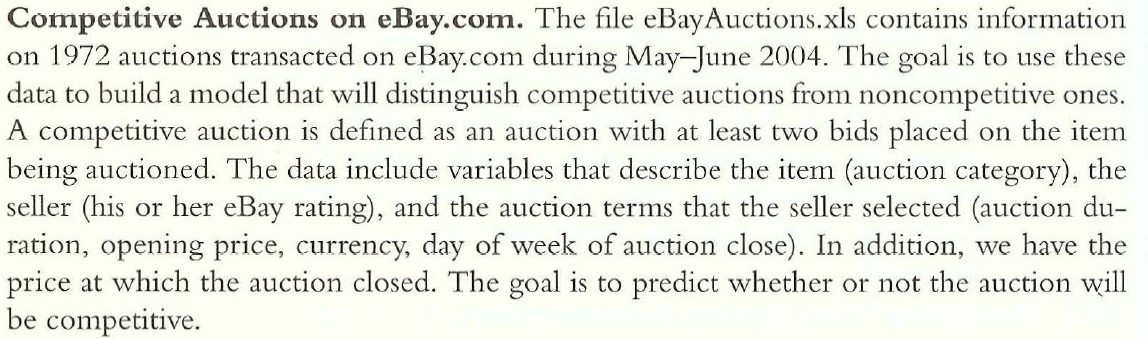
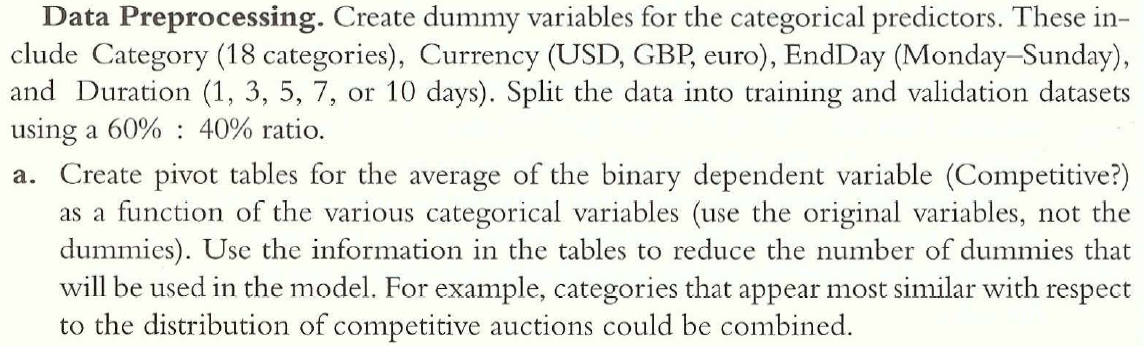
**Generalized Linear Model**

Name – Aman Chauhan

Student ID – 200208218

Unity ID – achauha3





Packages – readxl, dummies, caret, e1071, dplyr

Code –

library("readxl")

library("dummies")

library("caret")

library("e1071")

library("dplyr")

set.seed(13)

# load data

data <- as.data.frame(read\_excel("eBayAuctions.xls", sheet = 1))

# split into training and testing

train\_index <- createDataPartition(data$`Competitive?`, p=0.6, list=FALSE)

data.train <- data[train\_index,]

data.test <- data[-train\_index,]

# pivot tables for analysis

categories <- summarise(group\_by(data.train, Category), mean\_competitive=mean(`Competitive?`))

currencies <- summarise(group\_by(data.train, currency), mean\_competitive=mean(`Competitive?`))

durations <- summarise(group\_by(data.train, Duration), mean\_competitive=mean(`Competitive?`))

endingDays <- summarise(group\_by(data.train, endDay), mean\_competitive=mean(`Competitive?`))

Initial Pivot Tables –

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Category | Antique/Art/Craft | Automotive | Books | Business/Industrial | Clothing/Accessories | Coins/Stamps | Collectibles | Computer | Electronics |
| mean\_competitive | 0.5208 | 0.3333 | 0.5185 | 0.8 | 0.5256 | 0.3333 | 0.5617 | 0.6666 | 0.8 |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Category | EverythingElse | Health/Beauty | Home/Garden | Jewelry | Music/Movie/Game | Photography | Pottery/Glass | SportingGoods | Toys/Hobbies |
| mean\_competitive | 0.1111 | 0.2 | 0.6774 | 0.4667 | 0.5702 | 0.7778 | 0.3 | 0.6988 | 0.5071 |

|  |  |  |  |
| --- | --- | --- | --- |
| currency | EUR | GBP | US |
| mean\_competitive | 0.53184713 | 0.6746988 | 0.51969504 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| endDay | Fri | Mon | Sat | Sun | Thu | Tue | Wed |
| mean\_competitive | 0.46242775 | 0.6918239 | 0.42173913 | 0.48484848 | 0.5826087 | 0.48598131 | 0.46511628 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Duration | 1 | 3 | 5 | 7 | 10 |
| mean\_competitive | 0.52941176 | 0.40441176 | 0.70138889 | 0.47763864 | 0.53804348 |

Based on the pivot table values, lets combine the following categories –

1. Combine electronics and sporting goods
2. Combine collectibles and home/garden
3. Combine music/movie/game, computer and business/industrial
4. Combine toys/hobbies and antique/art/craft
5. Combine clothing/accessories and books
6. Combine pottery/glass, coins/stamps into everythingElse
7. Combine Euro and US

