Süleyman Demirel Üniversitesi
Fen Bilimleri Enstitüsü Dergisi
Cilt 23, Sayı 3, 840-849, 2019

Süleyman Demirel University
Journal of Natural and Applied Sciences
Volume 23, Issue 3, 840-849, 2019

DOI: 10.19113/sdufenbed.558620

# Sales History-based Demand Prediction using Generalized Linear Models

**Başar ÖZENBOY**[1], **Selma TEKİR**[*, 2]

[1]The Scientific and Technological Research Council of Turkey (TÜB˙ITAK), Advanced Technologies Research Institute, 06800, Ankara, Turkey
[2]Izmir Institute of Technology, Faculty of Engineering, Department of Computer Engineering, 35430, Izmir, Turkey
[1](ORCID:https://orcid.org/0000-0001-9809-7354)
[2](ORCID:https://orcid.org/0000-0002-0488-9682)

(Alınıs¸ / Received: 28.04.2019, Kabul / Accepted: 18.09.2019, Online Yayınlanma / Published Online: 30.12.2019)

**Keywords**
Sales demand prediction,
Gamma distribution,
Generalized linear models,
Categorization

**Abstract:** It's vital for commercial enterprises 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 is used to predict a numerical response variable like sale amount. In this category, linear models are simple and easy to interpret yet they permit generalization to very powerful and flexible families of models which are called Generalized linear models (GLM). The generalization potential over simple linear regression can be explained twofold: First, GLM 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 work models the sales amount prediction problem through the use of GLM. Unique company sales data are explored and the response variable, sale amount is fitted to the Gamma distribution. Then, inverse link function, which is the canonical one in the case of gamma-distributed response variable is used. The experimental results are compared with the other regression models and the classification algorithms. The model selection is performed via the use of MSE and AIC metrics respectively. The results show that GLM is better than the linear regression. As for the classification algorithms, Random Forest and GLM are the top performers. Moreover, categorization on the predictor variables improves model fitting results significantly.

# Genelleştirilmiş Doğrusal Modeller Kullanılarak Satış Geçmişine Dayalı Talep Tahminlemesi

**Anahtar Kelimeler**
Satıs¸ talep tahminlemesi,
Gama dağılımı,
Genelleştirilmiş doğrusal modeller,
Kategorizasyon

**Özet:** Ticari is¸letmeler için mevcut satıs¸ verilerini kullanarak talebi net olarak tahmin etmek önemlidir. Ş irketlerin karlılığı artırmak için karar destek sistemlerinin bir parçası olarak tahmin analitiği yapabiliyor olması gerekir.Tahmine yönelik veri analitiğinde, regresyon modelleri satış miktarı gibi sayısal bir bağımlı değişkenin tahmin edilmesinde kullanılır. Bu kategoride doğrusal modeller basittir, yorumlanması kolaydır ve aynı zamanda genelleştirilmiş doğrusal modeller (GLM) olarak adlandırılan çok güçlü ve esnek model ailelerine genelleştirme yapılmasını sağlar. Basit doğrusal regresyona göre genelleştirme potansiyeli iki katlı olarak açıklanabilir: İlk olarak GLM normal dağılımlı hata terimleri varsayımını yumuşatır. Ayrıca, tahmin değişkenleri kümesi ile bağımlı değişken arasındaki bağlantı fonksiyonunu özdeşlik fonksiyonu ile sınırlandırmaz. Bu çalışmada satış miktarı tahmin problemi GLM ile modellenmis¸tir. Model uyarlamasını eniyiles¸tirmek için bir s¸irkete ait satış verilerinin keşifsel analizi yapılmış ve bağımlı değişken olan satış miktarının dağılımı gama dağılımı olarak bulunmuştur. Sonrasında, gama dağılımlı bağımlı değişken için standart bağlantı fonksiyonu olan ters bağlantı fonksiyonu kullanılmıştır. Deneysel sonuçlar diğer regresyon modelleri ve sınıflandırma algoritmalarıyla karşılaştırılmıştır. Model seçiminde MSE ve AIC ölçütleri kullanılmıştır. Sonuçlar GLM'nin doğrusal regresyondan daha iyi olduğunu göstermektedir. Sınıflandırma algoritmaları açısından ise, rastgele orman ve GLM en üst performansı göstermiştir. Ayrıca, tahmin değişkenlerinin kategorizasyonunun model uyumunu iyileştirdiği görülmüştür.

*Corresponding author: selmatekir@iyte.edu.tr*

## 1. Introduction

Demand prediction is a vital activity for commercial companies. Companies should better manage current resources and plan for future needs in order not to lose their competitive advantage and reduce costs. The uncertainty in future makes the prediction hard. There are various methods for demand prediction yet the research community is still in search of more effective prediction techniques.

