SHAPELETS

Use Case

How to forecast energy pricing using energy generation data

Index

1. Introduction 03

What is included in this eBook and why 03

Who is this eBook for 03

2. Getting started 04

The Challenge 06

Data 07

Methodology 09

Metrics 10

Synthesised Resolution 11

Results 13

How does Shapelets help solve this challenge? 17

3. Conclusión 18

4. About the author 19

5. About Shapelets 20

How to forecast energy pricing using energy generation data

Introduction

What is included in this eBook and why

Shapelets is a powerful platform for data science and data analysis that helps you uncover hidden insights. Instead of telling you about our data analytics and visualization processes, we'll show you through this use case.

In this eBook, we will explore how Shapelets accelerated platform can be used to improve price prediction. We will show you how to build a Data App using historical energy price data together with energy generation data to find a relation between the two.

Shapelets has helped data scientists solve specific problems many times, and the results have been incredibly successful. Now it's your turn to check how it works.

Who is this eBook for

This eBook was written **specifically for data scientists and will be helpful for data engineers, data analysts and business users** who are interested in building, deploying and visualizing data models.

Getting started

In this use case, we provide a toy example of the construction of a simple data analysis solution aimed at energy pricing prediction using Shapelets. In particular, we are interested in finding the relationship between energy pricing data and power generation data.

The accurate forecasting of energy prices is crucial for the orientation of the energy market and can guide policymakers and market participants, such as businesses and individuals. In practice, energy prices depend on external factors, and accurate forecasting of energy prices is difficult. Energy prices are an important factor in the economy, and accurate forecasts can help enterprises and individuals make informed energy decisions.

While this case study focuses on the Spanish energy market, the following approach is applicable to any other international market, as long as large customer databases are available.

Furthermore, while the energy pricing market needs to follow each jurisdiction's rules, the primary objective of this use case is to obtain an understanding of the relationship between energy price data and energy generation data to improve its prediction.

The use case is based on a dataset extracted from the Operador del Mercado Ibérico-Polo español (OMIE) API. This API has **1400 indicators** available for analysis, among them information on scheduled power generation, real-time generation, and energy price including intraday market session prices.

The use case is organised as follows. First, **a high-level dataset review** is performed to understand the data available and its quality. Then, **an exploratory data analysis (EDA)** is performed in order to understand how it is distributed over time or if there are patterns that can be identified at a glance. Next comes the data **modelling stage**, in which predictive models are built and the performance of the model on new, unseen data is estimated.

Finally, we will use the model that we consider the most optimal to make predictions of future days. Using a validation dataset we can easily evaluate the performance.

Moreover, this use case serves as a true example of an energy price forecast that applies to any dataset. This is great as it allows us to help make better decisions on energy usage and investment in renewable energy, and identify the best time to buy or sell electricity.

The Challange

Several challenges arise in this case study, some of which are quite common to many data science studies:

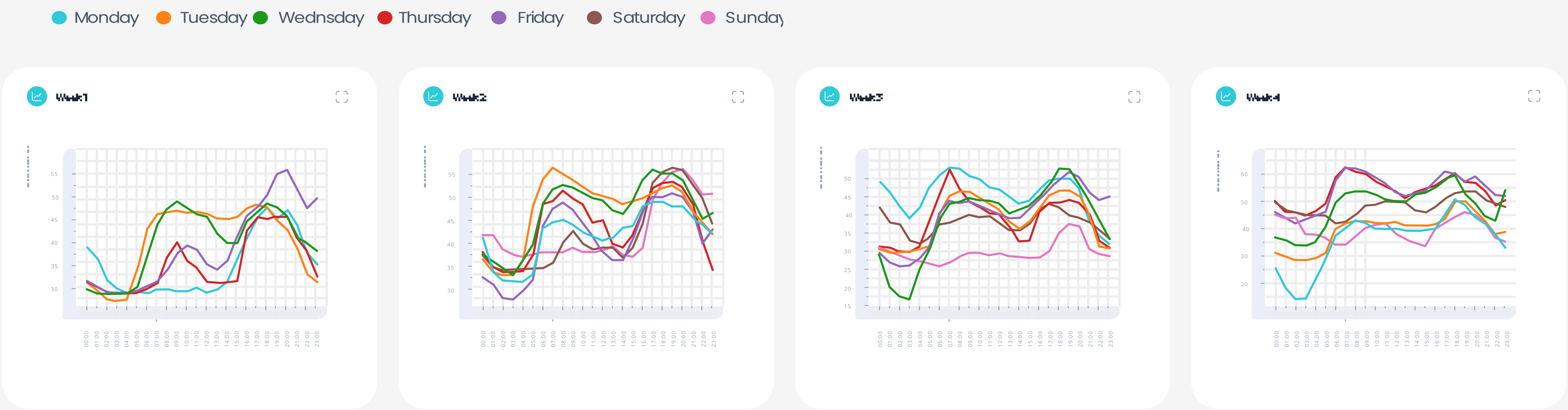
The main challenge solved with this Data App is to find some relationship between energy generation and energy price and discover that relationship using Machine Learning algorithms. Using different python libraries we implemented the best models to solve this problem.

