title

THE REGRESSORS

January 14th, 2021

Importing libraries:

```
library(dplyr)
library(ggplot2)
library(stringr)
library(gridExtra)
library(outliers)
library(PerformanceAnalytics)
library(foreach)
library(MASS)
library(e1071)
library(VGAM)
library(caret)
library(klaR)
library(arm)
library(caTools)
library(stepPlr)
library(LiblineaR)
library(caret)
library(Epi)
library(ROSE)
library(ResourceSelection)
```

Importing and manipulating the data:

```
credit <- read.csv('./data/credit.csv')
names(credit) <- tolower(names(credit))
```

Basic variable selection:

```
vars <- c("obs.","chk_acct","duration","history",
          "new_car","used_car","furniture","radio.tv",
          "education","retraining","amount","sav_acct",
          "employment","install_rate","male_div","male_single",
          "male_mar_or_wid","co.applicant","guarantor","present_resident",
          "real_estate","prop_unkn_none","age","other_install",
          "rent","own_res","num_credits","job",
          "num_dependents","telephone","foreign","response")

vars_to_remove <- c("own_res", "obs.", "real_estate")

credit <- credit %>% dplyr::select(setdiff(vars,vars_to_remove))
credit$response <- as.factor(credit$response)
names(credit)
```

```
#>  [1] "chk_acct"          "duration"          "history"           "new_car"
#>  [5] "used_car"          "furniture"         "radio.tv"          "education"
#>  [9] "retraining"        "amount"            "sav_acct"          "employment"
#> [13] "install_rate"      "male_div"          "male_single"       "male_mar_or_wid"
#> [17] "co.applicant"      "guarantor"         "present_resident" "prop_unkn_none"
#> [21] "age"               "other_install"     "rent"              "num_credits"
#> [25] "job"               "num_dependents"    "telephone"         "foreign"
#> [29] "response"
```

TTS:

```
set.seed(12)
spl = createDataPartition(credit$response, p = 0.8, list = FALSE)
Train = credit[spl,]
Test = credit[-spl,]
Train$response <- as.factor(Train$response)
Test$response <- as.factor(Test$response)
```

We define a function to obtain the formula of all models with and without a chosen amount of interactions (2-way, 3-way, etc):

```
model_formula <- function(data, combs, target, with_int=TRUE, all=FALSE) {
    formulas <- c()
    cols <- names(data)[2:(length(names(data))-1)]
    combinations <- combinat::combn(cols, combs)
    for (i in 1:length(combinations[1,])) {
        if (with_int == TRUE) {
            if (all == TRUE) {
                form_pst <- paste(combinations[,i], collapse="*")
                form <- stringr::str_interp("${target}~${form_pst}")
                formulas <- c(formulas, form)
            } else {
                form_pst <- paste(combinations[,i], collapse="+")
                form <- stringr::str_interp("${target}~(${form_pst})^${all}")
                formulas <- c(formulas, form)
            }
        } else {
            form_pst <- paste(combinations[,i], collapse="+")
            form <- stringr::str_interp("${target}~${form_pst}")
            formulas <- c(formulas, form)
        }
    }
    return(formulas)
}
```

Modelling function:

```
modelling <- function(data, formulas) {
    models <- list()
    for (i in 1:length(formulas)) {
        models[[i]] <- glm(formula=formulas[i], family=binomial, data=data)
    }
    return(models)
}
```

LRT for models with and without interactions:

```
test <- function(formulas_with, formulas_without, models_with_int, models_without_int) {
    p_vals <- c()
    for (i in 1:length(formulas_with)) {
        p_vals <- c(p_vals, anova(models_with_int[[i]], models_without_int[[i]], test="Chisq")$"Pr(>Chi
    }
    return(data.frame(formulas_with=formulas_with, formulas_without=formulas_without, pvals=p_vals))
}
```

Scoring function:

```
scoring <- function(data, testing, models, formulas) {
    accuracy <- c()
    roc_cutoff <- c()
    roc_auc <- c()
    roc_sensitivity <- c()
    roc_specificity <- c()
    # hoslem <- c()
    for (i in 1:length(models)) {
        # ROC curve
        roc1 <- Epi::ROC(form=formula(models[[i]]), data=data, plot="ROC", lw=3, cex=1.5)
        cutoff <- which.max(rowSums(roc1$res[, c("sens", "spec")]))

        # ROC params
        roc_cutoff <- c(roc_cutoff, roc1$res$lr.eta[cutoff])
        roc_auc <- c(roc_auc, roc1$AUC)
        roc_sensitivity <- c(roc_sensitivity, roc1$res$sens[cutoff])
        roc_specificity <- c(roc_specificity, roc1$res$spec[cutoff])

        # prediction using BEST cutoff
        prediction <- predict(models[[i]], newdata=testing, type="response")
        prediction <- ifelse(prediction > roc1$res$lr.eta[cutoff], 1, 0)
        pred <- as.factor(prediction)

        # target score
        real_vals <- as.factor(testing$response)

        # hosmer lemeshow goodness of fit test
        # hltest <- hoslem.test(real_vals, prediction)$p.value
        # hoslem <- c(hoslem, hltest)

        # confusion matrix score
        accuracy <- c(accuracy,confusionMatrix(pred, real_vals)$overall[1])
    }
    return(data.frame(formula=formulas,
                      accuracy=accuracy,
                      cutoff=roc_cutoff,
                      roc_auc=roc_auc,
                      sensitivity=roc_sensitivity,
                      specificity=roc_specificity))
}
```

Testing 2-variable models with and without interactions

We create models with all the combinations of 2 variables and then we perform LRT for models with and without interactions. Then we select models with an LRT p-value under 0.01, in order to keep the most important interactions.

```
formulas_with <- model_formula(credit, 2, "response", with_int=TRUE, all=2)
formulas_without <- model_formula(credit, 2, "response", with_int=FALSE)
models_with <- modelling(Train, formulas_with)
models_without <- modelling(Train, formulas_without)
```

We run the tests:

```
two_var_combs <- test(formulas_with, formulas_without, models_with, models_without)
```

We remove NAs, given that these interactions' product is 0 for all values, therefore, the LRT returns a p-value of 1 (meaning there's no difference between the models).

```
two_var_combs <- na.omit(two_var_combs[order(-two_var_combs$pvals),])
two_var_combs <- two_var_combs[two_var_combs$pvals < 0.01,]
```

We present the table showing the model formulas and the p-values:

```
knitr::kable(
    two_var_combs,
    booktabs=TRUE,
    longtable=TRUE,
    caption="best models"
)
```

Table 1: best models

	formulas_with	formulas_without	pvals
245	response~(install_rate+telephone)^2	response~install_rate+telephone	0.0092891
80	response~(used_car+amount)^2	response~used_car+amount	0.0063424
223	response~(employment+prop_unkn_none)^2	response~employment+prop_unkn_none	0.0039463
125	response~(radio.tv+employment)^2	response~radio.tv+employment	0.0034811
45	response~(history+other_install)^2	response~history+other_install	0.0032266
273	response~(male_single+foreign)^2	response~male_single+foreign	0.0025497
348	response~(job+foreign)^2	response~job+foreign	0.0008878
9	response~(duration+sav_acct)^2	response~duration+sav_acct	0.0008110
173	response~(retraining+age)^2	response~retraining+age	0.0004303

```
forms <- model_formula(Train, 3, "response", with_int=FALSE, all=3)
models <- modelling(Train, forms)
formulas_with <- model_formula(Train, 4, "response", with_int=TRUE, all=4)
formulas_without <- model_formula(Train, 4, "response", with_int=FALSE)
models_with <- modelling(Train, formulas_with)
models_without <- modelling(Train, formulas_without)
```

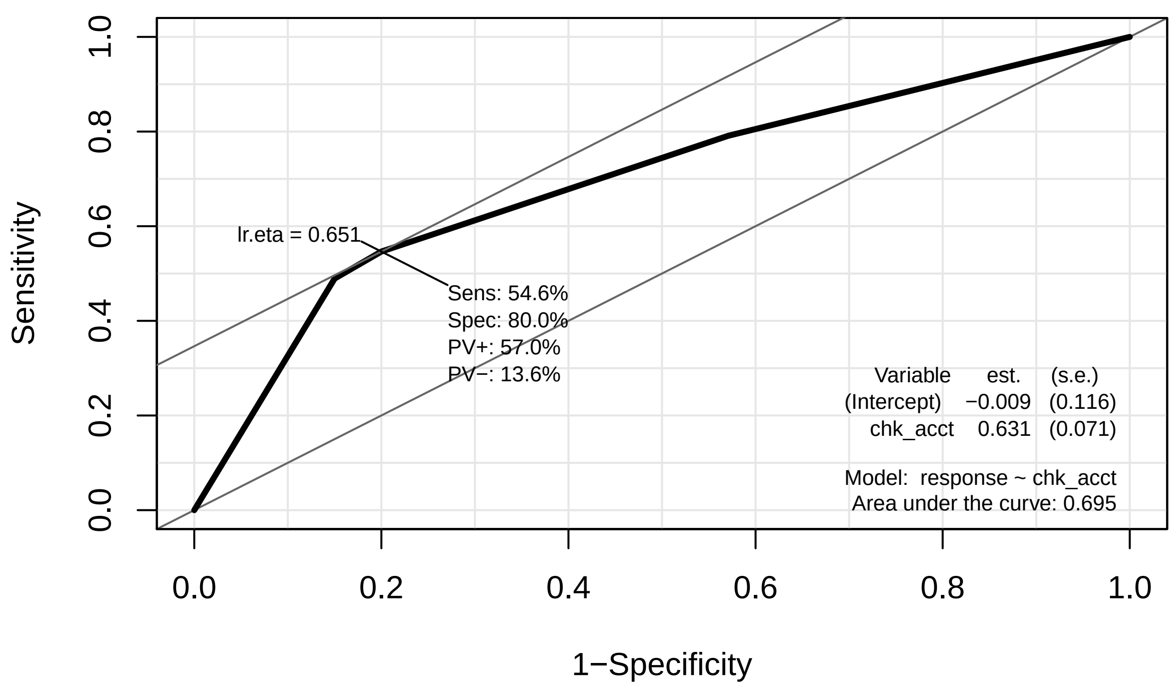
We remove *own_res* and *real_estate* as they represent the same (but opposite) as *rent* and *prop_unkn_none*.

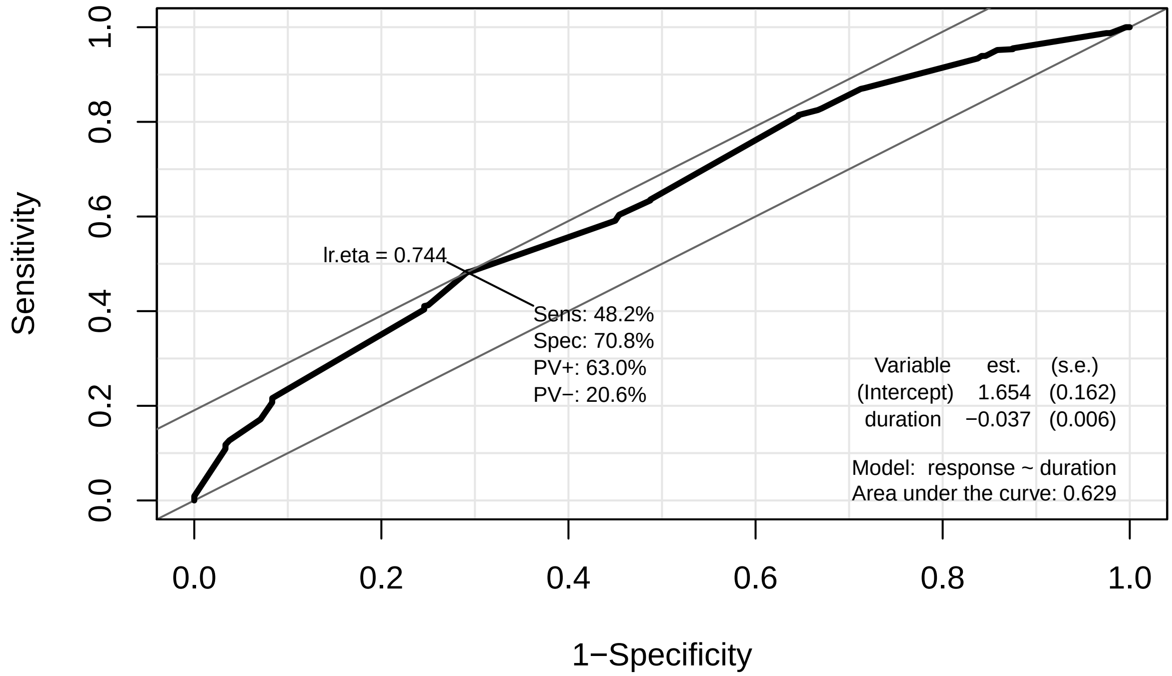
We run models using every single variable:

