

# Hinweise zur Prüfungsleistung ‘Vorhersagemodellierung’

Sebastian Sauer

2021-04-29

## Inhaltsverzeichnis

<b>1</b>	<b>tl;dr: Zusammenfassung</b>	<b>1</b>
<b>2</b>	<b>Vorhersage</b>	<b>2</b>
<b>3</b>	<b>Koeffizient zur Güte der Vorhersagegüte</b>	<b>3</b>
<b>4</b>	<b>Zum Aufbau Ihrer Prognosedatei im CSV-Format</b>	<b>3</b>
<b>5</b>	<b>Einzureichende Dateien</b>	<b>3</b>
<b>6</b>	<b>Tipps</b>	<b>3</b>
6.1	Tipps für eine gute Prognose . . . . .	3
6.2	Tipps zur Datenverarbeitung . . . . .	4
6.3	Tipps zum Aufbau des Analyseskripts . . . . .	4
<b>7</b>	<b>Bewertung</b>	<b>4</b>
<b>8</b>	<b>Hinweise</b>	<b>5</b>
<b>9</b>	<b>Formalia</b>	<b>6</b>
<b>10</b>	<b>Wo finde ich Beispiele?</b>	<b>7</b>
<b>11</b>	<b>Plagiatskontrolle</b>	<b>7</b>

## 1 tl;dr: Zusammenfassung

Vorhersagen sind eine praktische Sache, zumindest wenn Sie stimmen. Wenn Sie den DAX-Stand von morgen genau vorhersagen können, rufen Sie mich bitte sofort an. Genau das ist Ihre Aufgabe in dieser Prüfungsleistung: Sie sollen Werte vorhersagen.

Etwas konkreter: Stellen Sie sich ein paar Studentis vor; von allen wissen Sie, wie lange die Person für die Statistiklausur gelernt hat. Außerdem wissen Sie die Motivation jeder Person und vielleicht noch ein paar noten-relevante Infos. Und Sie wissen die Note jeder Person in der Statistiklausur. Auf dieser Basis fragt sie ein Student (Alois), der im kommenden Semester die Prüfung in Statistik schreiben ~~muss~~ will: “Sag mal, wenn ich 100 Stunden lerne und so mittel motiviert bin, welche Note kann ich dann erwarten?”. Mit Hilfe Ihrer Analyse können Sie diese Frage beantworten. Natürlich könnten Sie es sich leicht machen

und antworten: “Mei, der Notendurchschnitt war beim letzten Mal 2.7. Also ist das kein ganz doofer Tipp für deine Note.” Ja, das keine doofe Antwort, aber man kann es besser machen. Da hilft Ihnen die Statistik (doch, wirklich).

Kurz gesagt gehen Sie so vor: Importieren Sie die Daten in R, starten Sie die nötigen R-Pakete und schauen Sie sich die Daten unter verschiedenen Blickwinkeln an. Dann nehmen Sie die vielversprechendsten Prädiktoren in ein Regressionsmodell und schauen sich an, wie gut die Vorhersage ist. Wiederholen Sie das ein paar Mal, bis Sie ein Modell haben, das Sie brauchbar finden. Mit diesem Modell sagen Sie dann die Noten der neuen Studis (Alois und Co.) vorher. Je genauer Ihre Vorhersage, desto besser ist Ihr Prüfungsergebnis.

## 2 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 unbekannten Werte der Zielvariablen  $y$  *vorherzusagen*, z.B. für neue Kunden auf Basis von soziodemographischen Daten den *Kundenwert* – möglichst genau – zu prognostizieren. Dies geschieht auf Basis der vorhandenen Daten der Bestandskunden, d.h. inklusive des für diese Kunden bekannten Kundenwertes.

Ihnen werden *zwei Teildatenmengen* zur Verfügung gestellt: Zum einen gibt es die Trainingsdaten (auch *Lerndaten* genannt) und zum anderen gibt es Anwendungsdaten (auch *Testdaten* genannt), 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  $\hat{f}(\cdot)$  geschätzt.
2. Dieses geschätzte Modell ( $\hat{f}(\cdot)$ ) wird auf die Anwendungsdaten  $\mathbf{x}_0$ , für die (Ihnen) die Werte der Zielvariable unbekannt sind, angewendet, d.h., es wird  $\hat{y}_0 := \hat{f}(\mathbf{x}_0)$  berechnet. Der unbekannte Wert  $y_0$  der Zielvariable  $y$  wird durch  $\hat{y}_0$  prognostiziert.

Liegt zu einem noch späteren Zeitpunkt der eingetroffene Wert  $y_0$  der Zielvariable  $y$  vor, so kann die eigene Vorhersage  $\hat{y}_0$  evaluiert werden, d.h. z.B. kann der Fehler  $e = y_0 - \hat{y}_0$  zwischen prognostiziertem Wert  $\hat{y}_0$  und wahren Wert  $y_0$  analysiert werden.