To get this challenge it is necessary to take into account some problems such as working with datasets that have **different frequencies and also present anomalies** within the data can make it difficult to discover patterns in the data.

Data

The data used in this project is a dataset extracted from the OMIE API. This API has 1400 indicators available for analysis, among them information on scheduled power generation, real-time generation, and energy price including intraday market session prices.

For this use case, a few KPIs have been selected, because they are the most complete data (no missing data) and in the case of power generation indicators, they correspond to the energy sources that have the greatest impact on the Spanish energy market.



KPIs used for this study:

Spot Price: daily SPOT market price Freq **Hourly data**.

Real Demand: real energy demand in Spanish territory Freq **10 Mins data.**

Scheduled Generation Hydraulics: Scheduled hydroelectric power generation Freq **Hourly data.**

Scheduled generation Nuclear: Nuclear programmed power generation Freq **Hourly data.**

Scheduled Generation combined cycle: Combined cycle programmed power generation Freq **Hourly data.**

Scheduled Generation wind power: scheduled wind energy generation Freq **Hourly data**.

Scheduled Generation co-generation: programmed co-generation power generation Freq **Hourly data.**

Real-time generation Hydraulics: real-time generation of hydroelectric energy Freq **10 Mins data.**

Real-time generation Nuclear: real-time nuclear power generation Freq **10 Mins data.**

Real-time generation combined cycle: real-time combined cycle power generation Freq **10 Mins data.**

Methodology

The methodology **to predict the energy pricing** is based on three main steps commonly followed in Data Science studies:

An exploratory data analysis (EDA) in order to quickly discover important features or engineer them. In this case, we will understand the data available, its quality and how it is distributed over time or if there are patterns that can be identified at a glance.

A data processing stage to create the use case target variable, add new columns, and equalise the frequency. The objective of this use case is to predict the price of energy, so our target variable will be the Spot Price KPI, and our predictor variables, all the others. It is important to take into consideration that the data is not at the same frequency, so a resample of the data will be necessary.

Finally, we will perform a data modelling stage, in which predictive models are built and the performance of the model on new, unseen data is estimated. In this example, we will train three machine learning algorithms, Random Forest, LightGBM and XGBoost and compare their performance using regression error metrics. We will select the model that is optimal to make predictions of future time series. By using a validation dataset we can easily evaluate the performance.

Metrics

The chosen metrics are traditional regression error metrics to evaluate and report the performance of the models:

Mean Squared Error (MSE) - It penalizes larger errors because squaring larger numbers has a greater impact than squaring smaller numbers. The MSE is the sum of the squared errors divided by the number of observations.

Root Mean Squared Error (RMSE) - The square root of the MSE. RMSE is used to convert MSE back into the same units as the actual data.

Mean Absolute Error (MAE) - It is the average magnitude of the errors in a set of predictions, without considering their direction.

- **Mean Absolute Percent Error (MAPE) - The average of absolute errors divided by actual observation values.**

The target feature in this use case is the value that will be used by the Principal Component Analysis (PCA) algorithm to obtain the real price data so that a small error can mean a very large deviation in the real price data value. In the Shapelets team, we know that reducing the MAE value as much as possible is the best way to focus our efforts.

Synthesised Resolution

SEE THE FULL DATASET

This problem can be approached from different points of view.

- **As an autoregressive time series, predicting the value of the principal component of the daily data.**
- **As a multivariate time series problem, with the different KPIs, we have.**
- **As a regression problem.**

For this approach, we have treated the problem as a regression problem, in which, with the processing of the previous point, we predict the value of the principal component that we have calculated for the target.

We have used 3 algorithms:

- The **RandomForestRegressor** algorithm from the package.
- The **LightGBM** algorithm, from Microsoft.
- The **XGBoost** algorithm.

