

Fakultät Informatik

Analyse und Visualisierung von Datenqualität innerhalb eines Data Warehouse.

Bachelorarbeit im Studiengang Informatik

vorgelegt von

Lukas Welsch

Matrikelnummer 3201095

Erstgutachter:	Prof. Dr. Korbinian Riedhammer
Zweitgutachter:	Prof. Dr. Jens Albrecht
Betreuer:	Dipl. Ing. Andreas Sachs
Unternehmen:	Sopra Financial Technology GmbH

© 2021

Dieses Werk einschließlich seiner Teile ist **urheberrechtlich geschützt**. Jede Verwertung außerhalb der engen Grenzen des Urheberrechtsgesetzes ist ohne Zustimmung des Autors unzulässig und strafbar. Das gilt insbesondere für Vervielfältigungen, Übersetzungen, Mikroverfilmungen sowie die Einspeicherung und Verarbeitung in elektronischen Systemen.

Prüfungsrechtliche Erklärung der/des Studierenden

Angaben des bzw. der Studierenden:

Name: Welsch

Vorname: Lukas


Matrikel-Nr.: 3201095

Fakultät: Informatik



Studiengang: Informatik



Semester: Wintersemester  2020/21

Titel der Abschlussarbeit:

Analyse und Visualisierung der Datenqualität innerhalb eines Data Warehouse

Ich versichere, dass ich die Arbeit selbständig verfasst, nicht anderweitig für Prüfungszwecke vorgelegt, alle benutzten Quellen und Hilfsmittel angegeben sowie wörtliche und sinngemäße Zitate als solche gekennzeichnet habe.

Muss noch ausgefüllt

Ort, Datum, Unterschrift Studierende/Studierender

Erklärung zur Veröffentlichung der vorstehend bezeichneten Abschlussarbeit

Die Entscheidung über die vollständige oder auszugsweise Veröffentlichung der Abschlussarbeit liegt grundsätzlich erst einmal allein in der Zuständigkeit der/des studentischen Verfasserin/Verfassers. Nach dem Urheberrechtsgesetz (UrhG) erwirbt die Verfasserin/der Verfasser einer Abschlussarbeit mit Anfertigung ihrer/seiner Arbeit das alleinige Urheberrecht und grundsätzlich auch die hieraus resultierenden Nutzungsrechte wie z.B. Erstveröffentlichung (§ 12 UrhG), Verbreitung (§ 17 UrhG), Vervielfältigung (§ 16 UrhG), Online-Nutzung usw., also alle Rechte, die die nicht-kommerzielle oder kommerzielle Verwertung betreffen.

Die Hochschule und deren Beschäftigte werden Abschlussarbeiten oder Teile davon nicht ohne Zustimmung der/des studentischen Verfasserin/Verfassers veröffentlichen, insbesondere nicht öffentlich zugänglich in die Bibliothek der Hochschule einstellen.

Hiermit ☒ genehmige ich, wenn und soweit keine entgegenstehenden Vereinbarungen mit Dritten getroffen worden sind,
☐ genehmige ich nicht,

dass die oben genannte Abschlussarbeit durch die Technische Hochschule Nürnberg Georg Simon Ohm, ggf. nach Ablauf einer mittels eines auf der Abschlussarbeit aufgebrachten Sperrvermerks kenntlich gemachten Sperrfrist

von Jahren (0 - 5 Jahren ab Datum der Abgabe der Arbeit),

der Öffentlichkeit zugänglich gemacht wird. Im Falle der Genehmigung erfolgt diese unwiderruflich; hierzu wird der Abschlussarbeit ein Exemplar im digitalisierten PDF-Format auf einem Datenträger beigelegt. Bestimmungen der jeweils geltenden Studien- und Prüfungsordnung über Art und Umfang der im Rahmen der Arbeit abzugebenden Exemplare und Materialien werden hierdurch nicht berührt.

Ort, Datum, Unterschrift Studierende/Studierender

Formular drucken

Datenschutz: Die Antragstellung ist regelmäßig mit der Speicherung und Verarbeitung der von Ihnen mitgeteilten Daten durch die Technische Hochschule Nürnberg Georg Simon Ohm verbunden. Weitere Informationen zum Umgang der Technischen Hochschule Nürnberg mit Ihren personenbezogenen Daten sind unter nachfolgendem Link abrufbar: <https://www.th-nuernberg.de/datenschutz/>

Kurzdarstellung

Kurze Zusammenfassung der Arbeit, höchstens halbe Seite. Deutsche Fassung auch nötig, wenn die Arbeit auf Englisch angefertigt wird.

Dies hier ist ein Blindtext zum Testen von Textausgaben. Wer diesen Text liest, ist selbst schuld. Der Text gibt lediglich den Grauwert der Schrift an. Ist das wirklich so? Ist es gleichgültig, ob ich schreibe: „Dies ist ein Blindtext“ oder „Huardest gefburn“? Kjift – mitnichten! Ein Blindtext bietet mir wichtige Informationen. An ihm messe ich die Lesbarkeit einer Schrift, ihre Anmutung, wie harmonisch die Figuren zueinander stehen und prüfe, wie breit oder schmal sie läuft. Ein Blindtext sollte möglichst viele verschiedene Buchstaben enthalten und in der Originalsprache gesetzt sein. Er muss keinen Sinn ergeben, sollte aber lesbar sein. Fremdsprachige Texte wie „Lorem ipsum“ dienen nicht dem eigentlichen Zweck, da sie eine falsche Anmutung vermitteln.

Abstract

Only if thesis is written in English.

