

Machine Learning @HR.

Employee Churn Prediction bei Helvetia.

Karin Timcke
Head HR-Analytics & Payroll

Swiss HR Analytics
18. September 2025

einfach. klar. helvetia 
Ihre Schweizer Versicherung

Agenda.

Topic

01 Einleitung: Warum Churn Prediction wichtig ist

02 Projektziel & End-to-End Überblick

03 Datenarchitektur & Feature Engineering

04 Evaluation & Modeling Approach

05 Ergebnisse & Business Impact

06 Erfolgsfaktoren & Herausforderungen

Einleitung: Warum Churn Prediction wichtig ist.

Success Stories. Predictive Analytics.

Beispiele von Unternehmen, welche durch die Senkung der Fluktuation erhebliche Kosten eingespart haben, indem sie in Maßnahmen zur Mitarbeiterbindung investierten.



- IBM: Nutzt KI & Predictive Analytics zur Vorhersage von Fluktuation (bis 95 % Genauigkeit).
Ergebnis: gezielte Retentionsmassnahmen & Einsparungen von ca. 300 Mio. USD.



- Google: Analysiert Feedback, Arbeitszeiten & Teamdaten zur Früherkennung von Burnout.
Ergebnis: höhere Mitarbeiterzufriedenheit & -bindung.



- SAP: Predictive Analytics auf Basis von Employee Central (z. B. Gehalt, Leistung, Teamstrukturen).
Ergebnis: Frühzeitige Identifikation von Hochrisikobereichen & gezielte HR-Massnahmen.

Mitarbeiterabgänge. Scope & Definition

- **Definition:** Mitarbeiterfluktuation bezeichnet den Prozess, bei dem Mitarbeitende eine Stelle verlassen und deren Positionen neu besetzt werden müssen.
- **Types of Turnover:**

Intern: Transfers oder Beförderungen innerhalb des Unternehmens

Extern: Mitarbeiter verlassen das Unternehmen komplett

Scope: Freiwillige Mitarbeiterabgänge

Mitarbeiterkündigungen bedeuten: Wissensverlust, hohe Rekrutierungskosten und Brüche in der Unternehmenskultur.



Warum Mitarbeiterfluktuation vorhersagen?

- **Hohe Kosten** durch Mitarbeiterfluktuation – Rekrutierung, Onboarding, Produktivitätsverluste
- **Verlust von kritischem Wissen & Fähigkeiten** – Schlüsselrollen = Schlüsselrisiken
- **HR reagiert oft zu spät** – heutige Tools sind reaktiv, Feedback der Führungskräfte kommt nach der Entfremdung
- **Predictive Analytics ermöglicht frühzeitiges Eingreifen** → datenbasierte Retentionsstrategie
- **Wettbewerbsvorteil durch Mitarbeiterbindung** – Top-Talente zu halten ist kein „Nice-to-have“, sondern ein strategischer Differenzierungsfaktor

Projektziel & End-to-End Überblick.

Projektziel.

Um Mitarbeiterabwanderung proaktiv zu begegnen, haben wir ein Machine-Learning-Modell auf Basis historischer Personaldaten entwickelt. Das Modell prognostiziert die Wahrscheinlichkeit von Fluktuation in zentralen Mitarbeitersegmenten und ermöglicht es HR, gefährdete Gruppen frühzeitig zu identifizieren und gezielte Bindungsmassnahmen einzuleiten.

- **Typischer HR-Anwendungsfall von Predictive Analytics**
- Basierend auf **historischen Daten** zu Profilen, Verträgen, Gehalt u. a.
- ML erkennt **Muster**, die mit Kündigungsrisiken verbunden sind
- Ermöglicht **proaktive Bindungsmassnahmen**, bevor es zu spät ist



Prozessablauf & Vorgehen. Machine Learning.

**1. Daten
sammeln**

**2. Modell
erstellen**

3. Training

4. Testen

**5.
Vorhersagen**

Prozess-Schritte

Zunächst werden Daten gesammelt. Das können Zahlen, Bilder, Texte oder andere Arten von Informationen sein

Ein Modell ist eine Art mathematischer Algorithmus, der darauf trainiert wird, Muster in den Daten zu erkennen. Es gibt verschiedene Arten von Modellen, z. B. Entscheidungsbäume, neuronale Netze und Support-Vektor-Maschinen etc.

Das Modell wird mit den gesammelten Daten "trainiert". Das bedeutet, dass das Modell die Daten analysiert und versucht, Regeln oder Muster zu erkennen.

