Verbesserung der KI-gestützten Verspätungsursachenbestimmung

Amelie Schüler

Hochschule Darmstadt, Fachbereiche Mathematik und Naturwissenschaften & Informatik

Motivation

Im kooperierenden Unternehmen werden Zugverspätungen manuell nach Verspätungsursache kategorisiert. Dieser Prozess ist mit einem hohen Aufwand und einer gewissen Fehleranfälligkeit verbunden. Aus diesem Grund wurde ein auf maschinellem Lernen basierendes Empfehlungssystem entwickelt, das zu einer Verspätung eine Verspätungsursache empfiehlt und den Mitarbeitenden zusätzlich entscheidungsrelevante Daten zur Verfügung stellt. In der Testphase des Systems wurden jedoch zwei Verbesserungspotenziale identifiziert. Zum einen wurde festgestellt, dass die Verknüpfung zwischen Verspätungen und zusätzlichen Daten, die das Empfehlungssystem verwendet, fehlerhaft ist. Dies kann zu falschen Empfehlungen führen. Vor diesem Hintergrund soll untersucht werden, wie sich die gängigen Data-Matching-Ansätze - regelbasierte und Machine-Learning-Ansätze - hinsichtlich Verknüpfungsqualität, Interpretierbarkeit und Effizienz in diesem Praxisfall unterscheiden. Darüber hinaus wurden bisher ausschließlich Features generiert, die eine spezifische Verspätung betrachten, ohne die Verspätungshistorie oder den lokalen Kontext zu berücksichtigen. Daher soll untersucht werden, wie neue Features entwickelt werden können, die räumliche und zeitliche Aspekte von Verspätungen berücksichtigen. Dabei soll die Generierung möglichst automatisiert erfolgen, um eine möglichst datengetriebene Generierung zu ermöglichen. Darüber hinaus sollen die Features interpretierbar sein und die Modellgüte verbessern.

Vorgehen

Data Matching

Für die Untersuchung des regelbasierten und maschinell lernenden Data-Matching-Ansatzes wurde zunächst eine Datenanalyse durchgeführt. Dabei wurde festgestellt, dass die fehlerhafte Zuordnung von Verspätungen und Zusatzdaten durch eine fehlerhafte Ortsangabe in den zusätzlichen Daten verursacht wird. Die Ortsangabe wird automatisch am aktuellen Ort der Zugfahrt eingetragen, was nicht immer dem tatsächlichen Ort der Meldung/Verspätung entspricht. Durch weitere Analysen konnten insgesamt zwei deterministische Ansätze (Verknüpfungsregeln) und die Features für die Machine-Learning-Ansätze abgeleitet werden:

1. **Baseline** Die zusätzlichen Daten und die Verspätungsdaten werden verknüpft, wenn Tag, Ort und Zugnummer gleich sind.
2. **Erweiterung 1: Verwendung der Ortsangabe aus dem Freitext** Die Daten werden auch verknüpft, wenn Tag und Zugnummer gleich sind und der Ort der Verspätung im Freitextfeld der zusätzlichen Daten vorkommt.
3. **Erweiterung 2: Verzicht auf Ortsangabe** Die zusätzlichen Daten und die Verspätungsdaten werden verknüpft, wenn Tag und Zugnummer identisch sind, ohne dass die Ortsangabe berücksichtigt wird.
4. **Nutzung eines Klassifikationsmodells mit unterschiedlichen Klassifikationsgrenzen** Es wird ein binäres Klassifikationsmodell mit der Zielvariablen $Match = 1$ (richtige Verknüpfung) und $Non - Match = 0$ (falsche Verknüpfung) trainiert. Aus den Verspätungsdaten und den zusätzlichen Daten werden durch Anwendung der oben genannten zweiten Erweiterung Kandidatenpaare (mögliche Matches) gebildet. Die Kandidatenpaare werden dann in das Klassifikationsmodell eingegeben. Ist die Klassifikationswahrscheinlichkeit größer als ein Schwellenwert T_r, wird das Kandidatenpaar als $Match$ klassifiziert.
5. **Featuregewichtung mit Klassifikationswahrscheinlichkeit** Wie beim vierten Ansatz wird zunächst ein binäres Klassifikationsmodell trainiert und Kandidatenpaare gebildet. Statt eines festen Schwellwertes wird jedoch die Modellwahrscheinlichkeit zur Gewichtung der binären Features verwendet. Dadurch kann das KI-gestützte Empfehlungssystem den optimalen Schwellenwert durch die internen Splitkriterien selbst festlegen.