Dies hier ist ein Blindtext zum Testen von Textausgaben. Wer diesen Text liest, ist selbst schuld. Der Text gibt lediglich den Grauwert der Schrift an. Ist das wirklich so? Ist es gleichgültig, ob ich schreibe: „Dies ist ein Blindtext“ oder „Huardest gefburn“? Kjift – mitnichten! Ein Blindtext bietet mir wichtige Informationen. An ihm messe ich die Lesbarkeit einer Schrift, ihre Anmutung, wie harmonisch die Figuren zueinander stehen und prüfe, wie breit oder schmal sie läuft. Ein Blindtext sollte möglichst viele verschiedene Buchstaben enthalten und in der Originalsprache gesetzt sein. Er muss keinen Sinn ergeben, sollte aber lesbar sein. Fremdsprachige Texte wie „Lorem ipsum“ dienen nicht dem eigentlichen Zweck, da sie eine falsche Anmutung vermitteln.

Inhaltsverzeichnis

1. Einleitung	1
2. Grundlagen	3
2.1. Definition von Datenqualität	3
2.1.1. Datenqualitätsdimensionen	3
2.1.2. Definition / Metriken der Dimensionen	4
3. Daten	5
4. Methoden	9
4.1. Stakeholderanalyse	9
4.1.1. Stakeholder identifizieren	10
4.1.2. Betroffenheitsanalyse	11
4.2. (Statistisches Verfahren) ABT / Voranalyse	12
4.3. Machine Learning	12
4.3.1. Risikoscoring	13
4.3.2. Auswirkungen der einzelnen Merkmale auf das Gesamtscoreing	15
4.4. Visualisierung	15
4.5. Sicherstellung der korrekten Prozesse	15
5. Experimente	17
5.1. ML-Experimente / Modell Evaluation	17
5.1.1. Versuchsaufbau	17
5.1.2. Datenaufbereitung	18
6. Ausblick	19
7. Fazit	21
A. Supplemental Information	23
Abbildungsverzeichnis	25
Tabellenverzeichnis	27
Auffistung	29

Literaturverzeichnis 31

Glossar 33

Kapitel 1.

Einleitung

- Kann mit Hilfe von ML-Verfahren berechnet werden, ob Risikobewertungen über Kunden aktualisiert werden müssen? - Können die vorliegenden MetaDaten visualisiert werden, sodass Erkenntnisse zur (Prozess)-Qualität gewonnen werden können?

Kapitel 2.

Grundlagen

2.1. Definition von Datenqualität

Datenqualität wird in der Literatur auf verschiedene Arten definiert. Viele Datenqualitätsstudien verwenden Korrektheit als einziges (oder als Haupt-) Datenqualitätsmerkmal. [Wang 96] Datenqualität umfasst jedoch nicht nur die Korrektheit von Daten, sondern auch anderen Dimensionen. Ein Paper von 2011 zeigt den Einfluss von anderen Dimensionen, wie zb. Vollständigkeit auf die Erkennungsrate von Klassifikatoren [Esp11] Die Autoren schlagen vor den Aufbereitungsprozess und vor allem Datenqualität nicht nur auf die Korrektheit von Daten zu beziehen sondern auch auf die anderen Datenqualitätsdimensionen zu achten.

2.1.1. Datenqualitätsdimensionen

Die Dimensionen Richtigkeit, Vollständigkeit, Konsistenz und Aktualität werden in den meisten Veröffentlichungen genannt, allerdings gibt es keinen Standard, weder im Bezug auf die verwendeten Dimensionen, noch die Definition der Dimensionen. [Scan 02] **Begründung finden! In dieser Arbeit werden diese Definition von Datenqualität verwendet, da es sich am Besten anbietet?** Jede dieser Datenqualitätsdimensionen umfasst eine Facette der Datenqualität.

Richtigkeit Richtigkeit wird als Gleichheit zwischen zwei Werten definiert, sodass die Daten die Wirklichkeit korrekt repräsentieren. Beispiel: Eine Person mit dem Namen Max Mustermann ist als Mx Mustermann abgespeichert. Die gespeicherten Daten repräsentieren nicht die Wirklichkeit, sie sind somit nicht richtig.

Aktualität / temporäre Richtigkeit Als Spezialisierung der Richtigkeit kann die temporäre Richtigkeit gesehen werden, die richtige Repräsentation zu einem bestimmten Zeitpunkt oder Zeitspanne beschreibt. Die temporäre Richtigkeit zum aktuellen Zeitpunkt wird auch als Aktualität bezeichnet.

Vollständigkeit Die Daten sind umfangreich genug für die jeweilige Aufgabe. [Wang 96] Da ein Data Warehouse aus relationalen Datenbanken besteht, wird die Vollständigkeit im Folgenden auf relationale Datenbanken bezogen. Die Vollständigkeit wird von [?] in drei Klassen kategorisiert.

- Schema: Grad der Daten, die nicht im Schema fehlen
- Spalte: Anzahl der fehlenden Werte innerhalb einer Spalte
- Population: Wenn eine Spalte alle Institute (36,70..) beinhalten sollte, aber welche fehlen, dann herrscht Populationsunvollständigkeit

Konsistenz Die gleichen (redundanten) Datensätze haben die gleichen Werte in verschiedenen Tabellen. Ein Beispiel für Konsistenz ist die referentielle Integrität, die sicherstellt, dass Datensätze nur auf existierende Datensätze verweisen.

2.1.2. Definition / Metriken der Dimensionen

Um diese eher abstrakten Definitionen messbar zu machen haben sich einige Verfahren etabliert. Diese lassen sich grob in die Kategorien subjektiv und objektiv aufteilen. [?] Bei den subjektiven Verfahren wird die vorherrschende Datenqualität durch Experten geschätzt. Dieses Verfahren ist allerdings nicht nur zeitaufwendig für die Experten, sondern ist auch fehleranfällig. **HIERZITATANGEBEN** []

Eine weitere Möglichkeit besteht darin objektive Verfahren zu verwenden, bei denen mit Hilfe von mathematischen Funktionen die Datenqualität berechnet oder geschätzt wird.

Des weiteren gibt es ein kombiniertes Verfahren, bei denen Grenzwerte von Experten geschätzt und anschließend durch mathematische Funktionen abgeglichen werden. Das kombinierte Verfahren kann durch ein Quadrat visualisiert werden, das an der y-Achse die subjektive Bewertung und auf der x-Achse die objektive Bewertung zeigt. Bei einer guten Datenqualität liegt das Ergebnis im IV. Quadrant. [?]

