Making it explicit

Prognosefähigkeit von KI-Methoden im Vergleich

Julius Heitmann
Christoph Sonn

November 2019
Blogpost

Key Facts

- Um aussagefähige und hochwertige Datenanalysen zu implementieren, ist die Datenvorbereitung von überragender Bedeutung – sie ist jedoch zeitraubend und aufwändig
- Ein Teil der Datenvorbereitung stellt das Feature Engineering dar – ein optionaler Vorgang, der dem Modell implizit inherente Informationen explizit zugänglich machen soll
- Für das Feature Engineering ist der Einsatz von Domänen- bzw. Expertenwissen notwendig, um die Infomationen explizit zur Verfügung zu stellen
- Feature Engineering wird im Rahmen dieses Blogposts vor dem Hintergrund verschiedener Modelle anhand der Informationsextraktion aus zetlichen Daten untersucht
- Die Berücksichtigung von Domänenwissen für das Feature Engineering scheint auch unter Einbeziehung des damit verbundenen Aufwands sinnvoll und essentiell zur Erstellung von genaueren Analysen. Der Erfolg des Feature Engineerings ist allerdings vom verwendeten Algorithmus abhängig

Einführung

In Zeiten steigender Effizienz von Geschäftsprozessen infolge von Digitalisierung von Informationsflüssen stellt künstliche Intelligenz eine interessante Möglichkeit dar, weitere Effizienzausbeute zu realisieren. So besteht neben der medialen Aufmerksamkeit, die Artificial Intelligence (AI) umgibt, auch aus Unternehmensperspektive hohes Interesse am Einsatz von AI. Ein wichtiges Einsatzfeld dabei ist die Automatisierung von Entscheidungen. Um dieses Ziel verwirklichen zu können, müssen im Rahmen eines Datenanalyseprozesses Daten erhoben und vorbereitet werden, um im Anschluss das Training eines entsprechenden Modells zu ermöglichen. Der Prozess der Datenvorbereitung ist zeitraubend und aufwändig. Gemäß einer Studie des Magazins Forbes verbringen Datenanalysten rund 80% der Zeit mit Datenvorbereitung, worin die Schritte „Data Cleaning“, „Data Preparation“ und „Feature Engineering“ enthalten sind. Während die ersten beiden Elemente dafür benötigt werden, die Analyse erst zu ermöglichen, verbirgt sich hinter „Feature Engineering“ ein optionaler Vorgang, der dem Modell inherente Informationen zugänglich machen soll. Im Rahmen dieses Blogposts soll untersucht werden, ob ein solcher Prozessschritt, der unter anderem auch Domänenwissen und damit Fachkräfte benötigt, den getätigten Aufwand wert ist, ob heutige Modelle möglicherweise in der Lage sind, diese Informationen selbst zu extrahieren, oder ob dieser Schritt möglicherweise sogar essentiell für ein erfolgreiches Datenanalyseprojekt ist.

Das Feature Engineering ist der Prozess von Extraktion von Informationen (Features) aus Daten, die bereits implizit enthalten sind. Für diesen Vorgang wird Domänenwissen benötigt. Dabei werden durch die Anwendung von Modellen, die auf Expertenwissen beruhen, die Features extrahiert und dem Datensatz explizit hinzugefügt. So wird die Information auch für Algorithmen zugänglich. Ein gutes Beispiel, da jeder Leser des vorliegenden Blogposts Domänenwissen auf diesem Gebiet hat, ist die Extraktion von Features aus einem Datum, welches auch im späteren Verlauf des Blogposts weiter untersucht wird. Für einen Algorithmus ist eine Reihe von Daten eine geordnete Menge von Datenpunkten. Lediglich die Information zur relativen zeitlichen Position und möglicherweise die Informationen über Tag, Monat und Jahr sind zugänglich. Für

Menschen beinhaltet ein Datum Information über beispielsweise Wochentag, Arbeitstag oder Feiertag. Mit geeigneten Funktionen kann diese Information in Form eines zusätzlichen Attributs dem Datensatz hinzugefügt werden. Auch ein Algorithmus ist nun in der Lage beispielsweise ein Muster auf Basis des Wochentages zu entdecken.

