

Machine Learning 101

- an introduction

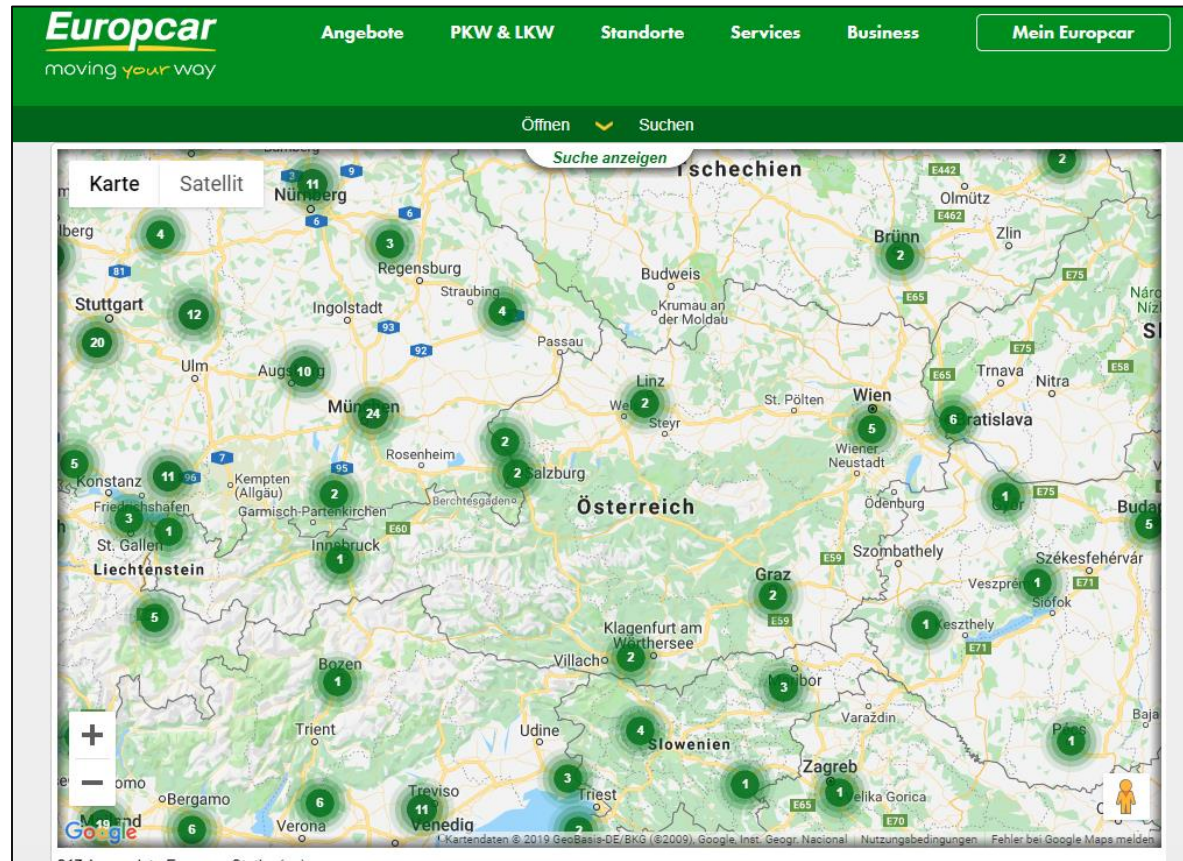
Inhalt

- Wer bin ich / wer sind wir?
- Machine Learning?
- Problemstellungen & Lösungsmöglichkeiten 😊
 - Wie ähnlich bzw. unähnlich sind Menschen?
 - Wie kann man vorhersagen, wie viele Besucher zum nächsten DEVTREFF kommen werden?
- Ausblick

Wer bin ich/wer sind wir?

- Laurenz Gröbner
 - Ausbildung
 - HTL Hochbau
 - BWL, Spezialisierung auf Operations Management
 - Diplomarbeit
 - Fuhrparkplanung in der Autovermietung
 - Preis- und Mengensteuerung
 - ~~Doktorat~~
 - Geschäftsführer von helloCash

Fuhrparkplanung in der Autovermietung?



Fuhrparkplanung in der Autovermietung?

PKW

LKW / Transporter

Filter minimieren

Filter zurücksetzen

Welches Fahrzeug benötigen Sie?

☐ Most Popular

☐ Klein

☐ Groß

☐ Automatik

☐ Luxus

☐ SUV & Allrad

☐ Family & Kleinbusse

☐ Transporter/LKW

Fahrzeugkategorie
Alle

Personen

Gepäck

Türen

CO₂
Emissions-Stufe

Transmission

Flotte

Wir haben 60 Ergebnisse gefunden.



FIAT 500

Mini (MCMR)

4 Personen

1 Gepäck

2 Türen

B

CO₂ 110 g/km

M

Manuell

Klimaanlage

jetzt Kategorie buchen



ABARTH 695C

Mini (NTMR)

4 Personen

1 Gepäck

2 Türen

D

CO₂ 155 g/km

M

Manuell

Klimaanlage

jetzt Kategorie buchen



VW POLO

Economy (EDMR)

5 Personen

1 Gepäck

5 Türen

B

CO₂ 106 g/km

M

Manuell

Klimaanlage

jetzt Kategorie buchen

Fuhrparkplanung in der Autovermietung?

- Erklärende Variablen (x_1, x_2, \dots)
 - Wintersaison
 - Feiertage
 - Bereits bekannte Reservierungen
 - Vergangene Nachfrage
 - Mietdauern (zwischen 2 Stunden und 30 Tagen)
 - ...
- Erklärte Variable (y)
 - Benötigte Fahrzeugmenge pro Kategorie x in Station y zum Zeitpunkt z

Fuhrparkplanung in der Autovermietung?

- (Predictive-Analytics)-Modelle
 - Prognosemodelle
 - Optimierungsmodelle
- Programmiersprache
 - SAS
- Ergebnisse



Predictive Analytics

Optimal in Fahrt

Wie Österreichs größte Autovermietung Europcar mit Hilfe der SAS Software ihr Flottenmanagement optimiert.

sorgen Europcar-Geschäftsführer Erich Windisch und sein Team dafür, dass alles reibungslos funktioniert und sämtliche Kundenwünsche erfüllt werden. Seit Ende des Vorjahres liefert dafür die SAS Software den optimalen logistischen Treibstoff.

„Unser wichtigstes Ziel ist es, das richtige Auto zur richtigen Zeit am richtigen Ort zu haben – und das natürlich immer mit Blick auf Kosten und Erträge.“ Erich Windisch, Geschäftsführer für Österreich und die Slowakei, bringt das Kernthema auf den Punkt. Wobei das, was aus dem Mund des Europcar-Chefs so simpel klingt, keineswegs einfach zu bewerkstelligen ist. Schließlich müssen unzählige Transaktions- und Kundendaten so gezielt analysiert werden, dass alle Europcar Stationen mit genau jenen Fahrzeugen bestückt sind, die auch tatsächlich von den Kunden vor Ort benötigt werden.

