.

2.2.2 Feature Engineering

Technical indicators improve the analysis of the price action, those variables depended of the close price in accumulative daily periods, including: Exponential Moving Average (EMA), Volatility, Average Directional Movement index(ADX), Moving Average Convergence Divergence(MACD), Relative Strength Index (RSI) and Bollinger Bands.

Regarding the mean, The Wavelet Filter, for an approximation of the price deviation from the mean (Sprint Effect), and The Box-Cox Transformation, to put out data that closely resembles a normal distribution, were included in the dataset.

Finally, an ISA's time serie decomposition of the trend, stationary and residual componets was calculated from the Box-Cox transformation added to describe the behavior of the price.

Describing result dataset

The final dataset results in a table with the following characteristics: DatetimeIndex: 74338 entries, 2003-07-03 to 2022-07-01, and Data columns: (total 33 columns). Table 2 Consolidated Dataset shows the composition of each of the columns.

Table 2: Dataset consolidated

#	Column	Non-Null	Count	Dtype
0	symbol	74338	non-null	object
1	last	74338	non-null	float64
2	max	74338	non-null	float64
3	mean	74338	non-null	float64
4	min	74338	non-null	float64
5	var_ab	74338	non-null	float64
6	var	74338	non-null	float64
7	quantity	51112	non-null	float64
8	vol	51112	non-null	float64
9	open	74338	non-null	float64
9	market	74338	non-null	object
10	id_market	74338	non-null	int64
11	vol_million	51112	non-null	float64
12	range	74338	non-null	float64
13	ema9	74338	non-null	float64
14	ema12	74338	non-null	float64
15	ema26	74338	non-null	float64
16	ema55	74338	non-null	float64
17	ema100	74338	non-null	float64
18	volatility1	74262	non-null	float64
19	volatility2	74224	non-null	float64
20	wavelet	64306	non-null	float64
21	-dm	74338	non-null	float64
22	+dm	74338	non-null	float64
23	adx14	74338	non-null	float64
24	macd	74338	non-null	float64
25	rsi14	74338	non-null	float64
26	transformed_close_boxcox	74338	non-null	float64
27	lambda_close_boxcox	74338	non-null	float64
28	dec_trend_365	60506	non-null	float64
29	dec_seasonal_365	74338	non-null	float64
30	dec_residuals_365	60506	non-null	float64
31	sma20	73616	non-null	float64
32	bollinger_up	73616	non-null	float64
33	bollinger_down	73616	non-null	float64

Figure 1 below shows some null data between columns. Describe it as follow:

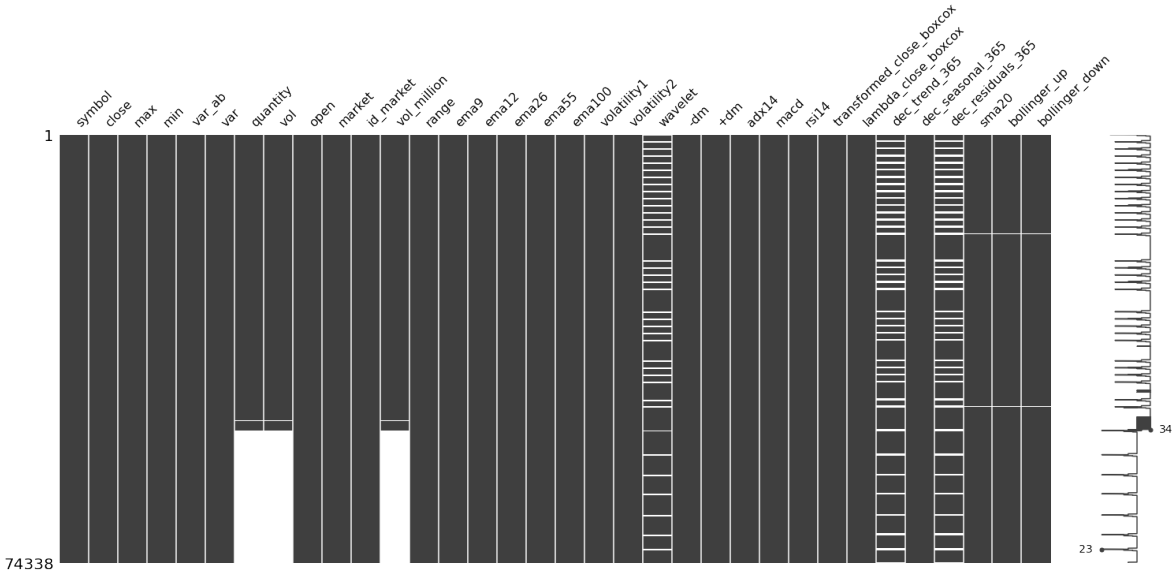


Figure 1: Nulls in dataframe

Therefore, the amount of nulls in Quantity and Vol correspond to the symbols TES or Public Debt Securities because there aren't quantities report, and some spaces in the indicators are necessary for their calculation.

2.2.3 Exploratory Data Analysis

Colombian Equity Market Context

Figure 2 shows the summary statistics of the market in last 5 years. The figure shows that the market has been stocks from 4 COP to 106.000 COP with a mean of 14.617, the maximum volume in COP was 327.000 million COP, the quantile 75 of variation of the market is 0.5%. ISA's stock price is higher than quantile 75 of the prices, its highest variation during these years was 12.9% and the lowest was -31.7%. Its stock price has a relevant part of the volume of the market; it exceeds the mean of the market. Finally, the maximum value of the stock was 29.250 COP and the lowest was 11.740 during those years.

Indicator	ISA Price COP	Market Price COP	ISA % Variation	Market % Variation	ISA Vol Mill. COP	Market Vol Mill. COP
Min	11740	4	-31.7	4	0	0
Mean	18371.8	14617.8	0	0	6789.5	3436.46
Q75	22165	22100	0.9	0.5	7832.2	2525.6
Max	29250	106000	12.9	100	132361.2	327000

Figure 2: Colombian Equity Market Last 5 Years

On figure 3, the size and color is the variation and the position depending of the price. It shows the top 5 stocks with highest volume, they are ECOPETROL, PFBCOLOMBIA, BCOLOMBIA, ISA, and PFAVAL. There are more common negative variation than positives, the highest mean negative variation is in the CONCONCRETO's stock price. Historically, BOGOTA was the stock with the highest price and FABRICATO with the lowest.

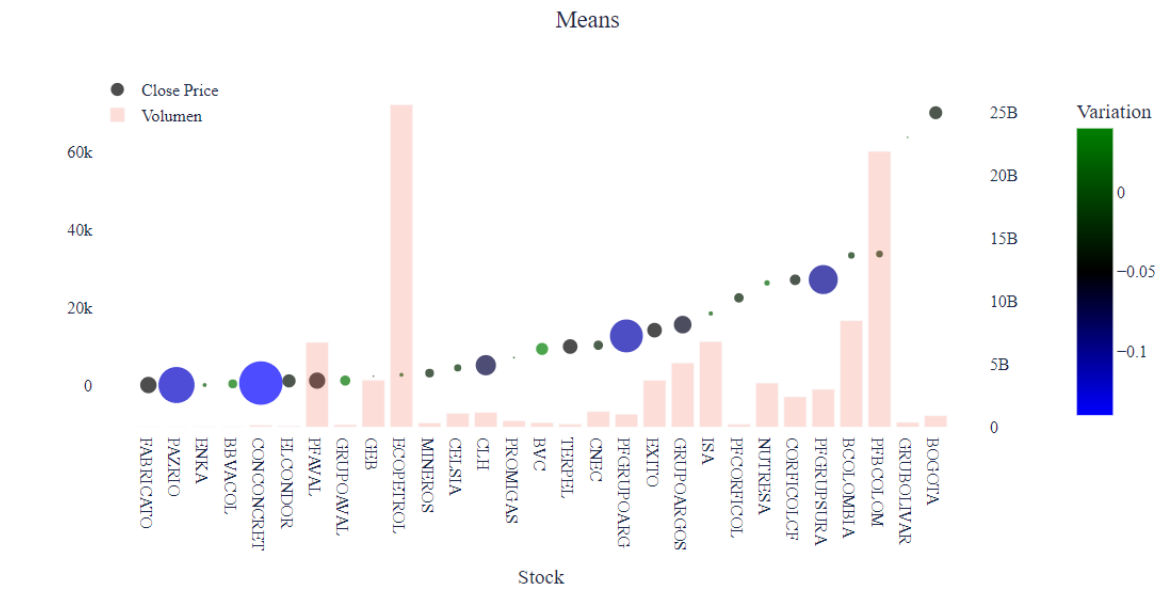


Figure 3: Colombian Equity Market Means Last 5 Years

We can decompose this serie in order to study trend, seasonality and residuals:

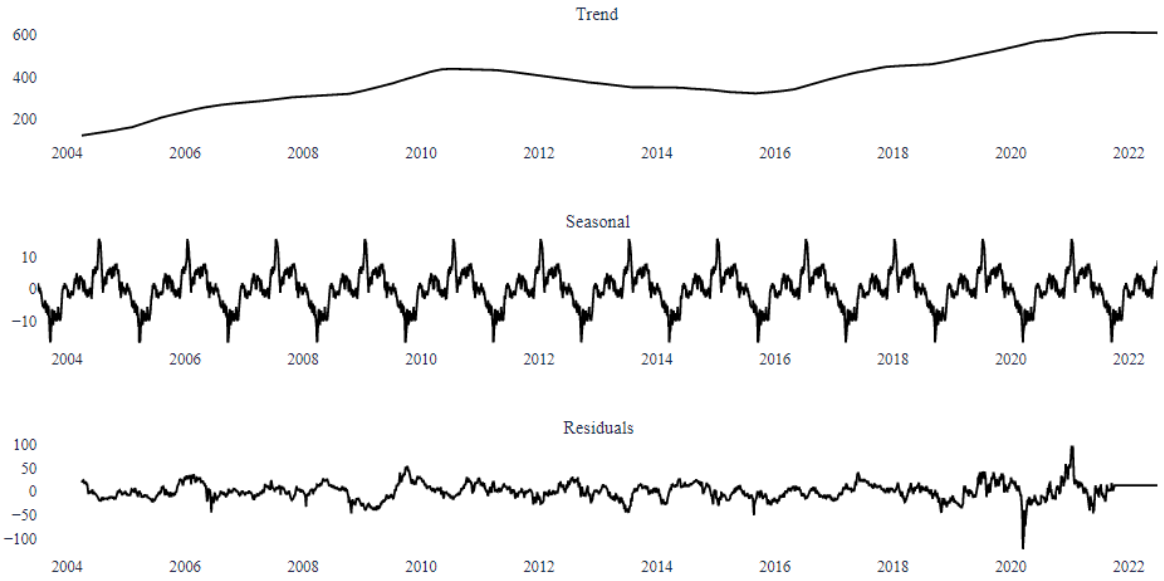


Figure 4: Decomposition of the serie

How are ISA's share prices related to others symbols?

Figure 5 presents the top 10 correlated symbols with ISA's share price during last 5 years.

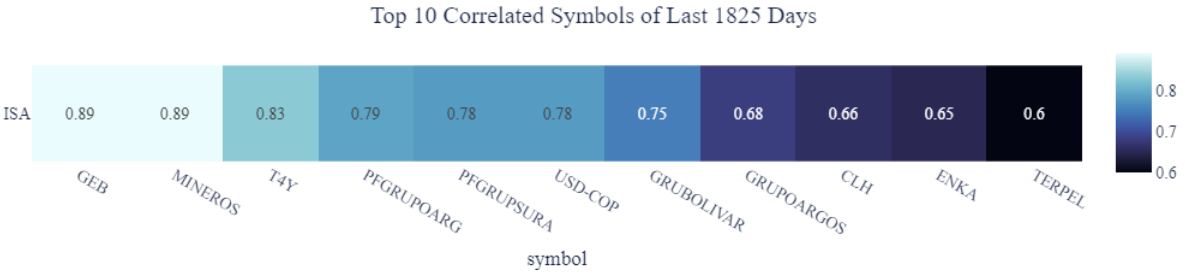


Figure 5: Pearson correlation of ISA's share price with others symbols.

This graph shows the high correlation ISA's share price keep with Grupo de Energia de Bogotá and Mineros from Stocks, and an important correlation also with USD-COP from forex market, and with T4Y from bonds market.

How is the behavior of ISA's share prices along years?

Figure 6: Behavior of ISA´s share price over time. Dots represent closing value of the stock, the color of each dot describes the price change (green dots refer to positive change while blue ones indicate negative change) and the size of the dots represents the volume of transactions.

The graph 6 suggest different patterns which are worth it to analyze:

1. Shows that in recent years there has been an increase in the volatility of the share price and in transaction volumes
2. Evidence some structural changes which can be due to either macro economical or fundamental variables.
3. Indicate how strong ISA´s share price behaved in front of recent pandemic crisis as it drops and shortly returned to its normal price.

Considering this information, it is necessary to analyze what additional factors are affecting this market dynamics.

How is the data distributed?

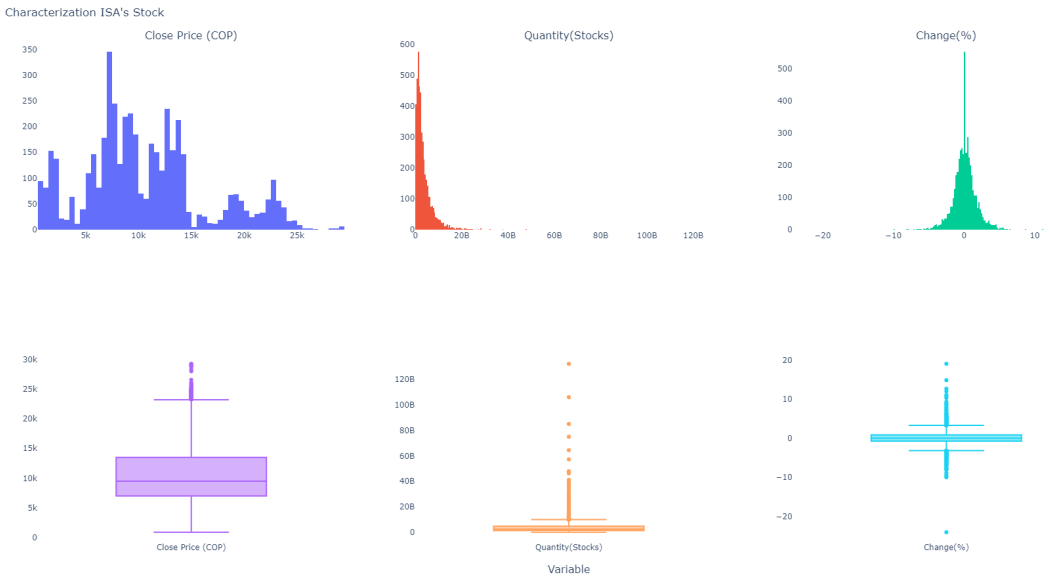


Figure 7: ISA stock data distribution.

The graph 7 provides important information that is going to be used when modeling the data and forecasting future prices:

1. Both the distribution and box-plot of the closing stock price evidence that most of the time, price has range between 5k and 15k, while higher prices than 25k has behaved as outliers in the series

2. Finally percentage change seems to behave normal (imbalance theory) and shows how positive change are normally compensated with negative ones.

These incipient graphs provide important insights but we will come back latter to this analysis in order to add information (like inflation indexes) and perform deeper conclusions.

