Best Practice Data-Science-Lifecycle

Data-Science-Projekte erfolgreich durchführen

Anhand des Data-Science-Lifecycles und eines Beispiels zur Entwicklung eines Vorhersagemodells zur Gewichtsveränderung bei Milchkühen nach der Kalbung, zeigen wir Ihnen in dieser Publikation, wie Data-Science-Projekte erfolgreich durchgeführt werden können.

Sie fragen sich nun bestimmt zu Recht, wie Data Science mit dem Bereich der Milchproduktion zusammenhängt. Das ist ganz einfach erklärt. Denn Milchviehbetriebe haben im Prinzip dasselbe Ziel wie Dienstleistungsunternehmen: sie möchten effektiv, effizient und rentabel sein. Dementsprechend gibt es auch ähnliche Problemstellungen, die im Bereich Data Science mit Daten gelöst werden können. Im Folgenden gehen wir hierauf näher ein.

BUSINESS.
DATA.
TECHNOLOGY.

CINTELLIC
CONSULTING GROUP

Ähnlichkeiten zwischen Unternehmen und Milchviehbetrieben

Wenn man Unternehmen im Allgemeinen und Milchviehbetriebe im Speziellen vergleicht, identifiziert man schnell übergeordnete und gemeinsame Ziele, nämlich die der Umsatz- und Ertragssteigerung bzw. der Kostensenkung. Unternehmen erreichen diese Ziele, indem Sie Produkte oder eben Dienstleistungen an Kunden verkaufen und die Kundenbedürfnisse befriedigen. Im Gegensatz dazu verkaufen Milchviehbetriebe ihr Rohprodukt – in dem Fall die Milch – überwiegend zur Weiterverarbeitung an Molkereien. Um dieses Rohprodukt zu erhalten, müssen sie die Bedürfnisse der Milchkühe befriedigen.

Analysemethoden zum Erreichen der Ziele

Im Dienstleistungssektor sind die Analysemethoden um CRM, CEM und Customer Analytics weitestgehend bekannt. Man versucht also über die Daten, welche ein Kunde hinterlässt, möglichst viele Informationen zu bekommen. Der Milchviehbetrieb hat hier andere Analysemethoden, die aber auch auf Daten wie zum Beispiel zum Tier selbst, zu dessen Leistung und Gesundheit, basieren. Solche Daten werden dann im so genannten Herdenmanagement hinterlegt.

	Unternehmen	Milchviehbetrieb
Ziel	• Umsatzsteigerung • Ertragssteigerung • Kostensenkung	• Umsatzsteigerung • Ertragssteigerung • Kostensenkung
Zielerreichung	• Verkauf von Produkten und Dienstleistungen an Kunden • Kundenbedürfnisse befriedigen	• Verkauf von Rohprodukten an Weiterverarbeitung (Molkereien) • Bedürfnisse der Milchkühe befriedigen

Analysemethoden	• CRM
	• CEM
	  • Customer Journey
	  • Customer Loyalty
	  • NBA
	• Customer Analytics
	  • Churn Prevention
	  • NBO
		• Herdenmanagement
		• CowEM
		  • Biologischer Zyklus
		  • Umwelt
		  • Routine
		• Cow Analytics
		  • Gesundheit
		  • Hohe Lebenserwartung

Data Science Lifecycle – 7 Projektphasen

In jedem Data-Science-Projekt gibt es gewisse Problemstellungen, welche anhand von Daten gelöst werden wollen. Die meisten solcher Projekte haben einen ähnlichen Arbeitsablauf, welcher genutzt werden kann, um das jeweilige Projekt in Phasen aufzuteilen, die es typischerweise durchläuft. Hierbei steht an oberster Stelle das übergeordnete Ziel, die jeweilige Problemstellung zu lösen.
Grundsätzlich wird der Data Science Lifecycle in sieben Phasen unterteilt, die wir nachfolgend näher erläutern:

1. Business Understanding

Der Ausgangspunkt des Data Science Lifecycles ist die erste Phase – das Business Understanding. Denn es ist entscheidend, dass die Problemstellung verstanden wird und die richtigen Fragen an Stakeholder gestellt werden, um letztendlich die richtigen Datensätze zu erhalten. Zudem muss auch überprüft werden, dass die Daten korrekt sind und aussagekräftige Erkenntnisse aus den Daten geschöpft werden können. Mit einem guten Business Understanding kann dann das Ziel des Projektes sowie die Variablen, die vorhergesagt werden sollen, identifiziert und definiert werden.

