

Machine Learning Canvas

Product:

Authors:

Date:

Version:

<div><div>Background</div><div><div>Describe the customer's goals and pains.</div><div>Die Nachfrage im städtischen Verkehr unterliegt starken zeitlichen Schwankungen. Für Taxiunternehmen oder Mobilitätsplattformen ist es daher wichtig, die Auslastung möglichst genau vorherzusagen. Historische Fahrtdaten der Yellow Cabs in New York City bieten eine solide Grundlage, um typische Nachfrageverläufe stundenweise zu modellieren.</div></div></div> <div><div>Value proposition</div><div><div>Propose the product with the value it creates and the pains it alleviates.</div><div>Ein datenbasiertes Vorhersagemodell kann helfen, betriebliche Entscheidungen (z. B. Fahrzeugverteilung, Personalplanung) besser abzustimmen. Auch andere städtische Akteure können von einem besseren Verständnis des Mobilitätsverhaltens profitieren, etwa in der Verkehrsplanung oder beim Infrastrukturmanagement.</div></div></div> <div><div>Objectives</div><div><div>Breakdown the product into key objectives that need to be delivered.</div><div>Ziel ist es, ein ML-Modell zu entwickeln, das die Anzahl an Fahrten pro Stunde prognostiziert. Dabei stehen neben der reinen Modellleistung auch die strukturelle Umsetzung, Tool-Nutzung (MLflow, Streamlit) und die Analyse möglicher Veränderungen im Datenverhalten (Data/Concept Drift) im Fokus.</div></div></div>	<div><div>Solution</div><div><div>Define the solution, including features, integration, constraints and what's out-of-scope</div><div>Die Lösung besteht aus einem Regressionsmodell auf Basis aggregierter Zeitfenster (Stunden-Buckets), ergänzt um eine Benutzeroberfläche zur interaktiven Nutzung. Die Umsetzung umfasst die vollständige ML-Pipeline: Datenaufbereitung, Feature Engineering, Modelltraining, Logging, Evaluation und Interface.</div></div></div> <div><div>Feasibility</div><div><div>Discuss the feasibility of the solution and if we have the required resources.</div><div>Die technischen Anforderungen sind durch die Nutzung etablierter Open-Source-Tools gut umsetzbar. Das Projekt wird zunächst lokal durchgeführt und in einem Deployment dann zumindest in Teilen öffentlich bereitgestellt. Die Datengrundlage ist öffentlich verfügbar. Grundkenntnisse in Datenanalyse, ML und Python sind vorhanden. Einschränkungen ergeben sich v. a. durch Rechenzeit und Speicherkapazität.</div></div></div>	<div><div>Data</div><div><div>Identify the training and production data sources, as well as the labeling process and decisions.</div><div>Verwendet werden Yellow-Taxi-Fahrtdaten der Jahre 2009–2017. Die Aggregation erfolgt auf Stundenebene. Für jede Zeiteinheit werden zentrale Merkmale wie Trip-Anzahl, Strecke, Passagiere, Fahrzeit und Preis berechnet. Ein manuelles Labeling ist nicht notwendig, da die Zielvariable (Trip Count) direkt vorliegt.</div></div></div> <div><div>Metrics</div><div><div>Prioritize key metrics that reflect the objectives.</div><div>Zentrale Metriken sind MAE, RMSE und R² zur Bewertung der Modellgüte. Zusätzlich wird beobachtet, wie stabil die Metriken über verschiedene Zeiträume hinweg bleiben (Robustheit gegenüber Drift)..</div></div></div> <div><div>Evaluation</div><div><div>Design offline and online evaluation criteria.</div><div>Offline-Evaluation, bzw. später auch im Deployment in reduzierter Form vorhanden, durch Aufteilung in Trainings- und Testzeiträume (2009–2014 / 2015–2017). Modelle werden über die Zeit hinweg verglichen, um etwaige Verschiebungen im Datenverhalten zu erkennen und einzuordnen.