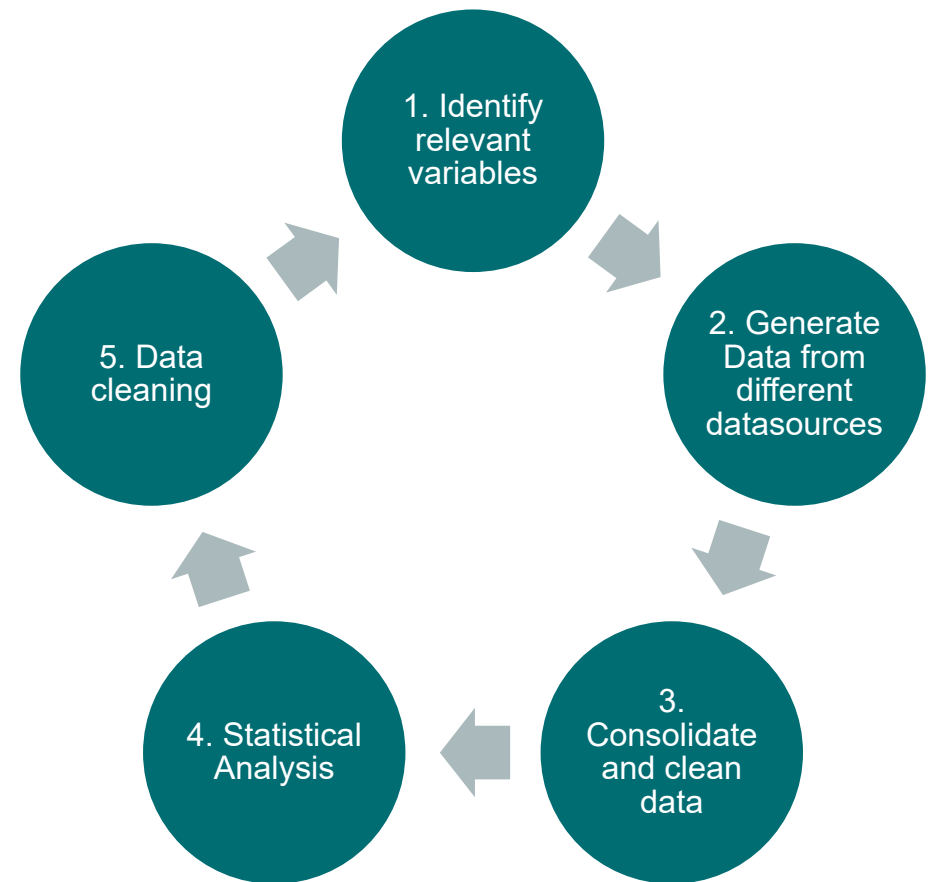
Nachdem das Modell trainiert wurde, wird es mit neuen, bisher ungesehenen Daten getestet, um zu überprüfen, wie gut es gelernt hat. Man misst die Genauigkeit und andere Leistungsmetriken.

Wenn das Modell gut genug ist, kann es verwendet werden, um Vorhersagen zu treffen oder Entscheidungen zu treffen.

Datenarchitektur & Feature Engineering.

Von Rohdaten zu Insights – Datenquellen & Umfang.

- Datenquelle: SAP HCM und SuccessFactors (Helvetia Schweiz)
- Zeitraum: Nur Mitarbeiterdaten aus 2022 und 2023 → Ältere Daten ausgeschlossen (Reorganisationen, Systemwechsel, inkonsistente Datenstrukturen)
- Variablenauswahl: basierend auf interner Datenverfügbarkeit, -qualität und -konsistenz
- Stichprobengröße



Modellanforderungen / Voraussetzungen.

Für den effektiven Einsatz von Daten in einem Machine-Learning-Modell müssen zentrale Voraussetzungen erfüllt sein. Diese sichern eine präzise Modellierung und verlässliche Prognosen:

- **Korrekte & vollständige Daten** – fehlende oder inkonsistente Einträge verfälschen Ergebnisse
- **Konsistente & saubere Datenstrukturen** über Zeiträume hinweg (Qualität & Struktur)
- **Relevante Features** – nur Variablen mit Bezug zur Fluktuation berücksichtigen
- **Klare Zielvariable** – z. B. „freiwillige Kündigung: Ja/Nein“
- **Ausreichendes Datenvolumen** – gross & repräsentativ
- **Rechtliche & ethische Compliance** – inkl. Datenschutz & Anonymisierung

Was ist Feature Engineering?

