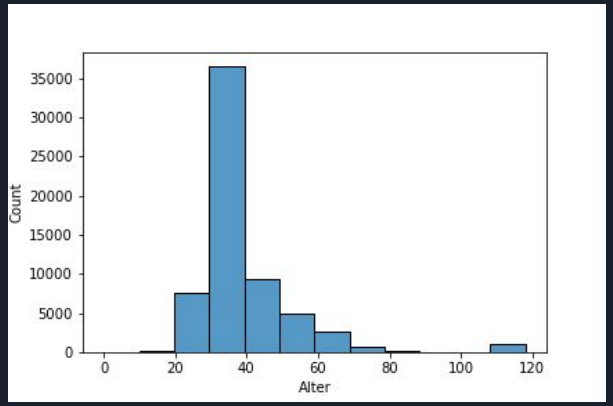
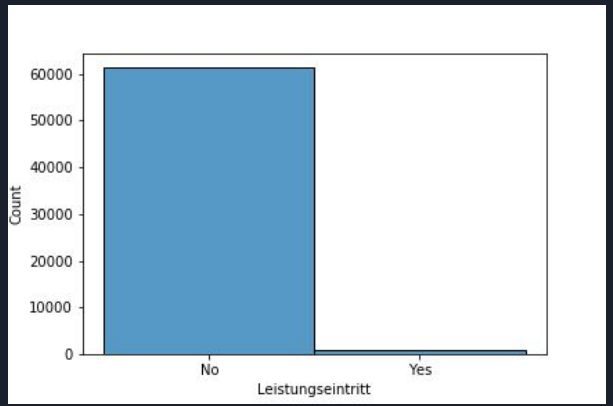
Casestudy Reiseversicherung

Johannes Zacherl

Explorative Datenanalyse

1. Null-Values in der Geschlecht-Spalte
   → Spalte vom Datensatz entfernen
2. Outlier-peak in der Altersverteilung bei über 100 Jahren
   → Zeilen mit Alter über 100 entfernen
3. Leistungseintritt nur in ~1,47% der Fälle
   → Imbalanced Dataset

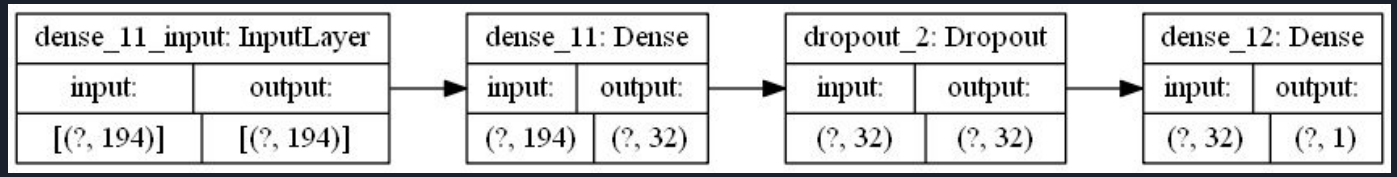




Preprocessing

1. Zielvariable von ‘Yes’ und ‘No’ auf binäre 1 und 0 konvertieren
2. Die übrigen binär kategorischen Variablen auf 1 und 0 konvertieren
3. Die kategorischen Variablen mit mehr als 2 Klassen via One hot encoding konvertieren
4. Die numerischen Variablen auf das Intervall [0,1] reskalieren.
5. Oversampling der unterrepräsentierten Klasse für die Trainingsdaten

Model



Einfaches DNN mit einem hidden Dense Layer und einem Dropout Layer zur Regularization um overfitting entgegen zu wirken.

Loss-Funktion: Binary Crossentropy

Zum Einsatz in einer Webapplikation nutzen der TensorFlow Funktionalität: saved_model und serving

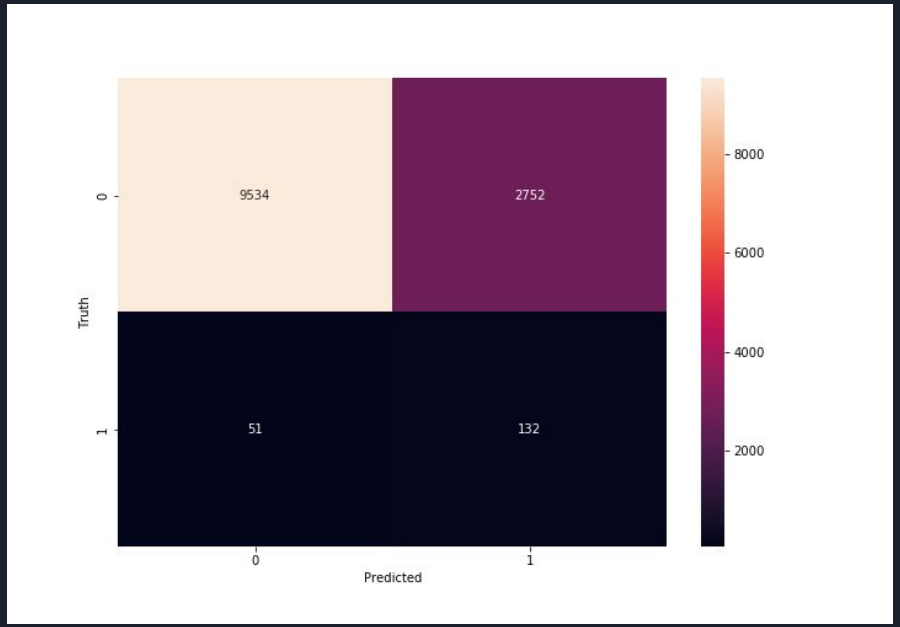
Evaluation

Accuracy: 0.78

Classification Report

Leistungs eintritt	Precision	Recall	f1-score	Support
No	0.99	0.78	0.87	12286
Yes	0.05	0.72	0.09	183

Confusion Matrix



Weiterführende Schritte

1. Preprocessing ins Modell inkludieren (z.B. TesnorFlow preprocessing layer)
2. Verbesserte Strategie um mit dem imbalanced Dataset umzugehen z.B. Synthetic Minority Oversampling Technique und/oder Focal Loss
3. Analyse welche Input Features den Größten Einfluss auf die Vorhersage haben z.B. via Analyse der Ableitungen