NOVA IMS Information Management School

MGI

Mestrado em Gestão de Informação
Master Program in Information Management

Retail forecasting under the influence of promotional discounts

André Neves de Almeida Roque Carreira

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre em Gestão de Informação

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

RETAIL FORECASTING UNDER THE INFLUENCE OF PROMOTIONAL DISCOUNTS

por

ANDRÉ NEVES DE ALMEIDA ROQUE CARREIRA

Dissertação como requisito parcial para a obtenção do grau de Mestre em Gestão de Informação, Especialização em Gestão do conhecimento e Business Intelligence

Orientador:PROF. DR. LEONARDO VANNESCHI

Fevereiro2017

AGRADECIMENTOS

Ao Departamento de Sistemas de Informação da Jerónimo Martins, pelo suporte oferecido durante o curso.

Aos meus pais que incessantemente me apoiam e incentivam a ser melhor.

À minha família, tios, avós e primas que tanto me alegram.

À Jovana, que encontro sempre a meu lado quando mais preciso.

RESUMO

Temos observado um aumento substancial da competitividade no negócio do retalho,onde as maiores empresas controlam o Mercado usando promoções para atrair e aumentar a fidelidade de clientes. Estas promoções causam um decréscimo significativo dos lucros, forçando um maior foco na otimização dos processos logísticos

Forçados a utilizar métodos estocásticos de modo a melhor alocar os recursos para eventos futuros, os responsáveis pela gestão de stock frequentemente recorrem a algoritmos de análise temporal ou simples extrapolações de dados históricos sob a assunção que as vendas podem ser consideradas séries estáveis e periódicas. Embora computacionalmente leves, estes algoritmos produzem resultados que estão sujeitos a uma maior incerteza devido à abordagem simplista.

Entretanto, o continuo aumento dedados disponíveis causado pelo desenvolvimento de sistemas automáticos de recolha de dados, permite a utilização de novas e mais complexas abordagens. Estas lidam com problemasde dimensionalidade, focadas em extrair conhecimento de potenciais valiosas fontes de informação.

Esta tese visa o desenvolvimento de uma solução escalável e compreensível utilizando algoritmos de machine learning para efetuar previsões das vendas diárias de artigos numa loja de retalho, sob a influencia de ações promocionais, de modo a suportar operações logísticas de gestão de stock. Este método preditivo tem como objetivo diminuir os custos de manutenção de stock e simultaneamente evitar falta de artigos, que impacta diretamente a satisfação do cliente com a marca.

O desenvolvimento de um sistema que modele automaticamente as vendas permitiria também aos retalhistas otimizar as suas estratégias promocionais com base dos resultados esperados de diferentes simulações.

Usando dados de uma das maiores companhias de retalho de Portugal, este projeto cai na definição de um problema de Big Data devido ao extenso histórico de vendas e aoelevado número de relações entre artigos que podem ser consideradas. A existência de discrepâncias entre a quantidade de artigos registada no sistema e a quantidade de artigos realmente disponível – phantom stock- será tida em conta tal como agentes externos relevantes que condicionam as vendas.

PALAVRAS-CHAVES

Gestão de stock; Big Data; Phantom stock; Machine Learning; Algoritmo

ABSTRACT

As we observe a rise of competitive pressure in retail business, major players control the market using promotions to attract and increase the fidelity of customers. These promotions cause a significant decrease in profit margin hence forcing a stronger focus on logistic processes efficiency.

Forced to make use of stochastic tools in order to better allocate resources for the future events, those responsible for assortment strategy frequently choose time series based algorithms and simple extrapolation of historical data, under the assumption that these events can be considered continuous, smooth and possibly periodic. While computationally light, these algorithms are subject to greater uncertainty due to the simplistic approach.

Meanwhile, the explosive growth of information and availability of data brought by improved automatic collection systems allow new and more complex approaches.These tackle the high dimensionality problem, focused on retrieving knowledge from potentially rich sources of information.

The work developed in this thesis aims to develop a comprehensive and scalable solution using machine learning algorithms to forecast daily sales of articles in a retail store, under the influence of discounts, as to support logistic storage allocation operations. This is done with the purpose of decreasing costs related to stock warehousing while simultaneously decreasing stock-outs as they directly influence client satisfaction with the brand.

The development of a successful automatic modelling system would simultaneously allow retailers to optimize their promotional schedules based on the expected results of different simulations.

Using real data from one of the biggest retailers in Portugal, this project's falls into the definition of Big Data due to extensive historical databases which cannot be simultaneously processed. The presence of discrepancies between registered stock and physical availability - Phantom stock - will be considered as well as relevant external events which affect the sales.

KEYWORDS

Storage allocation; Big Data; Phantom stock, Machine Learning; Algorithm

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

ANN	*Artificial Neural Network*
MLP	M*ultilayer perceptron*
AR	A*utoregressive model*
MA	M*oving average model*
ARIMA	A*utoregressive integrated moving average*
OSA	*On shelf availability*
SKU	*Stock-keeping unit*
RRP	*Recommended retail price*

CONTENTS

1. INTRODUCTION

1.1. Background and Problem Identification

In the latest years promotional discounts have emerged into retail businessencouraged by fierce competition. According to the Portuguese Association forDistribution Companies (APED - Associação Portuguesa de Empresas de Distribuição), promotions are getting more frequent. In 2014 more than a third of the sales in food retail business were associated with a promotional discounts.

These promotions are analysed, considered and renewed on a weekly, and even daily, basis in order to increase visibility and customer loyalty to the brand. Numerous studies have been developed in the field of marketresearch focused on promotions, their strategic implications and impact (Levy et al., 2004, Zhang et al., 2008).

However,beyond all strategic and planning decisions, a logistic demand planning framework must be implemented as to support the impact of such strategies(Trusov et al., 2006).

Is crucial to create an accurate demand forecasting of thousands of products, each one with its own specificity, in order to organize and plan production, purchase, transportation and labour force. Even though this aspect of promotion planning has received relatively little attention in marketing literature, possibly due to the more applied nature of process, the importance of accurate sales forecasts to efficient inventory management has long been recognized. Here, the precision of the implemented method can greatly contribute to lower storage costs and increase client satisfaction. However, a wrong estimative can cause losses either by excess or disruption.

While there are costs and losses in operational efficiency associated with stock holding, lower OSA levels caused by depletion of the product, not only decreases the sales amount as it can also greatly contribute to Consumer-Switching Behaviour(Corsten et al., 2003) once

costumers abandon the service in favour of a competitor's, progressively jeopardizing a firm's competitive position.

The impact of the improvement of these forecasts can be observed especially in big retailers, where optimization ROI (return of investment) can be more easily noticed.

Up until the last few decades, the use of the "last like" rule for ordering inventory for upcoming promotion events has been a common practice. Here, the quantity of goods ordered is the same as the quantity of goods sold during a similar past promotion (Cooper et al., 1999). However, recent advancements offered better ways for managers to handle the planning process.

Due to the strong trend and nonstationary seasonal patterns exhibited by Retail sales (Alon et al.l,2001), the most widespread methods have been Winters exponential smoothing, seasonal autoregressive integrated moving average (ARIMA) model and multiple regression which have the ability to model trend and seasonal fluctuations presented by aggregate retail sales.

These Linear time based models and regression algorithms have long supported logistic operations requirements with acceptable accuracy. However, the linearity in which these algorithms most effectively perform has been compromised. Due to economic instability and more fierce competition (Geurts, 1986 and Gür et al. 2009), recent retail sales time-series data show a higher degree of variability and nonlinearity, which decreases the accuracy of these models, hence justifying the use of nonlinear models such as the Artificial neural networks which have consistently showed to be more precise (Zhang et al., 2003 and Alon et al., 2001).

Regardless of the innovation brought from the mentioned works, external factors, frequent and irregular promotional discounts at the article level have not been thoroughly considered.

This work will focus on the analysis of some of the most relevant algorithms used for forecasting demand, comparing their performance.

Market tendency, external events and the discounts of not only the products analysed but also the chosen promotional strategy for all the store will be consideredin the interest of

estimating the consumption of customers. By using real data from the industry, the presence of Phantom Stock, Big Data, meaningful external influences and a wide range of promotions will characterize the challenge.

1.2. Study Objectives

The objective of this study is the creation of a scalable front end tool to improve on shelf availability (OSA). This will be achieved by creating models which allow an automatic and more accurate prediction to preventively avoid OOS and therefore, the loss of sales.

This approach will take into consideration several internal and external factors which affect demand. We will consider events which are not controlled by the business, such as the weather and main sport events, as well as variables which describe used marketing strategy.

The precision of different Machine Learning Algorithms will be tested in the interest of asserting the best model for this problem, their performance will be validated upon a training set of four weeks of daily data using nearly two hundred individual SKU (stock-keeping units).

This projects has also a strong visualization component where we intent to transmit complex results in a simple way. Graphics must be designed to support the user and allow easy comprehension of the reasons behind erratic results of the model.

Therefore, alongside the research development, the following objectives should be fulfilled:

- Literature review of previous works in Demand Forecasting using machine learning algorithms.

- Identify the external factors that should be considered and develop methods to extract them automatically.

- Identify empirically the level of aggregation that should be considered given the computational requirements and logistic characteristics.

- Consider to what extent the demand of a product can be affected by discounts in similar and different categories.

- Creation of a visual user interface which can analyse historical data and a promotional discount strategy and deliver a report on expected demand.

1.3. Theoretical Framework

The level of sophistication of the methods used by the retailers can greatly differ, being the most basic SKU sales forecasting methods, univariate forecasting models based on time series which analyse past sales history in order to extract a demand pattern that is projected into the future.

SinceRetail Sales data is rich in trends and seasonal patterns, the failure to account for these patterns may result in poor precision.

These time series techniques range from simpler moving averages and exponential smoothing to the more complicated box-jenkins ARIMA approach (Box and Jenkins 1974).

However these methods do not take external factors such as price changes and promotions into account (Alon et al., 2001).

(Gür et al. 2009) found that simple time series techniques perform well for periods without promotions. However, for periods with promotions, models with more inputs improve accuracy substantially. Therefore, univariate forecasting methods are usually adopted as a benchmark model in many studies (Gür et al. 2009, Huang at al 2015).

Numerous studies, in order to improve SKU sales forecasting in the presence of promotions, have integrated the focal product's promotional variables into their forecasting models. In practice, many retailers use a base-times-lift approach to forecast product sales at the SKU level (Cooper et al. 1999, Huang et al 2015). The approach is a two-step procedure which initially generates a baseline forecast from a simple time series models and then adjusts for any incoming promotional events. The adjustments are estimated based on the lift effect of the most recent price reduction and/or promotion (Ma et al. 2014). the limitation of these studies is that they overlook the potential importance of price reductions and promotions of other influential products.

Another stream of studies uses a model-based forecasting system to forecast product sales by directly considering the promotional information. A possible choice is to use multiple linear regression, but other methods capable of predicting a continuous target can be applied - including powerful non-linear methods such as support vector machines for regression, model trees (decision trees with linear regression functions at the leaves) and Neural Networks.

Often the most popular methods, both the ARIMA model (Box & Jenkins, 1974) and ANN have shown advantages forecasting trend and seasonal data, hence the motivation focused on these methods by the latest studies (Thiesing & Vornberger, 1997). (Zhang and QI 2005) concluded that the overall out-of-sample forecasting performance of ANNs in predicting retail sales is not better than ARIMA models without appropriate data pre-processing namely deseasonalization and detrending. Hence, the use of seasonal dummy variables as auxiliary inputs in the ANN (Zhang et al., 2003, Zhang, 2009 and Alon et al., 2001). Using this method, (Pan et al. 2014) demonstrates the potential of applying empirical mode decomposition (EMD) in forecasting aggregate retail sales. Here, he integrates EMD in neural networks showing how the new hybrid outperforms both the classical ANN model and seasonal ARIMA during the periods in which macroeconomic conditions are more volatile. Other Hybrids have been considered.(Aburto and Weber 2007) developed a hybrid intelligent system combining ARIMA type approaches and MLP-type neural networks for demand forecasting that showed improvements in forecasting accuracy. Their results led to fewer sales failures and lower inventory levels in a Chilean supermarket. (Au et al. 2008) studied the use of evolutionary neural networks (ENNs) for sales forecasting in fashion retailing, obtaining this way a faster convergence and more accurate forecasting than the fully connected neural network.

The collective results indicate that on average ANNs fare favourably in relation to the more traditional statistical methods, mainly due to their ability to capture the dynamic non-linear trend and seasonal patterns, as well as the interactions between them. However, despite its simplicity, the ARIMA model was shown to be a viable method for forecasting under relatively stable economic conditions, performing consistently well.

By using real data from the industry, the current work will analyse the performance of these widespread methods and compare them.

1.4. Study Relevance and Importance

The replenishment system is crucial to maintaining competitive advantage in Retail business and it can strongly be affected by forecasting models as they directly affect early order commitments (Zhao et al., 2001). Still, Out-of-stocks remain a large problem for retailers in the consumer goods industry. According to ECR Europe (2003), the latest advances in supply chain management, category management and investments in inventory-tracking technology have not reduced the overall level of out-of-stocks on store shelves from what was reported in previous studies. The Retail Chains are still far from delivering near-perfect fulfilment. Out-of-stock rates vary wildly among retailers but the majority tends to fall in the range of 5-10 percent.

Out-of-stock items relate to a waste of time and money, but more importantly, stock-outs largely contribute to consumers switching to other retail stores to fulfil their demands.

As to decrease costs related to stock warehousing while simultaneously decreasing stock outs as they directly influence client satisfaction with the brand, the proposed work will focus on the creation and performance evaluation of an adaptive models. This model will be used in the implementation of a front-end user interface which will allow the user to get an estimative of future demand given the specified promotional marketing strategy.

This project will take into account the market tendency, external events as well as discounts, not only of the products analysed, but also the chosen promotional strategy for all the store.

1.5. Methodology

This work will initially rely on an exhaustive literature review of the state of the art of Retail Demand forecastingand the most widespread predictive algorithms – Multivariate linear regression, ANN and ARIMA.

The following milestones werefollowedto complete the work:

- Extraction of data to local database (historical sales and Article dimension data) simultaneously masking it (due to legal obligations);
- Extraction of external relevant data (holidays, football games, weather data);
- Pre-processing the available data:
    - Extract raw data from the local database;
    - Structure input sales data, article information and promotional historical data;
    - Standardization and uniformization of data;
    - Identify and handle missing data;
    - Filter outliers;
    - Identify and handle Phantom stock.
- Development of the multivariate linear regressionalgorithm;
- Development of the Neural Network model;
- Development of the ARIMA model;
- Analysis of each of the models performance and limitations given out of sample data – Test Data;
- Development of profiling tools so that the behaviours and registered errors can be easily interpreted by visualization.
- Analysis of the performance of a combined solution.

The pre-processing steps are executed whenever there is a call to get data, either for training or testing. The data resides in its original state upon extraction. Permanently assuming the possibility of corrections and updates from the original source.

1.6. Thesis Outline

The rest of the document is organized as follows:

- Chapter 2 presents a project overview of predictive modelling andfocused on the pre-processing of the input variables required to improve data quality and secure viability of the trained models;
- Chapter 3 tackles the different input variables considered, listing the sources and the types of variables used;
- Chapter 4 presents and describes the used predictive models: Multiple linear regression, ARIMA and neural networks;
- Chapter 5 discusses the obtain results of the project through different perspectives including the results of a combined model;
- Chapter 6 contains the conclusions of this thesis;
- Chapter 7 mentions future work proposals;

2. PRE-PROCESSING

2.1 Predictive Modelling concepts

As we intend to estimate the value of an unavailable variable, based on experience, we enter the field of predictive modelling.

There are two main types based on the characteristics of the target variable:

Classification – The intent is to predict a discrete variable/assign a label;

Regression – Once the problem implies the prediction of a continuous variable as the daily amount of sales.

In order to achieve our goal, a descriptive model must be created.

The construction of this model will have into account passed history with known results. Hence this is classified as a supervised learning problem, we know the true value of each daily sales in the training dataset and a feedback for each training prediction exists.

As seen in the Figure 1, in order to create the model, initially we must gather training examples. Reliable data can be hard to gather and, atthis step, the assumptions we chose can derail the model, thus, a pre-processing is generally necessary in order to secure the quality of the input data.

The used attributes must be chosen carefully, generally through an iterative process.

Choosing a reduced number of attributes can result in insufficient descriptive power over the behaviour of the dependent variable. Choosing an abundant number of variables generally in the inclusion of spurious relations - relations with not translate any causality - which lead to mistakes in the future.

The choice of input space is often limited by the performance of the training step. The larger the input space the more data and computing power we need. This can be tackled with methods for size reduction of the input space as Principal Component Analysis (PCA).

When applied to the reality, the training data can define the viability of the model, not only the choice of attributes but also the considered data points can have severe impacts on its performance.

Real data suffers from several problems, such as, outliers, incorrect data, variables with different scales and omitted values, which can easily compromise the accuracy of the developed learning algorithm. The methods used to expose data in the best conditions are detailed in the following chapter.

Afterthe attainment of the input data set, we proceed to feed the training data to a choose an algorithm with intention of extracting the descriptive relations (knowledge) between the independent variables and the target, assumedly dependent variable.

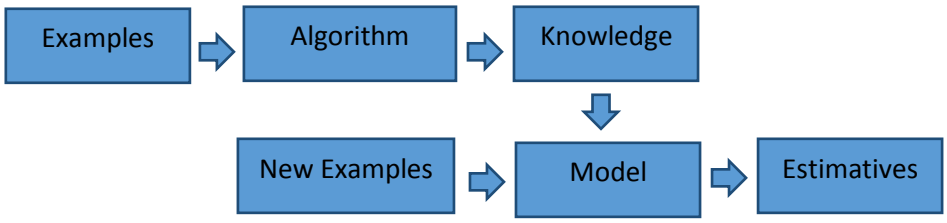


Figure 1Predictive model creation schema

One of the focus during the training phase is assuring that the model has generalization properties.

