Vorhersagemodellierung VORLAGE

Gabi Musterstudentin, Horst Vozeigestudent, Biene Maja

2020-08-23

HINWEIS

Dokumentieren Sie Ihre Analyse anhand dieser Vorlage. Die Dokumentation ist für die Prüfung einzureichen.

Einleitung

Die Einleitung in die Aufgabenstellung usw. wird bei Gruppenarbeiten von allen zusammen geschrieben.

Ungefährer Umfang: 0,5–1,5 Seiten.

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

Explorative Datenanalyse und Datenvorverarbeitung[1]

Daten einlesen:

```
train <- read.csv2("../data/tips_train.csv")
anwendung <- read.csv2("../data/tips_test.csv")

inspect(train)  # Hat das Einlesen funktioniert?
```

```
## Warning: `data_frame()` is deprecated as of tibble 1.1.0.
## Please use `tibble()` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.

##
## categorical variables:
##     name     class levels   n missing
```

[1] Gabi Musterstudentin

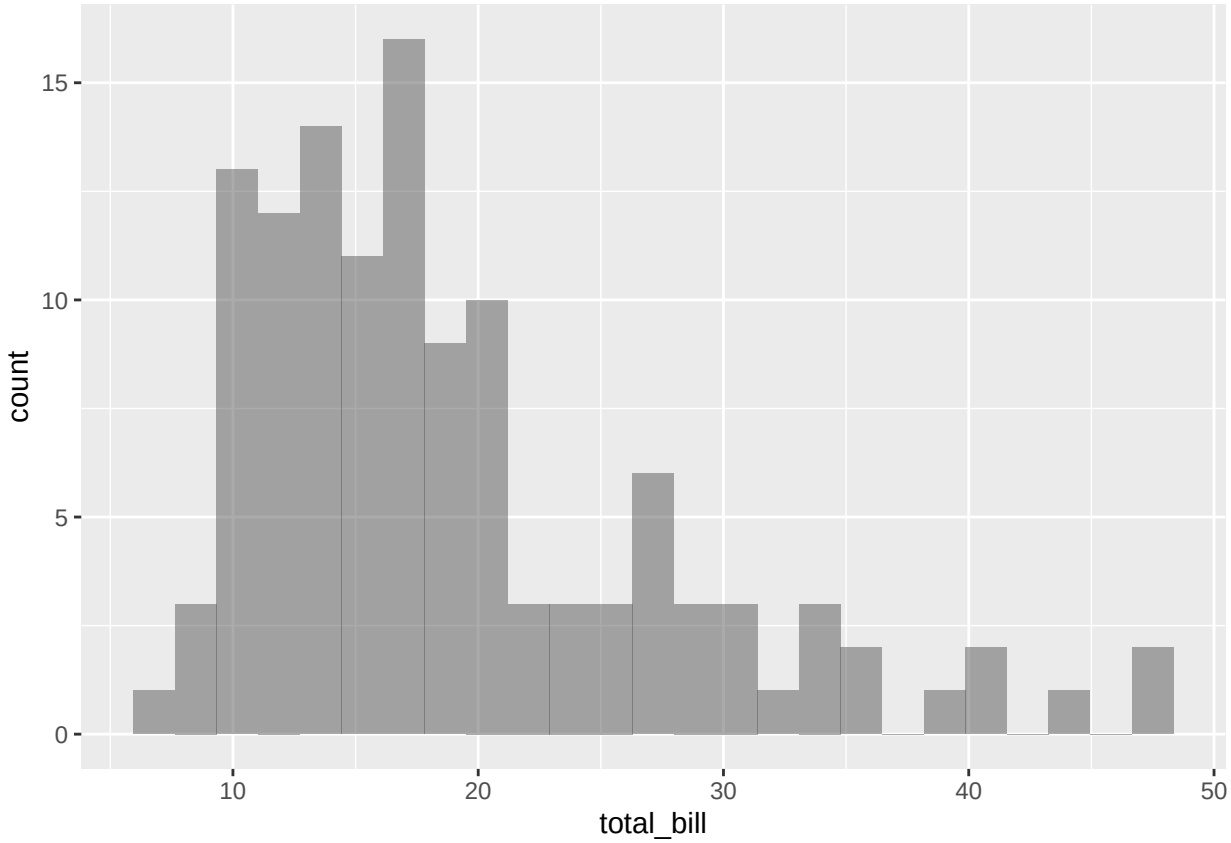
```
## 1    sex character      2 122       0
## 2 smoker character      2 122       0
## 3    day character      4 122       0
## 4   time character      2 122       0
##                                  distribution
## 1 Male (56.6%), Female (43.4%)
## 2 No (62.3%), Yes (37.7%)
## 3 Sat (35.2%), Sun (27%), Thur (25.4%) ...
## 4 Dinner (70.5%), Lunch (29.5%)
##
## quantitative variables:
##            name   class  min      Q1 median      Q3   max      mean        sd
## ...1 total_bill numeric 7.56 12.9575  16.48 22.8025 48.27 19.077131 8.7030565
## ...2        tip numeric 1.01  2.0000   2.68  3.4800  6.73  2.825656 1.1836496
## ...3       size integer 1.00  2.0000   2.00  3.0000  6.00  2.532787 0.9550598
##        n missing
## ...1 122       0
## ...2 122       0
## ...3 122       0
```