2. Data Mining

Aufbauend auf dem Business Understanding folgt Schritt zwei – das Data Mining. In dieser Phase werden die benötigten Daten gesammelt. Entweder liegen diese bereits vor und sie können aus bestehenden Datenquellen abgefragt werden, oder die Daten müssen zunächst erfasst werden.

Tipp/Hinweis: *Sollten die Daten erst erfasst werden müssen, sollte unbedingt die nötige Zeit hierfür eingeplant werden. Manchmal kann es Wochen und Monate dauern, bis alle relevanten Daten zusammengetragen sind!*

3. Data Cleaning

Nachdem alle relevanten Daten vorliegen, folgt Schritt drei – das Data Cleaning. Hier werden die Daten genauer unter die Lupe genommen, bereinigt und für weitere Analysen vorbereitet. Das Bereinigen bedeutet im Wesentlichen das Entfernen von Diskrepanzen also beispielsweise fehlende, falsche oder nicht benötigte Werte herauszunehmen. Im Großen und Ganzen wird hierbei also sortiert und strukturiert.

4. Data Exploration

Anschließend folgt Schritt vier – die Data Exploration. Hierbei werden häufig die Datenstatistiken wie Mittel-, Medial- und Extremwerte berechnet, um die Verteilung der Daten besser nachzuvollziehen. In diesem Schritt können bereits erste Visualisierungen vorgenommen werden. Beispielsweise Diagramme wie Histogramme, Punkt- oder Liniendiagramme. Letzteres hilft dabei, die Daten besser zu verstehen, aber auch, um viel-

leicht versteckte Muster zu erkennen, welche eine zusätzliche Information geben. Zum Beispiel Saisonalitäten, die vorher nicht bekannt waren.

Wissenswert: *Diese ersten Schritte (1 bis 4) nehmen 70 bis 90 Prozent der Projektzeit in Anspruch. Und das ist sehr wichtig, denn hier wird geschaut, ob die Daten, die vorhanden sind, tatsächlich repräsentativ sind und qualitativ so hochwertig, dass das vorangestellte Ziel erreicht werden kann.*

5. Feature Engineering

Mit den aus den ersten Schritten gewonnenen Erkenntnissen wird dann in Schritt fünf gestartet – Feature Engineering. Hier werden die Rohdaten so aufbereitet, dass sie sofort in Machine Learning Algorithmen verarbeitet werden können. Häufige Faktoren hierbei sind beispielsweise das Transformieren von Variablen, die Berechnung von Interaktionen oder die Erstellung von „Dummy-Variablen". Dieser Schritt hat das Ziel, gegenüber den Rohdaten einen Mehrwert zu schaffen. Die Herausforderung hierbei ist jedoch, dass keine unnötigen Daten verarbeitet werden, aber wichtige Daten ebenso nicht verloren gehen.

6. Predictive Modeling

Sind alle Rohdaten verarbeitet, folgt Schritt sechs – das Predictive Modeling. Hier werden die Machine Learning Algorithmen angewendet, um ein passendes Datenmodell zu erstellen. Basierend auf dem vorliegenden Problem, die Daten, welche gesammelt wurden, sowie dem definierten Ziel, gibt es verschiedene Arten von Algorithmen. Beispielsweise logistische und lineare Regressionen, Clustering-Methoden etc. Nachdem das Datenmodell in dieser Phase aufgesetzt wurde, wird die Performance evaluiert. Dafür wird das Modell auf bisher unbekannte Testdaten angewendet und über entsprechende Metriken wird dann die Performance sowie die Genauigkeit des Modells bestimmt, angepasst und verbessert.

7. Data Visualization

Ist ein Datenmodell aufgebaut, folgt der letzte Schritt – die Datenvisualisierung. Hier werden die Ergebnisse des Projektes mithilfe von Grafiken aber auch Storytelling dargestellt und den Stakeholdern vorgestellt.