Since the complexity of the training phase is not being limited, the model can memorize the training data, over focusing on its specific properties and idiosyncrasies. A phenomenon known as over-fitting occurs as it will base the forecast on unique particularities of the training set.

As a solution, a split of the original data is executed and the secondary disconnected set – validation set – is used to monitor the evolution error of the model across time. By testing the evolution of the error on this set, while using exclusively the training set to develop the model, we can realize when the model is becoming too complex and stop the training at that point in time. By doing so, we secure generalization.

A common practice is to divide the data to allow the creation of a third dataset, the testing set. The former will be used to produce an estimative of the expected error, providing the expected precision of the model when applied to new data.Unlikely the former sets, the estimative of the error given by the testing set will not suffer favourable skews since it was not used directly nor indirectly on the training phase.The sets are disjoint, random selection without reposition and the configuration of the partitioning by the three groups has an impact on the results.

Increasing the size of the training dataset ensures a better regression while decreasing it may increase the probability of finding properties which are not transportable to a bigger set.The enlargement of the validating set allows a better estimative of when to stop the training whereas the enlargement of the test set allows a better estimative of the expected error once applying the model to new instances.

There are no concrete answers to the optimized split of the data as it is dependent on the problem in hand. Still, many distributions are commonly choosen as rule of thumb as the 80-10-10 split. (80% of your data for training data, 10% for validation, 10% for tests) or the split 50-25-25.

The training step can be divided into two categories: Incremental (e.g. Neural networks) and batch (e.g. Multivariate linear regression). The former training method processes the input entries individually, provoking transformations sequentially while the latter simultaneously uses all available data points simultaneously.

2.2 Retail Idiosyncrasies

A particularity of retail data is the heterogeneity of the forecasting targets, the individual product sales. Products can have short or long expiry dates (e.g. meat and honey), they can present a higher seasonal component in sales (e.g. ice creams) and can exist solely during short periods of time where they will possess big discounts in the interest of attracting new clients.

Furthermore, the data is often unreliable mainly due to the human factor in the stores. Products can be transformed and sold under other labels (eg. fruits can be cut and sold in more convenient packages upon ripening), can be stolen or suffer damages which skew their stock values from reality and can be a target of promotions at store level if approaching the technical expiration date.

In orderto prepare the data and expose it in the best conditions to the model, preprocessing steps are required. This chapter will present the problems which have been tackled in order to expose the data in the best condition of being analysed.

2.3 Aggregated data

The used data consists of an aggregation of daily sales. It is the result of an SQL querry directed to a Stock fact table. Here the data is aggregated into the total values of the store per day per article, as seen in Table 1.

	DATE	STOCK	QUANTITY	SUM
1	16-09-2015	100	8	18
2	17-09-2015	92	5	10
3	18-09-2015	87	8	22
4	19-09-2015	79	29	43.5
5	20-09-2015	50	23	34.5
6	21-09-2015	27	5	18
7	22-09-2015	22	7	14
9	24-09-2015	15	5	26
10	25-09-2015	10	4	26

Table 1 Example of used sales data

The “STOCK” field indicates the theoretical stock in the store at the beginning of a given day. As discussed later, this information is not reliable, its error grows daily and is rectified by manual stock audits. The fields “QUANTITY” and “SUM” describe the daily totals of the sales quantity and the sum of the recommended retail price sales (RRP).

As a direct consequence of the aggregationinformation is lost. No precise identification of the promotional discount is known and as such it must be computationally estimated. We can estimate the daily RRP by dividing the sum by the quantity, as presented on expression [1]:

$$\hat{RRP}_i = \frac{SUM_i}{QUANTITY_i} \quad [1]$$

Taking Table 1 as an example, we observe that the individual RRP generated by the first three lines is 2, whereas the line 4 and 5 present an estimated RRP of 1.5. We can therefore compute the daily discount (DESC) as [2]:

$$\widehat{DESC}_i = \frac{RRP_{max} - \hat{RRP}_i}{RRP_{max}} \quad [2]$$

RRP_{max} is the maximum individual price registered for the given article across the whole dataset. It is worth noting that this computation brings an additional disadvantage as the computation is vulnerable to outliers in RRP_{max} .

Keeping in mind that the discount we compute is an estimative, which intuitively becomes one of the most meaningful predictors we consider. Unfortunately, there is plenty of information loss in this computation. We cannot distinguish promotional pack sales as a "take 2 pay 1" from a daily 50% discount,neither can we individualize the portion of sales which were caused due to promotions fomented by approaching expiration date. Nevertheless, the sales grow unidirectionally as the computed discount grows (Figure 2), making it a viable input for forecast prediction.

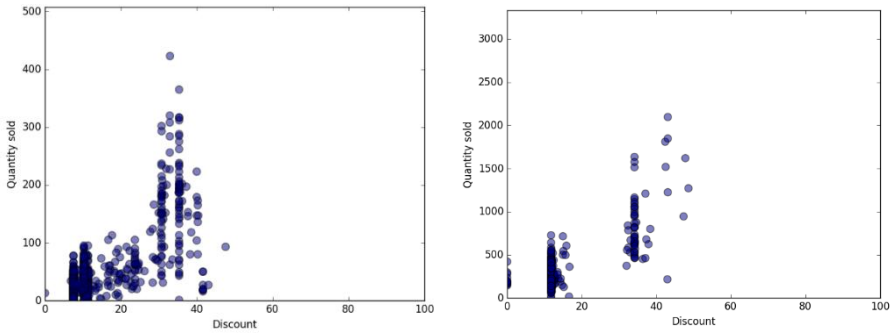


Figure 2 Sales variation with discounts in two articles

2.4 Missing Data

One of the characteristics of the used data is the lack of entries of days without sales as seen by the missing entry of the day 23 in Table 1. This is a characteristic imposed by SQL

filtering when extracting the data. Without the filter, the day 23 would present the following values:

DATE	STOCK	QUANTITY	SUM
23/09/2015	15	0	0

Table 2Filtered entry

Whichwould prevent the estimation of the daily RRP.

One solution to handle the missing data problem is by simply eliminating these entries. In the current paradigm, this solution cannot be considered as it would skew our dataset and lead to the loss of information of days without sales, severely compromising the viability of the model once applied to reality. Another reason of why these days cannot be omitted is due to the constructions of predictors, such as the previous month sales, which demand a continuous history.

Another solution is to fill the omittedRRP values with a measure of central tendency, such as the mean. This solution cannot be considered since it would imply setting a discount in each of these entries. We can, nevertheless, fill the missing values with the ones of similar individuals, or more precisely, by filling missing entries with values corresponding to linear regression of the values of individuals in both extremities – closest neighbours. Following the given example, we assign 2 to the RRP of the day 23. These estimated entries will be considered when rendering variables for future data points (such as weekly price variation) but, since no true sales value is known, the points themselves will not be considered as valid training or validation dataset.

2.5 Incorrect Data

When results suffer forced deviation by external variables, we should not consider all the dataset for training. Such is the effect of the stock uponthis project since the estimator is designed to consider full stock when forecasting future sales.

Phantom Stock

Building a reliable OSA performance indicator is hard to measure without a considerable human effort or auxiliary sensors to accurately identify its decreases.

Either due to logistic limitation or human error, articles have their sales reduced by not being available in good conditions in their shelfs at a given point in time. This contributes to faulty input data as the conditions described by its marketing strategy are disconnected from the sales result. Even if there is stock of a product in the store, it can be currently in the warehouse and not on the shelf. In fact, through the eyes of the HQ there is always uncertainty regarding the availability products in the shelf, as well as their conditions instore, the proper pricing, labelling and placement.

Either due to theft, damages or human error, stock reported to exist in the store, while in fact does not, is designated by Phantom Stock. This discrepancy between the theoretical stock and the real one can be seen as an error which gets periodically corrected by the store personal upon systematic inventory audits.

Incorrect Data

Independently of all other variables, the stock majorizes the sales. It forces the sales to zero when it simultaneously is zero and it limits the maximum sales amount at all times. Both this situations, caused by human errors or logistic constrains, temporarily disconnect the input configuration of all the remaining variables and the obtained result.

In order to protect the models from the skew effect caused by examples which are compromised due to stock limitation, we classify four different causes of deviated results which should be filtered:

- **Obvious stock out** – When the stock attribute reaches 0, we can securely assume the veracity of these situations and eliminate them.

- **Sales limitationby stock** – Although the degree of limitation cannot be verified (as the last product sold could have been the last to be sold even if there were more units to sell), these entries are identified when the sales amount reaches the stock amount. They will also be filtered.
- **Low shelf availability** – In the case not all the units are in the shelf and the ones that have been were sold. While not forcing the sales to zero, this will contribute to a decrease of the daily sales. These entries can easily be indistinguishable from usual entries, especially in the case of slow movers.
- **Phantom Stock** – In these situations the sales, just as in the previous case, are diminished by the lack of product in the shelf. Here the sales amount is little or inexistent due to the lack of product in the store when, theoretically, stock should be available. A supervised classification model needs to be created in order to label historical data as phantom stock, so that these entries can be filtered accordingly.

It is worth to notice that the implementation of a forecasting tool could easily identify punctual cases of low OSA and phantom stock on the go by creating stochastic alarms accordingly with the daily or hourly sales deviation from the estimative.

When considering historical data, the removal of these entries will most likely remove correct entries in which the null sales were caused by other agents.

2.6 Dimensionality

The problem of dimensionality is in fact a good problem since its opposite would imply lack of information to model the system.

The direction of this project assumes that the relevant data has been captured and distributed among a vast number of variables. It focuses in the sales of SKU level and, by doing so, faces the obstacle of ultra-dimensionality of the variable space.

We consider that the sales of an article are potentially affected by the promotions and stock of other products (e.g. the amount of fish sold is affected by the price of meat) and these inter and intra category effects need to have be taken into account when producing operational forecasts. Ma's empirical study (Ma et al., 2014) show that the models integrating more information perform significantly better than the baseline models which focus solely on the SKU's own predictors. Nevertheless, the identification of potentially influential categories from such a large set of possibilities(and further analysis of these effects) pose a serious modelling challenge. As (Ma et al., 2014) points out, the main challenge to be faced is that the dimensionality of promotional explanatory variables grows very rapidly when cross-product promotional information is considered, potentially much larger than the length of SKU sales.Even with the latest improvements in the standard computer performance, dimensionality can still lead to systems too complex to satisfy the maximum accepted time thresholds.

Product sales history, intra and inter category promotional schedules are all potential rich sources of information which may influence forecasting accuracy. The larger the input space the more data and computing power we need. A sub selection method is necessary in order to limit the size of the input space. So, which sources of information should be used as input into the forecasting model? As (Fan and Lv, 2008)pointed, the correlation between important predictors and unimportant ones grows with dimensionality, as does the maximum spurious correlations.

Trying to find the most influential variables is not an easy task. In the retail context, the similar products tend to be promoted simultaneously (e.g. several tastes of the same brand of yogurts) which will generate strong correlations, a spurious relation would be identified as an important predictive variable. Despite that, our decision, due to performance restrains, shall be a sub selection of aggregated indicators of the vast number of variables involved (chapter 3.1), at the cost of information.

2.7 Outliers

Given a dataset, observations beyond normal amplitudes are classified as outliers. If considered in the training process, these observations, similarly to those containing phantom stock, can skew the model. An example of outlier are the observations obtained during rare promotional events which cover all the stores (a universal discount throughout the first days of the store opening would classify as such). If no descriptive variables have been created to allow the model to treat these examples, then the model would consider them as one of the expected behaviours towards the given predictors and be significantly moved. We perform an outlier treatment, avoiding their significant weights.

Commonly the outliers are filtered from the data set, being treated individual, if a study of such situations is required, or just excluded. In the current work, for each article a top percentile of 1% in sales sum amount is established for a given training set. Any example containing a result of sales superior to that amount is automatically excluded, protecting the model from outliers.

2.8 Normalization

Among the predictors, scales of great amplitude tend to dominate and have higher weight onto the final classification.If one input has a range of 0 to 100.000 while a second input has a range from 0 to 10, the second input will be swapped by the first. Thus.A common practice in machine learning implementations is feature normalization, standardizing the range of independent variables. By putting all variables under the same scale and implicitly ensuring that all features will possess equally weighted representations, we protect our models from saturations and optimize their performance. One of the most common normalization methods is the generalized Max-Min method:

$$Y' = (\frac{(Y - Min_1)}{(Max_1 - Min_1)} . (Max_2 - Min_2) + Min_2 \quad [3]$$

where a normalized vector Y' is computed from a previous vector Y, given the maximum and minimum values of the original vector (Max_1 and Min_1 respectively) and the desired minimum and maximum boundaries of the newly normalized vector (Max_2 and Min_2). The Max-Min applied in the training uses as parameters the limit values of this initial set. All further normalization applied to new examples will use the same pre-established parameters, as to maintain coherency in the used data. In the current project, all the input variables have been normalized using the Max-Min and their values have been limited to the val$[-0.9,0.9]$.This way, we account for future observations of the data, which we intend to keep between $[-1,1]$.

3. INPUT DATA

3.1 Market Discount Strategy

The daily discount associated with each article is intuitively the biggest influencer of sales. As such, the daily discount has been taken into account (as explained in page 13). But the influence of the discount is not constant. One article which is presents the same discount throughout the week will probably have its impact progressively decreasing. In order to insert this information into the model, two auxiliary variables have been created: Week_desc_var reflects the difference of the discount once compared with the average of the previous week while Short_desc_var points out the variation of the discount when compared to the average of the last three days.

$$\widehat{Week\ Desc}\ Var_{\mathrm{t}} = \widehat{DESC}_{t-1} - \frac{\sum_{i=t-7}^{t-1} \widehat{DESC}_i}{7} \quad [4]$$

$$\widehat{Short\ Desc}\ Var_{\mathrm{t}} = \widehat{DESC}_{t-1} - \frac{\sum_{i=t-3}^{t-1} \widehat{DESC}_i}{3} \quad [5]$$

An example of these variables over time can be observed in Figure 3. Here the decline of the impact of the discount can be observed during the month of May, where the sales decreased over time together the variation variables, while the discount remained constant.

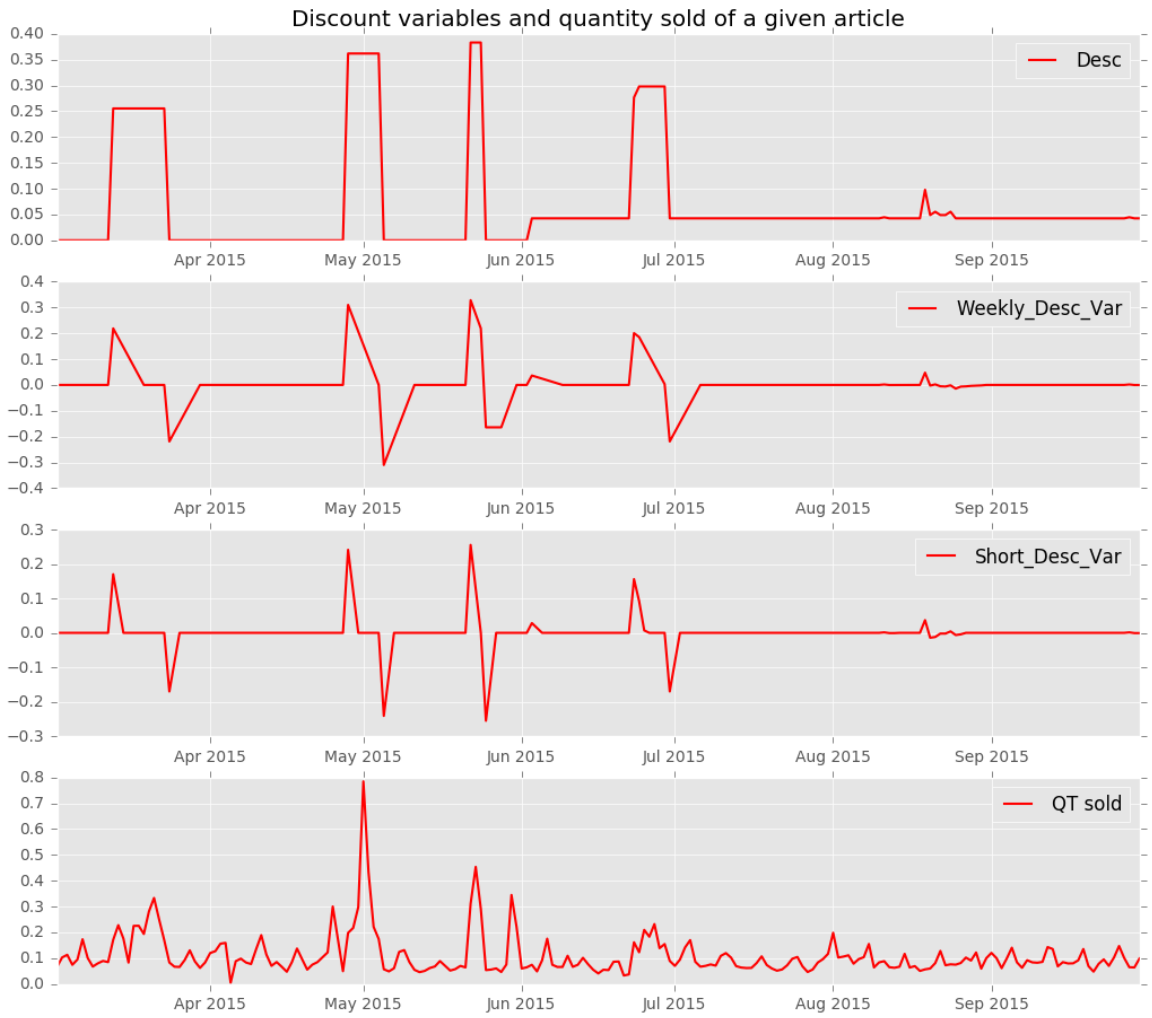


Figure 3 Article discount variables and respective amount sold

The influence of the discount strategy over the sales of an article is not restrained to the discount this article has. The discount of other articles within the store also have can have a considerable impact. When analysing consumer purchasing patterns, the relationships between articles can be classified as substitutability and complementarity. A relation of substitutability describes an inversely proportional impact in each other's sales, as the discount of a certain brand of yogurts negatively affect the sales of the similar yogurts of distinct brands. Whereas a complementary relationship describes a positive variation of one article once the price of the second article decreases, such is the case of pizza due and tomatoes or tortilla and taco sauce (Berman and Evans, 1989 andWalters et al.,1988). These relationships can be found at both intra-category level and inter-category level.

