# Gas price forecasting through various statistical methods

POLITECNICO MILANO 1863

***Group Members**: Leonardo Di Caterina, Alessandro Ferrante, Arthur Cosson, Andrea Mancini, Giulia Buttazzo*

***Course**: Applied Statistics | **Professors**: Piercesare Secchi & Guillaume Koechlin*

## 1. INTRODUCTION

Natural gas is a clean-burning source of energy that is used for heating, cooling, electricity generation, creating indispensable materials (such as steel and concrete) and more.
As a clean energy, natural gas plays an important role in the transformation of the world's energy system, offering extensive advantages in dealing with global climate change.
The reasonable and effective prediction of natural gas prices is helpful for researchers and decision makers in commodity trading and power production planning, allowing them to make better decisions and establish effective risk-avoidance mechanisms. Natural gas price forecasting is critical to the energy market orientation, and it can provide a reference for policymakers and market participants.

## 2. GOAL

The aim of this project is to make predictions on the price of natural gas using statistical modeling methods. In particular we studied the usage and comparison of the classification and regression models.

## 3. DATA ACQUISITION

Our dataset comes from the website *"finance.yahoo.com"*.

The sample units are the days: we set a time period that goes from 1st January 2017 to 6th June 2021.
The variables are:

- **CF=L** : ticker symbol for crude oil features;
- **^STOXX50E** : ticker symbol for the Euro Stoxx 50 index (the main stock index for Europe; it represents the 50 largest companies in the eurozone)
- **^GSPC** : ticker symbol for the S&P 500 index (the main stock index for USA)
- **^TNX** : ticker symbol for the10-year US Treasury yield (the reference yield of the US 10 years government bond);
- **NGAS.L (the target variable)** : WisdomTree Natural Gas ETC

We selected a mix of macro, equities and fixed income variables.

The fixed income market is important for commodities pricing as discounting inflation expectations and market risk sentiment. In a classic asset allocation framework, equities represent the risky assets.
In particular, we selected the main indices as a proxy for the equity risk factor.
Commodities are traded as risky assets and have historically exhibited time-varying correlation with equity.

Bonds are quoted in yield, so we had to infer the daily returns.
This was done through a standard simplification of fixed bond duration.
The price return of a bond is inversely related to the yield change.

## 4. DATA MANIPULATION

We checked for missing data and we filled them.

We calculated for each ticker:

- EMAs (exponential moving averages) with half-life of 5, 10, 30 days:
  $$\mathbf{EMA_t = (1-a)EMA_{t-1} + ax_t} \quad \textit{where } a = \frac{smoothing}{N+1},\ N = \textit{number of days in EMA},\ t = \textit{today},\ t-1. = \textit{yesterday},\ x_t = \textit{current price}$$
- fast and slow momentum:
  $$\mathbf{fast\_momentum = EMA_5 - EMA_{10}} \qquad \mathbf{slow\_momentum = EMA_{10} - EMA_{30}}$$
- rolling returns for 1, 7, 30 and 90 days.
  We calculated the logarithmic rate of return: $\mathbf{R} = \frac{\ln(\frac{V_f}{V_i})}{t}$ *where* $V_f = \textit{final value},\ V_i = \textit{initial value},\ t = \textit{length of time period}$
  We calculated the returns for ^TNX whit a different formula as this is a bond (bonds and stocks should be treated differently).