Bei einem Fuhrpark mit rund 2.000 Fahrzeugen, die in 20 Gruppen nach Größenordnung geclustert sind, ist dies eine durchaus komplexe logistische Herausforderung. Langfristige Reservierungen gilt es dabei ebenso abzudecken wie kurzfristige Nachfragen, die oft erst einen Tag vor dem gewünschten Datum eintreffen. Hinzu kommen unterschiedlichste Tarifmodelle, die nicht zuletzt auch davon abhängig sind, wie lange im Vorhinein ein Kunde bucht.

„Es geht um Fahrzeuge, Kunden und Tarife. Eine Unmenge an Daten also, die so abgebildet werden müssen, dass sie uns in die Lage versetzen, die Planung zu verbessern und die Erträge zu optimieren – und das immer mit größtem Fokus auf die Kundenzufriedenheit“,

Logistik erfolgreich optimiert

(von links): Europcar Österreich Geschäftsführer Erich Windisch, Yield Manager Laurenz Gröbner und Peter Pfeiszl, zentrale Fuhrparksteuerung

6 Jahre später kam ... helloCash



Wir arbeiten europaweit

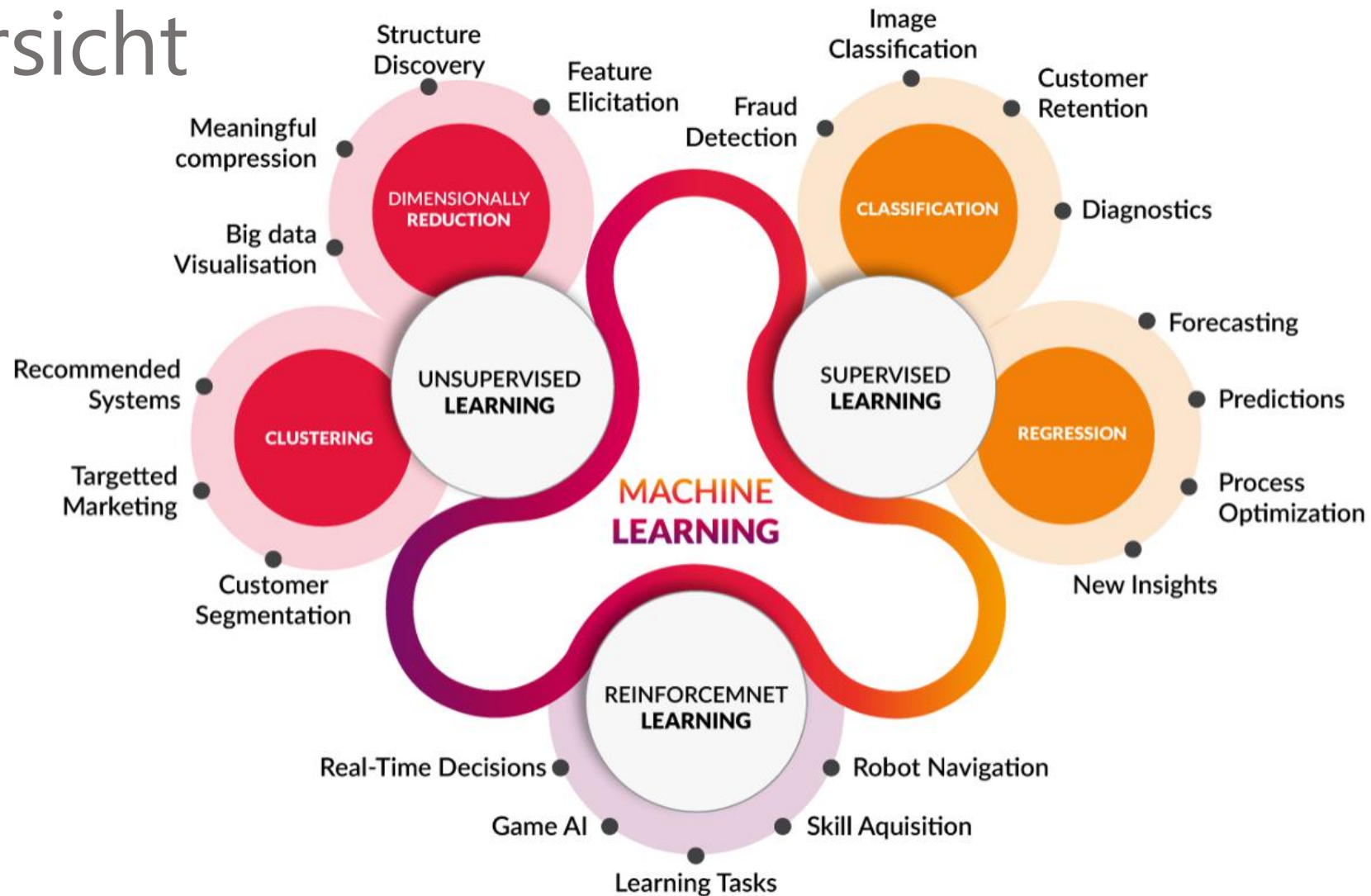
6 Länder

Österreich (Headquarter)	Polen
Deutschland	Spanien
Tschechien	Frankreich

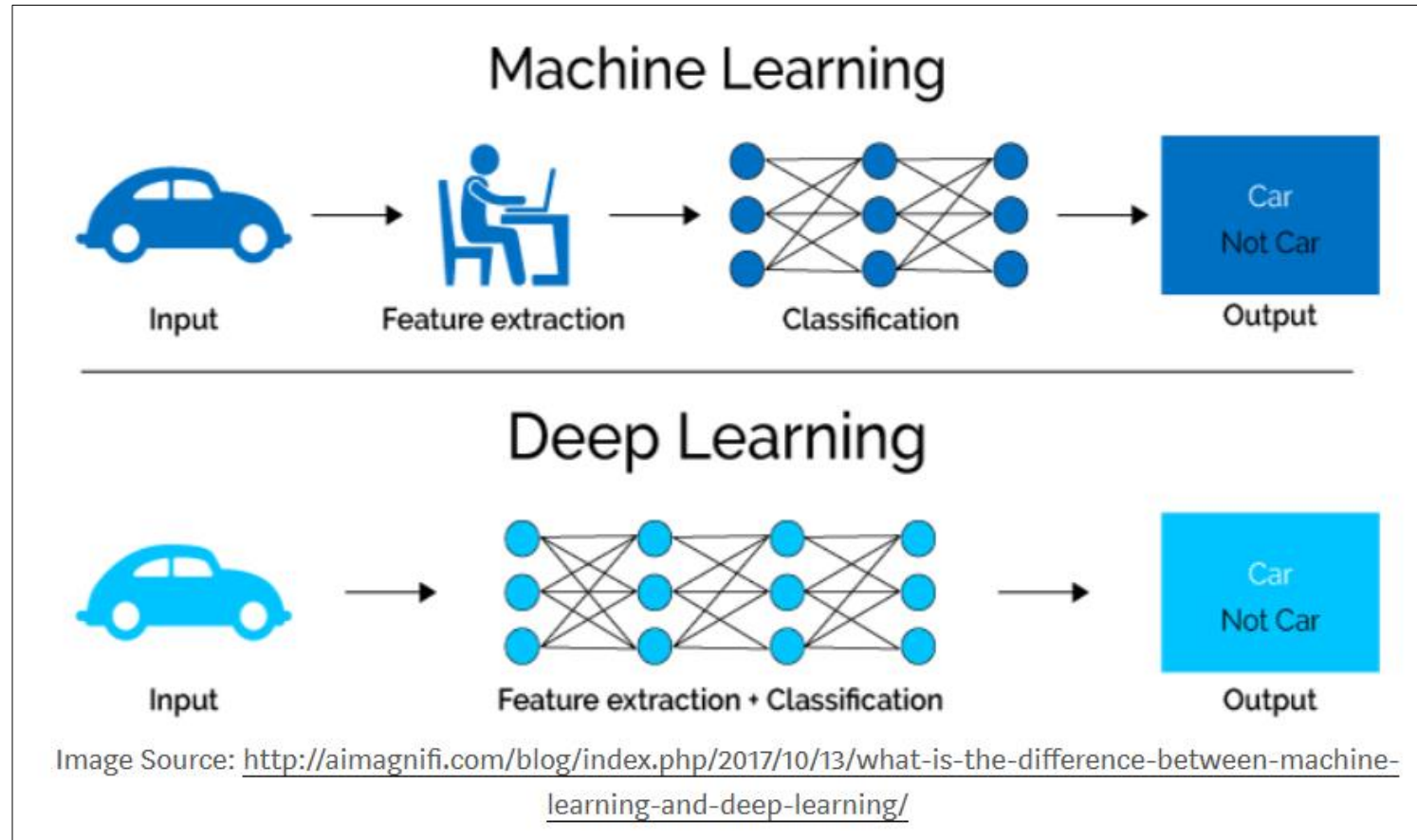
Machine Learning?

- What?
 - Datenaufbereitung
 - Datenanalyse
 - Datenverarbeitung
 - Erstellen analytischer Modelle
 - Bessere Entscheidungen

Übersicht

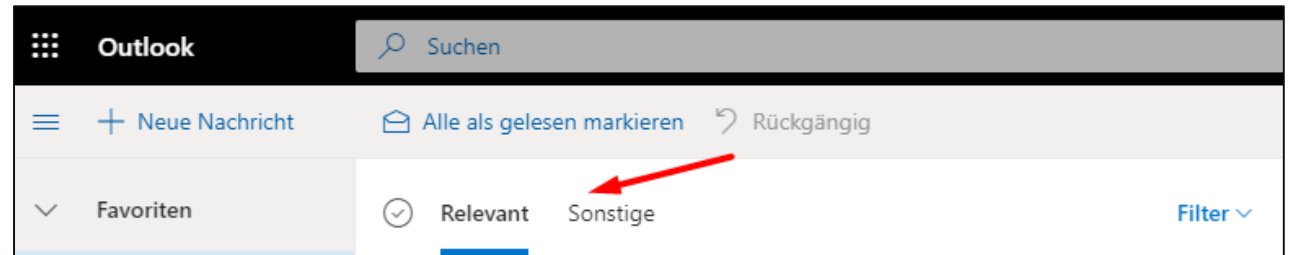


Übersicht



Übersicht

- Beispiele aus dem Alltag
 - Spam: Ja/nein
