Data Science Challenge: Water Level Prediction

Enrico Lauckner, 22.08.2016

Workflow

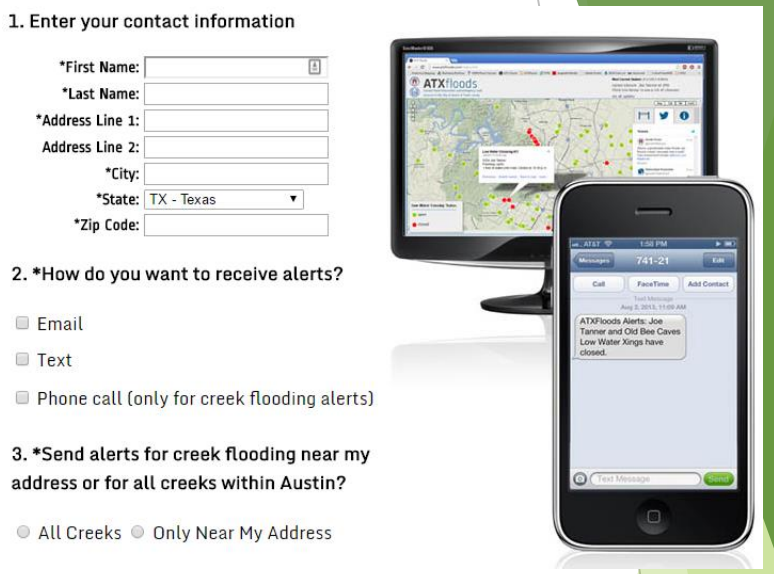
- Verstehen der Daten
- Gauss-Krüger-Koordinaten der Pegelstationen umwandeln in Längen- und Breitengrad (rgdal)
- Finden der nähesten Wetterstation pro Pegelstation (searchTree)
- Feature Engineering
- Modelle und Evaluation
- Shiny App

Erster Versuch mit Dresden gescheitert: Wetter am Ort selbst gar nicht so entscheidend

-> Tschechisches Gebirge und Nebenflüsse

Use Case - Landeshauptstadt Mainz

- Entwicklung eines Frühwarnsystem für Mainz mit Fokus auf **3 Tage Vorhersage** (+ Validierung einer 5 Tage Vorhersage)
- -> Der Pegel Mainz ist einer der wichtigen Messpegel für Rheinschifffahrt, Anwohner und flussabwärts gelegene Ufergebiete am Rhein [1]
- Vorteile:
  - (Automatisierte) Warnungen an Bewohner und Unternehmen
  - Sperren von Straßen ([3] Turn around, don't drown)
  - Rechtzeitige Beschaffung von Ressourcen (Rettungskräfte, Freiwillige Helfer, Sandsäcke)
  - Routenplanung für Frachtschiffe (uneingeschränkte Schifffahrt nur bei Pegelständen zwischen 2,60 m und 4,75 m möglich)



[2] Flut Frühwarnung – Beispiel Austin, Texas

--> Vermeiden von humanitären Katastrophen und wirtschaftlichen Schäden!

[1] https://de.wikipedia.org/wiki/Pegel_Mainz
[2] http://www.austintexas.gov/department/flood-early-warning-system
[3] https://www.youtube.com/watch?v=PJ-iZGtdNCs

Use Case - Landeshauptstadt Mainz

SWR FERNSEHEN

Landesschau AKTUELL RHEINLAND-PFALZ

STARTSEITE RHEINLAND-PFALZ KAISERSLAUTERN KOBLENZ MAINZ LUDWIGSH

Hochwasser: Runder Tisch in Mainz

[1] Stand: 15.6.2016, 8.13 Uhr

Bürgermeister fordert Frühwarnsystem

Viele rheinhessische Gemeinden sind noch immer gebeutelt von den Folgen des Unwetters vor zwei Wochen und hoffen auf Hilfe vom Land. Aber es geht ihnen auch um Erfahrungsaustausch.

