Appendix A: One purchase model process construction

One purchase model process construction

Pau Casas

Preliminary actions

Execute libraries

```
library(tidyverse)
library(gridExtra)
library(ResourceSelection)
library(ROCR)
```

Upload masterCustomers and univariate analysis for the one purchase's model datset

```
load("masterCustomers1pm.Rda")
load("univariate1pm.Rda")
```

Create testing and training dataset

```
smp_size <- floor(nrow(masterCustomers.1pm)*0.75)
set.seed(321)
train_ind <- sample(seq_len(nrow(masterCustomers.1pm)), size = smp_siz
e)
masterCustomers.1pm.train <- masterCustomers.1pm[train_ind, ]
masterCustomers.1pm.test <- masterCustomers.1pm[-train_ind, ]
```

Variable selection

1. Univariate Analysis

We use the following code to check the "p-value of inclusion"" for each variable at the Univariate stage.

```
p.value.univ.market <- pchisq(univ.market.1pm$null.deviance-univ.marke
t.1pm$deviance, univ.market.1pm$df.null-univ.market.1pm$df.residual, l
ower.tail = FALSE)

print(paste("p-value of inclusion for the market's univariate analysis
is", p.value.univ.market))

## [1] "p-value of inclusion for the market's univariate analysis is 2
.68659533369243e-32"
```

Below it can be seen the "p-value of inclusion"" for all the variables of the One purchase model

```
##              variable   p.value
## 1              market 0.0000000
## 2          digita_cli 0.0011377
## 3           fp.season 0.0403922
## 4            fp.month 0.1365230
```

```
## 5      firstp.entropy 0.0002319
## 6        firstp.lines 0.0000002
## 7     firstp.families 0.4112822
## 8           same.zone 0.0000000
## 9                dist 0.0000078
## 10 log10(totalvalue). 0.0000000
## 11           any.disc 0.2860217
```

2. First multivariate model

It is entered to the first multivariate analysis all the variables but the ones that have a "p-value of inclusion" in the Univariate analysis greater than 0.25: firstp.families and any.disc

```
multiv1pm.1 <- glm(good.customer ~ market + digita_cli + fp.season + f
p.month + firstp.entropy + firstp.lines + same.zone + dist + log10(tot
alvalue),
                   data=masterCustomers.1pm.train,
                   family="binomial")

summary(multiv1pm.1)

##
## Call:
## glm(formula = good.customer ~ market + digita_cli + fp.season +
##     fp.month + firstp.entropy + firstp.lines + same.zone + dist +
##     log10(totalvalue), family = "binomial", data = masterCustomers.
1pm.train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6099  -0.4424  -0.3502  -0.2878   4.6248
##
## Coefficients: (3 not defined because of singularities)
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -5.0224911  0.3104415 -16.179  < 2e-16 ***
## marketOEM          1.2898881  0.2256331   5.717 1.09e-08 ***
## marketSI           2.0711484  0.3437488   6.025 1.69e-09 ***
## digita_cliS        0.4230530  0.1822436   2.321   0.0203 *
## fp.seasonspring   -0.3543394  0.2715598  -1.305   0.1920
## fp.seasonsummer   -0.0404442  0.2772523  -0.146   0.8840
## fp.seasonwinter    0.3278046  0.2415090   1.357   0.1747
## fp.monthfebruary  -0.1648448  0.2318668  -0.711   0.4771
## fp.monthmarch      0.3117523  0.2639382   1.181   0.2375
## fp.monthapril      0.5431959  0.2852367   1.904   0.0569 .
## fp.monthmay               NA         NA      NA       NA
## fp.monthjune       0.3019874  0.2681785   1.126   0.2601
## fp.monthjuly      -0.0396707  0.2757202  -0.144   0.8856
## fp.monthaugust            NA         NA      NA       NA
## fp.monthseptember  0.2793342  0.2445496   1.142   0.2534
## fp.monthoctober   -0.1658312  0.2664510  -0.622   0.5337
## fp.monthnovember          NA         NA      NA       NA
## fp.monthdecember  -0.0229157  0.2469257  -0.093   0.9261
## firstp.entropy    -0.2147923  0.1535017  -1.399   0.1617
```

```
## firstp.lines       0.0457929  0.0213558   2.144   0.0320 *
## same.zoneTRUE     -0.2709428  0.1455544  -1.861   0.0627 .
## dist              -0.0002062  0.0001433  -1.439   0.1501
## log10(totalvalue)  1.6926636  0.1083972  15.615  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3128.4  on 4401  degrees of freedom
## Residual deviance: 2614.0  on 4382  degrees of freedom
## AIC: 2654
##
## Number of Fisher Scoring iterations: 5
```

