Einführung in die Datenanalyse

In der heutigen Zeit können Daten, beziehungsweise die Informationen darin, als wichtige Währung angeschaut werden. Praktisch überall fallen Daten an. Bei der Datenanalyse geht es darum, Daten in Erkenntnisse für bessere Geschäftsentscheidungen umzuwandeln.

In den Daten verbergen sich Geschäftseinblicke und Muster, die aufgedeckt werden sollen, um die eigenen Prozesse zu vereinfachen und zu verbessern. Die Datenanalyse beginnt mit den Fragen «Woher kommen die Daten?» und «Was soll aus den Daten gewonnen werden». In der Grafik wird der Lebenslauf der Datenanalyse verdeutlicht.

1. Finden Sie ein (geschäftliches) Problem, das Sie analysieren und lösen möchten.
2. Sammeln Sie Daten, die sich auf das Problem/Frage beziehen.
3. Bereinigen Sie die Daten und bereiten Sie sie für die Datenanalyse vor. Dieser Schritt wird auch als Data Scrubbing bezeichnet.
4. Analysieren Sie die Daten mithilfe von Modellen oder Algorithmen, um Trends und Muster zu erkennen.
5. Interpretieren Sie die Muster und leiten Sie Erkenntnisse ab.
6. Visualisieren Sie die Daten in Form von Diagrammen und Plots, um eine grafische Darstellung der Muster und Trends zu erhalten.

Risiken

Eine Datenanalyse enthält auch einige Risiken, die es zu beachten gilt. Fehlentscheide können bei Datenanalyseprojekten aufgrund der Qualität der erhobenen Daten auftreten. Obwohl es während des gesamten Lebenszyklus Risiken gibt, können die meisten davon der Datenerfassungsphase zugeschrieben werden.

Im Folgenden sind die Hauptursachen aufgeführt:

- Ausreisser: Ausreisser sind die Extremwerte, die ausserhalb der Normalverteilung der Daten liegen. Die Einbeziehung von Ausreissern als Teil eines Datensatzes könnte zu ungenauen Ergebnissen führen.
- Duplikate: Duplikate können aufgrund von Dateneingabefehlern auftreten. Duplikate führen bei statistischen Analysen zu falschen Erkenntnissen und müssen bei der Datenbereinigung beseitigt werden.
- Fehlende Werte: Fehlende Werte in den Daten stellen ein Risiko dar, wenn sie nicht behandelt werden. Fehlende Werte müssen durch Durchschnittswerte der Zeilen ersetzt werden. Manchmal müssen die fehlenden Werte vor der Analysephase ganz entfernt werden. Es muss jedoch darauf geachtet werden, dass die Datenintegrität gewahrt bleibt.
- Sicherheit: Ein Mangel an Datensicherheit kann zu grossen Risiken führen. Ungeschützte Daten können zu Missbrauch oder falscher Darstellung der Gesamtergebnisse führen. Die in den Datenanalyseprojekten verwendeten Daten müssen erkannt, autorisiert und genehmigt werden, bevor sie für die Analyse verwendet werden.
- Einhaltung von Vorschriften: Risiken der Nichteinhaltung von Vorschriften können durch nicht genehmigte Datenquellen, falsche Eigentumsverhältnisse, Probleme mit Kopierrechten usw. entstehen. Diese Risiken können zu rechtlichen Folgen führen.

Datenquellen

Ein wichtiger Schritt ist die geeignete Auswahl von Datenquellen. Klassisch werden drei Orte beigezogen:

- Interne Quellen: Die häufigste (und vermutlich auch wichtigste) Quelle der Datenerfassung sind bereits vorhandene Informationen. Zu diesen gehören unter anderem Aufzeichnungen von Kundeninformationen, die zum Beispiel in Kundenmanagementsystemen (CMS) gespeichert sind, Online-Anmeldeformulare, Logdaten aus Servern, Webanalysesoftware oder digitale Aufzeichnungen
- Öffentliche Datenquellen: Eine weitere Datenquelle sind öffentlich zugängliche Daten im Internet
- Nicht-traditionelle Datenquellen: Hierbei handelt es sich um eine alternative Datenquelle, die unter anderem in der Bank- und Finanzbranche üblich ist. Die Daten bestehen aus Anlagedaten, Berichten von Wertpapier- und Börsenaufsichtsbehörde usw.

