SALES HISTORY-BASED DEMAND PREDICTION BY USING GENERALIZED LINEAR MODELS

A Thesis Submitted to
The Graduate School of Engineering and Sciences of
İzmir Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of

MASTER OF SCIENCE

in Computer Engineering

by
Başar ÖZENBOY

July 2016
İZMİR

We approve the thesis of **Başar ÖZENBOY**

Examining Committee Members:

Assist. Prof. Dr. Derya BİRANT
Department of Computer Engineering, Dokuz Eylul University

Assist. Prof. Dr. Tolga AYAV
Department of Computer Engineering, Izmir Institute of Technology

Assist. Prof. Dr. Selma TEKİR
Department of Computer Engineering, Izmir Institute of Technology

21 July 2016

Assist. Prof. Dr. Selma TEKİR
Supervisor, Department of Computer Engineering,
Izmir Institute of Technology

Assoc. Prof. Dr. Y. Murat Erten
Head of Department of Computer
Engineering

Prof. Dr. Bilge KARAÇALI
Dean of Graduate School of
Engineering and Sciences

ACKNOWLEDGMENTS

I would like to express sincere appreciation to my supervisor Assist. Prof Dr. Selma TEKİR. I have learned a lot with her knowledge and enjoyed while working on this thesis.

I have been supported by TUBITAK “2210-C Domestic Graduate Scholarship Programme for Priority Areas”. I am also thankful to TUBITAK for financial support during my thesis period.

I would like to express my thanks to my employer Logo Yazılım Inc. and my directors for their support and encouragement during my thesis period and during lecture period.

I would also like to express my profound gratitude to my family, my wife Selen Okuroğlu Özenboy, for their support, understanding and patient during my studies. Their encouragement is invaluable for me.

ABSTRACT

SALES HISTORY-BASED DEMAND PREDICTION BY USING GENERALIZED LINEAR MODELS

Improved data collection and storage capabilities make vast amounts of data available in appropriate formats. Commercial enterprises store their sales data. It’s vital for companies to accurately predict demand by utilizing the existing sales data. Such predictive analytics is a crucial part of their decision support systems to increase the profitability of the company.

In predictive data analytics, the branch of regression modeling commonly is used to predict a numerical response variable like sales amount. In recent years, generalized linear models provide a generalization to better address the specificities of the problem at hand. To begin with, they relax the assumption of normally distributed error terms. Moreover, the relationship of the set of predictor variables and the response variable could be represented by a set of link functions rather than the sole choice of the identity function.

This thesis models the sales amount prediction problem through the use of generalized linear models. Unique company sales data are explored and fitted accordingly with the right distribution function of the response variable along with an appropriate link function. The experimental results are compared with the other regression models, classification algorithms, and time series models. The model selection is performed via the use of MSE and AIC metrics respectively.

ÖZET

GENELLEŞTİRİLMİŞ DOĞRUSAL MODELLER KULLANARAK SATIŞ GEÇMİŞİ TABANLI TALEP TAHMİNLEMESİ

Gelişmiş veri toplama ve depolama yetenekleri çok büyük miktarlardaki verileri uygun formatlarda erişilebilir hale getirmektedir. Birçok ticari firma kurumsal verilerini dijital ortamda saklayabilmektedir. Bu durumda, tahminleme analitiği firmaların karlılıklarını yükseltmek için karar destek sistemlerinin önemli bir parçası haline gelmiştir.

Tahminleme analitiğinde, regresyon modelleme dalı genellikle satış miktarı gibi nümerik yanıt değişkeninin tahminlemesinde kullanılır. Son yıllarda, genelleştirilmiş doğrusal modeller ele alınan problemleri daha iyi adresleyen bir genelleştirme sağlamak için kullanılmaya başlanmıştır. İlk olarak, modellerdeki hata terimlerinin normal dağıldığı varsayımından vazgeçilmiş, daha sonra, tahmin değişkenleri ile yanıt değişkeni arasındaki ilişki tek bir birim fonksiyonu yerine bağ fonksiyonları ile ifade edilmiştir.

Bu tezin kapsamında genelleştirilmiş doğrusal modeller kullanılarak satış miktarı tahminleme probleminin modellemesi çalışması yapılmıştır. Bir firmaya ait satış verileri keşifçi veri analizi teknikleri ile incelenmiştir. Yanıt değişkeninin uyum gösterdiği olasılık dağılımına göre uygun bir bağ fonksiyonu kullanılmıştır. Deneysel sonuçlar diğer regresyon modelleri, sınıflandırma algoritmaları ve zaman serileri modelleri ile karşılaştırılmıştır. Model seçimi Akaike ölçütü (AIC) ve ortalama hata kareleri (MSE) metrikleri kullanılarak uygulanmıştır.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Demand prediction is a vital study for commercial companies. Companies should use effective prediction techniques. The uncertainty in future makes the prediction hard. All the existing prediction techniques use past records but produced results vary.

As mentioned before, one of the important demand prediction element is past records. Therefore, companies should save their records in a data warehouse. The growth of these records brings about the need for computerized decision support. Determining sales policy according to predicted future demands could be one component of decision support systems.

Qualitative techniques and quantitative techniques are methods that are used to determine sales policy for decision support systems. Despite qualitative techniques are based on subjective methods because of lack of data availability, quantitative techniques are based on numerical data analysis with use of different statistical techniques. By the way, many of predicting techniques for example quantitative techniques, can be implemented as a part of decision support systems.

Linear regression is one of the statistical techniques that is commonly used for prediction for quantitative data. Linear regression models relate response (dependent) and predictor (independent) variables. For example, product sales can be predicted by referring a relationship between a response variable that is sales amount and predictor variables that are air temperature, product price etc.

Linear regression models have some assumptions to simplify the theory of analysis and real world situations. One of the assumptions is regarding error terms (ε). Linear regression models assume that the errors terms are normally distributed. The second assumption is such that response variables are independent normal random variables with the mean is the expected value of response variable and variance is σ^2 (Kutner, 2004).

Figure 1.1 visualizes the linear regression model with the stated assumptions. $E\{Y\} = aX + b$ implies linear regression model with parameters a, b and predictor variable

X. E{Y_i} implies the expected value of Y_i on the regression line, ε_i implies the error term with normal distribution. Y_i represents the real-valued response variable.

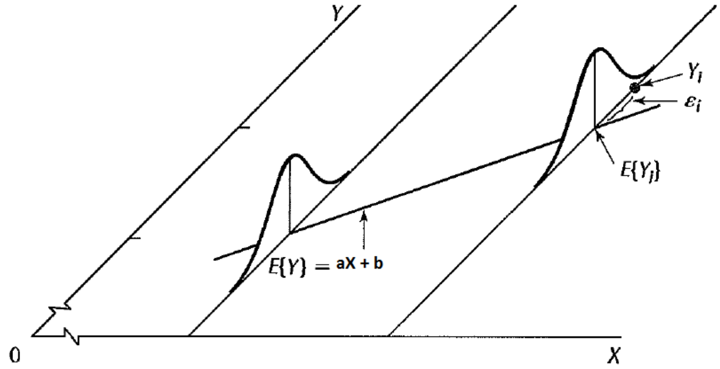


Figure 1.1. Linear regression model
Source: (Kutner, 2004).

In some real world applications, error terms and the expected value of response variables may not have normal distribution. In that case generalized linear models (GLM) can be used instead of linear regression models. GLM can be used for predicting the expected value of response variable which has a distribution from the exponential family and the individual values of the response variable are independent from each other. Link function is one of the GLM property which connects the parameters of the response variable distribution with the linear model (Nelder, 1972). So, if there exists an appropriate link function for fitting GLM then, the goodness of fit of GLM may produce better result than linear regression models.

In this study, we aim to predict the product sales for a company using corporate past sales data. GLM is used as the prediction technique.

The followed methodology can be described through its phases: Problem definition and exploratory data analysis, statistical evaluation, model fitting and the proposed contribution to existing literature.

Chapter 1 presents the purpose of this thesis and problem definition issues. Chapter 2 discusses methodological issues. Predictive methods which includes GLM, linear model, time series and classification techniques are presented. Model selection and model validation issues are discussed. Chapter 3 surveys related studies. Quantitative method which is one of the predictive method, are surveyed. Then classification and regression modelling studies are researched. In Chapter 4, data analysis issues are experienced with some useful research techniques. Exploratory data analysis are done for investigating the distribution of response variable. Data sets, response and predictor variables are introduced. Summary statistics are created and visualized by use of past years data sets. Model fitting and hypotheses testing issues are presented. GLM, linear model and predictive data mining methods are implemented. Then, coefficients of independent variables, which affects the response variable, are interpreted. Then, quality of fitted GLM is compared with relative the other models. At the end, thesis concludes with Chapter 5 which discusses results of this study.

CHAPTER 2

BACKGROUND

This chapter presents predictive methods which are included regression analysis techniques and classification techniques. Also at the end of this chapter, model selection and model validation issues are discussed.

2.1. Predictive Methods

Predictive methods are divided into two parts. The first one is qualitative methods which do not include measurement or statistical work, and the second one is quantitative methods which include statistical and mathematical techniques for empirical evaluation. Quantitative methods also are divided into two parts which are regression and classification. Regression analysis techniques is presented in the following section.

2.1.1. Regression Analysis Techniques

Regression analysis is a statistical technique to predict a response variable with one or more independent variables. This statistical technique is commonly used in commercial companies, natural and applied science. For example, a company’s product sales can be predicted with regression modeling by using predictors like air temperature, holidays, sales prices etc. (Kutner, 2004).

A regression model has two essentials with statistical relations. First, the response variable changes with the changes of predictor variables in a systematic manner. Second, when the independent variables are visualized in a scatter plot, points around the curve which shows independent variable, have statistical relationship. These essentials bring together two regression model concepts. First of these concepts, for each level of a predictor variable, there is a probability distribution for the response variable. Second the means of the probability distribution changes with the value of predictor variables (Kutner, 2004). Figure 2.1 visualizes these regression model concepts. The predictor variable X is located on regression curve and at that point, the response variable Y has probability distribution depends on the predictor variable value.

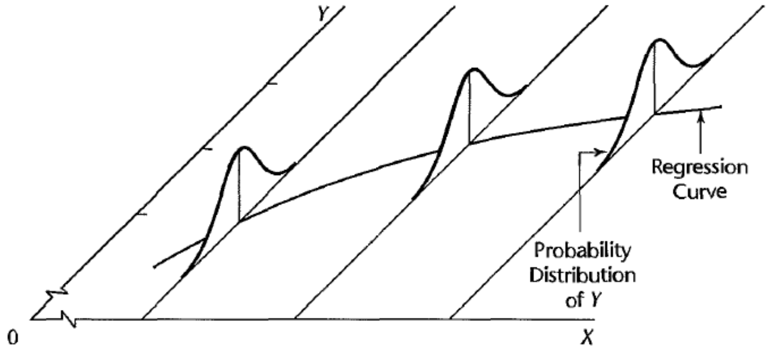


Figure 2.1. The probability distribution of the response variable with regression curve (Source: Kutner, 2004).

The next section presents linear models which are the special form of the regression models.

2.1.1. Linear Models

Linear models can be applied with one or more predictor variables to predict the response variable. The following formulation shows the regression model with one predictor variable:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{2.1}$$

Y_i shows the value of the response variable in the *i*th trial. β_0 and β_1 are the parameters of the linear regression model. X_i is the predictor variable in the *i*th trial. ε_i is a random error term.

Linear models have some assumptions to simplify the modeling problems. There exist some important assumptions about ε_i error term.

$$\varepsilon_i = Y_i - E[\, Y_i \,] \tag{2.2}$$

First, error terms are normally distributed and the expected value of them is zero, the mean is zero and the variance is σ^2. These assumptions simplify the theory of regression analysis and implementations of the real world applications.

$$E[\,\varepsilon_i\,] \;=\; 0 \qquad (2.3)$$

Second, error terms are uncorrelated so their covariance is zero. That means, the result of one trial does not affect another trial.

$$\sigma^2\{\,\varepsilon_i\,\} = \;\sigma^2 \; and \;\; \sigma\big\{\,\varepsilon_i, \varepsilon_j\,\big\} = \;0 \text{ for all i, j; i} = 1 = \text{j} \qquad (2.4)$$

Because the expected value of the error term is zero, the expected value of the response variable equations has the following formulation.

$$E[\,Y_i\,] = \beta_0 \;+\; \beta_1 X_i \qquad (2.5)$$

In summary, the response variable comes from the probability distribution with mean E{Y$_i$} and whose variance is σ^2 . And the response variables Y$_i$ are independent form each other (Kutner, 2004).

The assumptions of normal distribution for response variable is not acceptable for some real world problems. For example; error counts, the number of people that have illness, the amount of product sales are not distributed normally in the real world cases. In those cases, the model of the response variable which does not have a normal distribution, can be transformed with some techniques to acquire the normal distribution characteristics and the constant variance. Figure 2.2 illustrates inconstant variance on the X-Y surface. Especially, the logarithm transformation, the square root transformation, and the power transformation are common methods for transforming the data. However, there is no specific solution for transforming data to provide constant variance, normal distribution and simple model form.

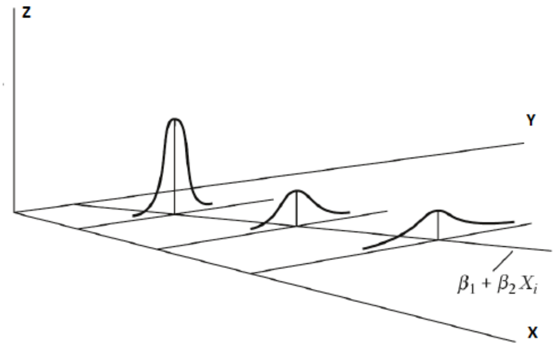


Figure 2.2. Inconstant variance

In Generalized linear models the response variables' distribution comes from the exponential family. Thus, first, exponential family is presented. After that, generalized linear models method for regression models which do not have to have normally distributed response variable are presented.

2.1.1.2. Exponential Family

Y is a single random variable and that variable's probability distribution is related with a single parameter (θ). The distribution of that variable is a member of the exponential family when it's defined in the form of the following formula:

$$f(y;\ \theta) = s(y)t(\theta)e^{a(y)b(\theta)} \qquad (2.6)$$

In the formula a, b, s, and t are known functions. There is a symmetry between y and θ. If the previous formulation is reformed, the following formulation is derived.

$$f(y;\ \theta) = \exp[a(y)b(\theta) + c(\theta) + d(y)]\, where \qquad (2.7)$$

$$s(y) = exp\ d(y) \qquad (2.8)$$

$$t(\theta) = exp\ c(\theta) \qquad (2.9)$$

The distribution is in the canonical form when $a(y) = y$ and $b(\theta)$ is the natural parameter of the distribution. In addition to θ parameter of formula, if there exist other parameters, those parameters are taken into account as nuisance parameters forming parts of the functions a, b, c and d (Dobson, 2008).

Many of the most common distributions which are Poisson, Bernoulli, gamma, normal, exponential, are included in the exponential family. The common property of these distributions is that they have certain parameters which are fixed and known. For example, binomial distribution has a parameter which is known and fixed over the number with the trials or negative binomial distribution has a parameter which is known and fixed number of failures etc. (Wikipedia - Exponential family, 2016)

2.1.1.3. Generalized Linear Models (GLM)

Generalized Linear models (GLM) is firstly presented by Nelder and Wedderburn in the year 1972. GLM is an alternative method of the other data transformation methods. That transformation has been done for a response variable that is not in the form of normal distribution.

GLM is a family of models which are regression models. That family includes normal error linear regression models, exponential, logistic, Poisson, log-linear, gamma, Gaussian and many other models (Kutner, 2004).

GLM has some essentials defined as below.

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{2.10}$$

I. Y_i's are independent from each other and these response variables come from the probability distribution with mean E[Y_i] = μ_i and that probability distribution is a member of the *exponential family*.

II. X_i's are predictor variables and X_i β is a *linear predictor* which is based on the predictor variables.

$$\eta = X_i' \beta = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{2.11}$$

III. g is the *link function* which relates the linear predictor to the mean of the response variables (Kutner,2004).

$$\eta = X_i' \beta = g(\mu_i) \tag{2.12}$$

The main difference between the linear model and the GLM is in the set-up of the estimated model function. It is important to specify that GLM transforms the expected value which is the mean of the response variables by the help of the link function, on the other hand, the linear model transforms itself. So the following linear model function estimates the expected value of g(y$_i$).

$$E[\,g(y_i)\,] = \beta_0 + \beta_1 X_i \tag{2.13}$$

And the following GLM function estimates g function value of the expected y$_i$ value (Balajir, 2011).

$$g(\,E[\,y_i\,]\,) = \beta_0 + \beta_1 X_i \tag{2.14}$$

The following table shows the common distributions with their link functions.

Table 2.1. Common distribution with link functions
(Source: De Jong, 2008)

Distribution	Link Name	Canonical Link Function g(μ$_i$)
Normal	Identity	$\eta = X\beta = \mu$
Exponential, Gamma	Inverse	$\eta = X\beta = \frac{1}{\mu}$
Inverse Gaussian	Log	$\eta = X\beta = \ln(\mu)$
Poisson	Log	$\eta = X\beta = \ln(\mu)$
Binomial, Bernoulli, Categorical, Multinomial	Logit	$\eta = X\beta = \ln(\frac{\mu}{1-\mu})$

To explain the concept of the link function some known distribution functions and link function relations are examined in the next sections.

2.1.1.3.1. Normal Distribution

First, the following formulation presents the probability density function of the normal distribution and the link relation with the mean value.

$$f(\mathrm{y}_i) \ = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i-\mu)^2}{2\sigma^2}} \qquad (2.15)$$

From the equation above, μ is the mean, σ is the standard deviation and σ^2 is the variance obtained. The following formulation presents the *identity link function.* That link function is appropriate for using the normally distributed response variable.

$$X\beta = \mu = E(\mathrm{y}_i) \qquad (2.16)$$

If the link function of GLM is the identity link function then, the following equation is derived which is the same with that of the linear model.

$$g\big(E(y_i)\big) = \mathrm{g}(\mu_i) = X_i\beta = \mu_i = \beta_0 + \ \beta_1\mathrm{X}_i \qquad (2.17)$$

2.1.1.3.2. Exponential Distribution

The following example shows the exponential distribution, and its link function. The following formulation shows the probability density function of exponential distribution (Myers, 2010).

$$f(y) = \frac{1}{\lambda} e^{-\frac{y}{\lambda}}, y \ \geq 0, \lambda \geq \ 0 \qquad (2.18)$$

If the previous function is written in the form of exponential family then the following function is derived:

$$f(y) = \exp\left\{[-1][y(\frac{1}{\lambda}) + ln\lambda]\right\} \qquad (2.19)$$

In that formula, $a(\Phi) = -1, \theta = \frac{1}{\lambda}, b(\theta) = ln\ \theta, c(.) = 0, \mu = \lambda\ and\ \sigma^2 = \lambda^2$ Φ is the dispersion parameter for the exponential family distribution.

The canonical link function shown in the following formula:

$$\frac{1}{\lambda} = X\beta \tag{2.20}$$

$$\mu = \lambda \tag{2.21}$$

The following formula is named as the *inverse link function*. That link function is appropriate for using exponential distributed response variable.

$$X\beta = \frac{1}{\mu} \tag{2.22}$$

$$\mu = \frac{1}{X\beta} \tag{2.23}$$

If the link function of GLM is the inverse link function then, the following equation is derived:

$$g\big(E(y_i)\big) = \mathrm{g}(\mu_i) = X_i\beta = \frac{1}{\mu_i} = \beta_0 + \beta_1 \mathrm{X}_i \tag{2.24}$$

2.1.1.3.3. Gamma Distribution

The following formulation shows the probability density function of the gamma distribution that is one of a special form of the exponential distribution (Balajir, 2011).

$$Y \sim gamma\ (\alpha, \beta) \tag{2.25}$$

$$f(y_i, \theta, \Phi) = \frac{1}{\beta_i^{\alpha_i}\Gamma(\alpha_i)} {y_i}^{(\alpha_i - 1)} e^{-(y_i/\beta_i)} \quad y_i, \alpha_i, \beta_i > 0 \tag{2.26}$$

$$\theta = \frac{-1}{\alpha \beta} \tag{2.27}$$

$$\Phi = \frac{1}{\alpha} \tag{2.28}$$

$$b(\theta) = -\log(-\theta) \tag{2.29}$$

$$a(\Phi) = \Phi \; and \; c(\Phi) = \alpha \log(\alpha) - \log\big(\Gamma(\alpha)\big) + (\alpha - 1)\log(y) \tag{2.30}$$

The canonical link function is shown in the following formula which is named as the *inverse link function* and which is appropriate for using gamma distributed response variable:

$$X\beta = \frac{1}{\mu} \tag{2.31}$$

β_i is the scale parameter and α_i is the shape parameter of the gamma distribution. Shape and scale parameters are positive real numbers. The expected value and variance equations are presented below. Generally, α_i values are constant with α.

$$E(y_i) = \mu_i = \alpha_i \beta_i \tag{2.32}$$

$$Var(y_i) = \alpha_i \beta_i^2 \tag{2.33}$$

The link function of the gamma distribution was represented as $\frac{1}{\mu}$ in the previous descriptions. As a result, the relation of the link function and the parameters of the gamma distribution can be seen from the following equation.

$$\frac{1}{\mu} = X\beta = g(\mu) = \frac{1}{\alpha \beta_i} \tag{2.34}$$

The shape parameter (α) of the gamma distribution is just a constant multiplier and it is equal to the inverse of the dispersion parameter (Φ). Following equation represents the dispersion parameter equation with shape parameter.

$$\Phi = \frac{1}{\alpha} \tag{2.35}$$

The relation between the mean and the variance of the gamma distribution is presented below.

$$\frac{var(y_i)}{E(y_i)} = \frac{\alpha_i \beta_i^2}{\alpha_i \beta_i} = \beta_i \tag{2.36}$$

The equation above presents the ratio of the mean to the variance, which is a constant and it is the scale parameter.

The relation between the mean and the standard deviation of the gamma distribution is presented by the help of the coefficient of variation equation:

$$CV = \frac{\sqrt{var(y_i)}}{E(y_i)} = \frac{\sqrt{\alpha_i \beta_i^2}}{\alpha_i \beta_i} = \frac{1}{\sqrt{\alpha_i}} \tag{2.37}$$

The equation above presents the ratio of the standard deviation to the expected value, which is related with a constant that is α_i parameter of the gamma distribution which is the shape parameter.

2.1.1.3.4. Inverse Gaussian Distribution

In some cases, the inverse link function causes negative values for the predicted response variables. In those cases, *log link function* can be used instead of the inverse link function. The following formulation shows the log link function. That link function is appropriate for using the inverse Gaussian distributed response variable.

$$X\beta = \ln(\mu) \tag{2.38}$$

If the link function of GLM is the log link function then, the following equation is derived.

$$g(E(y_i)) = \mathrm{g}(\mu_i) = X_i\beta = \ln(\mu_i) \ = \beta_0 + \ \beta_1 X_i \qquad (2.39)$$

2.1.1.3.5. Poisson Distribution

The following example demonstrates Poisson regression model with Y ~ Poisson (μ). Y are the observed events with the mean parameter μ. The probability distribution of a Poisson random variable Y is presented below:

$$f(Y, \theta, \Phi) = \frac{\mu^y}{Y!} e^{-\mu} \qquad (2.40)$$

The mean function of the Poisson model is:

$$\mu_i = \ \exp(X\beta) \qquad (2.41)$$

The link relations are:

$$\mathrm{g}(\mu_i) = \ \ln(\mu) = \ X\beta \qquad (2.42)$$

2.1.1.3.6. Binomial, Bernoulli, Categorical, and Multinomial Distributions

The following function shows *logit link function*. That link function is appropriate for using Bernoulli, binomial, categorical, multinomial distributed response variables.

$$X\beta = \ln(\frac{\mu}{1-\mu}) \qquad (2.43)$$

The other regression analysis techniques which are also termed as time series techniques moving averages, exponential smoothing, and trend fitting are presented in the next section.

2.1.1.4. Time Series Techniques

As mentioned in the regression analysis techniques, regression analysis models are set up based on an assumption that the response variable is predicted with one or more independent variables. However, in the case of time series techniques, the response variable is not predicted with independent variables, rather the response variable is predicted with its past behavior and the predicting relationship is set up with a function of time. The aim of time series techniques is to find out a pattern with past data set. The reasons for using time series are firstly, the behavior of system is hard to understand and secondly, the main interest is only predicting the future, not why it happens. For these reasons, the implementation of time series techniques is easier than the regressing analysis (Makridakis, 1998).

The next sections explain some known time series techniques like moving averages, exponential smoothing, and trend fitting.

2.1.1.4.1. Moving Averages

The idea of moving averages method is based on calculating the averages of observations which are close to the required point. In this manner, the prediction will more likely to be close values. Taking the averages of observation values causes to eliminate some randomness on data. Before using the moving averages method, the number of included data points must be decided. If three values of data points are included to the method, it is called as the moving average of order three (3 MA). The calculation method of 3 MA is represented below (Makridakis, 1998).

$$T_t = \frac{1}{3}(Y_{t-1} + Y_t + Y_{t+1}) \tag{2.44}$$

Where,

t Represents time series,

T is the predicted value and

Y is the observed value.

The important point of prediction with 3 MA is that there is no predicted value for the first and the last time period because there is no observation before the first and last time periods.

