**OVERVIEW OF SALES FORECASTING**

PROCESS Ideally the sales forecasting process needs different kinds of information from different departments. This could be done by integrating different departments by means of common information system. The business functions of departments like production, sales, purchasing, planning, finance and logistics, supply chain department are different ones, the strategies used by them also differ. While integrating the functions, all the departments share their information on a central information system which is joined with central database warehouse. As all functions work towards the same target, sharing information facilitates productive work. Coordination can be achieved by a central information system. Furthermore sales forecasting also runs on information like marketing, sales, production planning and logistics. All departments need the sales forecasting to plan their activities effectively. Integration of coordination may be the best way to achieve integrated and interactive forecast. The managers at the different functional areas will make the decisions using the data available in the central repository.

**IMPACT OF SCM ON SALES FORECASTING**

Supply chain is driven by customer needs. In order to ship products to customers who demand them in a dynamic and rapidly changing set of channels, strategic planning of the inventories becomes essential. The model products are available when the customer wants them. Smooth supply chain management is the result of well combined individual managerial functions like planning, organizing and controlling. They are important factors for effective supply chain management1. An exact estimation and reliable predictions of product volume and related services are therefore important for efficient functioning; and such estimations are nothing but forecasts. Demand / Sales forecasting is a crucial factor to any firm. It is a basic factor for the whole planning process and for the control of various sectors of a company such as production, supply, purchasing, marketing and finance. Decisions made in these sectors influence each other either directly or indirectly. Variability in product demand influences several factors which are coming under these sectors such as inventory needed for production (Production department), capital needs (Finance department) or special kind of strategies like outsourcing (Supply department). Forecasting is necessary for a variety of decisions, strategic decisions involving such things as constructing a new plant, developing more supplier capacity, expanding internationally, and other long-run company wide considerations. Forecasts for this type of decisions are highly aggregated estimates of general business trends over the long term.

The general principle indicated here is that the nature ofthe forecast must be matched with the nature ofthe decision. The level of aggregation, the amount of management review, the cost, and the time frame of the forecast needed really depends on the nature ofthe decisions being made3. Moreover, the source ofthe forecast can vary by need as well, as indicated by the useful techniques in the above table 3.2. The frequency and the number of forecasts made in the need for most short term operating decisions don’t warrant extensive management involvement, so computergenerated forecasts are utilized. Strategic decisions, on the other hand, are less frequent and involve more risk, thus justifying the use of more expensive procedures and management involvement .

**REASONS FOR UNDERTAKING SALES FORECASTS** **BY A SUPERMARKET**

Businesses are forced to look well ahead in order to plan their investments, launch new products, decide when to close or withdraw products and so on. The sales forecasting process is a critical one for most businesses and the critical areas where sales forecasting is necessary in a company are discussed in the following section.

**Sales forecasting need in Planning in a supermarket or a company , business organization**

Manufacturing industries work on principle to satisfy customer demand by appropriate supply. According to Mentzer and Moon (2005), companies consider the sales forecasting as integral part of this process. End customers create demand and it can be increased by activities like. Hence marketing focuses on end customers for creating demand. To meet this created demand supply should be sufficient and smooth. Different management functions like manufacturing, purchasing and logistics work together to maintain the supply. Different suppliers also play an important role in this chain. A constant flow of information runs through the complex structure of different management functions and the parties involved in this. According to Mentzer and Moon, (2005) this flow starts with demand and ends with supply functions. This flow of information is managed by sales and operational process (S&OP). Sales forecasting serves as the initial seeding to the S&OP process. The forecasting may originate from a study of past demand history. As the marketing function originates and manages the demand toward final customer, the necessity of sales forecasting arises from the demand side. Based on the sales forecasting, supply side prepares the capacity plan. The capacity plan is nothing but the capabilities to satisfy demand using maximum possible inputs via information net both forecasting and capacity plans studied out to consider strategies. Mentzer and Moon, (2005) describe two major plans in this process, operation plan and demand plan respectively. Considering different information collected from time to time and strategies undertaken, the demand plans are given out from S & OP system. The demand plans make marketing and supply departments understand future product launching and action needed to achieve corporate strategies. Based on the information available; the operation plan is given out from S&OP to supply functions. This plan consists of different functional plans. Smooth running of S&OP needs accurate forecasting. Continuing with Mentzer’s S&OP model, Armstrong (1983), shows how forecasting process is correlated to formal planning. Planning is a set of activities in company. Planning decides goals and action is taken accordingly. Wood, Robley (1980), Yen and Andrew (1980), Armstrong 97 (1983) describe four steps in planning: 1) specify objectives; 2) generate strategies; 3) evaluate strategies and take actions accordingly; 4) monitor results. Commitment towards the basic goai is a key towards success. But still the accurate forecast plays a major role in successful planning and achieving ofthe final goal. One needs to understand the difference between planning and forecasting. According to Armstrong (1983), forecasting is the process to give estimates and planning is the process to prepare strategies based on these estimates.

**Sales forecasting need in financial planning** **in a super market or a business organization**

Forecasting and planning are essential to good decision making, as a result financial manger has to use history oriented financial statement with a suitable model. One ofthe model that is followed is Francis-Rowell(FR) model. The objective of this model is to generate pro forma financial statements that describes the future financial condition ofthe firm for any assumed pattern ofsales. The FR model is composed often sectors they are 1) Industry sales 2) Production sector 3) Fixed capital stock requirements 4) Pricing 5) Production costs 6) Income 7) New financing required 8) Risks 9) Costs offinancing and 10) Commong stock valuation5.Forecasting and planning are essential to good decision making, as a result financial manger has to use history oriented financial statement with a suitable model. One ofthe model that is followed is Francis-Rowell(FR) model. The objective of this model is to generate pro forma financial statements that describes the future financial condition ofthe firm for any assumed pattern ofsales. The FR model is composed often sectors they are 1) Industry sales 2) Production sector 3) Fixed capital stock requirements 4) Pricing 5) Production costs 6) Income 7) New financing required 8) Risks 9) Costs offinancing and 10) Commong stock valuation.

**FACTORS FOR SELECTING THE FORECASTING MODEL** **BY A SUPER MARKET OR A BUSINESS ORGANIZATION**

**There are two major types of forecasting, which can be broadly described as macro and micro: Macro forecasting is concerned with forecasting markets in total. This is about determining the existing level of market demand and considering what will happen to market demand in the future. Micro forecasting is concerned with detailed unit sales forecasts. This is about determining a product’s market share in a particular industry and considering what will happen to that market share in the future. The selection ofwhich type offorecasting and the forecasting accuracy depend on several factors:**

**Time Horizon**

Most researchers agree that the longer the time horizon, the less accurate the forecast. However, most ofthe researchers disagree on which method to choose for a given time horizon. However Armstrong contends that simple econometric models are superior for long range forecasts. Chambers, Mullick and Smith7 argue that simple growth curves are also an accurate means of generating long term forecasts, whereas short term forecasts are provided by time series methods.

**Data availability**

Some researchers say that more accurate forecasts are possible when a greater amount of data is available8. However Markridakis and Hibson found no support for this contention in the series they studied. Possibly longer data series are more likely to involve changes in the underlying forces affecting the series that “confuse” a forecasting method and result in a loss of accuracy.

**Type of product**

Most of the authors agree that one of the key factors affecting forecast accuracy is the historical stability of a series. Forecasts ofunstable series are inaccurate. Stability can be measured in many ways, most simple isjudgemental assessment derived from an inspection of a scatter diagram of historical sales. More formal methods of assessing stability are also available, including autocorrelation and “runs” analysis.

