Schätzung von Versicherungsansprüchen für Krankenkassen in den USA

STUDIENARBEIT DATA MINING, MAREIKE LASS-HENNEMANN

MÄRZ/APRIL 2024

Geschäftsproblem

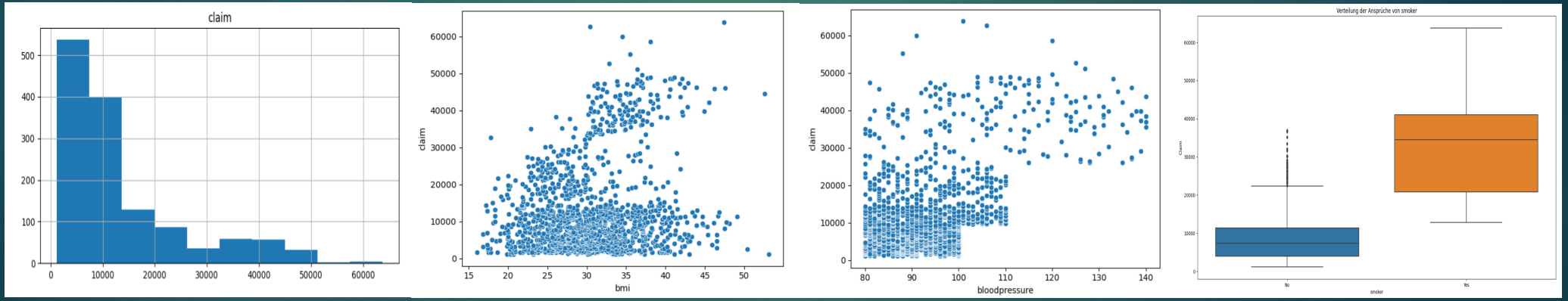
- Gibt es Zusammenhänge zwischen der Ausprägung verschiedener Merkmale wie Alter, Geschlecht, Blutdruck, BMI, Diabetesstatus, Raucherstatus, Kinderanzahl, Wohnregion und der Höhe der Versicherungsansprüche von Versicherten?
- Lassen sich die Versicherungsansprüche anhand eines Modells vorhersagen?
- Zukünftig: Einstufung der Versicherungsprämien (Beiträge, die von Versicherten zu zahlen sind) entsprechend der Vorhersagen

Daten

- Ursprung: Kaggle
- 1340 Zeilen, 10 Spalten.
- kategoriale Variablen: gender, diabetic, smoker, region
- numerische Variablen: Patient ID, age, bmi, bloodpressure, children, claim
  - Es soll “claim“ (Höhe der Ansprüche) als Zielvariable durch die anderen Merkmale erklärt werden.
  - Da der Datensatz durch „claim“ bereits gelabelt ist, kommen Modelle des supervised learning zur Lösung des Geschäftsproblems infrage.

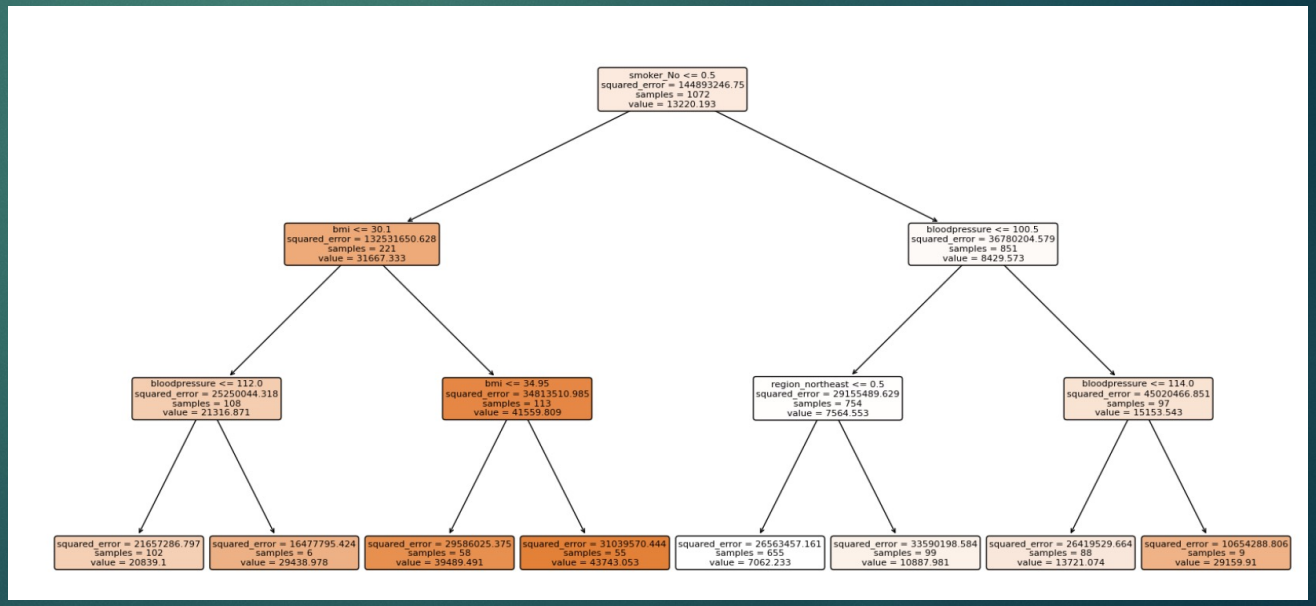
Daten verstehen

- Die Versicherungsansprüche reichen von 1000 bis knapp 64.000 US-Dollar, die mittlere Anspruchshöhe liegt bei 13252 US-Dollar.
- Vor allem die Merkmale BMI, Blutdruck und Raucherstatus scheinen einen Einfluss auf die Höhe der Versicherungsansprüche zu haben



Modell Decision Tree

- Der Decision Tree litt zunächst an starkem Overfitting und konnte nur 68% der Test-Daten erklären. Durch Anpassung der Hyperparameter konnte dies aber deutlich verbessert werden:
  - R2: 0.82
  - Train-MSE: 26936919
  - Test-MSE: 27792932
  - CV-MSE: 28170945
  - RMSE(Test): 5271



Modell Random Forest

- Der Random Forest zeigte ebenfalls zunächst Probleme und wurde mit Hyperparameter-Tuning sowie Feature-Auswahl begrenzt. Er erreichte dann ähnliche bzw. sogar leicht bessere Werte als der Decision Tree:
  - R2: 0.82
  - Train-MSE: 27474453
  - Test-MSE: 27540626
  - CV-MSE: 28888872
  - RMSE(Test): 5247

Bewertung/Evaluation

- mehr Daten!!!
- Der Random Forest könnte mit einem wachsenden Datensatz besser performen, da er durch die Kombination vieler Bäume grundsätzlich besser verallgemeinern kann und weniger anfällig ist für Overfitting als ein DecisionTree.

Implementierung? - Fazit

- Implementierung steht aus…
- mehr Daten - 1340 Fälle nur eine kleine Teilmenge der Gesamt-Versicherten
- Fehler (MSE/RMSE) weiter zu verringern - 5247 US-Dollar recht hoch
  - Zielwert für RMSE??
- Wie soll sich eine vorhergesagte Anspruchshöhe auf die zukünftige Versicherungsprämie der Versicherten auswirken??
- Einsatz“ort“: im Prozess zur Aufnahme neuer Versicherter z.B. direkt nach Abfrage über die bekannten Merkmale und in Verbindung mit der Berechnung der Versicherungsprämie >> direkt Information über die Höhe der Versicherungsprämie an den zukünftigen Versicherten