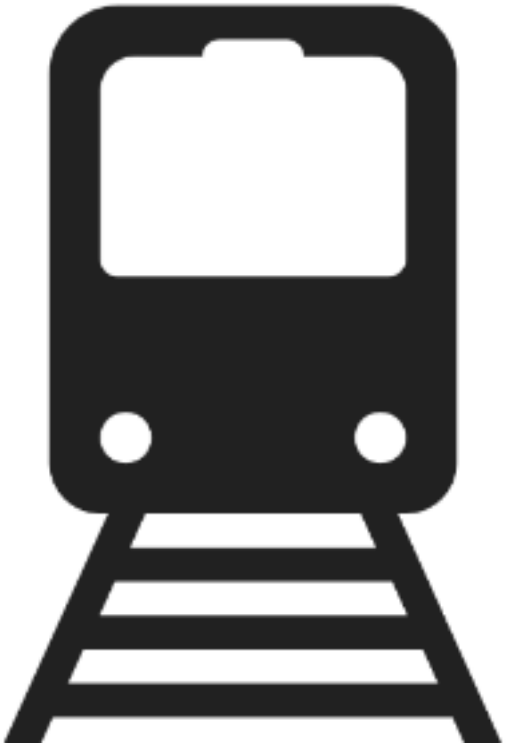
Projektarbeit

AUSTRIAN DATA HERO

Andreas Braun
Christoph Lintner
Florian Voglauer
Jan Vrablicz



Inhaltsverzeichnis

Erarbeitung

Projektidee

1. Predictive Analytics für die Zugauslastung basierend auf historischen Daten, Wetterdaten und Eventdaten.

   Szenario: Wir sind direkt bei der ÖBB angestellt - Abteilung: Innovation und Usability

   a. keine öffentliche Daten für Auslastungen
   b. Eventdaten nur durch Crawling (rechtliche Einschränkungen) verfügbar / Kooperation mit OETicket etc.
   c. Anzahl der ÖBB / VOR App Buchungen
   d. Standortdaten mobile Endgeräte (Wifi Hotspots an Bahnhöfen, nur Verbindungsdaten)
   e. Wetterdaten (Temperatur, Witterung)

   Ziele:
   - Verteilung der Auslastung
   - Kunden profitieren durch besseren Service (Reservierung nötig? Sitzplatz?)
   - Verbesserte Reiseplanung für den Kunden
   - Verbessertes Reiseerlebnis für den Kunden

   Kein Ziel:
   - Perfekte Vorhersage der Auslastung (auf die Person genau)
   - Planungsverbesserung durch die Bundesbahnen -> gezielte Erhöhung der Kapazitäten und Einsparungen bei nicht nötigen Waggons (Vorerst nicht als primäres Ziel deklariert)

Datenplanung

Der Datensatz muss folgende Attribute besitzen:

- Daten des Verkehrsmittels
    - Art des Verkehrsmittel
    - Platz im Verkehrsmittel
- Events (Daten müssen mit einer Zeitdimension abgebildet werden)
    - Ort
    - Zugangsmöglichkeiten (Zuganbindung etc./ wahrscheinlichste Route)
    - Reichweite (regional, national, international)
    - Interesse (basierend auf Social Media in größen wie Retweets, Likes, Zusagen)
    - Zielgruppe
- Störungen/Probleme
    - Art
    - Dauer/Größe
    - evt. Ersatzverkehr
    - Zeitpunkt
- Wetter
    - Temperatur
    - Witterung
    - Wind
    - Zeitpunkt
    - Latitude der Messstation
    - Longitude der Messstation
- Standort
    - Latitude
    - Longitude
    - Anzahl der Handys
- Verkehrsnetz
    - Name der Haltestelle
    - Koordinaten der Haltestelle
    - Polygonzug der Strecken

Data Lifecycle

- Create
    - Wie?
      ÖBB (Auslastung und Haltestellen), Crawling/OE-Ticket (Eventdaten), ZAMG (Wetter)

    - Wie oft?
      Auslastungsdaten (sekündlich), Haltestellen/Strecken (monatlich), Eventdaten (täglich)

