

ML in der IBNER Reservierung

Ein Fallbeispiel aus Kaggle

Dr. Nelvis Fornasin, Attila Gulyas
B&W Deloitte



DAA

DEUTSCHE
AKTUAR-AKADEMIE GmbH

Aktuelle ADS-Anwendungen... IM FOKUS, 16.11.2021

"Actuarial Loss Prediction" Wettbewerb: Hintergrund, Ziele

- **Dauer:** Dezember 2020 bis April 2021
- **Veranstalter:** Actuaries Institute of Australia, Institute and Faculty of Actuaries und Singapore Actuarial Society
- **Plattform:** Kaggle (die weltweit bekannteste Data-Science-Plattform)
- **Ziel:** Vorhersage der Reserven für die Arbeitsunfallversicherung auf Einzelschadenbasis, m. a. W. Entwicklung eines Einzelschaden-Reservierungsmodells für IBNeR (Incurred but not enough Reported)
- **Daten:** Synthetisch erzeugt ohne Bezug auf ein bestimmtes Rechtsgebiet oder Land. Der Datensatz enthielt u. a. anagraphische Daten der versicherten Person, eine Beschreibung des Schadens und eine erste Schätzung des Ultimates.
- **Bewertung eingereichtert Lösungen:** Mittlerer quadratischer Fehler (MSE)
- **Unser Ergebnis:** Der 2. Platz unter den 140 beteiligten Teams/Einzelpersonen

Die Daten im Überblick

- Details zur versicherten Person:

Age	Gender	MaritalStatus	DependentChildren	DependentsOther
43	F	M	1	0

- Details zum Beruf:

WeeklyWages	PartTimeFullTime	HoursWorkedperWeek	DaysWorkedperWeek
509.34	F	37.5	5

- Details zum Unfall:

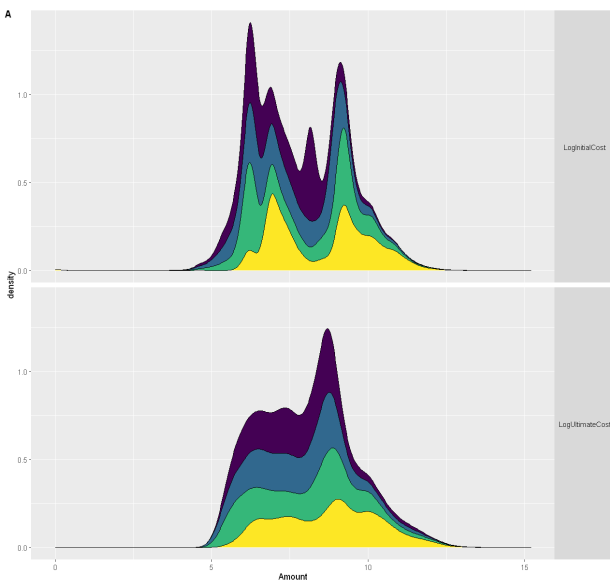
DateTimeofAccident	DateReported	ClaimDescription	InitialIncurredClaimCost
1999-01-07, 11:00:00	1999-01-20	CUT ON SHARP EDGE CUT LEFT THUMB	1500

- Zu schätzen ist das Ultimate:

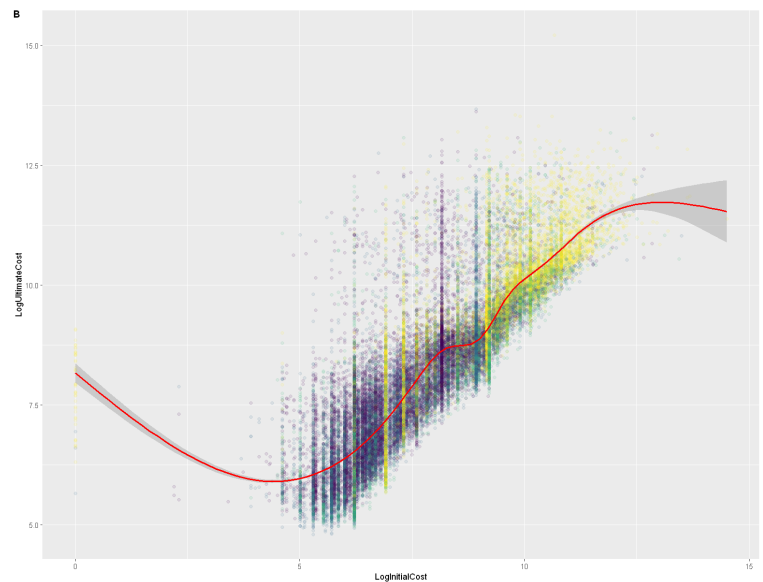
UltimateIncurredClaimCost

4748.203

Initial vs. Ultimate Claim Cost



Verteilung des initialen Ultimates (oben) und echten Ultimate (unten), nach Jahren getrennt

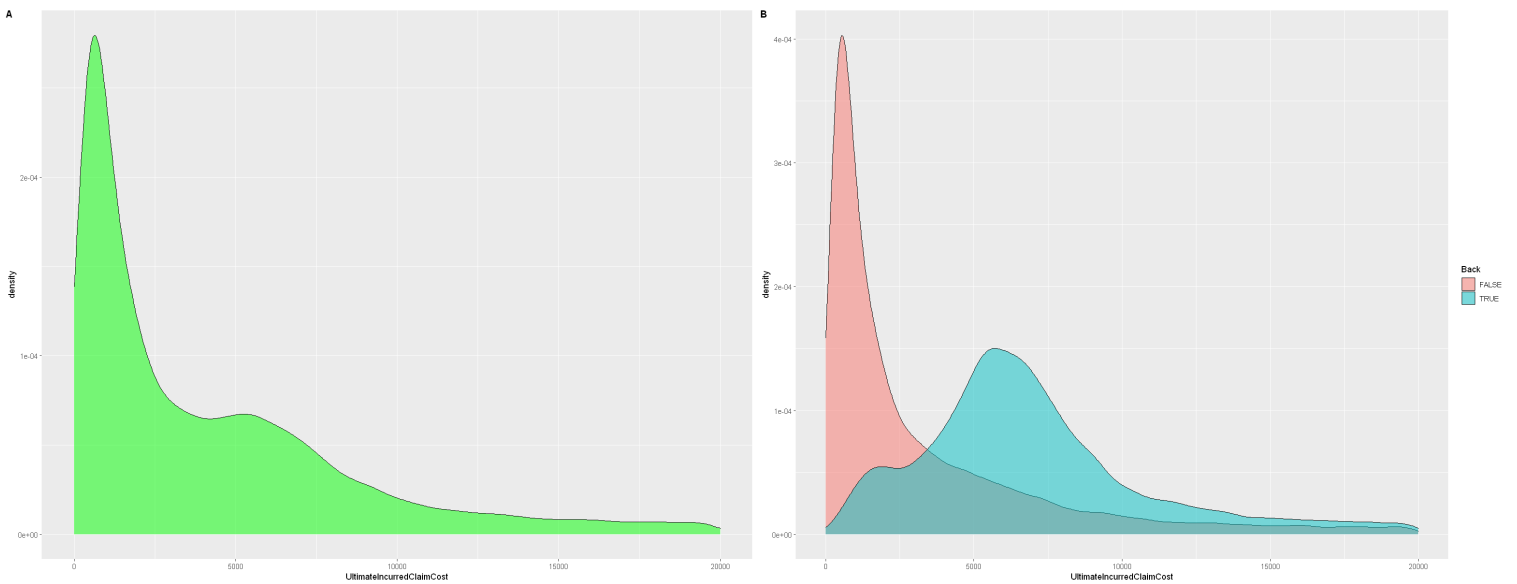


Initiales Ultimate vs. echtes Ultimate.
Herausforderung: Die vertikalen Linien erklären