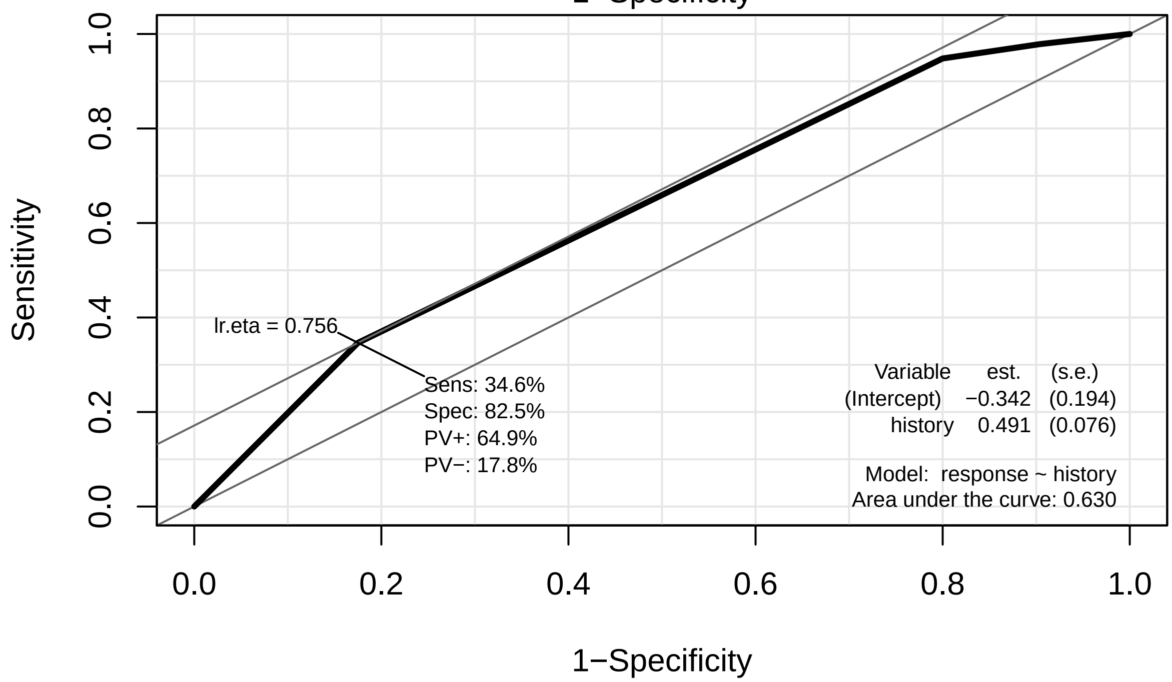
1-variable models:

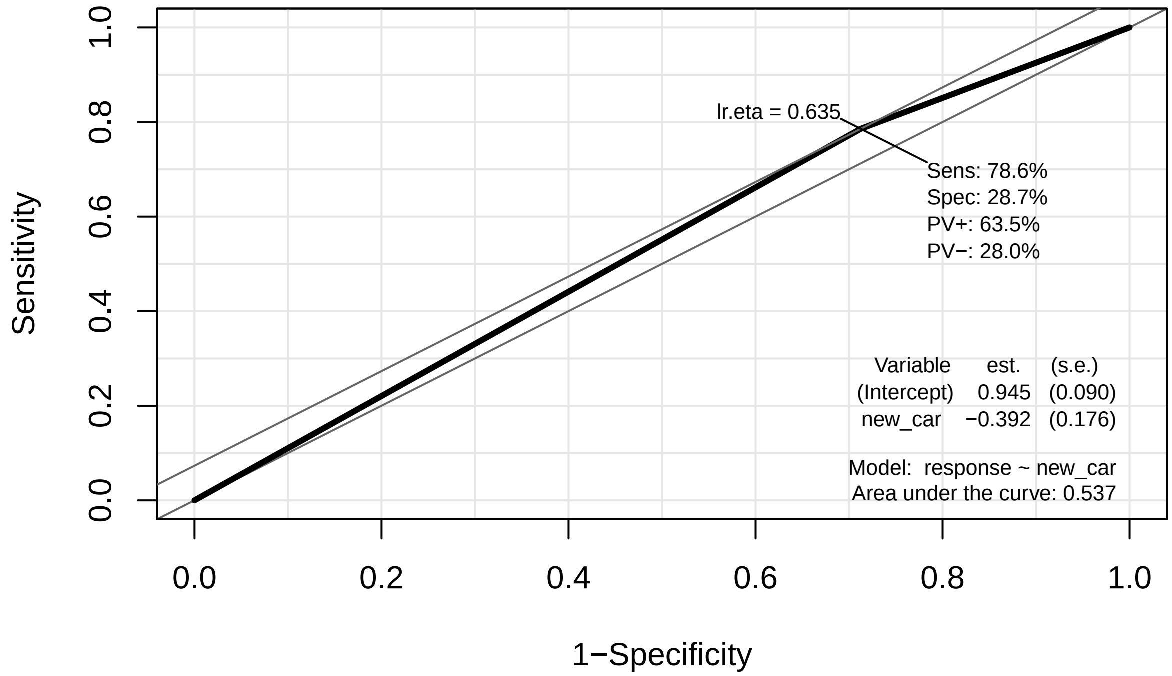
```
cols <- names(credit)[1:(length(names(credit))-1)]
vars <- c()
```

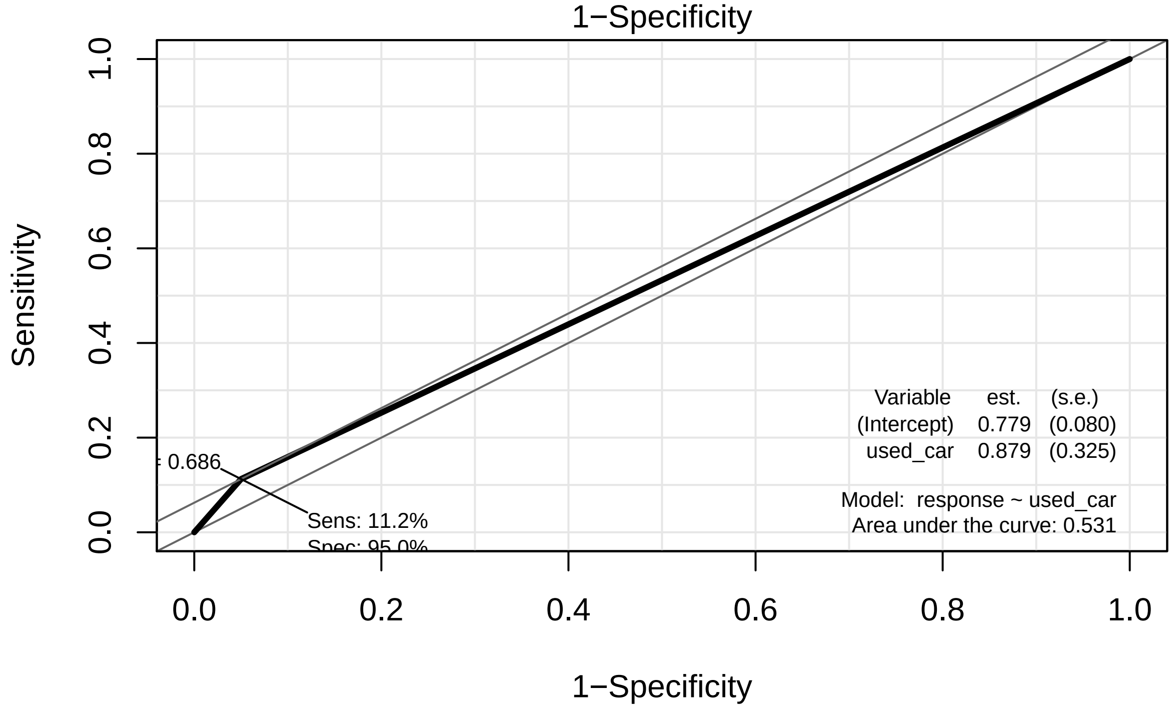
```r
acc <- c()
for (i in 1:length(cols)) {
    # formula and model
    form <- stringr::str_interp("response~${cols[i]}")
    mod <- glm(formula=form, family=binomial, data=Train)
    form <- formula(mod)

    # ROC curve and cutoff
    roc1 <- Epi::ROC(form=form, data=Train, plot="ROC", lw=3, cex=1.5)
    cutoff <- roc1$res$lr.eta[2]

    # prediction
    pred <- predict(mod, newdata=Test, type="response")
    pred <- ifelse(pred > cutoff, 1, 0)
    pred <- as.factor(pred)

    # confusion matrix and accuracy
    Accuracy <- confusionMatrix(pred, Test$response)$overall[1]

    # adding variables to prediction
    vars <- c(vars, cols[i])
    acc <- c(acc, Accuracy)
}
```

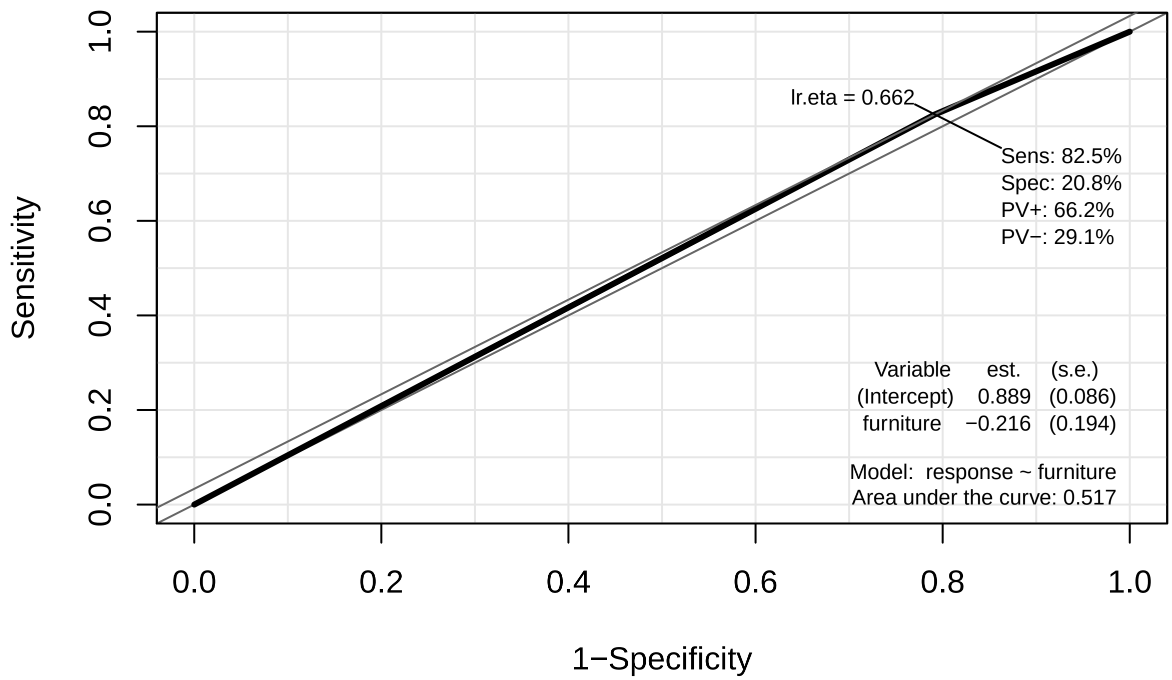


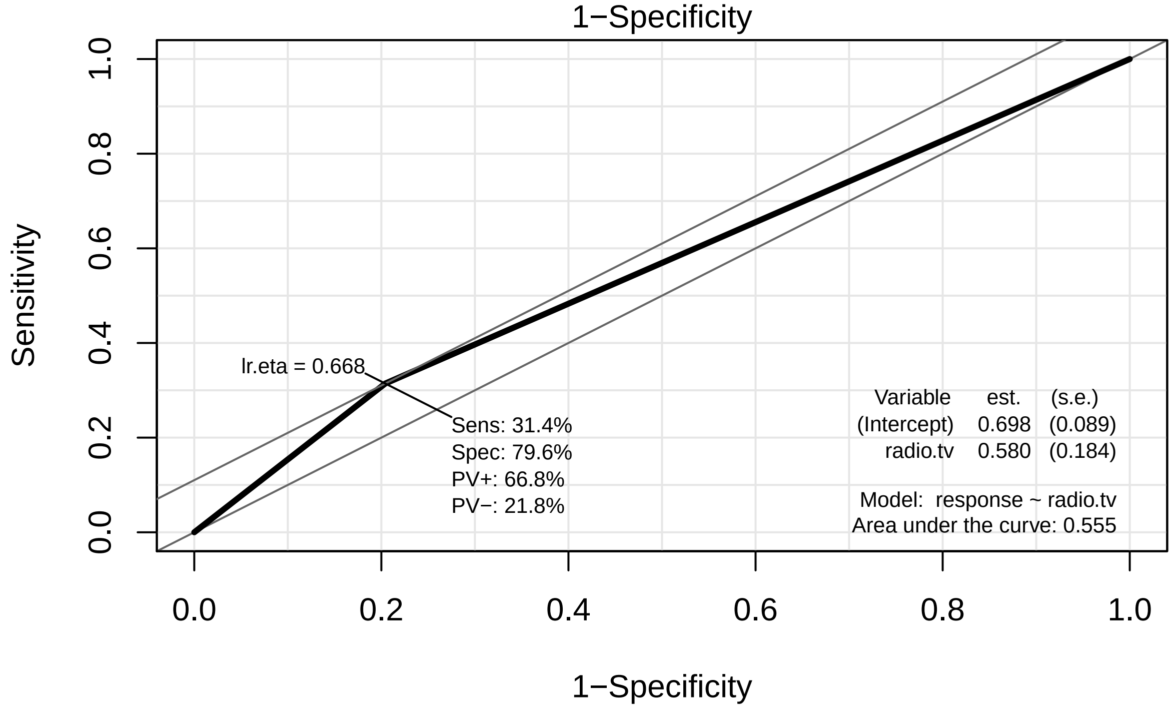

lr.eta = 0.744
Sens: 48.2%
Spec: 70.8%
PV+: 63.0%
PV−: 20.6%
Variable est. (s.e.)
(Intercept) 1.654 (0.162)
duration −0.037 (0.006)
Model: response ~ duration
Area under the curve: 0.629
Sensitivity
1−Specificity

