# Task 2 submission: Implementation

This is your submission document for [DataCapX Submission 3: Implementation](https://courses.edx.org/courses/course-v1:AdelaideX+DataCapX+3T2018/courseware/a7d3a888e3aa4b2689c5421bc4550619/c8d2611b443049c392c364d5bbbf492b/1?activate_block_id=block-v1%3AAdelaideX%2BDataCapX%2B3T2018%2Btype%40vertical%2Bblock%4022509f00a2814b41882462af3f6ddd3d).   
Save this document on your local machine and include all of your work within the relevant part of the assignment. Once you’ve completed every part of Task 2: Implementation, upload this document via the [Your Response area](https://courses.edx.org/courses/course-v1:AdelaideX+DataCapX+3T2018/courseware/a7d3a888e3aa4b2689c5421bc4550619/c8d2611b443049c392c364d5bbbf492b/1?activate_block_id=block-v1%3AAdelaideX%2BDataCapX%2B3T2018%2Btype%40vertical%2Bblock%4022509f00a2814b41882462af3f6ddd3d).

**Note: Don’t forget to save your scripts related to this task and submit them when you submit this document.**

# Checklist

* Have you answered every question?
* Have you shown all of your working, including evidence of your code?
* Have you included all R output as script (.r) files?
* Have you clearly stated conclusions where required?
* Have you saved your code in a script?

# Task 2 submission: Implementation

Prepare the dataset and build a classifier

1. Read in the reddit dataset correctly and provide evidence of your code.   
   [3 points]

|  |
| --- |
| Read Dataset   * Read in the reddit dataset and select the brand\_safe, num\_crossposts and the six attributes - over\_18, is\_self, is\_reddit\_media\_domain, is\_video, stickied and spoiler - to form a data frame.   Factor Analysis - Summarise the data frame:   * What are the types of those factors? * If the factor contains discrete variables, what are the possible values they can take? * If the factor contains continuous variables, what are the min, max, mean values of those factors? |

# read the source csv file

reddit <- read.csv("RS\_2017-09\_filtered70.csv")

reddit <- as\_tibble(reddit)

# select the subset of columns of interest

reddit.selection <- select(reddit, brand\_safe, num\_crossposts, over\_18, is\_self, is\_reddit\_media\_domain, is\_video, stickied, spoiler)

# summarize the data

reddit.selection %>%

  mutate\_if(is.character, as.factor) %>%

  summary

Table . Selected data summary.

|  |
| --- |
| brand\_safe num\_crossposts over\_18 is\_self is\_reddit\_media\_domain is\_video stickied  False:65141 Min. :0.0000000 False:127226 False:78627 False:122255 False:139379 False:139724  True :74682 1st Qu.:0.0000000 True : 12597 True :61196 True : 17568 True : 444 True : 99  Median :0.0000000  Mean :0.0007509  3rd Qu.:0.0000000  Max. :3.0000000  spoiler  False:138983  True : 840 |

**Observations**

* Of the 8 selected variables from the reddit dataset all of them are True/False values except for the num\_crosspost variable.
* The num\_crosspost variable has the following characteristics:
  + Min: 0
  + Max: 3
  + Mean: 0.0007509
* The following variable are almost uniform with less than 0.7 % in one of the two levels. This could be an indication that these variables might have only a small contribution towards the model predictive power:
  + is\_video
  + stickied
  + spoiler

Since all the columns except num\_crosspost are Boolean values convert them to the logical data type.

reddit.selection <- reddit.selection %>%

  mutate\_if(is.character, as.logical)

Include the name/s of any associated code (.r) script file/s related to Step 1 that are included in your .zip file:

* task02.R

1. Build a classifier based on the six factors and provide the classifier information. Provide the correct analysis about the factors.   
   [3 points]

|  |
| --- |
| * Build a classifier based on the six attributes to predict the brand\_safe factor. * Analyse the impact of the six input factors - Are those factors significant, and what is the impact of their value on the prediction? |

