Data Science Projekte, und zwar richtig!

Azenes Data Science Blog

Zug, 4. Februar 2023 | Version 1 – final

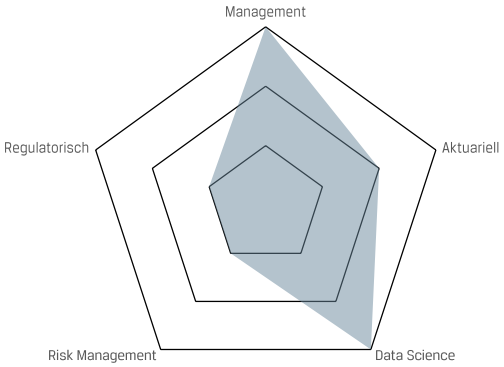
Azenes AG
Aegeristrasse 7 | 6300 Zug
+41 41 726 89 00
info@azenes.ch | www.azenes.ch

Management Summary

Data Science Projekte ohne strukturierte Herangehensweise führen zu Ineffizienzen, Mittelverschwendungen und, im schlimmsten Fall, falschen Geschäftsentscheidungen.

Cross-Industry Standard Process für Data Mining (CRISP-DM) ist eine robuste und bewährte Methodologie, um Data Science Projekte zu planen, organisieren und implementieren.

Azenes Rating für diesen Artikel



Rating		
Komplexität	tief	hoch
Zeithorizont	kurzfristig	langfristig
Impact	finanziell	operationell

Die Artikel von Azenes werden folgendermassen bewertet: Im oberen Teil wird die Relevanz für verschiedene Bereiche aufgezeigt, im unteren Teil wird die Implementation bewertet.

1 Problemspezifikation

Lösungen mit Data Science zu erarbeiten und im Geschäftsalltag einzuführen, ist ein langer Prozess. Es sind viele Schritte bis zum Ziel und bei jedem Schritt kann einiges schiefgehen. Zu oft ist das Motto eines Data Science Projekts «wir haben Daten, schauen wir, was wir daraus machen können»: Oft ist die Qualität der Daten nicht bekannt oder Zieldefinitionen werden ungenügend auf die Schwierigkeiten abgestimmt.

Das Ziel eines Data Science Projekts besteht darin, Prozesse zu verbessern und wertvolles Wissen aus Daten zu extrahieren, um dann richtige Entscheidungen für die Geschäftsstrategie fällen zu können. Dafür ist ein strukturierter Leitfaden, der sich in Theorie und Praxis bewährt hat, notwendig.

2 Lösung

Ein robuster Leitfaden ist der Cross-Industry Standard Process für Data Mining (CRISP-DM). CRISP-DM beschreibt einen iterativen Prozess, der branchenübergreifend genutzt wird, um einen Lebenszyklus eines Data Science Projekt zu beschreiben. Der Prozess teilt sich in sechs Phasen auf. Um die Resultate zu verbessern, können einzelne Phasen mehrfach zur Anwendung kommen. Im Folgenden stellen wir die einzelnen Phasen des CRISP-DM vor und *illustrieren die Phasen jeweils mit einem kleinen Praxisbeispiel, gekennzeichnet mit kursiver Schrift.*

Cross-Industry Standard Process für Data Mining (CRISP-DM)

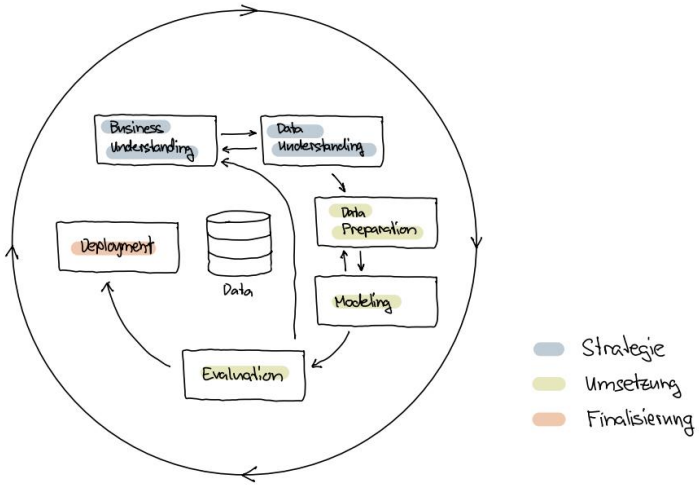


Abbildung 1: Eigene Darstellung nach Shearer (2000)