How have we implemented the algorithms?

Easy, look at this example.

With Shapelets you have all the power of Python and its different packages, so it is as easy as importing and running them!

To train the algorithm we have used data from all the KPIs from January 2015 to March 31, 2022.

To evaluate the performance of the models, we are going to split the training dataset to have a train and a test, selecting 30% of the data randomly, and calculating some metrics. You can see which model is working better below!

Predicting the price allow companies that purchase energy to adjust their budget more accurately.

We use data since January 2015 on scheduled power generation, real-time generation and prices. These data have different frequencies, there are some with hourly frequency and some with 10 Mins frequency. Visualizing the data, a price anomaly can be observed in March 2022 due to a sharp price rise.

Once the data is processed, three algorithms are proposed to predict the energy price which is evaluated using regression error metrics. These models have similar performance, giving some tighter predictions for each day.

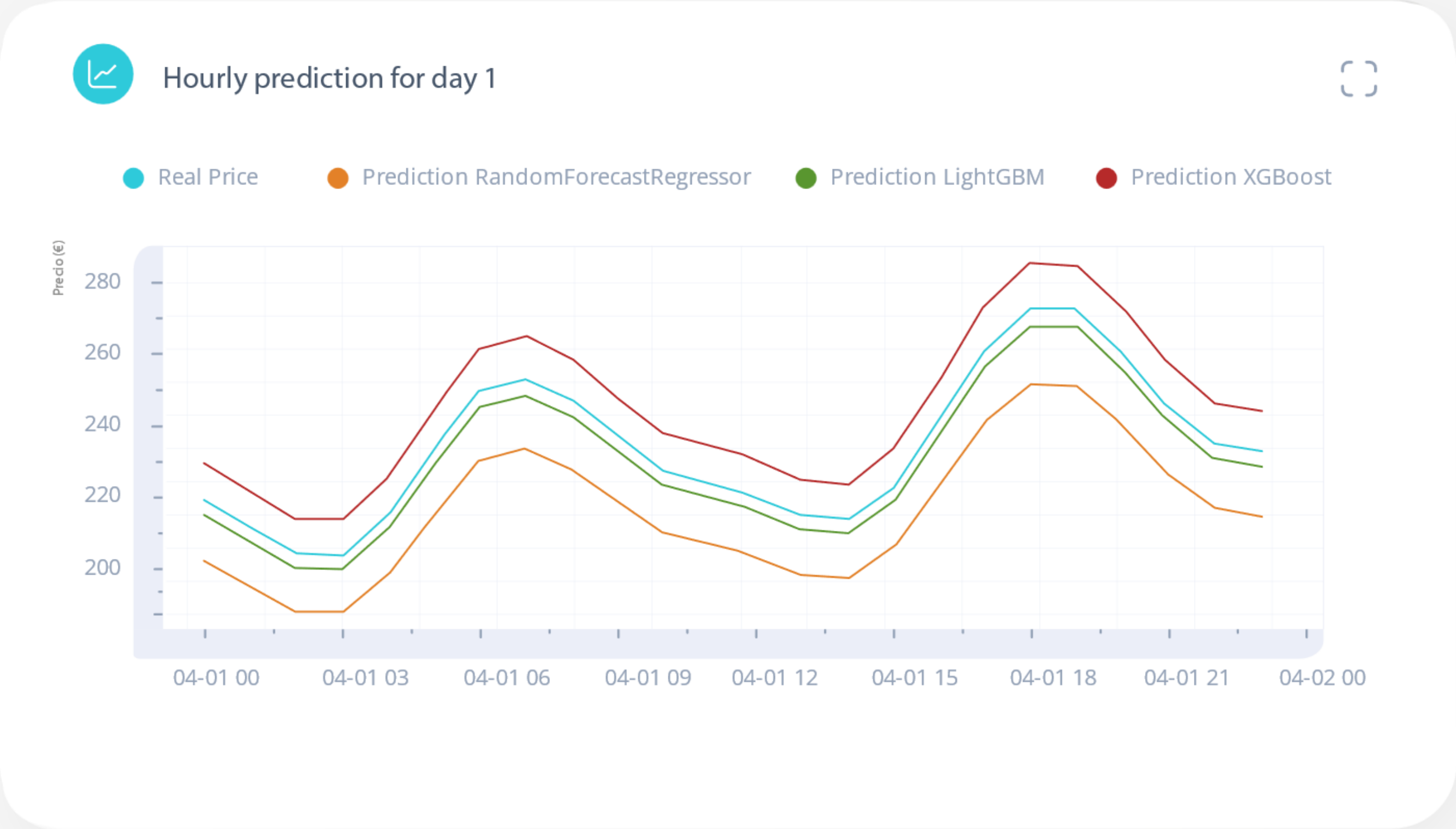
This data app implements a system to compare the predictions between them so that the user can choose which model best fits his business vision.