Sales demand prediction can vary due to short/intermediate/long range prediction, the characteristic of good such as durable or not, the type of response variable (binary, categorical, or numerical), and the model choice in the form of parametric vs. nonparametric.

Among existing models, linear models are simple and easy to interpret yet permit ready generalization to very powerful and flexible families of models. Generalized linear models (GLM) ([1]) represent such a powerful and flexible families of models. In this work, we predict demand by making a novel adaptation of GLM for unique company data.

To clarify the idea, it's useful to explain simple linear regression in conjunction with GLM:

Simple linear regression is commonly used to predict a numerical response variable like sales amount. It has some assumptions to simplify the theory of analysis. One of the assumptions is regarding error terms. Linear regression models assume that the error terms are normally distributed. The second assumption is such that response variables are independent normal random variables.

In some real world applications, error terms and response variables may not have normal distribution. In that case GLM can be used instead of linear regression models. GLM relax the assumption of normally distributed error terms. Moreover, GLM can be used for predicting the expected value of a response variable which has a distribution from the exponential family. Whenever the response variable is no more normally distributed, a constant change in a predictor variable does not lead to a constant change in the response variable. Thus, the relationship between the set of predictor variables and the response variable could be represented by a set of link functions rather than the identity function.

GLM can provide a solution for different types of response variable distributions. For a binary response variable, the two popular link functions are logit and probit. In demand prediction; besides estimating the amount of demand estimating the presence of demand can be crucial as well. Linear models which use logit link function to predict the probability of demand thus have common usage.

[2] compares and contrasts the probit ang logit link functions through the use of an example case.

[3] performs an empirical study on the cigarette demand problem. Cigarette demand problem traditionally is modeled as a mixed distribution: a logit specification to predict the decision to smoke and OLS for estimating the intensity of smoke. He tried to model the intensity of smoke in a population. The problem was modeled with both ordinary least squares (OLS) method and GLM. Results were compared to understand the importance of prediction bias due to omitting error terms while data transformation. The results show that OLS method overestimates the effect of price on the cigarette demand when compared to GLM. Because in the case of OLS, a logarithmic transformation is performed on the response variable whereas GLM performs logarithmic transformation on the expected value of the dependent variable. In other words, OLS with logarithmic transformation has a constant variance assumption which does not represent the truth. GLM, which has non-constant variance assumption thus performs better.

[4] conducts a study to predict voting behavior to Obama or Romney in 2012 American National Election. The study is based on logit model to evaluate the dichotomous dependent variable. The method performs well when the data set is sufficiently big.

We have five years' (2010, 2011, 2012, 2013, 2014) sales data for a cooling company. The data set includes sales data that consists of the variables of sale amount, the date of sale, item price, and air temperature.

In our sales demand prediction problem; the response variable, sale amount is found as gamma distributed. Thus, GLM with gamma distributed dependent variable is used. The canonical link function in this case is the inverse link function and it is also found to be best performing by the experimental evidence.

In our solution scheme, the combinations of three different predictor variables, "Days", "Temperature" and "Price" are analyzed. Then, the predictor variables are transformed into categorical variables for investigating the effect of categorization. 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 gives better fitting results than the linear model. Within the scope of classification techniques; in single predictor variable cases Random Forest is the best, but when "price" predictor variable is used in conjunction with other variable(s), GLM outperforms the others. The model fitting results are evaluated with respect to MSE and AIC metrics.

Our contribution comes in two different ways: Although

GLM is an old technique in modeling, its use in data mining is relatively not widespread almost restricted to the use of logistic regression. In fact, GLM is composed of a set of models that can be configured with respect to inherent characteristics of data. The generalization property is due to this. In our work, we used unique company data for sales demand prediction and adapted the GLM using data distribution (GLM with gamma distributed dependent variable). Moreover, we performed a comparative analysis with the linear model and other data mining algorithms considering the effect of feature selection and categorization.

## 2. Material and Method

In this section; first we explain GLM along with its adaptation to unique company sale demand prediction problem. Then, we go through the description of the classification and regression algorithms in our comparison base. Finally, we briefly describe model evaluation with respect to the sampling techniques (cross-validation and hold-out sampling) and evaluation metrics that are MSE and AIC respectively.

### 2.1. GLM

Linear regression models have some assumptions to simplify the theory of analysis. One of the assumptions is regarding error terms. Linear regression models assume that the error terms are normally distributed. The second assumption is such that response variables are independent normal random variables. [5].

Figure 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, $\varepsilon_i$ implies the error term with normal distribution. $Y_i$ represents the real-valued response variable.

The right-hand side $aX + b$ component is a functional form. The transformation linking the functional form to the expected value of the response variable is called a link function which is identity in the case of linear regression.