# split the data in training and testing data sets

reddit.selection <- reddit.selection %>%

  na.omit()

n <- dim(reddit.selection)[1]

ind <- sample(c(TRUE, FALSE), n, replace=TRUE, prob=c(0.7, 0.3))

reddit.train <- reddit.selection[ind, ]

reddit.test <- reddit.selection[!ind, ]

# select the factors to include in the model

factors <- c("over\_18", "is\_self", "is\_reddit\_media\_domain", "is\_video", "stickied", "spoiler")

formula <- as.formula(paste("brand\_safe ~ ", paste(factors, collapse="+")))

# build the base model

reddit.glm\_base <- glm(formula, family=binomial(), reddit.train)

summary(reddit.glm\_base)

Table . Model summary.

|  |
| --- |
| Call:  glm(formula = formula, family = binomial(), data = reddit.train)  Deviance Residuals:  Min 1Q Median 3Q Max  -2.6461 -1.0427 0.8387 0.8387 2.5544  Coefficients:  Estimate Std. Error z value Pr(>|z|)  (Intercept) -0.32554 0.01037 -31.39 < 2e-16 \*\*\*  over\_18TRUE -2.89781 0.04255 -68.11 < 2e-16 \*\*\*  is\_selfTRUE 1.18948 0.01496 79.51 < 2e-16 \*\*\*  is\_reddit\_media\_domainTRUE 1.01094 0.02232 45.29 < 2e-16 \*\*\*  is\_videoTRUE 2.78497 0.31364 8.88 < 2e-16 \*\*\*  stickiedTRUE -2.89678 0.40347 -7.18 6.99e-13 \*\*\*  spoilerTRUE 1.62038 0.12816 12.64 < 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: 135370 on 97988 degrees of freedom  Residual deviance: 118214 on 97982 degrees of freedom  AIC: 118228  Number of Fisher Scoring iterations: 5 |

**Analyse the impact of the six input factors**

broom::tidy(reddit.glm\_base) %>% filter(term != "(Intercept)") %>% arrange(p.value)

Table . Coefficient significance (ordered by p.value).

|  |
| --- |
| # A tibble: 6 x 5  term estimate std.error statistic p.value  <chr> <dbl> <dbl> <dbl> <dbl>  1 over\_18TRUE -2.90 0.0425 -68.1 0  2 is\_selfTRUE 1.19 0.0150 79.5 0  3 is\_reddit\_media\_domainTRUE 1.01 0.0223 45.3 0  4 spoilerTRUE 1.62 0.128 12.6 1.22e-36  5 is\_videoTRUE 2.78 0.314 8.88 6.71e-19  6 stickiedTRUE -2.90 0.403 -7.18 6.99e-13 |

* All six factors used to build the model are significant.
* The three variables over\_18, is\_self, and is\_reddit\_media\_domain is the most significant in the model.
* The stickied variable appears to be have the least impact on the model and is a possible candidate for removal.
* We should test what the impact will be when removing the least significant factors – the predictive power might be similar but the result will be a simpler model.

Include the name/s of any associated code (.r) script file/s related to Step 2 that are included in your .zip file:

1. Build a classifier based on the six factors and the additional factor. Provide the learned classifier information. Provide the correct comparison analysis between the built model and the model built in the previous step.   
   [2 points]

|  |
| --- |
| * Add the num\_crossposts factor to the six-factor model you built. * Check if this will result in a better classifier, giving justifications for your answer. |

factors <- c("num\_crossposts", "over\_18", "is\_self", "is\_reddit\_media\_domain", "is\_video", "stickied", "spoiler")

formula <- as.formula(paste("brand\_safe ~ ", paste(factors, collapse="+")))

reddit.glm\_crosspost <- glm(formula, family=binomial(), reddit.train)

summary(reddit.glm\_crosspost)

Table . Seven factor model summary.

