CSC 591 - ADBI
Unity ID - djgadre
Homework 1 - GLM

1) Single Predictor Regression Model

Predictor with highest estimate = Open Price
Let Open Price = X

a. Probabilities

$\beta 0 = 1.544$ $\beta 1 = -0.440$ n = 1972

$$p = \frac{1}{1 + e^{[-(\beta 0 - \beta 1 X)]}}$$

$$p = \frac{1}{1 + e^{[-(1.544 - 0.440X)]}}$$

$$\mu_y = np$$

$$\mu_y = \frac{1972}{1 + e^{[-(1.544 - 0.440X)]}}$$

b. Odds

$$\text{Odds} = \frac{\mu_y}{1 - \mu_y}$$

$$= \frac{1972}{e^{[-(1.544 - 0.440X)]} - 1971}$$

c. Logit

$$\text{Logit} = \log \frac{\mu_y}{1 - \mu_y}$$

$$= \log \frac{1972}{e^{[-(1.544 - 0.440X)]} - 1971}$$

$$= \log 1972 - \log \left(e^{[-(1.544 - 0.440X)]} - 1971\right)$$

$$= 7.5868 - \log \left(e^{[-(1.544 - 0.440X)]} - 1971\right)$$

2) **All Predictor Regression Model**

Predictors with highest absolute value estimates are Open Price, Close Price, Category_Category6, currency_Currency2.

Let Open Price = X1, Close Price = X2, Category_Category6 = X3, currency_Currency2 = X4.

a. Probabilities

$\beta 0 = 0.328 \quad \beta 1 = -1.014 \quad \beta 2 = 0.7898 \quad \beta 3 = 0.5528 \quad \beta 4 = 0.434$

$z = 0.328 - 1.014X1 + 0.7898X2 + 0.5528X3 + 0.434X4$

$$p = \frac{1}{1 + e^{-z}}$$

$$p = \frac{1}{1 + e^{[-(0.328 - 1.014X1 + 0.7898X2 + 0.5528X3 + 0.434X4)]}}$$

$\mu_y = np$

$$\mu_y = \frac{1972}{1 + e^{[-(0.328 - 1.014X1 + 0.7898X2 + 0.5528X3 + 0.434X4)]}}$$

b. Odds

$$Odds = \frac{\mu_y}{1 - \mu_y}$$

$$= \frac{1972}{e^{[-(0.328 - 1.014X1 + 0.7898X2 + 0.5528X3 + 0.434X4)]} - 1971}$$

c. Logit

$$Logit = \log \frac{\mu_y}{1 - \mu_y}$$

$$= \log \frac{1972}{e^{[-(0.328 - 1.014X1 + 0.7898X2 + 0.5528X3 + 0.434X4)]} - 1971}$$

$$= \log 1972 - \log \left(e^{[-(0.328 - 1.014X1 + 0.7898X2 + 0.5528X3 + 0.434X4)]} - 1971 \right)$$

$$= 7.5868 - \log \left(e^{[-(0.328 - 1.014X1 + 0.7898X2 + 0.5528X3 + 0.434X4)]} - 1971 \right)$$

3) **Odds Ratio**

The predictor with the highest estimate is Open Price.

$$\text{Odds Ratio} = \frac{\frac{1972}{e^{[-(0.328 - 1.014(X1+1) + 0.7898X2 + 0.5528X3 + 0.434X4)]} - 1971}}{\frac{1972}{e^{[-(0.328 - 1.014X1 + 0.7898X2 + 0.5528X3 + 0.434X4)]} - 1971}}$$

$$= \frac{e^{[-(0.328 - 1.014X1 + 0.7898X2 + 0.5528X3 + 0.434X4)]} - 1971}{e^{[-(0.328 - 1.014(X1+1) + 0.7898X2 + 0.5528X3 + 0.434X4)]} - 1971}$$

$$= \frac{e^{[-(0.328 - 1.014X1 + 0.7898X2 + 0.5528X3 + 0.434X4)]} - 1971}{e^{[-(0.682 - 1.014X1 + 0.7898X2 + 0.5528X3 + 0.434X4)]} - 1971}$$

$$= e^{coefficient\ of\ highest\ estimate\ predictor}$$

$$= e^{coefficient\ of\ Open\ Price}$$

$$= e^{-1.014}$$

$$= 0.3627$$

One unit increase in a predictor variable changes the log odds of response by the coefficient of predictor variable in the equation of logistic regression. By exponentiation, the odds ratio is $e^{coefficient\ of\ predictor\ with\ increased\ value}$.

The regression coefficients give the change in log(odds) in the response of a unit change in the predictor variable, holding all the other predictors constant.

For linear regression, the regression coefficients give the change in value of response variable for a unit change in the predictor variable, holding all other predictors constant.

4) **Reduced Logistic Regression Model**

Statistically significant predictors considered for reduced model are 'sellerRating', 'endDay_Day3', 'endDay_Day2', 'Category_Category2', 'endDay_Day1', 'Category_Category7', 'currency_Currency2', 'Category_Category6', 'ClosePrice', 'OpenPrice'.

Accuracy of model built with all predictors = 0.8035
Accuracy of reduced model = 0.8074

The reduced model is equivalent to the full model as both have nearly same accuracy. The insignificant parameters were discarded to build the the reduced model. As the accuracy of both models remains the same, it proves that the discarded parameters do not affect the response variable i.e. they are insignificant.

5) Dispersion of Model

Expected Variance of Model = 0.2482

[Expected Variance is computed using *df['Competitive?'].var()*]

Observed Variance of Model = 0.2472

[Observed Variance is computed using *y_pred_reduced.var()*]

$$Ø = \frac{Observed\ Variance}{Expected\ Variance} = \frac{0.2472}{0.2482} = 0.996$$

The Logistic Regression model is not overdispersed as $Ø < 1$.