J.E. Cairnes School of Business & Economics

NUI Galway
OÉ Gaillimh

Semester 2 Alternative Assessment 2019/20

End-of-Module Assignment

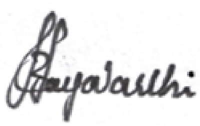
MS5108 Applied Customer Analytics

Student Name:	Jayakarthi Boovendran
Student ID:	19230487
Course:	MBY – MSc Business Analytics
Email:	J.Boovendran1@nuigalway.ie
Assignment Topic:	Generalised Linear Models in R

Declaration

"In submitting this work, I confirm that it is entirely my own. I acknowledge that I may be invited to undertake an online interview if there is any concern in relation to the integrity of my submission."

Signature: **Date:** 12 May 2020



Jayakarthi Boovendran

Generalized Linear Models in R

I. Dataset Information

- **Dataset**: House Sales in King County, USA
- Dataset Available @ https://www.kaggle.com/harlfoxem/housesalesprediction
- **Dataset Description**: The dataset includes house sales data in King County from May 2014 to May 2015. Some of the key attributes are Square feet living area, no. of bathrooms, no. of bedrooms, no. of floors, Grade and Condition. These attributes influence the price of the house.
- **No. of Transactions**: 21613 records; **No. of Data points** (attributes): 21
- **No. of Missing value records**: 0

II. Implementation of GLM

a) R Code

```
#Package Installation
install.packages("ggplot2")
install.packages("HistData")
install.packages('corrplot')
install.packages('caret')
install.packages('caTools')
install.packages('car')
install.packages('Metrics')
install.packages('e1071')

#Load Required Packages
library('ggplot2')
library('HistData')
library('corrplot')
library('MASS')
library('caret')
library('caTools')
library('car')
library('aws.s3')
library('Metrics')
library('e1071')
```

```r
# Loading Data
# The coding part is done using IBM Watson Studio's R Notebook
# The following code loads the dataset (kc_house_data.csv) from IBM Cloud Object Storage.
library('aws.s3')
obj <- get_object(
      object = "kc_house_data.csv",
      bucket = "generalizedlinearregression-donotdelete-pr-eh55msxy9iefdu",
      key = "ead47ea2fd854c4195760dec16c0846f",
      secret = "5abeb89065ef7fcbe1e8e1f16ea0689c381ecda0cf1ba735",
      check_region = FALSE,
      base_url = "s3.eu-geo.objectstorage.service.networklayer.com")

# The file is loaded into a raw vector & is processed using rawToChar()
house_sales_data <-  read.csv(text = rawToChar(obj))
head(house_sales_data)
```

```r
# Missing Data Handling
# Records with missing data are ignored from further processing
sapply(house_sales_data,function(x) sum(is.na(x)))
numberOfNA = length(which(is.na(house_sales_data) == T))
cat('No. of records with missing values: ', numberOfNA)
if(numberOfNA > 0)
{
  cat('\nRemoving missing values...')
  house_sales_data = house_sales_data [complete.cases(house_sales_data), ]
}
# No missing data records in the dataset;   # No. of records with missing values:0
```

```r
# Feature Selection
# Predictor Variables: sqft_living, bathroom
# Target Variable: price
```

```r
#Correlation Analysis
#Testing relationship between the predictor and target variables

#Scatter plot with regression line for sqft_living and price
ggplot(house_sales_data, aes(x=log(price), y=sqft_living)) +
geom_point(shape=18, color="blue")+
geom_smooth(method=lm, se=FALSE,color="red")+
xlab("Price(USD)")+
ylab('sqft_living')+
labs(title = "sqft_living  vs  Price",
subtitle='Positive Correlation Between sqft_living  &  Price')

#Pearson’s Test for Correlation
#Positive Correlation between sqft_living and price (p<0.05 , r=0.702)
cor.test(house_sales_data$price,house_sales_data$sqft_living,use='complete.obs',method=
'pearson')



#Scatter plot with regression line for bathrooms and price
ggplot(house_sales_data, aes(x=log(price), y=bathrooms)) +
geom_point(shape=18, color="blue")+
geom_smooth(method=lm, se=FALSE,color="red")+
xlab("Price(USD)")+
ylab('Bathrooms')+
labs(title = "Bathrooms  vs  Price",
subtitle='Positive Correlation Between Bathrooms  & Price')

#Pearson’s Test for Correlation
#Positive Correlation between sqft_living and price (p<0.05 , r=0.525)
cor.test(house_sales_data$price,house_sales_data$bathrooms,use='complete.obs',method
='pearson')


#Since both the variables independent variables have significant relationship with the target
#They can be used to build prediction models
```

```r
# Generalized linear models

# Model1 : price ~ sqft_living
# Since 'price' is numeric, gaussian link function is used instead of binomial ()
# p-value=0; Since p<0.05, the model has a very high significance in predicting the target
model1=glm(house_sales_data$price ~ house_sales_data$sqft_living, family=gaussian(link='
identity'))
summary(model1)
confint(model1, parm = "house_sales_data$sqft_living")
exp(coef(model1)["house_sales_data$sqft_living"])


# Model2 : price ~ sqft_living + bathrooms
# Since 'price' is numeric, gaussian link function is used instead of binomial ()
# p-value=0.14; Since p>0.05, the model has no significance in predicting the target
model2=glm(house_sales_data$price ~ house_sales_data$sqft_living+house_sales_data$ba
throoms, family=gaussian(link='identity'))
summary(model2)


# Comparing model1 and model2 using ANOVA
# p-value=0.14; Since p>0.05, there is no statistical significance between the models
# model 1 offers high predictive power than model 2
anova(model1,model2,test='Chisq')

prob<-predict(model1,type="response")
plot(sqft_living ~bathrooms,
   data = house_sales_data,
   pch = ".")
symbols(house_sales_data$sqft_living,
     house_sales_data$bathrooms,
     circles = prob,
     add = TRUE)
```