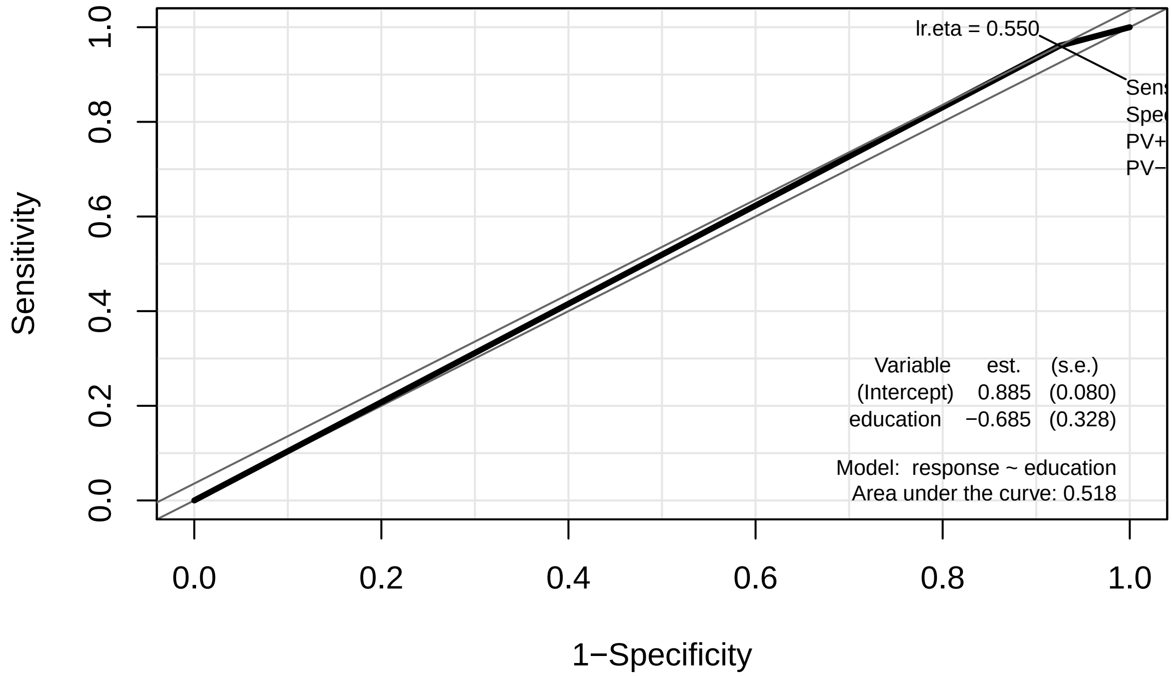

lr.eta = 0.756
Sens: 34.6%
Spec: 82.5%
PV+: 64.9%
PV−: 17.8%
Variable est. (s.e.)
(Intercept) −0.342 (0.194)
history 0.491 (0.076)
Model: response ~ history
Area under the curve: 0.630
Sensitivity
1−Specificity

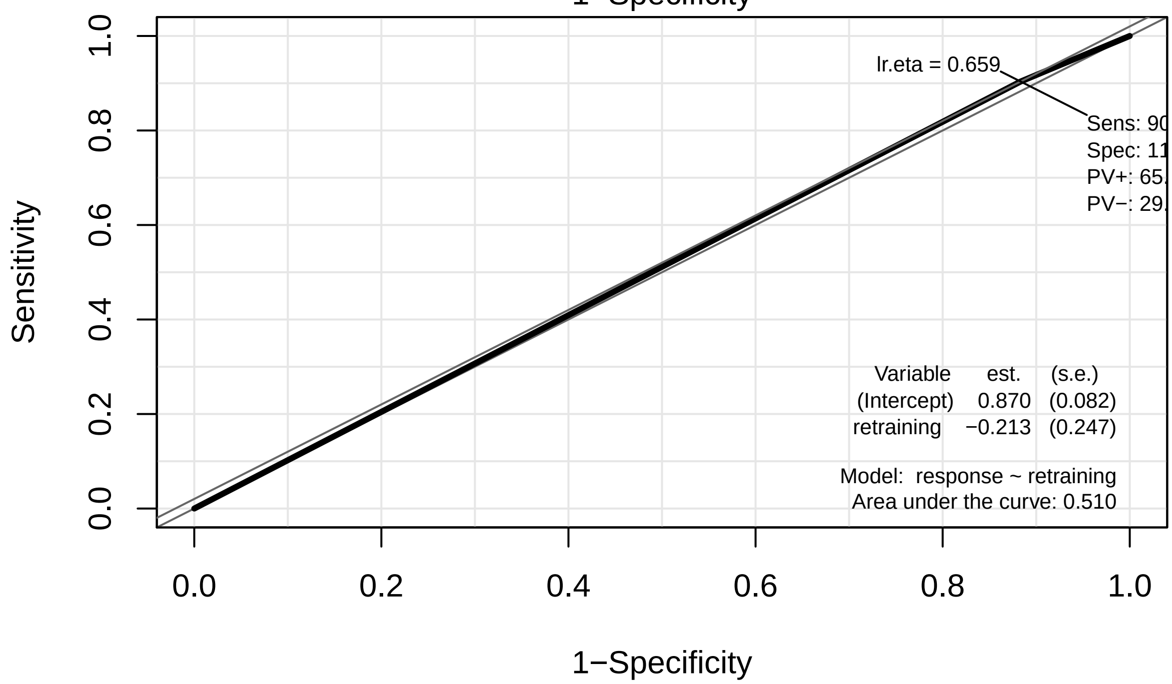

lr.eta = 0.635
Sens: 78.6%
Spec: 28.7%
PV+: 63.5%
PV−: 28.0%
Variable est. (s.e.)
(Intercept) 0.945 (0.090)
new_car −0.392 (0.176)
Model: response ~ new_car
Area under the curve: 0.537
Sensitivity
1−Specificity

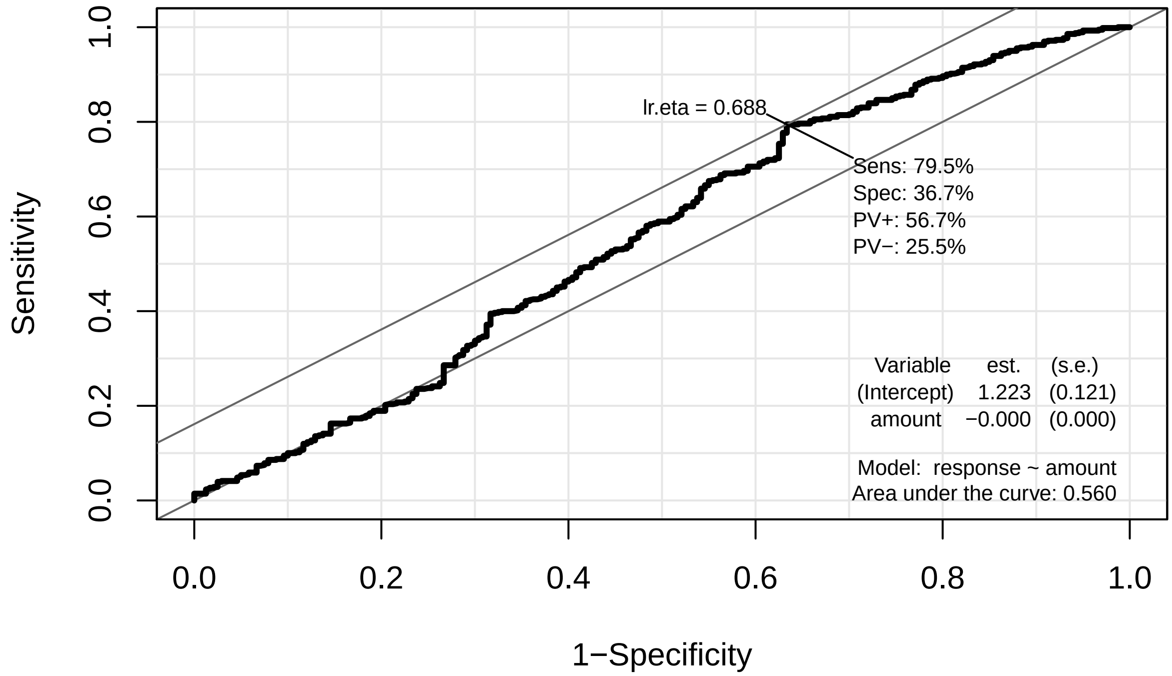

0.686
Sens: 11.2%
Variable est. (s.e.)
(Intercept) 0.779 (0.080)
used_car 0.879 (0.325)
Model: response ~ used_car
Area under the curve: 0.531
Sensitivity
1−Specificity


lr.eta = 0.662
Sens: 82.5%
Spec: 20.8%
PV+: 66.2%
PV−: 29.1%
Variable est. (s.e.)
(Intercept) 0.889 (0.086)
furniture −0.216 (0.194)
Model: response ~ furniture
Area under the curve: 0.517
Sensitivity
1−Specificity

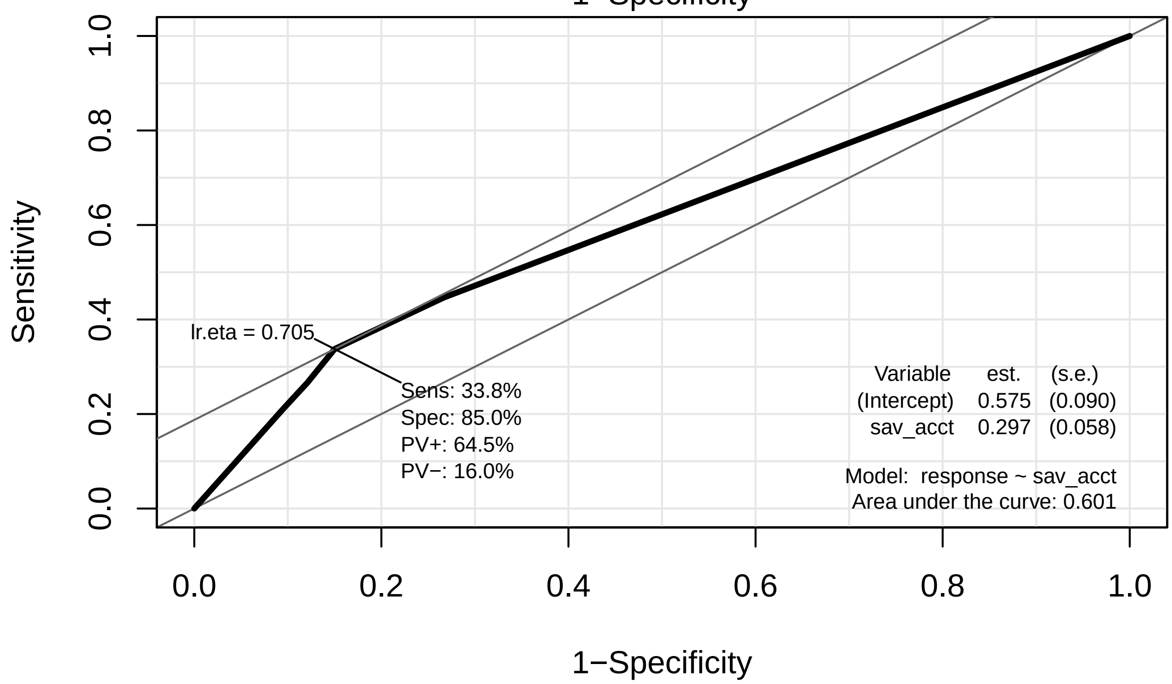

lr.eta = 0.668
Sens: 31.4%
Spec: 79.6%
PV+: 66.8%
PV−: 21.8%
Variable est. (s.e.)
(Intercept) 0.698 (0.089)
radio.tv 0.580 (0.184)
Model: response ~ radio.tv
Area under the curve: 0.555
Sensitivity
1−Specificity


lr.eta = 0.550
Sens
Spec
PV+
PV−
Variable est. (s.e.)
(Intercept) 0.885 (0.080)
education −0.685 (0.328)
Model: response ~ education
Area under the curve: 0.518
Sensitivity
1−Specificity

