TU Clausthal

Abschlussvortrag Masterarbeit Jordis Thibaut Kengne

„Definition und Umsetzung eines robusten MLOps Prozesses am Beispiel von wechselnden Datenformaten von Fahrzeugflottendaten“

Das Thema dieser Arbeit ist die Definition und Umsetzung eines robusten MLOps Prozesses am Beispiel von wechselnden Datenformaten von Fahrzeugflottendaten. Der Entwicklungslebenszyklus von Machine-Learning-Modellen variiert stark und dieser dauert abhängig von der Problemstellung, dem Modell und unternehmerischen Kontext teilweise sehr lange und viele solcher Anwendungen für zum Beispiel Vorhersagen schaffen es langsam in die Produktion, wenn überhaupt. Der Entwicklungslebenszyklus von Modellen wird heutzutage noch manuell durchgelaufen, deswegen bleiben viele Modelle immer noch in der Experimentationsphase. In dieser Arbeit wird der Prozess so definiert, dass Robustheit und semi-automatismus, bei sonst manuellen Aufgaben gefördert wird . Der Prozess wird so definiert und adaptiert, dass der Lebenszyklus zweistufig durchlaufen wird. Zünachst durchläuft man die Schritte 0 unter Anwendung des CRISP-DM (cross industry standard processes for data mining) Prozesses, bis ein Modell enstanden ist, das heißt es wird von der Datenvalidierung, zur Parameterauswahl bis hin zu der Modellevaluation eine Stufe 0 definiert. Danach wird in der Stufe 1, die gewonnenen Erkenntnisse der Stufe 0 als Komponente, so ausgelargert und mit Entwicklerfeeback definiert, dass einzeln ausführbare Funktionen für die Prozessphasen entstehen, die sich in einem Script so definieren lassen, dass sie (semi-)automatisiert ausführbar sind. Es wird dann eine Schnittmenge zwischen den Erkenntnissen der Stufe 0 und den anderen Daten aus der Datenmenge in dieser Arbeit als Beispiel aus einer Fahrzeugflotte, gebildet. Erkenntnisse werden in der Prozessdefinition so eingebaut, dass der Prozess adaptiv und lernend ist. Es werden Dateien entstehen, um das Gelernte zu speichern und es wird auch um Entwicklerfeedback an nötigen Stellen im Prozess gebeten. Am Ende ist man in der Lage Datensätze als einen Datensatz zu kombinieren trotz Unterschiede in Ihrer Beschaffenheit wie zum Beispiel Schwankungen vom Wertebereich und Parameternamenänderungen. Der Datensatz wird dann für die weitere Analyse benutzt, um das konkrete Ziel, zu erreichen.

Datensätze nach den Parametern und Eigenschaften für alle Daten in der Menge finden zu beurteilen, wird mit Entwicklerfeedback unterstützt. So kann ein Modell für eine Menge von Daten in kürzere Durchlaufzeiten entstehen, was die Effizienz und Produktivität in dem Kontext von Datenanalyse hier steigern kann, denn wiederholte Aufgaben werden vermiden und diese Zeit dann für den Verbesserung von den Phasen zur Zielerreichung eingesetzt.

Betreuer der Arbeit:	Prof. Dr. Steffen Herbold, Prof. Dr. Rüdiger Ehlers
Datum:	Mittwoch, 22. Juni 2022, 9:30 Uhr
Ort:	Online-Meeting über BBB
	Link: https://webconf.tu-clausthal.de/b/ste-w8t-1pf-sml