In der praktischen Anwendung können zeitlich drei aufeinanderfolgende Schritte 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):  $\hat{f}(\mathbf{x})$ . Dafür wird der Trainingsdatensatz genutzt.
2. In der folgenden *Anwendungsphase* sind nur die erklärenden Variablen ( $\mathbf{x}_0$ ) bekannt, nicht  $y_0$ . Auf Basis der Ergebnisse aus dem 1. Schritt wird  $\hat{y}_0 := \hat{f}(\mathbf{x}_0)$  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;  $(\mathbf{x}, \mathbf{y})$ ) 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.

### 3 Koeffizient zur Güte der Vorhersagegüte

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

Ihr(e) Dozent\*in kennt den Wert der Zielvariable ( $y_0$ ).

Zur Bewertung der Vorhersagegüte wird  $R^2$  herangezogen.

Von zwei Prognosemodellen zum gleichen Datensatz ist dasjenige Modell besser, das einen höheren  $R^2$ -Wert aufweist. Kurz gesagt: Höher ist besser.

### 4 Zum Aufbau Ihrer Prognosedatei im CSV-Format

1. Die CSV-Datei muss aus zwei Spalten mit (exakt) folgenden Spaltennamen bestehen:
  - a) **id**: Den ID-Wert jedes vorhergesagten Wertes
  - b) **pred**: Der vorhergesagte Wert.
2. Der Name der Datei muss wie folgt lauten: `Nachname_Vorname_Matrikelnummer_Prognose.csv`.  
Beispiel: `Sauer_Sebastian_0123456_Prognose.csv`.
3. Umlaute sind zu ersetzen (also `Süß` wird `Suess` etc.).
4. Die CSV-Datei muss als *Spaltentrennzeichen* ein *Komma* verwenden und als *Dezimaltrennzeichen* einen *Punkt* (d.h. also die *Standardformatierung* einer CSV-Datei; *nicht* die deutsche Formatierung).
5. Die CSV-Datei muss genau die Anzahl an Zeilen aufweisen, die der Zeilenlänge im Test-Datensatz entspricht.
6. Prüfen Sie, dass Ihre CSV-Datei sich problemlos lesen lässt. Falls keine (funktionstüchtige) CSV-Datei eingereicht (hochgeladen) wurde, ist die Prüfung nicht bestanden. Tipp: Öffnen Sie die CSV-Datei mit einem Texteditor und schauen Sie sich an, ob alles vernünftig aussieht. Achtung: Öffnen Sie die CSV-Datei besser nicht mit Excel, da Excel einen Bug hat, der CSV-Dateien verfälschen kann auch ohne dass man die Datei speichert.

### 5 Einzureichende Dateien

1. Folgende Dateien sind einzureichen:
  1. Ihre *Prognose-Datei* (CSV-Datei)
  2. Ihr *Analyseskript* (R-, Rmd- oder Rmd-Notebook-Datei)
2. Weitere Dateien sind nicht einzureichen.
3. Komprimieren Sie die Dateien *nicht* (z.B. via *zip*).

### 6 Tipps

#### 6.1 Tipps für eine gute Prognose

- Schauen Sie in die Literatur.
- Evtl. kann eine Datenvorverarbeitung (Variablentransformation, z.B. `log()` oder die Elimination von Ausreißern) helfen.

- Überlegen Sie sich Kriterien zur Modell- und/ oder Variablenauswahl. Auch hierfür gibt es Algorithmen und R-Funktionen.
- Vermeiden Sie Über-Anpassung (Overfitting).
- Vermeiden Sie viele fehlende Werte bei Ihrer Prognose. Fehlende Werte werden bei der Benotung mit dem Mittelwert (der vorhandenen Prognosewerte Ihrer Einreichung) aufgefüllt.

## 6.2 Tipps zur Datenverarbeitung

- Ein “deutsches” Excel kann Standard-CSV-Dateien nicht ohne Weiteres lesen. Online-Dienste wie Google Sheets können dies allerdings.

## 6.3 Tipps zum Aufbau des Analyseskripts

- Zu Beginn des Skripts sollten alle verwendeten R-Pakete mittels `library()` gestartet werden.
- Zu Beginn des Skripts sollten die Daten von der vom Dozenten bereitgestellten URL importiert werden (*nicht* von der eigenen Festplatte, da das Skript sonst bei Dritten, wie Ihrem Prüfer, nicht lauffähig ist).