The general formulation of the moving averages method is represented below:

$$T_t = \frac{1}{k} \sum_{j=-m}^{m} \left(Y_{t+j}\right) \tag{2.45}$$

$$m = \frac{k-1}{2} \tag{2.46}$$

Where,

k is an odd integer,

m is the *half width* which is defined as the number of points on start and end side that are included in the average.

An important point of this method is determining the length of a moving average. Large value of k increases the likelihood of the model because of decreased randomness. In spite of that, larger value of k causes to lose more prediction from beginning and end (Makridakis, 1998).

2.1.1.4.2. Exponential Smoothing

Moving average method predicts k observations with equal weight. But most recent data points which are more close to observation points, give the best information for predicting the response variable. So, prediction must be done with decreasing weights while observations get older. The method of predicting future with decreasing weights is named as exponential smoothing.

The formulation of single exponential smoothing is represented below (Makridakis, 1998):

$$F_{t+1} = F_t + \alpha(Y_t - F_t) \tag{2.47}$$

Where,

F_t Represents prediction on time t,

Y_t Represents observation on time t,

$(Y_t - F_t)$ Represents prediction error on time t,

α Represents the weight between the value 0 and 1.

As seen from the formulation above, the new prediction is calculated with adding previous prediction and adjustment for the error which is observed at the previous prediction. If α value is close to zero then, the adjustment value affects the new prediction very little. The effect of smoothing on prediction is to reduce randomness and random variation (Makridakis, 1998).

2.1.1.4.3. Trend Fitting

The value of the response variable can be predicted by fitting data set to a trend model. Trend model is set up by the equation which represents the trend. A trend model can be linear or nonlinear. The following formulation represents linear trend model equation:

$$X_t = a + bt \tag{2.48}$$

Where,

X_t is predicted value,

a is the constant value,

b is the slope of the line,

t is the time period.

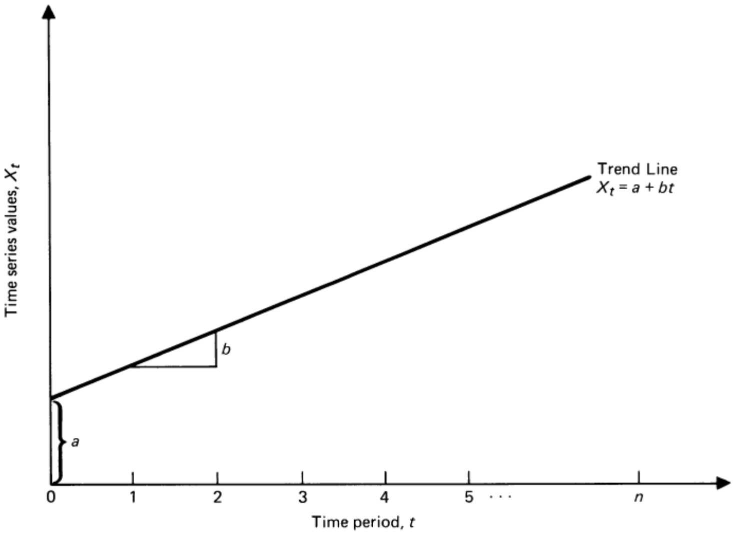


Figure 2.3. Linear trend for time series data
(Source: Makridakis, 1998)

Figure 2.3 represents the equation parameters on the graph. The statistical relationship with the response variable and the predictor variables are fitted on a straight line presented in Figure 2.3. The parameter of the linear trend model “a” presents the constant value and the other parameter of the linear trend model “b” presents the slope of the straight line.

2.2. Classification Techniques

Quantitative predictive techniques are divided into two parts. One of them is regression analysis techniques which are described in the previous sections. The second one is classification techniques.

In predictive modeling, classification techniques are used for predicting class of the response variable with predictor variables. So classification techniques are used for extracting models from the described classes.

The next sections explain some known classification techniques like Support Vector Machine, Decision Trees, and Random forest.

2.2.1. Support Vector Machine (SVM)

Support vectors represent the decision boundaries using the data which is obtained from the training the data set. Figure 2.4 illustrates two possible decision boundaries which divide data set into two different classes. Both of the decision boundaries provide zero training error. As the circle data points are perfectly separated from the rectangular data points. However, the decision boundary B$_1$ provides large margin than the other decision boundary B$_2$. This situation tends to have better generalization error. Also, small margin causes model overfitting (Steinbach, 2006).

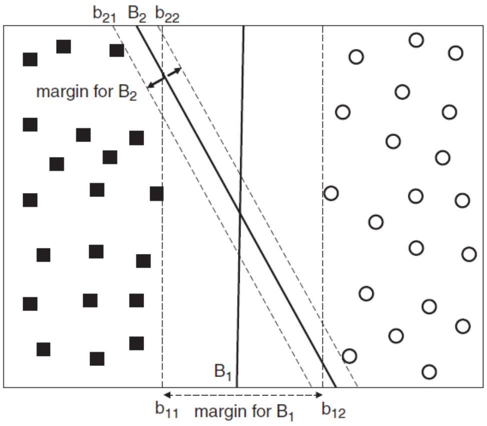


Figure 2.4. Possible decision boundaries
(Source: Steinbach, 2006)

The decision boundary B$_1$ which is related with a pair of hyperplanes demonstrated as b$_{11}$ and b$_{12}$, and which provides the largest margin, is represented by the largest hyperplane. The largest hyperplane is found with maximal margin classifier method. The following formulation represents the decision boundary of a linear classifier.

$$w.x + b = 0 \tag{2.49}$$

Where w and b are the parameters of the model. The decision boundary divides the training set into two parts. If we consider the binary classification problem, the following equations test the class of the data points.

$$y = \begin{cases} 1, & if\ w.z + b > 0 \\ -1, & if\ w.z + b < 0 \end{cases} \tag{2.50}$$

Where,

y is ∈ (1,-1),

z is the test sample which classifies the data points with -1 or 1.

$$b_{i1}: w.x + b = 1 \tag{2.51}$$

$$b_{i2}: w.x + b = -1 \tag{2.52}$$

b_{i1} and b_{i2} are two parallel hyperplanes. The distance between these two hyperplanes gives the margin of the decision boundary and it is represented by the following equation.

$$w.(x_1 - x_2) = 2 \tag{2.53}$$

$$\|w\| \times d = 2 \tag{2.54}$$

$$\therefore d = \frac{2}{\|w\|} \tag{2.55}$$

Where,

x_1 is located on b_{i1} and x_2 is located on b_{i2},

d is the margin of the decision boundary.

SVM requires the maximization of the margin of decision boundary. The margin is maximized by minimizing the following function:

$$f(w) = \frac{\|w\|^2}{2} \tag{2.56}$$

Minimization of the previous equation is a convex optimization problem and it can be solved by using the standard Lagrange multiplier method. The lagrangian form of the problem is given in the following equation:

$$L_p = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{N} \lambda_i (y_i(w.x_i + b) - 1) \qquad (2.57)$$

Where,

N is the number of training data,

y_i is ∈ (1,-1),

$\lambda_i$is the Lagrange multipliers.

The derivative of L_p with respect to w and b and setting the results to zero gives the minimization of the Lagrange equation.

$$\frac{\partial L_p}{\partial \mathrm{w}} = 0 \xrightarrow{yields} w = \sum_{i=1}^{N} \lambda_i y_i x_i \qquad (2.58)$$

$$\frac{\partial L_p}{\partial \mathrm{b}} = 0 \xrightarrow{yields} w = \sum_{i=1}^{N} \lambda_i y_i = 0 \qquad (2.59)$$

Finally, the decision boundary is represented by the following equation (Steinbach, 2006).

$$\left(\sum_{i=1}^{N} \lambda_i y_i x_i . x = 0\right) + b = 0 \qquad (2.60)$$

2.2.2. Decision Trees

For solving the classification problem, questions are asked consecutively until reaching the class label. These questions and their answers are designed as a form of decision tree. That decision tree has nodes and directed edges. The leaves of the decision tree indicate class labels (Steinbach, 2006).

A lot of decision trees can be constructed with the given set of attributes. Some decision trees could give more accurate results than the other combinations. Finding the global optimal tree might be hard because of the combination of the number of attributes.

So, some algorithms are focused on finding the locally optimal decision tree. Hunt's algorithm is one of the algorithms which induces decision tree (Steinbach, 2006).

Training records are partitioned into successively purer subsets while a decision tree is grown in a recursive form in Hunt's algorithm.

D_t is the set of training records which are related with node t.

y={y_1, y_2,… y_c,} are the class labels.

Steps of Hunt's algorithm is given below.

Step 1: t is a leaf node labeled as y_t, when all the training records (D_t) related to the same class y_t.

Step 2: For partitioning the records into smaller subsets, an attribute test condition is selected, when the training records (D_t) have records which are related to more than one class. A child node is generated for each output of the test condition. And training records are distributed to children as a result of the output. After that, the algorithm is applied to each child node recursively (Steinbach, 2006).

If every combination of attributes has unique class label and every combination of the attribute values exist in the training data then, Hunt's algorithm can be applied. Hunt's algorithm constructs the basis of many decision tree induction algorithms, for example classification and regression tree (CART), ID3 (Iterative Dichotomiser 3) algorithm, and C4.5 algorithm (Steinbach, 2006).

2.2.3. Random Forest

Random forest is designed for decision tree classifiers. Multiple decision trees are combined to make the prediction. Each decision tree is created based on random vectors which are shown in the figure below, and the probability distribution of the random vectors are fixed:

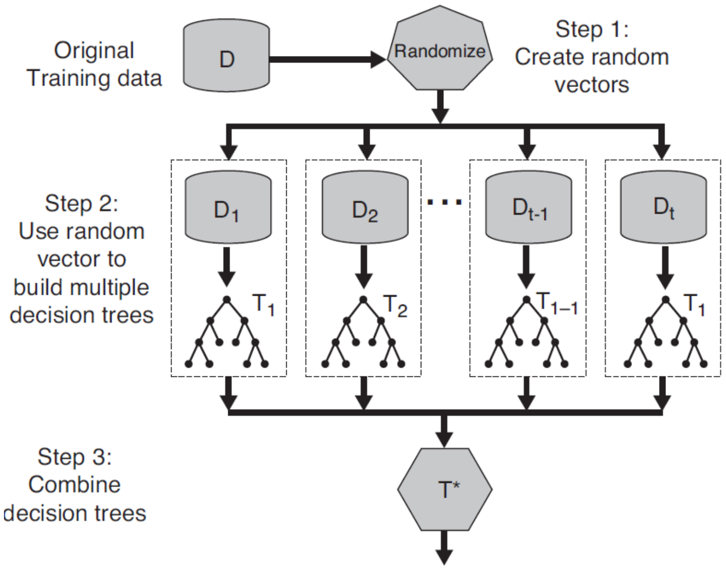


Figure 2.5. Random Forest
(Source: Steinbach, 2006)

Growing process of decision tree has been done with many ways. Firstly, input features (F) are randomly selected to split each node of the decision tree. In this manner, splitting decision of a node is determined by selected features thus there is no need to examine all the available features. As there is no pruning, the tree grows with its integrity. This method helps to reduce bias in the resulting tree. Majority voting schema is used to combine the predictions when the decision trees have been constructed. This method is known as Forest-RI and RI means random input selection. The correlation of random forest depends on the size of the features. If the size of features is big enough then, decision trees become more correlated. More correlation increases the strength of random forest. Randomly selected number of features is calculated by following equation (Steinbach, 2006).

$$F = log_2 d + 1 \tag{2.61}$$

Where,

F is the randomly selected number of features

d is the number of input features.

Selection of an independent set of random features is difficult when the number of input features (d) is small while building the decision tree. Creating linear combinations of the input features (L) solves the small feature space problem. Randomly choosing L of the input features generates a new feature for each node. With the range of [-1,1], uniform distribution coefficients are generated and then, input features are combined with these coefficients. Hereby, randomly combined new features are generated for each node and the best of the features are selected for splitting the node. This method is called as Forest-RC (Steinbach, 2006).

For each node of the decision tree, one of the F best splits are randomly selected for generating the random tree. Even if F is not big enough, this way leads to generate more correlated trees than Forest-RI and Forest-RC. However, this method does not give the results fast because of examining all the splitting features at the node of the decision tree (Steinbach, 2006).

2.3. Model Selection and Model Validation

The purpose of modeling is inferring mathematical or logical relationship from a system to understand its behavior (Law, 1991). The future behavior of a system can be predicted by using a model. A model is set up with a response variable which is the predicted variable, and predictor variables which represent inputs or causes.

In the modeling process, firstly, response and predictor variables are defined for the problem. Then, parameter estimation is performed for estimating the parameters of the model. After that, model fitting tests are performed.

This section firstly discusses the bias and variance trade-off and model complexity, then parameter estimation methods for choosing the appropriate predictor variables are presented. After that, model selection methods are presented for choosing the best model which represents the input data, finally model validation concept and generalized linear modelling steps are presented

2.3.1. Bias and Variance Trade-off

There are two issues which are error due to bias and error due to variance, considered while evaluating the goodness of fit of the predicted model. Minimizing both variance and bias is the required situation, however due to trade-off between bias and variance, it’s not achievable.

Error due to bias is defined as the difference between the expected value of prediction and the real value which is tried to be predicted. The distance between the real value and the prediction is evaluated by bias.

Error due to variance is defined as the variability of prediction for specified data points. In other words, error due to variance shows variety of estimates around its average (Fortmann-Roe S, 2013).

The following figure illustrates variance and bias for a predicted model.

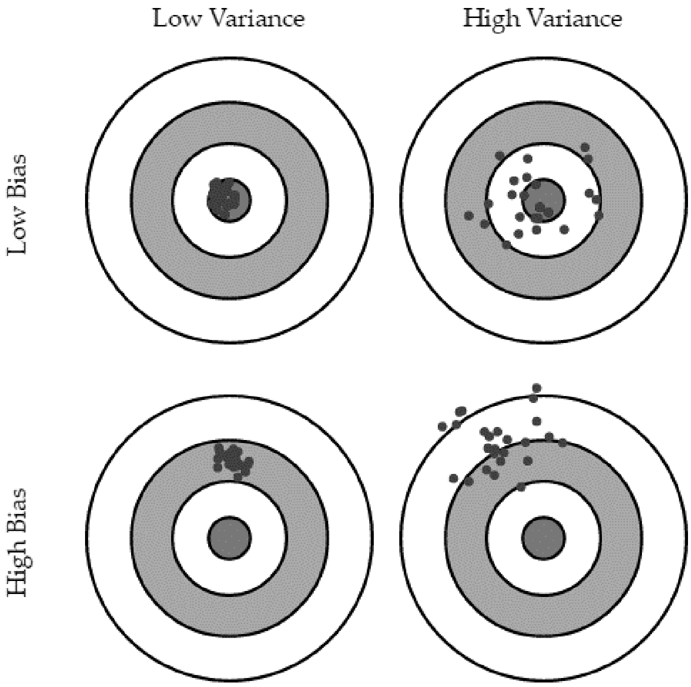


Figure 2.6. Illustration of bias and variance
(Source: Fortmann-Roe S, 2013)

In Figure 2.6, the center of circles state the real value which is tried to be predicted. Points away from the center mean a worse prediction. The group of points on circles implies repetitive prediction on the given predicted model. Spreads of prediction mean high variance and also worse prediction.

The following equation gives the bias definition of θ estimator which is the parameter of y_i's distribution (Lebanon, 2010).

$$Bias(\hat{\theta}) = E(\hat{\theta}) - \theta \tag{2.62}$$

The following equation gives the mean square error (MSE) definition of θ estimator (Lebanon, 2010).

$$E\left(\left\|\hat{\theta}-\theta\right\|\right) = Var\left(\hat{\theta}\right) + \left\|Bias(\hat{\theta})\right\|^2 \qquad (2.64)$$

The following theorem gives the relationship between MSE and variance, bias. The proof of that theorem can be viewed at Lebanon's "Bias, Variance, and MSE of Estimators" study (Lebanon, 2010).

$$E\left(\left\|\hat{\theta}-\theta\right\|\right) = Var\left(\hat{\theta}\right) + \left\|Bias(\hat{\theta})\right\|^2 \qquad (2.64)$$

As we see from the equations above, decreasing bias and variance also decreases MSE. To decrease bias, the expected value of prediction must be closer to real values by adding the parameters to model. Adding parameters increase model complexity and also causes over-fitting of the model to data (Fortmann-Roe S, 2013).

To decrease variance, the expected value of prediction must be closer to predicted values. Low variance provides simpler model but it causes under-fitting of the model to data (Fortmann-Roe S, 2013).

The following figure illustrates the trade-off between variance and bias and its relation with MSE.

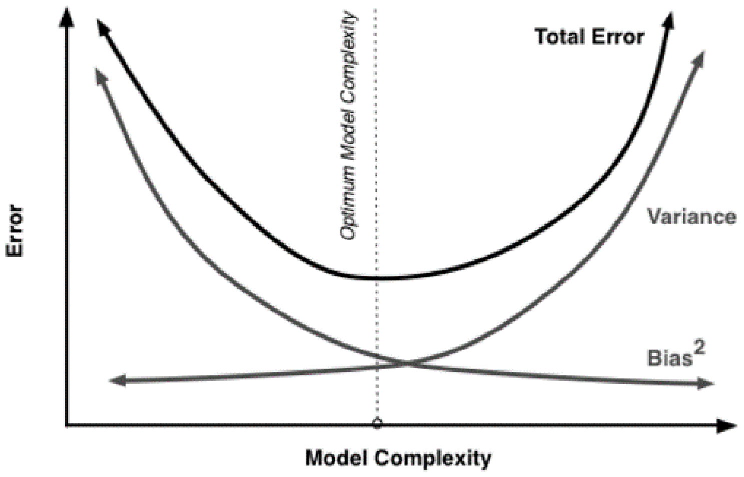


Figure 2.7. Bias and variance trade of
(Source: Fortmann-Roe S, 2013)

As we see from Figure 2.7, decreasing bias causes high variance and high model complexity as a result it increases error and MSE. Decreasing variance causes high bias and high error and high MSE. Optimum model complexity could be found while minimizing the total error.

Another illustration of the bias and variance trade-off with model complexity is presented in the following figure.

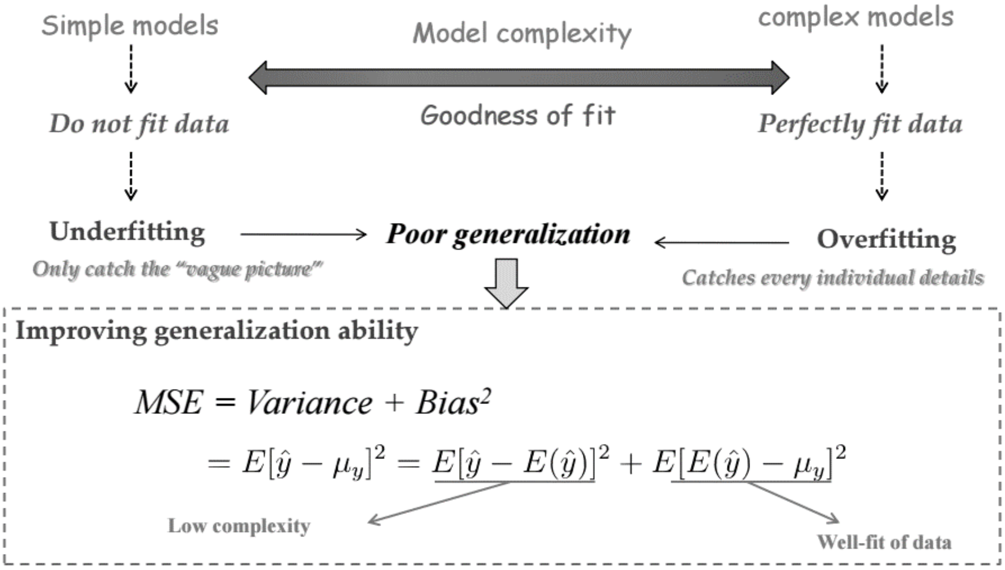

Simple models
Model complexity
complex models
Goodness of fit
Do not fit data
Perfectly fit data
Underfitting
Poor generalization
Overfitting
Only catch the "vague picture"
Catches every individual details
Improving generalization ability
$MSE = Variance + Bias^2$
$= E[\hat{y} - \mu_y]^2 = \underline{E[\hat{y} - E(\hat{y})]^2} + \underline{E[E(\hat{y}) - \mu_y]^2}$
Low complexity
Well-fit of data

Figure 2.8. Bias and Variance Trade-off with Model complexity
(Source: Li, 2011)

Making a good prediction with a simple model is the most desired situation. However, as seen in Figure 2.8, good prediction needs complex models which causes overfitting and simple model which do not well fit to the data causes under fitting. Model complexity trade-off is formulized as a cost function which needs to minimized, with bias and variance values. The low variance represents a low complexity and the low bias represents a well fit of data (Hand, 2001).

2.3.2. Parameter Estimation

In generalized linear modeling, estimating the parameters of the selected model is important for modeling and predicting the future data. This section presents the parameter estimation techniques, like method of moments, maximum likelihood estimation, and least squares method.

2.3.2.1. Method of Moments

Method of moments which is a heuristic method, is one of the parameter estimation methods. In a probability distribution, θ is the canonical parameter and Φ is the dispersion parameter. Finding those parameters by using method of moments, population mean is equal to $\dot{a}(\theta)$ (De Jong, 2008).

θ is the canonical parameter and Φ is the dispersion parameter of the exponential family probability function. $a(\theta)$ is the known function of the exponential family probability function. $\dot{a}(\theta)$ is the first derivative of the function $a(\theta)$ and $\ddot{a}(\theta)$ is the second derivative of the function $a(\theta)$.

Here $\bar{y}$ is the sample mean and the variance is equal to

$$\dot{a}(\theta) = \bar{y} \tag{2.65}$$

$$\Phi\,\ddot{a}(\theta) = \hat{\sigma}^2 \tag{2.66}$$

For example, if the distribution is the standard normal distribution then, method of estimators use the following equations.

$$\hat{\mu} = \bar{y} \tag{2.67}$$

$$\Phi = \hat{\sigma}^2 \tag{2.68}$$

According to these equations, the first derivative of $a(\theta)$ function is constant θ parameter and the second derivation of the $a(\theta)$ function is equal to 1.

$$\dot{a}(\theta) = \theta \tag{2.69}$$

$$\ddot{a}(\theta) = 1 \tag{2.70}$$

2.3.2.2. Maximum Likelihood Estimation (MLE)

Statistical model parameters are estimated by using the maximum likelihood estimation (MLE) method. For example, normally distributed data parameters which are mean and variance can be estimated by using the MLE method. MLE method chooses the model parameters which maximize the likelihood function.

Let $f(y_i;\ \theta, \Phi)$ is the probability distribution function. If the y_i are independent then the joint probability distribution is shown in the following formulation:

$$f(y;\ \theta, \Phi) = \prod_{i=1}^{n} f(y_i;\ \theta, \Phi) \tag{2.71}$$

The log likelihood function is the logarithm of the previous function. The log likelihood function is shown in the following formulation:

$$\ell(\theta, \Phi) \equiv \sum_{i=1}^{n} ln\, f(y_i;\ \theta, \Phi) \tag{2.72}$$

The maximum likelihood methods select the parameters $\theta,\ \Phi$ which maximize the likelihood function (De Jong, 2008).

2.3.2.3. Least Squares Method

Least squares method is generally used for estimating the regression coefficients in a multiple linear regression model. That method selects the parameters which are included in a model, and minimizes the sum of squares of error terms in the model. The function of least square is shown in the following formula:

$$S = \sum_{i=1}^{n} \varepsilon_i^2 \tag{2.73}$$

$$S = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2 \tag{2.74}$$

In the formula; β_j are the parameters of the model. The function of least squares must be minimized with regard to β_j parameters (Myers, 2010).

2.3.3. Model Selecting Methods

After fitting the data to candidate models, model selection study can be done for choosing the best model which represents the input data with different combinations of predictor variables. This section presents different model selection metrics which are sum of squares, r square, mean square error and Akaike's information criterion.

2.3.3.1. Sum of Squares

The sum of squares of vertical distances between real values of data points and the predicted values of data points is named as the sum of squared residuals (SS_{res}) or the sum of squared errors of prediction (SS_{error}). The formulation of SS_{res} is represented in the following formula (Cochen, 1995);

$$SS_{res} = \sum_{i=0}^{n} (y_i - \hat{y}_i)^2 \tag{2.75}$$

The sum of squares of vertical distances between the predicted data points and the mean of data points is named as the sum of squared regression (SS_{reg}). The formulation of SS_{reg} is represented below (Cochen, 1995);

$$SS_{reg} = \sum_{i=0}^{n} (\hat{y}_i - \bar{y})^2 \tag{2.76}$$

The sum of squares of vertical distances between the real values of data points and the mean of data points is named as the sum of squared total (SS_{total}). And also, sum of SS_{reg} and SS_{reg} gives the SS_{total} value. The formulation of SS_{total} is represented below (Cochen, 1995).

$$SS_{total} = SS_{reg} + \ SS_{res} \tag{2.77}$$

$$SS_{total} = \sum_{i=0}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=0}^{n} (y_i - \hat{y}_i)^2 \tag{2.77}$$

$$SS_{total} = \sum_{i=0}^{n} (y_i - \bar{y})^2 \tag{2.78}$$

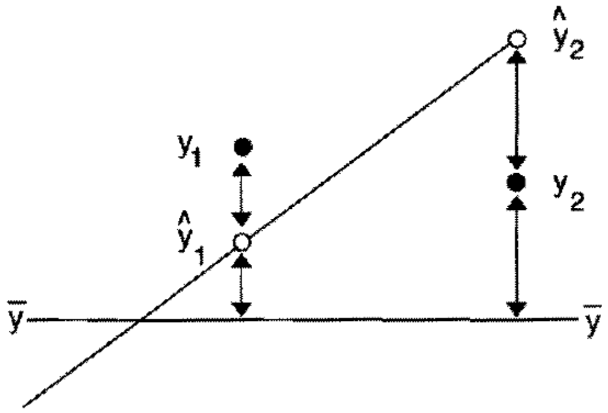


Figure 2.9. The decomposition of residuals from the mean (Source: Cohen, 1995).