 - Outlook
 - Recommender Systems
 - Netflix
 - Amazon
 - Forecasting-Modelle
 - Flug buchen
 - Kundensegmentierung
 - Bonusclubs...
 - Siri, Alexa



Machine Learning?

- Was haben alle Modelle gemeinsam?
 - Fehler 😊
 - Darum ist die Beurteilung der Modellqualität so wichtig...



The image shows a comparison of two movie recommendations from a machine learning model. On the left is the movie 'LIVE BY NIGHT' (2016, 16+, 2 Std. 8 Min.) with a 66% Übereinstimmung (agreement) score. On the right is the movie 'STATE OF PLAY' (2009, 16+, 2 Std. 7 Min.) with an 83% Übereinstimmung score. A red arrow points from the 'LIVE BY NIGHT' card to the 'STATE OF PLAY' card, indicating a higher agreement score. Below each movie card is a short synopsis. The 'ÜBERS' logo is visible in the bottom right corner of the right card.

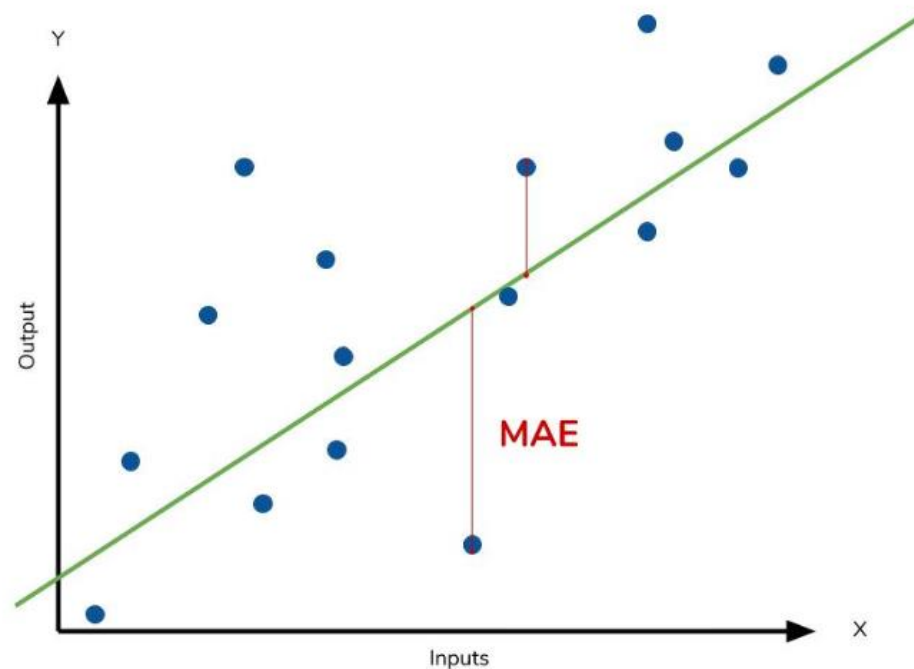
Movie Title	Year	Rating	Duration	Agreement Score	Synopsis
LIVE BY NIGHT	2016	16+	2 Std. 8 Min.	66 % Übereinstimmung	Der Sohn eines Polizisten arbeitet sich während der Prohibition durch Alkoholschmuggel an die Spitze von Floridas Unterwelt hoch. Doch er hat sich einen Feind gemacht.
STATE OF PLAY STAND DER DINGE	2009	16+	2 Std. 7 Min.	83 % Übereinstimmung	Beim dunklen Geheimnis eines Politikers geht es um Mord. Wie weit sollte man als Journalist für die Wahrheit gehen?

Ablauf einer ML-Aufgabe?