    - Woher?
      Auslastung (Mobilfunkanbieter, interne Wifi Stationen), Haltestellen (Staat/ÖBB/OpenData), Crawling (Soziale Medien), Wetter (Partnerschaften mit Wetterdiensten, ZAMG)

    - Prüfung?
      Durch Datenqualitätshandbuch
- Store
    - Wo?
      Cloudservice von AWS

    - Zugriffszeiten?
      5 Minuten

    - Wer?
      Data Engineer

    - Metadaten?
      Nach Dublin Core
- Use
    - Wo?
      Im Internet

    - Zugriffszeiten?
      30 Sekunden

    - Wer?
      Predictive Algorithmus

    - Lizenzen?
      Copyright, individuelle Lizenzen mit Social Media Plattformen
- Share
    - Es ist nicht geplant die Daten zu teilen.

- Archive
    - Wo?
      Eigene Server

    - Kosten?
      Server Anschaffung, Server Wartung

    - Gesetzliche Grundlagen?
      keine personenbezogenen Daten → keine DSGVO
- Destroy
    - Löschung von Daten?
      Speicherung auf eine bestimmte Zeit zur Verbesserung des Modells (5 Jahre)

    - gesetzliche Grundlagen?
      keine personenbezogenen Daten → keine DSGVO
      Bei Datenbeschaffung z.B. durch ÖBB App werden bei In-App Käufen keine persönlichen Daten wie Name des Käufers weitergeleitet

Datenqualität

Nach Datenqualitätshandbuch:

Nötiges Domänenwissen:
Meteorologe für Wetteranalysen. Social Media Analysten für Crawling.

Kostenschätzung

Grobe Schätzung:
Speicherarchitektur einmalig 50.000 €

laufende Kosten (pro Monat):

Daten	100.000 €
Menschen mit Domain Knowhow	20.000 €
Server Wartung	7.000 €
Gesamt (einmalig)	50.000€
Gesamt (laufend)	127. 000€

Lizenzen

- Wetter: kommerziell (Copyright)
- Auslastungsdaten/Störung/Verkehrsnetz/Verkehrsmittel: Werkstudenten der ÖBB, Firmenpolitik sind keine Lizenzen notwendig zwischen den Abteilungen
- Social Media: schriftliche Genehmigung von Nöten oder im Rahmen einer Open Source Lizenz (FB), individuelle Lizenzen (Twitter)

Algorithmus wird in Scotty eingebunden, daher ist keine zusätzliche Lizenz nötig.
Wir vergeben eine Copyright Lizenz. Begründung: Schutz unseres Algorithmus vor kommerzieller anderweitiger Nutzung.

Risikoanalyse

ID	Datumsreferenz	Risiko	P	Kosten (in €)	Eigentümer
R01	Wetterdaten	Daten kommen unvollständig/nicht rechtzeitig	5%	15000	Jan Vrablicz
R02	Verkehrsnetz	Daten kommen unvollständig	1%	5000	Christoph Lintner
R03	Standort	starke Abweichung von tatsächlicher Auslastung	20%	100000	Christoph Lintner
R04	Event	Event mit 500+ Leuten nicht erfasst	1%	25000	Florian Voglauer
R05	Event	Event ungenau (Abweichung um 250+ Personen) bzw. unvollständig erfasst	10%	15000	Jan Vrablicz
R06	Störung	Dauer inkorrekt erfasst (+/- 10 Minuten) und nicht rechtzeitig	30%	5000	Christoph Lintner
R07	Verkehrsmittel	Unvollständigkeit	1%	5000	Andreas Braun
		Risikobudget		24100	