Kapitel 3.

Daten

Die in der Arbeit verwendeten Daten stammen aus dem Data Warehouse einer deutschen Bank. Zunächst wird mit den Entwicklungsdaten gearbeitet, da diese im wesentlichen den produktiven Daten entsprechen. In den Entwicklungsdaten wurden personenbezogene Daten, aufgrund des Datenschutzes, bereits pseudonymisiert und anonymisiert.

Ein- und Ausschlusskriterien für die Daten:

In diesem Projekt werden zwei verschiedene Datensätze verwendet. Zum einen wird ein Datensatz benötigt, der für die Visualisierung verwendet werden kann. Dieser Datensatz benötigt Eigenschaften, die sich so visualisieren lassen, dass anhand der Visualisierung neue Erkenntnisse abgeleitet werden können. Am besten bieten sich Daten an, die nur wenige Dimensionen haben, da sich niedrig dimensionale Daten besser in einer Grafik darstellen lassen.

Des weiteren werden Daten benötigt, die für Machine Learning geeignet sind. Hierfür werden Daten benutzt, die feste Labels zu bestimmten Input Parametern besitzt. Ein Klassifikator kann anhand der Eigenschaften einer Ausprägung und dem dazugehörigem Label lernen, welche Eigenschaften zu welchem Label führen und so eine Klassifikation durchführen.

Für die Visualisierung können Metadaten zu den Prozessen verwendet werden. Diese beinhalten die Anzahl der verarbeiteten Datensätze pro Zeiteinheit.

Nach einer Recherche und Analyse der vorhandenen Daten bieten sich die Daten zum Risikoscoring am besten an. Diese Daten sind in einer Vielzahl vorhanden und erfüllen die gewünschten Anforderung an Machine Learning. Die Daten sind nachfolgend genauer beschrieben:

Das Label ist der Risikowert eines Geschäfts, wobei ein Geschäft in diesem Fall ein Darlehen ist. Dieser Risikowert hat einen Wertebereich von 0E bis 4E. Die Inputdaten bestehen aus den folgenden Eigenschaften, die zum Training des Klassifikators verwendet werden können.

Die Daten werden mit Hilfe eines ETL-Prozesses aus dem DataWarehouse extrahiert. Dieses Programm wurde im Rahmen der vorliegenden Arbeit entwickelt und ist in einer Versionsverwaltung abgelegt. Zunächst werden die benötigten Daten aus unterschiedlichen Quelltabellen geladen, hierzu zählen z.B. personenbezogene Daten oder auch die tatsächlichen Risikoscorings. Anschließend werden die Daten bereinigt und mit Hilfe einer Aggregatfunktion die Kredite, Saldo summiert und die Anzahl in einer zusätzlichen Spalte gespeichert.

Nach [Soko 05] verwenden Auskunfteien bestimmte Daten als Grundlage des Risikoscores. Darunter werden Kfz-Besitz, Wohndauer, verfügbares Einkommen, Berufsdaten und Haftende nicht im Data Warehouse gespeichert. Somit ergibt sich folgende Liste relevanter Daten für die ML-Experimente.

- Alter (Geburtsdatum)
- ausgeübter Beruf
- Bürge vorhanden
- Familienstand
- Geschlecht
- Kinderanzahl
- Kredite in Stück
- Nettokreditbetrag
- Nationalität
- Schufa
- Haushaltstyp

Im Rahmen der Recherche zu Daten, die sich zur Berechnung des Risikoscores anbieten konnten weitere Daten als nützlich identifiziert werden. Diese sind der aktuelle Rückstand der Kreditzahlung und Daten aus öffentlichem Schuldnerverzeichnissen.

Zunächst sollte sich eine Übersicht über die Daten beschafft werden:

```

1 SELECT VALIDDT, COUNT(VALIDDT)
2 FROM SCHEMA. bachelordaten
3 GROUP BY VALID_DT
4 ORDER BY VALID_DT
5 LIMIT 5

```

Listing 3.1: Überblick über die Daten

Anhand dieser Abfrage lassen sich die Anzahl der Daten zu den jeweiligen Tagen bestimmen. Beispielhaft wird mit einem aktuellen VALIDDT gearbeitet. Folgende Spalten sind in der Tabelle vorhanden:

Daten für die Visualisierung: Zur Visualisierung werden die Daten aus den Prozessen verwendet.

Kapitel 4.

Methoden

Um die nachfolgenden Methoden zur Analyse und Visualisierung von Datenqualität zu vergleichen, werden sowohl qualitative als auch quantitative Untersuchungen durchgeführt. Für eine Datenqualität, die sich in einem Unternehmen etablieren kann, ist es nötig diese von den Gesichtspunkten aller Stakeholdern zu betrachten. Anhand einer Stakeholderanalyse wird es möglich sein die Probleme der Datenqualität auf zwei Ebenen zu betrachten. Zum einen benötigen die Business-User gute Datenqualität, um die richtigen Zinsen anhand des Risikoscorings zu vergeben und auf der anderen Seite benötigen die Entwickler eine Möglichkeit, um ihre ETL-Strecken zu überprüfen. Für diese beiden Probleme können zwei verschiedene Verfahren eingesetzt werden. Zur Identifikation und Überprüfung des Risikoscores kann ein Machine Learning Algorithmus trainiert werden, der die korrekten Werte vorhersagt und somit einen Aufschluss darüber gibt, wie viele Datensätze aktualisiert werden sollten. Um die ETL-Strecken zu analysieren sind die Metadaten zu den Prozessen nötig, die anzeigen, wie viele Daten an jedem Tag extrahiert wurden. Auffälligkeiten können dann genutzt werden, um Maßnahmen einzuleiten. Hierfür bietet sich eine Visualisierung an, die in Echtzeit die Daten geliefert bekommt und anschließend interpretierbar darstellt. Die beiden beschriebenen Verfahren werden in dem Kapitel Experimente durchgeführt.