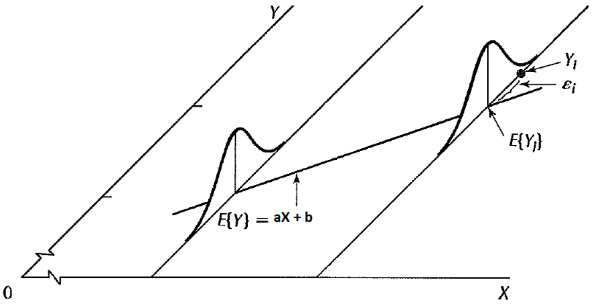

**Figure 1.** Linear regression model (Source: [5] ).

In some real world applications, error terms and response variables may not have normal distribution. In that case generalized linear models (GLM) can be used instead of linear regression models. GLM relax the assumption of normally distributed error terms. Moreover, 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 [1]. 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 other words, the issue is to find out the functional form-link function pair that is in accordance with the left-hand side variable's expected value distribution.

The gamma distribution, which is a member of the exponential family is widely used to model physical quantities that take positive values. Sale amount is such a quantity and can be modeled as a random variable denoted as $Y \sim Gamma(\alpha, \beta)$ where $\alpha$ is the shape parameter and $\beta$ is the scale parameter. Our model fitting results confirm that sale amount distribution is best fit to a gamma distribution. The probability density function of a gamma distribution is as follows:

$$f(y_i) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} y_i^{(\alpha-1)} e^{-(\frac{y_i}{\beta})} \quad y_i \geq 0; \alpha, \beta > 0 \tag{1}$$

The expected value and variance equations are provided below:

$$E\{y_i\} = \mu_i = \alpha\beta \tag{2}$$

$$Var\{y_i\} = \alpha\beta^2 \tag{3}$$

As seen from the formulations, when the parameter $\beta$ (scale parameter) is not varying so much, the expected value of the response variable is just dependent on the shape parameter $\alpha$. Then, the task becomes to find an appropriate link function connecting $\alpha$ to the right-hand side linear model. In accordance with this, the canonical parameter for the gamma distribution is $-\frac{1}{\mu}$. If the link function is chosen to be the function expressing the canonical parameter for the distribution being used as the linear sum, the fit becomes better. So the canonical link function is the inverse link [6].

The two common link functions for GLM are inverse link (canonical link) and log link functions where the former takes the inverse of the expected value of the response variable while the latter takes its logarithm. We used inverse link (the canonical) function for GLM fitting, it gave better performance (4).

$$E\{Y\} = \frac{1}{a + bX_1 + cX_2} \tag{4}$$

In the given equation, $a+bX_1+cX_2$ is our functional form, $X_1$ and $X_2$ are our predictor variables "Day" and "Temperature", we take the inverse of the functional form to connect with the expected value of the response variable $Y$ (Sale amount) distribution.

In [6], GLM with a Gamma-distributed dependent variable is analyzed through different combinations of link functions and functional forms.

- Identity function-Inverse on the functional form
- Inverse link function-Inverse on the functional form (log link equivalent)

are proposed as alternative ways of GLM fitting. Identity function-Inverse on the functional form can be represented by the following formula,

$$E\{Y\} = a + b\frac{1}{X_1} + c\frac{1}{X_2} \tag{5}$$

whereas Inverse link function-Inverse on the functional form (log link equivalent) can be stated as follows:

$$\frac{1}{E\{Y\}} = a + b\frac{1}{X_1} + c\frac{1}{X_2} \tag{6}$$

We applied all these variations. As stated before, inverse link function along with the standard functional form performed better (Equation 4).

## 2.2. Comparison base

### 2.2.1. GLMNet

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 $\ell1$ (the lasso) and $\ell2$ (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, $\ell1$ (the lasso) penalty term is used. Ridge regression ($\ell2$) shrinks the estimated coefficients with shrinking method which adds a penalty on coefficients. The mixture of $\ell1$ and $\ell2$ penalties is named as the elastic net regularized regression method which is GLMNet [7].

### 2.2.2. Gradient boosting method (GBM)

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 [8].

### 2.2.3. Principal component regression (PCR)

Principal component analysis (PCA) was invented by Karl Pearson [9]. With principal component analysis (PCA) method, independent variables are projected onto new variables such that the sample variance is maximized and the resultant linear combination is uncorrelated with the original one. The resultant variables are named as principle components. Principal component regression (PCR) predicts the dependent variable using linear regression on the principal components [10].

### 2.2.4. Support vector machine (SVM)

[11] introduced support vector machine learning method. The method is based on support vectors which represent decision boundaries on the training set. One desired characteristic of these decision boundaries is having a large margin as small margin causes model overfitting. Every such decision boundary can be associated with two hyperplanes and the task is to find out the maximal margin classifier that separates those two hyperplanes. The default classifier works with linear decision boundaries on the binary classification problem. Support vector machines generalize this to more complex surfaces by transformations from a linear decision surface into a nonlinear one.