- Daten erheben
- Features vorbereiten
- Explorative Datenanalyse
- Modellbildung
- Modellqualität
- Modellbildung...

Modellqualität

- Trainings- vs. Test-Daten → Fehlermaße



Was macht der Nutzer in der Kassa?

- Problemstellung
 - Sehr viele Nutzer mit unterschiedlichem Nutzungsverhalten
 - Wer nutzt was?
 - Wer nutzt was nicht?
 - Was sollen wir entwickeln, damit mehr Nutzer noch mehr nutzen (mehr Mehrwert haben)?
 - Welchen Newsletter senden wir an wen?
 - ...

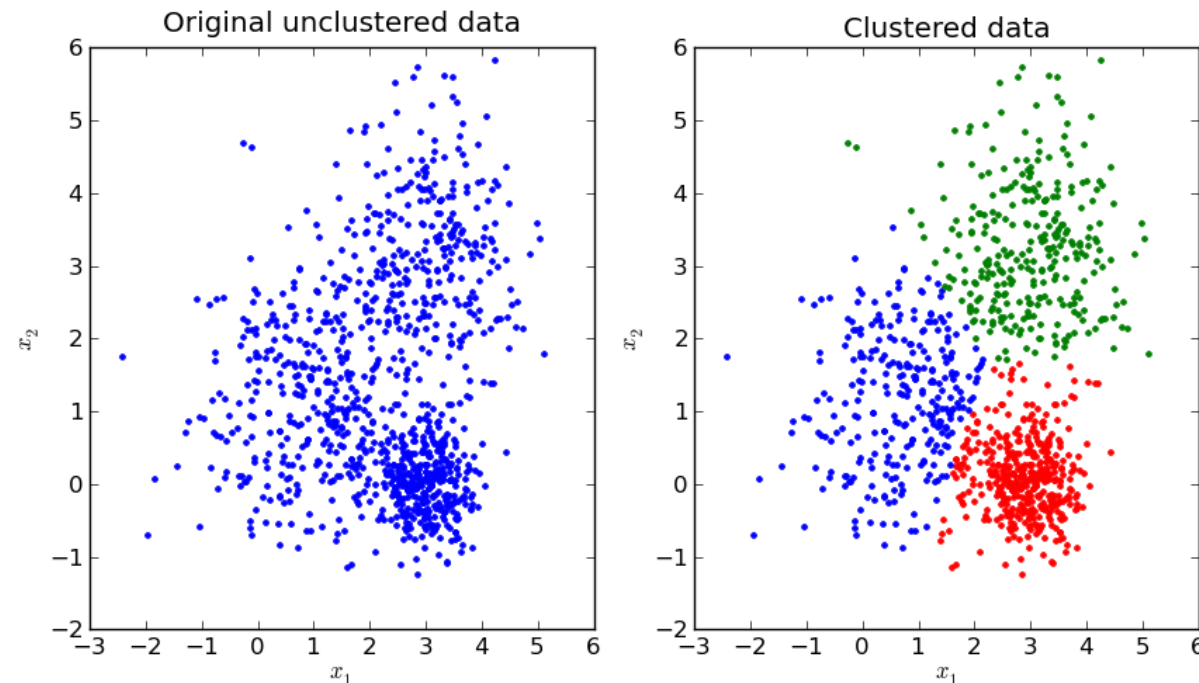
→ JÖ-Bonusclub → die stellen sich die gleichen Fragen

Lösungsansatz

- Man bildet Kundensegmente
 - Kunden mit ähnlichem
 - Nutzungsverhalten
 - Demographischen
 - Geographischen Merkmalen
- Ideale Situation
 - Jeder Nutzer ist ein Segment!
 - Herausforderung
 - Dann hätten wir viele tausende Segmente → das ist nicht handelbar...

Wie finden wir ähnliche Kundensegmente?

- ML-Modell
 - K-Means Clustering
 - Klassifikation (unsupervised)
 - zu welcher Klasse gehört welche Beobachtung?
 - Funktionsweise



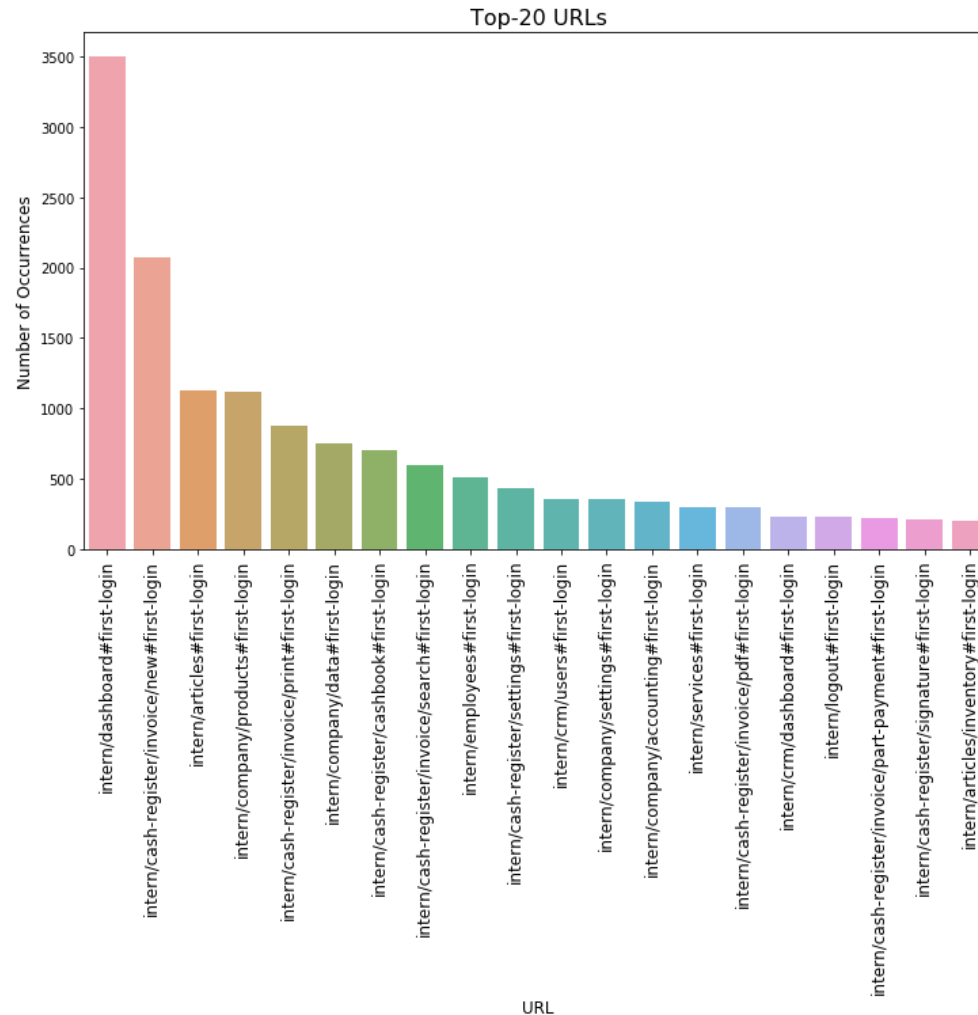
<https://towardsdatascience.com/k-means-data-clustering-bce3335d2203>

Wie finden wir ähnliche Kundensegmente?