In practice we can hardly identify the nature of the relations between articles since often these can become complex and change according to the situation, for example, the discount in a fruit can influence the buying to take it instead of another fruit but at the same time can influence another buyer to take them both.

In the following work, due to the exponential increase in the number of variables when having into account combinatory influences of discounts, composite variables have been created. Although they imply loss of information, these aggregated indicators empower the model with relevant information concerning the discount strategy at both intra-category level and inter-category level.

Intra Category information

At the inter-Category level two variables have been developed. Both focus on the rate of the article promotion when comparing with the others in the same category: the quantile of the Article's price (Cat_Price_Quantile) and the quantile of the Article's discount (Cat_Desc_Quantile). Quantiles are used in probability to divide observations into ranges of equal probabilities) are here applied as a comparison of the quality of the promotion when compared with similar articles. Figure 4 shows the prices of articles inside a category for a given day, here we can observe the information given by the quantile: the article sold at 3.60€ has a quantile value of 0.1 which indicates that 10% of the articles of the same category were cheaper in that day while the article at the quantile 0.8 was more expensive than 80% of the articles in this group.

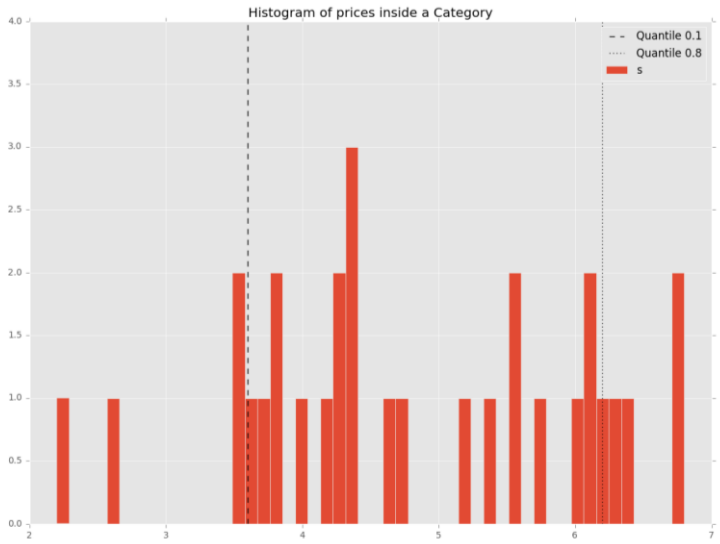


Figure 4 Histogram of prices of a category and visualization of quantiles

Inter Category information

Due to the sheer amount of information which can be used when exploring the impact of inter category discounts, aggregated variables were created. For each article, the correlation between its sales and the average discounts of all other subcategories was analysed and the most significant correspondences (the ones with the biggest correlation absolute value) were identified. The average discount of the chosen subcategories were then taken into account while producing the model.

3.2 Temporal data

In order to have into account information concerning the temporal effects, multiple temporal variables were taken into consideration.These Boolean variables help describing the time dimension of the sales and allow their impact to be taken into consideration by the model.

Figure 5 show the variables which allow the identification of weekdays, weekends and weeks of relevant holidays. Many others Boolean variables were taken into account as to identify and allow the model to distinguish between the days of the week, months and semesters

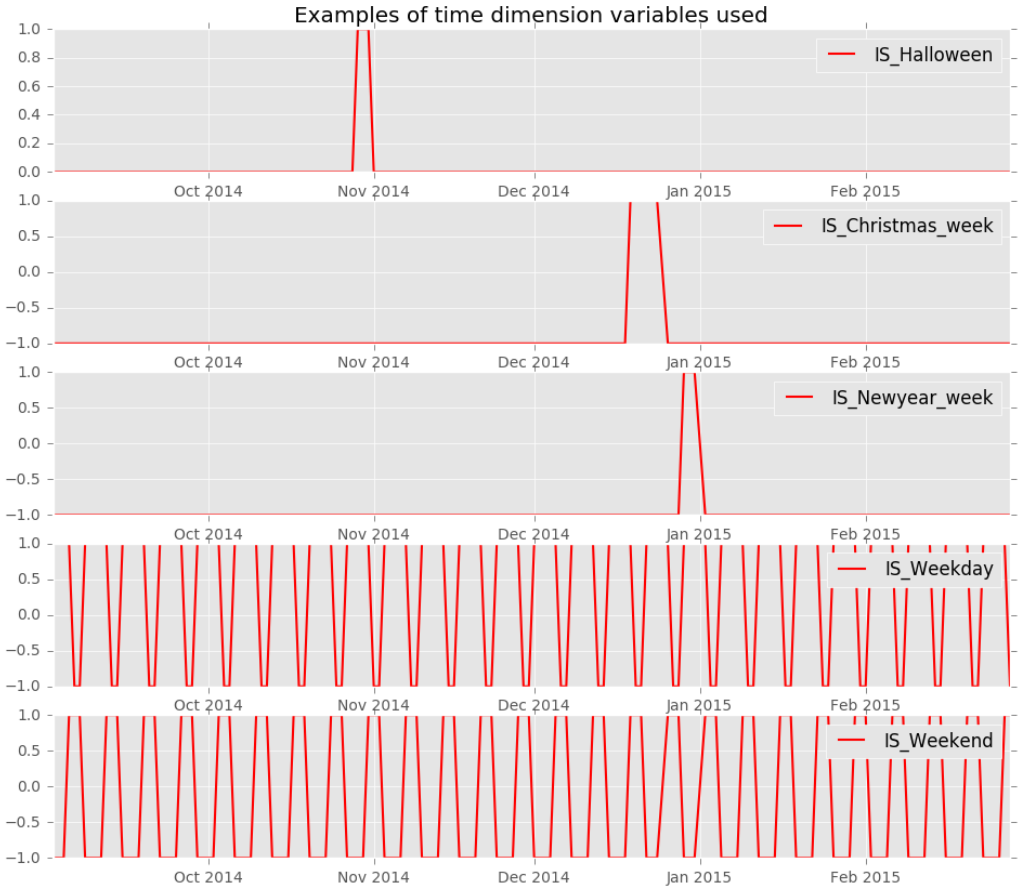


Figure 5 Examples of time dimension variables taken into account

3.3 Weather

Several weather attributes were considered and studies, from the intensity of the wind to the amount of rain and the temperature itself. These attributes have an influence in the consumption as they reflect cultural behaviours to the weather.

Although correct data is available during the training and testing of the models, in a real world implementation these attributes would be estimative in the shape of weather forecasts. The figures below present the variation over time of some of these variables as well as their histogram over an interval of two years.

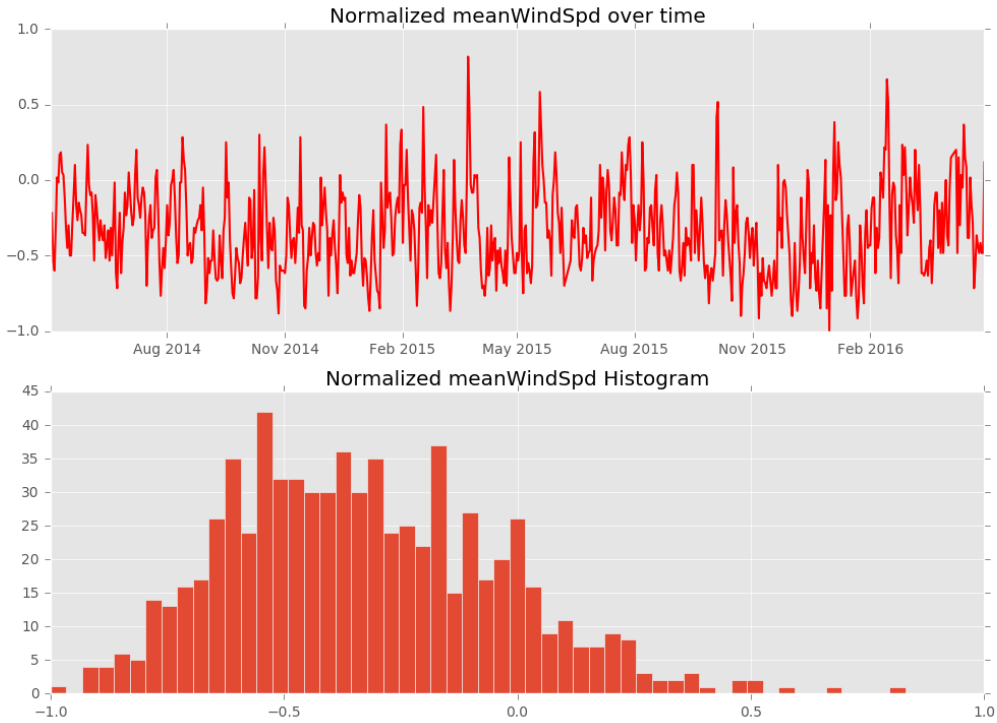


Figure 6 Variable representative of the daily strength of the wind

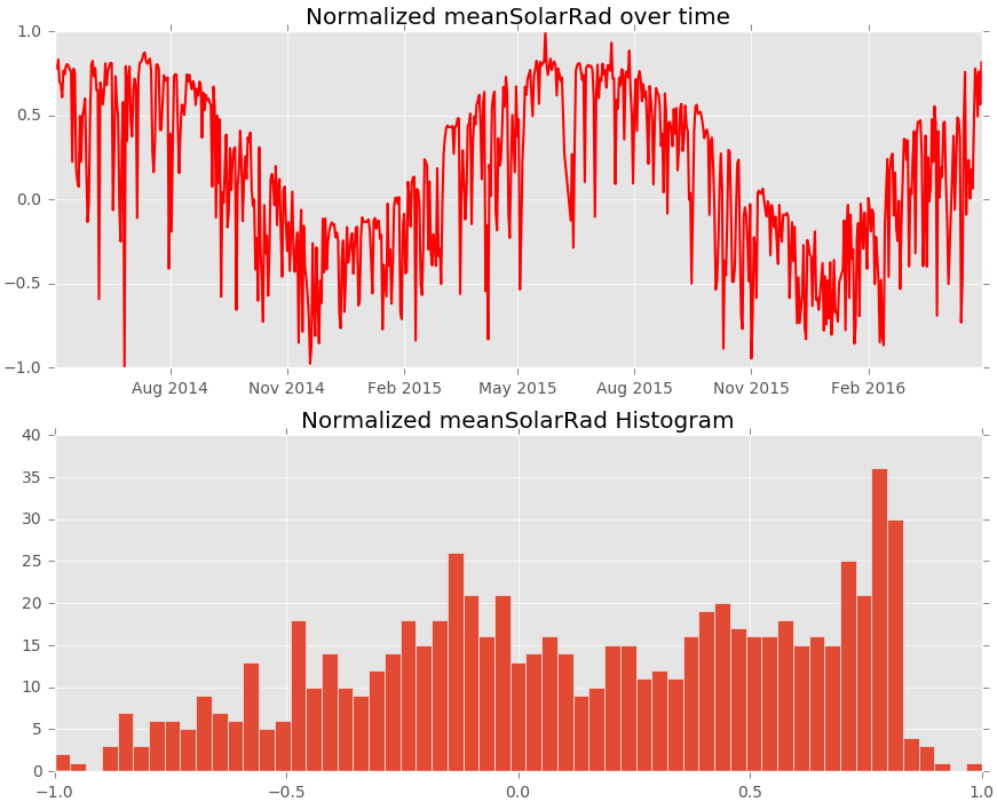


Figure 7 Variable representative of the daily strength of the sun's radiation

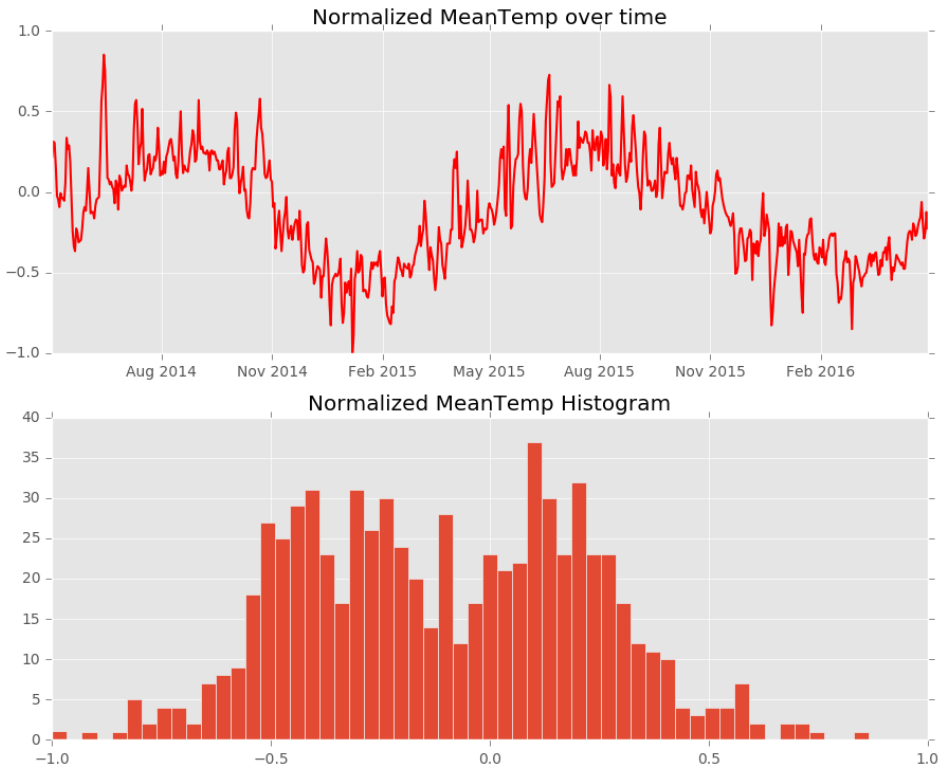


Figure 8 Variable representative of the daily mean temperature

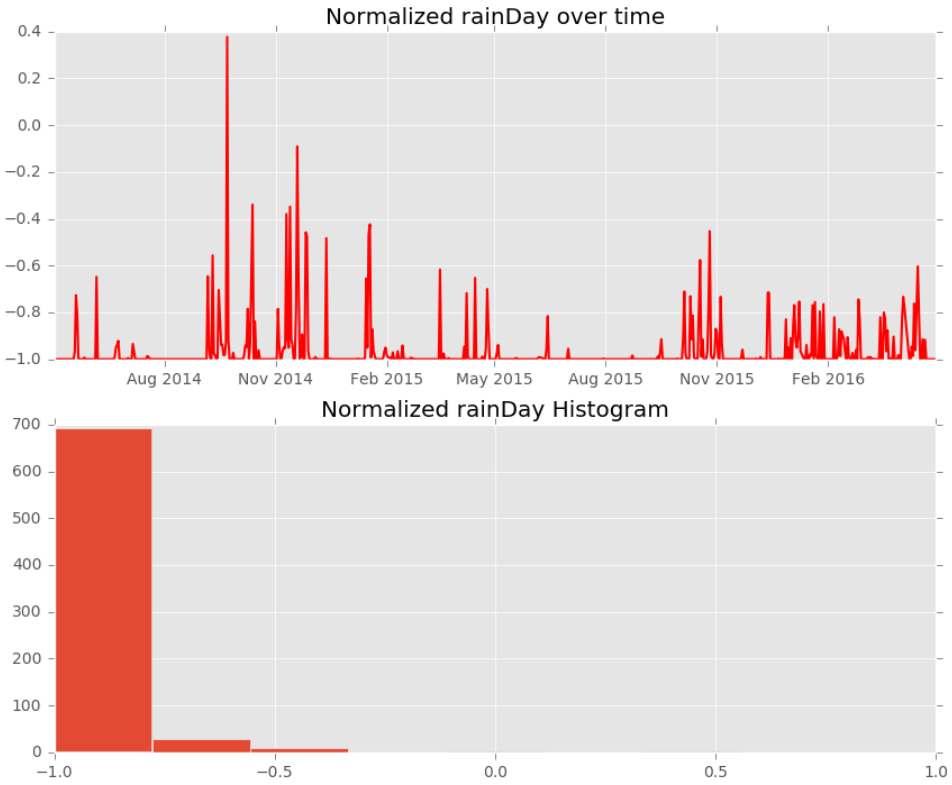


Figure 9 Variable representative of the daily amount of rain

3.4 Football matches

Football has become a sport of meaningful influence, it is closely followed by a great percentage of people who remotely watch these games. It isthen an important influencer in the behaviour of the client as it increases substantially sales in certain products as snacks and beer.

These games also can diminish the consumption of goods due the impact they have in the schedule of clients. In this project two attributes were considered based on these events

Game - the number of football games of national teams which occur per day.

