Auf einen Blick: Vorgehensweise bei einem ML-Projekt

Hier in diesem Dokument erhalten Sie die grundlegenden Schritte, die im Rahmen eines ML-Projekts durchlaufen werden und bekommen darüber hinaus Anregungen, wie beim jeweiligen Schritt zu verfahren ist.

Diese Liste ist ein Vorschlag, der zur Strukturierung eigener Projekte dienen soll. Sie ist dabei keineswegs erschöpfend oder allgemeingültig anwendbar. Einer der wichtigsten Faktoren ist und bleibt das Verständnis der eigenen Prozesse und Anforderungen.

1. Auf das große Ganze schauen

Bevor auch nur eine erste Zeile Code geschrieben, bevor auch nur ein erster Blick auf die Daten geworden wird, sollte zunächst ein Blick auf das Gesamtbild des Problems geworfen werden. So kann sichergestellt werden, dass alle beteiligten Parteien von denselben Annahmen ausgehen. In diesem Schritt sollte genau kommuniziert und identifiziert werden, was das Problem/die Aufgabe ist, wie die entwickelte Lösung weiterverwendet werden wird und welche Vorannahmen es hinsichtlich des Problems und/oder der potenziell entwickelten Lösung gibt.

Framing des Problems

1. Was genau ist das Ziel? Was soll erreicht werden?
2. Wie sieht die aktuelle Lösung aus? Gibt es wiederverwendbare Ansätze?
3. Weitere Vorüberlegungen:
   - Handelt es sich hier um ein Problem, was mit überwachtem/nicht-überwachtem/bestärkendem Lernen gelöst werden kann?
   - Ist es ein Klassifizierungsproblem? Eine Regression? ...
   - Ist eine Online- oder Offlinelösung besser?

Auswahl der Performanzmetrik

1. Wie wird der Erfolg meines Modells bemessen?
   - z.B. RSME (Root Mean Square Error), MAE (Mean Absolute Error), ...

Überprüfen der Vorannahmen

1. Welche Vorannahmen über das Problem habe ich getroffen?
2. Welche Vorannahmen über das Problem haben andere getroffen?
3. Überprüfen der Vorannahmen
   - ***Beispiel**: Ich gehe davon aus, dass mein Modell dazu verwendet wird, exakte Aussagen darüber zu treffen, wie belastbar ein Werkstück sein wird, dh. ich entwickle mein Modell so, dass es exakte, z.B. numerische Werte*

ausgibt (beispielsweise auf einer Skala von 0-100). Es stellt sich aber heraus, dass die Anforderung der anderen nur ist, dass gröbere Wertebereiche zusammengefasst werden (z.B. nicht belastbar, wenig belastbar, belastbar, sehr belastbar). Daraus kann ich schlussfolgern, dass eine exakte Wertvorhersage nicht notwendig ist.

2. Datenakquise

Der wichtigste Teil eines Machine Learning Projekts sind die Daten. Darum sollte die Datenakquise vorsichtig vorgenommen werden. In diesem Schritt geht es zunächst darum, zu definieren, welche Daten überhaupt benötigt werden und woher man diese bekommt. Gegebenenfalls müssen hier erst Authorisierungen (z.B. für eine Datenbank) eingeholt werden. Anschließend sollte man sich kurz mit den vorhandenen Daten vertraut machen. Dies wird in einem weiteren Schritt weiter vertieft. Zuletzt wird bereits hier ein Testset erstellt.

An die Daten herankommen

1. Welche Daten benötige ich zur Lösung des Problems?
2. Wo bekomme ich diese Daten her?
3. Daten herunterladen

Vertrautmachen mit den Daten

1. Erster Blick auf die Daten, z.B. mit Pandas
2. Um was für Daten handelt es sich?
3. Ggf. Entfernen oder Anonymisieren sensitiver Daten

Erstellen eines Testsets

1. Entfernen einer Teilmenge des Datensets. Dieses Testset kommt später wieder zum Einsatz.

3. Explorieren & Visualisieren der Daten

Nun sollte man sich genau mit den vorhandenen Daten vertraut machen. In diesen Schritt sollte ein großer Teil der Zeit investiert werden. Nur wenn man die Daten wirklich kennt, ist man in der Lage, eine sinnvolle Lösung zu finden.