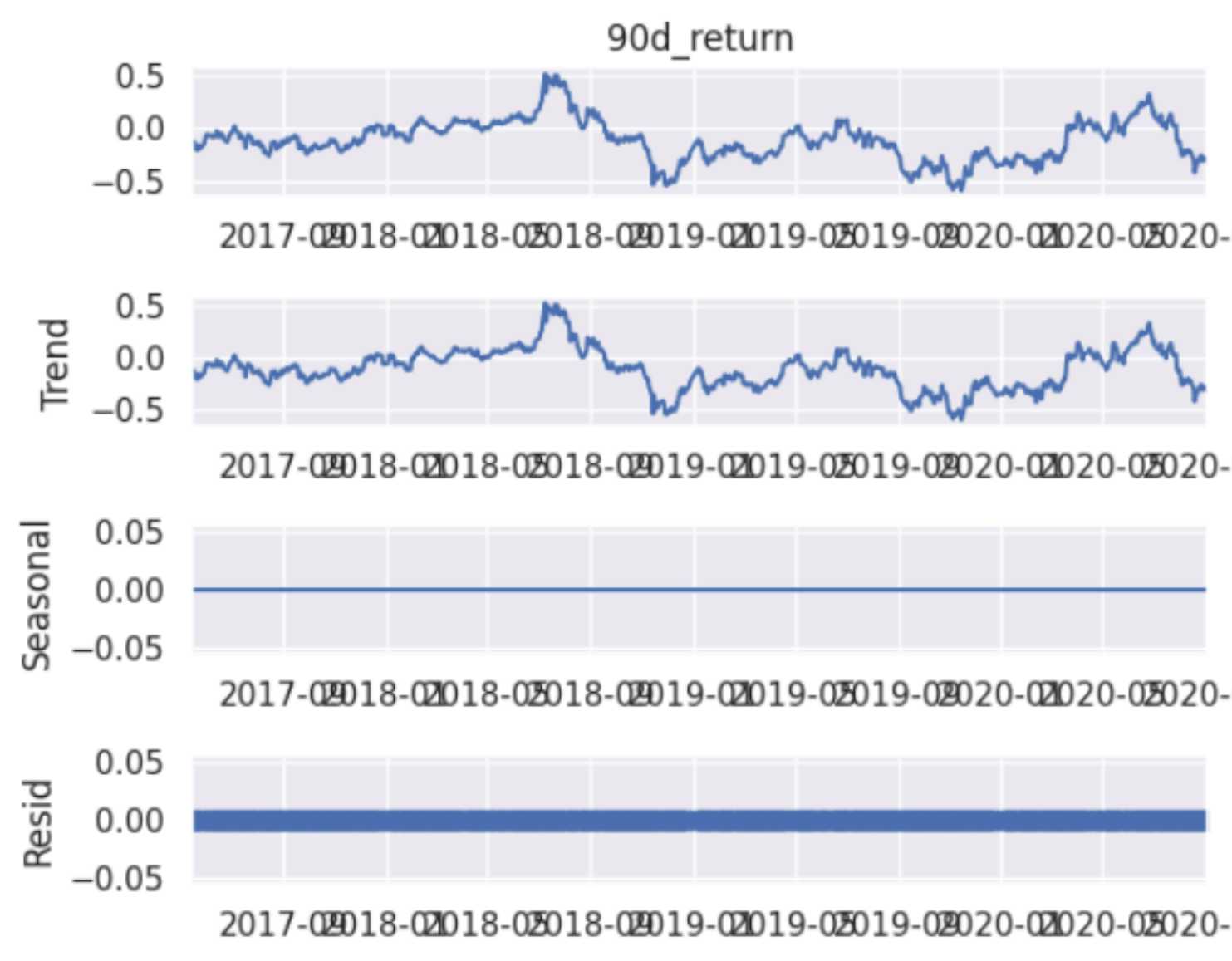
For the target variable we calculated the rolling returns for 30, 90 and 180 days.

The price time series is a random walk without any underlying distribution, we have to use the return, which has a gaussian distribution. In this project we use the log return, which performs better for finance data.

Further data manipulation involved calculating a proxy for momentum signals on each time series. For this, we used a commonly used method, which is the difference between two exponentially weighted moving averages with different half-lives. This is a simple proxy for the time series gradient, and the method is called "moving average crossover". The intuition behind this is very simple: if the recent moving average of prices is higher than the moving average with a longer window, the trend of the time series is upward. The trend has, over time, become an increasingly important factor as large market participants systematically trade trend-following strategies, also known as CTAs.