2.1 Strategie

Business Understanding

Wir beginnen den Prozess mit dem Geschäftsverständnis (Business Understanding). In dieser Phase werden konkrete Fragen diskutiert und Ziele definiert. Wichtig ist, dass die Geschäftsperspektive eingenommen wird und die Anforderungen aus dieser Sichtweise heraus formuliert werden. Output dieser Phase ist eine Problemdefinition und ein vorläufiger Projektplan inklusive der angedachten Verfahren und Ressourcen, um die Probleme anzugehen. Erfolgskriterien, um das Projektgelingen bewerten zu können, werden ebenfalls festgelegt.

Ein Versicherungsunternehmen möchte Data Mining einsetzen, um die Rentabilität von Versicherungsprodukten im Erwerbsersatz zu verbessern. Die Geschäftsleitung tauscht sich in einer ersten Sitzung aus, um die Probleme zu definieren und Soll-Zustände und deren Messbarkeit zu skizzieren, mögliche zu involvierende Abteilungen festzulegen, Verantwortlichkeiten zu verteilen und einen Zeitrahmen und das Budget abzustecken. Das Ganze wird in einem Projektplan festgehalten.

Es wird eine externe Beratungsfirma engagiert, die das Projekt organisiert und als Schnittstelle zu allen involvierten Parteien funktioniert. Die Beraterin führt Gespräche mit den Verantwortlichen; auch mit externen Dienstleistern, die beispielsweise für das Data Warehousing verantwortlich sind. Über mehrere Sitzungen wird ein Ist-Zustand skizziert, der die für das Projekt wichtigen Parteien, deren Relevanz und deren Verbindung zueinander zeigt. Der Projektplan wird im Austausch mit den Abteilungen verfeinert und bei Bedarf angepasst. Es stellt sich heraus, dass der angedachte Zeitplan nicht mit den verfügbaren Ressourcen der Tarifierungsabteilung vereinbar ist. Der Zeitplan und die Einbindung der Tarifierungsabteilung wird überarbeitet, um den neuen Informationen Rechnung zu tragen.

Projektplan

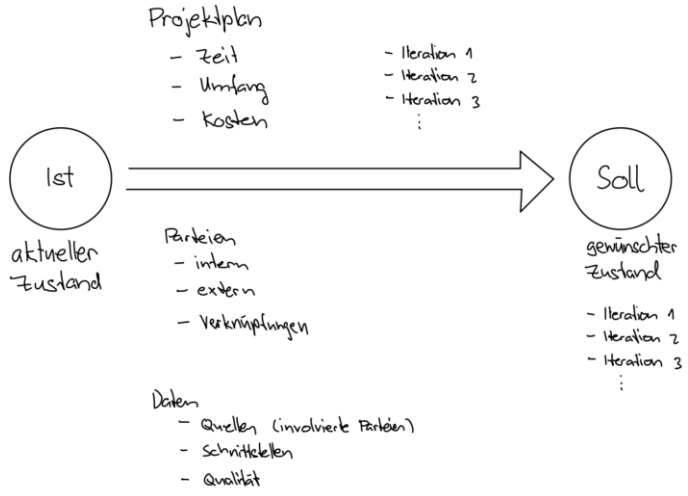


Abbildung 2: Erste Version des Projektplans.

Data Understanding

In der Phase des Datenverständnisses (Data Understanding) macht sich die Analystin mit den Daten vertraut. Dabei werden zum einen grundsätzliche Fragen gestellt: Welche Daten stehen zur Verfügung? Welche Charakteristiken wurden erhoben? Zum anderen überprüft die Analystin die Qualität und Verlässlichkeit der Daten. Die Geschäftsverständnisphase hängt eng mit dem Datenverständnis zusammen und die beiden stehen in einer Wechselwirkung zueinander: Weisen die Daten Eigenschaften auf, die in der vorläufigen Projektplanung nicht berücksichtigt worden sind, müssen Problemdefinition und Planung angepasst werden. Ein tieferes Verständnis sowohl für die Daten als auch für das Geschäft ermöglicht es der Analystin, präzise Methoden zu definieren, um die in der Planung definierten Informationen zu erlangen.