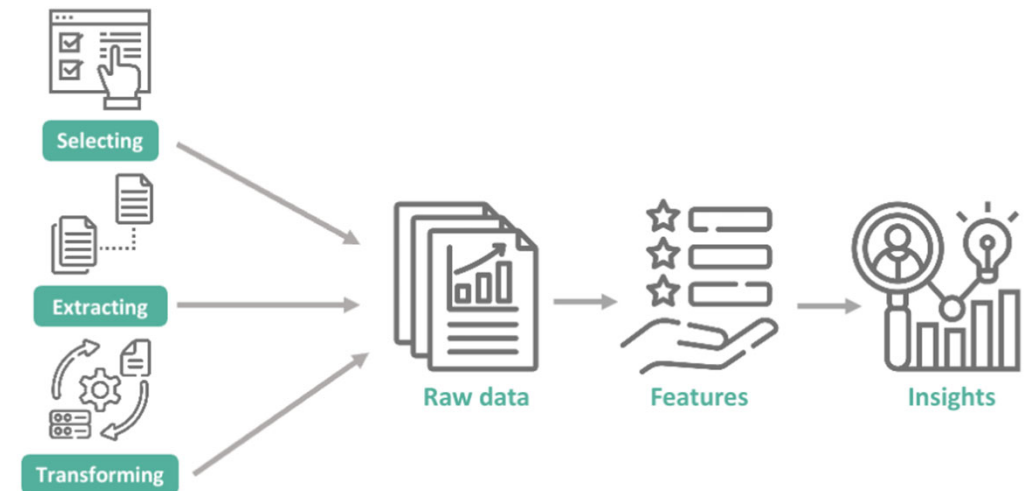
Faktoren, die Mitarbeiterkündigungen beeinflussen.

Warum Feature Engineering wichtig ist

- Steigert die Modellgenauigkeit
- Erhöht die Transparenz des Modells

Was Feature Engineering beinhaltet

- **Feature-Auswahl** → nur relevante Variablen berücksichtigen
- **Transformation von Variablen** → kategoriale Werte kodieren
- **Neue Features erstellen** → Variablen kombinieren oder ableiten
- **Irrelevante Features entfernen** → Variablen ausschliessen, die Rauschen erzeugen oder Multikollinearität verursachen



Evaluation & Modeling Approach.

Evaluation ML Algorithmen

	Eignung / Stärken	Wie es funktioniert?	Herausforderungen
XG-Boost	Hervorragend geeignet für HR-Prognosen, insbesondere zur Vorhersage von Mitarbeiterfluktuation. Hohe Genauigkeit, Effizienz und Interpretierbarkeit.	Erfasst nichtlineare Zusammenhänge und komplexe Wechselwirkungen.	Erfordert sorgfältiges Hyperparameter-Tuning und Parallelverarbeitung.
Random Forest	Robust bei Overfitting. Funktioniert gut bei nicht-linearen Daten.	Mehrere Entscheidungsbäume für Prognosen.	Langsamere Verarbeitung bei sehr grossen Datensätzen.
Support Vector Machines (SVM)	Wirksam für kleine bis mittelgrosse Datensätze. Gut geeignet bei klarer Klassentrennung.	Maximiert die Trennung zwischen Klassen durch optimierte Hyper-Ebenen.	Hoher Rechenaufwand & sensibel bei Parametern.
Deep Learning	Leistungsstark für komplexe, hochdimensionale Daten mit vielen Wechselwirkungen.	Verwendet neuronale Netze, um Muster in grossen Datensätzen zu erkennen.	Aufwendig, schwer erklärbar, braucht viele Daten.

Modeling Approach.

Machine Learning Ansatz

- XGBoost Classifier als Kernmodell – besonders geeignet bei unausgeglichene Daten, nichtlinearen Zusammenhängen und Feature-Interaktionen
- Umgang mit Klassenungleichgewicht durch SMOTE, Borderline-SMOTE und SMOTE-ENN
- Hyperparameter-Tuning mit GridSearchCV zur Bestimmung optimaler Modelleinstellungen
- Modellevaluation über Accuracy, Precision, Recall, F1-Score, ROC-AUC und Precision-Recall-Kurve

Interpretierbarkeit & Business-Relevanz

- Einsatz von SHAP (Shapley Additive Explanations) zur Identifikation der wichtigsten Einflussfaktoren Liefert transparente, managementtaugliche Insights zu den Modellvorhersagen