## 5. DATA ANALYSIS

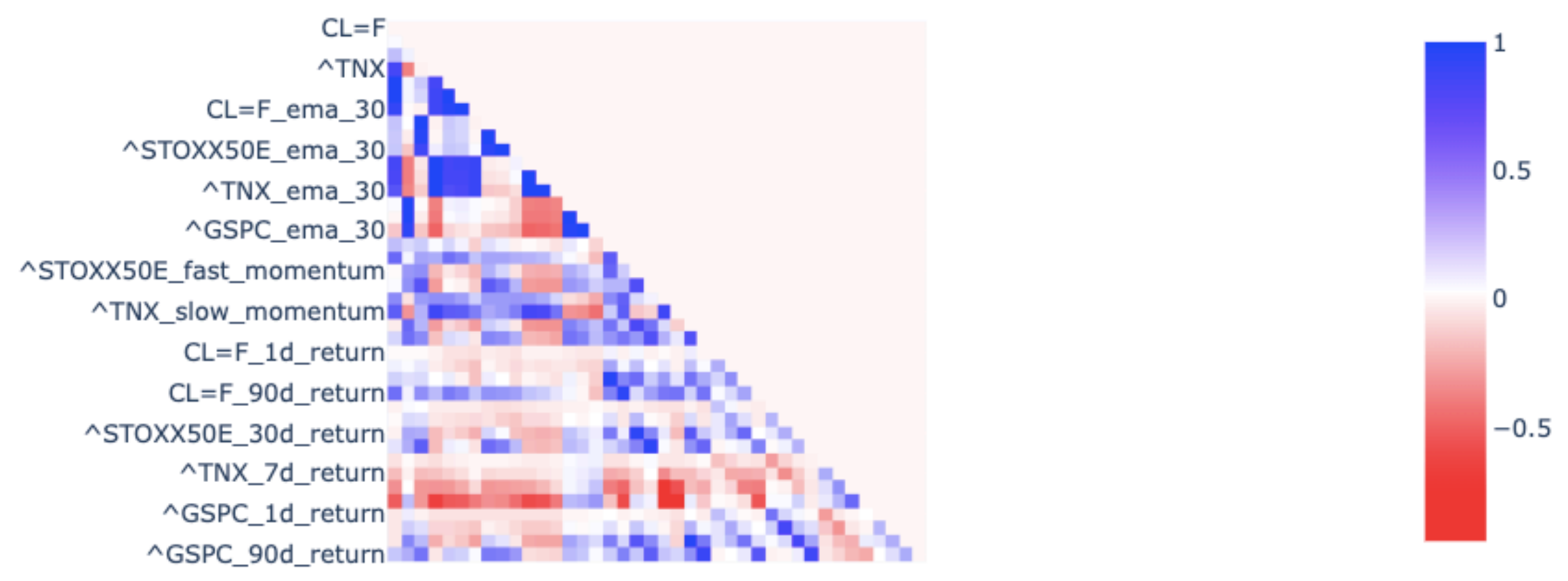
We checked for the periodicity of the time series of the returns of the natural gas price.
We report below the time series seasonal decomposition plot for the 90d return of NGAS.L .