|  |
| --- |
| Call:  glm(formula = formula, family = binomial(), data = reddit.train)  Deviance Residuals:  Min 1Q Median 3Q Max  -2.6461 -1.0423 0.8388 0.8388 2.5574  Coefficients:  Estimate Std. Error z value Pr(>|z|)  (Intercept) -0.32653 0.01037 -31.474 < 2e-16 \*\*\*  num\_crossposts 0.90095 0.25994 3.466 0.000528 \*\*\*  over\_18TRUE -2.89754 0.04255 -68.098 < 2e-16 \*\*\*  is\_selfTRUE 1.19028 0.01496 79.551 < 2e-16 \*\*\*  is\_reddit\_media\_domainTRUE 1.01058 0.02232 45.271 < 2e-16 \*\*\*  is\_videoTRUE 2.78628 0.31364 8.884 < 2e-16 \*\*\*  stickiedTRUE -2.90476 0.40331 -7.202 5.92e-13 \*\*\*  spoilerTRUE 1.62100 0.12817 12.647 < 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: 135370 on 97988 degrees of freedom  Residual deviance: 118200 on 97981 degrees of freedom  AIC: 118216  Number of Fisher Scoring iterations: 5 |

broom::tidy(reddit.glm\_crosspost) %>% filter(term != "(Intercept)") %>% arrange(p.value)

Table . Seven factor model: Coefficient significance (ordered by p.value).

|  |
| --- |
| # A tibble: 7 x 5  term estimate std.error statistic p.value  <chr> <dbl> <dbl> <dbl> <dbl>  1 over\_18TRUE -2.90 0.0425 -68.1 0  2 is\_selfTRUE 1.19 0.0150 79.6 0  3 is\_reddit\_media\_domainTRUE 1.01 0.0223 45.3 0  4 spoilerTRUE 1.62 0.128 12.6 1.16e-36  5 is\_videoTRUE 2.79 0.314 8.88 6.46e-19  6 stickiedTRUE -2.90 0.403 -7.20 5.92e-13  7 num\_crossposts 0.901 0.260 3.47 5.28e- 4 |

anova(reddit.glm\_base, reddit.glm\_crosspost, test="Chisq")

Table . Model Comparison.

|  |
| --- |
| Analysis of Deviance Table  Model 1: brand\_safe ~ over\_18 + is\_self + is\_reddit\_media\_domain + is\_video +  stickied + spoiler  Model 2: brand\_safe ~ num\_crossposts + over\_18 + is\_self + is\_reddit\_media\_domain +  is\_video + stickied + spoiler  Resid. Df Resid. Dev Df Deviance Pr(>Chi)  1 97982 118214  2 97981 118200 1 13.864 0.0001966 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1 |

**Observations**

* While the num\_crossposts factor is significant in the model, it is also the least significant and might not have a major impact on model performance.
* There is a reduction in deviance (overall quality of your model fit) for the second model which is statistically significant, however, the change is not very large so we do not expect the addition of this factor to greatly improve the prediction skill of the model.
* The AIC has decreased from 118228 to 118216 which is again a small change that indicates that the addition of the factor might not have a large enough impact to be included in the final model.
* Based on the above, addition of the factor will most likely result in a model that fits the data slightly better with slightly better predictions, but if we want to reduce the number of factors required to perform predictions this factor is a good candidate to be excluded.

