Seminararbeit AI-assisted programming and data analysis
Data Profiling with LLM's a Case Study on HR Data

Vincent-Konstantin Kapp[a,1], Prof. Dr. Martin Spindler[b,2]

[a] *Universität Hamburg, M.Sc. BWL, Street Address, Hamburg, Postal Code*

[b] *Universität Hamburg, Statistik mit Anwendung in der Betriebswirtschaftslehre, Street Address, Hamburg, Postal Code*

Abstract

Datengetriebene Prozesse können nicht nur Unternehmen beim Optimieren der Prozesse sondern auch bei der entwicklung neuer Strategien helfen. Aber uum festzustellen welche Daten in den Datenbanken vorliegen und wie diese zu einernander stehen braucht es Mitarbeiter*innen die einzelnd überprüfen welche Daten es gibt und in welchen zusammenhängen, ausgehend von Commen Sense, es gibt. Data Profiling stellt Unternehmen mit historisch gewachsene Datenbanksysteme vor Herausforderungen für die Dategetriebene Transformation da. Können die aktuellen entwicklunge von Large Language Modellen (LLM) dazu beitragen, autamtisiert die zusammenhängen von Spalten zu erkennen? Um diese Frage zu beantworten, widmet sich diese Seminararbeit aufbauen von der Arbeit von Trummer 2024, wie gut LLM, anhand einer Case Study, die Korrelation von Daten zu erkennen.

Keywords: LLM, Data Analysis, Data Profiling

Seminararbeit

Philipp Bach, Victor Chernozhukov, Martin Spindler (2024). Heterogeneity in the U.S. Gender Wage Gap. Journal of the Royal Statistical Society: Series A, 187(1), 209-230, available online.

Philipp Bach, Victor Chernozhukov, Martin Spindler, Closing the U.S. gender wage gap requires understanding its heterogeneity, Working Paper, available at arXiv, 2018.

*Corresponding author

Email addresses: vincent.kapp@studium.uni-hamburg.de (Vincent-Konstantin Kapp), martin.spindler@uni-hamburg.de (Prof. Dr. Martin Spindler)

[1]This is the author.

[2]This is the Editor

Sven Klaassen, Jan Teichert-Kluge, Philipp Bach, Victor Chernozhukov, Martin Spindler, Suhas Vijaykumar, DoubleMLDeep: Estimation of Causal Effects with Multimodal Data, available at arxiv, 2024.

(**?**)

1. Einleitung

2. Hintergrund

2.1. Data Profiling

2.2. Korrelation ermittelnt

2.3. Sprachmodelle

2.4. NLP for Data Bases

2.5. Benchmark

2.6. Benchmark Data

2.7. Benchnmark Metrics

3. Benchmark Analysis

4. Comparing Prediction Methods

4.1. Description of Methods

4.2. Experimental Setup

4.3. Comparison Results

5. Scenario Variants

6. Results Breakdown

7. Other Correlation Metrics

8. Column Types

9. Conclusion

References