e.g. eye, cornea => "eye" Cluster



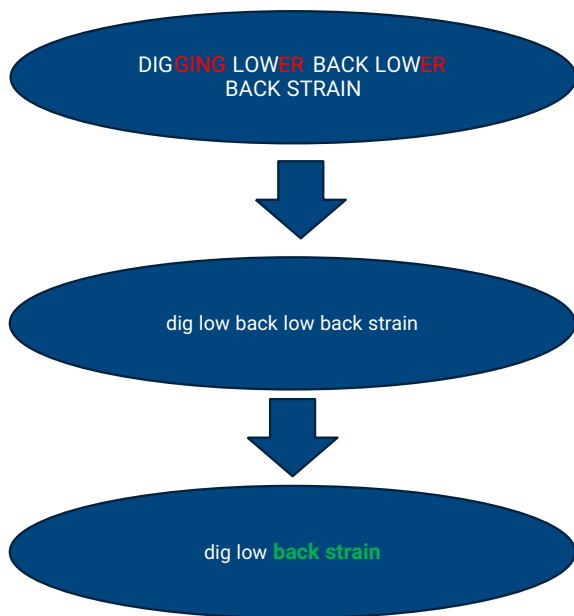
Claim Description: Effekt auf die Verteilung



Claim Description: Analyse


- Die (synthetisch erzeugten) Schadensbeschreibungen bieten nicht viel an grammatikalischer Struktur, manchmal einfach sinnlos, z.B. "TO RIGHT LEG RIGHT KNEE".
- Unser Ansatz zur Analyse:
 - Stoppwörter entfernen ("in", "auf", ...);
 - Lemmatisierung und Stemming der Wörter ("Füße" und "Fuß" werden auf "Fuß" abgebildet, "laceration" und "lacerated" auf "lacer");
 - Clustern und Gewichten von Wörtern nach Ultimate;
 - OHE für die häufigsten Wörter.
- Am Ende haben wir etwa 100 Wörter OHEncodiert und 30 Cluster erstellt.

Claim Description: Ein Beispiel



ClaimDescription

DIGGING LOWER BACK LOWER BACK STRAIN



back	eye	strain	...	Cluster_1	...	Cluster_25	..
1	0	1	...	0	...	1	..

Das Modell

- Unser Algorithmus setzte sich aus den folgenden Ensemble-Methoden zusammen:
 - **Boosting**: Gradient boosting mit xgboost
 - **Bagging**: Random forest als base learner
 - **Voting**: Kombination von Modellen aufgrund Expertschätzungen
- Die Einstellung der folgenden Modellparameter hat unsere Position auf dem Leaderboard deutlich verbessert:
 - *num_parallel_tree*: Die Einstellung dieses Parameters auf eine Zahl größer als 1 ermöglicht die Verwendung von Random Forest als Basismodell;
 - *monotone_constraints*: Das Parameter kann verwendet werden, um z.B. eine positive Zusammenhang zwischen der Anzahl der Kinder und des Schadenaufwands zu erzwingen;
 - *objective*: Einstellung auf reg:gamma und reg:tweedie.

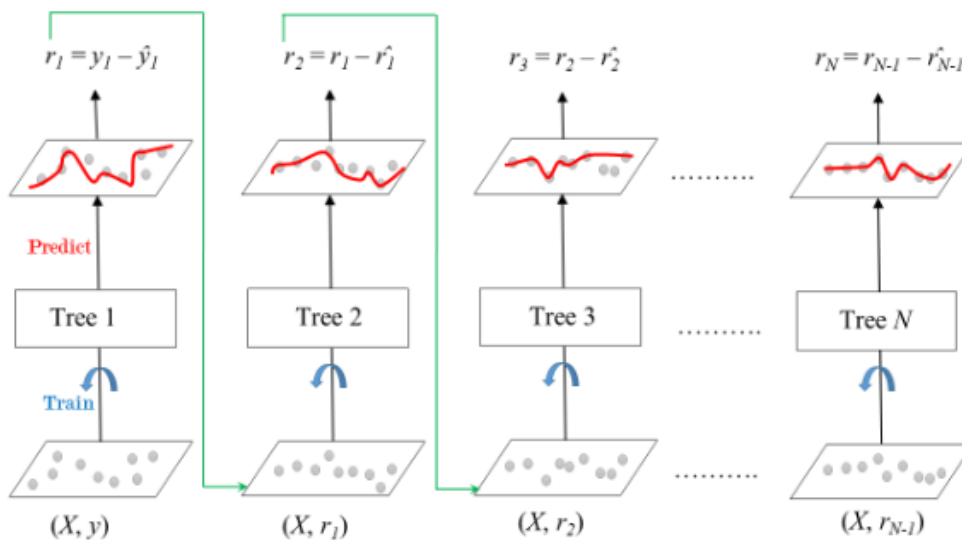
Was heißt Gradient Boosting?

