



Statistical methods used in forecasting White Paper

Quickborn Consulting LLC

February 2021



Forecasting in Retail



For a successful inventory planning in retail, it is crucial to conduct the forecasting accurately, to make sound and educated decisions in the business. The challenge to build accurate forecasts is mainly because of the difficulties in terms of validating the quality of historical data. In fact, forecasting items individually requires sufficient and consistent historical data, which is usually not available. In such events, it is desirable to forecast items according to their similarities, since that will solve the problem of low data quality caused by new items with no historical data. One solution is to pool historical data of items to one statistical model, ensuring the items behave similarly. Read on to find out how we can achieve that!

Statistical methods

One unique approach we have used in our practice is to apply statistical methods to determine what items behave similarly; by grouping items with and without history together. We estimate similarity by first conducting regression analysis on the sales profile of items against their descriptions' semantic features. Second, we predict sales by doing clustering and exponential smoothing of item descriptions. Third, we select the best fitting forecast (6-8 methods calculated per forecast) based on clusters identified in the previous step.

1st Step: Data cleaning and vectorizing- Application of the semantic latent analysis

The semantic latent analysis describes the occurrence of terms in a document. It is a sparse matrix whose rows are "terms" and whose columns are "documents".

This matrix is common in standard semantic models, such as the vector model, although its matrix form is not systematic, since the mathematical properties of matrices are rarely used.

Latent semantic analysis transforms the matrix into a relationship between terms and concepts, and a relationship between these concepts and documents. This way we can find similarities between documents.

Before processing, our data must be prepared. Hence, cleaning the data by standardizing, removing stop words and punctuations - among others - is necessary. In addition, the data is transformed into a list of separate words that is going to be transformed into a matrix of vectors.

This statistical method will help us measure the weight of a word in the description of an item according to its sales amount.

2nd Step: Similarity detection by cosine similarity and K-means

The cosine similarity, treats items as vectors based on matrix X and the similarity is computed as cosine angle between the vectors.

If \mathbf{d}_2 and \mathbf{q} are vectors, then

$$\cos \theta = \frac{\mathbf{d}_2 \cdot \mathbf{q}}{\|\mathbf{d}_2\| \|\mathbf{q}\|}$$

The purpose of finding similarities is to put items in different clusters before the Forecasting. In other words, we want to put similar products in one basket so that we can make better prediction on each product inventory stock.

Applying natural language processing by transforming words into vectors, and finding cosine similarity between items is the most appropriate method for clustering identification. In fact, it will solve the problem of items with different life cycle length and data sparsity.

We will put items into different cluster using the K-means method based on similarity scores.

The K-means is a method of partitioning data, where its algorithm is seeking to divide a set of data points into k groups, called clusters.







The algorithm will identify k number of centroids, and then allocates every data point to its nearest cluster, while keeping the centroids as small as possible. In fact, it will start first by selecting random centroids, and then perform iterative calculation to optimize their position. The number of clusters is only defined when there is no change in the values/positions of the centroids between successive iterations.

3rd Step: Applying Forecasting Methods and model selection

In this last step, we will define our parameters in the automated Machine learning forecast, and include the eligible forecasting models according to our type of data and desired outcome. It is an important requirement to maximize the performance of our model.

The forecasting models include, AutoArima, KNN, SeasonalAverage and ExponentialSmoothing among others.

After that we will include the training data that contains historical data along with the testing data, that will train and evaluate the selected models.

Method	MAE	MPE	MAPE	MASE	RMSE
					
BstDecTree	89.70721	1.577143	22.281254	1.538018	143.404917
DecFore	79.996899	-0.50047	21.588716	1.440935	125.592576
FastFore	85.78435	-0.171968	20.97652	1.447306	143.623429
snaive	97.537255	-5.255961	24.762828	1.649059	158.943538
STL_ETS	83.359802	0.0456	21.73414	1.458174	134.234719
STLF_ARIMA	73.802817	-4.140493	19.824139	1.283583	122.836633

The above table contains the error metrics

Source: AzureML Team, Microsoft, Retail Forecasting: Step 1 of 6, data pre-processing. 2015, www.gallery.azure.ai/

One crucial step is to decide on the best forecasting model for us. However, it becomes easier to make our decision, thanks to the descriptive statistics and measures of error. One of these is **MAPE**, which is the Mean Absolute Percentage Error, which we want to minimize. Therefore we want to choose the model with the smallest value in the MAPE column.

It is also important for us to identify the reasons for deviations, where adjustments to the forecast must be done with a certain level of confidence or a percentage of error tolerated.

Conclusion

Forecasting is a key issue that we face when we are planning inventory. It is very challenging for retailers to perform demand forecasting accurately. Accuracy of forecasting has a major impact on the business performance.

Applying the above strategy will help retailers produce an accurate demand forecast for items that have little or no history, providing business decision makers with the statistical support they need in driving the business towards its goals.

By evaluating and adjusting our model constantly, our goal is to apply the most advanced practices to be able to reach the highest level of accuracy on an ongoing basis.

Questions? Contact us:

<https://qbcs.com/contact-us/>

www.qbcs.com

sales@qbcs.com



QUICKBORN
www.qbcs.com