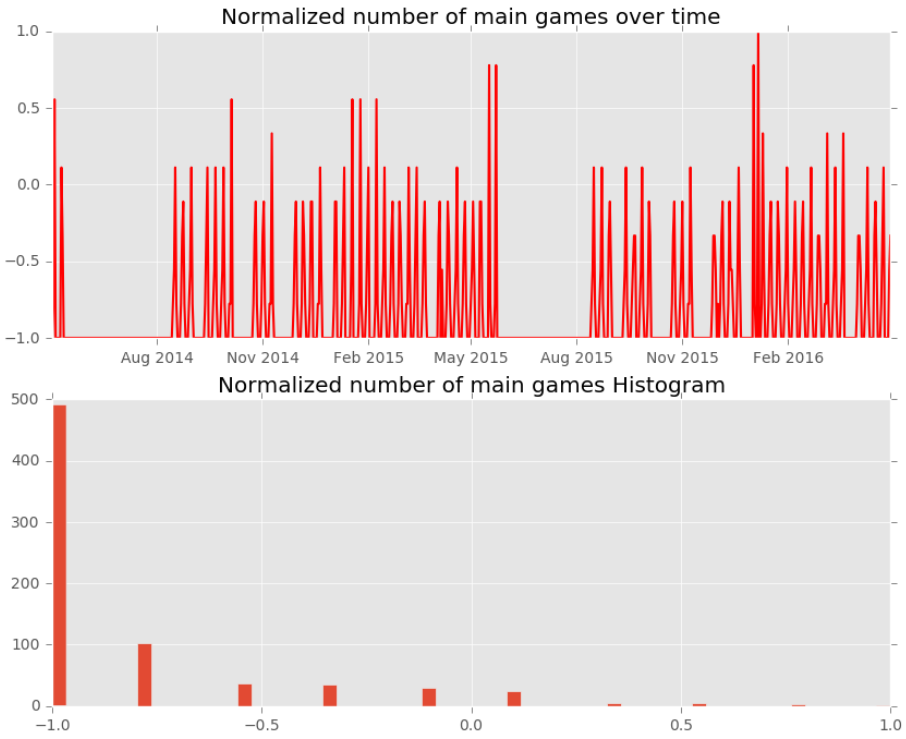


Figure 10 Variable representative of the number of football games per day

Game Main teams – this variable only takes three different values, it reflects the number of local teams playing at a given day.

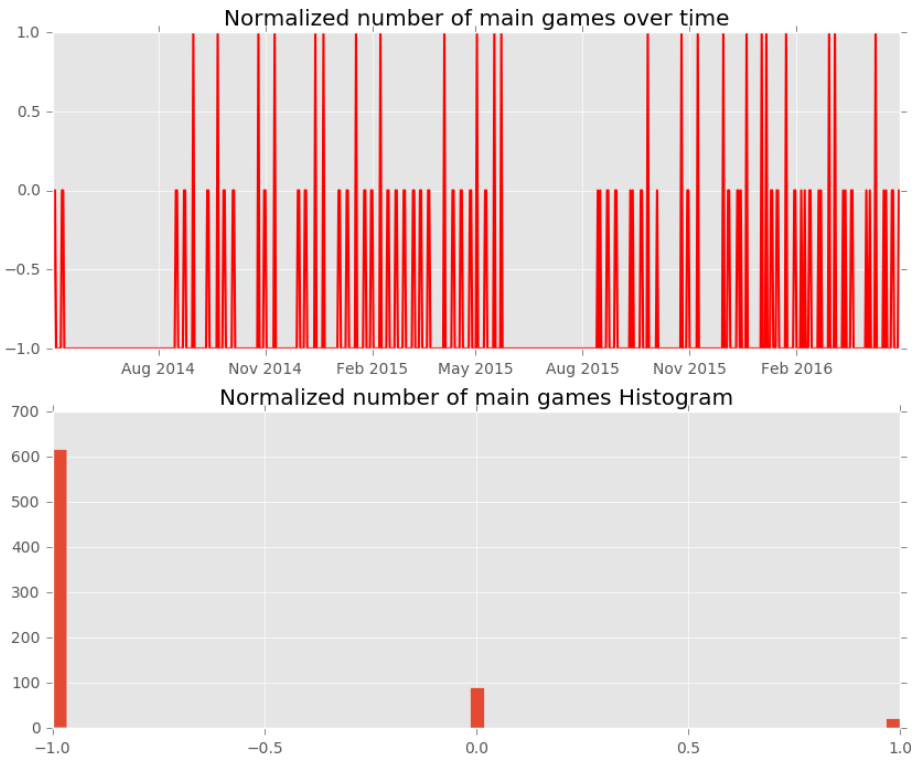


Figure 11 Variable representative of the number of football games of main teams per day

3.5 Model input data summary

An overview of the distinct sources of training data can be observed in Figure 12.

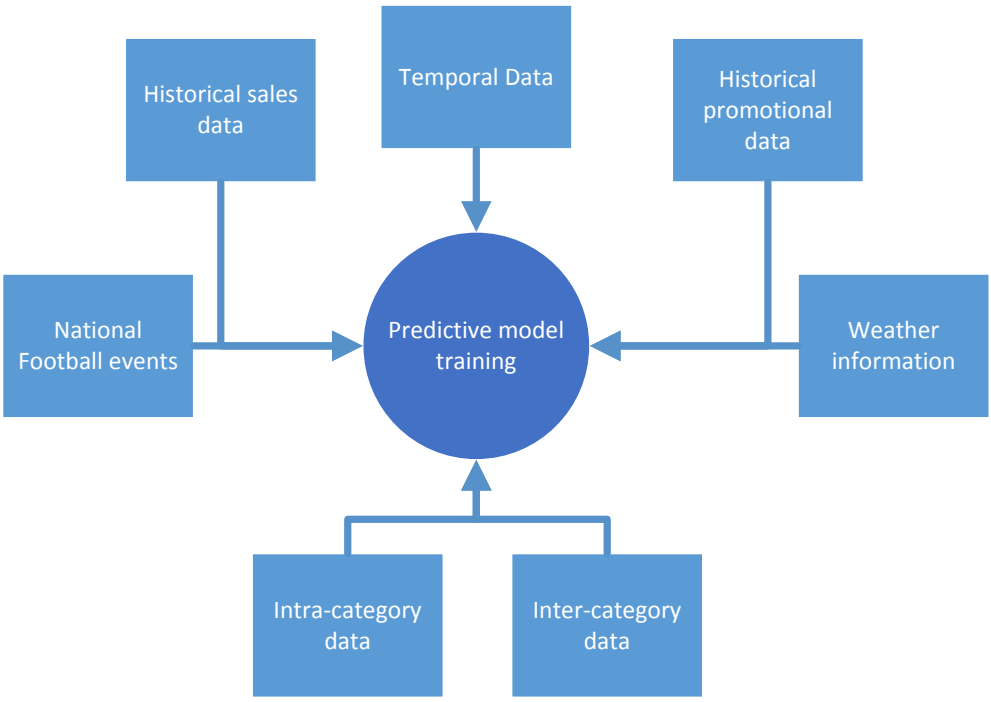


Figure 12 schema of input sources of data for training

In order to be processed by the models, Categorical variables (ordinal variables and nominal variables have been turned into Numeric variables by means of multiple discrete Boolean flags.

The list of variables is shown in Table 3.

For the evaluation of the performance of each algorithm per article, the training set corresponded to the temporal interval of 01.01.2012 until 11.10.2015 which roughly corresponds to 994 training data points (variant with daily data availability). Once trained, the models produced four weeks of forecast for daily sales (28 days). Nearly 200 articles were analysed.

Source	Input variable	Description	Behaviour
Historical sales data	QTY	Daily sales	Discrete variable
	QTY_PREV_YEAR	Sales of corresponding day of the month in previous year	Discrete variable
	QTY_PREV_YEAR_H	Sales of corresponding weekly day in previous year	Discrete variable
	QTY_PREV_MONTH_H	Sales of corresponding weekly day in previous month	Discrete variable
	STK	Theoretical Inventory amount	Discrete variable
Historical promotional data	DISC	Discount	Continuous variable (relative change)
	DISC_SEM_VAR	Weekly discount variation	Continuous variable (relative change)
	DISC_SHORT_VAR	Discount variation of the last 3 days	Continuous variable (relative change)
Temporal data	WEEKEND	Is weekend	Discrete variable (Boolean)
	YEAR	Year counter	Discrete Incremental variable (starts in 2012)
	MONTH	Month counter	Discrete Incremental variable (restarts yearly)
	FORTNIGHT	Fortnight counter	Discrete Incremental variable (restarts yearly)
	QUARTER	Quarter counter	Discrete Incremental variable (restarts yearly)
	QUADMONTH	Four months counter	Discrete Incremental variable (restarts yearly)
	SEMESTRE	Semester counter	Discrete Incremental variable (restarts yearly)
	HOLIDAY	Holiday	Discrete variable (Boolean)
	WORKDAY	Working day	Discrete variable (Boolean)
	SIN_WEEKEND	Sinusoidal function with maximum value on Saturdays	Continuous variable
	WEEK_HALLOWEEN	Halloween Week	Discrete variable (Boolean)
	WEEK_CHRISTMAS	Christmas Week	Discrete variable (Boolean)
	WEEK_NEWYEAR	New year's Week	Discrete variable (Boolean)
	MON	Monday	Discrete variable (Boolean)
	TUE	Tuesday	Discrete variable (Boolean)
	WED	Wednesday	Discrete variable (Boolean)
	THU	Thursday	Discrete variable (Boolean)
	FRI	Friday	Discrete variable (Boolean)
	SAT	Saturday	Discrete variable (Boolean)
	SUN	Sunday	Discrete variable (Boolean)
	JAN	January	Discrete variable (Boolean)
	FEB	February	Discrete variable (Boolean)
	MAR	March	Discrete variable (Boolean)
	APR	April	Discrete variable (Boolean)
	MAY	May	Discrete variable (Boolean)
	JUN	June	Discrete variable (Boolean)
	JUL	July	Discrete variable (Boolean)

Source	Input variable	Description	Behaviour
	AUG	August	Discrete variable (Boolean)
	SEP	September	Discrete variable (Boolean)
	OCT	October	Discrete variable (Boolean)
	NOV	November	Discrete variable (Boolean)
	DEC	December	Discrete variable (Boolean)
Weather information	MAXTEMP	Maximum Temperature	Continuous variable
	MINTEMP	Minimum Temperature	Continuous variable
	MEANTEMP	Average Temperature	Continuous variable
	MEANTEMPAP	Average apparent temperature	Continuous variable
	MEANHR	Average relative humidity	Continuous variable
	MAXWINDGUST	Maximum wind velocity	Continuous variable
	MEANWINDSPD	Average wind velocity	Continuous variable
	MEANPRES	Atmospheric pressure	Continuous variable
	MAXSOLARRAD	Maximum solar radiation	Continuous variable
	MEANSOLARRAD	Average Solar radiation	Continuous variable
	RAINDAY	Rainy day	Discrete variable (Boolean)
National Football events	GAME	Number of football games	Discrete variable
	GAME_MAIN	Number of football games with bigger impact	Discrete variable
Intra-category data	CAT_PVP_QUANT	Price's quantile among the Category	Continuous variable
	CAT_DISC_QUANT	Discount's quantile value among the Category	Continuous variable
Inter-category data	DISC_6648	The average discount of the 6 most correlated subcategories found	Continuous variable
	DISC_9462		Continuous variable
	DISC_6645		Continuous variable
	DISC_6646		Continuous variable
	DISC_6947		Continuous variable

Table 3 List of input variables

4. MODELS

4.1 Multiple Linear Regression

Years of development of the field allowed statisticians to develop more complex and advanced methods for numeric target prediction. Nevertheless, linear regression analysis continues to be the most widely used of all statistics techniques, mainly due to its simplicity and computational performance.

Linear regression attempts to establish a linear relationship between one independent variable $X$and a continuous outcome variable Y.Thus, fitting a line through a scatter plot (as seen inFigure 13).

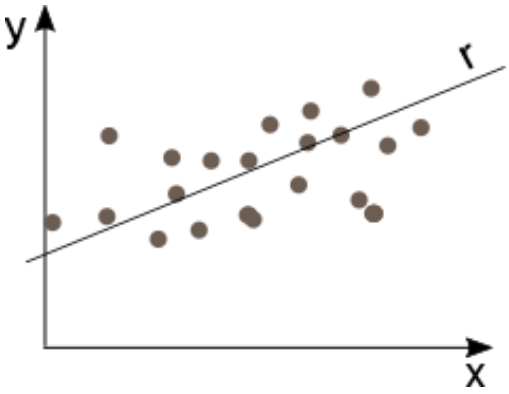


Figure 13 Example of a linear regression

On the other hand, multiple linear regression is a more complex and widely used form of linear regression. MLR can be described as fitting a line through a multi-dimensional cloud of data points

It is used to explain the relation between a continuous dependent variable and multiple independent variables. Here, the objective variable is the result of a straight-line function of each of the $\mathrm{X_i}$ variables, while holding the remaining fixed, their contributions are additive.

MLR notation

Given a data set of p data points $\{y_t, x_{t1}, x_{t2}, \dots, x_{tn}\}^{p}{}_{t=1}$ Let Y denote the dependent outcome whose values we intend to predict and let $X_1, X_2, \dots, X_n$ denote the n independent variables from with we wish to base our prediction. The equation of the general form of the multiple linear regression is given by the following expression:

$$Y_t = b_0 + b_1 X_{t1} + b_2 X_{t2} + b_3 X_{t3} + \cdots + b_n X_{tn} + \varepsilon_i \qquad [6]$$

Where:

t represents the number of the data point,

ε_i is called the error term or noise,

$b_1, b_2, \dots, b_n$ represent the individual slopes coefficients of the X_i variables, that is, the change in the predicted value $\hat{Y}$ per unit change in X_i, everything else remaining constant.

b_0 defines an additional constant intercept coefficient, represents the value the prediction would take if all the independent variables X_i were to be zero.

We can therefore pack all response values or the given observation into a t-dimensional vector – Response vector.

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \dots \\ \dots \\ \dots \\ Y_t \end{pmatrix} \qquad [7]$$

In order to represent the MLP in a matrix form we need to remove the intercept. Instead of doing so by centring the data, we can consider an extra column of 1's when packing the predictors into the matrix X – Design matrix ($\mathrm{t} \times \mathrm{n}+1$).

Hence,

$$\mathrm{X} = \begin{pmatrix} 1 & \mathrm{X}_{11} & \mathrm{X}_{12} & \cdots & \mathrm{X}_{1\mathrm{n}} \\ 1 & \mathrm{X}_{21} & \mathrm{X}_{22} & \ldots & \mathrm{X}_{2\mathrm{n}} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \mathrm{X}_{\mathrm{t}1} & \mathrm{X}_{\mathrm{t}2} & \cdots & \mathrm{X}_{\mathrm{tn}} \end{pmatrix} \quad [8]$$

The intercepts are packed into a $\mathrm{n}+1$ dimensional vector, the slope vector.

$$\mathrm{b} = \begin{pmatrix} \mathrm{b}_0 \\ \mathrm{b}_1 \\ \mathrm{b}_2 \\ \ldots \\ \ldots \\ \ldots \\ \mathrm{b}_\mathrm{n} \end{pmatrix} \quad [9]$$

Finally, the errors are packed into the error vector, a t-dimensional vector.

$$\epsilon = \begin{pmatrix} \epsilon_0 \\ \epsilon_1 \\ \epsilon_2 \\ \ldots \\ \ldots \\ \ldots \\ \epsilon_\mathrm{t} \end{pmatrix} \quad [10]$$

We can then compact the generalized expression into the matrix form:

$$\mathrm{Y} = \mathrm{Xb} + \epsilon \quad [11]$$

The coefficients and intercept contained in the slope vector can then be estimated by least squares. Here intend to obtain the unique values which minimize the sum of the squared errors,

$$e^T e = (Y - Xb)^T (Y - Xb) = Y^T Y - 2Y^T Xb + bX^T Xb \quad [12]$$

We can minimize this function by setting the partial derivative, with respect to the search variables $(b_0, b_1, b_2, \dots, b_n)$ to 0.

This will give us $n + 1$ equations with $n + 1$ unkowns where n is the number of variables in the model.

The solution in compact form is given by

$$b = \left(X^T X\right)^{-1} X^T Y \quad [13]$$

Where $(X^T X)^{-1}$ is a symmetric $n + 1 \times n + 1$ matrix and $X^T Y$ is a $n + 1$ dimensional vector.

We can therefore describe the fitted values

$$\hat{Y} = X\hat{b} \quad [14]$$

As

$$\hat{Y} = X\left(X^T X\right)^{-1} X^T Y \quad [15]$$

MLR considerations

Although this method assumes a strong linear relationship between the inputs and the dependent variable, the simplicity of the acquisition of the parameters b_i by computations in the matrix form allow great performances.

Multiple linear regression, by considering multiple input variables must have into account degenerations to the model caused by poor judgement in the selection of these variables.

Adding more independent variables to a multiple regression procedure does not imply that the model will perform better, in fact it can make things worse. This phenomenon is called overfitting.

In addition, the consideration of a bigger number of independent variables will also create more relationships between them. Not only the predictors will have relations with the outcome but they will also be correlated between themselves – multicollinearity.

The ideal is for all independent variables to be correlated with the dependent variable but not among each other.

Due to these phenomenon, which contributes to a decay in the MLR precision, a prior variable-selection step must be executed as to allow the model to assess the true relationship and descriptive power of each used variable.

Some independent variables, or sets of independent variables, are better at predicting the outcome than others, some of which can have no contribution at all.

We must decide which variables to include in our model and which ones to exclude.

Many subset selection methods have been studied for this purpose.

On this work, due to the large scale combinatorial problem, the genetic algorithms(GA) (Melab et all. 2006) have been used to identify the general subset of variables which will be considered on the individual MLR models.

GA Input variable selection

When developing a predictive model, the search of an optimal subset of relevant input variables is crucial.

Specially in problems with high volume of data, the collection of the ideal set can both time-consuming and computationally expensive. Numerous predictive variables can however be involved, not being possible to omit many of these without a significant loss of information.

The use of genetic algorithms is here explored as to automatically select the most relevant input variables set.