Erstellen einer Datenkopie

- Um das originale Datenset nicht versehentlich zu verfälschen, sollten immer Kopien erstellt werden, die anschließend manipuliert werden können.

Datenexploration

1. Welche Features beinhaltet mein Datenset?
2. Wie viele Instanzen existieren?
3. Gibt es Noise, Outlier, etc.?

Visualisieren

1. Plotten der Daten
2. Korrelationsanalyse

Feature-Engineering

- optional: oft ist es sinnvoll, verschiedene Features zu kombinieren, da diese Kombination eine höhere Aussagekraft hat, als die einzelnen Features.
- ***Beispiel****: Ein Algorithmus soll den Preis für Grundstücke vorhersagen. Als Features sind diesbezüglich im Datensatz unter anderem die Länge und die Breite des Grundstücks aufgeführt. Obwohl mit diesen Features separat gearbeitet werden könnte, bietet es sich doch an, Länge und Breite zur Gesamtfläche des Grundstücks zu kombinieren, die diese eine höhere Aussagekraft hat als beide Features für sich genommen.*

4. Datenaufbereitung

Nachdem die Daten gesichtet wurden, müssen sie noch für den Einsatz in einem Machine Learning Verfahren vorbereitet werden. Die grundlegenden Schritte zur Datenaufbereitung wurden bereits im **Modul *Daten & KI*, Kapitel: *Garbage in, Garbage out - warum die Qualität der Daten*** so wichtig ist besprochen.

- **Reinigen der Daten**
- **Transformieren**
- **Reduzieren**

Wichtig ist anzumerken, dass insbesondere bei der Datenaufbereitung darauf geachtet werden sollte, so viel wie möglich zu automatisieren. Das stellt sicher, dass ein Updaten des fertigen Modells mit neuen Daten so unkompliziert wie möglich verläuft.

5. Auswahl und Training des Modells

Nachdem die Daten vorbereitet wurden, muss noch das passende Modell gefunden werden. Idealerweise trainiert man nicht nur ein einziges Modell, sondern wählt eine Reihe verschiedener Modelle aus die dann jeweils trainiert werden. In einer

anschließenden Analyse werden Performanz und Fehler der Modelle genauer betrachtet. Zuletzt werden die besten Modelle ausgewählt.

Training von plausiblen Modellen

1. Welche Modelle könnten das Problem potenziell gut lösen
2. Trainieren einer Reihe von verschiedenen Modellen

Analyse

1. Analysieren der Modellperformanz
2. Was sind die Gründe für Fehler?
3. Wie kann man die Fehler potenziell beheben? Fehlen Daten?

Auswahl

- Auswählen der 3 bis 5 besten Modelle und Verwerfen der anderen

6. Finetuning des Modells

In diesem Schritt geht es darum, die Hyperparameter auf ihre bestmöglichen Werte zu bringen. Viele Machine Learning APIs bieten hierfür Hilfestellungen oder Automatisierungsmöglichkeiten. Sobald die Hyperparameter entsprechend eingestellt sind kann das fertige Modell, bzw. das Modellensemble am Testset validiert werden.

7. Launch des Modells in die Produktion

Nachdem das Modell fertiggestellt wurde muss es nur noch in die Produktionsumgebung integriert werden. Hier ist die Kollaboration mit den Expert*innen der Produktion gefragt.

8. Monitoring & Wartung

Mit der Übergabe des Modells in die Produktion ist die Arbeit noch nicht getan. Machine Learning Modelle müssen konstant überwacht und ggf. gewartet werden. Mit zunehmendem Alter der Datengrundlage neigen Modelle dazu, eine schlechtere Performanz zu zeigen. Außerdem müssen externe Bedingungen mit überwacht und in Betracht gezogen werden. Auch hier gilt: so viel Automatisierung wie möglich!

Monitoring

1. Code zur Überwachung des Systems
2. Wird mein Modell mit der Zeit schlechter?
3. Monitoring der Produktionsumgebung. Funktionieren alle Sensoren? Hat sich in der Umgebung grundlegend etwas verändert?

Wartung

- Aktualisieren der Datengrundlage bzw. Anpassen des gesamten Modells falls es eine sich ändernde Umgebung erfordert