3 Methods and Models

3.1 Domain and technical knowledge

Stock price prediction is the process by which future stock prices are forecasted on the basis of past prices. There are two strategies for predicting future prices of a stock: fundamental and technical analysis (historical data). The former one analyzes the intrinsic value of the company / commodity by studying most of the factors with an impact on the stock price of the company in future, such factors could be financial statement, management process, industry, etc. On the other hand, technical analysis uses past charts, patterns and trends to forecast the price movements of the entity in the next periods.

Using the data provided, it's possible to do technical analysis and a baseline forecast. The data is not independent, identically distributed, financial in nature, it suffers from several outliers and changes in level. There are some considerations to take into account regarding stocks market prediction. The first one is running live strategies requires a different approach, in this work strategy will run off-line. The second consideration is about data and code format, in this case, the vectorised format will be used; such format is a tabular format containing historical data and the code follows the algorithm.

Algorithm 1 ISA stock prediction Algorithm

```
i ← 0
Create stock prediction hypothesis
while i ≠ MaxIters do
    Implement a model on historical data
    Analyse the performance of the strategy                ▷ Backtesting
    if backtesting is not satisfactory then
        Adjust/tweak the model
        i ← i + 1
    end if
end while
if i = MaxIters then
    Change the stock prediction hypothesis
end if
```

For building the methods we use the most employed libraries respect to stocks forecasting which are:

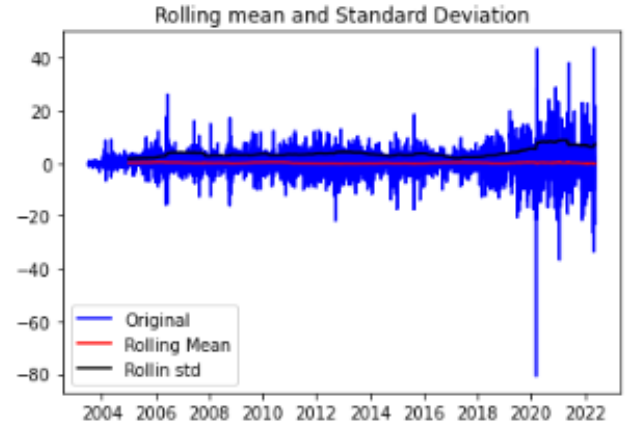
- Tensorflow,
- Keras,
- Pandas,
- Sklearn,
- Numpy,
- Matplotlib,
- Pycaret package (https://pycaret.org).

In the next section, we describe our models, the method used (slightly) and the results. Explicitly we study univariate and multivariate models. For the formers (ARIMA and LSTM) we just report the outcome, but we concentrate on multivariate dynamic models (FB prohpet, Random Forest and XG-Boost) as they have proved to behave better in production at forecasting. Mean Absolute Percentage Error (MAPE) is the metric we are going to use for comparing final models in competition, because it comes under percentage errors which are scale independent and can be used for comparing series on different scales. This metric is going to be the result of variant windows size and variables included in the models (as response of the correlations of their series respect to the one of ISA, the MAPE of its own series - which takes into account the randomness present in the series from its residuals - and technical indicators like exponential means and MACDI) in order to reach the target: Finding the lowest error between the forecast and real values.

3.2 Approach from univariate models

3.2.1 Auto-regressive integrated moving average(ARIMA)

ARIMA models suppose that serie get stationary status after being differentiated. After having used box-cox correction and one differentiation the plot and dicky-fuller tests show:




```
Result of Dickey-Fuller Test:
Test Statistic                  -32.372822
p-value                           0.000000
#Lags Used                        5.000000
Number of Observations Used    5444.000000
Critical Value (1%)              -3.431552
Critical Value (5%)              -2.862071
Critical Value (10%)             -2.567053
dtype: float64
```

Figure 8: ISA's stock data distribution.

Which means that we have a stationary series in mean and variance. This serie can be used to forecast the figure 8 (in spite of the inconstant variance yet). Now we can define the auto-regresive and MA components to build the ARIMA model. This procedure is done through the autocorrelation and partial autocorrelation function defined in graph 9:

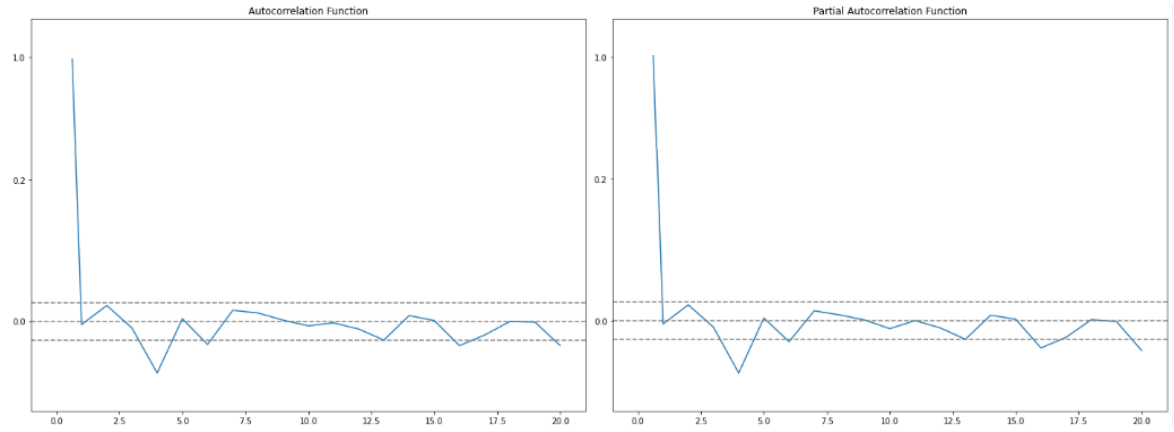


Figure 9: Autocorrelation and partial autocorrelation function.

This plot suggest that both auto-regressive and MA components are equal to 1 as well, so we would have an ARIMA (1,1,1) as the best model for our data, however the RSS is not good enough. When fitting this forecast to real data, what we obtain is:

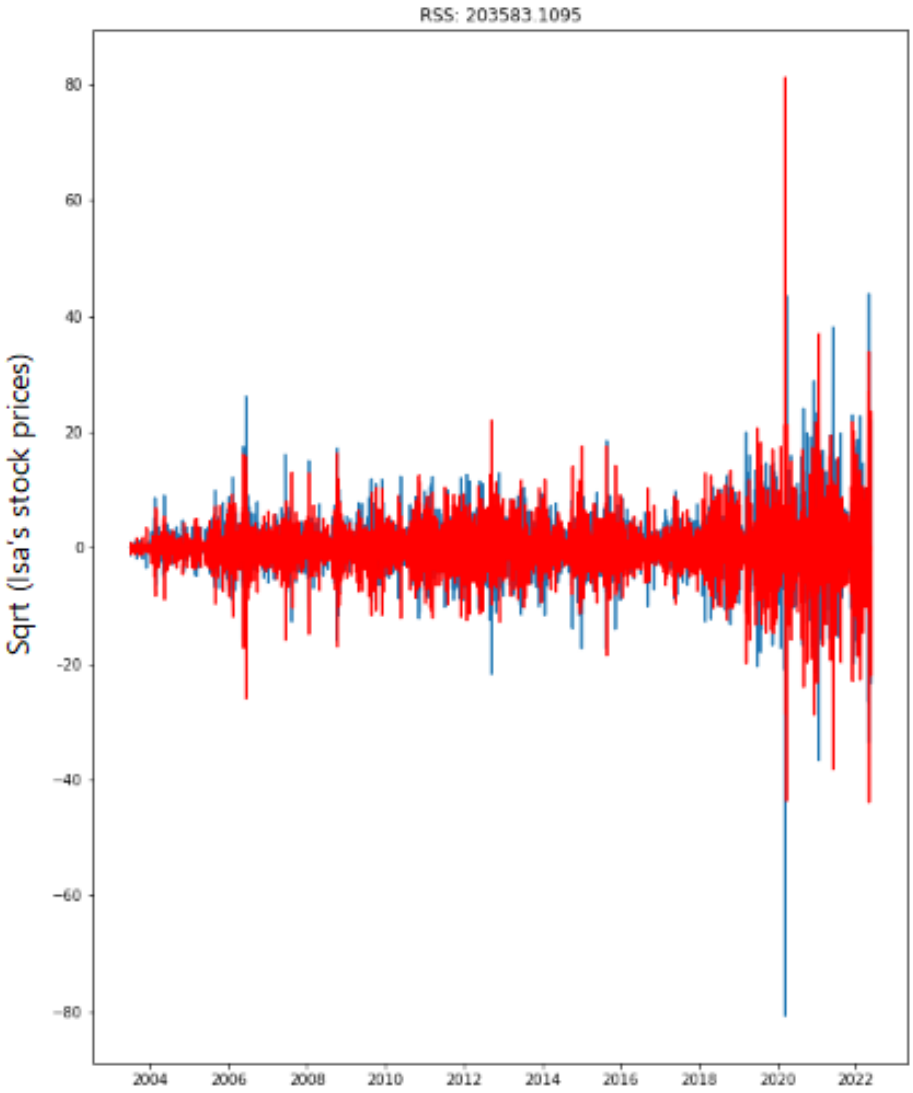


Figure 10: Forecasted and real ISA's stock price.

3.2.2 Long short-term memory Artificial Neural Network (LSTM)

Within deep learning models we take the univariate LTSM algorithm, which has shown efficiency in other projects. In this case, however, the output of this model had a RMSE which ranged from 931 to 454:

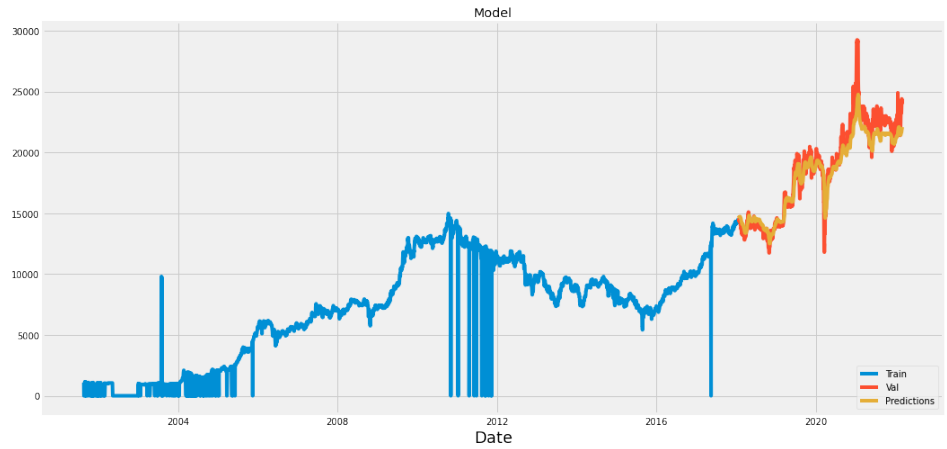


Figure 11: Model predictions for 60 days.

n_days	RMSE	Loss
30	931.77	9.32e-04
60	432.43	9.16e-04
90	454.84	9.95e-04

Table 3: RMSE and loss for three different time steps

We concentrate now on multivariate dynamic models (FB prophpet, Random Forest and XGBoost) as, how we said before, they have proven to behave better in production at forecasting.

3.3 Approach from multivariate models

In the case of these dynamic models, we got better results from the beggining. That´s why we continue working with them and perfect the methodology (which is being described in the next section) in order to obtain suitable results. Since the market changes over time, and there are more efficient models in certain scenarios, we decided to define a method that allows us to select the best forecast for the current scenario respect to to the algorithms that have shown a relevant performance, they are Facebook Prophet, Random Forest and XGBoost:

- Facebook Prophet: Prophet is a procedure for forecasting time serie data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. It works best with time serie that have strong seasonal effects and several seasons of historical data. Prophet is robust to missing data and shifts in the trend, and typically handles outliers well. For further information about this model visit `https://towardsdatascience.com/time-series-analysis-with-facebook-prophet-how-it-works-and-how-to-use-it-f15ecf2c0e3a`

- Random Forest: This model is a machine learning technique that's used to solve regression and classification problems. It utilizes ensemble learning, which is a technique that combines many classifiers to provide solutions to complex problems. A random forest algorithm consists of many decision trees. The 'forest' generated by the random forest algorithm is trained through bagging or bootstrap aggregating. Bagging is an ensemble meta-algorithm that improves the accuracy of machine learning algorithms. You can learn more about this one `https://medium.com/@madushankanipunajith/random-forest-in-machine-learning-91f7a890599e`

- XGBoost: Stands for Extreme Gradient Boosting, is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library. It provides parallel tree boosting and is the leading machine learning library for regression, classification, and ranking problems. It's vital to

an understanding of XGBoost to first grasp the machine learning concepts and algorithms that XGBoost builds upon: supervised machine learning, decision trees, ensemble learning, and gradient boosting. You can learn more about this technique `https://medium.com/p/edd9f99be63d`

3.4 Forecasting method

3.4.1 First approach

In an exploration process, models were developed as follows: first, we included a series with current stock prices of all the firms taken into account for building the dynamic model. Second, we used lists containing technical indicators (chosen regarding to what financial technical analysis has shown to work better for forecasting) for ISA and other correlated firms: 10-period and 55-period exponential moving average, 1-period volatility and seasonal effect included through a wavelet function. Finally, we included a one period lag of ISA's stock price.

It should be noted that this incipient concept of the prediction being made is based on the following assumptions:

- We know the prices of the firms and values of technical indicators used at the date of the prediction.
- We are going to predict the price of ISA's stock n days ahead.
- We first used the data from April and ahead as a test set in order to prove how good predictions are.
- The known value of ISA's stock price up to the date of the prediction is included as a training value.
- Each predicted data composed a data set for n days ahead.
- We do no use windows of stimation but the entire dataset.

In this exploratory approach, we had a wide scope concerning the ammount of days predicted: 1 to 45 days in the future. A different model would be available for each day in the future. As an example, the prediction made over some dates of the test group is shown in Figure 12.

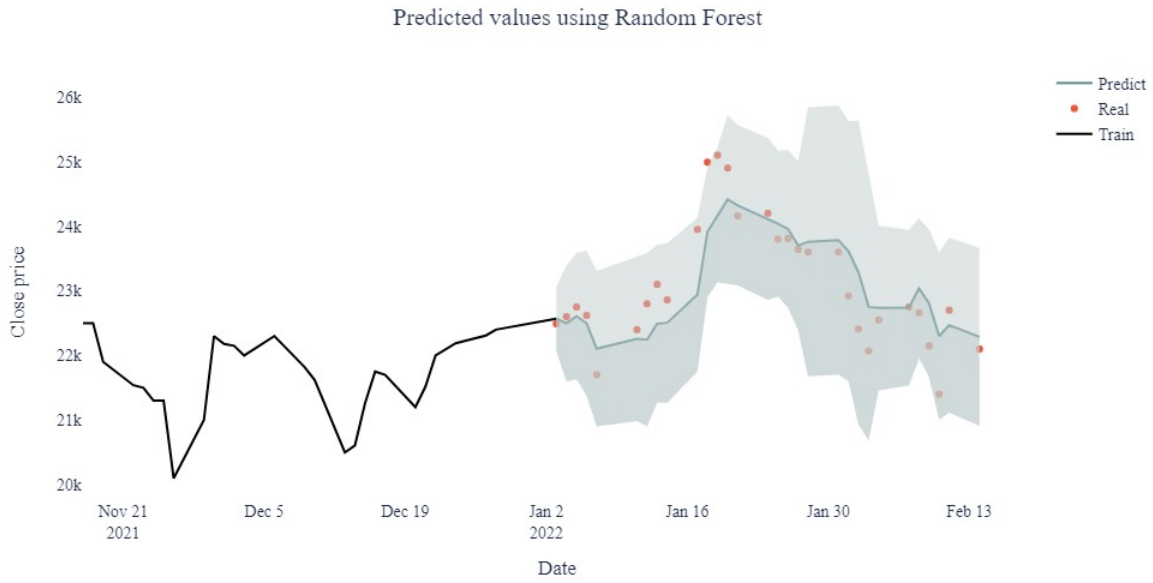


Figure 12: A month and a half of testing with a prediction from a specific date into the future.