Hochstätten (Kreis Bad Kreuznach) wurde besonders gebeutelt von den Fluten. (Archiv)

Umweltministerin Höfken (GRÜNE) hatte am Dienstag Landräte und Bürgermeister zu einem Runden Tisch geladen. Auch Vertreter aus Rheinhessen waren dabei, unter anderem der Bürgermeister der Verbandsgemeinde Alzey-Land, Steffen Unger. In seinem Verantwortungsbereich liegen die beiden Gemeinden Nieder-Wiesen und Flonheim.

Sie wurden bei dem heftigen Starkregen Ende Mai geradezu überflutet. Ähnlich wie in Hochstätten der Leischbach, war es hier der kleine Wiesbach, der plötzlich zum reißenden Fluss wurde.

21. Juni 2016 | 08.50 Uhr

Mainz [2]

Rhein-Schifffahrt wegen Hochwasser eingeschränkt

Teilen Twittern

Mainz. Hohe Wasserstände behindern weiterhin die Schifffahrt auf dem Rhein. Von Mannheim-Rheinau bis Köln dürfen Schiffe nur langsam und in der Mitte des Flusses fahren, wie das Hochwassermeldezentrum (HMZ) in Mainz mitteilte. Am Oberrhein kommen zahlreiche Schiffe gar nicht weiter: Der Abschnitt zwischen

--> Vermeiden von humanitären Katastrophen und wirtschaftlichen Schäden!

[1] http://www.swr.de/landesschau-aktuell/rp/mainz/hochwasser-runder-tisch-in-mainz-buergermeister-fordert-fruehwarnsystem/-/id=1662/did=17600272/nid=1662/xxqun6/
[2] http://www.rp-online.de/panorama/rhein-schifffahrt-wegen-hochwasser-eingeschraenkt-aid-1.6063303

Umsetzung in R und Shiny

Programmiersprache für statistische Berechnungen und Grafiken

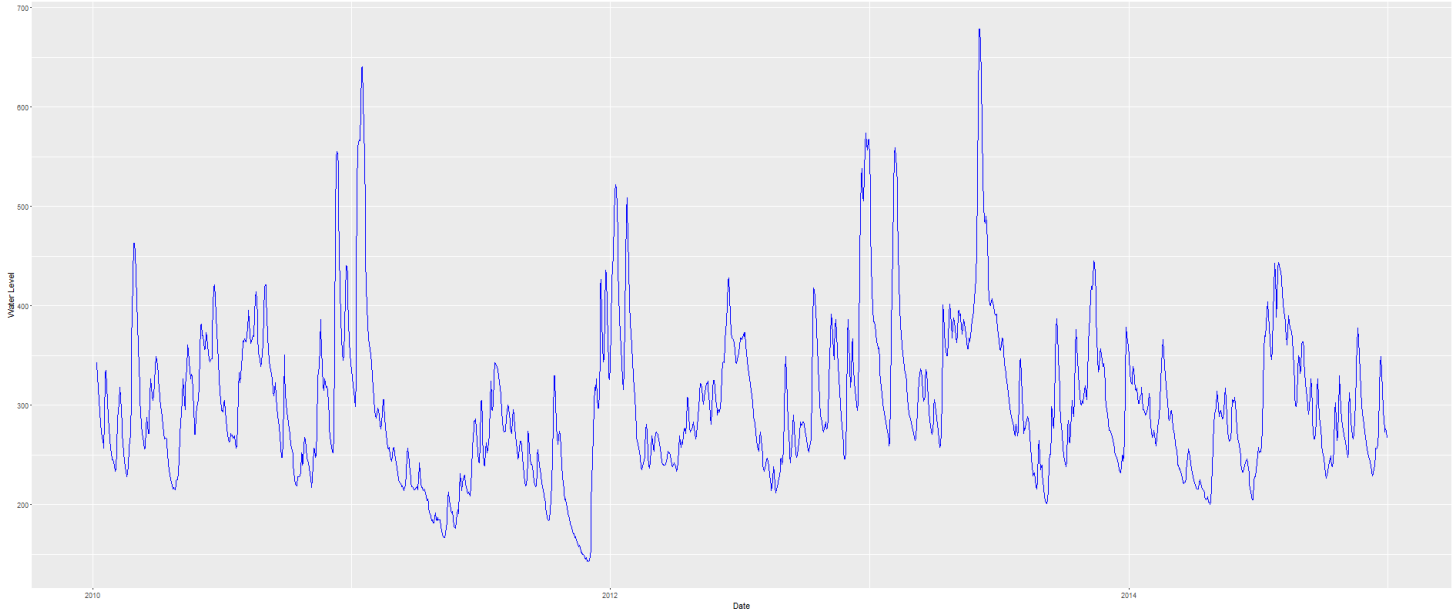
- Umwandlung von Geokoordinaten von Gauss-Krüger in Längen- und Breitengrade
- Vorbereitung des Model Input
- Erstellung von Modellen
- Evaluierung von Modellen
- Treffen und Evaluierung von Vorhersagen