Figure 2.9 represents the real values of data points $(y_1, y_2, \dots y_i)$, the predicted values of data points $(\hat{y}_1, \hat{y}_2, \dots \hat{y}_i)$ and the mean of data points $(\bar{y})$ which are used in SS_{total}, SS_{reg} and SS_{res} calculation.

Sum of square calculations are used for mean square error calculation, coefficient of determination (R^2) calculation and f test statistic calculation, so sum of square calculations are useful for model comparisons.

2.3.3.2. R^2

R^2 statistic shows how well models' predictors fit to the given input data. R^2 takes value between 0 and 1. If R^2 statistic takes value 1 then the model perfectly fits data and the fitting curve comes closer to data, else if R^2 statistic takes value 0 then the model does not fit data.

If R^2 takes value 0, the predictor variables do not help to predict the response variable. If R^2 takes value 1, predictor variables predict the response variable exactly. Even if R^2 takes a higher value, that does not mean that the fit is sensible, that means, the best fit values of the predictors may have infeasible values like the negative value of date time or confidence intervals may be very wide (Motulsky,2003).

R^2 formulation is represented below (Cohen, 1995):

$$R^2 = 1 - \frac{SS_{res}}{SS_{total}} = \frac{SS_{reg}}{SS_{total}} \tag{2.79}$$

As seen from the formula above, if the sum of squared regression value come close to sum of squared total then, R^2 value increases and the model fits to data better. On the other hand, if the sum of squared residual value come close to sum of squared total then, R^2 value decreases and model does not fit to data.

2.3.3.3. F Test Statistic

F test statistic is used for comparing the statistical models which are fitted to a data set. So F test statistic shows how well the predicted model fits to the given input data. F test statistic compares the mean square error and mean square regression (MSR). MSR is calculated by the sum of square regression which is divided by associated degrees of freedom. Following equations presents the MSR and F test statistics (Kutner, 2004):

$$MSR = \frac{SS_{reg}}{1} = \sum_{i=0}^{n} (\hat{y}_i - \bar{y})^2 \tag{2.80}$$

$$F^* = \frac{MSR}{MSE} \tag{2.81}$$

F test statistic tests the hypothesis that the suggested model fits the input data well. Following inequality test the H_0 hypothesis test.

H_0: There is no significant relationship between the predictor variables and the fitted model (β=0)

The alternative one is:

H_1: There is a significant relationship between the predictor variables and the fitted model (β≠0)

$$if\ F^* \leq\ F(1-\alpha; 1, n-2)\ conclude\ H_0 \tag{2.82}$$

Where α is the significance level which controls the risk of Type I error. Type I error can be define as the incorrect rejection of a true null hypothesis where n is the number of observations.

2.3.3.4. Mean Square Error (MSE)

Mean Square Error (MSE) statistic is another metric that shows the quality of predictors. MSE formulation is represented below:

$$MSE = \frac{SS_{res}}{n} \tag{2.83}$$

The relation between R^2 and MSE can be set up as the following formulation. That formulation shows that changes on MSE also affects R^2.

$$R^2 = 1 - \frac{MSE}{SS_{total}}\, n \tag{2.84}$$

As a result, the decrease of MSE increases R^2 and also increases the model fitting on the given data (Kutner, 2004). So R^2 and MSE give similar information about the model fitting.

Decreasing the MSE provides the well fit of data with chosen predictor variables but it causes the complex model fitting and overfitting of data. On the other hand, increasing the MSE causes the bad fit of data, but it provides simple models and low complexity. This condition is also explained in the manner of bias and variance trade-off.

2.3.3.5. Akaike's Information Criterion (AIC)

Akaike's Information Criterion (AIC) is the metric that is used to evaluate the goodness of fit of a model. That metric is used for comparing models which are fitted to any data with different parameters and coefficients. AIC can be used for nested or nonnested models. A model which is a simpler case of another model is named as a nested model (Motulsky,2003).

AIC formulation is represented below.

$$AIC = -2\ell + 2p \tag{2.85}$$

In the formulation, ℓ symbol is the log-likelihood function of the model that is fitted to data. The scatter of points around the regression curve determines the probability distribution of the response variable. According to probability distribution of the response variable, the likelihood function changes. Low value of AIC means good fit of model given data. p symbol means the number of parameters in the model. More parameter in the model increase the model complexity and AIC value. More parameters are penalized with higher AIC (De Jong, 2008).

2.3.4. Model Validation

One of the important step of model building is the validation of the selected model. Commonly, model validation is done by controlling a candidate model with independent data. There are three essential methods exist for validating a model:

I. Collection of new data and checking the model predictive ability is one of the methods of validation.

II. Empirical results are compared with the theoretical expectations and the simulation results.

III. Hold-out samples which are samples of data that are not used in fitting a model, are used to check the model predictive ability (Kutner, 2004).

2.3.4.1. Method of Checking Validity

One of the method of checking the validity of the predicted model is *re-estimating* a model with new data set which is a newly collected data set or simulated data set, and comparing regression coefficients and its characteristics with the model which is predicted before.

Another method of checking the validity of a predicted model is to use method of *mean square prediction error* (MSPR) calculation. Formulation of MSPR is represented below:

$$MSPR = \frac{\sum_{i=1}^{n^*}(Y_i - \hat{Y}_i)^2}{n^*} \tag{2.86}$$

Where,

Y_i Represents response variable value in i th validation case.

$\hat{Y}_i$ Represents predicted value in i th validation case.

n^* Represents number of cases in the validation data set

The result of MSPR calculation is compared with MSE, if two values are close to each other, then the predictive capability of the selected model is acceptable for using the prediction because the selected model is not critically biased.

In some conditions, collecting new data is not practical for validating the selected model. Because new data must be collected with the controlled experiment environment. The nature of data, controlled experiment environment cannot be provided. In that situation, observed data set is split in two sets where, one of them is the training set and another is the validation set. That method is named the *cross-validation*. The data must be large enough to use cross-validation method. Because the process of selecting the validation data set can cause to significant variation over data sets. If the data set is small then, the validation is done many times to reduce the variance. The next section presents cross-validation methodology.

2.3.4.2. Cross-Validation

The basic idea of the cross-validation method is to split the data into two parts which are the training part and the validation part. The model is set up and the parameters of the model are estimated by using the training part. The validation part is used to calculate the score function that implies how well the candidate model fits the data. The most commonly used score functions are sum of squares, r square, MSE and AIC which are described in the model selection methods in the previous sections. Cross-validation is the most preferred method in practice because it is robust and easy to implement (Hand, 2001). On the other hand, seasonal time series data can be problematic for implementing the cross-validation method. Because the validation data set and training data set must come from the same distribution for example the gamma distribution. If the validation data is selected randomly from a part of the whole data, the validation data may not have the same distribution with the training part. For example, if the data has four parts which represent the four different seasons' sales (summer, fall, winter, spring), and let the validation data set is selected as the summer sales, then the selected data set may not have the same distribution with the rest of the data.

The next section presents common types of cross-validation methods which are (Gutierrez, 2016):

- k-fold Cross-Validation,
- Random Subsampling

2.3.4.2.1. k-fold Cross-Validation

k-fold cross validation is a commonly used model validation method. The data set is randomly split into k equal sized subsamples. k-1 data are used for the training data and 1 subsample are used as the validation data. Validation evaluation is done for k times until each data set is used for the validation process. After that, score functions are averaged. The following figure illustrates the k-fold Cross-Validation method where k is equal to 4.

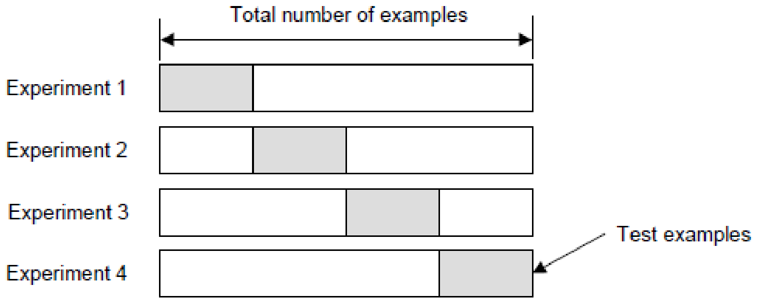


Figure 2.10. The decomposition of residuals from the mean
(Source: Gutierrez, 2016)

In Figure 2.9, gray parts are the test examples which represent the validation part. The rest of data set represents the training part.

The following equations are present the calculation of classifier error (Gutierrez, 2016).

$$E = \frac{1}{k}\sum_{i=1}^{k} E_i \tag{2.87}$$

Where,

E is the classification error.

k is the folds which divides the training examples to k times equal parts.

i is the positive integer which goes to k. $i = 1, \ldots k$:

2.3.4.2.2. Random Subsampling

The data set is randomly split into the training and the validation parts. After that, model is fitted to the training set and the average of score function is calculated. This method is known as the Monte Carlo cross-validation method. The following figure illustrates the random subsampling method:

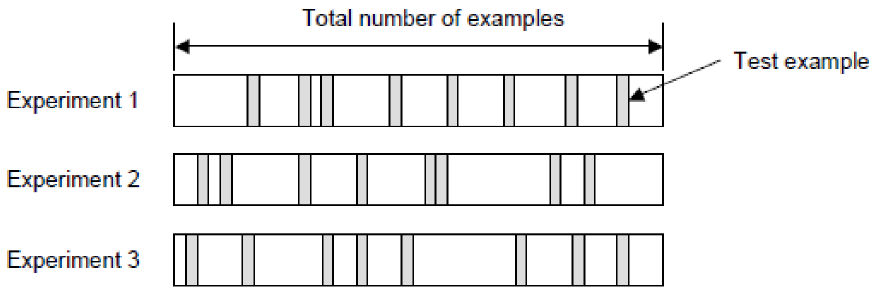


Figure 2.11. Random Subsampling illustration
(Source: Gutierrez, 2016)

In Figure 2.10, gray parts are the test examples which represent the validation part. The rest of the data set represents the training part. Experiments run several times to obtain accurate an estimate. The calculation of the classifier error is the same with the k-fold cross-validation method.

2.3.5. Generalized Linear Modeling Steps

Generalized linear modelling steps are similar beyond all modelling methods. Some distinctive features are included modelling steps for example predicting response variable distribution, and presented below (De Jong, 2008):

I. An appropriate distribution (exponential, gamma, normal distribution etc.) is defined for the response variable.

II. An appropriate link function is determined according to the response variable distribution.

III. Predictor variables are determined for the model.

IV. Observations are collected as response variables along with corresponding predictor variables.

V. The model is fitted by estimating the parameters of models. Maximum likelihood estimation or other estimation methods are used for parameter estimation.

VI. The fitted model is examined by appropriate model fitting metrics. For example, R square, F-test, Mean Square Error, Akaike’s information criterion etc. (De Jong, 2008)

CHAPTER 3

LITERATURE REVIEW

Data mining algorithms can be classified in two groups; Descriptive and predictive. In descriptive category, the techniques do not use predefined classes yet describe. Predictive methods, on the other hand, build a model using data instances and their predefined data classes then predict these class memberships for unseen data.

The predictive methods are described as qualitative or quantitative method based on the existence of a measurement schema. While qualitative methods do not include measurement or statistical work, quantitative methods include statistical and mathematical techniques for empirical evaluation.

In quantitative predictive data analytics, there are two main approaches which are classification and regression. The main distinction between these are in the classification case the dependent variable to predict is categorical while in regression it is numerical.

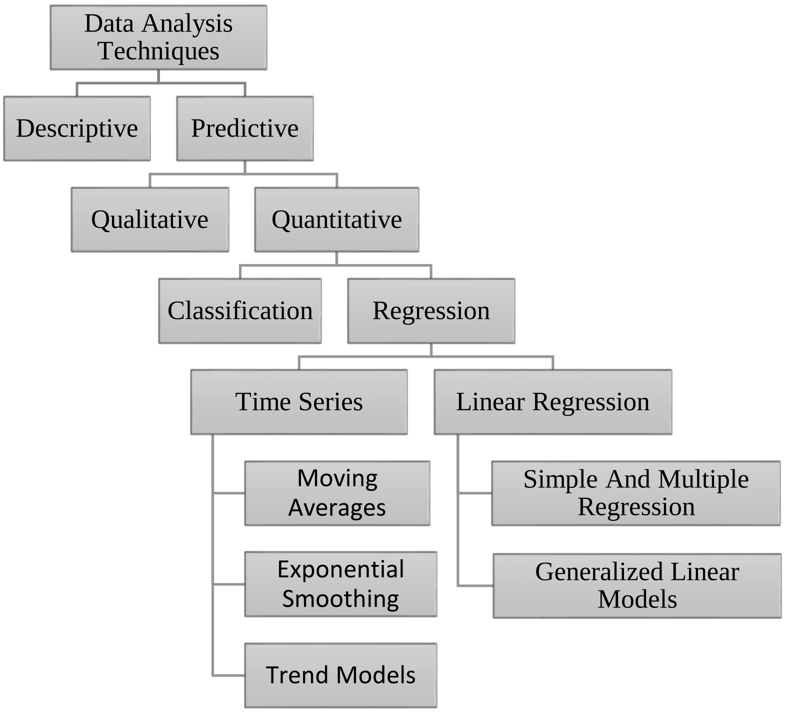


Figure 3.1. Data analysis techniques

Regression techniques can further be decomposed into two groups such as Linear Regression and Time Series (Figure 3.1).

In quantitative methods, regression methods can be divided in two parts. One of them is time series. Moving averages, exponential smoothing and trend models are investigated in time series part that are referred on Kirby's study. Second part of the regression methods is linear regression studies. Linear regression studies can be covered in two groups. One of them is Simple and Multiple Regression and the other is Generalized Linear Models.

Kirby, M. (1966) worked on Moving Average, Exponential Smoothing, Least Squares Model models which are in the quantitative methods. The aim of that his studies was comparing different methods accuracy to predict short and intermediate range prediction. He studied on synthetic series data because of cyclical or trend or noise components. He intended to predict intermediate and short range period like six month or one month. His studies was only with sales history data without independent variables which effect the sales demand. He thought that variables caused noise. For example, advertisements and some quality problems make noise on data. As a result of his study; medium range prediction produced more accurate result than short term prediction. But with synthetic cyclical data; the result was opposite with medium range result. That means short term prediction is more accurate than medium range prediction for synthetic cyclical data. In addition to that study, exponential smoothing was more accurate on cyclical data but moving averages method was more accurate on noise characteristics data. Exponential smoothing method was more accurate than moving averages in short term. Exponential smoothing and moving averages methods were more accurate than least squares methods in short and medium range. And also exponential smoothing and moving averages methods have same accuracy in medium range prediction. But least squares method is more accurate for long term prediction (Kirby, 1966).

Carlson and Umble (1980), tried to model customer purchase behavior in automobile market. And they investigate the energy crisis and its effect on automobile demand. They modelled car demand as regression model with linear structure (Carlson, 1980).

GLM studies are investigated under Linear Regression studies. GLM with data mining implementations, smoothing approaches and predictive methods are surveyed in the following studies.

Kolyshkina (2005) combined powerful side of data mining techniques and advantages of GLM in her studies. She bring together computational power of data mining tools and powerful properties of GLM in her study. The goal was choosing best fitted independent variables for model by making exploratory data analysis in that study. The choice of necessary input variables for GLM was supported by using MARS (multivariate adaptive regression splines) data mining method. When the output of MARS method was used as an input for GLM, the result was better than hand fitted model. In that study, the MARS method defined like that. First, different value intervals of each independent variables in a model are detected. Then, that intervals are divided sub values and different models are tried to fit each independent interval. According to that, significant degree of each independent values are calculated for each fitted model and independent variables and value intervals are determined (Kolyshkina, 2005). Another predictive methods which is similar to MARS method is Conditional Inference Tree (CTree) method which is also systematically trying all the combinations of the variables to select the right predictor variables. CTree creates a decision tree model. It generates splits iteratively. This splits are generated for most significantly related variable with the response variable. That response variable is evaluated by p values. Iterations finishes when there is no more significant p value available for remaining variables (Du, 2014).

Hothorn, Hornik, and Zeileis (2004) introduced CTree which is a regression trees and which is a member of non-parametric class. Also CTree is covering tree structured regression models. The technique of CTree can be performed to the regression models with a response variables which are nominal, ordinal, numeric and multivariate variables (Hothorn, 2004), (Hothorn, 2015).

Liu (2004) studied on GLM that is parametric model and nonparametric models. She present methods names smoothing spline’s for testing and comparing nonparametric models with parametric models because of avoiding misleading results of parametric models. GLM is a parametric model and it must be validated to avoid unexpected results (Liu, 2004).

O'Sullivan (1986) used nonparametric smooth method with parametric GLM together. Smoothing method provides that to avoid sharply increases in a function which is based on GLM and also it provides to observe close results in each other rather than sudden increases in interval probability values (O'Sullivan, 1986).

GLM can be modeled according to type of dependent variable distributions. Some of that GLM models are probit models and logtit models. In demand prediction, it is important not only estimating the amount of demand but also estimating the presence of demand. Linear models which use logit link function serves that goal and it provides the probability value of the demand prediction between zero and one. Thus, some abnormalities are prevented for example, the value of the predicted value is being under zero.

Tauras, A. (2005) studied on logit model which is subtype of GLM. He tried to model density of smokers to smoke in a population. The problem was modelled with both ordinary least squares (OLS) method and GLM. Results were compared in each other and tried to understand importance of prediction bias which appears on omitting of the error terms while data transformation. The result of OLS method, the effect of price which is independent variable is significantly much more than GLM result. The reason of that difference, GLM does not requires retransformation while OLS needs retransformation (Tauras, 2005).

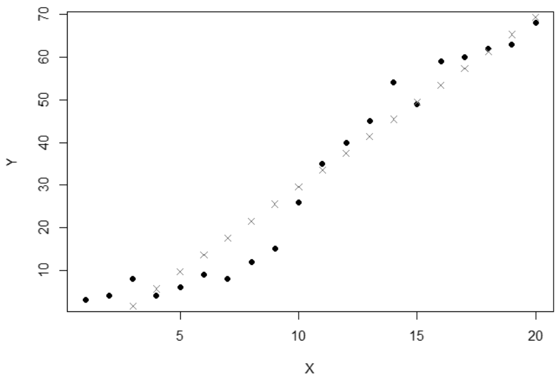
Another study about GLM with predictive data analysis is done by the Odum Instituted which tried to predict behavior of voting to Obama or Romney in 2012 American national election by using logit model which is in the family of GLM. In the study, dichotomous dependent variable was evaluated as a function of one or more independent variable in a logit model. Parameters of logit model was calculated by using maximum likelihood estimation (MLE) method. As a reason of that the Odum Instituted specifies, if the data set is sufficiently big enough then MLE method produces better results (The Odum Institute, 2015).

Classification is the second main approach for predictive data analytics techniques. There are some fundamental data mining methods like random forest, support vector machines, decision trees etc.

Random forest which is introduced by Leo Breiman (2001), is a machine learning method for classification in data mining algorithms. Random forest is constructed by collection of tree structured classifiers (Breiman, 2001). Lots of decision trees are combined by Random Forest algorithm. The aim of that algorithm is to reduce variance of predicting while keeping the bias at low level (Du, 2014). And also the algorithm does not over fit to sample data because of the Law of Large Numbers. Random events of Law of Large Numbers provide accurate classifiers and regressors by the help of individual predictor’s correlations (Breiman, 2001).

Corinna Cortes and Vapnik (1995) introduced support vector machine which is a machine learning method in 1995. That method used for two group classification problems. Main idea for algorithm is input vectors are mapped nonlinearly with a priority. The algorithm changes the solution from linear surface to nonlinear surface (Cortes, 1995).

In the following figures represents an example of fitting training data with linear model and support vector machine method. The graph in Figure 3.2 is the linear fitting of the training data, as well, in Figure 3.3 is the support vector machine fitting data. As seen from in Figures 3.2 and 3.3, support vector machine method is well fitted to training data than linear modelling method.



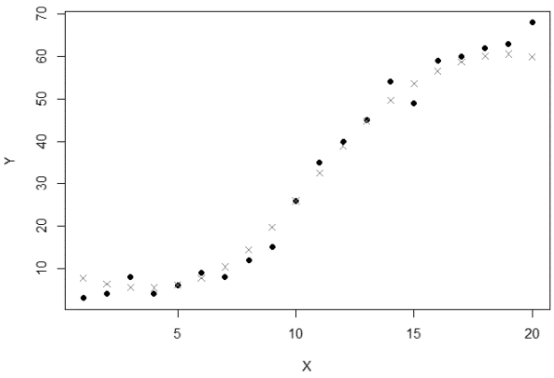


Figure 3.2. Linear fitting of training

Figure 3.3. Support vector fitting of training data

Principal component analysis (PCA) was invented by Karl Pearson (1901). With principal component analysis (PCA) method, independent variables in an observations which are possible correlated variables, are changed to linearly uncorrelated variables by using an orthogonal transformation. These changed variables are named principle components. Principal component regression (PCR) is a regression analysis method that is a form of PCA which is defined before. PCR model predicts dependent variable with a set of independent variables while using linear regression model however, PCR model uses principal components which are defined in PCA model (Jolliffe,1982), (Jolliffe,1986).

GLMNet is a regularized version of GLM. The regularization overcomes overfitting by adding terms to the cost function of the learning model. The addition of these terms in general push the parameters of the learning model towards a prior value. In the case of GLMNet there are two such terms namely ℓ1 (the lasso) and ℓ2 (ridge regression). The target regression problems in context are linear, two class logistic and multinomial regression model problems. To acquire a sparse solution for regression models, ℓ1 (the lasso) penalty term is used. Ridge regression (ℓ2) shrinks the estimated coefficients with shrinking method which adds a penalty on coefficients (Friedman, 2010). The mixture of ℓ1 and ℓ2 penalties is named as the elastic net regularized regression method which is GLMNet (Friedman, 2010).

The mixture of ℓ_1 and ℓ_2 penalties is named the elastic net which is a regularized regression method. Regularization method tries to prevent overfitting problems. For predicting a model which is fitted to data, some known methods are used. For example, least squares method. If the predicted model fits very well with the training data, over fitting problem arises. For solving that problem, a new term is included into the goal function which is minimized the error terms of the fitted model with the training data. The added term is named regularized term and the method is the regularization method.

Boosting is a method of machine learning which produces a combined strong classifier out of weak learners. The weak learner algorithm is run on the dataset and according to a loss function, an updated version of the weak learner is introduced. The data distribution is updated so that the misclassified points in the dataset get higher weights. Then, the updated weak learner algorithms are run repetitively in this manner. The result is a combination of weak classifiers weighted with respect to the loss function outcome per iteration. The basic assumption in a boosting scheme is that the selected weak learner algorithm is at least better than a random classifier. There are mainly two varying components namely the cost function and the weak learner in different boosting methods. In the case of Gradient Boosting Method, the weak learner is selected in the direction of the negative gradient of the loss function (Du, 2014).

CHAPTER 4

EXPERIMENTS AND COMPUTATIONAL ANALYSIS

In this chapter, first, exploratory data analysis studies are done for understanding the input data. After that, model fitting studies are done for making prediction and they are compared with each other.

4.1. Exploratory Data Analysis

This section starts with introducing the data set which is used for the model fitting. Then it continuous with the visual representation of the given data set. After that, the response variable distribution is predicted for implementing the Generalized Linear Model (GLM), method. At the end of this section, some preprocessing studies are done on the continuous predictor variables for investigating the effects of discretization.

4.1.1. Problem Definition and Data Description

We have five years' (2010, 2011, 2012, 2013, 2014) sales data for a company. Sales records include the date of sales, sales amount, product item price, in which city these products have been sold, and product codes. Using the date and city information, we added air temperature as an additional feature to the dataset.

Weather information is collected from World Whether Online web service by using Past Weather REST API method which allows us to access weather conditions from 1st July 2008 up until the present time. The API returns weather elements such as temperature, weather description, wind speed etc. (World Weather Online - Weather API, 2015). A sample form of data is represented in Table 4.1.

Table 4.1. Sample Data

Date	Sales Amount	Temperature	Price	City	Product Code
2010-01-06 00:00:00.000	1	10	536,114244	KONYA	A
2010-01-06 00:00:00.000	1	10	554,135625	KONYA	B
2010-01-11 00:00:00.000	1	13	1700,440678	ISTANBUL	C
2010-01-13 00:00:00.000	2	9	603,9029117	ISTANBUL	D
2010-01-13 00:00:00.000	2	9	769,6376662	ISTANBUL	E
2010-01-13 00:00:00.000	13	9	536,114244	KONYA	A
2010-01-13 00:00:00.000	6	6	536,114244	KUTAHYA	A
2010-01-27 00:00:00.000	1	2	1231,418677	ISTANBUL	F

There are different types of products for sale in this company namely refrigerators, deep freezers and freezers. These products can be categorized as commercial products that are business-to-consumer goods in shopping subcategory which is a type of product that requires research of customers and comparison of brands. (Figure 4.1). They are also end products in that they are assembled and/or processed with raw materials (Global Text Project, 2010).