Wichtig: *Die Stakeholder haben nicht immer einen technischen Hintergrund. Es sollte versucht werden, die Ergebnisse und Visualisierungen möglichst klar, einfach und nachvollziehbar darzustellen.*

Generell lässt sich sagen, dass der Data Science Lifecycle kein linearer Prozess ist. Unter Umständen müssen bestimmte Schritte iterativ ausgeführt werden, um bestmögliche Ergebnisse zu erzielen. Beispielsweise kann es vorkommen, dass in Schritt 4 festgestellt wird, dass zu wenig Daten vorhanden sind. Demnach müsste an dieser Stelle in Schritt 2 zurückgesprungen werden, um mehr Daten zu sammeln.

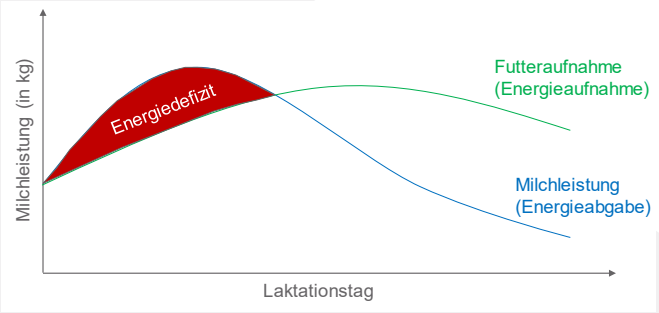
Best Practice: Erläuterung des Data Science Lifecycles anhand eines Beispiel-Projektes

Um das Vorhersagemodell zur Gewichtsveränderung bei Milchkühen nach der Kalbung aufzubauen, wurden alle sieben Schritte des Data Science Lifecycles durchgearbeitet.

1. Business Understanding

Im ersten Schritt, dem Business Understanding, wurden folgende Fakten aufgelistet:

- Generell ist es so, dass Säugetiere nach der Geburt Körperreserven mobilisieren, um die Milchproduktion aufrecht zu erhalten. Das heißt, dass in dieser Phase die Energieaufnahme kleiner als die Energieabgabe ist und somit eine negative Energiebilanz herrscht.
- Wenn aber eine übermäßige Mobilisierung von Körperreserven stattfindet, kann es für das Tier hinsichtlich Gesundheit und Fruchtbarkeit problematisch werden.



Das Ziel des Projektes war es, Milchkühe zu identifizieren, bei denen nach der Kalbung ein zu hoher Verlust am Körpergewicht zu erwarten ist. Denn wenn Säugetiere zu viel Körpergewicht verlieren, kann der Stoffwechsel entgleisen. Und es ist besser die Entgleisung des Stoffwechsels zu verhindern, als diese wieder einzuloten.
Die Herausforderung in der Praxis ist, dass eine routinemäßige und genaue Gewichtsmessung von Milchkühen schwierig ist.

Demnach wurde als übergeordnetes Ziel die Vorhersage zur Schätzung der Gewichtsveränderung gesetzt.

Wie kann dieses Ziel erreicht werden bzw. was heißt Mobilisierung von Körperreserven?
Die Mobilisierung von Körperreserven bedeutet, dass Körperfette mobilisiert werden. Daher wurde sich der Aufbau von Fetten zunächst nochmal hervorgerufen und Informationen zusammengetragen. Fette bestehen aus Triglyceriden (Glycerin und drei Fettsäuren) und sind auch ein Bestandteil der Milch. Daher kann das Fett sowie Fettsäuren in dieser gemessen werden. Doch wo kommen die Fettsäuren eigentlich her? Auf der einen Seite gibt es die „de novo"-Fettsäuren, welche im Euter synthetisiert werden. Auf der anderen Seite gibt es „preformed" -Fettsäuren, welche überwiegend aus den Körperfettreserven kommen. Darüber hinaus gibt es die „mixed"-Fettsäuren, die sowohl direkt im Euter synthetisiert werden als auch direkt aus den Körperfettreserven stammen können.

Aus diesen Erkenntnissen entstand die Idee, die Vorhersage der Gewichtsveränderung anhand des Milchfettsäureprofils zu erstellen.

