Lucerne University of
Applied Sciences and Arts

HOCHSCHULE
LUZERN

Informatik

Datenanalyse für das Internet der Dinge im digitalen Gebäude

Themenbereiche: Datenanalyse, Machine Learning

Studierende: Bernasconi Sandro, Brun Eric

Dozent: Prof. Dr. Michael Kaufmann

Experte: Martin Burri

Wirtschaftspartner: Siemens Schweiz AG

Keywords: Machine Learning, Datenanalyse, Data Mining, Regression, CRISP-DM, ANN, Sensoren, AWS

1. Aufgabenstellung

Die vorliegende Arbeit behandelt die Durchführung einer Datenanalyse unter Verwendung von Machine Learning (ML) Algorithmen zur Bestimmung der Raumbelegung in Bürogebäuden. So soll ein Algorithmus mit Sensorwerten wie CO_2-Konzentration in der Raumluft, Raumtemperatur und relativer Luftfeuchtigkeit mit möglichst hoher Genauigkeit bestimmen, ob ein Raum von Personen besetzt ist oder nicht.

In zahlreichen Gebäuden sind Sensoren und Systeme installiert, welche die genannten Werte aufzeichnen und primär für die Steuerung der Heizung, Lüftung und Klimatechnik (HVAC) vorgesehen sind. Viele Gebäude verfügen jedoch über keine Präsenzdetektoren oder anderweitige Sensoren zur Bestimmung der Personenanzahl.

Es soll geprüft werden, ob ein neuer Datenanalyse-Service eine Alternative zum Einkauf und der Installation von dedizierten Präsenzdetektoren darstellen kann. Im Zentrum stehen dabei die zu prüfenden Aspekte der Genauigkeit, Übertragbarkeit sowie die Wirtschaftlichkeit eines solchen Ansatzes.

2. Ergebnisse

Die Ergebnisse dieser Arbeit lassen sich, typisch für ein Data Mining Projekt, in Modelle und Erkenntnisse aufteilen.

Das auf den Daten von zwei Schulungsräumen trainierte und getestete Modell mit dem Logistic Regression Algorithmus, erreichte einen Gesamtgenauigkeitsgrad von 80 % für die Klassifizierung «Raum besetzt» und «Raum frei».

Die wichtigsten Erkenntnisse werden nachfolgend aufgelistet.

- Die untersuchten Präsenzdetektoren ermöglichen ein direktes Klassifizieren in «Raum belegt» und «Raum frei» mit einem Genauigkeitsgrad von > 96 %.
- Die in diesem Projekt trainierten ML-Modelle mit den Attributen CO_2, Temperatur inkl. Änderungsrate und der Raumgrösse ermöglichen eine bis zu 80 % genaue Voraussage des Präsenzdetektorwertes.
- Das Übertragen eines trainierten Modells auf Räume mit anderen Eigenschaften in Bezug auf Raumvolumen und Nutzungsprofil wirkt sich stark negativ auf die Qualität des Modells aus.
- Gebäudespezifische Eigenschaften, wie das Set von verfügbaren Sensoren, die Art der Datenstruktur und die Aufzeichnungsart (Intervall, Event), generieren einen beträchtlichen wiederholten Aufwand bei der Übertragung der Analysemethoden auf andere Gebäude.

- Die Einhaltung der Wirtschaftlichkeit bei dem ML-Ansatz erfordert eine grosse Anzahl von Räumen mit gleichen Eigenschaften und Einflussfaktoren um generalisierte Modellen zu entwickeln.
- Entscheidende Einflussfaktoren bezüglich der Genauigkeit sind zusammengefasst:
    - Lüftungssystem (Art und Funktionsweise)
    - Art der Raumnutzung in Bezug auf Anzahl Personen
    - Raumklima in benachbarten Räumen (Interferenzen zwischen Räumen)
    - Einstellung resp. Kalibrierung des Präsenzsensors
    - Status Türe und Fenster
    - Sonneneinstrahlung, Klimaeinwirkungen über Aussenfassade
    - Unterschiedliche Grundlevel der Werte in verschiedenen Gebäuden

3. Vorgehen / Lösungskonzept

Für diese Arbeit wurde das Vorgehensmodell CRISP-DM (Cross-industry standard process for data mining) verwendet, welches die Arbeiten wie in Abbildung 1 gezeigt in sechs Phasen teilt.

Abbildung 1: CRISP DM Cycle

Business Understanding

In dieser Phase wurden die Ziele aus Business-Sicht aufgenommen, präzisiert und priorisiert. Dabei wurde entschieden, dass der Use Case der binären Klassifizierung eines Raumes (besetzt / nicht besetzt) Vorrang hat gegenüber dem Use Case einer Klassifizierung mit mehreren Klassen (nicht belegt, schwach belegt, stark belegt).

Data Understanding

In einem nächsten Schritt wurden die verschiedenen Datenquellen eruiert und erschlossen. Die total rund 33 Mio. Logeinträge aus Aufzeichnungen über die Zeitperiode von knapp einem Jahr wurden analysiert und mithilfe der Tools Elasticsearch und Kibana in der Amazon Webservice (AWS) Cloud visualisiert.
Als Teil dieser Phase wurde zudem ein Experiment durchgeführt, um die Genauigkeit der Präsenzdetektoren zu überprüfen, da deren Werte als Label für das Training und die Validierung der ML-Algorithmen verwendet wurden. Ein weiteres Ziel des Experiments war, die Störquellen und Einflussfaktoren der HVAC Sensoren zu identifizieren.