Bei der Auswahl der Quellen gilt es immer den Datenschutz einzuhalten. Die Extraktion von Daten aus verschiedenen Quellen kann zu Problemen mit der Einhaltung von Datenschutzbestimmungen und dem Datenschutz führen. Daher gilt es schriftlich festzuhalten, wie die Daten erhoben wurden. Die Nichteinhaltung kann zu rechtlichen Folgen führen. Daher sollten Datenanalyseteams mit Rechtsexperten und Beratern zusammenarbeiten, um rechtliche Probleme im vorherein zu vermeiden.

Die Daten liegen in unterschiedlichen Formaten vor:

- Rohdaten: Daten in ihrem Anfangsstadium, die noch nicht verarbeitet oder manipuliert wurden, werden als Rohdaten bezeichnet. Diese Rohdaten müssen in strukturierte Daten formatiert werden, um genaue Ergebnisse zu erzielen.
- Strukturierte Daten: Daten, die in Form von Spalten und Zeilen dargestellt werden, nennt man strukturierte Daten. Dies können beispielsweise CSV-Dateien oder Excel-Tabellen sein. Oft liegen diese in Datenbanken. Strukturierte Daten sind viel einfacher zu durchsuchen und können durch Algorithmen effizient abgerufen und verarbeitet werden.
- Unstrukturierte Daten: Unstrukturierte Daten haben im Gegensatz zu strukturierten Daten kein Standardformat. Zum Beispiel gehören PDF-

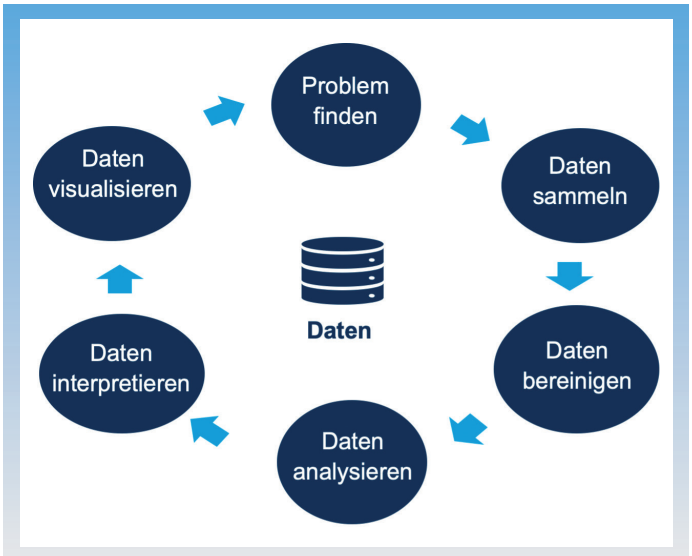

Bild: Andreas Wisler

Der Ablauf verdeutlicht den Lebenslauf der Datenanalyse.

ZUM AUTOR

Andreas Wisler, Dipl. Ing FH
goSecurity AG
Schulstrasse 11
CH-8542 Wiesendangen

T +41 (0)52 511 37 37
www.goSecurity.ch
wisler@gosecurity.ch

und Microsoft Word-Dokumente, Bilder, Videos, Audiodateien und E-Mails zu den unstrukturierten Daten.
– Big Data: Fortschritte bei der Speicherung, Geschwindigkeit und Leistung von Computern haben zu riesigen Datenbeständen geführt, die als Big Data bezeichnet werden. Diese gilt es ebenfalls in strukturierte Daten zu überführen.