Inspired from Charles Darwin's theory of evolution, genetic algorithms are used to solve elaborate optimisation problems (Melanie, 1995, Holland, 1975 and Goldberg, 1989) by finding the best solution in a very large space.

In a Genetic algorithm, the solutions are viewed as individuals, represented by vectors and characterized by their fitness, a value which describes their performance. These solutions compete and evolve over time, as to find an optimal solution.

Initially, the N individuals are randomly generated in the initial Population P. Their fitness is computed while an iterative selection (with reposition) phase executes until the intermediary population$P$' (which is initialized empty) has N individuals as well. The selection phase will give preference to the individuals with the highest fitness.

Once the intermediary population $P^{'}$ contains N individuals, the variation phase takes place (Figure 14). The individual will be selected in pairs and will be crossover with probability R_c (crossover rate). Once the crossover occurs, the vectors corresponding to the new individuals will detach and switch content.

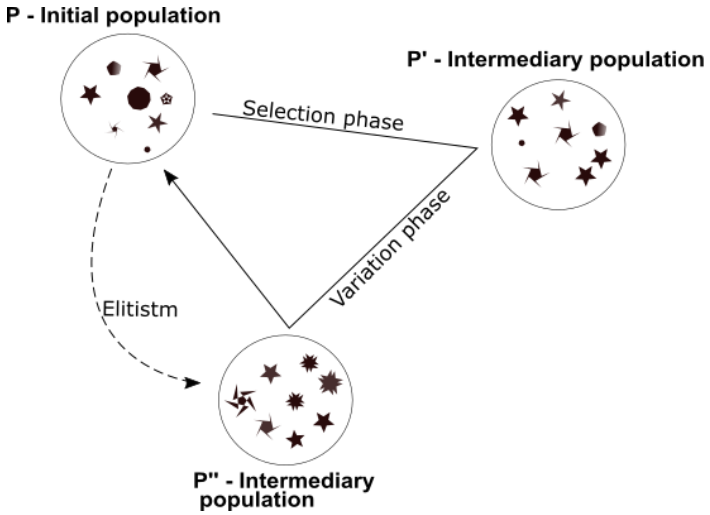


Figure 14 Genetic algorithm schema

Once the crossover phase ends, the mutation phase begins. Here, each position of each resulting vector will then mutate with a probability R_m (mutation rate).
The resulting duo will then be inserted into the new population P''.
Once P'' contains N individuals, it replaces the initial population P and the process iterates. One complete iteration is called generation. GA will stop either by a condition which infers the progressive improvement of the maximum fitness detected or by a pre specified number of generations. In the end, the solution will be the individual with the maximum fitness value. As to secure that the maximum fitness of a generation is always the same or bigger than the maximum fitness or the former, it is common to copy, without modification, the best individual of P to P'', this operator is called elitism.

Although widely used, these algorithms do not visit all of the solution space and therefore, do not guarantee the global optimum. Nevertheless, the use of wide populations and the use of the mutation phase, which introduces noise to the procedure, helps in avoiding local maximums.

The GA implemented in this project is defined as a search algorithm to find the best combination of input variables to be considered in the MLR. Here, an individual represents a set, defined as a vector V where each position represents an input variable with a binary encoding (Figure 15). Each position express if a variable was used (represented by "1") or not (represented by "0").

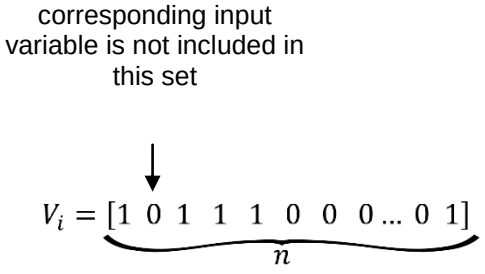


Figure 15 Example of an individual / solution vector

The fitness function which assesses the quality of a set, is using a subset of 10 articles. It is given by the inverse of the average absolute error of the MLR using the subset of articles:

$$Fitness(V) = -\frac{\sum_{i=1}^{10} \left| Error_{MLR_V}(Article_i) \right|}{10} \quad [16]$$

Although some preconceived notion about which of the predictors matter the most can be used, this method generally determines which predictors have the strongest impact on the outcome. Highly Correlated variables, which grouped do not bring a surplus of knowledge over their individual representation, can disappear here, without damaging the precision of the model.

4.2 ARIMA

Sometimes mentioned as Box-Jenkins after the original authors (Box and Jenkins, 1974), ARIMA stands for Autoregressive Integrated Moving Average and is a forecasting technique which belongs to the field of time series analysis.

Time series are set of observations on a particular value taken at different times (eg. Stock prices) over regular time intervals such as daily,monthly, weekly, quarterly or annually. Their analysis focus on the extrapolation of knowledge of patterns in the data.

Used in field such as statistics econometrics, mathematic finance, weather forecasting, earthquake prediction, ARIMA works as an alternative for multivariate analysis (such as the multivariate regression) once independent variables (x) are not available or are subject to restrictions.

$$y = a + bx \quad [17]$$

This technique uses solely the past values of the variable of interest (Y) to extrapolate to the future.Although it performs well when patterns are consistent over time with few outliers, it underperforms on highly volatile data.

One of the constrains of this model is its inability to tackle nonstationary.Series which present growing patterns must pass thought an intermediate step. This issue is commonly treated through one or two differentiations of the data, sequentially subtracting the observation in a period from the previous one.
As the name suggests, ARIMA incorporates autoregressive (AR) and moving average (MA) parameters to describe the behaviour of the, now stationary, time series.

An autoregressive (or AR) model is a model in which the variable at a given time, y_t, is a function of its past values plus a random error, thus,

$$y_t = f(y_{t-1}, y_{t-2}, y_{t-3}, y_{t-4}, \dots, \varepsilon_t) \quad [18]$$

Where ε_t is the error term. A common representation of an AR is given by:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \beta_3 y_{t-3} + \dots + \beta_n y_{t-n} + \varepsilon_t \quad [19]$$

The representation above would be described as an AR(p) where p refers to the number of past values or lag considered, usually not above 4. An AR model of order one (AR(1)) is represented bellow:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \varepsilon_t \quad [20]$$

The second component used in ARIMA tackles the moving average parameters.

A forecasting moving average model is one where y_t depends only on immediate past value error terms, i.e.

$$y_t = f(\varepsilon_t, \varepsilon_{t-1}, \varepsilon_{t-2}, \varepsilon_{t-3}, \dots) \quad [21]$$

Although the models appears similar to AR, instead of using past values they use only the error terms for forecasting. A common representation of a moving average model dependent on q past values (MA(q)) is be expressed by:

$$y_t = \phi_0 + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \phi_3 \varepsilon_{t-3} + \dots + \phi_n \varepsilon_{t-n} + \varepsilon_t \quad [22]$$

Following the example of AR models, the MA models can be extender to higher orders as to cover different moving average lengths.

The combination of the parameters of both models called AutoRegressive moving average model or ARMA(p, q), where p and q are the orders of the AR and MA respectively. Since the differenctiation induced to the time series must now be reverted as to provide the real forecast, an Integration step is included and such, the resulting model is stated as ARIMA(p, d, q) where d refers to the number of differentiating operators.

We still need to introduce the right specification to the ARIMA model, the correct values to the parameters$p$, d and q must be determined. Although there is substantial work focused on the identification of these parameters, it usually involve a component of graphical analysis of the autocorrelation of lags of the time series.

In the current thesis, given the low number of data points for a training (less than three years of daily data), the parameters have been computed by brute force given a limited interval of accepted values. The best ARIMA configuration (p, d, q) for the training data or each article is the representative of the ARIMA model and used in the validation set.

4.3 Neural Network

Artificial neural networks (ANN) are widely known for their use inmodelling complex problems, image recognition, speech recognition and deep learning as they provide a general method for learning from examples.

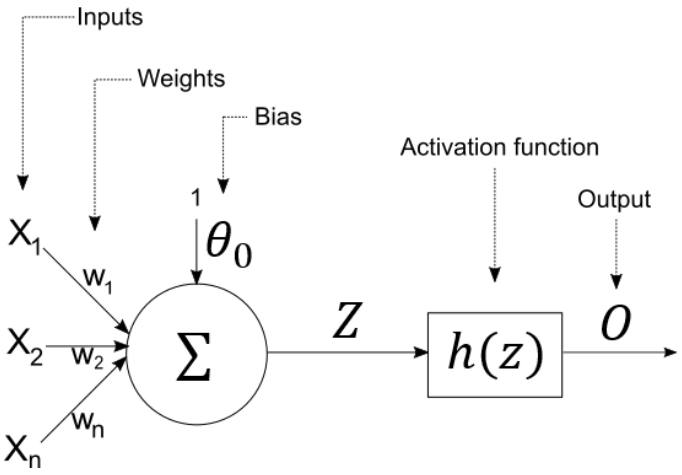
In theory, any class of statistical model can be designed through a neural network as they, with their adaptive weights and scalable complexity, can approximate complex non-linear functions.

ANN are easily adapted to solve regression problems. Suited therefore for problems where a more traditional regression model, such as the ones described in the last chapter, cannot find a solution.

A Neural network regression is a supervised learning method,which uses the known history and respective output to iteratively improve based on the individual error of each estimation. After the ANN being tuned and the non-linear relations of the system are approximated, it can then be used for estimation, predicting outcomesof new input examples.

The ANN follow a computational representation of how the brain is thought to work. In the brain, dendrites receive ions through synapses and the electrical signal is propagated to the cell body and if the neuron is sufficiently stimulated, it gets excited and will pass on the signal, stimulating its neighbours. This “activation” of a neuron is dependent on the strength of the signal brought by the connections from other neurons and its own sensitiveness to the signal. The Neural networks follow the same paradigm.

As seen in theFigure 16, a node in this method is the equivalent to a dendrite. It receives inputs X_i from previous layers and the weighted sum is then fed to an activation function which will define the output of the node, the node’s activation value.



$$Z = \sum_{i=1}^{n} w_i x_i + \theta_0$$

$$O = h(z)$$

Figure 16 The ANN node

As seen in the figure, every node (excluding the ones in the input layer) contain an extra weighted input connection from a unitary static source, this “bias”, generally expressed by θ, is used to express the tendency of the node.

A traditional neural network can be seen as a one directional graph containingset of nodes displayed among layers (as seen inFigure 17), connected by weighted edges. It contains one Input layer, which introduces the predictors to the network and one output layer, which outputs the estimated value(s) of the network. Theremaining layers that connect them both are called hidden layers.These are needed to introduce non-linearity into the network.

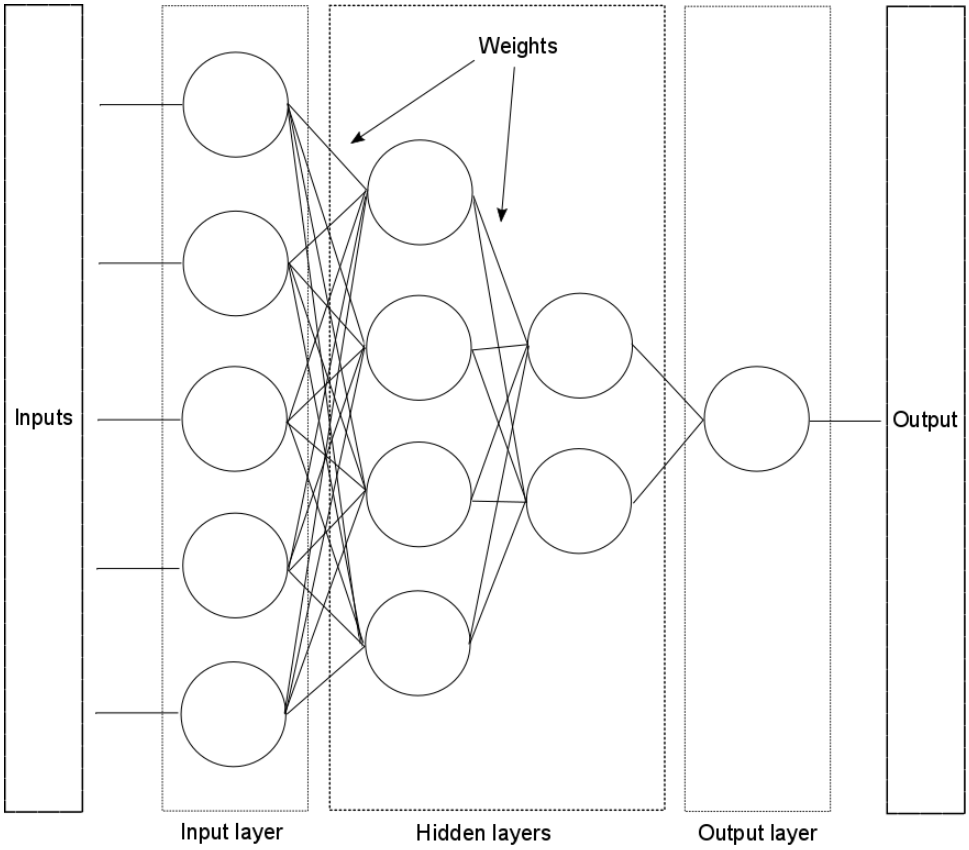


Figure 17 Traditional network schema

Designing a network requires setting many parameters as the number of hidden layers, nodes per each of the layers, considered activation functions, initial weights and learning rates (which will define how much the weights can be tuned in each iteration). Although the choice of these parameters can be critical, there is no fool proof recipe for determining them. Nevertheless, there are works which study heuristic process to assist this decision (Lecun et al.,1998).

Training Phase

Given a specified ANN, the training step follows a sequential concept; each training example will be processed individually by the network.

The result of the training process is the determination of the best set of weights for the specified layer. These represent the strength of the relationships between the predictors.

Unless some preconceived notion about the predictors relations can be considered, each layer is initialized with random weights.

The general idea of the training process is, similarly to the brain paradigm, to tune the weighs,reinforcing those that help achieving good and weakening the ones that lead to unwanted results.

Each example $\{y_i, x_{i1}, x_{i2}, \dots, x_{in}\}$is presented individually to the network and its consideration has two steps. Initially, a forward propagation of the signal occurs. The predictor values of the example $\{x_{i1}, x_{i2}, \dots, x_{in}\}$are introduced to the input layer and the activation values of each layer are sequentially computed. The value $\hat{y}_i$returned by the output layer, is the current estimative of the model for the given input data. $\hat{y}_i$ is then compared to the known output of the example y_i. The estimative error $\varepsilon_i = \hat{y}_i - y_i$ of the estimative is then computed and the backpropagation, second step of the training phase, starts.

The fundamental idea in the training is that the error obtained in each node output is a function of the weights, which generated the activation value of that node. Hence our focus in iteratively adjusting the weights in order to minimize the error while having into consideration their individual responsibility for the measured error. This can be done by computing the partial derivative of the result activation function relatively to the individual weights. This gradient, seen as the slope inFigure 18, characterized by a steepness and direction, gives an indication of how to move into the direction of a better weight.

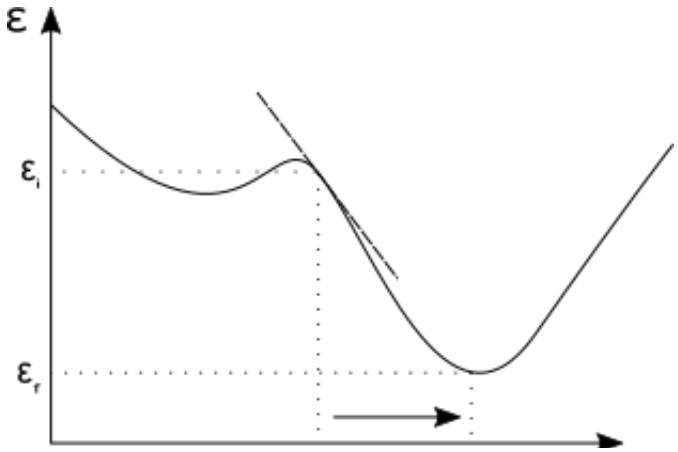


Figure 18 Error as a function of an individual weight

Backpropagation uses the gradient descent method in order to obtain the variation a weight must take in each step. According to this method, we define the error as the squared error function E:

$$E = \frac{1}{2}(\hat{y}_i - y_i)^2 \quad [23]$$

and this error is derived in respect to each of the $j$weights

$$\frac{\partial E}{\partial w_{ij}} \quad [24]$$

The differentiable logistic/sigmoid function was used as activation function $h(x)$:

$$h(x) = \frac{1}{1 + e^{-x}} \quad [25]$$

Hence, after the differentiation, the error of a given a node j(excluding the input layer), considers the weighed sum of the errors of the next layer $k$and is given by:

$$\delta_j = O_j(1 - O_j)\sum_k E_k w_{jk} \quad [26]$$

where O_j is the given output of the node j.This error is then used to update the weights from the $i$inputs of the node j as:

$$\Delta w_{ij} = \alpha \delta_j O_i \quad [27]$$

And,

$$w_{ij} = w_{ij} + \Delta w_{ij} \quad [28]$$

The defined learning rate $\alpha$is set in order tocontrol the amount of change a weight can suffer per iteration. It controls the speed of convergence of the system and protects the network from the influence of outliers but, if too big, can compromise the convergence of the network.

Considerations

Theoretically, neural networks can map any input-output function no matter how complex. Incrementing the number of hidden layers increments the complexity of models which can be fit by the network. Nevertheless, Neural networks can be computationally expensive, mainly due to the extensive number of weights it may contain. Although in many cases neural networks produce better results than other algorithms, obtaining such results may involve fair amount of iterations over these parameters. However, since the initialization is random and the algorithm persistently evolves as to decrease the error, the weights can be stuck into "local minima" if the error landscape (Figure 18) contains local hills. This problem can be tackled by multiple initializations and trainings of the algorithm or by using a momentum term η which allows a fraction of the previous update to the current:

$$\Delta w(t)_{ij} = \alpha \delta_j O_i + \eta \Delta w(t-1)_{ij} \quad [29]$$

On practice, this allows the attenuation of oscillations in the gradient descent, similarly to the learning rate, it can improve the speed convergence but it can also, if settled to high, make the system unstable.