Web Application Framework für R

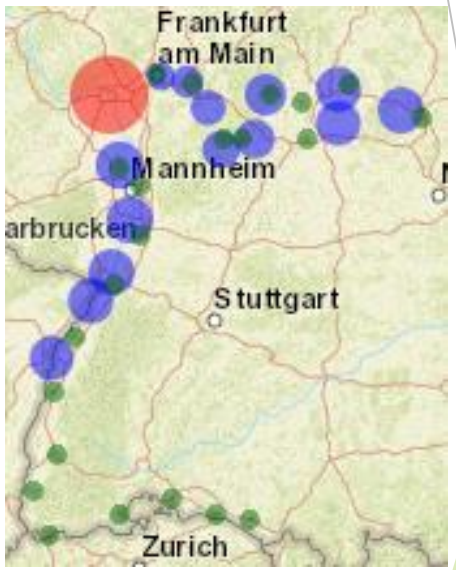
- Kartenvisualisierung zur Identifizierung von Pegel- und Wetterstationen
- Interaktive Weboberfläche zur Ergebnisdokumentation
- Bereitstellung im Web (siehe Link unten)

Zielvariable - Pegelstand Mainz

- Main und Rhein fließen kurz vor Mainz zusammen
- Viele vorgelagerte Pegel- und Wetterstationen in Deutschland



Tatsächlicher Verlauf des Mainzer Pegels Anfang 2010 bis Ende 2014 (R Darstellung)



Rot: Pegelstation Mainz
Blau: Pegelstationen
Grün: Wetterstationen
(Shiny Darstellung)

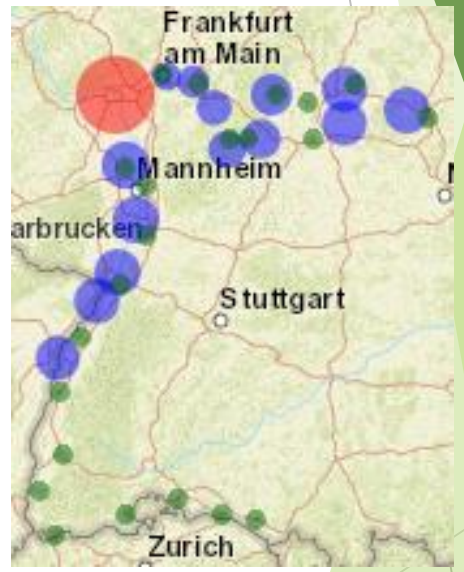
--> Modell finden was den Mainzer Pegelstand jeweils 3 Tage im Vorraus vorhersagt

Model Input – Pegelstationen (15 Variablen)

- Das Wissen über Pegelstände von flußaufwärts gelegenen Stationen nutzen

- Verwendete Pegelstationen Main (Flußabwärts geordnet, 9 Variablen):
  - Trunstadt
  - Schweinfurt neuer hafen
  - Astheim
  - Steinbach
  - Wertheim
  - Kleinheubach
  - Obernau
  - Krotzenburg
  - Frankfurt Osthafen

- Verwendete Pegelstationen Rhein (Flußabwärts geordnet, 6 Variablen):
  - Kehl-Kronenhof
  - Plittersdorf
  - Maxau
  - Speyer
  - Worms
  - Mainz -> Eigener Pegel als Vorhersage nutzbar?