**FORECASTING METHODS AND TECHNIQUES**

The ability to model and perform decision modelling and analysis is an essential feature of many real-world applications ranging from emergency medical treatment in intensive care units to military command and control systems. Existing formalisms and methods of inference have not been effective in real-time applications where tradeoffs between decision quality and computational tractability are essential. In practice, an effective approach to time-critical dynamic decision modelling should provide explicit support for the modelling of temporal processes and for t -I dealing with decisions . Almost all managerial decisions are based on forecasts. Every decision becomes operational at some point in the future. So it should be based on forecasts of future conditions. Forecasts are needed throughout an organization — and they should certainly not be produced by an isolated group of forecasters. Neither is forecasting ever "finished". Forecasts are needed continually, and as time moves on, the impact of the forecasts on actual performance is measured; original forecasts are updated; and decisions are modified, and so on. For example, many inventory systems cater to uncertain demand. The inventory parameters in these systems require estimates of the demand and forecast error distributions. The two stages of these systems, forecasting and inventory control, are often examined independently. Most studies tend to look at demand forecasting as if this were an end in itself, or at stock control models as if there were no preceding stages of computation. Nevertheless, it is important to understand the interaction between demand forecasting and inventory control since this influences the performance of the inventory system.

**FEW TERMS IN SALES FORECASTING**

Sales demand for an item or package is the projected total trades capacity in a business for a precise interval time in a well-defined location, under a clear advertising system or consumption. It is a term related with different stages of industry publicizing costs.

**Sales Prediction:**

Sales Prediction is the anticipated market request a particular level of industry promoting expenses.

**Sales Probable:**

Sales Probable is the determined market request, ensuing from a very superior level of business publicizing spending, where additionally increments in spending can have little impact on demand.

**Sales Potential:**

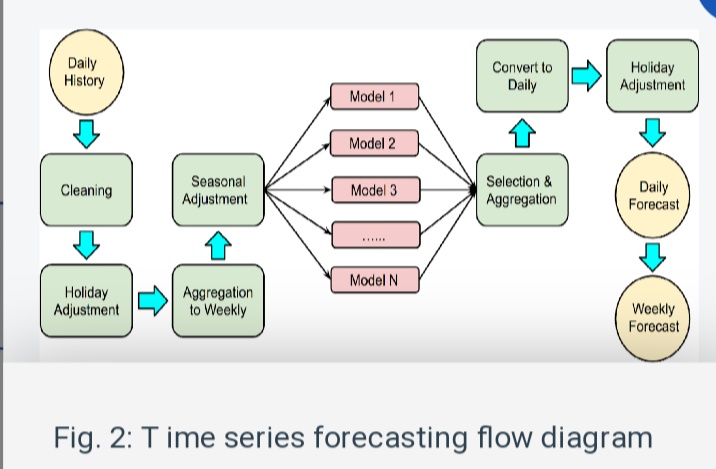
Sales potential is the expected business trades of a item or package, based on cut of market latent estimated by the firm.

**Sales Prediction:**

Sales Prediction is assessed business sales of a service or merchandise, based on a preferred publicizing costs plan, for an exact time interval, in an anticipated marketing atmosphere.

**Time Series:**

Moving Normal Model The Moving Normal model uses the last t time frames with a specific end goal to anticipate request in period t+1. There are two kinds of moving normal models: straightforward moving normal and weighted moving normal. The moving normal model presumption is that the most precise expectation of forthcoming request is a straightforward (direct) blend of previous request. In the simple moving normal models, the projection value is .



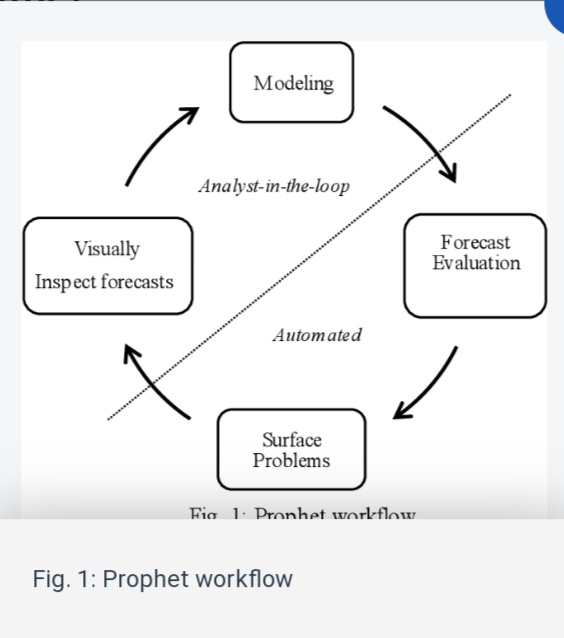
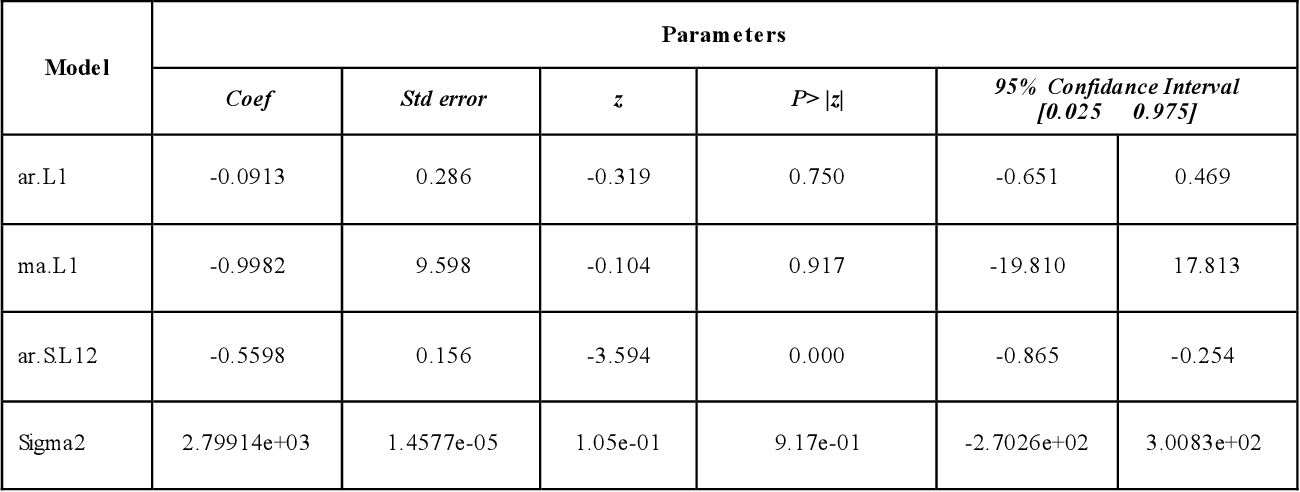
**Time series analysis**