# 7 Bewertung

- Es gibt drei Bewertungskriterien:
  - *Formalia*: u.a. Reproduzierbarkeit der Analyse, Lesbarkeit der Syntax, Übersichtlichkeit der Analyse.
  - *Methode*: u.a. methodischer Anspruch und Korrektheit in der Explorativen Datenanalyse, Datenvorverarbeitung, Variablenauswahl und Modellierungsmethode.
  - *Inhalt*: **Vorhersagegüte**.
- Das zentrale Bewertungskriterium ist *Inhalt*; die übrigen beiden Kriterien fließen nur bei besonders guter oder schlechter Leistung in die Gesamtnote ein.
- Zur Vorhersagegüte: Die Vorhersagegüte des Nullmodells entspricht einer 4,0, die eines (unbekannten) einfachen Referenzmodells Is Dozentis einer 2,0<sup>1</sup>. Ihre Bewertung erfolgt entsprechend Ihrer Vorhersagegüte, d.h., sind Sie besser als das Referenzmodell erhalten Sie hier in diesem Teilaspekt eine bessere Note als 2,0!
- Die quantitative Datenanalyse in Durchführung und Interpretation ist der Schwerpunkt dieser Arbeit. Zufälliges identisches Vorgehen, z.B. im R Code, ist sehr unwahrscheinlich und kann als **Plagiat** bewertet werden.
- Die Gesamtnote muss sich nicht als arithmetischer Mittelwert der Teilnoten ergeben.
- Es werden keine Teilnoten vergeben, sondern nur eine Gesamtnote.

---

<sup>1</sup>Gender-I: <https://gender-i.de/>

## 8 Hinweise

Sie haben freie Methodenwahl bei der Modellierung und Vorverarbeitung. Nutzen Sie den Stoff wie im Unterricht gelernt; Sie können aber auch auf weitere Inhalte, die nicht im Unterricht behandelt wurden, zugreifen.

Eine Einführung in verschiedene Methoden gibt es z.B. bei Sebastian Sauer (2019): *Moderne Datenanalyse mit  $R^2$*  aber auch bei Max Kuhn und Julia Silge (2021): *Tidy Modeling with R*.<sup>3</sup> 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-Science.

**Alles, was Sie tun, Datenvorverarbeitung, Modellierung und Anwenden, muss transparent und reproduzierbar sein.** Im Übrigen lautet die Aufgabe: Finden Sie ein Modell, von dem Sie glauben, dass es die Testdaten 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.

---

<sup>2</sup><https://link.springer.com/book/10.1007/978-3-658-21587-3>

<sup>3</sup><https://www.tnwr.org/>

## 9 Formalia

1. Es sind nur Einzelarbeiten zulässig.
2. In der Analyse muss als Ausgangspunkt der vom/von der Dozenten/in bereitgestellten Datensatz genutzt werden. Dazu muss der Datensatz zu Beginn des Skripts von der entsprechenden Internetadresse heruntergeladen werden.
3. Alle Analyseschritte bzw. alle Veränderungen an den Daten müssen im (eingereichten) *Analyseskript* nachvollziehbar (transparent und reproduzierbar) aufgeführt sein. Das Analyseskript ist als R-Skript, Rmd-Datei oder Rmd-Notebook-Datei abzugeben. Sie können die bereitgestellte Vorlage als Analyseskript nutzen (**Template-Dokumentation-Vorhersagemodellierung.Rmd**).
4. Das Analyseskript muss funktionstüchtig für den Prüfer sein: Alle Befehle müssen ohne Fehlermeldung durchlaufen (abgesehen von etwaiger Installation fehlender Pakete).
5. Es dürfen keine weiteren Informationen (Daten) als die vom Dozenten ausgegebenen verwendet werden. Sonstige Hilfe (z.B. von Dritten) ist ebenfalls unzulässig.
6. Nichtbeachtung der für dieses Modul formulierten Regeln kann zu Nichtbestehen oder Punkteabzug führen.
7. Der Schwerpunkt dieser Hausarbeit liegt auf der quantitativen Modellierung, der formale Anspruch liegt daher unter dem von anderen Hausarbeiten.
8. Es muss keine Literatur zitiert werden.
9. Ein ausgedrucktes Exemplar muss nicht abgegeben werden.

## 10 Wo finde ich Beispiele?

Eine Beispiel-Modellierung finden Sie in der Datei `Beispielanalyse-Prognose-Wettbewerb.Rmd`. Eine beispielhafte Vorlage (Template), die Sie als Richtschnur nutzen können, ist mit der Datei `Template-Vorhersagemodellierung.Rmd` bereitgestellt. Im Internet finden sich viele Fallstudien, von denen Sie sich inspirieren lassen können.

## 11 Plagiatskontrolle

Die eingereichten Arbeiten werden automatisiert auf Plagiate überprüft. Gibt es substanzielle Überschneidungen zwischen zwei (oder mehr) Arbeiten, werden alle diese Arbeiten mit *ungenügend* bewertet.