Final Pivot Tables –

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Category | Automotive | Clothing/Accessories/Books | Collectibles/Home/Garden | Electronics/SportingGoods | EverythingElse | Health/Beauty | Jewelry | Music/Movie/Game/Computer/Business/Industrial | Photography | Toys/Hobbies/Antique/Art/Craft |
| mean\_competitive | 0.3333 | 0.5238 | 0.5937 | 0.7288 | 0.275 | 0.2 | 0.4666 | 0.5842 | 0.7778 | 0.5127 |

|  |  |  |
| --- | --- | --- |
| currency | EUR/US | GBP |
| mean\_competitive | 0.52316076 | 0.6746988 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Duration | 1 | 3 | 5 | 7 | 10 |
| mean\_competitive | 0.52941176 | 0.40441176 | 0.70138889 | 0.47763864 | 0.53804348 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| endDay | Fri | Mon | Sat | Sun | Thu | Tue | Wed |
| mean\_competitive | 0.46242775 | 0.6918239 | 0.42173913 | 0.48484848 | 0.5826087 | 0.48598131 | 0.46511628 |

These will be the final dummy variables.

Code –

data.train<-dummy.data.frame(data.train,names=c("Category","currency","Duration","endDay"), sep="\_")

data.train<-data.train[colnames(data.train)]

data.test<-dummy.data.frame(data.test,names=c("Category","currency","Duration","endDay"),sep="\_")

data.test<-data.test[colnames(data.test)]

For the problem above, build the logistic regression model (fit.all) using all the predictors and answer the following questions by including the corresponding R code and showing all the required mathematical derivations used to answer these questions –

1. Let Xh be the predictor with the highest estimate (in terms of its absolute value) for its regression coefficient. Build a single predictor logistic regression model (fit.single) using Xh as the predictor. Write the equations relating the dependent variable (Response) to the explanatory variable in terms of –

Xh = currency\_EUR/US

* 1. Probabilities -
  2. Odds -
  3. Logit -

1. Write the estimated equation for the fit.all model in all three formats (if the number of predictors is more than four, then include only those four predictors whose absolute value estimates are the highest) –

The predictors are - currency\_EUR/US (-1.819), Category\_EverythingElse (-1.641),

Category\_Health/Beauty (-1.410), Category\_Clothing/Accessories/Books (-1.080)

* 1. Logit -
  2. Odds -
  3. Probability -

1. Let Xh be the predictor with the highest estimate (in terms of its absolute value) for its regression coefficient in the fit.all. Compute the odds ratio that estimated a single unit increase in Xh, holding the other predictors constant. Provide the interpretation for this regression coefficient. If it were a linear regression model, how would the interpretation change for a single unit increase in Xh.

X1 = currency\_EUR/US (-1.819)

X2 = Category\_EverythingElse (-1.641)

X3 = Category\_Health/Beauty (-1.41)

X4 = Category\_Clothing/Accessories/Books (-1.08)

This means that for a unit increase in the EUR/US currency, the odds decrease by a factor of 6.1657. The odds decrease because the coefficient is negative. In general, if the coefficient is positive, then odds increase exponentially on unit increase of predictor. If the coefficient is negative, then odds decrease exponentially on unit increase of predictor.

For linear regression,

So, for a unit increase in predictor, the response will decrease by a value of |.

1. Build a reduced logistic regression model (fit.reduced) using only the predictors that are statistically significant. Assess if the reduced model is equivalent to the full model. Justify your answer.

The predictors with significant coefficients (α≤0.1) from fit.all are –

Duration\_5 (0.5545)

Category\_Clothing/Accessories/Books (-0.9523)

Category\_EverythingElse (-1.312)

Category\_Health/Beauty (-1.182)

currency\_EUR/US (-0.8785)

sellerRating (0.0000315)

endDay\_Mon (0.4505)

ClosePrice (0.08237)

OpenPrice (-0.1122)

|  |  |
| --- | --- |
| Fit.reduced (validation data) | Fit.all (validation data) |
| Reference  Prediction 0 1  0 302 107   1. 65 314   Accuracy : 0.7817  95% CI : (0.7512,0.8101)  F1-Score = 0.7850  Null deviance: 1631.9 on 1183 degrees of freedom  Residual deviance: 1235.3 on 1174 degrees of freedom  AIC: 1255.3 | Reference  Prediction 0 1  0 277 97   1. 90 324   Accuracy : 0.7627  95% CI : (0.7314,0.792)  F1-Score = 0.7760  Null deviance: 1631.9 on 1183 degrees of freedom  Residual deviance: 1217.0 on 1160 degrees of freedom  AIC: 1265 |
| ANOVA between Fit.reduced and Fit.all  Resid. Df Resid. Dev Df Deviance Pr(>Chi)  1 1160 1217.0  2 1174 1235.3 -14 -18.32 0.1926 | |

If you compare different metrics, we can see most of the metrics for both reduced and full model are quite similar. Reduced and Fit are quite similar as seen from the non-significant p-value chi-square ANOVA test. Also, reduced model provides a better F1-score and better accuracy. Thus, we can say reduced model is equivalent to the full model, and provides better results.

1. Compute the dispersion of your model and run the dispersion diagnostic test. If the constructed model is overdispersed, then discuss the ways to deal with the issue.

From the summary of fit.reduced model,

Residual deviance: 1235.3 on 1174 degrees of freedom.

Overdispersion = Residual deviance/Residual df = 1.05 > 1.

Hence there is no overdispersion in the data.

If overdispersion were to be present, we could have removed by using the quasi-binomial family instead of binomial family.