3.4.2 Method design and model selection

After exploring a number of univariate and multivariate models, and changing model paramaters and features (variables and technical indicators) taken into account, we determined that predictive efficiency varies over time. Therefore, it is not convenient to generalize an only model for ISA's stock value prediction. In search of a dynamic solution, an optimization method was defined for parameters and features composing the models which was designed for getting predictions as close to real values as possible. The method is presented in figure 13 and can be deployed using as many models as one want them to compete. This method would be run every day to update the data on the web page.

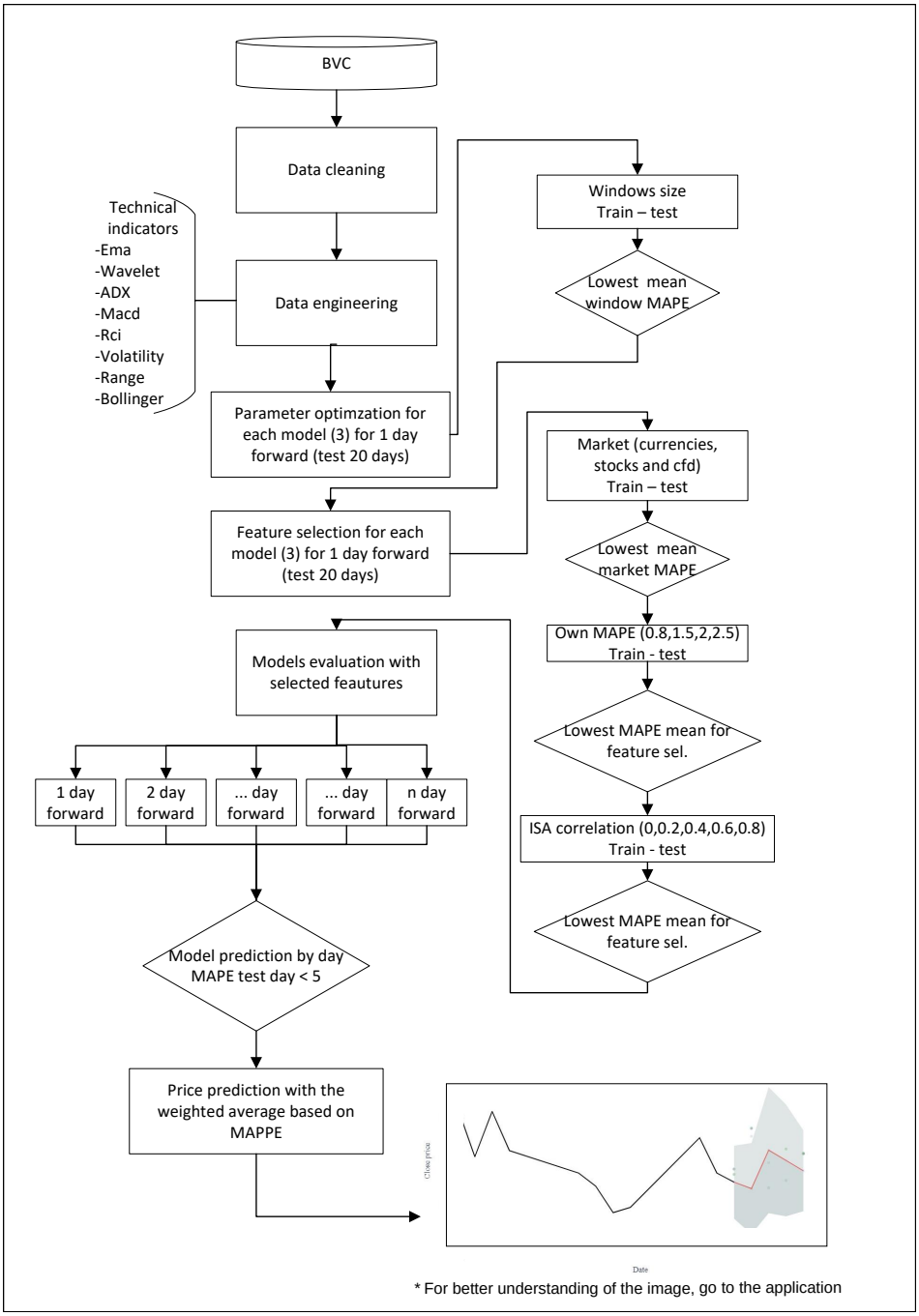


Figure 13: Method designed

Now we have defined the models we are going to use (section 3.3), we can talk about the method itself. Figure 13 describes our designing step by step. we have already described data cleaning and data engineering (aggregation of important technical indicator to the dynamic models) as part of our procedure. From this point on we begin surface tests: tests which are orientated to optimize hyper-parameter of the models (Figure 14), windows sizes (Figure 15), value provider firms (respect to the residuals when creating a multivariate auto-regressive model for each firm considered), correlated firm´s stock prices with the one of ISA, and suitable technical indicators. We describe indicators use for the ease of the reader:

- WAVELET: It is a tool for decomposing a serie by location and frequency.
- EMA: An exponential moving average (EMA) is a type of moving average (MA) that places a greater weight and significance on the most recent data points
- ADX: the average directional index (ADX) is a trend strength indicator, namely, it is used to determine when the price is trending strongly
- MACD: Moving Average Convergence Divergence (MACD) is a trend-following momentum indicator that shows the relationship between two moving averages of a security's price.
- RSI: The relative strength index (RSI) is a momentum indicator which measures the magnitude of recent price changes to evaluate overbought or oversold conditions in the price.
- Volatility: Refers to simple standard deviation
- Range: Refers to min and max value.
- Boilinger bands: These are envelopes plotted at a standard deviation level above and below a simple moving average of the price. Because the distance of the bands is based on standard deviation, they adjust to volatility swings in the underlying price.

We do this in a first stage which we call test stage, where we start generating estimation from 20 days before today which moves one day ahead for each new estimation and use MAPE of each option of the parameters/values studied from the surface tests in order to compare and decide which is the best option.

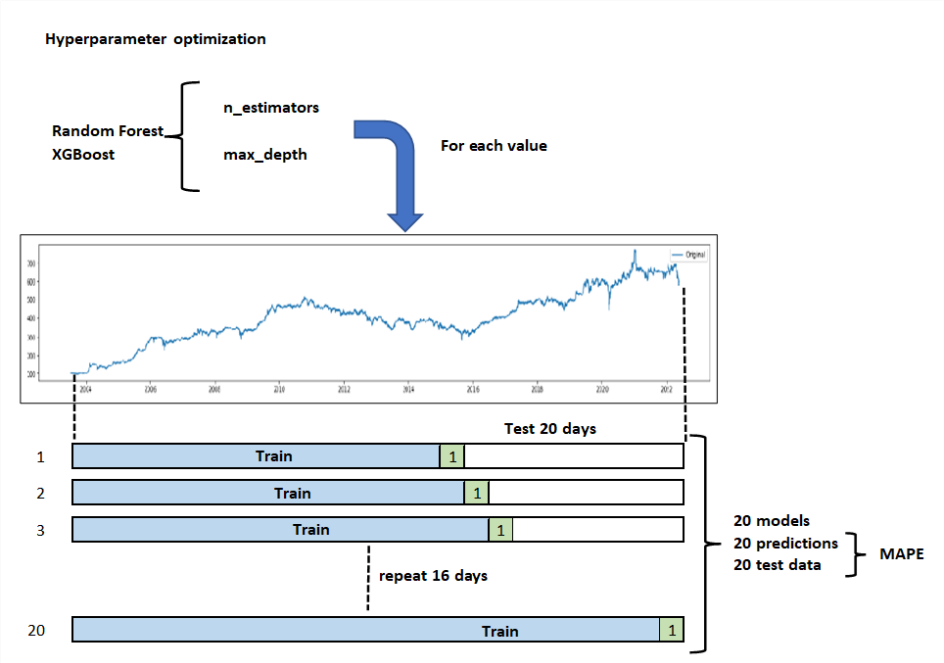


Figure 14: Hyper-parameter optimization.