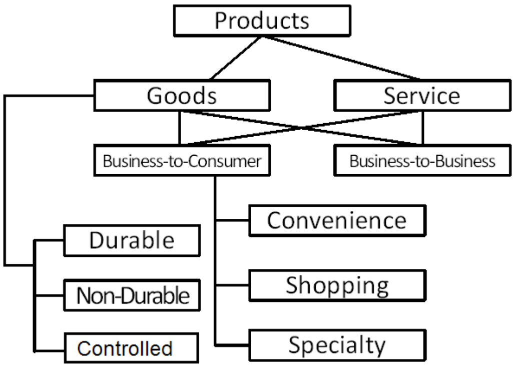


Figure 4.1. Classification of Products
(Source: Global Text Project, 2010)

In the context of exploratory data analysis, we calculated some summary statistics to describe data. One such summary statistic is the total product sales amount sorted according to city (Table 4.2). The city of Istanbul is the place with the highest number of product sales.

Table 4.2. Total Product Sales Sorted According to City

City	Sales Amount
ISTANBUL	22108
KONYA	17054
KUTAHYA	11695
IZMIR	8139
ANTALYA	6207
ANKARA	6132
AFYONKARAHISAR	6088
The Other Cities	Under 5000

The company has complex product codes in that simple product codes are used to represent original product codes namely A, B, C etc.

The most popular product is the product A with the maximum amount of sales. Product codes and their sales amounts are listed in Table 4.3.

Table 4.3. Individual Product Sales Amounts in All Cities

Product Code	Sales Amount
A	134247
G	12659
H	7804
I	6355
J	5342
K	4751
B	3829
L	3348
D	3090
M	1251
N	1189
O	1121
The Other Products	Under 1000

In this study, problem is defined as modelling the demand prediction of the most popular product which is the product “A” in Istanbul city which is the place with the highest number of product sales for the commercial company.

4.1.2. Visualizing Data

Table 4.4 represents total sales of product "A" between years 2010 and 2014. Sales are grouped by quarters. First quarter (January, February, March) sales of product "A" in the year of 2010 is 577 unit. And sum of first quarter's sales between years 2010 and 2014 is 17860 unit and so on.

Table 4.4. Total Product "A" sales between years 2010 and 2014 grouped by quarters

Quarters	2010 Sales Amount	2011 Sales Amount	2012 Sales Amount	2013 Sales Amount	2014 Sales Amount	Total Sales
1	577	2029	2837	6341	6076	17860
2	1451	2306	3711	7444	7541	22453
3	4737	8647	13930	17694	15756	60764
4	4040	6192	9368	8747	4823	33170

Bar graphs are helpful to compare specific categories in their own right. Relative sizes of quarter sales between years 2010 and 2014 are represented in Figure 4.2. The minimum sales are in the first quarter and the maximum sales are in the third quarter. Also Figure 4.2 shows sales increases and decreases while passing from one quarter to the other.

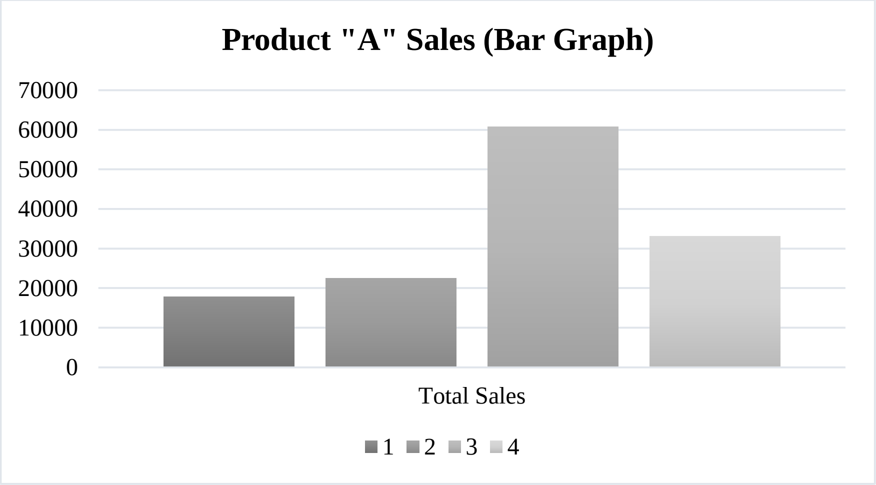


Figure 4.2. Sum of quarter sales between years 2010 and 2014

Table 4.2 and Table 4.3 represent that, product A and the city of Istanbul contain highest number of sales. In the following graphs, we focus on the detailed statistics regarding product A's sales information in Istanbul.

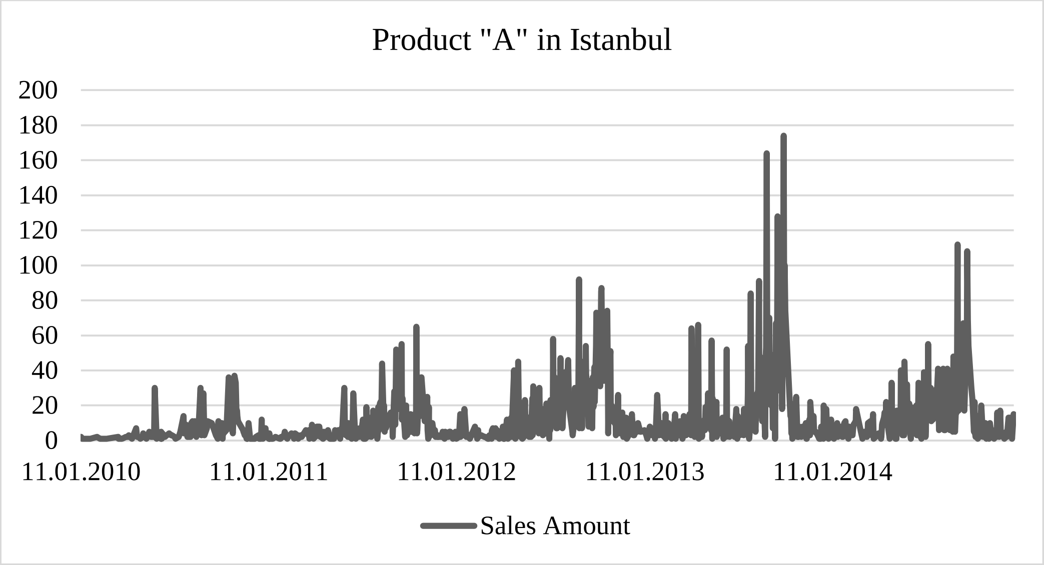


Figure 4.3. Sales amount graph for product “A” in Istanbul city between years 2010 and 2014

Figure 4.3 shows the sales between dates “01.01.2010” and “31.12.2014” for product “A” in the city of Istanbul. The data in Figure 4.3 exhibit periodic increases and decreases. Although this periodicity observed in years, the individual characteristics of rise and falls (being sharp/soft) differ. Thus, it is useful to investigate the details of the patterns for each year.

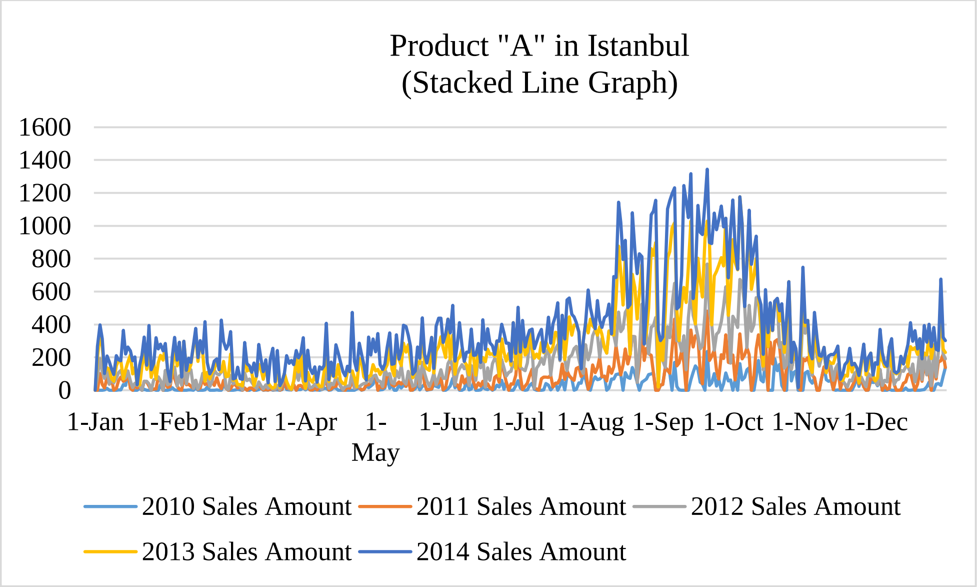


Figure 4.4. Stacked line graph for Istanbul data for product “A” and year in 2010 to 2014

The stacked line graphs are quite helpful to compare/contrast sales behavior according to different years.

Figure 4.4 exhibits data patterns for product “A”. In summertime sales increase for each year. In other seasons sales decrease. That pattern may be explained by the air temperature. Air temperature can be added to the model as a predictor variable which effects the response variable, sales amount.

4.1.3. Predicting the Response Variable Distribution

In this section, firstly the summary statistics of the response variable are presented then goodness of fit test study is done for investigating the probability distribution of the response variable.

4.1.3.1. Summary Statistics of Data

After visualizing the sales amounts, now we do research on it for each year separately. Table 4.5 gives summary statistics for each year.

Table 4.5. Summary Statistics of sales between years 2010 and 2014 for product “A” in Istanbul

Metrics	2010 Sales	2011 Sales	2012 Sales	2013 Sales	2014 Sales
N (Count of Data)	108	196	210	219	207
Mean	6,190	9,120	17,230	18,090	15,530
Std. Error of Mean	0,233	0,776	1,377	1,846	1,211
Median	3,5	5	10	8	10
Mode	1	2	3	1	2
Std. Deviation	7,554	10,863	19,950	27,316	17,424
Variance	57,068	117,995	398,017	746,148	303,590
Skewness	2,471	2,352	2,265	3,140	2,432
Minimum	1	1	1	1	1
Maximum	37	65	132	182	112
Coefficient of Variation (CV)	1,220	1,190	1,158	1,510	1,122

From the sales amount statistics we see that the means are different from the medians implying that the distributions are not symmetric (Law, 1991). For a non-symmetric distribution, skewness (v) measure tells us the direction of the skewness. If it’s greater than 0 the distribution is skewed to the right else if it’s negative, data points are skewed to the left. The skewness values of our data are always positive which means that yearly sales amount data distributions are skewed to the right. Figure 4.5 visualizes the skewness of the sales data.

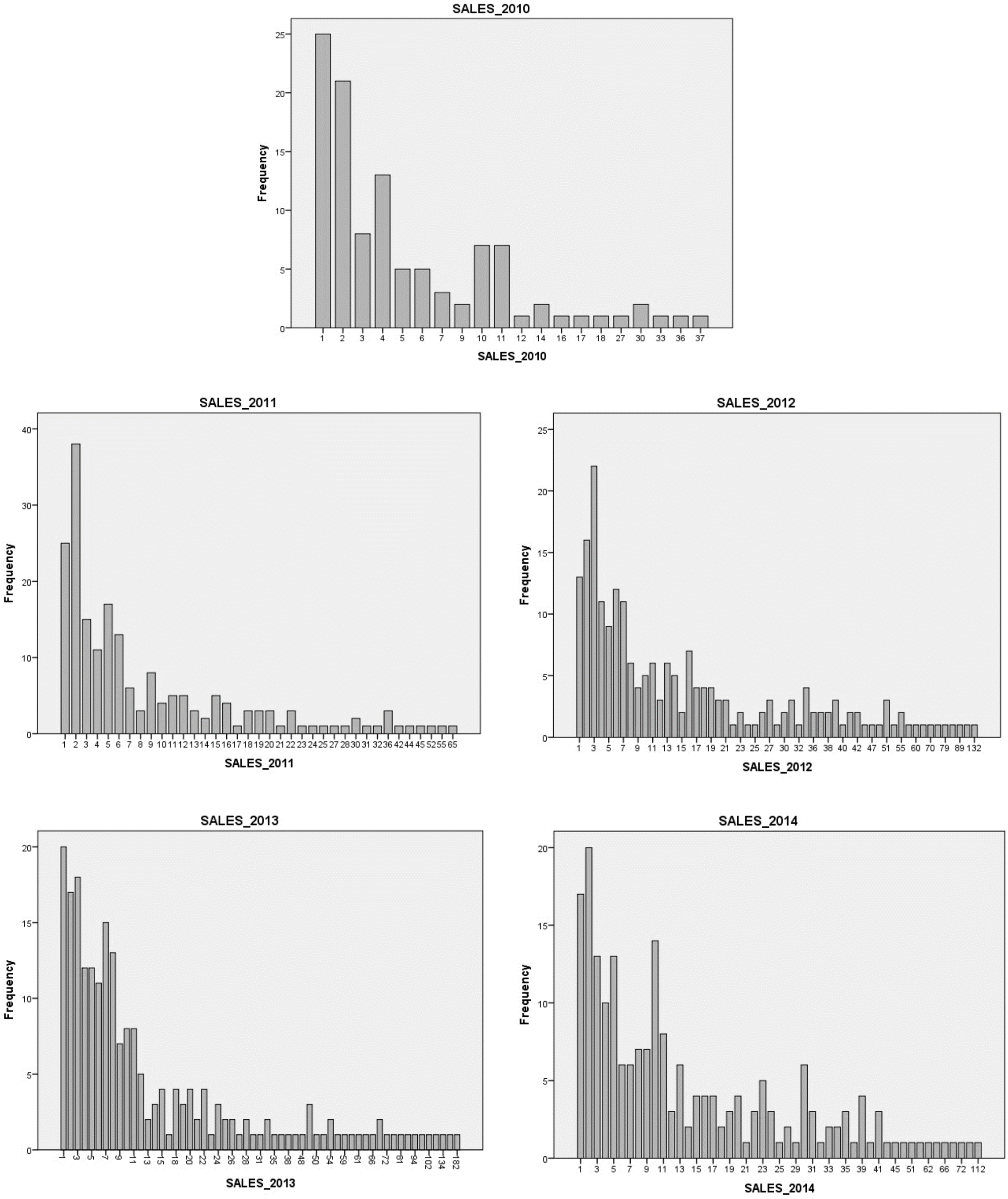


Figure 4.5. Bar Charts of individual sales for each year 2010 to 2014

We see from the Figure 4.5, all graphs are skewed to right. Graph values show individual sales amount in each sales. For example, in year 2010, first bar shows that in 25 days one product has been sold and second bar says in 21 days two products have been sold etc.

The coefficient of variation (CV) can sometimes provide useful information about the form of a continuous distribution. CV being close to 1 implies that the underlying distribution is exponential (Law, 1991). If CV equals to 1 then that means the distribution is the exponential distribution. None of the CV values of our data is equal to 1. That means yearly sales amount data distributions are not the exponential distribution.

4.1.3.2. Goodness of Fit Test of Data

Table 4.6 shows the goodness of fit test results for sales data between years 2010 and 2014. The results are acquired by the EasyFit statistical packed program. Yearly sales data are given to that program as an input to generate these results. The program applies the fit tests Kolmogorov-Smirnov, Anderson-Darling, and Chi-Squared. Chi-Squared technique is a very old technique (Pearson, 1900) that is applicable for continuous or discrete data and also it is asymptotically valid when N goes to infinity (Law, 1991). The results shown in Table 4.6 are generated according to the Chi-Squared fit test.

Table 4.6. Goodness of fit test result for sales data between years 2010 and 2014

Metrics	2010 Sales	2011 Sales	2012 Sales	2013 Sales	2014 Sales
Goodness of Fit	Log-Logistic	Gamma	Gamma (3P)	Lognormal	Gamma
Significance Level	0,05	0,05	0,05	0,05	0,05
P Value	0,21991	0,24049	0,57422	0,66265	0,22956
Reject?	no	no	no	no	no
Parameters	α=1,6499 β=3,5296	α=0,70528 β=12,935	α=0,71381 β=24,254 γ=1,0	σ=1,2341 μ=2,1291	α=0,79408 β=19,553

The null hypothesis in the context of goodness of fit test for the 2010 sales data is:

H_0: There is no difference between Sales2010 data distribution and the theoretical Log Logistic distribution

The alternative one is:

H_1: There is difference between Sales2010 data distribution and the theoretical Log Logistic distribution

If we reject the H_0, we are 21% wrong (p value is 0,21991). So we can say that, we cannot reject the Sales2010 data fitted to Log-Logistic distribution.

If we obtain p-value greater than 0,05, then we can conclude that the distribution of Sales2010 data and the theoretical Log-Logistic distribution are not significantly different. So we say that Sales2010 data fit to Log-Logistic Distribution.

Same hypothesis tests can be applied for the other years 2011, 2012, 2013, 2014 sales data. The other years' sales data have p-value greater than 0,05 so, Sales2011 data fit to Gamma Distribution, Sales2012 data fit to Gamma (3P) Distribution, Sales2013 data fit to Lognormal Distribution, Sales2014 data fit to Gamma Distribution.

4.1.4. Preprocessing on Data

Before fitting the data to a model, discretization studies are done on the continuous predictor variables. The discrete predictor variables are used also model fitting study and the effects of discretization are investigated on the goodness of fit of the statistical models. The forthcoming discretization section presents the categorization of predictor variables. Decomposition of time series section presents the visualization of the response variable by adapting to seasonality.

4.1.4.1. Discretization

Discretization may be explained as transferring the continuous variables into discrete variables with the categorization process. Predictor variables which are temperature, price and date are transferred into categorical variables for investigating the effect of categorization.

Predictor variables are categorized with the help of box plots which show upper and lower quartiles of data. The following figures show box plot diagrams of predictor variables and their quartile boundaries.

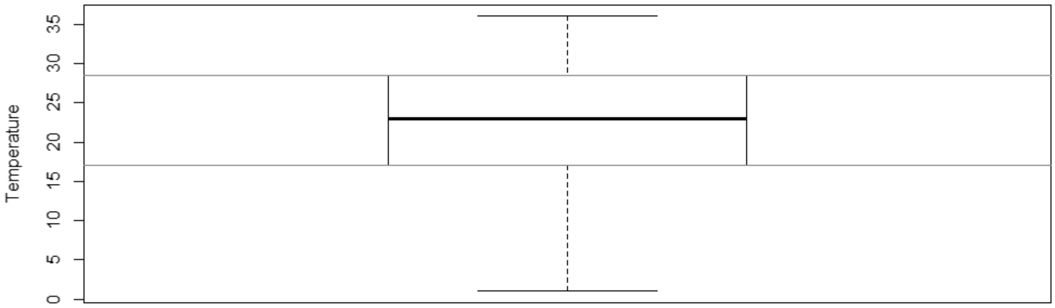


Figure 4.6. Box plot diagram of temperature predictor variable

Figure 4.6 illustrates the box plot diagram of the temperature predictor variable with the upper quartile and lower quartile. The following table presents the boundaries of categories of temperature predictor variable.

Table 4.7. Minimum and maximum value of temperature categories.

Temperature		
Category	**Min**	**Max**
Low Temp.	-	17
Medium Temp.	17	28,5
High Temp	28,5	-

As shown in the Table 4.7, the values under 17 degrees Celsius are categorized as “Low Temperature”, the value between 17 degrees Celsius and 28,5 degrees Celsius are categorized as “Medium Temperature” and the values equal to or above 28,5 degrees Celsius is categorized as “High Temperature”.

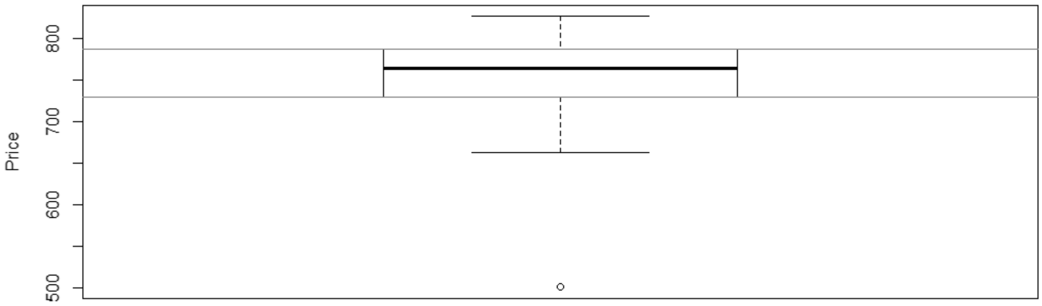


Figure 4.7. Box plot diagram of price predictor variable

Figure 4.7 illustrates the box plot diagram of the price predictor variable with the upper quartile and lower quartile. The following table presents the boundaries of categories of price predictor variable.

Table 4.8. Minimum and maximum value of price categories

Price Predictor Variable		
Category	**Min**	**Max**
Low Price	-	729,1154 (TL)
Medium Price	729,115 (TL)	787,0951 (TL)
High Price	787,095 (TL)	-

As shown in Table 4.8, the values under 729,1154 are categorized as “Low Price”, the values between 729,1154 and 787,0951 are categorized as “Medium Price” and the values that are equal to or above 787,0951 are categorized as “High Price”.

The date predictor variable is categorized with respect to quarters. That is, sales in January, February, and March are categorized as the first quarter. Sales in April, May, and June are categorized as the second quarter. Sales in July, August, and September are categorized as the third quarter. Sales in October, November, and December are categorized as the fourth quarter. These categories can be seen in the following table.

Table 4.9. Categorization of the date predictor variable

Date Predictor Variable	
Category	**Months**
First Quarter	January, February, March
Second Quarter	April, May, June
Third Quarter	July, August, September
Fourth Quarter	October, November, December

4.1.4.2. Decomposition of Time Series (Seasonality)

Time series are deconstructed into notional components with a statistical method which is named the decomposition of time series. The periodicity observed in years shows that the response variables can be investigated in a seasonal manner. The number of sales observations which is greater than or equal to one sales in a day, is shown in Table 4.10. For investigating the response variable in a seasonal manner, observations of response variables must have the same frequency for each year. For this reason, if there is no sales in any day, then the sales observation is assumed to have zero value.

Table 4.10. Number of sales observation in Istanbul city for product “A”

Year	Product Code	Observation Count	City
2010	A	108	Istanbul
2011	A	196	Istanbul
2012	A	210	Istanbul
2013	A	219	Istanbul
2014	A	207	Istanbul

The mathematical representation of the decomposition methodology is defined as follows (Makridakis, 1998).

$$Y_t = f(S_t; T_t; E_t) = S_t + T_t + E_t \tag{4.1}$$

Where,

t is the period,

Y_t is the time series value,

S_t is the seasonal component,

T_t is the trend component,

E_t is the random component,

The following figure illustrates the decomposition of time series. The observed value diagram illustrates the time series value, the trend value diagram illustrates the trend component, the seasonal value diagram illustrates the seasonal component, and the random value diagram illustrates the random component.

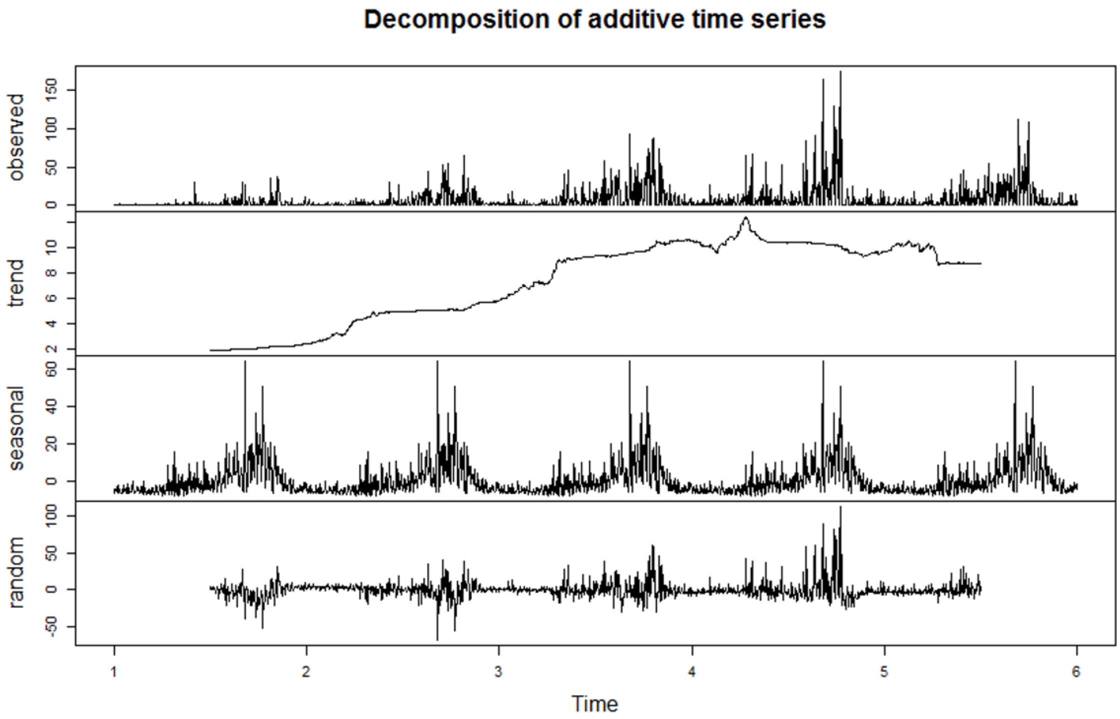


Figure 4.8. Decomposition of time series data for product “A” in Istanbul city

For the sales data, the trend component increases at first three time periods which indicate sales of 2010, 2011, 2012 years data. The trend component does not increase or decrease at the last two time periods which indicate sales of 2013 and 2014 sales data. The third graph which presents the seasonal component shows the seasonal pattern. The seasonal pattern at the start of the graph is the same until to the end of the graph. So sales date indicates same seasonal characteristic from beginning to the end. The last graph which presents the random component, shows random fluctuations. Fluctuations appear with the same time period with the increase of sales amount. That condition can be interpreted as, more sales causes more fluctuations and if the sales amount is close to the zero value, then fluctuations are close to the zero value also.