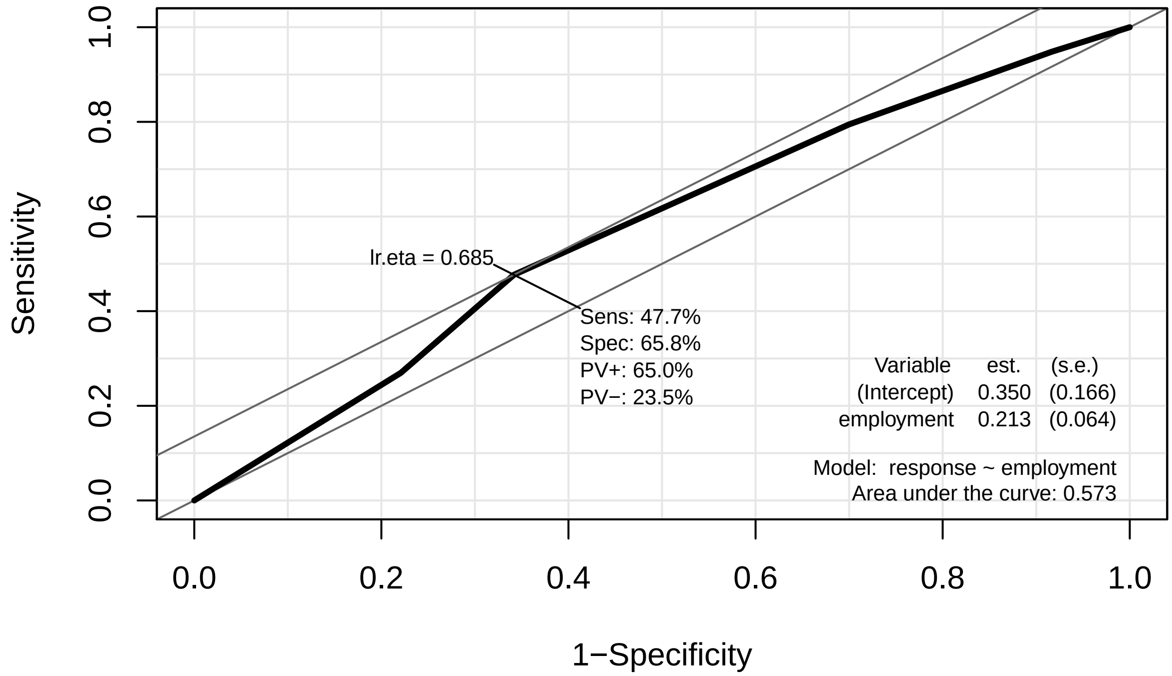

lr.eta = 0.659
Sens: 90
Spec: 11
PV+: 65.
PV−: 29.
Variable est. (s.e.)
(Intercept) 0.870 (0.082)
retraining −0.213 (0.247)
Model: response ~ retraining
Area under the curve: 0.510
Sensitivity
1−Specificity

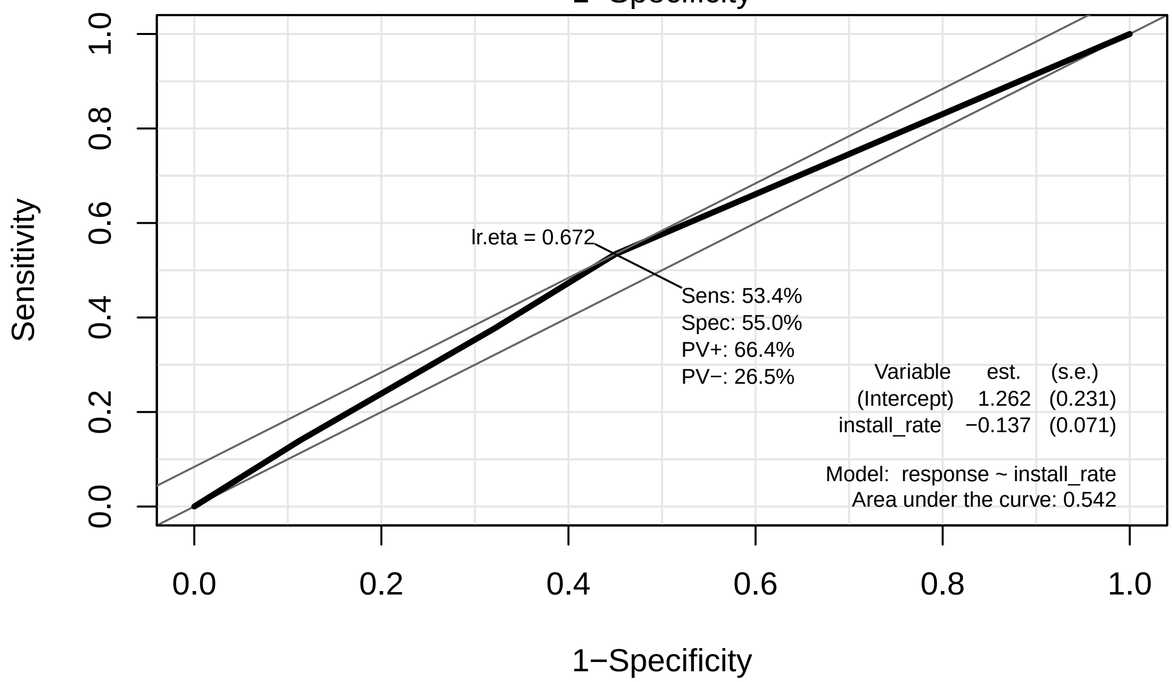


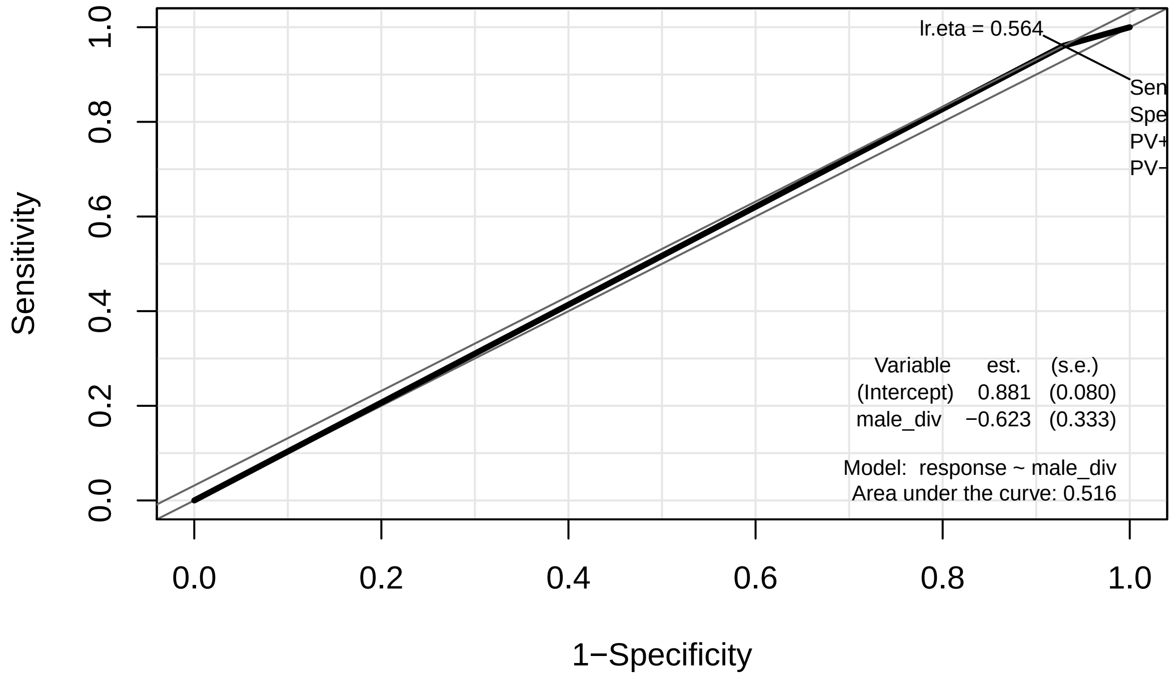
lr.eta = 0.688
Sens: 79.5%
Spec: 36.7%
PV+: 56.7%
PV−: 25.5%
Variable est. (s.e.)
(Intercept) 1.223 (0.121)
amount −0.000 (0.000)
Model: response ~ amount
Area under the curve: 0.560
Sensitivity
1−Specificity

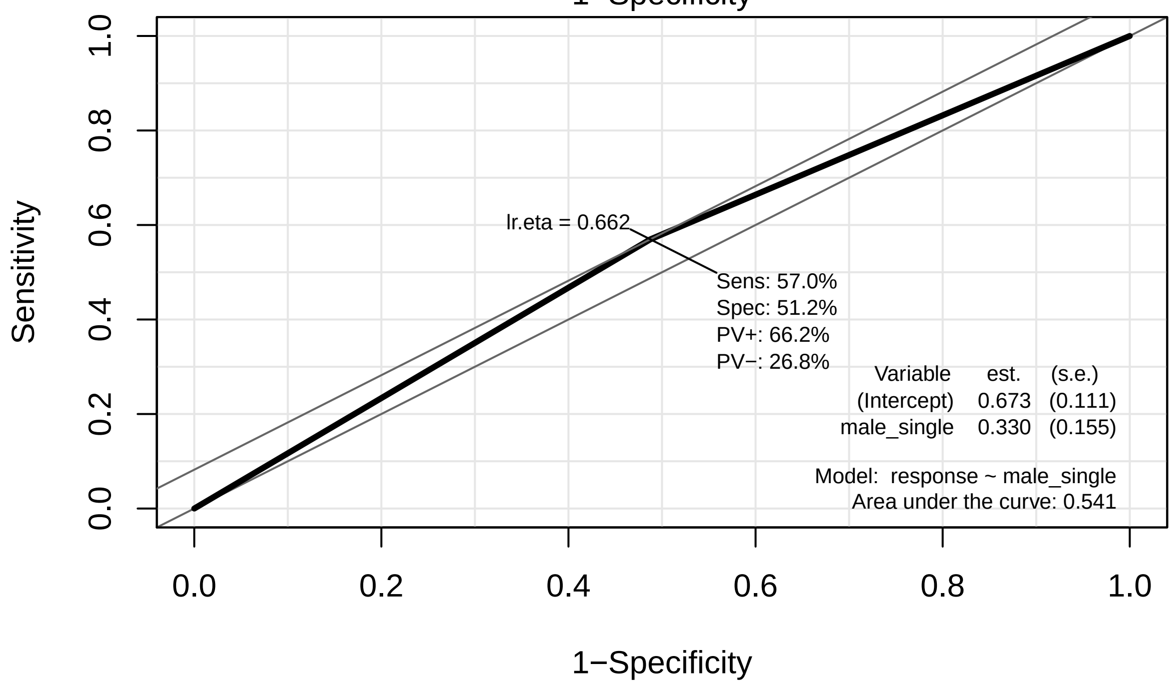


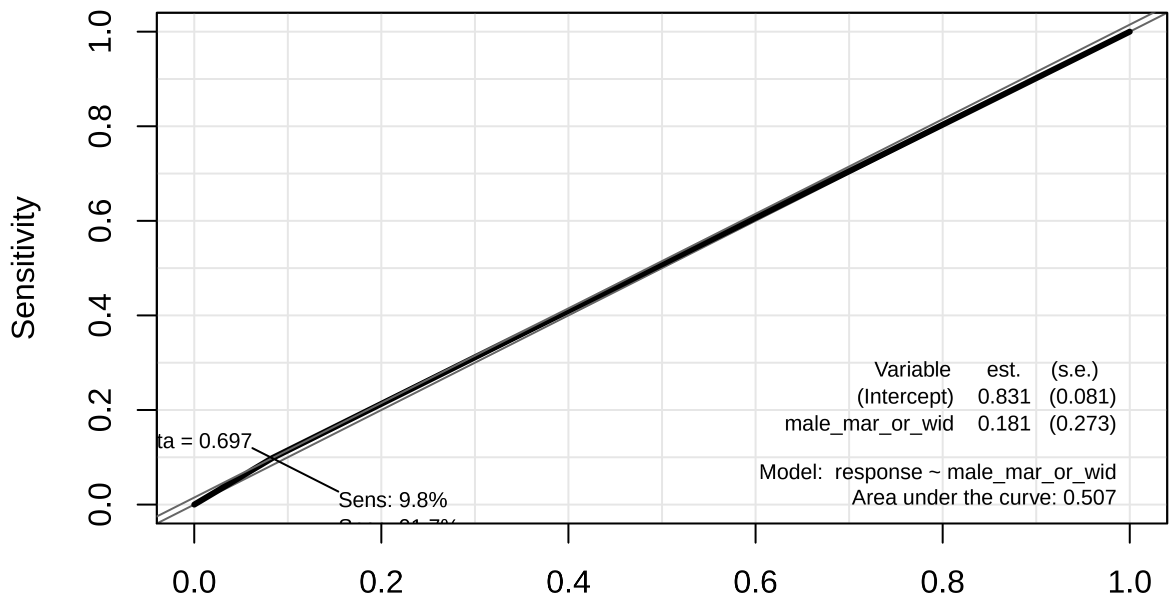
lr.eta = 0.705
Sens: 33.8%
Spec: 85.0%
PV+: 64.5%
PV−: 16.0%
Variable est. (s.e.)
(Intercept) 0.575 (0.090)
sav_acct 0.297 (0.058)
Model: response ~ sav_acct
Area under the curve: 0.601
Sensitivity
1−Specificity