Although their training process is understood, unlike models such as decision trees and regressions, the knowledge contained inside an ANN cannot be easily perceived. Thus, they are commonly mentioned as black boxes.

5. RESULTS

5.1 Visualization tools

For the thorough analysis of each case in order to improve the perception of the data problems and model performance, visual tools have been developed in the form of a diagnose dashboard which can be loaded upon the run of each model. An example of the produced dashboard for the result of a trained MLR model can be seen in Figure 19. All these graphs are equipped with zoom function.

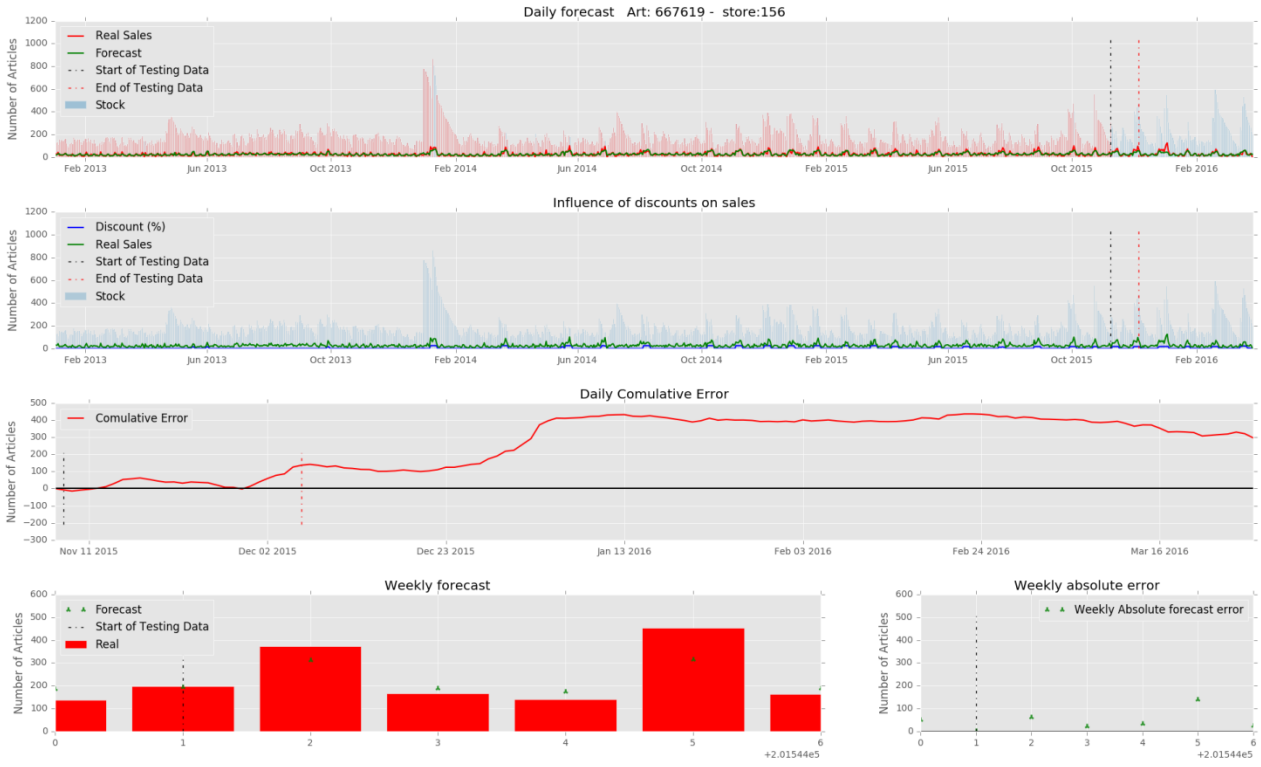


Figure 19 Model result dashboard

This dashboard is divided in 5 components which have demonstrated to be useful for a visual perception of the models performance.

The first component displays daily data. Here, the interval used for evaluation is delimited by dashed lines. The Inventory levels are displayed in bar plots, displayed in red in the case they belong to data points used for training. The real sales and the predicted value are displayed in a red and green bar respectively. By analysing the Figure 20 we can observe the behaviour of the sales prediction.Simultaneously, we can visualize small gaps in the data (as in the day 24 of October) and we can observe that the true value of sales of one of the days in the evaluation set (16 of November) was most likely limited due the limited inventory level.

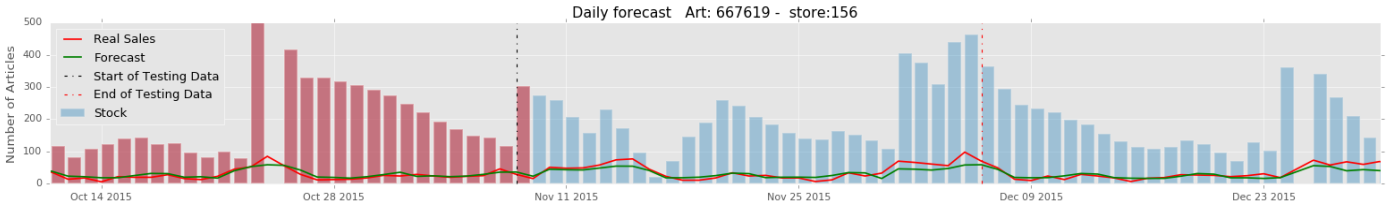


Figure 20 Model result dashboard – difference between forecast and the real sales

The solely objective of the second graph is discount analysis. Using this graph, we can justify previously identified abnormal behaviours. By using a close up displayed in the Figure 21 we quickly observe that the article is often subject to promotions of 25%, demonstrating consistent behaviours on sales.

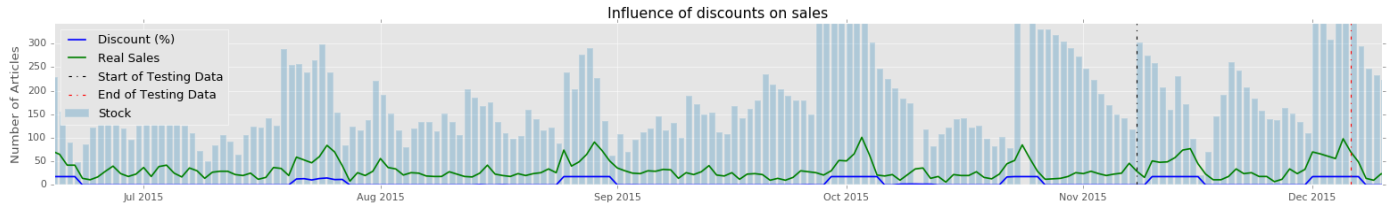


Figure 21 Model result dashboard – Influence of discounts on sales

The third graph displays the daily cumulative error, hence, it shows how the stock would accumulate if we strictly followed this prediction and provided daily inventory. It can be interesting to compare these levels with the levels of daily stock of the previous graphs. This graph (Figure 22) displays easily trends in stockpiling or depletion of the products.

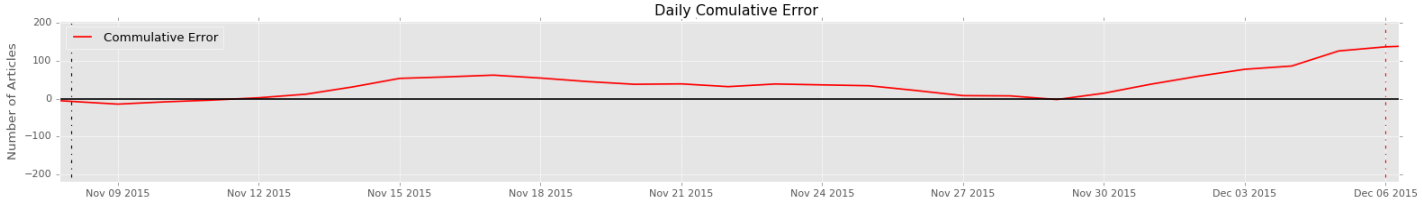


Figure 22 Model result dashboard – Daily cumulative error

The two last graphs (Figure 23) provide an overview of the weekly error. They present a comparison of the weekly forecast against the real amount. Here, the measure of performance of the project is displayed: the weekly absolute error, shown in the last figure and explained in the next chapter.

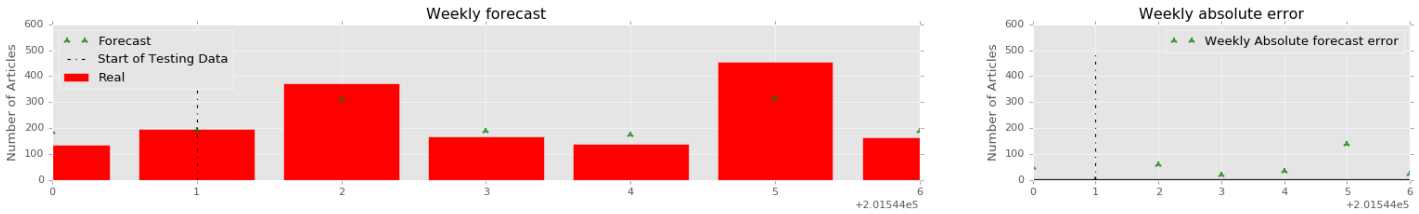


Figure 23 Model result dashboard – weekly analysis

5.2 Errors considered

After creating the models using data of three years (from 2012.01.01 to 2015.10.11), the precision on the forecasts was tested against the data of four weeks (not presented to the models during the training step). By comparison with the available real data, the precisions, strengths, and weaknesses of each one of the models were analysed.

The test data correspond to a subset of the most relevant products from five distinct article areas. As to limit the error impact caused by real data interference, products with one of the following properties were not analysed:

1) Products which had more than 20% entries (6 days) without sales in the test data
2) Products whose sales in the test data was limited by their low stock in more than 10% of used entries (3 days).

These conditions were set in order to maximize the validity of the results and control outliers caused by absence of stock or shelf availability.

The selection of the considered error plays a significant role in benchmarking the forecast quality. In order to facilitate the interpretation of the chosen error, two of the most popular measures have been used: the MPE (Mean Percentage Error) and MAPE (Mean Percentage Error). Where,

$$\epsilon_{MPE} = \frac{100\%}{n} * \sum_{i=1}^{n} \frac{\mathrm{Y_i} - \widehat{\mathrm{Y}}_i}{Y_i} \quad [30]$$

$$\epsilon_{MAPE} = \frac{100\%}{n} * \sum_{i=1}^{n} \left| \frac{\mathrm{Y_i} - \widehat{\mathrm{Y}}_i}{Y_i} \right| \quad [31]$$

A shortcoming of this choice is the sensibility these errors have towards low-volume data. To minimize this effect and simultaneously provide a more easily integrated solution, errors will be considered and analysed in the weekly basis ϵ_{week} . The daily estimates$\widehat{\mathrm{Y}}$ computed by every model will then be aggregated.Hence, we have:

$$\epsilon_{MPE_{week}} = \frac{100\%}{n} * \sum_{i=1}^{n} \frac{\mathrm{Y_{week\ i}} - \widehat{\mathrm{Y}}_{\mathrm{week\ i}}}{Y_{\mathrm{week\ i}}} \quad [32]$$

$$\epsilon_{MAPE_{\mathrm{week\ i}}} = \frac{100\%}{n} * \sum_{i=1}^{n} \left| \frac{\mathrm{Y_{week\ i}} - \widehat{\mathrm{Y}}_{\mathrm{week\ i}}}{Y_{\mathrm{week\ i}}} \right| \quad [33]$$

5.3 Global MPE and MAPE

By using the forecasted results for a period of four weeks and comparing it to the real data, we obtained the MPE and MAPE of each model's prediction by article. We can then map them into a histogram tovisualize the performance of each solution:

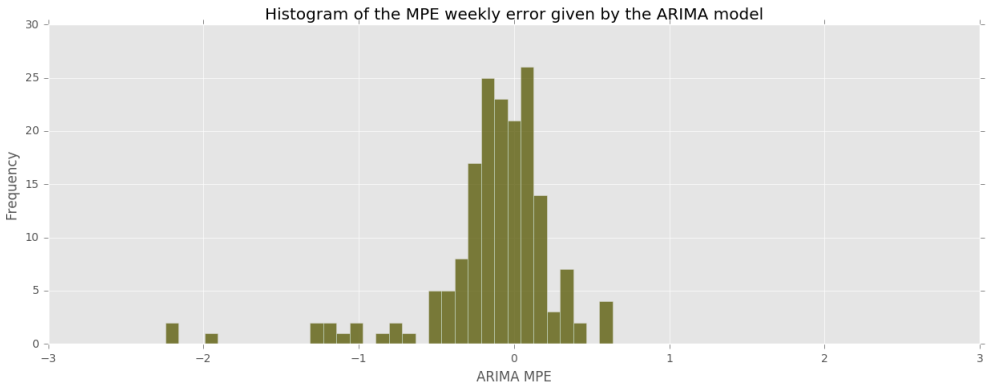


Figure 24 Histogram of the ARIMA weekly MPE

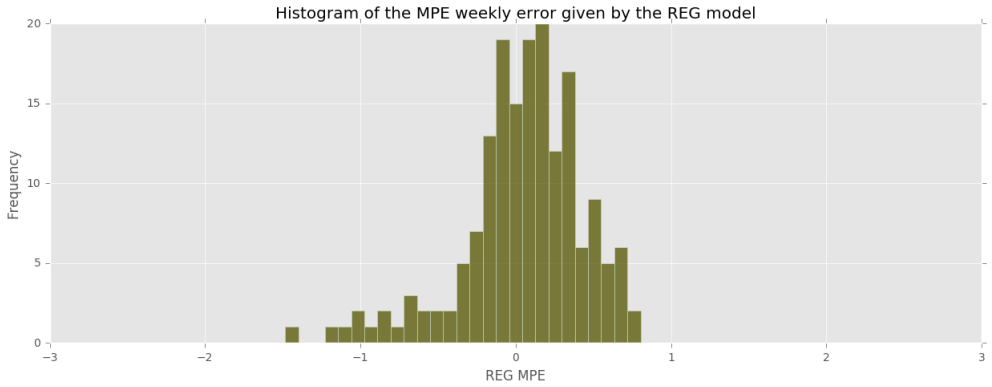


Figure 25 Histogram of the MLR weekly MPE

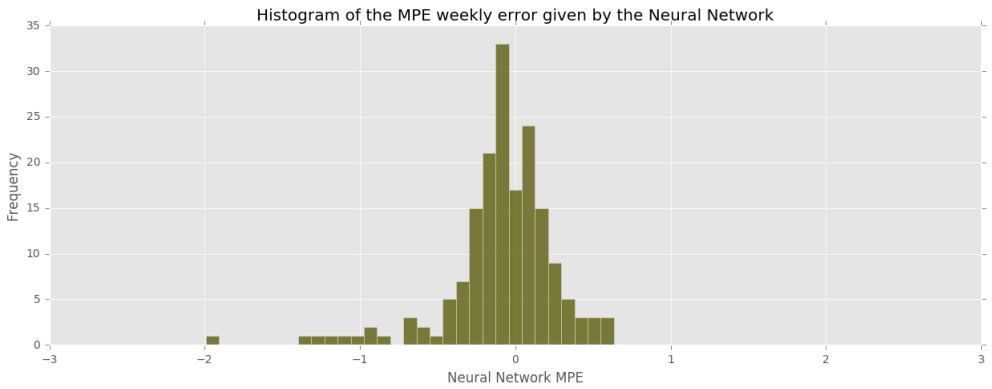


Figure 26 Histogram of the Neural Network weekly MPE

The expected value of the error's distribution (μ) and its standard deviation (σ) are displayed inTable 4.