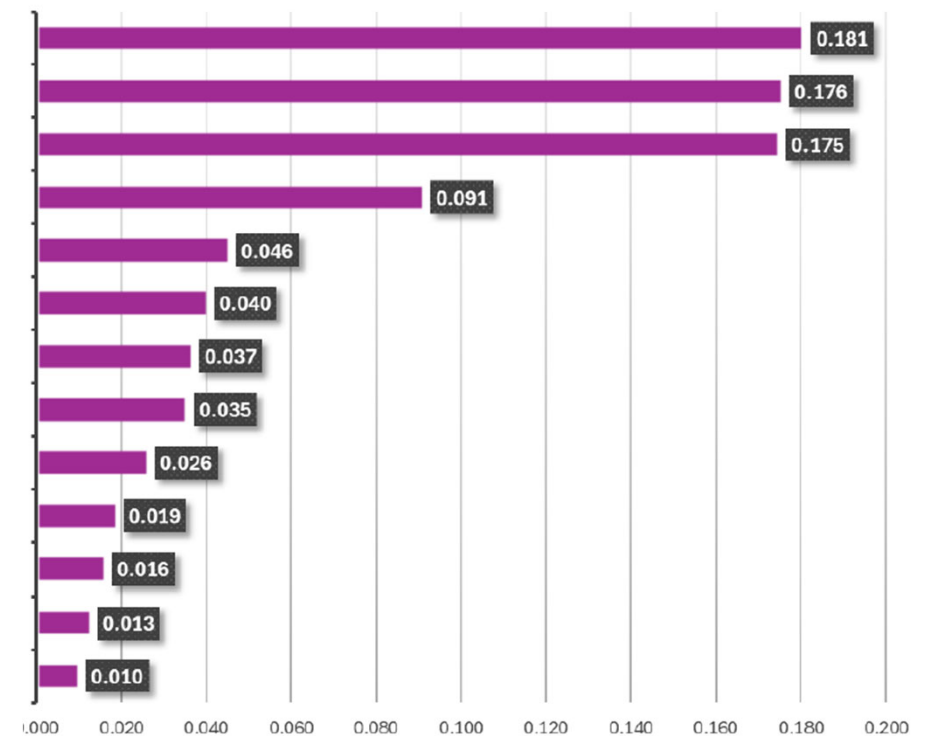


Ergebnisse & Business Impact.

Ergebnisse. Ranking der Einflussfaktoren von freiwilligen Mitarbeiterabgängen.

Das Balkendiagramm fasst die Ergebnisse nach Kategorien zusammen und veranschaulicht die relative Bedeutung der verschiedenen Faktoren.

To preserve anonymity, the influencing factors were concealed.



Ergebnisse. SHAP Values.

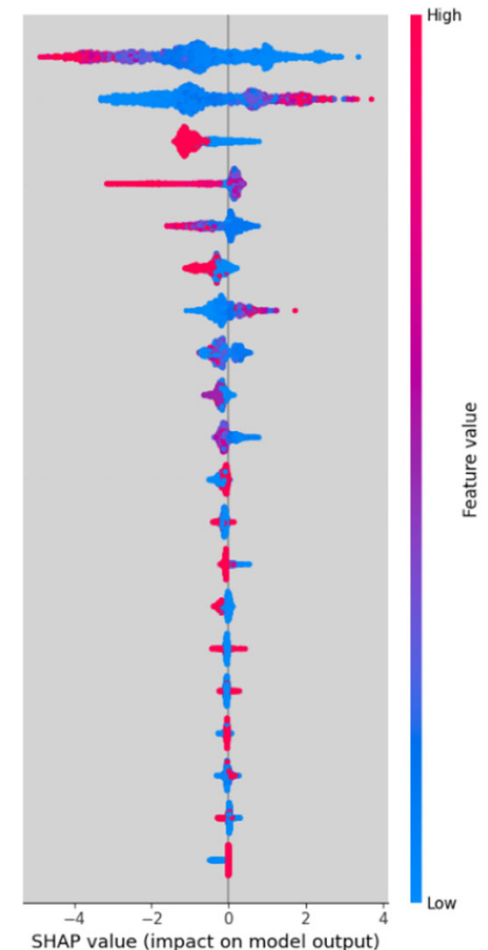
SHAP-Werte (Shapley Additive Explanations) machen die Vorhersagen von XGBoost zur Mitarbeiterfluktuation transparent, indem sie aufzeigen, welche Faktoren das Ergebnis beeinflussen.