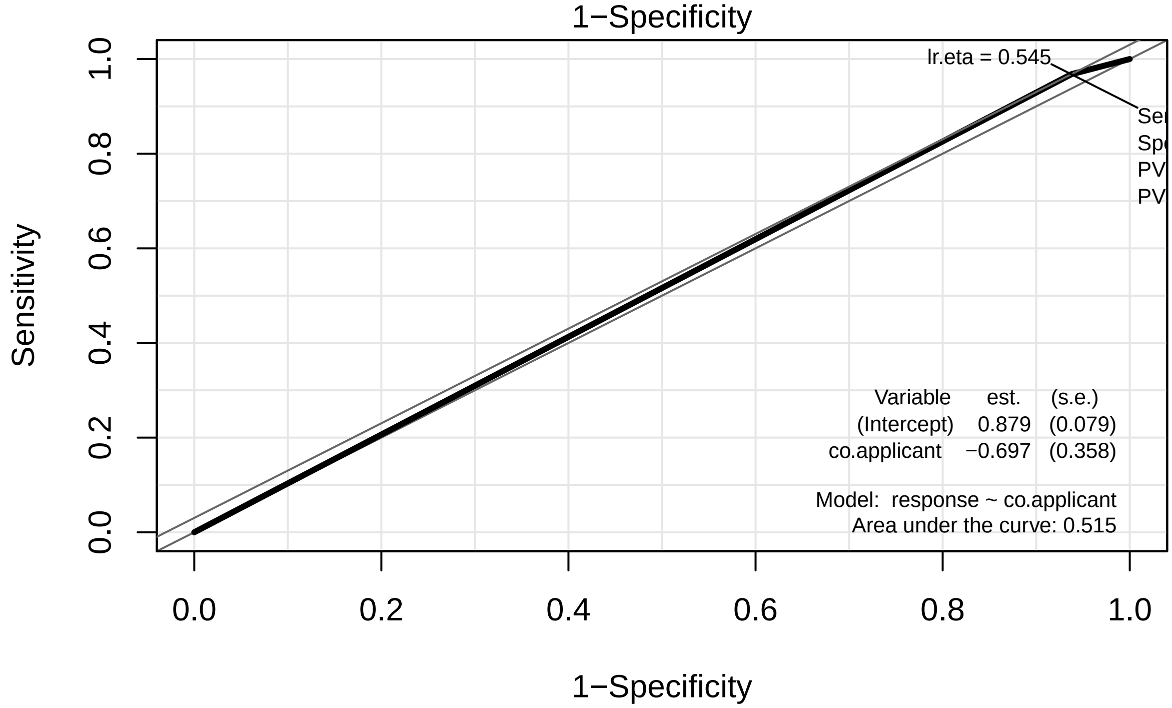

lr.eta = 0.685
Sens: 47.7%
Spec: 65.8%
PV+: 65.0%
PV−: 23.5%
Variable est. (s.e.)
(Intercept) 0.350 (0.166)
employment 0.213 (0.064)
Model: response ~ employment
Area under the curve: 0.573
Sensitivity
1−Specificity
0.0
0.2
0.4
0.6
0.8
1.0

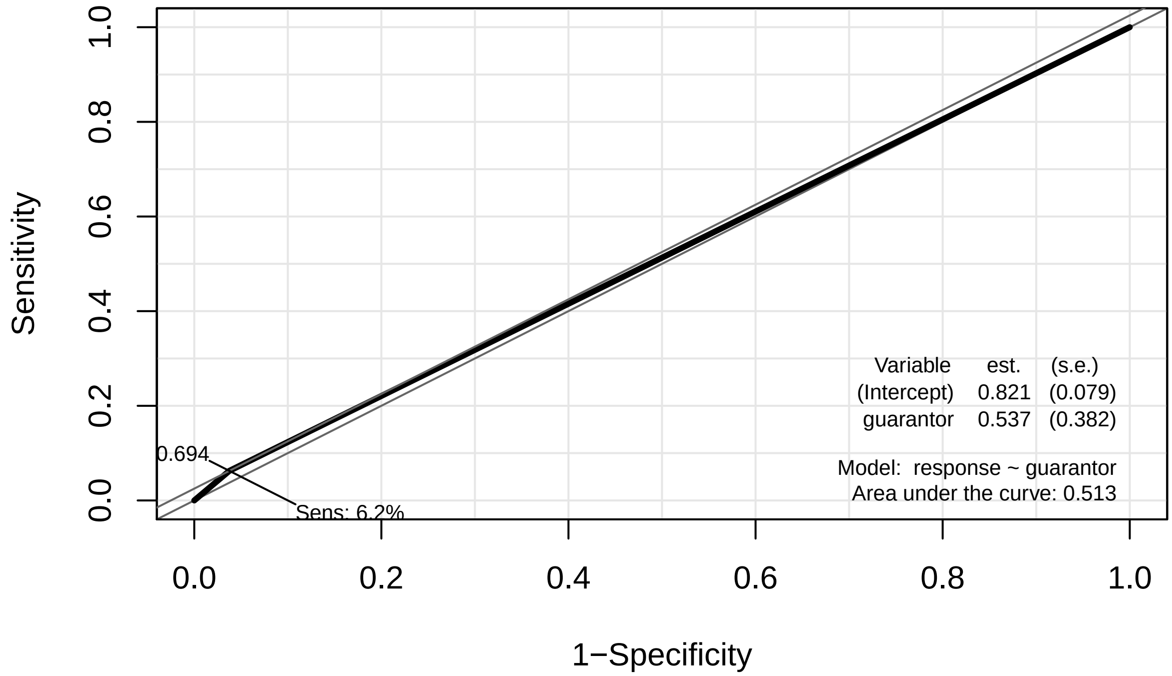

lr.eta = 0.672
Sens: 53.4%
Spec: 55.0%
PV+: 66.4%
PV−: 26.5%
Variable est. (s.e.)
(Intercept) 1.262 (0.231)
install_rate −0.137 (0.071)
Model: response ~ install_rate
Area under the curve: 0.542
Sensitivity
1−Specificity
0.0
0.2
0.4
0.6
0.8
1.0

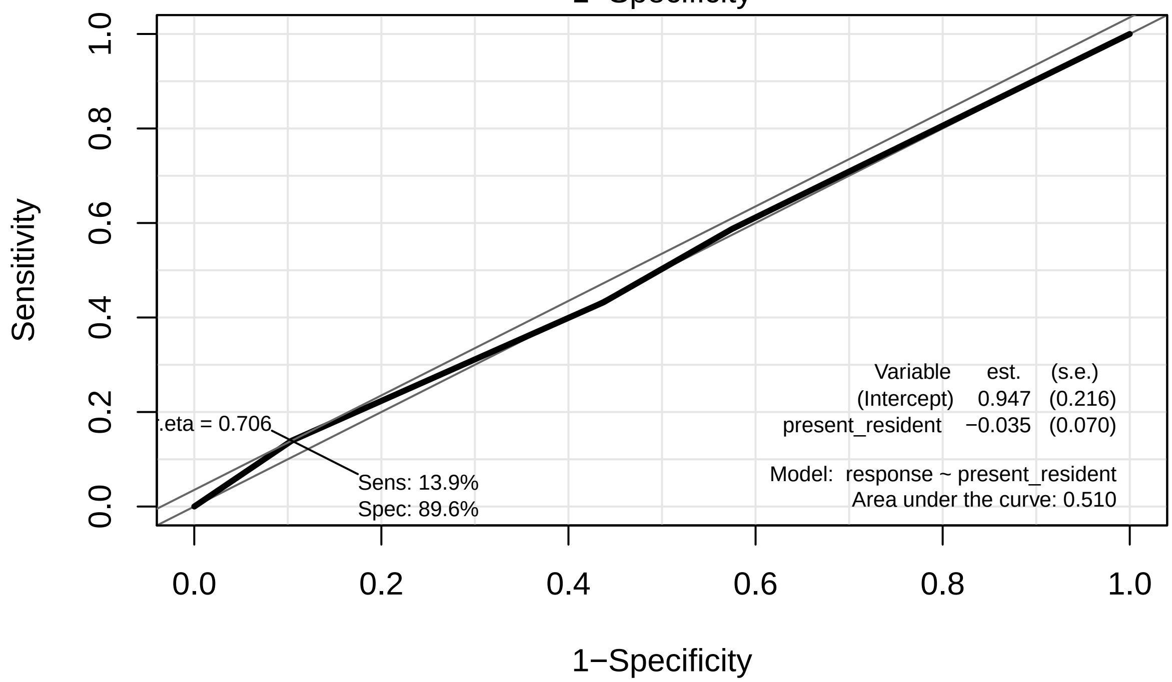

lr.eta = 0.564
Sen
Spe
PV+
PV−
Variable est. (s.e.)
(Intercept) 0.881 (0.080)
male_div −0.623 (0.333)
Model: response ~ male_div
Area under the curve: 0.516
Sensitivity
1−Specificity
0.0
0.2
0.4
0.6
0.8
1.0

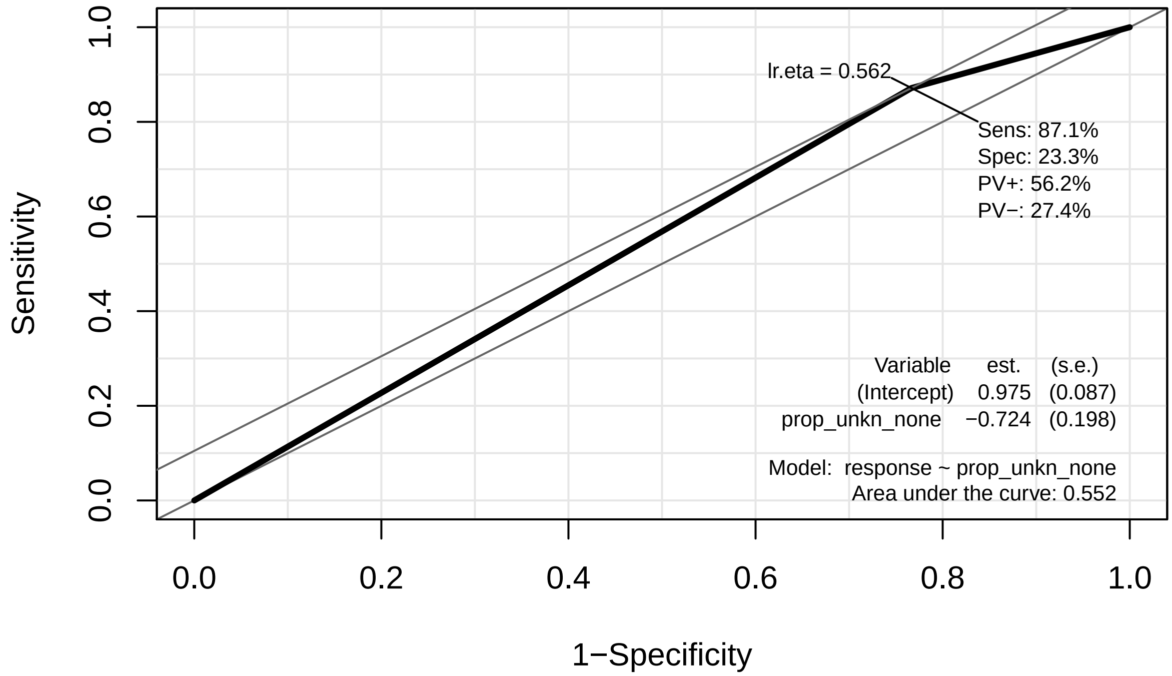

lr.eta = 0.662
Sens: 57.0%
Spec: 51.2%
PV+: 66.2%
PV−: 26.8%
Variable est. (s.e.)
(Intercept) 0.673 (0.111)
male_single 0.330 (0.155)
Model: response ~ male_single
Area under the curve: 0.541
Sensitivity
1−Specificity
0.0
0.2
0.4
0.6
0.8
1.0