- #1 - Data Cleansing
 - URL Aufrufe von Nutzern
 - Branchen
 - Akquisitionskanal
 - ...
- #2 – Feature Preparation
 - Wie müssen wir die Daten aufbereiten, dass wir diese verwendbar machen können?

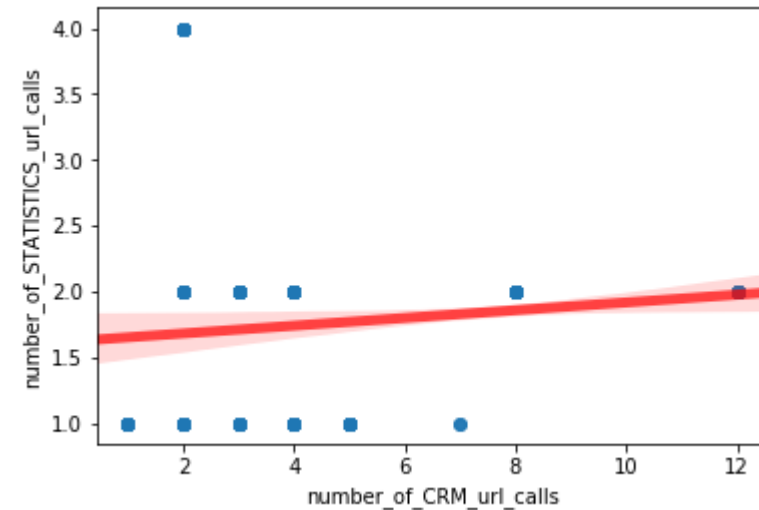
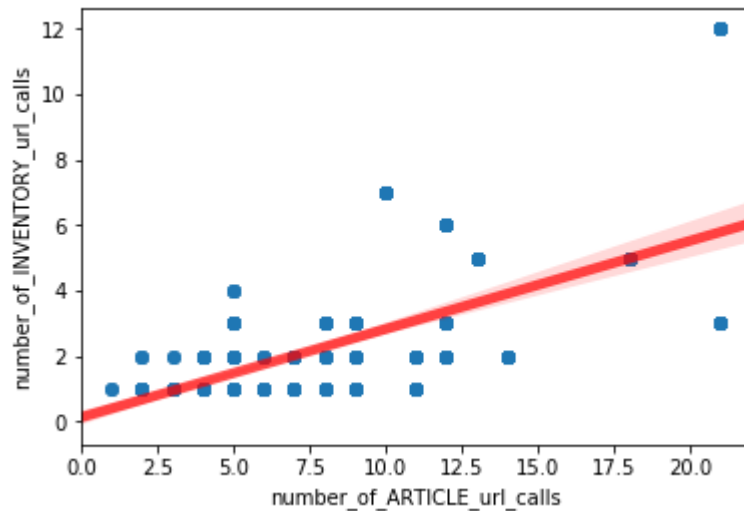
Wie finden wir ähnliche Kundensegmente?

- #3 – EDA

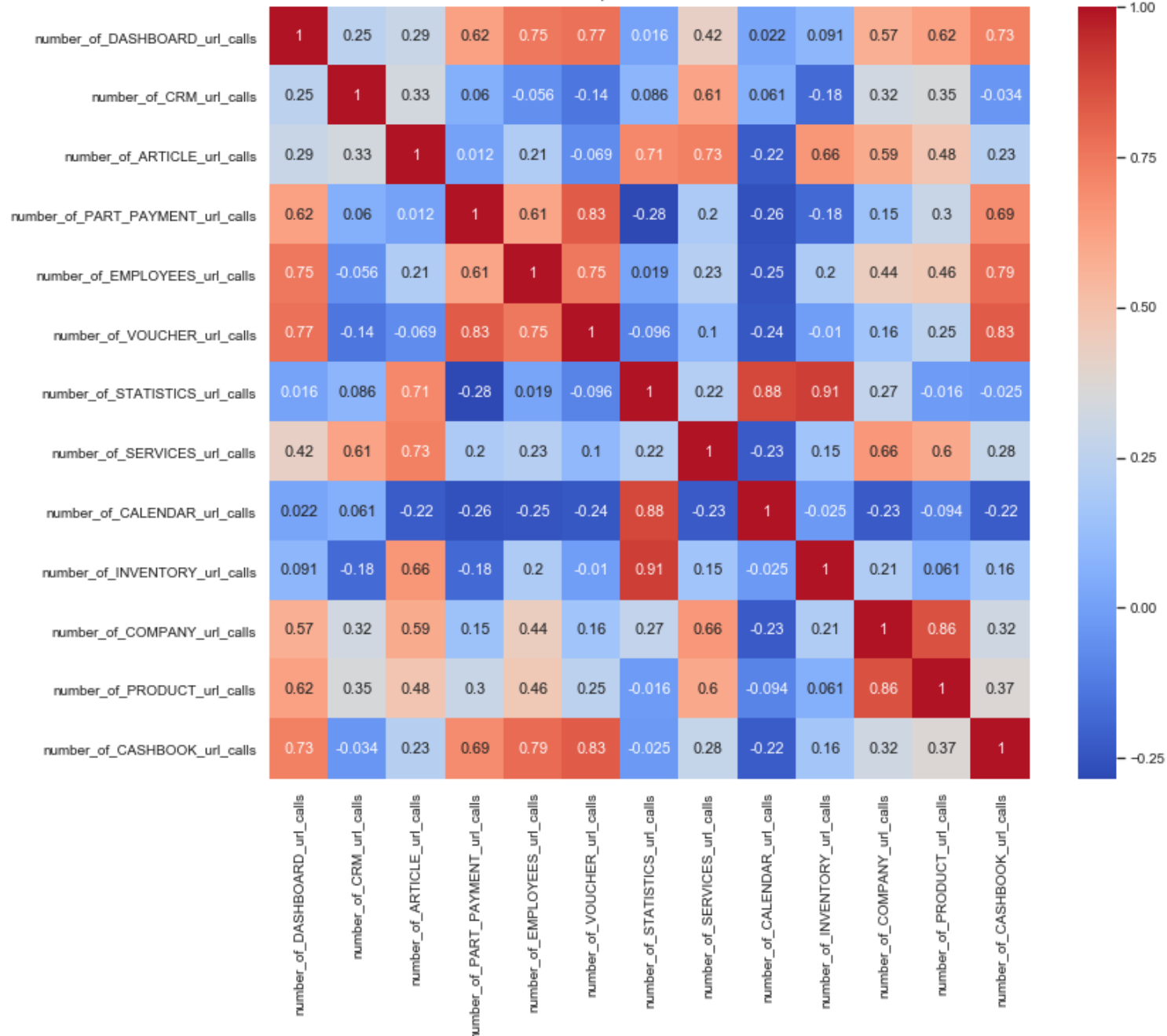


Wie finden wir ähnliche Kundensegmente?