SHAP berechnet den Beitrag jedes einzelnen Features zur Prognose eines Austritts – basierend auf spieltheoretischen Prinzipien.

Wie man SHAP-Werte liest:

- Jede Zeile = ein Feature, nach Wichtigkeit sortiert (oben = grösster Einfluss).
- X-Achse = SHAP-Wert → zeigt, ob das Feature die Kündigungswahrscheinlichkeit erhöht (rechts, positiv) oder senkt (links, negativ).
- Jeder Punkt = ein Mitarbeitender.
- Farbe: Ausprägung des Werts (rot = hoch, blau = niedrig). Position: Einfluss auf die Vorhersage.

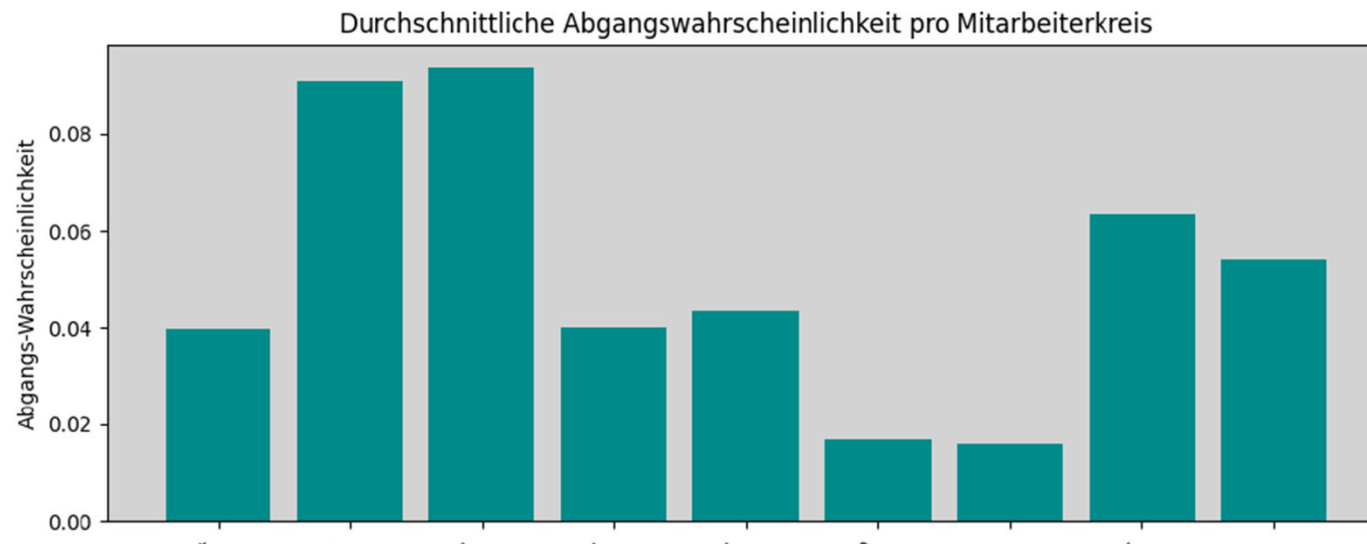
To preserve anonymity, the influencing factors were concealed.



Ergebnisse. Fluktuationswahrscheinlichkeit nach Mitarbeitergruppen.

Die im Modell berechneten Wahrscheinlichkeiten zeigen, wie wahrscheinlich es ist, dass ein Mitarbeitender aus einer bestimmten Kategorie das Unternehmen innerhalb eines Jahres verlässt.

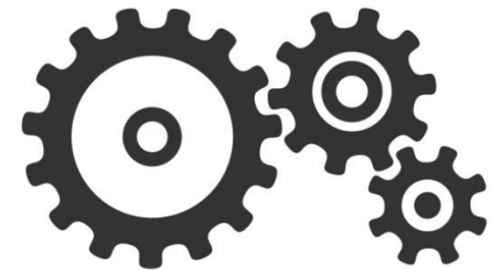
To preserve anonymity, the employee groups were concealed.



Business Impact.

