# Einführung in die Vorhersage-Modellierung

Matthias Gehrke

2024-03-21

## Inhaltsverzeichnis



## Vorhersage

Neben der erklärenden, rückwärtsgerichteten Modellierung spielt insbesondere in der Praxis die *vorhersageorientierte* Modellierung eine wichtige Rolle: Ziel ist es, bei gegebenen, neuen Beobachtungen die noch unbekannte Zielvariable $y$ *vorherzusagen*, z.B. für neue Kunden auf Basis von soziodemographischen Daten den Kundenwert zu prognostizieren. Dies geschieht auf Basis der vorhandenen Daten der Bestandskunden, d.h. inklusive des für diese Kunden bekannten Kundenwertes (Supervised Learning).

Es werden zwei Teildatenmengen unterschieden: Zum einen gibt es die Trainingsdaten (auch Lerndaten genannt), die aus einer Lern- oder Schätzstichprobe stammen, und zum anderen gibt es Anwendungsdaten, auf die man das Modell anwendet.

1. Bei den Trainingsdaten liegen sowohl die erklärenden Variablen $\mathbf{x} = (x_1, x_2, \dots, x_n)$ als auch die Zielvariable $y$ vor. Auf diesen Trainingsdaten wird das Modell $y = f(\mathbf{x}) + \epsilon = f(x_1, x_2, \dots, x_n) + \epsilon$ gebildet und durch $\widehat{f}(\cdot)$ geschätzt.

2. Dieses geschätzte Modell ($\widehat{f}(\cdot)$) wird auf die Anwendungsdaten $\mathbf{x}_0$, für die (zunächst) die Zielvariable unbekannt ist, angewendet, d.h., es wird $\widehat{y}_0 := \widehat{f}(\mathbf{x}_0)$ berechnet. Der unbekannte Wert $y_0$ der Zielvariable $y$ wird durch $\widehat{y}_0$ prognostiziert.

Eventuell liegt zu einem noch späteren Zeitpunkt der eingetroffene Wert $y_0$ der Zielvariable $y$ vor. Dann kann die eigene Vorhersage $\widehat{y}_0$ evaluiert werden, d.h. z.B. kann der Fehler $y_0 - \widehat{y}_0$ zwischen prognostiziertem Wert $\widehat{y}_0$ und wahrem Wert $y_0$ analysiert werden.

In der praktischen Anwendung können zeitlich drei aufeinanderfolgende Abschnitte unterschieden werden (vergleiche oben):

1. die Trainingsphase, d.h., die Phase für die sowohl erklärende ($\mathbf{x}$) als auch die erklärte Variable ($y$) bekannt sind. Hier wird das Modell geschätzt (gelernt): $\widehat{f}(\mathbf{x})$.
2. In der folgenden Anwendungsphase sind nur die erklärenden Variablen ($\mathbf{x_0}$) bekannt, nicht $y_0$. Auf Basis der Ergebnisses aus 1. wird $\widehat{y}_0 := \widehat{f}(\mathbf{x}_0)$ prognostiziert.
3. Evt. gibt es später noch die Evaluierungsphase, für die dann auch die Zielvariable ($y_0$) bekannt ist, so dass die Vorhersagegüte des Modells überprüft werden kann.

Im Computer kann man dieses Anwendungsszenario *simulieren*: man teilt die Datenmenge *zufällig* in eine Lern- bzw. Trainingsstichprobe (Trainingsdaten; $(\mathbf{x}, y)$) und eine Teststichprobe (Anwendungsdaten, ($\mathbf{x_0}$)) auf: Die Modellierung erfolgt auf den Trainingsdaten. Das Modell wird angewendet auf die Testdaten (Anwendungsdaten). Da man hier aber auch die Zielvariable ($y_0$) kennt, kann damit das Modell evaluiert werden.

## Wettbewerb

Ihre Aufgabe ist: Spielen Sie den Data-Scientist. Konstruieren Sie ein Modell auf Basis der Trainingsdaten $(\mathbf{x}, y)$) und sagen Sie für die Anwendungsdaten (($\mathbf{x_0}$)) die Zielvariable voraus ($\widehat{y}_0$).

Ihr(e) Dozent*in kennt den Wert der Zielvariable ($y_0$). Zur Bewertung der Vorhersagegüte wird der mittlere absolute Fehler MAE (**m**ean **a**bsolute **e**rror) auf die Anwendungsdaten herangezogen:

$$\mathrm{MAE}_{\mathrm{Test}} = \sum_{i=1}^{n_{\mathrm{Test}}} |y_i - \widehat{y}_i|$$

Dabei sind $y_i$ die wahren Werte, $\widehat{y}_i$ die prognostizierten Werte des geschätzten Modells $\widehat{f}(\cdot)$ und $n_{\mathrm{Test}}$ die Anzahl der Beobachtungen des Testdatensatzes (Anwendungsdatensatz). Für eine gute Prognose sollte daher $\mathrm{MAE}_{\mathrm{Test}}$ möglichst klein sein.

## Hinweise

Sie haben relativ freie Methodenwahl bei der Modellierung: Sie können z.B. eine lineare Regression mit Variablen Ihrer Wahl rechnen; Sie können aber auch Baumverfahren oder Neuronale Netze anwenden.

