GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Erkennung fehlerhafter Air Pressure System Sensordaten von Scania Lastkraftwagen

Lars Dupont, Marius Golombeck, Stefan Tübben

Fachhochschule Dortmund, Fachbereich Informatik, 44227 Dortmund, Deutschland

Einleitung

Diese Ausarbeitung ist das Ergebnis eines DiffPro-ML Miniprojekts im Rahmen des Moduls *Maschinelles Lernen* im Masterstudiengang *Wirtschaftsinformatik* der Fachhochschule Dortmund. Die Modulverantwortung und Betreuung dieses Projekts erfolgte durch Prof. Dr. Christoph Friedrich. Als Basis für die Bearbeitung dient der *„APS Failure at Scania Trucks"* Datensatz [1] der Firma *Scania CV AB Vagnmakarvägen, Stockholm, Schweden*, aus der gleichnamigen Kaggle Competition [1, 2].

Das Luftdrucksystem eines Lastkraftwagens, auch Air Pressure System (APS) genannt, übernimmt u.a. die Regelung kritischer Fahrzeugkomponenten, wie Bremsen und Getriebe. Das APS selbst besteht aus verschiedenen fehlerinduzierenden Komponenten, die eine komplexe Fehlerdiagnostik zur Folge haben. Um Eigendiagnose und automatische Intervention des Systems zu ermöglichen, ist es wichtig, dass APS-Fehler zuverlässig erkannt und von Fehlern anderer Komponenten abgegrenzt werden.

CRISP-DM

Die Ausarbeitung ist an den *„Cross-industry standard process for data mining"* (CRISP-DM) [3] angelehnt. Der Prozess wurde dem Projektumfang entsprechend angepasst (siehe Abbildung 1).

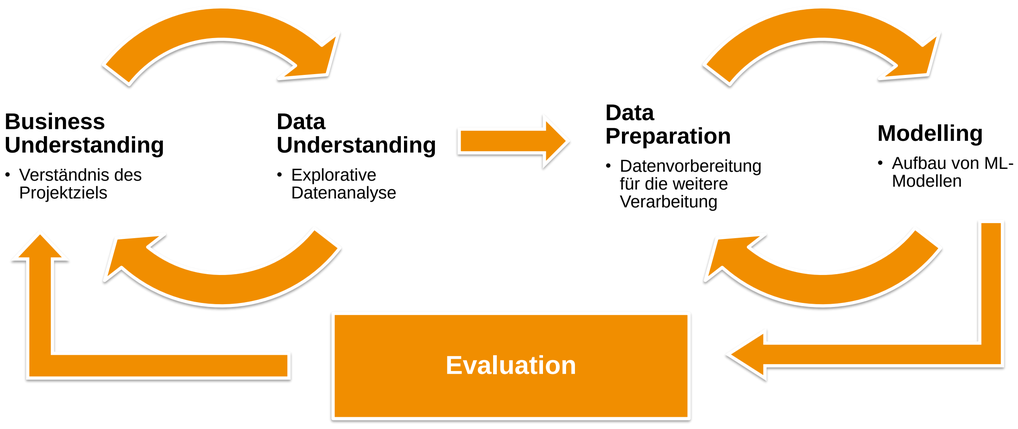

Abbildung 1: CRISP-DM Prozess (angepasst im Sinne des Projektumfangs) (angelehnt an [3])

Business Understanding

Abbildung 2: Business Understanding Prozess

Das APS verfügt über eine Vielzahl an Sensoren, welche dessen Zustand in Echtzeit messen. Ziel ist es, eine automatische Verarbeitung dieser Informationen zu erstellen, welche zuverlässig Fehlerzustände erkennt.
Wichtig ist hierbei, dass fälschlicherweise als Fehler erkannte Normalzustände (False-Positives) und besonders nicht erkannte Fehler (False-Negatives) zu minimieren sind.

Data Understanding

Daten
- Die Daten liegen im CSV-Format vor
- 60.000 Trainings-Datensätze
  - 59.000 Normalzustände
  - 1.000 Fehlerzustände (~1,57%)
- 16.000 Test-Datensätze

Attribute
- 170 Attribute
  - Numerische Werte
  - Keine Einheiten
  - Keine logische Interpretation möglich
  - Hohe Wertspreizung
- Datenqualität teilweise schlecht (bis zu 46.000 fehlende Werte eines Attributs)

Data Preparation

Ersetzen fehlender Attributwerte

Fehlende Attributwerte werden von vielen Modellen nicht unterstützt und müssen daher im Laufe der Data Preparation ersetzt werden. Je nach genutztem Modell werden die fehlenden Werte durch „0", den Durchschnittswert oder den Median ersetzt.