Three fp.month coefficients returns NA values due its correlation with fp.season variable. fp.month is then removed from the model.

```
multiv1pm.2 <- glm(good.customer ~ market + digita_cli + fp.season + f
irstp.entropy + firstp.lines + same.zone + dist + log10(totalvalue),
                   data=masterCustomers.1pm.train,
                   family="binomial")

summary(multiv1pm.2)

##
## Call:
## glm(formula = good.customer ~ market + digita_cli + fp.season +
##     firstp.entropy + firstp.lines + same.zone + dist + log10(totalv
alue),
##     family = "binomial", data = masterCustomers.1pm.train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5183  -0.4419  -0.3468  -0.3002   4.6271
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -4.9516929  0.2696649 -18.362  < 2e-16 ***
## marketOEM          1.2647944  0.2237678   5.652 1.58e-08 ***
## marketSI           2.0133176  0.3427869   5.873 4.27e-09 ***
## digita_cliS        0.4278416  0.1821461   2.349   0.0188 *
## fp.seasonspring   -0.1312306  0.1504669  -0.872   0.3831
## fp.seasonsummer    0.0052868  0.1475173   0.036   0.9714
## fp.seasonwinter    0.2106368  0.1427802   1.475   0.1401
## firstp.entropy    -0.2187112  0.1534079  -1.426   0.1540
## firstp.lines       0.0448769  0.0215449   2.083   0.0373 *
## same.zoneTRUE     -0.2758558  0.1451315  -1.901   0.0573 .
## dist              -0.0001935  0.0001420  -1.363   0.1729
## log10(totalvalue)  1.6871315  0.1080967  15.608  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 3128.4  on 4401  degrees of freedom
## Residual deviance: 2623.7  on 4390  degrees of freedom
## AIC: 2647.7
##
## Number of Fisher Scoring iterations: 5
```

3. Second multivariate model

firstp.entropy, same.zone, dist and fp.season are removed from the model as their "p-value of inclusion" is greater than 0.05.

It has also been checked the p-value of a new variable called "fp.seasonwinter" which is TRUE if the customers have born in winter and FALSE in any other case. It has been tested due its low p-value in the Wald test of the Univariate analysis. "P-value of inclusion" is 0.03 so it is added to the model.

Creation of the new variable:

```
masterCustomers.1pm.train <- masterCustomers.1pm.train %>%
  mutate(fp.seasonwinter=ifelse(fp.season=="winter",TRUE, FALSE))
masterCustomers.1pm.test <- masterCustomers.1pm.test %>%
  mutate(fp.seasonwinter=ifelse(fp.season=="winter",TRUE, FALSE))
```

```
multiv1pm.2 <- glm(good.customer ~ market + digita_cli + fp.seasonwint
er + firstp.lines +  log10(totalvalue),
                   data=masterCustomers.1pm.train,
                   family="binomial")

summary(multiv1pm.2)
```

```
##
## Call:
## glm(formula = good.customer ~ market + digita_cli + fp.seasonwinter
+
##     firstp.lines + log10(totalvalue), family = "binomial", data = m
asterCustomers.1pm.train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5319  -0.4432  -0.3483  -0.3076   4.6834
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -5.34732    0.19624 -27.249  < 2e-16 ***
## marketOEM            1.33384    0.22249   5.995 2.03e-09 ***
## marketSI             2.07360    0.33933   6.111 9.91e-10 ***
## digita_cliS          0.44711    0.18195   2.457   0.0140 *
## fp.seasonwinterTRUE  0.24749    0.11565   2.140   0.0324 *
## firstp.lines         0.03019    0.01903   1.586   0.1126
## log10(totalvalue)    1.73566    0.10162  17.079  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3128.4  on 4401  degrees of freedom
## Residual deviance: 2630.8  on 4395  degrees of freedom
## AIC: 2644.8
##
## Number of Fisher Scoring iterations: 5
```