4.2. Model Fitting of Data

To fit the yearly sales data with a GLM, the response variable (sales amount) must belong to the exponential family of probability distribution (Kutner, 2004). Thus; first, the probability distributions of response variables are determined for each year, then in this section, GLM fitting study has been performed.

4.2.1. GLM Fitting of Data

Sales2014 data are fitted to GLM. As stated in Table 4.6, the distribution of the response variable was determined as Gamma Distribution. Inverse link function is determined as the link function for GLM fitting.

4.2.1.1. Incremental Addition of Features

Day, temperature, and price are considered as predictor variables while constructing the fitted model to sales data for the year 2014. Predictor variables are added to the model incrementally for understanding their effects on significance model fitting. Adding any predictor variable into the problem increases the number of combinations of the fitted model. The next section discusses the number of combinations for the model fit.

Eight different models can be constructed with the combination of three predictor variables which are day, temperature and price. These seven different models are referred in Table 4.11. The eighth model is constructed with the empty model which has not any predictor variable. Only the intercept value is added for constructing it.

Table 4.11 presents GLM fitting results for sales data for the year 2014 but, Table 4.12 presents GLM fitting results for categorical sales data for the year 2014. Categorical sales data means predictor variables are categorical which is described in the discretization section previously.

GLM results and the effect of predictor variables are evaluated in the next section with predictor variable either categorical or not.

Table 4.11. GLM fitting results for the sales data for the year 2014

Sales2014 (GLM)								
Coefficients of Predictor Variables	Day	-0,0003	0	0	-0,0003	0	-0,0002	-0,0003
	Temperature	0	0,0039	0	-0,0041	-0,0041	0	-0,0041
	Price	0	0	-0,0004	0	-0,0004	0,0000	-0,0001
	Intercept	0,0957	0,1642	0,3874	0,2082	0,4974	0,1188	0,2698
MSE		288,6811	267,5052	287,1445	250,5749	259,6151	288,4999	249,1760
AIC		1539,5130	1515,8720	1545,0860	1504,9030	1512,0000	1541,4870	1506,7180
Null deviance		240,9493	240,9500	240,9500	240,9500	240,9500	240,9500	240,9500
Degrees of freedom		206	206	206	206	206	206	206
P Value		0,0479	0,0479	0,0479	0,0479	0,0479	0,0479	0,0479
Residual deviance		229,8613	208,3468	235,2086	197,3156	203,2985	229,8365	197,1618
Degrees of freedom		205	205	205	204	204	204	203
P Value		0,1124	0,4218	0,0726	0,6185	0,5007	0,1036	0,6023

4.2.1.1.1. Selecting the Right Predictor Variables with Non-Categorical Data

First, GLM fitting results are interpreted with non-categorical sales data for the year 2014. The results are presented in Table 4.11.

Null deviance shows that how well the response variable is predicted by the model with adding intercept to model but without the other predictor variables. Before adding predictor variables, the p value is calculated with null deviance and degrees of freedom which is 0,0479. Because it is under the value 0,05 the value is not significant. That means there is significantly difference between fitted values and observed values.

Deviance is calculated by -2 times the log likelihood ratio of the reduced model compared to the full model. Deviance is used for comparing the quality of fit statistics for two models.

To trace the improvement of fitted model, predictor variables are added to the model one by one. At the first step, only one predictor variable that is “Day” is added to our model. It decreases deviance from 240,95 to 229,86 and also decreases the degrees of freedom to 205.

P value is calculated for testing the H_0 hypothesis which is defined in the context of goodness of fit test. The following hypothesis is defined for the model fitting with the given data set.

H_0: There is no difference between observed values and fitted values

H_1: There is a difference between observed values and fitted values

The p value is 0,1124. Because it is over the value 0,05; the test hypothesis cannot be rejected. That means there is no significant difference between the observed values and the fitted values. And also, the model is well fitted to data with the "Day" predictor variable.

Value of MSE (mean square error) is 288,68 that we will use to compare the quality of the estimator. Value of AIC (The Akaike's Information Criterion) is 1539,51 that will also be used to compare the quality of the fitted model.

When we add only "Temperature" predictor variable to the model, deviance is decreased to 208,35. P value is calculated as 0,4218. Because it is over the value 0,05, the test hypothesis cannot be rejected. That means there is no significant difference between observed the values and the fitted values. And also, the model is well fitted to data with the "Temperature" predictor variable. Value of MSE is calculated as 267,50. AIC is calculated as 1515,87. MSE and AIC values are better than the previous model which is constructed with only "Day" predictor variable.

When we add only "Price" predictor variable to the model, deviance is decreased to 235,21. P value is calculated as 0,0726. Because it is over the value 0,05; the test hypothesis cannot be rejected. That means there is no significant difference between the observed values and the fitted values. And also, the model is well fitted to data with the "Price" predictor variable. Value of MSE is calculated as 287,14. AIC is calculated as 1545,08. However, MSE and AIC values give a worse result than the previous model which is constructed with only "Temperature" predictor variable.

When we add two predictor variables which are "Day" and "Temperature" to the model, the residual deviance decreases to 197,32 and the degrees of freedom decreases to 204. The p value is calculated as 0,6185. Because it is over the value 0,05; the test hypothesis cannot be rejected. That means there is no significant difference between the observed values and the fitted values. And also, the model is well fitted to data with the "Day" and "Temperature" predictor variables. MSE is calculated as 250,57 and AIC is calculated as 1504,90. When we look at these values, the result is much better than the previous conditions.

When we add two predictor variables which are "Temperature" and "Price" to the model, the residual deviance is calculated as 203,30 and the degrees of freedom decreases to 204. The p value is calculated as 0,5007. Because it is over the value 0,05; the test hypothesis cannot be rejected. That means there is no significant difference between the observed values and the fitted values. And also, the model is well fitted to data with the "Temperature" and "Price" predictor variables. MSE is calculated as 259,61 and AIC is calculated as 1512. When we look at these values, the result is not the best condition.

When we add two predictor variables which are "Day" and "Price" to the model, the residual deviance is calculated as 229,84 and the degrees of freedom decreases to 204. The p value is calculated as 0,1036. Because it is over the value 0,05; the test hypothesis cannot be rejected. That means there is no significant difference between the observed values and the fitted values. And also, the model is well fitted to data with the "Day" and "Price" predictor variables. MSE calculated as 288,49 and AIC calculated as 1541,48. When we look at these values, the result is also not the best condition.

When we add three predictor variables which are "Day", "Temperature" and "Price" to the model, the residual deviance decreases to 197,16 and the degrees of freedom decreases to 203. The p value is calculated as 0,6023. Because it is over the value 0,05; the test hypothesis cannot be rejected. That means there is no significant difference between the observed values and the fitted values. And also, the model is well fitted to data with the "Day", "Temperature" and "Price" predictor variables. MSE is calculated as 259,17 and AIC is calculated as 1506,71. When we look at these values, the result is almost the same with the model which is constructed with "Day" and "Temperature" predictor variable. And also, the result is better than the other conditions.

As a result, the model which is constructed with "Day" and "Temperature" gives the best result when we consider p values and AIC metrics comparatively. So, "Day" and "Temperature" are added to the model as predictor variables. Price is not added to the model.

To show the effects of adding the right predictor variable to the model, GLM fitting results are visualized in Figure 4.9. The worst model which has the largest AIC and MSE metric values and which is constructed with only "Price" predictor variable is visualized against the best model which has the smallest AIC metric and which is constructed with "Day" and "Temperature" predictor variable.

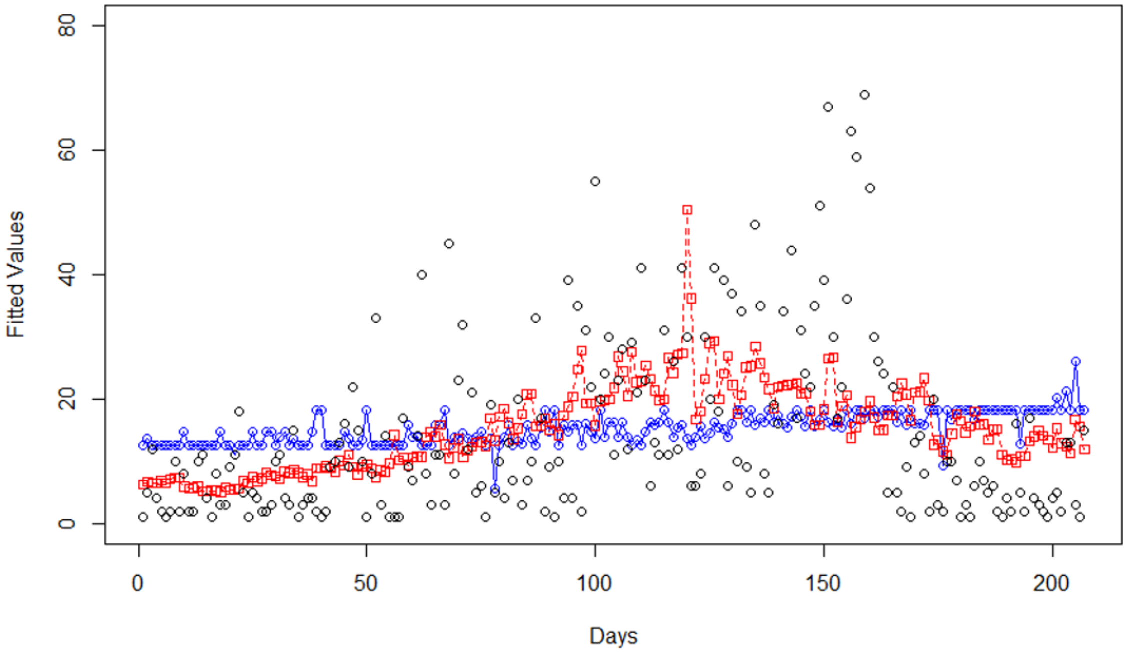


Figure 4.9. Visual representation of the GLM result

The blue curved line in Figure 4.9 shows the GLM fitting result while adding only one predictor variable which is the “Price” variable, to the model. After including the two predictor variables which are “Day” and “Temperature” variable, to the model, the red rectangular pointed line is formed. Real sales values in the year of 2014 are represented in Figure 4.9 as the black rounded points. As a result, the GLM fitted model that has two predictor variables, shows a better result than the GLM fitted model that has one predictor variable.

GLM fitted model which is constructed with the inverse link function and “Day” and “Temperature” predictor variables is presented below. Coefficients of the predictor variables which are “Day” and “Temperature” come from Table 4.11.

$$y = \frac{1}{0{,}2082 - 0{,}0003 * Day - 0{,}0041 * Temperature} \tag{4.2}$$

To show the effect of predictor variables on the response variable, a random point from input variables is chosen (Day, Temperature) = (15, 3).

First “Day” predictor variable is hold fixed at value 15 then, “Temperature” value is changed from 3 to 4 degrees Celsius.

$$\hat{y}_1 = \frac{1}{0,2082 - 0,0003 * 15 - 0,0041 * 3} = 5,2246 \tag{4.3}$$

$$\hat{y}_2 = \frac{1}{0,2082 - 0,0003 * 15 - 0,0041 * 4} = 5,3390 \tag{4.4}$$

As a result of calculations, by changing the “Temperature” from 3 to 4 degrees Celsius when the “Day” is fixed at 15, causes a sales predicton increase from 5,2246 to 5,3390.

Second “Temperature” predictor variable is hold fixed at value 3 degrees Celsius then, “Day” value is changed from 15 to 16.

$$\hat{y}_3 = \frac{1}{0,2082 - 0,0003 * 15 - 0,0041 * 3} = 5,2246 \tag{4.5}$$

$$\hat{y}_4 = \frac{1}{0,2082 - 0,0003 * 16 - 0,0041 * 3} = 5,2328 \tag{4.6}$$

As a result of calculations, by changing the “Day” from 15 to 16 when the “Temperature” is fixed at 3 degrees Celsius, sales prediction increases from 5,2246 to 5,2328.

4.2.1.1.2. Selecting the Right Predictor Variables with Categorical Data

Firstly GLM fitting results were interpreted with non-categorical sales data for the year 2014. Now, GLM fitting results are interpreted with categorical sales data for the year 2014. The results are presented in Table 4.12.

Table 4.12. GLM fitting results for categorical sales data for the year 2014

Categorical Sales2014 (GLM)								
Coefficients of Predictor Variables	Quarter1	0	0	0	0	0	0	0
	Quarter2	-0,0863	0	0	-0,0670	0	-0,0837	-0,0687
	Quarter3	-0,1378	0	0	-0,1209	0	-0,1317	-0,1216
	Quarter4	-0,0753	0	0	-0,0603	0	-0,0752	-0,0666
	Temperature1	0	-0,0867	0	-0,0279	-0,0744	0	-0,0245
	Temperature2	0	-0,0946	0	-0,0136	-0,0762	0	-0,0079
	Temperature3	0	0	0	0	0	0	0
	Price1	0	0	-0,0582	0	-0,0325	-0,0118	-0,0089
	Price2	0	0	-0,0298	0	-0,0148	-0,0010	0,0057
	Price3	0	0	0	0	0	0	0
	Intercept	0,1018	0,1444	0,1077	0,1791	0,1547	0,1785	0,1812
MSE		220,0774	263,6128	267,4917	212,0454	253,8025	220,1729	213,7004
AIC		1482,3530	1517,6320	1530,1930	1480,9990	1514,3340	1484,5400	1482,3370
Null deviance		240,9493	240,9493	240,9493	240,9493	240,9493	240,9493	240,9493
Degrees of freedom		206	206	206	206	206	206	206
P Value		0,0479	0,0479	0,0479	0,0479	0,0479	0,0479	0,0479
Residual deviance		177,8424	208,1400	219,3276	173,8299	201,8705	176,4747	171,8660
Degrees of freedom		203	204	204	201	201	201	199
P Value		0,8981	0,4065	0,2196	0,9174	0,4893	0,8931	0,9183

The empty model of the categorical data which has not any predictor variable is the same with non-categorical data. Because, only predictor variables are categorized for model fitting. So p value is calculated for the empty model with null deviance and degrees of freedom which is 0,0473. Because of it is under the value 0,05; the value is not significant. That means there is a significant difference between the fitted values and the observed values.

Again the improvement of fitted model is traced with adding predictor variables one by one. At the first step, only one predictor variable that is “Quarter” is added to the model. It decreases deviance from 240,95 to 177,84 and also decreases the degrees of freedom to 203.

P value is calculated for testing the H_0 hypothesis which is defined for in the context of goodness of fit test. The following hypothesis is defined for the model fitting with the given data set.

H_0: There is no difference between the observed values and the fitted values

H_1: There is difference between the observed values and the fitted values

The p value is 0,8981. Because it is over the value 0,05; the test hypothesis cannot be rejected. That means there is no significant difference between the observed values and the fitted values. And also, the model is well fitted to data with the "Quarter" predictor variable. The value of MSE is calculated as 220,07 and the value of AIC is calculated as 1482,35.

When we add only "Temperature" predictor variable to the model, deviance is decreased to 208,14. P value is calculated as 0,4065. Because it is over the value 0,05; the test hypothesis cannot be rejected. That means there is no significant difference between the observed values and the fitted values. And also, the model is well fitted to data with the "Temperature" predictor variable. The value of MSE is calculated as 263,61. AIC is calculated as 1517,63. MSE and AIC values are better than the previous model which is constructed with only "Quarter" predictor variable.

When we add only "Price" predictor variable to the model, deviance is decreased to 219,33. P value is calculated as 0,2196. Because it is over the value 0,05; the test hypothesis cannot be rejected. That means there is no significant difference between the observed values and the fitted values. And also, the model is well fitted to data with the "Price" predictor variable. The value of MSE is calculated as 267,49. AIC is calculated as 1530,19. MSE and AIC values give a worse result than the previous models.

When we add two predictor variables which are "Quarter" and "Temperature" to the model, the residual deviance decreases to 173,83 and the degrees of freedom decreases to 201. The p value is calculated as 0,9174. Because it is over the value 0,05; the test hypothesis cannot be rejected. That means there is no significant difference between the observed values and the fitted values. And also, the model is well fitted to data with the "Quarter" and "Temperature" predictor variables. MSE is calculated as 212,04 and AIC is calculated as 1480,99. When we look at these values, the result is much better than the previous conditions.

When we add two predictor variables which are "Temperature" and "Price" to the model, the residual deviance decreases to 201,87 and the degrees of freedom decreases to 201. The p value is calculated as 0,48. Because it is over the value 0,05; the test hypothesis cannot be rejected. That means there is no significant difference between the observed values and the fitted values. And also, the model is well fitted to data with the "Temperature" and "Price" predictor variables. MSE calculated as 253,80 and AIC calculated as 1514,33. When we look at these values, the result is not the best condition.

When we add two predictor variables which are “Quarter” and “Price” to the model, the residual deviance decreases to 176,47 and the degrees of freedom decreases to 201. The p value is calculated as 0,89. Because it is over the value 0,05; the test hypothesis cannot be rejected. That means there is no significant difference between the observed values and the fitted values. And also, the model is well fitted to data with the “Quarter” and “Price” predictor variables. MSE calculated as 220,17 and AIC calculated as 1514,33. When we look at these values, the result is also not the best condition.

When we add three predictor variables which are “Quarter”, “Temperature” and “Price” to the model, the residual deviance decreases to 171,87 and the degrees of freedom decreases to 199. The p value is calculated as 0,91. Because it is over the value 0,05; the test hypothesis cannot be rejected. That means there is no significant difference between the observed values and the fitted values. And also, the model is well fitted to data with the “Quarter”, “Temperature” and “Price” predictor variables. MSE is calculated as 213,70 and AIC is calculated as 1482,33. When we look at these values, the result is so close to the model which is constructed with “Quarter” and “Temperature” predictor variables. And also, the result which is constructed with “Quarter” and “Temperature” predictor variables is better than the other conditions.

As a result, the model which is constructed with “Quarter” and “Temperature” gives the best result when we consider MSE values and AIC metrics comparatively. So, “Quarter” and “Temperature” are added to the model as predictor variables. Price is not added to the model. Because adding the “Price” predictor variable does not affect the fitting result so much. So a simpler model which is constructed with two predictor variables is preferred as the fitted model.

To show the effects of adding the right predictor variable to the model, GLM fitting results are visualized in Figure 4.10. The worst model which has the largest AIC and MSE metric and which is constructed with only the “Price” predictor variable is visualized against the best model which has the smallest AIC metric and which is constructed with “Quarter” and “Temperature” predictor variables.

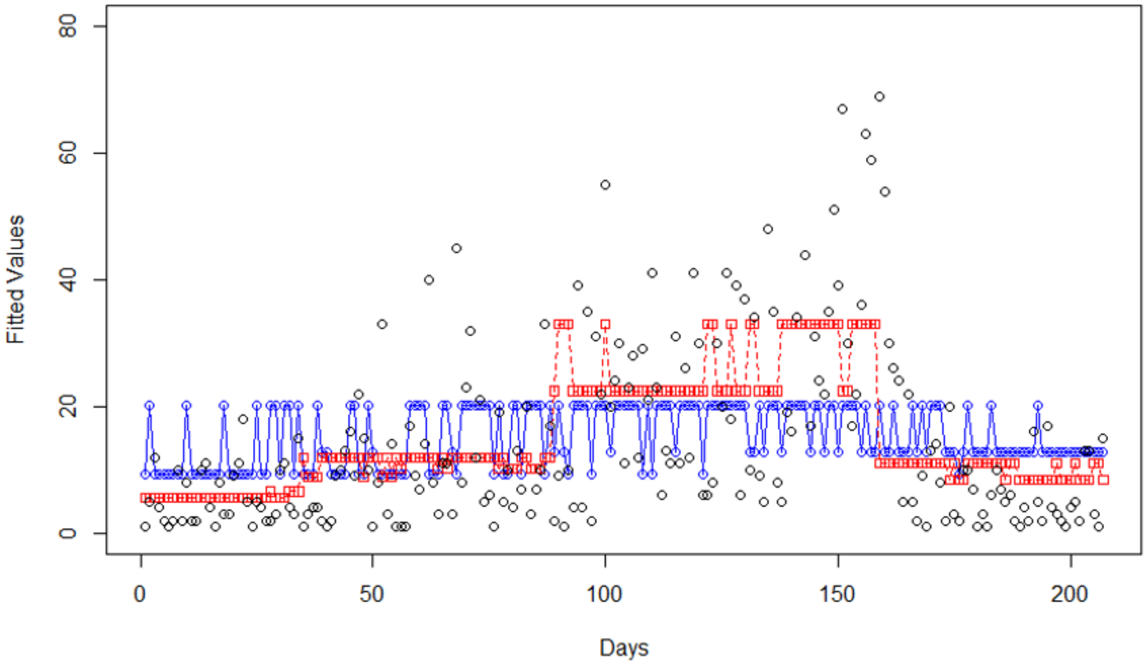


Figure 4.10. Visual representing of GLM result with categorical predictors

The blue curved line in Figure 4.10 shows the GLM fitting result while adding only one predictor variable which is the "Price" variable, to the model. The red rectangular pointed line presents the model fitted values which are constructed with the "Quarter" and "Temperature" predictor variables. Real sales values in the year of 2014 are represented in the Figure 4.10 as the black rounded points. As a result, second GLM fitted model that has two predictor variables shows a better result than the first GLM fitted model that has one predictor variable.

GLM fitted model which is constructed with the inverse link function and "Quarter" and "Temperature" predictor variables is presented below. Coefficients of the predictor variables which are "Quarter" and "Temperature" come from Table 4.12.

$$y = \frac{1}{\begin{array}{c}0{,}1790 - 0{,}0670 * \text{Quarter2} - 0{,}1209 * \text{Quarter3} \\ -0{,}0603 * \text{Quarter4} - 0{,}0279 * Temperature1 \\ -0{,}0136 * Temperature2\end{array}} \qquad (4.7)$$

To show the effect of predictor variables on the response variable a random point from input variables is chosen (Quarter, Temperature) = (1, Low Temperature).

First, “Quarter” predictor variable is hold fixed at value 1 then, “Temperature” value is varied from “Temperature1” to “Temperature2” which means low temperature to medium temperature.

$$\hat{y}_1 = \frac{1}{0{,}1790 - 0{,}0279} = 6{,}6181 \tag{4.8}$$

$$\hat{y}_2 = \frac{1}{0{,}1790 - 0{,}0136} = 6{,}0459 \tag{4.9}$$

As a result of calculations, by changing the “Temperature” from “Temperature1” to “Temperature2” when the “Quarter” is fixed at 1, sales prediction decreases from 6,6181 to 6,0459.

Second, “Temperature” predictor variable is hold fixed at value “Temperature1” then, “Quarter” value is changed from 1 to 2.

$$\hat{y}_3 = \frac{1}{0{,}1790 - 0{,}0279} = 6{,}6181 \tag{4.10}$$

$$\hat{y}_4 = \frac{1}{0{,}1790 - 0{,}0670 - 0{,}0279} = 11{,}8483 \tag{4.11}$$

As a of from calculations, by changing the “Quarter” from 1 to 2 when the “Temperature” is fixed at “Temperature1”, sales prediction increases from 6,6181 to 11,8483.

To show the effects of categorizing the predictor variables, two models of GLM fitting results are visualized in Figure 4.11 and compared in Table 4.13.