With Shapelets relevant metrics/KPIs can be monitored frequently and insights like the aforementioned ones can be instantly and seamlessly shared from the data scientist to all relevant departments in the organisation.

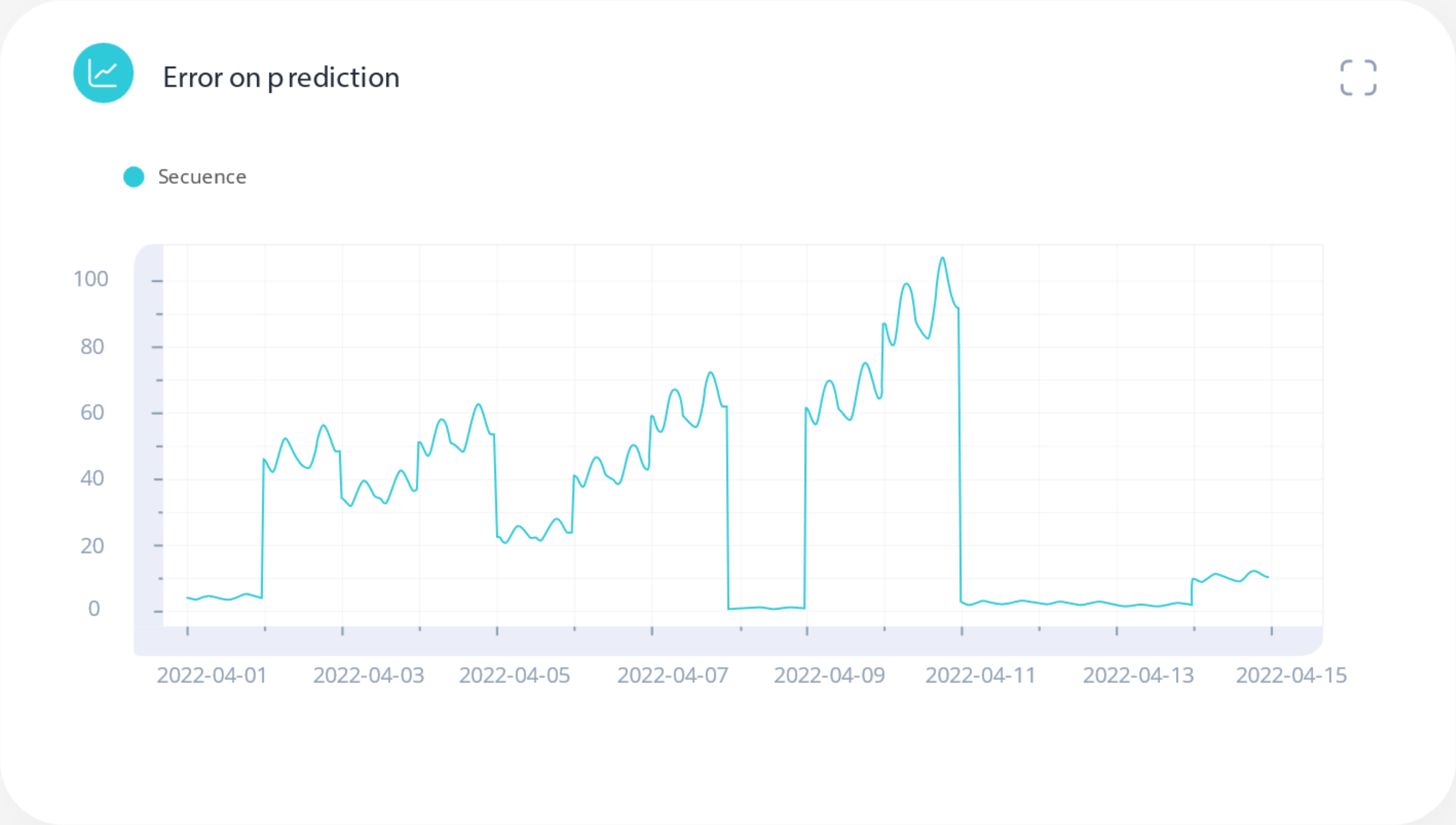
Results

Several interesting results arise from this study:

- **The first insight that is obtained is based on this study's results:** it is indeed possible to predict the price of energy with a very small error.
- **Looking at the MAE metric**, we see that the LightGBM algorithm is the model with the lowest error of the three.
- **Although it does not have as good a fit on some days as Random Forest**, on average it is the model that calculates the price with the lowest error.
- We have obtained an accurate energy price prediction **with a 1.12 average score in MAE on the target variable.**
- Additionally, if we look at it from a business perspective, **there is only a 35.77 € average difference in the pricing prediction.**



Hourly prediction for day 1



Error on p rediction

How does Shapelets help solve this challenge?

Shapelets is a great tool for solving data science and data analysis problems. Besides, it's easy to share solutions across the organization. The access to databases and distributed processing is immediate, seamless, and scalable. You don't need any coding or development experience to create powerful data apps that can be used by the entire organisation. Shapelets is well-suited for building data science solutions since it relies on several tools used by data scientists. In this particular case, the study has been developed using python and different python packages, like pandas and sci-kit-learn.

Thanks to Shapelets, we can fully develop a Data App in only 30 mins. This project serves as both an article and a dashboard to show results. This means that Shapelets allows the data professional to present a study to business stakeholders without worrying about the style, design or security of the article.

Additionally, the Data App is ready to go into production! With a few simple adjustments of no more than 10 minutes, a different approach can be settled and ready to use.

Conclusion

Find powerful insights in your big data.

Shapelets platform was designed to make your insights even more powerful and help you solve problems quickly. It was built for you and your data science team, giving you more collaboration, more analytics, and a faster way to solve problems unique to your business. We hope you found it useful and will try using Shapelets on your business projects.

If you need more info about this use case or Shapelets, please

CONTACT US

About the author

Adrián Carrio

LEAD DATA SCIENTIST

Adrián is the Lead Data Scientist at Shapelets. He received his degree in Industrial Engineering from the University of Oviedo and his PhD in Automation and Robotics (Cum Laude) from the Technical University of Madrid. Previously, he was a researcher in Arizona State University and the Massachusetts Institute of Technology (MIT) and has published more than 30 scientific publications and one patent. He combines strong expertise in Data Analysis, Machine Learning and Pattern Recognition and deploying these technologies in a variety of industrial settings. Furthermore, he has worked on numerous technology transfer projects using AI systems in various sectors for companies such as Arcelor Mittal and Airbus. He has also co-founded ThermoHuman (thermography for health and sports) and Dronomy (autonomous drones).

You can find more articles by Adrián in our Blog, and you can also follow him on Linkedin.

adrian.carrio@shapelets.io

+34 655 302 318

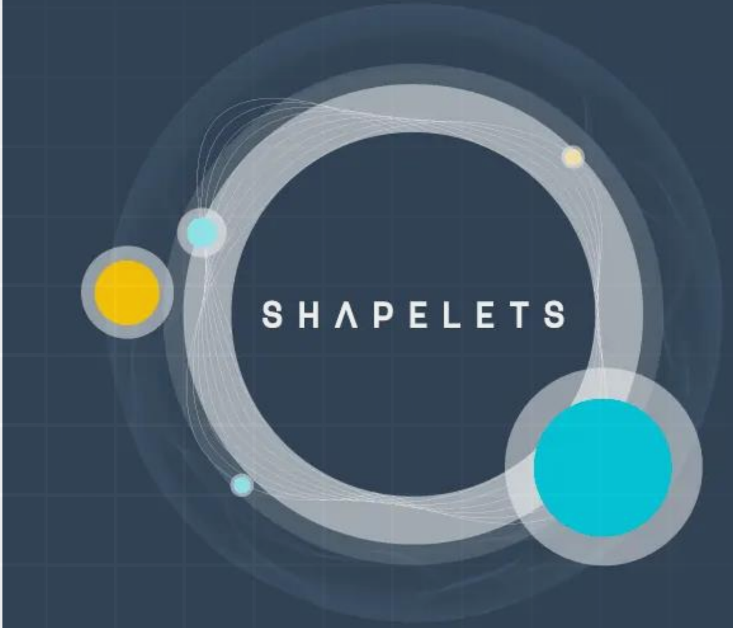
About Shapelets

Shapelets is a data-focused software company whose aim is to revolutionize Big Data analysis. Starting with an innovative analytics platform, we offer a comprehensive and flexible environment for modeling data behavior, especially time series data, using both our own and external algorithms. We leverage state-of-the-art data processing capabilities and cutting-edge visualizations.