2. Data Mining

Bei diesem Projekt lagen bereits Datensätze vor. Der erste Datensatz, die Milchanalysedaten, stammten aus der Milchkontrolle. Dabei werden die Milch sowie die Milchfettsäuren über Infrarotspektroskopie analysiert.
Der zweite Datensatz, die Gewichtsdaten, stammten aus einer Kooperation mit einem Melktechnik-Hersteller, wobei die Gewichte aus automatischen Melksystemen mit integriertem Wiegebogen gezogen wurden. Im Schnitt entstanden somit 2,5 Gewichtsmessungen pro Tag und pro Kuh.
Folgende beiden Datensätze haben sich daraus ergeben:

	Milchanalyse-daten	Gewichtsdaten
Einzelmessungen	197.058	28.581.762
Anzahl Milchkühe	34.870	35.787
Anzahl Herden	169	168
Messzeitraum	03/2015 – 03/2017	01/2015 – 09/2017
Messintervall/ Kuh	1x Monatlich	Ø 2,5x Täglich

3. Data Cleaning

In dieser Phase, dem Data Cleaning, wurden die Daten genauer unter die Lupe genommen und bereinigt. Bei den Milchanalysedaten wurde beispielsweise eine Entfernung von Beobachtung mit fehlenden Werten, abnormale Beobachtungen in den Milchfettsäurewerten sowie Daten der Laktationszahl größer als 3 und der Laktationstage kleiner als 5 sowie größer als 305 Tage vorgenommen.

Definition „Laktation" & „Laktationstage": *Eine Laktation bedeutet, dass die Kuh ein Kalb bekommen hat. Das heißt, sie kalbt und startet dann in die Milchproduktion also in die erste Laktation. Nachdem sie das zweite Kalb bekommen hat, startet sie in die zweite Laktation etc. (Im Projekt wurde die Laktationszahl auf 3 begrenzt, weil hierzu die meisten Daten vorlagen.) Die Laktationstage sind die „Tage in Milch pro Laktation".*

Im Hinblick auf die Gewichtsdaten wurde eine Entfernung von Beobachtungen mit abnormalen Beobachtungen in den Gewichten sowie ebenfalls der Laktationszahl größer als 3 und der Laktationstage kleiner als 5 und größer als 305 Tage vorgenommen.

4. Data Exploration

Hier wurde geschaut, ob die Daten repräsentativ sind, ob sie das darstellen, was zunächst im Business Understanding definiert wurde. Außerdem wurde geprüft, ob vielleicht sogar neue, unbekannte Muster der Daten vorhanden sind. Die Datenstatistik wurde ausgewertet, indem der Mittelwert, die Standardabweichung sowie Minimum und Maximum errechnet wurden. Letzteres galt beispielsweise für die Milch an sich, für den Fett- oder Proteinanteil oder für die Fettsäuregruppen und Gewichte.
Die ersten Daten – beispielsweise zu den Gewichten – wurden in Diagramme übertragen. Hier kam unter anderem heraus, dass die Kühe den Tiefstand des Gewichtsverlusts ca. am 30. Laktationstag erreichen und anschließend 60 bis 100 Tage benötigen, um das Ursprungsgewicht wieder zu erreichen.

5. Feature Engineering

Die Basis des Projektes, nämlich mit dem Modell die Körpergewichtsveränderung hervorzusagen, war an dieser Stelle noch nicht abgeschlossen bzw. erreicht. Denn die Daten boten nur die jeweiligen Gewichte der Kühe und noch keinen Aufschluss über dessen Veränderungen. Deshalb mussten letztere errechnet werden. Dies geschah, indem das Körpergewicht von Tag X minus das Körpergewichtes des Vortages errechnet wurde und anschließend nochmal durch das Gewicht des Vortages genommen wurde. Somit entstand eine relative tägliche Körpergewichtsveränderung in Gramm pro Kilo Körpergewicht. Der Grund für diese Berechnung war, dass eine standardisierte Metrik über alle Kühe hinweg erstellt werden sollte.
Die Ergebnisse wurden erneut in Diagramme gegossen. Anschließend wurden die beiden Datensätze miteinander verschmolzen (englisch: Datamerge). Dies geschah anhand der Tiernummer sowie anhand des jeweiligen Datums. Die Datensätze zur Milchanalyse lagen allerdings monatlich vor, während die Daten zur Gewichtsveränderung täglich vorlagen.
Die zentrale Frage lautete also: Welche und wie viele Körpergewichtsveränderungen werden durch das Milchfettsäureprofil am Tag der Milchanalyse eigentlich repräsentiert? Wie schnell ändert sich das Milchfettsäureprofil, wenn eine veränderte Energieversorgung vorliegt? Hierfür lagen noch keine Studien vor. Also wurde mit einer Professorin gesprochen, die ihr Feedback hierzu gegeben hat, auf dessen Basis Annahmen getroffen werden konnten.
Damit wurde ein finaler Datensatz entworfen:

	Milchanalyse-daten	Gewichts-daten	Modelldaten
Einzelmessungen	197.058	28.581.762	19.138
Anzahl Milchkühe	34.870	35.787	16.847
Anzahl Herden	169	168	165
Messzeitraum	03/2015 – 03/2017	01/2015 – 09/2017	03/2015 – 03/2017
Messintervall /Kuh	1x Monatlich	Ø 2,5x Täglich	1x Monatlich

6. Predictive Modelling

In der sechsten Phase, dem Predictive Modelling, wurde dann eine Variablenselektion anhand einer Principle Component Analysis (PCA) durchgeführt. Das Ziel hierbei war es, die wichtigsten Variablen für das Datenmodell zu erhalten. Im nächs-

ten Schritt wurden die Variablen normalisiert, denn sie lagen in unterschiedlichen Einheiten wie beispielsweise „Gramm“ und „Kilogramm“ vor. Im dritten Schritt wurde sich dann für das Vorhersagemodell des „Random Forest Algorithm“ entschieden. Denn dieser kann mit einer Vielzahl von linearen und nicht-linearen Beziehungen zwischen Variablen und einem hochdimensionalen Datensatz mit entsprechender Komplexität umgehen.

Exkurs: *Was ist der Random Forest Algorithm? Dieser Algorithmus basiert auf mehreren Entscheidungsbäumen. Dafür wird der gesamte Datensatz in mehrere kleine Datensätze aufgeteilt. An solche kleinen Stichproben werden Entscheidungsbäume angepasst. Dabei verwenden diese Entscheidungsbäume auch nicht immer alle Variablen, sondern eine Teilanzahl. Von dieser Teilanzahl wird schlussendlich ein finales Ergebnis erstellt.*

Im Projekt wurden die Datensätze nun also gesplittet und in ein Trainings- sowie einen Testdatensatz aufgeteilt. Hierbei wurden 80% in das Trainieren des Modells verwendet und 20% der Daten als Testdatensatz zur Seite gelegt. Es konnte allerdings nicht alles dem Zufall überlassen werden, denn es gab beim Projekt einen Einzeltier- sowie einen Herdeneffekt. Hier gibt es beispielweise Unterschiede in der Fütterung oder Haltung, was sich natürlich auf das Gewicht der Milchkühe auswirkt. Daher erfolgte die Datensplittung zum einen nach den Einzeltieren und zum anderen nach der Herde.
Außerdem wurde der Trainingsdatensatz anhand der 10-fachen Kreuzvalidierung erneut in weitere zehn kleinere Datensätze gesplittet. Nachdem das Modell aufgebaut wurde, wurde es außerdem an sieben verschiedenen Metriken evaluiert und die Performance gemessen.

7. Data Visualization

In der letzten Phase des Data Lifecycles wurde explizit bei diesem Projekt ein wissenschaftliches Paper erstellt. Die Ergebnisse wurden demnach in speziellen Grafiken aufbereitet.
Zu den ersten drei Variablen mit dem höchsten Einfluss auf das Modell zählen die kurzkettigen Fettsäuren („de novo“), die C18:0-Fettsäuren und die einfach-ungesättigten Fettsäuren (beide „preformed“). Hierfür wurde dementsprechend eine weitere Ergebnis-Grafik erstellt. Tiere mit einer negativen Körpergewichtsveränderung haben tendenziell niedrigere Werte in kurzkettigen Fettsäuren und höhere Werte in der C18:0- und den einfach-ungesättigten Fettsäuren. Dies spiegelt die Annahme aus dem Business Understanding wider. Es werden Körperreserven mobilisiert, um zu versuchen die negative Energie auszugleichen.