Die externe Beraterin setzt sich mit der IT-Abteilung und der externen Datenmanagementfirma zusammen und erstellt ein Mengengerüst zu den Daten. Dabei stellt die Beraterin fest, dass nur Daten der letzten drei Jahre über die externe Datenmanagerin verfügbar sind. Weitere sieben Jahre wären durch ein Altsystem verfügbar, müssen aber mit einem Mehraufwand aufbereitet werden. Ausserdem stellt die Beraterin fest, dass gewisse Attribute, welche in der Planung vorausgesetzt wurden, gar nicht erhoben wurden. Dafür sind Ortsangaben und Tätigkeitsinformation zu den Versicherten verfügbar, um die Daten sinnvoll mit externen Datenquellen, etwa dem Bundesamt für Statistik, zu kombinieren.

Der externen Beraterin stehen Daten zum Versicherungsportfolio, zu der Schadenabwicklung und den Rückstellungen, zum verbundenen Geschäft sowie Makler- und Schadenregulierungsinformationen zu Verfügung. Sie startet ihre Analyse mit einer Übersicht

über das Umfeld. Das bedeutet, die Beraterin betrachten den Markt des problematischen Geschäfts, vergleicht die Kundin mit anderen Marktteilnehmern und bespricht die Potenziale im Markt. Im Rahmen dieser Analyse wird bestätigt, dass die regulatorischen Rahmenbedingungen einen wichtigen Faktor im analysierten Geschäftsfeld spielen.

Rentabilitätspotential

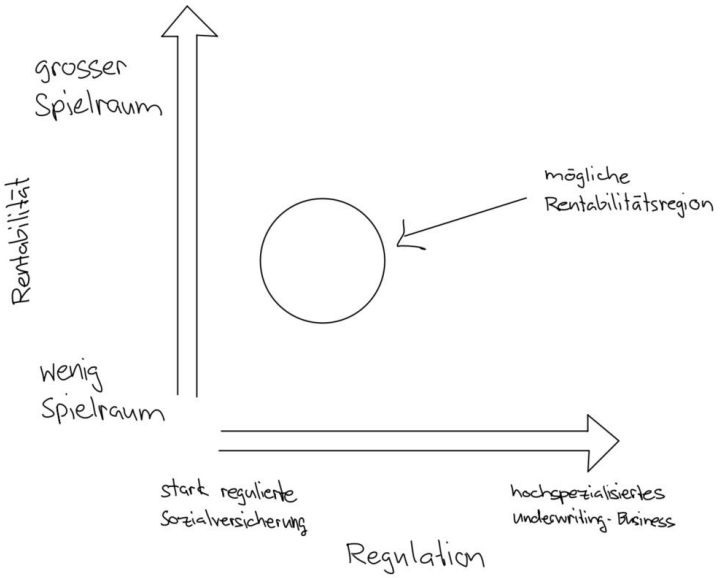


Abbildung 3: Innerhalb der Geschäfts- und Datenverständnisphase ermittelte Renditepotentiale.

Anschliessend werden die firmeneigenen Daten mit explorativen Methoden aufbereitet. Die Beraterin stellt fest, dass die involvierten Makler die Rentabilität in verschiedener Hinsicht beeinflussen. Eine Dimension, die bis anhin nicht vollumfänglich berücksichtigt wurde.

Die Beraterin erstellt einen Bericht und die Projektplanung wird um die neuen Erkenntnisse adjustiert.

2.2 Umsetzung[1]

Data Preparation

In der Phase der Datenvorbereitung (Data Preparation) erstellt die Analystin den in den Berechnungen verwendeten Datensatz. Dabei sind Codierungen/Transformationen von den in den vorangegangenen Phasen definierten Methoden abhängig.

[1] Aufgrund einer Verschwiegenheitserklärung verzichten wir im Umsetzungsteil auf das praktische Anwendungsbeispiel.

Qualitätsmängel in den Daten werden in dieser Phase mit geeigneten Verfahren bereinigt. Dazu werden verschiedene Datensätze kombiniert.