b) R Notebook

The given link contains the R code mentioned above along with their outputs. In addition to that, the Notebook also includes the prediction of the target variable using the selected model. IBM Watson Studio - R Notebook

III. Output

a) Sample Data

id	date	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition
7129300520	20141013T000000	3	1.00	1180	5650	1	0	0	3
6414100192	20141209T000000	3	2.25	2570	7242	2	0	0	3
5631500400	20150225T000000	2	1.00	770	10000	1	0	0	3
2487200875	20141209T000000	4	3.00	1960	5000	1	0	0	5
1954400510	20150218T000000	3	2.00	1680	8080	1	0	0	3
7237550310	20140512T000000	4	4.50	5420	101930	1	0	0	3

b) Feature Selection

sqft_living	bathrooms	price
1180	1.00	221900
2570	2.25	538000
770	1.00	180000
1960	3.00	604000
1680	2.00	510000
5420	4.50	1225000

Predictor variables: sqft_living , bathrooms

Target variable : price

c) Correlation Analysis

- ***Correlation analysis between 'sqft_living' and 'price'***

```
        Pearson's product-moment correlation

data:  house_sales_data$price and house_sales_data$sqft_living
t = 144.92, df = 21611, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6952099 0.7087336
sample estimates:
      cor
0.7020351
```

P< 0.05 and r =0.702 (+ve). Therefore, Statistically Significant Positive Correlation

Figure : Pearson's Correlation Analysis between sqft_living and price

- ***Correlation analysis between 'bathrooms' and 'price'***

```
        Pearson's product-moment correlation

data:  house_sales_data$price and house_sales_data$bathrooms
t = 90.714, df = 21611, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5154140 0.5347258
sample estimates:
      cor
0.5251375
```

P< 0.05 and r =0.525 (+ve). Therefore, Statistically Significant Positive Correlation