It is eliminated firstp.lines as its p-value of inclusion is greater than 0.05.

```
multiv1pm.2.1 <- glm(good.customer ~ market + digita_cli + fp.seasonwi
nter + log10(totalvalue),
                  data=masterCustomers.1pm.train,
                  family="binomial")

summary(multiv1pm.2.1)
```

```
##
## Call:
## glm(formula = good.customer ~ market + digita_cli + fp.seasonwinter
+
##     log10(totalvalue), family = "binomial", data = masterCustomers.
1pm.train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4980  -0.4419  -0.3488  -0.3097   4.7023
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -5.3137     0.1947 -27.296  < 2e-16 ***
## marketOEM             1.3143     0.2221   5.919 3.25e-09 ***
## marketSI              2.0632     0.3371   6.121 9.32e-10 ***
## digita_cliS           0.4285     0.1817   2.358   0.0184 *
## fp.seasonwinterTRUE   0.2470     0.1156   2.136   0.0327 *
## log10(totalvalue)     1.7626     0.1001  17.603  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3128.4  on 4401  degrees of freedom
## Residual deviance: 2633.3  on 4396  degrees of freedom
## AIC: 2645.3
##
## Number of Fisher Scoring iterations: 5
```

At this moment all the variables obtain a "p-value of inclusion" smaller than 0.05.

Now, it is checked the p-value for all the variables previously discarded in order to check if they are now significant. None of the variables is significant for a 5% level.

multiv1pm.2.1 is then the "Preliminary main effects model" for the One purchase model.

4. Interactions:

All the possible paired interactions are tested by checking its "p-value of inclusion" to the Preliminary main effects model.

It is only found significant for a 10% level one interaction: market x log10(totalvalue)

5. Final multivariate model

Interaction found significant in stage four is added to the "Preliminary main effects model"

```
multiv1pm.3 <- glm(good.customer ~ market + digita_cli + fp.seasonwint
er + log10(totalvalue) + market * log10(totalvalue),
                   data=masterCustomers.1pm.train,
                   family="binomial")

summary(multiv1pm.3)

##
## Call:
## glm(formula = good.customer ~ market + digita_cli + fp.seasonwinter
+
##     log10(totalvalue) + market * log10(totalvalue), family = "binom
ial",
##     data = masterCustomers.1pm.train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2366  -0.4327  -0.3315  -0.2934   4.9375
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -5.7068     0.2105 -27.111  < 2e-16 **
*
## marketOEM                       4.2519     0.5868   7.246 4.30e-13 **
*
## marketSI                        5.3117     0.9248   5.744 9.27e-09 **
*
## digita_cliS                     0.3877     0.1844   2.103 0.035483 *
## fp.seasonwinterTRUE             0.2415     0.1158   2.085 0.037084 *
## log10(totalvalue)               1.9788     0.1081  18.302  < 2e-16 **
*
## marketOEM:log10(totalvalue)    -1.4268     0.2683  -5.317 1.05e-07 **
*
## marketSI:log10(totalvalue)     -1.7192     0.4518  -3.805 0.000142 **
*
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3128.4  on 4401  degrees of freedom
```

```
## Residual deviance: 2601.2  on 4394  degrees of freedom
## AIC: 2617.2
##
## Number of Fisher Scoring iterations: 5
```

Both main effects and interactions are significant with a "p-value of inclusion" smaller than 0.05.

multiv1pm.3 is the "Final preliminary model".

Diagnostics

When residual diagnostics are performed for the "Preliminary final model" it is found a customer which has an unusually high dDev. It is the customer 732066.

This customer has a very low value in its first purchase but subsequent purchases led to the customer classification as a good one. In this case, a customer with very low first purchase value, and hence very low probability of being a good customer under the One purchase model, was classified as a good customer. Company reflected that the first transaction was a gift and not a purchase so that customer is removed from the dataset.