Additionally, our enhanced data visualization tools facilitate better communication for data scientists with all departments and stakeholders within an organization.

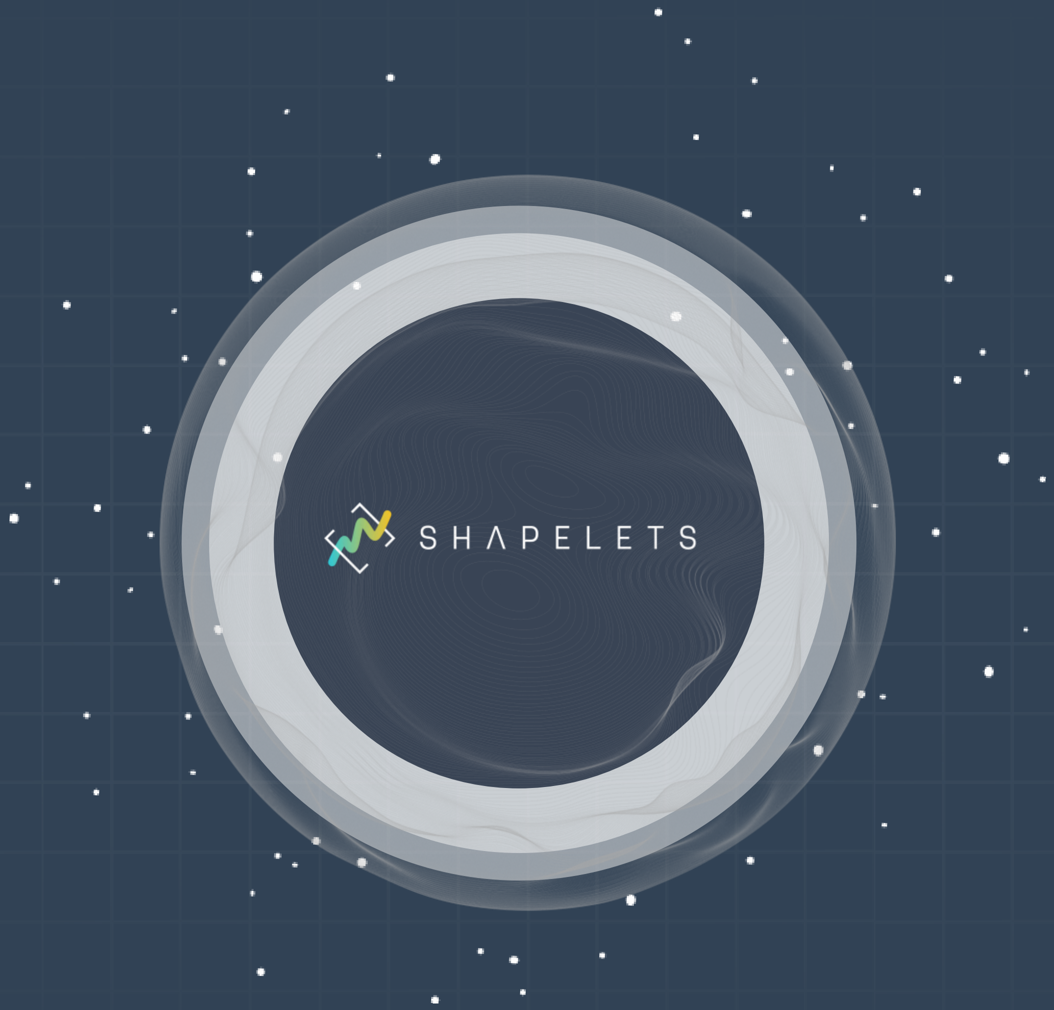
The innovation of Shapelets lies in being a solution focused on data ingestion, processing, and analysis on the platform. It can be seamlessly integrated with any local or cloud storage solution, automating modeling functions through machine learning algorithms and can operate in any environment, whether on-premises or in the cloud.

To learn more, visit our website or follow **Shapelets** on **LinkedIn, GitHub and YouTube.**



hello@shapelets.io

shapelets.io



SHAPELETS

shapelets.io