4.1. Stakeholderanalyse

Als Stakeholder Mit Hilfe einer Stakeholderanalyse ist es möglich, die negativen Einflüsse zu erkennen und die positiven Einflüsse zu nutzen. Zum Anderen können die Erwartungen der einzelnen Stakeholder erkannt und die Projektziele richtig gewichtet werden.

Die Stakeholderanalyse wird nach folgenden Schritten durchgeführt:

1. Stakeholder identifizieren
2. Stakeholder Einfluss analysieren
3. Aktionsplanung (Maßnahmen ableiten)

[?]

4.1.1. Stakeholder identifizieren

In diesem Schritt werden die Stakeholder genannt, kurz beschrieben und visualisiert. Aufgrund der Analyse des mittelständischen Unternehmens aus der Bankenbranche können folgende Stakeholder identifiziert werden:

- Entwickler
- Business User
- Bankmitarbeiter
- Product Owner

Als Stakeholder lassen sich zunächst die Entwickler identifizieren. Diese sind für die Einhaltung einer guten Datenqualität unmittelbar wichtig, da diese die Extraktionsprozesse entwickeln, die zu Fehlern in der Datenqualität führen können. Die Daten werden von einem legacy-System in das Data Warehouse übertragen, indem die Daten aus einer hierarchischen Datenbank als Datei abgelegt werden. Ein Prozess iteriert über alle Dateien und extrahiert die Daten, die anschließend neu strukturiert im Data Warehouse abgelegt werden. Aufgrund von einigen Problemen in der Extraktion kann es dazu führen, dass die Daten nicht korrekt extrahiert werden. Für diesen Fall gibt es eine Überwachung der Programme, die einen Fehler meldet. Wenn jedoch weniger oder mehr Daten verarbeitet werden als üblich deutet dies auch auf einen Fehler in der Extraktion hin.

Auf der anderen Seite gibt es die Business User, die beispielsweise die Gesamtbanksteuerung übernehmen. Bei diesem Teilgebiet wird errechnet, wie gut die Zinssätze sein dürfen, sodass die Bank Gewinn erzielt und gleichzeitig möglichst viele Kunden zufrieden stellen kann. Für diesen Stakeholder sind alle Dimensionen wichtig, da nur mit einer sehr guten Datenqualität eine optimale Gesamtbanksteuerung vorgenommen werden kann. Würde der Stakeholder nur über falsche, unvollständige und veraltete Daten verfügen, könnte dies Verluste für die Bank zur Folge haben. Besonders entscheidend ist für diesen Stakeholder, dass der Wert für den Risikoscore richtig und aktuell ist. Beim Risikoscore werden die Daten der Kunden an einen externen Dienstleister gesendet. Dieser berechnet einen Risikoscore, anhand dessen die Zinsvergabe erfolgt.

Des weiteren ist der Bankmitarbeiter Bestandteil der Kreditvergabe. Die Kreditvergabe erfolgt anhand einer Bewertung des Kunden, je nachdem wie gut der Risikoscore des Kunden ist, umso bessere Zinssätze bekommt dieser. Für den Bankmitarbeiter ist es wichtig, dass die Daten zum Risikoscore sowohl richtig als auch aktuell sind.

Die beschriebenen Stakeholder werden mit ihren Erwartungen an das Projekt in der folgenden Grafik dargestellt.

4.1.2. Betroffenheitsanalyse

ID	Wer	Betroffenheit	Erwartung	Macht	Einstellung	Maßnahmen
S1	Entwickler	m	Fehler in der Entwicklung werden angezeigt; Wenig Programmieraufwand	g	Neutral	Erklärung der Notwendigkeit, Zeitvorteil aufzeigen
S2	Business User	h	Daten sollten fehlerfrei und aktuell sein	g	Positiv	-
S3	Product-Owner	h	Fehler werden frühzeitig erkannt, sodass Kunde nichts merkt	g	Positiv	-
S4	Bankmitarbeiter	h	Daten müssen immer verfügbar und richtig sein	h	Positiv	-

Welchen Stakeholder interessiert welche Dimension?

Stakeholder	Dimension	Begründung
Entwickler	Vollständigkeit	Der Entwickler geht davon aus, dass die Daten die vom Quellsystem
Business User	Alle Dimensionen	
Datenmanager		
Kunde	Keine Dimension	Geht davon aus, dass die Datenqualität schon geprüft wurde

Für Welche Stakeholder kann sollte welches Verfahren angewendet werden? - Business User: Risikoscoring sollte immer aktuell sein, dabei kann ein ML Verfahren verwendet werden, um die Daten zu klassifizieren und bei großer Abweichung können diese Daten dann aktualisiert oder neu angefordert werden.

- Bank Mitarbeiter: ? -

In dieser Bachelorarbeit wird die Verwendung von Machine Learning Verfahren zur Verbesserung der Datenqualität untersucht. Deshalb bietet sich das Thema Risikoscoring am besten an, da es feste Ausgangswerte (Labels) bietet, die trainiert werden können. Der trainierte Klassifikator kann anschließend verwendet werden, um zu berechnen, ob ältere Risikoscorings vom Ergebnis des Klassifikators abweichen. Denn es können sich Attribute

aktualisieren, die einen positiven oder negativen Einfluss des Risikoscorings zur Folge haben. Allerdings wird das Risikoscoring nicht regelmäßig aktualisiert, da eine Aktualisierung Geld kostet. Grundvoraussetzung für dieses Vorhaben ist die Risikoscorings anhand von aktuellen Daten mit hoher **Präzision** vorherzusagen. Dafür werden nachfolgend Klassifikatoren ausgewählt und anschließend im Kapitel Experimente geprüft und ausgewertet.