Risikobudget: 24 100€
Verwendung: Versicherung 50%
Rücklagen: 50%

ID	Datensatz	Gültigkeit	Korrektheit	Aktualität	Vollständigkeit
Q01	Verkehrsnetz	in Ö+Grenzgebiete, Koordinaten; Strecke muss hinführen	durch Datenquelle gegeben	nur aktive Haltestellen (d.h. eine Haltestelle, die noch mind. 1x täglich befahren wird) des letzten Quartal	99% Name und Ort müssen ausgefüllt sein
Q02	Event	für ÖBB-Nutzer relevantes Event; Name und oder Standort, Ansammlung von Menschen (mind. 100) an einem Ort, der durch Bahnnetz erreichbar (im Umkreis von 10min Fußmarsch) ist	100 - 500 000 Besucher	täglich	-
Q03	Standort	Standort muss sich an Zug/Haltestelle binden lassen (Radius 10m)	durch Datenquelle gegeben	alle 5 Minuten	-
Q04	Wetter	geografischer Zusammenhang (Umkreis 2km) mit Haltestelle und oder Eventlocation	Temperatur zw -30° und 50°C, Wind in km/h	stündlich	99% Witterung müssen ausgefüllt sein
Q05	Störung	von ÖBB verifiziert, Beeinträchtigung (=Verspätung)	durch Datenquelle gegeben	alle 5 Minuten	99% Ort müssen ausgefüllt sein
Q06	Verkehrsmittel	wenn die Attribute Name, Art, Platzanzahl, bei Eigentumsübernahme durch die ÖBB abgeglichen mit den Herstellerdaten wurden	Die Platzanzahl muss größer 0 sein;	nur aktive Verkehrsmittel (d.h. ein Verkehrsmittel, welches im letzen Monat zumindest einmal gefahren ist.	99% der Namen und Platzanzahl müssen ausgefüllt sein

ID	Konsistenz	Prüfungszyklus	Prüfdatum	Prüfer
Q01	Koordinaten mit 15 Nachkommastellen, in Grad	wenn sich Quelle ändert	bei Implementierung in die Datenbasis	Jan Vrablicz
Q02	Auslastung in Besucherzahl (nicht %), Ort in Adresse (PLZ+Ortsname+Straßenname+Hausnr)	wöchentlich	bei Implementierung in die Datenbasis	Jan Vrablicz
Q03	Koordinaten mit 15 Nachkommastellen, in Grad	monatlich	bei Implementierung in die Datenbasis	Jan Vrablicz
Q04	Grad Celsius, konstante Zeitabstände(1 Messung/Stunde, gleiche Wetterstationen)	wöchentlich	bei Implementierung in die Datenbasis	Jan Vrablicz
Q05	Dauer in Minuten	wöchentlich	bei Implementierung in die Datenbasis	Jan Vrablicz
Q06	Platzanzahl in Personen	wenn sich die Quelle ändert	bei Implementierung in die Datenbasis	Jan Vrablicz

Metadaten (nach Dublin Core):

Wetterdatensatz

- Die mitgelieferten Metadaten der ZAMG werden übernommen.

Eventdatensatz

- ID: OEBB_DS_2
- Technische Daten
    - Format: CSV
    - Type (zb. Sound, Image, Collection...): File
    - Language: DE
- Beschreibung
    - Title: Eventdaten
    - Subject: Menschenansammlung
    - Coverage: Österreich und Grenzgebiete im Radius von 50 km zur Grenze
    - Description: In diesem Datensatz befinden sich die Namen der Events mit einem Zeitstempel, Besucheranzahl und Ort.
- Personen und Rechte
    - Creator: Susi Hauser (fiktiver Name)
    - Publisher: ÖBB Abteilung für Innovation und Usability
    - Contributor: OE-Ticket, Crawling-Menschen
    - Rights: Komerzielle Lizenz mit der OETicket und Open Source Lizenz mit Social Media Plattformen
- Vernetzung
    - Source: Link zum Datensatz
    - Relation:-
- Datum
    - 1.1.2020

Standortdatensatz

- Metadatensatz wird von Mobilfunkanbieter mitgeliefert und verwendet

Verkehrsnetzdatensatz

- ID: OEBB_DS_4
- Technische Daten
    - Format: Shapefile
    - Type (zb. Sound, Image, Collection…): Geofile
    - Language: DE
- Beschreibung
    - Title: Verkehrsnetz
    - Subject: Karte
    - Coverage: österr. Bahnnetz
    - Description: Polygonenzüge über Haltestellen und Streckenabschnitte des österreichischen Bahnnetzes
- Personen und Rechte
    - Creator: Mario Bauer (fiktiver Name)
    - Publisher: ÖBB
    - Contributor: ÖBB
    - Rights: SLA mit der IT Abteilung der ÖBB
- Vernetzung
    - Source: Link zum Datensatz
    - Relation: -
- Datum
    - 1.1.2020

