

Testversuch Masterarbeit Keller TUC

Business Case – Moderner Weinvertrieb:

In der Welt des Weinvertriebs ist es von entscheidender Bedeutung, potenzielle Kunden mit Bedacht auszuwählen und die Verkaufsstrategie entsprechend anzupassen. Vertriebsmitarbeiter stehen jedoch oft vor dem Problem unvollständiger oder inkonsistenter Daten potenzieller Kunden (Restaurants). David und Michael sind Vertriebsmitarbeiter desselben Weinguts „Weingut Bodensee“ und haben unterschiedliche Ansätze entwickelt, um das Datenproblem zu lösen.

David möchte mithilfe von maschinellem Lernen die unvollständigen und inkonsistenten Datensätze bereinigen, während Michael die traditionelle Art der Datenkorrektur verfolgt.

David vertritt die Ansicht, dass maschinelles Lernen nicht nur schneller, sondern auch skalierbarer in Bezug auf die benötigte Leistung sei. Während die manuelle Datenerfassung viel Zeit in Anspruch nimmt, kann er mit seinem Modell tausende Restaurants auf einmal analysieren. Zudem betonte er, dass das Modell ständig aktualisiert werden könne, um aktuelle Daten zu berücksichtigen.

Michael ist skeptisch. Er meint, dass es einfacher wäre, die Informationen händisch anzupassen, indem die Daten von ihm überprüft, bearbeitet und angepasst werden. Zudem sei sein Ansatz deutlich präziser und würde daher bessere Ergebnisse liefern.

Der Leiter des Weinguts entscheidet, dass beide Verfahren implementiert und anhand technischer sowie wirtschaftlicher Metriken und Verfahren evaluiert werden sollen. Dadurch soll das beste Verfahren ermittelt werden!

In diesem Szenario seid ihr Michael, welcher die traditionelle Art verfolgt, und versucht die Datenqualität mithilfe von euch gewählten Methoden zu verbessern!

Angaben Testperson:

- Beruf / Bildungsstand (Studiengang) (z. B. BW-Consultant, M.Sc. Wirtschaftsinformatik):
- Gute Kenntnisse in folgenden relevanten Programmiersprachen:
- Gute Kenntnisse in folgenden relevanten Tools:

Aufgabenstellung, Restriktionen, Ziel:

Aufgabenstellung:

Es existieren zwei Datensätze „Restaurants.csv“ und „Sales.csv“

- Die Datei „Restaurants.csv“ enthält Stammdaten; die Restaurantdaten inkl. der Adressen und der Information ob diese Alkohol ausschenken.
 - Es sollen doppelte Restauranteinträge gefunden und markiert werden. Es soll selbstständig herausgefunden in welchen Attributen die Doppelung erfolgt.
 - Es sollen die fehlenden Werte zum Attribut (Spalte) „Alcohol“ eingefügt werden. Mögliche Werte: „True“ oder „False“.
- Die Datei „Sales.csv“ enthält Transaktionsdaten; also z. B. die Bestellungen je Restaurant und Kunde. Diese enthält Bewertungen je Restaurantbesuch.
 - Es sollen die fehlenden Werte zum Attribut „OVERALLRATING“ eingefügt werden. Mögliche Werte: 0, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5.

Restriktionen:

- Es dürfen keine Machine Learning (ML) – Algorithmen eingesetzt werden.
- Es dürfen alle Tools und Programmiersprachen verwendet werden (bitte angeben unter Messwerte).

Ziel:

Es soll überprüft werden welche der beiden Verfahren schneller und präziser ist.

Hierzu werden die Ergebnisse beider Verfahren gegenübergestellt und z. B.

folgende Fragen beantwortet:

- Wie genau sind die von ML gefundenen Werte im Vgl. zu den von den Testpersonen gefundenen Werte?

- Wie lange dauerte die Implementierung beider Verfahren
- ...

Solche und weitere Fragen sollen mit den Ergebnissen beantwortet werden.

Daraus kann anschließend abgeleitet werden ob es sich lohnt ML im Datenqualitätsmanagement einzusetzen bzw. zu implementieren und an welchen Stellen Hürden und Chancen auftreten.

Messwerte:

- Eingesetzte Programmiersprache (z. B. Python):
- Eingesetzte Tools (z. B. Excel):
- Gesamtdauer in h (von Anfang bis Ende, d.h. Analyse, Implementierung, eigene Evaluation der Ergebnisse):
- Persönliche Schätzung der Genauigkeit in % (z. B. Ich glaube ich habe 30% der Duplikate gefunden, 70% richtig, ob ein Restaurant Alkohol ausschenkt und 50% richtig bei dem Rating):
 - Duplikate:
 - Attribut Alcohol:
 - Attribut OVERALLRATING:

Ergebnisse als Datei (z. B. .csv oder ähnlich) bitte mitschicken.

Vorgehen:

Bitte kurz in Stichworten das eigene Vorgehen bzw. Umsetzung beschreiben

Mögliche Fragen:

- Wurde eine spezielle Strategie eingesetzt?
- Welche Phase hat am längsten gedauert?
- Wo gab es Probleme? Was war schwierig, was war einfach?
- ...

Vielen Dank euch schonmal vorab!