Gradient Boosting ist ein Ensemble-Lernalgorithmus. Er fügt dem Ensemble einen Predictor nach dem anderen hinzu, so dass jeder neue Predictor den Fehler des vorherigen korrigiert.

Ein 3-Predictor Gradient Boosting Algorithmus könnte man so beschreiben:

- Fitte den 1. Predictor an X, y . Vorhersage von \hat{y}^1 .
- Rechne die Residuen des 1. Predictors: $y_2 = y - \hat{y}^1$. Fitte den 2. Predictor an X, y_2 .
Vorhersage von \hat{y}^2 .
- Rechne die Residuen des 2. Predictors: $y_3 = y_2 - \hat{y}^2$. Fitte den 3. Predictor an X, y_3 .

Illustration von Gradient Boosting



source: geeksforgeeks.org

Fazit

Lessons learned:

- Feature Engineering war wichtiger als Hyperparameter Tuning.
- Unsere Maschinen lernten schnell, aber es war zeitaufwendig sie zu pflegen!
- Stapelung der Modellen fuhrte zu Overfitting;
- Neuronale Netze lösen nicht jedes Problem.
- Großschäden hatten einen überproportionalen Einfluss auf die Vorhersagen. (MSE)

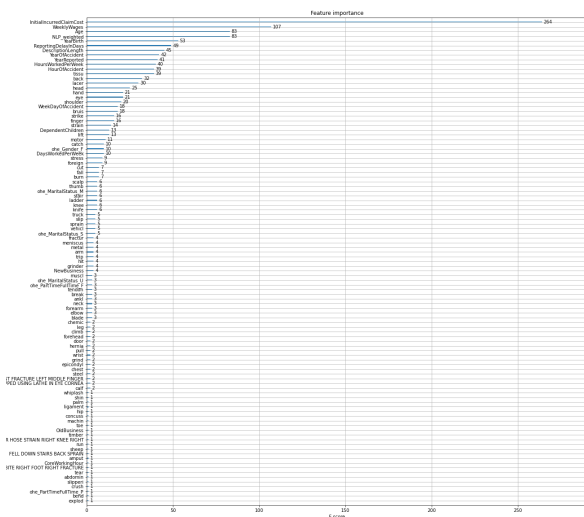
Vielen Dank für die Aufmerksamkeit!

Erste 15 Claim Descriptions

ClaimDescription

LIFTING TYRE INJURY TO RIGHT ARM AND WRIST INJURY
STEPPED AROUND CRATES AND TRUCK TRAY FRACTURE LEFT FOREARM
CUT ON SHARP EDGE CUT LEFT THUMB
DIGGING LOWER BACK LOWER BACK STRAIN
REACHING ABOVE SHOULDER LEVEL ACUTE MUSCLE STRAIN LEFT SIDE OF STOMACH
STRUCK HEAD ON HEAD LACERATED HEAD
FINGER BRUISED AND SWOLLEN LEFT ARM
CLEANING LEFT SHOULDER SPLINTER LEFT HAND
JACK SLIPPED CATCHING FINGER CUT LEFT LITTLE FINGER
STRUCK PINE DUST ABRASION LEFT EYE IRRITATION
STRAINED MUSCLE IN BACK STRAINED LOWER BACK PAIN
TO RIGHT LEG RIGHT KNEE
PICKING UP PARCELS BACK
STRUCK TIMBER RIGHT WRIST
EMPTYING BIN FISH BONE FOREIGN BODY EYE

Feature Importance



- Wie oft wird auf eine Variable im Entscheidungsbaum gesplittet?
- Wichtigste Variable:
 - Initial Cost
 - Weekly Wage
 - Age
- Wichtigste Einzelwörter:
 - Tissue
 - Back
 - Lacer(ation)
 - Head