### 2.2.5. Random forest (RF)

Random forest was introduced by [12]. It is a combination of multiple decision trees. In classifying a new instance, majority voting is applied on component decision trees. In order to construct every individual decision tree, a random training set using sampling with replacement is generated out of the original one. The performance of random forest is mainly dependent on the correlation between component trees and the strength of each individual tree.

### 2.2.6. Conditional inference trees (Ctree)

Another predictive method which is similar to MARS method is Conditional Inference Tree (CTree) method which also systematically tries all the combinations of the variables to select the right predictor variables. It is a tree structured regression model. CTree creates a decision tree. It generates splits iteratively. These splits are generated for most significantly related variable with the response variable. That response variable is evaluated by $p$ values. Iterations finish when there is no more significant $p$ value available for the remaining variables [13].

#### 2.2.7. Ensemble learning

In ensemble learning, a set of classifiers is combined to make a better prediction. The component classifiers can be identical or diverse. In general, the aggregate opinion of diverse classifiers is better in reducing variance. The aggregate opinion is formed using the weighted average of individual votes [14].

### 2.3. Model evaluation

#### 2.3.1. Cross-validation

It's a sub-sampling technique in which the existing data are split into training and test sets. The model is trained using the training part and validated on the test part. In k-fold cross-validation, data are divided into $k$ equal parts of size $n$. In every iteration, the $i^th$ set of $n$ items are used as the test and the remaining as the training. After $k$ iterations, the average performance from the $k$ sets is recorded as the resultant performance.

#### 2.3.2. Hold-out sampling

In hold-out sampling, a separate validation set is utilized in order to assess the predictive performance of the model on unseen data.

#### 2.3.3. Akaike information criterion (AIC)

Akaike Information Criterion (AIC) is a metric that is used to evaluate the goodness of fit of a model. Different models of different complexity can be compared using this metric. AIC is formulated as follows:

$$AIC = -2l + 2p \tag{7}$$

In the formulation, $l$ is the log-likelihood term that describes how well the data are described given the model. $p$ represents the model complexity in terms of the number of parameters. Lower AIC values are preferred and AIC favors simpler models that explain the data well.

#### 2.3.4. Mean squared error (MSE)

It is a measure of the deviations of real data points from the model predicted ones. The sum of squares of individual errors is taken and normalized with respect to the number of data points. The sum of squared errors (SSE) is calculated using the following formula;

$$S_{SSE} = \sum_{i}^{n} (y(i) - \hat{y}(i))^2 \tag{8}$$

where the first term inside the summation represents a real data point while the second is a model estimated data point.

### 2.4. GLM adaptation to data

#### 2.4.1. Data set

We have five years' (2010, 2011, 2012, 2013, 2014) sales data for a cooling company. There are more than twelve types of products that are sold in almost all cities of Turkey. The total number of sales in the product database is 185986. The most popular product is $A$ with 134247 total number of sales. For our demand prediction problem, we referred to product $A$ sales in the city of Istanbul. The total number of sales meeting this criterion is 12788. We queried the product database in order to filter it with respect to this criterion. As a result, our sales records include date, sale amount, product item price, city name, and product code. Using the date and city information, we added air temperature as an additional feature to the data set.

#### 2.4.2. Goodness of fit tests

As part of exploratory data analysis, we calculated some summary statistics. One such statistic is the total product sales amount for every product sorted with respect to city. Then, we based our demand prediction on the most popular product sold in Istanbul, which had the highest number of product sales.

We performed goodness of fit tests on the collected data to determine its distribution (Table 1). The results are acquired according to the Chi-Squared fit test.

The null hypothesis in the case of 2010 sales amount data can be stated as follows:

$H_0$ =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.

According to the $p$ value obtained (0.21991), if we reject the $H_0$, we are 21% wrong. Thus, we cannot reject that the Sales2010 data fit to Log-logistic distribution.

In a similar way; as all $p$ values for the other years' sales data are greater than 0.05, we can conclude that Sales2011, Sales2012, Sales2013, Sales2014 data fit to Gamma Distribution, Gamma (3P) Distribution, Lognormal Distribution, and Gamma Distribution respectively.

Please be reminded that in goodness of fit tests, Type 2 error (failing to reject a false null hypothesis) rather than Type 1 (p-value: the probability of incorrectly rejecting the null hypothesis) is common.

#### 2.4.3. Discretization

In order not to disregard the effect of discretization on the performance of GLM model fitting, we prepared the discretized versions of our predictor variables.