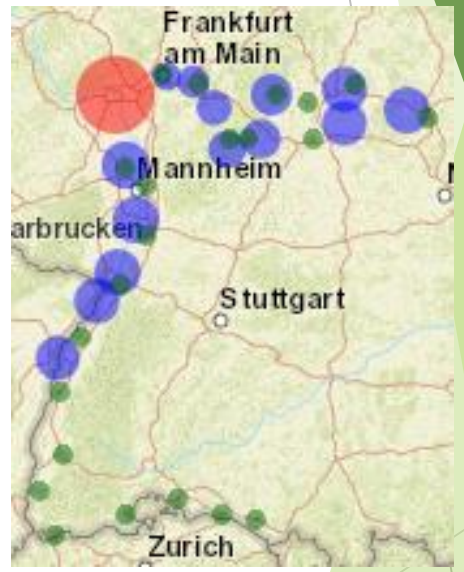


Rot: Pegelstation Mainz
Blau: Pegelstationen
Grün: Wetterstationen
(Shiny Darstellung)

→ Können Pegelstände von heute helfen Vorhersagen in 3 Tagen zu treffen?

Model Input – Wetterstationen (20 Variablen)

- Das Wissen über Wetter flußaufwärts nutzen
- 22 Wetterstationen entlang Main und Rhein
- Durchschnittsbildung über alle Stationen pro Variable und Fluß
- Einige Variablen wegen Korrelationen und Datenmangel (SCHNEEHÖHE) entfernt
- Verwendete Wettervariablen (10, jeweils für Main und Rhein):
  - LUFTTEMPERATUR
  - DAMPFDRUCK
  - BEDECKUNGSGRAD
  - LUFTDRUCK_STATIONSHOEHE
  - REL_FEUCHTE
  - WINDGESCHWINDIGKEIT
  - WINDSPITZE_MAXIMUM
  - NIEDERSCHLAGSHOEHE
  - NIEDERSCHLAGSHOEHE_IND
  - SONNENSCHEINDAUER