- #3 – EDA



Correlation-Heatmap: FIRST-LOGIN URL-Movements



Wie finden wir ähnliche Kundensegmente?

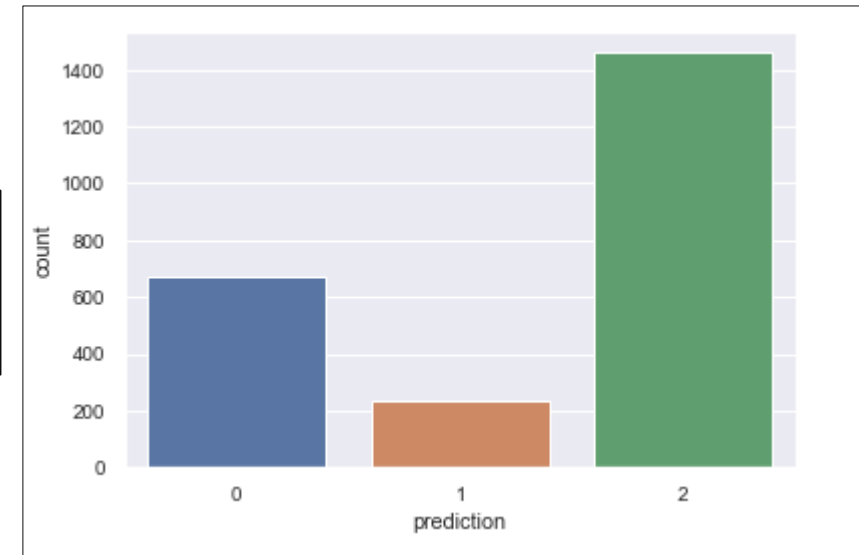
- #4 – Modellbildung

Datapreparation

```
# prepare dataframe
print(df_URL_movements_by_group_first_login_CORR.info())

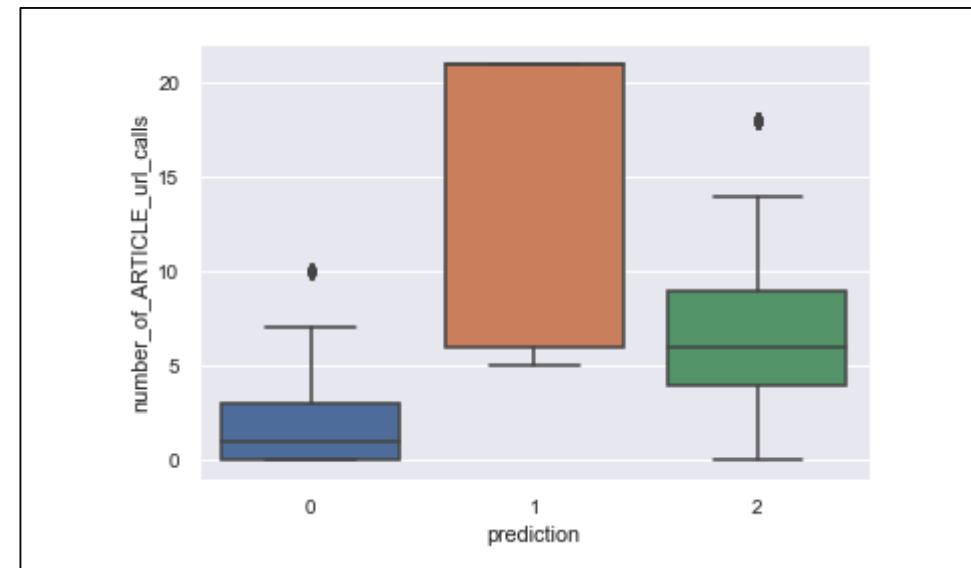
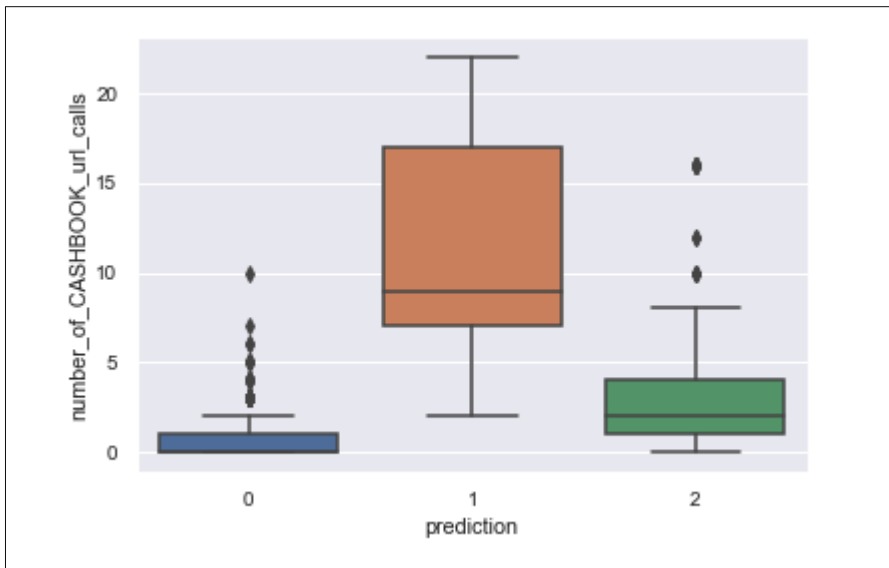
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2366 entries, 0 to 2365
Data columns (total 13 columns):
number_of_DASHBOARD_url_calls    2366 non-null int64
number_of_CRM_url_calls          1005 non-null float64
number_of_ARTICLE_url_calls      1721 non-null float64
number_of_PART_PAYMENT_url_calls 854 non-null float64
number_of_EMPLOYEES_url_calls    1187 non-null float64
number_of_VOUCHER_url_calls      698 non-null float64
number_of_STATISTICS_url_calls   397 non-null float64
number_of_SERVICES_url_calls     934 non-null float64
number_of_CALENDAR_url_calls     164 non-null float64
number_of_INVENTORY_url_calls    961 non-null float64
number_of_COMPANY_url_calls      1690 non-null float64
number_of_PRODUCT_url_calls      1496 non-null float64
number_of_CASHBOOK_url_calls     1224 non-null float64
dtypes: float64(12), int64(1)
```

```
from sklearn.cluster import KMeans
kmeans = KMeans(n_clusters = 3)
kmeans.fit(df_URL_movements_by_group_first_login_CORR)
kmeans.cluster_centers_
```



