Data mining project: BIP data analysis

PAOLO ANTONINI, DAVIDE AZZALINI, FABIO AZZALINI, AND VALERIA DECIANO

1. Introduction

Objective of our work was to analyse a dataset containing the selling volumes of two consumer products, in order to provide daily forecasts for the next ten days. A geolocalisation hierarchy and GPS coordinates are provided, so as to enable the investigation of possible interactions between sales areas.

After trying a time series analysis approach, which led to somewhat unsatisfactory results, we decided to adopt regression, which is a powerful machine learning technique. The results are evaluated using the *mean absolute percentage error*, or MAPE, which is defined as follows, where $\bar{A}_t$ is the average actual value and F_t is the forecast value:

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

2. Problem formulation and methodology

For both the two products we are considering, named '`Prodotto_1`' and '`Prodotto_2`' for confidentiality reasons, the dataset reports the selling volumes in 144 subareas, from 01/01/2014 to 19/05/2016. As a result, a total of 250 560 lines have to be analysed.

The nature of the dataset, containing both spatial and temporal data, suggested us to explore both aspects, starting with the temporal feature.

2.1. First attempt: time series analysis.

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 by a combination mathematical transformations, and it should also contain a substantial amount of data to work with. Stationarity is important, because when running regression (such as in ARIMA) the assumption is that all observations are independent of each other; however, in a time series observations are time dependent.

Thus, we evaluated the stationarity of the series with the Dickey-Fuller test, and tuned the parameters of the model by looking at the autocorrelation and partial autocorrelation graphs. A weekly trend was clearly evident, so we decided to use a seasonal ARIMA model, instead of a standard one, with the seasonality set as weekly.

Also, it seemed reasonable to correct the results with the spacial information we had: as a matter of fact, people tend to look for discounts and bargains, even though they have to move. So we applied a correction to the results of the ARIMA prediction, as follows: 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. This way, we thought we could capture the correlation between selling volumes between the subareas.

However, this correction actually worsened the ARIMA prediction. The average errors on the original prediction were: average MAPE of about 33.85 % over `Prodotto_1` prediction and 50.77 % over `Prodotto_2`. We believe that the main reason for such a significant error is that the values of the sells are very low, so even a single unit of difference introduces a great error.

2.2. **Second attempt: regression.** Even though using an ARIMA model could seem mathematically reasonable, the results were pretty poor, so we tried a different approach: we could exploit some powerful tools, developed in the machine learning field. Although these models are not specifically designed for time series analysis, they are so powerful that in some cases they perform better than others. We decided to try linear regression.

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

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, that performs both variable selection and regularisation in order to enhance the prediction accuracy and interpretability of the statistical model it produces. As a matter of fact, given its parameter $\alpha \leftarrow 0.1$ it was the model that performed the best.

3. Experiments and results

By using Lasso, we got the following MAPE error values: on `Prodotto_1`, the prediction error is 17.37 %, while over `Prodotto_2` we have around 36.73 %.

These results are pretty promising, even though still improvable. 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, and the product itself (luxury goods are more affected by the economic crisis than staple products; also, people are more or less likely to look for a bargain, even far from home, depending on the price of the product).

Another way of decreasing the error is to aggregate the forecasts into time interval longer than one day or into subareas: after all, such a fine grained insight is not needed when planning production.

Paolo Antonini (858242)
E-mail address: paolo1.antonini@mail.polimi.it

Davide Azzalini (855185)
E-mail address: davide.azzalini@mail.polimi.it

Fabio Azzalini (855182)
E-mail address: fabio.azzalini@mail.polimi.it

Valeria Deciano (864086)
E-mail address: valeria.deciano@mail.polimi.it