Vorgehen

Automatisierte Feature-Generierung

Für die automatische Generierung von Features, die räumliche und zeitliche Aspekte berücksichtigen, wurden zwei Ansätze verfolgt. Der erste Ansatz betrachtet für einen festen Tag und Ort zu einem Zeitpunkt t, welche Verspätungsursache an diesem Tag am häufigsten aufgetreten ist, um daraus Rückschlüsse für nachfolgende Verspätungen zu ziehen. Zu diesem Zweck wurde die Bibliothek *tsfresh* verwendet, die in der Lage ist, automatisch Features aus Zeitreihen oder auch zeitlich sortierten Daten zu extrahieren. Die daraus resultierenden Features wurden mit Hilfe der statistischen Featureauswahl von *tsfresh* und den Boruta-Algorithmen ausgewählt. Anschließend wurde das ursprüngliche Modell mit dem neuen Modell verglichen. Der zweite Ansatz analysiert, ob es Orte gibt, an denen bestimmte Verspätungsursachen besonders häufig auftreten. Um eine One-Hot-Encoding anwenden zu können, wurde die Anzahl der Orts-Features reduziert und anschließend die neuen binären Orts-Features mit dem Boruta-Algorithmus selektiert. Abschließend erfolgte ein Vergleich des Ausgangsmodells mit dem Modell der Orts-Features.

Ergebnisse

Data Matching

Der Vergleich der beiden Machine-Learning-Ansätze zeigt, dass für einen Schwellenwert $T_r = 0.1$ die Ansätze nahezu identisch sind. Daher wird im Folgenden nur der fünfte Ansatz ("Feature weighting") betrachtet.

- Der Machine-Learning-Ansatz (rot) hat mit einem AUC-Wert von 0.27 die höchste Modellgüte. Auch die *Erweiterung 1 der Baseline* (orange) ist mit einem Wert von 0.24 etwas besser als die Baseline (blau).
- Die *Erweiterung 2 der Baseline* (grün) und das Weglassen der zusätzlichen Daten (violett) zeigen eine deutliche Verschlechterung gegenüber der Baseline (blau).

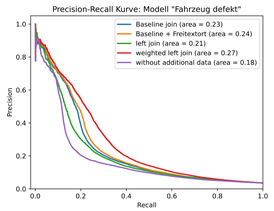

Figure 1. Vergleich der Daten-Matching Methoden

Die Betrachtung der Feature Importance zeigte einen plausiblen Anstieg der Wichtigkeit der neu gewichteten Features, was darauf hindeutet, dass der Anstieg der Modellgüte auf eine Verbesserung der Verknüpfung hinweist und nicht nur ein zufälliges Phänomen war.

Automatisierte Feature-Generierung

Das Empfehlungssystem besteht aus elf zusammengesetzten Modellen. Jedes Modell ist auf eine Verspätungsursache spezialisiert, daher kann hier nur eine verkürzte Version der Ergebnisse dargestellt werden:

1. **Betrachtung des Tagesgeschehens an einem festen Ort**: Die Feature-Auswahl hat drei neue Features ausgewählt. Das erste Feature beschreibt die durchschnittliche Häufigkeit eines Verspätungsgrundes an diesem Tag. Das zweite Feature beschreibt die absolute Verspätungsdauer an einem festen Ort und Tag. Und das dritte Feature beschreibt die Varianz der Verspätungsdauer an einem festen Ort und Tag. Die neu generierten Features führen zu einer starken Zunahme der Modellgenauigkeit. Insbesondere die Verspätungsgründe Signalstörung (+ 322,22 %), Weichenstörung (+ 3900 %) und gefährliche Ereignisse (+ 455,56 %) zeigen einen starken Anstieg.