Hinweis: Das direkte Einlesen ohne Pfadangabe gelingt nur, wenn die Daten im *aktuellen* Arbeitsverzeichnis (Befehl: `getwd()`) liegen. Ansonsten mit Pfadangabe einlesen. Den genauen Pfad koennen Sie z.B. über `file.choose()` und dann die Datei auswählen ermitteln.

Führen Sie hier eine Explorative Datenanalyse der von Ihnen verwendeten Variablen der Trainingsdaten durch. Auch eine evtl. Datenvorverarbeitung erfolgt in diesem Abschnitt.

Bei Einzelarbeiten sollte der reine Text (ohne Code, Abbildungen etc.) einen Umfang von ca. 1–2 Seiten haben, bei Gruppenarbeiten einen von ca. 2–4 Seiten.

```
gf_histogram( ~ total_bill, data = train)
```



Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

Methodenbeschreibung[2]

Bei Gruppenarbeiten mit weniger als drei Teilnehmer*innen entfällt dieser Abschnitt. Bei Gruppenarbeiten mit drei oder mehr Teilnehmer*innen: gehen Sie hier auf die verwendete Methode zur Modellierung, Variablen, und Modellauswahl ein. Zitieren Sie hier auch die methodische Literatur. Der Abschnitt sollte einen Umfang von 2–4 Seiten haben.

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

[2]Horst Vozeigestudent

Anwendung, Ergebnis und Vorhersage[3]

Wenden Sie hier Ihr Modell an und Interpretieren Sie Ihr Ergebnis. Bei Einzelarbeiten sollte der reine Text (ohne Code, Abbildungen etc.) einen Umfang von ca. 1–2 Seiten haben, bei Gruppenarbeiten einen von ca. 2–4 Seiten.

Modell schätzen:

```
lm.model <- lm(total_bill ~ size, data = train)
summary(lm.model)
```

```
##
## Call:
## lm(formula = total_bill ~ size, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.590  -4.478  -1.335   2.420  22.380
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.6671     1.6755   2.189   0.0306 *
## size          6.0842     0.6193   9.825   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.506 on 120 degrees of freedom
## Multiple R-squared:  0.4458, Adjusted R-squared:  0.4412
## F-statistic: 96.52 on 1 and 120 DF,  p-value: < 2.2e-16
```

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

Modell zur Vorhersage anwenden:

```
lm.predict <- predict(lm.model, newdata = anwendung)
```

Kontrolle und Export:

```
inspect(lm.predict)
```

```
## Warning: `...` is not empty.
##
## We detected these problematic arguments:
## * `needs_dots`
##
## These dots only exist to allow future extensions and should be empty.
## Did you misspecify an argument?
```

[3] Biene Maja

```
## # A tibble: 1 x 10
##   class     min    Q1 median    Q3   max  mean    sd     n missing
## * <chr>   <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <int>   <int>
## 1 numeric  9.75  15.8   15.8  21.9  40.2  19.5  5.78   122       0
```

```
write.csv2(lm.predict, file = "Prognose_IhrName.csv")
```

Zusammenfassung

Fassen Sie gemeinsam kurz die zentralen Ergebnisse zusammen (0,5–1 Seite). Gehen Sie auch auf die Grenzen Ihrer Analyse ein.

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

Literatur

Hier stehen die im Text verwendeten Quellen:

- Nachname Autor1, Anfangsbuchstabe Vorname Autor1, Nachname Autor2, Anfangsbuchstabe Vorname Autor2 1 & Nachname Autor3, Anfangsbuchstabe Vorname Autor3 (Jahr der Veröffentlichung). Titel des Beitrags. Weitere Publikationsinformationen.