Include the name/s of any associated code (.r) script file/s related to Step 3 that are included in your .zip file:

* task02.R

**Model/feature selection**

1. Perform model selection and produce the correct selection results. Provide your result and associated code.   
   [4 points]

|  |
| --- |
| * In this part, you will perform the model selection to reduce the six factors (over\_18, is\_self, is\_reddit\_media\_domain, is\_video, stickied, and spoiler) to three factors. You can use either “AIC” or “BIC” as the selection criterion. * Choose one model selection method to reduce the predictive factors from six to three. You will need to show:   + Your three chosen factors.   + The code used to calculate this result. |

The first method will be to simply select the three factors that are the most significant in the model.

broom::tidy(reddit.glm\_base) %>%

  select(term, p.value) %>%

  filter(term != "(Intercept)") %>%

  arrange(p.value) %>%

  head(3)

Table . The three most significant coefficients.

|  |
| --- |
| # A tibble: 3 x 2  term p.value  <chr> <dbl>  1 over\_18TRUE 0  2 is\_selfTRUE 0  3 is\_reddit\_media\_domainTRUE 0 |

factors <- c("over\_18", "is\_self", "is\_reddit\_media\_domain")

formula <- as.formula(paste("brand\_safe ~ ", paste(factors, collapse="+")))

reddit.glm\_sigcoef <- glm(formula, family=binomial(), reddit.train)

summary(reddit.glm\_sigcoef)

Table . Model summary using over\_18, is\_self, and is\_reddit\_media\_domain as factors.

|  |
| --- |
| Call:  glm(formula = formula, family = binomial(), data = reddit.train)  Deviance Residuals:  Min 1Q Median 3Q Max  -1.5620 -1.0445 0.8365 0.8365 2.5497  Coefficients:  Estimate Std. Error z value Pr(>|z|)  (Intercept) -0.32088 0.01035 -30.99 <2e-16 \*\*\*  over\_18TRUE -2.89012 0.04240 -68.16 <2e-16 \*\*\*  is\_selfTRUE 1.19101 0.01493 79.75 <2e-16 \*\*\*  is\_reddit\_media\_domainTRUE 1.04802 0.02217 47.28 <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: 135370 on 97988 degrees of freedom  Residual deviance: 118710 on 97985 degrees of freedom  AIC: 118718  Number of Fisher Scoring iterations: 5 |

Include the name/s of any associated code (.r) script file/s related to Step 1 that are included in your .zip file:

* task02.R

1. Perform model selection using another method and produce the correct selection results. Provide your result and associated code  
   [4 points]

|  |
| --- |
| Choose another model selection method to repeat the process in the previous step. Check if both methods produce the same results. Provide the selection result and code for this step. |

Stepwise Regression with forward selection will be used to determine the top 3 factors. The first three factor that are selected by this method will be chosen for the final model.

glm.null <- glm(brand\_safe ~ 1, data=reddit.train)

step(glm.null, scope=list(upper=glm(brand\_safe ~ ., data=reddit.train)), direction="forward")

Table . Stepwise regression output.