Im Rahmen dieses Blogposts soll die Wirksamkeit von Feature Engineering in Abhängigkeit verschiedener Modelle der Datenanalyse untersucht werden. Dazu untersuchte das COREAI

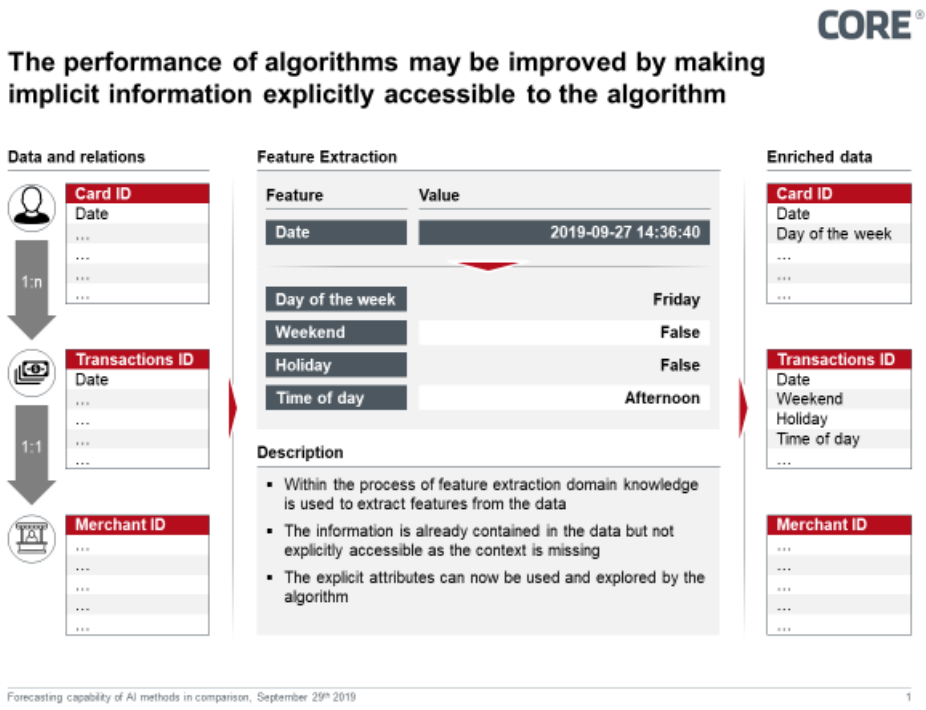

Abbildung 1: Feature Extraction Prozess in Elo Use Case

Team die Analyse und Auswertung von Kundenbedürfnissen auf Basis von Transaktionsdaten eines brasilianischen Zahlungsverkehrsanbieters. Die Thematik der Kundenbedürfnisvorhersage ist von entscheidender Bedeutung u.a. im Finanzumfeld. Der analysierte Datensatz wurde im Rahmen einer öffentlichen Ausschreibung zur Verfügung gestellt (Kaggle Challenge - Elo Merchant Category Recommendation). Die Challenge initiierte Elo, einer der größten Zahlungsverkehrsanbieter in Brasilien. Elo schloß mit einigen Händlern Partnerschaften, mit dem mittelfristigen Ziel Elo Kreditkarten-Kunden individualisierte Angebote zu unterbreiten. So wäre beispielsweise denkbar, den Karteninhabern Vorschläge für ein Restaurant, basierend auf den persönlichen Präferenzen und zusätzlich zu diesem Vorschlag noch einen Nachlass auf das Menü, unterbreiten zu können. In diesem Fall würde Elo u.a. durch anfallende Transaktionsgebühren profitieren.

Zunächst werden eine Auswahl von Algorithmen trainiert, auf Basis des Datensatzes die Konsumbereitschaft und damit ein Kundenbedürfnis zu prognostizieren. Diese trainierten Modelle werden im Anschluss vor dem Hintergrund von Feature Engineering näher betrachtet, um einen Diskurs zu Aufwand und Nutzen zu ermöglichen. Zur Prognose von Konsumbereitschaft sind verschiedene Regressionsmodelle (die zu prognostizierende Variable ist numerisch) vorstellbar.

Im Rahmen dieses Blogposts werden Analysen mit zwei aktuell sehr beliebten Modellen – Gradient Boosting und Deep Learning – vorgestellt und durchgeführt. Dazu wird zunächst ein Blick in die Daten geworfen. Art und Inhalt der Daten werden kurz erläutert. In einem zweiten Schritt wird der Vorgang des Feature Engineering kursorisch vorgestellt. Im darauf folgenden Versuch werden die Modelle Gradient Boosting sowie Deep Learning mit und ohne Feature Engineering an den Daten trainiert und validiert. Im Fazit werden die Ergebnisse reflektiert und somit ein Rückschluss auf Qualität und Eigenschaften der Modelle hinsichtlich des betrachteten Use Cases und des Feature Engineerings ermöglicht.