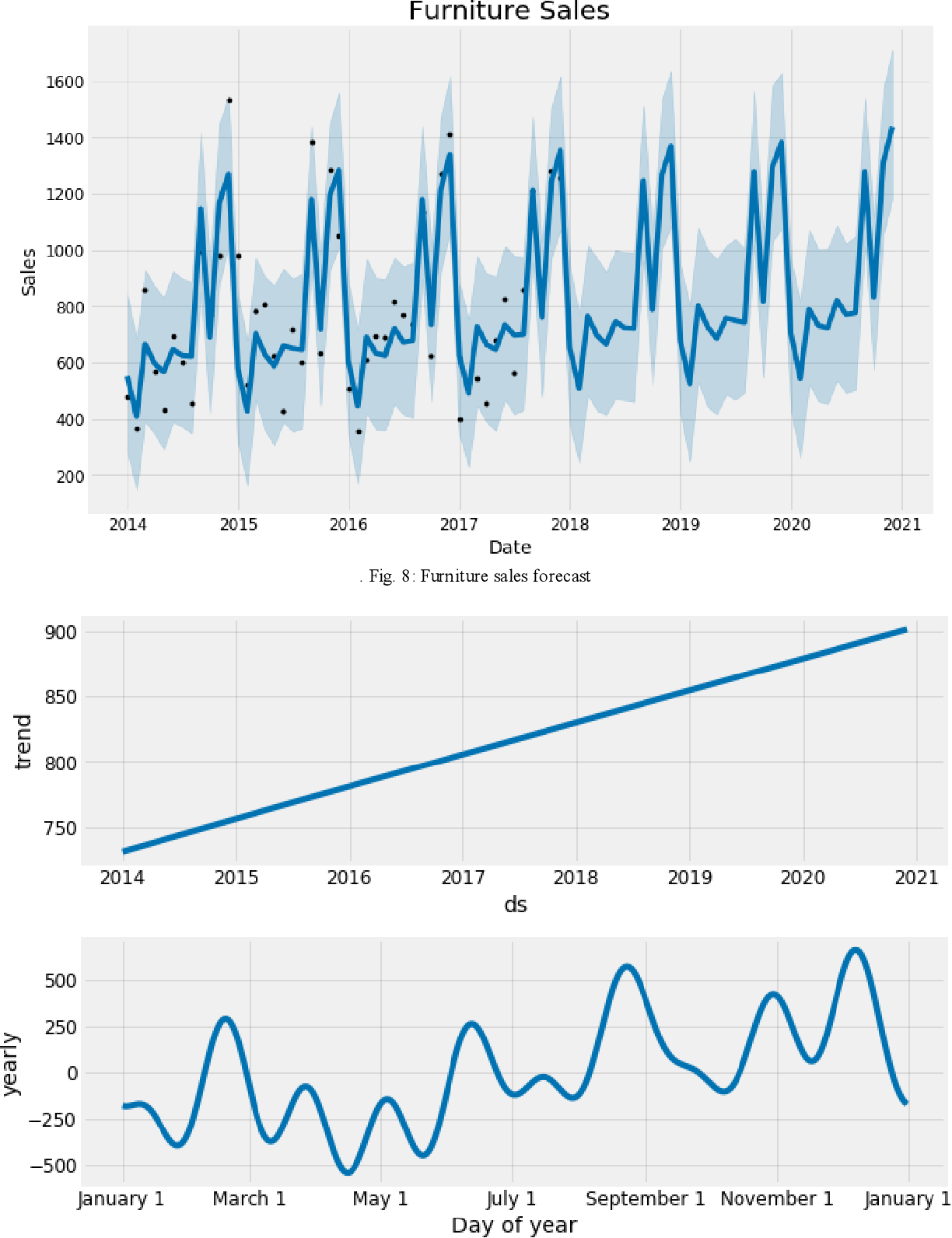
A time series is a set of sequential data sampled in a specific time unit, used to record a process output to enable the analysis of its evolution (CLEVELAND et al., 1990). Some examples of widely used time series include stock market price variation (LEUNG; MACKINNON; WANG, 2014; MONDAL; SHIT; GOSWAMI, 2014), daily temperature values (SANTER et al., 2000), and heart rate signals (CHRISTINI et al., 1995). A time series is also a signal that can be processed using signal processing and/or statistical methods. Indeed, several different modeling and decomposition approaches can be employed in this context. For instance, a hypothetical probability model can be set up to represent the data as a stochastic process (CHATFIELD, 1975). It is also possible to make seasonal adjustments so that long-term trends are noticed, filter out noise signals, predict future values, and perform hypothesis tests (BROCKWELL et al., 2016a). In this section, we discuss relevant preliminary concepts related to time series that are required to understand this work. Initially, we overview time series structure and processing. Later, important concepts for time series forecasting and well-established methods are briefly reviewed.

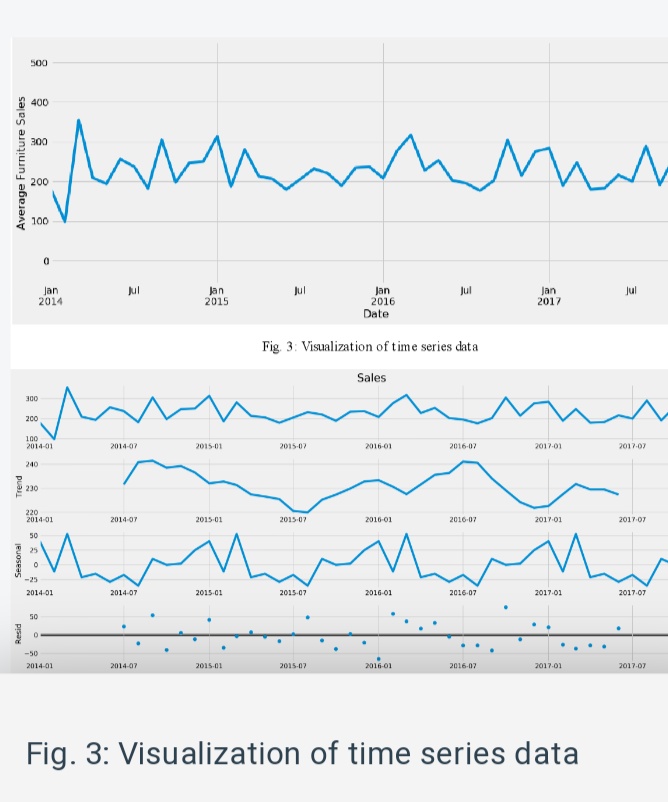
**Structure**

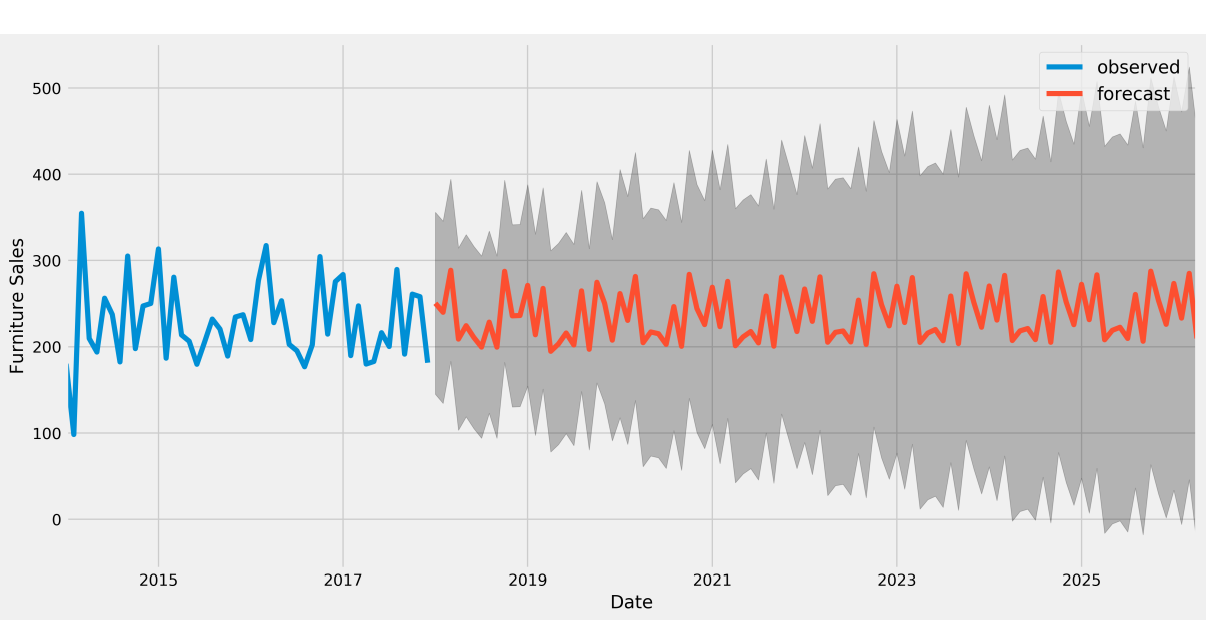
A time series model is an specification of the joint distributions of a sequence of random variables (BROCKWELL et al., 2016b). Effectively, for the purposes of this work, a time series model is a mathematical description of a set of random variables collected and indexed at fixed time intervals, with regards to its first- and second-order moments1 . The simplest time series model is the white noise, composed of values by time records with zero mean and constant variance, which are independent and identically distributed. Though it is not predictable, this type of series is important to model more complex time series. Another type of well-known time series model is the random walk. Unlike noise, the random walk is given by the cumulative sum of random variables. From these series, other types of relevant time series can be defined. A few components stand out in time series and can be used to model the processes, namely trend (T), seasonality (S), and residuals (R, or noise). These three components are the most present in a time series (BROCKWELL et al., 2016a), and comprise the formal definition given in Equation 2.1. We further detail each of these components below.