Normalisieren

Für die Optimierung des Datensatzes für ausgewählte Modelle, wurden die Attributwerte normalisiert.

Auswahl der Attribute

Die Attribute des Datensatzes wurden auf Basis ihres Informationsgehalts gefiltert. Attribute, bei denen die Klassifikation in keinem Zusammenhang mit dem Attributwert steht, wurden aus dem Datensatz entfernt (siehe „ba_009" in Abbildung 3).

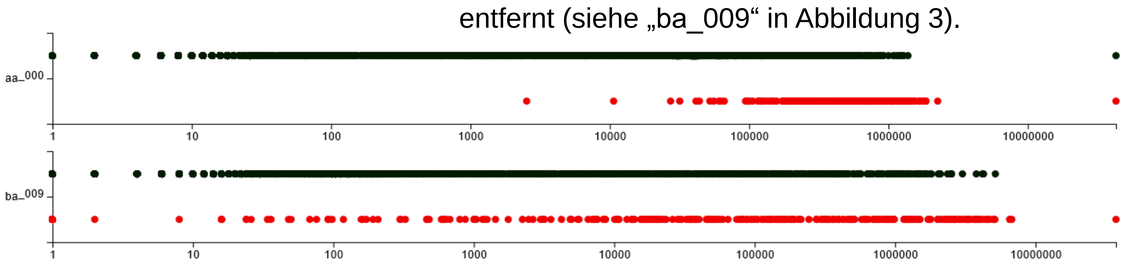

Abbildung 3: Wertverteilung der Attribute aa_000 und ba_009 (rot = Fehler, schwarz = kein Fehler)

Modelling

Die Modellierung erfolgt mit dem Data Science Tool *„KNIME"* [4]. Für die Realisierung wurden optimierte Workflows für die folgenden Modelle angelegt: *Decision Tree, Random Forest (siehe Abbildung 4), Isolation Forest, Probabilistic Neural Network, Multilayer Perceptron, Naive Bayes*.

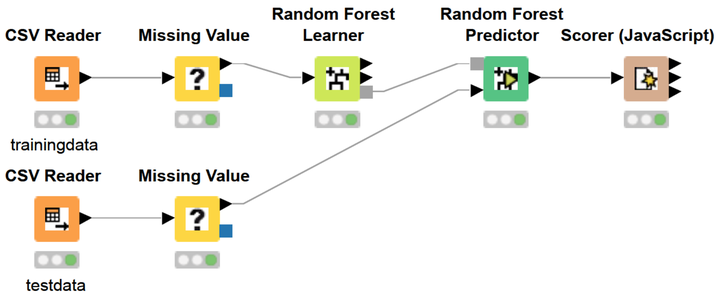

Abbildung 4: KNIME-Workflow des Random Forest Modells

Evaluation

Die implementierten Modelle verarbeiteten die Testdatensätze mit ähnlicher Qualität. Die erreichten Ergebnisse im Kontext der Kaggle Competition sind in Tabelle 1 dargestellt. Abbildung 5 zeigt die Genauigkeiten der Modelle als ROC-Kurven.

Accuracy
94% bis 99%
F-Measure
46% bis 71%
Fehlmeldungen
9 bis 266 Fehler
Nicht erkannte Fehler
26 bis 868