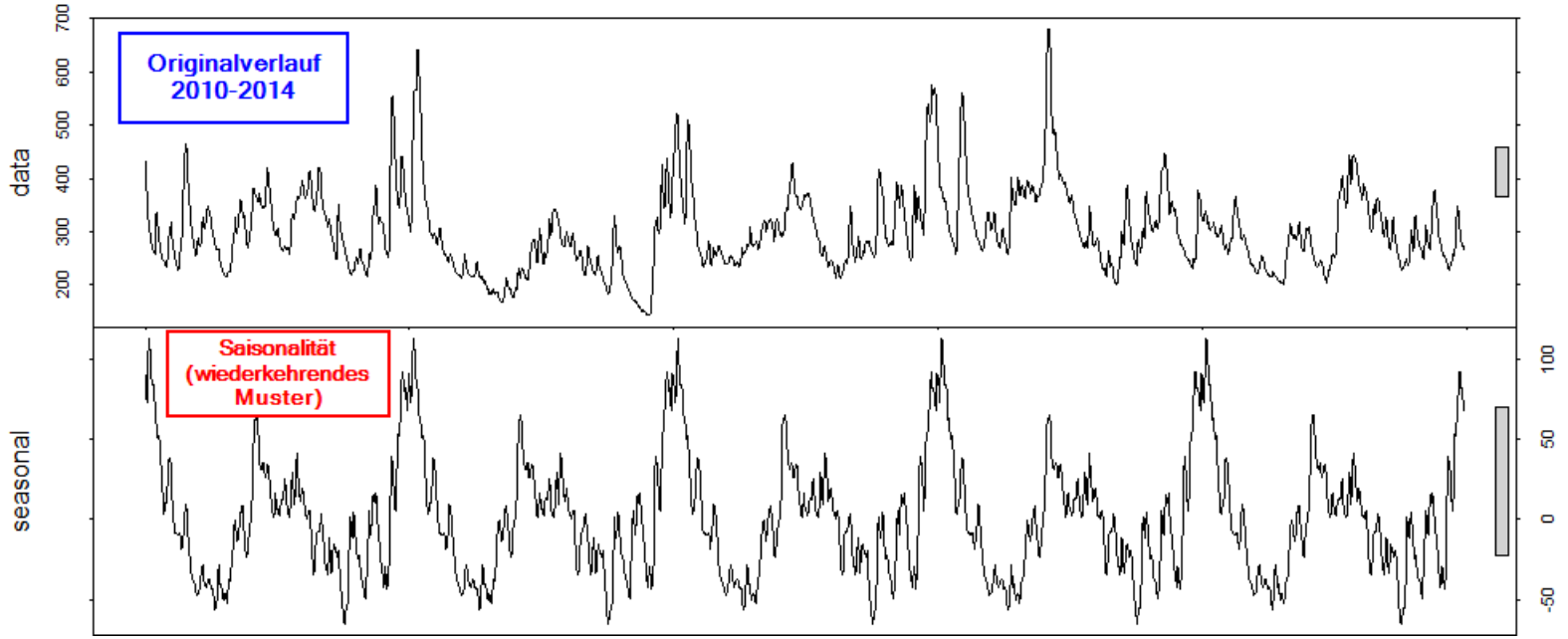


Rot: Pegelstation Mainz
Blau: Pegelstationen
Grün: Wetterstationen
(Shiny Darstellung)

→ Kann Wetter von heute helfen Vorhersagen in 3 Tagen zu treffen?

Model Input – Saisonalität (1 Variable)

Das Wissen über saisonale Effekte des Pegelstands nutzen
-> Schneeschmelzen im Frühjahr, Sommerregen, Vereisung im Winter



Vergleich tatsächlichem Verlaufs mit erkanntem saisonalem Muster (R Darstellung)

→ Kann die Saisonalität aus 5 Jahren Pegelbeobachtung helfen eine Vorhersage zu treffen?

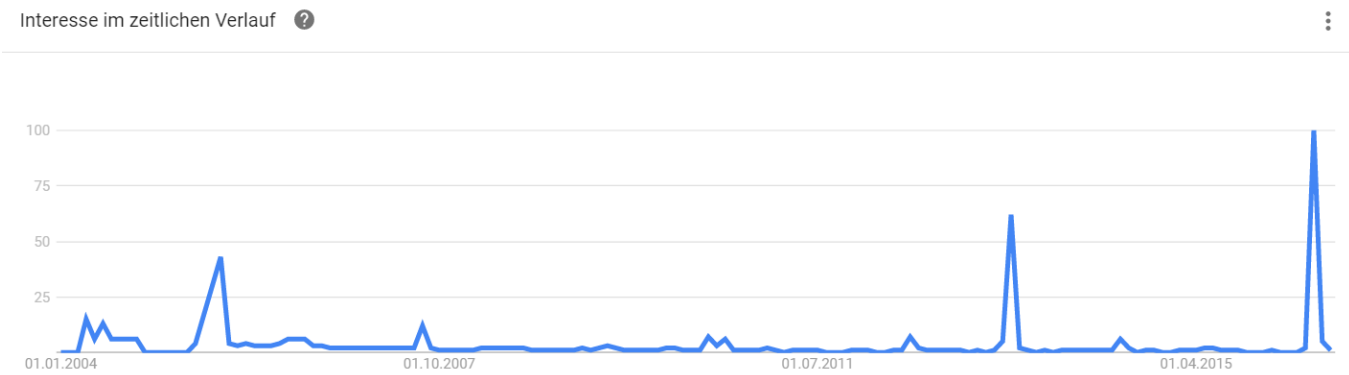
Model Input - Google Trends (1 Variable)

Das Wissen über Google-Suchanfragen zu regionalem Hochwasser nutzen

-> Google Trends

Suchterme:

- "Hochwasser Rhein“
- "Hochwasser Main“
- "Hochwasser Frankfurt Main“
- "Hochwasser Worms“
- "Hochwasser Bodensee"



Beispiel Google Trends: Suchanfrage „Hochwasser Bodensee“ 2004-heute

Historisiert (Zeitraum 2010-2014) liegen leider nur Daten auf Wochenebene vor, für aktuelle Vorhersagen jedoch deutlich granularer nutzbar. Zahl der Suchanfragen pro Woche werden jedem einzelnen Tag der Woche zugeordnet und um 4 Tage verschoben (Modell kennt z.B. ab Donnerstag gesamtes Interesse der nächsten Woche bereits)

-> Versuch tägliche Suchanfragen halbwegs zu simulieren

→ Können regionale Aktivitäten von Google-Suchanfragen zu Hochwasser helfen Vorhersagen zu treffen?