Störungsdatensatz

- ID: OEBB_DS_5
- Technische Daten
    - Format: CSV
    - Type (zb. Sound, Image, Collection…): File
    - Language: DE
- Beschreibung
    - Title: Störungen
    - Subject: Meldungen
    - Coverage: Störungen am österreichischen Bahnnetz
    - Description: Beschreibung der Störung mit einem Standort und einer ungefähren Dauer der Verzögerung
- Personen und Rechte
    - Creator: Fritz Hauer (fiktiver Name)
    - Publisher: ÖBB
    - Contributor: ÖBB
    - Rights: SLA mit der IT Abteilung der ÖBB
- Vernetzung
    - Source: Link zum Datensatz
    - Relation: -
- Datum
    - 1.1.2020

Verkehrsmitteldatensatz

- ID: OEBB_DS_6
- Technische Daten
    - Format: CSV
    - Type (zb. Sound, Image, Collection…): File
    - Language: DE
- Beschreibung
    - Title: Verkehrsmittel der ÖBB
    - Subject: Verkehrsmittel
    - Coverage: Verkehrsmittel die unter dem der ÖBB stehen
    - Description: In diesem Datensatz befinden sich die Namen und Art der Verkehrsmittel und die Platzanzahl.
- Personen und Rechte
    - Creator: Max Mustermann von der ÖBB (fiktiver Name)
    - Publisher: ÖBB
    - Contributor: Siemens, MAN
    - Rights: SLA mit der IT Abteilung der ÖBB
- Vernetzung
    - Source: Link zum Datensatz
    - Relation:-
- Datum
    - 1.1.2020

Technische Maßnahmen um unsere Daten sinnvoll zu schützen umfasst:

- Zugriffsbeschränkung durch Rollenvergaben (Minimalprinzip)
    - Developer (lesen und verarbeiten, Zugriff auf Testdaten)
    - Data Engineer (Lese- und Schreibrechte)
    - Data Owner (Leserechte)
    - Administrator (uneingeschränkter Datenzugriff)

- Schulungen
    - Jährliche Security Awareness Workshops für die Mitarbeiter
    - Interne Audits, die die Kompetenzen der Mitarbeiter überprüfen soll

- Datenaufbewahrung
    - Zentralisierte Datenbasis mit Hot Stand-By Server als Ersatz
    - Zyklische inkrementelle Backup-Erstellung (Monatlich)
    - Daten werden im RAID 5 Standard gespeichert
    - In jedem Serverraum wird ein Feuerlöscher bereitgestellt

- Zutrittsbeschränkungen Serverraum
    - ID-Karte für Zutritt zum Serverraum
    - Passwort (monatliche Änderung)

- Log File anlegen
    - Veränderungen und Löschung aufzeichnen
    - Rollen / Person der Veränderung

Datenschutz

Verfügbarkeit

Die Daten gelten als ausreichend verfügbar, wenn eine maximale Abrufzeit von 30 Sekunden gegeben ist.

Integrität

Die Daten gelten als integer, wenn sie echt vollständig und aktuell sind.
Daten gelten als echt, wenn sie die Primärquellen (ÖBB, ZAMG, OE-Ticket, Social Media) abbilden.
Daten gelten als vollständig, wenn sie die Qualitätskriterien erfüllen.
Daten gelten als aktuell, wenn sie die Qualitätskriterien erfüllen.

Vertraulichkeit

Die Daten gelten als vertraulich, wenn der Zugang durch Rollen geregelt ist.

Nichtverkettbarkeit

trifft nicht zu

Transparenz

trifft nicht zu

Intervenierbarkeit

trifft nicht zu

Datenschutzbeauftragter

Die Einhaltung der datenschutzrechtlichen Richtlinien wird vom Datenschutzbeauftragten der ÖBB kontrolliert - dieser ist auch für uns zuständig.