**Table 1.** Goodness of fit test results for sales amount data between 2010 and 2014.

| | 2010 | 2011 | 2012 | 2013 | 2014 |
|---|---|---|---|---|---|
| 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 | 0.52422 | 0.66265 | 0.22956 |
| Reject? | no | no | no | no | no |
| Parameters | $\alpha = 1.6499$<br>$\beta = 3.5296$ | $\alpha = 0.70528$<br>$\beta = 12.935$ | $\alpha = 0.71381$<br>$\beta = 24.254$<br>$\gamma = 1.0$ | $\alpha = 1.2341$<br>$\mu = 2.1291$ | $\alpha = 0.79408$<br>$\beta = 19.553$ |

We calculated the quantiles for the temperature and price predictor variables. Then, we converted the raw values into categories with respect to first, second, third, and fourth quartiles (Table 2 and 3).

**Table 2.** Categorization of the temperature predictor variable.

| Category | Min | Max |
|---|---|---|
| Temperature 1 | 1° | 17° |
| Temperature 2 | 17° | 23° |
| Temperature 3 | 23° | 28.5° |
| Temperature 4 | 28.5° | 36° |

**Table 3.** Categorization of the price predictor variable.

| Category | Min | Max |
|---|---|---|
| Price 1 | 500.6380TRY* | 729.1154TRY |
| Price 2 | 729.1154TRY | 763.9032TRY |
| Price 3 | 763.9032TRY | 787.0951TRY |
| Price 4 | 787.0951TRY | 826.8073TRY |

* Turkish Lira

Finally, the date predictor variable is categorized with respect to quarters (Table 4):

**Table 4.** Categorization of the 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 |

## 3. Results

In our modeling, we investigated the response variable with respect to three predictor variables, "Date", "Temperature" and "Price". We considered all combinations of these three variables in our experiments. Additionally, we transformed them into categorical variables to test the effect of categorization. In the following part, we first give the empirical results for non-categorical predictor variables then for their categorical counterparts.

### 3.1. Incremental addition of non-categorical predictor variables

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 testing their effects on model fitting in a controlled way.

To assess the performance of model fit, $p$ values are calculated with respect to the goodness of fit null hypothesis. The null hypothesis can be stated as follows:

$H_0$ =There is no difference between observed and fitted values.

When we analyze the results given in Table 5, p values and AIC metric verify that the model which is based on "Day" and "Temperature" predictor variables gives the best result. The calculated $p$ value, 0.6185 means that we cannot reject the null hypothesis, that is, the model fitting is valid statistically.

### 3.2. Incremental addition of categorical predictor variables

This time, GLM fitting results are interpreted with categorical sales data for the year 2014. The results are presented in Table 6.

The model which is constructed with "Quarter", "Temperature", and "Price" gives the best result when we consider MSE, AIC and $p$ values. So, "Quarter", "Temperature", and "Price" are added to the model as predictor variables.

### 3.3. GLM with non-categorical vs. categorical predictors

Table 7 represents the comparison of GLM results with non-categorical and categorical predictors. The model which is constructed with the categorical independent variables gives better results. The difference between two models can be explained by the effect of the "Quarter" predictor variable. It gives seasonal information and is more relevant when the sale of cooling goods is considered.

**Table 7.** Comparison of GLM with non-categorical and categorical predictors.

| Metrics | Non-categorical | Categorical |
|---|---|---|
| MSE | 249.1760 | 195.1147 |
| AIC | 1506.7180 | 1482.727 |
| Residual deviance | 197.1618 | 164.2224 |
| P value | 0.6023 | 0.9572 |

**Table 5.** GLM fitting results for sales 2014 data.

| | | | | | **Sales2014(GLM)** | | | |
|---|---|---|---|---|---|---|---|---|
| 4***Coefficients** | **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 |