Allerdings sollten auch für die anderen Stakeholder Verfahren entwickelt werden, um die Datenqualität zu verbessern. Dies beinhaltet unter anderem eine Peer-Review, um die Codequalität sicherzustellen, sowie statische Code-Analysen. Eine bessere Codequalität führt zu einer besseren Datenqualität, da keine bzw. weniger menschliche Fehler im System erzeugt werden. Des weiteren ist es notwendig die Datenbankschemas gut und exakt zu definieren, sodass keine null-Werte an Stellen eingefügt werden, an denen es keine null-Werte geben darf. Diese Lücken in den Daten sollten immer mit einem Default-Wert belegt werden und null-Values zu einem Abbruch der ETL-Strecken führen. Dadurch kann die Dimension der Vollständigkeit verbessert werden. Auch Log-Dateien die Aussagen über den ETL-Prozess liefern, können hilfreich sein und sollten ausgewertet werden. **Aufzeigen, warum es wichtig ist auch Visualisierungen zu verwenden**

4.2. (Statistisches Verfahren) ABT / Voranalyse

Mit Hilfe einer sogenannten Analytical Base Table ist es möglich erste Aussagen über die Datenqualität der zu prüfenden Daten zu treffen. Zunächst wird diese generiert, um festzustellen, ob die Daten für weitere Untersuchungen verwendet werden können. Anschließend kann ein Plan entwickelt werden, der darstellt, welche Maßnahmen bei unterschiedlichen Datenqualitätsproblemen getroffen werden kann.

4.3. Machine Learning

Im folgenden Kapitel werden die Verfahren vorgestellt und kurz beschrieben, die für eine Klassifikation benötigt werden. Als Beispiel dient der Risikoscore, da dieser feste Ausgangswerte besitzt und anhand von mehreren Eigenschaften bestimmt wird. Mit Hilfe von Machine Learning können die Risikoscores berechnet werden, ohne den Algorithmus zu kennen, der dahinter steckt. Dies ist interessant für dieses Projekt, da die Risikoscores bisher bei einer Auskunft eingekauft werden. Mit Hilfe eines trainierten Klassifikators können die Daten neu angefordert werden, die sich anhand der Ergebnisse der Klassifikation ändern müssten. Ein weiterer Ansatz besteht darin mit Hilfe von unüberwachtem Lernen mögliche Fehler in den Daten zu erkennen. Anschließend können die als ungewöhnlich markierten Datensätze einem Stakeholder vorgelegt werden. Nach dem der Stakeholder die Daten zugeordnet hat, die tatsächlich einem Fehler entsprechen, kann mit diesen Labels ein neuer Klassifikator trainiert werden. Da dieses Verfahren allerdings sehr viel manuellen Aufwand benötigt und

bei diesem Vorgehen Filter wesentlich zuverlässiger verwendet werden können, wird dieser Ansatz nicht weiter verfolgt. Allerdings ist dies eine gängige Praxis bei Machine-Learning Data Quality tools, wie z. B. talend.

4.3.1. Risikoscoring

Der Risikoscore wird in dem vorliegenden Unternehmen durch einen externen Dienstleister, eine Auskunftsbüro berechnet. Der Dienstleister erhält einige Daten über eine Schnittstelle und anschließend werden diese einem Risikoscore zugeordnet und dem Unternehmen zurückgesendet. Dieser Risikoscore kann mit Hilfe von Machine Learning Methoden berechnet werden, um zu überprüfen, welche Daten aktualisiert werden sollten. Bei den Daten, bei denen der Wert Abweichungen aufweist können neue Daten angefordert werden. Dies verbessert die Dimensionen der Richtigkeit und vor allem der Aktualität, die ein Teilgebiet der Richtigkeit darstellt.

Die Scores, die der Klassifikator erzeugt können nicht ausschließlich als Basis für Entscheidungen verwendet werden und kann deshalb nur dabei helfen zu entscheiden, welche Daten aktualisiert werden sollten. Dies liegt daran, dass Risikoscores nur mit Hilfe von Regressionen berechnet werden dürfen und nicht mit beispielsweise Neuronalen Netzen.

Es müssen feste Gruppen von 1A bis 4E vorhergesagt werden. Deshalb bieten sich Klassifikationsalgorithmen an. Nachfolgend werden die in den Experimenten verwendeten Klassifikatoren vorgestellt und darauf eingegangen, welche Parameter zur Verbesserung des Klassifikationsergebnisses diese haben.

Vorgehen

1. Daten extrahieren 2. Klassifikator trainieren 3. Mit Validierung besten auswählen 4. alte Daten und trainierten Klassifikator benutzen, um zu testen, ob es Abweichungen gibt 5. Veraltete Daten anzeigen -> diese sind nun Grundlage zum neu bestellen

KNN

Das Grundprinzip des K-Nearest-Neighbor Klassifikators besteht darin die Distanzen eines neuen Punktes zu allen anderen Punkten zu berechnen, um anschließend diesem einer Klasse zuzuordnen. Für die Zuordnung der Klassen werden die k-nächsten Punkte verwendet. Hierbei können verschiedene Distanzmetriken verwendet werden. Diese sind zum Beispiel die euklidische Distanz oder die Manhattan Distanz.

Support Vector Machine (SVM) (mit Multiklassenerweiterung)

Die Idee der SVM besteht darin eine ideale Trennlinie zwischen zwei Gruppen zu finden. Hierfür werden sogenannte Stützvektoren (Support Vectors) berechnet, indem eine mathematische Gleichung gelöst wird. Die Support Vector Machine ist in ihrer Grundform nur in der Lage Zweiklassen-Probleme zu lösen. Da es in dem vorliegenden Datensatz allerdings

mehrere Klassen gibt, ist es notwendig eine Multiklassenerweiterung zu verwenden. Dafür wird in der Implementierung von Sklearn One vs One verwendet. Bei diesem Verfahren werden Kombinationen gebildet, bei denen jede Klasse im Vergleich zu einer anderen trainiert wird. Um das Ergebnis einer Vorhersage zu erhalten, wird die gewählt, dessen Summe der einzelnen Vorhersagen am Größten ist.