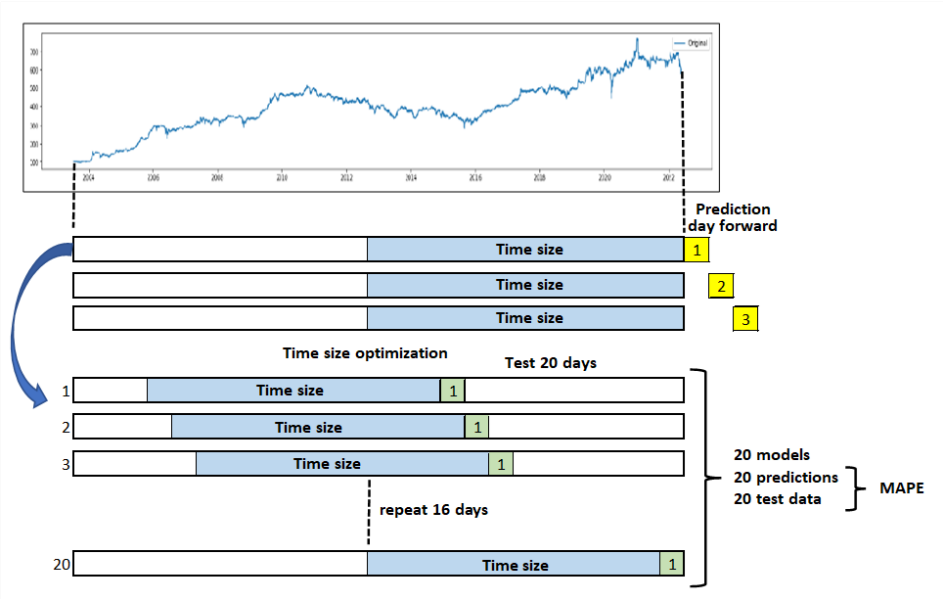


Figure 15: Window size optimization.

After selecting the best paramaters/ feature values (feature selection over correlation, own MAPE of firms used as regressors, windows size and relevant technical indicators), we can describe a final stage: Production. In this stage, we perform prediction with the optimized parameters and features and again, evaluate the accuracy of this prediction using MAPE as shown in the figure 16.

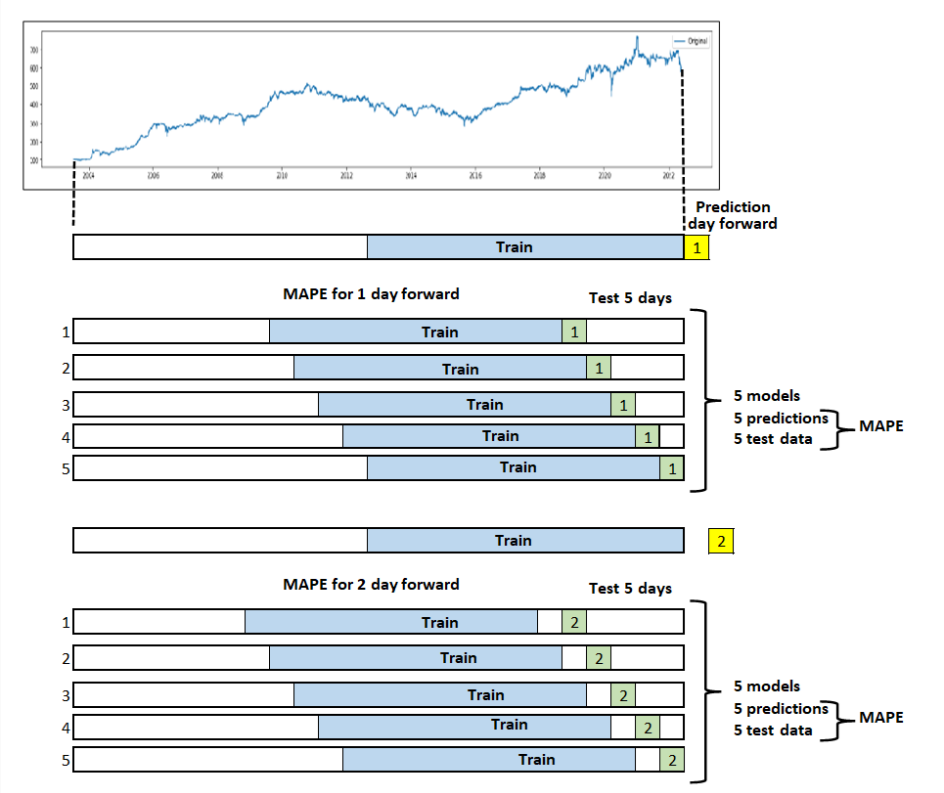


Figure 16: MAPE for prediction.

As shown in figure 13, before predicting, we perform a filter in order to decide which of the 3 models is going to contribute to the forecasting and how much is it going to be. To do this, we perform again a day-ahead prediction with a mobile window of 1 day for every model considered and for 5 days.

We then evaluate if the average MAPE of the predictions is less than 5. If it meets the condition, the value is used to make a prediction with the weighted value among all the stable models' predictions. Weighted values gives greater importance to the lower MAPE, generating more accurate predictions. If three of the models have a MAPE greater than 5, the prediction is omitted as we consider rising a caution about the potential bad estimation is better than generating a false or biased value.

4 Interface

The solution interface is composed of 4 main screens. These screens are titled: home, monitor, context and forecast. On each screen there is a drop-down menu with: the glossary, the project scoping and information about the team.

The home screen presents the information of each of the functionalities of the application and the objective of the present project. The figure 17 shows that this screen presents the colors contained in ISA's logo and there are multiple options to access the application's functionalities.

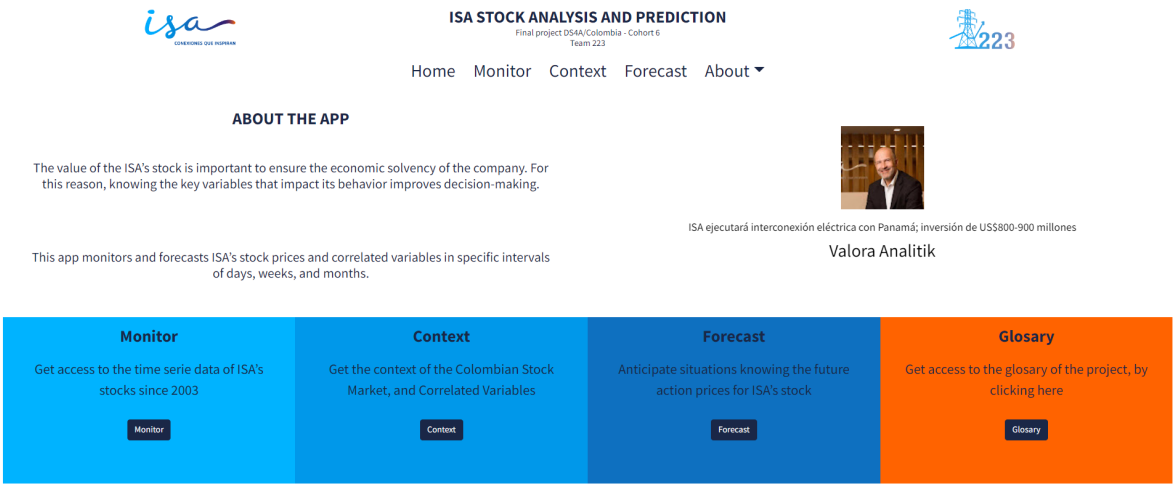


Figure 17: Home screen.