In einem weiteren Schritt wurden die beobachteten Gewichtsveränderungen mit denen aus dem Random Forest Modell vorhergesagten Gewichtsveränderungen in Bezug gesetzt. Die vorhergesagten Veränderungen unterschieden sich nicht signifikant von den beobachteten, was beweist, dass das Modell generell in der Lage ist, die Gewichtsveränderungen vorherzusagen.
Das letzte Ergebnis bezieht sich auf die Messung der Perfomance des Modells anhand der zurückgelegten Testdaten. Hier wurden die einzelnen Metriken aufgelistet und nach den Stratifikationen „Einzeltier“ sowie „Herde“ unterteilt. Die Performance des Testdatensatzes verliert allerdings im Gegensatz zu den Kreuzvalidierungs-Datensätzen an Performance. Das Modell eignet sich in der modellierten Form also noch nicht unbedingt für einen praktischen Einsatz.

Wissenswert: *Man sollte sich also Gedanken machen, warum der Testdatensatz nicht so gut performt. Woran kann das liegen und was kann hierbei noch optimiert werden?*

Fazit

Der Data Science Lifecycle hilft dabei, Projekte – egal wie exotisch und speziell diese sind – in ihren einzelnen Entwicklungsphasen zu strukturieren. Dies unterstützt bei der Planung einzelner Arbeitsschritte sowie bei der zeitlichen Umsetzung des Projektes. Des Weiteren ist der Data Science Lifecycle sehr nützlich in Bezug auf die Fokussierung und dient dazu, einen übersichtlichen Fahrplan zu haben.

Falls Sie mehr aus Ihren Daten machen möchten oder Unterstützung bei einem Data Science Projekt benötigen, zögern Sie bitte nicht, uns anzusprechen. Die CINTELLIC Consulting Group verfügt über jahrelange Erfahrung in den Bereichen Data Science, Data Mining und Big Data Analytics.

von Franziska Dettmann und Natalie Niebuhr,
CINTELLIC Consulting Group

Quellenangabe:
Dettmann, F.; Warner, D.; Buitenhuis, B.; Kargo, M.; Kjeldsen, A.M.H.; Nielsen, N.H.; Lefebvre, D.M.; Santschi, D.E. Fatty Acid Profiles from Routine Milk Recording as a Decision Tool for Body Weight Change of Dairy Cows after Calving. Animals 2020, 10, 1958. https://doi.org/10.3390/ani10111958

Ansprechpartner

Dr. Jörg Reinnarth
Geschäftsführer
CINTELLIC Consulting Group
joerg.reinnarth@cintellic.com

Stephan Klöckner
Senior Manager
CINTELLIC Consulting Group
stephan.kloeckner@cintellic.com

Cintellic im Social Web

Cintellic GmbH
Remigiusstraße 16
53111 Bonn
t +49 228 92 65 18 20
info@cintellic.com
www.cintellic.com

Über CINTELLIC

Die 2010 gegründete CINTELLIC Consulting Group ist eine auf digitales Kundenmanagement spezialisierte Unternehmensberatung, die ihre Klienten vom ersten Konzept bis zur Umsetzung in der Praxis ganzheitlich begleitet. An den Standorten in Bonn, Frankfurt am Main und München arbeiten über 70 Mitarbeiterinnen und Mitarbeiter.

Zu den Klienten zählen DAX-Konzerne, führende mittelständische Unternehmen und insbesondere zahlreiche sogenannte „Hidden Champions" mit den Branchenschwerpunkten Banken und Versicherungen, Telekommunikation, IT, Medien, Unterhaltung, Handel, E-Commerce, Versorger und Logistik.

www.cintellic.com

#jointheteam

CINTELLIC befindet sich auf Wachstumskurs. Vielleicht mit Ihnen? Jetzt Stellenanzeigen entdecken und bewerben!
https://www.cintellic.com/stellenangebote/

CINTELLIC
CONSULTING GROUP