Ergebnisse

1. Dies ist auch plausibel, da bei diesen Verspätungsursachen häufig mehrere Züge an einem Ort und Tag betroffen sind, während andere Verspätungsursachen wie z.B. Fahrzeugstörung (+ 44,44 %) meist nur einen Zug betreffen und daher einen geringeren Anstieg aufweisen.
2. **Betrachtung der Verspätungsorte** Auch die Hinzunahme der Ortsinformation zeigt eine Verbesserung der Modellgüte. Die größten Zuwächse sind bei den Verspätungsursachen Haltezeitüberschreitung (+61,54 %), Zugbereitstellung (+80 %), Anschluss (+69,23 %) und Behörden (+56,25 %) zu verzeichnen. Der Anstieg ist plausibel, da diese Verspätungsursachen im Vergleich zu den anderen Verspätungsursachen einen lokalen Bezug haben. So wurden beispielsweise für den Verspätungsgrund Behörden vor allem Grenzbahnhöfe ausgewählt, an denen die Bundespolizei häufiger Kontrollen durchführt. Bei den anderen Verspätungsgründen steigt die Modellgüte nur geringfügig, beim Verspätungsgrund Fremdeinwirkung (z.B. Personen im Gleis) sinkt sie sogar um 12,5 %. Hier zeigt sich, dass die Trainingsdaten nicht repräsentativ waren.

Diskussion und Fazit

Data Matching: Vergleich der Verknüpfungsqualität, Interpretierbarkeit und Effizienz

Die dargestellten Ergebnisse zeigen, dass der Machine-Learning-Ansatz die höchste Modellgüte aufweist, so dass davon ausgegangen werden kann, dass dieser Ansatz eine höhere Verknüpfungsqualität aufweist. Der regelbasierte Ansatz verbessert die Modellgüte nur geringfügig. Hinsichtlich der Nachvollziehbarkeit der Verknüpfungsentscheidung schneidet der regelbasierte Ansatz deutlich besser ab, da die Regel leicht nachvollziehbar ist. Der Machine-Learning-Ansatz würde für die Nachvollziehbarkeit zusätzliche Techniken wie LIME oder SHAP-Values benötigen. Auch im Bereich der Effizienzen ist der regelbasierte Ansatz im Vorteil. Im Gegensatz zum Regelbasierten Ansatz ist der Machine-Learning-Ansatz anpassungsfähiger, da mit neuen Trainingsdaten neue Verknüpfungsmuster erlernt werden können. Allerdings ist dieser Ansatz wartungsintensiver, da das Modell ständig auf "Data Drift" oder andere Fehler überprüft werden muss. Außerdem hat sich gezeigt, dass die aktuellen Trainingsdaten nicht repräsentativ sind, da einige Featureausprägungen in den Daten nicht vorkommen. Zusammenfassend kann gesagt werden, dass bei der derzeitigen Datenlage der Machine-Learning-Ansatz nicht empfohlen werden kann. Zu einem späteren Zeitpunkt könnte der Ansatz jedoch sinnvoll sein, da die Modellgüte signifikant höher ist.

Automatisierte Feature-Generierung

Die Ergebnisse zeigen, dass das Hinzufügen von Features, die räumliche und zeitliche Aspekte berücksichtigen, die Modellgüte signifikant verbessern kann. Darüber hinaus sind die meisten Features leicht zu interpretieren und können daher auch an die Nutzerinnen und Nutzer des Empfehlungssystems weitergegeben werden. Schwierig für die Nutzenden zu interpretieren ist die Varianz der Verspätungsdauer, die nicht ohne weitere Erklärung oder Transformation weitergegeben werden kann. Bei den Orts-Features ist fraglich, ob der Ort einer Verspätung ausreicht, um einen Verspätungsgrund zu empfehlen, daher sollte dieses Feature nur zur besseren Differenzierung eines Verspätungsgrundes verwendet werden.

Ausblick

In der Masterarbeit wurden bisher nur räumliche und zeitliche Aspekte betrachtet/berücksichtigt. Ein weiterer vielversprechender Ansatz ist die Betrachtung der Züge bzw. der Zugstrecke. Wie bei den Orts-Feature könnte es Zuglinien geben, die besonders häufig eine bestimmte Verspätungsursache haben. Aufgrund des Fahrplanwechsels im Dezember ist die Verwendung der Zugnummer bzw. Zugliniennummer nicht ohne weiteres möglich. Zukünftige Arbeiten könnten sich daher damit beschäftigen, wie Zugfahrten sinnvoll gruppiert werden können, um diese Information im Empfehlungssystem zu nutzen.

Referent: Prof. Dr. Markus Döhring

Koreferent: Prof. Dr. Timo Schürg