FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Deep learning approaches to sales forecasting of retail healthcare and wellness products

Mário Santos

Mestrado Integrado em Engenharia Informática e Computação

Supervisor: Alexandra Oliveira

Second Supervisor: Ana Paula Rocha

February 1, 2021

Deep learning approaches to sales forecasting of retail healthcare and wellness products

Mário Santos

Mestrado Integrado em Engenharia Informática e Computação

Approved in oral examination by the committee:

Chair: Prof. João Pedro Carvalho Leal Mendes Moreira
External Examiner: Prof. Joaquim José de Almeida Soares Gonçalves
Supervisor: Prof. Alexandra Alves Oliveira

February 1, 2021

Abstract

The healthcare and wellness products retail market can be characterized by the medium to long shelf-life of the products, the dynamism in the market, its competitiveness, and scale. In this market, considering the effects of promotions or promotional events is gaining importance, as it's been shown that they can greatly increase store traffic and sales volume [44][3]. Additionally, product demand can also be affected by other variables such as weather events, distance to nearest competitor, holidays, and more [2]. Currently, the main forecasting models used to produce sales predictions fall into 3 categories, the traditional time series models, such as the Auto-Regressive Integrated Moving Average (ARIMA) or Holt Winter's (HW), the machine learning models, like Support Vector Machines (SVM) or Tree-based models, and lastly the deep learning models, a subsection of the machine learning models, namely Recurrent Neural Networks (RNN) and Long Short Term Memory Networks (LSTM).

Using traditional models, in cases where demand is influenced by promotions, weather effects and special events like holidays, it isn't yet possible to generate accurate sales predictions. The abundance of promotional methods and their interactions can affect demand, further increasing the difficulty in forecasting product sales using the traditional methods. Furthermore, both traditional and newer approaches have difficulty when forecasting sales for intermittent and slow-moving products.

We compare traditional and state of the art approaches to the problem of sales forecasting, incorporating exogenous variables such as different promotional methods and types, weather effects, holidays, google analytics data, and more. We were particularly interested in a newer approach named Deep AR, using Auto-Regressive Recurrent Neural Networks (RNN), which seems to outperform its alternatives when applied to large volumes of data [43]. This approach allows for the incorporation of external variables, like promotions and weather events, and is also able to produce sales forecasts for products recently introduced in the market, with very limited sales histories. By applying each approach to the best of our ability, and selecting and identifying the most appropriate performance measures that suited our modeling process and data, we provide a fair comparison between the Deep AR, Prophet, SARIMAX, and Holt Winter's models. A performance analysis was conducted for multiple time series from multiple data sets that differed in scale, variability, seasonality, and context. One data set was acquired from a Kaggle competition, and it includes the store aggregated daily sales history, in euros, of several stores from the german drug stores chain Rossmann (store-level sales forecasting). The second data set was provided by Retail Consult and includes the daily sales history, in product units, of multiple products from multiple stores owned by a large Portuguese healthcare and wellness products retailer (product-store level sales forecasting). The time series selected for comparing the models were arranged into multiple groups to provide a more in-depth analysis of each models' performance (4 groups based on median daily sales of each store for the Rossmann data set, and 3 groups based on average daily sales of each product-store for the Retail data set). We assessed how the Deep AR model compared with other models when handling either seasonal or highly intermittent data, and when

supplied with and without external variables, by graphically analyzing the plots generated by each model (predictions and prediction intervals) and measuring their performance using the MASE, considered to be one of the most suitable error metrics for multiple series and particularly in the case of intermittent series [29].

Overall, results indicate that the Deep AR can be a highly suitable approach to the problem of daily sales forecasting when supplied with a large training set containing multiple time series, in particular for intermittent and slow-moving products. We also observed that it benefited from the inclusion of multiple external variables such as different promotions and events. It exhibited the best overall performance in the case of seasonal data and the case of highly intermittent data. In both situations, it generated the most accurate and precise prediction intervals. In the case of the Rossmann data set, the Deep AR reached the lowest average MASE values, out of all the models, for the low median daily sales (MASE = 0.208) and the medium-low median daily sales group (MASE = 0.184). Regarding the Retail data set, it achieved the lowest average MASE across all time series (MASE = 0.695). We conclude that in the conducted experiments the Deep AR model displayed the desirable traits of a valuable stock management and revenue optimization tool for retailers of healthcare and wellness products.

CCS Concepts: •**Applied computing** → Forecasting; •**Applied computing** → Marketing; •**Computing methodologies** → Machine learning algorithms;

Additional keywords and phrases: deep learning, forecasting, promotions, healthcare, wellness, market

Resumo

O mercado de retalho de produtos de saúde e bem estar pode ser caracterizado pelo tempo de vida médio a longo dos produtos, o dinamismo do mercado, a sua competitividade e escala. Neste mercado, considerar os efeitos de promoções e eventos promocionais é cada vez mais relevante pois já foi demonstrado que estes podem aumentar significativamente o número de clientes e o volume de vendas [44][3]. A procura também pode ser afetada por mais fatores tais como feriados e metereologia [2] entre outros. Atualmente, os métodos para previsão de vendas mais utilizados estão agrupados em 3 categorias, nomeadamente os métodos tradicionais, tais como o ARIMA ou Holt Winter's, métodos de machine learning, como Suppport Vector Machines (SVM) ou Tree-based, e finalmente os métodos de deep learning, uma sub-secção de machine learning, como por exemplo Recurrent Neural Networks (RNN) e Long Short Term Memory Networks (LSTM).

Em casos em que a procura é influenciada pelas variáveis mencionadas (promoções, metereologia, feriados), os métodos tradicionais são incapazes de produzir previsões com alta precisão. Adicionalmente, a grande variedade de métodos promocionais e as suas interações também podem afetar a procura, e consequemente dificultam o processo de produção de previsões. Ademais, tanto as abordagens tradicionais como as modernas apresentam dificuldades em casos de séries temporais intermitentes e de produtos de baixo-movimento.

Esta dissertação tem como objetivo comparar a recente abordagem Deep AR com as alternativas tradicionais e modernas, aplicadas ao problema de previsão de volumes de vendas, incorporando variáveis externas como efeitos promocionais, informação metereológica, dados recolhidos através de google analytics, entre outros. A literatura sugere que a abordagem Deep AR, usando Auto Regressive Neural Networks (ARNN), é capaz de um desempenho superior aos métodos alternativos, quando aplicada a grandes volumes de dados [43]. Esta abordagem permite a incorporação de variáveis externas, como promoções e feriados, e também é capaz de gerar previsões de volumes de vendas para produtos recentemente introduzidos no mercado, e portanto com históricos de vendas limitados. Aplicando cada abordagem o melhor que conseguimos, e identificando e selecionando as medidas de desempenho que melhor se adequam ao nosso processo modelativo e dados utilizados, providenciamos uma comparação justa entre os modelos Deep AR, Prophet, SARIMAX e Holt Winters. Analisamos o desempenho de cada abordagem, aplicada a múltiplas séries temporais de 2 conjuntos de dados distintos que variavam em escala, variabiabilidade, sasonalidade e contexto. Também analisamos o seu desempenho na presença e na ausência de variáveis externas como promoções, feriados, variáveis metereológicas, dados relacionados com a concorrência e dados de google analytics. Um dos conjuntos de dados foi obtido através de uma competição Kaggle, e inclui o histórico de vendas diárias, em euros, agregados por loja, de várias lojas da farmacêutica alemã Rossmann (previsões ao nível da loja). O segundo conjunto foi fornecido pela Retail Consult e inclui o histórico de vendas diárias de vários produtos em várias lojas, de produtos de saúde e bem-estar de um revendedor Português de grande dimensão. As séries temporais selecionadas para comparação foram organizadas em vários grupos para obtermos uma análise de desempenho mais detalhada (4 grupos baseados na mediana de vendas diárias

de cada loja para os dados Rossmann, e 3 grupos baseados na média de vendas diárias de cada produto em cada loja para os dados Retail). Comparamos os desempenhos de cada modelo através da medida MASE, considerada uma das mais adequadas quando comparando o desempenho de múltiplas séries temporais e especialmente no caso de séries intermitentes [29]. Adicionalmente, comparamos os gráficos das previsões e intervalos de previsão gerados pelos vários modelos.

Em geral, os resultados indicam que o Deep AR é uma abordagem adequada ao problema de previsão de vendas diárias quando lhe é fornecido um grande conjunto de dados para treino, contentendo múltiplas séries temporais, particularmente, para produtos de baixo-movimento e com vendas intermitentes. Também observamos que esta abordagem beneficia com a incorporação de variáveis externas, como promoções, feriados, dados metereológicos e mais. Na maior parte dos casos o Deep AR demonstrou o melhor desempenho, tanto na presença de séries temporais periódicas como séries temporais intermitentes. Em ambos os casos produziu as previsões e intervalos de previsão mais precisos e com menor variância. No caso dos dados Rossmann, o modelo Deep AR atingiu um valor médio de MASE mais baixo que os outros modelos, tanto para o grupo de de lojas com uma mediana de vendas diárias baixas (MASE = 0.208) como para o grupo com uma mediana de vendas diárias média-baixa (MASE = 0.184). No caso dos dados Retail, considerando todas as séreis temporais, o Deep AR obteve o valor MASE médio mais baixo (MASE = 0.695). Concluímos que as experiências realizadas demonstram que o Deep AR exibe as características desejáveis de uma valiosa ferramenta de gestão de stock e otimização de rendimento para revendedores de produtos de saúde e bem-estar.

Conceitos CCS: •**Computação aplicada** → Previsões; •**Computação aplicada** → Marketing; •**Metodologias de computação** → Algoritmos de machine learning;

Palavras-chave e frases adicionais: deep learning, previsões de vendas, promoções, saúde, bem-estar, mercado de retalho

Acknowledgements

First and foremost I wish to express my deepest gratitude to Alexandra Oliveira and Professor Ana Paula Rocha, who guided me throughout this whole process. The completion of this study would not have been possible without their continuous mentoring, guidance, and commitment.

My appreciation extends to our partner, Retail Consult, who provided material to conduct this dissertation. Vitor Rangel Rodrigues, Nuno Tiago Maia dos Santos, and Ana Rita Novo, the team from Retail Consult that accompanied this dissertation provided valuable insight, continuous support, and greatly assisted in its completion.

In addition, a thank you to Professor Luís Paulo Reis, head of the Laboratory of Artificial Intelligence and Computer Science, for his insightful comments and enthusiasm which had a lasting effect on me and further motivated my work.

Lastly, I'd like to thank the Gluon TS Python toolkit developer team for their quick and cooperative responses to all questions solicited during this study.

Mário Santos

"It doesn't matter how beautiful your theory is, it doesn't matter how smart you are. If it doesn't agree with experiment, it's wrong."

Richard P. Feynman

Contents

List of Figures

List of Tables

Abbreviations

ANN	Artificial Neural Network
AR	Auto Regressive
ARIMA	Auto Regressive Integrated Moving Average
ES	Exponential Smoothing
ETS	Error Trend and Seasonality
GBR	Gradient Boosting Regression (GBR)
HW	Holt Winter
HWES	Holt Winter's Exponential Smoothing
ISSM	Innovative State Space Model
LSTM	Long Short Term Memory
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MASE	Mean Absolute Scaled Error
MdAE	Median Absolute Error
MdAPE	Median Absolute Percentage Error
MdASE	Median Absolute Scaled Error
MLPNN	Multilayer Perceptron Neural Network
MLR	Multiple Linear Regression
MSE	Mean Square Error
QR	Quantile Regression
RBF	Radial Basis Functions
RMdSPE	Root Median Square Percentage Error
RMSE	Root Mean Square Error
RMSPE	Root Mean Square Percentage Error
RMSSE	Root Mean Squared Scaled Error
RNN	Recurrent Neural Networks
SARIMA	Seasonal ARIMA
SARIMAX	SARIMA with EXogenous regressors
SVM	Support Vector Machine
SVR	Support Vector Regression
WNN	Wavelets Neural Network
TS	Takagi-Sugeno
XGBR	Extreme Gradient Bossting Regression

Chapter 1

Introduction

1.1 Context

Retail consists of the sale of goods to the public for use or consumption rather than for resale. Within the retail market, our interest is in the healthcare and wellness products retail market, which can be described by the following characteristics:

- **Perishability** (shelf-life) of the products is medium to long
- **Dynamism**, given the frequent and large entry of new products into the market.
- **Competitiveness**, with retailers frequently adopting several strategies, namely promotions, to gain an advantage over competitors.
- **Dimension**, being one of the largest retail markets on a global scale, resulting in a very large number of retailers, products, and sales.

Sales forecasting is particularly useful for short and medium shelf-life product retailers since successful sales forecasting considerably reduces lost sales and product returns. This is crucial not only for revenue optimization but also due to the environmental factor since the returned product is usually discarded, resulting in waste.

The factors that may affect sales and demand were categorized by Arunraj *et al.* (2015) [12] as internal, partially internal, and external factors.

- **Internal Factors** - Promotions, Discounts, which are known to the retailer.
- **Partially Internal Factors** - cannibalization and complementarity.
- **External factors** - These can be observed, like sports events, festivals, holidays, abnormal events (health crisis, terror attacks), or can be forecasted like the national economy and the weather.

By affecting the consumers' demand, these factors may result in over-stocking, which leads to a loss of revenue and waste, or under-stocking, leading to situations where the consumer can't purchase the desired product [2]. This might make the customer switch to another option or, in the worst scenario, resort to a competitor retailer. As a result of these demand influencing factors, high volatility, skewness, multiple seasonal cycles, intermittence with zero sales, and stock-outs characterize the product-level sales data [21].

1.2 Motivation

Retailers must decide on their strategic development in a changing competitive and technological environment. The standard currently adopted market strategy and competitive factor defining elements within the developing technological environment is, normally, forecast dependent [33].

The purpose of product demand forecasting in a retail scenario is to generate predictions for a large quantity of data (time series), over a given period of time (forecast horizon). Accurately forecasting the demand for each product sold in each retail store is essential for the growth and viability of a retail chain. Given that many decisions, such as space allocation, availability, inventory management, ordering, and pricing for a product are directly related to its demand forecast (see Fildes et al. 2019 [21]). Decisions regarding ordering must avoid over-stocking or under-stocking to avoid high inventory costs, and particularly in the case of medium shelf-life products, the expiration of the product while still in the warehouse. Conversely, they must ensure that the inventory level isn't too low, to avoid high inventory costs or stock-outs and consequently loss of sales.

With regards to promotions, the regular and relative price discounts are variables that can be incorporated in a forecasting model to improve its performance [21]. Several retail software products incorporate these factors, as is the case of Systems Applications & Products in Data Processing (SAP). Also, different promotion types (e.g. buy-one-get-one-free vs 50% discount) which may seem similar to a customer, but aren't, result in effects that aren't reflected solely by the unit price. However, it isn't yet possible, with the traditional approaches, to accurately predict if a promotional event will increase or decrease the volume of sales of the promoted product.

Another key point is the sheer abundance of promotional methods, which can be divided into strategic (e.g. price discounts, card discounts, buy-one-get-one-free) or communicational (e.g. flyers, coupons, social media, different tag sizes, and displays). This becomes a challenge in forecasting product sales using the traditional methods, where incorporating complex input data (with many variables) does not result in more accurate sales forecasts (Gur Ali *et. al* 2009) [7]. All the possible combinations of promotion methods and types will interact differently, and retail marketers regularly introduce new promotional methods. This ensures that products have to be forecasted with mixed and varying promotional methods, not previously observed for the considered product.

1.3 Objectives

Our general objective is to compare the forecasting performance of state of the art and traditional sales forecasting methods when applied to healthcare and wellness products.

Our specific goals are:

- Train validate and apply the newer Deep AR approach to the problem of healthcare and wellness retail product sales forecasting.
- Train validate and apply common traditional and more recent approaches to the same problem, including Holt Winter's, SARIMA, SARIMAX, and Prophet.
- Compare the various approaches using multiple data sets with distinct contexts and structure:
  - Generating store aggregated daily sales volume forecasts in euros (forecasting at the store level). Regular time series with similar generating processes
  - Generating daily forecasts for individual products in particular stores in product units (forecasting at the product-store level). Irregular and highly distinct time series
  - Considering each approach, compare performance in time series grouped by close generating processes within the same data set
- Analyse and select performance measurements that best suit our modeling process and data
- Identifying the relationship between external variables, such as promotional or weather effects, and the target variable, daily sales
- Identifying when and if these features should be incorporated in the considered models.
- Assisting in the development of the tool kits utilized.

Recent literature seems to indicate that the new Deep AR approach can outperform alternative state of the art approaches, in forecasting accuracy, when analyzing large volumes of data and predicting sales for products recently introduced in the market, with very few time series available for training (see Salinas *et al.* 2019 [43]). We select the Deep AR model from GluonTS [5] [6], a Python toolkit for probabilistic time series modeling, built around Apache MXNet. This implementation has already been proven capable of generating accurate forecasts on large data sets [43].

1.4 Document Structure

This document is divided into five chapters. The current chapter, Introduction, describes the context and motivation of the work and enumerates its objectives. The following chapter, State of the Art 2, comprises our literature review, including common forecasting solutions 2.1, demand

influencing factors and their use as external variables 2.2, the introduction of machine learning approaches to the problem 2.3, the novelty deep learning approach Deep AR 2.5 and the most common error metrics used to assess forecasting performance2.6. Additionally in section 2.7 we provide an overview of the reviewed literature referenced in this section, regarding comparisons between different forecast approaches.

In chapter 3 3, methodology, provides an overview of the data sets and models used in this study.

The results are outlined in Chapter 4 4, both the given performance metric for each model and an analysis of the plotted forecasts and prediction intervals.

Lastly, in Chapter 5 5 we provide our conclusions on this study and recommendations for future work.

Chapter 2

State of the Art

The work mentioned in the following sections were selected out of the full literature review due to their particular interest to this specific dissertation. The remainder of the reviewed literature, which only indirectly addresses the related issues tackled in this dissertation, can be found in the appendix A in tables A.1 and A.2.

2.1 Common Solutions and Traditional Models

When forecasting sales, the expected demand can depend on changes in trend or seasonality. Simple exponential smoothing along with its variations that incorporate trend and seasonality, such as Seasonal Auto-Regressive Integrated Moving Average (SARIMA) and Holt-Winter's (HW) model, are the most commonly used market-level sales forecasting approaches [21]. Auto-Regressive, Exponential Smoothing, random-walk, and random-trend models are all special cases of $ARIMA\,(p,d,q)$ models. Research indicates that traditional time series models with stochastic trend, such as HW's Exponential Smoothing (HWES) variation and ARIMA, produced reliable forecasts if macroeconomic conditions were relatively stable [42]. However, the ARIMA model presumes certain conditions that might not be true (e.g. assumes the historical patterns of data won't vary in the forecast horizon [35])

The ARIMA model can take the following form:

$$(1-\phi_1 B-\ldots-\phi_p B^p)\,(1-B)^d\,y_t = c+(1+\theta_1 B+\ldots+\theta_q B^q)\,\varepsilon_t \tag{2.1}$$

And can be divided in 3 components:

$$(1-\phi_1 B-\ldots-\phi_p B^p) \rightarrow AR(p) \tag{2.2}$$

$$(1-B)^d\,y_t \rightarrow d\_differences \tag{2.3}$$

$$c+(1+\theta_1 B+\ldots+\theta_q B^q)\,\varepsilon_t \rightarrow MA(q) \tag{2.4}$$

Where:

- p is the order of the auto-regressive (AR) part
- d is the degree of first differencing involved (I),
- q is the order of the moving average part (MA)
- B is the backshift operator
- ε_t is the white noise
- θ is the slope coefficient

The Holt-Winters' seasonal method has two variations, the additive method, and the multiplicative method. When seasonal variations are approximately constant throughout the time series, the additive method should be employed, however, if the seasonal variations vary proportionally to the level of the series, the multiplicative method should be used.

The following is the additive component form:

$$\hat{Y}_{t+h|t} = l_t + hb_t + s_{t+h-m(k+1)} \tag{2.5}$$

Where:

$$l_t = \alpha\,(y_t - s_{t-m}) + (1-\alpha)\,(l_{t-1} + b_{t-1}) \tag{2.6}$$

$$b_t = \beta^*\,(l_t - l_{t-1}) + (1-\beta^*)\,b_{t-1} \tag{2.7}$$

$$s_t = \gamma(y_t - l_{t-1} - b_{t-1}) + (1-\gamma)\,s_{t-m} \tag{2.8}$$

and

l_t is the level smoothing equation, α = the corresponding smoothing parameter
b_t is the trend smoothing equation, β is the corresponding smoothing parameter
s_t is the seasonal component smoothing equation, γ is the corresponding smoothing parameter.
m is the frequency of seasonality (e.g. for monthly data m=12, for quarterly data m = 4)
k is the integer part of $(h-1)/m$, guarantees that the estimates of the seasonal indices use for forecasting come from the last year in the sample

The following is the multiplicative component form:

$$\hat{Y}_{t+h|t} = (l_t + hb_t)\,s_{t+h-m(k+1)} \tag{2.9}$$

Where:

$$l_t = \alpha \frac{y_t}{s_{t-m}} + (1-\alpha)(l_{t-1} + b_{t-1}) \tag{2.10}$$

$$b_t = \beta^*(l_t - l_{t-1}) + (1-\beta^*) b_{t-1} \tag{2.11}$$

$$s_t = \gamma \frac{y_t}{(l_{t-1} + b_{t-1})} + (1-\gamma) s_{t-m} \tag{2.12}$$

Veiga *et al.* (2016) [14] compared the performance between ARIMA and HW's models for the prediction of sales of a group of perishable dairy products (the dairy products weren't specified in the study), using historical sales data, in the period from 2005 to 2013, composed of 50 stock keeping units (SKUs). The study concluded that the ARIMA model performs well in terms of accuracy as well as the simple Holt-Winters. Their performance was measured using the Mean Absolute Percentage Error (MAPE) and Theil's inequality index (Theil's U value). They remarked that the forecast horizon of the predictions made with the HW's model shouldn't exceed the seasonal cycle of the series, as predictions with a larger forecast horizon generally have reduced accuracy when forecasting sales, due to mounting uncertainties. A limitation of this study is the lack of performance measures, relying solely on MAPE and Theil's U. The predictions produced by both models as well as the actual values they were attempting to predict are illustrated in Figure 2.1

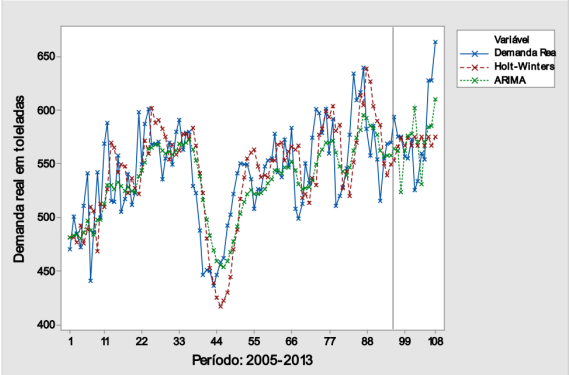

Figure 2.1: Actual Demand sales actual and forecast values by ARIMA and Holt Winter's (HW) [14]

In 2015 Ramos *et al.* [42] compared the forecasting performance of state-space models, namely Error Trend and Seasonality (ETS) model against a traditional model, ARIMA, using data containing the monthly sales of five categories of footwear including flats, sandals, booties, shoes, and boots over a period of 5 years resulting in 64 time points. Time points, of the form (t, y_t) are what describe a time series, t being the time and y_t the predicted value for that time. The performance measures used included MAPE, Mean Absolute Error (MAE), and Root Mean Square Error (RMSE), and the results show that the accuracy of predictions made for the testing data (out-of-sample) of the ETS and ARIMA models was similar. They found that, in general, ARIMA models don't produce more accurate forecasts than ETS models when predicting retail sales, and neither is the most appropriate for all scenarios, in fact, they are complementary approaches, since the ETS model is based on a description of trend and seasonality in the data, while the ARIMA

model attempts to describe auto-correlations present in the data. They also concluded that ARIMA tends to fit the data better (this doesn't necessarily mean it produces more accurate forecasts).

2.2 Demand Influencing Factors as External Variables

To overcome the limitations of the traditional models, a forecasting model that considers uncertainty and incorporates demand influencing factors as external variables in forecasts, (e.g. holidays, festivals, price reductions and weather during the promotion) is required [12]. It must also be mentioned that it is impossible to measure demand directly in retail stores since no orders are placed and the customers simply buy what is available on the shelves at the time. As a result of this, the actual demand is assumed to be the actual volume of sales. However, in the occurrence of a stock-out, the actual demand will be underestimated [13]. In addition to this, the price variations and weather may affect regular demand patterns of customers, and the demand peaks may be a consequence of promotions and holidays [13]. The following is an overview of factors that can affect expected demand and consequences they might generate [53]:

- Product presentation and displays - The presentation of the product (apparent quality, packaging), and the effects of displays have been thoroughly researched in the marketing literature. A general conclusion is that the presence of displays can increase sales significantly [3].
- Price reductions, variations and other promotional strategies - A price reduction may encourage customers to purchase more and increase volatility in demand, it can also be planned or unplanned. Warehousing and waste are two major consequences of an improperly planned price reduction (Armstrong, 2001 [10]). Customers demand patterns are affected by price variations, which may be caused by changes in the market.
- Holidays - Changes in demand as a result of festivals and holidays are expected and rely on cultural habits, location, demography, and the religion of the customers. Instances of retailers whose stores are close to touristic locations may experience higher demand variability during the festival and vacation season due to visiting tourists.
- Weather - The customers' purchase behavior can be affected by the weather (very hot or cold, snowfall, rainfall). In extreme cases (e.g. heavy rainfall, snowfall) customers are compelled to stay home or visit retail stores that are close to their residence. The quality of the weather forecast should also be taken into consideration;
- cannibalization and complementary - Customer demand patterns on a given product are affected by the introduction of new products or the application of promotions to related products. This effect can occur in the same store (in-store) or between a store and a nearby competitor (between-stores).
- Seasonal demand patterns and trend.

The forecast accuracy depends on the following characteristics of the data (time series) being forecast:

- Data quality: A forecasting model is largely reliant on the quality of its data. Both in-sample, the data in the training and validation sets (data used for fitting the model and hyper-parameter tuning respectively), and out-sample, the data in the testing set [13]. The quality of forecasted data like the weather is also crucial.
- Data availability: To understand the external factors affecting sales, longer and complete historical data must be available [13].
- Forecast horizon: Longer forecast horizons are likely to compound uncertainty which may result in a decrease in accuracy.

In general the inaccuracies of a sales forecast results in one of two problems [2]):

- Under-stocking - Results in stock-outs, reduced confidence in the retailer by the customer, and a worsening of market image.
- Over-stocking - Results in waste, a lack of shelf space, and shrinkage, particularly in products with a short shelf-life. The amount of waste is usually intensified by factors such as the short-life and inferior product quality.

In Figure 2.2, which is an edited version of Arunraj *et al.* (2015) [12] figure, the demand influencing factors are classified into events, weather, price, cannibalization, complementarity seasonality, product characteristics, and the number of customer visits. Our edited version includes complementarity, and groups these factors into 3 categories mentioned by the original authors, namely internal, partially internal and external. The internal forces include the price and the product characteristics. However, substitution and cannibalization effects are only partially internal as they can occur in-store or between-stores, as discussed in the context section 1.1. The external forces include events (regular holidays, festivals, and school vacations), weather (temperature, precipitation), seasonality, and the number of customer visits, which can be from regular customers, irregular customers, or special visitors, they are inherently uncontrollable.

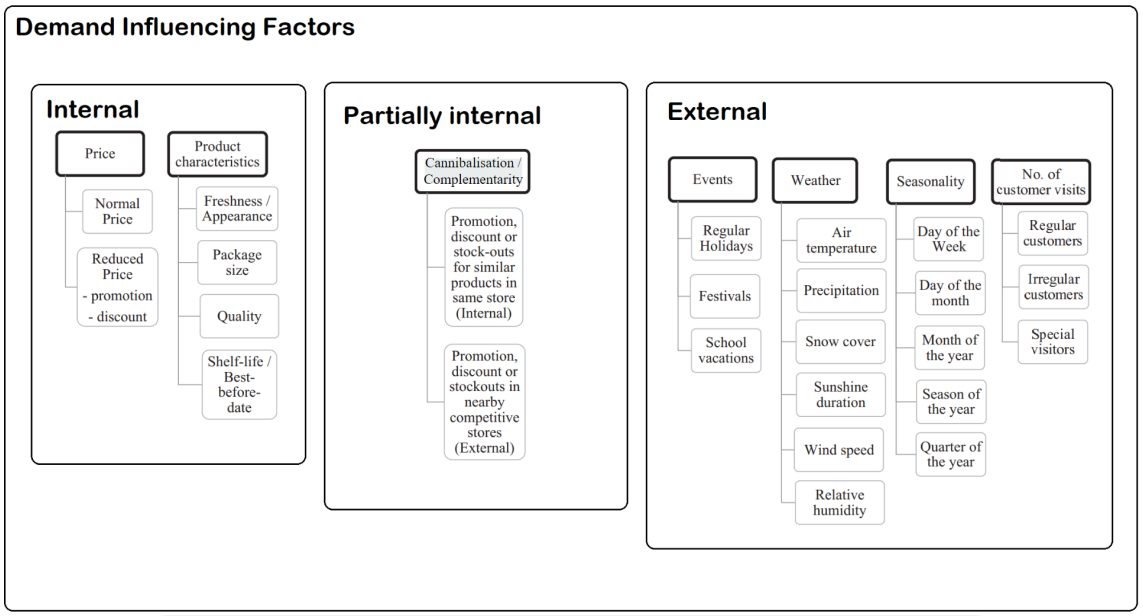

Figure 2.2: Classification and categorization of demand influencing factors.

There are many instances of these factors being considered as demand influencing, in 2010 the demand factors impacting sales on a leading soft drink company in the UK was discussed by Ramanathan and Muyldermans [40]. To understand the demand for different product types they considered holidays, festivals, size of promotion, type of promotion, duration of the promotion, temperature, week-in-year, the rank of the product and more. Additionally, the calendar effect of holidays was used as an external variable in Lee & Hamzah's (2010) [31] forecasting application and in a hybrid ARIMA Artificial Neural Network (ANN) forecasting model, researchers [1] used payment, intermediate payment, holidays, before holidays, festivals, school vacation, price and climate as input neurons. Also, Ali *et al.* (2009) [8] considered price, percentage of discount, and type of promotion in their regression tree forecast approach, Sharma & Sharma (2012) [46] used day-of-the-week, holidays, temperature and sale of alternate products in their ANN approach and Hasin *et al.* (2011) [25] also used recognized holidays, availability, consumption rate, price, promotional activities, climate and brand loyalty as input variables in an ANN based forecasting approach. Different demand influencing factors categorized into location characteristics, promotional variables, weather, national holidays and product characteristics were investigated by researchers [52] [37]. These external variables were analyzed in several linear regression analysis to predict promotional sales and again holidays, festivals, school vacation, weather (temperature, pressure, and rain), and promotions were also used as external prediction features in Zliobaite's *et al.* (2012) [56] intelligent model. Additionally, researchers analyzed the weather sensitivity of the food and drinks sector within the UK retail and distribution industry (Agnew and Thornes 1995) [2]. They discussed the feasibility and advantages of incorporating weather-derived variables in a sales forecasting model. Additional internal and external demand influencing factors were also considered.

- Internal forces - in-store promotions, advertising campaigns, merchandising, price changes, changes in retail outlets, and restructuring of management and operational systems.

- External forces - economic factors, political and legal considerations, technical development, social trend, seasons, holiday periods, and weather variability.

In 2016, Arunraj *et al.* [13]) utilized an extended version of ARIMA, which included seasonality (SARIMA) and another version that also included external variables (SARIMAX). The study analyzed SARIMAX's performance on a data set containing 5 years of daily sales of bananas, measured in kilograms from a typical food retail store.

The SARIMAX model is an extension of the SARIMA model with external variables. SARIMAX (p,d,q) (P,D,Q)s (X), where X is the vector of external variables can be modeled by the following multilinear regression equation:

$$Y_t = \beta_0 + \beta_1 X_{1,t} + \beta_2 X_{2,t} + ... + \beta_k X_{k,t} + \omega_t \tag{2.13}$$

where $X_{1,t}, X_{2,t}, ..., X_{k,t}$ are observations of k number of external variables corresponding to the dependent variable Y_t, $\beta_0, \beta_1, ..., \beta_k$ are regression coefficients of external variables and ω_t is the residual series that is independent of input series (stochastic residual) [13]. The stochastic residual is represented as follows:

$$\omega_t = \frac{\theta_q(B)\Theta_Q(B^S)}{\phi_p(B)\Phi_P(B^S)(1-B)^d(1-B^S)^D}\varepsilon_t \tag{2.14}$$

Substituting this equation for the stochatisc residual in the previous equation for the SARIMAX model gives us the general SARIMAX model equation:

$$Y_t = \beta_0 + \beta_1 X_{1,t} + \beta_2 X_{2,t} + ... + \beta_k X_{k,t} + \frac{\theta_q(B)\Theta_Q(B^S)}{\phi_p(B)\Phi_P(B^S)(1-B)^d(1-B^S)^D}\varepsilon_t \tag{2.15}$$

In the study, the external variables used included:

- Holiday effects, comprised of 4 categories, regular holidays, Christmas, Easter, and School vacations. Some examples of the dummy variables (0 or 1) are, the after holiday effect, the after Christmas effect, and the school vacations effect.

- The month effects are composed of 11 dummy variables (0 or 1) and January is used as the reference month.

- Lastly the price reduction effects are the discounted sales and the percentage of price reduction resulting from the promotion.

The study relied on the MAPE, and RMSE to measure the forecast accuracy, and Theil's U statistic to compare the considered model to a reference naive model. Based on the error measures researchers concluded that including the extra external variables, improves the forecast accuracy

of the SARIMAX model. The R^2 value of the forecasting model was used to assess its fitness. In general, an increase in R^2 indicates fitness improvement. The value of R^2 nearly doubled (from 0.386 to 0.613) when comparing the SARIMA to the SARIMAX model. Additionally, the Theil's U value for the SARIMAX and reference naive model were 0.60 and 1.004 respectively, implying the SARIMAX model produces more accurate forecasts.

In 2015 researchers [12] developed the Seasonal ARIMA using Multiple Linear Regression (SARIMA-MLR) and Seasonal ARIMA using Quantile Regression (SARIMA-QR) models to forecast the daily sales of bananas, measured in kilograms, in a German retail store, from January 2001 to April 2014. The banana was selected because of its short shelf life (between 2 and 4 days) and its availability (year-round). The time series was highly periodic, however, researchers noted that seasonal patterns were hard to observe. They found both models yielded more accurate forecasts for out-sample data (data the model wasn't trained on) when compared to seasonal naive forecasting models, namely SARIMA, and Multilayer Perceptron Neural Network (MLPNN), due to the fact that the SARIMA-MLR and QR models also consider demand influencing effects, such as promotional effect and discount effect. The promotions were split into two types, planned and unplanned. A planned promotion is announced in an advertisement, while the unplanned promotion is done on any day of the week, depending on the price of a related product from a competitor. However, in the model itself, the promotional effect variable is merely a percentage of price reduction. Researchers also considered the discount effect, which in this case is an unplanned price drop on a banana when its quality begins to deteriorate (after 2-3 days of shelf life). Using MAPE and RMSE as performance measures, they concluded that the MLPNN model produces less reliable and accurate predictions when the training data is insufficient, in comparison to traditional time series models. However, when complete data is available (inclusion of several external variables), the MLPNN model is more suitable to explain non-linear relationships between variables, unlike the SARIMA-MLR model that assumes linear relationships between variables. One of the main limitations of the SARIMA-MLR and QR models is that they only produce point forecasts, that is, the mean forecast, and given that the actual distribution of sales isn't a normal distribution, the estimation of prediction intervals from the extrapolation of higher quantiles from the mean forecast won't reflect reality [12].

To better understand the influence of temporary price discounts on the sales of perishable items, Donselar *et al.* (2016) [54] analyzed the potential threshold and saturation levels for the relative price discount of four perishable product categories, namely desserts, dairy drinks, cold cuts, and salads, using a large data set from a retailer encompassing over 100 stores. After which, a regression model was used to forecast the sales of those products. In total, the data contained 1447 promotions in 86 weeks for 407 perishable items. The data included sales quantities, prices, product information, weight or volume per consumer unit, the timing of promotions, additional promotion actions, and information whether the product was on display in the store or included in the store flyer. Researchers aggregated the data on the national level, that is, summed across all stores. Sales data was used instead of demand data, this is because the retailer carried safety stock to prevent out-of-stocks, The promotions ran simultaneously in all stores, generally for one

week. The sales data was also weekly, from the second week in 2010 until week 35 in 2011. Each sales week covers one promotion week. They found that desserts sell on average 14 times more during a week when they are being promoted, even though their average Shelf Life is only 1.5 weeks. They argue this result suggests that the success of promotion is more determined by the substitution effect (consumers swapping between different products due to lower price and other factors) than by the restriction to stockpile caused by the short shelf-life. They also noted that the weight or volume of the product was only relevant in the models for Cold Cuts and Salads, while Shelf Life was only significant in the models for desserts and dairy drinks. They suggest the addition of the interaction between these variables in the regression model may be used to further improve its accuracy. however,is important to note that the use of more detailed input data is advantageous only if more complex and advanced techniques are used, as shown by Gur Ali *et. al* (2009) [7], who noted that incorporating a linear regression model with many variables did not add any benefits. However, when employing a machine learning model the additional variables result in significant improvements to the forecasting accuracy.

In the real world, there are a plethora of factors that can be considered demand influencing, and we must consider that many of them are not readily available in the data, and present challenges in their measurement and handling. Thus in our work, we will focus on a particular group of factors, including different promotional strategies and communicational methods, such as price reductions and promotional displays respectively. Additionally, google analytics data, weather effects, competition data, as well as other factors that are obtainable from the data, such as the variables derived from dates will also be included.

2.3 Machine Learning models

Previous work indicates that classical sales forecasting models based on time series were. in some cases, unreliable at predicting aggregate retail sales, they identified signs of volatility and non-linearity in the market level retail sales data. As such, researchers have turned to non-linear models, particularly artificial neural networks (ANN) (Alon, *et al.* 2001 [9]), however,healthcare and wellness products sales forecasting isn't a widely covered topic within the machine learning literature [51].

In 2013 a study [17] was published on the use of machine learning with the purpose of forecasting the sales of a particular kind of pasta from a popular brand, for two stores, under promotions and its comparison to the statistical models. One of the stores considered had bad management which resulted in frequent stock-outs while the second store had good management and stock-outs were rare, this is important as it more accurately simulates a realistic environment. The data used for training and validation was from the years 2007 and 2008, while data from the year 2009 was used for forecasting and testing. The days in which the store was closed were eliminated and the year used for testing, 2009, was divided into 10 intervals of equal length. Store #1 had 10 intervals, 36 days each, and store #2 also has 10 intervals, but of 32 days each. 9 calendar attributes were used as input attributes linked to the specific day in which the output is given, month, day

of the month, day of the week. These attributes account for predictable human behavior (e.g. higher demand on a Saturday). They also used 4 problem-specific attributes, a Boolean which represents whether the product is being promoted or not, the number of hours that the store is open that day, the daily price of the product, and the total number of receipts released that day. It's important to note that the number of receipts released for a given day that the forecast was done is an unknown value, researchers used a Support Vector Machine (SVM) to forecast the number of receipts per day. This is an important point, as using a forecasted value as an attribute can decrease forecasting accuracy (depending on the accuracy of the forecasted attribute), however, researchers considered this attribute important enough to include it. The Neural Networks of Radial Basis Functions (RBF), Multilayer Artificial Neural Networks (ANN), and Support Vector Machines (SVM) models were compared against the traditional ARIMA, Exponential Smoothing (ES), and HWES. Results for store #1 show the best mean performance between the machine learning methods was given by a variation of the RBF (RBF4i was the variation of the RBF model which used the total receipts attribute), and between the statistical models it was produced by ARIMA. This was the store with frequent stock-outs, so the error measure, MAPE, presented high values. Yet, in every case, the machine learning models outperformed the statistical ones, resulting in a lower MAPE value. This suggests that learning machine models are superior to statistical methods when supplied with irregular data, in this case, a time series that presented various stock-outs. In the following figure 2.3 the sales forecasts produced by the RBF4i model are plotted as well as the actual sales.

For store #2, out of the machine learning models, the best result was produced by a variation of the SVM (SVM4i) using the total receipts attribute, and out of the statistical methods, the HWES generated the most accurate forecasts. In this situation, for each machine learning model, the version using the total receipts attribute produced the lowest MAPE, which means that the incorporation of the forecasted attribute beneficial. Given that this store did not frequently experience stock-outs, and the related time series were thus more regular, the statistical methods compared better against the machine learning methods. In the following figure, the sales forecasts produced by the SVM4i model are plotted as well as the actual sales.

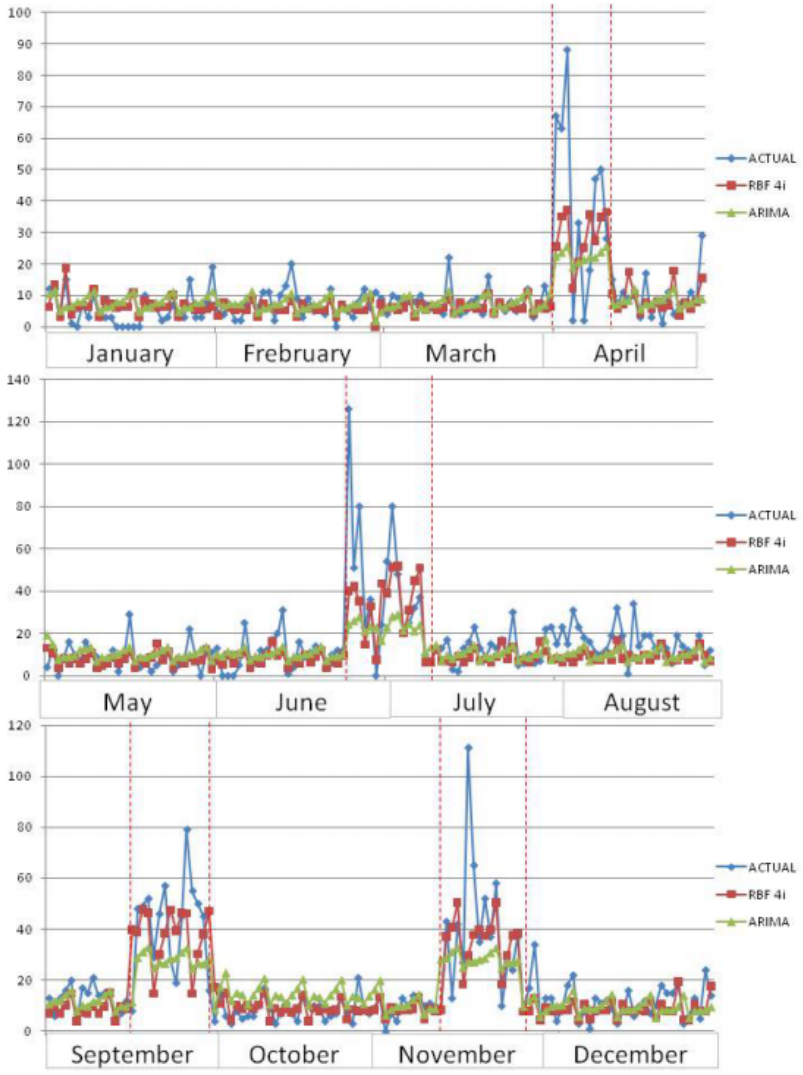

Figure 2.3: Predicted and actual sales values in store #1 during 2009 [17]

It would be interesting if different food products were included in the data, along with use of more error measures to more accurately assess the models performance. Again, cannibalization and complementarity effects were not considered.

In 2001, Alon *et al.* [9] found that given highly volatile economic conditions, resulting in rapid changes to the economic conditions, Artificial Neural Networks provided more accurate forecasts than the linear methods.

Veiga *et al.* [55] published a paper on the applicability of neural networks for food products retail forecasting, using a data set containing 63 monthly sales of liquid dairy products in 3 product groups, yogurt, fermented milk, and milk dessert, spanning 108 months, aggregated at the national level. These product groups accounted for 85% of the retailer's total sales. Using MAPE and Theil's as performance measures, they found that Wavelets Neural Networks provided the most accurate forecasts. Despite this, they discussed that its application required in-house experts, special software, and considerable computing time.