Modelling

In der Modellierungsphase (Modelling) wird mit den geeigneten Methoden versucht, aus den Daten Informationen zu ziehen. Die Art der Methoden können grundsätzlich in zwei Kategorien eingeteilt werden: vorhersagende und beschreibende Modelle. Für beide Kategorien gibt es eine Vielzahl an möglichen Methoden. Ein wichtiger Aspekt dieser Phase ist, die Stabilität der angewendeten Methoden zu überprüfen. Dabei variiert die Analystin meist Ausprägungen der Methode, um zu kontrollieren, ob die Resultate für jede Ausprägung die gleichen sind. Ebenfalls werden Kontrollmethoden benutzt, um sicherzustellen, dass die angewendeten Methoden geeignet sind, um die relevanten Informationen zu generieren. Die Resultate müssen sorgfältig dokumentiert werden, damit die einzelnen Schritte im Modellierungsprozess nachverfolgbar sind und Doppelspurigkeit vermieden werden können.

Dieser Schritt ist vielfach zentral in einem Data Science Projekt. Die Modellierungsphase muss jedoch keineswegs die aufwändigste Phase sein; falls die vorangegangenen Phasen genügend sorgfältig bearbeitet wurden.

Evaluation

In der Evaluationsphase werden die aus der Modellierungsphase gewonnen Informationen auf das Geschäftsproblem angewendet und die zuvor definierten Fragen werden beantwortet. Liefern die Verfahren zu wenige Informationen, um die Projektziele zu erreichen, muss der ganze Prozess oder Teilphasen iteriert werden. Dabei ist eine neue Iteration vielfach nicht unsorgfältiger Bearbeitung von vorherigen Phasen geschuldet. Vielmehr ist die iterative Struktur dieses Prozesses tief in Data Science Projekten verwurzelt. Nur durch die wiederholte Durchführung einer Ist-Zustandsanalyse, die Formulierung von Zielen sowie der in diesem Kapitel besprochene Umsetzung, können Erkenntnisse erlernt werden, die bessere Fragen und Zielsetzungen erlauben. Eine sorgfältige Dokumentation einer jeder Iteration hilft, das Projekt so effektiv und erfolgreich wie möglich zu gestalten.

2.3 Finalisierung

Deployment

Der beschriebene Prozess hat sich nun über mehrere Iterationen entwickelt. Gemeinsam haben sich die einzelnen Phasen ergänzt und die Entscheidungsträgerin ist mit der Problemspezifikation und mit dem aus dem Umsetzungsteil erlernten Wissen zufrieden. In der finalen Phase geht es darum, das gesammelte Wissen in die Geschäftsstrategie und schlussendlich in den Geschäftsalltag einzubinden. Diese Phase trägt den Titel Bereitstellungsphase (Deployment).

Der Aufwand der Bereitstellungsphase darf nicht unterschätzt werden. Vielfach bedarf es schwieriger Anpassung interner Geschäftsprozesse, um den potenziellen Mehrwert eines Data Science Projekts voll auszunutzen. Der zeitliche Aspekt, bis eine

strategische Ausrichtung formuliert ist und sich die Anpassungen im Geschäftsalltag bemerkbar machen, sollte ebenfalls nicht unterschätzt werden. All diese Punkte werden in der Bereitstellungsphase besprochen und die dazugehörige Strategie wird entwickelt.

Ebenfalls muss ein sorgfältiges Monitoring/Controlling Konzept ausgearbeitet werden, um eine reflektierte Implementierung voranzutreiben.

Die Beraterin formuliert klare Handlungsempfehlungen. Aufgrund der Ergebnisse wird ein Dashboard entwickelt, um die Rentabilität auf Stufe der Makler zu kontrollieren. Wichtige Rentabilitätskennzahlen, welche im Rahmen der Studie ermittelt wurden, werden im Dashboard übersichtlich dargestellt und das Versicherungsunternehmen hat immer die volle Kontrolle über ihr Versicherungsportfolio. Ein Monitoring/Controlling der Strategieänderung und der Rentabilität wird in wöchentlichen Sitzungen durchgeführt. Im Monitoring wird jeweils der Stand der Anpassungen überprüft, der durch die Strategieänderung erwirtschaftete Mehrwert wird gemessen und das Feedback der Mitarbeitenden wird eingeholt. Im Controlling bespricht die Geschäftsleitung die durch das Monitoring erhobene Daten und greift gegebenenfalls kontrollierend ein.