„Predictive-Analytics-Anwendungen geben uns Hinweise auf wahrscheinliche zukünftige Entwicklungen. Der entscheidende Schritt ist die Umsetzung – hier werden die Investitionen platziert.“ — Jac Fitz-Enz

- **Früherkennung von Risiken:** Datenbasierte HR-Strategien identifizieren und adressieren
- **Kostensenkung:** Geringere Fluktuation reduziert Rekrutierungs-, Onboarding- und Opportunitätskosten.
- **Wettbewerbsvorteil:** Stärkere Mitarbeiterbindung sichert Know-how, Stabilität und Arbeitgeberattraktivität.
- **Strategische Steuerung:** Frühe Risikoerkennung ermöglicht proaktive HR-Massnahmen statt reaktivem Krisenmanagement.



Erfolgsfaktoren & Herausforderungen.

Erfolgsfaktoren.

- **Datenqualität & -verfügbarkeit** – vollständig, konsistent, aktuell → Grundlage für valide Analysen
- **Technologie & Infrastruktur** – zuverlässige Tools und Pipelines für Datenaufbereitung → moderne Datenbanken beschleunigen Prozesse und sparen Ressourcen
- **Datenstrategie – frühzeitige Definition relevanter Daten** → erleichtert zukünftige Nutzung und schafft Entscheidungsbasis
- **Kompetenz** – interdisziplinäres Team mit technischer und HR-Expertise
- **Ressourcen – ausreichende Kapazität für Entwicklung, Testing & Implementierung** → ML-Modelle sind iterativ & ressourcenintensiv
- **Datengetriebene Kultur** – Schulungen zur Datenkompetenz, damit Mitarbeitende Analysen korrekt interpretieren und nutzen können

