HW8_Logistics_Regression_Yuefei_Chen

Yuefei Chen

2024-04-20

Model development

Running the following code, we build a multiple regression model based on rent house data. Its independent variables "area", "rooms", "bathroom", "parking spaces", "hoa", "property tax", "fire insurance". The dependent variable is "furniture".

```
library(readr)
library(ggplot2)
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
house_data <- read_csv("Dataset/Rent_House_random_200_multi_regression.csv")
```

```
## Rows: 200 Columns: 11

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (3): floor, animal, furniture
## dbl (8): area, rooms, bathroom, parking_spaces, hoa, rent_amount, property_t...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
house_data <- house_data[, c(1:4, 6:11)]
str(house_data)
```

```
## tibble [200 x 10] (S3: tbl_df/tbl/data.frame)
##  $ area          : num [1:200] 120 45 50 35 204 177 15 70 180 180 ...
##  $ rooms         : num [1:200] 3 1 2 1 4 3 1 2 3 4 ...
##  $ bathroom      : num [1:200] 4 1 1 1 4 3 1 2 3 4 ...
##  $ parking_spaces: num [1:200] 3 1 1 0 2 4 0 1 2 2 ...
```

```
##  $ animal        : chr [1:200] "acept" "not acept" "acept" "acept" ...
##  $ furniture     : chr [1:200] "not furnished" "furnished" "not furnished" "not furnished" ...
##  $ hoa           : num [1:200] 1350 3000 226 260 0 2700 0 1800 700 2600 ...
##  $ rent_amount   : num [1:200] 5600 5520 750 1400 3440 6900 1200 4200 2700 2000 ...
##  $ property_tax  : num [1:200] 560 0 0 0 100 509 0 250 175 584 ...
##  $ fire_insurance: num [1:200] 71 70 10 18 62 89 16 55 40 26 ...
```

```
reg_data <- house_data[,-c(5)]
reg_data$furniture <- as.factor(reg_data$furniture)
#reg_data$area <- as.factor(reg_data$area)
#reg_data$rooms <- as.factor(reg_data$rooms)
#reg_data$bathroom <- as.factor(reg_data$rooms)
#reg_data$rent_amount <- as.factor(reg_data$rent_amount)
logistic <- glm(furniture~rooms+bathroom+area+rent_amount+property_tax, data=reg_data, family="binomial
```

Model Acceptance and Residual Analysis

In the summary of the model, we focus on R squared value, coefficients, and P-value of each coefficient. The R-squared value is 0.1201421 and p value is 0.0004620982. The "rent amount" is a significant independent variable in this model. The Pseudo R-square shows there is an improvement and better than baseline model. The p-value based on Likelihood Ratio Test shows that the model improvement is significant. The AIC value is 175.21.

```
summary(logistic)
```

```
##
## Call:
## glm(formula = furniture ~ rooms + bathroom + area + rent_amount +
##     property_tax, family = "binomial", data = reg_data)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.459e+00  4.726e-01   3.087 0.002019 **
## rooms         5.615e-01  3.103e-01   1.809 0.070415 .
## bathroom     -3.747e-01  2.550e-01  -1.469 0.141727
## area          5.310e-03  4.543e-03   1.169 0.242468
## rent_amount  -3.126e-04  8.897e-05  -3.514 0.000442 ***
## property_tax  9.175e-04  7.725e-04   1.188 0.234948
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 185.49  on 199  degrees of freedom
## Residual deviance: 163.21  on 194  degrees of freedom
## AIC: 175.21
##
## Number of Fisher Scoring iterations: 6
```

```
ll.null <- logistic$null.deviance/-2
ll.proposed <- logistic$deviance/-2
(ll.null - ll.proposed) / ll.null
```

```
## [1] 0.1201421
```