**Table 6.** GLM fitting results for categorical sales 2014 data.

| | | | | | **Sales2014(GLM)** | | | |
|---|---|---|---|---|---|---|---|---|
| 11***Coefficients** | **Quarter1** | 0 | 0 | 0 | **0** | 0 | 0 | 0 |
| | **Quarter2** | -0.0844 | 0 | 0 | -0.0655 | 0 | -0.0766 | -0.0624 |
| | **Quarter3** | -0.1358 | 0 | 0 | -0.1216 | 0 | -0.1226 | -0.1171 |
| | **Quarter4** | -0.0752 | 0 | 0 | -0.0596 | 0 | -0.0674 | -0.0598 |
| | **Temperature1** | 0 | -0.07702 | 0 | -0.03195 | -0.06598 | 0 | -0.029915 |
| | **Temperature2** | 0 | -0.09405 | 0 | -0.0234 | -0.07895 | 0 | -0.0192 |
| | **Temperature3** | 0 | -0.09385 | 0 | -0.0117 | -0.07875 | 0 | -0.0065 |
| | **Temperature4** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | **Price1** | 0 | 0 | -0.0697 | 0 | -0.04985 | -0.03679 | -0.036082 |
| | **Price2** | 0 | 0 | -0.07196 | 0 | -0.04867 | -0.0301 | 0.0238 |
| | **Price3** | 0 | 0 | 0.03676 | 0 | 0.04621 | 0.0606 | 0.0660 |
| | **Price4** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | **Intercept** | 0.1726 | 0.1429 | 0.1260 | 0.1771 | 0.1728 | 0.1897 | 0.1941 |
| **MSE** | | 229.5703 | 271.2096 | 280.8429 | 209.3177 | 259.6477 | 221.4104 | **195.1147** |
| **AIC** | | 1489.3760 | 1523.7140 | 1535.0460 | 1488.8610 | 1514.3070 | 1484.2230 | **1482.7270** |
| **Null deviance** | | 241.0813 | 241.0813 | 241.0813 | 241.0813 | 241.0813 | 241.0813 | 241.0813 |
| **Degrees of freedom** | | 206 | 206 | 206 | 206 | 206 | 206 | 206 |
| **P value** | | 0.0473 | 0.0473 | 0.0473 | 0.0473 | 0.0473 | 0.0473 | 0.0473 |
| **Residual deviance** | | 177.8493 | 205.5993 | 215.5636 | 172.9775 | 192.7054 | 169.5828 | **164.2224** |
| **Degrees of freedom** | | 203 | 203 | 203 | 200 | 200 | 200 | 197 |
| **P value** | | 0.8980 | 0.4358 | 0.2598 | 0.9168 | 0.6315 | 0.9420 | **0.9572** |

### 3.4. Ordinary least squares estimator (OLS) versus GLM

For all the combinations of predictor variables GLM outperforms OLS. In order to give an idea of how they differentiate from each other, the OLS and GLM fitted models which are constructed with "Day" and "Temperature" are visualized in Figure 2 and compared in Table 8.

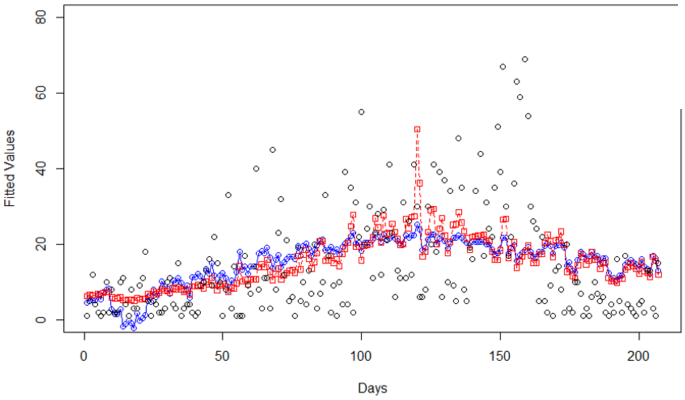

**Figure 2.** OLS vs. GLM.

The blue curved line shows the OLS fitting by adding two predictor variables which are "Day" and "Temperature" to the model. The red rectangular pointed line represents the GLM fitting result by the identical predictor variables. Real sales data are depicted as black rounded points. As seen from the figure, the OLS fitted values and GLM fitted values are similar.

In order to further analyze the difference between OLS and GLM model fitting, we can focus on their variances (Figure 3 and 4 respectively). The OLS method assumes that the residuals have the same variance which is named as homoscedasticity. Constant variance of the OLS method can be observed in Figure 3. GLM fitted model has non-constant variance across an entire range of values which is called heteroscedasticity (Figure 4). Fitting sales data by GLM provides less variance than the OLS method and GLM assumes different variances for each estimated response variable because of the error term.

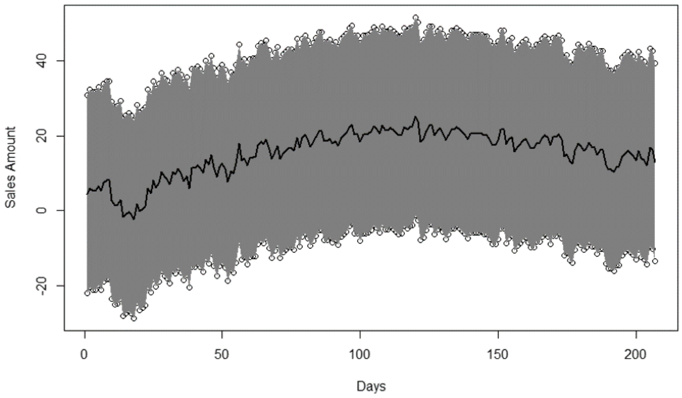

**Figure 3.** OLS variance.

Table 8 represents the comparison of OLS and GLM fitted models. GLM fitted model gives better AIC and MSE results. In addition to these, the residual to null deviance fraction gives a better result in the case of GLM. 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 the 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.

