Data mining project

BIP data analysis

P. Antonini D. Azzalini F. Azzalini V. Deciano

Politecnico di Milano

Project goal

Analyse a dataset containing the **selling volumes** of two consumer products in different sales areas, in order to provide **daily forecasts** for the next ten days.

For both the two products we are considering (‘`Prodotto_1`’ and ‘`Prodotto_2`’), the dataset reports the selling volumes in 144 subareas, from 01/01/2014 to 19/05/2016. A total of **250 560** lines have to be analysed.

The whole computation is performed with a Python 3.5 script.

Results evaluation

Results are evaluated using *mean absolute percentage error* (MAPE):

$$M = \frac{1}{n} \sum_{t=1}^{n} \left| \frac{\bar{A}_t - F_t}{\bar{A}_t} \right|$$

$\bar{A}_t$ is the mean actual value, F_t is the forecast value.

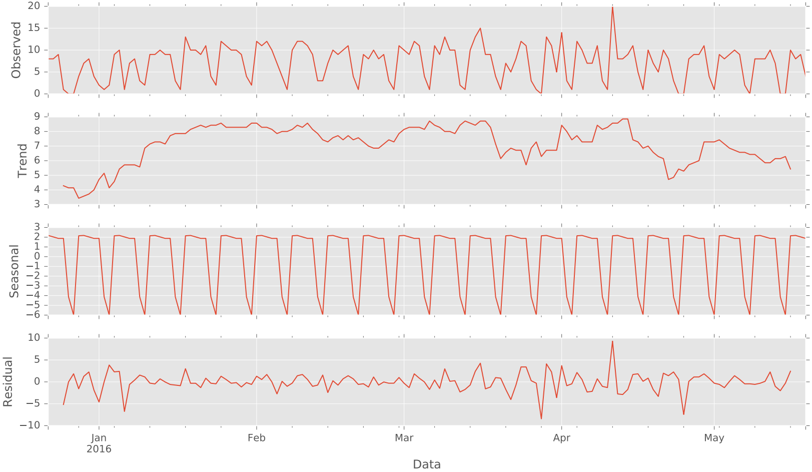
First approach

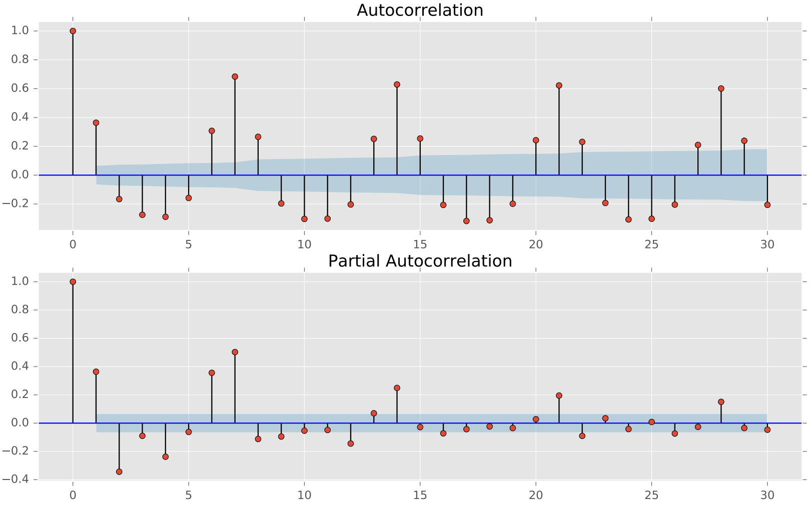
The most common way to cope with time series analysis and forecasting is by using a class of models called **autoregressive integrated moving average**, or ARIMA.

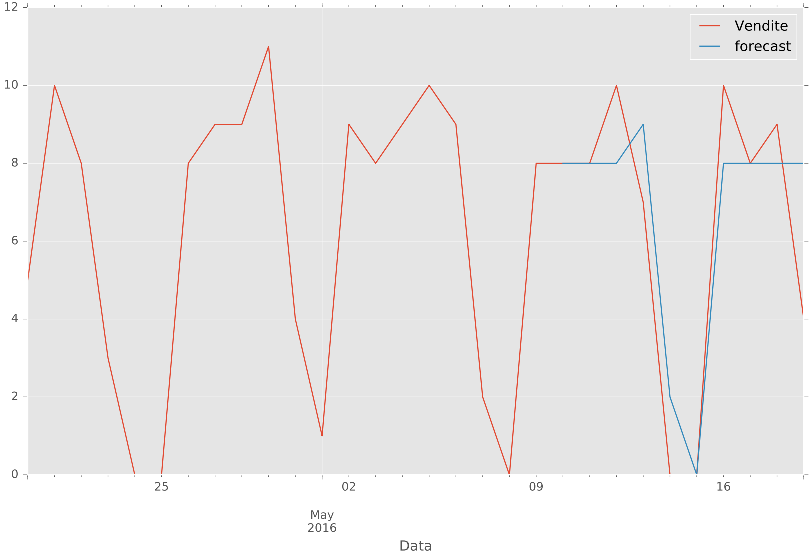
In order to be a good candidate for an ARIMA model, a series has to be made **stationary**. Stationarity is important, because when running regression we assume that all observations are independent of each other. However, in a time series observations are time dependent.

The phases

- We **evaluated the stationarity** of the series with the Dickey-Fuller test, and **tuned the parameters** of the model.
- A **weekly trend** was clearly evident, so we decided to use a seasonal ARIMA model with the seasonality set as weekly.