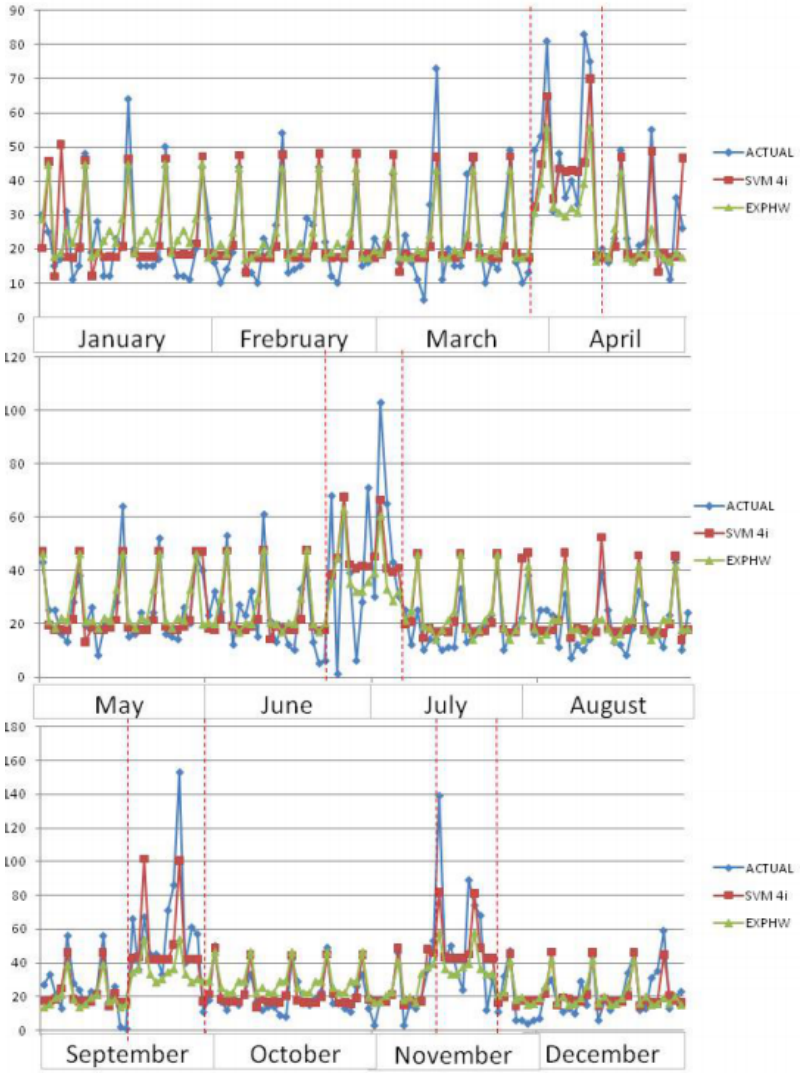

Figure 2.4: Predicted and actual sales values in store #2 during 2009 [17]

Loureiro *et al.* in 2018 [32] compared the sales predictions for the fashion retail market given by the deep learning approach with those given by a set of shallow techniques, such as Random Forest, Support Vector Regression, Decision Trees, Linear Regression and Artificial Neural Networks. The deep learning model generated accurate forecasts, but it did not significantly outperform some of the shallow techniques, such as Random Forest. It should be mentioned that the fashion and food retail market are very different, and abide by different rules, hence even though the results aren't significantly favorable for the deep learning approach, they cannot be generalized, and we expect a different outcome for our approach.

More recently, a study from Priyadarshi *et al.* (2019) [39] examined how the application of deep-learning techniques, such as Long Short-Term Memory (LSTM) networks, to forecasting, can overcome many of the limitations of the commonly used traditional approaches, like ARIMA, or the more simple Machine Learning methods like Random Forest. They used the daily sales data of vegetables, collected from a retail store. The vegetables were divided into 3 groups based

on their shelf-life, low, moderate, and long, in this case, tomatoes, onions, and potatoes respectively. The data was arranged in day-wise format for 22 weeks. This was done to minimize the effect of weekly seasonality on sales forecasts. It was then normalized using the min-max scalar. Results showed that long short-term memory (LSTM) networks and Support Vector Regression (SVR) models improved the accuracy of the sales forecasts, resulting in less waste of daily retail inventory and fresh produce, thereby increasing the daily revenue. However, the results obtained shouldn't be generalized as they were obtained from only one store. This study did not incorporate external variables in the models and used only the sales data. This research seems to indicate that deep learning models outperform the state of the art methods when external variables such as promotions and are supplied, although external variables like promotion effects or cannibalization weren't incorporated in the models, thus there is currently no research that compares the standard machine learning models with the novel deep learning models when both incorporate these variables.

2.4 Prophet

Prophet is a machine learning model for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects [20]. It works best with time series that have strong seasonal effects and multiple seasons of historical data [50]. It uses a decomposable time series model (Harvey & Peters 1990 [24]) with three main model components: trend, seasonality and holidays. These are combined in the following equation:

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t \tag{2.16}$$

Where g(t) is the trend function that models non-periodic changes in the value of the time series, s(t) represents periodic changes (e.g., weekly and yearly seasonality), and h(t) represents the effects of holidays that occur on potentially irregular schedules over one or more days. The error term ε_t represents any idiosyncratic changes that are not accommodated by the model.

2.5 Deep AR: Probabilistic Forecasting with Auto-Regressive Recurrent Networks

The Deep AR is a forecasting method based on Auto-Regressive Recurrent Neural Networks, that learns a global model from historical data of all the time series in the dataset [43].

The Deep AR approach, instead of fitting separate models for each time series, creates a global model from related time series to handle widely-varying scales through re-scaling and velocity-based sampling. They use an RNN architecture that incorporates a Gaussian/Negative Binomial likelihood to produce probabilistic forecasting. The figure below 2.5 shows the architecture for training and prediction:

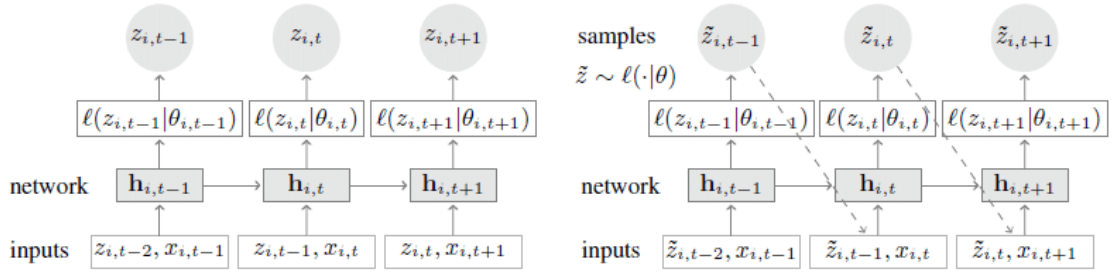

Figure 2.5: Deep AR Model Architecture

On the left, the goal is to predict the following time step. at each time step (forecast horizon of 1). The network must receive in input the previous observation (lag of 1) *z_t* − 1, along with a set of optional covariates *x_i*. The information is propagated to the hidden layer (represented in the figure by h) and up to the likelihood function. During training (the network on the left) the error is calculated using the current parametrization of the likelihood *theta*. This is represented by *mu* and *sigma* in the case of a Gaussian likelihood. This means that while performing backpropagation we are tuning the network parameters (weights w) which change the parametrization of every likelihood, until we converge to optimal values.

On the right, once we have trained the network weights, we can perform forward propagation using input *z_i* − 1 (along with [optional] covariates or encoded categorical features) and obtain distribution parameters mu and sigma. For predictions we start by drawing one sample from the output distribution of the first time step: that sample is the input to the second time step and so on. Every time we start from the beginning and sample up to the prediction horizon we create the equivalent of a Monte Carlo trace, which means that we can calculate the quantiles of the output distribution and assess the uncertainty of the predictions.

The likelihood model can be both Gaussian (with parametrization *mu* and *sigma*):

$$\ell_{\mathrm{G}}(z|\mu,\sigma) = (2\pi\sigma^2)^{-\frac{1}{2}}\exp(-(z-\mu)^2/(2\sigma^2))$$
$$\mu(\mathbf{h}_{i,t}) = \mathbf{w}_\mu^T\mathbf{h}_{i,t} + b_\mu \quad \text{and} \quad \sigma(\mathbf{h}_{i,t}) = \log(1+\exp(\mathbf{w}_\sigma^T\mathbf{h}_{i,t} + b_\sigma))\ .$$

Figure 2.6: Deep AR Gaussian Likelihood Model

or Negative Binomial (when dealing with time series of positive count data):

$$\ell_{\mathrm{NB}}(z|\mu,\alpha) = \frac{\Gamma(z+\frac{1}{\alpha})}{\Gamma(z+1)\Gamma(\frac{1}{\alpha})}\left(\frac{1}{1+\alpha\mu}\right)^{\frac{1}{\alpha}}\left(\frac{\alpha\mu}{1+\alpha\mu}\right)^{z}$$
$$\mu(\mathbf{h}_{i,t}) = \log(1+\exp(\mathbf{w}_\mu^T\mathbf{h}_{i,t} + b_\mu)) \quad \text{and} \quad \alpha(\mathbf{h}_{i,t}) = \log(1+\exp(\mathbf{w}_\alpha^T\mathbf{h}_{i,t} + b_\alpha))\ ,$$

Figure 2.7: Deep AR Negative Binomial Likelihood Model

Additionally, the model trains an embedding vector that learns the common properties of all the time series in the group. Another feature of the model is that Deep AR automatically creates additional feature time series depending on the granularity of the target time series. These can, for

instance, be time-series created for "day of the month" or "day of the year", and allow the model to learn time-dependent patterns. [36]

In a recent study Salinas *et al* 2019 [43] proposed the use of the Deep AR methodology to generate accurate probabilistic forecasts. They incorporated the model with data from related time series, such as the demand for related products, and discussed how this allowed for more complex models to be fitted while avoiding over-fitting, and simultaneously reducing the time and labor-intensive steps of selecting models and selecting and preparing covariates that traditional approaches require. Covariates are represented as $x_{i,t}$, where i is the item and t the time. In the case of the retail demand forecasting data sets, an item's covariate corresponds to a broad product category like electronics (e.g. $x_{i,t} = e$ where e represents electronics), while in the smaller data sets it corresponds to the item's identity, which allows the model to learn item-specific behaviors.

5 data sets were used, however, in this paper we will only address 3 of them, as they relate to the purpose of sales forecasting, while the remainder focused on forecasting electricity consumptions, and traffic (lane occupancy rates). The following is an overview of the considered data sets.

- parts - 1045 time series aligned and spanning 50 months (42 months for training, 8 for testing), containing the monthly sales of several items by an American car company.
- ec - 534884 time series, containing weekly item sales from Amazon.
- ec-sub - 39700 time series, containing weekly item sales from Amazon.

Both the ec and ec-sub data sets included novel products that were only introduced in the weeks before the forecast time. The Croston method [28], an ETS method by Hyndman *et al.*(2008) [27], the negative-binomial auto-regressive method of Snyder *et al.* (2012) [48] and the method of Seeger *et al.* (2016) [45] using an innovative state space model with covariate features (ISSM), along with 2 baseline RNN models were used for comparison, using the $p-$ risk metric (quantile loss). They asserted that their model provides a better forecast accuracy than previous methods. In addition to this, they stated the following advantages in comparison to the traditional methods:

- This approach produces probabilistic forecasts in the form of Monte Carlo samples. These samples can be used to compute consistent quantile estimates for all the sub-ranges in the forecast horizon.
- In cases where an item has minimal sales history available, it is still able to produce forecasts, while classical single-item forecasting methods fail.
- The model can include a wide range of likelihood functions, chosen by the user depending on the properties of the data.
- The model is able to learn seasonal behaviors and the effects on the given covariates across time series. Additionally, it captures complex group dependent behaviors with minimal manual intervention.

The proposed Deep AR model was able to effectively learn a global model from related time series, to handle varying scales through velocity-based sampling and re-scaling, to generate highly accurate forecasts, and to learn complex patterns such as uncertainty growth and seasonality over time. However, we note that the data used for forecasting wasn't related to healthcare and wellness products or daily sales, and it did not include demand influencing factors such as promotional methods and different promotion types.

2.6 Evaluation Measures

2.6.0.1 Absolute error based measures

For each product, the error for the forecast horizon in the instant t given by:

$$e_t = Y_t - \hat{Y}_t \tag{2.17}$$

Where Y_t is the actual value in the instant t and $\hat{Y}_t$ the predicted value in the instant t. Resulting in the following measures:

- Mean Absolute Error

$$MAE = mean(|e_t|) \tag{2.18}$$

- Median Absolute Error

$$MdAE = median\,(e_t) \tag{2.19}$$

- Mean Square Error

$$MSE = mean\left(e_t^2\right) \tag{2.20}$$

- Root Mean Square Error

$$RMSE = \sqrt{mean\left(e_t^2\right)} \tag{2.21}$$

The accuracy of these measures is dependent on the scale of the data [29]. Generally RMSE is preferred to the MSE as it is on the same scale as the data. RMSE and MSE have been popular, because of their theoretical relevance in statistical modelling. However, they are more sensitive to outliers than MAE or MdAE [29]. MAE is often used due to the ease of interpretation [38]. Also, these error measures are the most popular in various domains [47]. They have the following limitations (Shcherbakov *et al.* 2013 [47]):

- RMSE and MSE have a low reliability: results can be different depending on different fractions of data.
- Outliers have a high influence. These measures provide conservative values when data contains a maximal value outlier, which is common in real world tasks.
- They are dependant on scale. These measures can't be applied if the forecast contains objects with different scales or magnitudes.

2.6.0.2 Percentage Error based measures

Percentage error based measures are not scale dependent, and therefor are frequently used to compare forecasting performance across different data sets (with varying scales) [29]. The percentage error is given by:

$$p_t = 100\frac{e_t}{Y} \tag{2.22}$$

Resulting in the following measures:

- Mean Absolute Percentage Error

$$MAPE = mean(p_t) \tag{2.23}$$

- Root Mean Square Percentage Error

$$RMSPE = \sqrt{mean(p_t)^2} \tag{2.24}$$

- Median Absolute Percentage Error

$$MdAPE = median(p_t) \tag{2.25}$$

- Root Median Square Percentage Error

$$RMdSPE = \sqrt{median(p_t)^2} \tag{2.26}$$

These measures have the disadvantage of being infinite or undefined if $Y_t = 0$ for any t in the period of interest, and having an extremely skewed distribution when any value of Y_t is close to zero. Where the data involves small counts it is impossible to use these measures as zero values of Y_t occur frequently. MAPE is commonly used and as stated doesn't depend on scale, however,it does under-perform in cases where promotional activities took place (see Pinho, 2015 [38]). Thus, the use of percentage error measures isn't recommended for products with very few sales (Davydenko and Fildes, 2013 [15]). Their limitations can be summarized as [47]:

- When the actual value is equal to zero, results in an infinite value (division by zero),
- Non-symmetry. Depending on if the predicted value is bigger or smaller than the actual value the error values differ.
- Outliers significantly impact the result, particularly if outlier has a value much greater then the maximal value of the "normal" cases.
- May lead to an incorrect evaluation of the the forecasts performance since the error measures are biased.

2.6.0.3 Scaled error based measures

Hyndman and Koehler [29] proposed a new but related idea, by scaling the error based on the in-sample MAE from the naive (random walk) forecast method. Thus the scaled error is defined as:

$$q_t = \frac{|e_t|}{\frac{1}{m-1}\sum_{t=2}^{m}|Y_t - Y_{t-1}|} \tag{2.27}$$

A scaled error is less than one if the forecast is better than the average one-step naive forecast computed in-sample. It is larger than one if the forecast is worse than the average one-step naive forecast computed in-sample. In this case (/m)/ is the frequency of seasonality.

The Mean Absolute Scaled Error (MASE) is defined as:

$$MASE = mean\left(q_t\right) \tag{2.28}$$

The Mean Absolute Scaled Error (MASE) proposed by Hyndman and Koehler (2006) [29] is defined as the mean of the errors divided by the mean error that would be made using the naive (random walk) forecast. The naive method simply forecasts the same number of units observed in the last time period for the particular SKU-store. Analogously the Median Absolute Scaled Error (MdASE) and Root Mean Squared Scaled Error (RMSSE) are defined as:

$$MdASE = median\left(q_t\right) \tag{2.29}$$

$$RMSSE = \sqrt{mean\left(q_t^2\right)} \tag{2.30}$$

If all the time series in the data are on the same scale, the MAE can be used as it is easily explainable. If all the time series contain positive values much greater than zero, the MAPE can be used due to its simplicity. Although, in the instance of very different scales. including data close to zero or negative, the MASE is the best option according to Hyndman (2006) [29]. These measures are symmetrical and resistant to outliers. however,Shcherbakov *et al.* [47] considered 2 drawbacks:

- Results in an infinite value if the forecast horizon real values are equal to each other (division by zero).
- Weak bias estimates can be observed.

2.6.0.4 Selecting the appropriate measures

In 2013 [47] the following guidelines were proposed by Shcherbakov *et al.* for choosing the appropriate error measures. They based their guidelines on a systematic review regarding the available literature on error measurements in forecasting. [11] [34]

- If the forecast performance is evaluated for time series with a similar scale and the data was preprocessed, (detecting anomalies and data cleaning), it is appropriate to use RMSE, MAE, or MdAE.

- Percentage errors are commonly used, however, due to non-symmetry, they are not advised.
- Scaled measures, such as MASE, should be used if the data contains outliers. For this the following conditions should be met:
  - The forecast horizon should be large enough;
  - There should be no identical values;
  - The normalized factor should not be equal to zero.

2.7 Overview of reviewed literature

In this section we provide an overview of the focus, data, variables, evaluation measures and baseline methods used along with the limitations of the referenced literature regarding sales forecasting models in tables 2.1, 2.2, 2.3.

Table 2.1: Comparison between different forecasting approaches 1

Reference	Focus	Data	Variables*[a]	Evaluation measures	Baseline Methods	Issues and Limitations
Ramos *et al.* (2015) [42]	Comparing the forecasting performance of state space models and ARIMA models.	Monthly sales of five categories of women footwear, from January 2007 to April 2012 (64 observations).	Seasonality and trend.	RMSE, MAE and MAPE.	ETS and ARIMA.	Limitations of the MAPE, doesn't consider promotional effects and cross-effects.
Veiga *et al.* (2014) [14]	Comparing the performances between ARIMA and Holt-Winters models for the prediction of a time series formed by a group of perishable dairy products.	Historical demand data from a group of dairy products with short life cycle , in the period from 2005 to 2013, composed of 50 SKUS (StockKeepingUnits)	Seasonality and trend	MAPE, Theil	ARIMA, HW	Lack of performance measures.
Arunraj and Ahrens (2015) [12]	Comparing the performance of ARIMA based methods, SARIMA-MLR and SARIMA-QR, using seasonality and demand influencing factors interval for forecasting perishable food daily sales	5 years of daily sales of banana measured in kilograms from a typical food retail store	Promotion, discount, weather, weekday, month, holidays	MAPE, RMSE	SARIMA, SARIMA-MLR, SARIMA-QR, MLPNN and a Seasonal naive forecasting model	The SARIMA-MLR and QR models only produce point forecasts.
Arunraj and Ahrens (2016) [13]	SARIMAX is proposed to overcome the disadvantage of the traditional SARIMA model, in forecasting the daily sales of fresh foods in a retail store.	5 years of daily sales of banana measured in kilograms from a typical food retail store	Promotion, discount, weather, weekday, month, holidays	R2, MAPE and RMSE	SARIMA and SARIMAX	Weather effects, cannibalization and complementarity were not considered

[a] *In the variables column the value of previous sales volumes (lagged values) is implied for every case

Table 2.2: Comparison between different forecasting approaches 2

Reference	Focus	Data	Variables*[a]	Evaluation measures	Baseline Methods	Issues and Limitations
Pillo *et al.* (2013) [17]	An assessment of the use of Learning Machines for sales forecasting under promotions, and comparing it with the statistical methods.	The daily sales of a particular kind of pasta of a popular brand covering three years, 2007, 2008 and 2009.	Daily sales, 9 calendar attributes, month, day of the month and day of the week, 4 problem specific attributes, is there a promotion on the product, number of hours the store is open that day, the daily price of the product, the overall number of receipts released that day in the store and an attribute that indicates if in that day are expected high or low sales.	MAPE	Multilayer artificial neural networks (ANN), Neural networks of radial basis functions (RBF) and Support vector machines (SVM), ARIMA, Exponential Smoothing (ES), HWES.	Lacking in performance measures.
Donselaar *et al.* (2016) [54]	Analysing the potential threshold and saturation levels for the relative price discount of four perishable product categories, and using a regression model to forecast their sales.	Four perishable product categories: Desserts, Dairy Drinks, Cold Cuts and Salads, using a large empirical data set from a retailer operating more than 100 stores. The sales data is weekly (from week 2 in 2010 till week 35 in 2011) aggregated on national level, for 407 different items.	Sales quantities, prices, product information, weight or volume per consumer unit, timing of promotions, additional promotion action, whether the product was on display or in the store flyer, Lift Factor (Sales during a promotion/baseline sales), shelf life, holidays.	RMSE, MAPE and Average Bias.	Regression models.	Lacking in baseline methods and in performance measures.
Gur Ali *et al.* (2009) [8]	Identify methods of increasing complexity and data preparation cost yielding increasing improvements in forecasting accuracy, by varying the forecasting technique, input features and model scope.	The data set consists of 168 store-SKU combinations, and spans a period of 76 weeks.	Promotion type, price and discount, the week-number and the actual unit sales for the current and last four weeks.	MAE.	Stepwise linear regression, support vector regression (SVR), regression trees.	Lacking in evaluation and error measures.
Tsoumakas (2019) [51]	Reviewing existing machine learning approaches for food sales prediction.	Several data sets of food products sales histories.	Time, date, month, quarter of the year, season, brand, packaging information, promotion, price elasticity, expiration date, seasonality.	MSE, RMSE, RRMSE, MAE, MAPE, MASE.	Moving average, RBF network using Fuzzy means algorithm, Regression, SVM's	Did not consider different promotional types, nor cannibalization and complementarity effects.

[a] *In the variables column the value of previous sales volumes (lagged values) is implied for every case

Table 2.3: Comparison between different forecasting approaches 3

Reference	Focus	Data	Variables*[a]	Evaluation measures	Baseline Methods	Issues and Limitations
Pillo *et al.* (2016) [16]	Application of learning machines for sales forecasting under promotions	2 brands of pasta from two different retail stores, 3 years of daily sales	Calendar attributes, promotion, open hours, price, overall number of receipts	MAPE	ARIMA, ES, HWES, SVM.	Lacking in performance measures. Did not compare considered model to alternative state-of-the-art models.
Veiga *et al.* [55] (2016)	Applicability of natural computing approaches in foodstuff retail demand forecasting	63 liquid dairy products monthly sales in 3 product group, 108 months, aggregated at national level	Seasonality and trend	MAPE and Theil's U	ARIMA, HW, Takagi-Sugeno (TS), Wavelet Neural Network (WNN).	WNN has a very good performance when forecasting by using very accurate approximation properties but it is computationally difficult , requires special software, considerable computing time and in-house expertise.
Priyadarshi *et al.* (2019) [39]	Comparing the application of deep learning techniques to classical approaches, for the purpose of sales forecasting.	The daily sales data for three vegetables tomato, potato and onion, collected from a retail store.	Seasonality and trend.	RMSE, MAD, MAPE, BIAS, Tracking Signal (TS).	ARIMA, long short-term memory (LSTM) networks, support vector regression (SVR), random forest regression, gradient boosting regression (GBR) and extreme GBR (XGBoost/XGBR)	Demand influencing factors were not considered..
Salinas *et al.* (2019) [43]	Selecting the appropriate forecasting model at the retail stage for several groups of vegetables on the basis of performance analysis.	Parts by a US automobile company, spanning 50 months (1046 time series) hourly electricity consumptions, occupancy rates (traffic) and weekly item sales from Amazon (500,000 time series).	Seasonality, trend and the distance to the first observation in that time series (age).	$p-risk$, Theil's U, ND, NRMSE.	ISSM [45], Croston [28], Snyder [48], ETS [27]	No data regarding perishable food items sales. Did not consider promotional methods and types and various other demand influencing factors.

[a] *In the variables column the value of previous sales volumes (lagged values) is implied for every case

Chapter 3

Methodology and Preliminary Data Analysis

This chapter provides an overview of the methods used to conduct our experiments. The first section 3.1, details some characteristics and features of each model, and how their parameters were set. The two following sections, 3.2 and 3.3, give an in-depth description of each data set, including the available features as well as statistical analysis.

3.1 Models

3.1.1 Holt Winter's

We used the additive Holt Winter's implementation by statsmodel, a Python module that provides classes and functions for the estimation of many different statistical models, as well as for conducting statistical tests, and statistical data exploration. A seasonality period of 7 days, and 365 days for the Rossmann and Retail data sets respectively.

3.1.2 SARIMAX

The statsmodel SARIMAX model takes into account the parameters for the regular ARIMA model (p,d,q), as well as the seasonal ARIMA model (P,D,Q,s). These sets of parameters are arguments in our model called the order and the seasonal order, respectively. In order to find the best parameters we performed a grid search, generating all possible triplets of p,d,q, and seasonal P, D and Q, in a range from 0 to 2, and then selecting the best parameters ranked by the Akaike Information Criterion (AIC). Additionally, the SARIMAX model also includes the parameter (X), the vector of external variables. All external variables with the exception of static features were incorporated for training.

3.1.3 Prophet

The same exogenous variables were used in the Prophet model. Prophet includes functionality for time series cross-validation to measure forecast error using historical data. This is done by selecting cutoff points in the history, and for each of them fitting the model using data only up to that cutoff point. We can then compare the forecasted values to the actual values. This functionality was employed to find the optimal values for the 'changepoint prior scale' and 'seasonality prior scale' parameters, which were selected based on the lowest RMSE. Daily, weekly and yearly seasonality were set for the Rossmann data set, and yearly seasonality was set for the Retail data set.

3.1.4 Deep AR

Gluon TS's Deep AR model allows us to select a variety of likelihood functions that can be adapted to the statistical properties of the data allowing for data flexibility. We selected the Negative Binomial Distribution, which is a common choice for modeling time series of positive count data, such as daily sales, provided by the Gluon TS package [43].

Unlike the previous models which only included dynamic features, the Deep AR also takes into account static categorical features. Dynamic features are features that vary across time for a given time series, such as promotional and weather variables. On the other hand, static features are constant throughout a given time series, in this case, an example would be the State the store is located in. This allows the Deep AR model to group time series that have the same static features, improving forecasting accuracy. Additionally, a categorical feature is represented by a limited set of categories, and might not be a numerical value. As an example, Rain, Wind, and Snow are different categories of the weather variable Event. In this case, the feature must be encoded with a numerical value.

In Deep AR's case, we don't have to apply encoding to the static categorical features because the model applies entity embedding, an encoding method, and allows us to set the dimension of the embeddings for categorical features. We used the default value $[min(50, (cat + 1)//2)$ for cat in cardinality]). Entity embedding is especially useful when data sets contain features with high cardinality. It's also been demonstrated that the embeddings obtained from the trained neural network boost the performance of machine learning methods considerably when used as the input features. [22]

The Deep AR model expects a fair number of hyperparameters to be set, including the context length for training, number of epochs, dropout rate, and learning rate among others. In both experiments, a prediction length of 48 days and a context length of 96 days were set. The remaining parameters were left with the default values and changed during validation.

Additionally, the Deep AR model was the only considered model that used multiple time series as input to generate a global model. While in the case of the SARIMAX, Prophet and Holt Winter's approach, a model had to be trained for each time series. The global model for the Rossmann data

set was produced from 932 time series, and the global model for the Retail data set was generated from 2000 time series.

3.2 Rossmann Drug Stores Dataset

The following data was gathered from a Kaggle competition where the goal was to forecast the sales values (€) for the last 48 time steps (days) in the data. It contains the historical sales data for 1,115 Rossmann stores and supplemental information such as weather, promotional effects, competition-related features, and google trend related features, over a period of 990 days. Rossmann operates over 3,000 drug stores in 7 European countries. Currently, Rossmann store managers are tasked with predicting their daily sales for up to six weeks in advance. Store sales are influenced by many factors, including promotions, competition, school and state holidays, seasonality, and locality.

3.2.1 Data Description

The following data was used in all models that allowed for exogenous variables, the Deep AR, Prophet, and SARIMAX models namely.

- **Promotional features**
  - **Promo** - indicates whether a store is running a promo on that day
  - **Promo2** - Promo2 is a continuing and consecutive promotion for some stores: 0 = store is not participating, 1 = store is participating
  - **Promo2** Since Year/Week - describes the year and calendar week when the store started participating in Promo2
  - **After Promo** - gives the days since a school Promo
  - **Before Promo** - gives the days until another school Promo
  - **Promo2 Weeks** - describes how many weeks Promo2 has been ongoing
  - **Promo2 Days** - describes how many days Promo2 has been ongoing
  - **Promo Interval** - describes the consecutive intervals Promo2 is started, naming the months the promotion is started anew. E.g. "Feb,May,Aug,Nov" means each round starts in February, May, August, November of any given year for that store
  - **Promo fw** - the sum of occurrences of Promo in the following 7 day period
  - **Promo bw** - the sum of occurrences of Promo in the previous 7 day period
- **Time features**
  - **Day** - gives the day of the month
  - **Week** - gives the week of the year

  - **Month** - gives the month of the year
  - **Year** - gives the year
  - **Day of week** - gives the day of the week
  - **Day of year** - gives the day of the year
  - **Is month end** - indicates if it's the last day of the month
  - **Is month start** - indicates if it's the first day of the month
  - **Is quarter end** - indicates if it's the last day of the quarter
  - **Is quarter start** - indicates if it's the first day of the quarter
  - **Is year end** - indicates if it's the last day of the year
  - **Is year start** - indicates if it's the first day of the year

- **Holiday Features**
  - **State Holiday** - indicates a state holiday
  - **School Holiday** - indicates if the (Store, Date) was affected by the closure of public schools
  - **After School Holiday** - gives the days since seeing a school holiday
  - **Before School Holiday** - gives the days untill another school holiday
  - **After State Holiday** - gives the days since seeing a state holiday
  - **Before State Holiday** - gives the days untill another state holiday
  - **State Holiday fw** - the sum of occurrences of state holidays in the following 7 day period
  - **State Holiday bw** - the sum of occurrences of state holidays in the previous 7 day period
  - **School Holiday fw** - the sum of occurrences of school holidays in the following 7 day period
  - **School Holiday bw** - the sum of occurrences of school holidays in the previous 7 day period

- **Competition Features**
  - **Competition Distance** - distance in meters to the nearest competitor store
  - **Competition Open Since Month/Year** - gives the approximate year and month of the time the nearest competitor was opened
  - **Competition Months Open** - gives how many months the competition has been open for
  - **Competition Days Open** - gives how many days the competition has been open for

- **Google Trend Features**
  - **Trend** - Google trend value for a given state
  - **Trend DE** - Google trend value for the whole of Germany
- **Weather Features**
  - **Max/Mean/Min Temperature** - gives the Max/Mean/Min temperature in *Celsius*
  - **Max/Mean/Min Dew Point** - gives the Max/Mean/Min dew point in *Celsius*
  - **Max/Mean/Min Humidity** - gives the percentage Max/Mean/Min humidity
  - **Max/Mean/Min Sea Level Pressure** - gives the Max/Mean/Min Sea Level Pressure in *Pascal*
  - **Max/Mean/Min Wind Speed** - gives the Max/Mean/Min wind speed in *Km/h*
  - **Precipitation** - gives precipitation in *mm*
  - **Wind Direction** - gives wind direction in degrees
  - **Event** - describes weather events that occurred (*e.g. Fog, Fog-Rain, etc.*.)
- **Other**
  - **Open** - indicates if store was open

The models considered for these experiments produce a fitted model for each time series, except for the Deep AR model, which generates a global fitted model. Due to this, it can benefit from using static categorical features. These are features that are constant throughout a given time series but can vary between different time series, such as the store ID, store Type, product Class and Sub-class, and more. The following static categorical features were used exclusively by the Deep AR model:

- **Store ID** = Integer store identifier
- **Store Type** - differentiates between 4 different store models: a, b, c, d
- **Assortment** - describes an assortment level: a = basic, b = extra, c = extended
- **State** - the State in which the store is located
- **Competition Open Since Year** - the year since the Competition has been open

3.2.2 Data Cleaning and Feature Engineering

The following features were added:

- Included only time series which had a length greater than 942 days
- Clipped lower value of target variable (sales) to 0 (removing negative values)

- Created feature **Competition Days Open**, indicating the number of days the competition has been open for and **Competition Months Open**, with a maximum limit of 24 months to limit the number of unique categories.

- Created feature **Promotion 2 Days**, indicating the number of days the competition has been open for. Added feature **Promotion 2 Weeks**, with a maximum limit of 25 weeks to limit the number of unique categories.

- Features such as **Promotion 1** and **School Holiday** are called events. In time series data, it always helps to show how long it's been to and from an event as it helps the neural network pick certain patterns from them, these patterns could be a spike in sales just before a public holiday, a dip in sales before a promotion or how long it takes for sales to rise after a promotion has started, for example [26]. These don't need to be added to the data set as the neural network can learn them but providing this data makes the learning process easier. The features added were:

  - **After School Holiday**
  - **Before School Holiday**
  - **After State Holiday**
  - **Before School Holiday**
  - **After Promo**
  - **Before Promo**

- Another addition were rolling sums figures. A rolling sum figure was generated for the holidays and promotions within a window of 7 days. This gives an idea of the events status with a 7-day period. Again, this is useful information for the neural network that would allow it to pick certain patterns following a certain event. The rolling sum features added were:

  - **School Holiday bw**
  - **School Holiday fw**
  - **State Holiday bw**
  - **State Holiday fw**
  - **Promo bw**
  - **Promo fw**

  We also applied ***One-Hot Encoding*** to the weather feature **Event**, a categorical feature whose values describe an observed weather event, (e.g: 'Fog', 'Rain'), and hence requires encoding to convert it into a numerical value that the model can use. One of the limitations of traditional label encoding (e.g. mapping an integer to each different value the variable

can take), is that the model can interpret a false relationship between different 'Event' values. The following example demonstrates this: A given variable **Event** can take one of the following three values, '**Fog**', '**Rain**', '**Fog-Rain**', Using traditional label encoding, these values would be converted to '1', '2' and '3' respectively, which can lead to the relationship **Rain > Fog** which is false. ***One-Hot Encoding*** solves this by creating an additional feature (dummies) for each value the variable can take and assigning it a 1 or a 0 (*True* or *False*). In the previous example, we would create 3 additional features, Fog, Rain, and Fog-Rain, and assign them a value of 1 or 0.

A fraction of the total time series contained missing values (roughly 16%), this presented an obstacle because the Deep AR model used in the experiment can only handle missing values for the target variable (in this case, sales). Rather than imputing these values, we opted to use the complete data that we had and ignore the incomplete time series. This resulted in our data describing 932 time series (stores), each with 990 time points (days), with 57 features. The first 942 days were used as training input and the last 48 were used for testing.

3.2.3 Data Analysis

We performed data analysis to better understand the target variable behavior and its relationship with the external variables provided.

Firstly, some statistics regarding the target variable, the volume of daily sales, in euros:

- Mean = 5827.4293475542945 €
- Min = 0 €
- Max = 38722 €
- Median = 5783.0 €
- Standard Deviation = 3899.955241146479 €

The difference between the mean and median values is small, indicating most time series in this data set have similar generating processes (structure).

We analyzed the Customers feature, which gives the number of customers that visited the store on that day, by plotting the daily customer values and their corresponding sales values, illustrated by the following figure 3.1.

The plot displays a clear relationship between these variables, it can be seen that with the increase of the number of costumers, the currency/volume of total sales also increase, naturally. Given this strong relationship and the fact that the number of customers values are real values and not predictions, we didn't include this feature in the models. Its inclusion would have generated similar results as if the models were given the sales values for the testing data. Even though the feature wasn't included, we note that generating accurate predictions for the number of customers on a given day can be valuable data for forecasting models that can use that external variable.

Figure 3.1: Scatter plot of daily sales by number of customers

The following plot in Fig. 3.2 describes how sales vary according to the day of the week when the stores are open (days when stores are closed were ignored). It can be seen that there is a significant degree of variability in the daily sales values. The median daily sales values are the highest on the first and last days of the week, day 1 (Monday) and day 7 (Sunday). If we included the days when the stores were closed, day 7 would present the lowest median, given that most stores are closed on Sunday.

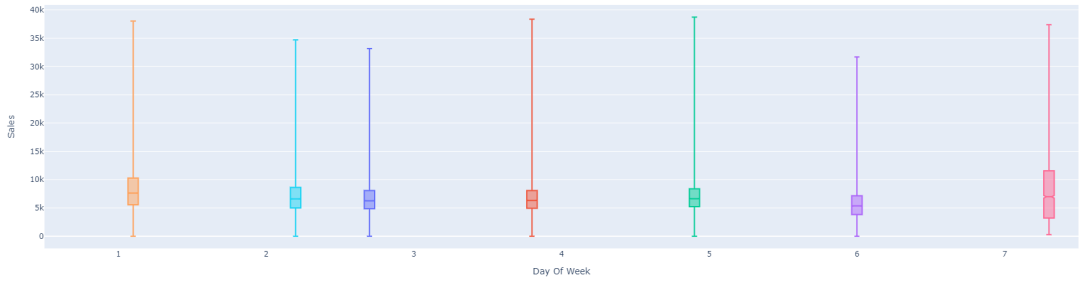

Figure 3.2: Box plots of daily sales for each day of the week

We then analyzed how each promotional feature affected sales. Starting by box plotting the daily sales values in case of promotion (in red) and in case of no promotion (in blue), for promotions 1 and 2, illustrated in figures 3.3 and 3.4. There's a greater median daily sales volume in days where promotion 1 occurs (5400 € vs 7700 €), and all quartiles are larger. However the difference in median daily sales between days with and without promotion 2 is very small (6700 € vs 6100 €), days with no promotion 2 have slightly higher median daily sales.

Figure 3.3: Box plots of daily sales by promotion 1

Figure 3.4: Box plots of daily sales by promotion 2

Next, we plotted the daily sales by the number of days since the last promotion, one of our engineered features, this is illustrated in figure 3.5. It appears that the longer it's been since a promotion, the lower the value of the daily sales. This makes sense since, in general, when promotions occur frequently, they are often advertised and thus increase products or brand exposure. In the absence of promotions, the opposite may occur, with brands and products being forgotten.

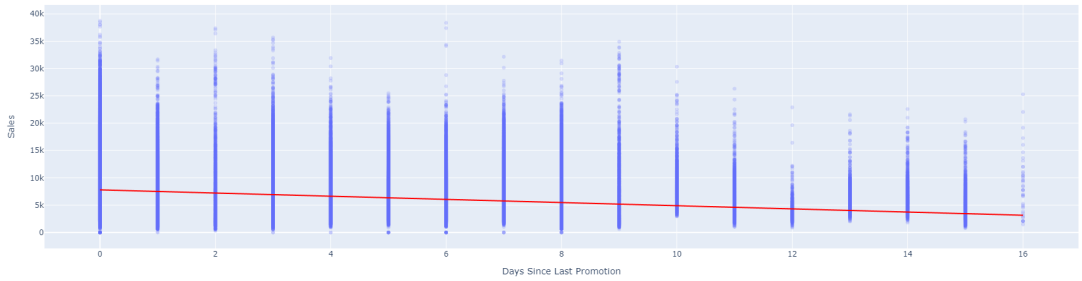

Figure 3.5: Box plot of daily sales by days since last promotion

The following box plots 3.6 illustrate the daily sales depending on if there was a State Holiday. During State Holidays both median and maximum daily sales volumes are higher.

Figure 3.6: Box plots of daily sales by State Holiday

The relationship between the daily sales and the occurrence of School Holidays is not clear, as illustrated by the following figure 3.7. During School Holidays, median daily sales values are roughly the same as on regular days. This might be explained by the fact that the majority of consumers of retail healthcare and wellness products are not children, which are the majority of the students, and as such are not impacted by School Holidays.

Another interesting finding is the relationship between sales and the weather *Events* variable. The following figure 3.8 illustrate a box plot of daily sales for each event. The events are ordered by descending median daily sales value. One observation is that days in which extreme events occurred, such as Fog-Rain-Snow-Hail and Rain-Snow-Hail-Thunderstorm display the highest median daily sales values. We believe this may be explained by such events mainly occurring in colder periods of the year, where flu and colds are more common, and naturally the demand for healthcare and wellness products increases.

Figure 3.7: Box plots of daily sales by School Holiday

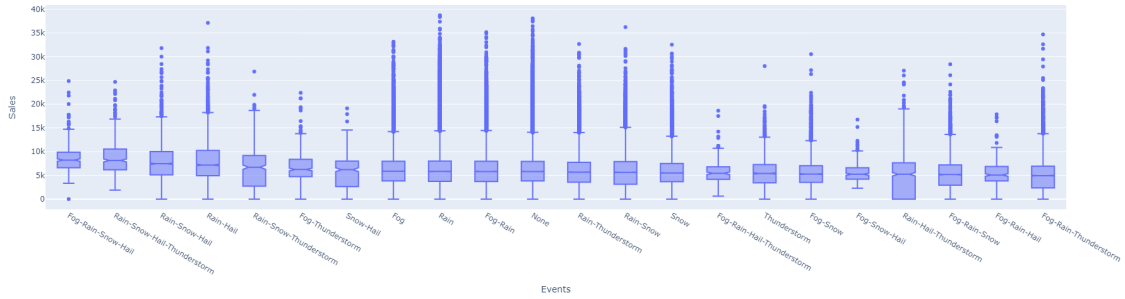

Figure 3.8: Box plot of daily sales for each weather event

3.2.4 Prepared Data Selection

In order to make it feasible to graphically compare and analyze the predictions of each model, a sub-set of 16 time series was selected from a total of 932. We set, for each time series, the median of the value of daily sales as our time series representation measurement, given that it's more robust towards outliers, in order to be used as a selection tool. This is illustrated by figure 3.9, a box-plot of the median daily sales value for each store.

We can see that there is much variability in the values of median daily sales values, by store, indicating that different stores have different dimensions and sales volumes, which translates to disparate seasonality and stability between time series. Considering this, we not only selected multiple individual time series but also grouped them based on the proximity of the median daily sales values.

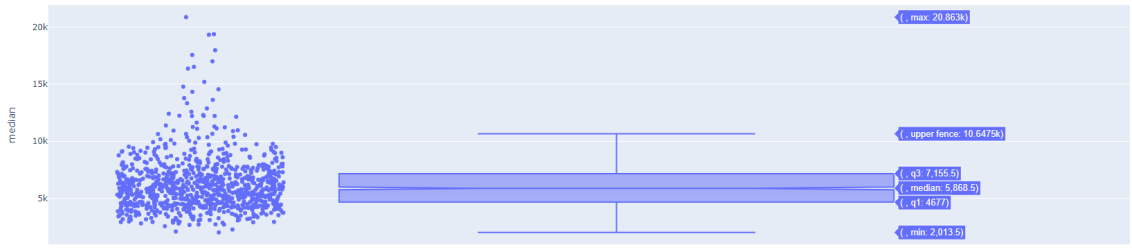

Figure 3.9: Box plot of median daily sales per store

Four time series from each inter-quartile interval were selected, making up the following four groups:

- **Low Median Daily Sales Group**
  - Four randomly selected series whose median values range from 0 to *q1* (4677 €), the time series (stores) in this group are the ones with the lowest median daily sales value out of all the groups.
- **Medium-Low Median Daily Sales Group**
  - Four randomly selected series whose median values range from *q1* (4677 €) to *q2* (5868 €).
- **Medium Median Daily Sales Group**
  - Four randomly selected series whose median values range from *q2* (5868 €) to *q3* (7155 €).
- **High Daily Sales Group**
  - Four randomly selected series whose median values range from *q3* (7155 €) to *q4* (10647 €).

These groupings allow us not only to assess a models' performance on individual time series but also to establish a relationship between model performance and the various groups, by identifying which model performed the best overall on a given group.

3.3 Retail Drug Stores Dataset

The following data was provided by Retail Consult, a group of professionals who specialize in technology solutions for retail from Portugal. The data is from a large Portuguese healthcare and wellness products retailer, with multiple stores across the country. The goal was to forecast the sales values (in product units) for the last 48 time steps (days). It contains the historical sales data for 195801 product-store pairs, over a period of 898 days, and supplemental information such as promotional events and discounts and out of stock data. For confidentiality purposes, categorical features were anonymized.

3.3.1 Data Description

The following features were used in all models that allowed for exogenous variables, the Deep AR, Prophet, and SARIMAX models namely.

- **Promotional features**

  - **High Impact Promotion 1** - indicates if the product in the store has promotion of type High Impact 1 on that day and the discount value (from 0.0 to 1.0)
  - **High Impact Promotion 2** - indicates if the product in the store has promotion of type High Impact 2 on that day and the discount value (from 0.0 to 1.0)
  - **Promotional Special Highlight w/ Discount** - indicates the discount for the product in the store (from 0.0 to 1.0)
  - **Promotional Display** - indicates if the product in the store had a promotional display
  - **Promotional Special Highlight** - indicates if the product in the store had a promotional special highlight
  - **Special Event Mobile Holiday** - indicates if a special event mobile holiday is occurring in a particular store for a particular product
  - **Promo2 Days** - describes how many days Promo2 has been ongoing
  - **After Promotional Display** - gives the days since the last promotional display
  - **Before Promotional Display** - gives the days until the next promotional display
  - **After Promotional Special Highlight** - gives the days since the last promotional special highlight
  - **Before Promotional Special Highlight** - gives the days until the next promotional special highlight
  - **After Special Event Mobile Holiday** - gives the days since the last special event mobile holiday
  - **Before Special Event Mobile Holiday** - gives the days until the next special event mobile holiday
  - **Promotional Display FW** - the sum of occurrences of promotional displays in the following 7 day period
  - **Promotional Display BW** - the sum of occurrences of promotional displays in the previous 7 day period
  - **Promotional Special Highlight FW** - the sum of occurrences of promotional special highlights in the following 7 day period
  - **Promotional Special Highlight BW** - the sum of occurrences of promotional special highlights in the previous 7 day period
  - **Special Event Mobile Holiday FW** - the sum of occurrences of special event mobile holidays in the following 7 day period
  - **Special Event Mobile Holiday BW** - the sum of occurrences of special event mobile holidays in the previous 7 day period

- **Time features**

  - **Day** - gives the day of the week
  - **Week** - gives the week of the year
  - **Month** - gives the month of the year
  - **Year** - gives the year
  - **Day of week** - gives the day of the week
  - **Day of year** - gives the day of the year
  - **Is month end** - indicates if it's the last day of the month
  - **Is month start** - indicates if it's the first day of the month
  - **Is quarter end** - indicates if it's the last day of the quarter
  - **Is quarter start** - indicates if it's the first day of the quarter
  - **Is year end** - indicates if it's the last day of the year
  - **Is year start** - indicates if it's the first day of the year
- **Other**
  - **Out Of Stock** - indicates if a product was/became out of stock on particular store

The models considered for these experiments produce a fitted model for each time series, except for the Deep AR model, which generates a global fitted model. Due to this, it can benefit from using static categorical features. These are features that are constant across time for a given time series but can vary between different time series, such as the store ID, store Type or State, and more, in this case. The following static categorical features were used exclusively by the Deep AR model

- **Product ID** - integer product identifier
- **Product Sub Class** - describes an assortment level: a = basic, b = extra, c = extended
- **Product Class** - the State in which the store is located
- **Product Department/Group/Division/Company** - these features were provided but were not used given that for the subset of selected time series, they have a cardinality of 1.
- **Store ID** - integer product identifier
- **Store City** - gives the city the store is located in
- **Store District** - gives the district the store is located in
- **Store Region** - gives the region the store is located in
- **Store Zone** - gives the zone the store is located in
- **Store Channel** - gives the the store's channel
- **Store Company** - gives store's company

3.3.2 Data Cleaning and Preparation

The following features were added:

- A similar approach as the one employed in the first experiment was used here. The features added were:
  - **After Promotional Display**
  - **Before Promotional Display**
  - **After Promotional Special Highlight**
  - **Before Promotional Special Highlight**
  - **After Promotional Special Event Mobile Holiday**
  - **Before Promotional Special Event Mobile Holiday**
- Another addition was rolling sums. A rolling sum figure was generated for the holidays and promotions within a window of 7 days. This gives an idea of the event's status with a 7-day period. Again, this is useful information for the neural network that would allow it to pick certain patterns following a certain event. The rolling sum features added were:
  - **Promotional Display BW**
  - **Promotional Display FW**
  - **Promotional Special Highlight BW**
  - **Promotional Special Highlight FW**
  - **Promotional Special Event Mobile Holiday BW**
  - **Promotional Special Event Mobile Holiday FW**

The data provided contained 195801 time series, each time series is identified by a store and a product. Due to hardware limitations, we selected a subset of time series which was used for data analysis and for training the models. This subset encompasses the 2000 series with the most data points (non-null target values). Rather than imputing these values we opted to use the complete data that we had and ignore the incomplete time series. This resulted in our data describing 2000 time series, identified by a product and a store, each with 898 time points (days), 31 dynamic features, and 10 static features. The first 850 days were used as training input and the last 48 were used for testing.