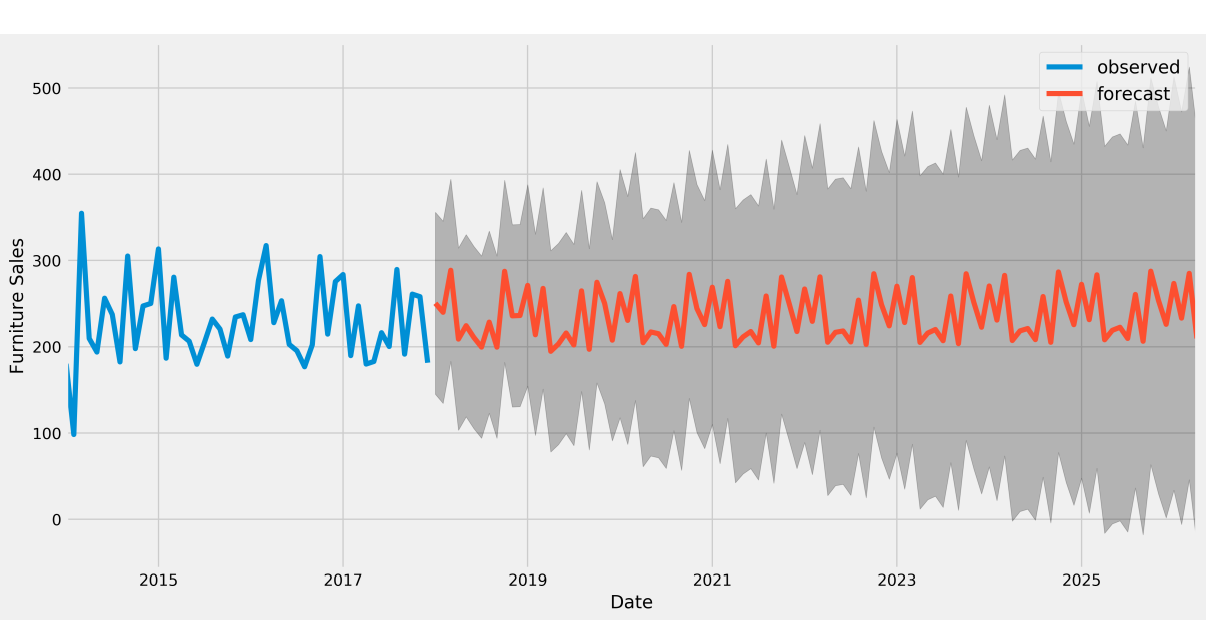
**Y (t) = T(t) + S(t) + R(t)**

****

****

****

****



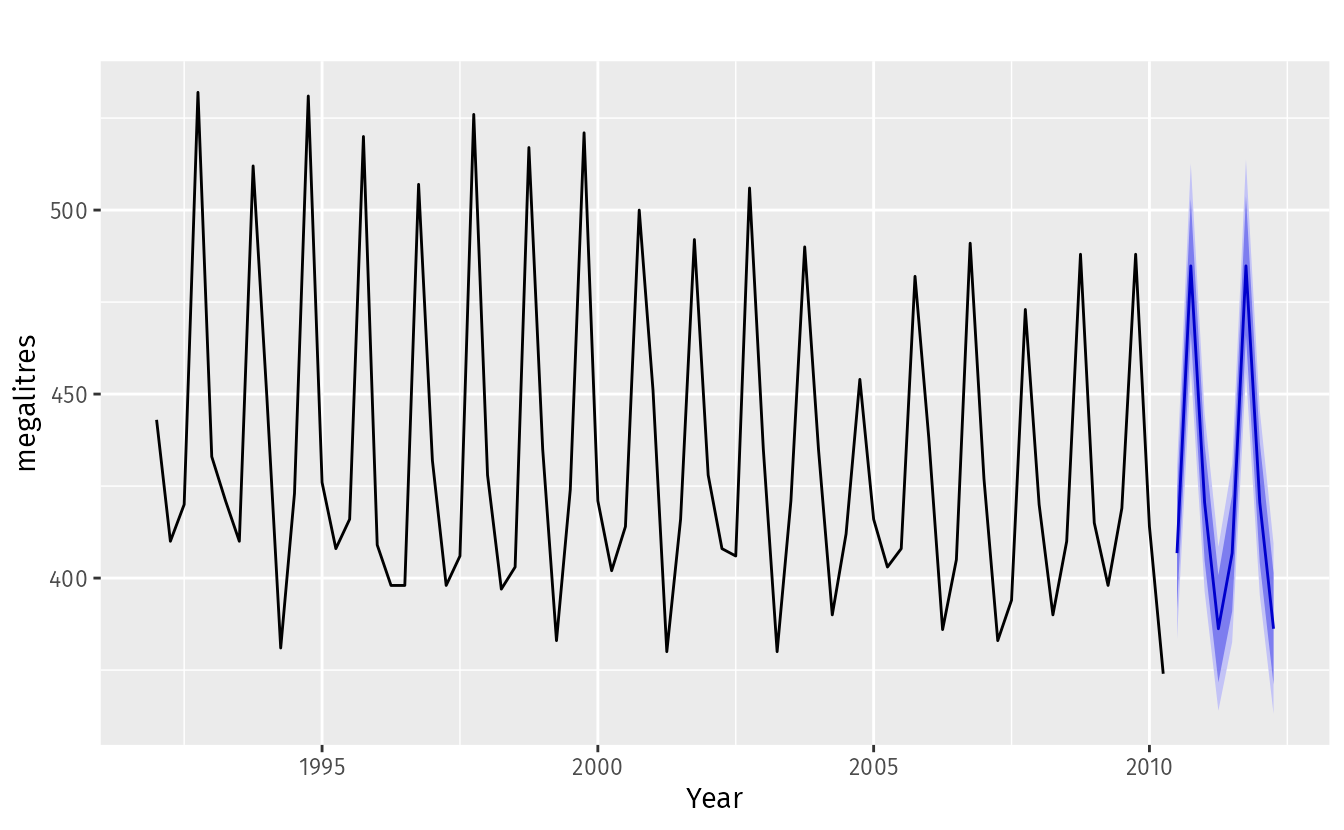
****

Figure 1.1: Australian quarterly beer production: 1992Q1–2010Q2, with two years of forecasts.

**Time plots**

For time series data, the obvious graph to start with is a time plot. That is, the observations are plotted against the time of observation, with consecutive observations joined by straight lines. Figure [2.1](https://otexts.com/fpp2/time-plots.html#fig:ansett) below shows the weekly economy passenger load on Ansett Airlines between Australia’s two largest cities.

autoplot(melsyd[,"Economy.Class"]) +

ggtitle("Economy class passengers: Melbourne-Sydney") +

xlab("Year") +

ylab("Thousands")