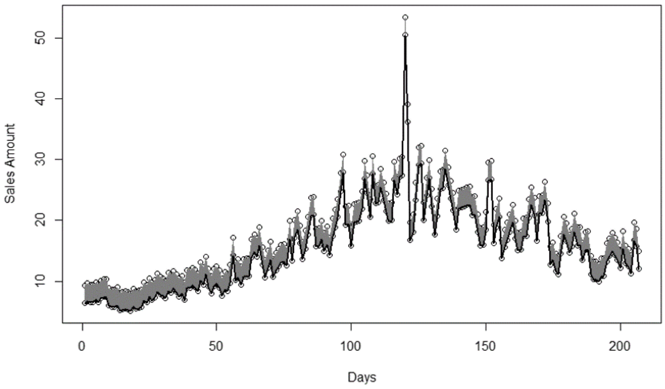

**Figure 4.** GLM variance.

**Table 8.** Comparison of OLS and GLM model fitting results.

| Metrics | OLS | GLM |
|---|---|---|
| | Days and Temperature | |
| MSE | 253.3723 | 250.5749 |
| AIC | 1741.157 | 1504.9 |
| Null deviance | 60025.91 | 240.95 |
| Residual deviance | 52448.07 | 197.32 |
| Residual to null deviance frac. | 0.8738 | 0.8189 |

### 3.5. Comparing GLM with predictive data mining methods

This section presents comparative results with selected data mining methods. As a comparison basis; Generalized Linear Models with Elastic Net Regularization (GLMNet) [7] , Gradient Boosting Method (GBM) [8], Principal Component Regression (PCR) [10], Support Vector Machine (SVM) [11], Random Forest (RF) [12], Conditional inference trees (CTree) [13], and Ensemble Learning (EL) [14] are used.

The validation of the selected model is one important step of model building. 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 set. Table 9 gives the comparison of the selected data mining methods and GLM with non-categorical predictor variables through the use of hold-out samples.

**Table 9.** Comparison of the selected data mining methods and GLM with non-categorical predictor variables.

| Non-categorical Sales2014 | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Factors** | **MSE of Methods** | | | | | | |
| **Day** | x | | | x | | x | x |
| **Temperature** | | x | | x | x | | x |
| **Price** | | | x | | x | x | x |
| **GLMNet** | 136.3917 | 136.6720 | 135.9715 | 136.6720 | 138.3373 | 135.9715 | 138.3373 |
| **GBM** | 169.6955 | **117.2606** | 121.9277 | 152.3425 | 114.9580 | 191.1965 | 147.7359 |
| **PCR** | **136.3344** | 155.5300 | 154.5678 | 136.4315 | 156.8812 | **135.8821** | 136.6662 |
| **SVM** | 188.3921 | 118.1168 | 141.2607 | 171.1725 | **106.7127** | 179.4780 | 165.0091 |
| **Random Forest** | 200.3054 | 188.3117 | **120.8555** | 201.0203 | 167.3700 | 180.3512 | 182.7320 |
| **Ctree** | 180.6332 | 150.7834 | 155.2848 | 151.0385 | 150.7834 | 180.6332 | 151.0385 |
| **Ensemble learning** | 177.8041 | 149.1145 | 134.1322 | 130.9450 | 154.6232 | 184.8372 | 128.5726 |
| **GLM** | 136.5303 | 214.8490 | 162.0401 | **122.3348** | 606.2935 | 136.3448 | **123.4321** |

**Table 10.** Comparison of the selected data mining methods and GLM with categorical predictor variables.

| Categorical Sales2014 | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Factors** | **MSE of Methods** | | | | | | |
| **Quarter** | x | | | x | | x | x |
| **Temperature** | | x | | x | x | | x |
| **Price** | | | x | | x | x | x |
| **GLMNet** | 235.9518 | 283.8525 | 287.758 | 235.6685 | 282.1223 | 236.6896 | 233.9616 |
| **GBM** | 1097.919 | 297.1452 | 500.7632 | 765.0678 | 445.0727 | 1232.807 | 725.4583 |
| **PCR** | 239.8311 | 289.2484 | 287.528 | 245.4711 | 280.0862 | 242.5003 | 240.7885 |
| **SVM** | 262.3833 | 323.2351 | 325.827 | 278.0332 | 321.076 | 275.4668 | 294.5758 |
| **Random Forest** | **215.8516** | **223.5914** | **239.4328** | **193.0158** | 279.3928 | 229.42 | 223.5908 |
| **Ctree** | 232.8376 | 275.8169 | 281.0322 | 220.7328 | 275.8169 | 232.8376 | 225.2724 |
| **Ensemble learning** | 237.558 | 285.4635 | 282.2418 | 218.3207 | 271.9933 | 230.829 | 241.1951 |
| **GLM** | 229.5703 | 271.2096 | 280.8429 | 209.3177 | **259.6477** | **221.4104** | **195.1147** |

As seen from Table 9, GLM gives the best result with "Days" and "Temperature" predictor variables and "Days", "Temperature", and "Price" predictor variables. GBM, PCR, SVM, and Random Forest models give better results in some other variable combinations (marked in bold).