|  |
| --- |
| Start: AIC=141770.8  brand\_safe ~ 1  Df Deviance AIC  + over\_18 1 22305 133059  + is\_self 1 23116 136558  + is\_reddit\_media\_domain 1 24233 141181  + is\_video 1 24320 141533  + spoiler 1 24324 141548  + stickied 1 24364 141710  + num\_crossposts 1 24377 141763  <none> 24380 141771  Step: AIC=133059.4  brand\_safe ~ over\_18  Df Deviance AIC  + is\_self 1 21222 128185  + is\_reddit\_media\_domain 1 22192 132562  + spoiler 1 22252 132827  + is\_video 1 22256 132845  + stickied 1 22290 132994  + num\_crossposts 1 22304 133054  <none> 22305 133059  Step: AIC=128185  brand\_safe ~ over\_18 + is\_self  Df Deviance AIC  + is\_reddit\_media\_domain 1 20689 125692  + is\_video 1 21146 127835  + spoiler 1 21183 128007  + stickied 1 21203 128097  + num\_crossposts 1 21219 128172  <none> 21222 128185  Step: AIC=125692.1  brand\_safe ~ over\_18 + is\_self + is\_reddit\_media\_domain  Df Deviance AIC  + spoiler 1 20651 125513  + is\_video 1 20658 125550  + stickied 1 20670 125605  + num\_crossposts 1 20686 125681  <none> 20689 125692  Step: AIC=125512.9  brand\_safe ~ over\_18 + is\_self + is\_reddit\_media\_domain + spoiler  Df Deviance AIC  + is\_video 1 20620 125370  + stickied 1 20631 125424  + num\_crossposts 1 20648 125502  <none> 20651 125513  Step: AIC=125369.9  brand\_safe ~ over\_18 + is\_self + is\_reddit\_media\_domain + spoiler +  is\_video  Df Deviance AIC  + stickied 1 20601 125281  + num\_crossposts 1 20617 125359  <none> 20620 125370  Step: AIC=125280.6  brand\_safe ~ over\_18 + is\_self + is\_reddit\_media\_domain + spoiler +  is\_video + stickied  Df Deviance AIC  + num\_crossposts 1 20598 125269  <none> 20601 125281  Step: AIC=125268.7  brand\_safe ~ over\_18 + is\_self + is\_reddit\_media\_domain + spoiler +  is\_video + stickied + num\_crossposts  Call: glm(formula = brand\_safe ~ over\_18 + is\_self + is\_reddit\_media\_domain +  spoiler + is\_video + stickied + num\_crossposts, data = reddit.train)  Coefficients:  (Intercept) over\_18TRUE is\_selfTRUE is\_reddit\_media\_domainTRUE  0.4300 -0.4716 0.2649 0.2282  spoilerTRUE is\_videoTRUE stickiedTRUE num\_crossposts  0.2559 0.3132 -0.4955 0.1824  Degrees of Freedom: 97988 Total (i.e. Null); 97981 Residual  Null Deviance: 24380  Residual Deviance: 20600 AIC: 125300 |

Observations

* The very last factor that is selected is num\_crossposts *(included in this procedure for completeness and because it is interesting to confirm earlier assertions),* which was also identified in previous sections as the factor that will most likely have the least impact on the model.
* If we consider all 7 (or if we omit num\_crossposts - 6) the result will be that all the factors we use for testing will be included in the final model. This is expected as we saw in earlier sections that these factors are all significant.
* Using stepwise regression the same three factors were selected fist as when we simply chose factors based on coefficient significance.
* Final model with 3 factors: brand\_safe ~ over\_18 + is\_self + is\_reddit\_media\_domain

Include the name/s of any associated code (.r) script file/s related to Step 2 that are included in your .zip file:

* task02.R

Verifying your chosen method

1. Provide an appropriate definition of the criterion. Provide code to test the performance of the learned classifier. Compare it with the baseline method. Provide evidence of your code.  
   [5 points]

|  |
| --- |
| In this part, you will test the performance of a model in predicting new data. You will test the performance of both a classifier model you built earlier and a feature selection process.   * Define criteria that can be used to evaluate performance of a classifier model in predicting new data. Note that the criteria need to produce a single scalar indicating the performance of the classifier model. (Note: Based on the feedback you will receive from a University of Adelaide staff member, this may be different to the design phase. * Provide the code you use to calculate the predictive performance of the classifier * Compare your classifier with a naive baseline which predicts brand\_safe as always True or False on the test data |

The Area under the ROC Curve will be used as the evaluation metric to evaluate the classification performance of a model. The following definition is taken from <https://datacadamia.com/data_mining/auc>:

*“The Area under the curve (AUC) is a performance metrics for a binary classifiers. By comparing the ROC curves with the area under the curve, or AUC, it captures the extent to which the curve is up in the Northwest corner. An higher AUC is good.*

*A score of 0.5 is no better than random guessing. 0.9 would be a very good model but a score of 0.9999 would be too good to be true and will indicate overfitting.”*

The AUC for both the training- and testing dataset will be calculated to allow for comparison and the detection of overfitting.

test\_model <- function(factors, train, test) {

  # build the model

  formula <- as.formula(paste("brand\_safe ~ ", paste(factors, collapse="+")))

  glm\_model <- glm(formula, family=binomial(), train)