</div></div></div>	<div><div>Modeling</div><div><div>List the iterative approach to model our task.</div><div>Zum Einsatz kommen verschiedene Regressionsverfahren (z. B. Linear Regression, Random Forest, XGBoost). Die Modellwahl erfolgt iterativ auf Basis von Vergleichsexperimenten, die mithilfe von MLflow dokumentiert werden. Feature-Auswahl und Hyperparameter werden systematisch angepasst und evaluiert.</div></div></div> <div><div>Inference</div><div><div>Decide whether we want to do batch (offline) or real-time (online) inference.</div><div>Die Anwendung erfolgt im Batch-Modus: Für vorliegende Stunden-Features werden stündliche Nachfragewerte (Trip Counts) auf Basis aggregierter Merkmale berechnet. Das Modell wird dabei nicht kontinuierlich, sondern abschnittsweise auf vorbereitete Daten angewendet – in diesem Fall auf (Mehr-)Jahresebene. Eine Echtzeitverarbeitung ist im Rahmen des vorliegenden Projekts nicht vorgesehen. Dies liegt nicht nur an den infrastrukturellen Anforderungen, sondern vor allem an der zugrunde liegenden Datenquelle: Die verwendeten Yellow-Cab-Daten stehen ausschließlich retrospektiv als Open Data zur Verfügung und bilden abgeschlossene Zeiträume ab. Ein Zugriff auf Echtzeitdaten ist damit strukturell nicht möglich.</div></div></div>	<div><div>Feedback</div><div><div>Outline sources of feedback from our system to use for iteration.</div><div>Feedbackquellen sind vorrangig technischer Natur: Modellmetriken, Vergleich über Zeiträume hinweg sowie visuelles Feedback über die Oberfläche. Potenzielle Drift-Anzeichen werden durch abnehmende Modellgüte im Evaluationszeitraum identifiziert.</div></div></div> <div><div>Project</div><div><div>Define the required team members, deliverables and projected timelines.</div><div>Das Projekt wird im Rahmen des Moduls MLOps-Seminars umgesetzt und in Einzelarbeit durchgeführt. Ziel ist die Entwicklung eines Prototyps zur stündlichen Vorhersage von Taxinachfrage auf Basis historischer NYC Yellow Cab Daten. Die Umsetzung umfasst sowohl technische als auch konzeptionelle Aspekte, mit besonderem Fokus auf Reproduzierbarkeit, Transparenz und Tool-Einsatz entlang des Machine-Learning-Lebenszyklus. Alle Arbeitsschritte – von der Datenaufbereitung bis zur Modellbereitstellung – werden nachvollziehbar in einem GitHub-Repository dokumentiert. Grundlage für die Dokumentation ist jedoch das vorliegende Notebook. Dabei kommen verschiedene MLOps-Praktiken zum Einsatz, bspw. für das Logging (MLflow), das User Interface (Streamlit) sowie zur Strukturierung des Workflows. Der Projektverlauf ist in sieben aufeinander aufbauende Phasen unterteilt: Business Requirements (Stichtag: 23.03.): Definition der Problemstellung, Zielsetzung und Nutzenanalyse mithilfe des ML Canvas. Daten (Stichtag: 30.03.): Beschaffung, Verständnis und erste Analyse der NYC Yellow Taxi Daten (Jahr 2024). Feature Engineering (Stichtag: 06.04.): Transformation der Rohdaten in geeignete Modellmerkmale. Modellentwicklung (Stichtag: 20.04.): Auswahl, Training und Optimierung eines Regressionsmodells. Test (Stichtag: 27.04.): Validierung der Modellleistung mit geeigneten Metriken und Daten. Deployment (Stichtag: 11.05.): Bereitstellung des Modells mithilfe einer MLOps-Plattform oder API. Monitoring & Lessons Learned (laufend): Beobachtung des Modells im Betrieb, Reflexion über den Entwicklungsprozess sowie Dokumentation der wichtigsten Erkenntnisse.</div></div></div>
--	--	--	--	---