	ARIMA	MLR	NN
μ	-0.155	0.010	-0.093
σ	0.466	0.511	0.354

Table 4 Models MPE - expected value and standard deviation

We can observe that both MLR and the Neural Networks models have a very well centred value of the expected MPE error, nevertheless, this value as little statistic information since the errors of oposite signals are cancelling each other. Being the most typical method for measuring forecast accuracy, we turn to the MAPE which gives us the average of percentage errors. A histogram of the MAPE results for each model was then computed:

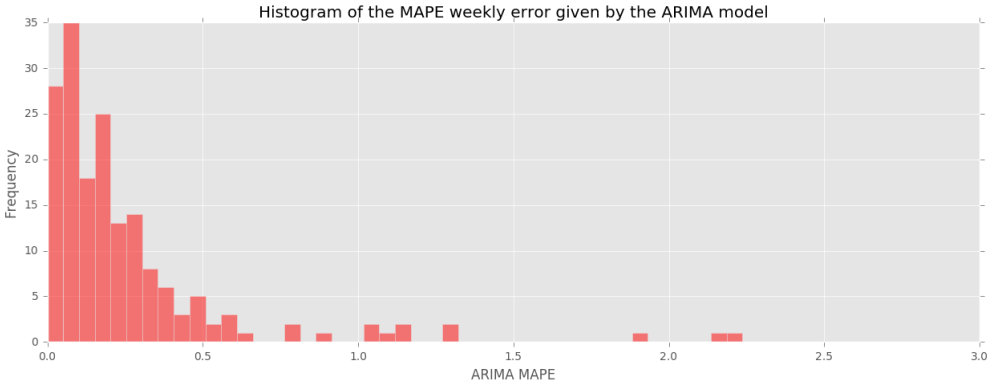


Figure 27 Histogram of the ARIMA weekly MAPE

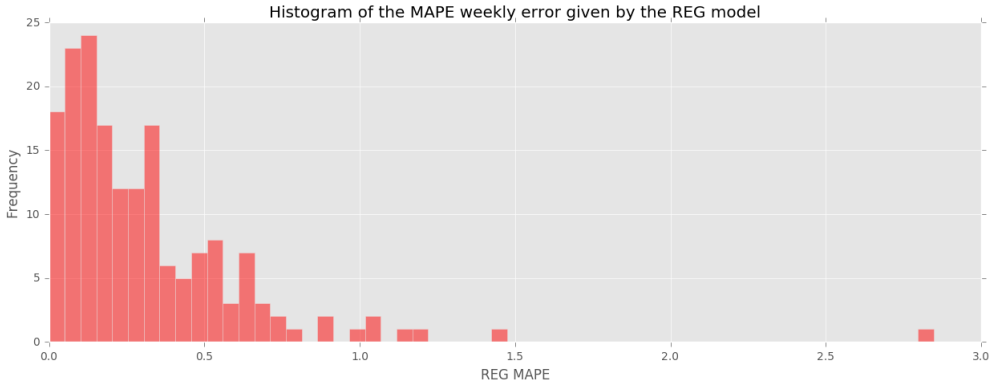


Figure 28 Histogram of the MLR weekly MAPE

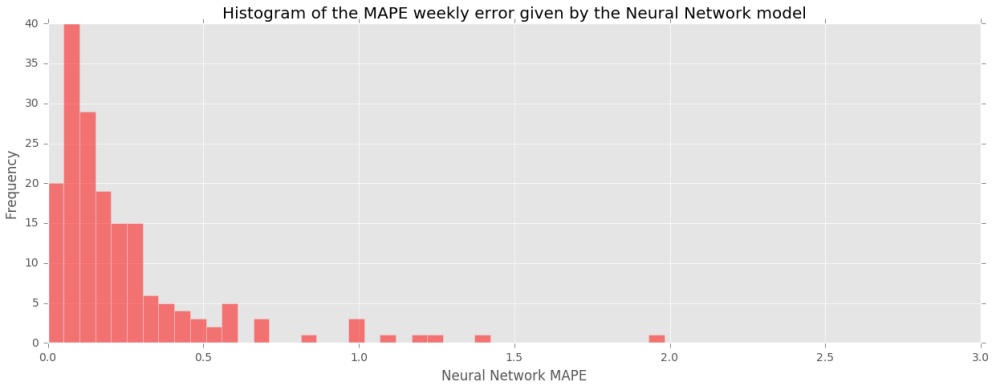


Figure 29 Histogram of the Neural Network weekly MAPE

The standard deviation (σ) is defined as the average distance a value has to the mean, hence it is greatly affected by outliers. Through an analysis of Table 5 we conclude that the most precise algorithm has been the Neural Networks which considered more information than its competitors. Although its average error can be compared with the one provided by ARIMA, it presents a significant improvement in accuracy as it shows a smaller deviation from the mean.

	ARIMA	**MLR**	**NN**
μ	0.280	0.327	0.242
σ	0.404	0.392	0.275

Table 5Models MAPE - expected value and standard deviation

As the overview of the results might hide advantages and disadvantages of the models towards properties of the dataset, we will now focus on analysing the models through different dimensions.

5.4 MAPE by Area

In the problem at hand, five different store areas were analysed. These present specific characteristics which can be better suited for a given model. Figure 30,Figure 31 and Figure 32 display the box plots of the resulting MAPE by Area for each model.

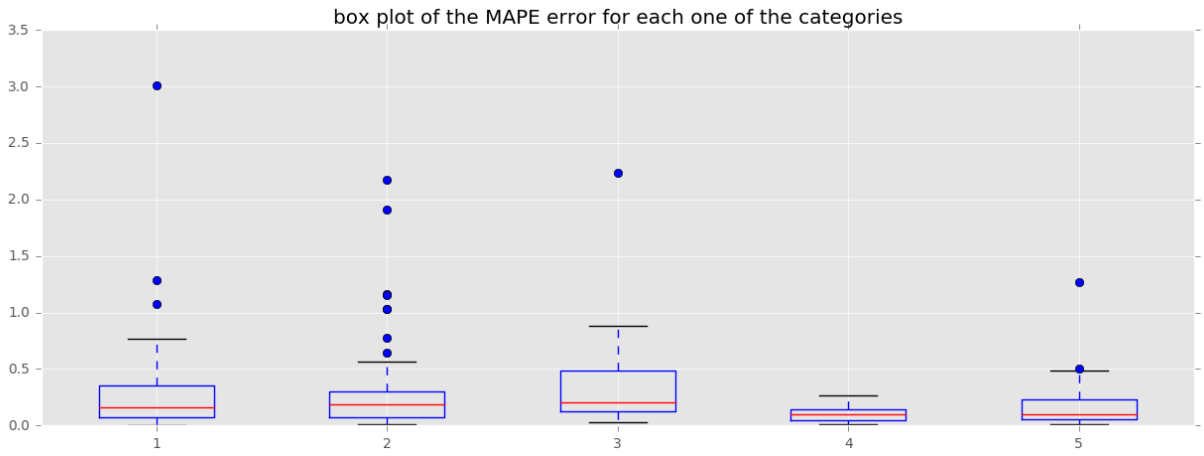
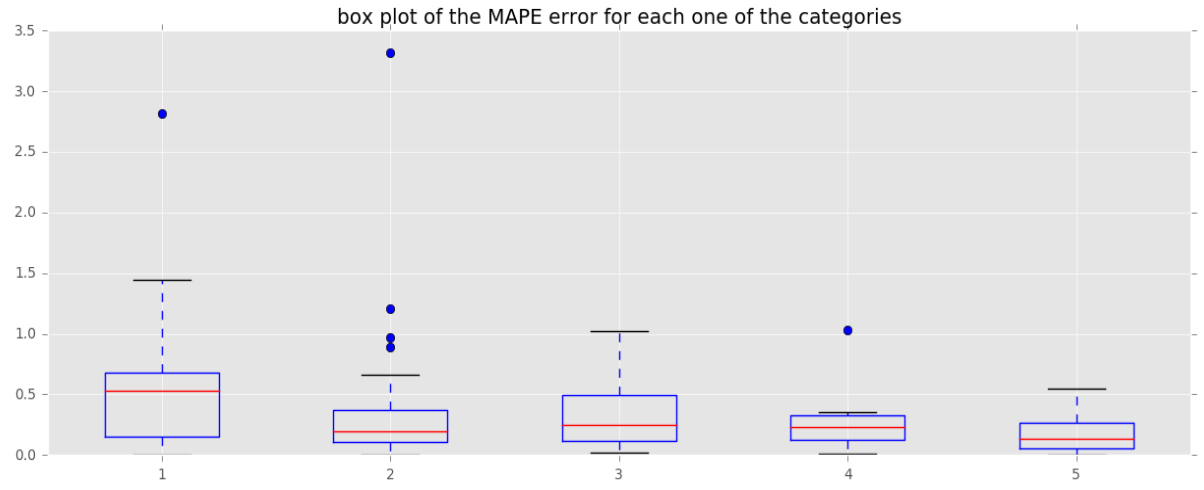


Figure 30 Box plot of the ARIMA MAPE by Area

Figure 31 Box plot of the MLR MAPE by Area



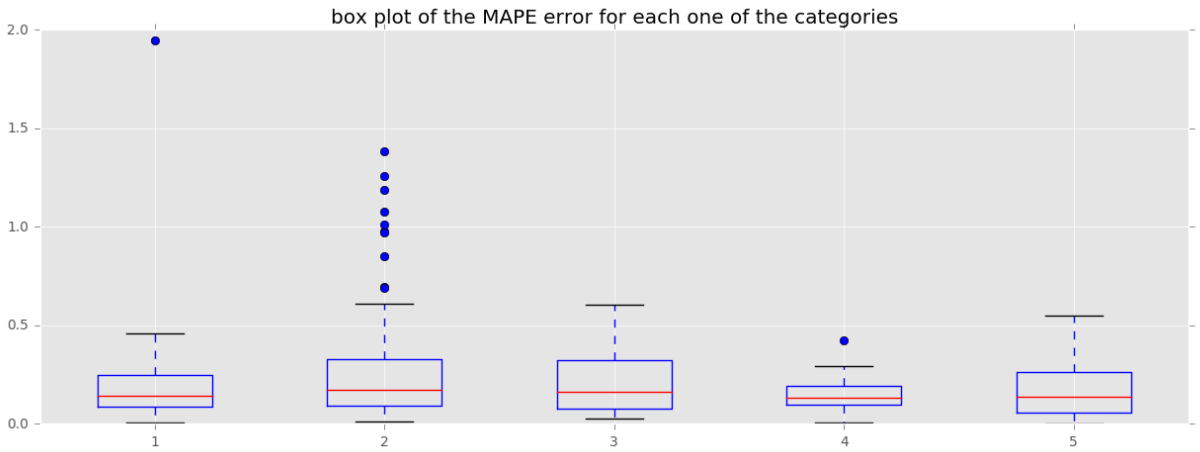


Figure 32 Box plot of the Neural Network MAPE by Area

	ARIMA		MLR		NN	
Areas	**μ**	**σ**	**μ**	**σ**	**μ**	**σ**
1	0.370	0.590	0.555	0.556	0.230	0.348
2	0.278	0.365	0.301	0.395	0.292	0.305
3	0.364	0.445	0.332	0.256	0.226	0.179
4	0.105	0.068	0.255	0.231	0.154	0.099
5	0.207	0.268	0.187	0.155	0.144	0.147

Table 6 Models MAPE by area - expected value and standard deviation

Through this analysis we conclude that, as expected, the Neural network model does not overcome all the others in all the areas. In fact, for Area 4, The ARIMA is significantly better suited. Here it presents a smaller absolute average perceptual error while simultaneously providing a smaller deviation for this expected value. This is most probable caused by the stronger correlation with temporaldimensions, condition in which this model excels the remaining. Although sometimes providing satisfactory results and even sporadically surpassing the ARIMA, the Multi linear regression never achieved the smallest errors. By concluding that there is no absolute best model across all the categories, we now turn to another perspective, making efforts to identify the characteristics which most likely influence the performance.

MAPE by Average Quantity

For each model, the MAPE of the analysed articles was mapped against their average quantity sold to search for patterns (Figure 33). Although the number of observations seems to be insufficient for concrete conclusions, all the results of the three models seem to vary inversely proportional to the average quantity. This would be expected since sensibility to lower number is one of the main drawbacks of the used error formal (MAPE).

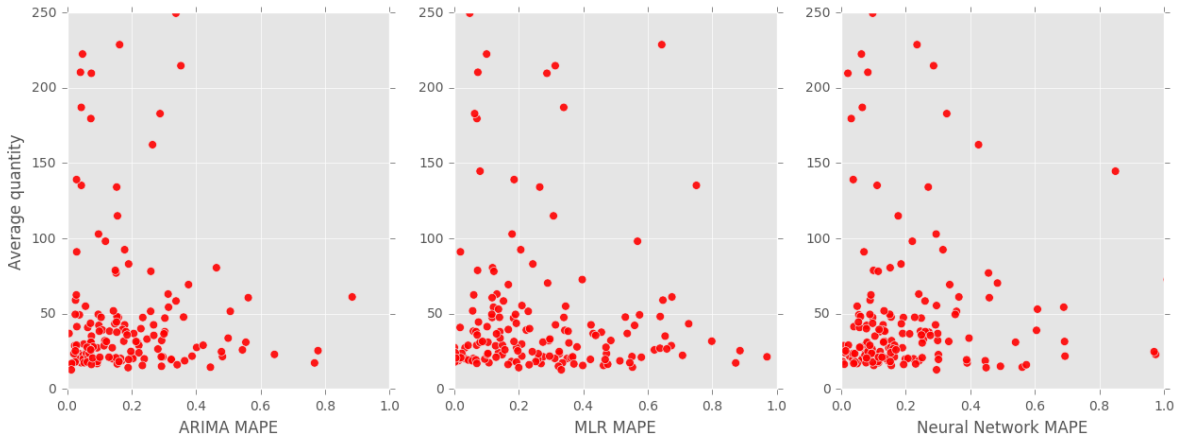


Figure 33 Models MAPE by average quantity

5.5 MAPE by discount fluctuation

Another dimension to be analysed is the fluctuation of the price, here described as a function of the standard deviation of the normalized discount.

$$\text{fluctuation of the price} = \sigma_{\text{DESC Normalized}} * 100 \quad [34]$$

By filtering the articles into three categories of price fluctuation we intend to assess the performance of the algorithms while eliminating progressively the articles most affected by promotional discounts. The following figures display the obtained result.

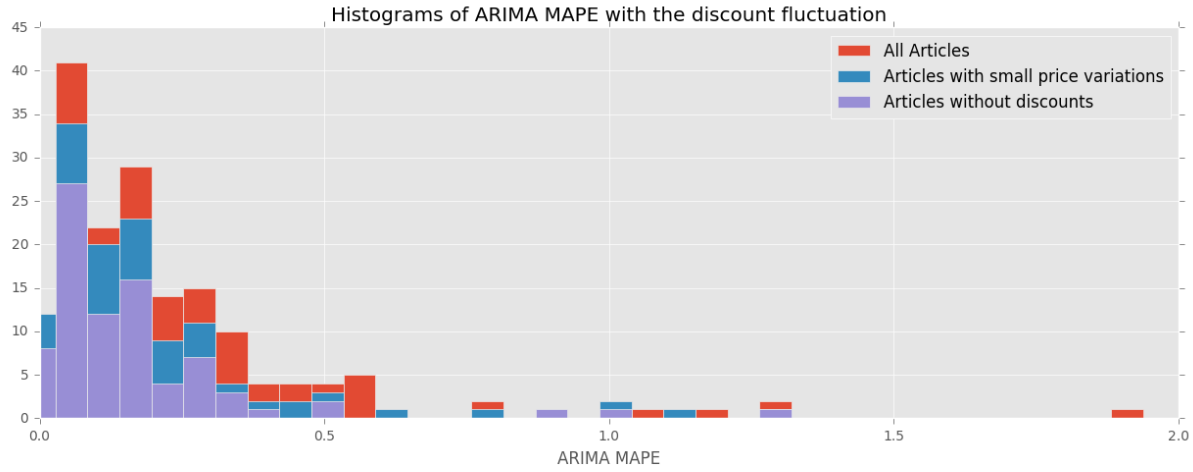


Figure 34 Histogram of the ARIMA MAPE by article promotional strategy

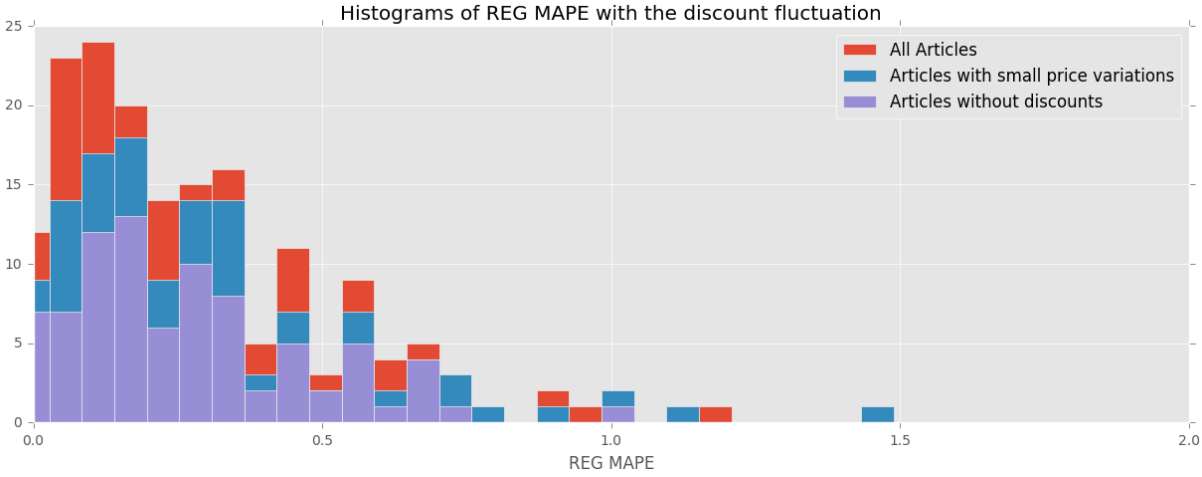


Figure 35 Histogram of the MLR MAPE by article promotional strategy

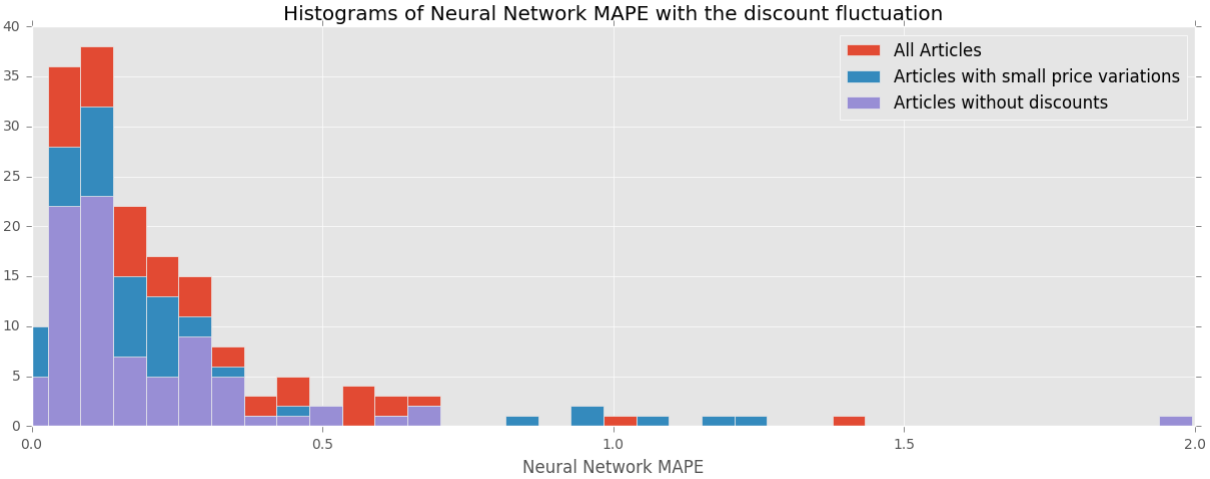


Figure 36 Histogram of the Neural Network MAPE by article promotional strategy

Through Table 7 We can observe that, for items without discount, although with less deviation, the Neural network achieved values similar to the ARIMA, suggesting that here, the extra information that this model considers does not give it a significant advantage over the temporal characterization of the sales. Both these models continue to surpass the results of the Multiple linear regression.