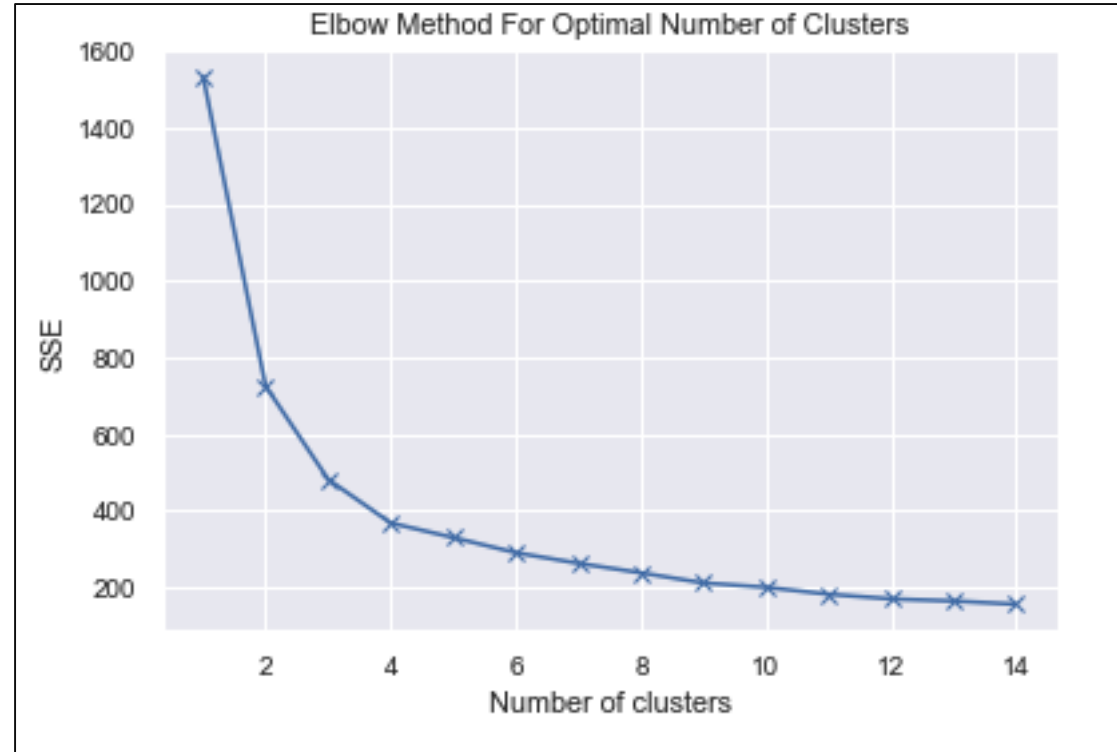
Wie finden wir ähnliche Kundensegmente?

- #4 – Modellbildung
 - Ergebnisse mit 3 Clustern
 - Verteilung der URL-Bewegungen pro Cluster



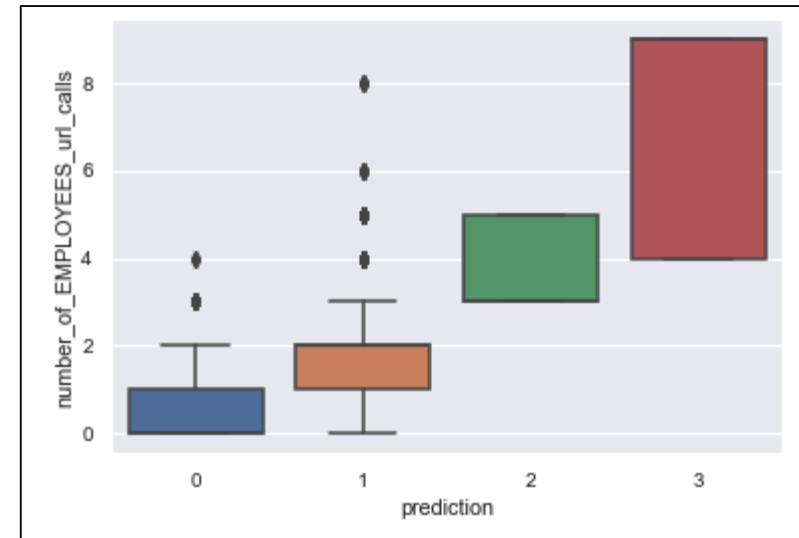
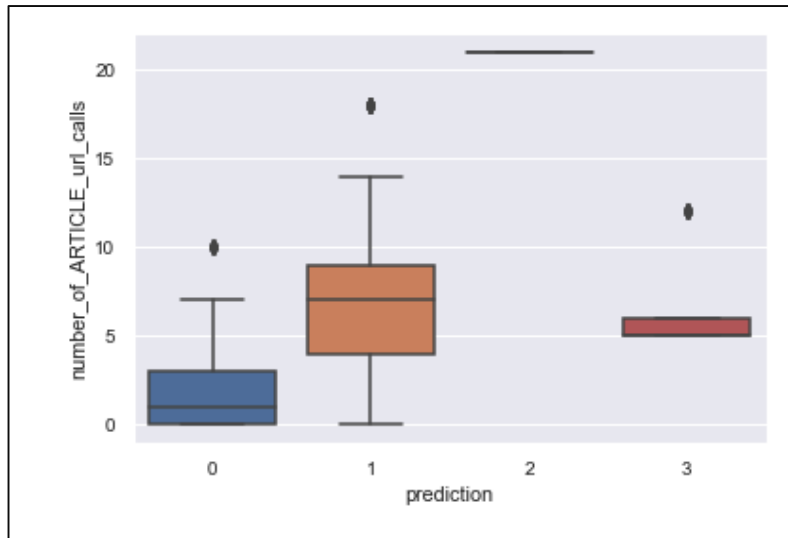
Wie finden wir ähnliche Kundensegmente?