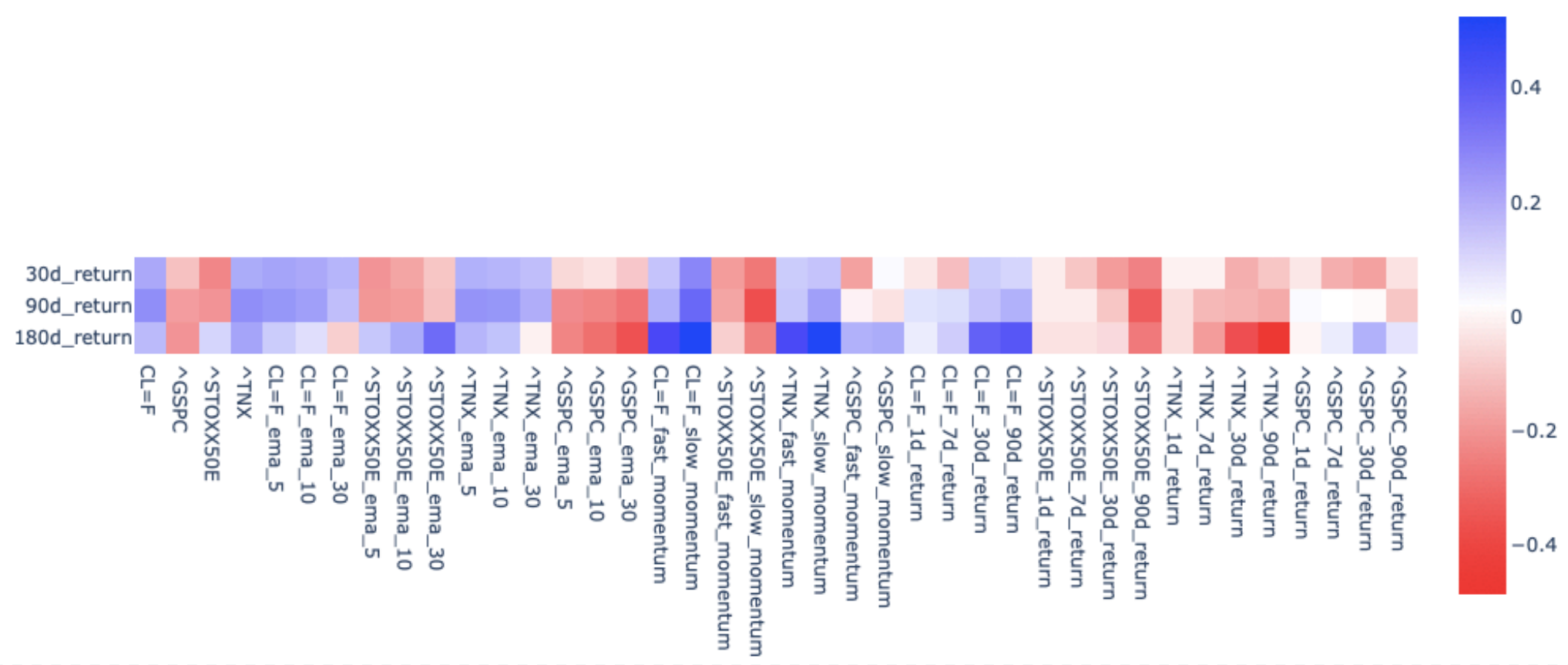
From the plots not the left we see that there are not seasonal trends of the returns.
This means that the returns are time independent. So, we have chosen a non time-dependent approach: we used specific time data, like EMA and momentum.

### CORRELATION ANALYSIS

We report below the plot of the correlation between the input variables.



We report also the plot of the correlation between the input variables and the target variables. We see that there is a good correlation. So, it makes sense to work with these data.



## 6. DATA PREPROCESSING

We have decided to discretize the forecasting problem. In practical financial applications, we would set a minimum expected return target before executing an order driven by the model. This would be driven by the strategy targeted volatility and liquidity of the underlying traded asset. Hence, we've decided to create a discrete framework of outcomes to classify. This would represent a theoretical systematic trading strategy.

We built a new target vector for classification.
We tried to predict the returns in 4 classes.

```
                 30d_return  90d_return  180d_return
Marginal Loss           365         310          324
Marginal Profit         186         182           75
High loss               161         264          363
High Profit              81          37           31
```

We shuffled the data for the training, so as to make them non-time dependent. In this way a better accuracy is achieved. Then, we split the data into training and testing sets.