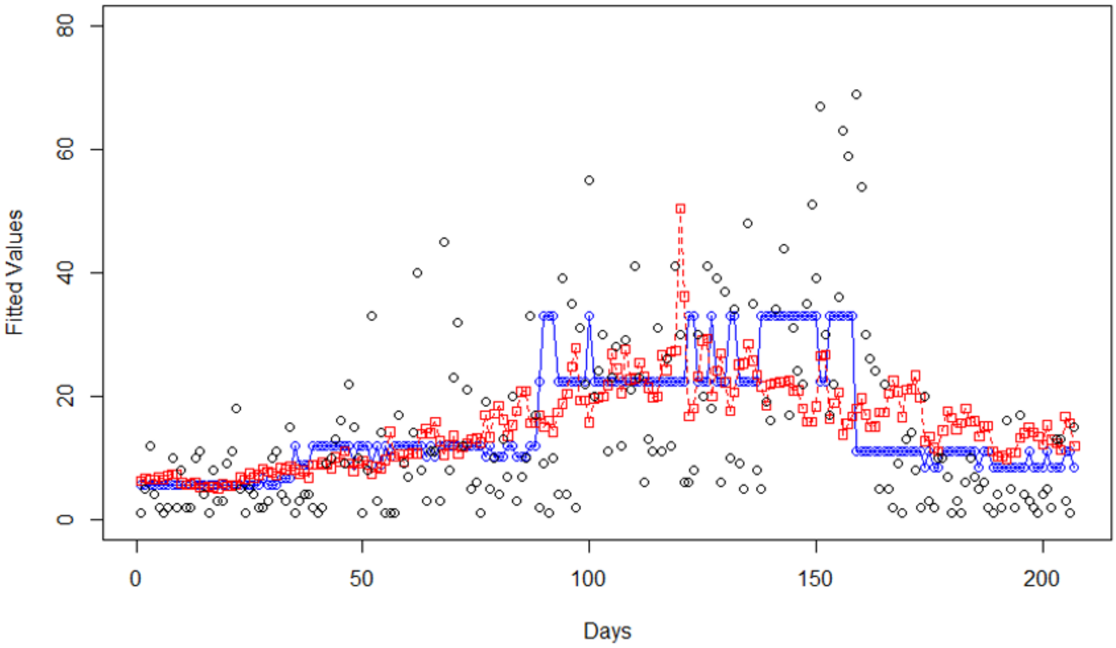


Figure 4.11. Visual representation of GLM result with non-categorical predictors and categorical predictors

One of the models which is constructed with the categorized predictor variables is shown with the blue curved line. The red rectangular pointed line presents the model fitted values which are constructed with the non-categorical predictors. Real sales values in the year of 2014 are represented as black rounded points. As seen from the figure, categorical predictor variables causes similar level predictions. For example, at most 12 level prediction can occur because four level quarter and three level temperature produces at most 12 combination predictions. On the other hand, non-categorical predictor variables cause many level predictions because too many combinations occur with the different temperature and day values.

Another inference from the figure is the fitting quality of predictor variables. Categorical and non-categorical predictor variables produce similar trend prediction but predictions from non-categorical predictors are slightly more close to the observed values.

Table 4.13. Comparison of the GLM result with categorical predictors and non-categorical predictors

Metrics	Predictor Variables	
	Not Categorical	Categorical
	Days And Temperature	Quarter and Temperature
MSE	250,5749	212,0454
AIC	1504,903	1480,999
Null deviance	240,9500	240,9493
Degrees of freedom	206	206
P Value	0,0479	0,0479
Residual deviance	197,3200	173,8299
Degrees of freedom	204	201
P Value	0,6185	0,9174

Table 4.13 represents the comparison of GLM results with categorical and non-categorical predictors. The model which is constructed with the categorical independent variables gives better results.

The difference between two models which are constructed with categorical and non-categorical predictor variables can be explained by the effect of the “Quarter” predictor variable. “Quarter” variable is more informative about the sales amount when compared to the “Day” predictor variable. Because “Day” predictor variable values are only series of numbers which implies time series. On the other hand, “Quarter” predictor variable values gives information about the seasons. In data definition section we see the effects of quarter changes on sales amounts. The change of “Quarter” predictor variable value from Quarter1 to Quarter2 refers season changes and also sales amount changes.

The difference between two models which are constructed with categorical and non-categorical predictor variables, can be explained by losing precision while categorizing predictor variables. The reason of categorizing the continuous predictor variable is simplicity and the easy interpretation of results. However, categorization causes a considerable loss of power and residual confounding. The use of categorized data induces a serious bias at the cut point (Royston, 2005). On the other hand, categorization provides reduced variance on the categorized variable (Cohen, 1995). For example, air temperature causes so much variance on the response variable. If the temperature levels are defined systematically then the particular range of temperature are covered successfully. For our investigated data, categorization provides more accurate prediction than the non-categorical data.

4.2.2. Ordinary Least Squares Estimator versus GLM

Unknown parameters in a linear regression model is estimated with ordinary least squares (OLS) method. The goal of the OLS method is to minimize the difference between the observed values and the predicted values through to use of the linear regression model. The assumption of OLS is that errors are normally distributed. That means if the GLM fits the model with the Gaussian distribution with the identity link function then, the data are fitted to linear regression model and parameters are estimated with OLS method.

Firstly, OLS fitting results are interpreted with the non-categorical sales data for the year 2014. The results are presented in Table 4.14.

Table 4.14. OLS fitting results for 2014 sales data

Sales2014 (OLS)								
Coefficients of Predictor Variables	Day	0,0590	0	0	0,0378	0	0,0563	0,0337
	Temperature	0	0,7492	0	0,6746	0,7215	0	0,6756
	Price	0	0	0,0754	0	0,0518	0,0084	0,0133
	Intercept	9,1302	-1,3848	-41,8914	-3,6572	-40,0194	3,0027	-13,3143
MSE		277,5722	258,1574	284,7068	253,3723	255,7137	277,5303	253,2688
AIC		1758,0390	1743,0290	1763,2930	1741,1570	1743,0610	1760,0080	1743,0720
Null deviance		60025,9100	60025,9100	60025,9100	60025,9100	60025,9100	60025,9100	60025,9100
Degrees of freedom		206	206	206	206	206	206	206
P Value		0,0001	0,6924	0,1551	0,3189	0,1528	0,9317	0,6935
Residual deviance		57457,4400	53438,5700	58934,3100	52448,0700	52932,7400	57448,7800	52426,6500
Degrees of freedom		205	205	205	204	204	204	203
P Value		0,0028	1,0840E-06	0,0527	1,0520E-06	2,6870E-06	0,0114	4,5250E-06
R Square		0,0428	0,1097	0,0182	0,1262	0,1182	0,0429	0,1266

R square statistic shows how well models fit to the given input data. R square takes a value between 0 and 1. If R square statistic takes the value 1 then the model perfectly fits data, on the other hand, if R square statistic takes the value 0 then the model does not fit data. Thus, different linear regression models can be compared with each other by the R square statistic.

In OLS method, deviance is calculated by the sum of squares of regression which is the sum of squares of vertical distances between the predicted data points and the mean of data points.

P value is calculated for testing the H_0 hypothesis which is defined in the context of goodness of fit test. The following hypothesis is defined for the model fitting with the given data set.

H_0: There is no significant relationship between the predictor variables and the fitted model ($\beta=0$)

The alternative one is:

H_1: There is a significant relationship between the predictor variables and the fitted model ($\beta\neq0$)

To trace the improvement of the fitted model, predictor variables are added to the model one by one. At the first step, only one predictor variable that is “Day” is added to our model. It decreases deviance from 60025 to 57457. R square value is calculated as 0,0428. MSE is calculated as 277,57. AIC is calculated as 1758,03. The p value is calculated under 0,05 so H_0 hypothesis is rejected, that means “Day” predictor variable contributes significantly to the fitted model.

When we add only “Temperature” predictor variable to the model, the deviance is decreased to 53438. R square value is calculated as 0,1097. MSE is calculated as 258,15. AIC is calculated as 1743,02. The p value is calculated under 0,05 so H_0 hypothesis is rejected, that means “Temperature” predictor variable contributes significantly to the fitted model. These statistics are better than the previous model fitting result. That means, this model fits to the data better than the previous model and “Temperature” predictor variable helps to predict the response variable better than “Day” predictor variable.

When we add only “Price” predictor variable to the model, the deviance is calculated as 58934. R square value is calculated as 0,0182. MSE is calculated as 284,70. AIC is calculated as 1763,29. The p value is calculated over 0,05 so H_0 hypothesis is not rejected, that means “Price” predictor variable does not contribute significantly to the fitted model. These statistics are worse than the previous model fitting result. That means, this model fits to the data worse than the previous models and the “Price” predictor variable does not help to predict the response variable better than the “Day” or “Temperature” predictor variable do.

When we add two predictor variables which are “Day” and “Temperature” to the model, the deviance is calculated as 52448. R square value is calculated as 0,1262. MSE is calculated as 253,37. AIC is calculated as 1741,15. The p value is calculated under 0,05 so H_0 hypothesis is rejected, that means “Day” and “Temperature” predictor variables contribute significantly to the fitted model. These statistics are better than the previous model fitting results. That means, this model fits to the data better than the previous models. “Day” and “Temperature” predictor variables together help to predict the response variable better than the other models which have only one predictor variable.

When we add two predictor variables which are “Temperature” and “Price” to the model, the deviance is calculated as 52937. R square value is calculated as 0,1182. MSE is calculated as 255,71. AIC is calculated as 1743,65. The p value is calculated under 0,05 so H_0 hypothesis is rejected, that means “Temperature” and “Price” predictor variables contribute significantly to the fitted model. These statistics do not give the best result in comparison to the previous model fitting results. That means “Temperature” and “Price” predictor variables do not give the best fitted model to predict the response variable.

When we add two predictor variables which are “Day” and “Price” to the model, the deviance is calculated as 57448. R square value is calculated as 0,0429. MSE is calculated as 277,53. AIC is calculated as 1760. The p value is calculated under 0,05 so H_0 hypothesis is rejected, that means “Day” and “Price” predictor variables contribute significantly to the fitted model. Also these statistics do not give the best result in comparison to the previous model fitting results. That means “Day” and “Price” predictor variables do not give the best fitted model to predict the response variable.

When we add three predictor variables which are “Day”, “Temperature”, and “Price” to the model, the deviance is calculated as 52426. R square value is calculated as 0,1266. MSE is calculated as 253,26. AIC is calculated as 1743,07. The p value is calculated under 0,05 so H_0 hypothesis is rejected, that means “Day”, “Temperature”, and “Price” predictor variables contribute significantly to the fitted model. MSE and deviance give the best result comparatively to the other models. When we look at these values, the result is almost the same with the model which is constructed with “Day” and “Temperature” predictor variables. And also, the result is better than the other fitted models. However, AIC is not the best value. Because adding new predictor variables make the model more complex and it increases the AIC value.

Linear fitted model which is constructed with “Day” and “Temperature” predictor variables is presented below. Coefficients of the predictor variables come from Table 4.14.

$$y = -3{,}6572 + 0{,}0378 * Day + 0{,}6746 * Temperature \qquad (4.12)$$

To show the effect of predictor variables on the response variable a random point from input variables is chosen (Day, Temperature) = (15, 3).

First, “Day” predictor variable is hold fixed at value 15 then, “Temperature” value is variated 3 to 4 degrees Celsius.

$$\hat{y}_1 = -3{,}6572 + 0{,}0378 * 15 + 0{,}6746 * 3 = -1{,}0664 \qquad (4.13)$$

$$\hat{y}_2 = -3{,}6572 + 0{,}0378 * 15 + 0{,}6746 * 4 = -0{,}3918 \qquad (4.14)$$

As a result of calculations, by changing the “Temperature” from 3 to 4 degrees Celsius when the “Day” is fixed at 15, sales prediction increases from -1,0664 to -0,3918. However, these predictions are not valid for sales data because these are under the zero value. These results show that, linear model does not give good results for sales prediction data.

Second, “Temperature” predictor variable is hold fixed at value 3 degrees Celsius then, “Day” value is variated from 15 to 16.

$$\hat{y}_3 = -3{,}6572 + 0{,}0378 * 15 + 0{,}6746 * 3 = -1{,}0664 \qquad (4.15)$$

$$\hat{y}_4 == -3{,}6572 + 0{,}0378 * 16 + 0{,}6746 * 3 = -1{,}0286 \qquad (4.16)$$

As a result of calculations, by changing the “Day” from 15 to 16 when the “Temperature” is fixed at 3 degrees Celsius, sales prediction increases from -1,0664 to -1,0286.

As a result, the OLS fitted model and the GLM fitted models which are constructed with “Day” and “Temperature” are visualized in Figure 4.12 and compared in Table 4.15.

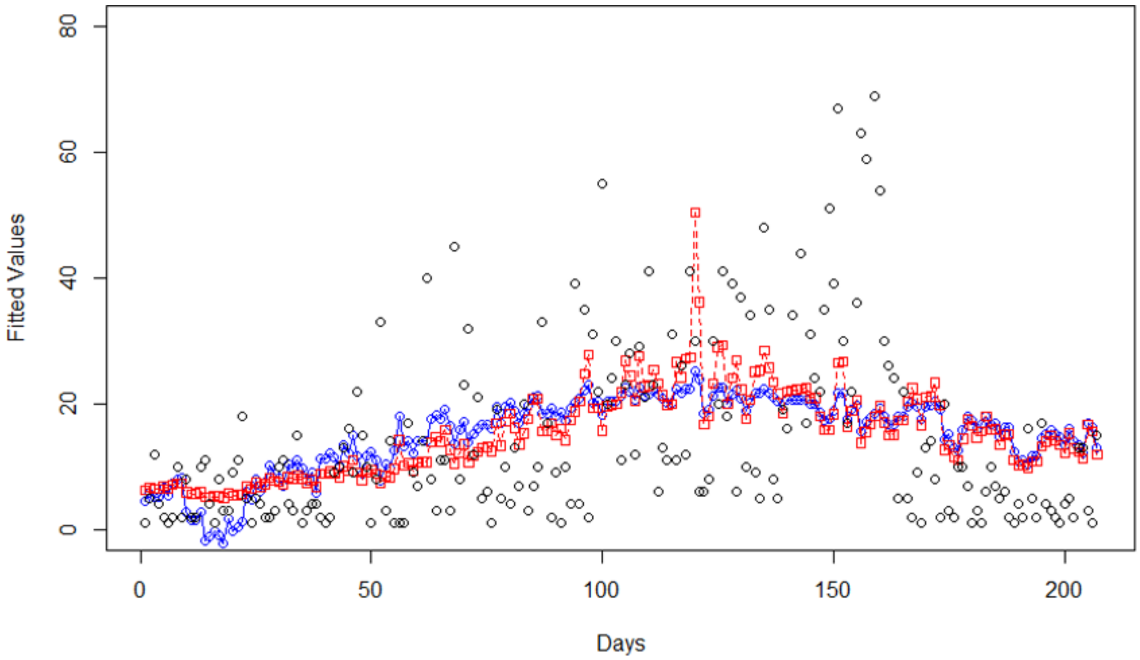


Figure 4.12. Visual representation of GLM and OLS method

The blue curved line in Figure 4.12 shows the OLS fitting result by adding two predictor variables which are “Day” and “Temperature”, to the model. The red rectangular pointed line is the visual of GLM fitting result by adding two predictor variables which are “Day” and “Temperature” variables. Real sales values in the year of 2014 are represented in Figure 4.12 as black rounded points. As seen from the figure, the OLS fitted values and GLM fitted values are similar. The first difference appears between 0 and 50 days interval. Some OLS fitted values are under zero value which is not possible for sales prediction. The second difference appears between 100 and 150 days interval. Some GLM fitted values are peaked. It looks like an outlier prediction but that prediction is also in the observed value range which is between 0 and seventy sales amounts.

Table 4.15. Comparison of GLM and OLS model fitting results

Metrics	**GLM**	**OLS**
	Days And Temperature	
MSE	250,5749	253,3723
AIC	1504,9	1741,157
Null deviance	240,95	60025,91
Residual deviance	197,32	52448,07
Residual / Null Deviance	0,8189	0,8738

Table 4.15 represents the comparison of GLM fitted model and OLS fitted model. GLM fitted model gives better AIC and MSE results. In addition to these, the residual to null deviance fraction gives a better result for GLM fitted model. That means adding "Day" and "Temperature" predictor variables to the GLM fitted model decreases null deviance more than that of the OLS fitted model. As a result, if we compare two models based on AIC, MSE and, the residual to null deviance fraction then, the GLM fitted model represents observed data better than the OLS fitted model.

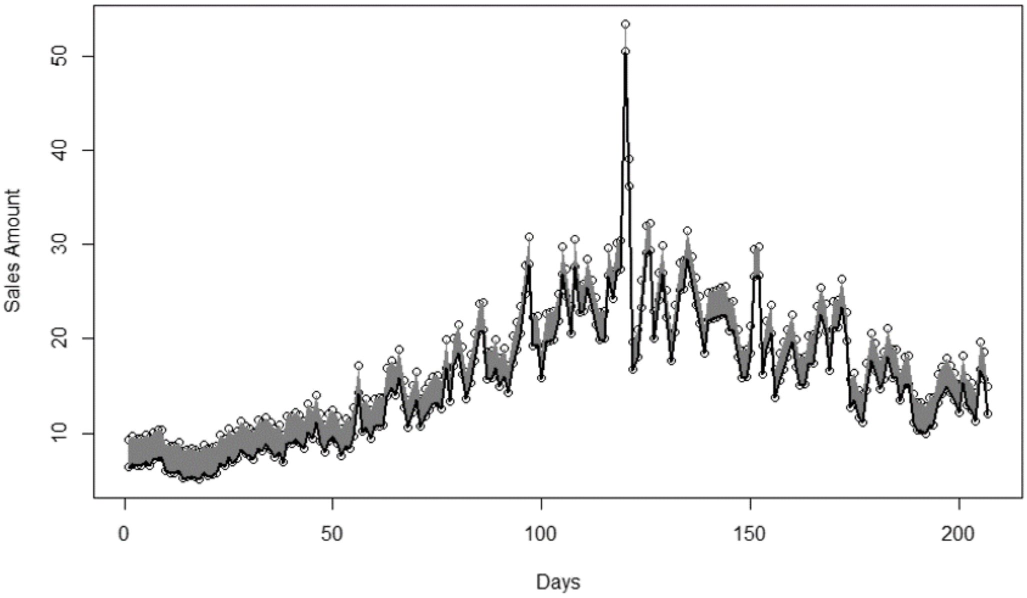


Figure 4.13. Visual representation of GLM variance

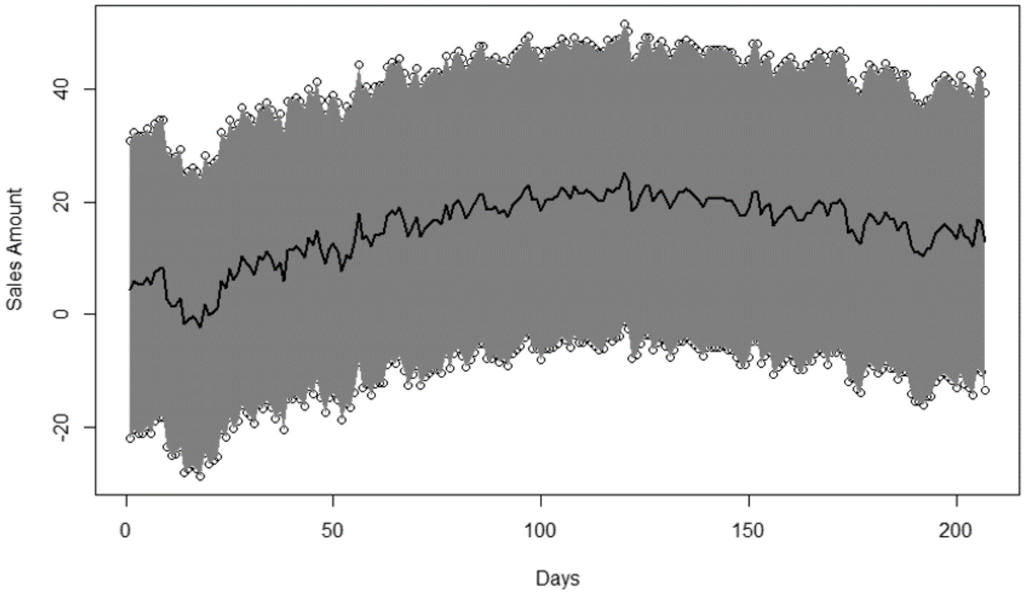


Figure 4.14. Visual representation of OLS variance

Figures 4.13 and 4.14 interpret the variances of the fitted models. The OLS method assumes that the residuals have the same variance that is named homoscedasticity. Constant variance of the OLS method is presented in Figure 4.14. GLM fitted model has non constant variance across an entire range of values that is named heteroscedasticity. Heteroscedasticity of the GLM fitted model is presented in Figure 4.13. Fitting the sales data by the GLM method provides less variance than the OLS method but GLM method causes heteroscedastic model which means different variances for each estimated response variable because of the error term. GLM method does not eliminate the error terms which is included in the regression equation. The transformed error terms are not constant and it causes heteroscedasticity.

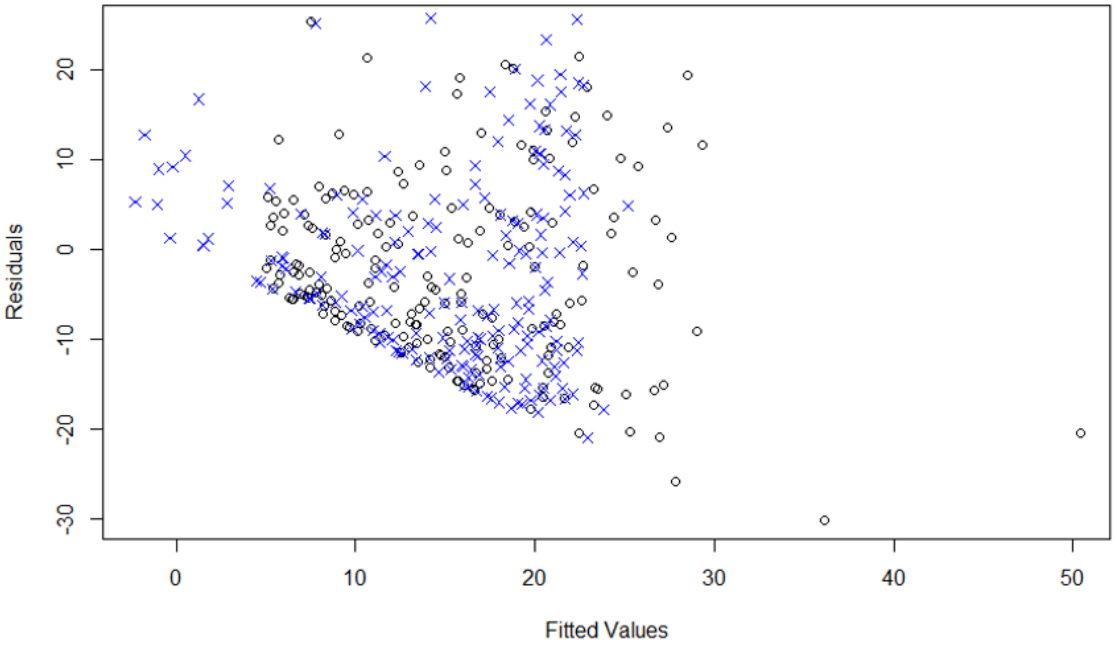


Figure 4.15. Visual representation of OLS residuals and GLM residuals

Figure 4.15 visualizes the residuals for each fitted values. The black o type points represent the GLM residuals for each fitted values and the blue x type points represent the OLS residuals for each fitted values. The residual dispersion and range is similar for each model, however, fitted values range is not similar for each model. Some of the fitted values are under the zero value for the OLS model. That is not feasible for the sales model. Because sales cannot be under the zero value.

First, OLS results are compared with GLM results. Now, OLS method with inverse transformation is used on the response variable. Data transformation work is done for improving the normality of variables, variance reduction and acquiring consistent results, for example eliminating prediction results which are under zero value. Inverse transformation result is compared both OLS and gamma GLM with inverse link function result. The following table presents the OLS model result for inverse response variable.

Table 4.16. OLS - inverse response variable fitting results for sales data for the year 2014

Sales2014 (OLS - Inverse response variable)								
Coefficients of Predictor Variables	**Day**	-0,0005452	0	0	-0,0002059	0	-0,0006993	-0,0003372
	Temperature	0	-0,011226	0	-0,0108194	-0,01124545	0	-0,010786
	Price	0	0	-0,0003	0	3,63098E-05	0,0005	0,0004
	Intercept	0,2817381	0,474455	0,4765465	0,48683296	0,4473677	-0,0806275	0,17986699
MSE		0,0754	0,0693	0,0764	0,0692	0,0693	0,0753	0,0691
AIC		58,4449	41,0395	61,0601	42,6153	43,0359	60,0429	44,3024
Null deviance		15,8341	15,8341	15,8341	15,8341	15,8341	15,8341	15,8341
Degrees of freedom		206	206	206	206	206	206	206
P Value		5,629E-12	1,578E-14	0,3227987	6,8316E-14	0,3311819	0,8889529	0,7471963
Residual deviance		15,6145	14,3552	15,8130	14,3258	14,3550	15,5842	14,3042
Degrees of freedom		205	205	205	204	204	204	203
P Value		0,0909871	7,535E-06	0,6011948	3,678E-05	4,52518E-05	0,197306	1,23E-04
R Square		0,0139	0,0934	0,0013	0,0953	0,0934	0,0158	0,0966

When we compare the OLS model result for inverse response variable with different predictor variables, the fitted model which is constructed with “Day”, “Temperature” and “Price” predictor variables achieves the best result with respect to MSE and R square, residual deviance metrics. On the other hand, the fitted model which is constructed with “Temperature” predictor variable achieves the best result with respect to AIC metric. So we go on to compare the fitted model which is constructed with “Temperature” predictor variable.

If the r squares of sales 2014 results are compared with Table 4.16 results, R Square results are gone to worse. That means transformation is not successful for the response variable. The other metrics which are MSE, AIC, deviances, are not calculated on the same scale with OLS model and OLS model with inverse response variable, so that metrics cannot be used for comparison. On the other hand, if the OLS model with inverse response variable fitted values are transformed to original scale then, MSE metrics and fitted value graphs can be compared. The following table and figure compares OLS methods and GLM method.