	ARIMA		MLR		NN	
	μ	**σ**	**μ**	**σ**	**μ**	**σ**
All Articles	0.280	0.404	0.327	0.392	0.242	0.275
Articles with small price variations	0.245	0.412	0.316	0.335	0.219	0.275
Articles without discounts	0.195	0.303	0.272	0.204	0.192	0.243

Table 7 Models MAPE by Article promotional strategy - expected value and standard deviation

Another conclusion can be taken from analysing the decay of the expected value of the ARIMA error. This value drops about 0.035 when removing the articles with most volatile prices. Bycomparing with the same variation of the neural network (0.023) and its values we conclude the Neural Network model is better fit to predict articles with promotional discount.

5.6 Combining Best model by article

Every article sales series has different properties. Some are more sensible to discounts or the week day, others might present very consistent seasonal behaviours. It is only natural that the performance of the used model is greatly influenced by the behaviour of each individual.

Methods of combining models in data mining are called Ensemble. Methods such as Boosting, Bayes model combination or bagging are such examples. These combine multiple predictive models into a usually more accurate and robust one. But the usage of these methods

come with a cost, than model interpretability is sacrificed in order to raise predictive accuracy.

In this thesis, we tackled the problem using a distinct model for each article, using the previous performance evaluations of the models.It can be seen as a rudimental decision tree ensemble where each node identifies one article and the model it uses. Further work and analysis of different ensemble methods is suggested in the chapter 7 since the combination of multiple model results for the same article can bring more robustness to results.

Using the results of the previous run where we forecasted the daily sales for four weeks (from 2016.10.12 to 2016.11.08) we can assert the models which seemingly perform better for each article. We can now proceed to use this correspondence of model-article to perform a new forecast of unknown data.

Using as training data the interval of [2012.01.01, 2016.11.08] we now compute the daily sales of each article for the four next weeks ([2016.11.09, 2016.12.06]). Once the models were trained, the errors of their forecast were computed. Their analysis can be observed in the following graphs:

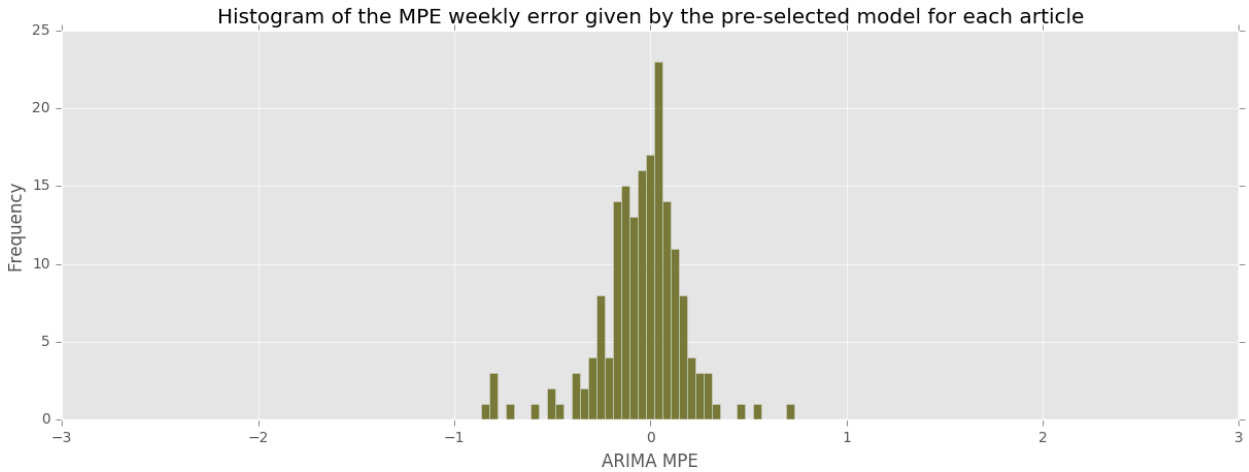


Figure 37 Histogram of the MPE of the pre-selected model

	Pre-selected model
μ	-0.049

σ	0.219

Table 8 MPE of the pre-selected model - expected value and standard deviation

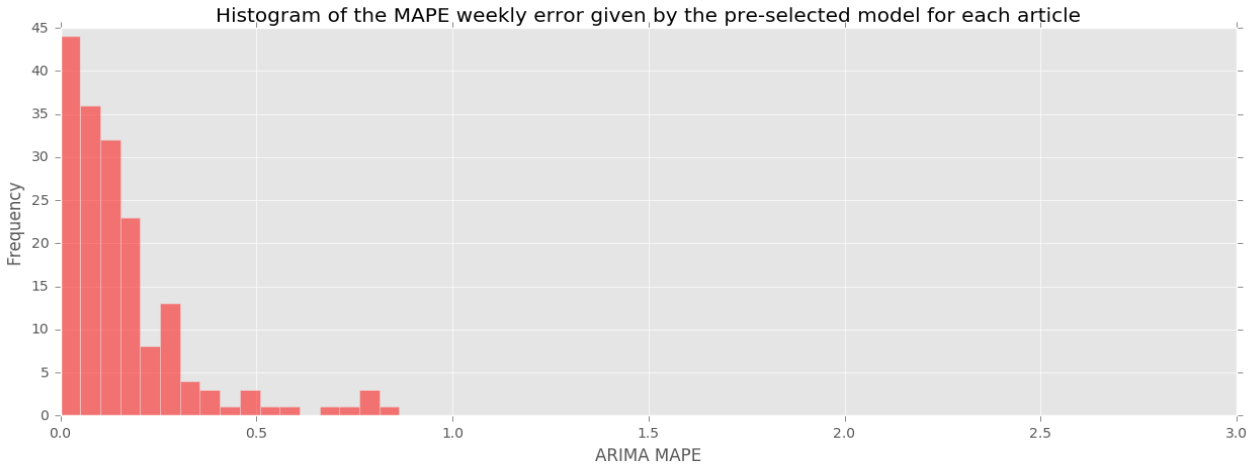


Figure 38 Histogram of the MAPE of the pre-selected model

	Pre-selected model
μ	0.155
σ	0.163

Table 9 MAPE of the pre-selected model - expected value and standard deviation

Even though, similarly to the initial runs (page 53) we are asserting the performance using data unknown to the models, contrary to what happened up until this point, we do not use the same model for all individuals, instead we use the knowledge acquired in the first run to know which model should be used to each individual.

As expected, we are lifting considerably the quality of the results since every model shall be used to compute the predictions where it excels.

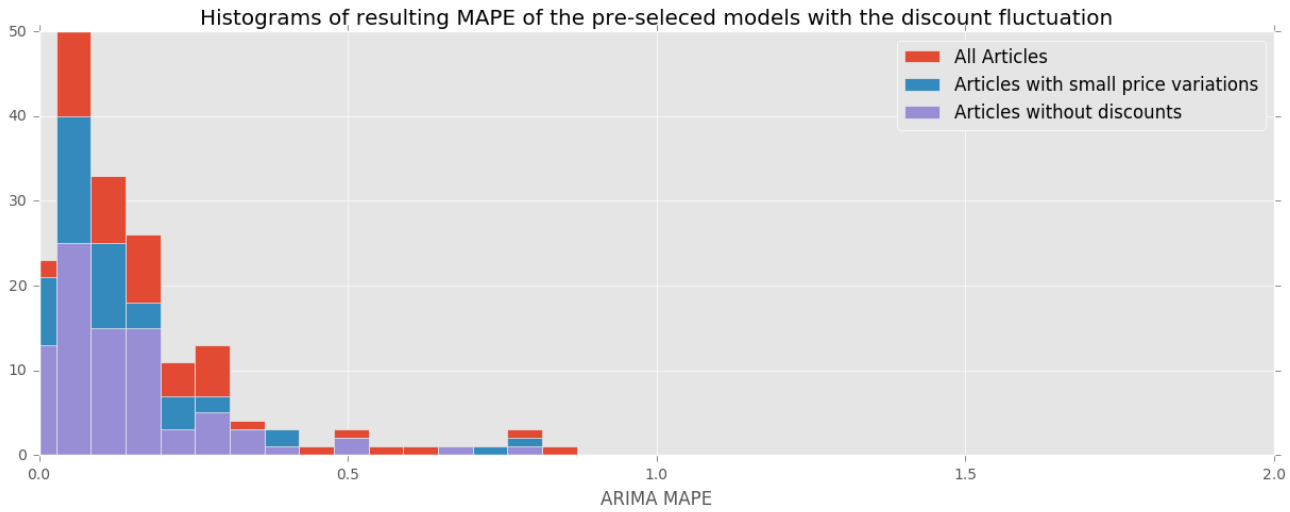


Figure 39 Histogram of the MAPE of the pre-selected model by article promotional strategy

Through the analysis of the Figure 39 and Table 10 we observe that the quality of the results became less dependent of the promotional strategy. This is also expected since here, different algorithms are being used appropriately, decreasing sensibility to volatility of prices.

	Pre-selected model	
	μ	**σ**
All Articles	0.155	0.163
Articles with small price variations	0.137	0.15
Articles without discounts	0.14	0.143

Table 10 MAPE of the pre-selected model by Article promotional strategy - expected value and standard deviation

6. CONCLUSIONS

In retail business, understanding the costumer behaviour and predicting demands is crucial for an efficient planning of production, transportation and purchasing as to optimize processes and avoid expenses due to inadequate stock levels.But Retail is a complex practice.Althoughsales have properties of time series, which are characterized by trend and seasonal patterns,unpredictive events and impact of external agents frequently characterize the problems.

In the literature regarding prediction, it is well established that no quantitative model would be ideal for all situations under all circumstances. One must analyse, experiment and conclude the proper configuration for a given problem.Many studies present different forecast models using linear and non-linear methods for quantitative demand forecast, following different paradigms from different areas.However, they focus usually on a less granular dataset byusing aggregated data, usually forecasting the total sales of entire categories and focus on optimizing one model for the entire population.

This project had as an objective the development of a fully-automatic algorithm that is capable of selecting key explanatory variables from very large data sets and using them in order to forecast demand.Support visual tools have been developed in the shape of dashboard in order to assist the assessment of each model's performance.

Tackling the problem of dimensionality, three models have been studied and applied. Later on, an evaluation of the forecasting of four weeks of daily sales has been executed.Assessing the performance of the models through different perspectives, we concluded that some models have better performances for different subset of articles. We then proceed to use the knowledge of the previous run and, by assigning the best model at article level, we boost the expected accuracy achieving an average absolute error of 15.5%. This error is con-

siderably smaller than 24%, the 2008 benchmark MAPE error of the retail industry for a one-month ahead forecast (Kolassa, 2008).

7. FUTURE WORK

This thesis focuses on the creation of a forecasting tool that automatically treats input data, pre-processing it and selecting the most relevant data.The tool provides a daily forecast for an interval of four weeks, which is plenty of time for a retail company to allocate resources and plan in advance. But, nevertheless, many improvements can be executed.

More sources of data should be considered as they can increase the tool's potential,either internal, such as shelf positioning and used marketing channels for advertising promotions, as well as external, such as the discount and marketing strategies of the competitors.These additional sources often entail extensive human resources to assemble.

Following the study of new ensemble models for the problem at hand, new methods for combining the models should be analysed in order to improve classifier accuracy and robustness, decreasing the deviation of the MAPE error.

Additionally, clustering algorithms such as K-means or self-organizing maps (SOM) could be used to labelarticles according to their behaviours towards the input variables and performance towards each model. This can potentially bring to light useful marketing knowledge of the article sets. Furthermore, following a clustering of articles by their behaviour in the variables of interest, as an extension of the project, the use of combining models would be advised, having into account that distinct models proved to produce significantly better forecasts for distinct subsets of articles.

Following the results of the thesis, a new automated project can be built on top of it in order to advise inventory replenishment and maintain efficient fill rates. This project would also bring another advantage since, by comparing the daily forecast with the real sales, cases of phantom inventory can potentially be identified at daily basis.

In this work, near two hundred articles were considered, this number still represents a small subset of the store and through more computational adequate tools, more articles and

more stores should be considered. Case the project faces enough articles and warranties a similar performance, by controlling and changing distinct inputs, such as the discount configuration of the next weeks, we can observe the changes in forecast. Potentially, we can get the combination of inputs which would maximize the sales. We can leave the field of predictive analytics and enter the field of prescriptive analytics.

8. BIBLIOGRAPHY

Aburto, L., & Weber, R. (2007). Improved supply chain management based on hybrid demand forecasts. *Applied Soft Computing*.

Alon, I., Qi, M., & Sadowski, R. J. (2001). Forecasting aggregate retail sales. *Journal of Retailing and Consumer Services*.

Au, K. F., Choi, T. M., & Yu, Y. (2008). Fashion retail forecasting by evolutionary neural networks. *International Journal of Production Economics*, 615-630.

Barry Berman, Joel R. Evans. (1989). *Retail management: a strategic approach.* Macmillan.

Box, G. E. P., and Jenkins, G. M. (1974). Time Series Analysis: Forecasting and Control. *Journal of Time Series Analysis, 3*.

Cooper, L. G., Baron, P., Levy, W., Swisher, M., & Gogos, P. (1999). A New Forecasting Method for Promotion Planning. *Marketing Science*.

Corsten, D. and Gruen, T. (2003). Desperately seeking shelf availability: an examination of the extent, the causes, and the efforts to address retail out-of-stocks. *International Journal of Retail & Distribution Management, 31*(12), 605-617.

David Walters, David White. (1988). *Retail Marketing Management.* Macmillan.

ECR Europe – Optimal Shelf Availability, Increasing Shopper Satisfaction at the Moment of Truth, ECR Europe, Brussels. (2003).

Geurts, M. D., & Patrick Kelly, J. (1986). Forecasting retail sales using alternative models. *International Journal of Forecasting*.

Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning.* MA: Addison Wesley.

Gür Ali, Ö., SayIn, S., van Woensel, T. and Fransoo, J. (n.d.). SKU demand forecasting in the presence of promotions. *Expert Systems with Applications, 36*(10), 12340-12348.

Holland, J. H. (1975). *Adaptation in Natural and Artificial Systems.* University of Michigan Press.

Ilan Alon , Min Qi , Robert J. Sadowski. (2001). Forecasting aggregate retail sales: A comparison of artifcial neural networks and traditional methods. *Journal of Retailing and Consumer Services, 8*, 147-156.

Jianqing Fan, Jinchi Lv. (2008). Sure independence screening for ultrahigh dimensional feature space. *Royal Statistical Society, Series B - Statistical Methodology, 70*(5), 849–911.

Kolassa, S. (2008). *Can we obtain valid benchmarks from published surveys of forecast accuracy?* (FORESIGHT, Ed.) Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.462.5576&rep=rep1&type=pdf

Levy, Michael, Dhruv Grewal, Praveen K. Kopalle and James Hess. (2004). Emerging Trends in Retail Pricing Practice: Implications for Research. *Journal of Retailing, 80*(3), 13–21.

Ma, S., Fildes, R. and Huang, T. (2014)). Lancaster University: The Department of Management Science. Demand forecasting with high dimensional data: the case of SKU retail sales forecasting with intra- and inter-category promotional information. *LUMS Working Paper 2014:, 9*.

Ma, Shaohui, Robert Fildes, Tao Huang. (2015). Demand forecasting with high dimensional data: The case of SKU retailsales forecasting with intra- and inter-category promotional information. *European Journal of Operational Research*.

Mitchell, M. (1995). Genetic Algorithms: An Overview. (W. O. Library, Ed.) *Complexity*, 31-39.

N. Melab, S. Cahon and E. Talbi. (2006). Grid computing for parallel bioinspired algorithms. *Journal of Parallel and Distributed Computing - Special issue on parallel bioinspired algorithms, 66*(8).

Pan, Y., Pohlen, T., & Manago, S. (n.d.). (2014). Hybrid neural network model in forecasting aggregate U.S. retail sales. *Business and Management Forecasting, 9*, 152-167.

Thiesing, F. M., & Vornberger, O. (1997). *Computational Intelligence Theory and Applications* (Vol. 26).

Trusov, M., Bodapati, A. V., & Cooper, L. G. (2006). Retailer promotion planning: Improving forecast accuracy and interpretability. *Journal of Interactive Marketing, 20*(3), 71-81.

Y. LeCun, L. Bottou, G. Orr and K. Muller. (1998). Efficient BackProp. (Springer, Ed.) *Neural Networks: Tricks of the trade*.

Zhang, G. P. and Chu, C. (2003). A comparative study of linear and nonlinear models for aggregate retail sales forecasting. *International Journal of Production Economics, 86*(3), 217-231.

Zhang, G. P. (n.d.). *Neural Networks for Retail Sales Forecasting.*

Zhang, G. P., & Qi, M. (2005). Neural network forecasting for seasonal and trend time series. *European Journal of Operational Research, 160*, 501-514.

Zhang,Cheng, Song, Peijian and Heng Xu. (2011). The influence of product integration on online advertising effectiveness. *Electronic Commerce Research and Applications, 10*, 288-303.

Zhao, X., Xie, J., & Lau, R. S. M. (2001). Improving the supply chain performance: Use of forecasting models versus early order commitments. *International Journal of Production Research*.