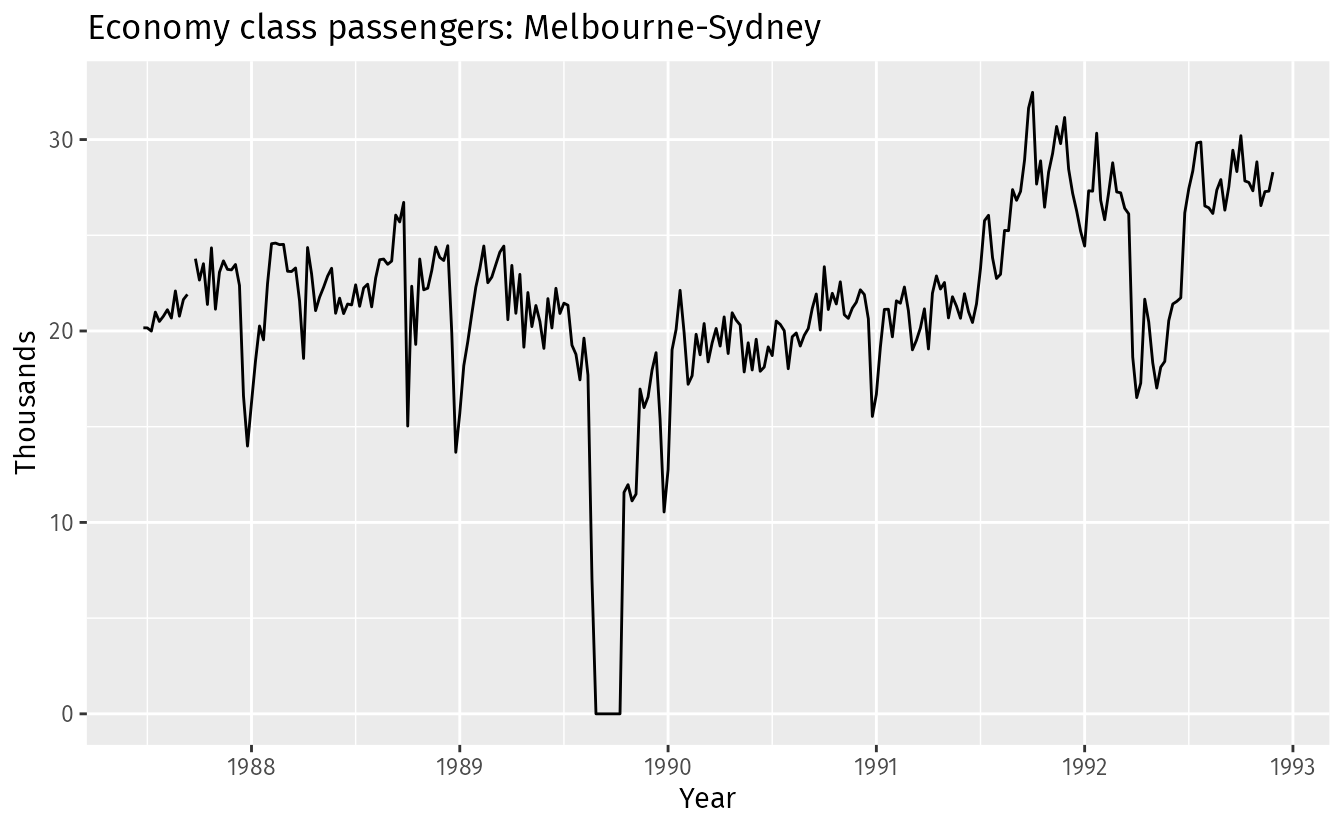
****

Figure 2.1: Weekly economy passenger load on Ansett Airlines.

**METHODOLOGY**

Accumulated retail data for unit sales was available at state, city, store and item level for each day between 2013 and 2017. Items are further classified by family, class and perishability. Each store was classified by its client defined cluster, type and location. Furthermore, holidays relevant by location and oil prices over the span of observation time were also included. Our main focus was to predict the sales of Bread/Bakery products as they are perishable in nature, making it important to understand the demand accurately to avoid shortage as well as wastage. For our study, we extracted data for one year, spanning across August 2016 to August 2017 as sales trends are best predicted by recent trends. The data available needed some preprocessing steps for it to be compatible with the different modelling techniques. • Pre-processing: o Pulled out observations relevant to the selected time frame (Aug-16 to Aug-17) and family of products (Bread / Bakery) o Identified outliers in the data

through exploratory data analysis and dropped observations where sales is less than 0 (indicating returns) or greater than 300

**MODEL(S) Linear Regression**

In linear regression, we fit a model based on the relationship between the dependent variable and the set of independent variables. The model identifies the best model by minimizing the mean of squared errors between the observed and predicted values. Since this method assumes the relationship to be linear, it tends to have a high bias, which we counter by training the model on 6 folds of data. Further we used backward as well as forward selection methods to identify the most important variables and choose the model with best performance measures. We differentiated the least squares formula and equate it to zero to get the model coefficients, leading to a closed form given by:

2 Gradient Boosting Method In this technique, predictors are chosen using decision trees which essentially divide the data set into smaller data sets based on the descriptive features until you reach a small enough set that contains data points that fall under one label with characteristic properties. • Boosting is an ensemble technique in which the predictors are not made independently, but sequentially. • Gradient Boosting is an example of boosting algorithm which employs the logic in which the subsequent predictors learn from the mistakes of the previous predictors. So, the intuition behind gradient boosting algorithm is to repetitively leverage the patterns in residuals and strengthen a model with weak predictions and make it better. The objective function of the algorithm is to minimize the residual error between the predicted value and the true observed value. • Tuning Parameters: o n.trees – Number of trees o interaction.depth (Maximum nodes per tree) - number of splits it has to perform on a tree o Shrinkage (Learning Rate) – It is considered as a learning rate. o n.minobsinnode - the minimum number of observations in trees' terminal nodes o bag.fraction (Subsampling fraction) - the fraction of the training set observations randomly selected to propose the next tree in the expansion. o train.fraction - The first train.fraction \* nrows(data) observations are used to fit the gbm and the remainder are used for computing out-of-sample estimates of the.

�" = �(∑ �)�)) − (∑ �)) (∑ �)) � ∑ �) - − (∑ �)) 2 �/0 (∑ �) -)(∑ �)) − (∑ �)) (∑ �)�)) � ∑ �) - − (∑ �)) 2

Gradient Boosting Method In this technique, predictors are chosen using decision trees which essentially divide the data set into smaller data sets based on the descriptive features until you reach a small enough set that contains data points that fall under one label with characteristic properties. • Boosting is an ensemble technique in which the predictors are not made independently, but sequentially. • Gradient Boosting is an example of boosting algorithm which employs the logic in which the subsequent predictors learn from the mistakes of the previous predictors. So, the intuition behind gradient boosting algorithm is to repetitively leverage the patterns in residuals and strengthen a model with weak predictions and make it better. The objective function of the algorithm is to minimize the residual error between the predicted value and the true observed value. • Tuning Parameters: o n.trees – Number of trees o interaction.depth (Maximum nodes per tree) - number of splits it has to perform on a tree o Shrinkage (Learning Rate) – It is considered as a learning rate. o n.minobsinnode - the minimum number of observations in trees' terminal nodes o bag.fraction (Subsampling fraction) - the fraction of the training set observations randomly selected to propose the next tree in the expansion. o train.fraction - The first train.fraction \* nrows(data) observations are used to fit the gbm and the remainder are used for computing out-of-sample estimates of the loss function Model Approach: 1. Initialize model with a constant value function .

**�/(�) = ���5 min9�(�), �) = )0" 2. For m = M ; M: number of iterations (for i = 1 to n) • Compute pseudo residuals �)> = − ? ��(�), �(�))) ��(�)) A B(C)0BDEF(C) • Fit a base learner (e.g. Tree) hm (x) to pseudo residuals (train using training set) {(�), �)>)})0" = • Compute multiplier �>r by solving the following one-dimensional optimization problem: �> = ���5 min9�(�), �>I"(�) + �ℎ>(�)) = )0" • Update the model �>(�) = �>I"(�) + �ℎ>(�) 3. Output �>(�) Neural Networks An Artificial Neural Network (ANN) is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. Neural networks, with their remarkable ability to derive meaning from complicated or imprecise data, can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. A neural network can learn how to do tasks based on the data given for training and it can create its own organization or representation of the information it receives during learning time. Our model implements one hidden layer that trains through responsibility assignment factor and we specify are tuning parameters through a tuning grid as follows: • Size = [1,2,3,4,5,6,7,8,9,10] • Decay = [0.1,0.2,0.3,0.4,0.5] Our business problem deals with predicting the volume of sales of bakery products per day, which essentially is a continuous variable. All of these methods perform well on regression type of problems GBM and neural network are non-parametric techniques that learn well in our problem. We can see from the results later that train and test models have similar performance, indicating no over-fitting, which is often a concern in non-parametric models, thus proving to be candidate models.**

# Supermarket Sales Prediction

### Supermarket Sales Prediction

#### Wecsley O. Prates

#### December, 11, 2020

# 1 Data Exploration & Cleaning

The data contains historical sales data for 45 stores located in different regions. Each store contains a number of departments, and you are tasked with predicting the department-wide sales for each store. The data is stored in 3 different csv files.

*stores.csv* This file contains anonymized information about the 45 stores, indicating the type and size of store.