Table 4.17. Comparison of GLM, OLS and OLS with inverse response variable

Sales 2014		
Predictor Variables	**Method**	**MSE**
Days and Temperature	Gamma GLM with Inverse Link	250,5749
Days and Temperature	OLS	253,3723
Temperature	OLS with Inverse Response Variable	376,6871

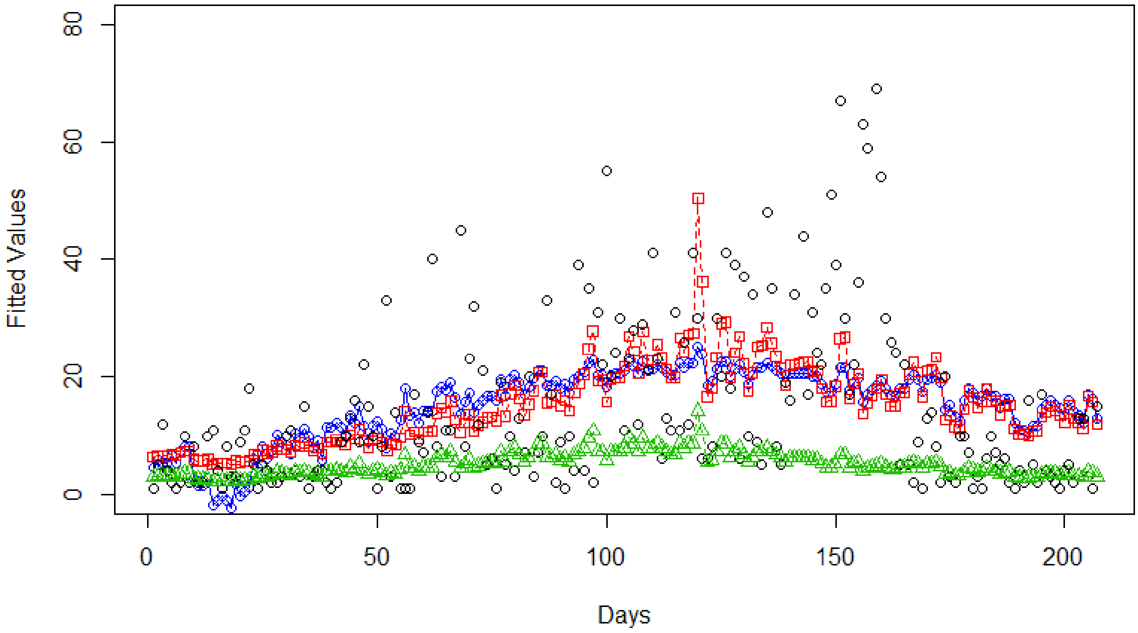


Figure 4.16. Visual representation of OLS and OLS with inverse response variable method

The blue curved line in Figure 4.16 shows the OLS fitting result by adding two predictor variables which are "Day" and "Temperature" variables, to the model. The green triangular pointed line is the visual of the OLS fitting result with inverse response variable by adding the same predictors. The red rectangular pointed line is the visual of the GLM fitting results with the same predictors. Real sales values in the year of 2014 are represented in Figure 4.16 as the black rounded points. As a result, the graphical representation and table comparison shows that OLS fitting result with inverse response variable is not better than the OLS or GLM method.

Firstly, OLS fitting results are interpreted with non-categorical sales data for the year 2014. Now, OLS fitting results are interpreted with categorical sales data for the year 2014. The results are presented in Table 4.18.

Table 4.18. OLS fitting results for categorical sales data for the year 2014

		Categorical Sales2014 (OLS)						
Coefficients of Predictor Variables	Quarter1	0	0	0	0	0	0	0
	Quarter2	5,5350	0	0	4,8480	0	5,1690	5,5800
	Quarter3	20,9800	0	0	23,1850	0	20,1610	24,2340
	Quarter4	4,2940	0	0	2,9380	0	5,5190	5,9820
	Temperature1	0	10,4130	0	2,7410	8,1530	0	1,7190
	Temperature2	0	13,1500	0	-4,9650	9,0700	0	-7,1590
	Temperature3	0	0	0	0	0	0	0
	Price1	0	0	10,9250	0	6,9630	2,3660	2,5240
	Price2	0	0	3,5540	0	1,7100	-1,3950	-3,0760
	Price3	0	0	0	0	0	0	0
	Intercept	5,7060	6,9270	9,2860	5,3830	5,3860	5,0790	4,8350
MSE		220,0774	263,6128	267,4917	212,4030	256,1228	218,0932	208,8401
AIC		1713,9940	1749,3580	1752,3820	1710,6470	1747,3920	1716,1200	1711,1450
Null deviance		60025,9100	60025,9100	60025,9100	60025,9100	60025,9100	60025,9100	60025,9100
Degrees of freedom		206	206	206	206	206	206	206
P Value		0,0275	0,0019	0,0000	0,0371	0,0344	0,0604	0,0700
Residual deviance		45556,0000	54567,8400	55370,7700	43967,4200	53017,4300	45145,2900	43229,9100
Degrees of freedom		203	204	204	201	202	201	199
P Value		3,9435E-12	5,9841E-05	2,6550E-04	2,8693E-12	4,5817E-05	3,6636E-11	8,9711E-12
R Square		0,2411	0,0909	0,0776	0,2675	0,1168	0,2479	0,2798

To trace the improvement of the fitted model, predictor variables are added to the model one by one. At the first step, only one predictor variable that is “Quarter” is added to our model. It decreases deviance from 60025 to 45556. R square value is calculated as 0,2411. MSE is calculated as 220,07. AIC is calculated as 1713,99.

When we add only “Temperature” predictor variable to the model, the deviance is calculated as 54567. R square value is calculated as 0,0909. MSE is calculated as 263,61. AIC is calculated as 1749,35. These statistics are not better than the previous model fitting results. That means, this model fits to the data not better than the previous model and “Temperature” predictor variable does not help to predict response variable better than the “Quarter” predictor variable.

When we add only “Price” predictor variable to the model, the deviance is calculated as 55370. R square value is calculated as 0,0776. MSE is calculated as 267,49. AIC is calculated as 1752,38. These statistics are worse than the previous model fitting result. That means, this model fits to the data worse than the previous models and “Price” predictor variable does not help to predict the response variable better than the “Quarter” or “Temperature” predictor variable do.

When we add two predictor variables which are “Quarter” and “Temperature” to the model, the deviance is calculated as 43967. R square value is calculated as 0,2675. MSE is calculated as 212,40. AIC is calculated as 1710,64. These statistics are better than the previous model fitting results. That means, this model fits to the data better than the previous model fits. The “Quarter” and “Temperature” predictor variables help to predict the response variable better than the other models which have only one predictor variable.

When we add two predictor variables which are “Temperature” and “Price” to the model, the deviance is calculated as 53017. R square value is calculated as 0,1168. MSE is calculated as 256,12. AIC is calculated as 1747,39. These statistics do not give the best results comparatively to the previous model fitting results. That means “Temperature” and “Price” predictor variables do not give the best fitted model to predict the response variable.

When we add two predictor variables which are “Quarter” and “Price” to the model, the deviance is calculated as 45145. R square value is calculated as 0,2479. MSE is calculated as 218,08. AIC is calculated as 1716,12. Also these statistics do not give the best results comparatively to the previous model fitting results. That means “Quarter” and “Price” predictor variables do not give the best fitted model to predict the response variable.

When we add three predictor variables which are “Quarter”, “Temperature” and “Price” to the model, the deviance is calculated as 43229. R square value is calculated as 0,2798. MSE is calculated as 208,84. AIC is calculated as 1711,14. Deviance gives the best result comparatively to the other models. When we look at MSE values, the result is almost the same with the model which is constructed with “Quarter” and “Temperature” predictor variables. And also, the result is better than the other fitted models. However, AIC is not the best value. Because adding a new predictor variable makes model more complex and it increases the AIC value.

Linear fitted model which is constructed with “Quarter” and “Temperature” predictor variables is presented below. Coefficients of the predictor variables come from Table 4.18.

$$y = \begin{matrix} 5{,}3830 + 4{,}8480 * \text{Quarter2} + 23{,}1850 * \text{Quarter3} \\ +2{,}9380 * \text{Quarter4} + \ 2{,}7410 * Temperature1 \\ -4{,}9650 * Temperature2 \end{matrix} \tag{4.17}$$

To show the effect of predictor variables on the response variables a random point from input variable is chosen (Quarter, Temperature) = (1, Low Temperature).

First, "Quarter" predictor variable is hold fixed at value 1 then, "Temperature" value is variated from "Temperature1" to "Temperature2" which means low temperature to medium temperature.

$$\hat{y}_1 = 5{,}3830 + \ 2{,}7410 = 8{,}124 \tag{4.18}$$

$$\hat{y}_2 = 5{,}3830 - 4{,}9650 = 0{,}418 \tag{4.19}$$

As a result of calculations, by changing the "Temperature" from "Temperature1" to "Temperature2" when the "Quarter" is fixed at 1, sales prediction decreases from 8,124 to 0,418.

Second, "Temperature" predictor variable is hold fixed at value "Temperature1" then, "Quarter" value is variated from 1 to 2.

$$\hat{y}_3 = 5{,}3830 + \ 2{,}7410 = 8{,}124 \tag{4.20}$$

$$\hat{y}_4 = 5{,}3830 + 23{,}1850 + \ 2{,}7410 = 31{,}309 \tag{4.21}$$

As a result of calculations, by changing the "Quarter" from 1 to 2 when the "Temperature" is fixed at "Temperature1", sales prediction increases from 8,124 to 31,309.

As a result, the OLS fitted model and the GLM fitted model which are constructed with "Quarter" and "Temperature" are visualized in Figure 4.17 and compared in Table 4.19.

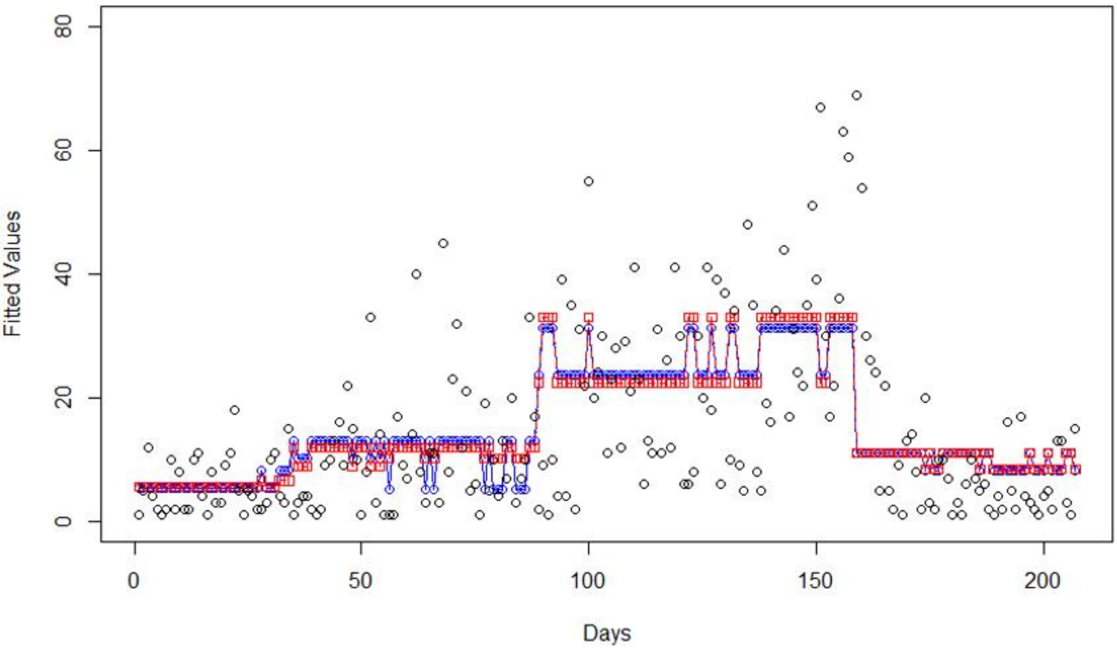


Figure 4.17. Visual representation of GLM and OLS method with categorical response variables

The blue curved line in Figure 4.17 shows the OLS fitting result by adding two predictor variables which are "Quarter" and "Temperature" variable to the model. The red rectangular pointed line is the visual of GLM fitting result by adding two predictor variables which are "Quarter" and "Temperature". Real sales values in the year of 2014 are represented in the Figure 4.17 as the black rounded points. The predicted results of GLM and OLS methods are seen very close in the figure. The MSE metric which is presented in the following table also supports that result. So we cannot say which result is better result by looking at the figure.

Table 4.19. Comparison of the GLM and the OLS methods with categorical response variables

Metrics	GLM	OLS
	Quarter And Temperature	
MSE	212,0454	212,403
AIC	1480,999	1710,647
Null deviance	240,9493	60025,91
Residual deviance	173,8299	43967,42
Residual / Null Deviance	0,721437663	0,732474027

Table 4.19 represents the comparison of the GLM fitted model and the OLS fitted model with categorical response variables. The GLM fitted model gives better AIC result.

The GLM fitted model and the OLS fitted model give almost the same MSE result. In addition to these, the residual to null deviance fraction gives better result for the GLM fitted model. That means adding “Quarter” and “Temperature” predictor variables to the GLM fitted model decreases null deviance more than that of the OLS fitted model. As a result, if we compare two models based on AIC and, the residual to null deviance fraction then, GLM fitted model represents observed data better than the OLS fitted model.

4.2.3. Comparing GLM with the Other Prediction Methods

Future sales prediction can be performed by predictive data mining algorithms as well as the regression models.

4.2.3.1. Predictive data Mining Methods

Predictive techniques of data mining are used for exploring the historical data and investigating the systematic relationship between the predictor variables. And then this relationship is used for predicting the future behavior of an unseen subset of data. This section presents common data mining prediction methods.

Data mining methods which are Generalized Linear Models with Elastic Net Regularization (GLMNet), Generalized Boosted Models (GBM), Principal Component Regression (PCR), Support Vector Machine (SVM), Random Forest (RF), Conditional inference trees (CTree) and Ensemble Learning (EL) are used for predicting the sales amount with the given categorical on non-categorical predictor variables.

4.2.3.1.1. Predictive Data Mining Methods with Non-Categorical Predictor Variables

Table 4.20 gives the comparison of 7 different data mining predicting methods and GLM method with non-categorical predictor variables.

Table 4.20. Comparison of the data mining methods' performance with the non-categorical predictor variables

Non-Categorical Sales2014							
Predictor Variables as Factor	**MSE of Methods**						
Days	x			x		x	x
Temperature		x		x	x		x
Price			x		x	x	x
GLMNet	281,9844	259,6785	284,6393	258,5426	260,4240	282,8659	259,3970
GBM	235,4106	266,9281	291,2424	269,6908	300,9476	229,9893	247,1120
PCR	277,5722	258,1574	284,7068	265,4049	270,1040	277,5638	269,3806
SVM	208,5947	269,1827	303,5295	218,4861	265,2296	218,0211	220,5765
Random Forest	182,4136	281,4718	**263,1377**	196,1034	271,1298	186,0975	**200,0894**
Ctree	**148,1135**	**244,6175**	289,9803	220,8652	244,6175	**148,1135**	220,8652
Ensemble Learning	227,5605	281,9034	303,6858	**183,0940**	**227,2966**	186,8814	224,5343
GLM	288,6811	267,5052	287,1445	250,5749	259,6151	288,4999	249,1760

Similar to the methodology followed in the previous sections, predictor variables are added to the model incrementally. At the first step, only one predictor variable that is "Day" is added to the models. CTree gives the best MSE metric with the "Day" predictor variable.

When we add only "Temperature" predictor variable to the model, MSE metrics of GBM, SVM, RF, CTree, EL give worse results, GLMNet, PCR and GLM give better result, then the previous model which is constructed with only the "Day" predictor variable. CTree outperforms the others with the "Temperature" predictor variable.

When we add only "Price" predictor variable to the model, MSE metrics are not the best with that predictor comparatively to the previous models. RF gives the best MSE metric with the "Price" predictor variable.

When we add two predictor variables which are "Day" and "Temperature" to the model, EL gives the best MSE metrics with "Day" and "Temperature" predictors. In addition to these, GLMNet and GLM give better results than the previous models which are constructed with only one predictor variable.

When we add two predictor variables which are “Temperature” and “Price” to the model, MSE metrics are not the best with that predictor comparatively to the previous models. EL model is the best with “Temperature” and “Price” predictor variables.

When we add two predictor variables which are “Day” and “Price” to the model, CTree model is the best with “Day” and “Price” predictor variables and also it is the best MSE value overall comparatively to previous models. And also, GBM and EL give their best MSE value.

When we add three predictor variables which are “Day”, “Temperature” and “Price” to the model, RF gives the best MSE metrics with “Day”, “Temperature” and “Price” predictors. And also, GLM gives the best MSE value which are constructed with the other combination of predictor variables and fitted with the GLM.

As a result, the CTree fitted model which is constructed with “Day” and the CTree fitted model which is constructed with “Day” and “Price” predictor variables give the best MSE value. The model which has “Day” predictor variable is chosen for comparison because of simple model complexity. CTree model is compared with the GLM model which is constructed with “Day” and “Temperature” predictor variables. Even the GLM model which is constructed with three predictor variable gives the best MSE value, the GLM model which has smaller AIC value is chosen. The visual representation of comparison is presented in Figure 4.18.

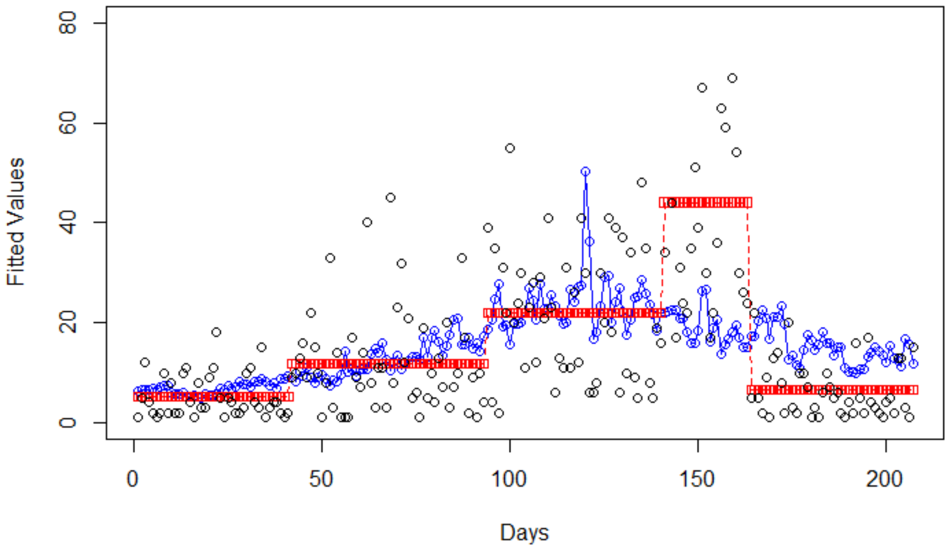


Figure 4.18. Visual representation of the GLM and CTree methods with the non-categorical variables

The blue curved line in Figure 4.18 shows the GLM fitting result by adding two predictor variables which are “Day” and “Temperature” variable, to the model. The red rectangular pointed line is the visual of the CTree fitting result by adding the predictor variables which is “Day”. Real sales values in the year of 2014 are represented in the Figure 4.18 as the black rounded points. As seen from the figure, between 0 to 150 days fitted values are similar with GLM and CTree. But between 150 to 200 days CTree fitted values are close to observed values comparatively to GLM. So MSE metric of the CTree model is smaller than MSE metric of the GLM. The “Day” factor better represents the data with CTree model than the “Day” and “Temperature” factors with GLM model does.

4.2.3.1.2. Predictive Data Mining Methods with Categorical Predictor Variables

Table 4.21 gives the comparison of these 7 different data mining predicting methods and GLM method with the categorical variables.

Table 4.21. Comparison of the data mining methods performance with the categorical predictor variables

Categorical Sales2014							
Predictor Variables as Factor	**MSE of Methods**						
Quarter	x			x		x	x
Temperature		x		x	x		x
Price			x		x	x	x
GLMNet	226,0597	275,8065	273,7088	221,025	264,6517	230,341	225,7461
GBM	1057,236	284,4265	304,6089	834,0854	295,9492	707,222	493,4558
PCR	230,0266	276,7929	271,9734	238,3523	266,8535	233,4338	230,1194
SVM	254,6224	307,2255	316,1198	261,8626	306,2575	264,9185	270,4253
Random Forest	225,9463	246,9802	**261,7086**	224,8032	259,2807	**178,9533**	**190,9274**
CTree	220,2684	264,8507	269,1998	223,2876	264,8507	223,2876	223,2876
Ensemble Learning	251,3017	**243,8648**	270,3831	235,7307	**247,372**	231,1219	212,4232
GLM	**220,0774**	263,6128	267,4917	**212,0454**	253,8025	220,1729	213,7004

Similar to the methodology followed in the previous sections, predictor variables are added to the model incrementally. At the first step, only one predictor variable that is

"Quarter" is added to the models. GLM gives the best MSE metric with the "Quarter" predictor variable.

When we add only "Temperature" predictor variable to the model, all the MSE metrics give worse results, except EL, then the previous model which is constructed with only the "Quarter" predictor variable except GBM. EL outperforms the others with the "Temperature" predictor variable.

When we add only "Price" predictor variable to the model, MSE metrics are not the best with that predictor comparatively to the previous models. RF gives the best MSE metric with the "Price" predictor variable.

When we add two predictor variables which are "Quarter" and "Temperature" to the model, GLM gives the best MSE metrics with "Quarter" and "Temperature" predictors. In addition to these, GLMNet, CTree, RF and EL give better results than the previous models which are constructed with only one predictor variable. That means, the "Quarter" and "Temperature" predictor variables are more significant than only one predictor variable which is "Quarter". The "Quarter" and "Temperature" predictor variables together fit the model better comparatively to only the "Quarter" predictor variable.

When we add two predictor variables which are "Temperature" and "Price" to the model, MSE metrics are not the best with that predictor comparatively to the previous models. EL model is the best with "Temperature" and "Price" predictor variables.

When we add two predictor variables which are "Quarter" and "Price" to the model, RF model is the best with "Quarter" and "Price" predictor variables. But MSE metrics are not the best with "Quarter" and "Price" predictors comparatively to the previous models.

When we add three predictor variables which are "Quarter", "Temperature" and "Price" to the model, RF gives the best MSE metrics with "Quarter", "Temperature" and "Price" predictors. And also, EL gives a better result comparatively to the other models which are constructed with the other combination of predictor variables and fitted with the EL.

As a result, the RF fitted model which is constructed with "Quarter" and "Price" predictor variables and GLM fitted model which is constructed with "Quarter" and "Temperature" are visualized in Figure 4.19.

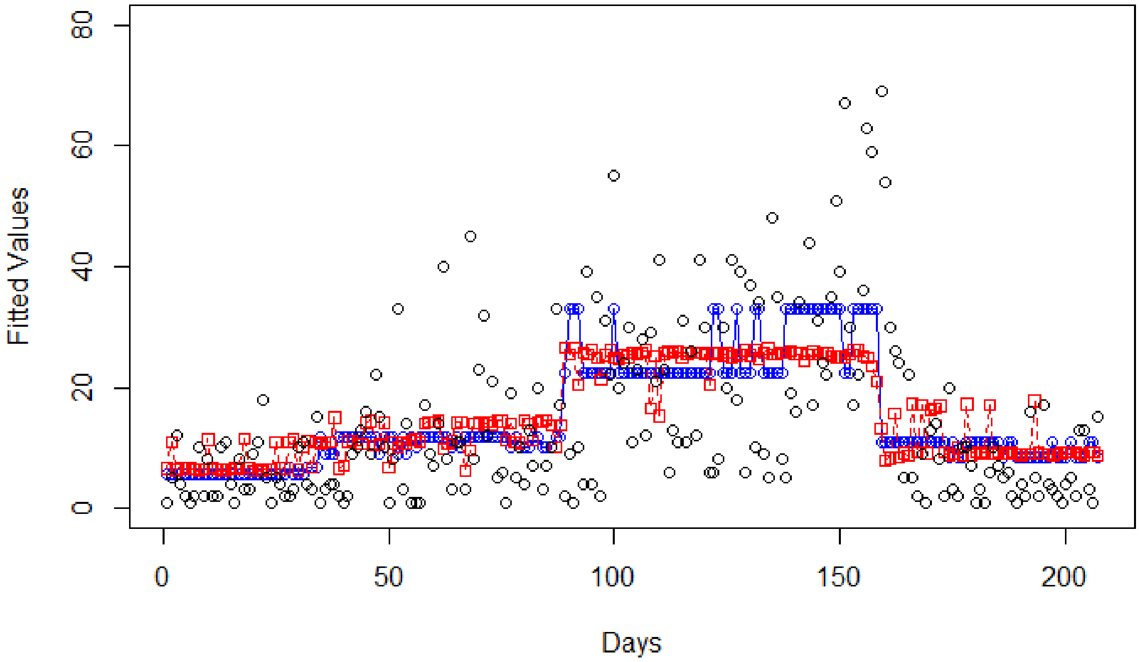


Figure 4.19. Visual representation of the GLM and Random Forest methods with the categorical variables

The blue curved line in Figure 4.19 shows the GLM fitting result by adding two predictor variables which are "Quarter" and "Temperature" variable to the model. The red rectangular pointed line is the visual of the RF fitting result by adding two predictor variables which are "Quarter" and "Price". Real sales values in the year of 2014 are represented in the Figure 4.19 as the black rounded points. As seen from the figure, the RF fitted values and GLM fitted values are similar between 0 and 100 days and between 150 and 200 days. The difference between is observed between 100 and 150 days. The predicted values of the RF model is close to observation, so MSE metric of the RF model is smaller than MSE metric of the GLM. The "Quarter" and "Price" factors better explain the data with RF model than, the "Quarter" and "Temperature" factors with GLM model does.