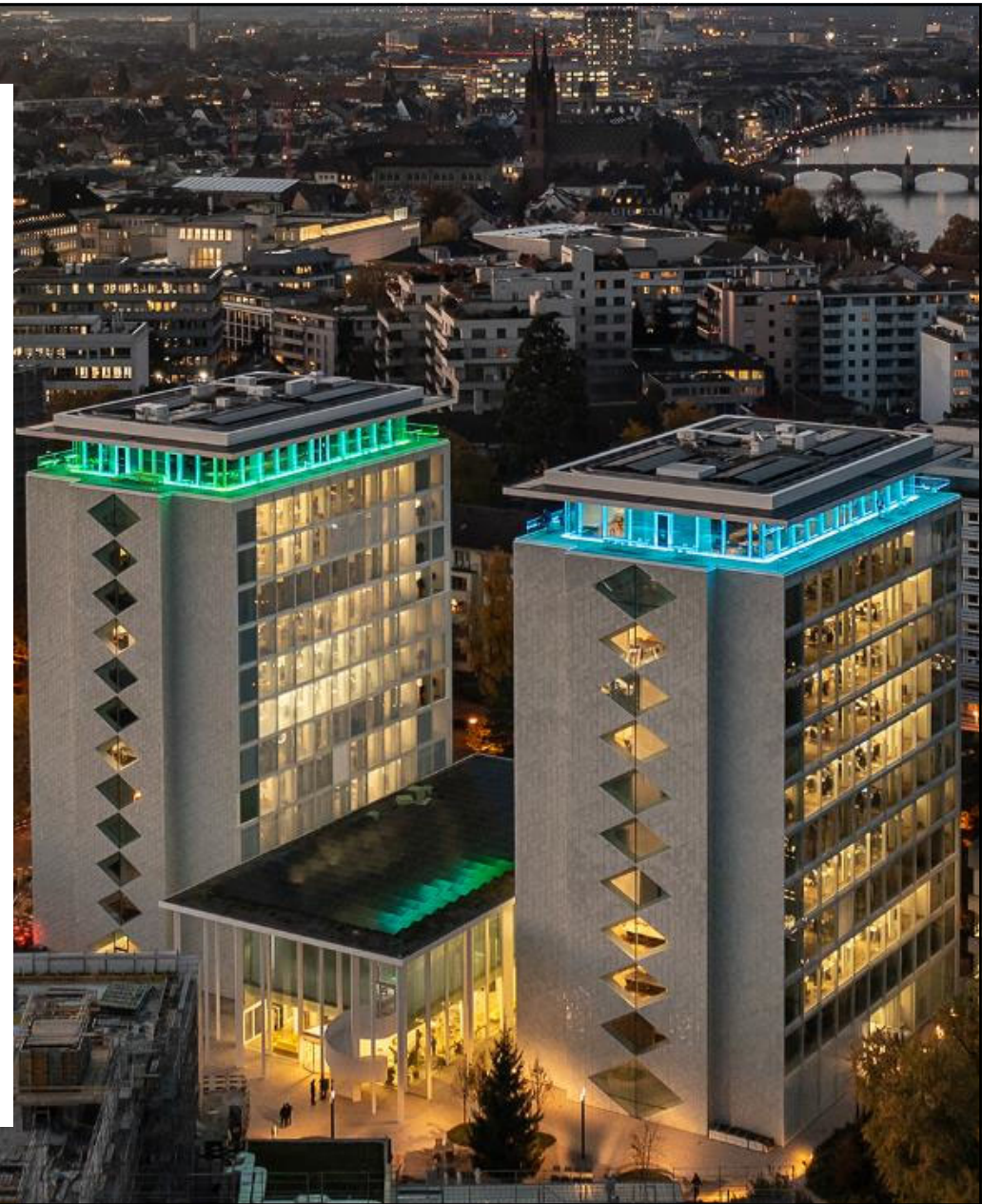


Herausforderungen.

- **Systemlandschaft:** Fragmentierte Systeme erschweren den einheitlichen Datenimport; manuelle Excel-Aufbereitung ist ressourcenintensiv & fehleranfällig; keine Echtzeitdaten verfügbar → nur Batch-Loads möglich.
- **Datenkonsistenz:** Organisatorische Veränderungen beeinträchtigen Vergleichbarkeit über Abteilungen und Zeiträume hinweg.
- **Modellhandling:** Stratifiziertes Splitting für repräsentative Daten GridSearchCV für Hyperparameter-Tuning Threshold-Anpassung & SMOTE gegen Klassenungleichgewicht und Recall-Probleme Early Stopping zur Vermeidung von Overfitting
- **Monitoring der Feature-Importance für Stabilität Bias-Risiko:** Unterschiede zwischen Vertrieb (Aussendienst) und Office-Personal können zu impliziten Verzerrungen führen (z. B. finanzielle Anreize wichtiger im Vertrieb, Arbeitsplatzsicherheit relevanter im Office).
- **Trade-off:** Separate Analysen pro Gruppe könnten Unterschiede aufzeigen, reduzieren aber die Stichprobengröße → Gefahr statistischer Instabilität.

**Vielen Dank
für Ihre
Aufmerksamkeit.**

einfach. klar. helvetia 
Ihre Schweizer Versicherung



Appendix

Where the magic happens. Technical base.

- Helvetia Analytics Platform
- Jupyter Notebook
- Python Code

The screenshot displays the Helvetia Analytics Platform dashboard. At the top, the browser address bar shows the URL: https://helvetia-analytics-platform.awhap.helv.cloud/?iss=https%2F%2Fidp.helvetia.com:443%2Fauth%2Foauth2%2Fclient_id=hap_dashboard_public_prod. The dashboard header includes the Helvetia logo and the text "Helvetia Analytics Platform".

Navigation: A sidebar menu on the left contains links for "Sandbox Home", "Wiki", "User Manual", "Service Now", and "Data Catalog".

Sandbox Status: The status is "running" (indicated by a green checkmark). A note states: "Please note that your sandbox is outdated, and will need to be recreated soon." Below this, there are buttons for "Refresh Status", "Stop", "Destroy", and "Start".

Host Information:

Host	i-0c8f150416d867bf0.sandbox.awhap.helv.cloud
IP address	10.177.130.132
Status	running
Size	t3.medium
Expiration	28/02/2025, 17:30
Release	stable
Password	*****

Tools: A row of buttons for "Jupyter", "RStudio", and "Code Server".

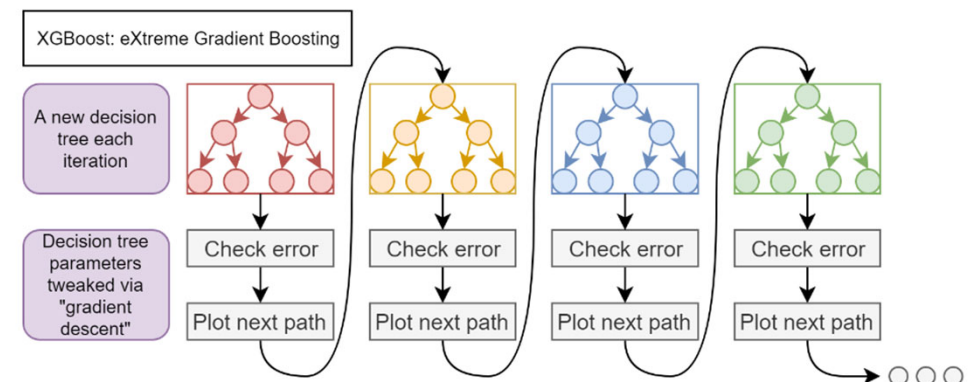
What's new? Version 31.01.2025. A welcome message states: "Welcome to the new version of the HAP Landing Page! For a better understanding of its features, the documentation can be found here: [HAP Landing Page](#) [Documentation](#)".