3.3.3 Data Analysis

Like in the previous experiment, we analyzed our data to assess the relationship between the available features and the target variable (daily sales).

Firstly, some statistics regarding the target variable, the volume of daily sales in product units (PU):

- Mean = 3.476 **PU**
- Min = 0.000 **PU**
- Max = 445.000 **PU**
- Median = 2.000 **PU**
- Standard Deviation = 6.389 **PU**

The median and mean values indicate that most of these products are slow-moving, with 2 to 3 units sold per day.

To overcome this the classical forecasting techniques commonly aggregate stock-keeping units (by store or by department for example), however, this aggregation is known to lead to poor performance in some cases [19].

Another problem with aggregation is the loss of information on the demand for each product, which is crucial for stock-management decisions. As such we were interested to see how the state of the art models handled time series at the product-store level and decided not to aggregate our data.

The following figures 3.10 and 3.11 show the history of total sales aggregated across all stores and products.

There are clear peaks in sales in the periods of September to October and December to January. Given that these are sales of healthcare and wellness retail products, this appears to be expected as flu and colds are more common in these periods.

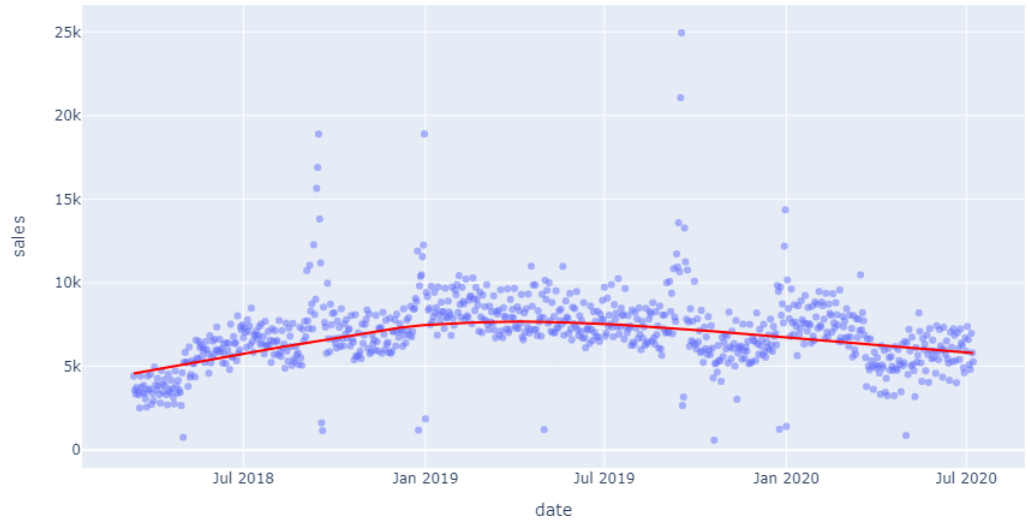

Figure 3.10: Aggregated total daily sales scatter chart

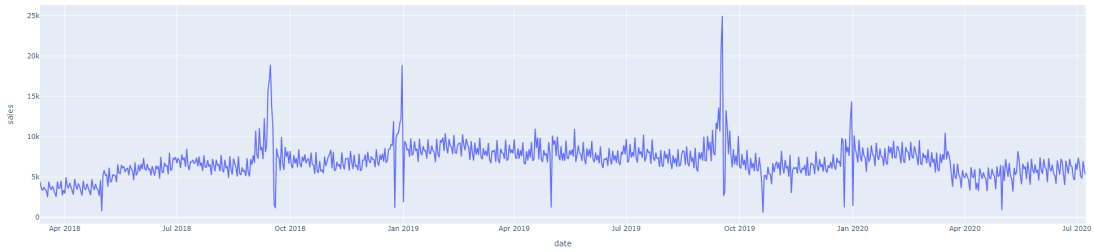

Figure 3.11: Aggregated total daily sales line chart

In comparison, when analyzing the plots of daily sales for individual time series, there are no clear seasonality or trend patterns, as illustrated by the following figures 3.12 and 3.13.

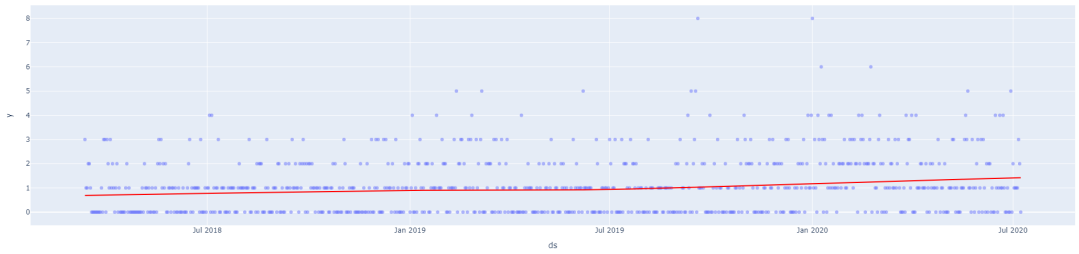

Figure 3.12: Daily sales volume for a single time series scatter chart

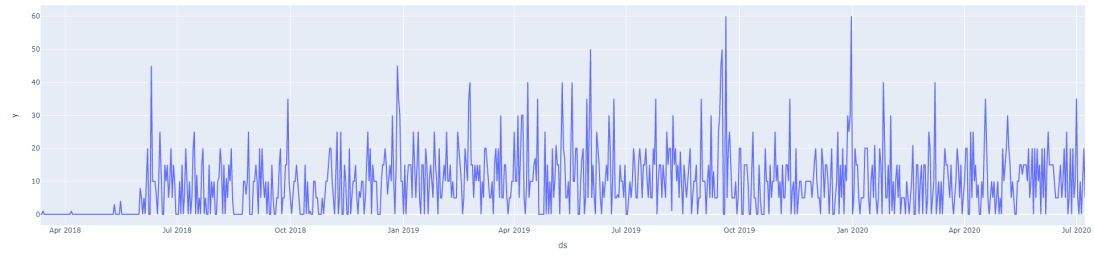

Figure 3.13: Daily sales volume for a single time serie line chart

We then analyzed the relationships between external features and the target variable. To determine if those relationships had a meaningful statistical value, we also analyzed the frequency of occurrences for each external dynamic variable.

First, we examined how out of stock occurrences affected the daily sales, illustrated by figures 3.14. Figure 3.15 displays the same box plots but zoomed in so that the median values can be seen.

Statistics:

- Number of Out of Stock occurrences = 14543
- Total number of time points = 1700000
- Out of Stock occurs in 0.86% of the time points

In spite of the low number of out-of-stock occurrences, this features' relationship with the value of the daily sales is clear. Its occurrence leads to an expected decrease in daily sales, given that the consumers are unable to purchase the product. When out of stock occurs, the median sales value is 0.

Figure 3.14: Box plots of daily sales by out of stock
Blue box plot - In Stock
Red box plot - Out of Stock

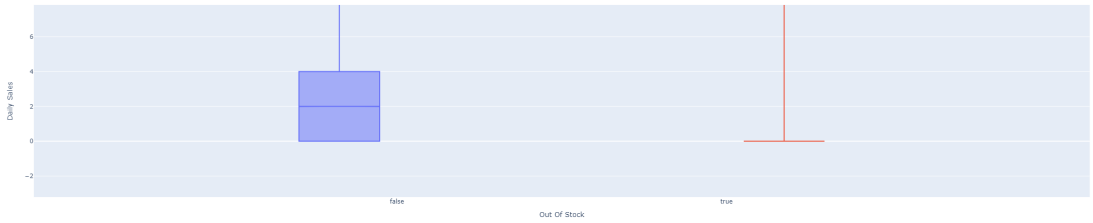

Figure 3.15: Box plots of daily sales by out of stock zoomed in
Blue box plot - In Stock
Red box plot - Out of Stock

The following box plots 3.16 illustrate the relationship between daily sales and the use of promotional displays. The same box plots are displayed in figure 3.17, zoomed so the differences in the lower ranges are perceivable.

Both the median and the maximal daily sales values are lower in the presence of a promotional display. We believe this may be explained by the very low frequency of promotional display occurrences in our sub-set of 2000 time series.

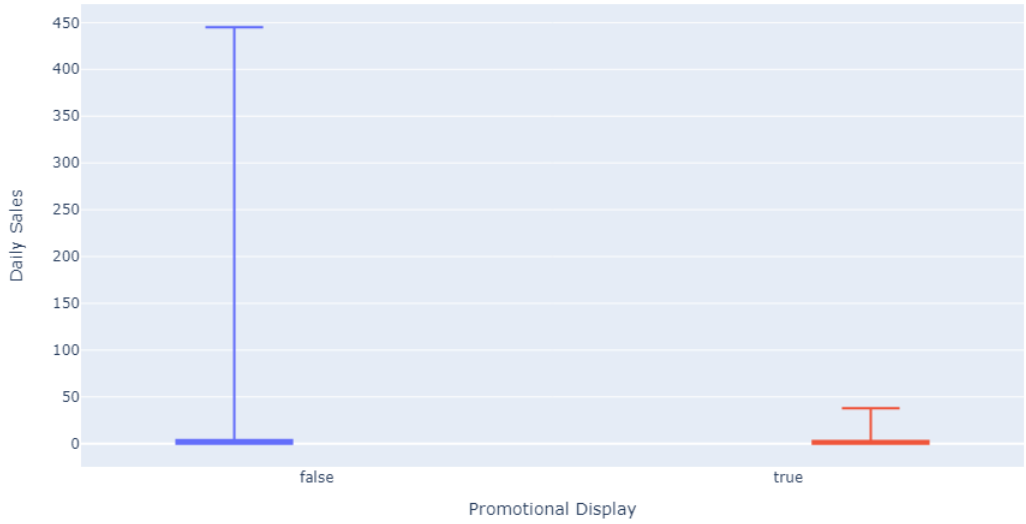

Figure 3.16: Box plots of daily sales by promotional display
Blue box plot - No Promotion
Red box plot - Promotion

Figure 3.17: Box plots of daily sales by promotional display zoomed in
Blue box plot - No Promotion
Red box plot - Promotion

Given that it is only observed in roughly 1.5% of time points, there isn't enough data to draw a conclusion on the relationship between this feature and the target variable. The low frequency of occurrences is common across all promotional features.

Statistics:

- Number of Promotional Display occurrences = 25402

- Total number of time points = 1700000
- Promotional Display occurs in 1.49% of the time points

The following box plots 3.18 illustrate the relationship between the daily sales and the occurrence of a Special Holiday Promotion. Figure 3.19 is the same as 3.18 except zoomed in.

Both median and maximum daily sales values are higher when there is no promotion, like in the previous features, this is likely due to the very low number of occurrences of this feature.

Figure 3.18: Box plots of daily sales by special holiday promotion
Blue box plot - No Promotion
Red box plot - Promotion

Figure 3.19: Box plots of daily sales by special holiday promotion zoomed in
Blue box plot - No Promotion
Red box plot - Promotion

Like the previous promotional feature, the special holiday promotion rarely occurs in the selected time series.

Statistics:

- Number of Special Holiday Promotion occurrences = 1142

- Total number of time points = 1700000
- Special Holiday Promotion occurs in 0.07% of the time points

When plotting the value of the daily sales by discount value, for features High Impact Promotional Discount 1 and 2 the relationship between each of these features and the daily sales is not clear, as illustrated by figures 3.20 and 3.21.

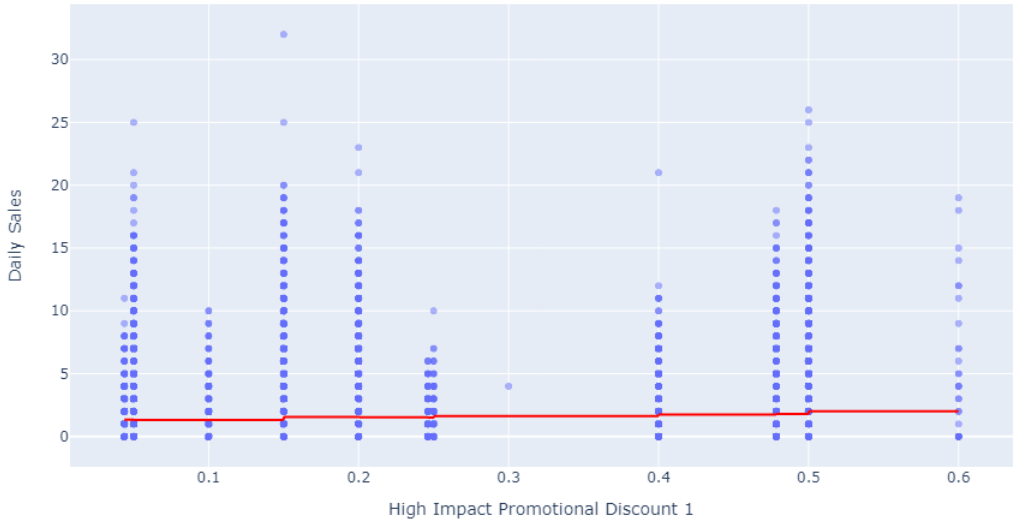

Figure 3.20: Plot of daily sales by high impact promotional discount 1

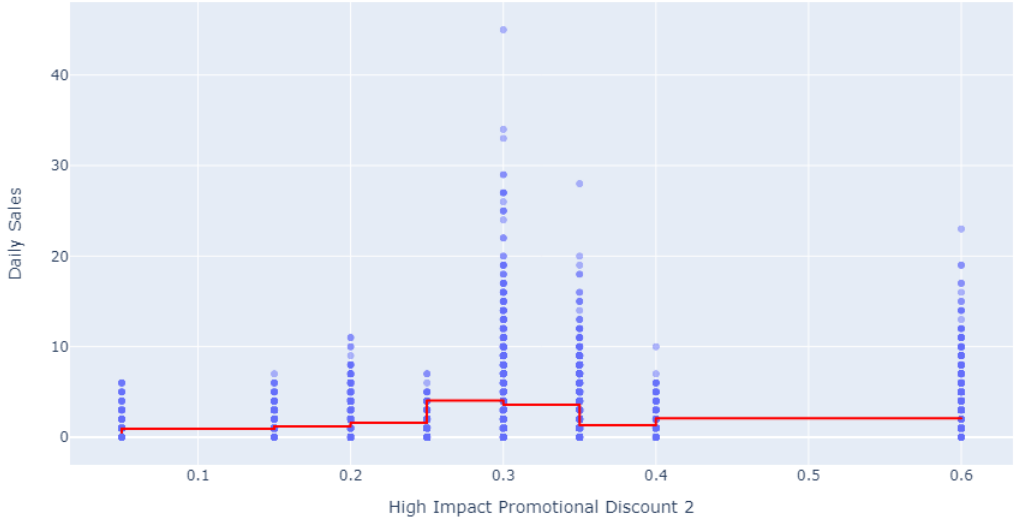

Figure 3.21: Plot of daily sales by high Impact promotional discount 2

Once more, we note that the frequency of occurrences of both of these features might explain their unclear relationship with the target variable.

Statistics:

- Number of High Impact 1 Promotional Discount occurrences = 130743
- Total number of time points = 1700000
- High Impact 1 Promotional Discount occurs in 7.69% of the time points

Statistics:

- Number of High Impact 2 Promotional Discount occurrences = 48991
- Total number of time points = 1700000
- High Impact 2 Promotional Discount occurs in 2.88% of the time points

We also analyzed the effect that the occurrence of these promotions had on daily sales, regardless of the discount value, this is illustrated in the following box plots, Fig. 3.22 and Fig. 3.23. In both cases the value of daily sales seem to decrease in the occurrence of promotions, again, given the low occurrence of these events, we cannot determine whether this means the promotions are not effective, or if we simply don't have enough data.

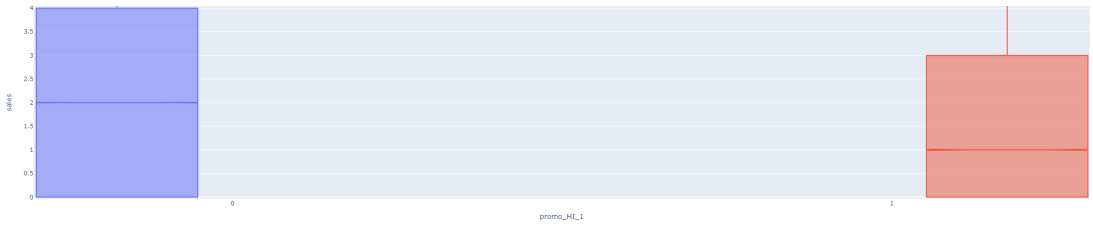

Figure 3.22: Box plots of daily sales by high impact promotional discount 1 occurrence
Blue box plot - No Promotion
Red box plot - Promotion

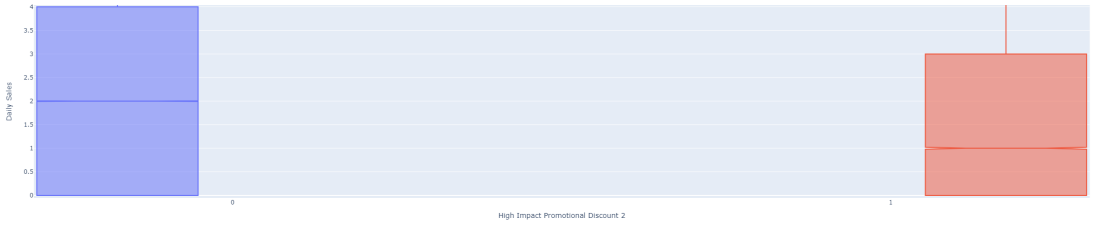

Figure 3.23: Box plots of daily sales by high impact promotional discount 2 occurrence
Blue box plot - No Promotion
Red box plot - Promotion

The following figures illustrate the relationship between the static external variables, such as product sub-class and class, and their relationship to the target variable.

Products of the sub-class H96xq have the highest demand, both in the median and the maximum values. Most sub-classes have the same median daily sales value, the largest variation is in the maximum value, as illustrated by 3.24

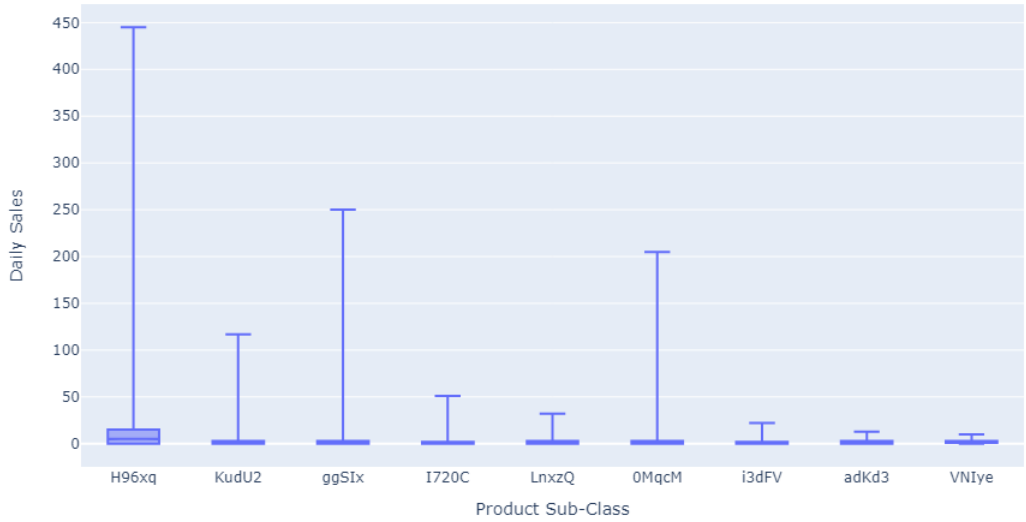

Figure 3.24: Box plots of daily sales by product sub class

Products of the class 5Isak have the highest demand, both in the median and the maximum values. Most classes have the same median daily sales value, the largest variation is in the maximum value.

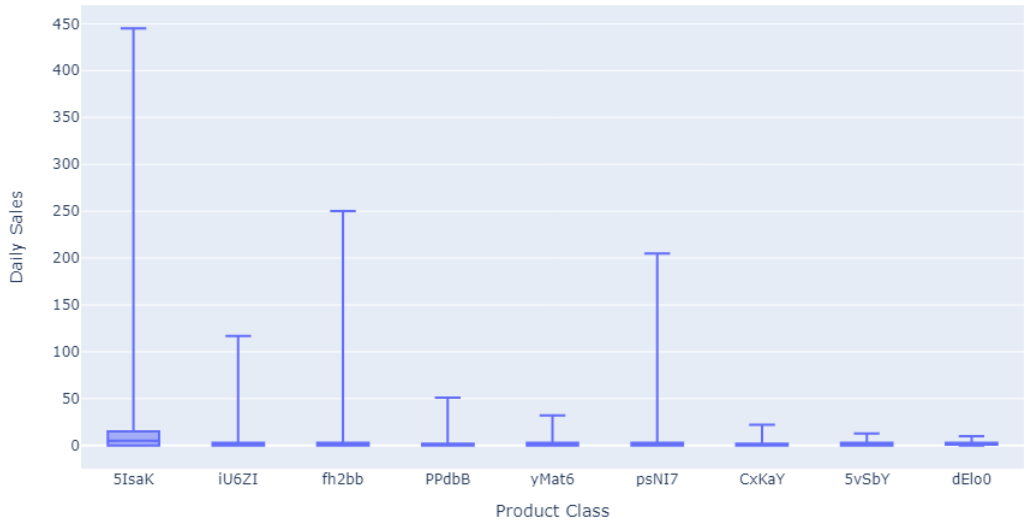

Figure 3.25: Box plots of daily sales by product class

There is a high variation of daily sales depending on the city, district and region that the stores are located in, both in median and maximum daily sales values, as illustrated by figures 3.26, 3.27,3.28, 3.29,3.32, 3.33,

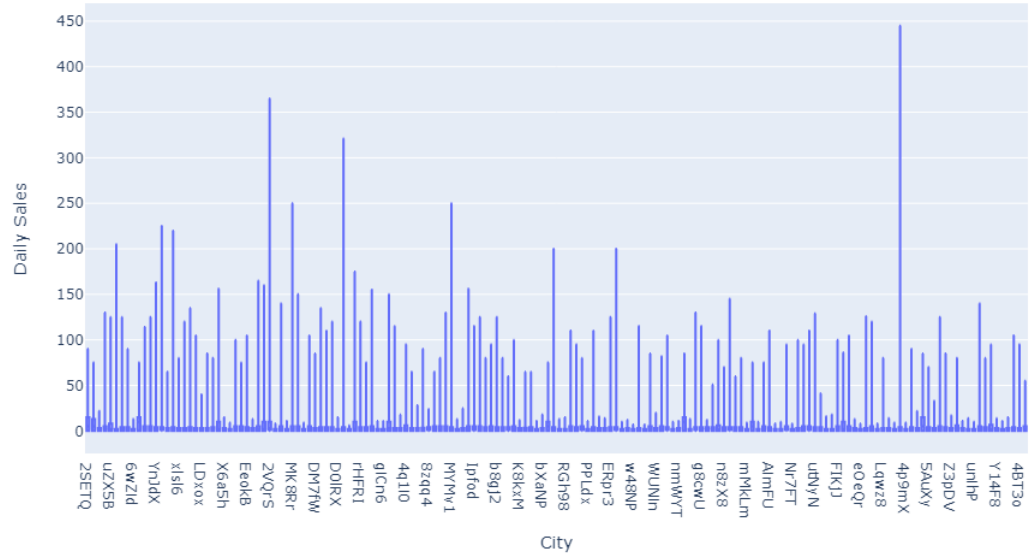

Figure 3.26: Box plots of daily sales by city

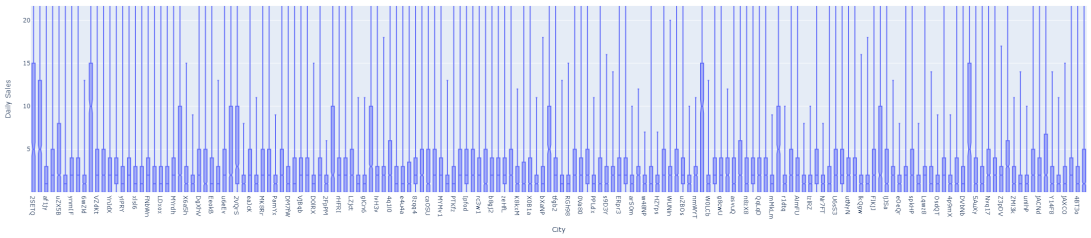

Figure 3.27: Box plots of daily sales by city zoomed in

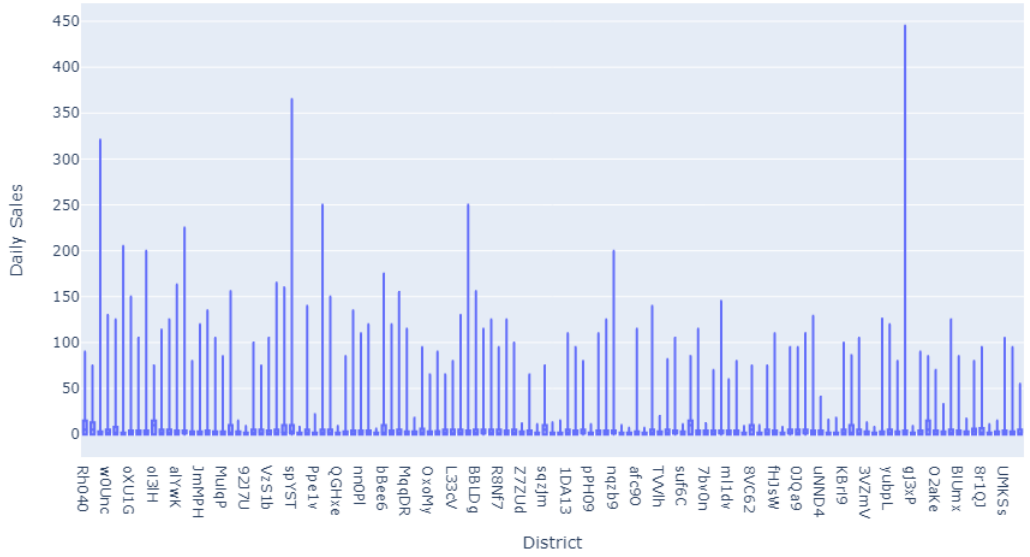

Figure 3.28: Box plots of daily sales by district

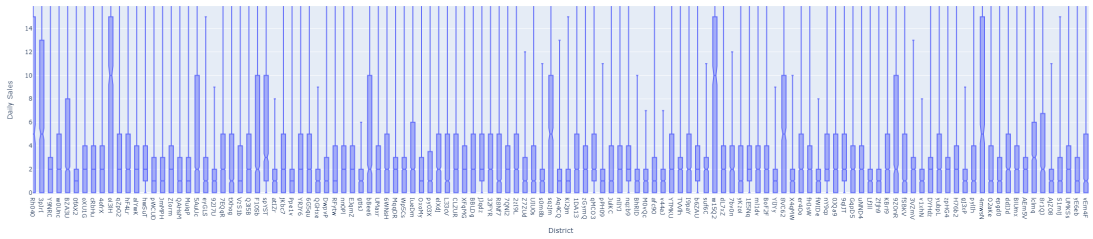

Figure 3.29: Box plots of daily sales by district zoomed in

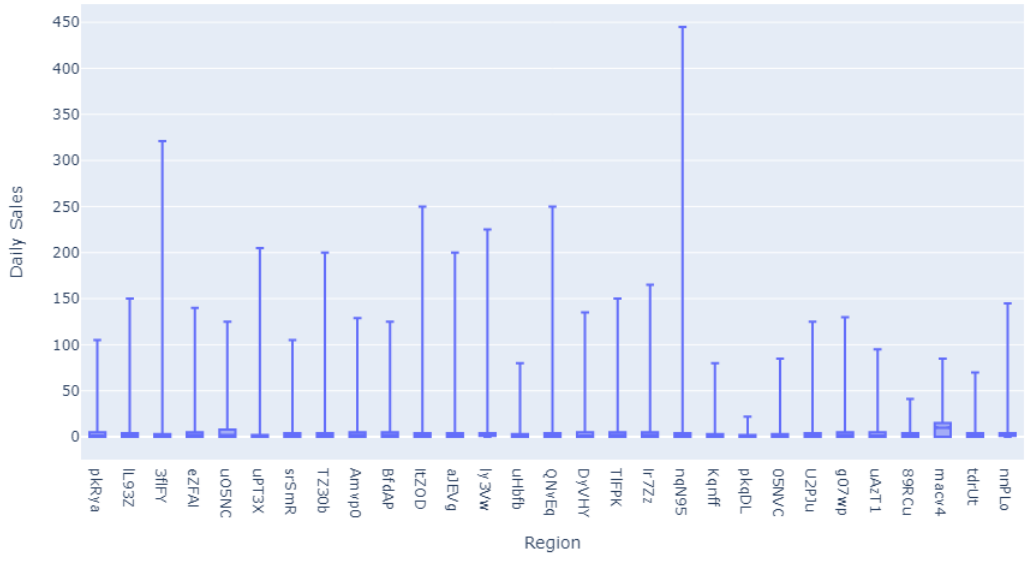

Figure 3.30: Box plots of daily sales by region

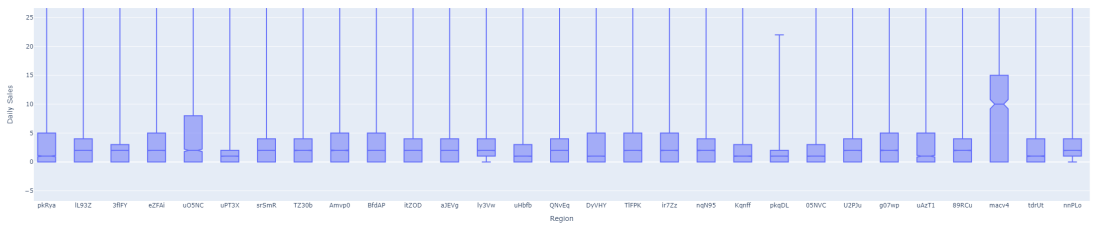

Figure 3.31: Box plots of daily sales by region zoomed in

Stores located in different zones have mostly the same median daily sales value, the variation lies in the maximum values as illustrated in figure 3.30 and 3.31

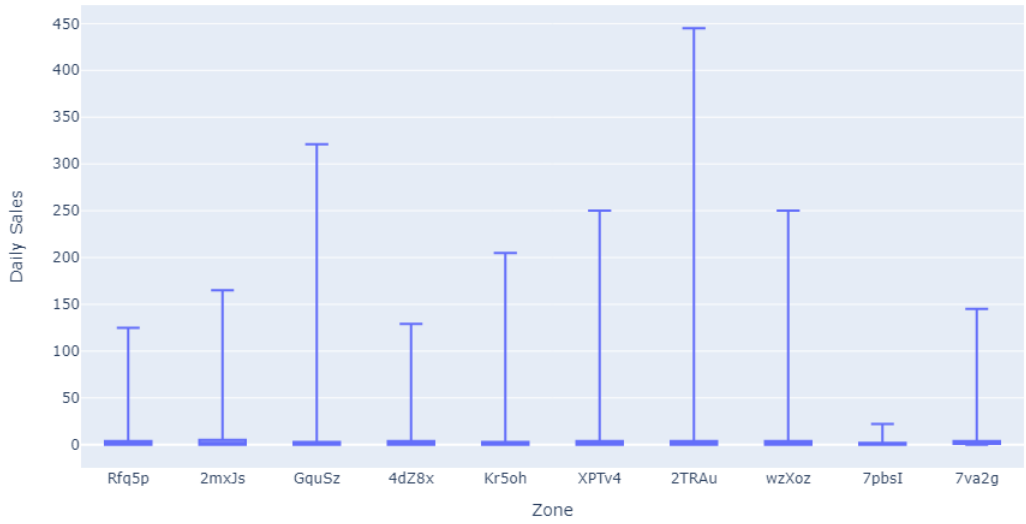

Figure 3.32: Box plots of daily sales by zone

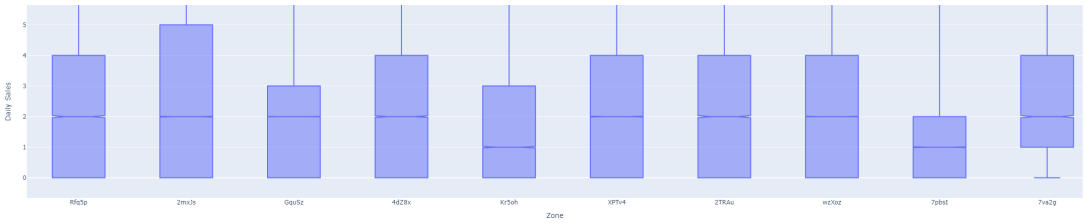

Figure 3.33: Box plots of daily sales by zone zoomed in

The median daily sales are the same for each channel, again the variation being in the maximum values, illustrated by figure 3.34

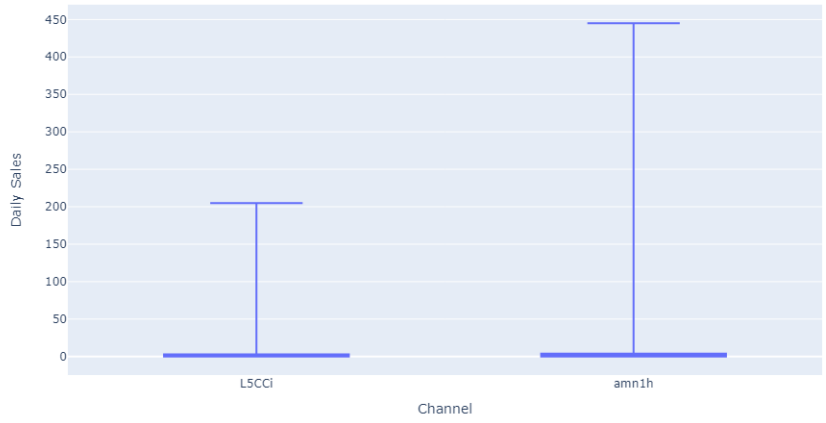

Figure 3.34: Box plots of daily sales by channel

Our data analysis leads us to conclude that for this experiment the static features are as valuable as the dynamic features if not more, and their relationship with the target variable is clearer.

3.3.4 Prepared Data Selection

The data selection process was similar to the one in the previous experiment, however, given the low variability of the median daily sales value for each time series, we instead set the average as our selection metric. The following figure 3.35 shows a box-plot of the average daily sales for each time series and we can observe significant variability in these values as well as a large number of outliers.

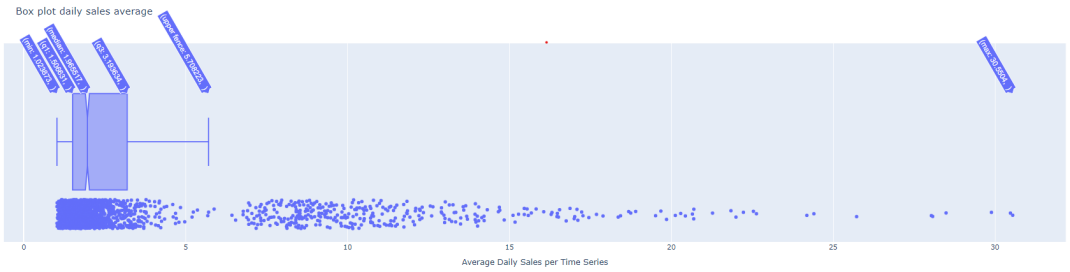

Figure 3.35: Box plot of average daily sales per Product-Store

Like in the previous experiment, this variability lead us to group the selected time series. They were grouped as follows:

- **High Average Daily Sales Group**
  - Five series with an average daily sales value ranging from 10 to 30 units.

- **Medium Average Daily Sales Group**
  - Two series with an average daily sales value ranging from 3 to 4 units.
- **Low Average Daily Sales Group**
  - Two series with an average daily sales value ranging from 1 to 2 units.

Chapter 4

Tests and Results

4.1 Rossmann Dataset Results

In this section, we will analyze and discuss the forecasts produced by each model and their performance. In the following tables, the cells highlighted from dark green to red, indicating the smallest and largest values. We selected the MASE measurement as our performance metric. Proposed Hyndman and Koehler (2006) [29], the MASE can be used to assess forecasting accuracy on individual time series as well as for multiple time series, and for all forecast methods and all types of series. It has been considered the best accuracy metric for intermittent demand studies and beyond by several data scientists [29]. The following is a list of commonly used performance metrics and our reasons [29] for not using them. A more in-depth explanation of each metric can be found in chapter 2 2.

- Scale dependant such as the MAE metrics weren't considered since all they are on the same scale as the data, and thus none of them are meaningful for assessing a method's accuracy across multiple series with varying scales. The RMSE wasn't considered when analyzing single time series due to its vulnerability towards extreme values.
- Error measurements based on percentage errors like the MAPE have the disadvantage of being infinite or undefined if there are zero values in a series, as is frequent for the considered data (intermittent). Additionally, percentage errors can have an extremely skewed distribution when actual values are close to zero.
- Relative errors weren't included given that when the errors are small, as they can be with intermittent series, the use of the naïve method as a benchmark is not possible because it would involve division by zero.

4.1.1 Average MASE measurements and Standard Deviation for each model for each group

The average MASE measurements for each group and the standard deviation are displayed in figure 4.1.

Regarding the first 2 groups, low and medium-low daily sales (Q1 and Q2), the Deep AR model achieved the lowest MASE measurements, indicating the best fit. The SARIMAX and Prophet models had very similar results, and the Holt Winter's model consistently produced the highest MASE values in all groups, indicating the worst fit.

The SARIMAX model has the best performance on the medium and high average groups (Q3 and Q4) reaching the lowest MASE values. In Q3 the Deep AR model reached the second-best fitness score, followed by Prophet, while in Q4 the opposite occurs. These results were very similar to the results for Q2.

Overall the standard deviations are very small, meaning the models' performance on time series of the same group is similar. The exception being in group Q4, where performance deviates the most for all models, in particular for the Deep AR model.

	MASE	
	Average	STD
Prophet - Q1	0.315	0.032
HW - Q1	1.086	0.029
SARIMAX - Q1	0.321	0.067
DeepAR - Q1	0.208	0.047
Prophet - Q2	0.222	0.032
HW - Q2	0.999	0.028
SARIMAX - Q2	0.186	0.028
DeepAR - Q2	0.184	0.031
Prophet - Q3	0.222	0.025
HW - Q3	1.070	0.134
SARIMAX - Q3	0.202	0.035
DeepAR - Q3	0.211	0.047
Prophet - Q4	0.383	0.211
HW - Q4	1.013	0.128
SARIMAX - Q4	0.353	0.207
DeepAR - Q4	0.557	0.453

Figure 4.1: Average metrics for each model for each quartile

In the following subsection, we analyze the performances of each model for the individual groups and time series.

4.1.2 Forecasts and Metrics for Low Median Group

For most time series in this group, the results in figure 4.2 show that the lowest MASE measurement was given by the Deep AR model. For the time series 157, the lowest MASE score was achieved by both the Deep AR and SARIMAX models, the difference between them being under 1%.

	MASE		MASE
Prophet - 157	0.299	Prophet - 697	0.291
HW - 157	1.054	HW - 697	1.093
SARIMAX - 157	0.232	SARIMAX - 697	0.352
DeepAR - 157	0.237	DeepAR - 697	0.201

	MASE		MASE
Prophet - 453	0.370	Prophet - 701	0.300
HW - 453	1.067	HW - 701	1.129
SARIMAX - 453	0.411	SARIMAX - 701	0.290
DeepAR - 453	0.259	DeepAR - 701	0.134

Figure 4.2: Metrics for each models forecasts for each time series from Low median group (Q1)

The forecasts and prediction intervals for store 701, for each model are illustrated in figures, 4.3,4.4, 4.5, and 4.6. analyzing the graphs, all models generate reliable forecasts, and in general produce forecasts close to the actual values, however, the Holt Winter's model notably produces the least accurate forecasts. Additionally, the prediction intervals generated by the Deep AR model display the highest accuracy (actual values lie within range) and precision (the intervals have smaller amplitudes) when compared to the ones generated by either the SARIMAX or Prophet models.

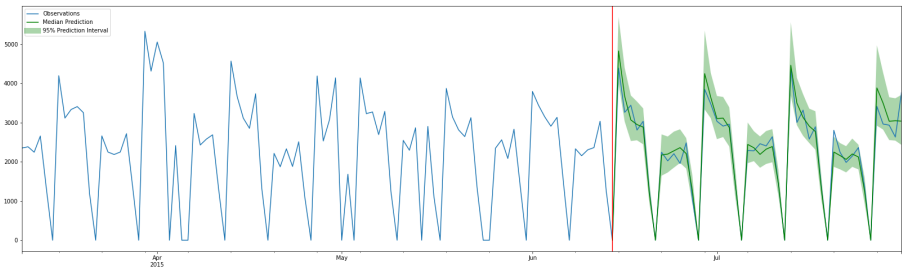

Figure 4.3: Deep AR forecasts, prediction intervals and actual values for store 701

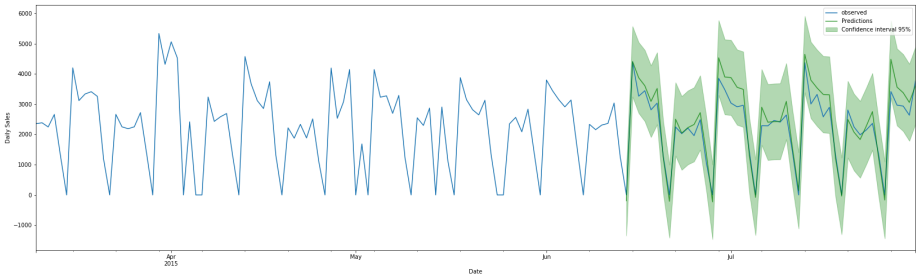

Figure 4.4: SARIMAX forecasts, prediction intervals and actual values for store 701

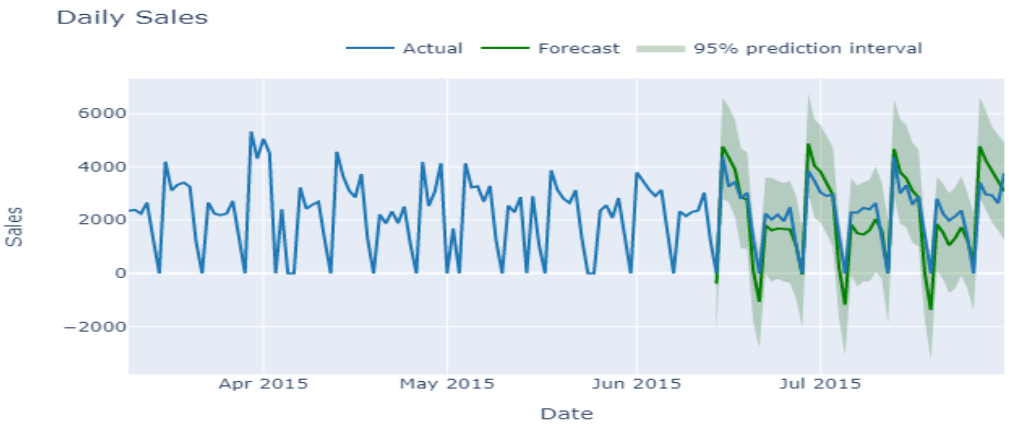

Figure 4.5: Prophet forecasts, prediction intervals and actual values for store 701

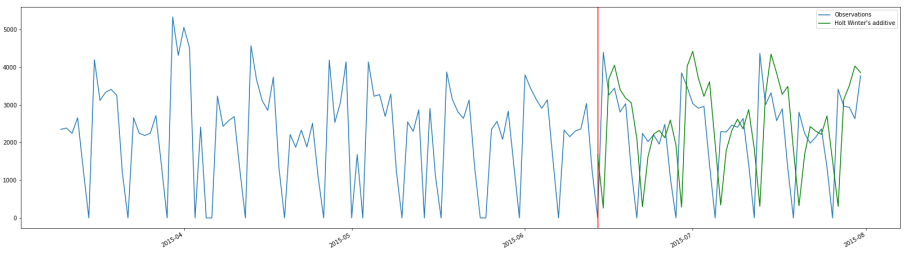

Figure 4.6: Holt Winter's forecasts and actual values for store 701

4.1.3 Forecasts and Metrics for Medium-Low Median Group

Overall, the results in figure 4.7 indicate that the lowest MASE values in group Q2 were given by the Deep AR and SARIMAX models. For stores 177 and 401, the Deep AR came ahead of SARIMAX, and conversely for stores 112 and 676. Like in the previous group the highest MASE values were reached by the Holt Winter's model.