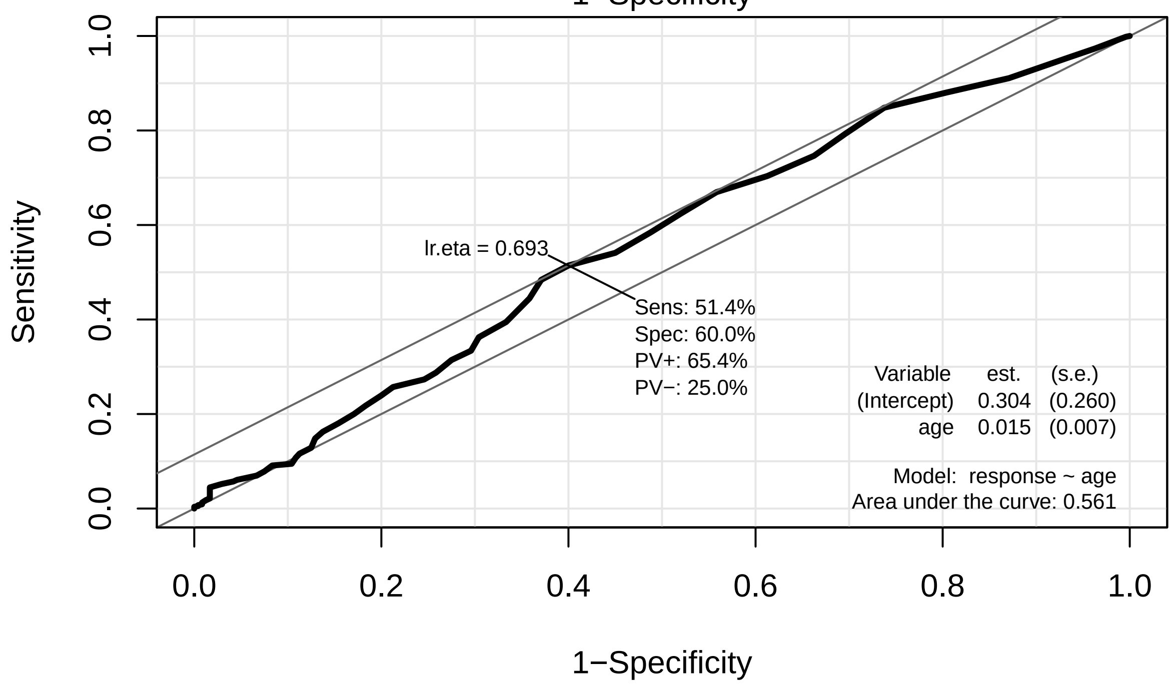

Sensitivity
1−Specificity
ta = 0.697
Sens: 9.8%
Variable est. (s.e.)
(Intercept) 0.831 (0.081)
male_mar_or_wid 0.181 (0.273)
Model: response ~ male_mar_or_wid
Area under the curve: 0.507

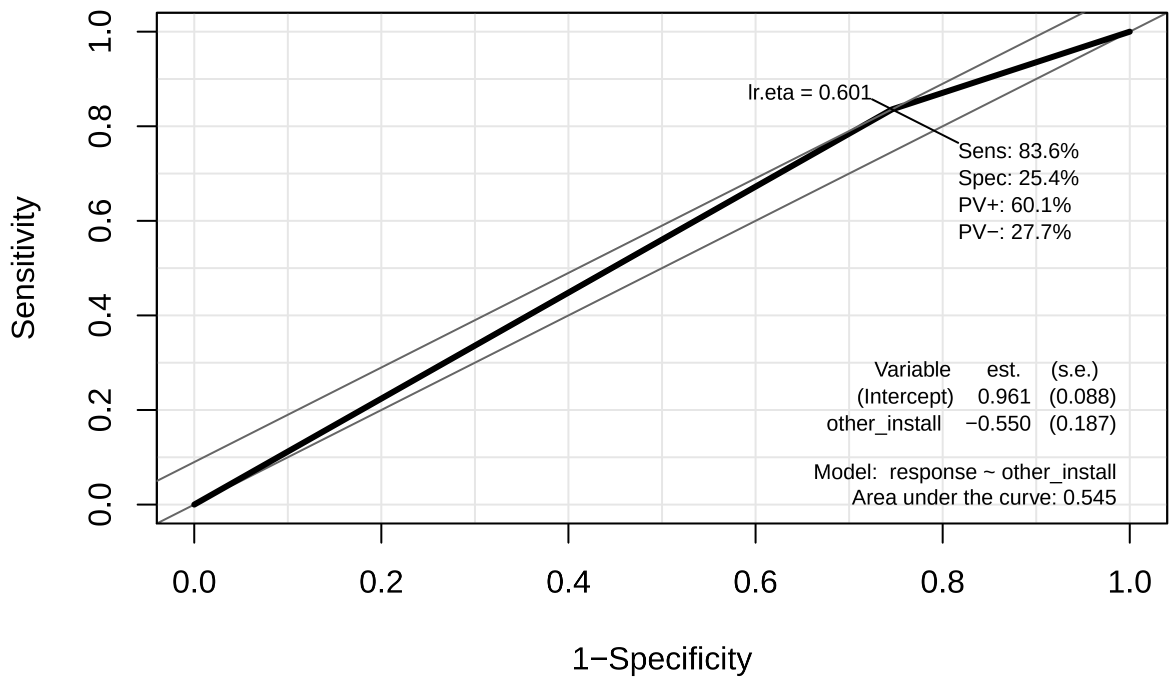

Sensitivity
1−Specificity
lr.eta = 0.545
Variable est. (s.e.)
(Intercept) 0.879 (0.079)
co.applicant −0.697 (0.358)
Model: response ~ co.applicant
Area under the curve: 0.515

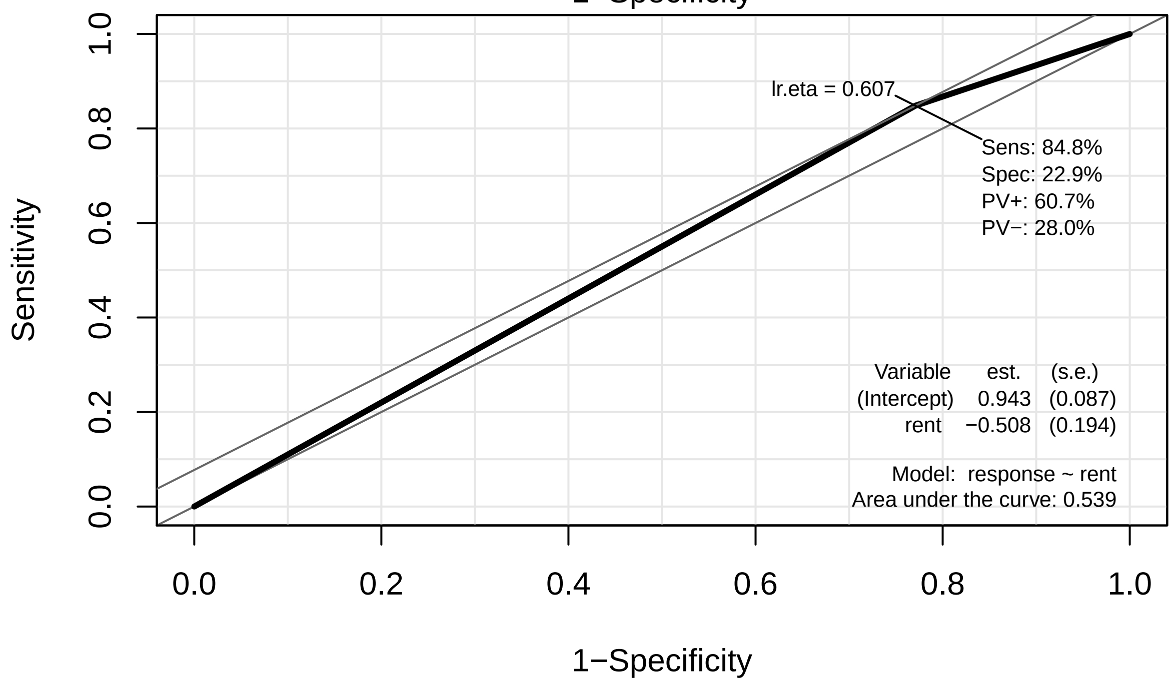

Sensitivity
1−Specificity
Variable est. (s.e.)
(Intercept) 0.821 (0.079)
guarantor 0.537 (0.382)
Model: response ~ guarantor
Area under the curve: 0.513
0.694
Sens: 6.2%

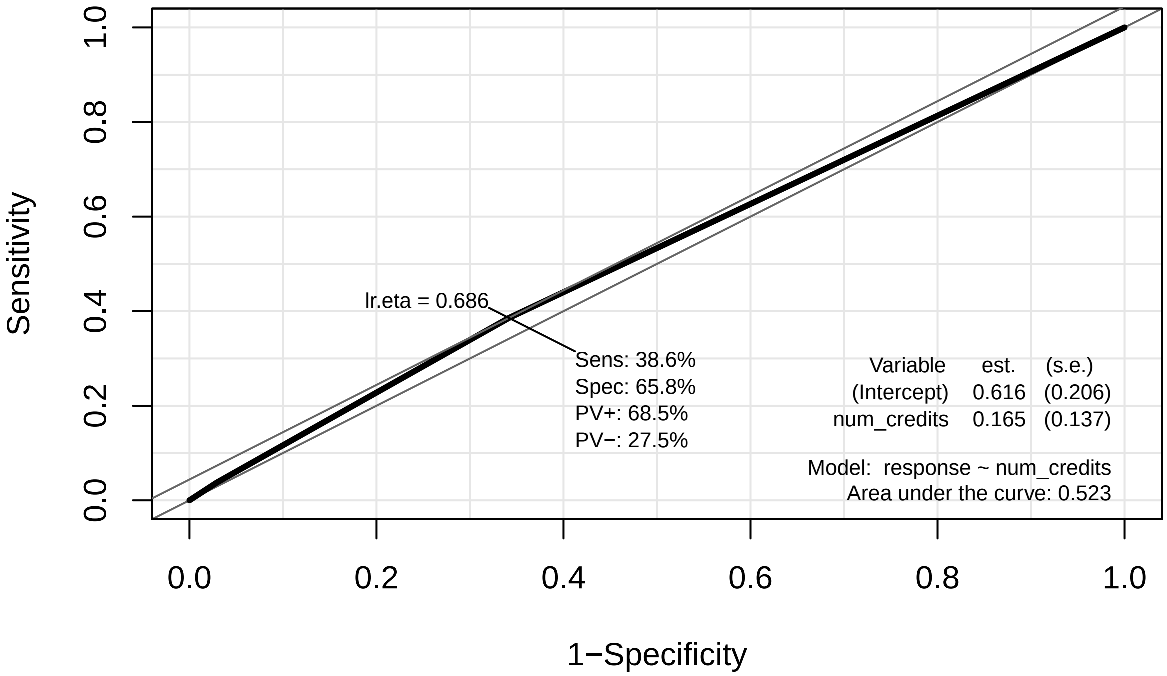

Sensitivity
1−Specificity
Variable est. (s.e.)
(Intercept) 0.947 (0.216)
present_resident −0.035 (0.070)
Model: response ~ present_resident
Area under the curve: 0.510
.eta = 0.706
Sens: 13.9%
Spec: 89.6%

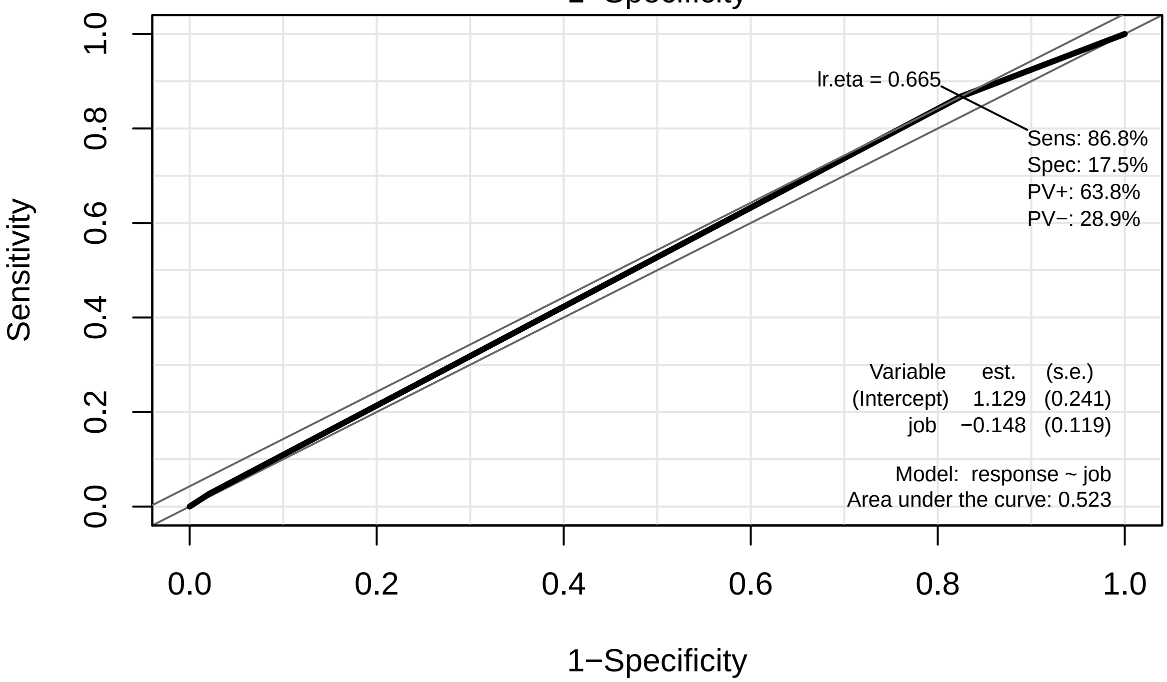

lr.eta = 0.562
Sens: 87.1%
Spec: 23.3%
PV+: 56.2%
PV−: 27.4%
Variable est. (s.e.)
(Intercept) 0.975 (0.087)
prop_unkn_none −0.724 (0.198)
Model: response ~ prop_unkn_none
Area under the curve: 0.552
Sensitivity
1−Specificity
0.0
0.2
0.4
0.6
0.8
1.0