Latest updates:

- **New UI:** This is a new version of HAP, with an improved interface.
- **More features:** We have the patch notes section and various user experience improvements.

Machine Learning Modell – XGBoost - Generell

- XGBoost nutzt eine Reihe von **Entscheidungsbäumen**, die schrittweise trainiert werden, um Vorhersagen zu verbessern.
- Die Bäume werden so angepasst, dass die Fehler der vorherigen Bäume minimiert werden.
- Das Modell lernt, wie es die Daten am besten aufteilen kann, um die gewünschten Ergebnisse zu erzielen, und kombiniert dann die Vorhersage der einzelnen Bäume, um ein endgültiges Ergebnis zu liefern.
- Es ist in der Lage, *komplexe Muster in den Daten* zu erkennen.
- Vereinfachte Darstellung:



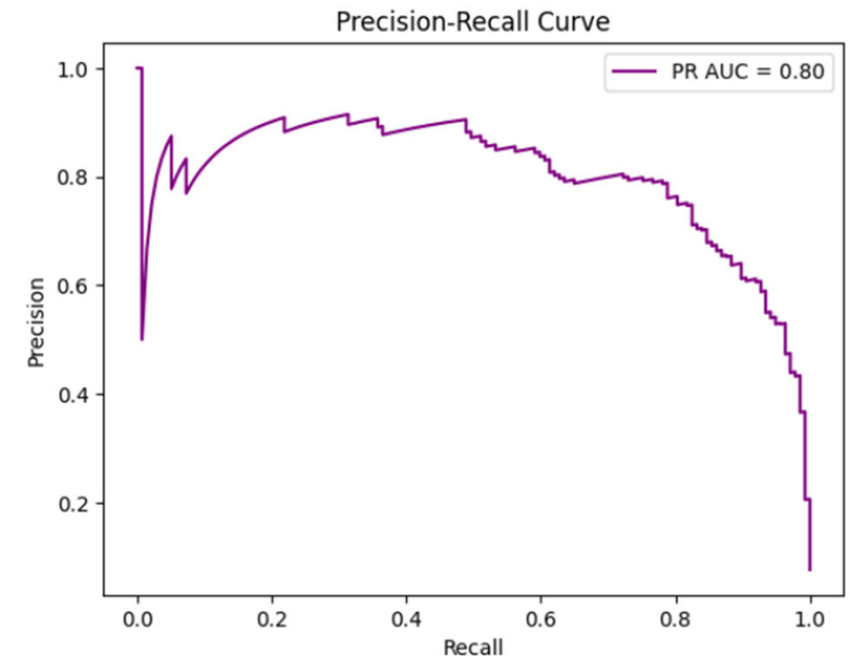
Quelle: http://www.mysmu.edu/faculty/jwwang/post/programming-with-data-fall-2021/SeminarNotes/Session_9s.html#72

Performance.

The goal is to achieve a precision-recall curve that rises as high and as far to the right as possible. Ideally, the curve should approach the top right corner, where both precision and recall are close to 1.

Caution: If values are too close to 1, it could indicate overfitting.→ It's essential to carefully evaluate model parameters and validation metrics.

Evaluation Metrics	What is being measured?
Accuracy	→ Proportion of correctly predicted cases ⚠️ Less meaningful with imbalanced datasets
Precision	→ Share of predicted leavers who actually left ✅ Good if >70%
Recall (Sensitivity)	→ Share of actual leavers correctly identified ✅ Acceptable if >60–70%
F1-Score	→ Balances precision and recall→ Useful for imbalanced data
AUC-ROC	→ Measures ability to distinguish between classes ✅ Target: 0.85–0.9
Log-Loss	→ Penalises wrong predictions with high confidence ✅ Ideal: close to 0, acceptable <0.2.



Further Info.

In parallel with the implementation of the machine learning model, I wrote my master's thesis on this topic as part of my Data Science studies. For those who would like to explore the subject in more depth, an anonymous version of the thesis can be found at:

<https://www.ffhs.ch/de/studium/thesen-openaccess/departement-informatik/mas-data-science>