After removing the customer the model is recalculated.

```
masterCustomers.1pm.train <- filter(masterCustomers.1pm.train, custome
r_id!=732066)
masterCustomers.1pm <- filter(masterCustomers.1pm, customer_id!=732066
)

multiv1pm.diag <- glm(good.customer ~ market + digita_cli + fp.seasonw
inter  + log10(totalvalue) + market*log10(totalvalue),
                  data=masterCustomers.1pm.train, family="binomial")

summary(multiv1pm.diag)

##
## Call:
## glm(formula = good.customer ~ market + digita_cli + fp.seasonwinter
+
##     log10(totalvalue) + market * log10(totalvalue), family = "binom
ial",
##     data = masterCustomers.1pm.train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2931  -0.4299  -0.3261  -0.2884   3.1769
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -5.8221     0.2130 -27.335  < 2e-16 **
*
## marketOEM                     4.3698     0.5878   7.435 1.05e-13 **
*
## marketSI                      5.4306     0.9252   5.870 4.37e-09 **
```

```
*
## digita_cliS                    0.3886     0.1851   2.099   0.0358 *
## fp.seasonwinterTRUE            0.2323     0.1165   1.994   0.0462 *
## log10(totalvalue)              2.0407     0.1091  18.698  < 2e-16 **
*
## marketOEM:log10(totalvalue)   -1.4885     0.2688  -5.538 3.06e-08 **
*
## marketSI:log10(totalvalue)    -1.7817     0.4520  -3.942 8.08e-05 **
*
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3124.1  on 4400  degrees of freedom
## Residual deviance: 2576.5  on 4393  degrees of freedom
## AIC: 2592.5
##
## Number of Fisher Scoring iterations: 5

multiv1pm.final <- multiv1pm.diag
```

Deviance of the model has decreased 24.7 point only removing this customer of the dataset (2,691.2 – 2,576.5)

At this moment there are no extreme values and only 3.56% of the customers have a dDev greater than 4 which is a crude approximation of the upper 95% percentile of the distribution, we can say then that the model fits reasonably well.

Assessing the fit of the data

AUC and ROC curve

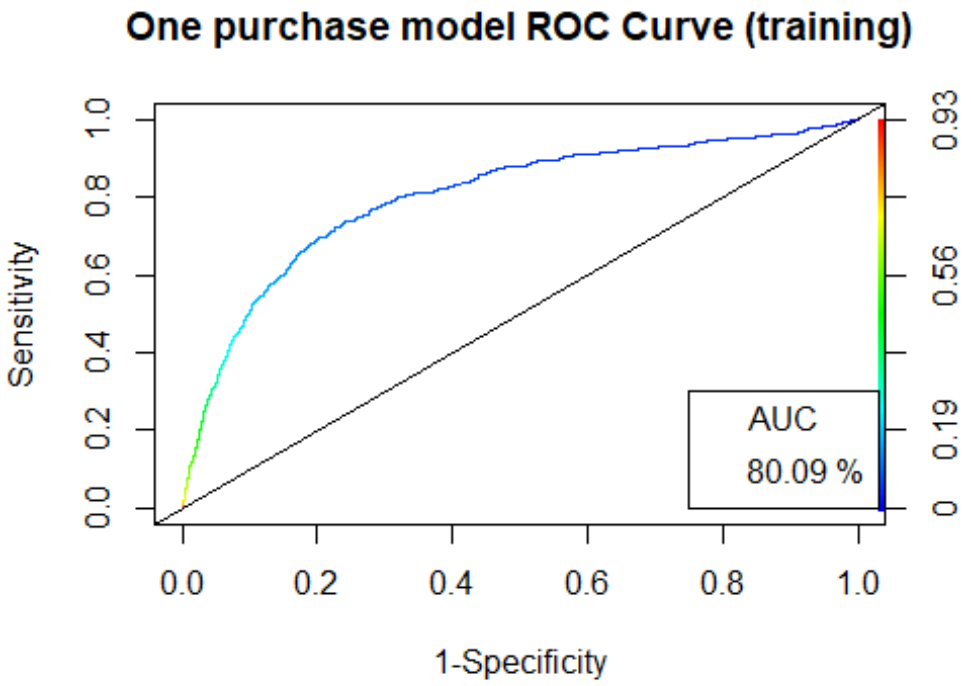
Training data

```
prediction.1pm.train <- predict(multiv1pm.final, masterCustomers.1pm.t
rain, type="response")
predvsreal.1pm.train <- prediction(prediction.1pm.train, masterCustome
rs.1pm.train$good.customer)

roc <- performance(predvsreal.1pm.train, "tpr", "fpr")
auc <- performance(predvsreal.1pm.train, "auc")
auc <- unlist(slot(auc, "y.values"))
print(c(AUC=auc*100))

##     AUC
## 80.0866

roc <- performance(predvsreal.1pm.train, "tpr", "fpr")
plot(roc, colorize=T, ylab="Sensitivity", xlab="1-Specificity", main="
One purchase model ROC Curve (training)")
abline(a=0, b=1)
legend(.75,.3,paste(round(auc*100,2),"%"), title="AUC")
```