	MASE		MASE
Prophet - 177	0.218	Prophet - 112	0.273
HW - 177	0.955	HW - 112	1.004
SARIMAX - 177	0.222	SARIMAX - 112	0.203
DeepAR - 177	0.214	DeepAR - 112	0.212

	MASE		MASE
Prophet - 401	0.209	Prophet - 676	0.186
HW - 401	1.003	HW - 676	1.033
SARIMAX - 401	0.164	SARIMAX - 676	0.155
DeepAR - 401	0.140	DeepAR - 676	0.170

Figure 4.7: Metrics for each models forecasts for each time series from Medium-Low median group (Q2)

The forecasts and prediction intervals for store 401, for each model are illustrated in figures, 4.8,4.9, 4.10, and 4.11.

analyzing the graphs, like in the previous group all models generated reliable forecasts, and in general produced forecasts close to the actual values, once more the Holt Winter's model produced the least accurate forecasts.

Once again the prediction intervals generated by the Deep AR model display the highest accuracy and precision when compared to the ones generated by either the SARIMAX or Prophet models.

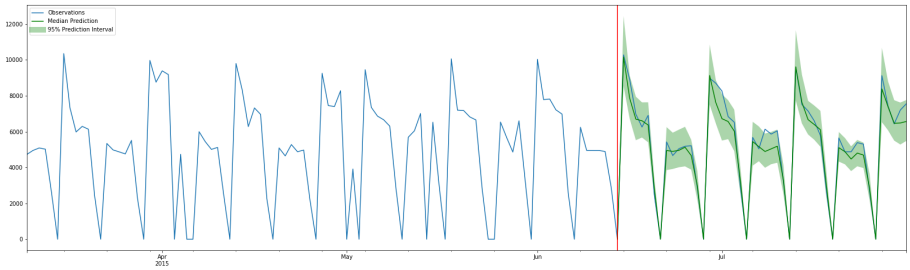

Figure 4.8: Deep AR forecasts, prediction intervals and actual values for store 401

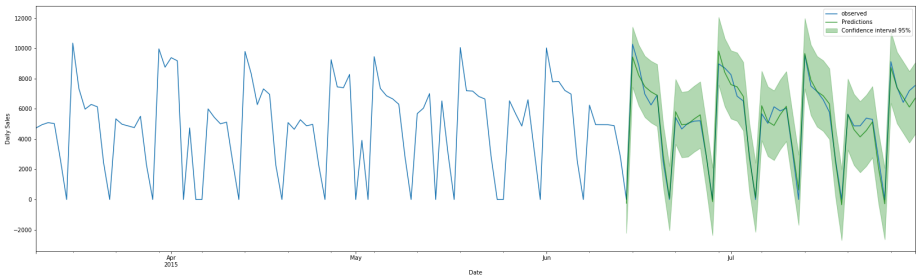

Figure 4.9: SARIMAX Forecast for store 401

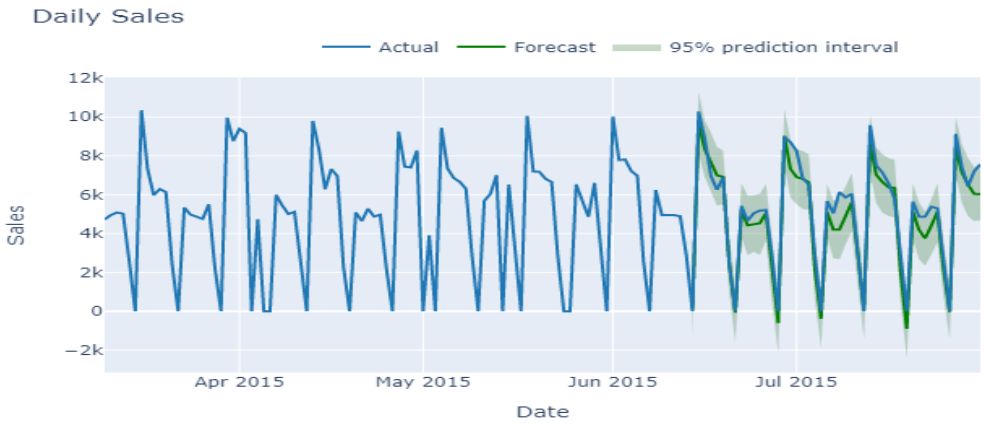

Figure 4.10: Prophet forecasts, prediction intervals and actual values for store 401

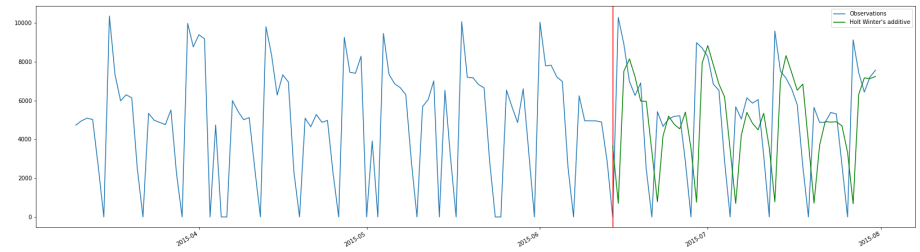

Figure 4.11: Holt Winter's forecasts and actual values for store 401

4.1.4 Forecasts and Metrics for Medium Median Group

In this group, the results are more mixed, as illustrated by figure 4.12. For store 850 the Deep AR model achieved the lowest MASE. The second-lowest MASE was given by the Prophet model followed very closely by SARIMAX.

For store 384 the lowest MASE was given by the SARIMAX, followed by Prophet and then Deep AR.

The lowest MASE for store 130 was again reached by the Deep AR model, followed by SARIMAX and Prophet.

Lastly, the SARIMAX model achieved the smallest MASE values for store 113. The second-lowest MASE was given by the Prophet model followed closely by Deep AR.

Once again the highest MASE values were given by the Holt Winter's model for all stores.

	MASE		MASE
Prophet - 850	0.241	Prophet - 384	0.178
HW - 850	1.283	HW - 384	0.917
SARIMAX - 850	0.255	SARIMAX - 384	0.161
DeepAR - 850	0.173	DeepAR - 384	0.253

	MASE		MASE
Prophet - 130	0.234	Prophet - 113	0.234
HW - 130	1.065	HW - 113	1.014
SARIMAX - 130	0.208	SARIMAX - 113	0.184
DeepAR - 130	0.157	DeepAR - 113	0.262

Figure 4.12: Metrics for each models forecasts for each time series from Medium median group (Q3)l

The forecasts and prediction intervals for store 113, for each model are illustrated in figures, 4.13,4.14, 4.15, and 4.16.

We can observe that in the case of store 113 the lowest MASE, and thus the best performance, was given by the SARIMAX model, however, the most accurate and precise prediction intervals were once more generated by the Deep AR model.

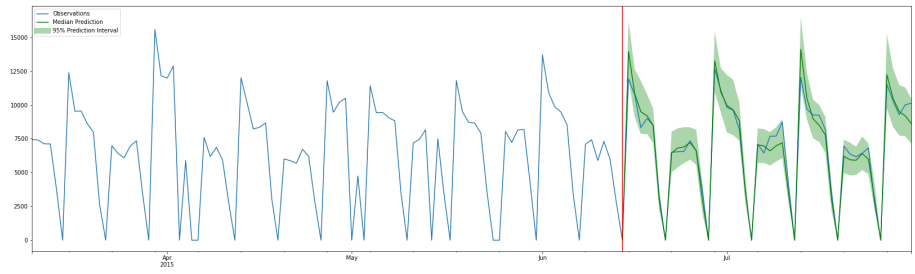

Figure 4.13: Deep AR forecasts, prediction intervals and actual values for store 113

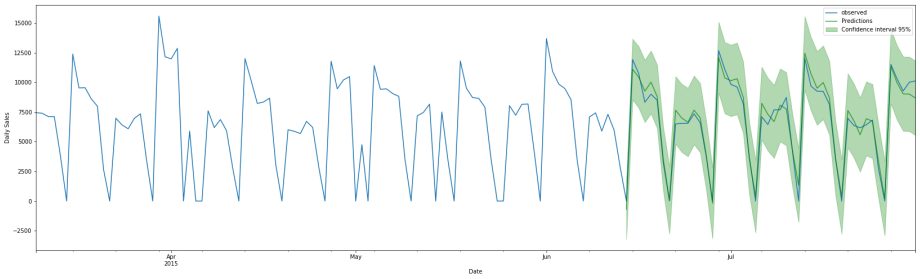

Figure 4.14: SARIMAX forecasts, prediction intervals and actual values for store 113

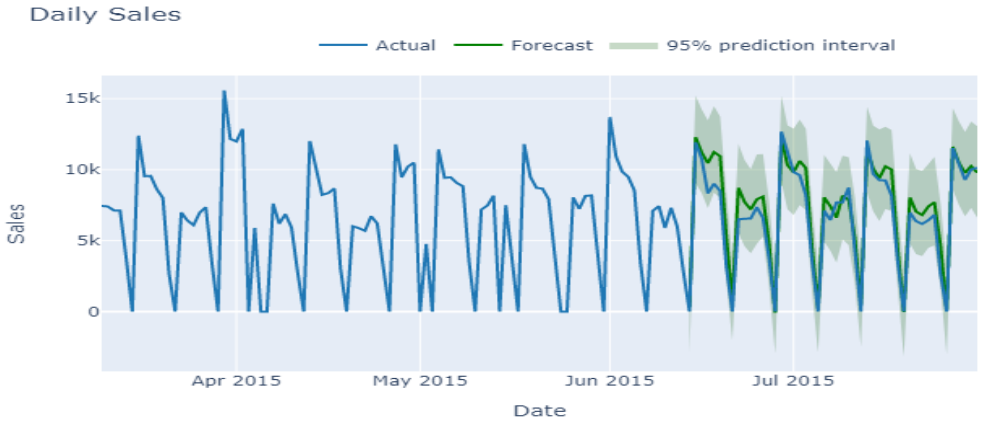

Figure 4.15: Prophet forecasts, prediction intervals and actual values for store 113

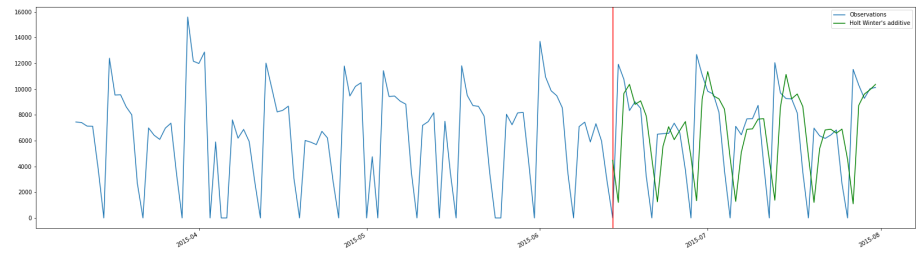

Figure 4.16: Holt Winter's forecasts and actual values for store 113

4.1.5 Forecasts and Metrics for High Median Group

Like in the previous group, the results, as illustrated by figure 4.17, are mixed, but even more so.

Store 639 is the only time series in which the Deep AR model gave the highest MASE value out of all the models. In comparison to previous results, all models performed the worst on this particular time series, indicating this series has a higher degree of irregularity and less perceivable trend and seasonality patterns. The SARIMAX model fits these series the best, achieving the lowest MASE, followed by Prophet and Holt Winter's. So far we've seen the best overall from the Deep AR and SARIMAX models, however, the results for store 639 differ from this trend. Unlike the previously considered stores, store 639 has no zero sales values, in fact, it is the only store that is never closed. Since the Deep AR generates a global model for all the time series it is given as input, it is expected that forecasts generated for time series with generating processes that vastly differ from the majority of the time series, supplied as training input, to be less accurate.

Regarding store 691 the Deep AR achieved the lowest MASE, followed by SARIMAX and Prophet models. All models reached low MASE values with except the Holt Winter's model.

The lowest MASE measurement for store 355 was given by the Prophet model, followed by SARIMAX.

For store 360 the SARIMAX and Prophet models gave the lowest MASE values, followed closely by the Deep AR model.

	MASE
Prophet - 639	0.721
HW - 639	0.922
SARIMAX - 639	0.683
DeepAR - 639	1.295

	MASE
Prophet - 691	0.242
HW - 691	1.206
SARIMAX - 691	0.197
DeepAR - 691	0.157

	MASE
Prophet - 355	0.393
HW - 355	0.875
SARIMAX - 355	0.372
DeepAR - 355	0.557

	MASE
Prophet - 360	0.174
HW - 360	1.047
SARIMAX - 360	0.160
DeepAR - 360	0.217

Figure 4.17: Metrics for each models forecasts for each time series from High median group (Q4)

The forecasts and prediction intervals for store 639, for each model are illustrated in figures 4.18,4.19, 4.20, and 4.21.

In this case, the most accurate precision intervals were generated by the SARIMAX model, although less precise. The Deep AR model still generates the most precise intervals although some of the actual values fall outside of the interval.

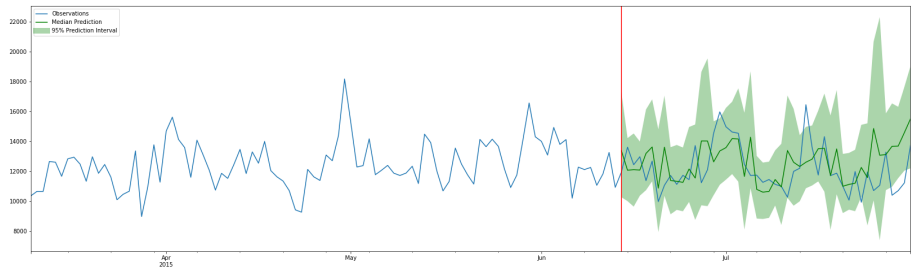

Figure 4.18: Deep AR forecasts, prediction intervals and actual values for store 639

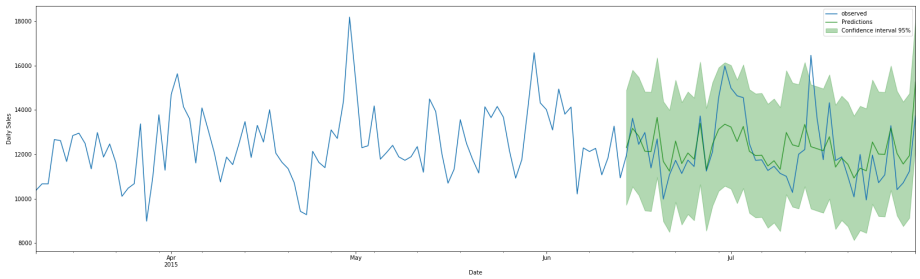

Figure 4.19: SARIMAX forecasts, prediction intervals and actual values for store 639

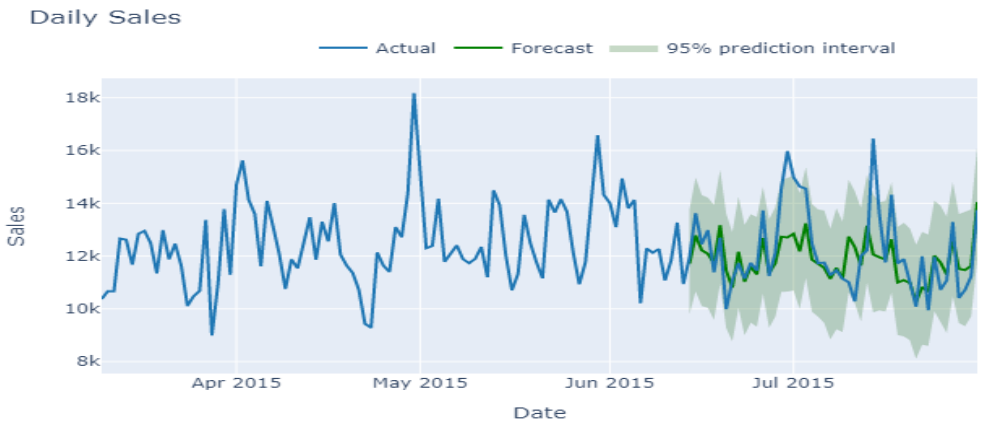

Figure 4.20: Prophet forecasts, prediction intervals and actual values for store 639

4.2 Retail Dataset Results

Given the data analysis outlined in the previous chapter 3 regarding the Retail data set, and the lack of observations of dynamic features across all time series, their value in generating predictions appears to be low, in fact, their inclusion in the models may result in less accurate forecasts. To account for this, we fitted the Deep AR, SARIMAX, and Prophet models with and without these features. The SARIMA, Prophet and Deep AR refer to the models fitted without external variables, and the SARIMAX, Deep AR - FDR (Features Dynamic Real) and ProphetX labels refer to the

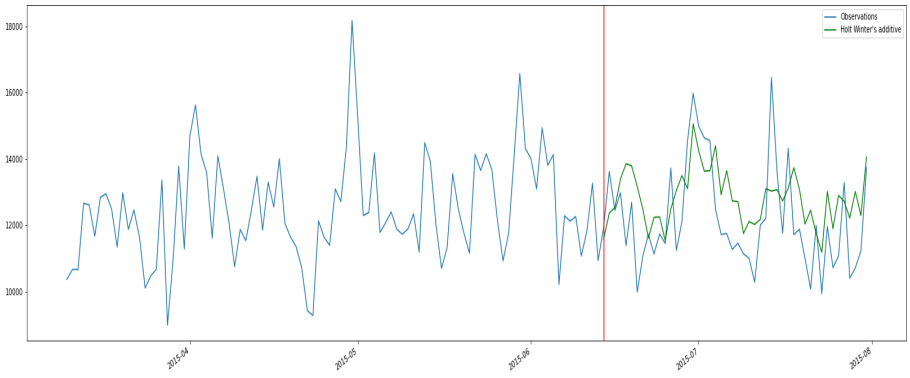

Figure 4.21: Holt Winter's forecasts and actual values for store 639

models that do incorporate them.

4.2.1 Average MASE measurements and Standard Deviation for each model

The lowest average MASE value across all time series was achieved by the Deep AR - FDR model and the highest by the Holt Winter's model. However when comparing both Deep AR model's standard deviations, the one including dynamic features has a higher performance variability.

The difference in the MASE averages for the Deep AR - FDR and SARIMAX models in comparison to their respective counterparts are small, indicating that the external features had little impact on the predicted values. However in Prophet's case, including these exogenous variables had a noticeable negative impact on forecast accuracy, as illustrated by figure A.1 Like in the previous experiment the highest MASE values were given by the Holt Winter's model forecasts.

	MASE	
	Average	STD
HW	1.024	0.205
SARIMA	0.760	0.102
SARIMAX	0.802	0.143
Prophet	0.853	0.208
ProphetX	0.997	0.315
DeepAR	0.695	0.135
DeepAR-FDR	0.661	0.208

Figure 4.22: Average metrics for each model for each quartile

4.2.2 Forecasts and Metrics for Low Average Group

Regarding the low average group, the SARIMA and SARIMAX models achieved the lowest MASE values, as illustrated by figure 4.23. In particular, for the time series 834, the lowest MASE was reached by SARIMAX. The second-lowest MASE was achieved by both SARIMA and Prophet models, followed by the Holt Winter's which displayed a better performance than both Deep AR models. For this time series, the worst performance was given by the Prophet model that incorporated external variables.

Concerning the time series 1496, the worst performances were given by the Deep AR model using dynamic features, followed by the Holt Winter's model and the Deep AR model without dynamic features. The lowest MASE was achieved by the SARIMA model, followed by both the SARIMAX and Prophet models, which reached the exact same performance. The ProphetX model performed slightly worse than Prophet.

	MASE		MASE
HW - 834	0.831	HW - 1496	0.940
SARIMA - 834	0.735	SARIMA - 1496	0.729
SARIMAX - 834	0.708	SARIMAX - 1496	0.754
Prophet - 834	0.735	Prophet - 1496	0.754
ProphetX - 834	1.703	ProphetX - 1496	0.777
DeepAR - 834	0.873	DeepAR - 1496	0.892
DeepAR-FDR - 834	0.931	DeepAR-FDR - 1496	1.110

Figure 4.23: MASE measurements for each models forecasts for each time series from the low average group

The forecasts and prediction intervals for time series 1496, for each model are illustrated in figures, 4.24,4.25, 4.26,4.27,4.28,4.29,4.30. The prediction intervals generated by the Deep AR are the most precise and also do not include negative values unlike the SARIMA and Prophet models. While the remaining models assume a Gaussian distribution, the Deep AR model allows us to set the negative binomial distribution as our likelihood function, which ensures that predicted values and prediction intervals are positive. We can also see that the prediction intervals fall short in accuracy across all models, there being multiple instances of actual values falling outside of the interval, mainly peaks.

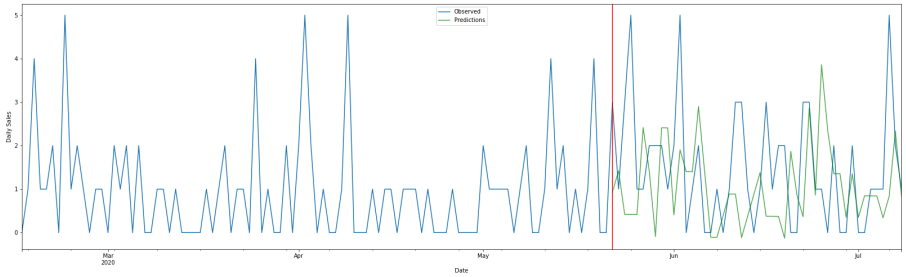

Figure 4.24: Holt Winter's forecasts and actual values for time series 1496

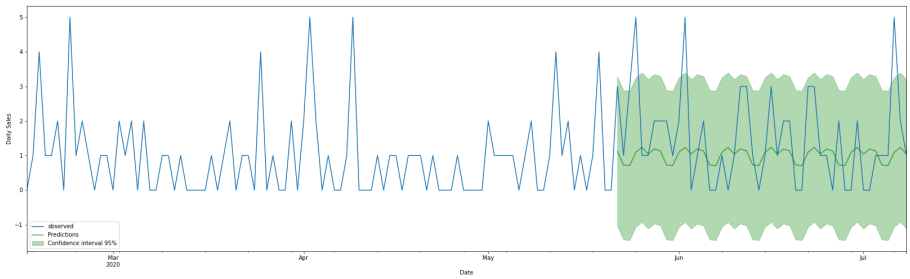

Figure 4.25: SARIMA forecasts, prediction intervals and actual values for time series 1496

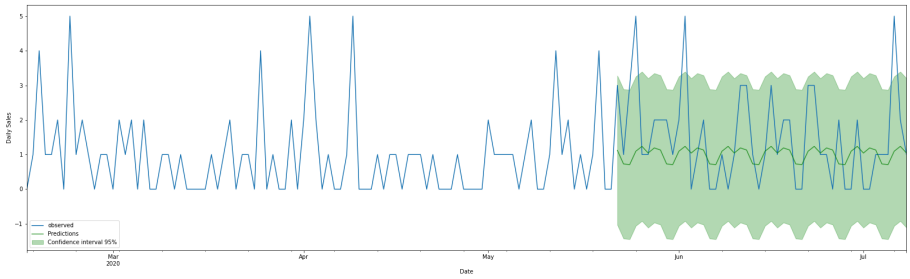

Figure 4.26: SARIMAX forecasts, prediction intervals and actual values for time series 1496

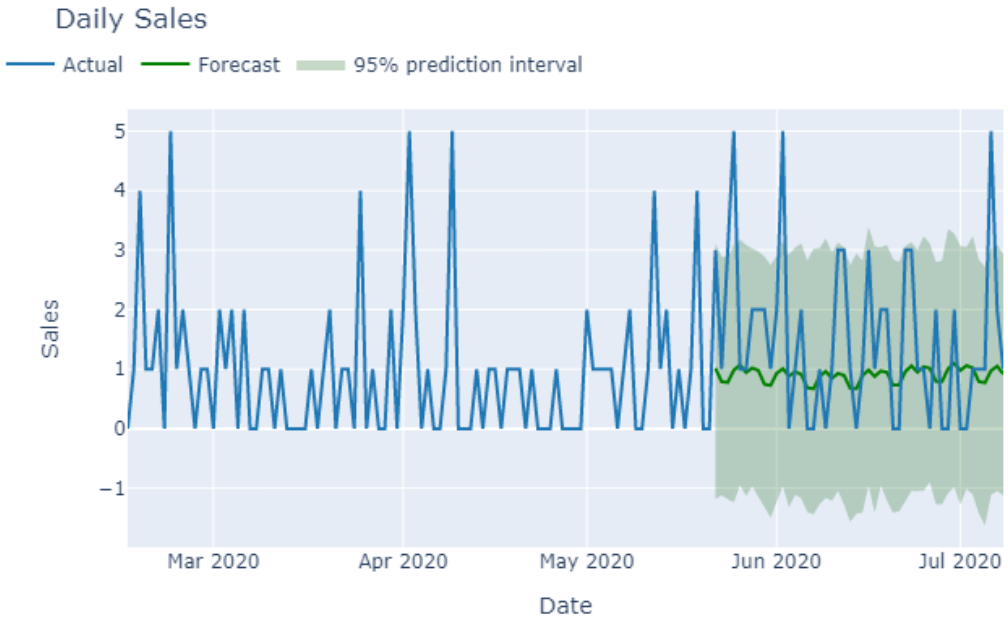

Figure 4.27: Prophet forecasts, prediction intervals and actual values for time series 1496

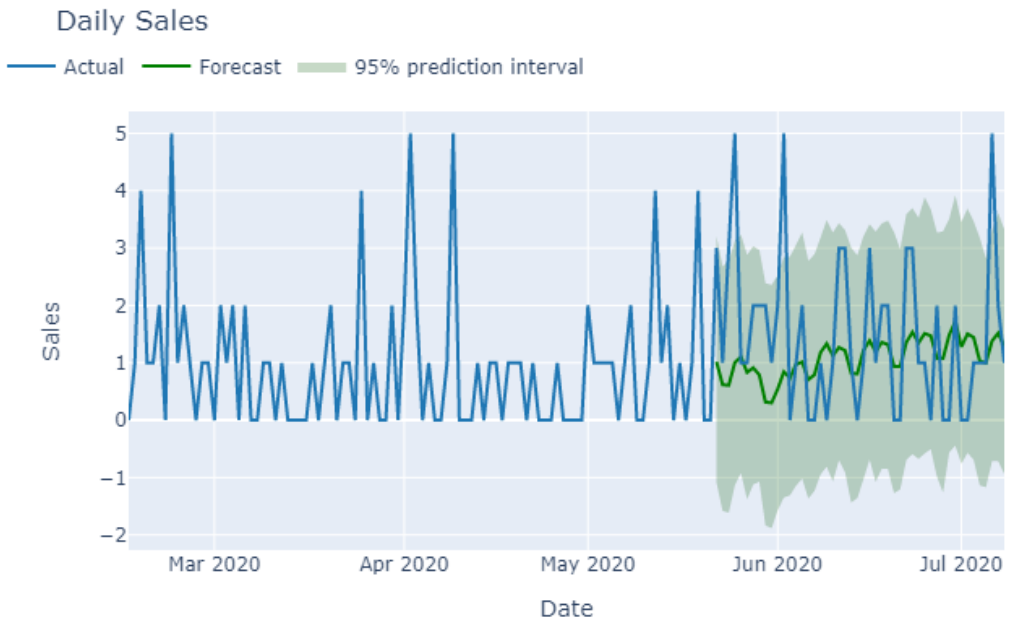

Figure 4.28: ProphetX forecasts, prediction intervals and actual values for time series 1496

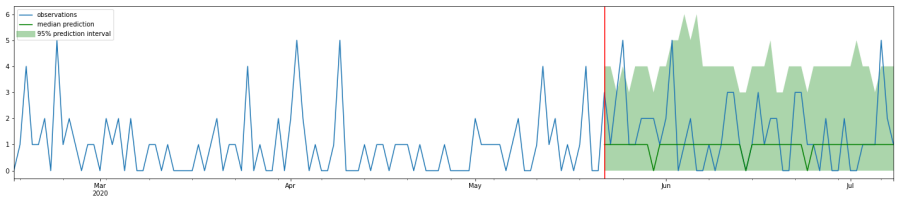

Figure 4.29: Deep AR forecasts, prediction intervals and actual values for time series 1496

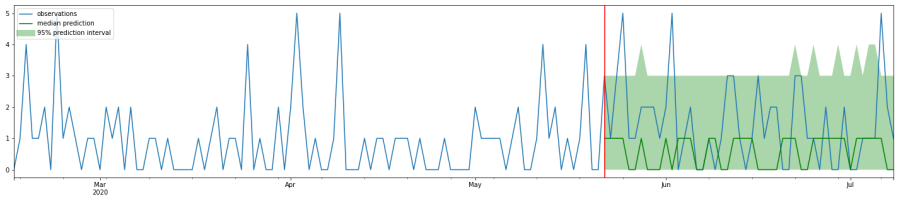

Figure 4.30: Deep AR - FDR forecasts, prediction intervals and actual values for time series 1496

4.2.3 Forecasts and Metrics for Medium Average Group

The lowest MASE values were achieved in both time series of the Medium Average group by the Deep AR model using dynamic external variables, and the second-lowest values by the Deep AR model using only static external values, as illustrated by figure A.74.

For time series 721 the Holt Winter's model performed the worst, followed closely by both of Prophets' models' variations (with and without external variables).

The highest MASE values in the case of time series 1675 were given by both Prophet models. The Holt Winter's and the two SARIMAX models had a similar performance.

	MASE		MASE
HW - 721	1.096	HW - 1675	0.886
SARIMA - 721	0.801	SARIMA - 1675	0.853
SARIMAX - 721	0.809	SARIMAX - 1675	0.861
Prophet - 721	0.994	Prophet - 1675	1.333
ProphetX - 721	0.993	ProphetX - 1675	1.381
DeepAR - 721	0.727	DeepAR - 1675	0.546
DeepAR-FDR - 721	0.551	DeepAR-FDR - 1675	0.546

Figure 4.31: Metrics for each models forecasts for each time series from Medium average group

The forecasts and prediction intervals for time series 1675, for each model are illustrated in figures, 4.32,4.33, 4.34,4.35,4.36,4.37,4.38. The generated prediction intervals are more accurate than in the previous group, once more the Deep AR model produced the most precise intervals, and both the SARIMAX and Prophet models produced the most accurate. Overall the best prediction intervals were generated by the SARIMAX model, even if less precise than Deep AR's.

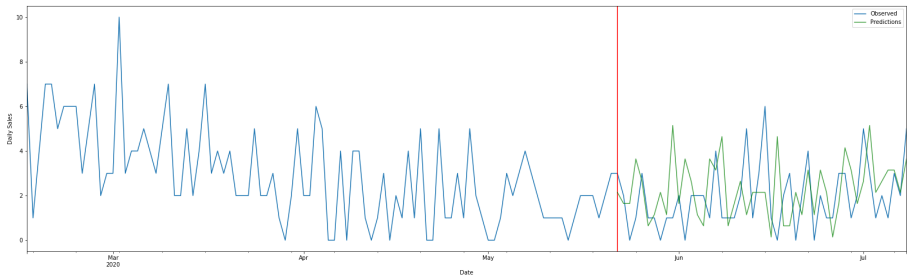

Figure 4.32: Holt Winter's forecasts and actual values for time series 1675

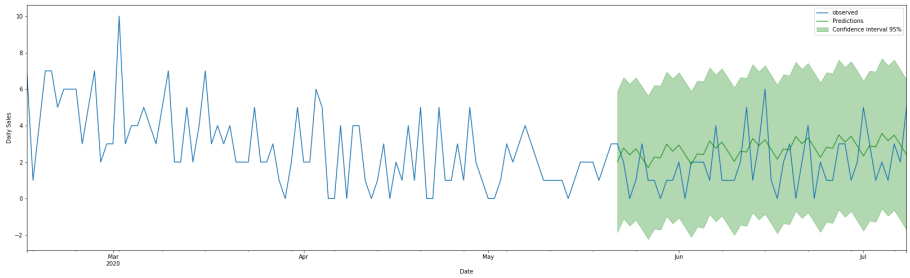

Figure 4.33: SARIMA forecasts, prediction intervals and actual values for time series 1675

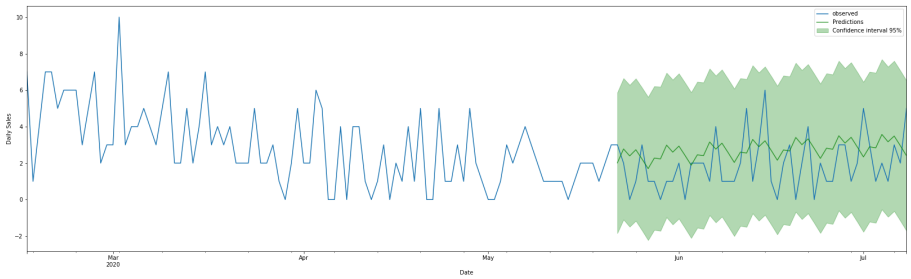

Figure 4.34: SARIMAX forecasts, prediction intervals and actual values for time series 1675

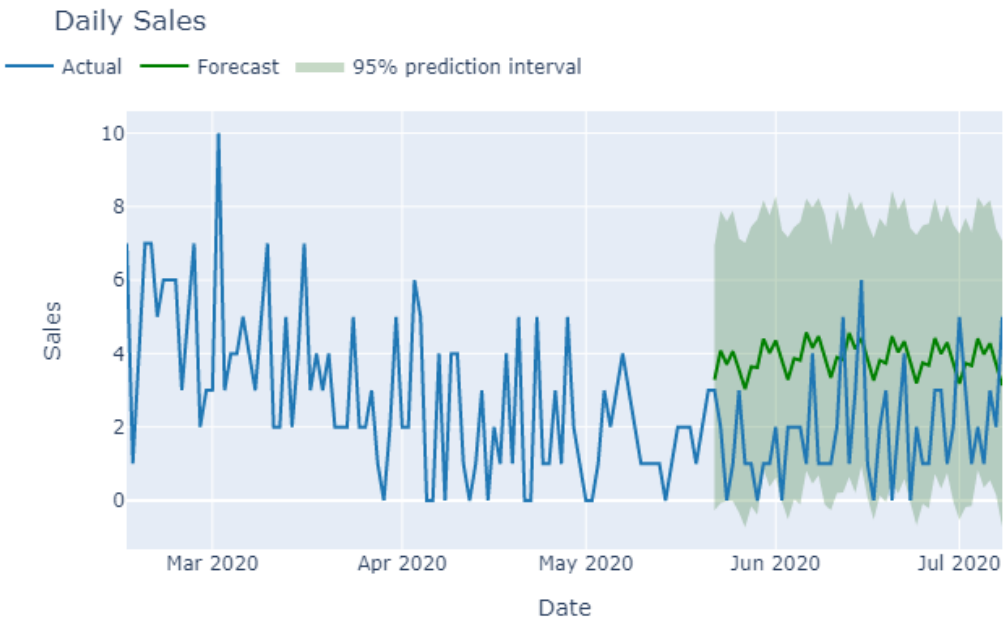

Figure 4.35: Prophet forecasts, prediction intervals and actual values for time series 1675

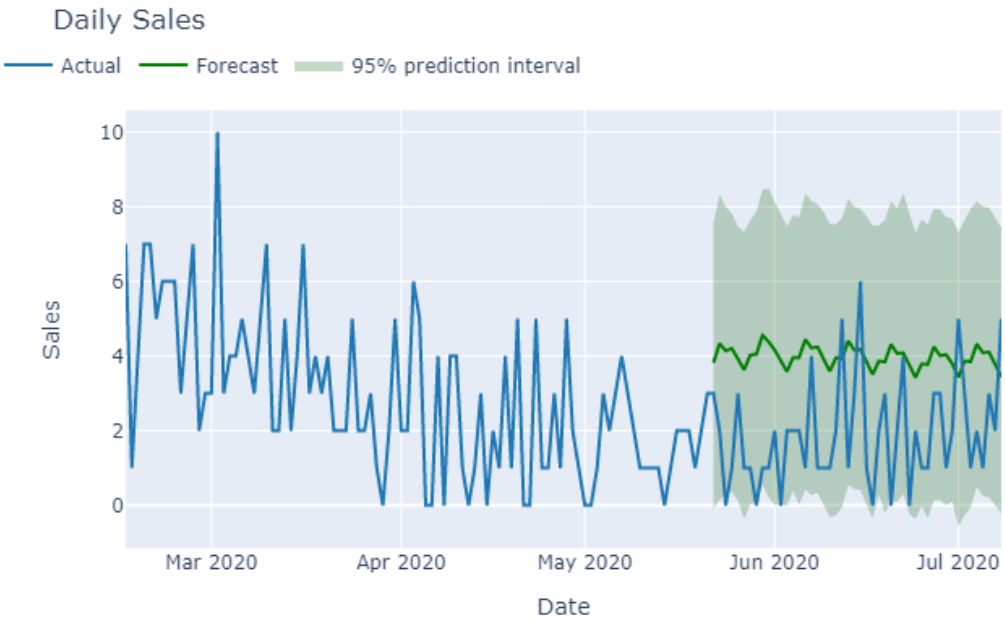

Figure 4.36: ProphetX forecasts, prediction intervals and actual values for time series 1675

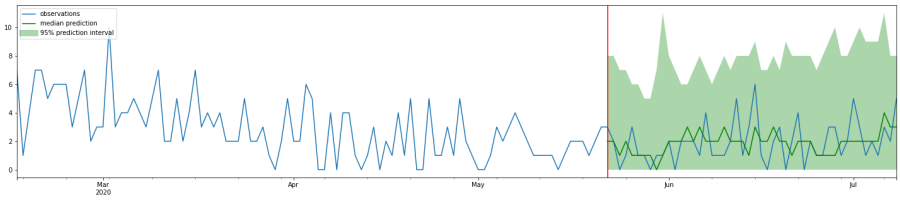

Figure 4.37: Deep AR forecasts, prediction intervals and actual values for time series 1675

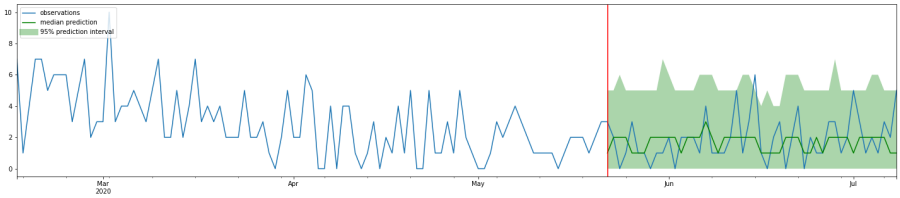

Figure 4.38: Deep AR - FDR forecasts, prediction intervals and actual values for time series 1675

4.2.4 Forecasts and Metrics for High Average Group

As illustrated by figure 4.39, overall the Deep AR models were responsible for the lowest MASE, except time series 1020 where the Deep AR model not incorporating dynamic features displayed the second-worst performance.

The second-lowest MASE measurements were given by the SARIMA model, which consistently outperformed the SARIMAX model.

In general, the highest MASE values were reached by the Holt Winter's model, except time series 1020, where the two worst performers were both Prophet models.

	MASE		MASE		MASE
HW - 1673	1.365	HW - 1020	0.809	HW - 1023	0.937
SARIMA - 1673	0.984	SARIMA - 1020	0.688	SARIMA - 1023	0.618
SARIMAX - 1673	1.156	SARIMAX - 1020	0.744	SARIMAX - 1023	0.627
Prophet - 1673	0.898	Prophet - 1020	0.680	Prophet - 1023	0.719
ProphetX - 1673	0.861	ProphetX - 1020	0.882	ProphetX - 1023	0.763
DeepAR - 1673	0.726	DeepAR - 1020	0.813	DeepAR - 1023	0.562
DeepAR-FDR - 1673	0.732	DeepAR-FDR - 1020	0.473	DeepAR-FDR - 1023	0.559

	MASE		MASE
HW - 1460	1.392	HW - 124	0.963
SARIMA - 1460	0.744	SARIMA - 124	0.691
SARIMAX - 1460	0.841	SARIMAX - 124	0.715
Prophet - 1460	0.953	Prophet - 124	0.615
ProphetX - 1460	0.941	ProphetX - 124	0.669
DeepAR - 1460	0.575	DeepAR - 124	0.541
DeepAR-FDR - 1460	0.491	DeepAR-FDR - 124	0.554

Figure 4.39: Metrics for each models forecasts for each time series from High average group

The forecasts and prediction intervals for time series 1020, for each model are illustrated in figures, 4.40,4.41, 4.42,4.43,4.44,4.45,4.46. The prediction intervals had similar accuracy however once more the Deep AR model produced the most precise intervals and the Deep AR model that did not incorporate dynamic features generated the most accurate intervals.

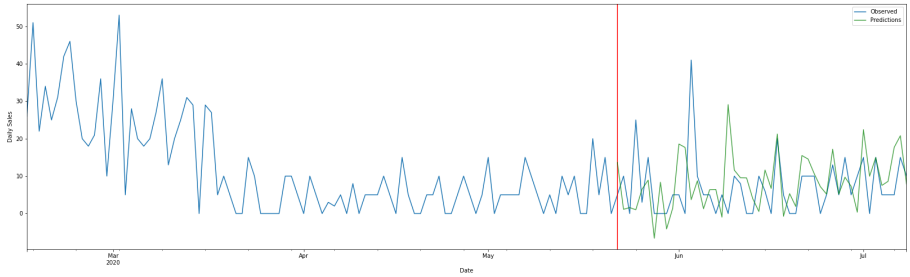

Figure 4.40: Holt Winter's forecasts and actual values for time series 1020

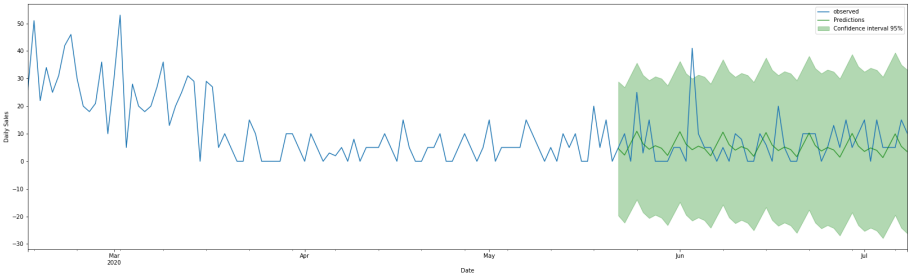

Figure 4.41: SARIMA forecasts, prediction intervals and actual values for time series 1020

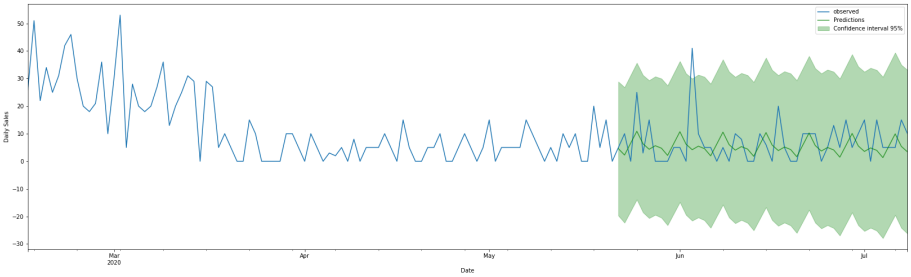

Figure 4.42: SARIMAX forecasts, prediction intervals and actual values for time series 1020

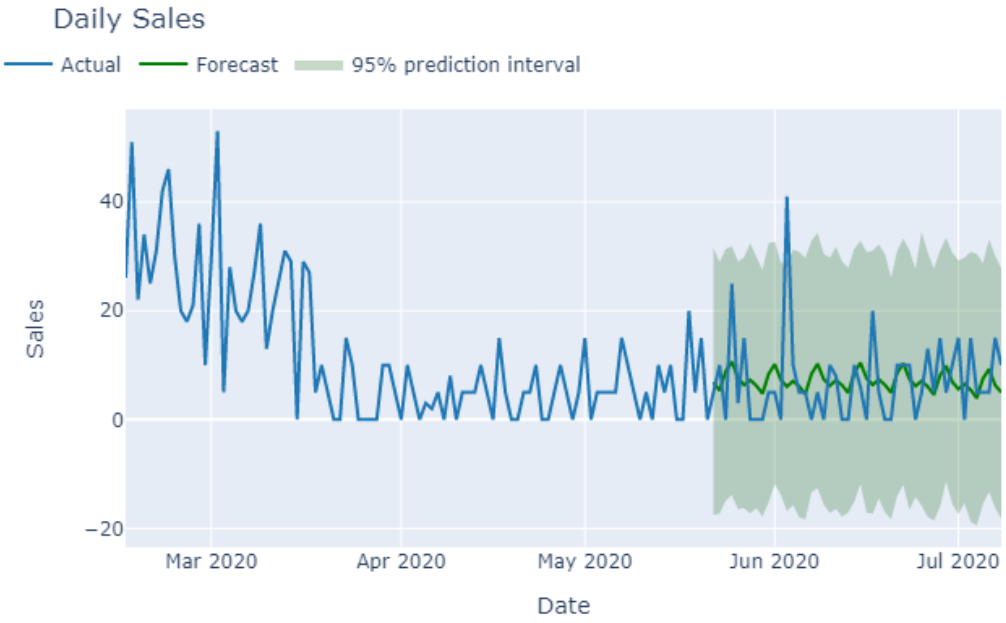

Figure 4.43: Prophet forecasts, prediction intervals and actual values for time series 1020

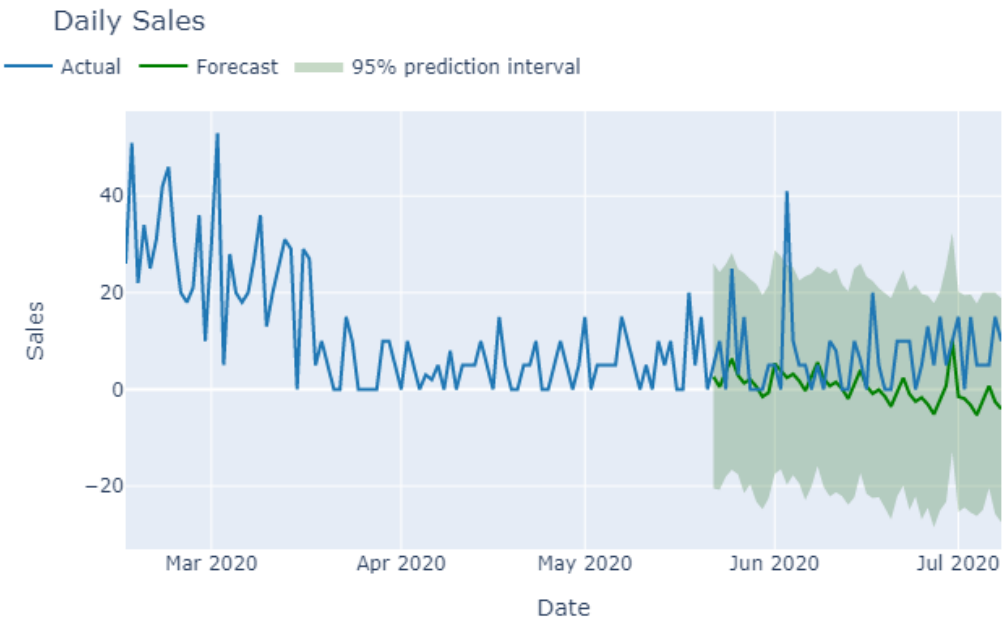

Figure 4.44: ProphetX forecasts, prediction intervals and actual values for time series 1020

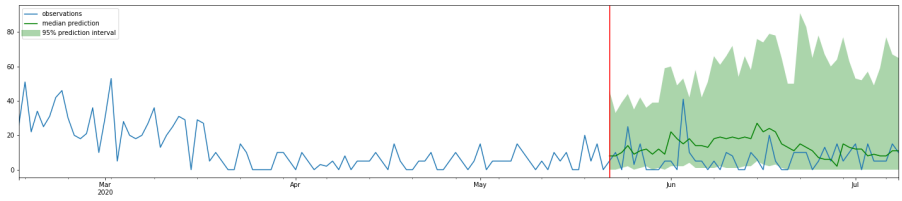

Figure 4.45: Deep AR forecasts, prediction intervals and actual values for time series 1020

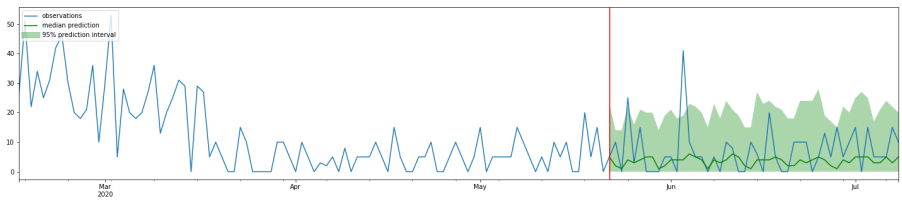

Figure 4.46: Deep AR - FDR forecasts, prediction intervals and actual values for time series 1020

Chapter 5

Conclusions

We compared the Deep AR approach to the problem of sales forecasting with traditional and state of the art alternatives. This comparison was done for two structurally distinct data sets and the MASE metric was selected to measure performance, based on the available literature and recommendation of multiple data scientists [29].