Table 10 gives the comparison of the selected data mining methods and GLM with categorical predictor variables:

As a result, in some variable combinations the RF fitted model is the best while in the remaining ones the GLM fitted model outperforms the others. To further characterize those variable combinations, in single predictor variable cases RF is the best, but when "price" predictor variable is used in conjunction with other variable(s), GLM is the top performer. The best values in every column are marked in bold.

Meanwhile, as an alternative we used cross-validation on the model building based on Istanbul sales data. The obtained results are in accordance with the results from the hold-out sampling.

## 4. Discussion and Conclusion

To deal with demand prediction, various techniques of regression analysis and data mining are used under the predictive methods. The purpose of this work is to make history-based demand prediction of sales by using generalized linear models.

The data set which is used for analysis is real company data. In our modeling, the response variable is the sale amount and the date of sale, item price, and air temperature are selected as the predictor variables. The distribution of sale 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 GLM modeling.

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.

We analyzed the combinations of three different predictor variables, "Days", "Temperature", and "Price". Then, the predictor variables are transformed into their categorical counterparts for investigating the effect of categorization. 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. Within the scope of classification techniques, RF and GLM are the top performers. When single predictor variables are used RF is the best while in the case of the usage of "price" with any other predictor variable(s) GLM is the best. The model fitting results are evaluated with respect to MSE and AIC metrics.

Classifier performance depends greatly on the characteristics of the data to be classified. Various classification algorithms are compared in order to find out the characteristics of data that explain their comparative performances. However, it's still an open problem.

Attribute error and concept size are good features (characteristics of data) to explain the performance of classification algorithms. Concept size is the proportion of concept space covered by positive instances while attribute error is the random substitution of attribute values [15].

In explaining the performance of RF, it can be said that RF is robust to attribute error as it performs random selection of a subset of features to grow each tree.

As for GLM, its performance can be attributed to our model adaptation in which we took the dependent variable, "sale demand" distribution (Gamma) into consideration.

Further research on this topic could have the following directions:

- Different discretization methods for categorizing the predictor variables can be used.
- Hybrid models [16] can be constructed such as the formulation of the GLM model along with an additive time series component.

## References

[1] Nelder, J.A., Wedderburn, R.W.M. 1972. Generalized linear models. Journal of the Royal Statistical Society, Series A, General, 135, 370-384.

[2] Razzaghi, M. 2013. The Probit Link Function in Generalized Linear Models for Data Mining Applications. Journal of Modern Applied Statistical Methods, 12(19), 164-169.

[3] Tauras, J.A. 2005. An Empirical Analysis of Adult Cigarette Demand. Eastern Economic Journal, 31(3), 361-375.

[4] The Odum Institute, 2015. Logistic Regression and the American National Election Study 2012: Vote Choice in the 2012 US Presidential Election. The Odum Institute.

[5] Kutner, M.H., Nachtsheim, C., Neter, J. 2004. Applied linear regression models. McGraw-Hill/Irwin.

[6] Johnson, P. 2006. GLM with Gamma-Distributed Dependent Variables (Access Date: 28.05.2018.

[7] Friedman, J., Hastie, T., Tibshirani, R. 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, Articles, 33(1), 1-22.

[8] Schapire, R.E., Freund, Y. 2012. Boosting: Foundations and Algorithms. MIT Press.

[9] Pearson, K. 1900. X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philosophical Magazine, 50(302), 157-175.

[10] Jolliffe, I.T. 1982. A Note on the Use of Principal Components in Regression. Journal of the Royal Statistical Society. Series C (Applied Statistics), 31(3), 300-303.

[11] Cortes, C., Vapnik, V. 1995. Support-Vector Networks. Mach. Learn., 20(3), 273-297.

[12] Breiman, L. 2001. Random Forests. Mach. Learn., 45(1), 5-32.

[13] Hothorn, T., Hornik, K., Zeileis, A. 2006. Unbiased Recursive Partitioning: A Conditional Inference Framework. Journal of Computational and Graphical Statistics, 15(3), 651-674.

[14] Zhou, Z. 2012. Ensemble Methods: Foundations and Algorithms. 1st. MIT Press. Chapman & Hall/CRC.

[15] Cohen, P.R. 1995. Empirical Methods for Artificial Intelligence. MIT Press, Cambridge, MA, USA.

[16] Bensoussan, A., Bertrand, P., Brouste, A. 2014. A generalized linear model approach to seasonal aspects of wind speed modeling. Journal of Applied Statistics, 41(8), 1694-1707.