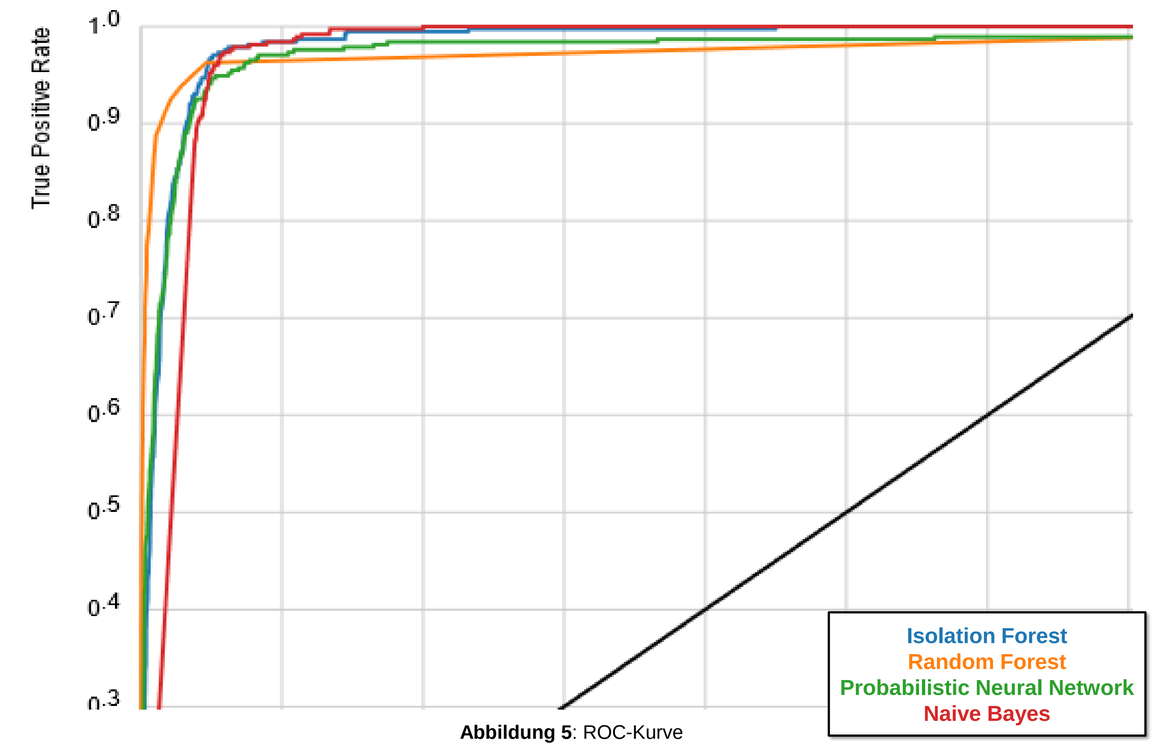

Abbildung 5: ROC-Kurve

Technologie	Accuracy	F-Measure	Typ-1-Fehler False-Positives	Typ-2-Fehler False-Negatives	Punkte (Typ-1-Fehler * 10 + Typ-2-Fehler * 500)
Decision Tree	98,70%	70%	126	84	43.260
Random Forest	98,80%	67%	178	18	10.780
Isolation Forest	**99,00%**	**71%**	148	17	**9.980**
PNN	98,20%	67%	**26**	266	133.260
MLP	98,71%	57%	179	37	20.290
Naive Bayes	94,49%	45%	869	**13**	15.190

Tabelle 1: Modellergebnisse und Vergleich zur Kaggle Competition

Fazit

Die realisierten Methoden können vor dem Hintergrund der Ergebnisse der Kaggle Competition (vgl. [1, 2]) verglichen werden. In diesem Kontext liefert die implementierte Isolation Forest Lösung das vielversprechendste Ergebnis. Die Lösung liefert deutlich weniger Typ-1-Fehler, jedoch schneidet sie bei den Typ-2-Fehlern schlechter ab. Im Vergleich zur Kaggle Competition erreichte die hier vorgestellte Lösung Platz 2.

Teilnehmer	Technologie	Typ-1-Fehler False-Positives	Typ-2-Fehler False-Negatives	Punkte (Typ-1-Fehler * 10 + Typ-2-Fehler * 500)
Costa et. al.	Random Forest	542	9	9.920
Dupont, Golombeck, Tübben	*Isolation Forest*	*148*	*17*	*9.980*
Gondek et. al.	Random Forest	490	12	10.900
Garnaik et. al.	Random Forest	398	15	11.480

Tabelle 2: Vergleich mit den besten drei Lösungen des Wettbewerbs [1]

Referenzen:
[1] APS Failure at Scania Trucks Data Set, https://www.kaggle.com/uciml/aps-failure-at-scania-trucks-data-set/version/2
[2] Advances in intelligent data analysis XV. 15th international symposium, IDA 2016, Stockholm, Sweden, October 13-15, 2016, ISBN: 978-3-319-46348-3
[3] Chapman et. al.: CRISP-DM 1.0, 1999, S. 10, ftp://ftp.software.ibm.com/software/analytics/spss/support/Modeler/Documentation/14/UserManual/CRISP-DM.pdf
[4] KNIME AG, https://www.knime.com

Projekt: Diese Posterpräsentation wurde im Rahmen von DiffPro-ML erstellt. DiffPro-ML wird gefördert durch das Bundesministerium für Bildung und Forschung unter der Förderkennzahl 01|S17075.

Kontakt:

Lars Dupont
ladup001@stud.fh-dortmund.de

Marius Golombeck
magol003@stud.fh-dortmund.de

Stefan Tübben
sttue002@stud.fh-dortmund.de