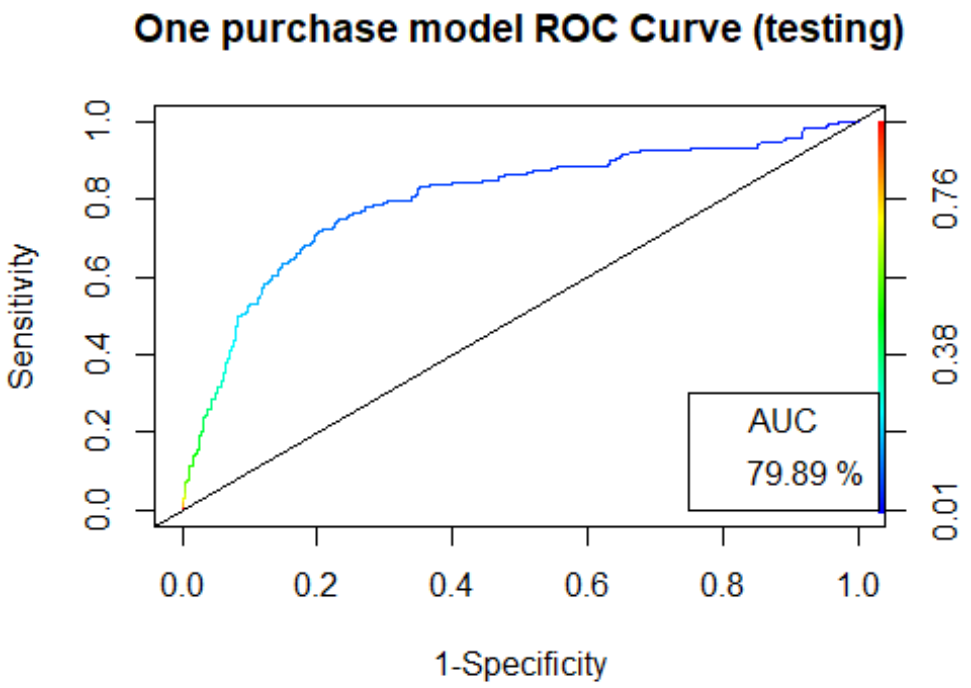
Testing data

```
prediction.1pm.test <- predict(multiv1pm.final, masterCustomers.1pm.te
st, type="response")
predvsreal.1pm.test <- prediction(prediction.1pm.test, masterCustomers
.1pm.test$good.customer)

roc <- performance(predvsreal.1pm.test, "tpr", "fpr")
auc <- performance(predvsreal.1pm.test, "auc")
auc <- unlist(slot(auc, "y.values"))
print(paste(c("AUC of the first purchase using testing dataset is", au
c*100)))

## [1] "AUC of the first purchase using testing dataset is"
## [2] "79.8874196265426"

roc <- performance(predvsreal.1pm.test, "tpr", "fpr")
plot(roc, colorize=T, ylab="Sensitivity", xlab="1-Specificity", main="
One purchase model ROC Curve (testing)")
abline(a=0, b=1)
legend(.75,.3,paste(round(auc*100,2),"%"), title="AUC")
```



We can see that AUC calculated using with the training and testing dataset are very similar. It means that the model can predict customers which has not been used in the model construction process.