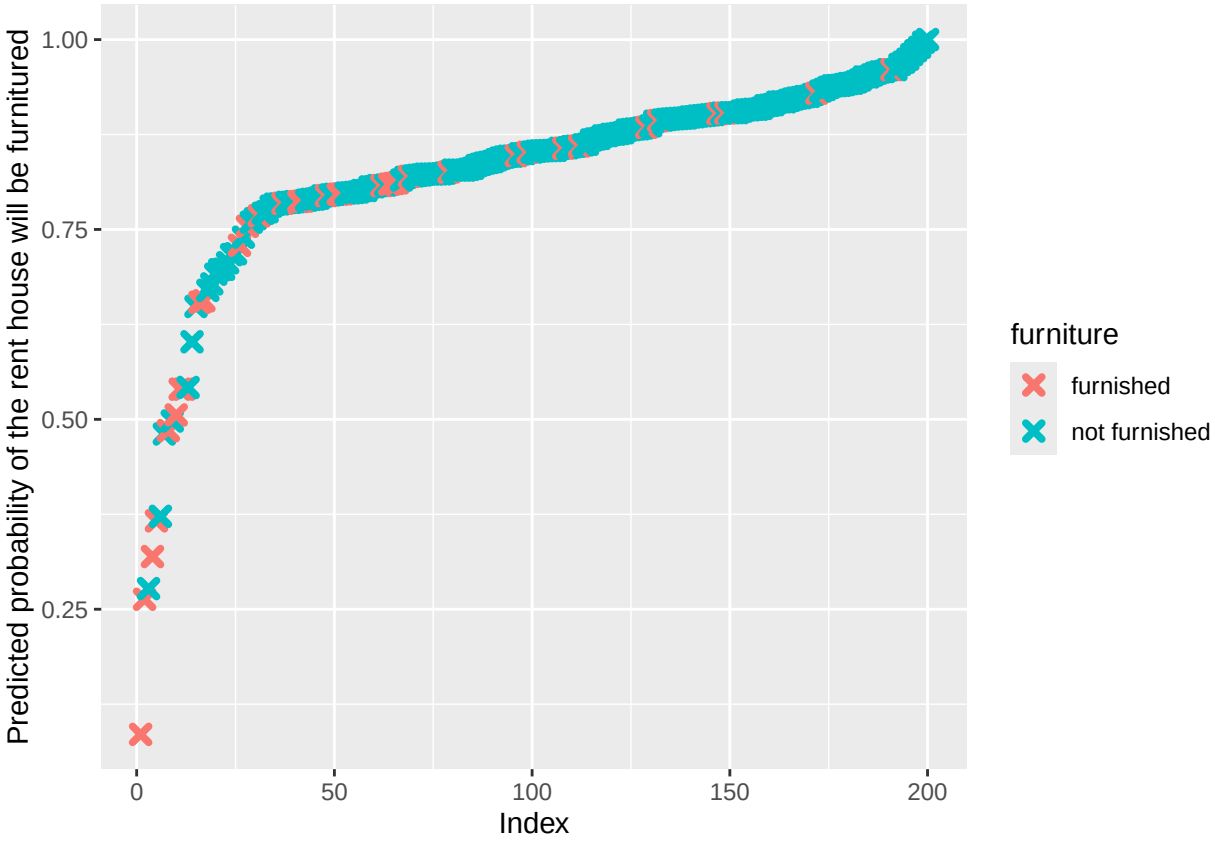
```
1 - pchisq(2*(ll.proposed - ll.null), df=(length(logistic$coefficients)-1))
```

```
## [1] 0.0004620982
```

Prediction

The data will be predicted in the model and the predicted probability of the rent house will be furnitured table is as follows.

```
predicted.data <- data.frame(probability.of.furniture=logistic$fitted.values,furniture=reg_data$furnitu
predicted.data <- predicted.data[order(predicted.data$probability.of.furniture, decreasing=FALSE),]
predicted.data$rank <- 1:nrow(predicted.data)
ggplot(data=predicted.data, aes(x=rank, y=probability.of.furniture)) +
geom_point(aes(color=furniture), alpha=1, shape=4, stroke=2) +
xlab("Index") +
ylab("Predicted probability of the rent house will be furnitured")
```