Zur Berechnung der benötigten Vergleiche kann folgende Formel verwendet werden:

$$(\text{NumClasses} * (\text{NumClasses} - 1)) / 2$$

logistische Regression Bei diesem Verfahren kann die Wahrscheinlichkeit vorausgesagt werden, welche Ausprägungen von unabhängigen Variablen zu einer abhängigen Variable führen. Dies wird über einen Threshold gelöst, der angibt bei welcher Wahrscheinlichkeit eine bestimmte Klasse vorhergesagt werden soll oder nicht. In unserem Anwendungsfall hat dies den Vorteil, dass bei unstimmgigen Datensätze (Wahrscheinlichkeit ist nicht eindeutig) auch neuangefordert werden könnte. Da dieses Verfahren darauf beruht, dass die Datensätze unabhängig voneinander sind, muss zunächst die Abhängigkeit der Variablen berechnet werden. Dies kann mit Hilfe des Pearson Korrelations Koeffizienten erledigt werden.

Klassifikation Tree Ein Klassifikationsbaum kann dafür verwendet werden, um Regeln zu generieren, die in ein SQL-Skript umgewandelt werden können. Dies hätte den Vorteil, dass dieser Klassifikator nur einmal trainiert und verwendet werden müsste. Anschließend kann das Ergebnis des Klassifikators in ein Skript überführt werden und zukünftig in die bestehende Infrastruktur mitabgelegt werden, ohne dass ein Mitarbeiter Kenntnisse für Python oder Machine Learning benötigt. Des Weiteren können die gewonnen Informationen verwendet werden, um herauszufinden, welche Eigenschaften besonders wichtig sind. Dieses Wissen kann verwendet werden, um die Daten, die einen großen Einfluss auf das Klassifikationsergebnis haben händisch zu überprüfen. So kann die Datenqualität verbessert werden, indem speziell die Eigenschaften der Daten überprüft werden, die tatsächlich für ein gutes Ergebnis benötigt werden.

Standardisierung Mit Hilfe der Min/Max-Normierung können Merkmale auf den Wertebereich $[0,1]$ abgebildet werden. Dadurch werden

Modellbewertung Mit Hilfe von einigen Metriken kann die Güte des Klassifikators bewertet werden. Dies hilft dabei die besten Parameter und das am besten für diesen Anwendungsfall geeigneten Klassifikator zu finden. Folgende Verfahren werden häufig bei der Bewertung eingesetzt: - Accuracy - Precision und Recall - Sensitivity und Specifity - F-Score - confusion matrix - auc - roc curve

Da die Confusion Matrix TP, FP, FN und TN angibt ist diese gut geeignet, da sowohl die Allerdings ist die Standard Confusion Matrix nur für ein Zweiklassen Problem geeignet und in diesem konkreten Fall wird mit einem Mehrklassenproblem gearbeitet. Deshalb muss die Confusion Matrix erweitert werden auf ein Mehrklassenproblem.

Holdout k-Fold stratified k-Fold

Optimierung von Hyperparametern -> Einteilen in Train/Dev/Test

Vergleichen der versch. ROC-Kurven: ROC / AUC

4.3.2. Auswirkungen der einzelnen Merkmale auf das Gesamtscoreing

Es ist wichtig zu wissen, wie sich das Scoring zusammensetzt, um gezielt die Daten zu verbessern, die einen großen Einfluss haben.

Wie groß sind die Auswirkungen der einzelnen Merkmale auf das Gesamtscoreing.

4.4. Visualisierung

Auch mit Hilfe einer Visualisierung können Informationen über die Daten gewonnen werden. In diesem Projekt werden die Daten visualisiert, die angeben, wie viele Daten extrahiert wurden. Dies geschieht in Echtzeit und werden anschließend in einem Dashboard dargestellt. Die Stakeholder können anhand diesem Dashboards sofort erkennen, ob die Datenmenge ungewöhnlich ist im Vergleich zu sonstigen in der Vergangenheit liegenden Datenlieferungen. Dies ist ein Indiz für Fehler, die in der Programmierung entstanden sind.

- Kibana Daten - Dashboard für Ampel logik

4.5. Sicherstellung der korrekten Prozesse

Ein weiterer wichtiger Aspekt guter Datenqualität besteht darin die Prozesse der Extraktionen so zu gestalten, dass diese fehlerfrei sind. Für diesen Anwendungsfall wird ein Dashboard erstellt, das live die Metadaten der Prozesse erhält und anschließend visualisiert. Es bieten sich folgende Visualisierungen an. Ein Liniendiagramm zur Darstellung des zeitlichen Verlaufs. Des weiteren ist es gut die mittlere Abweichung der Anzahl der extrahierten Daten mit darzustellen. Dadurch ist es möglich zu erkennen, ob ungewöhnlich viele bzw. wenige Daten verarbeitet wurden.

Die Abbildung zeigt den aktuellen Extraktions-Prozess des Data Warehouse. Bei diesem werden Daten aus einigen Quellsystemen abgeholt und in eine zentrale Datenbank, dem Data Warehouse extrahiert. Anschließend werden die Daten in das korrekte Format konvertiert und den Endsystemen so bereitgestellt, wie diese die Daten erwarten.