The Rossmann data set contained the daily sales history in euros, aggregated by store, for 1115 stores, over a period of 990 days. The majority of time series from this data set can be characterized by their high seasonality, and low intermittency (most stores were closed on Sunday). A total of 16 time series were selected from this data set, based on the median (more robust in the presence of outliers than the average) daily sales values and were then grouped in 4's as *low*, *low-medium*, *medium* and *high* median daily sales groups (Min-Q1, Q1-Q2, Q2-Q3,Q3-Max). These groupings allowed us to assess how the models performed overall between groups (low and medium-low median sales groups, small stores with less traffic, medium and high median sales groups, larger stores with more traffic), as well as assessing performance between time series of the same group. The models used the first 942 days as training input and the last 48 days were used for testing. After removing the time series that had missing values, the Deep AR model was supplied with a total of 932 time series for training. Considering a forecast horizon of 48 days, the Deep AR model displayed the best performance for both the *low* and *medium-low* median daily sales groups, achieving the lowest average MASE measurements (0.208 and 0.184). The SARIMAX model performed the best for the medium and high median daily sales group (MASE = 0.202 and 0.353). The Holt Winter's model consistently reached the highest MASE values and thus displayed the worst fit. One important take away from this experiment is that when supplying multiple time series as input to the Deep AR model, time series with vastly different generating processes (structure) will generate less accurate predictions. This was observed in the results from experiments on the group of stores with the highest medians of daily sales in particular for store 639. One way to overcome this is to train a separate model for the individual time series.

The second data set, from Retail Consult, contained the history of the daily sales volume, in product units, for a total of 195801 product-store pairs, over a period of 898 days. The time series in this data set were highly intermittent, non-seasonal, and the daily sales volumes were

consistently small, which is expected for slow-moving products, and thus, the target variable had a small range (around 1 to 5 daily sales), which generally leads to larger forecasting errors. From this data set 7 time series were selected and arranged in groups called *low*, *medium* and *high* average daily sales groups. The models used the first 850 days as training input and the last 48 days were used for testing, and the Deep AR model was supplied with a total of 2000 time series for training. For this case, we trained two SARIMAX, Prophet, and Deep AR models. The SARIMAX and Prophet models were trained with and without external variables (SARIMAX, ProphetX, SARIMA, and Prophet). The two Deep AR models were both supplied with external variables, however, one included all available features (dynamic and static) and the other static features only (Deep AR FDR and Deep AR respectively). Once more we set a forecast horizon of 48 days and, overall, the lowest average MASE values were achieved by both Deep AR models, the Deep AR FDR variation displaying a slightly lower MASE value(0.661 and 0.695). The second best results were given by the SARIMA and SARIMAX models respectively (0.760 and 0.802). Additionally, the Deep AR models generated the most accurate and precise prediction intervals for the majority of time series from both data sets, a highly valuable characteristic when considering stock management decision making.

The Gluon TS toolkit from which the Deep AR model was used is still in development, and as such, some valuable training features aren't yet available. For example, when trying to use a validation data set for training to compute the validation loss, if the training input consists of an individual or a small number of time series, it will throw an error. For our purposes, this did not present a problem as in both experiments over 900 time series were used as training input. Furthermore, the developer team is actively updating the tool kit. More in-depth data analysis and feature engineering should be performed on the Retail Dataset to improve performance. Given the very low number of observations of features, aggregating the products by class, subclass or department, for each store is a possible alternative to improve the prediction's accuracy while not losing as much information as when aggregating exclusively by store. Additionally, the data anonymization limited the analysis and feature engineering process. For example, if the city or region features weren't anonymized, it would have been possible to gather weather data which could further improve results. Additional features like google analytics, competition-related data, price, and relative price discount values are also important features that should be considered in the future if available. We'd also be interested in seeing future work regarding the Deep AR model using a wider range of accuracy measurements to provide a more complete view of the models' performance, such as CFE (Cumulative Forecast Error) paired with metrics like PIS (Periods in Stock) and SPEC (Stock-keeping-oriented Prediction Error Costs).

Our experiments showed that the Deep AR approach can be a highly suitable solution to the problem of daily sales forecasting when supplied with a large training set containing multiple time series, in particular for intermittent and slow-moving products. Additionally, we observed that it benefited from the inclusion of multiple external variables such as different promotions and events. It exhibited the best overall performance in the case of seasonal data and highly intermittent data and generated the most accurate and precise prediction intervals, and as such, we consider that

it displays the desirable traits of a valuable stock management and revenue optimization tool for healthcare and wellness retailers.

Appendix A

Additional literature review

Table A.1: Additional studies comparing forecasting methods 1

Reference	**Focus**	**Data: range and granularity in the product, location and time dimensions**	**Variables**	**Forecast horizon and evaluation**	**Baseline Methods**	**Issues and Limitations**	**Conclusion**
Ainscough and Aronson (1999)[4]	Examines neural networks as an alternative to traditional statistical methods for the analysis of scanner data	A national brand of strawberry yogurt, 575 weeks of sales data are aggregated from 6 stores	Lagged sales; Price, display, feature, and their two-way interactions.	20% of the sample as the validation set (appox. 20 weeks); MSE	OLS regression	The limitations of the study were the limitations of NN's.	It was found that three- and four-layer neural networks yielded significantly better predictions than OLS regression.
Taylor (2007)[49]	Forecasting daily supermarket sales using exponentially weighted quantile regression	256 time series of daily sales with median daily sales greater or equal to 5, in length from 72 to 1436 observations	None.	Horizon 1 to 14 days; 42,633 post-sample sales observations; Relative MAE, and coverage measures for prediction intervals.	Univariate models (Naïve, Holt's, HW, etc.) and company procedure.	Including trend or seasonal terms in the EWQR led to poor results	Exponentially Weighted Quantile Regression (EWQR) compared favorably with a variety of other univariate methods. Evaluates robust adaptions of EWQR.
Ramos and Fildes (2017)[41]	Comparative study on univariate and multivariate forecasting methods on storexSKU data	988 SKUs in 203 categories: intermittent demand excluded, 173 weeks	Seasonal data, holidays, promotional events	Horizon 1 week ahead. Rolling origin calculation with updated model parameters. MAPE, MdAPE, MRMAE, MdRMAE, GMRMAE, MRRMSE, MdRRMSE, GMRRMSE ,MASE, MdASE.	ETS, TBATS, ARIMA, Naïve: compared to multivariate alternatives & Lasso	Individual selection has potential to improve	Multivariate model including promotions performed best with TBATS the best of the univariate methods considered.
Lang et al. (2015)[30]	A hierarchical Bayesian semi-parametric approach to account simultaneously for heterogeneity and functional flexibility (non-linearity) in store sales models.	Weekly store-level scanner data for eight brands of orange juice .	The lowest price of a competing national (premium) brand in store s and week t, 11 other covariates such as age, education, family size, income, distances to nearest competing supermarkets or warehouses.	Predictive performance of the competing models measured by the Average Root Mean Squared Error (ARMSE) in holdout samples.	Extended flexible heterogeneous model (EFHetM) and other models, nested in EFHetM.	Only one validation method used. .	The new model class reveals improved predictive performance compared to a series of competing models. Allowing for multiplicative heterogeneity in addition to functional flexibility can improve the predictive validity of a store sales model.

Table A.2: Additional studies comparing forecasting methods 1

Reference	Focus	Data: range and granularity in the product, location and time dimensions	Variables	Forecast horizon and evaluation	Baseline Methods	Issues and Limitations	Conclusion
Gur Ali (2013)[23]	Introducing a new method, the Driver Moderator method for sales prediction, in the presence of promotions for existing and new products.	Daily unit sales, price, and feature information for five stores, and 115 products from September 6, 2006 to September 30, 2008. Weekly unit sales, prices, and display and feature information for 38 stores from five grocery store chains with 1020 total products, from 2002 to 2005.	Unit sales, average unit sales in the normalization period, moderator variable, driver variable, the main effect parameter for driver variable d, the interaction effect.	MAE and MASE.	Driver Moderator method, individual regressions, exponential smoothing, regression trees, stepwise regression, neural networks.	Data from two specific companies and categories, using a specific historical window to train the datasets. The proposed model requires high computational complexity and data collection and maintenance costs compared with the simple time series approaches.	The Driver Moderator method compared with more flexible learning algorithms, such as regression trees or neural networks, using similar variables, maintains comparable accuracy while providing much more interpretable models.
Hasin et al. (2011)[25]	Comparing traditional sales forecasting methods to ANN.	Five years data of a very fast moving item, Noodles.	Weekend, holiday, festival period, promotional activity, availability on shelf, price range, first or second half of the month, consumption rate, brand loyalty, climate or the season.	MAPE.	Holt Winter's model and fuzzy neural networks.	Lacking in evaluation and error measures.	The error level in advanced forecasting techniques such as HW's model increase in forecast frequency or alternatively decrease in forecast periods. In such a situation, fuzzy artificial neural network (ANN) can provide a better solution.
Doganis et al. (2006)[18]	Examining time series sales forecasting for short shelf-life food products based on artificial neural networks and evolutionary computing	Daily sales data of 11 milk pack for the first few months of the years 2001 and 2002 (108 days), provided by a leading manufacturer.	The six previous days of the current year, the six previous days of last year, the percentile change between the two years and the corresponding day of the previous year.	MAE.	Linear AR, HW, Neural Networks.	More validation and error measures. Promotional effects not considered.	The proposed models produced an accurate forecast. The performance of the methods can be further improved if adaptation capabilities are added to the neural network model, by accounting for recent incidents that have not been considered in originally.
Pinho (2015)[38]	To improve sales forecasting by incorporating promotional variables.	Daily sales data from 129081 products from Pingo Doce, 65% being textile products.	Price, number of days with promotion and lag of 1 and 2 períods of these variables, calendar events, last week of the month, Fourier terms, PCA of the price of products from the same category, PCA of the number of days with promotion of products from the same category and lag of 1 and 2 periods of these PCAs. 30 weeks 17,5% of total observations .	RMSE, MAE, MAPE and MASE.	ARIMA, Dynamic Regression.	The study only forecasts for a fixed forecast horizon of 30 weeks. No inter-category information and not using more error measures.	On average the dynamic regression model offer more accurate forecasts for this kind of forecasting.

Appendix A

List of Results and Plots

A.1 Rossmann Experiment

A.1.1 Metrics

	sMAPE	MASE	RMSE
Prophet - Q1 - Average	41.170	0.315	512.859
HW - Q1 - Average	70.880	1.086	1856.066
SARIMAX - Q1 - Average	42.204	0.321	500.847
DeepAR - Q1 - Average	9.700	0.208	418.853

	sMAPE	MASE	RMSE
Prophet - Q2 - Average	36.868	0.222	679.830
HW - Q2 - Average	59.476	0.999	3233.034
SARIMAX - Q2 - Average	35.779	0.186	583.541
DeepAR - Q2 - Average	7.400	0.184	650.304

	sMAPE	MASE	RMSE
Prophet - Q3 - Average	36.889	0.222	762.280
HW - Q3 - Average	60.846	1.070	3952.591
SARIMAX - Q3 - Average	36.198	0.202	677.465
DeepAR - Q3 - Average	9.700	0.211	961.246

	sMAPE	MASE	RMSE
Prophet - Q4 - Average	21.767	0.383	1210.855
HW - Q4 - Average	37.815	1.013	4163.109
SARIMAX - Q4 - Average	21.072	0.353	1061.676
DeepAR - Q4 - Average	9.500	0.557	1470.972

Figure A.1: Average metrics for each model for each quartil

	sMAPE	MASE	RMSE
Prophet - Q1 - STD	2.131	0.032	58.733
HW - Q1 - STD	7.258	0.029	217.017
SARIMAX - Q1 - STD	2.917	0.067	111.333
DeepAR - Q1 - STD	2.419	0.047	103.416

	sMAPE	MASE	RMSE
Prophet - Q2 - STD	0.894	0.032	120.202
HW - Q2 - STD	5.958	0.028	154.970
SARIMAX - Q2 - STD	0.747	0.028	103.614
DeepAR - Q2 - STD	0.792	0.031	113.745

	sMAPE	MASE	RMSE
Prophet - Q3 - STD	0.467	0.025	69.891
HW - Q3 - STD	5.160	0.134	286.152
SARIMAX - Q3 - STD	0.810	0.035	23.625
DeepAR - Q3 - STD	2.750	0.047	265.121

	sMAPE	MASE	RMSE
Prophet - Q4 - STD	14.362	0.211	212.137
HW - Q4 - STD	24.933	0.128	2204.132
SARIMAX - Q4 - STD	14.113	0.207	154.141
DeepAR - Q4 - STD	2.126	0.453	252.374

Figure A.2: Standard Deviation for each metric for each model for each quartil

	sMAPE	MASE	RMSE
Prophet - 157	44.143	0.299	504.367
HW - 157	81.115	1.054	1938.108
SARIMAX - 157	40.899	0.232	406.468
DeepAR - 157	12.400	0.237	498.288

	sMAPE	MASE	RMSE
Prophet - 453	42.171	0.370	584.764
HW - 453	68.785	1.067	1716.765
SARIMAX - 453	47.042	0.411	526.713
DeepAR - 453	11.900	0.259	525.122

	sMAPE	MASE	RMSE
Prophet - 697	38.719	0.291	538.361
HW - 697	60.959	1.093	2168.078
SARIMAX - 697	41.596	0.352	672.319
DeepAR - 697	7.700	0.201	388.988

	sMAPE	MASE	RMSE
Prophet - 701	39.648	0.300	423.946
HW - 701	72.659	1.129	1601.313
SARIMAX - 701	39.278	0.290	397.886
DeepAR - 701	7.000	0.134	263.017

Figure A.3: Metrics for each time series of low median group (Q1)

	sMAPE	MASE	RMSE
Prophet - 177	36.172	0.218	588.895
HW - 177	53.332	0.955	3086.749
SARIMAX - 177	36.702	0.222	666.246
DeepAR - 177	8.000	0.214	688.041

	sMAPE	MASE	RMSE
Prophet - 401	37.088	0.209	633.979
HW - 401	68.621	1.003	3360.513
SARIMAX - 401	35.231	0.164	489.388
DeepAR - 401	6.400	0.140	513.076

	sMAPE	MASE	RMSE
Prophet - 112	38.236	0.273	886.186
HW - 112	55.138	1.004	3413.077
SARIMAX - 112	36.302	0.203	705.826
DeepAR - 112	8.400	0.212	815.261

	sMAPE	MASE	RMSE
Prophet - 676	35.976	0.186	610.262
HW - 676	60.813	1.033	3071.797
SARIMAX - 676	34.881	0.155	472.703
DeepAR - 676	7.000	0.170	584.840

Figure A.4: Metrics for each time series of medium-low median group (Q2)

	sMAPE	MASE	RMSE
Prophet - 850	36.607	0.241	659.050
HW - 850	64.670	1.283	3556.948
SARIMAX - 850	37.480	0.255	705.139
DeepAR - 850	7.800	0.173	646.364

	sMAPE	MASE	RMSE
Prophet - 130	36.969	0.234	749.403
HW - 130	56.097	1.065	3803.024
SARIMAX - 130	36.103	0.208	660.333
DeepAR - 130	6.300	0.157	751.082

	sMAPE	MASE	RMSE
Prophet - 384	36.370	0.178	788.896
HW - 384	55.441	0.917	4224.374
SARIMAX - 384	35.970	0.161	695.832
DeepAR - 384	11.900	0.253	1223.256

	sMAPE	MASE	RMSE
Prophet - 113	37.609	0.234	851.769
HW - 113	67.178	1.014	4226.018
SARIMAX - 113	35.239	0.184	648.555
DeepAR - 113	12.900	0.262	1224.283

Figure A.5: Metrics for each time series of medium median group (Q3)

	sMAPE	MASE	RMSE
Prophet - 639	7.656	0.721	1281.365
HW - 639	9.758	0.922	1420.995
SARIMAX - 639	7.169	0.683	1195.965
DeepAR - 639	12.700	1.295	1848.533

	sMAPE	MASE	RMSE
Prophet - 355	7.216	0.393	1273.934
HW - 355	16.283	0.875	2634.192
SARIMAX - 355	6.756	0.372	1138.767
DeepAR - 355	9.800	0.557	1533.811

	sMAPE	MASE	RMSE
Prophet - 691	37.418	0.242	1428.782
HW - 691	64.359	1.206	6775.674
SARIMAX - 691	35.611	0.197	1112.063
DeepAR - 691	6.900	0.157	1175.624

	sMAPE	MASE	RMSE
Prophet - 360	34.777	0.174	859.337
HW - 360	60.861	1.047	5821.577
SARIMAX - 360	34.750	0.160	799.908
DeepAR - 360	8.500	0.217	1325.922

Figure A.6: Metrics for each time series of high median group (Q4)