*sales.csv* This is the historical sales data, which covers to 2010-02-05 to 2013-07-26. Within this file you will find the following fields:

* Store - the store number
* Dept - the department number
* Date - the week
* Weekly\_Sales - sales for the given department in the given store
* IsHoliday - whether the week is a special holiday week

*Features.csv* This file contains additional data related to the store, department, and regional activity for the given dates. It contains the following fields:

* Store - the store number
* Date - the week
* Temperature - average temperature in the region
* Fuel\_Price - cost of fuel in the region
* MarkDown1-5 - anonymized data related to promotional markdowns that Walmart is running. MarkDown data is only available after Nov 2011, and is not available for all stores all the time. Any missing value is marked with an NA.
* CPI - the consumer price index
* Unemployment - the unemployment rate
* IsHoliday - whether the week is a special holiday week
* [1.1 Read the Data Sets](https://rstudio-pubs-static.s3.amazonaws.com/703289_d02922583c14484ea3494646894f3bd4.html#read-the-data-sets)
* [1.2 Data Exploration & Manipulation](https://rstudio-pubs-static.s3.amazonaws.com/703289_d02922583c14484ea3494646894f3bd4.html#data-exploration-manipulation)

| **Store** | **Date** | | **Temperature** | | **Fuel\_Price** | **MarkDown1** | | **MarkDown2** | **MarkDown3** | | **MarkDown4** | **MarkDown5** | | **CPI** | **Unemployment** | **IsHoliday** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 05/02/2010 | | 42.31 | | 2.572 | NA | | NA | NA | | NA | NA | | 211.0964 | 8.106 | FALSE |
| 1 | 12/02/2010 | | 38.51 | | 2.548 | NA | | NA | NA | | NA | NA | | 211.2422 | 8.106 | TRUE |
| 1 | 19/02/2010 | | 39.93 | | 2.514 | NA | | NA | NA | | NA | NA | | 211.2891 | 8.106 | FALSE |
| 1 | 26/02/2010 | | 46.63 | | 2.561 | NA | | NA | NA | | NA | NA | | 211.3196 | 8.106 | FALSE |
| 1 | 05/03/2010 | | 46.50 | | 2.625 | NA | | NA | NA | | NA | NA | | 211.3501 | 8.106 | FALSE |
| 1 | 12/03/2010 | | 57.79 | | 2.667 | NA | | NA | NA | | NA | NA | | 211.3806 | 8.106 | FALSE |
| **Store** | | **Dept** | | **Date** | | | **Weekly\_Sales** | | | **IsHoliday** | | |
| 1 | | 1 | | 05/02/2010 | | | 24924.50 | | | FALSE | | |
| 1 | | 1 | | 12/02/2010 | | | 46039.49 | | | TRUE | | |
| 1 | | 1 | | 19/02/2010 | | | 41595.55 | | | FALSE | | |
| 1 | | 1 | | 26/02/2010 | | | 19403.54 | | | FALSE | | |
| 1 | | 1 | | 05/03/2010 | | | 21827.90 | | | FALSE | | |
| 1 | | 1 | | 12/03/2010 | | | 21043.39 | | | FALSE | | |

| **Store** | **Type** | **Size** |
| --- | --- | --- |
| 1 | A | 151315 |
| 2 | A | 202307 |
| 3 | B | 37392 |
| 4 | A | 205863 |
| 5 | BTRAAC.png | 34875 |
| 6 | A | 202505 |

# Machine Learning

Let´s now start the Machine Learning development.

Original Training data set is split into two random training and testing data sets.

* [3.1 Models Development](https://rstudio-pubs-static.s3.amazonaws.com/703289_d02922583c14484ea3494646894f3bd4.html#models-development)
* [3.2 Residual Analysis](https://rstudio-pubs-static.s3.amazonaws.com/703289_d02922583c14484ea3494646894f3bd4.html#residual-analysis)

##

## Call:

## lm(formula = Weekly\_Sales ~ Size + StoreGroup + IsHoliday, data = trainM)

##

## Residuals:

## Min 1Q Median 3Q Max

## -676576 -131333 -12932 105772 2173374

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 9.728e+05 1.649e+04 58.98 < 2e-16 \*\*\*

## Size 4.766e+00 7.632e-02 62.44 < 2e-16 \*\*\*

## StoreGroupLow -8.222e+05 1.693e+04 -48.58 < 2e-16 \*\*\*

## StoreGroupMedium -6.647e+05 9.978e+03 -66.62 < 2e-16 \*\*\*

## IsHolidayTRUE 6.813e+04 1.349e+04 5.05 4.6e-07 \*\*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 235900 on 4499 degrees of freedom

## Multiple R-squared: 0.8288, Adjusted R-squared: 0.8287

## F-statistic: 5445 on 4 and 4499 DF, p-value: < 2.2e-16

We can expect Holidays to positively impact Sales in general but as can be seem from above plot, we cannot strongly state that all Holiday Weeks result in higher Sales. Thanksgiving week seems to have jump in sales by Xmas week on the other hand shows a drop

We use Forward selection mechanism to pin point the most significant variables

Repeated with Backward Selection. In both cases, we ended up with similar models. Hence proceeding with this Model.

Using leaps package, we can check for different combinations of the independent variables and select the best combination on the basis or R-sq / Adjusted R-sq. We are using Adjusted R-sq here.

##

## Call:

## lm(formula = Weekly\_Sales ~ Type + Temperature + MarkDown3 +

## MarkDown5 + CPI + Unemployment + Size + WeekType + StoreGroup,

## data = trainM)

##

## Residuals:

## Min 1Q Median 3Q Max

## -696154 -123359 -1768 111566 1784907

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 2.578e+05 2.900e+04 8.892 < 2e-16 \*\*\*

## Type1 4.242e+04 1.083e+04 3.916 9.14e-05 \*\*\*

## Type2 3.658e+04 1.686e+04 2.169 0.030122 \*

## Temperature 7.270e+02 1.870e+02 3.887 0.000103 \*\*\*

## MarkDown3 4.027e+00 6.398e-01 6.295 3.38e-10 \*\*\*

## MarkDown5 5.878e+00 8.070e-01 7.285 3.79e-13 \*\*\*

## CPI -1.294e+03 8.549e+01 -15.135 < 2e-16 \*\*\*

## Unemployment -5.231e+03 1.847e+03 -2.832 0.004645 \*\*

## Size 5.072e+00 1.187e-01 42.742 < 2e-16 \*\*\*

## WeekType1 6.507e+04 7.266e+03 8.955 < 2e-16 \*\*\*

## WeekType2 3.997e+05 1.398e+04 28.597 < 2e-16 \*\*\*

## StoreGroup1 1.312e+05 1.338e+04 9.806 < 2e-16 \*\*\*

## StoreGroup2 7.924e+05 1.676e+04 47.278 < 2e-16 \*\*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 207500 on 4491 degrees of freedom

## Multiple R-squared: 0.8678, Adjusted R-squared: 0.8674

## F-statistic: 2456 on 12 and 4491 DF, p-value: < 2.2e-16

We had noticed that Size of the Store is high on large & Small sizes and few in between which is probably adding to the positive skewness. Hence we will further enhance the current Model with a log transformation of Size.