- #4 - Modellqualität



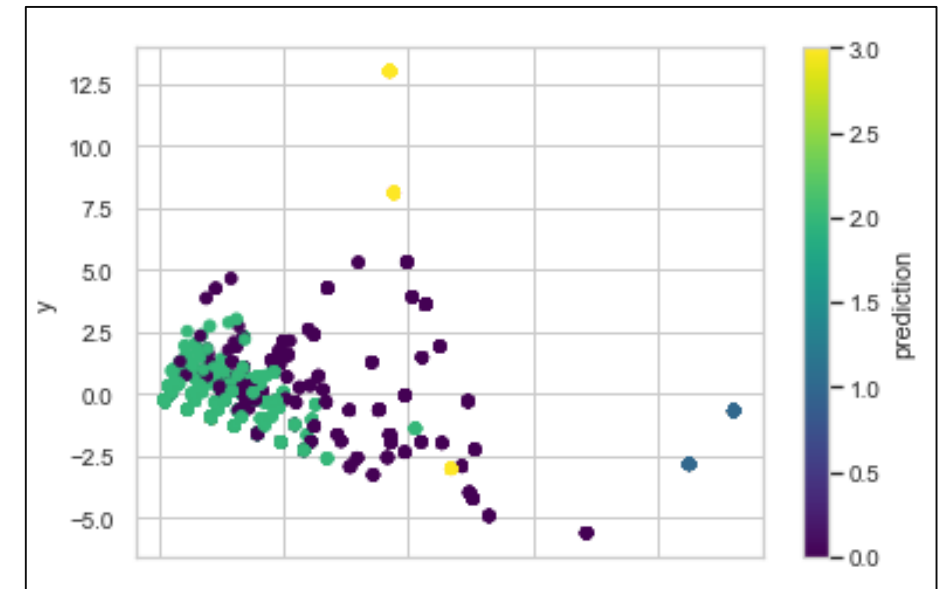
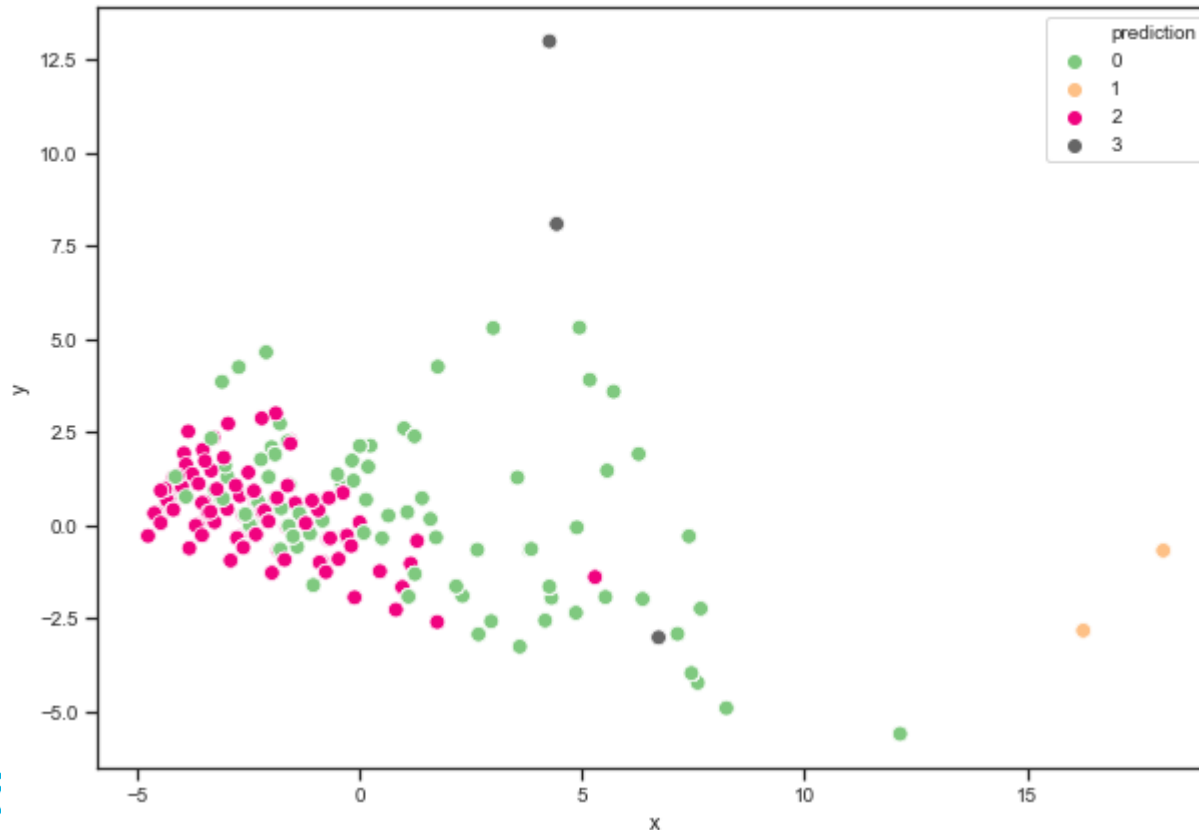
Wie finden wir ähnliche Kundensegmente?

- #4 – Modellbildung
 - Ergebnisse mit 4 Clustern



Wie finden wir ähnliche Kundensegmente?

- #4 – Modellbildung



Wie finden wir ähnliche Kundensegmente?

Cluster	Nutzungsverhalten	Clustergröße	Maßnahme
Heavy Users	jede Funktion einmal pro Woche
Heavy CRM-Users	jede kundenrelevante Funktion einmal pro Woche + Rechnungen
Invoice-Only Users	erstellen nur Rechnungen
...

Wie finden wir ähnliche Kundensegmente?

- Ergebnis & nächste Schritte
 - 4 Kundensegmente
 - ähnliche Nutzer innerhalb
 - Unterschiede zwischen den Segmenten
 - Tests, ob andere Variablen zu einem niedrigeren SSE (Fehlermaß) führen und gleichzeitig eine handelbare Clusteranzahl gefunden werden kann
 - Branche
 - Mitarbeiteranzahl
 - ...
- Ausreißerbereinigung

Wie viele Besucher kommen zum nächsten DEVTREFF?

- Erklärende Variablen (x_1, x_2, \dots)
 - Newsletter
 - Social Media Kanäle
 - Fans, Likes,...
 - Themen über die referiert wird
 - Außentemperatur
 - ...
- Erklärte Variable (y)
 - Anzahl der Besucher beim nächsten DEVTREFF

Wie viele Besucher kommen zum nächsten DEVTREFF?

- ML-Modell
 - Multiple Regression
 - Supervised Learning
- Ziel:
Anzahl DEVTREFF-Besucher = $y + x * \text{FB-Fans} - y * \text{Temperatur} + \dots$

Wie viele Besucher kommen zum nächsten DEVTREFF?

- Modell-Vergleich

Modell 1:

1 erklärende Variable: Anzahl Facebook Fans

Anzahl DEVTREFF-Besucher = $-20,3573 + 0,135614 * \text{FB_Fans}$

Durchschnittlicher Fehler = 52 Besucher

Modell 2:

> 10 erklärende Variablen: (siehe unten)

Anzahl DEVTREFF-Besucher = $-161,71 + 0,12401 * \text{FB_Fans} + \dots$

Durchschnittlicher Fehler = 18 Besucher

Hinweis:

fiktive und zuwenig Input-Daten

Modellvoraussetzungen?

Residuen-Analyse?

Ist Modell 2 klüger?

...

	Feature	Coefficients
0	temperature	-1.240108
1	Facebook_Fans	0.126521
2	Mostviertel_TV	-49.779601
3	topic_CSS	127.632475
4	topic_HTML	184.310695
5	topic_IOT	-14.751623
6	topic_JAVA	-175.201833
7	topic_KOTLIN	-145.359549
8	topic_LARAVEL	160.956605
9	topic_MYSQL	155.515596
10	topic_NUXT	218.047898
11	topic_PHP	165.517811
12	topic_PYTHON	65.210745

AUSBLICK & JOBS & DANKE 😊

JOBANGEBOTE BEI HELLOCASH

MARKETING- UND SALES-ASSISTENTIN GANZES INSERAT ANZEIGEN	SOFTWARE-ENTWICKLERIN GANZES INSERAT ANZEIGEN	MARKETINGASSISTENTIN (VZ/TZ) GANZES INSERAT ANZEIGEN
ONLINE MARKETING MANAGERIN GANZES INSERAT ANZEIGEN	WEBDESIGNER/IN GANZES INSERAT ANZEIGEN	weitere offene Jobs: <ul style="list-style-type: none">- Data Scientist- Support- ...

Bewerbungen an nina@hellocash.at