A.1.2 Plots

A.1.2.1 Q1

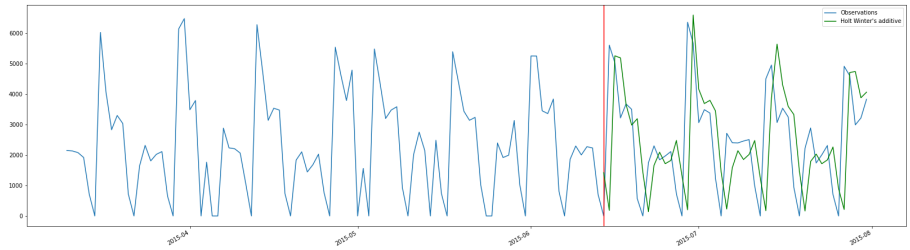

Figure A.7: Holt Winter's Forecasts for time series 157

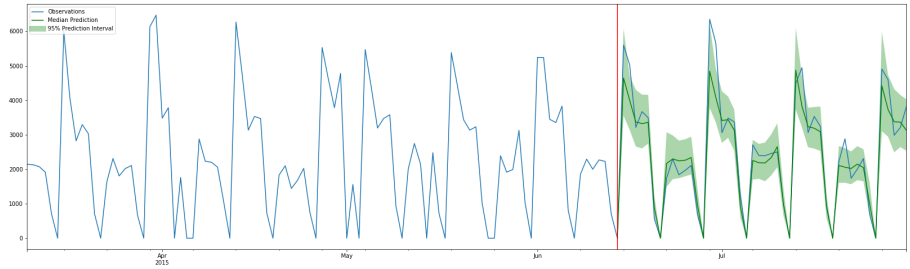

Figure A.8: Deep AR's Forecasts for time series 157

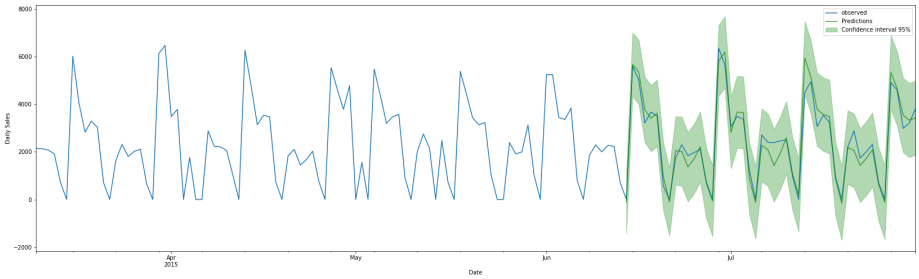

Figure A.9: SARIMAX's Forecasts for time series 157

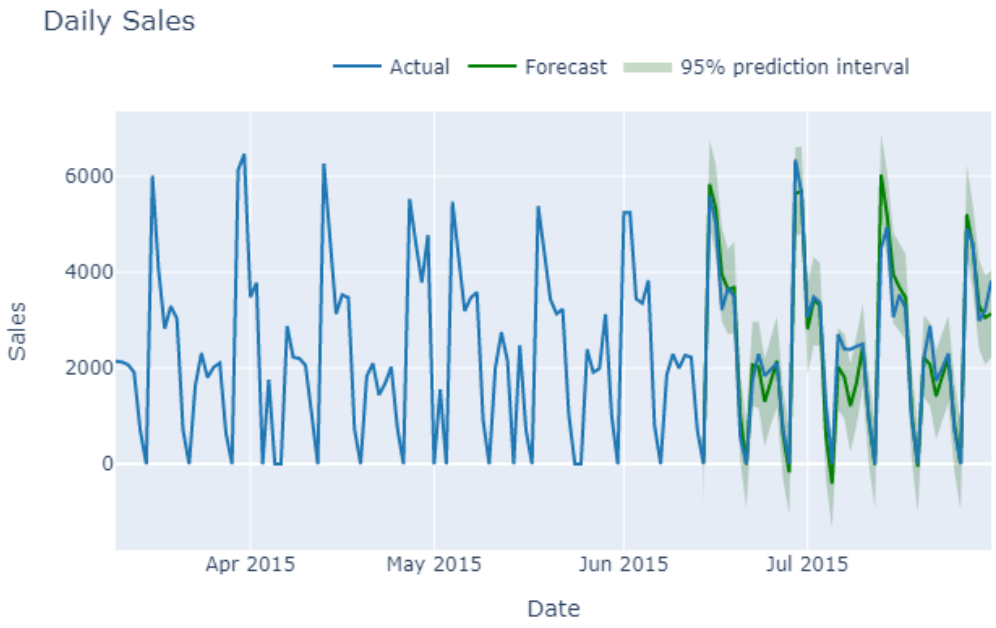

Figure A.10: Prophet's Forecasts for time series 157

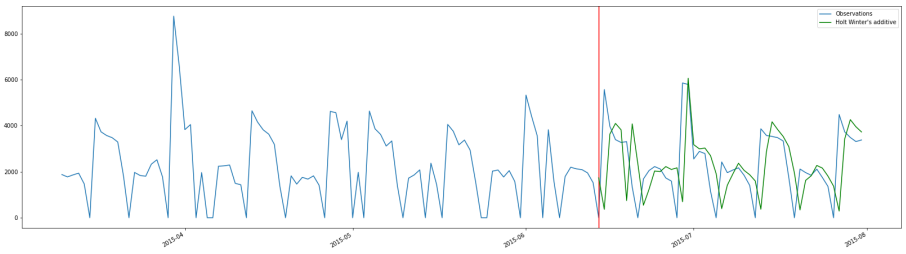

Figure A.11: Holt Winter's Forecasts for time series 453

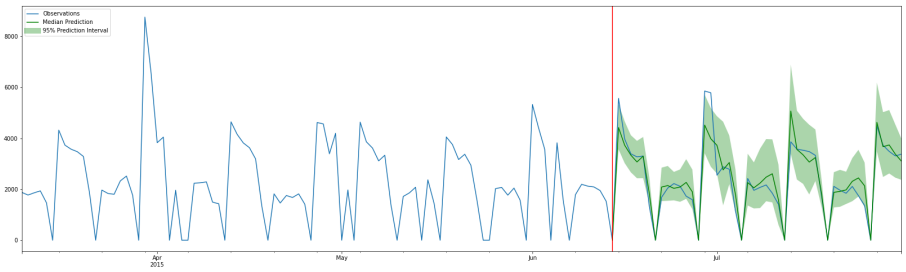

Figure A.12: Deep AR's Forecasts for time series 453

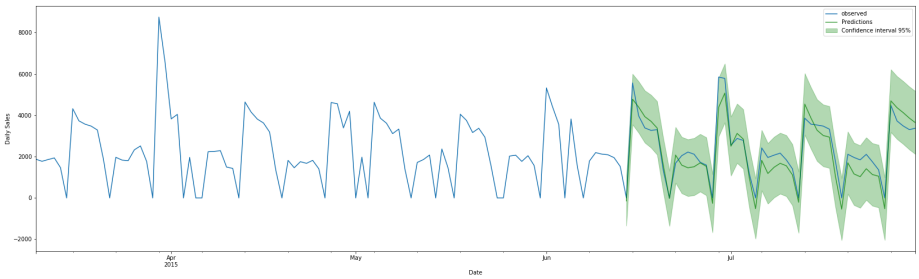

Figure A.13: SARIMAX's Forecasts for time series 453

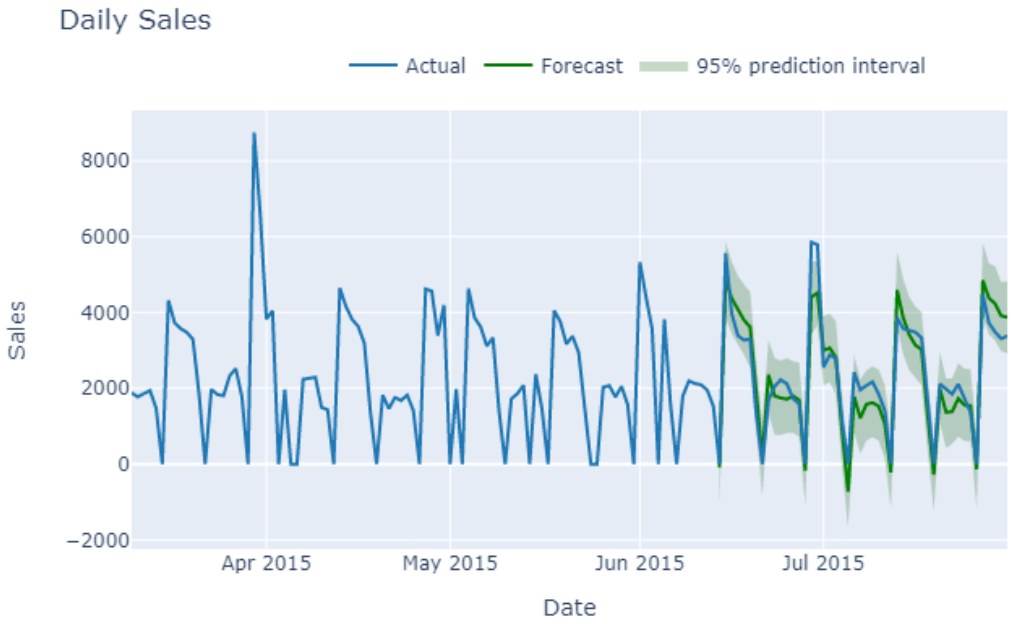

Figure A.14: Prophet's Forecasts for time series 453

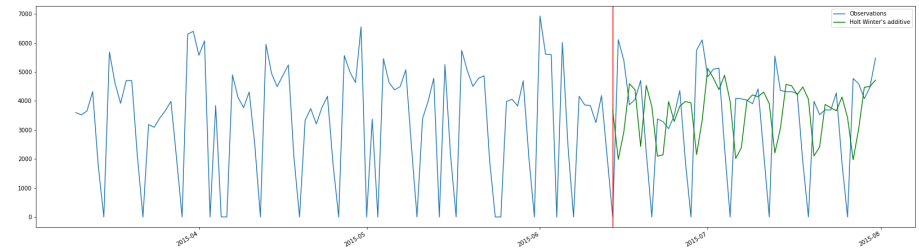

Figure A.15: Holt Winter's Forecasts for time series 697

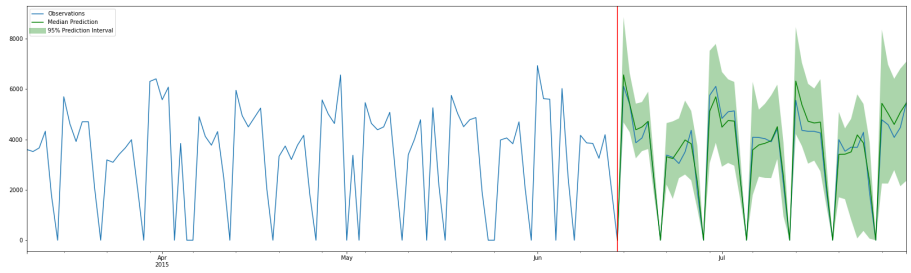

Figure A.16: Deep AR's Forecasts for time series 697

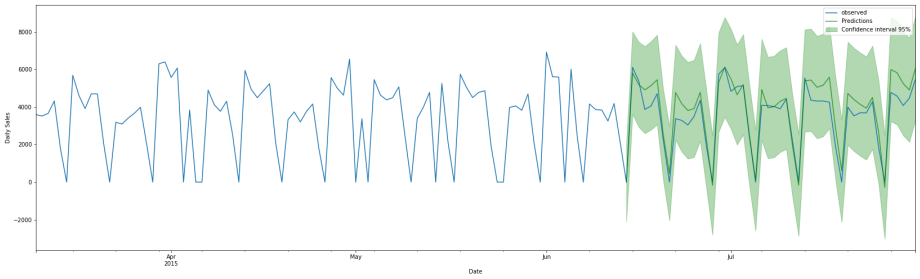

Figure A.17: SARIMAX's Forecasts for time series 697

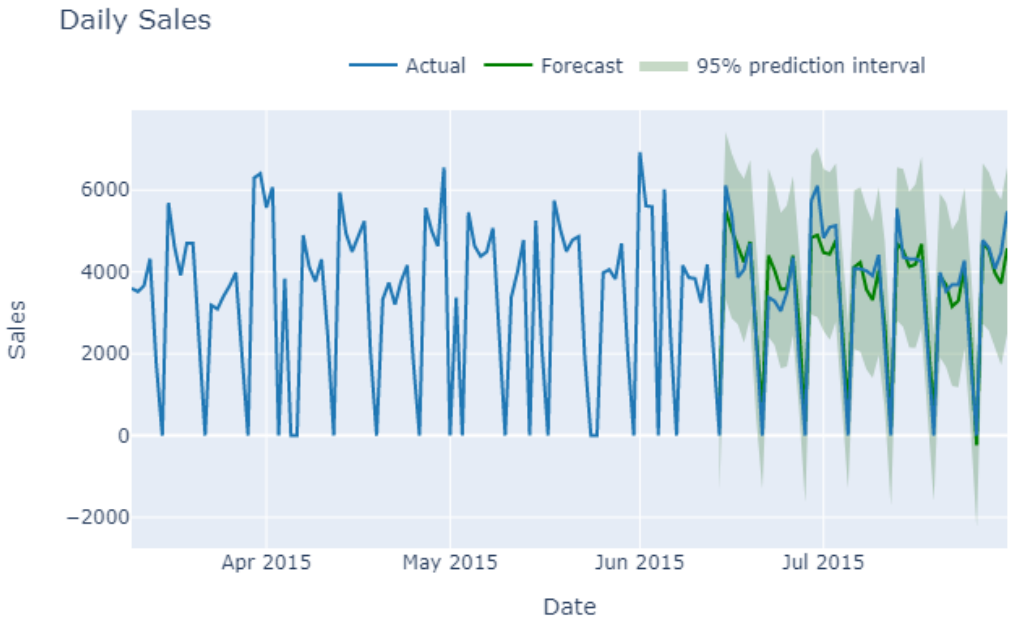

Figure A.18: Prophet's Forecasts for time series 697

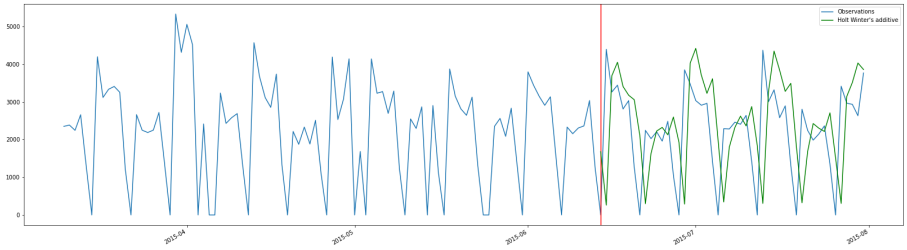

Figure A.19: Holt Winter's Forecasts for time series 701

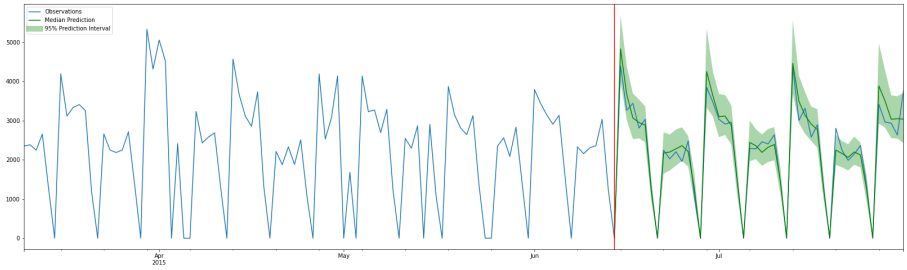

Figure A.20: Deep AR's Forecasts for time series 701

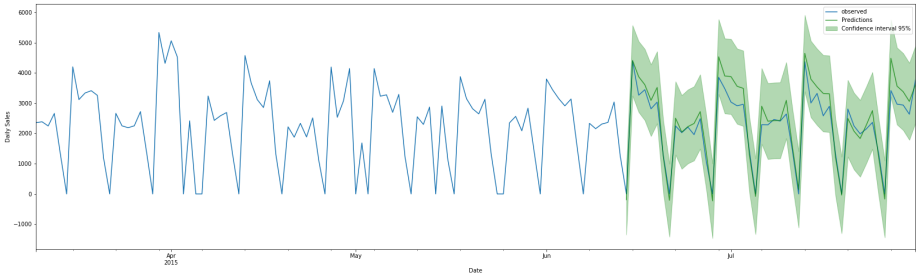

Figure A.21: SARIMAX's Forecasts for time series 701

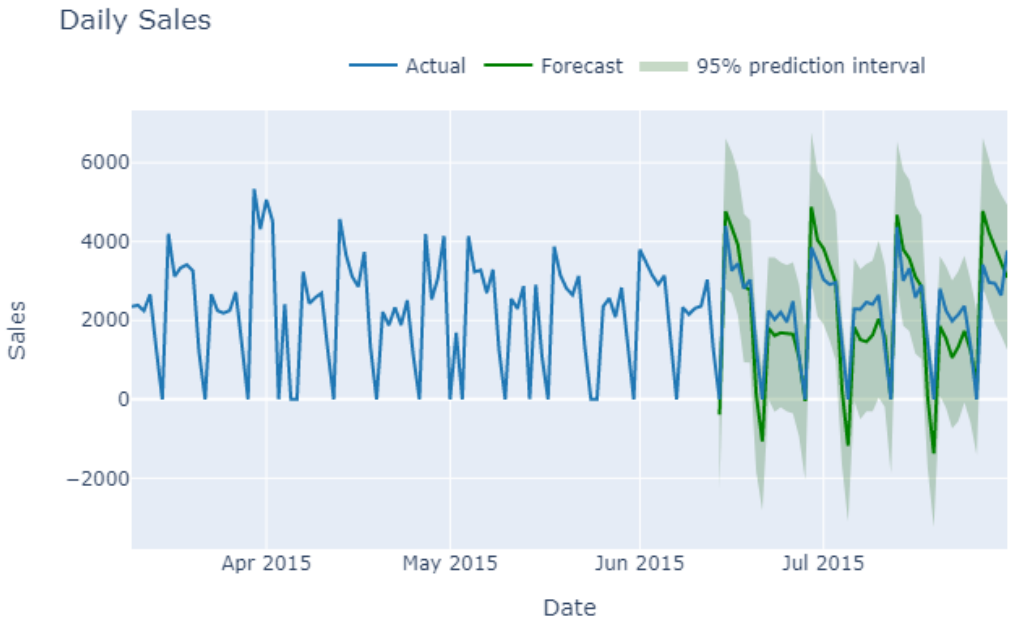

Figure A.22: Prophet's Forecasts for time series 701

A.1.2.2 Q2

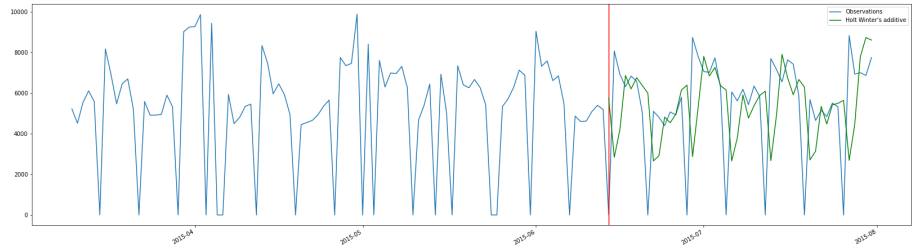

Figure A.23: Holt Winter's Forecasts for time series 177

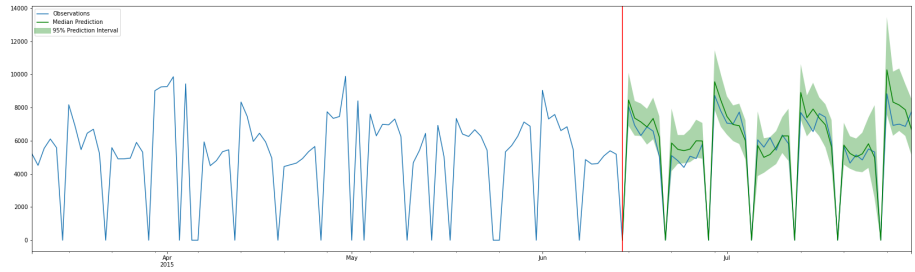

Figure A.24: Deep AR's Forecasts for time series 177

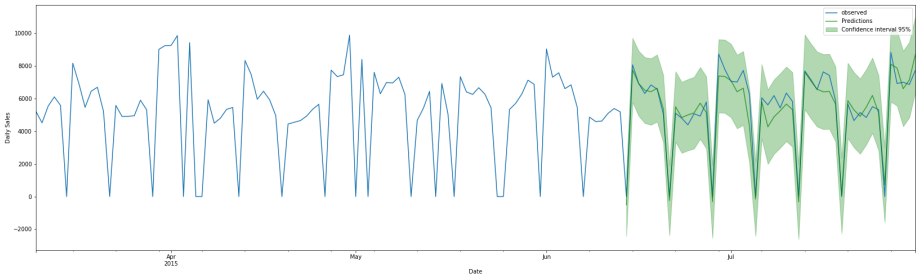

Figure A.25: SARIMAX's Forecasts for time series 177

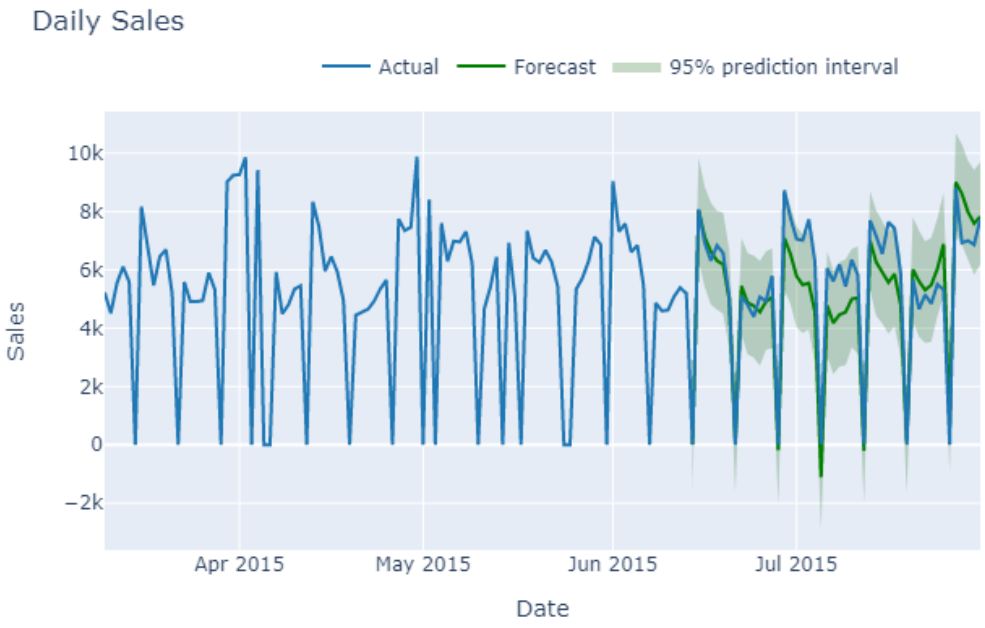

Figure A.26: Prophet's Forecasts for time series 177

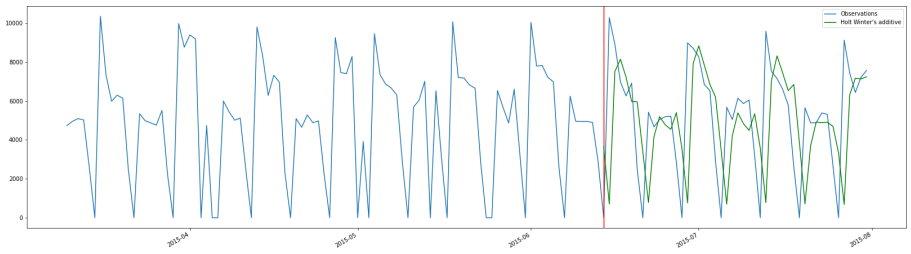

Figure A.27: Holt Winter's Forecasts for time series 401

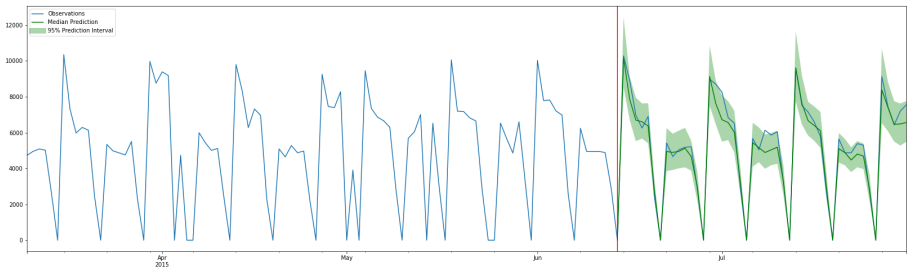

Figure A.28: Deep AR's Forecasts for time series 401

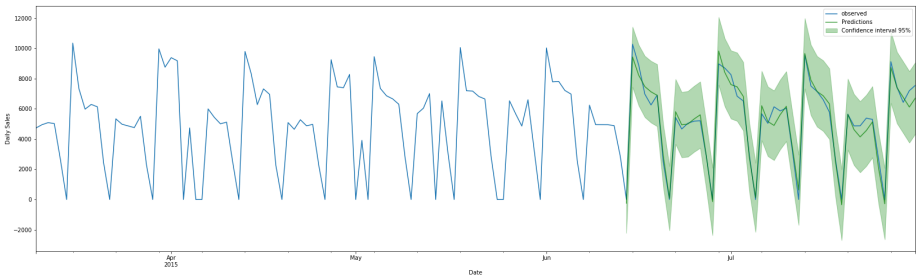

Figure A.29: SARIMAX's Forecasts for time series 401

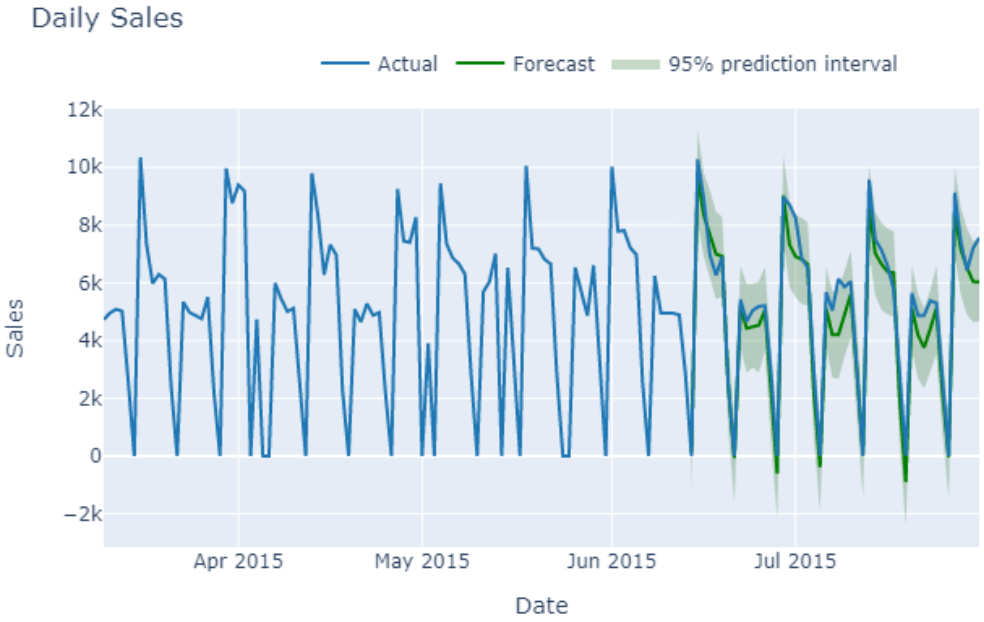

Figure A.30: Prophet's Forecasts for time series 401

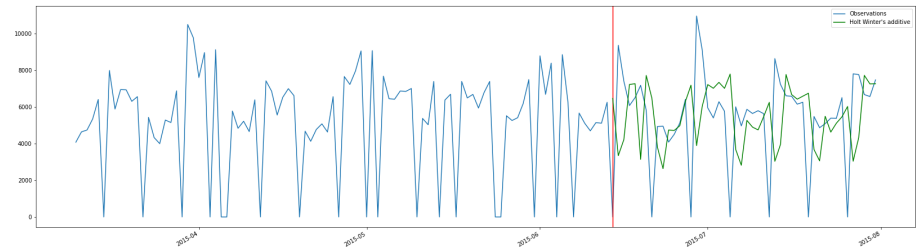

Figure A.31: Holt Winter's Forecasts for time series 112

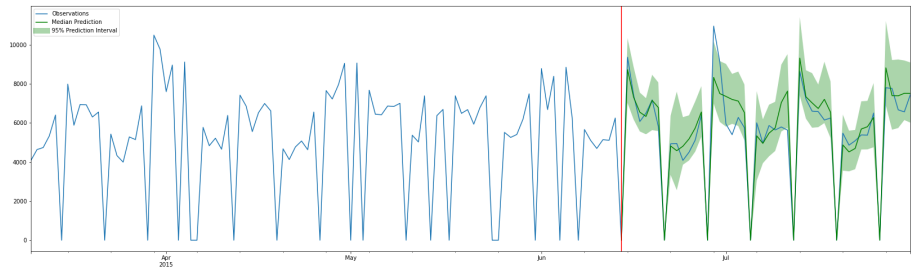

Figure A.32: Deep AR's Forecasts for time series 112

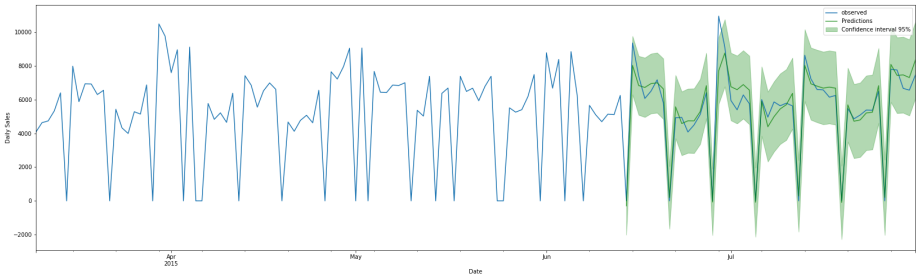

Figure A.33: SARIMAX's Forecasts for time series 112

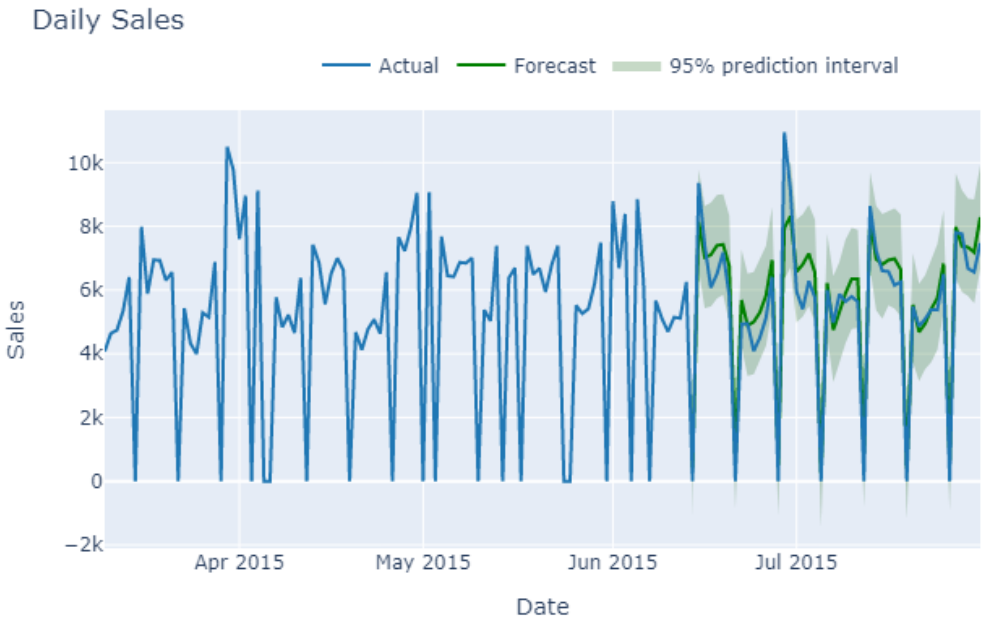

Figure A.34: Prophet's Forecasts for time series 112

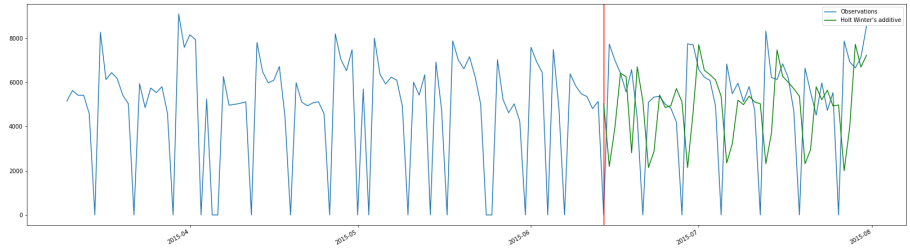

Figure A.35: Holt Winter's Forecasts for time series 676

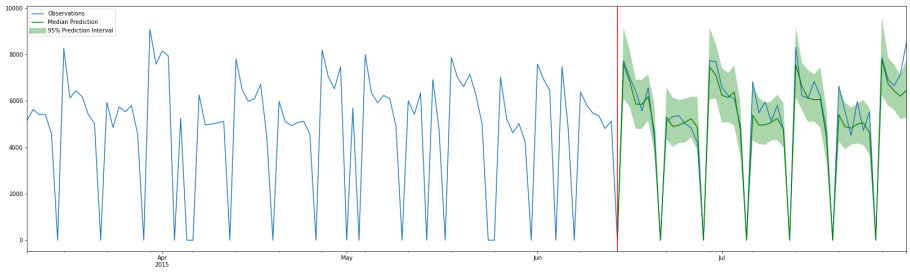

Figure A.36: Deep AR's Forecasts for time series 676

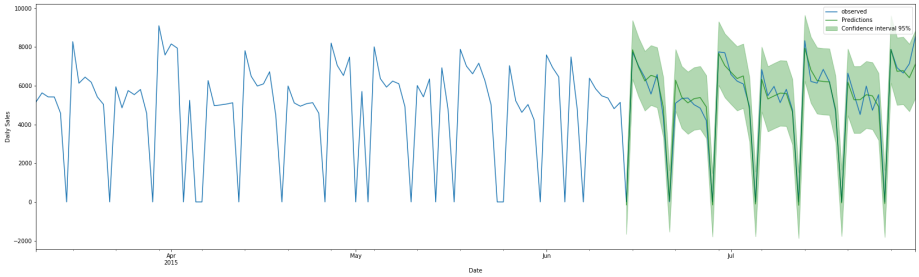

Figure A.37: SARIMAX's Forecasts for time series 676

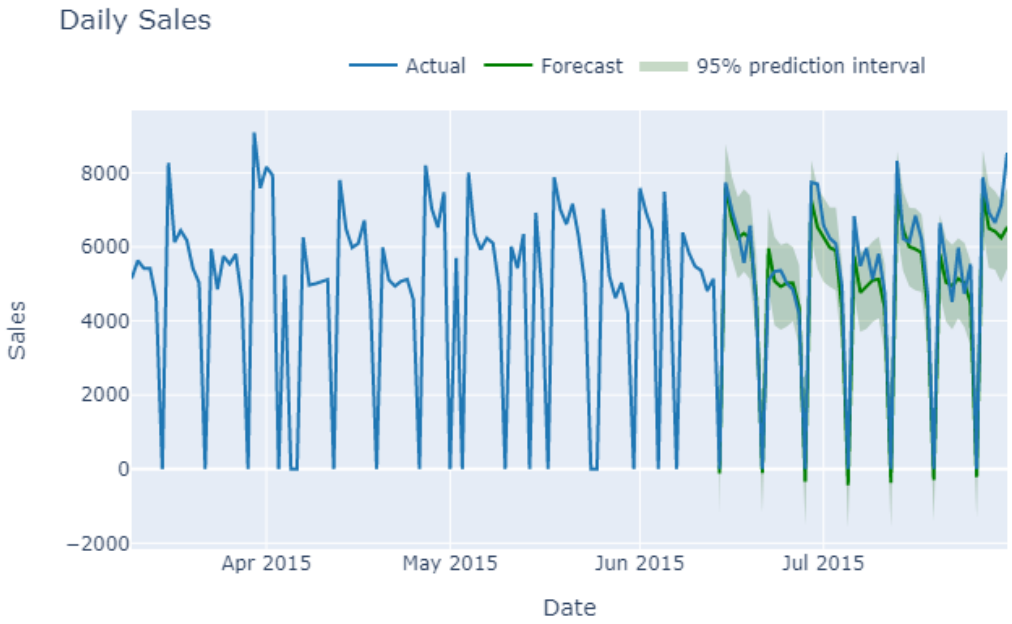

Figure A.38: Prophet's Forecasts for time series 676

A.1.2.3 Q3

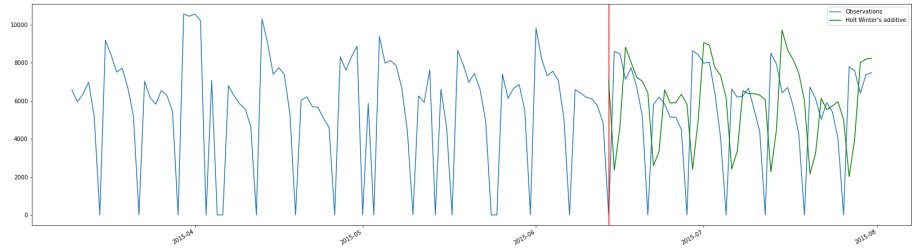

Figure A.39: Holt Winter's Forecasts for time series 850

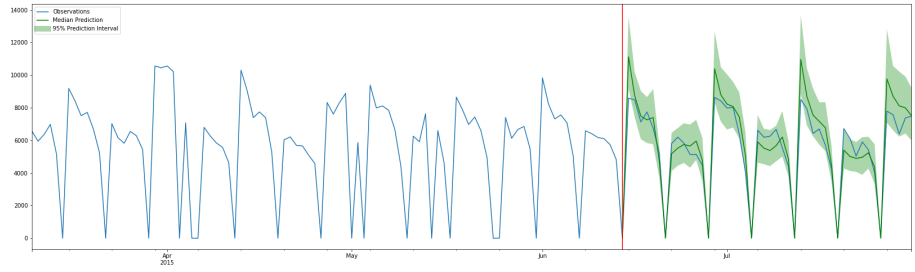

Figure A.40: Deep AR's Forecasts for time series 850

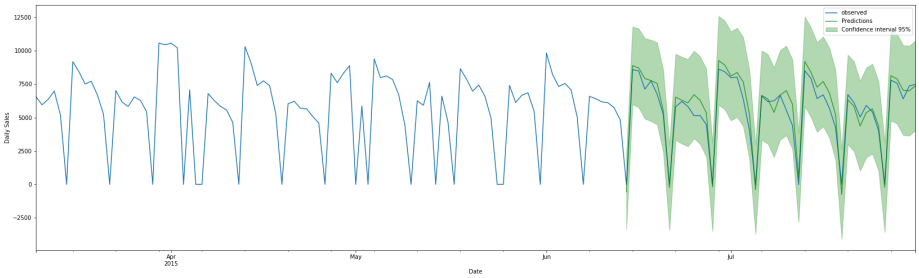

Figure A.41: SARIMAX's Forecasts for time series 850

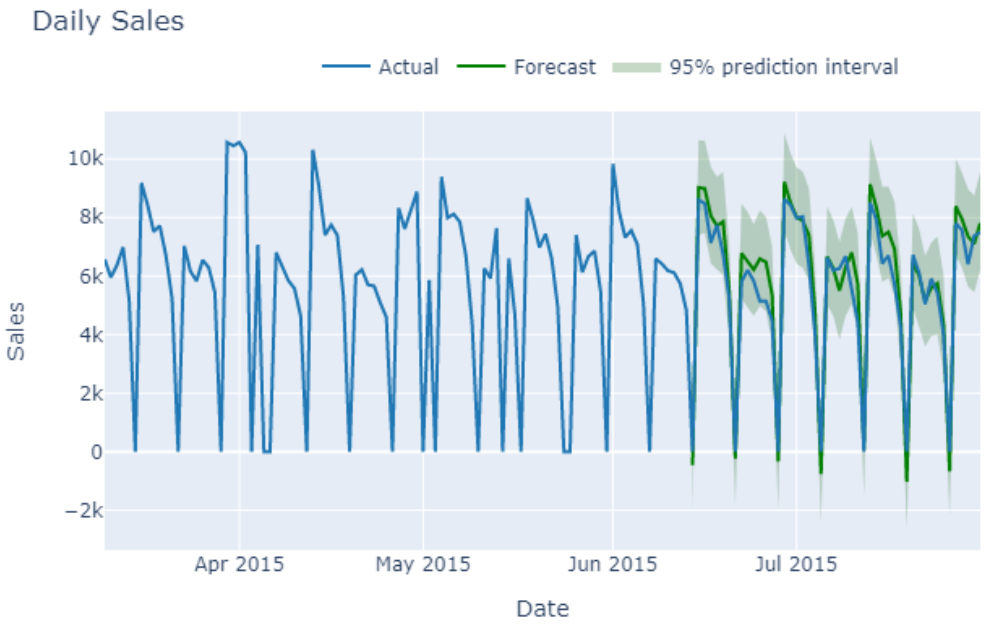

Figure A.42: Prophet's Forecasts for time series 850

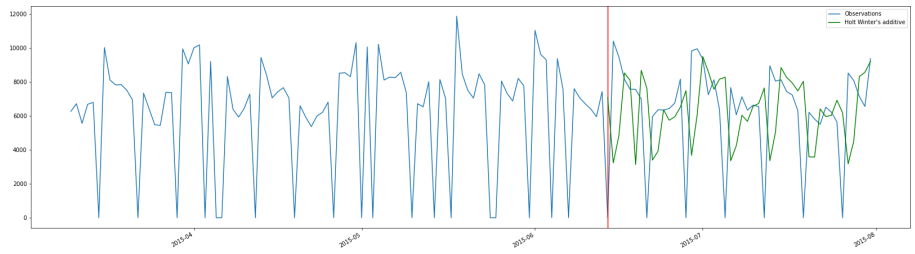

Figure A.43: Holt Winter's Forecasts for time series 130

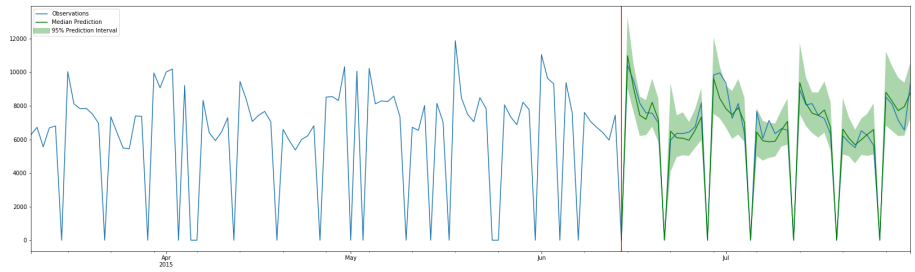

Figure A.44: Deep AR's Forecasts for time series 130

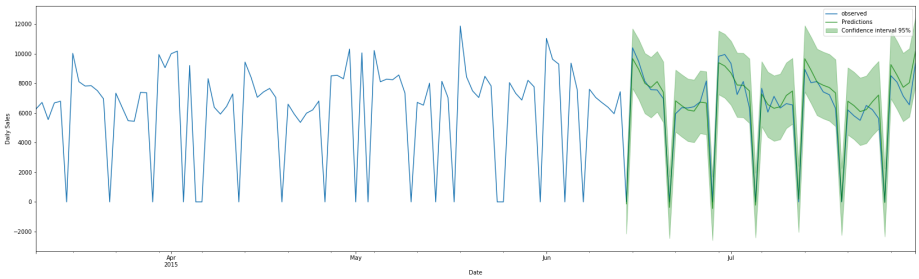

Figure A.45: SARIMAX's Forecasts for time series 130

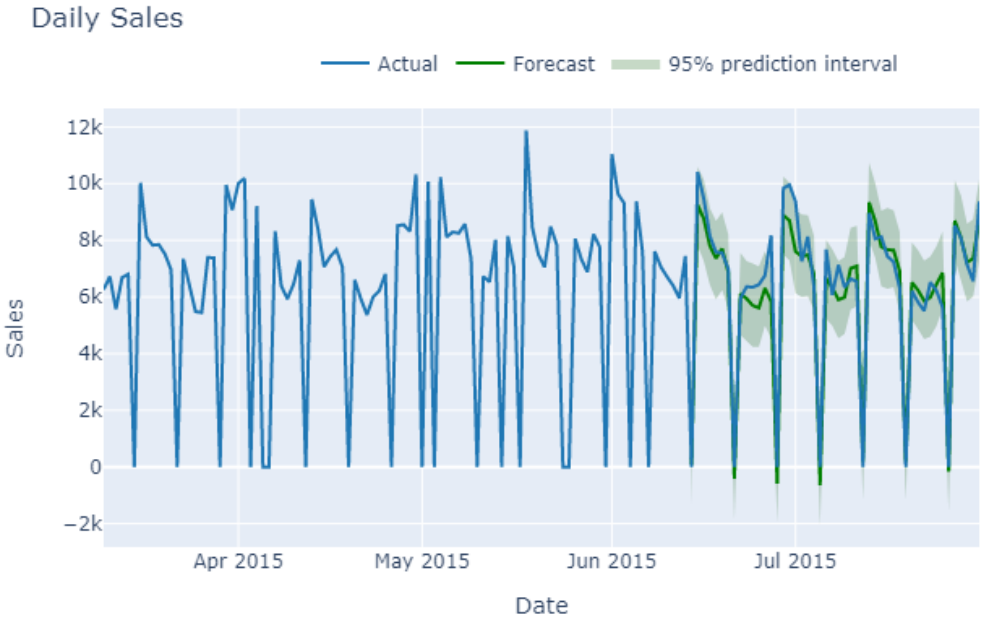

Figure A.46: Prophet's Forecasts for time series 130

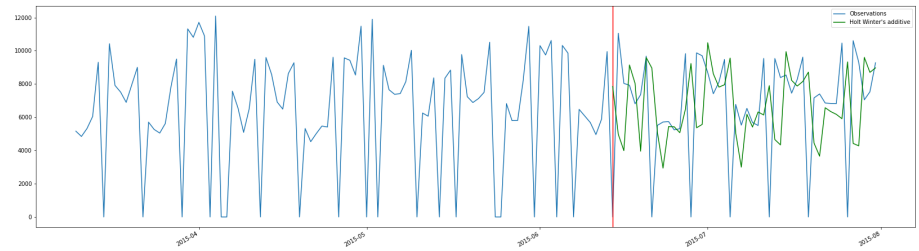

Figure A.47: Holt Winter's Forecasts for time series 384

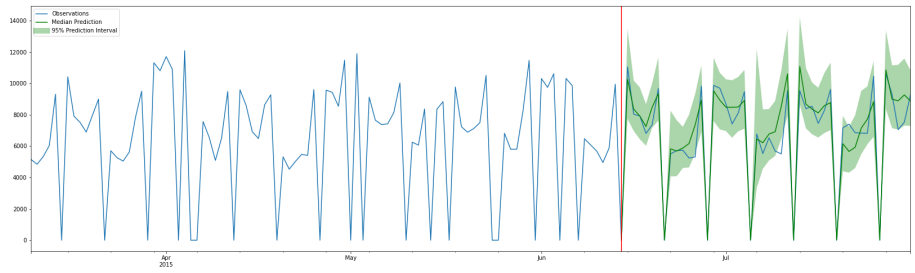

Figure A.48: Deep AR's Forecasts for time series 384

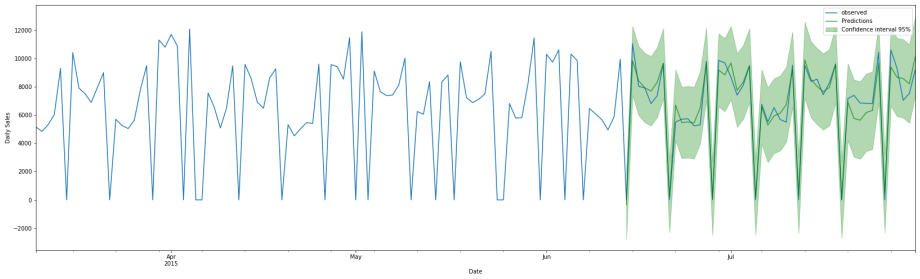

Figure A.49: SARIMAX's Forecasts for time series 384

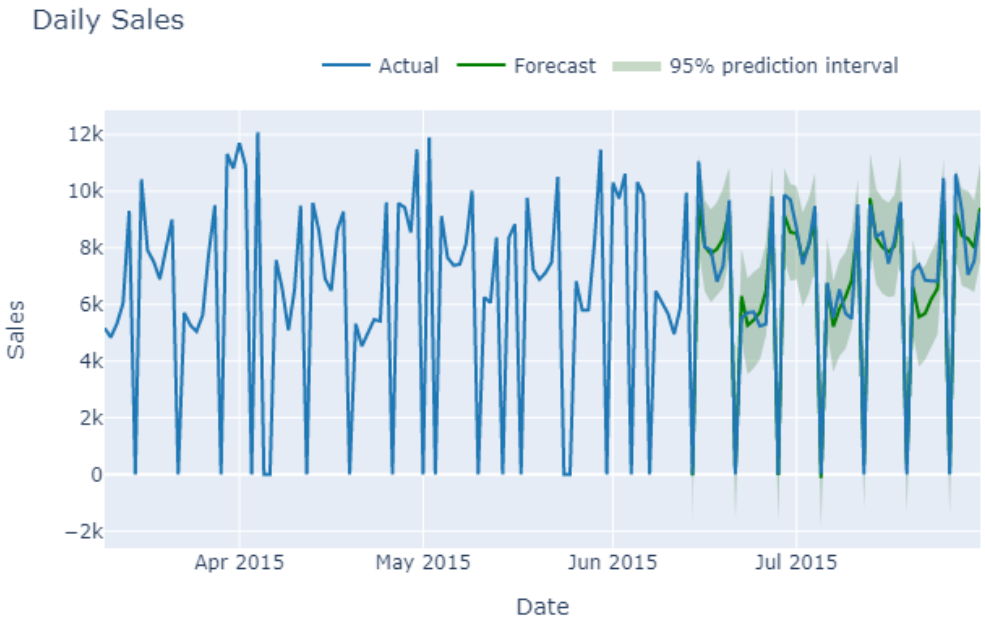

Figure A.50: Prophet's Forecasts for time series 384

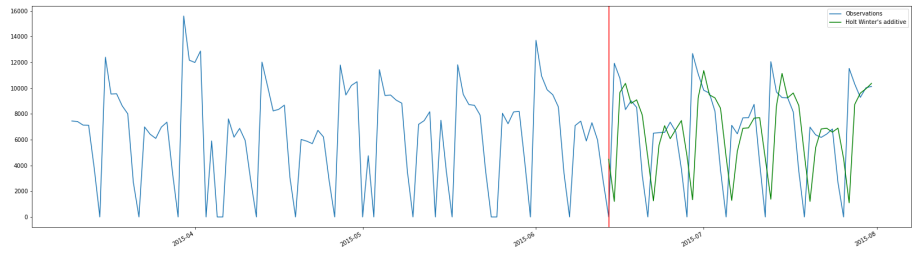

Figure A.51: Holt Winter's Forecasts for time series 113

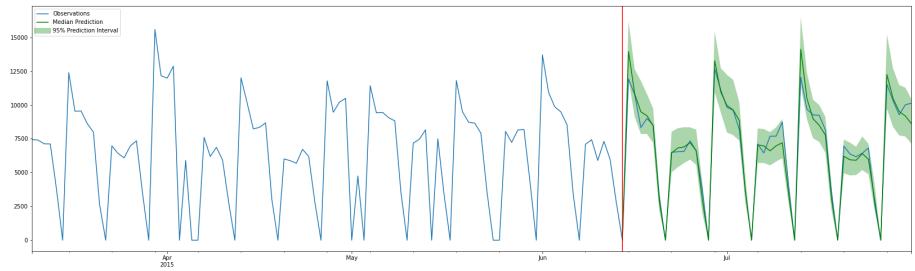

Figure A.52: Deep AR's Forecasts for time series 113

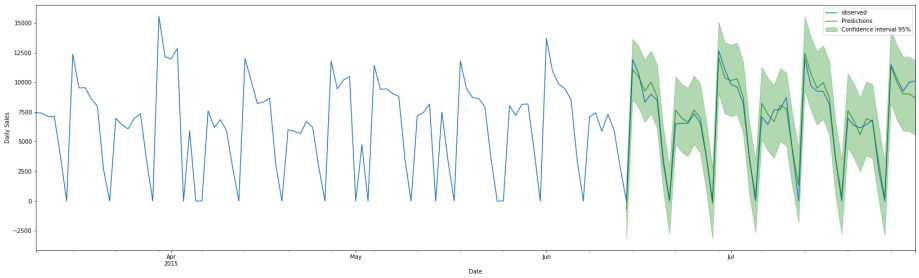

Figure A.53: SARIMAX's Forecasts for time series 113

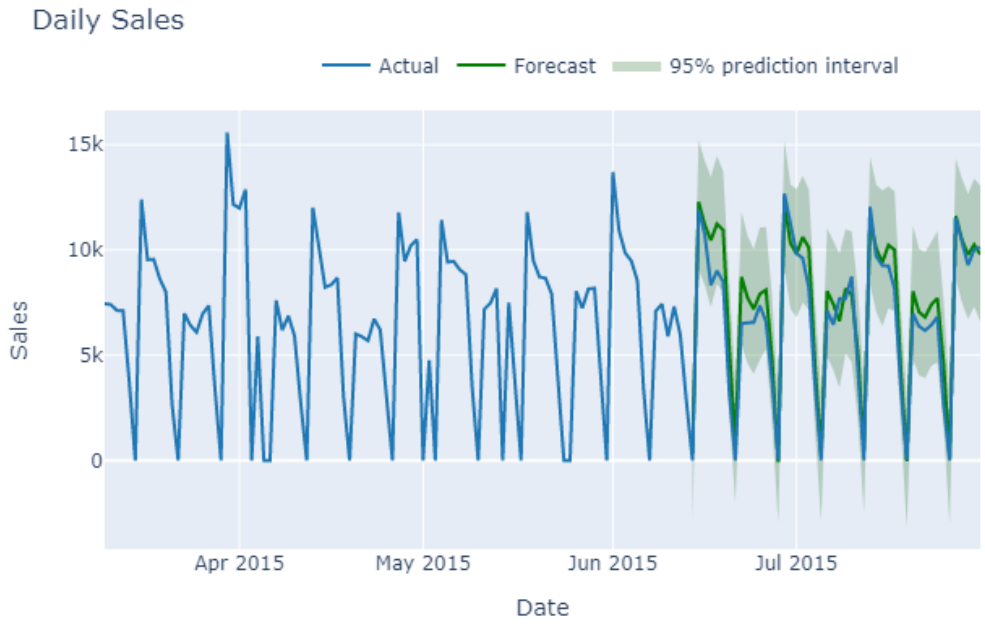

Figure A.54: Prophet's Forecasts for time series 113

A.1.2.4 Q4

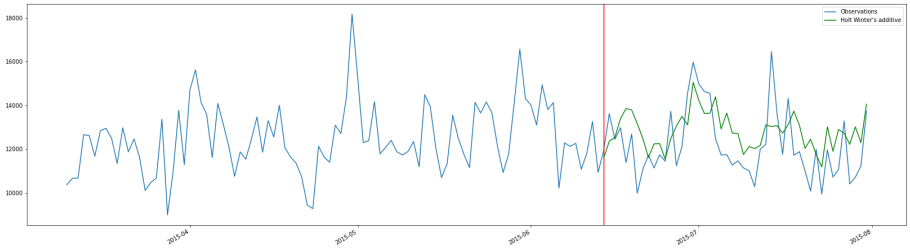

Figure A.55: Holt Winter's Forecasts for time series 639

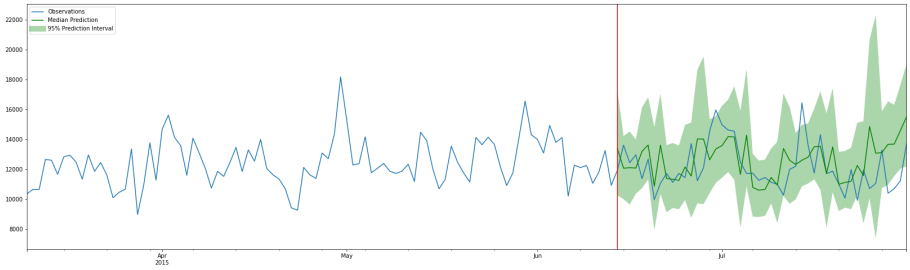

Figure A.56: Deep AR's Forecasts for time series 639

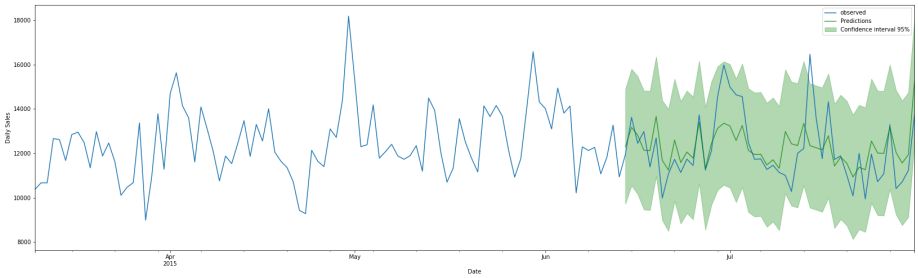

Figure A.57: SARIMAX's Forecasts for time series 639

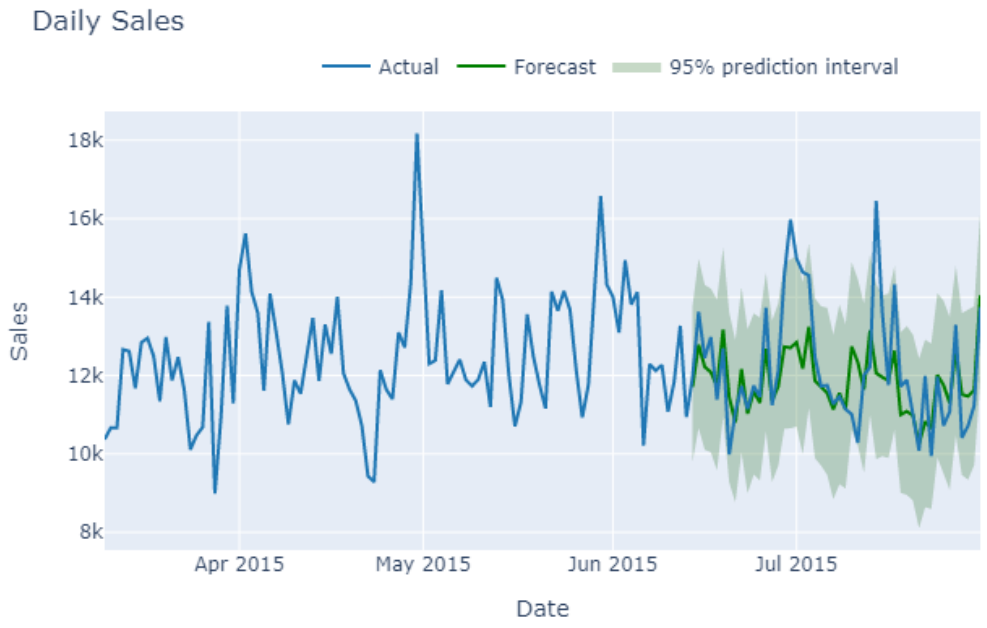

Figure A.58: Prophet's Forecasts for time series 639

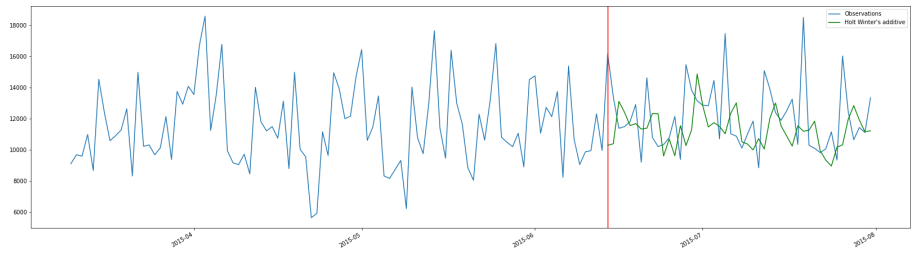

Figure A.59: Holt Winter's Forecasts for time series 355

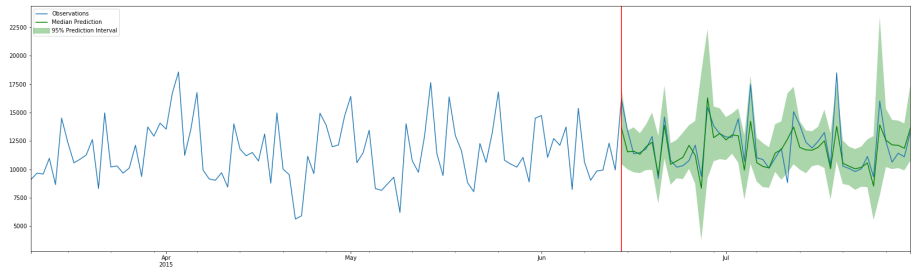

Figure A.60: Deep AR's Forecasts for time series 355

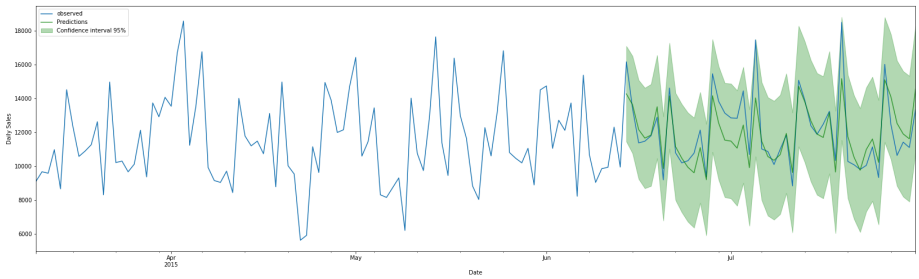

Figure A.61: SARIMAX's Forecasts for time series 355

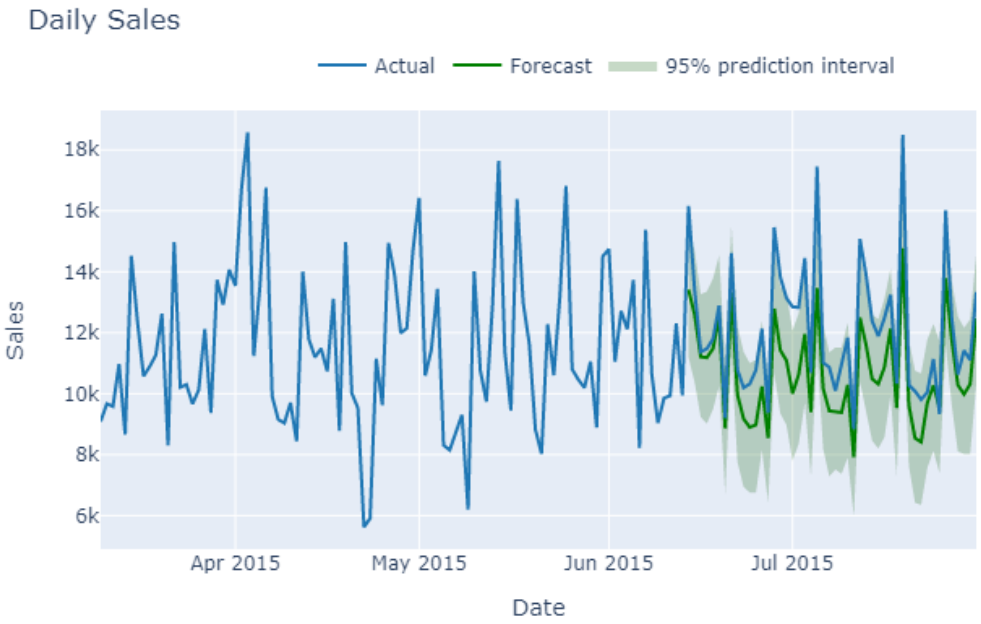

Figure A.62: Prophet's Forecasts for time series 355

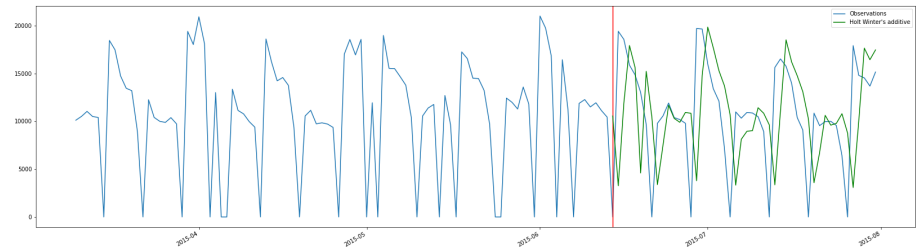

Figure A.63: Holt Winter's Forecasts for time series 691

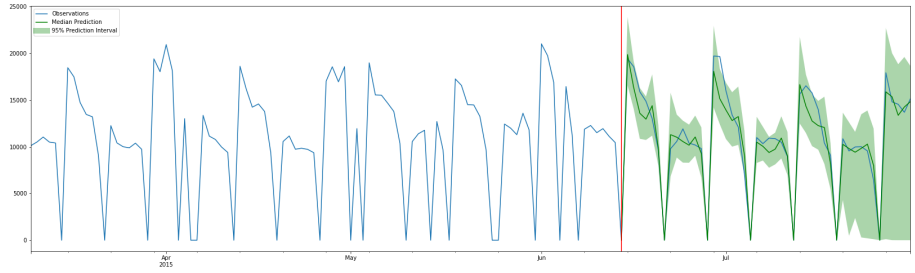

Figure A.64: Deep AR's Forecasts for time series 691

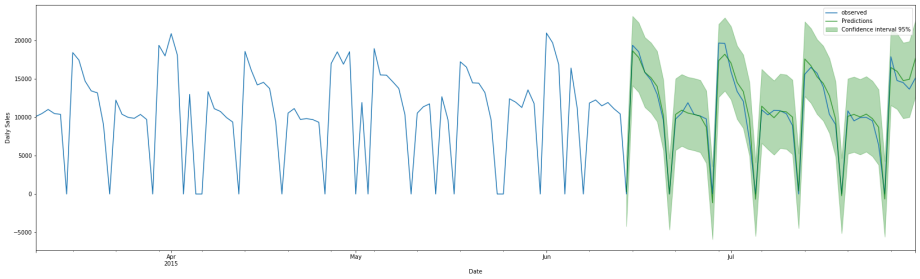

Figure A.65: SARIMAX's Forecasts for time series 691

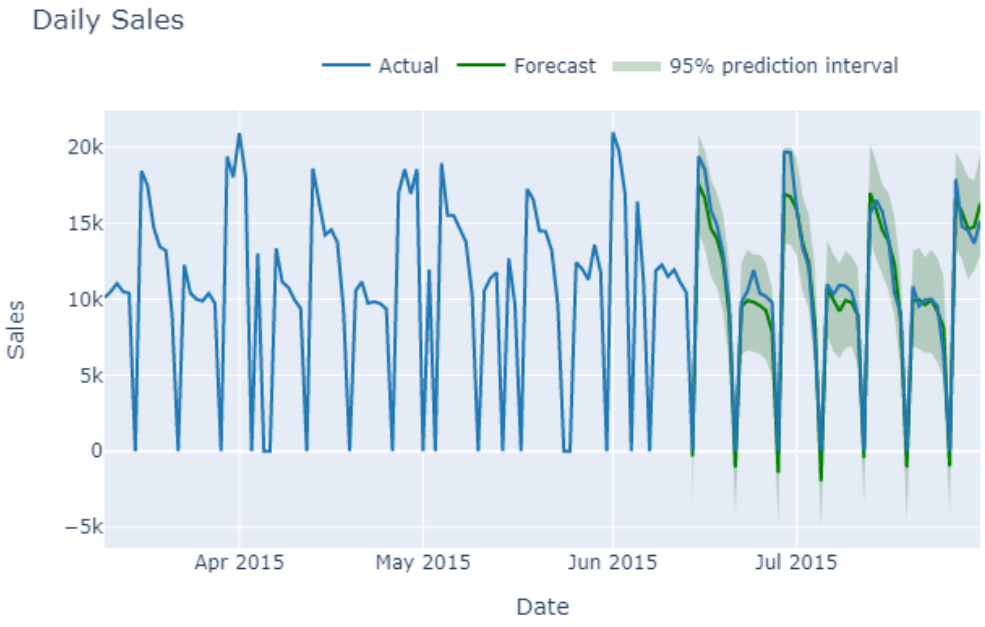

Figure A.66: Prophet's Forecasts for time series 691

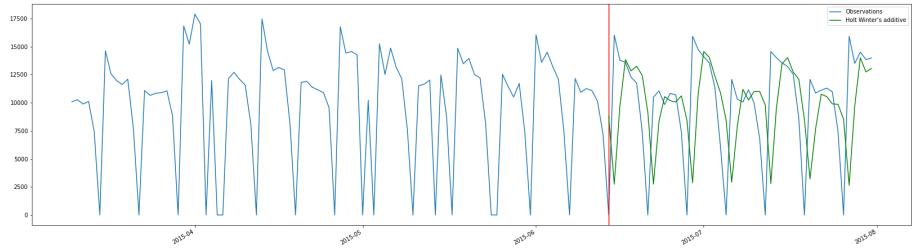

Figure A.67: Holt Winter's Forecasts for time series 360

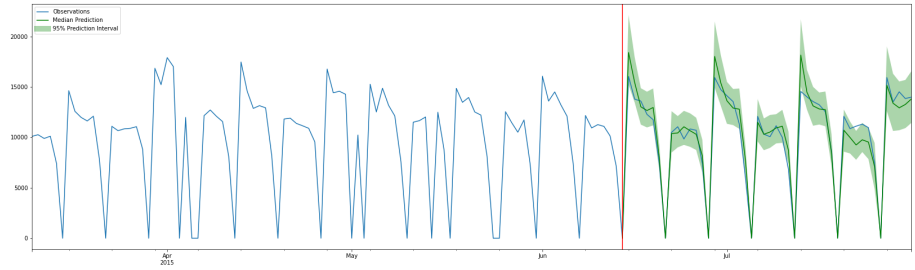

Figure A.68: Deep AR's Forecasts for time series 360

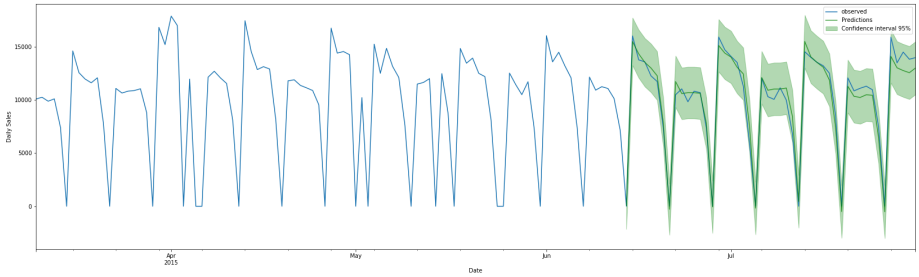

Figure A.69: SARIMAX's Forecasts for time series 360

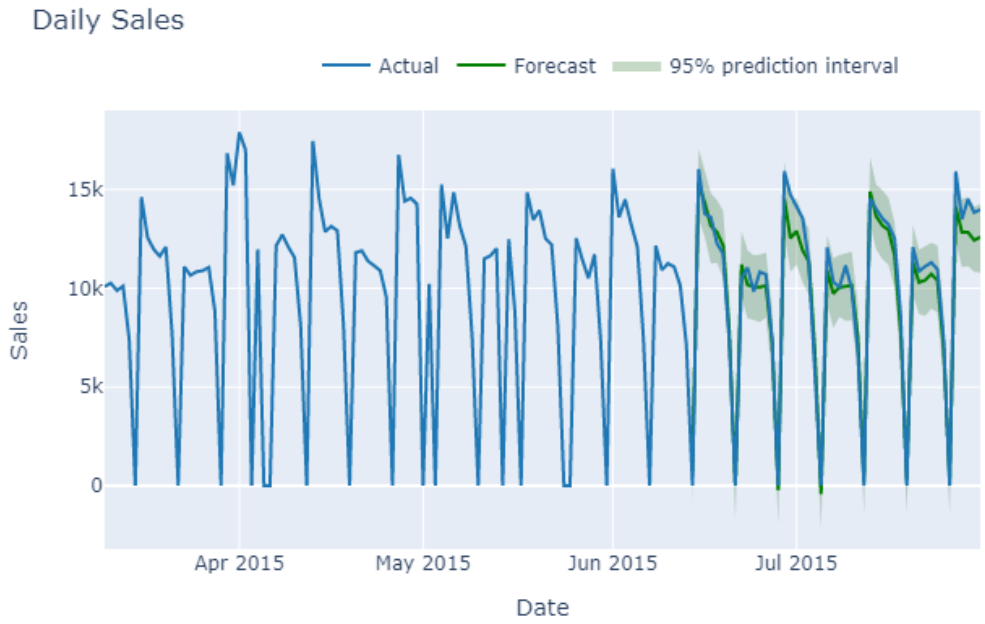

Figure A.70: Prophet's Forecasts for time series 360

A.2 Retail Experiment

A.2.1 Metrics

	sMAPE	MASE	RMSE
HW - Average	105.655	1.024	8.840
SARIMA - Average	85.669	0.760	6.270
SARIMAX - Average	87.021	0.802	6.758
Prophet - Average	92.216	0.853	6.531
ProphetX - Average	101.930	0.997	7.030
DeepAR - Average	85.566	0.695	7.203
DeepAR-FDR - Average	91.924	0.661	6.000