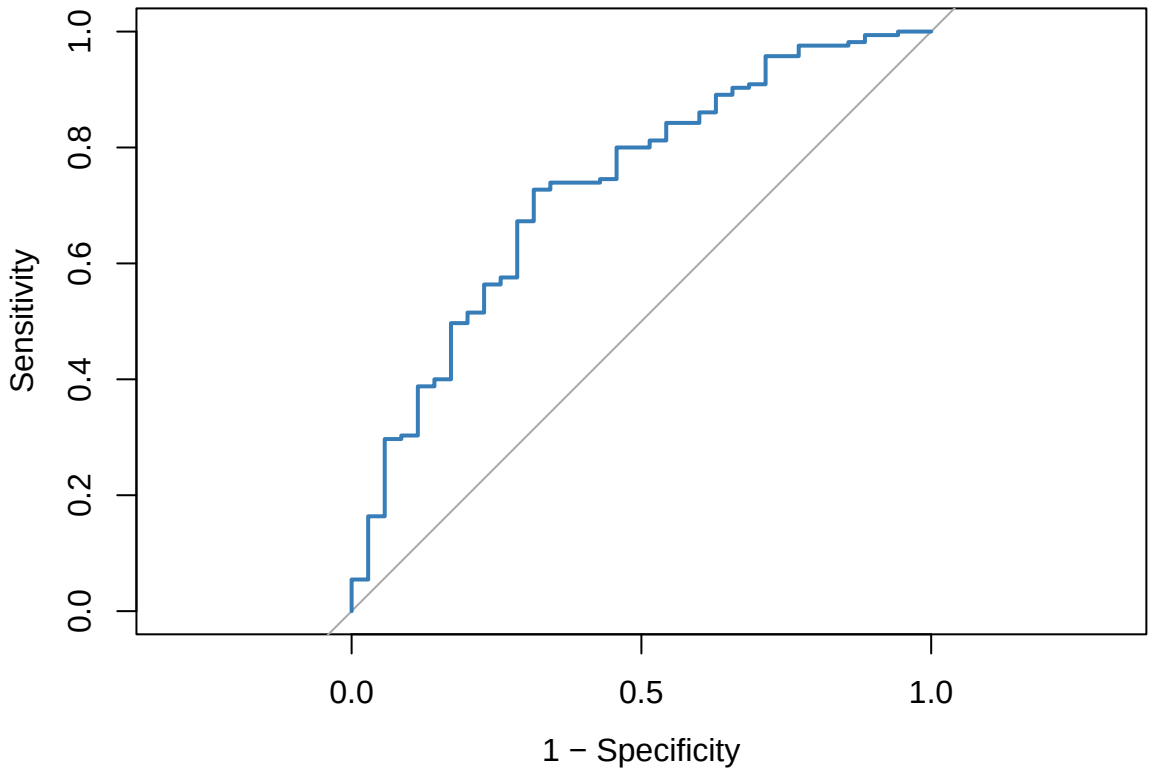
Model Accuracy

Receiver operating characteristic (ROC) curves is shown as follows. The area under the curve (AUC) is 0.7312, which means the accuracy of the model is considered acceptable.

```
roc(reg_data$furniture,logistic$fitted.values,plot=TRUE, legacy.axes=TRUE, col="#377eb8")
```

```
## Setting levels: control = furnished, case = not furnished
```

```
## Setting direction: controls < cases
```




```
##
## Call:
## roc.default(response = reg_data$furniture, predictor = logistic$fitted.values,     plot = TRUE, lega
##
## Data: logistic$fitted.values in 35 controls (reg_data$furniture furnished) < 165 cases (reg_data$furn
## Area under the curve: 0.7313
```