The monitor interface aims to track the time serie of ISA shares (Figure 18). The first graph evidences the movements of the stock since its listing in colombian stock exchange and allows to filter by periods of 15 days, 1 month, 3 months, 6 months or a year. This chart uses the OHLC format or candlesticks which allow you to see the price of open, high, low and close. This chart also includes indicators to track the stock. These technical indicators are: the 12 and 55 day exponential moving average. The y1 axis represents the closing price and the y2 axis represents the number of shares bought during each day. The axes are accompanied by a bar plot showing those moments with the highest number of transactions. Finally, graph 1 has the bollinger bands. These bands show the moments of volatility within the market.

The monitor has a second graph containing a decomposition of the serie. This decomposition presents the trend of the serie, the stationarity and the residual values. This plot was constructed after a Box-Cox transformation and using the seasonal_decompose library of statsmodels.

Figure 18: Monitor interface.

The context screen presents an overview of the stock market over the last 30 days (Figure 19). On this screen there is a table with the price differences that exist within the close, the percentage of variation and the volume of the ISA price with the equity market. Each stock of the stock market is presented with its minimum, average, 75th percentile and maximum values. The main figure shows ISA's performance in relation to the Colombian stock market. This graph allows us to identify several elements. First, which stocks have the highest trading volume. Second, what is the variation of the shares, and third, what is the location of the share price within the equity market. Finally, the symbols (assets) included in our database that have the highest correlation with the ISA share are presented.

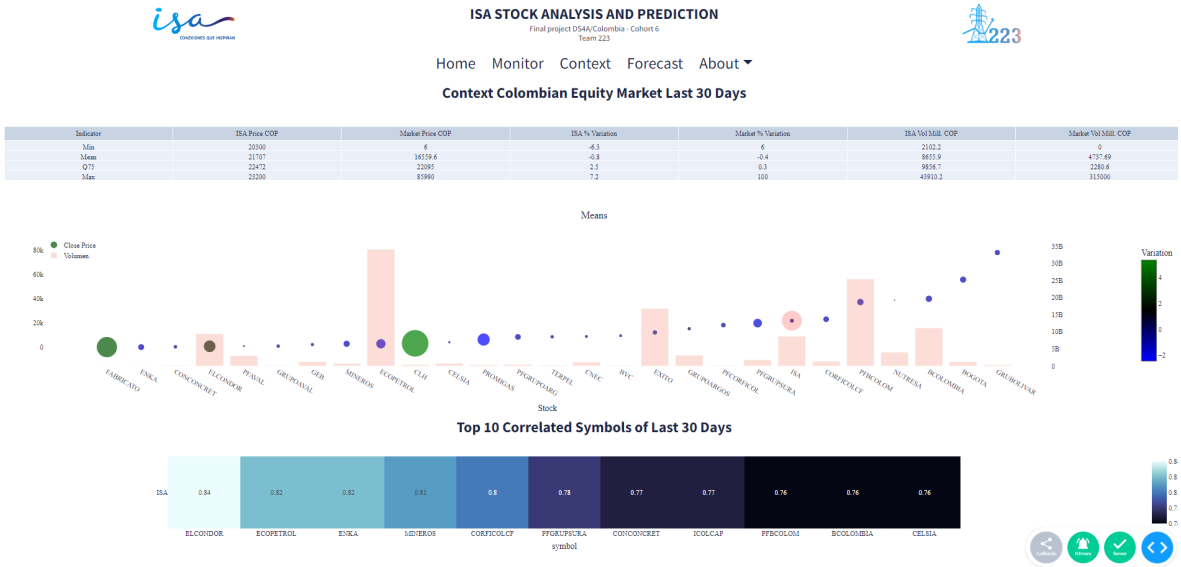


Figure 19: Context screen.

The forecast screen presents a line chart with the last year of historical ISA share price information, accompanied by a second line chart with the predictions for the following periods (Figure 20) . The forecast is accompanied by a gray shadow that represents the mean squared error of the model and represents our confidence interval.

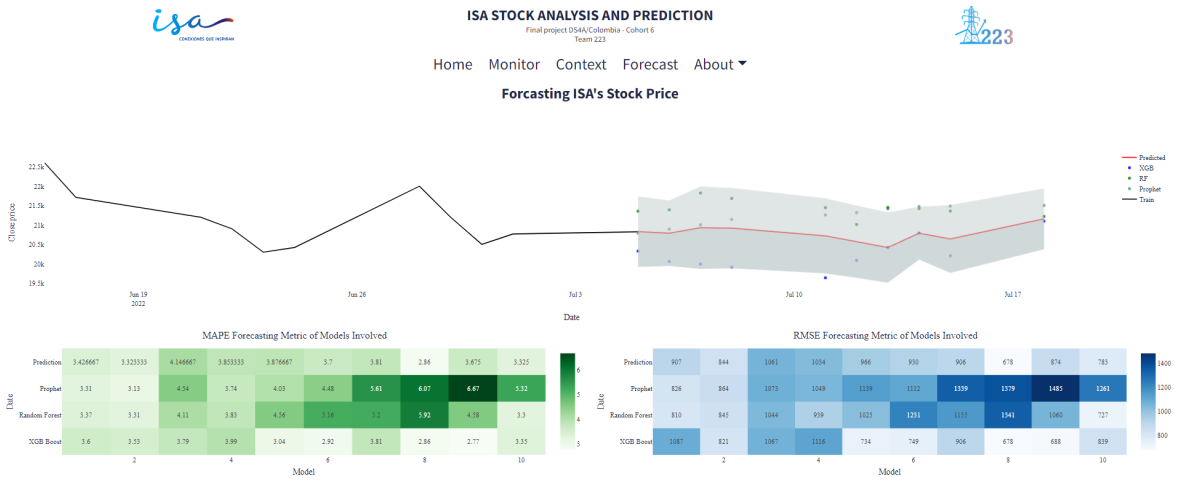


Figure 20: Forecast screen.

The about screen provides information to the user about the application. The information will be presented as a glossary with keywords and articles related to the project. This screen displays this document from the web site (Figure 21). Finally Team 223 will contain information about the authors of the project.

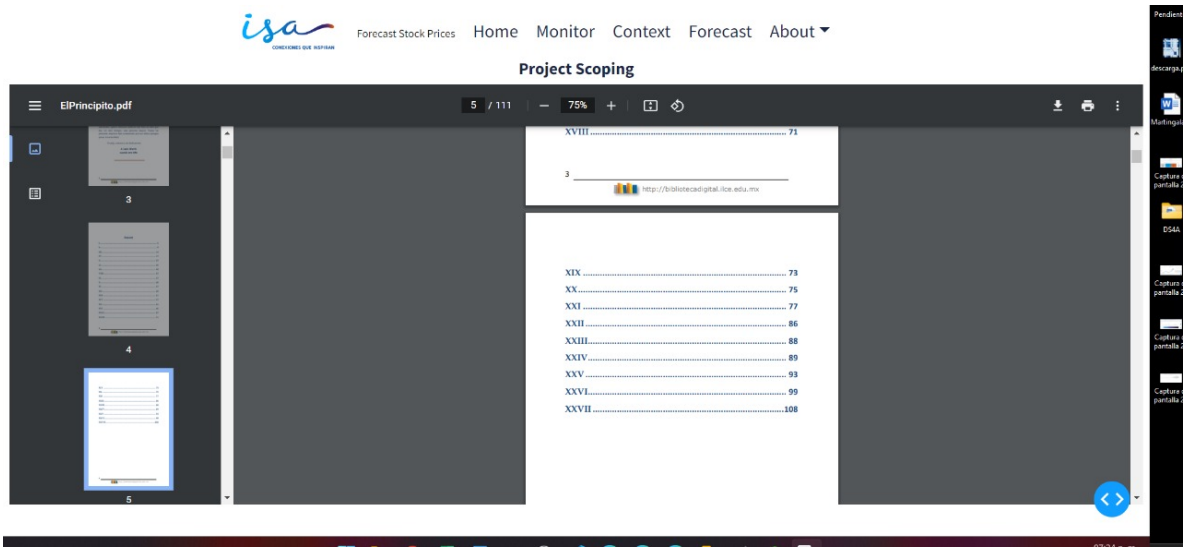


Figure 21: About screen.