Data Preparation

Die Daten wurden mit Python Frameworks in Jupyter Notebook aufbereitet und in die notwendige Form gebracht, um anschliessend als Input für die eingesetzten ML-Algorithmen zu dienen. Insbesondere wurden in dieser Phase die Daten und deren Attribute ausgewählt und Feature Extraction betrieben (also das explizite Ableiten von neuen Attributen wie z.B. der momentanen Steigung eines Wertes, basierend auf bestehenden Attributen).

Modeling

Folgende ML-Algorithmen wurden eingesetzt und mit unterschiedlichen Konfigurationen für die Problemstellung optimiert:

- Logistic Regression
- Random Forest
- Naive Bayes
- Linear Discriminant Analysis
- Gradient Boosting Classifier
- Aritficial Neural Networks (ANN)

Neben Genauigkeitsgrad wurden insbesondere die Metriken Sensitivität, Spezifität und Präzision der einzelnen Modelle verglichen. Die Abbildung 2 zeigt eine Auswahl der Ergebnisse.

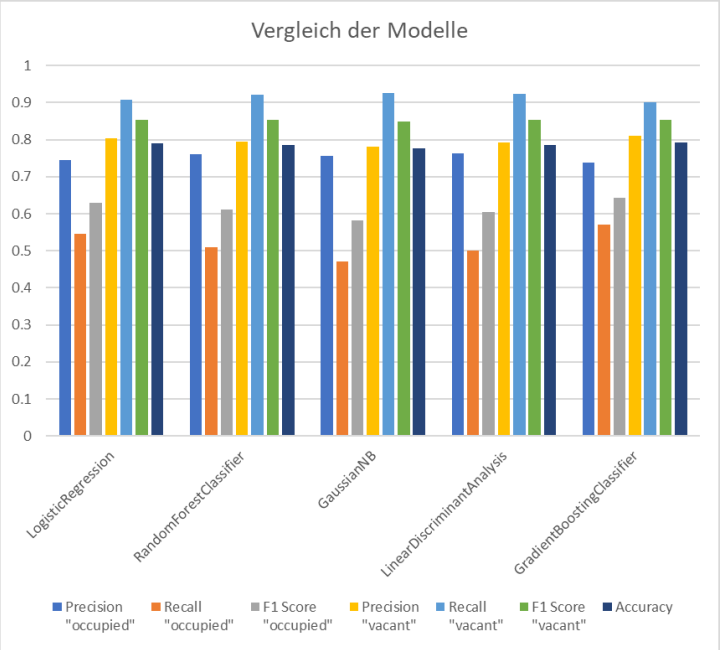

Abbildung 2: Vergleich der Modelle

Evaluation

In dieser Phase erfolgte die Gegenüberstellung der erstellten Modelle inkl. Erkenntnisse mit den Anforderungen aus der Phase Business Understanding. Obwohl mit den eingesetzten ML-Algorithmen im Testszenario ein Gesamtgenauigkeitsgrad von bis zu 80% erreicht wird, sind die Voraussetzungen für ein Deployment, beispielsweise als Web-Service für die Anbindung anderer Räume und Gebäude, noch nicht gegeben. Es wurden konkrete Empfehlungen abgegeben, um bei einem erneuten Durchlaufen des CRISP-DM Zyklus die Qualität der Modelle in Bezug auf die Übertragbarkeit zu steigern.

4. Spezielle Herausforderungen

Eine besondere Herausforderung stellte die Beschaffung von Informationen über die Daten dar. Anfänglich waren weder die Attributbezeichnungen noch die räumliche Verortung der aufgezeichneten Werte klar. Durch das Zuziehen von Fachpersonen und weiteren Datenquellen seitens Auftraggeber konnten diese Fragen geklärt werden und so das nötige Wissen über den vorliegenden Datenbestand erarbeitet werden.
Als sehr aufwendig stellte sich die Wahl der geeigneten Daten und Attribute, sowie deren Vorverarbeitung heraus. Weiter bietet jeder ML-Algorithmus verschiedene Optionen der Hyperparametrisierung was zu einer grossen Anzahl an möglichen Modellkonfigurationen führt. Ein strukturiertes Vorgehen bei der Konfiguration und Validierung half dabei die Resultate nachvollziehbar darzulegen. Im Rahmen der Validierung der Modelle stellte sich die Art des Testdesigns als entscheidender Faktor heraus. Eine Validierung mit einer 10-fachen Kreuzvalidierung erwies sich als deutlich aussagekräftiger als eine statische Aufteilung von Trainings und Testset.

5. Ausblick

Da die Anforderungen an die Qualität und die Wirtschaftlichkeit in Bezug auf die Übertragbarkeit der ML-Modelle auf andere Gebäude noch nicht erfüllt sind, wird ein erneutes Durchlaufen des CRISP-DM Zyklus empfohlen. Bei sehr standortspezifischen Einflussfaktoren wird hingegen der Einsatz von dedizierten Präsenzdetektoren empfohlen. An Standorten, an welchen sich viele Räume in Kategorien mit gleichen Eigenschaften zusammenfassen lassen, bietet das Erstellen von generalisierten ML-Modellen weiterhin Potential zur Optimierung. Weiter könnte durch das Zuziehen von weiteren Informationsquellen, welche auf die Präsenz von Personen schliessen lassen (z.B. WiFi und LAN Nutzungsdaten), die Genauigkeit verbessert werden.