### PCA

The variance of our data, due to their different nature, has different orders of magnitude.
So we performed a PCA on scale data and we took 12 principal components, that explain over 99% of the total variance.

## 7. CLASSIFICATION

We performed various classification methods. For each method we got the appropriate hyperparameters; we tuned them through cross validation in order to obtain models with optimal performance.
We calculated their accuracy, precision and recall (sensitivity). In particular we calculated the F1 score, that is the weighted harmonic mean of precision and recall. The closer the value of the F1 score is to 1.0, the better the expected performance of the model is.

We report below the table of the values of the accuracy on the left and the table of the values of the F1 score for the Marginal Profit on the right.

| | Accuracy | | |
|---|---|---|---|
| **Classification Models** | **30d return** | **90d return** | **180d return** |
| Random Forest | 0.88 | 0.91 | 0.91 |
| Logistic Regression | 0.42 | 0.53 | 0.67 |
| Naive Bayes | 0.50 | 0.53 | 0.61 |
| QDA | 0.73 | 0.77 | 0.77 |
| SVM | 0.39 | 0.52 | 0.61 |
| Perceptron | 0.41 | 0.35 | 0.56 |

| | F1 score for Marginal Profit | | |
|---|---|---|---|
| **Classification Models** | **30d return** | **90d return** | **180d return** |
| Random Forest | 0.89 | 0.91 | 0.67 |
| Logistic Regression | 0.21 | 0.32 | 0.00 |
| Naive Bayes | 0.43 | 0.32 | 0.38 |
| QDA | 0.50 | 0.70 | 0.38 |
| SVM | 0.00 | 0.00 | 0.15 |
| Perceptron | 0.34 | 0.19 | 0.00 |

We noticed that in every method for longer returns the accuracy is higher.
The inaccuracy of the short terms is linked to the fact that data is mostly noise for shorter periods.
In particular, QDA has a high and constant accuracy, and this is a good result.
About the F1 score for Marginal Loss, Random Forest and QDA methods seem to have a good performance in all the period terms.