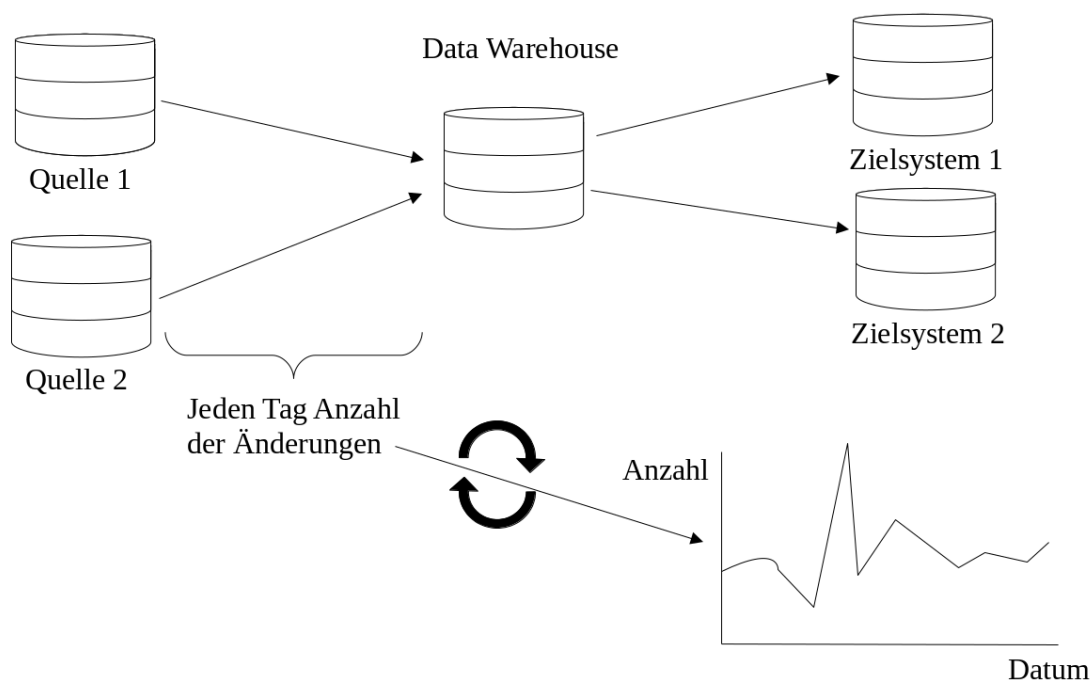


Abbildung 4.1.: Darstellung des aktuellen Extraktionsverfahrens und Integration der Visualisierung

Eine mögliche Visualisierung sollte die Meta-Daten, die von den Quellsystemen extrahiert werden erhalten und anschließend visualisieren. Hierfür wird per Trigger die Daten in Echtzeit an Elastic übertragen, dass die Daten anschließend mit Hilfe eines Kibana Dashboards visualisiert. Stakeholder sind dann in der Lage grafisch zu sehen, ob es große Abweichungen zu den Monaten / Wochen / Jahren davor gegeben hat und können so einschätzen, ob genauer nachgeforscht werden muss oder ob alles geklappt hat.

Kapitel 5.

Experimente

Die ML-Experimente bestehen darin die Methoden praktisch umzusetzen und die Ergebnisse zu interpretieren. Ziel der Experimente ist einen Klassifikator zu trainieren, der den Risikoscore zuverlässig vorhersagen kann. Mit dessen Hilfe ist es anschließend möglich die Aktualität bzw. Richtigkeit des Scores zu überprüfen und zu aktualisieren. Dies spart Geld, da nun nur noch die Daten aktualisiert werden müssen, bei denen der Klassifikator eine Abweichung feststellt. Auch für neue Kunden kann dieses Verfahren verwendet werden, um schnell ein Ergebnis zu erhalten, da nicht erst auf das Ergebnis der externen Ratingfirma gewartet werden muss. Allerdings sollten die Daten immer aktualisiert werden und nicht nur auf den Klassifikator vertraut werden. Um die Datenmenge zu reduzieren, werden zunächst nur die Daten von einem bestimmten Tag geladen und analysiert. Da die Klassifikatoren keine guten Ergebnisse aufweisen, wird als Lösung ein Upsampling der Daten vorgeschlagen, das die Lücken der Labels aus älteren Daten füllt.

5.1. ML-Experimente / Modell Evaluation

5.1.1. Versuchsaufbau

Die Experimente werden auf einem Windows-Rechner in einem Jupyter-Notebook durchgeführt. Als Python-Bibliotheken kommen pandas, numpy und sklearn zum Einsatz.

Vorgehensweise:

Zunächst werden die Daten mit Hilfe eines Python-Connectors von der Datenbank geladen, die wie im Kapitel 3 beschrieben, durch ein Skript erzeugt werden. Hierfür werden die Daten zunächst in ein python-Dictionary geladen und anschließend in einem DataFrame gespeichert.

5.1.2. Datenaufbereitung

Zunächst werden die Daten genauer untersucht um festzustellen ob die Daten bereinigt werden müssen oder ob sich die Klassenverteilung zu sehr unterscheidet. Zunächst zeigt eine Grafik die Verteilung der einzelnen Klassen. In der Abbildung ist zu erkennen, dass die Klassen sehr ungleichmäßig verteilt sind.

Klasse 0E ist deutlich häufiger vertreten als manche andere Klassen. Dies führt für einige Verfahren zu Problemen und sollte mit Hilfe von Under/Oversampling behoben werden.

Außerdem sind in dem vorliegen Datensatz einige textuelle Attribute, diese können durch einen OneHot Encoder so aufbereitet werden, dass diese in einem Klassifikator verwendet werden können. Da nach einem OneHot Encoding sehr viele Nullen in dem Datenset entstehen bietet sich zur weiteren Verarbeitung eine Sparse-Matrix an. Diese speichert nur die tatsächlich vorhandenen Daten und speichert keine Null-Werte.

Da manche Daten sehr von anderen abweichen (Varianz sehr groß) müssen die Daten skaliert werden.

Jedes Verfahren wird mit Upsampling brauchbar gemacht. Warum? (grob: was ist upsampling / oder Kapitel 4 verweisen)

KNN Ergebnisse mit Upsampling:

SVM

Random Forest

Integration der Daten in Kibana - ETL Prozess zur automatischen Erzeugung einer CSV-Datei. - Logstash? -

Kibana, Graphana

Reduktion der Datenmenge

- Welche Visualisierungen bieten sich an? - Gibt es evtl. Visualisierungen, die DQ-Probleme aufzeigen?

Kapitel 6.

Ausblick

Kapitel 7.

Fazit

Datenqualität ist unsichtbar, wenn alles richtig gemacht wird.

Anhang A.