Modellierung

- Trainingsdaten
- 2010-2012, 2014 für Vorhersage von **2013** (Testdaten) -> viele Hochwassertage
- 2010-2013 für Vorhersage von **2014** (Testdaten) -> keine Hochwassertage

- R package “caret” für Modellierung genutzt
- train() Funktion für Preprocessing, Cross-Validation und automatische Evaluierung über Tuningparameter der jeweiligen Modelle
- Vielzahl von verfügbaren Modellen aus anderen Paketen [1]
- Verwendete Modelle
  - Single Decision Tree (ctree)
  - Partial Least Squares (pls)
  - K-Nearest Neighbor (knn)
  - Support Vector Machine (svmPoly)
  - 2 x Boosted Tree (ranger, cubist)
  - 2 x Extreme Gradient Boosting (xgbTree, xgbLinear)
  - Ensemble aus besten Modellen (Bündeln von Stärken der einzelnen Modelle [2])

```
### Single Tree
  MActree <- train(x = MTrain3[c(4:18)], y = MTrain3[,2],
                   method = "ctree",
                   trControl = ctrl,
                   preProc = c("center", "scale", "BoxCox"))
```

Beispiel: Erstellung eines Decision Tree Models mit carets train()

[1] http://topepo.github.io/caret/bytag.html
[2] https://en.wikipedia.org/wiki/Ensemble_learning

Finden des besten Modells

Finales Modell

- Modelle mit sämtlichen Input Variablen liefern beste Ergebnisse auf Trainingsdaten
- Allerdings insgesamt schlechtere Ergebnisse mit Google-Hits-Input auf Testdaten

  -> finales Modell ohne diesen Input (auch wegen wöchentlicher Ebene)
- Cubist und Support Vector Machine (SVM) liefern insgesamt beste Ergebnisse

  -> im Ensemble noch präziser

Wichtigste Input-Variablen (Durchschnitt über SVM und Cubist mit allen Inputs)

1. NIEDERSCHLAG Rhein
2. Pegel Worms
3. Pegel Mainz
4. Pegel Kehl-Kronenhof
5. Lufttemperatur Main

…

9. Saisonalität

..

11. Google Hits

..

Schlusslicht: SONNENSCHEINDAUER Rhein

→ Finales Modell: Ensemble aus Cubist und SVM mit Inputs aus Pegelständen, Wetter und Saisonalität

Vorhersage mit bestem Modell 2013

Vergleich tatsächlichem Verlaufs (blau) mit Vorhersage (rot) (R Darstellung)

→ Kurve der Vorhersage sehr dicht an tatsächlichem Stand, Probleme bei Spitzen

Vorhersage mit bestem Modell 2014

Vergleich tatsächlichem Verlaufs (blau) mit Vorhersage (rot) (R Darstellung)

→
Weitere Ergebnisse
in Shiny-App
(2014, Kategorien)

→ Vorhersage Ergebnisse noch einmal besser als 2013, starker Ausreißer am Ende des Jahres

Weitere Ideen

- Komplexere Wettersimulationen -> z.B. Schneeschmelzen durch Mittelmeertiefs
- Wetterstationen
  - Wettervorhersagen (hier nur das aktuelle Wetter einbezogen)
  - Nicht zusammenfassen, einzeln als Input verwenden?
- Pegelstationen zu unterschiedlichen Tagen als Input verwenden, je nach Ankunft durch Fließgeschwindigkeit (hier alle an einem Tag betrachtet)
- Mehr/Andere Vorhersage-Modelle
  - Alleine in caret über 200 verfügbar
  - Neuronale Netze bei Kurzvalidierung sehr schlecht, woran lags?
- Modelle besser tunen (hier nur Autotuning benutzt)
- Saisonalität über mehrere Jahre (hier nur 5 Jahre)
- Tägliche (weitere) Social Media Daten -> z.B. Twitter, Google Trends (hier nur wöchentlich verfügbar)
- Dashboard mit Vorhersagen der nächsten Tage