Daten

Der analysierte Datensatz enthielt Daten von über 2 Millionen Kunden, zu über 300 Tausend Händlern sowie zu knapp 30 Millionen Transaktionen zwischen Händlern und Kunden. Die Daten sind soweit anonymisiert, dass keine Rückschlüsse auf Personen oder konkrete Zusammenhänge gezogen werden können, weshalb am Ende dieses Blogposts keine Aussage zu speziellem Kundenverhalten gegeben werden kann. Enthalten im Datensatz ist zu jedem Kunden eine Score, die die Loyalität in Form von Kaufbereitschaft des Kunden widerspiegelt. Dieses kundenspezifische Attribut gilt es zu prognostizieren. Die Score kann stellvertretend für andere kaufbezogene Eigenschaften (z.B. Markentreue) gesehen werden.

Feature Engineering

Der Prozess der Datenvorbereitung beinhaltet neben Datenbereinigung und -modellierung, die benötigt werden, damit ein Algorithmus die Daten überhaupt verarbeiten kann, den Schritt des Feature Engineering. Hier wird unter Zuhilfenahme von Domänen-/Expertenwissen Information, die bereits implizit in den Daten enthalten ist, extrahiert und explizit dem Datensatz in Form eines zusätzlichen Attributes beigefügt. Beispiele hierfür könnten die örtliche Anordnung von Geschäftsgebieten (z.B. könnte Gebiet A örtlich zwischen B und C liegen) oder bestimmtes Domänenwissen sein (z.B. Objekt A kann nur mit Objekt B gekauft werden). Diese Informationen gingen aufgrund einer starken Anonymisierung leider verloren. Die in diesem Blogpost behandelte Form von Feature Engineering ist die Extraktion von zeitlichen Features aus Daten. Für uns Menschen direkt intuitiv, beinhalten Daten neben einer zeitlichen Abfolge auch Informationen zu beispielsweise Wochentag, Monatsbeginn oder Arbeitstag. Gerade bei Zahlungsprozessen können hier weitere Informationen gewonnen werden (z.B. kann am Sonntag auf Grund von Ladenöffnung weniger gezahlt werden). Der Datensatz, der diese Informationen in Form von zusätzlichen Attributen enthält, wird im Folgenden als angereicherter Datensatz bezeichnet.

Versuch/Analyse

An den Daten können nun die verschiedenen Modelle getestet werden, wobei Neuronale Netze/Deep Learning in verschiedenen Tiefen (2, 3 und 4 Schichen tief) und Breiten (Schichten zwischen 50 und 100 Neuronen), erstellt mit dem KERAS Framework, trainiert werden. Zum Vergleich wird ein Modell zu Gradient Boosting, einer Ensemble Technik mit Entscheidungsbäumen, die aktuell sehr beliebt in der Datenanalysten Community ist, trainiert. Das verwendete Framework ist LightGBM.

Die Modelle werden dabei auf dem angereicherten sowie dem nicht angereicherten Datensatz getestet. Somit soll untersucht werden, welches der Modelle die extrahierten Informationen besser nutzt, und somit weiteren Mehrwert erzeugen kann. Als Qualitätsmaß wird das MSE (Mean squared error – die mittlere quadratische Abweichung vom Zielwert) gewählt. Ergebnisse mit einem höheren MSE haben eine höhere Abweichung zum Erwartungswert und sind somit als schlechter zu bewerten.

Im Folgenden sind die Ergebnisse der Modellierung abgebildet. Validiert wurde mittels 5-facher Kreuzvalidierung.

Model\Analyse	***Angereicherter Datensatz***	***Nicht angereicherter Datensatz***
Gradient Boosting	3,654782841672	3,698093670456
DL 2-Schichten	3,857178126259	3,857171424530
DL 3-Schichten	3,857898095313	3,857640795361
DL 4-Schichten	3,858468548452	3,858397482541

Es ist festzuhalten, dass der Gradient Boosting Ansatz stets die besseren Ergebnisse liefert. Zudem zeigte der Ansatz auch deutliche Vorteile in Lern- und Abfragegeschwindigkeit. Des Weiteren kann man erkennen, dass die Performance des Neuronalen Netzwerkes abnimmt, je tiefer das Netz wird. Von „Deep“ Learning kann in diesem Zusammenhang (2 Schichten tief) also kaum gesprochen werden. Zuletzt kann beobachtet werden, dass der Informationsgewinn durch Anreicherung des Datensatzes vom Gradient Boosting Ansatz besser genutzt werden kann. Es kann hier nicht auf die Allgemeinheit im Bezug auf Modellselektion geschlossen werden, doch scheint dieser Ansatz die zeitlichen Informationen besser mit der gesuchten Information in Zusammenhang bringen zu können.