Analyse

Wurden die Daten bereinigt und anschliessend in eine geeignete Form gebracht, folgt in einem weiteren Schritt die Auswertung. Bei der Datenanalyse geht es darum, «verborgene» Muster und Trends in den Daten zu finden und Vorhersagen zu treffen, die Unternehmen helfen, darauf aufbauend passende Entscheidungen zu treffen. Statistik ist dabei ein unverzichtbares Werkzeug für Datenanalysen. Die Statistik bietet verschiedene analytische Funktionen, die auf die Daten angewendet werden können, um Erkenntnisse zu gewinnen. Einige statistische Funktionen helfen bei der Zusammenfassung von Daten mit Funktionen wie Varianz, Mittelwert, Maximum, Modus oder Standardabweichung. Statistische Methoden können auch für Vorhersagen auf der Grundlage historischer Daten verwendet werden. Die Inferenzstatistik bietet die Möglichkeit, zukünftige Trends anhand von Stichprobendaten vorherzusagen. Eine weitere wichtige Funktion ist dabei Data Mining. Dies beinhaltet die Extraktion von Mustern in grossen Datenbeständen. Dabei werden verschiedene Techniken wie maschinelles Lernen, statistische Analyse und Mustererkennung eingesetzt. Dies wird auch als Knowledge Discovery in Databases (KDD) bezeichnet.

Visualisierung

Die Datenvisualisierung ist die Technik, um die Ergebnisse zu analysieren und sie den Beteiligten zu präsentieren, die dann Entscheidungen treffen und Strategien formulieren können. Im Rahmen der Datenvisualisierung werden Charts und Plots erstellt. Damit können beispielsweise Was-wäre-wenn-Analysen durchgeführt und die Daten auf explorative Weise untersucht werden. Zu den gängigsten Charts gehören:
– Linien-Diagramm: Liniendiagramme helfen bei der Visualisierung von Veränderungen einer Variable im Vergleich zu einer anderen, wie zum Beispiel Einkommen in verschiedenen Altersklassen oder Umsatz im Zeitverlauf.
– Balken-Diagramme: Balkendiagramme stellen Datenvariablen mithilfe von vertikalen oder horizontalen Balken dar und sind ideal für die Visualisierung diskreter Daten mit einer begrenzten Anzahl von Kategorien.
– Histogramm: Histogramme sind den Liniendiagrammen sehr ähnlich, enthalten jedoch keinen Rand oder Leerraum zwischen den Balken.
– Kreis-Diagramm: Kreisdiagramme stellen Variablenwerte im Verhältnis zu anderen Variablenwerten dar. Je grösser der Ausschnitt (denken Sie an ein Stück Pizza), desto grösser ist die Variable im Verhältnis zu den anderen Variablen.

Zusätzlich zu den Diagrammen werden auch verschiedene Plots verwendet:
– Streudiagramm: Das Streudiagramm (Scatter plot) wird zur Darstellung quantifizierbarer Beziehungen zwischen kontinuierlichen Variablen und zur Vermittlung von Informationen über die Beziehung zwischen diesen Variablen verwendet, zum Beispiel zur Identifizierung von Ausreissern und Anomalien, Datenvarianz, natürlichen Gruppen und Form (das heisst linear, nicht linear).
– Box-Diagramm: Das Box-Diagramm (Box plot) wird zur Zusammenfassung und Darstellung der Verteilung einer Reihe kontinuierlicher Daten verwendet. Sie sind nützlich für die Darstellung von Datenverteilung, Kürmmung und Ausreissern.
– Violin-Diagramm: Ähnlich wie die Box-Diagramme visualisieren Violinplots auch die Varianz der Daten, jedoch mit etwas mehr Details.
– Paar-Diagramm: Paardiagramme werden in explorativen Grafiken verwendet, um schnell Muster zwischen Variablen zu erkennen. Sie sind eine grossartige Möglichkeit, um herauszufinden, wie zwei gegebene Variablen miteinander verbunden sind.

Der Kreislauf einer Datenanalyse umfasst die Bestimmung der notwendigen Informationen, das anschliessende Sammeln und Bereinigen, bevor diese analysiert und die Ergebnisse visualisiert werden können. Das Resultat sollte die Antwort auf die zu Beginn definierte Frage sein. Mit diesem Vorgehen kann ein Unternehmen wichtige Erkenntnisse gewinnen um damit Prozesse verbessern oder vereinfachen zu können.