Figure A.71: Average metrics for each model

	sMAPE	MASE	RMSE
HW - STD	22.171	0.205	7.709
SARIMA - STD	19.852	0.102	5.184
SARIMAX - STD	16.514	0.143	6.059
Prophet - STD	17.553	0.208	4.962
ProphetX - STD	31.693	0.315	4.840
DeepAR - STD	25.088	0.135	5.431
DeepAR-FDR - STD	26.426	0.208	4.551

Figure A.72: Standard Deviation for each metric for each model

	sMAPE	MASE	RMSE
HW - 834	138.910	0.831	2.593
SARIMA - 834	111.941	0.735	2.439
SARIMAX - 834	113.542	0.708	2.466
Prophet - 834	115.306	0.735	2.442
ProphetX - 834	129.316	1.703	4.167
DeepAR - 834	105.456	0.873	2.526
DeepAR-FDR - 834	122.143	0.931	2.530

	sMAPE	MASE	RMSE
HW - 1496	115.784	0.940	1.684
SARIMA - 1496	89.681	0.729	1.370
SARIMAX - 1496	93.686	0.754	1.420
Prophet - 1496	93.434	0.754	1.432
ProphetX - 1496	93.200	0.777	1.422
DeepAR - 1496	89.583	0.892	1.367
DeepAR-FDR - 1496	125.000	1.110	1.504

Figure A.73: Metrics for each models forecasts for each time series from Low average group

	sMAPE	MASE	RMSE
HW - 721	118.134	1.096	2.483
SARIMA - 721	75.963	0.801	1.822
SARIMAX - 721	79.605	0.809	1.837
Prophet - 721	93.759	0.994	2.140
ProphetX - 721	90.727	0.993	2.136
DeepAR - 721	79.314	0.727	2.860
DeepAR-FDR - 721	76.346	0.551	1.755

	sMAPE	MASE	RMSE
HW - 1675	79.205	0.886	1.879
SARIMA - 1675	72.265	0.853	1.664
SARIMAX - 1675	72.690	0.861	1.694
Prophet - 1675	87.786	1.333	2.456
ProphetX - 1675	89.094	1.381	2.530
DeepAR - 1675	68.542	0.546	1.480
DeepAR-FDR - 1675	67.659	0.546	1.442

Figure A.74: Metrics for each models forecasts for each time series from medium average group

	sMAPE	MASE	RMSE
HW - 1673	73.613	1.365	26.146
SARIMA - 1673	63.789	0.984	18.360
SARIMAX - 1673	69.660	1.156	21.607
Prophet - 1673	60.055	0.898	16.990
ProphetX - 1673	58.789	0.861	16.197
DeepAR - 1673	54.676	0.726	15.462
DeepAR-FDR - 1673	54.970	0.732	15.450

	sMAPE	MASE	RMSE
HW - 1023	110.147	0.937	13.667
SARIMA - 1023	79.383	0.618	9.486
SARIMAX - 1023	78.135	0.627	9.488
Prophet - 1023	110.161	0.719	11.476
ProphetX - 1023	125.287	0.763	12.000
DeepAR - 1023	78.142	0.562	13.351
DeepAR-FDR - 1023	98.197	0.559	10.351

Figure A.75: Metrics for each models forecasts for each time series from high average group

	sMAPE	MASE	RMSE
HW - 1020	109.950	0.809	9.908
SARIMA - 1020	108.488	0.688	8.148
SARIMAX - 1020	104.649	0.744	8.217
Prophet - 1020	95.569	0.680	7.824
ProphetX - 1020	163.665	0.882	10.480
DeepAR - 1020	102.312	0.813	13.933
DeepAR-FDR - 1020	111.152	0.473	7.850

	sMAPE	MASE	RMSE
HW - 1460	127.831	1.392	14.180
SARIMA - 1460	112.380	0.744	7.554
SARIMAX - 1460	104.989	0.841	8.351
Prophet - 1460	106.849	0.953	8.882
ProphetX - 1460	106.413	0.941	8.945
DeepAR - 1460	137.355	0.575	8.714
DeepAR-FDR - 1460	114.206	0.491	7.804

Figure A.76: Metrics for each models forecasts for each time series from high average group

	sMAPE	MASE	RMSE
HW - 124	77.318	0.963	7.018
SARIMA - 124	57.131	0.691	5.591
SARIMAX - 124	66.234	0.715	5.740
Prophet - 124	67.026	0.615	5.139
ProphetX - 124	60.877	0.669	5.391
DeepAR - 124	54.711	0.541	5.136
DeepAR-FDR - 124	57.645	0.554	5.311

Figure A.77: Metrics for each models forecasts for each time series from high average group