  # perform the predictions

  train$prediction <- predict(glm\_model, type="response")

  test$prediction <- predict(glm\_model, type="response", newdata=test)

  # return the formula and AUC value

  return (c(

    paste(factors, collapse="+"),

    round(auc(train$brand\_safe, train$prediction), digits=5),

    round(auc(test$brand\_safe, test$prediction), digits=5),

    round(glm\_model$aic, digits=5)))

}

# compare different models against the baseline

auc\_results <- list(

  test\_model(c("1"), reddit.train, reddit.test),

  test\_model(c("over\_18", "is\_self", "is\_reddit\_media\_domain"), reddit.train, reddit.test),

  test\_model(c("over\_18", "is\_self", "is\_reddit\_media\_domain", "is\_video", "stickied", "spoiler"), reddit.train, reddit.test),

  test\_model(c("num\_crossposts", "over\_18", "is\_self", "is\_reddit\_media\_domain", "is\_video", "stickied", "spoiler"), reddit.train, reddit.test))

auc\_results <- data.frame(matrix(unlist(auc\_results), nrow=length(auc\_results), byrow=TRUE))

colnames(auc\_results) <- c("factors", "train.AUC", "test.AUC", "AIC")

auc\_results

Table . Model comparison.

|  |
| --- |
| factors train.AUC test.AUC AIC  1 1 0.5 0.5 135372.02378  2 over\_18+is\_self+is\_reddit\_media\_domain 0.69734 0.69726 118717.82914  3 over\_18+is\_self+is\_reddit\_media\_domain+is\_video+stickied+spoiler 0.70106 0.70072 118228.04964  4 num\_crossposts+over\_18+is\_self+is\_reddit\_media\_domain+is\_video+stickied+spoiler 0.7012 0.70083 118216.18602 |

Observations

* The baseline model with no skill has an AUC of 0.5 as expected.
* The AUC for predictions on both the training and testing data are very close to each other, an indication that the model has sufficiently generalized and does not have an overfitting problem.
* The model using num\_crossposts as an additional factor displays a very small improvement over the model using six factors which agrees with an earlier assertion that it could be omitted.
* The model using only 3 factors produced results that are very-very close to models using more factors and should thus be selected because it reduces model complexity.
* The model using only three factors perform significantly better than the baseline model with no skills and at an AUC of 0.7 could be sufficient depending on the intended application and accuracy required.

Include the name/s of any associated code (.r) script file/s related to Step 1 that are included in your .zip file:

* task02.R

1. Perform feature/model selection with the criterion developed in Step 1 of *Verifying your chosen method*. Conclude whether the selection results are the same as those in *Model/feature selection*. Provide your selection results and associated code. Provide your selection results and associated code.   
   [5 points]

|  |
| --- |
| * Now apply the criterion developed above to guide the feature selection procedure. Verify if the same feature selection results can be obtained if the selection criterion changes. * Note: You can choose any feature selection method developed in Part 2. |

The idea employed in this section is to apply a manual method similar to stepwise regression with forward selection by adding a variable at each step that increases the model predictive power the most. Instead of looking at AIC as before, the AUC value will be considered derived from the predicted values.

# step 1 - select the first variable with the most impact

auc\_results <- list(

  test\_model(c("num\_crossposts"), reddit.train, reddit.test),

  test\_model(c("over\_18"), reddit.train, reddit.test),

  test\_model(c("is\_self"), reddit.train, reddit.test),

  test\_model(c("is\_reddit\_media\_domain"), reddit.train, reddit.test),

  test\_model(c("is\_video"), reddit.train, reddit.test),

  test\_model(c("stickied"), reddit.train, reddit.test),

  test\_model(c("spoiler"), reddit.train, reddit.test))

auc\_results <- data.frame(matrix(unlist(auc\_results), nrow=length(auc\_results), byrow=TRUE))

colnames(auc\_results) <- c("factors", "train.AUC", "test.AUC", "AIC")

auc\_results %>%

  arrange(desc(train.AUC))