Polynomial Regression

##

## Call:

## lm(formula = Weekly\_Sales ~ Type + Temperature + MarkDown3 +

## MarkDown5 + CPI + Unemployment + I(log(Size)) + WeekType +

## StoreGroup, data = trainM)

##

## Residuals:

## Min 1Q Median 3Q Max

## -699985 -148795 3265 131375 1759117

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) -4.631e+06 1.512e+05 -30.633 < 2e-16 \*\*\*

## Type1 -6.443e+04 1.002e+04 -6.430 1.41e-10 \*\*\*

## Type2 1.676e+04 1.861e+04 0.901 0.3677

## Temperature 7.732e+02 1.952e+02 3.960 7.61e-05 \*\*\*

## MarkDown3 4.055e+00 6.679e-01 6.071 1.37e-09 \*\*\*

## MarkDown5 6.299e+00 8.422e-01 7.479 8.93e-14 \*\*\*

## CPI -1.049e+03 9.024e+01 -11.629 < 2e-16 \*\*\*

## Unemployment -4.608e+03 1.929e+03 -2.389 0.0169 \*

## I(log(Size)) 4.811e+05 1.331e+04 36.152 < 2e-16 \*\*\*

## WeekType1 6.335e+04 7.585e+03 8.353 < 2e-16 \*\*\*

## WeekType2 3.959e+05 1.459e+04 27.142 < 2e-16 \*\*\*

## StoreGroup1 7.428e+04 1.608e+04 4.619 3.97e-06 \*\*\*

## StoreGroup2 7.689e+05 1.921e+04 40.026 < 2e-16 \*\*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 216600 on 4491 degrees of freedom

## Multiple R-squared: 0.8559, Adjusted R-squared: 0.8555

## F-statistic: 2223 on 12 and 4491 DF, p-value: < 2.2e-16

# 4 Finalizing Model

Initially we tried Square Root transformation but not a major improvement in skewness.

Using Boxcox, we see lambda close to 0 and hence using log transformation.

##

## Call:

## lm(formula = log(Weekly\_Sales) ~ Size + Temperature + MarkDown3 +

## MarkDown5 + CPI + WeekType + StoreGroup, data = trainM)

##

## Residuals:

## Min 1Q Median 3Q Max

## -0.62833 -0.12823 -0.00425 0.12403 0.97004

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 1.261e+01 1.897e-02 664.939 < 2e-16 \*\*\*

## Size 5.329e-06 6.638e-08 80.271 < 2e-16 \*\*\*

## Temperature 8.137e-04 1.795e-04 4.532 5.98e-06 \*\*\*

## MarkDown3 2.313e-06 6.247e-07 3.703 0.000216 \*\*\*

## MarkDown5 5.363e-06 7.844e-07 6.837 9.16e-12 \*\*\*

## CPI -1.251e-03 7.840e-05 -15.957 < 2e-16 \*\*\*

## WeekType1 4.745e-02 6.953e-03 6.825 9.99e-12 \*\*\*

## WeekType2 3.033e-01 1.347e-02 22.514 < 2e-16 \*\*\*

## StoreGroup1 4.708e-01 1.099e-02 42.822 < 2e-16 \*\*\*

## StoreGroup2 9.111e-01 1.457e-02 62.541 < 2e-16 \*\*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 0.2027 on 4494 degrees of freedom

## Multiple R-squared: 0.8834, Adjusted R-squared: 0.8831

## F-statistic: 3782 on 9 and 4494 DF, p-value: < 2.2e-16

With this, we have build our final Model: Model with an adjusted R-sq = 88.24 and all variables are highly significant.

# 5 Prediction Using the Fit Model

# 6 Evaluation of Model by Test Data

Doing the prediction using the Test Data set, We got very good predictions and pretty closer with the real values.

For whole data set.

# 7 Forecasting

Other Models

Using now the ts data frame.