A.2.2 Plots

A.2.2.1 Low Average Group Plots

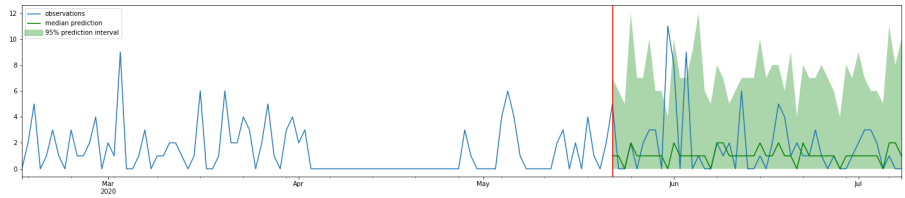

Figure A.78: Deep AR Forecasts not using dynamic features for time series 834

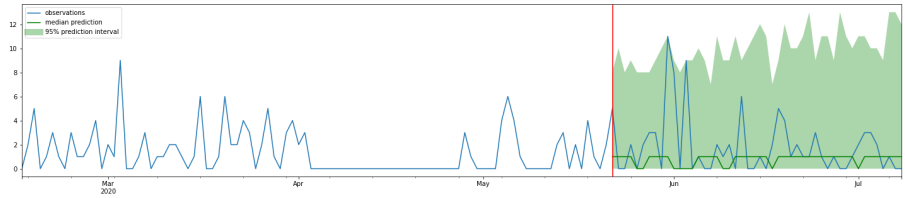

Figure A.79: Deep AR Forecasts using dynamic features for time series 834

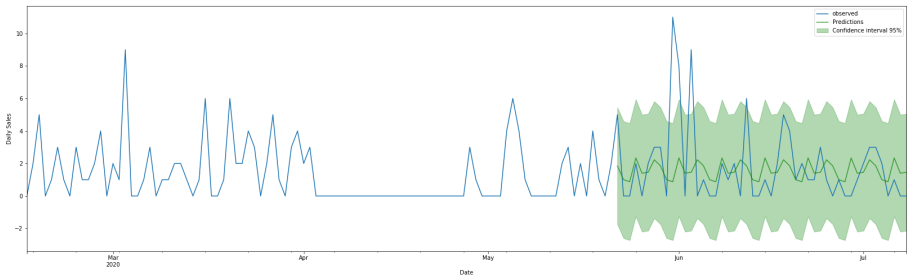

Figure A.80: SARIMA Forecasts for time series 834

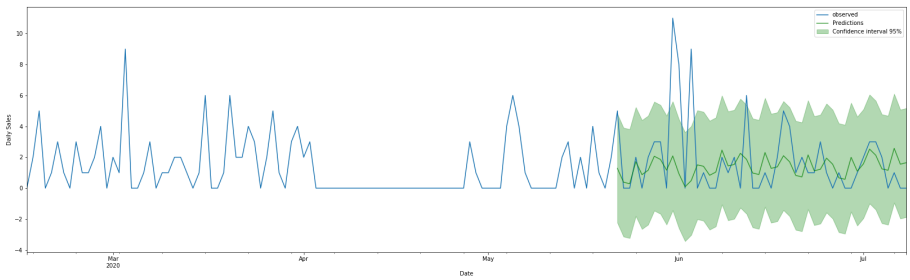

Figure A.81: SARIMAX Forecasts for time series 834

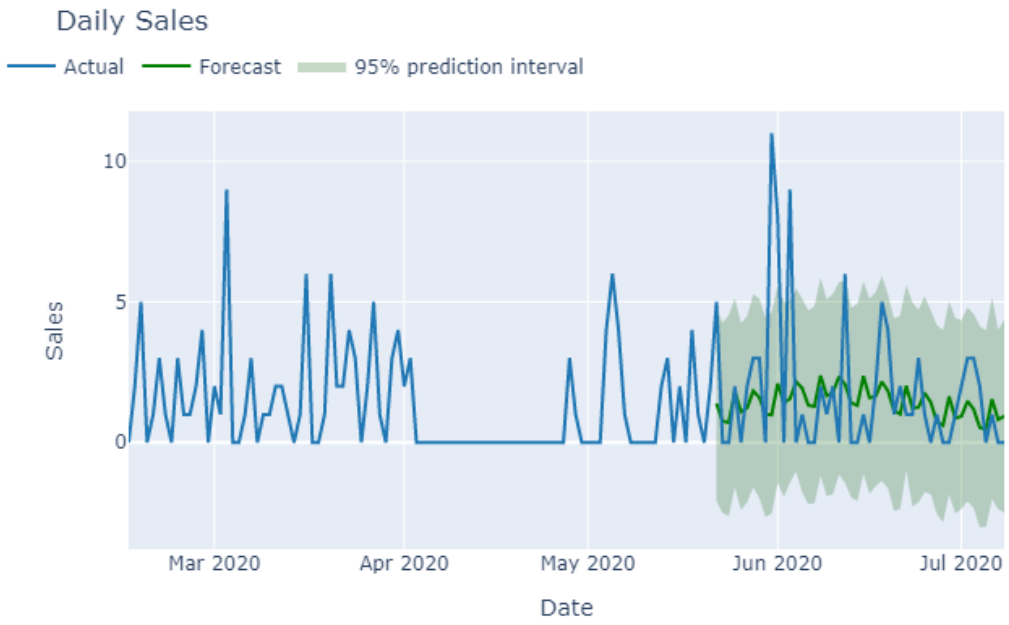

Figure A.82: Prophet Forecasts not using regressors for time series 834

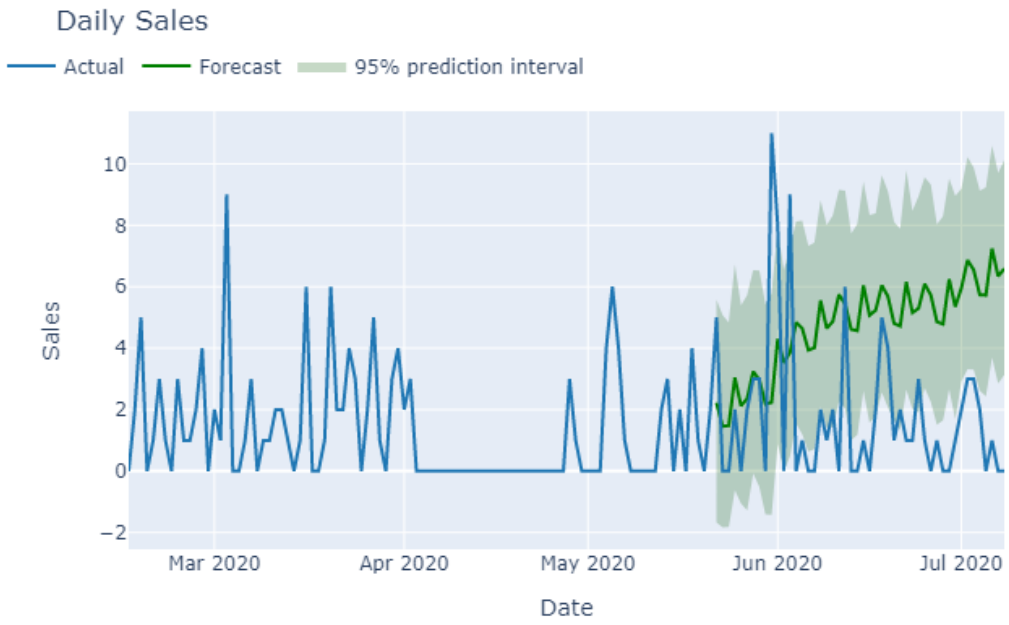

Figure A.83: Prophet Forecasts using regressors for time series 834

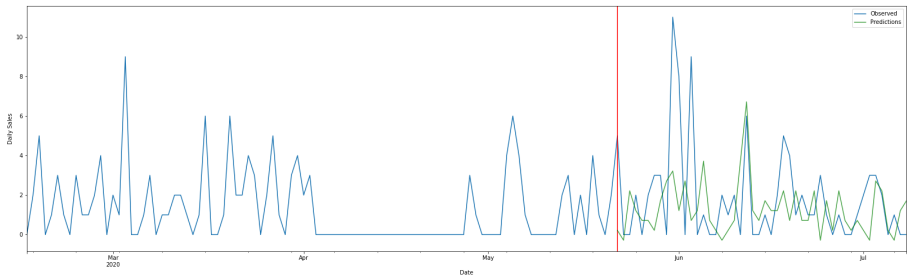

Figure A.84: Holt Winter's Forecasts using regressors for time series 834

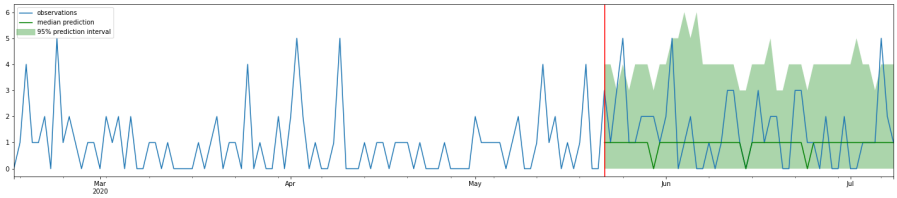

Figure A.85: Deep AR Forecasts not using dynamic features for time series 1496

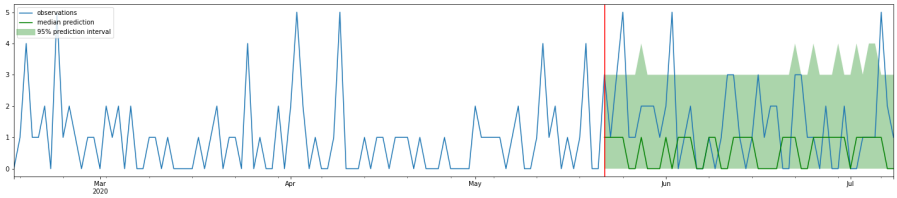

Figure A.86: Deep AR Forecasts using dynamic features for time series 1496

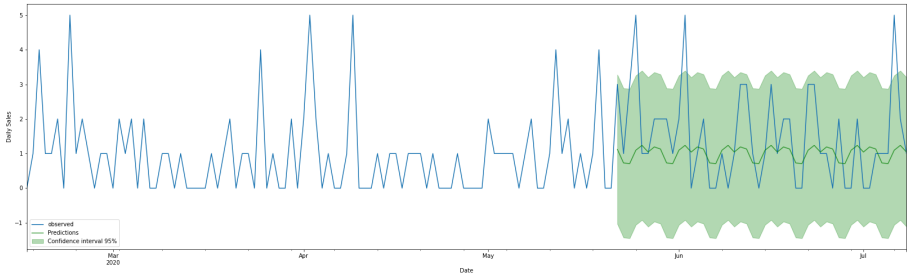

Figure A.87: SARIMA Forecasts for time series 1496

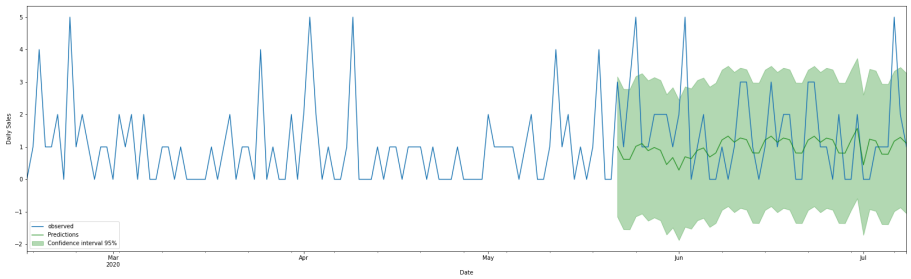

Figure A.88: SARIMAX Forecasts for time series 1496

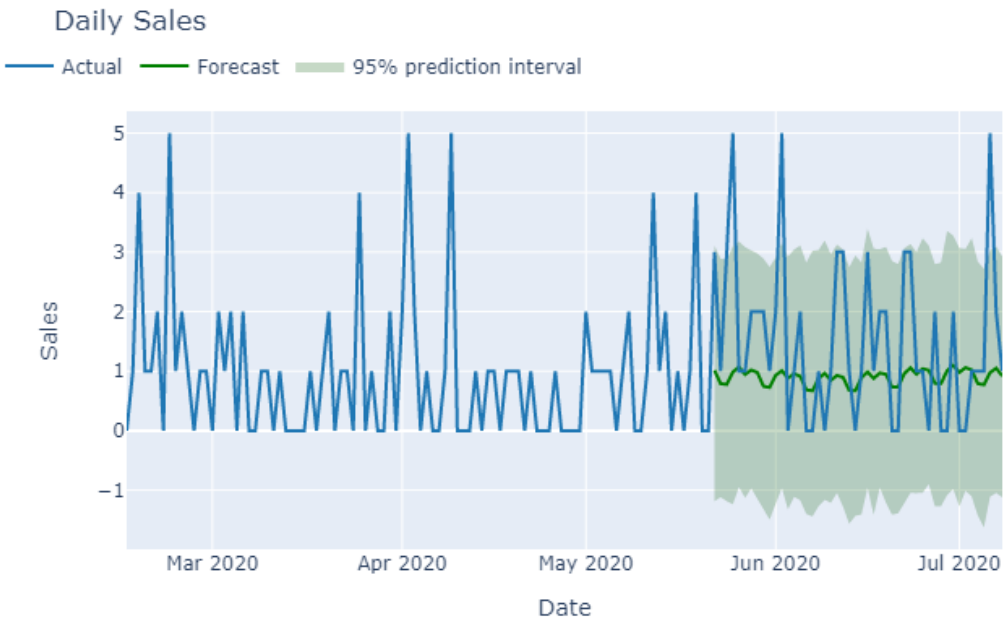

Figure A.89: Prophet Forecasts not using regressors for time series 1496

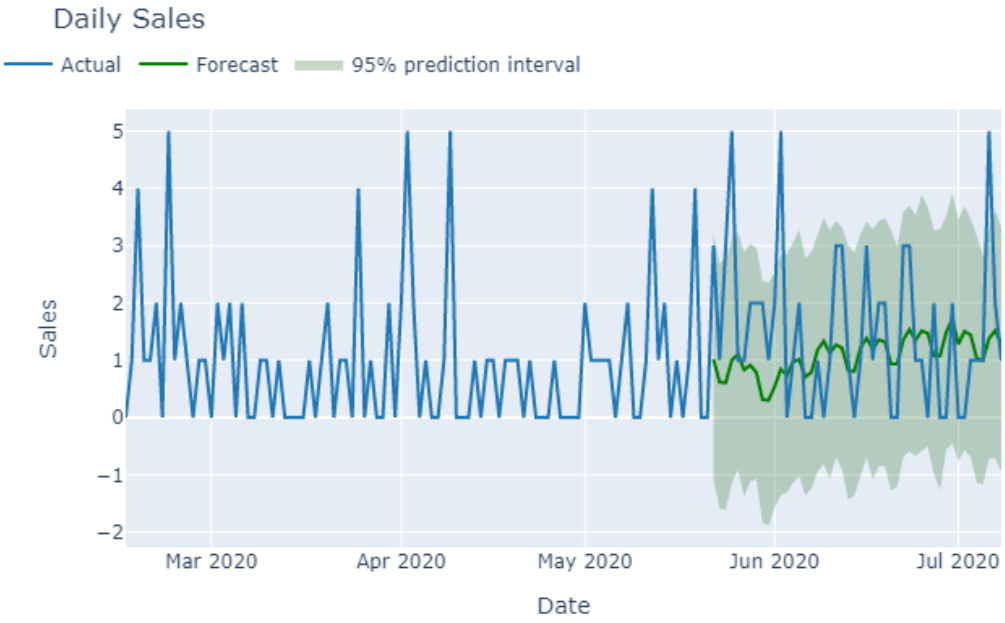

Figure A.90: Prophet Forecasts using regressors for time series 1496

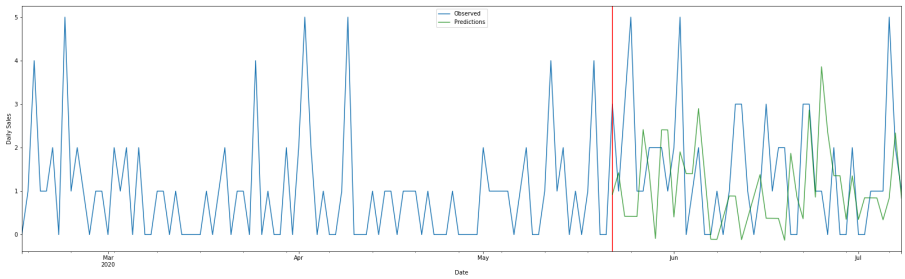

Figure A.91: Holt Winter's Forecasts using regressors for time series 1496

A.2.2.2 Medium Average Group Plots

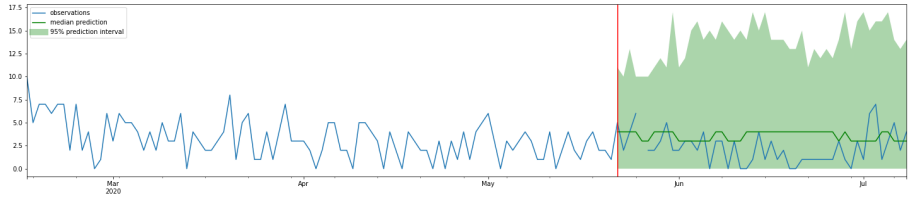

Figure A.92: Deep AR Forecasts not using dynamic features for time series 721

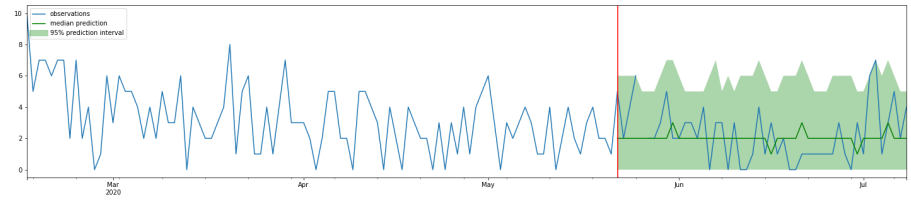

Figure A.93: Deep AR Forecasts using dynamic features for time series 721

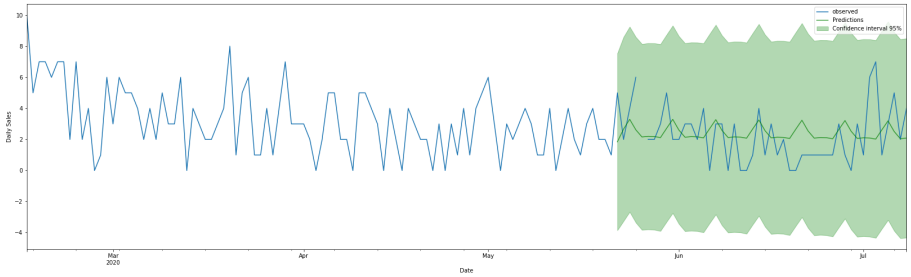

Figure A.94: SARIMA Forecasts for time series 721

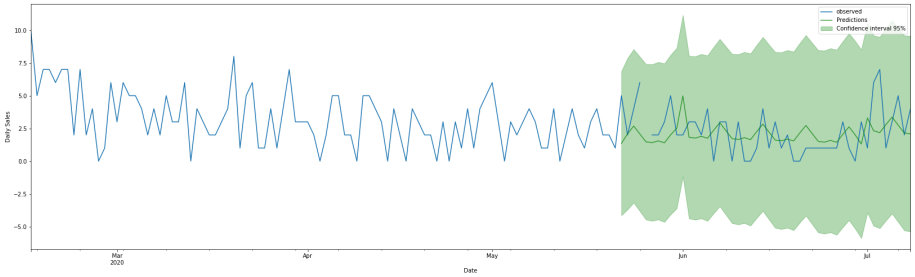

Figure A.95: SARIMAX Forecasts for time series 721

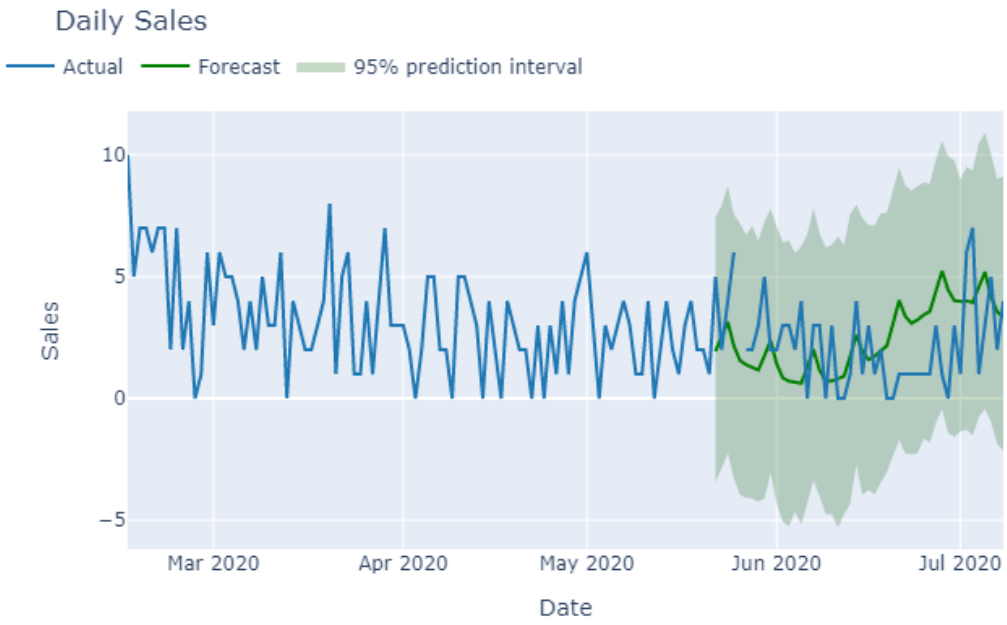

Figure A.96: Prophet Forecasts not using regressors for time series 721

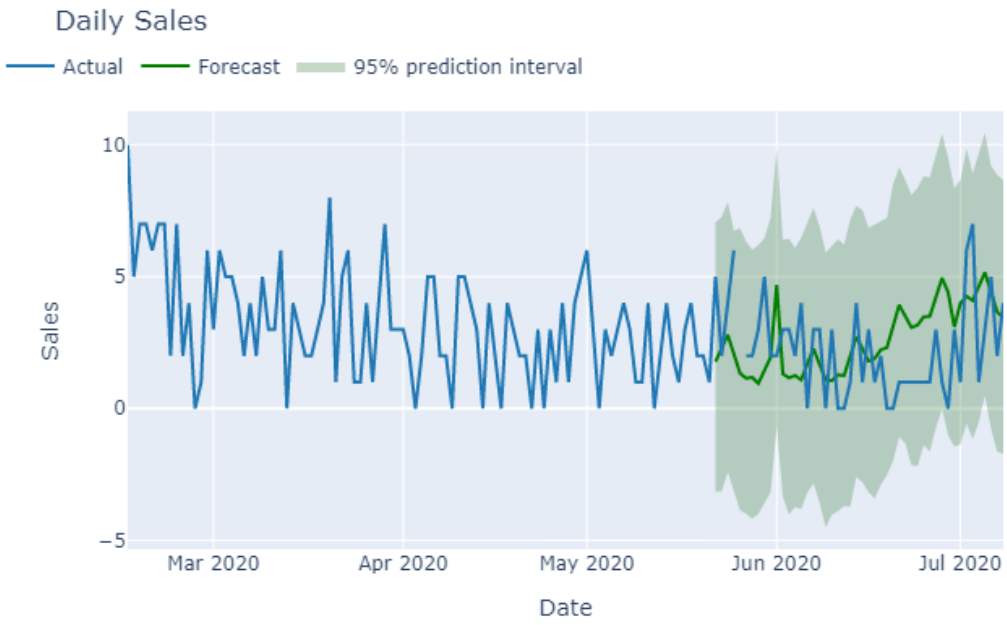

Figure A.97: Prophet Forecasts using regressors for time series 721

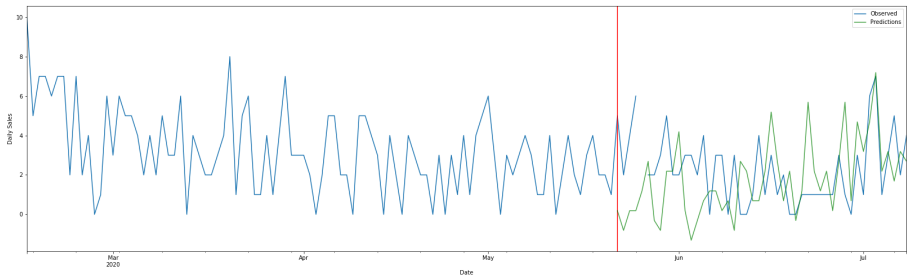

Figure A.98: Holt Winter's Forecasts using regressors for time series 721

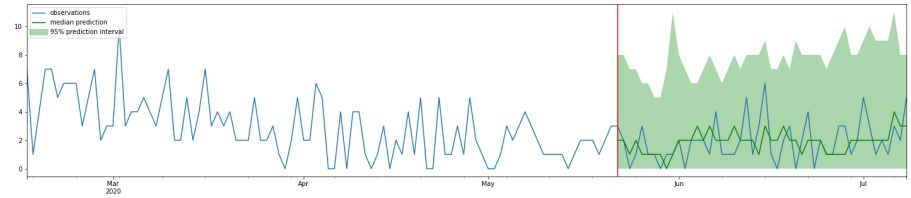

Figure A.99: Deep AR Forecasts not using dynamic features for time series 1675

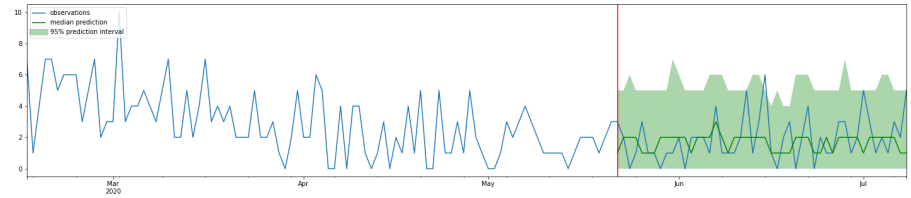

Figure A.100: Deep AR Forecasts using dynamic features for time series 1675

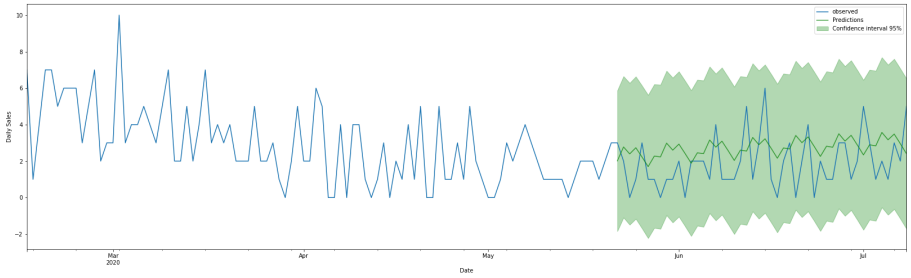

Figure A.101: SARIMA Forecasts for time series 1675

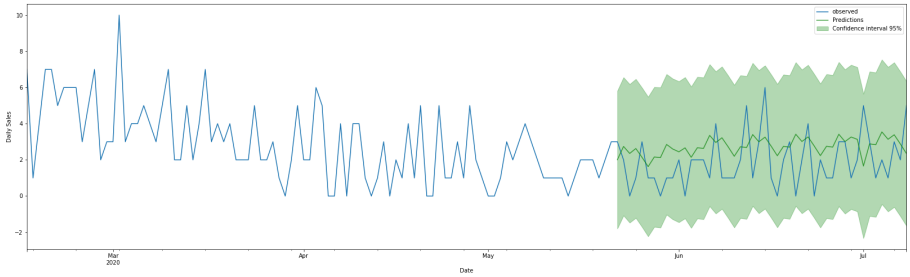

Figure A.102: SARIMAX Forecasts for time series 1675

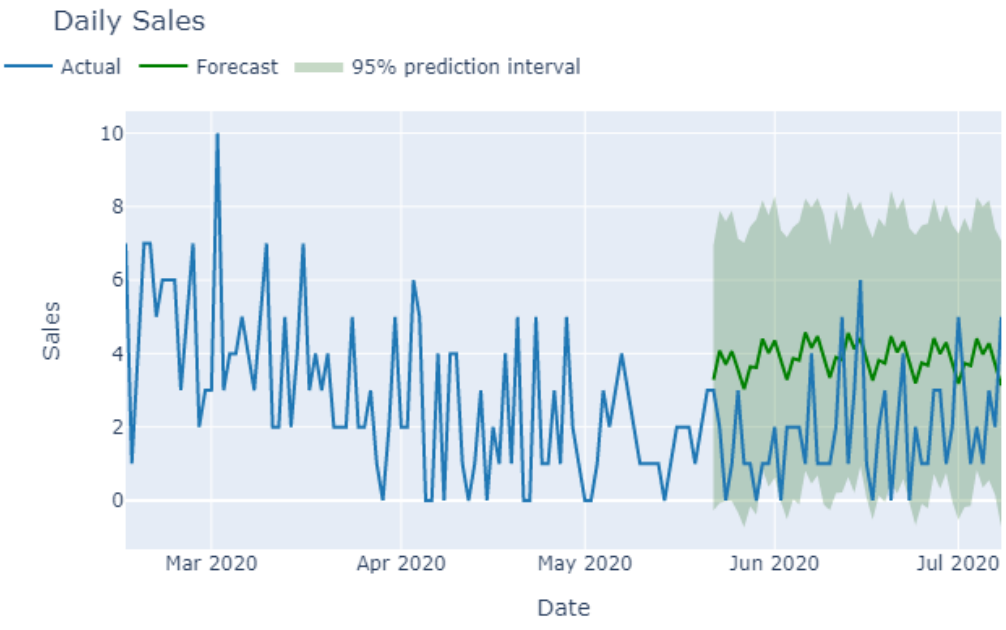

Figure A.103: Prophet Forecasts not using regressors for time series 1675

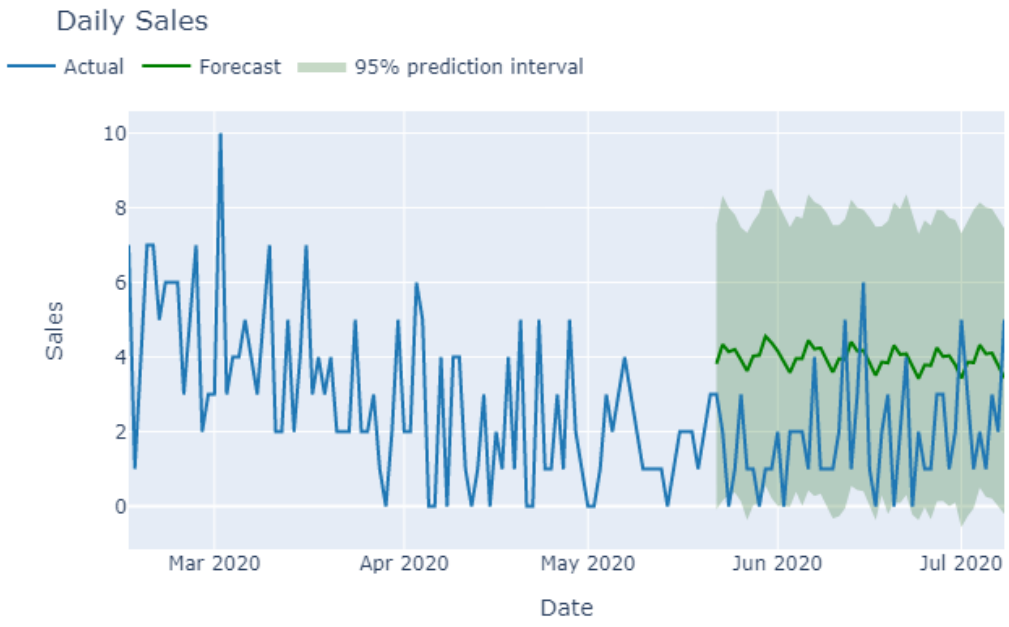

Figure A.104: Prophet Forecasts using regressors for time series 1675

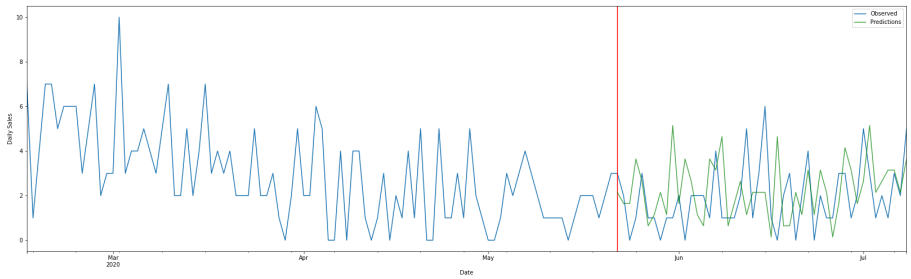

Figure A.105: Holt Winter's Forecasts using regressors for time series 1675

A.2.2.3 High Average Group Plots

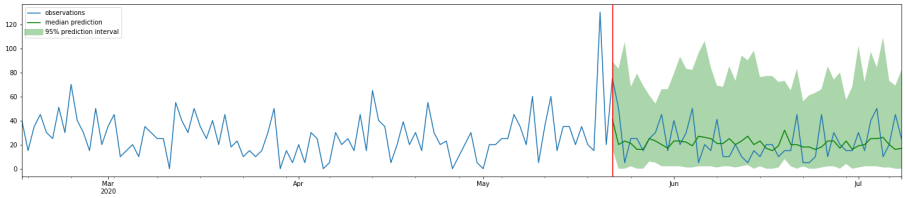

Figure A.106: Deep AR Forecasts not using dynamic features for time series 1673

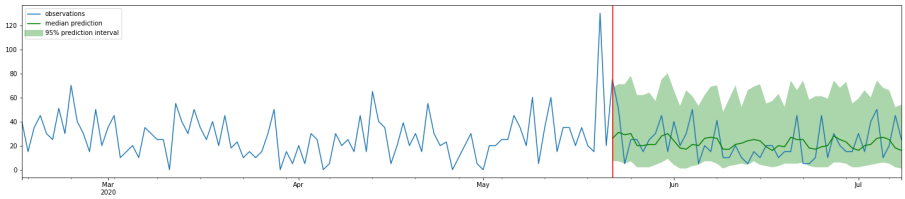

Figure A.107: Deep AR Forecasts using dynamic features for time series 1673

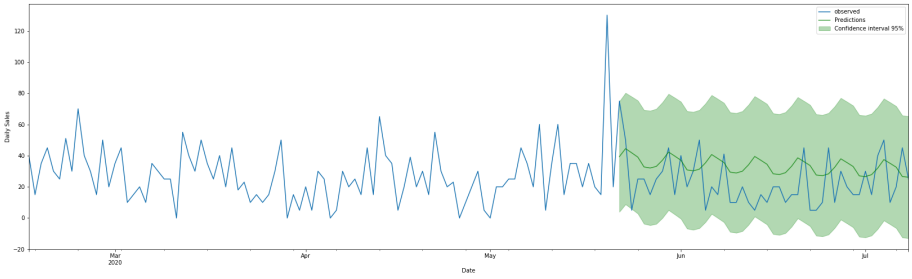

Figure A.108: SARIMA Forecasts for time series 1673

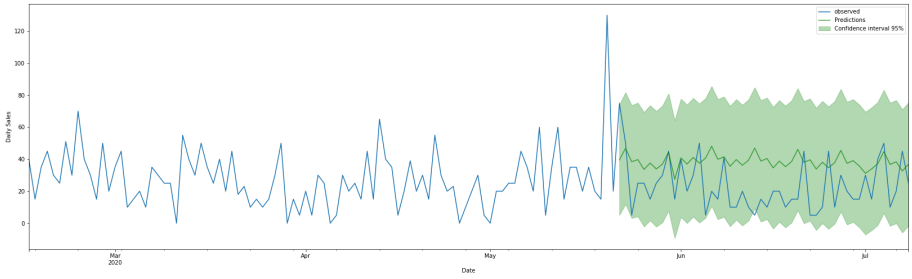

Figure A.109: SARIMAX Forecasts for time series 1673

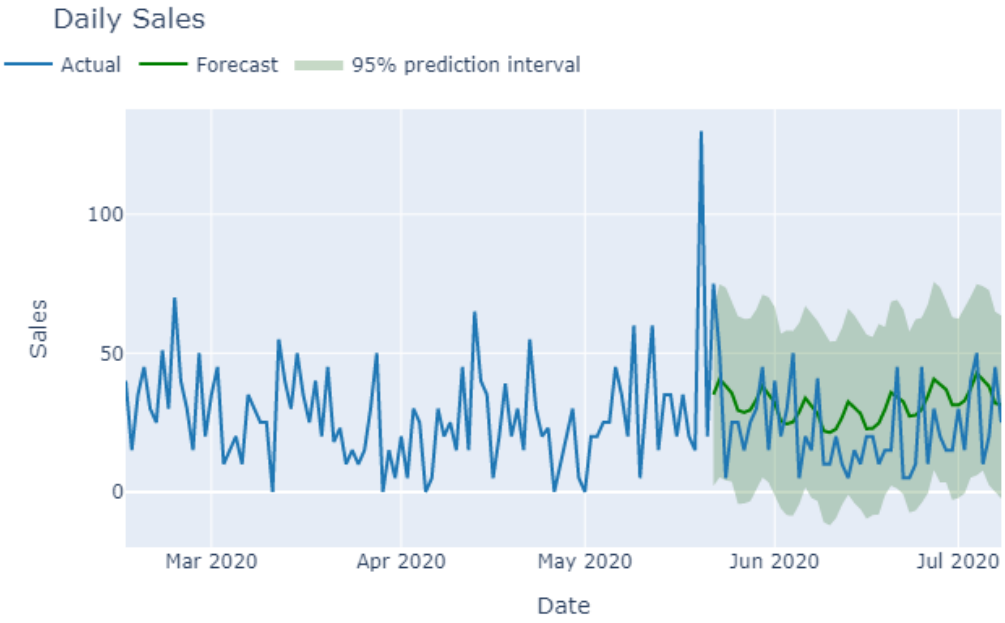

Figure A.110: Prophet Forecasts not using regressors for time series 1673

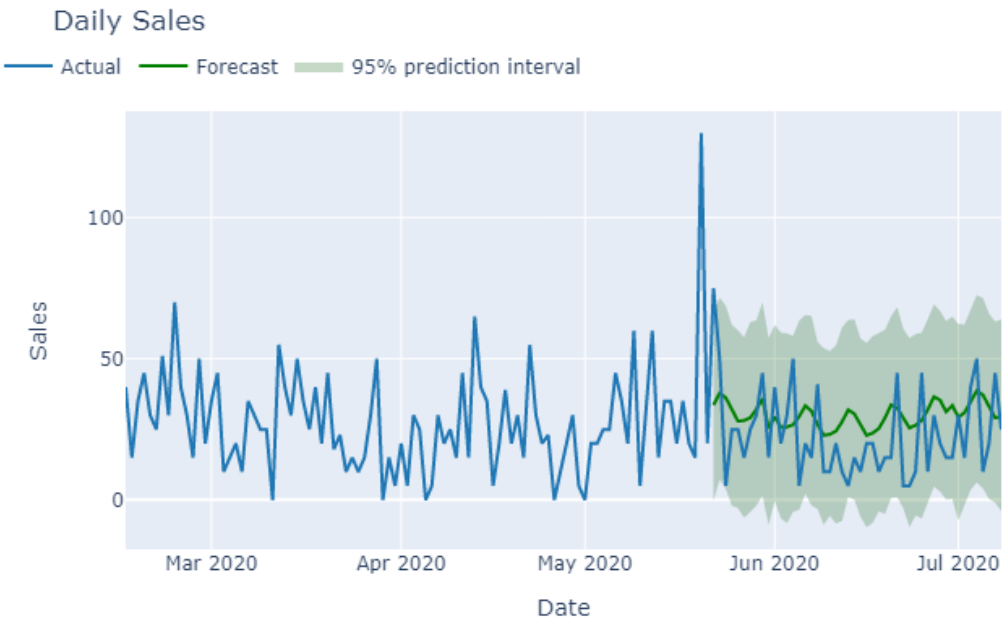

Figure A.111: Prophet Forecasts using regressors for time series 1673

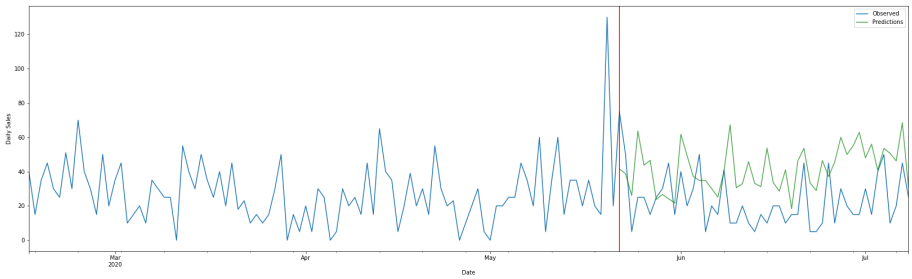

Figure A.112: Holt Winter's Forecasts using regressors for time series 1673

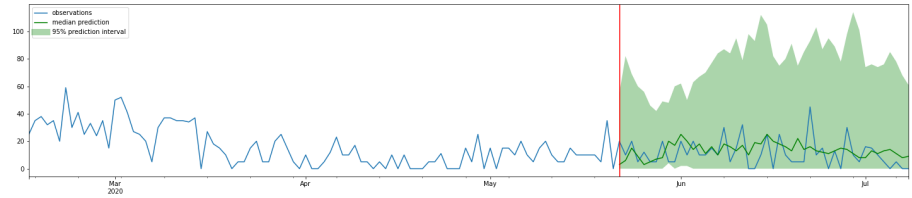

Figure A.113: Deep AR Forecasts not using dynamic features for time series 1023

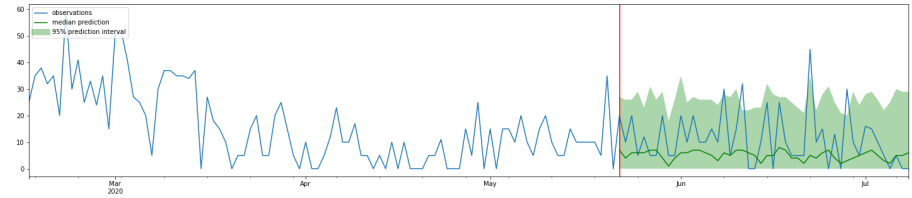

Figure A.114: Deep AR Forecasts using dynamic features for time series 1023

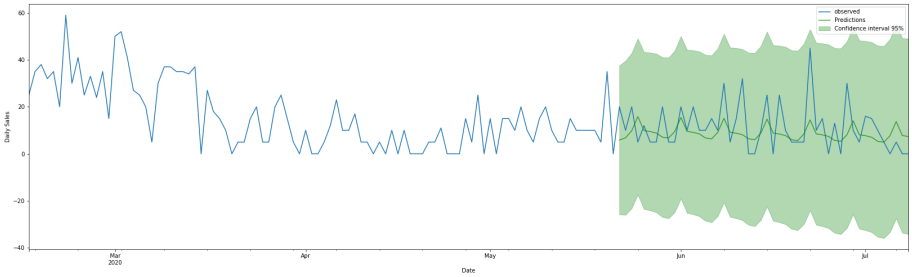

Figure A.115: SARIMA Forecasts for time series 1023

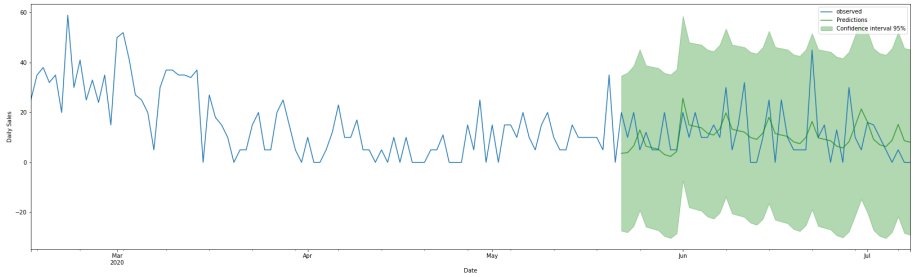

Figure A.116: SARIMAX Forecasts for time series 1023

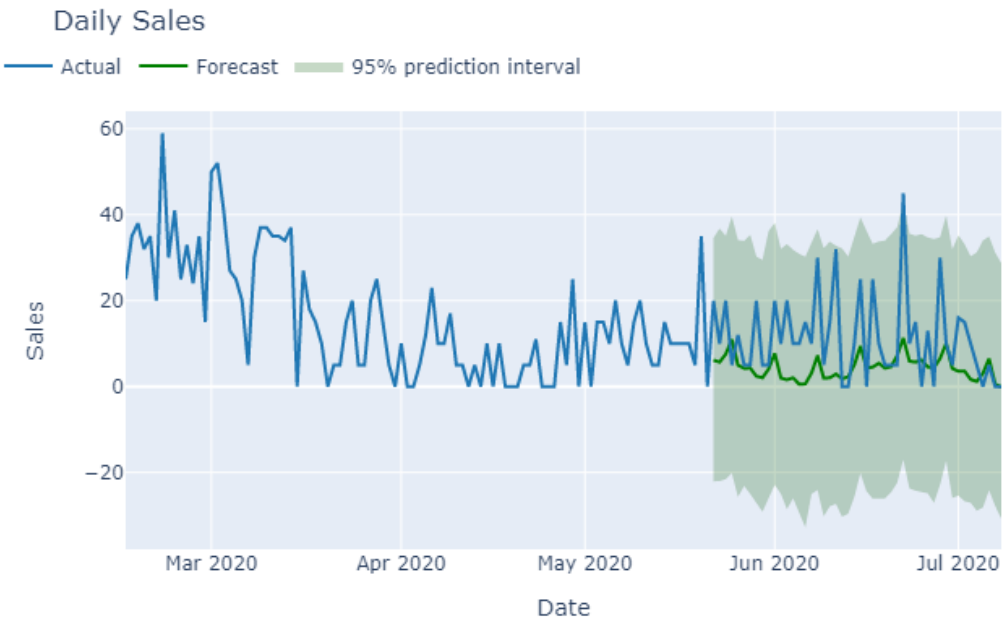

Figure A.117: Prophet Forecasts not using regressors for time series 1023

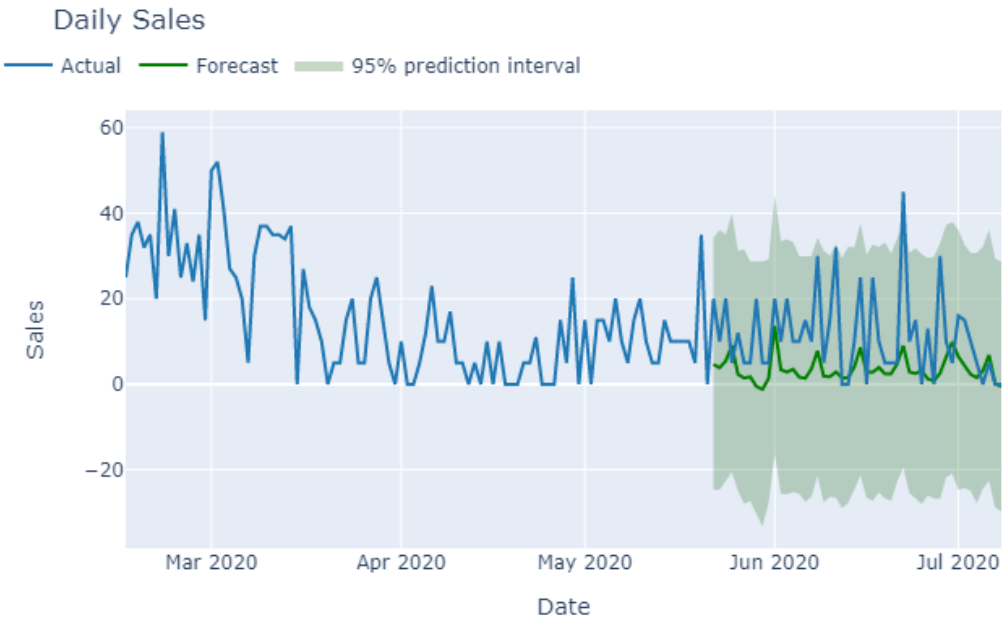

Figure A.118: Prophet Forecasts using regressors for time series 1023

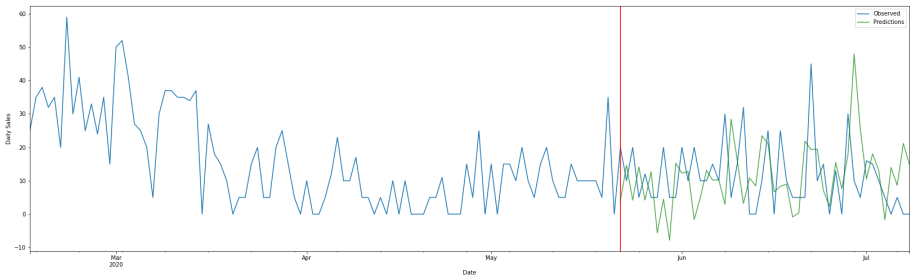

Figure A.119: Holt Winter's Forecasts using regressors for time series 1023

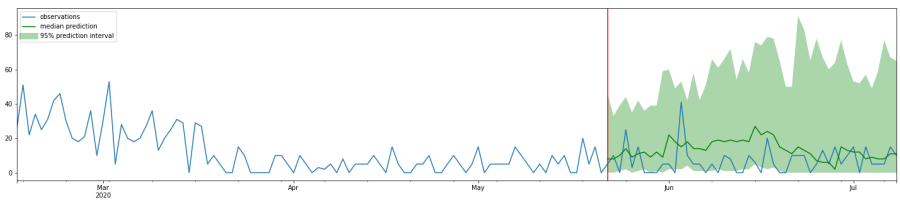

Figure A.120: Deep AR Forecasts not using dynamic features for time series 1020

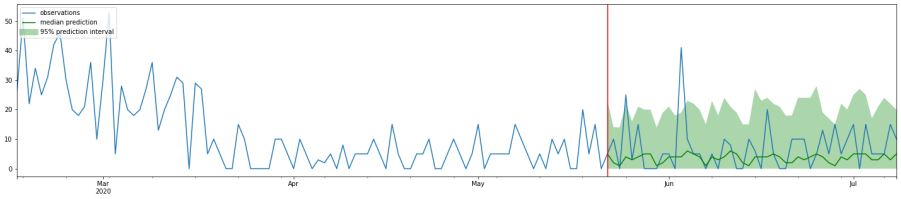

Figure A.121: Deep AR Forecasts using dynamic features for time series 1020

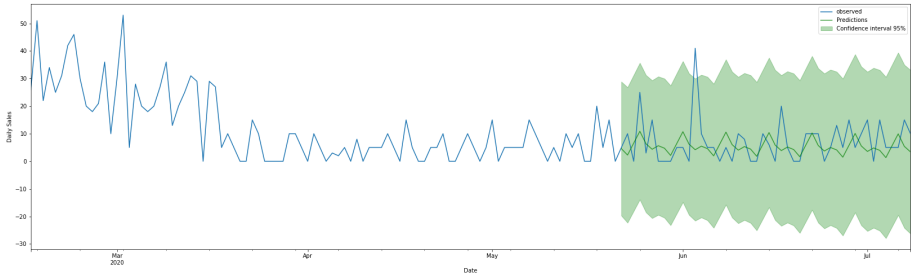

Figure A.122: SARIMA Forecasts for time series 1020

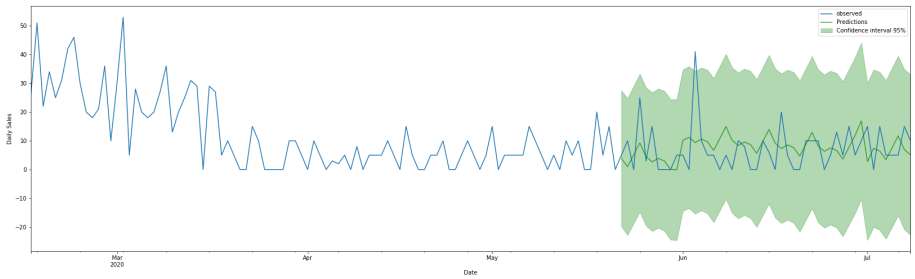

Figure A.123: SARIMAX Forecasts for time series 1020

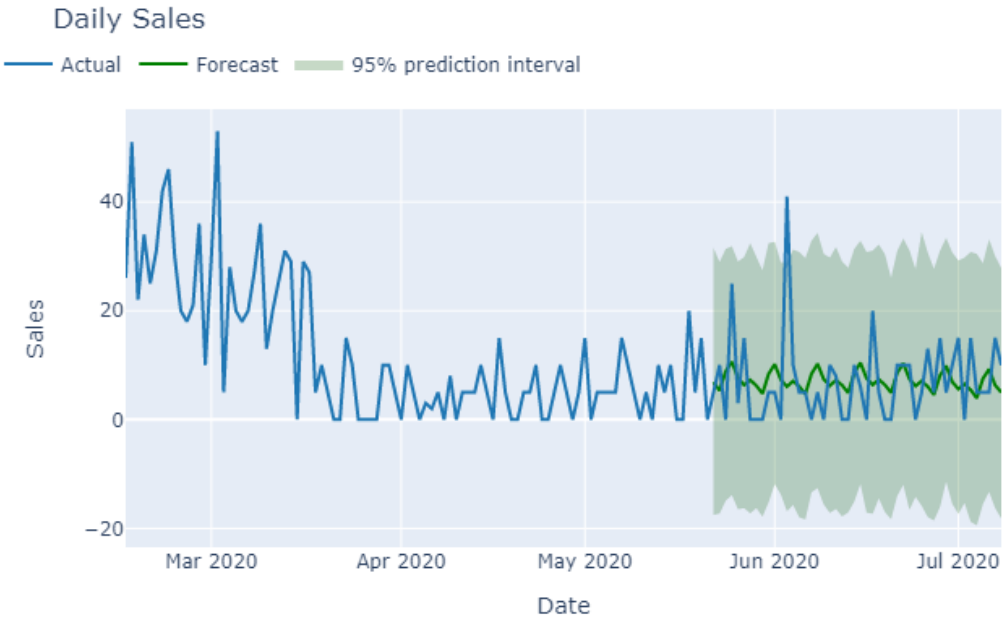

Figure A.124: Prophet Forecasts not using regressors for time series 1020

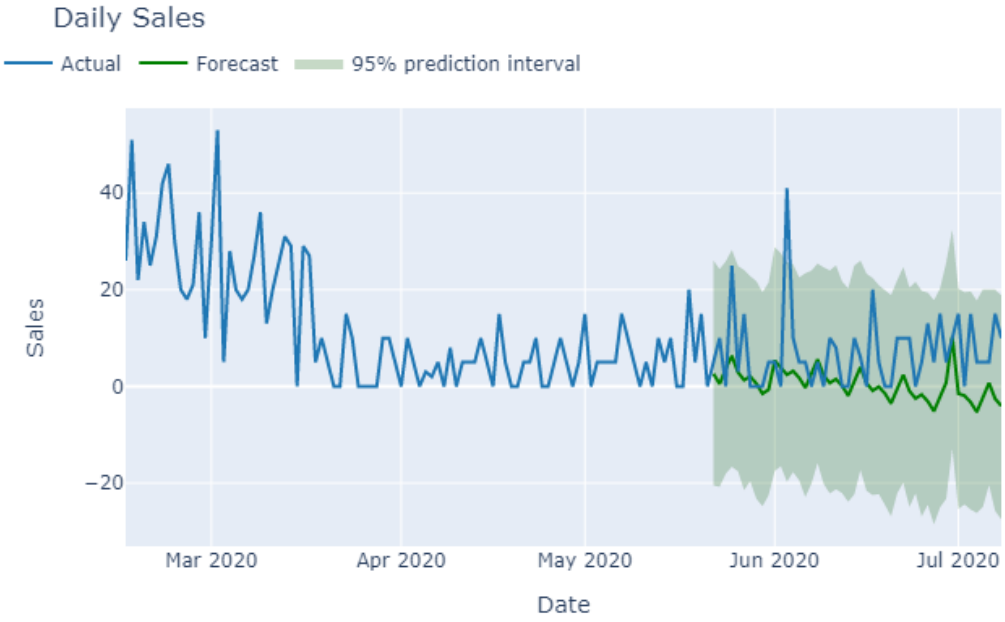

Figure A.125: Prophet Forecasts using regressors for time series 1020

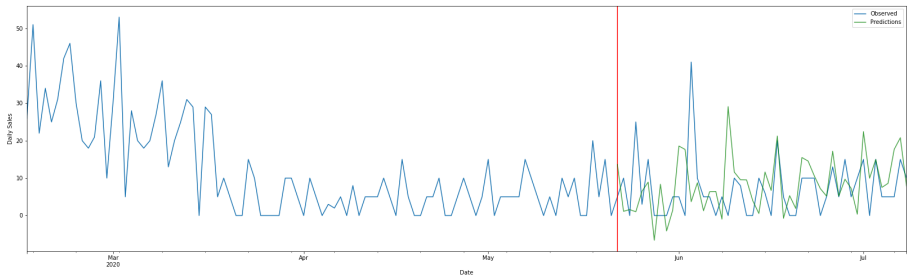

Figure A.126: Holt Winter's Forecasts using regressors for time series 1020

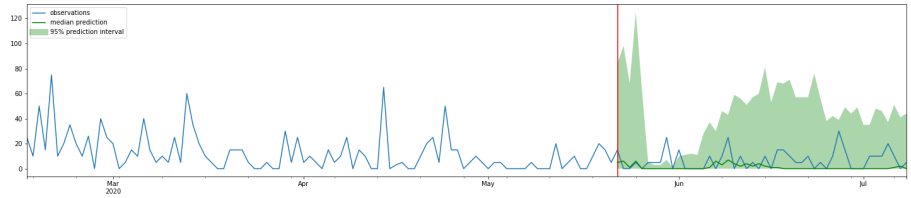

Figure A.127: Deep AR Forecasts not using dynamic features for time series 1460

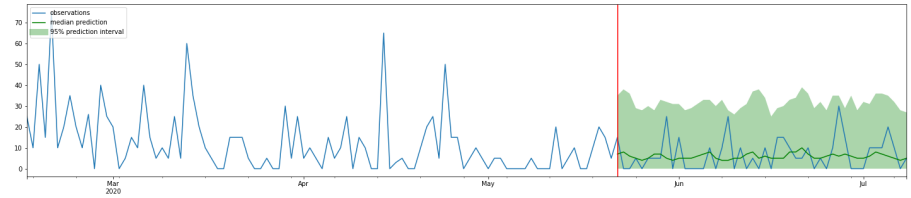

Figure A.128: Deep AR Forecasts using dynamic features for time series 1460

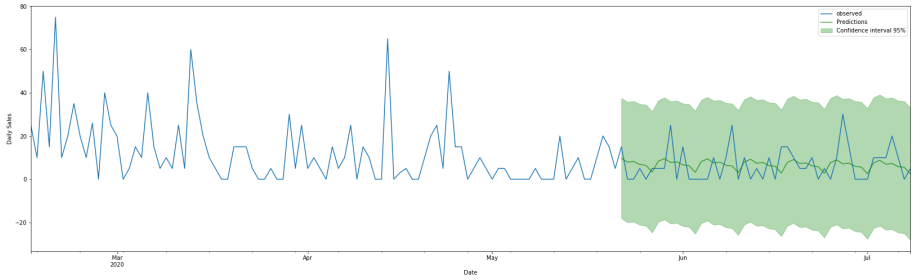

Figure A.129: SARIMA Forecasts for time series 1460

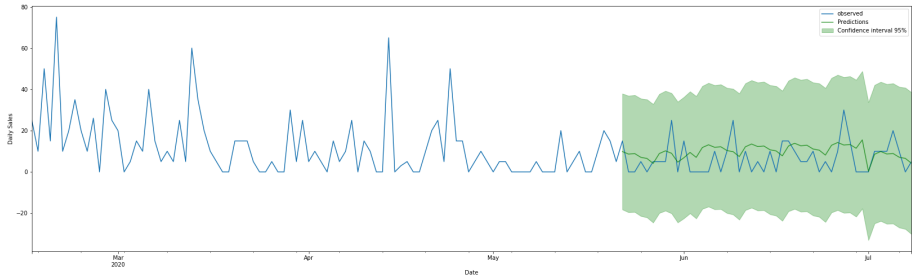

Figure A.130: SARIMAX Forecasts for time series 1460

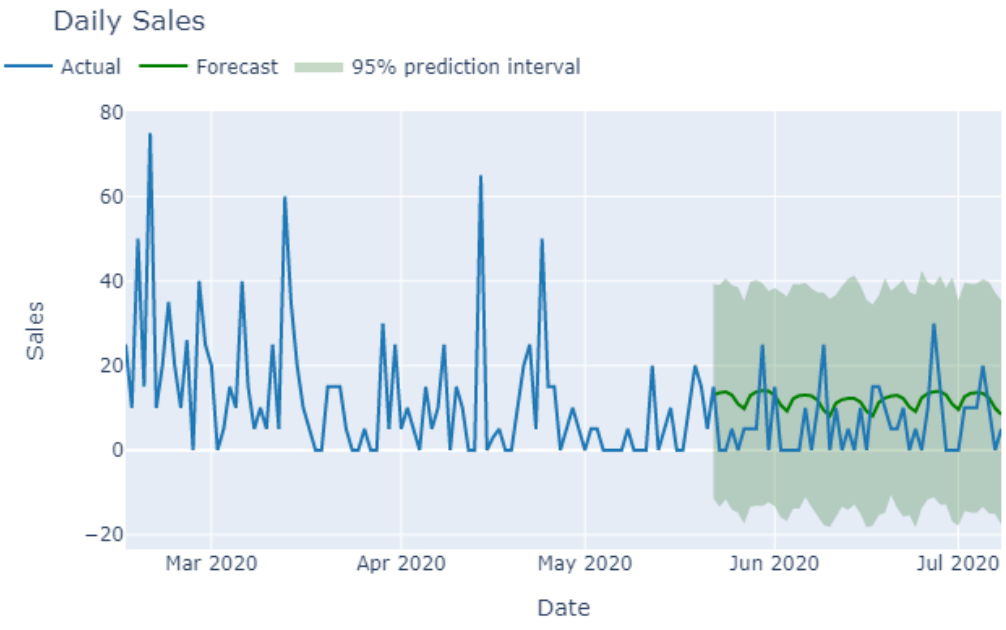

Figure A.131: Prophet Forecasts not using regressors for time series 1460

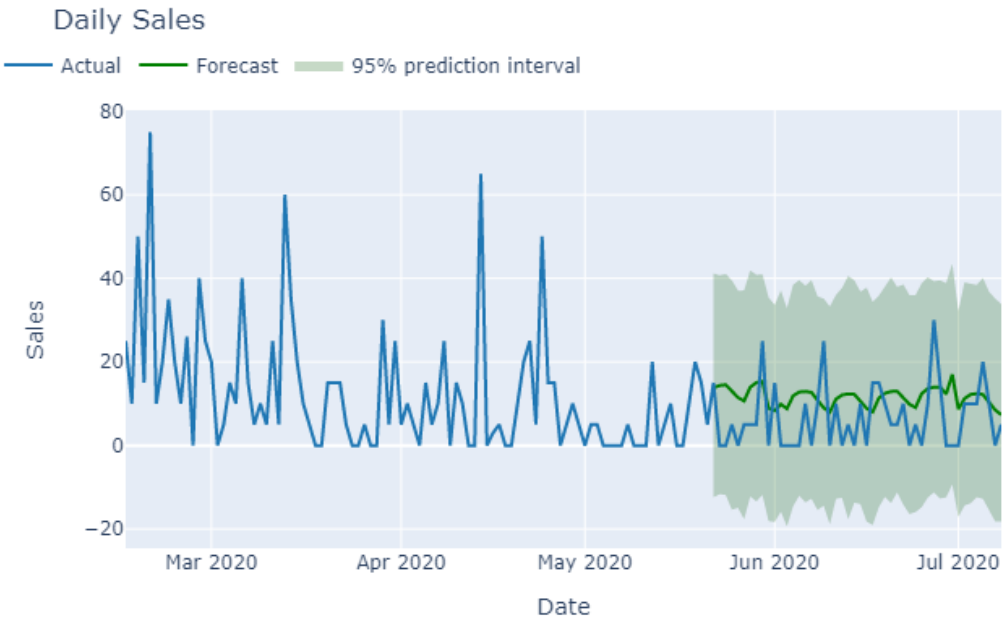

Figure A.132: Prophet Forecasts using regressors for time series 1460

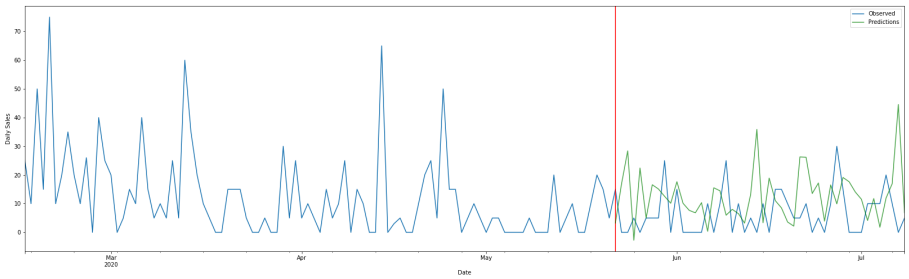

Figure A.133: Holt Winter's Forecasts using regressors for time series 1460

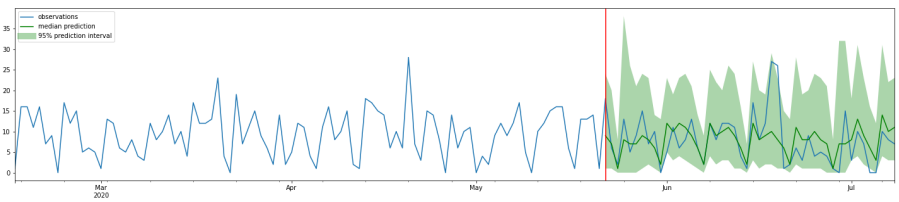

Figure A.134: Deep AR Forecasts not using dynamic features for time series 124

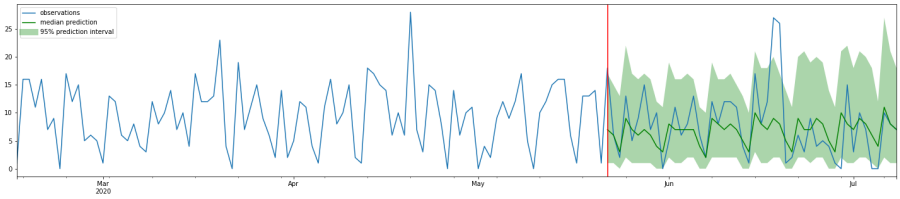

Figure A.135: Deep AR Forecasts using dynamic features for time series 124

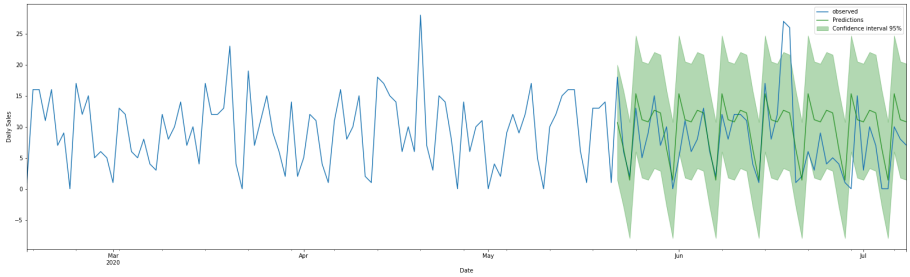

Figure A.136: SARIMA Forecasts for time series 124

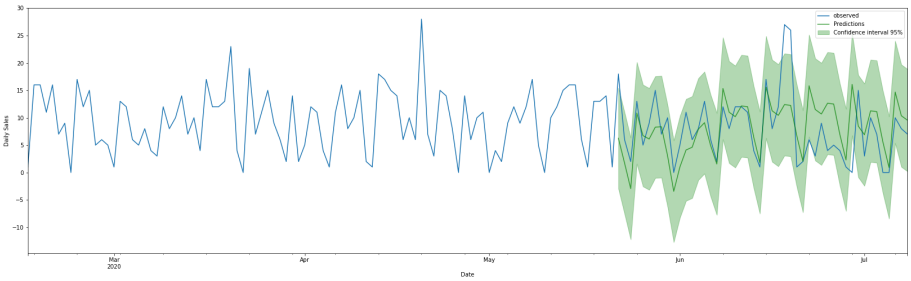

Figure A.137: SARIMAX Forecasts for time series 124

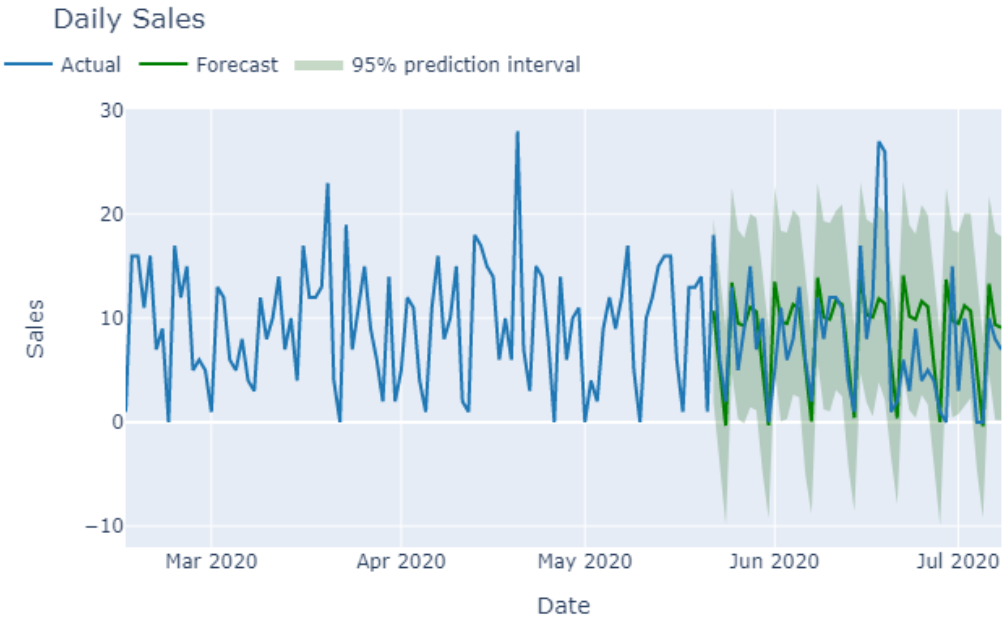

Figure A.138: Prophet Forecasts not using regressors for time series 124

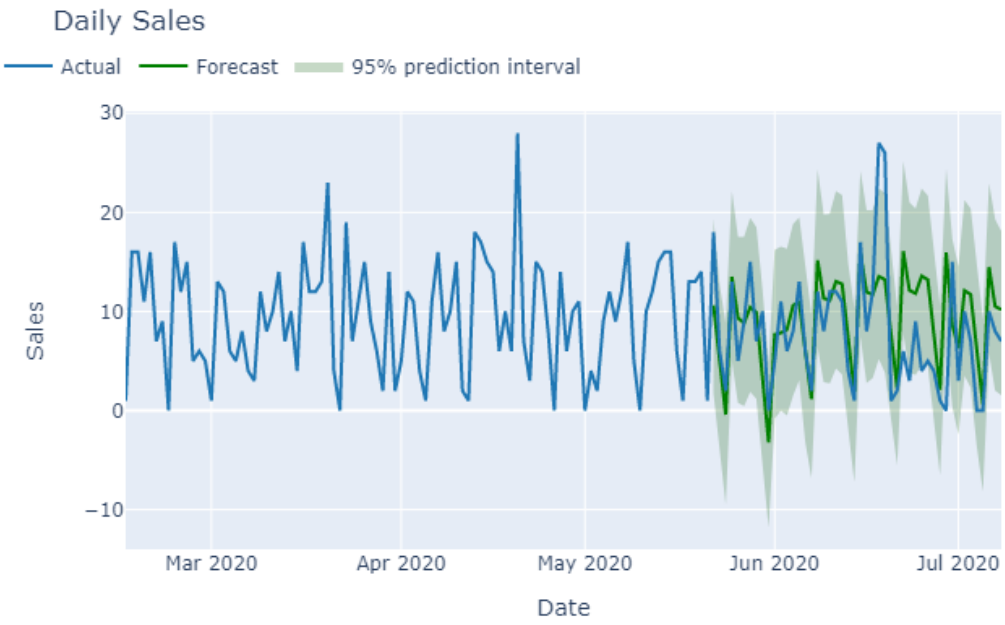

Figure A.139: Prophet Forecasts using regressors for time series 124

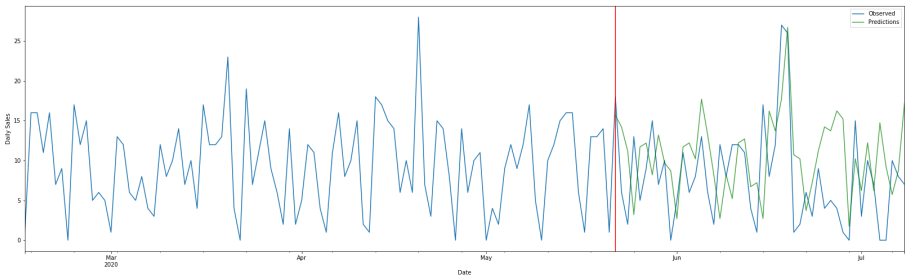

Figure A.140: Holt Winter's Forecasts using regressors for time series 124

References

[1] Luis Aburto and Richard Weber. Improved supply chain management based on hybrid demand forecasts. *Applied Soft Computing*, 7(1):136–144, 2007.

[2] Maureen D Agnew and John E Thornes. The weather sensitivity of the uk food retail and distribution industry. *Meteorological Applications*, 2(2):137–147, 1995.

[3] Kusum L. Ailawadi, Bari A. Harlam, Jacques César, and David Trounce. Promotion profitability for a retailer: The role of promotion, brand, category, and store characteristics. *Journal of Marketing Research*, 43(4):518–535, 2006.

[4] Thomas L Ainscough and Jay E Aronson. An empirical investigation and comparison of neural networks and regression for scanner data analysis. *Journal of Retailing and Consumer Services*, 6(4):205–217, 1999.

[5] A. Alexandrov, K. Benidis, M. Bohlke-Schneider, V. Flunkert, J. Gasthaus, T. Januschowski, D. C. Maddix, S. Rangapuram, D. Salinas, J. Schulz, L. Stella, A. C. Türkmen, and Y. Wang. GluonTS: Probabilistic Time Series Modeling in Python. *arXiv preprint arXiv:1906.05264*, 2019.

[6] Alexander Alexandrov, Konstantinos Benidis, Michael Bohlke-Schneider, Valentin Flunkert, Jan Gasthaus, Tim Januschowski, Danielle C. Maddix, Syama Rangapuram, David Salinas, Jasper Schulz, Lorenzo Stella, Ali Caner Türkmen, and Yuyang Wang. GluonTS: Probabilistic and Neural Time Series Modeling in Python. *Journal of Machine Learning Research*, 21(116):1–6, 2020.

[7] Özden Gür Ali, Serpil Sayın, Tom Van Woensel, and Jan Fransoo. Sku demand forecasting in the presence of promotions. *Expert Systems with Applications*, 36(10):12340–12348, 2009.

[8] Özden Gür Ali, Serpil Sayın, Tom Van Woensel, and Jan Fransoo. Sku demand forecasting in the presence of promotions. *Expert Systems with Applications*, 36(10):12340–12348, 2009.

[9] Ilan Alon, Min Qi, and Robert Sadowski. Forecasting aggregate retail sales: A comparison of artifcial neural networks and traditional methods. *Journal of Retailing and Consumer Services*, 8:147–156, 05 2001.

[10] J Scott Armstrong. Standards and practices for forecasting. In *Principles of Forecasting*, pages 679–732. Springer, 2001.

[11] J Scott Armstrong and Fred Collopy. Error measures for generalizing about forecasting methods: Empirical comparisons. *International journal of forecasting*, 8(1):69–80, 1992.

[12] Nari Sivanandam Arunraj and Diane Ahrens. A hybrid seasonal autoregressive integrated moving average and quantile regression for daily food sales forecasting. *International Journal of Production Economics*, 170:321–335, 2015.

[13] Nari Sivanandam Arunraj, Diane Ahrens, and Michael Fernandes. Application of sarimax model to forecast daily sales in food retail industry. *International Journal of Operations Research and Information Systems (IJORIS)*, 7(2):1–21, 2016.

[14] CLAUDIMAR PEREIRA Da Veiga, CÁSSIA RITA PEREIRA Da Veiga, Anderson Catapan, Ubiratã Tortato, and WESLEY VIEIRA Da Silva. Demand forecasting in food retail: A comparison between the holt-winters and arima models. *WSEAS transactions on business and economics*, 11(1):608–614, 2014.

[15] Andrey Davydenko and Robert Fildes. Measuring forecasting accuracy: The case of judgmental adjustments to sku-level demand forecasts. *International Journal of Forecasting*, 29:510–522, 07 2013.

[16] Gianni Di Pillo, Vittorio Latorre, Stefano Lucidi, and E Procacci. An application of support vector machines to sales forecasting under promotions. *4OR*, 14(3):309–325, 2016.

[17] Gianni Di Pillo, Vittorio Latorre, Stefano Lucidi, Enrico Procacci, et al. An application of learning machines to sales forecasting under promotions. *Control and Management Engineering*, 2013.

[18] Philip Doganis, Alex Alexandridis, Panagiotis Patrinos, and Haralambos Sarimveis. Time series sales forecasting for short shelf-life food products based on artificial neural networks and evolutionary computing. *Journal of Food Engineering*, 75(2):196–204, 2006.

[19] Alexandre Dolgui and Maksim Pashkevich. Demand forecasting for multiple slow-moving items with short requests history and unequal demand variance. *International Journal of Production Economics*, 112(2):885–894, 2008.

[20] Facebook. Prophet. https://facebook.github.io/prophet/. Accessed: 01/2021.

[21] Robert Fildes, Shaohui Ma, and Stephan Kolassa. Retail forecasting: Research and practice. *International Journal of Forecasting*, 2019.

[22] Cheng Guo and Felix Berkhahn. Entity embeddings of categorical variables. *arXiv preprint arXiv:1604.06737*, 2016.

[23] ÖZDEN GÜR ALI. Driver moderator method for retail sales prediction. *International Journal of Information Technology & Decision Making*, 12(06):1261–1286, 2013.

[24] Andrew C Harvey and Simon Peters. Estimation procedures for structural time series models. *Journal of Forecasting*, 9(2):89–108, 1990.

[25] M Ahsan Akhtar Hasin, Shuvo Ghosh, and Mahmud A Shareef. An ann approach to demand forecasting in retail trade in bangladesh. *International Journal of Trade, Economics and Finance*, 2(2):154, 2011.

[26] Jeremy Howard. Fastai course v3 - lesson 6. https://github.com/fastai/course-v3/blob/master/nbs/dl1/lesson6-rossmann.ipynb. Accessed: 01/2021.

[27] Rob Hyndman, Anne B Koehler, J Keith Ord, and Ralph D Snyder. *Forecasting with exponential smoothing: the state space approach*. Springer Science & Business Media, 2008.

[28] Rob J Hyndman, Yeasmin Khandakar, et al. *Automatic time series for forecasting: the forecast package for R*. Monash University, Department of Econometrics and Business Statistics ..., 2007.

[29] Rob J Hyndman and Anne B Koehler. Another look at measures of forecast accuracy. *International journal of forecasting*, 22(4):679–688, 2006.

[30] Stefan Lang, Winfried J Steiner, Anett Weber, and Peter Wechselberger. Accommodating heterogeneity and nonlinearity in price effects for predicting brand sales and profits. *European Journal of Operational Research*, 246(1):232–241, 2015.

[31] Muhammad Hisyam Lee and N Hamzah. Calendar variation model based on arimax for forecasting sales data with ramadhan effect. In *Proceedings of the Regional Conference on Statistical Sciences*, pages 349–361, 2010.

[32] Ana LD Loureiro, Vera L Miguéis, and Lucas FM da Silva. Exploring the use of deep neural networks for sales forecasting in fashion retail. *Decision Support Systems*, 114:81–93, 2018.

[33] Levy M., Barton A. Weitz, and Grewal D. *Retailing Management (8th Edition ed.)*. McGraw-Hill College, 2012.

[34] Essam Mahmoud. Accuracy in forecasting: A survey. *Journal of Forecasting*, 3(2):139–159, 1984.

[35] Spyros Makridakis, Steven C Wheelwright, and Rob J Hyndman. *Forecasting methods and applications*. John wiley & sons, 2008.

[36] Sam Mourad. Prophet vs deepar: Forecasting food demand. https://towardsdatascience.com/prophet-vs-deepar-forecasting-food-demand-2fdebfb8d282. Accessed: 01/2021.

[37] J Peters. Improving the promotional forecasting accuracy for perishable items at sligro food group bv. *Eindhoven University of Technology, Netherlands*, 2012.

[38] J Pinho. Previsão de vendas no setor do retalho sob o efeito de ações promocionais. *Universidade do Porto*, 2015.

[39] Rahul Priyadarshi, Akash Panigrahi, Srikanta Routroy, and Girish Kant Garg. Demand forecasting at retail stage for selected vegetables: a performance analysis. *Journal of Modelling in Management*, 2019.

[40] Usha Ramanathan and Luc Muyldermans. Identifying demand factors for promotional planning and forecasting: A case of a soft drink company in the uk. *International journal of production economics*, 128(2):538–545, 2010.

[41] P Ramos and R Fildes. Characterizing retail demand with promotional effects. In *International Symposium on Forecasting*. International Institute of Forecasters Cairns, Australia, 2017.

[42] Patrícia Ramos, Nicolau Santos, and Rui Rebelo. Performance of state space and arima models for consumer retail sales forecasting. *Robotics and computer-integrated manufacturing*, 34:151–163, 2015.

[43] David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 2019.

[44] Ardalan Sameti and Hamidreza Khalili. Influence of in-store and out-of-store creative advertising strategies on consumer attitude and purchase intention. *Intangible Capital*, 13:523–547, 07 2017.

[45] Matthias W Seeger, David Salinas, and Valentin Flunkert. Bayesian intermittent demand forecasting for large inventories. In *Advances in Neural Information Processing Systems*, pages 4646–4654, 2016.

[46] Suresh Kumar Sharma and Vinod Sharma. Comparative analysis of machine learning techniques in sale forecasting. *International Journal of Computer Applications*, 53(6), 2012.

[47] Maxim Vladimirovich Shcherbakov, Adriaan Brebels, Nataliya Lvovna Shcherbakova, Anton Pavlovich Tyukov, Timur Alexandrovich Janovsky, and Valeriy Anatol'evich Kamaev. A survey of forecast error measures. *World Applied Sciences Journal*, 24(24):171–176, 2013.

[48] Ralph D Snyder, J Keith Ord, and Adrian Beaumont. Forecasting the intermittent demand for slow-moving inventories: A modelling approach. *International Journal of Forecasting*, 28(2):485–496, 2012.

[49] James W Taylor. Forecasting daily supermarket sales using exponentially weighted quantile regression. *European Journal of Operational Research*, 178(1):154–167, 2007.

[50] Sean J Taylor and Benjamin Letham. Forecasting at scale. *The American Statistician*, 72(1):37–45, 2018.

[51] Grigorios Tsoumakas. A survey of machine learning techniques for food sales prediction. *Artificial Intelligence Review*, 52(1):441–447, 2019.

[52] MJ Thijs van der Poel. Improving the promotion forecasting accuracy at unilever netherlands. *Eindhoven University of Technology*, 2010.

[53] Jack GAJ van der Vorst, Andrie JM Beulens, W De Wit, and Paul van Beek. Supply chain management in food chains: Improving performance by reducing uncertainty. *International Transactions in Operational Research*, 5(6):487–499, 1998.

[54] Karel H Van Donselaar, Jordi Peters, Ad de Jong, and Rob ACM Broekmeulen. Analysis and forecasting of demand during promotions for perishable items. *International Journal of Production Economics*, 172:65–75, 2016.

[55] Claudimar Pereira da Veiga, CÃ¡ssia Rita Pereira Da, Veiga, Weslly Puchalski, Leandro dos Santos Coelho, and Tortato UbiratÃ£. Demand forecasting based on natural computing approaches applied to the foodstuff retail segment. *Journal of Retailing and Consumer Services*, 31(C):174–181, 2016.

[56] Indrė Žliobaitė, Jorn Bakker, and Mykola Pechenizkiy. Beating the baseline prediction in food sales: How intelligent an intelligent predictor is? *Expert Systems with Applications*, 39(1):806–815, 2012.