Daily gelato demand prediction for artisan shops

Melinda Demeter, Msc. in Business Analytics 2018, CEU

1 Abstract

This Capstone project in Business Analytics aims to determine the most important (non-economic) factors that affect the gelato sales of artisan gelato shops (parlours) in Hungary. To study this, Linear regression and Machine Learning based models have been built, based on sales data obtained from 9 shops located in different municipalities of Hungary.

Before conducting the analysis, my assumption was that the level of daily gelato consumption is associated with several factors beyond merely climate data. All the models built prove this, with varying degree of precision and variable importance.

Weather indeed plays a critical role in gelato production, storage and consumption. Currently most shop owners use temperature and calendar data (such as day of the week and events) only for estimating next days' consumption. Consumers are increasingly conscious and prefer products that are of high quality that is also defined by the gelato's freshness. Dealing with storage cost in case of left-over gelato, potential waste of perishable goods or lost customers in case of underproduction are a burden for gelato shop owners.

I believe that Machine Learning methods can help uncover relationships among these factors, greatly improving demand forecast, while carrying a great business potential. The results indicate that indeed daily gelato demand prediction can be improved with ML methods, as the outcomes were satisfactory, given the small dataset available..

The world artisan gelato market currently has a value of 15 billion euros, with an average annual growth of 4% between 2015 and 2018. There is an increasingly wide offer by artisan gelato parlours worldwide. [1] In Hungary in 2018 there are about 1500 gelato shops, so the studied market is quite significant at a global level.[2]

[1] https://en.sigep.it/media-room/press-releases?ext_url=/base/press_area/news_dettaglio.asp&codice=7912&Tipo=C&Pagina=%3Fext_url%3D%2Fbase%2Fpress_area%2Fnews_elenco%2Easp

[2] Landra Publishing House

2 Method

2.1 Data received:

This analysis was done based on data received from 9 gelato shop from 5 Hungarian cities. One gelato shop was able to provide data for 2 consecutive years (2016 & 2017), another for 5 consecutive years (2013-2017), while the rest for 2017. A total of 3721 daily observations were used for training and testing the model (this total excludes 0 values, i.e. closed days).

The data received from most shop owners included only the **day of sale** and the **value of sale** (for some, the HUF value, for others, the quantity in kilograms). Data was enriched with further variables were added, as detailed in the following section.

2.2 Feature engineering & Exploratory Data Analysis (EDA)

2.2.1 Target variable

The data received was in varying units and some owners substituted actual values with multiples of these, for data confidentiality reasons.

In order to make daily sales of different shops comparable, the variable G was created for assessing the level of sales of day compared to the entire season. Later this variable was converted back to the original units via a special formula.

2.2.2 Predictors

Gelato shop's profile:

As demand for gelato varies by the number of potential customers, the location of each shop was analyzed, along with other factors describing the surroundings of a particular shop that could influence the number of customers visiting the shop on a particular day.

Dummy variables were created representing the location type of a shop, the level of tourism and schools nearby, etc.

Calendar variables:

Daily operating hours, day of week variables were creates. Further dummies were added marking national holidays, events, academic calendar and long weekends.

Weather variables

The two sources of weather information used are: Wunderground and Utah State University's Climate Center. The latter was used for obtaining hourly sky coverage data, while the first was used for the rest. Data is accessible freely for non-commercial use, so it has been used for the purposes of this analysis. In case of an implemented business idea (with commercial purposes) use of an API (e.g. Wunderground) will be needed, depending on the software environment implemented. Only hourly data for the average opening hours was used in modeling.

3 Modeling – Best model

The data was split in Train / Test set in 3 ways, representing different business scenarios.

3.1 Performance of models

Different types of advanced Machine Learning models were built training set (each with several versions as means of validation and hyperparameter tuning). Caret package in R was used for modeling purposes.

Performance of these was assessed based on Root-Mean-Square-Error (RMSE) measurement. RMSE aims to measure the difference between values predicted by a model and observed values. The best model was picked based on the lowest RMSE, corresponding to the lowest residuals (RMSE for all models is listed in comparison tables).

Strong computational capacity was needed for running the most advanced models. The winning model's performance was close to the second best one and it was similar also in terms of variable importance.

3.2 Testing the model

The best one was further tested and the aforementioned conversion formula was applied.

The residual error of the winning model had a normal distribution centered around 0, as expected. Following the conversion the relative prediction error was calculated and for final assessment of model's reasonability the standard deviation of this relative error was used.

4. Error in the data

The data used is subject to various errors: errors in sales data recording; unmeasured variables; weather forecast errors. The size of the dataset

5. Future plans

More data should be collected while improving the business understanding and the model should be further tuned. The conversion formula can also be improved.