Hosmer Lemeshow test

Training data

```
Hosmer.1pm <- hoslem.test(multiv1pm.final$y, multiv1pm.final$fitted.va
lues, g=10)
Hosmer.1pm.df <- cbind(Hosmer.1pm$expected, Hosmer.1pm$observed)
Hosmer.1pm.df
```

```
##                       yhat0     yhat1  y0  y1
## [1.37e-05,0.0404] 450.8941  13.10593 445  19
## (0.0404,0.0419]   402.7986  17.20140 412   8
## (0.0419,0.0457]   417.9803  19.01972 425  12
## (0.0457,0.0519]   418.4353  21.56474 429  11
## (0.0519,0.0608]   417.2088  24.79122 424  18
## (0.0608,0.0752]   408.3162  29.68383 411  27
## (0.0752,0.0999]   402.1440  37.85598 403  37
## (0.0999,0.145]    387.0638  52.93623 371  69
## (0.145,0.27]      355.5022  84.49782 331 109
## (0.27,0.928]      238.6569 201.34312 248 192
```

```
print(paste("Hosmer p-value for the One purchase's model using trainin
g dataset is is ", Hosmer.1pm$p.value))
```

```
## [1] "Hosmer p-value for the One purchase's model using training dat
aset is is  5.2205395631888e-05"
```

Testing data

```
Hosmer.1pm <- hoslem.test(masterCustomers.1pm.test$good.customer, pred
iction.1pm.test, g=10)
Hosmer.1pm.df <- cbind(Hosmer.1pm$expected, Hosmer.1pm$observed)
Hosmer.1pm.df

##                       yhat0     yhat1  y0 y1
## [0.00568,0.0404] 143.84624  4.153758 141  7
## (0.0404,0.0421]  142.87327  6.126733 145  4
## (0.0421,0.0468]  137.65870  6.341305 142  2
## (0.0468,0.0525]  138.70999  7.290013 140  6
## (0.0525,0.0608]  138.70050  8.299500 141  6
## (0.0608,0.075]   137.15391  9.846085 141  6
## (0.075,0.101]    133.38930 12.610698 136 10
## (0.101,0.152]    128.73432 18.265680 125 22
## (0.152,0.273]    117.87874 29.121264 105 42
## (0.273,0.947]     82.86207 64.137928  85 62

print(paste("Hosmer p-value for the ONe purchase's model is ", Hosmer.
1pm$p.value))

## [1] "Hosmer p-value for the ONe purchase's model is 0.0290478541315
741"
```

Final coefficients calculation

After testing the model we calculate final coefficietns running the final model with the complete masterCustomers.1pm dataset:

```
masterCustomers.1pm <- masterCustomers.1pm %>%
  mutate(fp.seasonwinter=ifelse(fp.season=="winter",TRUE, FALSE))

multiv1pm.final.coefficients <- glm(good.customer ~ market + digita_cl
i + fp.seasonwinter  + log10(totalvalue) + market*log10(totalvalue),
                    data=masterCustomers.1pm, family="binomial")

summary(multiv1pm.final.coefficients)

##
## Call:
## glm(formula = good.customer ~ market + digita_cli + fp.seasonwinter
+
##     log10(totalvalue) + market * log10(totalvalue), family = "binom
ial",
##     data = masterCustomers.1pm)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3518  -0.4285  -0.3237  -0.2838   3.2013
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -5.93922    0.18818 -31.561  < 2e-16 **
```

```
*
## marketOEM                      4.81320    0.48551   9.914  < 2e-16 **
*
## marketSI                       5.38454    0.87825   6.131 8.73e-10 **
*
## digita_cliS                    0.33111    0.16467   2.011   0.0444 *
## fp.seasonwinterTRUE            0.25369    0.10046   2.525   0.0116 *
## log10(totalvalue)              2.10482    0.09638  21.839  < 2e-16 **
*
## marketOEM:log10(totalvalue) -1.69909    0.22157  -7.669 1.74e-14 **
*
## marketSI:log10(totalvalue)  -1.81356    0.43538  -4.165 3.11e-05 **
*
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4164.3  on 5868  degrees of freedom
## Residual deviance: 3424.5  on 5861  degrees of freedom
## AIC: 3440.5
##
## Number of Fisher Scoring iterations: 5
```