Observed
20
15
10
5
0
Trend
9
8
7
6
5
4
3
Seasonal
3
2
1
0
−1
−2
−3
−4
−5
−6
Residual
10
5
0
−5
−10
Jan
2016
Feb
Mar
Apr
May
Data

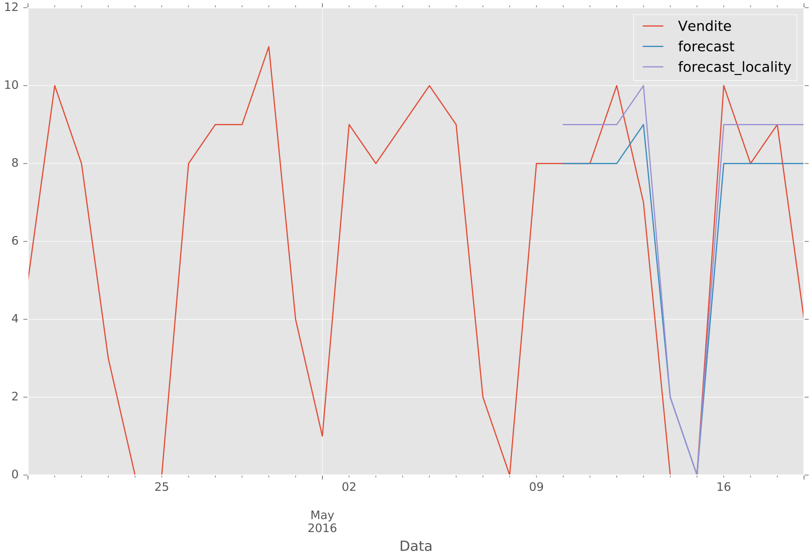

Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
0
5
10
15
20
25
30
Partial Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
0
5
10
15
20
25
30


Vendite
forecast
12
10
8
6
4
2
0
25
02
09
16
May
2016
Data

The correction

It seemed reasonable to **correct the results with the spacial information** we had.

We added the weighted average of the difference between the forecast in the considered subarea and the forecasts in the subareas nearby, divided by a given factor.


12
10
8
6
4
2
0
25
02
09
16
May
2016
Data
Vendite
forecast
forecast_locality

Results

The correction was not beneficial at all, as it actually **increased** the errors. The errors before the correction were:

- `Prodotto_1`: average MAPE of 33.85 %;
- `Prodotto_2`: average MAPE of 50.77 %;

Second approach

Using an ARIMA model could seem **mathematically reasonable**, but the **results were pretty poor**.

We exploited machine learning techniques: **linear regression**. Although these models are not specifically designed for time series analysis, they are so powerful that in some cases they perform better than others.

Data preparation

Before starting, we extracted some **additional features** from the data we have: we expanded the date hierarchy, by adding the day, the month, the year and the day of the week.

Also, we added the sell volumes for subareas within 75 km from the current area.

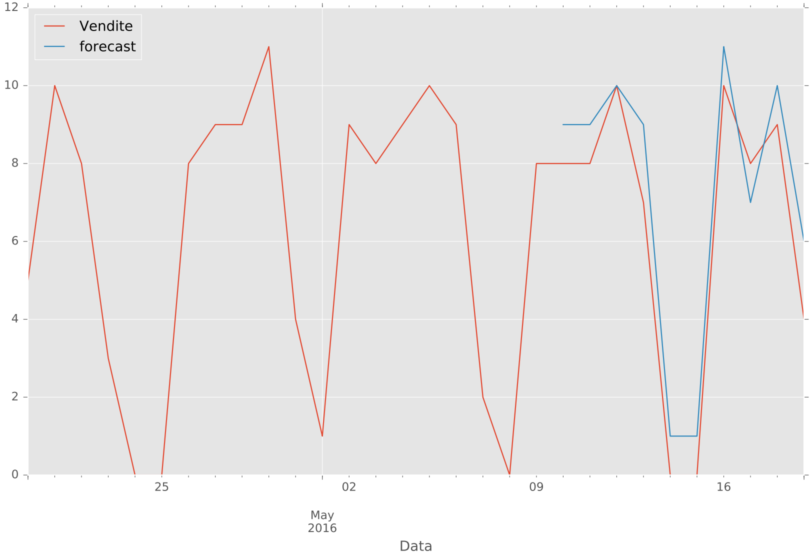
The model

We tried **different, increasingly complex, models** in the class of regression. The results were promising with every attempt.

Finally, we decided to abandon linearity and adopt **Lasso regression analysis method**.

```python
from sklearn import linear_model
model = linear_model.Lasso(alpha=0.1)
model.fit(X_train, y_train)
result = model.predict(X_test).transpose()
```

It was the model that performed **the best**.


Vendite
forecast
12
10
8
6
4
2
0
25
02
09
16
May
2016
Data

Results

By using Lasso, we got the following MAPE error values:

- Prodotto_1: average MAPE of 17.37 %;
- Prodotto_2: average MAPE of 36.73 %;

By reducing the time span to consider (instead of the whole dataset), slight improvements can be seen (Prodotto_1: 16.33 %; Prodotto_2: 35.64 %).

Final considerations

We believe that the error could be reduced even further by **taking into account other exogenous variables**, such as:

- the investments in marketing and advertisement;
- the market of the product;
- the product itself.

Also, the forecasts should be aggregated into **longer time intervals**: after all, such a fine grained insight is not needed when planning production.