4.1 Page components

The backend of the application has been implemented with python using the pandas and numpy libraries, since they allow data exploration and cleaning.
At the frontend level we have used dash to structure the interface that brings together the different visualizations. In particular we have implemented the dash_bootstrap_components library for a better distribution of the components ensuring a responsive behavior that allows access to the web from any device.
As for the visualizations, we have implemented histograms, scatterplots, line charts, bar charts and

boxplots using the plotly library, which offers interactive and dynamic graphics that make the interface a friendly space.

4.2 Application Link

http://50.116.45.165

5 Conclusion

Our porpoise was to correctly forecast the prices of ISA's stock prices based not only on the own serie of historical prices of the serie, but including valuable variables so that we can set a dynamic model whit suitable forecasting power. We first performed data cleaning and feature engendering and then we started optimizing several features in order to find the suitable components of the dynamic model: we look for optimization around the price of related companies (in order to decide which of them to take into account for forecasting), technical trading indicators (which help considering the strengh of the trend, the momentum, the weithed impact of recent prices and the volatility of the serie) and finally we also added transformations of the own ISA's price like the one suggested by wavelet and boxcox.

For having clear the variables we will take into account and the parameters of the models used, we first designed a test stage, where we adjusted these values for then running the selected models in the form of mobil windows (of 1 and 2 days) for predicting one and two days ahead (production stage). The models chosen were Facebook Prophet, Random Forest and XGBoost (this two last are part of the family of decision three decision models) as they were the ones which better behaved respect to other like ARIMA and neural networks at production systematically adequate estimations.

For evaluating which of the prediction is better in each of the iterations, we use MAPE and decided not to only used the one which behaved better in the last estimation but create a weighted average of those models (out of the 3) whose forecast produced a MAPE greater than 3. We have being tracking the estimations in real production from July the 1 and have agreed on this approach as a good method.

We hope our designing to be relevant respect to the requirements of ISA INTERCOLOMBIA and help them making valuable data-driven decisions over the forecasted prices but we would like to let two things clear:

1. The models used and the method design are optimized to answer the specific needs and data of ISA requirements. We do not recommend to use them with different firm's stocks or data in general
2. We would like to make clear that our method should be used only as a reference point about where future price will probable be, and recommend to pay more attention to the confidence intervals (designed through the estimation +/- RMSE) than to the specific value estimated.

6 Upcoming work

- It is expected to build a Backend with Fast-API that aims to request and update the information to a MongoDB database, then pass it through the method. Finally, the ETL-style responses would be returned to the same database with the results of the query or prediction in a different document (table).
- The extraction of the data from the symbols of other markets such as cryptos, S&P100 data or commodities (such as gold or oil) is interesting.

- We also expect to apply convolutional models like neural convolutional networks in order to prove their eficiency respect to our models for prediction.

7 Authors

This is the team's work 223

Buritica Angulo Rominger
Student of electrical engineering and data science with a strong orientation towards the application of engineering through practice and software, with knowledge of standards, regulations, and resolutions governing the practices and applications of electricity in low, medium, and high voltage. Currently, he is a collaborator of the research group GISE3 of the Universidad Distrital Francisco José de Caldas focused on the study of atmospheric electrical discharges and electromagnetic compatibility (EMC), making small contributions in the area of lightning parameters measurement. In the work environment, he works as an analyst of connection studies, focused on the behavior of the national electric power system, locating the possible areas of greatest interest to inject power from renewable energies considering the technical and economic feasibility of the projects to be submitted to the Unidad de Planeación Minero Energetica (UPME).

Cortés Cataño Carlos Felipe
Consultant and entrepreneur, he is an accountant from the Universidad del Manizales COL, in 2019 and a software engineering student from the Politécnico Grancolombiano COL, since 2020; in addition, he is studying for a Master's degree in Data Science at the Pontificia Universidad Javeriana COL, since 2022. He contributes to different companies in analytics and organization processes from his programming knowledge, statistics, finance, and taxes.

Lopera Parra Luis Miguel
Economist and scientific master in finance with an analytical and goal-oriented approach. Leadership and assertive communication have been developed through my experience as teacher (both in tecnológico de Antioquia and EAFIT) and during my work as lecturer for Banco de la República. Throughout my postgraduate studies, I worked as financial researcher in EAFIT and data scientist at Ingeniería y hogar and discovered my passion for data science world.

Mariño Osorio Nataly
Data analyst and project mangager in the Superintendence of Public Home Services. She has degree in systems engineer from the Universidad de Caldas, since 2012 and a Master's degree in software project management and development from the Universidad Autonoma de Manizales, since 2017. She has 8 years leading software development teams. Passionate about designing software products integrating market/customer needs with corporate strategies. Making the objectives possible through the coordination of the personnel involved in the product value chain from developers to support assistants.

Vargas Cañas Rubiel

Received the B.S. in Computer Sciences Engineering from the Universidad Industrial de Santander, Bucaramanga COL, in 2000, M.Sc. degree in engineering from the Universidad del Valle, Cali-COL, in 2010 and the Ph.D. degree in biomedical engineering from City University, London, UK, in 2012. From 2001 to 2002, he was the coordinator of the academic Databases for the library of Universidad Industrial de Santander. Since 2002, he has been a Professor within the Physics Department, Universidad del Cauca – COL. He is the author of two books and near fifteen academic papers. His research interests include digital signal processing, digital image processing, biomedical engineering, machine learning, deep learning, and internet of everything. He is a reviewer of several journals. ORCID: https://orcid.org/0000-0003-1548-942X

Vega Díaz Jhon Jairo

Researcher, teacher, consultant and entrepreneur, he has two engineering degrees, first the B.S. in agronomic engineering from the Universidad del Tolima COL, in 2001 and second the B.S. in computer sciences engineering from the Universidad del Tolima COL, in 2014; in addition a M.Sc. degree in Management from the Universidad Nacional, Manizales COL, in 2006 and the Ph.D. degree in Applied Science from Universidad Antonio Nariño, Bogota COL, in 2021. He is an expert in innovation methodologies and digital image processing. In innovation methodologies he uses TRIZ for the generation of innovative solutions and TRL to guide the development of technology, in this topic he has a patent application to identify the maturity of fruits in the tree through images. In image processing he develops classification or prediction models using RGB, multispectral, hyperspectral or thermal images; whether acquired on the ground, drones or satellites; these developments have been applied in precision agriculture and industrial inspection. He is currently working on the generation of predictive models in agriculture using the internet of things with LORA networks. ORCID: https://orcid.org/0000-0002-2165-8536

References

Bolsa de Valores de Colombia. (2022a). *Indices accionarios, Mercado de capitales.* Retrieved 2022-05-19, from `https://bvc.co/mercado-local-en-linea?tab=indices_accionarios`

Bolsa de Valores de Colombia. (2022b). *Renta variable, Mercado de capitales.* Retrieved 2022-05-19, from `https://bvc.co/mercado-local-en-linea?tab=renta-variable_mercado-local`