Figure : Pearson's Correlation Analysis between bathrooms and price

- *Scatter plots with regression line conforms the positive correlation*

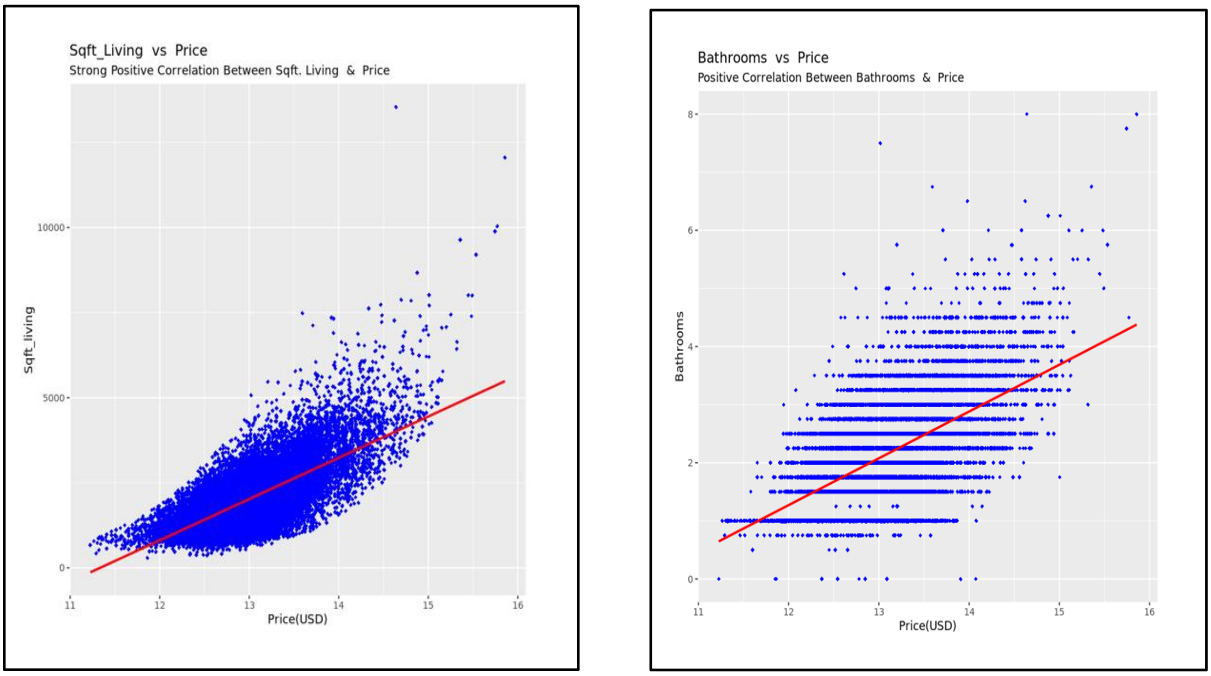


Figure : Scatter plot with regression line between the predictor and target variables

d) Building Generalized Linear Models

- ***Model 1 :*** *price ~ sqft_living*

```
Call:
glm(formula = house_sales_data$price ~ house_sales_data$sqft_living,
    family = gaussian(link = "identity"))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1476062   -147486    -24043    106182   4362067

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                 -43580.743   4402.690  -9.899   <2e-16 ***
house_sales_data$sqft_living    280.624      1.936 144.920   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 68357612435)

    Null deviance: 2.9129e+15  on 21612  degrees of freedom
Residual deviance: 1.4773e+15  on 21611  degrees of freedom
AIC: 600541

Number of Fisher Scoring iterations: 2
```

Figure : Summary of model 1 - price ~ sqft_living

- ***Model 2 :*** *price ~ sqft_living + bathrooms*

```
Call:
glm(formula = house_sales_data$price ~ house_sales_data$sqft_living +
    house_sales_data$bathrooms, family = gaussian(link = "identity"))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1483123   -147387    -24190    105951   4359876

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                 -39456.614   5223.129  -7.554 4.38e-14 ***
house_sales_data$sqft_living    283.892      2.951  96.194  < 2e-16 ***
house_sales_data$bathrooms    -5164.600   3519.452  -1.467    0.142
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 68353964336)

    Null deviance: 2.9129e+15  on 21612  degrees of freedom
Residual deviance: 1.4771e+15  on 21610  degrees of freedom
AIC: 600540

Number of Fisher Scoring iterations: 2
```

Figure : Summary of model 2 - price ~ sqft_living + bathrooms

e) Comparing the Models using ANOVA

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
21611	1.477276e+15	NA	NA	NA
21610	1.477129e+15	1	1.47193e+11	0.1422551

Figure : ANOVA results of model1 ~ model2

IV. Discussion

In predictive analytics, it is quintessential to use the predictor variables that statistically influence the predictive power of the model. Choosing an independent variable that correlates with the target variable highly improves the prediction.

In this study, we try to predict the sales price of houses based on the independent variables 'sqft_living' and 'bathrooms'. Therefore, to analyze the relationship between the independent and dependent variables, also, to assess the strength of the predictor variables, we make use of Generalized linear models.

As an initial step, we performed a correlation analysis between the dependent and independent variables to ensure whether there exists a relationship between them. The visual interpretation of scatter plots, along with the results of Pearson's correlation test proves that both the independent variables have a statistically significant positive correlation with the target variable. (sqft_living : $p<0.05$ and $r=0.703$; bathrooms: $p<0.05$ and $r=0.525$) Therefore, we can use them for building prediction models.

As a second step, we built a GLM (model 1) with one independent variable (sqft_living) predicting the target variable (price). The model generated a p-value of 0. The value of $p<0.05$ indicates that the predictor variable sqft_living has high statistical significance in predicting the target.

Next, we built a GLM (model 2) with two independent variables (sqft_living+ bathrooms) predicting the target (price). The value of $p>0.05$ confirms that the newly added predictor is not significantly associated with the target; therefore, it does not improve prediction.

Finally, we performed ANOVA test to check whether there is a significant difference between the models, The hypothesis would be,

- H_0 – There is no significant difference between the models (Null Hypotheses)
- H_1 – One model statistically outperforms the other (Alternate Hypotheses)

The Chi Square value of $p > 0.05$ ($p=0.14$) confirms that there is no significant difference between the models (Null Hypotheses is accepted). Further, applying model 2 does not offer a better prediction than model1. To be precise, in case of model 2 , the independent variables do not interact to predict the target. Therefore using a more simplistic model (model 1) expedites better prediction.