Table . Step 1, is\_self selected as the first factor to include.

|  |
| --- |
| factors train.AUC test.AUC AIC  1 is\_self 0.61323 0.61216 130232.60093  2 over\_18 0.58362 0.5833 125864.55206  3 is\_reddit\_media\_domain 0.52576 0.52653 134775.75846  4 spoiler 0.50373 0.50406 135123.7485  5 is\_video 0.50283 0.50259 135065.24718  6 stickied 0.50072 0.50046 135302.33557  7 num\_crossposts 0.50024 0.5003 135364.06246 |

# step 2 - is\_self as first variable

auc\_results <- list(

  test\_model(c("is\_self", "num\_crossposts"), reddit.train, reddit.test),

  test\_model(c("is\_self", "over\_18"), reddit.train, reddit.test),

  test\_model(c("is\_self", "is\_reddit\_media\_domain"), reddit.train, reddit.test),

  test\_model(c("is\_self", "is\_video"), reddit.train, reddit.test),

  test\_model(c("is\_self", "stickied"), reddit.train, reddit.test),

  test\_model(c("is\_self", "spoiler"), reddit.train, reddit.test))

auc\_results <- data.frame(matrix(unlist(auc\_results), nrow=length(auc\_results), byrow=TRUE))

colnames(auc\_results) <- c("factors", "train.AUC", "test.AUC", "AIC")

auc\_results %>%

  arrange(desc(train.AUC))

Table .Step 2, over\_18 selected as the second factor to include.

|  |
| --- |
| factors train.AUC test.AUC AIC  1 is\_self+over\_18 0.67434 0.67378 121069.54053  2 is\_self+is\_reddit\_media\_domain 0.64191 0.64125 127601.95836  3 is\_self+is\_video 0.61693 0.61555 129800.35579  4 is\_self+spoiler 0.61511 0.61444 130040.33536  5 is\_self+stickied 0.61405 0.61278 130141.26016  6 is\_self+num\_crossposts 0.61346 0.6125 130216.19881 |

# step 3 - over\_18 as second variable

auc\_results <- list(

  test\_model(c("is\_self", "over\_18", "num\_crossposts"), reddit.train, reddit.test),

  test\_model(c("is\_self", "over\_18", "is\_reddit\_media\_domain"), reddit.train, reddit.test),

  test\_model(c("is\_self", "over\_18", "is\_video"), reddit.train, reddit.test),

  test\_model(c("is\_self", "over\_18", "stickied"), reddit.train, reddit.test),

  test\_model(c("is\_self", "over\_18", "spoiler"), reddit.train, reddit.test))

auc\_results <- data.frame(matrix(unlist(auc\_results), nrow=length(auc\_results), byrow=TRUE))

colnames(auc\_results) <- c("factors", "train.AUC", "test.AUC", "AIC")

auc\_results %>%

  arrange(desc(train.AUC))

Table . Step 3, is\_reddit\_media\_domain selected as the second factor to include

|  |
| --- |
| factors train.AUC test.AUC AIC  1 is\_self+over\_18+is\_reddit\_media\_domain 0.69734 0.69726 118717.82914  2 is\_self+over\_18+is\_video 0.67736 0.67654 120681.98129  3 is\_self+over\_18+spoiler 0.67584 0.67537 120859.19539  4 is\_self+over\_18+stickied 0.67506 0.67437 120976.39652  5 is\_self+over\_18+num\_crossposts 0.67451 0.67394 121056.93794 |

* The three factors selected is once again the same factors identified using all previous methods.
* It is interesting to note that the order in which factors were selected are different than when AIC was used as an evaluation metric.

Include the name/s of any associated code (.r) script file/s related to Step 2 that are included in your .zip file:

* task02.R

Total points possible for Task 2: Implementation 26