Monitoring-Questionnaire

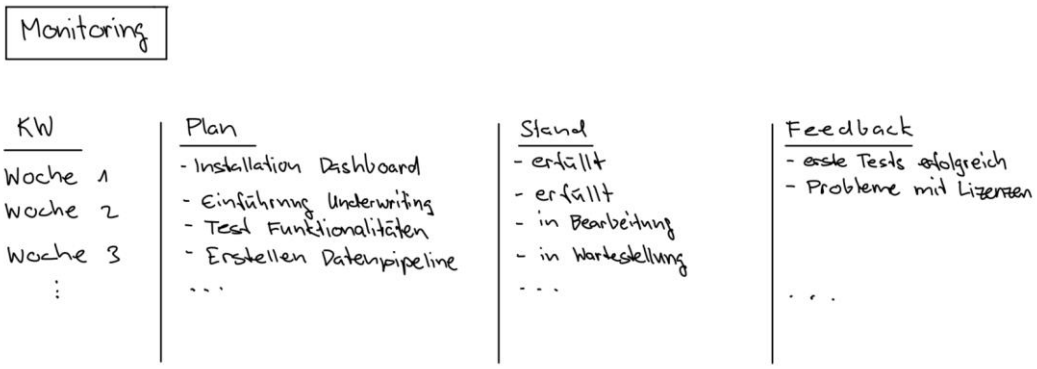


Abbildung 4: Einfacher Monitoring-Questionnaire, um die Strategieanpassung wöchentlich zu verfolgen.

Die Beraterin plant die laufende Wartung oder Aktualisierung. Sie richtet ein System für die regelmäßige Nachschulung des Modells ein, sobald neue Daten verfügbar sind, oder um Rückmeldungen der Benutzer einzubeziehen.

3 Alternative Ansätze

Es gibt verschiedene Alternativen zu CRISP-DM. Wir stellen im Folgenden eine Auswahl der wichtigsten Alternativen vor:

- Eine weiterentwickelte Variante des CRISP-DM ist **ASUM-DM (Analytics Solutions Unified Method for Data Mining)**. Im Vergleich zu CRISP-DM eignet sich ASUM-DM besser in einem agilen Umfeld. Der Fokus wird auf eine höhere Einbindung der Interessengruppen und deren Feedback gelegt und die einzelnen Schritte werden in kleinere Aufgaben unterteilt. Das iterative Element ist noch wichtiger als bei CRISP-DM.
- Eine vom SAS Institute entwickelte Methodik heisst **SEMMA (Sample, Explore, Modify, Model, and Assess)** und legt vergleichsweise mehr Wert auf spezifische Elemente sowie Aktivitäten in einem Data Science Projekt. Konkret werden vor allem die Phasen Datenverständnis und -vorbereitung, Modellierung und Evaluation detaillierter besprochen. CRISP-DM hingegen bietet einen generellen Rahmen für Data Science Projekte.
- **KDD (Knowledge Discovery in Databases)** ist ein Ansatz, der sich ebenfalls mehr auf die Phasen Datenverständnis bis Evaluation im CRISP-DM konzentriert; die Methodologie verbindet die Disziplinen Statistik, Machine Learning, Datenbanken und Visualisierungstheorie, um Wissen aufzubereiten.
- Als letzte Alternative wäre noch Google's entwickelte Methodik **OSEMN (Obtain, Scrub, Explore, Model, and iNterpret)** zu erwähnen. Gleichsam liefert OSEMN eine detaillierte Herangehensweise für die Phasen Datenverständnis bis Evaluation von CRISP-DM.

4 Fazit

CRISP-DM ist das am weitesten verbreite Modell für eine strukturierte Herangehensweise an Data Science Projekte. Dabei ist das Modell branchenübergreifend und kann auf mannigfaltige Probleme angewendet werden. Eine offensichtliche Schwäche von CRISP-DM ist, dass es nur den ganzheitlichen Prozess betrachtet, das Projektmanagement an sich ist mit CRISP-DM nicht direkt möglich.