Eine gute Einführung in verschiedene Methoden gibt es z.B. bei Sebastian Sauer (2019): *Moderne Datenanalyse mit R.* https://link.springer.com/book/10.1007/978-3-658-21587-3 aber auch bei Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani (2013): *An Introduction to Statistical Learning – with Applications in R*, http://www-bcf.usc.edu/~gareth/ISL/. Die Bücher beinhalten jeweils Beispiele und Anwendung mit R.

Auch ist es Ihnen überlassen, welche Variablen Sie zur Modellierung heranziehen – und ob Sie diese eventuell vorverarbeiten, d.h., transformieren, zusammenfassen, Ausreißer bereinigen o.ä.. Denken Sie nur daran, die Datentransformation, die Sie auf den Trainingsdaten durchführen, auch auf den Testdaten (Anwendungsdaten) durchzuführen.

Hinweise zur Modellwahl usw. gibt es auch in erwähnter Literatur, aber auch in vielen Büchern zum Thema Data-Mining/Data-Science.

Alles was Sie tun, Datenvorverarbeitung, Modellierung und Anwenden, muss transparent und reproduzierbar sein. Ansonsten lautet die Aufgabe: Finden Sie ein Modell, von dem Sie glauben, das es gut vorhersagt. $\hat{y} = 42$ tut es leider oft nicht. Eine gute Modellierung auf den Trainingsdaten (z.B. hohes $R^2$) bedeutet nicht zwangsläufig eine gute Vorhersage.

### Tipps für eine gute Prognose

- Vermeiden Sie Über-Anpassung.
- Evtl. kann eine Datenvorverarbeitung (Variablentransformation, z.B. log() oder die Elimination von Ausreißern) helfen.
- Überlegen Sie sich Kriterien zur Modell- und/oder Variablenauswahl.
- Schauen Sie in die Literatur.

## Bewertung

Gruppenarbeiten mit bis zu vier Personen sind möglich. In die Bewertung fließen u.a. ein:

- Methode: methodischer Anspruch und Korrektheit in der Explorativen Datenanalyse, Datenvorverarbeitung, Variablenauswahl und Modellierungsmethode
- Inhalt: inhaltliche Korrektheit in Beschreibung und Interpretation
- Vorhersagegüte: Die Vorhersagegüte des Nullmodells entspricht einer $4,0$, die eines (unbekannten) einfachen Referenzmodells Ihr(e)r Dozent*in einer $2,0$. Ihre Bewertung erfolgt entsprechend Ihrer Vorhersagegüte, d.h., sind Sie besser als das Referenzmodell erhalten Sie hier eine bessere Note als $2,0$!

## Organisatorisches

Der Schwerpunkt dieser Arbeit liegt auf der quantitativen Modellierung, der formale Anspruch, aber auch der Anspruch in Bezug auf Literatur etc. liegen daher unter dem von anderen schriftlichen Prüfungsleistungen. Um eine komplett transparente und reproduzierbare Analyse zu ermöglichen, muss das beigefügte qmd-Template verwendet werden (`Template-Vorhersagemodellierung.qmd`).

Dies wird dann direkt in eine pdf-Datei überführt (`render`) und im OC hochgeladen. Ein ausgedrucktes Exemplar muss nicht abgegeben werden.

Als Zusatzmaterial laden Sie die qmd-Datei und die csv-Datei mit den Daten Ihrer Prognose (`Prognose_IhrName.csv`) hoch.

## Checkliste

- Haben Sie eine Vorhersage für die 200 Anwendungsdaten erzeugt und als csv Datei exportiert: `Prognose_IhrName.csv` (Ihr Name entsprechend angepasst)?
- Entspricht diese in der Struktur dem Beispiel `Vorhersage_Zufall.csv`?
- Bei Gruppenarbeiten: Sind die individuellen Kapitelzuordnungen erkennbar?
- Läuft die qmd-Datei beim knitten durch?
- Haben Sie die pdf-Datei Ihrer Auswertung und als Zusatzmaterial die qmd- und die csv-Datei hochgeladen?

## Datenbeschreibung

Es liegen Daten einer Fahrradvermietung vor, Zielvariable $y$ ist die Anzahl der (täglichen) Vermietungen (`vermietungen`).

Als (potentiell) erklärende Variablen liegen folgende Daten des Tages vor:

- `einfuehrungsphase`: handelt es sich um die Einführungsphase, d.h., ist das Angebot neu?
- `jahreszeit`: Jahreszeit
- `wetter`: Wetterbeschreibung
- `arbeitstag`: Handelt es sich um einen Werktag oder um Wochenende bzw. Feier- oder Ferientag?
- `temperatur`: Temperatur in °C
- `windgeschwindigkeit`: Windgeschwindigkeit in km/h
- `luftfeuchtigkeit`: Luftfeuchtigkeit in %

Der Datensatz `train.csv` enthält die Zielvariable (`vermietungen`), anhand dieser Daten können Sie Ihr Modell entwickeln, angewendet wird es auf den Datensatz `anwendung.csv`. Dieser enthält die Zielvariable nicht. Die Aufteilung erfolgte zufällig. Erstellen Sie auf Basis der Beobachtungen `train.csv` ein Modell für die Anzahl Vermietungen, `vermietungen`. Wenden Sie Ihr Modell auf die Beobachtungen aus `anwendung.csv` an und erstellen Sie so für diese Beobachtungen eine Prognose für die Anzahl Vermietungen.

Exportieren Sie Ihre Prognose ebenfalls als `csv` Datei (`Vorhersage_IhrName.csv`, vergleiche `Vorhersage_Zufall.csv`, siehe Template `Template-Vorhersagemodellierung.Rmd`).