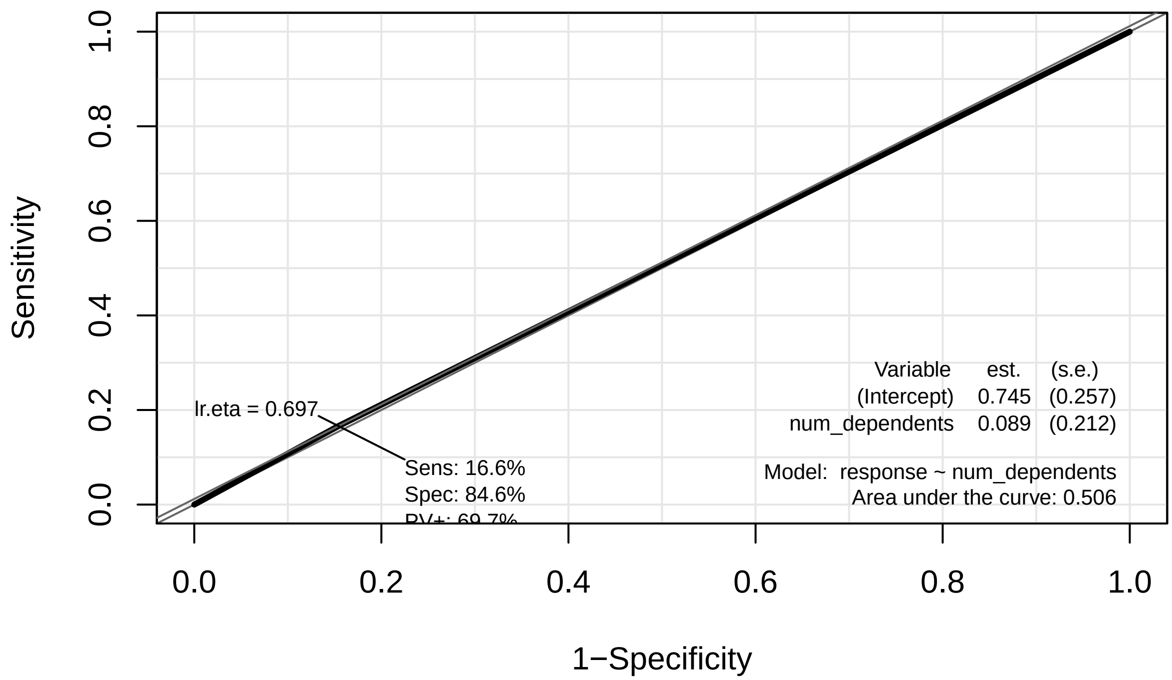

lr.eta = 0.693
Sens: 51.4%
Spec: 60.0%
PV+: 65.4%
PV−: 25.0%
Variable est. (s.e.)
(Intercept) 0.304 (0.260)
age 0.015 (0.007)
Model: response ~ age
Area under the curve: 0.561
Sensitivity
1−Specificity
0.0
0.2
0.4
0.6
0.8
1.0

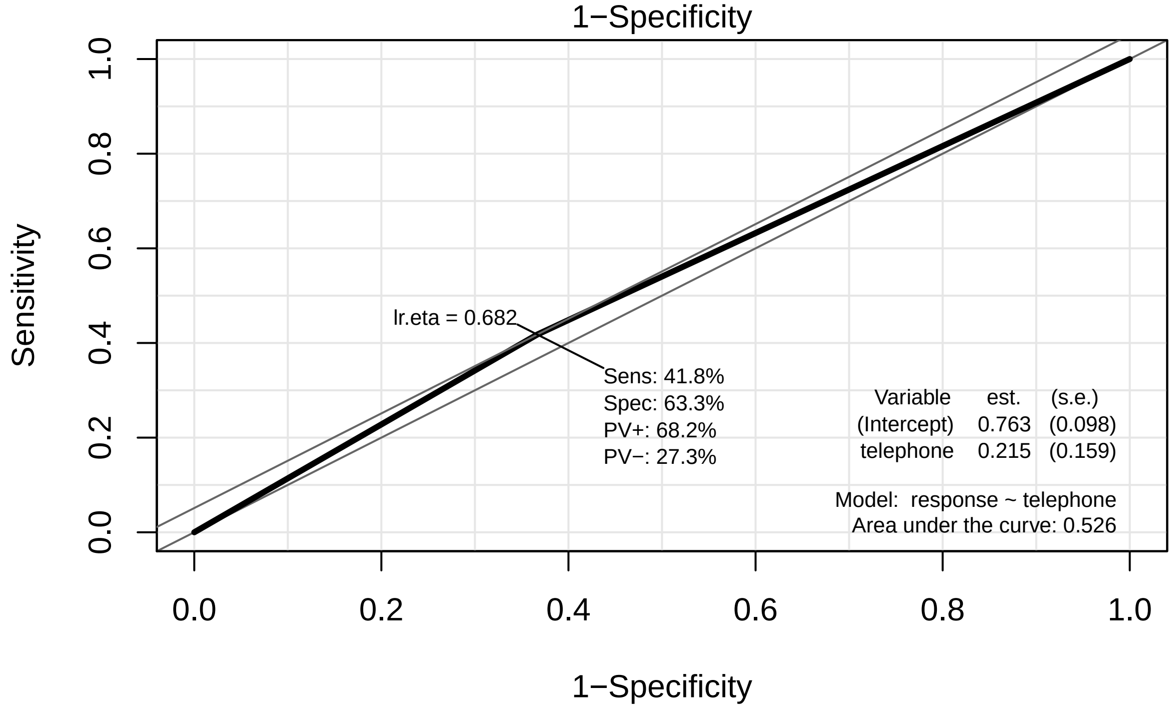

lr.eta = 0.601
Sens: 83.6%
Spec: 25.4%
PV+: 60.1%
PV−: 27.7%
Variable est. (s.e.)
(Intercept) 0.961 (0.088)
other_install −0.550 (0.187)
Model: response ~ other_install
Area under the curve: 0.545
Sensitivity
1−Specificity
0.0
0.2
0.4
0.6
0.8
1.0

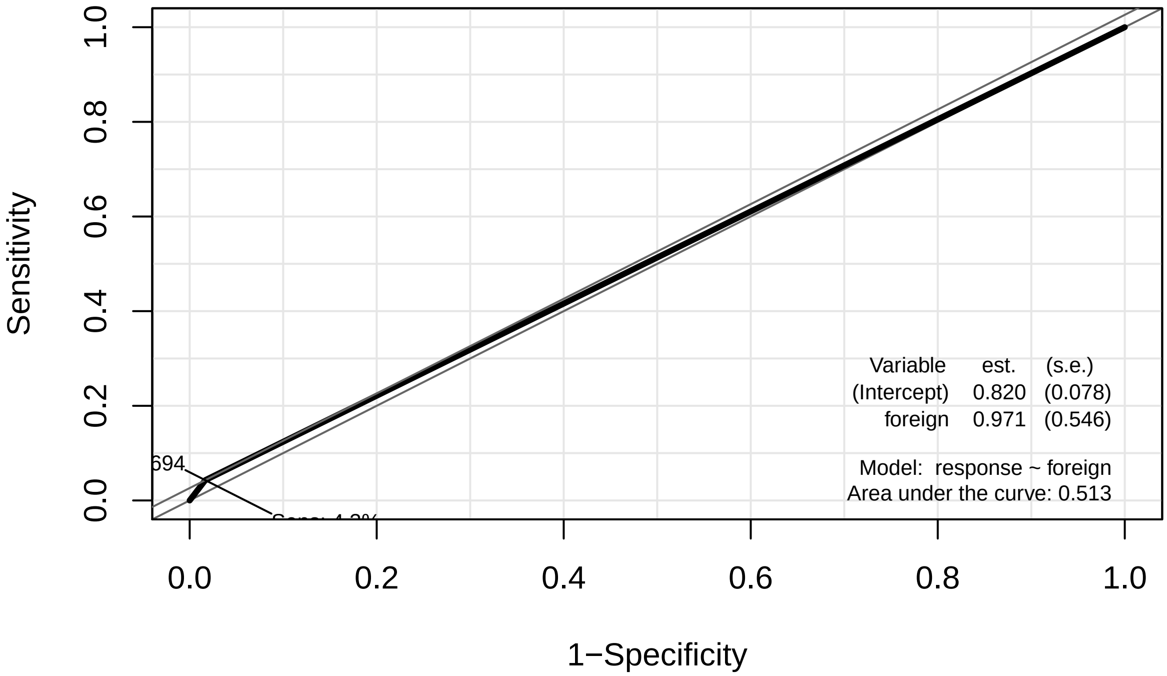

lr.eta = 0.607
Sens: 84.8%
Spec: 22.9%
PV+: 60.7%
PV−: 28.0%
Variable est. (s.e.)
(Intercept) 0.943 (0.087)
rent −0.508 (0.194)
Model: response ~ rent
Area under the curve: 0.539
Sensitivity
1−Specificity
0.0
0.2
0.4
0.6
0.8
1.0


Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
lr.eta = 0.686
Sens: 38.6%
Spec: 65.8%
PV+: 68.5%
PV−: 27.5%
Variable est. (s.e.)
(Intercept) 0.616 (0.206)
num_credits 0.165 (0.137)
Model: response ~ num_credits
Area under the curve: 0.523
1−Specificity


Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
lr.eta = 0.665
Sens: 86.8%
Spec: 17.5%
PV+: 63.8%
PV−: 28.9%
Variable est. (s.e.)
(Intercept) 1.129 (0.241)
job −0.148 (0.119)
Model: response ~ job
Area under the curve: 0.523
1−Specificity


Sensitivity
1−Specificity
lr.eta = 0.697
Sens: 16.6%
Spec: 84.6%
Variable est. (s.e.)
(Intercept) 0.745 (0.257)
num_dependents 0.089 (0.212)
Model: response ~ num_dependents
Area under the curve: 0.506
0.0
0.2
0.4
0.6
0.8
1.0