Supplemental Information

Dies hier ist ein Blindtext zum Testen von Textausgaben. Wer diesen Text liest, ist selbst schuld. Der Text gibt lediglich den Grauwert der Schrift an. Ist das wirklich so? Ist es gleichgültig, ob ich schreibe: „Dies ist ein Blindtext“ oder „Huardest gefburn“? Kjift – mitnichten! Ein Blindtext bietet mir wichtige Informationen. An ihm messe ich die Lesbarkeit einer Schrift, ihre Anmutung, wie harmonisch die Figuren zueinander stehen und prüfe, wie breit oder schmal sie läuft. Ein Blindtext sollte möglichst viele verschiedene Buchstaben enthalten und in der Originalsprache gesetzt sein. Er muss keinen Sinn ergeben, sollte aber lesbar sein. Fremdsprachige Texte wie „Lorem ipsum“ dienen nicht dem eigentlichen Zweck, da sie eine falsche Anmutung vermitteln.

Das hier ist der zweite Absatz. Dies hier ist ein Blindtext zum Testen von Textausgaben. Wer diesen Text liest, ist selbst schuld. Der Text gibt lediglich den Grauwert der Schrift an. Ist das wirklich so? Ist es gleichgültig, ob ich schreibe: „Dies ist ein Blindtext“ oder „Huardest gefburn“? Kjift – mitnichten! Ein Blindtext bietet mir wichtige Informationen. An ihm messe ich die Lesbarkeit einer Schrift, ihre Anmutung, wie harmonisch die Figuren zueinander stehen und prüfe, wie breit oder schmal sie läuft. Ein Blindtext sollte möglichst viele verschiedene Buchstaben enthalten und in der Originalsprache gesetzt sein. Er muss keinen Sinn ergeben, sollte aber lesbar sein. Fremdsprachige Texte wie „Lorem ipsum“ dienen nicht dem eigentlichen Zweck, da sie eine falsche Anmutung vermitteln.

Und nun folgt – ob man es glaubt oder nicht – der dritte Absatz. Dies hier ist ein Blindtext zum Testen von Textausgaben. Wer diesen Text liest, ist selbst schuld. Der Text gibt lediglich den Grauwert der Schrift an. Ist das wirklich so? Ist es gleichgültig, ob ich schreibe: „Dies ist ein Blindtext“ oder „Huardest gefburn“? Kjift – mitnichten! Ein Blindtext bietet mir wichtige Informationen. An ihm messe ich die Lesbarkeit einer Schrift, ihre Anmutung, wie harmonisch die Figuren zueinander stehen und prüfe, wie breit oder schmal sie läuft. Ein Blindtext sollte möglichst viele verschiedene Buchstaben enthalten und in der Originalsprache gesetzt sein. Er muss keinen Sinn ergeben, sollte aber lesbar sein. Fremdsprachige Texte wie „Lorem ipsum“ dienen nicht dem eigentlichen Zweck, da sie eine falsche Anmutung vermitteln.

Nach diesem vierten Absatz beginnen wir eine neue Zählung. Dies hier ist ein Blindtext zum Testen von Textausgaben. Wer diesen Text liest, ist selbst schuld. Der Text gibt lediglich den Grauwert der Schrift an. Ist das wirklich so? Ist es gleichgültig, ob ich schreibe: „Dies ist ein Blindtext“ oder „Huardest gefburn“? Kjift – mitnichten! Ein Blindtext bietet mir wichtige Informationen. An ihm messe ich die Lesbarkeit einer Schrift, ihre Anmutung, wie harmonisch die Figuren zueinander stehen und prüfe, wie breit oder schmal sie läuft. Ein Blindtext sollte möglichst viele verschiedene Buchstaben enthalten und in der Originalsprache gesetzt sein. Er muss keinen Sinn ergeben, sollte aber lesbar sein. Fremdsprachige Texte wie „Lorem ipsum“ dienen nicht dem eigentlichen Zweck, da sie eine falsche Anmutung vermitteln.

Dies hier ist ein Blindtext zum Testen von Textausgaben. Wer diesen Text liest, ist selbst schuld. Der Text gibt lediglich den Grauwert der Schrift an. Ist das wirklich so? Ist es gleichgültig, ob ich schreibe: „Dies ist ein Blindtext“ oder „Huardest gefburn“? Kjift – mitnichten! Ein Blindtext bietet mir wichtige Informationen. An ihm messe ich die Lesbarkeit einer Schrift, ihre Anmutung, wie harmonisch die Figuren zueinander stehen und prüfe, wie breit oder schmal sie läuft. Ein Blindtext sollte möglichst viele verschiedene Buchstaben enthalten und in der Originalsprache gesetzt sein. Er muss keinen Sinn ergeben, sollte aber lesbar sein. Fremdsprachige Texte wie „Lorem ipsum“ dienen nicht dem eigentlichen Zweck, da sie eine falsche Anmutung vermitteln.

Abbildungsverzeichnis

4.1. Darstellung des aktuellen Extraktionsverfahrens und Integration der Visualisierung	16
---	----

Tabellenverzeichnis

Auflistung

3.1. Überblick über die Daten	6
---	---

Literaturverzeichnis

- [Espi 11] R. Espinosa Oliva, J. Zubcoff, and J.-N. Mazón. *A Set of Experiments to Consider Data Quality Criteria in Classification Techniques for Data Mining*. June 2011.
- [Scan 02] M. Scannapieco and T. Catarci. “Data Quality under a Computer Science Perspective”. *Journal of The ACM - JACM*, Vol. 2, Jan. 2002.
- [Soko 05] B. Sokol. “Living by numbers Leben zwischen Statistik und Wirklichkeit”. p. 149, 2005.
- [Wang 96] R. Y. Wang and D. M. Strong. “Beyond Accuracy: What Data Quality Means to Data Consumers”. *Journal of Management Information Systems*, Vol. 12, No. 4, pp. 5–33, March 1996.

Glossar

library A suite of reusable code inside of a programming language for software development. i

shell Terminal of a Linux/Unix system for entering commands. i