The main goal of the RF algorithm is to reduce the variance of the prediction while keeping the bias at a low level. MSE metric depends on the combination of the bias and variance of the predicted values. RF model has the lowest MSE value in comparison to data mining models and GLM model.

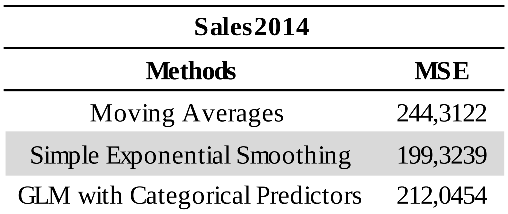
4.2.3.2. Time Series Forecasting Methods

Time series forecasting methods are the other approach of predicting the future behavior of a given system. First, we discussed explanatory models which try to predict the response variable with one or more predictor variables in an explanatory relationship. Distinct from the explanatory prediction methods, time series forecasting methods make prediction without any predictor variables which may affect the system. There are two reasons for using time series forecasting methods. First, understanding the system or measuring the relationship with the system and predictor variables may be difficult. Second, the main goal is only predicting the response variable without understanding how it happens (Makridakis, 1998).

This section presents time series forecasting methods which are Moving Averages (MA), Simple Exponential Smoothing (ES) and Time Series with seasonality (TS) for predicting the sales amount with the given sales history.

Table 4.22 presents the predicted results of the sales data by using the year of 2014. And the visual representation of the GLM and the ES predicted results are shown in Figure 4.20.

Table 4.22. Comparison of time series forecasting methods with using the year of 2014 sales data



Sales2014	
Methods	**MSE**
Moving Averages	244,3122
Simple Exponential Smoothing	199,3239
GLM with Categorical Predictors	212,0454

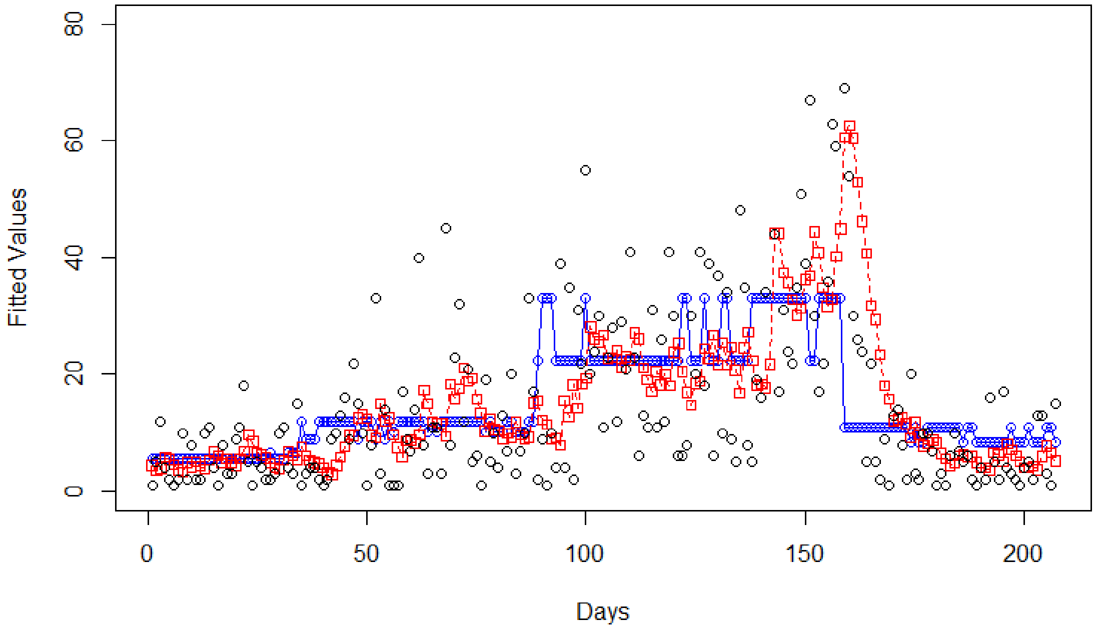


Figure 4.20. Visual representation of GLM with categorical response variables and Simple Exponential Smoothing method

The blue curved line in Figure 4.20 shows the GLM fitting result by adding two predictor variables which are the “Quarter” and “Temperature” to model. The red rectangular pointed line is the visual of the ES fitting result. Real sales values in the year of 2014 are represented in Figure 4.20 as the black rounded points.

As a result, ES method gives better a result than the other methods. Because the recent observations of sales are relatively more weighted in prediction than the older observations. So a few past observation affects the recent prediction than the other observations. On the other hand, ES has a lag of ability to apply the trend model and cyclical or seasonal data. And also, ES does not include the effect of external factors into the model, for example, air temperature or the other variables.

Table 4.23 presents the predicted results of sales data by using the years between 2010 and 2014 with TS method and predicted results of sales data by using the year 2014 with GLM method with “Quarter” and “Temperature” predictor variables. The visual representation of GLM method with “Quarter” and “Temperature” predictor variables and TS method are presented in Figure 4.21 and Figure 4.22.

Table 4.23. Comparison of TS method and GLM method

Method	MSE
Time Series (Seasonality)	126,8212
GLM with Categorical Predictors	212,0454

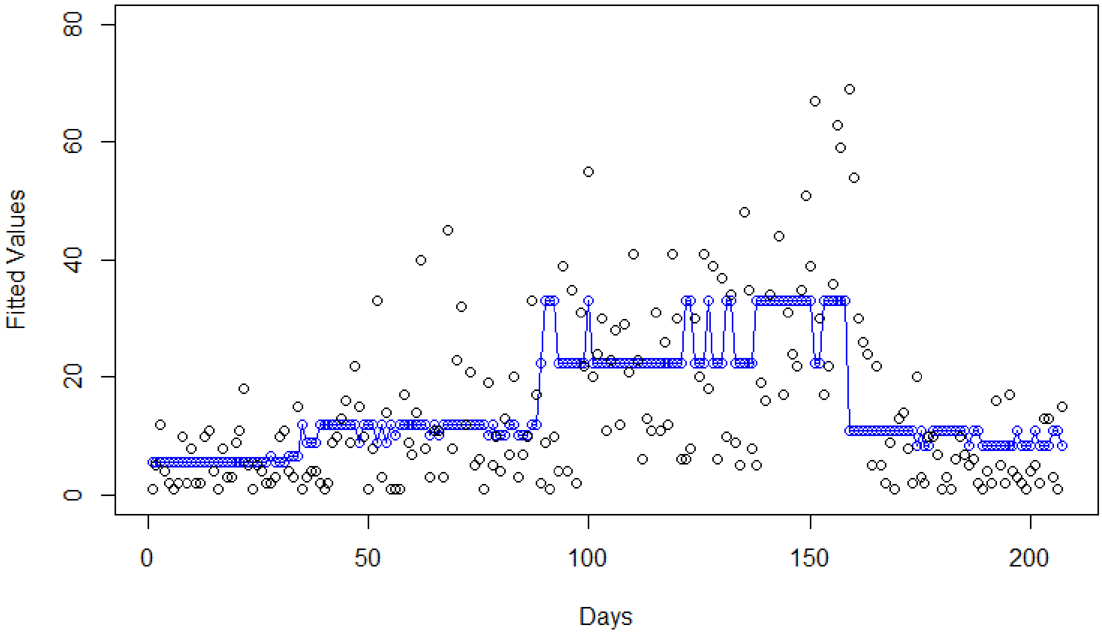


Figure 4.21. Visual representing of GLM method with categorical response variables

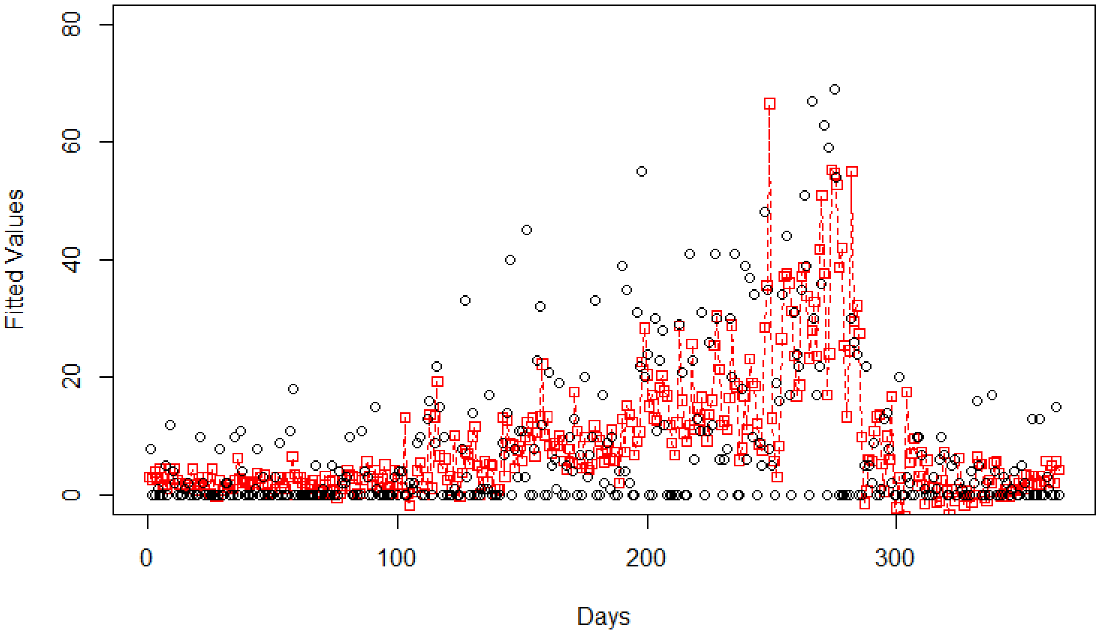


Figure 4.22. Visual representing of TS method

The blue curved line in Figure 4.21 shows the GLM fitting result by adding two predictor variables which are “Quarter” and “Temperature” variables. The red rectangular pointed line in Figure 4.22 is the visual of the TS fitting result. Real sales values are represented as the black rounded points in both figures. TS method is implemented with five years data and 365 days cycle. So Figure 4.22 presents 365 points which represent individual sales amounts for the year 2014.

The seasonal pattern of the data was presented in “Decomposition of Time Series (Seasonality)” section previously. This pattern of the seasonal variations better contributes to the prediction of the future sales than GLM does. Because seasonal pattern is established by using past five years’ data and TS prediction eliminates unwanted external factors for example outliers. In addition to these, the best known seasonal pattern is the outdoor temperature with seasons of the years effect the seasonal pattern. The best GLM result with the factors which are “Quarter” and “Temperature” presented before supports also the seasonal effects on the sales and TS results.

4.3. Model Validation

The validation of the selected model is one of the important step of model building. The details of the model validation is explained in chapter two under the model validation section. In this section we do validation on the fitted models which are constructed by non-categorical predictor variables. Hold-out samples which are samples of data that are not used in fitting a model are used to validate the fitted models. Istanbul sales are used as the training data and Izmir sales are used as the validation data. The results are presented in the following table.

Table 4.24. Validation of non-categorical sales data

Non-Categorical Sales2014								
Predictor Variables as Factor		**MSE of Methods**						
Days		x			**x**		x	x
Temperature			x		**x**	x		x
Price				x		x	x	x
GLMNet	Training	281,9844	259,6785	284,6393	258,5426	260,4240	282,8659	259,3970
	Validation	136,3917	136,6720	135,9715	136,6720	138,3373	135,9715	138,3373
GBM	Training	235,4106	266,9281	291,2424	269,6908	300,9476	229,9893	247,1120
	Validation	169,6955	**117,2606**	**121,9277**	152,3425	**114,9580**	191,1965	147,7359
PCR	Training	277,5722	258,1574	284,7068	265,4049	270,1040	277,5638	269,3806
	Validation	136,3344	155,5300	154,5678	136,4315	156,8812	135,8821	136,6662
SVM	Training	208,5947	269,1827	303,5295	218,4861	265,2296	218,0211	220,5765
	Validation	188,3921	**118,1168**	141,2607	171,1725	**106,7127**	179,4780	165,0091
Random Forest	Training	182,4136	281,4718	263,1377	196,1034	271,1298	186,0975	200,0894
	Validation	200,3054	188,3117	**120,8555**	201,0203	167,3700	180,3512	182,7320
CTree	Training	148,1135	244,6175	289,9803	220,8652	244,6175	148,1135	220,8652
	Validation	180,6332	150,7834	155,2848	151,0385	150,7834	180,6332	151,0385
Ensemble Learning	Training	227,5605	281,9034	303,6858	183,0940	227,2966	186,8814	224,5343
	Validation	177,8041	149,1145	134,1322	130,9450	154,6232	184,8372	128,5726
GLM	Training	288,6811	267,5052	287,1445	250,5749	259,6151	288,4999	249,1760
	Validation	136,5303	214,8490	162,0401	**122,3348**	**606,2935**	136,3448	123,4321
LM	Training	277,5722	258,1574	284,7068	253,3723	255,7137	277,5303	253,2688
	Validation	136,3344	155,5300	154,5678	136,6564	159,0696	135,9014	138,4332
Moving Averages	Training	244,3122	244,3122	244,3122	244,3122	244,3122	244,3122	244,3122
	Validation	187,4593	187,4593	187,4593	187,4593	187,4593	187,4593	187,4593
Simple Exponential Smoothing	Training	199,3239	199,3239	199,3239	199,3239	199,3239	199,3239	199,3239
	Validation	187,0722	187,0722	187,0722	187,0722	187,0722	187,0722	187,0722
Time Series (Seasonality)	Training	126,8212	126,8212	126,8212	126,8212	126,8212	126,8212	126,8212
	Validation	146,7634	146,7634	146,7634	146,7634	146,7634	146,7634	146,7634

As seen from Table 4.24, the MSE values of the models which are constructed by training data and validation data are close to each other except the GLM validation model which is constructed with "Temperature" and "Price" predictor variables.

The following figure visualizes the GLM validation model which is constructed with "Temperature" and "Price" predictor variables.

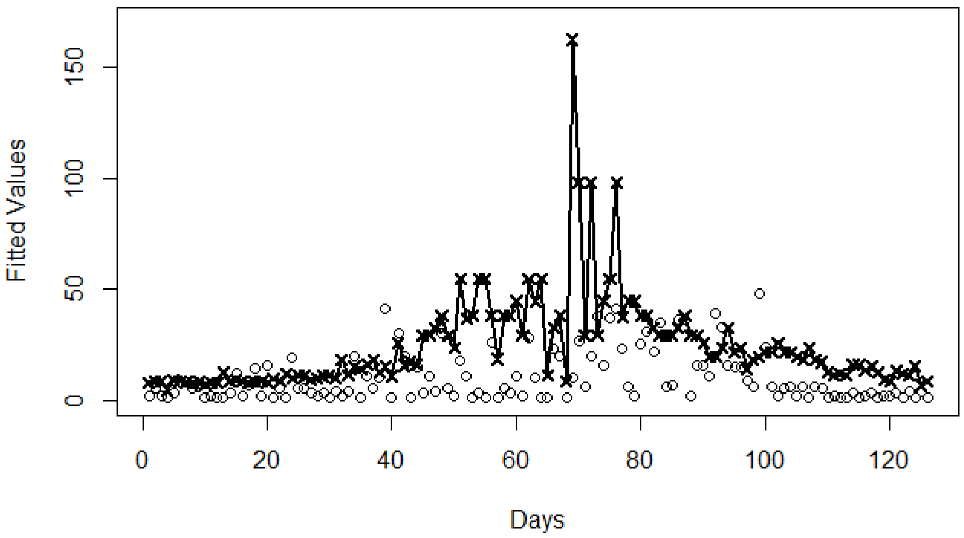


Figure 4.23. Visual representing of GLM method with "Temperature" and "Price" predictor variables

Figure 4.23 represents the fitted results of validation data with x type point and thick line, validation data is represented with o type points. As seen from the figure, there are 4 outliers for fitted model with validation data. So the MSE value of the model which is constructed with validation data is not close to the MSE value of the model which is constructed with training data. If the outlier are trimmed from the fitted results then the fitted result can be close to validation data. As a result, the fitted models can be used for predicting the future sales.

GLM with "Days" and "Temperature" predictor variables validation result is the best result when the other models with same predictor variables are compared with it. Also, only six models which are GBM, SVM and Random Forest models give better result than GLM with "Days" and "Temperature" predictor variables. The results of these models are remarked bold character in Table 4.24.

CHAPTER 5

CONCLUSION

Prediction is needed to determine the quantity of a demand when it occurs so that appropriate planning can be taken before the demand appears. Making an accurate demand prediction has a crucial role for any commercial company as it is an important part of the decision-making activities of management. To deal with demand prediction, various techniques of regression analysis and data mining are used under the predictive methods. The purpose of this thesis is to make sales history-based demand prediction by using generalized linear models.

The data set which is used for analysis is real company data. The data set includes sales data that consists of the variables of sales amount, the date of sales, item price, and air temperature. And also product codes and sales locations are included to data for exploratory data analysis. We particularly utilize the most popular product's sales data for the city of Istanbul. In our modeling, the response variable is the sales amount and the date of sales, item price, and air temperature are selected as the predictor variables.

First, exploratory data analysis techniques are used for understanding the basic statistical characteristics and the distribution of data. The distribution of sales amount which is the response variable for real customer data is discovered as the gamma distribution. Because of the response variable distribution, GLM method is used with the gamma distribution which is a member of the exponential family and inverse link function is used. Investigating the data, choosing the GLM setting in accordance with the response variable distribution along with an appropriate link function are crucial steps for characterizing the data through the use of modeling like GLM.

In any modeling case, the target model is fitted to the whole set of data points. Thus, the variance of data points and the bias of them from their real population should be considered. In order to govern the total variance due to the predictor variables, if there are a few variables like our case, all the possible combinations of them should be taken into consideration. If the original data types of variables and their ranges cause a lot of variance, it can be an option to apply categorization to the variables in order to better identify their impact on the response variable.

The combinations of three different predictor variables, “Days”, “Temperature” and “Price” are analyzed. Then, the predictor variables are transferred into categorical variables for investigating the effect of categorization. The “Temperature” and “Price” predictor variables are categorized in three levels which are low, medium and high. The “Days” predictor variable is categorized as four levels which includes three sequential months. The combinations of three different categorical predictor variables, “Quarter”, “Temperature” and “Price” are analyzed. “Days” and “Temperature” predictor variables are selected as the factors that mostly affect the response variable in both categorical and non-categorical predictors. When categorical predictor variable fitting results are compared with that of non-categorical, the former one gives better results than the latter. Thus, categorization provides more accurate prediction mostly due to variance reduction on predictor variables.

When the GLM result is compared with the other predictive data analysis techniques, our findings are as follows: Within the scope of regression techniques, GLM is compared with the linear model both for default and inverse response variables. As a result of the comparison, GLM gives better fitting results than the linear model in both cases. When it comes to time series methods, time series with seasonality method gives better prediction results than the GLM. Although time series method does not include any predictor variable, seasonal pattern of the data better contributes to the prediction of the future sales. Within the scope of classification techniques, RF gives the best fitting result. This performance can be explained by the fact that RF algorithm reduces variance of prediction while keeping the bias at a low level. However, when “Days” and “Temperature” predictor variables are selected, the GLM method outperforms the others.

The model fitting results are evaluated with respect to MSE and AIC metrics.

In our sales demand prediction problem; the solution set can be formed by the combinations of the predictor variables, data discretization, and different modeling methods. Adding a new predictor variable, using another discretization method or investigating another modeling method increase the candidate models which are to be evaluated.

Further research on this topic could have the following directions:

- The results can be validated on new data. For example the next year’s sales data can be collected and the prediction results are compared with the new data.
- Different discretization methods for categorizing the predictor variables can be used.

- Hybrid models can be constructed such as the formulation of the GLM model along with an additive time series component.
- In order to run all these analysis on the used ERP program, R scripts can be integrated with it. A way for using R scripts in the ERP program is using COM (Component Object Model) objects which enable the R functions in the ERP program.

Finally, the contribution of this study to literature is summarized. First, the real company data in enterprise resource planning (ERP) form are used for making the demand prediction. Second; GLM is used for modeling the demand prediction and its fitting performance is compared with that of OLS and predictive data mining methods. The results show that the GLM has important potential in this modeling task. Third, new predictor variables which are “Days”, “Quarter”, “Temperature”, and “Price” are evaluated for demand prediction within the scope of the given company data.

REFERENCES

Balajir, R. (2011). Advanced Data Analysis Techniques CVEN 6833 Lecture Notes, http://civil.colorado.edu/~balajir/CVEN6833/lectures/GammaGLM-01.pdf

Breiman, L. (2001). Random Forests, Machine Learning, Kluwer Academic Publishers

Breiman, L. (2015). Bias, Variance, and Arcing Classifiers, Statistics Department University of California

Cohen, P. (1995). Empirical Methods for Artificial Intelligence, The MIT Press

Cornwell, B. (2010). Adding products and services: Where does innovation come from, Global Text Project, USA

Cortes, C., Vapnik, V. (1995). Support-Vector Networks, AT&T Bell Labs USA

De Jong, P., Heller, G. (2008). Generalized Linear Models for Insurance Data, Cambridge University Press, London

Dobson, A. (2008). An Introduction to Generalized Linear Models. 3rd ed., Chapman & Hall

Du, J., He, R., Zhechev, Z. (2014). Forecasting Bike Rental Demand, Stanford University

Exponential family. (n.d.). In Wikipedia. Retrieved June 01, 2016, from https://en.wikipedia.org/wiki/Exponential_family

Fortmann-Roe S. (2013). Understanding the Bias-Variance Tradeoff, http://scott.fortmann-roe.com/docs/BiasVariance.html

Friedman, J., Hastie, T., Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent, Journal of Statistical Software

Gutierrez, R., (2016). Validation Lecture Notes, Inteligent Sensor Systems, Wright State University, research.cs.tamu.edu/prism/lectures/iss/iss_l13.pdf

Hand, D., Mannila, H., Smyth, P. (2001). Principles of Data Mininng, The MIT Press

Hothorn, T., Hornik, K., Zeileis, A. (2004). Unbiased Recursive Partitioning: A Conditional Inference Framework, Wirtschaftsuniversität Wien

Hothorn, T., Hornik, K., Zeileis, A. (2015). CTree: Conditional Inference Trees, The Comprehensive R Archive Network

Jolliffe, Ian T. (1982). A note on the Use of Principal Components in Regression, Journal of the Royal Statistical Society

Jolliffe, Ian T. (1986). Principal Component Analysis Springer-Verlag

Kolyshkina, I. Wong, S. And Lim, S. (2005). Enhancing Generalised Linear Models with Data Mining: PricewaterhouseCoopers

Kutner, M., H., Nachtsheim, C., Neter, J., Li, W. (2004) Applied Linear Statistical Models, McGraw-Hill

Law, A., Kelton W. (1991). Simulation Modelling & Analysis 2nd ed., McGraw-Hill

Lebanon, G. (2010). Bias, Variance, and MSE of Estimators, Georgia Institute of Technology

Li, M. (2011). Data mining fall lecture notes, Chapter 7: Score Functions for Data Mining Algorithms, Department of Computer Science and Technology, Nanjing University

Liu, A., Meiring, W., Wang, Y. (2004) Testing generalized linear models using Smoothing spline methods, Statistica Sinica 15

Makridakis S., Wheelwright, S., Hyndman S. (1998). Forecasting: Methods and Applications, 3rd Edition, John Wiley & Sons, Inc.

Motulsky, H., Christopoulos, A. (2003). Fitting Models to Biological Data using Linear and Nonlinear Regression, A practical guide to curve fitting, GraphPad PRISM Version 4.0

Myers, R. H., Montgomery, D. C., Vining, G. G, (2010). Generalized Linear Models: with Applications in Engineering and the Sciences 2nd ed., John Willey & Sons

Nelder, J., Wedderburn, R. (1972). Generalized Linear Models, Journal of the Royal Statistical Society, Blackwell Publishing

O'Sullivan, A., Yandell, S., Raynor, W. (1986) Automatic Smoothing of Regression Functions in Generalized Linear Models: Journal of the American Statistical Association

Pearson, K., (1900). On the criteria that a given system of deviations from the probable in the case of correlated system of variables is such that it can be reasonable supposed to have arisen from random sampling. Phil. Mag., Vol. 50, PP. 157-172.

Royston, P., Altman, D., Sauerbrei, W. (2005). Dichotomizing continuous predictors in multiple regression: a bad idea, Wiley InterScience

Shapire, R., Freund, Y. (2014). Boosting: Foundations and Algorithms, The MIT Press

Tan, P., Steinbach, M., Kumar, V. (2006). Introduction to Data Mining, Pearson International Edition, Pearson Education Inc.

Tauras, A. (2005). An Empirical Analysis of Adult Cigarette Demand: Eastern Economic Journal

The Odum Institute (2015). Logistic Regression and the American National Election Study 2012: Vote Choice in the 2012 US Presidential Election, SAGE Publications, Ltd.

Weather API. (n.d.). In World Weather Online Developer Portal. Retrieved August 01, 2015, from http://developer.worldweatheronline.com/api/

Zhou, Z. (2012). Ensemble Methods: Foundations and Algorithms, Chapman and Hall