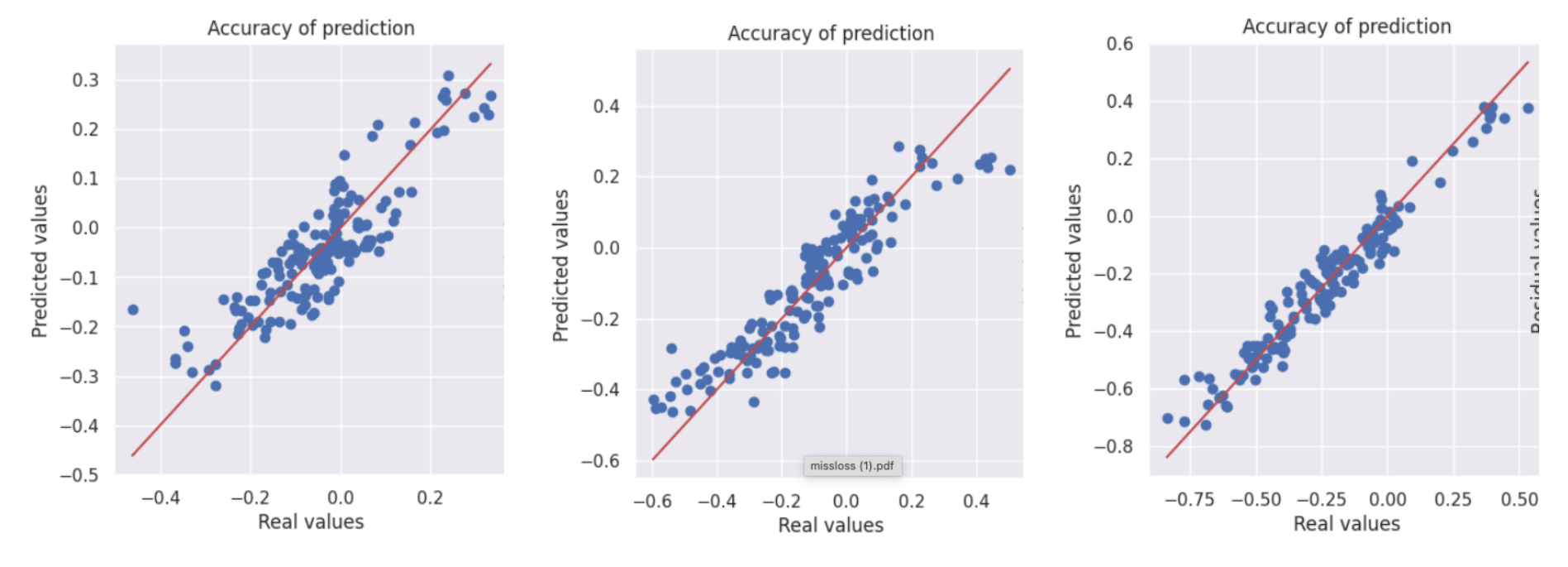
## 8. REGRESSION

We evaluated the results of our analysis in the continuous frame through various regression methods.
For each method we got the best hyperparameters through cross validation.

We evaluated their accuracy through the coefficient of determination R-Squared. From the table of values in the right we noticed that in every method for longer returns the accuracy is higher.

| | R^2 | | |
|---|---|---|---|
| **Regression Models** | **30d return** | **90d return** | **180d return** |
| Ridge Regression | 0.49 | 0.52 | 0.77 |
| Lasso Regression | 0.44 | 0.30 | 0.60 |
| SVM Regression | 0.76 | 0.85 | 0.94 |
| Gradient Boosting | 0.82 | 0.90 | 0.92 |
| XGBoost | 0.84 | 0.88 | 0.91 |

In support of the thesis we have reported below the prediction diagram of the SVM Regression method for 30_d return, 90_d return and 180_d return.



We see that for longer returns the points are more aligned and closer to the red line.

We calculated also the % of the misclassified profit (predict profit but actually is loss) and the % of the misclassified loss (predict loss but actually is profit), that are two important tools in the context of natural gas forecasting. We have reported below the tables of values of these two quantities.

| | % of Misclassified Profit | | |
|---|---|---|---|
| **Regression Models** | **30d return** | **90d return** | **180d return** |
| Ridge Regression | 16 | 8 | 4 |
| Lasso Regression | 18 | 9 | 3 |
| SVM Regression | 9 | 3 | 2 |
| Gradient Boosting | 8 | 2 | 1 |
| XGBoost | 4 | 0 | 3 |

| | % of Misclassified Loss | | |
|---|---|---|---|
| **Regression Models** | **30d return** | **90d return** | **180d return** |
| Ridge Regression | 11 | 11 | 4 |
| Lasso Regression | 11 | 18 | 5 |
| SVM Regression | 10 | 6 | 4 |
| Gradient Boosting | 5 | 8 | 4 |
| XGBoost | 7 | 6 | 4 |

We see that SVM Regression, Gradient Boosting and XGBoost have a better performance

## 9. CONCLUSIONS

We have seen how regression and classification can be used on the same data set to solve prediction problems. In particular, from the results it can be seen that regression is better than classification. However, overall QDA reveals better prediction performance compared with the other classification methods. The results demonstrate that three regression methods used have decent performance in forecasting natural gas price; these are SVM Regression, Gradient Boosting and XGBoost. These three methods obviously outperforms the other methods while Ridge and Lasso Regression are the worst. It has always been a difficult task to predict the exact price of natural gas. Many factors such as political events, general economic conditions, and traders' expectations may have an influence on it. But here, based on the past and present traits, we were able to achieve up to a good accuracy in predicting the price of any given date.