Fazit

Wie bereits im vorangegangenen Blogpost, in dem ein „Churn Case“ mit den beiden genannten Modellen untersucht wurde, zeigt Gradient Boosting bei Klassifikationsproblemen, die sich nicht um Bild, Sprache oder Handschrift drehen, starke Ergebnisse gegenüber Deep Learning. Dies zeigte sich in der Vergangenheit auch in internen COREAI Projekten sowie in der Community der Datenanalysten. Neuronale Netzwerke sind aufgrund ihrer flexiblen und potentiell mächtigen Architektur in der Lage, hoch-komplexe Muster in sehr komplexen Datentypen (Video, Bild oder Sprache) zu entdecken, doch sind sie nicht zwingend auch bei niederdimensionaleren Datentypen in Führung.

Ziel des Blogposts war es, die beiden Modelle vor dem Hintergrund des Feature Engineering zu betrachten und die Frage zu beantworten, ob der Aufwand hierfür gerechtfertigt ist. Neben dem Umstand, dass Feature Engineering in der Vergangenheit ein essentielles Werkzeug der Datenanalyse war, das dem Algorithmus ermöglicht, Informationen zu nutzen, die implizit in den Daten enthalten sind, zeigte die Untersuchung, dass der Erfolg des Arbeitsschrittes abhängig vom verwendeten Alogorithmus ist. Es scheint zum Einen sinnvoll und essentiell, Domänenwissen in die Datenanalyse von Beginn an einfließen zu lassen, auch wenn dies einen oft nicht unbeträchtlichen Aufwand der Fachseite bedarf. Das Bild des einsamen Datenanalysten in Kellerräumen lässt sich also nicht rechtfertigen. Zum Anderen ist es wichtig, eine Entscheidung über den Algorithmus im Zielbild nicht ohne den Faktor des Feature Engineering zu treffen, da möglicherweise eine Leistungszunahme grundsätzlich abhängig hiervon ist. Mit ansteigender Zahl von Datenquellen, was in heutigen Projektkontexten nicht ungewöhnlich ist, steigt das Potential auf implizit enthaltene Informationen, da die Daten mit hoher Wahrscheinlichkeit nicht für diesen Zweck vorgesehen waren. Das zu hebende Potential findet sich im gewissenhaften Feature Engineering.

Quellen

1. **Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says**
   https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#6ca5c1016f63
2. **Elo Merchant Category Recommendation**
   https://www.kaggle.com/c/elo-merchant-category-recommendation#description)
3. **This is going to be epic... sit back, relax and enjoy!**
   https://www.kaggle.com/raddar/target-true-meaning-revealed
4. **Concepts in predictive machine learning**
   http://www.davidwind.dk/wp-content/uploads/2014/07/main.pdf
5. **Gradient Boosting vs. Deep Learning. Möglichkeiten des Einsatzes Künstlicher Intelligenz im Banking**
   https://core.se/de/techmonitor/gradient-boosting-vs-deep-learning-moeglichkeiten-des-einsatzes-kuenstlicher-intelligenz-im-banking

Als Expert Manager unterstützt **Julius Heitmann** bei CORE im Schwerpunkt die Konzeption komplexer IT-Transformationen. In der Automobilindustrie hat er umfangreiche Erfahrungen im Umgang mit Big Data und Data Analytics sammeln können. Diese bringt er insbesondere in Projekten zur Analyse und Optimierung großer Datenmengen ein und entwickelt passgenaue Steuerungssysteme für die Datenverarbeitung vor allem im Finanzsektor.

Mail: julius.heitmann@core.se

Christoph Sonn ist Engineering Manager bei CORE. Er verfügt über einen doppelten Abschluss in Wirtschaftswissenschaften und Chemie und hat seinen fachlichen Schwerpunkt im Bereich der Entwicklung innovativer Produkte und Digitalisierungslösungen für Banken, Versicherungen und Biotechnologieunternehmen. Von der strategischen Konzeption bis zum Go-live innovativer IT Transformationen entwickelt Christoph Lösungen für Unternehmen, um sich im Markt zu positionieren.

Mail: christoph.sonn@core.se

CORE SE
Am Sandwerder 21-23
14109 Berlin | Germany
https://core.se/
Phone: +49 30 263 440 20
office@core.se

COREtransform GmbH
Am Sandwerder 21-23
14109 Berlin | Germany
https://core.se/
Phone: +49 30 263 440 20
office@core.se

COREtransform GmbH
Limmatquai 1
8001 Zürich | Helvetia
https://core.se/
Phone: +41 44 261 0143
office@core.se

COREtransform Ltd.
Canary Wharf, One Canada Square
London E14 5DY | Great Britain
https://core.se/
Phone: +44 20 328 563 61
office@core.se

COREtransform MEA LLC
DIFC – 105, Currency
House, Tower 1
P.O. Box 506656
Dubai I UAE Emirates
https://core.se/
Phone: +97 14 323 0633