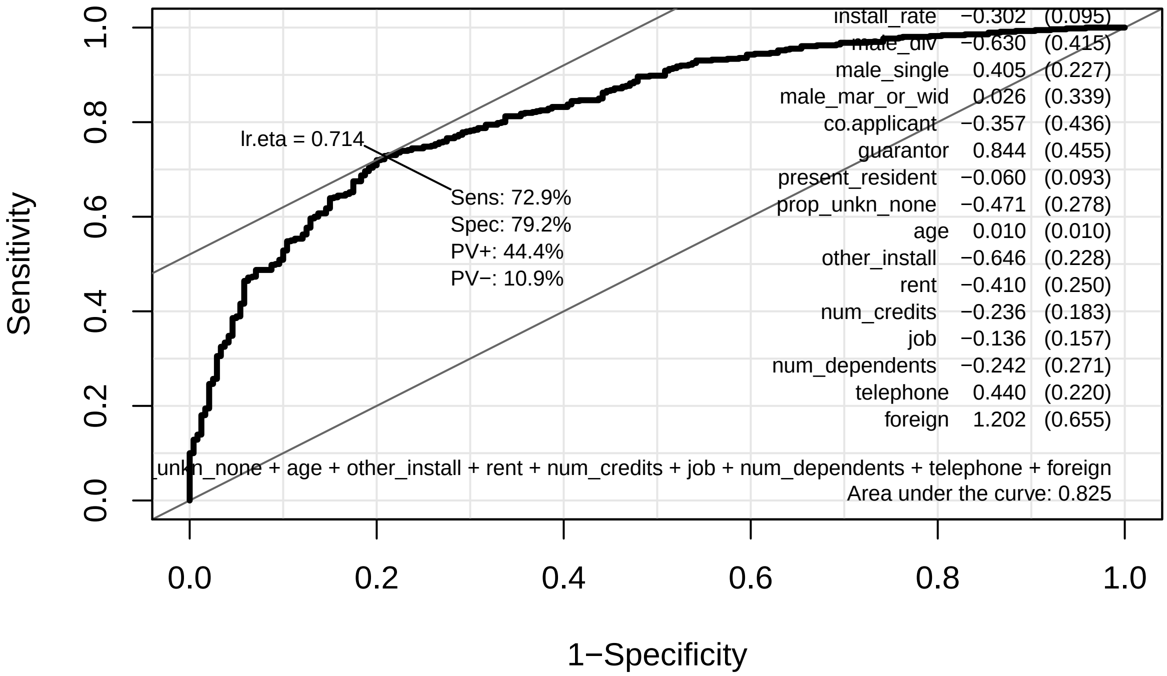

Sensitivity
1−Specificity
lr.eta = 0.682
Sens: 41.8%
Spec: 63.3%
PV+: 68.2%
PV−: 27.3%
Variable est. (s.e.)
(Intercept) 0.763 (0.098)
telephone 0.215 (0.159)
Model: response ~ telephone
Area under the curve: 0.526
0.0
0.2
0.4
0.6
0.8
1.0




```
df <- data.frame(vars=vars, accuracy=acc)
df[order(df$accuracy),]
#>                 vars accuracy
#> 18         guarantor    0.330
#> 16   male_mar_or_wid    0.335
#> 26    num_dependents    0.335
#> 28           foreign    0.345
#> 5           used_car    0.390
#> 7           radio.tv    0.445
#> 27         telephone    0.460
#> 24       num_credits    0.465
#> 19  present_resident    0.520
#> 11          sav_acct    0.525
#> 13      install_rate    0.555
#> 15       male_single    0.570
#> 6          furniture    0.615
#> 25               job    0.630
#> 23              rent    0.655
#> 12        employment    0.665
#> 4            new_car    0.675
#> 14          male_div    0.675
#> 20    prop_unkn_none    0.680
#> 9         retraining    0.685
#> 22     other_install    0.685
#> 8          education    0.690
#> 17      co.applicant    0.690
#> 3            history    0.695
#> 2           duration    0.700
#> 10            amount    0.700
#> 21               age    0.700
#> 1           chk_acct    0.750
```

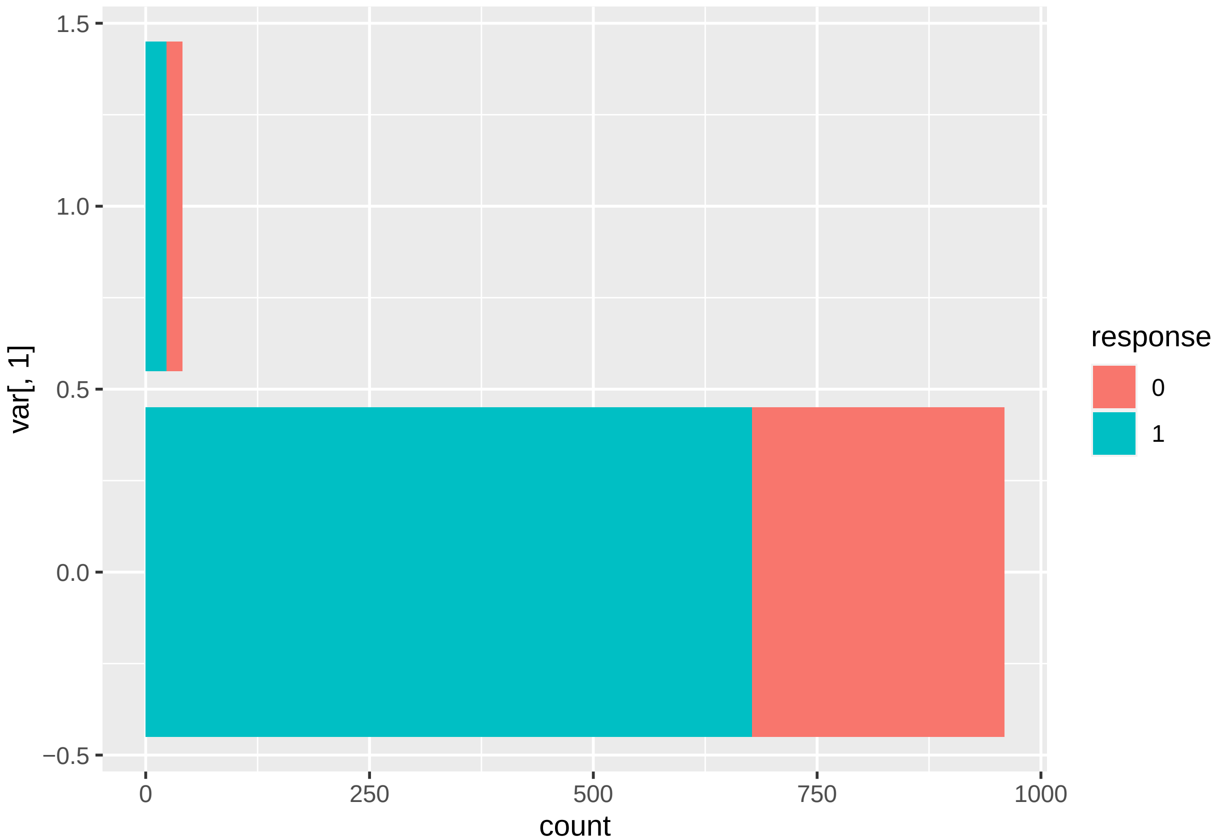
We run a model with *all the variables*:

```r
mod <- glm(formula=response~., family=binomial, data=Train)
roc1 <- Epi::ROC(form=formula(mod), data=Train, plot="ROC", lw=3, cex=1.5)
```

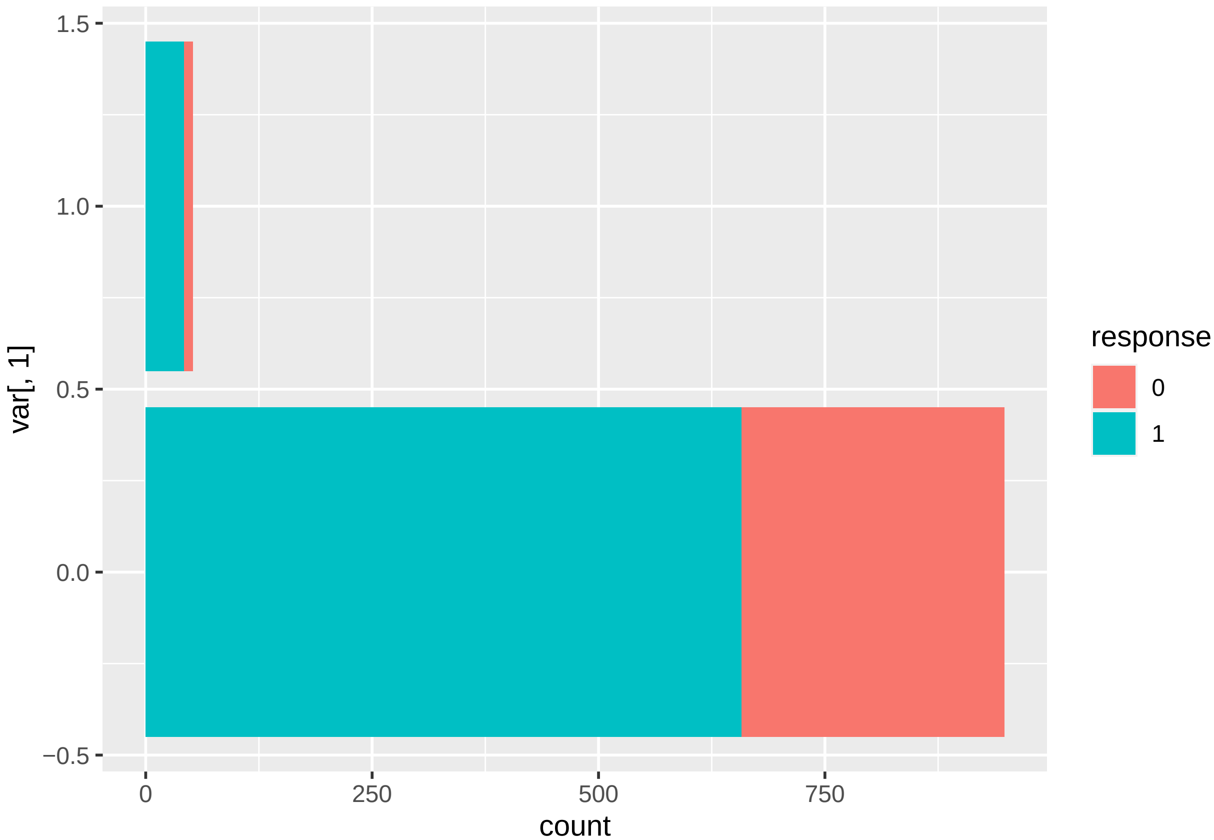



```r
cutoff <- which.max(rowSums(roc1$res[, c("sens", "spec")]))
prediction <- predict(mod, newdata=Test, type="response")
prediction <- ifelse(prediction > roc1$res$lr.eta[cutoff], 1, 0)
pred <- as.factor(prediction)
confusionMatrix(pred, Test$response)$overall[1]
#> Accuracy
#>    0.785
```

```r
plt <- function(col) {
    print(col)
    var <- credit %>% dplyr::select(col)
    ggplot(credit, aes(y=var[,1], fill=response)) + geom_bar()
}

plt('co.applicant')
#> [1] "co.applicant"
```




```
plt('guarantor')
#> [1] "guarantor"
```



Checking 2-variable models and interactions:

```
cool_stuff <- na.omit(cool_stuff[cool_stuff$pvals < 0.02,])
model_numbers <- as.numeric(rownames(cool_stuff))
all_vars <- list()
for (i in 1:length(model_numbers)) {
    all_vars[[i]] <- all.vars(formula(models_with[[model_numbers[i]]])[-2])
    all_vars[[i]] <- c(all_vars[[i]], paste(all_vars[[i]],collapse=":"))
}
vars <- c()
for (i in 1:length(all_vars)) {
    vars <- c(vars, all_vars[[i]][1], all_vars[[i]][3])
}

test_model <- glm(form=str_interp("response~(${paste(vars, collapse='+')})^2"), family=binomial, data=T
staic <- stepAIC(test_model)
```

```
scores <- read.csv('./outputs/scores_with.csv')
scores_without <- read.csv('./outputs/scores_without.csv')
p_value <- read.csv('./outputs/2_var_models_LRT.csv')
scores = scores %>% rename("formulas_with"="formula")
a = merge(p_value, scores, by = "formulas_with")

scores_without = scores_without %>% rename("formulas_without"="formula")
b = merge(a,scores_without, by =  "formulas_without")
cool_stuff_2 <- b[b$accuracy.x>0.7,]
model_numbers <- as.numeric(rownames(cool_stuff_2))
all_vars <- list()
```

```r
for (i in 1:length(model_numbers)) {
    all_vars[[i]] <- all.vars(formula(models_with[[model_numbers[i]]])[-2])
    all_vars[[i]] <- c(all_vars[[i]], paste(all_vars[[i]],collapse=":"))
}
vars <- c()
for (i in 1:length(all_vars)) {
    vars <- c(vars, all_vars[[i]][1],all_vars[[i]][3])
}
vars <- unique(vars)

test_model <- glm(form=str_interp("response~${paste(vars, collapse='+')}"), family=binomial, data=Train
staic <- stepAIC(test_model)

roc1 <- Epi::ROC(form=formula(staic), data=Train, plot="ROC", lw=3, cex=1.5)
cutoff <- which.max(rowSums(roc1$res[, c("sens", "spec")]))

prediction <- predict(staic, newdata=Test, type="response")
prediction <- ifelse(prediction > roc1$res$lr.eta[cutoff], 1, 0)
pred <- as.factor(prediction)
real_vals <- Test$response
confusionMatrix(pred, real_vals)

credit_2 <- credit %>% dplyr::select(vars)
```

```r
df <- data.frame(vars=vars, accuracy=acc)
df[order(df$accuracy),]
#>                 vars accuracy
#> 18         guarantor    0.330
#> 16   male_mar_or_wid    0.335
#> 26    num_dependents    0.335
#> 28           foreign    0.345
#> 5           used_car    0.390
#> 7           radio.tv    0.445
#> 27         telephone    0.460
#> 24       num_credits    0.465
#> 19 present_resident     0.520
#> 11          sav_acct    0.525
#> 13      install_rate    0.555
#> 15       male_single    0.570
#> 6          furniture    0.615
#> 25               job    0.630
#> 23              rent    0.655
#> 12        employment    0.665
#> 4            new_car    0.675
#> 14          male_div    0.675
#> 20    prop_unkn_none    0.680
#> 9         retraining    0.685
#> 22     other_install    0.685
#> 8          education    0.690
#> 17      co.applicant    0.690
#> 3            history    0.695
#> 2           duration    0.700
#> 10            amount    0.700
#> 21               age    0.700
```

```
#> 1           chk_acct     0.750
```