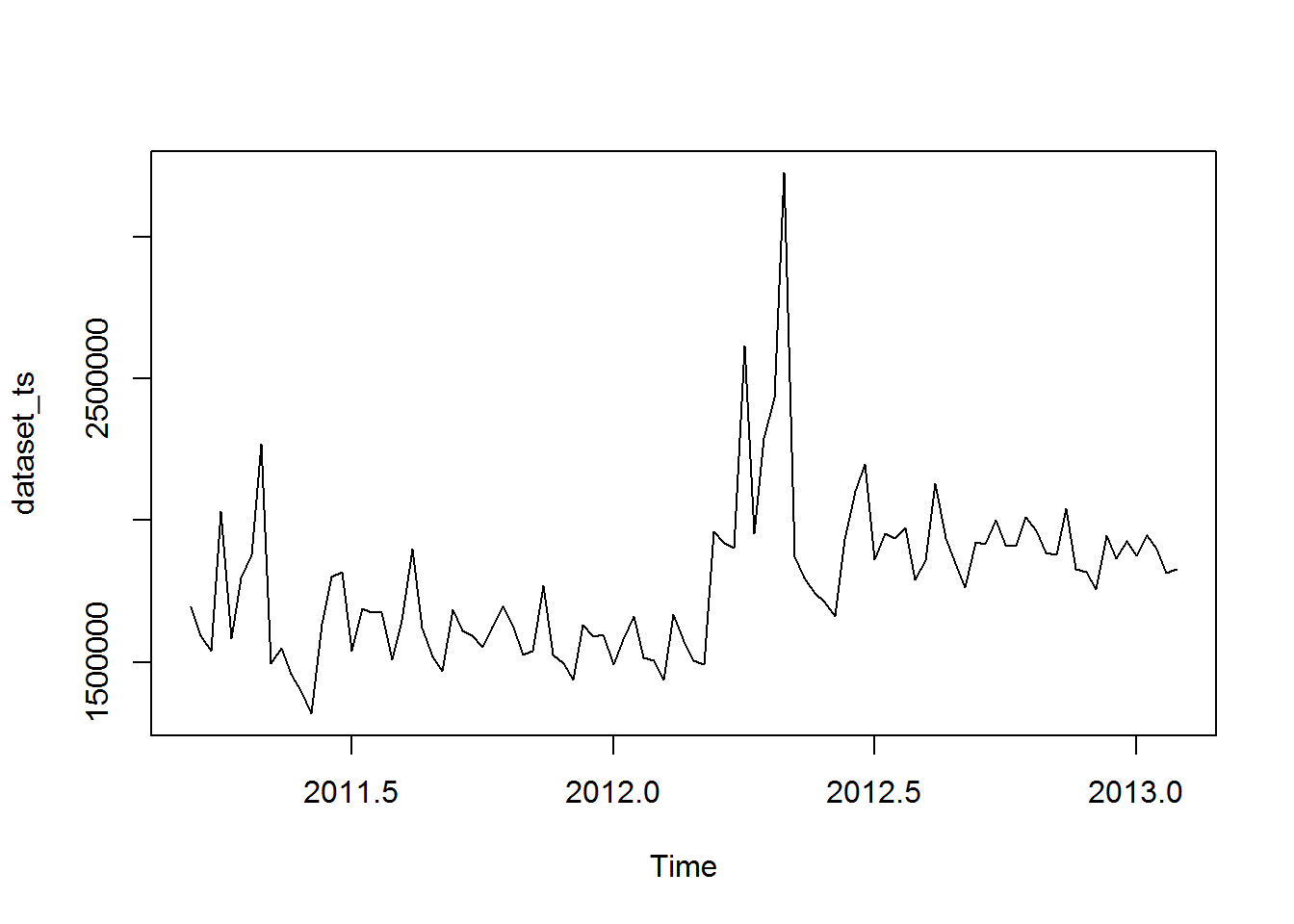
# 8 Conslusion

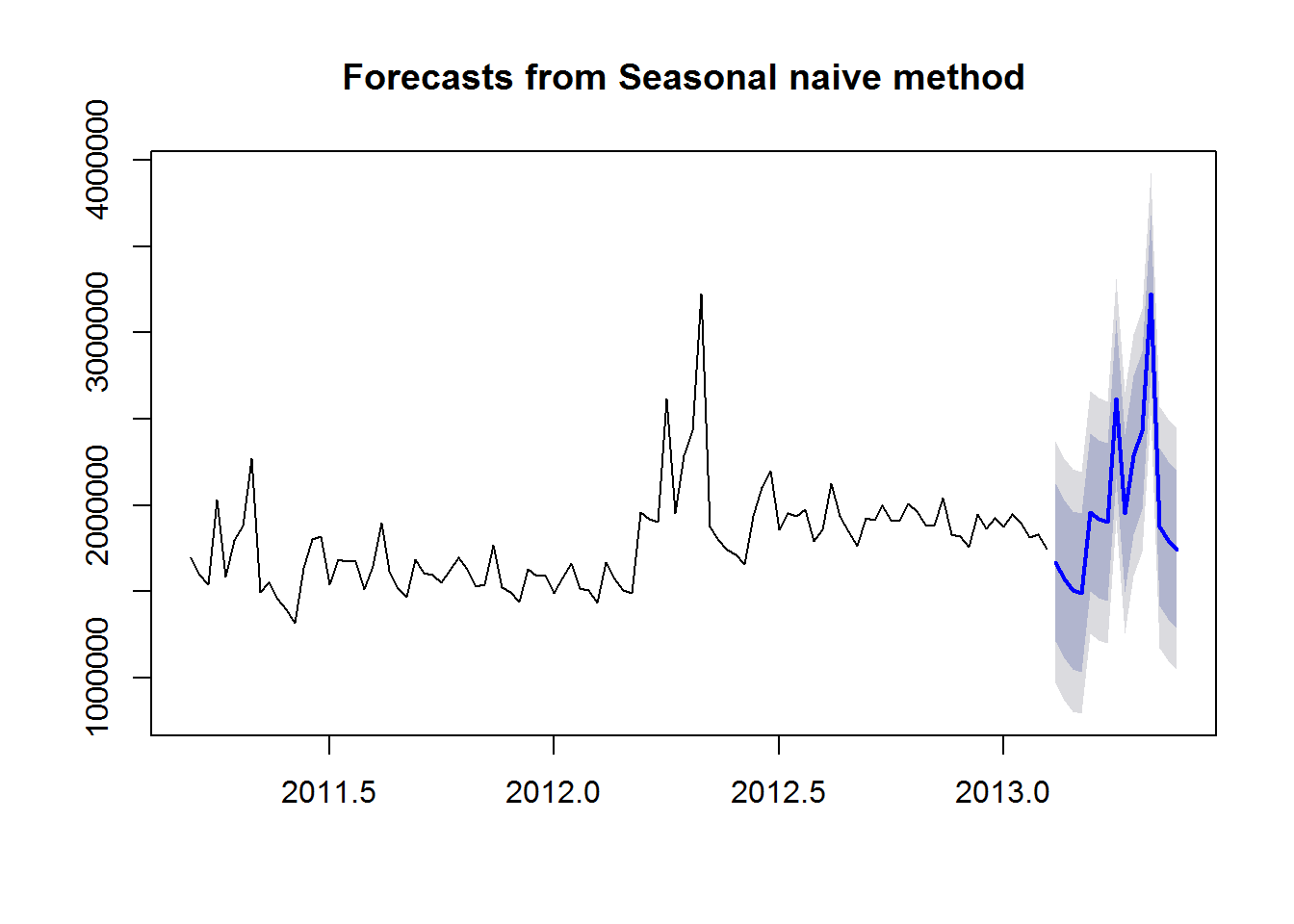
We were able to build this model with Adjusted R-sq = 88% and Predicted R-Sq = 88%. The variables “Store Group” & “Week Type” constructed from the data set are the most significant variables. With this Model, we can predict a good estimate of the weekly Sales for Stores.

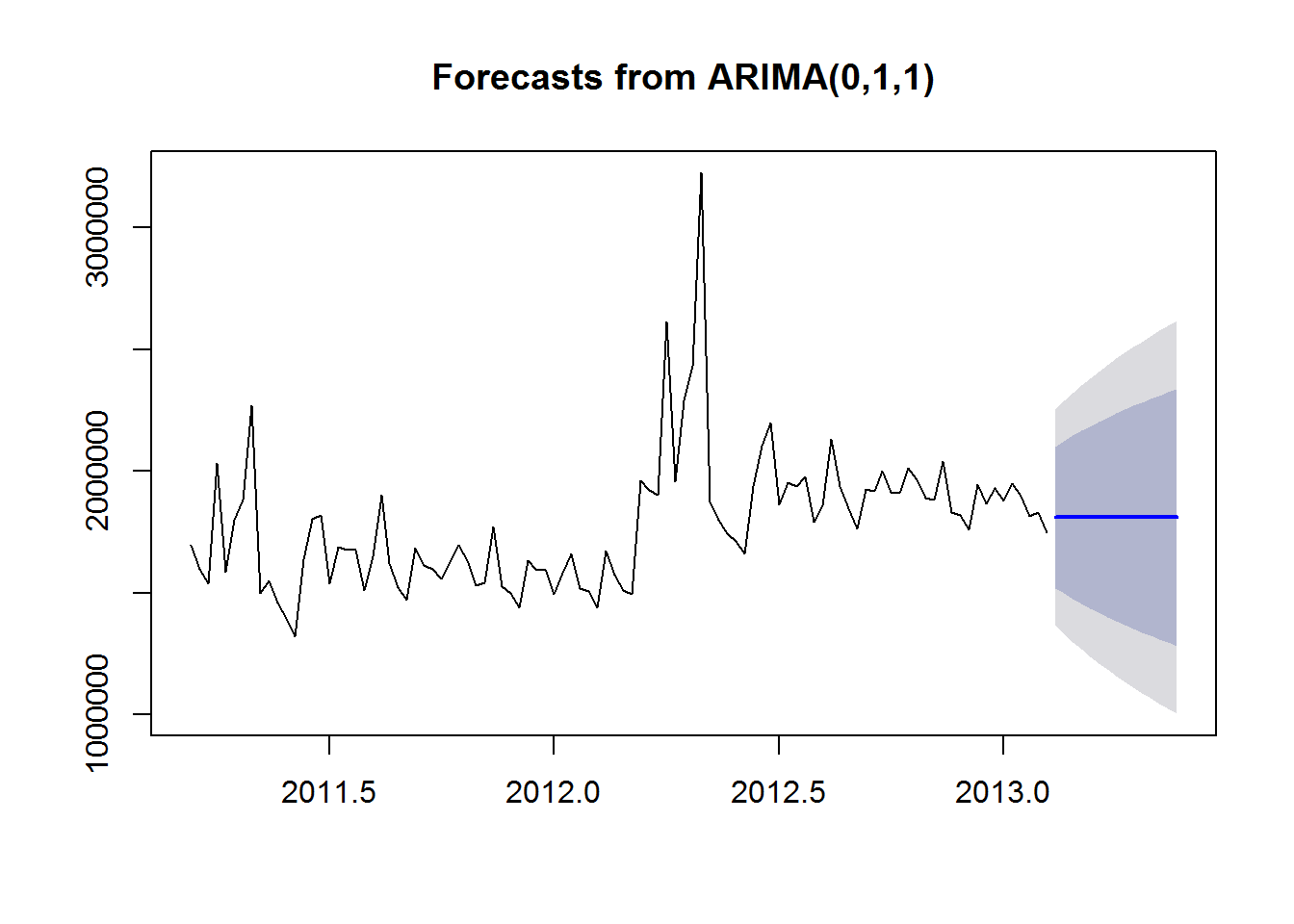
# 9 Extra: Pivot Table

With this Pivot Table We can manipulate and see the distribution for whatever variables combination. It´s totally interactive Pivot Table. We can drag the variables to the columns or rows, choose the categories, make filters, choose which metrics or graphs.

|  |  |  |
| --- | --- | --- |
|  | ↕↔ |  Year ▾ |
|  Store ▾   Type ▾   Week ▾   Month ▾   Day ▾   IsHoliday ▾   Weekly\_Sales ▾   Temperature ▾   Fuel\_Price ▾   MarkDown1 ▾   MarkDown2 ▾   MarkDown3 ▾   MarkDown4 ▾   MarkDown5 ▾   CPI ▾   Unemployment ▾   Size ▾   StoreGroup ▾   sales\_class ▾ |  WeekType ▾ | |  | **Year** | **2010** | **2011** | **2012** | **Totals** | | --- | --- | --- | --- | --- | --- | | **WeekType** |  | | **High** | | 215 | 215 | 18 | **448** | | **Low** | | 511 | 673 | 667 | **1,851** | | **Medium** | | 1,434 | 1,452 | 1,250 | **4,136** | | **Totals** | | **2,160** | **2,340** | **1,935** | **6,435** | |







# Conslusion

We were able to build this model with Adjusted R-sq = 88% and Predicted R-Sq = 88%. The variables “Store Group” & “Week Type” constructed from the data set are the most significant variables. With this Model, we can predict a good estimate of the weekly Sales for Stores.

https://github.com/8765197073/forecast.git