

# 1 Zusammenfassung

In der folgenden Masterarbeit dreht es sich um die Frage, inwiefern sich die Digitalisierung allgemein auf konventielle Berufe und im Speziellen auf den Beruf des Kriminalanalysten auswirkt. Die im Zuge der Digitalisierung häufig angewandten maschinellen Lernmethoden besitzen dabei das Potential, riesige Datenmengen in einer angemessenen Zeit nach Mustern zu durchsuchen und dabei Hinweise zu geben, wie einer Problematik entgegen gewirkt werden kann. So könnten diese Analysen auch bei der Ergreifung von Verbrechern behilflich sein (*Predictive Policing*).

Bereits seit den 20er-Jahren wird versucht, die Wahrscheinlichkeit der Rückfälligkeit von Tätern bei möglichen Bewährungsstrafen miteinzubeziehen und diese zu dadurch zu klassifizieren. Auch bei der Ermittlung eines angemessenen Strafmaßes wird diese Wahrscheinlichkeit in Betracht gezogen. Die daraus abzuleitende Hauptfrage der folgenden Arbeit wird es sein, inwiefern sich alte Methoden (bspw. *logistische Regression*) durch neuartige (bspw. *Random Forests*) ergänzen bzw. gänzlich ersetzen lassen.

Auch um die Frage einer angemessenen Veranschaulichung des Modells wird es in der folgenden Arbeit gehen, da die Resultate für Personen mit Entscheidungsgewalt (*Stakeholder*) so aufschlussreich wie möglich dargestellt werden sollten, um diesen die Legitimation ihrer Entscheidung zu vereinfachen. Dabei wird es zu Limitationen kommen, da sich durch maschinelle Lernmethoden häufig Umstände ergeben, die ein genaues Erfassen der Vorgehensweise dieser Methoden erschweren und so Rückschlüsse auf einzelne Variablen nur bedingt möglich sein werden, sodass die Vor- und Nachteile dieser neuartigen Methoden vor dem Hintergrund dieser Intransparenz kritisch aufgewogen werden müssen.

Wichtige Angelegenheiten wie bspw. der Datenschutz und die damit einhergehende Wahrung der Privatsphäre müssen im Folgenden zusätzlich diskutiert werden, da es den Menschen Unbehagen bereitet, wenn mit ihren Daten nicht sorgfältig und anonym umgegangen wird. Die Frage, was dabei rechtens ist und wie sich Gesetze im Bereich der Digitalisierung anpassen bzw. entwickeln sollten, um die perfekte Balance zwischen Persönlichkeitsrechten und öffentlicher Sicherheit zu finden, wird vom Autor ebenfalls berücksichtigt. So soll letztendlich mit dieser Arbeit dazu beigetragen werden, den Polizeiämtern im Hinblick auf begrenzte, personelle Ressourcen eine Hilfestellung bei der Distribution dieser Ressourcen zu geben. Straftaten zu verhindern ist dabei in jeglicher Hinsicht kostengünstiger, als sie im Nachhinein aufzuklären.

## **2 Einleitung**

### **3 Grundlagen und Notwendigkeit**

## 4 Maschinelle Lernmethoden zum Forecasting

### 4.1 Klassifikation

- Random Forest (kein Overfit/ adaptiv/ Rückschlüsse auf einzelne Variablen durch den 'Increase in Forecasting Error/ Partial Dependence Plot' möglich (Berk, 2013)),
- Neurales Netzwerk (Overfit), stochastic gradient boosting (Overfit/ adaptiv), Logistische Regression (nicht adaptiv, nur zwei Outputklassen)
- wichtig, dass nicht Apples mit Oranges verglichen werden (siehe Tuning-Parameter der einzelnen Methoden)
  - muss vergleichbar sein
- allgemein: siehe Hastie, Perry

#### 4.1.1 Linearität und Nichtlinearität

- je mehr Nichtlinearität, desto genauer, jedoch Gefahr des Overfittings

#### 4.1.2 Behandlung der Forecast-Errors

- sollten diese gleich behandelt werden bei Klassifikationen?
- Betrachtung der Auswirkung einer Fehleinschätzung
- wie könnte man Fehler unterschiedlich gewichten? Bestimmung der Cost Ratio (False Negatives versus False Positives)

### 4.2 Regression

- Neurales Netzwerk
- allgemein: siehe Hastie, Perry
- wichtig, dass nicht Apples mit Oranges verglichen werden (siehe Tuning-Parameter der einzelnen Methoden)
  - muss vergleichbar sein

#### 4.2.1 Linearität und Nichtlinearität

- je mehr Nichtlinearität, desto genauer, jedoch Gefahr des Overfittings

### 4.3 Empirie

- Datensets: Kriminaldatensätze aus Kaggle (San Francisco, Chicago)
- > welche Variablen sind vorhanden? Sind diese relevant? Einschätzung!
- Auswertungen mit Confusion Table bei Klassifikationsproblemen
- sind Ergebnisse relevant (in absoluten Zahlen?)

## 5 Diskussion, Limitationen und Rechtliches

- auf was sollte stets bei der Datenbeschaffung geachtet werden?
- auf fragliche Paragrafen eingehen? oder zu rechtlich?
- auf neu geschaffenes 'Digitalministerium' eingehen
- Facebook-Skandal mit Cambridge Analytica (auch in Einleitung, um Aktualität hervorzuheben)

## 6 Schluss